

Incorporating structural topic modeling into short text analysis

[結構化主題模型運用於短文本分析]

Po-Ya Angela Wang [王伯雅] and Shu-Kai Hsieh [謝舒凱]
National Taiwan University [國立臺灣大學]

The past few decades have seen the rapid development of topic modeling. So far, research has been more concerned with determining the ideal number of topics or meaningful topic clustering words than with applying topic modeling techniques to evaluate linguistic theories. This study proposes the Structural Topic Model (STM)-led framework to facilitate the interpretation of topic modeling results and standardize text analysis. STM encompasses various model training mechanisms, thereby requiring systematic designs to properly combine language studies. “Structural” in STM refers to the inclusion of metadata structure. Unlike the corpus-based keyness approach, STM can capture contextual cues and meta-information for the interpretation of topical results. Besides, STM can make cross-corpora comparisons via topical contrast, a challenging task for corpus-driven related models such as the Biterm Topic Model (BTM). Stylistic variations in song lyrics are taken as an illustration to show how to use the suggested framework to delve into the linguistic theory proposed by Pennebaker (2013). The topical model and iterable model in the proposed paradigm can clarify how pronouns affect style distinction. We believe the proposed STM-led framework can shed light on text analysis by conducting a reproducible cross-corpora comparison on short texts.

Keywords: structural topic modeling, biterm topic model, Chinese lyrics, corpus linguistics, keyness

關鍵詞: 結構化主題模型, 雙詞主題模型, 中文歌詞, 語料庫語言學, 關鍵詞

<https://doi.org/10.1075/cons1.22026.wan>

Concentric 49:1 (2023), pp. 96–138. ISSN 1810-7478 | E-ISSN 2589-5230



Available under the CC BY-NC 4.0 license. © 2023 Department of English, National Taiwan Normal University

1. Introduction

From the perspective of linguistics, in particular, a text is never “unstructured,” for sentences and discourse all have their own intricate internal structures. Texts, on the other hand, are unstructured data in the context of data processing. In the data science and text mining disciplines, it is conventional to describe texts as “unstructured” since they are not easily accessible to computers. For example, a text does not have a fixed structure (e.g., a fixed number of paragraphs, a fixed number of sentences, etc.), nor does it have a fixed vocabulary (e.g., the same word can have different meanings in different contexts), making it difficult for computers to easily approach it with simple algorithms. In recent years, corpus-based quantitative methods have been proposed to capture the essence of textual content in both supervised and unsupervised manners. Supervised methods require correct answers (labeled inputs) to facilitate model training, whereas unsupervised methods capture the hidden structure of data without human annotation. Using advanced machine learning techniques, for instance, we can automatically extract abstract information from large amounts of data to generate reading comprehension answers (Rajpurkar et al. 2016), analyze sentiment (Zhang et al. 2018), mine public opinion (Zhang, Wang & Liu 2018), classify documents (Yao, Mao & Luo 2019), or paraphrase texts (Shahmohammadi, Dezfoulian & Mansoorizadeh 2021). Most unsupervised methods usually depend on a vast quantity of data to discover patterns. Content from different text genres, however, pose limitations to these methods. Due to probable overfitting issues in machine learning approaches, social media posts of restricted word length, hashtags, emoticons, and other meta information are all likely to be underrepresented. Concerns about overfitting can be alleviated since topic modeling does not require large quantities of data to infer statistical structures. In this study, we leverage a recent unsupervised corpus method called the Structural Topic Model (henceforth STM, Roberts, Stewart & Tingley 2019) to form a STM-led framework to probe linguistic issues. This framework aims to construct a reproducible architecture for language research that can extract essential information from texts with a limited amount of data and document-level covariate information, yielding accountable classification results.

“Structure” has multiple senses in data mining, linguistics, and STM. In data mining, “structured data” refers to well-formatted, searchable, and quantifiable information. Categorical information (for example, a name or phone number) is a common structured data type (Snyder 2015). Qualitative data are typically viewed as unstructured data, for they are frequently stored in their native formats, including survey responses, social media posts, and blog articles. Texts are varied in their display formats, so in the field of data mining, they are deemed unstruc-

tured not because of their content but because of their format. To linguists, such qualitative texts are valuable for their internal structures. The linguistic structures include the following varieties: morphology, word order, syntax, and pragmatics. Topic modeling techniques can help efficiently capture linguistic structural information from a large corpus of texts without requiring human labeling labor. It enables unstructured text formats to be structured and accessible. In addition, the “structural” sense of STM signifies its incorporation of metadata structure in topic modeling. The metadata (for example, text genre, author, text length, time) are similar to the annotations for each text. In a manner comparable to supervised method, it provides contextual information to comprehend the interplay between texts and other variables. In spite of its evident advantages, STM does not specify how it can be systematically implemented in linguistic research. Therefore, a framework based on STM is proposed here to outline schematic training procedures that can be integrated with linguistic insights.

We will use song lyrics as a case study to illustrate the proposed framework, as lyrics are assumed to include delicate non-structured and structured features that are tricky to address via machine learning approaches: (a) the diverse content formats are due to arrangement considerations; (b) the amount of single singers’ works is unlikely to be substantial; that is, they may have just hundreds rather than thousands of songs; (c) various lyricists can contribute to the songs of a singer; and (d) lyric content is limited in length, full of repeated words, or composed using innovative word choices to shape conceptual themes. Applications, such as music recommendation systems or song popularity predictions, demand efficient and effective information retrieval from song lyrics. To achieve this, topic modeling can appropriately handle unstructured data formats and limited amounts of data. Lyricist distribution (how many different lyricists are involved), singer type (who sings the songs), or lyric length can be employed as metadata in STM to grasp how lyric topics may be affected by these factors. The theme of the contents is going to be highlighted through the proposed procedures. Lyrics are taken as an example to demonstrate how the proposed STM-led framework classifies lyrics based on linguistic hypotheses. In addition, other corpus-based measures such as *keyness* and the corpus-driven *Biterm Topic Model* (henceforth BTM, Yan et al. 2013) are compared to illustrate how the STM framework can facilitate the conducting of a reproducible corpus linguistic study capable of explaining textual heterogeneity between corpora.

2. Literature review

2.1 Lyrics and linguistics

Lyrics are a cross-disciplinary art form encompassing aspects from music and literature to culture and history (Eckstein 2010). With the help of the Multidimensional Analysis Tagger, Werner (2021) identified lyrics as an exclusive register compared to written or spoken registers based on n-gram and concordancing properties. Such connotations of prestige and genre prominence evince its diverse research possibilities from both qualitative and quantitative angles. Linguistic-oriented studies tend to analyze language styles and pragmatic implications in lyrics, such as corpus-based lyric style analysis (Kreyer & Mukherjee 2007; Petrie, Pennebaker & Sivertsen 2008), diachronic pragmatic strategies in lyrics (Sophiadi 2014), vocabulary size evaluation in lyrics (Tegge 2017), negation processing in lyric contents (Nahajec 2019), and figurative language used in lyric creation (Arifah 2016; Setiawati & Maryani 2018; Dewi, Hidayat & Alek 2020). Quantitative approaches focus more on applications, such as experiments regarding music and speech processing in the brain (Besson et al. 1998), lyric segmentation approach development (Baratè, Ludovico & Santucci 2013), topic modeling methods in lyric interpretation (Sasaki et al. 2014; Sterckx 2014; Laoh, Surjandari & Febirautami 2018), lyric generation assisting system (Watanabe et al. 2017), lyric emotion detection (Akella & Moh 2019; Devi & Saharia 2020; Sharma et al. 2020), music recommendation (Schedl 2019), and song popularity prediction (North, Krause & Ritchie 2020).

Both quantitative and qualitative approaches provide insights that attest to cognitive mechanisms beyond lyrics. With such insights, two research directions can be delved into more thoroughly. First, variables involved in lyric perception demand more attention. For instance, Barradas & Sakka (2021) indicate that lyrics can elicit negative emotions in participants from a particular cultural background. Additionally, a steady reduction in the complexity of lyrics over time is a manifestation of chronological changes in cultural settings (Varnum et al. 2021). The studies mentioned exemplify that different covariates can provide more insight into the mechanisms driving musical compositions. Second, it is hard to strike a balance between the qualitative and quantitative aspects of lyric research. For example, removing stop words such as function words is presumed for data pre-processing in quantitative analysis, while in qualitative studies, function words such as pronouns and quantifiers can offer linguistic insights and encode pragmatic implications. This is discussed in Pennebaker (2013). With the proposed STM-led framework, this study intends to invite more contextual information, meta-data variables and apply linguistic insights in classifying lyric themes, bene-

ficial to applications like song recommendation, author detection, popularity prediction, and multimodal data alignments.

2.2 Corpus-based approaches

Previous corpus-based studies on text analysis, such as on lyrics, novels, or emails, primarily adopted stylometry measures (Whissell 1996; Hoover 2007) (i.e., word usage, word length, word repetition, frequency-based cluster analysis), or used word-based content analysis resources and tools like Linguistic Inquiry and Word Count Analysis (LIWC) (Pettijohn & Sacco 2009), Jaccard N-Gram Lexical Evaluator (Jangle), and Wordsmith Tools (Wright 2014) to conduct emotion and author identification tasks. The majority of these statistical counts are *local*, limiting the ability to investigate the intricate interactions between words and documents; therefore, *global* statistics measures should be involved for an analytics scenario.

With global statistics, frequency can be leveraged in either a top-down or bottom-up manner. Gabrielatos (2018) proposed two primary frequency approaches: the top-down “focused frequency comparisons” utilizing predetermined wordlists to prove hypotheses, and the bottom-up “exploratory frequency comparisons” by recognizing terms for looking into texts.

The keyness measure is a widely used bottom-up global statistical measure for keyword analysis, which can be used either to characterize a genre or to recognize crucial beliefs in a target text. As a bottom-up approach, the keyness measure is suitable for comparing two corpora. It can realize “social, institutional, linguistic, and other factors which distinguish one culture from another” (Leech & Fallon 1992), as shown in their comparison of Brown and LOB corpora. Kilgariff (1997) also states that keyness can portray differences between two corpora and characterize an entire target corpus. The Chi-Squared test can sketch frequency difference (Aarts 1971). Pojanapunya & Todd (2018) indicate that log-likelihood (LL) (a probability statistic) can identify the general purposes of a genre, and odds ratio (OR) (an effect size statistic) can signify words for certain purposes, reflecting discourse level information of target texts, as discussed in Gabrielatos (2018). The primary calculation methods for capturing keyness are summarized below:

Chi-squared (χ^2) (Aarts 1971):

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Difference Coefficient (Leech & Fallon 1992):

$$(a-c) / (a+c)$$

Relative Frequency Ratio (Damerau 1993):

$$(a/c) / [(a+b)/(c+d)]$$

Log-likelihood Ratio (G^2) (Dunning 1993):

$$G^2 = 2 \sum_i O_{ij} \log \left(\frac{O_{ij}}{E_{ij}} \right)$$

$$E_{ij} = \frac{N_i N_j}{N}$$

Figure 1. Calculation methods for capturing keyness (Aarts 1971; Leech & Fallon 1992; Damerau 1993; Dunning 1993)

In this study, we take chi-squared (χ^2) and log-likelihood ratio (G^2) as keyness measures. In both formulas, the word differential between a target corpus and a reference corpus can be identified by calculating the observed (O_{ij}) and expected frequency (E_{ij}) of a word. Statistical evaluation measurements like chi-squared and log-likelihood consider samples random, which conflicts with the nature of language since language is by nature never random (Kilgarriff 2005). This study takes advice from Rayson (2019) not to adopt keyness values to decide lexical item significance but to obtain the keyword ranking as a comparison to the proposed framework.

2.3 Corpus-driven approaches

This section reviews a probabilistic corpus-driven methodology called topic modeling, an excellent method of exploring unstructured data from a global perspective.

2.3.1 Topic modeling

Topic modeling is an unsupervised clustering approach to infer hidden variables (topics) given what has been observed (documents). It is noted that *topic* here refers to groups of related words, which are estimated to cohere into different latent themes by identifying co-occurrence patterns within documents in the

way they co-occur within documents (Schweinberger, Haugh & Hames 2021). By employing this technique, the general picture of document collections can be clarified for further applications.

Topic word grouping works differently from clear-cut k-mean clustering in that the same words can repeatedly appear in different topics, and different documents can take more than one topic in topic modeling results. This reflects resonant feelings across topics and documents in natural language communication, which can be further applied to humanizing information retrieval tasks. To obtain such data representation and gauge applicable modeling results, topic modeling adopts different mechanisms to disclose and evaluate statistical structures within and across the documents. Wallach (2006, 2008) notes two main topic modeling types and proposed model evaluation methods (Wallach et al. 2009). Topic modeling models can be divided into bag-of-words-based models and local-linguistic-structure-based models. Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999), and Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan 2003) are bag-of-words-based probabilistic topic models, drawing inferences from word-level correlation while ignoring word order. N-gram language models reviewed in Chen & Goodman (1999) calculate statistical word distribution with surrounding lexical items, taking the local syntactic structure into consideration. To overcome potential constraints of previous models, Wallach (2008) incorporates word order, sentence-level structure, and document structure into topic variables to form “structured topic models.” These language-internal structure-oriented models are advantageous for capturing linguistic characteristics, yet contextual information is required for comprehending the modeling findings for future applications.

Despite not involving a syntax structure-oriented design like the structured topic models, the STM, as a variant of LDA, includes metadata structure to meet the contextual needs of topic understanding. In STM, syntactic independence and contextual consideration might still benefit various task requirements (for instance, text-to-image prompting or search engines). At first sight, lexicon-based STM, resting on word-level information instead of syntactic structure, may produce results with a lower degree of granularity. However, lower granularity is not synonymous with lower quality. The granularity level of a task depends on its purpose. Tasks like query engines, common trait analysis, or chatbot triggers need efficient retrieval of information from massive data. Faced with numerous texts, topic words can neatly represent the common ground of the texts, benefiting downstream applications. For instance, “I don’t love you” and “I want to love you” could both be classified under the topic of “love.” This syntax-independence processing way can facilitate information retrieval efficiency in coping with big data. Besides, STM can provide fine-grained analysis based on the general trends

of text collection due to the inclusion of metadata. Calculations concerning metadata enable researchers to detect specific differences in perspectives on the same topic. Namely, by integrating metadata structure into topic modeling, STM results can highlight how one topic is presented differently across documents using various lexical items. Such an advance comes from the employment of the correlated topic model (CTM) (Blei & Lafferty 2007), the Dirichlet-Multinomial Regression (DMR) topic model (Mimno & McCallum 2008), and the Sparse Additive Generative (SAGE) topic model (Eisenstein, Ahmed & Xing 2011), and spectral initialization (Roberts et al. 2013). With higher interpretability in qualitative analysis, STM has been widely applied in a broad range of investigations (Lindstedt 2019; Chen et al. 2020; Aranda et al. 2021).

In short, STM: (a) takes topic correlation and metadata into consideration; (b) incorporates multimodality (in statistical discussion) and big data issues into its methodological design (Roberts, Stewart & Tingley 2016); (c) provides more dimensions to evaluate the corpus. Such characteristics and quantities of interest can be visualized using the following graphical model, adopted from Roberts et al. (2013):

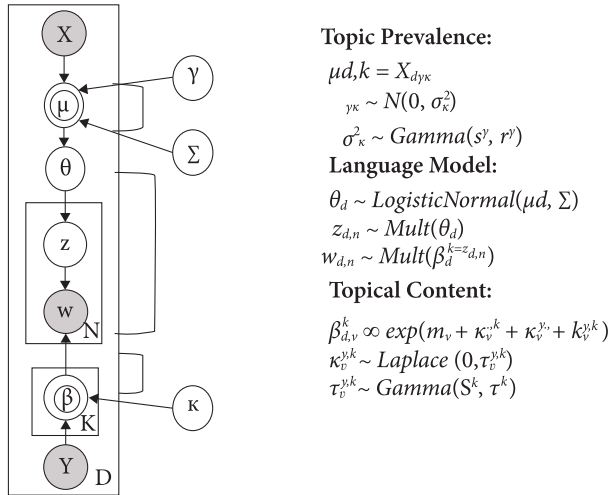


Figure 2. STM graphical representation

At the document analysis level, the document-topic proportion θ is of interest (Roberts et al. 2014). Corpus analysis level calculates topic-word proportions, topical prevalence covariate effects, and topical content covariate effects (Roberts et al. 2014). Topic-word proportions depict the contribution of each word to a topic. Topical prevalence refers to the association between a document and a topic (Roberts, Stewart & Tingley 2019). The topical prevalence covariate refers to the

metadata accounting for such an association. Topical prevalence covariate effects thereby signal the relationship between a document covariate X and the mean probability of every topic discussed in the document. Likewise, the topical content covariate explains topical content, defined as the words employed in a topic (Roberts, Stewart & Tingley 2019). Topical content covariate effects describe the relationship between a document covariate and the word frequency used in a particular topic.

As mentioned earlier, STM is a variant of LDA, a probability-based topic discovery model based on the occurrence frequency of terms. By understanding the architecture of LDA and the challenges it meets, the advances of STM can be illustrated. The LDA graphical model is shown in Figure 3, adopted from Blei (2012):

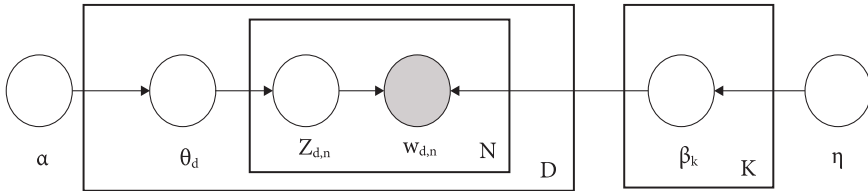


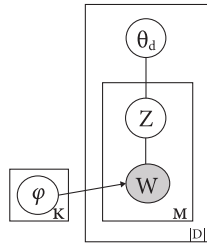
Figure 3. LDA graphical representation

θ_d indicates the probability distribution of topics within documents, and β_k shows the probability distribution of terms within topics K . Z represents a word assigned to the topic. These values assume that every document includes more than one topic but that a couple of topics are more dominant than others. Every word contributes in varying degrees to every topic. Words are ranked according to their representativeness in delineating specific topics. In other words, documents can overlap in specific topics, and the contribution percentage of each word in different topics is divergent. This characteristic distinguishes topic modeling from the cluster method, which does not tolerate overlap. “Document-topic proportions” and “topic-specific distributions” are inference tasks provided by topic models.

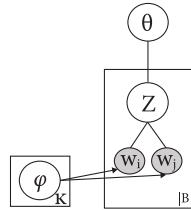
Difficulties arise, however, when an attempt is made to implement LDA to detect topics in short texts. The dependence on “document-level word co-occurrence” results in sparsity when the texts are short. Numerous follow-up LDA studies try to address this issue. The two most widely used approaches are data aggregation and data simplification. For the former, the tweets of a single user are viewed as a single document (Weng et al. 2010), or tweets including identical words are collected as a single document (Hong & Davison 2010). Nevertheless, such user-based or word-based aggregation is restricted in terms of data collection instability in that the number of tweets any single individual posts is inconsistent.

Concerning data simplification, this considers each short text has only one topic (Zhao et al. 2011). Yan et al. (2013) argued that this viewpoint is an oversimplification and has overfitting problems.

To tackle short texts, BTM is proposed (Yan et al. 2013). In contrast with “document-level word co-occurrence” in LDA, they advocate calculating the “global word co-occurrence” at the entire corpus level. To be specific, every document is so brief that the BTM approach treats them all as a single document: the corpus. In this way, what is calculated is not the topic word occurrence of a single document, but that of the entire corpus. The order of words is not taken into consideration in this strategy since the emphasis is on the general topic pattern. The graphical models to compare (a) LDA and (b) BTM are shown below, according to Yan et al. (2013):



a.



b.

Figure 4. Graphical representation: LDA vs. BTM

In BTM, the probability distribution of topics θ is within the entire corpus rather than within the documents. Unlike the document-based calculation in LDA, the biterns are generated for the whole collection. The concept of “bitern” is not equal to “bigram” in the sense that the former is a term composed of two words, while the latter is a term composed of two consecutive units. This bitern-pattern aggregation of the entire corpus resolves the sparsity concern. The BTM model obtains not the topic distribution per document but the topic distribution

within the entire corpus. The topic distribution of every single document can be extracted from the trained model. BTM seems to be an effective tool for capturing the topic trends of short texts. It has potential limitations resulting from a disregard for word order. In addition, it is hard to conduct cross-corpus comparisons with BTM. Similar to structured topic models (Wallach et al. 2009), it lacks contextual aids for interpreting modeling results. As opposed to BTM's focus on word co-occurrences, STM focuses on incorporating metadata information about corpus structure (Roberts, Stewart & Tingley 2016). By integrating document-level meta-information, STM possesses advanced interpretability in uncovering the relationship between covariates and the (latent) topics by including document-level meta-information.

Recently, with the tremendous advance in deep learning, Top2vec (Angelov 2020) and BERTopic (Grootendorst 2022) have emerged to enhance topic modeling results. With UMAP (Narayan, Berger & Cho 2021), Top2vec solves the sparsity problem and clusters document vectors. In this manner, topic number tuning in LDA is no longer needed, whereas the comprehensibility of the resulting topics is still not easy. Extending the Top2vec approach, BERTopic (Grootendorst 2022) incorporates context-based BERT embeddings and adopts TF-IDF to highlight more characteristic words in the resulting topics. Abuzayed & Al-Khalifa (2021) compared topic results from LDA and BERTopic, which suggests that the quality of the resulting topics remains unsatisfactory. Ebeling et al. (2021) propose a framework encompassing LDA and BERTopic. In their proposed framework, LDA evaluates topic coherence, and BERTopic selects the most illustrative ones from the LDA results. They attempt to enhance the interpretability of topics by leveraging *agglomerations* (to add similar groups together). Progress has been made with these techniques; however, they seldom give sufficient attention to how to interpret the results qualitatively.

The main weakness with current deep learning methodology is how to analyze the results from balanced quantitative and qualitative perspectives. STM, instead, is a contributing approach to overcoming such a disadvantage by considering the following three aspects: word contexts, topic correlations, and metadata information. The contexts of topics can be retrieved for further linguistic analysis through different wordlists. Several wordlist options are available in STM. *FREX* introduces words based on the frequency and uniqueness of the assigned topic. *Lift words* are chosen if they appear less in another topic cluster, which is advantageous to the research in that word choice innovation should be stressed in lyric creation. The *Score* value is similar to the Lift value, yet it adopts log frequencies. Documents corresponding to specific topics can also be retrieved to read between the lines. These perspectives assist the perceiving of topics more comprehensively. Topic correlation benefits to interpret any subtle stance divergence

toward the same topic. For instance, some texts describe the topic “love” from a positive angle, but others may highlight its negative features. With metadata, Roberts et al. (2013), in their exploration of political attitudes revealed by open-ended questions or media languages, have identified how categorical or numerical meta information of collected texts sheds light on the text mining analysis employed.

In short, STM serves the same aim as LDA and BTM, which is to establish a structured direction for unstructured data. In contrast to other approaches, it is more than an exploratory step as a result of its two advantages. First, STM incorporates contextual information (metadata). This provides insights into interpreting the resulting topics since the effect of the metadata variable on the resulting topics may be assessed statistically. Second, STM provides more accountable results than supervised methods owing to its diverse topic wordlist selection and topic correlation attributes.

STM can benefit linguistic analysis due to its inclusion of metadata structure, as it encompasses divergent training methods. How to employ models that can satisfy linguistic analysis purposes remains undefined. It requires linguistic insights to integrate various methods into organized assessment steps. Hence, we propose the STM-led framework for linguistic analysis and hypothesis evaluation. Unlike the original STM, the proposed framework systematizes steps to highlight contextual aspects (topic coherence and metadata structure) while analyzing findings. In every phase, linguistic considerations are incorporated. The results of BTM will also be presented as a comparison to highlight the capabilities of STM.

2.3.2 Model evaluation

Since topic modeling is an unsupervised, corpus-driven method, the evaluation of models becomes crucial for subsequent analysis. Such evaluation mostly focuses on finding optimal clustering topic number k , for the more appropriate this cluster topic number is, the more semantic comprehensibility the resulting topics own. In order not to cluster words arbitrarily or interpret results like “reading tea leaves” (Chang et al. 2009), quantitative strategies employed in the literature include *perplexity*, *semantic coherence* (as a coherence measure), and *exclusivity*.

In terms of perplexity, the lower the perplexity value is, the better the model’s performance. Traditional topic model training tends to take perplexity reduction as a way to ensure topic quality. For example, Hofmann (1999) adopted perplexity in evaluating the widely used traditional topic model, Probabilistic Latent Semantic Indexing (PLSA). The employment of perplexity was first challenged by Chang et al. (2009). They pointed out that models with better perplexity results do not have more comprehensible topics. Instead, they introduced “word intru-

sion” and “topic intrusion” from the perspective of human judgment to highlight the coherence of the resulting topics.

Corresponding to the attention given to coherence in Chang et al. (2009), Newman et al. (2010) adopted Pointwise Mutual Information (PMI) to evaluate the semantic coherence of resulting topics automatically. Aware that too many topic clusters can lead to senseless topics, Mimno et al. (2011:266) further improved the PMI evaluation approach to define topic coherence by emphasizing “the conditional probability of each word given each of the higher-ranked words in the topic.” In other words, each possible word on a topic is conditioned by other more likely ones. Mimno et al. (2011) defined topic coherence as formulated in Figure 5:

$$C(t; V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \left(\frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \right)$$

Figure 5. Topic coherence equation (Mimno et al. 2011)

For words v_m and v_l in the M possible words in topic t , the number of documents containing both v_m and v_l is divided by the number of documents having at least one token of word type v_l . The logarithm of the calculated result is added 1 in case the result is zero. Following this line of thought, Arora et al. (2013) specified that the *coherence* in Mimno et al. (2011) is to evaluate the within-topic semantic quality. Regarding *inter-topic similarity*, Arora et al. (2013) underlined that words in one topic cluster should have less overlap with the other resulting topics. Specifically, words within a topic should be as similar as possible; whilst words across topics should be as divergent as possible.

Coherence measures have evolved further. Röder, Both & Hinneburg (2015) proposed a framework for flexibly incorporating different coherence measures. The main measures¹ reviewed are: C_V , C_P , C_{UCI} , C_{UMASS} , C_{NPMI} , and C_A . C_{UCI} used by Röder, Both & Hinneburg (2015) refers to the calculation proposed by Newman et al. (2010), and C_{UMASS} means the *topic coherence* defined by Mimno et al. (2011). C_V , according to Röder, Both & Hinneburg (2015), is to integrate “indirect cosine measure,” “Normalized Pointwise Mutual Information (NPMI),” and “Boolean sliding window.” In the discussion of Röder, Both & Hinneburg (2015), C_V is the most accurate, and C_{UMASS} is the fastest. Trenquier (2018) compared C_V , C_{UMASS} , and word2vec coherence. The proposed word2vec coherence calculation in Trenquier (2018) assumes that words within a topic should be

1. For the reader’s reference, these measures are illustrated on this website: <https://palmetto.demos.dice-research.org/>

highly similar, yet the differences between topics should be as significant as possible. The former situation is termed “intra_topic_similarity” in Trenquier (2018:10), and the latter condition is referred to as “inter_topic_similarity” in Trenquier (2018:11). Both values are calculated. This provides semantic representation trained on Google News, which can provide an external angle of semantic similarity information. The extrinsic word2vec model is limited in tackling the “out of vocabulary” (OOV) issue owing to its embedding calculation method. Trenquier (2018) pointed out that some nonsense character combinations are assigned high values. This limitation implies that semantic coherence measures are still a better choice to gauge topic modeling quality. Different topic modeling packages choose different evaluation heuristics. Python Gensim package for LDA² can adopt C_V , C_{UCP} , C_{UMASS} , or C_{NPMI} to tune topic modeling results. The coherence evaluation method adopted in the Python Biterm topic model package is based on the C_{UMASS} coherence calculation.

As reviewed above, various semantic coherence measures are proposed to ensure the quality of intra-and-inter topic words. It is noted that such approaches can easily fail to address the “exclusivity” of topic words. That is, the yielded topic terms tend to be the same for more than one topic. Lack of exclusivity can lead to common words’ preponderance in resulting topics. Bischof & Airolodi (2012), critical of the tendency to depend solely on frequency or exclusivity, proposed Frequency-Exclusivity (FREX) calculation to reduce the inclusion of words occurring in every topic. Following this insight, Roberts et al. (2013, 2014) and Roberts, Stewart & Tingley (2016, 2019) also considered the frequency of words and their uniqueness to a topic by taking the weighted harmonic mean of the vocabulary’s ranking hierarchy. As they argued, exclusivity acknowledges both the frequency and uniqueness of a word, which can counterbalance possible frequency effects in semantic coherence results. The adopted formula (Roberts, Stewart & Tingley 2019) is shown in Figure 6:³

$$FREX_{k,v} = \left(\frac{\omega}{ECDF(\beta_{k,v} / \sum_{j=1}^k \beta_{j,v})} + \frac{1-\omega}{ECDF(\beta_{k,v})} \right)^{-1}$$

Figure 6. FREX equation (Roberts, Stewart & Tingley 2019)

In sum, the semantic coherence measure is maximized when the most probable words in a given topic frequently co-occur; while the exclusivity (in terms of FREX) measure is maximized when a topic includes many exclusive terms. In

2. <https://radimrehurek.com/gensim/models/coherencemodel.html>

3. In the STM R package (<https://cran.r-project.org/web/packages/stm/index.html>) we used in our experiment, the weight is set to 0.7 to support exclusivity.

addition, in the STM R package, other diagnostic metrics such as held-out likelihood (Wallach et al. 2009), which can indicate prediction ability of the model to unseen documents, and residual analysis (Taddy 2012) are also included. This can be incorporated at the exploratory stage in comparing the performances of different models to find out the most appropriate topic number k .⁴

3. A proposed STM-led analytics framework

This section introduces a proposed STM-led framework incorporating linguistic supervision, as schematized in Figure 7. In this way, topic modeling is no longer merely an exploratory step to sketch the structure of datasets but can be a comparison technique to provide statistical validity and accountability on linguistic assumptions even when the number of datasets is limited. The results of the *BTM* and *keyness* measures will be employed to justify the advantages of this proposed STM framework.

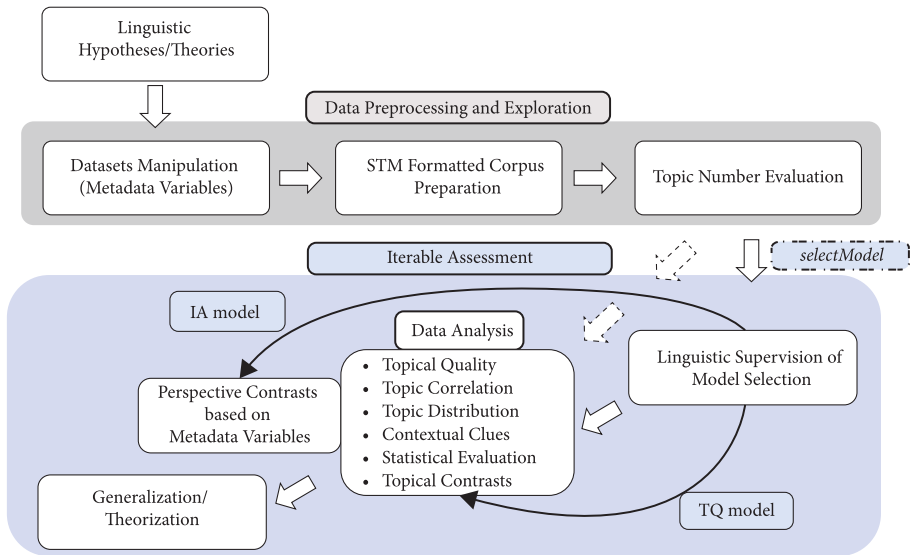


Figure 7. The proposed STM framework

4. The coherence evaluation method adopted in the Python Biterm topic model package is based on the UMass coherence calculation. In the STM R package, the employed evaluation metrics are semantic coherence (Mimno et al. 2011), exclusivity, held-out likelihood, and residual analysis. In the study, we follow the demonstration on Model search across numbers of topics in Roberts, Stewart & Tingley (2019) by focusing on the tradeoffs between semantic coherence and exclusivity to choose the model that meets our assumptions.

The framework falls into two stages: the Data Preprocessing and Exploration stage and the Iterable Assessment stage. In the first stage, the target metadata variables of the datasets are set based on a chosen linguistic hypothesis. These metadata variables encode contextual information related to the texts. After organizing the texts into formatted data, model evaluation diagnosis assists in choosing the appropriate topic classification number. Following this “Topic Number Evaluation” step, we can go through the dotted arrow to run the default STM spectral learning method.

Alternatively, to better interpret modeling results from a linguistic standpoint, the Iterable Assessment stage is taken. At this stage, two models are identified to handle linguistic oriented analysis: *Topical Quality model (TQ model)* and the *Iterable Assessment model (IA model)*. To comprehend topic coherence, the TQ model is trained in the following steps. At the “Linguistic Supervision of Model Selection” step, the *selectModel* function retains only 20% of models with the highest likelihood (Roberts, Stewart & Tingley 2019). In addition to ensuring the model’s performance, this is where linguistic supervision can step in. Based on 20% of preserved models, the exclusivity-semantic coherent distribution indicates how clustered words correspond to the clusters’ *topical quality* (exclusive or coherent). The internal cross-document structures (topic coherence) are disclosed in this step. Researchers can select clusters with a certain *topical quality* (highly exclusive or highly coherent) to compare how they perform in a follow-up data analysis. To facilitate data analysis, the TQ model provides the following information: topic correlation, contextual clues (topical qualities and various topic wordlists), topic distribution, statistical evaluation (estimated effects of the chosen meta-variables), and topical contrasts (words contrasting different topics).

To embrace a diverse range of contextual information, the chosen meta-variable(s) can be iterately assessed by training a new model, the IA model. The TQ model can highlight the diversities among different topics; the IA model is to pinpoint the topic contrasts resulting from specified metadata variables. The same steps mentioned in training TQ are repeated, yet the content covariate parameter in *selectModel* is specified with the chosen metavariable(s). The levels of the chosen variable are regarded as different *perspectives* in IA. From the viewpoint of the corpus, these predefined perspectives act as labels signifying different corpora. Along these lines, IA can identify how perspectives (different levels of the target variable) contribute to topical contrasts. This step illustrates that the proposed STM model is iterable to dig into how statistically assessed variables commit to topical differences. Through doing so, cross-corpus comparison is available. In the end, the proposed linguistic theories can be ratified via both statistical assessment and contextual analysis. The functions in the STM package can be assembled to constitute a linguistic-oriented framework. Each step in the

proposed framework is paired with a sophisticated linguistic consideration. This makes for a well-balanced STM framework based on language.

Machine learning methods applied in lyric studies usually aim to conduct sentiment analysis (Xia, Gu & Lu 2019) or lyric generation tasks (Wang & Zhao 2019). While STM has been widely adopted to detect topics in education research, social movements, or management discourse (Lindstedt 2019; Chen et al. 2020; Aranda et al. 2021), how to apply linguistic insights to distinguishing lyric topics from different singers is rarely approached. The integration of STM practice and lyric studies in the proposed framework could contribute to areas such as song recommendation systems, author detection, and multimodal analysis. For example, instead of looking for keywords in a song title, listeners could benefit from finding songs by typing a single topic word. To attain such topic query efficiency, a model capable of comparing and contrasting songs by various performers is required. The classification tasks in the following case study section will demonstrate using the proposed framework to draw distinctions in lyric style according to linguistic theories. The particular linguistic focus in this case study is the function words. In text-mining and NLP tasks, function words are frequently overlooked in pre-processing steps, while Pennebaker (2013) argued that function words, such as pronouns, quantifiers, or propositions, can capture language users' styles, so they are "style words." Pennebaker (2013) calculated the style matching degree of two given texts by evaluating their function word rates. This Language Style Matching (LSM) measurement is a *top-down* method that focuses primarily on how often function words are used, not how they are used in context.

$$1 - \frac{|\% \text{ Person 1's pronouns} - \% \text{ Person 2's pronouns}|}{\% \text{ Person 1's pronouns} + \% \text{ Person 2's pronouns}}$$

Figure 8. Language style matching (LSM) measurement (Pennebaker 2013)

This study aims to revisit the same style difference issue from a *bottom-up* view. Instead of searching for predefined style words, setting preset style-difference meta-variables enables the proposed STM-led framework to identify words contributing to styling contrasts in topical contexts. The topical contexts of this corpus-driven method comprise topical qualities (exclusive or coherent) and topical meanings (four types of topic wordlists). Incorporating topical contexts offers contextual clues to clarify the role style words play. As Pennebaker (2013: 22) pointed out, "A good way to think about style words is that, by themselves, they really don't have any meaning to anyone." In this manner, the relationship among style-difference meta-variables, word types, and topical contexts

can be statistically evaluated to understand the style word theory proposed in Pennebaker (2013).

In the next section, lyrics are used as a case study to demonstrate how the proposed STM architecture outperforms existing approaches (*keyness* measure and *BTM*). The framework applied to lyric analysis will be examined via TQ and IA models. The TQ model demonstrates how contextual clues facilitate result analysis and how style-difference meta-variables are related to topical contexts. The IA model identifies words contributing to style differences based on topical contrast results. With these two models from the proposed STM framework, we can expound on the style difference issue from two perspectives: (a) How do style differences interact with topical qualities? (b) What are the predominant style words for capturing style differences?

4. Lyrics analytics as a case study

This section demonstrates how the proposed framework assesses assertions proposed in Pennebaker (2013). Results from *BTM* and *keyness* will be compared to prove the interpretable and systematical advantages brought by STM in linguistic analysis.

4.1 Data pre-processing and exploration

Word pre-processing in short texts plays a vital role in subsequent analysis. Trenquier (2018) indicates that appropriate pre-processing processes can benefit topic modeling results. The stemming step can increase the statistical significance of words' semantic importance, and the part-of-speech tagging step can decrease the distracting influence of function words. Since in this research short texts are Chinese songs, the stemming step was not required. The data was only cleaned by removing non-Chinese words and possible crawling errors (e.g., the lyricists' or composers' names).

CKIP Tagger (Li, Fu & Ma 2020)⁵ was utilized further for word segmentation and part-of-speech tagging. Nouns and verbs were chosen from the segmented results. Pronouns (Nh), numbers (Neu), and classifiers (Nf) are function words kept in the corpora, for they are the highlights of the linguistic hypothesis (Pennebaker 2013) targeted here.⁶ Other content words, verbs and the remainder

5. <https://github.com/ckiplab/ckiptagger>

6. POS tags in CKIP tagger are listed in <https://github.com/ckiplab/ckiptagger/wiki/POS-Tags>

of the nouns, are retained as they give contextual clues for subsequent comparison and analysis. The data sets are taken from the work of two popular Taiwanese singers: Jolin Tsai⁷ and Cheer Chen.⁸ The lyrics of Tsai and of Chen were crawled from the Mojim website.⁹ The redundant data was removed, but the suite (to combine several individual songs into one single piece) was kept in that putting together the lyrics from several songs into one song is regarded as a new creation. Lyrics of the same song but formatted differently and appealing on different albums were also preserved. The data sets comprised of a total of 306 songs from Tsai and 122 songs from Chen. The song title, lyricist, and composer were separately saved in different columns. The two datasets were aggregated into the Combination dataset to go through the proposed topic modeling framework. The 20 most frequent words in each data set are shown in Figure 9.

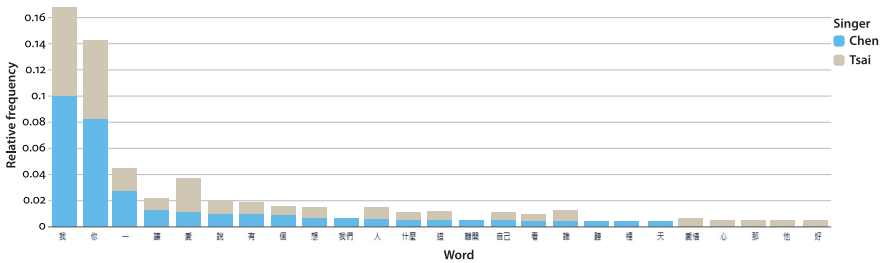


Figure 9. 20 most frequent words in the 2 datasets

We selected Tsai and Chen based on the linguistic insights involved: stylistic distinctions. To control the comparison variable, we selected vocalists of the same gender but different song characteristics. If we had wanted to evaluate different linguistic theories with the STM-led framework, we would have had to choose different subject types. In performance, Tsai usually combines singing with dancing, which is quite dissimilar to Chen. According to Spotify statistics,¹⁰ both artists belong to “taiwan pop,” “mandopop,” and “taiwan singer-songwriter” genres. However, the two singers also have different style tags. Tsai is labeled as “singaporean mandopop,” “mainland chinese pop,” and “taiwan idol pop,” which differ from Chen’s style labels (“chinese indie,” and “taiwan indie”). To revisit

7. https://en.wikipedia.org/wiki/Jolin_Tsai

8. https://en.wikipedia.org/wiki/Cheer_Chen

9. <https://mojim.com/>

10. Every Noise at Once is a listenable visualized acoustic music style website, utilizing information gathered and processed by Spotify as of 2022-12-01 for 5,987 genre-shaped differences. (<https://everynoise.com/>)

the style distinction hypothesis proposed by Pennebaker (2013), the metadata variables set in this study concern language style differences rather than singing styles. The stylistic difference defined here encompasses both non-verbal and verbal aspects: lyric length and lyricist distribution. Lyric length is a possible non-verbal factor in sketching different styles. Different lyricists may lead to diverse verbal styles in terms of word choice and topic shaping. Tsai's songs are composed by different lyricists, whereas Chen's songs have a pretty homogeneous lyricist distribution. This can be seen in Figure 10. Thus, the lyricist distribution disparity is treated under the *singer type* heading with two levels: Tsai and Chen. Such a singer-type comparison is equivalent to comparing the corpora of Tsai and Chen.

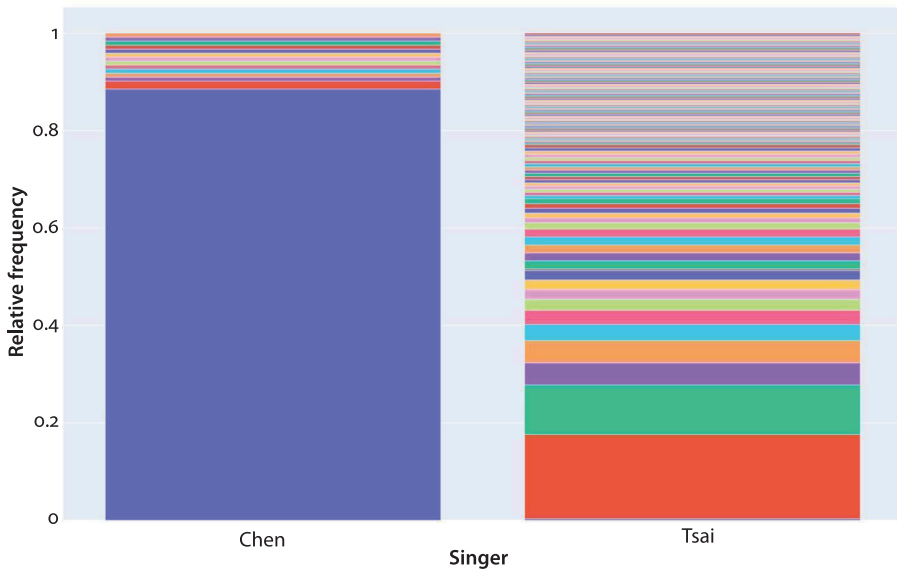


Figure 10. Lyricist distribution in Tsai and Chen

With the BTM's¹¹ word-occurrence correlation calculation, our remarks on the verbal style differences contributed by *singer type* (lyricist distribution) can be quantified. As reviewed in Section 2.3, the BTM is a legitimate generative model that can learn topics from short texts by attending to the correlation of word-occurrence patterns in a corpus. By employing this correlation idea in the BTM, we could examine the correspondence between word choices and lyricist diversity. Figure 11 depicts the correlation between the top 60 bi-terms in each dataset,

11. We used both BTM R and Python packages for visualization and evaluation in our study. The complete codes are available at: https://github.com/diff94/STM_shorttext

excluding zero correlations. This connection strength implies a disparity of word choice in Tsai, where the lyricist distribution is diverse. Strikingly different from Tsai, the words within Chen are highly correlated. It can be inferred that the diversity of lyricists may relate to the terms’ correlation within a dataset: words employed by Chen may be highly consistent across different songs. This discrepancy between the two singers supports choosing *singer type* as the covariate of style differences.

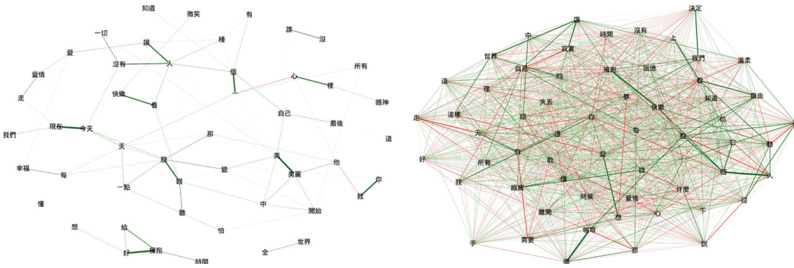


Figure 11. Top 60 term correlations in 2 datasets: The one on the left is from Tsai, and one on the right is from Chen

After determining the metadata covariates (*lyric length* and *singer type*), we constructed a formatted corpus and selected the initial topic number (cf. Figure 7). The Combination data was adopted to construct an STM formatted corpus with *lyric length* and *singer type* as metadata variables. The resulting corpus has 428 documents, 3483 terms, and 25911 tokens. 2632 terms were removed from 6116 terms.

To select the appropriate topic number k , the diagnostic values for topics ranging from 4 to 100 are shown in the left part of Figure 12. The semantic coherence looks appropriate between models with 4 and 16 topics. In the exclusivity-semantic coherence distribution plot, the points are annotated with topic labels, and topics from the same models are marked with the same color. The models with 12 or 16 topics tend to have higher exclusivity. The model with 4 topics has one outlier with relatively lower semantic coherence. The model with 8 topics has all the topics centered around higher semantic coherence and medium-level exclusivity. Thus, $k=8$ was chosen for the latter model training.

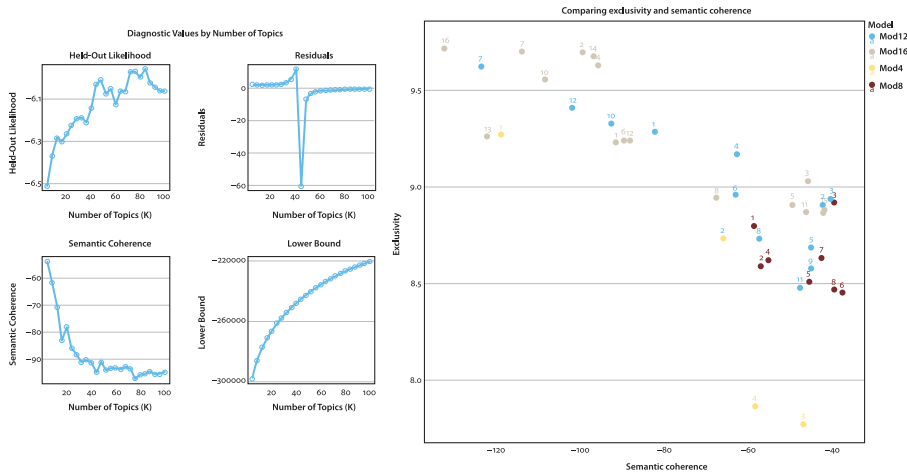


Figure 12. Diagnostic values by “number of topics” and exclusivity-semantic coherence distribution

4.2 Iterable assessment

4.2.1 Linguistic supervision of model selection

The topic number evaluation in the previous step is similar to what BTM does when calculating semantic coherence. As shown in the left plot of Figure 13, without setting the background topic, the first stark peak of the Combination dataset is when $k=12$, after which the trend obviously drops. As for setting the first topic as the background topic, except for iteration equals to 40, the peak is reached when $k=12$ in other iteration situations.

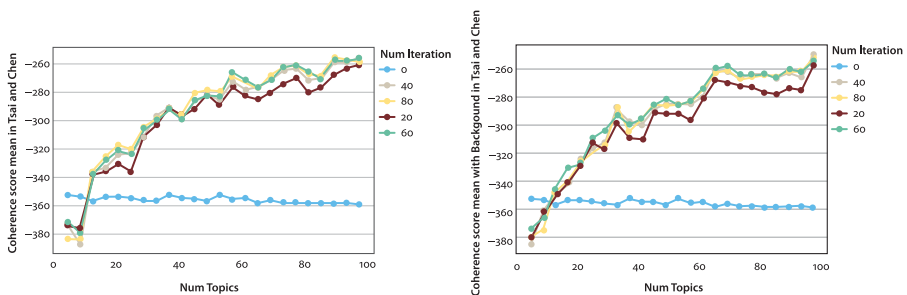


Figure 13. Topic number evaluation in BTM

Compared with BTM, STM works differently in inviting exclusivity, which provides room for linguistic supervision insights, especially when the *selectModel* function is adopted. After running *selectModel* with $k=8$, four models were pre-

served. In Figure 14, models 2 and 3 both have higher exclusivity and semantic coherence in overall topic distribution. Model 3 was chosen to be the TQ model because it includes both the topic with the highest exclusivity and the topic with the highest semantic coherence. Style differences between the highly semantically coherent topic and the highly exclusive topic could thus be investigated.

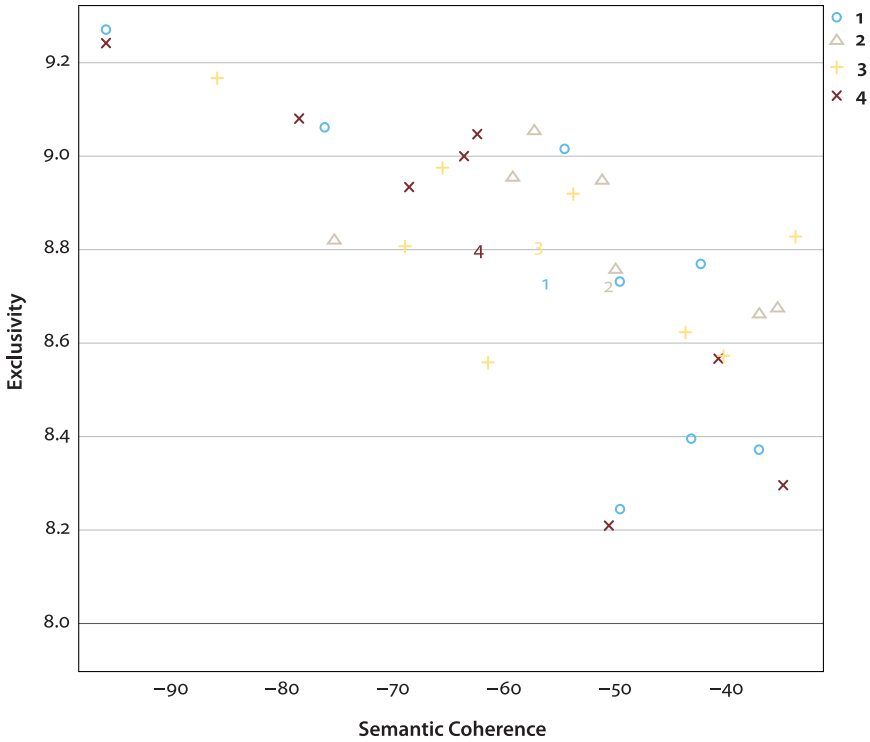


Figure 14. Topic evaluation for model selection in STM

4.2.2 Topical quality model (TQ model)

In this section, we further explore the way style differences interact with topical qualities. To demonstrate how the proposed STM framework better illustrates this question than keyness and BTM, the highlighted word clusters in keyness and BTM are shown in Figure 15. Based on previous literature reviews on global statistics measures, the keyness wordlist highlights uniqueness across corpora. The *ni* ‘female you’ and *ai* ‘love’ are highlighted in Tsai. In contrast, *wo* ‘I’ and *ge* ‘song’ are indicative of the uniqueness of Chen’s lyrics.

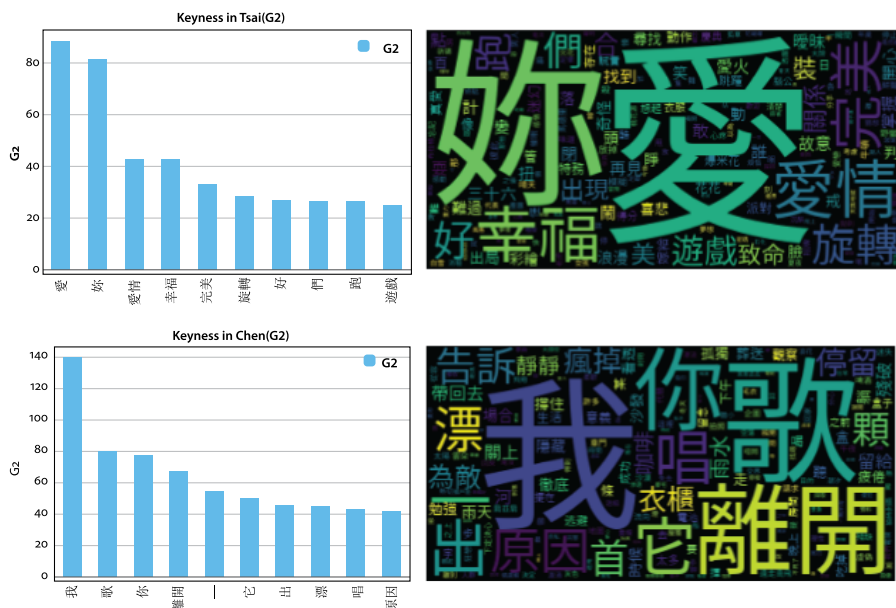


Figure 15. Keyness wordlists

In contrast to keyness, the resulting topic words from BTM in each singer’s corpus denote the main themes in the individual corpus. The 15 most frequent topic words in each dataset are determined. “Self,” “without,” “love,” and “what” are shared topics in these two corpora.

Chen with setting background topic

topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7
雨水	世界	原因	我們	什麼	我們	目的	離開
懂懂	所有	離開	告訴	時間	喜歡	希望	寂寞
我們	失去	什麼	離開	靜靜	回家	陌生	時候
沙發	自己	欣賞	最後	生活	想要	什麼	留給
崇拜	為數	表情	看見	知道	冬天	關上	咖啡
塵埃	什麼	舞台	停留	以後	靜靜	觀察	一點
自己	這樣	多少	溫柔	沒有	今天	帶回去	告訴
下午	太陽	雨天	生命	太多	發現	卻還	自由
決定	需要	許多	天空	自己	哪裡	離開	忍不住
盒子	獲得	回憶	備香	開始	衣櫃	離開	自己

Tsai with setting background topic

topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7
愛情	我們	旋轉	改變	愛情	什麼	我們	愛情
遊戲	什麼	所有	世界	完美	自己	想說	天使
自己	沒有	電話	完美	特殊	愛火	今天	幸福
三十六	自己	寫著	再見	牛頓	致命	自己	那些
關係	天空	臉公	一點	奧爾	動作	愛情	消極
什麼	喜歡	爆米花	迷幻	定律	愛情	昨天	遊戲
保持	離開	從前	開始	消失	未來	幸福	等候
掌握	時間	跳躍	花花	自己	世界	關係	以後
出現	世界	什麼	女孩	擁抱	暴力	明天	騎士
想念	最後	回到	愛情	放掉	現在	慢慢	星星

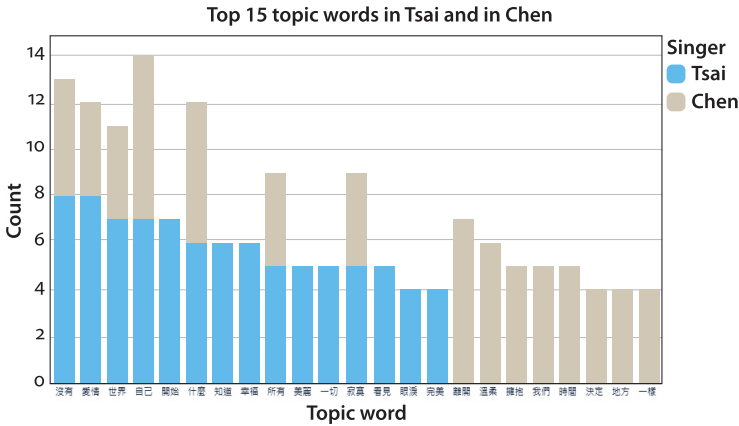
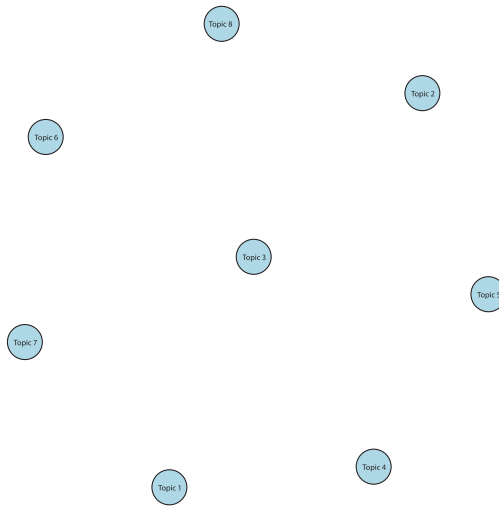


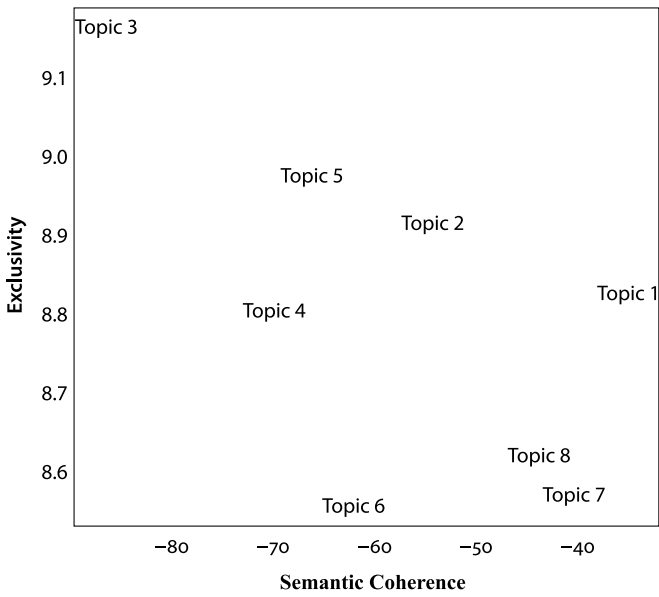
Figure 16. BTM topic wordlists

Comparing Figure 15 and 16, we see that keyness is advantageous in sketching differences between the target and reference corpora, but it cannot provide contextual information to interpret the result. BTM, on the other hand, presents topic clusters for each corpus while failing to compare and contrast across corpora. Consequently, neither would be suitable to detect style variations in the corpora. STM, in contrast, provides information about cross-topic linkages, contextual clues (topical qualities and various wordlists), topic distribution, statistical evaluation, and topical contrasts to facilitate data analysis.

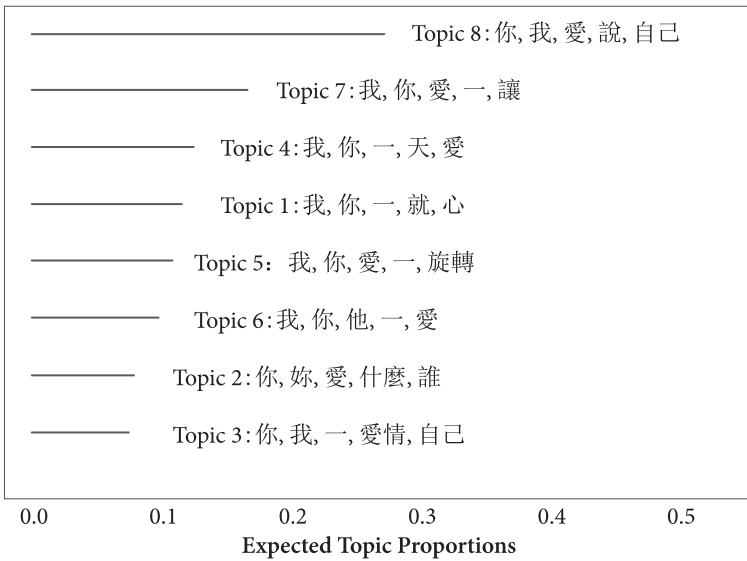
Instead of listing words, STM clarifies the relationship among each topic as well as offering contextual information. In Figure 17, topic cluster coherence degrees, diverse calculation-based wordlists, statistical evaluation, and topical contrasts exhibit different aspects of expounding the topic modeling results. The resulting 8 topics from the TQ are clear-cut without correlation. Of these 8 topics, the most exclusive but least coherent is topic 3, whereas the more semantic coherent are topics 1, 7, and 8. The most probable wordlists denote the dominance of pronouns in every topic. Topical qualities and the FREX, Score, and Lift wordlists provide more interpretative clues to the resulting topics. The topic words most associated with each topic cluster are selected based on calculated values, and the most widely distributed topic cluster is placed at the top. Compared to other topics, the most exclusive topic 3 in the TQ is explained by its innovative word usages, *sanshiliu* ‘thirty-six,’ shown in the FREX and the Score valued words.



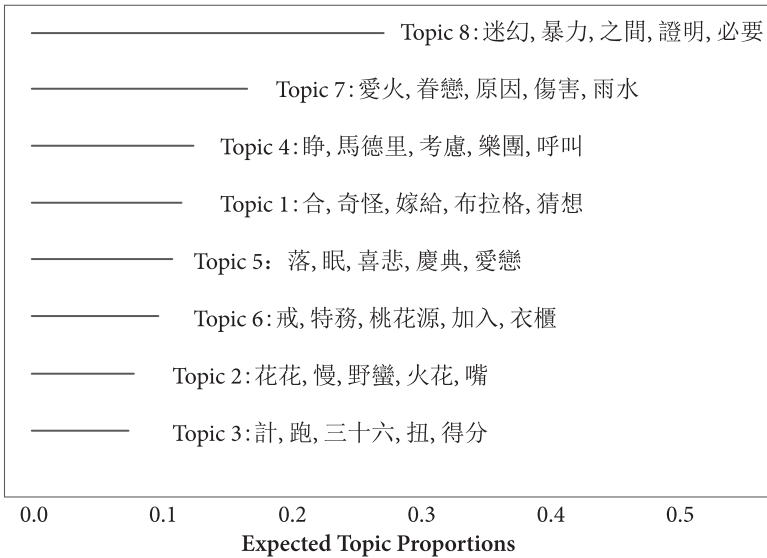
a. Topic correlation



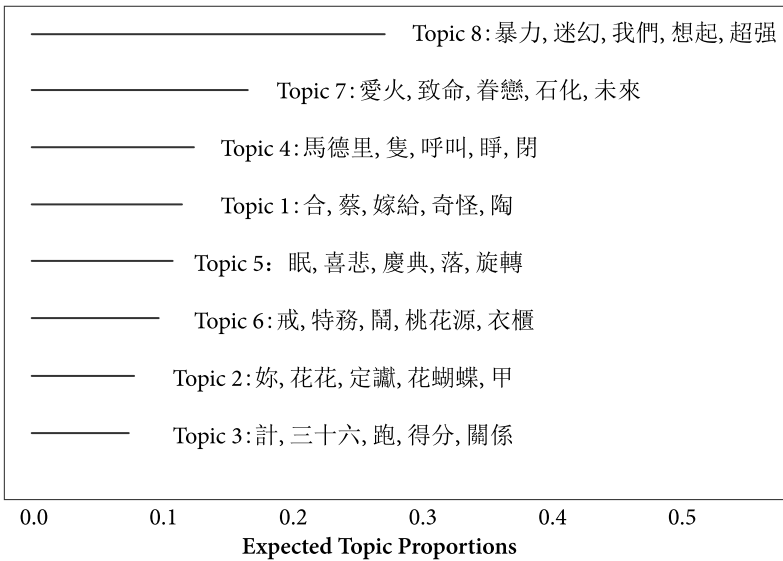
b. Exclusivity-semantic coherence distribution



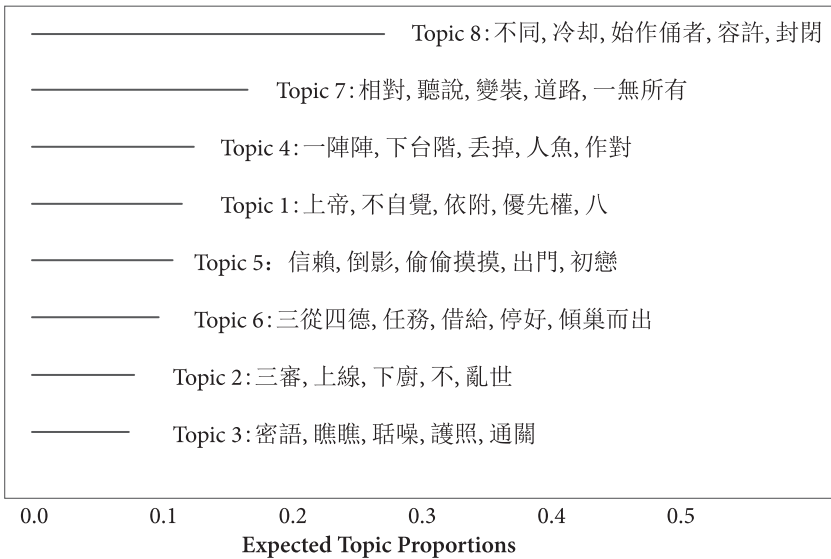
c1. 8 topics based on the highest probability



c2. 8 topics based on the FREX value



c3. 8 topics based on the score value



c4. 8 topics based on the lift value

Figure 17. (a) Topic correlation of the TQ model, (b) Exclusivity-semantic coherence distribution, and (c) Resulting topics

With the statistical evaluation of meta variables, STM can identify how style-difference meta-variables influence resulting topics. The result in Figure 18 shows that non-verbal *lyric length* only reaches significance in topic 3. *Singer type* is not so significant except in topics 2 and 7, which implies that the work of the two singers, differing in lyricist diversity, are statistically similar for most topics. The style differences show significance in the most semantically coherent topic 7.

Topic 2:					Topic 7:				
Coefficients:					Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)		Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.06848	0.13955	0.49	0.6239	(Intercept)	0.22375	0.17402	1.29	0.1992
singersai	0.07147	0.02529	2.83	0.0049 **	singersai	-0.08917	0.03356	-2.66	0.0082 **
s(length)1	-0.00879	0.30664	-0.03	0.9771	s(length)1	-0.37663	0.38424	-0.98	0.3283
s(length)2	-0.09844	0.15692	-0.63	0.5284	s(length)2	0.15286	0.20275	0.75	0.4513
s(length)3	-0.05118	0.17146	-0.30	0.7655	s(length)3	-0.00753	0.20184	-0.04	0.9703
s(length)4	-0.01951	0.14446	-0.14	0.8926	s(length)4	-0.09826	0.18816	-0.52	0.6018
s(length)5	-0.07984	0.15651	-0.51	0.6103	s(length)5	0.01291	0.19239	0.07	0.9465
s(length)6	0.00860	0.14725	0.06	0.9534	s(length)6	0.06156	0.18496	0.33	0.7395
s(length)7	-0.09350	0.15261	-0.61	0.5484	s(length)7	0.11378	0.18978	0.60	0.5492
s(length)8	0.08980	0.21961	0.41	0.6828	s(length)8	-0.37766	0.26467	-1.43	0.1544
s(length)9	0.29733	0.38446	0.77	0.4397	s(length)9	0.16491	0.42645	0.39	0.6992
s(length)10	-0.16653	0.25024	-0.67	0.5061	s(length)10	-0.19089	0.30781	-0.62	0.5355
---					---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

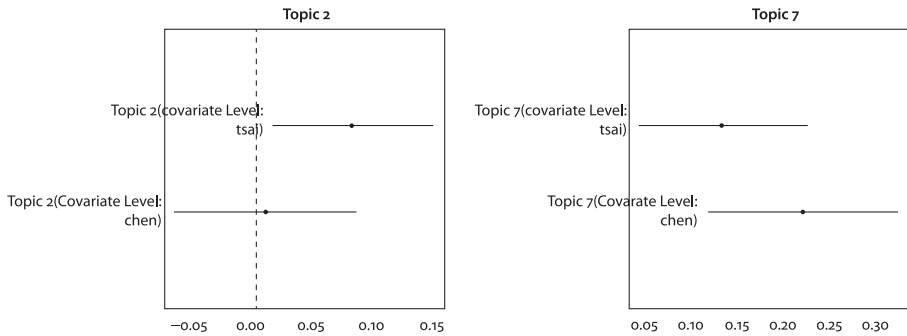


Figure 18. Statistical significance of metadata variables and their estimated effects

Take a closer look at the two significant topics, topic 2 is influenced more by Tsai, while topic 7 is authored by Chen. This statistical significance in topic 2 conforms to the cross-corpora keyness comparison. Topic 2 is the only TQ topic with “female you,” which is signified in the keyness wordlist as it is mostly derived from Tsai’s work as well as confirmed in statistically estimated effects.

In addition to statistical evaluation, topical contrasts can illustrate the dominant contrasting word in sketching disparity. The topical disparity is shown in the following graphics (Figure 19 and Figure 20), where the larger the font size is, the more dominant it is in the topic. The dotted line indicates the boundary of the compared topics. Words with a stronger affinity for one topic are located closer to it horizontally. Compared to the most coherent topic 7, the most exclusive topic 3, and the most distributive topic 8, “female you” is always the dominant one in

topic 2. Other topics may earn the dominance of other pronouns. This distinction is clear.

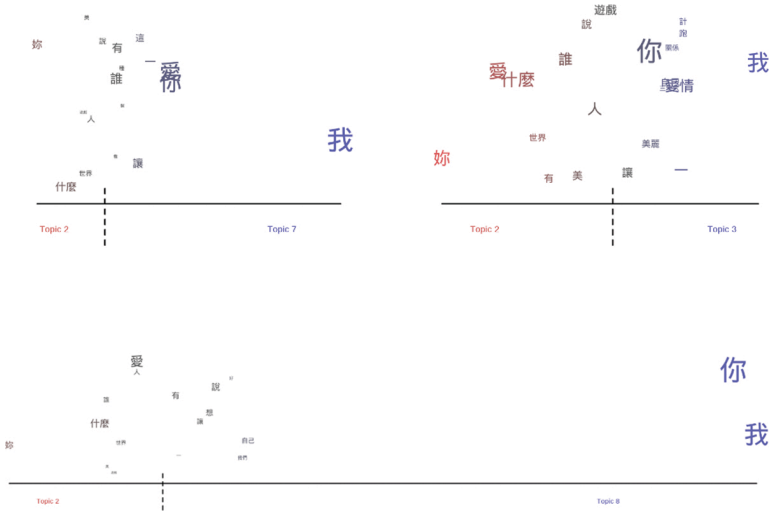


Figure 19. Topical contrasts between topic 2 and topic 3, topic 7, topic 8

As shown in Figure 20, in contrast to the most exclusive topic 3, “I,” “love,” and “you” are the most prevalent terms in topic 7. They all favor topic 7. Contrarily, content words are more clustered around the exclusive topic 3.

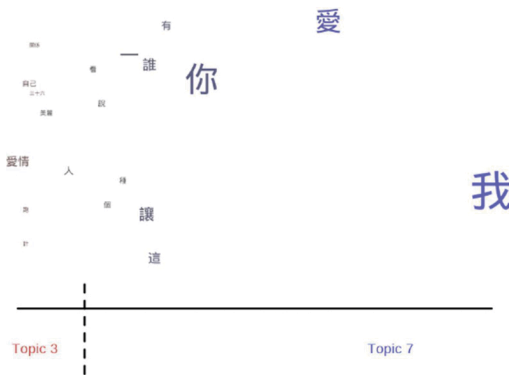


Figure 20. Topical contrasts between topic 3 and topic 7

To be specific, the TQ model clarifies the cross-topic relationship using topic cluster coherence degrees, diverse calculation-based wordlists, statistical evaluation, and topical contrasts. Different topics show contrasting topic cluster coherence degrees. The wordlists in Figure 19 indicate different content perspectives

under the same topic. For instance, “love” seems to recur in the majority of topics. According to FREX, score, and lift value, topic 8 is more concerned with “violence, evidence, difference, the initiator of evil, or tolerance;” nevertheless, topic 7 depicts “love, fire, hurt, raindrops, or the future of love.” Various viewpoints on love are presented. The topical contrasts indicate that most topics overlap in their usages of pronouns, even though the statistical assessment indicates the independence of the topics. Besides, meta information (singer type) is significant only in topics 2 and 7. This implies the influence of singer type on topics is limited, yet such style-difference meta-variables significantly influence the coherent topic. The exclusive topic, however, is clustered around content words based on the comparison between the exclusive topic 3 and the coherent topic 7. The previous comparison of bags of words from *BTM* and *keyness* make accounting for differences difficult. In contrast, *STM* allows us to compare similar topics and capture subtle differences.

Results from this section highlight the role pronouns play in highly coherent topics, whereas whether function words such as pronouns are related to style differences still remains unclear. In the next section, the iterable assessment model is trained to identify words contributing to style differences based on topical contrast results.

4.2.3 Iterable assessment model (IA model)

The TQ model indicates the dissimilarities of different topics. The IA model is applied to determine if the metadata variable “*singer type*” (word consistency) leads to topical contrasts. The model training procedure in this stage is similar to previous steps, but exclusivity is specified within the content covariate parameter so there is no exclusivity-semantic coherence distribution figure. As stated in the manual,¹² the resulting sparsity should be higher than 0.5, so model 2 is chosen as the IA model for its higher sparsity.

```
[1] "Model 1 has on average -72.6553878980206 semantic coherence and 0.926808266360505 sparsity"
[1] "Model 2 has on average -71.5237151082896 semantic coherence and 0.927780966960071 sparsity"
[1] "Model 3 has on average -70.406812862862 semantic coherence and 0.924942594718714 sparsity"
[1] "Model 4 has on average -70.809340476223 semantic coherence and 0.927956371986223 sparsity"
```

Figure 21. Semantic coherence and sparsity value for trained models

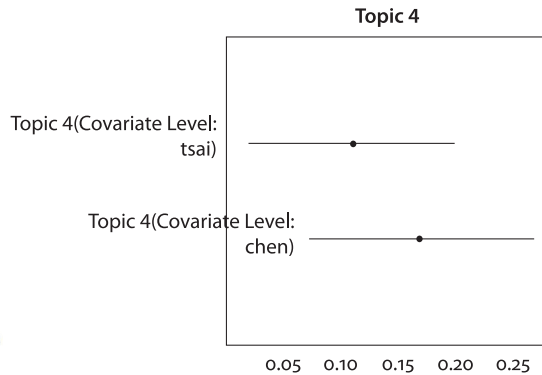
This new model results in new topical distributions, topical contrasts, and estimated effects. The topical contrasts in the IA model denote the contrasting perspectives of the covariate *singer type*. Based on the statistical assessment shown in Figure 22, topic 4 is marginally affected by *singer type* and *lyric length*. Topic

12. <https://cran.r-project.org/web/packages/stm/stm.pdf>

7 and topic 8 are more associated with the *singer type* variable. As depicted in Figure 23, the perspective contrast between the two singers in topics 4, 7, and 8 elucidates that pronouns remain vital in style differences. According to Figure 23 and 24, topic 7 is about “wither,” fallen,” “tide,” “harm,” and “taste.” Tsai’s perspective on this topic is more concerned with these fallen qualities’ association with the content word “love.” Similarly, Tsai’s perspective on topic 8 is more concerned with how “love” is involved with “indulge,” “heal,” “energy,” “realize,” and “make trouble out of nothing.” Unlike Tsai’s perspective, the link between “you” and “I” is emphasized in Chen’s perspective on both topics. To sum up, except for particular topics, the metadata variable effects (*singer type* and *lyric length*) are still not significant. These two particular topics (topics 7 and 8), which are widely distributed and more associated with the *singer type* variable, display that pronouns are strongly related to Chen’s perspective. To be specific, if differentiation is signified, then it is pronouns that lead to perspective contrasts between the two singers. In most cases, Chen pays more attention to “you” and “I” compared to Tsai’s work. In this regard, style differences are indeed encapsulated in pronoun dominance.

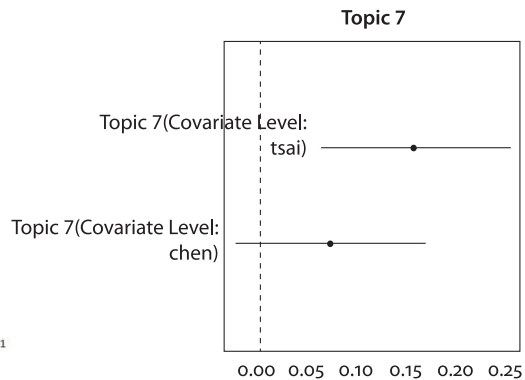
```

Topic 4:
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07278   0.14970   0.49   0.627
singertsai  -0.05758   0.03232  -1.78   0.076 .
s(length)1   0.13734   0.36548   0.38   0.707
s(length)2   0.39273   0.19859   1.98   0.049 *
s(length)3  -0.04632   0.17838  -0.26   0.795
s(length)4   0.15736   0.15436   1.02   0.309
s(length)5   0.10466   0.17165   0.61   0.542
s(length)6   0.00583   0.15830   0.04   0.971
s(length)7   0.06737   0.16376   0.41   0.681
s(length)8   0.06669   0.23325   0.29   0.775
s(length)9   0.10344   0.41261   0.25   0.802
s(length)10 -0.04396   0.27481  -0.16   0.873
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
    
```



```

Topic 7:
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07378   0.18355   0.40   0.688
singertsai   0.08373   0.03549   2.36   0.019 **
s(length)1  -0.03893   0.43072  -0.09   0.928
s(length)2   0.04853   0.20477   0.24   0.813
s(length)3   0.00993   0.21330   0.05   0.963
s(length)4   0.03095   0.19424   0.16   0.873
s(length)5  -0.03500   0.20447  -0.17   0.864
s(length)6   0.10083   0.19435   0.52   0.604
s(length)7   0.07728   0.20458   0.38   0.706
s(length)8  -0.29881   0.27177  -1.10   0.272
s(length)9   0.11320   0.43969   0.26   0.797
s(length)10 -0.21592   0.31546  -0.68   0.494
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```



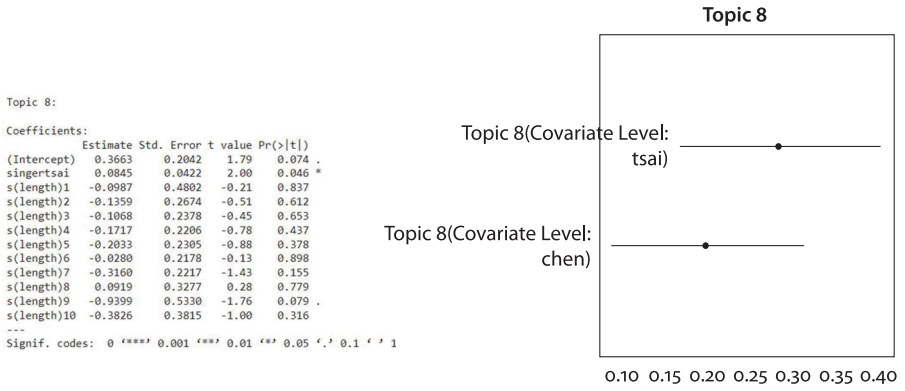
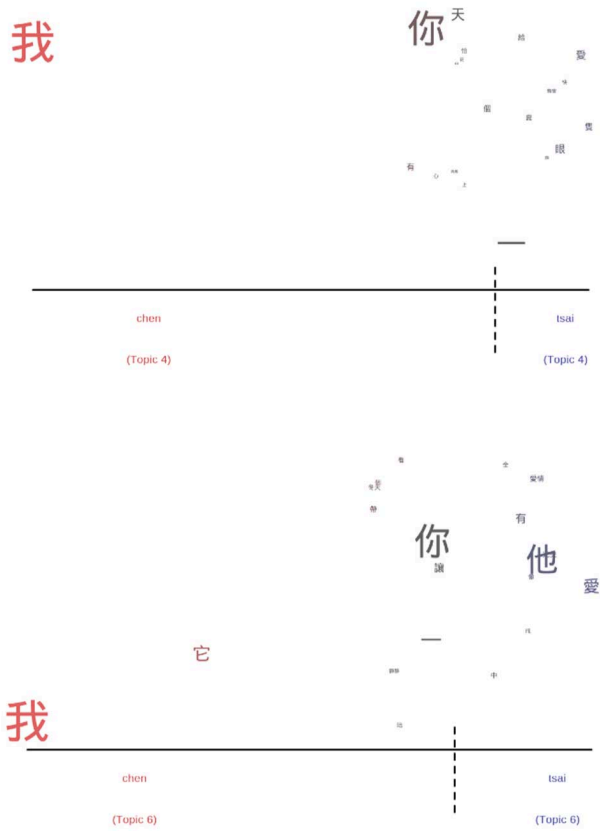


Figure 22. Statistical significant results of metadata variables and their estimated effects on metadata variables



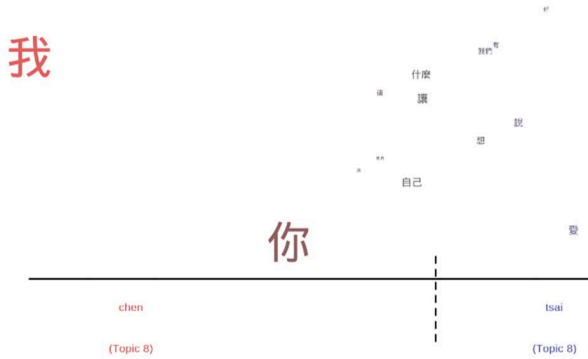


Figure 23. Perspective contrasts of the specified singer type variable

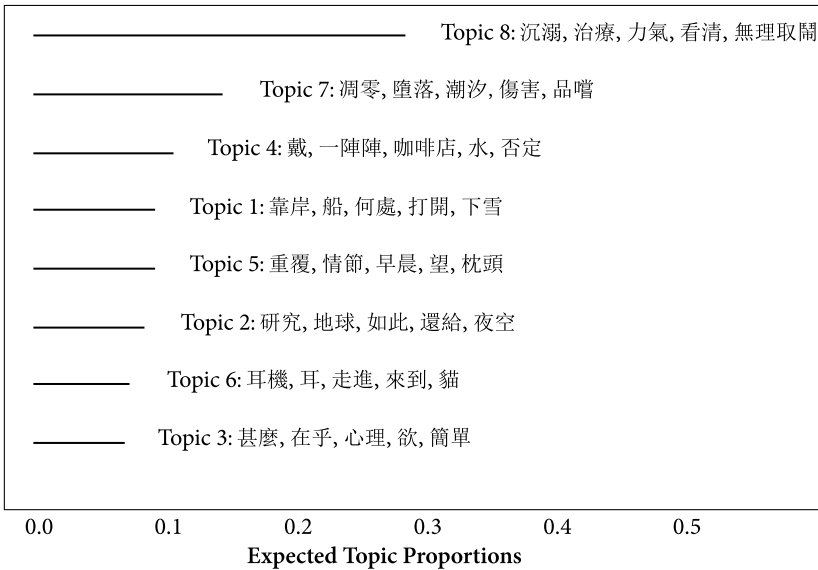


Figure 24. Expected topic clusters and topic proportions in the IA model

4.3 Generalization

Pennebaker (2013) suggested that gender preference, social implications, personal traits, ways of thinking, and affection sharing can be depicted in pronoun usages. In his discussion, Pennebaker (2013: 291) identified function words as style words by pointing out, “A good rule of thumb is that people who pay a great deal of attention to other people tend to use personal pronouns at high rates.” Pronouns

account for about 20% to 15% of Chen’s and Tsai’s lyrics, as illustrated in Figure 25. The caring quality of Chen’s lyrics is highlighted.

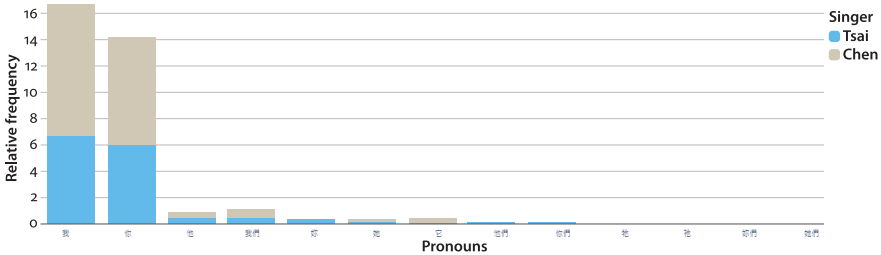


Figure 25. Pronoun distribution in Tsai and in Chen

With the assistance of the proposed STM-led framework, we can further explicate style differences from two perspectives: (a) How do style differences interact with topical qualities? (b) What style words are dominant in capturing style differences? The TQ model illustrates how style differences relate to topical contexts and how topic contexts assist in the analysis. The statistical results reveal that it is the coherent topic that the style-difference meta-variable (singer type) may significantly influence. The topical contrast result between the most coherent topic and the most exclusive topic denotes that pronouns predominantly depict topical contrast and are in favor of the coherent topic. Content words, however, cluster around the exclusive topic.

To identify the “style words,” the IA model is trained to obtain perspective contrasts based on style-difference meta-variables. The topical contrast results illustrate pronouns’ dominance in style contrast, coinciding with views in Pennebaker (2013), “quiet words can say more.” Word correlations (word consistency) may differ according to lyricist distributions as seen in Section 4.1, whereas the significance of style disparity can only be depicted in specific topics. This implies that topics from diverse lyricists are pretty similar. When it comes to distinguishing styles, pronouns indeed take the lead. The conclusion drawn here enriches the details of the “style word” hypothesis proposed by Pennebaker (2013). The STM-led framework can specifically adduce the intricate mechanisms underlying pronouns.

5. Conclusion

This study proposes the STM-led analytics framework to overcome possible limitations of previous corpus linguistic approaches. The corpus-based *keyness*

approach captures words occurring differently in the target and reference corpora. The main weakness of this approach is that it can only detect local word-level information without contextual clues, thus hindering the qualitative analysis of the results. Although advantageous in depicting global corpus-level themes in short texts, the corpus-driven *BTM* can only capture the topics of a single corpus at a time, which obstructs drawing a direct and valid comparison between two corpora. In contrast to these two approaches, the proposed STM-led framework can systematically compare and contrast data sets by leveraging adequate contextual clues and statistical evidence.

The case study on style differences in lyrics demonstrates how to apply the proposed framework for studying linguistic theory. The linguistic issue in question, style difference, is signified by two metavariables: *lyric length* and *singer type*. The *singer type* variable denotes the different lyricist distribution in two corpora (Tsai and Chen). To address the research question, the proposed STM-led framework employs two models (TQ and IA models) for the systematic evaluation of linguistic theories. In this manner, the TQ model provides insights from topical qualities, various wordlists, and statistical evaluation to indicate that style differences are especially significant in coherent topics and that pronouns are prominent in topical contrast. In addition to depicting topical contrasts, the IA model delineates how pronouns can notably discern perspective contrasts on the same topic based on specified metadata variables. Seldom do studies employ topic modeling techniques to probe into linguistic theories based on lyric data. The contribution of this study is that, despite differences in word consistency and lyric length across the two corpora, the core theme identified by the TQ is almost the same: it is love. Yet, the IA model further clarifies the perspective contrasts on this theme. The results from the proposed STM-led framework extend the style differences issue discussed by Pennebaker (2013) in a more data-driven yet robust way. In this way, the STM-led framework is more than a statistics-based STM. With the methodical design of model training stages and linguistic insights, the STM-led framework is capable of revealing the shared themes among different corpora, as well as exposing the cross-corpora comparisons of topic perspective contrasts on a common topic. This is a paradigm from a *bottom-up* view, so it is not limited to corpora with consistent or inconsistent styles. On the contrary, the purpose is to identify whether the corpora are similar or different, and from what aspects. The proposed STM-led architecture provides two vital functions: identifying similarities between different corpora and revealing differences within a topic. Thus, even if datasets behave in the same way in their resulting topics, the proposed framework can still disclose the differences in talking about those topics.

Texts with limited length and limited amounts of data, such as lyrics, social media posts, hashtag contents, and online conversations, are part of our daily





lives. Such data is often marginalized in deep learning methods. How to retrieve accountable classifications on such texts in an unsupervised manner is an issue that must be addressed. Topic modeling has been commonly employed for exploring unstructured data. We believe that the proposed STM-led paradigm incorporating linguistic supervision propels this exploratory method into an evaluative model that can better draw comparisons among clustering results.

Roberts, Stewart & Tingley (2016) suggested “more direct supervision” to facilitate STM. Linguistic insights can enrich the supervision parts of this unsupervised method. The proposed STM-led framework with linguistic supervision elucidates that linguistic theories and topic modeling techniques can cooperate reasonably. Such collaboration can assist in carrying out reproducible and interpretable future research in lyricist division detection, gender factors in lyric creation, and temporal influence on lyric conception. Adopting both qualitative and quantitative points of view to address linguistic data of shorter text length and texts with a limited amount of data is no longer a challenge, but full of opportunities.

Acknowledgements

The authors are very appreciative to the anonymous reviewers for their insightful remarks and recommendations on several aspects of the article. Additionally, the author would like to thank the editors for their correction remarks.

References






-  Aarts, F.G.A.M. 1971. On the distribution of noun-phrase types in English clause structure. *Lingua* 26.3:281–293.
-  Abuzayed, Abeer, and Hend Al-Khalifa. 2021. BERT for Arabic topic modeling: An experimental study on BERTopic technique. *Procedia Computer Science* 189:191–194.
-  Akella, Revanth, and Teng-Sheng Moh. 2019. Mood classification with lyrics and ConvNet. *Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, ed. by M.A. Wani, 511–514. Los Alamitos, CA: IEEE Computer Society.
- Angelov, Dimo. 2020. *Top2vec: Distributed Representations of Topics*. Retrieved January 14th, 2023, from <https://arxiv.org/abs/2008.09470>
-  Aranda, Ana M., Kathrin Sele, Helen Etchanchu, Jonne Y. Guyt, and Eero Vaara. 2021. From big data to rich theory: Integrating critical discourse analysis with structural topic modeling. *European Management Review* 18:197–214.
- Arifah, Khadijah. 2016. *Figurative Language Analysis in Five John Legend’s Song*. Doctoral dissertation, Maulana Malik Ibrahim State Islamic University, Malang, Indonesia.




- Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. *Proceedings of the 30th International Conference on Machine Learning*, ed. by Sanjoy Dasgupta and David McAllester, 280–288. Atlanta, GA: JMLR.org.
- doi** Baratè, Adriano, Luca A. Ludovico, and Enrica Santucci. 2013. A semantics-driven approach to lyrics segmentation. *Proceedings of the 2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, ed. by Randall Bilof, 73–79. Los Alamitos, CA: IEEE Computer Society.
- Barradas, Gonçalo T., and Laura S. Sakka. 2021. When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions. *Psychology of Music* 50.2:650–669.
- doi** Besson, Mireille, Frederique Faita, Isabelle Peretz, A-M. Bonnel, and Jean Requin. 1998. Singing in the brain: Independence of lyrics and tunes. *Psychological Science* 9.6:494–498.
- Bischof, Jonathan, and Edoardo M. Airoidi. 2012. Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ed. by John Langford and Joelle Pineau, 201–208. Madison, WI: Omnipress.
- doi** Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55.4:77–84.
- doi** Blei, David M., and John D. Lafferty. 2007. A correlated topic model of *Science*. *The Annals of Applied Statistics* 1.1:17–35.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.
- Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems* 32:288–296.
- doi** Chen, Stanley F., and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13.4:359–394.
- doi** Chen, Xieling, Di Zou, Gary Cheng, and Haoran Xie. 2020. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of *Computers & Education*. *Computers & Education* 151:103855.
- doi** Damerau, Fred J. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management* 29.4:433–447.
- doi** Devi, Maibam Debina, and Navanath Saharia. 2020. Exploiting topic modelling to classify sentiment from lyrics. *Proceedings of the 2nd International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (MIND 2020)*, ed. by Arup Bhattacharjee, Samir Kr. Borgohain, Badal Soni, Gyanendra Verma and Xiao-Zhi Gao, 411–423. Singapore: Springer.
- doi** Dewi, Erniyanti Nur Fatahela, Didin Nuruddin Hidayat, and Alek Alek. 2020. Investigating figurative language in “Lose You to Love Me” song lyric. *Loquen: English Studies Journal* 13.1:6–16.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.1:61–74.

- doi Ebeling, Régis, Carlos Abel Córdova Sáenz, Jeferson Campos Nobre, and Karin Becker. 2021. The effect of political polarization on social distance stances in the Brazilian COVID-19 scenario. *Journal of Information and Data Management* 12.1:86–108.
- doi Eckstein, Lars. 2010. *Reading Song Lyrics*. Leiden: Brill.
- Eisenstein, Jacob, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ed. by Lise Getoor and Tobias Scheffer, 1041–1048. Madison, WI: Omnipress.
- doi Gabrielatos, Costas. 2018. Keyness analysis: Nature, metrics and techniques. *Corpus Approaches to Discourse: A Critical Review*, ed. by Charlotte Taylor and Anna Marchi, 225–258. London: Routledge.
- Grootendorst, Maarten. 2022. *BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure*. Retrieved May 7th, 2022, from <https://arxiv.org/abs/2203.05794>
- doi Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ed. by Fredric Gey, Marti Hearst and Richard Tong, 50–57. New York, NY: Association for Computing Machinery.
- doi Hong, Liangjie, and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. *Proceedings of the 1st Workshop on Social Media Analytics*, ed. by Prem Melville, Jure Leskovec and Foster Provost, 80–88. New York, NY: Association for Computing Machinery.
- Hoover, David L. 2007. Corpus stylistics, stylometry, and the styles of Henry James. *Style* 41.2:174–203.
- Kilgarriff, Adam. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings of the 5th ACL Workshop on Very Large Corpora*, ed. by Joe Zhou and Kenneth Church, 231–245. Beijing and Hong Kong: Tsinghua University and The Hong Kong University of Science and Technology.
- doi Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1.2:263–276.
- doi Kreyer, Rolf, and Joybrato Mukherjee. 2007. The style of pop song lyrics: A corpus-linguistic pilot study. *Anglia. Journal of English Philology* 125.1:31–58.
- doi Laoh, Enrico, Isti Surjandari, and Limisgy Ramadhina Febirautami. 2018. Indonesians' song lyrics topic modelling using latent dirichlet allocation. *Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, ed. by Shaozi Li, Ying Dai and Yun Cheng, 270–274. Los Alamitos, CA: IEEE Computer Society.
- Leech, Geoffrey, and Roger Fallon. 1992. Computer corpora-what do they tell us about culture? *ICAME Journal* 16:29–50.
- doi Li, Peng-Hsuan, Tsu-Jui Fu, and Wei-Yun Ma. 2020. Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. *Proceedings of the AAAI Conference on Artificial Intelligence*, ed. by Vincent Conitzer and Fei Sha, 8236–8244. California, USA: AAAI Press, Palo Alto.
- doi Lindstedt, Nathan C. 2019. Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. *Social Currents* 6.4:307–318.


- Mimno, David M., and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, ed. by David McAllester and Petri Myllymaki, 411–418. Arlington, VA: AUAI Press.
- Mimno, David M., Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, ed. by Regina Barzilay and Mark Johnson, 262–272. Edinburgh, Scotland, UK: Association for Computational Linguistics.
- [doi](#) Nahajec, Lisa. 2019. Song lyrics and the disruption of pragmatic processing: An analysis of linguistic negation in 10CC's 'I'm Not in Love'. *Language and Literature* 28.1:23–40.
- [doi](#) Narayan, Ashwin, Bonnie Berger, and Hyunghoon Cho. 2021. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology* 39:765–774.
- [doi](#) Newman, David, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, ed. by Jane Hunter, 215–224. New York, NY: Association for Computing Machinery.
- North, Adrian C., Amanda E. Krause, and David Ritchie. 2020. The relationship between pop music and lyrics: A computerized content analysis of the United Kingdom's weekly top five singles, 1999–2013. *Psychology of Music* 49.4:735–758.
- Pennebaker, James W. 2013. *The Secret Life of Pronouns: What Our Words Say About Us*. London: Bloomsbury Publishing.
- [doi](#) Petrie, Keith J., James W. Pennebaker, and Borge Sivertsen. 2008. Things we said today: A linguistic analysis of the Beatles. *Psychology of Aesthetics, Creativity, and the Arts* 2.4:97–202.
- [doi](#) Pettijohn, Terry F., and Donald F. Sacco Jr. 2009. The language of lyrics: An analysis of popular Billboard songs across conditions of social and economic threat. *Journal of Language and Social Psychology* 28.3:297–311.
- [doi](#) Pojanapunya, Punjaporn, and Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory* 14.1:133–167.
- [doi](#) Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ed. by Jian Su, Kevin Duh and Xavier Carreras, 2383–2392. Stroudsburg, PA: Association for Computational Linguistics.
- Rayson, Paul. 2019. Corpus analysis of key words. *The Concise Encyclopedia of Applied Linguistics*, ed. by Carol Ann Chapelle, 320–326. Oxford: John Wiley & Sons.
- [doi](#) Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2016. Navigating the local modes of big data: The case of topic models. *Computational Social Science: Discovery and Prediction*, ed. by R. Michael Alvarez, 51–97. New York: Cambridge University Press.
- [doi](#) Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. Stm: An R package for structural topic models. *Journal of Statistical Software* 91.2:1–40.

- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation* 4:1–20.
-  Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58.4:1064–1082.
-  Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, ed. by Xueqi Cheng and Hang Li, 399–408. New York, NY: Association for Computing Machinery.
- Sasaki, Shoto, Kazuyoshi Yoshii, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima. 2014. LyricsRadar: A lyrics retrieval system based on latent topics of lyrics. *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, ed. by Hsin-Min Wang, Yi-Hsuan Yang and Jin Ha Lee, 585–590. Taipei: International Society for Music Information Retrieval.
-  Schedl, Markus. 2019. Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics* 5:44.
-  Schweinberger, Martin, Michael Haugh, and Sam Hames. 2021. Analysing discourse around COVID-19 in the Australian Twittersphere: A real-time corpus-based analysis. *Big Data & Society* 8.1:1–17.
-  Setiawati, Wilya, and Maryani Maryani. 2018. An analysis of figurative language in Taylor Swift's song lyrics. *PROJECT (Professional Journal of English Education)* 1.3:261–268.
-  Shahmohammadi, Hassan, MirHossein Dezfoulian, and Muharram Mansoorzadeh. 2021. Paraphrase detection using LSTM networks and handcrafted features. *Multimedia Tools and Applications* 80.4:6479–6492.
-  Sharma, Hardik, Shelly Gupta, Yukti Sharma, and Archana Purwar. 2020. A new model for emotion prediction in music. *Proceedings of the 2020 6th International Conference on Signal Processing and Communication (ICSC)*, ed. by Jitendra Mohan and Abhinav Gupta, 156–161. Los Alamitos, CA: IEEE Computer Society.
- Snyder, Robin M. 2015. An introduction to topic modeling as an unsupervised machine learning way to organize text information. Paper presented at the Annual Meeting of the Association Supporting Computer Users in Education (ASCUE), Myrtle Beach, SC.
-  Sophiadi, Angelina. 2014. The song remains the same... or not? A pragmatic approach to the lyrics of rock music. *Major Trends in Theoretical and Applied Linguistics*, vol. 2, ed. by Nikolaos Lavidas, Thomai Alexiou and Areti-Maria Sougari, 125–142. London: De Gruyter Open Poland.
- Sterckx, Lucas. 2014. Topic Detection in a Million Songs. Doctoral dissertation, Ghent University, Ghent, Belgium.
- Taddy, Matt. 2012. On estimation and selection for topic models. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, ed. by Neil D. Lawrence and Mark Girolami, 1184–1193. Retrieved May 27th, 2022, from <https://proceedings.mlr.press/v22/taddy12.html>

-  Tegge, Friederike. 2017. The lexical coverage of popular songs in English language teaching. *System* 67:87–98.
- Trenquier, Henri. 2018. Improving Semantic Quality of Topic Models for Forensic Investigation. Doctoral dissertation, University of Amsterdam, Amsterdam, Netherlands.
-  Varnum, Michael E. W., Jaimie Arona Krems, Colin Morris, Alexandra Wormley, and Igor Grossmann. 2021. Why are song lyrics becoming simpler? A time series analysis of lyrical complexity in six decades of American popular music. *PLOS ONE* 16.1:0244576.
-  Wallach, Hanna Megan. 2006. Topic modeling: beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning*, ed. by William W. Cohen and Andrew Moore, 977–984. New York, NY: Association for Computing Machinery.
- Wallach, Hanna Megan. 2008. Structured Topic Models for Language. Doctoral dissertation, University of Cambridge, Cambridge, UK.
-  Wallach, Hanna Megan, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, ed. by Andrea Danyluk, 1105–1112. New York, NY: Association for Computing Machinery.
- Wang, Jie, and Xinyan Zhao. 2019. *Theme-Aware Generation Model for Chinese Lyrics*. Retrieved September 20th, 2022, from <https://arxiv.org/abs/1906.02134>
-  Watanabe, Kento, Matsubayashi Yuichiroh, Inui Kentaro, Nakano Tomoyasu, Fukayama Satoru, and Goto Masataka. 2017. Lyrisys: An interactive support system for writing lyrics based on topic transition. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, ed. by George A. Papadopoulos and Tsvi Kuflik, 559–563. New York, NY: Association for Computing Machinery.
-  Weng, Jianshu, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twiterrank: Finding topic-sensitive influential twitterers. *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, ed. by Brian D. Davison and Torsten Suel, 261–270. New York, NY: Association for Computing Machinery.
-  Werner, Valentin. 2021. Catchy and conversational? A register analysis of pop lyrics. *Corpora* 16.2:237–270.
-  Whissell, Cynthia. 1996. Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities* 30.3:257–265.
- Wright, David. 2014. Stylistics Versus Statistics: A Corpus Linguistic Approach to Combining Techniques in Forensic Authorship Analysis Using Enron Emails. Doctoral dissertation, University of Leeds, Leeds, England.
-  Xia, Xiaoling, Xin Gu, and Qinyang Lu. 2019. Research on the model of lyric emotion algorithm. *Journal of Physics: Conference Series* 1213:042004.
-  Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web*, ed. by Daniel Schwabe, Virgilio Almeida and Hartmut Glaser, 1445–1456. New York, NY: Association for Computing Machinery.
-  Yao, Liang, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, ed. by The Association for the Advancement of Artificial Intelligence, 7370–7377. Palo Alto, CA: AAAI Press.

-  Zhang, Lei, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4:1253.
-  Zhang, Liang, Keli Xiao, Hengshu Zhu, Chuanren Liu, Jingyuan Yang, and Bo Jin. 2018. CADEN: A context-aware deep embedding network for financial opinions mining. *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, ed. by Lisa O’Conner, 757–766. Los Alamitos, CA: IEEE Computer Society.
-  Zhao, Wayne Xin, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. *Advances in Information Retrieval: Proceedings of the 33rd European Conference on IR Research*, ed. by Paul Clough, Colum Foley, Cathal Gurrin, Gareth J.F. Jones, Wessel Kraaij, Hyowon Lee and Vanessa Mudoch, 338–349. Heidelberg: Springer Berlin.

Address for correspondence

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
shukaihsieh@ntu.edu.tw
 <https://orcid.org/0000-0001-9674-1249>

Co-author information

Po-Ya Angela Wang
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
differe94nt@gmail.com

Publication history

Date received: 8 July 2022
Date revised: 5 October 2022
Date accepted: 18 January 2023