International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Automatic Discovery of Nativelike Selections through Topic Modeling: an Experiment Based on Online Reviews

Huasheng Zhang
College of Liberal Arts, Sichuan Normal University, China
Corresponding Author: Huasheng Zhang, **E-mail**: zhang.huasheng@foxmail.com

| ABSTRACT

Nativelike selections are linguistic expressions favored by native speakers over other grammatically correct alternatives to express a certain concept. They are important items for language teaching and lexicography. This study proposes an automatic method for discovering them through the application of the topic modeling technique of BERTopic which discovers the frequent linguistic expressions that express the same topic. The method's effectiveness and efficiency are demonstrated through an experiment based on online reviews of a dish soap product in English, Chinese, and Japanese.

| KEYWORDS

nativelike selection; topic modeling; BERTopic; online reviews

| ARTICLE INFORMATION

ACCEPTED: 02 January 2025 **PUBLISHED:** 15 February 2025 **DOI:** 10.32996/ijllt.2025.8.2.11

1. Introduction

Pawley & Syder (1983: 191) were the first linguists to call attention to the phenomenon of nativelike selection as "the ability of the native speaker routinely to convey his meaning by an expression that is not only grammatical but also nativelike; what is puzzling about this is how he selects a sentence that is natural and idiomatic from among the range of grammatically correct paraphrases, many of which are non-nativelike or highly marked usages". For illustration, Pawley & Syder (1983) demonstrated that (1a) is more nativelike than (1b), (1c), etc., even though they are all grammatically correct and convey "[t]he same objective information". Following Foster (2009) and Zaytseva (2016), such linguistic expressions favored by native speakers to express a certain concept, are referred to as the term "nativelike selection" in this article.

(1) a. I want to marry you.
b. I wish to be wedded to you.
c. I want marriage with you.
…

Over the years, collocations have been considered key to nativelike selection (Pawley & Syder, 1983; Erman, 2009; Wray, 2012; Erman et al., 2016). The automatic discovery of collocations largely relies on statistical measures such as mutual information (Deng & Liu, 2022). However, these calculation methods essentially analyze the target linguistic expression in isolation rather than comparing it with other specific expressions of the same concept. Collocation can be defined as a series of words or terms that co-occur more often than would be expected by chance; a word combination can be considered as a collocation if its mutual information is above a threshold of 3 (Gries, 2022: 17), which does not directly compare the synonymous expressions of the same concept.

In this study, an automatic method for discovering nativelike selections is proposed, which discovers the frequent linguistic expressions that express the same topic and the corresponding concept, using the topic modeling technique of BERTopic (Grootendorst, 2022). This method better aligns with the definition of nativelike selection, and this perspective is beneficial for developing language learning resources, given the learners' need to sound nativelike when discussing various topics in another language. Moreover, this automatic method offers an efficiency advantage over previous manual studies, which will be discussed in the following section.

2. Literature Review

2.1 Manual Discovery of Nativelike Selections

Previous studies used manual analysis to discover linguistic expressions favored by native speakers to express a certain concept. Various types of data have been used. Geeraerts et al. (1994) utilized fashion magazines and was targeted at the words denoting the clothes in magazine images. Data can also be obtained by elicitation tasks. Foster (2009) used cartoon re-telling tasks and studied how native speakers as well as learners described the different "frames" in the cartoons. Similarly, Zaytseva (2016) studied both native speakers and learners' usages, using written and oral tasks. Different from these studies, in Smiskova et al. (2012), the nativelikeness of the expressions on the same theme was found out by native speakers' rating and reference corpus frequency check. Finally, the nativelike selections for expressing a given concept can be discovered through controlled tasks such as the forced-choice task (Liu, 2013) and the discourse completion task (Ortaçtepe, 2013). These manual methods require effort to discover noteworthy aspects of native speakers' language use, as well as to collect and compare synonymous expressions. These can be facilitated by an automatic method which can also be more easily applied to larger scale data.

2.2 Topic Modeling Technique of BERTopic

Nativelike selections can be discovered by examining the favored expressions under the same topic as demonstrated in the manual methods above. Retrieving topics and corresponding topic words from a collection of documents is a well-developed field in natural language processing, known as topic modeling. The topic modeling technique of BERTopic (Grootendorst, 2022) is used for automatic discovery of nativelike selections in this study. Additionally, OCTIS is a universal framework for optimizing topic models, and is compatible with BERTopic.

BERTopic leverages language models to generate semantic vectors for the documents in the source data. These semantic vectors encode a document's semantic content within a semantic space, where vectors in closer proximity indicate greater semantic similarity between documents. These documents are then grouped into different topics by clustering these semantic vectors. Finally, topic words for each topic are identified by the ranking of their frequencies and association with the topic. For a detailed introduction to semantic vectors, please refer to Lenci (2018).

BERTopic has been recognized as the state-of-the-art topic modeling technique (Chen et al., 2023) and can handle short noisy data, which poses challenges for traditional methods (Egger & Yu, 2022). Previously, BERTopic has been widely used in social science to understand public discourses such as Twitter discussions on COVID-19 (Xu et al., 2022), and changes in discourse on climate change and economic growth in news (Mervaala, 2025). However, this study explores the potential of applying BERTopic in linguistics, specifically for discovering nativelike selections from source data.

3. Materials and Methods

Data was collected from customer reviews of a dish soap sold internationally across the United Kingdom, China, and Japan. Customer reviews were collected from three online marketplaces: Amazon.co.uk (United Kingdom), JD.com (China), and Amazon.co.jp (Japan). Each review was split into individual sentences. The data structure is as follows: English (23,709 words from 2,521 sentences), Chinese (65,839 characters from 3,487 sentences), and Japanese (103,116 characters from 3,880 sentences).

The topic modeling technique of BERTopic (Grootendorst, 2022) was applied to the data. The process includes the following steps: text vectorization, dimensionality reduction of semantic vectors, clustering semantic vectors, topics and topic words identification, result optimization, and creating descriptions. The language models employed were GTE-large and GTE-large-zh (Li et al., 2023) for English and Chinese respectively, and sup-simcse-ja-large (Tsukagoshi et al., 2023) for Japanese.

4. Result and Discussion

Respectively, 17, 23, and 23 topics were extracted for the English, Chinese, and Japanese data as a result of BERTopic. Owing to space limitations, only 4 topics and their corresponding topic words are selected for display here for each language, as summarized in Table 1 – 3, focusing on overall opinions, environmental concerns, scent, and skin feel. The topic words are ranked by frequency and the association with the topic (i.e., c-TF-IDF value). The superscript numbers indicate the raw frequencies of the words across the entire data.

Table 1: Topic modeling results for the English data.

Topic	Topic words
Overall opinion	*product*[318], *good*[466], *thank*[28], *great*[279], *love*[93]
Environmental concern	*eco*[122], *environment*[64], *friendly*[97], *planet*[21], *kind*[49]
Scent	*smell*[199], *scent*[128], *lovely*[69], *nice*[80], *amazing*[36]
Skin feel	*hand*[141], *skin*[159], *sensitive*[84], *gentle*[41], *dry*[45]

Table 2: Topic modeling results for the Chinese data.

Topic	Topic words
Overall opinion	*bucuo*[507] 'not bad', *hao*[997] 'good', *haoping*[48] 'positive review, praise', *manyi*[84] 'satisfied', *ting*[171] 'quite'
Environmental concern	*huanbao*[240] (lit 'environmentally protective'), *fangxin*[121] 'not worry', *anquan*[101] 'safe', *chanpin*[253] 'product', *pinpai*[129] 'brand'
Scent	*xiangwei*[167] 'fragrance', *weidao*[242] 'taste', *chanpin*[253] 'product', *meiyou*[216] 'there is no', *wu*[122] 'no'
Skin feel	*ganjing*[237] 'clean', *shou*[253] 'hand', *shang*[222] 'harm', *qingxi*[109] 'wash', *xi*[433] 'wash'

Table 3: Topic modeling results for the Japanese data.

Topic	Topic words
Overall opinion	*tsukau*[703] 'use', *koonyuu*[129] 'purchase', *ii*[744] 'good', *shoohin*[160] 'product', *nedan*[53] 'price'
Environmental concern	*kankyoo*[135] 'environment', *eko*[82] 'eco (friendly)', *shizen*[52] 'nature', *yasashii*[82] 'gentle', *hairyo*[23] 'consideration'
Scent	*kaori*[452] 'fragrance', *nioi*[89] 'smell', *kooryoo*[82] '(aromatic) essence', *mu*[72] 'no', *ii*[744] 'good'
Skin feel	*teare*[252] 'chapped hand', *hada*[207] 'skin', *areru*[133] 'chap', *te*[379] 'hand', *hadaare*[43] 'chapped skin'

Several conclusions can be drawn from the results of this experiment. First, this method effectively and efficiently discovers nativelike selections by automatically discovering groups of frequent words that express the same concept within a narrow topic; the topics are well-organized, and the topic words within the same topic are semantically coherent. For example, *lovely*, *nice*, and *amazing* are expressions favored by English speakers to praise the scent of the dish soap. Also, this method helps discover latent topics from a large collection of texts without the need of prior knowledge, an important step for the discovery of nativelike selections. One can easily think of "overall opinion" as a recurring topic in the texts, but other latent topics are not obvious like it.

Furthermore, the crosslinguistic comparison makes it salient that these topic words form nativelike selections in their respective languages. To express the same concept within a given topic, different languages may prefer different expressions. This means that in each language, certain expressions are preferred over alternatives that may instead be nativelike selections in another language. The most prominent differences are found in "skin feel", where *gentle*, *shang* 'harm', and *-are-* 'chap' serve as the respective nativelike selections for English, Chinese, and Japanese to praise the product's skin feel (i.e., *gentle on hands*, *bu shang shou* 'not harm hands', and *te ga are-nai* 'hands do not chap'). An extreme example is that the translation equivalents of *shang* (*harm* for English, and *kizutsukeru* for Japanese) appear only once in the English data, and 0 time in the Japanese data to describe the skin feel.

Regarding "overall opinion", *good* (*hao* for Chinese and *ii* for Japanese) is a nativelike selection in all three languages to express the positive overall opinion. However, the individual language shows its own additional nativelike selections of this concept such as *love* in English, and *bucuo* 'not bad' in Chinese. Under the topic of "environmental concern", the respective nativelike selections for English, Chinese, and Japanese are *eco/environment friendly*, *huanbao* (lit 'environmentally protective'), and *kankyoo/shizen ni yasashii* 'gentle to environment/nature'. Finally, compared with Chinese and Japanese, English native speakers prefer some special nativelike selections to describe the good scent of this product, which are *lovely*, *nice*, and *amazing* (e.g., *lovely smell*). It is worthwhile to note that *lovely* and *amazing* are hard to translate into Chinese and Japanese in this context without losing their nuances.

5. Conclusion

This study applied the topic modeling technique of BERTopic to the automatic discovery of nativelike selections in a multilingual context, using customer reviews of a dish soap product. The method's practicality was validated by the experiment's results. It effectively discovers frequent words used to express the same concept within a narrow topic and efficiently processes large text collections to identify recurring latent topics, underlying concepts, and corresponding frequent expressions—tasks that would be labor-intensive if performed manually. The limitations of this study should be acknowledged. The generalizability of this method can be further studied by applying it to text materials across diverse genres. In summary, this study offers practical benefits for developing language learning resources and lexicography, and also contributes to our understanding of the phenomenon of nativelike selection in our daily language life.

Funding:This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Chen, W., Rabhi, F., Liao, W., & Al-Qudah, I. (2023). Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study. *Electronics*, *12*(12). https://doi.org/10.3390/electronics12122605

[2] Deng, Y., & Liu, D. (2022). A multi-dimensional comparison of the effectiveness and efficiency of association measures in collocation extraction. *International Journal of Corpus Linguistics*, *27*(2), 191–219. https://doi.org/10.1075/ijcl.19111.den

[3] Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

[4] Erman, B. (2009). Formulaic language from a learner perspective: What the learner needs to know. In Roberta Corriga, Edith A. Moravcsik, Hamid Ouali, & Kathleen Wheatley (Eds.), *Formulaic languagee: Vol. 2. Acquisition, loss, psychological reality, and functional explanations* (pp. 323–346). John Benjamins.

[5] Erman, B., Forsberg Lundell, F., & Lewis, M. (2016). Formulaic language in advanced second language acquisition and use. In K. Hyltenstam (Ed.), *Studies on Language Acquisition* (pp. 111–148). De Gruyter Mouton.

[6] Foster, P. (2009). Lexical diversity and native-like selection: The bonus of studying abroad. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 91–106). Palgrave Macmillan. https://doi.org/10.1057/9780230242258_6

[7] Geeraerts, D., Stefan Grondelaers, & Peter Bakema. (1994). *The structure of lexical variation: Meaning, naming, and context*. De Gruyter Mouton.

[8] Gries, S. T. (2022). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1), 1–33. https://doi.org/10.1075/jsls.21028.gri

[9] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://doi.org/10.48550/arXiv.2203.05794

[10] Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

[11] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). *Towards general text embeddings with multi-stage contrastive learning*. arXiv. https://doi.org/10.48550/arXiv.2308.03281

[12] Liu, D. (2013). Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns. *Cognitive Linguistics*, *24*(1), 67–113. https://doi.org/10.1515/cog-2013-0003

[13] Mervaala, E. (2025). Climate change versus economic growth: Quantifying, identifying and comparing articulations in news media using dynamic topic modeling. *Environmental Communication*, *19*(1), 1–23. https://doi.org/10.1080/17524032.2025.2458222

[14] Ortaçtepe, D. (2013). Formulaic language and conceptual socialization: The route to becoming nativelike in L2. *System*, *41*(3), 852–865. https://doi.org/10.1016/j.system.2013.08.006

[15] Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication*. Longman.

[16] Smiskova, H., Verspoor, M., & Lowie, W. (2012). Conventionalized ways of saying things (CWOSTs) and L2 development. *Dutch Journal of Applied Linguistics*, *1*(1), 125–142. https://doi.org/10.1075/dujal.1.1.09smi

[17] Tsukagoshi, H., Sasano, R., & Takeda, K. (2023). *Japanese SimCSE technical report*. arXiv. https://doi.org/10.48550/arXiv.2310.19349

[18] Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*, 231–254. Cambridge Core. https://doi.org/10.1017/S026719051200013X

[19] Xu, W. W., Tshimula, J. M., Dubé, È., Graham, J. E., Greyson, D., MacDonald, N. E., & Meyer, S. B. (2022). Unmasking the twitter discourses on masks during the COVID-19 pandemic: User cluster–based BERT topic modeling approach. *JMIR Infodemiology*, *2*(2), e41198. https://doi.org/10.2196/41198

[20] Zaytseva, V. (2016). *Vocabulary acquisition in study abroad and formal instruction: An investigation on oral and written lexical development* [Dissertation]. Universitat Pompeu Fabra.