Scientific Research Publishing

Open Journal of Modern Linguistics, 2020, 10, 722-739
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

A Comparative Study of Keywords and Sentiments of Abstracts by Python Programs

Penghua Zhang[1*], Yi Pan[2]

[1]School of Foreign Languages, Xidian University, Xi'an, China
[2]School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China
Email: *roczp919@gmail.com, panyilook@163.com

How to cite this paper: Zhang, P. H., & Pan, Y. (2020). A Comparative Study of Keywords and Sentiments of Abstracts by Python Programs. *Open Journal of Modern Linguistics, 10,* 722-739.
https://doi.org/10.4236/ojml.2020.106044

Received: October 16, 2020
Accepted: November 22, 2020
Published: November 25, 2020

Abstract

Four corpora are created to investigate the self-mentions, keywords and sentiment of abstracts. First, self-mentions are categorized to examine the authorial interactions with the reader. Then, the study of high-frequency words and keywords is conducted with different Python programs and the software AntConc. The keywords generated with WordCloud and TF-IDF-LDA methods show a definite relation with high-frequency words generated by Jieba_Counter and NLTK FreqDist. Further, the sentiment analysis is performed with SnowNLP and TextBlob yielding different results, which verifies the authorial interactions with the reader and increased factual information respectively. Finally, the verification by reference corpora validates the consistency of the sentiment analysis by these two methods. The research suggests that the methods for high-frequency words generation, keywords generation and sentiment analysis be selected discriminatively since different methods generate different results; meanwhile, the study verifies that the objectivity remains in the writing of abstracts. The investigation is conducive to the choices of keywords generation and self-mentions in writing.

Keywords

Self-Mentions, Generation of Keywords, Sentiments, SnowNLP, TextBlob, Objectivity

1. Introduction

Abstracts, as an indispensable part of academic writing, serve as a stepstone for the reader to find important information and to decide whether to read the whole paper or not. They are studied by the Swalesian School under the framework of move patterns IMRD and more variations of move patterns

ranging from 5 to 13 moves. The research materials cover either single subject or the multi-disciplines from soft sciences to hard sciences. Some scholars have conducted researches on various moves and their realization schemes; and some others use corpus linguistic method to explore such micro-features as frequencies of words, formulaic expressions, voices, tenses, and syntactic patterns under different move patterns. Still, some researchers compare abstracts interculturally from the perspectives of two or more different languages.

As a quite important part of academic writing, keywords of the paper are required by almost all the journals. They are usually chosen intuitively by the author by words' frequency or importance; or some authors generate them automatically by some software, though they are significant for the reader to obtain key information in the paper. However, little literature on the automatic generation of keywords by Python has been presented to show their reliability.

Furthermore, the views on objectivity and subjectivity of such academic writing as abstract writing have undergone great changes in the recent three decades. However, the views are seldom verified quantitively to show the objectivity or subjectivity, as Hyland (2005) holds that the authorial interactions with the reader in abstracts are realized through various self-referencing strategies. Some related sentiment analysis can be found in these scholar's works (Batool et al., 2013; Chong et al., 2014; Tyagi & Tripathi, 2019; Dubey, 2020). However, few researches have been performed on the sentiments of abstracts to show the interactions between the author and the reader.

Considering these situations in academic writing of abstracts and keywords, we select abstracts from several international agricultural journals to investigate the variations of keywords and the differences of sentiments of abstracts by different libraries of Python.

The previous studies on keywords generation can be traced in these scholars' works (Joshi & Motwani, 2006; Thomaidou & Vazirgiannis, 2011; Hussey et al., 2012; Liu et al., 2014; Savva et al., 2014; Scholz, et al., 2019; Arora & Kumar, 2019; Zheng & Sun, 2019; Thushara et al., 2019). Among the works by these researchers, Scholz, et al. (2019) propose an automated approach for generating keywords for Sponsored Search Advertising based on his keyword generation algorithms. Zheng & Sun (2019) utilize the three properties of relevance, coverage, and evolvement of candidate keywords by using active learning and multiple-instance learning to follow up the main topics of tweets along the development of events. Few researchers have assessed the reliability of the keyword generated by different Python programs.

The self-mention, or self-reference referred in some literature, includes personal pronouns in Hyland's (2001, 2005) framework. The categories of self-mentions can be found in different researchers' works (Ivanic, 1998; Tang & John, 1999; Hunston, 2000; Liu, 2011). We selected the terms as the indication of self-mentions listed in **Table 1**, which are considered as the strategies for achieving authorial interactions with the reader, showing the authorial sentiments as well. In

this table, we have excluded the concealed entity *It is + past participle/adj* because of its rare occurrences as proved by our previous studies.

Table 1. Classification of self-mentions and their markers.

Entity		Marker of entity
Human entity	First person pronoun & determiner	I, we, me, us, my, our
	Third person pronoun	author, writer, researcher
Inanimate entity	Research-oriented noun	study, research
	Discoursal noun	paper, thesis, article

Regarding the research on the self-reference in abstracts, the interdisciplinary-oriented self-reference research includes the works by these scholars (Yeo & Ting, 2014; Khedri, 2016; Lancaster, 2016; McGrath, 2016; Seoane & Hundt, 2017). And the researches cover the disciplines of both social science—politics (Albalat-Mascarell & Carrió-Pastor, 2019), anthropology and history (McGrath, 2016), sociology (Bruce, 2010; Işık-Taş, 2018), business management (Mur-Dueñas, 2011), marketing (Khedri, 2016), economics (Carter-Thomas & Chambers, 2012; Lancaster, 2016), language studies (Chen, 2020), and applied linguistics (Bonn & Swales, 2007; Molino, 2010; Zareva, 2013; Karahan, 2013; Walková, 2019), and natural science—electrical engineering (Hyland, 2000), computing science (Shehzad, 2007; Soler-Monreal, 2015), medical science (Li & Ge, 2009), biomedical science (Carciu, 2009; Kanoksilapatham, 2015), and agricultural engineering (Gheinani & Tabatabaei, 2017). Although the self-reference in research abstracts has been studied from the perspectives of social sciences and natural sciences, and the study of abstracts in agricultural engineering can somehow be represented by the other subjects in natural sciences, the self-reference in agriculture-specific abstracts is still under-researched.

In terms of genres, the study of self-mentions ranges from research articles (Khedri, 2016; McGrath, 2016; Wu & Zhu, 2015; Chen, 2020), research article abstracts (Friginal & Mustafa, 2017; Bonn & Swales, 2007), presentations (Zareva, 2013), speech (Albalat-Mascarell & Carrió-Pastor, 2019) and introduction part in academic writing (Loi, 2010; Wang & Yang, 2015; Tankó, 2017), to literature review (Soler-Monreal, 2015), and personal statement (Li & Deng, 2019). However, these studies of self-mentions have not covered the sentiments of abstracts and their self-reference.

This paper intends to compare the keywords generated by WordCloud and TF-IDF-LDA and to explore the sentiments of abstracts generated by SnowNLP and TextBlob to show whether the programs are reliable or not, and whether authorial interactions with the reader can be improved by self-mentions.

The next part Section 2 of the paper is the research materials and methods of the study, followed by the results of the high-frequency words by Jieba_Counter, NLTK FreqDist and AntConc (v.3.5.8) (Anthony, 2019), WordCloud keywords TF-IDF-LDA topic keywords and sentiment analysis by SnowNLP and TextBlob.

Section 4 is the discussion of the research results and the last part is the conclusion of the paper.

2. Research Materials and Methods

2.1. Materials

The research materials we selected are the abstracts from international agricultural journals. The journal names, country names of the author and quantity of abstracts in each journal are listed in Table 2.

Table 2. Journals surveyed in the study.

Journal Title	Quantity CHC (468)	Quantity INC (460)	Authors Area in INC
Agronomy Journal (bimonthly)	200	200	US/Canada/UK/Australia/Ireland
Journal of Agricultural Science and Technology (quarterly)	68	60	US/Canada/UK/Australia/New Zealand
Agronomy (quarterly)	200	200	US/UK/Australia/Canada

The raw materials include the information of the author, journal name, volume, year, URL, ISSN and DOI, which are deleted with Python programs to retain abstracts' content only in order to obtain keywords and sentiments of abstracts. Altogether, 460 abstracts have been selected from the journals and 451 abstracts have been obtained after being processed by programs. A corpus named as INC is created with these abstracts. In order to verify the sentiment results we build another corpus named as CHC as reference corpus, which is made of 462 abstracts processed from 468 abstracts by Python programs. These two corpora are raw corpus without annotation and POS tags.

The overall statistics of the two corpora are listed in Table 3:

Table 3. Overview of corpora INC & CHC.

Corpora	Words	Sentences	ASL	LD
INC	128,411	4328	29.67	11.79
CHC	134,202	4365	30.75	11.69

ASL—Average sentence length; LD—Lexical diversity.

These data in the table are obtained by using NLTK sentence tokenizer and word tokenizer. The data show that the two corpora are largely comparable and therefore CHC can serve as the reference corpus for INC.

2.2. Methods

After building the corpus of INC, we investigate the high-frequency words and keywords of INC by using various methods comparatively. The high-frequency words are generated by Jieba_Counter and NLTK FreqDist, and then cross-checked

and confirmed by the software AntConc. The keywords are calculated by WordCloud and TF-IDF-LDA methods respectively.

We adopt Jieba_Counter and NLTK FreqDist methods as per the following procedures. First, Jieba is used to split words with NLTK stopwords loaded, and then the symbols and punctuations are deleted, and finally Counter is imported to generate a list of high-frequency words. NLTK FreqDist high-frequency words are generated with NLTK Word Tokenize and FreqDist, with NLTK stopwords and these additional stopwords and symbols "'.', '(', 'NO', 'The', ')', 'N', '%', '&', ';', ',', '1', 'l.', 'n', 'kg'" loaded, and then the high-frequency words are visualized by Numpy and Matplotlib. Only the first 15 high-frequency words are selected for the visualization. The results of Jieba_Counter and NLTK FreqDist are cross-checked and confirmed by the wordlist and frequency of AntConc.

WordCloud keywords and TF-IDF-LDA topic keywords are produced in the following procedures. First, NLTK Word Tokenize is used to split the abstract into words with NLTK stopwords and additional stopwords "'and', 'of', 'the', 'in the', 'for', 'in', 'with', 'at', 'by', 'under', 'wa', 'were', 'as', 'on', 'to the', 'kg', 'ha'" loaded, and then Python WordCloud and Matplotlib are imported to visualize the first 15 high-frequency words. TF-IDF-LDA topic keywords are generated by using TF-IDF and LDA model with NLTK stopwords loaded. Five topics are designed for generating 10 keywords in each topic upon several tests with more stopwords added.

In order to calculate the sentiments of texts with self-mentions, first we extract the whole sentences and phrases with self-mentions and then create two corpora named as INSM and CHSM for INC and CHC respectively. The frequencies of each self-mention in INC and CHC are shown in Table 4.

Table 4. Self-mentions in INSM & CHSM.

Corpora	We	Us	Our	Study	Research	Paper	Article	TTL
CHSM	103	4	29	232	27	13	2	410
INSM	179	0	78	252	73	7	1	590

We have excluded those self-mentions that do not refer to the author to ensure that the self-mentions reflect the authorial interactions with the reader. The overall statistics of INSM and CHSM are listed in the following Table 5.

Table 5. Overview of corpora INSM & CHSM.

Corpora	Words	Sentences	ASL	LD
INSM	10,957	590	18.57	4.28
CHSM	7423	410	18.11	3.87

The statistics in this table are calculated by the same way as those in INC and CHC. These data suggest CHSM can be used as reference corpus in terms of the comparability.

Sentiment analysis is conducted with SnowNLP and TextBlob respectively.

These two libraries are for sentiment analysis, but they have different scoring systems to show the sentiments of texts; therefore, we use both libraries to compare the sentiment results quantitively. Firstly, the sentiment analysis of INC and CHC are performed, and then the sentiment analysis of INSM and CHSM are conducted respectively in order to explore whether the sentiment scores reflect the authorial interactions with the reader. The average sentiments are also calculated by Numpy and visualized by Matplotlib with the visualized sentiments of the texts in these corpora.

3. Results

3.1. High-Frequency Words and Keywords in INC

The high-frequency words generated by NLTK FreqDist in INC are *yield*, *soil*, *crop*, *water*, *study*, *production*, *cover*, *increased*, *minus*, *results*, *growing*, *grain*, *biomass*, *compacted*, and *two* as shown in Figure 1.

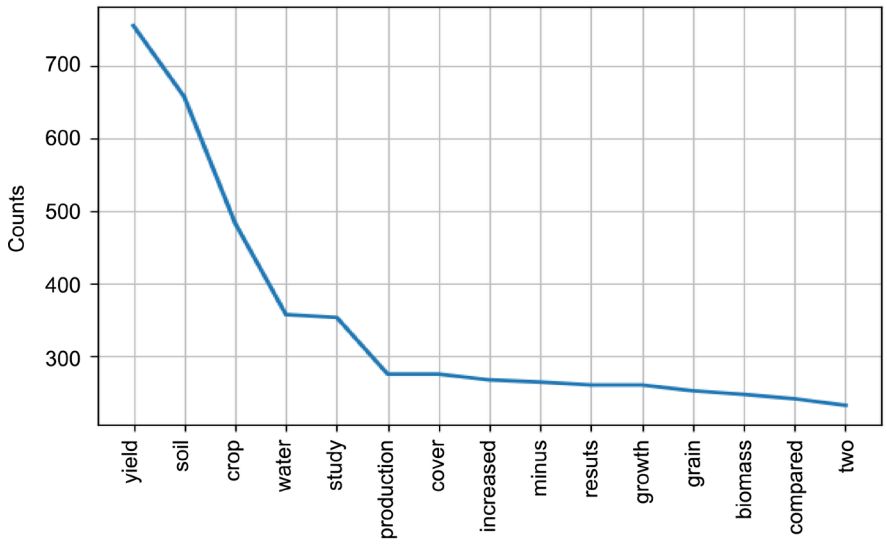

Figure 1. INC high-frequency words by NLTK FreqDist.

The keywords generated by WordCloud in INC are *yield*, *soil*, *plant*, *cover*, *crop*, *rate*, *treatment*, *using*, *cultivar*, *model*, *increased*, *this study*, *year*, *field*, and *high*, which are shown in Figure 2.

Figure 2. INC WordCloud keywords.

The high-frequency words calculated by Jieba_Counter, AntConc, NLTK FreqDist and WordCloud keywords are listed in Table 6.

Table 6. High-frequency words by Jieba_Counter, AntConc, FreqDist and keywords by WordCloud.

Jieba_C	Frequency	AntConc	Frequency	FDist	WordCloud
yield	767	yield	768	yield	yield
soil	678	soil	678	soil	soil
crop	506	crop	506	crop	plant
water	368	water	368	water	cover
study	354	study	355	study	crop
cover	280	cover	280	production	rate
production	277	production	277	cover	treatment
increased	267	increased	267	increased	using
minus	265	minus	265	minus	cultivar
growth	264	growth	264	results	model
plant	262	plant	263	growing	increased
results	260	results	260	grain	this study
grain	254	grain	253	biomass	year
two	248	biomass	248	compacted	filed
biomass	248	two	248	two	high

Jieba_C—JiebaCounter words; FDist—FreqDist words; WordCloud—WordCloud keywords.

The keywords with their respective TF-IDF values generated by TF-IDF-LDA are listed in the following.

(0, "0.017 * 'stress' + 0.014 * 'growth' + 0.014 * 'data' + 0.013 * 'crop' + 0.011 * 'weight' + 0.011 * 'heat' + 0.009 * 'agricultural' + 0.007 * 'fruit' + 0.007 * 'except' + 0.007 * 'effect'"),

(1, "0.018 * 'model' + 0.012 * 'among' + 0.012 * 'moisture' + 0.010 * 'potential' + 0.009 * 'content' + 0.009 * 'lower' + 0.008 * 'date' + 0.007 * 'increase' + 0.007 * 'field' + 0.007 * 'biomass'"),

(2, "0.028 * 'results' + 0.026 * 'water' + 0.022 * 'different' + 0.016 * 'showed' + 0.014 * 'used' + 0.012 * 'tolerance' + 0.011 * 'significant' + 0.010 * 'observed' + 0.010 * 'plants' + 0.010 * 'based'"),

(3, "0.039 * 'soil' + 0.020 * 'yield' + 0.018 * 'using' + 0.015 * 'high' + 0.015 * 'compared' + 0.013 * 'higher' + 0.013 * 'rainfall' + 0.012 * 'we' + 0.012 * 'temperature' + 0.011 * 'increased'"),

(4, "0.016 * 'study' + 0.014 * 'two' + 0.014 * 'plant' + 0.012 * 'production' + 0.011 * 'total' + 0.010 * 'significantly' + 0.010 * 'clay' + 0.009 * 'values' + 0.008 * 'four' + 0.008 * 'wild'").

The first three words with highest TF-IDF values in each topic are *stress*, *growth*, *data*, *model*, *moisture*, *potential*, *results*, *water*, *different*, *soil*, *yield*, *using*, *study*, *two*, and *plant.*

3.2. Sentiment Results

The sentiment analysis results of INC and INSM obtained by SnowNLP are

shown in Figure 3 and Figure 4. The sentiment scores are between the limits of 0 and 1, 0 being negative and 1 positive in sentiments.

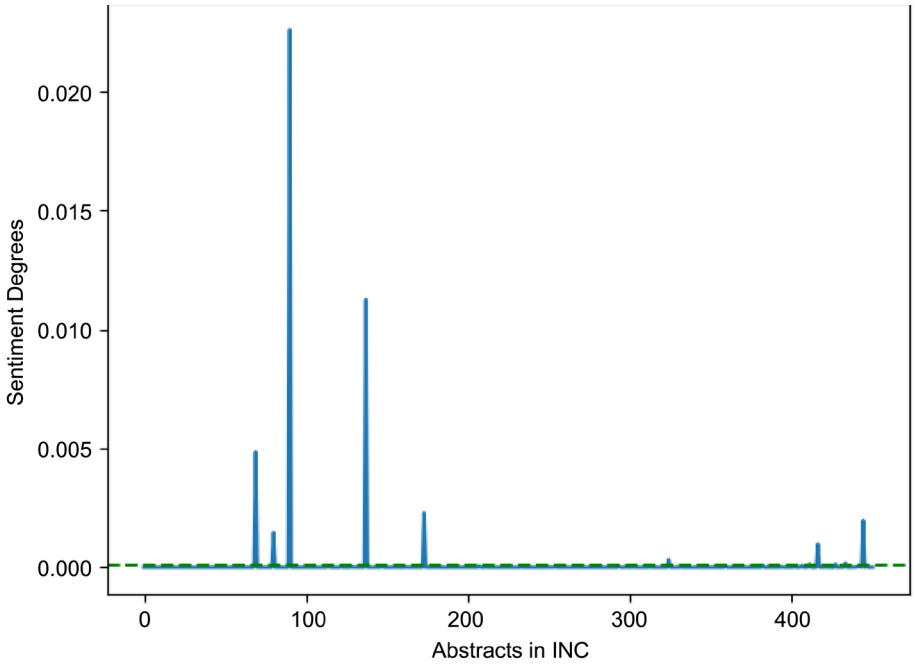

Figure 3. Sentiments of abstracts in INC by SnowNLP.

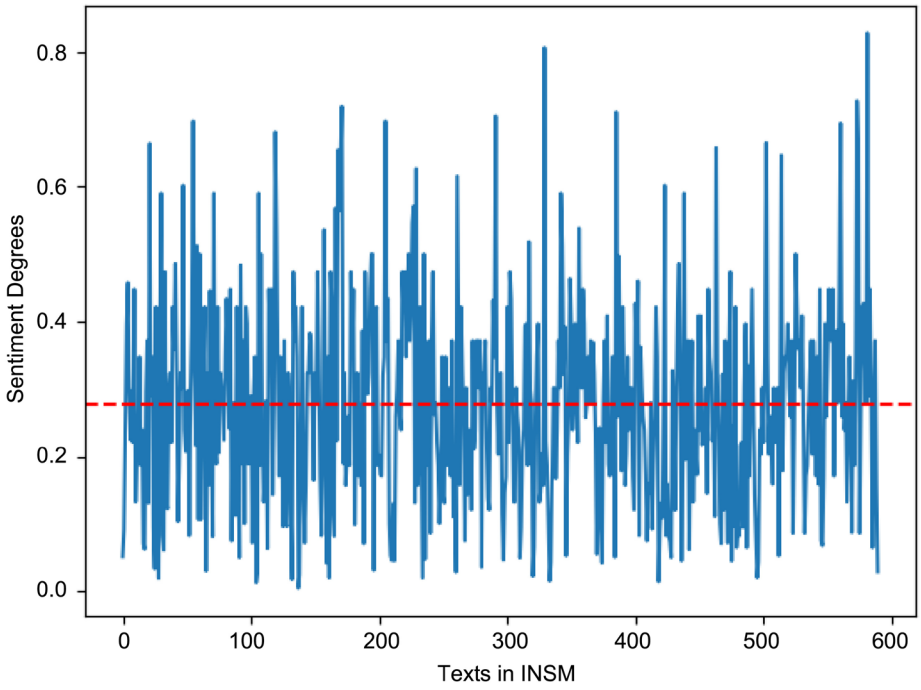

Figure 4. Sentiments of texts in INSM by SnowNLP.

The green dotted line in the figure shows the average sentiment (0.00010516297819778477) of the abstracts in INC.

The red dotted line shows the average sentiment (0.27732966047183893) of

texts in INSM.

The sentiment scores obtained from TextBlob are shown in Figure 5 and Figure 6, of which the blue lines indicate the polarity and yellow lines indicate the subjectivity of the texts.

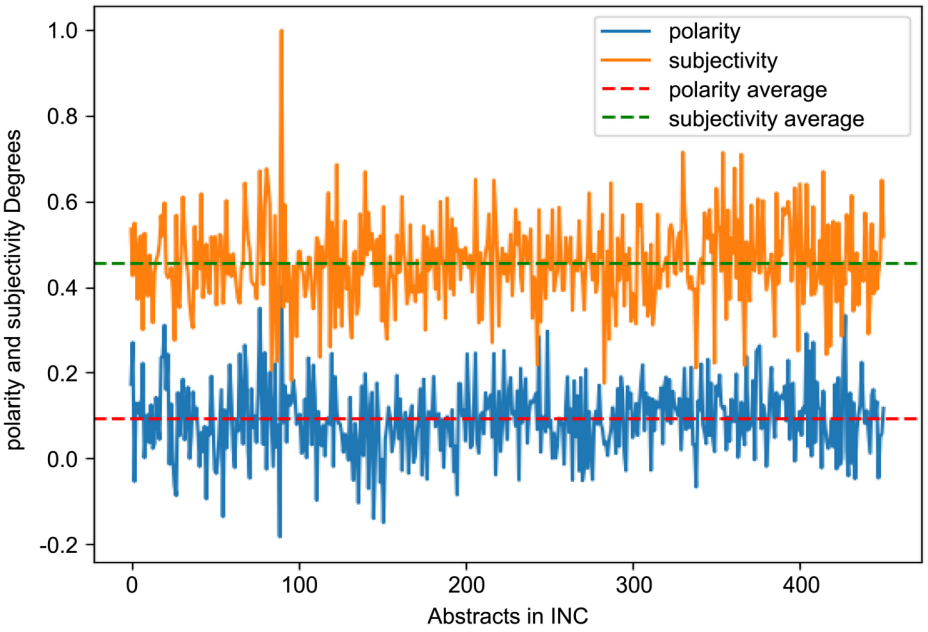

Figure 5. Sentiments of abstracts in INC by TextBlob.

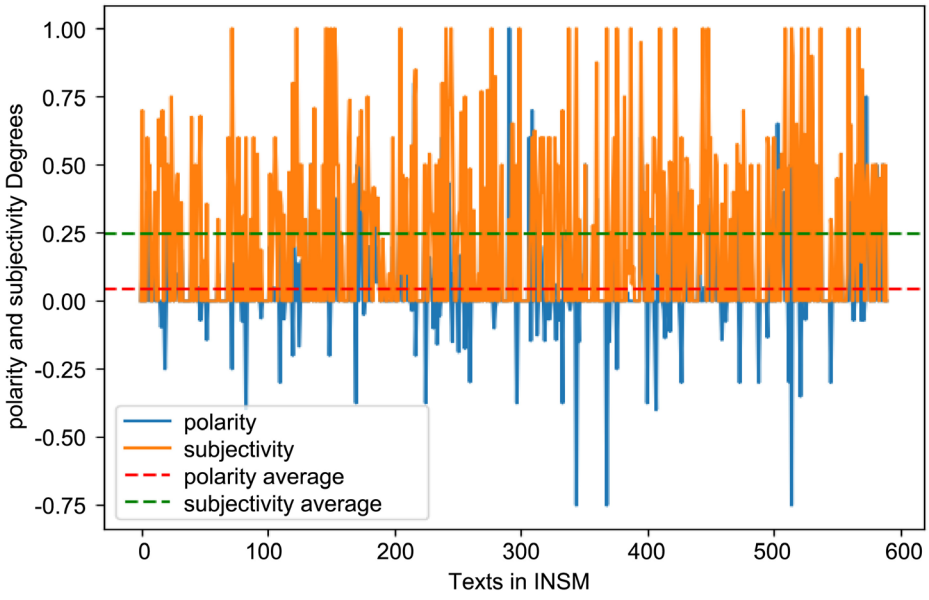

Figure 6. Sentiments of texts in INSM by TextBlob.

The green dotted lines show the average subjectivity (INC average subjectivity 0.456681, INSM average subjectivity 0.245443) and the red dotted lines the average polarity (INC average polarity 0.090902, INSM average polarity 0.042876). The subjectivity suggests the proportion of facts and opinions and the polarity shows

sentiments of texts. The values of subjectivity are between 0 and 1, 0 indicating more facts and 1 more personal opinions. The limits of polarity are between −1 and 1, the higher value signifying positive and the lower value negative in sentiments.

The sentiments of the texts in reference corpora CHC and CHSM by SnowNLP and TextBlob are shown in Figures 7-10.

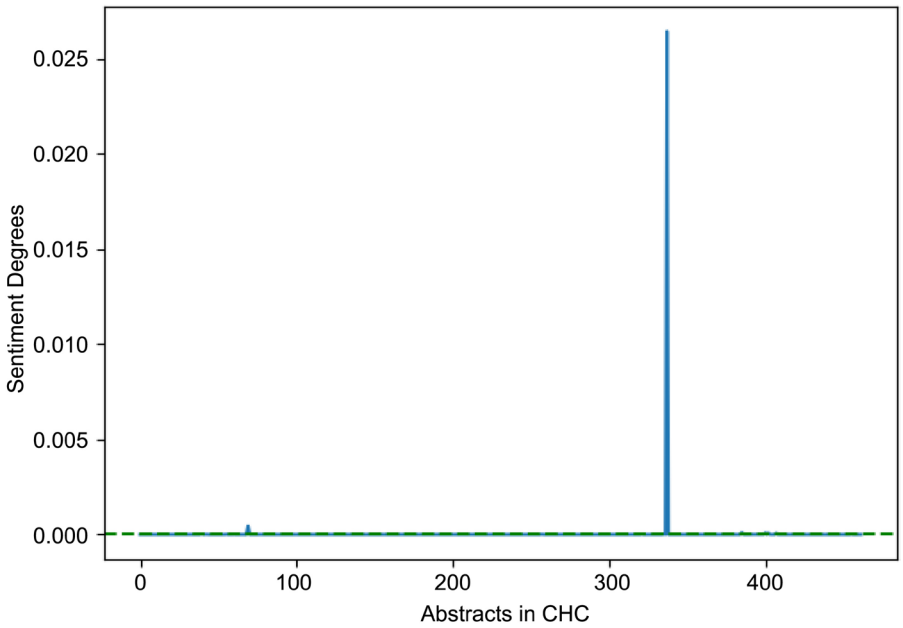

Figure 7. Sentiments of abstracts in CHC by SnowNLP.

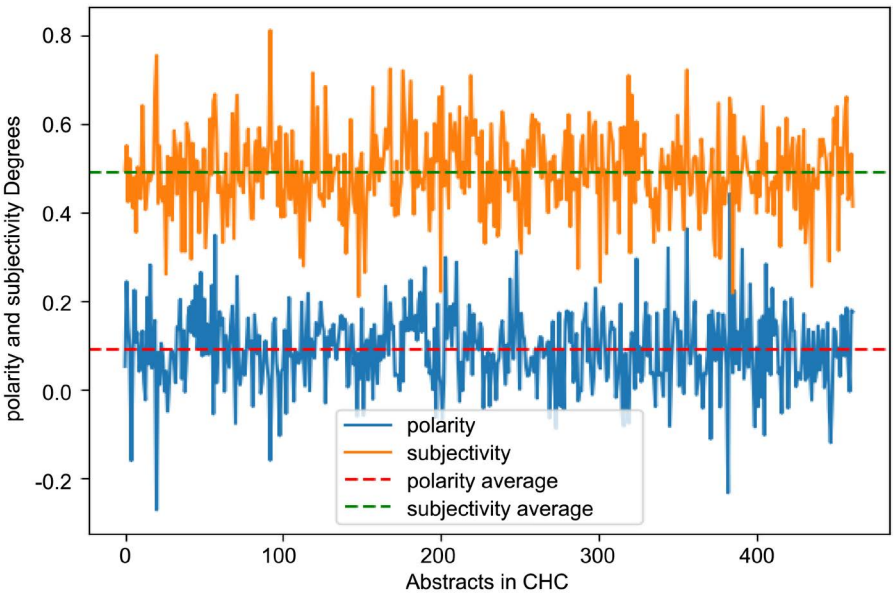

Figure 8. Sentiments of texts in CHSM by SnowNLP.

The average sentiments of CHC and CHSM by SnowNLP are 0.00005938709819871512 and 0.2656879456177752 respectively.

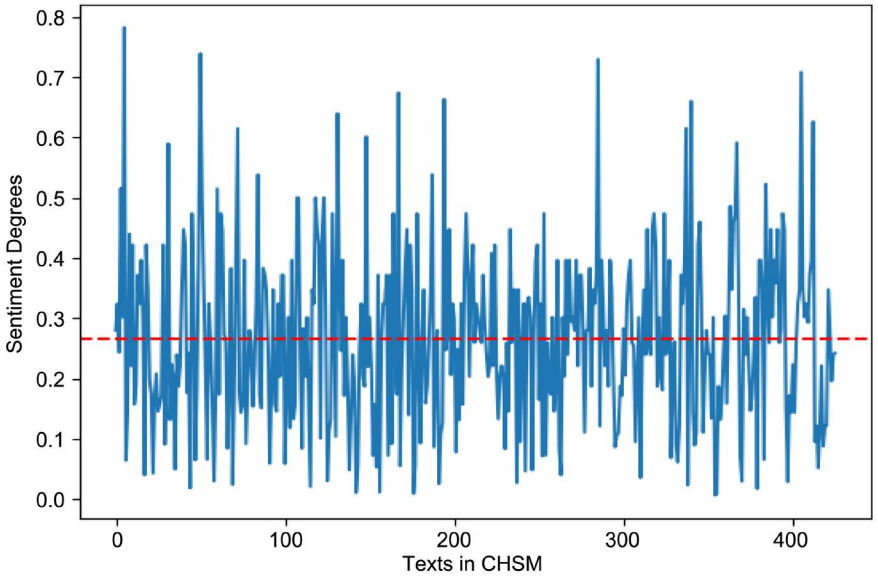

Figure 9. Sentiments of abstracts in CHC by TextBlob.

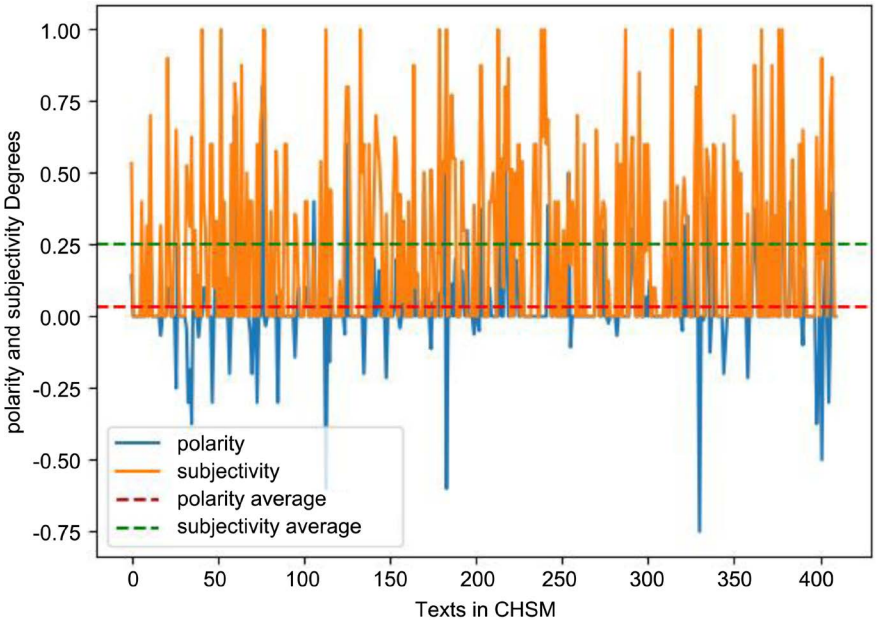

Figure 10. Sentiments of texts in CHSM by TextBlob.

The average polarity values of CHC and CHSM by TextBlob are 0.089434 and 0.032679 respectively and the average subjectivity values of CHC and CHSM are 0.490393 and 0.250492 respectively.

4. Discussions

4.1. High-Frequency Words and Keywords in INC

The high frequency words generated by Jieba_C and AntConc are the same, with frequency differences in words *yield*, *study*, and *plant* greater than 1, and *grain* less than 1 by AntConc. We search the words with Notepad manually to find

that the respective counts of these words are 767, 353, 258, and 253, of which both methods give wrong counts for *study* and *plant*, and that Jieba_C and AntConc give the right frequency of words *yield* and *grain* respectively. The high-frequency words generated by FDist are nearly the same as those given by Jieba_C and AntConc except for the word *compacted* in the place of *plant* by the Jieba_C and AntConc. These data show that the high-frequency words by Jieba_C and AntConc are reliable; however, their frequencies may need further verification, and the words generated by FDist needs further confirmation.

The keywords generated by WordCloud and TF-IDF-LDA are overlapping with those high-frequency words given by Jieba_C and AntConc, which shows that the keywords have a definite relation with the high-frequency words. WordCloud generates 6 keywords *yield*, *soil*, *plant*, *cover*, *crop*, and *increased*, same as those in high-frequency words, and yields these 9 different words and phrases *rate*, *treatment*, *using*, *cultivar*, *model*, *increased*, *this study*, *year*, and *filed* from those of high-frequency words.

Of the 15 keywords out of the first 3 words in each topic, the keywords that are the same as high-frequency words are the 8 words *growth*, *results*, *water*, *soil*, *yield*, *study*, *two*, and *plant* and those that are not in the high frequency words are the 7 words *stress*, *data*, *model*, *moisture*, *potential*, *different*, and *using*. The two common keywords are *using* and *model* by these two methods. The keywords from those high-frequency words by WordCloud and TF-IDF-LDA take 40% and 53.3% of the total high-frequency words respectively, though the topic keywords are calculated with different methods.

The topic keywords may provide us with more perspectives to show the key information in a text, while WordCloud keywords may be eye-catching. Topic keywords may change with different topics assigned, which need tests and trials and human judgement on the different outcomes. As the TF-IDF values are given with the keywords in each topic, we can judge and compare the keywords quantitively, which is an advantage over the WordCloud keywords. The dynamic change of keywords with topics makes it possible the multi-views of keywords in different topics. With different language models applied, there unfolds a different picture of keywords, which will definitely deepen our understanding of the text.

4.2. Sentiments by SnowNLP and TextBlob

The average sentiments of INC are rather low compared with those of INSM because the texts in INSM are with self-mentions. These self-mentions definitely raise the personal opinions of the texts, which are shown by the average sentiment score in INSM. SnowNLP scores take 0.5 as neutral, while the average sentiment score of INSM is way below the neutral point, which shows that the texts in INSM are objective, though the sentiments have been raised fairly high with self-mentions in the texts.

Compared with the sentiments of CHC and CHSM, the average sentiment

level of CHSM has also been raised with the use of self-mentions in CHSM, which can confirm that self-mentions in texts improve the authorial interactions with the reader. Further study of the sentiment data of INSM and CHSM shows that they are of the same order, both being around 0.27, with the score of INSM slightly higher than that of INC.

Noticeably, the average sentiments of INC and CHC are rather low, with the average score of CHC being nearly half that of INC. Both scores have been raised over 2641 times with the use of self-mentions in INSM and CHSM, which may suggest that self-mentions are a powerful device to highlight personal opinions. The rather low values of CHC and INC may indicate little sentiments in these two corpora; however, the greatly-increased values of INSM and CHSM are still about half the neutral point. This may also imply that self-mentions function limitedly as a strategy for improving the authorial interactions with the reader.

The sentiment results by TextBlob show that the average subjectivity of INC has been lowered by 53.7% with the use of self-mentions in INSM, which indicates that the texts in CHSM are with fewer opinions and more facts in contrast with CHC. This can be reasonable in that the self-mentions in CHSM contain more inanimate entities than human entities, which may lower the subjectivity of the sentiment scores. Meanwhile, this may also suggest that self-mentions in CHSM might have strengthened the factual statements.

The average polarity of INC by TextBlob has also been decreased by 47.2% with the use of self-mentions in INSM, which suggests that the self-mentions in INSM serve as a strategy for lowering the sentiments of texts, thus showing little authorial interactions with the reader. In view of the self-mentions' constituents, the inanimate entities take more proportion than the human entities, which might cause the decrease of sentiments.

In order to confirm the sentiment results by TextBlob, tests of sentiment analysis have been performed of CHC and CHSM, which results in similar data. The average subjectivity and polarity of CHC and CHSM are of the same order as those in INC and INSM. The average polarity in CHC is decreased by 36.5% and the average subjectivity is decreased by 51.1% by the use of self-mentions in CHSM. The tests with CHC and CHSM may validate that the sentiment analysis by TextBlob is consistent with that of INC and INSM. However, the problem arises which sentiment analysis is reliable since both SnowNLP and TextBlob yield seemingly reasonable results. The other problem is also thought-provoking that self-mentions are devices for raising the sentiments of texts with SnowNLP, but they function as a strategy for lowering the subjectivity and polarity with TextBlob. Why do these two methods yield different outcomes?

We have checked the program many times to ensure the programs are correct and the materials in these corpora are proper. As a result, no faults have been found with the programs and materials. Then we reexamined the source codes of SnowNLP and TextBlob to find that TextBlob is NLTK-based, while SnowNLP is not, although both methods are Naïve-Bayers-based. In addition, another library

Pattern Analyzer of TextBlob may also contribute to the differences of the two methods. Theoretically speaking, the sentiments results by SnowNLP agree with Hyland (2005) claim self-mentions improve the authorial interactions with the reader, while the sentiment results by TextBlob suggest the personal opinions have been decreased with the use of self-mentions. The seemingly paradoxical results given by the two methods may attribute to their different mechanisms since the research materials are the same in these tests. Therefore, it is hard to conclude which method is more reliable or better than the other one. Undoubtedly, we may find different perspectives of sentiments by different methods.

5. Conclusion

We have examined different methods for generating high-frequency words, keywords and sentiments of abstracts. The high-frequency words generated by Jieba_C and AntConc are reliable; however, the frequencies need further verification and FDist high-frequency words are mostly reliable, with some needing further confirmation. The keywords by WordCloud and TF-IDF-LDA overlap with some of those high-frequency words, which shows a definite relation between high-frequency words and keywords.

Different methods for generating keywords yield different results; however, the TF-IDF-LDA method can demonstrate dynamic topic keywords for us to select the best combinations. LDA language model is one of the language models for generating keywords, and other language models can also be applied to generate keywords, which will show more scenarios. Keywords generation is somehow like casting dices, which may result in different results, and can show us a broader view of the text. The high-frequency words in combination with keywords can help us obtain the key information in the text and they may help the writer to write the keywords for the paper.

The sentiment analysis by SnowNLP and TextBlob yields different results, which can function differently, since they have different parameters for us to see through the sentiments of texts. SnowNLP results agree with the theoretical assumption that self-mentions improve the authorial interactions with the reader. This finding also suggests that more opinions exist in the texts with the use of self-mentions. TextBlob sentiment results support more factual information and fewer opinions are in the texts with self-mentions. These two methods support different views on self-mentions, which are seemingly controversial.

This paradox can be due to the mechanism of the two methods, and the controversy may also be caused by the different constituents of self-mentions as analyzed in the discussion. Recognizing the differences of the two methods, we may use the methods discriminatively. Based on the findings in sentiments by SnowNLP and TextBlob, we may conclude that the objectivity of abstracts remains as before, though the authorial interactions with the reader have been greatly increased with the use of self-mentions.

This study can help the writer select proper approaches to generating the

keywords of the paper on the one hand; on the other hand, it may also give some implications for the writer to choose proper self-mentions in order to enhance the authorial interactions with reader. Meanwhile, this approach may broaden the researchers' horizon to adopt Python programs to study the automatic keywords generation and sentiments of academic texts. More libraries and language models can be imported to explore the keywords and sentiments of academic texts quantitively.

Fund

The study is supported by the cross-disciplinary project in humanities and information (project name: Research on the Corpus-based Translation Universals of Abstracts and Their Application in Computer-aided Translation) of Xidian University (project No. RW180180).

Acknowledgements

The authors would like to thank Xidian University for his funding the research upon which the article is based.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Albalat-Mascarell, A., & Carrio-Pastor, M. L. (2019). Self-Representation in Political Campaign Talk: A Functional Metadiscourse Approach to Self-Mentions in Televised Presidential Debates. *Journal of Pragmatics, 147,* 86-99. https://doi.org/10.1016/j.pragma.2019.05.011

Anthony, L. (2019). *AntConc (Version 3.5.8) [Computer Software]*. Tokyo: Waseda University. https://www.laurenceanthony.net/software

Arora, B., & Kumar, N. S. (2019). Automatic Keyword Extraction and Crossword Generation Tool for Indian Languages: SEEKH. *2019 IEEE Tenth International Conference on Technology for Education (T4E),* Goa, 9-11 December 2019, 272-273. https://doi.org/10.1109/T4E.2019.00070

Batool, R., Khattak, A. M., Maqbool, J., & Lee, S. (2013). Precise Tweet Classification and Sentiment Analysis. *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS),* Niigata, 16-20 June 2013, 461-466. https://doi.org/10.1109/ICIS.2013.6607883

Bonn, S. V., & Swales, J. M. (2007). English and French Journal Abstracts in the Language Sciences: Three Exploratory Studies. *Journal of English for Academic Purposes, 6,* 93-108. https://doi.org/10.1016/j.jeap.2007.04.001

Bruce, I. (2010). Textual and Discoursal Resources Used in the Essay Genre in Sociology and English. *Journal of English for Academic Purposes, 9,* 153-166. https://doi.org/10.1016/j.jeap.2010.02.011

Carciu, O. M. (2009). An Intercultural Study of First-Person Plural References in Biomedical Writing. *Ibérica, 18,* 71-92.

Carter-Thomas, S., & Chambers, A. (2012). From Text to Corpus: A Contrastive Analysis

of First Person Pronouns in Economics Article Introductions in French and English. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.), *Corpus-Informed Research and Learning in ESP* (pp. 17-39). Amsterdam: Benjamins. https://doi.org/10.1075/scl.52.02car

Chen, R. (2020). Single Author Self-Reference: Identity Construction and Pragmatic Competence. *Journal of English for Academic Purposes, 45,* Article ID: 100856. https://doi.org/10.1016/j.jeap.2020.100856

Chong, W. Y., Selvaretnam, B., & Soon, L.-K. (2014). Natural Language Processing for Sentiment Analysis: An Exploratory Analysis on Tweets. *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology,* Kota Kinabalu, 212-217. https://doi.org/10.1109/ICAIET.2014.43

Dubey, A. D. (2020). *Twitter Sentiment Analysis during COVID-19 Outbreak*. https://ssrn.com/abstract=3572023 https://doi.org/10.2139/ssrn.3572023

Friginal, E., & Mustafa, S. S. (2017). A Comparison of U.S.-Based and Iraqi English Research Article Abstracts Using Corpora. *Journal of English for Academic Purposes, 25,* 45-57. https://doi.org/10.1016/j.jeap.2016.11.004

Gheinani, M. T., & Tabatabaei, O. (2017). A Structural Move Analysis of the Abstract Section of ISI Articles of Iranian and Native Scholars in the Field of Agricultural Engineering. *International Journal of Research Studies in Language Learning, 7,* 3. https://doi.org/10.5861/ijrsll.2017.1864

Hunston, S. (2000). Evaluation and the Planes of Discourse: Status and Value in Persuasive Texts. In S. Hunston, & G. Thompson (Eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse* (pp. 176-207). Oxford: Oxford University Press.

Hussey, S. W., Mitchell, R., & Field, I. (2012). A Comparison of Automated Keyphrase Extraction Techniques and of Automatic Evaluation vs. Human Evaluation. *International Journal on Advances in Life Sciences, 4,* 136-153.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing.* London: Longman.

Hyland, K. (2001). Humble Servants of the Discipline? Self-Mention in Research Articles. *English for Specific Purposes, 20,* 207-226. https://doi.org/10.1016/S0889-4906(00)00012-0

Hyland, K. (2005). Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies, 7,* 173-192. https://doi.org/10.1177/1461445605050365

Işık-Taş, E. E. (2018). Authorial Identity in Turkish Language and English Language Research Articles in Sociology: The Role of Publication Context in Academic Writers' Discourse Choices. *English for Specific Purposes, 49,* 26-38. https://doi.org/10.1016/j.esp.2017.10.003

Ivanic, R. (1998). *Writing and Identity: The Discoursal Construction of Identity in Academic Writing*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/swll.5

Joshi, A., & Motwani, R. (2006). Keyword Generation for Search Engine Advertising. *Proceedings of the Sixth IEEE International Conference on Data Mining,* Hong Kong, 18-22 December 2006, 490-496. https://doi.org/10.1109/ICDMW.2006.104

Kanoksilapatham, B. (2015). Distinguishing Textual Features Characterizing Structural Variation in Research Articles across Three Engineering Sub-Discipline Corpora. *English for Specific Purposes, 37,* 74-86. https://doi.org/10.1016/j.esp.2014.06.008

Karahan, P. (2013). Self-Mention in Scientific Articles Written by Turkish and Non-Turkish Authors. *Procedia—Social and Behavioral Sciences, 70,* 305-322.

https://doi.org/10.1016/j.sbspro.2013.01.068

Khedri, M. (2016). Are We Visible? An Interdisciplinary Data-Based Study of Self-Mention in Research Articles. *Poznan Studies in Contemporary Linguistics, 52,* 403-430. https://doi.org/10.1515/psicl-2016-0017

Lancaster, Z. (2016). Expressing Stance in Undergraduate Writing: Discipline-Specific and General Qualities. *Journal of English for Academic Purposes, 23,* 16-30. https://doi.org/10.1016/j.jeap.2016.05.006

Li, L.-J., & Ge, G.-C. (2009). Genre Analysis: Structural and Linguistic Evolution of the English-Medium Medical Research Article (1985-2004). *English for Specific Purposes, 28,* 93-104. https://doi.org/10.1016/j.esp.2008.12.004

Li, Y., & Deng, L. (2019). I Am What I Have Written: A Case Study of Identity Construction in and Through Personal Statement Writing. *Journal of English for Academic Purposes, 37,* 70-87. https://doi.org/10.1016/j.jeap.2018.11.005

Liu, P., Azimi, J., & Zhang, R. (2014). Automatic Keywords Generation for Contextual Advertising. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14 Companion)* (pp. 345-346). New York: Association for Computing Machinery. https://doi.org/10.1145/2567948.2577361

Liu, S.-F. (2011). Author's Self-Mentions and Identity Construction in Chinese and English Abstracts. *Contemporary Rhetoric, 166,* 85-88.

Loi, C. K. (2010). Research Article Introductions in Chinese and English: A Comparative Genre-Based Study. *Journal of English for Academic Purposes, 9,* 267-279. https://doi.org/10.1016/j.jeap.2010.09.004

McGrath, L. (2016). Self-Mentions in Anthropology and History Research Articles: Variation between and within Disciplines. *Journal of English for Academic Purposes, 21,* 86-98. https://doi.org/10.1016/j.jeap.2015.11.004

Molino, A. (2010). Personal and Impersonal Authorial References: A Contrastive Study of English and Italian Linguistics Research Articles. *Journal of English for Academic Purposes, 9,* 86-101. https://doi.org/10.1016/j.jeap.2010.02.007

Mur-Dueñas, P. (2011). An Intercultural Analysis of Metadiscourse Features in Research Articles Written in English and in Spanish. *Journal of Pragmatics, 43,* 3068-3079. https://doi.org/10.1016/j.pragma.2011.05.002

Savva, M., Chang, A. X., Manning, C. D., & Hanrahan, P. (2014). TransPhoner: Automated Mnemonic Keyword Generation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)* (pp. 3725-3734). New York: Association for Computing Machinery. https://doi.org/10.1145/2556288.2556985

Scholz, M., Brenner, C., & Hinz, O. (2019). AKEGIS: Automatic Keyword Generation for Sponsored Search Advertising in Online Retailing. *Decision Support Systems, 119,* 96-106. https://doi.org/10.1016/j.dss.2019.02.001

Seoane, E., & Hundt, M. (2017). Voice Alternation and Authorial Presence: Variation across Disciplinary Areas in Academic English. *Journal of English Linguistics, 46,* 3-22. https://doi.org/10.1177/0075424217740938

Shehzad, W. (2007). Explicit Author in the Scientific Discourse: A Corpus-Based Study of the Author's Voice. *Malaysian Journal of ELT Research, 3,* 56-73.

Soler-Monreal, C. (2015). Announcing One's Work in Phd Theses in Computer Science: A Comparison of Move 3 in Literature Reviews Written in English L1, English L2 and Spanish L1. *English for Specific Purposes, 40,* 27-41. https://doi.org/10.1016/j.esp.2015.07.004

Tang, R., & John, S. (1999). The "I" in Identity: Exploring Writer Identity in Student

Academic Writing through the First Person Pronoun, *English for Specific Purposes, 18,* 23-39. https://doi.org/10.1016/S0889-4906(99)00009-5

Tankó, G. (2017). Literary Research Article Abstracts: An Analysis of Rhetorical Moves and Their Linguistic Realizations. *Journal of English for Academic Purposes, 27,* 42-55. https://doi.org/10.1016/j.jeap.2017.04.003

Thomaidou, S., & Vazirgiannis, M. (2011). Multiword Keyword Recommendation System for Online Advertising. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining* (pp. 423-427). Washington DC: IEEE Computer Society. https://doi.org/10.1109/ASONAM.2011.70

Thushara, M. G., Mownika, T., & Mangamuru, R. (2019). A Comparative Study on Different Keyword Extraction Algorithms. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC),* Erode, 27-29 March 2019, 969-973. https://doi.org/10.1109/ICCMC.2019.8819630

Tyagi, P., & Tripathi, R. C. (2019). A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data (February 8, 2019). *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE),* Sultanpur, India, 8-9 February 2019, 91-95. https://ssrn.com/abstract=3349569 https://doi.org/10.2139/ssrn.3368718

Walková, M. (2019). A Three-Dimensional Model of Personal Self-Mention in Research Papers. *English for Specific Purposes, 53,* 60-73. https://doi.org/10.1016/j.esp.2018.09.003

Wang, W., & Yang, C. (2015). Claiming Centrality as Promotion in Applied Linguistics Research Article Introductions. *Journal of English for Academic Purposes, 20,* 162-175. https://doi.org/10.1016/j.jeap.2015.05.002

Wu, G.-Q., & Zhu, Y. (2015). Self-Mention and Authorial Identity Construction in English and Chinese Research Articles: A Contrastive Study. *Linguistics & the Human Sciences, 10,* 133-158. https://doi.org/10.1558/lhs.v10i2.28557

Yeo, J. Y., & Ting, S.-H. (2014). Personal Pronouns for Student Engagement in Arts and Science Lecture Introductions. *English for Specific Purposes, 34,* 26-37. https://doi.org/10.1016/j.esp.2013.11.001

Zareva, A. (2013). Self-Mention and the Projection of Multiple Identity Roles in TESOL Graduate Student Presentations: The Influence of the Written Academic Genres. *English for Specific Purposes, 32,* 72-83. https://doi.org/10.1016/j.esp.2012.11.001

Zheng, X., & Sun, A. (2019). Collecting Event-Related Tweets from Twitter Stream. *Journal of the American Society for Information Science & Technology, 70,* 176-186. https://doi.org/10.1002/asi.24096