CIT. Journal of Computing and Information Technology, Vol. 30, No. 3, September 2022, 193–204
doi: 10.20532/cit.2022.1005566

Research on Keywords Variations in Linguistics Based on TF-IDF and N-gram

Yuyao Li[1], Xueyi Wen[2] and Xingyu Liu[3]

[1]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
[2]The Institute of Corpus Studies and Applications, Shanghai International Studies University, Shanghai, China
[3]Faculty of English Language and Culture, Guangdong University of Foreign Studies, Guangzhou, China

The rapid development of natural language processing (NLP) holds great promise for bridging the divide among languages. One of its main innovative applications is to use broad data to explore the historical trend of a subject. However, since Saussure pioneered modern linguistics, there is relatively inadequate research work done in the linguistic research on the field's variations to comprehensively reveal the linguistic trends. To trace the changes in linguistic research hotspots, we use a dataset of more than 30,000 linguistics-related literature with their titles from the Web of Science and apply NLP techniques to the data consisting of their keywords and publication years. It is found that the co-occurrence relationship between keywords, NGRAM, and their relationship with years can effectively present changes in linguistic research themes. This research is supposed to provide further insights and new methods that can be applied in the field of linguistics and related disciplines.

ACM CCS (2012) Classification: Applied computing → Arts and Humanities → Language Translation

Keywords: keyword extraction, TF-IDF, N-Gram, Linear Discriminant Analysis (LDA)

1. Introduction

ChatGPT has demonstrated the power of algorithms in processing human language. With the power of technology, linguistic research has reached a new tipping point. Linguistics is an important discipline. However, there are few systematic literature reviews of its development except a few phased studies on the development of some linguistic branches, most of which have been done based on qualitative research. With the development of linguistics and the trend of disciplinary integration, there is a wealth of accessible linguistics-related literature. Keyword extraction from literature at different stages can help linguists to sort out the weight and changes of various subfields in linguistics and to have a better overall understanding of it.

This paper uses natural language processing (NLP) techniques to extract keywords from the data of 37,890 literature titles and presents the historical development of linguistic research according to the temporal relationship of keywords. This paper is helpful for researchers to understand the development process of language research.

2. Related Work

Citespace is a commonly used tool for overview studies of specialist fields. Several scholars use Citespace to conduct research on specific areas of linguistics, for example, lexicography [1], language learning [2], and discourse studies [3]. However, there is little research on the changes in keywords in the field of linguistics over time. Some research on the development of linguistic subfields is based on technical resources, such as corpora. However, most research in this field is based on manual collection and summarized roughly by different stages. For example,

S. Hua *et al.* [4] have outlined three stages of linguistics development based on existing literature: linguistics studies before scientific methods were introduced, historical linguistics, and modern linguistics. The current research still lacks a comprehensive review of the subject. Malkiel in 1953 [5] provided an overview of the history of language evolution as well as historical linguistics. Jones (1999) [6] presented the development of different linguistic civilizations and linguistic features at the sociocultural level. George van Driem [7] and others critically discussed the misrepresentation of the history of linguistics by LaPolla *et al.* [8] and history, such as the Tibetan-Burmese language family, is presented. As it can be seen, the majority of the studies are based on manual analysis and are overviews of a particular branch of linguistics (*e.g.*, historical linguistics), or the history of linguistics at a certain time. The overarching analysis of linguistic research hotspots is not sufficiently up-to-date and generalized, and the study of the historical changes in keywords through big data analysis is still relatively rare.

This research aims to fill the gap in the analysis of how keywords change in the field of linguistics. We use a crawler to collect and extract relevant information from linguistics papers in major academic databases, such as authors, titles, abstracts, and references. Then we extracted the keywords in literature from 1800 to the present according to different periods to find out the tipping points of linguistics and the changes of emphasis to know how linguistics develops, from a thorough and complete perspective. We also extracted keywords from the linguistics literature from 1884 to 2022 and modeled their themes to verify the accuracy of the keywords extraction and to analyze the main themes involved in linguistics.

3. Description of the Data

The data was obtained from the Web of Science, a comprehensive database of scholarly information resources covering the largest number of disciplines in the world. We downloaded the title, author, year of publication, and abstract of 37,890 results using the keyword "语言学/linguistics" to search. Since the number of papers from 1884–2022 is not balanced across different periods, we decided to classify the linguistic papers according to the following periods: 1884 to 2007 with 8,556 articles; 2008 to 2015 with 13,191 articles; 2016 to 2022 with 16,143 articles.

We combined the titles and abstracts of the 37,890 articles and pre-processed the texts using the nltk package, including sentence segmentation, removal of punctuation, tokenization, removal of stop words, and stemming/lemmatization to obtain independent topic words and avoid repetitive counting of the same words.

4. Research Methodology

We used the TF-IDF algorithm, N-gram extraction, and LDA to investigate the disciplinary hotspot changes.

TF-IDF is a statistical method for assessing the importance of a word within a corpus. It is often used in search engine applications and keyword lookup. The main idea is the balance between the frequency of a word and the number of documents containing it, *i.e.*, a word or phrase is of high importance if it occurs with high TF frequency in one article and low TF frequency in other articles. The IDF is proposed because if the number of documents containing a word is low, it means that the word is more distinguishable.

TF-IDF is a commonly used supervised method for keyword extraction. Its calculation is mainly based on the following formula. Suppose d is the keyword extraction text, t is the candidate word and D is the corpus, then the TF-IDF feature value W_{t-D} is calculated by the following equation:

$$W_{t-D} = TF \cdot IDF = \frac{f_{t,d}}{|d|} \cdot \log \frac{|D|}{f_{t,d}} \qquad (1)$$

where TF represents the term frequency of the candidate word t, $f_{t,d}$ represents the number of occurrences of candidate word t in text d, and $|D|$ represents the number of texts in the corpus. According to the TF-IDF algorithm, the higher the frequency of a candidate word in a single document and the lower the frequency in the overall corpus, the higher the weight it has.

TF-IDF algorithm is simple to calculate, widely used, and suitable for various kinds of corpus building [9].

However, the TF-IDF algorithm yields results as independent tokens with separate information and neglects contextual information. This is well compensated by the *N*-Gram[10], which extracts a consecutive *N*-number of words from a sentence, allowing for a richer set of features and incorporating word order information. As *N* increases, the dimension of the word becomes higher, creating the sparsity problem. This study retrieves 2-Gram words for analysis in this research.

In addition, we can calculate the number of occurrences of co-occurring words in different documents by TF-IDF. However, this method neglects semantic association. Concerning this, we also used the LDA topic modeling as a reference.

LDA topic modeling is based on the relative distribution of document-topic, topic-word, and document-word to represent the abstract concepts using a series of topic words [11]. The higher the probability of occurrence of a word, the higher its relevance to the topic. The probability *p* of occurrence of a word is expressed by Equation 2.

$$p(word|document) = \frac{}{\sum_{theme} p(word|document) \cdot p(word|document)} \quad (2)$$

Therefore, we combined the titles and abstracts of 37,890 items. After text pre-processing, we use them as the input to LDA. Then, LDA outputs the top 7 high-frequency topic words for 10 topics. The results are used to make a comparison with the results of topic extraction with TF-IDF statistics to verify the TF-IDF validity.

The result of English word separation, in general, are independent words, which assume that words are independent of each other and do not consider their order. In contrast, *N*-gram (*N*-gram grammar) extracts a set of *N* consecutive words from a sentence and obtains the information before and after the words. It is more common to use 2-gram or 3-gram, and in this paper, we also use 2-gram and 3-gram for the analysis. Using *N*-gram for the analysis can obtain richer features and combine the word order information, but with the increase of *N*, the word list dimension will become higher and produce the data sparsity problem.

5. Research Findings

Using sklearn's feature_extraction tools, CountVectorizer and TfidfTransformer, we vectorized the corpus and then calculated its TF-IDF and N-Gram values to sort out the changing history of linguistic research hotspots.

5.1. Change in Research Hotspots from 1884 to 2007

Figure 1 shows the main keywords used between 1884 and 2007, including "language", "speech", "model", "grammar" "sentence", and so on. From this, we can see that during this period, researchers focused on theoretical linguistics, language models, syntax, and lexicography. Apart from that, the keywords also include "brain", "adult", "children", "medic(al)", "problem", "aphasia", and "comprehend", which shows that the during this period the researchers also focused the contrast between adult and child language, such as child language acquisition, as well as the connections between linguistics and brain neuroscience, such as psycholinguistics, and aphasia. Another rather common keyword is "English", which remains the main object of linguistic research from 1884 to 2007., showing the researched language dominated by "English".

From the perspective of single keyword variations, the words "speech" and "linguistic" have remained to be high-frequency words from 1884 to 2007, showing that language and speech has been the focal point of linguistic studies. However, the word "model" only appears as a high-frequency word between 1884 and 1949, while "children" has been among the top 3 keywords since 1950. This indicates that, after 1950, the focus of research shifted from language models to children's language acquisition and language disorders. Thus, the frequency of keywords for language disorders (*e.g.*, "aphasia") and medical treatment (*e.g.*, "medic(al)") have become higher.

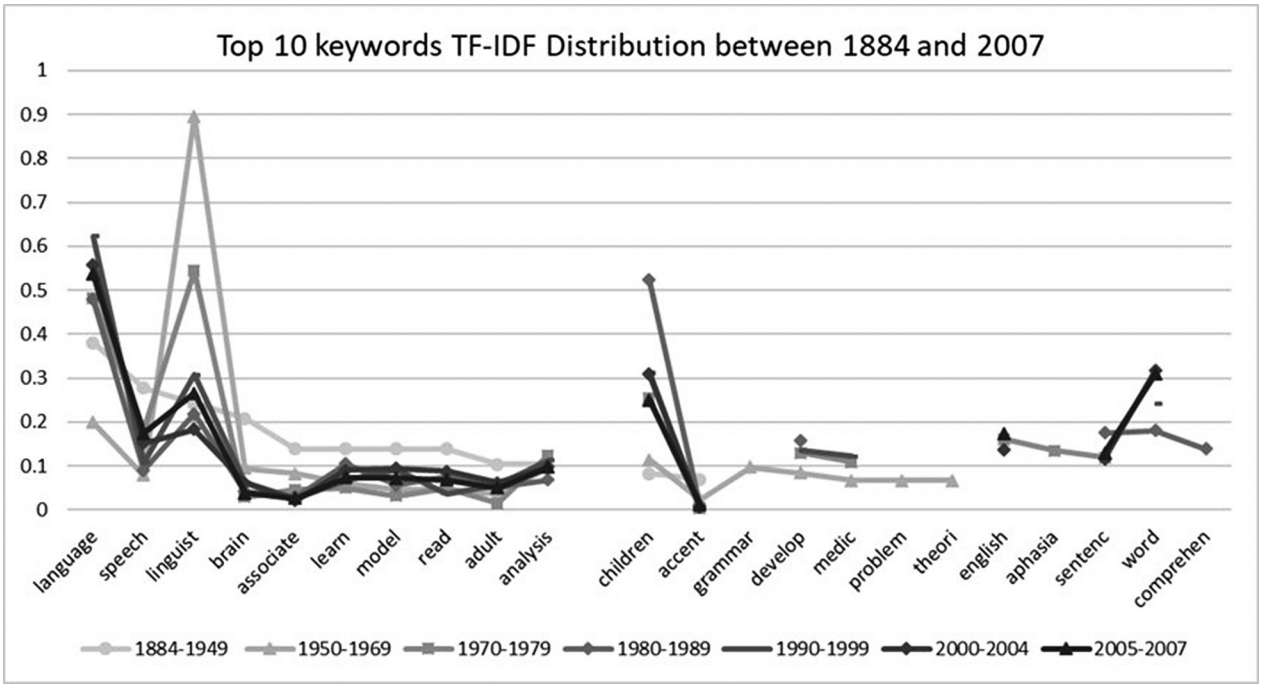

Figure 1. TF-IDF distribution of the top 10 keywords between 1884 and 2007.

These TF-IDF statistic results can be supplemented by N-Gram tokenization results over the period 1884–2007, as shown in Table 1. The results of 2-gram tokenization are largely in line with the previous analysis of TF-IDF topic words. From 1884 to 1949, the focus was mainly on "brain connectivity", "language model", and so on. After 1950, the focus began to be centered on "age group" and "children", with child-related keywords such as "deaf child", "impaired child", "normal child", "language disorder child", "autistic child", "child language", and "develop child". It shows that the study of language disorders in children and child language acquisition has become increasingly frequent after 1950. Thus, the period 1950–2007 may be a period when linguists focused on language disorders in children starting with age and studied language disorders, which further supports the previous analysis of the subject words.

Another category of keywords related to "language impairment" is "brain damage", "aphasic patient", "Broca's area" and "language impair". Broca's area is the motor center of the brain that controls language. Lesions in Broca's area can cause motor aphasia or expressive aphasia, which means that the patients have difficulty in articulation. From the keywords which indicate the disease's progress, we can see that a breakthrough was made between 1990 and 2007 regarding the relationship between language and the brain.

In addition, the *N*-Gram segmentation has extended more information based on the word "English", such as "black English", "English language" and "Korean language". The focus of research in this period was on different branches of English and the analysis of the Korean language.

5.2. Change in Research Hotspots from 2008 to 2015

From 2008 to 2015, the keyword curves overlap a lot, which means that keyword trends are largely consistent over time. Compared to the period 1884–2007, the TF-IDF values for "language" and "linguist(ic)" show an increase and then remain stable at around 0.6. The words "Korean", "cognit", "speech", "analysis", "study", "research" and "children", which are topic words from 1884–2007, also appear in the top ten keywords from 2008–2015. It shows that in this period, researchers focused more on cognitive linguistics and Korean linguistics as well as shifting their focus to experimentation and analysis.

Table 1. TF-IDF distribution of the top 10 keywords in the *N*-Gram between 1884 and 2007.

Top 10 Keywords by Years								
	1884-1949	**TF-IDF value**	**1950-1969**	**TF-IDF value**	**1970-1979**	**TF-IDF value**	**1980-1989**	**TF-IDF value**
1	language	0.38	linguist	0.90	linguist	0.54	children	0.52
2	speech	0.28	language	0.20	language	0.48	language	0.48
3	linguist	0.24	children	0.11	children	0.25	linguist	0.22
4	brain	0.21	analysis	0.10	speech	0.18	word	0.18
5	associate	0.14	grammar	0.10	English	0.16	sentence	0.18
6	learn	0.14	develop	0.08	aphasia	0.13	develop	0.16
7	model	0.14	speech	0.08	develop	0.13	comprehend	0.14
8	read	0.14	medic	0.07	analysis	0.12	acquisition	0.13
9	adult	0.10	problem	0.07	sentence	0.12	child	0.10
10	analysis	0.10	theory	0.07	medic	0.11	learn	0.10
Total Papers	311		476		965		1304	
	1990-1999	**TF-IDF value**	**2000-2004**	**TF-IDF value**	**2005-2007**	**TF-IDF value**		
1	language	0.62	language	0.56	language	0.54		
2	children	0.31	word	0.32	word	0.31		
3	linguist	0.31	children	0.31	linguist	0.27		
4	word	0.24	linguist	0.18	children	0.25		
5	develop	0.14	expert	0.16	Korean	0.18		
6	medic	0.12	speech	0.15	English	0.17		
7	analysis	0.11	English	0.14	speech	0.17		
8	speech	0.11	process	0.14	process	0.15		
9	process	0.10	effect	0.12	sentence	0.13		
10	group	0.09	sentence	0.12	active	0.12		
Total Papers	1586		2585		2785			

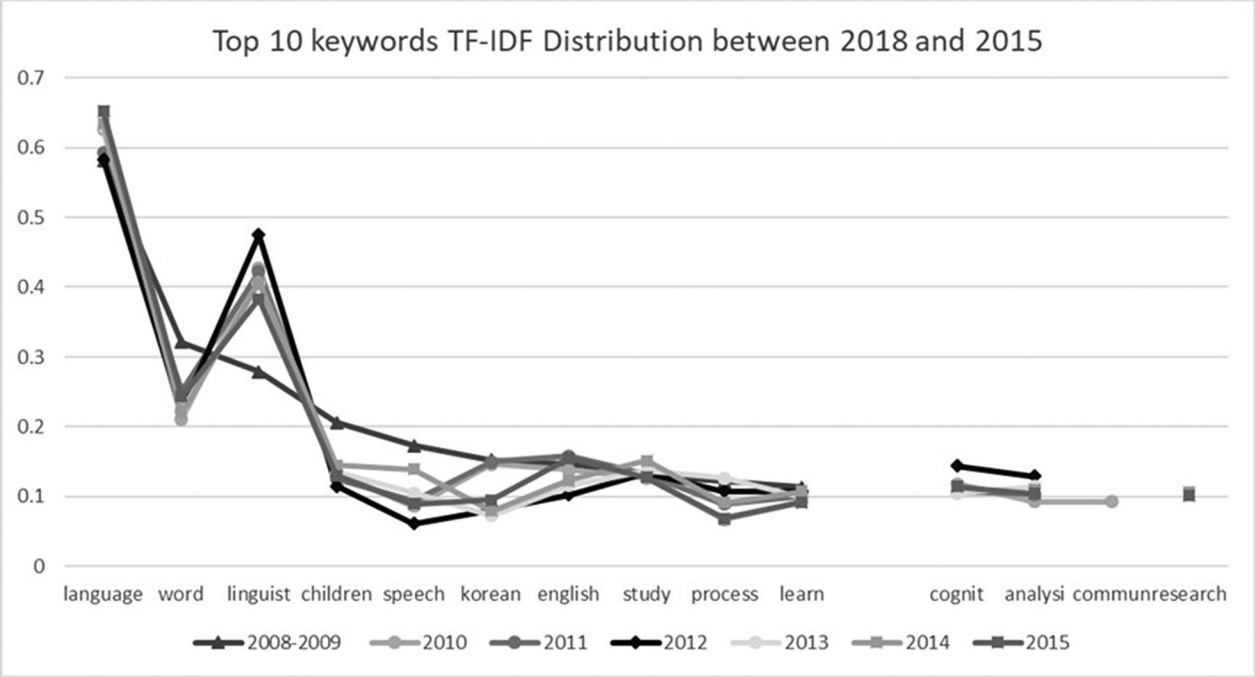

Figure 2. TF-IDF distribution of top 10 keywords between 2008 and 2015.

The *N*-Gram tokenization in Table 2 reveals several terms derived from the keyword "linguistics" such as "foreign language", "(K) Korean linguistics", "(C) Chinese character", and so on. It can support the fact that the focus of research in this period began to shift from English linguistics to other languages such as Korean and Chinese. Various branches of linguistics are also identified during this period, such as the terms "frontal gyrus" and "inferior frontal (gyrus)". In the study of the brain and language between 1884 and 1949, more in-depth integration of language, biology, and neurology studies can be found. In cognitive linguistics, the keywords "conceptual metaphor" and "cognitive linguistics" appeared in the top 10 TF-IDF keyword lists. It shows that the brain's perception of the world and the relationship between metaphor and cognition were also major topics in linguistic studies during this period. The TF-IDF score of another major branch of linguistics term, "applied linguistics", also increased in frequency during this period. This indicates a gradual shift of emphasis from theoretical to applied linguistics between 2008 and 2015.

In addition, new keywords also appear in this period compared to 1884–2007, such as "sign language", "bilingual child", "control group", "target word", *etc*. During this period, the study of children's language has shifted from the study of language disorders to the study of bilingual children, with the increasing importance of experimentation in linguistic research.

5.3. Keyword Change from 2016 to 2022

As shown in Figure 3, the trends for most of the keywords also overlap with each other between 2016 and 2022. The focus of the research remained on "children", "cognitive" and "apply". At the same time, the research on topics including "English", "Korean" and "children" remains mainstream. For some words that do not overlap with each other, we can indicate new research topics during this period. For instance, in 2019, the word "corpus" appears in the top ten TF-IDF keywords, which shows that corpus linguistics is a hot topic in 2019. In 2022, the word "sentence" returns to the keyword queue once again and it can be assumed that the study of syntax will be a hot topic in 2022.

Table 2. TF-IDF distribution of the top 10 keywords in the *N*-Gram for the period 2008-2015.

Top 10 Keywords by Years								
	2008-2009	**TF-IDF value**	**2010**	**TF-IDF value**	**2011**	**TF-IDF value**	**2012**	**TF-IDF value**
1	language	0.58	language	0.63	language	0.59	language	0.58
2	word	0.32	linguist	0.43	linguist	0.42	linguist	0.47
3	linguist	0.28	word	0.21	word	0.25	word	0.22
4	children	0.21	Korean	0.15	English	0.16	cognit	0.14
5	speech	0.17	English	0.14	Korean	0.15	study	0.13
6	Korean	0.15	children	0.13	study	0.13	analysis	0.13
7	English	0.15	study	0.13	children	0.13	children	0.12
8	study	0.13	cognit	0.12	cognit	0.11	learn	0.11
9	process	0.12	analysis	0.09	analysis	0.10	process	0.11
10	learn	0.11	community	0.09	learn	0.10	English	0.10
Total Papers	2259		2179		2286		2207	
	2013	**TF-IDF value**	**2014**	**TF-IDF value**	**2015**	**TF-IDF value**		
1	language	0.63	language	0.63	language	0.65		
2	linguist	0.41	linguist	0.41	linguist	0.38		
3	word	0.22	word	0.22	word	0.24		
4	study	0.14	study	0.15	english	0.15		
5	children	0.13	children	0.14	children	0.13		
6	process	0.13	speech	0.14	study	0.13		
7	english	0.12	english	0.12	cognit	0.11		
8	analysi	0.11	analysi	0.11	analysi	0.10		
9	cognit	0.11	learn	0.11	research	0.10		
10	speech	0.11	research	0.11	korean	0.10		
Total Papers	2349		2439		2542			

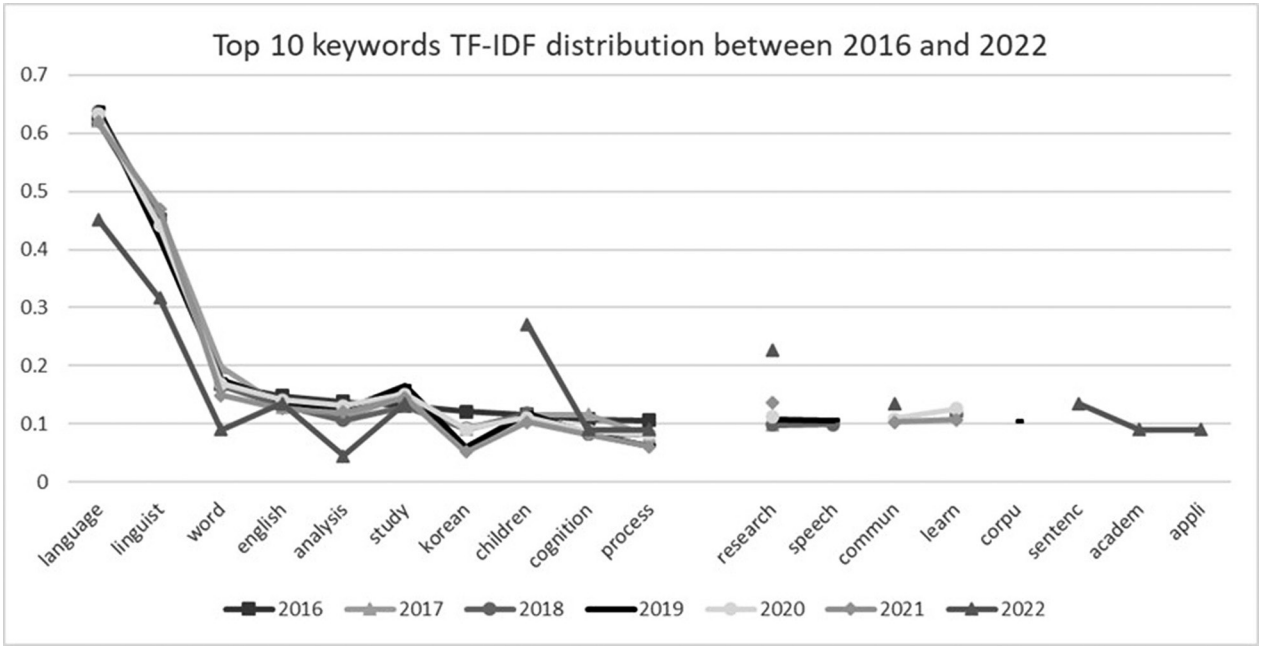

Figure 3. TF-IDF distribution of top 10 keywords during 2016-2022.

The TF-IDF analysis with N-Gram segmentation gives the corresponding data, as shown in Table 3. In the period 2016–2022, the keywords remain to be "linguistics", with the ranking of "apply (applied) linguistics" being more prominent than in the previous two time periods and is spread over almost all the time between 2016 and 2022. "Cognitive linguistics", "conceptual metaphor" and "sign language" remained keywords in linguistics. At the same time, China-related keywords, such as "Chinese character" and "Hong Kong" emerged during this period. In addition, in foreign linguistics, "English language", "Korean language", "foreign language" and "American English" are also keywords that can be seen as trends in the study of foreign linguistics, such as American English, English, and Korean. This can show some rising linguistic studies of applied linguistics and different languages in different countries, such as China and Korea.

6. Topic Extraction

To compare and validate the hotspot extraction results, this experiment generated a corpus of processed title and abstract data, weighted by TF-IDF, and later passed through the LDA topic extraction model. A total of 10 topics were extracted in this design and the seven words with the highest frequency were selected for each topic. The theme extraction results are shown in Figure 4.

As shown in Table 4, the keywords in the first topic include "verb", "sentence", "syntactic", "semantics", and "grammar", which are related branches of theoretical linguistics. Thus, we can conclude that the first topic is presumably theoretical linguistics. The keywords in the second topic include "children", "language(e)", and "test", indicating that the second topic is presumably related to children's language acquisition and language testing. The keywords in the third topic include "noun", "pronoun" and "refer", leading to the fact that the third topic is related to the accusative pronoun. The keywords in the fourth topic include "word", "experiment", "lexical" and "read", signaling the fourth topic to be lexicography. The keywords for the fifth topic include "language", "brain", "function" and "neural", indicating the fifth topic to be presumably neurolinguistics, *i.e.* the study of language and the brain. The keywords in the sixth topic include "language" and "culture ", so we can assume the sixth topic to be language and culture. Similarly, the seventh topic contains the keywords "linguist" and "analysis", so it could be linguistic research and textual analysis. The eighth topic contains the keywords "student", "education", "study" and "learner", leading it to be language education and learning. The key-

Table 3. TF-IDF distribution of top 10 keywords for *N*-Gram between 2016 and 2022.

Top 10 Keywords by Years								
	2016.00	**TF-IDF value**	**2017**	**TF-IDF value**	**2018**	**TF-IDF value**	**2019**	**TF-IDF value**
1	language	0.62	language	0.62	language	0.64	language	0.64
2	linguist	0.45	linguist	0.46	linguist	0.45	linguist	0.42
3	word	0.17	word	0.20	word	0.17	word	0.18
4	English	0.15	study	0.15	English	0.13	study	0.16
5	analysis	0.14	English	0.13	study	0.13	English	0.14
6	study	0.13	learn	0.12	children	0.12	analysis	0.12
7	Korean	0.12	cognit	0.12	learn	0.12	children	0.12
8	children	0.12	children	0.12	analysis	0.11	research	0.11
9	cognit	0.11	analysis	0.11	speech	0.10	speech	0.11
10	process	0.11	research	0.10	research	0.10	corpus	0.11
Total Papers	2872		2942		3088		3087	
	2020	**TF-IDF value**	**2021**	**TF-IDF value**	**2022**	**TF-IDF value**		
1	language	0.63	language	0.62	language	0.45		
2	linguist	0.44	linguist	0.47	linguist	0.32		
3	word	0.17	word	0.15	children	0.27		
4	study	0.15	study	0.14	research	0.23		
5	English	0.14	research	0.14	communicate	0.14		
6	analysis	0.13	English	0.13	English	0.14		
7	learn	0.13	analysis	0.12	sentence	0.14		
8	research	0.11	learn	0.11	study	0.14		
9	children	0.11	children	0.10	academic	0.09		
10	commun	0.11	commun	0.10	appli	0.09		
Total Papers	3191		3096		183			

words in the ninth topic include "speech", "measure" and "health", which leads to the assumption that the ninth topic is language testing and language disorders. The keywords for the tenth topic include "English", "speaker", "bilingual" and "native", so the tenth topic is presumed to be bilingual language use.

Overall, the above ten topics are largely consistent with the results of the TF-IDF analysis, covering syntax, semantics, lexicology, language testing, child language acquisition, neurolinguistics, language and culture, language education, and the study of bilingual acquisition and foreign linguistics.

Table 4. LDA theme extraction results.

topic	value-word	value-word	value-word	value-word	value-word	value-word	value-word
1	0.023* verb	0.022* sentence	0.019* syntactic	0.018* structure	0.017* grammatic	0.017* semantic	0.015* grammar
2	0.041* children	0.034* language	0.020* age	0.019* develop	0.019* group	0.013* study	0.012* test
3	0.072* noun	0.052* gender	0.043* name	0.035* person	0.029* object	0.025* pronoun	0.025* refer
4	0.060* word	0.021* effect	0.020* process	0.019* experiment	0.014* task	0.013* lexical	0.011* read
5	0.054* language	0.021* process	0.017* active	0.016* network	0.016* brain	0.015* function	0.014* neural
6	0.038* language	0.021* culture	0.020* linguist	0.020* translate	0.019* korean	0.019* study	0.017* chinese
7	0.027* linguist	0.012* study	0.012* research	0.011* analysis	0.010* text	0.009* language	0.009* base
8	0.047* language	0.037* learn	0.032* student	0.029* english	0.021* education	0.018* study	0.018* learner
9	0.035* speech	0.017* patient	0.013* emotion	0.010* result	0.009* method	0.009* measure	0.009* health
10	0.032* english	0.029* speaker	0.024* language	0.024* bilingual	0.023* speech	0.019* native	0.015* vowel

7. Discussion and Conclusions

In general, the TF-IDF with *N*-Gram outputs keywords on the changing hotspots of linguistic research, which are broadly in line with the general trends in linguistics but provide a more specific variation of the topics. The method applied in this study is different from the qualitative study by S. Hua *et al.* [4]. Instead of focusing on some generalized trending or single linguistic domain, this study aims to demonstrate the overall shift of more than 10 words for each examined time period and dig into the linguistic hotspots studies and reasons behind them. In terms of the methods, compared with the generalization of previous literature and qualitative studies focusing on specific areas of vocabulary or discourse analysis, the disciplinary analysis of the field of linguistics using big data and computer technology can show the whole picture of linguistic hotspots and development more objectively, systematically, comprehensively, and meticulously. This study also demonstrates an innovative exploration of the cross-field methods in the field of computing and linguistics.

In detail, from 1884 to 1949, linguistic studies focus on historical-comparative linguistics and structuralist linguistics. According to the TF-IDF and N-gram calculations, we can explore the detailed variation behind the broad trending: the hotspots of research from 1884 to 1949

were language models and theoretical linguistic concerns of grammar and vocabulary, in line with historical-comparative linguistics. This indicates emphasis on language, grammatical and lexical correspondences, and comparative studies. At the beginning of the 20th century, the hotspots variations show the birth of structuralist linguistics, which has a separate set of relational structures, such as Saussure's belief that linguistic behavior is social and individual.

Between 1950–2015, the linguistic study enters its third period, focusing on the integrated study of phonology, syntax, semantics, and pragmatics, viewing language as a complex information system. Experimental linguistics divides language structure into levels of words and phrases, syllables, and phonemes using mathematical symbol patterns and formal deduction methods. As we can see, the TF-IDF and the N-Gram model again provide detailed variation and evidence, calculating the hotspots of research between 1950 and 1969. During that period grammar, phonology, theories of language acquisition, language disorders, and the study of language and culture were in focus. This is also in line with the change in 1957. In 1957, an american linguist Chomsky introduced transformational generative grammar, which placed syntactic relations at the center of language structure and assumed that people had a mechanism for language acquisition. During this period, language became widely integrated with mathematics, sociology, philosophy, psychology, and computer science.

TF-IDF and *N*-Gram models generated that "language and society" is the research hotspot between 2015–2022. This is consistent with the fourth stage of linguistics, the systemic functional linguistics. The hotspots generated show that during this period, scholars start to consider language as a social symbol system and emphasizes the basic functions of language communication and exchange, as well as the relationship between language and society and specific language combinations.

Thus, the TF-IDF model as well as the *N*-gram algorithm can greatly help us to sort out and explore the changes in linguistic research hotspots between 1884 and 2022, with details and specific changes of each hotspot with the help of a large database. The prospects of the research are to further increase the amount of data extracted, which can be combined with multiple databases or applied to other areas of hotspot variation research, as well as use multiple keyword extraction techniques, semantic-based machine learning, and deep learning models to analyze and compare. More importantly, the research expects to further visualize the trending changes of each topic from 1884–2022 and explore the reasons behind the change in research hotspots.

References

[1] L. Liu *et al.*,"The Research on Antonymous Compound Word in Modern Chinese Language from 1978–2021: Visualization Maps and The Analysis Based on CiteSpace", *The Journal of Chinese Language and Literature*, vol. 133, pp. 171–200, 2022. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002836572

[2] N. Sinhye, "Comparison of Korean Vocabulary Usage Patterns Depending on the Mother Tongue of Korean Language Learners: Using TF-IDF Scores", *Language & Information Society*, vol. 46, pp. 1–26, 2022. https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002864755

[3] G. Wang *et al.*, "A Bibliometric Study of News Discourse Analysis (1988–2020)", *Discourse & Communication*, vol. 16, pp. 110–128, 2021. https://doi.org/10.1177/17504813211043725

[4] S. Hua *et al.*, "Talking About the Stages of Linguistic Development", *Eurasian Humanities Studies (Chinese and Russian)*, (04):61–63+87+91, 2021. (In Chinese)

[5] Y. Malkiel, "Language History and Historical Linguistics", *Romance Philology*, vol. 7, no. 1, pp. 65–76, 1953. http://www.jstor.org/stable/44938289

[6] G. Jones, "Strange Talk: The Politics of Dialect Literature in Gilded Age America", Univ of California Press, 1999.

[7] G. V. Driem, "Linguistic History and Historical Linguistics", *Linguistics of the Tibeto-Burman Area*, vol. 41, pp. 106–127, 2018. https://doi.org/10.1075/ltba.18005.dri

[8] R. J. LaPolla, "Once Again on Methodology and Argumentation in Linguistics: Problems with the Arguments for Recasting Sino-Tibetan as 'Trans-Himalayan'", *Linguistics of the Tibeto-Burman Area*, vol. 39, no. 2, pp. 282–297, 2016. https://doi.org/10.1075/ltba.39.2.03lap

[9] Z. Ting and S. S. Ge, "An Improved TF-IDF Algorithm Based on Class Discriminative Strength for Text Categorization on Desensitized Data", in *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, 2019, pp. 39–44. https://doi.org/10.1145/3319921.3319924

[10] S. S. M. M. Rahman *et al.*, "An Investigation and Evaluation of *N*-Gram, TF-IDF and Ensemble Methods in Sentiment Classification", in *Proc. of the Cyber Security and Computer Science: 2nd EAI International Conference, ICONCS 2020, Dhaka, Bangladesh*, 2020, pp. 391–402. https://doi.org/10.1007/978-3-030-52856-0_31

[11] M. Steyvers *et al.*, "Probabilistic Author-topic Models for Information Discovery", in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 306–315, 2004. https://doi.org/10.1145/1014052.1014087

Received: May 2023
Revised: July 2023
Accepted: August 2023

Contact addresses:
Yuyao Li*
School of Information Science and Technology
Guangdong University of Foreign Studies
Guangzhou
China
e-mail: everybit@163.com
*Corresponding author

Xueyi Wen
The Institute of Corpus Studies and Applications
Shanghai International Studies University
Shanghai
China
e-mail: wenxueyi@shisu.edu.cn

Xingyu Liu
Faculty of English Language and Culture
Guangdong University of Foreign Studies
Guangzhou
China
e-mail: 20210100160@gdufs.edu.cn

Yuyao Li is a lecturer from Guangdong University of Foreign Studies, and he is the head of the Department of Information Technology Teaching in the School of Information Science and Technology (School of Cyberspace Security) of the university. He graduated from South China Normal University, China. His research interests focus on education and humanities data mining.

Xueyi Wen received BA degree in English Language and Culture (Linguistics) from Guangdong University of Foreign Studies, China, in 2022 and is currently a postgraduate student major in Language Data Science and Applications in the Institute of Corpus Research, Shanghai International Studies University. Her research interests are computational linguistics/natural language processing.

Xingyu Liu is a student of English and software engineering at Guangdong University of Foreign Studies, Guangzhou, China. His research interest is natural language processing.