

Four decades of research on multiword expressions: A bibliometric analysis

Wei Yi* , Yihan Zhao, Wanxin Wang

School of Chinese as a Second Language, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing 100871, China

*Corresponding author. School of Chinese as a Second Language, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing 100871, China. E-mail: weiyisla@pku.edu.cn

Multiword expressions have been a prominent research topic in applied linguistics and second language acquisition. However, no bibliometric analysis has been conducted to examine the landscape of this field. Using CiteSpace, we analyzed 1,302 research articles published from 1983 to 2024, retrieved from the Web of Science database. We performed various bibliometric analyses based on citation, co-citation information, and keywords from the publications to identify the most influential authors, scholarly documents, and journals in the field. We also visualized collaboration networks between authors, research institutions, and countries. Additionally, we identified the key research areas and the most frequently discussed topics over different time periods. By tracking and visualizing the patterns in research on multiword expressions over the past four decades, our study offers insights into the intellectual structure and dynamics of this field, thereby guiding future research on larger-than-word units.

Introduction

Language is represented by linguistic units of varying grain sizes, including single words and word combinations. For decades, Chomsky's generative grammar dominated the field of linguistics, establishing a tradition that viewed language as consisting of single words and grammatical rules. Following the words-and-rules approach, multiword units were considered peripheral because they were believed to be generated by combining words based on syntactic manipulations. However, applied linguists such as [Pawley and Syder \(1983\)](#) recognized that multiword units (referred to as 'complex lexical items'), particularly idioms, do not fit into Chomsky's generative grammar, as these larger-than-word units are not subject to syntactic analysis. Additionally, multiword units are frequently used by language users and play an important role in achieving nativelikeness.

Early studies of multiword units paved the way for increasing interest in larger-than-word units, which have been given many different names, such as lexical phrases ([Nattinger and DeCarrico 1992](#)), formulaic sequences ([Wray and Perkins 2000](#); [Wray 2002](#)), multiword sequences (e.g. [Yi et al. 2017, 2023](#); [Yi and Zhong 2024](#)), multiword phrases ([Arnon and Snider 2010](#)), and multiword expressions ([Siyanova-Chanturia and Martinez 2015](#)). To be theoretically neutral and to maintain consistency, we will refer to recurrent word combinations as multiword expressions

throughout this article. Multiword expressions are heterogeneous in terms of structural, syntactic, and semantic characteristics. They include larger-than-word units such as proverbs (e.g. 'actions speak louder than words'), idioms (e.g. 'kick the bucket'), phrasal verbs (e.g. 'take off'), speech formulae (e.g. 'what's up?'), binomials (e.g. 'bride and groom'), collocations (e.g. 'make progress'), and lexical bundles (e.g. 'is one of the'). Usage-based theories (e.g. Tomasello 2000) suggest that multiword expressions are integral building blocks of language. They are crucial for first language development (Wray 2002) and contribute to the development of second language skills and nativelikeness. From a cognitive perspective, multiword expressions are processed significantly faster than free-word combinations (Yi and Zhong 2024). Therefore, they are said to save language processing effort, reduce time pressure, and circumvent the limitations of cognitive capacity (Christiansen and Chater 2016).

Over the past decades, numerous empirical studies have explored various aspects of multiword expressions, resulting in divergent findings. Several systematic reviews have attempted to synthesize research methodologies, findings, and pedagogical practices related to multiword expressions (e.g. Weinert 1995; Wray 2000, 2012; Wood 2002; Shaoul and Westbury 2011; Siyanova-Chanturia 2013; Siyanova-Chanturia and Martinez 2015; Boers 2020). However, despite these efforts, the field remains fragmented, with studies differing in focus, theoretical approaches, and methodologies. This diversity makes it difficult to capture the overall landscape of research on multiword expressions without employing techniques such as bibliometric analysis. Bibliometric analysis is a quantitative method for evaluating the impact and trends of scientific literature by analyzing publication patterns, citation counts, and authorship metrics. In the field of applied linguistics and second language acquisition, bibliometric analysis has only recently begun capturing the attention of researchers, with a handful of articles being published in recent years (e.g. Lei and Liu 2019; Zhang 2020). Compared with narrative reviews and meta-analyses, bibliometric analysis has several advantages worth mentioning. First, unlike narrative reviews and meta-analyses, bibliometric analysis can handle big data and process large amounts of literature. Second, using citation and co-citation analysis, bibliometric analysis can help identify influential works, collaboration networks, key research areas, frequently explored topics, and emerging trends. Third, bibliometric analysis can visualize patterns and trends in the research landscape, which allows for a clearer and deeper understanding of how a field has evolved over time. Fourth, bibliometric analysis is largely data-driven. Therefore, results obtained from this technique can maintain objectivity and reduce bias to a maximum degree.

As far as we know, no bibliometric analysis has been conducted to illustrate the landscape and dynamics of research on multiword expressions. To fill this gap, this study presents a comprehensive bibliometric review using CiteSpace as the primary analytic tool (Chen 2006). CiteSpace is a Java-based software developed for analyzing and visualizing research themes and collaborative networks as well as emerging trends in scientific literature. In this study, we aim to gain a comprehensive overview of research on multiword expressions by addressing the following questions:

1. How has the number of publications changed over time?
2. Which scholars, references, and sources of publications are most influential?
3. Which scholars, institutions, and countries have collaborated with each other?
4. What are the key areas of research on multiword expressions?
5. What are the most frequently explored topics and emerging trends?

Methodology

Data collection

To conduct the bibliometric analysis, we performed a topic search on the Web of Science Core Collection database using the following search query, which included various terms commonly used in the field to refer to multiword expressions: TS = ('binomial*' OR 'collocation*' OR 'idiom*' OR 'N-gram*' OR 'congram' OR 'chunk*' OR 'phraseology' OR 'phrasal verb*' OR 'frozen phrase*'

OR 'formulaic speech' OR 'formulaic language' OR 'formulaic sequences' OR 'prefabricated language' OR 'prefabricated patterns' OR 'word combination*' OR 'word sequence*' OR 'lexical bundles' OR 'multiword construction*' OR 'multiword expression*' OR 'multiword unit*' OR 'multiword phrase*' OR 'multiword sequence*'). We then refined our search according to the following procedure. First, we limited the document type to include only articles, early-access publications, and proceeding papers. Second, we focused on papers categorized under 'Linguistics' in the research area. Finally, we specified that the language of the publications should be English. Our search returned 4,277 publications. The first and third authors of this study then screened all the publications to ensure that the articles were indeed linguistic research published in English and that their focus of study was multiword expressions. Based on this procedure, 1,302 records were selected and downloaded for the bibliometric analysis, with publication years ranging from 1983 to 2024.

Data analysis

We used the CiteSpace software (version 6.2.R4 Advanced) to analyze the bibliometric data. Specifically, we employed the following techniques. First, to examine the overall developmental pattern of research on multiword expressions over the past four decades, we plotted the number of publications per year against the publication year. Second, to identify high-impact authors, papers, and journals, we analyzed their citation counts and conducted burst analysis to determine the active periods of highly cited authors and papers. Third, to investigate the distribution of research papers as well as the collaboration patterns, we performed a collaboration network analysis on authors, institutions, and countries. Fourth, to explore key research themes in the field of multiword expressions, we conducted co-citation analysis of references. Finally, to identify hot topics and emerging trends in the field, we employed co-word analysis by extracting terms from titles, abstracts, and keywords of the publications, and combined this with burst analysis to examine the active periods of different research topics.

Before running the bibliometric analysis, we configured the CiteSpace software to include publications from 1983 to 2024, with the dataset segmented into annual intervals (i.e. 1 year per slice). To filter out noise from the dataset and improve the quality of science mapping, we used the *g*-index, which emphasizes highly cited works. We set $k = 25$ for the document co-citation analysis, meaning the top 25 most-cited articles for each scholar were considered. We set $k = 15$ for the co-word analysis, such that the top 15 most-relevant terms for each reference were used for detecting significant themes in research on multiword expressions. Additionally, we set the selection criteria to 'Top 50' for the collaboration network analyses, ensuring that the top 50 most productive authors, countries, and institutions per slice were the focus of our analysis. To enhance the clarity and interpretability of the networks generated by CiteSpace, we applied the pathfinder algorithm and pruned the sliced networks.

Results and discussion

Growth in publications over time

Figure 1 illustrates the publication trend in the field of multiword expressions over the past four decades. The line graph reveals two distinct stages. The first stage, spanning from 1983 to 2007, is characterized by an average of fewer than 10 publications per year ($M = 9.6$, $SD = 5.8$), with a cumulative total of less than 250. During this period, research on multiword expressions was scattered, resulting in low and fluctuating yearly publication counts and a slow accumulation of total publications. This phase aligns with Kuhn's pre-science and paradigm formation phase (Kuhn 1962), where research was sporadic and various theoretical models were proposed without a dominant paradigm. Since 2008, there has been a significant increase in both yearly publication counts and the cumulative total of publications. This period reflects a more mature and expanding research field with established paradigms, diverse methodologies, and a substantial rise in research output and highly cited publications, corresponding to Kuhn's conventional science and

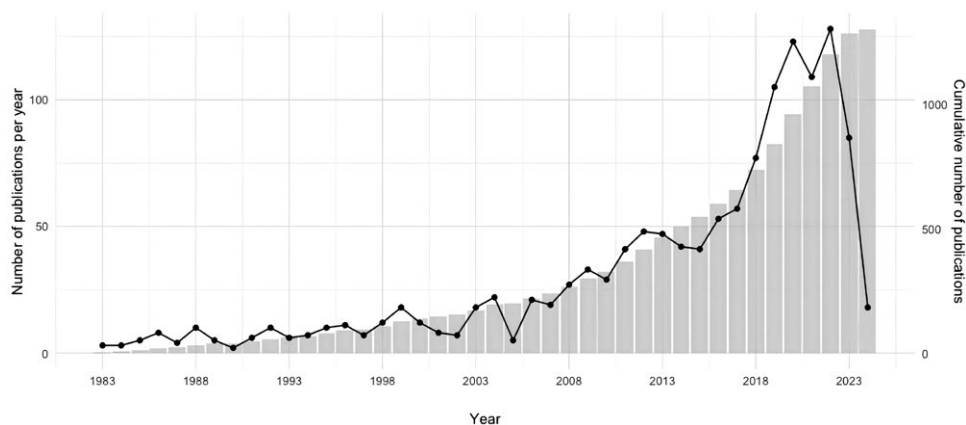


Figure 1. Growth in publications on multiword expressions from 1983 to 2024.

paradigm accumulation phase. The graph shows a steep upward trajectory in cumulative publications and steady growth in yearly publication counts, indicating that the field has gained momentum and coherence. This trend is also evident in special issues on multiword expressions published in linguistics journals, such as the *Annual Review of Applied Linguistics* (Wray 2012), *The Mental Lexicon* (Wulff and Titone 2014), *Language Teaching Research* (Siyanova-Chanturia 2017), and *Topics in Cognitive Science* (Christiansen and Amon 2017).

Influential authors, references, and journals

Most-cited authors.

Table 1 lists the top 25 most-cited authors in the field of multiword expressions. The majority of these scholars are affiliated with research institutions in the UK or USA, with exceptions including Batia Laufer, Britt Erman, Cristina Cacciari, and Michael Stubbs. Most authors are retired or nearing retirement age, though a few younger scholars such as Ana Siyanova-Chanturia and Philip Durrant are also noted. These authors have made significant contributions across various topics and approaches: phraseology or lexicography (e.g. Edward Finegan, James Nattinger, Peter Howarth, Tony Cowie), vocabulary acquisition (e.g. Alison Wray, Batia Laufer, Frank Boers, and Norbert Schmitt), corpus linguistics (e.g. Douglas Biber, John Sinclair, Michael Stubbs, Nadja Nesselhauf, Rosamund Moon, and Sylviane Granger), cognitive processing (Ana Siyanova-Chanturia, Cristina Cacciari, Joan Bybee, Nick Ellis, Philip Durrant, and Raymond Gibbs), and academic writing (e.g. Ken Hyland). Their foundational research has gained wide recognition and citations within the field. Several of these scholars also exhibit high centrality or sigma values. Centrality scores in CiteSpace measure an author's capacity to bridge diverse areas of study, with values typically ranging from 0 to 1. Higher centrality scores indicate greater influence and stronger connections across disciplines. Sigma scores, which can exceed 1 and have no upper limit, combine centrality with citation burstness and reflect both sustained influence and periods of significantly increased citations. Compared to other highly cited authors, those with high centrality or sigma scores are more influential because their works are cited by researchers across various subfields and domains, largely due to their innovative research methodologies, novel findings, and interdisciplinary backgrounds.

To enrich the overview discussed above, we also conducted burst analysis using CiteSpace. Figure 2 shows the top 25 authors with the strongest citation bursts. Citation burst strength typically ranges from values close to 0 to much higher numbers, depending on the intensity of the surge in citations during specific periods. Higher burst values indicate more significant and rapid increases in citation activity, reflecting heightened attention to the authors' work. From the 1980s to the early 2000s, the focus of citation bursts in the field centered

Table 1. Top 25 most-cited authors

Author	Citation	Centrality	Sigma	Degree
Wray A	226	0.01	1.00	24
Sinclair J	209	0.02	1.14	36
Gibbs R	179	0.17	142.85	83
Ellis N	159	0.02	1.00	28
Biber D	149	0.01	1.00	26
Pawley A	123	0.04	1.00	38
Finegan E	122	0.02	1.00	27
Schmitt N	120	0.01	1.04	19
Granger S	118	0.02	1.00	32
Nesselhauf N	102	0.03	1.00	30
Cacciari C	98	0.09	2.67	81
Durrant P	95	0	1.04	12
Cowie A (Tony)	91	0.15	1.00	47
Laufer B	89	0.03	1.15	20
Siyanova A	89	0.01	1.09	19
Nation P	88	0.02	1.10	22
Boers F	87	0.04	1.18	45
Howarth P	87	0.04	1.00	34
Bybee J	87	0.04	1.19	28
Erman B	87	0.01	1.03	18
Hyland K	86	0	1.02	16
Moon R	84	0.05	1.33	24
Hunston S	84	0.01	1.05	22
Stubbs M	82	0.05	1.37	29
Nattinger J	81	0.01	1.00	32

Note. Bolded authors are highly cited and have high centrality or sigma scores (top 25). Citation refers to the number of times an author is cited in our bibliometric database (not citations in Web of Science), when this author acts as the first author of a research paper.

on scholars specializing in phraseology (i.e. Dwight Bolinger, Maxine T. Boatner, and Adam Makkai) and the comprehension of idioms, metaphors, or figurative language (i.e. Andrew Ortony, Sam Bobrow, Brian Ackerman, David Swinney, George Lakoff, Wendy Schweigert, Marilyn Nippold, Cristina Cacciari, Philip Prinz, Sam Glucksberg, and M.C. Levorato). Since 2005, these bursts have shifted towards corpus analysis of multiword expressions (Michael Stubbs, John Sinclair), and more recently towards the acquisition of collocations (Stuart Webb, Brent Wolter) and the processing of word combinations (Michael Hoey, Ana Siyanova-Chanturia, and Philip Durrant). As illustrated in Fig. 2, early studies on phraseology and the comprehension of idioms and figurative language have sustained long-lasting citation bursts,

Top 25 Cited Authors with the Strongest Citation Bursts

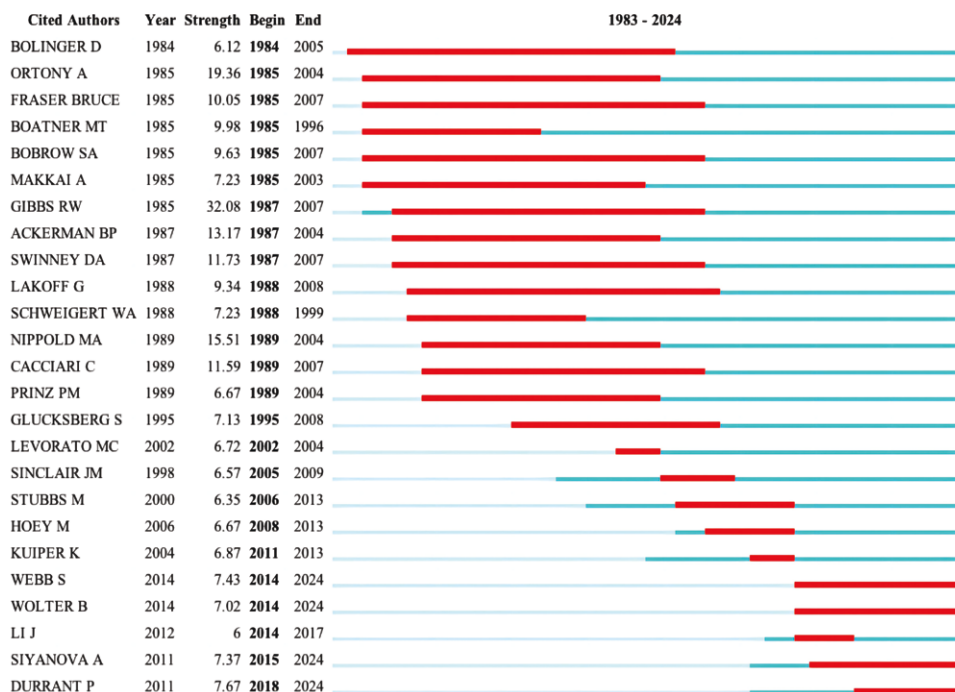


Figure 2. Top 25 cited authors with the strongest citation bursts.

indicating enduring impact in the field, particularly before the early 2000s. In contrast, scholars focusing on corpus analysis (e.g. Michael Stubbs, John Sinclair), despite their high citation rates, exhibit shorter burst durations, suggesting that the field of multiword expressions is more dynamic and undergoes more rapid shifts in research interests.

Most-cited references.

Table 2 presents the top 25 most-cited references in the field of multiword expressions (only the first author is shown, see [Supplementary Appendix S1](#) for reference details). To ensure comparability in assessing scholarly impact and original contribution, our analysis excluded books and book chapters. The citation counts provided by CiteSpace are based on the 1,302 papers included in our dataset. For a comprehensive evaluation of each reference's citation impact, we cross-referenced these counts with data from Web of Science. Overall, the ranking of the top 25 references by citation count in CiteSpace closely corresponds with their ranking in Web of Science. Both CiteSpace and Web of Science highlight that the following nine papers have the highest citation counts: Hyland (2008), Simpson-Vlach (2010), Arnon (2010), Ellis (2008), Conklin (2008), Laufer (2011), Chen (2010), Biber (2009), and Webb (2013). After excluding two review articles (Boers 2012; [Siyanova-Chanturia 2015](#)), the topics covered by the top most-cited references include: (1) development of phrasal expression lists using corpora (Simpson-Vlach 2010; Martinez 2012; Ackermann 2013); (2) language users' use of multiword expressions in writing (Hyland 2008; Biber 2009; Chen 2010; Laufer 2011; Adel 2012; Paquot 2012); (3) cognitive processing of multiword expressions (Jiang 2007; Conklin 2008; Ellis 2008; Arnon 2010; Durrant 2010; Siyanova-Chanturia 2011a, 2011b; Tremblay 2011; Wolter 2011; Wolter 2013); and (4) acquisition and teaching of multiword expressions (Webb 2013; Boers 2014). Except for Jiang (2007), [Siyanova-Chanturia \(2015\)](#), and Tremblay (2011), all references listed in [Table 2](#) also ranked within the top 25 in terms of

Table 2. Top 25 most-cited references

Publication	Citation_CSP	Citation_WOS	Centrality	Sigma
Webb S (2013)	24	190	0.02	1.13
Simpson-Vlach R (2010)	22	406	0.02	1.15
Chen YH (2010)	21	254	0.04	1.35
Ellis NC (2008)	20	290	0.07	1.78
Conklin K (2008)	19	277	0.01	1.06
Biber D (2009)	19	216	0.01	1.07
Siyanova-Chanturia A (2011a)	18	171	0.02	1.09
Laufer B (2011)	18	274	0.01	1.11
Arnon I (2010)	17	358	0.04	1.25
Hyland K (2008)	17	460	0.01	1.07
Wolter B (2013)	17	114	0.01	1.11
Wolter B (2011)	14	121	0.02	1.15
Siyanova-Chanturia A (2011b)	13	173	0.01	1.07
Adel A (2012)	13	173	0.01	1.04
Martinez R (2012)	12	167	0.01	1.05
Ackermann K (2013)	12	138	0.01	1.05
Boers F (2012)	12	106	0.01	1.04
Durrant P (2010)	11	111	0.02	1.07
Siyanova-Chanturia A (2015)	11	86	0	1.02
Jiang N (2007)	11	137	0	1.01
Boers F (2014)	10	86	0.01	1.03
Durrant P (2010)	10	111	0.02	1.08
Paquot M (2012)	10	126	0.01	1.03
Tremblay A (2011)	10	179	0	1.00
Biber D (2007)	9	46	0.02	1.09

Note. Citation counts from CiteSpace (Citation_CSP) are based on co-citation patterns within our dataset of publications on multiword expressions. Citation_WOS refers to the number of citations a publication receives based on Web of Science. Discrepancies in citation counts and rankings may occur when compared to Google Scholar, which includes a broader range of sources such as books, conference papers, and unpublished works. Bolded authors are highly cited, yet their centrality or sigma scores do not rank in the top 25.

centrality or sigma values, which indicates that they play a crucial role in connecting different subfields in research on multiword expressions.

Figure 3 presents the top 25 references exhibiting the strongest citation bursts. (only the first author is shown, see [Supplementary Appendix S2](#) for reference details) Except for a few papers (Biber 2007; Durrant 2010; Tremblay 2011; Martinez 2012; Boers 2014; [Siyanova-Chanturia 2015](#)), most papers listed in [Table 2](#) are included in this list. This indicates that these studies sparked citation bursts and attracted significant research interest during different time periods. Among

Top 25 References with the Strongest Citation Bursts

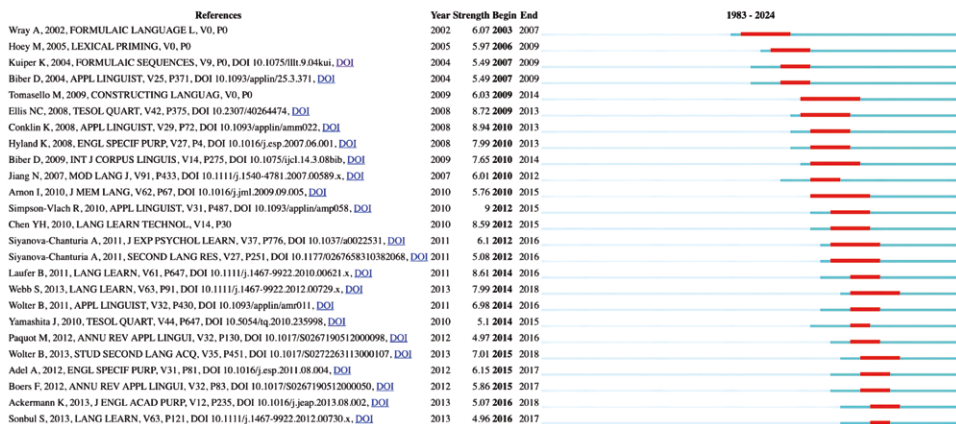


Figure 3. Top 25 references with the strongest citation bursts.

the papers experiencing citation bursts, those with the strongest burst strength include Ellis (2008), Conklin (2008), Chen (2010), Laufer (2011), and Hyland (2008). This suggests that topics such as cognitive processing of multiword expressions and their use in writing, as explored by these references, attracted the highest attention from researchers in the field. Regarding the duration of citation bursts, Ellis (2008), Chen (2010), Hyland (2008), Ackerman (2013), and Webb (2013) maintained bursts lasting more than 5 years. This indicates enduring interest and ongoing research in areas such as cognitive processing of multiword expressions, their use in academic writing, and incidental learning of collocations. In addition to references overlapping with those in Table 2, Fig. 3 also highlights several influential books focusing on or related to multiword expressions (Wray 2002; Hoey 2005; Tomasello 2009). Furthermore, references such as Kuiper (2004), Biber (2004), Yamashita (2010), and Sonbul (2013) are featured, covering topics such as multiword expressions in speech, treatment of multiword expressions in textbooks, congruency effects in the processing of multiword expressions, and the acquisition of multiword expressions under different conditions. Lastly, Fig. 3 also illustrates the evolution of intellectual background in the field of multiword expressions. Specifically, the field initially concentrated on theoretical works on multiword expressions, including Wray's (2002) on formulaic language, Hoey's (2005) on lexical priming models, and Tomasello's (2009) on usage-based theory. Subsequently, references with citation bursts began to address more specific issues such as the processing of multiword expressions, the development of phrasal materials for writing, and the acquisition of multiword expressions. This evolutionary pattern aligns with Kuhn's concept of the accumulation of conventional science and paradigms (Kuhn 1962), suggesting that research on multiword expressions has established theoretical frameworks and expanded rapidly into empirical research around the years 2008–9, as discussed in the previous section.

Most-cited journals.

Table 3 lists the top 25 most-cited journals in the field of multiword expressions. As can be seen, research on multiword expressions frequently cites studies published in prestigious linguistics and applied linguistics journals, such as *Applied Linguistics*, *Journal of Memory and Language*, and *Language Learning*. Additionally, journals specializing in subfields of research, such as pragmatics (*Journal of Pragmatics*), corpus linguistics (*International Journal of Corpus Linguistics*), cognitive linguistics (*Cognitive Linguistics*), and language acquisition (e.g. *Studies in Second Language Acquisition*, *Journal of Child Language*), are also highly referenced. Furthermore, several interdisciplinary journals focusing on psycholinguistics and experimental psychology are also frequently cited,

Table 3. Top 25 most-cited journals

Journal	Citation counts	Centrality
Applied Linguistics	334	0.15
TESOL Quarterly	216	0.08
Language Learning	200	0.02
Language	191	0.14
Journal of Memory and Language	184	0.06
Language & Communication	152	0.09
Studies in Second Language Acquisition	150	0.02
International Journal of Corpus Linguistics	150	0.04
Memory and Cognition	147	0.05
Journal of Pragmatics	134	0.08
Journal of Verbal Learning and Verbal Behavior	127	0.06
English for Specific Purposes	124	0.02
Journal of Experimental Psychology: Learning, Memory, and Cognition	118	0.04
System	116	0.03
International Review of Applied Linguistics	116	0.05
Cognition	106	0.07
Second Language Research	97	0.02
ELT Journal	94	0.04
Modern Language Journal	91	0.02
Brain and Language	89	0.05
Journal of Child Language	89	0.05
Annual Review of Applied Linguistics	87	0
Journal of Psycholinguistic Research	85	0.06
Cognitive Linguistics	84	0.03
Language Teaching Research	83	0.04

including *Memory and Cognition*, *Cognition*, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *Journal of Psycholinguistic Research*, and *Brain and Language*. Among the top 10 most-cited journals, seven also exhibit high centrality scores: *Applied Linguistics*, *Language*, *Language & Communication*, *TESOL Quarterly*, *Journal of Pragmatics*, *Journal of Memory and Language*, and *Memory and Cognition*. Conversely, three journals, namely, *International Journal of Corpus Linguistics*, *Language Learning*, and *Studies in Second Language Acquisition*, are among the top 10 most-cited but have relatively low centrality scores. This suggests that while these journals are highly influential in publishing research on multiword expressions, their work is less connected with other research areas. This phenomenon might be attributable to the fact that publications on multiword expressions in these journals originate from a relatively small group of researchers or that these journals exert more influence on specific topics on multiword expressions.

Collaboration networks between authors, institutions, and countries

Collaboration networks between authors.

Using CiteSpace, we analyzed the collaboration patterns among authors in the field of multiword expressions and illustrated these patterns with network maps. We set the threshold at four publications, such that authors with fewer than four publications related to multiword expressions were not labeled on the map. In the network map, each node represents an individual author, and the size of the node indicates the author's productivity, measured by the number of publications. Nodes are connected by links, indicating that two authors have co-authored one or more publications. The clusters are groups of interconnected nodes and represent groups of authors who frequently collaborate with each other. Colors are used to represent different time periods, with older collaborations shown in blue and more recent collaborations in warmer colors (i.e. green and yellow).

As shown in Fig. 4, the network map reveals multiple clusters, each representing a group of authors who frequently collaborate. Clusters where the number of publications per author is fewer than four were not labeled. Among the labeled clusters, most are centered around a single leading scholar, with other authors in the cluster being less influential and having fewer publications. These clusters are represented by scholars including B. Keysar, C. D. Qualls, Diana Van Lancker Sidtis, Douglas Biber, M. A. Nippold, Nick C. Ellis, R. W. Gibbs, and W. A. Schweigert. Many of these clusters are in blue (except for those led by Diana Van Lancker Sidtis and Nick C. Ellis), suggesting that these collaborations were formed before or around the early 2000s. These small-size clusters consist of relatively few authors, and the topics they address include corpus-based analysis of multiword expressions (Douglas Biber), cognitive processing of multiword expressions (Nick C. Ellis, Diana Van Lancker Sidtis), comprehension of idioms (M. A. Nippold, W. A. Schweigert, C. D. Qualls), and comprehension of metaphors (B. Keysar, R. W. Gibbs). Interestingly, despite sharing a research interest in the comprehension of idioms, the three clusters led by C. D. Qualls, M. A. Nippold, and W.A. Schweigert are not connected with each other. This lack of connection may be due to these authors taking different approaches or focusing on different aspects of idiom comprehension. For example, Qualls focused on the comprehension of idioms in

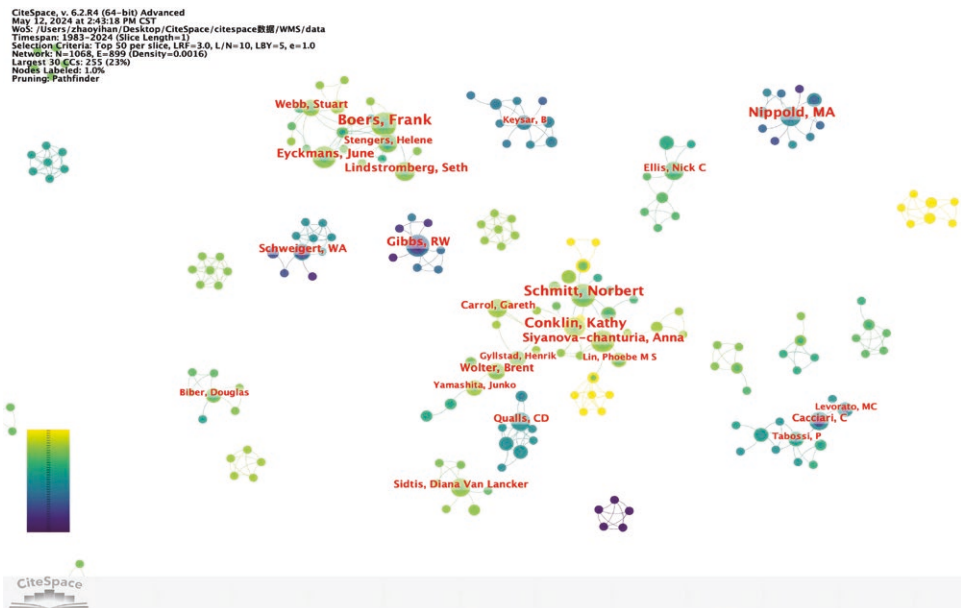


Figure 4. Collaboration networks between authors.

adolescents with learning disabilities, whereas Nippold focused on the comprehension of idioms in adolescents with mental health problems.

Figure 4 presents three larger-sized clusters of authors. The first cluster located in the bottom right corner is led by Cristina Cacciari and her two collaborators, M. C. Levorato and P. Tabossi. This group of researchers focused on the comprehension of idioms by native Italian speakers, including aphasic patients. This collaboration network is in dark blue, indicating it was formed in the early 2000s and is not currently active, likely due to the retirement of Cristina Cacciari. The second cluster in the upper left corner consists of five leading scholars: Frank Boers, Helene Stengers, June Eyckmans, Seth Lindstromberg, and Stuart Webb. The nodes representing these scholars are relatively large, indicating they are very productive and influential in the field. The topics addressed by these scholars mainly concern the acquisition and teaching of multiword expressions. The green nodes in this cluster suggest that their collaborations are more recent (around the 2010s). The third cluster is the largest among all collaboration networks in the field of multiword expression research. This cluster includes a large number of authors, with the most productive and influential being Norbert Schmitt, Kathy Conklin, Anna Siyanova-Chanturia, Gareth Carroll, and Brent Wolter. Upon closer examination, this cluster can be broken down into two smaller, relatively independent collaboration networks. Scholars and former PhD alumni from the University of Nottingham (Norbert Schmitt, Kathy Conklin, Ana Siyanova-Chanturia, Phoebe Lin, and Gareth Carroll) frequently collaborate with each other, focusing on the acquisition and processing of multiword expressions. In contrast, Henrik Gyllstad, Brent Wolter, and Junko Yamashita lead another related but more independent research group with a more specialized focus, namely, the congruency effect on the processing of collocations. Henrik Gyllstad's research on the processing of multiword expressions and the testing of vocabulary knowledge positions him as a bridging figure connecting the two groups of researchers, as depicted in the figure. Lastly, the third cluster is in green and yellow, indicating that these collaborations were formed more recently with new scholars joining the collaboration network, making this cluster of research the most active.

Collaboration networks between institutions.

A total of 524 institutions were identified from our bibliometric database. We set a threshold such that institutions with fewer than eight publications on multiword expressions were not labeled on the map. Figure 5 illustrates the collaboration patterns among research institutions that publish studies on multiword expressions, with the 15 most productive institutions labeled. As shown in this figure, these institutions are located in the USA (5), the UK (3), Belgium (3), Hong Kong (2), Israel (1), and New Zealand (1). Among them, the top three—the University of Nottingham, Victoria University of Wellington, and the University of California System—have the largest node sizes and the greatest number of links connecting to other nodes, suggesting that they significantly contribute to the field of multiword expressions due to their high productivity and extensive global collaborations. The nodes representing the top 15 research institutions are in green, indicating that these institutions remain highly active in generating research on multiword expressions. Some unlabeled institutions appear in dark blue, suggesting that they are no longer active in the field. In contrast, some nodes are in bright yellow, indicating that these institutions have recently joined the collaboration network and have started publishing papers on multiword expressions in recent years.

Collaboration networks between countries.

A total of 58 countries were identified in our bibliometric database. We set a threshold such that countries with fewer than three publications on multiword expressions were excluded from the analysis. Figure 6 presents the collaboration networks between countries in the field of research on multiword expressions. As can be seen, multiword expressions have attracted global interest, as the countries labeled in the figure are distributed worldwide. In terms of research productivity, the top ten countries include the USA, the UK, China, Spain, Belgium, Germany, New Zealand,

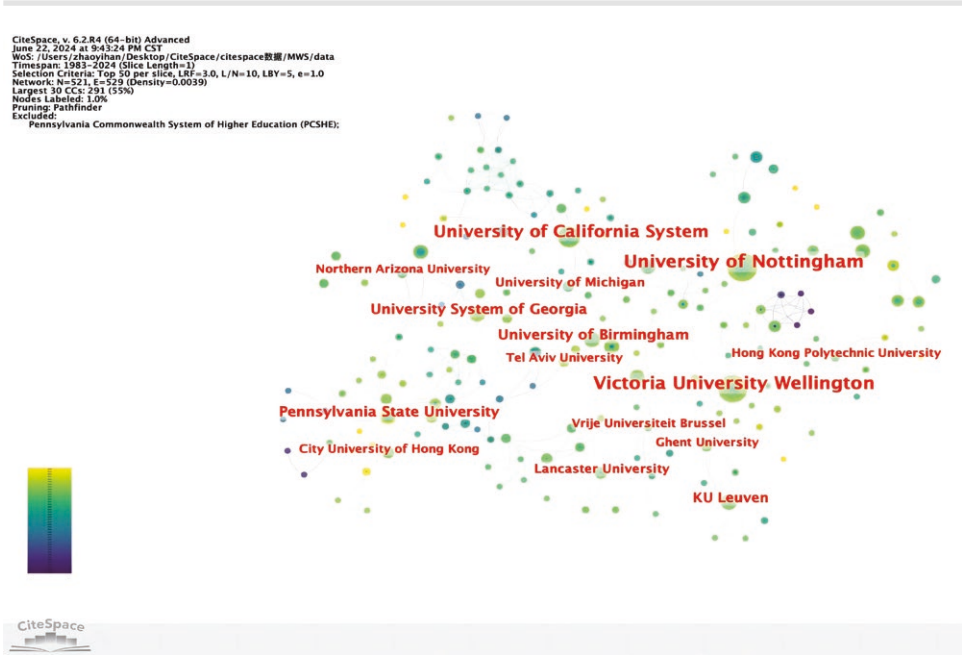


Figure 5. Collaboration networks between institutions.

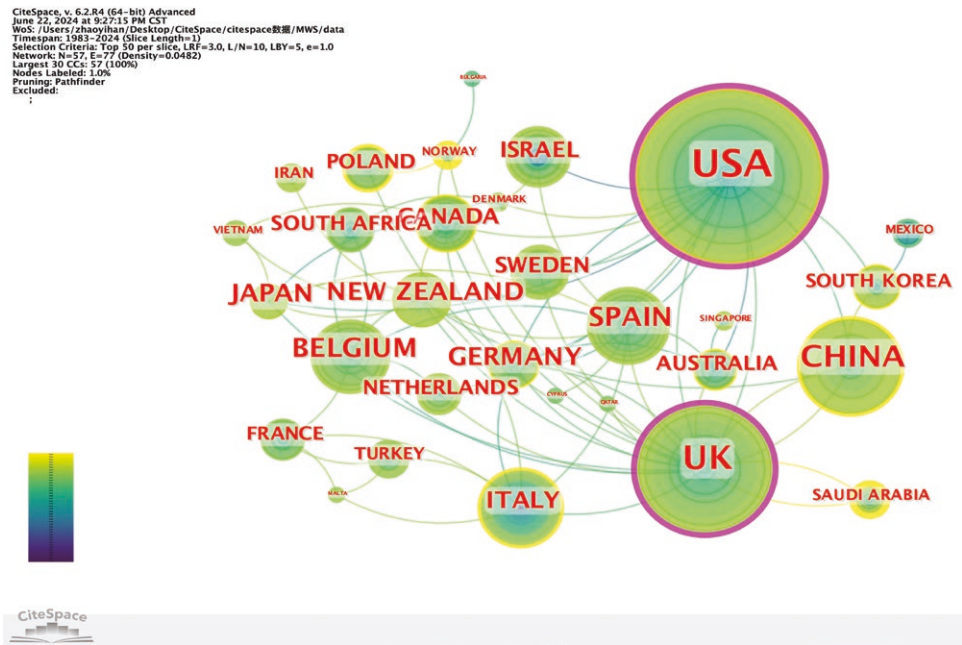


Figure 6. Collaboration networks between countries.

Japan, Italy, and Israel. Regarding centrality in the collaboration network, the top 10 central countries are the UK, the USA, Spain, Italy, New Zealand, Canada, Germany, Belgium, Australia, and South Korea. Comparing the centrality rankings against the productivity rankings, it is evident that countries such as China and Japan, although very productive, play a less central role in the

research network, due to fewer international collaborations with other countries. Conversely, Canada, despite being less productive, is positioned more centrally due to its extensive international collaborations. Overall, the USA and the UK act as the most prominent countries, with the largest node sizes and surrounded by thick, purple rings. This suggests that these two countries are central in the collaboration network, exhibiting the highest production of research and the most extensive international collaborations in the field. Notably, countries such as Saudi Arabia, Norway, Poland, and China have nodes with yellow outer rings, indicating they have been very active in the field of research on multiword expressions in recent years.

Key research areas

To identify key research areas in the study of multiword expressions, we conducted a co-citation analysis of references using CiteSpace. This analysis identified 132 clusters, indicating distinct research themes within the field. The modularity Q value of 0.88 suggests that the clusters in the co-citation network are well-separated, whereas the weighted mean silhouette score of 0.95 suggests that the references within each cluster are coherent in the research theme and distinct from other clusters. After excluding clusters with few cited references or citing articles, as well as those overlapping in research themes, we identified five major clusters representing distinct research themes, as illustrated in Fig. 7.

Among the five major research themes, idiom comprehension (cluster_2) emerged earliest, predominantly before the 2000s. This line of research focuses on the processing of figurative and literal meanings of idioms (e.g. Popiel and McRae 1988), the decomposition and analyzability of idioms (e.g. Gibbs et al. 1989), and the processing of idioms among brain-damaged participants (e.g. Tompkins et al. 1992). Another early-established area is labeled 'corpus analysis' (cluster_4), which involves studies using corpus-based techniques to identify different types of multiword expressions (e.g. Biber et al. 2004; Durrant 2009) and studies validating the processing advantage of multiword expressions extracted from corpora (e.g. Ellis et al. 2008). Cluster 0 encompasses

```
CiteSpace, v. 6.2.R4 (64-bit) Advanced
July 25, 2024 at 11:01:12 PM CST
WOS: /Users/zhaoyihan/Desktop/CiteSpace/citespace数据/MWS/data
Timespan: 1983-2024 (Slice Length=1)
Selection Criteria: q-index (k=25), LRF=3.0, L/N=10, LBY=5, e=1.0
Network: N=1059, E=2823 (Density=0.005)
Nodes Labeled: 1.0%
Pruning: Pathfinder
Modularity Q=0.8837
Weighted Mean Silhouette S=0.9473
Harmonic Mean(Q, S)=0.9144
Excluded:
:
```

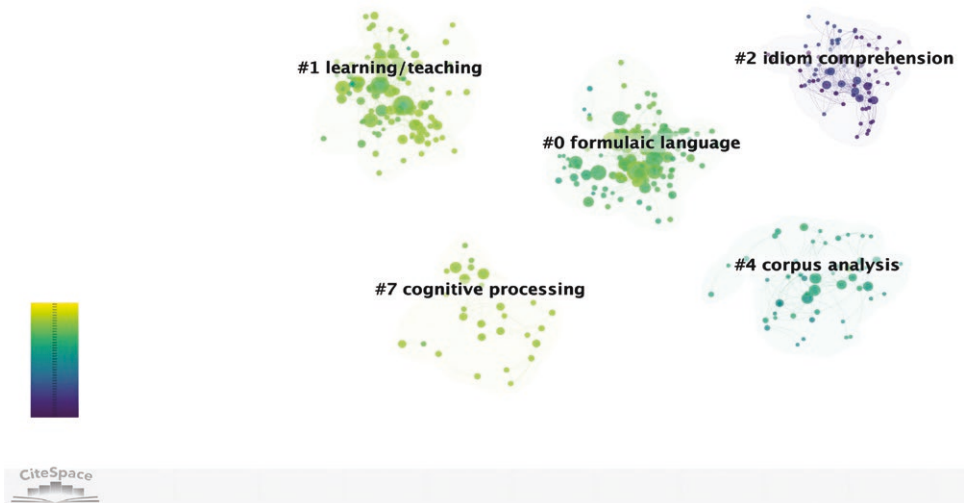


Figure 7. Key research areas in the field of study on multiword expressions.

studies conducted under the umbrella term of ‘formulaic language’. Coined by Wray (2002), formulaic language refers to both multiword and single-word items, which are believed to be stored and retrieved as holistic units from memory at the time of use, rather than being generated or analyzed through grammar. This definition has gained wide acceptance, though it narrows the scope of multiword expressions to more fixed linguistic units and equates the processing advantage with holistic storage, which may not always hold true (Siyanova-Chanturia 2015; Yi and Zhong 2024). Research in this cluster includes the development of phrasal expression lists (e.g. Simpson-Vlach and Ellis 2010; Martinez and Schmitt 2012) and the processing of various types of larger-than-word units (e.g. Millar 2011; Conklin and Schmitt 2012). In addition to these themes, two other themes, including the learning and teaching of multiword expressions (cluster_1) and the cognitive processing of multiword expressions (cluster_7), also attract considerable attention in the field. Multiword expressions are frequently used by language users, yet they are notoriously difficult for L2 speakers to acquire (Boers et al. 2014). Studies in cluster 1 explored topics such as factors influencing the learning of collocations (e.g. Peters 2016), incidental learning of collocations (e.g. Pellicer-Sánchez 2017), and the use of multiword expressions in L2 writing (e.g. Siyanova-Chanturia 2015). Regarding the cognitive processing of multiword expressions (cluster_7), this research theme is relatively new but rapidly expanding. Studies in this cluster have addressed topics such as L1-L2 congruency effects in collocation processing (e.g. Yamashita 2018), semantic transparency and the processing of collocations (e.g. Gyllstad and Wolter 2016), and frequency effects in collocation processing (e.g. Wolter and Yamashita 2018).

Frequently explored topics and emerging trends

To examine the most frequently explored topics in the study of multiword expressions, we conducted a co-occurrence analysis of terms using CiteSpace. This technique analyzes terms extracted from the titles, abstracts, keywords, and full texts of publications in our bibliometric database. After testing various thresholds, we determined that selecting terms appearing more than six times provided the best balance between clarity and coverage. Lower thresholds included too many terms, which obscured discernible patterns, while higher thresholds excluded significant terms, reducing the comprehensiveness of the analysis. The terms were mapped in a network, as shown in Fig. 8. Our analysis identified the following terms with the largest nodes, indicating that they are the most frequently explored topics: formulaic language/formulaic sequences, multiword units/multiword units, lexical bundles, phrasal verbs, idiomatic expression/idiom comprehension/figurative language, corpus linguistics, academic writing, native speakers/first language, non-native speakers/foreign language/second language/second language acquisition/EFL learners/L2 learners, mental lexicon, and cognitive linguistics. Among these frequently explored topics, several had purple rings surrounding their nodes, indicating high betweenness-centrality scores: lexical bundles, non-native speakers/second language learners, idiomatic expression/literal meaning/idiom comprehension/figurative language. These topics are most prominent and play an important role in connecting studies from various theoretical and methodological perspectives.

Formulaic sequences/language and multiword units are two umbrella terms used to refer to multiword expressions, although the former also includes single-word items, such as expletives (e.g. *damn*, *bastard*) and exclamations (e.g. *wow*, *oops*), conversational speech formulas (e.g. *hello*, *thanks*), and other formulaic items at the word level. Figure 8 shows that the term formulaic language/sequences has been used much earlier and more frequently than multiword units. According to Fig. 8, lexical bundles, phrasal verbs, and idioms are the most explored subcategories of multiword expressions, with the interest in collocations growing in recent years. Compared with studies using methods such as corpus analysis, comparative/contrastive analysis, or case reports, a growing number of studies explore the processing and representation of multiword expressions in the mental lexicon using experimental tasks such as lexical decision. Noticeably, although the node for cognitive linguistics is positioned less centrally than other research topics, its relatively large size and the mixture of green and yellow colors suggest that this branch of

research dates back several decades and is growing in recent years. Lastly, in terms of the evolution of research topics, the following patterns can be observed. First, research on multiword expressions seemed to focus more on native speakers (including L1 children) before the 2000s, yet more studies have been carried out targeting L2 speakers in recent years. Second, research on language development and idiom comprehension (especially among brain-damaged subjects and aging adults) received much attention before the 2000s (indicated by dark blue colors), yet these lines of study seem to be no longer active in recent years. By contrast, topics such as incidental learning of multiword expressions, phraseological competence, and the use of multiword expressions in academic writing are gaining popularity and receiving increasing attention, as indicated by bright yellow colors for their nodes.

To evaluate the amount of attention received by each topic, the duration of interest in each topic, and the evolution of research interest in the field of multiword expressions over time, we conducted a burst detection analysis of terms extracted from our database. Burst detection in CiteSpace reports the strength of the burst as well as the beginning and ending years for each term. The strength of the burst quantifies the attention received by each topic, while the beginning and ending years mark the duration of interest. Figure 9 presents the top 25 terms with the strongest citation bursts. The following terms, as shown in large nodes in Fig. 8, also had strong bursts: lexical bundles, idiom comprehension, academic writing, formulaic language, foreign language/second language, corpus linguistics, and phrasal verbs. This indicates that these terms represent the most heated topics in the study of multiword expressions. Research interest in idiom comprehension surged from 1998 to 2010, but this topic no longer receives continuing attention. Interestingly, corpus-based studies of multiword expressions first saw a surge in citations in 2004, particularly those utilizing corpora such as the British National Corpus compiled from native speakers. Although learner corpora had been used earlier in research on multiword expressions, our citation burst analysis shows a significant increase in attention to learner corpora for studying multiword units starting only in 2019, much later than the surge in the use of native speaker corpus data. As mentioned earlier, researchers also examined the processing

CiteSpace, v. 6.2.R4 (64-bit) Advanced
June 30, 2024 at 5:37:18 PM CST
WOS: /Users/zhaozhan/Desktop/CiteSpace/citespace数据/MWS/data
Timespan: 1983–2024 (Slice Length=1)
Selection Criteria: g-index (k=15), LRF=3.0, L/N=10, LBY=5, e=1.0
Network: N=628, E=965 (Density=0.0049)
Largest CC: 506 (80%)
Nodes Labeled: 1.0%
Pruning: Pathfinder
Excluded:

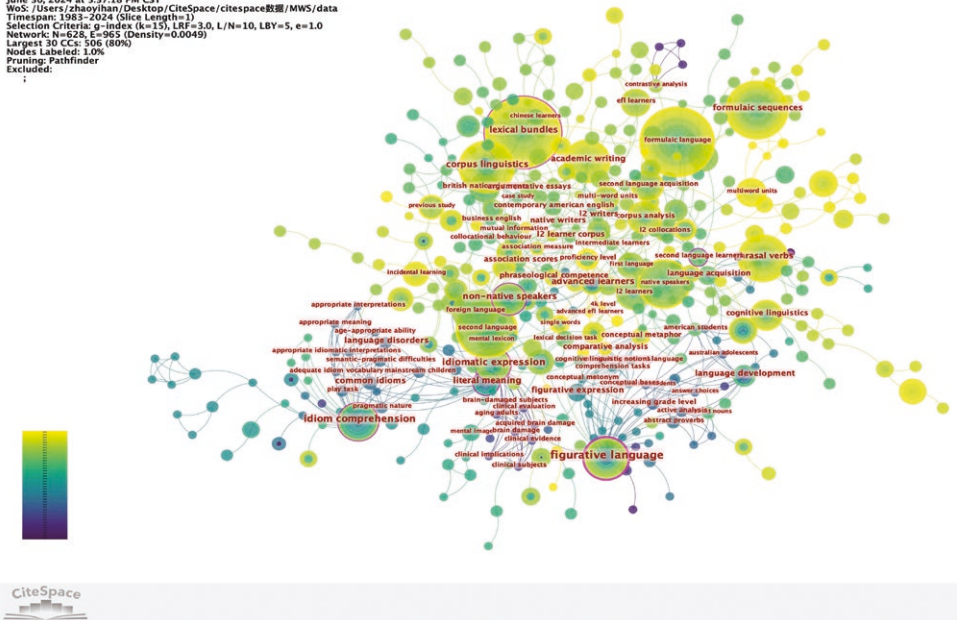


Figure 8. Co-occurrence analysis of terms.

Top 25 Terms with the Strongest Citation Bursts



Figure 9. Top 25 terms with the strongest citation bursts.

of multiword expressions using controlled experiments, and such a practice gained a surge in attention from 2013 to 2017. Regarding the evolution of research interests and emerging trends, according to Fig. 9, early research focused on idiom comprehension among native speakers. By contrast, later studies expanded to broader categories of multiword expressions (under the umbrella term ‘formulaic language’) and to second language learners. Figure 9 shows that the most recent interests lie in topics such as incidental learning of multiword expressions, lexical bundles, the use of multiword expressions in academic writing, corpus analysis of larger-than-word units, and the study of multiword expressions from the theoretical perspective of construction grammar. Needless to say, these topics are very likely to emerge as trends and guide research on multiword expressions in the coming decades.

Conclusion

In this study, we present a comprehensive overview of research on multiword expressions. Using bibliometric analysis, we examined a large body of research articles published over the past four decades, tracing the growth of publications, identifying influential authors, research papers, and academic journals, and mapping collaborative networks among scholars, institutions, and countries. Our analysis also highlights key research areas, frequently explored topics, and emerging trends in the field. The results show that multiword expressions will remain a significant topic in applied linguistics. This area will also serve as a testing ground for theoretical frameworks, such as usage-based theories. The growing global interest suggests that research on multiword expressions in languages other than English will likely increase, with central topics including L2 speakers’ acquisition of multiword units under different learning conditions, cognitive mechanisms involved in processing multiword units, and the development of phraseological competence in relation to language skills. Our study also demonstrates the usefulness of bibliometric analysis as a tool for examining the research landscape within specific linguistic domains. It also offers valuable insights for graduate students seeking to identify supervisors with expertise in multiword expressions and helps early-career scholars navigate the literature and select research topics to explore in this field of study. Additionally, this analysis can inform the design of both undergraduate- and graduate-level courses that focus on multiword units, ensuring that the curricula reflect the most-relevant and up-to-date knowledge in the field.

Our findings reveal that research topics in this field have evolved from a narrow focus on idioms and fixed expressions to a broader exploration of other subcategories of multiword units, such as lexical bundles and collocations. This shift has been accompanied by a surge in publications, with hundreds of studies being published each year, exploring diverse topics related to multiword expressions. The development and accessibility of large corpora in recent years have played a crucial role in this process, and there are strong reasons to believe that corpus-driven or corpus-informed research on multiword expressions will continue to drive the growth of the literature and lead to the identification of additional subcategories of multiword units, particularly in languages other than English. Our bibliometric analysis also reveals two notable trends in the study of multiword expressions: the increasing focus on second language acquisition and the interdisciplinary nature of this research area. The growing emphasis on the acquisition and processing of multiword units by L2 learners is largely due to the widespread recognition of their importance for second language proficiency. Given that second language pedagogy has long focused on the teaching of words and syntactic rules, the expanding literature on L2 learners’ acquisition and processing of multiword expressions is likely to inspire innovative teaching practices in the near future. Regarding the interdisciplinary nature of the field, as research on multiword units incorporates insights from various perspectives such as corpus linguistics, psycholinguistics, cognitive linguistics, second language acquisition, and language pathology, we believe this field will witness advancements not only in research methodologies but also in deeper insights into the understanding of larger-than-word units.

As research on multiword expressions continues to expand, several directions for future studies are worth highlighting. First, researchers should place greater emphasis on languages other than English, particularly on subcategories of multiword expressions that remain underexplored. Second, in order to deepen our understanding of how multiword expressions are comprehended, acquired, and processed, as well as the roles they play in language and human cognition, research should integrate theoretical frameworks and methodologies from various disciplines. Third, research interest groups and academic institutions should actively support and initiate international collaborations to enhance the scale and quality of research on multiword expressions and drive innovation in the field. Fourth, the communication between different branches of research on multiword expressions should also be strengthened, which helps translate research findings more effectively into innovations in language pedagogy. Finally, more effort should be directed toward building and testing theoretical models that unify existing research findings, providing a solid framework for guiding future studies.

Supplementary data

Supplementary data is available at *Applied Linguistics* online.

Notes on contributors

Wei Yi is an assistant professor at Peking University. He holds a Ph.D. in Second Language Acquisition from the University of Maryland, College Park. His research focuses on vocabulary acquisition and lexical processing. He has published in the journals *Studies in Second Language Acquisition*, *Language Learning*, *Second Language Research*, *System*, and *Applied Psycholinguistics*.

Yihan Zhao is currently a graduate student supervised by Wei Yi. She is interested in the role of language aptitude and individual differences in second language acquisition.

Wanxin Wang is a graduate student supervised by Wei Yi. Her primary interest is the acquisition and processing of multiword expressions.

Acknowledgement

This research was funded by a grant from the National Social Science Fund of China (No. 23CYY057) to Wei Yi.

References

- Armon, I., and Snider, N. (2010) 'More than Words: Frequency Effects for Multi-Word Phrases', *Journal of Memory and Language*, 62: 67–82. <https://doi.org/10.1016/j.jml.2009.09.005>
- Biber, D., Conrad, S., and Cortes, V. (2004) 'If You Look At...: Lexical Bundles in University Teaching and Textbooks', *Applied Linguistics*, 25: 371–405.
- Boers, F. (2020) 'Factors Affecting the Learning of Multiword Items', in S. Webb (ed.) *The Routledge Handbook of Vocabulary Studies*, pp. 143–57. London: Routledge.
- Boers, F., et al. (2014) 'Gauging the Effects of Exercises on Verb-Noun Collocations', *Language Teaching Research*, 18: 54–74.
- Chen, C. M. (2006) 'CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature', *Journal of the American Society for Information Science and Technology*, 57: 359–77.
- Christiansen, M. H., and Armon, I. (2017) 'More than Words: The Role of Multiword Sequences in Language Learning and Use', *Topics in Cognitive Science*, 9: 542–51. <https://doi.org/10.1111/tops.12274>
- Christiansen, M. H., and Chater, N. (2016) 'The Now-or-Never Bottleneck: A Fundamental Constraint on Language', *Behavioral and Brain Sciences*, 39: 1–72.

- Conklin, K., and Schmitt, N. (2012) 'The Processing of Formulaic Language', *Annual Review of Applied Linguistics*, 32: 45–61. <https://doi.org/10.1017/s0267190512000074>
- Durrant, P. (2009) 'Investigating the Viability of a Collocation List for Students of English for Academic Purposes', *English for Specific Purposes*, 28: 157–69. <https://doi.org/10.1016/j.esp.2009.02.002>
- Ellis, N. C., Simpson-Vlach, R., and Maynard, C. (2008) 'Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL', *TESOL Quarterly*, 42: 375–96.
- Gibbs, R. W., Nayak, N., and Cutting, C. (1989) 'How to Kick the Bucket and Not Decompose: Analyzability and Idiom Processing', *Journal of Memory and Language*, 28: 576–93.
- Gyllstad, H., and Wolter, B. (2016) 'Collocational Processing in Light of the Phraseological Continuum Model: Does Semantic Transparency Matter?', *Language Learning*, 66: 296–323.
- Kuhn, T. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Lei, L., and Liu, D. (2019) 'Research Trends in Applied Linguistics from 2005 to 2016: A Bibliometric Analysis and Its Implications', *Applied Linguistics*, 40: 540–61.
- Martinez, R., and Schmitt, N. (2012) 'A Phrasal Expressions List', *Applied Linguistics*, 33: 299–320. <https://doi.org/10.1093/applin/ams010>
- Millar, N. (2011) 'The Processing of Malformed Formulaic Language', *Applied Linguistics*, 32: 129–48.
- Nattinger, J., and DeCarrico, J. (1992) *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
- Pawley, A., and Syder, F. H. (1983) 'Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency', in J. Richards and R. Schmidt (eds) *Language and Communication*, pp. 191–225. London: Routledge.
- Pellicer-Sánchez, A. (2017) 'Learning L2 Collocations Incidentally from Reading', *Language Teaching Research*, 21: 381–402.
- Peters, E. (2016) 'The Learning Burden of Collocations: The Role of Interlexical and Intralexical Factors', *Language Teaching Research*, 20: 113–38.
- Popiel, S. J., and Mcrae, K. (1988) 'The Figurative and Literal Senses of Idioms, or All Idioms are Not Used Equally', *Journal of Psycholinguistic Research*, 17: 475–87. <https://doi.org/10.1007/bf01067912>
- Shaoul, C., and Westbury, C. (2011) 'Formulaic Sequences: Do They Exist and Do They Matter?', *The Mental Lexicon*, 6: 171–96. <https://doi.org/10.1075/ml.6.1.07sha>
- Simpson-Vlach, R., and Ellis, N. C. (2010) 'An Academic Formulas List: New Methods in Phraseology Research', *Applied Linguistics*, 31: 487–512. <https://doi.org/10.1093/applin/amp058>
- Siyanova-Chanturia, A. (2015) 'Collocation in Beginner Learner Writing: A Longitudinal Study', *System*, 53: 148–60. <https://doi.org/10.1016/j.system.2015.07.003>
- Siyanova-Chanturia, A. (2013) 'Eye-Tracking and ERPs in Multi-Word Expression Research: A State-of-the-Art Review of the Method and Findings', *The Mental Lexicon*, 8: 245–68. <https://doi.org/10.1075/ml.8.2.06siy>
- Siyanova-Chanturia, A. (2017) 'Researching the Teaching and Learning of Multi-Word Expressions', *Language Teaching Research*, 21: 289–97. <https://doi.org/10.1177/1362168817706842>
- Siyanova-Chanturia, A., and Martinez, R. (2015) 'The Idiom Principle Revisited', *Applied Linguistics*, 36: 549–69.
- Tomasello, M. (2000) 'First Steps Toward a Usage-Based Theory of Language Acquisition', *Cognitive Linguistics*, 11: 61–82.
- Tompkins, C. A., Boada, R., and McGarry, K. (1992) 'The Access and Processing of Familiar Idioms by Brain-Damaged and Normally Aging Adults', *Journal of Speech and Hearing Research*, 35: 626–37. <https://doi.org/10.1044/jshr.3503.626>
- Weinert, R. (1995) 'The Role of Formulaic Language in Second Language Acquisition: A Review', *Applied Linguistics*, 16: 180–205. <https://doi.org/10.1093/applin/16.2.180>
- Wolter, B., and Yamashita, J. (2018) 'Word Frequency, Collocational Frequency, L1 Congruency, and Proficiency in L2 Collocational Processing What Accounts for L2 Performance?', *Studies in Second Language Acquisition*, 40: 395–416.

- Wood, D. (2002) 'Formulaic Language in Acquisition and Production: Implications for Teaching', *TESL Canada Journal*, 20: 01–15. <https://doi.org/10.18806/tesl.v20i1.935>
- Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge, UK: Cambridge University Press.
- Wray, A. (2000) 'Formulaic Sequences in Second Language Teaching: Principle and Practice', *Applied Linguistics*, 21: 463–89. <https://doi.org/10.1093/applin/21.4.463>
- 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: 231–54. <https://doi.org/10.1017/s026719051200013x>
- Wray, A., and Perkins, M. R. (2000) 'The Functions of Formulaic Language: An Integrated Model', *Language and Communication*, 20: 1–28. [https://doi.org/10.1016/s0271-5309\(99\)00015-4](https://doi.org/10.1016/s0271-5309(99)00015-4)
- Wulff, S., and Titone, D. (2014) 'Bridging the Methodological Divide: Linguistic and Psycholinguistic Approaches to Formulaic Language', *The Mental Lexicon*, 9: 371–6. <https://doi.org/10.1075/ml.9.3.001int>
- Yamashita, J. (2018) 'Possibility of Semantic Involvement in the L1-L2 Congruency Effect in the Processing of L2 Collocations', *Journal of Second Language Studies*, 1: 60–78. <https://doi.org/10.1075/jsls.17024.yam>
- Yi, W., Lu, S., and Ma, G. (2017) 'Frequency, Contingency and Online Processing of Multiword Sequences: An Eye-Tracking Study', *Second Language Research*, 33: 519–49. <https://doi.org/10.1177/0267658317708009>
- Yi, W., Man, K., and Maie, R. (2023) 'Investigating L1 and L2 Speaker Intuitions of Phrasal Frequency and Association Strength of Multiword Sequences', *Language Learning*, 73: 266–300.
- Yi, W., and Zhong, Y. (2024) 'The Processing Advantage of Multiword Sequences: A Meta-Analysis', *Studies in Second Language Acquisition*, 46: 427–52.
- Zhang, X. (2020) 'A Bibliometric Analysis of Second Language Acquisition between 1997 and 2018', *Studies in Second Language Acquisition*, 42: 199–222.