Journal of Research in Applied Linguistics

ISSN: 2345-3303 – E-ISSN: 2588-3887 – http://rals.scu.ac.ir

Published by Shahid Chamran University of Ahvaz

Please cite this paper as follows:

Shokri, A., Khany, R., & Aliakbari, M. (2022). Two decades of research articles keywords in corpus-based studies in International Journal of Corpus Linguistics. *Journal of Research in Applied Linguistics*, *13*(1), 70-83. https://doi.org/10.22055/RALS.2022.17426

Shahid Chamran University of Ahvaz

Research Paper

Two Decades of Research Articles Keywords in Corpus-Based Studies in International Journal of Corpus Linguistics

Afsaneh Shokri[1], Reza Khany[2], & Mohammad Aliakbari[3]

[1]English Department, Ilam University, Ilam, Iran; *a.shokri2@ilam.ac.ir*

[2]Corresponding author; English Department, Ilam University, Ilam, Iran; *r.khany@ilam.ac.ir*

[3]English Department, Ilam University, Ilam, Iran; *m.aliakbari@ilam.ac.ir*

Received: 11/06/2021 Accepted: 15/01/2022

Abstract

Keywords accompanying abstracts are the metadata and topic representative features of research articles. They can enhance the retrieval, citation, and sorting of the studies in academia. Whereas keywords have been of interest to researchers in various fields of study, few, if any, studies have addressed them in a corpus-driven analysis. The present diachronic case study, then, aimed at analyzing 245 corpus-based studies to investigate their form, relevance, source, and frequency. Selected studies were published in *International Journal of Corpus Linguistics* from 1996 to 2016. Descriptive analysis of the data revealed that domain-specific keywords were the most frequent ones supporting the journal's more specific taste. High percentage of keywords with one representation across the two decades alerted the chaos in selecting appropriate keywords. Lack of criteria in selecting the most relevant keywords and asymmetrical forms were among the salient results. Findings may be useful to the researchers, authors of research articles, as well as editors and publishers.

Keywords: Corpus-Based Studies; Keyword Indexing; Research Articles Keywords; Domain-Specific Keywords

1. Introduction

Citation as the building block of the impact factor serves as a proxy in measuring the quality of journals and research articles (Sohrabi & Iraj, 2017). Moreover, easy access to an article, as well as its high citation, is a crucial goal of the author as well as the publisher. However, in the era of electronic publication, scientific databases, and data explosion, finding highly valid and relevant metadata is a "long-standing issue" (Hurt, 2010, p. 81). According to Garcia, Gattaz, and Gattaz (2019), scientific publications could be considered similar to an iceberg, as the popular articles are visible and easily accessible, located at the top, and less-known ones are in the hidden parts. However, these invisible articles may hold valuable information. So, the essential question is this: Why and how do some research articles get cited more than other articles with the same quality and even more valuable information? Being aware of the characteristics that guarantee the high retrieval and citation of an article is among the first and foremost priorities for authors and journals.

According to Gil-Leiva and Alonso-Arroyo (2007), indexing is a method for choosing those ideas that best indicate the content of documents and therefore pave the way for their appropriate storage and retrieval. In other words, indexing is the description of a document concerning its content and it is necessary for retrieval and storing of information (Gil-Leiva, 2017). In the past, indexing was carried out based on title, abstract, both, or the whole text. However, the results of recent studies (e.g., Ansari, 2005; Gil Leiva & Alonso-Arroyo, 2007; Strader, 2009) revealed that keywords, as an essential step in indexing, can also provide fruitful information about the documents.

Overall, keywords, as metadata features, form a small-scale version of the full manuscript and represent the content and purpose of the study. Further, they are important components in the process of retrieval and citation of an article. They can guide researchers toward papers in the hidden parts of publication that may not come into their considerations in the typical course of studies.

1.1. Keywords: Definition and Importance

Keyword, sometimes called index, descriptor, or subject heading (Hartley & Kostoff, 2003; Howcroft, 2007), represents the topic of the documents or research questions. Despite its name, the keyword can be a single word, a combination of words, or an alphanumerical expression (Howcroft, 2007). According to Hartley and Kostoff (2003), keywords can be chosen by authors, editors, referees, publishers, or computer programs. They are created by manual or automatic analysis of the documents (Howcroft, 2007) and derived from the titles or the body of documents reflecting their topics and contents (Hurt, 2010).

According to GburJr and Trumbo (1995), in line with titles, keywords can act as a "mini-abstract" (p. 32) in research articles. Keywords play a complementary role in retrieving the documents in search queries. They are classified into three groups: general, intermediate, and specific (Lebrun, 2007). General keywords delineate the domain/genre of the study and their differentiation power is very low. Intermediate keywords have higher power and depict the methods or the subdomains of the field. Specific keywords, on the other hand, have the highest differentiating power and are well defined for the researchers in a given journal (Lebrun, 2007).

Hartley and Kostoff (2003) mention five crucial roles for keywords: (1) enabling readers to ensure about the relevance of research article, (2) helping readers to find related articles by using proper terms, (3) providing editors and indexers to bring related materials together, (4) enabling researchers and editors to record changes in the domain in question over time, and (5) connecting specific topics and concerns to higher metalevel topics. Due to the substantial and 5influential roles of keywords in research articles and in response to the lack of inclusive literature in applied linguistics in this regard, the present study aimed at analyzing the articles in the domain of corpus linguistics (as an essential part of applied linguistics). Therefore, the main purpose of the study was to investigate keywords used in research articles to find out further information about their structures, relevance, sources, and frequency in corpus-based studies. Having considered the importance of keywords in research, a number of researchers have investigated them in different fields of study to gain further knowledge about their nature and fundamental roles in research articles

2. Literature Review

According to Gil-Leiva and Alonso-Arroyo (2007), the subject of studies on keywords can be classified into the following categories: retrieval efficiency (e.g., Voorbij, 1998), author/ editor-assigned keywords (e.g., GhurJr & Trumbo, 1995; Hartley & Kostoff, 2003), metatag keywords (Alimohammadi, 2004; Craven, 2004), automatic extraction (e.g., Gil-Leiva, 2017; Turney, 2000), and comparison of author-assigned and descriptor assigned keywords (e.g., Ansari, 2005; Strader, 2009). These studies have been carried out in different fields (e.g., information studies, library, medical sciences, and applied linguistics).

To compare the value of title keywords and subject descriptors as subject search entries, Voorbij (1998) carried out two studies. In the first study, 12 librarians of the International Library compared the subject descriptors and title keywords of 475 records in humanities and social sciences. The results showed that subject descriptors considerably enhanced 37% of records. In the second study, the librarians utilized title keywords and subject descriptors to investigate their capabilities in recalling the records. The results showed that the relative recalls for searches by title keywords and subject descriptors were 48 % and 87 %, respectively. Based on the results, Voorbij concluded that titles could not provide sufficient information for searching compared to subject descriptors that prove better results.

In their seminal study, Hartley and Kostoff (2003) investigated keywords' generation, usability, and value in research articles from various disciplines. They analyzed 230 research articles in arts, education, psychology, science, medicine, and statistics regarding the usability and value of keywords. The results revealed that there was no formal requirement, rule, or guidance for the generation and use of keywords in the disciplines in question. Hartley and Kostoff also asked 35 editors to mention the advantages and disadvantages of using keywords in the research articles. The results revealed that there was no formal requirement, rule, or guidance for the generation and use of keywords in the disciplines in question.

To answer the question of "what might we lose if subject headings were not added to bibliographic records" (p. 224), Gross and Taylor (2005) investigated the importance of keywords in the subject heading field in retrieving the records. Out of 3,397 keywords from the catalog of Winthrop University library, 227 keywords were selected as the

sample and analyzed. This finding showed that more than 35% (about one-third) of hits were lost in the absence of subject-heading searching.

Gil-Leiva and Alonso-Arroyo (2007) analyzed 640 research articles from 24 scientific journals to investigate the presence of keywords assigned by authors in the descriptors proposed by indexers. The research articles derived from four databases (i.e., CAB, ISTA, LISTA, & INSPFC) were analyzed in terms of quantitative and semantic relations between keywords and descriptors. The results showed that the keywords were crucial in the data descriptors proposed by indexers (about 46%). Their results also revealed a lack of literature about the role of keywords assigned by the authors in indexing.

In his matching study, Strader (2009) investigated the overlap and uniqueness degrees between keywords assigned by authors and cataloguer-assigned Library of Congress Subject Heading (LCSH). He applied six categories of matching to determine the match between the two methods: exact-match, all-present (but not in exact order), partial-match (keywords covered by two LCSH), and variant (e.g., variant of cross-reference, variant is abbreviation), and no-match (not present). Strader (2009) analyzed 285 electronic theses and dissertations of doctoral candidates at Ohio State University. He found a strong relationship between the two methods and concluded that author-assigned and cataloguer-assigned terms could increase the discoverability of theses and dissertations.

Hurt (2010) explored the differences between keywords assigned by authors and automatically generated keywords. He examined 22 research articles drawn from *Journal of Applied Polymer Science*. The results showed that there was no statistically significant difference between the two methods. Based on his findings, Hurt concluded that author-assigned keywords are not necessarily better choices for selecting suitable keywords.

Schwing, McCutcheon, and Maurer (2012) replicated Strader's (2009) study to investigate the overlap between author-assigned keywords in electronic theses and dissertations (ETD) bibliographic records in terms of uniqueness, matching, and complementary natures. They analyzed 95 bibliographic records. The results were in line with Strader's (2009) findings and supported his conclusions about the crucial roles of author-assigned and cataloguer-assigned keywords.

In recent years, researchers have also investigated keywords in fields such as applied linguistics, library science, and medical education. In the applied linguistics domain, Babaii and Taase (2013) investigated the characteristics of the keywords assigned by authors in 200 research articles of applied linguistics journals. They explored the keywords' domain, degree of specificity, and their relationship to titles. They also asked the authors with publishing experience in applied linguistics to cite their strategies for selecting keywords. The results showed that there was a considerable match between titles and keywords in terms of field-specific keywords. The viewpoints collected from the academics justified that the reasons mentioned by the respondents were not mainly related to the primary roles of the keywords. Babaii and Taase stressed that authors' viewpoints about the strategies of keywords selection could illuminate the complex nature of keywords in the applied linguistics domain.

In the library science domain, Akbari, Rezaei, and Beheshti (2018) studied the common mistakes in the author-assigned keywords in medical education. They investigated 10,965 English and Persian keywords extracted from 3,194 articles of 13 medical journals. They found that the most common mistake was the problem of finding accurate equivalents for Persian and English keywords. Akbari et al. stressed that the authors and editors should consider the important rule of keywords in data storage and retrieval efficiency.

In medical education, Mazaheri, Mostafavi, and Geraie (2019) used author keywords and index terms to compare the intellectual structure knowledge of articles in *International Journal of Preventive Medicine* (IJPM) with medical subject headings (MeSH). Using coword techniques, they analyzed 1,104 articles published in IJPM to determine the compatibility of author keywords with index terms. The results showed that about 59% of the keywords were classified in the no-match category. They concluded that the Iranian authors of the mentioned journal did not have sufficient knowledge on selecting proper keywords concerning MeSH, and they need some training about the selection and use of keywords in medical education.

As the above studies revealed, keywords have been mainly studied in information science, library science, and medical education. Despite their importance and functionality, little attention has, up to the present, been paid to keywords

and their fundamental roles in applied linguistics, in general, and corpus linguistics, in particular. It seems that keywords are undervalued and ignored by researchers. Thus, further in this regard is warranted.

2.1. Statement of the Problem

Keywords "as a gateway to the text" (Garcia et al., 2019, p. 5) can significantly increase the visibility (Kalwij & Smit, 2013), efficiency (GburJr & Trumbo, 1995), impact and discoverability (Strader, 2009), the frequency of citation (Chicoo, 2017), electronic information retrieval (Hartley & Kostoff, 2003), and the chance to advertise (Kalwij & Smit, 2013) the research articles. They can also remove a large body of unwanted data.

Based on the earlier review done on keywords functions in research articles (e.g., GburJr & Trumbo, 1995; Hartley & Kostoff, 2003), it can be concluded that the keywords serve a two-fold purpose in two points: Keywords can help authors find out the most relevant and proper article(s) before they write their research article. Second, when the authors write their articles, keywords can facilitate the efficient retrieval of their articles by interested readers and researchers. However, despite their pivotal roles, keywords are generally underestimated by the authors and even the editors of journals. It is more problematic in applied linguistics journals than other journals because they only provide general roles and keywords guidelines. *International Journal of Corpus Linguistics* (IJCL), for example, describes the instructions on keywords as follows: "Please provide up to five keywords; separated by commas, which are not too general—e.g., corpus linguistics, corpus, corpora, etc. are too general for IJCL" (*International Journal of Corpus Linguistics, Format & Style Guideline,* 2020, p. 2). Besides, it seems that most researchers in applied linguistics deemphasize the influential roles of keywords. For instance, in an online questionnaire, about 70% of the respondents (well-known academics in applied linguistics) were not acquainted with the functions of keywords (Babaii & Taase, 2013).

Despite the aforementioned research, keywords have not been addressed adequately in applied linguistics studies, in general, and corpus-based studies, in particular. They still lack a comprehensive investigation on some fundamental features of keywords. First, the structures of the keywords and the criteria in selecting the most relevant keywords have not been addressed adequately. Second, the source of the keywords (e.g., the domain they were derived from) is another area of research that has received relatively little attention among researchers. Further, few studies, if any, have examined the most frequent words and phrases as well as their numbers in corpus-based studies in order to find out a trend in corpus-based studies diachronically. Accordingly, this study was an attempt to answer the following questions:

1. What are the common formats and structures of keywords in IJCL corpus-based studies?
2. To what extent are keywords relevant to the title and abstract in IJCL corpus-based studies?
3. What are the primary sources of keywords in IJCL corpus-based studies?
4. What are the most frequent words or phrases in the keywords of IJCL corpus-based studies?

3. Methodology

This study investigated the overall structure of keywords in corpus-based studies in IJCL. In this section, the corpus, the design, the main stages in compiling the corpus and the procedure are explained.

3.1. Corpus

The corpus consisted of all the corpus-based studies (245 research articles), published in IJCL over the past two decades (1996-2016). The rationales behind selecting IJCL were as follows: First, IJCL is a quarterly peer-reviewed journal that publishes research articles, short notes, and book reviews on corpus linguistics, mainly in the field of applied linguistics. Second, IJCL has been published since 1996, so, as a leading journal, it can provide a dynamic view of the trends of corpus-based studies from the last two decades. In the present study, only research articles that showed title, abstract, and keywords were chosen.

3.2. Data Collection

This study adopted a corpus-based approach to scrutinize the nature and structure of keywords in corpus-based studies. A diachronic and descriptive analysis attempted to understand the trend of keywords used in corpus-based studies and attain an overall picture of their form, relevance, source, and frequency in the last two decades (i.e., 1986-2016). To answer the research questions, first, the e-versions of the research articles published in IJCL were downloaded, collected, and saved in 20 files according to the year of publication. Because we wanted to get an idea of the original and regular articles, special issues were excluded. Reviews, forums, editorials, letters to the editors, short notes, and research articles without keywords or abstracts were also excluded. A total of 245 research articles were collected, starting with the first issue of the journal in January 1996 (to 2016). Then, the title, abstract, and keywords of each research article were saved in four files specifying four 5-year periods: 1996-2001, 2001-2006, 2006-2011, and 2011-2016. Each article was saved with a name specifying the assigned number, the year of publication, and the journal.

3.3. Procedure

The research articles were analyzed in terms of form, relevance, sources, and frequency across four 5-year periods.

3.3.1. Form of Keywords

The form category consisted of the following subcategories:

1. Number of the words in the keywords section in the entire 20 years addressed in the study and in each four 5-year periods, for example: In the example below, the number of words was 8:

   IJCL.138.19.1.2014:

   Keywords: educational history, lexical difficulty, lexical diversity, reading textbooks

2. The number of the keywords in the entire 20 years addressed in the study as well as in each four 5-year periods; for instance, in the above example, the number of keywords was 4.

3. The structure of the keywords in the entire 20 years as well as in each four 5-year periods (e, g., word, phrase)

3.3.2 Relevance of Keywords

In the relevance category, the overlap between keywords and titles and keywords and abstracts were analyzed. In his study, Strader (2009) utilized match categories to codify the overlap between author-assigned and cataloger-assigned keywords (see section 3, paragraph 3). In this study, 60 research articles were piloted (three randomly selected research articles from each year), and the following categories were selected in line with the purposes of the study:

1. **Exact-match**: All the terms in the keyword match title or abstract, term-for-term.

   For example:

   Article.62.2001.1. IJCL.

   Title: Functions of Actually in a Corpus of Intercultural Conversations

   Keywords: discourse analysis, discourse marker, actually, Hong Kong, corpus, intercultural communication, naturally occurring conversation

   In the above example, *actually* and *corpus* were classified in the exact-match category

2. **Partial-match**: At least one of the terms in the keyword match across title or abstract. In the exact-match example, *intercultural communication* was classified as in the partial-match category.

3. **All-present:** All the terms in the keyword match across the title or abstract, but not in the exact order.

4. **Variant-match:** They are variations at the base of keywords (e.g., singular/plural, abbreviations, etc.).

5. **No-match**: As the name suggests, no terms in the keywords match title or abstract; *discourse analysis* in the above example was classified in the no-match category

3.3.3 Source of Keywords

In the source category, the keywords of the research articles were scrutinized to discover the domain they were derived from. At first, the generality and domain-specificity of keywords were determined following Babaii and Taase's (2013) procedure. In their study, Babaii and Taase consulted the last version of *Longman Dictionary of Language Teaching and Applied Linguistics* (Richards & Schmidt, 2002) and *Encyclopedic Dictionary of Applied Linguistics* (Johnson & Johnson, 1998). In the present study, we utilized *Longman Dictionary of Language Teaching and Applied Linguistics* (5th ed.; Richards & Schmidt, 2013), *Encyclopedic Dictionary of Applied Linguistics* (Johnson & Johnson, 1998), and *Glossary of Corpus Linguistics* (Baker, 2006).

Then, the specific keywords were examined to determine their sources. To identify the different types of specific keywords, we independently analyzed 60 research articles (three randomly selected research articles from each year) containing 56 keywords. After completing the draft, the results were compared, and some items and definitions were revised, leading to the final version of the source categories. The results of the pilot study indicated that about 60% of the analyzed keywords dealt with the specific features of language that were the focus of the research studies. Twenty percent of the keywords belonged to the main language components, and the remaining 20% dealt with different types of corpora and diverse categories (10% & 10%, respectively).

The final version of the sources consisted of:

1. **Target feature keywords**: This item investigated those target features that were the subjects of the study (e.g., collocations, supplementive clauses).
2. **General feature keywords**: They referred to the keywords related to the main component of language in general (e.g., syntax, semantics).
3. **Corpus features keywords**: Two types of corpora were found in the present study. Type 1 consisted of general types of corpora (general corpora, specific corpora, raw corpora, annotated corpora, diachronic or synchronic corpora, monolingual and parallel corpora, statistic and dynamic corpora, spoken or written corpora, learner corpora). Type 2 included the types of corpora analyzed by the author/s of the research studies (e.g., research articles, English dictionary).
4. **Other:** Categories such as analytic techniques keywords (e.g., tagging, annotation), participant keywords (e.g., Chinese learners, advanced British learners), tool and software keywords, context keywords (e.g., high school, university), statistical analysis keywords (e.g., chi-square, factor analysis), and language.

3.3.4 Frequency of Keywords

In the last part, the most frequent nouns and phrases in the keywords were calculated using Lancsbox, which is Lancaster University corpus toolbox. Lancsbox is a new generation software package to analyze language data and different kinds of corpora. For the purpose of the present study, its N-gram tool was utilized.

Three researchers and two trained raters (Ph.D. students in applied linguistics) coded the 60 randomly selected research articles (i.e., three articles from each year) to estimate the reliability of the classifications (i.e., form, relevance, & source). The results revealed that the overall percentages were high for form, relevance, and sources (97%, 88%, & 84%, respectively).

4. Results

A total of 245 corpus-based research articles were analyzed across four 5-year periods (1996-2001, 2001-2006, 2006-2011, and 2011-2016). Table 1 presents the frequency and percentage of research articles in each time period:

Table 1. *Frequency and Percentage of Research Articles Across Four Time Periods*

Journal	Time Period				Total
IJCL	1996-2001	2001-2006	2006-2011	2011-2016	
F (%)	46 (19)	67 (27)	63 (26)	69 (28)	245

Note: IJCL = International Journal of Corpus Linguistics

It is worth noting that IJCL was published biannually during 1996-2004. Then, it was published 3 times a year from 2004 onward (except special issues). This can explain the lower frequency of corpus-based studies in the first time period compared with the subsequent time periods. In this section, the findings of the study regarding the aforementioned research questions are reported.

4.1. Form of Keywords

The number of words, keywords, and the structure of the keywords were analyzed, and their means were calculated to answer the first question of the study (see Table 2):

Table 2. *Frequency and Mean of Words, Keywords, Noun, and Noun Phrases Across Time Periods*

Time Periods	Words *F* ($\tilde{x}$)	Keywords *F* ($\tilde{x}$)	Structure	
			Single word *F* ($\tilde{x}$)	Phrase *F* ($\tilde{x}$)
1996-2001	368 (8)	208 (4.5)	88 (1.91)	120 (2.60)
2001-2006	612 (9.1)	348 (5.19)	145 (2.16)	203 (3.02)
2006-2011	696 (11)	382 (6.06)	140 (2.22)	242 (3.42)
2011-2016	536 (7.76)	333 (4.82)	145 (2.10)	188 (2.72)
Total	2,212 (9)	1,271 (5)	518 (2)	753 (3)

Note: F = frequency, $\tilde{x}$ = mean

The number of the keywords over the past 20 years and in each time period was near the standard defined by the journal (i.e., five keywords). The structure of the keywords consisted of single words and phrases with two or more words. In the first time period (1996-2001), about 42% of the keywords comprised single words and 58% included phrases. The minimum/maximum numbers of the keywords were 2 and 11, respectively. Across the second period (2001-2006), the percentages of the single words and phrases as well as the minimum/maximum number of the keywords remained consistent. In the third time period, the single words constituted 38% of the keywords, a slightly lower percentage than the former time periods. The minimum/maximum numbers of the keywords were 3 and 11, respectively. Finally, in the last time period (2011-2016), about 44% of the keywords involved single words and the phrases amounted to 56% of the keywords. Generally speaking, the number of phrases was slightly higher than the words in the past two decades and in each time period. A detailed analysis of the corpus of the study revealed that the single words mainly were from the nouns category (more than 98%), and the noun phrases were the most frequent types of phrases in each time period (83%, 78%, 75%, and 95%, respectively).

4.2. Relevance of Keywords

Concerning the second question, the overlap between the keywords and the titles and the keywords and the abstracts were scrutinized in terms of exact-match, partial-match, all-present, variant, and no-match. Tables 3 and 4 present the findings:

Table 3. *Frequency and Percentage of Different Types of Matching in Keywords and Titles*

Time Periods	Exact-Match *F* (%)	Partial-Match *F* (%)	All-Present *F* (%)	Variant *F* (%)	No-Match *F* (%)	Total *F* (%)
1996-2001	40 (19)	29 (14)	0	8 (4)	131 (63)	208 (17)
2001-2006	179 (51)	56 (16)	3 (1)	17 (5)	93 (27)	348 (27)
2006-2011	92 (24)	58 (15)	1 (0)	19 (6)	212 (55)	382 (30)
2011-2016	103 (31)	52 (16)	0	14 (4)	164 (49)	333 (26)
Total	414 (33)	195 (15)	4 (1)	58 (4)	600 (47)	1,271

Note: F = frequency, %: percent

As Table 3 presents, except for the second period (2001-2006), the most frequent category was no-match, followed by exact-match in the three other periods and all over the two decades. This result is consistent with Strader's (2009) and Schwing et al.'s (2012) results. The exact-match category was more frequent in both studies (55% & 48%, respectively). However, the findings showed that the exact-match category had increased steadily over time (except for the second period), whereas the no-match category had decreased. Moreover, the results revealed that the exact-match category was more frequent than the no-match category in the second period:

Table 4. *Frequency and Percentage of Different Types of Matching in Keywords and Abstracts*

Time periods	Exact-Match *F* (%)	Partial-Match *F* (%)	All-Present *F* (%)	Variant *F* (%)	No-Match *F* (%)	Total *F* (%)
1996-2001	105 (50)	24 (12)	0	3 (1)	76 (36)	208
2001-2006	224 (64)	63 (18)	0	3 (1)	58 (17)	348
2006-2011	243 (63)	78 (20)	0	5 (2)	56 (15)	382
2011-2016	252 (75)	42 (13)	0	3 (1)	36 (11)	333
Total	824 (65)	207 (16)	0	14 (1)	226 (18)	1,271

Note: F = frequency, % = percent, All = all present

The results in Table 4 reveal that in the abstract and keywords overlap, the most frequent type of matching is the exact-match category in all of the four time-periods and across the two decades. The partial-match and no-match categories had almost the same percentage across the past 20 years, but their frequencies varied in different time periods.

4.3. Sources of Keywords

To answer the third question, the frequencies of the general and domain-specific keywords were calculated (see Table 5):

Table 5. *Frequency and Percentage of General and Domain-Specific Keywords*

Time Periods	General	Domain-Specific	Total
	F (%)	*F* (%)	*F* (%)
1996-2001	7 (3)	201 (93)	208
2001-2006	8 (2)	340 (98)	348
2006-2011	7 (2)	375 (98)	382
2011-2016	12 (4)	321 (96)	333
Total	34 (3)	1,237 (97)	1271

Note: F = Frequency, % = percent

The general keywords were rare in corpus-based studies in all four periods and across the two decades of research addressed in the present study. The domain-specific keywords had a very high frequency over the two decades (94%) and in all time periods. The high frequency of the domain-specific keywords is consistent with Babaii and Taase (2013). They found that more than 80% of the keywords were specialized keywords in their corpus. Then, the sources for the keywords in domain-specific keywords were scrutinized (see Table 6):

Table 6. *Sources of Domain-Specific Keywords*

	Time Periods				
	1996-2001	2001-2006	2006-2011	2011-2016	Total
Keywords Sources	*F* (%)	*F* (%)	*F* (%)	*F* (%)	*F* (%)
Target Features	105 (50)	199 (57)	214 (56)	162 (49)	680 (55)
General Features	34 (17)	53 (15)	68 (17)	69 (21)	224 (18)

Corpus Features					
Type1	12 (6)	14 (4)	14 (4)	6 (2)	131 (11)
Type2	8 (4)	21 (6)	24 (6)	32 (10)	
Other	42 (21)	53 (26)	55 (27)	52 (26)	202 (16)
Analytic Techniques	21 (10)	16 (5)	21 (6)	11 (4)	69 (6)
Context	4 (2)	23 (7)	13 (3)	15 (5)	55 (3)
Statistical Analysis	10 (5)	8 (2)	12 (3)	15 (5)	45 (4)
Tools & Software	6 (3)	3 (1)	4 (1)	5 (2)	18 (1)
Language & Nationality	1 (0)	2 (1)	3 (1)	2 (1)	8 (1)
Participant	0	1 (0.2)	2 (1)	3 (1)	6 (0.5)
Design	0	0	0	1 (0.3)	1 (0)
Total	201 (16)	340 (28)	375 (30)	321 (26)	1,237

Note: F = frequency, % = percent

Concerning the type of specific keywords, about half of the keywords fell into the Target Features category in all of the four periods of time and over the two decades. It was followed by General Features keywords. The third category with high frequency was the Corpus Features category. As can be seen, the Type 2 subcategory was about double that of the Type 1 subcategory. Finally, about 16% of the keywords belonged to the Other category.

4.4. Frequency of Keywords

To answer the last question concerning the most frequent words and phrases in the keywords, the N-gram tool of Lancsbox was used (see Table 7):

Table 7. *Most Frequent Nouns and Phrases in the Keywords*

Time Periods	Nouns (*F*)	Phrases (*F*)
1996-2001	corpus (19), corpora (9), language (8), English (7), analysis (7), tagging (5) parallel (5), parsing (4), vocabulary (4), alignment (4)	parallel corpora (3), corpus analysis (2), spoken English (2), corpus linguistics (2), English vocabulary (2), lexical semantic (2), British national corpus (2), spoken corpus (2), national corpus (2)
2001-2006	corpus (26), corpora (11), linguistics (11), English (8), analysis (8), collocation (7), translation (7), of (6), language (6), discourse (6)	corpus linguistics (5), language corpora (4), spoken corpus (3), spoken language (3), corpus idioms (3), Hong Kong (3), translation studies (2), semantic prosody (2), part-of-speech tagging (2), African language (2)
2006-2011	corpus (20), corpora (15), language (14), analysis (12), frequency (9), lexical (9), variation (8), English (7), tagging (7), semantic (7)	syntactic complexity (4), comparable corpora (3) bilingual corpora (2), spoken corpus (2), learner corpus (2), POS tagging (2), semantic prosody (2), register variation (2) Chinese unknown words (2), corpus linguistics (2)
2011-2016	corpus (17), discourse (14), analysis (13) English (12), language (10), writing (10), lexical (7), research (6), learner (6), academic (6)	discourse analysis (4), academic discourse (3), research article (3), lexical priming (2), corpus analysis (2), sense disambiguation (2), cluster analysis (2), learner corpus (2), lexical diversity (2), lexical bundle (2), corpus linguistics (2)
total	corpus (83), analysis (40), language (38), corpora (38), English (34), discourse (25), linguistics (24), lexical (23), of (22), translation (18)	corpus linguistics (10), discourse analysis (7), semantic prosody (5), academic writing (5), language corpora (5), parallel corpora (5), spoken corpora (5), comparable corpora (4), academic discourse (4)

Note: F = frequency

As Table 7 presents, the most frequent word in each of the four periods and the two decades was *corpus*. The six other words among the top 10 words in each four-year period and the two decades were: *corpora*, *English*, *analysis*, *linguistics*, *language*, and *discourse*. As Table 7 indicates, the most frequent nouns were language-features keywords. A detailed analysis reveals that the number of the words with one or two frequencies was common among the keywords.

For example, in the four-time periods, 50%, 43%, 57%, and 53% of the words repeated only once or twice, respectively. Over the 20 years, about 57% of the keywords appeared only once or twice and 47% (600 keywords) only had 1 representation. Regarding the phrases, no particular pattern was observed. In the four-time periods, 87%, 94%, 98%, and 97% of the keywords had only 1 or 2 representations. Over the two decades, more than 84% of the keywords appeared only once or twice.

5. Discussion

The purpose of this study was to investigate the keywords in IJCL corpus-based studies in terms of form, relevance, sources, and frequency. This section is divided into four parts: first, the form of keywords and their structure are discussed. This is followed by the discussion of the relevance and sources of keywords, respectively. The final section is allocated to the most frequent keywords.

5.1. Form of Keywords

As Table 2 shows, the average number of the keywords in the corpus was about 5 across the past 20 years and in each of the four 5-year periods addressed in the present study (except for the second period that was 6). It follows that the authors of the research articles respected the author's guidelines provided by IJCL about the maximum number of keywords (i.e., 5). Though the journal and publisher did not explain why they think that the maximum of 5 keywords was appropriate, 5 keywords are likely to represent the essence of the article sufficiently. Reviewing the recommendations of 24 journals on keywords, Gil-Leiva and Alonso-Arroyo (2007) concluded that 3-6 keywords can cover the cardinal concerns in the research articles. Due to the rough requirements for manuscript submission, the authors may not play a serious role in determining the number of keywords. However, the authors' decisions about their numbers may vary. Some authors prefer using as many keywords as possible to promote the chance of visibility of their article in the retrieving process.

Concerning the structure of keywords, all the keywords were either single words or single phrases with two or more words. However, the phrases were more frequent than the single words in all of the time periods and across the two decades. This result is in line with the guidelines provided by the scholars in this domain (e.g., GburJr & Trumbo, 1995). As the result represents, the most frequent types of single words and phrases were nouns and noun phrases, respectively. According to Mahdabi, Andersson, Keikha, and Crestani (2012), noun phrases can improve retrieval effectiveness in search queries up to 20%. So, it can be claimed that the authors of IJCL chose the proper type of phrases deliberately or accidentally. The most common structures were NN + NN(S) followed by ADJ + NN(S). A close examination also showed that the keywords with two words were more frequent in the past two decades of analysis ($\tilde{x}$:74%). Chicoo (2017) suggested that the authors should avoid the too short and too long keywords. In his opinion, too short keywords bring about specific and limited searches, whereas too long keywords may result in broad searches. In their study, GburJr and Trumbo (1995) provided similar recommendations.

5.2. Relevance of Keywords

5.2.1. Title and Keywords Overlap

According to Alcaraz and Méndez (2016), a title is the first part of the research article that "captures the readers' eyes" (p. 133). Rath (2010) holds that the title summarizes the research article evoking the readers to download it. As "a kind of architectural design" (Kosharnaya, Chumak-Zhun, Plotnikova, Maltseva, & Boldyreva, 2019, p. 406), titles can promote the precision of the retrieval and the chance to be read by the interested readers. However, because the wording of titles is limited (Schwing et al., 2012), keywords can complement them to play their roles better. GburJr and Trumbo (1995) believe that title and keywords can act as a "mini-abstract" (p. 32) in research articles. Concerning the similarity of keywords and title, two viewpoints exist: Researchers such as Mack (2012) believe that keywords should duplicate the words of the title to increase the chance of retrieving. The second group of scholars insists that keywords should not include the title's words (e.g., Chicoo, 2017; GburJr& Trumbo, 1995; Kalwij & Smit, 2013). Hartley and Kostoff (2003) point out that because the title's words "picked up anyway" (p. 437), researchers should not waste the keywords space by repeating title. In other words, the role of keywords is to complement the role of the title (GburJr& Trumbo, 1995) in retrieving the related articles.

As Table 3 shows, the no-match category was the most frequent category, followed by the exact-match category. Strader (2009) mentions two reasons for the high percentage of the no-match category. Firstly, he believes that the high frequency of the no-match category can be explained in terms of the inherently limited nature of titles in wording. As a result, the titles cannot cover all of the words in the keywords section, so the frequency of matching decreases. Secondly, he relates this result to the use of different words and phrases for similar concepts. These phrases and concepts can be discussed in terms of semantic relations and need more investigation. However, as mentioned previously, the matching category is not superior to the no-match category. Schwing et al. (2012) stress that, sometimes, the no-match category can boost the uniqueness of nature and promote the discoverability and impact of the article and the journal. Very interesting points in Table 3 were the decreasing trend in the exact-match category as well as an increasing trend in the no-match category over the two decades of analysis (with the exception of the second-time period). This can be explained in terms of the increasing use of domain-specific keywords in corpus-based studies in the recent 20 years. According to Babaii and Taase (2013), it seems that "keyword-title match is mostly due to the presence of specialized terminologies rather than general terms" (p. 7). Accordingly, the high frequency of the exact-match category in the second-time period might be due to the higher rate of domain-specific keywords used during this period.

5.2.2. Abstract and Keywords Overlap

Regarding the abstract-keywords overlap, the result is exactly opposite to the title- keywords findings. The percentage of the exact-match category was more than the no-match category. Following Strader's (2009) explanation about the title, we conclude that abstract is not as limited as the title in terms of wording and consequently can cover more similar words. Schwing et al. (2012) pointed out that the title, the abstract, and the keywords sections are all written by the same authors who share the same vocabularies and grammatical structures, so the similarity is inevitable.

5.3. Source of Keywords

Concerning specific and general keywords, it seems that one of the crucial differences between a novice researcher and an expert is that the experienced researcher utilizes the domain-specific terms in his or her queries for needed documents. Hölscher and Strube (2000) point out that the domain-specific terms can help the researchers quickly access the related information and, consequently, optimize the discoverability of the research studies. It can be concluded that specific keywords can maximize the differentiation, resulting in most related documents (Lebrun, 2007). Therefore, researchers should be equipped with the specific knowledge of their domain as an important strategy for efficient and successful information search.

Regarding domain-specific keywords, Hartley and Kostoff (2003) stress that one of the salient steps in choosing the proper keywords is recognizing the central concept and the purpose of the study. According to Lebrun (2007), general keywords can describe the domain or type of a study, but its differentiating power is very low and may retrieve a long list of unrelated papers (GburJr & Trumbo, 1995). Our findings show that the authors of IJCL research articles followed the journal instructions for the authors and did not use general feature keywords (e.g., corpus and corpus linguistics).

Concerning the type of corpora, Lebrun (2007) mentions that keywords can be different from one field to another and from one discipline to another. In corpus-based studies, corpus and type of corpus have a very fundamental role, so it seems eligible and logical to use them as keywords. Future research can concentrate on the investigation of the genre-related and filed/disciplines keywords.

5.4. Frequency of Keywords

More than 57% of the words and 84% of the phrases in the keywords subsection appeared only once or twice all over the past 20 years. This fluctuation raises doubts over the existence of any criterion or category for selecting keywords in IJCL corpus-based studies. According to Hurt (2010), some journals utilize a predetermined group of keywords for authors. He believes that a predetermined set of keywords increases the keyness values. Hartley and Kostoff (2003) suggest that providing a list of categories and subcategories can lead to selecting proper keywords. They proposed to create a database such as MeSH, which has a predetermined list provided by the publisher or journal from which the authors can select the proper and related keywords.

6. Summary and Conclusion

The present study aimed at investigating the keywords in IJCL corpus-based studies in terms of form, relevance, source, and frequency. All in all, the following conclusions can be highlighted: First, concerning the form, the authors of IJCL followed the author's guidelines proposed by the journal about the maximum number of keywords (i.e., 5). Besides, the results revealed that single nouns and noun phrases were the most commonly used structure. The most common structures were NN + NN(S), followed by ADJ + NN(S).

Second, regarding relevance, the titles and keywords of the research studies were scrutinized. The results showed that the no-match category was the most frequent category followed by the exact-match category. In the case of keywords and abstract overlap, however, the findings were contrary to the results of titles and keywords overlap, meaning that the frequency of the exact-match category was more than the no-match category. The difference can be explained in terms of the nature and purposes of titles and abstracts in research articles.

Third, considering the source of keywords, the findings manifested that the specific language-related keywords were the most frequent, followed by general language keywords and the types of corpus. We conclude that the authors were in line with the guidelines provided by scholars about the privilege of specific keywords to the general keywords.

Finally, considering the frequency, the high percentages of the keywords with one representation alerted the chaos in selecting keywords in applied linguistics, in general, and corpus-based studies, in particular. Moreover, the most frequent words in keywords were *corpus*, *analysis*, *language*, *English*, *corpora*, *discourse*, and *linguistics* over the past 20 years. Concerning the phrases, no pattern was observed.

The results of this study are invaluable for researchers, authors of research articles, editors, and publishers. Keywords have a dual role for researchers and authors: (1) as a tool for finding and retrieving the most relevant articles and (2) as a tool for helping interested readers to find their articles. Further, the findings can assist researchers and authors to be more informed about keywords, their forms and structures, their relevance to titles and abstracts, their sources, their frequency, and their crucial functions in research articles. The findings can also help editors and publishers provide more comprehensive author's guidelines about keywords. They should be aware of the different types of keywords in different kinds of studies and encourage authors to be familiar with the functions of keywords and exploit the most appropriate keywords in their research studies.

In conducting the present study, some limitations were tolerated: This research was a case study limited to one journal (i.e., IJCL). Accordingly, the findings cannot be generalized to other corpus-based studies, in general, and corpus-based studies in applied linguistics, in particular. Further, the corpus was relatively limited. A larger corpus derived from corpus-based studies from different journals might bring about more exact results.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

All the authors contributed to the study's conception and design. The first author wrote the first draft of the manuscript, and all the authors commented on the previous versions of the manuscript. All the authors read and approved the final manuscript.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

References

Akbari, E., Rezaei, R., & Beheshti, Z. (2018). Common mistakes in English keywords of articles in the field of medical sciences education. *Journal of Mazandaran University of Medical Sciences, 28*(165), 150-158. http://jmums.mazums.ac.ir/article-1-10954-en.html

Alcaraz, M. Á., & Méndez, D. I. (2016). When astrophysics meets lay and specialized audiences: Titles in popular and scientific papers. *Language & Communication, 3*(2), 133-146. http://hdl.handle.net/10045/60150

Alimohammadi, D. (2004). Measurement of the presence of keywords and description metatags on a selected number of Iranian Web sites. *Online Information Review, 28*(3), 220-223. https://doi.org/10.1108/14684520410543661

Ansari, M. (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review, 54*(7), 410-414. https://doi.org/10.1108/00242530510611901

Babaii, E., & Taase, Y. (2013). Author-assigned keywords in research articles: where do they come from? *Iranian Journal of Applied Linguistics, 16*(2), 1-19. URL: http://ijal.khu.ac.ir/article-1-1786-fa.html

Baker, P. (2006). *Glossary of corpus linguistics*. Edinburgh University Press.

Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox v. 4.x [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining, 10*(35), 1-17. https://doi.org/10.1186/s13040-017-0155-3

Craven, T. C. (2004). Variations in use of meta tag keywords by web pages in different languages. *Journal of Information Science, 30*(3), 268-279. https://doi.org/10.1177/0165551504042805

Formatting and Style Guidelines (2020). *International Journal of Corpus Linguistics (IJCL).* London: John Benjamins Publishing Company.

Garcia, D. C. F., Gattaz, C. C., & Gattaz, N. C. (2019). The relevance of title, abstract and keywords for scientific paper writing. *Revista de Administração Contemporânea, 23*(3), 1-9. https://doi.org/10.1590/1982-7849rac2019190178

GburJr, E. E., & Trumbo, B. E. (1995). Key words and phrases—The key to scholarly visibility and efficiency in an information explosion. *The American Statistician, 49*(1), 29-33. https://doi.org/10.1080/00031305.1995.10476108

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology, 58*(8), 1175-1187. https://doi.org/10.1002/asi.20595

Gil-Leiva, I. (2017). SISA-automatic indexing system for scientific articles: Experiments with location heuristics rules vs. TF-IDF rules. *Knowledge Organization, 44*(3), 139-162. https://doi.org/10.5771/0943-7444-2017-3-139

Gross, T., & Taylor, A.G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries, 66*(3), 212-30. https://doi.org/10.5860/crl.66.3.212

Hartley, J., & Kostoff, R. N. (2003). How useful are keywords' in scientific journals? *Journal of Information Science, 29*(5), 433-438. https://doi.org/10.1177/01655515030295008

Holster, C., & Strobe, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks, 33*(6), 337-346. https://doi.org/10.1016/S1389-1286(00)00031-1

Hopcroft, G. (2007). A beginner's guide to metadata and keywords. *Editors' Bulletin, 3*(3), 75-77. https://doi.org/10.1080/17521740701788437

Hurt, C. D. (2010). Automatically generated keywords: A comparison to author-generated keywords in the sciences. *Journal of Information and Organizational Sciences, 34*(1), 81-88.

Johnson, K., & Johnson, H. (Eds.). (1998). *Encyclopedic dictionary of applied linguistics*. Oxford: Blackwell Publishing.

Kalwij, J. M., & Smit, C. (2013). How authors can maximize the chance of manuscript acceptance and article visibility. *Learned Publishing, 26*(1), 28-31. https://doi.org/10.1087/20130106

Kosharnaya, S. I., Chumak-Zhun, I. I., Plotnikova, L. Y., Maltseva, G. M., & Boldyreva, S. (2019). *The title of a literary text as a discursive phenomenon.* Proceedings of the 6th International Conference on Applied Linguistics Issues (ALI 2019), Saint Petersburg, Russia. https://doi.org/10.22055/rals.2019.14711

Lebrun, J. L. (2007). *Scientific writing 2.0: A reader and writer's guide*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Mack, C. A. (2012). How to write a good scientific paper: Title, abstract, and keywords. *Journal of Micro/Nanolithography, MEMS, and MOEMS, 11*(2), 020101.1-4. https://doi.org/10.1117/1.JMM.11.2.020101

Mahdabi, P., Andersson, L., Keikha, M., & Crestani, F. (2012). *Automatic refinement of patent queries using concept importance predictors*. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, USA.

Mazaheri, E., Mostafavi, I., & Geraie, E. (2019). Comparison of intellectual structure of knowledge in international journal of preventive medicine with MeSH: A coword analysis. *International Journal of Preventive Medicine, 10*(201), 1-5.

Rath, A. (2010). Dual function of first position nominal groups in research article titles: Describing methods and structuring summary. *Journal of Research in Applied Linguistics*, *1*(2), 5-23.

Richards, J. C., & Schmidt, R. W. (2013). Longman dictionary of language teaching and applied linguistics. London: Routledge.

Schwing, T., McCutcheon, S., & Maurer, M. B. (2012). Uniqueness matters: Author-supplied keywords and LCSH in the library catalog. *Cataloging & Classification Quarterly, 50*(8), 903-928. https://doi.org/10.1080/01639374.2012.703164

Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics, 110*(1), 243-251. https://doi.org/10.1007/s11192-016-2161-5

Strader, C. R. (2009). Author-assigned keywords vs. Library of Congress subject headings. *Library Sources & Technical Services, 53*(4), 243-250. https://doi.org/10.5860/lrts.53n4.243

Turney, P. D. (2000). Learning algorithms for key phrase extraction. *Information Retrieval, 2*(4), 303-336. https://doi.org/10.1023/A:1009976227802

Voorbij, H. J. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of documentation, 54*(4), 466-476. https://doi.org/10.1108/EUM0000000007178