Linguistic Research 34(3), 427-457
DOI: 10.17250/khisli.34.3.201712.008

Corpus linguistics research trends from 1997 to 2016: A co-citation analysis

Hyejin Park*[a] · Daehyeon Nam[b]**
(University at Albany, SUNY[a] · Ulsan National Institute of Science and Technology[b])

Park, Hyejin and Daehyeon Nam. 2017. Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. *Linguistic Research* 34(3), 427-457. Corpus linguistics is one of the fastest growing areas of linguistics because of its interface with neighboring academic disciplines and the data-processing capability of a large amount of empirical linguistic data. This study reviews research trends from the last two decades within the corpus linguistics fields. Specifically, the study applied systematic citation analysis procedures to summarize and identify the salient research themes and publications from citation-reference data of peer-reviewed research articles published and indexed in the Web of Science (WoS) between 1997 and 2016. The co-citation analysis of 5,600 research articles and their 172,352 references indicated that, over the four time spans of five years, the corpus linguistics research articles have cited works ranging from general linguistics journal titles to specialized journal titles and individual books. In terms of the research themes of corpus linguistics, the topics of the linguistics research have rapidly changed over the time spans. More recently, the development of web-based large monitor corpora and corpus analysis software has contributed significantly to the dynamic and productive interaction of research in the discipline. This may indicate the evolving and juvenile nature of corpus linguistics and its possibility of growing into a multi-disciplinary field. Although there are exceptions to all of the research patterns found in the co-citation analysis, the current study also discusses the most up-to-date research trends and the future directions of corpus linguistics. **(University at Albany, SUNY · Ulsan National Institute of Science and Technology)**

Keywords corpus linguistics, co-citation analysis, citation, reference, corpora

1. Introduction

Over the last decades, corpus linguistics has been developed in an effort to empirically describe and extensively analyze language uses based on naturally

* First author
** Corresponding author

occurring linguistic data (Baker 2009). Especially with technological advancements, the study of language also allows linguists to explore large amounts of text data to answer the tendency of semantic prosody of words and phrases in certain environments, which would not have been possible with the manual examination of texts (Louw 1993; Sinclair 1991). Corpus linguistics, therefore, has evolved into a rigorous methodology used to describe structural, lexical, and variational linguistic phenomena (Kennedy 1998). The "young and restless" linguistic methodology provides room for novel and alternative analysis methods, bridging neighboring or even heterogeneous academic disciplines, such as linguistics and computer science, into a multi-disciplinary nature of the body of knowledge (Jurafsky and Martin 2008; Manning and Schütze 2001). Because of the multi-faceted applicability and all-around adaptability of the methodology across academic disciplines, it may not be easy to capture how a certain area of literature has influenced other areas and vice versa.

Recent advances in knowledge building are not the result of one or two leading academic disciplines; rather, they are a product of active and dynamic interactions in all walks of academia. Corpus linguistics plays a role among academic disciplines. Given the recent overarching knowledge-building practices and the methodological roles of corpus linguistics, it is necessary to review how a certain body of knowledge has been created according to the common denominator of corpus linguistics.

Ideas can be developed by others' ideas, but in some cases new ideas are generated from an individual's mind. They can also be generated from existing information of others. In academic areas, it is common to borrow others' ideas (Case and Higgins 2000). When referring to other scholars' knowledge, it is required to reveal where these ideas come from and whose ideas they are (i.e., citing information).

Investigating citation structures reveals the interaction patterns in a scholarly domain (Kuo and Yang 2012). Specifically, co-cited studies are examined to discover the important research and academic issues (Chen, Ibekwe-SanJuan, and Hou 2010; Tang et al. 2015). In addition, studies that address similar research questions tend to be located proximately in a research network; thus, clustering allows researchers to look further into intellectual structures and important research topics (Anderberg 1973; Kaufman and Rousseeuw 2009; Kuo and Yang 2012; Small 2003).

In the current study, we analyzed co-cited documents published in corpus linguistics during the past twenty years. The aims of this study are to discover the dominant themes

and publications and to investigate how they changed during the target period. We also demonstrate the clusters visually and further discuss the implications of the findings.

2. Literature Review

2.1 Corpus Linguistics

A corpus is a body of systematically gathered texts or transcribed speech to represent a particular function of a language that can serve as the basis for linguistic analysis and description. Corpus linguistics, therefore, is one of the sources for describing the structure and uses of languages as well as for different applications, such as natural language processing in computer science or language teaching and learning in language education (Kennedy 1998). Because corpus linguistics is based on bodies of large texts and the observation of the frequency of certain linguistic units (e.g., words or grammatical categories in the corpora), the corpus linguistic method enables the linguistic analysis of performance rather than competence, description rather than linguistic universals, quantitative and qualitative models of language, and a more empiricist rather than rationalist view of scientific inquiry (Leech 1992). Because of the powerful and versatile methodological approach, the scope of corpus linguistics, the boundaries of corpus linguistics, and other/neighboring areas of linguistics have become blurred and interdisciplinary in nature.

Corpus linguistics has a relatively shorter history because it is considered to have emerged with the development of computer technology. Along with the core role of linguistic research in the description and explanation of linguistic phenomena, additional research activities are specifically pertinent to corpus linguistics, including corpus design and compilation (Sinclair 1991; Fries, Tottie, and Schneider 1994), analytic tools development (Anthony 2009), probabilistic descriptive linguistic investigation (Oakes 1998; Halliday 1991), the application of linguistic descriptions such as language learning and teaching (O'Keeffe, McCarthy, and Carter 2007; Timmis 2015), and natural language processing, including translation studies (Oakes and Meng 2012) and speech recognition (Jurafsky and Martin 2008).

More recently, the scope of corpus linguistics has been subdivided and the number of interfaces with other academic disciplines has increased. Corpus

linguistics, however, has become an indispensable methodology throughout the field of linguistics and its neighboring disciplines. McEnery and Hardie (2012), for example, predict the third stage of corpus linguistics after the first stage of struggling and setting in during the late 1980s in the face of the Chomskian view of linguistics and the second stage of establishing corpus linguistics as a semi-independent sub-field of linguistics until present day. In their prediction, corpus linguistics would become increasingly integrated with other disciplines. Furthermore, the methodologies in corpus linguistics will be a crucial step for enhancing the rigor of incorporation into all kinds of linguistic and non-linguistic research.

It has been a decade since the book *Corpus Linguistics 25 Years on* (Facchinetti 2007) was published. The book surveys the corpus linguistics discipline, providing a brief overview of 25 years of corpus linguistics studies, including descriptive corpus studies of syntax and semantics, as well as second language acquisition with specialized corpora. As has been observed, corpus linguistics is a fairly new and rapidly growing discipline. The influence and impact of the new methodology is huge not only for related sub-fields in linguistics, but also in the major fields of the humanities, social sciences, and science and engineering. Given the considerable interest in utilizing the corpus linguistic approach, in addition to the dynamic and interdisciplinary nature of current studies involving partnerships among disciplines, a comprehensive and systematic overview of the development of and relationships among individual research in the fields of corpus linguistics is called for. Thus, the pressing academic quest is to review past achievements as well as future directions of corpus linguistics.

2.2 Citation Analysis

A citation analysis investigates the structures of the ideas being disseminated. In particular, co-citation patterns, which are generated when different items are simultaneously referenced in the same article, reveal the topics and features that are shared in that domain (Chen et al. 2010; Tang et al. 2015). Analyzing co-citation patterns explores existing relevant studies that offer prospective ideas for further additional investigation (Gmür 2003; Jankovic, Kaufmann and Kindler 2008). Such an analysis can be utilized to figure out phenomena in a certain field identified by researchers who appear in the same references (Zhao and Strotmann 2008). It also

allows for the determination of comparable areas cited by the same publications (Özçınar 2015). Thus, co-citation analysis defines the characteristics of a particular discipline (Kuo and Yang 2012) that are not easily discovered in the references at first sight. In particular, it is conducted to analyze documents' citation patterns, which is known as document co-citation analysis (DCA; Chen et al. 2010; Tang et al. 2015). DCA, according to Chen et al. (2010), is to investigate the documents which are shown in the references at the same time. That is, DCA examines the documents referenced in the same documents, assuming that the co-cited documents can uncover the arrangement of the academic knowledge in the domains.[1]

Investigating co-cited documents reveals clusters built in the intellectual community. A cluster in a network analysis indicates group results with a similar theme and present theme to the user in a more concise form (Tan, Steinbach, and Kumar 2006). An analysis of the clusters explores data to reveal specific groups in them (Kaufman and Rousseeuw 2009). This analysis method categorizes datasets into smaller clusters (Tan et al. 2006), demonstrating the specialties of the constructed groups of the significantly co-cited documents (Jankovic et al. 2008). In this way, it identifies the prominent themes existing in knowledge networks and facilitates the comprehension of the intellectual world (Anderberg 1973; Chen et al. 2010; Jalali and Park 2017).

Moreover, the clustering of intellectual communities reveals the collaboration structures between the groups within a community. The explored thematic clusters can be shown in the form of visualization. The visualized document clusters indicate the association among them, revealing the characteristics of the network from a broader point of view (Small 2003). Özçınar (2015) investigated how the intellectual communities of teacher education have changed, along with the illustrations of the main research themes reflected in the co-cited documents of the field. Such intellectual collaboration can be mapped out using a social network analysis (SNA) (Ronda-Pupo, Sánchez, and Cerpa

1 Citation occurs as borrowing ideas from other researchers in intellectual resources. Notably, in terms of academic papers, the lists of articles, books, proceedings, and other resources are shown in the references to demonstrate the original ideas that are used to support the findings in the papers. On the one hand, co-citation indicates cases of when any two different documents are cited in one same paper (Tang et al., 2015). That is, the investigation of co-citation shows the network of the intellectual resources that are simultaneously referenced in the same academic papers, assuming that they share the similar areas of knowledge. In particular, Author Co-citation Analysis (ACA) is to investigate the authors of the co-cited documents, while Document Co-citation Analysis (DCA) is to examine the co-cited documents focusing on the specific information of the documents (Chen et al., 2010). For more information, refer to Bellis (2009).

2015; Jalali and Park 2017; Park and Leydesdorft 2013). A growing number of studies on social networks have been conducted (Gmür 2003) to discover significant figures and academic works with the advent of various analysis tools (Liang, Liu, Yang, and Wang 2008).

2.3 Applications of Citation Analysis

Various disciplines have been studied using citation analysis to investigate the research trends. A recent article used citation as an indicator to evaluate the academic performances of research teams (Popova, Romanov, Drozdoz, and Gerashchenko 2017). With regard to academic disciplines, Gmür (2003) investigated the structure of co-citations in sociology. The co-citation patterns of the authors (White and McCain 1998) and clusters along with the co-citation patterns (Shiau, Dwivedi, and Yang 2017) were examined in information science. In addition, in the area of healthcare, studies have investigated the risk management system (Bradea, Delcea, and Paun 2015) and research impacts of the academic papers (Mori and Nakayama 2013). Even the structure of the patents in medicine was examined (Song, Han, Jeong, and Yoon 2017). Furthermore, researchers in the field of business management utilized citation analysis to comprehend the structures of the business models (Kuo and Yang 2012) or major themes in entrepreneurship (Schildt, Zahra, and Sillanpää 2006).

Meanwhile, in the education fields, along with the semantic and knowledge network analysis of second language acquisition domain (e.g., Jang, Wood, and Khan 2017), citation analysis has been used to assess research trends. Budd and Magnuson (2010) examined the citation patterns in higher education journals and listed the most cited authors, publication forms, and publication years. The most cited journals and books were further investigated. Park (2012) replicated Budd and Magnuson's (2010) study to examine the citation patterns in computer-assisted language learning. More recently, Griffin (2017) analyzed randomly chosen dissertations in educational leadership fields. In that study, the most popular serial titles were represented, and the researcher determined that the items published within the most recent decade were frequently cited, and studies in various disciplines were mentioned in educational leadership. Another trial visualized the citation patterns in teacher education within the past two decades (Özçınar 2015). Özçınar (2015) utilized DCA to reveal the flow of the change in the teacher education

field. Specific disciplines were found and mapped in three different time spans. Meanwhile, Fazel and Shi (2015) conducted qualitative research to identify the reasons for the citation. They interviewed graduate students, asking why they reference other scholars' works when writing proposals for research grants. This study was meaningful in the sense that the graduate students' recorded utterances were coded according to the purpose of the citation and subsections in the proposals.

In the field of corpus linguistics and its neighboring field of language education, only a few attempts have been made to explore a comprehensive and bird's eye view of the interaction among published research articles on the subject. Gilquin and Gries (2009), for example, examined how corpus linguists and psycholinguists use corpus linguistic methodology in their research. The authors collected 81 papers from the then-most recent issues of representative corpus linguistics journals. They also collected 85 papers from bibliographical databases using specific Boolean keywords —namely, "corpus," "experiment," and "elicitation." From the review study of corpus linguistic research compared to psycholinguistics, the authors found that the researchers in the two disciplines use corpora in different ways: for corpus linguists, exploratory and descriptive study; for psycholinguists, hypothesis-testing. Boulton and Cobb (2017), who were interested in the effectiveness of using the corpus tools and techniques of data-driven learning, conducted another comprehensive study of corpus linguistics. In their meta-analysis of 64 separate studies, the researchers revealed a positive relationship between the sample sizes and effect sizes. However, few studies have explored certain key areas of durability/transfer of learning; these areas are recommended for future data-driven learning studies. Although Gilquiin and Gries (2009) and Boulton and Cobb (2017) offered revealing insights into how corpus linguistic studies are similar to or different from neighboring fields and the "common truth" behind the conceptually similar studies (e.g., overall effects), they may not reveal the overall trends of mapping the conceptual relationship and developments of the individual studies in the corpus linguistics discipline.

In order to contribute new insights to explore and uncover research trends in corpus linguistics from the past two decades, the purpose of the current study is address the following questions:

1) In corpus linguistics literature, what are the salient works over the last two decades?

2) What salient issues have been discussed in corpus linguistics research during the past 20 years?
3) How have these issues emerged and faded during this period?

3. Methodology

3.1 Data

A series of small-scale pilot studies of the co-citation analysis was carried out to evaluate the technical feasibility of processing a large amount of citation data and identify any initial patterns of the co-citation analysis results. First, in order to ensure a robust and systematic investigation, *CiteSpace*, a program for visualizing and analyzing trends and patterns in scientific literature, was employed to conduct the co-citation analysis (Chen et al. 2010). The citation data were automatically downloaded by Boolean search terms related to corpus linguistics (e.g., "corpus" OR "corpora" OR "corpus linguistics") from Web of Science (WoS). The data consisted of research articles published in 2015 or 2016, which included 1,272 articles. The pilot studies revealed that clusters included somewhat unrelated themes to either linguistics or language, such as "government work report." To ensure the purposes and directions of the current research, and since the use of corpus has been closely related to language pedagogy (Römer 2011), such related search terms were also included in the main research as: (1) "corpus" OR "corpora" OR "corpus-based" OR "corpus-driven"; (2) "English for academic purposes" OR "EAP" OR "English for specific purposes" OR "ESP." Second, the citation data from the journal articles published from 1993 to 2016 in linguistics and language-related fields were downloaded from the databases of Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), and Arts and Humanities Citation Index (AandHCI). Within this period, 5,698 articles were pulled and tested to examine the reference co-citation patterns; however, the patterns were blurry and not clear enough to appropriately interpret the results, especially at the beginning of the period. Finally, the citation data were analyzed according to four different time spans from 1997 to 2016. The complete final data was downloaded from the WoS on June 29, 2017. Five-year spans were purposefully divided for the convenience of the analysis. As a result, the 5,600 research articles and their 172,352 references were analyzed: (1) 1997-2001 (286 documents with 9,215 references); (2)

2002-2006 (669 documents with 18,985 references); (3) 2007-2011 (1,792 documents with 52,382 references); and (4) 2012-2016 (2,853 documents with 91,770 references).

3.2 Analysis

Researchers refer to other researchers' ideas and borrow their findings from the published scholarly works. There might be various reasons for referencing other researchers' publications. One of the reasons is that the referenced academic resources are used to support their research ideas (Case and Higgins 2000). During the process, a co-citation occurs when two different documents are cited together by the same document (Tang et al. 2015). According to Chen et al. (2010), the primary purpose of the co-citation analysis is to look into patterns of identified groups of individuals who share similar themes.

Examining the clusters based on the specialties found in the references provides insights into the targeted scholarly world. References that share similar purposes can be grouped through cluster analysis (Kaufman and Rousseeuw 2009). Cluster analysis is a method used to investigate a phenomenon not revealed from a one-dimensional data set. Consequently, this method brings about insightful interpretations from the sections. When drawing insights from the data, the proximity between individual documents is mostly considered to evaluate the similarities for building groups (Anderberg 1973). The degree of the similarities between the documents is calculated using cosine coefficients. For instance, "if A is the set of papers that cites i and B is the set of papers that cite j, then $W_{ij}=\frac{|A\cap B|}{\sqrt{|A|\times|B|}}$, where $|A|$ and $|B|$ are the citation counts of i and j, respectively; and $|AB|$ is the co-citation count, i.e. the number of times they are cited together" (Chen et al. 2010: 8-9).

The grouped individual documents show the thematic closeness between them. Documents are likely to be clustered together with other documents dealing with the relevant themes (Kaufman and Rousseeuw 2009). Furthermore, the clusters' social structures are visually depicted via SNA (Ronda-Pupo et al. 2015; Jalali and Park 2017; Park and Leydesdorft 2013). Many programs for co-citation patterns exist (Liang et al. 2008) based on graph theory, which consists of "sets of nodes (vertices) and links (arc and edges)" (Park, Yoon, and Leydesdorff 2016: 1020). In other words, in co-citation clusters, nodes represent co-cited documents or authors while links indicate a co-cited

relationship. In addition, visualized clusters can be defined according to the significant themes found in the documents within the clusters (Chan et al. 2010). In this study, the clusters were constructed according to the indexing terms in the citing documents. In order to address the research issues in the current study, we utilized *CiteSpace* to investigate co-citation patterns and illustrate the co-citation groups of specialties (Chen et al. 2010; Liang et al. 2008). We focused in particular on the co-citation relations between documents as they reveal the networks of the discipline more distinctively (Özçınar 2015).

4. Results

4.1 Salient Publications

Two sources of important citation data, most-cited journal titles and the individual publications, were analyzed to understand citation patterns in corpus linguistics. Every journal title has its own aims and scope of publication; therefore, the analysis of the cited journal titles identified trends in how corpus linguistics research has been communicated within the research community. Furthermore, the analysis of the most-cited publications was expected to identify influential and leading works within the given time span.

Table 1 shows the 10 most-cited journals of every five-year time span since 1997 to serve as an overview of the journals investigated in the current study. Four journals appeared in all of the time spans: *Language*, *Journal of Pragmatics*, *Linguistics*, and *Applied Linguistics*. Although *Applied Linguistics* is relatively new (founded in 1980), the remaining three journals have contributed to the linguistic community for more than 40 years. *Language*, *Linguistics*, and *Journal of Pragmatics* were founded in 1925, 1963, and 1977, respectively. These journals have been privileged journals in the linguistics community for several decades and have produced much of body of linguistic knowledge; thus, it is interesting to observe how corpus linguistics-related studies have cited these journals over time. For example, the citation frequency of *Language* (see Table 1) has grown 10 times since the 1997-2001 time span. The other journals show similar patterns in terms of the citation frequency.

Meanwhile, other journals from the list disappeared over time. *Computational Linguistics* was ranked at the top in the 1997-2001 time span, but the number of citations did not grow much; the ranking of the journal declined, and eventually it dropped off the

list in the recent time span. Furthermore, some journals that appeared in the middle of the time span became actively cited by the corpus linguistics papers since then. *English for Specific Purposes* and *TESOL Quarterly*, founded in 1980 and 1967, respectively, appeared after the turn of the century and have been consistently cited by the corpus linguistics papers.

Another trend worth noting is the emergence of specialized journals on the list. Between 2007 and 2011, *International Journal of Corpus Linguistics*, which was first published in 1996 and covers the areas of linguistics, applied linguistics, and translation studies, began to be cited widely, ranking until even recent years. *Journal of English for Academic Purposes*, along with *English for Specific Purposes* and *Cognitive Linguistics*, is another example of a specialized journal actively being cited by corpus linguistics papers. Therefore, based on the journals most cited by corpus linguistics papers, corpus linguistics studies not only refer to general linguistic studies, but also specialized journal papers for new areas of language education, cognitive linguistics, and corpus linguistics itself.

Table 1. Most Cited Journal Titles between 1997 and 2016

Time span	Journals
1997-2001	Computational Linguistics [89]; Language [67]; Journal of Pragmatics [38]; Applied Linguistics [37]; Cognition [34]; Lingua [29]; Linguistic Inquiry [28]; Language and Speech [28]; Language in Society [26]
2002-2006	Language [125]; Journal of Pragmatics [105]; Computational Linguistics [101]; English for Specific Purposes [84]; Linguistics [79]; Applied Linguistics [77]; TESOL Quarterly [73]; Cognition [54]; Text [51]; Journal of Linguistics [47]
2007-2011	Language [420]; Journal of Pragmatics [369]; Applied Linguistics [296]; English for Specific Purposes [225]; International Journal of Corpus Linguistics [210]; Linguistics [210]; TESOL Quarterly [192]; Computational Linguistics [190]; Language in Society [172]; Text [146]
2012-2016	Language [683]; Journal of Pragmatics [626]; Applied Linguistics [518]; English for Specific Purposes [398]; International Journal of Corpus Linguistics [385]; Linguistics [359]/TESOL Quarterly [359]; Lingua [309]; Journal of English for Academic Purposes [288]

Note. The number in brackets indicates the number of citations.

When it comes to the most cited works by corpus linguistics papers, the works of researchers like John Sinclair, Douglas Biber, Michael Halliday, and Randolph Quirk would be expected to appear in all time spans examined in the current

research because these works contributed to the development of corpus linguistics. However, according to the results of the current study, relatively new names were found, especially in later time spans. For example, Mike Scott is well known for his computer concordancer *WordSmith Tools*; Ken Hyland for his works in disciplinary discourse analysis and the writing of English for academic purposes; and John Swales for rhetoric, discourse analysis, and English for specific purposes. Table 2 presents the 10 most cited publications according to the five-year time spans since 1997. The most cited publications in each time span show both similarities and differences, which may represent research trends in the given periods.

Table 2. Most Cited Publications between 1997 and 2016

Time span	Publications
1997-2001	Markus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2): 313-333. [16]; Halliday, M. A. K. 1994. *Introduction to functional grammar* (2nd ed.). London: Arnold. [9]; Chaff, Wallace. 1994. *Discourse, consciousness, and time: the flow and displacement of conscious experience in speaking and writing*. Chicago, IL: University of Chicago Press. [7] / Hopper, Paul J. and Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge, UK: Cambridge University Press. [7] / Quinlan, J. Ross. 1993. *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann. [7]; Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *International Journal of Translation Studies* 7(2): 223-243. [6] / Baker, Mona, Gill Francis, and Elena Tognini-Bonelli (eds.). 1993. *Text and technology: In honour of John Sinclair*. Amsterdam: John Benjamins [6] / Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman. [6] / Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22(2): 249-254. [6] / Laviosa, Sara. 1998. The English comparable corpus: A resource and a methodology. In Lynne Bowker, Michael Cronin, Dorothy Kenny, and Jennifer Pearson (eds.), *Unity in diversity? Current trends in translation studies* (101-112). Manchester: St. Jerome Publishing. [6] / Stubbs, Michael. 1996. *Text and corpus analysis*. Oxford: Blackwell. [6]
2002-2006	Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar and spoken and written English*. London: Longman. [42]; Hyland, Ken. 2000. *Disciplinary discourse*. Cambridge: Cambridge University Press. [22]; Biber, Douglas, Susan Conrad, and Randi

	Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press. [15]; Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. [13] / Hyland, Ken. 2001. Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* 20(3): 207-226. [13]; Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(3): 213-238. [12] / Kennedy, Graeme D. 1998. *An introduction to corpus linguistics*. London: Longman. [12] / Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: The MIT Press. [12] / Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press. [11] / MacWhinney, Brian. 2000. *The CHILDES project: The database*. London: Lawrence Erlbaum. [11] / Miller, George and Christiane Fellbaum. 1998. *Wordnet: An electronic lexical database*. Cambridge: The MIT Press. [11] / Moon, Rosamund. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press. [11]
2007-2011	Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. [46]; Swales, John M. 2004. *Research genres: Exploration and applications*. Cambridge: Cambridge University Press. [44]; Halliday, M. A. K. 2004. *Introduction to functional grammar* (3rd ed.). London: Arnold. [39]; Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243. [38]; Goldberg, Adele. 2006. *Constructions at work*. Oxford: Oxford University Press. [37]; Croft, William and D. Alan Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press. [34]; Hopper, Paul J. and Elizabeth Closs Traugott. 2003. *Grammaticalization* (2nd ed.). Cambridge, UK. Cambridge University Press. [32]; Tomasello, Michael. 2003. *Constructing a language: A usage-based account of language acquisition*. Cambridge: Cambridge University Press. [31]; Hoey, Michael. (2005). *Lexical priming: A new theory of words and language*. New York: Routledge. [30]; Carter, Ronald and Michael McCarthy. 2006. *Cambridge grammar of English: A comprehensive guide - spoken and written English grammar usage*. Cambridge: Cambridge University Press. [28] / Wray, Alison. 2002. *Formulaic language*. Cambridge: Cambridge University Press. [28]
2012-2016	Davies, Mark. 2008. The corpus of contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/ [64]; Baayen Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. [59]; Scott, Mike. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software. [51]; Goldberg, Adele. 2006.

Constructions work. *Cognitive Linguistics* 20(1): 201-224. [40] / Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press. [40] / Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21. [36]; Baayen, Harald, Douglas Davidson, and Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390-412. [35]; Simpson-Vlach, Rita and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487-512. [32]; Bybee, Joan. 2006. From usage to grammar: Mind's response to repetition. *Language* 82(4): 711-722. [30]; Langacker, Ronald. 2008. *Investigation in cognitive grammar*. New York: Mouton de Gruyter. [29]

Note. The number in brackets indicates the number of citations.

During the earliest period of the current research scope, between 1997 and 2001, the research topics related to new perspectives on grammar were the central issue in corpus linguistics. This trend likely stemmed from the publication of grammar references based on usage-based explanations or empirical corpus data, such as *Introduction to Functional Grammar* and *Longman Grammar of Spoken and Written English*. In tandem with the advancement of technology and large-scale empirical data, researchers attempted to explain grammar with the parsed corpora of the large-scale empirical data: *Penn Treebank*.

Between 2002 and 2006, although researchers still cited corpus-based grammar references for their studies (e.g., *Cambridge Grammar of the English Language*), one group of researchers made use of newly developed datasets, both large or small, such as *The CHILDES Corpus*, *Wordnet*, and *A New Academic Word List*. Another group of researchers, possibly novices in the discipline, cited general references to corpus linguistics, such as *Introduction to Corpus Linguistics*, *Corpora in Applied Linguistics*, and *Foundations of Statistical Natural Language*. An alternative explanation for the emergence of the introductory references of corpus linguistics would be because the period was the optimal time for establishing corpus linguistics as a part of linguistics after a "hodgepodge" multi-directional development of corpus linguistics.

Despite the small inception of specific and narrowed-down research topics, based on the cited publications between 2002 and 2006 (e.g., discipline and research articles by Ken Hyland and fixed expressions and idioms by Rosamund Moon), between 2007 and 2011, the corpus linguists continuously developed, expanding the areas of corpus linguistics research by citing ideas from *Collostructions* by Anatol Stefonowitsch and Stephan Gries, *Cognitive Linguistics* by William Croft and Alan Cruse, *Constructing a*

Language: A Usage-based Account of Language Acquisition by Michael Tomasello, and *Lexical Priming: A New Theory of Words and Language* by Michael Hoey.

Most recently, between 2012 and 2016, the two most interesting citation trends reflected the ease of accessibility and convenient manageability of corpora. Two sources stand out: Mark Davies' *Corpus of Contemporary American English* (COCA) and Mike Scott's *WordSmith Tools*. These two sources are not research papers, but are the introduction to a new type of corpus and a corpus analysis concordancer software suite program, respectively. The emergence of the COCA meant that researchers, teachers, and students are able to study ongoing changes in linguistics, including morphology, syntax, semantics, and lexis with language, using an unprecedented large-scale corpus, which may not have been possible with the previous "Brown family" of corpora, let alone the familiar architecture and interface of the Google search engine, which is familiar to most internet users. In fact, technology has been one of the main barriers of entry in corpus linguistics practice. Thanks to *WordSmith Tools*, many linguists, teachers, and even students are able to compile personal corpora and analyze the statistical and concordance data. Thus, during the last five years, the thresholds of practicing corpus linguistics for language researchers and practitioners have lowered.

Research between 2012 and 2016 highlighted recurrent sequences of word forms, fixed expressions, lexical bundles, and n-grams. Earlier, between 1997 and 2001, *LGSWE* discussed lexical bundle structures as a part of lexical expressions. Between 2002 and 2006, researchers cited *LGSWE* along with the fixed expressions of Rosamund Moon. In addition, between 2012 and 2016, the area of corpus linguistics was further developed through discipline study and language pedagogy research in Ken Hyland's discipline study and Simpson-Vlach and Ellis' academic formula list.

4.2 Salient Academic Research Themes

Table 3 presents 49 prominent clustered themes from the five-year time spans over the last 20 years: nine themes in the first time span (1997-2001), 17 themes in the second (2002-2006), 13 themes in the third (2007-2011), and 10 themes in the fourth (2012-2016). The main themes for each cluster were "algorithmically organized according to hierarchical relations derived from co-occurring concepts" (Chen 2017:17). In the co-citation analysis with *CiteSpace*, a cluster is defined and extracted as a group of shared themes of title terms, keywords, and abstract terms of citing articles (Chen 2017;

Tan et al. 2006). For more accurate clustering, the analysis also include the keywords of the references' titles (Thomson Reuters 2010). For this reason, one of the themes between 2007 and 2011 is "weird" (see Table 3) because Ignacio M. Martínez and Paloma Pertejo's (2014) article "Strategies Used by English and Spanish Teenagers to Intensify Language" cited three references containing the word "weird." In total, 13 references included the word; the skewed frequency of the words affected the clustering of the themes, which is a noise and ignorable in the current study.

Table 3. Clustered Themes between 1997 and 2016

Time span	Clustered Themes
1997-2001	Processing definite description [21; 1994]; Translation studies [20; 1993]; Academic speech [18; 1997]; Research article [15; 1995]; MDL principle [13; 1992]; Functional opposition [13; 1994]; Learning system [11; 1995]; Noun verb problem [10; 1992]; Spanish text [6; 1997]
2002-2006	Pedagogic issue [44; 1998]; Discourse marker [22; 2001]; Research article [31; 1998]; Doctoral student [31; 2000]; Indirect object [24; 2000]; Summarizing scientific article [22; 1996]; Unseen bigram [22; 1998]; Academic writing [21; 1999]; Urban Nigerian Arab [20; 2000]; Syntactic hierarchical configuration information [17; 1999]; ESP classroom [17; 1998]; Economics metaphor [15; 2001]; Signalizing delay [9; 2003]; Interactional language [9; 1998]; Semantic relation [7; 2000]; Competing motivation [6; 1999]; Grammaticalization phenomenon [5; 1998]
2007-2011	Formulaic sequence [33; 2003]; Metadiscourse [33; 2002]; Schema abstraction [27; 2004]; Dative alternation [23; 2005]; Corpus consultation [22; 2005]; Weird [N/A; N/A] [21; 2002]; Metaphor [19; 2004]; Sense [15; 2002]; Politeness [11; 2002]; Language contact [11; 2003]; Machine [9; 2004]; Work [6; 2002]; Localization [5; 2006]
2012-2016	Statistical model [55; 2008]; Formulaic language [44; 2007]; Corpus-based study [37; 2007]; Academic writing [31; 2008]; Data-driven learning [29; 2008]; Second language development [20; 2009]; Progressive aspect [16; 2008]; Phrasal verb [5; 2008]; Grammar checker [3; 2010]; Pragmatic marker [2; 2010]

Note. The numbers delimited by a semicolons in the brackets indicate the cluster size (i.e., the number of references) and the mean average publication year of the references within the cluster, respectively.

Research papers and books have been published to introduce corpus linguistics and define how to use it or build a small corpus for pedagogic purposes. These introductory publications advanced the usage of the corpus for language teaching. In the late 2000s and

early 2010s, studies about exploiting corpus were conducted concerning referencing-compiled corpus for language learning, especially in academic writing. For instance, students' attitudes or reactions were examined regarding consultation with corpus while they were writing.

Considering the relationship among most sited publications and the salient academic research themes, it seems that the corpus linguistics has become a linchpin of certain academic disciplines. This is visually witnessed as shown in Figure 1.

At the beginning, the clusters are sparse with distance, however, in the recent time span, the cluster has become a tight and dense network, which suggests productive research themes in corpus linguistics and active interaction among publication citation. More details about prominent clusters will be presented in the following sections, focusing on five years at a time.

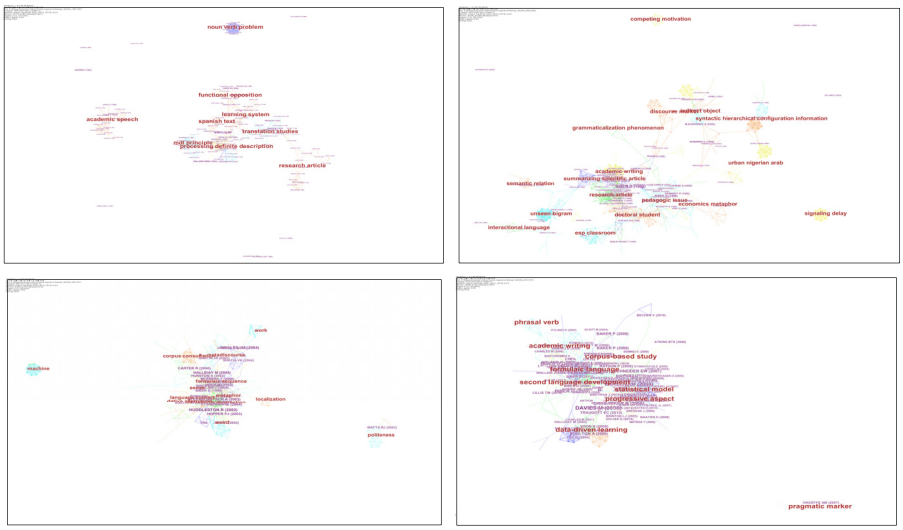

Figure 1. Overview of cluster network of the co-cited references and themes between 1997 and 2001 (top left), 2002–2006 (top right), 2007–2011 (bottom right), and 2012–2016 (bottom left).

4.2.1 1997–2001

Between 1997 and 2001, 9,215 co-cited references were obtained from the 286 articles. Nine keyword themes were derived from the references: "processing definite description"; "translation studies"; "academic speech"; "research article"; "MDL

principle"; "functional opposition"; "learning system"; "noun verb problem"; and "Spanish text." Figure 2 shows the network of the clustered themes in this time span. In *CiteSpace*, the colors of the clusters represent the year when the co-citation first occurred during the time span. For example, blue clusters were first created in 1998, while green and yellow ones emerged later, in 1999 and 2001, respectively. In addition, nodes, depicted by circles, present the co-cited references, and each link, which is line, shows the status of the co-citation between two different co-cited references. In addition, the moment of sudden increase of co-citations is shown by a bigger node size or in pink (Chen 2017).

During this period, a new view emerged toward the corpus dataset to be used for studies on grammar: Large text data compiled with the help of computational corpus enhanced the research on parsed grammar, resulting in publications such as *Introduction to Functional Grammar* by Halliday (1994) and *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, and Finegan 1999). The latter book in particular was salient in the cluster-themed "academic speech." This cluster did not share any co-citations with documents in other clusters; instead, documents within this cluster mostly shared citations in the same cluster label. In addition, Rosamund Moon's (1998) book *Fixed Expressions and Idioms in English: A Corpus-based Approach* was located in the central position within the cluster. The author emphasized that idioms are comprehensible based on the discourse with them. This publication also ranked at the top of the co-cited references in the second time span.

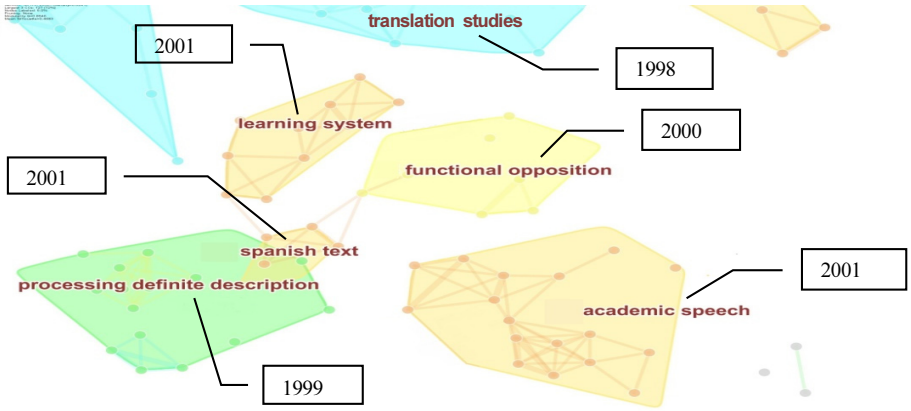

Figure 2. Cluster network of the co-cited references between 1997 and 2001. The color of the cluster indicates the year when the first co-citation occurred in the cluster: blue in 1998, green in 1999, yellow in 2000, and orange in 2001.

4.2.2 2002–2006

Between 2002 and 2006, 669 articles were analyzed as the dataset, resulting in 18,985 co-cited references that revealed 17 keyword themes: “pedagogic issue”; “discourse marker”; “research article”; “doctoral student”; “indirect object”; “summarizing scientific article”; “unseen bigram”; :“cademic writing” “urban Nigerian Arab”; “syntactic hierarchical configuration information”; “ESP classroom”; “economics metaphor”; “signalizing delay”; “interactional language”; “semantic relation”; “competing motivation”; and “grammaticalization phenomenon.” The most notable themes were “academic writing” and “research article” (see Figure 3). The two clusters consisted of citing articles published primarily in 2004. In particular, when it comes to the cluster “research article,” Ken Hyland’s (2000) book *Disciplinary Discourses: Social Interactions in Academic Writing* was co-cited with articles in “academic writing.” Hyland discussed communicational features reflected in research in terms of different disciplines and forms of publications. He also discussed academic writing and cultural aspects of texts in different fields (Hyland, 2000).

On the other hand, Biber et al.’s (1999) book *Longman Grammar of Spoken and Written English* was also dominant in the co-citation network (see Figure 3). This book was in the cluster of “summarizing scientific article”-citing articles published in 2003 in general. The book describes the syntactic aspects of English in four types of discourses through quantitative analysis using corpus (Biber et al. 1999; Biber et al. 2002). It was co-cited with Ken Hyland’s publications in other clusters, such as *Disciplinary Discourses: Social Interactions in Academic Writing* (Hyland, 2000) in the “research article” cluster and the journal article “Humble Servants of the Discipline? Self-mention in Research Article” (Hyland, 2001) in the “academic writing” cluster. Furthermore, articles giving co-citations to these articles covered the topics of learner corpus in EAP (Lee and Swales 2006), thesis writers’ language use (Charles 2006), linguistic features (Hewings and Hewings 2002), and metadiscourse and social engagement in academic writing (Tse and Hyland 2006) across various disciplines. These studies were conducted using compiled corpus.

Another significant cluster during this time span, “pedagogic issue,” had a wider range in the cluster and of overlapped parts with other clusters, such as “research article,” “academic writing,” and “summarizing scientific writing” (see Figure 3).

This cluster included citing articles that were mostly published in 2003. This cluster included introductory corpus-related books (Biber, Conrad, and Reppen 1998; Kennedy 1998) and academic word lists (AWLs) (Coxhead 2000). More specifically, *Corpus Linguistics: Investigating Language Structure and Use* (Biber et al. 1998) provided information on corpus-based academic writing and speaking in terms of lexicography, word frequency, and word usage across registers. *An Introduction to Corpus Linguistics* (Kennedy 1998) presented fundamental concepts of corpus linguistics and types of corpus programs; this book also illustrated the corpus-based research which focuses on grammar and lexicon and further elaborated on corpus analysis. Coxhead's (2000) article "A New Academic Word List" examined an academic English vocabulary list built through a series of corpus analyses. The AWL in this article was developed by referring to principles in Biber et al.'s (1998) and Kennedy's (1998) (Coxhead 2000). This finding may imply that the two referenced books helped create AWL, further leading to the publications of the academic works regarding pedagogic issues in corpus linguistics and EAP.

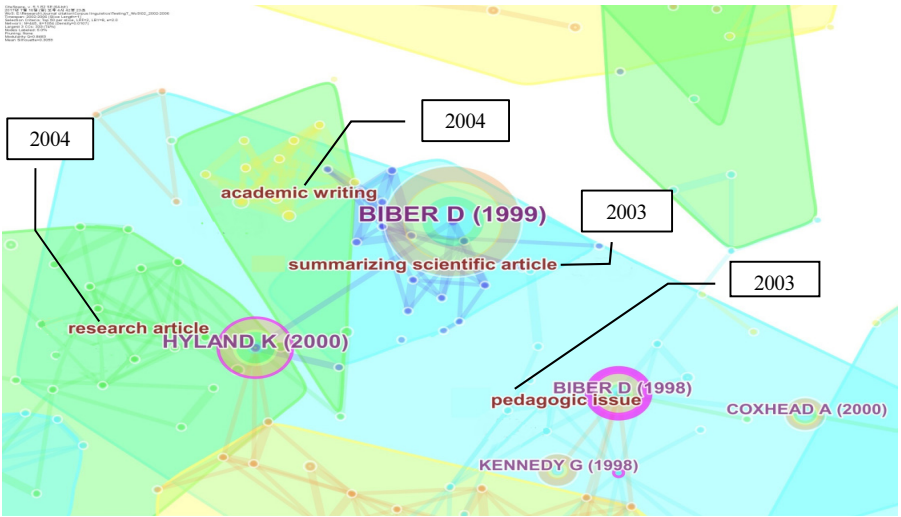

Figure 3. Cluster network of the co-cited documents between 2002 and 2006. The color of the cluster indicates the year when the first co-citation occurred in the cluster: blue in 2003 and green in 2004. The thickness of a ring is proportional to the number of citations in the time span.

4.2.3 2007–2011

In 1,792 articles from 2007 to 2011, 52,382 co-cited references were identified. The references showed 13 keyword themes: “formulaic sequence”; “metadiscourse”; “schema abstraction”; “dative alternation”; “corpus consultation”; “weird”; “metaphor”; “sense”; “politeness”; “language contact”; “machine”; “work”; and “localization.” During this period, corpus linguistics was used to figure out a fixed series of words and use them for students’ language learning, which was distinctively depicted in the cluster network (see Figure 4). In the “formulaic sequence” cluster, Michael Hoey’s (2005) book *Lexical Priming: A New Theory of Words and Language* was shown to be an influential publication. According to this book, language learners acknowledge and acquire a rule related to a combination of words. Lexical priming indicates a moment when a language learner recognizes a series of words that probably occur in terms of collocation, grammatical rules, and location in the discourse. Referring to the concordances that the author built, the author claimed that lexical priming could happen semantically and contextually (Hoey 2005; Khamkhien 2015). Another distinctive publication in the formulaic word sequence cluster (see Figure 3) was Wray’s (2002) book *The Transition to Language*, which consists of studies on the evolution of language delivered at an international conference in 2000. In general, these two books were co-cited by articles published in 2008, whose addressed issues included formulaic sequences of words (Ellis, Simpson-Vlach, and Maynard 2008), prefabricated lexical chunks on the web (Shei 2008), discovery of lexis regarding gaming (Ooi 2008), and relation between primed lexis and its position in a text (Hoey and O’Donnell 2008). Hoey’s (2005) and Wray’s (2002) publications were co-cited by those in the clustered themes in schema abstraction. For example, Hoey (2005) was co-cited with *Cognitive Linguistics* by Croft and Cruse (2004) whereas Wray (2002) was co-cited with *Collostructions: Investigating the Interaction of Words and Constructions* (Stefanowitsch and Gries 2003) and *Constructing a Language: A Usage-based Account of Language Acquisition* (Tomasello 2003). The co-citations of these articles and books may indicate that cognitive aspects can be related to the word sequences.

Another cluster, “corpus consultation” (see Figure 4), included two publications by Angela Chambers: one describing the use of corpora and concordances as

immediate consultation tools for students' language learning (Chambers 2005), and the other presenting research on corpus consultation conducted both qualitatively and quantitatively as well as the pros and cons of the corpus consultation (Chambers 2007). In addition, Yoon and Hirvela (2004) observed attitudes of ESL learners when they were writing by referring to corpus and found that referencing concordances was beneficial for writing. Furthermore, in Lee and Swales's (2006) article, students compiled personal corpus with their writing and compared that to professional writing. In general, these four studies researched students' use of corpus for academic writing after an investigation of the lexical items in an enormous size of texts. They were co-cited by a review paper exploring corpus in L2 writing (Yoon 2011), an article that examined the students' reactions to the existence of guidance on their corpus use (Pérez-Paredes, Sánchez-Tornel, Calero, and Jimenez 2011), and research about using specialized corpus in a language course (Rodgers, Chambers, and Le 2011). In line with the co-cited articles, these citing studies also investigated exploiting corpus as a referential tool for learning and using language.

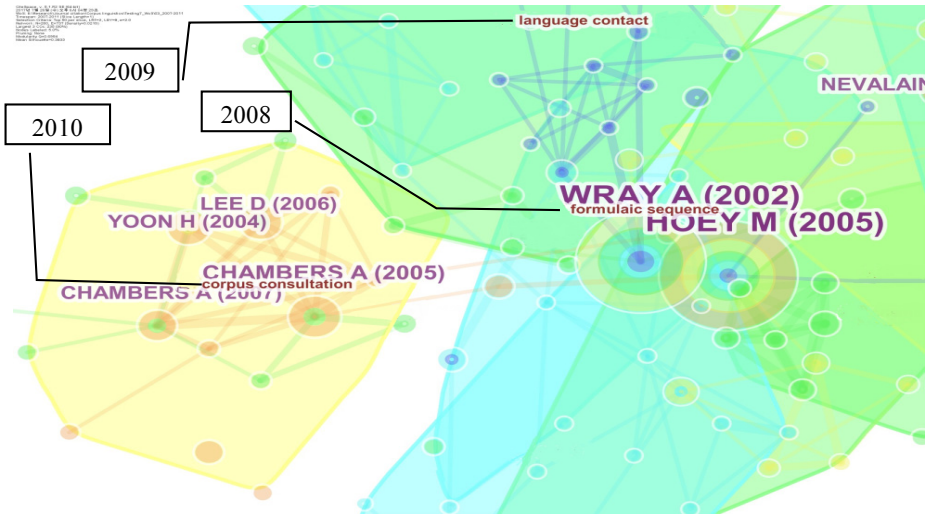

Figure 4. Cluster network of the co-cited documents between 2007 and 2011. The color of the cluster indicates the year when the first co-citation occurred in the cluster: blue in 2008, green in 2009, and yellow in 2010. The thickness of a ring is proportional to the number of citations in the time span.

4.2.4 2012-2016

A total of 10 keyword themes were extracted from the 91,770 co-cited references between 2012 and 2016: “statistical model”; “formulaic language”; “corpus-based study”; “academic writing”; “data-driven learning”; “second language development”; “progressive aspect”; “phrasal verb”; “grammar checker”; and “pragmatic marker.” The references were included in 2,853 articles. Of the keyword themes, “formulaic language” was a prominent theme related to corpus linguistics and EAP or ESP (see Figure 5). This theme was related to the “formulaic sequence” theme found in the 2007-2011 period. In this cluster, an article presented the *Academic Formulaic List* (AFL) of Simpson-Vlach and Ellis (2010). AFL was developed through the process of analyzing academic spoken and written corpora compared to general corpora. AFL provided a recurrent series of lexis dominantly appearing in academic environments; thus, it was the list of the lexical items beyond the single-word unit. Hyland’s (2008) paper in this cluster emphasized recurrent sequences of words, especially four-word combinations, and their implications in EAP. Hyland analyzed dissertations and academic articles to discover that the lexical bundles shape and differentiate the features of the discourse in academic disciplines.

Another distinctive cluster was “data-driven learning,” in which *COCA* (Davies 2008) appeared as the most influentially cited item. *COCA* is not a published article; rather, it is a web-based corpus complied of a 520 million-word database from newspapers, magazines, fiction, and other academic text documents. *WordSmith Tools* (Scott 2008) was dominant in the “corpus-based study” cluster. *WordSmith Tools* is not a published article either; it is downloadable software for analyzing huge texts. *COCA* and *WordSmith Tools* enabled educators to compile their own text data and analyze the data in specific aspects of linguistics, such as lexis, syntax, semantics, or morphology. These corpus programs reduce the work of looking into the personal text data of educators and even students, which helped increase the related research, such as corpus-based research, data-driven learning, academic writing, or second language development.

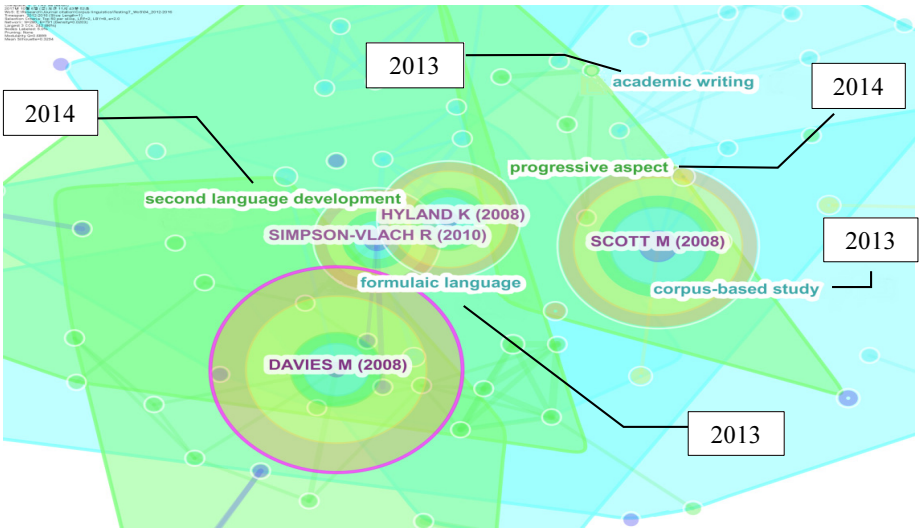

Figure 5. Cluster network of the co-cited documents between 2012 and 2016. The color of the cluster indicates the year when the first co-citation occurred in the cluster: blue in 2013 and green in 2014. The thickness of a ring is proportional to the number of citations in the time span.

5. Discussion and Conclusion

The present study explored the research trends in corpus linguistics over the past 20 years. Co-citation analysis was conducted to examine the salient themes of the clusters every five years, along with the significant co-cited publications within the themes. Changes among the most-cited journals and publications were also investigated. The findings revealed that four journals were steadily referred to during the two decades examined: *Language*, *Journal of Pragmatics*, *Linguistics*, and *Applied Linguistics*. Furthermore, journals that first ranked on the list after the early or late 2000s and remained dominant in corpus linguistics until now were *English for Specific Purposes*, *TESOL Quarterly, International Journal of Corpus Linguistics,* and *Cognitive Linguistics.* Thus, during the last two decades, in addition to general linguistics research being consistently referenced, specialized academic journals on corpus linguistics, cognitive linguistics, pragmatics, sociolinguistics, and language education have also been actively cited.

Corpus linguists' most-cited publications which served as the foundations of

corpus linguistics were constantly referenced. For instance, publications by John Sinclair, Douglas Biber, Michael Halliday, and Randolph Quirk were steadily cited throughout the 20 years. Meanwhile, publications by Mike Scott, Ken Hyland, and John Swales were first found in the middle time spans; their publications on corpus tools and discourse analysis in ESP or EAP were frequently cited. The pattern of cited publications also suggests that corpus linguistics has been specialized and branched out.

Co-citation patterns in each time span also revealed significantly clustered themes as well as publications within the clusters. In the years before 2000 and the early 2000s, corpus linguistics tended to be studied by using enormous empirical datasets. During that time, corpora enabled researchers to conduct studies on grammar using corpus-based parsers, such as the Penn Treebank and its parts of speech tags. Consequently, researchers were able to computationally build corpus data, focusing primarily on grammatical or semantic aspects in discourses. In addition, publications in academic writing, academic speech, pedagogic issues, and research articles appeared as important themes during this period. Most recently (i.e., since the late 2000s), corpus tools were more commonly used by various groups, including not only researchers, but also language teachers and students, who finally had direct access to corpus. Moreover, these corpus practitioners started to compile their own personal corpora for analyzing and investigating linguistic features and examples of expressions by using corpus analysis software packages. Another recent trend in the corpus linguistics research, according to the current study, was the emergence of large web-based corpus (e.g., *COCA*). With the development of the technology and the easy accessibility to the internet, it is also worth noting the new practices in corpus linguistics research (e.g., Gato 2014; Hundt, Nesselhauf, and Biewer 2007; Schäfer and Bildhauer 2013). With the lowered barriers to building corpora and using corpus analysis tools, corpus linguistics have recently been expanded to the wider range of research areas, even including pedagogical aspects. In particular, during the recent period, corpus linguists broadened the research topics into a series of words and put more emphasis on the pedagogical perspectives of corpus linguistics. Researchers have become more interested in the areas of formulaic sequence, corpus consultation, and data-driven learning when utilizing corpus tools.

Few existing studies have explored research patterns and trends in corpus linguistics from a bird's-eye view. The findings of the current study demonstrate

how quickly corpus linguistics has grown through diverse and distinctive, yet collaborative research endeavors. Future researchers should explore a wider range of facets of the co-citation patterns of corpus linguistics. Indeed, unlike the current study of co-citation patterns of the referenced publications, future studies should illustrate the corpus linguistics trends and research network among the co-cited authors (Chen et al. 2010; He and Hui 2002; White and McCain 1998). Such analyses should compare and contrast the results of the current study to better understand research patterns and trends.

References

Anderberg, Michael R. 1973. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. New York, NY: Academic Press.

Anthony, Laurence. 2009. Issues in the design and development of software tools for corpus studies: The case for collaboration. In Paul Baker (ed.), *Contemporary corpus linguistics*, 87-104. New York: Continuum.

Baker, Paul (ed.). 2009. *Contemporary corpus linguistics*. New York: Continuum.

Bellis, Nicola De. 2009. *Bibliometrics and citation analysis: From the Science Citation Index to Cybermetrics*. Lanham, MD: Scarecrow Press.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Graeme Hirst. 2002. The Longman grammar of spoken and written English. *TESOL Quarterly* 34(4): 132-139.

Boulton, Alex and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67(2): 348-393.

Bradea, Ioana, Camelia Delcea, and Ramona Paun. 2015. Healthcare risk management analysis－A bibliometric approach. *Journal of Eastern Europe Research in Business and Economics* 2015: 1-11.

Budd, John M. and Lauren Magnuson. 2010. Higher education literature revisited: Citation patterns examined. *Research in Higher Education* 51(3): 294-304.

Case, Donald O. and Georgeann M. Higgins. 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science and Technology* 51(7): 635-645.

Chambers, Angela. 2005. Integrating corpus consultation in language studies. *Language Learning and Technology* 9(2): 111-125.

Chambers, Angela. 2007. Popularising corpus consultation by language learners and teachers. In Encarnación Hidalgo Tenorio, Luis Quereda, and Juan Santana (eds.), *Corpora in the foreign language classroom: Selected papers from the sixth international conference on teaching and learning corpora*, 3-16. Amsterdam: Rodopi.

Charles, Maggie. 2006. The construction of stance in reporting clauses: A cross-disciplinary study of theses. *Applied Linguistics* 27(3): 492-518.

Chen, Chaomei. 2017. Science mapping: A systematic review of the literature. *Journal of Data and Information Science* 2(2): 1-40.

Chen, Chaomei, Fidelia Ibekwe-SanJuan, and Jianhua Hou. 2010. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology* 61: 1386-1409.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-238.

Croft, William and D. Alan Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.

Davies, Mark. 2008. The corpus of contemporary American English: 450 million words, 1990-present. Retreived from http://corpus.byu.edu/coca/

Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguists, corpus linguists, and TESOL. *TESOL Quarterly* 42(3): 375-396.

Faccinetti, Roberta (ed.). 2007. *Corpus linguistics 25 years on*. Amsterdam: Rodopi.

Fazel, Ismaeli and Ling Shi. 2015. Citation behaviors of graduate students in grant proposal writing. *Journal of English for Academic Purposes* 20: 203-214.

Fries, Udo, Gunnel Tottie, and Peter Schneider (eds.). 1994. *Creating and using English language corpora: Papers from the fourteenth international conference on English language research on computerized corpora, Zürich 1993*. Amsterdam: Rodopi.

Gato, Maristella. 2014. *Web as corpus: Theory and practice*. New York: Bloomsbury.

Gilquin, Gaëtanelle and Stefan Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1-26.

Gmür Markus. 2003. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics* 57(1): 27-57.

Griffin, Karin L. 2017. Citation analysis for core journals in educational leadership. *Collection Building* 35(1): 12-15.

Halliday, M. A. K. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer and Bengt Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*, 30-43. UK: Longman.

Halliday, M. A. K. 1994. *Introduction to functional grammar* (2nd ed.). London: Arnold.

He, Yulan and Siu Cheung Hui. 2002. Mining a web citation database for author co-citation

analysis. *Information Processing and Management* 38: 491-508.

Hewings, Martin and Ann Hewings. 2002. "It is interesting to note that…": a comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes* 21(4): 367-383.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. New York: Routledge.

Hoey, Michael and Matthew B. O'Donnell. 2008. Lexicography, grammar, and textual position. *International Journal of Lexicography* 21(3): 293-309.

Hundt, Marianne, Nadja Nesselhauf, and Carolin Biewer (eds.). 2007. *Corpus linguistics and the web*. Amsterdam: Rodopi.

Hyland, Ken. 2000. *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

Hyland, Ken. 2001. Humble servants of the discipline? Self-mention in research article. *English for Specific Purposes* 20: 207-226.

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.

Jalali, Seyed Mohammad Jafar and Han Woo Park. 2017. State of the art in business analytics: themes and collaborations. *Quality and Quantity* 1-7.

Jang, Haejin, Jacob Wood, and Gohar Feroz Khan. 2017. A social network analysis of knowledge infrastructure in the second language acquisition domain. *Linguistic Research* 34(Special edition), 125-160.

Jankovic, Milan P., Mark Kaufmann, and Christoph H. Kindler. 2008. Active research fields in anesthesia: A document co-citation analysis of the anesthetic literature. *International Anesthesia Research Society* 106(5): 1524-1533.

Jurafsky, Dan and James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). New York: Prentice Hall.

Kaufman, Leonard and Peter J. Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Hoboken, NJ: John Wiley and Sons, Inc.

Kennedy, Graeme D. 1998. *An introduction to corpus linguistics*. London: Longman.

Khamkhien, Attapol. 2015. Review of "Lexical priming: A new theory of Words and Language." *Electronic Journal of Foreign Language Teaching* 12(1): 135-138.

Kuo, Hsiu-Kuei and Chyan Yang. 2012. An intellectual structure of activity-based costing: a co-citation analysis. *The Electronic Library* 32(1): 31-46.

Lee, David and John Swales. 2006. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes* 25(1): 56-75.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In Jan Svartvik (ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm,*

4-8 August 1991, 105-122. Berlin: Mouton de Gruyter.

Liang, Yongxia, Zeyuan Liu, Zhongkai Yang, and Xianwen Wang. 2008. Knowledge mapping of citation analysis domains. In *Proceedings of the 4th International Conference on Webometrics, Informetrics and Scientometrics and 9th COLLNET Meeting*.

Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 157-175. Amsterdam: John Benjamins.

Manning, Christopher D. and Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Martínez, Ignacio and Paloma Pertejo. 2014. Strategies used by English and Spanish teenagers to intensify language. A contrastive corpus-based study. *Spanish in Context* 11(2): 175-201.

McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. UK: Cambridge University Press.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press.

Mori, Hiroko and Takeo Nakayama. 2013. Academic impact of qualitative studies in healthcare: Bibliometric analysis. *Plos One* 8(3). Retrieved from https://doi.org/10.1371/journal.pone.0057371

Oakes, Michael. 1998. *Statistics for corpus linguistics*. UK: Edinburgh University Press.

Oakes, Michael and Ji Meng (eds.). 2012. *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*. UK: John Benjamins.

O'Keeffe, Anne, Michael McCarthy, and Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. UK: Cambridge University Press.

Ooi, Beng Yeow Vincent. 2008. The lexis of electronic gaming on the web: A Sinclairian approach. *International Journal of Lexicography* 21(3): 311-323.

Özçınar, Hüseyin. 2015. Mapping teacher education domain: A document co-citation analysis from 1992 to 2012. *Teaching and Teacher Education* 47: 42-61.

Park, Hyejin. 2012. Analyzing article citation patterns in CALL journals. *Multimedia-Assisted Language Learning* 15(2): 143-166.

Park, Han Woo and Loet Leydesdorft. 2013. Decomposing social and semantic networks in emerging "big data" research. *Journal of Informetrics* 7(3): 756-765.

Park, Han Woo, Jungwon Yoon, and Loet Leydesdorff. 2016. The normalization of co-authorship networks in the bibliometric evaluation: The government stimulation programs of China and Korea. *Scientometrics* 109(2): 1017-1036.

Pérez-Paredes, Pascual, María Sánchez-Tornel, Jose Maria Alcaraz Calero, and Pilar Aguado. 2011. Tracking learners' actual use of corpora: guided vs non-guided corpus consultation. *Computer Assisted Language Learning* 24(3): 233-253.

Popova, Olga, Dmitry Romanov, Alexander Drozdoz, and Alexander Gerashchenko. 2017.

Citation-based criteria of the significance of the research activity of scientific teams. *Scientometrics* 112(3): 1179-1202.

Rodgers, Ornaith, Angela Chambers, and Florence Le Baron-Earle. 2011. Corpora in the LSP classroom: A learner-centered corpus of French for biotechnologists. *International Journal of Corpus Linguistics* 16(3): 391-411.

Römer, Ute. 2011. Corpus research application in second language teaching. *Annual Review of Applied Linguistics* 31: 205-225.

Ronda-Pupo, Guillermo, Aurora Sánchez, and Narciso Cerpa. 2015. Mapping the structure of international research collaboration network and knowledge domains on electronic commerce in the journal of theoretical and applied electronic commerce research. *Journal of Theoretical and Applied Electronic Commerce Research* 10(3): i-ix.

Schäfer, Roland and Felix Bildhauer. 2013. *Web corpus construction*. https://doi.org/10.2200/S00508ED1V01Y201305HLT022

Schildt, Henri A., Shaker A. Zahra, and Antti Sillanpää. 2006. Scholarly communities in entrepreneurship research: A co-citation analysis. *Entrepreneurship: Theory and Practice* 3: 399-415.

Scott, Mike. 2008. *WordSmith Tools Version 5*. Lexical Analysis Software. Oxford: Oxford University Press.

Shei, Chi-Chiang. 2008. Discovering the hidden treasure on the Internet: using Google to uncover the veil of phraseology. *Computer Assisted Language Learning* 21(1): 67-85.

Shiau, Wen-Lung, Yogesh K. Dwivedi, and Han Suan Yang. 2017. Co-citation and cluster analysis of extant literature on social networks. *International Journal of Information Management* 37(5): 390-399.

Simpson-Vlach, Rita and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487-512.

Sinclair, John. 1991. *Corpus concordancer collocation*. UK: Oxford University Press.

Small, Henry. 2003. Paradigms, citations, and maps of science: A personal history. *Journal of the American Society for Information Science and Technology* 54(5): 394-399.

Song, Chie Hoon, Jeung-Whan Han, Byeongki Jeong, and Janghyeok Yoon. 2017. Mapping the patent landscape in the field of personalized medicine. *Journal of Pharmaceutical Innovation* 12(3): 238-248.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to data mining*. Boston, MA: Addison-Wesley Longman Publishing Co.

Tang, Kai-Yu, Chia-Yu Wang, Hsin-Yi Chang, Sufen Chen, Hao-Chang Lo, and Chin-Chung Tsai. 2015. The intellectual structure of metacognitive scaffolding in science education: A co-citation network analysis. *International Journal of Science and*

Math Education 14(2): 249-262.
Thomson Reuters. 2010. Web of Science. Quick reference card. Retrieved from http://ips.clarivate.com/m/pdfs/mgr/webofscienceqrc.pdf
Timmis, Ivor. 2015. *Corpus linguistics for ELT*. NY: Routledge.
Tomasello, Michael. 2003. *Constructing a language: A usage-based account of language acquisition*. Cambridge: Cambridge University Press.
Tse, Polly and Ken Hyland. 2006. 'So what is the problem this book address?': Interactions in academic book reviews. *Test and Talk* 26(6): 767-790.
White, Howard and Katherine McCain. 1998. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society of Information Science* 49(4): 327-355.
Wray, Alison. 2002. *The transition to language*. Oxford: Oxford University Press.
Yoon, Choongil. 2011. Concordancing in L2 writing: An overview of research and issues. *Journal of English for Academic Purposes* 10(3): 130-139.
Yoon, Hyunsook and Alan Hirvela. 2004. ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing* 13(4): 257-283.
Zhao, Dangzhi and Andreas Strotmann. 2008. Author bibliographic coupling: Another approach to citation-based author knowledge network analysis. *Proceedings of the Association for Information Science and Technology* 45(1): 1-10.

Hyejin Park
Department of Educational Theory and Practice
University at Albany, SUNY
1400 Washington Ave, Albany, NY, USA
E-mail: ilspring36@gmail.com

Daehyeon Nam
Division of General Studies
Ulsan National Institute of Science and Technology
Business Administration Building, Rm. #406-7
50 UNIST-gil, Ulsan 44919, Republic of Korea
E-mail: dnam@unist.ac.kr

Received: 2017. 10. 03.
Revised: 2017. 11. 04.
Accepted: 2017. 11. 10.