Ph.D. Dissertation

Mapping the Knowledge Structure of Korean Humanities

– Bibliographic data analysis of humanities journal articles in the Korea citation index, 2004~2019–

Byungjun Kim

Department of Interaction Science

The Graduate School

Sungkyunkwan University

Ph.D. Dissertation

Mapping the Knowledge Structure of Korean Humanities

2022

Byungjun Kim

Ph.D. Dissertation

Mapping the Knowledge Structure of Korean Humanities

– Bibliographic data analysis of humanities journal articles in the Korea citation index, 2004~2019 –

Byungjun Kim

Department of Interaction Science

The Graduate School

Sungkyunkwan University

Mapping the Knowledge Structure of Korean Humanities

– Bibliographic data analysis of humanities journal articles in the Korea citation index, 2004~2019 –

Byungjun Kim

A Ph.D. Dissertation Submitted to the Department of Interaction Science and the Graduate School of Sungkyunkwan University in partial fulfillment of the requirements for the degree of Ph.D. in Interaction Science

April 2022

Supervised by

Jang Hyun Kim

Major Advisor

This certifies that the Ph.D. Dissertation

of Byungjun Kim is approved.

[signature]

Committee Chair :

[signature]

Committee Member :

[signature]

Committee Member :

[signature]

Committee Member :

[signature]

Major Advisor :

The Graduate School

Sungkyunkwan University

June 2022

Table of Contents

List of Tables

List of Figures

Abstract

Mapping the Knowledge Structure of Korean Humanities: Bibliographic data analysis of humanities journal articles in the Korea citation index, 2004~2019

Using a digital humanities technique, this study analyzes the bibliographic information of around 250,000 KCI humanities papers published between 2004 and 2019 in order to comprehend the knowledge structure of Korean humanities over the previous 15 years. Bibliographic information used in the analysis includes text information such as the thesis's title, abstract, and keywords, citation information such as the number of citations and references, and demographic information such as the gender, age, and academic institution of the researcher who wrote the thesis. This study consists of four chapters: 1) Exploratory data analysis of KCI thesis bibliographic information and each researcher's demographic and sociological information;,2) research topic cluster analysis utilizing the structural topic model; and 3) research topography analysis utilizing co-citation analysis of references. 4) Policy

suggestions for the future of Korean humanities based on three research outcomes.

In Study 1, I will use exploratory data analysis to analyze the number of papers published by year, the change in paper output over time in the reference list, and the generation and gender of researchers. In Study 2, topic changes are studied based on each thesis's textual content and its meta-information. I study which study topics have risen and declined in popularity throughout time, as well as which topics will garner interest in the future. In Study 3, groups of publications are extracted based on their citation relationships utilizing simultaneous citation analysis of references, and key works that led to the study cluster are selected and analyzed by cluster. In conclusion, the structure of knowledge generation and diffusion in the Korean humanities over the past 15 years is disclosed based on the above four data analysis results, and the future of the Korean humanities is considered. In particular, the future path of Korean humanities and the agenda of open science and digital humanities are proposed.

Keywords: Scholar Communication, Korean Humanities, Knowledge Structure, Digital Humanities, KCI

Chapter 1. Introduction

1. The need for and purpose of this study

1) The need for this study

Since November 2007, when the National Research Foundation of Korea (NRF) piloted the Korean Citation Index system (henceforth referred to as KCI), KCI has had a considerable impact on the humanities in Korea. Scholarly communication, research performance monitoring, and professor appointments were all influenced by the KCI in the Korean humanities community. In 2004, over 10,000 humanities papers were published in the KCI; as of 2019, that number stood at 18,000.

However, rather than recognize the quantitative rise of the humanities and conclude that the Korean humanities had expanded qualitatively, it has been claimed by critics that the KCI acted as a hegemony that controls the Korean humanities (Cheon, 2010). The rise of the KCI is connected to the neoliberal system and the university evaluation system. In other words, neoliberalism created the current academic environment by allowing the application of economic logic to the evaluation of university performance. This system compels humanities professors to create only academic writing for thesis, thereby shaping the humanities world. Despite criticism from members in the

humanities community, the KCI method has become the academic standard. Particularly, state-led research promotion initiatives such as BK (Brain Korea), HK (Humanities Korea), and SSK (Social Sciences Korea) strengthened the KCI system. This is because these organizations have mandated the submission of quantitative thesis results. Moreover, it is believed that the so-called "Hakjin" system has had a negative impact on the humanities in Korea and is the source of the "crisis" in the humanities (Jung, 2013). This crisis in the humanities is already a worldwide phenomenon, not only in Korea. According to Kaufmann (1977), when quantitative performance assessment was applied to the humanities in the 1970s in the United States, the concern with microscope and overspecialization led to the fragmentation of knowledge, which precipitated a crisis in the humanities.

Aside from the fact that the KCI system had a significant impact on how humanities scholars wrote their research and theses, the research that quantitatively analyzed the academic communication and knowledge structure of the Korean humanities through KCI over the past two decades found that the KCI system had a significant impact on the academic communication and knowledge structure of the Korean humanities. It was unusual. For the most part, library and information science scholars have done studies to clarify the knowledge structure of the Korean humanities. In the library and information science research, important humanities fields were evaluated (Jeong, 2020b; J. Y. Lee, 2015, 2015, 2021; Song, 2015), although the study did not

encompass all areas of Korean humanities. Although other studies had previously demonstrated the knowledge structure of Korean studies, this was the first to use bibliographic data from the international publishers SCOPUS (H. Kim, 2020a, 2020b).

In the humanities, macroscopic studies on knowledge structures began when access to thesis bibliographic material was facilitated and anyone could easily employ computational power and programming language. Particularly, as the field of digital humanities is currently in the limelight, research on research trend analysis is being performed in several prominent institutions. These analyses focus on the dissertations of academic disciplines (Kim & Cheon, 2020) or several decades of journal articles (Seol et al., 2020). As a result, bibliometric-based research, which was hitherto exclusive to library and information science, will likely be implemented in the humanities and social sciences soon (Kim, 2021).

Alternatively, the demographic information of the scholars who wrote the thesis can be considered as the KCI bibliographic information. This is since knowledge structure and scholarly communication do not consist solely of papers but are also directly tied to the social background of academic researchers. The NRF manages the research performance of Korean researchers using the Korean Research Information (KRI) system. KRI includes demographic data such as a researcher's gender, date of birth, and specialized field of study. Very few studies have utilized this information to

examine academia. A study was conducted on Japanese studies, which have entered a period of decline in South Korea, using the paper output of Japanese studies researchers in Korea and the demographic data of researchers (Jin, 2020). In addition, the effect of research group factors such as gender and major on the performance of convergence research was examined (Lee, 2016).

To summarize, in order to comprehend the knowledge structure of the Korean humanities, it is necessary to assess the research results released by KCI and the material provided by researchers who participated in the study. Because the knowledge system of the Korean humanities involves two axes, namely research and researcher, this is the case. Recent attempts to employ bibliographic or researcher information through bibliometrics or digital humanities approaches have been limited to certain key humanities fields or research that does not even use researcher information. Since the establishment of KCI, it is necessary to conduct a macro analysis of the Korean humanities using bibliographic data and researcher information from the entire humanities field.

2) Purpose of the study

The purpose of this study is to investigate the knowledge structure of the Korean humanities using KCI bibliographic data and digital humanities methodology and based on this, to examine research conducted over the past

two decades to demonstrate the possibility of future research employing bibliographic methodologies in the humanities. Through descriptive statistical analysis of KCI and KRI data, the current status of Korean humanities research and researchers is first determined. Second, I study the structural topic modeling of the thesis text information in order to examine the relationship between the evolution of the research topic's content and its topic as time passes. Using the study of bibliographical co-citation, I extract and cluster the most influential humanities research publications over the past two decades. In addition, utilizing key indicators from the citation network, I will assess the significance of the works in Korean humanities study and the resulting paradigm shift in knowledge. On the basis of the results of the three preceding analyses, the knowledge structure of the Korean humanities is appraised as a paradigm shift and diffusion theory, and future recommendations for the Korean humanities are provided. The specifics of the aforementioned four study objectives are provided below.

First, after integrating the bibliographic information of each KCI humanities thesis with the author's KRI researcher information, the overall status of the Korean humanities community is analyzed using EDA. KCI-related statistics include the number of papers published, the number of citations, and particular fields of study, whereas KRI-related statistics include the researcher's personal information, including gender, age, and generation (10-year). Also retrieved are statistics connecting KCI and KRI data.

Second, by employing structural topic modeling, the research topic latent in the thesis text data (title, abstract, and keyword) is identified. In the humanities studies conducted over the past two decades, I identify the conceptual terms that have been employed and the subject clusters that these terms constitute. By studying the changes in research patterns over time by infusing time series meta-information as a covariate of structural topic modeling, topics that will attract attention in the future and topics that are fading are studied.

Thirdly, key references referenced by scholars in humanities publications are extracted by analyzing bibliographic co-citations. By clustering the extracted references, it is determined which sort of research is most frequently cited in humanities research, and the features of the top-level references, such as betweenness centrality and citation burst, are assessed based on the index criteria of the citation network.

Fourth, the knowledge structure of the Korean humanities and the features of academic communication will be explored by considering the results of the three analyses presented before. Currently, the bibliometric approach utilized primarily in science and technology is utilized to explain the Korean humanities' trends. In addition, I provide policy recommendations for the future of Korean humanities centered on KCI.

2. Method and composition of the study

1) Method of the Study

This study intends to demonstrate the potential of digital humanities research utilizing bibliometrics in future humanities research by gaining a thorough understanding of the Korean humanities' knowledge structure. To achieve this objective, data analysis was conducted on the bibliographic information of all articles classed as humanities by KCI from 2004 to 2019 and the researcher information of the authors who wrote the papers. The acquired data was refined in the computer languages Python and R, and EDA, structural topic modeling, and simultaneous citation analysis were conducted to examine the knowledge structure of Korean humanities from multiple perspectives. Python's Pandas and Matplotlib programs were used to evaluate and illustrate descriptive statistics, while R's stm package was utilized for structural topic modeling. In addition, for simultaneous citation analysis, CiteSpace, software for bibliographic analysis, was used for analysis and visualization.

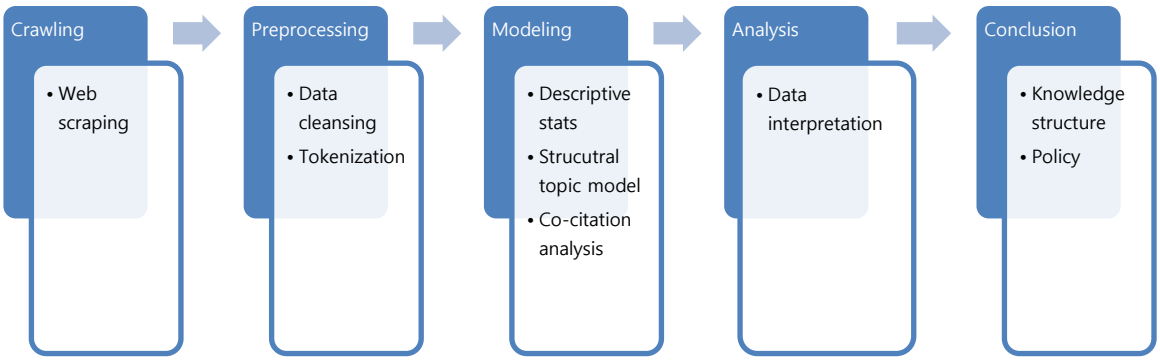

Figure 1. The research processes.

2) Structure of the Study

This study is comprised of five chapters, each of which is described below.

Chapter 1's Introduction covers the study's necessity and aim, as well as the research methodology.

In Chapter 2, Theoretical Background, the relationship between the Korean humanities and the KCI system, as well as past research, is analyzed. In addition, the concept of knowledge organization was discussed from the standpoint of academic communication and prior research. With the introduction of digital scholarship, both the concept of open science and the academic environment transformed. In addition, theoretical foundations for bibliometrics and digital humanities as analysis tools were presented, and relevant earlier research was examined.

The third chapter, Methods, will explain the research model and demonstrate the data collecting, preprocessing, and analysis procedures. In addition, the research model's EDA, structural topic modeling, and simultaneous citation analysis techniques were described.

In Chapter 4, Research Results, the analytical outcomes of three research models for research topics are presented.

In the concluding section of Chapter 5, I attempt to interpret the results

from a variety of perspectives so as to facilitate an understanding of the knowledge structure of the Korean humanities and to increase the utility of the research results by offering future policy recommendations for the Korean humanities. In addition, the limitations of this study and future research directions are outlined.

Chapter 2. Background and Literature Review

This chapter summarizes the relevant contexts for comprehending the knowledge structure of the Korean humanities and associated earlier research. To comprehend the knowledge structure of Korean humanities, one must comprehend the meaning of KCI bibliographic data and the idea of knowledge structure. First, the KCI system is used to arrange scholarly communication and digital scholarships, which have had a significant impact on the modern academic environment. In addition, the process of constructing the knowledge structure is understood from the standpoint of academic communication, and past research that have sought to assess the knowledge structure based on these two notions that have been categorized as bibliometrics and digital humanities.

1. Scholarly communication

1) Scholarly communication

Scholarly communication is communication in the academic realm, encompassing the production, evaluation, distribution, preservation, and reuse of research results (Klain-Gabbay & Shoham, 2018). Academic advancement and further transition to a "knowledge society" are made possible by scholarly

communication (Mukherjee, 2009). Although there are numerous definitions of academic communication, it can be understood as a link that allows scholars to express their perspectives (De Solla Price, 1965). Borgman (2000) defines scholarly communication as the process through which academics seek, utilize, and disseminate information. Even though there are numerous definitions of scholarly communication, it is evident that the communication process is often associated with it.

Currently, scholarly communication is separated into formal and informal communication. Public domain documents, such as theses and books, constitute official communication. Examples of representative official communication actions include the citation of other scholars' articles or books and the publication and dissemination of research results. Through formal communication activities, scholars can accumulate their academic accomplishments and contribute to the preservation of knowledge (Mukherjee, 2009). On the other hand, informal communication occurs rapidly and effortlessly, including face-to-face talks, e-mail exchanges, preprints, and, more recently, social media (see Table 1).

Table 1

Features and Benefits of Informal and Formal Scholarly Communication

Type	Feature	Advantages
Informal Scholarly Communication	● Communication partners know each other. ● One-to-one communication is used, from face-to-face discussions to exchanging opinions via e-mail.	● Usually quick and easy. ● A wide range of information exchanged.
Formal Scholarly Communication	● Tools of a research and public nature, such as monograph journals or journal papers, are used.	● Conveys information to a large readership via public communication. ● Simple to obtain specific information. ● The material that has been thoroughly evaluated and may be cited when necessary. ● Provides a rationale for giving academic performance priority. ● Serves the function of an archive.

Source: Compiled from Ji (2020), p.17.

Formal and informal communication are crucial to the growth of the academic environment. Roosendaal and Guerts define scientific (scholarly) communication in four ways (Roosendaal & Geurts, 1997). Registration,

archiving, certification, and awareness are the four functions (see Figure 2). Registration is the function of having precedence over previous academic discoveries. The purpose of archiving is to preserve academic documents and literature. Certification is a function that confers validity on documented academic accomplishments. Scholars are aware of new claims and discoveries.

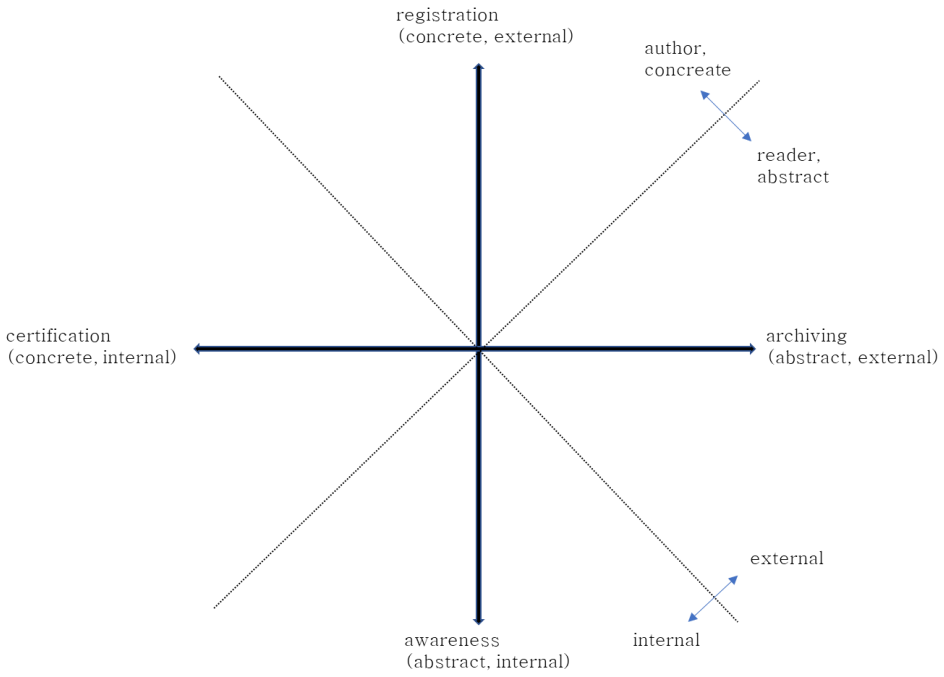

Figure 2. The Four Functions of Scientific Communication.
Source: Compiled from Roosendaal & Geurts(1997), p.14.

2) The process of scholarly communication

The development of scholarly communication varies by academic subject, and several models exist based on various scholarly interpretations. However, the model of Garvey & Griffith(1972) is representative as a paradigm illustrating the development of academic communication. The

Garvey–Griffith model was founded on psychologists' research, but its relevance to physics and the social sciences has been demonstrated by them. The model comprises three phases: research start, implementation, and conclusion. At each stage, researchers engage in distinct modes of communication. At the beginning of the research process, researchers discuss their research problems informally with their colleagues. During the research performance phase, while they are writing their own manuscript, they communicate with other researchers and engage in informal conversation. In the completion stage, research is disseminated through conference presentations, etc., and feedback is obtained prior to submission. Currently, formal communication takes place through preprints and other means (see Figure 3).

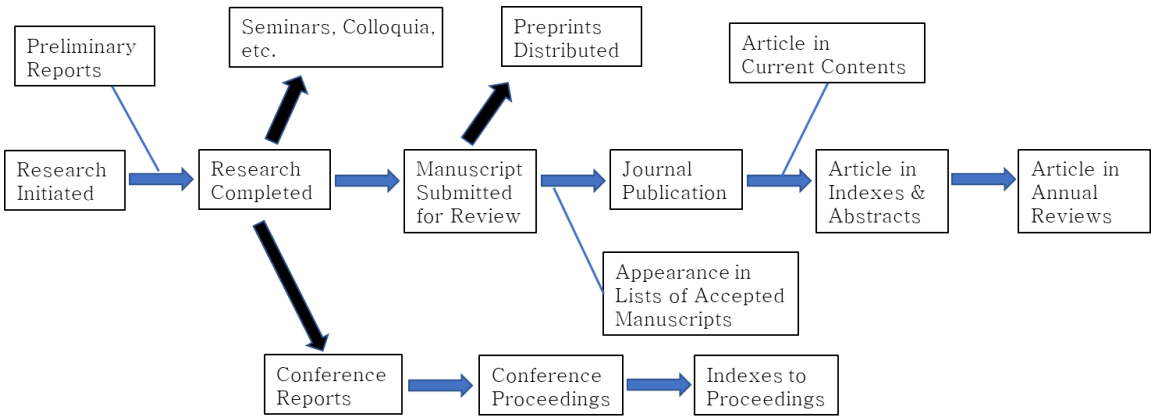

Figure 3. The Garvey–Griffith Model.
Source: Compiled from Garvey & Griffith (1972), p.127.

The Garvey–Griffith model is a case study of foreign scholars, but

there is also a case study of the scholarly communication process among Korean humanities researchers. To comprehend the humanities research process, Yu (2016) conducted in-depth interviews with twelve Korean history scholars. Similar to the Garvey-Griffith model, the Yu model includes informal or formal academic contact in the intermediate stage of the research process (see Figure 4). Unlike earlier models, humanities research revealed an organic relationship between ideas, facts, and logic throughout the research conception stage. This is an exclusive trait of research in the humanities. The cyclical form of this study is a distinguishing feature of the humanities research process compared to the social sciences, natural sciences, and engineering (Knöchelmann, 2019).

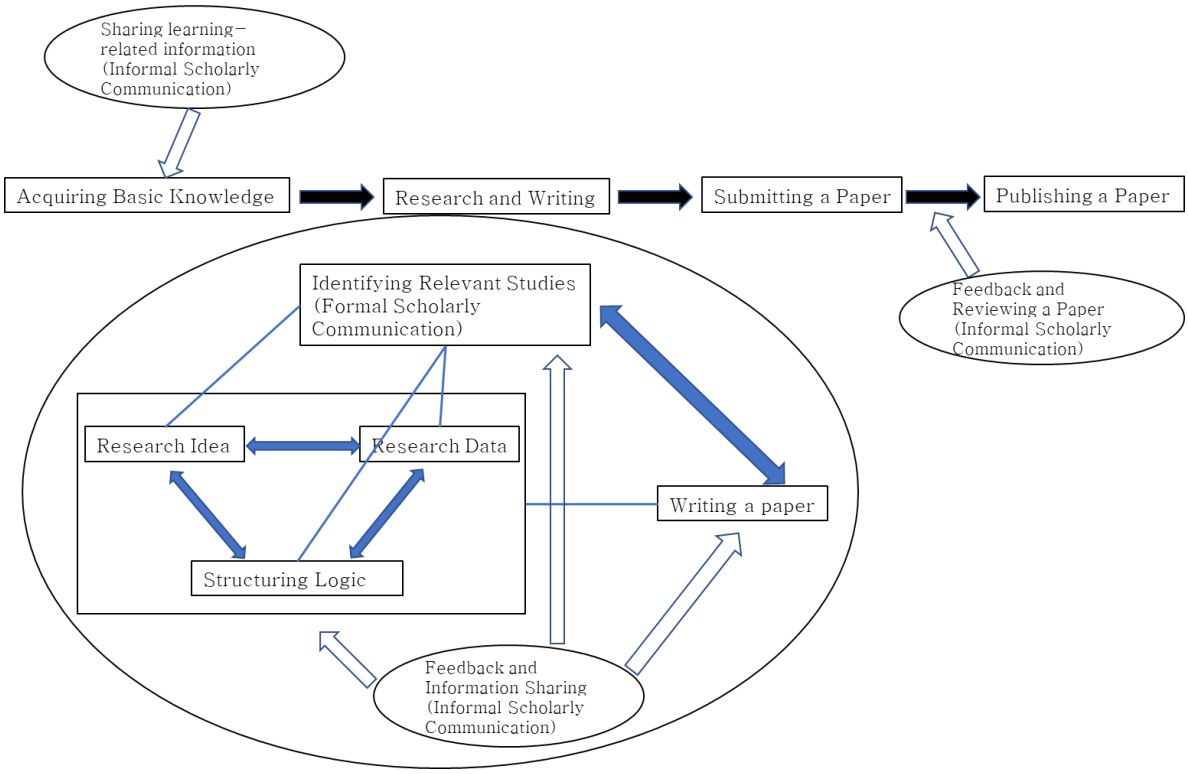

Figure 4. Humanities Research Process Identified through In-depth Interview. Source: Compiled from Yu (2016), p.338.

Due to the advancement of digital technology, scholarly communication has taken on a new form. Now, academics conduct studies utilizing digital research data and tools and communicate the findings digitally (Assante et al., 2015). Hurd (2000) projected the research process of academic communication in 2020 (see Figure 5). According to Hurd, electronic journal systems and other digital media would replace the current print-based approach. As he anticipated almost twenty years ago, digital scholarship has generated a new academic wave. The following section elaborates upon this trend.

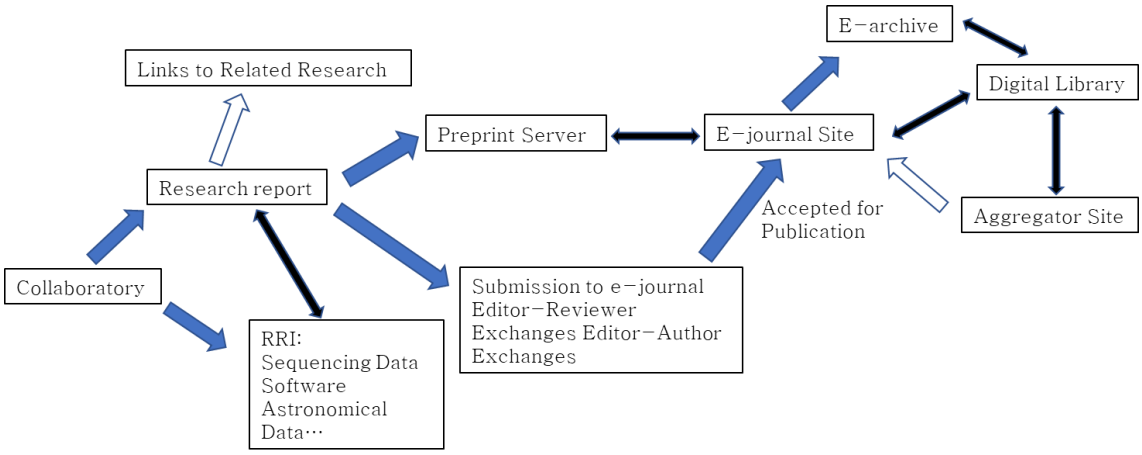

Figure 5. Scientific Communication : A Model for 2020.
Source: Compiled from Hurd (2000), p. 1281.

2. Digital scholarship and open science

1) Digital scholarship

Digital scholarship is infrastructure adapted to changes in the academic environment, often known as digital transformation. This new paradigm might be described as "multidisciplinary, open, network-centric, and highly reliant on internet technology" (Thanos, 2014). As the existing print-based academic communication environment transitions to a digital environment that can be accessed by anyone, various disciplines can easily meet (multidisciplinary), research and research-related data can be accessed (open), and connected, creating a ripple effect (network-centric). Cyberscholarship, like digital scholarship, refers to academic environment changes in a comparable context (Arms, 2008; Larsen, 2008). Arms suggested that cyberscholarship is only possible with a web-based research environment and available study data. Larsen believes that machine-readable and accessible application program

interfaces (APIs) will become the norm in the new infrastructure supporting cyberscholarship. For instance, instead of a format centered on human reading, such as PDF, a data-friendly format like XML is required.

How would immediate remote academic communication effect humanities research in a digital scholarship environment? Compared to other academic subjects, it is known that "book-oriented academic communication" dominates the study of the humanities (Yu, 2016 as cited in). In addition, the rate of digital change appears to be slower than in other fields of study due to the nature of humanities research, which focuses mostly on independent study and conventional literature research. Due to the emergence of digital humanities and an increase in collaborative research, however, the humanities' primary research subject, classical literature, has been converted into electronic data, and humanities scholarly communication is taking on a new form (Knöchelmann, 2019; Longley Arthur & Hearn, 2021; Riande et al., 2020; Shim et al., 2015). According to Knöchelmann (2019), the time has come for a new humanities communication known as open humanities paired with open science, while keeping the communication approach of humanities research that is undertaken organically by reinterpreting and disseminating current research.

2) Open science (Humanities)

Open science is a broader idea that encompasses the digital

scholarship outlined previously (Steinerová, 2016). Beyond 'open-access publication,' where anyone can read research results, there is a movement for anybody to access all study results by opening up research data and note records (UNESCO, 2020). Additionally, complete accessibility to study results might serve as a mechanism to ensure research transparency and repeatability (Riande et al., 2020). Below is the UNESCO (2020) declaration on open science.

Driven by unprecedented advances in our digital world, the transition to Open Science allows scientific information, data and outputs to be more widely accessible (Open Access) and more reliably harnessed (Open Data) with the active engagement of all relevant stakeholders (Open to Society). However, in the fragmented scientific and policy environment, a global understanding of the meaning, opportunities and challenges of Open Science is still missing[1].

As mentioned in the preceding declaration, fragmented knowledge systems and policies do not facilitate the spread of open science. As modern natural scientific or engineering research has grown in scope, collaborative research and data sharing have become the norm, and open science has

[1] https://en.unesco.org/news/unesco-launches-global-consultation-develop-standard-setting-instrument-open-science

become widespread. As a result, open science cannot be generalized. Specifically, this is due to the fact that the humanities did not consider their study outputs to be data and did not record them in a digital format that anybody could utilize (Reigersberg, 2015).

The question of how to apply open science to the humanities appeared as a connection with digital humanities. Collaboration is inevitable due to the nature of digital humanities, which has a research breadth that a single researcher cannot cover (Kretzschmar & Gray Potter, 2010). During the partnership, it was inevitable that researchers would investigate standardized data formats and open methodologies (Riande et al., 2020). Several digital humanities projects, for instance, have published research data and outcomes in data repositories like the Open Science Framework (OSF) (Foster & Deardorff, 2017).

3) KCI system

The advent of the KCI system had such a significant impact that it determined the academic communication approach of everyone in Korean humanities' field. In contrast to the company-led SCI, the KCI, also known as the so-called academic system, was part of the state's academic promotion agenda. Prior to the introduction of the KCI system, each university's academic society published its own academic publications, but now all journals in Korea are categorized as listed, candidate, or unlisted by the government.

Because of the KCI's ranking system, Korean humanities publications are under state control which means NRF has the authority to decide whether to list each journal in the KCI. Consequently, numerous researchers condemn the KCI on the grounds that it hinders academic autonomy and originality because of the evaluation system that emphasizes only the number of papers(Cheon, 2010; K. Hwang et al., 2014; Jung, 2013).

On the other hand, the introduction of the KCI contributed significantly to the advancement of digital scholarship. For instance, it led to the standardization of the thesis format and the unification of bibliographic material in digital format. For instance, if a researcher is writing a thesis intended for publication in the KCI, he or she must adhere to the conventional bibliographic format, which includes the title, keywords, abstract, and bibliography. With the introduction of standard bibliographic formats for theses, quantitative analysis of Korean theses in the humanities became possible. From the standpoint of open science, however, there are still numerous flaws. KCI serves as an open access platform but lacks the ability to exchange research data via a data repository. In addition, the National Research Foundation (NRF) operates the Korean Research Memory (KRM) data repository, but it is still unfamiliar to the humanities and lacks several features that important for the facilitation of data exchange. Shim et al. (2015) recommended a reform to KRM for sharing humanities-related research data in the interest of open science. According to the survey, most of Korean

humanities experts continue to use analog research methods and are reluctant to share their data. The necessity of a policy solution to develop a humanities-centered repository was also underlined.

3. Knowledge structure

1) Concept of knowledge structure

There does not appear to be a consensus definition of what a knowledge structure is. The concept of knowledge structure varies based on the researcher's area of study, and the type of knowledge may likewise vary by discipline. For instance, humanities knowledge frameworks need to incorporate historical studies or frameworks such as intellectual history and conceptual history. This study examined the knowledge structure in library and information science because the knowledge structure cannot differ among academic disciplines[2].

Classification is a theory of knowledge structure, according to Farradane(1950), who defined knowledge structure as the link between knowledge in various domains. In contrast, De Solla Price (1965) stated that academics build research fronts by mentioning only a few current works relevant to their interests. The research fronts that he proposes are a type of

[2] There are several synonyms for knowledge structure, which are intellectual structure, structure of knowledge, and so on. In this study, the synonyms are also regarded as knowledge structures.

knowledge structure, i.e. an ongoing research topic. Small (1976) noticed that knowledge organization is developed during the research process in order to generate and disseminate new information. According to him, the structuring of knowledge is dependent on the interaction of scholarly and informal communication. However, since the interaction is unseen, it has been proposed to examine the knowledge structure by examining the journal citation pattern.

If I examine the viewpoints of LIS scholars on the knowledge structure in this manner, this study finds that the prevailing idea is "relation" (Song, 2015). One piece of knowledge is formed by contact with other knowledge, with the bibliographic data of the thesis and the researcher's information serving as good markers of this process. Because, for instance, the relationship between the thesis and the thesis and the researcher and the researcher can be retrieved from the data. Specifically, it will be feasible to explore the structure of knowledge by studying the relationship between knowledge through citation information given in references.

2) Paradigm shifts

Having previously described the concept of the knowledge structure, it is now time to describe modifications to the knowledge structure. How do the knowledge structures that publications, researchers, and citations represent evolve? The growth of scientific knowledge, according to Kuhn (2010), is not done by knowledge accumulation alone, but through paradigm shifts. In other

words, what was once deemed "exceptional science" in academia is only temporarily revolutionary and soon becomes "regular science" due to paradigm shifts. In accordance with Kuhn's theory, Chen & Song (2017) assert that paradigm shifts can be proven using bibliometrics. They note the lack of evidence in Kuhns argument for the framework he portrays as a conflict between paradigms. The framework also states that earth-shaking revolutions are extremely unlikely to occur. According to Chen and Song, Kuhnian paradigm transformations occur at "many levels of granularity" rather than a single level. And the key is to be able to use bibliometric research to examine the paradigm shift of multiple layers.

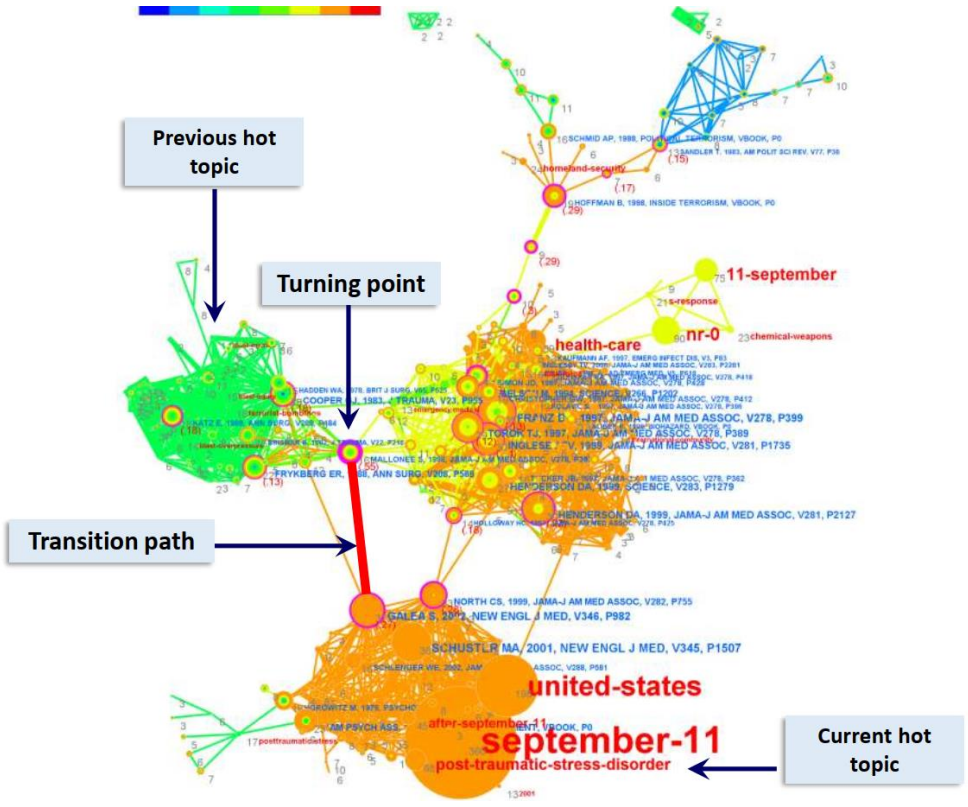

Figure 6. Turning point of research topic.
Source: Compiled from Chen (2021), p.74.

Using bibliometric analysis, Chen tries to analyze he study topic depicted in the graphic. This indicates that there is a path from the prior hot issue to the present hot topic through the paper's turning point. Although not as drastic as paradigm shifts, it can demonstrate that the structure of knowledge is also changeable.

4. Literature review

This section describes previous research on the Korean humanities that utilized bibliometrics or digital humanities approaches based on bibliographic data or researcher information. Bibliometrics is the statistical analysis of literary works such as books, journals, and other publications. Bibliography is typically applied in the field of library and information science, but with the introduction of big data and the growth of computational power, it is being used to assess research patterns in other academic disciplines. Digital humanities is "the study of computerizing humanities data and organizing them in digital media or of gaining new insights through data science analysis"(Cha, 2020). Text mining, natural language processing, and data science are therefore included in the digital humanities technique described here.

1) Bibliometrics

Song (2015) derived a knowledge structure by applying a bibliometric

method and FGI (Focus Group Interview) to academic thesis data in the field of Korean studies at home and abroad, and analyzes the bibliographic characteristics of the research topic, scope, and number of citations to demonstrate the distinction between domestic and international Korean studies. Ji (2020) provided a summary of the notion of scholarly communication, conducted a quantitative analysis of connected publications, and then examined specific themes using LDA topic modeling and network analysis. In two works, Kim identified the knowledge structure of Korean studies. First, the bibliographical information on Korean studies–related articles from KCI and SCOPUS was separated into distinctions between subject recognition and other recognition to demonstrate variances in study subjects (H. Kim, 2020a). In addition, using direct citation analysis and primary path analysis, SCOPUS's Korean studies–related studies uncovered research themes like ancient Korean agricultural culture and Koreans' English acquisition (H. Kim, 2020b).

2) Digital humanities

Kim & Cheon (2020) gathered bibliographic information from 1,528 doctoral dissertations from those majoring in modern literature in the Department of Korean Language and Literature from 2000 to 2019 in Research Information Sharing Service (RISS) which is a platform for scholarly communication in order to examine and forecast research trends. Keyword

analysis based on Term Frequency-Inverse Document Frequency (TF-IDF) and the dynamic topic model were utilized to investigate the evolution of research topics over time. Using bibliographic information (text data) of articles linked to Korean modern literature published in KCI from 1980 to 2019 and demographic information of academics, Lee & Kim (2020) discovered a consistent generation gap in research topics between scholars in their 20s and 50s. Hwang (2012) conducted a macroscopic analysis by utilizing bibliographic data from 845 papers of Sangheohakbo (journal) and Research on the History of Minjok Literature, which are representative journals for research on Korean modern literature, analyzing changes in keywords and foreign researchers who are frequently cited. He identified which foreign scholars were considered important in the study of Korean literature. Through a survey of Japanese study researchers in Korea and KRI data, Jin (2020) interviewed a total of 1,654 Japanese-related PhD degree holders. A Japanese researcher has demonstrated that the country entered a phase of slow decline after the 2010s, with an aging and highly specialized reproductive system.

Although not in the humanities, there have been previous studies utilizing bibliographic and researcher information in the social sciences. Kim et al. (2008) investigated the pattern of knowledge generation and distribution using 18,000 references of papers and articles published between 1996 and 2005 in the Korean Journal of Journalism. According to this study, the citation

frequency of publications exhibited a power function distribution, with Internet and mobile communication–related studies at the hub of the network. Kim & Song (2020) contrasted the discourses of Korean sociology and international sociology using bibliographic data from 2011 to 2018 from KCI and SSCI sociology journals. When the study topics were compared utilizing the structural topic model, the difference in discourse between the two academic circles became apparent.

Chapter 3. Methodology

1. Research model

This study provides a broad understanding of the knowledge structure of the Korean humanities based on KCI bibliographic data, which has had a significant impact on the Korean humanities and suggests the feasibility of employing bibliographical approaches in humanities research. How have humanities and humanities researchers in Korea evolved over the past fifteen years? What research subjects in the Korean humanities pique public interest? Specifically, which publications dominated the academic sphere and garnered notice in the Korean humanities? This section provides a summary of this study's research topics and models (see Table 2).

RQ #1: What have been the quantitative achievements of Korean humanities research over the past fifteen years, and how does this performance relate to demographic characteristics of researchers?

RQ #1-1: According to the bibliographic data of KCI, what type of quantitative growth and change did the Korean humanities demonstrate?

RQ #1-2: How differently do Korean humanities researchers engage in academic activities based on their gender or generation, according to KRI's demographic data?

RQ #2: What have been the primary research topics in the Korean humanities over the past fifteen years?

RQ #2-1: What were the distinguishing elements of hot topics that were in the spotlight and cold topics that were less popular among the key research issues in Korean humanities?

RQ #2-2: When and why did the inflection point of the rising-and-falling research topics in the Korean humanities occur?

RQ #3: In the past fifteen years, what have been the most important research references in the field of Korean humanities? What type of cluster did the references comprise?

RQ #3-1: What are the most prominent reference clusters produced from the co-citation analysis, and what are their academic characteristics?

RQ #3-2: What were the leading studies that shaped the knowledge structure of the Korean humanities in the reference network, and what are the features of these papers?

RQ#4: What is the knowledge structure of the Korean humanities, and how is it connected to digital humanities and open science?

RQ#4-1: What are the characteristics of the knowledge structure of the Korean humanities? How is it different from the knowledge structure of other disciplines?

RQ#4-2: How do the digital humanities and open science, proposed as a solution to the crisis in the humanities, relate to the findings of an analysis of the humanities' knowledge structure?

Table 2

Research Models

Research Questions	Processes	Methods	Chapters
RQ #1	a. Collecting KCI thesis bibliographic data and KRI researcher information. b. Mapping bibliographic data and researchers' data. c. EDA based on a consolidated table.	EDA	4.1
RQ #2	a. Extracting nouns with Korean morpheme analyzer after combining title, abstract, and keyword.	Structural topic model	4.2

	b. Analyzing research trends over the past 15 years through the structural topic model.		
RQ #3	a. Converting research data into Web of Science format and inputting it into CiteSpace. b. Analyzing topic clusters and major references in the Korean humanities through simultaneous citation analysis.	Co-citation analysis	4.3
RQ #4	a. Understanding the characteristics of the knowledge structure of the Korean humanities based on the three analysis results. b. Suggestions on KCI System and Korean Humanities Research Policy.	Data Interpretation	5

2. Data

1) Data crawling

A. KCI

By selecting 'humanities' as the primary subject category on the KCI

website[3], just the humanities papers were extracted. The specific search conditions in this instance are as follows:

Table 3

Search Rquirement in KCI

Search Requirement	Value
Categories	Humanities
Index	All : KCI, KCI candidate, non-official
Publication Year	2004[4] ~ 2019

KCI provides a method for exporting bibliographic information from the page displaying thesis search results. Up to 3,000 bibliographic records can be exported at once, and Excel, TXT, and XML file formats are available. In this investigation, bibliographic data was downloaded as an Excel file, and the procedure was automated using the selenium[5] Python program.

[3] https://www.kci.go.kr/kciportal/po/search/poArtiSear.kci

[4] According to the KCI website, this date was chosen as the starting point because studies published after 2004 have been fully databased.

[5] https://www.selenium.dev/

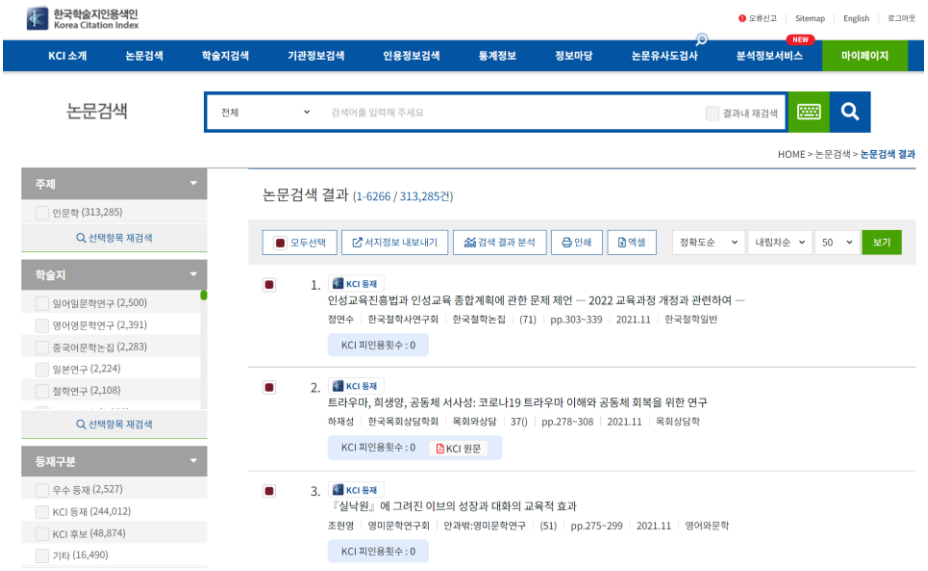

Figure 7. The results of paper search in KCI.

The thesis bibliographic data downloaded as an Excel file provides the thesis's title, author, journal, and citation year, among other basic bibliographic information. However, the file lacked essential bibliographic information such as abstracts, references, and cited works. Accessing the detail page of each paper was required to acquire insufficient bibliographic data. Both Selenium and Beautifulsoup4[6] packages were used to obtain the collection of detail pages.

[6] https://www.crummy.com/software/BeautifulSoup/

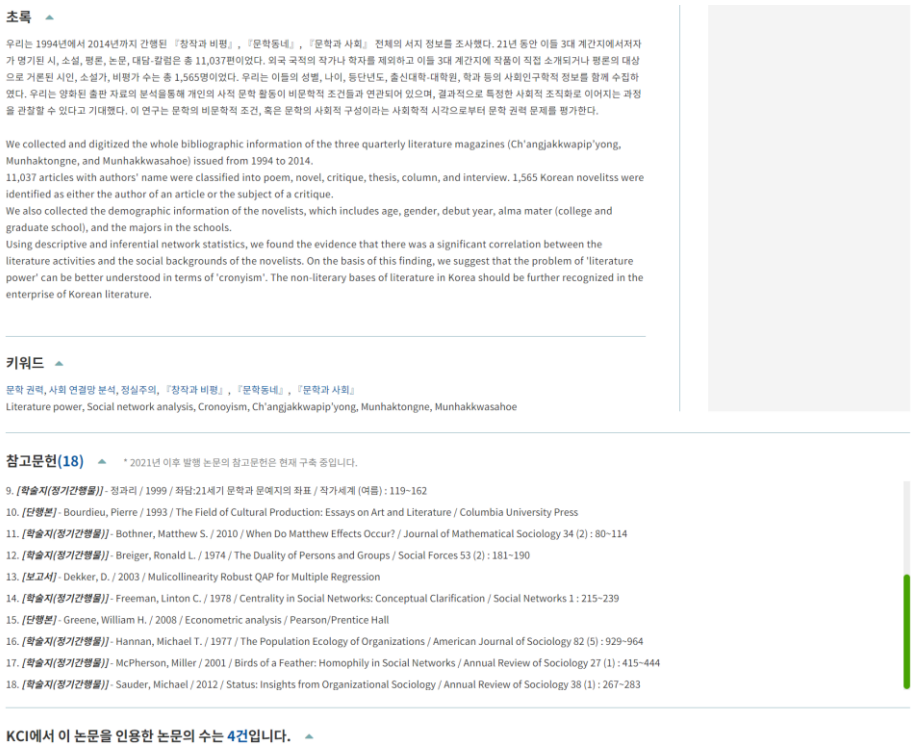

Figure 8. The paper detail page.

The KCI bibliographic data collection was conducted for around one month beginning in December 2021, and the bibliographic data of 249,661 manuscripts was collected.

B. KRI

After acquiring the bibliographic information of around 250,000 publications, KRI was used to collect the researcher information of the papers' authors. When the thesis was a collaboration between two or more people, only the first author's researcher information was obtained. Using the researcher's unique number obtained from KCI and the researcher search

function of KRI, the researcher's demographic information, such as gender, year of birth, and specific major, was retrieved. Similarly, the packages Selenium and Beautifulsoup4 were used to collect data.

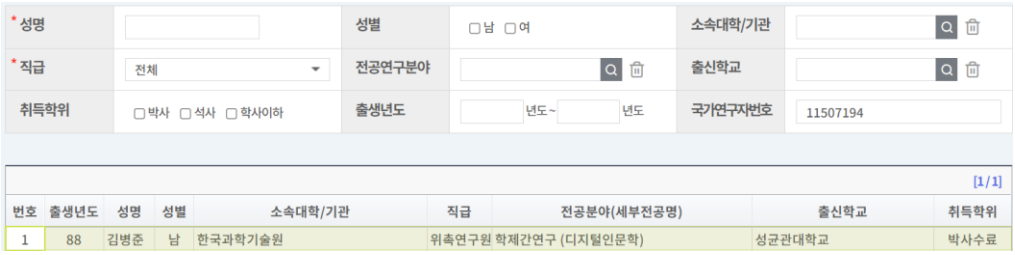

Figure 9. The researcher search page in KRI.

This data collection was carried out for about two weeks from December 2021, and the information of 28,042 researchers was secured.

2) Data preprocess

Data preprocessing was the most time-consuming aspect of this investigation. In this case, preparation comprises the data collecting method outlined above. In other words, if there were any missing or incorrectly gathered portions during data preprocessing, the collecting procedure was repeated. From August to September 2021, a roughly two-month-long preprocessing was administered.

A. KCI-KRI table integration

The process of combining the KCI bibliographic table and the KRI researcher information table into a single table was required for the research

to continue in a single table. Some international scholars or graduate students lacked a national researcher identity number; hence the researcher information column was left empty. Due to table integration, a total of 202,101 KCI papers contained KRI information.

B. Stemming

The thesis text data consists of the title, abstract, and keywords. Since only Korean data are examined in this study, English titles, abstracts, and keywords were omitted. Using kiwi[7], a Korean morpheme analyzer, nouns were retrieved from the combined title, abstract, and keyword column. Other elements of speech were excluded from the study because only nouns in academic literature were deemed to contain the concept of the text most effectively (Kim et al., 2017; Kim & Cheon, 2020). However, numerals denoting the research period (e.g., 1930, 18th century) were analyzed as nouns. Numerous Chinese characters appear in Korean humanities works, but all Chinese characters were replaced with Hangul and analyzed using Hangul. In this instance, the Python hanja[8] module was utilized. Additionally, two words that appeared consecutively more than 1,000 times were consolidated into a single word (bigram). For instance, because the terms 'Joseon' and 'era' appear frequently, 'Joseon era' is treated as a single word by employing an

[7] https://github.com/bab2min/kiwipiepy
[8] https://github.com/suminb/hanja

underbar (' '). The list of stopwords[9] was finalized by adding and removing cliches and grammatical words that occur frequently in the thesis text. For instance, it is difficult to see 'research' and 'result' as words that reflect distinct publications because they appear in practically every paper.

C. Converting to Web of science bibliographic data format

In the case of EDA and structural topic modeling, the KCI-KRI integrated table established in the preceding stage can be utilized for analysis. In the third analysis model, simultaneous citation analysis is performed with CiteSpace, but KCI bibliographic data format cannot be input at this time. It is therefore required to convert the KCI format[10] to the Web of Science bibliography format so that it can be imported into CiteSpace. First, the column names used in the Web of Science format must be standardized. For instance, the references column's name gets transformed to CR. And CiteSpace translates table data into the plain text format of Web of Sciences using a preprocessing tool.

[9] The stopwords list has been added to the appendix.
[10] https://images.webofknowledge.com/images/help/WOS/hs_wos_fieldtags.html

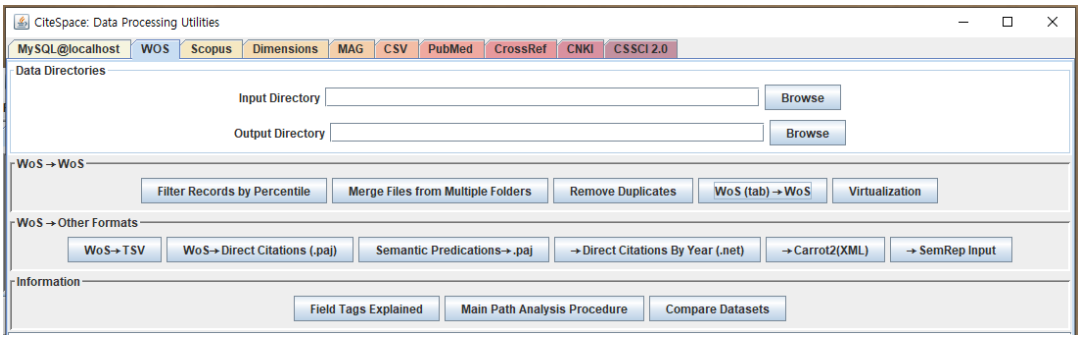

Figure 10. Data processing utilities in CiteSpace.

PT J
AU 10049568, 조
AF 10049568, 조남호
TI 주시경 제자 단군 이해
SO 仙道文化
DT Article
DE 주시경;단군;대종교;개천절;홍익인간;단군조선;기자조선
AB 주시경 제자 단군 이해 종교 문화 언어 역사 측면 주시경 일본 침략 정신 보존 결심 기독교
CR 조남호 10049568, 2014, SER000001102, VOL000058346, P0, DOI ART001908249
이현희 10128924, 2012, SER000011460, VOL000042897, P0, DOI ART001730667
오현수 10005539, 2012, 000956, VOL000040738, P0, DOI ART001704421
조남호 10049568, 2012, SER000001102, VOL000039861, P0, DOI ART001694473
김동환 10004730, 2011, SER000001102, VOL000025401, P0, DOI ART001532073
崔起榮 CRT000336797, 2003, 001258, VOL000000006, P0, DOI ART000898279
NR 6
TC 0
PU 국학연구원
PI
PA
SN
EI
J9
PD AUG 15
PY 2015
VL
IS
BP
EP
DI
PG
WC 철학
SC 기타철학일반
UT KJD:ART002028144

Figure 11. Web of Science bibliographic file format.

3) Data columns

After the data collection and preprocessing stated previously, the following data items were used for the study.

Table 4

Data Columns

Number	Column	Description
1	Id	Unique article id
2	Title	Korean title
3	Author	Author's name
4	Year	Publication year
5	Institute	Journal of institute
6	Journal	Journal name
7	Abstract	Korean abstract
8	Keywords	Korean keywords
9	Token	Nouns extracted from title, keywords, and abstract
10	Citations	The number of citations
11	References	The articles in the reference list
12	Cited papers	The articles citing the article in KCI
13	KRI number	Unique KRI id
14	Gender	Male / Female
15	Birth	The year of birth
16	Univ	Affiliation
17	Major	Specific major
18	Graduation	Graduation school
19	Diploma	Last academic degree

3. Data analysis method

1) Exploratory data analysis (EDA)

Descriptive statistics is a statistical method that identifies the features of the acquired data by summarizing, interpreting, and arranging the collected data (Mann, 2007). This contrasts with the objective of inferential statistics, which is statistical inference. Typically, fundamental indicators such as mean, standard deviation, and scatterplots can determine the nature of the analytical objective. Similar to descriptive statistics, exploratory data analysis (EDA) exists as a non-traditional statistical model inside the data science analysis process (Tukey, 1977). Prior to large-scale data science modeling, EDA utilizes data visualization using box plot, histogram, and scatter plot.

EDA has a high reputation as an index that can be consulted prior to full-scale modeling; nonetheless, a complete chapter is devoted to this investigation. It was determined that the 15 years of data included in this study were significant in and of themselves. Few studies have analyzed trends in a particular major by combining KRI and KCI data since the introduction of KCI (Jin, 2020).

The variable of greatest interest in this study is the time series variable. Specifically, I attempted to illustrate the amount of change in

various statistics from 2004[11] to 2019, when the thesis was developed in earnest at KCI. Since the KCI has had a significant impact on the Korean humanities community during the past 15 years, EDA can be used to forecast the future knowledge structure of the Korean humanities.

2) Structural topic model

This study applies the structural topic model (STM) to assess the emergence and demise of research trends in the Korean humanities. Topic modeling is a technique for evaluating the probability of occurrence of subjects and phrases in a literary group using a matrix of documents and words (Blei et al., 2003). Latent Dirichlet Allocation (LDA), the earliest and most popular approach of topic modeling, identifies latent topics in a corpus as clusters of linked words. During the writing process, the author predetermines several subjects that comprise the complete text, selects words relevant to each topic, and composes the final text. LDA is a method for identifying a topic with terms that appear in previously produced material by reversing the writing process (Jeong, 2020a).

[11] https://www.kci.go.kr/kciportal/aboutKci.kci

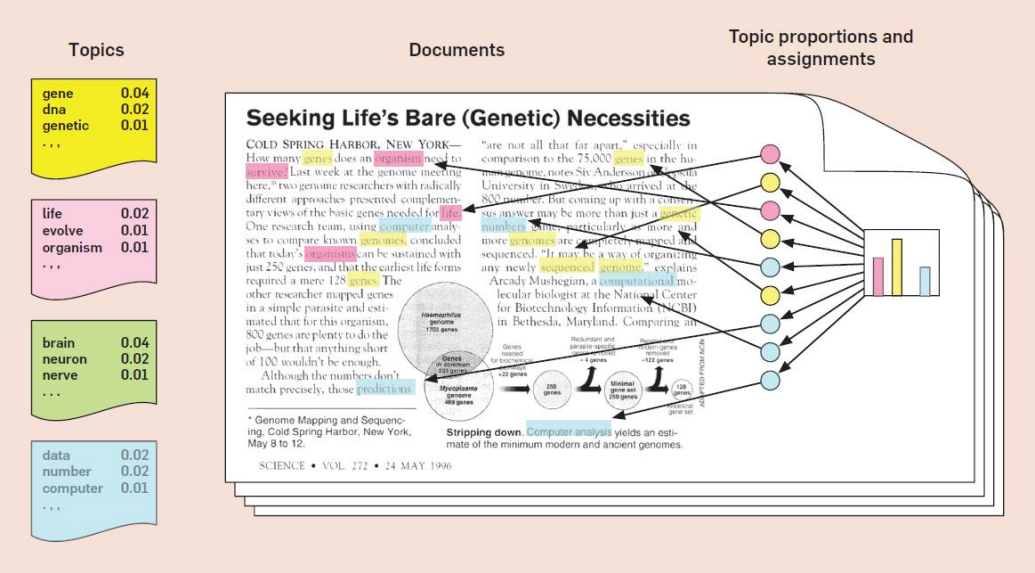

Figure 12. Training process of topic models.
Source: Compiled from Blei (2012), Figure 1.

Examining the topic proportions and assignment on the left, authors select how much weight to give each topic when creating an article (see Figure 12). The picture depicts the training process for a subject expressed in pink, yellow, and blue. Thus, topic modeling is an optimal way for discovering hidden subjects inside a big corpus of literature texts that cannot typically be recognized/noticed by humans.

In order to extract topics, the researcher is required to arbitrarily specify hyperparameters while using LDA in text mining research. Also, there are criticisms that LDA just displays the themes of the literature and cannot make additional statistical inferences; however, these criticisms are merely impressionistic. Consequently, after the introduction of LDA, topic modeling that quantified it emerged, and the STM methodology was used for this work.

The STM forecasts the occurrence likelihood of a topic in the literature's metadata (Roberts et al., 2014). In this context, the metadata of a document refers to its bibliographic data. Metadata includes information such as a document's publication date, number of citations, and author.

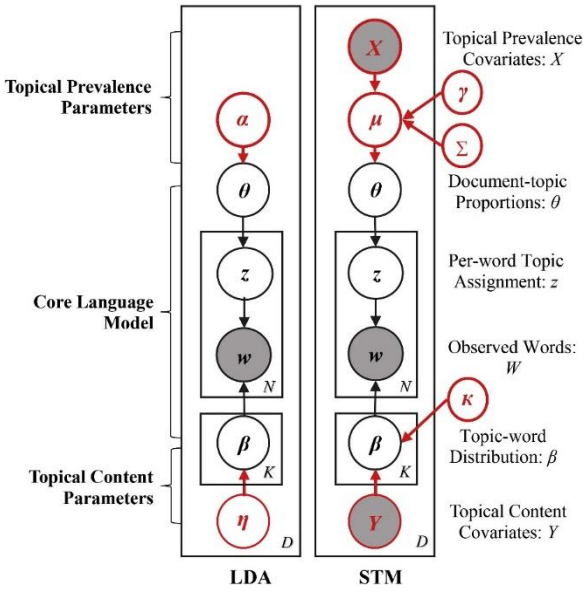

Figure 13. Plate diagram comparison of LDA and STM.
Source: Compiled from Hu et al. (2019), p.420.

In Figure 13, D in each model means the entire corpus, N is the documents constituting the corpus, w is the words in the documents, z is the topics embedded therein, and β is when a word is generated from the topic. α means the probability value of θ is a parameter related to the topic, and α is an exogenous variable that affects θ. The point where STM is differentiated from LDA is the process of forming a topic (ζ). In LDA, a pre-determined variable α forms a topic ζ through θ. In STM, a topic (ζ) is formed by

inputting literature metadata (X and μ). In this study, the year of publication of the paper was input as metadata for analysis.

3) Co-citation analysis

Co-citation analysis is one of the most frequently employed approaches for identifying the knowledge structure of a certain area, and it classifies co-cited documents in the reference list of the work. It has been demonstrated to be effective in studying the research front, an academic field with current research (Lee, 2015; Small, 1973). Co-citation analysis is easier to comprehend when contrasted with bibliographic coupling. Bibliographic coupling examines two works by linking them if they share a common citation (Kessler, 1963). For instance, in the figure below, if both document 1 and document 4 cite document 2 as a source, this is referred to as bibliographic coupling. In contrast, co-citation is a citation connection from a reference's perspective. In the diagram below, Document 1 mentioned Documents 3 and 2 collectively. In other words, Document 3 and Document 2 appear together in the reference list of Document 1, and the relationship between them is known as a co-citation. In this instance, the co-citation or bibliographic link could be with an author or a journal instead of a document. However, Literature was the subject of examination in this study. This technique is known as document co-citation analysis (DCA). This is because the objective is to determine

which works have been deemed significant in the Korean humanities, particularly on the knowledge front, which means areas where the latest research is actively conducted .

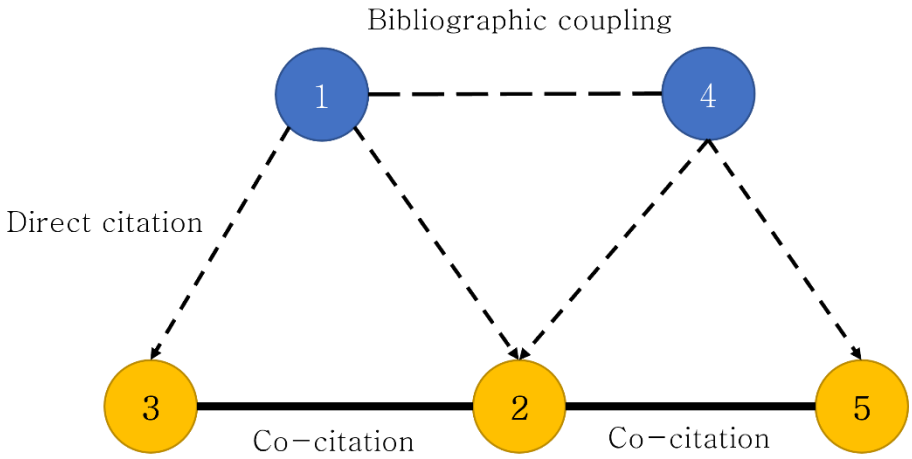

Figure 14. Type of citation network[12].

A. CiteSpace

CiteSpace[13] is a software for bibliometric analysis and visualization developed by Professor Chaomei Chen. It is a tool optimized for simultaneous citation analysis, such as literature citation analysis and author simultaneous citation analysis, and it is also capable of predicting future academic trends by assessing research trends and the uncertainties of scientific knowledge in a certain subject. Recently, trends in COVID-19-related publications and types of illness uncertainty were studied (Chen, 2020) (see Figure 15).

[12] https://pythonhosted.org/tethne/tutorial.bibliocoupling.html
[13] https://citespace.podia.com/

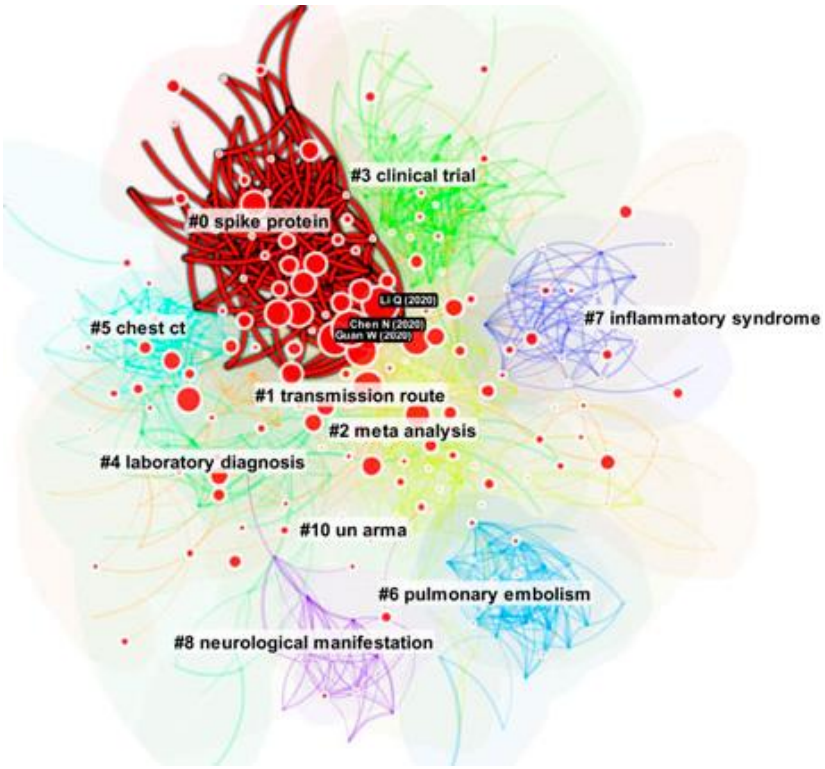

Figure 15. Clusters of COVID-19 literature.
Source: Compiled from Chen (2020), p.11.

CiteSpace's philosophy is also associated with Thomas Kuhn's theory of paradigm shift. In the opening to their book, Chen and Song assert that the growth of scientific knowledge is not linear but rather extremely complex, and that there is also a point of revolutionary transformation (Chen & Song, 2017). And Chen and Song asserts that it is possible to quantitatively demonstrate Thomas Kuhn's paradigm shift with the development of bibliographies and big data processing technology, such as CiteSpace. This study will apply their theories to the subject of Korean humanities, despite the fact that the numerous situations from Chen and Song's books center on science and technology.

Chapter 4. Results

This chapter summarizes the analysis results of the three research models. There is a descriptive statistical analysis, STM, and co-citation analysis. Each study includes an Introduction that discusses the model, Results that explain the analysis outcomes, and a Conclusion that interprets the model results. In addition, the R/Python analysis code for the three models was published on Github[14].

1. Study1: EDA

1) Introduction

A total of 249,661 KCI humanities publications were analyzed during the data preprocessing procedure, and the data are current as of 30 December 2021. A total of 28,042 demographic and sociological backgrounds were gathered by gathering the information of one author and the researcher for these works from KRI. The figure is 202,101 excluding documents lacking KRI information. Regardless of the availability or absence of KRI information, all data were included in the

[14] https://github.com/ByungjunKim/KnowledgeStructureOfKoreanHumanities

descriptive statistical analysis of KCI articles. Only papers containing KRI-related material were included in the descriptive statistical analysis related to KRI. Following is a summary table of the aforementioned conditions.

Table 5

The Scope of Analysis

Condition	Frequency
The number of KCI articles	249,661
The number of the first authors with KRI information	28,042
The number of articles with KRI id of the first author	202,101

2) Results

A. KCI

i. Publication

Since 2004, the number of papers published each year has increased, from 9,794 in 2004 to 18,658 in 2015 with a minor dip – to 17,822 – in 2019. These figures demonstrate the expansion of the Korean humanities.

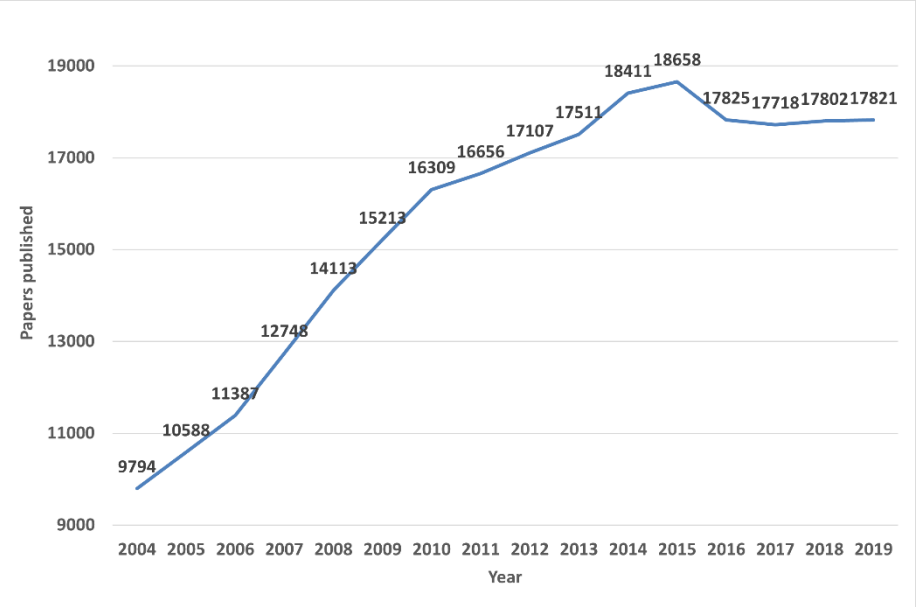

Figure 16. Papers published by year.

Using the KRI number, the number of publications per researcher each year was derived. Only publications with a confirmed KRI researcher number were included in the statistical analysis. In 2004, the average number of papers published per individual was 1.61; in 2009, it was 1.61; and in 2019, it was 1.56.

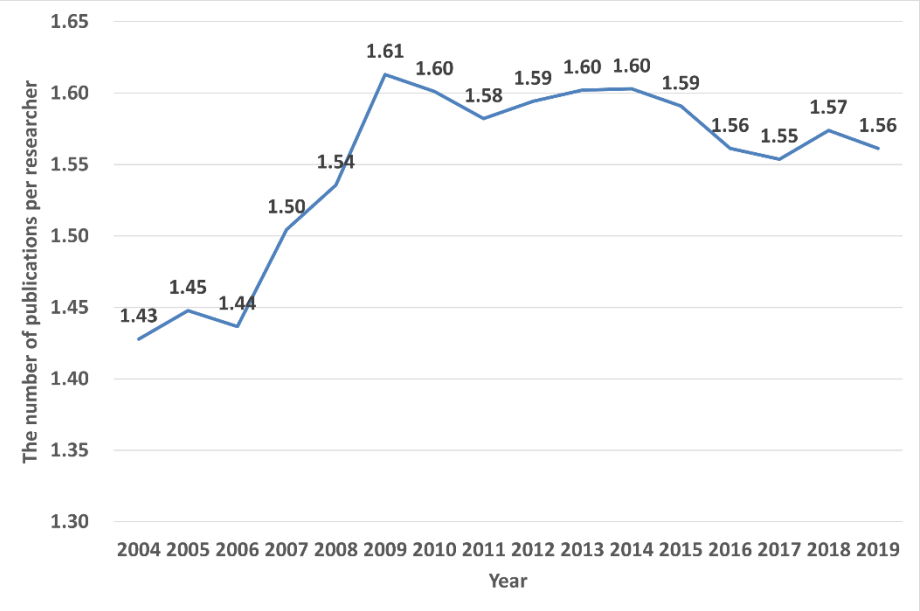

Figure 17. The annual average number of publications per researcher.

KCI includes medium and small categories in addition to the major categories ('Humanities,' 'Social Science,' 'Engineering,' etc.). The table[1516] below displays the quantity and proportion of humanities papers published in the medium category. The top three medium categories ('Korean and Literature,' 'History,' and 'English and Literature') represented almost 42% of the total. When considering the proportion of publications that study Korean history, such as 'Korean Language and Literature' and 'History,' it was determined that Korean studies–related studies accounted for the biggest proportion.

[15] All medium categories with less than 500 items are classified as "Other."
[16] The KCI's medium classification system lacks the support of humanities scholars. Specifically, academic journals that are difficult to identify are included to the categories 'other humanities,' 'humanities,' and 'literature.'

Table 6

Papers Published by Medium Category[17]

Category	Freq	Percent	Category	Freq	Percent
Korean L&L	46,239	18.52%	Religious studies	4,419	1.77%
History	37,246	14.92%	Education	3,246	1.30%
English L&L	21,318	8.54%	Buddhism	3,121	1.25%
Other humanities	20,708	8.29%	Other	2,938	1.18%
Philosophy	18,074	7.24%	Russian L&L	2,766	1.11%
Chinese L&L	16,757	6.71%	Interpretation & translation	1,875	0.75%
Japanese L&L	15,909	6.37%	Confucianism	1,413	0.57%
Christian theology	12,226	4.90%	Oriental L&L	875	0.35%
Linguistics	11,796	4.72%	Catholic theology	866	0.35%
Humanities	9,284	3.72%	Geography	676	0.27%
German L&L	6,146	2.46%	Interdisciplinary studies	628	0.25%
French L&L	5,489	2.20%	Spanish L&L	617	0.25%
Literature	5,029	2.01%	Total	249,661	100.00%

This study also examined the number of articles that fell into the medium category by year. As shown in the graph below, the number of papers published in the fields of 'Korean Language and Literature' and 'History' has consistently increased, with 3,327 articles published in 'Korean Language and Literature' and 2,871 papers published in 'History' as of 2019. This constitutes around 35 percent of all papers published in 2019. Since 2011, 'Other Humanities' has surpassed 'English and Literature' to take third place.

[17] L&L means 'Language and Literature' .

However, the decline in 'English and Literature' and the classification of all academic journals that are difficult to include in any academic division within the KCI system as 'Other Humanities' could be considered a phenomenon.

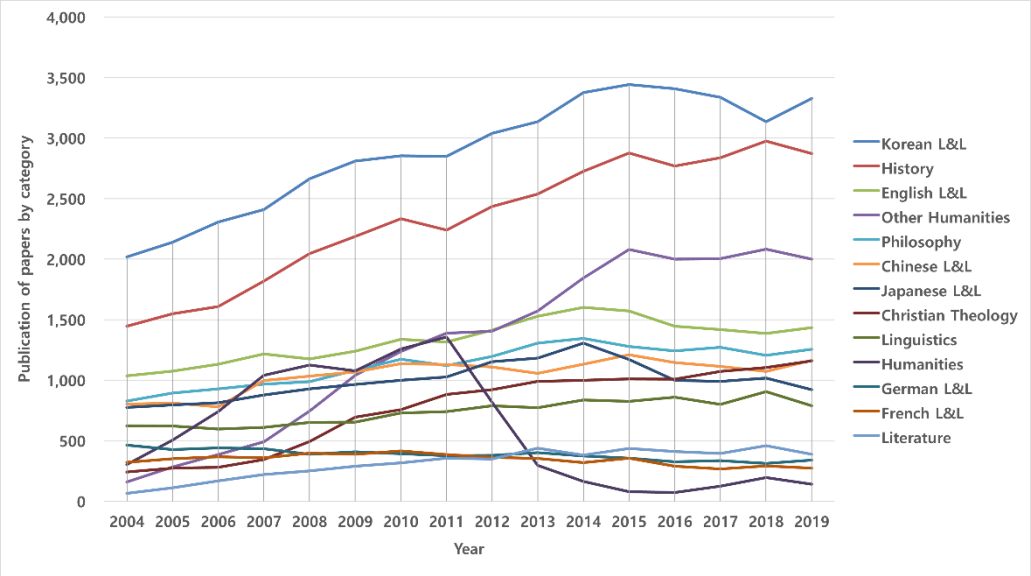

Figure 18. The annual publication of papers by category.

ii. References

On every paper description page, KCI gives a list of references. Although not every publication has a reference list, the majority of papers published since 2008[18] have a well-established reference list and can be utilized as statistics. The graph below displays the annual average number of references. In 2008, publications cited an average of approximately 25 references, and this number continues to climb; in 2019, papers cited an

[18] From 2004 to 2007, more than 1,000 papers without a reference list appeared every year among published papers, so they were excluded from the analysis.

average of about 29 references. It is logical for the number of citations to increase as the KCI's annual thesis output rises. Most likely, academic databases such as RISS, KCI, and DBpia (https://www.dbpia.co.kr/) have provided researchers greater access to academic literature, causing them to cite more sources in their works than previously.

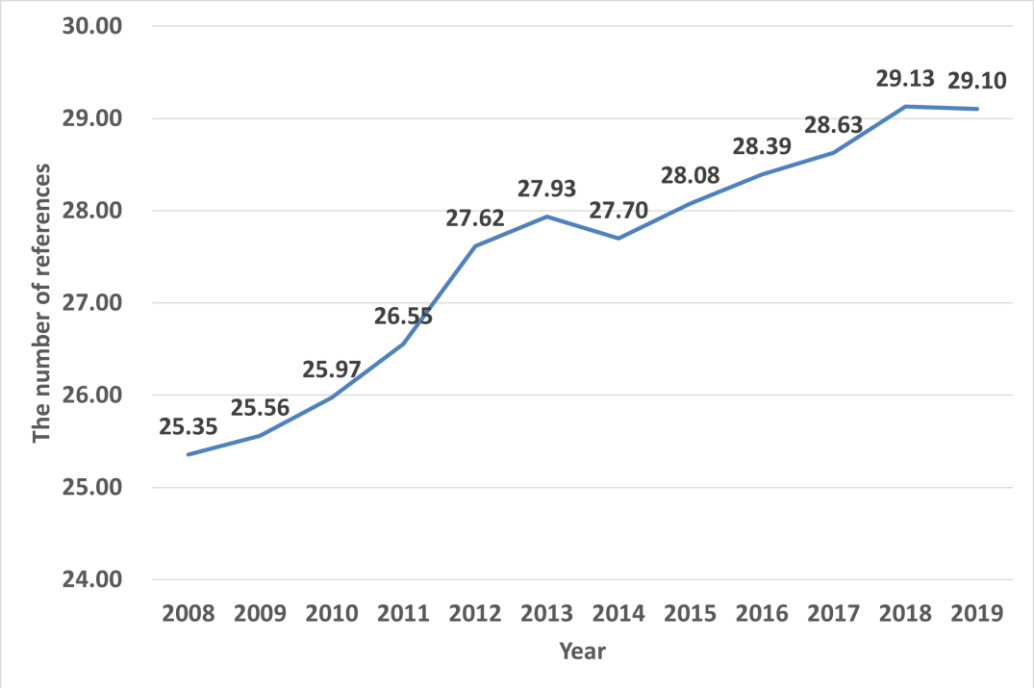

Figure 19. Average number of references by year.

As the number of references continues to rise, it is vital to assess which types of references are mentioned the most. Six distinct sorts of references have been defined in the KCI. Observing the graph below, it can be seen that the proportion of books among references was approximately 53% in 2008, but continued to decline until it reached 44% in 2019. In contrast,

journal papers (including conference papers) accounted for approximately 27 percent in 2008, but have progressively risen to approximately 39 percent in 2019. If humanities and academic communications in Korea were traditionally delivered mostly in books, this is rapidly changing to academic articles. It is also related to the increase in citations discussed above. The increase in the proportion of journal papers contributed to the rise in citations among humanities scholars. Furthermore, since anyone can quickly look for and access journal papers, their significance increases naturally.

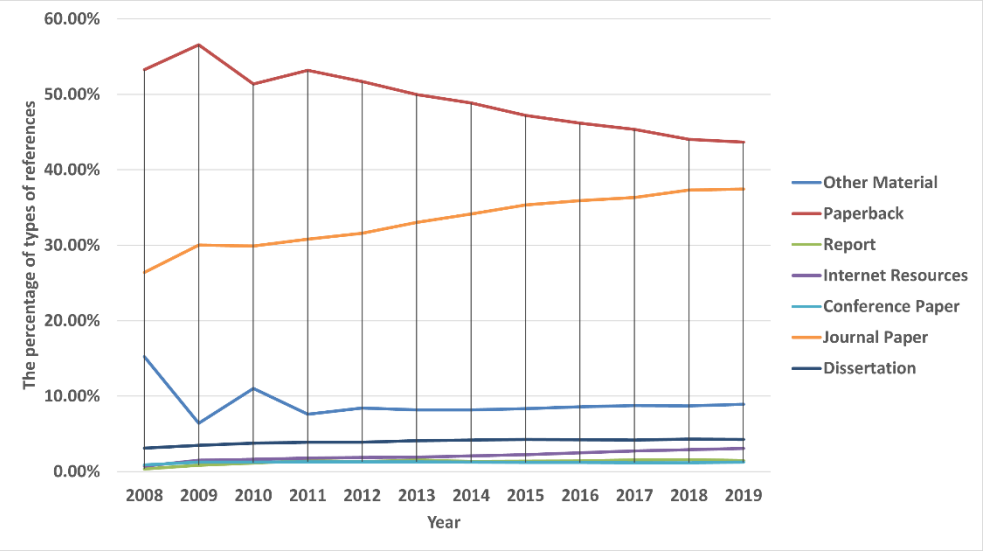

Figure 20. Trends in the percentage of types of references by year.

iii. Citation

Citation is an indicator of a paper's, researcher's, or journal's influence in academic communication. Additionally, it might be understood in terms of academic communication as an indicator of active journal distribution. The KCI only provides citation information in instances where KCI papers cite KCI

articles. The average number of citations for 249,661 KCI humanities works is 3.44, median is 2, standard deviation is 5.28, minimum is 0, and maximum is 404. Since the median value is 2, almost half of the papers acquire fewer than two citations, resulting in a power law distribution[19] with a relatively long right tail, as depicted in the following figure.

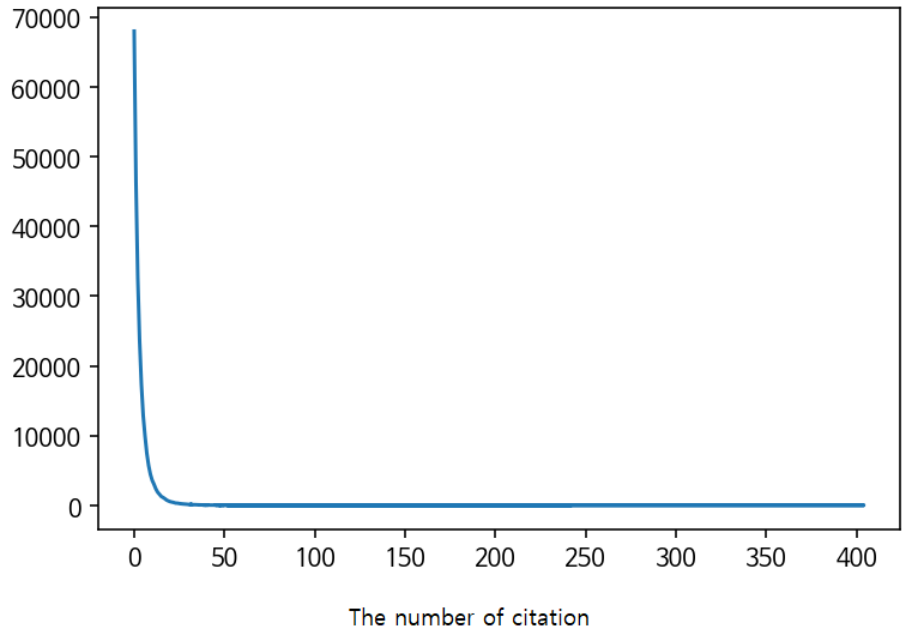

Figure 21. Histogram shows the number of citations to papers (X-axis: the number of papers).

Since the number of citations increases with the length of time since the paper's publication, it is important to normalize it according to the paper's period of publication. Given that KCI journals typically release new issues every quarter, the number of citations was normalized based on the number of

[19] https://en.wikipedia.org/wiki/Power_law

quarters since the publication date of the paper. For instance, if the number of citations for a paper published in December 2019 is 2, the normalized number of citations is 0.25[20], as eight quarters (two years) have passed since the date of data collection (December 2021). The calculation of the normalized number of citations yielded a mean of 0.1, a median of 0.06, a standard deviation of 0.14, a minimum of 0 and a maximum of 6.52. A normalized citation count average of 0.1 indicates that citations occur every 10 quarters on average (2.5 years). Although a comparison with social science and natural scientific publications is required, it is evident that the rate of knowledge transmission in humanities papers is modest. As shown in the graph below, the normalized average number of citations per year is rising rapidly. This indicates that more recent articles are mentioned more frequently, and that the number of citations increases along with the number of KCI papers produced.

[20] 2/8 = 0.25

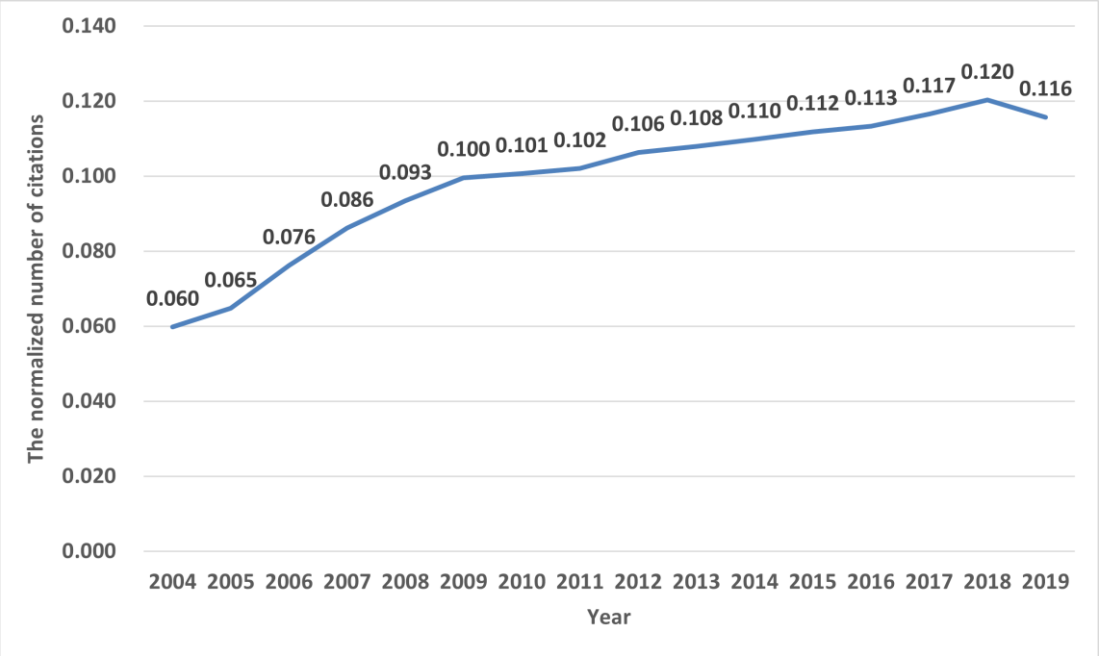

Figure 22. Average normalized number of citations by year.

The indicator to focus on in the following table is the standard deviation. As the average number of normalized citations climbs each year, so does the standard deviation, indicating that citations are becoming increasingly polarized. In other words, a small number of papers receive the majority of citations.

Table 7

Normalized Statistics for Citation Count by Year

Year	Frequency	Mean	SD	Min	25%	50%	75%	Max
2004	9,794	0.060	0.096	0.000	0.000	0.029	0.074	2.143
2005	10,588	0.065	0.102	0.000	0.000	0.031	0.078	2.621
2006	11,387	0.076	0.132	0.000	0.016	0.033	0.098	6.516
2007	12,748	0.086	0.121	0.000	0.017	0.052	0.107	2.203

2008	14,113	0.093	0.133	0.000	0.019	0.056	0.127	4.566
2009	15,213	0.100	0.134	0.000	0.020	0.060	0.137	2.458
2010	16,309	0.101	0.132	0.000	0.022	0.065	0.133	1.761
2011	16,656	0.102	0.136	0.000	0.024	0.050	0.143	4.524
2012	17,107	0.106	0.139	0.000	0.026	0.056	0.139	2.528
2013	17,511	0.108	0.137	0.000	0.029	0.063	0.147	2.429
2014	18,411	0.110	0.141	0.000	0.032	0.069	0.143	4.267
2015	18,658	0.112	0.147	0.000	0.000	0.077	0.160	2.778
2016	17,825	0.113	0.150	0.000	0.000	0.050	0.150	4.571
2017	17,718	0.117	0.155	0.000	0.000	0.063	0.167	2.188
2018	17,802	0.120	0.162	0.000	0.000	0.077	0.167	3.067
2019	17,821	0.116	0.168	0.000	0.000	0.091	0.182	2.273

This study investigated the citations and normalized averages of citations by medium category. In the table below, 'Educational Studies' and 'Interpretation and Translation' are rated first since their major is a small subject with fewer than 2,000 published articles, and a few works have increased the average value. It is also possible that citations to the topic came from disciplines outside the humanities (social science, natural science, etc.). As described previously, 'History' and 'Korean Language and Literature' are the categories in which the greatest number of papers are produced, but they also have the highest citation rankings, indicating active academic contacts. In

contrast, over the past 15 years, 'English and Literature,' 'Japanese language and literature,' and 'Chinese language and literature' produced more than 15,000 publications, although citations were low. Due to the fact that the relevant topic is foreign language literature, there are few citations between KCI works and numerous citations from foreign journals.

Table 8

Citation/Normalized Citation Average by Medium Category

Category	Citation	Normalized Citation	Category	Citation	Normalized Citation
Education	6.120	0.165	Religious studies	2.814	0.081
Interpretation & translation	4.701	0.162	Regional studies	1.651	0.080
History	**4.695**	**0.144**	**English L&L**	**2.014**	**0.060**
Korean L&L	**5.005**	**0.141**	Catholic theology	1.564	0.059
Other	3.761	0.123	**Japanese L&L**	**1.644**	**0.053**
Humanities	5.143	0.112	Oriental L&L	1.697	0.053
Christian theology	2.986	0.109	German L&L	1.816	0.048
Linguistics	3.873	0.108	**Chinese L&L**	**1.637**	**0.048**
Other humanities	3.010	0.105	French L&L	1.648	0.043
Philosophy	3.231	0.092	Russian L&L	1.583	0.042

Buddhism	2.516	0.089	Interdisciplinary studies	1.790	0.037
Literature	2.506	0.085	Spanish L&L	0.922	0.032
Confucianism	2.980	0.082	Total	3.447	0.102

iv. Self-citation

Self-citation is the practice of citing one's own previous works. The KCI generates the Impact Factor as an index for measuring the effect of journals; however, self-citations are disregarded to ensure a fair evaluation. The rate of self-citation is also a component of the KCI journal management evaluation metric. The academic perspective on self-citation is inconsistent. One side thinks that self-citation is a 'acceptable' act to summarize the researcher's previous study, while the other side believes it is a 'inappropriate' move to purposely enhance the quantitative indicator of himself or herself and the journal (Ioannidis, 2015). Self-citation is employed as an indicator of scholarly communication for analysis in this study. Increased self-citations indicate that the researcher has conducted extensive research in the past, but also other researchers are not interested in his or her research.

Even if only one author overlapped in the papers citing each other, it was considered self-citation for the sake of this research. In other words, if thesis A was mentioned from thesis B, regardless of whether the first authors

of each thesis's author coincide or not, and if two theses were the result of collaborative research, even if just one author of the theses overlapped, it was considered self-citation. The mean number of self-citations is 0.497, the median is 0, and the standard deviation is 1.149, as shown in the table below. Similar to citations, the majority of papers recorded 0 or 1 self-citations, while just a few publications had self-citations in excess of ten.

Table 9

Statistics by Citation Type

Statistic	Citations	Self-citations	Excluding self-citations from citations
Mean	3.447	0.497	2.950
Median	2	0	1
SD	5.284	1.149	4.971
Min	0	0	0
Max	404	28	404
Freq	249,661		

Similar to citations, the number of normalized self-citations is presented each year. As indicated in the graph below, the average normalized number of self-citations increased annually as well.

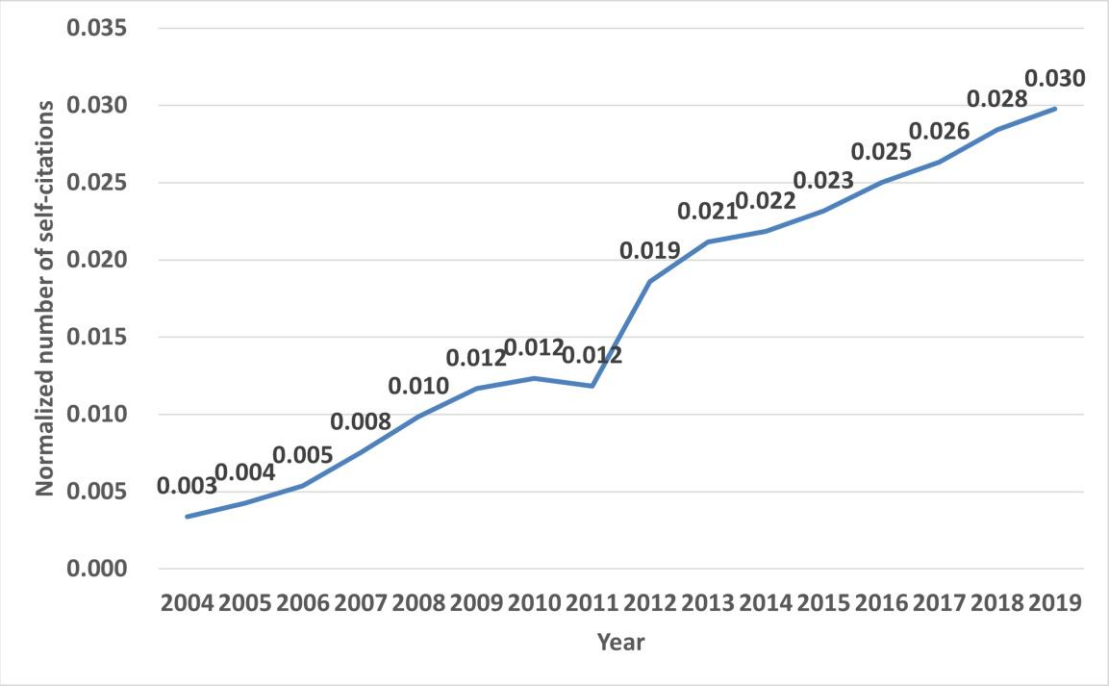

Figure 23. Average normalized number of self-citations by year.

The proportion of publications that had just self-citations among those that were cited several times (a total of 181,698 cases) is seen in the graph below. The fraction of articles containing just self-citations has steadily climbed each year, reaching approximately 18 percent in 2019. Obviously, the more recently published an article is, the more likely it is to contain exclusively self-citations. However, the rise in the fraction of papers containing just self-citations renders journals incapable of serving as a public forum. This is because active scholarly communication increases the likelihood that the most recent works will be actively cited.

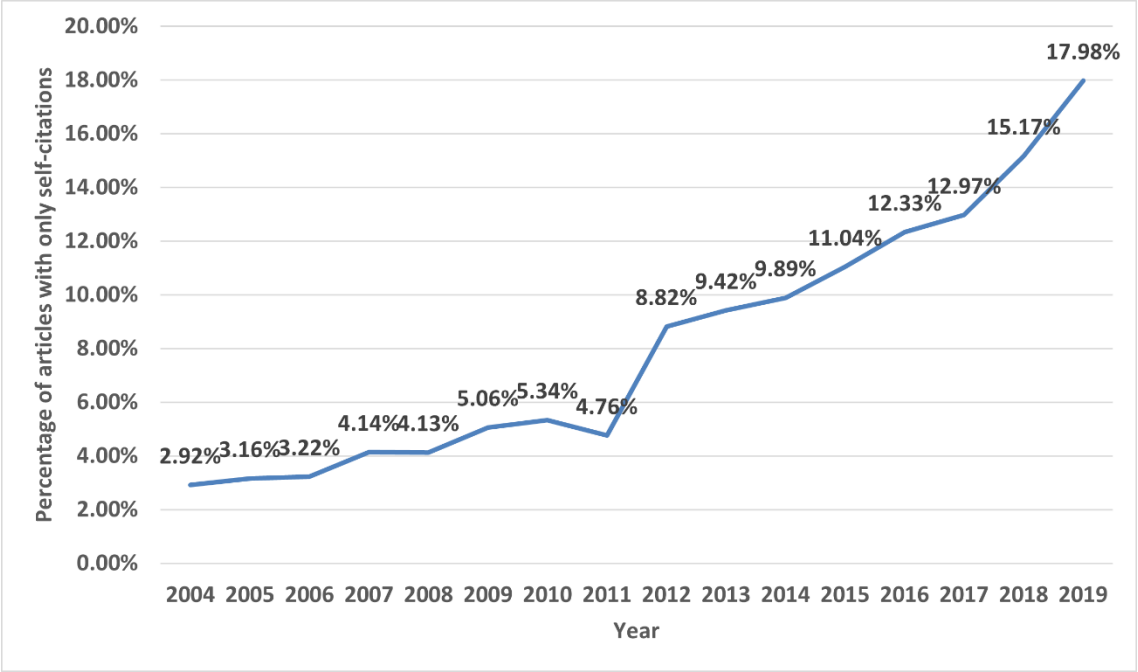

Figure 24. Percentage of articles with only self-citations by year.

B. KCI - KRI

i. Gender

Of the 28,042 researchers whose KRI numbers were recognized, 27,939 had gender markers, including 15,803 males (56.6%) and 12,136 females (43.4 percent). If the ratio is estimated based on the number of publications rather than the number of male and female researchers who have published papers, the number of males is 126,514 (62.8 percent) and the number of females is 74,807. (37.2 percent). The year-by-year trend of thesis production by gender is depicted in the following graph. In 2004, the ratio of male to female thesis production was 70:30, but by 2019, the ratio had decreased to 57.4:42.6%. This appears to be a result of the steady influx of

fresh female scholars (see Figure 25).

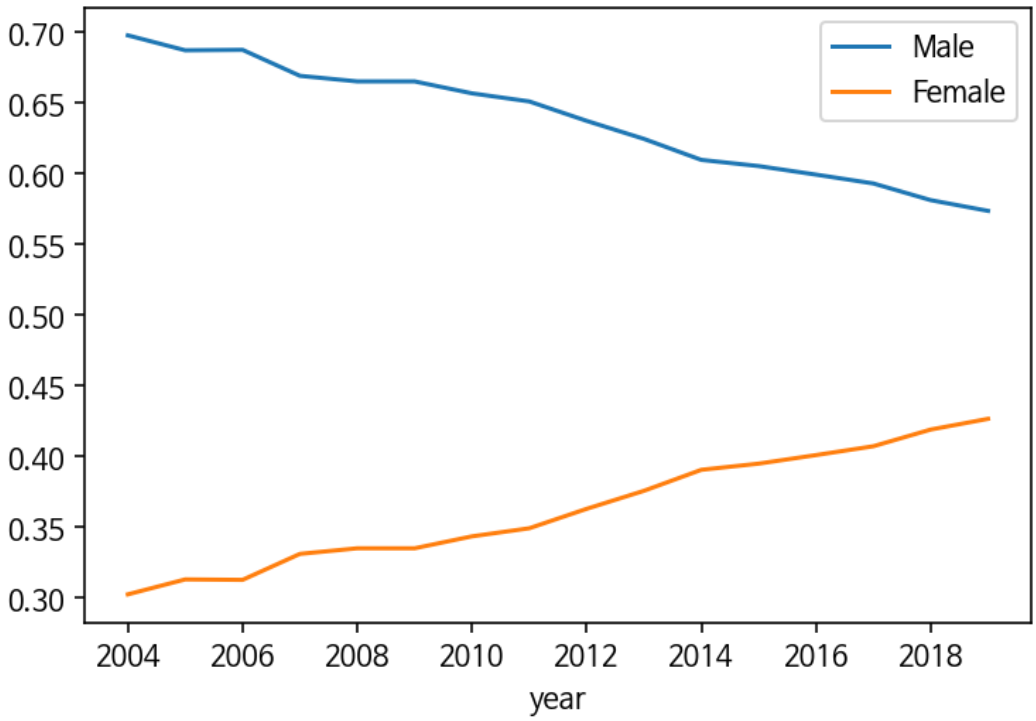

Figure 25. Ratio of male and female's paper production by year (X-axis: the percent of papers).

ii. Age

A statistical analysis of the ages of 27,823 researchers whose birth year was listed in the KRI was conducted (see Figure 26). Here, age refers to the age at the time of publication. Thus, a person born in 1988 who published a paper in 2019 was considered 31 years old. First, I examined the number of researchers who participated by age and year. As indicated in the graph below, people in their 40s, who have comprised over 40 percent of the population since 2004, will fall below 30 percent for the first time in 2015 and

be overtaken by those in their 50s in 2019. As of 2019, more than 80 percent of researchers are in their 40s or older, a trend that is accelerating significantly.

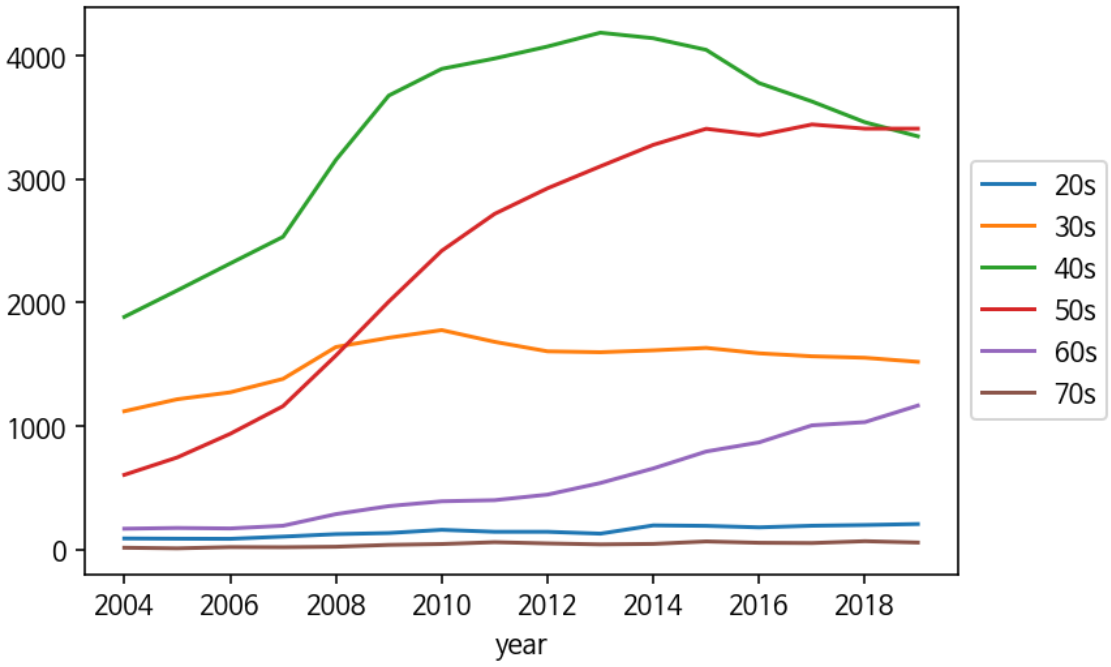

Figure 26. Trends in the number of researchers by age group (X-axis: the number of papers).

The same holds true for the number of published papers by age. In 2009, the number of papers published by individuals in their 50s exceeded that of those in their 30s, as indicated in the graph below (see Figure 27). As of 2019, the percentage of thesis authors in their 40s is approximately 37% and the percentage of thesis authors in their 50s is approximately 35%.

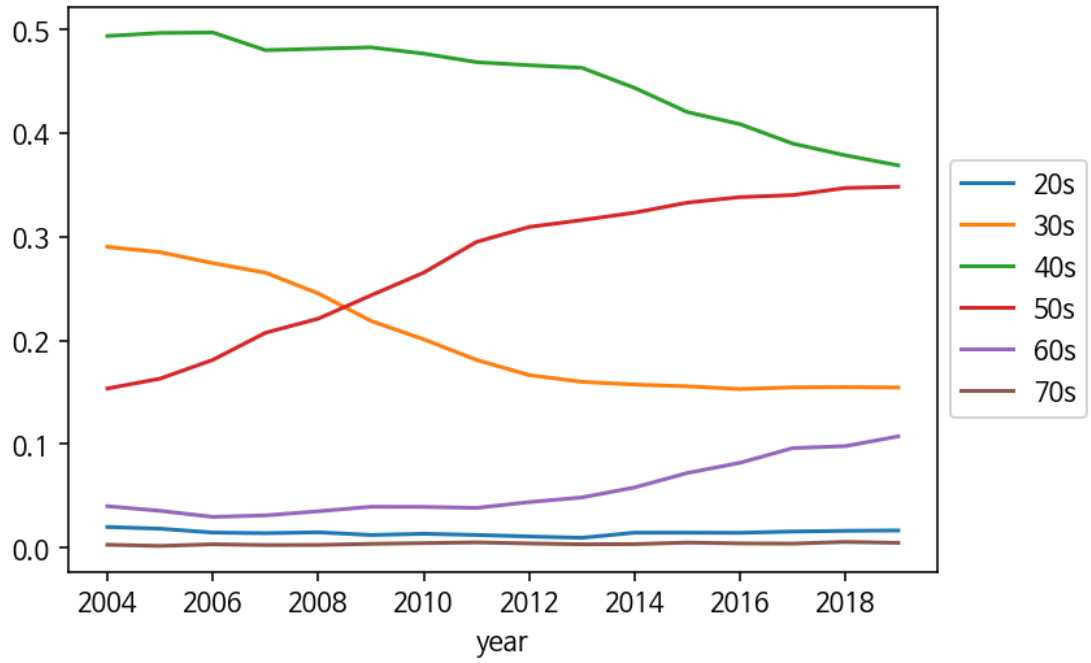

Figure 27. Trends in the percent of publication by age group (X-axis: the percent of papers).

iii. Generation

The generations were categorized by birth year, and the analysis was conducted similarly to that of the age groups. As indicated in the graph below (see Figure 28), persons born in the 1960s comprised the biggest proportion of the population after 2004. In 2011, the number of researchers born in the 1970s outnumbered those born in the 1950s, and in 2018, the number of researchers born in the 1980s outnumbered those born in the 1950s. And as of 2019, the percentage of researchers born in the 1960s is approximately 35.1%, while the percentage of researchers born in the 1970s is

approximately 34.5%; therefore, it is anticipated that the number of researchers born in the 1970s will soon surpass those born in the 1960s.

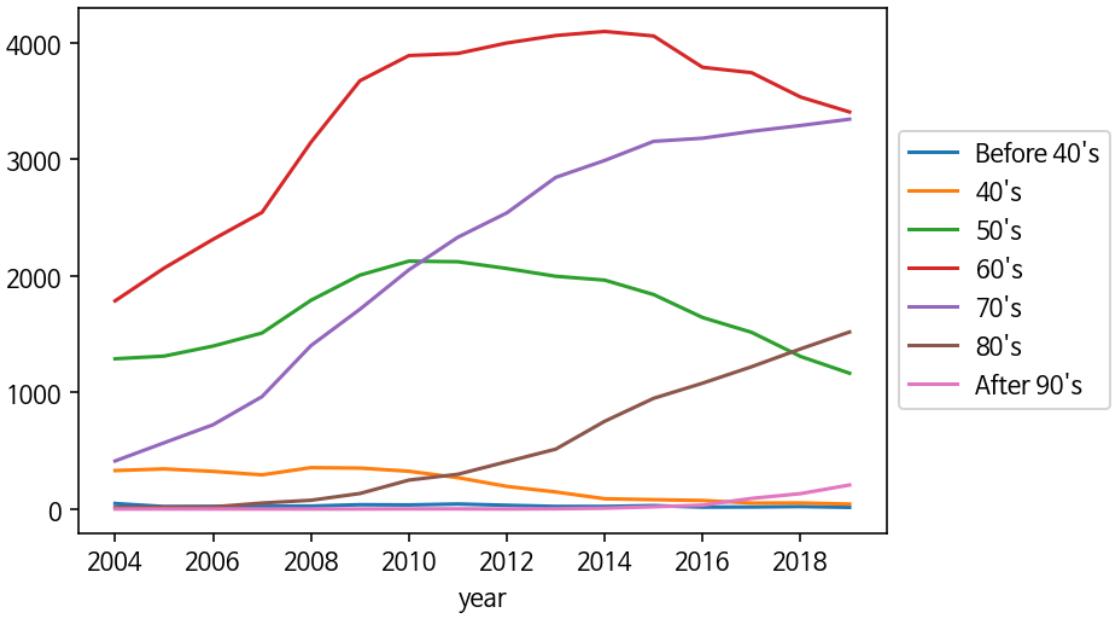

Figure 28. Trends in the number of publications by generation (X-axis: the number of papers).

The trend of the number of researchers by generation is paralleled by the number of papers produced by generation. As demonstrated in the graph below (see Figure 29), scholars born in the 1960s regularly placed first in terms of manuscript production; but, in 2019, they reversed the proportion of theses published by those born in the 1970s. In their analysis of Korean

modern literature, Lee & Kim (2020) also found a substantial proportion of the 586 generation[21] in scholarly fields.

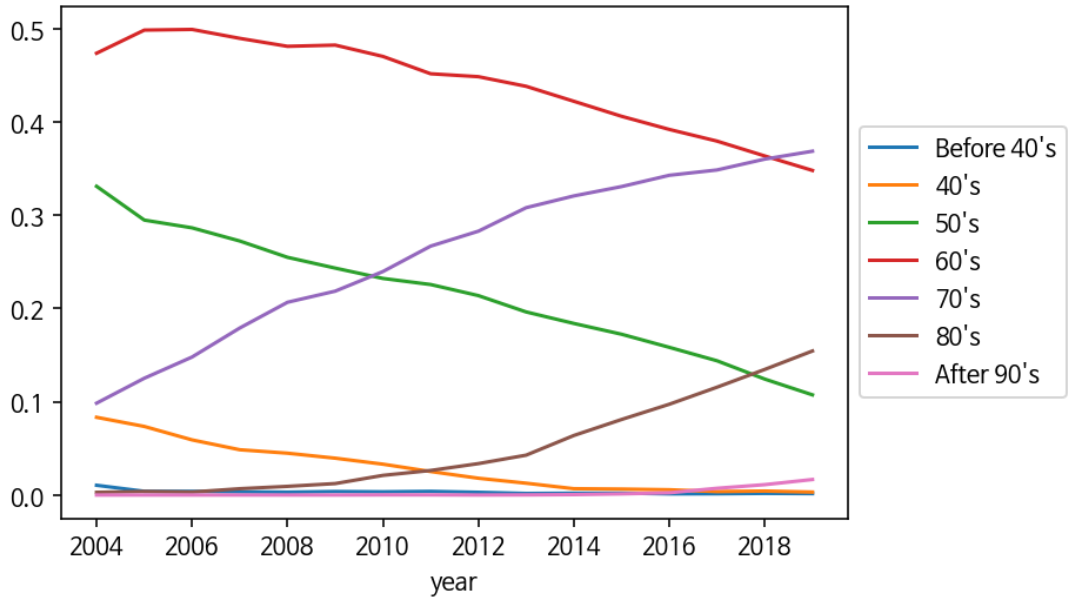

Figure 29. Trends in the percent of publication by generation (X-axis:the percent of papers).

3) Conclusions

In the first study, the knowledge structure of the Korean humanities was determined using descriptive statistical analysis on the basis of bibliographic data from KCI humanities papers and information from KRI researchers. Here, I discovered that the knowledge structure of the Korean humanities had the following features. After the establishment of KCI 15 years ago, the Korean humanities community has seen an increase in the number of papers and

[21] As of 2021, it refers to those in their 50s, those admitted to college in the 1980s, and those born in the 1960s.

researchers. The publication output, the number of researchers involved, citations, and references continued to rise. Second, it is questionable whether the qualitative expansion of the humanities in Korea was encouraged despite the numeric expansion. There was a polarization of articles and researchers according to the results. Only a small number of papers are cited, the trend toward specific academic fields such as Korean studies is ongoing, and the rate of self-citations is gradually increasing. Despite their quantitative expansion, it is now time to question whether the KCI system serves as a forum for the academic community. Thirdly, differentiation according to the researcher's demographic and sociological background was discovered. The disparity in paper publication volume based on gender has steadily diminished, while the proportion of a certain age and generation still constituted the majority. In addition, a rapid aging of researchers was observed. As a result of state-led humanities promotion policies such as BK and HK, the humanities community in Korea has been able to develop the output of 'thesis production.' However, the qualitative growth of scholarly communication and academic virtuous cycle through the next generation of academics has not kept pace with the production of outcomes.

2. Study2: Structural topic model

1) Introduction

In Study2, the Structural Topic Model (STM) was utilized to study the evolution of research topics using 249,661 publications analyzed in Study1. The text preprocessing and modeling process of the bibliographic data of papers is described below.

i. Text preprocessing

Python was used to perform text preparation by separating the operation into three steps (see Figure 30). First, the text-related KCI paper columns 'title,' 'keywords,' and 'abstract' were combined into a single text. At this time, titles were present in all publications, but keywords and abstracts were absent in around 8 percent[22] of instances, and just the titles were included in this case. Using the kiwi morpheme-analyzer, only nouns were retrieved from the Korean text column prepared in step 1 using only nouns. Kiwi features a user dictionary capability that enables the user to individually add proper names and abbreviations that are not recognized by the morpheme analyzer. Thirdly, bigrams that appear over a thousand times sequentially among the extracted nouns were grouped into a single noun. Also, terms that appeared in nearly all paper manuscripts were deemed stop words and

[22] 21572 cases without keywords and 21222 cases without abstract.

eliminated (see Appendix 1). In addition, the Hanja[23] module was used to transform all Chinese characters to Hangul.

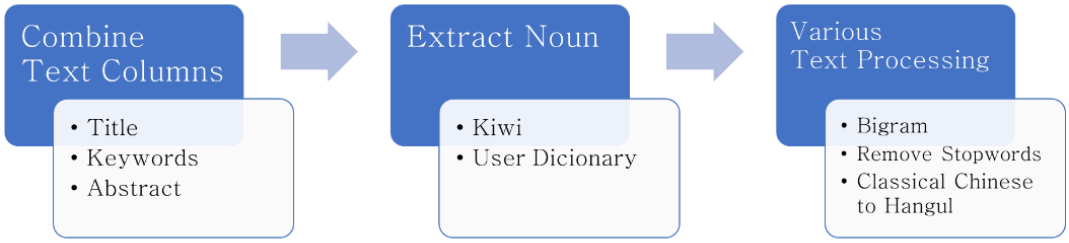

Figure 30. The text preprocesses.

ii. Modeling

Text was imported into R for modeling after being preprocessed with Python. STM is distinguished by its capacity to utilize document meta-data as covariates. The year was included as a covariate in this study's model. The model formula looks like this:

$$Topic = s(year)$$

In the following formula, the dependent variable (T) is the distribution of the topic, and the covariate (year) is not entered directly, but rather into

[23] https://github.com/suminb/hanja

the s function of the B-spline. The STM package's B-spline function curves a continuous variable as a straight line (Kamari et al., 2019). This subsequently helps to reveal variations in the proportions of topics throughout time.

Since topic modeling is unsupervised learning, hyperparameters are crucial. Specifically, K, the number of topics, should conform to the standard. searchK is a function made available by the stm package. A researcher can establish the ideal number of topics according to four criteria (Held-Out Likelihood, Residuals, Semantic Coherence, and Lower Bound). When depicted in the graph below, the model's evaluation improves as all variables, excluding residuals, increase. Semantic coherence specifically refers to the cohesion of content within a topic. When the number of topics reaches 15, it increases again, so 15 might be considered the optimal number of topics.

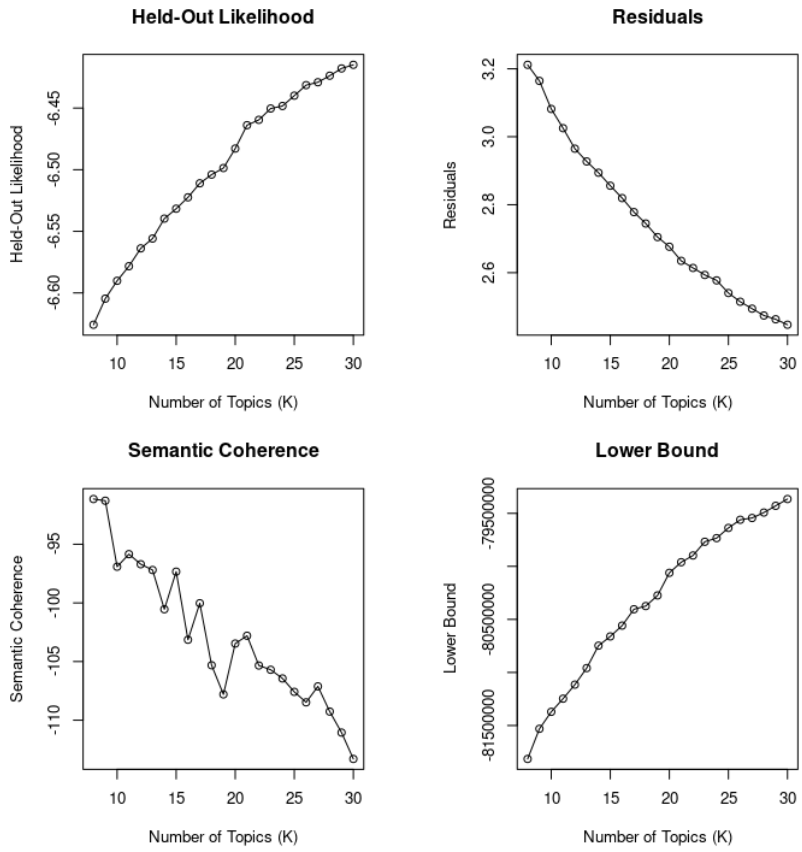

Figure 31. Diagnostic values by number of topics.

The hyperparameters of the model used in the study are as follows.

Table 10

Hyperparameters of STM

Hyperparameters	Value
Minimum number of documents	1,000
The number of topics (K)	15
Metadata	Year
Init type	Spectral

2) Results

Four total keyword weights are provided by STM (Highest Probability, FREX, Lift, and Score). In this scenario, the recommended package weight is

Frequency and Exclusivity (FREX). FREX is a method for weighting words that demonstrate exclusivity between topics while taking into account their frequency. For instance, it is a strategy for providing greater weight to distinctive terms that are uncommon in other topics among words that occur frequently within a topic. The ratio of topics' keywords to those chosen based on FREX is depicted in the graph below (descending order). Topic 12 had the highest proportion, accounting for almost 10%, while Topic 6 had the lowest proportion, accounting for approximately 4%.

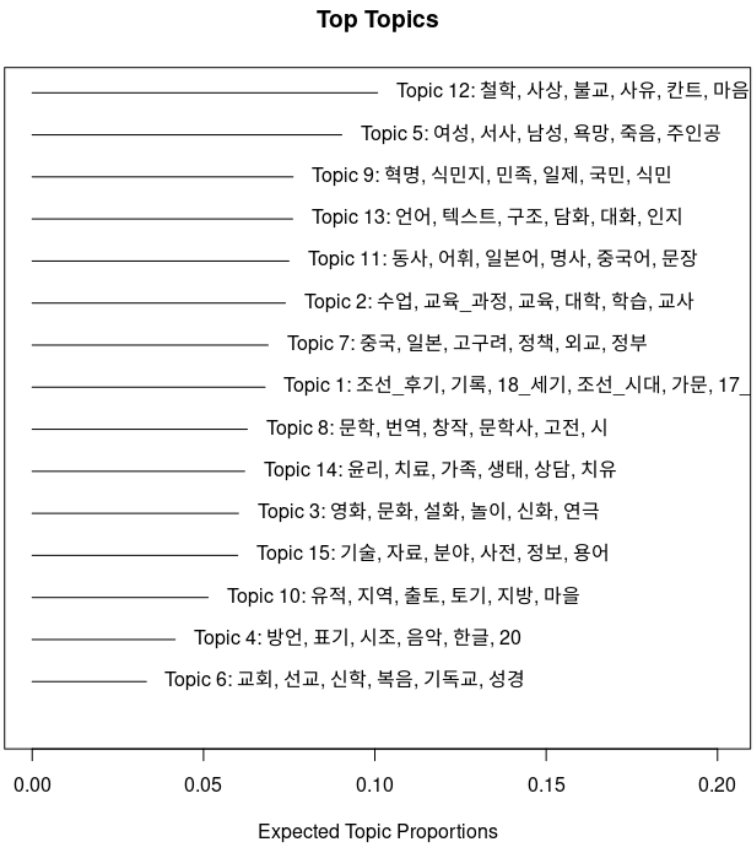

Figure 32. Topic distribution and keywords.

The results of labeling 15 topics based on the top 10 words of the four keyword weights (see Appendix 2) are as follows.

T12 (philosophy), T5 (narrative), T9 (colonial/modernity research), T13 (language), T11 (foreign language), T2 (education), T7 (Korean ancient history), T1 (late Joseon Dynasty), T8 (literary research), T14 (counseling/treatment), T3 (cultural contents), T15 (research data), T10 (historical data), T4 (Korean writing), T6 (theology)

The following diagram illustrates how the weight of fifteen themes varies per year. There are three basic categories for topics: hot with an uptrend, cold with a downturn, and neutral with neither a clear uptrend nor a clear downtrend. Hot topics are T1 (Late Joseon dynasty), T6 (Theology), T7 (Korean ancient history), T10 (Historical data), T14 (Counselling and treatment), T15 (Research data). Cold topics are T3 (Cultural contents), T4 (Korean notation), T5 (Narrative), T8 (Literature study), T11 (Foreign language), T12 (Philosophy). Sideway topics are T2 (Education), T9 (Colonial/Modernity studies), T13 (Linguistics).

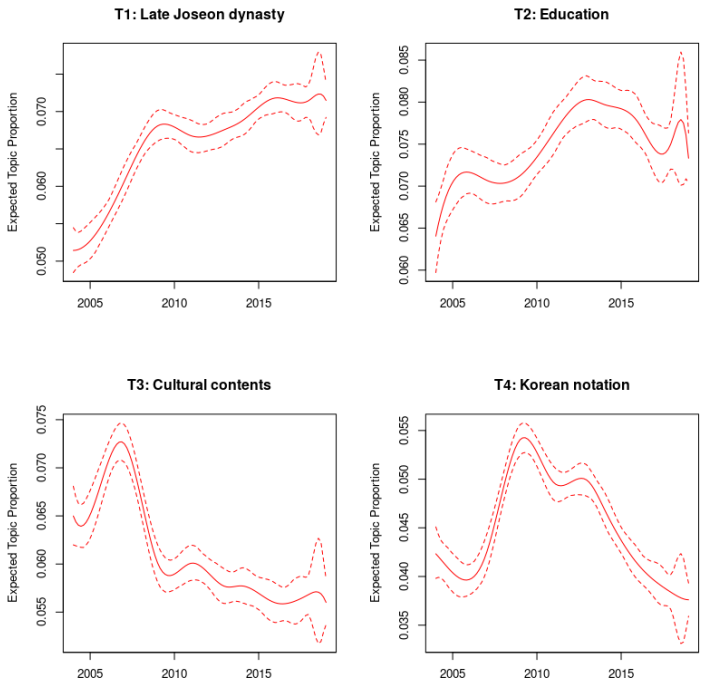

Figure 33. Trends in topic proportion from T1 to T4.

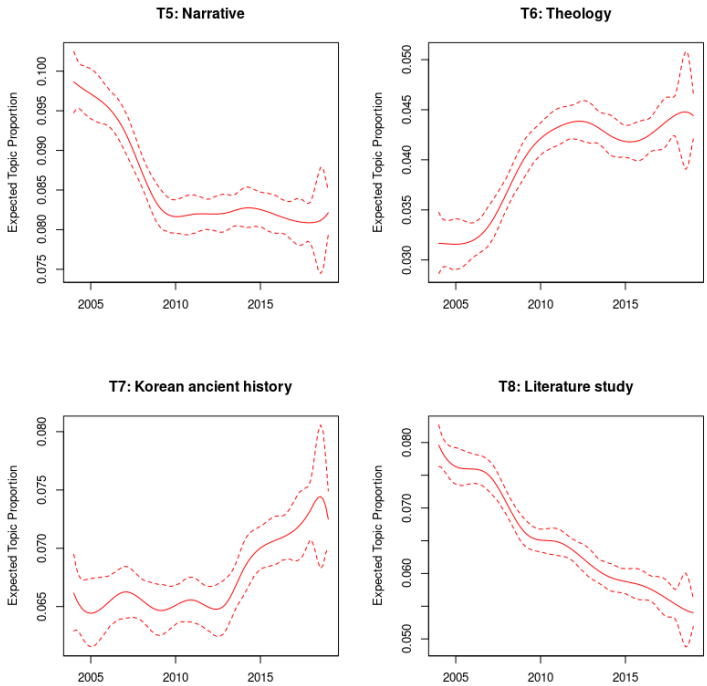

Figure 34. Trends in topic proportion from T5 to T8.

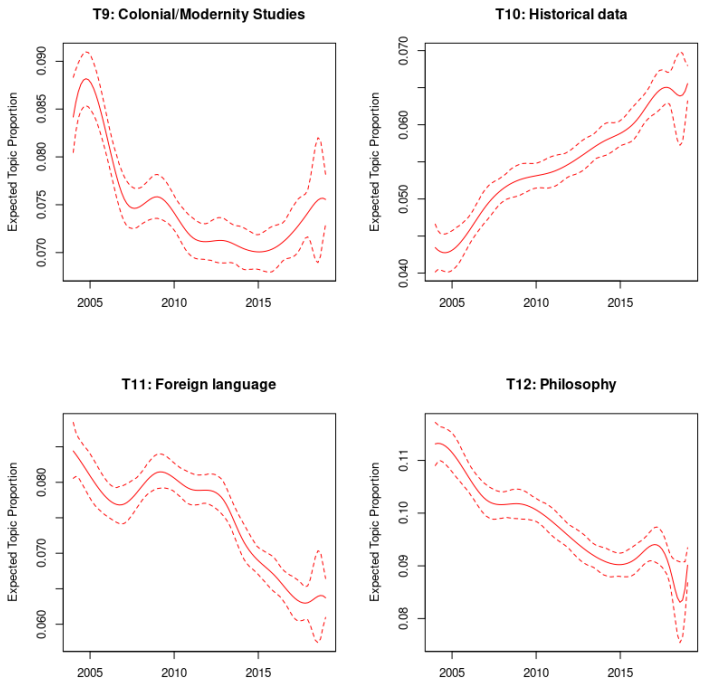

Figure 35. Trends in topic proportion from T9 to T12.

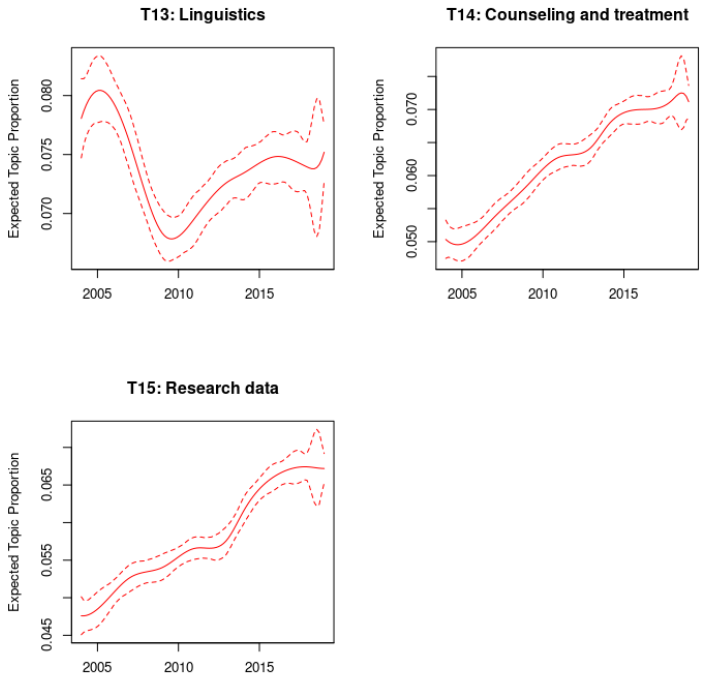

Figure 36. Trends in topic proportion from T13 to T15.

The above graphs depict a confidence interval, which is the outcome of assessing statistical significance by dividing the 15-year time span into 10 intervals. The effect of temporal factors on 15 topic models was evaluated using regression analysis given by STM (see Appendix 3). In other words, the p-values of 10 parts are examined to see whether the movement of each section is statistically significant. At this time, nine out of fifteen subjects with eight or more of ten portions were statistically significant (T1, T5, T6, T8, T9, T10, T12, T14, T15). T1 (late Joseon dynasty) and T9 (colonial/modernity studies) are utilized as examples, and statistically significant analysis is undertaken in relation to the inflection point of trend shift. T1, a subject relating to the history of the late Joseon Dynasty, rose in popularity until approximately 2010. According to Kwon (2011), on the study trend of the late Joseon Dynasty 2009-2010 was a time when numerous approaches of researching late Joseon history were attempted and advanced. This outcome corresponds to the topic modeling outcome. On the other side, 2007-2008 is the turning point for T9 study subjects relating to colonization and modernity. From this point forward, the comparable study demonstrated a decreasing trend followed by a minor increase. According to B. Kim & Cheon (2020), studies on colonialism and modernity in the study of Korean literature have decreased during the 2010s.

3) Conclusions

As a consequence of identifying research topics in the Korean humanities using STM, the three major domains, literature, history, and philosophy, as well as education and counseling-related topics, were discovered. Although dynamically depicted in the above image, it was difficult to discern the phenomena of concentration on a particular topic among the 15 topics. Additionally, the variability of the topic ratio over time was approximately 1 percent, making it difficult to notice a significant shift in the research topic over the past 15 years. However, subjects relating to Korean studies (Korean history, Korean literature, etc.) continued to occupy a significant percentage and exhibited an upward tendency. In addition, it was discovered that numerous research topics within the Korean humanities are in a perpetual state of flux.

3. Study3: Co-citation analysis

1) Introduction

In the first study, this paper examined general statistics pertaining to the Korean humanities; in the second study, it examined the shift in research topic. Using the reference list provided by KCI, study 3 examines which works have garnered specific attention.

To import KCI bibliographic information into CiteSpace, it must first be

translated into Web of Science (WoS) format, as described in the Method chapter. The KCI article number was used as a unique identifier for the reference column, which is an important column in this analysis. Because it was difficult to discern them, references lacking KCI paper numbers were omitted from the study. CiteSpace was configured with the necessary data after conversion to WoS format.

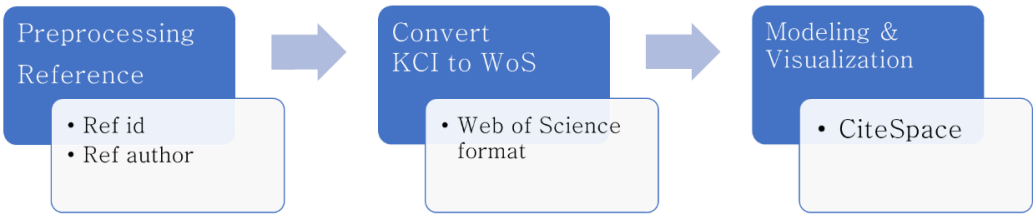

Figure 37. Preprocessing for co-citation analysis.

CiteSpace offers the ability to configure analytic units, network nodes, and other elements. The analysis unit can be configured as a thesis, a reference, the author of the thesis, the author of references, a keyword, a nation, or an affiliated institution (university). References were employed as the unit of analysis in this study. This is due to the fact that KCI papers were evaluated in study 2, and the simultaneous citation analysis is conducted because

references are the focus of analysis, and the genealogy of Korean humanities may be proved by observing which references have been in the spotlight.

Because using all references to be analyzed for analysis is inefficient and it is difficult to recognize them at a glance, only the top N references were analyzed. Document Co-citation Analysis (DCA) was conducted with the default K value of 25 for the g-index, one of the key reference selection criteria offered by CiteSpace. G-index is a measure of authors proposed by Egghe (2006). If the G-index is 10, it indicates that the author's 10 most cited works must have at least 100 citations. The operation resulted in the extraction of a total of 1,800 main references, which were related by 8,931 edges.

2) Results

i. Clusters

Fifteen clusters were discovered including more than 10 nodes. They are summarized in the following tables and figures. In the following table, Size represents the number of documents (nodes) belonging to the cluster, while Mean (Year) represents the average publication year of the articles. There were three clusters (#0, #1, and #10) pertaining to Korean language education that accounted for a substantial fraction. There were also three clusters related to the Korean language (#2, #8, #9), and Korean language education and Korean language studies were so closely associated that six

clusters were viewed as one neighboring cluster. A cluster of pan-Korean language and literature relating to writing, popular culture, colonial studies, and literary therapy was also present. It demonstrates that the majority of references used in Korean humanities are skewed toward Korean studies.

Table 11

Clusters and Labels by DCA

Cluster	Label	Size	Mean (Year)
#0	Korean language education 1	145	2006
#1	Korean language education 2	113	2007
#2	Korean linguistics 1	105	2009
#3	English education	101	2008
#5	Korean writing	77	2008
#7	Colonial studies	61	2006
#8	Korean linguistics 2	51	2008
#9	Korean linguistics 3	48	2006
#10	Korean language education	47	2009
#12	Korean history and comparative language	26	2009
#13	Interpretation and translation	26	2016
#15	Education	20	2013
#16	Christian theology	19	2014
#17	Popular culture	17	2011
#18	Literature education	12	2013

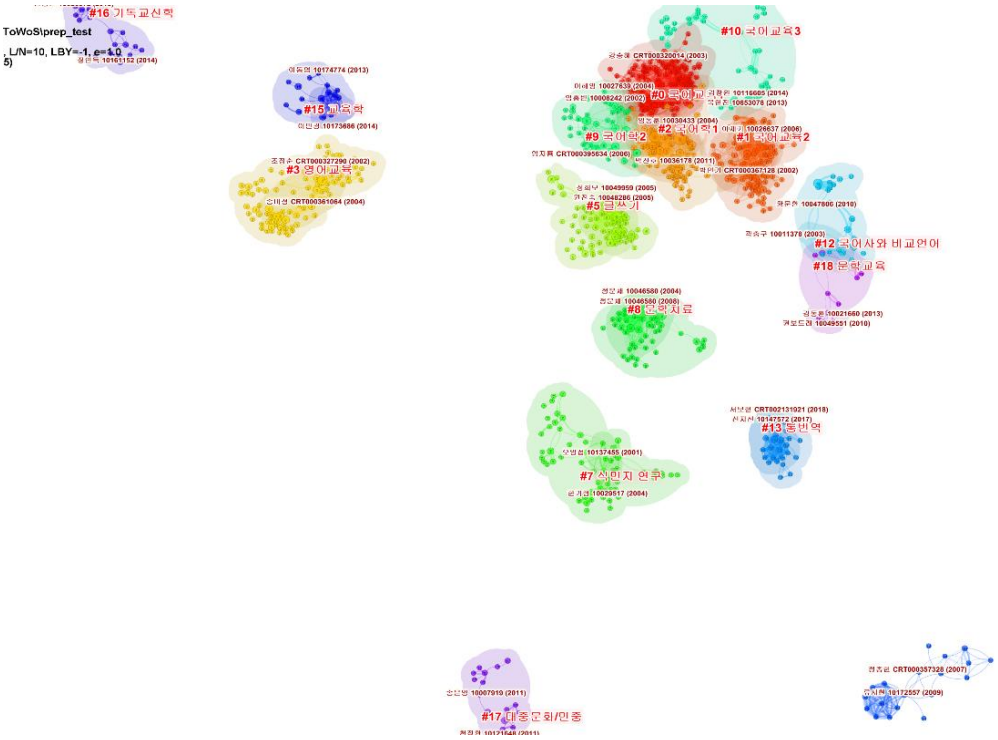

Figure 38. Clusters by DCA.

ii. Main nodes

The table 12 and figure 39 below provide a summary of the top 20 papers according to sigma, a statistic that combines betweenness centrality and citation burst in the simultaneous citation network. The more a paper's betweenness centrality, the greater its importance in co-citation analysis, as it serves as a hub between other articles. In other words, it signifies that the publication that becomes the hub is cited in numerous domains. Also, similar to the citation explosion, the strength index rises as a result of the rapid concentration of citations. If the citations are concentrated in a short period of time, it is likely that the thesis focuses on a prominent academic issue at the

moment. Therefore, the sigma top papers that summarize the indications are very likely to be research that drive the academic agenda in a short period of time and serve as a center for the citation network. Two cases are developed from the primary studies presented in the table below. Primarily, the highest fraction of research papers is devoted to methodology. These are papers that introduce approaches or concepts that are fundamental to several studies. For instance, Park's (2011) “Tense, Aspect, and Modality” was able to surge to the top because it addressed concepts and research methodology that are crucial to Korean language academics. Second, the thesis establishes a new research direction. It is a study that began as a trailblazer in a previously understudied field. Han (2004), for instance, proposed a new paradigm for magazine research within the field of Korean contemporary literature. Prior to that study, contemporary magazines and newspapers were not the primary focus of literary research; nevertheless, media research has since been the primary focus of Korean literary research (J. ho Kang, 2013).

Table 12

Top Sigma 20 Papers

Freq	Burst	Between	Sigma	Author	Year	Title	Cluster
144	**11.36**	**0.09**	**2.77**	**박진호**	**2011**	**시제, 상, 양태**	**2**
36	12.66	0.08	2.64	이민경	2014	거꾸로 교실(Flipped Classroom)의 효과와 의미에 대한 사례 연구	15
45	7.89	0.09	2.03	구본관	2011	원리 중심의 문법 교육에 대한 연구	2
73	11.3	0.06	1.86	정희모	2005	대학 글쓰기 교육과 사고력 학습에 관한 연구	5
34	10.95	0.04	1.55	민현식	2010	統合的 文法 敎育의 意義와 方向	1
80	10.1	0.03	1.38	이해영	2004	학문 목적 한국어 교과과정 설계 연구	0
82	7.66	0.04	1.31	조정순	2002	대학 교양영어 교육의 현황과 방향성 탐구	3
188	4.35	0.06	1.3	정운채	2008	문학치료학의 서사이론	8
59	5.5	0.05	1.28	박인기	2002	문화적 문식성의 국어교육적 재개념화	1
38	11.69	0.02	1.27	김성애	2002	장애 대학생 학업성취 실태 및 대학생활 욕구 분석	3
33	8.73	0.02	1.23	장경현	2003	문어/문어체 · 구어/구어체 재정립을 위한 시론	0
31	8.22	0.02	1.18	강소연	2005	English as the Medium of Instruction in Korean Engineering Education	3
25	**11.3**	**0.01**	**1.18**	**한기형**	**2004**	**최남선의 잡지 발간과 초기 근대문학의 재편-『소년』, 『청춘』의 문학사적 역할과 위상**	**7**
33	9.42	0.02	1.17	최숙기	2007	자기 표현적 글쓰기(expressive writing)의 교육적 함의	5
40	5.52	0.03	1.15	박진호	2007	유형론적 관점에서 본 한국어 대명사 체계의 특징	2
34	10.41	0.01	1.15	원진숙	2007	다문화 시대 국어교육의 역할	1
16	6.14	0.02	1.13	정근식	2005	일제하 검열기구와 검열관의 변동	7

26	6.58	0.02	1.12	민현식	2002	‘부사성’의 문법적 의미	1
15	6.61	0.02	1.12	김현정	2012	교양교육으로서의 글쓰기 교과의 본질과 방향	5
14	7.43	0.02	1.12	이봉범	2010	1950 년대 잡지저널리즘과 문학 – 대중잡지를 중심으로	17

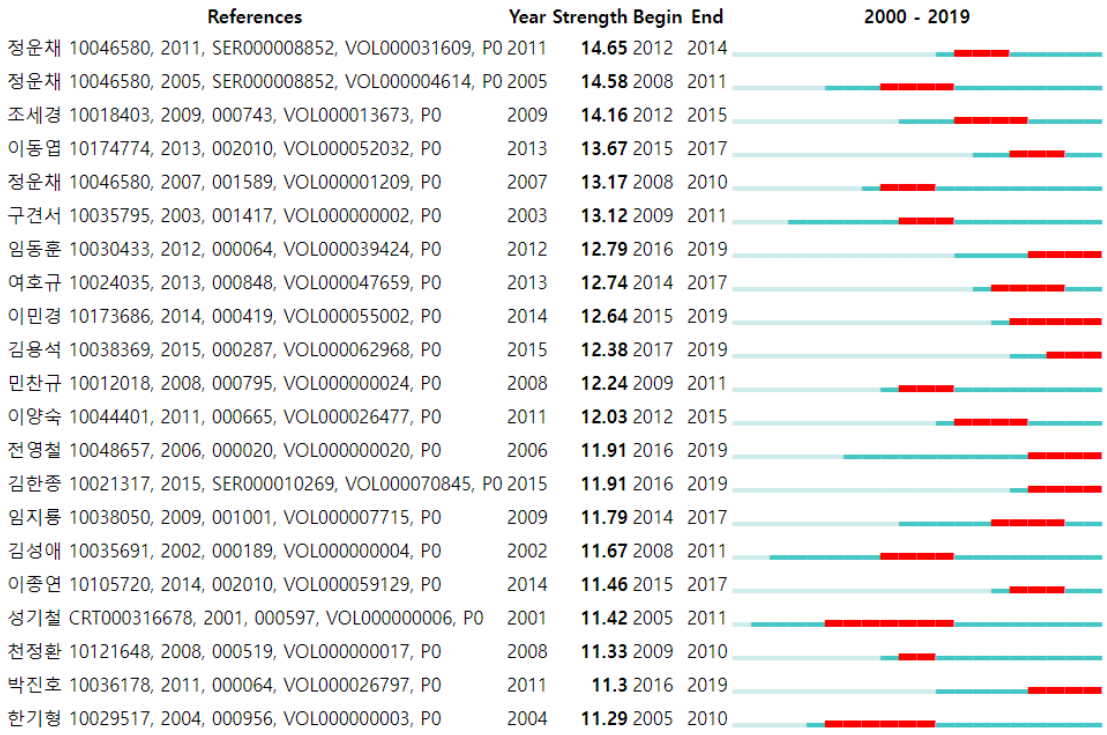

Figure 39. Top citation burst papers.

3) Conclusions

In the preceding Study 1, the phenomena of citation concentration are discussed in a small number of works. It was not possible to determine which papers were cited for what purpose. Through numerous indications on the citation network, this study explains why the manuscript garnered attention, as opposed to merely the number of citations. In other words, the co-citation analysis of study 3 could reveal a shift in the academic paradigm, which was not observed well in study 2. In conclusion, articles with high betweenness centrality and high citation burst are methodological (Peritz, 1983) or field-leading(Newman, 2009) papers, and in the future, the discussion of leading papers may become a key academic priority. Although it is impossible to observe a paradigm change as rapidly as in natural science or engineering, this study can quantitatively examine the existence of articles that serve as turning points in Korea's humanities study field.

Chapter 5. Conclusions

1. Summary

Since the foundation of the KCI system, quantitative growth in the Korean humanities has continued. On the other hand, the dialogue over the crises in the humanities continued. The Korean academic community's criticism that the humanities thesis used to satisfy the performance was "mass-produced" continues. After the 'crisis of the humanities' declared by the delegation of the national liberal arts schools in 2006, controversy and inquiry regarding the 'crisis' of the Korean humanities have continued in the media and in scholarly journals (Son, 2011). With KCI humanities bibliographic data and researcher information, the most active academic communication sectors in the Korean humanities, no macroscopic studies based on bibliographic approaches have been done.

Consequently, the goal of this study was to assess the knowledge structure of the Korean humanities utilizing a variety of quantitative approaches based on the KCI humanities theses published to date and the researcher data. An overall statistical analysis of the knowledge structure of the Korean humanities, an analysis of research subjects, and a

characterization analysis of the references in the network's center were conducted.

In the first study, EDA were employed to expose the knowledge structure using each researcher's demographic information, such as the number of KCI publications and citations produced, gender, and age during a fifteen-year period. As a result, paper publication in the humanities in Korea has constantly expanded but has become increasingly specialized. There was also a gender and age structure-based structural bias. Using the structural topic model, Study 2 evaluated the rise and fall of thesis research subjects. Literature, history, philosophy, and education comprise the majority of research topics in the Korean humanities, with Korean studies accounting for the majority of other topics. In the Korean humanities, there were hot topics and cold topics over the 15-year period, but there was no dramatic transformation. Using literature co-citation analysis, key papers were identified from a network created in Study 3. Using indices of the citation network, such as betweenness centrality and citation burst, I analyzed the papers that became the turning points of the study. Consequently, papers proposing a study approach or identifying a new research topic were evaluated highly. This was a characteristic that has also appeared in other fields of study, such as natural science and engineering.

2. Discussion

'Polarization' is a result shared by the three studies previously described. Analysis of research topic and citation networks revealed, in addition to bibliographic data and researcher information, a centralization phenomenon. In network theory, the phenomena of the power function are natural, but excessive concentration is also a concern (Barabási & Albert, 1999). It was discovered that specialization on a certain academic discipline or research issue, as well as on a particular age group, inhibits the diversity of the Korean humanities. It is uncertain whether national research strategies will encourage studies in neglected research topics instead of the more mainstream subjects. The purpose of this investigation was not to determine the influence of individual papers or researchers. Rather, the objective is to reflect on the preceding 15 years by examining a knowledge system that I would not have observed if I had not examined the data.

Additionally, it is vital to reconsider the claim that the KCI system violates academic freedom within the humanities. Since the birth of KCI, humanities scholars have been compelled to publish articles and research proposals just for funding, and long-term research has not been conducted correctly, according to Kang (2013). I believe the criticism effectively illustrates the contradiction in the contemporary Korean humanities environment. However, the proposed answer is an unrealistic Luddite

movement. His point is to reduce the influence of the KCI system, but given the reality of academic communication via KCI, this is nearly impossible. Even though the KCI system is biased, its effect is steadily growing in the academic ecosystem of the humanities, and the goal of open access to academic data in the context of digital (cyber) scholarship has been realized. It is anticipated that academic communication will become more active because of KCI, and I believe that it will serve as the seed for digital humanities and open science. As depicted in the diagram below, Shim et al. (2015) found that Korean Research Memory (KRM), an open scientific platform similar to KCI, can aid in the formation of a virtuous cycle structure for humanities research.

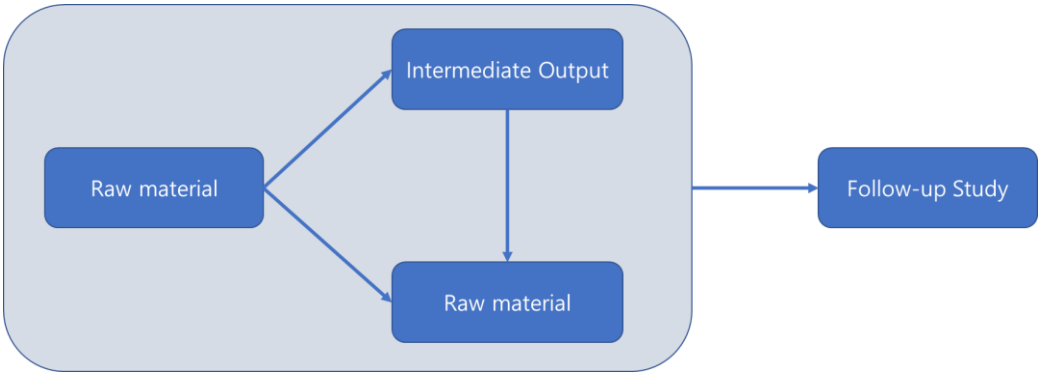

Figure 40. Circulation and subsequent linkage of humanities materials.
Source: Compiled from Shim et al. (2015), p.171.

I intend to conclude by discussing this study's contribution to the digital humanities and open science, which are the future goals of the humanities. As a result of excessive segmentation and specialization, the current dilemma in humanities research and education, according to Coleman (2009), is "learning more and more about less and less." She suggested

humanities education and research as a solution to this issue, highlighting "quantitative reasoning" and "connection" between fields as an example. This is similar to open science, which aspires to establish a link between the digital humanities, a data-driven quantitative research method, and academic fields.

If true, how can this paper contribute directly to digital humanities research and open science? First, this study shows the ability to use KCI thesis bibliographic information and researchers' demographic information to create a new intellectual history research approach. Intellectual history can play a role in facilitating inter-disciplinary dialogue in the before noted excessively fragmented humanities system (Min, 2017). This is due to the fact that intellectual history encompasses all humanities knowledge of the period. Intellectual history routinely dealt with tens of thousands of books and historical resources that might be merged with digital humanities-related approaches (Edelstein, 2016). As KCI already functions as a mediator of academic communication and tens of thousands of articles will be archived in the future, it is anticipated that study on intellectual history will be performed in the context of digital humanities. Second, this study demonstrates the possibilities of using open science in the Korean humanities and rejuvenating digital humanities on a broad scale. This is because this study employs the open-access KCI system and has significance as a digital humanities study employing open data in the humanities. In other words, while humanities research continues to concentrate on historical materials and classical works,

this investigation represents a novel effort and opportunity. Connected Paper[24], for instance, utilized open-access bibliographic information to offer researchers a network-based thesis search system (see Figure 41). Using KCI bibliographic data, Ryu et al., (2021) also processed and showed humanities texts in the form of network ontology[25] (see Figure 42). This network ontology prototype was developed as part of the open access movement[26] to alleviate the problem that humanities knowledge is not effectively disseminated throughout society, including academia. I expect that this work will serve as a steppingstone toward the realization of digital humanities and open science in the Korean humanities (M. K. Kim, 2018; Y. Kim, 2017; S. H. Park & Jung, 2022).

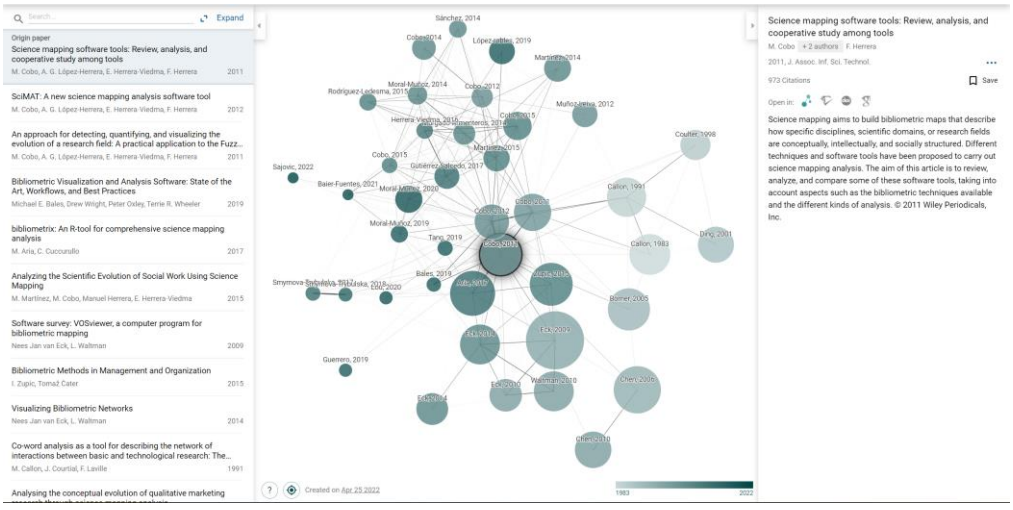

Figure 41. Connected Paper.

[24] https://www.connectedpapers.com/
[25] https://github.com/ByungjunKim/PaperKnowledgeGraph
[26] https://knowledgecommoning.org/

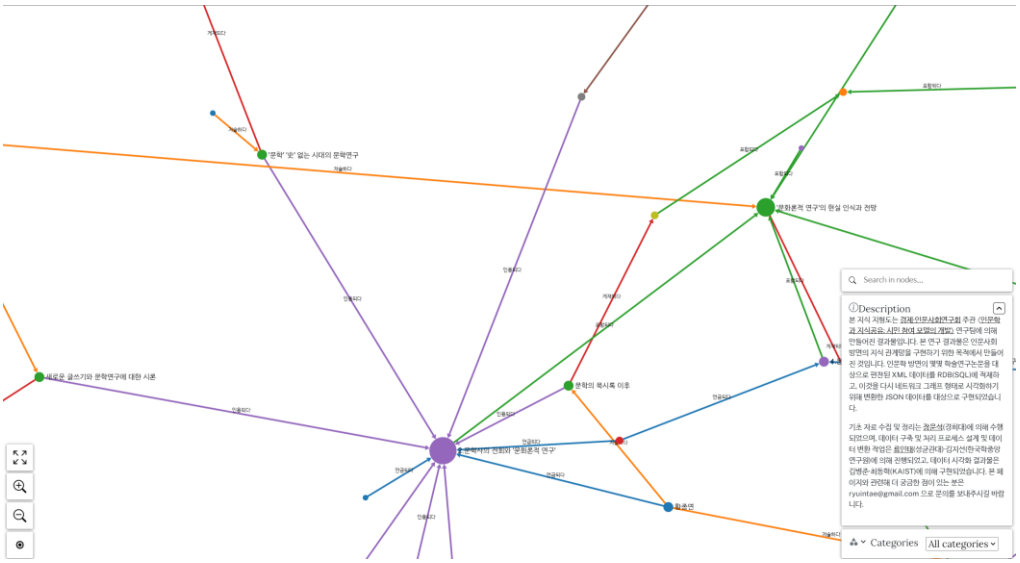

Figure 42. KCI Papers Network Ontology.

3. Limitation

The interpretation of the outcomes of the data analysis is this study's shortcoming. As in other studies (Song, 2015), a more nuanced interpretation would have resulted if the bibliographic analysis results had been shown to a scholars majoring in humanities and interpreted jointly. Additionally, it would be preferable if a study were undertaken on the influence of social background on citation behavior. In other words, it will be more beneficial to comprehend the knowledge structure of Korean humanities if I examine how origin school, gender, and generation impact citation.

I intend to pursue future study in two distinct areas[27]. First, by expanding the breadth of KCI thesis bibliographic data and comparing KCI with worldwide publications to confirm the existence of Korea's distinctive humanities and social sciences (L. Kim & Song, 2020). This will be a study that validates the contribution of Korean humanities and social sciences to the international academic community. Second, I plan to utilize KCI data for academic uniqueness (diversity) study. Using the word-embedding of abstracts, Shibayama et al (2021) created a method for measuring the uniqueness of research. On this basis, I aim to analyze the evolution and development of science in Korean humanities.

[27] KCI data up to 2019 was chosen for analysis since, at the time of collection (2021), publications published beyond 2020 could not be gathered due to the absence of reference data.

References

Arms, W. Y. (2008). Cyberscholarship: High Performance Computing Meets Digital Libraries. *Journal of Electronic Publishing*, *11*(1). https://doi.org/10.3998/3336451.0011.103

Assante, M., Candela, L., Castelli, D., Manghi, P., Pagano, P., & Nazionale, C. (2015). Science 2.0 repositories: Time for a change in scholarly communication. *D-Lib Magazine*, *21*(1/2), 1-14.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439). https://doi.org/10.1126/science.286.5439.509

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Borgman, C. L. (2000). Digital libraries and the continuum of scholarly communication. *Journal of Documentation*.

Cha, J. (2020). Big Data and the Future of Humanities. *Civilization and Borders*, *3*, 43-77.

Chen, C. (2020). A Glimpse of the First Eight Months of the COVID-19 Literature on Microsoft Academic Graph: Themes, Citation Contexts, and

Uncertainties. *Frontiers in Research Metrics and Analytics*, *5*. https://doi.org/10.3389/frma.2020.607286

Chen, C. (2021). *Delineating the Scholarly Landscape of a Research Field*. *June*. https://doi.org/10.13140/RG.2.2.20144.69127

Chen, C., & Song, M. (2017). *Representing Scientific Knowledge*.

Cheon, J. (2010). Evaluation System of University and Writing in the Era of Neo-Liberalism. *Critical Review of History*, *92*, 185-209.

Coleman, L. (2009). *Liz Coleman's call to reinvent liberal arts education*. TED. https://youtu.be/syqScVtnKuU

De Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*(3683). https://doi.org/10.1126/science.149.3683.510

Edelstein, D. (2016). Intellectual history and digital humanities. *Modern Intellectual History*, *13*(1), 237-246.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1). https://doi.org/10.1007/s11192-006-0144-7

Farradane, J. E. L. (1950). A scientific theory of classification and indexing and its practical applications. *Journal of Documentation*, *6*(2). https://doi.org/10.1108/eb026155

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association: JMLA*, *105*(2), 203.

Garvey, W. D., & Griffith, B. C. (1972). Communication and information processing within scientific disciplines: Empirical findings for psychology.

Information Storage and Retrieval, *8*(3), 123-136.

Han, K.-H. (2004). Choi Nam-son's Magazine Publication and Relocation of Early Modern Literature-SoNyun, ChungChun's role and position in literature. *DAEDONG MUNHWA YEON'GU*, *11*(45), 221-260.

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417-426. https://doi.org/10.1016/J.TOURMAN.2019.01.002

Hurd, J. M. (2000). The transformation of scientific communication: A model for 2020. *Journal of the American Society for Information Science*, *51*(14), 1279-1283. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1044>3.0.CO;2-1

Hwang, H. (2012). Inbursts from the Outside, Geopolitics of the Vernacular Studies — Korean modern literature studies, graphs, maps and trees on the problematics. *Sanghur Hakbo: The Journal of Korean Modern Literature*, *35*, 53-115. https://doi.org/10.22936/sh.35..201206.002

Hwang, K., Keehyeung, L., Yoon, S. H., & Deok, C. K. (2014). What is the Matter with Scholarly Writing? *Communication Theories*, *10*(1), 59-126.

Ioannidis, J. P. A. (2015). A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *Journal of Psychosomatic Research*, *78*(1), 7-11. https://doi.org/10.1016/j.jpsychores.2014.11.008

Jeong, Y. K. (2020a). Analyzing Topic Trends of News Articles Using

Structural Topic Modeling and Time Series Network Analysis in Hwangsungsinmun. *Critical Studies on Modern Korean History*, *24*(1), 133–159. https://doi.org/10.36432/CSMKH.43.202004.4

Jeong, Y. K. (2020b). Journal Citation Network Analysis of Library and Information Science Field in Korea. *Journal of the Korean Library and Information Science*, *54*(4), 221–238.

Ji, H. (2020). *Topic Analysis of Scholarly Communication Research: Using LDA Topic Modeling and Network Analysis* [Master, Ewha Womans University]. http://www.riss.kr/link?id=T15475535

Jin, C.-S. (2020). Study of Japanese Studies in Korea Entering a Period of Decline from a Stagnation. *The Korea Journal of Japanese Studies*, *51*, 125–156.

Jung, M. (2013). Transformation of Korean Intellectual Field and a Genealogy of Graduate Student, 1980–2012. *The Korean Journal of Cultural Sociology*, *15*, 7–78. https://doi.org/10.17328/kjcs.2013.15..001

Kamari, H., Huet, S., & Taupin, M.-L. (2019). *stm: R Package for Structural Topic Models*. *VV*(Ii), 1–41. https://doi.org/10.18637/jss.v000.i00

Kang, J. ho. (2013). Literature, Society, and Study of Literature—20th anniversary of Sanghur Study and Korean language study. *Sanghur Hakbo: The Journal of Korean Modern Literature*, *37*, 11–45. https://doi.org/10.22936/sh.37..201302.001

Kang, M. G. (2013). *Silent Factory*. A thousand years of imagination.

http://www.riss.kr/link?id=M13161705

Kaufmann, W. (1977). *The future of the humanities*. Transaction Publishers.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1). https://doi.org/10.1002/asi.5090140103

Kim, B. (2021). The Possibility of Bibliometric-based Conceptual History Research. *CONCEPT AND COMMUNICATION*, *28*, 233-246.

Kim, B., & Cheon, J. (2020). The Changes and Prospects of Studies on Modern Korean Literature Data Analysis of Doctoral Dissertations from 2000 throughout 2019. *Sanghur Hakbo: The Journal of Korean Modern Literature*, *60*, 443-517. https://doi.org/10.22936/sh.60..202010.012

Kim, B., Jun, B. G., & Lee, W. (2017). The Shifting Aspects of Concept Words Shown in Criticism in Literature Magazines, 1995 ~ 2015. *Journal of Korean Modern Literature*, *61*, 49-102. https://doi.org/10.35419/kmlit.2017..61.002

Kim, H. (2020a). Detection of Knowledge Structure of Korean Studies Using Document Co-citation Analysis: The Difference between Self-perception and Others' Perception. *Journal of Korean Library and Information Science Society*, *51*(1), 179-200. https://doi.org/10.16981/kliss.51.1.202003.179

Kim, H. (2020b). The Main Path Analysis of Korean Studies Using Text Mining: Based on SCOPUS Literature Containing 'Korea' as a Keyword. *Journal of the Korean Society for Information Management*, *37*(3), 253-274. https://doi.org/10.3743/KOSIM.2020.37.3.253

Kim, L., & Song, S.-Y. (2020). Is Korean Academia Unique? A Comparison of

Knowledge Discourses between Korean and International Sociology. *Korean Journal of Sociology*, *54*(4), 1-40. https://doi.org/10.21562/kjs.2020.11.54.4.1

Kim, M. K. (2018). Suggestions for Establishing Digital Humanities Policy in Korea. *Humanities Contents*, *51*, 41-67. https://doi.org/10.18658/humancon.2018.12.51.41

Kim, Y. (2017). Digital Humanities in Korea: Crisis, Hope, and Reality. *The Journal of Criticism and Theory*, *22*(2), 41-62.

Kim, Y. H., Kim, Y., & Kim, Y. S. (2008). The Structure of Production and Diffusion of Knowledge in Korean Communication Studies. *Korean Journal of Journalism & Communication Studies*, *52*(1), 117-140.

Klain-Gabbay, L., & Shoham, S. (2018). Scholarly communication and the academic library: Perceptions and recent developments. In *A Complex Systems Perspective of Communication from Cells to Societies* (pp. 1-22). IntechOpen London.

Knöchelmann, M. (2019). Open science in the humanities, or: Open humanities? *Publications*, *7*(4), 65.

Kretzschmar, W. A., Jr, & Gray Potter, W. (2010). Library collaboration with large digital humanities projects. *Literary and Linguistic Computing*, *25*(4), 439-445. https://doi.org/10.1093/llc/fqq022

Kuhn, T. (2010). *The Structure of Scientific Revolutions* (M. Ruse, Ed.; pp. 176-177). Princeton University Press. https://doi.org/10.1515/9781400831296-024

Kwon, N. H. (2011). Trends in Historical Studies of Late Joseon Dynasty. *YOKSA HAKBO*, *211*, 113-137.

Larsen, R. L. (2008). On the Threshold of Cyberscholarship. *Journal of Electronic Publishing*, *11*(1). https://doi.org/10.3998/3336451.0011.102

Lee, B. (2016). *An empirical study on analyzing the characteristics of R&D group effecting to convergence of R&D outputs: Advanced research center projects*. Sungkyunkwan University.

Lee, J. Y. (2015). Identifying the Research Fronts in Korean Library and Information Science by Document Co-citation Analysis. *Journal of the Korean Society for Information Management*, *32*(4), 77-106. https://doi.org/10.3743/KOSIM.2015.32.4.077

Lee, J. Y. (2021). Analyzing the Multidisciplinarity and Interdisciplinarity of Humanities and Social Science Journals in KCI: Focusing on General and Miscellaneous Fields. *Journal of the Korean Society for Information Management*, *38*(2), 227-250. https://doi.org/10.3743/KOSIM.2021.38.2.227

Lee, W., & Kim, B. (2020). *The generation gap doesn't disappear: 1980—2019 Korean Contemporary Literature Text Analysis*. 6-24. DBpia. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10477881

Longley Arthur, P., & Hearn, L. (2021). Toward Open Research: A Narrative Review of the Challenges and Opportunities for Open Humanities. *Journal of Communication*, *71*(5), 827-853. https://doi.org/10.1093/joc/jqab028

Mann, P. S. (2007). *Introductory statistics*. John Wiley & Sons.

Min, B. H. (2017). The Return of Intellectual History and a Challenging Future – Some Reflections on the State of the Field with a Focus on Recent Anglo-American Scholarship –. *Journal of Humanities*, *64*, 425-465. https://doi.org/10.31310/HUM.064.13

Mukherjee, B. (2009). Scholarly communication: A journey from print to web. *Library Philosophy and Practice*, 1.

Newman, M. E. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001.

Park, J. (2011). Tense, Aspect, and Modality. *Journal of Korean Linguistics*, *60*, 289-322. https://doi.org/10.15811/jkl.2011..60.011

Park, S. H., & Jung, K. H. (2022). *Share Knowledge: Korea Open Access Movement*. Red Salt.

Peritz, B. (1983). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, *5*(4), 211-218.

Reigersberg, M. S. (2015). Problematizing digital research evaluation using DOIs in practice-based arts, humanities and social science research. *F1000Research*, *4*.

Riande, G. del R., Tóth-Czifra, E., Wuttke, U., & Moranville, Y. (2020). *OpenMethods: A Compass for a More Open Digital Humanities*. https://doi.org/10.20944/preprints202003.0016.v1

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models

for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064-1082. https://doi.org/10.1111/ajps.12103

Roosendaal, H. E., & Geurts, P. (1997). Forces and functions in scientific communication: An analysis of their interplay. *Cooperative Research Information Systems in Physics*, *31*, 97.

Ryu, I. T., Kim, B., & Choi, D. H. (2021). *Humanities and Knowledge Sharing: Development of Citizen Participation Models*. http://www.klbksk.com/wiki/index.php/HumanitiesDataArchive

Seol, D.-H., Jaehun, K., Yoo, S.-H., & Lee, K. J. (2020). Content Analysis of Papers Published and Presented at the Korean Association for Survey Research, 1999-2019: Continuation and Change in Disciplines, Methodologies, Research Objects, and Techniques of Data Analysis. *Survey Research*, *21*(1), 1-32.

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PLOS ONE*, *16*(7), e0254034. https://doi.org/10.1371/JOURNAL.PONE.0254034

Shim, W., Ahn, H. Y., & Byun, J. (2015). Strategies for Improving the Collection and Use of Research Data in the Humanities. *Journal of the Korean Library and Information Science*, *49*(3), 155-183. https://doi.org/10.4275/KSLIS.2015.49.3.155

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for*

Information Science, *24*(4), 265-269.

Small, H. G. (1976). Structural dynamics of scientific literature. *KNOWLEDGE ORGANIZATION*, *3*(2), 67-74. https://doi.org/10.5771/0943-7444-1976-2-67

Son, Y.-S. (2011). A study on Groundworks for the Promotion of the Humanities in Korea. *Journal of the Daedong Philosophical Association*, *57*, 291-313.

Song, M. (2015). *Intellectual structure analysis in Korean studies* [Ph.D. Dissertation]. Sungkyunkwan University.

Steinerová, J. (2016). Open Science and the Research Information Literacy Framework. In S. Kurbanoğlu, J. Boustany, S. Špiranec, E. Grassian, D. Mizrachi, L. Roy, & T. Çakmak (Eds.), *Information Literacy: Key to an Inclusive Society* (pp. 277-285). Springer International Publishing. https://doi.org/10.1007/978-3-319-52162-6_27

Thanos, C. (2014). The Future of Digital Scholarship. *Procedia Computer Science*, *38*, 22-27. https://doi.org/10.1016/j.procs.2014.10.005

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

UNESCO. (2020, February 25). *UNESCO launches a global consultation to develop a standard-setting instrument on Open Science*. UNESCO. https://en.unesco.org/news/unesco-launches-global-consultation-develop-standard-setting-instrument-open-science

Yu, S.-Y. (2016). Exploratory Study of Characterizing Scholarly Communication Patterns in Humanities for Facilitating Consilience in Cyberscholarship Environment: Based on Historians' Research Activities. *Journal of the Korean Society for Library and Information Science*, *50*(1), 331-351. https://doi.org/10.4275/KSLIS.2016.50.1.331

Appendix

<Appendix 1> Stopword list

['것','수','등','년','속','중','연구','중심','의미','점','문제','관련','결과','논문','때','글',
'논의', '말','ㅁ','1_2','문제_해결','데_목적','데_기여','기여','데','데_도움
','2_3','1','2','3','4','때문','간','히','5','가운데','6','7','8','9','뿐',
'0', '고', '취', '다', '언','입', '견', '용', '기', '대', '자', '상','이', '연','록','씨',
'22','화','사', '명', '장', '일', '관', '지','제','부', '주','제이','행','특','출','신','본','학',
'설','편','집','서','경','권','정','본고','세','년대']

<Appendix 2> STM Keywords

Topic	Weight	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
1	Highest Prob	조선	기록	당시	인물	내용	조선_후기	이후	시기	활동	책
1	FREX	조선_후기	기록	18_세기	조선_시대	가문	17_세기	일기	간행	정조	책
1	Lift	차례	교유	일기	사대부	조선_후기	선비	정조	가문	문집	양반
1	Score	차례	조선	기록	조선_후기	문인	인물	간행	가문	조선_시대	편찬
2	Highest Prob	교육	학습	대학	평가	수업	학생	교과서	내용	학교	교사
2	FREX	수업	교육_과정	교육	대학	학습	교사	교과서	학생	학교	학습자
2	Lift	초등	수업	교육_과정	교과	초등학교	교수_학습	중학교	국어과	학년	과목
2	Score	초등	교육	학습	수업	교육_과정	교사	교과서	학습자	학생	대학
3	Highest Prob	문화	영화	공간	한국	전통	도시	예술	신화	설화	대중
3	FREX	영화	문화	설화	놀이	신화	연극	도시	산업	민속	공연
3	Lift	축제	공연	문화_콘텐츠	민속	연극	놀이	관광	스토리텔링	음식	미디어
3	Score	축제	문화	영화	도시	신화	공간	예술	설화	콘텐츠	연극
4	Highest Prob	변화	시조	음악	20	방언	표기	노래	한글	19	가
4	FREX	방언	표기	시조	음악	한글	20	가사	19	노래	음운
4	Lift	방언	모음	음운	가사	지명	표기	음	소리	발음	시조
4	Score	방언	시조	음운	표기	음악	모음	국어	음절	발음	한글

5	Highest Prob	여성	서사	소설	주체	의식	현실	인물	이야기	죽음	세계
5	FREX	여성	서사	남성	욕망	죽음	주인공	기억	형상	환상	비극
5	Lift	수사학	비극	환상	젠더	주인공	남성	결말	영웅	여성	모티프
5	Score	수사학	여성	서사	소설	욕망	죽음	남성	작품	사랑	인물
6	Highest Prob	종교	교회	기독교	신학	신앙	선교	하나님	영성	성경	영
6	FREX	교회	선교	신학	복음	기독교	성경	하나님	예수	종교	한국_교회
6	Lift	복음	선교	예수	한국_교회	교회	성서	그리스도	신학	성경	선교사
6	Score	선교	교회	신학	기독교	하나님	종교	신앙	복음	한국_교회	그리스도
7	Highest Prob	일본	중국	정책	전쟁	한국	조선	국가	관계	신라	정부
7	FREX	중국	일본	고구려	정책	외교	정부	군사	동아시아	전쟁	국
7	Lift	고구려	외교	왕권	전투	파견	군사	대외	황제	중국	일본
7	Score	고구려	일본	중국	신라	전쟁	정책	조선	외교	정부	백제
8	Highest Prob	문학	시	작품	소설	번역	창작	작가	한국	고전	비평
8	FREX	문학	번역	창작	문학사	고전	시	비평	작품	작가	장르
8	Lift	번역	문학사	문단	문학	문예	창작	시가	모더니즘	산문	희곡
8	Score	번역	문학	작품	시	소설	작가	창작	문학사	시인	비평
9	Highest Prob	근대	정치	운동	민족	사회	국가	담론	역사	식민지	한국
9	FREX	혁명	식민지	민족	일제	국민	식민	북한	해방	민족주의	운동
9	Lift	동학	민족주의	식민	민주주의	혁명	식민주의	사회주의	인민	민주	식민지
9	Score	동학	운동	근대	민족	정치	식민지	혁명	식민	국가	북한

10	Highest Prob	지역	시기	확인	지방	변화	유적	세기	백제	제작	마을
10	FREX	유적	지역	출토	토기	지방	마을	조성	사업	세기	제작
10	Lift	유적	토기	출토	축조	고고학	시설	유역	부산	지역	남부
10	Score	토기	지역	백제	유적	출토	지방	마을	유역	신라	축조
11	Highest Prob	어휘	한국어	문법	사용	고찰	문장	문	일본어	동사	유형
11	FREX	동사	어휘	일본어	명사	중국어	문장	구문	문법	단어	통사
11	Lift	장합	형용사	동사	명사	구문	용법	통사	일본어	타	중국어
11	Score	형용사	어휘	한국어	문법	동사	일본어	명사	구문	중국어	문장
12	Highest Prob	사상	철학	불교	존재	개념	비판	인간	해석	이해	인식
12	FREX	철학	사상	불교	사유	칸트	마음	이성	유가	비판	주장
12	Lift	장자	맹자	형이상학	만물	칸트	공자	심성	철학	주희	철학자
12	Score	장자	철학	사상	불교	칸트	인간	도덕	형이상학	개념	사유
13	Highest Prob	언어	분석	텍스트	구조	구성	표현	기능	개념	방식	사용
13	FREX	언어	텍스트	구조	담화	대화	인지	기능	은유	표현	발화
13	Lift	표지	담화	발화	은유	수사	언어	설득	텍스트	내러티브	가설
13	Score	표지	언어	텍스트	담화	화자	은유	발화	대화	구조	표현
14	Highest Prob	사회	인간	윤리	삶	관계	공동체	개인	가족	가치	사람
14	FREX	윤리	치료	가족	생태	상담	치유	생명	행복	가정	부모
14	Lift	목회	치료	상담	질병	윤리	행복	생태	부모	치유	의학
14	Score	목회	인간	윤리	사회	상담	치료	도덕	가족	생명	삶

15	Highest Prob	역사	자료	기술	방법	검토	분석	필요	내용	지식	활용
15	FREX	기술	자료	분야	사전	정보	용어	정리	지식	연구자	학술
15	Lift	사전	학술	연구_성과	분야	수집	용어	수정	연구자	한자	기술
15	Score	사전	자료	한자	역사	정보	지식	기술	한문	문헌	용어

<Appendix 3> STM Regression Results

Topic 1:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.140e-02	1.513e-03	33.968	< 2e-16	***
s(year)1	8.044e-05	3.514e-03	0.023	0.9817	
s(year)2	7.127e-03	3.109e-03	2.292	0.0219	*
s(year)3	1.930e-02	2.780e-03	6.942	3.88e-12	***
s(year)4	1.415e-02	2.681e-03	5.277	1.31e-07	***
s(year)5	1.592e-02	2.688e-03	5.922	3.19e-09	***
s(year)6	1.743e-02	2.671e-03	6.526	6.75e-11	***
s(year)7	2.213e-02	3.519e-03	6.289	3.21e-10	***
s(year)8	1.827e-02	4.156e-03	4.396	1.10e-05	***
s(year)9	2.184e-02	5.258e-03	4.153	3.29e-05	***
s(year)10	2.009e-02	1.965e-03	10.229	< 2e-16	***

Topic 2:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.064076	0.002160	29.670	< 2e-16	***
s(year)1	0.010074	0.004990	2.019	0.043483	*
s(year)2	0.005980	0.003988	1.499	0.133793	

s(year)3	0.006132	0.003751	1.635	0.102109
s(year)4	0.012342	0.003468	3.559	0.000372 ***
s(year)5	0.017087	0.003601	4.745	2.09e-06 ***
s(year)6	0.015108	0.003470	4.354	1.34e-05 ***
s(year)7	0.015629	0.004747	3.292	0.000995 ***
s(year)8	0.005649	0.005769	0.979	0.327446
s(year)9	0.016822	0.007245	2.322	0.020241 *
s(year)10	0.009203	0.002659	3.461	0.000539 ***

Topic 3:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.065010	0.001581	41.110	< 2e-16 ***
s(year)1	-0.005133	0.003588	-1.431	0.152474
s(year)2	0.018915	0.002711	6.977	3.03e-12 ***
s(year)3	-0.010363	0.002678	-3.869	0.000109 ***
s(year)4	-0.002536	0.002365	-1.072	0.283591
s(year)5	-0.008046	0.002443	-3.294	0.000987 ***
s(year)6	-0.006721	0.002473	-2.718	0.006561 **
s(year)7	-0.009988	0.003482	-2.868	0.004129 **
s(year)8	-0.008460	0.004159	-2.034	0.041950 *
s(year)9	-0.007561	0.005135	-1.473	0.140856
s(year)10	-0.009033	0.002029	-4.452	8.52e-06 ***

Topic 4:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.042310	0.001335	31.700	< 2e-16	***
s(year)1	-0.001821	0.003005	-0.606	0.54444	
s(year)2	-0.007166	0.002658	-2.696	0.00701	**
s(year)3	0.017598	0.002058	8.551	< 2e-16	***
s(year)4	0.003924	0.002345	1.674	0.09419	.
s(year)5	0.009617	0.002063	4.661	3.15e-06	***
s(year)6	0.003287	0.002156	1.525	0.12734	
s(year)7	-0.001728	0.002691	-0.642	0.52083	
s(year)8	-0.003585	0.003389	-1.058	0.29021	
s(year)9	-0.004763	0.004332	-1.099	0.27157	
s(year)10	-0.004714	0.001546	-3.050	0.00229	**

Topic 5:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.098659	0.001978	49.889	< 2e-16	***
s(year)1	-0.002018	0.004627	-0.436	0.662727	
s(year)2	-0.002026	0.003942	-0.514	0.607341	
s(year)3	-0.018328	0.003300	-5.554	2.80e-08	***

s(year)4	−0.016243	0.003446	−4.714	2.43e−06 ***
s(year)5	−0.016945	0.003073	−5.514	3.51e−08 ***
s(year)6	−0.015512	0.003230	−4.803	1.57e−06 ***
s(year)7	−0.016815	0.003914	−4.296	1.74e−05 ***
s(year)8	−0.018181	0.005153	−3.528	0.000419 ***
s(year)9	−0.017408	0.006081	−2.863	0.004202 **
s(year)10	−0.016529	0.002315	−7.140	9.38e−13 ***

Topic 6:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.163e−02	1.569e−03	20.159	< 2e−16 ***
s(year)1	−3.987e−05	3.686e−03	−0.011	0.991370
s(year)2	−8.819e−04	2.861e−03	−0.308	0.757888
s(year)3	9.424e−03	2.734e−03	3.447	0.000566 ***
s(year)4	1.186e−02	2.412e−03	4.918	8.75e−07 ***
s(year)5	1.268e−02	2.556e−03	4.963	6.94e−07 ***
s(year)6	1.031e−02	2.518e−03	4.096	4.20e−05 ***
s(year)7	9.695e−03	3.333e−03	2.908	0.003632 **
s(year)8	1.277e−02	4.131e−03	3.092	0.001990 **
s(year)9	1.355e−02	5.377e−03	2.520	0.011733 *
s(year)10	1.277e−02	1.950e−03	6.547	5.87e−11 ***

Topic 7:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0661161	0.0017103	38.658	< 2e−16	***
s(year)1	−0.0038023	0.0042866	−0.887	0.37506	
s(year)2	0.0025590	0.0033485	0.764	0.44474	
s(year)3	−0.0028006	0.0029894	−0.937	0.34885	
s(year)4	0.0007718	0.0027400	0.282	0.77819	
s(year)5	−0.0028449	0.0027442	−1.037	0.29989	
s(year)6	0.0038540	0.0029078	1.325	0.18503	
s(year)7	0.0045489	0.0036001	1.264	0.20638	
s(year)8	0.0057443	0.0047774	1.202	0.22922	
s(year)9	0.0093481	0.0058368	1.602	0.10925	
s(year)10	0.0063822	0.0020243	3.153	0.00162	**

Topic 8:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0795151	0.0016101	49.385	< 2e−16	***
s(year)1	−0.0058449	0.0039848	−1.467	0.142	
s(year)2	0.0006603	0.0033141	0.199	0.842	
s(year)3	−0.0158500	0.0027331	−5.799	6.67e−09	***
s(year)4	−0.0134251	0.0027039	−4.965	6.87e−07	***

s(year)5	−0.0178315	0.0027774	−6.420	1.36e−10 ***
s(year)6	−0.0207595	0.0025177	−8.245	< 2e−16 ***
s(year)7	−0.0207689	0.0036796	−5.644	1.66e−08 ***
s(year)8	−0.0238099	0.0045170	−5.271	1.36e−07 ***
s(year)9	−0.0253568	0.0052985	−4.786	1.71e−06 ***
s(year)10	−0.0255169	0.0019788	−12.895	< 2e−16 ***

Topic 9:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.084203	0.001989	42.331	< 2e−16 ***
s(year)1	0.011283	0.004293	2.628	0.008590 **
s(year)2	−0.015098	0.003585	−4.211	2.54e−05 ***
s(year)3	−0.005537	0.003342	−1.657	0.097519 .
s(year)4	−0.014340	0.002993	−4.791	1.66e−06 ***
s(year)5	−0.012367	0.003000	−4.123	3.75e−05 ***
s(year)6	−0.014195	0.002845	−4.989	6.06e−07 ***
s(year)7	−0.014401	0.003965	−3.632	0.000281 ***
s(year)8	−0.010942	0.004910	−2.229	0.025831 *
s(year)9	−0.008041	0.005945	−1.353	0.176156
s(year)10	−0.008622	0.002383	−3.618	0.000297 ***

Topic 10:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.043449	0.001602	27.122	< 2e−16	***
s(year)1	−0.002531	0.003762	−0.673	0.501172	
s(year)2	0.005882	0.003378	1.741	0.081617	.
s(year)3	0.009663	0.002919	3.310	0.000933	***
s(year)4	0.009714	0.002753	3.528	0.000418	***
s(year)5	0.012333	0.002697	4.573	4.82e−06	***
s(year)6	0.014920	0.002732	5.461	4.74e−08	***
s(year)7	0.015786	0.003683	4.286	1.82e−05	***
s(year)8	0.024163	0.004506	5.362	8.22e−08	***
s(year)9	0.019100	0.005846	3.267	0.001086	**
s(year)10	0.022075	0.002045	10.795	< 2e−16	***

Topic 11:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0844595	0.0019964	42.305	< 2e−16	***
s(year)1	−0.0031085	0.0046341	−0.671	0.50236	
s(year)2	−0.0120055	0.0036776	−3.264	0.00110	**
s(year)3	−0.0001298	0.0034363	−0.038	0.96988	
s(year)4	−0.0073139	0.0031104	−2.351	0.01870	*

s(year)5	−0.0040316	0.0030500	−1.322	0.18623
s(year)6	−0.0150854	0.0032351	−4.663	3.12e−06 ***
s(year)7	−0.0166457	0.0041002	−4.060	4.91e−05 ***
s(year)8	−0.0233493	0.0049724	−4.696	2.66e−06 ***
s(year)9	−0.0198993	0.0062144	−3.202	0.00136 **
s(year)10	−0.0206964	0.0025097	−8.247	< 2e−16 ***

Topic 12:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.113143	0.002112	53.583	< 2e−16 ***
s(year)1	0.001114	0.005126	0.217	0.82800
s(year)2	−0.013922	0.004348	−3.202	0.00136 **
s(year)3	−0.010078	0.003795	−2.656	0.00792 **
s(year)4	−0.014618	0.003613	−4.046	5.21e−05 ***
s(year)5	−0.019793	0.003305	−5.989	2.11e−09 ***
s(year)6	−0.022968	0.003330	−6.897	5.32e−12 ***
s(year)7	−0.023682	0.004608	−5.139	2.76e−07 ***
s(year)8	−0.014052	0.005658	−2.484	0.01301 *
s(year)9	−0.035365	0.007152	−4.945	7.64e−07 ***
s(year)10	−0.022992	0.002694	−8.535	< 2e−16 ***

Topic 13:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0780581	0.0017146	45.526	< 2e−16	***
s(year)1	0.0043209	0.0040074	1.078	0.28094	
s(year)2	0.0006003	0.0033727	0.178	0.85872	
s(year)3	−0.0121547	0.0028753	−4.227	2.37e−05	***
s(year)4	−0.0084380	0.0029138	−2.896	0.00378	**
s(year)5	−0.0054545	0.0026979	−2.022	0.04320	*
s(year)6	−0.0045613	0.0026794	−1.702	0.08869	.
s(year)7	−0.0025673	0.0038069	−0.674	0.50007	
s(year)8	−0.0039466	0.0041602	−0.949	0.34279	
s(year)9	−0.0045150	0.0056351	−0.801	0.42300	
s(year)10	−0.0028417	0.0020858	−1.362	0.17307	

Topic 14:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.050333	0.001550	32.478	< 2e−16	***
s(year)1	−0.002314	0.003665	−0.631	0.52786	
s(year)2	0.003060	0.003049	1.004	0.31548	
s(year)3	0.007735	0.002636	2.935	0.00334	**
s(year)4	0.013870	0.002773	5.002	5.68e−07	***

s(year)5	0.011967	0.002426	4.932	8.13e-07 ***
s(year)6	0.019093	0.002634	7.248	4.23e-13 ***
s(year)7	0.019984	0.003327	6.006	1.90e-09 ***
s(year)8	0.019283	0.004272	4.514	6.38e-06 ***
s(year)9	0.023300	0.005094	4.574	4.79e-06 ***
s(year)10	0.020833	0.002033	10.249	< 2e-16 ***

Topic 15:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0476107	0.0012599	37.790	< 2e-16 ***
s(year)1	-0.0003307	0.0029678	-0.111	0.911286
s(year)2	0.0063979	0.0024682	2.592	0.009539 **
s(year)3	0.0054817	0.0022037	2.488	0.012864 *
s(year)4	0.0102230	0.0021204	4.821	1.43e-06 ***
s(year)5	0.0078527	0.0021155	3.712	0.000206 ***
s(year)6	0.0157085	0.0020951	7.498	6.51e-14 ***
s(year)7	0.0190680	0.0029545	6.454	1.09e-10 ***
s(year)8	0.0201467	0.0036703	5.489	4.05e-08 ***
s(year)9	0.0194511	0.0049154	3.957	7.59e-05 ***
s(year)10	0.0195882	0.0015906	12.315	< 2e-16 ***

논문요약

한국 인문학의 지식구조 파악:
KCI 인문학 논문의 서지 데이터 분석, 2004~2019

김병준

인터랙션사이언스학과

성균관대학교

이 연구는 2004년부터 2019년까지 발간된 약 25만건의 KCI 인문학 논문 서지정보를 디지털인문학 방법론으로 분석해 지난 15년간의 한국 인문학의 지식구조를 파악한다. 분석에 활용한 서지정보는 논문의 제목, 초록, 주제어 등의 텍스트 정보와 피인용 수, 참고문헌 등의 인용 정보, 그리고 해당 논문을 쓴 연구자의 성별, 나이, 출신 학교 등의 인구 사회학 정보이다. 본 연구는 크게 네 장으로 구성된다. 1) KCI 논문 서지정보와 연구자의 인구 사회학 정보에 대한 탐색적 데이터 분석, 2) 구조적 토픽 모형을 활용한 연구 주제 군집 분석, 3) 참고문헌 동시 인용 분석을 통한 연구 지형도 분석이다. 4) 세가지 연구 결과를 토대로 앞으로의 한국 인문학에 대한 정책적 제언이다.

Study1에서는 연도별 논문 게재량과 참고문헌 목록에서 시간에 따른 논문 생산량의 변화와 연구자의 세대 및 성별 변화 등을 탐색적 데이터 분석으로 살펴볼

것이다. Study 2에서는 논문의 텍스트 정보를 바탕으로 논문 메타정보에 따른 토픽 변화를 분석한다. 어떤 연구 주제가 시간에 흐름에 따라 부상하고 하락했는지 확인하고 앞으로 주목받을 주제는 무엇인지 확인한다. Study 3에서는 참고문헌 동시 인용 분석을 활용해 인용 관계에 따른 논문의 군집을 추출하고, 군집별로 해당 연구 군집을 이끈 중요한 저작을 뽑아 분석한다. 결론에서는 위 네 가지 데이터 분석 결과를 바탕으로 지난 15년간의 한국 인문학의 지식 생산과 확산의 구조를 밝히고, 앞으로의 한국 인문학계를 조망한다. 특히 오픈 사이언스와 디지털 인문학이라는 아젠다와 앞으로의 한국 인문학의 방향을 제언한다.

주제어: 학술 커뮤니케이션, 한국 인문학, 지식구조, 디지털인문학, KCI