

기무라 에이이치(木村英一) 장군(章群) 이론의 계량적 검토^{***}

— 『논어』1~5편의 군집 분석과 인문학적 재검토 구간의 식별

김서윤^{***}

-
1. 서론
 2. 기무라의 장군 이론
 3. 장군 이론의 정형화
 4. 이중 층위 군집 분석과 장군 분류의 타당성 진단
 - 1) 분석 방법 및 이중 층위 설계의 목적
 - 2) Jiayan 형태소 기반 군집 분석 결과
 - 3) GuwenBERT 임베딩 기반 군집 분석 결과
 - 4) 정합·부분 일치·인문학적 재검토
 5. 결론 및 후속 연구
-

■ 국문요약

본 연구는 기무라 에이이치(木村英一)의 장군(章群) 이론을 계량적으로 검토하고 그 타당성을 진단하는 것을 목적으로 한다. 기무라는 『논어』 각 편 내에

* 이 논문은 2024년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2024S1A5B5A17040150).

** 이 논문은 2026년 한국경학학회 춘계학술대회에서 발표된 논문을 수정·보완하여 작성되었음.

*** 한국학중앙연구원 디지털인문학연구소 연구원 / E-mail: sy527991@gmail.com

서 의미·형태적으로 연관된 장들을 군집 단위로 묶고 그 형식을 유형화함으로써 『논어』 문헌 연구를 구조적 재편성의 단계로 심화시켰다. 그러나 그의 이론은 방대한 분량의 산문 형식으로 기술되어 있어 구조화된 데이터로 전환하기 어렵고, 그 정성적 판단이 계량 분석과 얼마나 부합하는지의 타당성은 지금껏 검증된 바 없다.

이에 본 연구는 두 단계의 작업을 수행하였다. 첫째, 기무라의 이론을 대분류·소분류의 두 층위로 정형화하여 『논어』 제1~5편에 대한 분류 모델을 수립하였다. 이는 산문 형식으로 기술된 기무라의 판단을 정형화된 데이터 구조로 옮기는 작업이다. 둘째, 이중 층위의 군집 분석을 수행하여 기무라의 분류와 대조하였다. 고대 한문 형태소 분석기 Jiayan을 활용한 TF-IDF 분석으로 어휘 패턴의 형태적 유사성을, 고대 한문 언어모델 GuwenBERT 임베딩 분석으로 문맥적 의미의 유사성을 각각 측정하였다.

분석 결과는 단순한 일치·불일치의 이분법으로 환원되지 않고 세 가지 구간으로 식별되었다. 곧 기무라의 판단이 어휘·의미 두 층위의 분석 결과와 모두 일치하는 ‘정합 구간’, 어휘와 의미 중 한 측면에서만 부합하는 ‘부분 일치 구간’, 두 계량 분석이 분기되어 기무라의 분류와 어긋남으로써 비판적 정독을 통한 추가 검토가 요구되는 ‘인문학적 재검토 구간’이 그것이다. 이 세 구간 체계는 계량적 분석이 인문학적 판단을 대체하는 것이 아니라, 재검토가 필요한 지점을 식별하는 진단 도구로 기능한다는 본 연구의 입장을 보여준다. 본 연구는 1~5편을 대상으로 한 파일럿으로서, 여기서 확립된 방법론은 향후 『논어』 20편 전체로 확장될 수 있다.

주제어: 『논어』, 장군 이론, 기무라 에이이치, 군집 분석, Jiayan, GuwenBERT

1. 서론

『논어』의 편장 구조가 체계적이지 못하다는 것은 오랫동안 자명한

전체로 받아들여져 왔다. 중국의 역사학자 錢穆(1895~1990)는 “『논어』의 편차나 체례에는 정해진 규칙이 없고, 편장의 선후에도 특별한 의의가 없다.”라고 했고,¹⁾ 김용옥(1948~) 역시 『논어』가 전체적으로 통일된 주제 없이 단편적인 조합 또는 집합으로 이루어졌다고 지적한 바 있다.²⁾ 이러한 평가는 단순한 형식상의 문제가 아니라, 공자 사상의 온전한 파악을 가로막는 본질적 장애라는 점에서 중요하다. 오랜 시간에 걸쳐 각 편의 내용이 재배치되고 여러 자료가 후대에 삽입되면서 공자의 어록과 제자들의 언행이 뒤섞여 있기 때문이다.³⁾ 더욱이 현행 20편 체계는 사상적 완결성을 기준으로 편찬된 것이 아니라, 前漢 말기 張禹(?~기원전 5년)가 魯論을 중심으로 齊論을 절충하여 확정한 이른바 ‘張侯論’이 그대로 정착된 결과일 뿐이다.⁴⁾

이러한 인식 위에서 현행 20편 체계에 정면으로 문제를 제기한 연구들이 등장하였다. 이들의 흐름은 크게 두 가지로 나뉘는데, 하나는 개별 편·장 단위의 미시적 차원에서 문헌 비평적 신뢰도를 따지는 방향이고, 다른 하나는 20편 전체를 거시적으로 조망하여 상론·하론의 이분 구조로 재구분하는 방향이다.

전자의 대표적 사례가 청대 고증학자 崔述(1740~1816)의 『論語餘說』 및 『洙泗考信錄』이다. 그는 공자 및 제자들의 행적을 낱낱이 고증한 『수사고신록』에서 『논어』 후반 5편에 대한 후대 삽입 가능성을 제기하였고, 『논어여설』에서 본격적으로 각 편의 문헌 비평적 신뢰도를 논증하였다. 이는 『논어』를 불변의 고정된 경전으로 전제하던 전통적 인식에 대해 처음 이의를 제기한 것으로, 『논어』를 유동적 텍스트로 인식하는

1) 錢穆(1990), 16면.

2) 김용옥(2001), 154면.

3) 김영기(2012a), 174면.

4) 김용옥(2000), 76면.

관점의 출발점이라 할 수 있다.

후자의 흐름은 『論語集注』에서 주희가 胡寅(1098~1156)의 설을 수용해 “대체로 이 책의 후반 10편에는 빠지거나 잘못된 부분이 많다. [大抵此書後十篇多闕誤]”⁵⁾고 지적한 것에서 비롯되었으며, 명말에 羅喻義가 이를 전10편·후10편의 이분 구조로 정식화했다.⁶⁾ 이후 일본 古學의 대표 학자인 이토 진사이(伊藤仁齋, 1627~1705)가 『논어』 20편을 上論 10편과 下論 10편으로 나누며 본격화되었다고 정리된다. 이토 진사이는 상론 10편을 공자 사상의 본의를 담은 원형으로, 하론 10편을 후대에 부가된 속집으로 판별하였는데, 이는 공자 사상의 원형(상론)과 후대 변용(하론)을 구분하려 했던 사상사적 층위 구분의 초기 시도였다. 이후 오규 소라이(荻生徂徠, 1666~1728)·다자이 순다이(太宰春臺, 1680~1747) 등도 이 구분에 동의하였고, 근대 학자 다케우치 요시오(武內義雄, 1886~1966)는 『논어의 연구』⁷⁾에서 이토 진사이의 설과 최술의 설을 함께 검토하며 상론·하론 이분 구조를 재논증하였다. 이상의 두 흐름은 분석 층위는 다르지만, 모두 문헌 신뢰도 판별을 통해 『논어』의 사상적 원형에 접근하려 했다는 공통의 지향을 지닌다.

이러한 두 개의 흐름은 20세기 이후 특정 편을 중심으로 하는 연구로 이어지고 있는데, 대표적인 연구로는 1958년 사토 이치로(佐藤一郎)의 〈자장〉편을 대상으로 한 단락군 분석이 있다.⁸⁾ 이후 2008년에 김용욱 역시 『논어한글역주』에서 〈태백〉편을 단락 단위로 구분하였으나, 분류 근거에 대한 체계적인 분석은 제시하지 않았다.⁹⁾ 국내에서도 지금까지

5) 『論語集注』 16-12.

6) 『經義考』 卷221: 나유의 「논어분편」 2권 현존. 나유의의 자서에서 『논어』 20편은 둘로 나뉜다고 했다. [羅氏喻義論語分篇 二卷 存 喻義自序曰 論語二十篇 其別爲二分] [https://zh.wikisource.org/zh-hant/經義考_\(四庫全書本\)/卷221](https://zh.wikisource.org/zh-hant/經義考_(四庫全書本)/卷221).

7) 武內義雄(1939) 참조.

8) 佐藤一郎(1958) 참조.

특정 편을 중심으로 편장 체계의 성립 과정에 대한 연구들이 이루어져 왔다.¹⁰⁾ 그러나 이러한 연구들은 편장 구성에 대한 문제의식을 공유하면서도 『논어』를 고정된 경전으로 전제하는 전통적 인식의 틀 안에서 이루어진 것이어서, 20편 전체를 넘나드는 새로운 편장 구조의 재편성으로는 나아가지는 못했다. 편 경계로 분리된 장들이 실제로는 같은 주제 맥락에 속하는지, 혹은 같은 편 안에 묶여 있는 장들이 과연 의미적으로 응집된 단위를 이루는지에 답하기 위해서는 특정 편의 내부 구성을 검토하는 데 그치지 않고, 『논어』 전체를 하나의 코퍼스로 놓고 편 경계를 가로지르는 의미 단위를 탐색하는 접근이 필요하다.

이에 필자는 최술의 『논어여설』을 다룬 선행 논문에서, 『논어』의 원형을 파악하기 위해 현행 20편 체제의 재구성이 필요함을 제기하고, 현행 원본을 유지하면서도 최술의 변증 체계를 비롯한 다양한 비평 정보를 입력할 수 있는 디지털 비판 정보 구축 방법론을 제시한 바 있다.¹¹⁾ 그러나 최술의 비평은 신뢰도 평가에만 집중한 나머지, ‘이 장은 어떤 단위에 속해야 맞는가’라는 구조적 재편성의 문제로 나아가지 못했다는 한계가 있다. 최술이 나아가지 못했던 『논어』 전체의 재구성을 가장 체계적으로 시도한 연구가 바로 기무라 에이이치(木村英一, 1906~1981)¹²⁾의 章群 이론이다. 기무라는 편 경계를 넘나드는 재구성에는 이르지 못

9) 김용욱(2008) 참조.

10) 김윤조(2008); 박재복(2009); 김영기(2011); 김영기(2012b); 송병렬(2015); 이예성·이강재(2018) 참조.

11) 김서운(2025) 참조.

12) 기무라 에이이치는 일본의 중국철학 연구자다. 1930년 교토제국대학 문학부 중국철학사학과를 졸업하고, 1960년 「노자의 새로운 연구」로 박사학위를 취득하였다. 교토대학 조교수를 거쳐 1949년부터 오사카대학 문학부에서 중국철학을 강의하였으며, 1967년 정년퇴임 후 오사카대학 명예교수 및 오테몬가쿠인대학 교수를 역임하였다. 주요 저서로 『공자와 논어(孔子と論語)』, 『노자의 새로운 연구(老子の新研究)』, 『중국 철학의 탐구(中國哲學の探求)』 등이 있다.

했으나, 20편 전체를 대상으로 각 편 안에서 의미적으로 연관된 장들의 묶음을 명시적으로 분류하였고, 그 연관 방식을 연쇄·병렬·대응·집중·부록의 다섯 형식으로 유형화하여 일관된 기준으로 재구조화했다. 그러나 기무라의 원저 『공자와 논어』(1971)¹³⁾는 그 근거를 방대한 분량의 산문 서술 방식으로 기술하고 있어, 처음부터 끝까지 검토하는 것이 결코 쉽지 않다. 이러한 방식은 후속 연구자가 동일한 기준으로 이를 재현하거나 데이터로 구조화하기 어려우며, 정성적 독해에만 근거하고 있어 그 신뢰성이 계량적으로 검토되지 못한다는 한계를 지닌다.

이에 본 연구는 기무라의 장군 분류가 실제로 얼마나 유효한지를 측정하기 위해, 먼저 기무라의 서술을 대분류·소분류 두 층위로 구조화하여 『논어』 1~5편(<학이>~<공야장>)의 분류표를 작성하였다. 이는 기무라의 텍스트를 컴퓨터가 연산가능한 정형 데이터로 치환한 것이며, 이 중 층위의 군집 분석을 통해 정성적 비평과 정량적 분석 사이의 상관관계를 측정하기 위함이다.

첫 번째 층위는 고대 한문 형태소 분석기 Jiayan을 활용한 TF-IDF 기반 군집 분석으로 어휘 패턴의 유사성을, 두 번째 층위는 고대 문헌을 사전 학습한 언어모델 GuwenBERT 임베딩 기반 군집 분석으로 문맥적 의미의 유사성을 각각 측정한다. 1~5편을 대상으로 한 이 파일럿 연구에서 확립된 방법론은 이후 20편 전체에 적용될 것이며, 본 연구의 결과로 산출되는 타당성 정보는 향후 『논어』의 비판적 정본 재구성을 위한 객관적 지표로 활용될 것이다.

이상의 논의는 다음과 같이 진행된다. 2장에서는 기무라 장군 이론의 내용을 단장·장군·장군연관의 중층 구조와 다섯 가지 연관 형식을 중심으로 검토하고, 그 분류가 산문 형식으로 기술되어 재현·구조화가 어

13) 木村英一(1971) 참조.

렵다는 한계를 짚는다. 3장에서는 이 한계를 넘어서기 위해 1~5편의 장군 분류를 대분류·소분류의 두 층위로 정형화한다. 4장에서는 형태소·임베딩 두 층위의 군집 분석을 수행한 뒤 기무라의 분류와 대조하여, 그 타당성을 진단하는 방법을 제안한다. 5장에서는 분석 결과를 종합하고 후속 연구의 전망을 제시한다.

2. 기무라의 장군 이론

기무라는 『논어』의 각 편 내에서 의미적으로 서로 연관된 장들을 ㉠·㉡·㉢ 형태로 분류하는 장군 이론을 제시한 바 있다. 이 이론에서 그는 『논어』 20편 500여 장의 존재 방식을 두 가지로 구분한다.¹⁴⁾ 하나는 앞뒤 장과 관계없이 고립된 상태로 존재하는 ‘單章’이고, 다른 하나는 앞뒤 장과 의미 면에서 연관을 갖는 ‘章群’이다. 장군들 사이에도 층위가 있는데, 의미 면에서 앞뒤가 서로 이어진 일련의 장군을 ‘章群[聯關]’이라 부른다. 단장·장군·장군연관·복수의 장군연관이라는 중층 구조에 의해 각 편의 성격과 구조가 형성된다는 것이 기무라의 핵심 주장이다.

기무라는 ‘장군’이 발생한 이유를 두 가지로 해석한다. 하나는 전승을 위한 기억의 편의를 위함이다. 스승의 언행에 관한 단편적인 지식이 공자 문하에서 구전으로 전승되면서, 전승자가 의미상 유사하거나 서로 관

14) 『논어집주』의 경우 20편은 총 482장으로 이루어져 있으며, 기무라 역시 주희의 편장체계를 따랐다. 그러나 본고는 향후 분장 체계가 상이한 복수의 주석서를 비교 검토하는 연속 연구를 전제하므로, 특정 주석서의 분장 체계를 기준으로 삼을 경우 비교 작업의 일관성이 훼손될 수 있다. 이에 본고는 공개 접근이 가능하고 분석의 재현 가능성이 담보되는 Chinese Text Project(ctext.org)의 『논어』 편장 체계를 분석 저본으로 채택하였다.

런 있는 단문을 모아 기억하도록 했기 때문에 장군이 발생했다는 것이다. 다른 하나는 문헌 편집 단계에서의 의도적인 배열이다. 『논어』가 문자로 정착되는 과정에서 편찬자들이 산재한 단문들을 단순히 수집하는데 그치지 않고, 주제적 연관성과 논리적 흐름을 고려하여 의식적으로 장의 순서와 배치를 조정했다는 것이다. 기무라는 위의 두 요인이 상호 작용하며 현행 구조를 형성했다고 보았으며, 장군연관이 형성되는 방식을 아래와 같이 다섯 가지 형식으로 분류한다.¹⁵⁾

연쇄 방식은 인접한 장들이 특정 글자나 개념을 매개로 순차적으로 이어지는 형식이다. 〈헌문〉 제14편 35~38장이 그 예시이다. 35장과 36장은 ‘德’이라는 글자로 연결되고, 36장과 37장은 ‘怨’자를 통해 이어받으며, 37장과 38장은 다시 ‘天’과 ‘命’을 통해 연결된다.¹⁶⁾ 이처럼 연쇄 방식은 내용상의 통일성보다 글자 단위의 연결을 매개로 장군이 형성된다. 〈안연〉 제12편 1~5장에서도 같은 방식이 확인된다. 1~3장은 仁을 주제로 한 사제간 문답으로 장군㉓¹⁷⁾를 이루고, 3~5장은 사마우와 공자·자하의 문답으로 장군㉔¹⁸⁾를 이루는데, 장군㉓와 ㉔에 동시에 해당되는 3장이 두 장군의 점착제가 되어 1~5장 전체의 장군연관이 연쇄적으로 성립한다.

병렬 방식은 동일한 형식을 갖는 장군들이 나란히 배열되는 형식이

15) 기무라 에이이치(2020), 348~364면 참조.

16) 14-35. 子曰 驥不稱其力 稱其德也

14-36. 或曰 以德報怨 何如 子曰 何以報德 以直報怨 以德報德

14-37. 子曰 莫我知也夫 子貢曰 何爲其莫知子也 子曰 不怨天 不尤人 下學而上達 知我者其天乎

14-38. 公伯寮愬子路於季孫...子曰 道之將行也與 命也 道之將廢也與 命也 公伯寮其如命何

17) 장군㉓: 12-1. 顏淵問仁. 子曰... 12-2. 仲弓問仁. 子曰... 12-3. 司馬牛問仁. 子曰...

18) 장군㉔: 12-3. 司馬牛問仁. 子曰... 12-4. 司馬牛問君子. 子曰... 12-5. 司馬牛憂曰. 人皆有兄弟. 我獨亡. 子夏曰...

다. 〈술이〉 제7편처럼 ‘子曰’로 시작하여 공자의 태도 묘사로 끝나는 패턴이 반복적으로 나열되는 방식으로, 장군㉓ · ㉔ · ㉕ … ㉑가 동일한 형식으로 병렬되면서 편 전체의 구조를 이룬다.¹⁹⁾

대응 방식은 한 편의 처음과 끝에 주체적으로 호응하는 장이 배치되는 형식으로, 수미상관 구조라고도 할 수 있다. 〈학이〉 제1편이 그 전형이다. 1장은 “사람들이 알아주지 않더라도 서운해하지 않는다면 군자가 아니겠는가[人不知而不慍，不亦君子乎?]”로 끝나고, 16장은 “남이 자신을 알아주지 못함을 걱정하지 말고, 내가 남을 알지 못함을 걱정해야 한다.[不患人之不己知，患不知人也.]”로 끝난다. 서로 유사한 공자의 두 격언이 하나의 장군이 되어야 하겠지만, 이 문장들이 편수와 편말에 나누어 배치되면서 〈학이〉 제1편 하나를 통합하고 있다. 첫 장과 마지막 장이 군자의 태도를 보여준다는 면에서 주체적으로 호응하며, 편 전체를 하나의 완결된 의미 단위로 통합하는 역할을 한다. 아울러, 〈위정〉 제2편에서도 1장(爲政以德)과 21장(其爲爲政)이 덕치의 선언과 완성이라는 형태로 수미상관을 이룬다. 기무라는 이러한 수미상관 구조가 『논어』 전체에 걸쳐 반복적으로 나타난다는 점에서, 편찬자의 의도적 배열이 작용한 형식적 증거로 보았다.

집중 방식은 중심 장군㉓를 기준으로 이후의 장군㉔ · ㉕ · ㉖가 각각 장군㉓와 관련을 맺으며 순차적으로 병렬되는 형식인데, 〈위정〉 제2편이 그 대표적 사례이다. 1~4장(장군㉓)은 덕치 · 예치 및 시경의 정신과 호학하는 자세를 서술하는 중심 장군이다.²⁰⁾ 이후 5~8장의 孝에 대

19) 장군㉓: 7-1. 子曰 述而不作… 7-2. 子曰 默而識之… 7-3. 子曰 德之不脩… 7-4. 子之燕居 申申如也 夭夭如也

장군㉔: 7-5. 子曰 甚矣吾衰也… 7-6. 子曰 志於道… 7-7. 子曰 自行束脩以上… 7-8. 子曰 不憤不啓… 7-9. 子食於有喪者之側… 7-10. 子於是日哭 則不歌

20) 장군㉓: 2-1. 子曰 爲政以德… 2-2. 子曰 詩三百… 2-3. 子曰 道之以政… 2-4. 子曰 吾十有五而志於學…

한 문답(장군㉑²¹⁾)은 장군㉑ 3장의 禮 개념과 연결되고,²²⁾ 9~11장의 안회 칭찬·사람을 보는 법(장군㉒²³⁾)은 장군㉑ 4장의 호학 자세와 상응하며, 12~14장의 君子 주제(장군㉓²⁴⁾)는 장군㉑ 1·3장의 군자론과 다시 이어진다. 기무라는 <위정>편의 여섯 개 장군이 모두 장군㉑와 연관을 맺으며 순차적으로 병렬된 것으로 파악한다. 이처럼 집중 방식은 편 전체의 의미가 하나의 중심 장군으로 수렴되는 방사형 구조를 이룬다.

부록 방식은 일련의 장군연관이 일단 완결된 뒤에 다른 장 혹은 장군이 부록으로 덧붙여진 형식이다. 기무라는 <위정> 제2편의 경우, 1~21장에 걸친 장군㉑~㉑의 편찬이 일단 완결된 뒤에 정치의 기본조건을 논하는 22장과 예에 관한 23~24장이 나중에 부가되었을 것이라고 설명하면서, 22장²⁵⁾을 단장으로 23~24장을 별도의 장군㉒²⁶⁾로 분류하였다. 또한 <팔일> 제3편이 계씨와 삼가의 참례를 논하는 장으로 시작되다는 점에서, 기무라는 <위정>편의 장군㉒를 본래 <팔일>편에 소속되어야 할 장으로 보았다. 이는 주희가 이 두 장을 <팔일>편의 錯簡으로 지목한 견해를 따른 것이다.

기무라의 장군 이론은 두 가지 면에서 선구적 의의를 갖는다.

21) 장군㉑: 2-5. 孟懿子問孝 子曰 無違... 2-6. 孟武伯問孝... 2-7. 子游問孝... 2-8. 子夏問孝...

22) <위정> 제2편 3장에서 인도하기를 덕으로 하고, 가지런히 하기를 예로써 하면 백성들이 수치를 알고 바로잡히게 될 것이다.[道之以德 齊之以禮 有恥且格]라는 문장이 있다. 장군㉑의 머리 장인 5장의 뒤에 오는 "살아 계시면 예로 섬기고 돌아가시면 예로 장사 지내며, 예로 제사 지내는 것이다[生事之以禮 死葬之以禮 祭之以禮]"라는 말을 근거로, 효라는 것에 대해 공자가 말한 "위배되지 않는다[無違]"를 '예에 위배되지 않는 것'으로 규정할 수 있다. 이러한 점에서 기무라는 장군㉑의 5장이 장군㉑의 제3장을 이어받아 연결하고 있는 것으로 해석했다. (이하 기무라 에이이치(2020), 417면 참조.)

23) 장군㉒ 2-9. 子曰 吾與回言終日... 2-10. 子曰 視其所以... 2-11. 子曰 溫故而知新...

24) 장군㉓ 2-12. 子曰 君子不器 2-13. 子貢問君子... 2-14. 子曰 君子周而不比...

25) 단장 2-22. 子曰 人而無信 不知其可也...

26) 장군㉒ 2-23 子張問十世可知也... 2-24. 子曰 非其鬼而祭之 詔也...

첫째, 최술의 변증이나 사토 이치로·김용옥 등의 단락 분석이 특정 편이나 범위에 한정되었던 것과 달리, 기무라는 20편 전체를 일관된 기준으로 재구조화하는 작업을 처음으로 수행하였다. 이는 최술이 신뢰도 평가에 집중하여 구조적 재편성의 문제로 나아가지 못했던 한계를 넘어선 것으로, 『논어』를 유동적 텍스트로 인식하는 관점을 넘어 구조적 재편성의 단계로 심화시킨 시도라 평가할 수 있다.

둘째, 기무라는 장군연관을 연쇄·병렬·대응·집중·부록의 다섯 형식으로 유형화함으로써, 개별 직관에 머물던 분류를 명시적 기준의 문제로 전환하였다. 이는 정성적 독해를 일정한 규칙으로 환원하려는 시도로서, 그 기준이 후속 연구자에 의해 검증·재현되거나 데이터로 구조화될 수 있는 이론적 토대를 마련하였다는 데 의의가 있다. 다만 기무라 자신은 그 기준을 방대한 산문 서술 속에 분산시켜 뒀으로써 이 가능성을 충분히 실현하지는 못하였다. 본 연구가 그의 분류를 정형 데이터로 옮기고 계량적으로 교차 검증하려는 것은, 바로 이 미실현된 재현 가능성을 이어받아 구체화하려는 작업이다.

3. 장군 이론의 정형화

본 장에서는 기무라의 서술을 토대로 장군 분류를 대분류·소분류의 두 층위로 정형화하여 『논어』 제1~5편의 분류표를 구성한다. 기무라의 원저는 장군 분류를 산문적 서술 속에 분산하여 기술하고 있어, 이를 정량적으로 검증하기 위해서는 기계 가독형(Machine-readable) 구조로의 치환이 선행되어야 한다. 이 정형화 작업은 기무라 장군론의 첫 번째 디지털 구조화 시도이며, 4장에서 수행할 군집 분석의 직접적인 비교 준

거로 기능한다.

기무라 이론의 정형화 모델은 각 편 내의 편집 층위와 주제적 응집도를 반영하여 대분류·소분류 두 단계의 계층 구조로 설계하였다. 대분류는 각 편 내에서 성격이 확연히 구별되는 상위 군집을 가리킨다. 이 모델에서는 기무라가 명시적으로 구분하지 않은 구간이라도 그의 분석적 논지에 부합한다면 전체 흐름에 따라 동일한 방식으로 편입하여 분류의 일관성을 확보하였다. 예컨대 〈학이〉편은 대분류를 따로 두지 않았으나, 기무라는 제1장부터 제16장까지를 하나의 온전한 편장 구성으로 인식하고 있으므로 01A 하나로 구분했고, 〈위정〉편은 02A(주요 장군연관)와 02B(착간)으로 구분되므로 두 개의 대분류로 확정하였다. 대분류가 편 내 이질적 구간의 존재를 표시한다면, 소분류는 그 안에서 공통 주제나 장군연관 방식으로 묶이는 세부 군집을 나타낸다. 이 대분류·소분류 체계로 〈학이〉부터 〈공야장〉까지 다섯 편에 적용하여 정리한 것이 〈표 1〉이다.²⁷⁾

대분류	소분류	장 범위	주요내용
01A	01a	LY0101~0102	군자 양성·효제와 인의 기본
	01b	LY0103~0104	군자가 하지 말아야 할 것·군자의 성실함 예시(증자)
	01c	LY0105~0107	정치·도덕 실천·인격완성
	01d	LY0108~0110	군자의 인간상·정교 효과
	01e	LY0111~0113	예의 실천·쓰임·보편적 가치
	01f	LY0114~0116	군자의 호학·호례
02A	02a	LY0201~0204	덕치·예치·군자의 학문 기반
	02b	LY0205~0208	효에 대한 문답

27) 기무라 에이이치(2020), 367~472면 참조. 아울러, 본 연구에서 『논어』 각 장의 표기는 편의상 'LY+편번호+장번호'의 형식을 따른다. 예컨대 LY0101은 제1편 제1장을, LY0516은 제5편 제16장을 가리킨다.

대분류	소분류	장 범위	주요내용
	02c	LY0209~0211	안연의 호학·사람 보는 법
	02d	LY0212~0214	군자의 학문 태도
	02e	LY0215~0218	학문의 방법
	02f	LY0219~0221	정치에 대한 문답
	-	LY0222	정치의 기본조건(단장)
02B	02g	LY0223~0224	착간
03A	03a	LY0301~0303	참례 비판·예악과 인간 본질
	03b	LY0304~0306	예의 근본·질서 정신
	03c	LY0307~0308	군자의 경쟁·예와 시경
	03d	LY0309~0313	체제사·제사 태도
	03e	LY0314~0318	주나라 예 기준·고려 실천
	-	LY0319	군신 관계의 예(단장)
03B	03f	LY0320~0322	1차 부록
	03g	LY0323~0326	2차 부록
04A	04a	LY0401~0407	인에 대한 집중 서술
	04b	LY0408~0409	도에 대한 설명
	04c	LY0410~0411	군자의 태도
	04d	LY0412~0414	공리주의·예치·군자의 걱정
04B	04e	LY0415~0425	증자 문인 편집본·군자의 일
	-	LY0426	간언의 절도(단장)
05A	05a	LY0501~0512	공자의 제자 평
	05b	LY0513~0514	자공·자로 부록
05B	05c	LY0515~0521	열국 대부에 대한 평
	05d	LY0522~0525	제자·인물 총괄 비평 부록
	05e	LY0526~0528	공자 자신의 뜻 서술

〈표 1〉 기무라 장군 분류표 (『논어』 제1~5편)

각 편의 구조적 특징을 살펴보면 다음과 같다.

〈학이〉편은 01a부터 01f까지 여섯 소분류가 모두 A군 본문으로만 구성되어 있다. 이 흐름은 단순한 나열이 아니라 군자의 내면 덕목에서 출발하여 사회적 실천으로 확장되는 논리적 완결성을 보여주며, B군 없이 A군만으로 완결된 구조는 〈학이〉편이 편집상 가장 정제된 형태임을 시사한다.

〈위정〉편은 02a~02f의 A군 안에서 덕치·효·호학·학문 태도·학

문 방법·정치 문답이라는 복합 주제가 전개된다. 여기에 LY0222 단장과 착간 02g가 B군으로 덧붙여진 구조는, 〈학이〉편보다 편찬 층위가 복잡함을 보여준다. 이는 논술의 핵심부 이외에 이질적 전송 자료가 더 많이 통합된 것으로 보인다.

〈팔일〉편은 A군(03a~03e)에서 禮의 근본 정신을 집중적으로 논한 뒤, B군에서 군신 관계의 예라는 단장(LY0319)과 두 차례의 부록(03f·03g)을 배치한다. 부록이 1차와 2차로 구분된다는 점은 텍스트 형성 과정이 최소 두 단계의 보충 편찬을 거쳤음을 함의한다.

〈이인〉편은 인(仁) 개념의 집중 서술(04a~04b)로 A군이 형성되고, B군에서 다양한 화자와 주제의 어록이 뒤따른다. 이 다양성은 仁이라는 핵심 주제 주변에 여러 제자 계통의 자료가 보충 편집된 구조임을 보여준다.

〈공야장〉편은 공자의 제자 평(05a)과 자공·자로 부록(05b)으로 A군이 구성되고, B군에서 열국 대부 평·제자 인물 총괄 비평·공자 자신의 뜻 서술이 순차적으로 배치된다. A군과 B군의 경계가 앞의 네 편보다 훨씬 세밀하게 나뉜다는 점이 두드러지는데, 이는 〈공야장〉편의 편찬이 다단계 보충 과정을 통해 이루어졌음을 보여준다.

이러한 정형화 과정을 통해, 필자는 『논어』를 단순한 어록 집성이 아닌 특정한 편찬 원리에 의해 형성된 구조적 텍스트로 인식하게 되었다. 결국 기무라의 장군 이론은 『논어』가 어떤 순서로, 어떤 원리에 의해, 몇 단계의 편찬 과정을 통해 지금의 형태를 갖게 되었는지를 텍스트 내부 구조로부터 논증하고자 했던 것으로 보인다.

4. 이중 층위 군집 분석과 장군 분류의 타당성 진단

1) 분석 방법 및 이중 층위 설계의 목적

본 연구가 지향하는 편장 구조의 재편성은 형태적·사상적 일관성에 근거한 체계화를 목표로 한다. 이를 위해 본고는 현행 장들을 의미 단위로 해체하여 재편성하되, 기무라의 장군 이론을 그 출발점으로 삼는다. 그러나 기무라의 분류는 정성적 독해에 근거한 것으로, 다른 연구자가 동일한 텍스트를 읽고 다른 결론에 도달할 가능성이 있다. 예컨대 김운조는 사토 이치로(佐藤一郎)라는 일본학자의 연구방법론을 계승하여 『논어』 편찬 과정을 고찰하는 과정에서 기무라의 이론을 참고하면서도, 〈학이〉편 전체를 ‘孔門學生便覽’ 내지 ‘孔門學校規則’으로 규정한 기무라의 논지에는 공감하기 어렵다고 기술한 바 있다.²⁸⁾ 이처럼 동일한 텍스트를 두고 연구자마다 상이한 결론에 도달하는 정황은, 기무라의 분류를 채택하든 거부하든 그 판단이 정성적 독해에만 머무를 경우 객관적 재현이 어렵다는 점을 보여 준다. 더욱이 본 연구는 기무라의 장군론을 편 경계를 넘어서는 개념 단위 분류로 확장하려 하므로, 그 확장의 근거를 정성적 판단에만 의존할 수는 없다.

이러한 문제를 해결하기 위해 본 연구는 군집 분석(Clustering)을 외부 준거로 도입한다. 군집 분석은 텍스트 사이의 어휘·의미 유사성을 계량적으로 측정하여 유사한 것끼리 자동으로 묶는 방법으로, 연구자의 사전 판단이나 개입 없이 계산 결과만으로 분류가 이루어진다. 다시 말해 연구자가 장군 이론과 같은 분류 기준을 사전에 설정하지 않아도 텍스트 데이터 자체의 패턴에 따라 군집이 형성된다는 점에서, 정성적 독

28) 김운조(2008), 167면 참조.

해를 보완하는 객관적 재현 가능성을 확보할 수 있다.

이때 두 층위의 분석 결과가 기무라의 분류와 맞물리는 양상은 단순한 일치·불일치의 이분 구조가 아니라 세 가지 구간으로 갈라지는데, 그 구체적 식별과 판정은 4절에서 다룬다. 이러한 구간 구분은 계량적 분석이 인문학적 판단을 대체하는 것이 아니라, 어느 지점에서 인간 연구자의 개입이 필요한지를 가시화하는 진단 도구로 기능한다는 본 연구의 입장을 구체화한 것이다.

구체적으로 이 검증은 세 가지 질문에 답하는 것을 목적으로 한다. 첫째, 기무라가 동일 장군으로 확정한 장들이 어휘·의미 층위에서도 실제로 유사한가. 이는 기무라가 장군으로 묶은 근거가 계량적으로 지지되는지를 묻는 것이다. 둘째, 기무라가 단장 또는 부록으로 분류한 장들이 본편 장군과 계량적으로도 이질적으로 분리되는가. 이는 착간·연문으로 의심되어 온 장들의 이질성이 계량적으로도 확인되는지를 묻는 것이다. 셋째, 클러스터링 결과와 기무라의 분류가 불일치하는 구간은 어디이며, 그 불일치는 무엇을 의미하는가. 이는 군집 분석을 정답으로 전제하지 않되, 양자가 갈라지는 지점을 전통 주석과 선행 연구에 비추어 어떻게 해석할지를 묻는 것이다. 첫째와 둘째 질문이 기무라 분류의 검증에 해당한다면, 셋째 질문은 그 검증을 넘어 새로운 분류 체계의 확정과 조정으로 나아가기 위한 해석적 작업에 해당한다.

이러한 연구 질문에 대한 답을 찾기 위해 검증 방법을 '형태소 분석 기반'과 '임베딩 기반'의 두 층위로 설계하였다. 형태소 분석은 어휘 사용 패턴의 표면적 유사성을 포착하는 데 유리하며, 언어모델 임베딩은 텍스트 이면의 문맥적·철학적 유사성을 측정하는 데 탁월하다. 이처럼 서로 다른 측면을 포착하는 두 층위를 교차 활용함으로써, 기무라의 분류가 어휘 층위와 의미 층위에서 각각 어떻게 지지되거나 어긋나는지를 구분

해 살피고자 한다.

아울러, 본 연구의 분석 대상을 1~5편으로 한정할 데는 방법론적 근거가 있다. 예비 분석으로 상론 전체인 1~10편 258장을 대상으로 예비 분석을 수행한 결과, TF-IDF²⁹⁾ 기반 군집 결과에서 전체 장의 82%에 달하는 212장이 단일 군집으로 흡수되어 사실상 변별력을 상실하였다. 이는 분석 범위가 확대될수록 동일한 단어의 가중치가 달라지는 TF-IDF의 구조적 특성에서 비롯된다. 즉, 코퍼스가 커질수록 『논어』 전반에 두루 등장하는 夫子·孔子·弟子 등 호칭어와 仁·禮·學·君子 같은 사상 어휘는 '흔한 단어'로 간주되어 페널티를 받게 되며, 그 결과 변별 키워드가 사라진 다수의 장이 하나의 군집으로 흡수된다. 이때 분석 결과는 텍스트 자체의 성격이 아니라 분석자가 설정한 범위에 좌우되므로, 객관적 재현성을 확보하기 위해서는 코퍼스 규모를 적절히 통제할 필요가 있다. 반면 다섯 편 분량의 군집 분석에서는 편별 주제가 비교적 뚜렷하게 구분되며 주제적 응집도가 높아 계량 분석의 변별력이 충분히 확보되었다. 따라서 본고는 1~5편을 파일럿 범위로 설정하여 방법론을 확립하고, 이후 다섯 편씩 단계적으로 나누어 20편 전체로 확장 적용하는 접근을 택하였다.³⁰⁾

29) TF-IDF(Term Frequency-Inverse Document Frequency)는 문서 집합에서 특정 단어의 중요도를 측정하는 가중치 산출 방식이다. TF(단어 빈도)는 해당 문서 안에서 특정 단어가 얼마나 자주 등장하는지를 측정하고, IDF(역문서 빈도)는 그 단어가 전체 문서 집합에서 얼마나 드물게 등장하는지를 측정한다. 두 값을 곱함으로써, 특정 문서에는 자주 등장하지만 다른 문서에는 잘 등장하지 않는 단어—즉 해당 문서의 개념적 특성을 잘 대표하는 단어—에 높은 가중치가 부여된다. 본 연구에서는 『논어』 각 장을 하나의 문서로 간주하고, 각 장의 개념적 특성을 대표하는 핵심 어휘에 높은 가중치를 부여하는 방식으로 장들 사이의 어휘 유사성을 측정하는 데 활용하였다.

30) 본 논문의 군집분석에 사용된 데이터와 코드는 다음 깃허브 저장소에서 확인할 수 있다: <https://github.com/sy527991/Analects-Clustering>

2) Jiayan 형태소 기반 군집 분석 결과

첫 번째 층위는 고대 한문 형태소 분석기 Jiayan(甲言)³¹⁾을 활용한 TF-IDF 기반 군집 분석이다. ‘형태소 분석기’란 각 장을 품사 단위로 분절하는 도구를 말한다. 본 연구에서는 분절 결과 가운데 실제 의미를 담는 명사류(n, nh, ns)만을 추출하여 개념 단위의 어휘 분포를 측정하였다. 즉, 허사·접속사 등 문법 기능어를 제거한 상태에서 각 장을 문서로 간주하고 단어의 출현 빈도와 역문서 빈도를 결합한 TF-IDF 벡터를 산출함으로써 표면적 어휘 패턴의 유사성을 측정하는 방식이다. 다만, 기능어를 제거한 명사류 벡터로 군집을 산출하되, 군집이 형성된 뒤에는 각 군집의 문형 경향을 별도로 관찰했다. 따라서 군집을 가르는 입력값은 명사류에 한정되며, 뒤에 제시할 <표 2>의 문형 특징은 이렇게 형성된 군집을 사후적으로 관찰한 결과이다.

형태소 분석 기반의 이 층위는 같은 글자가 반복 등장하는 장들이 실제로 묶이는지를 확인하는 데 유효하며, 기무라가 연쇄 방식의 근거로 제시한 공유 글자의 역할을 계량적으로 검토할 수 있다. 예컨대 <학이>편 첫 장의 형태소 분석 결과는 다음과 같이 출력된다.³²⁾

31) Jiayan(甲言)은 ‘갑골문언(甲骨文言)’에서 이름을 딴 고대 한문 자연어 처리(NLP) 전문 툴킷이다. 기존의 중국어 NLP 툴은 대부분 현대 중국어(백화문)를 기준으로 설계되어 있어, 조사나 어휘 체계가 완전히 다른 고대 한문을 처리할 때 분절 오류가 심하고 정확도가 매우 떨어졌다. Jiayan은 이러한 한계를 극복하고, 고대 문헌 데이터를 디지털 정보로 쉽게 전환(인공지능 분석, 텍스트 마이닝 등)할 수 있도록 돕기 위해 만들어졌다. 현재 오픈소스 공유 플랫폼인 GitHub를 통해 전 세계 개발자와 연구자들에게 제공되고 있다. <https://github.com/jiaeyan/Jiayan>

32) Jiayan 고전 한문 품사 태그셋은 다음의 url에서 확인할 수 있다. <https://github.com/jiaeyan/Jiayan/tree/master/jiayan/postagger> 형태소 분석 결과를 살펴보면, 동사(verb)인 學이 명사(general noun)로 표시되고, 부사(adverb)인 不가 명사로 표시되는 것과 같이, 문맥에 따라 품사가 달라지는 글자에 태깅 오류가 종종 발생한다. Jiayan과 같은 통계적·딥러닝 기반 모델은 학습 데이터에서 學이 ‘학문’·‘배움’이라는 명사로 쓰인 사례가 많을 경우, 문맥을 완전히 파악하지 못하고

子(n) 曰(v) 學(n) 而(c) 時(v) 習(v) 之(u) 不(n) 亦(d) 說(v) 乎(u) 有(v) 朋(n) 自(r) 遠(v) 方(n) 來(v) 不(n) 亦(d) 樂(v) 乎(u) 人(n) 不(d) 知(v) 而(c) 不(v) 慍(n) 不(n) 亦(d) 君子(n) 乎(u)

군집화는 K-Means 알고리즘으로 수행하였으며, 군집 수 K는 엘보우 방법(Elbow Method)에 따라 K=5로 설정하였다.³³⁾ 1~5편 전체 121장을 5개 군집으로 분류한 결과는 다음의 <표 2>와 같다.

군집	장수	핵심어휘 및 문형특징	대표 개념
Cluster 0 ³⁴⁾	25	여기조사 矣 활용, 상태·판단의 단정적 서술	道·孝
Cluster 1 ³⁵⁾	10	‘~하는 자(者)’, ‘~할 수 있다(能)’ 결합	仁
Cluster 2 ³⁶⁾	52	也·之·其 빈출, 진행적 평서문 구조	人·知
Cluster 3 ³⁷⁾	26	不·乎·則 빈출, 질문·조건·통치 논의	民
Cluster 4 ³⁸⁾	7	행동(行)과 이익(利)의 대조 구조	君子·小人

<표 2> Jiayan 형태소 분석 기반 클러스터링 결과

동사 용법에서도 명사로 예측하는 것이다. 그러나 이러한 오류는 말뭉치(corpus) 전반에 걸쳐 일관되게 나타나는 체계적 오류로서 장들 사이의 상대적 어휘 분포 패턴 비교를 목적으로 하는 TF-IDF 분석의 결과에 결정적 영향을 미치지 않는다. 다만 어휘 층위 분석이 갖는 이러한 한계를 보완하기 위해 GuwenBERT 임베딩 기반 분석을 병용하였다.

- 33) K-Means는 데이터를 사전에 지정한 K개의 군집으로 분류하는 클러스터링 알고리즘이다. 각 데이터 포인트를 가장 가까운 군집 중심(centroid)에 배정하고, 배정된 데이터들의 평균값으로 군집 중심을 갱신하는 과정을 반복하여 군집 내 유사성을 최대화한다. K값의 결정은 군집 수를 늘려도 군집 내 분산이 크게 줄어들지 않는 지점을 시각화하는 엘보우 방법과, 각 장이 자신의 군집에 얼마나 적합하게 배치되었는지를 수치화하는 실루엣 계수(silhouette coefficient)를 함께 적용하여 K=5를 최적값으로 선정하였다.
- 34) Cluster 0 군집에 해당하는 25개의 장은 다음과 같다.(이하, 같은 방식으로 나열)
LY0103, LY0108, LY0111, LY0113, LY0114, LY0202, LY0203, LY0205, LY0206, LY0207, LY0211, LY0301, LY0302, LY0316, LY0325, LY0404, LY0408, LY0409, LY0420, LY0423, LY0502, LY0506, LY0513, LY0515, LY0526
- 35) Cluster 1: LY0112, LY0322, LY0401, LY0402, LY0403, LY0405, LY0406, LY0413, LY0511, LY0519

각 군집의 성격을 살펴보면, Cluster 0은 공자의 발언 중에서도 道와 孝를 핵심 개념으로 하며, 어기조사 矣를 통해 어떤 상태나 판단을 단정적으로 서술하는 문형 패턴이 두드러진다. Cluster 1은 仁과 관련된 장들이 강하게 묶였으며, ‘~하는 자(者)’·‘~할 수 있다(能)’ 등의 표현과 결합하여 인을 실천하는 인물이나 그 가능성을 서술하는 패턴을 보인다. Cluster 2는 사람(人)에 대한 知(知)을 핵심 개념으로 다루되, 也·之·其 등의 대명사와 종결사가 빈번하게 나타나는 고전 한문의 가장 전형적인 평서문 구조로 묶인다. Cluster 3은 民을 다스리는 방법에 대한 논의를 핵심 개념으로 하며, 부정부사 不·의문조사 乎·조건절 則을 통해 질문을 던지거나 조건을 따지는 문형 패턴이 두드러진다. Cluster 4는 君子와 小人이 핵심 키워드로 등장하며, 이들의 행동(行)과 이익(利)을 뚜렷하게 대조하는 특수한 구조의 장들로 구성된다. 이 군집은 1~5편에서는 소규모로 나타나지만, 20편 전체를 분석 대상으로 삼을 경우 더 뚜렷한 군집을 형성할 것으로 예상된다.

이 결과를 기무라의 장군 분류와 대조하면 몇 가지 주목할 만한 지점이 확인된다. 우선 Cluster 1은 장군 이론 04a(LY0401~0407)³⁹⁾를

36) Cluster 2: LY0102, LY0110, LY0115, LY0116, LY0201, LY0208, LY0210, LY0213, LY0216, LY0217, LY0218, LY0221, LY0222, LY0223, LY0224, LY0304, LY0305, LY0307, LY0308, LY0309, LY0310, LY0311, LY0312, LY0313, LY0315, LY0317, LY0318, LY0319, LY0323, LY0324, LY0407, LY0410, LY0414, LY0421, LY0422, LY0426, LY0501, LY0503, LY0504, LY0505, LY0507, LY0508, LY0509, LY0512, LY0514, LY0516, LY0518, LY0521, LY0522, LY0524, LY0525, LY0528

37) Cluster 3: LY0101, LY0104, LY0105, LY0106, LY0107, LY0109, LY0204, LY0209, LY0212, LY0214, LY0215, LY0219, LY0220, LY0303, LY0314, LY0320, LY0321, LY0326, LY0415, LY0417, LY0418, LY0419, LY0425, LY0517, LY0523, LY0527

38) Cluster 4: LY0306, LY0411, LY0412, LY0416, LY0424, LY0510, LY0520

39) 04a<이인>: 4-1. 子曰 里仁爲美... 4-2. 子曰 不仁者不可以久處約... 4-3. 子曰 唯仁者能好人... 4-4. 子曰 苟志於仁矣... 4-5. 子曰 富與貴 是人之所以所欲也... 4-6. 子曰 我未見好仁者... 4-7. 子曰 人之過也...

중심으로 기무라가 仁 개념의 집중 서술로 묶은 장군과 높은 일치율을 보인다. 이는 기무라의 정성적 분류가 어휘 층위에서도 지지됨을 보여주는 사례다. 반면 Cluster 2는 52개 장이 하나의 군집으로 묶인 가장 큰 군집으로, 기무라가 서로 다른 장군으로 분류한 장들이 혼재되어 있다. 이는 TF-IDF 기반 클러스터링이 표면적 어휘 패턴의 유사성을 측정하는 방식인 만큼, 고전 한문의 전형적 문형을 공유하는 장들이 내용과 무관하게 하나로 묶이는 한계를 보여준다. 이러한 한계는 의미 층위의 유사성을 측정하는 임베딩 기반 분석으로 보완될 필요가 있다. 기무라 분류와의 정밀한 대조는 의미 층위의 유사성을 측정하는 임베딩 기반 클러스터링 결과와 함께 제4절 ‘일치·불일치 구간의 해석과 신뢰도 판정’에서 종합적으로 검토한다.

3) GuwenBERT 임베딩 기반 군집 분석 결과

두 번째 층위는 고대 한문을 사전 학습한 언어모델(Pre-trained Language Models)인 GuwenBERT⁴⁰⁾를 활용한 임베딩 기반 군집 분석이다. 각 장은 768차원의 고차원 벡터로 표현되며, 이 벡터는 어휘 분포의 표면 패턴을 넘어 문장 전체의 문맥과 철학적 의미를 함께 반영한다. 따라서 본 층위는 仁과 禮처럼 글자는 다르지만 개념적으로 긴밀히 연결된 장들의 관계를 드러낼 수 있다는 강점을 지닌다. 군집화는 형태소 분석과의 비교를 위해 동일하게 K-Means 알고리즘으로 5개 군집(K=5)을 산출하였으며, 그 결과는 다음의 <표 3>과 같다.

40) GuwenBERT(古文BERT)는 始知閣古代文獻(중국의 고전 문헌 디지털화 프로젝트) 코퍼스를 기반으로 학습된 모델로, 고대 문헌 15,694종, 약 17억 자(字)의 데이터를 포함하며 모든 번체자는 간체자로 변환 처리되었다. <https://github.com/Ethan-yt/guwenbert>

군집	장수	특징	대표 개념
Cluster 0 ⁴¹⁾	16	학문의 목적과 도덕적 실천을 통한 군자됨의 당위적 논의	學・君子
Cluster 1 ⁴²⁾	26	덕치・효도・인물 평가 등 관계 속 역할 수행의 철학적 사유	德治・孝
Cluster 2 ⁴³⁾	22	仁・禮의 본질 논의 및 제자의 학문 태도・정교 효과에 대한 공자의 평가	仁・禮
Cluster 3 ⁴⁴⁾	8	단호한 경구와 경계해야 할 태도의 선언적 발화	경구・단정
Cluster 4 ⁴⁵⁾	48	道・知에 관한 보편적 담론. 서술 맥락의 일반성이 응집 배경	道・知

〈표 3〉 GuwenBERT 임베딩 기반 클러스터링 결과

각 군집의 성격을 살펴보면, Cluster 0은 學과 君子를 핵심 개념으로 하여, ‘어떻게 행동해야 군자인가’라는 당위적 물음을 공유하는 장들이 모인 군집이다. 학문의 목적과 도덕적 실천이 하나의 의미 축을 이루며, 교육적 논의가 응집의 중심을 이룬다. Cluster 1은 德治와 孝를 핵심

41) Cluster 0: LY0101, LY0104, LY0106, LY0108, LY0218, LY0219, LY0220, LY0222, LY0304, LY0313, LY0317, LY0326, LY0402, LY0413, LY0418, LY0522

42) Cluster 1: LY0116, LY0201, LY0202, LY0206, LY0210, LY0213, LY0214, LY0301, LY0302, LY0305, LY0314, LY0316, LY0319, LY0403, LY0414, LY0416, LY0420, LY0422, LY0424, LY0426, LY0514, LY0517, LY0518, LY0520, LY0524, LY0528

43) Cluster 2: LY0109, LY0211, LY0215, LY0224, LY0303, LY0310, LY0312, LY0318, LY0320, LY0325, LY0401, LY0407, LY0409, LY0411, LY0417, LY0419, LY0421, LY0503, LY0506, LY0511, LY0523, LY0527

44) Cluster 3: LY0103, LY0212, LY0216, LY0404, LY0408, LY0412, LY0423, LY0425

45) Cluster 4: LY0102, LY0105, LY0107, LY0110, LY0111, LY0112, LY0113, LY0114, LY0115, LY0203, LY0204, LY0205, LY0207, LY0208, LY0209, LY0217, LY0221, LY0223, LY0306, LY0307, LY0308, LY0309, LY0311, LY0315, LY0321, LY0322, LY0323, LY0324, LY0405, LY0406, LY0410, LY0415, LY0501, LY0502, LY0504, LY0505, LY0507, LY0508, LY0509, LY0510, LY0512, LY0513, LY0515, LY0516, LY0519, LY0521, LY0525, LY0526

개념으로 하여, 정치·사회 속에서 인간과 군자의 역할을 다루는 군집이다. 덕치주의·인물 평가·효도 등 관계 중심의 철학적 사유를 담은 장들이 집중적으로 분포한다. Cluster 2는 유학의 핵심 개념인 仁과 禮의 본질에 대한 논의가 집중된 군집으로, 공자 자신의 시각에서 제자들의 학문적 태도와 政敎의 효과를 평가하고 관찰하는 내용이 주를 이룬다. Cluster 3은 규모가 가장 작으나 가장 특징적인 군집으로, “巧言令色(LY0103)”, “君子不器(LY0212)”, “朝聞道, 夕死可矣(LY0408)” 등 공자의 단호한 경구와 강하게 경계해야 할 태도를 선언하는 장들이 발화 방식의 유사성을 매개로 응집된다. Cluster 4는 전체 장의 절반 가까이 포함된 가장 큰 군집으로, 道와 知에 관한 『논어』의 가장 보편적인 담론들이 해당하며, 특정 주제보다는 서술 맥락의 일반성이 응집의 배경으로 작용한다.

이상의 결과를 형태소 분석 기반 군집과 비교하면, 양자는 서로 다른 차원의 텍스트 구조를 드러낸다. 형태소 분석 기반 군집이 어기조사(也·乎·矣)나 부정어(不)의 배치와 같은 문법적 구조에 민감하게 반응했다면, GuwenBERT 군집은 ‘이 장이 군자다움의 당위를 논하는가, 관계 속의 역할을 다루는가, 선언적 경구인가’와 같이 텍스트 이면의 주제적·철학적 의미 유사도를 포착하여 분류한다. 이로써 『논어』 텍스트가 어휘 층위와 의미 층위에서 각각 다른 응집 구조를 갖는 다층적 텍스트임을 보여준다.

이 결과를 기무라의 장군 분류와 대조하면 형태소 분석 기반 군집과는 다른 양상이 확인된다. 가장 주목할 만한 사례는 Cluster 2의 仁·禮 관련 장들이다. 이 군집은 기무라가 〈이인〉편과 〈팔일〉편에서 각각 별도의 장군으로 분류한 장들을 편 경계를 넘어 하나의 군집으로 묶는다. 이는 기무라가 편 내부 구조 분석에서 포착하지 못한 편 경계를 넘나드는

개념적 연관을 의미 층위에서 확인한 사례이며, 본 연구가 기무라의 장군론을 편 경계를 넘어서는 개념 단위 분류로 확장하려는 방향과도 부합한다.⁴⁶⁾ 반면 Cluster 3의 경구·단정 군집은 기무라의 장군 분류와 직접 대응하는 단위를 찾기 어려운데, 이는 공자의 경구적 발화 방식이라는 문체적 특성이 기무라의 의미 연관 분류와는 다른 기준으로 장들을 묶기 때문으로 해석된다. 그러므로 임베딩 기반 분석은 기무라가 다루지 않은 문체 차원의 군집까지 식별해 낼 수 있음을 보여준다.

4) 정합·부분 일치·인문학적 재검토

앞에서 수행한 두 층위의 군집 분석 결과를 기무라의 장군 분류와 대조하면, 분석 결과는 단순한 일치나 불일치의 이분 구조가 아니라 정합 구간, 부분 일치 구간, 인문학적 재검토 구간 이렇게 세 가지 구간으로 식별된다.

기무라의 정성적 판단이 형태·의미 두 계량 분석 결과와 모두 일치하는 정합 구간은, 분류의 타당성이 어휘 층위와 의미 층위에서 독립적으로 검증된 사례에 해당한다.

가장 뚜렷한 사례는 〈이인〉 제4편 04a의 仁 집중 서술 구간이다. 이 장군은 Jiayan 형태소 기반 분석에서 Cluster 1로, GuwenBERT 기반 분석에서 Cluster 2로 묶이며 두 층위 모두에서 높은 응집도를 보였다.

46) 클러스터링 수치만 놓고 보면 LY0319의 이질성이 즉각적으로 드러나지 않는 것처럼 보일 수 있으나, 좀 더 세밀하게 살펴보면 형태소 분석에서 LY0319는 LY0314·0316장과 다른 군집에, 임베딩 분석에서는 LY0315·0317·0318장과 다른 군집에 배치된다. 두 층위 모두에서 LY0319가 인접 장들과 연관되게 다른 군집에 놓인다는 점은 주목할 만하며, 여기에 내용적으로도 주나라 예의 기준과 고레 실천을 논하는 제14~18장과 달리, LY0319는 군신 관계를 다루며 발화 형식도 구별된다는 점을 함께 고려하면, 기무라의 단장 판단이 어휘·의미 층위에서 부분적으로 지지된다고 볼 수 있다.

곧 어휘 차원에서는 仁·者·能 등의 핵심 어휘가 반복되고, 의미 차원에서는 仁의 본질에 대한 논의라는 주제적 응집성이 함께 확인된다. 이는 기무라가 정성적 독해에 근거하여 제시한 분류 기준이 두 층위에서 동시에 지지되는 가장 강력한 사례이며, 본 연구의 분류 체계에서 확정적 단위로 채택될 수 있다. 또한 <팔일> 제3편 03B 부록 구간⁴⁷⁾과 일부 단장 후보들은 두 분석 모두에서 본편 장군과 이질적인 군집으로 분류되어, 기무라가 본편과 분리한 정성적 판단이 계량적으로 지지되었다.

두 계량 분석 가운데 한 측면에서만 기무라의 분류와 부합하는 부분 일치 구간은, 기무라의 분류가 어휘 또는 의미 중 한 층위에서만 변별력을 갖는 사례에 해당한다. 이는 두 분석 방법이 서로 다른 차원의 텍스트 구조를 포착함으로써 드러나는 자연스러운 현상이다.

대표적 사례가 LY0319⁴⁸⁾이다. 이 장은 형태소 분석에서 인접 장 LY0314·0316과 다른 군집에, 임베딩 분석에서 인접 장 LY0315·0317·0318과 다른 군집에 배치된다. 곧 두 층위 모두에서 인접 장들과 일관되게 다른 군집에 놓이지만, 어휘 층위와 의미 층위에서 이질성을 드러내는 양상은 서로 다르다. 내용적으로도 LY0319는 주나라 예의 기준과 고레 실천을 논하는 인접 장(LY0314~0318)⁴⁹⁾과 달리 군신 관계를 다루며 발화 형식도 구별된다. 이러한 패턴은 기무라의 단장 판단이 어휘·의미 층위에서 부분적으로 지지됨을 보여주는 사례로, 본 연구의 분류 체계에서 단장 후보로 잠정 분류될 수 있다.

47) 03B의 1차부록 구간(03f): 3-20. 子曰 關雎 樂而不淫... 3-21. 哀公問社於宰我... 3-22. 子曰 管仲之器小哉...

03B의 2차부록 구간(03g): 3-23. 子語魯大師樂曰 樂其可知也... 3-24. 儀封人請見曰... 3-25. 子謂韶 盡美矣... 3-26. 子曰 居上不寬...

48) LY0319: <팔일> 3-19. 定公問 君使臣 臣事君 如之何 孔子對曰 君使臣以禮 臣事君以忠

49) 3-14. 子曰 周監於二代... 3-15. 子入太廟 每事問... 3-16. 子曰 射不主皮... 3-17. 子貢欲去告朔之餼羊... 3-18. 子曰 事君盡禮 人以爲諂也

또 다른 사례로 GuwenBERT의 Cluster 2(仁·禮의 본질 논의)는 기무라가 〈이인〉 제4편과 〈팔일〉 제3편에서 각각 별도의 장군으로 분류한 장들을 편 경계를 넘어 하나의 군집으로 묶는다. 이는 의미 층위에서는 통합적 군집이 형성되지만 어휘 층위에서는 분리되어 있는 사례로, 기무라가 편 내부 구조 분석에 머무름으로써 포착하지 못한 편 경계 횡단적 연관을 본 연구가 새롭게 식별한 지점이다.

두 계량 분석이 모두 기무라의 분류와 어긋나는 인문학적 재검토 구간은, 정성적 판단과 계량적 측정이 포착하는 텍스트 구조의 차원이 서로 다름을 시사하는 지점이며, 비판적 정독을 통한 추가 검토가 요구된다. 대표적으로 GuwenBERT의 Cluster 3(경구·단정)은 기무라의 장군 분류와 직접 대응하는 단위를 찾기 어렵다. 이 구간은 향후 비판적 정독을 통해 기무라의 분류를 유지할지, 새로운 기준을 추가할지를 판정해야 할 지점으로 남는다.⁵⁰⁾ 예컨대, 이 군집은 “君子不器”(LY0212), “朝聞道, 夕死可矣”(LY0408) 등 공자의 단호한 경구로 묶인 군집인데, 기무라의 장군 분류 기준이 의미 연관에 있다면 이 군집은 문체적 특성이라는 다른 기준으로 장들을 묶은 셈이다. 이는 기무라 이론의 반증이 아니라, 기무라가 채택한 의미 연관 기준 외에도 문체 차원의 군집 기준이 『논어』 텍스트에 잠재해 있음을 보여주는 사례다.

이러한 문체 차원의 군집 기준은 분석 범위를 20편 전체로 확장할 때 한층 뚜렷해질 것으로 예상된다. 대표적인 경우가 〈향당〉 제10편이다. 이 편은 공자의 말(語)을 문답 형식으로 전하는 『논어』의 일반적 서술 방식과 달리, 공자의 용모·의복·식사·행동거지를 객관적으로 묘사

50) 여기서 새로운 기준이란 기무라의 의미 연관 외에 다른 연구자들의 성과를 토대로 도입될 수 있는 판단 기준으로, 최술의 문헌 신뢰도 평가, 역대 주석가들이 착간·연문으로 지목한 장에 대한 검토, 경구형·단정형 등 발화 방식의 유형화가 그 후보에 해당한다.

하는 문장들로 이루어져 있어, 전통적으로도 여타 편과 그 성격이 구별되는 것으로 인식되어 왔다. 이와 유사하게 공자의 행동을 묘사하는 구절들은 <술이> 제7편의 “子之燕居 申申如也 夭夭如也”(LY0704)처럼 여러 편에 산재해 있는데, 의미 연관이 아니라 서술 양식을 공유하는 이 장들은, 기무라의 장군 분류로는 포착되지 않지만 임베딩 기반 군집에서는 하나의 문체적 단위로 결집될 가능성이 높다. 이러한 문체 군집에 대해서도 의미 기반 군집과 어떻게 위계적으로 결합할 것인지 20편 전체를 대상으로 한 후속 연구에서 별도로 다루어야 할 과제이다.

이상의 세 구간 체계를 통해 본 연구는 기무라의 장군 이론에 대해 단순한 채택이나 각각의 판정이 아니라, 층위별 변별력에 따른 차등적 타당성 판정을 수행하였다. 정합 구간은 분류의 확정적 단위로, 부분 일치 구간은 잠정적 단위 또는 편 경계 횡단적 재편성의 후보로, 인문학적 재검토 구간은 정독을 통한 추가 판정 대상으로 각각 식별된다. 이 차등적 판정 체계는 기무라의 분류를 출발점으로 삼되 거기에 머무르지 않고, 계량적 외부 준거를 통해 검증·조정·확장하는 분류 방법의 기반을 마련하기 위함이다.

5. 결론 및 후속 연구

본 연구는 기무라 에이이치의 장군 이론을 군집 분석을 통해 계량적으로 검증하는 것을 목적으로 하였다. 기무라는 『논어』 각 편 내에서 의미·형태적으로 연관된 장들을 군집 단위로 묶고 그 형식을 유형화함으로써 『논어』 문헌 연구를 구조적 재편성의 단계로 심화시켰으나, 그의 분류 근거는 산문 서술 형식으로 기술되어 있어 정형화된 데이터로 구조

화하기 어렵고, 그 정성적 판단이 계량 분석과 얼마나 부합하는지의 타당성은 지금껏 검증된 바 없다.

이에 본 연구는 두 단계의 작업을 수행하였다. 먼저 기무라의 분류를 대분류·소분류의 두 층위로 정형화하여 『논어』 제1~5편의 분류 모델을 수립하였고, 이어 고대 한문 형태소 분석기 Jiayan을 활용한 TF-IDF 분석과 고대 한문 언어모델 GuwenBERT 임베딩 분석으로 어휘 층위와 의미 층위의 유사성을 각각 측정하여 기무라의 분류와 대조하였다. 분석 결과는 단순한 일치·불일치의 이분 구조가 아니라 기무라의 판단이 두 계량 분석 결과와 모두 일치하는 정합 구간, 한 측면에서만 부합하는 부분 일치 구간, 두 결과 모두 분기되어 비판적 정독이 요구되는 인문학적 재검토 구간 이렇게 세 가지 구간으로 식별되었다. 이러한 차등적 타당성 판정을 통해 본 연구는 기무라의 장군 이론을 채택하거나 기각하는 이분법적 평가를 넘어, 그의 분류를 검증·조정·확장할 수 있는 방법론적 기반을 마련하였다.

아울러, 본 연구는 1~5편을 대상으로 한 파일럿 연구로서, 향후 다음과 같은 방향으로 확장될 것이다.

첫째, 분석 범위의 확장이다. 본 연구에서 확립된 방법론은 이후 다섯 편 단위로 단계적으로 적용되어 『논어』 20편 전체로 확장될 것임을 앞에서 밝힌 바 있다. 그러나 다섯 편씩 단계적으로 적용할 경우, 구간마다 IDF가 독립적으로 산출되므로 동일한 어휘의 변별 비중이 구간에 따라 달라질 수 있다는 한계는 남는다. 이에 후속 연구에서는 기술적 보정보다 각 구간의 군집 결과를 연구자가 직접 검토하는 ‘인문학적 재검토 구간’을 별도로 분류하고 정성적으로 재검토하는 절차를 병행하고자 한다. 이 방식으로 본 연구에서 확립된 방법론을 20편 전체로 단계적으로 확장하여, 공자 사상을 가장 잘 드러내는 최적의 편장 체계 선정을 위한

객관적 지표를 마련하는 것이 향후의 목표이다. 분석 범위가 확장되면 선행 연구로 서술한 최술·나유의·이토 진사이 등 선행 경학사 논의와 계량 분석 결과가 어떤 점에서 상응하는지에 대한 체계적인 비교도 가능해질 것이다.

둘째, 분석 결과의 데이터 구조화이다. 본 연구가 산출한 타당성 정보, 즉 각 장의 군집 배치, 기무라 분류와의 정합도, 세 구간 분류는 TEI/XML⁵¹⁾ 문서로 구조화될 것이다. 이를 통해 산문으로만 기술되어 데이터로 활용되기 어려웠던 기무라의 분류는 물론, 선행 연구에서 디지털화한 최술의 문헌 비평 정보까지 하나의 통합된 디지털 비판 정보 안에 결합될 수 있다. 다시 말해, 현행 20편 원본은 그대로 유지하면서도 각 장에 대해 최술의 신뢰도 평가, 기무라의 장군 분류, 본 연구의 군집 분석 결과가 다층적인 주석으로 부착된 형태의 정본이 구축되는 것이다.

셋째, 공유와 협업 기반의 마련이다. 이렇게 구축된 디지털 비판 정본은 GitHub 등 오픈 플랫폼에 공개하여, 후속 연구자가 동일한 작업을 처음부터 반복하지 않고도 기존 성과 위에서 자신의 연구를 이어 갈 수 있는 토대로 기능하게 할 것이다. 텍스트와 주석 데이터를 누구나 내려받고, 비판하고, 확장할 수 있는 형태로 제공함으로써, 디지털 비판 정본은 한 연구자의 성과물이 아니라 학술 공동체의 공동 자산으로 기능하

51) XML(Extensible Markup Language)은 W3C가 표준화한 마크업 언어로, 데이터에 일정한 구조와 의미를 부여하는 태그 체계를 통해 인간의 가독성과 기계의 처리 가능성을 동시에 확보한다. 다만 XML은 마크업의 문법적 틀을 제공할 뿐 무엇을 어떻게 표시할지는 규정하지 않는데, 이 틀 위에서 인문학 텍스트의 표현 방식을 표준화한 국제 규약이 TEI(Text Encoding Initiative)이다. TEI는 1987년 출범한 국제 협력체이자 그 결과물인 가이드라인을 가리키며, 원문·이체자·교감·주석·인물 및 지명 정보 등 고전 문헌 연구에 필요한 요소들을 표준화된 태그로 기술하는 방법을 규정한다. 본 연구가 산출한 타당성 정보를 TEI 가이드라인에 따라 부호화하면, 특정 도구나 개인에 종속되지 않고 연구자 사이에서 공유·재사용·상호검증이 가능한 형식으로 정본 데이터를 축적할 수 있다. 경학자료를 대상으로 한 TEI를 준수하는 XML 데이터 설계 연구는 김서운(2024) 논문을 참조.

게 된다.

또한 계량적인 분석을 위해 기무라의 이론을 구조화하는 작업은 본 연구의 방법론이 『논어』를 넘어 『맹자』·『대학』·『중용』 등 다른 경서 연구에도 적용 가능한 모델이 될 수 있음을 시사한다. 정성적 비평을 정형 데이터로 명시화하고, 계량 분석으로 그 타당성을 교차 검증하며, 결과를 표준 형식으로 구조화하여 공유하는 일련의 절차는 텍스트를 다루는 여러 학문 전반에 적용될 수 있기 때문이다. 동일한 절차를 다른 경서에 적용함으로써 동아시아 경학 연구의 디지털 인프라가 단계적으로 축적될 수 있다.

요컨대 본 연구가 지향하는 바는 단지 기무라 이론의 검증이나 『논어』 한 편의 재편성에 그치지 않는다. 정성적 비평의 정형화, 계량적 교차 검증, 표준 형식의 데이터 구조화, 그리고 오픈 플랫폼을 통한 공유라는 일련의 순환 구조를 통해, 개별 연구자의 정성적 통찰이 학술 공동체의 누적 가능한 자산으로 전환되는 길을 모색하는 것, 이것이 본 연구가 제안하는 디지털 경학의 방법론적 틀이자, 공유와 협업을 지향하는 디지털 인문학의 연구방법론이다.

참고문헌

1. 원전 자료

- 伊藤仁齋(1712), 『論語古義』, 奎文館, 1712.
荻生徂徠(1737), 『論語徵』, 大和屋孫兵衛, 1737.
太宰春臺(1745), 『論語古訓外傳』, 嵩山房, 1745.
朱熹(1779), 『論語集注』, 乾隆御覽四庫全書薈要本, 1779.
朱彝尊, 『經義考』, 臺灣商務印書館, 1983.

2. 연구 논저

- 단행본

- 武內義雄(1939), 『論語の研究』, 岩波書店.
木村英一(1971), 『孔子と論語』(東洋學叢書), 創文社.
錢穆(1990), 『四書釋義』, 學生書局.
김용옥(2000), 『논어한글역주1』, 통나무.
김용옥(2008), 『논어한글역주3』, 통나무.
崔述, 박준원 역(2013), 『論語餘說』, 지식을만드는지식.
기무라 에이이치, 나종석 역(2020), 『공자와 논어(孔子と論語)』, 예코리브르.

- 한국어

- 김서윤(2024), 「한국경학자료시스템 XML 검토 및 데이터 설계」, 『대동문화연구』 제128집, 성균관대학교 대동문화연구원.
김서윤(2025), 「崔述의 『論語餘說』과 디지털 비판 정본-청대 고증학의 현대적 계승을 위한 시론-」, 『민족문화』 제71집, 한국고전번역원.
김영기(2011), 「공자 사후 공문 후계구도와 《논어》의 편찬-학이편을 중심으로-」, 『중국어문학』 제58집, 영남중국어문학회.
김영기(2012a), 「공자의 지향과 《논어》 편차(篇次)에 관한 소고(小考) -〈학이〉, 〈위정〉, 〈팔일〉을 중심으로-」, 『동아인문학』 제22집, 동아인문학회.
김영기(2012b), 「최술의 『논어』 학이편 제2장 其爲人也孝弟章 考信에 대한 검토」, 『중국어문학』 제60집, 영남중국어문학회.

김윤조(2008), 「論語 學而篇 研究 —논어 편찬 과정에 대한 一 管見—」, 『동양한문학연구』 제27집, 동양한문학회.

박재복(2009), 「출토문헌에 보이는 『論語』 고찰 — 定州 漢墓와 敦煌에서 발견된 『論語』 「述而」편을 중심으로 —」, 『동양고전연구』 제36집, 동양고전학회.

송병렬(2015), 「『論語』 「泰伯」편 構成과 主題에 관한 研究」, 『동방한문학』 제65집, 동방한문학회.

이예성·이강재(2018), 「『논어정의』에 나타난 노론(魯論)」, 『인문논총』 제75권 제2호, 서울대학교 인문학연구원.

- 기타 동양어

佐藤一郎(1958), 「論語 子張篇の研究—論語原典批判 其—」, 『日本中國學會報』, 第10期, 日本中國學會.

3. 웹사이트

한국고전번역원, 『한국고전종합DB』, <https://db.itkc.or.kr>

전통문화연구회, 『동양고전종합DB』, <https://db.cyberseodang.or.kr>

維基文庫 <https://zh.wikisource.org>

CTEXT, <https://ctext.org>

殆知閣, <https://daizhige.org/>

甲言(Jiayan), <https://github.com/jiaeyan/Jiayan>

古文BERT(GuwenBERT), <https://github.com/Ethan-yt/guwenbert>

TEI(Text Encoding Initiative), <https://tei-c.org/>

Extensible Markup Language (XML), <https://www.w3.org/XML/>

A Quantitative Examination of Kimura Eiichi's (木村英一)
Chapter-cluster (章群) Theory:
A Cluster Analysis of Books 1 -5 of the Analects and the Identification
of Humanistic Reexamination Zones

Kim, Seoyun *

This study aims to examine Kimura Eiichi's (木村英一) theory of chapter-clusters (章群) quantitatively and to diagnose its validity. By grouping semantically and formally related chapters within each book of the Analects into clusters and typologizing the modes of their connection, Kimura advanced the philological study of the Analects into the stage of structural reorganization. His theory, however, is set out in extensive prose, which makes it difficult to convert into structured data, and the validity of his qualitative judgments—the degree to which they accord with quantitative analysis—has never been tested.

To address this, the present study carried out two stages of work. First, Kimura's theory was formalized into a two-tier model of primary and secondary categories, on the basis of which a classification model for Books 1-5 of the Analects was established. This step translates Kimura's prose-based judgments into a structured data format. Second, a dual-layer cluster analysis was performed and its results were compared with Kimura's classification: a TF-IDF analysis using the classical Chinese morphological analyzer Jiayan measured formal similarity in lexical patterns, while an embedding-based analysis using the classical Chinese language model GuwenBERT measured contextual similarity in meaning.

* Researcher, Center for Digital Humanities, the Academy of Korean Studies
/ E-mail: sy527991@gmail.com

The results cannot be reduced to a binary of agreement or disagreement; rather, they are identified along three distinct zones. These are the zone of convergence, in which Kimura's judgment aligns with both the morphological and the semantic results; the zone of partial convergence, in which it accords with only one of the two layers; and the zone calling for humanistic reexamination, in which the two quantitative analyses diverge and depart from Kimura's classification, thereby requiring further critical close reading. This three-zone framework shows the position of this study: that quantitative analysis does not replace humanistic judgment but functions as a diagnostic instrument for identifying the points that call for reexamination. As a pilot study of Books 1-5, the methodology established here can be extended to all twenty books of the Analects in future work.

Key Words

Analects, chapter-cluster theory, Kimura Eiichi, cluster analysis, Jiayan, GuwenBERT

논문접수일: 2026.05.11. 심사완료일: 2026.06.07. 게재확정일: 2026.06.24.
--