中國語文學 第83輯(2020. 4) 295-344쪽

코퍼스 분석 도구를 이용한 중국어 교재 텍스트의 어휘와 정형 표현 분석*

강 병 규**

<目 次>

Ⅰ. 서 론

본고는 코퍼스 언어학적 관점에서 한국에서 출판된 중국어 교재 텍스트의 언어적 특징을 탐색하는 것을 목표로 한다. 소위 '대외한어교재(对外汉语教材)'로 불리는 중국어 교재는 어떤 특징이 있는가? 교재별로 차이가 있다면 구체적으로 어떠한 양상을 보이는가? 또한 중국에서 출판된 교재에 비해 한국에서 출판된 교재 텍스트가 가지는 언어적 특징은 무엇인가? 본고는 이러한 주제를 중심으로 논의를 진행해 가고자 한다.

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A5A2A03050117)
** 서강대학교 중국문화전공 교수

한국에서 출판된 중국어 교재는 다양한 유형이 존재하며 저마다의 특징이 있다. 이미경(2018:228)에 따르면 1960년대 이후로 출판된 중국어 학습 관련 서적은 3,896권에 달한다.[1] 이러한 교재들은 중국과 유사한 면도 있지만 한국적 상황을 반영한 참신한 형식과 내용을 담고 있기도 하다.

본 연구는 김나래 외(2016), 이미경(2018)의 연구와 연관된다. 김나래 외(2016)와 이미경(2018)은 중국과 한국에서 출판된 중국어 교재 목록을 유형별, 시기별로 정리하였다. 본 연구는 이에 대한 후속 연구로서 교재 텍스트 안에 반영된 언어적 요소가 어떠한지를 탐색하고자 한다. 전자가 중국어 교재에 대한 목록학적인 측면에 중점이 있었다면 후자는 교재의 언어적 특징을 분석하는데 초점이 있다.

그동안의 중국어 교재 연구는 주로 중등학교 교과서와 같은 규범적인 유형에 집중되어 왔다. 황인옥(2001), 김종호(2004), 강선주(2012), 배재석·김강림(2013), 배은영(2015), 김난미(2016), 임재민(2019) 등의 연구자들은 중고등학교 중국어 교과서를 대상으로 어휘, 문형, 문화요소 등에 대한 다양한 분석을 시도하였다. 그러나 이에 비해 대학이나 일반인들을 대상으로 하는 '대외한어교재'에 대한 관심은 상대적으로 부족했다. 물론 소수이기는 하지만 류다리(2006), 손정애(2012), 김현철 외(2018), 윤상희(2017), 김나래(2018) 등에서 중국어 교재를 대상으로 한 연구가 있기는 하다. 그러나 이러한 연구는 대개 교재에서 특정 어휘나 문법 항목을 분석한 것이다. 중국어 교재 전체에 대한 분석이 진행된 것은 거의 없다. 중국에서도 邱爽(2014), 金佳垠(2017), 郑贤淑(2018), 王淑珍(2019) 등에서도 대부분 두세 권의 교재를 대상으로 연구했을 뿐이다. 중국어 교재에 대한 전반적인 분석을 위해서는 교재의 수량을 어느 정도 확보하여 연구하는 것이 필요한데 기존의 연구에서는 그런 시도가 거의 이루어지지 않았다고 해도 과언이 아니다. 이에 본고에서는 중국어 교재의 언어적 특징을 관찰

1) 이미경(2018:228)에 따르면 1960년대 이후 출판된 중국어 교재는 모두 3,896권인데 이중에 학습서가 2,454권, 목적서가 416권, 수험서가 663권, 교과서가 363권이다.

하기 위해 다양한 텍스트를 확보하는 것이 필요하다는 전제하에 국내에서 출판된 중국어 교재와 중국에서 출판된 교재를 찾아 분석하고 그 결과에 대해 논의하고자 한다.

Ⅱ. 연구대상 및 방법

본고에서는 코퍼스 언어학에서 사용되는 계량적 분석 방법을 통해 단어 사용빈도, 어휘 등급, 품사적 특징, 키워드, 정형화된 표현 등을 분석하고 중국어 교재가 가지는 언어적 특징을 살펴보는 방식으로 연구를 진행하고자 한다.

1. 연구대상

본고에서 연구대상으로 삼은 것은 크게 세 가지이다. 첫째는 한국에서 출판된 중국어 교재이다. 둘째는 중국에서 출판된 중국어 교재이다. 셋째는 일상 언어의 사용 양상을 관찰할 수 있는 참조 코퍼스 자료이다. 이 세 종류의 코퍼스를 연구대상으로 하되 가장 중점을 두어 분석한 자료는 한국에서 출판된 중국어 교재이다. 나머지 두 종류의 자료는 한국에서 출판된 중국어 교재와 비교하기 위해 사용되었다.

(1) 한국에서 출판된 중국어 교재 텍스트

한국에서 출판된 중국어 교재 텍스트는 2000년대 이후로 한국 대학이나 중국어 교육기관에서 많이 사용되는 교재 중에서 선별되었다. 본고에서 선별한 교재는 모두 35권이다. 이 교재 중에는 주저자가 한국인인 것이 20권이고, 중국 원서를 한국에서 새롭게 편역한 교재가 15권이다. 본고에서는 전자를 편의상 'A 유형 교재'로 분류하고 후자를 'B 유형 교재'로 분류하기로 한다.

〈표 1〉 한국에서 출판된 중국어 교재 텍스트 표본

A 유형: 주저자가 한국인인 교재		B 유형: 중국 원서를 편역한 교재	
교재명(출판사)	수량	교재명(출판사)	수량
①≪중국어 마스터≫(다락원)	6권	①≪신공략중국어≫(다락원)	6권
②≪베이직 중국어≫(동양북스)	3권	②≪한어구어≫(동양북스)	4권
③≪완전성공중국어≫(시사)	2권	③≪한어구어345구≫(시사)	2권
④≪맛있는중국어≫(JRC)	3권	④≪301구중국어회화≫(다락원)	2권
⑤≪스마트중국어≫(동양북스)	4권	⑤≪한어교정≫(시사)	1권
⑥≪중국어교실≫(넥서스)	2권		
합계	20권	합계	15권

(2) 중국에서 출판된 교재 코퍼스

중국에서 출판된 교재 텍스트는 북경사범대학 중국어정보처리연구소(北京师范大学中文信处理研究所)에서 구축한 코퍼스를 이용하였다. 이 코퍼스는 'CTC(Corpus of Teaching Chinese as a Second Language: 汉语国际教育动态语料库)'라고 불린다. CTC 코퍼스는 "대규모 텍스트 다층구조 지식 표현 및 중국어 텍스트 이해 응용 시스템 연구 개발(海量文本多层次知识表示及中文文本理解应用系统研制)"이라는 863 연구과제의 일환으로 구축된 언어자원이다. 이 코퍼스는 대외한어교재 텍스트를 분석하여 연구와 교육에 활용하기 위해 만들어졌다. 코퍼스 구축에 사용된 교재는 대부분 2000년대 이후에 중국에서 출판된 것들이다. CTC 코퍼스는 모두 263권의 중국어 교재로 구성되어 있다. 이 중에서 대외한어 교재는 197권이고 HSK 교재는 66권이다. 본고에서는 이 코퍼스를 편의상 'C 유형 교재'로 칭하겠다.

C 유형의 CTC 코퍼스는 온라인 자료라서 교재별로 예문을 모으는 것이 힘들다. 대신 예문을 문장 단위로 검색하여 저장할 수는 있다. 본고에서는 웹 크롤링(web crawling) 방법을 사용하여 중복되지 않게 여러 문장을 내려받아 컴퓨터에 저장하였다. 그리고 컴퓨터에 저장된 중국어 문장 중에서 무작위 추출(random sampling) 방식으로 문장을 1,000개 내외로

뽑는 과정을 반복하였고 이 중 15개의 파일을 표본으로 만들었다.[2] 이렇게 추출한 자료(1,000문장×15개 파일)에 대해 순서대로 “CTC1”, “CTC2”, “CTC3”, “CTC4”, ……“CTC15”라는 명칭을 부여하고 분석에 활용하였다.

(3) 일반 참조 코퍼스

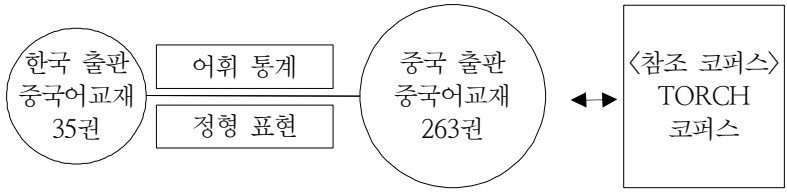
일반 참조 코퍼스로 삼은 것은 TORCH 코퍼스이다. TORCH(Texts Of Recent Chinese) 코퍼스는 170만자의 소규모이지만 다양한 영역의 텍스트 자료가 원본 파일로 제공되기에 정밀한 분석을 하는데 편리하다. 이 코퍼스는 중국 64개 대학의 교수, 대학원생이 미국의 Brown 코퍼스의 형태를 참조하여 공동으로 구축한 언어자원이다. 코퍼스 파일은 신문보도, 사설, 종교, 예술, 전기, 정부문서, 학술서적, 중국어 교재작품(일반소설, 판타지 소설, 추리소설), 희곡 등의 영역으로 나누어져 있다.[3]

2. 연구 방법 및 분석 도구

본고의 연구는 아래 그림과 같은 방식으로 진행되었다. 먼저 한국에서 출판된 중국어 교재(A 유형, B 유형) 텍스트의 어휘와 어휘 다발로 구성된 표현을 분석하였다. 아울러 중국에서 출판된 교재 코퍼스(C 유형)의 어휘적 특징을 분석하였다. 그리고 각 유형의 교재 텍스트의 단어사용빈도, 결합유형, 분포적 특징 등을 비교하였다. 더 나아가 일반 영역의 참조 코퍼스에 비해 중국어 교재 텍스트가 가지는 언어적 특징을 거시적으로 탐색해 보았다.

2) 문장의 수량을 1,000개 단위로 뽑은 이유는 통계 분석과정에서 표본 추출(sampling)할 때 무작위로 1,000개 내외의 데이터를 추출하는 관례가 있기 때문이다. 통계학에서는 일반적으로 어떤 모집단에서 일정한 표본(1,000개 내외)을 적절하게 선택하면 모집단의 특성을 정확하고 효과적으로 보여준다고 알려져 있다.

3) 현재 구축된 것은 TORCH2009, TORCH2014 등이 있다. 자세한 것은 참고문헌에 제시된 홈페이지를 참고하기 바람.



〈그림 1〉 중국어 교재 분석 내용

(1) 중국어 교재 텍스트의 수집과 전처리(preprocessing)

중국어 교재 문장은 다양한 방식으로 수집될 수 있다. 본고는 그 중에서 문서를 스캔한 뒤 문자로 디지털화하는 방법을 사용하였다. 문자인식 프로그램을 활용하여 각 과별로 제시된 기본 문형과 본문(회화, 독해) 문장을 디지털화하고 오탈자를 교정하여 문장 단위로 저장하였다.

중국어 텍스트 분석은 단어와 품사 정보를 기초로 진행되었다. 띄어쓰기가 없는 원문을 그대로 사용하면 단순 검색만이 가능하다. 만약 텍스트에 대한 단어·품사 정보가 없으면 연어 분석이나 문형 분석을 기대하기 힘들다. 이에 본고에서는 중국어 교재 텍스트에 대해서 단어와 품사 분석을 별도로 진행하였다. 이 과정에서 사용한 프로그램은 중국어 품사분석기인 'CorpusWordParser'와 'ICTCLAS'이다.[4] 이 프로그램은 파일에 저장된 중국어 문장을 자동으로 읽어 들여 단어 분리와 품사를 부착해주는 기능이 있다. 'CorpusWordParser', 'ICTCLAS' 프로그램으로 35권의 중국어 교재 텍스트 파일에 대해서 자동 태깅을 한 다음 오류가 있는 부분은 수정 작업을 진행하였다.

(2) 중국어 코퍼스 분석 도구

각 교재별로 단어의 사용빈도와 연어, 말덩어리(chunk), 문형 등을 조사하기 위해 사용한 코퍼스 분석도구는 AntConc와 WS(WordSmith Tools)

4) 이 프로그램은 아래의 사이트에서 각각 내려받을 수 있다.
① http://corpus.zhonghuayuwen.org/resources.aspx, ② http://ictclas.nlpir.org

프로그램이다. 이 프로그램은 코퍼스 분석 도구로서 용례 검색, 어휘 통계, 정형 표현(cluster)·연어·키워드 추출 기능이 있다. 이외에 본고에서는 Excel, Tableau, SPSS를 활용하여 시각화 분석과 통계 분석을 진행하였다.

이러한 분석 과정은 다음과 같이 요약할 수 있다.

〈표 2〉 중국어 교재 텍스트의 분석 방법

중국어 교재 본문 자료 입력	OCR 프로그램
↓	
중국어 단어 분석 및 품사 태깅	단어분석 프로그램
↓	
교재 텍스트별 단어, 키워드, 정형 표현 추출	AntConc, WS
↓	
어휘와 정형 표현에 대한 통계 분석	Excel, SPSS

Ⅲ. 중국어 교재 텍스트에 대한 기초 통계

이 장에서는 코퍼스 분석 결과를 바탕으로 중국어 교재에 나타난 기본적인 특징을 살펴보고자 한다. 일반적으로 코퍼스 언어학에서는 언어 요소가 어느 정도의 사용빈도를 보이는지에 관심을 가진다. 이를 바탕으로 텍스트에서 높은 빈도를 보이는 단어와 문장의 경향성을 파악하는 것이다.

중국어 교재 텍스트 분석 과정에서는 단어의 수량과 문장의 수량을 기초로 전반적인 특징을 파악할 수 있다. 전문기·임인재(2009), 권지혜·이석재(2019) 등에서 사용한 코메트릭스(Coh-Metrix)의 분석 방식에 따르면 교재 텍스트의 수준과 난이도는 단어수, 문장수, 평균 문장길이를 통해서 어느 정도 가늠해 볼 수 있다.[5)]

5) 코메트릭스(Coh - Metrix, 3.0)는 미국 멤피스 대학교 지능형 시스템 연구소에서 개발한 언어 분석 프로그램이다. 웹기반 영어 텍스트 분석 도구로 어휘 다

단어빈도수는 특정 단어가 몇 번 사용되는지를 측정하여 나타낸 수치이다. 반복되는 단어가 많을수록 난이도가 낮은 문장이라 할 수 있으며 더 쉽게 읽히는 문장이라 할 수 있다. 단어 유형은 한 텍스트에서 얼마나 다양한 단어들이 사용되는지를 나타내준다. 문장의 수량과 길이도 교재 텍스트 분석 과정에서 난이도를 측정하는 지표가 될 수 있다. 코메트릭스 평가 기준에 따르면 일반적으로 어려운 텍스트일수록 문장의 수량이 많고 한 문장이 여러 단어로 구성되어 길어지는 경향이 있다.

본고는 이러한 분석 방법에 따라 중국어 교재 텍스트에서 사용된 단어와 문장의 사용빈도를 조사하였다. 교재 텍스트는 크게 3가지 유형으로 나누어 분석하였다. (1) A 유형 교재(한국에서 출판되고 한국인이 주저자인 중국어 교재), (2) B 유형 교재(중국 원서를 한국에서 편역하여 출판한 교재), (3) C 유형 교재(중국에서 출판된 중국어 교재)로 나누어 조사하였다. 아래의 표는 세 종류의 중국어 교재에서 사용된 단어수, 단어 유형, 단어 길이, 문장수, 문장 길이를 나타낸다.

〈표 3〉 교재별 단어 및 문장 수량 및 문장 평균 길이

		A 유형 교재	B 유형 교재	C 유형 교재
단어 층위	단어수(token)	69,399	124,201	225,797
	단어유형(type)	4,817	7,873	19,478
	평균 단어 길이	1.48	1.47	1.57
문장 층위	문장수	9,421	14,228	15,455
	평균 문장 길이	7.37	8.73	14.61

위의 분석 결과 단어의 수량은 'A 유형 〈 B 유형 〈 C 유형'의 순서를 가진다. 이를 통해 한국인 주저자가 집필한 중국어 교재가 상대적으로 단

양성부터 응집성과 정합성까지 다양한 요소들을 분석할 수 있다. 그 중에서 '기초산출치'라는 항목이 있는데 이는 단어 수, 문장 길이 등을 보여준다. 측정치가 높게 나타날수록 텍스트의 난이도가 높다고 볼 수 있다.

어 사용량이 적고 난이도가 낮을 것으로 예상할 수 있다. 이에 비해 중국인 저자가 쓴 교재일수록 단어량이나 유형이 많아진다.

중국어 교재에 나타난 단어의 음절수는 평균 1.5개 내외로서 교재별로 큰 차이가 없었다. 조사 결과에 따르면 A 유형 교재는 평균 음절수가 1.48이다. B 유형 교재는 1.47음절이고 C 유형 교재는 1.57 음절로 조사되었다. 아래의 표는 각 교재별로 사용된 단어의 음절수를 보여준다.

〈표 4〉 교재별 단어의 평균 음절수

	A유형 교재		B유형 교재		C유형 교재	
1음절	39,895	57.5%	71,706	57.7%	10,4931	49.1%
2음절	26,235	37.8%	47,292	38.1%	98,039	45.8%
3음절	2,808	4.0%	4,309	3.5%	8,119	3.8%
4음절	447	0.6%	839	0.7%	2,510	1.2%
5음절 이상	14	0.0%	55	0.0%	228	0.1%
평균음절수	1.48		1.47		1.57	

그러나 문장의 길이는 단어와는 달리 교재 유형에 따라 큰 차이가 난다. 문장의 길이는 한 문장 안에 사용된 단어를 기준으로 계산할 수 있다. 한 문장이 평균적으로 몇 개의 단어로 구성되었는지를 조사한 결과에 따르면 A 유형 교재는 평균적으로 7.37개의 단어로 구성된다. B 유형 교재는 문장의 평균 길이가 8.73이다. 주목할 것은 중국에서 출판된 교재인 C 유형의 텍스트는 문장의 평균 길이가 14.61로서 확연한 차이가 있다는 점이다. 이는 중국에서 출판된 교재가 상대적으로 긴 문장으로 구성되어 있음을 의미한다. 아래의 예를 보자.

(1) 你们 一天 工作 多长 时间？ 一天 工作 七 个 小时。 (문장길이: 5)
≪중국어마스터 2≫

(2) 你们 在 中国 才 学习 了 一 年, 汉语 水平 就 提高 得 这么 快, 主要是 因为 你们 学习 都 很 努力。 (문장길이: 24) ≪新实用汉语课本 2≫

(1)에서 A 유형 교재는 초중급 수준에서 비교적 짧은 단문 중심의 문장으로 본문을 구성한 예이다. (2)는 중국에서 출판된 교재인데 문장의 길이가 상당히 긴 예이다.

그렇다면 수준별로 나누어 보았을 때 단어의 수량과 문장의 길이는 어떠한 상관성이 있을까? 이를 알아보기 위해 본고에서는 한국에서 출판된 A 유형 교재와 B 유형 교재를 초급과 중고급으로 나누어 관찰해 보았다. 수준을 나눌 때는 편의상 시리즈 교재 중에 입문과정이나 시리즈1·2에 해당하는 것을 초급으로 분류하였고 시리즈 3 이상의 교재를 중고급 교재로 분류하였다.

〈표 5〉 교재의 수준별 분류

A 유형 교재 텍스트		B 유형 교재 텍스트	
초급 교재	베이직중국어 1·2 성공중국어 1·2 JRC맛있는중국어 1·2 중국어마스터STEP 1·2 넥서스중국어 1·2 스마트중국어 1·2	초급 교재	301구 중국어회화 1·2 한어구어345구 1·2 한어구어 1·2 신공략중국어 기초·초급
중고급 교재	베이직중국어3 JRC맛있는중국어3 중국어마스터STEP3~6 스마트중국어3·4	중고급 교재	한어구어5·6 신공략중국어프리토킹 신공략중국어실력향상 신공략중국어고급

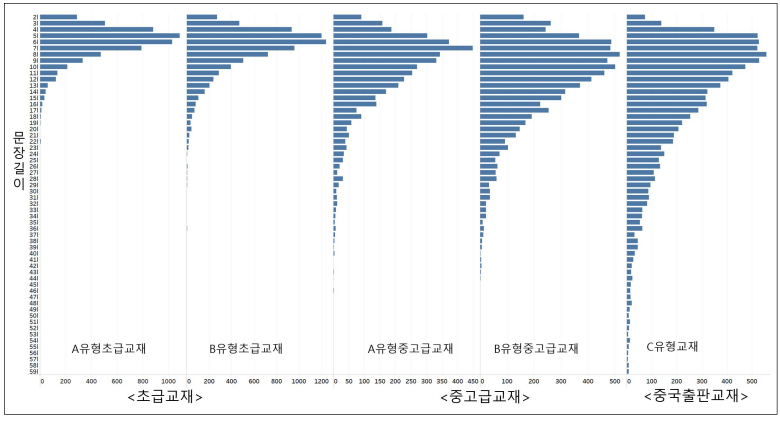
조사 결과 초급교재는 A 유형 텍스트가 평균 6.19 단어로 한 문장이 구성되고 B 유형 텍스트는 평균 7.55 단어로 한 문장이 구성된다. 반면에 중고급 교재의 경우에는 문장의 평균 길이가 증가한다. A 유형 중고급 교재는 한 문장당 10.89개의 단어가 사용된다. B 유형 중고급 교재는 한 문장당 12.23개의 단어가 사용된다. 중고급 교재의 표준편차도 초급교재보다 더 크다. 이는 중고급 수준으로 갈수록 단어의 양이 많고 복잡한 수식어를 사용하거나 복문을 많이 사용하는 것을 알 수 있다. 중국에서 출판된 C

유형 교재 텍스트는 문장의 평균 길이가 더 증가하여 14.61이고 표준편차도 11.45나 된다.

〈표 6〉 교재별 문장의 평균 길이

	A 유형 교재		B 유형 교재		C 유형 교재
	초급	중고급	초급	중고급	종합
평균문장길이	6.19	10.89	7.55	12.23	14.61
표준편차	2.77	6.81	4.71	7.11	11.45
가장 짧은 문장	1단어	1단어	1단어	1단어	1단어
가장 긴 문장	22단어	73단어	42단어	49단어	98단어
10단어 이상 사용비율	10.9%	47.7%	22.1%	58.5%	60.0%
20단어 이상 사용비율	0.2%	9.7%	2.2%	14.3%	24.9%

분석 결과에 따르면 초급 교재는 한 문장이 10단어 이내로 구성되는 것이 다수를 차지한다. 이에 비해 중고급 교재로 갈수록 10단어 이상으로 구성되는 문장의 수량이 많다. 교재의 수준이 높을수록 한 문장을 구성하는 단어의 수량이 증가한다. 아래의 그래프는 교재 수준별로 문장을 구성하는 단어의 수량에 어떠한 차이가 있는지를 보여준다.



〈그림 2〉 교재별 문장의 길이 비율 그래프

위의 그래프를 보면 초급 수준의 교재에서는 10 단어 이내로 이루어진 문장이 다수를 차지한다. A 유형 초급 교재는 10 단어 이내로 구성되는 문장의 비율이 89%에 달한다. B 유형 초급 교재도 78%의 문장이 10 단어 이내로 구성된다. 이에 비해 중고급 교재에서는 이러한 비율이 감소한다. A 유형 중고급 교재는 10 단어 이내로 구성되는 비율이 대략 53% 정도이다. B 유형 중고급 교재는 그 비율이 42% 내외이다. C 유형 교재의 경우에는 그 비율이 더 감소하여 40%에 불과하다. 나머지 60%의 문장은 모두 10 단어 이상으로 구성된다. 20 단어 이상으로 구성되는 문장의 비율도 24.9%에 달한다. 이를 통해 우리는 교재의 수준에 따라 문장을 구성하는 단어의 수량에 차이가 있음을 알 수 있다. 즉 교재의 수준이 높아지면 단어의 수량도 증가하며 문장도 길어지고 복잡해진다.

Ⅳ. 어휘 사용의 다양성 분석

어휘의 다양성을 나타내주는 기본적인 지표로는 'TTR(type-token ratio)'을 들 수 있다. 이 수치는 교재 텍스트에서 단어의 출현빈도(token frequency)와 유형빈도(type frequency)의 비율을 백분율로 나타낸 수치이다. 이 수치는 어휘의 다양성과 난이도를 측정하는 지표로 널리 사용되고 있다.[6] TTR 값이 높다는 것은 교재 텍스트에 다양한 단어가 사용된 것을 의미하므로 어휘의 다양성이 높다고 해석된다. 반면에 TTR 값이 낮다는 것은 사용된 단어가 상대적으로 적어 어휘의 다양성이 낮다고 할 수 있다. 다른 측면에서 보자면 TTR 수치가 높으면 다양한 단어가 사용되어 어휘 학습의 난이도가 높은 것을 의미한다. 반대로 TTR 수치가 낮으면 같은 단어가 여러 번 반복되어 사용된 것을 의미하므로 어휘 학습의 난이도가 상대적으로 낮아진다.

그러나 TTR 수치의 단점은 텍스트 크기에 따라 많은 영향을 받는다는

6) 이규형·김하응·이용훈(2015:4-5) 참조.

점이다. 텍스트의 크기가 비슷한 경우에는 TTR 값으로 상호 비교를 해도 괜찮지만 크기가 다르면 동일한 기준으로 평가하기 어렵다. 이러한 단점을 보완하기 위해서는 크기를 표준화해서 TTR 수치를 계산하여 비교하는 것이 필요하다. 표준화된 타입-토큰 비율은 일반적으로 'STTR(Standarized TTR)'이라고 불린다. 이것은 한 텍스트를 1,000 단어 단위로 쪼개서 TTR 값을 계산한 다음 평균을 낸 것이다. 이렇게 하면 문서의 크기에 상관없이 평균 1,000 단어 중에서 얼마나 다양한 유형이 존재하는지 알 수 있어 객관적인 비교가 가능하다.

본고는 중국어 교재별로 TTR 수치와 STTR 수치를 조사하였다. 아래의 표는 교재별로 사용된 단어의 출현빈도, 유형빈도, TTR, STTR 수치를 정리한 것이다.

〈표 7〉 A 유형 중국어 교재 텍스트의 어휘 다양성

교재 텍스트	Tokens	Types	TTR	STTR
베이직중국어1	1910	313	16.39	19.00
베이직중국어2	2826	527	18.65	25.40
베이직중국어3	3632	725	19.96	31.33
성공중국어1	2211	293	13.25	21.05
성공중국어2	4292	532	12.40	27.65
JRC맛있는중국어1	2033	338	16.63	21.85
JRC맛있는중국어2	3773	564	14.95	25.43
JRC맛있는중국어3	5489	805	14.67	29.12
중국어마스터STEP1	1921	354	18.43	19.50
중국어마스터STEP2	3587	708	19.74	31.37
중국어마스터STEP3	4404	911	20.69	34.85
중국어마스터STEP4	5707	1091	19.12	36.16
중국어마스터STEP5	6341	1367	21.56	39.68
중국어마스터STEP6	5867	1522	25.94	44.84
넥서스중국어1	1445	289	20.00	20.30

교재 텍스트	Tokens	Types	TTR	STTR
넥서스중국어2	2833	657	23.19	32.60
스마트중국어1	1226	226	18.43	21.20
스마트중국어2	2120	434	20.47	27.05
스마트중국어3	2936	675	22.99	36.40
스마트중국어4	4846	1069	22.06	38.03

〈표 8〉 B 유형 중국어 교재 텍스트의 어휘 다양성

교재 텍스트	Tokens	Types	TTR	STTR
301구 중국어회화1	5116	567	11.08	24.44
301구 중국어회화2	6196	1055	17.03	34.77
한어교정1	6104	741	12.14	26.22
한어구어345구1	3107	387	12.46	20.50
한어구어345구2	6989	932	13.34	29.57
한어구어1	1590	294	18.49	19.50
한어구어2	5182	791	15.26	30.22
한어구어5	8504	1537	18.07	36.46
한어구어6	11545	2200	19.06	40.32
신공략중국어기초	5350	679	12.69	27.96
신공략중국어초급	10374	1202	11.59	33.75
신공략중국어프리토킹	17719	3023	17.06	42.00
신공략중국어실력향상1	10491	1522	14.51	36.70
신공략중국어실력향상2	7773	1586	20.40	40.69
신공략중국어고급	18161	3744	20.62	47.31

〈표 9〉 C 유형 중국어 교재 텍스트의 어휘 다양성

교재 텍스트	Tokens	Types	TTR	STTR
CTC 텍스트1	15124	4718	31.20	62.29
CTC 텍스트2	16350	4835	29.57	61.01

교재 텍스트	Tokens	Types	TTR	STTR
CTC 텍스트3	15183	4578	30.15	61.39
CTC 텍스트4	15917	4690	29.47	61.39
CTC 텍스트5	15445	4716	30.53	61.41
CTC 텍스트6	19173	5310	27.70	61.64
CTC 텍스트7	15348	4718	30.74	61.39
CTC 텍스트8	16884	5004	29.64	61.34
CTC 텍스트9	18323	5294	28.89	60.96
CTC 텍스트10	16542	4994	30.19	61.44
CTC 텍스트11	8559	2058	24.04	48.85
CTC 텍스트12	8992	2301	25.59	50.41
CTC 텍스트13	9000	2325	25.83	50.26
CTC 텍스트14	9625	2752	28.59	55.34
CTC 텍스트15	11764	3110	26.44	55.13

위의 표에 근거할 때 중국어 교재별로 어휘 사용량과 TTR 수치는 다양한 양상을 보인다. 예를 들어 한국에서 출판된 중국어 교재(A 유형)의 경우, ≪베이직중국어1≫은 단어 출현빈도가 1,910회, 유형빈도가 313회, STTR 수치가 19.00으로서 어휘 사용량이 적은 편이다. ≪다락원중국어마스터STEP1≫도 출현빈도(1,921), 유형빈도(354), STTR(19.50)의 수치를 통해 봤을 때 어휘 사용량이 적다. 이 두 교재는 300개~400개 정도의 단어로 전체 본문이 구성되었으며 어휘 다양성이 낮아 초급 교재에 해당한다고 할 수 있다. 이에 비해 ≪다락원중국어마스터STEP6≫는 1,522개의 단어로 전체 본문이 구성되었고 STTR 수치도 44.84에 해당하여 A 유형 교재 중에서는 가장 높은 수준에 해당한다.

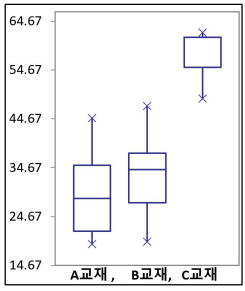
한국에서 출판된 B 유형 교재의 어휘 사용 양상도 교재별로 다양하다. ≪한어구어1≫은 단어 출현빈도(1,590), 유형빈도(294), STTR (19.50) 수치를 통해 볼 때 적은 어휘가 사용된 초급 수준의 교재임을 알 수 있다. 이에 비해 ≪신공략중국어고급≫, ≪신공략중국어프리토킹≫ 교재는 사용

된 단어가 3,000개 이상이고 STTR 수치도 47.31과 42.00으로서 어휘 다양성이 높은 수준에 해당한다.

한편, 중국에서 출판된 중국어 교재를 종합한 C 유형의 교재는 어휘 다양성이 더 높은 편이다. 본고에서 1,000문장 단위로 표본을 추출하여 조사한 15개의 텍스트를 보면 평균적으로 2,000~5,000 단어가 사용되었으며 STTR 수치도 50~60에 달한다.[7] 이는 한국에서 출판된 중국어 교재 표본보다는 높은 수치이다.

아래의 표는 세 종류의 교재를 비교하기 위해 어휘 다양성 수치인 STTR을 기준으로 평균값, 중간값, 표준편차 등을 정리한 것이다.

〈표 10〉 중국어 교재별 어휘 STTR



A 유형 교재		B 유형 교재		C 유형 교재	
평균	29.14	평균	32.94	평균	58.28
중앙값	28.39	중앙값	34.26	중앙값	61.34
최대값	44.84	최대값	47.31	최대값	61.64
최소값	19.00	최소값	19.50	최소값	48.85
표준편차	7.55	표준편차	7.93	표준편차	4.88
F값 : 82.44, 유의수준 : p〈0.001**(유의함)					

위의 표를 종합적으로 살펴볼 때 어휘의 다양성의 측면에서 A 유형 교

7) 외국어 교재별로 어떤 STTR 값이 적당한지에 대한 객관적인 기준은 없다. 다만 영어 교재 연구의 사례를 통해 볼 때 어휘수와 STTR 값이 어느 정도 범위에 속하는지는 참고할 수 있다. 물론 중국어 교재와 영어 교재를 단순 비교하는 것을 문제가 있겠지만 참고용으로 검토할만한 가치는 있다고 판단된다. 예를 들어 김재은·최인철(2015), 권지혜·이석재(2019) 등의 연구를 보면 우리나라 고등학교 영어 교과서는 단어 사용량이 대체로 3,000단어 내외이고 STTR 값이 40.00 내외이다. 이를 통해 외국어 교재 편찬과 연구에서 어휘 사용량과 어휘 다양성 등에 관한 기준 등을 고민해 볼 수도 있으리라 생각된다. 그러나 이규형·김하응·이용훈(2015:7-8)에서 지적한 것처럼 어휘량과 STTR 값이 어휘 다양성과 난이도를 결정하는 절대적인 지표는 아니므로 각각의 장단점을 이해하고 사용하는 것이 필요하다.

재가 상대적으로 낮고 B 유형 교재가 그 다음이며 C 유형 교재가 가장 높다고 할 수 있다. 어휘 다양성이 높으면 그만큼 어휘 학습량의 증가하고 난이도가 높아지게 된다.

Ⅴ. 어휘 사용 등급 분석: HSK 어휘표와 비교

본고에서는 교재별 어휘 사용의 수준을 파악하기 위해 HSK 어휘 등급표와 비교해 보았다. HSK 어휘 등급표는 외국인들을 대상으로 하는 중국어 교육과 평가 과정에서 사용되는 어휘 목록이다. ≪国际汉语教学通用课程大纲≫에는 등급별로 어휘의 수량과 목록이 제시되어 있다. 이에 따르면 어휘 목록은 수준에 따라 1급 어휘부터 6급 어휘로 구분된다. 어휘의 수량 면에서는 1급 어휘가 150개, 2급 어휘가 150개, 3급 어휘가 300개, 4급 어휘가 600개, 5급 어휘가 1,300개, 6급 어휘가 2,500개이다.

HSK 어휘 등급은 중국어 시험이나 외국인들의 중국어 수준을 나타내주는 참고 지표로 사용된다.[8] 예를 들어 HSK 3급에 해당하는 외국인은 매주 2~3시간씩 3학기(120~180시간) 정도의 중국어를 학습하고, 600개의 상용어휘를 구사할 수 있다고 규정된다. HSK 4급은 매주 2~4시간씩 4학기(190~400시간)정도의 중국어를 학습하고, 1,200개의 상용어휘를 익숙하게 활용할 수 있어야 한다고 되어 있다.[9] 이러한 기준을 중국어 교재에 적용해 보면 각 교재의 HSK 어휘 사용 비율을 통해 어휘 수준을 어느 정도 파악할 수 있다.

본고는 이러한 가정하에 교재별로 사용된 어휘와 HSK 어휘 목록을 비교하였다. 아래의 표는 교재별로 사용된 어휘가 HSK 어휘 목록의 어느 수준과 대응되는지를 비율로 나타낸 것이다.

먼저 한국인 저자들이 집필한 중국어 교재(A 유형)를 살펴보기로 하자.

8) 이지은·신수영(2015:162), 진현(2019:5) 참조.
9) HSK 한국사무국 홈페이지 참고 http://www.hsk.or.kr/

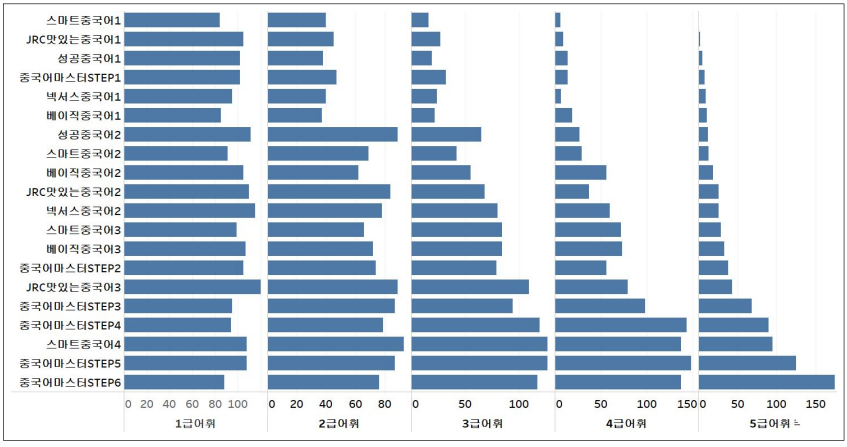
아래의 표는 각 교재에서 1급 어휘 ~ 6급 어휘가 몇 개 사용되었는지 조사한 결과이다.

〈표 11〉 A 유형 교재와 HSK 어휘 등급표 비교

교재 텍스트	1급	2급	3급	4급	5급	6급	기타	합계
베이직중국어1	85	37	22	19	11	7	132	313
베이직중국어2	105	62	55	56	19	7	223	527
베이직중국어3	107	72	84	73	33	12	344	725
성공중국어1	102	38	19	14	5	1	114	293
성공중국어2	111	89	65	27	12	7	221	532
JRC맛있는중국어1	105	45	27	9	2	0	150	338
JRC맛있는중국어2	110	84	68	37	26	5	234	564
JRC맛있는중국어3	120	89	109	79	43	9	356	805
중국어마스터STEP1	102	47	32	14	8	1	150	354
중국어마스터STEP2	105	74	79	56	38	17	339	708
중국어마스터STEP3	95	87	94	98	68	28	441	911
중국어마스터STEP4	94	79	119	143	90	52	514	1091
중국어마스터STEP5	108	87	126	148	125	75	698	1367
중국어마스터STEP6	88	76	117	137	175	84	845	1522
넥서스중국어1	95	40	24	7	9	3	111	289
넥서스중국어2	115	78	80	60	26	15	283	657
스마트중국어1	84	40	16	6	1	3	76	226
스마트중국어2	91	69	42	29	13	8	182	434
스마트중국어3	99	66	84	72	29	22	303	675
스마트중국어4	108	93	126	137	95	28	482	1069

위의 표에서 알 수 있듯이 교재별로 사용되는 어휘 중에 HSK 어휘 등급에 포함되는 비율은 다양한 양상을 보인다. 예를 들어 ≪베이직중국어1≫은 1급~3급 어휘가 144개이고, 4급~5급 어휘가 40개, 6급 어휘가 7개, 기타 어휘가 132개를 차지한다. ≪중국어마스터STEP6≫은 1급~3급 어휘가

281개이고, 4급~5급 어휘가 312개, 6급 어휘가 84개, 기타 어휘가 845개를 차지한다. 전반적으로 보면 초급 수준의 교재일수록 1급~3급 어휘의 비율이 높고 중고급 수준의 교재일수록 4급~5급 수준의 어휘의 비율이 늘어난다.



〈그림 3〉 A 유형 교재별 HSK 등급(1급~5급) 어휘 분포

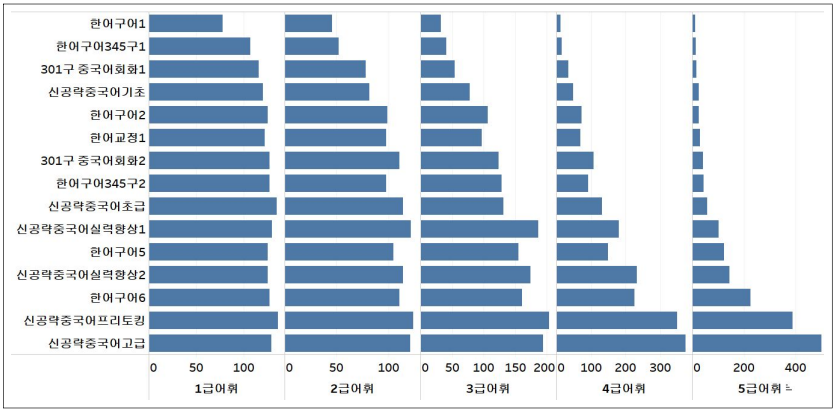
중국에서 출판된 원서를 한국에서 편역한 중국어 교재(B 유형)의 어휘 사용 양상도 다양한 특징을 보인다. 아래의 표는 B 유형 교재에서 HSK 어휘가 얼마나 사용되었는지를 조사한 결과이다.

〈표 12〉 B 유형 교재와 HSK 어휘 등급표 비교

교재 텍스트	1급	2급	3급	4급	5급	6급	기타	합계
301구 중국어회화1	116	78	54	35	16	3	265	567
301구 중국어회화2	127	111	124	107	43	10	533	1055
한어교정1	122	98	97	70	31	7	316	741
한어구어345구1	107	52	41	15	14	2	156	387
한어구어345구2	127	98	128	92	45	4	438	932

교재 텍스트	1급	2급	3급	4급	5급	6급	기타	합계
한어구어1	78	46	32	13	12	1	112	294
한어구어2	125	99	107	72	26	8	354	791
한어구어5	125	105	155	150	124	52	826	1537
한어구어6	127	111	161	225	227	98	1251	2200
신공략중국어기초	120	82	78	48	26	7	318	679
신공략중국어초급	135	114	131	131	58	11	622	1202
신공략중국어프리토킹	136	124	204	348	389	169	1653	3023
신공략중국어실력향상1	130	122	186	181	103	35	765	1522
신공략중국어실력향상2	125	114	174	232	145	38	758	1586
신공략중국어고급	129	121	194	373	501	323	2103	3744

B 유형의 교재에서도 초급 수준일수록 1급~3급 어휘의 비율이 높고 중고급 수준의 교재일수록 4급~5급 어휘의 비율이 늘어난다. 예를 들어 ≪301구로끝내는중국어회화1≫은 1급~3급 어휘가 248개이고, 4급~5급 어휘가 51개, 6급 어휘가 3개, 기타 어휘가 265개를 차지한다. 반면에서 ≪신공략중국어고급≫은 1급~3급 어휘가 444개이고, 4급~5급 어휘가 874개, 6급 어휘가 323개, 기타 어휘가 2,103개를 차지한다.



〈그림 4〉 B 유형 교재별 HSK 등급(1급~5급) 어휘 분포

위의 도표에서 보이듯이 A 유형 교재와 마찬가지로 B 유형의 교재에서 HSK 어휘의 사용 비율은 교재의 수준에 비례하는 경향이 있다. 초급 수준의 교재일수록 1급~3급 어휘의 비율이 높고 중고급 수준의 교재일수록 4급~5급 수준의 어휘의 비율이 늘어난다.

그러나 한가지 주목할 점은 중국어 교재에 사용된 어휘 중에서 HSK 어휘목록에 제시되지 않은 비율도 매우 높다는 것이다. 각 교재의 난이도에 상관없이 1급~6급 어휘에 포함되지 않은 것도 상당히 많다. 이러한 경향은 진현(2019:13)의 분석 결과와도 유사하다. 진현(2019:13)에서 조사한 결과에 따르면 중국어 교재에서 등급외 어휘가 차지하는 비중은 50%를 넘는다. 본고의 조사 결과를 살펴보면 A 유형 교재에서 HSK 어휘표에 없는 단어의 사용비율이 평균적으로 44.1%에 이른다. B 유형 교재는 평균적으로 48.5%의 단어가 HSK 어휘표에 없는 것들이다. 아래의 표는 A 유형 교재, B 유형 교재에서 등급외 비율을 정리한 결과이다.[10]

〈표 13〉 교재별 HSK 어휘 등급 외 사용 비율

A 유형 교재		B 유형 교재	
교재 텍스트	등급 외 비율	교재 텍스트	등급 외 비율
베이직중국어1	42.2%	301구 중국어회화1	46.7%
베이직중국어2	42.3%	301구 중국어회화2	50.5%
베이직중국어3	47.4%	한어교정1	42.6%
성공중국어1	38.9%	한어구어345구1	40.3%
성공중국어2	41.5%	한어구어345구2	47.0%
JRC맛있는중국어1	44.4%	한어구어1	38.1%
JRC맛있는중국어2	41.5%	한어구어2	44.8%
JRC맛있는중국어3	44.2%	한어구어5	53.7%
중국어마스터STEP1	42.4%	한어구어6	56.9%

10) 본고에서는 HSK 어휘등급표에 없는 단어를 일괄적으로 등급 외 어휘에 포함시켰다. 예컨대 인명, 지명, 기관명 등과 같은 어휘도 등급 외 어휘에 포함시켜 비율을 계산하였다.

A 유형 교재		B 유형 교재	
교재 텍스트	등급 외 비율	교재 텍스트	등급 외 비율
중국어마스터STEP2	47.9%	신공략중국어기초	46.8%
중국어마스터STEP3	48.4%	신공략중국어초급	51.7%
중국어마스터STEP4	47.1%	신공략중국어프리토킹	54.7%
중국어마스터STEP5	51.1%	신공략중국어실력향상1	50.3%
중국어마스터STEP6	55.5%	신공략중국어실력향상2	47.8%
넥서스중국어1	38.4%	신공략중국어고급	56.2%
넥서스중국어2	43.1%		
스마트중국어1	33.6%		
스마트중국어2	41.9%		
스마트중국어3	44.9%		
스마트중국어4	45.1%		
평균	44.1%	평균	48.5%

위의 표에서 보이듯이 중국어 교재에서 HSK 어휘표와 일치하지 않은 비율이 상당히 높다. 이를 통해 우리는 저자들이 교재를 편찬할 때 HSK 어휘 등급과 수량으로만 한정하지 않고 다양한 문화어휘, 신조어, 관용어 등을 사용한다는 것을 알 수 있다. 이러한 특징은 기본어휘표나 정해진 어휘 규범을 반드시 준수해야 하는 중고등학교 교과서의 사용 양상과는 다른 점이다.

Ⅵ. 주요 어휘 범주의 사용 양상

1. 중국어 교재 텍스트의 키워드 추출

본고에서는 중국어 교재 텍스트에서 자주 사용되는 기능어나 내용어가 무엇인지를 파악하기 위해 코퍼스 키워드 분석 방법을 사용해 보았다. 키워드(keyword)는 어떤 텍스트의 특징을 핵심적으로 보여주는 단어를 의미

한다. 통계적 관점에서 키워드는 텍스트에서 유난히 강조되거나 자주 사용되는 단어이다.

키워드를 추출하기 위해서는 두 종류의 코퍼스가 필요하다. 첫째는 참조 코퍼스(reference corpus)로서 비교의 대상이 되는 코퍼스이다. 둘째는 연구자가 분석하고자 하는 관찰 코퍼스이다. 본고에서는 참조 코퍼스로 TORCH 2009·2014 코퍼스를 활용하였다. 그리고 관찰 코퍼스로는 A 유형, B 유형, C 유형 교재 텍스트를 이용하였다.

아래는 키워드 분석을 통해서 추출된 어휘 중에서 키워드 수치(로그우도비)가 높은 100개의 단어를 정리한 결과이다.

〈표 14〉 중국어 교재에 나타난 주요 키워드 100개

단어	키워드수치	단어	키워드수치	단어	키워드수치
你	9921.22	一点儿	728.27	高兴	382.62
我	8857.14	现在	714.14	问	382.48
吗	4817.80	多	684.49	不	381.17
去	4120.33	可是	675.77	生日	380.31
吧	2746.93	昨天	671.56	他	380.06
什么	2636.76	那儿	642.74	时候	375.98
呢	2093.89	真	620.21	贵	374.76
好	2050.96	以后	554.41	觉得	374.54
很	1981.38	太	547.4	都	373.09
哪儿	1961.31	上课	546.13	再见	369.08
了	1914.5	商店	534.83	找	364.92
您	1815.59	妈妈	500.84	这么	364.30
汉语	1704.49	好吃	497.42	多少	363.15
买	1624.08	要是	494.66	没	361.43
吃	1509.57	周末	489.96	家	353.26
今天	1433.01	一起	489.45	跟	347.33
看	1424.68	坐	487.87	呀	346.66
喜欢	1327.83	课	487.86	休息	346.39

단어	키워드수치	단어	키워드수치	단어	키워드수치
怎么样	1306.38	听说	485.61	不错	340.79
这儿	1154.12	忙	468.25	星期	339.41
老师	1147.72	吃饭	453.73	自行车	336.60
明天	1130.11	咱们	449.47	喝	331.55
想	1018.91	晚上	449.02	邮局	324.27
你们	991.98	请问	437.56	有点儿	322.64
谢谢	965.87	玩儿	426.65	做	321.91
怎么	952.64	几	425.97	宿舍	320.12
啊	878.06	便宜	421.04	下雨	320.05
菜	841.93	有	417.63	快	312.11
朋友	813.48	还是	408.24	公共汽车	311.00
我们	804.46	打算	399.27	钱	307.96
请	775.42	衣服	396.54	大夫	305.65
不太	767.56	留学生	392.35	天气	305.59
得	729.05	爸爸	382.99	叫	299.01

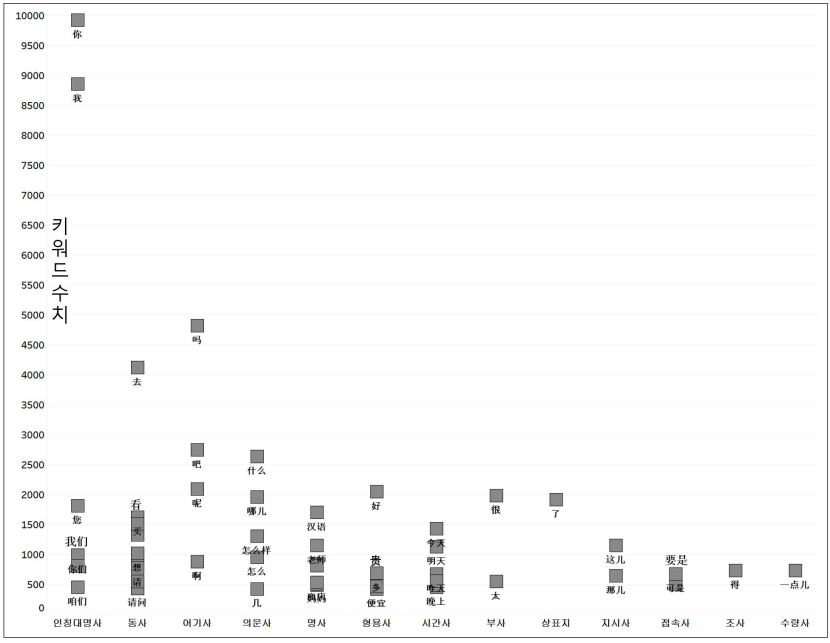
위의 표에서 키워드로 추출된 것은 일반 영역의 텍스트에 비해서 유난히 많이 사용된다고 추정되는 단어이다. 예를 들어 '我', '你', '我们', '咱们', '吧', '呢', '什么', '谁', '哪儿', '汉语', '老师', '今天', '星期' 등의 단어는 일반 영역(신문, 문학작품, 방송대본, 학술자료, 백과사전)의 텍스트보다 훨씬 높은 빈도로 사용된다. 위의 키워드 목록을 토대로 드러나는 어휘적 특징은 몇 가지로 요약할 수 있다.

① 인칭대명사의 사용이 매우 빈번하게 관찰된다. 특히 '我', '你', '我们', '咱们', '你们' 등의 1인칭·2인칭 대명사의 사용이 두드러진다.
② 중국어 어기사 중에는 의문을 나타내는 '吗'의 사용이 가장 두드러진다. 이 밖에도 '吧', '呢', '啊' 등의 어기사가 자주 사용된다.
③ 일반 텍스트에 비해 의문사의 사용 비율이 매우 높다. '什么', '谁', '哪儿', '怎么', '怎么样' 등이 그러하다.

④ 중국어 교재 텍스트에서 자주 사용되는 명사류, 동사류, 형용사류는 일상생활 관련 구어체 단어나 중국어 학습과 관련된 단어이다. 예를 들어 '汉语', '老师', '留学生', '教室', '去', '买', '吃', '学习', '上课', '休息', '贵', '便宜', '忙', '累' 등이 그러하다.

2. 주요 어휘 범주의 사용 양상

본고에서는 어휘 범주별로 키워드 수치가 높은 단어를 그래프로 분석해 보았다. 아래의 그래프에서 보이듯이 중국어 교재 텍스트에서 인칭대명사(你, 我), 의문조사(吗), 의문사(什么, 哪儿) 등의 키워드 수치(로그우도비)가 매우 높은 것을 알 수 있다. 이밖에도 중국어 교재 텍스트에서는 각 어휘 범주별로 자주 사용되는 명사류, 동사류, 형용사류 등이 있다.



〈그림 5〉 중국어 교재에서 키워드 수치가 높은 어휘 범주

(1) 인칭대명사, 의문사, 지시사

중국어 교재 텍스트에서는 1인칭, 2인칭 대명사가 아주 많이 사용된다. 그리고 중국어 교재 텍스트에서 의문사와 지시사의 사용비율도 높다. 이는 교재 본문이 대화문 형태로 구성된 것과 밀접한 관련이 있다. 특히 초급 교재일수록 본문이 간단한 대화문 형식으로 짜여 있어 인칭대명사 및 의문사의 사용비율이 매우 높다. 예를 들어 ≪스마트 중국어 2≫의 본문 구성을 보면 대화체 형식으로 문장이 제시되고 인칭대명사와 의문사가 많이 사용되는 것을 알 수 있다.

(3) 朴民秀: 你的爱好是什么？
张　京: 我的爱好是游泳。
朴民秀: 你为什么喜欢游泳？
张　京: 游泳对身体好，还可以减肥呢。(≪스마트 중국어 2≫)

〈표 15〉 중국어 교재에서 상용되는 인칭대명사, 의문사, 지시사

你(9921.22), 我(8857.14), 您(1815.59), 你们(991.98), 我们(804.46), 咱们(449.47), 他(380.06), 他们(293.44), 什么(2636.76), 哪儿(1961.31), 怎么样(1306.38), 这儿(1154.12), 怎么(952.64), 几(425.97), 多少(363.15), 谁(232.45), 哪(207.01)

(2) 명사

중국어 교재에서 가장 많은 비중을 차지하는 품사는 명사이다. 교재 텍스트에서 고빈도로 출현하는 명사만을 보더라도 그 교재에서 주로 다루는 내용이 무엇인지 파악할 수 있다.

명사는 교재 유형과 수준별로 사용되는 특징이 다르다. 한국인 저자가 집필한 교재와 중국인 저자가 집필한 교재가 다르고 초급 교재와 중고급 교재의 차이가 있다. 예를 들어 A 유형 교재와 B 유형 교재에서 두드러지는 명사는 일정한 차이를 보인다.

〈표 16〉 교재 유형별·수준별 명사 키워드의 차이

A 유형 교재		B 유형 교재	
초급	중고급	초급	중고급
(고유명사)汉语, 韩国, 韩国人, 中国, 北京; (친족명) 哥哥, 爸爸, 妈妈; (일반명사): 家, 学校, 银行, 老师, 咖啡, 手机, 生日, 电影; (교통수단)地铁, 飞机, 火车; (시간사)今天, 明天, 今年 등	(고유명사)汉语, 韩国, 韩国人, 中国, 中国人; (일반명사)普通话, 性格, 外貌, 红色, 谐音, 沙尘暴, 学期, 老师, 年糕, 手术, 宠物, 电影, 职员, 同屋, 新郎 新娘, 天气, 收据, 钱包 ; (시간사)今天, 周末 등	(고유명사)汉语, 汉字, 中国, 中国人; (친족명)爸爸, 妈妈, 爱人; (일반명사)老师, 商店, 邮局, 宿舍, 教室, 留学生, 食堂, 房间, 毛衣, (교통수단)自行车, (시간사) 今天, 明天, 今年 등	(고유명사)汉语, 汉字 中国人; (친족명) 爱人, 伯母, 伯父; (일반명사) 留学生, 孩子, 广告, 生活质量, 安乐死, 老师, 生日, 银行 时候, 飞机, 房子, 早饭, 旅行; (시간사)今天, 周末, 昨天 등

위의 표에서 보이듯이 고유명사 중에는 공통적으로 '汉语', '中国' 등이 많이 사용된다. 차이점은 A 유형 교재에서는 고유명사 중에 '韩国', '韩国人' 등의 사용이 두드러지고 B 유형 교재에서는 '中国人', '汉字'라는 단어가 두드러진다. 일반명사 중에서도 A 유형 교재에서는 '咖啡', '电影', '银行' 등이 자주 사용된다. 또한 '外貌', '性格', '宠物', '沙尘暴', '手术', '年糕' 등이 두드러진다. 이에 비해 B 유형 교재에서는 이러한 단어들이 자주 관찰되지 않는다. B 유형 교재에서는 친족명 중에 '爱人', '伯父', '伯母' 등이 자주 사용되는 것이 특징이다. A 유형 교재에서는 이러한 단어들의 빈도가 높지 않다. 그리고 학업과 관련해서는 '留学生', '宿舍', '食堂' 등의 단어가 많이 사용된다. 이는 유학생이 중국어를 배운다는 상황 설정으로 교재가 구성되었음을 보여준다. 또한 교통수단과 관련하여 A 유형 교재에서는 '地铁', '飞机', '公共汽车'가 많이 등장한다. 반면에 B 유형 교재에서는 '自行车'가 많이 사용된다. 종합적으로 보면 A 유형 교재는 한국인의 관점에서 문장을 구성하였기 때문에 한국인들에게 익숙한 상황에 해당하는 명사들이 특징적으로 추가되었다고 할 수 있다. 이에 비해 B 유형 교

재는 중국인의 관점에서 편찬되었기 때문에 중국의 전통문화, 생활방식, 중국의 교육환경과 관련된 명사들이 자주 사용된 것을 알 수 있다.

(3) 동사, 형용사, 부사

중국어 교재 텍스트에서 동사는 명사와 더불어 많이 사용되는 품사이다. 그 중에서도 두드러지는 단어는 방향동사 '去'이다. 이 외에도 학업·운동·쇼핑·여행·전화 등과 관계된 동사들이 키워드로 추출되었다.

〈표 17〉 중국어 교재에서 자주 사용되는 동사

去(4120.33), 买(1624.08), 吃(1509.57), 看(1424.68), 喜欢(1327.83), 想(1018.91), 谢谢(965.87), 请(775.42), 上课(546.13), 坐(487.87), 听说(485.61), 吃饭(453.73), 请问(437.56), 玩儿(426.65), 有(417.63), 打算(399.27), 问(382.48), 觉得(374.54), 再见(369.08), 找(364.92), 休息(346.39), 喝(331.55), 做(321.91), 下雨(320.05), 叫(299.01), 学(297.01), 试(295.83), 来(254.91), 听(249.98), 尝(243.27), 学习(231.81), 旅行(219.8), 说(215.73), 考试(208.96), 打电话(200.46), 祝(194.82), 上网(181.8), 懂(178.81), 结婚(176.94), 知道(176.8), 打(176.07)

본고의 조사 결과에 따르면 동사는 교재 유형별로 큰 차이를 보이지는 않는다. A 유형 교재와 B 유형 교재는 모두 위의 표에서 사용되는 동사들이 자주 사용된다. 물론 A 유형 교재에서 두드러지게 관찰되는 동사들도 있다. 예를 들어 카드결제, 성형수술 등을 나타내는 동사는 A 유형 교재에서 자주 관찰된다. 반대로 B 유형 교재에서 더 자주 사용되는 단어도 있다. 예컨대 흡연 관련 단어는 B 유형 교재에서 자주 관찰된다.

(4) 迈　克: 请问，这儿可以刷卡吗？
　　售货员: 我们这儿不能刷卡，只能用人民币。≪JRC중국어3≫
(5) 抽烟是从抽喜烟开始的。我本来不抽烟，人家结婚的时候，递上一支喜烟，你不抽不抽也得抽一支吧。≪신공략중국어고급≫

중국어 교재에서 자주 사용되는 형용사와 부사도 교재 유형별로 큰 차이를 보이지 않는다. 일반 영역과 비교했을 때 교재 텍스트에서 자주 사용되는 형용사는 '好', '多', '好吃', '便宜', '贵', '漂亮', '累', '忙' 등이다. 사용빈도가 높은 부사로는 정도부사, 범위부사, 빈도부사 등이 있다.

〈표 18〉 중국어 교재에서 자주 사용되는 형용사 및 부사

好(2050.96), 多(684.49), 好吃(497.42), 忙(468.25), 便宜(421.04), 高兴(382.62), 贵(374.76), 不错(340.79), 快(312.11), 舒服(296.1), 有意思(239.18), 漂亮(233.42), 麻烦(224.59), 累(201.85), 好看(200.84), 辣(183.55), 很(1981.38), 不太(767.56), 真(620.21), 太(547.4), 一起(489.45), 还是(408.24), 都(373.09), 这么(364.3), 有点儿(322.64), 就(310.32), 常常(284.98), 挺(251.07)

(4) 기능어: 전치사, 상표지, 어기사

중국어 교재 텍스트에서 전치사, 상표지, 어기사 등의 사용빈도를 조사해 보면 일반 영역의 코퍼스와 비교할 때 일정한 차이를 보인다. 그러나 교재 유형별로는 큰 차이를 보이지 않는다. A 유형 교재, B 유형 교재에서 공통으로 관찰되는 기능어 사용상의 몇 가지 특징을 정리하면 다음과 같다.

전치사는 '跟', '给', '比' 등의 사용빈도가 높다. 그러나 일반 영역의 코퍼스에 비해 '把', '被', '对' 등의 전치사는 상대적으로 적게 출현한다.

(6) 사용빈도가 상대적으로 높은 전치사
跟(347.33), 给(277.39), 比(251.89)
(7) 사용빈도가 상대적으로 낮은 전치사[11]
把, 被, 对, 对于, 为, 用, 与, 于

11) 교재 코퍼스와 일반 영역 코퍼스를 비교했을 때 사용빈도가 두드러지게 높은 것은 긍정적 키워드(positive keyword)라고 한다. 반대로 보통의 사용빈도보다 두드러지게 낮은 것은 부정적 키워드(negative keyword)라고 한다. 위에 제시된 전치사들은 모두 값이 음수(−)로 나오는 것들이다.

위의 예를 통해 중국어 교재에서는 수혜격 구문(给字句)와 비교구문(比字句)의 사용빈도가 높은 것을 알 수 있다. 반대로 처치구문(把字句), 피동문(被字句) 등은 일반 영역의 코퍼스에 비해 사용빈도가 낮다고 볼 수 있다. 이밖에도 문어체 전치사의 사용빈도가 낮다.

중국어 상표지 중에는 동작의 완료와 상태의 변화를 나타내는 '了'의 사용빈도가 두드러진다. 중국어 교재 텍스트에서 '了'의 사용빈도는 일반 영역의 코퍼스보다 훨씬 높다. 한편 진행을 나타내는 '正在', '在'의 사용도 두드러지며 경험을 나타내는 '过'도 빈번하게 사용된다. 그러나 이에 비해 '着'의 사용빈도는 교재의 유형에 상관없이 낮은 빈도를 보인다. 일반 영역의 자료에서는 '着'의 사용빈도가 아주 높으나 중국어 교재 텍스트에서는 그렇지 않다.

(8) 사용빈도가 상대적으로 높은 상표지
了(1914.5), (正)在(195.23), 过(30.12)
(9) 사용빈도가 상대적으로 낮은 상표지
着

문말에 사용되는 어기사 중에는 의문 표지로 사용되는 '吗'의 사용이 두드러진다. 본고의 조사 결과 '吗'의 키워드 수치는 4817.8이나 된다. 이는 중국어 교재 텍스트에서 '吗'로 구성된 문장이 매우 많다는 것을 의미한다. 중국어 교재에서는 구어 형식의 묻고 답하는 대화문 구성이 많아 의문 표지가 자주 사용된다. 이외에도 명령, 의문, 추측의 의미를 나타내는 '吧'의 사용빈도도 높다. 뿐만 아니라 '呢', '啊', '呀' 등과 같은 어기사도 빈번하게 사용된다.

(10) 사용빈도가 상대적으로 높은 상표지
吗(4817.8), 吧(2746.93), 呢(2093.89), 啊(878.06), 呀(346.66)

Ⅶ. 고빈도 정형 표현의 추출과 분류

1. N-gram 모델을 이용한 정형 표현 추출하기

최근 외국어 교육에서는 2 단어 이상의 어휘 덩어리(chunk) 형태로 나타나는 정형 표현(formulaic expression)이 주목을 받고 있다. 특히 고빈도로 출현하는 정형 표현은 단어보다는 크지만 어휘처럼 학습되고 사용된다는 점에서 의사소통 능력을 구성하는 중요한 요소로 인정받고 있다.[12] 정형 표현은 일반적으로 자주 사용되는 고정적·관습적 표현을 가리킨다. 이는 정형화된 연결어구(formulaic sequences), 어휘 다발(lexical bundles), 어휘 덩어리(lexical chunks) 등으로도 불리기도 한다.

본고에서는 고빈도로 나타나는 2 단어 이상의 어휘 결합을 정형 표현으로 정의하고 중국어 교재에서 자주 사용되는 유형을 정리해 보았다.

일반적으로 코퍼스에서 고빈도의 정형 표현을 추출하는 방법은 'n-gram' 모델을 기초로 한다. 하나의 어휘 단위를 '1-gram'이라고 가정하면 2개의 단어 결합은 '2-gram'이 된다. 3개의 단어 결합은 '3-gram'이 된다. 4개의 단어 결합은 '4-gram'이 된다. 예를 들어 한 문장은 {我/r(1-gram), 我/r-要/v(2-gram), 我/r-要/v-先/d(3-gram), 我/r-要/v-先/d-走/v(4-gram).....}과 같이 분석될 수 있다.

이러한 방식으로 n-gram 조합을 추출하면 중국어 교재 텍스트에서 사용된 단어와 품사 배열 패턴을 모두 조사할 수 있다. 정형화된 언어 표현도 단어의 배열을 기초로 하는 것이므로 고빈도로 출현하는 단어의 연쇄나 품사의 연쇄는 정형 표현을 추출하는데 기초자료가 된다. 예를 들어 다음의 문장을 보기로 하자.

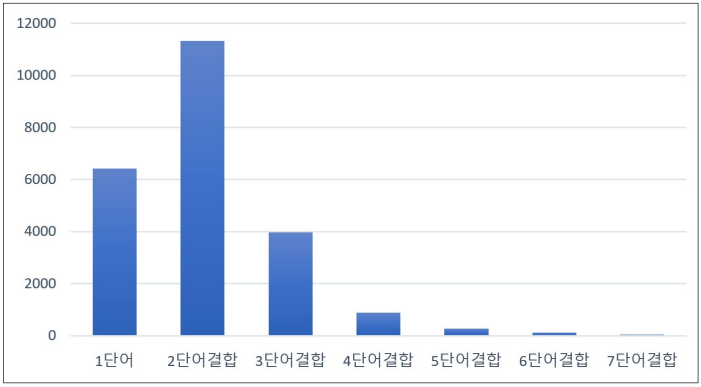
(11) a. 难怪/d 你/r 的/u 汉语/n 那么/r 好/a 。/w

12) 강현화(2007), 이유미(2019) 참조.

b. 단어 배열 연쇄 : 难怪 你 的 汉语 那么 好 。
c. 품사 배열 연쇄 : d r u n r a

(11.a)는 ≪중국어마스터2≫에 나온 문장을 태깅한 것이다. (11.b)는 이 중에서 단어 배열의 정보를 제시한 것이다. (11.c)는 품사 배열의 순서를 제시한 것이다. 본고에서는 단어 배열 정보와 품사 배열 정보를 기초로 고빈도의 n-gram을 모두 추출하였다.

'n-gram' 통계 방법을 이용하여 중국어 교재 텍스트에서 자주 사용되는 단어 결합 패턴을 조사한 결과 사용빈도가 높은 단어 결합은 1-gram(한 단어), 2-gram(두 단어 결합)이다. 그 다음으로 3-gram(세 단어 결합), 4-gram(네 단어 결합) 패턴이다. 예를 들어 본고에서 대상으로 삼은 중국어 교재 텍스트 중에 5회 이상 사용된 개별 단어(1-gram)은 6,411개이다. 그리고 이 단어들이 결합된 2단어 조합(2-gram)은 11,318개에 달한다. 3단어 조합(3-gram)은 3,963개에 달한다. 이는 텍스트에서 개별 단어뿐만 아니라 그 단어가 결합되어 만들어진 정형 표현이 많다는 것을 의미한다. 아래의 표는 n-gram 통계 방법으로 5회 이상 출현한 단어의 조합을 그래프로 나타낸 것이다.



〈그림 6〉 교재 텍스트의 n-gram 단어 유형

n-gram 통계 방법을 이용하면 단어의 조합뿐만 아니라 품사기호로 배열된 수많은 조합을 찾을 수 있다. 중국어 교재 텍스트 중에는 사용빈도가 1,000회를 넘어서는 품사 배열 패턴도 많이 관찰된다. 아래는 품사 배열 패턴 중에서 사용빈도가 1,000회 이상 사용된 것을 정리한 결과이다.[13]

〈표 19〉 중국어 교재 텍스트에서 자주 사용되는 품사 배열

동사(v)+조사(u)(11079), 부사(d)+동사(v)(11043), 대명사(r)+동사(v)(10826), 동사(v)+명사(n)(10739), 동사(v)+동사(v)(9401), 명사(n)+동사(v)(9115), 조사(u)+명사(n)(9048), 동사(v)+대명사(r)(7209), 명사(n)+명사(n)(6410), 명사(n)+조사(u)(6140), 대명사(r)+명사(n)(5893), 부사(d)+형용사(a)(5486), 대명사(r)+부사(d)(5326), 명사(n)+부사(d)(5305), 형용사(a)+조사(u)(4667), 대명사(r)+조사(u)(3558), 대명사(r)+부사(d)+동사(v)(2762), 형용사(a)+명사(n)(2714), 부사(d)+부사(d)(2655), 동사(v)+부사(d)(2585), 전치사(p)+대명사(r)(2487), 전치사(p)+명사(n)(2446), 조사(u)+동사(v)(2407), 동사(v)+형용사(a)(2365), 명사(n)+형용사(a)(2359), 대명사(r)+형용사(a)(2244), 동사(v)+조사(u)+명사(n)(2170), 수사(m)+명사(n)(2137), 명사(n)+조사(u)+명사(n)(2127), 형용사(a)+동사(v)(2122), 부사(d)+동사(v)+동사(v)(2102), 대명사(r)+동사(v)+조사(u)(2063), 부사(d)+동사(v)+조사(u)(1991), 대명사(r)+동사(v)+동사(v)(1931), 명사(n)+동사(v)+조사(u)(1851), 동사(v)+명사(n)+조사(u)(1835), 동사(v)+동사(v)+명사(n)(1825), 명사(n)+부사(d)+동사(v)(1823), 대명사(r)+조사(u)+명사(n)(1822), 동사(v)+수사(m)(1821), 수사(m)+양사(q)(1763), 대명사(r)+대명사(r)(1730), 대명사(r)+동사(v)+대명사(r)(1613), 명사(n)+명사(n)부사(d)(1592), 명사(n)+대명사(r)(1580), 시간사(nt)+동사(v)(1574), 동사(v)+대명사(r)+명사(n)(1515), 부사(d)+명사(n)(1497), 대명사(r)+동사(v)+명사(n)(1480), 부사(d)+동사(v)+명사(n)(1473), 명사(n)+부사(d)+형용사(a)(1452), 동사(v)+동사(v)+조사(u)(1436), 대명사(r)+계사(vl)(1415), 명사(n)+전치사(p)(1393), 대명사(r)+시간사(nt)(1338), 명사(n)+동사(v)+명사(n)(1331), 대명사(r)+전치사(p)(1318), 부사(d)+계사(vl)(1292), 부사(d)+형용사(a)+조사(u)(1281), 계사(vl)+명사(n)(1277), 부사(d)+부사(d)+동사(v)(1253), 형용사(a)+조사(u)+명사(n)(1243), 동사(v)+대명사(r)+동사(v)(1237), 조사(u)+형용사(a)(1226), 명사(n)+동사(v)+동사(v)(1195), 동사(v)+처소사(ns)(1188), 양사(q)+명사(n)(1183), 계사(vl)+대명사(r)(1177), 명사(n)+c(1148), 조사(u)+명사(n)+동사(v)(1143), 동사(v)+대명사(r)+조사(u)

13) 괄호 안의 숫자는 사용빈도를 나타낸다.

(1136), 조사(u)+대명사(r)(1124), 대명사(r)+명사(n)+동사(v)(1112), 조사(u)+명사(n)+부사(d)(1107), 동사(v)+부사(d)+동사(v)(1107), 형용사(a)+형용사(a)(1105), 동사(v)+명사(n)+명사(n)(1083), 부사(d)+전치사(p)(1082), 동사(v)+전치사(p)(1070), 부사(d)+동사(v)+대명사(r)(1063), 시간사(nt)+대명사(r)(1055), 처소사(ns)+명사(n)(1029), 동사(v)+동사(v)+대명사(r)(1024)

위의 표에 제시된 품사 배열 패턴은 중국어 교재 텍스트에서 추출된 것이다. 이 중에는 통사적으로 완전한 구성도 있지만 그렇지 않은 것도 존재한다. 예를 들어 '[부사(d)+동사(v)]', '[부사(d)+형용사(a)]'는 동사구나 형용사구가 되겠지만 '[부사(d)+부사(d)]', '[조사(u)+명사(n)]'는 완전한 구를 이루지 못한다. 그러나 이러한 품사 배열 패턴은 통사구조의 특징이나 고빈도 정형 표현을 파악하는데 기초자료로 활용될 수 있다. 예를 들어 '[부사(d)+형용사(a)]' 구성 중에 '很(부사)+형용사(a)'로 사용되는 정형 표현을 조사해 보면 다음과 같다.

(12) '[很(d)+형용사(a)]' 결합 유형:
很 多(287), 很 好(245), 很 高兴(118), 很 大(110), 很 忙(77), 很 难(54), 很 好吃(45), 很 有意思(39), 很 漂亮(36), 很 重要(36), 很 容易(35), 很 高(35), 很 好看(33), 很 不错(32), 很 近(31), 很 远(26), 很 长(25), 很 冷(22), 很 便宜(21), 很 流利(21), 很 累(19), 很 贵(18), 很 厉害(16), 很 晚(16), 很 方便(15)

중국어 교재에는 의문사의 사용이 빈번한데 그 중에 '什么+명사(n)'구조 사용되는 표현도 상당히 많다. 또한 가능보어 부정형식의 정형 표현도 자주 관찰된다.

(13) '[의문사(r)+명사(n)]' 결합 유형:
什么 时候(224), 什么 工作(61), 什么 地方(50), 什么 意思(50), 什么 名字(46), 什么 东西(31), 什么 问题(29), 什么 颜色(22), 什么 运动

(21), 什么 书(19), 什么 菜(18), 什么 关系(16), 什么 礼物(16)

(14) '[동사(v)+不(d)+동사/형용사(v/a)]' 결합 유형:
对 不 起(143), 差 不 多(110), 说 不 定(38), 听 不 懂(34), 舍 不 得(26), 受 不 了(24), 找 不 到(21), 看 不 懂(19), 怪 不 得(18), 忘 不 了(17), 吃 不 了(16), 睡 不 着(16), 了 不 起(15), 去 不 了(14), 忍 不 住(13), 看 不 出(11), 离 不 开(11), 吃 不 下(10), 认 不 出(10), 谈 不 上(10)

위의 예에서 보이듯이 중국어 교재 텍스트에서 반복적으로 사용되는 단어의 조합이 존재한다. 이들은 문장에서 한 '덩어리(chunk)'를 이루어서 자주 사용되는 경향이 있다. 그중에서 출현빈도가 높고 수량이 많은 표현은 중국어 학습의 기초자료가 된다는 측면에서 관심을 가질 필요가 있다.

2. 정형 표현에 대한 유형 분류

중국어 교재 텍스트에서 어떤 표현이 자주 사용된다는 것은 그것이 일종의 관습적인 성격을 가진다는 것을 의미한다. 관습적인 성격을 가진다는 것은 언어사용자가 언어 표현을 자주 쓰다 보니 자연스럽게 한 덩어리처럼 굳어졌음을 의미한다. 하나의 덩어리로 굳어지게 되면 하나의 단위처럼 인식되고 점점 정형화된 표현으로 발전하게 된다. 예를 들어 중국어에서 새해인사를 하거나 생일축하를 할 때는 '新年快乐', '生日快乐'라고 하는데 '기쁘다'는 의미를 고려하면 이 표현만 쓸 수 있는 것이 아니다. '新年愉快', '新年高兴, '生日愉快', '生日高兴'과 같은 표현도 떠올릴 수 있다. 물론 '愉快'나 '高兴'은 기뻐하는 주체가 사람이어야 한다는 의미적 제약이 존재하지만 실제 언어 사용의 관점에서는 사람들이 자주 '新年快乐', '生日快乐'라는 표현을 써왔기 때문에 관습적으로 굳어진 것일 수도 있다.

중국어 교재 텍스트에서 관찰되는 정형 표현은 연어, 관용구, 자유결합, 문법적 표현 문형 등 다양한 유형이 존재한다.[14] 그 중에서도 사용빈도와

수량면에서 많은 비중을 차지하는 것은 자유결합과 문법적 표현 문형이다.

(1) 자유결합(free combination)

두 개 이상의 단어가 관습적으로 자주 결합되어 사용되더라도 이들은 대개 자유 결합 형식이 많으며 의미적으로도 투명한 것이 일반적이다. 본 연구에서 분석 대상으로 삼은 중국어 교재 텍스트에서도 고빈도의 정형화된 표현은 대개 자유결합 형식의 통사구조를 가지며 어휘소의 축자적인 의미의 합으로 해석해 낼 수 있다. 예를 들어 중국어 교재 텍스트에서 명사구나 동사구의 정형 표현을 보면 자유결합 형태가 다수를 차지한다.

(14) 명사구 고빈도 자유결합
中国朋友(68회), 学校食堂(25회), 公司职员(11회),
一个人(187회), 一个月(52회), 一个问题(25회), 这个问题(63)

(15) 동사구 고빈도 자유결합
在哪儿(282회), 做什么(266),一起去(219), 去哪儿(213회),
在这儿(112회), 看电视(98회), 看电影(97회), 坐飞机(65회)

우리가 이러한 표현을 정형 표현이라고 말할 수 있는 것은 그것이 통사적인 측면에서 고정되어 있고 출현빈도가 높기 때문이다. 하나의 통사구조가 텍스트 안에서 한 덩어리로 자주 결합되는 특성 때문에 관습적인 성격을 띠기도 한다. 예를 들어 '这时候', '那时候', '什么时候'처럼 'X+时候'구조, '很长时间', '多长时间', '业余时间'처럼 'X+时间'로 사용된 표현을 모두 단어로 볼 수는 없다. 그러나 이들은 중국어 교재 텍스트에서 한 단위처럼 자주 어울려 사용된다.

(2) 문법적 표현 문형(expressive sentence pattern)

정형 표현에 대한 논의는 대개 어휘적 요소의 결합체에 초점을 맞추는

14) 이에 대한 더 자세한 논의는 강병규(2009)를 참고하기 바람.

경우가 많지만 어떤 경우에는 문법적 요소가 결합하여 일정한 문법적 기능을 나타내기도 한다. 이러한 것을 외국어 교육학계에서 흔히 표현 문형(expressive sentence pattern)이라고 부른다. 한국어 교육 분야에서는 표현 문형을 중요한 문법 교육 단위로 삼고 있다. 표현 문형은 여러 한국어 교재에서 문법 항목의 절반 이상을 차지할 만큼 비중이 높다.[15] 표현 문형은 형태소, 단어, 구 등의 형식으로 완전히 정의하기는 어렵다. 이들은 2개 이상의 형태소 또는 단어가 결합되어 특정한 문법적 기능을 한다. 예를 들어 한국어 교재에서는 '-(으)ㄹ 뿐만 아니라', '-기 때문에', '-은(는) 바람에' 등과 같은 것을 표현 문형으로 정의하고 전체 표현이 가지는 의미 기능을 중심으로 선행 요소와 후행요소의 제약조건을 설명하는 방식으로 구성된다. 최근 들어 표현 문형이 한국어 교육에서 주목을 받는 이유는 문법적 기능이 다양하고 생산성이 높기 때문이다.

본고에서 조사한 결과에 따르면 중국어 교재 텍스트에서도 일정한 문법적 기능을 하는 표현 문형으로 들 수 있는 것이 적지 않다. 예를 들어 'X的时候(-할 때에)'라는 구성은 'X的时间(-하는 시간)'과는 다르다. 'X的时间'은 명사적인 기능만 하지만, 'X的时候'는 선행절 뒤에 사용되어 동작의 시점을 나타내는 기능도 할 수 있다. 'X的时候'는 '的'와 '时候'를 분리하는 것보다 전체로서 하나의 의미를 가지고 선행요소와 후행요소에 일정한 제약 조건이 있는 것으로 이해하는 것이 합리적이다.

(16) 休息的时候，他去喝一杯咖啡，吃一点儿东西。≪한어교정2≫
(17) 她工作很忙，学习的时间少了，也不能睡懒觉了。≪신공략초급≫

코퍼스에서 n-gram 통계 방식으로 2 단어 이상의 결합을 조사하다 보면 생산성이 높은 정형화된 표현을 많이 찾아내게 된다. 예컨대 '并不……',

15) 원운하·박덕유(2017), 이유미(2019) 등에 따르면 한국어 교재에는 정형화된 표현 형태로 이루어진 여러 표현 문형이 제시되어 어휘 학습과 문법 학습에 활용된다.

‘根本不……’, ‘从来不……’ 등의 복합 구성도 일정한 문법 기능을 가지며 생산성이 높은 유형에 속한다.

〈표 20〉 ‘X不/没…’ 형태의 복합 구성

유형	사용빈도	유형	사용빈도
也 不……	287	从来 没……	23
都 不……	126	从来 不……	19
还 没……	124	根本 不……	14
能 不……	88	却 不……	12
还 不……	79	怎么 还 不……	12
也 没……	78	再也 不……	11
都 没……	46	怎么 还 没……	9
一点儿 也 不……	44	谁 也 不……	8
怎么 不……	39	倒 不……	7
好 不……	38	怎么 能 不……	7

위와 같은 복합 구성을 모두 특정한 문법적 기능을 담당하는 표현 문형으로 볼 수는 없다. 반복적으로 사용되는 용례를 검토하다 보면 어떤 것이 자유결합이고 어떤 것이 문법적 기능을 하는 표현 문형인지 판단하기 어려운 경우도 많다. 그러나 이 중에서 일부는 점점 그 형태가 고정되어서 일정한 문법적 기능을 하는 방향으로 발전할 가능성이 있다. 따라서 정형화된 표현에 대한 분류는 이분법적인 것이 아니라 정도성의 측면에서 접근하는 것이 바람직하다. 중요한 것은 이러한 표현이 하나의 덩어리로 저장되어 어휘처럼 관습적으로 자주 사용된다는 점이다.

그동안 중국어 교육에서 어휘 덩어리로 나타나는 정형화된 표현은 많은 주목을 받지 못하였다. 대신 개별 단어와 문법 항목을 언어의 구성 요소로 간주하여 교재에 반영하였다. 그러나 교재 텍스트의 언어 내용을 조사해 보면 단어 이상의 어휘적 단위(lexical unit)가 많이 출현하는 것을 알 수

있다. 한국어 교재나 영어 교재에서 고빈도로 사용되는 정형 표현이 점점 중시되는 것처럼 중국어 교재 연구에 있어서도 정형 표현에 대해 관심을 가질 필요가 있다.

Ⅷ. 어휘 분포와 결합 정보에 기초한 중국어 교재의 군집분석

중국어 교재 텍스트의 어휘 사용빈도, 어휘 다양성 수치, 어휘 범주, 정형 표현, 문장의 길이 등을 종합해 보면 교재별로 일정한 유사점과 차이점이 존재한다. 앞 장에서 언급된 것처럼 한국인 저자가 쓴 A 유형 교재, 중국 원서를 편역한 B 유형 교재, 중국에서 사용되고 있는 C 유형 교재는 유사한 특징도 있지만 여러 가지 측면에서 다른 점도 많다.

본고에서는 어휘 분포와 결합 정보에 기초하여 중국어 교재의 유사점과 차이점을 찾아 유형을 분류하는 작업을 시도해 보았다. 유형 분류에 사용된 방법은 통계적 요인분석(factor analysis)[16]과 군집분석(cluster analysis)[17] 모델이다. 우선 요인분석 방법을 사용하여 중국어 교재에 나타나는 여러 가지 변수들을 합쳐서 소수의 요인으로 축약하였다. 요인분석 결과 중국어 교재의 정보들은 단어 사용빈도, 어휘 다양성, 문장의 길이, 정형 표현 등의 몇 가지 요인으로 축약할 수 있었다. 그리고 요인분석에 따라 선택된 변수를 중심으로 유사한 특징이 있는 교재들을 통계적 군집분석 방법을 통해 군집으로 나누는 작업을 진행하였다. 아래의 그림은 중국어 교재의 유사성 척도에 근거하여 Ward 연결 방법[18]으로 계층적 군집분석을 실시

16) 요인분석은 다양한 변수를 가지는 자료에서 변수들의 상관관계를 조사하여 상호 연관성이 있는 변수들을 묶어서 소수의 요인으로 축약하여 자료를 해석하는 통계적 방법이다.

17) 군집분석은 다양한 특성을 가진 자료에서 유사성과 차이점을 중심으로 유사한 개체끼리 묶어내는 통계적 방법이다. 군집분석은 자료를 유형화화하거나 분류하기 위한 목적으로 많이 활용되는 방법이다.

18) 계층적 군집분석 방법 중에 Ward 연결 방법(Ward linkage)은 군집 평균과 군집 내 유클리디언 거리를 최소한도로 증가시키면서 군집화하는 방법이다. 이

하고 그 결과를 수형도 형태의 일종인 덴드로그램 방식으로 시각화한 것이다. 이 덴드로그램의 수평 축에서 왼쪽에 묶인 개체일수록 유사한 군집이다.

〈그림 7〉 중국어 교재에 대한 군집 분석 결과

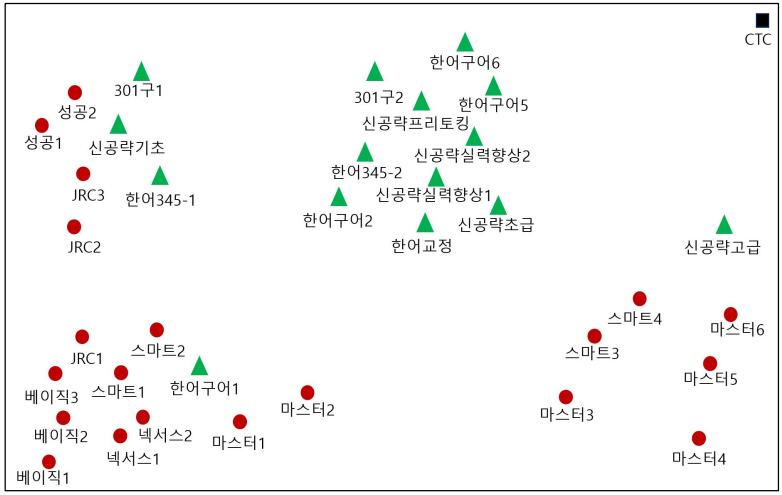
군집분석 결과에 따르면 본고에서 조사한 중국어 교재는 크게 5가지 유형으로 분류된다. 그리고 각 군집별로 단어 유형빈도, TTR, STTR, 단어평균음절수, 문장평균길이, 어휘 등급, 인칭대명사, 어기사, 동사, 전치사, 정형표현의 평균 사용빈도 등의 항목에서 모두 일정한 차이가 존재한다. 아래의 표에서 보이듯이 각 항목의 분산분석 결과에 따르면 F값과 유의확률이 모두 통계적으로 유의미한 차이가 있다는 것을 나타내준다.

방법은 비슷한 크기의 군집끼리 잘 묶어내는 장점이 있어 자주 사용된다.

〈표 21〉 교재 군집에 따른 어휘 사용 통계 정보

유형		1군집	2군집	3군집	4군집	5군집
단어유형(Types)		487.45	546.71	1409.20	1482.71	19478.00
단어평균길이		1.42	1.44	1.45	1.54	1.57
어휘 다양성	TTR	18.83	13.07	16.17	21.85	9.18
	STTR	25.44	25.16	34.00	39.61	59.38
문장평균길이		4.87	5.73	8.06	11.36	14.61
어휘등급	1급어휘	97.45	112.29	127.90	103.00	148.00
	2급어휘	55.45	73.14	109.60	87.00	149.00
	3급어휘	44.82	62.00	146.70	122.86	295.00
	4급어휘	31.09	36.43	160.80	158.29	595.00
	5급어휘	15.64	20.29	119.10	154.71	1220.00
	6급어휘	6.73	4.86	43.20	87.43	1700.00
인칭대명사 평균빈도		108.55	145.29	250.80	106.00	2465.00
어기사 평균빈도		41.09	67.86	69.10	19.29	751.00
동사 평균빈도		39.27	78.00	128.60	42.86	957.00
전치사 평균빈도		6.01	11.73	25.30	19.00	478.00
정형표현 평균빈도		3.19	6.80	8.61	12.00	260.00

[군집별 분산 분석(F): 16.8〈F〈591.9, 유의확률(p〈0.001)]



〈그림 8〉 중국어 교재에 대한 유형 분류 시각화

위의 그림은 군집분석 결과를 기초로 교재의 유형을 2차원 공간상에 재배치해 본 결과이다. 위의 그림에서 원(●)으로 표시된 것은 A 유형 중국어 교재 텍스트를 나타낸다. 삼각형(▲)은 B 유형 교재 텍스트를 나타낸다. 그리고 사각형(■)은 C 유형 교재 텍스트를 나타낸다.

① 제1유형: 1군집

첫 번째 군집에 속하는 교재는 한국에서 출판된 초급 교재(A유형)가 다수를 차지한다. 예를 들어 ≪JRC맛있는중국어1≫, ≪중국어마스터1·2≫, ≪넥서스1·2≫, ≪베이직중국어1·2·3≫, ≪스마트중국어1·2≫ 등이다. B 유형 교재로는 ≪한어구어1≫이 포함된다. 이 유형에 속하는 교재는 평균 487개의 단어가 사용되었고, 어휘 다양성 수치가 비교적 낮고(STTR:25.44), 문장의 평균 길이도 짧고(4.87단어), 어휘 등급도 1급·2급 어휘가 다수를 차지한다.

② 제2유형 :2군집

두 번째 유형에 속하는 교재는 A 유형 교재 중에 ≪JRC맛있는중국어2·3≫, ≪성공중국어1·2≫가 포함되고 B 유형 교재 중에 ≪301구로끝내는중국어회화1≫, ≪신공략중국어기초≫, ≪한어구어345-1≫가 포함된다. 이 유형에 속하는 교재는 평균 546개의 단어가 사용되었고, 어휘 다양성 수치가 비교적 낮고(STTR:25.17), 문장의 평균 길이도 짧고(5.73단어), 어휘 등급도 1급·2급 어휘가 다수를 차지한다.

③ 제3유형: 3군집

세 번째 유형에 속하는 교재는 중국 원서를 편역하여 출판한 B 유형에 속한다. 이 유형에 속하는 교재는 ≪한어구어345-2≫, ≪301구로끝내는중국어회화2≫, ≪한어교정≫, ≪신공략중국어프리토킹≫, ≪신공략중국어실력향상≫, ≪한어구어2·5·6≫ 등이다. 이 유형에 속하는 교재는 평균 1,409개의 단어가 사용되었고, 어휘 다양성 수치가 중간 수준(STTR:34.00)이고, 문장의 평균 길이는 1·2유형보다 다소 길며(8.06) 어휘 등급은 1급·2급 이외에 3급·4급·5급 어휘도 많이 사용된다.

④ 제4유형: 4군집

네 번째 유형에 속하는 교재는 A유형 중고급 교재와 B유형 고급 교재이다. 이 유형에 속하는 교재는 ≪다락원중국어마스터3·4·5·6≫과 ≪신공략중국어고급≫이다. 이 유형에 속하는 교재는 평균 1,482개의 단어가 사용되었고, 어휘 다양성 수치가 1~3 군집보다 높은 수준(STTR:39.61)이고, 문장의 평균 길이도 1~3 군집보다 길며(11.36), 어휘 등급은 1급·2급보다는 3급·4급·5급 어휘가 다수를 차지한다.

⑤ 제5유형: 5군집

다섯 번째 유형에 속하는 것은 중국에서 출판된 교재 텍스트 자료이다. 이 자료는 단어의 수량도 많고, 어휘 다양성 수치도 매우 높고(STTR: 59.38), 문장의 평균 길이도 14.74에 달하며 어휘 등급도 4급~6급 단어가 많은 비중을 차지하여 난이도가 높다고 할 수 있다.

군집분석 결과를 통해서도 중국어 교재는 언어적 측면에서 다양한 특징을 가진다는 것을 알 수 있다. 교재별로 어휘 사용의 수준이 다르고 자주 사용되는 내용어, 기능어 등이 다르다. 통계적 군집분석 결과를 통해서 볼 때 한국에서 출판된 교재와 중국에서 출판된 교재는 언어적 특징에 일정한 차이를 보인다. 한국인이 주저자인 교재와 중국인이 주저자인 교재는 내용어, 기능어, 정형 표현의 사용 양상이 다르다. 동일한 초급교재라도 한국인 주저자가 출판한 교재는 중국 교재와는 다른 특징을 가진다. 그리고 초급·중급·고급처럼 중국어 교재의 수준에 따라 단어의 사용빈도·다양성·등급, 문장의 길이 등의 측면에서 뚜렷한 차이가 존재한다.

Ⅸ. 결론

한국과 중국에서 제2언어 학습자를 대상으로 사용되고 있는 중국어 교재는 다양한 유형이 존재한다. 그리고 중국어 교재별로 편찬 목적에 따라 어휘 사용, 문장 구성 등이 서로 다른 양상을 보인다.

본고에서는 코퍼스 언어학에서 사용되는 계량적 분석 방법을 사용하여

중국어 교재의 어휘 사용빈도·다양성·수준, 어휘 범주, 정형화된 표현 등을 분석하여 교재별로 존재하는 언어적 특징을 살펴보았다. 소위 '대외한어교재'로 불리는 중국어 교재는 어떤 특징이 있는지, 교재별로 차이가 있다면 구체적으로 어떠한 양상을 보이는지 등을 고찰하였다. 특히 중국에서 출판된 교재에 비해 한국에서 편찬된 교재 텍스트가 가지는 언어적 특징이 무엇인지를 중점적으로 탐색하였다.

본고는 중국어 교재 텍스트에 대한 전반적인 특징을 분석하기 위해서 다양한 종류의 교재를 수집하는 것이 필요하다는 전제하에 국내에서 출판된 중국어 교재 35권과 중국에서 출판된 교재 263권의 자료를 데이터화하는 작업을 진행하였다. 그리고 이 교재를 세 종류로 나누어 비교 분석하였다. 이 세 부류는 한국인 주저자가 편찬한 교재(A 유형), 중국 교재를 한국에서 편역한 교재(B 유형), 중국에서 출판된 교재(C 유형)이다.

첫 번째로 본고는 중국어 교재 텍스트에서 사용된 단어와 문장의 사용빈도를 조사하였다. 분석 결과에 따르면 단어의 출현빈도와 유형빈도를 볼 때 'A 유형 〈 B 유형 〈 C 유형'의 순서를 가진다. 중국어 교재에 사용된 단어의 음절수는 평균 '±1.5'음절로서 교재별로 큰 차이가 없었다. 그러나 문장의 평균적인 길이는 교재 유형과 수준별로 큰 차이를 보였다. 교재 유형별로는 A 유형 교재의 문장 길이가 가장 짧고 C 유형 교재의 문장 길이가 가장 길었다. 특히 C 유형 텍스트는 단어의 평균 길이가 14.61로서 확연한 차이가 난다. 수준별로 나누어 보면 초급 교재일수록 문장의 길이가 짧고 중고급 교재일수록 문장이 길어진다. 즉 교재의 수준이 높아지면 단어의 수량도 증가하며 문장도 길고 복잡해진다.

두 번째로 중국어 교재별로 STTR 수치를 조사하여 얼마나 다양한 어휘가 사용되는지를 비교해 보았다. 분석 결과 A 유형 교재의 STTR 값이 가장 낮고 B 유형 교재가 그 다음이며 C 유형 교재가 가장 높다. STTR 값이 낮다는 것은 사용된 단어가 상대적으로 적고 어휘의 다양성이 낮다고 할 수 있다. 반면에 STTR 값이 높으면 교재 텍스트의 다양성이 높다고 해석된다. 어휘 다양성이 높으면 그만큼 어휘 학습량이 증가하고 난이도가

높아지게 된다.

세 번째로 교재별 어휘 사용의 수준을 파악하기 위해 HSK 어휘 등급표와 비교해 보았다. 전반적으로 보면 어휘 등급의 비율은 교재 유형별로 큰 차이가 없으나 수준별로는 차이를 보였다. 즉, 교재의 어휘 등급 수준은 교재의 수준에 비례하는 경향이 있었다. 초급 수준의 교재일수록 1급~3급 어휘의 비율이 높고 중고급 수준의 교재일수록 4급~5급 어휘의 사용비율이 늘어난다. 한가지 주목할 점은 중국어 교재에 사용된 어휘 중에서 HSK 어휘 목록에 제시되지 않은 비율도 높다는 것이다. 이를 통해 우리는 저자들이 교재를 편찬할 때 규범적인 어휘 등급으로 한정하지 않고 다양한 고유명사, 문화 어휘, 신조어, 관용어 등을 사용한다는 것을 알 수 있다. 이러한 특징은 기본어휘표나 정해진 어휘 규범을 준수해야 하는 중고등학교 교과서의 사용 양상과는 다른 점이다.

네 번째로 중국어 교재 텍스트에서 자주 사용되는 어휘 범주가 무엇인지 파악하기 위해 키워드 분석을 진행하였다. 분석 결과에 따르면 중국어 교재 텍스트에서 인칭대명사, 의문조사, 의문사 등의 키워드 수치가 매우 높은 것을 알 수 있었다. 이밖에도 각 어휘 범주별로 자주 사용되는 명사류, 동사류, 형용사류 등이 존재한다. 교재 유형별로 보면 A 유형 교재는 한국인의 관점에서 문장이 구성되었기 때문에 한국적 담화 상황을 반영한 어휘 범주가 특징적으로 사용된 측면이 있다. 이에 비해 B 유형 교재는 중국인의 관점에서 편찬되었기 때문에 중국의 전통문화, 생활방식, 중국의 교육환경과 관련된 단어들이 자주 사용된 것을 알 수 있다.

다섯 번째로 고빈도로 나타나는 2 단어 이상의 어휘 결합을 정형 표현으로 정의하고 중국어 교재에서 자주 사용되는 유형을 정리해 보았다. 분석 결과에서도 알 수 있듯이 중국어 교재 텍스트에도 반복적으로 사용되는 단어의 조합이 존재한다. 이들은 문장에서 한 '덩어리'를 이루어서 자주 사용되는 경향이 있다. 특히 고빈도로 출현하는 정형 표현은 단어보다는 크지만 어휘처럼 학습되고 사용된다는 점에서 의사소통 능력을 구성하는 중요한 요소가 된다. 따라서 출현빈도와 생산성이 높은 정형 표현은 중

국어 학습의 기초자료가 된다는 측면에서 관심을 가질 필요가 있다.

마지막으로 어휘 분포와 결합 정보에 기초하여 중국어 교재의 유사점과 차이점을 찾아 유형을 분류하는 작업을 시도해 보았다. 우선 요인분석 방법을 사용하여 중국어 교재에 나타나는 여러 변수를 합쳐서 소수의 요인으로 축약하였다. 그리고 요인분석을 통해 선택된 변수를 중심으로 유사한 특징이 있는 교재들을 분류하였다. 통계적 군집분석 결과에 따르면 본고에서 표본으로 선택한 중국어 교재는 5가지 군집 유형으로 나누어진다. 5개의 군집 유형은 교재의 주저자가 한국인인지 중국인인지에 따라 나누어지고 교재의 수준이 초급인지 중고급인지에 따라 구분된다.

이상에서 살펴본 바와 같이 중국어 교재는 언어적 측면에서 다양한 유형이 존재하며 저마다의 특징이 있음을 알 수 있다. 교재별로 어휘 사용의 수준이 다르고 자주 사용되는 내용어, 기능어가 다르다. 한국에서 출판된 교재와 중국에서 출판된 교재는 일정한 언어적 특징의 차이가 존재한다. 한국인이 주저자인 교재와 중국인이 주저자인 교재는 내용어, 기능어, 정형 표현의 사용 양상이 다르다. 이러한 분석 결과는 향후 중국어 교재 텍스트를 분류하고 연구하는데 참고자료로 활용할 수 있을 것이다.

＜참고문헌＞

강병규, 〈중국문학 텍스트 번역을 위한 상용구 자동 추출과 언어학적인 고찰〉, ≪중국어문논역총간≫ 25집, 2009.

강선주, 〈고등학교 9 종 중국어 교과서의 문화영역 분석〉, ≪중국학논총≫ 37집, 2012.

강현화, 〈한국어 표현문형 담화기능과의 상관성 분석 연구〉, ≪이중언어학≫ 34집, 2007.

권지혜·이석재, 〈한국과 아시아권 고등학교 1학년 영어 교과서 코퍼스 대조 분석〉, ≪Secondary English Education≫ 12집, 2019.

김나래, 〈중국어 교재에 나타난 현대중국어 부사 '又'의 의미와 용법〉, ≪중국어문논역총간≫ 43집, 2018.

김나래·김석영·박종한·손남호·신원철·이미경·이연숙, 〈중국의 중국어 교재 분포와 개발 현황 분석〉, ≪중국중국어 교재≫ 88집, 2016.

김난미, 〈고등학교 중국어Ⅰ 교과서에 반영된 중국 문화 내용 분석〉, ≪중국문화연구≫ 33집, 2016.

김재은·최인철, 〈고등학교 영어 교과서, EBS 수능 연계 교재, 대학수학능력시험의 코퍼스기반 난이도 비교 분석〉, ≪Multimedia-Assisted Language Learning≫ 18집, 2016.

김정은, 〈한, 미 고등학교 중국어 교재 비교 분석〉, ≪중국어중국어 교재≫ 68집, 2015.

김종호, 〈고등학교 중국어 교과서 '본문 (text)'의 素材 분석〉, ≪중국어중국어 교재논집≫ 27집, 2004.

김현철·이준섭·권순자, 〈한국 출판 중국어 회화교재의 문장 성분 생략 현상 연구〉, ≪중국어중국어 교재지≫ 65집, 2018.

류다리, 〈교재 ≪신공략중국어(초급편)≫ 비교문형 제시에 관한 고찰〉, ≪중국어중국어 교재≫ 47집, 2006.

배은영, ≪2009 개정 고등학교 〈중국어〉 교과서 어휘 분석≫, 이화여대 석사논문, 2015.

배재석·김강림, 〈중·고등학교 중국어 교과서 주제별 어휘 비교 분석〉, ≪중국어문학논집≫ 79집, 2013.

손정애, 〈의사소통 기반의 중국어 교재에서 문법항목의 선정과 배열 연구: 성인 학습자용 초급회화 교재를 중심으로〉, ≪중국언어연구≫ 42집, 2012.

손정애·김나래, 〈중국어교육을 위한 현대중국어 부사 '환(還)'의 의미와 용법 연구〉, ≪중국중국어 교재≫ 88집, 2016.

원운하·박덕유, 〈한·중 한국어 교재에 제시된 표현문형 고찰〉, ≪교육문화연구≫ 23집, 2017.

이미경, 〈한국의 중국어 교재 분포와 개발 현황 분석〉, ≪중국어교육과연구≫ 28집, 2018.

이용훈·이규형·김하응, 〈표준화된 타입-토큰 비율, 어휘성장곡선, 그리고 영어교재분석〉, ≪영어학연구≫ 21집, 2015.

이유미, 〈한국어 학습자의 숙달도에 따른 표현 문형 사용 양상 연구〉, ≪언어학≫ 27집, 2019.

이지은·신수영, 〈신HSK 5·6급 독해 영역 어휘의 코퍼스 분석〉, ≪언어와정보사회≫ 25집, 2015.

임재민, 〈중국어교과 교육과정과 교과서 비교 연구〉, ≪중국언어연구≫ 83집, 2019.

전문기·임인재, 〈코메트릭스를 이용한 중학교 1 학년 개정 영어 교과서의 코퍼스 언어학적 비교 분석〉, ≪영어교육연구≫ 21집, 2009.

진 현, 〈중국어 3급 교재 어휘 분석-国际汉语教学通用大纲과 비교하여〉, ≪중국어문학≫ 80집, 2009.

황인옥, ≪중국어 교재 개발을 위한 중국어 문형 연구: 중, 고등학교 중국어 교과서 분석을 중심으로≫, 이화여대 석사논문, 2001.

金佳垠(2017), ≪中韩汉语教材词汇· 语法项目编排的比较研究≫, 上海师范大学 硕士论文.

邱 爽, ≪对外汉语教材词汇的计量研究≫, 黑龙江大学 硕士论文, 2014.

王淑珍, ≪对外汉语教材中的构式研究≫, 鲁东大学 硕士论文, 2019.

郑贤淑, ≪中韩汉语教材比较研究≫, 四川师范大学 硕士论文, 2018.

柏崎有里, ≪日本本土初级汉语教材研究-以编写者差异为视角≫, 曲阜师范大学 硕士论文, 2018.

중국어 코퍼스 사이트

国际汉语教学助手网 http://www.aihanyu.org/index.aspx

北外语料库语言学 http://corpus.bfsu.edu.cn/index.htm

<Abstract>

This study examined the linguistic features of Chinese textbooks using quantitative analysis methods used in corpus linguistics. In particular, we focused on finding out the linguistic features of textbooks compiled in Korea compared to textbooks published in China.

In this paper we collected data from 35 Chinese textbooks published in Korea and 263 textbooks published in China. And this textbook was divided into three categories. They are texts compiled by Korean authors (type A), texts translated from Chinese textbooks in Korea (type B), and textbooks published in China (type C).

First, we investigated the frequency of words and sentences used in Chinese textbooks. According to the analysis result, when looking at the token frequency and type frequency of a word, it has the order of 'A type 〈B type 〈C type'. The average length of sentences varies greatly depending on the textbook type and level.

Second, the STTR statistics for each Chinese textbook were examined to compare how various vocabulary words are used. As a result of analysis, the STTR value of the A-type textbook is the lowest, the B-type textbook is the next, and the C-type textbook is the highest.

Third, to understand the level of vocabulary use for each textbook, we compared it with the HSK Vocabulary Rating Table. Overall, the proportion of vocabulary grades did not differ significantly by textbook type, but by level.

Fourth, keyword analysis was conducted to understand what vocabulary categories are frequently used in Chinese textbooks. According to the results of the analysis, it was found that the keyness of personal

pronouns and interrogations in Chinese textbooks was very high.

Fifth, formulaic expressions frequently used in Chinese textbooks were extracted. As can be seen from the results of the analysis, there is a combination of words repeatedly used in Chinese textbook text. The formulaic expression needs to be studied important in that it becomes the basic data of Chinese education.

Finally, in this paper, cluster analysis was performed based on the linguistic characteristics of Chinese textbooks. According to the results of statistical cluster analysis, Chinese textbooks are divided into five cluster types. The results of this analysis can be used as a reference to study texts in Chinese textbooks in the future.

Key Words : 중국어 교재(chinese textbooks), 출현빈도(token frequency), 유형빈도(type frequency), 단어 음절수(word syllables), 문장 길이(sentence length), 어휘 다양성(vocabulary diversity), TTR, STTR, 어휘 등급(vocabulary grade), 키워드(key words), 정형표현(formulaic expression), 군집분석(cluster analysis)