ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/372512682

□□□ □□□□ □□ −≪□□□□≫ □□□□ □□□□

Article *in* ALTAI HAKPO · June 2023

DOI: 10.15816/ask.2023..33.003

CITATIONS
0

READS
5

3 authors, including:



Jeong-up Do
Seoul National University
22 PUBLICATIONS **2** CITATIONS
SEE PROFILE



Jung sung hoon
Seoul National University
91 PUBLICATIONS **1,345** CITATIONS
SEE PROFILE

All content following this page was uploaded by Jeong-up Do on 22 July 2023.

The user has requested enhancement of the downloaded file.

만주어 코퍼스의 구축
— ≪滿文老檔≫ 태조편을 대상으로 —

최운호(국립목포대학교 부교수)*
정성훈(국립목포대학교 조교수)
도정업(서울대학교 선임연구원)

Choi, Woonho, Sunghoon Jung & Jeongup Do. 2023. Construction of the Manchu corpus: Focusing on *Manwen laodang Taidzu*. *Altai Hakpo* 33. 67–87. The Altaic Society of Korea.

By using *Manwen laodang Taidzu* as the target material, this study seeks a way to construct a morphologically annotated corpus for linguistic and cultural research. The Manchu corpus constructed in this study is an annotated corpus, and the annotation level is set to morphological annotation. In this study, we construct a morphologically annotated corpus by building a rule-based stemmer and tagging system using existing dictionaries, collectible vocabulary lists, and language knowledge of Manchu experts. *Manwen laodang* electronic document for annotated *Manwen laodang* corpus construction is Kim et al. (2019) was used. The purpose of the system constructed in this study is to produce a corpus that can be used for research by accurately tagging a large amount of data, which is a historical corpus. In other words, the goal of this study is to produce an annotated corpus by applying various heuristics to build a reusable annotated corpus. We anticipate that different approaches will also be used in Manchu and Altaic studies. Furthermore, we anticipate that the corpus released as a result of this study will serve as the basis for data-centered Manchu language research.

Keywords: Written Manchu, *Manwen laodang*, morphologically annotated corpus, stemmer, tagging system

* 최운호: 제1저자, 정성훈: 공동저자, 도정업: 교신저자

https://doi.org/10.15816/ask.2023..33.003

www.kci.go.kr

1. 들어가는 말

코퍼스는 언어 연구의 목적을 위해 균형성과 대표성을 고려하여 전산화된 자료를 말한다(강범모 2008: 7). 코퍼스는 언어모델(language model) 구성, 공기어(共起語) 분석, 언어 단위의 분포 연구, 사전 편찬 등을 위한 기초 자료이며 언어의 기능, 의미 연구를 위한 자원이라고 할 수 있다. 특히 최근 기계 학습(machine learning)과 심층 학습(deep learning) 기법이 발달하고 보급되면서 언어자료를 수집하여 코퍼스를 구축하는 것이 더욱 중요해졌다.

언어 연구뿐만 아니라 문학, 역사학 등 인문학의 다양한 분야에서도 코퍼스를 활용한 연구가 활발히 이루어지고 있다. Jockers(2013)에서는 아일랜드계 미국 작가의 소설 출판 현황에 대한 메타데이터 분석을 수행해서 문학의 사회학 연구를 수행하였고 기존 문학사 연구와는 다른 관점을 제시하였다. 우리 학계에서도 최근에는 역사학, 문학, 사회학, 문헌정보학 등 언어 연구 이외의 다양한 분야에서 코퍼스 자료를 활용하는 디지털 인문학(digital humanities) 연구가 활성화되고 있다.[1)]

만주어를 비롯한 알타이어 연구에서도 자료 중심의 연구를 위하여서는 활용 가능한 코퍼스의 구축이 필요하다. 지금까지 한국의 알타이언어 연구에서는 절멸 위기의 알타이언어의 수집과 보존을 위한 알타이언어 아카이브 구축 연구 수행 사례가 있다.[2)] 이 연구에서는 문서화(documentation)라는 방법론으로 자료의 수집과 구축을 추진하였다는 점에서 절멸위기 언어의 연구를 위한 자료의 우선 확보라는 목적을 충실히 수행해 왔다고 할 수 있다. 문서화가 자료 수집과 보존이라는 목적을 위한 것이라면 자료의 주석(annotation)을 통해 언어 분석, 텍스트 마이닝 등의 분석 방법을 활용한 개념어의 역사 의미 연구 등으로 알타이언어의 연구 주제를 확장할 필요가 있다. 이를 위해서는 주석 코퍼스(annotated corpus)의 구축이 필수적이다.

만주어 코퍼스는 자료가 만들어지던 시기의 언어 현상을 잘 담고 있을 뿐만 아니라, 역사, 문화 연구를 위한 원천 자료로서도 활용 가치가 높다. 이 연구는 ≪滿文老檔≫ 태조편을 대상 자료로 삼아서 언어, 문화 연구를 위한 형태주석 코퍼스(morphologically annotated corpus)를 구축하는 것을 목적으로 한다. 코퍼스의 구축을 위한 전자사전, 시스템 구성, 그리고 형태주석 태그셋에 대한 정보와 함께 언어, 문화 연구를 위하여 만주어 코퍼스 자료를 활용하는 방안을 이 연구에서 함께

1) 이태훈·정유경(2020)은 디지털 인문학 연구방법론을 적용한 다양한 사례들을 담고 있다.

2) 김주원(2008: 352)에서는 절멸위기 알타이언어 조사를 통한 자료 수집과 디지털 아카이브 구축 과정을 상세히 설명하고 있다.

www.kci.go.kr

제시하고자 한다.

2. 코퍼스 구축 대상 자료와 방법

2.1. ≪만문노당≫ 소개

≪만문노당≫은 유권점 만주문으로 쓰인 만문당안을 일컫는다. 이 만문당안은 원래 'tongki fuka sindaha hergen i dangse(유권점 당안)'이라고 적혀있을 뿐 한자로 된 서명은 없었는데, 1905년에 이 당안을 발견한 나이토 토라지로(內藤虎次郎)가 "滿文老檔"이라고 이름을 붙인 것이 지금의 명칭으로 굳어졌다(閻崇年 1998: 59-60). 무권점 만주문자를 읽을 수 있는 사람이 없어지는 것을 염려한 건륭제는 건륭 6년(1741)에 오르타이(Ortai, 鄂爾泰)와 시오이 위완멍(Sioi Yuwanmeng, 徐元夢)에게 무권점 만주문자로 표기된 어휘의 사전을 편찬하게 하였는데,[3] 이 사전을 만들 때 활용한 자료가 만주문자로 기록된 공문서인 만문당안이었다. 이 당안이 오래되어서 파손된 것을 다시 잘 보존하기 위해서 표구하여 제본하기도 하였다(성백인 1978: 98-103). 이렇게 하여 1607년부터 1636년까지의 기록이 포함된 입관전 만문당안 37책이 새롭게 장정되었다. 이 당안을 슈허더(Šuhede, 舒赫德)의 건의에 따라 건륭 40년(1775)부터 중초하기 시작하여 건륭 43년(1778)에 완성하였다. 먼저 만문당안을 무권점자와 유권점자의 흘림체로 각각 중초하였고, 이것을 저본으로 하여 다시 무권점자와 유권점자 정자체로 정리하였다. 이 4종은 북경(北京) 내각(內閣)에 보관하였다. 건륭 43년(1778)부터 다시 정자체로 된 중초본을 정본으로 하여 부본(副本)을 만들었다. 이 부본은 건륭 45년(1780)에 성경(盛京)의 숭모각(崇謨閣)에 보관하였다(김두현 2008: 262). 본 연구에서 사용한 문헌은 내각장본 ≪만문노당≫의 태조편이다. 내각장본 ≪만문노당≫의 간략한 서지 사항은 다음과 같다.

≪滿文老檔≫ 內閣藏本 tongki fuka sindaha hergen i dangse
四周雙邊, 白口, 單魚尾, 無行格, 半葉 7行
광곽 크기 28.5cm×18.3cm
26函 180冊(太祖朝 10函 81冊, 太宗朝 16函 99冊)

앞서 언급한 바와 같이 ≪만문노당≫은 무권점자 당안을 포함하고 있는 만문당안을 중초하였기 때문에 권점을 적는 방식은 ≪어제증정청문감(御製增訂淸文鑑≫

3) 사전의 이름은 ≪無圈點滿文 十二子頭(tongki fuka akū hergen i bithe)≫이다.

www.kci.go.kr

(1771 序)의 방식을 따르고 있으나, 어휘나 문법은 전반적으로 ≪만문원당(滿文原檔)≫의 만주어를 그대로 따르고 있다. 이러한 이유로 ≪만문노당≫은 ≪만문원당≫의 내용을 발췌하여 옮겨 쓰는 과정에서 이해하기 어려운 부분을 그대로 베끼고, 난상에 있는 두주(頭註)에서 해당 부분을 설명하는 방식을 택하였다. 따라서 이 문헌은 18세기 후반에 편찬되었으나 17세기 초반의 문법 현상을 반영하고 있다고 볼 수 있다.

2.2. 코퍼스 구축 방법

이 연구에서 구축하는 만주어 코퍼스는 주석 코퍼스(annotated corpus)이며, 주석의 단계는 형태 주석으로 설정하였다. 형태 주석 코퍼스는 단어를 형태소 단위로 분절(segmentation)하고 분절된 형태소에 적절한 형태 표지를 부착하여 제작한다. 형태 주석을 부착하는 방법은 여러 가지가 있을 수 있다. 우선 대상 언어를 잘 이해하고 분석할 수 있는 전문가가 단어 하나하나를 분절하고 주석을 부착하여 코퍼스를 만드는 방법이 있을 수 있다. 형태 분석을 수행해서 주석을 부착해야 하는 원시 코퍼스(raw corpus)의 양이 많을 때에는 전문가의 노력만으로는 짧은 시간에 많은 자료에 주석을 부착하는 것이 어렵다. 코퍼스 구축 시간을 줄이기 위해서는 형태소 분석기(morphological analyzer)[4] 또는 태거(tagger)를 제작하여 활용해야 한다. 기존에 이미 구축해 놓은 주석 말뭉치의 양이 상당한 경우에는 용례 기반 형태소 분석기를 구성하고 이미 분석한 경험이 있는 단어를 자동으로 용례에서 추출하여 태깅을 할 수 있다. 또는 주석 코퍼스를 훈련 데이터로 사용하여서 기계 학습 방법을 활용한 태거를 제작하여 활용하는 방법도 가능하다. 하지만 활용 가능한 주석 코퍼스가 없는 경우에 다량의 자료를 분석하기 위한 형태소 분석기 또는 태거를 제작하기 위해서는 규칙 기반 시스템(rule-based systems)을 제작하여 활용하는 방법이 효율적이다. 그리고 분석 대상 언어 전문가의 피드백을 통하여 시스템을 개선하면서 주석 코퍼스의 양과 정확도를 늘려나가면 짧은 기간에 다량의 주석 코퍼스를 구축하는 것이 가능하다.

이 연구에서는 기존에 활용 가능한 사전과 수집 가능한 어휘 목록, 그리고 만주어 전문가의 언어 지식을 활용한 규칙 기반 형태소 분석기와 태깅 시스템을 구축하

4) 만주어 자료만을 위한 형태소 분석 시스템은 기존에 발표된 것이 없으며, 형태 주석 말뭉치도 따로 구축되어서 공개된 것을 조사해 보았지만 발견하지 못했다. 이와는 별개로 로마자 표기법으로 만주어 자료를 구축할 때 활용할 수 있도록 공개 소프트웨어의 확장을 통해서 사용할 수 있는 맞춤법 검사기는 You(2014)에서 공개된 바 있다.

www.kci.go.kr

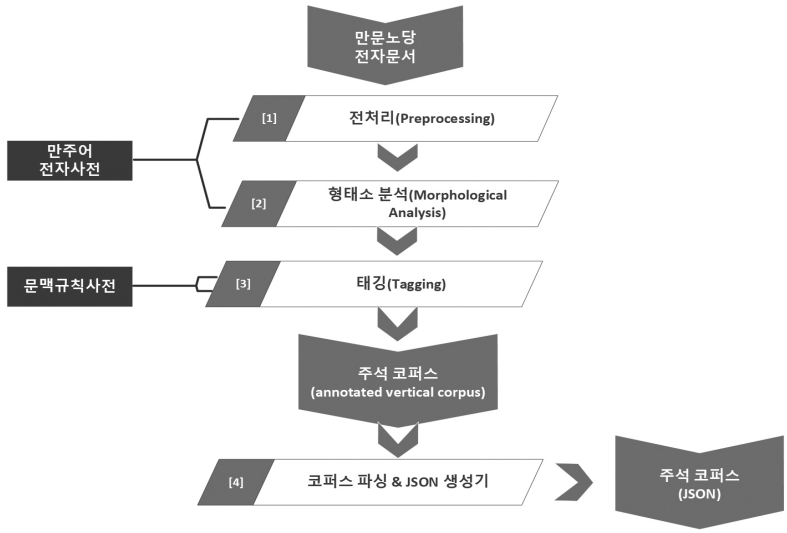


그림 1. *Manwen laodang* (滿文老檔) corpus annotation systems

여 형태 주석 코퍼스를 구축하도록 한다. 우선 규칙 기반 만주어 형태소 분석기[5]의 구현을 위해서 Norman(2013)의 자료를 기본 데이터로 입력하여 시스템 사전에 활용하였다. 그리고 김주원 외(2019)의 부록으로 기록된 고유명사를 복합어로 시스템 사전에 추가하였다. 〈그림 1〉은 "≪만문노당≫ 주석 말뭉치 구축 시스템"의 구성을 보여주는 시스템 구성도이다.

≪만문노당≫ 주석 코퍼스 구축을 위한 ≪만문노당≫ 전자문서는 김주원 외(2019)의 자료를 활용하였다. 이 연구에서 구축하는 시스템은 일반적인 자연어처리 시스템처럼 비정형 텍스트(free text)에 강건하게 대응해야 하는 시스템이 아니다. 역사 말뭉치에 해당하는 대량의 자료를 정확하게 태깅해서 연구에 활용할 수 있는 코퍼스를 생산하는 것이 목적이다. 요약하면, 강건한(robust) 자연어처리 시스템 구축이 목적이 아니라 활용 가능한 주석 코퍼스(reusable annotated corpus) 구축이 목적이고 이에 맞춰서 다양한 경험적 방법(heuristics)을 적용하여 주석 코퍼스를 생산해 내는 것이 연구의 목표이다. ≪만문노당≫ 주석 코퍼스 구축의 각 과정과

5) 이 연구에 참여하신 연구자들과의 논의를 통해서 ≪만문노당≫ 주석 코퍼스 구축 시스템과 그 결과로 생산된 주석 코퍼스를 통칭하여 '까막별'이라는 별칭으로 부르기로 하였다. '까막별'은 2022년 돌아가신 경성대 한국한자연구소의 연규동 교수님이 사용하셨던 별칭이다.

www.kci.go.kr

사용된 사전에 대한 설명은 3장에서 상세하게 제시한다.

3. 만주어 주석 코퍼스 구축

3.1 전자사전과 형태소 분석

≪만문노당≫ 자료를 형태 주석 코퍼스로 바꾸기 전의 원시 코퍼스는 김주원 외(2019)의 자료의 만주어 원문으로 구성하였다. 이 원시 코퍼스 자료는 묄렌도르프(Möllendorff) 전사법에 따른 로마자 표기법으로 작성되었다. 전자 사전의 구축과 분석 시스템의 편의를 위해서 시스템의 입력으로 사용되는 코퍼스, 그리고 전자 사전에서 š, ū는 각각 x, v로 변환하여 사용하였으며 배포용 자료에서는 이를 다시 š, ū로 변환하였다. ≪만문노당≫ 태조편은 모두 81권인데 〈표 1〉은 원시 코퍼스를 가공하여서 전처리기(preprocessor)로 복수어 단위 인식을 수행한 뒤 주석 코퍼스 제작을 위한 입력 자료[6]로 변환하여 놓은 것이다.

≪만문노당≫ 원시 코퍼스는 전처리와 형태소 분석 단계를 거쳐서 형태 분석 코

```
##!BOOK##!01
tongki fuka sindaha hergen i dangse .
○ cooha be waki seme tumen cooha be unggifi tosoho , tere
  tosoho cooha be acaha manggi , hvrhan_hiya ini gajire sunja
  tanggv boigon be , alin i ninggude jase jafafi , emu tanggv
  cooha be tucibufi boigon tuwakiyabuha , cooha gaifi genehe
  ilan beile de , ula i cooha heturehebi seme amasi niyalma
  takvraha , tere dobori , ula i tumen ujihe , muse tuttu ujifi
  ula i gurun de unggifi ejen obuha niyalma kai , ere bujantai
  musei gala ci tucike niyalma kai , jalan goidahakvbi , beye
  halahakvbi , ere cooha be geren seme ume gvnire , muse de
⇦ abkai gosime buhe amba horon bi , jai
⇦ ama han i gelecuke amba gebu bi , ere cooha be muse absi
  akv gidambi seme henduhe manggi , geren coohai niyalma gemu
  urgunjeme afaki saciki seme jabufi , tere bira be dooha ,
  doofi ,
```

표 1. *Manwen laodang* raw corpus

6) 앞에서 설명하였듯이, 묄렌도르프 전사법은 배포용 주석 코퍼스 생산 단계에서 적용된다. 전처리기를 통과한 코퍼스에서는 š, ū는 각각 x, v로 표기되어 있다. 또한 복수어 단위는 밑줄 '_'로 서로 연결되어서 하나의 어휘 단위로 형태소 분석 단계에서 사용되도록 미리 처리되었다.

www.kci.go.kr



그림 2. Lexicon of *Manwen laodang* corpus annotation systems

전처리 사전	표제어 수	표제어 예
MWU_6	1	uje dogon i xa ho xo
MWU_5	6	an giya wan ho wei, ba giya ho cang tun
MWU_4	61	bejang guwan tun gaxan, ca doo ho wei
MWU_3	615	ai ho wei, ajige age beile
MWU_2	987	abai nakcu, abatai taiji

표 2. Example entries in Multi-Word Unit Lexicon

퍼스로 만들어진다. 전처리와 형태소 분석 단계에서 사용하는 전자사전의 구성은 〈그림 2〉와 같다.

전처리(preprocessing) 사전은 인명, 지명 등 고유명사 복수어 단위(MWU, multi-word unit)를 형태 분석 단계에 진입하기 전의 전처리 단계에서 한 어휘 항목으로 처리하는 데 사용된다. MWU_6는 6개 단일어 연쇄로 이루어진 복수어 단위로 만주어 형태 주석 시스템의 표제어로 1개가 등록되어 있다. MWU_2는 2개의 단일어로 이루어진 복수어 단위로 모두 987개가 표제어로 등록되어 있다. 각 전처리 사전 항목의 예는 다음과 같다.

형태 분석 단계에서 단어 형성 규칙으로 처리하기 어려운 어휘들, 또는 시스템에서 형태 분석을 수행하는 것보다 기분석 사전의 참조를 통해서 분석 결과를 대체하는 것이 효율적인 어휘들은 따로 기분석 사전의 표제어로 등록하였다. 기분석 사전

www.kci.go.kr

어미(ending)	태그(tag)	어미(ending)	태그(tag)
ci	e::cve	mbi	e::fve.npst
cibe	e::cve	cina	e::fve.opt
fi	e::cve	ki	e::fve.opt
hai	e::cve	kini	e::fve.opt
hei	e::cve	habi	e::fve.prs.prf
hoi	e::cve	hebi	e::fve.prs.prf
mbime	e::cve	hobi	e::fve.prs.prf
me	e::cve	kabi	e::fve.prs.prf
nggala	e::cve	kebi	e::fve.prs.prf
nggele	e::cve	kobi	e::fve.prs.prf
nggolo	e::cve	hakvbi	e::fve.prs.prf.neg
pi	e::cve	hekvbi	e::fve.prs.prf.neg
tala	e::cve	kakvbi	e::fve.prs.prf.neg
tele	e::cve	kekvbi	e::fve.prs.prf.neg
tolo	e::cve	rahv	e::fve.prv
		mbihe	e::fve.pst.ipfv
ra	e::npst.ptcp	ha	e::pst.ptcp
re	e::npst.ptcp	he	e::pst.ptcp
ro	e::npst.ptcp	ho	e::pst.ptcp
rakv	e::npst.ptcp.neg	ka	e::pst.ptcp
		ke	e::pst.ptcp
		ko	e::pst.ptcp
		hakv	e::pst.ptcp.neg
		hekv	e::pst.ptcp.neg
		kakv	e::pst.ptcp.neg
		kekv	e::pst.ptcp.neg

표 3. Ending list of *Manwen laodang* corpus annotation systems

은 모두 917개의 어휘가 등록되어 있다. 시스템 구성 단계에서 1차로 구성된 목록과 2차로 구성된 목록이 물리적으로는 분할되어 있지만, 논리적으로는 하나의 사전으로 통합하여서 기분석 사전을 운영하게 된다.

www.kci.go.kr

어미 태그	설명
cve	**c**on**v**erb **e**nding
npst.ptcp	**n**on-**p**ast **p**artici**p**le
npst.ptcp.neg	**n**on-**p**ast **p**artici**p**le **neg**ative
fve.npst	**f**inite **v**erb **e**nding, **n**on-**p**ast
fve.opt	**f**inite **v**erb **e**nding, **opt**ative
fve.prs.prf	**f**inite **v**erb **e**nding, **pre**sent **p**er**f**ect
fve.prs.prf.neg	**f**inite **v**erb **e**nding, **pre**sent **p**er**f**ect, **neg**ative
fve.prv	**f**inite **v**erb **e**nding, **prev**entive
fve.pst.ipfv	**f**inite **v**erb **e**nding, **p**ast, **imp**er**f**ecti**v**e
pst.ptcp	**p**ast **p**artici**p**le
pst.ptcp.neg	**p**ast **p**artici**p**le, **neg**ative

표 4. Tag list of Manchu verbal endings

≪만문노당≫ 형태 주석 시스템의 어미 사전에는 모두 45개의 형태가 등록되어 있으며 그 목록은 〈표 3〉과 같다.

〈표 4〉에서는 어미의 분류에 사용된 태그를 제시하고 있다. 어미는 크게 부동사형 어미(cve, Con Verb Ending), 정동사형 어미(fve, Finite Verb Ending), 비과거형 어미(npst, Non-PaST), 과거형 어미(pst, PaST)로 구분한다.

≪만문노당≫ 형태 주석 시스템에서는 어휘 형태를 동사(verbal)와 동사가 아닌 비동사(non-verbal) 2개의 부류로 구분하였다. 이 구분은 형태, 의미, 기능 중 형태 중심의 분류를 사용하여 시스템의 효율을 높이기 위해 채택한 방법이다. 어휘 형태 목록은 Norman(2013)의 어휘 목록을 기본 자료로 하여 구성되었다. 동사 사전에는 동사 표제어로 등록된 어휘 항목 5,795개가 등록되어 있으며, 비동사(non-verbal) 사전에는 동사를 제외한 어휘 14,481개가 표제어로 등록되어 있다. 비동사 사전에는 미식별어(unknown)를 포함한 9개의 기호 표현과 고유명사, 형용사, 명사 또는 명사 상당 어휘들을 포함한 어휘들이 표제어로 등록되어 있다.

〈그림 3〉은 ≪만문노당≫ 코퍼스를 대상으로 설계된 형태소 분석 시스템의 구성도이다. loadLex는 전자사전을 관리하고 검색해주는 모듈이다. mcma[7]는 형태소 분절 및 분석을 수행해 주는 모듈인데 규칙 기반 시스템으로 설계되었다. 그리고 형

7) mcma는 '**M**an**C**hu **M**orphological **A**nalyzer'의 약자이다.

www.kci.go.kr

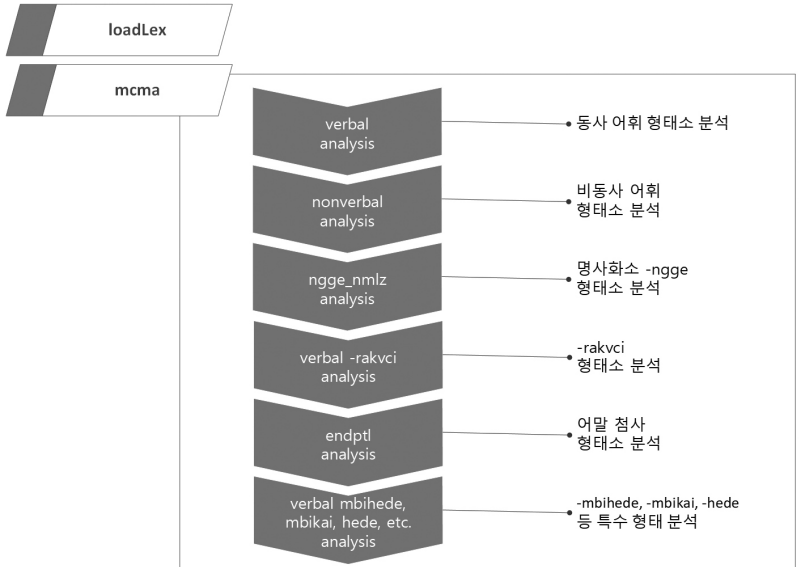


그림 3. Manwen laodang morphological analysis systems

단어	형태 분석 후보
acuhiyadabu	acuhiyadabu/verb::vv + ∅/e::fve.imperative
bayasai	bayasa/nv::nv + i/case::gen
	bayan/nv::nv + sa/nv::nv + i/case::gen
bahakvbi	ba/verb::vv + hakvbi/e::fve.prs.prf.neg
	baha/verb::vv + hakvbi/e::fve.prs.prf.neg
bisirei	bisire/nv::nv + i/case::gen
	bisi/verb::vv + re/e::npst.ptcp + i/case::gen
bisirengge	bisirengge/nv::nv
	bisi/verb::vv + re/e::npst.ptcp + ngge/nmlz::nmlz
burakvbi	bu/verb::vv+ rakv/e::npst.ptcp.neg + bi/verb::vi.npst

표 5. Examples from mcma (Manchu morphological analyzer)

태소 분석 모듈 안의 다양한 처리 함수들이 순차적으로 배치되어서 실행되어 입력 단어를 처리한다. mcma는 〈그림 3〉에 있는 순서를 거쳐서 ≪만문노당≫에서 사용된 단어들의 가능한 분석 결과를 모두 생성해내는 최대 분석의 방법으로 복수의 형

www.kci.go.kr

태 분석 결과를 생성해낸다. 〈표 5〉는 mcma에서 가능한 모든 분석 결과를 생성해서 복수의 형태 분석 결과를 산출해낸 예시이다. 형태 분석은 단어 하나만 관찰하여 그 단어의 분석 결과를 규칙적으로 복원해 내는 것이기 때문에 당연히 복수의 분석 결과가 산출된다. 이 복수의 분석 결과에서 문맥에 따라 가장 적합한 단어를 선택하는 과정을 다음 절에서 설명하기로 한다.

3.2 태깅과 주석 코퍼스의 생산

≪만문노당≫ 형태 분석 결과로 산출된 결과물에는 복수의 분석 후보가 포함되어 있다. 즉 중의성이 포함되어 있다. 이러한 중의성을 제거하기 위해서 ≪만문노당≫이라는 텍스트의 특성에 맞게 복수의 분석 후보들에 대한 문맥 규칙을 작성하여 주석 말뭉치를 생산하였다. 문맥 규칙은 중의성을 가지는 단어의 좌우 어휘 문맥 의존 규칙 리스트를 작성하여서 복수 분석 가능한 것 중 하나를 선택하는 규칙으로 작성하며, 모두 68개의 규칙을 운영한다. 그 중 *tere*를 예시로 태그 선택 과정을 설명해 보겠다. *tere*는 "*tere*/nv::nv"와 "*te*/verb::vv + *re*/e::npst.ptcp" 두 가지로 분석이 가능하다. 이 중 두 가지 분석 중 하나를 선택하는 데 어떤 문맥이 결정적 요인으로 작용하는지를 규칙화하기 위해서 중의적 맥락을 자동으로 추출하고 그중에서 태그를 결정하는 결정 분포를 규칙 체계로 활용하였다.

〈표 6〉은 *tere*가 "*te* + *re*"로 분석되는 맥락을 ≪만문노당≫에서 추출하여 규칙으로 구성한 결과를 보여준다. 〈표 6〉에서 L과 R는 좌, 우 문맥에 출현하는 단어의 수를 가리킨다. 그리고 〈표 6〉에서는 좌우 출현하는 단어들의 맥락에서는 이때 '*tere*'를 '*te* + *re*'로 분석하여야 한다는 태그 선택 제약으로 사용된다. 선택 제약 규칙으로 적용되지 않는 환경에서는 기본값인 '*tere*/nv::nv'가 지정된다.

태그 선택 과정을 거친 결과는 수직 코퍼스(vertical corpus) 형태로 만들어지고, 다시 이 수직 코퍼스는 배포용 JSON(JavaScript Object Notation) 파일로 변환된다.

〈표 7〉은 수직 코퍼스 형식으로 만들어진 ≪만문노당≫ 형태 주석 코퍼스이며, 이 코퍼스는 연구용 자료로 배포하기 위해서 JSON 구조로 변환이 된다. JSON 파일의 구조는 〈그림 4〉와 같다. 〈그림 4〉에서 '+' 기호는 반복되어 출현할 수 있음을 나타내는 연산자(closure operator)이다. 〈표 8〉은 배포용으로 제작된 JSON 파일 중 일부를 추출한 예제이다.

3.3 ≪만문노당≫ 형태 주석 코퍼스 통계

이 연구에서 제작한 ≪만문노당≫ 태깅 시스템을 활용하여서 구축한 ≪만문노당≫

www.kci.go.kr

L	R	L2	L1	W	R1	R2	R3	R4	R5
0	3			tere	boo	usin	be		
0	3			tere	boo	,	tebure		
0	5			tere	boo	,	jetere	jeku	,
0	5			tere	boo	,	tarire	usin	,
1	1		bi	tere	seme				
1	1		sini	tere	ligiya				
1	2		acan	tere	,	jeku			
1	2		anafu	tere	,	gurun			
1	2		baitakv	tere	ba	arafi			
1	3		emgi	tere	,	emgi	yabure		
1	3		durifi	tere	,	muterakvci	,		
2	2	jeku	,	tere	boo	,			
2	2	han	i	tere	ilan	boo			
2	2	bade	bi	tere	,	sini			
2	2	genehebi	,	tere	boo	,			
2	2	jeku	,	tere	boo	,			
2	3	tuwame	,	tere	boo	,	jetere		
2	3	usin	,	tere	boo	be	getuken		
2	3	horho	,	tere	boo	,	tetun		

표 6. The co-occurrence distribution of the word '*te*/verb∷vv'

```
BOOK##!01
tongki fuka sindaha hergen i dangse .
    tongki          tongki/nv::nv
    fuka            fuka/nv::nv
    sindaha         sinda/verb::vv + ha/e::pst.ptcp
    hergen          hergen/nv::nv
    i               i/cm::gen
    dangse          dangse/nv::nv
    .               ./mark::period
```

표 7. Morphologically annotated vertical corpus of *Manwen laodang*

www.kci.go.kr

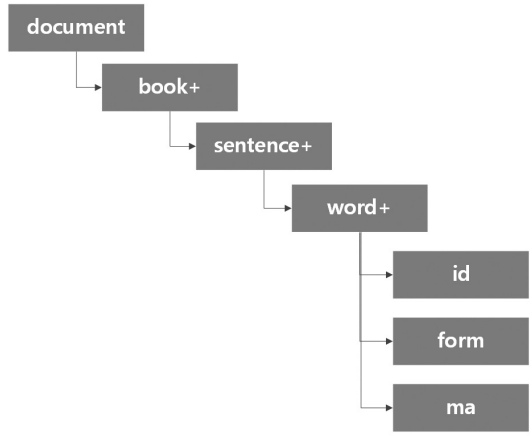


그림 4. JSON structure of *Manwen laodang* corpus

형태 주석 코퍼스에 사용된 형태소들의 분포를 정리하면 〈표 9〉와 같다.

형태 통계 정보는 두 가지 구획으로 구분하여 제시되어 있다. 상단 구획에는 중분류 태그까지 합산 빈도를 추출하여 제시하였다. 하단에는 문장부호를 제외한 형태소 빈도를 제시하였다. 하단 빈도를 관찰하면, verb(동사)와 e(어미)의 빈도가 거의 비슷하다. 반면 cm(격표지)의 빈도는 다른 nv(비동사) 계열 형태에 비해 빈도가 작다. 많은 비동사 계열 어휘들이 격표지 없이 사용되는 경우가 많다는 것을 알 수 있다. 〈표 10〉부터는 각 소분류 형태소의 빈도가 제시되어 있다.

≪만문노당≫ 형태 주석 코퍼스의 품사 태그는 크게 내용어와 기능어로 분류하였고, 문장부호와 미식별단어를 표시하는 기타 태그가 있다.

내용어는 verb(동사), pn(대명사), nn(고유명사), nv(비동사)로 세분화하였고, 기능어는 cm(격표지), e(어미), nmlz(명사화소), ptcl(첨사), sfx(접사)로 세분화였다. 내용어 중 중분류 verb는 vv(동사)와 vv.npst(비과거 동사)로 다시 분류하였다. 동사에는 *wa*, *unggi*, *aca* 등이 있고 비과거 동사에는 *bi*, *aina*, *bio* 등이 있다. 중분류 pn은 pronominal로서, 1pl.excl(배타적 1인칭 복수), 1pl.incl(포괄적 1인칭 복수), 1sg(1인칭 단수), 2pl(2인칭 복수), 2sg(2인칭 단수), 3pl(3인칭 복수), 3sg(3인칭 단수)로 구분하였다. ≪만문노당≫에서는 배타적 1인칭 복수 *be*, *men* 두 가지 종류가 출현하였고, 포괄적 1인칭 복수는 *muse*만이 나타났다. 2인칭 복수에는 *suwe*와 *suwen*, 3인칭 복수에는 *ce*와 *cen*이 나타났다. 한편 1인칭 단수는 *bi*와 *min*, 2인칭 단수는 *si*와 *sin*, 3인칭 단수는 *i*와 *in*이었다. 중분류 nn은 nominal noun으로서, prpn(일반 고유명사), prpn.person(인명), prpn.plc(장소명), prpn.

www.kci.go.kr

```
{
   "id": "KMPDJC2302100010",
   "document": [
        {
            "id": "MMNTMADOC0001",
            "book": [
                {
                    "id": "MMNTMADOC0001.b01",
                    "sentence": [
                        {
                            "id": "MMNTMADOC0001.b01.1",
                            "form": "tongki fuka sindaha hergen i dangse .",
                            "word": [
                                {
                                    "id": "1",
                                    "form": "tongki",
                                    "ma": "tongki/nv::nv"
                                },
                                {
                                    "id": "2",
                                    "form": "fuka",
                                    "ma": "fuka/nv::nv"
                                },
                                {
                                    "id": "3",
                                    "form": "sindaha",
                                    "ma": "sinda/verb::vv + ha/e::pst.ptcp"
                                },
                                {
                                    "id": "4",
                                    "form": "hergen",
                                    "ma": "hergen/nv::nv"
                                },
                                {
                                    "id": "5",
                                    "form": "i",
                                    "ma": "i/cm::gen"
                                },
                                {
                                    "id": "6",
                                    "form": "dangse",
                                    "ma": "dangse/nv::nv"
                                },
                                {
                                    "id": "7",
                                    "form": ".",
                                    "ma": "./mark::period"
                                }
                            ]
                        }
                    ],
                    ...이하 생략...
```


표 8. JSON example of *Manwen laodang* corpus

www.kci.go.kr

형태소 통계			
내용어	verb (verb)	40,521	16.04%
	pn (pronominal)	3,612	1.43%
	nn (nominal noun)	11,455	4.53%
	nv (non verbal)	83,207	32.94%
기능어	cm (case marker)	35,203	13.93%
	e (ending)	40,426	16.00%
	nmlz (nominalizer)	442	0.18%
	ptcl (particle)	1,266	0.50%
	sfx (suffix)	45	0.02%
문장부호	mark	36,468	14.43%
합		252,645	100%
문장부호를 제외한 형태소 통계			
내용어	verb (verb)	40,521	18.75%
	pn (pronominal)	3,612	1.67%
	nn (nominal noun)	11,455	5.30%
	nv (non verbal)	83,207	38.49%
기능어	cm (case marker)	35,203	16.28%
	e (ending)	40,426	18.70%
	nmlz (nominalizer)	442	0.20%
	ptcl (particle)	1,266	0.59%
	sfx (suffix)	45	0.02%
합		216,177	100%

표 9. Frequency of Manwen Laodang morphemes

gen(고유명사 소유격)으로 세분하였다. prpn은 일반적인 고유명사[8]로서 *hvrhan_hiya*, *sure_kundulen_han*, *bokdo_beile* 등이 나타났으며, prpn.person은 인명으로서 *bujantai*, *cangxu*, *nacibu* 등이 ≪만문노당≫에 출현하였다. prpn.plc는 장소명으로서 *warka*, *hesihe*, *hoifa* 등이 나타났으며, prpn.gen는 고유명사의 소유격 *i*가

8) 고유명사 중 복수어 단위(Multi-Word Unit)는 전처리 과정에서 밑줄(_)을 이용하여 복수어를 한 단위로 묶어 주었다.

www.kci.go.kr

대분류	중분류	소분류	빈도
내용어	verb (verb)	vv(verb)	40,242
		vv.npst(verb non-past)	279
		소계	40,521
	pn (pronominal)	1pl.excl(1st plural exclusive)	277
		1pl.incl(1st plural inclusive)	152
		1sg(1st singular)	1,235
		2pl(2nd plural)	767
		2sg(2nd singular)	713
		3pl(3rd plural)	55
		3sg(3rd singular)	413
		소계	3,612
	nn (nominal noun)	prpn(proper noun)	5,622
		prpn.person(proper noun, personal)	4,418
		prpn.plc(proper noun, place)	1,415
		소계	11,455
	nv (non verbal)	adj(adjective)	6,730
		nv(non verbal)	76,477
		소계	83,207

표 10. Frequency of Manwen Laodang content words

*baharai*에서 딱 한번만 나타났다. 중분류 nv는 동사가 아닌 것들을 묶어서 비동사(非動詞) 부류로 형태 처리를 하고 사전을 구성하였으며, adj(형용사)와 nv(비동사)으로 구분하였다. adj에는 *geren*, *amba*, *gelecuke* 등이 있고, nv에는 *tongki*, *fuka*, *hergen* 등이 있다.

기능어 중 중분류 cm은 abl(탈격), acc(대격), dat(여격), gen(소유격)으로 다시 분류하였다. 탈격은 *ci*, 대격은 *be*, 여격은 *de*, 소유격은 *i*와 *ni*가 ≪만문노당≫에 나타났다. 중분류 e에는 cve(부동사 어미), fve.imperative(명령형 정동사어미), fve.npst(비과거 정동사어미), fve.opt(원망형 정동사어미), fve.prs.prf(현재완료 정동사어미), fve.prs.prf.neg(부정형 현재완료 정동사어미), fve.prv(염려형 정동사어미), fve.pst.iptv(과거 비완망상 정동사어미), npst.ptcp(비과거 분사어미), npst.ptcp.neg(부정형 비과거 분사어미), pst.ptcp(과거 분사어미), pst.ptcp.neg(부정

www.kci.go.kr

대분류	중분류	소분류	빈도
기능어	cm (case marker)	abl(ablative)	1,168
		acc(accusative)	9,368
		dat(dative)	9,779
		gen(genitive)	14,888
		소계	35,203
	e (ending)	cve(converb ending)	15,502
		fve.imperative(finite verb ending, imperative)	1,938
		fve.npst(finite verb ending, non-past)	2,431
		fve.opt(finite verb ending, optative)	1,338
		fve.prs.prf(finite verb ending, present perfect)	554
		fve.prs.prf.neg(finite verb ending, present perfect negative)	44
		fve.prv(finite verb ending, preventive)	117
		fve.pst.iptv(finite verb ending, past imperfective)	103
		npst.ptcp(non-past participle)	4,220
		npst.ptcp.neg(non-past participle negative)	2,002
		pst.ptcp(past participle)	11,657
		pst.ptcp.neg(past participle negative)	520
		소계	40,426
	nmlz (nominalizer)	nmlz(nominalizer)	442
		소계	442
	ptcl (particle)	ptcl(particle)	1,266
		소계	1,266
	sfx (suffix)	sfx(suffix)	45
		소계	45

표 11. Frequency of *Manwen Laodang* function words

형 과거 분사어미)가 있다. 부동사 어미는 ***ci***, ***fi***, ***me*** 등 14가지가 존재한다. 명령형 정동사어미는 ∅이고, 비과거 정동사어미는 ***mbi***이며, 원망형 정동사어미는 ***ki***, ***kini***, ***cina***이다. 현재완료 정동사어미에는 ***hebi***, ***habi***, ***hobi***, ***kabi***, ***kebi***가 있으며, 부정형 현재완료 정동사어미에는 ***hakvbi***, ***hekvbi***, ***kekvbi***가 있다. 염려형 정동사어미

www.kci.go.kr

대분류	중분류	소분류	빈도
문장 부호	mark	circle	1,851
		comma	32,285
		lparen(left parenthesis)	27
		rparen(right parenthesis)	26
		period	2,275
		unknown	4
		소계	36,468

표 12. Frequency of *Manwen laodang* marks

는 *rahv*, 과거 비완망상 정동사어미는 *mbihe*이다. 비과거 분사어미에는 *ra*, *re*, *ro*가 있으며, 부정형 비과거 분사어미에는 *rakv*가 있다. 또한 과거 분사에는 *ha*, *he*, *ho*, *ka*, *ke*, *ko*가 있으며, 부정형 과거 분사어미에는 *hakv*, *hekv*, *kakv*, *kekv*가 있다. 중분류 nmlz, ptcl, sfx는 각각 nmlz(명사화소), ptcl(첨사), sfx(접사)로 재분류하였는데, nmlz에는 *ngge*가 존재하며, ptcl에는 *kai*, *dere*, *o*, *vn*, *ni*, *nikai*, *dabala*가, sfx에는 *bu*와 *mbu*가 존재한다.

기타 태그로는 marker(문장부호)가 있다. 문장부호는 다시 circle(원), comma(쉼표), lparen(왼쪽 괄호), rparen(오른쪽 괄호), period(점)으로 구분하였으며, □□□와 ■■■는 원문 미식별어를 나타내는 표기여서 모두 unknown으로 구분하였다.

4. 마치며

지금까지 김주원 외(2019)에 수록된 ≪만문노당≫ 태조편 자료의 만주어 원문을 대상으로 형태 주석 코퍼스를 구축하는 과정과 빈도 분석 결과를 간단하게 설명하였다. 요즘의 코퍼스 언어학과 디지털 인문학(digital humanities) 등의 자료 중심 언어 연구에는 현재 유행하는 다양한 기계학습(machine learning), 심층학습(deep learning) 기법들을 활용하고 텍스트 마이닝의 다양한 방법론을 적용하여 연구의 수준을 높여가고 있다. 그렇지만 알타이언어 자료에서는 주석 체계 등을 갖춘 정련된 코퍼스를 찾아보기가 쉽지 않다. "구슬이 서 말이라도 꿰어야 보배"라는 말은 자료를 다루는 연구 분야에서는 언제나 진리인 것으로 보인다. 특히 만주어 자료는 많은 문헌 기록들이 남아 있어서, 언어 연구, 청대 역사 연구, 문명 교류 등의

www.kci.go.kr

연구를 위한 귀한 자료임에도 디지털화 과정이 더디게 진행되고 있다. 그리고 검색(retrieval), 텍스트 마이닝(text mining) 등을 위한 만주어 형태 주석 코퍼스는 거의 존재하지 않았다.

이 연구를 시발점으로 삼아서 다양한 만주어 자료의 주석 코퍼스 구축이 계획되어 있다. ≪만문노당≫ '태조편'의 주석 코퍼스 작업이 이번 연구에서 이루어졌고, 이후 '태종편'을 비롯한 만주어 원문이 준비되는 대로 ≪만문노당≫ 형태 주석 코퍼스 구축이 지속적으로 이루어지고, 또 배포될 예정이다. 이번 연구의 결과로 구축된 ≪만문노당≫ 태조편 형태 주석 코퍼스는 GitHub를 통해서 공개될 예정이다.[9] 그리고 ≪만문노당≫ 형태 주석 코퍼스는 앞으로 기계 학습 방식에 기반한 태거를 새로 제작하는 데 훈련용 코퍼스로 다시 사용될 계획이다. 만주학, 그리고 알타이학에서도 다양한 기술들이 적용되기를 바라며, 이번 연구의 결과로 공개되는 코퍼스가 자료 중심 만주어 연구의 기초 역할을 하기를 바란다.

References

Jockers, M. L. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Kang, Beommo 강범모. 2008. Ene kiswulul wihan khophesuuy kwuchwukkwa pinto (thongkyey) hwalyong 언어 기술을 위한 코퍼스의 구축과 빈도(통계) 활용 [Building Corpora and Making Use of Frequency (Statistics) for Linguistic Descriptions]. *Journal of Korealex* 12. 7-40. The Korean Association For Lexicography (Korealex).

Khakimov, B., Salimov, F., & Ramazanova, D. 2015. Building dialectological corpora for Turkic languages: Mishar dialect of Tatar. *Procedia – Social and Behavioral Sciences* 198. 218-225. Amsterdam: Elsevier.

Kim, Doohyun 김두현. 2008. Manmwunwentangkwa ipkwancen chengsayenkwu ≪滿文原檔≫과 入關前 淸史硏究 ["*Man-wen yuan-dang*" and the History of Early Qing]. *Journal of Ming-Qing Historical Studies* 30. 259-286. The Society For Ming-Qing Historical Studies.

Kim, Juwon et al. (trans.) 김주원 외 역주. 2019. *Manmwunnotang: thayco 1, 2* 만문노당: 태조 1, 2 [The chronicles of early Qing dynasty: Taidzu 1, 2]. Seoul: SNU Press.

Kim, Juwon 김주원. 2008. Althaieneuy saylowun yenkwu panghyangey tayhaye 알

9) https://github.com/Kkamakpyel/manwenlaodang 에서 JSON 형식의 코퍼스가 공개되어 있다.

www.kci.go.kr

타이언어의 새로운 연구 방향에 대하여 [A New Direction of Research on Altaic Languages]. *Hangeul* 282. 343-367. Hangeul Hakhoe (The Korean Language Society).

Lee, Taehun & Yookyung Jeong (eds.) 이태훈·정유경 엮음. 2020. *Ticithel inmwunhakkwa kuntayhankwukhak: ticithelɯy nwunulo hankwukhakul ilkta* 디지털 인문학과 근대한국학: 디지털의 눈으로 한국학을 읽다 [Digital humanities and morden Korean studies: Methods of Korean studies in the digital age]. Seoul: Somyong Publishing.

Norman, Jerry. 2013. *A Comprehensive Manchu-English Dictionary*. Cambridge and London: Havard University.

Seong, Baegin 성백인. 1978. Manmwun mwukwencem sipicatwu yenkwu 만문 무권점 십이자두 연구 [A study of the 12 initials without dots and circles in the Manchu script]. *Language* 말 3. 97-134. Seoul: Yonsei University Korean Language Institute.

Yan, Chongnian 閻崇年. 1998. Lun *Manzhou laodang* 論 ≪滿洲老檔≫ [Discussion of *Manzhou laodang*]. *Manchu studies* 滿學研究 4. 1-84. Manchu institute 滿學研究所 in Beijing academy of social sciences.

You, Hyun-Jo. 2014. A Manchu speller: With a practical introduction to the natural language processing of minority languages. *Altai Hakpo* 24. 39-67. The Altaic Society of Korea.

Choi Woonho
Department of Korean Language and Literature
Mokpo National University
1666 Yeongsan-ro, Cheonggye-myeon, Muan-gun, Jeollanam-do 58554, KOREA
<whchoi@mnu.ac.kr>

Jung Sunghoon
Department of Korean Language and Literature
Mokpo National University
1666 Yeongsan-ro, Cheonggye-myeon, Muan-gun, Jeollanam-do 58554, KOREA
<kobe99@mnu.ac.kr>

Do Jeongup
Department of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, KOREA
<dju1st@naver.com>

www.kci.go.kr

Received 31 March 2023;
revision received 10 June 2023;
accepted 10 June 2023.

www.kci.go.kr

www.kci.go.kr

View publication stats