국립국어원 2018-01-41

발간등록번호
11-1371028-000738-01

2018년 국어 기초 어휘 선정 및 어휘 등급화 연구

연구책임자: 이삼형

국립국어원

제 출 문

국립국어원장 귀하

“2018년 국어 기초 어휘 선정 및 어휘 등급화 연구”에 관하여 귀 원과 체결한 연구용역 계약에 의하여 연구보고서를 작성하여 제출합니다.

2018년 12월 22일

연구 책임자: 이삼형(한양대학교)

연구 기관 한양대학교 산학협력단

연구 책임자 이삼형

공동 연구원 박진호, 최형용, 김정선, 이승연,
이현주, 신명선, 이기연, 김시정

연구 보조원 허인영, 김혜지, 김수지, 이윤희

보 조 원 양세문

요약문

1. 과업명: 2018년 국어 기초 어휘 선정 및 어휘 등급화 연구

2. 과업의 목적

본 연구의 목적은 국어의 기초 어휘를 선정 및 등급화 방법론을 구축하기 위하여 기초 어휘의 구축 및 활용과 관련된 해외 사례를 조사하고 등급화를 위한 방법론을 수립하며 기초 어휘를 추출하기 위한 목적의 언어 자료를 정련화하는 데 있다.

3. 과업의 배경

"2018년 국어 기초 어휘 선정 및 어휘 등급화 연구"의 필요성은 다음과 같다.

첫째, 2017년 "국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구"에서 구축한 언어 자료 보완의 필요성

둘째, 어휘 등급화에 관한 통계적, 정성적 방법론의 체계화 및 실증의 필요성

셋째, 연구 사업의 추진 과정의 쟁점에 대한 확인 및 해결 방안 모색의 필요성

4. 연구 내용 및 방법

본 연구의 내용은 다음과 같이 요약된다.

1) 기초 어휘 선정 및 어휘 등급화를 위한 사례 조사
   ○ 기초 어휘 선정 및 어휘 등급화 국내외 사례 연구
   ○ 교과서 어휘 연구 사례 조사 및 기초 어휘와의 상관성 검토
2) 기초 어휘 추출 목적의 언어 자료 정제와 처리 과정
   ○ 언어 자료의 양적·질적 보완 및 정제를 통한 통계적 신뢰성 제고
   ○ 형태소 분석 모델 개선을 통한 기초 어휘 추출의 정확도 제고
3) 어휘 등급화의 통계적 방법론 수립
   ○ 직관 판단 실험을 통한 빈도, 범위, 산포도에 기반한 어휘 점수 산출
   ○ MANULEX의 U값에 따른 기초 어휘 순위와의 비교·분석
4) 어휘 등급화의 정성적 방법론 수립
   ○ 예비 어휘 목록에 대한 세부 쟁점별 전문가 검토
   ○ 초등 교과서 어휘 목록과 예비 어휘 목록의 비교

본 연구의 방법은 다음과 같이 요약된다.

1) 문헌 연구

국민의 어휘 능력과 어휘 평정에 관한 기초 어휘 선정 및 등급화 기준 설정의 토대를 마련한다.

2) 사례 연구

국내외 기초 어휘 선정 및 어휘 등급화의 사례를 수집하여 언어 자료 구축에서부터 활용 체계에 이르는 시사점을 얻는 데 활용한다.

3) 언어 자료 분석

예비 어휘 목록을 마련하고 이를 분석하여 어휘 등급화의 정성적 방법론 수립에 활용한다.

4) 실험

개별 어휘의 중요도와 어휘 목록의 타당성 검증을 위해 한국어 모어 화자의 직관 판단 실험을 시행한다.

5. 연구 결과

본 연구의 결과를 정리하면 다음과 같다.

첫째, 기초 어휘 선정 및 어휘 등급화의 국내 사례 조사를 통해 기초 어휘 선정 및 어휘 등급화 방향을 수립하였다. 국외 사례의 경우 2017년 연구의 후속 조사로써 프랑스어와 중국어 사례를 통해 기초 어휘 선정 및 어휘 등급화에 대한 시사점을 얻었다. 특히 프랑스어 사례 조사를 통해 본 연구의 어휘 통계 방식과의 비교에 활용하였다. 또한, 교과서 어휘 목록 사례를 추가로 조사하여 기초 어휘 목록과 교과서 어휘 목록 사이의 상관성을 확인하였다.

둘째, 언어 자료를 양적으로 보완하여 총 45억 어절의 언어 자료를 수집하였다. 이에 대해 줄 바꿈 문자의 처리, 어문 규범 위반 사례의 처리, 장르 체계화 등 통계 수치의 신뢰성을 제고할 수 있는 방향으로 정제하였다. 이들 언어 자료로부터 기초 어휘를 추출하기 위한 형태소 분석 모델을 개선하고 후처리 모듈을 개발하여 기초 어휘의 정확도를 제고하였다.

셋째, 언어 자료를 기반으로 하여 기초 어휘를 추출할 때 활용하는 통계적 방법론을 수립하고 이를 정교화하였다. 어휘 형태소의 상대빈도, 범위, 산포도에 가중치를 부여한 어휘 점수를 산출하였으며, 한국어 모어 화자에 대한 직관 판단 실험을 실시하여 이를 정교화하였다.

넷째, 어휘 자료에 대한 전문가의 정성적 검토를 통해 어휘 등급화의 정성적 방법론 수립 시 고려해야 할 쟁점을 추출하였다. 이들 중 동형어 처리 및 접사 처리에 대한 전문가 검토를 통해 어휘 등급화 방향을 제시하였다. 또한, 초등 교과서 어휘 목록과의 비교를 통해 예비 어휘 목록을 검토함으로써 개선 방향을 제시하였다.

6. 결과물: 최종 보고서 50부, 시디 10매, 대용량 저장장치 1개

Abstract

1. Title of Task: A 2018 Study on the Selection and Grading of Basic Vocabulary in Korean

2. Goals of Task

The purposes of this study were to examine overseas cases related to the establishment and utilization of basic vocabulary to build a methodology of selecting and grading basic vocabulary in Korean, establish a methodology for grading, and refine language materials to extract basic vocabulary.

3. Backgrounds of Task

There was a need for "A 2018 Study on the Selection and Grading of Basic Vocabulary in Korean" in the following aspects:

First, there was a need to supplement the language materials accumulated in "A Basic Study on the Selection and Grading of Basic Vocabulary in Korean" in 2017.

Second, there was a need to systemize and demonstrate statistical and qualitative methodologies for vocabulary grading.

Third, there was a need to check issues with the process of pushing forward a research project and search for solutions for them.

4. Content and Methods of Research

The content of the study can be summarized as follows:

1) Case studies to select basic vocabulary and grade vocabulary

○ Case studies home and broad for the selection of basic vocabulary and the grading of vocabulary

○ Case studies on vocabulary in textbooks and review of its correlations with basic vocabulary

2) Process of refining and treating language materials to extract basic vocabulary

○ Higher statistical reliability through the supplementation and refinement of language materials both in quantity and quality

○ Higher accuracy of extracting basic vocabulary through the improvement of a morpheme analysis model

3) Statistical methodologies for the grading of vocabulary

○ Obtaining vocabulary scores based on frequency, scope and dispersion in an intuitive judgment experiment

○ Comparison and analysis with the ranking of basic vocabulary according to the U value of MANULEX

4) Qualitative methodologies for the grading of vocabulary

○ Expert review for issues with a preliminary vocabulary list in details

○ Comparison between the vocabulary list for elementary school textbooks and a preliminary vocabulary list

The methodologies can be summarized as follows:

1) Literature study

The study would build a foundation for the criteria of selecting and grading basic vocabulary with regard to the vocabulary abilities and rating of people.

2) Case study

The study would collect cases of selecting basic vocabulary and grading vocabulary home and abroad and use them to obtain implications for the establishment of language materials and their utilization system.

3) Analysis of language materials

The study would make and analyze a preliminary vocabulary list and use it to establish a qualitative methodology for vocabulary grading.

4) Experimentation

The study would conduct an intuitive judgment experiment with native Korean speakers to test the importance of individual vocabulary and the validity of vocabulary lists.

5. Findings

The findings were arranged as follows:

First, the study examined cases of selecting basic vocabulary and grading vocabulary in the nation, thus establishing a direction for them. For overseas cases, the study examined French and Chinese cases as part of follow-up study after the 2017 study, having implications for selecting basic vocabulary and grading vocabulary. The French case was especially compared with the method of vocabulary statistics introduced in the present study and utilized further. In addition, case lists of textbook vocabulary were examined to check correlations between the list of basic vocabulary and that of textbook vocabulary.

Secondly, the study supplemented language materials in quantity, collecting the language materials of total 4.5 billion syntactic words. They were then refined in a direction of increasing the reliability of statistical numbers such as the treatment of line-breaking characters, processing of violations of language norms, and genre systemization. A morpheme analysis model was improved to extract basic vocabulary from these language materials, and a post-treatment module was developed to increase the accuracy of basic vocabulary.

Thirdly, the study established and elaborated a statistical methodology used to extract basic vocabulary based on language materials. Weights were granted to the relative frequency, scope and dispersion of vocabulary morphemes to calculate vocabulary scores. They were then elaborated with an intuitive judgment experiment with native Korean speakers.

Finally, the study identified issues that should be taken into consideration to establish a qualitative methodology for vocabulary grading through a qualitative review of vocabulary materials by experts. A direction for vocabulary grading was proposed through the expert review of homonym and affix treatment. In addition, a preliminary vocabulary list was compared with the list of vocabulary in elementary school textbooks to propose a direction for its improvement.

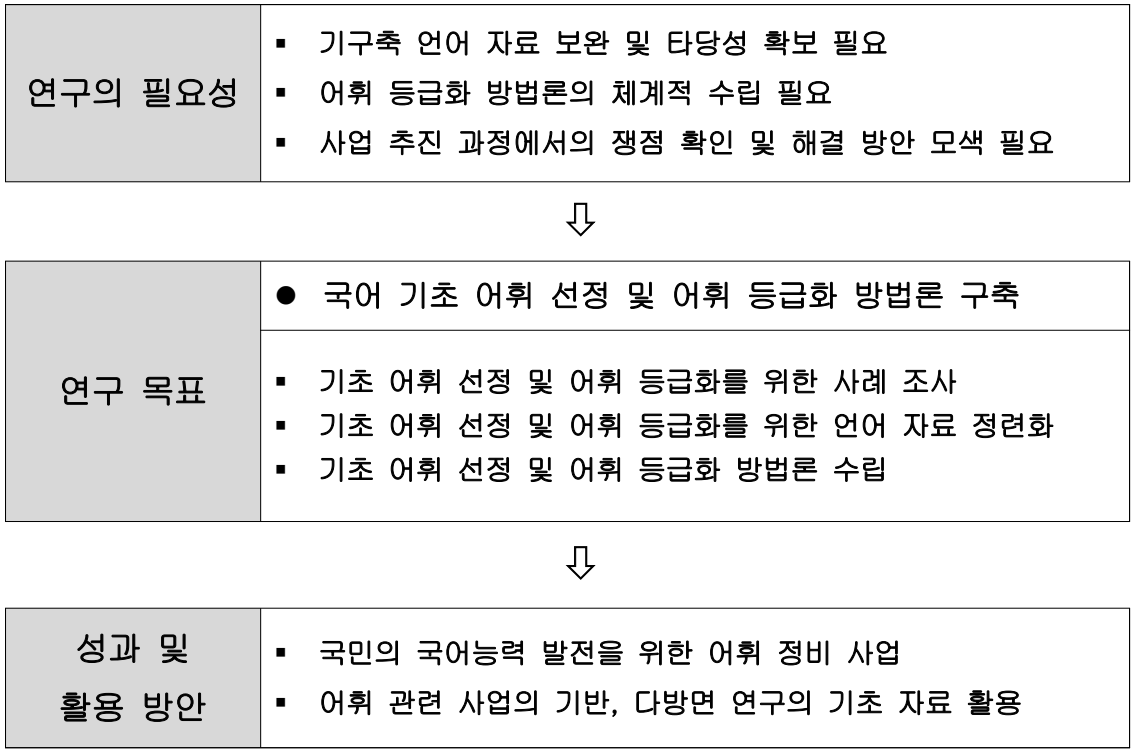
6. Outcomes: 50 copies of final report, ten CDs, and a large-capacity storage device.

목 차

<표 차례>

<그림 차례>

Ⅰ. 서론



[그림 1] 연구 개요

1. 연구의 목적과 필요성

1.1. 연구의 목적

본 연구의 목적은 국어의 기초 어휘를 선정하고 등급화하기 위한 방법론을 구축하기 위하여 기초 어휘의 구축 및 활용과 관련된 해외 사례를 조사하고 등급화를 위한 방법론을 수립하며 기초 어휘를 추출하기 위한 목적의 언어 자료를 정련화하는 데 있다.

1.2. 연구의 필요성

“2018년 국어 기초 어휘 선정 및 어휘 등급화 연구”는 국민 전체를 대상으로 한 어휘 정비 사업에 대한 요구, 국가고사, 공문서 정비 등 다양한 목적으로 널리 활용

될 수 있는 어휘 목록 구축에 대한 요구 등을 충족시키는 사업으로서 기획, 추진되고 있다. 이 가운데 당해 연도에 시행되는 연구의 필요성을 살펴보면 다음과 같다.

첫째, 2017년에 수행되었던 “국어 기초 어휘 선정 및 어휘 등급화를 위한 기초연구”에서 구축한 언어 자료를 보완하기 위함이다. 이는 결과적으로 기초 어휘 목록의 타당성을 높이기 위한 작업이다. 2017년에 구축한 언어 자료는 매우 방대한 분량이지만 장르, 시기의 편중성 등의 보완 사항을 지니고 있었다. 이에 따라 언어 자료를 보완하여 정련화함으로써 언어 자료 자체의 표준성을 높임은 물론이고 연구 결과의 타당성을 높여야 할 필요가 있다.

둘째, 어휘 등급화의 방법론을 체계적으로 수립하기 위함이다. 전년도에 수행된 사업에서는 기초 어휘 목록의 추출 및 등급화를 위한 이론적 기반을 마련하고 이를 이룰 수 있는 방향을 탐색하였다면, 당해 연도의 작업은 실제 기초 어휘를 추출하는 통계적, 정성적 방법론을 체계화하고 실증하는 데 초점을 맞춘다. 이를 위하여 실제 언어 자료에 대한 통계적, 정성적 검토에 더하여 교과서 어휘와의 비교 연구, 직관 실험 등 다양한 측면의 접근을 시도할 필요가 있다.

셋째, 연구 사업의 추진 과정에서 해결하여야 할 쟁점 등을 확인하고 이를 해결할 수 있는 방안을 모색할 필요가 있다. 이는 기초 연구 단계에서 이루어졌던 전반적 고찰을 심화하여, 언어 자료를 기반으로 기초 어휘 목록을 추출할 때 고려해야 할 사항 등을 확인하고, 그것의 활용을 중심으로 최종 연구 성과물이 제공해야 할 정보, 자료의 성격 등을 명확화하는 것을 의미한다. 또한 향후 연구의 수행에서 해결해야 하는 구체적 과제들을 안내함으로써 이후의 연구를 예비할 필요가 있다.

2. 연구의 내용과 방법

2.1. 연구 내용

연구 내용	1) 기초 어휘 및 어휘 등급화를 위한 사례 조사 2) 기초 어휘 추출 목적의 언어 자료 정제와 처리 과정 3) 어휘 등급화의 통계적 방법론 수립 4) 어휘 등급화의 정성적 방법론 수립

[그림 2] 연구 내용

1) 기초 어휘 및 어휘 등급화를 위한 사례 조사

(1) 국내 사례

2017년 연구에 이어 국내 사례 조사는 기초 어휘에 대한 연구와 국어교육용 · 한국어교육용 기초 어휘 목록, 등급화에 대한 사례 조사를 실시한다. 아울러 기초 어휘의 검증 기준으로 활용 가능한 교과서 어휘에 대한 사례 조사도 실시한다. 조사의 주요 내용은 다음과 같다.

○ 기초 어휘 연구 사례
- 기초 어휘의 개념과 특성
- 기초 어휘 선정의 필요성

○ 기초 어휘 목록 및 어휘 등급화 사례
- 국어교육용 어휘 목록 및 등급화
- 한국어교육용 어휘 목록 및 등급화

○ 교과서 어휘 사례
- 교과서 어휘 정책 연구 성과

(2) 국외 사례

국외 사례 조사는 2017년도에 조사한 영어, 일본어 기초 어휘 사례 조사를 확장하여 프랑스어, 중국어의 사례를 조사한다. 조사 내용은 다음과 같다.

○ 프랑스어 사례
- 기초 어휘 목록
- 어휘 등급화 관련 선행 연구 조사

○ 중국어 사례
- 기초 어휘, 상용 어휘, 일반 어휘의 구분
- 상용 어휘 등급화

2) 기초 어휘 추출 목적의 언어 자료 정제와 처리 과정

(1) 기초 어휘 추출 목적의 언어 자료 정제

전년도에 구축한 언어 자료에 금년도에 새로 수집한 언어 자료를 합치면 약 45억 어절을 상회한다. 구축한 언어 자료의 정확한 통계 처리를 위하여 보완 작업을 수행한다.

○ 기초 어휘 추출 목적의 언어 자료 구축 현황
- 1차 연도에 구축한 언어 자료: 약 29억 어절
- 2차 연도에 구축한 언어 자료: 약 16억 어절
- 총 구축한 언어 자료: 약 45억 어절

○ 기초 어휘 추출 목적의 언어 자료 보완
- 줄 바꿈 문자 처리
- 어문 규범 위반 사례 처리
- 신문 장르 세분을 위한 실험

(2) 언어 자료 처리 과정

언어 자료의 처리 과정은 크게 형태소 분석 이전 단계, 형태소 분석 단계. 형태소 분석 이후 단계로 구분하여 시행한다.

○ 형태소 분석 이전 단계
- 줄 바꿈 문자 처리
- 어문 규범 위반 사례 처리
- 장르 체계 수정/정비 및 이에 따른 파일 및 디렉토리 구조 정비

○ 형태소 분석 단계
- UTagger의 문제점과 한계 검토
- 형태소 분석 처리를 위한 훈련 데이터 구축
- 분절, 품사 부착, 동형어 구분 모델 만들기
- 모델의 오류 모니터링 및 최적화
- 추후 작업 계획

○ 형태소 분석 이후 단계
- 형태소 분절에 대한 후처리
- 형태소에 부여된 품사에 대한 후처리

3) 어휘 등급화의 통계적 방법론 수립

어휘 등급화를 위한 통계적 방법론을 수립하여, 언어 자료의 통계적 분석을 수행한다. 우선, 총 언어 자료를 대상으로 형태소 분석 처리를 시행한 후, 장르를 고려하여 각 어휘 형태소의 상대빈도, 범위, 산포도를 산출한다. 다음으로, 어휘 점수 산출을 위한 지표 개발 및 정교화 작업을 수행한다. 그리고 빈도, 범위, 산포도와 가중치를 바탕으로 어휘 점수를 산출하고, 이 점수에 따라 단어들을 배열하여 순위에 따른 어휘 목록을 추출한다. 본 연구에서는 어휘 점수의 객관성과 타당도를 확보하기 위하여 가중치 산출을 위한 실험을 수행한다.

4) 어휘 등급화의 정성적 방법론 수립

(1) 정성적 분석의 검토 항목

통계적 분석으로는 개별 어휘들의 특성과 어휘 간의 관계에 대한 정교한 분석이 어렵기 때문에 이들에 대한 정성적 분석을 수행한다. 정성적 분석을 위해서는 검토 사항을 정리한 후, 이에 대한 선행 연구를 검토하고 주요 검토 사항별 구체적인 어휘를 가지고 판단한다.

○ 주요 검토 사항 정리
- 동형어 처리: 품사 통용어, 본용언과 보조용언의 처리, 어근 어휘와 파생어 어휘의 처리, 상위 품사와 하위 품사의 처리
- 기초 어휘의 단위: 조사와 어미의 포함 여부, 접사의 포함 여부, 고유 명사의 포함 여부, 감탄사의 포함 여부

○ 기존 연구 검토
- 국어교육과 한국어교육에서의 주요 검토 사항별 처리 검토

(2) 동형어 처리와 접사 처리

주요 검토 사항 가운데 금년도에는 동형어 처리와 접사 처리를 50,000위까지의 어휘를 대상으로 검토하고, 이들 어휘 처리에 대한 방향을 모색한다. 특히 접사는 생산성이 높은 접사를 대상으로 이들을 기초 어휘로 선정할 때 발생할 수 있는 문제에 대해 살펴본다.

○ 동형어 처리 검토: 품사 통용어
- 체언 기준 동형어 검토
- 용언 기준 동형어 검토
- 수식언 기준 동형어 검토

○ 접사 처리 검토
- 접두사 '미-, 불-, 비-' 검토
- 접미사 '-적, -성, -화' 검토
- 접미사 '-스럽-, -롭-, -답-' 검토

(3) 교과서 어휘 목록을 통한 검토

초등학교 교과서에서 추출한 어휘 목록을 활용하여 본 연구에서 수집한 언어 자료 기반의 기초 어휘 예비 목록을 비교 검토한다.

2.2. 연구 방법

1) 문헌 연구

기초 어휘 선정과 등급화의 기준 설정의 토대를 마련하기 위해 문헌 연구를 실시한다. 문헌 연구의 주요 내용을 간추리면 다음과 같다.

○ 언어학 · 국어학: 기초 어휘의 개념, 어휘 분석의 단위
○ 언어 교육: 기초 어휘 선정 방법, 어휘 선정의 기준, 어휘 등급화 방법
○ 국어정보학: 언어 자료 구축 방법론, 형태소 분석, 어휘 추출 방법

2) 사례 조사 연구

본 연구에서는 국내외의 기초 어휘에 대한 다양한 사례를 조사하고 분석함으로써 언어 자료 구축에서부터 활용 체계에 이르는 시사점을 얻는 데 활용한다.

○ 국어교육과 한국어교육에서의 어휘 목록
○ 프랑스어의 기본 어휘 목록과 등급화 방안
○ 중국어의 기본 어휘와 상용 어휘 목록

3) 언어 자료 분석

본 연구는 기초 어휘 추출 목적으로 수집·구축한 언어 자료를 빈도, 범위, 산포도 등 통계적 기준을 활용하여 예비 어휘 목록을 마련한다. 예비 어휘 목록을 다수의 전문가가 체계적으로 분석하여 어휘 등급화의 정성적 방법론 수립에 활용한다.

4) 실험

본 연구에서는 개별 어휘의 중요도나 어휘 목록의 타당성을 검증하기 위하여 고도의 직관을 가진 실험자 집단을 섭외하여 어휘 목록 및 등급화에 관한 직관 판단 실험을 시행한다. 실험의 개요는 다음과 같다.

○ 실험 목적: 모국어 화자의 직관을 고려한 가중치 결정
○ 실험 대상: 서울 소재 대학교 한국인 재학생 146명
○ 실험 기간: 2018.11.19.~2018.11.24.
○ 실험 내용: 1인당 50개 단어 목록 2개를 받아 직관에 따라 기초 순위로 배열

3. 연구 추진 과정

3.1. 연구 추진 일정

	과업의 내용	세부 연구 내용	월						
			6	7	8	9	10	11	12
1	기초 어휘 사례 연구	어휘 등급화 관련 선행 연구 검토	○						
		해외 사례 조사	○	○					
2	기초 어휘 목록 추출 및 등급화 방법론 수립	기초 어휘 예비 목록 추출 및 공유	○						
		어휘 등급의 통계적 방법 체계화 (어휘 등급화를 위한 지표 정교화)	○	○					
		어휘 등급의 정성적 방법 수립 (예비 목록 검토를 통한 등급화 쟁점 도출)	○	○	○				
3	어휘 예비 목록 검토 및 어휘 등급화	기초 어휘 예비 목록 검토			○	○	○		
		기초 어휘 예비 목록의 등급화 및 검토				○	○		
4	교과서 어휘 목록 연구	교과서 자료 및 어휘 목록 확보	○	○	○	○			
		기초 어휘 판단의 적절성 검증		○	○	○			
		기초 어휘 목록과 교과서 어휘 목록의 비교				○	○		
5	언어 자료 정제	언어 자료 규범상 오류 정제	○	○	○				
		형태소 분석기 성능 개선			○	○			
		정성적 검토에 따른 언어 자료 정제				○	○	○	
6	종합	결과보고서 작성						○	○

3.2. 주요 협의회 내용

본 연구진은 매달 1회의 전체 회의와 다수의 분과별 회의를 정기적으로 진행하였으며, 총 3회의 보고회를 개최하였다. 보고회 회의 내용을 일자별로 정리하여 제시하면 다음과 같다.

1) 착수 보고회

- 날짜: 2018년 6월 25일(월)
- 장소: 국립국어원
- 참여자: 국립국어원 관계자, 연구진, 자문위원단

○ 주요 내용 및 결과

- 연구 용역 사업 추진 계획 보고, 전체 사업 진행 방향 및 일정 점검.
- 등급화 방법론의 질적 검토에 있어 신뢰성 확보가 중요함.
- 언어 자료의 성격과 어휘 목록의 연관성을 고려하여 언어 자료의 정제 작업이 진행되어야 함.
- 기초 어휘 선정 시 어휘 단위, 어휘 범위, 어종, 조어법 상의 특성, 부적절한 표현의 처리, 본말과 준말의 문제 등을 고려할 필요가 있음.
- 업무 분장에 있어 각 분과 간의 상호 검토 체제가 잘 되어 있음.

2) 중간 보고회

- 날짜: 2018년 9월 19일(수)
- 장소: 국립국어원
- 참여자: 국립국어원 관계자, 연구진, 자문위원단

○ 주요 내용 및 결과

- 연구 용역 사업 진행 사항 보고, 이후 사업 진행 방향 및 일정 점검.
- 충분한 인원의 실험자를 대상으로 직관 실험을 시행함으로써 모국어 화자의 직관에 잘 맞는 가중치를 산출할 수 있을 것으로 예상함.
- 정성적 방법의 경우 고려할 것이 많아 올해 모든 결정을 내리기는 어려움. 충분한 시간을 들이고 다각도로 고려하여 적정한 방법을 찾아야 할 것임.
- 예비 어휘 목록이 도출되면 비교 대상으로 교과서 어휘 목록을 활용할 예정임.
- 교육과정평가원에서 진행 중인 교과서 어휘 관련 연구가 완결되면 본 연구에 도움이 될 것으로 생각됨.

3) 최종 보고회

- 날짜: 2018년 12월 19일(수)
- 장소: 국립국어원
- 참여자: 국립국어원 관계자, 연구진, 자문위원단

○ 주요 내용 및 결과

- 연구 용역 사업 결과 보고.
- 통계적 방법론의 경우, 작년 연구에 비하여 가중치 산출이 세밀화되었음.
- 정성적 방법론의 경우, 동형어 처리와 접사 처리 등 일부 쟁점 사항에 대한 방법론을 제시하였음.
- 후속 연구에서 정성적 방법론의 쟁점 사항, 전문어의 문제 등 어휘 목록 추출을 위해 결정해야 할 사항이 많음.

Ⅱ. 기초 어휘 및 어휘 등급화를 위한 사례 조사

1. 국내 사례

1.1. 기초 어휘 및 어휘 등급화 사례

지금까지 행해진 기초 어휘에 관한 연구로는 기초 어휘의 개념과 기초 어휘 선정의 중요성을 밝힌 연구, 기초 어휘 선정 방법을 밝히고 기초 어휘 목록을 제시한 연구, 기초 어휘 평정을 통해 등급을 구분한 연구 등이 있다.

1.1.1. 기초 어휘 연구 사례

기초 어휘란 일반적으로 의사소통에서 가장 기본적이고 핵심적인 어휘로, 그 언어에서 상대적으로 중요도가 높은 어휘를 말한다. 기초 어휘에 속하는 단어는 일상 언어생활에서 널리 쓰이며, 다른 단어로 대체하기 어렵다는 특성이 있다.

국내의 기초 어휘 관련 연구의 시작은 문교부에서 1956년에 수행한 「우리말 말수 사용의 잦기 조사」이며, 이후로 꾸준히 국어 기초 어휘의 중요성이 논의되어 왔다. 그러나 그 중요성이 강조되어 온 것에 비하여 실제 연구가 활발하게 이루어지지는 않았다. 기초 어휘의 개념과 특성에 관한 연구는 1990년대 이후 시기에 집중되어 있다. 기초 어휘의 개념에 관한 연구는 기초 어휘, 기본 어휘, 기간 어휘 등 유사 개념이 혼재한 상태에서 이루어졌다. 이삼형 외(2017a)에서는 기초 어휘, 기본 어휘에 관한 기존 성과를 요약하며 학계에서 기초 어휘와 기본 어휘가 혼용되어 쓰여왔음을 지적한 바 있다. 기초 어휘와 기본 어휘의 개념을 구분하여 쓰는 연구(임지룡, 1991; 김종학, 2001 등), 기초 어휘와 기본 어휘를 구분하지 않고 혼용하여 쓰는 연구(김광해, 1988; 이충우, 1994a 등), 하나의 개념으로 통일하여 쓰는 연구(성광수, 1999)가 모두 있었다. 기초 어휘의 개념과 특성에 관한 주된 선행 연구를 살펴보면 다음과 같다.

이충우(1990)는 기본 어휘란 대개 일상생활을 하는 데 불편 없이 언어를 구사할 수 있는, 상당수의 어휘 가운데 정상적인 기본 생활을 하는 데 필요하다고 간주되는 어휘를 말하며, 대개 2,000~3,000 단어 정도 된다고 하였다.

임지룡(1991)은 기초 어휘, 기본 어휘, 기간 어휘의 개념을 구분하였다. 기초 어휘란 특정 언어 가운데 그 중추적 부분으로서 구조적으로 존재하는 어의 부분 집단이며, 기본 어휘란 어떤 목적에 따라 인위적으로 선정되며 공리성을 지닌 어의 집

단이며, 기간 어휘란 어떤 특정 집단을 대상으로 한 어휘 조사에서 직접적으로 얻어지는 그 단어 집단의 골격적인 부분 집단이라 하였다. 그중에서도 기초 어휘는 언어생활에서 빈도수가 높고 사용 범위가 넓으며 파생이나 합성 등 2차 조어의 근간이 되는 최소한의 필수어로 규정할 수 있으며, 기초 어휘의 선정 및 체계화가 요긴하고 시급한 작업이라고 밝힌 바 있다.

성광수(1999)는 기초 어휘와 기본 어휘의 개념을 구분하지 않고 기초 어휘로 통일하여 사용하고 있는데, 기초 어휘란 언어병리학, 교육학 등의 일정한 목적에 따라 선정된 일정 수의 한정적인 어휘라 하였다.

김종학(2001)은 기초 어휘를 통시적 개념으로 보아 사회의 변천이나 문화적인 환경의 영향을 거의 받지 않고 우리 민족의 사고나 생활에 필수적인 의미를 지닌 어휘소들의 집합체로 규정하고, 기본 어휘를 공시적으로 사용 빈도와 사용 범위를 기준으로 선정한 어휘 목록이라 하였다.

이희자(2003)는 기초 어휘란 가장 기본적이고 핵심적이며 일상적으로 널리 쓰이는 단어들의 총체이며, 기본 어휘란 제한된 범위 내에서의 목적에 따라 인위적으로 선정하는 기본 어휘 선정 작업과 관련된 것이라 하였다.

개별 연구에서 정의하는 개념의 차이가 있으나 일반적으로 기초 어휘는 일상 언어생활에서 사용하는 핵심적인 어휘의 집합으로, 기본 어휘는 일정한 기준에 의해 수집, 선정한 어휘의 집합으로 볼 수 있다.

본 연구는 국립국어원에서 2017년에 수행한 「국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구」의 연장선상에 위치하므로, 2017년 연구의 기초 어휘 개념을 따른다.

1.1.2. 기초 어휘 목록 및 어휘 등급화 사례

국어 기초 어휘의 중요성과 필요성에 대한 논의는 언어학 분야와 교육학 분야에서 이루어져 왔다. 그러나 향후 과제로서 국어 기초 어휘의 중요성을 강조해왔음에도 국어 기초 어휘 선정 방법론을 밝히거나 실제 어휘 목록을 선정하고 어휘 등급을 평정한 사례는 많지 않다.

기초 어휘 목록을 제시한 연구 성과는 교육용, 학습용 어휘에 집중되어 있으며, 제1언어 교육으로서의 국어교육보다는 제2언어 교육으로서의 한국어교육에서 주로 수행되었다. 기초 어휘 목록은 사회 다양한 분야에서 실용적으로 활용 가능하나 국내 연구에서는 교육을 목적으로 한 사례가 대부분이다.

한국어교육용 어휘 목록은 실제 언어 자료를 바탕으로 구축한 어휘 자료를 활용하여 양적 분석 방법을 중심으로, 전문가의 평정을 통한 질적 분석 방법이 함께 사용되었다. 주요 성과로는 서상규 외(2000)와 조남호(2003)을 들 수 있다. 서상규

외(2000)는 1998년도에 구성된 한국어교육용 말뭉치를 보완하여 주요 연구 대상으로 삼아 한국어교육을 위한 기초 어휘 의미 빈도 사전 개발을 추진하였다. 이때 활용된 한국어교육용 말뭉치의 구성은 세종 글말 말뭉치와 세종 입말 말뭉치가 각 55만 어절, 12.5만 어절 규모였으며, 여기에 제6차 교육과정 초등학교 교과서 26.7만 어절과 한국어 교재 10.5만 어절을 포함하여 총 100만 어절의 규모로 구축되었다. 조남호(2003)는 한국어 학습용 어휘 선정 연구에서 현대 국어 어휘 사용 빈도 조사를 우선적으로 수행한 후에 전문가 선정위원들의 등급 판정 작업을 거침으로써 양적 분석과 질적 분석을 함께 사용하였다. 구체적으로는 현대 국어 사용 빈도 조사 목록 59,000여 개 중 출현 빈도 15회 이상인 어휘 10,352개를 추리고, 이들 목록을 선정위원에게 배포하였다. 선정위원들은 1단계, 2단계, 3단계에 속할 수 있다고 생각되는 어휘를 각각 A, B, C로, A, B, C 3단계 어딘가에 속할 수 있다고 생각되는 후보 단어를 D로, 배제할 단어를 E로 판정하였다. 이 결과를 수합하여 약 5,965개 어휘를 최종 선정하였다. 최종 결과물인 어휘 목록에는 순위, 단어 형태, 품사, 동음이의어 구별을 위한 풀이, 등급 등의 정보를 함께 제공하였다.

국어교육용 어휘 목록의 주요 성과로는 김광해(2003)를 들 수 있다. 김광해(2003)는 제1언어, 제2언어 교육의 영역 모두에 적용할 수 있는 등급으로 평정한 어휘 목록을 선정하여 제시하였다. 어휘 선정 및 평정 작업에서 메타 계량 방법을 활용하였다. 메타 계량 방법을 사용하며 고려한 변인은 분포와 자료의 타당도이다. 이 연구에서 메타 계량의 대상은 조현용(2000), 임지룡(1991), 이충우(1992) 등과 <연세한국어사전>, 고려대 국어사전 표제어 자료, 21세기 세종계획의 전자사전 개발 자료 등의 사전 자료를 포함한 총 14건이다. 이 연구에서 선정한 총 어휘의 양은 모두 237,990어이며, 이들을 교육적 중요도에 따라 총 7등급의 집합으로 묶어 제시하였다. 7등급으로 등급화된 이 연구 성과는 이후 ㈜낱말에서 9등급 체제로 보완하였는데, 이는 기존의 3, 4 등급을 3, 4, 5, 6 등급으로 세분화하는 과정을 거친 것이다.

국어교육용 어휘 목록 관련 연구는 한국어교육용 어휘 목록에 비하여 그 수가 적고, 실제 어휘 목록을 제시한 연구보다는 기초 조사 성격의 연구가 많은 편이다. 국어 어휘에 대한 기초 조사는 문교부(1956)를 시초로 한다. 문교부(1956)에서는 “우리말 말수(어휘)가 사용되는 잦기(빈도)의 실태를 조사하여, 과학적인 국어의 기본 형태를 파악하고, 우리말의 합리적인 사용을 꾀하며, 국어의 정상적인 발달 및 정화 운동을 목표하는 교과서 편집이나 계몽을 활용하고, 나아가서는 국어학 연구의 참고 자료로 제공하려 함”이라고 조사 목적을 밝히고 있다(이삼형 외, 2017a).

이후 다수의 기초 조사가 수행되었으나 연구 목적이나 대상이 한정적이라는 한계가 있다. 연구 대상이 교과서, 아동 도서, 문학 도서 등의 일부 출판물이나 학습자의 작문 자료 정도여서 실생활에서의 언어 사용 실태를 온전히 반영하지 못하고 있는 것이다. 또한, 대부분의 연구가 문어(文語)를 대상으로 삼고 있으며, 구어(口語)

를 대상으로 한 연구[1]를 찾아보기 어렵다. 아직까지 국어교육용 어휘 선정을 위한 전면적인 조사는 이루어지지 않았으며, 이러한 목적으로 수행된 기초 조사로는 이삼형 외(2017b)가 있다. 이삼형 외(2017b)는 기초 어휘 선정 작업의 적실성과 실제적인 방안을 마련하는 데 목적을 둔 기초 연구이다. 이를 위해 기초 어휘와 등급화에 대한 제반 이론을 수립하고, 기 구축 말뭉치 및 어휘 목록 사례를 검토하여 이론적 기반을 마련하였다. 그리고 말뭉치에 기반한 기초 어휘 선정 및 등급화를 위한 샘플 말뭉치를 구축하였고, 기초 어휘 사업의 중장기 계획을 수립하였다.

1.2. 교과서 어휘 사례

본 연구에서는 기초 어휘 선정 및 어휘 등급 평정의 검토 작업으로 교과서 어휘와의 비교 분석을 시행한다. 교과서는 정규 교육과정에 따라 편찬된 교재로, 전 국민이 학교 교육을 통해 사용하는 국가적 출판물이다. 초·중·고등학교 교과서는 학습 도구이므로 가장 기본적인 어휘를 포함하고 있으며, 과목과 영역에 따라 전문 영역의 어휘도 포함하고 있다. 따라서 기초 어휘의 상당수가 교과서 어휘로 등장할 것이므로, 국어 기초 어휘 선정 및 어휘 등급 평정 작업에 있어 교과서 어휘와의 비교 분석 작업은 의미가 있을 것이다. 교과서 어휘 비교 분석에 대해서는 Ⅴ장에서 다룰 것이며 여기에서는 예비 작업으로 교과서 어휘와 관련된 선행연구 사례를 살펴보고자 한다.

교과서 어휘에 관한 선행연구는 국가기관에서 수행한 정책 연구와 교육학, 국어학 분야의 개인 연구가 모두 이루어져 왔다. 개인 연구의 경우 정책 연구보다 연구 대상이 다양하고 그 양이 방대하다. 그러나 여기에서는 국어 기초 어휘 목록 선정에 있어 비교 대상으로서 교과서 어휘 목록의 타당성과 합당성을 살펴보는 것이 주목적이므로 정책 연구 성과를 중심으로 살펴보기로 한다. 주요 교과서 어휘 정책 연구 성과를 간단히 정리하면 다음과 같다. 교과서 어휘에 관한 정책 연구는 주로 국립국어원, 한국교육과정평가원, 교육부 주관으로 이루어졌다.[2]

우선 교과서 수록 어휘 실태에 주목한 연구를 살펴보면 다음과 같다. 국어연구소(1986, 1987)의 「국민학교 교육용 어휘」는 초등학교 아동의 이해 어휘와 사용 어휘의 실태를 조사하여 아동의 발달 단계에 맞는 교육용 어휘를 선정하고 초등학교 교과서 편찬에 참고하는 것을 목적으로 하였다. 국어연구소(1988, 1989)의 「중학교 교과서 어휘」 연구는 「국민학교 교육용 어휘」의 후속 연구로, 두 연구 모두 교과서 수록 어휘의 빈도를 제시하였다. 김한샘(2009)은 국어교육용 어휘의 단계별 선정

1) 구어를 대상으로 한 연구로는 장경희 외(2012)가 있다.
2) 주관 기관에 따라 국립국어원 주관 연구(국어연구소, 1986; 1987; 1988; 1989; 서종학, 2000; 민현식, 2003; 2004; 김문오, 2007, 박재현, 2007; 김한샘, 2009), 한국교육과정평가원 주관 연구(양정실, 2015; 윤지훈, 2016, 서지영, 2017), 교육부 주관 연구(이관규, 2016)로 나눌 수 있다.

사업의 일환으로 초등학교 교과서 어휘 조사 연구를 수행하였다. 초등학교 전 학년 교과서 말뭉치를 분석하여 어휘 빈도를 조사하였으며, 목록에서 순위, 학년별·과목별 빈도, 항목, 풀이, 품사 등의 정보를 제시하였다. 양정실(2015)은 초등학교 교과서의 어휘 실태를 분석하였다. 2009 개정 교육과정의 초등학교 교과서 115권을 대상으로 하였으며 전체 어휘는 907,989개로 학년이 올라감에 따라 어휘 빈도수가 증가하고 어휘 종수가 다양해짐을 실증하였다.

그리고 교과서 어휘 표기, 지침 마련과 관련된 연구를 살펴보면 다음과 같다. 박재현(2007)은 교과서 표기의 일관성을 위해 국정 및 검인정 교과서 감수 시 활용할 수 있는 교과서 표기 감수 지침 시안을 마련하였다. 윤지훈(2016)과 서지영(2017)은 교과용 도서 어휘의 표준화 방안을 마련하는 연구를 수행하였다. 표준화 대상 어휘를 선정하고 목록화하였으며, 어휘 표기의 표준 및 허용 범위를 설정하였다. 이관규(2016)는 초등학교 교과서의 어휘 표현 중 우리말로 순화할 수 있는 어휘 1,341개를 선정하여 대체 순화어를 제시하였다.

그 외에 교과서에 수록된 한자어를 분석한 연구, 남북 교과서의 학술 용어를 비교한 연구, 교과서 어휘의 조사단위를 설정한 연구 등이 있었다. 이처럼 교과서 어휘에 관한 연구는 초·중·고등학교 교과서를 대상으로 다양한 관점에서 수행되었음을 알 수 있다. 대부분의 연구가 교과서 수록 어휘는 학습어휘의 성격을 띠고 있으며, 표준적인 어휘라는 관점을 공통적으로 취한다는 점이 특징적이며, 이는 기초 어휘 목록과 교과서 어휘 목록 사이의 상관성을 방증한다.

본 연구에서는 특히 교과서 어휘 관련 선행연구 중에서 교과서에 수록된 어휘를 추출하고 빈도, 품사 등의 어휘 정보를 포함한 어휘 목록을 연구 결과로 제시한 연구들이 있어, 이를 본 연구의 검토 작업의 도구로 활용할 수 있다. 예비 어휘 목록과의 비교 분석 작업은 Ⅴ장에서 다룰 것이다.

2. 국외 사례

2017년 연구에서는 국외 기초 어휘 관련 연구 사례로 영어와 일본어 사례를 살펴보았다. 금년도 연구에서는 유럽권의 프랑스어 사례와 아시아권의 중국어 사례를 조사하였다.

2.1. 프랑스어 사례

2.1.1. 프랑스어 기초 어휘 목록

1) 구게나임(Gougenheim) 2.0 기본 프랑스어

기초 어휘나 기본 어휘에 대한 해외의 연구는 1930년 오그덴과 리차드(Ogden & Richards)가 고안한 '기본 영어Basic English'의 성공과 쟁점화에 힘입어 40~50년대부터 크게 증가한다. 특히 프랑스어에서의 기초 프랑스어, 기본 프랑스어에 대한 개념과 관점, 논의들은 이후 독일과 스페인 등 유럽의 다른 국가의 어휘 교육에 여러 측면에서 영향을 미친다. 기본 영어는 1931년 일본인 영문학자인 후쿠하라 린타로가 오그덴을 직접 방문하여 기본 어휘에 대한 개념을 익히고 일본에 소개하면서 실제 외국어로서의 영어 교육에 사용되었고 이에 따라 실제로 매우 빠르고 효과적으로 언어 습득이 가능하게 되었다. 그러나 850개 어휘로 이루어진 기본 영어는 논리학자들에 의해 고안된 것으로서, 여러 한계를 보인다. 기본 영어라기보다는 영어를 기본으로 한 논리언어, 보편언어의 성격이 강해서, 명사 이외의 품사는 현저히 부족하다.[3] 그러므로 기본 영어의 학습을 통해서 말하는 영어는 실제 영어와는 거리가 멀었으며, 이에 따라 언어의 실제 사용 및 활용이라는 기준에 대한 필요성이 더욱 대두되었다.

프랑스에서도 기본 영어에 영향을 받아 기본 프랑스어(Français fondamental)의 개념을 고안하였다. 그러나 기본 프랑스어 목록을 구축한 구게나임(Gougenheim)은 초반부터 기본 영어를 논리학적 방법론에 따른 단순화된 어휘 목록으로 간주하고 이를 비판하면서 이와는 다른 기준을 통해 어휘 목록을 구축하고자 하였다. 즉, 우선은 통계적 방법이 근간이 되어야 함을 주창하고[4] 철학적 도구어가 아닌 실제 구

3) 850개 중 동사는 18개, 부사와 전치사를 합하여 20개, 그 이외에 812개는 모두 명사이다.

4) 오그덴과 리차드의 기초 영어가 30~40년대에 전 세계의 외국어로서의 영어 교육과 교수법에 지대한 영향을 미치고 효과를 거두었으며 당시 응용언어학 분야가 꽃피울 수 있었던 주요 요인이었다는 사실을 부정하는 이는 아무도 없을 것이다. 그러나 한편으로는 기초 영어가 최소의 언어로 세계를 최대한으로 표상하고자 하는 준 인공어(semi artificial language)이며 다른 언어들은 기초 영어를 보편언어(universal language) 관점에서 단순히 옮겨서 목록화할 위험이 있음을 경고하는 입장도 격렬히

어를 자료로 삼아야 한다고 주장한다. 이때부터 프랑스에서는 구어를 대상으로 한 기본 프랑스어에 대한 개념이 구축되었다. 기본 프랑스어는 프랑스어의 보호와 확산을 위해 UNESCO와 프랑스 교육부가 함께 진행한 사업으로 프랑스나 프랑스어권 나라, 그 밖의 외국에서의 프랑스어 교육을 위해 1950년대에 정립한 단어 및 문법 표지의 목록이다. 즉, 외국어로서의 프랑스어와 모국어로서의 프랑스어 차원에서 두루 사용되는 학습의 기본 도구이다.

초기의 기본 프랑스어는 163개의 대화를 근거로 그 대화 안에 등장하는 312,135개의 단어를 표본으로 삼는다. 이 가운데, 7,995개의 레마형(표제항)이 수집되었고, 이 중에서 각각 1,475개의 1등급 어휘(빈도수>20)와 1700개의 2등급 어휘(빈도수<20) 목록을 구축하였다. <표 1>은 해당 단어의 문법범주(semgram), 대화에 등장하는 빈도수(lemfreq), 그리고 그 단어가 163개의 대화 중 몇 개에 등장하는지를 보이는 분포도(repartition)를 보여주고 있다. 상위 빈도수의 어휘 1,000개에 대해서 270개는 문법적 어휘, 380개는 명사, 200개의 동사, 100개의 형용사, 기타 50개의 타 품사로 구성된다.

<표 1> Français fondamental 목록의 예

mots	semgram	lemfreq	repartition
être	(verbe .)	14083	163
avoir		11552	163
de		10503	163
je		7905	162
il	(ou ils)	7515	160
ce	(pronom)	6846	163
la	(article)	5374	163
pas	(négation)	5308	158
à	(prepos.)	5236	163
et		5082	161
le	(article)	4957	163
on		4266	128
vous		4202	154
un	(article)	4188	162
ça	(pronom démonstratif)	3972	159
les	(article)	3815	162
que	(conj.)	3537	162

현재 서비스되고 있는 어휘 목록은 초기 기본 프랑스어의 연구자인 구게나임(G. Gougenheim)의 이름을 본따서 구게나임 2.0 데이터베이스로 불리고 있으며, 이전의 버전보다 조금 더 보강되어 275명의 인터뷰를 대상으로 총 8,774개의 어휘를 제공하고 있다.

제기되기도 하였다.(Lõpez, 2006: 97~98)

구게나임(Gougenheim, 1955: 407~410)은 어휘 선정에 있어 빈도수의 한계를 지적하기도 한다. 어떤 언어이든 빈도 목록에서는 기능어(문법어)들이 높은 빈도를 차지하고 보조용언, 동사들이 그 뒤를 따르며 명사는 그 이후에 온다. 특히 명사 중에 높은 빈도를 보이는 것은 총칭명사들('시간', '사람', '남자', '여자', '아이', '물건'('것') 등)이며 구체명사('팔꿈치', '단추', '포크' 등)는 약한 빈도를 보이거나 텍스트 구성에 따라 불규칙적이고 불안정한 특징을 보인다. 구체명사는 그 자체로 빈도를 가지고 있다기보다는 텍스트나 구어 자료가 다루는 주제와 밀접한 관련이 있다. '시내버스', '전철'[5], '포크' 등과 같은 비교적 낮은 빈도의 구체 '상용어휘'를 포함하기 위해 '접근성(availability)'이라는 기준을 도입한다. 대화 중에 자주 등장하지는 않지만 필요할 때 지속적으로 제공되어야 하는 것을 일컫는다. 접근성의 기준은 빈도와 관련이 없으며, 자동적으로 산출되지 않는다. 접근성은 접근도(degree of availability)로 결정되는데, 구게나임은 접근성의 확보를 위해 16개의 관심범주(신체, 의식주 등)를 상정하고 충분히 거리가 있는 서로 다른 지역 4개의 초·중등교육기관들에 각 관심범주에 가장 많이 사용된다고 생각하는 20개의 단어를 쓰게 하고 이들 중 공통되게 접근성이 높은 목록을 수립한다. 이렇게 고빈도 어휘와 고 접근성 어휘 목록으로 기초 프랑스어 목록을 작성한다.[6] 그러므로 기본 프랑스어 목록은 다음의 세 단계의 방법론을 따른다고 할 수 있겠다.

1) 어휘 빈도 목록 수립
2) 단어 접근도 조사
3) 능력 있는 화자 집단에 의한 자료 정제

기본 프랑스어 목록은 이후 구게나임(1958)의 『기본 프랑스어 사전』(Dictionnaire fondamental de la langue française), 그리고 기본 어휘에 대한 구문 구조 등과 관련된 기본 프랑스어 문법 교육 등으로 이어지며,[7] 이후 구어 위주의 언어교육에 대한 매우 중요한 언어학계 논쟁의 시발점이 된다.[8] 이와 같은 흐름에 따라 기본 어휘 연구는 자국어의 교육과 외국어로서의 언어 교육뿐 아니라, 구어 연구, 코퍼스 연구, 기본 문법(기본 어휘 구문)에 대한 개념, 철자 개혁, 문해력, 사전학 더 나아가 언어계획과 정책 등과 밀접한 관련을 맺게 된다.

---

5) 구게나임(Gougenheim, 1955: 408)은 '전철(métro)'이 파리 지역의 구어 자료에서조차 높은 빈도를 보이지 않음을 지적한다.
6) 16개의 관심범주가 모든 분야를 포괄하지 못하므로 정부 부처에 의뢰하여 보건의료 분야 등의 목록을 추가하기도 한다.
7) Gougenheim, G., Michea, M., Rivenc, P., Sauvageot, A. (1956/1964), L'elaboration du Francais fondamental (1er degre). Etude sur l'elaboration d'un vocabulaire et d'une grammaire de base, Paris : Didier.
8) 지난 2006년 Gougenheim et al. 이 발표한 기본 프랑스어(le français fondamental)의 50주년을 기념하는 학술회의가 열리면서 그 의의가 재조명되었다.

2) 구게나임 이후의 어휘 데이터베이스

구게나임 이후 기초 어휘 학습과 관련하여 프랑스에서 네 개의 어휘 데이터베이스가 차례로 등장한다. BRULEX, LEXIQUE, NOVLEX와 MANULEX 등이다. BRULEX나 LEXIQUE 2.0이 일반 성인의 프랑스어 어휘 목록을 다루는 데 비해, NOVLEX와 MANULEX는 초등교육에 활용할 기초 어휘 목록을 제시하고 있다.

데이터베이스 BRULEX는 1990년에 Content *et al.*이 만든 프랑스어 어휘 빈도 목록으로, 표제어는 총 35,746개로 구성되어 있는데, 자료 자체는 기존 사전의 것을 활용하였다. 즉, 『로베르 소사전』(Le Petit Robert, 1986)의 표제어를 어휘 자료로 선정하였는데 그 이유는, 이 사전이 동시대의 통상어, 구어 자료가 제일 잘 반영되어 있으며 이뿐 아니라 필수적인 전문어, 그리고 고전문학을 읽는 데 필요한 고어도 포함하고 있기 때문이라고 밝히고 있다. BRULEX는 35,000여개의 어휘에 대해서 TLF(Trésor de la langue française)의 대규모 말뭉치인 '프랜텍스트(Frantext)'[9]를 통해 빈도를 추출하고, 사전적 기술의 대상이 되는 '기본 정보(basic informations)'와 '추출 정보(generated informations)'를 제공한다. 추출 정보는 철자형 형태(orthographic forms) 빈도(FRFRM)와 레마형 어휘 빈도(FRLEX), 철자나 발음, 음운, 음절에 대한 통계치 등을 포함한다. BRULEX가 제시하는 빈도수는 TLF 사전의 빈도 정보를 활용하고 있는데, 이 빈도수는 1919년~1964년 사이의 문학텍스트 코퍼스에서 추출한 2,600만 개의 단어를 대상으로 한 것이다. 이렇듯 BRULEX는 기존의 사전과 코퍼스 빈도를 재활용한 데 그쳤다는 한계가 있다.

2004년에 New, Pallier, Ferrand, Matos 등에 의해 개발된 LEXIQUE 현대 프랑스어 어휘목록 역시 프랑스의 가장 대표적인 프랜텍스트(Frantext) 말뭉치를 사용하였다. 이중 1950년~2000년까지의 문헌과 이에 덧붙여 1,500만 개의 프랑스 웹페이지를 통해 빈도를 계산하였다. 프랜텍스트(Frantext) 말뭉치가 문학과 학문적 텍스트 위주로 구성되어 있으므로 이중의 코퍼스를 사용하였는데, 구체적으로는, 프랜텍스트(Frantext)에서 추출한 13만 항목의 철자형 형태에 대하여 패스트서치(FastSearch)[10] 검색 엔진을 통해 1,500만 개의 웹페이지에 등장하는 어휘를 추출하였다. 각 단어에 대해 그것이 등장하는 웹페이지의 수를 확보하고 프랜텍스트(Frantext), 패스트서치(FastSearch), 그리고 TLF의 빈도수 상관도를 계산하여 최종적으로 73,000여개의 빈도별 어휘 목록을 구축하였다.

LEXIQUE는 크게 세 부분으로 구성되어 있는데, 철자형 형태로 구성된 '문자소 데이터베이스(Graphèmes DB)'와 대표형 표제항인 레마로 구성된 '사전등재형 데이

9) 프랑스 ATILF에서 개발한 프랑스어의 대표적인 코퍼스로 19~20세기의 문학(소설, 시 에세이 등) 및 과학기술 문헌 약 3,200개 담고 있다. 단어는 약 3,100만개. http://zeus.inalf.fr/frantext.htm

10) http://www.alltheweb.com

터베이스(Lemmes DB)'와 철자, 2철자 단어, 3철자 단어, 음소, 어절 등의 빈도 통계를 요약한 파일로 구성되어 있다. 문자소 데이터베이스와 사전등재형 데이터베이스에서 제공하는 정보들은 각각 아래 [그림 3], [그림 4]와 같다. 현재 버전 3이 서비스되고 있으며[11] 온라인상에서는 공개 사전(Open Lexique)을 통해 다른 19개의 어휘 데이터베이스의 정보도 함께 검색하여 비교할 수 있도록 하고 있다.

Field Name	Description
graph	Orthographic representation
phon	Phonological representation
cgram	Grammatical category
genre	Gender
nombre	Number
lemme	Lemma
rand	Random number
frantfreqparm	Frantext frequency
fsfreqparm	Fastsearch frequency
nblettres	Number of letters
nbphons	Number of phonemes
cvcv	Orthographic abstract representation
pcvcv	Phonological abstract representation
puorth	Orthographic uniqueness point
puphon	Phonological uniqueness point
syll	Syllabified form
nbsyll	Number of syllables
syllcv	Syllabified abstract form
voisorth	Number of orthographic neighbours
voisphon	Number of phonological neighbours
orthrenv	Reverse orthographic representation
phonrenv	Reverse phonological representation

[그림 3] 철자형 영역과 기술 용어

Field Name	Description
lem	Orthographic representation of the lemma
graph	Inflectional family
phon	Phonological family
cgram	Grammatical class family
genre	Gender family
nombre	Number family
rand	Random number
frantfreqcum	Inflectional frantext cumulative frequency
frantfreqgraph	Inflectional frantext frequency family
fsfreqcum	Inflectional fastsearch cumulative frequency
fsfreqgraph	Inflectional fastsearch frequency family

[그림 4] 레마형 영역과 기술 용어

1_ortho	2_phono	3_lemme	4_cgram	5_genre	6_nombre	7_freqlemfilms	8_freqlemlivres	9_freqfilms	10_freqlivres	11_infover	12_nbhomogr	13_nbhomoph	14_islem
dansant	d@s@	danser	VER			108.14	92.57	2.34	5.54	par:pas:	2	3	0
dansante	d@s@t	dansant	ADJ	f	s	1.65	6.89	0.48	1.76		1	2	0
dansantes	d@s@t	dansant	ADJ	f	p	1.65	6.89	0.21	1.96		1	2	0
dansants	d@s@	dansant	ADJ	m	p	1.65	6.89	0.37	0.61		1	3	0
danse	d@s	danse	NOM	f	s	41.06	35.14	38.62	29.19		2	8	1
danse	d@s	danser	VER			108.14	92.57	18.46	9.8	imp:pre:2s:	2	8	0
dansé	d@se	danser	VER	m	s	108.14	92.57	5.27	4.32	par:pas:	1	4	0
dansée	d@se	danser	VER	f	s	108.14	92.57	0.11	0.27	par:pas:	1	4	0
dansent	d@s	danser	VER			108.14	92.57	3.14	5.54	ind:pre:3p:	1	8	0

11) http://www.lexique.org

1_ortho	15_nblettres	16_nbphon	17_cvcv	18_p_cvcv	19_voisorth	20_voisphor	21_puorth	22_puphor	23_syll	24_nbsyll	25_cv-cv	26_orthrenv	27_phonrenv	28_orthosyll
dansant	7	4	CVCCVCC	CVCV	3	14	5	4	d@-s@	2	CV-CV	tnasnad	@s@d	dan-sant
dansante	8	5	CVCCVCCV	CVCVC	1	3	0	0	d@-s@t	2	CV-CVC	etnasnad	t@s@d	dan-san-te
dansantes	9	5	CVCCVCCVC	CVCVC	0	3	0	0	d@-s@t	2	CV-CVC	setnasnad	t@s@d	dan-san-tes
dansants	8	4	CVCCVCCC	CVCV	1	14	0	4	d@-s@	2	CV-CV	stnasnad	@s@d	dan-sants
danse	5	3	CVCCV	CVC	6	18	5	3	d@s	1	CVC	esnad	s@d	dan-se
danse	5	3	CVCCV	CVC	6	18	5	3	d@s	1	CVC	esnad	s@d	dan-se
dansé	5	4	CVCCé	CVCV	4	54	0	4	d@-se	2	CV-CV	ésnad	es@d	dan-sé
dansée	6	4	CVCCéV	CVCV	2	54	0	4	d@-se	2	CV-CV	eésnad	es@d	dan-sée
dansent	7	3	CVCCVCC	CVC	2	18	0	3	d@s	1	CVC	tnesnad	s@d	dan-sent

[그림 5] danse의 철자형 형태의 정보 예 Lexique3.txt

[표지]: **ortho**: 단어; **phon**: 단어의 음운형태; **lemme**: 단어의 레마형; **cgram**: 문법범주; **genre**: 성; **nombre**: 수; **freqlemfilms**: 영화자막 코퍼스에서의 단어 레마형 빈도 (백분위수); **freqlemlivres**: 서적 코퍼스에서의 단어 레마형 빈도 (백분위수) **freqfilms**: 영화자막 코퍼스에서의 단어 빈도 (백분위수); **freqlivres**: 서적 코퍼스에서의 단어 빈도 (백분위수);**infover**: 동사의 법, 시제, 인칭; **nbhomogr**: 동형어 개수; **nbhomoph**: 동음어 개수; **islem**: 레마형인지 아닌지 표시; **nblettres**: 철자 수; **nbphons**: 음소 수; **cvcv**: 철자구조; **p-cvcv**: 음운구조; **voisorth**: 철자변이형 개수; **voisphon**: 발음 변이형 개수; **puorth**: 철자 통합점; **puphon**: 음운통합점; **sylll**: 어절화된 음운형태; **nbsyll**: 어절 수; **cv-cv**: 어절화된 음운구조; **orthrenv**: 역철자형태; **phonrenv**: 역음운형태; **orthosyll**: 어절화된 철자형태]

2.1.2. 어휘 등급화 관련 선행 연구 조사

프랑스어에서 단계별 등급화가 이루어진 어휘 목록이 따로 존재하지는 않으나, 어휘 학습의 단계 중 초등 단계를 염두에 두고 개발된 어휘 데이터베이스가 있다. 앞서 살펴본 두 데이터베이스, 즉, BRULEX나 LEXIQUE가 성인을 대상으로 한 어휘목록인 반면, NOVLEX와 MANULEX는 어린이 및 청소년의 어휘 학습 도구로 개발된 기초 어휘 DB이다.[12] 기존의 기초 어휘가 학습의 단계를 고려하지 않은 한 언어의 기본 어휘를 다루었다면 NOVLEX, MANULEX는 나이대별로 맞는 어휘량과 어휘습득 절차 및 방식을 반영하려 했다고 볼 수 있다.

인간의 어휘 저장 능력은 20세~25세 이상은 더 이상 유의미한 발전을 하지 않지만[13] 어린 아이의 어휘 습득은 9개월~15개월에서 시작해서 2세에 가속되고 특히 읽기를 배우기 시작할 때부터 급증한다. 프랑스의 CE1(만 7세)학년과 CM2(만 10세)학년 사이에 어휘력의 발전이 주목할 만한 수준에 이른다(Ehrlich et al. 1978). 7세~10세 기간에는 그 이전 1세에서 7세 사이에 습득한 전체 어휘량보다 매해 50%씩 증가 추세를 보인다고 한다. 그러므로 어휘습득량과 각 단계에 알맞은 어휘 목록을 구성하는 것이 필요할 것이다.

우선 2001년에 NOVLEX(Lambert & Chesnet, 2001)의 등장으로 어린이를 위한 글에서의 프랑스 어휘 빈도가 측정되었다. CE2(만 8세)의 교재 19종과 어린이용 문학서적(총 417,000 단어)을 선정하여 20,600개의 철자형 단어와 9,300개의 레마형 단어에 대해 빈도수 데이터를 제공한다. 반면, MANULEX는 6세~11세에 해당하는 학년을 크게 3단계로 나누는데, CP학년(만 6세)은 음운의 매개로 어휘를 구성하는 단계, CE1학년은 읽기/쓰기의 단어를 점차적으로 인지하면서 어휘의 철자를

12) http://unpc.univ-lyon2.fr/~lete/manulex/index.htm
13) Lété (2004: 242)

구성할 수 있는 단계이고 그 후 3단계 학년(3 cycle, 8세~11세)은 글에 지속적으로 노출됨에 따라 어휘 축적이 일어나는 시기이다. 각 학년별 학습교재와 프랑스어 교재 총 54종에서 48,886개의 철자형 단어와 23,812개의 레마형 단어를 구성하며 각각의 단계별 빈도와 총 어휘에 대한 빈도를 아래 <표 2>와 같이 제공한다.[14)]

<표 2> Tableau 2: Nombre de mots et nombre d´entrées du lexique des formes orthographiques et du lexique des lemmes à chaque niveau

	CP	CE1	CYCLE 3	TOTAL
Mots	172 400	351 100	1 386 600	1 910 000
Lexique des formes orthographiques	11 400	19 100	45 600	48 900
Lexique des lemmes	6 800	10 500	22 500	23 900

학년 별로 보면, CP학년에서 6,800개, CE1학년에서 10,500개, 초등학교 고학년인 Cycle 3에서 22,500개, 그리고 6~11세에는 총 23,900개의 레마형 단어가 선정되었다. 그러나 이 어휘수(lexicon)는 실제 어린이들이 읽기/쓰기에 동원하는 어휘량(vocabulary)과는 차이가 있다. 각 학년 교재들이 수록하는 어휘와 알퐁스 도데의 단편소설 「방앗간 소식」의 어휘를 비교하여 실제 학년별로 읽기에 활용되는 어휘량을 산출한 Lété는 CP학년이 1,900개, CE1이 2,900개, Cycle3학년들이 4,900개의 레마형을 실제 글을 이해하는 데 활용한다고 추정한다.[15)] 이외에 학년별, 나이별 어휘 습득, 어휘 활용의 양을 추정하는 연구들이 많이 있는데, 대표적으로 Ehrlich et *al.*(1978)과 Pothier & Pothier(2003)이 있다. Ehrlich et al.(1978)는 7세(CE1)부터 11세(CM2)까지의 학생을 대상으로 학생 당 450개의 단어에 대해 5점 척도로 단어를 알고 있는 정도를 표시하게 하여 13,500개의 어휘 표본을 만들었으며, Pothier & Pothier(2003)는 CP학년부터 CM2학년까지 48,900명의 초등학생들을 대상으로 11,700개의 단어(학생 당 50단어)를 쓰고 그 철자를 확인했는데, 75% 이상의 학생이 철자를 정확하게 알고 있는 단어들을 목록화하였다. 이상의 어휘량을 비교하면 아래 <표 3>과 같다.

14) 한편, 빈도수 계산에 있어서는 Caroll et al(1971)의 *The American Heritage Word Frequency Book*을 따라서 각 4단계의 두 어휘유형에 대해 세가지 지표로 계산한다. 지표 F는 해당 코퍼스의 원시 빈도, 지표 D는 교재들의 단어 분포, U는 D에 의해 백분위수를 계산한 빈도에 해당하며 지표 U가 가장 언어현실에 부합하는 빈도수라 할 수 있다.
- D = [log(∑pi)-[(∑pi log pi/∑pi]] / log(n) (n은 각 학년별 교재 수(CP의 n=13, CE1의 n=13, Cycle3의 n=28, 코퍼스 총체의 n=54), i는 교재 번호(i=1, 2,..., n), pi는 i번 교재에서의 해당 단어 빈도(F), pi=0이면, pi log pi = 0)
- U = (1,000,000/N)[FD + (1-D)*fmin] (N은 코퍼스에서 총 단어수(172,348(CP); 351,024(CE1); 1,386,546(Cycle3); 1,909,918(CP~Cycle3), F는 코퍼스에서의 해당 단어 빈도, D는 분포도, fmin은 fi와 si의 합에 대한 1/N , fi는 교재 i에서의 빈도수이고 si는 그 교재에서 해당 단어 수)

15) Lété(2004)의 초등학생의 어휘사용량을 추정한 바에 따르면 cycle 3학년(11세)에는 5,000개 가량의 단어(레마형)가 노출된다고 할 수 있다. 하지만 실제 성인의 일반 글이 약 75,000개의 표제형을 포함한다고 할 때(New et al., 2001) 이는 매우 부족하며 상당한 독서로 이를 채워나가야 할 것이다.

<표 3> 연구별 어휘량 비교

	CP	CE1	CE2	CM1	CM2
1) MANULEX의 어휘 수	6,800	10,500	22,400	22,500	22,500
2) 독해를 위한 실제 어휘 추정	1,900	2,900	4,900	4,900	4,900
3) S. Ehrlich et al.(1978)에서 '매우 잘 아는 단어'라고 판단한 어휘량	-	2,400	2,900	3,000	3,300
4) Pothier & Pothier(2003)에서 올바른 철자 쓰기로 추정한(75% 이상 학생의 정확도) 어휘 수	400	1,100	2,200	3,700	5,100

프랑스어의 경우, 모국어 화자를 위한 기초 어휘(학습용 어휘)의 활용에서 가장 두드러진 분야는 어휘 교육 분야, 특히 철자 연습이다. 유럽 언어들이 거의 모두 레마형과 실제 활용이 매우 괴리가 있기에 앞서 살펴본 어휘 목록들은 모두 레마형과 동사/형용사/명사 활용형이나 격변화형을 포함한 철자형(변이형)을 구분하여 다루고 있다.

프랑스어 사전의 경우, 학습용 어휘를 사용자 그룹별, 나이별로 구분하여 제공하는데, 사용자 그룹을 구분하여 표제어를 선별한 사전류가 1990년대에 급증하는데, 대략의 예는 아래 <표4>와 같다.

<표 4> 프랑스어 학습용 어휘 사전 사례

출판사	사전명	연령대	정보
Hachette	Hachette benjamin		6,000단어, 600 삽화
	Hachette juniors	CE2, CM학년	20,000단어
Larousse	Mini Débutants	CP-CE1학년(6-8세)	5400단어, 500삽화
	Maxi Débutants	CE2, CM학년(7-10세)	5,400단어, 500삽화
	Super Major	CM1, CM2(9-12세)	23,800단어
	Dictionnaire Larousse du collège	중학생	35,000단어
Robert	Le Robert Benjamin	CP, CE학년	6,000단어, 640삽화
	Le Robert Junior	CE~6e학년	20,000단어, 1,000삽화
	Le Robert des jeunes	CE~6e학년	20,000단어
	Le Robert Collège	12~15세	40,000단어
	Le Robert Micro		35,000 단어와 그 어족

사실 사전은 수요자 및 시장과도 밀접한 관련이 있어 그 다양성을 유형화하는 데 어려움이 있다. Larousse사에서 출판된 종이사전 중 표제어 50,000개 미만의 사전만 나열해도 아래와 같이 다양하다.[16]

16) 어린이용 백과사전, 어린이용 부문 사전 제외

<표 5> Larousse사에서 출판된 프랑스어 사전 목록

사전명	수록 내용
Mini Débutants	5400단어, 500삽화
Maxi Débutants	5,400단어, 500삽화
Super Major	23,800단어
Dictionnaire Larousse du collège	35,000단어
Dictionnaire français	6,000단어, 640삽화
Petit dictionnaire français	20,000단어, 1,000삽화
Petit dictionnaire français	20,000단어
Dictionnaire noms communs/noms propres	38,000 단어/관용어/표현 5,000 고유명사
Larousse de poche, Dicionnaire noms communs/noms propres	38,000 단어/관용어/표현 5,000 고유명사
Dictionnaire de la langue française	38,000 단어, 5,000 고유명사
Le Plus Petit Larousse	38,000 단어/관용어/표현, 5,000 고유명사
Dictionnaire de la langue française	38,000 단어, 50,000 동의어, 20,000 관용어
Le Plus Petit Larousse	16,000 단어
Mini Dictionnaire de français	25,000 단어
Le Larousse des 1 000 mots (Hors collection Jeunesse)	

2.2. 중국어 사례

중국의 기초 어휘 연구는 '기초 어휘'와 '상용 어휘'로 나뉘어 진행되었으며, 중국어의 문자인 한자가 표의문자라는 점에서 상용 한자의 선정도 어휘 선정의 성격을 띠기도 한다. 중국어 연구에서 '기본 어휘'라는 용어는 1947년 손복원(孫伏園)의 「기본사휘연구술요(基本詞彙研究述要)」라는 논문에서 처음 등장하였다. 여기서 제시한 기본 어휘는 일상생활과 상용성을 그 특징으로 하며, 중국어를 모어로 하는 화자는 교육을 받은 사람이든 그렇지 않은 사람이든 모두 알 수 있는 단어이다(김현철·조은경, 2010). 이후 러시아의 기본 어휘 학설, 특히 스탈린의 「마르크스주의와 언어학 문제」에서 "모든 단어가 공통적으로 어휘를 구성하지만, 그 중 기본 어휘는 일반 어휘보다 그 양은 적지만 생명력은 더 강하여 몇백 년이 지난 후에도 존재하며 새로운 단어를 만드는 기초가 된다"(楊同用, 2003:23)는 주장이 중국에 받아들여지면서, 기초 어휘 연구가 활성화되었다.

중국어의 기초 어휘 연구는 크게 세 가지로 구분되는데, 첫째는 스탈린의 논문에 제시된 '기본 어휘(основной словарный фонд)'를 중국어의 특성에 맞게 번역하기 위한 논쟁이다. 흔히 중국 학계에서는 '기본 어휘(基本語彙)'라고 표현하였으나, 중국어에서 기초 어휘가 모두 단음절 어휘라는 점에서 '기본 자휘(基本字彙)'라는 용어가 제시되기도 했다. 그러나 어휘 연구에 있어 문자를 기본으로 해서는 안 된다는 반박이 다시 힘을 얻으면서, 논쟁이 일단락되었다(梁伍鎭, 2005). 이를 통해 알 수 있는 점은 중국어는 그 특성상 '문자(字)'와 '단어(詞)'의 구분이 모호하다는 것인데, 이 때문에 기초 어휘에 대한 정확한 선정 기준이나 언어 계량이 이루어지지 않고 있다(국립국어원, 2009).

그럼에도 불구하고 중국어 기초 어휘의 선정 방법에 대한 논의는 蘇培成(1993)에서 일부 제시되었는데, 그는 기초 어휘가 통시적 범주에 속하기 때문에 기초 어휘를 논할 때에는 반드시 시대적 구분이 필요함을 주장했다. 그 구체적인 방법으로 먼저 공시적 측면의 사용 빈도 통계에 따라 시기별 상용 어휘를 선정하고, 시기별 상용 어휘 중 공통된 부분을 바탕으로 기초 어휘를 선정하는 것을 제안하였다(김현철·종은경, 2010). 이를 통해 중국에서는 빈도 기반의 '상용 어휘'와 의사소통의 핵심 요소가 되는 '기초 어휘'를 확실하게 구분하고 있음을 알 수 있다. 이와 관련하여 상용 어휘는 공시적 측면에서 높은 빈도를 보이는 단어들로 시대성이 높은 한편, 기초 어휘는 사용 빈도가 높으면서도 통시적으로 사용된다고 설명하기도 한다(曹煒, 2004). 또한, 기초 어휘는 내용어만을 포함하는 한편 상용 어휘는 단순히 빈도에 근거하므로 기능어도 포함한다(이상도, 1995:139).

한편, 기초 어휘와 일반 어휘를 구분하기 위해 세 가지 기초 어휘의 특징인 보편성·생산성·안정성에 대한 논의 또한 기초 어휘 선정 문제와 관련하여 줄곧 이어져

오고 있다. 이들 세 가지 특징이 얼핏 보기에는 동시에 구비할 수 있는 조건으로 보이지만, 세 가지 조건을 모두 필수 조건으로 한다면 사람들이 실제 어감으로 느끼는 기초 어휘의 상당 부분이 배제될 것이며, 고정된 어휘를 기초 어휘의 범주에 포함시키기도 어려울 것이라는 문제 제기(周薦, 1987; 梁伍鎭, 2005에서 재인용)가 대표적이다. 그러나 이러한 논쟁과 연구의 흐름에도 불구하고, 무엇보다 중요한 과제인 기초 어휘의 선정 기준에 대한 논의와 구체적 선정 작업이 아직까지 시도되고 있지는 않다.

빈도 기반의 '상용 어휘'에 대해서는 그 통계 작업이 상당히 이루어진 편이다. 중국어 상용 어휘의 범위와 관련한 주요 통계 자료[17]는 다음과 같다.

<표 6> 중국어 상용 어휘에 관한 주요 통계

연도	결과물	어휘 수	연구자	대상
1959	普通話三千常用詞表	3,000개	文改會 漢字組	구어 어휘
1960	兩千雙字詞表	2,000개	<文字改革> 발표	구어 어휘
1964	外國學生用四千詞表	4,000개	北京語言學院	구어 어휘
1973	Frequency Dictionary of Chinese Words	3,000개	E.Shen Liu	
1981	外國人實用漢語常用詞表	3,040개	北京語言學院	구어 어휘
1983	報刊詞語三千六百條	3,600개	北京語言學院	구어 어휘
1983	中小學文科教學七千詞表	7,000개	承德醫學院, 中國人民大學	구어 어휘
1983	現代漢語七千詞表	7,000개	中國人民大學	구어 어휘
1983	擬制文件六千詞表	6,800개	燕山計算機應用研究中心	구어 어휘
1985	信息處理用現代漢語五千詞表	2음절 이상 단어 5,639개	現代漢語工程實用詞庫國家標準研制組	구어 어휘
1985	現代漢語頻率詞典	상용어휘 8,548개	北京語言學院	- 4분야: 신문·잡지, 과학기술문헌, 구어 자료, 문학 작품 - 전체 어휘량: 131만 - 개별어: 31,159 - 한자 수: 180만
1986	對外漢語教學常用詞表	4,000개	北京語言學院	구어 어휘
1988	漢語水平等級標準和等級大綱	甲乙丙 3급 5,168개	中國對外漢語教學學會	구어 어휘
1989	現代漢語常用詞詞頻詞典	상용어휘 9,000개	北京航空航天大學 등	구어 어휘

17) 楊同用, 2003: 23; 국립국어원, 2009에서 재인용함.

1990	現代漢語常用詞庫	상용어휘 9,000개	山東大學	구어 어휘
1990	中小學漢語常用詞表	상용어휘 8,197개	北京師範大學 現代教育技術研究所	구어 어휘
1991	北京口語調查	상용어휘 6,966개	北京語言學院	구어 어휘
1992 (2001, 修正本)	漢語水平詞彙與漢字等級大綱	8,822개 (갑급 1033, 을급 2,018, 병급 2,202, 정급 3,569)	中國對外漢語教學領導小組辦公室 漢語水平考試部	대표 어휘집 4종(現代漢語頻率詞典, 現代漢語常用詞詞頻詞典, 中小學漢語常用詞表, 現代漢語常用詞庫)과 어휘표 등을 근거로 분석

이들 상용 어휘를 활용하여 제2언어로서의 중국어 교육에서 다양한 어휘 목록을 제시하였다. <표 6>의 마지막에 제시된 <漢語水平詞彙與漢字等級大綱>는 외국인 대상 중국어 시험인 HSK(한어수평고시, 漢語水平考試)의 기반으로 활용되었으며, 중국 소학교 및 중학교의 어휘 교육 주요 지침서로 사용되었다(張和生, 2010). 이들 어휘는 다음의 네 가지 관점과 여덟 가지 세부 원칙을 기준으로 선정되어 등급화되었다(김현철·조은경, 2010:267).

<표 7> <漢語水平詞彙與漢字等級大綱>의 어휘 선정 및 등급화 원칙

주요원칙	내용
빈도 통계 관점	- 상용성 원칙 - 균등성 원칙
언어학적 관점	- 과학성 원칙 - 규범성 원칙
제2언어로서의 중국어 교육 관점	- 실용성 원칙 - 연상성 원칙
학생 언어 습득의 관점	- 포용성 원칙 - 서열성 원칙

<漢語水平詞彙與漢字等級大綱>의 등급 구분은 중국의 자국어 교육에도 활용되었다는 점에서 의미가 있다. 2000년 3월에 반포된 <漢語水平詞彙與漢字等級大綱>의 개정판에서는 6년제 소학교(한국의 초등학교에 해당) 단계에 갑, 을구에 해당하는 상용 어휘 3,000자를 익히되 2,500자는 쓸 줄 알아야 하며 실제 사용되는 의미도 파악하여야 한다고 규정하였고, 중학교에서는 8,000개 정도를 어휘 교육의 목표로 삼았다(梁伍鎭, 2003). 이 3,000개의 실제 언어 자료에서의 분포율(coverage)은 86%에 달하며, 상위 8,000개는 95%인 것으로 나타나고 있다(梁伍鎭, 2005).

2009년에는 국가한판·공자학원총부에서 이를 수정, 보완하여 어휘를 1급부터 6급까지 6개 등급으로 구분하였다. 구HSK의 등급에 대응하는 신 HSK의 등급을 다음 <표 8>과 같이 제시하고 있으나, 이에 대한 자세한 선정 기준은 밝히지 않았다.

<표 8> 신HSK와 구HSK의 등급 구분(임학준, 2016)

학습 단계	신HSK			구HSK		
	등급	학기수(개)	어휘량(개)	등급	학년(시수)	어휘량(개)
초급	1급	1학기	150	갑	1학년(800)	1,033
	2급	2학기	300	을		2,018
중급	3급	3학기	600	병	2학년(1,600)	2,202
	4급	4학기	1,200			
고급	5급	4학기 이상	2,500	정	3·4학년(3,000)	3,659
	6급		5,000 이상			

이후 중국이 경제성장과 함께 다른 국가와의 교류를 활발히 함에 따라서 중국어 학습의 요구가 늘어남에 따라 '상무한어(常務漢語)', 즉 비즈니스 중국어를 위한 어휘 선정도 이루어졌다. 주된 어휘 선정 결과는 북경대학에서 2006년 출판된 <BCT大綱>, 2014년 국가한판·공자학원총부에서 출판된 <新BCT大綱>을 꼽을 수 있다. 이중 <新BCT大綱>은 총 4,648개의 단어를 포함하고 있으며, 구체적인 등급 구분 없이 A·B권으로 나뉘어 수록되어 있다(임학준, 2016). 여기에는 전문 용어가 다수 수록되어 있어 기초 어휘에 해당하지 않는 단어가 많으나, 직장 생활이라는 일상생활에 필요한 어휘라는 점에서 많은 수요를 보이고 있다.

이상으로 살펴본 바에 의하면, 중국에서의 기초 어휘 사례는 아직 명확하게 제시된 바가 없으며, 이는 중국어의 특성에서 기인하는 것으로 보인다. 한편 상용 어휘와 어휘 등급화 사례는 제2언어로서의 중국어 교육에 집중되어 있다. 주목할 만한 점은 기초 어휘와 상용 어휘의 차이를 구분하면서, 상용 어휘를 기반으로 다양한 교육용 어휘가 제시되고 있기도 하며 이러한 교육용 어휘가 제2언어로서의 언어 교육뿐만 아니라 자국민의 표준어 평가, 어린이들의 조기 언어 교육에도 널리 활용되고 있다는 점이다.

Ⅲ. 기초 어휘 추출 목적의 언어 자료 정제와 처리 과정

1. 기초 어휘 추출 목적의 언어 자료 정제

기초 어휘 추출 목적을 위해 2017년에 구축한 언어 자료 현황을 정리하고, 본 연구에서의 보완 계획에 대해 살펴보기로 한다.

1.1. 기초 어휘 추출 목적의 언어 자료 구축 현황

1.1.1. 2017년에 구축한 언어 자료 현황

2017년도 어휘 등급화 작업을 위해 구축된 언어 자료는 다음과 같다.

<표 9> 언어 자료 전체 통계

종류	비고	장르 수	어절 수	용량 (MB)
세종 말뭉치	신문, 소설 제외	10	32,407,263	374
도서 자료	세종 말뭉치의 소설을 도서 자료의 소설과 통합함.	24	173,730,053	1,640
잡지 자료	성격이 비슷하고 분량이 작은 것은 한데 묶음.	25	408,273,167	3,344
블로그 자료	N사: 35개 장르, 222개 블로그 A사: 1개 장르, 72개 블로그 L사: 1개 장르	37	245,628,059	2,642
드라마 자료		1	39,314,345	380
신문 자료	11개 신문 1990~2015년 기사 세종 말뭉치의 신문을 이 항목에 통합함.	1	961,256,124	17,849
위키 백과		1	166,040,163	1,775
방송 뉴스	2017년 8~10월 뉴스	1	18,919,129	193
계		100	2,045,568,303	29,197

각 하위 자료(sub-corpus)의 세부 내역은 다음과 같다.

<표 10> 세종 말뭉치 장르별 통계

코드	장르	어절 수
12	잡지	7,060,533
130	책-총류	1,855,494
132	책-교육	4,240,064
134	책-체험	3,142,657
135	책-인문	5,032,111
136	책-사회	2,526,948
137	책-자연	1,391,485
138	책-예술	2,840,098
139,14,15,19	기타 출판물	1,813,490
21~29	대화(희곡, 회의)	2,494,383
계		32,407,263

<표 11> 도서 자료 장르별 통계

장르	어절 수	장르	어절 수
건강의학	13,520,110	시집	1,643,696
과학	3,923,733	신화	769,919
교양	8,495,989	심리	534,116
기독교	822,506	역사	10,933,587
기타	54,348,232	영어	843,123
문화	6,027,327	영화	344,617
법률	839,672	요리	1,324,929
불교	1,380,352	음악	814,238
소설_역사	8,745,935	잡지	372,444
소설_일반	45,703,533	지식	4,705,155
수필	1,667,516	지혜	4,621,911
시O	1,032,816	컴퓨터	314,668
계			173,730,053

<표 12> 잡지 자료 장르별 통계

제목		어절 수	제목		어절 수
프OOO		90,976,514	우OOO		6,603,139
매OOOOO		70,107,812	ㅍOOO		6,320,907
시OOO		46,314,682	기독교잡지 6,115,736	기독교 사상	5,895,114
씨OOO		29,524,004		청어람 매거진	220,622
대학 학보 28,665,696	A대학	5,224,946	트OO		5,725,496
	B대학	7,450,907	행OOOOOO		5,380,007
	C대학	6,829,818	인터넷 언론 매체 4,941,436	A	3,411,282
	D대학	2,971,524		B	1,058,782
	E대학	6,188,501			471,372
지방지 19,123,996	A사	18,660,211		C	3,813,572
	B사	463,785			1,683,046
딴OOO		14,954,069	컴OOOO		3,813,572
과학잡지 14,290,268	과OOO	13,859,979	객O		1,683,046
	어OOOOOO	430,289	뉴OOO		1,292,360
이OOOOO		13,980,607	르OOOOOOOO		1,187,661
시OO		13,901,344	문OOO		1,098,865
월OOO		13,686,395	쎄O		698,304
레OOOO		7,307,010	올OO		580,241
계					408,273,167

<표 13> 블로그 자료 장르별 통계

장르		어절 수	장르		어절 수	장르		어절 수
N사	영화	9,305,123	N사	원예,재배	4,761,881	N사	생활공예	3,841,037
	음향,영상	9,141,098		가구,인테리어	4,642,122		캠핑	3,836,286
	스포츠	6,258,676		자필문학,에세이	4,634,411		사진	3,814,408
	요리	6,247,570		지역,해외생활	4,622,894		공연,전시,문화	3,804,911
	여행	5,787,029		자동차	4,436,957		차,커피,디저트	3,796,392
	게임	5,526,043		육아	4,334,085		웹툰,일러스트	3,787,565
	토이,모형,수집	5,515,000		등산,낚시,레저	4,294,118		미술,디자인	3,782,002
	책	5,465,840		일상	4,199,011		와인,술	3,770,214
	만화,애니	5,375,365		교육,외국어	4,142,060		패션,뷰티	3,768,092
	맛집	5,329,475		애완,반려동물	3,991,474	A사		75,370,881
	시사,인문,경제	5,219,988		철도,항공,교통	3,940,012			
	드라마,방송	5,214,048		과학,자연관찰	3,899,150	L사		1,082,524
	음악	4,834,591		IT,웹,프로그램	3,855,726			
계								245,628,059

1차년도의 어휘 등급화 작업에 반영하지는 못했으나, 1차년도 말미에 추가로 구축한 보완 언어 자료의 세부 내역은 다음과 같다.

<표 14> 보완 언어 자료의 구성

대부류	소부류	어절 수	비고
블로그	효O	1,808,330	
잡지	피OO	407,271	
	인OOO	17,393,051	
	교OOO	12,417,961	
인터뷰	C사 뉴OO	9,274,808	
	C사 시OOO	8,431,664	
	Y사 뉴OOOO	3,702,902	
	Y사 뉴OOOOO	5,020,996	
	Y사 당OOOOO	1,059,081	
	Y사 생OOO	3,325,544	
	Y사 수OOOOO	1,817,856	
	Y사 열OOOO	102,093	
	Y사 출OOOO	7,792,344	
	M사 뉴OOOO	826,989	
	M사 시OOO	8,536,501	
	M사 세OOOOO	6,965,041	
	S사 시OOOO	3,981,014	
	T사 뉴OOO	1,059,747	
	T사 색OOOO	1,274,939	
	K사 윤OOOOO	2,992,345	
	B사 아OOO	6,147,149	
	P사 열OOOOO	12,276,471	
	G사 세OOOO	160,486	
	Y사 만OOOOOO	3,263,644	
	P사 김OOOOOOO	846,316	
	위OOOO	1,258,702	
국회속기록	최근 회의록	7,261,047	
인터넷 게시판	D사-정치	382,642,212	
	클OO	53,386,055	
	루OO	20,081,251	
	보OOO	11,109,313	
	디OOOOO	9,035,067	
	엠OOOO	4,746,331	
	뽐O	53,548,664	
	오O	36,392,355	
라디오 드라마	K사	5,870,321	
영화·드라마 자막	씨OOO 영화	154,743,546	
	씨OOO 드라마	3,388,610	
	곰OO 드라마	26,570,174	
계		864,348,017	

1.1.2. 2018년에 추가 구축한 언어 자료 현황

2018년에 추가로 구축한 언어 자료 현황은 다음과 같다.

<표 15> 2018년 구축 언어 자료 개황

대부류	소부류	어절 수
청OOOOOO		26,262,567
자기소개서	잡OOO	16,049,948
	사OO	4,495,121
D사 영화 시놉시스		9,807,506
H사 여행기		1,689,079
가요 가사		4,077,767
강의		69,823
N사 지OO		83,843,784
N사 뉴스		1,332,026,152
법령		35,756,225
판례		84,273,770
홈쇼핑		60,912,159
계		1,659,263,901

이 중 대부류 'N사 뉴스'의 세부 내역은 다음과 같다.

<표 16> 'N사 뉴스' 자료 세부 내역

대부류	소부류	어절 수	어절 수 소계
100정치	264청와대	19,488,145	
	265국회/정당	47,062,182	
	266행정	10,610,295	
	267국방/외교	21,256,184	
	268북한	22,486,408	
	269정치일반	114,485,955	
소계			235,389,169
101경제	258증권	75,998,430	
	259금융	26,838,426	
	260부동산	26,454,528	
	261산업/재계	70,226,052	
	262글로벌경제	7,737,428	
	262글로벌경제2017	13,594,738	
	263경제일반	98,425,373	
	310생활경제	17,888,986	
	771중기/벤처	11,870,495	

소계			349,034,456
102사회	249사건사고	12,876,825	
	250교육	15,642,383	
	251노동	9,668,983	
	252환경	8,978,018	
	254언론	2,818,443	
	254언론2017	4,442,177	
	254언론2016	3,675,019	
	255식품/의료	9,366,764	
	256지역	138,168,538	
	257사회일반	78,993,453	
	276인물	8,648,390	
	59b인권/복지	4356378	
	59b인권/복지2017	5132759	
소계			302,768,130
103생활문화	237여행/레저	12,059,560	
	238음식/맛집	2,753,597	
	238음식/맛집2017	4,323,149	
	238음식/맛집2016	3,536,678	
	239자동차시승기	11,998,287	
	240도로교통	1,848,277	
	241건강정보	11,965,871	
	242공연/전시	11,046,548	
	243책	9,131,833	
	244종교	4,424,917	
	244종교2017	7,349,374	
	245생활문화일반	47,257,015	
	248날씨	7,287,017	
	376패션/뷰티	3,518,960	
	376패션/뷰티2017	5,296,158	
	376패션/뷰티2016	7,230,184	
소계			151,027,425
104세계	231아시아/호주	17,663,973	
	232미국/중남미	18,934,168	
	233유럽	8,027,479	
	233유럽2017	11,996,408	
	234중동/아프리카	3,363,930	
	234중동/아프리카2017	4,068,250	
	234중동/아프리카2016	3,742,346	
	322세계일반	16,301,507	
소계			84,098,061
105IT/과학	226인터넷/SNS	6,160,948	
	226인터넷/SNS2017	9,357,565	
	227통신/뉴미디어	7,371,230	
	227통신/뉴미디어2017	12,697,888	
	228과학일반	4,720,274	
	228과학일반2017	5,987,625	
	229게임리뷰	7,466,687	
	229게임리뷰2017	7,317,592	
	230IT일반	29,528,611	
	283컴퓨터	3,907,381	
	283컴퓨터2017	6,095,463	
	283컴퓨터2016	6,553,095	
	731모바일	3,602,856	

	731모바일2017	5,351,170	
	731모바일2016	5,874,579	
	732보안/해킹	2,191,498	
	732보안/해킹2017	3,565,826	
	732보안/해킹2016	3,376,698	
소계			131,126,986
오피니언	사설	39,192,031	
	칼럼	11,735,103	
소계			50,927,134
TV/연예		9,769,334	9,769,334
스포츠		17,885,457	17,885,457
계		2,636,397,513	27,654,791

각 소부류별로 일단 2018년 1월부터 9월까지의 자료를 수집하였다. 2018년도 자료만으로는 하나의 소부류의 분량이 타 소부류에 비해 현저하게 적을 경우, 2017년 자료도 수집하였고, 그것으로도 부족하면 2016년 자료까지 수집하였다.

이 중 대부류 'N사 지OO'의 세부 내역은 다음과 같다.

<표 17> 'N사 지OO' 자료 세부 내역

구별번호	대부류	어절 수
1	컴퓨터통신	8,533,281
3	엔터테인먼트,예술	4,462,563
4	경제	3,568,839
5	쇼핑	2,804,509
6	사회,정치	7,744,581
7	건강	4,891,967
8-1	가족,의식,취미	9,795,750
8-2	주거,교통	9,845,737
9	여행	1,362,952
10	스포츠,레저	889,458
11-1	교육	15,386,350
11-2	학문	13,873,372
12	지역,플레이스	270,801
13	쥬OO	413,624
계		83,843,784

위의 자료 가운데 구별번호 2번 '게임'은 특수한 단어가 많이 출현하기 때문에 제외하였으며, 8번 및 11번 범주는 타 범주에 비해 분량이 매우 커서 둘로 나누었다.

이상의 1차년도 구축분과 2차년도 구축분을 합치면 46억 어절(4,595,750,395 어절)에 육박한다. 이 가운데 인터넷 URL 등 어휘 통계와 관계없는 것들을 제외하면 4,528,211,008 어절이 된다.

1, 2차년도에 구축된 언어 자료 전체의 장르별 통계는 다음과 같다.

<표 18> 언어 자료 전체의 장르별 통계

장르번호	장르명	어절 수
1	세종_12_잡지	7,060,533
2	세종_130_책_총류	1,855,494
3	세종_132_책_교육	4,240,064
4	세종_134_책_체험	3,152,657
5	세종_135_책_인문	5,032,111
6	세종_136_책_사회	2,526,948
7	세종_137_책_자연	1,391,485
8	세종_138_책_예술	2,840,098
9	세종_기타출판물	1,813,489
10	세종_2_대화	2,494,383
11	도서_건강의학	13,217,761
12	도서_과학	3,845,028
13	도서_교양	8,272,791
14	도서_기독교	820,643
15	도서_기타	52,868,903
16	도서_문화	5,879,214
17	도서_법률	824,199
18	도서_불교	1,354,230
19	도서_소설_역사	8,497,313
20	도서_소설_일반	34,489,504
21	도서_수필	1,622,960
22	도서_시O	1,004,416
23	도서_시집	1,600,164
24	도서_신화	745,254
25	도서_심리	522,131
26	도서_역사	10,625,699
27	도서_영어	835,456
28	도서_영화	338,932
29	도서_요리	1,281,759
30	도서_음악	791,584
31	도서_잡지	361,974
32	도서_지식	4,573,892
33	도서_지혜	4,549,246
34	도서_컴퓨터	307,152
35	잡지_객O	1,683,036
36	잡지_과학	14,290,169
37	잡지_기독교	6,115,782
38	잡지_뉴OOO	1,292,364
39	잡지_대학학보	28,665,572
40	잡지_딴OOO	14,953,933
41	잡지_레OOOO	7,307,028

42	잡지_르OOOOOOOO	1,187,527
43	잡지_매OOOOO	70,097,918
44	잡지_문OOO	1,098,868
45	잡지_시OO	13,901,446
46	잡지_시OOO	46,314,922
47	잡지_쎄O	698,360
48	잡지_씨OOO	29,523,117
49	잡지_올OO	580,244
50	잡지_우OOO	6,602,965
51	잡지_월OOO	13,686,254
52	잡지_이OOOOO	13,980,649
53	잡지_매체	4,941,317
54	잡지_지방지	19,089,888
55	잡지_컴OOOO	3,813,536
56	잡지_트OO	5,725,791
57	잡지_ㅍOOO	6,320,779
58	잡지_프OOO	90,973,207
59	잡지_행OOOOOO	5,386,429
60	블로그_A사	75,374,783
61	블로그_IT웹프로그램	3,855,805
62	블로그_L사	1,082,179
63	블로그_가구인테리어	4,642,136
64	블로그_게임	5,526,047
65	블로그_공연전시문화예술	3,804,882
66	블로그_과학자연관찰	3,899,147
67	블로그_교육외국어	4,141,969
68	블로그_드라마방송	5,212,374
69	블로그_등산낚시레저	4,294,615
70	블로그_만화애니	5,374,517
71	블로그_맛집	5,329,410
72	블로그_미술디자인	3,781,794
73	블로그_사진	3,814,221
74	블로그_생활공예	3,841,036
75	블로그_스포츠	6,258,693
76	블로그_시사인문경제	5,219,989
77	블로그_애완반려동물	3,990,823
78	블로그_여행	5,787,122
79	블로그_영화	9,305,848
80	블로그_와인술	3,771,740
81	블로그_요리	6,247,523
82	블로그_원예재배	4,751,751
83	블로그_웹툰페인팅일러스트	3,787,699
84	블로그_육아	4,333,326
85	블로그_음악	4,834,009
86	블로그_음향영상	9,137,227
87	블로그_일상	4,198,880

88	블로그_자동차	4,437,129
89	블로그_자필문학에세이	4,634,383
90	블로그_지역해외생활	4,622,450
91	블로그_차커피디저트	3,796,663
92	블로그_책	5,465,875
93	블로그_철도항공교통	3,940,735
94	블로그_캠핑	3,836,568
95	블로그_토이모형수집	5,514,972
96	블로그_패션뷰티	3,768,103
97	드라마	45,184,666
98	신문_1990-2015	961,256,124
99	위OO	166,040,473
100	방송뉴스_에OO	18,919,129
101	인터뷰	91,908,199
102	인터넷게시판	577,929,468
103	영화드라마자막	154,713,793
104	국회속기록	7,261,085
105	라디오드라마	5,870,144
106	잡지_피OO	407,270
107	잡지_인OOO	17,393,452
108	잡지_교OOO	12,418,023
109	블로그_효O	1,808,337
110	청OOOOOO	26,261,513
111	자기소개서	20,544,994
112	영화시놉시스	9,807,438
113	여행기	1,689,060
114	가요	4,077,767
115	강의	119,282
116	지OO_컴퓨터통신	8,533,481
117	지OO_엔터테인먼트,예술	4,462,537
118	지OO_경제	3,568,693
119	지OO_쇼핑	2,804,508
120	지OO_사회,정치	7,744,555
121	지OO_건강	4,891,955
122	지OO_가족,의식,취미	9,795,708
123	지OO_주거,교통	9,845,839
124	지OO_여행	1,362,915
125	지OO_스포츠,레저	889,449
126	지OO_교육	15,385,323
127	지OO_학문	13,867,888
128	지OO_지역,플레이스	270,754
129	지OO_쥬OO	413,530
130	N사 뉴스_100정치_264청와대	19,488,300
131	N사 뉴스_100정치_265국회정당	47,062,329
132	N사 뉴스_100정치_266행정	10,610,438
133	N사 뉴스_100정치_267국방외교	21,256,365

134	N사 뉴스_100정치_268북한	22,486,623
135	N사 뉴스_100정치_269정치일반	114,486,997
136	N사 뉴스_101경제_258증권	74,584,500
137	N사 뉴스_101경제_259금융	26,838,543
138	N사 뉴스_101경제_260부동산	26,458,074
139	N사 뉴스_101경제_261산업재계	39,649,018
140	N사 뉴스_101경제_262글로벌경제	21,332,310
141	N사 뉴스_101경제_263경제일반	97,349,657
142	N사 뉴스_101경제_310생활경제	17,889,087
143	N사 뉴스_101경제_771중기벤처	11,870,605
144	N사 뉴스_102사회_249사건사고	12,876,861
145	N사 뉴스_102사회_250교육	15,643,114
146	N사 뉴스_102사회_251노동	9,669,103
147	N사 뉴스_102사회_252환경	8,977,579
148	N사 뉴스_102사회_254언론	10,935,610
149	N사 뉴스_102사회_255식품의료	9,366,852
150	N사 뉴스_102사회_256지역	138,170,954
151	N사 뉴스_102사회_257사회일반	77,694,590
152	N사 뉴스_102사회_276인물	8,648,775
153	N사 뉴스_102사회_59b인권복지	9,489,141
154	N사 뉴스_103생활문화_237여행레저	12,059,586
155	N사 뉴스_103생활문화_238음식맛집	10,614,044
156	N사 뉴스_103생활문화_239자동차시승기	11,998,383
157	N사 뉴스_103생활문화_240도로교통	1,848,290
158	N사 뉴스_103생활문화_241건강정보	11,965,944
159	N사 뉴스_103생활문화_242공연전시	11,046,895
160	N사 뉴스_103생활문화_243책	9,132,200
161	N사 뉴스_103생활문화_244종교	11,774,326
162	N사 뉴스_103생활문화_245생활문화일반	47,260,909
163	N사 뉴스_103생활문화_248날씨	7,287,040
164	N사 뉴스_103생활문화_376패션뷰티	16,045,382
165	N사 뉴스_104세계_231아시아호주	17,664,755
166	N사 뉴스_104세계_232미국중남미	18,934,322
167	N사 뉴스_104세계_233유럽	20,024,022
168	N사 뉴스_104세계_234중동아프리카	11,174,536
169	N사 뉴스_104세계_322세계일반	16,301,556
170	N사 뉴스_105IT과학_226인터넷SNS	15,518,532
171	N사 뉴스_105IT과학_227통신뉴미디어	20,069,203
172	N사 뉴스_105IT과학_228과학일반	10,708,376
173	N사 뉴스_105IT과학_229게임리뷰	14,784,279
174	N사 뉴스_105IT과학_230IT일반	29,529,459
175	N사 뉴스_105IT과학_283컴퓨터	16,555,938
176	N사 뉴스_105IT과학_731모바일	14,828,600
177	N사 뉴스_105IT과학_732보안해킹	9,134,011
178	N사 뉴스_스포츠	15,725,451
179	N사 뉴스_연예_TV연예	4,914,761

180	N사 뉴스_연예_영화	4,583,054
181	N사 뉴스_오피니언_사설	11,553,772
182	N사 뉴스_오피니언_칼럼	38,466,038
183	법령	35,755,774
184	판례	84,271,638
185	홈쇼핑	60,912,164
계		4,528,211,008

1.2. 기초 어휘 추출 목적의 언어 자료 보완

1.2.1. 줄 바꿈 문자의 처리

기초 어휘 추출을 위해 언어 자료에 대한 보완 작업이 필요하다. 구축된 언어 자료 중 「도서」 장르의 경우, 인쇄-출판된 책의 형태에서 줄이 바뀌는 위치에 줄 바꿈 문자(newline character, hard return)가 들어가 있는 경우가 많이 있다. 「도서」 언어 자료의 「건강/의학」 장르에 포함된 텍스트 일부를 보이면 다음 [그림 6]과 같은 식이다. '요리', '포식', '비롯하여', '뇌혈관계', '육류에', '두드러지게' 등의 어절이 줄 바꿈 문자에 의해 두 줄에 나뉘어 있다. 이 상태대로 처리하면 '요'와 '리', '포'와 '식', '비롯하'와 '여' 등이 각각 별도의 어절인 것처럼 처리된다.

이러한 오류를 수정하는 일은, 딥러닝으로 쉽게 해결할 수 있다. 줄 바꿈 문자를 만났을 때 이것이 진짜(내용상의) 줄 바꿈 문자인지, 아니면 인쇄된 책의 형태에 의해 야기된 가짜 줄 바꿈 문자인지 판정하는 과제인 셈인데, 이에 대해 정답 label이 부착되어 있는 훈련 데이터(training data)를 쉽게 만들 수 있으므로, 이 훈련 데이터를 바탕으로 신경망을 훈련시키면 매우 높은 정확도를 얻을 수 있다. 「도서」 장르에서 「신화」, 「잡지」, 「지혜」, 「요리」에 속하는 약 2천만 어절을 훈련 데이터로 하여 신경망을 1 epoch 동안 훈련시킨 결과 약 96.15%의 정확도를 얻었다. 더 많은 데이터를 가지고 더 많은 epoch 동안 훈련시키면 정확도가 더 올라가겠으나, 일단 현 수준에서 만족하고 훈련을 중지하였다. 이 훈련된 모델을 「도서」 장르의 모든 텍스트에 적용하여 수정을 완료하였다.

지금 우리나라는 바야흐로 식도락의 시대를 맞이하고 있다. TV의 채널을 돌려도 요리 프로그램이 줄을 잇고 있으며 잡지를 펼쳐 보아도 맛있는 음식을 만드는 전문음식점이나 식당에 대한 안내가 눈을 어지럽히고 있다. 프랑스요리를 비롯해서 세계 각국의 산해진미가 맛을 돋구고 있는데 실로<음식 문화의 전성기>에 접어들고 있다.
그리고 우리들의 식생활은 육식 중심주의의 서구형으로 크게 변해가고 있으며 또한 포식 시대를 맞이하고 있다. 이에 따라 사람의 생명을 빼앗는 질병의 종류도 크게 변해가고 있으며 최근 30년 동안 사망률 상위 3위를 계속차지하고 있는 것은 암을 비롯하여 협심증, 심근경색 등과 같은 심장부의 질환과 뇌혈전증, 뇌일혈 등과 같은 뇌혈관계 질환인 것이다. 그리고 해방 전에는 문제도 되지 않았던 혈액과 혈관에 관련된 이들 병에 의해 세 사람 중에 한 사람이 목숨을 빼앗기고 있는 것이 현실이다.
<식은 문화이다>라고도 할 수 있다. 육식을 중심으로 한 식도락을 만끽하기 위해서는 그것을 받아들일 수 있을만한 생활 체제와 지혜가 필요한 것이다. 예를 들어 육류에 있는 기름이나 지방분은 공기 속에서 산화되기 쉽고 과잉산화가 되어 과산화지질로 바뀐다. 이것이 동맥경화를 비롯하여 각종 성인병을 일으키고 있으며 뇌의 노화를 두드러지게 촉진시키는 원흉이 되고 있다.

[그림 6]「도서」장르의 줄 바꿈 문자 문제 예시

지금 우리나라는 바야흐로 식도락의 시대를 맞이하고 있다. TV의 채널을 돌려도 요리 프로그램이 줄을 잇고 있으며 잡지를 펼쳐 보아도 맛있는 음식을 만드는 전문음식점이나 식당에 대한 안내가 눈을 어지럽히고 있다. 프랑스요리를 비롯해서 세계 각국의 산해진미가 맛을 돋구고 있는데 실로<음식 문화의 전성기>에 접어들고 있다.
그리고 우리들의 식생활은 육식 중심주의의 서구형으로 크게 변해가고 있으며 또한 포식 시대를 맞이하고 있다. 이에 따라 사람의 생명을 빼앗는 질병의 종류도 크게 변해
가고 있으며 최근 30년 동안 사망률 상위 3위를 계속차지하고 있는 것은 암을 비롯하여 협심증, 심근경색 등과 같은 심장부의 질환과 뇌혈전증, 뇌일혈 등과 같은 뇌혈관계 질환인 것이다. 그리고 해방 전에는 문제도 되지 않았던 혈액과 혈관에 관련된
이들 병에 의해 세 사람 중에 한 사람이 목숨을 빼앗기고 있는 것이 현실이다.
<식은 문화이다>라고도 할 수 있다. 육식을 중심으로 한 식도락을 만끽하기 위해서는 그것을 받아들일 수 있을만한 생활 체제와 지혜가 필요한 것이다. 예를 들어 육류에 있는 기름이나 지방분은 공기 속에서 산화되기 쉽고 과잉산화가 되어 과산화지질로
바뀐다. 이것이 동맥경화를 비롯하여 각종 성인병을 일으키고 있으며 뇌의 노화를 두드러지게 촉진시키는 원흉이 되고 있다.

[그림 7]「도서」장르의 줄 바꿈 문자 변환 후

「홈쇼핑」장르도 마찬가지의 문제를 안고 있다. 딥러닝으로 가짜 줄 바꿈 문자를 판별하는 신경망을 훈련시킨 결과, 테스트 데이터에 대해 97.13%의 정확도를 얻었다. 이 신경망 모델로「홈쇼핑」장르를 모두 수정하였다.

-자, 여러분 어서 오세요.
꿈은 이루어진다.
그래서 저희가 드디어 꿈 같은 혜
택을 준비를 했습니다.
롯데홈쇼핑 어플이 더 새로워지
면서 오늘은요.
방송 상품을 사실 때마다 10%의
적립을 받아보실 수가 있습니다.
5만 원이면 5000원이겠죠?
자, 그래서 꿈은 이루어질 수 있
게끔 여러분의 적립금을 꿈처럼
쌓아드리도록 하겠습니다.
지금은 저희는 시원하게 인견 란
쥬와 팬티 시작하도록 하겠습니
다.

[그림 8]「홈쇼핑」 장르의 줄 바꿈 문자 문제 예시

-자, 여러분 어서 오세요.
꿈은 이루어진다.
그래서 저희가 드디어 꿈 같은 혜택을 준비를 했습니다.
롯데홈쇼핑 어플이 더 새로워지면서 오늘은요.
방송 상품을 사실 때마다 10%의 적립을 받아보실 수가 있습니다.
5만 원이면 5000원이겠죠?
자, 그래서 꿈은 이루어질 수 있게끔 여러분의 적립금을 꿈처럼 쌓아드리도록 하겠습니다.
지금은 저희는 시원하게 인견 란쥬와 팬티 시작하도록 하겠습니다.

[그림 9]「홈쇼핑」 장르의 줄 바꿈 문자 변환 후

1.2.2. 어문 규범 위반 사례의 처리

구축된 언어 자료에 들어 있는 대부분의 텍스트는 이미 어문 규범에 맞게 되어 있다. 다만「인터넷 게시판」과「N사 지OO」의 경우 어문 규범에 어긋나는 표기가 매우 자주 나타난다. 특히 띄어쓰기의 경우 이러한 경향이 매우 심하다.

우리집 주변에 동네분들이 길냥이 집이랑 화장실 만들고 먹이도 **주먄서** 돌보고 계시는데 원래 거기서 지내는 고양이가 딱 4마리로 **고정돼있었는데얼마전에** 좀 추워지기 시작할때 못보던 길냥이 **한마리가 나타나더니며칠안있어서** 길냥이 집 중 하나에 **자리잡고** 새끼를 **낳았더라구원래** 더 많았던것같은데 지금은 **네마리만** 남아있음ㅜ**남은애들은** 다행히 어미가 잘 돌보고 있는것같아 눈 아주 쪼금 **뜬것같음모처럼** 어미가 나와있길래 새끼 모습 **찍었는데옆에서 안절부절하고는있는데** 성질내거나 경계는 **안하더라추운날씨에 새끼낳아서** 안쓰럽다 그나마 **담요깔린 박스집 있고먹이도** 먹을수 **있는곳으로** 잘 찾아와서 다행이지만..**애기들** 잘 클수 있을지 걱정이네ㅜㅜ나중에 **젖뗄때까지** 얘들 건강하면 어미 **티엔알시키고애기들만이라도 무료분양 알아보는게** 좋을까..이궁 애기들 막 단춧구멍 눈뜨고 꼬물거리는거 넘 귀엽다맘같아선 젖떼면 **울집에 데려오고싶은데 사정이안되네** 휴치즈하나 올깜하나 **고등어두마리** 이렇게 **넷인듯**

[그림 10]「인터넷 게시판」 장르의「디OOOOO」의 글

이렇게 어문 규범에 어긋난 오류를 그대로 둔 채 형태소 분석기, 예컨대 UTagger로 분석하면 형태소 분석 오류가 많이 생기게 된다. 일반 언어 사용자의 어문 규범 오류는 그 자체로서 의미 있는 자료이고, 맞춤법 검사기, 맞춤법 교정기 등의 소프트웨어를 개발할 때 기초 자료로 이용될 수도 있다. 그러나 어휘 등급화를 목적으로 하는 본 연구에서는 통계를 왜곡시키는 노이즈로 작용한다고 볼 수 있다. 따라서 보다 타당성 높은 연구 결과를 얻기 위해서는 이러한 오류를 수정한 뒤에 통계를 도출하는 것이 바람직하다고 판단된다.

이러한 어문 규범상의 오류를 자동으로 수정해 주는 소프트웨어가 여럿 나와 있다. 예컨대 N사 맞춤법 검사기를 바탕으로 한 py-hanspell이라는 라이브러리로 위의 글을 처리하면 다음과 같이 된다.

우리 집 주변에 동네분들이 길냥이 집이랑 화장실 만들고 먹이도 **주면서** 돌보고 계시는데 원래 거기서 지내는 고양이가 딱 4마리로 **고정돼있었는데 얼마 전에** 좀 추워지기 시작할 때 못 보던 길냥이 **한 마리가 나타나더니 며칠 안 있어서** 길냥이 집 중 하나에 **자리 잡고** 새끼를 **낳았더라고 원래** 더 많았던 것 같은데 지금은 **네 마리만** 남아있음ㅜ**남은 애들은** 다행히 어미가 잘 돌보고 있는 것 같아 눈 아주 쪼금 **뜬 것 같음 모처럼** 어미가 나와있길래 새끼 모습 **찍었는데 옆에서 안절부절하고는 있는데** 성질내거나 경계는 **안 하더라 추운 날씨에 새끼 낳아서** 안쓰럽다 그나마 **담요 깔린 박스 집 있고 먹이도** 먹을 수 **있는 곳으로** 잘 찾아와서 다행이지만.. **아기들** 잘 클 수 있을지 걱정이네ㅜㅜ나중에 **젖뗄 때까지** 얘들 건강하면 어미 **티엔 알 시키고 아기들만이라도 무료 분양 알아보는 게** 좋을까.. 이 궁 아기들 막 단춧구멍 눈뜨고 꼬물거리는 거 너무 귀엽다 맘 같아선 젖떼면 **울 집에 데려오고 싶은데 사정이 안 되네** 휴치 조하나 올 검 하나 **고등어 두 마리** 이렇게 **넷인 듯**

[그림 11] 위의 글을 py-hanspell로 수정한 결과

완벽하지는 않지만, 입력 텍스트의 상당수 띄어쓰기를 교정해 줌으로써 형태소 분석시의 오류도 상당히 줄일 수 있고, 이는 언어 자료 전체의 통계 수치의 신뢰성

을 더욱 높이는 데 기여할 수 있다. 여러 맞춤법 검사기들을 테스트하여 본 결과, py-hanspell이 가장 성능이 좋아서, 이를 사용하여 「인터넷 게시판」과 「N사 지OO」의 모든 텍스트를 수정하였다.

1.2.3. 「신문」 장르 세분을 위한 실험

이미 구축된 언어 자료 가운데 「신문기사」는 시기상 1990년~2015년에 해당하며, 약 9억 어절 정도 되는데, 범주 구분이 전혀 되어 있지 않다. 「N사 뉴스」는 2018년(일부 범주는 2017년, 2016년도 포함) 기사들 약 13억 어절(약 728만 기사)을 수집하였는데, 53개의 범주로 구분되어 있다. 이 728만 개 기사를 훈련 데이터로 삼아, 임의의 신문기사를 입력하면 자동으로 53개 범주 중 어디에 속하는지 판단해 주는 신경망 모델을 만들 수 있다.

그런데 훈련 데이터의 양이 많아서 학습에 시간이 매우 많이 걸렸다. 이 데이터 전체를 컴퓨터 메모리에 탑재할 수 없기 때문에, 256개씩의 미니배치(mini-batch)를 load하여 학습시키는 방식을 취할 수밖에 없다. 이렇게 할 때 1 epoch당 90분 정도 소요된다.

딥러닝의 신경망 학습 시에 모델의 가중치를 매우 조금씩 수정하는데, optimizer가 가중치를 수정하는 방향을 알려 주는 역할을 한다. SGD라는 optimizer는 평범하게 입력 데이터의 모든 차원들 중 loss를 가장 많이 줄일 수 있는 방향으로 가중치를 변경하는 데 비해, Adam 등의 보다 진화된 optimizer는 momentum이라는 개념을 이용하여, 최소점(minimum)에 보다 빨리 도달할 수 있도록 하는 기능을 수행한다. 그러나 optimizer가 판단을 잘못하여 loss가 급격히 증가하는 방향으로 가중치가 변경될 위험도 있다.

「N사 뉴스」 기사의 범주화 모델 학습 시에도 바로 그런 문제가 발생하였다. 그래서 Adam 등의 보다 진화된 optimizer를 쓰는 것을 포기하고, SGD optimizer를 사용하여 학습을 진행하였다. SGD는 loss가 급격하게 증가하는 등의 불안정성은 거의 없으나, 학습의 진행 속도가 매우 느리다는 단점이 있다.

181 epoch까지 학습시킨 결과 약 42%의 정확도에 도달하였다. 첫 epoch 때 정확도가 약 15%였으므로, 1 epoch당 정확도가 평균 0.15% 포인트씩 향상된 셈이다. 정확도가 높아질수록 정확도의 향상 폭도 줄어드는 경향이 있으므로, 90% 이상의 정확도에 도달하려면 앞으로도 320 epoch (약 480시간 = 20일) 이상 더 훈련시켜야 한다는 계산이 나온다. 이 모델의 학습을 앞으로 틈나는 대로 진행하였으나, 올해의 작업에는 반영이 어려웠다. 2016~2018년 「N사 뉴스」 기사 자료는 53개의 장르로 구분하고, 1990년~2015년 「신문기사」 자료는 그러한 장르 구분 없이 하나의 장르로 간주하여 올해의 어휘 통계 작업을 진행하였다.

2. 언어 자료 처리 과정

언어 자료 처리 과정은 형태소 분석 이전 단계, 형태소 분석 단계, 형태소 분석 이후 단계로 구분하여 진행하였다.

2.1. 형태소 분석 이전 단계

형태소 분석이 이루어지기 이전의 언어 자료 처리 과정을 간략히 정리하면 다음과 같다.

① 줄 바꿈 문자 처리
② 어문 규범 위반 사례 처리
③ 장르 체계 수정/정비 및 이에 따른 파일 및 디렉토리 구조 정비

위의 ①~③에 대해서는 앞의 '1.2. 기초 어휘 추출 목적의 언어 자료 보완'에서 자세히 설명하였다.

2.2. 형태소 분석 단계

2.2.1. UTagger의 문제점과 한계

1차년도에는 언어 자료 통계 분석에 앞서 UTagger로 형태소 분석을 실시하였다. UTagger가 수행하는 작업은 다음의 세 단계로 나누어 생각할 수 있다.

① 제1단계: 분절(segmentation): 어절을 구성 형태소로 나누는 것.
학교에만은요 = 학교+ 에+ 만+ 은+ 요
흘렀겠더군요 = 흐르+ 었+ 겠+ 더+ 군+ 요
② 제2단계: 품사 부여(tagging, labeling): 각 형태소의 품사 판정.
학교에만은요 = 학교/NNG+ 에/JKB+ 만/JX+ 은/JX+ 요/JX
흘렀겠더군요 = 흐르/VV+ 었/EP+ 겠/EP+ 더군요/EF
③ 제3단계: 동형어 구분: 표준국어대사전의 어깨번호 부여.
사과를 = 사과__01/NNG+ 를/JKO
섰다 = 쓰__02/VV+ 었/EP+ 다/EF

어절을 구성 형태소로 분절하고 각 형태소에 품사를 붙여 줄 뿐 아니라 동형어의 경우 표준국어대사전의 동형어 번호까지 붙여 주는 기능은 UTagger 외의 다른 형태소 분석기에서는 제공하지 않으므로, 선택의 여지가 별로 없었다고 할 수 있다.

UTagger가 현재 나와 있는 형태소 분석기들 중 가장 낫기는 하나, 정확도에 아직 개선의 여지가 있다. 그런데 UTagger의 소스코드가 공개되어 있지 않아, UTagger 자체의 성능 개선을 제3자가 수행할 수 없게 되어 있다. 그래서 제2차년도에는 독자적인 형태소 분석기를 개발하여 UTagger를 대체할 수 있도록 하였다.

2.2.2. 형태소 분석 처리를 위한 훈련 데이터 구축

딥러닝을 통해 UTagger와 동일하게 제1단계~제3단계의 기능을 모두 수행하는 형태소 분석기를 만들기 위해서는, 이 세 단계에 해당하는 정보를 모두 포함하는 언어 자료가 필요하다. 세종 형태의미분석 말뭉치는 어절의 형태소로의 분절(제1단계), 형태소의 품사 부착(제2단계)뿐 아니라 각 형태소의 동형어 번호 부착(제3단계)까지 되어 있어서, 이 목적을 위해 사용하기에 가장 적합한 언어 자료이다.

그런데 국립국어원에서 배포하고 있는 세종 형태의미분석 말뭉치에는 많은 오류가 포함되어 있다. 예컨대 본래의 어절이 방언형 '얼매'를 포함하는 '얼매만입니꺼?'라면 형태소 분석 결과도 '얼매/NNG+만__01/NNB+이/VCP+ㅂ니꺼/EF+?/SF'와 같이 제시되어야 하는데 '얼마/NNG+만__01/NNB+이/VCP+ㅂ니꺼/EF+?/SF'와 같이 제시되어 있는 것이다.

언어 자료의 오류 중 빈번히 나타나는 몇 가지 유형을 들면 다음과 같다.

① 태그 오류: 문단 닫는 태그 를 로 잘못 입력.
② 문자코드 오류: 입력 어절과 분석 어절에 들어 있는 서로 대응하는 문장부호나 문자가 겉보기에는 비슷해 보이나 문자코드가 다른 경우. 예: “”와 ""
③ 분석 어절에서의 누락: 입력 어절이 많은 형태소로 이루어져 길이가 긴 경우, 분석 어절에서 뒷부분의 형태소들이 누락된 경우. 예: 국민연금관리공단(國民年金管理工團)에서는→국민/NNG+연금/NNG+관리__04/NNG+공단/NNG+(/SS+國民/SH+年金/SH+管理/SH+工團/SH+)/SS
④ 비일관적 분석: 예컨대 '더운'을 '덥+은'으로 분석한 경우도 있고 '덥+ㄴ'으로 분석한 경우도 있음. 또한 '(이)라도', '(이)며', '(이)니' 등은 긍정지정사의 활용형으로 보아 '이+라도'로 분석해야 하는 경우도 있고, 한 덩어리의 조사로 보아야 하는 경우도 있는데, 후자의 경우 두 가지 처리가 일관된 원칙에 따라 되어 있지 않음.
⑤ 불규칙한 비표준형 축약형의 과도한 원형 복원: 예컨대 '가부렀어'를 '가+아+버리+었+어'로 분석한 경우.

세종 형태의미분석 말뭉치 중 명백한 오류에 해당하는 것은 89,461 어절이다. 이는 총 약 1,200만 어절의 0.74%에 해당하는 수치이다. 한 어절 내에 복수의 오류가 들어 있는 경우도 꽤 있으므로, 오류 하나하나를 세면 이보다 더 많다. 그 외에도 관점에 따라 오류로 볼 수 있는 사례들까지 합산하면 훨씬 더 많은 오류가 있다. 이러한 오류를 수정하는 데 많은 시간과 노력이 소요되었다.

2.2.3. 분절, 품사 부착, 동형어 구분 모델 만들기

영어 같은 언어의 경우 띄어쓰기 단위로 구분된 각 토큰을 분석할 필요 없이 한 덩어리로 취급하여 품사만 부여하면 되기 때문에 딥러닝의 sequence labeling 방법을 이용하면 비교적 쉽게 품사 태거를 개발할 수 있는 데 비해, 한국어는 띄어쓰기 단위로 구분된 각 토큰(즉, 어절)을 우선 형태소로 분절(segment)해야 하기 때문에 난이도가 훨씬 높다. 즉 앞에서 구분한 제1~3단계 중 제1단계가 극복해야 할 중요한 난관이다.

제1단계 분절에서는 '흘렀겠더군요' 같은 입력 어절로부터 '흐르+었+겠+더+군+요' 같은 출력 결과를 얻어야 하는데, 이것이 그리 간단한 일은 아니다. 음절과 음절 사이에 형태소 경계가 들어가야 하는지를 결정해야 할 뿐 아니라, '흘러'→'흐르+어'처럼 입력의 2음절이 출력의 3음절이 되기도 하고 '가'→'가+아'처럼 입력의 1음절이 출력의 2음절이 되기도 하고 '간'→'가+ㄴ'처럼 입력 1음절의 초·중성과 종성을 둘로 쪼개야 하는 경우도 있다. 형태소와 형태소 사이의 경계가 음절/글자 경계와 일치하는 경우도 있지만, 하나의 음절을 초·중·종성으로 분리한 뒤 이들 자소 사이에 형태소 경계가 놓이는 일도 있고(예: 흐른=흐르+ㄴ, 흘렀다=흐르+었+다), 입력 어절에는 존재하지 않는 음절/글자를 추가해야 하는 경우도 있어(예: 친구다=친구+이+다) 문제를 더 복잡하게 한다.

그렇기는 하나, 입력 어절(글자 연쇄)과 형태소 분절이 완료된 출력 글자 연쇄를 광범위하게 비교·조사하면 입력 글자와 출력 문자열 사이의 대응 관계를 유형화할 수 있다. 이러한 조사를 한 결과 201개 유형이 있음을 알게 되었다.

딥러닝을 통해 형태소 분절 과제를 신경망에게 학습시킬 때에는 입력된 각 글자가 이 201개 유형 중 어느 것에 해당하는지를 판단하게 하면 된다. 이 201개 유형 중 일부를 보이면 다음과 같다.

0 입력과 출력이 동일. 경계 없음. 예: 빨0리0=빨리, 배0우1고0=배우+고
1 입력과 출력 글자가 동일하나 뒤에 경계 있음. 예: 먹1어0=먹+어0, 먹1었1다0=먹+었+다
2 ㅂ 불규칙 어간. 예: 더2워8=덥+어, 도2와8=돕+아
3 ㅅ 불규칙 어간. 예: 이3어0=잇+어, 이3으0면0=잇+으면
4 ㄷ 불규칙 어간. 예: 들4어0=듣+어, 들4으0면0=듣+으면
5 ㄹ 탈락 어간. 예: 가5는0=갈+는
6 러 불규칙 어미. 예: 이0르1러6서0=이르+어서
7 러 불규칙 어미. 뒤에 경계. 예: 이0르1렀7다0=이르+었+다
8 ㅂ 불규칙 어미. 예: 더2워8=덥+어, 도2와8=돕+아
9 ㅂ 불규칙 어간+매개모음어미1: 더9우128면0=덥+으면
10 초·중성과 종성 ㄴ 사이에 경계. 예: 간10=가+ㄴ, 먹1어0선10=먹+어서+ㄴ
……
180 의존명사, 계사 생략: 겝니다=거+이+ㅂ니다
181 잖=하+잖
182 계사 생략, 인용 축약: 랄=이+라+하+ㄹ
183 계사 생략, 인용 축약: 랄=이+라+하+ㄹ+
184 하 ㅎ 탈락, 사이에 경계: 진/친=하+지+ㄴ, 질/칠=하+지+ㄹ
185 의존명사 거: 겔=거+이+ㄹ
186 계사 생략: N+ㄴ+
187 인용 하 생략: 더=하+더
188 인용 축약: 댄=다+하+ㄴ, 랜=라+하+ㄴ. 잰=자+하+ㄴ, 낸=냐+하+ㄴ
189 잖=하+잖+
190 잖=하+지+않
191 잖=하+지+않+
192 '하' 불규칙: ㅐ>엏+어,사이에 경계 그래=그렇+어, 어때=어떻+어
193 '하' 불규칙: ㅐ>엏+어,사이에 경계, 뒤에 경계 그래=그렇+어+, 어때=어떻+어+
194 '하' 불규칙 과거형: 앴>엏+었+ 그랬=그렇+었+, 어땠=어떻+었+
195 '하' 불규칙 과거형: 앴>엏+었+ 그랬=그렇+었, 어땠=어떻+었
196 의문사: 뉘=누구+이+
197 인용 하 생략: 는=하+는
198 인용 하 생략: 는=하+는+
199 하 축약 탄=하+다+ㄴ
200 인용 하 생략: 셔=하+시+어+

[그림 12] 형태소 분절 시의 201개 음절 변화 유형

분절, 품사부착, 동형어구분의 세 단계로 나누어, 각각 딥러닝을 통해 LSTM(Long Short-Term Memory) 신경망 모델을 만들고 학습시켰다. 이 3가지 신경망 모형 각각의 테스트 데이터에 대한 성능(정확도)은 다음과 같다.

① 분절 모형: 글자 단위로 count했을 때 99.79%
② 품사부착 모형: 형태소 단위로 count했을 때 99.89%
③ 동형어 구분 모형: 형태소 단위로 count했을 때 99.89%

그러나 하나의 어절 내에 오류가 하나라도 있으면 그 어절에 대한 분석이 실패한 것으로 간주하고, 하나의 어절에 대한 분절·품사부착·동형어 구분이 완전히 들어맞아야 분석이 성공한 것으로 간주하는 식으로 계산하면, 위의 수치보다는 성능이 약간 떨어진다.

세종 형태의미분석 말뭉치의 약 92만 6천개 문장 중 앞부분의 1만개 문장(약 16만 어절)을 가지고 어절 단위로 정확도를 테스트한 결과 96.53%(158,041/163,720)의 정확도를 보였다. 신문 사설 텍스트에 대한 형태소 분석 결과의 일부를 보이면 다음과 같다.

올림픽/NNG 참가__01/NNG+·/SP+남북/NNP 대화__06/NNG 공개__02/NNG 제안__02/NNG //SP '/SS+핵/NNG 개발/NNG 시간__04/NNG 벌__02/VV+기/ETN+'/SS 노림수/NNG 의심__03/NNG //SP 한/NNP+·/SP+미/NNP+동맹__01/NNG 강화__04/NNG 어느/MM 때__01/NNG+보다/JKB 중요__02/NNG
북한/NNP 김정/NNP+은/JX 노동당/NNP 위원장/NNG+이/JKS 어제__01/NNG 신년사/NNG+에서/JKB 평창/NNP 동계__01/NNG+올림픽/NNG+에/JKB 대하__02/VV+아/EC "/SS+대표단/NNG 파견__01/NNG+을/JKO 포함__02/NNG+하/XSV+아/EC 필요/NNG+하/XSA+ㄴ/ETM 조치__04/NNG+를/JKO 취하__01/VV+ㄹ/ETM 용의__01/NNG+가/JKS 있/VV+으며/EC 이/NP+를/JKO 위하__01/VV+아/EC 북남/NNP 당국__02/NNG+이/JKS 시급히/MAG 만나/VV+ㄹ/ETM 수__02/NNB+도/JX 있/VV+을/ETM 것/NNB+"/SS+이라고/JKQ 말__01/NNG+하/XSV+았/EP+다/EF+./SF "/SS+진정__05/NNG+으로/JKB 민족/NNG+적/XSN 화해__02/NNG+와/JC 단합/NNG+을/JKO 원하__02/VV+ㄴ다면/EC 남조선/NNP+의/JKG 집권__01/NNG+여당__01/NNG+은/JX 물론/MAG 야당/NNG+들/XSN+,/SP 각계각층/NNG 단체__02/NNG+들/XSN+과/JC 개별/NNG+적/XSN 인사__01/NNG+들/XSN+을/JKO 포함__02/NNG+하/XSV+아/EC 그/MM 누구/NP+에게/JKB+도/JX 대화__06/NNG+와/JKB 접촉/NNG+,/SP 내왕/NNG+의/JKG 길__01/NNG+을/JKO 열__02/VV+어/EC+놓/VX+을/ETM 것/NNB+"/SS+이라고/JKQ+도/JX 하/VV+았/EP+다/EF+./SF 평창/NNP+올림픽/NNG+을/JKO 계기__04/NNG+로/JKB 남북/NNP 대화__06/NNG+와/JC 교류__01/NNG+를/JKO 재개__04/NNG+하/XSV+겠/EP+다는/ETM 의사__02/NNG+를/JKO 공개__02/NNG+적/XSN+으로/JKB 밝히/VV+ㄴ/ETM 것/NNB+이/VCP+다/EF+./SF
북한/NNP+의/JKG 평창/NNP+올림픽/NNG 참가__01/NNG+는/JX 우리/NP+로서/JKB+ㄴ/JX 환영__02/NNG+하/XSV+ㄹ/ETM 일__01/NNG+이/VCP+다/EF+./SF 북한/NNP 도발/NNG 우려/NNG+로/JKB 인하__01/VV+ㄴ/ETM 각국/NNG 선수단__02/NNG 안전__03/NNG+문제__06/NNG+에/JKB 청신호/NNG+가/JKS 켜__01/VV+어/EC+지/VX+었/EP+다는/ETM 사실__04/NNG+만/JX+으로/JKB+도/JX 다행/NNG+스럽/XSA+ㄴ/ETM 일__01/NNG+이/JKC 아니/VCN+ㄹ/ETM 수__02/NNB 없/VA+다/EF+./SF 남북/NNP+대화__06/NNG 제__06/NNG+의/JKG 역시/MAG 마찬가지/NNG+이/VCP+다/EF+./SF 하지만/MAJ 김정/NNP+은/JX 신년사/NNG+에/JKB+는/JX 평화__02/NNG+적/XSN 신호__01/NNG+보다/JKB+는/JX 평화__02/NNG+를/JKO 위협/NNG+하/XSV+는/ETM 섬뜩/MAG+하/XSA+ㄴ/ETM 발언__02/NNG+들/XSN+이/JKS 더/MAG 많/VA+다/EF+./SF

[그림 13] 신문 사설 형태소 분석 결과

UTagger와의 비교는 다음과 같다.

2	27	“/SS+ 진정__05/NNG+ 으로/JKB	“/SS+ 진정__05/NNG+ 으로/JKB
#2	28	민족/NNG+ 적/XSN	민족적/NNG
2	29	화해__02/NNG+ 와/JC	화해__02/NNG+ 와/JC
2	30	단합/NNG+ 을/JKO	단합/NNG+ 을/JKO
$2	31	원하__02/VV+ ㄴ다면/EC	원__15/NNG+ 하/XSV+ ㄴ다면/EC
2	32	남조선/NNP+ 의/JKG	남조선/NNP+ 의/JKG
2	33	집권__01/NNG+ 여당__01/NNG+ 은/JX	집 권 _ _ 0 1 / N N G + 여 당 __01/NNG+ 은/JX
&2	34	물론/MAG	물론__01/MAG
2	35	야당/NNG+ 들/XSN+ ,/SP	야당/NNG+ 들__09/XSN+ ,/SP
2	36	각계각층/NNG	각계각층/NNG
^2	37	단체__02/NNG+ 들/XSN+ 과/JC	단체__02/NNG+ 들__09/XSN+ 과/JKB
#2	38	개별/NNG+ 적/XSN	개별적/NNG
2	39	인사__01/NNG+ 들/XSN+ 을/JKO	인사__01/NNG+ 들__09/XSN+ 을/JKO
2	40	포함__02/NNG+ 하/XSV+ 아/EC	포함__02/NNG+ 하/XSV+ 여/EC
2	41	그/MM	그__01/MM
2	42	누구/NP+ 에게/JKB+ 도/JX	누구/NP+ 에게/JKB+ 도/JX
2	43	대화__06/NNG+ 와/JKB	대화__06/NNG+ 와/JKB
2	44	접촉/NNG+ ,/SP	접촉/NNG+ ,/SP
2	45	내왕/NNG+ 의/JKG	내왕/NNG+ 의/JKG
2	46	길__01/NNG+ 을/JKO	길__01/NNG+ 을/JKO
2	47	열__02/VV+ 어/EC+ 놓/VX+ 을/ETM	열 _ _ 0 2 / V V + 어 / E C + 놓 __01/VX+ 을/ETM
2	48	것/NNB+ ”/SS+ 이라고/JKQ+ 도/JX	것__01/NNB+ ”/SS+ 이라고/JKQ+ 도/JX
2	49	하/VV+ 았/EP+ 다/EF+ ./SF	하__01/VV+ 였/EP+ 다/EF+ ./SF

<범례>

! : '뚜렷/XR+하/XSA' 대 '뚜렷하/VA'처럼 'X하-' 같은 용언의 분석 차이

@ : '초/XPN+일류/NNG' 대 '초일류/NNG' 같은 접두사 분석 차이

: '과학/NNG+성/XSN' 대 '과학성/NNG' 같은 접미사 분석 차이

(이상은 쉽게 상호 변환 가능하므로 어느 쪽도 오류로 간주하지 않음.)

$: (위의 경우 이외의) 형태소의 수 차이

% : 표기 형태의 차이 ('하여/해'를 '하+아' 또는 '하+여'로 하는 것과 같은 차이는 무시)

^ : 품사 차이 (있'을 VA 또는 VV로 하는 차이는 무시)

& : 동형어번호 차이 ((준)문법적 요소, '있', '없' 등의 동형어번호 차이는 무시)

[그림 14] UTagger와의 비교

&로 표시된 경우의 상당수는, UTagger의 분석 결과에는 번호가 있고, 우리 형태소 분석기의 분석 결과에는 번호가 없는 경우인데 우리 형태소 분석기의 오류/미달 분석인 경우도 있지만 품사 정보 등을 바탕으로 충분히 추론 가능한 경우도 많다. 앞으로, 분석 결과를 보면서 후처리 모듈 추가, 사전을 참조하여 분석을 수정하거나 보완하는 모듈 추가 등을 고려하고 있다.

새로 개발된 형태소 분석기는 이미 UTagger와 대등하거나 조금 나은 성능을 보이고 있기는 하나, 속도가 매우 느리다는 것이 문제이다. 45억 어절이 넘는 언어 자료를 모두 분석하려면 꽤 많은 시간이 소요된다. 따라서, 형태소 분석기의 성능 향상을 위한 노력은 앞으로도 틈나는 대로 계속 하되, 올해의 작업에는 2017년처럼 UTagger를 사용하였다. 2017년에 이미 UTagger로 형태소 분석을 해 놓은 자료를 최대한 활용하되, 줄 바꿈 문자나 어문 규범 오류를 수정한 언어 자료는 UTagger로 다시 분석을 실시하였다.

2.2.4. 모델의 오류 모니터링 및 최적화

언어 자료를 훈련 데이터로 하여 딥러닝을 통해 형태소 분석기를 만들면, 형태 분석의 중의성(ambiguity)이 있는 경우 그 해결의 단서를 문맥으로부터 자동으로 추출하여 상당히 정확하게 중의성 해소를 할 수 있다는 장점이 있다. 인간이 해결의 단서들을 찾아서 일일이 코딩해 주어야 했던 과거의 규칙 기반 방식에 비해 훨씬 더 효율적이면서도 높은 성능을 거두고 있다.

그런데 이 방식은 언어 자료에 존재하는 패턴을 최대한 끌어낼 수 있다는 장점이 있는 반면에, 만약 언어 자료에 오류나 문제점이 존재한다면 이를 바탕으로 만들어진 시스템도 이러한 오류나 문제점을 고스란히 떠안는다는 한계가 있다.

현재 세종 형태의미분석 말뭉치는 수많은 오류를 수정했음에도 불구하고, 여전히 많은 문제점이 남아 있다. 예컨대 '-이라도' 또는 '-라도'라는 문자열은 긍정 지정사 '-이다'의 활용형으로서 '이/VCP+라도/EC'로 태깅되어야 하는 경우도 있고(예: 그는 외국인이라도 한국어를 잘 한다) 한 덩어리의 보조사(JX)로 태깅되어야 하는 경우도 있다(예: 꿩 대신 닭이라도 잡아야겠다). 그런데 현재 세종 형태의미분석 말뭉치에서는 '-이라도'나 '-라도'의 대다수가 긍정 지정사의 활용형으로 태깅되어 있다. 이에 대한 전수조사를 통해 보조사의 예는 품사 태그를 수정해야 하는데, 이러한 작업은 아직 극히 일부만 이루어져 있는 상태이다.

현재 형태소 분석기의 오류를 체계적으로 모니터링하여, 언어 자료를 수정할 것은 수정하고, 후처리 모듈에 넣어야 할 것은 넣는 식으로 오류에 대처해 나아가야 한다. 이러한 작업은 2차년도뿐 아니라 그 이후에도 지속적으로 이루어질 예정이다.

2.2.5. 추후 작업 계획

현재의 형태소 분석기는 구글에서 만든 기계학습 라이브러리인 텐서플로우(Tensorflow)를 백엔드(back-end)로, 파이썬으로 작성된 오픈소스 신경망 라이브러리인 케라스(Keras)를 프런트엔드(front-end)로 하여 개발되어 있다.[18] 이에 대

해 몇 가지 추가 고려사항이 있다.

우선 현재의 텐서플로우(Tensorflow)+ 케라스(Keras) 프레임워크(framework, 개발 환경)를 유지하면서 마스킹 레이어(Masking layer)를 추가할 필요가 있다. 딥러닝 신경망 학습 시에는 모든 입력 문장의 길이가 동일하게 조정되어야 한다. 예컨대 분절 모델에서 입력 문장의 글자 수를 120개로 설정하면, 이보다 길이가 더 긴(글자 수가 더 많은) 문장의 경우 120개를 초과하는 글자들은 삭제되며, 글자 수가 120개에 미달하는 문장의 경우 패딩(padding)이라 불리는 무의미 글자를 채워 넣어서 120개를 맞춘다. 입력 문장 중에는 글자 수가 120개에 현저히 미달하는 것이 상당수이므로, 신경망 학습에 입력되는 문장에는 매우 많은 수의 패딩이 들어가게 된다. 당연히 이 패딩은 신경망이 쉽게 판단할 수 있으므로, 신경망의 성능을 측정하는 손실 함수(loss function)가 이 패딩까지 포함해서 계산을 하게 되면, 성능이 실제보다 높게 평가되는 결과를 낳는다. 정확도가 99.7~99.8% 정도에 이르게 되면 더 이상 올리기가 매우 어렵게 되는데, 패딩이 계산에 포함되어 쉽게 이러한 수치에 도달해서 성능 향상이 더 이상 이루어지지 않는다면, 이는 바람직하지 않다. 따라서 손실 함수가 패딩을 계산에 넣지 않도로 하는 조치가 필요한데, 이것이 바로 마스킹 레이어가 하는 역할이다.

텐서플로우(Tensorflow)는 딥러닝 프레임워크로서 널리 쓰이고 매우 우수한 것은 사실이나, 애초부터 태생적으로 지니고 있는 문제점이 있다. 텐서플로우는 신경망 모델을 만들어 훈련에 들어가기 전에 미리 신경망 모델에 대한 연산 그래프(computation graph)를 정적(static)으로 완전히 구성한다. 그래서 입력 문장의 길이가 미리 하나의 수치로 정해져야 하는 것이다. 반면에 파이토치(Pytorch)는 연산 그래프를 그때그때 동적(dynamic)으로 구성한다. 그래서 가변 길이 입력(variable length input)에 더 유연하게 대처할 수 있다. 예컨대 훈련 데이터로서 100만개 문장이 있다고 하면, 이 100만개 문장 전체에 대해 하나의 길이를 정해 놓는 것이 아니라, 각 배치(batch, 데이터 입력 묶음)마다 입력 문장의 길이가 달라도 된다. 배치의 크기가 128이라면 이 128개의 문장끼리만 일정한 길이를 맞추면 된다. 형태소분석기에 들어가는 입력 문장들은 길이가 매우 가변적이고 그 차이가 크므로, 이에 유연하게 대처할 수 있는 파이토치(Pytorch)가 장점이 있는 것이다. 텐서플로우(Tensorflow)+ 케라스(Keras)를 기반으로 만들어진 기존 형태소분석기를 앞으로 파이토치(Pytorch)+ 알렌엔엘피(AllenNLP)를 기반으로 다시 만들 계획이다. 또한 파이썬(Python)으로 개발한 형태소 분석기는 속도가 느리므로, 이를 C++로 포팅하여 속도 향상을 도모한다.

18) 프런트엔드(front-end)와 백엔드(back-end)는 소프트웨어의 구조를 가리키는 말인데, 프런트엔드는 사용자로부터 다양한 형태의 입력을 받아 백엔드가 사용할 수 있는 규격으로 처리하고, 백엔드는 프런트엔드에서 입력된 명령을 처리한다.

2.3. 형태소 분석 이후 단계

형태소 분석 이후의 언어 자료 처리 과정을 간략히 정리하면 다음과 같다.

① 형태소 분절에 대한 후처리
② 형태소에 부여된 품사에 대한 후처리
③ 장르별 절대 빈도 추출
④ 각 장르의 규모를 고려한 상대 빈도 산출
⑤ 각 단어의 범위와 산포도 산출
⑥ 빈도, 범위, 산포도 이외의 추가적인 지표 산출
⑦ 언어 자료로부터 추출된 통계 지표를 바탕으로 한 어휘 점수 산출

위의 ③~⑦에 대해서는 'Ⅳ. 어휘 등급화의 통계적 방법론 수립'에서 자세히 설명할 것이다. 여기에서 위의 ①과 ②에 대해서 살펴보겠다.

2.3.1. 형태소 분절에 대한 후처리

이 사업에서 개발된 형태소 분석기는 철저하게 세종 형태의미분석 말뭉치를 바탕으로 하여 귀납적으로 만들어졌기 때문에, 어떤 문자열을 어떤 방식으로 분절할 것인가 하는 문제에 대해 철저하게 이 언어 자료의 처리 방식을 따르게 된다. 예컨대 '공부한다'를 '공부+하+ㄴ+다'(1안)로 분절할 수도 있고, '공부+하+ㄴ다'(2안)로 분절할 수도 있고, '공부하+ㄴ+다'(3안)로 분절할 수도 있고, '공부하+ㄴ다'(4안)로 분절할 수도 있는데, 세종 형태의미분석 말뭉치는 2안을 따르고 있기 때문에, 형태소 분석기도 어쩔 수 없이 2안을 따르게 된다.

이러한 분절의 문제는 이 외에도 어미와 인용 조사('다고' 대 '다+고'), 선어말어미와 어말어미('더라' 대 '더+라'), 어미와 격조사('기에' 대 '기+에'), 긍정 지정사와 어미('이라고' 대 '이+라고', '이라도' 대 '이+라도', '이니' 대 '이+니') 등 다양한 경우에 발생할 수 있다.

이러한 문제에 대해 세종 형태의미분석 말뭉치의 처리가 나름대로 일리가 있으나 모든 경우에 해당하는 것은 아니다. 언어 자료를 사용하고자 하는 목적에 따라 얼마든지 다른 처분을 선호할 수도 있다. 예컨대 '공부하-'를 '공부+하'로 분절하면 동사 '공부하-'에 포함된 '공부'와 명사 '공부'(예컨대 '공부가 제일 쉬웠다'의 '공부')가 함께 명사로 처리되어 명사 '공부'의 빈도에 반영되고 동사 '공부하-'는 따로 기록되지 않게 된다. 이런 처리는 명사 '공부'와 동사 '공부하-'를 묶어서 처리하는 word family의 방식과 비슷해지는 셈이 된다. 통사적인 측면보다는 어휘적인 측면

에 초점을 맞추는 연구에서는 이 방식이 좋다고 할 수 있다.

반면에 통사적인 측면에 초점을 맞추는 연구에서는 명사 '공부'와 동사 '공부하-'의 통사적 행태가 매우 다르므로 이 둘을 따로 처리하는 것을 선호할 수 있다. 이렇게 처리하려면, 형태소 분석기가 내놓은 출력을 약간 가공해야 한다.

처리하는 것의 범위를 정해야 한다. 'X하-'뿐 아니라 'X되-'도 이렇게 처리하자는 데 많은 사람들이 동의할 것 같지만, 'X시키-', 'X받-', 'X당하-' 등은 어떻게 할 것인가에 대해서는 사람들의 견해가 갈릴 가능성이 있다. 이런 문제에 대해 하나하나 지침을 정하여 언어 자료 후처리를 해야 한다.

올해의 작업에서는 UTagger의 기본 설정에서 정한 대로 '공부하-' 등을 한 덩어리의 용언으로 처리하였다.

2.3.2. 형태소에 부여된 품사에 대한 후처리

각 형태소에 어떠한 품사를 부여할 것인가 하는 문제도 역시 세종 형태의미분석 말뭉치의 처분을 형태소 분석기가 그대로 따르게 된다. 예컨대 '~다가 ~다가 하-'의 '하'를 동사(VV)로 태깅할 수도 있고(1안) 보조용언(VX)로 태깅할 수도 있는데(2안), 세종 형태의미분석 말뭉치는 이 문제에 대해 배포된 버전에서는 일관성 없이 1안과 2안이 혼재되어 있었다. 이 사업에서는 수작업을 통해 2안으로 통일시켰다. 따라서 형태소분석기도 이 경우의 '하'를 보조용언으로 태깅한다.

동일한 형태의 용언이 본용언으로도 쓰이고 보조용언으로도 쓰일 때, 형태소 분석기는 입력 문장에 들어 있는 이 용언이 본용언으로 쓰인 것인지 보조용언으로 쓰인 것인지 나름대로 구별하여 판정을 내린다. 이 구별을 그대로 가져다 쓰는 것이 좋을 때도 있고, 이 구별을 없애고 하나로 합쳐서 처리하는 것이 좋을 때도 있다. 만약 후자로 결정한다면, 언어 자료에 대한 약간의 후처리가 필요하게 된다.

이러한 품사 통용어의 처리 문제(품사를 구별할 것인가 구별 안 하고 합칠 것인가)는 본용언과 보조용언뿐 아니라 명사-부사(예: 오늘), 대명사-부사(예: 여기, 언제), 동사-형용사(예: 크다, 지나치다) 등 다양하게 제기된다. 이에 대해서도 하나하나 지침을 마련하여 세심하게 결정해야 하고, 이 결정대로 언어 자료 후처리가 이루어져야 한다. 올해의 사업에서는 이러한 지침 마련 작업과 언어 자료의 통계적 처리가 각각 이루어졌다.

Ⅳ. 어휘 등급화의 통계적 방법론 수립

1. 장르별 어휘 통계 추출

장르별 어휘 통계를 추출하는 과정은 다음과 같다.

○ 형태소분석기가 어느 정도 안정화 단계에 접어들면, 이를 사용하여 43억 어절 언어 자료 전체를 형태소 분석할 것이다.
○ 형태소 분석된 언어 자료를 바탕으로 각 장르별로 어휘 형태소(조사, 어미 제외)의 절대빈도를 추출한다.
○ 각 장르의 규모를 고려하여 각 형태소의 상대빈도를 추출한다. 1차년도의 경우 각 장르의 크기를 1천만 어절로 설정하여 환산하였다. 2차년도에도 각 장르의 크기를 고려하여 평균에 가까운 수치를 사용할 예정이다.
○ 각 형태소에 대해, 이것이 출현하는 장르의 수, 즉 범위(range)를 산출한다.
○ 각 형태소에 대해, 이것이 각 장르에서 얼마나 골고루 출현하는지, 즉 산포도(dispersion)를 산출한다.
○ 상대빈도, 범위, 산포도 이 3가지가 어휘 점수를 산출하는 기반이 된다.

2. 어휘 점수 산출을 위한 지표 개발 및 정교화

어휘의 중요도/기본도(basicness)를 빈도, 범위, 산포도의 3가지 변수를 바탕으로 산출하는 것은 Nation 교수 및 그를 중심으로 하는 학파에서 제안한 이론이다. 세계의 영어 교육계에서는 이 이론이 매우 널리 사용되고 있다.

그러나 어휘의 기본도를 측정하는 방법에 이것만 있는 것은 아니다. 예컨대 J. B. Carroll, P. Davies and B. Richman (eds.) (1971), The American Heritage Word-Frequency Book (Boston: Houghton Mifflin)에서 개발하고 프랑스 MANULEX 프로젝트에서 발전시킨 어휘 지표는 다음과 같다.

지표 F: 언어 자료에서의 raw frequency

지표 D: 각 학년별 교재에서의 단어 분포. dispersion이라고 부르나, Nation 학파에서 말하는 dispersion과는 계산 방식이 다르다.

$$D = \left[\log\left(\sum p_i\right) - \left[\left(\sum p_i \log p_i\right) / \sum p_i\right]\right] / \log(n)$$

(n: 학년별 교재의 수, i: 각 교재를 가리키는 index,
p_i: 해당 교재에서의 해당 단어의 출현 빈도 F를 바탕으로 한 확률)

지표 U: D에 의해 백분위수를 계산한 빈도

$$U = \left(1{,}000{,}000 / N\right)\left[FD + \left(1 - D\right) * f_{\min}\right]$$

(N: 언어 자료의 총 단어 수, F: 언어 자료에서의 해당 단어 빈도,
f_{min}: f_i와 s_i의 곱을 N으로 나눈 값, f_i: 교재 i에서의 빈도, s_i: 그 교재의 단어 수)

여기서 '교재'라고 한 것을, 우리 작업에서는 '장르'로 대치하면 된다. 이 U 값에 따라 단어들을 배열하였을 때 1위~50위까지를 보이면 다음과 같다.

<표 19> U 값에 따라 소팅된 결과 (1위~50위)

순위	단어	빈도	범위	산포도	D	Fmin	U
1	이/VCP	6.263524e+07	185	97.901647	3.306229	1.383223e-02	4.573250e+04
2	것__01/NNB	2.935206e+07	185	97.250537	3.161035	6.482043e-03	2.048996e+04
3	있__01/VA	2.349591e+07	185	97.658056	3.118406	5.188785e-03	1.618074e+04
4	하__01/VV	2.303348e+07	185	96.222510	3.114598	5.086662e-03	1.584291e+04
5	들__09/XSN	2.293627e+07	185	97.013502	3.113788	5.065195e-03	1.577194e+04
6	있__01/VX	1.702836e+07	185	97.910607	3.056735	3.760505e-03	1.149487e+04
7	수__02/NNB	1.379084e+07	185	97.472548	3.016340	3.045538e-03	9.186377e+03
8	되__01/VV	1.279957e+07	185	96.646520	3.002051	2.826630e-03	8.485686e+03
9	하__01/VX	1.104825e+07	185	97.148067	2.973865	2.439870e-03	7.255845e+03
10	않/VX	9.250350e+06	185	97.365518	2.939843	2.042827e-03	6.005589e+03
11	없__01/VA	8.330225e+06	185	96.648931	2.919773	1.839628e-03	5.371297e+03
12	등__05/NNB	8.076319e+06	185	94.150374	2.913843	1.783556e-03	5.197004e+03
13	년__02/NNB	8.049318e+06	185	95.517585	2.913202	1.777593e-03	5.178488e+03
14	보__01/VV	7.733004e+06	185	96.400334	2.905522	1.707739e-03	4.961875e+03
15	주__01/VX	7.519042e+06	185	93.716919	2.900148	1.660488e-03	4.815661e+03
16	이__05/MM	7.493935e+06	185	96.802853	2.899507	1.654944e-03	4.798521e+03
17	일__07/NNB	7.453266e+06	185	92.365927	2.898464	1.645963e-03	4.770764e+03
18	그__01/MM	6.881215e+06	185	94.264658	2.883167	1.519632e-03	4.381353e+03
19	아니/VCN	5.740243e+06	185	96.514357	2.848439	1.267662e-03	3.610859e+03
20	같/VA	5.636001e+06	185	96.177301	2.844928	1.244642e-03	3.540917e+03
21	대하__02/VV	5.409056e+06	185	95.685319	2.837055	1.194524e-03	3.388930e+03
22	나__03/NP	5.344994e+06	185	87.486799	2.834773	1.180377e-03	3.346100e+03
23	때__01/NNG	5.341417e+06	185	96.265684	2.834645	1.179587e-03	3.343709e+03
24	사람/NNG	5.286161e+06	185	94.856139	2.832653	1.167384e-03	3.306793e+03
25	가__01/VV	5.239848e+06	185	92.936488	2.830967	1.157156e-03	3.275871e+03
26	보__01/VX	5.064950e+06	185	93.976392	2.824464	1.118532e-03	3.159254e+03
27	위하__01/VV	4.872595e+06	185	96.260518	2.817047	1.076053e-03	3.031292e+03
28	받__01/VV	4.801091e+06	185	96.378062	2.814216	1.060262e-03	2.983806e+03
29	그__01/NP	4.780389e+06	185	92.351564	2.813388	1.055691e-03	2.970067e+03
30	한__01/MM	4.660812e+06	185	97.039539	2.808535	1.029283e-03	2.890779e+03
31	월__02/NNB	4.559293e+06	185	94.807422	2.804317	1.006864e-03	2.823566e+03
32	지__04/VX	4.351881e+06	185	96.800604	2.795398	9.610596e-04	2.686544e+03
33	좋__01/VA	4.349589e+06	185	93.608846	2.795297	9.605536e-04	2.685032e+03
34	이__05/NP	4.001529e+06	185	95.327322	2.779320	8.836888e-04	2.456054e+03
35	기자__05/NNG	3.995195e+06	185	89.512766	2.779017	8.822900e-04	2.451899e+03
36	원__01/NNB	3.892006e+06	185	91.423919	2.774004	8.595019e-04	2.384262e+03
37	더__01/MAG	3.781318e+06	185	96.510850	2.768477	8.350577e-04	2.311838e+03
38	우리__03/NP	3.683200e+06	185	94.278802	2.763441	8.133897e-04	2.247754e+03
39	및/MAG	3.606642e+06	185	89.887405	2.759417	7.964829e-04	2.197829e+03
40	만__06/NR	3.529538e+06	185	91.028800	2.755278	7.794554e-04	2.147616e+03
41	만들/VV	3.488443e+06	185	93.822135	2.753034	7.703800e-04	2.120883e+03
42	많/VA	3.466735e+06	185	96.947704	2.751839	7.655860e-04	2.106769e+03
43	중__04/NNB	3.454154e+06	185	96.743851	2.751142	7.628076e-04	2.098592e+03
44	때문/NNB	3.435845e+06	185	96.241444	2.750124	7.587643e-04	2.086696e+03
45	말하/VV	3.389152e+06	185	95.535981	2.747503	7.484527e-04	2.056376e+03
46	사진__07/NNG	3.263222e+06	185	90.376681	2.740250	7.206427e-04	1.974741e+03
47	제__21/XPN	3.235897e+06	184	61.389731	2.738639	7.146083e-04	1.957054e+03
48	명__03/NNB	3.177283e+06	185	92.938267	2.735137	7.016640e-04	1.919147e+03
49	알/VV	3.049444e+06	185	94.349359	2.727271	6.734324e-04	1.836632e+03
50	뉴스/NNG	3.011591e+06	183	87.190021	2.724878	6.650730e-04	1.812243e+03

3. 빈도, 범위, 산포도의 가중치 산출을 위한 실험

Nation 학파에서 빈도, 범위, 산포도를 어휘 점수 산출의 지표로 제안하기는 했으나, 이 지표들을 어떻게 조합하여 최종적인 어휘 점수를 산출할 것인지에 대해서는, 모든 경우에 통용될 수 있는 해답이 나와 있는 것은 아니다.

복수의 지표들을 종합하는 가장 상식적인 방법은, 각 지표들의 선형 함수(linear function)로 표상하는 것이다. 즉 각 지표에 가중치를 곱하여 더하는 것이다(즉 weighted sum). 이때 모든 가중치의 합은 1이 된다. 문제는 가중치를 어떻게 얻어낼 것인가이다.

가중치들을 다양하게 설정하여 어휘 점수를 산출하여 이 점수에 따라 단어들을 배열한 뒤, 이 배열 순서가 모어 화자들의 단어 기본도에 대한 직관과 얼마나 일치하는지를 알아볼 수 있다. 모어 화자들은 단어 기본도에 대해 어느 정도 정확한 직관을 가지고 있다고 가정할 수 있고, 이러한 직관에 개인차가 있겠지만 충분한 수의 화자들로 표본을 구성하여 평균을 내면, 어느 정도 객관성이 있는 단어 순위를 얻을 수 있다. 이것을 기준으로 삼아, 어휘 점수의 타당도(validity)를 확인할 수 있는 것이다.

이러한 직관 실험은 대학생 146명을 피험자로 선정하여 실시하였다. 1차년도에 얻은 어휘 점수에 따라 1위~3만위의 단어들 중 적절히 50개의 구간을 설정하여, 각 구간에서 1개씩 총 50개의 단어를 추출하였다. (순위가 높을 때는 구간의 크기를 작게 하고, 순위가 낮아질수록 구간의 크기를 크게 한다. 가장 기본적인 단어들에 대해서는 모어 화자들의 직관이 더 뚜렷한 반면에, 순위가 내려갈수록 모어 화자들의 직관도 흐릿해지기 때문이다.) 이러한 50 단어 목록을 10개 추출하였다.

이 10개의 50 단어 목록을 146명의 피험자에게 배분하여 한 사람이 2개의 목록을 담당하게 하였다. 예컨대 주민등록상 생년월일의 끝자리에 따라 다음과 같이 배분하였다.

끝자리 1: 1번, 6번 파일
끝자리 2: 2번, 7번 파일
끝자리 3: 3번, 8번 파일
끝자리 4: 4번, 9번 파일
끝자리 5: 5번, 10번 파일
끝자리 6: 6번, 5번 파일
끝자리 7: 7번, 4번 파일
끝자리 8: 8번, 3번 파일
끝자리 9: 9번, 2번 파일
끝자리 0: 10번, 1번 파일

피험자에게 목록 안의 50 단어를 직관에 따라 기본도의 순으로 배열하게 하였다. 같은 단어 목록을 사용한 피험자의 응답을 종합하여(단, 가나다순 거의 그대로 제출한 무성의한 답변은 제외하였음.) 평균을 구하였다.

50 단어의 빈도, 범위, 산포도와 실험을 통해 얻은 등위 평균을 하나의 파일로 결합하였다. 1번 파일만 보이면 다음과 같다.

<표 20> 1번 50 단어 목록

	단어	빈도	범위	산포도	D	Fmin	등위 평균
1	가리__03/VV	81428.27840106464	182	95.85593396384424	2.033258089537666	1.798243903766965e-05	28.7
2	고래__01/NNG	29898.716212608848	179	89.82270162575338	1.841334951860237	6.602765674962303e-06	42.67741935483871
3	교외__01/NNG	11397.477825284326	170	90.97833632380753	1.656592093 6661145	2.516993533461311 4e-06	15.96666666 6666667
4	교체하/VV	64057.90490708519	180	84.76522418653197	1.987296629 6788825	1.414640457211776 4e-05	25.83333333 3333332
5	구리__02/NNG	32266.94403809571	176	89.7343786150165	1.855936975 345215	7.125759815761595e-06	22.83333333 3333332
6	균열__02/NNG	16393.367272611926	171	91.815385610689	1.726220437 6106058	3.620274594900 06635e-06	15.66666666 6666666
7	그것/NP	1904928.8580998515	185	92.370934972499	2.637139785 4378307	0.00042068023215667506	43.2
8	그냥/MAG	766456.8704883244	182	90.52580185539033	2.462741421 8150667	0.00016926262250902696	42.7
9	기어이/MAG	7036.715738330812	149	87.92554325739944	1.564213229 694216	1.553972578994007e-06	22.96666666 6666665
10	당상__01/NNG	2055.1289177682656	125	77.2314692976431	1.328443292 0826519	4.538500776879578e-07	5.766666666 666667
11	대__06/NNB	469900.7225693587	183	94.50635755169303	2.369020418 651829	0.00010377182550441756	23.83333333 3333332
12	되__01/VV	12799574.81928756	185	96.64652012885242	3.002051173 4561506	0.00282662950 04085555	37.23333333 3333334
13	등록증/NNG	16517.807330000539	167	74.7342448884643	1.727669038 6066988	3.647755659094363e-06	23.96666666 6666665
14	만들/VV	3488443.1725429776	185	93.82213499988335	2.753034354 9929928	0.00077037999 47440474	40.83870967 741935
15	멘__88/NNG	15788.47326423476	173	92.63232827113343	1.719018498 4829433	3.486691153822385e-06	13.6
16	목__10/NNG	19155.16183484891	170	78.48419155098767	1.756045111 1781434	4.230183134356 47e-06	39.33333333 3333336
17	미물__03/NNG	1530.2001733272127	99	86.20176668406965	1.271945223 6589885	3.379259868022503e-07	8.733333333 333333
18	바로__02/MAG	2111821.7108506057	185	95.14076491197841	2.656890600 415351	0.00046636998 7423211	34.9
19	부수__12/NNG	7286.1795008330755	156	82.45955108460524	1.570886680 504229	1.609063598838607e-06	11.8
20	북상하/VV	10424.702634577468	140	58.17502261111315	1.639502483 781489	2.302168034166 28e-06	10.6
21	빛바래/VA	3262.987226506507	142	87.52008738733221	1.417001340 3765682	7.205908074384742e-07	16.06666666 6666666
22	뽀뽀/NNG	7017.622410999357	148	82.27562182406614	1.563692752 8935078	1.549756051253 2011e-06	44.43333333 333333

23	서류__02/NNG	138310.47812888908	180	85.13448235463358	2.134741284196301	3.0544176913252424e-05	27.566666666666666
24	선생님/NNG	287126.52578950964	181	90.11139135906122	2.274659328086023	6.340838032552869e-05	43.633333333333333
25	속하__02/VV	180757.82658293203	183	90.5445398471298	2.186013107439767	3.9918154490500117e-05	27.448275862068964
26	수리하__02/VV	19327.356235915733	173	80.46278156318775	1.757759417127797	4.268210161092328e-06	27.666666666666668
27	시달리/VV	68841.08400451383	181	94.44722381869643	2.0010913367594645	1.5202711155220493e-05	21.966666666666665
28	쓰리/VA	9940.280506881	171	86.31528558529895	1.6303875873390195	2.1951893340039775e-06	25.533333333333335
29	아침저녁/NNG	7808.040910315638	142	77.77952609627327	1.5841376458152745	1.7243103063265285e-06	37.1
30	알리/VV	1015455.2054697061	185	83.4016492864768	2.5166292888588506	0.000224250858379987	34.6
31	염두__01/NNG	46863.197441112196	181	94.71442807712793	1.9274243174669363	1.0349163799639934e-05	11.933333333333334
32	왜냐하면/MAG	108460.68661134382	180	85.93952021138236	2.0881710434365854	2.3952215658661796e-05	34.266666666666666
33	음흉하/VA	3222.462851042211	128	86.72128306967592	1.4146074086976671	7.116414949190925e-07	14.666666666666666
34	이름/NNG	1051724.0729279108	184	94.08725069283193	2.5233517811072245	0.00023226039402091193	46.533333333333333
35	이봐/IC	17252.958445358476	127	47.213233103694684	1.7360103477500 4	3.810104788597006e-06	33.733333333333334
36	장만하/VV	18566.91660880159	175	88.58994027558624	1.7500702489631 63	4.100276373163 57e-06	19.466666666666665
37	적당하__02/VA	110059.59796939508	183	90.85192902621328	2.0909743479127942	2.4305315669908614e-05	27.6
38	절대__05/NNG	69139.86466782891	182	90.24224289848344	2.0019209275629417	1.5268693209234146e-05	31.933333333333334
39	종이__01/NNG	138982.85516207648	183	93.01627104207728	2.1356702595 30703	3.0692663154728825e-05	43.266666666666666
40	주민/NNG	377066.334387188	183	90.40838695732847	2.3268584806652823	8.327048667145 24e-05	28.7
41	주석__01/NNG	52506.794479750046	166	71.23185731186108	1.9492063413426675	1.1595483158136e-05	8.03448275862069
42	중__01/NNG	30528.145740865588	180	91.44528277930603	1.8453257767 22984	6.741767485422266e-06	20.5
43	질의__01/NNG	39292.62562065119	178	65.23341265980869	1.8936726322293296	8.677295636451753e-06	13.033333333333333
44	천직__01/NNG	2227.504986180541	132	87.76896310761595	1.3438720371407145	4.919172234344211e-07	9.5666666666666666
45	촬영하/VV	144325.7814830553	178	78.9014829899194	2.1428963061021937	3.187258306383573e-05	22.9
46	크나크/VA	7721.372053661105	166	90.22462784742727	1.5819994748540998	1.7051705496982673e-06	25.1
47	퇴사__04/NNG	16883.237976851357	154	60.74951280570935	1.7318607572754503	3.728456546531004e-06	14.033333333333333
48	파고__01/NNG	10402.623955507288	164	83.89019078385364	1.6390963495351671	2.2972922280187366e-06	2.8333333333333335
49	할배/NNG	7669.804521321884	134	64.9188087459697	1.5807158561293693	1.6937824910923153e-06	32.866666666666667
50	효율적/NNG	101792.64352162858	180	89.8329675096 2572	2.0760167120 85653	2.2479659923486624e-05	15.266666666666667

1차년도에 했던 것처럼, 가중치의 여러 조합에 따른 단어 순위와 피험자가 응답한 단어 순위를 비교하여, Spearman 상관계수를 구하여, 가장 상관관계가 높은 가중치 조합을 고르는 방법도 가능하다.

그러나 위의 방법은 무수히 많은 가중치의 다양한 조합들을 모두 조사할 수 없다는 단점이 있다. 그래서 3개의 가중치 값을 random하게 설정한 뒤, 가중치에 따른 어휘 점수 및 단어 순위와 피험자가 응답한 단어 순위와의 차이를 측정하는 손실함수를 최소화하는 경사하강법(gradient descent)을 사용하여 가중치 값을 얻는 것이 더 정밀한 방법이다. 10개의 50 단어 목록에 대해 각각 3번씩 경사하강법을 실시하여 얻은 가중치는 다음과 같다.

<표 21> 경사하강법으로 얻은 세 변수의 가중치

목록	반복	구분	빈도	범위	산포도
1	1	초기값	0.020643484413166213	0.49506325006659313	0.48429326552024066
1	1	결과값	0.3646846763880179	0.7070288544107577	-0.07171353079877552
1	2	초기값	0.24650774401480957	0.225205778417927	0.5282864775672634
1	2	결과값	0.36608161609462425	0.7095306642222291	-0.07561228031685324
1	3	초기값	0.08707331406859331	0.3875796743218586	0.5253470116095481
1	3	결과값	0.36350356291645697	0.7061022378739067	-0.06960580079036362
2	1	초기값	0.30774938566662924	0.5892973615457417	0.10295325278762901
2	1	결과값	0.23941053683422134	0.5572004279556256	0.2033890352101529
2	2	초기값	0.2452056890387455	0.2478606054959709	0.5069337054652836
2	2	결과값	0.24526962279387096	0.5516676511470674	0.2030627260590617
2	3	초기값	0.10360137029073246	0.23936000589822293	0.6570386238110446
2	3	결과값	0.2397135502201397	0.5565814883464874	0.20370496143337286
3	1	초기값	0.6067231579479004	0.2835206109292253	0.10975623112287425
3	1	결과값	0.4539397446033022	0.41457283856263494	0.13148741683406287
3	2	초기값	0.001331785017107734	0.7197363637987775	0.2789318511841148
3	2	결과값	0.45241108169975347	0.4159115367485596	0.13167738155168698
3	3	초기값	0.4183035353751394	0.05017141561692351	0.5315250490079371
3	3	결과값	0.4547312798894476	0.41414958739250013	0.1311191327180523
4	1	초기값	0.8596136254047705	0.01757679021203784	0.12280958438319167
4	1	결과값	0.371021480389144	1.0423202424747366	-0.41334172286388055
4	2	초기값	0.09554405087901374	0.3823228649703543	0.522133084150632
4	2	결과값	0.3740790103743298	1.0554761045409575	-0.42955511491528736
4	3	초기값	0.22042130675747662	0.5343595328122006	0.24521916043032277
4	3	결과값	0.37425781502494304	1.050041541256702	-0.42429935628164483
5	1	초기값	0.6150129038763305	0.2065711359902137	0.17841596013345584
5	1	결과값	0.4002405673663959	0.9425362182280965	-0.3427767855944924
5	2	초기값	0.06503213401532648	0.3400178625827386	0.5949500034019349
5	2	결과값	0.4000108405108005	0.9385467654534684	-0.3385576059642689
5	3	초기값	0.256387882346681	0.6058101957791069	0.13780192187421214
5	3	결과값	0.39763862613260775	0.9393406904225692	-0.3369793165551771
6	1	초기값	0.26453227890245046	0.4839106157612547	0.25155710533629483
6	1	결과값	0.41064425835951335	0.8349426772176859	-0.24558693557719916
6	2	초기값	0.019315564719607825	0.2978180979627033	0.6828663373176889
6	2	결과값	0.40615106950577773	0.833509935721195	-0.2396610052269726
6	3	초기값	0.3430243271824174	0.547986161377674	0.10898951143990865
6	3	결과값	0.41444345695657303	0.8374217498295318	-0.25186520678610474
7	1	초기값	0.4923191800977199	0.07059936275587475	0.43708145714640534
7	1	결과값	0.3146049526581701	0.8072832656681724	-0.12188821832634249
7	2	초기값	0.5222155670169149	0.22363924090561504	0.2541451920774701

7	2	결과값	0.3132979777175549	0.8111012535883579	-0.124399231305913
7	3	초기값	0.643627106363124	0.1468227276140679	0.20955016602280807
7	3	결과값	0.31597301732571115	0.809187007460866	-0.1251600247865771
8	1	초기값	0.03601907664670345	0.4445918758459987	0.5193890475072979
8	1	결과값	0.03513902329035066	0.6394850339346563	0.3253759427749931
8	2	초기값	0.4402211008008571	0.30663377806970904	0.2531451211294339
8	2	결과값	0.04754451906137011	0.6394699441552981	0.31298553678333174
8	3	초기값	0.4696454509273139	0.2801990449553796	0.25015550411730647
8	3	결과값	0.035674971638847516	0.639732474442751	0.3245925539184014
9	1	초기값	0.051223064754320036	0.4658801500739711	0.48289678517170886
9	1	결과값	0.16768948298689276	0.8084028243023805	0.023907692710726833
9	2	초기값	0.1753117611425693	0.5235839068974592	0.30110433195997155
9	2	결과값	0.17238087090769635	0.800400703043825	0.027218426048478683
9	3	초기값	0.4043122502568567	0.48066182921226375	0.11502592053087957
9	3	결과값	0.1688969788001558	0.8075617041570806	0.02354131704276344
10	1	초기값	0.38759020144974565	0.46601607229926467	0.14639372625098968
10	1	결과값	0.28082273650892214	0.5366919489258095	0.1824853145652683
10	2	초기값	0.19324745471727933	0.005010812663831299	0.8017417326188894
10	2	결과값	0.28040891938123 77	0.5349133249271788	0.18467775569158373
10	3	초기값	0.7411487290017387	0.11645998229556243	0.1423912887026989
10	3	결과값	0.2827390562087232	0.5351943393742215	0.18206660441705538

10개의 목록 각각을 보면, 3번의 시행을 비교해 볼 때, 초기값이 어떻게 설정되든 상관없이 결과값이 거의 비슷하게 수렴함을 알 수 있다.

경사하강법에 따르면 가중치가 음수가 될 수도 있다. 그래서 산포도에 대한 가중치가 음수가 된 경우가 있다. 수학적으로는 얼마든지 가능한 일이다. 즉 해당 변수가 높을수록 전체 점수가 낮아지는 관계인 것이다. 그런데 Nation 학파의 이론에 따르면, 산포도가 높을수록 여러 장르에서 골고루 나타나므로 단어의 기본도와 양의 상관관계가 있을 것으로 기대할 수 있다. 따라서 산포도에 대한 가중치가 상당히 큰 음수 값을 갖게 된 4, 5, 6번 목록의 경우는 실험 결과의 타당성에 영향을 줄 것으로 추정되어 그 후의 계산에서 제외하였다. 남은 7개의 단어 목록을 가지고 평균을 낸 결과는 다음과 같다.

빈도: 0.266425　　범위: 0.638199　　산포도: 0.0953768

1차년도의 간이 실험을 통해 얻은 가중치 0.2, 0.7, 0.1과 크게 다르지는 않은 수치이다. 계산의 편의를 위해 이 세 가중치를 0.265, 0.64, 0.095로 하여 그 후의 통계 처리를 실시하였다.

4. 빈도, 범위, 산포도와 가중치를 바탕으로 한 어휘 점수 산출

앞에서 살펴본 방법에 따라 빈도, 범위, 산포도에 대해 얻어진 가중치를 이용하여 어휘 점수를 산출하고, 이 점수에 따라 단어들을 배열하여 단어 순위를 얻었다. 1위~50위를 제시하면 다음과 같다.

<표 22> 빈도, 범위, 산포도의 가중치에 따른 순위 (1위~50위)

순위	단어	빈도	범위	산포도	weighted sum
1	이/VCP	6.263524e+07	185	97.901647	47.507826
2	것__01/NNB	2.935206e+07	185	97.250537	24.539074
3	있__01/VA	2.349591e+07	185	97.658056	20.474817
4	하__01/VV	2.303348e+07	185	96.222510	20.138308
5	들__09/XSN	2.293627e+07	185	97.013502	20.078605
6	있__01/VX	1.702836e+07	185	97.910607	15.960139
7	수__02/NNB	1.379084e+07	185	97.472548	13.678752
8	되__01/VV	1.279957e+07	185	96.646520	12.970070
9	하__01/VX	1.104825e+07	185	97.148067	11.732968
10	않/VX	9.250350e+06	185	97.365518	10.451565
11	없__01/VA	8.330225e+06	185	96.648931	9.783010
12	등__05/NNB	8.076319e+06	185	94.150374	9.574612
13	년__02/NNB	8.049318e+06	185	95.517585	9.568944
14	보__01/VV	7.733004e+06	185	96.400334	9.349242
15	이__05/MM	7.493935e+06	185	96.802853	9.180167
16	주__01/VX	7.519042e+06	185	93.716919	9.167161
17	일__07/NNB	7.453266e+06	185	92.365927	9.105815
18	그__01/MM	6.881215e+06	185	94.264658	8.709217
19	아니/VCN	5.740243e+06	185	96.514357	7.895680
20	같/VA	5.636001e+06	185	96.177301	7.815308
21	대하__02/VV	5.409056e+06	185	95.685319	7.642398
22	때__01/NNG	5.341417e+06	185	96.265684	7.598109
23	사람/NNG	5.286161e+06	185	94.856139	7.542844
24	나__03/NP	5.344994e+06	185	87.486799	7.511993
25	가__01/VV	5.239848e+06	185	92.936488	7.489031
26	보__01/VX	5.064950e+06	185	93.976392	7.369406
27	위하__01/VV	4.872595e+06	185	96.260518	7.248935
28	받__01/VV	4.801091e+06	185	96.378062	7.196631
29	그__01/NP	4.780389e+06	185	92.351564	7.140417
30	한__01/MM	4.660812e+06	185	97.039539	7.098173
31	월__02/NNB	4.559293e+06	185	94.807422	6.999329
32	지__04/VX	4.351881e+06	185	96.800604	6.863156
33	좋__01/VA	4.349589e+06	185	93.608846	6.829151
34	이__05/NP	4.001529e+06	185	95.327322	6.582526
35	기자__05/NNG	3.995195e+06	185	89.512766	6.518905
36	원__01/NNB	3.892006e+06	185	91.423919	6.459506
37	더__01/MAG	3.781318e+06	185	96.510850	6.426251
38	우리__03/NP	3.683200e+06	185	94.278802	6.328380
39	및/MAG	3.606642e+06	185	89.887405	6.225065
40	많/VA	3.466735e+06	185	96.947704	6.188428
41	만__06/NR	3.529538e+06	185	91.028800	6.177132
42	중__04/NNB	3.454154e+06	185	96.743851	6.176628
43	만들/VV	3.488443e+06	185	93.822135	6.173618
44	때문/NNB	3.435845e+06	185	96.241444	6.157367
45	말하/VV	3.389152e+06	185	95.535981	6.114028
46	사진__07/NNG	3.263222e+06	185	90.376681	5.963909

47	명__03/NNB	3.177283e+06	185	92.938267	5.922699
48	알/VV	3.049444e+06	185	94.349359	5.836702
49	크__01/VA	2.984442e+06	185	95.083587	5.792935
50	뉴스/NNG	3.011591e+06	183	87.190021	5.734515

앞서 <표 19>에 제시한, MANULEX의 U 값에 따른 1위~50위 목록과 비교하면, 몇몇 단어의 등위에 약간의 차이가 있을 뿐, 대동소이함을 볼 수 있다. 이는 두 가지 단어 점수 산출 방식이 (적어도 최고 등위 단어의 경우에는) 수렴함을 보여 준다.

이렇게 얻어진 단어 순위는 순전히 언어 자료 기반으로 통계적 방법론에 입각해서 얻어진 것이므로, 언어 전문가의 눈으로 보면 미흡한 점이 눈에 띌 수 있다. 그러한 미흡한 점은 통계적 방법론 자체에서 기인하는 것도 있을 수 있고, 기반 언어 자료가 지닌 한계에서 비롯된 것도 있을 수 있다. 미흡한 점을 찾아내고 그 원인을 알아내야, 추후의 과제도 알 수 있다. 통계적 방법론에서 기인한다고 판단되면 방법론상의 개선점을 모색해야 하고, 언어 자료에서 기인한다고 판단되면 언어 자료의 보완점을 모색해야 할 것이다. 예컨대 의존명사 '월(月)'(31위), '일(日)'(17위), 일반명사 '사진'(46위), '뉴스'(50위), '기자'(35위), '무단'(110위), '배포'(113위), '전재(轉載)'(126위), 고유명사 '뉴시스'(144위), '연합뉴스'(190위), '페이스북'(357위) 등이 상위에 랭크된 것은, 뉴스 및 신문 기사가 우리의 언어 자료에서 매우 큰 비중을 차지하기 때문이라고 할 수 있다.

양적 방법론에 따라 도출된 어휘 목록에 대한 정성적 검토를 위해서는 언어 전문가의 직관에 입각한 검토뿐 아니라, 다른 어휘 목록과의 비교도 생각할 수 있다. 국어 교육계, 한국어 교육계 등에서 개발한 등급별/연령별/학년별 어휘 목록, 기존 사전의 수록 어휘 목록 등과의 체계적 비교를 통해서도 언어 자료 기반 어휘 목록의 문제점을 발견할 수 있을 것이다.

Ⅴ. 어휘 등급화의 정성적 방법론 수립

1. 정성적 분석의 검토 항목

앞서 논의한 빈도, 분포, 산포도를 고려하여 형태소 분석기를 통해 등급화된 어휘가 그대로 기초 어휘의 선정으로 이어지는 것은 아니다. 통계적 분석으로는 개별 어휘들이 지닌 특성과 어휘 간의 관계들에 대해 정교한 분석이 어렵기 때문에 이들 어휘에 대한 정성적 분석이 필요하다. 더불어 기초 어휘의 예비 후보에 대한 검토 작업도 이루어져야 할 것이다. 한편 기초 어휘 목록의 교육적 활용 가능성을 고려해 교과서 어휘와의 비교 검토 작업을 실시하기로 한다.

정성적 분석을 위해 검토해야 하는 항목은 우선 크게 두 가지로 나눌 수 있다. 하나는 동형어라고 부를 수 있는 것들의 존재를 어떻게 처리할 수 있느냐 하는 것이다. 동형어란 같은 형태를 가지지만 같은 어휘가 아니라 다른 어휘로 처리될 수 있는 어휘들을 말한다. 여기에서 파생되는 쟁점들을 몇 가지로 정리하면 다음과 같다.

(1) 가. 품사 통용의 처리
　　나. 본용언과 보조 용언의 처리
　　다. 어근 어휘와 파생어 어휘의 처리
　　라. 상위 품사와 하위 품사의 처리

(1가)는 가령 ‘오늘’이라는 어휘가 명사와 부사로 쓰인다고 할 때 이를 통합하여 하나의 어휘로 처리할 것인지 아니면 서로 다른 어휘로 처리할 것인지 하는 문제와 관련된다. 앞의 것은 어휘에 대한 다의어 처리를 의미하고 뒤의 것은 어휘에 대한 동음이의어 처리를 의미한다.

(1나)는 용언이 본용언으로 쓰일 경우와 보조용언으로 쓰일 경우 이 두 가지를 통합하여 처리할 것인지 아니면 분리하여 처리할 것인지 하는 문제와 관련된다. 용언 가운데는 본용언으로만 쓰이는 것도 있고 보조 용언으로만 쓰이는 것도 있기 때문에 이러한 것들은 큰 문제가 되지 않는다. 그러나 용언 가운데는 본용언으로 쓰이는 것도 있고 보조 용언으로 쓰이는 것들도 적지 않으므로 이들을 각각 분리하여 처리할지 아니면 통합하여 처리할지 고민할 필요가 있다.

(1다)는 어떤 어휘가 다른 어휘와 단어 형성의 관계에 놓여 있다고 할 때 이들을 하나의 어휘로 처리할 것인지 아니면 단어 형성의 관계에 놓여 있는 어휘들도 함께 처리할 것인지 하는 문제와 관련된다. 이 문제는 특히 생산성이 높은 접사의 처리와 관련하여 쟁점이 된다. 가령 ‘공부’와 ‘공부하다’라는 단어가 있을 때 이 두 어휘

를 별개의 어휘로 처리할 수도 있지만 ‘공부’에 ‘공부하다’라는 단어를 포함시켜 ‘공부’ 하나로 통합할 수도 있다. ‘합리’와 ‘합리적’도 마찬가지 문제를 제기한다.

(1라)는 품사는 동일하지만 하위 품사의 쓰임이 상위 품사의 쓰임과 현저히 다를 경우 이를 별도의 어휘로 처리할 것인지 하는 문제와 관련된다. 가령 일반 명사로 쓰이는 단어가 의존 명사로 발달한 경우 그 의미에도 일정한 차이가 발생하여 사전에 따라서는 이를 별도의 단어로 동음이의어 처리하는 경우가 있으므로 이들을 하나의 어휘로 다룰 것인지 아니면 별도의 단어로 다룰 것인지 하는 것이 쟁점이 될 수 있다.

이상의 내용들은 기초 어휘의 선정과 등급화에 큰 영향을 미치는 것들이라고 할 수 있다. 하나의 어휘로 처리할 경우 상대적으로 다양한 어휘들을 기초 어휘로 선정할 수 있고 빈도, 분포, 산포도에 따른 순위가 합산되므로 등급도 상향될 가능성이 높다. 서로 다른 어휘로 처리할 경우 상대적으로 기초 어휘로 선정되는 어휘들의 다양성이 떨어질 수 있고 빈도, 분포, 산포도에 따른 순위가 개별적으로 계산되므로 등급도 하향될 가능성이 있다.

다음으로는 기초 어휘의 범위와 관련되는 쟁점을 들 수 있다. 이것은 기초 어휘를 어떤 단위로 삼을 것인가 하는 문제와 기초 어휘에 포함되는 어휘의 속성과 관련되는 두 가지 문제를 포함한다. 이와 관련된 세부 쟁점들을 정리하면 다음과 같다.

(2) 가. 조사와 어미의 포함 여부
나. 접사의 포함 여부
다. 고유 명사의 포함 여부
라. 감탄사의 포함 여부

(2가), (2나)는 기초 어휘의 단위와 관련된 쟁점이고 (2다), (2라)는 기초 어휘의 속성과 관련된 쟁점이다.

(2가)에서 제시한 조사와 어미는 통상 어휘의 범주에서는 제외되는 것이 일반적이다. 그러나 조사와 어미는 국어의 문법적 특성을 반영하는 대표적인 두 요소이므로 이들을 기초 어휘에 포함시키는 경우도 생각해 볼 수 있다.

(2나)에서 제시한 접사는 접두사와 접미사로 구체화될 수 있는데 모든 접두사와 접미사가 기초 어휘에 포함되는 것은 아니지만 생산성이 높은 접두사나 접미사의 경우 이를 기초 어휘에 포함시켜 다루는 것도 가능하다.

(2다)의 고유 명사는 기초 어휘로서의 가치가 일반 명사에 비해 현저히 떨어지는 것이다. 따라서 이를 기초 어휘에 포함시켜야 할지 만약 포함시킨다면 어느 정도까지 포함시켜야 할지 하는 문제를 검토해 보아야 할 필요가 있다.

(2라)의 감탄사는 빈도, 분포, 산포도라는 관점에서만 보면 상당히 높은 순위를

가지는 것들이 나타나는 것이 일반적이다. 그러나 고유 명사와 마찬가지로 감탄사도 기초 어휘로서의 가치가 떨어지는 것이라고 할 수 있다는 점에서 역시 이를 포함시켜야 할지 만약 포함시킨다면 어느 정도까지 포함시켜야 할지 하는 문제를 검토해 보아야 할 필요가 있다.

이상의 정성적 검토 내용에 대해 방향을 정하기 위해서는 두 가지 방향으로의 해결안을 모색할 필요가 있다. 하나는 기존의 연구들에서 이들 쟁점을 어떻게 처리하고 있는지 살펴보는 것이고 다른 하나는 구체적 어휘들을 대상으로 이들 각 쟁점에 따른 결과가 어떤 양상으로 나타나는지 살펴보는 것이다.

마지막으로 통계적 절차에 따라 추출한 기초 어휘 예비 목록을 초등학교 교과서에서 추출한 어휘 23,280개와 비교 검토한다. 본 연구에서는 전년도 연구 결과에 따라 1등급 기초 어휘를 3,000개로 선정할 예정이며, 2등급 이상 기초 어휘는 50,000개 정도로 선정할 예정이다.[19] 통계적으로 추출한 이들 어휘와 교과서 어휘와의 비교를 통해 개별 어휘에 대한 정성적 검토를 실시하여 최종 선정하는 기초 어휘 목록에 반영할 사항들을 도출하고자 한다.

19) 이삼형 외(2017b)에서는 텍스트 포괄 범위, 선행 연구 등을 참고해 1등급 기초 어휘로 3,000개, 2등급 이상 기초 어휘로 50,000개를 예상한 바 있다.

2. 기존 연구 검토

앞에서 제시한 쟁점들에 대해 선행 연구들에서는 어떤 입장을 표명하고 있는지 검토해 볼 필요가 있다. 이를 위해 본 연구에서 살펴볼 선행 연구 목록을 국어교육과 한국어교육으로 나누어 제시하면 다음과 같다.

영역	번호	선행연구 목록
국어 교육	1	임지룡(1991), 「국어의 기초 어휘에 대한 연구」, 『국어교육연구』 23-1, 국어교육학회.
	2	김광해(2003), 『등급별 국어교육용 어휘』, 박이정
	3	김한샘(2005), 『현대 국어 사용 빈도 조사』, 국립국어원.
	4	서정미(2008), 「말뭉치를 활용한 고등학교 국어사전의 편찬을 위한 기초 연구」, 경기대 박사학위논문.
	5	김한샘(2009), 『초등학교 교과서 어휘 조사 연구』, 국립국어원.
	6	장경희 외(2012), 『초·중·고등학생의 구어 어휘 조사』, 지식과교양.
한국어 교육	7	이충우(1994), 「한국어 어휘 교육을 위한 대표 어휘 선정」, 『국어교육』 85, 한국어교육학회.
	8	조현용(2000), 『한국어 어휘교육 연구』, 박이정.
	9	임칠성(2002), 「초급 한국어 교육용 어휘 선정 연구」, 『국어교육학연구』 14, 국어교육학회.
	10	조남호(2003), 『한국어 학습용 어휘 선정 결과 보고서』, 국립국어원.
	11	배주채(2010), 『한국어 기초어휘집』, 한국문화사.
	12	서상규(2013), 『한국어 기본어휘 연구』, 한국문화사.
	13	서상규(2014), 『한국어 기본어휘 빈도 사전』, 한국문화사.
	14	한송화(2015), 『한국어 교육 어휘 내용 개발(4단계)』, 국립국어원.

이들 선행 연구에서 앞서 제시한 쟁점인 동형어 구분, 어근 어휘와 파생어 어휘 구분, 품사별 어휘 목록의 등재 여부에 대하여 어떻게 처리하고 있는지 살펴보기로 한다.[20]

20) 선행 연구 중에는 어휘 처리에 대한 지침을 제시한 연구도 있고 그렇지 않은 연구도 있었다. 그리고 어휘 목록을 제시한 연구도 있고 그렇지 않은 연구도 있었다. 지침이나 어휘 목록이 제시되어 있지 않은 경우, 제시된 정보만을 바탕으로 쟁점 사항의 처리 방식을 확인하였다. 선행연구 중 지침을 제시한 연구로는 김광해(2003), 김한샘(2005), 서정미(2008), 김한샘(2009), 장경희(2012), 이충우(1994), 조현용(2000), 임칠성(2002), 조남호(2003), 서상규(2013), 한송화(2015)가 있었다. 전체 어휘 목록을 제시한 연구로는 임지룡(1991), 김한샘(2005), 서정미(2008), 김한샘(2009), 장경희(2012), 이충우(1994), 조현용(2000), 임칠성(2002), 조남호(2003), 배주해(2010), 서상규(2013), 서상규(2014), 한송화(2015)가 있었다.

2.1. 동형어 구분

동형어 구분에 있어서는 품사통용 어휘의 구분, 본용언과 보조용언의 구분 여부를 확인한다. 품사통용 어휘는 명사-부사 통용, 명사-의존명사 통용, 부사-접속부사 통용을 확인한다.

품사통용 어휘의 구분 여부를 확인하면 다음과 같다. 전체 14개 연구 중 명사-부사 통용을 구분한 연구는 8개 연구, 명사-의존명사 통용을 구분한 연구는 11개 연구, 부사-접속부사 통용을 구분한 연구는 3개 연구가 있었다. 명사-의존명사 통용의 경우, 특히 국어교육 영역의 모든 연구에서 구분을 하고 있었고, 부사-접속부사 통용의 경우, 특히 한국어교육 영역의 모든 연구에서 구분을 하지 않아 연구 목적에 따라 처리 방식이 달라짐을 알 수 있었다.

○ 명사-부사 통용
- 어휘를 구분한 연구: 김한샘(2005), 서정미(2008), 김한샘(2009), 장경희(2012), 임칠성(2002), 조남호(2003), 서상규(2014), 한송화(2015)
- 구분하지 않은 연구: 임지룡(1991), 김광해(2003), 이충우(1994), 조현용(2000), 배주채(2010), 서상규(2013)

○ 명사-의존명사 통용
- 어휘를 구분한 연구: 임지룡(1991), 김광해(2003), 김한샘(2005), 서정미(2008), 김한샘(2009), 장경희(2012), 임칠성(2002), 조남호(2003), 배주채(2010), 서상규(2013), 한송화(2015)
- 구분하지 않은 연구: 이충우(1994), 조현용(2000), 서상규(2014)

○ 부사-접속부사 통용
- 어휘를 구분한 연구: 김한샘(2005), 김한샘(2009), 장경희(2012)
- 구분하지 않은 연구: 임지룡(1991), 김광해(2003), 서정미(2008), 이충우(1994), 조현용(2000), 임칠성(2002), 조남호(2003), 배주채(2010), 서상규(2013), 서상규(2014), 한송화(2015)

본용언과 보조용언의 구분 여부를 확인하면, 전체 14개 연구 중 어휘를 구분한 연구는 8개 연구, 통합하여 처리한 연구는 5개 연구, 구분 여부를 확인하지 못한 연구는 1개 연구가 있었다. 국어교육 영역의 경우 확인하지 못한 1개 연구를 제외한 모든 연구가 구분하여 처리하고 있고, 한국어교육 영역의 경우 전체 8개 연구 중 3개 연구만이 구분하여 처리하고 5개 연구가 통합하여 처리하고 있어 연구 목

적에 따라 처리 방식이 달라짐을 확인할 수 있었다.

○ 본용언-보조용언
- 어휘를 구분한 연구: 임지룡(1991), 김광해(2003), 김한샘(2005), 김한샘(2009), 장경희(2012), 조남호(2003), 서상규(2013), 서상규(2014)
- 구분하지 않은 연구: 이충우(1994), 조현용(2000), 임칠성(2002), 배주채(2010), 한송화(2015)
- 구분 여부를 제시하지 않은 연구: 서정미(2008)

2.2. 파생어 어휘 구분

어근 어휘와 파생어 어휘의 구분 여부를 생산성이 높은 접사인 '-적'과 '-하다'를 통해 확인한다. 선행 연구에서 '명사+적' 형태의 파생어 어휘와 명사 형태의 어근 어휘, '명사+하다' 형태의 파생어 어휘와 명사 형태의 어근 어휘를 구분하고 있는지를 확인한다.

'명사+적' 어휘와 명사 어휘의 구분 여부를 확인하면, 전체 14개 연구 중 어휘를 구분한 연구는 6개 연구, 통합하여 처리한 연구는 5개 연구, 구분 여부를 확인하지 못한 연구는 1개 연구가 있었다.

○ '명사+적'-명사
- 어휘를 구분한 연구: 김광해(2003), 김한샘(2005), 김한샘(2009), 장경희(2012), 조남호(2003), 배주채(2010), 서상규(2014), 한송화(2015)
- 구분하지 않은 연구: 임지룡(1991), 이충우(1994), 조현용(2000), 임칠성(2002), 서상규(2013)
- 구분 여부를 제시하지 않은 연구: 서정미(2008)

'명사+하다' 어휘와 명사 어휘의 구분 여부를 확인하면, 전체 14개 연구 중 어휘를 구분한 연구는 10개 연구, 통합하여 처리한 연구는 4개 연구가 있었다.

○ '명사+하다'-명사
- 어휘를 구분한 연구: 김광해(2003), 김한샘(2005), 서정미(2008), 김한샘(2009), 장경희(2012), 임칠성(2002), 조남호(2003), 배주채(2010), 서상규(2013), 서상규(2014)
- 구분하지 않은 연구: 임지룡(1991), 이충우(1994), 조현용(2000), 한송화(2015)

2.3. 어휘 목록의 등재 요소

어휘 목록의 등재 요소에 있어서는 품사 중 조사, 어미, 접사, 고유명사, 감탄사를 어휘 목록에 포함하고 있는지를 확인한다.

조사와 어미를 어휘 목록에 포함한 연구는 전체 14개 연구 중 2개 연구가 있었다. 한국어교육 영역의 경우 조사를 하나의 어휘로 제시한 경우가 없었다.

○ 조사, 어미의 목록 등재 여부
- 목록에 포함한 연구: 서정미(2008), 장경희(2012)
- 포함하지 않은 연구: 임지룡(1991), 김광해(2003), 김한샘(2005), 김한샘(2009), 이충우(1994), 조현용(2000), 임칠성(2002), 조남호(2003), 배주채(2010), 서상규(2013), 서상규(2014), 한송화(2015)

접사를 어휘 목록에 포함한 연구는 전체 14개 연구 중 6개 연구가 있었다. 특이사항을 살펴보면, 서상규(2014)에서는『한국어 기본어휘 빈도 사전』을 만들 때 “실질 어휘만을 올림말로 하고(일부 접사 포함), 조사와 어미는 단략한 동형어 빈도수만을 보이기로 한다.”라고 밝히고 있다. 이때 일부 접사는 “-가, -간, -감, -같다, -경, -권, -께, -끼리, -네, -님, -답다, 대, -들, -류, -만하다, -명, -발, -별, 부-, 비-, -상, -생, -석, -선, -세, -식, 신-, -싶다, -씨, -씩, 양-, -어치, -여, -용, -이, -인, -재, -적, 제-, -제, -지, -질, -짜리, -째, -쯤, -차, -채, -측, -투성이, -하, -하다, -형, -화”임을 명시하고 있다.

○ 접사의 목록 등재 여부
- 목록에 포함한 연구: 서정미(2008), 김한샘(2009), 이충우(1994), 서상규(2013), 서상규(2014), 한송화(2015)
- 포함하지 않은 연구: 임지룡(1991), 김광해(2003), 김한샘(2005), 장경희(2012), 조현용(2000), 임칠성(2002), 조남호(2003), 배주채(2010)

고유명사를 어휘 목록에 포함한 연구는 전체 14개 연구 중 8개 연구가 있었다. 특이사항을 살펴보면, 조현용(2000)의 경우 ‘태극기’, ‘한국’, ‘러시아’, ‘중국’, ‘태권도’ 등을 어휘 목록에 포함하고 있기는 하나 고유명사라는 정보를 제시하지는 않았다. 임칠성(2003)의 경우 지침에서 고유 명사 중 인명만을 제외한다고 하였으며 실제 어휘 목록에 ‘일본’, ‘제주도’, ‘한국’, ‘영국’ 등 국가나 지명을 어휘 목록에 포함하고 있었다.

○ 고유명사의 목록 등재 여부
- 목록에 포함한 연구: 김광해(2003), 장경희(2012), 조현용(2000), 임칠성(2002), 조남호(2003), 배주채(2010), 서상규(2013), 서상규(2014)
- 포함하지 않은 연구: 임지룡(1991), 김한샘(2005), 서정미(2008), 김한샘(2009), 이충우(1994), 한송화(2015)

감탄사를 어휘 목록에 포함한 연구는 전체 14개 연구 중 11개 연구가 있었다.

○ 감탄사의 목록 등재 여부
- 목록에 포함한 연구: 김광해(2003), 김한샘(2005), 서정미(2008), 김한샘(2009), 조현용(2000), 임칠성(2002), 조남호(2003), 배주채(2010), 서상규(2013), 서상규(2014), 한송화(2015)
- 포함하지 않은 연구: 임지룡(1991), 장경희(2012), 이충우(1994)

이상의 내용을 표로 정리하여 제시하면 다음과 같다.

<표 23> 기존 연구의 어휘 처리 지침

선행연구		국어교육						한국어교육							
		임지룡 (1991)	김광해 (2003)	김한샘 (2005)	서정미 (2008)	김한샘 (2009)	장경희 (2012)	이충우 (1994)	조현용 (2000)	임칠성 (2002)	조남호 (2003)	배주채 (2010)	서상규 (2013)	서상규 (2014)	한송화 (2015)
동형어 구분	명사/부사	X	X	O	O	O	O	X	X	O	O	X	X	O	O
	명사/의존명사	O	O	O	O	O	O	X	X	O	O	O	O	X	O
	부사/접속부사	X	X	O	X	O	O	X	X	X	X	X	X	X	X
	본용언/보조용언	O	O	O	?	O	O	X	X	X	O	X	O	O	X
파생어 어휘 구분	OO/OO+적	X	O	O	?	O	O	X	X	X	O	O	X	O	O
	OO/OO하다	X	O	O	O	O	O	X	X	O	O	O	O	O	X
어휘 목록 포함	조사	X	X	X	O	X	O	X	X	X	X	X	X	X	X
	어미	X	X	X	O	X	O	X	X	X	X	X	X	X	X
	접사	X	X	X	O	O	X	O	X	X	X	X	O	O	O
	고유명사	X	O	X	X	X	O	X	△	△	O	O	O	O	X
	감탄사	X	O	O	O	O	X	X	O	O	O	O	O	O	O

3. 항목별 검토

여기에서는 실제 어휘 목록을 대상으로 앞에서 쟁점이 되었던 항목들 가운데 동형어와 접사 처리에 대해 검토해 보기로 한다. 고유명사와 품사는 이들을 처리할 구체적인 지침이 필요하며, 본 연구에서 다루는 접사 이외의 수많은 접사들은 출현 양상에 대한 실제를 검토하여 처리지침을 마련하여야 한다. 실제 어휘 목록은 이삼형 외(2017b)에서 제시되었던 방법론에 따라 샘플 언어 자료를 대상으로 형태소 분석을 거친 후 빈도, 분포, 산포도의 가중치를 반영하여 추출한 50,000위 어휘를 대상으로 한다.

3.1. 동형어 처리

먼저 50,000위 가운데 동형어의 양상을 살펴보면서 이를 처리할 수 있는 방안에 대해 모색해 보고자 한다. 50,000위 어휘 중 동형어는 총 655개가 출현하였으며,[21] 이들을 체언, 용언, 수식언 기준으로 정리해 보이기로 한다.[22]

3.1.1. 체언 기준 동형어 처리 검토

1) 명사-부사 동형어 검토

명사-부사는 50,000위까지 중 총 97개의 동형어가 있다. 이 가운데 부사가 더 높은 순위에 있는 것은 59개, 명사가 더 높은 순위에 있는 것은 38개이다.

순위	단어	품사태그	순위	단어	품사태그
4282	가로__01	NNG	11099	가로__01	MAG
2615	각자__02	NNG	3515	각자__02	MAG
10179	금새	NNG	30101	금새	MAG
1321	내일	NNG	16033	내일	MAG
9057	매사__01	NNG	37587	매사__01	MAG
7187	밤낮	NNG	19127	밤낮	MAG
12313	뱅	NNG	33316	뱅	MAG
864	보통	NNG	2843	보통	MAG
2127	비교적	NNG	4636	비교적	MAG

21) 655개의 동형어 중 고유명사와의 동형어는 311개와 접사와의 동형어에 대한 분석은 제외한다. 또한 동형어가 1개만 있거나 품사 부착 오류로 판단되는 것도 제외한다. 특히 고유명사는 3,000위 기초 어휘에 포함할 어휘 목록을 따로 선정할 필요가 있고, 말뭉치에서 추출한 고유명사 어휘 목록은 사전에 등재되지 않은 일반명사를 활용한 상호명, 프로그램명 등이 다수 포함되어 있어 별도의 선정 지침이 필요하다고 본다.

22) '체언, 용언, 수식언 기준'이라 함은 두 개의 품사 중 앞에 제시하는 품사를 기준으로 정리한 것을 말한다. '본용언-보조용언'의 동형어는 편의상 '용언 기준 동형어'에서 다루기도 한다.

32802	뻑	NNG	38201	뻑	MAG
164	사실__04	NNG	430	사실__04	MAG
2165	사실상	NNG	42210	사실상	MAG
2088	상호__04	NNG	37800	상호__04	MAG
708	실제__02	NNG	13838	실제__02	MAG
29810	쌩	NNG	30064	쌩	MAG
15772	앤	NNG	36315	앤	MAG
1588	어제__01	NNG	2275	어제__01	MAG
176	오늘	NNG	2692	오늘	MAG
14163	왈	NNG	15886	왈	MAG
16853	이만저만	NNG	25790	이만저만	MAG
38500	인자__01	NNG	39348	인자__01	MAG
5216	일시__01	NNG	7751	일시__01	MAG
6546	일체__01	NNG	9339	일체__01	MAG
447	자연__01	NNG	24414	자연__01	MAG
2181	잘못	NNG	2387	잘못	MAG
2187	전부__05	NNG	2398	전부__05	MAG
2764	절대__05	NNG	1771	절대__05	MAG
8903	제각각	NNG	12772	제각각	MAG
20860	종래	NNG	37350	종래	MAG
178	지금__03	NNG	296	지금__03	MAG
970	진짜	NNG	4454	진짜	MAG
33013	짜장	NNG	33490	짜장	MAG
12827	천만__01	NNG	29061	천만__01	MAG
6122	통상__02	NNG	20275	통상__02	MAG
9487	하루하루	NNG	10079	하루하루	MAG
978	한번	NNG	11851	한번	MAG
1118	한편	NNG	1267	한편	MAG
29882	흔들	NNG	41519	흔들	MAG
11280	가급적	NNG	5744	가급적	MAG
3914	가까이	NNG	1583	가까이	MAG
2923	각각__01	NNG	998	각각__01	MAG
33655	각기__02	NNG	2829	각기__02	MAG
24566	거의__01	NNG	307	거의__01	MAG
11389	계속__04	NNG	442	계속__04	MAG
34471	군데군데	NNG	8944	군데군데	MAG
30107	그날그날	NNG	20912	그날그날	MAG
19112	그때그때	NNG	10026	그때그때	MAG
17892	그만큼	NNG	1413	그만큼	MAG
31998	기왕	NNG	17102	기왕	MAG
43884	늘	NNG	564	늘	MAG
1689	다__03	NNG	83	다__03	MAG
11076	다소__01	NNG	1234	다소__01	MAG
37005	다소간	NNG	35865	다소간	MAG
11821	대강__02	NNG	10013	대강__02	MAG
5580	대개__03	NNG	2438	대개__03	MAG
18655	대략	NNG	2617	대략	MAG
30141	대체__02	NNG	4962	대체__02	MAG
33592	대폭__01	NNG	6984	대폭__01	MAG
22391	또	NNG	55	또	MAG
45314	매년	NNG	2186	매년	MAG

35364	매달	NNG	4674	매달	MAG
32059	매월	NNG	7818	매월	MAG
5566	매일	NNG	915	매일	MAG
33227	매주__01	NNG	4336	매주__01	MAG
6960	먼저	NNG	303	먼저	MAG
704	모두__01	NNG	134	모두__01	MAG
25441	무조건	NNG	2061	무조건	MAG
2351	물론__01	NNG	197	물론__01	MAG
19644	방금__01	NNG	3845	방금__01	MAG
1386	서로__01	NNG	468	서로__01	MAG
46409	서로서로	NNG	16618	서로서로	MAG
18621	순간순간	NNG	16470	순간순간	MAG
1632	스스로	NNG	718	스스로	MAG
49919	시종__02	NNG	16926	시종__02	MAG
8835	실상__01	NNG	8248	실상__01	MAG
6355	아까	NNG	4866	아까	MAG
1552	약간	NNG	790	약간	MAG
21567	이만큼	NNG	9766	이만큼	MAG
25637	이왕__02	NNG	6875	이왕__02	MAG
466	이제__01	NNG	241	이제__01	MAG
11081	이쯤	NNG	7585	이쯤	MAG
24560	인제__01	NNG	18058	인제__01	MAG
4016	잠깐	NNG	1939	잠깐	MAG
5655	잠시	NNG	737	잠시	MAG
11895	저마다	NNG	5273	저마다	MAG
37325	저만치	NNG	18441	저만치	MAG
5725	정말__01	NNG	160	정말__01	MAG
2440	제일__04	NNG	1284	제일__04	MAG
49373	조각조각	NNG	17423	조각조각	MAG
767	조금__01	NNG	308	조금__01	MAG
4943	직접	NNG	320	직접	MAG
7611	참__01	NNG	319	참__01	MAG
36949	천방지축	NNG	30811	천방지축	MAG
3420	하나하나	NNG	3161	하나하나	MAG
37853	한바탕	NNG	6765	한바탕	MAG
4009	한창__01	NNG	3821	한창__01	MAG
638	현재__02	NNG	348	현재__02	MAG

이들을 다시 3,000위 기준으로 살펴보면 다음의 세 가지 유형으로 구분할 수 있다. 3,000위 안에 드는 어휘들은 기초 어휘 가운데서도 1등급에 포함될 가능성이 그만큼 높은 것들에 해당한다.

구분	어휘 수
부사와 명사 모두 3,000위 안에 드는 경우	20개
부사만 3,000위 안에 드는 경우	19개
명사만 3,000위 안에 드는 경우	9개
부사와 명사 모두 3,000위 밖에 드는 경우	49개
계	97개

가. 부사와 명사 모두 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
2923	각각__01	NNG	998	각각__01	MAG
1689	다__03	NNG	83	다__03	MAG
704	모두__01	NNG	134	모두__01	MAG
2351	물론__01	NNG	197	물론__01	MAG
864	보통	NNG	2843	보통	MAG
164	사실__04	NNG	430	사실__04	MAG
1386	서로__01	NNG	468	서로__01	MAG
1632	스스로	NNG	718	스스로	MAG
1552	약간	NNG	790	약간	MAG
1588	어제__01	NNG	2275	어제__01	MAG
176	오늘	NNG	2692	오늘	MAG
466	이제__01	NNG	241	이제__01	MAG
2181	잘못	NNG	2387	잘못	MAG
2187	전부__05	NNG	2398	전부__05	MAG
2764	절대__05	NNG	1771	절대__05	MAG
2440	제일__04	NNG	1284	제일__04	MAG
767	조금__01	NNG	308	조금__01	MAG
178	지금__03	NNG	296	지금__03	MAG
1118	한편	NNG	1267	한편	MAG
638	현재__02	NNG	348	현재__02	MAG

명사와 부사 모두 3,000위 안에 드는 어휘를 보면, 3,000위 안에서도 비교적 높은 순위를 보이는 어휘가 많고(모두01, 사실04, 오늘, 이제01, 조금01, 지금03, 현재02), 명사와 부사의 순위가 높지 않은 것들도 눈에 띈다(사실04, 어제01, 잘못, 전부05, 지금03, 한편, 현재02).

나. 부사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
3914	가까이	NNG	1583	가까이	MAG
33655	각기__02	NNG	2829	각기__02	MAG
24566	거의__01	NNG	307	거의__01	MAG
11389	계속__04	NNG	442	계속__04	MAG
17892	그만큼	NNG	1413	그만큼	MAG
43884	늘	NNG	564	늘	MAG
11076	다소__01	NNG	1234	다소__01	MAG
5580	대개__03	NNG	2438	대개__03	MAG
18655	대략	NNG	2617	대략	MAG
22391	또	NNG	55	또	MAG
45314	매년	NNG	2186	매년	MAG
5566	매일	NNG	915	매일	MAG
6960	먼저	NNG	303	먼저	MAG
25441	무조건	NNG	2061	무조건	MAG
4016	잠깐	NNG	1939	잠깐	MAG
5655	잠시	NNG	737	잠시	MAG
5725	정말__01	NNG	160	정말__01	MAG
4943	직접	NNG	320	직접	MAG
7611	참__01	NNG	319	참__01	MAG

부사만 3,000위 안에 드는 경우는 부사의 순위와 명사의 순위가 큰 차이가 있었다.

다. 명사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
2615	각자__02	NNG	3515	각자__02	MAG
1321	내일	NNG	16033	내일	MAG
2127	비교적	NNG	4636	비교적	MAG
2165	사실상	NNG	42210	사실상	MAG
2088	상호__04	NNG	37800	상호__04	MAG
708	실제__02	NNG	13838	실제__02	MAG
447	자연__01	NNG	24414	자연__01	MAG
970	진짜	NNG	4454	진짜	MAG
978	한번	NNG	11851	한번	MAG

명사만 3,000위 안에 드는 경우도 명사로만 처리할 수 있겠으나, 의미 계열 관계를 고려해 보아야 할 어휘들이 있다. '어제, 오늘'은 모두 3,000위 안에 있어 명사와 부사 각각으로 처리한다고 할 때, '내일'의 처리도 이들과 동일선상에서 이루어지는 것이 바람직하다고 할 수 있다.

2) 명사-의존명사 동형어 검토

50,000위 안에 상위 품사와 하위 품사의 형태가 동일한 명사-의존명사 동형어는 총 85개이고, 이 가운데 명사가 더 높은 순위에 있는 어휘가 29개, 의존명사가 더 높은 순위에 있는 어휘가 56개이다.

순위	단어	품사태그	순위	단어	품사태그
33511	거리__02	NNG	3873	거리__02	NNB
13577	건__04	NNG	1809	건__04	NNB
5907	격__01	NNG	3267	격__01	NNB
30495	그램	NNG	9440	그램	NNB
35988	그루__01	NNG	5214	그루__01	NNB
44660	남짓	NNG	8448	남짓	NNB
38571	내__09	NNG	445	내__09	NNB
35501	년__02	NNG	16	년__02	NNB
35138	년도	NNG	2725	년도	NNB
19912	놈__01	NNG	1170	놈__01	NNB
831	달__05	NNG	701	달__05	NNB
10431	달러	NNG	879	달러	NNB
5657	대__06	NNG	690	대__06	NNB
32840	대__15	NNG	1477	대__15	NNB
9336	도__05	NNG	1270	도__05	NNB
17957	되__01	NNG	8128	되__01	NNB
22915	등__04	NNG	4927	등__04	NNB
4513	마련	NNG	1987	마련	NNB

40366	말__03	NNG	16375	말__03	NNB
44026	메가	NNG	14251	메가	NNB
26931	배__05	NNG	4365	배__05	NNB
40582	배럴	NNG	39897	배럴	NNB
29996	번__04	NNG	77	번__04	NNB
38335	벌__02	NNG	4062	벌__02	NNB
1681	부__15	NNG	1353	부__15	NNB
22009	분__08	NNG	204	분__08	NNB
34048	불__08	NNG	12621	불__08	NNB
32420	뼘__02	NNG	12788	뼘__02	NNB
35316	선__15	NNG	33663	선__15	NNB
25979	세__13	NNG	349	세__13	NNB
47462	센트	NNG	9603	센트	NNB
16554	셈__01	NNG	1027	셈__01	NNB
6400	수__02	NNG	7	수__02	NNB
12388	승__12	NNG	5145	승__12	NNB
4980	시__10	NNG	186	시__10	NNB
3371	식__04	NNG	904	식__04	NNB
27063	아름__01	NNG	11472	아름__01	NNB
3775	월__02	NNG	39	월__02	NNB
29850	장__22	NNG	858	장__22	NNB
30508	조__20	NNG	17856	조__20	NNB
1657	주__26	NNG	1482	주__26	NNB
25625	주일__03	NNG	3066	주일__03	NNB
1370	중__04	NNG	51	중__04	NNB
43113	지경__02	NNG	3357	지경__02	NNB
23789	쪽__03	NNG	14028	쪽__03	NNB
31872	쪽__05	NNG	350	쪽__05	NNB
39887	차__03	NNG	429	차__03	NNB
34736	참__03	NNG	6702	참__03	NNB
5897	초__03	NNG	1058	초__03	NNB
48279	측	NNG	1008	측	NNB
45342	톤__01	NNG	3934	톤__01	NNB
3484	판__01	NNG	3407	판__01	NNB
1950	편__04	NNG	551	편__04	NNB
27223	편__09	NNG	654	편__09	NNB
47450	평__02	NNG	1873	평__02	NNB
44760	프로__01	NNG	28011	프로__01	NNB
2646	단__06	NNG	24582	단__06	NNB
8105	대__01	NNG	4680	대__01	NNB
7317	돌__01	NNG	30447	돌__01	NNB
11267	동__15	NNG	12165	동__15	NNB
2510	등급	NNG	32373	등급	NNB
8563	랜드	NNG	38253	랜드	NNB
33173	루피	NNG	35310	루피	NNB
2143	마당	NNG	11366	마당	NNB
2979	막__05	NNG	8447	막__05	NNB
370	모양__02	NNG	19849	모양__02	NNB
506	바람__01	NNG	2231	바람__01	NNB
4478	바퀴__01	NNG	5968	바퀴__01	NNB
25368	바트__01	NNG	44754	바트__01	NNB

427	법__01	NNG	1035	법__01	NNB
14364	빨	NNG	40892	빨	NNB
9093	수__04	NNG	15042	수__04	NNB
64	시간__04	NNG	545	시간__04	NNB
17477	실__05	NNG	27494	실__05	NNB
1705	알__01	NNG	43999	알__01	NNB
3638	엔__01	NNG	4452	엔__01	NNB
6387	유로__06	NNG	25971	유로__06	NNB
146	점__10	NNG	679	점__10	NNB
1882	조__15	NNG	18089	조__15	NNB
4312	주간__05	NNG	12527	주간__05	NNB
1026	줄__01	NNG	25972	줄__01	NNB
13130	칼로리	NNG	46080	칼로리	NNB
13932	탕__07	NNG	35415	탕__07	NNB
8368	폰	NNG	33007	폰	NNB
628	해__01	NNG	1756	해__01	NNB

명사-의존명사 동형어 85개 중 3,000위 기준으로 살펴보면 다음의 유형으로 구분할 수 있다.

구분	어휘 수
명사와 의존명사 모두 3,000위 안에 드는 경우	10개
명사만 3,000위 안에 드는 경우	7개
의존명사만 3,000위 안에 드는 경우	25개
명사와 의존명사 모두 3,000위 밖에 드는 경우	43개
계	85개

가. 명사와 의존명사 모두 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
506	바람__01	NNG	2231	바람__01	NNB
427	법__01	NNG	1035	법__01	NNB
64	시간__04	NNG	545	시간__04	NNB
146	점__10	NNG	679	점__10	NNB
628	해__01	NNG	1756	해__01	NNB
831	달__05	NNG	701	달__05	NNB
1681	부__15	NNG	1353	부__15	NNB
1657	주__26	NNG	1482	주__26	NNB
1370	중__04	NNG	51	중__04	NNB
1950	편__04	NNG	551	편__04	NNB

이들을 보면 명사와 의존명사 모두 3000위 기본 어휘에 포함하는 데 큰 문제가 없다고 할 수 있겠다.

나. 명사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
2646	단__06	NNG	24582	단__06	NNB
2510	등급	NNG	32373	등급	NNB
2143	마당	NNG	11366	마당	NNB
2979	막__05	NNG	8447	막__05	NNB
370	모양__02	NNG	19849	모양__02	NNB
1705	알__01	NNG	43999	알__01	NNB
1026	줄__01	NNG	25972	줄__01	NNB

명사와 의존명사의 순위 차이가 크게 나므로 명사만 기본 어휘에 포함할지에 대해서는 고려해 보아야 할 것이다.

다. 의존명사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
13577	건__04	NNG	1809	건__04	NNB
38571	내__09	NNG	445	내__09	NNB
35501	년__02	NNG	16	년__02	NNB
35138	년도	NNG	2725	년도	NNB
19912	놈__01	NNG	1170	놈__01	NNB
10431	달러	NNG	879	달러	NNB
5657	대__06	NNG	690	대__06	NNB
32840	대__15	NNG	1477	대__15	NNB
9336	도__05	NNG	1270	도__05	NNB
4513	마련	NNG	1987	마련	NNB
29996	번__04	NNG	77	번__04	NNB
22009	분__08	NNG	204	분__08	NNB
25979	세__13	NNG	349	세__13	NNB
16554	셈__01	NNG	1027	셈__01	NNB
6400	수__02	NNG	7	수__02	NNB
4980	시__10	NNG	186	시__10	NNB
3371	식__04	NNG	904	식__04	NNB
3775	월__02	NNG	39	월__02	NNB
29850	장__22	NNG	858	장__22	NNB
31872	쪽__05	NNG	350	쪽__05	NNB
39887	차__03	NNG	429	차__03	NNB
5897	초__03	NNG	1058	초__03	NNB
48279	측	NNG	1008	측	NNB
27223	편__09	NNG	654	편__09	NNB
47450	평__02	NNG	1873	평__02	NNB

의존명사만 3,000위 안에 드는 경우는 대부분 단위성 의존명사이므로 의존명사로 처리해도 될 듯하다.

3) 명사-관형사 동형어 검토

명사-관형사 동형어는 50,000위 안에 총 7개의 어휘가 있다. 명사가 순위가 높은 경우는 3개, 관형사가 순위가 높은 경우는 4개이다.

순위	단어	품사태그	순위	단어	품사태그
587	수__26	NNG	3438	수__26	MM
95	전__08	NNG	376	전__08	MM
5614	주__03	NNG	6746	주__03	MM
4266	만__07	NNG	2555	만__07	MM
33576	어떤	NNG	114	어떤	MM
9849	타__01	NNG	4358	타__01	MM
25597	헌	NNG	5966	헌	MM

명사-관형사 동형어 7개 중 3000위 안에 둘 다 있는 경우는 '전_08' 하나뿐이었으며, 관형사만 있는 경우는 '만_07', '어떤'이 있고, 명사만 있는 경우는 '수_26' 하나이다.

4) 명사-감탄사 동형어 검토

50,000위 안에 있는 명사-감탄사 동형어는 총 14개로, 명사가 순위가 더 높은 경우가 6개, 감탄사가 순위가 높은 경우가 8개이다.

순위	단어	품사태그	순위	단어	품사태그
6532	만세__04	NNG	35090	만세__04	IC
3762	머	NNG	18800	머	IC
11777	쉬__02	NNG	47327	쉬__02	IC
2828	으	NNG	44504	으	IC
18265	이크	NNG	39731	이크	IC
9363	흐	NNG	33978	흐	IC
7905	안녕	NNG	7358	안녕	IC
21622	에고	NNG	18299	에고	IC
25537	에그	NNG	10331	에그	IC
30957	우와	NNG	15267	우와	IC
30854	파이팅	NNG	22105	파이팅	IC
47482	호오	NNG	35153	호오	IC
36472	화이팅	NNG	26993	화이팅	IC
9276	휴	NNG	6983	휴	IC

이들 중 3,000위 안에 드는 형태는 명사 '으' 이외에는 없었다. 명사-감탄사 동형어로 처리된 것들은 여러 변이형이 가능하고, 대개 감정을 표현하는 것들이어서 3,000위 기본 어휘에 포함해야 할지는 좀 더 고려해 보아야 할 것이다.

5) 대명사-감탄사 동형어 검토

50,000위 안의 대명사-감탄사 동형어는 4개의 어휘가 있으며, 모두 대명사의 순위가 높았다.

순위	단어	품사태그	순위	단어	품사태그
35447	거시기	NP	36681	거시기	IC
5067	무어__01	NP	19692	무어__01	IC
214	뭐	NP	993	뭐	IC
253	어디__01	NP	7336	어디__01	IC

3,000위 기준으로 보면, '뭐'가 감탄사와 대명사 모두 3,000위 안에 들고, '어디_01'이 대명사일 때 3,000위 안에 들었다. '뭐'가 감탄사와 대명사 모두 3,000위 안에 들고 있지만 감탄사를 기본 어휘에서 어떻게 처리할지에 따라 3,000위 안에 포함 여부가 달라질 것이다.

6) 수사-관형사 동형어 검토

수사-관형사 동형어는 50,000위 가운데 총 11개이고, 수사가 순위가 높은 경우는 9개, 관형사가 높은 경우는 2개이다.

순위	단어	품사태그	순위	단어	품사태그
10214	구__01	NR	36977	구__01	MM
2222	삼__06	NR	9533	삼__06	MM
14486	삼십	NR	43239	삼십	MM
2164	십	NR	17219	십	MM
2798	오__04	NR	41520	오__04	MM
12263	육__02	NR	39782	육__02	MM
2421	이__09	NR	18485	이__09	MM
1996	일__05	NR	6222	일__05	MM
22008	칠십	NR	44204	칠십	MM
4190	몇	NR	189	몇	MM
6094	몇몇	NR	2111	몇몇	MM

이들 가운데 3,000위 안에 드는 어휘 11개는 다음과 같이 유형을 구분해 볼 수 있다.

구분	어휘 수
수사와 관형사 모두 3,000위 안에 드는 경우	0개
수사만 3,000위 안에 드는 경우	5개
관형사만 3,000위 안에 드는 경우	2개
수사와 관형사 모두 3,000위 밖에 드는 경우	4개
계	11개

가. 수사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
2222	삼__06	NR	9533	삼__06	MM
2164	십	NR	17219	십	MM
2798	오__04	NR	41520	오__04	MM
2421	이__09	NR	18485	이__09	MM
1996	일__05	NR	6222	일__05	MM

수사와 관형사의 순위 차이가 크기 때문에 수사만 기본 어휘에 포함해야 할지에 대해서는 고려해 보아야 할 것이다.

나. 관형사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
4190	몇	NR	189	몇	MM
6094	몇몇	NR	2111	몇몇	MM

‘몇’, ‘몇몇’은 3,000위 기초 어휘에 포함할 수 있겠다.

3.1.2. 용언 기준 동형어 처리 검토

1) 동사-형용사 동형어 검토

50,000위 안에 드는 어휘 중 동사-형용사의 동형어는 32개이다. 이 가운데 동사가 순위가 더 높은 경우가 12개, 형용사가 순위가 더 높은 경우가 20개이다.

순위	단어	품사태그	순위	단어	품사태그
1147	감사하__05	VV	44932	감사하__05	VA
16225	거세지	VV	40376	거세지	VA
17072	너무하	VV	17328	너무하	VA
801	당하__01	VV	33010	당하__01	VA
814	더하	VV	10258	더하	VA
18026	도드라지	VV	22170	도드라지	VA
5110	만하	VV	30890	만하	VA
577	못하	VV	3743	못하	VA
3693	무리하	VV	9162	무리하	VA
9722	밝	VV	1185	밝	VA
36079	설__01	VV	37597	설__01	VA
10028	충족하	VV	34259	충족하	VA
7888	건조하__02	VV	6499	건조하__02	VA
9279	격하__02	VV	9010	격하__02	VA
7871	굳	VV	3599	굳	VA
11042	굽__02	VV	9925	굽__02	VA
41206	근질근질하	VV	34232	근질근질하	VA

42311	끈적끈적하	VV	15334	끈적끈적하	VA
45326	너덜너덜하	VV	20182	너덜너덜하	VA
11633	늦	VV	1030	늦	VA
7525	두드러지	VV	4862	두드러지	VA
7862	만족하	VV	2167	만족하	VA
33121	바삭하	VV	44813	바삭하	VA
47440	반질반질하	VV	42485	반질반질하	VA
17955	분주하__05	VV	4133	분주하__05	VA
38059	애통하__01	VV	31254	애통하__01	VA
44368	엉거주춤하	VV	42594	엉거주춤하	VA
19373	있__01	VV	5	있__01	VA
3505	지나치	VV	1472	지나치	VA
21840	찡하__01	VV	19581	찡하__01	VA
6565	크__01	VV	42	크__01	VA
42566	화끈하	VV	10944	화끈하	VA

3,000위 기준으로 32개의 동형어를 구체적으로 살펴보면 다음과 같이 유형을 구분하여 정리해 볼 수 있다.

구분	어휘 수
동사와 형용사 모두 3,000위 안에 드는 경우	0개
동사만 3,000위 안에 드는 경우	4개
형용사만 3,000위 안에 드는 경우	6개
동사와 형용사 모두 3,000위 밖에 드는 경우	22개
계	32개

가. 동사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
1147	감사하__05	VV	44932	감사하__05	VA
801	당하__01	VV	33010	당하__01	VA
814	더하	VV	10258	더하	VA
577	못하	VV	3743	못하	VA

‘못하다’를 제외하고는 모두 동사와 형용사의 순위 차이가 크다. ‘못하다’는 선행용언에 따라 달라지는 것이므로 동사와 형용사 모두로 처리하는 것이 바람직할 듯하다.

나. 형용사만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
9722	밝	VV	1185	밝	VA
11633	늦	VV	1030	늦	VA
7862	만족하	VV	2167	만족하	VA
19373	있__01	VV	5	있__01	VA
3505	지나치	VV	1472	지나치	VA
6565	크__01	VV	42	크__01	VA

3,000위 기본 어휘에는 형용사만 드는 경우 동사의 순위도 모두 20,000위 안에 들고 있다.

2) 본용언-보조용언 동형어 검토

본용언-보조용언은 50,000위까지 중 총 38개의 동형어가 있다.[23] 이 중 본용언이 더 높은 순위에 있는 것이 19개, 보조용언이 더 높은 순위에 있는 것이 19개이다.

순위	단어	품사태그	순위	단어	품사태그
29	가__01	VV	161	가__01	VX
86	가지	VV	4890	가지	VX
249	갖__01	VV	37571	갖__01	VX
1489	계시	VV	2135	계시	VX
115	나__01	VV	444	나__01	VX
335	나가	VV	503	나가	VX
1232	대__01	VV	1702	대__01	VX
238	두__01	VV	384	두__01	VX
90	들__01	VV	2377	들__01	VX
23686	뜨리	VV	38628	뜨리	VX
46	먹__02	VV	2673	먹__02	VX
13	보__01	VV	23	보__01	VX
630	빠지__02	VV	9413	빠지__02	VX
45	오__01	VV	133	오__01	VX
26950	잇__04	VV	47409	잇__04	VX
260	죽__01	VV	17384	죽__01	VX
4455	치우__01	VV	6766	치우__01	VX
1422	터지	VV	24408	터지	VX
30639	프	VV	43088	프	VX
4	하__01	VV	9	하__01	VX
171	내__02	VV	159	내__02	VX
518	놓__01	VV	167	놓__01	VX
3478	달__05	VV	1016	달__05	VX
709	드리__01	VV	694	드리__01	VX
46596	듯싶	VV	8902	듯싶	VX
8742	듯하	VV	398	듯하	VX
5110	만하	VV	464	만하	VX
660	말__03	VV	174	말__03	VX
577	못하	VV	60	못하	VX
843	버리__01	VV	187	버리__01	VX
21052	법하	VV	6061	법하	VX
21139	뻔하__01	VV	5443	뻔하__01	VX
2221	싶	VA	57	싶	VX
20873	아니하	VV	2620	아니하	VX
916	않	VV	10	않	VX

23) 이는 태그 오류 2개를 제외한 수치이다.

19373	있__01	VV	6	있__01	VX
98	주__01	VV	15	주__01	VX
1550	지__04	VV	26	지__04	VX
32520	척하__01	VV	5106	척하__01	VX
21384	체하__01	VV	11936	체하__01	VX

이들을 다시 3,000 순위 기준으로 살펴보면 다음과 같이 구분할 수 있다.

구분	어휘 수
본용언과 보조용언 모두 3,000위 안에 드는 경우	20개
본용언만 3,000위 안에 드는 경우	5개
보조용언만 3,000위 안에 드는 경우	5개
본용언과 보조용언 모두 3,000위 밖에 드는 경우	8개
계	38개

가. 본용언과 보조용언 모두 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
29	가__01	VV	161	가__01	VX
1489	계시	VV	2135	계시	VX
115	나__01	VV	444	나__01	VX
335	나가	VV	503	나가	VX
171	내__02	VV	159	내__02	VX
518	놓__01	VV	167	놓__01	VX
1232	대__01	VV	1702	대__01	VX
238	두__01	VV	384	두__01	VX
709	드리__01	VV	694	드리__01	VX
90	들__01	VV	2377	들__01	VX
660	말__03	VV	174	말__03	VX
46	먹__02	VV	2673	먹__02	VX
577	못하	VV	60	못하	VX
843	버리__01	VV	187	버리__01	VX
13	보__01	VV	23	보__01	VX
916	않	VV	10	않	VX
45	오__01	VV	133	오__01	VX
98	주__01	VV	15	주__01	VX
1550	지__04	VV	26	지__04	VX
4	하__01	VV	9	하__01	VX

위의 목록들은 본용언의 순위와 보조용언의 순위가 모두 높아 각각 매우 활발한 빈도로 사용되는 것으로 추정된다. 이들 각각을 개별 어휘로 처리하는 것이 좋을 듯하다.

나. 본용언만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
86	가지	VV	4890	가지	VX
249	갖__01	VV	37571	갖__01	VX

630	빠지__02	VV	9413	빠지__02	VX
260	죽__01	VV	17384	죽__01	VX
1422	터지	VV	24408	터지	VX

본용언의 순위와 보조용언의 순위가 크게 차이가 있어 3,000위 기본 어휘에는 본용언으로만 처리해도 무방할 것으로 보인다.

다. 보조용언만 3,000위 안에 드는 경우

순위	단어	품사태그	순위	단어	품사태그
3478	달__05	VV	1016	달__05	VX
8742	듯하	VV	398	듯하	VX
5110	만하	VV	464	만하	VX
20873	아니하	VV	2620	아니하	VX
19373	있__01	VV	6	있__01	VX

보조용언만 3,000위 안에 드는 경우 중 '듯하다, 만하다, 아니하다'는 보조용언으로만 쓰이는 것이므로 보조용언으로 처리하고, '달다_05, 있다_01'은 본용언, 보조용언으로 두루 사용되는 것이므로 본용언, 보조용언으로 처리하여도 큰 문제가 없을 듯하다.

3.1.3. 수식언 기준 동형어 처리 검토

1) 부사-접속부사 동형어 검토

부사-접속부사는 50,000위 가운데 총 8개의 동형어가 있다. 이 가운데 부사가 순위가 더 높은 경우는 3개, 접속부사가 순위가 높은 경우는 5개이다.

순위	단어	품사태그	순위	단어	품사태그
1203	단__05	MAG	7624	단__05	MAJ
7714	하긴	MAG	16606	하긴	MAJ
11843	허나	MAG	29263	허나	MAJ
18912	그럼__01	MAG	722	그럼__01	MAJ
1755	그리고	MAG	50	그리고	MAJ
19327	하물며	MAG	4775	하물며	MAJ
9628	하지만	MAG	158	하지만	MAJ
21187	한데__03	MAG	15406	한데__03	MAJ

8개의 동형어 중 3,000위 안에 부사와 접속부사가 모두 드는 경우는 '그리고' 하나였으며, '단_05'는 부사만 3,000위 안에 들고, '그럼_01, 하지만'은 접속 부사만이 3,000위 안에 든다. 이들은 모두 3,000위 기본 어휘에서는 '부사'로 처리하여도 무방할 듯하다.

2) 부사-감탄사 동형어 검토

50,000위 안에 드는 부사-감탄사 동형어는 총 3개였으며, 모두 부사의 순위가 높았다. 이 가운데 3,000위 안에 드는 어휘는 2개뿐이고 이들은 부사로 3,000위 기본 어휘에 포함할 수 있을 듯하다.

순위	단어	품사태그	순위	단어	품사태그
244	왜__02	MAG	34073	왜__02	IC
319	참__01	MAG	6223	참__01	IC
16844	헉	MAG	30155	헉	IC

3.1.4 동형어 처리에 대한 방향 모색

이상에서 살펴본 바와 같이 동형어의 처리는 전체를 일률적으로 처리하기 어려운 부분이 적지 않다. 따라서 동형어 처리에 대한 전체적인 원칙을 설정하고 이와는 별도로 구체적 사안에서는 각기 다른 처리 방식을 취하는 방안을 고려할 필요가 있다.

앞선 선행 연구를 검토해 보면 대체적으로 국어교육에서는 동형어를 구분하는 것이 우세하였고, 한국어교육에서는 명사-의존명사를 제외하고는 통합하여 하나로 처리하는 것이 우세하였다. 그러나 한국어교육에서도 최근 연구인 2015년 '한국어 어휘 교육 내용 개발 1단계' 연구와 2017년 '국제 통용 한국어 표준 교육과정 적용 연구 4단계'에서 다음에서 보는 바와 같이 서로 다른 원칙을 제시하고 있어 개별 유형에 대한 고찰이 필요한 것으로 보인다.

2015년	(1) 품사 통용어 ○ 하나의 어휘가 두 개 이상의 품사를 가지는 경우 **품사별로 개별 처리하는 것을 원칙**으로 하되, **어휘 간의 의미 차이가 적어 구분이 어려운 경우에는 해당 품사를 모두 제시하고 하나의 목록으로 처리**하였다. 예를 들어, 지시 대명사 '어디01'와 '벼르거나 다짐할 때, 되물어 강조할 때, 남의 주의를 끌 때 등'에 사용하는 감탄사 '어디01'는 각각의 목록으로 처리하였다. 이때 각 어휘별 등급을 고려하여 대명사의 '어디01'는 초급 단계에 포함되나 감탄사의 '어디01'는 초급 단계에서 포함되지 않았다. 한편, 수사와 관형사로 사용되는 '다섯'은 실제 사용에서 문법적인 지위는 달라지지만 의미에 차이가 없으므로 하나의 목록으로 처리하였다.
2017년	1. 품사별 동형어 처리 방법 * 분류 원칙: 동형어 중 상위 빈도(혹은 이른 등급)에 있는 어휘를 중심으로 결합, 등급화하여 제시

본 연구에서는 위의 견해들을 참고하여 품사 통용되는 어휘 중 3000위 기본 어휘에 드는 경우에는 해당 품사를 모두 제시하고 하나의 목록으로 제시하는 방법을 택하고자 한다. 품사 통용 어휘의 품사에 대한 정보는 모국어 화자이기 때문에 모두 제시하여도 크게 무리가 가지 않을 것으로 판단된다.

더욱이 모국어 화자를 위한 기초 어휘이므로 동형어가 포함된 3,000위 목록보다 동형어를 하나로 묶은 목록이 서로 다른 어휘를 더 많이 포함할 수 있으므로 효율성 면에서 낫다고 볼 수 있다.

이와 함께, 언어 자료에서 추출한 목록에서는 동형어로 처리하였지만 의존명사, 보조용언과 같이 하나의 품사로만 사용되는 경우는 하나의 품사로 제시하는 것이 바람직할 것이다. 부사-접속부사와 같이 서로 층위가 다른 품사 정보로 인한 동형어도 마찬가지이다.

3.2. 접사 처리

선행 연구에서는 접사를 기초 어휘로 선정한 경우도 있고 그렇지 않은 경우도 있었다. 이때 기초 어휘로 선정된 접사는 접두사이든 접미사이든 생산성이 높은 경우에 해당하므로 여기에서도 생산성이 높은 접두사와 접미사를 대상으로 이들을 별도의 기초 어휘로 선정할 때 발생할 수 있는 문제에 대해 살펴보기로 한다.

3.2.1. 접두사 ‘미-, 불-, 비-’ 처리 검토

먼저 접두사는, ‘부정(否定)’이라는 동일한 의미 관계를 나타내면서 상대적으로 순위가 높은 ‘미-, 불-, 비-’를 대상으로 이들이 50,000위 어휘 목록에서 단어 형성에 참여하는 양상을 비교해 접사 처리에 대한 방향을 모색해 보기로 한다. 50,000위 어휘 목록에 ‘미-, 불-, 비-’가 붙어 이루어진 단어와 어근 동형어가 모두 포함된 어휘 쌍의 비율은 다음과 같다.

‘미(未)-’ 파생어	‘불(不)-’ 파생어	‘비(非)-’ 파생어	계
12개(12.3%)	60개(61.2%)	26개(26.5%)	98개

1) ‘미-’ 접두 파생어 검토

접두사 ‘미-’가 붙어 이루어진 파생어가 어근 동형어를 가지는 경우는 50,000위 안에 다음과 같이 12개가 발견된다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
8433	미완성	NNG	1818	완성__01	NNG
25431	미확인	NNG	1587	확인__02	NNG
26160	미성숙	NNG	7790	성숙	NNG
28744	미성년	NNG	14886	성년__01	NNG
29102	미성숙하다	VA	4928	성숙하다	VV
33539	미달하다	VV	1643	달하다__01	VV
34013	미취학	NNG	29338	취학	NNG
37496	미해결	NNG	1820	해결__02	NNG
43701	미발표	NNG	1774	발표__01	NNG
44286	미개척	NNG	7840	개척	NNG
45206	미완성되	VV	1826	완성되	VV
46278	미개발	NNG	583	개발	NNG

접두사 '미-'의 경우, 접두사가 붙지 않은 형태를 접두사가 붙은 말과 비교할 때, 접두사가 붙지 않은 어근의 순위가 매우 높은 편이라고 할 수 있으며 파생어와의 순위 차이도 매우 큰 편이다. 이는 가령 '미개발'이 46,278위인데 반해 '개발'은 583위에 해당한다는 사실을 통해 단적으로 알 수 있다. 또한 '거짓말'과 '거짓'의 경우와 같이 더 복잡한 어휘의 순위가 덜 복잡한 어휘보다 순위가 높은 경우가 발견되지 않는다는 점도 '미-' 파생어의 특징이라고 할 수 있다. 따라서 이 경우에는 접두사를 별도의 어휘로 처리해도 큰 문제가 생기지 않는다고 할 수 있다. '미-' 파생어의 경우에는 3,000위 안에 드는 것은 발견되지 않는다.

2) '불-' 접두 파생어 검토

접두사 '불-'이 붙어 파생어가 어근 동형어를 가지는 경우는 50,000위 안에 다음과 같이 60개가 있다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
1460	불가능하	VA	212	가능하	VA
1475	불편하__01	VA	924	편하	VA
2931	불법__01	NNG	1035	법__01	NNG
3683	불필요하	VA	227	필요하	VA
4978	불안정하	VA	30004	안정하__01	VV
5251	불가능	NNG	1781	가능	NNG
5400	불분명하	VA	1162	분명하__01	VA
6243	불완전하	VA	2569	완전하__01	VA
6956	불투명하	VA	3579	투명하__02	VA
6998	불확실하	VA	1398	확실하	VA
7067	불균형	NNG	2107	균형	NNG
8383	불쾌감	NNG	7232	쾌감	NNG

12613	불이익	NNG	1389	이익__02	NNG
12930	불안정	NNG	2625	안정__01	NNG
13358	불공평하	VA	6979	공평하__01	VA
15158	불확실성	NNG	30789	확실성	NNG
15299	불명예	NNG	2536	명예__01	NNG
15597	불규칙적	NNG	6271	규칙적	NNG
16022	불특정	NNG	1717	특정	NNG
17260	불일치	NNG	4750	일치__01	NNG
17897	불만족	NNG	2311	만족__01	NNG
18042	불만족스럽	VA	3405	만족스럽	VA
18175	불평등	NNG	9633	평등	NNG
18546	불친절하	VA	2659	친절하	VA
18561	불성실하	VA	4381	성실하__02	VA
18708	불완전	NNG	1367	완전__01	NNG
19385	불법적	NNG	3436	법적__01	NNG
19435	불충분하	VA	1239	충분하__01	VA
20564	불명확하	VA	3072	명확하	VA
22006	불투명	NNG	4229	투명__02	NNG
23009	불합리	NNG	34318	합리__01	NNG
23451	불합격	NNG	6619	합격	NNG
24532	불효__01	NNG	5627	효__02	NNG
24539	불공정	NNG	8575	공정__02	NNG
24683	불평등하	VA	6228	평등하	VA
24906	불괘	NNG	27496	괘__04	NNG
25352	불필요	NNG	365	필요	NNG
25438	불규칙	NNG	2590	규칙__02	NNG
26035	불공정하	VA	5350	공정하__01	VA
29101	불공평	NNG	22940	공평__01	NNG
30853	불친절	NNG	6975	친절	NNG
31521	불충분	NNG	9101	충분__01	NNG
32684	불안정성	NNG	6248	안정성	NNG
34405	불효자	NNG	7884	효자__01	NNG
34912	불확실	NNG	17789	확실	NNG
35690	불포화	NNG	10182	포화__07	NNG
36232	불건전하	VA	4861	건전하	VA
37975	불복종	NNG	12876	복종__01	NNG
37983	불가능성	NNG	514	가능성	NNG
38155	불분명	NNG	1580	분명__01	MAG
39285	불이행	NNG	10870	이행__08	NNG
40025	불가결	NNG	39008	가결__01	NNG
40470	불안전하	VA	1989	안전하	VA
40624	불만족하	VA	2167	만족하	VA
40624	불만족하	VA	7862	만족하	VV
43355	불합리성	NNG	17011	합리성	NNG
44455	불구속	NNG	6423	구속__02	NNG
46447	불균형적	NNG	27641	균형적	NNG
46899	불연속적	NNG	8230	연속적	NNG
47470	불성실	NNG	13026	성실__02	NNG

접두사 ‘불-’의 경우에도 접두사 ‘미-’의 경우와 유사한 양상을 보였으나, 이른바 ‘거짓말 유형’에 해당하는 것들이 일부 나타났다는 특징을 발견할 수 있다. ‘불안정하-, 불확실성, 불합리, 불쾌’와 같은 어휘는 접두사가 붙지 않은 어근이 접두사가 붙은 파생어에 비해 순위가 더 낮은 것이다. 이러한 사실을 중시한다면, 접두사를 따로 떼어 독립 항목으로 처리하기보다는 각 어휘가 지닌 사용 양상을 반영하여 개별 어휘 단위로 처리하는 것이 타당할 것으로 여겨진다. 다만 3,000위 안에 드는 ‘불-’ 파생어는 ‘불가능하-, 불편하-, 불법’의 세 개인데 이들은 구조의 복잡성에 따라 순위가 역전되지 않는다는 것을 알 수 있다.

3) ‘비-’ 접두 파생어 검토

접두사 ‘비-’이 붙어 파생어가 어근 동형어를 가지는 경우는 50,000위 안에 다음과 같이 26개가 있다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
6918	비정상적	NNG	2993	정상적	NNG
10719	비현실적	NNG	2169	현실적	NNG
10809	비정상	NNG	4605	정상__02	NNG
13067	비공식	NNG	2141	공식__01	NNG
14176	비공개	NNG	1949	공개__02	NNG
14462	비효율	NNG	4377	효율	NNG
19387	비인간적	NNG	3463	인간적	NNG
19471	비합리적	NNG	2359	합리적	NNG
20091	비공식적	NNG	4585	공식적	NNG
20280	비전문가	NNG	1119	전문가	NNG
20827	비대칭	NNG	7834	대칭__02	NNG
24181	비무장	NNG	6589	무장__06	NNG
30785	비위생적	NNG	13297	위생적	NNG
31519	비협조적	NNG	24423	협조적	NNG
32444	비이성적	NNG	7107	이성적	NNG
34674	비포장도로	NNG	37547	포장도로	NNG
36888	비인간	NNG	172	인간__01	NNG
37285	비정규직	NNG	36393	정규직	NNG
38992	비정규	NNG	7277	정규__01	NNG
40019	비양심적	NNG	11849	양심적	NNG
40718	비정형	NNG	8545	정형__05	NNG
42305	비회원	NNG	1706	회원	NNG
46489	비협조	NNG	20570	협조__02	NNG
48772	비민주적	NNG	18341	민주적	NNG
49068	비과세	NNG	25741	과세__04	NNG
49348	비합법적	NNG	9629	합법적	NNG

접두사 '비-'가 붙어 이루어진 말도 단어 구조의 복잡성에 따라 순위가 역전되는 현상이 나타나지 않아 그 양상은 '미-'의 경우와 흡사하다는 것을 알 수 있다. '비-' 파생어 가운데도 3,000위 안에 드는 것은 발견되지 않는다.

3.2.2. 접미사 처리 검토

접미사는 접두사에 비해 파생어의 수가 매우 많다. 따라서 이를 다시 한자 접미사와 고유어 접미사로 나누고 한자 접미사는 '-적, -성, -화'를 대상으로, 고유어 접미사는 '-스럽-, -롭-, -답-'을 대상으로 살펴보기로 한다. 이들은 모두 생산성이 높은 접미사의 예들이라는 점에서 공통성을 지닌다.

1) 접미사 '-적, -성, -화' 검토

접미사의 경우에는 접두사의 경우와는 달리 파생어의 수가 상당히 많이 존재하므로 모든 예들을 검토 대상으로 삼는 것은 적절하지 않다. 따라서 여기에서는 50,000위 어휘 내 '-적(的)', '-성(性)', '-화(化)' 파생어 중 50,000위 어휘 내에 어근 동형어가 있는 경우를 1,000개를 추출하고, 다시 이들 가운데 3,000위 안에 드는 어휘를 대상으로 그 양상을 살펴보기로 한다.

먼저 1,000개 어휘 가운데 '-적(的)', '-성(性)', '-화(化)' 파생어의 비중은 각각 다음과 같다.

'-적(的)' 파생어	'-성(性)' 파생어	'-화(化)' 파생어	계
630개(63%)	255개(25.5%)	115개(11.5%)	1,000개

이들 가운데 3,000위 안에 드는 어휘는 일차적으로는 '-적(的)', '-성(性)', '-화(化)' 파생어만 대상으로 삼으면 되지만 기초 어휘 선정을 위해서는 이들 접미사를 제외한 어근도 살펴볼 필요가 있다. 따라서 그 경우의 수를 '접사 포함 단어만 3,000위 안에 드는 경우', '어근 동형어만 3,000위 안에 드는 경우', '접사 포함 단어와 어근 동형어 모두 3,000위 안에 드는 경우'의 세 가지로 나누기로 한다. 이들 각각의 비중은 다음과 같다.

	'-적(的)'	'-성(性)'	'-화(化)'
가. 접사 포함 단어만 3,000위 안에 드는 경우	12개	0개	0개
나. 어근 동형어만 3,000위 안에 드는 경우	168개	67개	37개
다. 접사 포함 단어와 어근 동형어 모두 3,000위 안에 드는 경우	20개	2개	0개
파생어의 수	32개	2개	0개

‘파생어의 수’는 3,000위 안에 드는 ‘-적(的)’, ‘-성(性)’, ‘-화(化)’ 파생어를 의미하는데 50,000위 안에서 모두 1,000개가 들어 있다는 점과 비교하면 3,000위 안에는 모두 34개가 들어 있어 그 비중이 각각 2%와 1.1%로 차이가 있다. 더욱이 3,000위 안에 존재하는 ‘-적(的)’, ‘-성(性)’, ‘-화(化)’ 파생어 가운데 대부분이 ‘-적(的)’ 파생어라는 점도 주목할 필요가 있다.

(1) ‘-적’ 접미 파생어 검토

가. 접사 포함 단어만 3,000위 안에 드는 경우

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
1289	구체적	NNG	23095	구체__02	NNG
2353	근본적	NNG	4187	근본	NNG
2158	긍정적	NNG	5835	긍정	NNG
1646	본격적	NNG	10338	본격	NNG
2547	부정적	NNG	5284	부정__02	NNG
2701	전반적	NNG	5190	전반__03	NNG
2799	전형적	NNG	7654	전형__04	NNG
2799	전형적	NNG	23873	전형__07	NNG
2993	정상적	NNG	4605	정상__02	NNG
2286	지속적	NNG	4385	지속__01	NNG
2538	직접적	NNG	4943	직접	NNG
2359	합리적	NNG	34318	합리__01	NNG

‘-적(的)’ 파생어만 3,000위 안에 드는 경우는 ‘거짓말’과 ‘거짓’의 관계와 흡사하다. 즉 단어 구조에서는 더 복잡하지만 어근이 되는 단어보다 파생어의 순위가 더 높은 것이다. 이러한 어휘들의 존재는 어근과 파생어를 하나의 어휘로 처리하기는 어렵고 서로 개별적으로 처리해야 한다는 사실을 말해 준다.

한편 위의 ‘-적(的)’ 파생어들은 3,000위 안에 들기는 하지만 그 순위가 대부분 2,000위를 넘고 있고 1,000위 안에 드는 것은 없다는 점도 특징이다.

나. 어근 동형어만 3,000위 안에 드는 경우

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
17883	가정적__01	NNG	1070	가정__06	NNG
19545	가족적	NNG	318	가족__01	NNG
7139	감각적	NNG	1590	감각__02	NNG
5535	감동적	NNG	1721	감동__02	NNG
8648	감성적	NNG	2526	감성__02	NNG
6214	감정적__01	NNG	895	감정__06	NNG
22797	개념적	NNG	848	개념	NNG
15354	개성적	NNG	2131	개성__03	NNG

38191	개혁적	NNG	2712	개혁	NNG
21321	건설적	NNG	1855	건설	NNG
3291	결과적	NNG	265	결과__02	NNG
11561	경쟁적	NNG	1142	경쟁	NNG
18086	경험적	NNG	615	경험	NNG
23406	계산적	NNG	2782	계산__01	NNG
9185	계속적	NNG	442	계속__04	MAG
24394	계절적	NNG	2086	계절__01	NNG
9038	계획적	NNG	422	계획__01	NNG
5880	고전적	NNG	2559	고전__02	NNG
12711	고정적	NNG	2861	고정__06	NNG
11732	공간적	NNG	272	공간__05	NNG
7837	공개적	NNG	1949	공개__02	NNG
7519	공격적	NNG	1290	공격__02	NNG
4585	공식적	NNG	2141	공식__01	NNG
9993	교육적	NNG	328	교육	NNG
46861	구성적	NNG	824	구성__07	NNG
6331	구조적	NNG	463	구조__08	NNG
6607	국가적	NNG	281	국가__01	NNG
42536	국내적	NNG	382	국내__02	NNG
11526	국민적	NNG	408	국민	NNG
5641	국제적	NNG	700	국제__02	NNG
13663	군사적	NNG	2869	군사__04	NNG
6271	규칙적	NNG	2590	규칙__02	NNG
27641	균형적	NNG	2107	균형	NNG
19146	근대적	NNG	2765	근대__03	NNG
6784	기계적	NNG	1562	기계__07	NNG
7910	기능적	NNG	339	기능__03	NNG
17971	기록적	NNG	759	기록__02	NNG
5788	기술적__01	NNG	270	기술__01	NNG
40563	기업적	NNG	304	기업__01	NNG
10888	남성적	NNG	1060	남성__01	NNG
11484	내부적	NNG	784	내부__04	NNG
24070	내용적__01	NNG	251	내용__02	NNG
4905	논리적__01	NNG	1946	논리	NNG
10737	단계적	NNG	652	단계__03	NNG
11995	대략적	NNG	2617	대략	MAG
4965	대중적	NNG	1253	대중__02	NNG
22274	도시적	NNG	393	도시__03	NNG
12661	도전적	NNG	2399	도전__04	NNG
5464	독립적	NNG	1714	독립	NNG
17429	동물적	NNG	791	동물	NNG
36495	동시적	NNG	557	동시__02	NNG
8336	동양적	NNG	2591	동양__03	NNG
9267	무조건적	NNG	2061	무조건	MAG
44668	문명적	NNG	1868	문명__03	NNG
12175	문학적	NNG	818	문학__01	NNG
3080	문화적	NNG	233	문화__01	NNG
6679	물질적	NNG	1250	물질__02	NNG
32817	미술적	NNG	1957	미술	NNG

9422	미적__01	NNG	2733	미__14	NNG
19676	민족적	NNG	1211	민족	NNG
24754	발생적	NNG	2116	발생	NNG
19380	발전적	NNG	698	발전__01	NNG
48406	범죄적	NNG	1881	범죄	NNG
3436	법적__01	NNG	427	법__01	NNG
16814	병적__01	NNG	905	병__04	NNG
4299	본질적	NNG	2809	본질__02	NNG
3840	부분적	NNG	157	부분__01	NNG
17490	분석적	NNG	933	분석__02	NNG
19385	불법적	NNG	2931	불법__01	NNG
6991	비판적	NNG	1857	비판__01	NNG
17731	사상적	NNG	1458	사상__15	NNG
8611	사실적	NNG	164	사실__04	NNG
29215	산업적	NNG	598	산업	NNG
34041	상상적	NNG	1688	상상__07	NNG
4476	상식적	NNG	2977	상식__06	NNG
5137	상징적	NNG	2289	상징	NNG
9851	생산적	NNG	1095	생산	NNG
34205	서양적	NNG	1902	서양	NNG
10293	선택적	NNG	800	선택	NNG
46792	설명적	NNG	736	설명	NNG
29559	성격적	NNG	853	성격__02	NNG
5206	성적__01	NNG	1143	성__07	NNG
26898	세기적	NNG	617	세기__03	NNG
17769	수적__01	NNG	587	수__26	NNG
4864	순간적	NNG	419	순간__03	NNG
7323	습관적	NNG	2160	습관	NNG
4931	시간적	NNG	64	시간__04	NNG
11987	시기적	NNG	665	시기__04	NNG
5766	시대적	NNG	250	시대__02	NNG
11205	시적__01	NNG	473	시__13	NNG
46398	신분적	NNG	2132	신분__02	NNG
22092	실천적	NNG	2792	실천__01	NNG
9720	실험적	NNG	1348	실험	NNG
4168	심리적	NNG	2349	심리__01	NNG
5466	안정적	NNG	2625	안정__01	NNG
38503	암적__01	NNG	1793	암__08	NNG
11374	양적__01	NNG	854	양__20	NNG
18466	언어적	NNG	827	언어__01	NNG
11890	여성적	NNG	280	여성__01	NNG
32184	연극적	NNG	1449	연극	NNG
8230	연속적	NNG	2027	연속__02	NNG
6466	열정적	NNG	2035	열정__02	NNG
18966	영웅적	NNG	2728	영웅	NNG
33999	영화적	NNG	85	영화__01	NNG
5649	예술적	NNG	606	예술	NNG
10941	예외적	NNG	2896	예외	NNG
17679	옛적	NNG	1507	옛__01	MM
14473	외부적	NNG	1329	외부__02	NNG

11133	운명적	NNG	1727	운명__01	NNG
45786	원리적	NNG	1746	원리__02	NNG
7454	원칙적	NNG	1555	원칙	NNG
47169	유형적__02	NNG	2386	유형__07	NNG
13760	음악적	NNG	215	음악__01	NNG
9292	의식적	NNG	932	의식__03	NNG
13999	의학적	NNG	2734	의학__02	NNG
5653	이론적	NNG	952	이론__01	NNG
3463	인간적	NNG	172	인간__01	NNG
10126	인공적	NNG	2864	인공__01	NNG
3053	인상적	NNG	2152	인상__06	NNG
3406	일상적	NNG	1169	일상__04	NNG
7255	자동적	NNG	2248	자동__01	NNG
5492	자연적	NNG	447	자연__01	NNG
38629	작가적	NNG	362	작가__01	NNG
5465	전국적	NNG	1042	전국__03	NNG
10772	전략적	NNG	1112	전략__03	NNG
3443	전문적	NNG	909	전문__08	NNG
16392	전투적	NNG	2361	전투	NNG
45628	점차적	NNG	1920	점차__02	MAG
8369	정서적	NNG	2897	정서__06	NNG
19372	정책적	NNG	528	정책__02	NNG
14822	제도적	NNG	838	제도__01	NNG
8577	제한적	NNG	2277	제한__01	NNG
9694	조직적	NNG	842	조직	NNG
6461	종교적	NNG	1269	종교	NNG
5828	종합적	NNG	2000	종합	NNG
17981	주목적	NNG	2239	주목__03	NNG
21076	주의적	NNG	2263	주의__02	NNG
14571	주체적	NNG	2758	주체__02	NNG
28511	중간적	NNG	797	중간__01	NNG
7147	중심적	NNG	435	중심__01	NNG
25576	지구적__01	NNG	812	지구__04	NNG
45446	지도적	NNG	1715	지도__09	NNG
6673	지배적	NNG	2578	지배__01	NNG
10627	지역적	NNG	194	지역__03	NNG
11900	직업적	NNG	1541	직업	NNG
4919	질적	NNG	2076	질__08	NNG
15741	집단적	NNG	1340	집단	NNG
3712	집중적	NNG	2341	집중__02	NNG
16000	차별적	NNG	2476	차별	NNG
6270	창조적	NNG	2759	창조__03	NNG
7606	철학적	NNG	1298	철학	NNG
4103	체계적	NNG	1377	체계__03	NNG
16664	추가적	NNG	1335	추가__02	NNG
6201	충격적	NNG	1828	충격__02	NNG
37369	통일적	NNG	2091	통일__02	NNG
6064	특징적	NNG	773	특징	NNG
13330	파괴적	NNG	2938	파괴	NNG
8100	평균적	NNG	1354	평균	NNG

22470	평화적	NNG	1599	평화__02	NNG
9662	폭력적	NNG	2191	폭력	NNG
7891	표면적__01	NNG	2357	표면	NNG
4957	필수적	NNG	2661	필수__02	NNG
8573	한국적	NNG	101	한국__05	NNP
7573	핵심적	NNG	1327	핵심	NNG
40674	행동적	NNG	626	행동	NNG
20470	행정적	NNG	2205	행정__01	NNG
9697	혁명적	NNG	1671	혁명	NNG
4889	현대적	NNG	1233	현대__01	NNG

이들은, 어근이 되는 단어는 3,000위 안에 들지만 '-적(的)' 파생어는 그렇지 못한 경우이다. 이는 단어 구조가 더 복잡한 것이 더 순위가 낮은 경우로서 세 가지 가운데 그 비중이 가장 높은 것에서도 알 수 있듯이 가장 일반적인 경우라고 할 수 있다. 따라서 하나의 어휘로 묶어 처리하든 어휘 개별적으로 처리하든 큰 문제가 없다고 할 수 있다.

다. 접사 포함 단어와 어근 동형어 모두 3,000위 안에 드는 경우

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
787	개인적	NNG	523	개인__02	NNG
2674	결정적	NNG	1201	결정__01	NNG
1600	경제적	NNG	333	경제__04	NNG
2663	과학적	NNG	641	과학	NNG
1303	기본적	NNG	678	기본	NNG
1148	대표적	NNG	343	대표	NNG
1675	매력적	NNG	857	매력	NNG
2127	비교적	NNG	1800	비교__01	NNG
985	사회적	NNG	152	사회__07	NNG
2242	상대적	NNG	965	상대__04	NNG
2718	성공적	NNG	1182	성공__01	NNG
1635	세계적	NNG	140	세계__02	NNG
2253	역사적	NNG	352	역사__04	NNG
885	일반적	NNG	675	일반__02	NNG
1494	전체적	NNG	421	전체__01	NNG
2097	전통적	NNG	884	전통__06	NNG
2737	절대적	NNG	2764	절대__05	NNG
2290	정신적	NNG	456	정신__12	NNG
1618	정치적	NNG	478	정치__03	NNG
2538	직접적	NNG	320	직접	MAG

이들은 '-적(的)' 파생어는 물론 어근이 되는 단어도 3,000위 안에 드는 경우이다. 어휘들을 자세히 비교해 보면 '절대/절대적'의 경우만 그 순위가 역전되어 있고 나머지는 모두 어근이 되는 단어의 순위가 더 높다는 점에 주목할 필요가 있다. 따라서 넓게는 바로 앞의 경우와 일맥상통한다는 점을 알 수 있다.

또한 양쪽 모두 1등급 어휘로 선정될 가능성이 높기 때문에 어떤 경우보다도 하나의 어휘로의 처리를 지지하는 것으로 해석할 수 있게 한다. 다만 그러한 경우 어느 한 쪽이 3,000위 안에 들어 있는 경우보다 합산 순위가 높아질 가능성이 있기 때문에 세부 등급을 나눌 경우 순위가 상승할 가능성이 그만큼 높아진다. 또한 하나의 어휘로의 처리 때문에 개별적인 어휘로 처리할 때와는 달리 다른 후순위의 어휘가 1등급 어휘 안으로 포함될 수 있다는 점도 염두에 둘 필요가 있다.

(2) '-성' 접미 파생어 검토

가. 접사 포함 단어만 3,000위 안에 드는 경우

'-성(性)' 파생어만 3,000위 안에 드는 경우는 존재하지 않았다.

나. 어근 동형어만 3,000위 안에 드는 경우

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
42506	건강성	NNG	949	건강__03	NNG
18015	경제성	NNG	333	경제__04	NNG
38280	경향성	NNG	1862	경향__02	NNG
32669	계획성	NNG	422	계획__01	NNG
29695	고유성	NNG	2942	고유__03	NNG
15523	공격성	NNG	1290	공격__02	NNG
34524	공공성	NNG	1895	공공__02	NNG
37910	과학성	NNG	641	과학	NNG
16861	관련성	NNG	454	관련	NNG
34988	구조화	NNG	463	구조__08	NNG
23598	국민성	NNG	408	국민	NNG
26373	규칙성	NNG	2590	규칙__02	NNG
49559	근대성	NNG	2765	근대__03	NNG
42213	노인성__01	NNG	1825	노인__01	NNG
32312	논리성	NNG	1946	논리	NNG
17252	대중성	NNG	1253	대중__02	NNG
27833	대표성	NNG	343	대표	NNG
25949	덕성__01	NNG	1779	덕__05	NNG
18576	도시화	NNG	393	도시__03	NNG
16939	독립성	NNG	1714	독립	NNG
24331	동물성	NNG	791	동물	NNG
46155	동시성	NNG	557	동시__02	NNG
23303	마성	NNG	2174	마__10	NNG
31460	목적성	NNG	734	목적__03	NNG
42731	문제성	NNG	82	문제__06	NNG
37561	문학성	NNG	818	문학__01	NNG
29726	민족성	NNG	1211	민족	NNG
12406	방향성__01	NNG	597	방향__01	NNG

22533	사실성	NNG	164	사실__04	NNG
39874	사업성	NNG	369	사업__04	NNG
10187	사회성	NNG	152	사회__07	NNG
20972	상대성	NNG	965	상대__04	NNG
12505	상징성	NNG	2289	상징	NNG
25500	상품성	NNG	891	상품__03	NNG
12043	생산성	NNG	1095	생산	NNG
27799	시장성	NNG	235	시장__04	NNG
12108	신뢰성	NNG	2558	신뢰__02	NNG
8530	악성__01	NNG	2217	악__04	NNG
14100	안전성	NNG	1523	안전__03	NNG
6248	안정성	NNG	2625	안정__01	NNG
20705	역사성	NNG	352	역사__04	NNG
14750	연속성	NNG	2027	연속__02	NNG
30048	운동성	NNG	336	운동__02	NNG
6563	위험성	NNG	1379	위험	NNG
26336	음악성	NNG	215	음악__01	NNG
24022	이동성	NNG	1243	이동__03	NNG
8992	인간성	NNG	172	인간__01	NNG
41307	인사성	NNG	1817	인사__02	NNG
38193	일상성	NNG	1169	일상__04	NNG
12195	전문성	NNG	909	전문__08	NNG
42637	절대성	NNG	2764	절대__05	NNG
23418	접근성	NNG	2407	접근	NNG
37128	정치성	NNG	478	정치__03	NNG
49231	종교성	NNG	1269	종교	NNG
17371	주체성	NNG	2758	주체__02	NNG
33324	준비성	NNG	685	준비	NNG
36853	지역성	NNG	194	지역__03	NNG
13013	진실성	NNG	1699	진실__02	NNG
16208	창조성	NNG	2759	창조__03	NNG
12683	특수성	NNG	2234	특수__02	NNG
20420	폭력성	NNG	2191	폭력	NNG
27272	한계성	NNG	1420	한계	NNG
42886	항상성	NNG	639	항상	MAG
49233	현대성	NNG	1233	현대__01	NNG
10708	현실성	NNG	458	현실__02	NNG
20836	활동성	NNG	401	활동__02	NNG

‘-적(的)’ 파생어와 마찬가지로 ‘-성(性)’ 파생어도, 파생어는 3,000위 안에 들지 않지만 어근이 되는 단어만 3,000위 안에 드는 경우가 가장 많다. 역시 이 경우는 하나의 어휘로의 처리나 어휘 개별 처리 어느 쪽으로도 큰 문제가 생기지 않는다고 할 수 있다.

다. 접사 포함 단어와 어근 동형어 모두 3,000위 안에 드는 경우

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
514	가능성	NNG	1781	가능	NNG
2882	필요성	NNG	365	필요	NNG

‘-성(性)’ 파생어의 경우는 위의 두 예만 접사 포함 단어와 어근 단어가 모두 3,000위 안에 들었다. ‘가능성-가능’은 파생어의 순위가 더 높은 경우이고 ‘필요성-필요’는 어근 단어의 순위가 더 높은 경우이다. 역시 하나의 어휘로 처리할 경우 이들의 빈자리를 다른 어휘가 차지하게 된다는 점에 주의할 필요가 있다.

(3) ‘-화’ 접미 파생어 검토

가. 접사 포함 단어만 3,000위 안에 드는 경우

‘-성(性)’ 파생어처럼 ‘-화(化)’ 파생어만 3,000위 안에 드는 경우는 존재하지 않았다.

나. 어근 동형어만 3,000위 안에 드는 경우

접사 포함 단어			어근동형어 (없는 경우 공란)		
순위	단어	품사	순위	단어	품사
19769	고급화	NNG	1911	고급__02	NNG
38953	고도화	NNG	2886	고도__08	NNG
44344	공동화	NNG	1064	공동__02	NNG
28399	공식화	NNG	2141	공식__01	NNG
38768	과학화	NNG	641	과학	NNG
20671	국제화	NNG	700	국제__02	NNG
13389	근대화	NNG	2765	근대__03	NNG
23478	기계화	NNG	1562	기계__07	NNG
9702	대중화	NNG	1253	대중__02	NNG
29975	대형화	NNG	1349	대형__04	NNG
46293	문제화	NNG	82	문제__06	NNG
14699	미화__01	NNG	2733	미__14	NNG
35523	사회화	NNG	152	사회__07	NNG
12326	산업화	NNG	598	산업	NNG
46292	상징화	NNG	2289	상징	NNG
15318	상품화	NNG	891	상품__03	NNG
25356	생활화	NNG	315	생활	NNG
12837	세계화	NNG	140	세계__02	NNG
37122	습관화	NNG	2160	습관	NNG
7460	악화__01	NNG	2217	악__04	NNG
19647	안정화	NNG	2625	안정__01	NNG
37746	영화화	NNG	85	영화__01	NNG

48377	의식화	NNG	932	의식__03	NNG
48464	인간화	NNG	172	인간__01	NNG
13083	일반화	NNG	675	일반__02	NNG
28465	일상화	NNG	1169	일상__04	NNG
16808	자동화	NNG	2248	자동__01	NNG
25321	자유화__01	NNG	653	자유__03	NNG
22795	전문화	NNG	909	전문__08	NNG
39129	제도화	NNG	838	제도__01	NNG
37556	조직화	NNG	842	조직	NNG
36979	집중화	NNG	2341	집중__02	NNG
10529	차별화	NNG	2476	차별	NNG
20759	체계화	NNG	1377	체계__03	NNG
32825	최대화	NNG	880	최대	NNG
11921	현대화	NNG	1233	현대__01	NNG
15314	현실화	NNG	458	현실__02	NNG

‘-적(的)’, ‘-성(性)’ 파생어와 마찬가지로 가장 많은 경우가 이에 해당한다. 역시 하나의 어휘로 처리하든 어휘 개별적으로 처리하든 큰 문제가 없다고 할 수 있다. 3,000위 안에 드는 파생어는 보이지 않지만, 어근은 3,000위 안에 드는 것이 다수라는 사실도 알 수 있다.

다. 접사 포함 단어와 어근 동형어 모두 3,000위 안에 드는 경우

‘-화(化)’ 파생어와 어근 단어가 모두 3,000위 안에 드는 경우는 존재하지 않았다.

2) 접미사 ‘-스럽-, -롭-, -답-’ 검토

여기에서는 50,000위 어휘 가운데 고유어 접미사 ‘-스럽-, -롭-, -답-’이 결합된 파생어를 대상으로 살펴보도록 한다. 대상 어휘의 분포는 다음과 같다.

‘-스럽-’ 파생어	‘-롭-’ 파생어	‘-답-’ 파생어	계
108개(71%)	41개(27%)	3개(2%)	152개

(1) ‘-스럽-’ 접미 파생어 검토

50,000위 안에 ‘-스럽-’이 결합한 파생어와 그 어근이 함께 나타난 것은 모두 108개이다. 이 가운데 ‘-스럽-’이 결합된 파생어의 순위가 높은 경우는 15건, 어근의 순위가 더 높은 경우는 91건으로, 어근의 순위가 파생어보다 압도적으로 높게 나타난다.

파생어의 순위가 높은 단어 목록을 우선 살펴보도록 한다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
3095	조심스럽	VA	7010	조심__02	NNG
8135	당황스럽	VA	10651	당황	NNG
8676	촌스럽	VA	11647	촌__03	NNG
9201	변덕스럽	VA	10016	변덕	NNG
9370	성스럽	VA	29395	성__14	NNG
10921	익살스럽	VA	18940	익살	NNG
12718	곤혹스럽	VA	30695	곤혹	NNG
13686	고풍스럽	VA	33367	고풍__01	NNG
16076	죄송스럽	VA	28330	죄송	NNG
20447	맛깔스럽	VA	25761	맛깔	NNG
22075	능청스럽	VA	43982	능청__01	NNG
25176	억척스럽	VA	37868	억척	NNG
27434	유난스럽	VA	30018	유난	NNG
27692	의아스럽	VA	36470	의아	NNG
29248	까탈스럽	VA	42841	까탈	NNG

이 부류의 어휘들은 파생어와 어근의 순위가 비교적 큰 차이를 보인다. 특히 어근 '고풍, 곤혹, 까탈, 억척' 등은 발화 내에서 독립적으로 사용되는 경우가 드물다. 이 목록에 포함된 어휘들은 순위가 전반적으로 매우 낮고, 3,000위 안에 드는 어휘가 없으므로 어떤 방식으로 처리해도 영향을 받지 않을 것으로 보인다.

다음은 동형 어근의 순위가 파생어의 순위보다 높은 단어들의 목록을 살펴보도록 한다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
1033	자연스럽	VA	447	자연__01	NNG
3405	만족스럽	VA	2311	만족__01	NNG
3464	사랑스럽	VA	222	사랑__01	NNG
4058	혼란스럽__01	VA	2662	혼란__02	NNG
4075	자랑스럽	VA	3290	자랑__01	NNG
4264	의심스럽	VA	2753	의심__03	NNG
5298	부담스럽	VA	1160	부담__01	NNG
5850	고통스럽	VA	1187	고통	NNG
5967	다행스럽	VA	2465	다행	NNG
6263	새삼스럽	VA	4165	새삼__01	MAG
7267	비밀스럽	VA	1366	비밀	NNG
7576	고급스럽	VA	1911	고급__02	NNG
8134	정성스럽	VA	2665	정성__11	NNG
8275	걱정스럽	VA	1054	걱정	NNG
8769	실망스럽	VA	3486	실망__02	NNG
9193	소란스럽	VA	8178	소란__04	NNG
9955	신비스럽	VA	7047	신비__02	NNG
10029	사치스럽	VA	5869	사치__03	NNG
10780	고집스럽	VA	3832	고집__02	NNG
11215	영광스럽	VA	3192	영광__01	NNG

11866	탐스럽	VA	8868	탐__02	NNG
12099	탐욕스럽	VA	7050	탐욕	NNG
12778	볼만스럽	VA	2146	볼만	NNG
13220	유감스럽	VA	6139	유감__04	NNG
13661	의문스럽	VA	1913	의문__02	NNG
13808	장난스럽	VA	2875	장난	NNG
14629	자유스럽	VA	653	자유__03	NNG
14647	혐오스럽	VA	8391	혐오__02	NNG
15322	당혹스럽	VA	14639	당혹	NNG
15661	짜증스럽	VA	3807	짜증	NNG
16230	원망스럽	VA	9690	원망__01	NNG
16293	후회스럽	VA	3542	후회__01	NNG
16338	시원스럽	VA	13465	시원__02	NNG
16639	억지스럽	VA	8636	억지__01	NNG
16927	수치스럽	VA	10521	수치__03	NNG
16932	감격스럽	VA	9709	감격	NNG
17786	수다스럽	VA	5211	수다__01	NNG
17836	멋스럽	VA	3101	멋__01	NNG
17963	호사스럽	VA	11795	호사__06	NNG
18042	볼만족스럽	VA	17897	볼만족	NNG
18394	공포스럽	VA	1918	공포__08	NNG
19413	고생스럽	VA	2790	고생	NNG
19461	요란스럽	VA	17782	요란__01	NNG
19492	호화스럽	VA	9964	호화__02	NNG
19571	어른스럽	VA	1292	어른__01	NNG
20396	고민스럽	VA	1133	고민	NNG
20455	바보스럽	VA	2832	바보	NNG
22036	충성스럽	VA	6694	충성__01	NNG
23031	염려스럽	VA	3788	염려__01	NNG
23615	위험스럽	VA	1379	위험	NNG
24318	수고스럽	VA	3968	수고__01	NNG
24362	죄스럽	VA	1518	죄__03	NNG
24668	잡스럽	VA	14444	잡__03	XPN
24802	용맹스럽	VA	23225	용맹	NNG
25633	한스럽	VA	3345	한__05	NNG
26554	흉물스럽	VA	32547	흉물	NNG
26674	치욕스럽	VA	10188	치욕__01	NNG
26709	별스럽	VA	1434	별__02	MM
26738	불명예스럽	VA	15299	불명예	NNG
28331	평화스럽	VA	1599	평화__02	NNG
28366	극성스럽	VA	10053	극성__05	NNG
30074	호들갑스럽	VA	14256	호들갑	NNG
30202	애교스럽	VA	10104	애교__02	NNG
30334	불경스럽	VA	20519	불경__04	NNG
30731	거북스럽	VA	11563	거북__02	NNG
30826	괴기스럽	VA	23250	괴기__02	NNG
30998	부산스럽	VA	27622	부산__01	NNG
31270	예스럽	VA	4401	예__01	NNG
32309	복스럽	VA	2708	복__13	NNG
33570	이상스럽	VA	3002	이상__12	NNG

33710	경사스럽	VA	14141	경사__12	NNG
34496	감탄스럽	VA	4747	감탄	NNG
35902	신령스럽	VA	29713	신령__02	NNG
36190	우려스럽	VA	2096	우려__01	NNG
37692	존경스럽	VA	3601	존경	NNG
38234	경악스럽	VA	10123	경악__02	NNG
39643	태평스럽	VA	20766	태평	NNG
40514	절망스럽	VA	5922	절망__02	NNG
41154	여성스럽	VA	280	여성__01	NNG
42469	다정스럽	VA	30551	다정__01	NNG
43009	근심스럽	VA	5989	근심__01	NNG
45349	개탄스럽	VA	43168	개탄	NNG
45410	불량스럽	VA	8911	불량__01	NNG
45520	야만스럽	VA	14684	야만__02	NNG
46472	유머스럽	VA	5354	유머	NNG
46935	끔찍스럽	VA	42738	끔찍	NNG
48368	짐스럽	VA	2079	짐__01	NNG
49309	옛스럽	VA	1507	옛__01	MM
49594	한탄스럽	VA	9367	한탄	NNG
49684	저주스럽	VA	6025	저주__03	NNG
49777	부자유스럽	VA	38630	부자유	NNG

이 목록의 어근들은 사용 빈도가 높은 어휘들을 다수 포함하고 있는데, 어근 중 30개가 3,000위 내에 들어 기초 어휘에 들 것으로 보인다. 그러나 파생어 중에서는 '자연스럽-' 1개만 3,000위 내에 든다. 이상의 내용으로 볼 때 '-스럽-'이 결합된 단어와 그 어근은 어휘군으로 처리하든, 개별 어휘로 처리하든 결과가 크게 달라질 것으로 보이지 않는다.

(2) '-롭-' 접미 파생어 검토

50,000위 안에 '-롭-'이 결합한 파생어와 그 어근이 함께 나타난 것은 모두 41개이다. 이 가운데 '-롭-' 파생어의 순위가 높은 경우는 18건, 동형 어근의 순위가 높은 경우는 19건으로 두 유형의 단어들 사이에 우열이 드러나지 않는다.

다음은 파생어의 순위가 동형 어근보다 높은 단어 목록이다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
153	새롭	VA	712	새__06	MM
2404	흥미롭	VA	2660	흥미	NNG
2786	외롭	VA	40112	외__06	XPN
4579	해롭	VA	5022	해__11	NNG
4601	풍요롭	VA	10202	풍요__02	NNG
5488	신비롭	VA	7047	신비__02	NNG
6737	감미롭	VA	46527	감미__01	NNG

6741	순조롭	VA	32209	순조__01	NNP
7090	경이롭	VA	14431	경이__04	NNG
7129	대수롭	VA	17132	대수__01	NNG
7662	위태롭	VA	46955	위태__01	NNG
7703	단조롭	VA	25719	단조__04	NNG
10227	슬기롭	VA	17949	슬기__01	NNG
10855	한가롭	VA	26647	한가__04	NNG
12385	예사롭	VA	17416	예사__01	NNG
18818	이채롭	VA	43022	이채__03	NNG
22434	자애롭	VA	30363	자애__03	NNG
27931	상서롭	VA	39141	상서__01	NNG

위 표의 단어들은 파생어의 순위가 높은 부류에 속한다. 3,000위 내에 속하는 파생어는 '새롭다, 외롭다, 흥미롭다' 세 단어이며, 이 중 '새롭다'의 어근 '새', '흥미롭다'의 어근 '흥미'가 함께 3,000위 내에 든다. 기초 어휘 목록의 처리 방식에 따라 두세 개 정도 차이가 발생하나 큰 문제라고 보기는 어렵다.

다음은 어근의 순위가 '-롭-' 파생어의 순위보다 높은 어휘의 목록이다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
1226	자유롭	VA	653	자유__03	NNG
4113	평화롭	VA	1599	평화__02	NNG
6564	조화롭	VA	1885	조화__07	NNG
7370	지혜롭	VA	2153	지혜__02	NNG
7462	여유롭	VA	1445	여유	NNG
8778	향기롭	VA	1626	향기__01	NNG
10990	호화롭	VA	9964	호화__02	NNG
11101	정의롭	VA	2259	정의__03	NNG
15268	명예롭	VA	2536	명예__01	NNG
16720	자비롭	VA	7052	자비__09	NNG
18534	의롭	VA	2827	의__05	NNG
23685	수고롭	VA	3968	수고__01	NNG
26375	은혜롭	VA	8386	은혜	NNG
29070	호기롭	VA	21343	호기__14	NNG
33330	권태롭	VA	21815	권태__01	NNG
33542	영예롭	VA	11548	영예__03	NNG
44561	보배롭	VA	16836	보배	NNG
48407	허허롭	VA	31348	허허__04	NNG
49798	영화롭	VA	26897	영화__04	NNG

위 목록을 통해 확인할 수 있듯이 이 부류의 단어들은 전반적으로 어근의 순위가 '-롭-' 파생어보다 상위에 있다. 그리고 19개 중 9개의 어근이 3,000위 이내에 든다. 반면에 '-롭-' 접사가 결합된 단어 중 '자유롭다'만이 3,000위 내에 들고 대부분의 파생어가 기초 어휘에 들지 못할 것으로 보인다. 따라서 어휘군이든 개별 어휘이든 처리 방식이 목록에 큰 차이를 가져오지 않을 것으로 보인다.

(3) '-답-' 접미 파생어 검토

'-답-'이 결합한 파생어와 그 어근이 50,000위 안에 함께 나타난 단어는 다음의 3쌍에 불과하다.

접사 포함 단어			어근동형어		
순위	단어	품사	순위	단어	품사
11242	정답	VA	2470	정__20	NNG
18492	참답	VA	7611	참__01	NNG
18788	꽃답	VA	278	꽃__01	NNG

위 표의 단어에서는 어근의 순위가 공통적으로 월등히 높고, 어근의 의미에서 파생어가 쉽게 유추된다는 특징이 보인다. 따라서 이들 목록은 어휘군으로 처리하는 편이 낫다고 판단할 수 있다.

3.2.3. 접사 처리에 대한 방향 모색

지금까지 살펴본 접사들의 양상을 종합하면 대체로 어근보다 파생어의 순위가 낮아 하나의 어휘로 처리하든 개별적 어휘로 처리하든 큰 문제가 발생하지 않는다는 것으로 요약할 수 있다.

하나의 어휘로 처리한다는 것은 가령 '공부'라는 어휘 하나에 '공부하다'를 합쳐 처리한다는 것을 의미한다. 따라서 이 경우 접사 처리에 대해 두 가지 방법이 존재한다는 것을 알 수 있다. 하나는 접사들을 별도의 어휘 항목으로 따로 처리하지 않는 것이고 다른 하나는 접사들을 별도의 어휘 항목으로 따로 처리하는 것이다. 앞의 방법은 '공부'만 어휘 항목이 되고 뒤의 방법은 '공부' 외에 '-하-'도 별도의 어휘 목록이 된다. 이들 방법 가운데 '공부'만 어휘 항목으로 처리하는 경우는 상대적으로 가장 다양한 어휘를 포괄할 수 있다는 장점을 가진다. '-하-'도 별도의 어휘 목록으로 처리하는 경우 접사를 어휘의 하나로 포괄하는 데 따른 부담을 감소해야 한다는 점은 문제이지만 높은 생산성을 지니는 접사를 별도로 제시해 줄 수 있다는 점에서 효용성의 측면에서는 장점으로 작용할 수 있다.

다만 이러한 두 방향으로의 처리는, 많지는 않지만 파생어의 순위가 어근보다 더 높은 경우가 존재한다는 점이 부담이 될 수 있다. 앞에서 살펴본 접사들의 경우 접두사는 '불-' 파생어, 접미사는 '-적' 파생어, '-성' 파생어, '-스럽-' 파생어, '-롭-' 파생어에서 이러한 경우가 발견된 바 있다. 이들은 파생어의 순위가 더 높게 나타나 어근을 중심으로 상관관계에 있는 어휘를 처리하는 것이 불합리하다는 것을 의미한다.

한편 개별 어휘로 처리한다는 것도 접사의 처리와 관련하여 다시 그 경우의 수를 두 가지로 나눌 수 있다. 하나는 개별 어휘로만 처리하고 접사들을 별도의 어휘 항목으로 따로 처리하지는 않는 것이다. 이는 '공부'와 '공부하다'를 별도의 어휘로 처리하되 '-하-'라는 접사를 별도의 어휘로 처리하지는 않는다는 것을 의미한다. 이러한 처리는 어근보다 파생어의 순위가 높게 나타나는 경우에도 아무런 문제를 발생시키지 않는다는 점에서 장점이 있다. 다만 '공부'와 '공부하다'가 서로 의미와 형식의 측면에서 상관관계에 있는데 이를 별도의 어휘로 처리하게 되면서 '공부'만 제시하는 것보다는 상대적으로 어휘의 다양성을 확보할 수 없다는 단점이 있다.

다른 하나는 개별 어휘로 처리하되 접사들을 별도의 어휘 항목으로도 처리하는 것이다. 이는 '공부'와 '공부하다'를 별도의 어휘로 처리하면서 '-하-'라는 접사도 별도의 어휘로 처리한다는 것을 의미한다. 이는 어근보다 파생어의 순위가 높게 나타나는 경우도 문제 삼지 않을 뿐만 아니라 접미사 '-하-'를 통해 '공부'와 '공부하다'의 상관관계를 포착할 수 있다는 장점은 있지만 '-하-'를 어휘의 하나로 처리해야 한다는 점에서 앞서 접사를 어휘로 처리하는 데 따른 부담감을 안아야 한다는 단점이 있다. 또한, 앞의 어떤 방법보다도 어휘의 다양성이 가장 줄어들게 된다는 것은 약점이라고 할 수 있다. 그 결과 서로 상관관계에 놓인 단어들 사이의 잉여성도 부담이 될 수밖에 없다.

기초 어휘에 포함되는 어휘의 다양성 측면에서만 보면 앞의 처리들은 다음과 같이 순서를 매길 수 있을 것이다.

'공부'만 어휘로 처리 > '공부'와 '-하-'를 어휘로 처리 > '공부'와 '공부하다'를 어휘로 처리 > '공부', '공부하다', '-하-'를 모두 어휘로 처리

기존의 연구들을 보면 국어교육이나 한국어교육 모두에서 개별 어휘로의 처리가 더 많다는 것은 아무래도 어휘 항목으로 접사를 포함시키는 데 따른 부담이 작용한 결과로 해석할 수 있다. 그러나 그 차이가 매우 근소하다는 점에도 주목할 필요가 있다. 이는 어휘 항목으로 접사를 포함시키는 데 따른 부담보다는 접사를 따로 처리하는 데서 오는 효용성에 주목한 결과로 해석할 수 있기 때문이다. 따라서 어느 한 방법만을 적용하여 단일한 어휘 순위를 산출하는 대신 이들 각각의 방법에 따라 어휘의 순위를 산출하여 어휘 목록에 대한 다양성을 확보하는 방안을 적극적으로 고려할 필요가 있다.

4. 교과서 어휘 목록을 통한 검토

대규모 언어 자료를 대상으로 하여 이를 통계적 절차에 따라 추출한 어휘 목록은 자료가 지닌 편향성 등 다양한 문제점을 지닐 가능성이 있다. 따라서 이 절에서는 어휘 목록이 지닌 기초성을 검토하기 위하여 초등학교 교과서에 수록된 어휘 목록과 본 연구에서 잠정적으로 선정한 어휘 목록을 비교, 검토하는 작업을 실시하고자 한다.

학교교육에서 사용되고 있는 초등 교과서는 전 국민을 대상으로 한 국가적 출판물로, 가장 초급 단계의 교육 내용을 반영하고 있다는 점에서 기초적인 어휘를 포함하고 있다고 볼 수 있다. 이러한 점에서 초등 교과서에 수록된 어휘 목록은 본 연구에서 선정하고 등급화하고자 하는 기초 어휘의 목록을 검토하기 위한 기준으로서 의미를 지니고 있다.

본 연구에서 사용하고자 하는 초등 교과서의 어휘 목록은 국립국어원에서 기수행한 바 있는 김한샘(2009)의 연구 결과이다. 김한샘(2009)은 초등학교 교과서 13개 과목의 18종 교과서 총 127권을 <표준국어대사전>의 동형어 정보를 기준으로 분류하여 목록화 한 연구로 초등학교 전 과목의 교과서를 포함하고 있다. 본 연구의 기초 어휘 목록 역시도 <표준국어대사전>의 동형어 정보를 반영하고 있고 최대한 다양한 장르의 텍스트를 수집하여 언어 자료를 마련하였기 때문에 두 목록을 통한 상호 비교는 수월하게 진행될 수 있다.

그런데 두 목록이 설정하고 있는 분석 단위의 설정에서 양자의 차이가 있으므로 이에 대해 밝혀 두고자 한다. 김한샘(2009)에서의 목록은 21세기 세종계획의 형식을 그대로 따라 '조사, 어미' 등의 문법소를 포함하여 본 연구와는 차이를 보인다. 본 연구에서는 분석 단위를 '어휘' 단위로 한정하여, 형식 형태소인 '조사' 등은 목록에서 제외하고 예비 목록을 생성하였으므로 양자의 비교를 위해서는 본 연구에서 목록에 등재하지 않은 요소들을 제외하는 전처리 과정을 거친 뒤에 두 목록을 비교하는 작업을 수행하였다.

두 목록의 비교·검토 작업은 편의상 1등급에 속하는 기초 어휘 3,000개 목록을 중심으로 실시하였다. 한편 초등 교과서의 어휘 목록은 모두 23,279개로 확인되었는데, 이 목록과 1등급 기초 어휘 목록을 비교하였다.

1등급 기초 어휘로 선정한 3,000개 목록과 초등 교과서의 어휘 목록을 비교한 결과, 1등급에 속하는 어휘 목록 3,000개 중 총 2,554개의 어휘가 교과서의 어휘 목록에도 포함되는 것으로 확인되었다. 비율상 1등급 기초 어휘의 약 85%의 어휘가 교과서 어휘 목록에 포함되어 있었다. 두 목록에서 공통으로 출현한 어휘를 출현 빈도를 기준으로 각각 100개씩 정리하여 표로 보이면 다음과 같다.

<표 24> 교과서 어휘와 3,000위 어휘 비교: 공통 어휘 100개 목록(출현빈도순)

번호	단어	품사	교과서 어휘 순위	3,000위 어휘 순위	단어	품사	교과서 어휘 순위	3,000위 어휘 순위
1	보__01	VX	1	26	것__01	NNB	4	2
2	하__01	VV	2	4	있__01	VA	3	3
3	있__01	VA	3	3	하__01	VV	2	4
4	것__01	NNB	4	2	있__01	VX	5	6
5	있__01	VX	5	6	수__02	NNB	6	7
6	수__02	NNB	6	7	되__01	VV	9	8
7	하__01	VX	7	9	하__01	VX	7	9
8	나__03	NP	8	24	않	VX	21	10
9	되__01	VV	9	8	없__01	VA	43	11
10	우리__03	NP	10	38	등__05	NNB	52	12
11	알아보	VV	11	1210	년__02	NNB	84	13
12	사람	NNG	12	23	보__01	VV	18	14
13	때__01	NNG	13	22	이__05	MM	54	15
14	생각하	VV	13	86	주__01	VX	31	16
15	만들	VV	15	43	일__07	NNB	202	17
16	읽	VV	16	319	그__01	MM	36	18
17	쓰__01	VV	17	282	같	VA	44	20
18	보__01	VV	18	14	대하__02	VV	32	21
19	친구__02	NNG	19	185	때__01	NNG	13	22
20	말하	VV	20	45	사람	NNG	12	23
21	않	VX	21	10	나__03	NP	8	24
22	일__01	NNG	22	58	가__01	VV	39	25
23	한__01	MM	23	30	보__01	VX	1	26
24	어떤	MM	24	131	위하__01	VV	37	27
25	수__26	NNG	25	595	받__01	VV	116	28
26	개__10	NNB	26	57	그__01	NP	168	29
27	여러	MM	27	229	한__01	MM	23	30
28	어떻	VA	28	114	월__02	NNB	869	31
29	방법	NNG	29	170	지__04	VX	42	32
30	몇	MM	30	226	좋__01	VA	51	33
31	주__01	VX	31	16	이__05	NP	161	34
32	대하__02	VV	32	21	기자__05	NNG	2121	35
33	더__01	MAG	33	37	원__01	NNB	155	36
34	말__01	NNG	34	53	더__01	MAG	33	37
35	내용__02	NNG	35	206	우리__03	NP	10	38
36	그__01	MM	36	18	많	VA	57	40
37	위하__01	VV	37	27	만__06	NR	2724	41

38	무엇	NP	38	201	중__04	NNB	109	42
39	가__01	VV	39	25	만들	VV	15	43
40	잘__02	MAG	40	66	때문	NNB	107	44
41	가지__04	NNB	41	176	말하	VV	20	45
42	지__04	VX	42	32	사진__07	NNG	302	46
43	없__01	VA	43	11	명__03	NNB	101	47
44	같	VA	44	20	알	VV	47	48
45	활동__02	NNG	45	346	크__01	VA	66	49
46	글	NNG	46	267	뉴스	NNG	2086	50
47	알	VV	47	48	오__01	VV	67	51
48	생활	NNG	48	283	따르__01	VV	71	52
49	그림__01	NNG	49	444	말__01	NNG	34	53
50	다음__01	NNG	50	161	통하	VV	208	54
51	좋__01	VA	51	33	나오	VV	97	56
52	등__05	NNB	52	12	개__10	NNB	26	57
53	이야기	NNG	53	153	일__01	NNG	22	58
54	이__05	MM	54	15	시간__04	NNG	124	62
55	모양__02	NNG	55	652	그렇	VA	82	63
56	나타내	VV	56	1294	경우__03	NNG	154	65
57	많	VA	57	40	잘__02	MAG	40	66
58	물__01	NNG	58	227	정도__11	NNG	311	67
59	점__10	NNG	59	154	그리고	MAJ	60	68
60	그리고	MAJ	60	68	싶	VX	74	69
61	듣__01	VV	61	162	보이__01	VV	245	70
62	집__01	NNG	62	97	문제__06	NNG	91	71
63	생각__01	NNG	63	117	함께	MAG	81	72
64	두__01	MM	64	79	먹__02	VV	114	73
65	모두__01	MAG	65	152	지나	VV	304	75
66	크__01	VA	66	49	곳__01	NNG	76	76
67	오__01	VV	67	51	다른	MM	71	77
68	나라__01	NNG	68	334	안__02	MAG	244	78
69	모습__01	NNG	69	128	두__01	MM	64	79
70	따르__01	VV	71	52	못하	VX	151	80
71	다른	MM	71	77	바로__02	MAG	551	81
72	찾	VV	73	137	후__08	NNG	143	82
73	싶	VX	74	69	이런__01	MM	341	83
74	안__01	NNG	75	212	가장__01	MAG	100	84
75	곳__01	NNG	76	76	저__03	NP	159	85
76	살__01	VV	77	102	생각하	VV	13	86
77	사용하__03	VV	78	139	이번__01	NNG	677	87

78	가지	VV	79	109	번__04	NNB	157	88
79	우리나라	NNG	80	565	전__08	NNG	250	89
80	함께	MAG	81	72	그것	NP	407	91
81	그렇	VA	82	63	자신__01	NNG	246	92
82	이용하__01	VV	83	268	좀__02	MAG	218	93
83	년__02	NNB	84	13	많이	MAG	92	95
84	어머니__01	NNG	85	692	속__01	NNG	130	96
85	그리__02	VV	86	426	집__01	NNG	62	97
86	마음__01	NNG	89	145	주__01	VV	98	98
87	위__01	NNG	90	186	다시__01	MAG	94	99
88	문제__06	NNG	91	71	금지__04	NNG	7630	100
89	많이	MAG	92	95	모르	VV	231	101
90	어느__01	MM	93	240	살__01	VV	77	102
91	다시__01	MAG	94	99	아이__01	NNG	148	103
92	왜__02	MAG	95	259	ㅋ	NNG	9389	104
93	넣	VV	96	228	밝히	VV	1483	105
94	나오	VV	97	56	데__01	NNB	141	106
95	주__01	VV	98	98	이렇	VA	235	107
96	나누	VV	99	393	모든	MM	354	108
97	가장__01	MAG	100	84	가지	VV	79	109
98	명__03	NNB	101	47	무단__02	NNG	12683	110
99	놀이__01	NNG	102	2279	대통령	NNG	1229	112
100	이야기하	VV	103	1358	배포__01	NNG	16068	113

초등 교과서는 교육의 대상이 되는 초등학생의 수준을 고려하여 집필되므로 교과서 집필자의 직관을 토대로 기초성이 있는 어휘를 최대한 살려 쓰는 방법으로 마련된 텍스트라고 가정할 수 있다. 그런데 본 연구에서 기반으로 삼은 언어 자료는 일반 국민이 다양한 목적에서 향유하는 텍스트 모둠을 주로 활용하였으므로 두 목록은 상당한 차이를 보일 가능성도 있었다. 그러나 결과적으로 두 목록이 85% 정도의 높은 일치도를 보임으로써 본 연구의 기초 어휘 목록이 지닌 기초성을 방증할 만한 결과를 보였다.

위의 <표 24>를 중심으로 살펴볼 때, 교과서에서 높은 출현 빈도를 보이는 어휘는 대체로 기초 어휘 목록으로 선정된 것들 가운데에서도 출현 빈도가 매우 높은 말들이 대부분임을 알 수 있다. 그러나 교과서에서는 매우 높은 빈도를 보이지만 언어 자료를 기반으로 선정한 기초 어휘 목록에서는 상대적으로 순위가 낮은 어휘가 일부 존재하는데 이는 교과서에서 흔히 사용되는 표현 방식이나 발문, 교육 활동 용어 등 교과서의 장르성에 영향을 받은 데 따른 것으로 판단된다. 예컨대, '알아보-'는 교과서의 어휘 순위에서는 11위로 매우 높은 출현 빈도를 보이는 어휘에

해당하지만 기초 어휘 목록에서는 1,210위로 상대적으로 매우 떨어지는 출현 빈도를 보인다. 이는 1~10위에 해당하는 어휘들의 경우에 비해 큰 차이를 보이는데, 이는 교과서 특유의 교육용 발문이나 표현 방식, 용어 등으로 인하여 자주 반복되는 표현이라는 것에서 말미암은 것으로 보인다. 이러한 예에는 '알아보-' 외에도 '나타내-, 놀이01' 등의 어휘도 해당된다.

한편 3,000위 안에 드는 목록을 중심으로 비교할 때, 기초 어휘의 목록에는 없지만 교과서에만 있는 목록은 모두 20,720개이며, 반대로 기초 어휘에만 있는 목록은 447개였다. 전자는 교과서가 교육용 텍스트이므로 전문어를 수록한다든가 하는 연유로 인하여 기초성이 떨어지는 목록이라고 판단할 여지가 크나 본 연구에서 사용한 언어 자료가 과소 추정한 어휘일 가능성도 없지 않으므로 향후 정성적 판단이 필요한 목록이라고 할 수 있다. 그리고 후자는 초등학교 교과서에는 3,000위 안에 들지 않았으나 기초 어휘 목록에는 포함되는 것이므로 이 역시도 어휘의 기초성이 과대 추정되지 않았는지에 대한 향후의 판단이 필요하다.

다음 표는 본 연구의 목록에는 없지만, 초등 교과서에만 있는 목록을 출현 빈도에 따라 50개만 제시해 본 것이다.

<표 25> 교과서 어휘에만 있는 단어의 예

번호	교과서 어휘 순위	단어	품사
1	70	알맞	VA
2	87	또	MAJ
3	88	쪽__02	NNG
4	108	까닭	NNG
5	119	아니	VA
6	134	센티미터	NNB
7	147	공__01	NNG
8	172	모둠	NNG
9	176	분수__06	NNG
10	179	미터__02	NNB
11	180	계산하	VV
12	185	써넣	VV
13	192	식__04	NNG
14	198	규칙__02	NNG
15	199	물음__01	NNG
16	203	익히__02	VV
17	206	도형__03	NNG
18	209	낱말__02	NNG
19	212	표__02	NNG
20	218	고장__01	NNG
21	238	소수__04	NNG

22	253	관찰하	VV
23	261	넓이	NNG
24	275	걸음	NNG
25	283	모형__05	NNG
26	289	주고받	VV
27	307	조상__07	NNG
28	317	삼각형	NNG
29	325	실천하__01	VV
30	326	재__02	VV
31	350	되돌아보	VV
32	353	주의하__01	VV
33	355	교실	NNG
34	359	변__09	NNG
35	362	상자__10	NNG
36	400	보기__01	NNG
37	404	곱셈	NNG
38	407	물체	NNG
39	412	각__02	NNG
40	435	짝__01	NNG
41	438	생물__01	NNG
42	438	원__12	NNG
43	465	세로__01	NNG
44	468	가로__01	NNG
45	468	킬로그램	NNB
46	473	퍼센트	NNB
47	478	문화재	NNG
48	483	백성	NNG
49	483	글쓴이	NNG
50	483	나눗셈	NNG

이러한 목록을 보면 양자의 불일치는 본 연구의 어휘 목록이 지닌 문제라기보다는 교과서에서 자주 사용되는 관습적 표현, 각 교과목의 교육용 전문어가 일상 언어에서는 빈번하게 사용되지 않음에 따라 일어난 현상인 것으로 보인다.

<표 26> 3,000위 어휘에만 있는 단어의 예

번호	3,000위 어휘 순위	단어	품사
1	1	이	VCP
2	5	들__09	XSN
3	19	아니	VCN

4	39	및	MAG
5	55	제__21	XPN
6	59	한국__05	NNP
7	60	서울__01	NNP
8	61	하	XSV
9	64	또	MAG
10	74	미국__03	NNP
11	90	재__17	XPN
12	94	시키	XSV
13	111	적__18	XSN
14	126	전재__10	NNG
15	142	일본__02	NNP
16	144	뉴시스	NNP
17	146	중국__01	NNP
18	168	제__01	NP
19	178	님__04	XSN
20	189	내__04	NP
21	190	연합뉴스	NNP
22	213	되	XSV
23	231	북한__03	NNP
24	233	제보	NNG
25	239	성__17	XSN
26	256	여__27	XSN
27	265	자__31	XSN
28	311	한__11	NNP
29	314	부__23	XSN
30	320	화__16	XSN
31	323	간__16	XSN
32	345	김__06	NNP
33	357	페이스북	NNP
34	369	중__05	NNP
35	373	씩__03	XSN
36	377	저작권자	NNG
37	411	장__41	XSN
38	416	용__11	XSN
39	435	미__14	NNP
40	437	네이버	NNP
41	456	알려지	VV
42	460	상__26	XSN
43	461	드리	XSV

44	475	공감	NNG
45	486	사__41	XSN
46	490	별__04	XSN
47	507	형__08	XSN
48	519	데일리	NNP
49	527	비__32	XPN
50	532	부산__02	NNP

반면에 위의 표는 1등급 기초 어휘 목록에는 있지만 교과서에서는 존재하지 않는 어휘 목록을 제시한 것이다. 본 연구에서 포함하고 있는 언어 자료는 신문, 언론 자료를 대량으로 사용하여 일부 언론사나 포털 사이트의 명칭이 포함되어 있으며, 일상 언어생활에서 흔히 쓰이는 어휘, 지역명 등도 포함되어 있음이 확인된다. 특히 고유명사에 해당하는 언론사 명칭 등의 문제는 향후 정성적 검토를 통해 목록을 정교화 하여야 할 것이다.

참고로 1등급에 해당하는 어휘 목록 외에 2등급 이상 기초 어휘 목록과 교과서 어휘 목록 50,000개와 교과서 어휘 목록에 공통으로 출현한 어휘는 총 16,760개였다. 이는 비율상 2등급 이상 기초 어휘 50,000개의 약 36%의 어휘가 교과서 어휘 목록과 겹치는 것이라고 할 수 있다. 다만, 2등급 이하의 기초 어휘 목록은 기초성을 따지기에 매우 방대한 목록이므로, 여기서는 출현 빈도를 중심으로 순위 100위에 이르는 목록만을 각각 보인다.

<표 27> 교과서 어휘와 50,000위 어휘 비교: 공통 어휘 100개 목록(출현빈도순)

번호	단어	품사	교과서 어휘 순위	50,000위 어휘 순위	단어	품사	교과서 어휘 순위	50,000위 어휘 순위
1	보__01	VX	1	26	것__01	NNB	4	2
2	하__01	VV	2	4	있__01	VA	3	3
3	있__01	VA	3	3	하__01	VV	2	4
4	것__01	NNB	4	2	있__01	VX	5	6
5	있__01	VX	5	6	수__02	NNB	6	7
6	수__02	NNB	6	7	되__01	VV	9	8
7	하__01	VX	7	9	하__01	VX	7	9
8	나__03	NP	8	24	않	VX	21	10
9	되__01	VV	9	8	없__01	VA	43	11
10	우리__03	NP	10	38	등__05	NNB	52	12
11	알아보	VV	11	1210	년__02	NNB	84	13
12	사람	NNG	12	23	보__01	VV	18	14
13	때__01	NNG	13	22	이__05	MM	54	15
14	생각하	VV	13	86	주__01	VX	31	16
15	만들	VV	15	43	일__07	NNB	202	17
16	읽	VV	16	319	그__01	MM	36	18
17	쓰__01	VV	17	282	같	VA	44	20

18	보__01	VV	18	14	대하__02	VV	32	21
19	친구__02	NNG	19	185	때__01	NNG	13	22
20	말하	VV	20	45	사람	NNG	12	23
21	않	VX	21	10	나__03	NP	8	24
22	일__01	NNG	22	58	가__01	VV	39	25
23	한__01	MM	23	30	보__01	VX	1	26
24	어떤	MM	24	131	위하__01	VV	37	27
25	수__26	NNG	25	595	받__01	VV	116	28
26	개__10	NNB	26	57	그__01	NP	168	29
27	여러	MM	27	229	한__01	MM	23	30
28	어떻	VA	28	114	월__02	NNB	869	31
29	방법	NNG	29	170	지__04	VX	42	32
30	몇	MM	30	226	좋__01	VA	51	33
31	주__01	VX	31	16	이__05	NP	161	34
32	대하__02	VV	32	21	기자__05	NNG	2121	35
33	더__01	MAG	33	37	원__01	NNB	155	36
34	말__01	NNG	34	53	더__01	MAG	33	37
35	내용__02	NNG	35	206	우리__03	NP	10	38
36	그__01	MM	36	18	많	VA	57	40
37	위하__01	VV	37	27	만__06	NR	2724	41
38	무엇	NP	38	201	중__04	NNB	109	42
39	가__01	VV	39	25	만들	VV	15	43
40	잘__02	MAG	40	66	때문	NNB	107	44
41	가지__04	NNB	41	176	말하	VV	20	45
42	지__04	VX	42	32	사진__07	NNG	302	46
43	없__01	VA	43	11	명__03	NNB	101	47
44	같	VA	44	20	알	VV	47	48
45	활동__02	NNG	45	346	크__01	VA	66	49
46	글	NNG	46	267	뉴스	NNG	2086	50
47	알	VV	47	48	오__01	VV	67	51
48	생활	NNG	48	283	따르__01	VV	71	52
49	그림__01	NNG	49	444	말__01	NNG	34	53
50	다음__01	NNG	50	161	통하	VV	208	54
51	좋__01	VA	51	33	나오	VV	97	56
52	등__05	NNB	52	12	개__10	NNB	26	57
53	이야기	NNG	53	153	일__01	NNG	22	58
54	이__05	MM	54	15	시간__04	NNG	124	62
55	모양__02	NNG	55	652	그렇	VA	82	63
56	나타내	VV	56	1294	경우__03	NNG	154	65
57	많	VA	57	40	잘__02	MAG	40	66
58	물__01	NNG	58	227	정도__11	NNG	311	67
59	점__10	NNG	59	154	그리고	MAJ	60	68
60	그리고	MAJ	60	68	싶	VX	74	69
61	듣__01	VV	61	162	보이__01	VV	245	70
62	집__01	NNG	62	97	문제__06	NNG	91	71
63	생각__01	NNG	63	117	함께	MAG	81	72
64	두__01	MM	64	79	먹__02	VV	114	73
65	모두__01	MAG	65	152	지나	VV	304	75

66	크__01	VA	66	49	곳__01	NNG	76	76
67	오__01	VV	67	51	다른	MM	71	77
68	나라__01	NNG	68	334	안__02	MAG	244	78
69	모습__01	NNG	69	128	두__01	MM	64	79
70	알맞	VA	70	5647	못하	VX	151	80
71	따르__01	VV	71	52	바로__02	MAG	551	81
72	다른	MM	71	77	후__08	NNG	143	82
73	찾	VV	73	137	이런__01	MM	341	83
74	싶	VX	74	69	가장__01	MAG	100	84
75	안__01	NNG	75	212	저__03	NP	159	85
76	곳__01	NNG	76	76	생각하	VV	13	86
77	살__01	VV	77	102	이번__01	NNG	677	87
78	사용하__03	VV	78	139	번__04	NNB	157	88
79	가지	VV	79	109	전__08	NNG	250	89
80	우리나라	NNG	80	565	그것	NP	407	91
81	함께	MAG	81	72	자신__01	NNG	246	92
82	그렇	VA	82	63	좀__02	MAG	218	93
83	이용하__01	VV	83	268	많이	MAG	92	95
84	년__02	NNB	84	13	속__01	NNG	130	96
85	어머니__01	NNG	85	692	집__01	NNG	62	97
86	그리__02	VV	86	426	주__01	VV	98	98
87	쪽__02	NNG	88	3495	다시__01	MAG	94	99
88	마음__01	NNG	89	145	금지__04	NNG	7630	100
89	위__01	NNG	90	186	모르	VV	231	101
90	문제__06	NNG	91	71	살__01	VV	77	102
91	많이	MAG	92	95	아이__01	NNG	148	103
92	어느__01	MM	93	240	ㅋ	NNG	9389	104
93	다시__01	MAG	94	99	밝히	VV	1483	105
94	왜__02	MAG	95	259	데__01	NNB	141	106
95	넣	VV	96	228	이렇	VA	235	107
96	나오	VV	97	56	모든	MM	354	108
97	주__01	VV	98	98	가지	VV	79	109
98	나누	VV	99	393	대통령	NNG	1229	112
99	가장__01	MAG	100	84	어떻	VA	28	114
100	명__03	NNB	101	47	앞	NNG	126	115

Ⅵ. 종합 및 제언

1. 요약

본 연구에서는 기초 어휘 선정 및 등급화 사례를 검토하여 기초 어휘 선정 및 등급화의 방법론을 구축하고, 이를 추출하기 위한 목적의 언어 자료를 정제하였다. 이를 통해 어휘 등급화 방법론을 체계적으로 수립하고, 2017년 연구에서 구축된 언어 자료를 보완함으로써 기초 어휘 목록의 타당성을 높이는 데 그 목적이 있다. 연구의 주요 내용을 정리하면 다음과 같다.

1) 기초 어휘 선정 및 어휘 등급화 사례

본 연구는 2017년에 수행된 “기초 어휘 선정 및 어휘 등급화를 위한 기초 연구”의 후속 연구로서 해당 연구에서 진행한 기초 어휘 선정 및 어휘 등급화의 사례 조사에 내용을 추가하고 보완하는 방향으로 조사 및 분석하여 기초 어휘 선정 및 어휘 등급화 방향을 수립하였다. 먼저, 기초 어휘 연구 사례를 추가 검토하여 2017년 연구에서 제시된 정의인 ‘일상생활에서 사용 빈도가 높고 파생이나 합성들의 조어에 고빈도로 참여하여 다른 단어로 대체하기 어려운 특성을 지닌 단어’를 따르고자 하였다. 또한 기초 어휘 선정 및 어휘 등급 평정의 검토 작업으로 교과서 어휘와의 비교 분석을 위해 교과서 어휘 연구 및 어휘 목록의 사례를 검토한 결과 기초 어휘 목록과 교과서 어휘 목록 사이의 상관성을 확인하였고, 이를 기반으로 어휘 등급화 방법론 수립에 활용하였다. 마지막으로 국외 사례는 2017년 연구에서 영어와 일본어 사례를 살핀 것에 더하여 유럽권의 프랑스어 사례와 아시아권의 중국어 사례를 조사하였으며, 특히 프랑스어 사례에서 MANULEX의 어휘 통계 방식을 본 연구의 통계 방식과 비교하는 데에 활용하였다.

2) 기초 어휘 추출 목적의 언어 자료 정제와 처리 과정

2017년 연구에서 구축한 20억 어절의 샘플 언어 자료를 아래와 같이 양적·질적으로 보완하였다.

○ 2017년 연구의 언어 자료는 문어적 성격이 큰 기존의 말뭉치들(세종 현대 문어 원시 말뭉치, 도서 말뭉치 등)을 활용한 것이었으므로, 구어적 성격을 보완하기 위하여 인터넷 게시판, 홈쇼핑 구어 자료 등을 수집하여 총 45억 어절의 언어 자료를 수집하였다.

○ 줄 바꿈 문자의 처리, 어문 규범 위반 사례의 처리 등 언어 자료의 오류를 수정하고, 통계 수치에 영향을 줄 수 있는 장르 구분을 체계화하기 위한 「신문」 장르 세분 실험 등을 실시하여 통계 수치의 신뢰성을 제고할 수 있는 방향으로 언어 자료를 질적으로 보완하였다.

○ 언어 자료로부터 어휘 자료를 추출하기 위한 형태소 분석 모델을 개선하고, 후처리 모듈을 개발하여 어휘 자료 추출의 정확도를 제고하였다.

향후 연구에서는 언어 자료의 양적 규모의 확장보다는 장르 체계화와 형태소 분석 모델 개선에 초점을 맞추어야 할 것이다. 특히 형태소 분석 모델은 지침 마련 작업과 언어 자료의 통계적 처리가 유기적으로 이루어져야 할 것이므로, 기초 어휘 선정 및 어휘 등급화 방법론과 쟁점을 고려한 형태소 분석 모델이 만들어질 수 있도록 다양한 경우를 고려해 볼 수 있을 것이다.

3) 어휘 등급화의 통계적 방법론 수립

언어 자료를 기반으로 하여 기초 어휘를 추출할 때 기반 자료로 활용하는 통계적 방법론을 수립하고 이를 정교화하였다. 어휘 형태소의 상대 빈도, 범위, 산포도에 가중치를 부여한 어휘 점수를 산출하였으며, 이를 정교화하기 위하여 한국어를 모국어로 하는 화자들의 직관을 활용하였다. 그 결과 '0.265*빈도+0.64*범위+0.095*산포도'가 한국어 모어 화자와 가장 일치하는 어휘 점수임을 밝혔다. 또한, 이를 기반으로 하여 추출한 어휘를 MANULEX의 U값과 비교한 결과, 두 가지 단어 점수 산출 방식이 수렴함을 확인하였다.

통계적 방법으로 얻어진 단어 순위는 순전히 언어 자료 기반이므로 언어 전문가의 눈으로 보면 미흡한 점이 눈에 띌 수 있다. 이러한 미흡한 점이 언어 자료에서 비롯하는지, 통계적 방법론 자체에서 비롯하는지 확인하고 그 개선점을 모색해야 할 것이다. 양적 방법론에 따라 도출된 어휘 목록에 대한 정성적 검토를 위해 언어 전문가의 직관 및 다른 어휘 목록과의 비교가 가능할 것이다.

4) 어휘 등급화의 정성적 방법론 수립

통계적 분석으로는 개별 어휘들이 지닌 특성과 어휘 간의 관계들에 대해 정교한 분석이 어려우므로 이들 어휘에 대한 정성적 분석이 필요하다. 정성적 분석에는 전문가들의 직관에 따른 분석과 교과서 어휘와의 비교 작업을 실시하였다. 먼저, 정성적 분석을 위해 고려해야 하는 쟁점을 '동형어 처리'와 '기초 어휘의 범위'로 나누었

는데, 그 세부 쟁점을 정리하면 다음과 같다.

동형어 처리 관련 세부 쟁점	기초 어휘 범위 관련 세부 쟁점
가. 품사 통용의 처리	가. 조사와 어미의 포함 여부
나. 본용언과 보조 용언의 처리	나. 접사의 포함 여부
다. 어근 어휘와 파생어 어휘의 처리	다. 고유 명사의 포함 여부
라. 상위 품사와 하위 품사의 처리	라. 감탄사의 포함 여부

이들 세부 쟁점의 일부를 검토한 결과 다음과 같은 등급화 방향을 제시하였다.

○ 품사 통용 어휘 중 3000위 기초 어휘에 드는 경우 목록의 다양성 확보를 위해 해당 품사를 모두 제시하고 하나의 목록으로 제시한다.
○ 의존명사, 보조용언과 같이 하나의 품사로만 사용되거나 부사-접속부사와 같이 서로 층위가 다른 품사 정보로 인한 동형어는 하나의 품사로 처리한다.
○ 접사의 경우 어근보다 파생어의 순위가 낮아 이들을 하나의 어휘로 처리하든 개별적 어휘로 처리하든 큰 문제가 발생하지 않는다. 그러나 기초 어휘에 포함되는 어휘의 다양성을 확보하는 방안을 적극적으로 고려할 필요가 있다.

마지막으로, 김한샘(2009)의 초등 교과서 어휘 목록과 본 연구의 통계적 방법론을 적용한 예비 어휘 목록을 비교함으로써 시사점을 제시하였다. 초등 교과서 어휘 목록과 본 연구의 예비 어휘 목록은 85% 정도의 높은 일치도를 보임으로써 본 연구의 예비 어휘 목록이 지닌 기초성을 방증하는 데 주목할 만한 결과를 보였다. 다만, 일치하지 않는 단어들 중 과소 추정된 어휘가 있을 가능성이 있으므로 정성적 판단이 필요하다. 특히 본 연구의 언어 자료가 신문, 언론 자료를 대량으로 사용하여 일부 언론사나 포털 사이트의 명칭, 지역명 등이 포함되어 있음이 확인되므로 향후 정성적 검토를 통한 목록의 정교화가 필요하다.

2. 제언

2.1 향후 추진 내용

본 연구는 기초 어휘 사업의 중장기 계획에 따르면, ‘토대 확보 단계’, ‘발전 및 확장 단계’, ‘지속 발전 가능 단계’ 중 ‘토대 확보 단계’에 해당한다. 향후 이루어져야 할 연구 내용은 차년도 수행 내용과 그 이후의 수행 내용으로 나누어 살피기로 한다.

차년도 수행 내용은 크게 두 가지로 이루어진다고 볼 수 있다. 첫째, 당해 연도에 이루어진 정성적 방법론에서 다루어지지 못한 구체적 쟁점을 확인하고 결정할 필요가 있다. 언어 자료를 기반으로 하고 대용량의 언어 자료를 기반으로 하며 자국민을 대상으로 한 기초 어휘의 목록 추출 및 등급화는 현재까지 거의 이루어지지 못한 사업 분야라고 할 수 있다. 이에 따라 목록의 구성 및 순위 등에 영향을 미치는 각각의 쟁점을 확인하고 그것을 편의적으로 해결하기보다는 각 결정이 영향을 미치는 범위를 파악하여 면밀하게 결정함으로써 이후의 연구에 대한 토대를 쌓을 필요가 있다. 당해 연도에 모두 결정하지 못한 사항(예: 감탄사, 고유명사, 접사 등), 특히 정성적 방법론(Ⅴ장 참조)의 쟁점 사항을 추가로 확정하되 각 쟁점 사항에 대한 해결 과정, 예상 결과 등도 함께 검토되어야 할 것이다.

둘째, 모국어 화자를 대상으로 하는 기초 어휘는 전문어가 필연적으로 포함될 것이다. 차년도에는 국민의 의사소통에 필요한 전문어를 중심으로 분야별 기초 어휘 선정 작업이 이루어져야 할 것이다. 전 분야의 구분, 전문어의 유형, 전문어의 규모 등이 주요 연구 내용이다.

셋째, 기초 어휘 추출 목적의 언어 자료는 지속적으로 정련화되어야 하며 언어 자료 자체에 대한 타당성 검증도 필요하다. 기초 어휘를 추출하기 위한 목적의 언어 자료는 국민의 언어 사용 양상을 충실히 반영하여야 하므로 예비 목록 추출 이전까지의 지속적 보완이 이루어져야 하며, 기구축된 언어 자료에 대한 검증, 수정, 보완 또한 필수적이다. 이를 위해 전문가 집단을 편성하여 말뭉치의 대표성 등을 검증하고, 기초 어휘 예비 목록을 추출했을 때의 대표성을 높여갈 수 있는 방안을 강구할 필요가 있다. 또한 당해 연도까지 이루어진 미국, 유럽 등에서의 사례 연구에 더하여, 어휘 목록의 활용, 활용 체계의 개발이라는 관점에서 그 접근의 폭을 넓혀 향후 활용 체계가 적실하게 개발될 수 있도록 하여야 할 것이다.

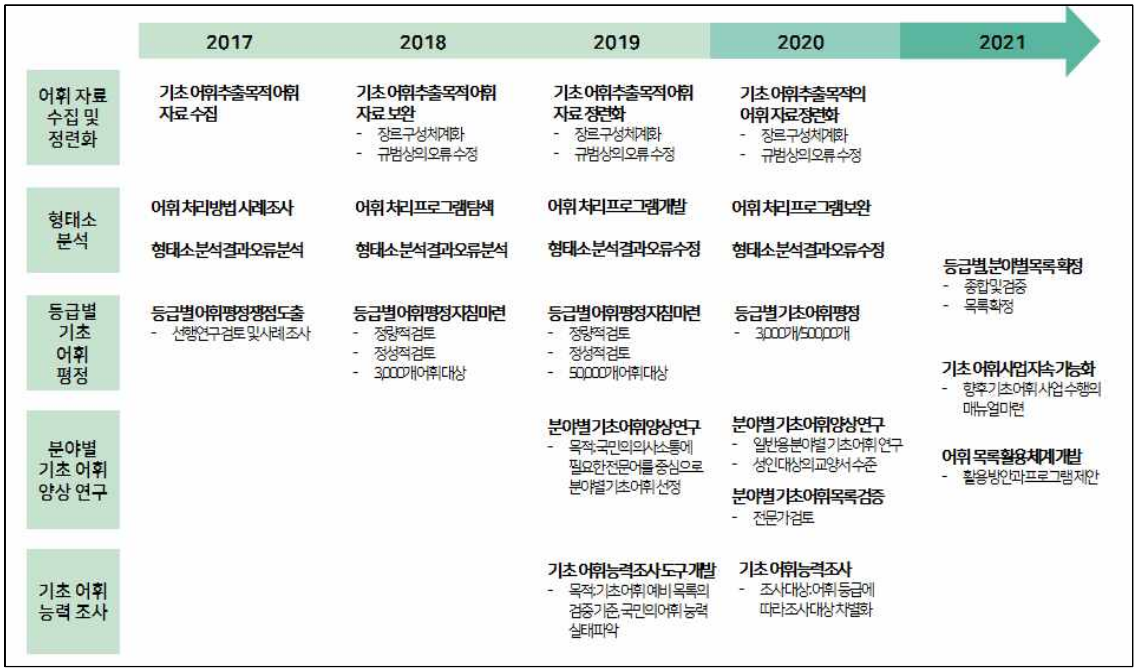
차년도 이후의 연구는 중장기 계획상 발전 및 확장 단계, 지속 발전 가능 단계에 해당한다. 이에 각 단계에서 수행하여야 할 핵심 과업을 제시해 보면 다음과 같다.

먼저, 발전 및 확장 단계에서는 기초 어휘 예비 목록을 추출하고 이를 토대로 등

급화를 완수하는 것이 필요하다. 다만 이 단계에서 제안된 목록은 최종 목록이 아니며 각계 전문가 집단의 판단이나 향후 활용 체계 등을 고려하여 수정될 수 있으며 이를 위한 하나의 판단 기준으로서 국민의 어휘 능력 조사를 병행하여 실시하거나 관련 자료를 통해 보완할 수 있을 것이다.

다음으로, 지속 발전 가능 단계에서는 전 국민을 대상으로 한 기초 어휘의 웹기반 활용 체계를 개발하기 위한 연구를 수행하는 것이 필수적이다. 기초 어휘 사업은 일회성으로 이루어지기보다는 지속적으로 변화하는 국민의 언어 실태를 반영하고 각 시기의 언어 실태를 파악하는 누적적 자료로 활용됨이 타당하다. 따라서 국민이 손쉽게 다양한 목적으로 활용할 수 있는 기초 어휘 및 관련 자료의 검색, 활용 체계를 구안하여 제시하고, 이를 언어 변화에 따라 지속적으로 보완할 수 있는 방안을 마련하는 것도 중요한 과업이다.

이상의 내용을 도식화하여 보이면 다음과 같다.



[그림 15] 기초 어휘 사업 개요

2.2 정책 제언

본 연구는 국민의 국어 능력 발전을 궁극적 목적으로 하여 국어의 기초 어휘를 선정하고 등급화하기 위한 사업의 일환으로 수행되었다. 이에 따라 사업의 본래 목적과 향후 이루어질 사업의 내용에 주목하여 정책 방향을 제언해 보도록 한다.

○ 중장기 계획 및 기존 사업의 성과에 기반한 연구 추진

2017~2018년도에 수행된 연구 사업의 성과를 충실히 이해하고 이에 기반한 연구의 추진이 필요하다. 이를 위해서는 본 연구에서 고려한 쟁점 사항 등을 면밀히 고려하여 이를 보완할 수 있는 차원에서 향후 연구 사업 내용이 선정되고 추진되어야 할 것이다(6.2.1. 참조).

또한 연구의 수행의 효율성을 제고하기 위한 공조 체제의 구축도 요청된다. 이를 위하여 국립국어원이 보유한 다양한 성과물을 활용할 수 있도록 적극적으로 지원하고 유사한 사업 수행 실적이 있는 각급 기관과의 공조를 확대할 수 있는 체계를 마련하여 시행할 필요가 있다.

○ 어휘에 기반한 국어 능력 발전 사업의 기획 및 추진

어휘 능력은 국어 능력의 근간으로서 그 중요성에 비해 접근이 덜하였던 대표적 분야라고 할 수 있다. 이러한 점에서 국어 기초 어휘에 관한 사업 이외에도 어휘를 기반으로 국민의 국어 능력을 증진시킬 수 있는 영역은 적지 않다. 예컨대 직능 등 각 분야별로 익혀야 할 어휘 목록을 개발한다거나 국민의 연령별, 성별에 따른 어휘 사용 양상 분석, 연령별 어휘 능력 조사 및 도구 개발 등도 이제껏 미처 관심을 기울이지 못했으나 향후 연구가 추진되어야 할 분야라고 할 수 있다.

참고문헌

강병규·손민정(2016),「2015 개정 중국어교육과정의 기본 어휘 선정 및 활용」,『중국언어연구』63, 한국중국언어학회.

고영근·구본관(2007),『우리말문법론』, 집문당.

국립국어연구원(2002),『기본 어휘 선정 및 사용 실태 조사를 위한 기초 연구』, 국립국어연구원.

국립국어연구원(2002),『현대 국어 사용 빈도 조사-한국어 학습용 어휘 선정을 위한 기초 조사』, 국립국어연구원.

국립국어연구원(2003),『한국어 학습용 어휘 선정 결과 보고서』, 국립국어연구원.

국립국어원(2015a),『2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구』, 국립국어원.

국립국어원(2015b),『2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업』, 국립국어원.

국어연구소(1986),『국민학교 교육용 어휘 (1·2·3학년용)』, 국어연구소.

국어연구소(1987),『국민학교 교육용 어휘 (4·5·6학년용)』, 국어연구소.

국어연구소(1988),『중학교 교과서 어휘 (국어·국사)』, 국어연구소.

국어연구소(1989),『중학교 교과서 어휘 (도덕·사회)』, 국어연구소.

권민재(2016),「독일어 교육과정에서 기본 어휘 선정의 문제」,『외국어로서의 독일어』, 39, 한국독일어교육학회.

김경선(1998),「초등학교 2학년 국어 읽기 교과서의 어휘 조사」,『초등국어교육』8, 서울교육대학교 초등교육연구소.

김광해(1988),「이차 어휘의 교육에 대하여」,『先淸語文』16(1), 서울대학교 국어교육과.

김광해(1989),『고유어와 한자어의 대응 현상』, 탑출판사.

김광해(1993),『국어 어휘론 개설』, 집문당.

김광해(2003),「국어교육용 어휘와 한국어교육용 어휘」,『국어교육』111, 한국어교육학회.

김광해(2003),『등급별 국어교육용 어휘』, 박이정.

김문오(2007),『남북 교과서 학술 용어 비교 연구』, 국립국어원.

김석영(2014),「중국어 어휘론에서 기본 어휘의 문제-중국식 기본 어휘 개념의 어휘론적 유용성에 대한 비판적 검토」,『중국언어연구』52, 한국중국언어학회.

김언자(2006),「고등학교 교육과정에서의 프랑스어 기본 어휘 선정에 대하여」,『프랑스어문교육』23, 한국프랑스어문교육학회.

김종로(1990),「기본 불어의 연구」,『불어불문학연구』28호, 한국불어불문학회.
김종학(2001),『한국어의 기초 어휘 연구』, 박이정.
김한샘(2003),『한국 현대 소설의 어휘 조사 연구』, 국립국어연구원.
김한샘(2005),『현대 국어 사용 빈도 조사2』, 국립국어원.
김한샘(2009),『초등학교 교과서 어휘 조사 연구』, 국립국어원.
김한샘(2010),「국어교육용 어휘 선정을 위한 교과서 어휘 조사 연구-초등학교 교과서 어휘 분석」,『국어교육연구』47, 국어교육학회.
김한샘(2011),「교육용 어휘 선정을 위한 단어족 분석 연구」,『한말연구』29, 한말연구학회.
김한샘(2012a),「어휘 교육을 위한 사용 어휘 분석 연구 -초등학교 작문 어휘 조사를 기반으로」,『겨레어문학』48, 겨레어문학회.
김한샘(2012b),「한국어 어휘 계량 연구의 성과」,『한민족문화연구』41, 한민족문화학회.
김한샘(2013),「교육용 어휘 선정을 위한 접미사의 생산성 연구-고유어 명사 파생 접미사의 분석」,『한국어의미학』40, 한국어의미학회.
김한샘(2013),「교육용 접사 선정을 위한 명사 파생 접미사 빈도 연구」,『언어와 문화』9(1), 한국언어문화교육학회.
김한샘(2014),「교육용 어휘 선정을 위한 접미사의 의미 예측성 연구」,『한국어의미학』44, 한국어의미학회.
김한샘(2015),「교육용 어휘 선정을 위한 접두사의 생산성 연구」,『우리말 글』65, 우리말글학회.
김한샘·서상규(1998),「말뭉치의 구축과 활용 -연세 말뭉치 1의 구상과 실제-」,『언어정보개발연구』1, 연세대학교 언어정보연구원.
김현철·조은경(2010),「중국학계의 중국어 기본 어휘 선정 현황과 활용방안 연구」,『중국어문학논집』60, 중국어문학연구회.
김흥규·강범모(1995),「고려대학교 한국어 말모둠 1(KOREA-1 CORPUS): 설계 및 구성」,『한국어학』3, 한국어학회.
김희진(1990),「중학교 교육용 어휘에 대한 연구」,『국어교육』71, 한국국어교육연구회.
남기심·고영근(2014),『표준국어문법론(4판)』, 박이정.
문교부(1956),『우리말 말수 사용의 잦기 조사』, 문교부.
민경모(2011),「해외 청소년 대상 교육용 어휘 선정을 위한 기초 연구: 해외 청소년용 교재에 나타난 어휘의 계량적 분석을 중심으로」,『언어와 문화』7(2), 한국언어문화교육학회.
민현식(2004),『초등학교 교과서 한자어 및 한자 분석 연구』, 국립국어원.

민현식(2004),『중학교 교과서 한자어 및 한자 분석 연구』, 국립국어원.
박재현(2007),『교과서 표기 감수 지침 시안』, 국립국어원.
배재석·임승규(2005),「고교 중국어 기본 어휘 만족도 조사 연구」,『중국어문학논집』32, 중국어문학연구회.
배주채(2010),『한국어 기초어휘집』, 한국문화사.
서덕현(1990),「기본 어휘의 개념과 기초 어휘의 위상: 교육용 어휘를 중심으로」,『국어교육』71, 한국국어교육연구회.
서상규(2009),『교육용 기본 어휘 선정을 위한 기초 연구』, 국립국어원.
서상규(2013),『한국어 기본어휘 연구』, 한국문화사.
서상규(2014),『한국어 기본어휘 빈도 사전』, 한국문화사.
서상규·남윤진·진기호(1998),『한국어교육을 위한 기초 어휘 선정 (1) 기초 어휘 빈도 조사 결과』, 문화관광부·한국어세계화추진위원회.
서상규·백봉자·강현화·김홍범·남길임·유현경·정희정·한송화(2004),『외국인을 위한 한국어 학습사전(보고서)』, 문화관광부.
서상규 외(2000),『"한국어 교육 기초 어휘 의미 빈도 사전의 개발" 사업 보고서』, 문화관광부 한국어 세계화 추진 위원회.
서정미(2008),「말뭉치를 활용한 고등학교 국어사전의 편찬을 위한 기초 연구」, 경기대 박사학위논문.
서지영(2017),『교과용도서의 교과별 어휘 표준 구축 방안Ⅱ』, 한국교육과정평가원.
서종학(2000),『교과서 어휘의 조사단위 연구』, 국립국어원.
서종학·김주필(1999),『교과서의 어휘 분석 연구: 초등학교 국어 교과서를 중심으로』, 국립국어연구원.
석영(2014),「중국어 어휘론에서 기본어휘의 문제 –중국식 기본어휘 개념의 어휘론적 유용성에 대한 비판적 검토」,『中國言語硏究』52, 한국중국언어학회.
성광수(1999),「어휘부의 구조와 기초 어휘의 활용」,『선청어문』27(1), 서울대학교 국어교육과.
송철의 외(2008),『한국 근대 초기의 어휘』, 서울대학교출판부.
신동광(2011),「기본 어휘의 선정 기준: 영어 어휘를 중심으로」,『국어교육학연구』40, 국어교육학회.
신명선(2004),「어휘 교육의 목표로서의 어휘 능력(lexical competence)에 대한 연구」,『국어교육』113, 한국어교육학회.
신명선(2008),『의미, 텍스트, 교육』, 한국문화사.
신명선(2011),「국어과 어휘 교육 내용의 유형화에 관한 연구」,『국어교육학연구』

40, 국어교육학회.
신자영(2011),「DELE 코퍼스 구축 및 등급별 스페인어 기본 어휘 선정」,『이베로아메리카연구』22(2), 서울대학교 라틴아메리카연구소.
심재기 외(2016),『국어 어휘론 개설』, 박이정.
심재기(1990),「국어 어휘의 특성에 대하여」,『국어생활』22, 국어연구소.
안동환 역(2010),『코퍼스언어학 개론』, 한국문화사.
안의정(2012),「한국어 빈도 사전 편찬을 위한 기초 연구」,『한국사전학』20, 한국사전학회.
양명희(2010),「고급 한국어 어휘 교재 개발을 위한 기초 연구」,『반교어문연구』29, 반교어문학회.
양오진(2005),「중국어 기초 어휘·상용어휘와 단계별 어휘 교육에 대하여」,『중국언어연구』20, 한국중국언어학회.
양정실(2015),『초등학교 교과서의 어휘 실태 분석 연구』, 한국교육과정평가원.
연세대학교 언어정보개발연구원(2007),『연세한국어사전』, 두산동아.
유현경 외(2010),『전문 용어 자료 구축 및 정비를 위한 연구』, 국립국어원.
윤경선·이유미(2014),「유아 한글 교육용 어휘 목록 선정을 위한 연구」,『어문논집』59, 중앙어문학회.
윤지훈(2016),『교과용도서의 교과별 어휘 표준 구축 방안Ⅰ』, 한국교육과정평가원.
윤혜경(2016),「한국어교육용 구어 어휘 선정 연구」,『인문과학연구』50, 강원대학교 인문과학연구소.
이경수(2011),「초등학교 국어 어휘교육에 대한 소고 –우리와 프랑스의 학업성취도 평가 유형을 중심으로–」,『국어교육학연구』40, 국어교육학회.
이관규(2016),『교과서 어휘의 우리말 순화 연구』, 교육부.
이문복·신동광(2015),「2015 영어과 교육과정 기본 어휘 목록 개발」,『영어교과교육』14(4), 한국영어교과교육학회.
이삼형·김시정(2016),「구어 말뭉치의 어휘 분석을 통한 인지적 사고 발달 양상 연구」,『한국언어문화』59, 한국언어문화학회.
이삼형 외(2017a),「국어 기본 어휘 선정을 위한 기초 연구 –현황과 과제를 중심으로–」,『국어교육』156, 한국어교육학회.
이삼형 외(2017b),『국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구』, 국립국어원.
이상도(1995),「중국어 필수 어휘 선정에 관한 몇 가지 의견」,『中國言語硏究』3, 한국중국언어학회.
이상도·오영식·오문의·박정구(2002),「중국어 학습용 어휘 선정」,『중국학』19, 대

한중국학회.
이유경·최호철(2015),「학문 목적 한국어 어휘학습 교재 개발을 위한 기초 연구」,『어문논집』74, 민족어문학회.
이익환 외(2002),『기본 어휘 선정 및 사용 실태 조사를 위한 기초 연구』. 국립국어연구원.
이종은(2005),「한국어교육을 위한 의존용언 표현의 어휘항목 선정」,『이중언어학』28, 이중언어학회.
이종철(2011),「작문 교육과 어휘 교육」,『국어교육학연구』40, 국어교육학회.
이준호(2008),「한국어 어휘 교육 연구사: 학위 논문을 중심으로」,『문법 교육』9, 한국문법교육학회.
이지욱(2009),「외국인을 위한 한국어 파생어 교육」,『이화어문논집』27, 이화여자대학교 한국어문학연구소.
이진아·편도원·곽승철(2011),「발달장애아동의 기초 학습어휘 선정에 관한 연구: 유치원 및 초등학교 아동을 중심으로」,『특수교육학연구 』46(2), 한국특수교육학회.
이진영(2008),『'국어 초등학습용어사전' 편찬을 위한 초등 국어교과 기본 어휘 연구』, 경인교대 교육대학원 석사학위논문.
이충우(1990),「어휘교육의 기본과제」,『국어교육』71, 한국국어교육연구회.
이충우(1991),「초등학교 1, 2학년 국어과 교과서 어휘 조사 연구」,『관동어문학』7, 관동어문학회.
이충우(1992),『國語 敎育用 語彙 硏究 : 國民學校·中學校 國語科 敎育用 語彙 選定을 중심으로』, 서울대 박사학위논문.
이충우(1994a),「한국어 어휘 교육을 위한 대표 어휘 선정」,『국어교육』85, 한국국어교육연구회.
이충우(1994b),『한국어교육용 어휘 연구』, 국학자료원.
이충우(1998),「국어 어휘 교육론 개발을 위한 기초 연구 (1) -어휘 교육의 이론과 실제-」,『국어교육』98, 한국국어교육연구회.
이충우(1999),「국어 어휘 교육론 개발을 위한 기초 연구 (2) -『어휘교육론』의 내용-」,『국어교육학연구』9, 국어교육학회.
이해윤(2006),「'외국어로서의 독일어' 기본 어휘 선정에 대하여」,『외국어로서의 독일어』19, 한국독일어교육학회.
이현정(2014),「한국어교육용 외래어 선정을 위한 기초 연구 –중복도, 빈도의 객관적 지표와 전문가 평정을 바탕으로」,『시학과 언어학』27, 시학과 언어학회.
이현정·최영롱(2013),「한국어교육용 연결어미 선정을 위한 기초 연구: 구어·문어

빈도 및 교재 중복도 등의 객관적 지표를 중심으로」,『언어와 문화』 9(3), 한국언어문화교육학회.
이현주·조동성(2011),「학술 전문용어 정비 및 표준화의 특징 및 과제」,『한국어의미학』35, 한국어의미학회.
이희자(2003),「국어의 기초 어휘 및 기본 어휘 연구사」,『새국어생활』13(3). 국립국어원.
임지룡(1991),「국어의 기초 어휘에 대한 연구」,『국어교육연구』23-1, 국어교육학회.
임지룡(2002),「현대 국어 어휘의 사용 실태와 조어론적 특성」,『배달말』30, 배달말학회.
임지룡(2010),「국어 어휘교육의 과제와 방향」,『한국어 의미학』33, 한국어의미학회.
임지아(2005),「한국어 교재에 나타난 교육용 어휘 분석: 유의어를 중심으로」,『국어국문학』24, 동아대학교 국어국문학과.
임칠성(2002),「초급 한국어 교육용 어휘 선정 연구」,『국어교육학연구』14, 국어교육학회.
임칠성(2003),「기본 어휘 선정 방법론」,『새국어생활』13(3), 국립국어원.
임학준(2016a),「現代商務漢語 중급교재 어휘 특징 연구」,『동아인문학』37, 동아인문학회.
임학준(2016b),「현대상무한어(現代商務漢語) 초급교재 어휘 특징 연구 -≪신사로 초급속성상무한어(新絲路初級速成商務漢語)≫(Ⅰ·Ⅱ)를 중심으로-」,『중국어문학』73, 영남중국어문학회.
임홍빈·한재영(1993),『국어 어휘의 분류 목록에 대한 연구』, 국립국어연구원.
장경희·조성문·김명희·김순자·김정선·이필영·임유종·안미리·김응모·김태경(2005),『한국인의 의사소통 능력 발달 단계에 관한 연구』, 한양대학교.
장경희·이삼형·이필영·김명희·김태경·김정선·전은진(2012),『초·중·고등학생의 구어 어휘 조사』, 지식과교양.
장현진·전희숙·신명선·김효정(2014),「초등학생 교육용 기초 어휘 선정 연구: 저학년 중심으로」,『언어치료연구』23(1), 한국언어치료학회.
조남호(2002),「국어 어휘의 분야별 분포 양상」,『冠嶽語文硏究』27, 서울대학교 국어국문학과.
조남호(2003),「말뭉치를 활용한 어휘 빈도 조사」,『텍스트언어학』15, 한국텍스트언어학회.
조남호(2003),『한국어 학습용 어휘 선정 결과 보고서』, 국립국어원.
조성문(1997),「한국어 초급 교재의 기초 어휘 선정에 관하여」,『한국언어문화』

15, 한국언어문화학회.
조창규(2002),「교육용 어휘의 단위」,『국어교육학연구』14, 국어교육학회.
조현용(1999),「한국어교육용 기본 어휘 선정에 관한 연구」,『고황논집』25, 경희대학교 대학원
조현용(2000),「어휘 중심 한국어교육 방법 연구」, 경희대 박사학위논문.
조현용(2000),『한국어 어휘교육 연구』, 박이정.
조형일(2013),「교육용 외래어·외국어 표현 선정과 표기 방안 연구」,『한국언어문화학』10(1), 국제한국언어문화학회.
주형미(2011),「국가 영어과 교육과정 기본 어휘 목록 개선 연구」,『영어학연구』17(1), 한국영어학학회.
채영숙·채영희(2002),「기초 어휘 선정을 위한 초등학교 국어 교과서에 등장하는 어휘 분석 방안」,『한국정보과학회 언어공학연구회 학술발표 논문집』2002(10), 한국정보과학회 언어공학연구회.
최기선(2004),『(21세기 세종계획)전문용어의 정비』, 문화관광부.
최기선·송영빈·신효식(2000),『전문용어연구』, 전문용어언어공학연구센터.
최성규(1999),「장애아동의 어휘지도를 위한 일반아동의 기초 어휘 난이도 분석」,『특수교육연구』6, 국립특수교육원.
최형용(2013),『한국어 형태론의 유형론』, 박이정.
최형용(2016),『한국어 형태론』, 역락.
한송화(2015),『한국어 교육 어휘 내용 개발(4단계)』, 국립국어원.
한영균(2006),「한국어 어휘 교육ㆍ학습 자료 개발을 위한 계량적 분석의 한 방향: 어휘 빈도 조사 방법의 개선을 위하여」,『어문학』94, 한국어문학회.
한정한(2012),「한국어교육에서의 어휘와 문법-조사, 어미의 기본 어휘 선정 과정을 중심으로-」,『한국어학』57, 한국어학회.
황용주(2016),「21세기 세종 말뭉치 제대로 살펴보기」,『새국어생활』26(2), 국립국어원.
황유모·김정훈(2015),「개방형 한국어 지식 대사전 전문용어 신분류 체계 설정 및 재분류」,『전기학회논문지』64(2), 대한전기학회.

金鉉哲(2008),「關於韓國的漢語教學現狀與發展」,『중국어문학논집』53, 중국어문학연구회.
梁伍鎭(2003),「中國의 自國語 교육」,『語文硏究』31(4), 한국어문교육연구회.
梁伍鎭(2005),「중국어 기초어휘·상용어휘와 단계별 어휘 교육에 대하여」.『中國言語硏究』20, 한국중국언어학회.
林丕莪·裴宰奭(2010), 「中國汉语基本词彙與韩国高中中国语课程的基本词彙劃分與分

析」,『중국어문학논집』63, 중국어문학회.
蘇培成(1993),「關於基本詞彙的一些思考」, 本书编辑组(1995),『词汇学新研究－首届全国现代汉语词汇学术讨论会选集』, 語文出版社.
楊同用(2003),「基本詞彙問題的重新思考」,『語文研究』3, 山西省社科院.
張能甫(1999),「漢語基本詞彙研究的回顧與展望」,『四川師範大學學報』 2, 四川師範大學.
曹煒(2004),『現代漢語詞彙研究』, 北京大學出版社.
周薦(1987),「基本詞彙與一般詞彙劃分芻議」,『南開學報』3, 南開大學學報編輯部.

Bandelier A. & Cortier c. (2006) "Vocabulaires fondamentaux et Francais fondamental : applications à l'apprentissage de l'orthographe", Documents pour l'histoire du français langue étrangère ou seconde, 36.

Catach N., Jejcic F. & HESO groupe (1984) Les Listes orthographiques de base du français. Les mots les plus fréquents et leurs formes fléchies les plus fréquentes, Paris, Nathan.

Cortier, C. & Parpette, C. (2006) *De quelques enjeux et usages historiques du Français fondamental, Documents pour l'histoire du français langue étrangère ou seconde*, n. 36. https://journals.openedition.org/dhfles/1178

Coxhead, A. (2000) A new Academic Word List, *TESOL Quarterly,* 34, 2, pp. 213-238.

Dumais, C., Stanké, B., Moreau, A.C. & Beaudoin, M. (2014) "L'enseignement de l'orthographe lexicale: rélfexion sur les bases de données", Les Cahiers de l'AQPF, 4(3).

Gougenheim, G. (1955) Le français élémentaire, *International Review of Education,* Springer.

Gougenheim, G., Michea, M., Rivenc, P., Sauvageot, A. (1956/1964), *L'elaboration du Francaisfondamental (1er degre). Etude sur l'elaboration d'un vocabulaire et d'une grammaire de base*, Paris : Didier.

Ministère de l'Education Nationale, ENS de Saint Cloud, (1972), *Le Français fondamental 1er Degré, 2eme degré,* Institut National de Recherche et de Documentation Pédagogique.

Nation, I. S. P.(1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P.(2001). *Learning vocabulary in another language.* Cambridge:

Cambridge University Press.

Nation, I. S. P.(2004). “A study of the most frequent word families in the British National Corpus”. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing*. Amsterdam: John Benjamins.

Nation, I. S. P.(2006) “How large a vocabulary is needed for reading and listening?” *Canadian Modern Language Review,* 63(1).

Nation, I. S. P.(2001). *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, I. S. P., & Webb, S.(2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Nation, I. S. P.(2013) *Learning Vocabulary in Another Language,* Cambridge University Press.

Noyau, C. (2008) Place des verbes dans le Français Fondamental, acquisition du lexique verbal en français langue seconde, et didactique du lexique, *in* R. Bouchard & C. Cortier, eds. *Pratiques et représentations de l'oral en FLES, 50 ans après le français fondamental. Le Français dans le monde –Recherches et applications,* n. spécial, pp. 87-101.

Peereman, R., Lété, B. & Matos, R. (200) "MANULEX-Infra: Grade-level statistics upon grapheme-phoneme associations from child-directed written material", Behavior Research Methods,

Pothier, B. & Pothier, P. (2003) EOLE: Echelle d'acquisition en orthographe lexicale (du CP au CM2), Paris, Retz.

Ters F., Mayer G., & Reichenbach D. (1969) L'échelle Dubois-Buyse d'orthographe uselle française, Neuchâtel, Messeliller.

연구 책임자: 이삼형 (한양대학교 국어교육과 교수)
공동 연구원: 박진호 (서울대학교 국어국문학과 교수)
최형용 (이화여자대학교 국어국문학과 교수)
김정선 (한양대학교 국어교육과 교수)
이승연 (서울시립대학교 교양교육부 객원교수)
이현주 (인천대학교 불어불문학과 교수)
신명선 (인하대학교 국어교육과 교수)
이기연 (국립국어원 학예연구사)
김시정 (수원대학교 교양학부 객원교수)
연구 보조원: 허인영 (고려대학교 국어국문학과 박사수료)
김혜지 (이화여자대학교 국어국문학과 박사수료)
김수지 (한양대학교 국어교육과 박사수료)
이윤희 (한양대학교 국어교육과 박사과정)
보 조 원: 양세문 (한양대학교 국어교육과 석사과정)
담당 연구원: 이기연 (국립국어원 학예연구사)

발 행 인 소강춘

발 행 처 국립국어원
서울시 강서구 금낭화로 154(방화 3동 827)
전화: 02-2669-9775 전송: 02-2669-9747

인 쇄 일 2018년 12월 22일

발 행 일 2018년 12월 22일

*** 이 책은 국립국어원의 용역비로 수행한 '2018년 국어 기초 어휘 선정 및 어휘 등급화 연구' 사업의 결과물을 발간한 것입니다.**