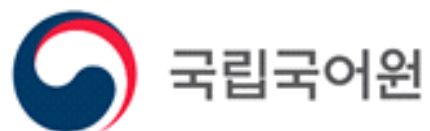


국립국어원 2017-01-50

발간등록번호
11-1371028-000695-01

# 국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구

연구책임자: 이삼형





# 제 출 문

## 국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 “국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구”에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2017년 5월 ~ 2017년 12월

2017년 12월 20일

연구 책임자: 이삼형(한양대학교)

연구 기관    한양대학교 산학협력단

연구 책임자    이삼형

공동 연구원    박진호, 최형용, 김정선, 신명선  
                    신동광, 강남옥, 이기연, 김시정

연구 보조원    김수지

보    조    원    이윤희, 양세문



# 요약문

1. **과업명:** 국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구

## 2. 과업의 목적

본 연구는 기초 어휘와 등급화에 대한 제반 이론을 검토하고, 기 구축 말뭉치 및 어휘 목록 사례를 검토하여 이론적 기반을 마련한다. 그리고 말뭉치에 기반한 기초 어휘 선정 및 등급화를 위한 샘플 말뭉치를 개발, 실증함으로써 국어 기초 어휘 선정 작업의 구체적 방안과 중장기 계획을 수립하는 데 그 목적이 있다.

## 3. 과업의 배경

국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구는 다음의 국가·사회적 요구를 배경으로 하고 있다.

- 첫째, 국민의 국어 능력 발전을 위한 어휘 정비 작업의 필요성
- 둘째, 말뭉치에 기반한 체계적 기초 어휘 선정과 활용 체계 개발
- 셋째, 기초 어휘 평정 작업에 대한 중장기 계획 수립의 필요성

## 4. 연구 내용 및 방법

본 연구의 내용은 다음과 같이 요약된다.

### 1) 기초 연구

- 기초 어휘, 어휘 평정 및 등급화 관련 선행 연구 검토
- 국내외 말뭉치 및 어휘 목록의 사례 연구

### 2) 말뭉치 구축

- 장르 및 시간/연대를 중심으로 한 기준 마련
- 말뭉치 보완 계획 수립

### 3) 기초 어휘 선정 및 어휘 등급화

- 말뭉치 기반 기초 어휘 목록의 추출 및 검증 절차 마련
- 어휘 등급화 및 검증

### 4) 기초 어휘 사업의 중장기 계획 수립

- 중장기 계획 목표 수립
- 중장기 단계 설정 및 실행 전략 수립

본 연구의 방법은 다음과 같이 요약된다.

### 1) 문헌 연구

국민의 어휘 능력과 어휘 평정에 관한 문헌 고찰을 통해 기초 어휘 선정과 어휘 등급화 기준 설정의 토대를 마련한다.

## 2) 사례 연구

국내외 말뭉치 및 어휘 목록의 사례를 수집하여 말뭉치 설계 방법과 어휘 목록 추출 방법 수립에 활용한다.

## 3) 조사 연구

형태소 분석기에서 나타나는 추출 어휘의 오류 패턴을 조사하여 연구의 정밀성을 높인다.

## 4) 전문가 자문

연구 결과물로서의 샘플 말뭉치 및 어휘 선정 기준에 대한 실험 과정, 어휘 등급의 기준 및 방향 등을 각계 전문가와 검토함으로써 연구의 타당도를 제고한다.

# 5. 연구 결과

본 연구의 결과를 정리하면 다음과 같다.

첫째, 기초 어휘의 개념을 정립하고 어휘 평정 및 등급화의 이론적 토대를 마련하였다. 기초 어휘, 어휘 평정, 어휘 등급과 관련된 선행 연구와 국내외 말뭉치 및 어휘 목록 사례를 검토함으로써 추후 수행될 기초 어휘 선정 및 어휘 등급화 연구의 실효적 방법론을 수립하기 위한 제반 이론을 확충하였다.

둘째, 기초 어휘 선정 및 어휘 등급화를 위한 샘플 말뭉치를 구축하고 검증 및 어휘 목록 추출 절차를 실증함으로써 말뭉치 보완 계획을 수립하였다. 등급화 대상 어휘 목록을 추출하기 위해서는 대규모 말뭉치가 필요하므로 인터넷을 통해 대량의 실제 언어 자료를 수집하고 형태소 분석기를 사용하여 가공하는 등 말뭉치의 구축 및 보완 방법론을 제시하였다.

셋째, 기초 어휘 선정 및 어휘 등급화의 절차를 수립하였다. 먼저 1등급 기초 어휘 선정 절차는 대규모 말뭉치로부터 정선된 균형 말뭉치로부터 양적 방법과 질적 방법을 상호 보완적으로 사용하여 어휘 목록을 추출하고 이를 검증하는 순으로 이루어진다. 그리고 2등급 이상 기초 어휘의 등급화는 대규모 말뭉치로부터 추출된 어휘 목록을 ‘교육용 도서와의 비교, 어휘 능력 검사, 전문가 자문’ 등의 방법을 통해 검증, 보완함으로써 최종 어휘 목록을 산출한다.

넷째, 기초 어휘 선정 및 어휘 등급화 사업의 중장기 계획을 수립하였다. 본 연구의 연구 결과에 따라 대규모 말뭉치와 어휘 목록을 회귀적으로 검토한다. 구체적으로 ‘어휘 목록을 제시하는 토대 확보 단계, 추출된 어휘 목록과 등급화 방법을 검증하고 정교화하는 발전 및 확장 단계, 그리고 만들어진 어휘 목록을 지속적으로 관리·보완하고 전국민이 활용할 수 있는 체계를 마련하는 지속 발전 가능 단계’의 순서로 사업을 심화, 확장할 수 있도록 중·장기 발전 계획을 제시하였다.

## 6. 결과물: 최종 보고서 50부, 시디(CD) 10매

# Abstract

**1. Task Title:** Basic Research for Lexical Grading and Selection of Basic Vocabulary in Korean

**2. Purpose of Task:** This study desires to build the theoretical foundation and establish the relevance and working plan for basic vocabulary selection by demonstrating sample corpus for selection and grading of basic vocabulary based on corpus. Also, this research not only develops all sorts of theories about basic vocabulary and grading, but also examines corpus and vocabulary list cases that have been established in the past.

**3. Background of Task:** Necessities of the basic research for lexical grading and selection of basic vocabulary in Korean are as follows:

First, it requires maintenance work on the vocabularies for the whole people.

Second, it is necessary to select basic vocabularies based on corpus.

Third, it needs a medium and long term plan for rating basic vocabularies.

## **4. Study Contents and Methods**

Study contents of 'Basic Research for Lexical Grading and Selection of Basic Vocabulary in Korean' are as follows:

### 1) Fundamental research

- Review of preceding studies on basic vocabulary, lexical rating and grading
- Case study for corpora and vocabulary lists at domestic and foreign language.

### 2) Construction of a corpus

- Consideration of genre and time / era
- Supplement plan for corpus

### 3) Lexical grading and selection of basic vocabularies

- Extraction and verification procedures of a corpus-based basic vocabulary list
- Lexical grading and verification procedures

### 4) Establishment for long term plan of this basic vocabulary project

- Goal setting for long term plan
- Construction of phase settings and execution strategy during long term

Study methods of 'Basic Research for Lexical Grading and Selection of Basic Vocabulary in Korean' are as follows:

1) Reference Study

It provides the foundation of the basic vocabulary selection and the setting criteria for vocabulary grading through literature reviews on the vocabulary ability and lexical rating of the people.

2) Case Study

It examines the methods for the corpus design and the extraction of lexical list by collecting examples of corpus and vocabulary lists at domestic and foreign language.

3) Survey Study

It investigates the error patterns of corpus-based lexical extraction using morpheme analyzer and reinforces the accuracy of this study.

4) Experts' Advice

It enhances the validity of this study by examining the sample corpus as a research result, the experiment process on the vocabulary selection criteria, and the criteria and direction of the vocabulary class with experts from various fields.

**5. Research Finding:** The results of this study are as follows:

First, this study established the concept of basic vocabulary and provided the theoretical foundation of lexical rating and grading. Moreover, it expanded all sorts of theories to establish the effective methodology of lexical grading research and the basic vocabulary selection to be carried out, by examining not only the preceding studies related to the basic vocabulary, the vocabulary rating, and the vocabulary class, but also the example of the vocabulary lists and the corpora of the domestic and foreign language

Second, this study constructed a sample corpus for basic vocabulary selection and lexical grading, and established supplementary planning about a corpus by demonstrating and verifying extraction procedures for vocabulary list. In this study, a large corpus is required to extract the list of appropriate vocabulary for grading. Therefore, it collected a large number of actual language data through the Internet and produced the data using a morphological analyzer. Through this process, this study suggested a methodology for constructing and supplementing corpus.

Third, this study established the procedures for basic vocabulary selection and lexical grading. First of all, it took the selection procedure about the first grade basic vocabulary, according to the order of extracting large corpus from the vocabulary list and verifying it, by making complementary uses of quantitative and qualitative method. In order to grade the basic vocabulary more than grade 2, the vocabulary list extracted from the large corpus is repeatedly verified through



comparison with educational books, vocabulary ability test, expert advice, etc., and the final list is produced.

Fourth, this study has established a long term plan for the selection of basic vocabulary and lexical grading project. There are three phases in this plan. Based on the results of this study, the first step is to establish a basis for presenting the vocabulary list by recursively reviewing the large corpus and the vocabulary list, and the second step is development and expansion stage in which verify and refine the extracted vocabulary list and grading methods. Lastly, this plan suggested the road-map to deepen and expand the project in Worder of sustainable development that establishes a system that can be utilized by the whole people and continuously manage the list of vocabulary created.

**6. Output:** 50 final reports, 10 CDs including final reports

# 목 차

I. 서론 .....	1
1. 연구의 목적과 필요성 .....	1
1.1. 연구의 목적 .....	1
1.2. 연구의 필요성 .....	2
2. 연구 내용과 방법 .....	4
2.1. 연구 내용 .....	4
2.2. 연구 방법 .....	7
3. 연구 추진 과정 .....	9
3.1. 연구 추진 일정 .....	9
3.2. 주요 협의회 내용 .....	10
II. 기초 연구 .....	13
1. 이론 연구 .....	13
1.1. 기초 어휘의 개념 .....	13
1.2. 어휘 평정 및 등급화의 개념 .....	16
1.3. 말뭉치 기반 기초 어휘 선정 및 등급화 방법론 .....	17
1.4. 기초 어휘 및 어휘 등급과 관련된 문법적 쟁점 .....	24
2. 사례 연구 .....	29
2.1. 국내 말뭉치 및 어휘 목록 .....	29
2.2. 해외 말뭉치 및 어휘 목록 .....	48
III. 말뭉치 구축 .....	65
1. 장르에 대한 고려 .....	65
2. 시간/연대에 대한 고려 .....	67
3. 말뭉치 구축 현황 .....	70
4. 말뭉치 보완 계획 .....	74
IV. 기초 어휘 선정 및 어휘 등급화 .....	77
1. 말뭉치 기반 어휘 목록의 추출 .....	77
1.1. 형태소 분석 .....	78

1.2. 장르별 단어 빈도, 범위, 산포도 추출 .....	82
1.3. 세 변수 사이의 관계에 대한 통계적 고찰 .....	84
1.4. 변수들에 가중치를 부여하여 단어 순위 결정 .....	88
1.5. 다른 어휘 목록과의 비교 .....	99
1.6. 텍스트 포괄 범위 조사 .....	101
1.7. 향후 과제 .....	102
2. 어휘 등급화 절차 .....	104
2.1. 1등급 어휘 선정 절차 .....	104
2.2. 2등급 이상 어휘 선정 절차 .....	110
 V. 기초 어휘 사업의 중장기 계획 수립 .....	114
1. 중장기 계획 수립의 경위 .....	114
2. 중장기 계획 수립 .....	117
 VI. 종합 및 제언 .....	122
1. 종합 .....	122
2. 정책 제언 .....	125
 참고 문헌 .....	127
 <부록 1> 세종 태그셋 .....	139
<부록 2> UTagger 오류 패턴 .....	140

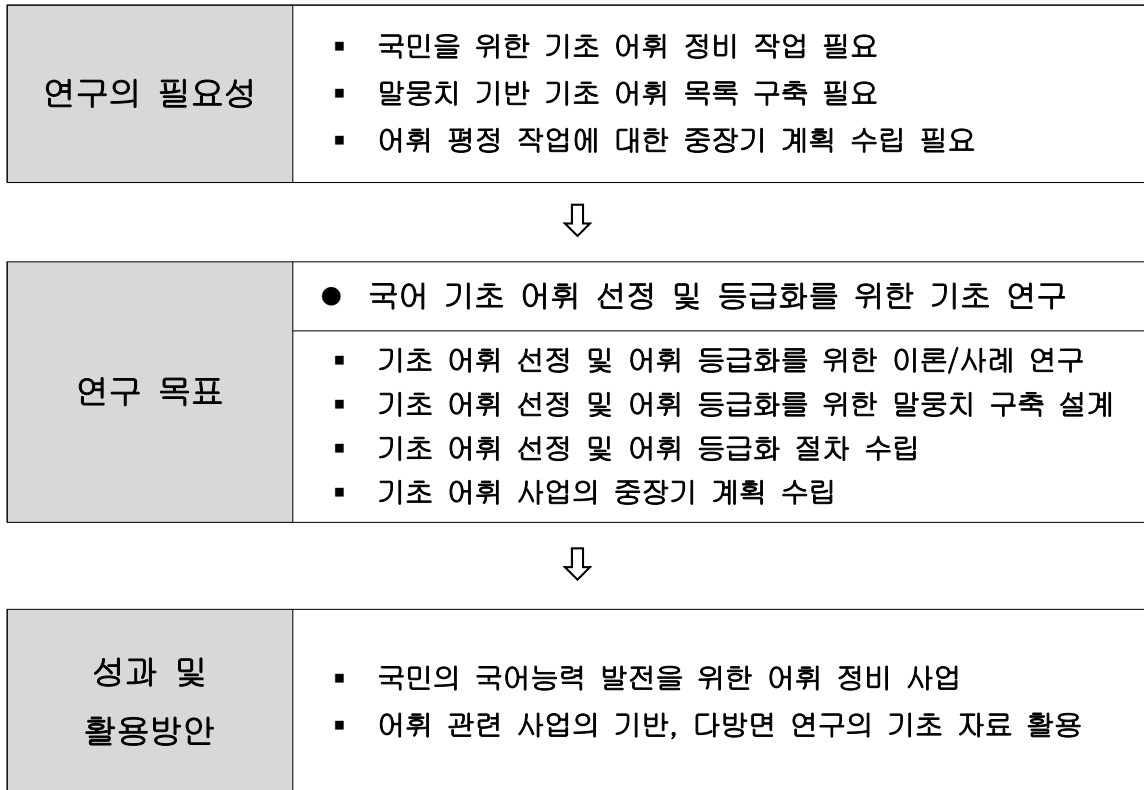
## 〈표 차례〉

<표 1> 국어교육용 어휘 목록 선정의 주요 성과 .....	14
<표 2> 어휘력을 구성하는 어휘의 양(Nation, 1990: 19) .....	19
<표 3> 어휘력의 관점에서 본 어휘의 구조(김광해, 2003) .....	19
<표 4> 문자 및 음성 언어에서의 텍스트 포괄 범위 비교(Nation, 2006: 79) ...	20
<표 5> 산포도 공식(신동광, 2011 참조) .....	21
<표 6> 어휘 친숙도 측정 척도(신동광, 2011) .....	21
<표 7> Waring(2000)의 어휘 지식 측정 척도 .....	22
<표 8> 어휘 목록 제작 시 어휘 선정 기준 및 적용 순서(신동광, 2011: 225) ...	22
<표 9> 등급별 어휘량의 변화 .....	23
<표 10> 국어 말뭉치의 목록 .....	29
<표 11> 국어 말뭉치의 구성과 특징 .....	30
<표 12> 국내 어휘 선정 연구 성과(연도순) .....	33
<표 13> 등급별 어휘의 상황(김광해, 2003: 27 수정 인용) .....	46
<표 14> 한국어 학습용 어휘 목록 품사별 분포(조남호, 2003: 11) .....	47
<표 15> Brown Corpus의 주제 영역 및 구성 단어 수 .....	49
<표 16> BNC의 대(大)주제 영역 및 구성 단어 수 .....	50
<표 17> BNC 문어 자료의 세부 주제 영역 및 구성 단어 수 .....	50
<표 18> COCA 주제 영역별 단어 수 .....	51
<표 19> BNC-COCA 상위 2,000 단어 선정을 위한 말뭉치 구성 .....	52
<표 20> 2015 영어과 교육과정 기본 어휘 목록 개발을 위한 말뭉치 구성 .....	53
<표 21> 말뭉치 분석 기반 영어 어휘 목록 .....	57
<표 22> 문자 및 음성 언어에서의 텍스트 포괄 범위 비교(Nation, 2006: 79) ...	61
<표 23> 2015 영어과 교육과정 과목별 어휘 수 .....	63
<표 24> 신문 말뭉치 연도별 빈도 통계: 지정사 ‘이다’의 경우 .....	67
<표 25> 말뭉치 전체 통계 .....	71
<표 26> 세종 말뭉치 장르별 통계 .....	71
<표 27> 도서 말뭉치 장르별 통계 .....	72
<표 28> 잡지 말뭉치 장르별 통계 .....	72
<표 29> 블로그 말뭉치 장르별 통계 .....	73
<표 30> 보완 말뭉치의 구성 .....	76
<표 31> 검토 대상 자료 .....	79
<표 32> 1명의 직관에 따른 등위와 가중치 부여 점수의 비교 .....	93
<표 33> 100 단어 목록 1에 대한 실험 결과 .....	95
<표 34> 100 단어 목록 2에 대한 실험 결과 .....	95
<표 35> 100 단어 목록 3에 대한 실험 결과 .....	96
<표 36> 점수(결정된 weighted sum), 빈도, 범위, 산포도 사이의 관계 .....	97
<표 37> <연세한국어사전> 표제어와의 말뭉치 등위 비교 .....	101
<표 38> 텍스트 포괄 범위 .....	102

## 〈그림 차례〉

[그림 1] 연구 개요 .....	1
[그림 2] 연구의 내용 .....	4
[그림 3] 어휘 목록의 추출 및 검증 절차 .....	6
[그림 4] 이해 어휘량의 발달(玉村文郎編, 1989: 159) .....	16
[그림 5] Dickins의 GSL 예시 .....	57
[그림 6] ERF의 어휘 등급 척도 .....	64
[그림 7] 순위-빈도 그래프 .....	65
[그림 8] 연도별 빈도 최상위 5개 단어의 절대빈도 추이 .....	68
[그림 9] 연도별 빈도 최상위 5개 단어의 상대빈도 추이 .....	68
[그림 10] 말뭉치 기반 어휘 목록의 추출 절차 .....	77
[그림 11] 100개 장르에서의 단어별 절대빈도 .....	82
[그림 12] 단어별 빈도, 범위, 산포도(빈도순 소팅 결과의 첫부분) .....	83
[그림 13] 범위(가로축)와 산포도(세로축)의 관계(청색 선: 회귀 곡선) .....	84
[그림 14] 빈도(가로축, 로그 변환)와 산포도(세로축)의 관계(청색 선: 회귀 곡선) .....	85
[그림 15] 범위(가로축)와 빈도(세로축, 로그 변형)의 관계(청색 선: 회귀 곡선) .....	87
[그림 16] random 100 단어 목록 1 .....	89
[그림 17] random 100 단어 목록 2 .....	90
[그림 18] random 100 단어 목록 3 .....	91
[그림 19] 100 단어 목록 1에 대한 정보 통합 결과 .....	92
[그림 20] 100 단어 목록 1에 대한 6인이 부여한 등위 통합 결과 .....	94
[그림 21] $\text{weighted sum}(=0.2 \times \text{빈도} + 0.7 \times \text{범위} + 0.1 \times \text{산포도})$ 에 따른 소팅 결과 .....	96
[그림 22] 범위와 점수(로그 변형)의 관계 .....	97
[그림 23] 산포도와 점수(로그 변형)의 관계 .....	98
[그림 24] 빈도(로그 변형)와 점수(로그 변형)의 관계 .....	98
[그림 25] <연세한국어사전> 표제어와 말뭉치 통계 결과의 비교 .....	100
[그림 26] 기초 어휘 선정용 말뭉치 구축 후 절차 .....	106
[그림 27] 선정된 기초 어휘의 검증 방향 .....	108
[그림 28] 어휘 등급화를 위한 결정 요소 .....	110
[그림 29] 어휘 등급화 절차 .....	111
[그림 30] 등급별 어휘의 타당성 검토 방안 .....	113
[그림 31] 기초 어휘 사업의 중장기 로드맵 .....	117
[그림 32] 단계별 발전 방향 .....	124

# I. 서론



[그림 1] 연구 개요

## 1. 연구의 목적 및 필요성

### 1.1. 연구의 목적

본 연구는 기초 어휘와 등급화에 대한 제반 이론을 검토하고, 기 구축 말뭉치 및 어휘 목록 사례를 검토하여 이론적 기반을 마련한다. 그리고 말뭉치에 기반한 기초 어휘 선정 및 등급화를 위한 샘플 말뭉치를 실증함으로써 기초 어휘 선정 작업의 적실성과 방안을 마련하는 데 연구의 목적이 있다.

## 1.2. 연구의 필요성

### 1) 국민 전체를 대상으로 한 어휘 정비 작업이 필요하다

어휘는 지식을 습득하고, 사고하며 의사소통하는 수단으로 언어를 배우는 유아기부터 성인에 이르기까지 국어 능력의 핵심 요소이다. 어휘 능력은 듣기, 읽기, 말하기 쓰기의 4대 언어 기능을 직접적으로 지원하므로 국민의 언어 생활에 토대가 되는 능력이다. 기초 어휘는 의사소통에서 가장 기본적이고 핵심적인 어휘로, 국어 정책 수립의 기준이 된다. 국민의 국어 능력 향상을 위해서는 근간이 되는 어휘 능력 향상이 가장 중요한데, 이를 측정할 수 있는 기준이 되는 것도 기초 어휘이다. 즉 기초 어휘 목록 구축은 그 자체만으로도 어휘 사용 실태를 보여준다는 측면에서 의의가 있지만, 나아가 교과서 및 교육 자료 개발 및 감수, 국가 고사 출제, 국민의 국어능력 평가를 비롯한 각종 기초 조사 근거로 활용될 수 있다는 점에서 특히 중요하다.

국내의 기초 어휘는 교육적 필요성에 의해 한국어교육에서 중점적으로 이루어져 왔다. 한국어 등급별 어휘, 한국어 학습 어휘와 같이 학습자의 요구에 부응하는 연구 성과들이 축적되었으나(조남호, 2003; 서상규, 2013), 정작 모국어 어휘에 대한 연구는 미진한 편이다. 이마저도 모국어 어휘에 대한 연구는 어휘 발달이 두드러진 유아와 초등 시기에 집중되어 있어(장현진 외, 2014; 김한샘, 2010), 국민 전체를 대상으로 한 어휘 사용 실태, 어휘 목록 구축 등의 연구는 거의 전무한 실정이다. 또한 기존에 수행된 연구의 성격도 기초 조사에 머물고 있어 구체적인 성과물로 나아가지 못하였다는 점에서 국민을 위한 기초 어휘 정비 작업이 필요한 시점이라고 할 수 있다.

### 2) 말뭉치에 기반한 기초 어휘 선정이 필요하다

말뭉치는 언어의 본질을 연구하기 위해 기계 가독 형태로 수집한 대량의 언어 자료로, 사전 제작 또는 문법 기술의 근거를 마련하는 데 주로 사용될 뿐만 아니라, 해당 언어의 실상을 보여준다는 점에서 어휘 정비 작업에서도 적극적으로 활용되고 있다. 어휘 목록을 추출하고 등급화하는 데에는 말뭉치 구축이 선행되어야 하는바, 말뭉치 구축은 인적, 물적 자원이 필요한 작업이므로, 장기간의 연구와 구축, 검증 등의 작업이 요구된다. 본 연구에서는 기초 어휘 선정과 등급화를 위해 기존 말뭉치와 최신 언어 자료 등을 적극 수집, 종합하여 말뭉치를 재정비한다. 이를 위해 지금까지 구축된 국내외 말뭉치 현황을 파악하여 기초 어휘 선정에 필요한 말뭉치 구축 방법 및 사례를 살핀 뒤, 언어 실태를 객관적으로 보여주는 말뭉치의 규모와 구성 비율을 설정하여, 말뭉치 기반의 기초 어휘 선정이 타당하게 이루어지도록 한다.

### 3) 기초 어휘 평정 작업에 대한 중장기 계획이 필요하다

기초 어휘는 그 중요성과 필요성에 대해서는 일찍부터 인식하였지만 이에 대한 장기적인 계획 없이 파편적으로 연구들이 진행되어 왔다(서상규, 2001). 그 결과 동일한 주제의 연구가 반복될 뿐, 구체적 자료 수집으로 이어지지 않는 등 실효성 있는 결과로 나아가지 못하였다.

국민의 언어 생활과 관련한 정책 수립의 지표가 되는 기초 어휘 연구는 대규모 말뭉치를 기반으로 하여 언어 실태 파악, 어휘 선정, 어휘 등급화 등의 일련의 과정을 거쳐야 하고 국어정보학, 어휘, 국어교육 분야의 전문가들의 협업이 필요하다. 이에 작업의 복잡성을 고려해, 각 과정마다의 철저한 사전 준비와 실행, 결과에 대한 검증이 필수적이다. 또한 최종 결과물이 지니는 가치를 생각할 때 장기적으로 사업을 발전, 지속해 갈 수 있는 중장기 계획이 필요하다.



## 2. 연구 내용과 방법

### 2.1. 연구 내용

본 연구는 다음과 같은 세부 과제로 구성되어 있다.

연구 내용	1) 기초 연구 2) 샘플 말뭉치 구축 3) 기초 어휘 선정 및 어휘 등급화 4) 기초 어휘 사업의 중장기 계획 수립
-------	--

[그림 2] 연구 내용

#### 1) 기초 연구

##### (1) 기초 어휘, 어휘 평정 및 등급화 관련 선행 연구 검토

기초 어휘의 개념과 성격, 어휘 평정, 어휘 등급화의 개념, 등급화 방법론, 언어학적 쟁점 등에 대한 이론적 검토를 실시한다. 검토의 주요 내용은 다음과 같다.

- 기초 어휘의 개념
  - 기초 어휘의 개념 범위, 성격
  - 기초 어휘와 기본 어휘의 관계
- 어휘 평정 및 등급화의 개념
  - 어휘 평정과 등급화의 개념 구분
  - 국어 어휘 평정의 예
- 등급화 방법론
  - 어휘 등급 설정을 위한 어휘 평정 기준 수립
  - 어휘 등급의 기준 수립을 위한 고려 요소

○ 기초 어휘 및 어휘 등급과 관련한 문법적 쟁점

- 분석 단위 설정
- 어휘군의 처리
- 어휘군, 상하 관계, 유의 관계와 어휘 등급

(2) 국내외 말뭉치 및 어휘 목록의 사례 연구

기존에 구축된 국내외 말뭉치와 어휘 목록을 수집하여 성격과 규모, 목적 등을 분석하여 본 연구가 구축하고자 하는 말뭉치와 어휘 목록의 개발 방향을 수립한다. 본 연구에서 분석한 주요 내용은 다음과 같다.

○ 국내 말뭉치 및 어휘 목록 수집 및 분석

- 국내 말뭉치 구축 현황: 구성과 특징, 규모
- 어휘 목록 선정 현황
- 등급화 사례

○ 해외 말뭉치 및 어휘 목록 수집 및 분석

- 해외 말뭉치 구축 현황: 구성과 특징, 규모
- 어휘 목록 선정 현황
- 어휘 학습량 및 어휘 등급화 단위 설정

2) 말뭉치 구축

(1) 장르 및 시간/연대를 고려한 말뭉치 구축

말뭉치 기반 어휘 목록 추출을 위해서는 말뭉치 구축에 있어서도 다양한 요인을 고려해 구축해야 한다. 본 연구에서는 우선적으로 ‘장르와 시간/연대’를 주요 고려 요소로 파악하여 샘플 말뭉치를 설계하고 실증한다.

(2) 말뭉치 구축 현황

등급화된 어휘 목록 추출을 위해서는 실제 언어 생활을 포착할 수 있는 대규모 말뭉치 구축이 필요하다. 본 연구에서는 세종말뭉치를 비롯하여 기존에 구축된 말뭉치를 활용해 샘플 말뭉치를 마련한다.

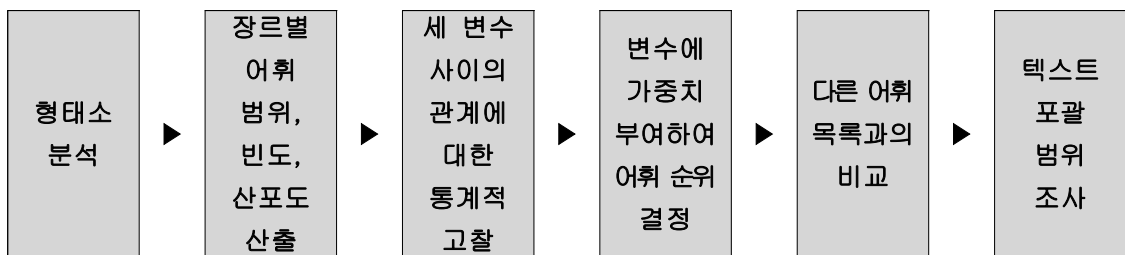
### (3) 말뭉치 보완 계획

현재 구축한 샘플 말뭉치에 나타난 문제를 파악하여 향후 사업에서 보완할 말뭉치를 계획한다.

## 3) 기초 어휘 선정 및 어휘 등급화

### (1) 말뭉치 기반 기초 어휘 목록의 추출 및 검증 절차

샘플 말뭉치를 기반으로 기초 어휘 목록을 추출하여 검증하는 절차는 다음과 같다. 이러한 절차는 기초 어휘 선정 및 어휘 등급화를 위한 구체적인 방법론을 적용해 본 것으로 대규모 말뭉치를 기반으로 하여 기초 어휘를 선정하는 내년도 사업의 예비 수행의 성격을 지닌다.



[그림 3] 어휘 목록의 추출 및 검증 절차

### (2) 어휘 등급화 및 검증 절차

말뭉치 기반 어휘 추출 후 기초 어휘를 선정하고 이를 등급화하여 한다. 이때 기초 어휘의 규모와 성격 등을 규정해야 하고, 이를 위해 양적 분석 단계와 질적 분석 단계를 거친다. 선정된 기초 어휘의 검증 방향에 따라 기초 어휘를 선정하고, 어휘 등급화를 위한 결정 요소에 따라 등급화와 이에 대한 검증 절차를 진행한다.

#### 4) 기초 어휘 사업의 중장기 계획 수립

##### (1) 중장기 계획 목표 수립

국민 전체를 대상으로 하는 기초 어휘의 선정 및 등급화를 목표로 하는 본 사업은 중장기적 계획 하에 수행되어야 하는 특수성을 지니고 있다. 이에 본 연구에서는 최종 목표 설정과 이에 달성하기 위한 구체적인 중장기 계획을 수립하고자 한다. 이는 국어기본법, 국어발전 기본 계획 등에 입각한 기존 정책의 연속선상에서 수립되는 계획이다.

##### (2) 중장기 단계 설정 및 실행 전략 수립

중장기 계획은 ‘확보 단계 → 발전 및 확장 단계 → 지속 발전 가능 단계’로 구분하여 세부 목표와 과제를 설정하고, 이를 추진하기 위한 실행 전략을 수립한다. 본 연구에서 제안한 실행 전략으로는 말뭉치 구축 과업에 걸맞은 재정 규모 확보, 기초 어휘 과업의 성격에 맞는 전문가 인력 풀 구축 등이 있다.

## 2.2. 연구 방법

본 연구는 이론적 토대 마련을 위한 문헌 연구와 사례 연구, 대규모의 말뭉치에서 어휘를 추출하기 위해 활용될 형태소 분석기에 대한 조사 연구, 연구 내용과 관련하여 전문가의 자문을 듣는 등의 방법으로 수행하였다.

### 1) 문헌 연구

국민의 어휘 능력과 어휘 평정에 관한 문헌 고찰을 통해 기초 어휘 선정과 어휘 등급화의 기준 설정의 토대를 마련한다. 문헌 연구의 주요 내용을 간추리면 다음과 같다.

- 언어학·국어학: 어휘의 유형 및 구조, 기초 어휘와 기본 어휘, 어휘 분석의 단위
- 언어교육(국어교육, 한국어교육, 영어교육): 기초 어휘 선정 방법, 어휘 선정의 기준, 어휘 등급화의 개념과 방법
- 국어정보학: 말뭉치 구축 방법론, 말뭉치 기반 어휘 추출 방법

## 2) 사례 연구

국내외 말뭉치 및 어휘 목록의 사례를 수집하여 말뭉치 설계 방법과 어휘 목록 추출 방법을 검토한다.

- 국어 기구축 말뭉치 현황 분석
- 국어교육과 한국어교육 분야에서의 어휘 목록
- 영어교육에서의 말뭉치 구축 현황과 기초 어휘 목록

## 3) 조사 연구

본 연구에서는 말뭉치를 기반으로 어휘를 추출하여 형태소 분석 작업을 거쳐 어휘의 빈도를 조사한다. 대규모 어휘 분석을 위해 형태소 분석기(U-tagger)를 사용하는데 이때 일정 비율의 오류가 출현한다. 본 연구에서는 형태 분석 결과의 오류를 수정하기 위해 오류 패턴을 조사하여 정리함으로써 기초 어휘 선정 작업에 활용하도록 하였다. 조사 개요는 다음과 같다.

- 조사 내용: 형태소 분석기를 활용한 말뭉치 기반 추출 어휘의 오류 패턴
- 조사 방법: 직접 조사(말뭉치 자료 수집→형태소 태그 부착→분석→패턴화)
- 조사 규모: 약 16,000어절(한 파일 당 2,000어절씩, 문어 5개 파일, 구어 3개 파일)

## 4) 전문가 자문

본 연구에서는 국어정보학, 어휘 등급화 등의 관련 전문가를 초빙하여 연구의 타당도를 제고하였다.

- 전문가: 김일환(고려대), 김한샘(연세대), 최운호(목포대), 송영빈(이화여대)
- 검토 내용: 최종보고서 세부 내용 검토
- 검토 기간: 2017년 11월 28일(화)~2017년 12월 7일(목) (10일)

### 3. 연구 추진 과정

#### 3.1. 연구 추진 일정

단계 수립	세부 연구 내용	월							
		5	6	7	8	9	10	11	12
계획 수립	전체 진행 계획 확정	○							
	분과별 업무 분담	○							
기초 연구	선행 연구 검토	○	○						
	기초 어휘의 개념과 유형 정립		○						
말뭉치 구축 방법론 수립	말뭉치 성격 및 범위 설정		○	○					
	말뭉치 구축 및 관리 방법론 수립		○	○					
	전문가 자문				○				
어휘 평정	분석 단위 설정 및 검토				○				
	기초 어휘 선정 기준 마련				○	○			
	전문가 자문						○		
샘플 말뭉치 구축	원시 말뭉치 구축					○	○		
	주석 말뭉치 구축					○	○		
	어휘 추출 및 평정 작업						○		
중장기 계획 설계	중장기 계획 목표 수립						○		
	세부 목표 및 전략 수립						○	○	
종합	전문가 의견 수렴							○	○
	결과 보고서 작성							○	○

### 3.2. 주요 협의회 내용

본 연구진은 매달 1회 전체 회의를 정기적으로 진행하였으며, 총 3회의 보고회를 개최하였다. 그리고 연구 결론에 대해서는 전문가 자문회의를 실시하였다. 내용을 일자별로 정리하여 제시하면 다음과 같다.

#### 1) 착수 보고회

- 날짜: 2017년 6월 1일(목)
- 장소: 국립국어원
- 참여자: 국립국어원 관계자, 연구진, 자문위원단

##### ○ 주요 내용 및 결과

- ‘기본 어휘’ 개념의 모호성에서 벗어나기 위해 국어능력의 기초가 되는 어휘라는 관점에서 ‘기초 어휘’로 연구 과제명을 변경. 이에 따라 본 연구는 ‘기초 어휘 선정 및 등급화’를 목적으로 함.
- 언어정보과에서 기 구축된 말뭉치 및 차년부터 구축 예정인 말뭉치를 이용 가능함. 말뭉치 구축을 지속적인 사업으로 진행 예정에 있으므로 계획 설계에 있어 고려 바람.
- 어휘 평정에서 전문어의 경우 우리말샘 DB를 활용 가능함.
- 말뭉치 구축 및 어휘 목록 작성에 있어서 지적 재산권 등 데이터에 대한 법적 문제를 검토하는 작업이 기초 연구에서 추가되어야 함.

#### 2) 중간 보고회

- 날짜: 2017년 9월 25일(월)
- 장소: 국립국어원
- 참여자: 국립국어원 관계자, 연구진, 자문위원단

##### ○ 주요 내용 및 결과

- 연구 용역 사업 진행 사항 보고, 이후 사업 진행 방향 및 일정 점검
- 연구 기간과 연구 예산을 고려하여 현실적으로 구축 가능한 말뭉치 규모를 제시

- 할 필요가 있음. 실제로 기초 어휘는 대개 사용 빈도가 매우 높으므로 규모에 관계없이 나타난다는 점에서 대규모 말뭉치의 활용을 재고할 필요가 있음.
- 어휘 등급화를 위해서는 저빈도어까지 포함되어야 하므로 대규모 말뭉치가 필요함. 그러나 현재 어휘 등급화에 대한 논의가 더욱 풍부하게 이루어지지 않고 있어 그 필요성이 드러나지 않음.
  - 연구의 궁극적인 목표가 말뭉치 구축이 아닌 기초 어휘 선정과 어휘 등급화에 있음을 염두에 두어야 함.
  - 말뭉치 구축 이후의 여러 가지 비용도 함께 고려하여 중장기계획을 세워야 함.

### 3) 전문가 자문 회의

#### ○ 주요 내용 및 결과

- 총평: 연구 보고서를 보면 기존 한국에서의 이 주제에 대해 이루어진 연구를 충분히 검토하고 이를 극복하는 방안을 모색하려고 하는 점이 충분히 전달됨. 영어를 중심으로 한 이론적인 토대를 충실히 반영하고 있다는 점도 확인할 수 있었음. 이 연구를 통해 한국어의 특성이 밝혀지길 바램. 또한 단발적인 연구가 아닌 지속적인 연구로 결실을 맺기를 바램.
- 수정 및 권고사항
  - (Ⅱ장) 어휘 등급 기준 수립 시 고려 요소들을 실제 작업에서 어떻게 적용할 것인지의 방법론이 추가되기를 바램.
  - (Ⅲ장) 말뭉치 구축 계획에 초중 교과서는 높은 가중치를 부여받아 말뭉치에 우선적으로 포함되었으면 함.
  - (Ⅳ장) 어휘 선정을 위한 통계적 고려에서는 주로 ‘빈도, 범위, 산포도’라는 3개의 통계적 변수를 기반으로 한 실험 내용을 구체적으로 서술하고 있음. 이러한 방식에 문제가 있다기보다는 어휘 선정에서 고려되는 좀 더 포괄적인 통계적 접근에 대한 소개가 이루어진 다음 실험에 대한 세부 내용으로 넘어가는 것이 전체를 이해하는 데에 도움이 될 듯함.
  - (Ⅴ장) ‘2+ 1+ 1’의 중·장기 계획이 사업에 목적에 맞게 충분히 잘 수립되었음. 하지만 이 연구 사업의 목표와 목적한 바를 수행하기에는 사업 기간이 짧아 보이기도 함. 보고서에서는 ‘중장기 계획’이라고 되어있지만 수행되는 연구의 양적 부담을 감안하면 중·단기 계획으로 판단됨.



#### 4) 최종 보고회

- 날짜: 2017년 12월 12일(화)
- 장소: 국립국어원
- 참여자: 국립국어원 관계자, 연구진, 자문위원단

##### ○ 주요 내용 및 결과

- 연구 용역 사업 결과 보고
- 올해 연구에서 수집한 샘플 말뭉치는 문어에 치중되어 있어 구어 자료의 보완이 필요함. 순수 구어 자료(세종 구어 말뭉치, 한양대 연령별 말뭉치 등)와 함께 구어적 성격이 상대적으로 많이 나타나는 자료(인터넷 게시판, 뉴스 인터뷰 기사, 영화 자막 등)를 수집하여 검토, 활용할 계획임.
- 어휘 등급의 수는 추후 연구를 진행하는 과정에서 확정될 것임. 어휘 등급의 수는 목적과 활용 방향을 고려하여 결정될 것이며, 세부적인 구분이 이루어질 것으로 예상함.
- 향후 4개년 사업 이후 국립국어원에서 자체적으로 어휘 목록을 갱신하고 보완할 수 있도록 보완 계획 혹은 절차가 함께 제시되기를 바램. 이러한 보완 계획은 실질적 방향제시와 함께 대외적인 차원에서 필요성과 가능성을 제시할 근거가 될 것임.

## II. 기초 연구

이 장에서는 기초 어휘 선정 및 어휘 등급화의 방향 설정을 위하여 선행 연구를 검토하고 기 구축 말뭉치와 어휘 목록의 사례를 분석한다. 먼저, 이론 연구에서는 기초 어휘, 어휘 평정, 어휘 등급화 등 본 연구의 주요 개념에 관한 기존 논의를 검토하고, 이를 기반으로 본 연구에서 사용할 개념을 설정하였다. 또한 말뭉치 기반 기초 어휘의 선정 및 등급화의 방법론을 검토하여 시사점을 발견하고, 기초 어휘 및 어휘 등급과 관련된 문법적 쟁점을 검토하였다. 다음으로, 사례 연구에서는 국내와 해외에서 구축된 말뭉치와 말뭉치 기반 어휘 목록의 특성을 분석함으로써 본 연구에서 말뭉치 기반 어휘 목록을 작성할 때 고려해야 할 사항을 제시하였다.

### 1. 이론 연구

#### 1.1. 기초 어휘의 개념

##### 1) 기초 어휘에 대한 접근 방식

일반적으로 기초 어휘란 해당 언어의 의사소통에서 가장 기본적이고 핵심적인 어휘를 이른다. 어휘는 단어에 대하여 집합 개념을 말하므로 한 언어의 어휘 가운데 중요도가 상대적으로 높은 단어들의 집합으로 정의하기도 한다. 기초 어휘에 속하는 단어는 일상생활에서 사용빈도가 높고 파생이나 합성 등의 조어(造語)에 고빈도로 참여하며 다른 단어로 대체하기 어려운 특성을 지닌다고 알려져 있다.

한 언어의 어휘를 구성하는 수십만 개의 단어는 그 중요도가 각기 서로 다르다고 할 수 있다. 어떤 단어는 그 언어에 익숙하지 않은 어린이나 외국인도 거의 대다수 알고 있기도 하며, 어떤 단어는 평생 그 언어를 사용한 화자도 거의 접해 보지 못한 것도 있다.

국내외의 기초 어휘에 대한 접근 방식은 어휘 선정의 작업을 전제로 하므로, 이를 근거로 나누어 볼 수 있다. 심재기 외(2016: 127~128)에 따르면, 어휘 교육을 위한 관점, 외국인의 원활한 의사소통을 위한 관점, 한 언어의 근간이 되는 어휘를 밝히고자 하는 관점, 역사언어학적 관점이 그것이다.

이러한 네 가지 관점은 기초 어휘에 대한 다양한 접근의 방식을 대표하고 있지만 이 네 가지 관점의 접근 방식에 따른 결과물이 크게 차이가 나는 것은 아니다. 기

초 어휘에 대한 판정 기준이나 접근의 태도가 상이하다고 하여, 완전히 다른 어휘 목록이 제시되는 경우는 드물기 때문이다. 가령 해당 언어의 근간이 되는 어휘는 마찬가지로 계통적으로도 중요할 가능성이 높으며, 교육이나 언어 발달적인 견지에서도 유의미할 수 있을 것이다.

국어에서 기초 어휘는 어휘론, 국어교육 및 한국어교육에서 향후 발전 과제로 꾸준히 언급되어 왔던 중점 연구 영역이다. 그러나 국어에서 기초 어휘에 대해 논의해 온 중요성에 비해, 구체적인 목록을 제시하고, 이에 근거하여 기초 어휘의 개념론이나 그 성격을 논의한 연구는 많지 않은 편이며 실제 한국어 화자 전체에 대한 대표성을 목표로 이루어진 연구는 거의 전무한 편이다.

이삼형 외(2016: 76~79)에서는 기존의 기초 어휘 연구는 주로 한국어교육에서 선평되었다고 지적하면서, 다음과 같이 주요 성과를 요약하고 있다.

<표 1> 국어교육용 어휘 목록 선정의 주요 성과

연구 성과	규모 및 선정 기준
국어의 기초 어휘에 대한 연구 (임지룡, 1991)	- 1,500개 - 선정원리: 절충적 방법(기존 연구의 고빈도어와 필자의 주관적 판정) <sup>1)</sup>
등급별 국어교육용 어휘 (김광해, 2003)	- 제1언어, 제2언어 교육의 영역 모두에 적용할 수 있는 등급으로 평정한 어휘 목록 - 어휘 등급의 평정: 메타 계량(기존 계량 처리된 자료를 충분히 구하여 그 분포 상황과 자료의 타당도를 함께 고려하면서 비교함으로써 중요도를 구하는 방법. 어휘 목록 14종에 대한 타당도 순위를 정한 다음 비교하여 등급화 수행)
초등학생 교육용 기초 어휘 선정 연구 (장현진 외, 2014)	- 초등학교 저학년의 기초 어휘 511개 - 선행 연구를 토대로 기초 어휘 533개 선정 후 교사 10명의 중요도 평정 - 품사별, 의미영역별 분류

위의 표에서 볼 수 있는 바와 같이, 국어의 기초 어휘에 대한 기존 성과는 대체로 언어학 분야보다는 교육 분야에서 실용적인 목적으로 사용하기 위하여 성하였다고 할 수 있다. 그러나 우리말에 서투른 외국인이나 초등학교 수준의 대상을 위한 연구에 머물러 있는 상황이며, 국어의 사용에 전반적으로 관여한다고 할 수 있을 만한 어휘 목록 확정이나 규모의 설계, 체계화된 연구 절차 수립, 대규모의 언어 자료에 근거한 연구 등 제반 분야에 대한 정교화가 필요한 실정이다.

1) 임지룡(1991: 43)은 1,500 단어를 기초 어휘로 선정한 뒤, 합성과 파생의 이차 어휘를 포함하면 대략 5,000개의 어휘로 확장될 수 있을 것으로 예상하였다.

## 2) 기초 어휘의 개념과 성격

기초 어휘에 대한 기존의 논의를 살펴보면, 현재 학계에서 기초 어휘의 개념은 기본 어휘와 혼용되어 사용되고 있다. 이 두 개념을 구분하여 사용하는 연구가 있는가 하면(임지룡, 1991; 김광해, 1993; 김중학, 2001), 구별하지 않고 혼용하여 쓰는 것도 있고(이충우, 1994; 김광해, 1988), 기초 어휘로 통일해서 쓰기도 한다(성광수, 1999).

임지룡(1991: 2~3)은 기초 어휘와 관련된 유관 개념을 통해 기초 어휘를 설명하고 있는바, 기초 어휘란 특정 언어 가운데 그 중추적 부분으로서 구조적으로 존재하는 어의 부분 집단이며, 기본 어휘란 어떤 목적에 따라 인위적으로 선정되어 공리성을 지닌 어의 집단이고, 기간 어휘란 어떤 특정집단을 대상으로 한 어휘조사에서 직접적으로 얻어지는 어집단의 골격적인 부분집단으로 보았다.

이를 통해 볼 때, 기초 어휘는 해당 언어 자체에 대한 어휘 연구라는 점에서 기본 어휘나 기간 어휘와 변별되는데, 이러한 점에서 국어의 전반을 다루는 본 연구의 성격과 관련되는 것은 기초 어휘의 개념에 가까운 것임을 알 수 있다. 또한 기초 어휘가 해당 언어의 쓰임에서 기저적인 영역을 이루는 것이라면, 기본 어휘는 그 목적에 따라 선정될 수 있으므로 일정 부분 겹치는 목록이 있을 수 있으나 근본적으로 목표로 하는 언어 기반이 다르므로, 양자는 서로 변별된다고 할 수 있다.<sup>2)</sup> 즉 기본 어휘의 목록을 따라서 말뭉치와 통계를 중심으로 볼 때, 기본 어휘는 어휘 빈도의 통계치에 따르고 기초 어휘는 이보다는 균형이 있고 체계적이라는 데 그 특징이 있다(이희자, 2003: 4).

한편, 이희자(2003)에 따르면, 기초 어휘는 다음과 같은 특성을 지닌다. 기초 어휘(basic vocabulary)는 ‘가장 기본적이고 핵심적이며 일상적으로 널리 쓰이는 단어들의 총체’로서 즉 ‘일상생활에서 쓰이는 횟수가 매우 잦고 사용 범위가 넓으며 따라서 역사적으로 보았을 때에도 그 목록이 잘 변하지 않는 특성을 지닌 것으로서 형태론적으로는 파생이나 합성 등 이차 조어의 근간이 되는 최소한의 필수어’(임지룡, 1991; 김중학, 1995 외 참조)라고 규정할 수 있다. 또한 기초 어휘로 선정된 것들은 일반적으로 모국어의 제1차 습득 어휘 목록으로 간주되는 것들인데 이는 외국어로서의 한국어교육에 있어서 필수 ‘습득’ 어휘 목록으로 활용되는가 하면, 김중학(2001)에서는 이를 어원 연구, 어휘 체계의 변천 양상 파악 등 통시적인 어휘 연구의 기초 자료로도 활용된다고 밝히는 등 그 쓰임새가 언어 연구와 언어 교육 양방면에서 매우 크다고 하였다.

2) 영어의 경우, 통상적으로 기초 어휘란 1,000~2,000 개 정도의 목록을 대상으로 한 것을 가리키는 데, Ogden 등이 선정한 기초 영어 체계(The system of Basic English)의 850개 목록 등이 유명하다.

## 1.2. 어휘 평정 및 등급화의 개념

일반적으로 ‘어휘 평정’은 ‘빈도 등급화’로 치환되어 이해되는 경향이 있다. 그러나 ‘어휘 평정’이란 어휘소들을 그 중요한 정도에 따라 등급을 매겨 배열하는 것(김광해, 2003)으로서, 기본 개념에 충실하게 판단한다면 어휘 평정과 빈도 등급화가 똑같은 개념은 아니다. 빈도 등급화는 어휘 평정 작업의 일환이 될 수 있으나 두 개가 완전히 같은 개념은 아님에 유의해야 한다. 빈도를 고려하지 않은 채 특정 목적에의 중요도에 따라 어휘를 등급화하여 순위를 매겨 어휘 평정 작업을 하는 것도 이론상 가능하다.

국어 어휘 평정의 대표적 예는 김광해(2003)가 있다. 이 연구에서 이루어진 어휘 평정에서는 1차 어휘는 대개 1등급, 2차 어휘는 대개 2~5등급, 나머지가 3차 어휘에 대응되는 결과를 보여 주었다. 이들 어휘들은 대개 아동의 어휘 발달과도 긴밀한 관련을 맺는데, 아동이 어휘를 획득해 가는 과정을 보면 대개 1차 어휘를 먼저 배우고 이후 학교 교육이 이루어지는 과정에서 점차 2차 어휘에 익숙해지며 가장 나중에서야 3차 어휘를 획득해 가는 양상을 띠는 것으로 상정해 볼 수 있다.

구분 연령	어휘량		연간 증가율	
	남	여	남	여
6	5,606	5,158	18.7	21.6
7	6,655	6,250	17.5	16.5
8	7,822	7,283	25.4	27.1
9	9,812	9,256	31.0	36.5
10	12,863	12,635	35.0	42.9
11	17,359	18,057	33.0	39.9
12	23,085	25,254	24.8	22.9
13	28,809	31,035	19.3	15.1
14	34,379	35,726	14.8	11.2
15	39,475	39,722	11.2	6.9
16	43,886	42,447	8.7	4.0
17	47,721	44,161	4.9	2.3
18	50,069	45,190	1.8	0.6
19	50,985	45,467	0.4	0.1
20	51,176	45,496	-	-

[그림 4] 이해 어휘량의 발달(玉村文郎編, 1989: 159)

위 표에서 제시된 것은 일본 아동의 이해 어휘량 발달 과정을 보여 주는데, 어휘 수 정보만 있기 때문에 목록의 나열 외에 그 활용을 위한 실질적 정보가 적은 편이라는 한계가 있다.

이러한 점에서 볼 때 어휘 평정 결과는 최종적으로 그 어휘를 사용하는 집단, 특히 아동의 어휘량 발달 과정과 일정한 관계를 맺도록 설계되어야 할 것으로 보인다. 현재 우리나라의 경우에는 신뢰할 수 있는 대규모의 어휘량 발달 결과가 없는데, 어

휘 평정이 이루어지고 나면 아동 어휘량 발달 조사 과정이 훨씬 수월해질 것이다.

한편 ‘어휘 평정’과 관련하여 또 하나 주의해야 할 점은 등급화가 곧 수준별 유형화를 의미하는 것은 아니라는 점이다. 어휘의 ‘수준’이라는 개념도 모호할 뿐만 아니라 등급화와 수준은 동일한 개념으로 이해하기 어렵다. ‘1등급은 쉽고 7등급은 어렵다’ 등과 같은 단편적인 이해는 해당 등급 어휘들의 특성을 있는 그대로 이해하는 것을 방해할 수 있음에 유의해야 한다.

### 1.3. 말뭉치 기반 기초 어휘 선정 및 등급화 방법론

#### 1) 어휘 등급 설정을 위한 어휘 평정 기준 수립

어휘 평정을 하기 위해서는 평정의 기준을 수립해야 한다. 이때 중요하게 부각되는 문제가 ‘중요한 정도에 따라 등급을 매긴다’고 할 때에 논의되는 ‘중요성’의 개념이다. ‘중요성’을 어떻게 판단하느냐의 문제는 곧 어휘 평정 기준 수립과 관련된다.

어휘 평정 기준 수립은 어휘 평정의 목적에 따라 달라질 수 있다. 만일 한국어 의사소통을 목적으로 한국어를 배우는 외국인 학생들을 위한 사전 편찬용으로 어휘 평정을 시도한다면 한국어 기초 의사소통에의 유용성에 초점을 두어 중요성을 판단할 수 있다. 이 경우 구어 의사소통에서의 어휘 사용 빈도는 매우 유용한 중요성 판단 기준이 될 수 있다. 그러나 만일 한국어라는 전체 어휘의 분포도에 초점을 두어 어휘 평정을 시도한다면 빈도 외 다른 기준이 더 요구될 수 있다.

등급화하려는 어휘 집합의 성격 역시 어휘 평정 기준 수립에 영향을 미칠 수 있다. 예컨대 과학 학술 분야의 어휘를 대상으로 어휘 평정을 시도한다면 과학 학술 분야라는 특성이 고려되어야 한다. ‘과학, 학술’ 등의 성격과 관련하여 사고도구어(academic words)나 과학 전문어(technical words) 등을 어떤 방식으로 평정할 것인지가 어휘 평정 기준 수립에 영향을 미칠 수 있다. 특히 과학 전문어의 경우 대개의 경우 저빈도어로서 한국어 전체를 대상으로 평정할 경우에는 그 빈도가 매우 적게 나오겠지만, 과학 학술 분야 텍스트에서는 상대적으로 빈도가 높게 나올 수 있다. 또한 신명선(2004)에서 논의된 것처럼 ‘사려된다’와 같은 특정 사고도구어가 과학 학술 분야에서 특히 높게 사용되는 현상이 있을 경우 이러한 단어들을 어떻게 평정할 것인지를 논의되어야 한다.

본 연구진은 이번 연구를 첫 단계로 하여 장기적으로 한국어 전체 어휘를 대상으로 어휘 평정을 목표로 하여 세부적인 절차를 다듬어 가고자 한다. 이 작업은 한국어 전체 어휘를 대상으로 하는 만큼 우리말 어휘의 전반적인 특성을 드러낼 수 있고, 어휘 평정 결과가 국어 대사전 편찬이나 교육용 교재 편찬 등에도 유용하게 이

용될 수 있도록 제시되어야 한다. 이러한 작업은 전술한 것처럼 어휘 평정의 목적을 보다 구체화하고 우리말 어휘 집합이 갖는 전반적인 성격을 꼼꼼하게 분석할 때에 가능하다. 현재 본 연구를 통해 우리말(한국어) 어휘를 전체적으로 표집하고 그 방법론을 모색하는 과정에서, 예상할 수 있고 또 과정 중 설정하는 방식에 의해 상당히 차이를 보일 결과값을 예상해 보면서 이론적인 탐구를 수행하였다.

이번 연구 과정을 통해 국외의 다양한 어휘 평정 작업의 방식에 대한 이론적 탐구를 진행하고, 또 몇 가지 설정을 두고 진행한 말뭉치 선정 작업을 통해 예상되는 결과를 바탕으로 두고 평정의 내용론을 구체화하였다.

## 2) 어휘 등급의 기준 수립을 위한 고려 요소

어휘 등급 기준 수립을 위해 고려할 요소들을 정리하면 다음과 같다.

### (1) 등급 대상 단어의 수

한국어 총어휘가 <표준국어대사전>의 경우 50여 만, 우리말샘의 경우 100만이 넘는다. 그러나 등급화 대상이 되는 총 단어는 결국 최대치로 잡아도 20만 수준 안팎일 것으로 추정된다. 따라서 대개의 단어는 무등급 표시로 갈 수밖에 없다고 판단된다.

한편 ‘당대 사람들이 사용하는 단어는 총어휘 목록의 10~15% 정도(김광해, 2003: 16)’임을 고려해 본다면 사실 집중 등급 대상 목록은 5만~6만 개 정도로 추정해 볼 수 있다. 민간 차원에서 어휘 평정이 이루어진 ‘낱말 v.2001’의 경우 1~5 등급의 어휘량이 66,751개로서 일본어의 이해어휘량 발달<sup>3)</sup>로 판단해 볼 때 나름의 타당성을 확보해 구축되었음을 알 수 있다.

### (2) 어휘력의 구조

Nation(1990)은 어휘력을 설명하기 위해 어휘의 유형을 ‘고빈도어, 사고도구어, 전문어, 저빈도어’로 나누고 이들이 텍스트에서 차지하는 비율을 다음과 같이 제시한 바 있다.

3) 오가와라 히토시(2016)에 따르면, 일본 대학생 4학년의 사용 어휘는 약 3만 어, 이해 어휘는 약 4만 5천이라고 한다. (荻原 廣(2016), 「大学4年生の日本語の使用語彙は平均約3万語、理解語彙は平均約4万5千語」, 『京都語文』 23, 佛敎大学国語国文学会.)

<표 2> 어휘력을 구성하는 어휘의 양(Nation, 1990: 19)

유형	단어 수	빈도	텍스트 등장비율	어원	교수, 학습을 위한 참고
고빈도어	2천	모든 텍스트에서 자주 등장	텍스트 전체 어휘의 87%	절반 정도가 라틴어, 불어, 그리스어	이 단어들의 학습에 많은 시간을 투입해야 함. 꼭 알아야 함.
사고도구어	8백	대부분의 학술적 텍스트에서 자주 등장	학술적 텍스트 전체 어휘의 8%	2/3 정도가 라틴어, 불어, 그리스어	고등교육을 받으려면 이 단어들을 위해 많은 시간 투입해야 함. 꼭 알아야 함.
전문어	주제별로 1천-2천	전문적 텍스트에서 경우에 따라 자주 등장	전문적 텍스트 전체 어휘의 3%		어떤 과목을 학습한다는 것은 해당 교과과의 어휘 학습을 포함함. 각 교과 교사가 이 단어들을 다룰 수 있음. 단 국어교사는 학습 전략을 도울 수 있음.
저빈도어	약12만 3천	자주 등장하지 않음	어떤 텍스트든 2% 이상		이 단어들을 다루기 위한 교수 전략 필요. 이 단어들만을 가르치기 위해 시간을 할애할 필요성은 적음.

고빈도어는 말 그대로 빈도가 가장 높은 것이며 사고도구어, 전문어, 저빈도어로 갈수록 빈도가 낮아진다. 김광해(2003)는 저빈도어의 경우 말 그대로 빈도가 매우 낮아 그 유용성이 떨어지는 만큼 고빈도어, 사고도구어, 전문어만을 어휘력을 구성하는 요소로 삼았다. 김광해(2003)에서는 고빈도어를 1차 어휘로, 사고도구어를 2차 어휘로, 전문어를 3차 어휘로 재규정한 뒤 다음과 같은 표를 제시하기도 하였다.

<표 3> 어휘력의 관점에서 본 어휘의 구조(김광해, 2003)

문학	...	3차 어휘 전문어	과학	수학
역사	2차 어휘 사고 도구어			...
...				공학
사회	1차 어휘 고빈도어			기술
철학				.....
	思考 및 논리 전개를 위한 도구 기본어			
	여러 종류의 전문어			



이와 같이, 기초 어휘 선정을 목적으로 하는 본 연구에서도 1차 어휘, 2차 어휘, 3차 어휘의 발달 과정에 대한 분석이 이루어질 수 있도록 어휘 평정이 이루어지도록 할 필요가 있다. 또한 어휘 평정 시 ‘고빈도어, 사고도구어, 전문어, 저빈도어’와 같은 어휘 유형들과 그 특징을 적극 고려하여 어휘량의 발달 과정을 분석할 수 있고 어휘력의 구조를 좀 더 구체화할 수 있도록 해야 할 것이다.

### (3) 기초 어휘(1등급) 목록 선정과 말뭉치

기초 어휘(1등급) 목록 선정과 관련하여 말뭉치를 어떻게 설정할지는 적극적인 검토가 필요하다. 기초 어휘 선정 방법은 다음과 같은 두 가지 방법이 가정될 수 있다.

- ① 이원화 방법: 기초 어휘의 경우는 일상 구어 생활 텍스트로 구성된 말뭉치에서 추출
- ② 일원화 방법: 동일 균형 말뭉치 구축 후 기초 어휘를 추출하는 방법. 영어권의 경우를 참조하여 전체 텍스트의 80% 정도를 포괄하는 단어를 기초 어휘로 선정할 수 있음.

### (4) 텍스트 포괄 범위(Text Coverage)

어휘의 텍스트 포괄 범위에 대한 좀 더 구체적인 자료로는 다음과 같은 것이 있다. 이에 따르면 1등급 어휘가 텍스트의 대략 80%를 차지하는데, 그 포괄 범위는 등급이 올라갈수록 낮아진다. 여기에서 1등급 어휘는 대개 기존 논의에서 자주 다루어졌던 기초 어휘 개념에 해당한다.

<표 4> 문자 및 음성 언어에서의 텍스트 포괄 범위 비교(Nation, 2006: 79)

어휘 등급	등급 수	텍스트(구어) 포괄 범위 (%)	텍스트(문어) 포괄 범위 (%)
1st 1,000	1	78-81	81-84
2nd 1,000	1	8-9	5-6
3rd 1,000	1	3-5	2-3
4th-5th 1,000	2	3	1.5-3
6th-9th 1,000	4	2	0.75-1
10th-14th 1,000	5	< 1	0.5
고유명사	1	2-4	1-1.5
그 외	1	1-3	1

(5) 빈도(Frequency), 사용 범위(Range), 산포도(Dispersion)<sup>4)</sup>

말뭉치 기반 어휘 목록 선정 작업에서 가장 중요하게 고려되는 요소는 빈도이다. 빈도는 어휘 평정 시 고려되는 매우 중요한 요소인데다 어휘 평정의 결과가 사전 편찬이나 교재 편찬 등에서 그 유용성을 확보하기 위해서는 주요하게 고려할 수밖에 없다.

사용 범위는 하나의 어휘가 얼마나 다양한 말뭉치에서 사용되는지를 측정하는 어휘 선정 기준이다. 예컨대 인문, 사회, 과학, 예술의 4개 분야로 나누어 네 개의 말뭉치를 구축하였는데 어떤 하나의 어휘가 인문 말뭉치에서만 고빈도로 등장하고 과학 분야 말뭉치에서는 등장하지 않는다면 해당 단어는 사용 범위가 낮은 단어가 된다.

산포도는 사용 범위가 갖는 한계를 극복하기 위해 제안된 개념으로, 사용 범위가 단순히 한 단어의 빈도가 한 번이라도 몇 개의 다른 말뭉치에서 사용됐는지를 측정하는 것이라면, 산포도는 해당 단어가 한 말뭉치 안에서 일정한 빈도를 유지하는 정도를 측정하는 것이다. 어휘의 산포도를 구하는 공식으로는 다음과 같은 것이 있다.

<표 5> 산포도 공식(신동광, 2011 참조)

$$\text{Dispersion(산포도)} = 100 \times [1 - (V / \text{사용한 corpus의 수} - 1)]$$

\* V = 각 corpus에서 나타나는 type들이 가지는 빈도의 표준편차/각 corpus를 구성하는 token의 평균

## (6) 친숙도(Familiarity)

신동광(2011)에서는 어휘 사용 정도의 친숙도에 초점을 두어 척도를 개발하였는데, 어휘 친숙도 측정 척도는 다음과 같다. 이러한 방식은 기존의 친숙도 척도가 다음 Waring(2000)과 같이 어휘 지식의 깊이에 초점을 둔 것과는 구별된다.

<표 6> 어휘 친숙도 측정 척도(신동광, 2011)

척도	기준 설명
5	표현의 사용이 매우 친숙하다.
4	표현의 사용이 비교적 친숙하다.
3	표현을 잘 사용하지 않지만 의미는 잘 알고 있다.
2	표현을 거의 사용한 적이 없으나 의미는 대략 알거 같다.
1	표현을 듣거나 본적이 없어 의미를 전혀 알지 못한다.

4) Nation의 이론에서 기초 어휘 선정에 위한 세 기준 중 ‘산포도(dispersion)’라 불리는 것은 ‘분포빈도수/사용분포(spread frequency)’라고 하기도 한다. ‘산포도’라는 용어가 통계학에서 다른 의미로 쓰여서 혼란의 여지가 있기는 하나, 해당 단어가 특정 장르에 편중되지 않고 여러 장르에 골고루 퍼져서 나타난다는 의미를 살리기에는 ‘산포도’라는 번역어가 좋다고 생각한다.

<표 7> Waring(2000)의 어휘 지식 측정 척도

등급	친숙도 정도
1단계	단어를 알지 못한다.
2단계	단어의 의미를 이해한다고 생각하지만 실제 사용하지는 못한다.
3단계	단어의 의미를 이해하지만 실제 사용하지는 못한다.
4단계	단어의 의미를 이해한다고 생각하고 실제 사용할 수도 있다.
5단계	단어의 의미를 이해하고 실제 사용할 수도 있다.

신동광(2011)에서 언급된 것처럼 친숙도 척도는 빈도, 사용 범위, 산포도와는 달리 객관적인 수치로 나타내기 어렵다. 관련 전문가들을 대상으로 한 광범위한 조사를 바탕으로 하더라도 주관성이 개입된다는 문제점이 여전히 제기될 수 있다. 그러나 언어를 배우고 사용하는 환경에 따라 유용성과 필요성 면에서 친숙도가 다른 어휘가 분명 존재할 수 있다는 점에서 검토가 필요한 기준이다. 예컨대 한국에서 영어를 배우는 학생들에게 ‘classroom, textbook, livingroom, watermelon, cute, song’ 등(교실영어나 우리나라 환경에 친숙한 과일이나 음식명 등)은 그 친숙도가 원어민보다 더 높을 수 있다.

이러한 양상은 국내 국어교육 상황에서도 발견된다. 교과서에 특정 제재가 반복해서 등장함으로써 해당 제재에 나오는 특정 단어의 친숙도가 매우 높아지는 경우 등에서 나타난다. 어휘에 따라 해당 단어가 전문어로서 그 개념 이해가 매우 어려운 단어라 할지라도 학생들은 해당 단어를 쉽게 이해하고 사용하고 있을 수 있다. 이는 교육적 국면 외 특정 시기의 사회적 이슈와 관련해서도 불거지는 문제이다. 이러한 점을 고려할 때에 ‘친숙도’ 역시 어휘 평정 시 신중하게 고려될 필요가 있다.

## (7) 어휘 평정 기준의 적용 순서

어휘 평정 기준은 위에서 언급된 다양한 요소들에 대한 본격적인 검토를 통해 설정된다. 그런데 이들 적용 기준이 설정되더라도 이들을 어떤 순서로 적용할 것인가의 문제가 남는다. 다음은 주요 영어 어휘 목록 제작에서 어휘 선정 기준을 적용한 순서를 정리한 것이다.

<표 8> 어휘 목록 제작 시 어휘 선정 기준 및 적용 순서(신동광, 2011: 225)

어휘 목록	어휘 선정 기준 및 적용 순서			
NEAT3000	사용 범위	→	산포도	→ 빈도
BNC3000	사용 범위	→	빈도	→ 산포도
OXFORD3000	빈도	→	사용 범위	→ 친숙도

### (8) 어휘 평정과 분야별 등급화

한국어 전체 어휘를 대상으로 어휘 평정 작업을 진행하지만, 그 과정에서 인문, 사회, 과학, 예술 등 분야별 등급화 역시 고려해야 한다. 분야별 등급화의 결과 도출될 1등급 어휘는 한국어 전체 어휘를 대상으로 할 때 설정되는 1등급 어휘와 다소 다를 것으로 판단된다. 전자가 기초 어휘의 성격을 강하게 띠고 있다면, 후자는 해당 분야의 기반이 되는 어휘라는 점에서 한국어 기초 의사소통을 하기 위해 꼭 필요한 양자는 차별화된다.

한편 분야별 등급화를 하기 위해서는 어떻게 분야를 나누어야 하며 몇 개로 나누어야 하는지 역시 문제가 될 수 있다. 이는 말뭉치 규모와 분야별 어휘 특징에 대한 적극적인 분석을 통해 수행되어야 한다.

### (9) 어휘 평정 등급

어휘 평정의 결과는 어휘의 등급화로 귀결되는데, 이 과정에서 주요하게 부각되는 문제가 등급의 수이다. 김광해(2003)의 경우 총어휘 237,990개를 교육적 중요도에 따라 총 7등급으로 나누었다. 이후 ㈜날말은 김광해(2003)의 7등급 어휘 체계(아래에서는 날말 v.2001로 표기됨)를 9등급 체계로 보완하였다. 다음에서 알 수 있듯이 김광해(2003)의 3, 4, 5등급이 ‘㈜날말’에서는 3, 4, 5, 6, 7등급으로 좀 더 세분화되었다.

<표 9> 등급별 어휘량의 변화

날말 v. 2001				날말 v. 2004			
어휘 등급	어휘 량	누계	개념	어휘 등급	어휘 량	누계	개념
1등급	1,839	1,839	기초 어휘	1등급	1,675	1,675	기초 어휘
2등급	4,228	6,067	정규교육 이전	2등급	4,063	5,738	초등 1,2학년
3등급	8,361	14,428	정규교육 개시 + 사춘기 이전	3등급	7,736	13,474	초등 3,4학년
4등급	19,377	33,805	사춘기 이후 - 급격한 지적 성장	4등급	8,753	22,227	초등 5,6학년
5등급	32,946	66,751	전문화된 지적 성장 다량의 전문어	5등급	9,697	31,924	중등 1,2학년
6등급	45,569	112,320	저 빈도어 대학이상 전문어	6등급	11,552	43,476	중등 3학년, 고등 1학년
7등급	125,670	237,990	누락어 분야별 전문어	7등급	16,528	60,004	고등 2,3학년
등급외	234,725	472,715		8등급	52,987	112,991	저빈도어 대학이상 전문어
전체	472,715	472,715		9등급	106,615	219,606	누락어 분야별 전문어
등급 표기	237,990			등급외	253,109	472,715	
				전체	472,715	472,715	
				등급 표기	219,606		

몇 개의 등급으로 나누는 것이 타당한지에 대한 판단은 말뭉치 구축 이후, 기초 어휘에 대한 수차례에 걸친 어휘량 분석 결과와 샘플 말뭉치 구축 과정을 통해 얻은 결과를 바탕으로 연구진과 전문가들의 의견을 최종 수합하여 결정해야 할 것이다.

## 1.4. 기초 어휘 및 어휘 등급과 관련된 문법적 쟁점

### 1) 기초 어휘와 관련된 문법적 쟁점

#### (1) 기초 어휘 선정을 위한 분석 단위의 설정

##### ① 빈도를 고려한 품사의 설정

기초 어휘 선정에서 가장 중요하게 고려해야 하는 것은 빈도이므로 이를 고려하여 선정된 기초 어휘를 분류할 때 명사의 하위 범주로 의존 명사를 따로 설정하고 부사에서도 접속 부사를 따로 독립시킬 필요가 있다. 또한 ‘이다’와 ‘아니다’를 별도의 범주로 간주하는 등의 조치가 필요하다.

예를 들어 현대 국어의 사용 빈도를 조사한 국립국어원(2002, 2005)에서는 단어를 ‘일반 명사, 의존 명사, 대명사, 수사, 동사, 형용사, 보조 용언, 부정 지정사, 관형사, 일반 부사, 접속 부사, 감탄사’로 품사를 분류하였다.

##### ② 조사와 어미의 기초 어휘 포함 여부 및 처리 방향

기초 어휘 선정을 위한 해외 연구를 참조하면 텍스트 포괄 범위를 기준으로 한 경우가 적지 않은데 이때 기준이 되는 것은 단어형(word form)인 경우가 많다. 국어에서는 단어형이 조사 결합형과 어미 결합형에 해당한다. 따라서 기초 어휘가 조사 및 어미와 가지는 상관성에 대해 미리 정리할 필요가 있다. 이를 처리하는 방법은 여러 가지가 있을 수 있는데 결국 조사와 어미를 개별적으로 목록화하여 추출하되 이를 기초 어휘에서는 제외하는 방법이 가장 타당하다고 판단된다.

이와 관련된 예를 살펴보면, 국립국어원(2002, 2005)에서는 조사와 어미를 구분하여 범주화하되 격 조사는 세분하고 보조사와 접속 조사는 하나의 범주로 다루었다. ‘이다’는 ‘긍정 지정사’로 간주하여 조사의 아래에 포함시켰다. 그리고 어미는 선어말 어미, 연결 어미, 종결 어미는 세분하지 않고 전성 어미의 경우만 이를 명사형 어미와 관형사형 어미로 세분하였다.

##### ③ 생산성이 높은 접사의 기초 어휘 포함 여부

일부 기초 어휘 관련 연구에서는 생산성이 높은 접미사를 기초 어휘에 포함시키는 경우가 있으나 생산성이 높은 접사라도 어휘의 일부일 뿐이고 생산성이 높은 접사가 결합하고 있어도 결합하는 어근에 따라 기초 어휘 포함 여부가 달라질 수 있

으므로 접사만을 따로 기초 어휘에 포함시키는 것은 바람직하다고 판단되지 않는다. 이 문제는 분석 단위의 문제뿐만이 아니라 하나의 어휘를 중심으로 이와 형태론적인 측면에서 관련을 가지는 일군의 무리인 어휘군(word family)을 고려할 필요가 있다.<sup>5)</sup>

#### ④ 기타 단위의 처리

보조 용언은 대부분 독립된 항목으로 간주되어야 하나 ‘지다’와 ‘하다’의 경우는 ‘-아/어’가 결합하였을 때 하나의 단위로 처리하는 것이 바람직하다. 예를 들어 ‘이루어지다’, ‘행복해하다’는 하나의 단위로 처리하는 것이 바람직하다.

부사의 반복은 우리말의 단어 형성 방법 가운데 매우 생산적인 것이므로 하나의 단위로 처리하지만 부사형의 반복은 문장 구성 방법으로 간주되므로 하나의 단위로 처리하지 않는 것이 좋다. 예를 들어 ‘많이많이’는 하나의 단위로 처리하고 ‘늦게 늦게’는 하나의 단위로 처리하지 않아야 한다.

동일한 형식을 지니는 어떤 어휘를 다의어로 볼 것인지 아니면 동음이의어로 간주할 것인지에 따라 어휘 항목 수가 달라질 수 있다. 일차적으로는 다의어와 동음이의어는 <표준국어대사전>을 기준으로 판정하는 것이 바람직하다고 판단된다.

### (2) 기초 어휘 선정을 위한 어휘군의 처리

#### ① 파생어와 관련된 문제

빈도를 기준으로 기초 어휘를 선정할 때 생산성이 낮은 접사가 포함된 단어가 선정될 수 있고 반대로 생산성이 높은 접사가 포함된 단어가 선정되지 않을 수도 있다. 예를 들어 국립국어원(2005)에는 ‘살림, 살림꾼’, ‘소리, 소리꾼’, ‘심부름, 심부름꾼’ 쌍은 모두 포함되어 있으나 ‘씨름’과 ‘씨름꾼’에 대해서는 ‘씨름’만 있고 ‘씨름꾼’은 포함되어 있지 않다. 따라서 파생어와 관련하여서는 이를 어휘 등급 산정의 기준으로 적용하는 방안을 고민할 필요가 있다.

예를 들면, ‘살림, 살림꾼’, ‘소리, 소리꾼’, ‘심부름, 심부름꾼’은 모두 기초 어휘에 포함되지만 기초 어휘의 등급에서 ‘살림’, ‘소리’, ‘심부름’이 ‘살림꾼’, ‘소리꾼’, ‘심부름꾼’보다 더 기초적인 어휘 등급에 포함될 가능성이 높다. 다만 이 경우의 등급

5) ‘어휘군’은 달리 ‘단어족’이라고도 한다. 그런데 ‘단어’는 한국어의 경우 ‘조사’도 그 대상에 포함하는 경우가 있어 특히 단어 가운데 어휘적 단어에만 국한한다는 점을 강조하기 위해 ‘word family’를 ‘어휘군’으로 번역하여 사용하기로 한다. 다만 ‘어휘군’을 ‘하나의 어휘를 중심으로 이와 형태론적인 측면에서 관련을 가지는 일군의 무리’로 정의한다고 할 때 특히 어근이 모두 어휘의 자격을 가지는 합성어의 경우는 다른 어휘에서도 중복적으로 포착되므로 그 범위를 정하는 것이 쉽지 않을 때가 있다.

확정은 어근보다 파생어의 빈도가 더 높을 수 있으므로 상호 관련성을 고려해서 결정해야 한다. ‘씨름, 씨름꾼’의 경우에는 ‘씨름’만 기초 어휘에 포함되고 ‘씨름꾼’은 기초 어휘에 포함되지 않을 것이다. 따라서 ‘씨름꾼’은 더 높은 어휘 등급을 부여받을 것이다. 이러한 처리는 어휘 등급을 결정할 때 접사의 분석 처리가 영향을 미칠 수 있음을 의미한다. 따라서 접사를 구분하지 않고 파생어를 하나의 단위로 계산할 경우에는 저빈도 어휘이던 것이 접사를 구분하여 처리하면 고빈도 어휘가 되는 경우가 생길 수 있음을 고려해야 할 필요가 있다.

## ② 합성어와 관련된 문제

합성어에서는 단어와 구의 구별이 가장 중요한 문제인데 그 판정 여부는 일차적으로는 <표준국어대사전>을 참조할 필요가 있다. <표준국어대사전>에는 어근의 의미의 합이 형성된 합성어의 의미의 합과 다른 정도인 의미 합성성(semantic compositionality)뿐만이 아니라 의미 투명성(semantic transparency)이 높은 것들도 빈도를 고려하여 등재된 것들이 적지 않다. 따라서 이 두 가지 기준을 고려하여 비록 <표준국어대사전>에 등재되어 있지 않은 경우라도 합성어로 판정되는 것은 합성어로 판정하여 하나의 어휘로 처리하는 방안을 검토할 필요가 있다.

합성어 형성에서 자주 출현하는 단위 가운데는 독립된 단어로 존재하지 않는 경우가 있을 수 있다. 이때는 이를 따로 분석 단위로 삼아 의미에 따라 구분해야 할 필요성이 있는지 검토할 필요가 있다. 예로 ‘큰가시고기’, ‘큰마음’, ‘큰방’의 ‘큰’과 ‘큰어머니’, ‘큰아버지’, ‘큰처남’의 ‘큰’은 모두 독립된 단어로 존재하지 않지만 그 의미 차이가 분명하므로 이를 구분해서 처리해야 하는지 검토할 필요성이 있다.

## 2) 어휘 등급과 관련된 형태론적 쟁점

지금까지의 어휘 등급의 판정은 빈도를 일차적인 기준으로 삼되 여러 연구들에서 중복적으로 다루고 있는지를 고려하는 메타 계량 등의 방식을 동원하였다. 그러나 빈도를 일차적으로 고려하는 방식은 어휘 등급을 어휘의 중요도와 혼동하는 결과를 낳는 단점도 가지고 있다는 점에 유의할 필요가 있다. 빈도가 높은 단어가 반드시 교육 등에서 우선시되어야 하는 중요한 단어와 일대일로 대응하는 것은 아니기 때문이다.

가령 ‘법’과 ‘법률’, ‘거짓’과 ‘거짓말’의 경우 모두 기초 어휘에 속하는 것으로 간주되어 왔다. ‘법’과 ‘법률’의 경우 ‘법’보다 ‘법률’의 빈도가 현저히 낮으므로 ‘법’이 ‘법률’보다 기초 어휘의 등급이 더 높고 교육의 선후에서도 ‘법’이 먼저라는 데 이견이 없다. 그러나 ‘거짓’과 ‘거짓말’은 상황이 반대여서 ‘거짓’보다 ‘거짓말’이 빈도가

더 높지만 형태론적 관점에서 교육의 선후를 따진다면 ‘거짓말’이 ‘거짓’보다 먼저 배워야 할 단어 곧 더 중요한 단어로 간주되는 것은 문제가 있을 수 있다. 이것은 곧 어휘 등급 판정이나 중요도 결정에 있어 빈도 외에 형태론적 관점에서도 고려해야 할 사항이 있다는 것을 의미하는데 이러한 것 가운데 대표적인 것으로 어휘군과 의미 관계를 들 수 있다.

### (1) 어휘군과 어휘 등급

분석 단위를 어휘군으로 선정할 경우, 어휘군의 문제는 기초 어휘 선정에서도 매우 중요한 역할을 담당하게 된다. 또한, 선정된 어휘의 등급을 나누는 문제에서도 큰 역할을 담당한다고 할 수 있다. ‘뱀’과 ‘뱀물’, ‘거짓’과 ‘거짓말’은 모두 어휘군의 관계에 놓여 있는데 빈도가 어휘군의 복잡성과 일치하지 않는 ‘거짓’과 ‘거짓말’의 경우 보다 단순한 단어 구조를 가지는 단어인 ‘거짓’을 보다 복잡한 구조를 가지는 ‘거짓말’보다 더 높은 어휘 등급을 부여할 수 있는 판단 근거가 된다. 따라서 어휘 등급을 결정할 때 빈도와 형태론적 복잡성 가운데 어떤 것을 더 상위에 두어야 할 것인지에 대해 천착할 필요가 있다.

### (2) 상하 관계와 어휘 등급

어휘와 어휘 사이의 상하 관계는 두 가지로 나누어 분석할 필요가 있는데 하나는 공통되는 요소를 가지지 않는 경우이고 다른 하나는 공통되는 요소를 가지는 경우이다. ‘동물’, ‘개’, ‘삽사리’의 경우는 공통되는 요소를 가지지 않는 경우의 상하 관계에 해당하는데 ‘동물’과 ‘개’에서는 ‘동물’이 상위어이고 ‘개’와 ‘삽사리’에서는 ‘개’가 상위어에 속한다. 그러나 이들 가운데 기초 어휘로서의 등급이 가장 높은 것은 ‘개’인데 이는 상하 관계에서 가장 포괄적인 의미를 가지는 최상위어가 가장 등급이 높은 기초 어휘와 일대일로 대응하는 것이 아님을 보여 준다.

‘빛’, ‘푸른빛’, ‘연푸른빛’의 경우는 공통되는 요소를 가지는 경우의 상하 관계에 해당하는데 이러한 경우는 예외 없이 보다 짧은 형식이 상위어에 해당하고 또한 기초 어휘에 해당할 가능성이 높다. 따라서 ‘빛’이 ‘푸른빛’보다 기초 어휘로서의 등급이 높고 ‘푸른빛’이 ‘연푸른빛’보다 등급이 높으며 이들 가운데 ‘빛’이 기초 어휘로서의 등급이 가장 높다.

공통되는 요소를 가지는 경우의 상하 관계는 어휘군과도 매우 밀접한 연관을 가지는데 보다 복잡한 구조를 가지는 단어가 그렇지 않은 단어보다 기초 어휘의 등급이 높다고 보기는 어렵기 때문에 빈도에만 기대야 하는 단점을 어휘군과 상하 관계가 보완할 수 있다. 앞서 제시한 ‘거짓’과 ‘거짓말’의 관계가 이러한 예를 잘 보여 준다.



### (3) 유의 관계와 어휘 등급

국어의 경우 유의 관계는 고유어와 한자어 사이에 존재하는 것이 가장 일반적인데 이러한 어종도 어휘 등급 판정에 고려할 필요가 있다. 가령 김광해(1989)에서는 고유어와 한자어가 일대다의 대응 관계를 보이면서 서로 유의 관계에 놓이는 다양한 경우에 대해 언급하고 있는데 고유어 ‘땅’에 대해 ‘육지, 영토, 대륙’ 등이 이러한 관계를 갖는다고 보았다. 이는 곧 비슷한 의미를 나타내는 여러 어휘가 있을 경우 그 등급을 산정할 때 고유어가 한자어보다 기초 어휘의 등급이 더 높은 것으로 판단할 필요가 있음을 의미한다. 가령 유의 관계에 놓인 ‘금년’과 ‘올해’, ‘전부’와 ‘모두’의 경우 기초 어휘에 포함되어 왔고 한자어 ‘금년’, ‘전부’보다 고유어 ‘올해’, ‘모두’의 빈도가 높아 기초 어휘로서의 등급이 더 높은 것이 고유어에 해당하는 경우도 있다. 그러나 유의 관계에 놓인 ‘바뀌다’와 ‘변하다’의 경우 모두 기초 어휘에 속하지만 한자가 포함된 ‘변하다’가 빈도가 더 높다고 하더라도 이것이 기초 어휘로서의 등급이 더 높다고 볼 수 있는지 논의해 볼 필요가 있다.

이상의 논의들은 어휘 등급을 산정할 때 빈도 이외에도 어휘 사이의 계열적 관련성이 고려되어야 한다는 것을 의미하는 것으로 어휘 등급이 오로지 양적인 특성에 의해서만 결정되는 것은 아니라 질적인 측면에서도 타당성을 확보해야 한다는 점을 강조하기 위한 것이다.

## 2. 사례 연구

### 2.1. 국내 말뭉치 및 어휘 목록

#### 1) 국내 말뭉치의 구축 현황

국내에서 말뭉치 구축과 활용에 대한 연구는 1990년대 이후 본격적으로 이루어졌다. 국내 말뭉치 연구 중 많은 경우 한국어교육 영역에서 이루어져 왔으며, 구축 대상 자료는 주로 문어를 중심으로 구축된 경향이 있다.

현재까지 국내에서 구축된 주요 말뭉치를 제시해 보면 다음과 같다.

<표 10> 국어 말뭉치의 목록

기관	기간	말뭉치	목적	말뭉치 규모 (어절)
연세대학교	1990 ~ 2013	연세 말뭉치1	• 낱말 빈도 조사 • 사전 표제어 확정	288만
		연세 말뭉치2	• 국어 어휘 통계적 특성 파악	110만
		연세 말뭉치 3, 5, 6, 7	• 시대별 낱말 인지도 파악 (1960~1990년대)	3,548만
		연세 말뭉치4	• 구어 어휘 특성 파악	77만
		연세 말뭉치8	• 교육용 어휘 체계화	87만
		연세 말뭉치9	• 아동 교육용 어휘 파악	150만
		연세 구어 말뭉치		99만
고려대학교	1995	고려대 한국어 말모듬1	• 국어 연구를 위한 한국어 데 이터 베이스 구축 • 국어사전 편찬	1,000만
KAIST	1997	KAIST 코퍼스III	• 정보화 사회의 자연어 처리 • 언어연구와 사전 편찬	7,000만
국립국어원	1998 ~ 2007	세종말뭉치	• 기초 언어 말뭉치 개발 • 통합적 국가 말뭉치 구축	6,200만
한양대학교	2002 ~ 2003	연령별 구어 말뭉치	• 의사소통 능력의 발달 단계 연구	350만
고려대학교	2008 ~ 2013	물결 21 코퍼스	• 21세기 국어의 어휘 사용 양상 연구	60,000만
국립국어원	2010	2010년 한국어 학습자 말뭉치	• 한국어교육 연구를 위한 기반 자료 구축	400만
	2015	2015년 한국어 학습자 말뭉치	• 한국어교육 연구를 위한 기반 자료 구축	370만

위의 목록은 주로 한국어 어휘 측정을 파악하기 위한 기초 자료로서 구축된 것이다. 이 중 특수 목적의 말뭉치로는, 빈도 조사를 통한 사전 편찬을 위해 구축된 연세말뭉치1과 고려대 한국어 말모듬1, 교육용 어휘 체계화 연구를 위해 구축된 연세말뭉치8, 9가 있다. 그리고 연령별 구어 말뭉치는 의사소통 능력의 발달 단계 연구를 위해, 한국어 학습자 말뭉치는 한국어교육 연구를 위해 구축된 것이다. 이처럼 국내 말뭉치 연구는 언어 실태 파악과 교육용 어휘 선정을 주요 목적으로 구축되었음을 확인할 수 있다. 이처럼 말뭉치 연구의 목적과 말뭉치의 구성은 밀접한 관련성이 있다.

국어 말뭉치의 구성을 연령, 문어와 구어의 비율을 기준으로 살펴보면 다음과 같다.

<표 11> 국어 말뭉치의 구성과 특징

말뭉치	연령별 구분					문어/구어 구분		
	영유아	초	중	고	성인	문어	준구어	구어
연세 말뭉치1, 2		○	○	○	●	●	○	
연세 말뭉치3, 5, 6, 7		○	○	○	●	●		
연세 말뭉치4					●		●	●
연세 말뭉치8		●	●	●		●		
연세 말뭉치9	●	●				●		
연세 구어 말뭉치			○	○	●			●
고려대 한국어 말모듬1		○	○	○	●	●	●	●
KAIST 코퍼스Ⅲ					●	●	○	
세종말뭉치		○	○	○	●	●	○	○
연령별 구어 말뭉치	●	●	●	●	●			●
물결 21 코퍼스					●	●		
한국어 학습자 말뭉치					●	●		●

※ ○ : 전체 자료의 5% 미만인 경우 표기

위의 <표 11>은 국내 선행연구에서 구축된 국어 말뭉치를 대상으로 대상 연령과 문어, 구어 자료 포함 여부를 기준으로 정리한 것이다. 연령을 기준으로 볼 때, 대부분의 말뭉치가 성인 언어 자료에 집중되어 있음을 확인할 수 있다. 실제 영·유아 및 청소년의 사용 언어를 다룬 말뭉치는 한양대학교의 ‘연령별 구어 말뭉치’가 거의 유일하다. 연세 말뭉치1, 3, 5, 6, 7과 고려대학교 한국어 말모듬1, 세종 말뭉치에

서 구어를 일부 포함하고 있으나, 교과서와 동화 등을 대상 자료로 포함한 것으로 집필자가 성인이라는 점에서 실제 영·유아 및 청소년의 사용 언어를 반영했다고 보기는 어렵다. 다만 교과서와 동화의 예상독자가 영·유아 및 청소년이라는 점은 고려할 만하다. 한편 고려대학교 한국어 말모듬1에서 중학생의 작문 자료를 대상으로 하고 있다는 점도 특징적인 부분이다.

문어와 구어 자료 포함 여부를 기준으로 볼 때, 대부분의 말뭉치가 문어 자료에 집중되어 있다. 고려대학교 한국어 말모듬1과 세종말뭉치의 문어와 구어의 비율은 약 9:1로 구어의 비율이 낮은 편이다. 세종말뭉치의 경우 2007년 보고서 기준 원시 말뭉치 총 6,200만 어절 중 약 340만 어절만이 구어이다. 그리고 한양대학교의 연령별 구어 말뭉치와 연세 구어 말뭉치는 말뭉치 규모는 작은 편이지만 순구어 자료만을 대상으로 하였다는 점이 특징적이다.

한편 한국어 학습자 말뭉치의 경우 실제 학습자의 학습 자료 및 인터뷰/설문 자료를 바탕으로 구축한 말뭉치로 이에 대한 학습자의 오류 분석 말뭉치가 따로 설계되었다는 점이 특징적이다.

## 2) 어휘 목록

지금까지 국내에서 어휘 목록을 제시한 연구 성과는 교육용 어휘 목록에 집중되어 있으며, 국어교육보다는 한국어교육 영역에서 주로 수행되었다. 어휘 목록은 다양한 방향으로 활용 가능하나, 국내 연구에서는 언어 교육의 목적으로 제시된 사례가 대다수이다.

교육용 어휘 선정 절차에서 언어 사실을 반영한 기초 자료를 얻어내기 위하여 활용할 수 있는 요소로는 대규모의 말뭉치, 교과서, 기본 어휘 목록 및 학습 사전들의 표제어/중요 어휘 목록 등이다(서상규 외, 2009). 국내 한국어교육용 어휘 목록의 선정은 실제 언어 자료를 바탕으로 구축한 말뭉치에 기반한 양적 분석 방법이 주로 활용되었으며, 전문가의 평정을 통한 질적 분석 방법이 함께 사용되어 왔다. 구체적인 연구 성과를 살펴보면, 한국어교육용으로 선정된 최초의 어휘 목록인 한국어 능력 평가용 기본 어휘표는 연세 말뭉치에 바탕을 둔 빈도 조사 결과를 활용하여 만든 것이다. 서상규 외(2000)는 1998년도에 구성된 한국어교육용 말뭉치를 보완하여 주요 연구대상으로 삼아 한국어교육을 위한 기초 어휘 의미 빈도 사전 개발을 추진하였다. 이때 활용된 한국어교육용 말뭉치의 구성은 세종 글말뭉치와 세종 입말뭉치가 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)은 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.

국어교육용 어휘 목록과 관련한 연구는 한국어교육용 어휘 목록에 비해 그 수가 적으며, 기존 어휘 목록을 거듭 활용하는 방법을 주로 사용하였다는 특징이 있다. 그리고 구체적 어휘 목록을 제시한 연구보다는 기초 조사 성격의 연구가 많다. 국어 어휘에 대한 기초 조사는 문교부(1956)를 시초로 한다. 문교부(1956)에서는 “우리말 말수(어휘)가 사용되는 찾기(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, 2017). 이후 국립국어원 연구 등의 기초 조사가 시행되었으나 연구대상이 교과서 등의 출판물, 학생 작문 자료 등 문어를 주 대상으로 하여 언어 사용 실태를 오롯이 반영하지 못하였다는 한계가 있다. 국어교육용 어휘 선정의 직접적 활용 영역인 학령기 학습자에 대한 기초 조사는 국립국어원(2009)의 초등학교, 그것도 교과서나 글쓰기에서의 어휘 조사에 한정되어 있고, 초중고를 모두 포괄한 연구는 구어만을 대상으로 하고 있어(장경희 외, 2012) 국어교육용 어휘 선정을 위한 전면적인 조사는 아직 이루어지지 않은 실정이다(이삼형 외, 2017).

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. 이 연구에서 선정한 총어휘의 양은 모두 237,990어이며, 이들을 교육적 중요도에 따라 총 7등급의 집합으로 묶어 제시하였다.

한편, 교육적 활용을 주된 목적으로 두지 않고 어휘 목록을 선정한 연구로는 송철의 외(2008)의 한국 근대 초기의 어휘 연구와 한유석(2010)의 일한 분류어휘 비교 연구 등이 있다. 송철의 외(2008)는 한국 근대 초기에 해당하는 1901년부터 1910년 사이의 신문, 잡지, 소설, 교과서, 문법서 등을 바탕으로 말뭉치를 구축한 뒤, 당시의 시대상을 보여주는 어휘를 선정하여 제시하였다. 한유석(2010)은 일본 국립국어연구소의 『分類語彙表』에 입각하여 <연세한국어사전>의 표제어와 부표제어를 분류한 뒤, 양 언어를 서로 비교하기 쉽도록 병렬적으로 배치하여 제시하였다. 한편 전문 용어와 관련된 어휘 연구로는 최기선(2000), 유현경 외(2010) 등을 찾아볼 수 있다.

이처럼 국내 연구가 특정 영역에 치우친 경향이 있기는 하지만 어휘 목록 선정과 관련된 국내 연구들은 기초 어휘 선정을 목적으로 하는 본 연구와 밀접한 관련성을 가지므로 각 연구의 내용과 성과를 세밀히 살펴보도록 한다.

먼저 국내에서 수행된 어휘 선정 연구의 주요 성과의 목록을 간략히 제시하면 다음과 같다.

<표 12> 국내 어휘 선정 연구 성과(연도순)

연구자	연구 년도	연구 제목(성과)
문교부	1956	우리말 말수 사용의 찾기 조사
임지룡	1991	국어의 기초 어휘에 대한 연구
임홍빈	1993	국어 어휘의 분류 목록에 대한 연구
이충우	1994	한국어 어휘 교육을 위한 대표 어휘 선정
한국어능력 평가위원회	1997	한국어 능력 평가용 기본 어휘표
최길시	1998	외국인을 위한 한국어교육의 실제
서상규 외	1998	한국어교육을 위한 기초 어휘 선정-기초 어휘 빈도 조사 결과
연세대 언어 정보연구원	1998	연세한국어사전
서상규 외	2000	한국어교육 기초 어휘 의미 빈도 사전 개발
조현용	2000	한국어 어휘 교육 연구
임철성	2002	초급 한국어교육용 어휘 선정 연구
이익환	2002	기본 어휘 선정 및 사용 실태 조사를 위한 기초 연구
김광해	2003	등급별 국어교육용 어휘
김한샘	2003	한국 현대 소설의 어휘 조사 연구
조남호	2003	한국어 학습용 어휘 선정 결과 보고서
최기선	2004	(21세기 세종계획) 전문용어의 정비
김한샘	2005	현대 국어 사용 빈도 조사
서상규 외	2006	한국어 학습 사전 편찬과 기본 어휘의 선정을 위한 기초연구
서지영	2007	말뭉치를 활용한 중학교 국어과 학습사전 편찬을 위한 기초 연구
서정미	2008	말뭉치를 활용한 고등학교 국어사전의 편찬을 위한 기초 연구
이진영	2008	‘국어 초등 학습 용어 사전’ 편찬을 위한 국어교과 기본 어휘 연구
송철의 외	2008	한국 근대 초기의 어휘
김한샘	2009	초등학생 교과서 어휘 조사 연구
배주채	2010	한국어 기초 어휘집
국립국어원	2010	초등학생 글쓰기 어휘 조사 연구
한유석	2010	일한 분류어휘 비교
장경희 외	2012	초중고등학생의 구어 어휘 조사

이상의 연구 사례를 국어교육 분야와 한국어교육 분야로 나누어 그 개요를 살펴 보면 다음과 같다.<sup>6)</sup>

6) 연구 사례의 목록은 김광해(2003), 조남호(2003), 서상규 외(2006), 이삼형·김정선·김시정(2017)에서 주로 참조하였다.

## (1) 국어교육 영역

문교부	1956	우리말 말수 사용의 찾기 조사
연구 목적		<ul style="list-style-type: none"> <li>우리말 말수(어휘)가 사용되는 찾기(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.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>한글의 경우 1,259음절, 19만 2,463 빈도의 글자를 조사하였고, 음절에 따른 찾기조사의 결과표, 단자에 따른 찾기조사 집계표를 제시하였음.</li> <li>한자의 경우 5만 6,485어, 221만 8,727 빈도의 글자를 조사하였고, 한자 사용의 찾기 차례표, 한자 사용의 부수 차례표, 한자 조사 집계표를 제시하였음.</li> </ul>

임지룡	1991	국어의 기초 어휘에 대한 연구
연구 목적		<ul style="list-style-type: none"> <li>일상생활에서 사용되는 어휘 가운데 가장 기초적이고도 핵심적인 어휘의 목록을 작성하고 현실적인 문제로서 어휘교육을 위한 기초적인 목록을 작성하고자 함</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>기초 어휘는 언어생활에서 빈도가 높고, 분포가 넓으며, 파생이나 합성 등 이차조어의 근간이 되는 최소한의 필수어를 의미함. 기본 어휘는 어떤 목적에 따라 인위적으로 선정되며 공리성을 지닌 어휘의 집단을 의미함.</li> <li>기존 어휘 빈도 조사의 고빈도어를 참조하여 필자의 주관적 판정을 가미하여 어휘를 선정하는 ‘절충적 방법’을 사용하였음. 참조한 기존 어휘 빈도 조사로는 문교부(1956)의 ‘우리말 말수 사용의 찾기 조사’, 서연국(1969)의 ‘국어 기본 어휘의 연구’, 이응백(1972)의 ‘국민학교 학습용 기본 어휘’, 이연변 외(1980)의 ‘한국 아동의 어휘 발달 연구’, 정우산(1987)의 ‘국민학교 교과서 어휘 연구’, 국어연구소(1988)의 ‘중학교 교과서 어휘(국어)’, 임지룡의 ‘제5차 국민학교 국어 교과서 어휘 빈도 조사’가 있음.</li> <li>최근 8종 분류 어휘집을 참조하여 사람, 의식주, 사회생활, 교육 및 예체능, 자연계, 감각 및 인식, 동작, 상태, 기타의 9가지 의미 분야를 정하였음.</li> </ul>

임흥빈	1993	국어 어휘의 분류 목록에 대한 연구
연구 목적		<ul style="list-style-type: none"> <li>• 국어 사전 편찬 및 집필을 위한 필요 도구의 하나로써, 국어의 어휘에 대한 분류 목록을 작성하는 것을 목적으로 함.</li> <li>• 첫째, 어휘 분류의 개념을 분명히 함. 둘째, 분류 원리와 분류 방법론의 확립을 위하여 어휘 분류의 역사적인 측면에 주의를 기울이기로 함. 셋째, 본 연구에서 채택될 분류의 방법론을 제시하고 그 분류 목록을 보이기로 함.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>• 어휘 분류의 조건               <ul style="list-style-type: none"> <li>- 대상 조건: 어휘 분류의 대상은 ‘단어’이다.</li> <li>- 귀속 조건: 분류의 대상이 되는 ‘단어’는 동일 언어 체계에 속해야 한다.</li> <li>- 동질 조건: 하나의 부류에 포함된 단어는 동질적이어야 한다.</li> <li>- 기준 조건: 분류의 원리는 일정해야 한다.</li> </ul> </li> </ul>

이익환	2002	기본 어휘 선정 및 사용 실태 조사를 위한 기초 연구
연구 목적		<ul style="list-style-type: none"> <li>• ‘한국어의 기본 어휘를 선정하고 그 사용 실태를 조사’하기 위한 기초 자료로서의 말뭉치를 실제적으로 구축하기 위한 연구를 목적으로 함.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>• 쓸모 있는 말뭉치의 요건               <ul style="list-style-type: none"> <li>- 말뭉치를 만들기 위해 텍스트를 수집하거나 입력하는 과정에서, 원래의 내용이나 형태가 고의로든 실수로든 달라지거나 누락되지 않아야 함.</li> <li>- 말뭉치는 관찰하고자 하는 분야의 언어 사용의 축소판으로서 언어의 변이가 최대한 반영되도록 설계되는 것이 바람직함.</li> <li>- 말뭉치를 분석한 결과가 그 분야의 언어 사용에 대한 분석으로서 타당성을 가지기 위해서는, 해당 분야 언어의 다양한 특성을 고루 보여줄 수 있을 만큼 충분한 양의 텍스트가 수집되어야 함.</li> </ul> </li> </ul>



김광해	2003	등급별 국어교육용 어휘
연구 목적	<ul style="list-style-type: none"> <li>우리말에서 사용되는 실제 어휘에 가장 근접하는 목록 확보</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>어휘 선정이란 우리가 원하는 어휘집합을 구성할 것이라고 판단되는 어휘소들의 목록을 확정하는 작업을 말하며, 어휘 평정은 어휘소들을 그 중요한 정도에 따라 등급을 매겨 배열하는 작업을 말함.</li> <li>이 연구에서 어휘 평정을 위해 채택한 자료 처리 방법은 ‘메타 계량 방법’임. 더 높은 객관성을 확보하기 위해 구상된 것으로서, 직접 말뭉치를 계량하는 것이 아니라 기존에 계량 처리된 자료들을 충분히 구하여 그 분포 상황과 자료의 타당도를 함께 고려하면서 비교함으로써 중요도를 정하는 방법을 말함. 개별적인 계량 연구 하나하나가 그 빈도순 목록을 신뢰하는 데 문제가 있는 경우가 있지만, 이러한 개별 연구들을 충분히 모은 뒤 기본적으로 분포의 폭을 고려하되 타당도를 함께 고려하여 단어의 중요도를 정하는 것이 현재의 역량으로써 할 수 있는 가장 객관적인 중요도 판정 방법이라 할 수 있음.</li> <li>어휘 평정 단계에서 주요하게 고려된 변인은 분포와 자료의 타당도 두 가지임.</li> <li>이 연구에서 선정한 총어휘의 양은 모두 237,990어이며, 이들은 교육적 중요도에 따라 총 7등급의 집합으로 묶임.</li> </ul>	

김한샘	2003	한국 현대 소설의 어휘 조사 연구
연구 목적	<ul style="list-style-type: none"> <li>국어교육용 어휘의 단계별 선정 사업의 일환. 이 사업의 최종 목표는 국어교육 현장에서 이용할 수 있는 신뢰도 있는 국어교육용 기본 어휘 목록을 단계별로 확정하는 것임. 이는 대량의 문헌 자료를 대상으로 한 어휘 사용 실태 조사 결과와 기존 연구 검토 결과, 전문가 자문을 종합하는 매우 방대한 작업임.</li> <li>어휘 분석이 완료된 소설 문헌 약 100만 어절을 대상으로 하여 국어 사용 빈도 조사를 결과를 제시하고자 함.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>연구 대상은 초판이 1900년 이후 출판된 소설임.</li> <li>표본 추출 방법으로 확률 통계적 방법과 판단 추출법을 사용했음. 표본 후보를 형태 분석하여 빈도를 구하고, 1,000만 어절 말뭉치에서 누적 빈도 90%인, 상위 5,000여 개의 목록과 비교하여 그 사용률을 지표로 활용하였음.</li> <li>표본 선정 원칙                             <ul style="list-style-type: none"> <li>번역 자료는 조사 대상에 포함하지 않는다.</li> <li>외국어, 방언, 전문 용어 등이 많이 섞인 문헌은 배제한다.</li> <li>한문체이거나 고어 문체인 것은 배제한다.</li> <li>다양한 표본이 포함되도록 하기 위해서 한 작가의 작품이 3편을 초과하지 않도록 선정한다.</li> <li>입력 오류가 적고 파일의 수집이 용이한 ‘21세기 세종계획’의 자료를 최대한 활용한다.</li> </ul> </li> </ul>	

최기선	2005	(21세기 세종계획) 전문용어의 정비
연구 목적		<ul style="list-style-type: none"> <li>• 전문용어의 정비는 지식정보화 사회에서 요구되는 효율적인 정보의 교환, 확산을 목적으로 함. 정보의 기본단위인 용어의 사용에 대한 사회적 약속이 투명할 때 정확한 정보의 유통이 가능함. 현재와 같은 다원화된 지식공급의 체계 하에서는 용어들의 난립현상이 더욱 두드러지고 있음.</li> <li>• 연구의 필요성               <ul style="list-style-type: none"> <li>- 유럽, 일본을 비롯한 선진국은 오랜 전통 속에서 전문용어의 중요성을 인식하고 용어 정비 작업에 대한 기술을 축적함에 비해 우리나라는 전문용어 연구에 대한 전통이 일천함.</li> <li>- 현재의 지식정보화 사회를 맞이하여 정보의 효율적 유통을 위해 용어의 정비는 국가 산업의 인프라로 인식됨.</li> <li>- 용어의 정비는 물론 새로운 용어의 생성과 확장을 위한 전문용어 관리시스템의 확립과 개발공정의 체제 구축이 필요함.</li> <li>- 전문분야 정보의 대중화로 인한 용어정보의 서비스 필요성이 증대됨.</li> <li>- 전문분야별 용어학적 특성 연구를 바탕으로 정보산업 응용기술 개발에 기여</li> </ul> </li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>• 전문용어 기초자료 데이터베이스 구축</li> <li>• 전문용어의 표준화</li> <li>• 전문용어 표준화를 위한 통합검색 시스템의 확립 및 전문용어 관련 응용 툴 개발</li> <li>• 전문용어학적 연구를 위한 국내외 활동 강화</li> </ul>

김한샘	2005	현대 국어 사용 빈도 조사
연구 목적		<ul style="list-style-type: none"> <li>• 국어교육용 어휘의 단계별 선정 사업의 일환</li> <li>• 1900년 이후 출판된 문헌을 대상으로 한 현대 국어 사용 빈도 조사</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>• 되도록 다양한 문헌을 포함하여 조사 결과가 편중되지 않도록 하기 위하여 표본으로 선정된 문헌의 앞부분부터 약 5,000 어절 씩 끊어 300만 어절을 목표로 문헌 자료를 구성했음. 크기가 5,000 어절에 못 미치는 표본이 상당수 포함되어 있어 최종적으로 672개의 표본, 300만 어절의 문헌 자료를 수집하여 어휘 조사를 진행했음.</li> <li>• 결과적으로 300만 어절의 어휘 분석 말뭉치를 분석하여 자모, 음절, 일반 어휘, 조사, 어미, 어절 구성, 구, 어휘 범주, 활용형, 규범 오류형 등의 빈도를 조사하였음.</li> </ul>

서지영	2007	말뭉치를 활용한 중학교 국어과 학습사전 편찬을 위한 기초 연구
연구 목적	<ul style="list-style-type: none"> <li>중학교 국어 교과서 말뭉치를 활용하여 중학교 국어과 학습 사전을 편찬하는 데 필요한 표제어의 선정과 정보 기술 방법을 알아보고, 그 예를 제시하여 학습자들의 국어과 학습에 도움이 되는 국어과 학습 사전 편찬의 기반을 제공함.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>제7차 교육과정의 7, 8, 9학년 국어 교과서 6권을 대상으로 교과서 말뭉치를 구축하고, ‘깜짝새’를 이용하여 품사별, 어절별 색인, 어절별 용례 사전을 작성하였음. 학년별, 빈도별, 품사별 어휘를 분석하여 표제어를 선정하고, 국어과 학습 사전을 편찬하기 위한 지침을 마련하였음.</li> <li>최종 결과물로 15,431개 어휘를 제시하였음.</li> </ul>	

서정미	2008	말뭉치를 활용한 고등학교 국어사전의 편찬을 위한 기초 연구
연구 목적	<ul style="list-style-type: none"> <li>고등학교 국어사전 편찬에 필요한 사전의 형식과 내용을 마련하고자 함.</li> <li>고등학생이 고등학교 교과서를 공부하는 중에 반드시 이해해야 할 어휘에 관한 정보를 찾고자 할 때, 고등학생 학습자가 학습을 할 때 유용하게 활용할 수 있는 사전을 만드는 데 목표를 둠.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>7차 교육과정 국정 교과서 22책, 검정 교과서 56책 도합 78권의 본문 내용을 중심으로 말뭉치를 구축하였음. 색인 프로그램(글잡이 II)을 이용하여 표제항을 선정하고, 개별 단어의 쓰임을 보였음. 깜짝새를 이용하여 어절 색인과 용례 사전을 작성하였음.</li> </ul>	

송철의 외	2008	한국 근대 초기의 어휘
연구 목적	<ul style="list-style-type: none"> <li>이 연구의 목적은 시대상과 사회상을 잘 보여주는 어휘들을 고찰하는 데 있음. 개항기 이후 특히 20세기 초기는 우리 사회가 급속히 근대 사회적 면모를 갖추어 나아가는 시기로, 이 시기에는 급격한 외래 문화 수용과 더불어 새로운 단어들도 급속히 유입되기 시작했는데, 이러한 새로운 어휘들의 유입은 당시 시대상과 사회상을 잘 반영하고 있음.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>어휘 조사는 20세기의 첫 10년, 즉 1901년부터 1910년 사이의 신문, 잡지, 소설, 교과서, 문법서 등을 바탕으로 말뭉치를 구축한 뒤 이를 대상으로 집필 대상이 되는 표제어를 선정하는 방식을 취하였음.</li> <li>선정된 어휘에 대해서는 각각 표제 기본형, 원어, 이표기, 품사, 의미, 예문을 제시하였음.</li> </ul>	

김한샘	2009	초등학교 교과서 어휘 조사 연구
연구 목적	<ul style="list-style-type: none"> <li>• 국어교육용 어휘의 단계별 선정 사업의 일환</li> <li>• 초등학교 교과서 어휘 조사 연구</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>• 초등학교 전 학년 교과서 말뭉치를 분석하여 어휘 빈도를 조사하였음. 목록에서 순위, 학년별·과목별 빈도, 항목, 풀이, 품사 등의 범주 등을 제시하였음.</li> </ul>	
유현경 외	2010	전문 용어 자료 구축 및 정비를 위한 연구
연구 목적	<ul style="list-style-type: none"> <li>• 전문 용어 검색 시스템 탑재를 위한 풍부한 양의 자료 구축과 전문 용어에 대한 어문 정비 연구의 기틀 마련</li> <li>• 구체적인 목표는 다음과 같음. <ul style="list-style-type: none"> <li>- 기존 전문 용어 자료 보완 및 신규 행정용 전문 용어 자료 수집을 통한 통합적 자료 구축</li> <li>- 행정용 전문 용어의 사용 양상 분석을 통해 전문 용어 정비 및 개선 촉구</li> <li>- 기 구축 및 신규 수집된 행정용 전문 용어의 어문 정비를 통한 일관성 있는 전문 용어 형태 보급 방안 모색</li> <li>- 전문 용어 검색 시스템 탑재를 위한 구축 자료들의 형식 정제 및 변환</li> <li>- 구축 자료의 탑재를 통한 향후 전문 용어 통합 검색 시스템 활용 방안 모색</li> </ul> </li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>• 기 구축 및 신규 수집된 행정용 전문 용어를 수합하여 통합적으로 구축하고, 구축된 자료를 정비하여 ‘전문 용어 통합 검색 시스템’에 탑재하였음.</li> <li>• 행정용 전문 용어를 수집하여 신규 자료를 구축한 방법은 크게 네 가지가 있음. 첫째, 해당 기관에 요청하여 자료집을 제공받음. 둘째, 해당 기관 홈페이지를 통해 목록을 내려받음. 셋째, 해당 기관 홈페이지를 통해 크롤링함. 넷째, 해당 기관 홈페이지에서 용어 개별 목록들을 수작업으로 복사하여 구축함.</li> </ul>	

한유석	2010	일한 분류어휘비교
연구 목적		<ul style="list-style-type: none"> <li>일본 국립국어연구서의 『分類語彙表』의 분류체계에 입각하여 한국어를 분류한 뒤, 양 언어를 서로 비교하기 쉽도록 병렬적으로 배치하였음.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>분류어휘표는 ‘류, 부문, 중항목, 분류항목, 단락, 행, 분류어’의 7 단계의 구조를 가짐.</li> <li>한국어의 분류는 &lt;연세한국어사전&gt;의 표제어와 부표제어 전체를 대상으로 삼았음.</li> <li>본 연구의 의의는 다음과 같음. <ul style="list-style-type: none"> <li>한국어 시소러스 구축을 위한 기반 데이터로서의 역할</li> <li>양 언어 어휘 비교는 물론, 언어학의 제 분야의 비교연구의 용이</li> <li>다언어 정보검색, 자동통번역 등의 정보처리 분야 응용</li> <li>대역사전의 대응어 선정 및 국어사전의 의미 기술 정교화에 기여</li> <li>일본어 교육, 한국어교육 응용</li> </ul> </li> </ul>

장경희 외	2012	초·중·고등학생의 구어 어휘 조사
연구 목적		<ul style="list-style-type: none"> <li>한국 초·중·고등학생들이 실제 사용하고 있는 구어 어휘를 형태적, 의미적 측면에서 심층적으로 살펴보는 것을 목적으로 함.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>분석 대상은 한양대학교 한국교육문제연구소의 ‘연령별 구어 말뭉치’임. 분석 대상 자료의 화자 수는 총 478명(남 230명, 여 248명), 학년별 남녀 각 20명임. 녹음은 학교 교실 등 빈 공간에서 조사 대상자 둘 또는 세 명이 짝이어 자유롭게 일상적인 대화를 나누도록 하였고, 조사 분량은 각 대화 세트별로 1회에 1시간 동안 연속으로 녹음하는 것을 원칙으로 하였음.</li> <li>녹음 조사 후 음성 자료를 문자로 전사하여 기계가독형 말뭉치로 전환하였음. 그 후 형태 정보 주석, 의미 정보 주석을 하여 분석할 자료 형태로 가공하였음.</li> <li>학생들의 어휘 사용 특징이 파악될 수 있도록 단순 사용 빈도 외에도 사용 화자 수, 형태·의미의 하위 유형 등 여러 차원에서 분석을 수행하였음. 특히 고빈도 어휘를 중심으로 학교급별로 그 사용 추이를 분석하여 학교급에 따른 변화 양상을 구체적으로 살펴보았음.</li> <li>초·중·고등학생이라는 특정 집단의 언어를 대상으로 한다는 점에서 언어학적 의의가 큼. 또한 어휘 사용에 대한 관찰을 통하여 그 사용자의 인지 특성이나 그들이 속한 집단의 문화적 특성 등을 가늠할 수 있다는 점에서 국어교육적 의의가 큼.</li> </ul>

(2) 한국어교육 영역

이충우	1994	한국어 어휘 교육을 위한 대표 어휘 선정
연구 목적	<ul style="list-style-type: none"> <li>어휘 능력을 신장시키기 위한 일환으로 교육용 어휘를 구성하는 대표 어휘라 할 수 있는 어휘소의 선정 기준 수립과 어휘 선정을 목적으로 함.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>기초 어휘는 한정된 소수의 어휘 자료에 의해서 가장 기본이 되는 일상생활 각 영역에서의 필요가 충족될 수 있도록 계획적으로 선정된 것을 의미함. 기본 어휘는 일상생활에서 일반적으로 사용하고 사용 빈도가 높은 어휘 가운데 모든 사람들에게 공통되는 어휘 중 그 사회의 구성원으로서 정상적인 기본 생활을 하는 데 필요하다고 간주되는 것임. 학습용 기본 어휘와 기본 어휘를 구별하지 않음.</li> <li>어휘 선정 기준으로는 다음 8가지를 사용하였음. <ul style="list-style-type: none"> <li>- 사용 빈도가 높아야 한다.</li> <li>- 사용 범위가 넓은 어휘여야 한다.</li> <li>- 교육에 기초적인 어휘여야 한다.</li> <li>- 조어력이 높은 어휘여야 한다.</li> <li>- 학습자의 발달 단계에 맞는 어휘여야 한다.</li> <li>- 적용성이 큰 어휘여야 한다.</li> <li>- 시대가 요구하는 어휘여야 한다.</li> <li>- 고유명사, 계급명, 의성어, 의태어, 은어, 비속어, 유행어, 방언, 고어 등은 한정된 범위에서 선정해야 한다.</li> </ul> </li> <li>교과서 어휘 연구를 참조하여 한국어 학습용 어휘를 선정하였으며, 선정 결과로는 대표 어근 65어, 대표 접사 175어, 대표 한자어 형성소 175어를 제시하였음.</li> </ul>	

한국어능력 평가위원회	1997	한국어 능력 평가용 기본 어휘표
연구 목적	<ul style="list-style-type: none"> <li>한국어교육 및 한국어 능력 평가의 모형 수립에서의 기초적 자료로서 작성된 현대 한국어의 기본 어휘 목록 제시</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>연세대학교 한국어 사전 편찬실에서 만든 ‘현대 한국어 총 어휘 빈도표’를 사용하여 기본 어휘표를 작성하였음. 이 연구소에 비치된 약 4,300만 어절의 말뭉치를 형태소 분석하여 얻어낸 자료임.</li> <li>형태소 분석기의 정확도가 대체로 90% 전후로 일컬어지고 있으나, 실제 언어자료를 분석하였을 경우 적지 않은 오류를 포함하는 것으로 확인되어 사후 처리를 통해 오류를 제거하였음.</li> <li>총 10,740개 어휘를 제시하고 있으며, 품사에 따라 그 비율을 보면 명사가 약 57%, 동사가 약 20%, 형용사가 약 7%, 부사가 6% 순으로 높은 비율을 차지하고 있었음.</li> </ul>	

최길시	1998	외국인을 위한 한국어교육의 실제
연구 목적	<ul style="list-style-type: none"> <li>외국인을 위한 효율적인 한국어교육을 위해 교육용 기초 어휘와, 외국인이 한국에서 언어생활을 하는 데 있어서 필수적으로 습득해야 할 기본 어휘를 선정하고자 함.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>최길시(1994)의 일본어를 모어로 하는 사람들을 위한 한국어교육 방법 연구에서 발표한 1,000개 기초 어휘와, 각종 어휘 빈도 조사 자료를 참고하여 한국어교육 현장 경험을 바탕으로 기본 어휘를 선정하였음.</li> <li>어휘 빈도 조사 자료에서 사용 빈도가 높은 것을 기본으로, 요즘 일상생활어로 자주 쓰이는 신생어, 외래어들은 빈도 조사 자료와 상관없이 선정하는 방식으로 2,000개 어휘를 선정하여 제시하였음.</li> </ul>	

서상규 외	1998	외국어로서의 한국어교육을 위한 기초 어휘 선정 -기초 어휘 빈도 조사 결과
연구 목적		<ul style="list-style-type: none"> <li>외국어로서의 한국어교육을 위한 교육용 기초 어휘를 대규모의 실제 언어 자료에 대한 국어학적, 정보학적, 계량적 분석을 통하여 선정하는 것을 목적으로 함.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>대규모 연세 말뭉치와 교육 분야 텍스트의 어휘 빈도를 조사하여 두 결과의 상관성을 분석하고, &lt;조선어 빈도수 사전&gt;의 어휘 빈도의 분포와 대조하였음. 어휘 사용률과 어휘 증가율을 고려하여 어휘 분포와 단계를 분석하고 기본 어휘 후보 목록을 확정하였음. 어휘 목록의 검증과 보완 과정을 거쳐 기본 어휘 구간과 목록을 확정하였음.</li> <li>어휘 선정 시 누적 빈도율이 90%인 지점인 약 5,000개 어휘를 ‘기초 어휘 후보군’으로 추출한 후, 교과서, 조선어 빈도수 사전, 한국어 교재에 중복하여 등장한 어휘 목록과 비교하여 최종 어휘를 선정하였음.</li> </ul>

서상규 외	2000	한국어교육 기초 어휘 의미 빈도 사전 개발
연구 목적		<ul style="list-style-type: none"> <li>교재 개발, 학습 사전 편찬, 교과 과정의 개발 등 한국어교육 분야에서 응용될 수 있는 기반 정보인, 한국어 기초 어휘 의미빈도 제시</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>서상규 외(1998)에서 구성된 ‘한국어교육용 말뭉치’를 더욱 보완하여 1999년에는 총 100만 어절의 표준적 한국어교육용 말뭉치 구성을 완료하고, 이 말뭉치를 대상으로 2년간 의미주석 작업을 수행하였음.</li> </ul>

조현용	2000	한국어 어휘 교육 연구
연구 목적		<ul style="list-style-type: none"> <li>제7차 교육과정의 초등학교 국어 교과서 어휘를 조사하고 분석함. 분석 결과는 이후 기초 어휘 선정이나 교육과정 편찬에 이용될 수 있으리라 생각됨.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>객관적 말뭉치에 근거한 자료와 교육 전문가의 경험적 방법을 절충한 방법을 택하였음. 먼저 기존 자료(서상규 외(1998)의 어휘 빈도 자료, 최길시(1998)의 어휘 목록, 연세대 한국어학당 교재 편찬위원의 기본 어휘 자료)의 공통된 어휘 543개를 추출하였음. 그리고 누락된 어휘 등 182개 어휘를 추가하여 총 725개의 한국어교육용 기본 어휘 목록을 선정하여 제시하였음.</li> </ul>



임철성	2002	초급 한국어교육용 어휘 선정 연구
연구 목적		<ul style="list-style-type: none"> <li>• 초급 한국어교육용 어휘 선정</li> <li>• 연구의 목표는 다음과 같음. <ul style="list-style-type: none"> <li>- 빈도와 분산을 함께 고려하는 계량의 방법을 사용하여 객관성을 유지한다.</li> <li>- 가능한 한 직관에 일치하는 결과가 나오도록 한다.</li> <li>- 어휘의 체계를 고려한다.</li> <li>- 1,000개 상당의 어휘를 추출한다.</li> </ul> </li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>• 초급 한국어교육용 어휘는 한국어를 외국어로서 배우고자 하는 초급 단계의 학습자들이 익혀야 하는, 교육의 대상이 되는 어휘를 의미함.</li> <li>• 먼저 연세 말뭉치에서 빈도 5,000 이상인 어휘 500~600개 정도를 추출하였고, 다음으로 국내 초급용 한국어교육용 교재들에 등장하는 어휘에 대한 계량 분석을 실시하여 어휘를 추출하였음. 두 방법을 통해 얻어진 어휘를 바탕으로 연구자의 직관과 연세 빈도를 고려하여 어휘를 조정하여 총 1,038개 어휘를 선정, 제시하였음.</li> </ul>

조남호	2003	한국어 학습용 어휘 선정 결과 보고서
연구 목적		<ul style="list-style-type: none"> <li>• 우리말을 배우고자 하는 외국어 화자가 우리말을 잘 익히도록 하기 위해서는 한국어교육에 필요한 어휘를 선정하는 일이 먼저 이루어져야 하므로 한국어 학습용 어휘 목록을 선정하고자 함.</li> </ul>
주요 내용		<ul style="list-style-type: none"> <li>• 2000~2002년에는 150만 어절 규모의 말뭉치에 대한 빈도 조사를 수행하였음. 2002~2003년에는 한국어교육 전문가 6인에게 주관적인 어휘 평정을 의뢰하고 결과를 종합하여 어휘 등급을 결정하였음.</li> <li>• 현대 국어 사용 빈도 조사 목록 59,000여 개 중 출현 빈도 15회 이상인 단어 10,352개를 추리고, 이를 대상으로 등급 판정을 한 목록을 선정위원회에 배포하였음. 선정위원들이 단계별로 A, B, C, D(후보 단어), E(배제할 단어)로 판정한 결과를 기준으로 약 6,000여 개의 어휘 목록을 선정하였음.</li> <li>• 어휘 목록에서 조사, 어미는 제외하였으며, 복합어와 품사통용어는 전부 개별 단어로 처리하였음.</li> <li>• 최종 결과물은 5,965개 단어로 등급별로 보면 A등급 982개, B등급 2,111개, C등급 2,872개였음. 어휘 목록에는 순위, 단어 형태, 품사, 동음이의어 구별을 위한 풀이, 등급 등의 정보를 함께 제공하였음.</li> </ul>

서상규 외	2006	한국어 학습 사전 편찬과 기본 어휘의 선정을 위한 기초 연구
연구 목적	<ul style="list-style-type: none"> <li>말뭉치를 분석함으로써 얻을 수 있는 여러 가지 기초적 언어 정보가 한국어교육을 위한 기본 어휘의 선정을 위한 연구, 한국어 학습 사전의 기본 어휘 표제어나 중요 어휘의 선별 과정에 과연 도움을 줄 수 있는지, 어떠한 방법으로 이를 활용할 수 있는지를 모색함.</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>한국어교육 기본 어휘는 제2언어로 한국어를 학습하는 이들이 익혀야 할 기본적 단어로, 생활을 위한 필수 어휘와 한국어 학습에 기본이 되는 어휘를 아울러 포괄하는 것을 의미함.</li> <li>5개 말뭉치(연세 말뭉치, 한국어교육 표준 말뭉치, 한국어 교재 말뭉치, 초등학교 전체 교과서 말뭉치, &lt;연세한국어사전&gt; 뜻풀이 말뭉치)에서 중복도 3이상인 어휘 중 고유명사와 복합수사를 제외한 2,028개 어휘를 중요 어휘 1로 선정하였음. 한국어 교재 사용 어휘에서 중복도와 어휘 빈도를 고려하여 2,025개 중요 어휘 2를 선정하였음. 10종 기본 어휘 목록과 학습 사전의 중복도를 고려하여 1,977개 중요 어휘 3을 선정하였음. 이 세 종류의 어휘 목록을 대조, 통합하여 최종적으로 총 2,995개 어휘 목록을 제시하였음.</li> </ul>	

배주채	2010	한국어 기초 어휘집
연구 목적	<ul style="list-style-type: none"> <li>한국어능력 1급부터 4급까지 배우기에 알맞은 기초 어휘 제시</li> </ul>	
주요 내용	<ul style="list-style-type: none"> <li>1급 단어 300개, 2급 단어 600개, 3급 단어 800개, 4급 단어 1,000개로 총 2,700개 단어를 제시하였음.</li> <li>동철이의어를 구별할 때 국어사전에서는 어깨번호를 사용하지만 여기서는 확인어를 사용했음.</li> </ul>	

### 3) 어휘 등급화 단위 설정

국내 어휘 선정 연구 중 어휘의 등급화를 시도한 것은 국어교육 영역의 김광해(2003)와 한국어교육 영역의 조남호(2003) 등이 있다. 두 연구에서 등급화 단위를 어떻게 설정하고 있는지에 주목하여 살펴보겠다.

김광해(2003)에서는 우리말에서 사용되는 실제 어휘에 가장 근접한 목록 확보를 목표로, 이른바 ‘메타 계량 방법’을 사용하여 어휘 평정 작업을 하였다. 메타 계량 방법은 더 높은 객관성을 확보하기 위해 구상된 것으로서, 직접 말뭉치를 계량하는 것이 아니라 기존에 계량 처리된 자료들을 충분히 수집하여 그 분포 상황과 자료의

타당도를 함께 고려하면서 상호 비교함으로써 중요도를 결정하는 방법이다. 메타 계량 방법을 사용하며 고려된 변인은 ‘분포’와 ‘자료의 타당도’ 두 가지이다. 분포란 매우 자명한 것으로 하나의 단어가 얼마나 많은 자료들에서 등장하고 있는가 하는 양적인 측면을 중요도 판정의 변인으로 고려한 것이다. 그리고 자료의 타당도란 입수한 연구물에 나타난 어휘 목록이 얼마나 보편타당성을 지니고 있는가 하는 질적인 측면의 변인으로 고려한 것이다.

이 연구에서 메타 계량의 대상은 조현용(2000), 임지룡(1991), 이충우(1992) 등의 연구와 <연세한국어사전>, 고려대의 국어사전 표제어 자료, 21세기 세종계획의 전자사전 개발 자료 등 사전 자료 총 14건으로 개별 어휘소 합계는 약 9만 8천여 개이다. 이 연구에서는 등급별 어휘의 양을 결정하기 위해 일본, 러시아, 미국, 독일어의 사례를 참고하였으며 최종적으로 선정한 어휘의 양은 23만 8,010어로 어휘를 모두 7등급으로 분류하였다.

<표 13> 등급별 어휘의 상황(김광해, 2003: 27 수정 인용)

어휘량	누계	국어교육용		한국어교육용			비고	
		등급	개념	4구분	6구분	개념		
1,845	1,845	1	기초 어휘	초급	1	자국인과 어휘량을 일치시키는 방향으로 조절함	소사전	중사전
4,245	6,090	2	정규 교육 이전	중급	2			
8,358	14,448	3	정규 교육 개시 - 사춘기 이전, 사고 도구어 일부 포함	상급	3			
					4			
19,377	33,825	4	사춘기 이후 - 급격한 지적 성장, 사고도구어 포함	고급	5			
					6			
32,946	66,771	5	전문화된 지적 성장 단계, 다량의 전문어 포함					
45,569	112,340	6	저빈도어: 대학 이상, 전문어 (기존 계량 자료 등장 어휘 + 누락어 14,424어 추가)					
125,670	238,010	7	누락어: 분야별 전문어, 기존 계량 자료 누락어휘					

이 연구의 성과 중 하나는 그간 학계에서 확정되어 있지 않은 등급별 어휘량을 결정하여 제시하였다는 점이다. 이때 국어교육용 등급별 어휘량과 한국어교육용 등급별 어휘량이 반드시 일치한다고 보기는 어려우나 이를 일치시킨다는 방침 아래 작업을 수행하였다.

7등급으로 등급화된 김광해(2003)의 연구 성과는 이후 ㈜낱말에서 9등급 체계로 보완하였는데, 이는 기존의 3, 4등급을 3, 4, 5, 6등급으로 세분화하는 과정을 거친 것이다.

한편 조남호(2003)에서는 한국어 학습 단계를 3단계로 나누어 1, 2, 3단계의 한국어 학습용 어휘를 선정하는 방식으로 제시하였다. 이 연구의 어휘 선정 및 등급화 과정은 다음과 같다. 2000~2002년에는 150만 어절 규모의 말뭉치에 대한 빈도 조사를 수행하고, 2002~2003년에는 한국어교육 전문가 6인에게 주관적인 어휘 평정을 의뢰하고 결과를 종합하여 어휘 등급을 결정하였다. 구체적으로는 현대 국어 사용 빈도 조사 목록 59,000여 개 중 출현 빈도 15회 이상인 어휘 10,352개를 추리고, 이를 대상으로 등급 판정을 한 목록을 선정위원회에 배포하였다. 선정위원들은 1단계, 2단계, 3단계에 속할 수 있다고 생각되는 어휘를 각각 A, B, C로, 3단계에 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. 최종적으로 선정된 한국어 학습용 어휘의 등급별 개수를 품사를 구분하여 정리하면 다음과 같다.

<표 14> 한국어 학습용 어휘 목록 품사별 분포(조남호, 2003: 11)

등급 품사	A	B	C	계
명사	497	1,199	1,708	3,404
고유 명사	21	27	15	63
의존 명사	33	44	53	130
대명사	32	5	10	47
수사	45	2	-	47
동사	155	501	689	1,345
형용사	75	132	169	376
보조 용언	18	5	10	33
관형사	27	19	23	69
부사	65	146	182	393
감탄사	12	22	10	44
분석 불능	2	9	3	14
계	982	2,111	2,872	5,965

어휘 목록의 선정과 어휘 등급화 작업은 어휘 목록의 목적이 무엇인지와 밀접한 관련이 있다. 위에서 살펴본 두 사례에서도 국어교육용 어휘와 한국어교육용 어휘의 목적이 다른 만큼 선정되는 어휘의 양과 등급의 기준 등이 차이가 나는 것을 확인할 수 있다. 위의 사례에서는 학습 단계를 기준으로 등급을 분류하고 그에 맞추어 어휘를 판정한 결과물을 보이고 있는데, 반대로 등급별 어휘 수를 우선적으로 결정하고 계량적으로 등급을 판정하는 방법도 가능할 것이다.

이처럼 국내 어휘 연구에서 어휘 등급화와 직접적인 관련이 있는 연구가 아직 미진한 편이고, 충분한 연구 성과가 축적되지 못했으므로 어휘 선정 및 등급화 관련 해외 사례를 적극 활용해 방법론을 수립하여야 할 것이다.

## 2.2. 해외 말뭉치 및 어휘 목록

### 1) 해외 말뭉치의 구축 현황

영어권에서는 일찍이 어휘 교육의 효율성을 위해 말뭉치 분석을 다양한 방식으로 교수·학습에 활용하였다. 초기 말뭉치의 활용은 말뭉치로부터 유용한 표현을 추출하거나 사전의 예문 발췌에 주로 적용되었다. 이런 작업은 컴퓨터를 바탕으로 한 말뭉치 분석 기술이 활성화되면서 가능하게 되었으며 이와 같은 초기 말뭉치 활용 기법을 소위 “Behind-the-Scenes Approach”라고 한다. 이 접근법은 말 그대로 현장에서 한걸음 물러나 미리 준비한다는 의미로 온라인으로 말뭉치를 실시간 활용하는 것이 아니라 오프라인으로 말뭉치를 분석하여 빈도가 높은 어휘 같은 학습내용을 선정하거나 이를 바탕으로 교재를 개발하는데 활용하는 것이다. 그러나 말뭉치 활용에 관한 연구가 가속화되고 웹기반 인프라가 급격히 발전함에 따라 말뭉치 활용에 대한 새로운 요구들이 생겨나기 시작했다. 이 새로운 접근법은 말뭉치의 샘플을 온라인으로 직접 연결하여 그 예를 학생들에게 즉각적으로 제시하거나 말뭉치를 바탕으로 예시문항을 현장에서 바로 제작하여 활용하는 것으로 소위 “On Stage Approach”라 한다. 이 접근법은 Johns(1986, 1988, 1991)가 제안한 “Data-Driven Learning(DDL)”으로 더 널리 알려져 있다. 이는 말 그대로 언어 데이터인 말뭉치에서 자료를 끌어와 교수·학습에 활용하는 것으로 학습자들은 많은 샘플을 보면서 특정 표현의 의미를 유추해 보거나 그 실제적 쓰임을 직접 접하면서 문법 또는 어휘를 귀납적으로 습득하는 발견식 학습을 그 원리로 하고 있다.

본 연구는 국어과 기본 어휘 개발을 목표로 하고 있는바, 말뭉치 분석을 바탕으로 어휘 목록 개발에 활용하는 “Behind-the-Scenes Approach”에 기반하고 있다고 말할 수 있다.

미국에서는 1964년 브라운 대학(Brown University)에서 프란시스(Francis)와 쿠세라(Kučera)의 주도로 ‘Brown Corpus’라는 컴퓨터로 분석 가능한 최초의 전자 말뭉치(electronic corpus)를 구축하였다. Brown Corpus는 언어의 대표성을 확보하기 위해 다음의 <표 15>와 같이 15개의 영역에서 데이터를 수집하였다. 이는 각 2,000 단어로 구성된 500개의 문서로 총 100만 단어의 말뭉치이다. 최근의 말뭉치 구축 추세는 말뭉치가 거대화되는 추세이기는 하지만 Brown Corpus는 이후 말뭉치 구축의 모델이 되어온바, 90년대까지 대부분의 영어 말뭉치는 100만 단어의 규모로 구축되는 것이 일반적이었다(예: LOB, ACE, WSC, WWC).

<표 15> Brown Corpus의 주제 영역 및 구성 단어 수

주제 영역	규모 (단어 수)	주제 영역	규모 (단어 수)	주제 영역	규모 (단어 수)
1. Press Reportage	89,000	6. Popular Lore	87,000	11. Fiction Mystery	48,000
2. Press Editorial	55,000	7. Biography & Memoirs	152,000	12. Fiction Humor	18,000
3. Press Reviews	35,000	8. Government & Industry	63,000	13. Fiction Sci-Fi	12,000
4. Religion	21,000	9. Learned & Academic	163,000	14. Fiction Adventure	58,000
5. Skills & Hobbies	73,000	10. Fiction General	58,000	15. Fiction Romance	59,000

Brown Corpus와 같이 다양한 주제 영역에서 데이터를 구축할 경우 언어 사용의 대표성을 확보할 수 있을 뿐만 아니라 세부 영역별로 목적에 따라 활용하기에도 용이하다는 장점이 있다. 실제 Brown Corpus의 주요 영역만을 보면 크게 ‘Press(1, 2, 3)-179,000 단어,’ ‘Academic(9)-163,000 단어,’ ‘Fiction(10, 14, 15)-175,000 단어’로 거의 같은 규모로 균형 있게 구성되어 있다.

1978년 영국에서는 1964년에 미국에서 제작된 위의 Brown Corpus에 자극을 받아 Brown Corpus와 같은 형식에, 같은 크기로 Lancaster Oslo/Bergen(LOB) corpus를 구축하였다. 이는 Brown Corpus와 같은 구성과 규모라는 점에서 미국과 영국 영어를 비교할 때 주요 자료 사용된다. 또한 London Lund Corpus(Survey of English Usage)는 1959년에 University College London에서 퀴크(Quirk)가 주도하여 구축한 Survey of English Usage(SEU)의 자료를 디지털화한 후 1975년에 Svartvik이 Lund University에서 제작된 Survey of Spoken English(SSE) 자료와 통합하여 구축한 대표적인 영국 말뭉치이다. 이 말뭉치는 50만 단어 규모로 1953~1987년의 구어 자료를 바탕으로 구성되어 있다.

말뭉치 분석에 기반한 어휘 연구 중 상업적 성공과 대중화에 성공한 사례는 Collins COBUILD English Language Dictionary가 최초의 사례이다. Birmingham University의 Sinclair(1987)가 주도한 ‘COBUILD(Collins Birmingham University International Language Database) Project’는 ‘Bank of English’라는 대규모 영어 데이터베이스를 구축하였고 1980년대 말뭉치 기반 사전인 Collins COBUILD English Language Dictionary를 출판하여 큰 성공을 거두었다. 당시 구축된 Bank of English의 크기는 2억 단어 정도였지만 현재는 4억 5천만 단어가 넘는 대규모 데이터베이스가 되었다. 이 말뭉치는 25%가 구어이며 75%는 문어로 구성되어 있으며 70%는 영국 영어이고 20%는 미국 영어 그리고 10%는 그 밖의 영어권 국가에서 수집되었다. 또한 Collins COBUILD English Language Dictionary에 수록된 모든 예문은 이 말뭉치에서 발췌한 대표성 있는 실제적 표현이어서 이것이 상업적 성공의 이유가 되었다.

또 하나의 대규모 말뭉치는 British National Corpus(BNC)로, 이는 가장 보편적으로 사용되고 있는 1억 단어의 영국 말뭉치이다. 1991~1994년 사이에 수집된 데이터로 10%는 구어이고 90%는 문어로 4,049개의 문서로 구성되어 있다. BNC를 구성하는 주제별 비율은 다음의 <표 16>과 같으며 말뭉치의 대부분을 차지하는 문어 자료의 세부 주제 영역은 <표 17>과 같다.

<표 16> BNC의 대(大)주제 영역 및 구성 단어 수

주제 영역	문서 수	규모(단어 수)	%
Spoken demographic	153	4,233,955	4.30
Spoken context-governed	755	6,175,896	6.27
Written books and periodicals	2,685	79,238,146	80.55
Written-to-be-spoken	35	1,278,618	1.29
Written miscellaneous	421	7,437,168	7.56

<표 17> BNC 문어 자료의 세부 주제 영역 및 구성 단어 수

주제 영역	문서 수	규모(단어 수)	%
Imaginative	476	16,496,420	18.75
Informative: natural & pure science	146	3,821,902	4.34
Informative: applied science	370	7,174,152	8.15
Informative: social science	526	14,025,537	15.94
Informative: world affairs	483	17,244,534	19.60
Informative: commerce & finance	295	7,341,163	8.34
Informative: arts	261	6,574,857	7.47
Informative: belief & thought	146	3,037,533	3.45
Informative: leisure	438	12,237,834	13.91

전 세계적으로 구어 말뭉치 자료는 수집뿐 아니라 녹음 및 전사 그리고 데이터클리닝 같은 추가적인 작업으로 인하여, 시간 및 비용 소요가 문어 말뭉치 구축에 비해 상대적으로 크기 때문에 규모가 작고 구축된 사례도 많지 않다.

Cambridge Nottingham Corpus of the Discourse of English(CANCODE)는 구어 말뭉치로 1995~2000년 사이 수집된 500만 단어의 영국 말뭉치이다. 현재는 제작사인 Cambridge University Press와 매카시(McCarthy)나 슈미트(Schmitt)와 같은 참여 연구자들이 독점하고 있으며 일반 연구자에게는 공개가 되지 않는 자료이다.

이 밖에도 International Corpus of English(ICE)는 영국, 미국, 홍콩, 싱가포르, 남아공, 인도, 필리핀 등 여러 영어권 국가에서 1990년부터 수집이 시작된 말뭉치 세트로 Brown Corpus와 같이 각 2,000단어로 구성된 500개의 문서로 구성되어 있다. 구어와 문어의 구성 비율은 60%가 구어이고 40%는 문어 자료가 차지하고 있다.

상대적으로 말뭉치 연구에 소극적이었던 미국은 최근에 이르러 세계 최대의 말뭉치를 구축하였다. 1990~2015년 사이에 Brigham Young University의 Mark Davies가 최신의 자료로 구축한 이 말뭉치는 Corpus of Contemporary American English(COCA)이며 총 5개의 대영역으로 구성되어 있고 2015년 기준으로 총 5억 3천만 단어 규모에 이른다. 그 주제 영역과 규모는 다음과 같다.

<표 18> COCA 주제 영역별 단어 수

주제 영역		규모(단어수)
Spoken	TV 및 라디오 프로그램 대본	109,391,643
Fiction	단편소설 및 문학잡지	104,900,827
Popular Magazines	다양한 영역의 잡지	110,110,637
Newspapers	다양한 주제의 신문 기사	105,963,844
Academic Journals	다양한 영역의 학술지	103,421,981

말뭉치의 구축에는 상당한 인력과 시간이 소요되는 만큼 COCA를 구축한 마크 데이비스(Mark Davies)는 기존의 말뭉치 구축 방식과는 달리 인터넷 자료를 바탕으로 COCA를 구축하였다. 단시간에 세계 최대 규모의 말뭉치를 구축할 수 있었던 이유도 여기에 있다. COCA는 최대 규모의 최신 자료를 바탕으로 구축되었다는 장점이 있는 반면 인터넷 자료라는 한계로 인해 주로 인터넷 상의 신문이나 TV 방송 대본 등 언론 매체의 자료에 지나치게 의존한 나머지 일반 영어 사용의 대표성에는 다소 문제가 있는 것으로 나타났다(신동광·전유아·이신웅·박명수, 2017).

지금까지는 특정 목적을 가지고 구축된 말뭉치가 아니라 하나의 언중(言衆)의 언어 사용을 대표하기 위해 구축된 영어 말뭉치에 대해 살펴보았다. 최근의 말뭉치



연구의 추세는 특정한 목적에 최적화된 ESP(English for Specific Purposes) 말뭉치를 구축하거나 기존의 말뭉치를 원하는 목적에 맞는 비율로 재구성하여 대규모 말뭉치를 구성하는 방식을 채택하고 있다. 이러한 말뭉치 구축 방식은 최신의 어휘 개발 프로젝트에 폭넓게 적용되고 있다.

현존 영어 어휘 목록 가운데 가장 타당성이 높다(신동광, 2014)고 판단되는 BNC-COCA 25000(Nation & Webb, 2011)은 기존의 어휘 목록 개발에서는 적용되지 않은 차별화된 방식의 말뭉치 구축을 채택하였다. 사실 하나의 위계 등급으로 아동과 성인의 어휘 학습을 모두 포괄하는 어휘 목록을 개발하는 것은 쉽지 않다. 그 이유는 지금까지 구축된 대부분의 말뭉치는 연구윤리나 편의상 성인의 언어 사용 데이터를 기반으로 구축되었기 때문에 아무리 빈도와 같은 객관적인 어휘 선정 기준을 적용해도 아동의 관심과 학습의 실제성을 모두 반영하는 데는 한계가 있기 때문이다.

아동은 보통 이미지 형상화가 가능하고 의미가 구체적인 명사와 같은 내용어를 쉽게 습득(Morrison, Chappell & Ellis, 1997; Sandhofer, Smith & Luo, 2000)하지만 성인 말뭉치를 구성하는 어휘를 빈도에 따라 정렬하면 상위빈도는 대부분은 관사나 전치사와 같은 기능어이기 때문에 이들 어휘를 위주로 아동이 학습할 기초 어휘를 선정하는 것은 현실적으로 문제가 많다. Nation & Webb(2011)은 이러한 문제를 해결하기 위해 기초 어휘를 위한 말뭉치를 별도로 구성하였다. 즉 총 25,000개의 어휘군 목록을 개발하면서 기초 핵심 어휘라고 일컫는 상위 2,000개의 어휘 선정에는 구어 자료나 일상적인 내용이 풍부한 문어 자료로 말뭉치를 구축하고 3,000~25,000 단어 즉 23,000개의 어휘군의 선정에는 영국의 대표 말뭉치 BNC와 미국의 대표 말뭉치 COCA를 합한 초대형 말뭉치(mega-corpus)에서 추출한 어휘 정보를 사용하였다.

여기서 상위 2,000 단어 선정을 위해 구축된 말뭉치는 구어 자료가 6, 문어 자료가 4의 비율로 구어 자료의 비율이 더 많고 총 말뭉치의 규모는 1,000만 단어에 이른다

<표 19> BNC-COCA 상위 2,000 단어 선정을 위한 말뭉치 구성

미국 자료		영국/뉴질랜드 자료	
구어	규모(단어 수)	구어	규모(단어 수)
1. AmNC spoken face to face, telephone 1	1,107,602	4. BNC 1	1,036,097
2. AmNC spoken face to face, telephone 2	1,029,831	5. BNC 2	1,125,523
3. Movies and TV	1,000,000	6. BNC Plus half of WSC	1,132,620
문어	규모(단어 수)	문어	규모(단어 수)
7. AmNC written fiction, letters 1	1,145,081	9. School journals	1,028,842
8. AmNC written fiction, letters 2	939,407	10. BNC fiction	1,040,204

위의 <표 19>에서 보듯 BNC-COCA 상위 2,000 단어 선정을 위한 말뭉치 구성에는 총 10개의 하위 말뭉치가 사용되었고 각 하위 말뭉치의 규모는 약 100만 단어로 균일한 것을 알 수 있다. 미국과 영국<sup>7)</sup>의 자료 비율도 구어나 문어 모두에서 5:5로 균형을 이루고 있다. 문어의 경우에도 학습자들의 학교 과제물을 바탕으로 구성된 말뭉치를 포함하여 기초 수준을 유지하고자 하였다. 이를 통해 기존 상위 1,000 단어 수준에서는 볼 수 없었던 ‘alright, pardon, hello, dad, bye’ 등의 일상적인 단어들이 상위 1,000 단어에 포함될 수 있게 되었다.

끝으로 동일한 문제를 해결하기 위해 이와는 다소 다른 방식을 채택한 2015 영어과 교육과정 기본 어휘 목록 개발의 사례도 참조된다.

2015 영어과 교육과정 기본 어휘는 영어의 다양성을 인정하되 기본 어휘의 경우 유용성과 일반성을 강조하여 미국과 영국, 호주와 뉴질랜드 말뭉치를 모두 포함하는 구성을 채택하였고 대입시험을 준비하는 과정까지 활용할 수 있도록 문어의 비율을 높이면서도 구어에 포함된 일상적인 표현을 최대한 선정하기 위해 구어와 문어의 비율을 5:3으로 설정하였다.

<표 20> 2015 영어과 교육과정 기본 어휘 목록 개발을 위한 말뭉치 구성

말뭉치 종류	문어/구어	말뭉치 규모
Freiburg-Brown Corpus	미국 문어	100만 단어
Freiburg-LOB Corpus	영국 문어	100만 단어
Australian Corpus of English	호주 문어	100만 단어
British National Corpus Written Sampler	영국 문어	100만 단어
Wellington Written Corpus	뉴질랜드 문어	100만 단어
Corpus of Contemporary American English Spoken Sampler	미국 구어	100만 단어
British National Corpus Spoken Sampler	영국 구어	100만 단어
Wellington Spoken Corpus	뉴질랜드 구어	100만 단어

<표 20>에서 보듯 모든 하위 말뭉치의 규모는 100만 단어로 동일하며 총 800만 단어로 구성된 말뭉치를 구성하였다. 그러나 이러한 말뭉치 구성이 BNC-COCA 25000의 개발 사례와 마찬가지로 아동을 위한 기초 어휘 선정의 모든 문제를 해결할 수는 없다. 따라서 이러한 기초 어휘 선정을 위해 2015 영어과 교육과정 기본 어휘 선정에는 빈도와 사용 범위와 같은 객관적인 기준 외에도 친숙도와 초등 추천 어휘라는 기준을 추가 적용하였다. 하지만 원어민과의 의사소통 시 유용성을 고려하여 영어과 교육과정 어휘 3,000개의 선정에 앞서 빈도와 사용 범위라는 객관적

7) 뉴질랜드 영어는 영국 영어와 매우 유사하기 때문에 영국 자료와 동일하게 간주한다.

기준을 통해 3,500 단어를 먼저 선정하였다. 그리고 3,500개의 어휘군 풀 안에서 친숙도와 초등 추천 기준을 우선 적용하여 상위 3,000 단어 수준을 넘어서더라도 친숙도와 초등 추천 기준에서 높은 수치를 보여 그 순위가 3,000 단어 내에 포함되면 기본 어휘로 선정하였다(이문복·신동광, 2015). 이는 BNC-COCA 25000의 개발 시 이원화된 말뭉치 구축과는 차별되는 대안이라고 할 수 있다.

BNC-COCA 25000과 2015 영어과 교육과정 기본 어휘 목록 개발 사례는 초등과 중등 영어 어휘 더 나아가 성인 국어 기본 어휘 선정에 필요한 말뭉치 구축은 물론 어휘 선정 방식에 중요한 시사를 제공하고 있다.

## 2) 어휘 목록

어휘 목록의 기원은 기원전 3400-3300년 현재의 이라크 지역에 자리 잡았던 수메르인의 문자 체계로 거슬러 올라간다. 이 문명의 유물 중 쐐기문자가 세겨진 경판에 기록된 문자 중 15%가 어휘 목록이었다(Gnanadesikan, 2009: 15). 그리고 나머지 85%는 행정과 회계에 관련한 문자였다.

지금까지 알려진 최초의 핵심 어휘에 초점을 둔 어휘 목록은 16세기에 들어서면서 나타난다. 런던에서 불어를 가르치던 클라디우스 홀리밴드(Claudius Holyband)는 *The Schoolmaster*와 *The Littleton*이라는 대표적인 외국어 서적을 집필한다. 두 권의 책 모두에는 주제별로 정리된 상당한 양의 어휘 목록이 포함되어 있었다(Howatt & Widdowson, 2004: 27). 이 어휘들은 주로 상황 중심의 대화 내용에 초점을 두고 있었다. 이 책들은 현대의 언어 교재와 마찬가지로 학교, 가정, 여행, 비즈니스 등의 주제로 구성되어 있다. 비즈니스에 관련한 어휘는 어린 학습자에게는 부적절할 수 있다. 사실 Nation(1990: 20~21)은 이와 같은 직관적인 어휘 목록은 실제 빈도 기반으로 제작된 어휘 목록과 비교할 때 상당한 문제가 있다고 지적한 바 있다. 하지만 홀리밴드의 수업을 들던 학생들은 대부분 상인들의 자제들로 그의 책이나 어휘 목록은 수요자의 요구를 충족시켰다. 즉 ESP의 입장에서 수요자의 요구에 부합하는 어휘 목록이었다.

1588년에는 티모테 브라이트(Timothe Bright)가 속기에 관련한 도서를 영어로 출판하였는데 이 책에는 559개 어휘로 구성된 목록이 포함되어 있었다. 그리고 이 어휘들은 핵심 어휘의 의미에 초점을 두고 개발되었다. 웨스트 파머(West Palmer)와 오그덴(Ogden)은 후에 이러한 주요 의미를 커버하는 것을 핵심 어휘 선정 기준으로 제안하기도 하였다(Richards, 1974: 73~75).

17세기에는 혁신적인 라틴어 교사였던 요한 아모스 코메니우스(Jan Amos Comenius)가 외국어 학습에 있어 4단계의 교육과정을 개발·제안하였다(Howatt & Widdowson, 2004: 47). 코메니우스 교육과정 개발의 첫 번째 단계는 일상생활의

간단한 대화에서 활용하기에 충분한 몇 백 개의 어휘로 구성된 목록을 개발하여야 한다는 것이다. 두 번째 단계는 교육적으로 가치가 있고 학습자의 자발적인 관심을 이끌어 낼 수 있는 수준별 교재에 포함된 8,000개 정도의 어휘를 초점을 두고 지도해야 한다는 것이다. 세 번째와 네 번째 단계를 위한 교재는 스타일, 어법, 번역 등을 위한 것이었지만 코메니우스는 이 교재들을 완성하지는 못했다. 대신 그는 Orbis Sensualium Pictus라는 이중언어(bilingual) 그림 사전을 출간하였다. 요즘과는 달리 이러한 그림 사전은 이 시기에 매우 독창적인 발상이었다. 코메니우스는 이 교재를 토론 수업 중심으로 개발하고 그 핵심 어휘에 학습의 우선순위를 두었다. 그는 어휘를 나름대로 등급화하였고 가급적 불필요한 저빈도 어휘는 사용하지 않으려 했다. 이는 핵심적인 어휘에 먼저 초점을 두고 어휘 학습의 양을 몇 천 단어까지 확대해 나가는 현대의 어휘 교육과도 일맥상통한다.

1755년에는 사무엘 존슨(Samuel Johnson)이 ‘Dictionary of the English Language’를 제작하는데 이는 일종의 어휘 목록으로 이해될 수 있다. 그는 1750대의 실제적 문어 표현을 그대로 발췌하려고 하였다. 그 당시만 해도 구어적 표현은 단어의 발음정도만 정리하는데 머물고 있었다. 그는 말뭉치를 기반으로 하지는 않았지만 실제적 표현을 그대로 발췌함으로써 사전에 수록된 표현의 실제성을 확보하였다(Howatt & Widdowson, 2004: 110~116). 존슨의 사전은 45,000개의 기본형을 수록하였고 19세기에 이르러 1828년에는 노아 웹스터(Noah Webster)가 존슨의 사전을 70,000단어까지 확대 수록하였다.

이후 문법번역식 교수법이 18세기 후반부터 19세기 초 독일에서 주요 교수법으로 자리를 잡으면서 어휘 목록 활용도가 함께 증대되었다. 그러나 문법번역식 교수법에서는 어휘 선정에 있어 의사소통의 유용정보보다는 문법구조의 설명에 필요한 어휘를 보다 중요하게 고려하였다(Zimmerman, 1997: 6). 이에 따라 1870년대 Berlitz가 직접식 교수법을 전파하기 전까지 의사소통을 위한 구어 표현이나 어휘는 경시되는 경향을 보였다. 그러나 19세기 후반 토마스 프렌더개스트(Thomas Prendergast)는 광범위한 의사소통 기능을 커버할 수 있는 한정된 수의 어휘 목록을 구성하였다. 그가 구성한 214개의 핵심 어휘 중 82%는 Thorndike-Lorge(1944)의 어휘 목록 중 가장 빈도가 높은 어휘에 포함되어 있다. 나머지 14% 또한 상위빈도 1,000단어 안에 포함되어 있다(Howatt & Widdowson, 2004: 176). 토마스 프렌더개스트는 독어, 불어, 스페인어, 라틴어, 희랍어에도 동일한 기준을 적용하여 언어별 어휘 목록을 개발하였다.

그러나 19세기까지의 어휘 목록은 개발자의 직관에 바탕을 두고 개발되었다. 당시에는 여전히 어휘의 빈도를 산출하여 어휘를 선정하는 것이 쉽지 않았다. 실제 객관적인 기준을 적용하여 어휘를 선정하는 방법은 당시의 교육자들에게는 생각하기도 힘든 아이디어였다. 그럼에도 불구하고 몇몇 연구자들은 작은 규모이기는 하

지만 빈도를 산출하여 어휘를 선정하는 프로젝트를 시작하였다. 1837년에 피트먼(Pitman)은 속기 시스템을 개발하면서 10,000개의 어휘로 구성된 목록을 만들었고 바로 직전인 1820년에는 프리먼(Freeman)이 성인 학습자들 위한 20,000개의 단어로 구성된 어휘 목록을 말뭉치를 기반으로 개발하였다(McArthur, 1998: 51~52). 1904년, 영국에서는 놀스(Knowles)가 맹인들을 위한 읽기 시스템을 개발하기 위해 10만 개 단어의 어휘 목록을 구축하였고 이 중 353개 단어는 전체 영어 읽기의 75%를 커버한다고 주장하기도 하였다(McArthur, 1998: 52).

19세기말에는 교육개혁운동과 함께 언어교육에서 구어교육의 중요성이 증대되었다. 이는 앞에서 언급한 듣기, 말하기와 같은 기본적인 의사소통에 초점을 둔 직접식 교수법 출현의 계기가 되었다(Schmitt, 2000: 12). 1902년 벨기에에서 직접식 교수법에 기반한 영어교육을 시작하면서 영어교육의 효율성을 위해 빈도가 가장 높은 어휘를 먼저 학습하는 방법이 헤롤드 파머(Herold Palmer)에 의해 제안되었다(Palmer, 1936: 371). 하지만 어휘 선별 시 활용되는 말뭉치 구성의 불균형 또는 주관성의 논란으로 인해 그의 어휘 선정은 비판을 받기도 하였다. 이러한 문제는 20세기 Thorndike & Lorge(1944)의 어휘집 The Teacher's Wordbook of 30,000 Words 개정판에서도 여전히 문제로 지적되었다.

1930년에 오그덴은 소위 British American Scientific International Commercial(BASIC) English라는 850개로 구성된 어휘 목록을 출간하였다. 오그덴은 BASIC이 국제적인 의사소통의 수단으로 그 역할을 충분히 수행할 수 있다고 주장하였다. 그러나 West, Swenson, Fawkes, Russell과 de Magellanes Wilf(1934)는 BASIC에 대하여 오그덴은 BASIC이 일주일 또는 최대 한 달이면 모두 학습할 수 있는 제한된 수의 어휘 목록이라고 강조했지만 그것은 어휘 목록에 포함된 많은 다의어를 고려한지 않은 것이라고 지적하였다. 실제 Fries & Traver(1960)의 분석에 따르면 오그덴의 BASIC은 850개의 어휘로 구성되어 있지만 그 어휘들은 무려 12,425개의 다양한 의미를 지닌 것으로 나타났다. 따라서 850개의 어휘 형태는 한 달 안에 학습할 수 있을지는 모르지만 그 어휘의 모든 의미를 그 시간 안에 학습하는 것은 불가능하다고 주장하였다.

20세기 중후반에 이르러 말뭉치 분석을 활용한 어휘 목록 개발의 시도가 지속적으로 나타난다. 기존의 시행착오를 거쳐 1953년에는 드디어 웨스트에 의해 영어권 말뭉치 기반 어휘 목록의 고전이라고 할 수 있는 가장 대표적인 어휘 목록인 General Service List(GSL)이 개발되었다. 다음의 <표 21>은 GSL을 포함하여 최근까지 개발된 대표적인 말뭉치 분석 기반 영어 어휘 목록을 정리한 것이다.

&lt;표 21&gt; 말뭉치 분석 기반 영어 어휘 목록

어휘 목록/개발자(연도)	어휘 목록 규모 및 단위	어휘 선정 기준	활용 말뭉치
GSL(General Service List)/ West(1953)	2,000 어휘군	빈도	-
AWL/Coxhead(1998, 2000)	570 어휘군	빈도, 사용범위, 산포도	Academic Corpus
BNC 14000/Nation(2004)	14,000 어휘군	사용범위, 빈도, 산포도	BNC Spoken
BNC-COCA 25000/Nation & Webb(2011)	25,000 어휘군	빈도, 사용범위	BNC COCA
NGSL(New General Service List)/Browne(2013)	2,818 사전 등재형	산포도, 표준 빈도 지표	CEC
NGSL/Brezina & Gablasova(2013)	2,494 사전 등재형	빈도, 평균 압축 빈도	LOB BNC BE06 EnTenTen12

GSL은 최초로 말뭉치의 어휘 빈도를 체계적으로 분석하여 개발한 어휘 목록이다. 총 2,000단어로 구성된 GSL은 Thorndike & Lorge(1944)가 분석한 말뭉치의 어휘 빈도 정보를 활용하여 개발하였으며 2009 영어과 교육과정 기본 어휘까지 우리나라 영어 교육과정 기본 어휘의 주축을 이룰 정도로 오랜 역사와 함께 가장 보편적으로 활용되던 어휘 목록이다. 일반적으로 GSL을 기준으로 할 때 영어의 2,000단어(어휘군)는 문어에서 80%, 구어에서 90% 이상을 커버하는 학습 효율성을 담보하는 핵심 어휘라고 말한다(Nation, 2001).

이러한 이유로 최근까지도 이 GSL을 개정하려는 여러 시도가 있었다. Dickins(n.d.)는 웨스트의 GSL를 바탕으로 Longman Lexicon of Contemporary English(LLCE, McArthur, 1981)의 의미 분류 기준을 적용하였다. 의미의 분류는 품사변화를 기준(사전등재형 기준, 문법적 변화형을 모두 포괄하는 대표형)으로 하였으며 [그림 5]과 같이 2,000단어를 각 세부 의미로 나누고 그 빈도 비율의 정보를 제시하였다.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
7	able	[Y]	able	x	940	940	100%	940	GenSerList	X				
8	able	[G042][N2	able	adj	940	940	89%	836	GenSerList	(1) Able to take care of				
9	able	[G042][N2	able	adj	940	940	11%	103	GenSerList	(2) An able man				
10	able	[M163]	able	adj	940	940	0.10%	1	GenSerList	Able-bodied				
11	able	[G043]	ability	x	340	340	100%	340	GenSerList	X				
12	able	[G043]	ability	n	340	340	58%	197	GenSerList	(1) (cleverness) A man of great ability. Musical ability				
13	able	[G043]	ability	n	340	340	40%	136	GenSerList	(2) (power to) Ability to learn				
14	able	[G043]	ability	n	340	340	2%	6	GenSerList	legal sense				

[그림 5] Dickins의 GSL 예시

[그림 5]에서 보듯 'able [Y]'은 어휘군의 표제어(기본형, 대표형)를 의미하고 이 어휘군은 'able x'과 'ability x'라는 사전등재형(lemma)으로 구분되며 각 어휘의 전체 사용정도를 100%라고 할 때 세부 의미별 사용비율은 89%와 11% 그리고 58%, 40%, 2%라는 것을 의미한다. K열은 각 분류 항목에 따른 세부 의미를 제시하고 있다.

이 밖에도 소위 New General Service List(NGSL)라는 이름으로 두 개의 어휘 목록이 추가로 개발되었다. 2013년 브라운(Browne)는 Cambridge English Corpus(CEC)의 하위 말뭉치에서 여러 영역을 균형 있게 선별하여 총 2억 7천만 단어의 말뭉치에서 2,818개의 사전등재형을 추출하고 이를 NGSL이라 명명하였다. 이 NGSL 개발에는 여러 말뭉치의 크기를 균일하게 하여 특정 영역의 말뭉치가 가지는 언어적 특성을 최대한 배제하려고 노력하였다. 하지만 언어적으로 오류가 많은 학습자 말뭉치를 어휘 목록 개발에 포함한 것은 보편타당한 기초 어휘를 선정하고자 한 NGSL의 개발 취지에는 부적절하다고 판단된다.

NGSL의 어휘 선정에는 100만 단어별 어휘의 추정 빈도(/million) 즉, 산포도와 Carroll, Davies & Richman(1971)이 제시한 표준 빈도 지표(Standard Frequency Index)를 적용하였고 이후 고유명사, 약어, 은어, 요일, 달, 숫자 등은 제외시킨 뒤, 전문가 그룹과의 토의를 거쳐 기존의 주요 어휘 목록과 비교하여 추가할 어휘와 제외할 어휘를 최종 결정하였다.

같은 해인 2013년에는 또 다른 NGSL이 브레지나(Brezina)와 가블라스바(Gablasova)에 의해 개발되었다. 이 NGSL의 개발에는 4개의 말뭉치가 사용되었고 이들 말뭉치의 총 크기는 121억 단어의 말뭉치 자료가 사용되었다. 전체 말뭉치에서 구어 자료의 양은 미미하여 문어적인 색채가 다소 짙고 어휘 선정의 기준에는 4개의 말뭉치에 각각 나타는 품사별 어휘의 빈도와 평균 압축 빈도(Average Reduced Frequency, ARF)<sup>8)</sup>를 사용하였다. ARF는 크기가 다양한 말뭉치에서 어휘의 보편성을 고려하여 어휘를 선정할 때 효과적인 기준이다. 그리고 어휘 목록 개발에서 고유명사와 알파벳은 제외되었다.

지금까지는 GSL에 기반한 어휘 목록을 살펴보고 2,000단어 수준을 넘어서는 대규모의 어휘 목록으로는 Nation(2004)의 BNC 14000을 들 수 있다. 1억 단어로 구성된 British National Corpus(BNC)의 10%는 구어 자료인바, 네이션은 1000만 단어의 구어 말뭉치에서 14,000개의 어휘군<sup>9)</sup>을 추출하였다. 그리고 이를 1,000개의 어휘군 단위로 나누어 14개 등급으로 세분화하여 제시하였다.

Nation(2001)에 따르면 영어 원어민 화자는 자연스러운 언어 노출을 통해 1년에

8) ARF는 특정 단어의 출현 빈도로 말뭉치를 구성하는 단어 수를 나누고 그 값을 구간 수로 하여 그 구간 중 그 특정 단어를 포함한 구간의 숫자를 산출한 후 그 평균값을 구한 수치이다. 이는 비교하고자 하는 말뭉치 간 규모의 차이가 클 때 사용되는 절대 빈도이다(Hlaváčová, 2006).

9) [http://www.victoria.ac.nz/lals/resources/range\\_BNC.zip](http://www.victoria.ac.nz/lals/resources/range_BNC.zip)에서 다운로드 가능하다.

약 1,000개의 어휘군을 습득한다고 한다. 이러한 이유로 어휘 목록은 1,000개의 단어를 한 개 등급으로 구분되었으며 한 개의 등급은 동일한 수준으로 간주된다. BNC 14000의 특징으로는 국가명과 그 나라의 국민이나 언어를 가리키는 어휘는 어휘 목록에 포함되어 있으나 사람의 고유한 이름과 도시 이름, 로마자, 감탄사 등은 예외적인 어휘 목록으로 14,000단어와는 별도로 구성하였다. 이 어휘 목록이 주목받는 주요 이유 중 하나는 어휘 목록 개발에 가장 우선하여 적용된 어휘 선정 기준이 빈도가 아니라 여러 말뭉치를 적용할 경우 하나의 어휘가 얼마나 다양한 말뭉치에서 사용되고 있는지를 측정하는 수치인 사용 범위이기 때문이다.

대부분의 어휘 목록은 빈도를 첫 번째 어휘 선정 기준으로 적용한다(Nation, 2001). 그것은 빈도가 가지는 노출의 기회와 높은 빈도가 말해주는 유용성에 대한 전통적인 믿음에서 비롯되었다. 실제로, 빈도를 최우선 순위로 두었을 때 텍스트 포괄 범위는 가장 높게 나타난다(신동광, 2011). 그럼에도 불구하고 네이션은 BNC 14000 어휘 목록 개발 시에 한 어휘가 얼마나 다양한 텍스트에서 사용되는지를 측정하는 사용범위를 총 사용 빈도보다 우선하여 적용하였고 마지막으로 한 단어가 각 텍스트에서 얼마나 일정 빈도 이상으로 사용되는 지를 측정하는 산포도(dispersion, spread frequency)를 추가하여 적용하였다. 이는 단순한 텍스트 포괄 범위의 수치보다는 특정 영역에 국한되지 않고 보편적으로 사용될 수 있는 어휘를 우선하여 선정하고자 했기 때문이다.

하지만 네이션은 BNC 14000이 영국 말뭉치만을 사용하여 개발하였기 때문에 미국 영어와는 다소 차이가 있다는 것을 인식하고 있었다. 이러한 BNC 14000의 한계를 극복하기 위해 2011년 네이션과 웹(Webb)은 영국 영어를 대표하는 1억 단어의 BNC와 미국 영어를 대표하는 개발 당시 4억 5천 단어의 COCA에 나타난 어휘 빈도와 분포 통계를 통합하여 25,000개의 어휘군 목록<sup>10)</sup>을 개발하였다. 이는 미국 영어와 영국 영어 자료를 모두 적용하여 개발하였다는 점에서 의미가 있다. 그리고 1,000단어 단위의 등급으로 구성된 25개의 어휘 목록 중 상위 2개 등급 즉, 상위 2,000개의 어휘군은 BNC-COCA가 아닌 별도의 말뭉치를 구축하여 개발하였다.

GSL의 소개에서도 언급한 바와 상위 2,000단어는 학습의 기본이 되기 때문에 중요한 의미를 가진다. 하지만 두 개의 대규모 말뭉치를 합친다는 것이 기초 어휘의 대표성을 보장하는 것은 아니다. 좀 더 기초적인 일상적인 2,000개의 어휘를 산출하기 위해 네이션과 웹은 이 2,000개의 어휘 목록 개발을 위해 구어와 문어를 6:4로 구성한 총 1,000만 단어의 말뭉치를 새로 구축한다. 이 말뭉치에는 전화표현이나 영화 및 TV와 같이 일반적이고 쉬운 구어 자료를 문어 자료에 비해 상대적으로 더 많이 포함시켰다. 또한 문어 자료의 경우에도 아이들을 위한 쉬운 언어 자료를

10) [http://www.victoria.ac.nz/lals/about/staff/publications/BNC\\_COCA\\_25000.zip](http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip)에서 다운로드가 가능하다.



포함시켰다. 따라서 순수한 BNC-COCA 어휘는 3,000단어 수준부터라고 할 수 있다. 또한 BNC-COCA 25000은 25,000단어 외에 3,000개의 합성어, 로마자, 고유명사 등을 별도의 어휘 목록으로 구성하였다. 결과적으로 BNC-COCA 25000은 가장 최근에 개발된 대규모의 등급화된 일반 영어 어휘 목록이라고 할 수 있다.

일반적인 영어 어휘의 사용을 대표하는 목록이 BNC-COCA 25000이라면 특수 목적을 위해 개발된, 특히 학술 목적을 위해 개발된 어휘 목록의 대표는 Coxhead(1998, 2000)가 개발한 Academic Word List(AWL)를 들 수 있다. 콕스헤드(Coxhead)는 AWL을 개발하기 위해 약 3백 5천만 단어의 Academic Corpus를 구축하였다. Academic Corpus는 크게 4개 대영역(Arts, Commerce, Law, Science)으로 세분화되고 각 대영역은 7개의 대학전공(예-Arts: Education, History, Linguistics, Philosophy, Politics, Psychology, Sociology)으로 다시 구분된다. 즉 AWL 개발에는 28개의 대학 전공 자료가 사용되었다. 이 자료는 학술지 저널, 전공 교재, 실험 매뉴얼, 학생 노트 등을 포함하고 있다. 어휘 선정 기준으로 는 빈도, 사용 범위, 산포도를 적용하였다.

먼저 일반 영어 어휘를 제외하기 위해 GSL에 포함된 최상위 2,000단어는 목록에서 삭제한 후 Academic Corpus에서 나타나는 최상위 빈도의 단어를 추출하였다. 또한 AWL은 Academic Corpus에 포함된 28개의 대학전공 중 최소 15개의 교과목에서 사용되는 어휘이며 끝으로 4개의 대영역 모두에서 각 10번 이상의 빈도를 가지는 단어라는 조건을 충족해야 한다. 이러한 과정을 통해 콕스헤드는 570개 어휘군 목록을 개발하였다. Nation(2001)은 GSL이 일상생활의 기초 영어 학습에 유용한 반면 AWL은 학술적인 영역을 보완한다는 면에서 이상적인 조합이라고 말한다.

### 3) 어휘 학습량 및 어휘 등급화 단위 설정

어휘 목록 개발에서는 지금까지 살펴본 여러 요인 외에도 크게 두 가지 핵심 요인을 고려할 필요가 있다. 첫 번째는 원어민 또는 외국인 학습자의 입장에서 도달해야 할 학습의 목표, 다시 말해 총 어휘 학습량을 어느 정도로 설정하고 어휘 목록을 개발한 것인가의 문제이다. 두 번째는 어휘 목록을 여러 등급으로 세분화할 경우 등급의 규모를 어느 정도로 설정해야 등급별 적당한 학습량이 될 것인가에 대한 문제이다. 다음의 두 절에서는 이러한 두 가지 요인을 고찰해 볼 것이다.

#### (1) 어휘 학습량 설정

현재까지 체계적인 연구가 부족하기는 하지만 일반적으로 20세까지 적절한 교육을 받은(well-educated) 영어 원어민 화자는 20,000개의 어휘군에 대한 어휘지식

을 가지고 있다고 한다(Nation, 2001). 하지만 원어민 화자도 20여년에 걸쳐 원어민의 언어 환경에서 습득할 수 있는 20,000 단어를 학습 목표로 설정한 사례는 찾아보기 힘들며 대개는 언어 학습에 있어 어떤 목적을 가지고 있는가가 중요한 학습 목표 설정의 기준이 된다.

Hu & Nation(2000)은 언어사용에 있어 최소한으로 알아야 할 기본 어휘량을 텍스트의 80%로 보고 있으며 98% 이상을 알아야 사전(dictionary)과 같은 외부의 어떠한 자료를 참고하지 않고도 문맥에서 단어의 뜻을 유추하며 독립적으로 글을 읽을 수 있다고 주장한다.

Francis & Kucera(1982)는 2,000개의 어휘가 문어 자료(written text)에서 약 80%를 커버한다고 하였고 Schonell, Meddleton과 Shaw(1956)는 2,000 단어가 비공식적인 구어 자료(informal spoken text)에서 무려 96%를 포괄할 수 있다고 주장하기도 하였다. Nation(2006)은 British National Corpus(BNC)에 기반한 14개 등급 총 14,000개 어휘 목록을 활용하여 다양한 자료의 텍스트 포괄 범위를 분석하였다.

예를 들어 “슈렉(Shreck)”이라는 애니메이션 대본을 분석하면 텍스트의 98% 이상을 커버하기 위해서는 고유명사를 제외하고 총 7,000 단어 정도가 필요하고 일반 소설에서는 8,000~9,000 단어의 지식이 요구된다는 결과를 보여주었다. 이뿐만 아니라 Nation은 일반적인 구어 자료에서는 98%의 텍스트 포괄 범위를 확보하기 위해 6,000~7,000 단어가 필요하다고 주장하기도 하였다.

아래 표에서 보듯 98%의 텍스트 포괄 범위를 안정적으로 확보하기 위해서는 문어 자료에서 8,000~9,000개의 어휘가 필요하고 구어에서는 6,000~7,000개의 어휘를 요구된다는 것을 알 수 있다.

<표 22> 문자 및 음성 언어에서의 텍스트 포괄 범위 비교(Nation, 2006: 79)

어휘 등급	등급 수	텍스트(구어) 포괄 범위 (%)	텍스트(문어) 포괄 범위 (%)
1st 1,000	1	78-81	81-84
2nd 1,000	1	8-9	5-6
3rd 1,000	1	3-5	2-3
4th-5th 1,000	2	3	1.5-3
6th-9th 1,000	4	2	0.75-1
10th-14th 1,000	5	< 1	0.5
고유명사	1	2-4	1-1.5
그 외	1	1-3	1

하지만 위의 분석은 영어권 원어민 화자의 언어사용 데이터에 기반한 결과이며 영어를 외국어로 학습하는 한국인의 기준에는 도달하기 쉽지 않은 학습량이라고 볼

수 있다. 현재 보편적으로 제안되는 텍스트 포괄 범위는 95%이다(Read, 2004). 위의 표를 기준으로 하면 5,000 단어(어휘군) 수준으로 추정할 수 있다(구어: 92~98%, 문어: 89.5~96%). Laufer(1992) 또한 영어로 강의가 진행되는 대학교육에서 요구되는 어휘 수는 5,000개로 대학교재의 개발이나 일상적인 의사소통, 학문적인 과제를 해결하기 위해 반드시 필요하다고 주장하였다.

그러나 우리나라 영어과 교육과정의 기본 어휘 수는 이보다도 적은 3,000 단어로 설정되어 있기도 하다. Nation(2001)에 따르면 영어를 외국어로 배우는 학습자가 일주일에 매 50분으로 구성된 4시간의 영어 수업을 5~6년 간 받을 경우 약 1,000 시간 안팎의 영어 노출 시간이 주어지고 1,000 시간의 영어 수업을 통해 습득될 수 있는 어휘는 약 3,000 단어 정도라고 한다. 이러한 주장에 근거한다면 우리나라 영어과 교육과정에서 설정한 3,000 단어 정도의 기본 어휘는 시간대비 학습량으로는 적당하다고도 볼 수 있다.

위의 사례를 국어 어휘에 그대로 적용해 보면 한국어 원어민인 우리나라 국민의 어휘 학습 목표는 어휘군 기준으로 9,000 단어, 외국인 학습자는 3000~5000 단어 정도가 될 수 있을 것이다. 국어에서는 파생과 굴절형을 모두 포함하는 어휘군 단위가 아닌 보통 품사가 다르면 별개의 어휘로 산정하는 사전등재형을 어휘 단위로 보고 일반적으로 사용하기 때문에 위의 수치는 사전등재형 수로 재산정할 필요가 있다. 대략 한 개 어휘군이 평균적으로 4개의 품사 즉, 명사, 동사, 형용사, 부사로 구성되어 있다고 가정하고 4배의 수를 산출하면 우리나라 국민의 어휘 학습 목표는 28,000개의 사전등재형, 외국인 학습자는 12,000~20,000개의 사전등재형을 학습해야 구어와 문어를 포함하여 98% 이상의 텍스트 포괄 범위를 확보할 수 있을 것으로 추정된다. 다만 이 또한 대표성이 있는 국어 말뭉치의 텍스트 포괄 범위(95~98%)를 분석해야만 정확한 어휘 학습량 설정이 가능할 것이다.

### (3) 어휘 등급화 단위 설정

최근에 개발된 영어 어휘 목록(예, BNC-COCA 25000)의 등급은 1,000개 (어휘군)를 한 등급의 단위로 설정하고 있다. 그 이유는 영어 원어민의 경우, 자연스러운 언어 노출에 따라 1년에 약 1,000개 단어를 습득한다고 보고 이 1,000개 단어는 동일한 수준으로 간주하기 때문이다(Nation, 2001). 하지만 연간 어휘 습득량과 교육목적으로 한 어휘 등급 규모가 항상 일치하는 것은 아니다. 이는 개인별 발달 차이나 여러 변인들을 고려해야 하기 때문에 실제 학습을 목표로 하는 어휘 등급의 규모는 실제 이론으로 제시된 기준보다는 작게 설정되는 것이 일반적이다.

대체로 학습자의 연령이 낮을수록 학습량은 적게 설정한다. Stenach & Williams(1988)은 기존에 구축된 다양한 아동 말뭉치 연구(Hopkins, 1979;

Murphy, 1957; Moe, Hopkins & Rush, 1982; Wepman & Hass, 1969) 데이터를 활용하여 2,500개의 아동용 어휘 목록을 개발하였다. 이 어휘 목록은 어휘의 빈도 정보를 바탕으로 10개의 등급으로 구분되었고 각 등급은 250개의 어휘로 구성되어 있다. 스테마치(Stemach)와 윌리엄스(Williams)는 이중 상위 2개 등급 즉 500개의 어휘는 초등 1학년 학생의 구어 사용에서 85%를 차지한다고 주장하였다. 이는 현재 우리나라 초등 영어과 교육과정에서 제시하고 있는 교과서 개발 어휘 사용지침과도 거의 일치한다. Gyllstad, Vilkaite & Schmitt(2015) 또한 실제 출판사들의 수준별 교재(graded reader)의 등급 구분이 500개 미만의 어휘 단위로 되어 있으며 이는 학습 수준을 구분 짓는 어휘 등급의 단위로서 1,000개의 어휘는 너무 많다는 점을 지적하였다. Gyllstad, Vilkaite & Schmitt(2015)에 따르면 Cambridge Discovery Reader 시리즈의 경우, Starter: 250개 어휘, Grade 1: 400개 어휘, Grade 2, 800개 어휘, Grade 3: 1,300개 어휘, Grade 4: 1,900개 어휘, Grade 5: 2,800개 어휘, Grade 6: 3,800개 어휘로 등급을 구분하고 있다.

우리나라의 2015 영어과 개정 교육과정 기본 어휘는 크게 3개 등급으로 구분되어 있다. 그 3개의 등급은 800개의 초등 권장 어휘, 1,800개의 중등 일반선택 권장 어휘(누적 2,600개), 400개의 진로·전문교과 I 권장 어휘로 총 3,000개의 어휘가 기본 어휘로 설정되어 있으며 학년별 또는 학년군별로 다음의 표와 같이 사용 어휘 수를 통제하고 있다(이문복·신동광, 2015). 또한 교과서 개발 시에는 지정된 어휘 제한 수에 따라 3개의 등급의 권장 어휘풀에서 해당 등급 어휘의 90% 이상을 사용하도록 규제하고 있다. 다만 고등학교 진로선택 및 전문교과 I 부터는 학년별 권장 어휘 중 80% 이상을 사용하여 교과서를 개발하도록 일부 완화하여 적용하고 있다.

<표 23> 2015 영어과 교육과정 과목별 어휘 수

학년군/과목명		어휘 수	
초등학교	초등학교 3~4 학년군	240 내외	
	초등학교 5~6 학년군	500 내외	
중학교	중학교 1~3 학년군	1,250 내외	
고등학교	공통	영어	1,800 내외
		영어 회화	1,500 이내
	일반선택	영어	2,000 이내
		영어 독해와 작문	2,200 이내
		영어 II	2,500 이내
	진로선택	실용 영어	2,000 이내
		영어권문화	2,200 이내
		진로 영어	2,500 이내
		영미 문학 읽기	3,000 이내
	전문교과 I	심화 영어 회화 I	1,800 이내
		심화 영어 회화 II	2,000 이내
		심화 영어 I	2,500 이내
		심화 영어 II	2,800 이내
		심화 영어 독해 I	3,300 이내
		심화 영어 독해 II	3,500 이내
		심화 영어 작문 I	2,000 이내
		심화 영어 작문 II	2,300 이내

위의 표에서 볼 수 있듯이 교과서 간의 어휘 차이는 500개 미만으로 Gyllstad, Vilkaitė & Schmitt(2015)의 조사와 일치하는 것을 확인할 수 있다.

끝으로 다독재단(Extensive Reading Foundation, ERF)에서 제공하는 다독 교재의 어휘 등급을 보면 다음과 같다(ERF, n.d.).



## The Extensive Reading Foundation Grading Scale

Beginner				Elementary			Intermediate			Upper Intermediate			Advanced			Bridge			Near Native
Alphabet	Early	Mid	High	Early	Mid	High	Early	Mid	High	Early	Mid	High	Early	Mid	High	Early	Mid	High	
1	51	101	201	301	401	601	801	1001	1251	1501	1801	2101	2401	3001	3601	4501	6001	8001	12001-18000 and above
50	100	200	300	400	600	800	1000	1250	1500	1800	2100	2400	3000	3600	4500	6000	8000	12000	

This scale is only for approximate leveling of Language Learner Literature by headword<sup>1</sup> counts by series. Some individual titles may need to move up or down as necessary. A list of where each publisher's Graded Reader series fits this scale is available at the website.

[그림 6] ERF의 어휘 등급 척도

ERF는 영어 원어민 화자의 수준을 최상위 수준의 단계로 하여 총 7개의 등급으로 구분하고 각 등급을 다시 수준별로 3~4개(예, Early, Mid, High)까지 세분화하고 있다. 7개 등급의 총 누적 어휘 수는 300 - 800 - 1,500 - 2,400 - 4,500 - 12,000 - 18,000이며 등급 간 어휘 수 증가는 500 - 700 - 900 - 2,100 - 7,500 - 6,000로 점진적으로 확대되는 것을 볼 수 있다.

지금까지 살펴본 어휘 등급 규모의 사례를 종합하면 250~500개의 어휘가 현실적인 어휘 등급의 단위로 적당하다고 판단되며 사전등재형으로 어휘 단위를 변환하여 추정하면 500개 어휘군을 기준으로 할 때 2,000개의 사전등재형, 250개의 어휘군으로 기준으로 하면 1,000개의 사전등재형으로 등급을 구분할 수 있을 것으로 판단된다. 그러나 실제 우리나라 국민의 어휘 지식의 규모를 연령별로 측정해야 연간 한국어 원어민 학습자의 어휘 습득량을 확인할 수 있고 이를 바탕으로 최종적인 어휘 등급의 규모 등도 확정될 것이다.

### Ⅲ. 말뭉치 구축

이 장에서는 기초 어휘 선정 및 어휘 등급화에 사용되는 어휘 목록을 추출하기 위한 말뭉치의 고려 사항을 제시하고, 현재 구축한 현황 및 보완 계획을 제시한다. 먼저, 말뭉치를 실제 구축하는 과정에서 얻게 된 실험 결과를 토대로 장르에 대한 고려와 시간/연대에 대한 고려 사항을 파악하였다. 다음으로, 당해 연도에 구축한 말뭉치의 현황을 보이고 이 말뭉치의 어휘 목록 추출의 신뢰도와 타당도를 높이기 위한 보완 계획을 마련하였다.

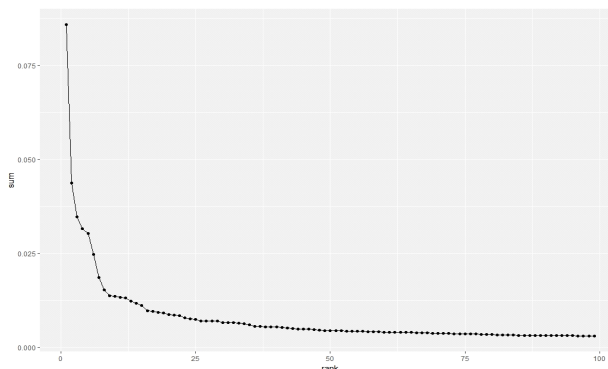
#### 1. 장르에 대한 고려

기초 어휘에 포함되는 단어는 다음의 3가지 조건을 충족시켜야 한다.

- ① 빈도: 자주 쓰여야 한다.
- ② 범위: 장르 면에서 널리 쓰여야 한다. 다양한 장르에서 쓰여야 한다.
- ③ 산포도: 장르 면에서 골고루 쓰여야 한다. 특정 장르에만 편중되지 않고, 여러 장르에서 고른 빈도를 보여야 한다.

①의 조건을 충족시키는 단어들을 알아내기 위해서는 대규모 말뭉치가 필요하다. 단어뿐 아니라 어떤 것이든 통계적으로 신뢰할 만한 수치를 얻을 수 있으려면, 연구 대상이 되는 사건이 일정 수준 이상의 빈도로 출현해야 한다. 너무 낮은 빈도는 통계적으로 신뢰할 수 없다.

문제는 단어의 빈도 분포가 극도로 편중되어 있어서, 극소수의 단어가 말뭉치의 대부분을 차지하고, 나머지 단어들은 매우 낮은 빈도로만 나타난다는 사실이다. 본 사업에서 사용한 말뭉치에서 빈도 1위부터 100위까지의 단어에 대해, 순위를 가로축에, 상대빈도를 세로축에 표시하면 다음 그래프와 같이 된다.



[그림 7] 순위-빈도 그래프

1위에 비해 2위는 빈도 값이 거의 절반으로 떨어지고, 순위가 내려갈수록 비슷한 식으로 빈도가 급격히 낮아져서, 금세 수평선에 가까운 형태가 된다. 어느 말뭉치는 순위-빈도 그래프를 그려 보면 대개 이와 비슷한 모양이 된다. 이것은 Zipf의 법칙으로 잘 알려져 있다.

이렇게 말뭉치에 나타나는 단어의 대다수가 저빈도어이므로, 이런 저빈도어에 대해서도 통계적으로 신뢰할 만한 수치를 얻기 위해서는 말뭉치의 규모가 엄청나게 커야 한다. 그러려면 말뭉치의 규모가 도대체 어느 정도 되어야 하는가 하는 질문에 대해서는 아직 뚜렷한 정답이 나와 있지 않고 연구 목적에 따라 대답이 달라질 수 있으나, 1억 어절 정도의 말뭉치로도 부족하다는 것이 중론이다. 최근 연구 경향은 최소한 10억 어절 이상의 말뭉치를 사용하는 추세이다.

그런데 말뭉치의 크기가 단순히 큰 것만으로는 부족하고, ②와 ③의 조건을 충족시키는 단어들을 알아내기 위해서는 말뭉치가 다양한 장르로 구성되어 있어야 한다. 장르를 몇 개로 어떻게 나눌 것인가에 대해서도 논란이 많고 아직 정설이 확립되어 있지는 않다.

이상적으로는, 전국민의 언어 사용 실태를 조사하여, 일상생활에서 면대면 대화가 몇 퍼센트, 일방향적인 강의·연설 등을 듣는 것이 몇 퍼센트, 신문을 읽는 것이 몇 퍼센트, 소설을 읽는 것이 몇 퍼센트, 텔레비전 드라마를 시청하는 것이 몇 퍼센트, …… 하는 식으로 장르의 비중을 조사하여 이에 맞게 말뭉치의 장르의 비중을 정하면 좋을 것이다. 그러나 그러한 조사가 제대로 시행되지 않았을 뿐더러, 그런 장르 비중을 객관적으로 조사할 수 있는지조차도 의문이다.

이런 상황에서 말뭉치의 장르 비율이 어떠해야 한다는 원칙을 미리 정해 놓고 이에 따라 말뭉치를 구축하는 방법론[하향식 접근(top-down approach)]보다는, 우선 쉽게 사용·접근·수집할 수 있는 자료들을 모아서 말뭉치를 구축하고 이로부터 얻어진 결과를 바탕으로 말뭉치의 장르 비율을 조정하는 방법론[상향식 접근(bottom-up approach)]이 더 현실적이라고 생각된다.

본 연구에서는 이러한 상향식 접근법에 따라 우선 가용 자원부터 수집하여 정리하였다. 가장 먼저 이용할 수 있는 대표적인 말뭉치는 세종 말뭉치이다. 세종 말뭉치 가운데서도 가장 규모가 크고 다양한 장르로 구성되어 있는 현대 문어 말뭉치를 사용하였다. 그 다음으로는 출판사에서 편집용으로 제작한 도서 자료들을 수집하였다. 셋째로는 인터넷을 통해 신문, 잡지, 드라마, 블로그, 위키 백과 등의 자료를 수집하였다. 잡지와 블로그는 다양한 분야의 것들이 있기 때문에, 그 자체로서 장르의 다양성을 확보하는 데 크게 기여할 수 있다. 이렇게 해서 총 100개 장르로 이루어진 약 20억 어절 규모의 말뭉치를 구축하였다. (Ⅲ. 3. 참조)

## 2. 시간/연대에 대한 고려

말뭉치를 구성하는 텍스트들이 어느 시기의 것인가도 중요한 고려사항이다. 시기에 따라 단어의 사용 빈도에 차이가 있다면 더더욱 시기에 대한 고려가 필요하다.

신문 말뭉치는 1990년부터 2015년까지 연도별로 구분되어 있어, 이 시기 동안의 시간적 빈도 추이를 살펴볼 수 있다. 빈도 순위가 시기에 따라 달라질 수도 있기 때문에 시기별 변화 추이에 대한 고찰이 필요하다.

이를 위해 신문 말뭉치에서 총 빈도가 1,000,000 이상인 단어 118개를 뽑아서 연도별 빈도를 추출하였다. 1위 단어인 지정사 ‘이다’의 경우만 표로 제시하면 다음과 같다.

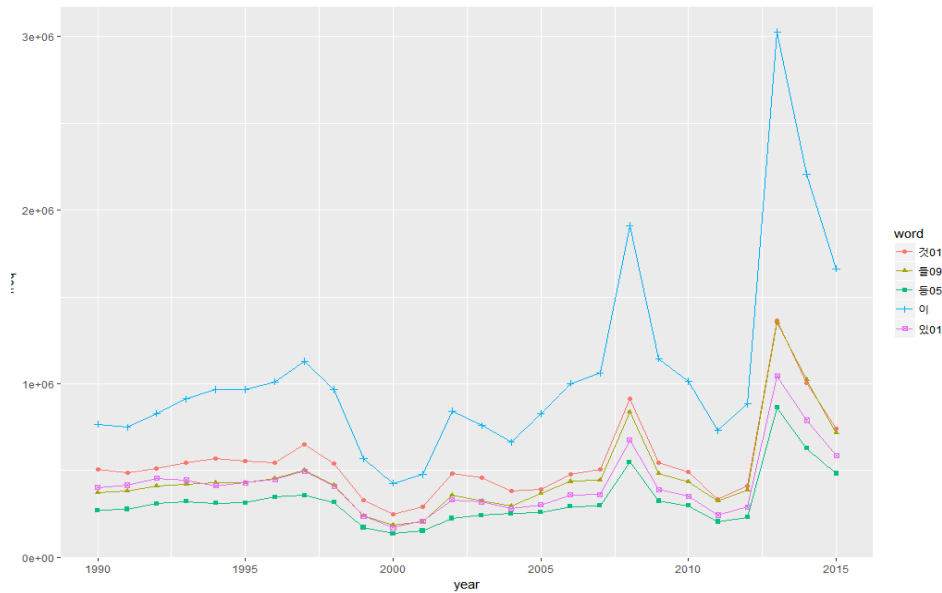
<표 24> 신문 말뭉치 연도별 빈도 통계: 지정사 ‘이다’의 경우

연도	절대빈도	말뭉치크기	상대빈도
1990	766847	25055399	0.030606058199272740
1991	748854	25010645	0.029941410947218673
1992	828378	27352266	0.030285534661003956
1993	911049	28524636	0.031939022815225405
1994	965286	29068510	0.033207274813879350
1995	965579	29671140	0.032542699741230030
1996	1010662	31653420	0.031928998509481755
1997	1129599	35534082	0.031789170745989720
1998	966871	31135770	0.031053383295161800
1999	566501	18258035	0.031027490088610300
2000	426354	14065581	0.030311865539006173
2001	477974	16200095	0.029504394881634952
2002	843027	27502304	0.030652959112080210
2003	757970	26471250	0.028633706379562734
2004	662480	24456428	0.027088174937075847
2005	827059	29701535	0.027845665215619327
2006	1001849	35262321	0.028411317564717310
2007	1060451	37085212	0.028594982819567002
2008	1910804	67208737	0.028430886894958315
2009	1144662	40748042	0.028091214787694583
2010	1014115	35612083	0.028476711120773250
2011	729914	25640518	0.028467209593815540
2012	885112	31365665	0.028219137072336902
2013	3026320	114624934	0.026401934504058252
2014	2209188	83172165	0.026561626717303800
2015	1662168	63098291	0.026342520116749278

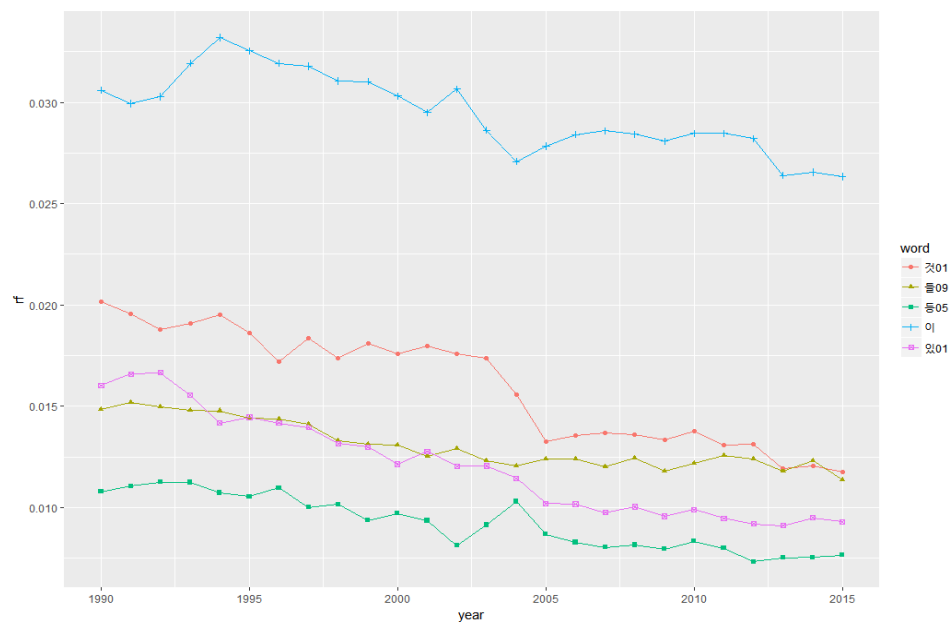
1위부터 5위까지에 해당하는 지정사 ‘이다’, 의존명사 ‘것01’, 복수 접미사 ‘들09’, 보조용언 ‘있다01’, 의존명사 ‘등05’만을 대상으로 하여 연도별 빈도 추이를 그래프로 그려 보면 다음과 같다.



절대빈도는 상대빈도를 알아보는 데에 큰 차이는 없다. 1위와 2위는 순위 변동이 거의 없지만, 3위와 4위는 시기에 따른 순위 변동이 있다. 초기에는 보조용언 ‘있다01’이 접미사 ‘들09’보다 빈도가 높다가, 후기로 가면서 이 둘의 순위가 뒤집힘을 볼 수 있다. 2위 ‘것01’과 3위 ‘들09’의 경우, 순위가 뒤집히는 정도까지는 아니지만, 최근으로 올수록 둘 사이의 격차가 좁혀짐을 볼 수 있다.



[그림 8] 연도별 빈도 최상위 5개 단어의 절대빈도 추이



[그림 9] 연도별 빈도 최상위 5개 단어의 상대빈도 추이

일정한 시간적 범위를 설정해 놓고 그 시기 전체에서의 빈도를 바탕으로 하여 어휘를 선정할 때와, 그 범위 중 특히 최근에 해당하는 짧은 기간을 바탕으로 하여 어휘를 선정할 때 결과가 달라질 수 있다는 사실을 알 수 있다.

이는 기초 어휘 선정 작업에도 시사하는 바가 크다. 기초 어휘 선정의 목적에 따라 선정 작업의 기초가 되는 말뭉치도 상당히 달라야 한다. 만약 20세기 초부터 지금까지의 꽤 넓은 시간 폭을 갖는 현대 국어를 대상으로 하여 기초 어휘를 선정한다고 하면, 20세기 전반기나 중반기의 자료를 말뭉치에 포함하는 것이 타당할 것이다. 반면에 미래 지향적으로 앞으로의 국어 정책, 국어교육 등의 목적을 위해 기초 어휘를 선정한다고 하면, 20세기 전·중반기의 자료는 현재의 언어 사실과 사뭇 다를 가능성이 있기 때문에 말뭉치에 포함하지 않는 편이 나을 것이다. 후자의 목적을 중시한다면 현재에 가까운 자료를 가능한 한 많이 넣고, 현재로부터 시간적으로 거리가 있는 자료일수록 아예 안 넣거나 넣더라도 적은 양을 신중하게 넣어야 할 것이다.

이런 점을 고려하면 세종 말뭉치를 그대로 사용하는 것은 문제가 있을 수 있다. 세종 말뭉치에 포함된 텍스트는 1990년대의 신문, 잡지, 도서가 주종을 이루고 있으나, 소설 같은 경우 시대가 훨씬 올라가는 작품들도 꽤 수록되어 있다. 본 사업에서 사용한 도서 말뭉치도 1980~90년대 자료가 주종을 이루고 있다.

1980~90년대 자료를 말뭉치에서 아예 빠지는 않더라도 시기별 균형을 맞추려면 최신 자료의 양/비중을 대폭 늘릴 필요가 있다. 본 사업에서 최근의 신문과 잡지를 대량 수집한 것도 그런 취지에서 긍정적인 역할을 할 것으로 기대된다. 신문의 경우 한국언론재단에서 구축한 자료 및 이를 바탕으로 한 본 사업의 신문 말뭉치가 1990년부터 최근까지 두루 있기는 하지만, 최근 자료의 양이 비교적 많다.

앞으로 신문, 잡지뿐 아니라 최신 자료를 보다 폭넓게 수집할 수 있는 방안을 강구할 필요가 있다.

### 3. 말뭉치 구축 현황

등급화된 어휘 목록을 추출하기 위해서는 실제 언어생활의 다양한 측면을 포착할 수 있는 말뭉치가 필요하다. 이 말뭉치는 고빈도어뿐 아니라 저빈도어도 포함할 수 있도록 규모가 커야 하고, 다양한 주제·분야의 어휘를 포괄할 수 있기 위해 여러 장르를 망라해야 한다.

이러한 목적을 달성하기 위해 가장 이상적인 방법은, 상당수의 국어 화자들을 대상으로 하여 일상생활에서 어떠한 방식으로 의사소통하고 어떤 매체를 얼마나 오랫동안/자주 접하는지, 즉 언어 사용 실태를 광범위하게 조사하여 한국인의 언어생활에서 각 장르가 어느 정도의 비중을 차지하는지 알아낸 뒤, 이 장르 비율을 바탕으로 말뭉치를 구축하는 것이다. 그러나 이러한 조사를 체계적이고 정확하게 수행하는 데에도 시간과 비용이 많이 들고, 그러한 조사가 이루어진다 하더라도 그 조사 결과 얻어진 장르 비율을 충실히 반영하는 말뭉치를 구축하는 데에도 많은 시간과 비용이 소요될 것이다. 따라서, 실행 가능한 차선택은, 우선 시간과 비용을 많이 들이지 않고 수집할 수 있는 여러 장르의 텍스트 자료를 가능한 한 폭넓게 수집하여 이를 바탕으로 등급화된 어휘 목록을 추출한 뒤, 이 어휘 목록이 질적 방법으로 구축된 기존 어휘 목록과 비교하여 어떤 문제점이 있는지 파악하여, 그 문제점을 보완할 수 있도록 말뭉치를 보완하는 것이다.

이러한 전제 하에, 본 연구에서는 아래와 같이 말뭉치 자료를 수집하고 정리하였다.

- ① 이미 구축되어 공개된 대규모 말뭉치를 최대한 이용하였다. 21세기 세종계획에서 구축된 현대 문어 원시 말뭉치가 이에 해당한다. 현대 구어 말뭉치, 형태 분석 말뭉치 등은 규모가 작아서 일단 제외하였으나, 1차년도의 결과를 검토한 뒤 차년도의 작업에서 포함시킬 수 있다.
- ② 인터넷에서 웹 크롤링 기법을 이용하여 자동 수집할 수 있는 자료를 폭넓게 수집하였다. 드라마·시나리오, 신문, 잡지, 블로그 등이 이에 해당한다.
- ③ 출판사에서 출판 과정에서 만든 편집용 파일, PC 통신 등에서 연재된 소설 등의 각종 도서 자료를 폭넓게 수집하였다.

수집된 자료를 장르별로 정리하여 규모를 제시하면 아래와 같다.

<표 25> 말뭉치 전체 통계

종류	내용	비고	장르 수	어절 수	용량 (MB)
세종 말뭉치	21세기 세종계획 현대 문어 원시 말뭉치	신문, 소설 제외	10	32,407,263	374
도서 말뭉치	국내 출판 각종 도서	세종 말뭉치의 소설을 여 기의 소설에 합침	24	173,730,053	1,640
잡지 말뭉치	인터넷 서비스되는 잡지 34종	성격이 비슷하고 분량이 작은 것은 한데 묶음	25	408,273,167	3,344
블로그 말뭉치	네이버, 알라딘, LG CNS	네이버: 35개 장르 222개 파워블로그 알라딘: 72개 블로그	37	245,628,059	2,642
드라마 말뭉치	드라마 대본, 영화 시나리 오		1	39,314,345	380
신문 말뭉치	한국언론재단에서 제공하 는 신문 기사	11개 신문 1990~2015년 기사, 세종 말뭉치의 신 문을 여기에 합침	1	961,256,124	17,849
위키 백과	한국어 위키피디아		1	166,040,163	1,775
방송 뉴스	SBS 뉴스	2017 8~10월	1	18,919,129	193
계			100	2,045,568,303	29,197

신문 말뭉치에 포함된 11개 신문은 경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레신문, 한국일보이다.

세종 말뭉치, 도서 말뭉치, 잡지 말뭉치, 블로그 말뭉치는 장르별로 세분할 수 있다. 이 네 말뭉치의 장르별 통계는 다음과 같다.

<표 26> 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		32,397,263

<표 27> 도서 말뭉치 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
시공디스커버리총서	1,032,816	컴퓨터	314,668
계		173,730,124	

<표 28> 잡지 말뭉치 장르별 통계

제목		어절 수	제목		어절 수
프레시안		90,976,514	우먼센스		6,603,139
매경이코노미		70,107,812	표표스스		6,320,907
시사저널		46,314,682	기독교잡지 6,115,736	기독교 사상	5,895,114
씨네21		29,524,004		청어람 매거진	220,622
대학 학보 28,665,696	서울대 대학신문	5,224,946	트레비		5,725,496
	연세대 연세춘추	7,450,907	행복이가득한집		5,380,007
	고려대 고대신문	6,829,818	좌파매체 4,941,436	레디앙	3,411,282
	한양대 한대신문	2,971,524		시민교육센터	1,058,782
	이대 이대학보	6,188,501			471,372
지방지 <sup>11)</sup> 19,123,996	인천일보	18,660,211		민중의 소리	3,813,572
	토마토	463,785	1,683,046		
딴지일보		14,954,069	컴퓨터월드		3,813,572
과학잡지 14,290,268	과학동아	13,859,979	객석		1,683,046
	어린이 과학동아	430,289	뉴스위크		1,292,360
이코노미스트		13,980,607	르몽드디플로마티크		1,187,661
시사인		13,901,344	문학신문		1,098,865
월간중앙		13,686,395	쎬쎬		698,304
레이디경향		7,307,010	올리브		580,241
			계		413,769,785

11) 대학 학보와 지방지는 관점에 따라 잡지가 아니라 신문에 포함시켜야 한다고 볼 수도 있다. 그러나 본 연구의 신문 말뭉치에 포함된 11개 신문은 전국적으로 발행되는 일간지여서, 대학 학보와 지방지를 여기에 포함시키기에는 이질적이라고 판단된다. 본 연구의 잡지 말뭉치에 포함된 것들은 그 종류가 매우 다양하여, 대학 학보나 지방지의 성격에 매우 가까운 것들도 포함되어 있다. 이러한 성격 및 발행 주기 등을 고려하여 잡지 말뭉치에 포함시켜 둔다. 통계 수치를 뽑을 때는 대학 학보가 하나의 장르로 간주되고 지방지도 하나의 장르로 간주되므로, 이 둘이 잡지 말뭉치에 포함

<표 29> 블로그 말뭉치 장르별 통계

장르		어절 수	장르		어절 수	장르		어절 수
네이 버블 로그	영화	9,305,123	네이 버블 로그	원예,재배	4,761,881	네이 버블 로그	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		알라딘	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	LG CNS	1,082,524	
	음악	4,834,591		IT,웹,프로그램	3,855,726			
						계	245,628,059	

본 연구에서 구축하여 어휘 등급화에 사용한 말뭉치의 특징은 다음과 같이 정리할 수 있다.

- ① 규모: 기존의 어떤 말뭉치보다 규모가 월등히 크다.
- ② 구성: 주로 문어에 한정되어 있기는 하나, 문어의 매우 다양한 장르를 포함한다.
- ③ 시대적 성격: 현대 정보화 사회에서의 국민들의 언어생활상의 특징을 반영하여, 블로그, 인터넷 잡지 등 인터넷 매체의 최근의 텍스트들이 대폭 포함되어 있다.

되어 있다는 사실은 큰 의미가 없다.

#### 4. 말뭉치 보완 계획

본 연구에서는 이미 확보한 말뭉치를 바탕으로 각 단어의 빈도, 범위, 산포도를 산출하고 이들 변수를 바탕으로 단어의 순위를 정한 뒤, <연세한국어사전> 등의 기존 어휘 목록과 비교하는 식으로 연구를 진행하였다. (IV.1. 참조)

특히 기존 어휘 목록과의 비교 작업을 통해 현재 확보하고 있는 말뭉치의 문제점을 알아볼 수 있었다. 최근에 만들어진 어휘 목록은 대개 말뭉치를 기반으로 한 것인데, 모든 말뭉치는 한국어라는 모집단을 완벽하게 대표할 수는 없고 선정된 텍스트가 무엇인가에 따라 편향성을 가질 수밖에 없다. 따라서 기존 어휘 목록들이 완벽하다고 할 수는 없으나, 여러 기존 어휘 목록을 비교함으로써 이러한 편향성을 어느 정도는 극복할 수 있다. 또한 기존 어휘 목록 중 정성적인 방법으로 만들어진 것은, 정량적인 방법에 의해 추출된 어휘 목록을 보완하는 역할을 할 수 있다. 즉 여러 기존 어휘 목록들에는 공통적으로 높은 순위로 들어 있는 단어가 우리의 말뭉치에서는 나타나지 않는다거나 나타나더라도 매우 낮은 순위로 나타난다면, 우리의 말뭉치에 심각한 결함이 있다는 증거가 될 것이다. 그런 단어가 일정한 의미 범주에서 많이 나타난다면, 그 의미 범주의 단어들이 흔히 나타나는 장르가 현재 말뭉치에서 빠져 있거나 매우 과소평가(under-represented)되어 있다는 뜻이 될 것이다. 그런 장르가 발견된다면 해당 장르를 보완해야 할 것이다.

당해 연도에는 실험을 위한 기존 어휘 목록으로서 <연세 한국어 사전>만을 이용했는데, 그 외에도 다양한 어휘 목록을 사용할 필요가 있다. <초등 국어사전>(두산동아), 국립국어원의 기초 어휘 통계 조사, (주)날말의 등급화된 어휘 목록 등을 생각할 수 있다. 질적·양적 방법론을 함께 사용하여 구축된 이들 기존 어휘 목록과의 일치율을 높일 수 있도록 노력하고, 일치하지 않는 부분을 찾아서 그 원인을 규명함으로써 기존 말뭉치의 한계와 문제점을 발견하고 보완해 나아갈 것이다.

현재 말뭉치의 장르 구성도 더 다듬을 필요가 있다. 규모가 너무 작은 장르들을 비슷한 것들끼리 묶어서 하나의 장르로 통합할 만한 것들도 있고, 반대로 규모가 매우 크고 이질적인 것들이 섞여 있어서 나눌 만한 것들도 있을 것이다. 규모나 내용 면에서 두 개 이상의 장르로 나눌 만하다고 판단된 것들 가운데에도, 단어의 빈도 분포는 상당히 비슷하게 나오는 경우가 있을 수 있다. 다변량 통계 분석을 통해 단어 빈도 분포가 일정 정도 이상으로 비슷한 패턴을 보이는 장르들을 찾아내어 하나로 통합하는 작업을 할 필요가 있다.

현재의 말뭉치는 문어에 매우 편중되어 있으므로, 앞으로 구어 장르를 대폭 보완할 필요가 있다. 완전한 구어의 성격을 지니는 대화는 말뭉치로 구축하는 데 시간과 비용이 많이 소요되므로, 대안으로서 구어적 성격을 상당히 지니는 자료를 수집할 수도 있다. 영화·드라마 대사, 인터뷰, 인터넷 게시판, 속기록 등이 그러한 예들

이다. 이러한 점을 감안하여 말뭉치를 <표 30>과 같이 보완하였다.<sup>12)</sup>

아래의 보완 말뭉치에는 구어적 성격을 어느 정도 띤 텍스트들이 많이 포함되어 있어, 내년도 사업에서 이것을 말뭉치에 반영해 다시 통계 작업을 할 경우, 단어들의 순위 변동이 기대된다.

내년도에는 말뭉치의 규모를 확장하기보다는, 기존 20억 어절 규모 말뭉치로부터 얻은 어휘 목록, 약 9억 어절 규모의 보완 말뭉치로부터 얻은 어휘 목록, 그리고 기존 말뭉치와 보완 말뭉치를 합친 29억 어절 규모 말뭉치로부터 얻은 어휘 목록 이 3자를 비교하는 데 초점을 맞출 것이다. 기존 20억 어절 규모 말뭉치는 구어가 미미하고 문어에 치중되어 있는 반면에, 9억 어절 규모의 보완 말뭉치는 구어적 성격을 상당히 띠고 있는 자료이므로, 이것을 포함했을 때 어휘 등급에 상당한 변동이 생길 수 있다. 기존 말뭉치를 바탕으로 한 결과에서는 낮은 등급을 차지했던 단어가 보완 말뭉치를 바탕으로 얻은 결과에서는 그보다 훨씬 높은 등급으로 나타난다면, 이들 단어는 문어보다 구어 자료에서 고빈도로 나타나는 단어, 즉 구어적 성격이 강한 단어임을 알 수 있다. 기존 말뭉치와 보완 말뭉치를 합친 말뭉치를 바탕으로 어휘 등급화를 시행하면, 그 결과는 문어와 구어를 적절히 아우른 결과가 될 수 있을 것이다.

이러한 작업을 거친 뒤에도 다시 정성적으로 구축된 어휘 목록과의 비교도 할 필요가 있고 연구자들의 직관을 동원한 정성적 검토 작업을 해야 한다. 그 결과 일부 단어들이 과대 평가(over-represent)되었다든지 일부 단어들이 과소 평가(under-represent)되어 있다고 판단된다면, 과소 평가된 단어들의 반영 비중이 높아질 수 있는 방향에서 말뭉치 보완 계획이 수립되어야 할 것이다. 현재까지는 인터넷 등에서 손쉽게 구할 수 있는지 여부를 바탕으로 말뭉치의 규모를 늘리는 작업을 해 왔으나, 제2차년도의 연구 결과를 바탕으로 하여 그 후에는 과소 반영된 장르를 보완하는 방향으로 작업이 진행되어야 하며, 이를 위해서는 말뭉치 기반 어휘 목록에 대한 정성적 검토가 향후 작업에서 이루어져야 하는 것이다.

12) 통계 작업을 마친 뒤여서, IV.1.의 통계에는 반영되어 있지 않다.



<표 30> 보완 말뭉치의 구성

대부류	소부류	어절 수	비고
블로그	효성	1,808,330	
잡지	pitchone	407,271	
	insight	17,393,051	
	교수신문	12,417,961	
인터뷰	CBS 뉴스쇼	9,274,808	
	CBS 시사자키	8,431,664	
	YTN 뉴스투데이	3,702,902	
	YTN 뉴스정면승부	5,020,996	
	YTN 당신의전성기	1,059,081	
	YTN 생생경제	3,325,544	
	YTN 수도권투데이	1,817,856	
	YTN 열린라디오	102,093	
	YTN 출발새아침	7,792,344	
	MBC 뉴스의광장	826,989	
	MBC 시선집중	8,536,501	
	MBC 세계는우리는	6,965,041	
	SBS 시사전망대	3,981,014	
	TBS 뉴스공장	1,059,747	
	TBS 색다른시선	1,274,939	
	KBS 윤준호입니다	2,992,345	
	불교방송 아침저널	6,147,149	
	평화방송 열린세상오늘	12,276,471	
	경인방송 세상을연다	160,486	
	YES24 만나고싶었어요	3,263,644	
	폴리뉴스 김능구의정국진단	846,316	
	위클리피플	1,258,702	
국회속기록	최근회의록	7,261,047	
인터넷게시판	다음아고라-정치	382,642,212	
	클리앙-모두의공원	53,386,055	
	루리웹	20,081,251	
	보배드림-베스트글	11,109,313	
	디시인사이드-초개념	9,035,067	
	mlbpark-today best-볼펜추천	4,746,331	
	뽀뿌	53,548,664	
	오유	36,392,355	
라디오드라마	KBS 한민족방송 라디오극장	5,870,321	
영화·드라마 자막	씨네스트 영화	154,743,546	
	씨네스트 드라마	3,388,610	
	공TV 드라마	26,570,174	
신문	동아일보 한중대역 기사	10,353,263	중국어 부분은 띄어쓰기가 없으므로, 한국어 부분만 계산해도 1천만 어절에 가까움
계		901,271,454	기존 말뭉치와 합치면 29억 어절을 상회함.

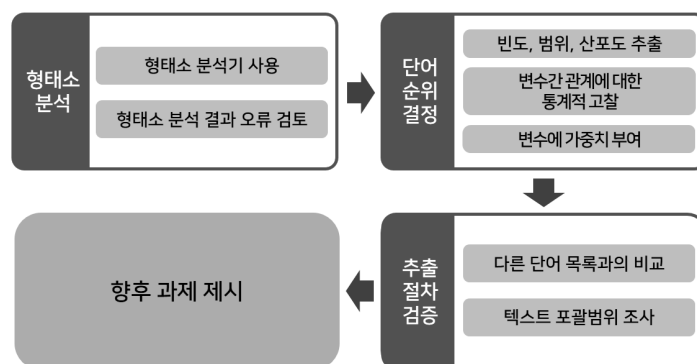
## IV. 기초 어휘 선정 및 어휘 등급화

이 장에서는 말뭉치를 기반으로 하여 기초 어휘 추출을 위한 어휘 목록 선정 및 등급화에 대해 논의한다. 이 장의 구성은 크게 두 부분으로 나뉘는데, 먼저 III장에서 구축한 말뭉치를 토대로 어휘 목록을 추출하는 과정을 전반적으로 수행한 내용을 제시한다. 이러한 과정에서는 어휘 목록 선정에 관여하는 변수들 간의 관계 분석, 기존 어휘 목록과의 비교 등을 통해 어휘 목록 선정에 필요한 요소들을 점검하고 파악하였다.

다음으로 기초 어휘 목록 선정과 등급화를 위해 구체적 논의를 진행한다. 본 연구에서는 등급화 대상 어휘를 일단은 기초 어휘와 비기초 어휘라 명명하고 논의를 진행하였다. 특히 1등급 기초 어휘 목록은 이론 연구와 연구진 회의를 통해 말뭉치 구축이 별도로 이루어져야 한다는 점이 지적된 바 있으므로 이에 대한 과정에 대해 설명하였다. 그리고 2등급 이상 어휘 목록은 IV장 1절에서 논의한 결과를 토대로 어휘 목록 수를 정하고 이를 대상으로 검증하는 절차를 제시하였다.

### 1. 말뭉치 기반 어휘 목록의 추출

본 연구에서는 사전등재형을 어휘 추출 단위로 정하고 일차로 형태소 분석 작업을 거친 후, 장르별 빈도, 범위, 산포도를 추출하고 변수들 사이의 관계에 대한 통계적 고찰을 실시한다. 이후 변수들에 가중치를 부여하여 단어 순위를 결정하고 그 결과를 다른 어휘 목록과 비교하고 텍스트 포괄 범위를 조사하여 검증하였다. 이는 샘플 말뭉치를 활용하여 구체적인 절차를 실행해본 것으로, 각각의 절차를 실제로 수행해 봄으로써 실제 기초 어휘 선정 및 어휘 등급화 사업을 진행할 때의 유의점이나 시사점을 도출하였다. 이러한 절차를 도식화하면 다음과 같다.



[그림 10] 말뭉치 기반 어휘 목록의 추출 절차

## 1.1. 형태소 분석

### 1) 형태소 분석기 사용

원시 말뭉치만으로는 각 단어의 빈도를 정확히 낼 수 없다. 각 어절을 체언과 조사, 용언 어간과 어미로 분석하고, 동형어가 있는 경우 동형어의 구분도 해야 한다. 대규모 말뭉치에 대해서 이것을 수작업으로 할 수는 없다.

울산대학교 한국어 처리 연구실에서 개발한 UTagger가 형태소 분석과 동형어 구분을 자동으로 해 주기 때문에, 이 작업에 이용하기에 적합하다. 한국어 자동 형태소 분석기가 여럿 개발되어 있으나, UTagger가 정확도가 가장 우수하고,<sup>13)</sup> 게다가 동형어 구분까지 해 주는 것은 UTagger밖에 없기 때문에, 현재로서 다른 대안은 없다고 할 수 있다. UTagger는 형태소 분석을 한 뒤, 동형어가 있을 경우 <표준국어대사전(종이 사전)>의 어깨번호를 붙여 준다.

UTagger의 형태소 분석 결과를 예시하면 다음과 같다.

입력	경제난이 새로운 한반도 불안정 요소로 떠오르고 있는 실정이다. 하지만 자력회복이 어려운 상태여서 남북 경제 교류, 협력이 가장 현실적이고 효과적인 경제난 회복방법이다.
출력	경제난/NNG+01/JKS 새롭/VX+ㄴ/ETM 한반도/NNP 불안정/NNG 요소__04/NNG+로/JKB 떠오르/VV+고/EC 있__01/VX+는/ETM 실정__04/NNG+01/VCP+다/EF+./SF 하지만/MAJ 자력__01/NNG+회복/NNG+01/JKS 어렵/VX+ㄴ/ETM 상태__01/NNG+01/VCP+어서/EC 남북/NNP 경제__04/NNG 교류__01/NNG+./SF+협력/NNG+01/JKS 가장__01/MAG 현실적/NNG+01/VCP+고/EC 효과적/NNG+01/VCP+ㄴ/ETM 경제난/NNG 회복/NNG+방법/NNG+01/VCP+다/EF+./SF

13) 본 연구에서는 UTagger, 한나눔 형태소분석기(KAIST 개발), 꼬꼬마(서울대 Intelligent Database System 연구실 개발), Komoran(Shineware 개발), Twitter 분석기(트위터 개발), 은전한늬(일명 Mecab, Atlassian 개발)의 성능을 비교 분석하였으며, UTagger의 성능이 가장 우수함을 확인하였다.

## 2) 형태소 분석기 오류 검토

### (1) 형태소 분석기 오류 검토의 필요성

기초 어휘는 대규모 말뭉치를 구축하고 형태소 분석기를 통해 빈도와 분포 등을 고려하여 선정되므로 형태소 분석기의 오류를 사전에 검토하여 예상되는 문제점을 검토할 필요가 있다. 이를 위해 동음이의어 분석이 가능한 UTagger를 활용하여 문어 10,000어절, 구어 6,000어절 규모를 분석하여 오류 유형을 살펴보기로 한다. 문어와 구어는 모두 세종 균형 말뭉치의 분야별 자료로서 그 구체적인 내용은 다음과 같다.<sup>14)</sup>

<표 31> 검토 대상 자료

구분	영역	매체	제목	출판연도
문어	책	책:상상적 텍스트-일반	날마다 축제	2004
		책:자연-기초자연과학	유전자가 세상을 바꾼다	2004
	신문	신문:보도해설-사회	조선일보 2003년 기사: 사회	2003
	잡지	잡지:사회-일반	주간조선 1770호	2003
		잡지:생활	여성중앙21, 전자파일	2000
구어	강연	강연	아이발달	
	토론	토론	세계화세미나	
	회의	회의	총학생회전체회의	

### (2) 오류 유형 결과 분석

문어, 구어 자료의 형태소 분석 오류 유형 별 사례는 <부록 2>에서 제시하는 것으로 대신하고, 이를 종합적으로 분석한 내용을 정리하면 다음과 같다.

먼저 문어와 구어에 따른 오류 유형의 차이를 살펴보면 다음과 같다. 말뭉치를 분석한 결과 문어의 오류 유형은 크게 동음이의어 분석 오류, 품사 분석 오류, 품사 통용 관련 오류, 오분석에 따른 오류, 직접 성분 분석에 따른 오류로 나눌 수 있다. 그리고 구어의 오류 유형은 크게 동음이의어 분석 오류, 품사 분석 오류, 오분석에 따른 오류, 직접 성분 분석에 따른 오류, 구어적 특성에 따른 오류로 나눌 수 있다.

전체적으로 오류가 나타나는 경우는 표본 말뭉치의 수가 더 적은 구어에서 보이는 오류 항목이 표본 말뭉치의 수가 더 많은 문어에서 보이는 오류 항목보다 훨씬 많았다. 이는 기본적으로 형태소 분석기가 한글 맞춤법에 기반한 문어 지향적 속성을 띠기 때문이다. 그러나 구어의 경우는 동일한 항목 오류가 반복되는 경우가 적

14) UTagger는 프로그램의 특성상 세종 태그셋을 활용하여 품사 태그를 부착한다. 세종 태그셋은 <부록 1>에 제시되어 있다.

지 않아 전체 오류 항목 수가 늘어난 것임을 염두에 둘 필요가 있다.

문어에서도 구어적 특성에 따른 오류가 간간히 발견되기는 하였지만 이를 따로 오류의 유형으로 구분할 정도는 아니다. 이것은 표본이 되는 문어 말뭉치의 특성이자 할 수 있고 앞으로 문어 가운데 대화와 같은 구어적 특성이 풍부한 자료가 포함될 경우 따로 오류 유형으로 나눌 가능성이 있다.

표본으로 삼은 말뭉치를 분석한 결과 구어에서는 문어에서 보이는 품사 통용 관련 오류가 발견되지 않았는데 이것은 표본이 되는 구어 말뭉치의 특성일 뿐이므로 보다 대규모의 말뭉치를 대상으로 할 경우 충분히 발견될 수 있는 오류에 해당한다.

따라서 문어와 구어를 모두 염두에 둘 때 오류 유형은 동음이의어 분석 오류, 품사 분석 오류, 품사 통용 관련 오류, 오분석에 따른 오류, 직접 성분 분석에 따른 오류, 구어적 특성에 따른 오류 등으로 그 범위를 확대할 가능성이 높다.

다음으로, 오류 유형과 기초 어휘 선정 및 등급과의 관계를 살펴 보면 다음과 같다. 조사된 오류는 형태소 분석기 측면에서는 모두 동일한 가치를 지니는 오류이지만 기초 어휘 선정 및 등급 선정과 관련해서는 그 가치가 각각 다르다. 즉 형태소 분석 오류 가운데는 기초 어휘 선정 및 선정된 어휘의 등급과 관련해 주목해야 할 오류와 무시해도 좋을 오류가 포함되어 있다.

동음이의어 분석 오류는 구어뿐만 아니라 문어에서도 가장 높은 비중을 차지하고 있는데 이는 기초 어휘 선정 및 선정된 어휘의 등급에 가장 큰 영향을 미치는 오류 유형이라고 할 수 있다. 전술한 바와 같이 여러 형태소 분석기 가운데 UTagger를 활용하려는 가장 큰 이유도 이 동음이의어 분석 기능에 있음을 상기할 때 동음이의어 분석 오류를 집중적으로 검토할 필요가 있다. UTagger를 통한 동음이의어는 고유한 어깨번호를 가지고 있으므로 일차적으로 형태소 분석기를 거쳐 기초 어휘로 선정된 어휘들 가운데 어깨번호를 가지고 있는 것들을 대상으로 그 오류 양상을 분석하고 반영하는 절차를 거칠 필요가 있다.

품사 분석 오류는 문어에만 한정되는 것도 있고 구어에만 한정되는 경우도 있지만 보다 중요한 것은 기초 어휘 선정 및 선정된 어휘의 등급과 관련하여 주목해야 할 것들을 선별하는 것이다. 우선 조사와 어미는 그것이 결합한 선행 어휘의 분석에 필요한 정보이고 그 자체가 기초 어휘에 해당하지는 않으므로 조사나 어미와 관련된 품사 분석 오류는 중요성을 가지지 않는다. 이에 따라 부사격 조사와 접속 조사 사이의 오류는 무시해도 된다. 특히 문어와 구어 모두에서 종결 어미와 연결 어미 사이의 오류가 적지 않은데 이 역시 기초 어휘 선정에 영향을 미치지 못하므로 무시해도 되는 오류에 해당한다.

한편 구어의 경우 감탄사와 관련된 오류가 문어의 경우에 비해 압도적으로 높지만 감탄사는 소리를 임의적으로 전사한 경우가 대부분이므로 기초 어휘에서는 배제

되는 일이 일반적이라는 사실을 감안하면 이 오류도 크게 신경 쓸 필요는 없어 보인다. 반면 나머지 품사 분석 오류는 구어와 문어 사이에 차이를 보이기는 하지만 결과적으로 앞의 동음이의어 분석 오류와 동일한 효과를 가진다는 점에서 선정된 기초 어휘의 구분 및 등급 산정 때 이를 검토하여 반영할 필요가 있다.

품사 통용 오류는 한 어휘의 다의어적 쓰임에 다른 것이므로 기초 어휘 선정 및 등급을 부여할 때 큰 중요성을 가지지 않는다. 다만 동일한 어휘의 다른 품사는 이를 합산해서 기초 어휘 선정 및 등급 부여에 반영해야 한다는 점에 주의할 필요가 있다. 가령 표본 말뭉치 분석에서 두드러진 품사 통용 오류는 부사와 명사 사이에서 나타났는데 이것이 동음이의어로 처리되어 서로 다른 어휘로 간주되지 않도록 주의할 필요가 있다.

오분석에 따른 오류는 다시 과분석에 따른 오류와 미분석에 따른 오류로 나누었는데 과분석은 하나의 어휘를 지나치게 분석한 것이고 미분석은 둘 이상의 어휘로 나눌 수 있는 것을 분석하지 않은 것이다. 표본 말뭉치 분석 결과 이들 각각에 해당하는 경우가 적지 않았는데 과분석이나 미분석 모두 기초 어휘 선정 및 등급 결정에 영향을 미칠 수 있다. 따라서 선정된 기초 어휘에서 과분석이나 미분석에 해당하는 어휘가 존재하는지 여부를 확인할 필요가 있다.

직접 성분 분석에 따른 오류는 그 경우의 수가 많지 않으나 특히 전사된 자료가 띄어쓰기 측면에서 오류를 가질 경우에 직접 성분 분석에 따른 오류가 발생하는 경우가 많다. 방대한 양의 말뭉치를 처리하는 과정에서 이를 바로 잡기는 매우 어려운 일이지만 기초 어휘 선정에서 고려해야 하는 부분에 해당한다.

전술한 바와 같이 표본 말뭉치 분석에서 드러난 구어적 특성에 따른 오류는 특히 주로 구어 말뭉치에 한정되는 속성을 갖는다. 먼저 더듬거림에 따른 오류는 그 경우도 많지 않을 뿐만 아니라 한 어휘의 일부에 품사를 부여하게 되므로 기초 어휘 선정 과정에서 크게 문제가 되지 않아 보인다. 또한 조사나 어미가 구어적 특성에 따라 변형되는 경우도 앞서 언급한 바와 같이 기초 어휘 선정에서 문제될 것이 없다. 그러나 준말을 포함하여 체언이나 용언 혹은 수식언이 구어적 변형을 거치는 경우에는 이를 문어로 환원하여 분석해야 하므로 기초 어휘 선정 및 등급 판정에 영향을 미칠 수 있다. 따라서 이들에 대해서는 기초 어휘 선정 및 등급 판정 과정에서 주목할 필요가 있다.

향후 기초 어휘 산정을 위해 처리하게 될 말뭉치 규모는 오류 유형 분석을 위해 선정한 것과는 비교가 되지 않을 정도로 크기 때문에 이들 각각에 대한 오류를 일일이 바로 잡는 것은 불가능하다. 따라서 유형별 오류들을 바로 잡을 수 있도록 별도의 조치를 강구할 필요가 있다. 이는 두 가지 측면에서 가능할 것으로 보이는 바 한 가지 방법은 오류 유형을 반영하여 현행 UTagger의 성능을 업그레이드하는 것이다. 그러나 이는 UTagger 개발자와의 협력이 필요하다는 문제가 있다. 따라서 가

능한 대안은 연구팀이 오류 유형을 보다 정밀하게 분류하여 기초 어휘로 선정 가능한 단어들을 대상으로 오류 유형을 바로잡을 수 있는 프로그램을 자체적으로 개발하는 것이다.

## 1.2. 장르별 단어 빈도, 범위, 산포도 추출

세종 말뚝치의 장르들 중 분량이 적은 21~29는 ‘대화’라는 하나의 장르로 묶어서 처리하였다. 그 결과 본 연구의 분석 말뚝치는 세종 말뚝치 장르 10개, 도서 말뚝치 장르 24개, 잡지 말뚝치 장르 25개, 블로그 말뚝치 장르 37개, 드라마 말뚝치·신문 말뚝치·위키백과·방송뉴스 각 장르 1개로, 총 100개 장르로 구성되었다.<sup>15)</sup>

이 100개 장르 각각에 대해 각 단어(형태소)의 빈도를 추출한 다음, 이 100개의 파일을 하나의 파일로 통합하였다. 이 파일의 한 부분을 보이면 다음과 같다.

[illegible]

[그림 11] 100개 장르에서의 단어별 절대빈도

이 파일에 기록되어 있는 수치는 절대빈도이다. 그런데 각 장르의 규모가 천차만 별이므로 이 절대빈도를 그대로 사용할 수는 없고, 각 장르의 규모가 동일하다고 전제했을 때의 빈도로 환산해야 한다. 본 작업에서는 각 장르가 1000만 어절 규모라고 전제했을 때의 빈도로 환산한 뒤, 각 단어에 대해 이 환산된 빈도의 합, 범위, 산포도를 구하였다. 이를 빈도순대로 소팅한 결과의 첫 부분을 보면 다음과 같다.

15) Ⅲ.3.에 말뚝치 전체 통계 및 장르별 통계를 자세히 수록하였다.

	1	2	3	4
1	이/VCP	37453589.37048225	100	97.91168445837461↓
2	것__01/NNB	17321134.381407566	100	97.1353364461081↓
3	들__09/XSN	14411994.588581333	100	96.79985763802394↓
4	하__01/VV	13611982.89881008	100	96.46163209799118↓
5	있__01/VA	13293198.987567576	100	97.55577888606076↓
6	있__01/VX	9439647.91418627	100	97.11971731433736↓
7	수__02/NNB	7897218.146092479	100	96.77145949762249↓
8	되__01/VV	7601453.126661401	100	97.1139049461496↓
9	하__01/VX	6331603.328728959	100	97.58000976570487↓
10	않/VX	5765145.55906604	100	97.60526878368962↓
11	없__01/VA	5128633.710224747	100	96.86039529651202↓
12	그__01/MM	5003000.740099675	100	94.22767636136007↓
13	보__01/VV	4720936.4623573385	100	95.09741953140339↓
14	이__05/MM	4515948.791985744	100	96.3232406130526↓
15	주__01/VX	4104260.5412761383	100	91.68707160205717↓
16	년__02/NNB	4031396.655156952	100	92.41200641025212↓
17	사람/NG	3675291.7517757164	100	94.60715023483269↓
18	그__01/NP	3553930.9195911456	100	91.50372445990051↓
19	아니/VCN	3534725.405859018	100	96.64294082689419↓
20	나__03/NP	3489093.1659047366	100	90.14068442538168↓
21	같/VA	3407444.1643081196	100	96.6929428576926↓
22	때__01/NG	3402034.2011654675	100	96.68317564688775↓
23	보__01/VX	3383820.0048366017	100	92.92837659015434↓
24	한__01/MM	2915412.9751448943	100	96.70222424705834↓
25	등__05/NNB	2897570.5914934683	100	92.09151444889113↓
26	지__04/VX	2868520.6139228893	100	96.83119199295825↓
27	좋__01/VA	2719953.0952579714	100	92.28245283578762↓
28	대하__02/VV	2665759.5080093984	100	94.01179333403859↓
29	가__01/VV	2585176.77012815	100	93.6682859758379↓
30	우리__03/NP	2447510.159288957	100	93.95763397235274↓
31	만들/VV	2416263.907959582	100	92.09762084093647↓
32	위하__01/VV	2310453.4263258358	100	95.69948813700994↓
33	더__01/MAG	2292376.6866783467	100	95.61645642947927↓
34	때문/NNB	2196356.0796374925	100	95.83741793105409↓
35	맡/VA	2122948.192297459	100	96.67827168833306↓
36	받__01/VV	2097003.3051201273	100	96.99968825727436↓

[그림 12] 단어별 빈도, 범위, 산포도(빈도순 소팅 결과의 첫부분)



### 1.3. 세 변수 사이의 관계에 대한 통계적 고찰

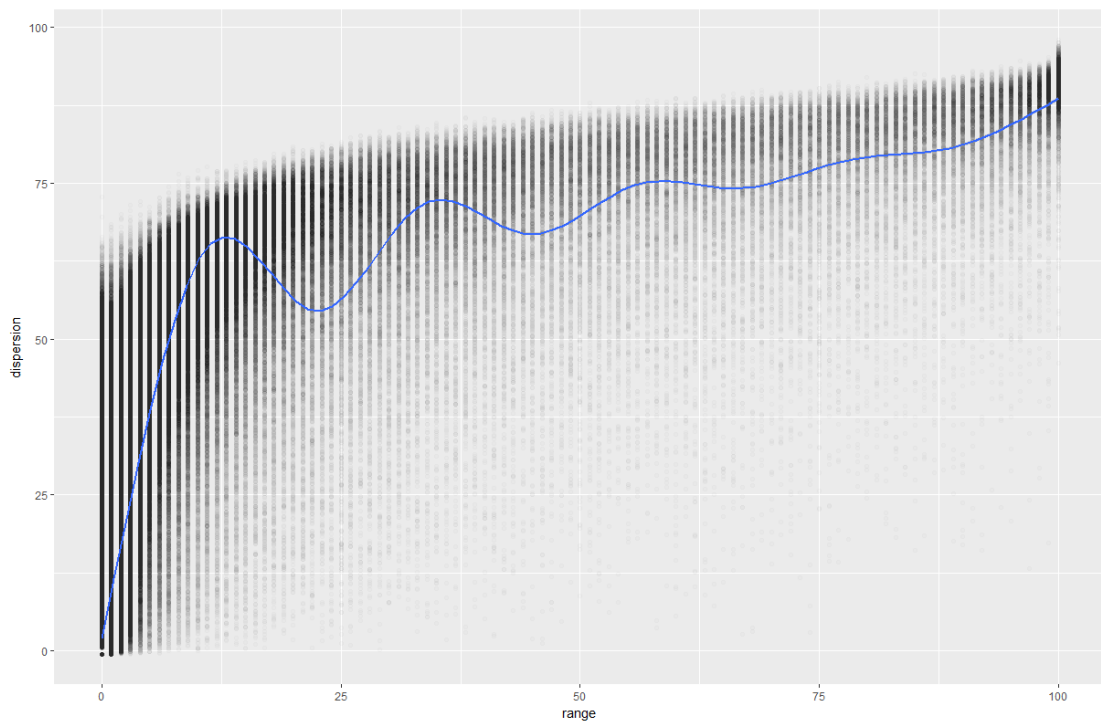
앞의 결과에서 빈도가 최상위인 단어들은 범위가 대체로 만점이 100에 가깝고 산포도도 만점인 100에 가까움을 볼 수 있다. 그러나 이 세 변수가 완전히 일치하는 것은 아니기 때문에, 빈도·범위·산포도 중 어느 것에 어느 정도의 가중치를 두느냐에 따라 단어의 순위가 달라질 수 있다. 예컨대 위의 그림에서 산포도에 가장 큰 가중치를 준다면 단어들의 순위가 상당히 달라질 것이다.

변수들 사이의 상관관계의 강도를 나타내는 상관계수는 다음과 같다.

- 빈도~범위	:	0.09476222
- 범위~산포도	:	0.6769272
- 빈도~산포도	:	0.0448866

위의 상관계수를 통해 빈도와 범위 사이에, 그리고 빈도와 산포도 사이에는 유의미한 상관관계가 없으나 범위와 산포도 사이에는 유의미한 양의 상관관계가 있음을 알 수 있다.

세 변수들 중 각각의 두 변수들 사이의 관계를 좀 더 자세히 알아보기 위해 그래프를 그려 보면 다음과 같다.



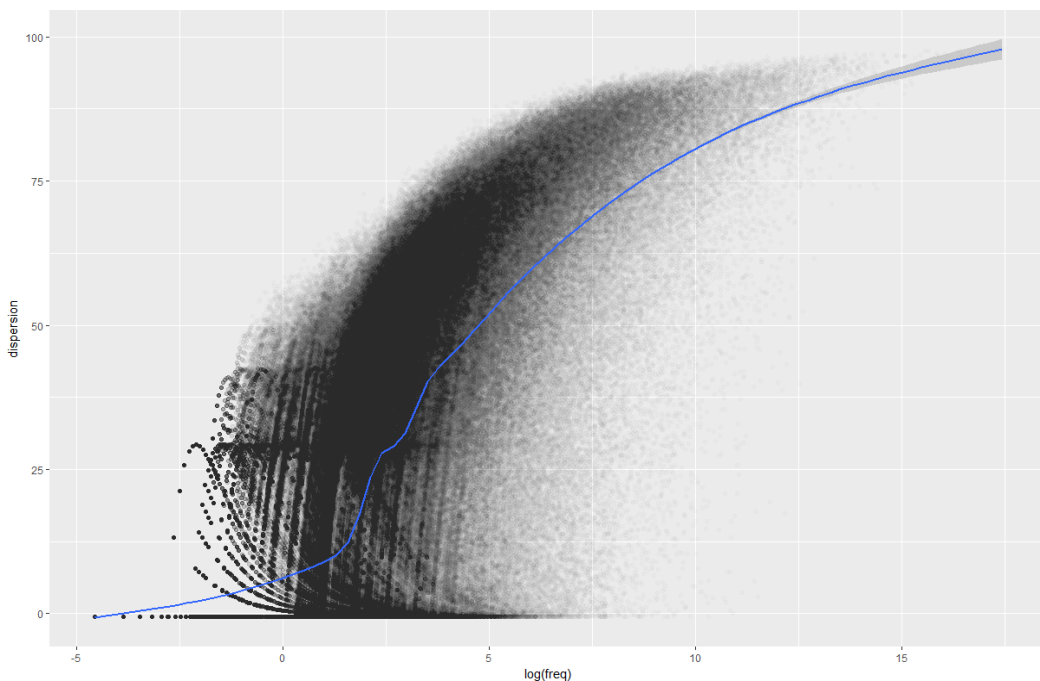
[그림 13] 범위(가로축)와 산포도(세로축)의 관계(청색 선: 회귀 곡선)

[그림 13]에서 보듯이, 산포도의 편차가 상당히 크기는 하나, 범위 수치가 큰 단어들은 산포도가 높으면서 편차가 적은 편이고(막대가 위로 치우쳐 있으면서 짧음), 범위 수치가 작은 단어는 산포도의 편차가 큰 편이다(막대의 길이가 길).

범위와 산포도가 정말 양의 상관관계를 보이는지는 위의 그래프의 회귀 곡선을 보면 더 잘 알 수 있다. 회귀 곡선을 보면, 범위 수치가 중간쯤 되는 구간에서 아래 위로 출렁이는 부분이 있기는 하나, 전체적으로는 좌하에서 우상으로 향하고 있음을 알 수 있다. 국소적으로 좌상에서 우하로 향하는 구간 때문에 상관계수를 낮추기는 했지만, 그럼에도 불구하고 상관계수가 높게 나왔다는 것은 이 두 변수 사이의 관계가 긴밀함을 말해 준다.

두 변수가 전체적으로 양의 상관관계를 보이는 것은 분명한 사실이지만, 범위가 낮은 구간(좌측 부분)에서의 회귀 곡선의 기울기와 범위가 높은 구간(우측 부분)에서의 회귀 곡선의 기울기가 큰 차이가 있음에도 유의할 필요가 있다. 즉 범위가 0~12 정도 되는 구간에서는 범위 점수의 증가에 따라 산포도 점수도 가파르게 증가하는 반면에, 범위 점수가 12 이상 되는 단어들의 경우에는 범위 점수가 증가한다고 해서 산포도 점수가 반드시 증가하는 것도 아니고(회귀 곡선의 출렁임), 증가한다고 해도 매우 완만하게 증가하는 것이다. 즉 범위 점수가 12 이하인 단어들의 경우에는 범위보다는 산포도가 변별력이 있지만, 범위 점수가 일정 수준 이상인 단어들의 경우에는 산포도보다는 범위가 변별력이 있다고 할 수 있다.

다음으로 빈도와 산포도의 관계를 살펴보자.



[그림 14] 빈도(가로축, 로그 변환)와 산포도(세로축)의 관계(청색 선: 회귀 곡선)

빈도에 대해 아무런 조치를 취하지 않고 그냥 그래프를 그리면 빈도가 낮은 단어들이 매우 많고 고빈도 단어들은 매우 적어서, 관측점들이 왼쪽 가장자리에 밀집되어 나오게 된다. 자료의 더 자세한 패턴을 알아보기 위해서는 밀집된 관측점들을 분산시킬 필요가 있다. 이를 위해 빈도에 로그 변형을 하여 그래프를 그렸다.

[그림 14]를 보면, ‘빈도와 산포도가 양의 상관관계를 보이는 패턴’(산포도 0~100에서 좌하~우상으로 향하는 선들)과 ‘빈도와 산포도가 음의 상관관계를 보이는 패턴’(산포도 40 이하에서 좌상~우하로 향하는 선들)이 공존하고 있음을 볼 수 있다. 즉 산포도의 전 구간에서 두 변수 사이에 양의 상관관계가 있지만, 산포도가 일정 수준 이하인 단어들 중 일부에서는 오히려 두 변수 사이에 음의 상관관계가 있는 것이다. 특히 산포도가 거의 0인 단어들 가운데는 빈도가 낮은 것부터 높은 것까지 골고루 나타나고 있다.

또한 오른쪽으로 갈수록 회귀 곡선의 기울기가 완만해지기는 하지만, 전반적으로 좌하에서 우상으로 향하고 있음을 알 수 있다. 위의 범위와 산포도의 관계 그래프와 마찬가지로, 빈도 점수가 높아질수록(우측으로 갈수록) 회귀 곡선의 기울기가 급격히 완만해지므로, 빈도가 매우 높은 단어들의 경우에는 산포도가 별로 변별력이 없다고 할 수 있다.

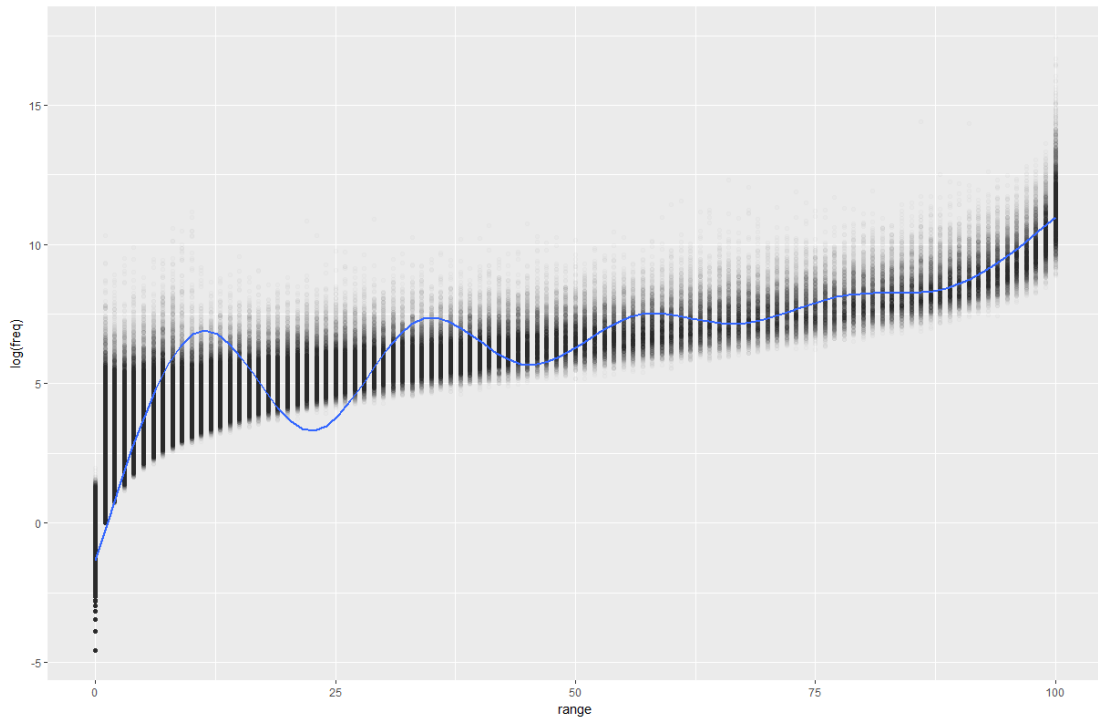
산포도 40 이하인 단어들의 패턴은 더 자세히 살펴볼 필요가 있다. 이 구간에서도 음의 상관관계를 보이는 패턴(좌상에서 우하로 향하는 선들)이 눈에 분명히 보이는 것은 하나, 그에 못지않게 양의 상관관계를 보이는 패턴(좌하에서 우상으로 향하는 선들)도 꽤 강하게 존재하기 때문에, 전체적인 회귀 곡선은 좌하에서 우상으로 향하는 형태로 나타난다. 그리고 위의 그래프에 비해 기울기가 훨씬 완만함에 주목할 필요가 있다.

산포도와 빈도가 일정 수준 이하인 단어들에서 빈도와 산포도 사이에 음의 상관관계의 패턴이 나타나는 원인은, 특정 장르에서만 엄청나게 고빈도로 나타나서 말뭉치 전체의 빈도도 어느 정도는 되지만 특정 장르에 편중되어 있기 때문에 산포도 점수는 낮은 단어들(편중 단어)이 꽤 있기 때문이다. 예컨대 ‘하나님, 교회, 예수’와 같은 단어는 잡지 ‘기독교사상’에서 매우 높은 순위를 차지하며(각각 12위, 19위, 23위) 이로 인해 전체 말뭉치에서의 순위도 어느 정도 되지만(각각 521위, 641위, 1095위) 산포도는 그리 높지 않다(각각 43.41, 56.12, 49.45).

기초 어휘 선정시 빈도에 매우 높은 가중치를 부여하고 산포도에 매우 낮은 가중치를 부여한다면, 이러한 편중 단어들이 높은 순위를 차지할 가능성이 생기고, 가중치를 반대로 부여한다면 편중 단어들의 순위가 매우 낮아지게 된다. 둘 중 어느 쪽이 항상 옳다고 하기는 어렵고, 편중 단어들을 얼마나 중요시할 것인가에 따라 선택이 달라질 수 있다. 특정 장르에 편중되어 있는 단어라도 빈도가 매우 높고 사고 도구어로서 중요한 기능을 한다면 중시할 수도 있다. 고학년에 올라갈수록 교육에

서 이런 단어가 중요한 기능을 하게 된다. 반면에 그런 편중 단어는 일상생활에서 흔히 쓰이는 기초 단어는 아닐 가능성이 높다.

범위와 빈도의 관계를 그래프로 그리면 다음과 같다.



[그림 15] 범위(가로축)와 빈도(세로축, 로그 변형)의 관계(청색 선: 회귀 곡선)

대부분의 단어가 저빈도에 너무 몰려 있으므로, 빈도에 아무런 조치를 취하지 않으면 관측점들이 아래쪽에 극단적으로 몰리게 된다. 그래서 관측점들을 분산시키기 위해 로그 변형을 하였다. 두 변수가 음의 상관관계를 보이는 구간이 있기는 하나, 전반적으로는 회귀 곡선이 좌하에서 우상으로 향하고 있다. 전자의 구간 때문에 상관계수를 많이 까먹어서 두 변수 사이의 상관계수가 매우 낮게 나온 것으로 보인다. 범위 0~12 구간에서는 회귀 곡선의 기울기가 가파르다. 즉 범위가 증가함에 따라 빈도도 가파르게 증가한다. 반면에 범위가 12 이상인 구간에서는 회귀 곡선이 하강하는 구간도 있고, 상승한다 하더라도 그 기울기가 매우 완만하다. 즉 범위 12 이하의 단어들의 경우에는 빈도가 매우 변별력이 높은 변수이지만, 범위 12 이상인 단어들의 경우에는 빈도보다는 범위가 더 변별력이 높다고 할 수 있다. 그러나 범위가 85를 넘어서면 다시 회귀 곡선의 기울기가 가팔라져서, 빈도의 변별력이 다시 높아짐을 알 수 있다.

세 변수 사이의 상관관계가 매우 높다면 셋 중 어느 하나만 조사해도 되겠지만, 이들 사이의 상관관계가 낮다는 것은 이 세 변수가 서로 상당히 독립적이며 이들

변수를 모두 조사할 필요가 있음을 의미한다. 만약 어떤 사정에 의해 2개의 변수만 고려해야 한다면, 나머지 두 변수와 상관관계가 별로 없는 빈도는 반드시 포함되어야 하며 서로 상관관계가 비교적 높은 범위와 산포도 중 하나를 선택하면 될 것이다. 물론 세 변수를 모두 고려하는 것이 가장 이상적이다.

또한 위의 고찰은 빈도·범위·산포도 세 변수에 가중치를 어떻게 부여할 것인지에 대해서도 시사하는 바가 있다. 가중치를 모든 구간에 대해 일률적으로 부여하기보다는, 변별력을 고려하여 차별적으로 부여할 필요가 있음을 보여 준다. 세 변수의 값이 낮은 구간과 높은 구간에서 변별력에 차이가 있기 때문이다.

#### 1.4. 변수들에 가중치를 부여하여 단어 순위 결정

세 변수에 가중치를 어떻게 부여하는 것이 가장 좋을지 알아보려면, 가중치를 다양하게 부여해서 점수를 산출해 보고 이 점수에 따라 단어들을 배열하여, 이들 중 어느 것이 가장 이상적인지 알아보아야 할 것이다. 이를 위해서는 세 변수에 가중치를 부여하여 얻은 점수에 따른 배열 결과를 비교할 기준이 필요하다. 이 기준을 마련하는 방법에 대해서도 많은 논의가 필요하겠으나, 가장 손쉽게 생각하면 일반인들의 직관을 바탕으로 각 단어의 기본성(basicness)을 판단하여 배열한 단어 목록을 생각할 수 있다.

이 기준 단어 목록을 만들기 위해, 우선 빈도 1위~50500위 사이의 단어 100개를 random하게 추출하였다. 이 100개의 단어를 추출할 때, 각 구간의 크기는, 등위가 내려갈수록 점점 더 크게 하였다. 사람이 직관에 의해 단어의 기본성을 판단할 때(즉 이 단어가 저 단어보다 더 기본적인 단어라고 판단할 때), 등위가 높은 단어들끼리 판단하는 것이 낮은 단어들끼리 판단하는 것에 비해 상대적으로 쉬울 것이라고 생각하기 때문이다. 각 구간(1~10위, 11~30위, 31~60위, 61~100위, 101~150위, 48511~49500위, 49501~50500위)에서 1개씩 random하게 추출하였다. 이런 random한 100 단어 목록은 얼마든지 추출할 수 있으나, 일단 3개의 목록만 추출하였다. 이 3개의 목록의 앞 부분을 보이면 다음과 같다.

- 1: 연번
- 2: 하한
- 3: 구간 크기
- 4: 상한
- 5: 빈도 등위
- 6: 단어

	1	2	3	4	5	6
1	1	0	10	10	6	수__02/NNB↓
2	2	10	20	30	20	갈/VA↓
3	3	30	30	60	54	또/MAG↓
4	4	60	40	100	97	마음__01/NNG↓
5	5	100	50	150	100	한국__05/NNP↓
6	6	150	60	210	154	새롭/VA↓
7	7	210	70	280	273	남자__02/NNG↓
8	8	280	80	360	355	앉/VV↓
9	9	360	90	450	426	이루__01/VV↓
10	10	450	100	550	505	가치__06/NNG↓
11	11	550	110	660	598	못하/VV↓
12	12	660	120	780	765	설명/NNG↓
13	13	780	130	910	814	조금__01/NNG↓
14	14	910	140	1050	961	기록하/VV↓
15	15	1050	150	1200	1126	기__13/NNG↓
16	16	1200	160	1360	1355	사이즈__01/NNG↓
17	17	1360	170	1530	1409	동생__01/NNG↓
18	18	1530	180	1710	1641	불편하__01/VA↓
19	19	1710	190	1900	1822	출발하/VV↓
20	20	1900	200	2100	2068	과제__04/NNG↓
21	21	2100	210	2310	2284	시내__03/NNG↓
22	22	2310	220	2530	2461	정확히__01/MAG↓
23	23	2530	230	2760	2730	용__09/NNG↓
24	24	2760	240	3000	2823	성인__01/NNG↓
25	25	3000	250	3250	3195	만__07/MM↓
26	26	3250	260	3510	3502	연방__09/NNG↓
27	27	3510	270	3780	3580	단백질/NNG↓
28	28	3780	280	4060	3884	사망__04/NNG↓
29	29	4060	290	4350	4229	굳/VA↓
30	30	4350	300	4650	4393	솔로__01/NNG↓
31	31	4650	310	4960	4719	겪/VV↓
32	32	4960	320	5280	5136	곤란하/VA↓
33	33	5280	330	5610	5406	씩우__01/VV↓
34	34	5610	340	5950	5689	아버님/NNG↓

[그림 16] random 100 단어 목록 1

	1	2	3	4	5	6
1	1	0	10	10	4	있__01/VA↓
2	2	10	20	30	21	때__01/NNG↓
3	3	30	30	60	41	크__01/VA↓
4	4	60	40	100	77	만__06/NR↓
5	5	100	50	150	138	만나/VV↓
6	6	150	60	210	182	시__10/NNB↓
7	7	210	70	280	211	아직__01/MAG↓
8	8	280	80	360	356	역사__04/NNG↓
9	9	360	90	450	386	아들/NNG↓
10	10	450	100	550	451	나__01/VX↓
11	11	550	110	660	566	하/XSA↓
12	12	660	120	780	747	모두__01/NNG↓
13	13	780	130	910	851	컵/NNG↓
14	14	910	140	1050	1034	한번/NNG↓
15	15	1050	150	1200	1070	달리__04/VV↓
16	16	1200	160	1360	1256	국회/NNG↓
17	17	1360	170	1530	1464	계시/VV↓
18	18	1530	180	1710	1670	무료__01/NNG↓
19	19	1710	190	1900	1759	성분__01/NNG↓
20	20	1900	200	2100	2051	자세히/MAG↓
21	21	2100	210	2310	2256	기르/VV↓
22	22	2310	220	2530	2467	흥미롭/VA↓
23	23	2530	230	2760	2723	유명__01/NNG↓
24	24	2760	240	3000	2951	부족__01/NNG↓
25	25	3000	250	3250	3019	수입__01/NNG↓
26	26	3250	260	3510	3338	감소__01/NNG↓
27	27	3510	270	3780	3760	여유롭/VA↓
28	28	3780	280	4060	3870	정체성__01/NNG↓
29	29	4060	290	4350	4131	측정하__01/VV↓
30	30	4350	300	4650	4414	엄격하__02/VA↓
31	31	4650	310	4960	4940	한우__03/NNG↓
32	32	4960	320	5280	5248	증세__01/NNG↓
33	33	5280	330	5610	5319	툼/NNP↓
34	34	5610	340	5950	5741	광범위하/VA↓

[그림 17] random 100 단어 목록 2



	1	2	3	4	5	6
1	1	0	10	10	2	들__09/XSN↓
2	2	10	20	30	17	그__01/NP↓
3	3	30	30	60	33	때문/NNB↓
4	4	60	40	100	64	생각하/VV↓
5	5	100	50	150	100	한국__05/NNP↓
6	6	150	60	210	167	처음/NNG↓
7	7	210	70	280	223	사랑__01/NNG↓
8	8	280	80	360	314	참__01/MAG↓
9	9	360	90	450	397	도시__03/NNG↓
10	10	450	100	550	528	개인__02/NNG↓
11	11	550	110	660	633	역__14/NNG↓
12	12	660	120	780	699	이웃/NNG↓
13	13	780	130	910	828	진행하/VV↓
14	14	910	140	1050	1026	궁금하__01/VA↓
15	15	1050	150	1200	1053	바탕__01/NNG↓
16	16	1200	160	1360	1243	돕/VV↓
17	17	1360	170	1530	1394	값/NNG↓
18	18	1530	180	1710	1698	식품__01/NNG↓
19	19	1710	190	1900	1714	민주당/NNP↓
20	20	1900	200	2100	2011	의자__03/NNG↓
21	21	2100	210	2310	2123	솔루션/NNG↓
22	22	2310	220	2530	2451	구매하__02/VV↓
23	23	2530	230	2760	2530	질__08/NNG↓
24	24	2760	240	3000	2930	이사__11/NNG↓
25	25	3000	250	3250	3005	판결/NNG↓
26	26	3250	260	3510	3500	하__05/NNP↓
27	27	3510	270	3780	3664	교__88/NNG↓
28	28	3780	280	4060	4022	정상적/NNG↓
29	29	4060	290	4350	4115	실시간/NNG↓
30	30	4350	300	4650	4403	노트북/NNG↓
31	31	4650	310	4960	4680	명절__01/NNG↓
32	32	4960	320	5280	4992	작업실/NNG↓
33	33	5280	330	5610	5446	국어__01/NNG↓
34	34	5610	340	5950	5841	수확__02/NNG↓

[그림 18] random 100 단어 목록 3



이 가운데 목록 1에 대해, 연구진 중 1명이 직관을 바탕으로 중요도/기본성 순서에 따라 배열해 보도록 하였다. 100개 단어에 대해 산출한 이러한 직관 순위와 기타 변수를 통합하여 다음과 같은 자료를 만들었다.

- 1: 단어
- 2: 빈도 순위
- 3: 직관에 따라 매긴 순위
- 4: 빈도
- 5: 범위
- 6: 산포도

	1	2	3	4	5	6
1	단어	빈도순위	직관순위	빈도	범위	산포도↓
2	가치__06/NNG	10	24	276358.1121541043	100	91.65332470062805↓
3	갈/VA	2	2	3407444.1643081196	100	96.6929428576926↓
4	거미줄/NNG	53	29	7230.490900495131	94	87.77988673189553↓
5	계획서/NNG	62	44	4572.867336572876	85	82.81158199817449↓
6	고성능/NNG	41	43	14293.268603997352	87	66.95157117243377↓
7	고집스럽/VA	66	36	3781.495082177106	93	91.53555114802569↓
8	곤란하/VA	32	18	26677.251048252096	100	93.94366570055503↓
9	과제__04/NNG	20	14	76864.56558001183	100	88.72503152646249↓
10	관리직/NNG	81	45	1912.7103212745494	70	78.41779115307007↓
11	교주__06/NNG	64	72	4037.9208676354424	82	78.79384739990974↓
12	굳/VA	29	37	33991.884171994236	99	90.6915319683782↓
13	기__13/NNG	15	53	142383.17111883784	100	91.40477695454115↓
14	기거하__02/VV	65	73	3860.278861404104	90	88.4867727393346↓
15	기록하/VV	14	38	163850.52716934998	100	87.33689899818361↓
16	김__04/NNB	40	12	15742.11653150019	99	90.35465554851432↓
17	깨기/NNG	96	96	1064.0424949693727	78	76.38071044354783↓
18	껌/VV	31	19	29590.716166837825	99	91.30758403604072↓
19	끼얌/VV	42	51	13580.795529674999	96	53.31713557726658↓
20	남자__02/NNG	7	6	473818.0025193209	100	91.41131993026144↓
21	누릴/VA	45	15	11557.894360848526	93	85.0148068862439↓
22	단백질/NNG	27	42	41896.44982740568	92	75.03264110326684↓
23	대표성/NNG	74	46	2576.1723789999483	73	81.33030008723162↓
24	도모하/VV	38	49	18103.74053480715	99	89.21062058786491↓
25	동생__01/NNG	17	8	113443.86908638106	99	91.19833998185864↓
26	동의어/NNG	71	82	2931.8067752312586	93	89.33394586054193↓
27	또/MAG	3	4	1415163.1884241782	100	96.0519631883795↓
28	라니/NNP	82	99	1884.5491782489432	58	39.9866044231621↓
29	러스트/NNP	90	97	1359.0505742167147	52	66.79930684010846↓
30	류현진/NNP	54	92	6808.455950352933	47	53.38161801965742↓
31	마음__01/NNG	4	7	1040155.781978277	100	92.05689258793223↓
32	만__07/MM	25	39	47812.206080281416	100	91.27782547143983↓
33	매독__02/NNG	75	87	2551.3094340009393	72	73.22286345541914↓
34	먹통__01/NNG	88	47	1438.1080582451211	82	76.77974603870703↓
35	며느라/NNG	47	91	9977.122409949297	80	72.7538486982579↓

[그림 19] 100 단어 목록 1에 대한 정보 통합 결과

이 순위표의 단어 등위와 세 변수에 적절히 가중치를 부여하여 얻어지는 weighted sum에 따른 단어 등위를 비교하여, 가중치를 어떻게 부여하면 직관 등위와 가장 일치하는지를 알아보려는 것이다. 이러한 조사의 결과는 다음과 같다.

<표 32> 1명의 직관에 따른 등위와 가중치 부여 점수의 비교

변수1	변수2	Spearman 상관계수
직관 등위	빈도 등위	0.7249325
직관 등위	빈도	0.7249325
<b>직관 등위</b>	<b>범위</b>	<b>0.802154</b>
직관 등위	산포도	0.6966577
직관 등위	0.4*빈도 + 0.3*범위 + 0.3*산포도	0.7980087
직관 등위	0.5*빈도 + 0.25*범위 + 0.25*산포도	0.7987039
직관 등위	0.6*빈도 + 0.2*범위 + 0.2*산포도	0.7994959
직관 등위	0.3*빈도 + 0.4*범위 + 0.3*산포도	0.8022082
직관 등위	0.25*빈도 + 0.5*범위 + 0.25*산포도	0.8058206
직관 등위	0.2*빈도 + 0.6*범위 + 0.2*산포도	0.8109091
<b>직관 등위</b>	<b>0.2*빈도 + 0.7*범위 + 0.1*산포도</b>	<b>0.8157336</b>
직관 등위	0.1*빈도 + 0.8*범위 + 0.1*산포도	0.8150735
직관 등위	0.15*빈도 + 0.7*범위 + 0.15*산포도	0.8142814

Spearman 상관 계수는 변수의 실제 값이 아니라 등위만 고려한다. 그리고 세 변수들의 weighted sum(변수2)을 구할 때, 세 변수의 규모가 다르므로 R의 scale 함수를 이용하여 normalize한 뒤에 weighted sum을 계산하였다. 빈도·범위·산포도의 세 변수를 단순히 직관에 따른 등위와 비교하였을 때, 세 변수 중 범위와의 상관관계가 가장 높게 나타났다. 이에 따라 세 변수 중 범위에 가장 큰 가중치를 부여하고 나머지 두 변수에는 작은 가중치를 부여하여 얻어지는 weighted sum이 직관 등위와 높은 상관관계를 보인다. 다양한 가중치 부여 결과를 비교하여, [0.2\*빈도 + 0.7\*범위 + 0.1\*산포도]가 직관 등위와 가장 높은 상관관계를 보임을 알 수 있었다.

그런데 위의 결과는 1명의 직관을 바탕으로 하여 얻어진 결과이므로 객관성을 갖추었다고 하기 어렵다. 이에 연구진 6명이 3개의 단어 목록에 대해 직관을 바탕으로 등위를 부여하여 위와 같은 실험을 다시 실시하였다. 6명이 부여한 등위를 평균하여 이 평균값에 따라 100개의 단어를 배열하는 것이다. 단어 목록 1에 대해 이렇게 배열한 결과는 다음과 같다.

	A	B	C	D	E	F
1	단어	빈도	범위	산포도	등위평균	직관등위
2	가치__06/NNG	276358.112154104	100	91.6533247006	34.8333	31
3	갈/VA	3407444.16430812	100	96.6929428577	5.66667	4
4	거미줄/NNG	7230.4909004951	94	87.7798867319	28.1667	23
5	계획서/NNG	4572.8673365729	85	82.8115819982	44.5	42
6	고성능/NNG	14293.2686039974	87	66.9515711724	56.6667	58
7	고집스럽/VA	3781.4950821771	93	91.535551148	32.3333	28
8	곤란하/VA	26677.2510482521	100	93.9436657006	38.8333	37
9	과제__04/NNG	76864.5655800118	100	88.7250315265	40.3333	38
10	관리직/NNG	1912.7103212746	70	78.4177911531	59.1667	60
11	교주__06/NNG	4037.9208676355	82	78.7938473999	72.6667	78
12	굳/VA	33991.8841719942	99	90.6915319684	23.1667	19
13	기__13/NNG	142383.171118838	100	91.4047769545	50.6667	51
14	기거하__02/VV	3860.2788614041	90	88.4867727393	69.8333	75
15	기록하/VV	163850.52716935	100	87.3368989982	38.1667	36
16	김__04/NNB	15742.1165315002	99	90.3546555485	41.6667	39
17	깨기/NNG	1064.0424949694	78	76.3807104435	61.3333	64
18	껌/VV	29590.7161668378	99	91.307584036	22.5	17
19	끼얌/VV	13580.795529675	96	53.3171355773	43.6667	40
20	남자__02/NNG	473818.002519321	100	91.4113199303	3.33333	1
21	누렇/VA	11557.8943608485	93	85.0148068862	20.5	14
22	단백질/NNG	41896.4498274057	92	75.0326411033	61.8333	65
23	대표성/NNG	2576.172379	73	81.3303000872	55	56
24	도모하/VV	18103.7405348072	99	89.2106205879	62.6667	66
25	동생__01/NNG	113443.869086381	99	91.1983399819	4.16667	2
26	동의어/NNG	2931.8067752313	93	89.3339458605	78.1667	83
27	또/MAG	1415163.18842418	100	96.0519631884	6.16667	6
28	라니/NNP	1884.549178249	58	39.9866044232	76.5	81
29	러스트/NNP	1359.0505742167	52	66.7993068401	98.1667	100
30	류현진/NNP	6808.4559503529	47	53.3816180197	93.5	93
31	마음__01/NNG	1040155.78197828	100	92.0568925879	5.66667	5
32	만__07/MM	47812.2060802814	100	91.2778254714	50.3333	50
33	매독__02/NNG	2551.3094340009	72	73.2228634554	85.8333	88
34	먹통__01/NNG	1438.1080582451	82	76.7797460387	55.3333	57
35	명나라/NNG	9977.1224099493	80	72.7538486983	78.8333	84
36	못하/VV	242466.157279456	100	93.922298135	13.5	8
37	문수봉/NNP	1579.4755766655	17	7.2942597635	96.3333	96
38	물난리/NNG	1658.5976815542	73	57.6136517598	50	49

[그림 20] 100 단어 목록 1에 대한 6인이 부여한 등위 통합 결과

단어 목록 3개에 대해 다양한 가중치를 실험한 결과는 다음과 같다.

<표 33> 100 단어 목록 1에 대한 실험 결과

변수1	변수2	Spearman 상관계수
직관 등위	빈도	0.6706751
<b>직관 등위</b>	<b>범위</b>	<b>0.7831403</b>
직관 등위	산포도	0.7061266
직관 등위	0.4*빈도 + 0.3*범위 + 0.3*산포도	0.6688798
직관 등위	0.5*빈도 + 0.25*범위 + 0.25*산포도	0.625656
직관 등위	0.6*빈도 + 0.2*범위 + 0.2*산포도	0.5671431
직관 등위	0.3*빈도 + 0.4*범위 + 0.3*산포도	0.7006287
직관 등위	0.25*빈도 + 0.5*범위 + 0.25*산포도	0.7138504
직관 등위	0.2*빈도 + 0.6*범위 + 0.2*산포도	0.7201523
<b>직관 등위</b>	<b>0.2*빈도 + 0.7*범위 + 0.1*산포도</b>	<b>0.722592</b>
직관 등위	0.1*빈도 + 0.8*범위 + 0.1*산포도	0.7165654
직관 등위	0.15*빈도 + 0.7*범위 + 0.15*산포도	0.7206424

<표 34> 100 단어 목록 2에 대한 실험 결과

변수1	변수2	Spearman 상관계수
직관 등위	빈도	0.2750149
<b>직관 등위</b>	<b>범위</b>	<b>0.677099</b>
직관 등위	산포도	0.5358898
직관 등위	0.4*빈도 + 0.3*범위 + 0.3*산포도	0.6350649
직관 등위	0.5*빈도 + 0.25*범위 + 0.25*산포도	0.5927839
직관 등위	0.6*빈도 + 0.2*범위 + 0.2*산포도	0.5352243
직관 등위	0.3*빈도 + 0.4*범위 + 0.3*산포도	0.6679084
직관 등위	0.25*빈도 + 0.5*범위 + 0.25*산포도	0.6835296
직관 등위	0.2*빈도 + 0.6*범위 + 0.2*산포도	0.6918831
<b>직관 등위</b>	<b>0.2*빈도 + 0.7*범위 + 0.1*산포도</b>	<b>0.6976845</b>
직관 등위	0.1*빈도 + 0.8*범위 + 0.1*산포도	0.6916086
직관 등위	0.15*빈도 + 0.7*범위 + 0.15*산포도	0.694136



<표 35> 100 단어 목록 3에 대한 실험 결과

변수1	변수2	Spearman 상관계수
직관 등위	빈도	0.2825612
<b>직관 등위</b>	<b>범위</b>	<b>0.6646282</b>
직관 등위	산포도	0.566077
직관 등위	0.4*빈도 + 0.3*범위 + 0.3*산포도	0.6394283
직관 등위	0.5*빈도 + 0.25*범위 + 0.25*산포도	0.599292
직관 등위	0.6*빈도 + 0.2*범위 + 0.2*산포도	0.5432708
직관 등위	0.3*빈도 + 0.4*범위 + 0.3*산포도	0.6672772
직관 등위	0.25*빈도 + 0.5*범위 + 0.25*산포도	0.6786746
직관 등위	0.2*빈도 + 0.6*범위 + 0.2*산포도	0.6839466
<b>직관 등위</b>	<b>0.2*빈도 + 0.7*범위 + 0.1*산포도</b>	<b>0.6879461</b>
직관 등위	0.1*빈도 + 0.8*범위 + 0.1*산포도	0.6802731
직관 등위	0.15*빈도 + 0.7*범위 + 0.15*산포도	0.6841112

각 실험에서 구체적인 수치는 약간의 차이는 있으나, 모두 일관되게 세 변수 중 범위와의 상관관계가 가장 높게 나왔고, 다양한 weighted sum 가운데 [0.2\*빈도 + 0.7\*범위 + 0.1\*산포도]와 같은 가중치가 직관 등위와 가장 높은 상관관계를 보였다. 따라서 단어 등급화에 이 weighted sum을 이용하기로 결정하였다.

이 식에 따라 각 단어의 점수를 구하여 이 점수에 따라 정렬한 결과는 다음과 같다.

1	2	3	4	5
1 단어,	빈도,	범위,	산포도,	점수=0.2*빈도+0.7*범위+0.1*산포도
2 이/VCP,	37453589.37048225,	100,	97.91168445837461,	10.917236766033596↓
3 것_01/NNB,	17321134.381407566,	100,	97.1353364461081,	7.896761054124062↓
4 들_09/XSN,	14411994.588581333,	100,	96.79985763802394,	7.459335089513478↓
5 하_01/VV,	13611982.89881008,	100,	96.46163209799118,	7.337977788664442↓
6 있_01/VA,	13293198.987567576,	100,	97.55577888606076,	7.294943168212786↓
7 수_01/VX,	9439647.91418627,	100,	97.11971731433736,	6.7155490851389334↓
8 우_02/NNB,	7897218.146092479,	100,	96.77145949762249,	6.4828872242950055↓
9 되_01/VV,	7601453.126661401,	100,	97.1139049461496,	6.440046264859098↓
10 하_01/VX,	6331603.328728959,	100,	97.58000976570487,	6.251761536540967↓
11 알/VX,	5765145.55906604,	100,	97.60526878368962,	6.1669798065726935↓
12 없_01/VA,	5128633.710224747,	100,	96.86039529651202,	6.068363694606207↓
13 그_01/MM,	5003000.740099675,	100,	94.22767636136007,	6.0381322228364↓
14 보_01/VV,	4720936.4623573385,	100,	95.09741953140339,	5.999628509335308↓
15 이_05/MM,	4515948.791985744,	100,	96.3232406130526,	5.97421810303891↓
16 주_01/VX,	4104260.5412761383,	100,	91.68707160205717,	5.892439469786606↓
17 년_02/NNB,	4031396.655156952,	100,	92.41200641025212,	5.884659941154138↓
18 사람/NNG,	3675291.7517757164,	100,	94.60715023483269,	5.840801289361156↓
19 아나/VCN,	3534725.405859018,	100,	96.64294082689419,	5.828553678225008↓
20 갈/VA,	3407444.1643081196,	100,	96.6929428576926,	5.809695497261425↓
21 그_01/NP,	3553930.9195911456,	100,	91.50372445990051,	5.809171169358395↓
22 때_01/NNG,	3402034.2011654675,	100,	96.68317564688775,	5.80884243619128↓
23 나_03/NP,	3489093.1659047366,	100,	90.14068442538168,	5.793550321191976↓
24 보_01/VX,	3383820.0048366017,	100,	92.92837659015434,	5.78984874729743↓
25 한_01/MM,	2915412.9751448943,	100,	96.70222424705834,	5.735998360726915↓
26 지_04/VX,	2868520.6139228893,	100,	96.83119199295825,	5.729529552810628↓
27 등_05/NNB,	2897570.5914934683,	100,	92.09151444889113,	5.713352983532642↓
28 줄_01/VA,	2719953.0952579714,	100,	92.28245283578762,	5.687561724479185↓
29 대하_02/VV,	2665759.5080093984,	100,	94.01179333403859,	5.686930806197562↓
30 가_01/VV,	2585176.77012815,	100,	93.6682859758379,	5.673366509084565↓
31 우리_03/NP,	2447510.159288957,	100,	93.95763397235274,	5.653988680080458↓
32 만들/VV,	2416263.907959582,	100,	92.09762084093647,	5.641249302438438↓
33 위하_01/VV,	2310453.4263258358,	100,	95.69948813700994,	5.6409938339596994↓
34 더_01/MAG,	2292376.6866783467,	100,	95.61645642947927,	5.637925142045724↓
35 때론/NNB,	2196356.0796374925,	100,	95.83741793105409,	5.624492295307179↓
36 많/VA,	2122948.192297459,	100,	96.67827168833306,	5.61713334819692↓
37 받_01/VV,	2097003.3051201273,	100,	96.99968825727436,	5.614637393785918↓

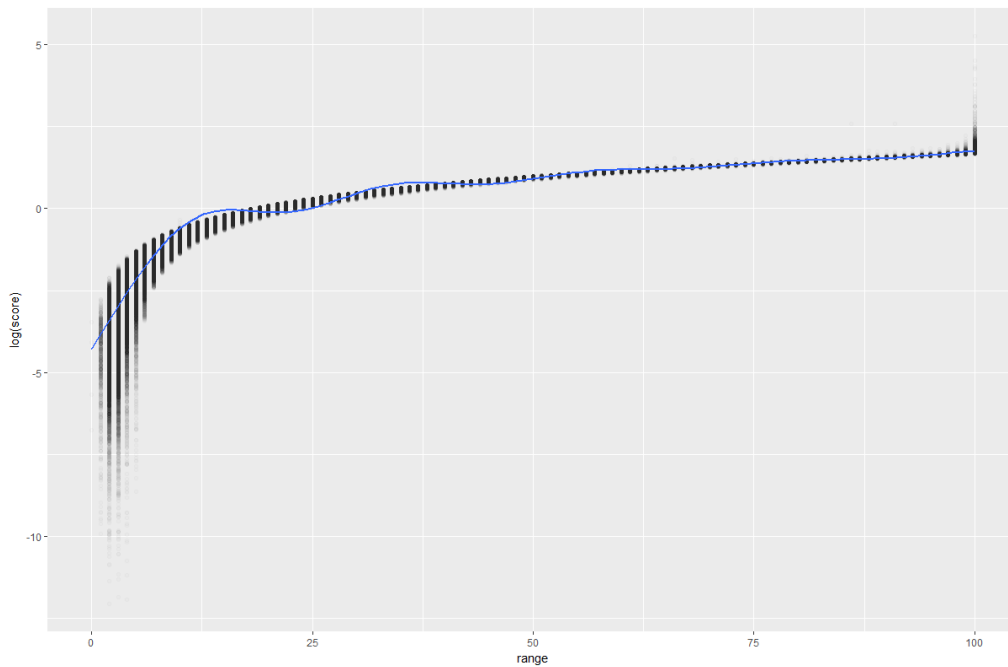
[그림 21] weighted sum(=0.2\*빈도+0.7\*범위+0.1\*산포도)에 따른 소팅 결과

이 weighted sum, 빈도, 범위, 산포도 사이의 관계는 다음과 같다.

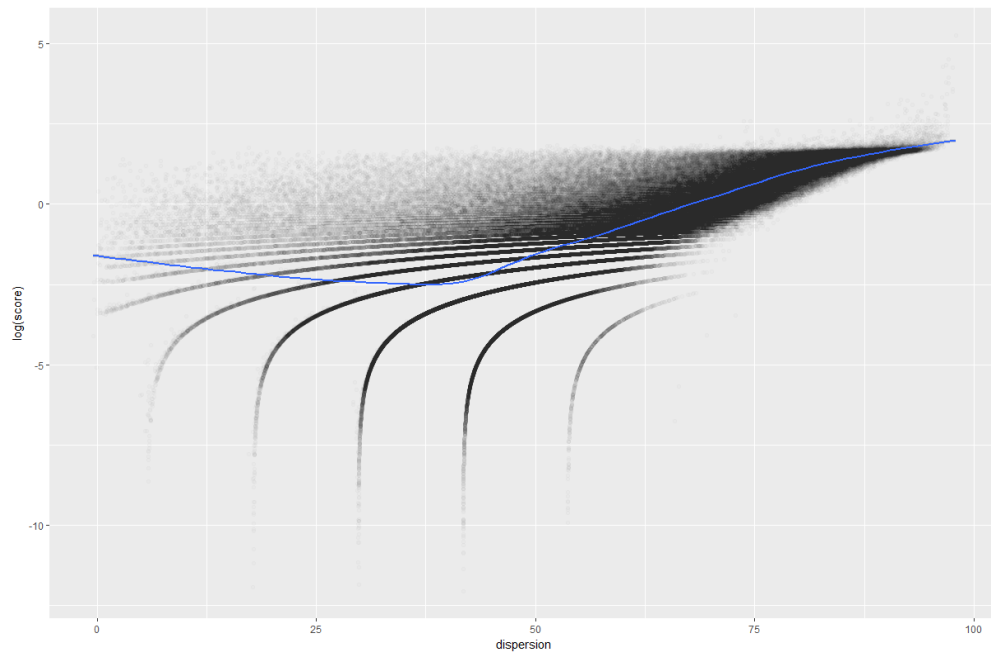
<표 36> 점수(결정된 weighted sum), 빈도, 범위, 산포도 사이의 관계

변수1	변수2	Person 상관계수
점수	빈도	0.3325793
<b>점수</b>	<b>범위</b>	<b>0.9660282</b>
점수	산포도	0.7157315
빈도	범위	0.09476222
빈도	산포도	0.0448866
<b>범위</b>	<b>산포도</b>	<b>0.6769272</b>

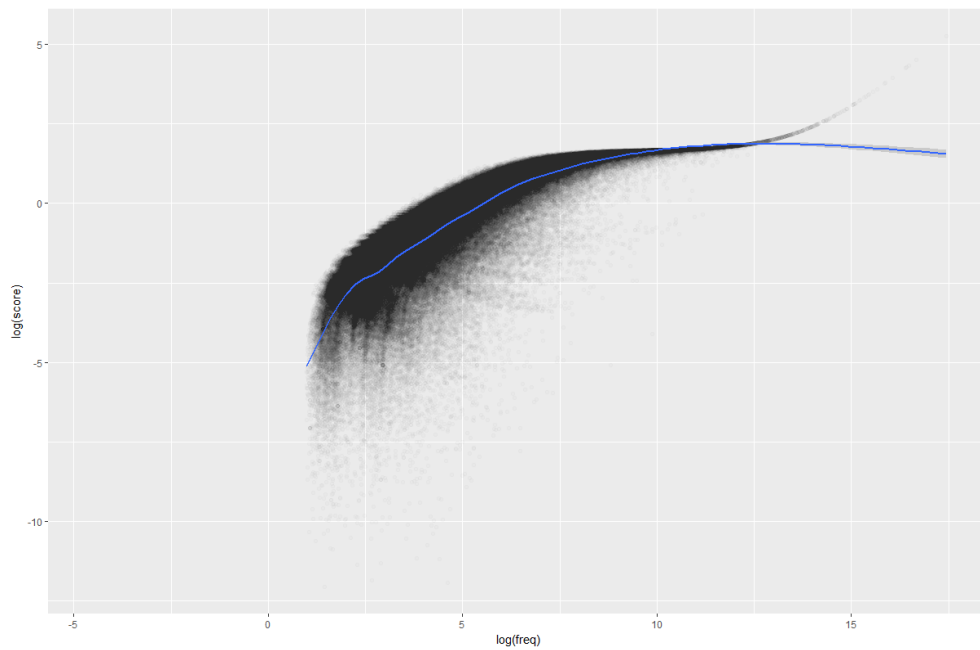
점수를 산출할 때 세 변수 중 범위에 가장 큰 가중치를 부여했으므로 점수와 범위 사이의 상관관계가 가장 높게 나오는 것이 당연하다. 세 변수와 점수 사이의 관계를 그래프로 그려 보면 다음과 같다.



[그림 22] 범위와 점수(로그 변형)의 관계



[그림 23] 산포도와 점수(로그 변형)의 관계



[그림 24] 빈도(로그 변형)와 점수(로그 변형)의 관계

빈도-점수 그래프에서 관측점들이 일정한 선을 상한선으로 하여 그 아래로는 넓게 분포해 있으나, 이 선 위로는 넘어가지 못하는 양상을 보인다. 이는 빈도에 0.2라는 낮은 가중치를 부여했기 때문에, 빈도가 아무리 높아도 점수에 반영되는 데 한계가 있게 된 것이다.

### 1.5. 다른 어휘 목록과의 비교

지금까지의 작업은 순전히 말뭉치에서 추출한 통계를 바탕으로 하였기 때문에, 기반 말뭉치가 가진 편향성이나 기타 문제를 고스란히 반영하고 있을 것이고, 말뭉치로부터 얻은 수량적 속성 외의 다른 속성은 제대로 반영하지 못할 가능성이 높다. 따라서 말뭉치를 바탕으로 양적 방법에 의해 얻어진 단어 목록을 질적 방법에 의해 얻어진 다른 단어 목록과 비교할 필요가 있다.

이 목적을 위해 본 작업에서는 <연세한국어사전>의 표제어 목록과 비교하였다. <연세한국어사전>도 기본적으로 말뭉치의 빈도를 바탕으로 하였으나, 이를 그대로 사용하지 않고 사전 편찬자들의 직관을 동원하여 빠진 단어를 보충하는 등 질적 방법을 혼용하여 표제어 목록을 선정하였다.<sup>16)</sup> 따라서 본 작업에서 비교 대상으로 삼기에 어느 정도 적합하다고 할 수 있다.

<연세한국어사전>에는 약 5만 개의 표제어가 실려 있으나, 본 작업과 별로 상관이 없는 조사, 어미 등의 문법요소를 제외하고 47,299개의 단어 목록을 추출하였다. <연세한국어사전>의 품사 표시는 말뭉치와의 비교를 위해 UTagger에서 부여되는 세종 말뭉치 방식의 품사 태그로 교체하였다.

그 다음에 두 목록의 비교에 있어서 큰 걸림돌은 동형어 번호이다. UTagger에서 부여되는 동형어 번호는 <표준국어대사전>의 어깨번호이나, <연세한국어사전>의 동형어 번호는 이와 상관없이 독자적으로 붙여진 번호이다. 본 작업을 위해서는, 동형어들 중 말뭉치 빈도순에 따라 <연세한국어사전>의 어깨번호와 대응시켰다. 즉 <연세한국어사전>에 ‘가격1’과 ‘가격2’라는 동형의 명사가 2개 실려 있는데, 두 단어 중 말뭉치에서 빈도가 더 높은 것을 ‘가격1’에 대응시키고, 빈도가 그 다음인 것을 ‘가격2’에 대응시킨 것이다.<sup>17)</sup>

이렇게 대응시킨 결과의 일부를 보이면 다음과 같다.

- |  |
|--|
| 1: <연세한국어사전> 표제어, 어깨번호, 품사<br>2: 점수 등위<br>3: 말뭉치에서 UTagger가 부여한 동형어 번호, 품사 태그<br>4: 빈도<br>5: 범위<br>6: 산포도<br>7: 점수 |
|--|

16) 당시 기술 수준의 한계로 동형어를 구별하여 빈도를 추출한 것은 아니라고 한다. 그러나 정성적 방법으로 그러한 한계와 문제점을 어느 정도는 극복한 것으로 판단된다.

17) <연세한국어사전>에서 동형어들을 배열할 때 사용 빈도를 어느 정도 고려하여 순서를 정한 것으로 보인다. 그래서 본 연구에서 말뭉치로부터 추출한 동형어들의 사용 빈도 순위와 대체로 일치한다.



	1	2	3	4	5	6	7
1	가__1/NNG	1713	가__01/NNG	27675.55444744633	100	91.96563532696686	5.2827162501352545↓
2	가__2/NNG	6737	가__88/NNG	22982.049022871066	97	83.86034907775097	5.0920519998920435↓
3	가__4/XP	9245	가__15/XP	17871.306179588395	94	89.59601149186803	4.961277806767393↓
4	가__5/XSN	1108	가__16/XSN	95004.01109021797	100	92.20209030390536	5.293830517651961↓
5	가__6/XSN	2559	가__13/XSN	38943.94060696617	100	86.48142929952573	5.26064992281574↓
6	가__7/XSN	2615	가__12/XSN	21551.685565494674	100	86.47525071894965	5.258016702101756↓
7	가__8/XSN	2634	가__18/XSN	32688.25380828662	100	85.8644448056726	5.257039934587116↓
8	가__9/XSN	11329	가__14/XSN	3559.5818397869502	92	88.63107613069249	4.85171834199833↓
9	가__10/XSN	22065	가__17/XSN	1421.3997665817385	80	82.87319734910554	4.207047372783703↓
10	가가호호/NNG	93732	가가호호/NNG	151.05072497444985	23	70.48273669877432	1.2109894299016657↓
11	가감/NNG	14265	가감__01/NNG	3114.1325246706683	89	86.99015083072437	4.68969133946271↓
12	가감승제/NNG	146381	가감승제/NNG	51.74232570030524	11	65.15043356926252	0.5684674095168667↓
13	가감하/VV	25590	가감하__01/VV	1656.5608206338172	77	66.46914026812897	3.981175312440321↓
14	가건물/NNG	33444	가건물/NNG	1087.1797653750816	66	83.81133679783005	3.488415792537821↓
15	가계/NNG	2284	가계/NNG	107162.47054568012	100	86.58866387394359	5.271337856516343↓
16	가갯집/NNG	86593	가갯집/NNG	191.4959721590321	26	69.45726424409183	1.3614061358813134↓
17	가갯__1/NNG	389	가갯__03/NNG	462572.9039358211	100	87.24960665999018	5.327463678588015↓
18	가갯__2/NNG	52695	가갯__01/NNG	633.4294265829609	48	60.42167851852549	2.457919512841076↓
19	가갯표__1/NNG	22105	가갯표/NNG	2413.8663673274864	80	82.1840605677375	4.204211085260832↓
20	가갯표__2/NNG↓						
21	가갯하/VV	33426	가갯하/VV	824.0832384516409	66	84.0186235914253	3.489274234547668↓
22	가결/NNG	39024	가결__01/NNG	1936.933812350343	60	79.38673675428942	3.1596727852852093↓
23	가결되/VV	48579	가결되__01/VV	1662.7455624006527	50	81.14650502967326	2.651079140085207↓
24	가결하/VV	52257	가결하__01/VV	888.9393047799256	47	77.86451947179603	2.4818946152015036↓
25	가계__1/NNG	7830	가계__08/NNG	18297.404740954044	96	83.46269675530532	5.038009992869963↓
26	가계__2/NNG	28619	가계__06/NNG	1458.9844981271003	72	80.86172694517208	3.7854002010817687↓
27	가계부/NNG	24304	가계부/NNG	2559.3049465985955	77	85.61506455960182	4.064241886034078↓
28	가계비/NNG	145405	가계비/NNG	99.58490233548247	13	42.88115600086483	0.5752493838488064↓
29	가계약/NNG	66742	가계약/NNG	233.63595827659404	36	77.16370193948934	1.9109682300061408↓
30	가곡/NNG	23633	가곡__01/NNG	7233.656711190619	80	59.74292420733569	4.107728774612845↓
31	가공__1/NNG	10055	가공__01/NNG	24070.210657751777	94	79.3888482504677	4.917994085586538↓
32	가공__2/NNG	23839	가공__04/NNG	1762.8692705101978	78	80.71393528760545	4.09451061099273↓
33	가공되/VV	15051	가공되/VV	2918.3586758225897	88	87.9741290476414	4.642306632899841↓
34	가공업/NNG	46722	가공업/NNG	533.9075480039737	53	67.34868987566391	2.745996734584141↓
35	가공적__1/NNG	95889	가공적/NNG	228.80833628860398	22	73.06889301802654	1.170585609860107↓
36	가공품/NNG	33122	가공품/NNG	1498.4960021948261	67	75.64855790350973	3.5047375579392845↓
37	가공하__1/VV	10360	가공하__01/VV	7914.445490076528	93	88.48049711910073	4.903336242577906↓
38	가공하__2/VA	18403	가공하__03/VA	3384.678153296777	84	87.94937084041925	4.435799253326499↓
39	가관/NNG	14879	가관__02/NNG	2443.2123591748104	88	89.9907328067907	4.650970420415594↓
40	가교__1/NNG	17483	가교__02/NNG	2405.0269272520977	85	89.3995843187849	4.493551599159362↓
41	가교__2/NNG	112444	가교__06/NNG	46.45991798057435	17	72.47691499538847	0.9099065794376369↓
42	가구__1/NNG	5459	가구__03/NNG	120867.91448054838	99	69.82158396342277	5.149147059424784↓

[그림 25] <연세한국어사전> 표제어와 말뭉치 통계 결과의 비교

비교 대상이 된 <연세한국어사전>의 47,299개 표제어 중, 말뭉치 빈도 순위 몇 위 안에 드는 것이 몇 개인지를 살펴보면 다음과 같다.

<표 37> <연세한국어사전> 표제어와의 말뭉치 순위 비교

말뭉치 순위	표제어 수	말뭉치 순위	표제어 수	말뭉치 순위	표제어 수
< 10,000	11,256	< 90,000	41,220	< 170,000	44,510
< 20,000	19,312	< 100,000	42,051	< 180,000	44,688
< 30,000	25,849	< 110,000	41,618	< 190,000	44,843
< 40,000	30,721	< 120,000	43,066	< 200,000	44,976
< 50,000	34,336	< 130,000	43,487	< 300,000	45,661
< 60,000	36,945	< 140,000	43,818	< 400,000	45,929
< 70,000	38,872	< 150,000	44,080	< 1,000,000	46,352
< 80,000	40,252	< 160,000	44,323	등외	2,274

<표 37>에서 볼 수 있듯이, 말뭉치 순위 4만등 안에 드는 것이 약 3만개, 8만등 안에 드는 것이 약 4만개 있고, 나머지가 1만개 가까이 있음을 알 수 있다. <연세한국어사전>에 실려 있으나 말뭉치에서는 포착되지 않은 것도 2,274개 나타났다. 이들 중 일부는 어깨번호 부여 방식에 있어서 <표준국어대사전>과 <연세한국어사전>이 다른 경우이다. ‘가격표’에 대해 <표준국어대사전>에서는 하나의 표제어만 실려 있으나, <연세한국어사전>에는 어깨번호 1과 2가 달린 2개의 표제어가 실려 있다. 이에 따라 UTagger는 ‘가격표’에 어깨번호를 부여하지 않았고, <연세한국어사전>의 ‘가격표2’는 말뭉치에 대응 단어가 없는 것으로 나타난 것이다.

이러한 비교 결과를 면밀히 검토하여, <연세한국어사전>에 실려 있는데도 말뭉치에서 순위가 지나치게 하위로 나타난 단어들이 어떤 것들인지 알아보고, 이 단어들이 실제로는 보다 높은 순위를 부여 받아야 마땅하다고 판단된다면, 이를 위해 말뭉치를 어떻게 보완해야 할지에 대한 단서를 얻을 수 있다.

## 1.6. 텍스트 포괄 범위 조사

위에서 설명한 방법에 따라 각 단어의 점수를 산출하고 이 점수에 따라 배열함으로써 각 단어의 순위가 부여된다. Zipf의 법칙에 따라 순위가 높은 단어는 빈도가 매우 높으므로, 상위의 소수의 단어들이 텍스트의 token 중 대다수를 차지하게 된다. 이러한 양상을 표로 정리하면 다음과 같다.

<표 38> 텍스트 포괄 범위

Text Coverage	점수 상위 단어 수
90%	26,912
95%	49,426
97%	75,454
98%	103,503
99%	174,659
99.5%	289,306
99.9%	750,525
99.99%	1,691,695
99.999%	1,872,175
99.9999%	1,893,569

말뭉치의 어휘형태소 type 수는 1,895,946개이고, 말뭉치의 어휘형태소 token 수는 2,377,114,012개이다. 상위 26,912개 단어가 말뭉치 전체 token의 90%를 cover하고, 상위 49,426개 단어가 말뭉치 전체 token의 95%를 cover하며, 상위 174,659개 단어가 말뭉치 전체 token의 99%를 cover한다. 대략 텍스트의 token의 95% 정도를 차지하는 약 5만 개의 상위 빈도 단어가 어휘 등급화의 주된 대상이 된다.

## 1.7. 향후 과제

### 1) 실험의 기반이 되는 단어 목록 보완

우선 실험을 더욱 객관화할 필요가 있다. 6명보다 훨씬 더 많은 사람의 판단을 종합할 필요가 있다. 실험 대상이 된 단어 목록의 단어 수도 보완할 필요가 있다. 100개보다 더 많은 단어를 추출하여 실험하면 신뢰도가 더 높아질 수 있다. 단어 목록도 다양화할 필요가 있다. 3개의 단어 목록이 아니라 보다 많은 수의 목록을 만들어서 실험하면 더 좋을 것이다.

## 2) 점수 산출 방식의 다양한 실험

세 변수(빈도, 범위, 산포도)에 다양한 가중치를 부여하여 얻어지는 점수 순위와 직관에 바탕을 둔 단어 순위 사이의 상관관계 고찰하였는데, 가중치 부여 방식을 가급적 다양하게 하여 실험하기는 하였으나, 이를 좀 더 체계적으로 실시할 수 있는 방법론을 모색할 필요가 있다. 미처 시도해 보지 않은 가중치 부여 패턴이 의외의 결과를 낳을지도 모르기 때문이다.

## 3) 점수 최상위 단어들에 대한 정성적 검토

통계적 방법으로 부여된 점수에 따라 배열된 단어들을 최상위부터 검토하여, 혹시 부당하게 높은 등급을 부여받은 단어들이 없는지 알아볼 필요가 있다. 그런 단어가 발견되면 장르별 빈도를 검토하여, 특정 장르가 부당하게 과대평가(over-represent)되지는 않았는지 점검할 필요가 있다.

## 4) 기존 기본 어휘 목록에 대한 정성적 검토

기존 기본 어휘 목록에는 높은 등급으로 올라 있는데, 통계적 방법으로는 낮은 점수/등급을 부여받은 단어들이 혹시 있는지 검토할 필요가 있다.

## 2. 어휘 등급화 절차

지금까지의 검토 결과를 바탕으로 할 때에 어휘 등급화 작업 시 기초 어휘와 나머지 단어들의 선정 절차는 달라야 한다. 이는 첫째 기초 어휘가 갖는 중요성 때문이고 둘째는 기초 어휘가 갖는 독특한 성격 때문이다. 기초 어휘는 아동 언어 발달의 첫 시작을 알리는 단어들이며 국어교육적 중요성이 크다. 또한 한국어교육 시에 첫 교육 대상이 되는 단어들이다. 한국어 전체를 대상으로 어휘론적 입장에서 기초 어휘를 살펴보면 기초 어휘는 독특한 특성을 갖는다. 비유컨대 기초 어휘는 한국어의 뿌리이거나 기반에 해당하는 단어들이다. 한국어 전체 집합이 일정한 층을 형성하고 있다면 기초 어휘는 그 가장 아래에서 한국어 언어생활의 심연을 담당한다.<sup>18)</sup>

기초 어휘가 갖는 이러한 성격을 고려할 때에 기초 어휘 선정은 신중해야 한다. 이러한 점을 고려하여 기초 어휘 선정 및 검증 절차와 나머지 단어들의 선정 및 검증 절차를 차별화할 필요가 있다. 본 연구에서 기초 어휘는 등급 구간을 나누어 숫자를 매긴다면 1등급에 해당한다. 따라서 나머지 단어들은 2등급 이상 단어들에 해당한다. 편의상 1등급 어휘와 2등급 이상 어휘로 칭하도록 한다.<sup>19)</sup>

### 2.1. 1등급 어휘 선정 절차

#### 1) 1등급 어휘 선정용 말뭉치의 별도 구축

1등급 어휘는 대개 구어 의사소통 상황에서 사용되며 범위가 산포도가 높다. 만일 특정 분야에서만 집중적으로 사용된다면 1등급 어휘가 아닐 가능성이 높다. 따라서 1등급 어휘 추출을 위해서는 말뭉치 구성 시 다음과 같은 몇 가지 점에 유의해야 한다.

18) 기초 어휘의 성격은 기초 어휘가 가져야 할 다음과 같은 조건을 통해 명료하게 드러난다. ① 그 어휘를 사용하지 않고 다른 단어를 대용하는 일이 불가능하여 문장을 작성하는 일이 불가능해지며 다른 단어를 대용한다고 하더라도 오히려 그것이 더 불편해진다. ② 그 단어들을 서로 조합하여 다른 복잡한 개념이나 새로운 명명이 필요한 개념 등을 나타내는 단어를 쉽게 만들 수 있다. ③ 기초 어휘에 속하지 않은 단어를 설명하는 경우 결국에는 기초 어휘의 범위에 들어 있는 단어들에 의지하는 일이 대개 가능하다. ④ 그 단어들의 많은 것은 오랜 옛날부터 사용되어 오던 것이며 앞으로도 계속 사용될 가능성이 크다. ⑤ 여러 방면의 화제에 흔하게 사용된다. (田中章夫, 1988:79, 김광해, 1993:48에서 재인용)

19) 등급 구간에 따른 어휘 구분을 할 때 각 구간별 단어 집단을 지칭하는 용어는 신중하게 검토하여 결정하여야 한다. 이 연구에서는 잠정적으로 1등급 어휘와 2등급 이상 어휘로 구분하여 사용하며, 추후 연구 진행에 따라 보다 정확한 용어로 수정할 것이다.

첫째, 현존하는 말뭉치들이 문어적 성격이 농후하다는 점을 고려할 때에 구어 말뭉치의 보강이 요구된다. 앞서 기술한 것처럼 우리는 인터넷이나 잡지 자료 등을 첨가하여 구어적 성격이 높은 말뭉치를 대거 보강하였다. 그러나 추후 실제 구어 자료(일상 대화나 전화 통화 자료 등)의 보강이 요구된다. 1등급 어휘 선정을 위한 말뭉치는 구어와 문어의 비율을 6:4나 7:3 정도로 하여 문어보다 구어의 비율을 더 높게 하는 것이 타당하다. 참고로 네이션과 웹은 기초 어휘 선정을 위해 구어와 문어 비율을 6:4로 한 바 있다(관련 내용은 본 보고서의 II장 참조).

둘째, 다양한 국어 생활 자료를 포함하여 국어 장르가 골고루 들어가도록 해야 한다. 이는 범위나 산포도가 높은 단어를 선택해야만 1등급 어휘의 성격을 지니기 때문이다. 영어의 BNC 14000의 경우 어휘 목록 개발 시 빈도가 아니라 범위를 제1기준으로 선정한 바 있다. 특정 자료에서만 집중적으로 사용되는 단어들을 제외하기 위해서는 빈도와 산포도를 적극 고려할 필요가 있다. 물론 빈도가 높은 단어들이 텍스트 포괄 범위가 높다는 점에서 빈도에 대한 고려도 필요하다. 빈도, 범위, 산포도 중 어디에 우선 순위를 두어야 하는지는 실제 말뭉치 통계 자료에 대한 검토를 통해 결정될 것이다. 분명한 것은 1등급 어휘 선정 시에는 범위와 산포도의 비중을 크게 늘려야 한다는 점이다.

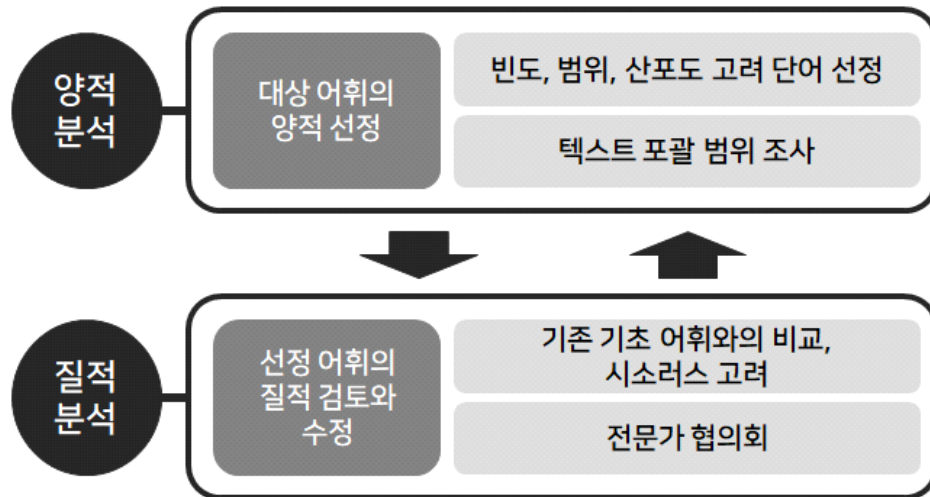
셋째, 1등급 어휘는 대개 사용 빈도가 매우 높아 말뭉치 규모와 상관없이 등장한다는 점을 고려할 필요가 있다. 영어의 Basic English가 850 단어에 불과하며 GSL(General Service List)이나 기 선정된 한국어 기초 어휘들이 2000여 단어밖에 안 되는 점을 생각하면 이 단어들은 대개의 일상생활에서 자주 사용되는 단어들이다. 이는 바꿔 말하면 1등급 어휘 선정을 위한 말뭉치의 규모는 그리 클 필요가 없음을 보여 준다.<sup>20)</sup>

따라서 1등급 어휘 선정을 위한 말뭉치는 적은 규모여도 상관없으나 구어의 비율을 높이고 다양한 장르가 들어가 있는 균형 말뭉치의 형태여야 한다. 본 연구는 이러한 점을 고려하여 1등급 어휘 선정을 위한 말뭉치를 따로 구축하고자 한다.

20) 중간 보고 회의와 이후의 연구진 회의를 통해 기초 어휘 선정만을 위한 말뭉치의 규모는 클 필요가 없으나, 균형 말뭉치의 형태여야 함을 결정하였다.

## 2) 1등급 어휘 선정용 말뭉치 구축 후 절차

1등급 어휘 선정을 위한 말뭉치를 따로 구축한 후에는 다음과 같은 절차를 거쳐 1등급 어휘를 선정하고자 한다.



[그림 26] 기초 어휘 선정용 말뭉치 구축 후 절차

### (1) 양적 분석

먼저 1등급 어휘 선정용 말뭉치를 기반으로 양적인 분석을 통하여 단어를 추출한다. 이때에는 빈도, 범위, 산포도를 종합적으로 고려할 것이며 이들 세 기준의 적용 순서와 적용 비율은 실제 말뭉치 통계 자료에 대한 검토를 통해 추후 결정될 것이다.

다음, 선정된 단어들의 텍스트 포괄 범위를 대규모 말뭉치를 통해 양적으로 분석할 것이다. 1등급 어휘 선정용 말뭉치와 별도로 우리는 2등급 이상 어휘 선정용 말뭉치를 구축할 예정이며 후자의 경우 전자에 비해 대규모로 구축될 예정이다. 우선 선정된 단어들의 텍스트 포괄 범위를 조사할 것이다.

1등급 어휘는 대개 텍스트 포괄 범위가 매우 높다. Nation(2001)은 기초 어휘의 텍스트 포괄 범위가 80%를 넘는다고 보고한 바 있다. 한국어 1등급 어휘의 텍스트 포괄 범위에 대한 명확한 조사는 없지만 최소 60~70% 이상은 확보되어야 1등급 어휘라고 부를 수 있을 것이다.

따라서 이에 해당하는 어휘를 대상으로 단어들의 텍스트 포괄 범위를 조사하여 이것이 낮은 단어는 제외한다. 다시 빈도, 범위, 산포도를 조사하여 텍스트 포괄 범위가 높은 단어를 선정하는 작업을 반복하도록 한다.

## (2) 질적 분석

1단계 양적 분석을 통해 선정한 단어들을 대상으로 질적 분석을 수행한다. 이것은 크게 이론적 타당성 검토와 실제적 타당성 검토로 나누어 진행한다.

먼저 이론적 타당성 검토를 위해 시소러스 검토와 기존 기초 어휘와의 비교 작업을 수행하고자 한다. 이중 전자는 현실적으로 어려움이 많은 것이 사실이다. 현재 한국어 전체 어휘를 대상으로 하는 체계적인 시소러스 구축 작업은 이루어지지 못했기 때문이다. Basic English의 경우 사물어 600개, 성질어 150개, 작용어 100개로 구성된 것으로 알려져 있다. 국어 기초 어휘가 타당성을 확보하기 위해서는 기초 어휘 자체가 일정한 체계를 지닐 필요가 있을 것이며 이를 위해서는 한국어 시소러스에 대한 검토가 요구된다.

다음 한국어교육용 기초 어휘나 국어교육용 기초 어휘 형태로 제시된 기존 기초 어휘 목록과의 비교 검토 작업을 수행하고자 한다. 이러한 작업은 이 연구에서 선정한 1등급 어휘의 특징을 드러내고 그 한계와 장점을 부각시키기 위해서도 필요할 것으로 사려된다.

이와 같은 작업을 통해 최종 선정된 단어들은 3000개 이내가 될 것으로 예상된다. 우리는 이 단어들을 대상으로 전문가의 직관을 바탕으로 한 질적 검토 작업을 수행하고자 한다. 연구자의 직관뿐만 아니라 광범위한 전문가들과의 협의를 통해 어휘 목록의 직관적 적합성을 확보하고자 한다. 그리고 질적 검토 과정에서는 다음과 같은 요소들도 적극 논의될 것이다.

고유명사, 1,2,3,4와 같은 수, 요일명 등의 처리

- 대개 고유명사는 제외하는 것이 원칙이나 문제가 되는 어휘가 있는지 검토가 필요하다.
- 1,2,3,4..., 첫째, 둘째, 셋째... 등과 같은 수는 포함할 것인지를 여부를 결정해야 하며 만일 포함한다면 어디까지 포함할 것인지 검토해야 한다. 만일 빼다면 이들 어휘들을 어떻게 처리할 지도 결정해야 한다.
- 요일명 등 기타 문제가 되는 단어들의 처리 여부가 검토되어야 한다.

1등급 어휘가 갖는 의미적 특성 검토

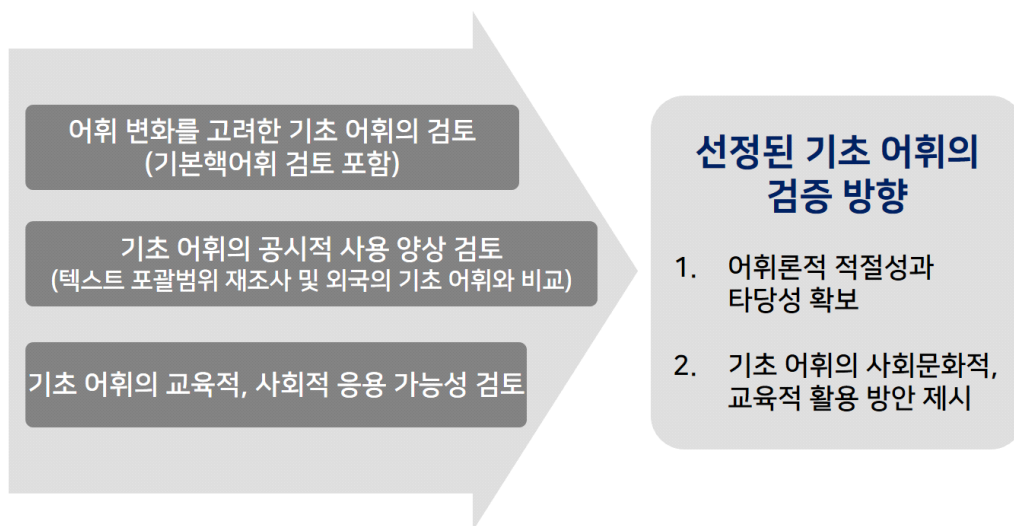
1등급 어휘는 대개 다의성을 지닌다. West의 Basic English의 경우 한 단어가 갖는 의미의 개수가 18.171개로 알려져 있다. 김광해(1993)에서 논의된 것처럼 1등급 어휘는 비전문어적 성격이 강해 해당 단어의 의미 영역이 넓다. 대개 한국어의 기초 어휘들은 고유어들일 가능성이 있다. 그런데 한국어 어휘는 고유어와 한자어가 일대다 대응 현상을 보이는 만큼 그 의미 영역이 넓고 비전문적이다. 따라서 1등급 어휘 목록 검토 시 1등급 어휘가 갖는 이러한 의미적 특성이 적극 고려되어야 할 것이다.



질적 분석 이후에는 다시 양적 분석으로 돌아가 선정된 단어들의 통계적 특성을 검토하는 작업이 반복될 것이다. 본 연구는 양적 분석과 질적 분석을 넘나들으로써 최대한 1등급 어휘 선정의 타당성을 높이고자 한다.

### 3) 선정된 기초 어휘의 검증 절차

위와 같은 과정을 통해 선정된 1등급 어휘들은 다시 그 타당성을 다음과 같은 방식으로 검증할 것이다.



[그림 27] 선정된 기초 어휘의 검증 방향

기본 핵어휘는 원래 언어 연대학에서 나온 개념이다. 언어 간 친족 관계를 확인하기 위해 기초적인 어휘 목록 2백여 개를 정하고 이 단어들의 소실과 보존 비율을 통계적으로 비교하기 위해 설정된 개념이다. 그런데 이때 사용되는 기본 핵어휘 목록은 한 언어의 근간을 이루는 단어들이므로 1등급 어휘이거나 1등급 어휘와 긴밀한 관련을 맺는다.

본 연구에서는 이 개념을 우리가 선정한 1등급 어휘 목록의 타당성 검토 시 이용할 수 있다. 말뭉치 확보의 어려움을 고려할 때 100여 년 전의 자료를 구하기는 쉽지 않을 것으로 보인다. 현실적인 대안으로 몇 십 년 전 자료 예컨대 20세기 중반 말뭉치 자료를 확보할 수 있다면 이들 자료에 우리가 선정한 1등급 어휘가 어느 정도로(빈도와 범위) 등장하는지 검토할 수 있다. 이는 선정된 1등급 어휘의 특성을 드러내기에도 유용하다.

말뭉치 확보가 가능하다면 우리가 선정한 1등급 어휘가 통시적으로 어느 정도의 텍스트 포괄 범위를 보여 주는지 어느 정도의 빈도로 사용되었는지 조사할 수 있

다. 가령 1960~80년대의 말뭉치와 1980~2000년대의 말뭉치에서 우리가 선정한 기초 어휘의 빈도 및 범위를 조사하여 그 추이를 정리함으로써 선정된 1등급 어휘의 특성을 드러낼 수 있다. 위와 같은 작업을 통해 필요하다면 어휘 목록의 수정도 이루어질 것이다.

1등급 어휘 선정의 타당성을 검토하기 위해 1등급 어휘가 광범위하게 사용될 것으로 예상되는 자료를 중심으로 기초 어휘의 텍스트 포괄 범위를 재조사할 필요가 있다. 그러한 자료들로는 한국어교육용 초급 교재, 유치원이나 초등학교 1,2학년 교재(교과서, 동화, 보조자료 등) 등이 대표적이다. 이러한 자료들에서 1등급 어휘가 어느 정도의 텍스트 포괄 범위를 갖는지 조사할 필요가 있다. 또 구어 말뭉치에서의 텍스트 포괄 범위와 문어 말뭉치에서의 텍스트 포괄 범위를 비교하여 구어 말뭉치에서의 텍스트 포괄 범위가 더 높은지 확인할 필요가 있다.

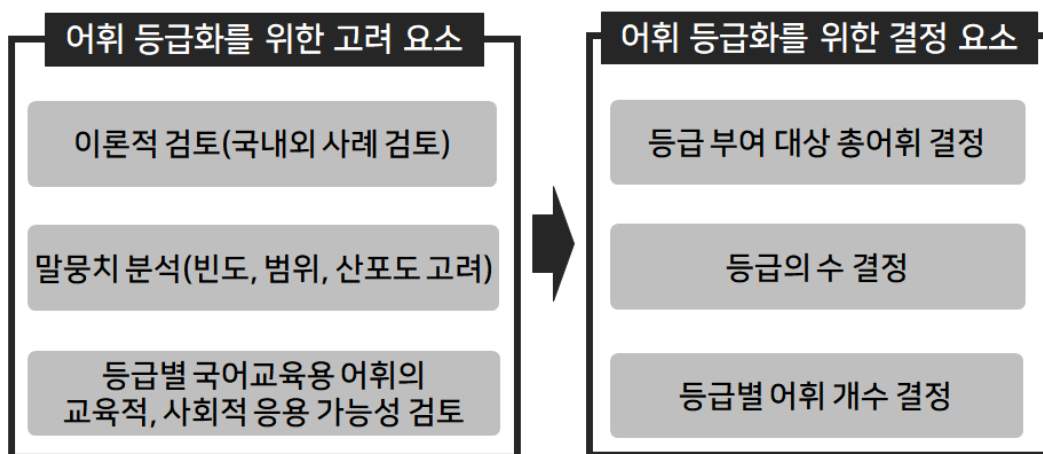
가능하다면 영어나 일본어의 1등급 어휘 목록과의 비교와 대조도 요구된다. 번역의 한계는 상존하지만, 기초 어휘로 제시된 단어들의 시소러스 체계와 제시된 단어들의 의미역 검토가 필요하다.

마지막으로 1등급 어휘가 갖는 교육적 사회적 중요성을 고려하여 이 목록의 응용 가능성을 다방면으로 조사할 필요가 있다. 이는 이 목록이 타당한지 여부를 증명하는 길이면서 추후 활용 가능성을 높이는 길도 된다. 한국어교육용 초급 교재나 유치원용 교재, 초등 1,2학년 교재에서의 텍스트 포괄 범위 조사는 이런 관점에서도 요구된다. 이 부분에 대해서는 관련 전문가들의 폭넓은 자문이 필요하다.

## 2.2. 2등급 이상 어휘 선정 절차

### 1) 어휘 등급화를 위해 결정해야 할 선행 요소

2등급 이상 어휘의 등급화를 위해서는 등급 부여 대상 총 단어 수와 등급의 개수, 등급 내 단어 수 등을 결정해야 한다. 우리는 국내외 등급화 사례를 검토하고 말뭉치 분석 결과를 참조하되, 어휘 등급화 결과물의 교육적 사회적 응용 가능성도 고려하고자 한다.



[그림 28] 어휘 등급화를 위한 결정 요소

올해 사업에서 정확한 수치를 제시하기는 어려우나 잠정적으로 등급 부여 대상 총 어휘 수는 5만 여개일 것으로 예상된다. 이는 일반 성인의 어휘량이 대략 5만 내외라는 논의(김광해, 1993: 309)를 참조한 것이다. 또한, 이와 관련하여 앞서 분석한 결과에서 텍스트 포괄 범위 95% 수준의 어휘수가 49,426개임을 확인한 바 있다. 국립국어원의 <표준국어대사전>에 수록되어 있는 어휘의 수는 대략 50여 만에 이르지만 이중 대다수는 그 사용 빈도가 매우 적다. 저빈도어들의 경우 등급 자체가 무의미하다. 등급 부여가 유의미하기 위해서는 해당 단어의 활발한 사용이 전제되어야 한다.

한편 이들 단어들을 몇 개의 등급으로 나누며 각 등급 내 몇 개 정도의 단어를 제시할 것인가도 어려운 문제이다. 김광해(2003)의 경우 총어휘 237,990를 교육적 중요도에 따라 총 7등급으로 나누었다. 이후 (주)낱말은 김광해(2003)의 7등급 어휘 체계를 9등급 체계로 보완하였는데 이는 교육적 활용도를 높이기 위함이었다.

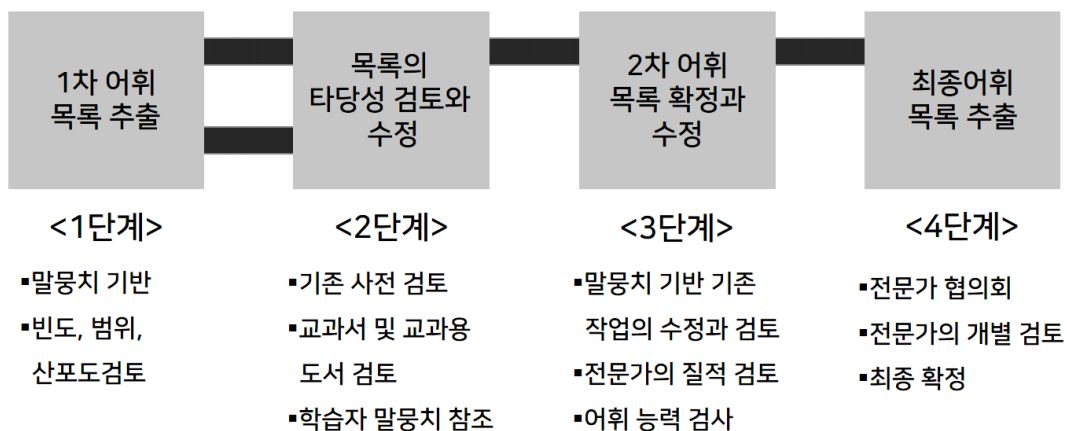
본 연구의 II장에서 논의한 것처럼 Nation(2001)은 영어 원어민 화자가 자연스러운 노출을 통해 1년에 약 1,000개의 어휘군을 습득한다고 보고 한 개의 등급 내

1,000개의 단어를 귀속시켰다. 그러나 이 작업은 BNC 14000 작업의 일환으로 이루어진 것으로서 총 단어 수가 14,000 어휘군에 불과하다는 점을 고려해야 한다. 이 작업의 결과를 우리 연구에 그대로 적용하기는 어렵다.

본 연구에서는 이와 같은 연구 결과들을 참조하되 말뭉치 분석을 통해 통계적으로 유의미한 지점을 포착하여 등급의 수와 등급 내 단어 수 결정의 1차적 기준으로 삼고자 한다. 그 다음 어휘 목록의 교육적, 사회적 활용도를 고려하여 이를 최종 결정하고자 한다. 소사전, 중사전 개발이나 교과용 교재 개발, 아동의 언어 발달 등을 적극 고려할 것이다.

## 2) 어휘 등급화의 절차와 검증

먼저 어휘 등급화의 절차와 검증 단계를 정리하여 제시하면 다음 그림과 같다.



[그림 29] 어휘 등급화 절차

본 보고서의 앞장에서 기술한 바 우리는 대규모 말뭉치를 구축하여 어휘 등급화 시 이용하고자 한다. 어휘 등급화를 위한 <1단계> 작업은 대규모 말뭉치에 대한 분석으로 시작할 것이다. 이 단계는 양적 분석 작업으로 1차적 검토 대상이 되는 5만여 개의 단어 추출에 집중될 것이다. 앞 장의 실험적 분석에서 드러난 것처럼 빈도, 범위, 산포도의 적용 순서와 그 비율은 1차적으로는 통계적 수치에 대한 검토를 통해 결정될 것이다. 말뭉치의 규모와 구성에 따라 선정 단어가 달라진다는 점에서 말뭉치 구축의 양적, 질적 정교화가 이루어질 것이다.

말뭉치에 기반하여 1차적으로 5만여 개의 단어를 추출한 뒤 이들 단어들의 타당성 검토 작업을 다방면으로 진행할 것이다(<2단계>). 우선 기존 소사전에 수록된 단어의 수가 대략 5만여 개인 점을 고려하여 소사전 어휘 목록과의 비교·검토를 수

행할 것이다. 그리고 이들 목록의 교육적 사회적 응용 가능성을 높이면서도 어휘 등급화의 학년별 수준별 타당성을 검토하기 위해 학년별, 수준별로 나누어져 있는 교과용 도서나 권장도서의 어휘 목록을 검토할 것이다. 교과용 도서의 경우 학년별 학습자의 수준을 고려하여 개발된다는 점에서 참고할 가치가 있다. 그러나 이들 교과용 도서 역시 개발자의 추정과 어휘적 직관에 의해 개발되었다는 점에서 100% 신뢰하기는 어렵다. 따라서 최종 판결은 전문가의 검토와 합의에 의해 결정되는 것이 타당할 것이다.

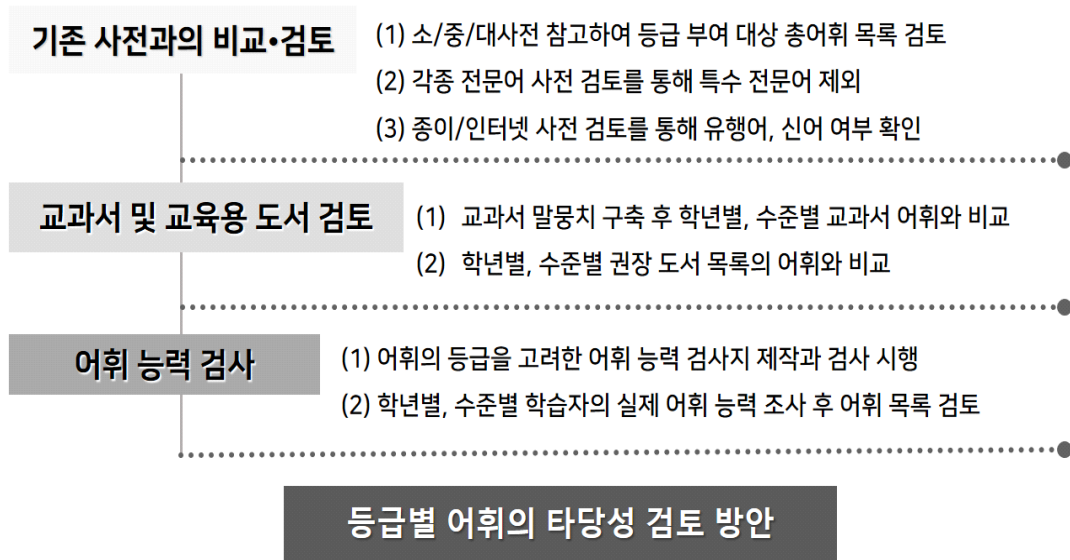
어휘 등급화 대상이 되는 단어에 특수 전문어나 유행어, 신어 등은 제외하는 것이 타당하다. 그런데 특수 전문어나 유행어, 신어 여부의 판단이 쉽지 않다. 전문가의 검토와 더불어 인터넷사전과 각종 특수 전문어 사전을 참고하여 이를 판단하고자 한다.

만일 학습자 쓰기 자료로 된 말뭉치가 있다면 등급별 어휘의 타당성 검토가 용이할 것이다. 그러나 현재 학습자 쓰기 자료는 학년별, 수준별, 지역별, 장르별 등을 고려하여 체계적으로 구축되어 있지 않다. 따라서 대규모 쓰기 말뭉치를 바탕으로 한 체계적인 검토는 불가능하다. 그러나 우리가 선정한 어휘의 타당성 검토를 위해 일정 부분 학습자 쓰기 자료를 참조하는 방안은 모색할 수 있을 것이다.

대규모 말뭉치를 기반으로 한 5만여 어휘의 선정과 그 타당성 검토를 위한 소사전과 교과용 도서 어휘와의 비교 작업은 회귀적으로 반복될 것이다. 따라서 <1단계> 작업과 <2단계> 작업은 순환적이다.

<1단계>, <2단계> 작업이 어느 정도 완료되면 2차 어휘 목록을 확정된 후 전문가 검토를 받을 예정이다. 그리고 어휘 능력 검사지를 개발하여 이 목록의 타당성을 실제적으로 검증하고자 한다. 학습자마다 개인차는 존재하겠지만 연령의 증가와 더불어 어휘량 역시 증가한다. 따라서 우리가 선정한 어휘 목록을 바탕으로 어휘 검사지를 만들어 유치원부터 초·중·고, 대학생을 대상으로 어휘 능력 검사를 시행하면 우리 목록의 타당성 검토가 일정 부분 가능하다.

지금까지의 논의를 통해 드러난바 우리는 어휘 등급화 결과의 타당성을 높이기 위해 다양한 검증 장치를 고려하고 있다. 이를 그림으로 정리하면 다음과 같다.



[그림 30] 등급별 어휘의 타당성 검토 방안

위 그림과 같이, 등급화된 어휘 목록은 전문가들의 개별적 질적 검토와 협의회를 통해 최종 결정하고자 한다. 앞의 3단계 작업 결과를 바탕으로 선정된 어휘 목록을 연구진 내에서 수정한 후 다시 전문가 자문을 통해 검토할 예정이다. 따라서 최종 어휘 결정은 개별 전문가들의 숙고와 전문가 협의회를 거쳐 이루어지게 된다. 이는 전문가의 어휘적 직관뿐만 아니라 어휘 등급화의 교육적, 사회적 응용 가능성도 충분히 검토하여 최종 확정하는 것이 연구 성과를 높이는 데 기여할 것이기 때문이다.

## V. 기초 어휘 사업의 중장기 계획 수립

### 1. 중장기 계획 수립의 경위

#### 1.1. 국민의 국어 능력과 어문생활에 기여하는 정책적 수단

기초 어휘 사업은 국어기본법에서 주창하고 있는 것과 같이, 변화하는 언어 환경에 능동적으로 대응하고 국민의 국어 능력을 실효성 있게 향상할 수 있는 방안으로 마련된 것이다.

기초 어휘 사업은 국어의 수십 만 어휘 중에서도 언어 사용의 근간을 이루는 부류로, 그것의 목록과 성격, 활용 방안 등을 개발하여 보급하면 국민 전체의 어문 생활을 진작시키는 데 직접 기여할 수 있을 뿐만 아니라, 국민의 국어 능력 발전을 위한 언어 자원으로써 다방면에 활용될 여지가 매우 크므로 시급한 과제이다. 그러나 기초 어휘 사업의 성립을 위해서는 문자 언어와 음성 언어를 모두 망라해야 하고, 남성과 여성, 성인과 노년층 등 국민 전체의 언어 사용 양상을 대표할 수 있는 대규모 말뭉치를 구축, 분석, 관리하여야 하므로 국가 수준의 예산 출연과 정책적 지원이 뒷받침되어야 한다.

2017년 현재 시점에 이르기까지 국립국어원에서 이루어져 온 어휘 관련 사업은 분야별 전문 용어 정비 등 각종 전문 분야 및 직능(職能) 등에 기반한 어휘 사업이 있었으나, 정작 국민의 국어 능력 발전과 관련된 어휘적 토대에 해당하는 기초 어휘 사업이 수행되지 못한 실정이다. 이에 따라, 기초 어휘의 개념 및 성격, 국내외 사례의 심층 조사 등과 같은 이론 연구에서부터 기초 어휘 목록 추출을 위한 말뭉치 구축 및 관리 방법론, 기초 어휘의 검증 및 타당화 도구 개발, 기초 어휘의 보급 및 활용 체계, 향후 말뭉치 및 기초 어휘 목록 보완 계획 등 장기간에 걸친 치밀하고 체계적인 계획이 마련될 필요가 있다.

#### 1.2. 국어 사용의 현황 파악 및 정비 사업 수행

기초 어휘 사업의 수행은 국민의 국어능력 평가, 교과서 및 교육자료 감수, 공무원 시험이나 대학수학능력시험 등의 국가 고사 출제에 기초 자료로 활용될 수 있는 목록 및 사용 체계 마련이 필요하다. 국가 공문서의 전달 어휘 선정, 국가고사의 타

당도 등 기초 어휘를 통해 공공 언어 과업의 효율성을 제고할 수 있으며, 어휘의 목록 및 사용 체계를 개발, 보급함으로써 국민 일반의 언어생활 향상에 크게 기여할 수 있다.

그런데 기초 어휘 사업은 그 자체의 성과물로서도 큰 가치를 지니지만, 국민의 국어 사용 양상 변화를 파악하고, 향후 어문 정책을 수립하는 근거로서 체계적으로 수집된 어휘 및 그 근거 자료로서의 말씀치가 마련된다는 점에서 국민 언어의 정비 사업과 긴밀히 관련되므로 체계화된 중장기 계획의 마련이 절실하다.

국어기본법에서도 밝힌바, ‘국가와 지방자치단체는 변화하는 언어 사용 환경에 능동적으로 대응하고, 국민의 국어 능력 향상과 국어의 발전을 위해 노력(국어기본법 제4조 참조)’해야 하는데, 이는 언어 현실의 변화를 민감하게 파악하고 국가가 이를 긍정적인 방향으로 추동해야 함을 의미한다. 어휘는 국어 능력의 기반이 되며, 언어 사용의 변화 양상을 매우 민감하게 반영하는 언어 단위이다. 또한 이 가운데서 기초 어휘는 국민의 국어 생활 전반을 기능적으로 부담하므로 국민의 언어 생활을 파악하고 향상시키는 데 매우 긴요하게 사용될 수 있다.

즉 기초 어휘의 조사 및 분석을 통하여 당대 국어를 사용하는 국민의 어휘 구성 및 사용 양상 전반을 파악하는 데도 활용할 수 있고, 어휘를 기반으로 한 국민의 국어 능력 수준 및 향후 발전 방향 등을 가늠해보는 도구로도 기능할 수 있으며, 기초 어휘 평정을 위해 수집된 말씀치나 언어 자료는 중장기적 차원에서 국민의 국어 능력 변화 추이를 파악하는 DB(데이터베이스)로 적재되는 등 다방면의 사용이 가능하다.

### 1.3. 언어 발전 계획의 세계적 수준 도달

세계 각국은 자국의 언어 문화를 발전시키고 그 보급을 위해 상당한 투자를 하고 있을 뿐만 아니라, 국민의 언어 사용 능력을 발전시키기 위한 정책적 노력을 다각도로 기울이고 있다.

미국의 경우, 일찍부터 기초 어휘 및 다양한 용도의 어휘 목록이 활발하게 개발되어 어문 생활에서 교육 및 각종 사업 분야에 이르기까지 이러한 성과물이 다용도로 활용되고 있으며, 영어의 영향력 및 사용 확대에 적지 않게 기여하고 있는 것이 사실이다. 각국의 언어 발전 성과는 곧 그 언어의 위상과 해당 국가의 경쟁력과도 밀접한 관련을 맺고 있기 때문이다.

기초 어휘 사업을 통해 어휘 및 기초 어휘와 관련된 선진 각국의 언어 발전 수준에 근접하게 되는 중요한 사명이 있고, 국가 고사, 출판, 교육 등 다방면에서 국민이 활용할 수 있는 언어 자원을 제공하여 국어의 세계화 및 한국어교육 등의 발전



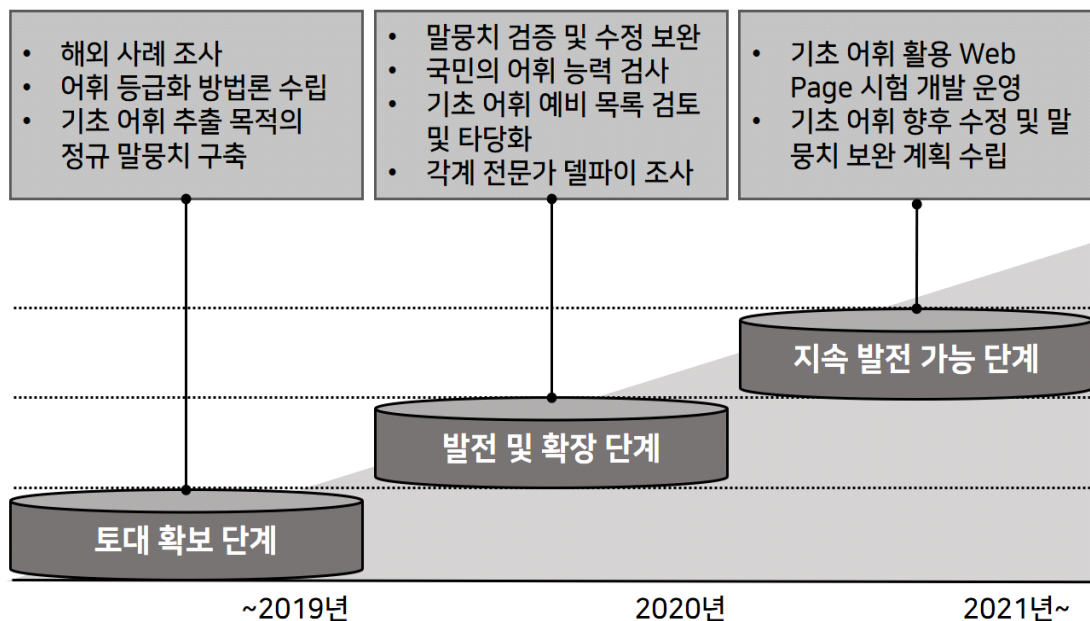
에도 기여할 여지가 매우 크므로 체계화된 계획 마련이 요구된다. 또한 기초 어휘 사업의 성과물이 국어 문화 공동체 전반에 보급되고 보다 다방면에 활용되도록 하기 위해서는 단발성 사업의 일환으로 추진되기보다는 장기적 관점에서 그 성과 및 향후 활용 계획, 전략 등이 포함된 중장기형 사업 추진 계획이 필요하다.

## 2. 중장기 계획 수립(향후 4개년)

본 연구는 기초 어휘 사업의 일환에서 수행된 연구로, 그 명칭은 ‘국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구’이다. 이 연구의 수행 결과를 바탕으로 향후 기초 어휘 사업 전반의 계획을 중장기 로드맵으로 제시해 보면 다음과 같다.

**2021년까지 ‘토대 확보 단계’→‘발전 및 확장 단계’→‘지속 발전 가능 단계’의 3단계에 걸쳐 발전 방향을 설정하여, 사업 목적 달성의 로드맵을 구성함.**

### ❖ 단계별 발전 방향



[그림 31] 기초 어휘 사업의 중장기 로드맵

앞서 본 장의 1.1에서도 논의한 바와 같이, 기초 어휘 사업은 그 사업의 성격상 중장기형으로 추진함으로써 국민의 국어 능력 발전과 언어 환경 개선에 지속 가능한 사업으로 자리매김 할 수 있다.

위 그림에서 나타나는 기초 어휘 사업 계획의 특징은 2+1+1의 형태로 사업 계획을 정교화 함으로써 토대 확보 단계(2년)에서부터 발전 및 확장 단계(1년), 지속 발전 가능 단계(1년)의 순서로 사업을 심화, 확장할 수 있도록 로드맵이 마련되어 있다는 점이다. 다음은 각 단계에 해당하는 세부 계획을 단계별로 살펴보도록 한다.

## 2.1. 토대 확보 단계

### [사업 개요]

- 전 국민 대상 기초 어휘 목록의 마련 및 활용 체계 마련을 위한 토대 수립 목적의 연구
- 기초 어휘 사업의 핵심 요소인 등급화 방법론 개발 및 말뭉치 자료 구축, 기초 어휘 예비 목록의 검증 수단 개발 연구

### [사업 목적]

- 기초 어휘 사업의 첫 단계로서 정량적 방법의 예비 어휘 목록을 타당화할 수 있는 등급화 기준 등을 설정
- 해외의 관련 어휘 연구 및 기초 어휘에 대한 폭넓은 검토를 통해 활용 체계 및 어휘 등급화에 대한 정보 수집
- 기초 어휘 예비 목록의 검증 기준으로 적용 가능한 각종 교과서 등 학술 텍스트를 통한 어휘 목록 개발
- 국어 기초 어휘의 목록 추출에 초점화 된 말뭉치를 구축하고 말뭉치의 수집, 전사와 관리 방법론 수립함으로써 향후 사업이 수행될 수 있는 토대 마련

### [세부 사업 내용]

- 【1단계】 토대 확보 단계 (2018년)
  - 기초 어휘 해외 사례 조사: (기초)어휘가 활용되는 해외 각 분야의 사례를 조사하고 국어 기본 어휘 개발에 적용할 수 있는 시사점 마련
  - 기초 어휘 목록의 등급화 방법론 수립: 정량적, 통계적 방법론으로 추출(당해 연도 연구 결과)되는 기초 어휘 예비 목록을 다양한 기준으로 타당화할 수 있는 구체적 방법론 수립

- 교과서 어휘 목록 연구: 교과서에서 추출된 어휘 목록을 활용하여 말뭉치 기반의 기초 어휘 예비 목록을 검증, 보완할 수 있는 어휘 목록 추출. 기존에 구축된 교과서 어휘 목록 등을 적극 활용
- 기초 어휘 추출 목적의 말뭉치 정련화 (1): 국어 기본 어휘의 목록 추출에 초점화 된 말뭉치를 구축하고 말뭉치의 수집, 전사와 관리 방법론 수립함으로써 향후 사업이 수행될 수 있는 토대 마련

○ 【2단계】 토대 확보 단계 (2019년)

- 기초 어휘 추출 목적의 말뭉치 정련화 (2): 국어 기본 어휘의 목록 추출에 초점화 된 말뭉치를 구축하고 타당성을 검증함으로써, 사업 전체의 타당성과 어휘 정비 사업으로서의 자료적 가치 제고

## 2.2. 발전 및 확장 단계

### [사업 개요]

- 전 국민 대상 기초 어휘 목록에 대한 타당화 목적의 연구
- 기초 어휘 사업의 성과물인 기초 어휘 예비 목록과 말뭉치를 검증, 수정, 보완하기 위한 연구

### [사업 목적]

- 토대 확보 단계에서 마련된 대규모 말뭉치를 활용하여 기초 어휘 예비 목록 및 등급화 도출
- 기초 어휘 예비 목록에 대한 각종 검증 및 타당화 필요
- 기본 어휘 예비 목록의 검증 기준으로 적용할 수 있도록 타당화가 가능한 규모의 표본을 통해 국민의 어휘 능력 평가

- 검증 결과에 따른 연구 성과물의 수정·보완을 통해 사업을 통해 산출된 자료의 가치와 활용 가능성 제고
- 기초 어휘 사업의 성과 확산을 위한 기초 어휘 목록을 통한 활용 체계의 청사진 및 개발안 마련

[세부 사업 내용]

- 발전 및 확장 단계 계획(2020)
  - 기초 어휘 예비 목록 및 등급화: 말뭉치를 통한 정량적, 통계적 방법을 통해 등급화 된 기초 어휘 예비 목록을 도출
  - 기초 어휘 예비 목록의 검증 및 타당화 작업: 기초 어휘 예비 목록에 대해 성인의 국어 능력 평가, 교과서 어휘 조사 결과, 각계 전문가 델파이 조사 등을 통해 검증하고 필요 시 수정·보완함.
  - 기초 어휘를 통해 제공 가능한 활용 체계 구안: 기초 어휘를 통해 제공 가능한 활용 체계에 대한 개발안 마련
  - 성인의 어휘 능력 조사: 말뭉치의 정량적 분석을 통해 수립된 기초 어휘 예비 목록을 검증하는 기준으로 활용될 수 있는 어휘 능력 조사 결과 도출. 성인의 어휘 능력을 평가할 수 있는 도구 개발 및 다양한 직종 및 분야의 국민을 적정 규모로 표집, 조사

## 2.3. 지속 발전 가능 단계

[사업 개요]

- 전 국민을 대상의 기초 어휘 기반 활용 체계(공적 언어 서비스) 개발 연구
- 기초 어휘 사업의 성과물을 확장시키고 지속 가능한 사업으로 정착시키기 위한 목적의 연구

[사업 목적]

- 기초 어휘 사업의 성과물 홍보 및 보급
- 기초 어휘 응용 사업의 개발 계획 수립(국어 능력 평가 도구, 텍스트 난도 평정, 각종 국가 고사의 적용 가능 도구 개발 등)
- 기초 어휘 사업이 지속 가능 하도록 하는 말뭉치 및 어휘 목록 보완 사업 시행 계획안 마련

[세부 사업 내용]

- 발전 및 확장 단계 계획(2021)
- 기초 어휘 활용 체계 개발: 웹페이지 또는 모바일 형태의 기초 어휘 활용 체계 개발
- 기초 어휘 활용 체계 시험 운용 및 보완: 일정 기간의 기초 어휘 활용 체계에 대한 시범 운영을 통해 운용 노하우를 확보하고, 국민의 국어 능력 발전에 기여할 수 있도록 활용 체계를 수정·보완
- 기초 어휘 응용 사업의 개발 계획 수립: 기초 어휘 사업의 성과를 통해 각 분야에 적용 가능한 응용 사업의 계획을 수립함.<sup>21)</sup>
- 기초 어휘 사업의 지속 가능화: 기초 어휘를 통해 제공 가능한 활용 체계에 대한 개발안 마련

21) 연구의 최종 결과물인 어휘 목록의 활용 방향은 크게 두 가지로 나누어볼 수 있다. 첫째, 이 연구는 국민의 국어능력 발전을 위한 어휘 정비 사업의 일환으로써, 어휘 목록과 함께 제시될 어휘 등급 정보, 분야 정보, 대체 어휘 정보 등을 바탕으로 다양한 분야에 활용이 가능하다. 둘째, 이 연구의 성과물인 어휘 목록은 향후 어휘 관련 사업의 기반으로 다양한 어휘 연구의 기초 자료로서의 중요한 역할을 하게 될 것이다. 구체적인 활용 분야로는 크게 언어 환경 개선, 교육 분야, 출판 분야, 임상 분야의 네 영역으로 나누어 볼 수 있으며, 각 분야별 활용 방향은 다음과 같은 것들이 가능하다고 본다.

- 언어 환경 개선: 국가 주도의 표준화 검사 개발, 문식성 지수 개발, 공문서 어휘의 적합성 평가
- 교육 분야: 기초 학력 교육 및 소외계층 언어 능력 지원, 해외 동포·새터민·다문화 배경 국민의 교육 기초 자료 개발, 기초 직무 교육 개발의 기본 자료
- 출판 분야: 수준별, 목적별 도서 선정 및 도서 편찬의 기초 자료, 다용도 사전 편찬의 기초 자료
- 임상 분야: 언어 장애의 진단과 치료 도구 개발, 임상 검사의 어휘 적절성 검토의 기초 자료

## VI. 종합 및 제언

### 1. 종합

본 연구에서는 기초 어휘와 등급화에 대한 제반 이론을 수립하고, 기 구축 말뭉치 및 어휘 목록 사례를 검토하여 이론적 기반을 마련하며 말뭉치에 기반한 기초 어휘 선정 및 등급화를 위한 샘플 말뭉치를 구축하였다. 이를 통해 기초 어휘 선정 작업의 적실성과 실제적인 방안을 마련하는 데 그 목적이 있었다. 연구의 주요 내용을 정리하면 다음과 같다.

#### 1) 기초 어휘 선정 및 등급화를 위한 기초 연구

본 연구는 기초 어휘 선정 및 등급화를 위한 기초연구로서 기초 어휘, 어휘 평정 및 등급화 관련 선행 연구를 검토하고, 기구축된 국내외 말뭉치와 어휘 목록을 수집, 분석하여 기초 어휘 목록 개발 방향을 수립하였다. 먼저 어휘 평정 및 등급화 관련 선행연구의 검토를 통해 기초 어휘의 개념을 ‘일상생활에서 사용 빈도가 높고 파생이나 합성들의 조어(造語)에 고빈도로 참여하여 다른 단어로 대체하기 어려운 특성을 지닌 단어’로 정의하였는데, 이는 기존의 기초 어휘 개념을 좀더 적극적으로 해석한 것이다. 또한 국내외 말뭉치 및 어휘 목록을 정리하여 특징을 분석하였다. 특히 국내 어휘 목록은 한국어 교육 영역에서 교육용으로 선정한 것이 대표적이었으며 정책적 목적으로 국민 전체를 대상으로 한 어휘 목록의 구축이나 등급화는 실제 이루어진 사례가 없었다.

#### 2) 샘플 말뭉치 구축

기초 어휘 및 등급화를 위한 말뭉치는 실제 언어생활의 다양한 측면을 포착해야 하므로 규모 및 장르에 대한 고려와 시간/연대에 대한 고려가 함께 이루어져야 한다. 본 연구에서는 아래와 같이 말뭉치 자료를 수집하고 정리하여 총 20억 어절 규모의 말뭉치를 구축하였다.

- 이미 구축되어 공개된 대규모 말뭉치를 최대한 이용하였다. 21세기 세종계획에서 구축된 현대 문어 원시 말뭉치가 이에 해당한다. 현대 구어 말뭉치, 형

태 분석 말뭉치 등은 규모가 작아서 일단 제외하였으나, 1차년도에 결과를 검토한 뒤 차년도의 작업에서 포함 가능하다.

- 인터넷에서 웹 크롤링 기법을 이용하여 자동 수집할 수 있는 자료를 폭넓게 수집하였다. 드라마·시나리오, 신문, 잡지, 블로그 등이 이에 해당한다.
- 출판사에서 출판 과정에서 만든 편집용 파일, PC 통신 등에서 연재된 소설 등의 각종 도서 자료를 폭넓게 수집하였다.

향후 연구에서는 말뭉치의 양적 규모를 확장하기보다는 기존 말뭉치와 보완 말뭉치로부터 얻은 어휘 목록의 비교에 초점을 맞추어야 할 것이다. 기존 말뭉치와 보완 말뭉치를 합친 총 29억 어절 규모의 말뭉치를 바탕으로 어휘 등급화를 시행하면, 그 결과는 문어와 구어의 균형을 추구할 수 있을 것이다.

### 3) 기초 어휘 선정 및 등급화

말뭉치를 기반으로 하여 기초 어휘 추출을 위한 어휘 목록 선정 및 등급화를 위해 본 연구에서는 크게 두 부분으로 구분하여 연구를 진행하였다. 하나는 샘플 말뭉치를 토대로 어휘 목록을 추출하는 과정을 전반적으로 수행해 보는 것과 다른 하나는 기초 어휘 선정과 등급화의 절차를 수립한 것이다.

샘플 말뭉치를 토대로 어휘 목록을 추출하는 과정은 사전등재형을 어휘 추출 단위로 정하고 일차로 형태소 분석 작업을 거친 후, 장르별 빈도, 범위, 산포도를 추출하고 변수들 사이의 관계에 대한 통계적 고찰을 실시하였다. 그 다음 변수들에 가중치를 부여하여 단어 순위를 결정하고 그 결과를 다른 어휘 목록과 비교하고 텍스트 포괄 범위를 조사하여 검증하였다. 조사 결과 상위 49,426개 단어가 말뭉치 전체 token의 95%를 포괄하였다. 대략 텍스트 token의 95% 정도를 차지하는 약 5만 개의 상위 빈도 단어가 어휘 등급화의 주된 대상일 될 것으로 예상되었다. 이러한 작업들은 샘플 말뭉치를 활용하여 구체적인 절차를 실행해 본 것으로, 각각의 절차를 실제로 수행해 봄으로써 실제 기초 어휘 선정 및 어휘 등급화 사업을 진행할 때의 유의점이나 보안점 등을 도출할 수 있었다.

- 가중치 부여 실험의 기반이 되는 단어 목록 보완
- 점수 산출 방식의 다양한 실험
- 점수 최상위 단어들에 대한 정성적 검토
- 기존 기초 어휘 목록에 대한 정성적 검토

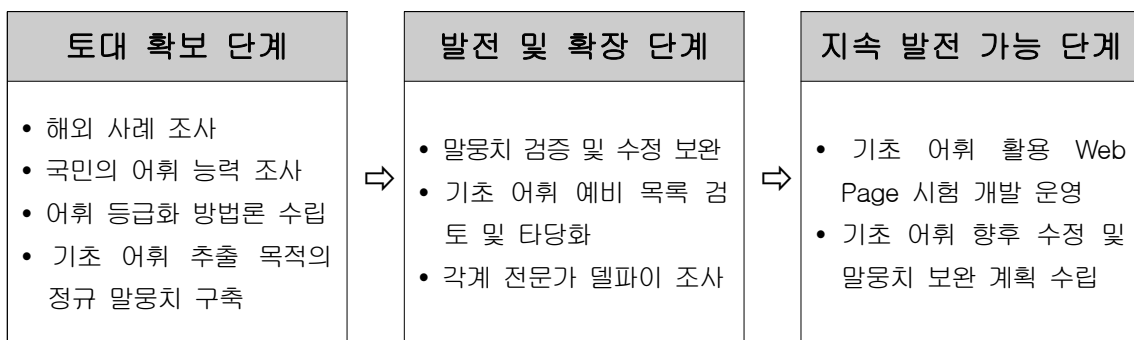


기초 어휘 선정과 등급화의 절차는 1등급 기초 어휘와 2등급 이상 기초 어휘로 구분하여 논의를 진행하였다. 1등급 기초 어휘는 대개 사용 빈도가 높고 말뭉치 규모와 상관없이 등장하므로 그 기반 말뭉치가 적은 규모여도 상관 없으나, 균형적인 장르가 반영되어야 한다. 현재 구어적 성격이 높은 언어 자료를 대거 보강하였으나, 추후 실제 구어 자료의 보강이 필요하다. 1등급 기초 어휘 선정용 말뭉치 구축 후 빈도, 범위, 산포도를 종합적으로 고려하여 단어를 선정하고, 텍스트 포괄 범위를 분석하여 최소 60~70%를 확보하는지 확인하는 양적 분석과 기존 기초 어휘와의 비교, 전문가 자문 및 협의 등 질적 분석을 통해, 양적 분석만으로 선정된 단어를 수정하는 절차를 거친다.

2등급 이상 기초 어휘는 현재 말뭉치 자료의 텍스트 포괄 범위 95%를 차지하는 어휘 수가 49,426개이므로 잠정적으로 약 5만 여개 단어로 예상한다. 말뭉치로부터 추출한 어휘 목록을 통해 통계적으로 유의미한 지점을 포착하여 등급의 수와 등급 내 단어의 수를 1차로 결정하고, 그 다음 어휘 목록의 교육적·사회적 활용도를 고려하여 최종 결정할 것이다.

#### 4) 기초 어휘 사업의 중장기 계획 수립

기초 어휘 사업은 그 사업의 성격상 중장기형으로 추진함으로써 국민의 국어 능력 발전과 언어 환경 개선에 지속 가능한 사업으로 자리매김할 수 있다. 본 연구의 수행 결과에 따라 향후 기초 어휘 사업 전반의 계획은 2+1+1의 형태로 사업 계획을 정교화 함으로써 토대 확보 단계(2년)에서부터 발전 및 확장 단계(1년), 지속 발전 가능 단계(1년)의 순서로 사업을 심화, 확장할 수 있도록 로드맵을 마련하였다.



[그림 32] 단계별 발전 방향

## 2. 정책 제언

본 연구는 국어 기초 어휘의 선정 및 등급화를 위한 기초 연구를 수행한 성과로서 궁극적으로 국민의 국어 능력을 발전시키고자 하는 사업 목적에서 수행된 것이다.

이러한 일환에서 국민의 국어 능력과 국어 어휘 사업의 확장 및 발전 가능성에 주목하여 향후 사업에 대한 정책 방향을 제언하도록 한다.

### ○ 사업 규모의 적정화 및 사업의 지속 가능성 확보

본 연구에서 수행된 사례 연구, 전문가 자문 등에서도 여러 번 지적된 것과 같이 국어 어휘 사업은 단기, 일회성의 사업으로 시행되어서는 그 성과를 확산하기 어려우며 국민의 국어 능력 발전을 위한 토대를 쌓는다는 점에서 접근될 필요가 있다. 이를 위하여 국어 기본 어휘에 영향을 미치는 사회적, 언어적 변인을 다양하게 고려하고 다방면의 교차 검증과 실험 등을 시행할 수 있도록 사업 규모를 적정화하는 것이 필수적이다.

또한 어휘는 본질적으로 시대상의 변화에 민감하게 변화하는 언어 단위이므로, 이러한 특성을 반영해 이 사업을 지속 가능한 형태로 어휘 말뭉치와 목록을 보완해 가는 방식(중장기 계획 참조)으로 추진할 필요가 있다. 또한 장기적으로는 사업의 범위를 보다 확장하여 기초 어휘에 포함되는 다수의 전문어 등 타 분야의 어휘 조사 및 정비 계획도 수립하여 전체 사업의 범위를 확장시킬 필요가 있다.

### ○ 어휘 기반 국어 능력 발전 사업의 기획 및 추진

당해 연도 수행된 사업의 성과는 국어 기초 어휘의 목록을 추출하고 활용 체계를 마련하기 위한 기반을 닦는 것으로 요약된다. 이러한 관점에서 활용 체계를 국어 능력 발전 사업이라는 관점에서 보다 적실한 형태로 제공하기 위해서는 그 자체로 독자성을 지니면서도 어휘 사업의 틀에서 성과를 공유할 수 있는 세부 사업을 정교화 하는 것이 필요하다.

예컨대 직능에 따른 어휘의 조사나 국어적 특성에 맞는 어휘 조사 도구 개발과 검증 등 어휘 기반 국어 능력 발전에 선행적으로 수행되어야 하는 사업이 다수인바, 이러한 사업 수요를 조사, 예측하여 사업을 추진할 필요가 있다. 이뿐만 아니라 어휘 사업의 성과물은 다른 국어 관련 사업에 비해 언어 교육 등 타 분야로의 전파 및 활용 가능성이 매우 크므로 사업적 관심을 적극 확대할 필요가 있다.

### ○ 수행 기관 및 각급 기관과의 공조 체제 강화

국어 기초 어휘 목록의 선정 및 등급화는 전 국민을 대상으로 자료를 수집하는 등의 방식은 지양하지만 그 목록 등을 정교화 하는 데 있어 다수의 실험이나 검증 절차를 포함할 필요성이 점차 증가할 것이다. 이러한 점에서 국립국어원이 보유한 성과물을 활용할 수 있도록 지원하고, 각급 공공기관 및 학교급과의 공조 체계를 마련하여 연구의 성과 및 검증에 조력할 필요가 있다.

## 참고 문헌

- 강범모(2011), 『개정판 언어, 컴퓨터, 코퍼스언어학』, 고려대학교출판부.
- 강병규·손민정(2016), 「2015 개정 중국어교육과정의 기본 어휘 선정 및 활용」, 『중국어언어연구』 63, 한국중국어언어학회.
- 강충열(1999), 「초등학교 1-6학년 아동의 기초 어휘 이해 발달 상황」, 『한국심리학회지 발달』 12(2), 한국심리학회.
- 강현화(2001), 「국어교육용 기초 한자어에 대한 기초 연구: 한국어 교재에 나타난 어휘를 바탕으로」, 『한국어교육』 12(2), 국제한국어교육학회.
- 강현화(2010), 「한국어 어휘학습 교재 개발을 위한 기초 연구: 학습자 요구분석을 중심으로」, 『언어와 문화』 6(2), 한국언어문화교육학회.
- 강현화(2014), 「한국어교육용 중급 어휘 선정에 대한 연구」, 『외국어로서의 한국어교육』 40, 연세대학교 언어연구교육원 한국어학당.
- 강현화(2014), 「한국어교육용 초급 어휘 선정 연구」, 『문법 교육』 21, 한국문법교육학회.
- 강현화(2015), 「한국어교육용 고급 어휘 선정에 대한 연구」, 『외국어로서의 한국어교육』 42, 연세대학교 언어연구교육원 한국어학당.
- 강현화(2015), 「한국어능력시험 어휘 목록 개발 연구」, 『한국언어문화교육학회 학술대회』 2015(4), 한국언어문화교육학회.
- 강현화·원미진(2012), 「한국어학습자를 위한 『한국어기초사전』 구축 방안 연구」, 『한국사전학』 20, 한국사전학회.
- 강현화·원미진(2015), 「언어학습자를 위한 『한국어기초사전』 편찬의 원리와 실제」, 『민족문화연구』 67, 고려대학교 민족문화연구원.
- 고석주·남윤진·서상규(1998), 「한국어교육을 위한 기초 어휘 의미 빈도 사전의 개발」, 『언어정보개발연구』 1, 연세대학교 언어정보연구원.
- 고영근·구본관(2007), 『우리말문법론』, 집문당.
- 곽재용(2004), 「초등학교 국어교과서의 어휘 분석」, 『우리말글』 32, 우리말글학회.
- 곽재용(2010), 「초등학교 저학년 국어 교과서에 나타난 어휘 분석」, 『한글』 290, 한글학회.
- 곽재용(2012), 「“한국어교육과정” 부록에 제시된 어휘 목록 분석」, 『배달말』 51, 배달말학회.
- 구본관(2011), 「어휘 교육의 목표와 의의」, 『국어교육학연구』 40, 국어교육학회.
- 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(2005), 『현대 국어 사용 빈도 조사2』, 국립국어연구원.

- 국립국어원(2009), 『교육용 기본 어휘 선정을 위한 기초 연구』, 국립국어원.
- 국립국어원(2010), 『초등학생 글쓰기 어휘 조사 연구』, 국립국어원.
- 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(1988), 『중학교 교과서 어휘』, 국어연구소.
- 권재일·고동호(2004), 『국어학 고유 용어 분류 체계에 관한 연구』, 국립국어원.
- 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.
- 김동일·안예지·황지영·박소영·김봉년(2016), 「교육과정중심 어휘검사 개발을 위한 기초연구: 초등학교 4-6학년 교육용 어휘 선정을 중심으로」, 『학습장애연구』, 13(3), 한국학습장애학회.
- 김석영(2014), 「중국어 어휘론에서 기본 어휘의 문제-중국식 기본 어휘 개념의 어휘론적 유용성에 대한 비판적 검토」, 『중국어언어연구』 52, 한국중국어언어학회.
- 김언자(2006), 「고등학교 교육과정에서의 프랑스어 기본 어휘 선정에 대하여」, 『프랑스어문교육』 23, 한국프랑스어문교육학회.
- 김유미·강현화(2008), 「학문목적 학습자를 위한 학술 전문어휘 선정 연구 -한국어, 문학, 경영학, 컴퓨터공학 전공을 대상으로-」, 『한국어교육』 19(3), 국제한국어교육학회.
- 김윤주(2016), 「『한국어교육과정』 어휘 목록 분석 -『국제 통용 한국어교육 표준 모형』과의 비교를 중심으로-」, 『우리어문연구』 54, 우리어문학회.
- 김정렬·이선아(2004), 「통합영어교육을 위한 타교과 기초 어휘 조사 연구」, 『영어교과교육』 3(1), 한국영어교과교육학회.
- 김정우(2011), 「문학교육과 어휘교육」, 『국어교육학연구』 40, 국어교육학회.
- 김중학(2001), 『한국어의 기초 어휘 연구』, 박이정.
- 김중신(2011), 「어휘를 통한 정의적 텍스트 생산 전략」, 『국어교육학연구』 40, 국어교육학회.
- 김지영(2004), 「한국어 어휘 교육 항목 선정을 위한 기초 연구」, 『한국어교육』 15(2), 국제한국어교육학회.
- 김창원(2012), 「고등학교 어휘 교육의 위상과 어휘교육론의 과제」, 『국어교육학연구』 44, 국어교육학회.

- 김한샘(2000), 「한국어 명사의 어휘망 구축에 대한 기초 연구: 『연세한국어사전』의 분석을 중심으로」, 『사전편찬학 연구』 10(1), 연세대학교 언어정보개발원.
- 김한샘(2003a), 『한국 현대 소설의 어휘 조사 연구』, 국립국어연구원.
- 김한샘(2003b), 『현대 국어 사용 빈도 조사』, 국립국어원.
- 김한샘(2004), 「국어 어휘 분석 말뭉치의 구축과 활용」, 『한말연구』 14, 한말연구학회.
- 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, 우리말 글학회.
- 김한샘·국립국어원(2009), 『초등학교 교과서 어휘 조사 연구』, 국립국어원.
- 김한샘·서상규(1998), 「말뭉치의 구축과 활용 -연세 말뭉치 1의 구상과 실제-」, 『언어정보 개발연구』 1, 연세대학교 언어정보연구원.
- 김현철·조은경(2010), 「중국어학계의 중국어 기본 어휘 선정 현황과 활용방안 연구」, 『중국어 문학논집』 60, 중국어문학연구회.
- 김화수·이숙·서지희·정다운·천정민·최경윤(2015), 「초등학교 1-3학년 국어 교과서 어휘 분석」, 『언어치료연구』 24(4), 한국언어치료학회.
- 김홍규·강범모(1995), 「고려대학교 한국어 말모듬 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.
- 남길임(2005), 「말뭉치 기반 사전 편찬의 동향과 지향 방향: 최근 30년간의 사전 편찬 방법론을 중심으로」, 『한말연구』 16, 한말연구학회.
- 목정수(2014), 「사전과 코퍼스 두껍게 읽기 -서상규, 『한국어 기본 어휘 의미빈도 사전』(한국문화사, 2014)」, 『어문논총』 60, 한국문학언어학회.
- 문교부(1956), 『우리말 말수 사용의 찾기 조사』, 문교부.

- 문금현(2011), 「어휘장을 활용한 한국어 어휘 교육」, 『우리말교육현장연구』 5(2), 우리말교육현장학회.
- 민경모(2011), 「해외 청소년 대상 교육용 어휘 선정을 위한 기초 연구: 해외 청소년용 교재에 나타난 어휘의 계량적 분석을 중심으로」, 『언어와 문화』 7(2), 한국언어문화교육학회.
- 박미숙(2011), 「중학생들의 어휘능력 실태 연구」, 『국어교과교육연구』 19, 국어교과교육학회.
- 배재석·임승규(2005), 「고교 중국어 기본 어휘 만족도 조사 연구」, 『중국어문학논집』 32, 중국어문학연구회.
- 배주채(2010), 『한국어 기초 어휘집』, 한국문화사.
- 변은주(2005), 「초등학교 국어 교과서의 어휘 분석」, 진주교대 교육대학원 석사학위논문.
- 서덕현(1990), 「기본 어휘의 개념과 기초 어휘의 위상: 교육용 어휘를 중심으로」, 『국어교육』 71, 한국국어교육연구회.
- 서상규(2012), 『한국어 말뭉치의 구축과 과제: 한국어와 정보화』, 태학사.
- 서상규(2013), 『한국어 기본 어휘 연구』, 한국문화사.
- 서상규(2014), 「한국어 기본 어휘 검증에 관한 일고찰 -연세대 한국어학당 <1급단어목록>과의 대조 분석-」, 『외국어로서의 한국어교육』 40, 연세대학교 언어연구교육원 한국어학당.
- 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·한영균(1999), 『국어정보학 입문』, 태학사.
- 서정국(1968), 「국어 기본 어휘의 연구」, 고려대 교육대학원 석사학위논문.
- 서정미(2008), 「말뭉치를 활용한 고등학교 국어사전의 편찬을 위한 기초 연구」, 경기대 박사학위논문.
- 서종학·김주필(1999), 『교과서의 어휘 분석 연구: 초등학교 국어 교과서를 중심으로』, 국립국어연구원.
- 서지영(2007), 「말뭉치를 활용한 중학교 국어과 학습사전 편찬을 위한 기초 연구」, 경기대 교육대학원 석사학위논문.
- 설성수(2013), 「개방형 한국어 지식대사전 분야 분류」, 『한국사전학』 22, 한국사전학회.
- 성광수(1999), 「어휘부의 구조와 기초 어휘의 활용」, 『선청어문』 27(1), 서울대학교 국어교육과.
- 송영빈(2011), 「일본에서의 한자 교육 -초등학교를 중심으로-」, 『국어교육학연구』 40, 국어교육학회.
- 송철의 외(2008), 『한국 근대 초기의 어휘』, 서울대학교출판부.
- 신동광(2011), 「기본 어휘의 선정 기준: 영어 어휘를 중심으로」, 『국어교육학연구』 40, 국

- 어교육학회.
- 신동광(2014). 「우리는 어떤 영어 어휘를 학습하고 있는가?: 국내외 대표 영어 어휘목록 비교·분석」, 『외국학연구』 30, 외국학연구소.
- 신동광·전유아·이신웅·박명수(2017), 「영어 연어 능숙도 검사지 개발」, 『멀티미디어언어교육』 20(2), 한국멀티미디어언어교육학회.
- 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.
- 신민철(2011), 「어휘이론과 교육기본 어휘」, 『일본어교육연구』 21, 일본어교육학회.
- 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.
- 심혜령(2007), 「학습사전에서의 접사 처리의 문제」, 『겨레어문학』 38, 겨레어문학회.
- 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」, 『중국어언어연구』 20, 한국중국어언어학회.
- 연세대학교 언어정보개발연구원(2007), 『연세한국어사전』, 두산동아.
- 유해준(2010), 「한국어교육용 어휘 문법 항목의 위계화 방안 -언어적 구성 항목을 중심으로-」, 『국제한국어교육학회 학술대회 논문집 2010』, 국제한국어교육학회.
- 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.
- 윤혜경(2016), 「한국어교육용 구어 어휘 선정 연구」, 『인문과학연구』 50, 강원대학교 인문과학연구소.
- 이경수(2011), 「초등학교 국어 어휘교육에 대한 소고 -우리과 프랑스의 학업성취도 평가 유형을 중심으로-」, 『국어교육학연구』 40, 국어교육학회.
- 이관규(2011), 「문법 교육과 어휘 교육」, 『국어교육학연구』 40, 국어교육학회.
- 이도영(2011), 「어휘 교육 평가의 이론적 고찰 -목표와 내용 타당도를 중심으로-」, 『국어교육학연구』 40, 국어교육학회.
- 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·김시정·김정선(2017), 「국어 기본 어휘 선정을 위한 기초 연구 -현황과 과제를 중심으로-」, 『국어교육』 156, 한국어교육학회.
- 이숙의(2014), 「의미부류를 활용한 한국어교육용 온톨로지 구축: 고급 단계 [인지적 행위] 동사를 중심으로」, 『한국어 의미학』 43, 한국어의미학회.
- 이옥형(1995), 「성인 지능발달에 관한 일고찰」, 『교육연구』 29, 성신여자대학교 교육문제연구소.
- 이유경(2005), 「외국인의 대학 수학을 위한 어휘 목록 선정의 필요성 연구 - 한국어교육 전공 논문의 명사 어휘 중심으로」, 『이중언어학』 29, 이중언어학회.
- 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.
- 이중은(2005), 「한국어교육을 위한 의존용언 표현의 어휘항목 선정」, 『이중언어학』 28, 이중언어학회.
- 이중철(2011), 「작문 교육과 어휘 교육」, 『국어교육학연구』 40, 국어교육학회.
- 이준호(2008), 「한국어 어휘 교육 연구사: 학위 논문을 중심으로」, 『문법 교육』 9, 한국문법교육학회.
- 이지옥(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.
- 이창수(2011), 「어휘 교육 연구 어디까지 왔나」, 『우리말교육현장연구』 5(2), 우리말교육현장학회.
- 이충우(1991), 「초등학교 1, 2학년 국어과 교과서 어휘 조사 연구」, 『관동어문학』 7, 관동어문학회.
- 이충우(1992), 『國語 教育用 語彙 研究 : 國民學校·中學校 國語科 教育用 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·최영룡(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」, 『한국어 의미학』 35, 한국어의미학회.
- 이현진·김주필(2004), 「유아용 동화책의 어휘 분석 연구」, 『Communication Sciences and Disorders』 9(1), 한국언어청각임상학회.
- 이희자(2003), 「국어의 기초 어휘 및 기본 어휘 연구사」, 『새국어생활』 13(3), 국립국어원.
- 임지룡(1991), 「국어의 기초 어휘에 대한 연구」, 『국어교육연구』 23, 국어교육학회.
- 임지룡(1998), 「어휘력 평가의 기본 개념」, 『국어교육연구』 30(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 한국어교육용 어휘 선정 연구」, 『국어교육학연구』 14, 국어교육학회.
- 임칠성(2003), 「기본 어휘 선정 방법론」, 『새국어생활』 13(3), 국립국어원.
- 임홍빈·한재영(1993), 『국어 어휘의 분류 목록에 대한 연구』, 국립국어연구원.
- 장경숙·정규태·이병천(2011), 「2009 교육과정 개정에 따른 영어과 기본 어휘 목록 및 지침 개발을 위한 기초 연구」, 『현대영어교육』 12(2), 현대영어교육학회.
- 장경희·조성문·김명희·김순자·김정선·이필영·임유중·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.
- 장유진·홍희정(2005), 「국어사전의 전문용어에 관한 연구」, 『한글』 270, 한글학회.
- 장현진·전희숙·신명선·김효정(2013), 「영·유아의 기초 어휘 선정 연구」, 『언어치료연구』 22(3), 한국언어치료학회.
- 장현진·전희숙·신명선·김효정(2014), 「초등학생 교육용 기초 어휘 선정 연구: 저학년 중심으로」, 『언어치료연구』 23(1), 한국언어치료학회.
- 전미순·이병운(2009), 「초급 단계 문화 어휘 선정과 문화 교육 방안」, 『한국언어문화학』 6(1), 국제한국언어문화학회.

- 전영주·김은성(2015), 「교과 어휘의 기능 분류에 따른 교과 학습 어휘의 개념 설정 -국어, 사회과 교과서 어휘 구성 비교 분석을 중심으로-」, 『새국어교육』 104, 한국국어교육학회.
- 전은주(2012), 「중학교 어휘 교육의 위상과 개선 방안」, 『새국어교육』 93, 한국국어교육학회.
- 정찬섭·이상섭·남기심·한중철·최영주(1990), 「우리말 낱말 빈도 조사 표본의 선정 기준」, 『사전편찬학연구』 3(1), 연세대학교 언어정보개발원.
- 정한진·이옥분·서경희(2007), 「성인용 동사 이름대기 평가 어휘 목록 -말뭉치를 기반으로 한 기초 어휘 연구」, 『언어치료연구』 16(2), 한국언어치료학회.
- 조남호(2002), 「국어 어휘의 분야별 분포 양상」, 『冠嶽語文研究』 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.
- 조철원(1997), 「고등학교 국어 교과서 어휘 연구」, 경남대 교육대학원 석사학위논문.
- 조현용(1999), 「한국어교육용 기본 어휘 선정에 관한 연구」, 『고향논집』 25, 경희대학교 대학원.
- 조현용(2000), 「어휘 중심 한국어교육 방법 연구」, 경희대 박사학위논문.
- 조형일(2013), 「교육용 외래어·외국어 표현 선정과 표기 방안 연구」, 『한국언어문화학』 10(1), 국제한국언어문화학회.
- 주세형(2005), 「국어과 어휘 교육의 발전 방향」, 『독서연구』 14, 한국독서학회.
- 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
- 최길시(1998), 『외국인을 위한 한국어교육의 실제』, 태학사.
- 최상재(1993), 「고등학교 국어 교과서 어휘 연구」, 경남대 교육대학원 석사학위논문.
- 최성규(1999), 「장애아동의 어휘지도를 위한 일반아동의 기초 어휘 난이도 분석」, 『특수교육연구』 6, 국립특수교육원.
- 최진영(2001), 「중학교 국어 교과서의 어휘 연구」, 계명대 교육대학원 석사학위논문.
- 최충일(1999), 「6차 교육과정 국어(하)교과서 어휘 연구」, 경남대 교육대학원 석사학위논문.

- 최형용(2013), 『한국어 형태론의 유형론』, 박이정.
- 최형용(2016), 『한국어 형태론』, 역락.
- 한송화(2015), 『한국어 학습용 어휘 목록』, 국립국어원.
- 한영균(2006), 「한국어 어휘 교육·학습 자료 개발을 위한 계량적 분석의 한 방향: 어휘 빈도 조사 방법의 개선을 위하여」, 『어문학』 94, 한국어문학회.
- 한영균(2009), 「코퍼스에 기반한 한·일 기본 어휘의 연어 구성 대조 분석 연구」, 『국어학』 55, 국어학회.
- 한유석(2010), 「일한 분류어휘비교표의 구성과 전망」, 『한국일본어문학회 학술발표대회논문집』 2010(4), 한국일본어문학회.
- 한정환(2012), 「한국어교육에서의 어휘와 문법-조사, 어미의 기본 어휘 선정 과정을 중심으로-」, 『한국어학』 57, 한국어학회.
- 허재영(2012), 「국어 어휘 분류 체계의 역사적 흐름」, 『겨레어문학』 48, 겨레어문학회.
- 홍윤표(2002), 『한국어와 정보화』, 태학사.
- 황용주(2016), 「21세기 세종 말뭉치 제대로 살펴보기」, 『새국어생활』 26(2), 국립국어원.
- 황유모·김정훈(2015), 「개방형 한국어 지식 대사건 전문용어 신분류 체계 설정 및 재분류」, 『전기학회논문지』 64(2), 대한전기학회.

- Bauer & Nation, I. S. P.(1993). “Families”. *International Journal of Lexicography*, 6.
- Brezina V., & Gablasova, D.(2013). “Is there a core general vocabulary?: Introducing the New General Service List”. *Applied Linguistics*.
- Browne, C.(2013). “The New General Service List: Celebrating 60 years of vocabulary learning”. *The Language Teacher*, 7(34).
- Carroll, J. B., Davies, P., & Richman, B. (eds).(1971). *The American heritage word frequency book*. Boston, MA: Houghton Mifflin.
- Coxhead, A.J.(1998). The development and evaluation of an academic word list. Unpublished M.A. Thesis. Wellington: Victoria University of Wellington.
- Coxhead, A.J.(2000). “A new academic word list”. *TESOL Quarterly*, 34(2).
- Davies, M., & Gardner, D.(2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. London: Routledge.
- Dickins, J. (n.d.). *Extended version of a General Service List of English words*. Retrieved from <http://www.leeds.ac.uk/arts/downloads/file/2205/gslforweb27513xlsx>.
- Extensive Reading Foundation. (n.d.). *The extensive Reading Foundation grading scale*. Retrieved from [http://erfoundation.org/wordpress/wp-content/uploads/2012/07/ERF\\_Scale\\_2013.png](http://erfoundation.org/wordpress/wp-content/uploads/2012/07/ERF_Scale_2013.png)

- Francis, W. N., & Kucera, H.(1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin Company.
- Francis, W. N., & Kucera, H.(1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin Company.
- Fries, C. C., & Traver, A. A.(1960). *English word lists*. Ann Arbor: George Wahr.
- Gnanadesikan, A. E. (2009). *The writing revolution: Cuneiform to the internet*. Chichester, UK: Wiley-Blackwell.
- Gyllstad, H., Vilkaitė, L., Schmitt, N.(2015). “Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates”. *ITL-International Journal of Applied Linguistics*, 166(2).
- Hlaváčová, J.(2006). *New approach to frequency dictionaries: Czech example*. Retrieved from [http://www.lrec-conf.org/proceedings/lrec2006/pdf/11\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/11_pdf.pdf).
- Hopkins, C.(1979). “The spontaneous oral vocabulary of children in grade 1”. *The Elementary School Journal*, 79(4).
- Howatt, A. P. R., & Widdowson, H. G.(2004). *A history of English language teaching*. Oxford : Oxford University Press.
- Hu, M. H-C., & Nation, I. S. P.(2000). “Unknown vocabulary density and reading comprehension”. *Reading in a Foreign Language*, 13(1).
- Johns, T.(1986). “Micro-concord: A language learner's research tool”. *System*, 14(2).
- Johns, T.(1988). “Whence and whither classroom concordancing?” In T. Bongaerts, P. de Haan, S. Lobbe & H. Wekker (Eds.), *Computer applications in language learning*. London: Foris.
- Johns, T.(1991). “Should you be persuaded - two examples of data-driven learning”. In T. Johns & P. King (Eds.), *Classroom concordancing: English Language Research Journal*, 4. University of Birmingham: Centre for English Language Studies.
- Laufer, B.(1992). “How much lexis is necessary for reading comprehension?” In Arnaud, P. J. L. and Bejlint, H. (Eds.), *Vocabulary and applied linguistics*. London: Macmillan.
- McArthur, T.(1981), *Longman lexicon of contemporary English*. Harlow: Longman.
- McArthur, T.(1998). *Living words: Language, lexicography and the knowledge revolution*. Exeter: University of Exeter Press.
- Moe, A. J., Hopkins, C. J., & Rush, R. T. (1982). *Vocabulary of first grade children*. Springfield, IL: Charles C. Thomas.
- Morrison, C. M., Chappell, T. D., & Ellis, A. W.(1997). “Age of acquisition norms

- for a large set of object names and their relation to adult estimates and other variables”. *Quarterly Journal of Experimental Psychology*, 50(A).
- Murphy, H.(1957). “The spontaneous speaking vocabulary of children in primary grades”. *The Journal of Educational Research*, 140.
- 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.
- Ogden, C. K.(1930). *Basic English: A general introduction with rules and grammar*. London: Kegan Paul.
- Palmer, H. E.(1936). “The history and present state of the movement towards vocabulary control”. In R. C. Smith (Ed.), *Teaching English as a foreign language, 1912-1936: Pioneers of ELT Vol. V*. Routledge.
- Read, J.(2004). “Research in teaching vocabulary”. *Annual Review of Applied Linguistics*, 24.
- Richards, J. C.(1974). “Word lists: Problems and prospects”. *RELC Journal*, 5(2).
- Sandhofer, C. M., Smith, L. B., & Luo, J.(2000). “Counting nouns and verbs in the input: Differential frequencies, different kinds of learning?” *Journal of Child Language*, 27.
- Schmitt, N.(2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Schonell, F. J., Meddleton, K. G., & Shaw, B. A.(1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.
- Stemach, G., & Williams, W. B.(1988). *WordExpress: The first 2,500 words of spoken English*. Novato, CA: Academic Therapy Publications.

- Thorndike, E. L., & Lorge, I.(1944). *The teacher's word book of 30,000 words*. New York: Teachers' College Columbia University.
- Wepman, J. M., & Hass, W.(1969). *A spoken word count (children-ages 5, 6 and 7)*. Chicago: Language Research Associates.
- West, M.(1953). *A general service list of English words*. London: Longman, Green & Co.
- West, M., Swenson, E., Fawkes, K., Russell, F., & de Magellanes Wilf, J.(1934). *A Critical Examination of Basic English*. Toronto, The University of Toronto Press.
- Zimmerman, C. B.(1997). "Historical trends in second language vocabulary instruction". In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition*. New York: Cambridge University Press.
- 萩原 廣(2014), 「日本人の語彙量(理解語彙、使用語彙)調査を行うにあたっての基礎的研究」, 『京都語文』 21, 佛教大学.
- 萩原 廣(2016), 「大学4年生の日本語の使用語彙は平均約3万語、理解語彙は平均約4万5千語」, 『京都語文』 23, 佛教大学国語国文学会.

## <부록 1> 세종 태그셋

체언	명사	NN	일반명사	NNG
			고유명사	NNP
			의존명사	NNB
	대명사	NP	대명사	NP
용언	수사	NR	수사	NR
	동사	VV	동사	VV
	형용사	VA	형용사	VA
	보조용언	VX	보조용언	VX
	지정사	VC	긍정지정사	VCP
			부정지정사	VCN
	관형사	MM	관형사	MM
수식언	부사	MA	일반부사	MAG
			접속부사	MAJ
독립언	감탄사	IC	감탄사	IC
관계언	격조사	JK	주격조사	JKS
			보격조사	JKC
			관형격조사	JKG
			목적격조사	JKO
			부사격조사	JKB
			호격조사	JKV
			인용격조사	JKQ
	보조사	JX	보조사	JX
의존형태	접속조사	JC	접속조사	JC
			선어말어미	EP
			종결어미	EF
			연결어미	EC
			명사형전성어미	ETN
			관형형전성어미	ETM
	접두사	XP	체언접두사	XPN
	접미사	XS	명사파생접미사	XSN
			동사파생접미사	XSV
			형용사파생접미사	XSA
	어근	XR	어근	XR
기호	마침표, 물음표, 느낌표			SF
	쉼표, 가운뎃점, 콤마, 빗금			SP
	따옴표, 괄호표, 줄표			SS
	줄임표			SE
	불임표(물결, 숨김, 빠짐)			SO
	외국어			SL
	한자			SH
	기타기호(논리 수학기호, 화폐기호 등)			SW
	명사추정범주			NF
	용언추정범주			NV
	숫자			XN
	분석불능범주			NA



## <부록 2> UTagger 오류 패턴 검토

### 1) 단어의 오류 유형

#### ① 동음이의어 분석 오류

1	오분석	사대_10/NNG+교수_06/NNG+협_88/NNG
	정분석	사대_04/NNG+교수_06/NNG+협_88/NNG
	설명	→ ‘사대’는 ‘사범대학’의 의미이므로 ‘사대_04’로 수정되어야 함.
2	오분석	오늘/NNG 150+개_10/NNB+대_04/NNG
	정분석	오늘/NNG 150+개_10/NNB+대_06/NNG
	설명	→ ‘대’는 ‘대학’의 의미이므로 ‘대_06’으로 수정되어야 함.
3	오분석	평가_03/NNG+단_20/XSN 결성/NNG
	정분석	평가_03/NNG+단_22/XSN 결성/NNG
	설명	→ ‘단’은 ‘단체’의 의미이므로 ‘단_22’로 수정되어야 함.
4	오분석	총장_01/NNG+평가_03/NNG+제_19/XSN
	정분석	총장_01/NNG+평가_03/NNG+제_20/XSN
	설명	→ ‘제’는 ‘제도’의 의미이므로 ‘제_20’으로 수정되어야 함.
5	오분석	교수_06/NNG+대_20/XPN+의회_02/NNG
	정분석	교수_06/NNG+대_04/NNG+의회_02/NNG
	설명	→ ‘대’는 ‘큰’의 의미이므로 ‘대_04/NNG’로 수정되어야 함.
6	오분석	광명_03/NNP+천안_01/NNP 연결_01/NNG 민자/NNG+고속도_02/NNG 검토/NNG
	정분석	광명_03/NNP+천안_01/NNP 연결_01/NNG 민자/NNG+고속도/NNG 검토/NNG
	설명	→ ‘고속도’는 ‘고속도로’의 의미이므로 ‘속도가 높음’의 의미를 가지는 ‘고속도_02’와 구별되어야 함.
7	오분석	서울_01/NNP 서부_01/NNG+광명_02/NNG+에서/JKB
	정분석	서울_01/NNP 서부_01/NNG+광명_02/NNP+에서/JKB
	설명	→ ‘광명’은 고유 명사에 해당함.
8	오분석	복원_02/NNG+지_28/XSN
	정분석	복원_02/NNG+지_25/XSN
	설명	→ ‘복원지’의 ‘지’는 장소의 의미이므로 ‘잡지’의 의미를 가지는 ‘지_28/XSN’는 ‘지_25/XSN’로 수정되어야 함.
9	오분석	이_05/NP+는/JX 인공_01/NNG+복원_02/NNG+시_06/NNG
	정분석	이_05/NP+는/JX 인공_01/NNG+복원_02/NNG+시_10/NNB
	설명	→ ‘시_06/NNG’는 ‘행정 구역’을 의미하므로 ‘시기’를 의미하는 ‘시_10/NNB’으로 수정되어야 함.
10	오분석	수피_03/NNG
	정분석	수피_02/NNG
	설명	→ ‘나무의 껍질’을 뜻하는 명사는 ‘수피_02’에 해당함. ‘수피_03’은 ‘동물의 가죽’의 의미임.
11	오분석	재_17/XPN+사정_07/NNG
	정분석	재_17/XPN+사정_11/NNG
	설명	→ ‘심사하여 결정한다’는 뜻은 ‘사정_11’이고 ‘사정_07’은 ‘일의 형편이나 까닭’의 의미임.
12	오분석	반올림하/VV+ㄴ/ETM 정수_11/NNG
	정분석	반올림하/VV+ㄴ/ETM 정수_21/NNG
	설명	→ 수학 용어인 ‘정수_21’을 ‘물을 맑게 함’을 뜻하는 ‘정수_11’로 잘못 분석함.

13	오분석	2+차_03/NNB 전형_07/NNG
	정분석	2+차_04/NNB 전형_07/NNG
	설명	→ ‘번, 차례’의 의미를 가진 것은 ‘차_04’에 해당함.
14	오분석	무효/NNG+소_03/NNG
	정분석	무효/NNG+소_15/NNG
	설명	→ ‘소송’을 뜻하는 ‘소_15’를 동물 ‘소_03’로 잘못 분석함.
15	오분석	눈_01/NNG+이/JKS 이어지/VV+ㄹ/ETM 것_01/NNB+이라고/JKQ
	정분석	눈_04/NNG+이/JKS 이어지/VV+ㄹ/ETM 것_01/NNB+이라고/JKQ
	설명	→ 눈_01’은 ‘신체 기관’의 의미이고 ‘내리는 눈’의 의미는 ‘눈_04’에 해당함.
16	오분석	월동/NNG 장구_01/NNG+를/JKO 갖추/VV+ㄴ/ETM
	정분석	월동/NNG 장구_14/NNG+를/JKO 갖추/VV+ㄴ/ETM
	설명	→ ‘장구_01’은 ‘악기’에 해당하고 여기서의 ‘장구’는 ‘도구’의 의미를 가지는 ‘장구_14’로 분석해야 함.
17	오분석	위_05/NNB+뱀/SH+를/JKO 대폭_01/MAG
	정분석	위_06/NNG+뱀/SH+를/JKO 대폭_01/MAG
	설명	→ ‘장기’의 하나인 ‘위_06/NNG’를 단위성 의존 명사 ‘위_05/NNB’로 잘못 분석함.
18	오분석	GM/SL+의/JKG 대주주/NNG+이/VCP+ㄴ/ETM 모_11/NNB 장학_01/NNG+회_14/XSN
	정분석	GM/SL+의/JKG 대주주/NNG+이/VCP+ㄴ/ETM 모/MM 장학_01/NNG+회_14/XSN
	설명	→ 관형사 ‘모’가 의존 명사로 잘못 분석됨.
19	오분석	경찰_04/NNG 수사권/NNG 독립/NNG+안_01/NNG+을/JKO
	정분석	경찰_04/NNG 수사권/NNG 독립/NNG+안_04/NNG+을/JKO
	설명	→ ‘안건’을 뜻하는 ‘안_04/NNG’을 ‘안쪽’을 뜻하는 ‘안_01/NNG’으로 잘못 분석함.
20	오분석	흐르_01/VV+어/EC+나가/VV+면서/EC 경찰_04/NNG+대_04/NNG 출신/NNG 간부_05/NNG+들_09/XSN+은/JX
	정분석	흐르_01/VV+어/EC+나가/VV+면서/EC 경찰_04/NNG+대_06/NNG 출신/NNG 간부_05/NNG+들_09/XSN+은/JX
	설명	→ ‘경찰대/NNP’ 혹은 ‘경찰_04/NNG+대_06’으로 분석되어야 하나, ‘크다’의 의미를 가진 ‘대’로 분석함.
21	오분석	최근/NNG 검사_03/NNG+실_05/NNG+별_04/XSN+로/JKB
	정분석	최근/NNG 검사_03/NNG+실_12/XSN+별_04/XSN+로/JKB
	설명	→ 접미사 ‘실_12/XSN’을 방을 뜻하는 명사 ‘실_05/NNG’로 잘못 분석함.
22	오분석	참여/NNG+복지_10/NNG
	정분석	참여/NNG+복지_09/NNG
	설명	→ ‘행복한 삶’을 뜻하는 ‘복지_09’를 ‘복덕과 지혜’를 뜻하는 ‘복지_10’으로 잘못 분석함.
23	오분석	과장_07/NNG+사무관/NNG+급_04/NNG 직원_03/NNG
	정분석	과장_07/NNG+사무관/NNG+급_06/XSN 직원_03/NNG
	설명	→ 접미사 ‘급_06/XSN’을 일반 명사 ‘급_04/NNG’으로 잘못 분석함.
24	오분석	국_01/NNG+局/SH+별_04/XSN+로/JKB
	정분석	국_02/NNG+局/SH+별_04/XSN+로/JKB
	설명	→ ‘부서’를 의미하는 ‘국_02’가 마시는 ‘국_01’으로 잘못 분석됨.
25	오분석	금주_01/NNG 중_05/NNP 과장_01/NNG+급_04/NNG+들_09/XSN+만/JX+을/JKO
	정분석	금주_01/NNG 중_04/NNB 과장_01/NNG+급_04/NNG+들_09/XSN+만/JX+을/JKO
	설명	→ 의존 명사 ‘중_04’를 ‘중국’을 뜻하는 고유 명사 ‘중_05’로 잘못 분석함.

26	오분석	7+차례_01/NNG
	정분석	7+차례_02/NNG
	설명	→ ‘순서’를 의미하는 ‘차례_02’가 제사를 의미하는 ‘차례_01’로 잘못 분석됨.
27	오분석	제_21/XPN+1+출국/NNG+장_45/XSN+에서/JKB
	정분석	제_22/XPN+1+출국/NNG+장_45/XSN+에서/JKB
	설명	→ ‘차례’의 뜻을 지닌 접두사는 ‘제_22’이고 ‘제_21’은 접미사에 해당함.
28	오분석	대한항공/NNP KE/SL+653+편_09/NNB
	정분석	대한항공/NNP KE/SL+653+편_05/NNB
	설명	→ ‘이동 수단’을 뜻하는 의존 명사 ‘편_05’ 대신 ‘책을 세는 단위’인 의존 명사 ‘편_09’로 잘못 분석함.
29	오분석	청풍/NNP+에/JKB 민물_01/NNG+비빔/NNG+회_14/XSN
	정분석	청풍/NNP+에/JKB 민물_01/NNG+비빔/NNG+회/NNG
	설명	→ ‘음식’을 뜻하는 일반 명사 ‘회’로 분석되어야 하나 모임을 뜻하는 접미사 ‘회_14/XSN’로 잘못 분석함.
30	오분석	간_10/NNB 유전자/NNG+를/JKO
	정분석	간_08/NNG 유전자/NNG+를/JKO
	설명	→ 일반 명사 ‘간_08/NNG’을 의존 명사로 잘못 분석함.
31	오분석	재빨리/MAG 배_02/NNG 위_01/NNG+에/JKB 올라타/VV+았/EP+다/EF
	정분석	재빨리/MAG 배_01/NNG 위_01/NNG+에/JKB 올라타/VV+았/EP+다/EF
	설명	→ ‘운송 수단’인 ‘배_01’을 ‘신체 기관’으로 잘못 분석함.
32	오분석	일자_05/NNG 원피스/NNG+를/JKO 입_01/VV+고/EC
	정분석	일자_02/NNG 원피스/NNG+를/JKO 입_01/VV+고/EC
	설명	→ ‘일자 모양’을 뜻하는 ‘일자_02’를 ‘날짜’의 의미로 분석함.
33	오분석	명함/NNG+을/JKO 한_01/MM+장_22/NNB
	정분석	명함/NNG+을/JKO 한_01/MM+장_21/NNB
	설명	→ ‘장_22’는 ‘성씨’의 의미이므로 종이를 세는 단위인 ‘장_21’로 수정되어야 함.
34	오분석	할인_01/NNG+가_18/XSN+로/JKB 사/VV+아/EC+오_01/VV+ㄴ/ETM
	정분석	할인_01/NNG+가_19/XSN+로/JKB 사/VV+아/EC+오_01/VV+ㄴ/ETM
	설명	→ ‘가_18’은 ‘노래’를 의미하므로 ‘값’의 의미를 가지는 ‘가_19/XSN’로 수정되어야 함.
35	오분석	어학_02/NNG 연수_08/NNG 간_10/NNB 8+개월/NNB
	정분석	어학_02/NNG 연수_09/NNG 간_10/NNB 8+개월/NNB
	설명	→ ‘출장’의 의미는 ‘연수_09’에 해당함. ‘연수_08’은 ‘비용’의 의미.
36	오분석	위안_03/NNG+의/JKG 대상_11/NNG+이/JKS 없_01/VA+었/EP+다/EF
	정분석	위안_02/NNG+의/JKG 대상_11/NNG+이/JKS 없_01/VA+었/EP+다/EF
	설명	→ ‘위안_03’은 ‘중국 화폐의 단위’이므로 ‘위로하고 마음을 편하게 함’의 의미인 ‘위안_02/NNG’로 수정되어야 함.

## ② 품사 분석 오류

### ○ 일반 명사와 고유 명사

37	오분석	주택/NNG+공사_07/NNG+가/JKS 동백_01/NNG 택지_01/NNG+개발/NNG+지구_03/NNG 내_09/NNB
	정분석	주택/NNG+공사_07/NNG+가/JKS 동백/NNP 택지_01/NNG+개발/NNG+지구_03/NNG 내_09/NNB
	설명	→ ‘동백’은 고유 명사이므로 ‘동백/NNP’로 분석해야 함.
38	오분석	동해안/NNG 지역_03/NNG 2+만_06/NR+4000+여_27/XSN+ha/SL+를/JKO
	정분석	동해안/NNP 지역_03/NNG 2+만_06/NR+4000+여_27/XSN+ha/SL+를/JKO
	설명	→ ‘동해안’은 고유 명사이므로 ‘동해안/NNP’로 분석해야 함.

39	오분석	구성재/NNG
	정분석	구성재/NNP
	설명	→ 인명 ‘구성재’가 일반 명사로 분석됨.
40	오분석	강원/NNG 산간/NNG+과/JC
	정분석	강원/NNP 산간/NNG+과/JC
	설명	→ ‘강원’은 고유 명사인데 일반 명사로 분석됨.
41	오분석	초당_02/NNG+할머니/NNG+순두부/NNG+033+652+2058+가/JKS
	정분석	초당/NNP+할머니/NNG+순두부/NNG+033+652+2058+가/JKS
	설명	→ ‘초당’은 고유 명사인데 일반 명사로 분석됨.
42	오분석	삼교_02/NNG 원조_06/NNG 동치미/NNG 막국수/NNG+033+661+5396+는/JX
	정분석	삼교/NNP 원조_06/NNG 동치미/NNG 막국수/NNG+033+661+5396+는/JX
	설명	→ ‘삼교’는 고유 명사인데 일반 명사로 분석됨.
43	오분석	황수정/NNP+이/JKS 박히/VV+ㄴ/ETM 백금_01/NNG+반지_02/NNG
	정분석	황수정/NNG+이/JKS 박히/VV+ㄴ/ETM 백금_01/NNG+반지_02/NNG
	설명	→ ‘황수정’은 일반 명사인데 고유 명사로 분석됨.
44	오분석	그_01/MM 일_08/NNP+신동엽/NNP 대마초/NNG 구속_02/NNG 사건_01/NNG+이/JKS 있_01/VA+고/EC 나_01/VX+아서/EC
	정분석	그_01/MM 일/NNB+신동엽/NNP 대마초/NNG 구속_02/NNG 사건_01/NNG+이/JKS 있_01/VA+고/EC 나_01/VX+아서/EC
	설명	→ 일반 명사 ‘일’을 고유 명사로 분석함.

○ 부사격 조사와 접속 조사

45	오분석	법제화/NNG 추진_02/NNG 움직임/NNG+과/JC 맞물리/VV+면서/EC
	정분석	법제화/NNG 추진_02/NNG 움직임/NNG+과/JKB 맞물리/VV+면서/EC
	설명	→ ‘과’는 부사격 조사이므로 ‘과/JKB’로 분석되어야 함.
46	오분석	훌륭하/VA+ㄴ/ETM 총장_01/NNG+이사장/NNG+과/JKB +비리_08/NNG+비_32/XPN+민주_02/NNG 총장_01/NNG+이사장/NNG+을/JKO
	정분석	훌륭하/VA+ㄴ/ETM 총장_01/NNG+이사장/NNG+과/JC +비리_08/NNG+비_32/XPN+민주_02/NNG 총장_01/NNG+이사장/NNG+을/JKO
	설명	→ ‘과’는 접속 조사이므로 ‘과/JC’로 분석해야 함.
47	오분석	자료_03/NNG 분석_02/NNG+과_04/NNG+설문_01/NNG+조사_30/NNG+를/JKO
	정분석	자료_03/NNG 분석_02/NNG+과/JC+설문_01/NNG+조사_30/NNG+를/JKO
	설명	→ ‘과’는 접속 조사이므로 ‘과/JC’로 분석해야 함.
48	오분석	2+월_02/NNB+까지/JX 신문_10/NNG+협회/NNG+와/JC 체결하/VV+겠/EP+다/EF+고/JKQ
	정분석	2+월_02/NNB+까지/JX 신문_10/NNG+협회/NNG+와/JKB 체결하/VV+겠/EP+다/EF+고/JKQ
	설명	→ ‘와’는 부사격 조사이므로 ‘와/JKB’로 분석해야 함.
49	오분석	화이트/NNG 큐빅/NNG+과/JC 어울리/VV+어/EC
	정분석	화이트/NNG 큐빅/NNG+과/JKB 어울리/VV+어/EC
	설명	→ ‘과’는 부사격 조사인데 접속 조사로 분석됨.

○ 부사와 접미사

50	오분석	전국_03/NNG+평가_03/NNG+교수_06/NNG+단_05/MAG+을/JKO 15+일_07/NNB 결성하/VV+고/EC
	정분석	전국_03/NNG+평가_03/NNG+교수_06/NNG+단_22/XSN+을/JKO 15+일_07/NNB 결성하/VV+고/EC
	설명	→ ‘단’은 부사가 아니라 접미사이므로 ‘단_22/XSN’로 분석해야 함.

○ 관형사와 대명사

51	오분석	민간/NNG+사업자/NNG+들_09/XSN+이/JKS 이_05/MM 갈/VA+은/ETM
	정분석	민간/NNG+사업자/NNG+들_09/XSN+이/JKS 이/NP 갈/VA+은/ETM
	설명	→ 대명사 ‘이’를 관형사로 잘못 분석함.

○ 명사와 동사

52	오분석	개표기/NNG 등_05/NNB+에/JKB 대한_07/NNP+증거/NNG+보전_03/NNG 신청_01/NNG+을/JKO
	정분석	개표기/NNG 등_05/NNB+에/JKB 대하_02/VV+ㄴ/ETM+증거/NNG+보전_03/NNG 신청_01/NNG+을/JKO
	설명	→ ‘대한’은 ‘대한민국’의 의미가 아니라 ‘대하다’의 활용형이므로 ‘대하_02/VV+ㄴ/ETM’로 분석되어야 함.
53	오분석	남_01/NNG+의/JKG 집_01/NNG 처마_01/NNG+밑_01/NNG+에서/JKB 자_14/NNG
	정분석	남_01/NNG+의/JKG 집_01/NNG 처마_01/NNG+밑_01/NNG+에서/JKB 자/VV
	설명	→ ‘자’는 동사의 활용형인데 일반 명사로 분석됨.

○ 명사와 접사

54	오분석	재_01/NNG+개표_03/NNG 등_05/NNB
	정분석	재_17/XPN+개표_03/NNG 등_05/NNB
	설명	→ ‘재개표’의 ‘재’는 ‘다시’를 뜻하는 접두사로서 ‘재_17/XPN’로 처리해야 함.
55	오분석	조혜련/NNP+의/JKG 그네_02/NNG+분만_01/NNG 코믹/NNG 출산_02/NNG+기_13/NNG
	정분석	조혜련/NNP+의/JKG 그네_02/NNG+분만_01/NNG 코믹/NNG 출산_02/NNG+기/XSN
	설명	→ ‘기’는 접미사인데 일반 명사로 분석됨.

○ 부사와 고유 명사

56	오분석	인제_01/MAG+고성_06/NNP+을/JKO
	정분석	인제/NNP+고성_06/NNP+을/JKO
	설명	→ 고유 명사 ‘인제’를 부사로 잘못 분석함.

○ 조사와 접미사

57	오분석	그래픽/NNG+ 비만_01/NNG+수술_05/NNG 개념/NNG+도/JX
	정분석	그래픽/NNG+ 비만_01/NNG+수술_05/NNG 개념/NNG+도/XSN
	설명	→ ‘그림’의 의미를 가지는 접미사 ‘도’를 보조사로 분석함.

○ 명사와 조사

58	오분석	김영배/NNP+金令培/SH+서울_01/NNP 양천/NNP+을/JKO+ 의원_05/NNG+에게/JKB
	정분석	김영배/NNP+金令培/SH+서울_01/NNP 양천/NNP+을_01/NNG+ 의원_05/NNG+에게/JKB
	설명	→ 선거구 ‘을’을 의미하는 ‘을_01/NNG’을 목적격조사 ‘을/JKO’로 잘못 분석함.

○ 일반 명사와 어미

59	오분석	이재운/NNP+씨_07/NNB+다_89/NNG
	정분석	이재운/NNP+씨_07/NNB+다/EF
	설명	→ 어미 ‘다’를 일반 명사로 잘못 분석함.

○ 동사와 접미사

60	오분석	영국_01/NNP+의/JKG 웰컴/NNG 살_01/VV+는/ETM
	정분석	영국_01/NNP+의/JKG 웰컴/NNG 살_01/VV+는/JKC
	설명	→ ‘웰컴사는’의 ‘사’는 접미사인데 동사 어간으로 분석하고 이어지는 ‘는’도 조사인데 관형사형 어미로 잘못 분석함.

○ 일반 명사와 의존 명사

61	오분석	인기_01/NNG 절정_03/NNG 꽃_01/NNG 모양_02/NNG 두_01/MM 가지_04/NNB 톤_01/NNB+의/JKG
	정분석	인기_01/NNG 절정_03/NNG 꽃_01/NNG 모양_02/NNG 두_01/MM 가지_04/NNB 톤/NNB+의/JKG
	설명	→ 일반 명사 ‘톤’이 의존 명사로 분석됨.

○ 종결 어미와 연결 어미

62	오분석	전화_07/NNG+걸_02/VV+어/EC 보_01/VX+겠/EP+조/EC
	정분석	전화_07/NNG+걸_02/VV+어/EC 보_01/VX+겠/EP+조/EF
	설명	→ ‘조’는 ‘지요’의 준말로서 종결 어미인데 연결 어미로 분석함.

○ 축약과 관련된 품사 분석 오류

63	오분석	총_06/MM+학장_02/NNG+이사장/NNG
	정분석	총/NNG+학장_02/NNG+이사장/NNG
	설명	→ ‘총’은 ‘총장’의 의미이므로 관형사가 아님.
64	오분석	최근/NNG 교육부/NNG+의/JKG 국_01/NNG+공립_01/NNG+대_04/NNG
	정분석	최근/NNG 교육부/NNG+의/JKG 국/NNG+공립_01/NNG+대_04/NNG
	설명	→ ‘국’은 ‘국립’의 의미이므로 ‘마시는 국’의 의미를 가지는 ‘국_01’이 아님.
65	오분석	전_08/MM+현직_01/NNG 은행_02/NNG 직원_03/NNG+이/JKS
	정분석	전/NNG+현직_01/NNG 은행_02/NNG 직원_03/NNG+이/JKS
	설명	→ ‘전’은 ‘전직’의 의미이므로 관형사로 분석할 수 없음.

③ 품사 통용 관련 오류

66	오분석	완전_01/NNG 자율/NNG 규제/NNG+에서/JKB 직접/MAG 규제/NNG+를/JKO 강화하_02/VV+ㄴ/ETM
	정분석	완전_01/NNG 자율/NNG 규제/NNG+에서/JKB 직접/NNG 규제/NNG+를/JKO 강화하_02/VV+ㄴ/ETM
	설명	→ ‘직접’은 명사 앞에 쓰였으므로 부사가 아니라 명사로 분석해야 함.
67	오분석	한편/NNG+ 서울_01/NNP+고법_02/NNG 형사_02/NNG+10+부_15/NNB+는/JX
	정분석	한편/MAG+ 서울_01/NNP+고법_02/NNG 형사_02/NNG+10+부_15/NNB+는/JX
	설명	→ ‘한편’은 ‘앞서 말한 측면과 다른 측면’의 의미이므로 부사인데 명사로 분석함.

68	오분석	아니_02/IC 이것/NP+이/JKS 진짜/NNG
	정분석	아니_02/IC 이것/NP+이/JKS 진짜/MAG
	설명	→ 부사 ‘진짜’를 일반 명사로 분석함.
69	오분석	유통_04/NNG+기한_03/NNG+이/JKS 한참/NNG 지나/VV+았/EP+을/ETM 것_01/NNB
	정분석	유통_04/NNG+기한_03/NNG+이/JKS 한참/MAG 지나/VV+았/EP+을/ETM 것_01/NNB
	설명	→ 부사 ‘한참’을 일반 명사로 분석함.
70	오분석	그_01/NP+는/JX 혼자_01/NNG 여름휴가/NNG+를/JKO 보내/VV+던/ETM
	정분석	그_01/NP+는/JX 혼자/MAG 여름휴가/NNG+를/JKO 보내/VV+던/ETM
	설명	→ 부사 ‘혼자’를 일반 명사로 분석함.
71	오분석	그리하/VV+여도/EC 영세민/NNG+이/JKS 젤/NNG 편_07/NNG+하/XSA+ 여/EF
	정분석	그리하/VV+여도/EC 영세민/NNG+이/JKS 젤/MAG 편_07/NNG+하/XSA+ 여/EF
	설명	→ ‘젤’은 부사인데 일반 명사로 분석함.

#### ④ 오분석에 따른 오류

##### ○ 과분석

72	오분석	나_03/NP+ㄴ/JX+亂/SH+개발/NNG
	정분석	난개발/NNG
	설명	→ ‘난개발’은 ‘난개발/NNG’로 하나의 단어임.
73	오분석	서_06/NNG+수원_01/NNP+과/JC
	정분석	서수원/NNP
	설명	→ ‘서수원’은 전체가 하나의 고유 명사로 분석되어야 함.
74	오분석	4+6+차_03/NNB+로/JKB 고속도로/NNG+를/JKO
	정분석	4+6+차로/NNG 고속도로/NNG+를/JKO
	설명	→ 일반 명사 ‘차로’를 의존 명사와 부사격 조사로 분석함.
75	오분석	불_15/XPN+합격하/VV+였/EP+다며/EC 불합격/NNG+취소_01/NNG 소송 _01/NNG+과/JKB
	정분석	불합격하/VV+였/EP+다며/EC 불합격/NNG+취소_01/NNG 소송_01/NNG+과/JKB
	설명	→ 사전에 ‘불합격하다’가 등재되어 있으므로 이를 접두사와 동사로 더 분석할 필요가 없음.
76	오분석	따르_01/VV+아서/EC 재_17/XPN+검표/NNG 여부_01/NNG+가/JKS
	정분석	따라서/MAJ 재_17/XPN+검표/NNG 여부_01/NNG+가/JKS
	설명	→ 접속 부사 ‘따라서’를 동사와 어미로 분석함.
77	오분석	재_17/XPN+검표/NNG+를/JKO 실시_03/NNG+해_01/NNG
	정분석	재_17/XPN+검표/NNG+를/JKO 실시해/VV
	설명	→ ‘실시하다’의 활용형 ‘실시해’의 ‘해’를 ‘태양’의 의미를 가지는 일반 명사로 분석함.
78	오분석	차량_01/NNG+에/JKB 한_01/MM+해_01/NNB
	정분석	차량_01/NNG+에/JKB 한해/VV
	설명	→ ‘한하다’의 활용형 ‘한해’의 ‘해’를 ‘태양’의 의미를 가지는 일반 명사 로 분석하고 그 앞의 ‘한’은 관형사로 분석함.
79	오분석	일부_02/NNG 대학_01/NNG+이/JKS 자체_02/NNG+적_18/XSN+으로 /JKB +총장_01/NNG+평가_03/NNG+제_19/XSN+ 도입/NNG+을/JKO
	정분석	일부_02/NNG 대학_01/NNG+이/JKS 자체적/MM+으로/JKB +총장 _01/NNG+평가_03/NNG+제_19/XSN+ 도입/NNG+을/JKO
	설명	→ ‘자체적’은 하나의 단어인데 이를 일반 명사 ‘자체’와 접미사 ‘적’으로 분석함.

80	오분석	고혈압/NNG+당뇨병/NNG 등_05/NNB 비_01/NNG+만/JX 관련/NNG 합병증/NNG+을/JKO
	정분석	고혈압/NNG+당뇨병/NNG 등_05/NNB 비만/NNG 관련/NNG 합병증/NNG+을/JKO
	설명	→ ‘비만’은 일반 명사인데 이를 ‘자연 현상’인 ‘비’와 보조사 ‘만’으로 분석함.
81	오분석	이환/NNP+의/JKG 前/SH+한나라_02/NNG 부총재/NNG
	정분석	이환의/NNP 前/SH+한나라_02/NNG 부총재/NNG
	설명	→ 고유 명사 ‘이환의’를 ‘이환’과 관형격 조사 ‘의’로 분석함.
82	오분석	서울_01/NNP+지검_02/NNG 나_03/NP+의/JKG 컴퓨터/NNG+가/JKS
	정분석	서울_01/NNP+지검_02/NNG 내/NNG 컴퓨터/NNG+가/JKS
	설명	→ ‘안’의 의미를 가지는 일반 명사 ‘내’를 ‘나의’로 분석함.
83	오분석	연락_02/NNG+을/JKO 취_04/NNG+하/XSV+여/EC 모이_01/VV+시/EP+ㄴ/ETM
	정분석	연락_02/NNG+을/JKO 취하/VV+여/EC 모이_01/VV+시/EP+ㄴ/ETM
	설명	→ 동사 ‘취하다’가 존재함.
84	오분석	도_11/NNG+계장_03/NNG+이/JKS 많/Va+았/EP+기/ETN 때문/NNB
	정분석	도계장/NNG+이/JKS 많/Va+았/EP+기/ETN 때문/NNB
	설명	→ 일반 명사 ‘도계장’을 ‘도’와 ‘계장’으로 분석함.
85	오분석	황남/NNG+땡_01/NNG+으로/JKB 유명_01/NNG+하/XSA+다/EF
	정분석	황남땡/NNP+으로/JKB 유명_01/NNG+하/XSA+다/EF
	설명	→ ‘황남땡’ 전체를 고유 명사로 분석해야 하나 이를 일반 명사 ‘황남’과 ‘땡’으로 분석함.
86	오분석	목석/NNG+원_18/XSN+가든_88/NNG
	정분석	목석원/NNP+가든_88/NNG
	설명	→ ‘목석원’ 전체를 고유 명사로 분석해야 하나 이를 일반 명사 접미사로 분석함.
87	오분석	옥류/NNG+정_33/NNB
	정분석	옥류정/NNP
	설명	→ ‘옥류정’ 전체를 고유 명사로 분석해야 하나 이를 일반 명사와 의존 명사로 분석함.
88	오분석	추_02/VV+어/EF+鰕魚/SH+에/JKB
	정분석	추어/NNG+鰕魚/SH+에/JKB
	설명	→ ‘추어’는 일반 명사인데 이를 동사와 어미로 분석함.
89	오분석	팔_01/NNG+영루_01/NNG+횃집_02/NNG+043+647+8632+은/JX
	정분석	팔영루/NNP+횃집_02/NNG+043+647+8632+은/JX
	설명	→ ‘팔영루’는 고유 명사인데 이를 일반 명사와 일반 명사로 분석함.
90	오분석	하_01/VV+지만/EC 한편/NNG+에서/JKB+는/JX
	정분석	하지만/MAJ 한편/NNG+에서/JKB+는/JX
	설명	→ 접속 부사 ‘하지만’을 어간과 어미로 분석함.
91	오분석	질_02/VA+은/ETM 보_01/VX+라/EC+
	정분석	질_02/VA+은/ETM 보라/NNG
	설명	→ ‘보라’는 일반 명사인데 보조 용언 어간과 어미로 분석됨.
92	오분석	파란색/NNG 큐빅/NNG 피_01/VV+ㄴ/ETM 4+천_03/NR+원_01/NNB+
	정분석	파란색/NNG 큐빅/NNG 핀/NNG 4+천_03/NR+원_01/NNB+
	설명	→ ‘핀’은 일반 명사인데 동사 어간과 어미로 분석됨.
93	오분석	커다랗/Va+ㄴ/ETM 살구/NNG+색_03/NNG 알_01/NNG+이/JKS 신비스럽/Va+ㄴ/ETM 느낌/NNG+이/VCP+다/EF
	정분석	커다랗/Va+ㄴ/ETM 살구색/NNG 알_01/NNG+이/JKS 신비스럽/Va+ㄴ/ETM 느낌/NNG+이/VCP+다/EF
	설명	→ ‘살구색’은 하나의 일반 명사인데 ‘살구’와 ‘색’으로 나뉘어 각각 일반 명사로 분석됨.



94	오분석	의도_02/NNG+적_18/XSN+으로/JKB 가지/VV+ㄴ/ETM 휴식기/NNG+가/JKC 아니/VCN+라/EC
	정분석	의도적/NNG으로/JKB 가지/VV+ㄴ/ETM 휴식기/NNG+가/JKC 아니/VCN+라/EC
	설명	→ ‘의도적’은 하나의 단어인데 일반 명사 ‘의도’와 접미사 ‘적’으로 분석됨.
95	오분석	신동엽/NNP+은/JX 방송_01/NNG+계_19/XSN+에서/JKB
	정분석	신동엽/NNP+은/JX 방송계/NNG+에서/JKB
	설명	→ ‘방송계’는 하나의 일반 명사인데 일반 명사 ‘방송’과 접미사 ‘계’로 분석됨.
96	오분석	백_05/NR+배_09/NNG 사죄_04/NNG+하/XSV+고/EC
	정분석	백배/MAG 사죄_04/NNG+하/XSV+고/EC
	설명	→ ‘백배’는 하나의 부사인데 수사 ‘백’과 일반 명사 ‘배’로 분석됨.
97	오분석	좌식_01/NNG+분_08/NNB+만_01/NNB+의/JKG 일종_03/NNG+이/VCP+ㄴ/ETM
	정분석	좌식_01/NNG+분만/NNG+의/JKG 일종_03/NNG+이/VCP+ㄴ/ETM
	설명	→ ‘분만’은 일반 명사인데 이를 각각 의존 명사로 분석함.
98	오분석	산모_02/NNG+의/JKG 신체_02/NNG+적_18/XSN+이/VCP+ㄴ/ETM 조건_02/NNG+으로/JKB
	정분석	산모_02/NNG+의/JKG 신체적/NNG+이/VCP+ㄴ/ETM 조건_02/NNG+으로/JKB
	설명	→ ‘신체적’은 한 단어인데 이를 일반 명사 ‘신체’와 접미사 ‘적’으로 분석함.
99	오분석	신문_10/NNG+사_41/XSN 직접/NNG+규제/NNG+2+월_02/NNB+까지/JX
	정분석	신문사/NNG 직접/NNG+규제/NNG+2+월_02/NNB+까지/JX
	설명	→ ‘신문사’는 한 단어인데 이를 ‘신문_10/NNG+사_41/XSN’로 분석함.

### ○ 미분석

100	오분석	교육부/NNG+는/JX +반성중/NNP
	정분석	교육부/NNG+는/JX +반성_01/NNG+중_04/NNB
	설명	→ ‘반성_01/NNG+중_04/NNB’으로 분석해야 할 것을 전체로 하나의 고유 명사로 분석.
101	오분석	서면조사/NNG+를/JKO
	정분석	서면/NNG+조사/NNG+를/JKO
	설명	→ 일반 명사 ‘서면’과 ‘조사’로 더 분석해야 함.
102	오분석	현금_04/NNG+지급기/NNG+에서/JKB
	정분석	현금_04/NNG+지급_01/NNG+기_44/XSN+에서/JKB
	설명	→ ‘지급기’는 사전에 등재되어 있지 않으므로 ‘지급_01/NNG+기_44/XSN’로 분석해야 함.
103	오분석	몇/MM+분간/NNG+이나/JX 지속_01/NNG+되/XSV+었/EP+을까/EF
	정분석	몇/MM+분/NNB+간/XSN+이나/JX 지속_01/NNG+되/XSV+었/EP+을까/EF
	설명	→ ‘분간’은 의존명사 ‘분’과 접미사 ‘간’으로 분석해야 하는데 일반 명사 ‘분간’으로 분석함.
104	오분석	반짝이/NNG 컬러_01/NNG+ 큐빅/NNG 귀티_02/NNG+나_01/VV+는/ETM 백금은/NNP 여기_01/NP
	정분석	반짝이/NNG 컬러_01/NNG+ 큐빅/NNG 귀티_02/NNG+나_01/VV+는/ETM 백금/NNG+은/NNG 여기_01/NP
	설명	→ 일반 명사 ‘백금’과 보조사 ‘은’으로 분석되어야 할 것이 고유 명사로 분석됨.
105	오분석	업스타일/NNG+할_02/NNB 때_01/NNG
	정분석	업스타일/NNG+하/XSV+ㄹ/ETM 때_01/NNG
	설명	→ ‘할’은 ‘하/XSV+ㄹ/ETM’로 분석되어야 하는데 의존 명사로 분석됨.

106	오분석	검찰_02/NNG+이/JKS 금융감독원/NNG+의/JKG 사전_13/NNG 조사_30/NNG+과정_03/NNG+을/JKO
	정분석	검찰_02/NNG+이/JKS 금융감독원/NNG+의/JKG 사전_13/NNG 조사_30/NNG+과정_03/NNG+을/JKO
	설명	→ ‘금융 감독원’은 구인데 이를 일반 명사로 분석함.
107	오분석	글쎄_01/IC+ 잘_02/MAG 있_01/VA+다고/EC 하_01/VV+여야/EC 하_01/VX+≒지/EC 모르/VV+겠/EP+네요/EF
	정분석	글쎄_01/IC+ 잘_02/MAG 있_01/VA+다고/EC 하_01/VV+여야/EC 하_01/VX+≒지/EC 모르/VV+겠/EP+네/EF+요/JX
	설명	→ ‘네요’는 종결 어미 ‘네’와 보조사 ‘요’로 분석되어야 함.
108	오분석	그거/NP 보_01/VV+고/EC 있_01/VX+엿/EP+다니까요/EF
	정분석	그거/NP 보_01/VV+고/EC 있_01/VX+엿/EP+다니까/EF+요/JX
	설명	→ ‘다니까요’는 종결 어미 ‘다니까’와 보조사 ‘요’로 분석되어야 함.

##### ⑤ 직접 성분 분석과 관련된 오류

###### ○ 단순 직접 성분 분석 오류

109	오분석	전국_03/NNG+평가_03/NNG+교수_06/NNG+단장_08/NNG+은/JX
	정분석	전국_03/NNG+평가_03/NNG+교수단/NNG+장_13/NNG+은/JX
	설명	→ ‘교수단장’은 ‘교수단/NNG+장_13/NNG’으로 분석해야 함.
110	오분석	교육/NNG+인적자원/NNG+부_15/NNG+는/JX
	정분석	교육/NNG+인적/NNG+자원/NNG+부/XSN+는/JX
	설명	→ ‘교육 인적 자원부’로 분석되어야 할 것이 ‘교육 인적자원 부’로 잘못 분석됨. 또한 ‘부’는 일반 명사가 아니라 접미사에 해당함.
111	오분석	이_05/MM+모양_02/NNG+의/JKG
	정분석	이/NNP+모/NP+양/XSN
	설명	→ ‘모양’을 ‘模樣’으로 간주하여 하나의 명사로 분석하였으나 ‘모’는 ‘某’의 의미를 가지는 대명사이고 ‘양’은 ‘孃’의 의미를 가지는 접미사이므로 ‘이/NNP+모/NP+양/XSN’으로 분석되어야 함.
112	오분석	지역_03/NNG+구민_03/NNG+을/JKO
	정분석	지역구/NNG+민_07/XSN+을/JKO
	설명	→ ‘지역구/NNG+민_07/XSN’으로 분석되어야 할 것을 ‘지역’과 ‘구민’으로 잘못 분석함.

###### ○ 띄어쓰기 오류에 따른 직접 성분 분석 오류

113	오분석	표토_01/NNG+총_02/NNG+이대량/NNP 제거되/VV+기/ETN
	정분석	표토_01/NNG+총_02/NNG+이/JKS+대량/MAG 제거되/VV+기/ETN
	설명	→ 원래 ‘표토총이 대량 제거되기’로 되어야 할 것이 띄어쓰기 오류로 붙여 적혀 분석 오류가 발생함.

## 2) 구어의 오류 유형

### ① 품사 분석 오류

#### ○ 일반 명사와 고유 명사

125	오분석	하나/NNP+를/JKO 중심_01/NNG+으로/JKB
	정분석	하나/NNG+를/JKO 중심_01/NNG+으로/JKB
	설명	→ ‘하나’는 일반 명사인데 고유 명사로 분석함.
126	오분석	신세계/NNP 혼돈_01/NNG+에서/JKB
	정분석	신세계/NNG 혼돈_01/NNG+에서/JKB
	설명	→ ‘신세계’는 일반 명사인데 고유 명사로 분석함.
127	오분석	위_01/NNG+에다/JKB 안_03/NNP
	정분석	위_01/NNG+에다/JKB 안/NNG
	설명	→ 일반 명사 ‘안’을 고유 명사로 분석함.

#### ○ 명사와 동사

128	오분석	번다/NNG
	정분석	번다/VV
	설명	→ 동사 ‘벌다’의 활용형을 일반 명사로 분석함.
129	오분석	가고_01/NNG
	정분석	가고/VV
	설명	→ 동사 ‘가다’의 활용형을 일반 명사로 분석함.
130	오분석	식_04/NNG+을/JKO 만하/VV+ㄴ/ETM 것_01/NNB+이/VCP+ㄴ 데/EC
	정분석	식을/VV+ 만하/VV+ㄴ/ETM 것_01/NNB+이/VCP+ㄴ 데/EC
	설명	→ 동사 ‘식다’의 활용형을 일반 명사와 조사의 결합으로 분석함.

#### ○ 명사와 접사

131	오분석	사_18/NNG 분_01/NNB+의/JKG
	정분석	사_18/NNG 분/XSN+의/JKG
	설명	→ ‘분’은 접미사인데 의존 명사로 분석함.
132	오분석	심지어/MAG 반_07/NNG+종교적/NNG+이/JKC
	정분석	심지어/MAG 반/XPN+종교적/NNG+이/JKC
	설명	→ ‘반’은 접두사인데 이를 일반 명사로 분석함.

#### ○ 명사와 조사

133	오분석	빨리/MAG 철학/NNG+과/JKB 좀_02/MAG 소개하_01/VV+여/EC 보_01/VX+아요/EF
	정분석	빨리/MAG 철학/NNG+과/NNG 좀_02/MAG 소개하_01/VV+여/EC 보_01/VX+아요/EF
	설명	→ ‘과’는 일반 명사인데 부사격 조사로 분석함.
134	오분석	철학/NNG+과/JC
	정분석	철학/NNG+과/NNG
	설명	→ ‘과’는 일반 명사인데 접속 조사로 분석함.

## ○ 종결 어미와 연결 어미

135	오분석	아니/VCN+ㄴ가/EC+요/JX
	정분석	아니/VCN+ㄴ가/EF+요/JX
	설명	→ ‘ㄴ가’는 종결 어미인데 연결 어미로 분석함.
136	오분석	어_02/IC 그거/NP 맞_01/VV+는/ETM 거_01/NNB 같/Va+애/EC
	정분석	어_02/IC 그거/NP 맞_01/VV+는/ETM 거_01/NNB 같/Va+애/EF
	설명	→ ‘애’는 구어적 표현이기는 하지만 종결 어미인데 연결 어미로 분석함.
137	오분석	설득력/NNG 있_01/Va+네/EC
	정분석	설득력/NNG 있_01/Va+네/EF
	설명	→ ‘네’는 종결 어미인데 연결 어미로 분석함.
138	오분석	어떻/Va+게/EC 좀_02/MAG 도움/NNG+을/JKO 받_01/VV+으면/EC 좋_01/Va+을/ETM 것_01/NNB+이/VCP+냐/EC
	정분석	어떻/Va+게/EC 좀_02/MAG 도움/NNG+을/JKO 받_01/VV+으면/EC 좋_01/Va+을/ETM 것_01/NNB+이/VCP+냐/EF
	설명	→ ‘냐’는 종결 어미인데 연결 어미로 분석함.
139	오분석	나이_01/NNG+가/JKS 들_01/VV+면서/EC 점점_01/MAG 크_01/Va+어요/EC
	정분석	나이_01/NNG+가/JKS 들_01/VV+면서/EC 점점_01/MAG 크_01/Va+어요/EF
	설명	→ ‘어요’는 종결 어미인데 연결 어미로 분석함.
140	오분석	이_05/MM+것_01/NNB+이/VCP+에요/EC
	정분석	이_05/MM+것_01/NNB+이/VCP+에요/EF
	설명	→ ‘에요’는 종결 어미인데 연결 어미로 분석함.
141	오분석	하_01/VV+였/EP+나/EC
	정분석	하_01/VV+였/EP+나/EF
	설명	→ 종결 어미 ‘나’를 연결 어미로 분석함.
142	오분석	바로_02/MAG 이런_01/MM 이유_04/NNG+이/VCP+에요/EC
	정분석	바로_02/MAG 이런_01/MM 이유_04/NNG+이/VCP+에요/EF
	설명	→ ‘에요’는 종결 어미인데 연결 어미로 분석함.
143	오분석	이거_01/NP 분석하_02/VV+였/EP+다/EC
	정분석	이거_01/NP 분석하_02/VV+였/EP+다/EF
	설명	→ ‘다’는 종결 어미인데 연결 어미로 분석함.
144	오분석	계산하/VV+는/ETM 것_01/NNB 있_01/Va+지/EC+않/VX+아요/EC
	정분석	계산하/VV+는/ETM 것_01/NNB 있_01/Va+지/EC+않/VX+아요/EF
	설명	→ ‘아요’는 종결 어미인데 연결 어미로 분석함.
145	오분석	되_01/VV+ㄹ/ETM 수_02/NNB+가/JKS 있_01/Va+겠/EP+느냐/EC
	정분석	되_01/VV+ㄹ/ETM 수_02/NNB+가/JKS 있_01/Va+겠/EP+느냐/EF
	설명	→ 종결 어미 ‘느냐’를 연결 어미로 분석함.
146	오분석	광장히/MAG 중요하_02/Va+ㅁ니다/EC
	정분석	광장히/MAG 중요하_02/Va+ㅁ니다/EF
	설명	→ 종결 어미 ‘ㅁ니다’를 연결 어미로 분석함.
147	오분석	그렇/Va+지/EC 않/VX+겠/EP+습니까/EC
	정분석	그렇/Va+지/EC 않/VX+겠/EP+습니까/EF
	설명	→ 종결 어미 ‘습니까’를 연결 어미로 분석함.
148	오분석	그_01/MM+ 붙잡/VV+자/EC 놀이_01/NNG+를/JKO 하_01/VV+거나/EC
	정분석	그_01/MM+ 붙잡/VV+자/EF 놀이_01/NNG+를/JKO 하_01/VV+거나/EC
	설명	→ 종결 어미 ‘자’를 연결 어미로 분석함.
149	오분석	사람/NNG 정말_01/MAG 힘들/Va+거든요/EC
	정분석	사람/NNG 정말_01/MAG 힘들/Va+거든요/EF
	설명	→ 종결 어미 ‘거든요’를 연결 어미로 분석함.

150	오분석	애기/NNG 좀_02/MAG 하_01/VV+여/EC 주_01/VX+세요/EC
	정분석	애기/NNG 좀_02/MAG 하_01/VV+여/EC 주_01/VX+세요/EF
	설명	→ 종결 어미 ‘세요’를 연결 어미로 분석함.
151	오분석	어떨/VA+게/EC 동기_07/NNG+부여_04/NNG+가/JKC 되_01/VV+ㄹ까/EC
	정분석	어떨/VA+게/EC 동기_07/NNG+부여_04/NNG+가/JKC 되_01/VV+ㄹ까/EF
	설명	→ 종결 어미 ‘ㄹ까’를 연결 어미로 분석함.
152	오분석	뒤/NP+이/VCP+야/EC
	정분석	뒤/NP+이/VCP+야/EF
	설명	→ 종결 어미 ‘야’를 연결 어미 ‘야’로 분석함.

○ 축약과 관련된 품사 분석 오류

153	오분석	석_88/NNG+박사_01/NNG 쿼터_02/NNG+제_19/XSN 없이/MAG
	정분석	석사/NNG+박사_01/NNG 쿼터_02/NNG+제_19/XSN 없이/MAG
	설명	→ ‘석사’의 축약인 ‘석’을 일반 명사로 분석함.

○ 접속 부사와 일반 부사

154	오분석	근까/MAG
	정분석	그러니까/MAJ
	설명	→ ‘근까’는 ‘그러니까’의 줄임말로 일반 부사가 아니라 접속 부사로 분석되어야 함.

○ 감탄사와 명사

155	오분석	음_04/NNG+ 것_01/NNB+도/JX 나타나/VV+고/EF
	정분석	음/IC+ 것_01/NNB+도/JX 나타나/VV+고/EF
	설명	→ 감탄사 ‘음’을 일반 명사로 분석함.
156	오분석	예_88/NNG
	정분석	예/IC
	설명	→ 감탄사 ‘예’를 일반 명사로 분석함.
157	오분석	하_05/NNP 그것/NP+이/JKS
	정분석	하/IC 그것/NP+이/JKS
	설명	→ 감탄사 ‘하’를 고유 명사로 분석함.
158	오분석	예예/NNP
	정분석	예예/IC
	설명	→ ‘예예’는 감탄사인데 고유 명사로 분석됨.
159	오분석	자_14/NNG 그_01/MM 다음_01/NNG 레벨_01/NNG+이/JKS
	정분석	자/IC 그_01/MM 다음_01/NNG 레벨_01/NNG+이/JKS
	설명	→ 감탄사 ‘자’를 일반 명사로 분석함.
160	오분석	아우_01/NNG 이상하/VA+여/EC
	정분석	아우/IC 이상하/VA+여/EC
	설명	→ ‘아우’는 감탄사인데 일반 명사로 분석함.

○ 감탄사와 조사

161	오분석	예/JKB 그것/NP+이/JKS
	정분석	예/IC 그것/NP+이/JKS
	설명	→ 감탄사 ‘예’를 부사격 조사로 분석함.

○ 감탄사와 접미사

162	오분석	어_08/XSN
	정분석	어/IC
	설명	→ 감탄사 ‘어’를 접미사로 분석함.

○ 감탄사와 대명사

163	오분석	어디_01/IC
	정분석	어디/NP
	설명	→ 대명사 ‘어디’를 감탄사로 분석함.

○ 감탄사와 어미

164	오분석	으/EC+ 제_01/NP+가/JKS 만들/VV+ㄴ/ETM
	정분석	으/IC+ 제_01/NP+가/JKS 만들/VV+ㄴ/ETM
	설명	→ 감탄사 ‘으’를 어미로 분석함.

○ 수사와 일반 명사

165	오분석	사_18/NNG 분_01/NNB+의/JKG
	정분석	사/NR 분_01/NNB+의/JKG
	설명	→ ‘사’는 수사인데 일반 명사로 분석함.

○ 지정사와 접미사

166	오분석	지배_01/NNG 계층/NNG+인_17/XSN+가/JKS
	정분석	지배_01/NNG 계층/NNG+이/VCP+ㄴ가/EF
	설명	→ 지정사 ‘이’와 어미 ‘ㄴ가’를 접미사 ‘인’과 주격 조사로 분석함.

○ 지정사와 명사

167	오분석	이거_01/NP 찍_02/VV+어/EC 이/VCP+ㄴ지/EC 치료/NNG
	정분석	이거_01/NP 찍_02/VV+어/EC 인지/NNG 치료/NNG
	설명	→ ‘인지’는 일반 명사인데 지정사 ‘이’와 어미 ‘ㄴ지’로 분석함.

○ 관형사와 명사

168	오분석	한_11/NNP
	정분석	한/MM
	설명	→ 관형사 ‘한’을 고유 명사로 분석함.

○ 조사와 어미

169	오분석	어떻게/VA+게/EC 기르/VV+았/EP+나/JC
	정분석	어떻게/VA+게/EC 기르/VV+았/EP+나/EF
	설명	→ ‘나’는 종결 어미인데 접속 조사로 분석함..

○ 조사와 관형사

170	오분석	반_07/NNG 정도_11/NNG+에다가/JKB 이제_01/MAG 이런_01/MM 이런_01/MM 안_04/NNG+들_09/XSN+이_05/MM 있_01/VA+습니다/EF
	정분석	반_07/NNG 정도_11/NNG+에다가/JKB 이제_01/MAG 이런_01/MM 이런_01/MM 안_04/NNG+들_09/XSN+이/JKS 있_01/VA+습니다/EF
	설명	→ ‘이’는 주격 조사인데 관형사로 분석함.

② 오분석에 따른 오류

○ 과분석

171	오분석	잘_02/MAG 자_01/VV+ㄹ/ETM 세계화/NNG 모르/VV+아요/EF
	정분석	잘_02/MAG 잘_02/MAG 세계화/NNG 모르/VV+아요/EF
	설명	→ 부사 ‘잘’을 동사 어간과 어미로 분석함.
172	오분석	세계_02/NNG 이_05/MM 세계_02/NNG+화_16/XSN 아니/VCN+ㄴ 데/EC
	정분석	세계_02/NNG 이_05/MM 세계화/NNG 아니/VCN+ㄴ 데/EC
	설명	→ ‘세계화’는 한 단어인데 이를 일반 명사 ‘세계’와 접미사 ‘화’로 분석함.
173	오분석	비_32/XPN+종교/NNG+적_18/XSN+이/JKC
	정분석	비_32/XPN+종교적/NNG+이/JKC
	설명	→ ‘비종교적’은 접두사 ‘비’와 명사 ‘종교적’으로 분석해야 하는데 ‘종교적’도 일반 명사 ‘종교’와 접미사 ‘적’으로 더 분석함.
174	오분석	그치/IC 그릴/VA+지/EC
	정분석	그치/IC 그릴지/IC
	설명	→ 감탄사 ‘그릴지’를 형용사 어간과 어미로 분석함.
175	오분석	근본/NNG+주의_02/NNG 자체_02/NNG+는/JX
	정분석	근본주의/NNG 자체_02/NNG+는/JX
	설명	→ ‘근본주의’의 하나의 단어로서 일반 명사인 데 이를 각각 일반 명사 ‘근본’과 ‘주의’로 분석함.
176	오분석	많/VV+은/ETM 사회_07/NNG+적_18/XSN+이/VCP+ㄴ /ETM
	정분석	많/VV+은/ETM 사회적/NNG+이/VCP+ㄴ /ETM
	설명	→ ‘사회적’은 한 단어인데 일반 명사 ‘사회’와 접미사 ‘적’으로 분석함.
177	오분석	말_03/VV+아/EC+구_15/NNG 생각_01/NNG+을/JKO 하_01/VV+느라고/EC
	정분석	마구/MAG 생각_01/NNG+을/JKO 하_01/VV+느라고/EC
	설명	→ 부사 ‘마구’를 동사 어간과 어미, 일반 명사의 연결로 분석함.
178	오분석	중요시/NNG 하_01/VV+는/ETM 부분_01/NNG+이/VCP+구요/EC
	정분석	중요시하/VV+는/ETM 부분_01/NNG+이/VCP+구요/EC
	설명	→ ‘중요시하다’는 하나의 단어인데 일반 명사 ‘중요시’와 동사 ‘하다’로 분석함.
179	오분석	발_01/NNG+음_04/NNG+이/JKS 나쁜_01/VA+ㄴ /ETM 애_02/NNG+들_09/XSN
	정분석	발음/NNG+이/JKS 나쁜_01/VA+ㄴ /ETM 애_02/NNG+들_09/XSN
	설명	→ 일반 명사 ‘발음’을 일반 명사 ‘발’과 ‘음’으로 분석함.
180	오분석	뇌_03/NNG+성_17/XSN+마비_02/NNG
	정분석	뇌성/NNG+마비_02/NNG
	설명	→ 일반 명사 ‘뇌성’을 일반 명사 ‘뇌’와 접미사 ‘성’으로 분석함.
181	오분석	가만/MAG+히_07/XSB 있_01/VA+는/ETM 아이_01/NNG+들_09/XSN
	정분석	가만히/MAG 있_01/VA+는/ETM 아이_01/NNG+들_09/XSN
	설명	→ ‘가만히’는 하나의 단어인데 이를 부사 ‘가만’과 접미사 ‘히’로 분석함.
182	오분석	상호_04/NNG 작용_01/NNG 되_01/VV+고/EC 나_03/NP+ㄴ /JX
	정분석	상호_04/NNG 작용되/VV+고/EC 나_03/NP+ㄴ /JX
	설명	→ ‘작용되고’는 하나의 동사인 데 일반 명사 ‘작용’과 동사로 분석함.

183	오분석	너_01/NP+의/JKG 문서/NNG 정리_09/NNG 작업_01/NNG
	정분석	네/IC 문서/NNG 정리_09/NNG 작업_01/NNG
	설명	→ 감탄사 ‘네’를 대명사와 조사로 분석함.
184	오분석	전체적/NNG+으로/JKB 하_01/VV+ㄴ/ETM 세_01/MM 명_03/NNB 정도_11/NNG 예상하/VV+고/EC 있_01/VX+습니다/EF
	정분석	전체적/NNG+으로/JKB 한/MM 세_01/MM 명_03/NNB 정도_11/NNG 예상하/VV+고/EC 있_01/VX+습니다/EF
	설명	→ 관형사 ‘한’을 동사 어간과 관형형 전성 어미로 분석함.
185	오분석	개별/NNG+적_13/NNG+으로/JKB 접촉하/VV+는/ETM 방법/NNG
	정분석	개별적/NNG+으로/JKB 접촉하/VV+는/ETM 방법/NNG
	설명	→ ‘개별적’은 하나의 단어인데 이를 일반 명사 ‘개별’과 접미사 ‘적’으로 분석함.
186	오분석	추계_01/NNG 학술/NNG+제_20/XSN 회_08/NNB+의/JKG 지난번/NNG+이/JKS
	정분석	추계_01/NNG 학술/NNG+제_20/XSN 회의/NNG 지난번/NNG+이/JKS
	설명	→ 일반 명사 ‘회의’를 의존 명사 ‘회’와 조사 ‘의’로 분석함.
187	오분석	어제_01/MAG 뭐/IC 문의_03/NNG+하_01/VV+러/EC 오_01/VV+았었/EP+거든/EF
	정분석	어제_01/MAG 뭐/IC 문의하/VV+러/EC 오_01/VV+았었/EP+거든/EF
	설명	→ ‘문의하다’는 하나의 동사인데 일반 명사 ‘문의’와 동사 ‘하다’로 분석함.
188	오분석	논문/NNG 작성_01/NNG+자_14/NNG 전용_04/NNG 공간_05/NNG+화_16/XSN+로/JKB 하_01/VV+는/ETM 것_01/NNB+는/JX
	정분석	논문/NNG 작성자/NNG 전용_04/NNG 공간_05/NNG+화_16/XSN+로/JKB 하_01/VV+는/ETM 것_01/NNB+는/JX
	설명	→ ‘작성자’는 하나의 일반 명사인데 일반 명사 ‘작성’과 ‘자’로 분석함.
189	오분석	일일이_02/MAG 일일이_02/MAG 일_05/NR 대_11/NNB 일_05/NR+로/JKB 다_03/MAG+수_02/NNB 없_01/VA+으니까/EC
	정분석	일일이_02/MAG 일일이_02/MAG 일대일/NNG+로/JKB 다_03/MAG+수_02/NNB 없_01/VA+으니까/EC
	설명	→ ‘일대일’은 하나의 일반 명사인데 수사, 의존 명사 수사로 분석함.

○ 미분석

190	오분석	노동률/NNP+이/JKS
	정분석	노동/NNP+률/XSN+이/JKS
	설명	→ ‘노동률’은 사전에 없으므로 일반 명사 ‘노동’과 접미사 ‘률’로 분석되어야 하나 이를 고유 명사로 분석함.
191	오분석	그것/NP+이/JKS 국제화랑/NNP
	정분석	그것/NP+이/JKS 국제화/NNG+랑/JKB
	설명	→ ‘국제화랑’은 ‘국제화/NNG+랑/JKB’으로 분석되어야 함.
192	오분석	이렇/VA+게/EC 떼_01/NNG 붙이/VV+ㄴ다면/EC
	정분석	이렇/VA+게/EC 떼_01/VV+어/EC 붙이/VV+ㄴ다면/EC
	설명	→ ‘떼’는 ‘떼_01/VV+어/EC’로 분석되어야 하는데 일반 명사로 처리됨.
193	오분석	애_03/NP+네_08/XSN+들_09/XSN+도/JX 그_01/MM 근본주원데/NNG
	정분석	애_03/NP+네_08/XSN+들_09/XSN+도/JX 그_01/MM 근본주의/NNG+ㄴ데/EC
	설명	→ ‘근본주원데’는 일반 명사 ‘근본주의’와 어미 ‘ㄴ데’로 분석되어야 함.
194	오분석	노력하_01/VV+여/EC 온_88/NNG
	정분석	노력하_01/VV+여/EC 오/VV+ㄴ/ETM
	설명	→ ‘온’은 어간 ‘오’와 관형형 전성 어미 ‘ㄴ’으로 분석되어야 하는데 일반 명사로 처리됨.



195	오분석	이것/NP+이/JKS 더_01/MAG 설득력/NNG+이/JKS 있_01/VA+조/EC
	정분석	이것/NP+이/JKS 더_01/MAG 설득력/NNG+이/JKS 있_01/VA+지/EF+요/JX
	설명	→ ‘조’는 종결 어미 ‘지’와 보조사 ‘요’로 분석해야 함.
196	오분석	어_02/IC 많이/MAG 보_01/VV+시/EP+엇/EP+네요/EC
	정분석	어_02/IC 많이/MAG 보_01/VV+시/EP+엇/EP+네/EF+요/JX
	설명	→ ‘네요’는 종결 어미 ‘네’와 보조사 ‘요’로 분석해야 함.
197	오분석	기지_12/NNG+도/JX 못하/VV+고/EC
	정분석	기/VV+지/EC+도/JX 못하/VV+고/EC
	설명	→ ‘기지’는 동사 어간과 어미로 분석해야 하는데 일반 명사로 분석함.
198	오분석	연락처/NNG+를/JKO 절_08/NNG 일단_01/MAG 주_01/VV+세요/EF
	정분석	연락처/NNG+를/JKO 저/NP+를/JKO 일단_01/MAG 주_01/VV+세요/EF
	설명	→ ‘절’은 대명사 ‘저’와 조사 ‘를’로 분석해야 하는데 전체를 일반 명사로 분석함.
199	오분석	지지난/MM+준가_02/NNG 비_01/NNG 많이/MAG 오_01/VV+았/EP+을/ETM 때_01/NNG+도/JX
	정분석	지지난/MM+주/NNG+이/VCP+ㄴ가/EF 비_01/NNG 많이/MAG 오_01/VV+았/EP+을/ETM 때_01/NNG+도/JX
	설명	→ ‘준가’는 일반 명사 ‘주’, 지정사, 어미 ‘ㄴ가’로 분석되어야 함.
200	오분석	수요_06/NNG 조사_30/NNG+를/JKO 어떨/VA+게/EC 하_01/VV+시/EP+ㄹ/ETM 생각_01/NNG+이신지/NNP 모르/VV+겠/EP+는데/EC
	정분석	수요_06/NNG 조사_30/NNG+를/JKO 어떨/VA+게/EC 하_01/VV+시/EP+ㄹ/ETM 생각_01/NNG+이/VCP+시/EP+ㄴ지/EC 모르/VV+겠/EP+는데/EC
	설명	→ ‘이신지’는 지정사와 어미의 결합인데 고유 명사로 분석함.

### ③ 직접 성분 분석과 관련된 오류

#### ○ 단순 직접 성분 분석 오류

201	오분석	누구/NP+가/JKS 주_33/XPN+첸/NNP+데_01/NNB
	정분석	누구/NP+가/JKS 주체/NNG+이/VCP+ㄴ데/EC
	설명	→ ‘주첸데’는 일반 명사 ‘주체’와 어미 ‘ㄴ데’로 분석되어야 함.

### ④ 구어적 특성에 따른 오류

#### ○ 줄임말 관련 분석 오류

202	오분석	그_01/MM+ 곳_01/VV+ㄹ까/EC
	정분석	그/IC+ 그러니까/MAJ
	설명	→ 부사 ‘그러니까’의 줄임말 ‘그까’를 동사 어간과 어미로 분석함.
203	오분석	민족/NNG 국_10/XSN+간_10/NNB
	정분석	민족/NNG 국가/NNG+이/VCP+ㄴ/JX
	설명	→ ‘민족 국가인’의 ‘국가인’을 ‘국간’으로 줄여 적으므로 일반 명사 ‘국가’와 지정사 ‘이’, 관형형 전성 어미 ‘ㄴ’을 접미사 ‘국’과 의존 명사 ‘간’으로 분석함.
204	오분석	아저씨/NNG+하_01/VV+겠/EP+지/EC+X
	정분석	아저씨/NNG+이/VCP+겠/EP+지/EC+X
	설명	→ ‘아저씨겠지’는 ‘아저씨이겠지’의 준말인데 ‘이’ 대신 ‘하’를 복원하여 분석함.
205	오분석	이케/MAG 어렵/VA+ㄴ/ETM 영어_02/NNP+가/JKS 나오/VV+아도/EC
	정분석	이러하/VA+게/EC 어렵/VA+ㄴ/ETM 영어_02/NNP+가/JKS 나오/VV+아도/EC
	설명	→ ‘이케’는 ‘이렇게’의 준말이므로 ‘이러하/VA+게/EC’로 분석되어야 하는데 부사로 분석됨.

206	오분석	대부분/NNG 어떤/MM 심저_01/NNG 어떤/MM 엄마/NNG+들_09/XSN+은/JX
	정분석	대부분/NNG 어떤/MM 심저어/MAG 어떤/MM 엄마/NNG+들_09/XSN+은/JX
	설명	→ ‘심저’는 부사 ‘심저어’의 준말인데 이를 일반 명사로 분석함.
207	오분석	그_01/NP+이/VCP+니까/EC
	정분석	그러니까/MAJ
	설명	→ ‘그니까’는 부사 ‘그러니까’의 준말인데 대명사 ‘그’와 지정사 ‘이’, 어미 ‘니까’로 분석함.
208	오분석	마지막/NNG 전체_01/NNG 회원데/NNG
	정분석	마지막/NNG 전체_01/NNG 회의/NNG+이/VCP+ㄴ 데/EC
	설명	→ ‘회원데’는 ‘회의인데’의 준말로서 일반 명사 ‘회의’와 지정사, 어미 ‘ㄴ 데’로 분석되어야 함.

○ 말 더듬에 따른 분석 오류

209	오분석	성_07/NNG 성립/NNG+이/VCP+라기/ETN+보다/JKB
	정분석	성/IC 성립/NNG+이/VCP+라기/ETN+보다/JKB
	설명	→ ‘성립’의 ‘성’을 더듬은 것인데 이를 일반 명사로 분석함.
210	오분석	제_21/XPN 제_01/NP+가/JKS
	정분석	제/IC 제_01/NP+가/JKS
	설명	→ ‘제가’의 ‘제’를 더듬은 것인데 이를 접두사로 분석함.
211	오분석	칠십/NR+년대/NNB 시_06/NNG 시작되_01/VV+ㄴ/ETM 이슬람교/NNG
	정분석	칠십/NR+년대/NNB 시/IC 시작되_01/VV+ㄴ/ETM 이슬람교/NNG
	설명	→ ‘시작된’의 ‘시’를 더듬은 것인데 이를 일반 명사로 분석함.
212	오분석	또/MAG 일부_02/NNG+만/JX 제_01/NP+가/JKS 쓰_01/VV+어/EC 놓_01/VX+은/ETM+논_01/NNG 것_01/NNB+이/VCP+에요/EF
	정분석	또/MAG 일부_02/NNG+만/JX 제_01/NP+가/JKS 쓰_01/VV+어/EC 놓_01/VX+은/ETM+놓_01/VX+은/ETM 것_01/NNB+이/VCP+에요/EF
	설명	→ ‘논’은 ‘놓은’을 더듬은 것인데 일반 명사로 분석함.
213	오분석	유_05/NNP 유지하_02/VV+ㄴ 다는/ETM 것_01/NNB+는/JX
	정분석	유/IC 유지하_02/VV+ㄴ 다는/ETM 것_01/NNB+는/JX
	설명	→ ‘유’는 ‘유지하다’를 더듬은 것인데 고유 명사로 분석함.

○ 말 끊음에 따른 분석 오류

214	오분석	저_04/MM
	정분석	저/NP
	설명	→ 문맥을 보면 ‘저’는 대명사인데 관형사로 분석함.
215	오분석	아무튼/MAG 그럴/VA+였/EP+고/EC 제_21/XPN
	정분석	아무튼/MAG 그럴/VA+였/EP+고/EC 저/NP+의/JKG
	설명	→ ‘제’는 ‘저의’의 준말인데 접두사로 분석함.

○ 경음 표현에 따른 분석 오류

216	오분석	그러니까/MAJ 쫓_02/VV+ㅁ/ETN
	정분석	그러니까/MAJ 쫓/MAJ
	설명	→ ‘쫓’은 부사 ‘쫓’의 구어 표현인데 이를 동사 어간과 어미로 분석함.
217	오분석	다른/MM+따르_01/VV+ㄴ/ETM 정신_12/NNG+과/JC 문제_06/NNG+나/JC
	정분석	다른/MM+다른/MM 정신_12/NNG+과/JC 문제_06/NNG+나/JC
	설명	→ ‘따른’은 관형사 ‘다른’의 구어 표현인데 이를 동사 어간과 어미로 분석함.

218	오분석	머리_01/NNG+가/JKS      꼬꿈/NNG+씩_03/XSN      좇_01/VA+아/EC+지_04/VX+기/ETN 시작하_01/VV+여요/EF
	정분석	머리_01/NNG+가/JKS      조금/NNG+씩_03/XSN      좇_01/VA+아/EC+지_04/VX+기/ETN 시작하_01/VV+여요/EF
	설명	→ ‘꼬꿈’은 ‘조금’의 구어 표현인데 이를 일반 명사로 분석함.

○ 기타

219	오분석	보통/NNG 아이_01/NNG+들_09/XSN+두_01/MM
	정분석	보통/NNG 아이_01/NNG+들_09/XSN+도/JX
	설명	→ ‘두’는 조사 ‘도’의 구어체 표현인데 관형사로 분석함.
220	오분석	요고_01/NNG 요고_01/NNG 다_03/MAG 관촬/VA+거든요/EF
	정분석	요거/NP 요거/NP 다_03/MAG 관촬/VA+거든요/EF
	설명	→ ‘요고’는 대명사 ‘요거’의 구어적 표현인데 이를 일반 명사로 분석함.
221	오분석	고론_02/NNG 애_02/NNG+들_09/XSN+이/JKS
	정분석	그런/MM 애_02/NNG+들_09/XSN+이/JKS
	설명	→ ‘고론’은 관형사 ‘그런’의 구어적 표현인데 이를 일반 명사로 분석함.
222	오분석	그러은/IC+요/JX
	정분석	그러면/MAG+요/JX
	설명	→ 부사 ‘그러면’의 구어적 표현 ‘그러은’을 감탄사로 분석함.
223	오분석	그르_01/VA+ㄴ/ETM 식_04/NNB+으로/JKB
	정분석	그런/MM 식_04/NNB+으로/JKB
	설명	→ ‘그른’은 관형사 ‘그런’의 구어적 표현인데 이를 동사 어간과 어미로 분석함.
224	오분석	다_03/MAG 정리_09/NNG 됐그든/NNG+요/JX
	정분석	다_03/MAG 정리_09/NNG 되/VV+였/EP+-거든/EF+요/JX
	설명	→ ‘됐그든’은 용언 활용형 ‘됐거든’의 구어적 표현인데 이를 일반 명사로 분석함.
225	오분석	차등_03/NNG+을/JKO 돌려/NNG+고/JKQ 하_01/VV+고/EC
	정분석	차등_03/NNG+을/JKO 두/VV+려/EC+고/JKQ 하_01/VV+고/EC
	설명	→ ‘돌려고’는 ‘두려고’의 구어적 표현인데 이를 일반 명사 ‘돌려’와 인용격 조사로 분석함.
226	오분석	어_02/IC 무슨/MM 말_01/NNG+이/VCP+ㄴ 지/EC 몰_88/NNG+를/JKO 거_01/NNB 같/VA+애서/EC
	정분석	어_02/IC 무슨/MM 말_01/NNG+이/VCP+ㄴ 지/EC 모르/VV+ㄹ/JKG 거_01/NNB 같/VA+애서/EC
	설명	→ ‘몰를’은 ‘모를’을 구어적 표현인데 이를 일반 명사와 조사로 분석함.

연구 책임자: 이삼형 (한양대학교 국어교육과 교수)  
 공동 연구원: 박진호 (서울대학교 국어국문학과 교수)  
                   최형용 (이화여자대학교 국어국문학과 교수)  
                   김정선 (한양대학교 국어교육과 교수)  
                   신명선 (인하대학교 국어교육과 교수)  
                   신동광 (광주교육대학교 영어교육과 교수)  
                   강남옥 (경인교육대학교 국어교육과 교수)  
                   이기연 (국립국어원 학예연구사)  
                   김시정 (수원대학교 교양학부 객원교수)  
 연구 보조원: 김수지 (한양대학교 국어교육과 박사과정)  
 보 조 원: 이윤희 (한양대학교 국어교육과 석사과정)  
                   양세문 (한양대학교 국어교육과 석사과정)  
 담당 연구원: 이기연 (국립국어원 학예연구사)

---

발 행 인	송 철 의
발 행 처	국립국어원 서울시 강서구 금남화로 154 전화: 02-2669-9775    전송: 02-2669-9727
인 쇄 일	2017년 12월 20일
발 행 일	2017년 12월 20일

---

\* 이 책은 국립국어원의 용역비로 수행한 ‘국어 기초 어휘 선정 및 어휘 등급화를 위한 기초 연구’ 사업의 결과물을 발간한 것입니다.