

국립국어원 2015-01-42

발간등록번호
11-1371028-000579-01

# 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업

연구 책임자  
강 현 화



# 제 출 문

국립국어원장 귀하

“2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업”에 관하여 귀 원과 체결한 연구용역 계약에 의하여 연구보고서를 작성하여 제출합니다.

2015년 12월 18일

연구 책임자: 강현화(연세대학교)

연구 기관	연세대학교 산학협력단
연구 책임자	강현화(연세대)
공동 연구원	김선정(계명대), 김일환(고려대), 김정숙(고려대), 안경화(서울대), 이동은(국민대), 이정희(경희대), 한송화(연세대), 황용주(국립국어원)
연구 보조원	홍혜란(연세대), 강민석(고려대), 공나형(연세대), 김경아(연세대), 김형주(경희대), 송지혜(연세대), 신범숙(서울대), 이민아(국민대) 홍종호(계명대), 유성희(연세대)



## 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업

이 연구는 한국어 학습자 말뭉치 구축을 위한 기본 계획을 수립하고, 기초 연구를 통해 말뭉치 구축 지침을 마련한 후 그에 따라 약 35만 어절 규모의 말뭉치를 시험 구축하는 데에 주요한 목적이 있다.

**한국어 학습자 말뭉치 구축 기본 계획 수립:** 한국어 학습자 말뭉치는 3단계 6년간의 사업을 통해 총 370만 어절 규모의 말뭉치(문어 300만 어절, 구어 70만 어절)를 수집한다. 수집 대상은 국내 학습자, 이주민, 국외 학습자이다.

**한국어 학습자 말뭉치 구축 지침 수립:** 한국어 학습자 말뭉치 구축 지침으로 수집 지침, 자료 처리 지침, 문어 입력 지침, 구어 전사 지침, 형태 주석 지침, 오류 주석 지침을 마련하였다. 각 단계별 지침은 국가 주도 말뭉치로서의 일관성과 체계성 유지, 비모어 화자 자료와의 비교를 위해 <21세기 세종 한국어 균형 말뭉치>의 지침을 근간으로 하되, 비모어 화자의 텍스트/말화 특성을 고려하여 수정·보완하였다.

**한국어 학습자 말뭉치 실제 수집, 구축, 가공:** 2015년 한국어 학습자 말뭉치는 앞선 단계에서 마련한 지침을 토대로 원시 말뭉치 35만 어절(문어 30만 어절, 구어 5만 어절), 형태 주석 말뭉치 22만(문어 20만, 구어 2만 어절), 오류 주석 5만 어절(문어 4만, 구어 1만)이 구축되었다.

그 외에도 체계적인 말뭉치 구축을 위하여 구축/가공 인력 실무 교육 계획을 수립하고 한국어 학습자 말뭉치 사업을 널리 알리기 위하여 홍보 계획을 마련하였다. 홍보 계획에는 사용자들의 활용 능력 제고를 위한 교육 프로그램도 포함된다. 그 외에도 자료의 효율적 활용을 위해 한국어 학습자 말뭉치 활용 방안을 제시하였다.

<한국어 학습자 말뭉치>는 <21세기 세종 한국어 균형 말뭉치> 이후로 구축되는 국가 주도의 말뭉치로서의 위상을 가진다. <한국어 학습자 말뭉치>는 한국어 비모어 화자인 외국인 학습자의 자료를 수집하여 구축한 것으로

연구, 교육, 환경에 다음과 같이 작용함으로써 한국어의 세계화 및 국제 경쟁력 강화에 이바지할 것이다.

- [연구] 학습자 말뭉치 활용 연구를 통한 국제 수준의 학술 교류 기반 조성
- [교육] 한국어교육 이론의 체계화 및 교육 자료 구축의 기반 조성
- [학습] 한국어 학습자의 다양화에 따른 교수·학습 환경의 과학화

주요어: 한국어 학습자 말뭉치, 문어 말뭉치, 구어 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치

# 차 례

I. 연구 개요 .....	1
1. 연구의 목적 및 필요성 .....	1
1.1. 연구의 목적 .....	1
1.2. 연구의 필요성 .....	1
2. 연구의 범위 .....	2
3. 연구 방법 .....	2
4. 연구 추진 일정 .....	3
5. 연구 결과 .....	4
II. 말뚝치 구축 기본 계획 수립 .....	5
1. 중장기 구축 계획 .....	5
1.1. 기본 계획 .....	5
1.2. 주요 쟁점과 계획 수립의 방향 .....	9
2. 사업 조직의 구성 및 운영 .....	18
2.1. 기본 계획 .....	18
2.2. 주요 쟁점과 계획 수립의 기본 방향 .....	20
3. 총 예산 측정 .....	21
3.1. 기본 계획 .....	21
3.2. 주요 쟁점과 계획 수립의 기본 방향 .....	36
4. 자료 수집의 방향 .....	37
4.1. 자료 수집 방법 .....	37
4.2. 수집 범위 .....	39
5. 말뚝치 구축/가공 인력 실무 교육 및 홍보 .....	42
5.1. 교육 .....	42
5.2. 홍보 .....	44

### Ⅲ. 말뭉치 구축 지침 수립 ..... 46

1. 수집 지침 .....	46
1.1. 주요 쟁점 .....	46
1.2. 지침 수립의 기본 방향 .....	46
2. 자료 처리 지침 .....	52
2.1. 주요 쟁점 .....	52
2.2. 지침 수립의 기본 방향 .....	53
3. 문어 자료 입력 지침 .....	56
3.1. 주요 쟁점 .....	56
3.2. 지침 수립의 기본 방향 .....	56
4. 구어 자료 전사 지침 .....	58
4.1. 주요 쟁점 .....	58
4.2. 지침 수립의 기본 방향 .....	59
4.3. 2016년 구어 전사 지침 및 기구축 자료 보완 방향 .....	73
5. 마크업 지침 .....	76
5.1. 주요 쟁점 .....	76
5.2. 지침 수립의 기본 방향 .....	77
6. 형태 주석 지침 .....	80
6.1. 주요 쟁점 .....	80
6.2. 지침 수립의 기본 방향 .....	81
6.3. 2016년 형태 주석 지침 및 기구축 자료 보완 방향 .....	88
7. 오류 주석 틀(태그 세트) 및 주석 작업 지침 .....	88
7.1. 주요 쟁점 .....	88
7.2. 지침 수립의 기본 방향 .....	89
7.3. 2016년 오류 주석 지침 및 기구축 자료 보완 방향 .....	108

### Ⅵ. 1차 연도 한국어 학습자 말뭉치 구축의 실제 ..... 108

1. 구축 과정 및 절차 .....	108
2. 단계별 세부 구축 내용 및 결과 .....	109
2.1. 자료 수집 .....	109
2.2. 자료 처리 .....	115

2.3. 원시 말뭉치 구축 .....	115
2.4. 형태 주석 말뭉치 구축 .....	121
2.5. 오류 주석 말뭉치 구축 .....	133
3. 온라인 구축 시스템 개발을 위한 협의 작업 .....	149
3.1. 한국어 학습자 말뭉치 구축과 온라인 구축 시스템의 연계 모형 .....	149
3.2. 한국어 학습자 말뭉치 구축 도구의 활용 .....	152
3.3. 2016년 한국어 학습자 말뭉치 구축 도구 활용 및 추가 기능 제안 .....	155
4. 말뭉치 구축/가공 인력 실무 교육 및 홍보 .....	155
4.1. 단계별 수집 지침 배포 및 온/오프라인 교육, 워크숍 .....	155
4.2. 학술대회 발표 .....	156

## **V. 말뭉치 활용 방안 연구 ..... 157**

1. 사용자 집단별 활용 모형 .....	157
1.1. 연구자를 위한 활용 모형 .....	157
1.2. 한국어 교사를 위한 활용 모형 .....	162
1.3. 학습자를 위한 활용 모형 .....	163
2. 말뭉치 활용을 위한 검색 기능 설계(안) .....	165
2.1. 말뭉치 검색 기능 요약 .....	165
2.2. 검색 기능에 관한 세부 내용 .....	167

## **VI. 결론 및 제언 ..... 176**

1. 연구 요약 .....	176
2. 연구의 의의 및 기대 효과 .....	178
3. 보고서 활용 방안 .....	180
4. 정책 제언 .....	181

## **참고문헌**

## 표 차례

<표 1> 연구의 범위와 세부 과업 .....	2
<표 2> 연구 수행 방법 .....	2
<표 3> 단계별 한국어 학습자 말뭉치 구축 일정 .....	3
<표 4> 한국어 학습자 말뭉치 구축 결과 .....	4
<표 5> 학습자 말뭉치 연차별 구축 계획 .....	5
<표 6> 학습자 말뭉치 연차별 자료 구축 계획 .....	7
<표 7> 수집 범위 및 대상 .....	9
<표 8> 한국어 학습자 말뭉치 총 예산 .....	21
<표 9> 한국어 학습자 말뭉치 총 예산과 수집 규모 .....	21
<표 10> 한국어 학습자 말뭉치 구축 사업비 총예산 .....	22
<표 11> 한국어 학습자 말뭉치 2차 연도 예산과 수집 규모 .....	25
<표 12> 한국어 학습자 말뭉치 구축 사업비 2차 연도 예산 .....	25
<표 13> 한국어 학습자 말뭉치 3차 연도 예산과 수집 규모 .....	27
<표 14> 한국어 학습자 말뭉치 구축 사업비 3차 연도 예산 .....	27
<표 15> 한국어 학습자 말뭉치 4차 연도 예산과 수집 규모 .....	29
<표 16> 한국어 학습자 말뭉치 구축 사업비 4차 연도 예산 .....	29
<표 17> 한국어 학습자 말뭉치 5차 연도 예산과 수집 규모 .....	31
<표 18> 한국어 학습자 말뭉치 구축 사업비 5차 연도 예산 .....	31
<표 19> 한국어 학습자 말뭉치 6차 연도 예산과 수집 규모 .....	32
<표 20> 한국어 학습자 말뭉치 구축 사업비 6차 연도 예산 .....	33
<표 21> 한국어 학습자 말뭉치 구축 사업비 모듈 단위 예산: 문어 10만 어절 .....	34
<표 22> 한국어 학습자 말뭉치 구축 사업비 모듈 단위 예산: 구어 1만 어절 .....	35
<표 23> 한국어 학습자 말뭉치의 수집 경로 .....	37
<표 24> 수집 대상 기관과 수집 방법에 따른 학습자 말뭉치 .....	40
<표 25> 수집 대상 기관과 수집 방법에 따른 학습자 말뭉치 .....	41
<표 26> 말뭉치 구축/가공 인력 교육 내용 .....	42
<표 27> 한국어 학습자 말뭉치 활용 아카데미 개최 계획 .....	45
<표 28> 연구 윤리 준수 조건을 고려한 말뭉치 구축 지침 .....	47
<표 29> 말뭉치 구축 대상 자료 선별 지침 .....	53
<표 30> 한국어 학습자 말뭉치 파일명 코드 .....	55

<표 31> <21세기 세종 균형 말뭉치>와 <2015 한국어 학습자 말뭉치> 전사 지침의 비교 .....	59
<표 32> 전사 마크업 체계 수정안 .....	73
<표 33> 한국어 학습자 말뭉치의 헤더 정보 항목 .....	77
<표 34> 형태소 분석 표지 .....	82
<표 35> 형태소 분석 결과 중 분석 불능 및 오분석 예시 .....	84
<표 36> 오류가 발생한 부분의 형태소 분석 결과와 처리 방안 .....	85
<표 37> 형태소 분석 결과 중 분석 불능 및 오분석 수정 예시 .....	86
<표 38> 한국어 학습자 말뭉치의 오류 주석 틀 및 주석 표지 .....	91
<표 39> 국외 학습자 말뭉치의 오류 주석 체계 .....	97
<표 40> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-음운 층위 .....	102
<표 41> Corder(1981)의 오류 유형 .....	102
<표 42> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-형태 층위 .....	104
<표 43> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-통사 층위 .....	105
<표 44> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-담화 층위 .....	106
<표 45> 오류 판정과 교정의 기준 .....	107
<표 46> 한국어 학습자 말뭉치 구축 과정 및 절차 .....	108
<표 47> 수집 기관별 실무 책임자 및 실무 교사 명단 .....	110
<표 48> 국적별 종적 말뭉치 수집 대상자 분포 .....	114
<표 49> 원시 말뭉치의 수준별 자료 규모 .....	118
<표 50> 원시 말뭉치의 국적별 자료 규모 .....	118
<표 51> 형태 주석 말뭉치의 수준별 자료 규모 .....	124
<표 52> 형태 주석 말뭉치의 국적별 자료 규모 .....	124
<표 53> 형태 주석 결과 분석: 전체 말뭉치의 수준별 품사 분포 .....	127
<표 54> 형태 주석 결과 분석: 문어 말뭉치의 수준별 품사 분포 .....	129
<표 55> 형태 주석 결과 분석: 구어 말뭉치의 수준별 품사 분포 .....	131
<표 56> 오류 주석 말뭉치의 수준별 자료 규모 .....	136
<표 57> 오류 주석 말뭉치의 국적별 자료 규모 .....	136
<표 58> 오류 주석 결과의 분석: 분석 여부 .....	137
<표 59> 오류 주석 결과의 분석: 전체 말뭉치의 오류 현상 .....	138
<표 60> 오류 주석 결과의 분석: 문어 말뭉치의 오류 현상 .....	138
<표 61> 오류 주석 결과의 분석: 구어 말뭉치의 오류 현상 .....	139
<표 62> 오류 주석 결과의 분석: 전체 말뭉치의 오류 층위 .....	139

<표 63> 오류 주식 결과의 분석: 문어 말뭉치의 오류 층위 .....	140
<표 64> 오류 주식 결과의 분석: 구어 말뭉치의 오류 층위 .....	140
<표 65> 오류 주식 결과의 분석: 전체 말뭉치의 오류 층위-발음 .....	141
<표 66> 오류 주식 결과의 분석: 문어 말뭉치의 오류 층위-발음 .....	141
<표 67> 오류 주식 결과의 분석: 구어 말뭉치의 오류 층위-발음 .....	142
<표 68> 오류 주식 결과의 분석: 전체 말뭉치의 오류 층위-어휘 .....	142
<표 69> 오류 주식 결과의 분석: 문어 말뭉치의 오류 층위-어휘 .....	143
<표 70> 오류 주식 결과의 분석: 구어 말뭉치의 오류 층위-어휘 .....	144
<표 71> 오류 주식 결과의 분석: 전체 말뭉치의 오류 층위-문법 .....	145
<표 72> 오류 주식 결과의 분석: 문어 말뭉치의 오류 층위-문법 .....	146
<표 73> 오류 주식 결과의 분석: 구어 말뭉치의 오류 층위-문법 .....	147
<표 74> 오류 주식 결과의 분석: 전체 말뭉치의 오류 층위-담화 .....	148
<표 75> 오류 주식 결과의 분석: 문어 말뭉치의 오류 층위-담화 .....	148
<표 76> 온라인 구축 시스템을 활용한 작업의 범위 .....	151
<표 77> 말뭉치 구축/가공 인력 실무 교육 및 홍보 내용 .....	155
<표 78> 한국어 학습자 말뭉치를 활용한 연구 사례 .....	157
<표 79> 말뭉치 검색 기능 설계(안) .....	166
<표 80> 조건 검색의 세부 변인 .....	167

## 그림 차례

<그림 1> 자료 수집 조직 체계 .....	18
<그림 2> 자료 구축 조직 체계 .....	19
<그림 3> 지역별 거점 기관을 통한 오프라인 방식의 자료 수집 .....	37
<그림 4> 온라인 자료 수집 시스템을 활용한 수집 체계도 .....	38
<그림 5> 한국어 학습자 말뭉치 자료 처리 절차 .....	53
<그림 6> 한국어 학습자 말뭉치 파일명 부여 체계 .....	54
<그림 7> 학습자 자료 예시: 한글에 없는 글자 .....	57
<그림 8> 학습자 자료 예시: 띄어쓰기 구분이 어려운 경우 .....	58
<그림 9> 헤더 마크업 예시 .....	79
<그림 10> 형태소 단위 중심의 오류 주석 예시 .....	90
<그림 11> 지역별 거점 기관을 중심으로 한 협의체 구성 .....	110
<그림 12> 자료 수집 체계도 .....	112
<그림 13> 파일 등록 예시 자료 .....	115
<그림 14> 학습자 작문 원본 스캔 파일 예시 .....	116
<그림 15> 학습자 작문 입력 파일 예시 .....	117
<그림 16> 전사 도구 엘란의 실행 화면 .....	117
<그림 17> 전사 도구 엘란으로 전사한 후 출력한 파일의 예시 .....	117
<그림 18> 형태 주석 결과의 예시 .....	122
<그림 19> XML 기반의 학습자 말뭉치 출력 파일 예시-형태 주석 .....	123
<그림 20> 오류 주석 작업 화면 예시 .....	134
<그림 21> XML 기반의 학습자 말뭉치 출력 파일 예시-오류 주석 .....	135
<그림 22> 파일 업로드 화면 예시 .....	152
<그림 23> 학습자 개인 정보 입력 화면 예시 .....	153
<그림 24> 파일 정보 입력 화면 예시 .....	153
<그림 25> 파일명 자동 생성 화면 예시 .....	154
<그림 26> 오류 주석 화면 예시 .....	154
<그림 27> 학습자 말뭉치를 활용한 학습 절차 .....	164
<그림 28> 문맥 색인 분석 예시 .....	169
<그림 29> 연어 분석 예시 .....	170
<그림 30> 연어 분석 결과의 문맥 색인 정보 예시 .....	170
<그림 31> 군집 분석 예시 .....	171

<그림 32> 키워드 분석 예시 .....	172
<그림 33> 사용 어휘 목록 예시: 형태 주석 말뭉치의 경우 .....	173
<그림 34> 사용 어휘 목록의 문맥 색인 정보 예시: 형태 주석 말뭉치의 경우 .....	174
<그림 35> 사용 어휘 목록 예시: 원시 말뭉치의 경우 .....	174
<그림 36> 사용 어휘 목록의 문맥 색인 정보 예시: 원시 말뭉치의 경우 .....	175

# I. 연구 개요

## 1. 연구의 목적 및 필요성

### 1.1. 연구의 목적

○ 본 연구의 목적은 한국어 학습자의 수준별·언어권별 문어·구어 말뭉치 자료 구축을 통해 과학적이고 객관적인 방법인 한국어교육 연구 환경을 조성하는 데에 있다. 아울러 한국어교육 현장에서 교재와 교수 자료를 개발하거나 교수·학습 방법을 구안하는 데에 활용하여 보다 체계적인 교육 기반을 조성하는 데에 있다. 이를 위한 세부 목표는 다음과 같다.

- 한국어 학습자 말뭉치 구축 기본 계획 수립
- 한국어 학습자 말뭉치 구축 지침 수립
- 한국어 학습자 말뭉치 실제 수집, 구축, 가공
- 한국어 학습자 말뭉치 구축/가공 인력 실무 교육 및 홍보
- 한국어 학습자 말뭉치 활용 방안 연구

### 1.2. 연구의 필요성

- 국가 주도의 한국어 학습자 균형 말뭉치 구축
- 학습자 말뭉치 활용 연구를 통한 국제 수준의 학술 교류에 대한 요구
- 학습자 말뭉치를 활용한 연구 방법론과 도구 사용 교육에 대한 요구
- 한국어교육 이론의 체계화 및 교육 자료 구축의 기반 마련
- 한국어 학습자의 다양화에 따른 교수·학습 환경의 과학화 필요

## 2. 연구의 범위

- 본 연구의 범위와 세부 내용은 다음과 같다.

<표 1> 연구의 범위와 세부 과업

연구의 범위	세부 내용
말뭉치 기초 연구 교육	○ 말뭉치 구축 기본 계획 수립(수집, 구축 설계) ○ 말뭉치 구축 지침 수립(태그셋 포함) ○ 말뭉치 구축/가공 인력 실무 교육 및 홍보 ○ 말뭉치 활용 방안 연구(연구 방법론 등 개발)
말뭉치 수집 및 구축 가공	○ 한국어 학습자 실제 말뭉치 자료 수집 및 구축 ○ 구축된 말뭉치 가공(형태분석, 오류분석)

## 3. 연구 방법

- 본 연구는 기초 연구와 실제 말뭉치 구축 연구가 순환적으로 이루어진다. 기초 연구는 문헌 연구, 사례 조사, 현황 조사, 전문가 의견 수렴의 네 가지 방법에 따른다. 각 연구 방법에 따른 연구 내용과 목적은 다음과 같다.

<표 2> 연구 수행 방법

	내용	목적 및 의의
문헌 연구	○ 학습자 말뭉치 설계 및 구축 이론 ○ 학습자 말뭉치의 활용, 학습자 언어 특성에 관한 선행 연구	○ 연구 방향 및 체계 수립, 본 계획 수립의 기초 자료로 활용 ○ 말뭉치 구축 기본 계획 수립의 타당성 확보 ○ 말뭉치의 활용도 제고
사례 분석	○ 국내외 학습자 말뭉치 구축 사례	○ 연구 방향 및 체계 수립, 기본 계획 수립의 기초 자료로 활용
현황 조사	○ 한국어 학습자 통계	○ 현실적인 자료 수집 가능성

	○ 수집 대상 기관 목록, 교육 과정, 학습자 분포	검토
의견 수렴	○ 한국어 학습자 말뭉치 협의체 구성을 통한 자문 및 의견수렴 ○ 사업 착수 보고회의, 중간 보고회의, 최종보고회의를 통한 정기 자문 및 평가	○ 학계의 참여 유도 ○ 연구 방향 및 방법, 절차의 타당성 검증

## 4. 연구 추진 일정

<표 3> 단계별 한국어 학습자 말뭉치 구축 일정

단계	작업 내용	1 개 월	2 개 월	3 개 월	4 개 월	5 개 월	6 개 월	7 개 월	8 개 월
기초 연구	문헌연구	●	●	●	●	●	●	●	
	사례분석	●	●	●	●	●	●	●	
	현황조사	●				●	●		
수집	자료 수집 지침 개발 및 교육	●	●	●	●	●	●	●	
	자료 수집		●	●	●	●	●	●	
구축	말뭉치 입력 및 전사 지침 개발 및 교육		●	●	●	●	●	●	
	파일 분류 및 파일링			●	●	●	●	●	
	기계 가독형 자료로의 변환(입력 및 전사)				●	●	●	●	
가공	말뭉치 가공 지침 개발 및 교육		●	●	●	●	●	●	
	형태 및 T 단위 주석				●	●	●	●	

	오류 주석				●	●	●	●	
마무리	1차 연도 사업 마무리 및 2차 연도 기초 계획 수립								●

## 5. 연구 결과

○ 연구 결과는 다음과 같다.

- 한국어 학습자 말뭉치 수집 지침
- 한국어 학습자 말뭉치 수집 자료 처리 지침
- 한국어 학습자 말뭉치 문어 입력 지침
- 한국어 학습자 말뭉치 구어 전사 지침
- 한국어 학습자 말뭉치 형태 주석 지침
- 한국어 학습자 말뭉치 오류 주석 지침
- 한국어 학습자 말뭉치 구축 결과물

<표 4> 한국어 학습자 말뭉치 구축 결과

	구분		1급	2급	3급	4급	5급	6급	합계
원시 말뭉치	문 어	어절 수	51,446	55,869	53,839	50,477	51,676	50,274	313,581
		파일 수	723	496	454	389	363	321	2,746
	구 어	어절 수	6,875	13,751	12,702	15,717	29,328	27,438	105,811
		파일 수	30	17	19	17	30	17	130
형태 주석 말뭉치	문 어	어절 수	34,077	30,109	35,377	35,136	35,231	31,618	201,548
		파일 수	483	271	312	271	248	208	1,793
	구 어	어절 수	4,251	4,722	4,789	4,725	7,070	4,844	30,401
		파일 수	17	6	7	8	8	4	50
오류 주석 말뭉치	문 어	어절 수	7,469	7,796	8,269	7,033	7,246	7,188	45,001
		파일 수	103	67	62	50	48	46	376
	구 어	어절 수	2,333	2,454	3,270	2,083	3,123	2,247	15,510
		파일 수	10	3	5	4	4	2	28

## II. 말뭉치 구축 기본 계획 수립

### 1. 중장기 구축 계획

#### 1.1. 기본 계획

##### 1) 연차별 구축 계획

- 본 연구는 3단계 6년간의 연구로 수집 대상 및 연차별 구축 계획은 다음과 같다. 각 단계별로 국내 학습자의 자료, 이주민 자료, 국외 학습자의 자료를 각각 중점적으로 수집하도록 한다. 이때 체계적인 자료 수집을 위해 전년도에 수집 설계와 시험 수집을 하고, 중점 수집 기간 이후에는 수집한 자료의 구축과 가공, 수준별·언어권별 균형을 맞추기 위한 추가 수집 작업을 한다. 또한 최신의 교수 환경과 학습자 특성을 반영한 자료를 추가 수집하도록 한다.

<표 5> 학습자 말뭉치 연차별 구축 계획

	1단계		2단계		3단계		합계
	2015년	2016년	2017년	2018년	2019년	2020년	
수집 대상 및 규모	[수집 설계 및 시험 구축/가공]	국내 학습자 [집중 수집]		[추가 수집 및 구축/가공]			문어 160만/ 구어 40만 목표(문어 160-200만 조정 가능)
		[수집 설계 및 시험 구축/가공]	이주민 [집중 수집]		[구축/가공]		문어 70만/ 구어 15만 목표(문어 50-70만 조정 가능)
			수집 설계 및 시험 구축/가공 ▪ 수집: 문어 10만/구 어 5만	국외 학습자 [집중 수집]		[구축/가공]	문어 70만/ 구어 15만 목표(문어 50-70만 조정 가능)

자료 수집	문 어	30만+a	50만+a	80만+a	90만+a	50만+a	-	300만+a
	구 어	5만+a	15만+a	20만+a	20만+a	10만+a	-	70만+a
원시 구축	문 어	30만	40만	40만	30만	20만	-	300만
	구 어	5만	10만	10만	10만	5만	-	70만
형태 주석	문 어	20만	30만	45만	50만	90만	70만	300만
	구 어	2만	8만	15만	20만	20만	5만	70만
오류 주석	문 어	4만	11만	25만	30만	25만	25만	120만
	구 어	1만	4만	10만	15만	10만	-	40만

- 자료의 수집은 학습자 군의 규모에 비례하여, 국내 학습자에 초점을 두어 동일 집단(대표 집단)의 말뭉치로서의 균형성을 높인다. 이주민의 경우에는 상대적으로 자연스러운 습득의 환경에 있으므로, 교육과정에 종속되지 않은 구어에 비중을 두어 수집한다. 아울러 국외 학습자의 경우에는 국내 학습자의 언어권의 다양성을 보완할 수 있는 언어권에 초점을 두며 동포 학습자 자료 확보에도 비중을 둔다.
- 자료의 수집과 구축은 [집중 수집] 기간 전후에 [수집 설계와 시험 구축], [추가 수집 및 구축] 기간을 두어 사업 기간 동안 최신 자료가 꾸준히 포함되도록 하며, 자료의 균형성을 확보하도록 한다. 이 중 [수집 설계와 시험 구축]은 본격적인 수집과 구축 작업에 앞서 실제 수집 환경과 자료의 질적 측면을 사전 점검하여 양적·질적으로 우수한 자료를 체계적으로 수집하기 위한 것이다.
- 이주민 자료와 국외 자료의 경우 국내 자료에 비해 수집 조건(수집 네트워크의 긴밀성, 학습 환경 등)이나 물리적인 환경(국외)이 좋지 않으므로 수집 중 수집 기간 내에 수집을 완료하는 것을 목표로 한다.

○ 연차별 세부 구축 계획은 다음과 같다.

<표 6> 학습자 말뭉치 연차별 자료 구축 계획

			2015년	2016	2017	2018	2019	2020	합계
수집	국내	문어	30만	40만	40만	30만	20만	-	160만
		구어	5만	10만	10만	10만	5만	-	40만
	이주	문어	-	10만	30만	30만	-	-	70만
		구어	-	5만	5만	5만	-	-	15만
	국외	문어	-	-	10만	30만	30만	-	70만
		구어	-	-	5만	5만	5만	-	15만
	합계	문어	30만	50만	80만	90만	50만	-	300만
		구어	5만	15만	20만	20만	10만	-	70만
구축	국내	문어	30만	40만	40만	30만	20만	-	160만
		구어	5만	10만	10만	10만	5만	-	40만
	이주	문어	-	10만	30만	30만	-	-	70만
		구어	-	5만	5만	5만	-	-	15만
	국외	문어	-		10만	30만	30만	-	70만
		구어	-	-	5만	5만	5만	-	15만
	합계	문어	30만	50만	80만	90만	50만	-	300만
		구어	5만	15만	20만	20만	10만	-	70만
형태 주석	국내	문어	20만	30만	30만	30만	30만	20만	160만
		구어	2만	8만	10만	10만	10만	-	40만
	이주	문어	-	-	10만	10만	30만	20만	70만
		구어	-	-	5만	5만	5만	-	15만
	국외	문어	-	-	-	10만	30만	30만	70만
		구어	-	-	-	5만	5만	5만	15만
	합계	문어	20만	30만	40만	50만	90만	70만	300만
		구어	2만	8만	15만	20만	20만	5만	70만
오류 주석	국내	문어	4만	11만	20만	20만	15만	10만	80만
		구어	1만	4만	5만	5만	5만	-	20만

	이주	문어	-	-	5만	5만	5만	5만	20만
		구어	-	-	5만	5만	-	-	10만
	국외	문어	-	-	-	5만	5만	10만	20만
		구어	-	-	-	5만	5만	-	10만
	합계	문어	4만	11만	25만	30만	25만	25만	120만
		구어	1만	4만	10만	15만	10만	-	40만

- 사업 전반과 중반에는 수집과 구축을 집중적으로 하고, 사업 후반에는 가공 작업을 중점적으로 하면서 필요 시 약간의 추가 수집을 할 수 있다. 마지막 연차인 2020년은 사업의 마지막 해로 잔여 분량의 가공 작업과 기구축 자료의 보완 작업을 주로 한다.

## 2) 자료 수집 범위 및 대상

- 자료 수집은 국내 학습자, 이주민, 국외 학습자를 대상으로 한다. 이주민 자료의 수집 사례 분석, 국외 학습자 자료의 시험 수집, 말뭉치 구축에 관한 설문조사 결과를 통해 실제 자료 수집에서 다음과 같은 제약이 있을 수 있음을 알 수 있었다.
  - **결혼이민자:** 정규 교육을 받지 않은 경우가 주를 이루며, 자연 발화 환경에 노출되어 자연 습득을 하는 경우가 많다. 또한 집체 교육 기관의 경우 자료 수집에 대한 보상이 필요하다. 따라서 방문 교사나 개인 교수자 등을 통해 문어 자료보다는 구어 자료를 수집하는 것이 합리적이다.
  - **다문화가정 자녀, 한글학교, 한국학교:** 아동 학습자의 자료는 아동 본인 뿐만 아니라 학교, 담임교사, 부모의 동의서를 받아야 하는데, 현실적으로 부모를 일일이 대면하여 말뭉치에 대해 설명하고 동의서를 받기가 쉽지 않다.
- 아울러 수집 대상이 다양해질수록 각 집단별 자료의 규모가 적어짐으로써 변인별 연구 시 자료의 규모가 적어질 우려가 있다. 이러한 점을 고려하여 본 연구에서는 우선 수집 대상이나 수집 자료의 유형을 다음과 같이 한정하기로 한다.

<표 7> 수집 범위 및 대상

구분	국내 학습자	이주민	국외 학습자
대상	○ 언어교육원 어학 연수생 ○ 대학, 대학원 진학 유학생	○ 결혼이민자 (교육과정 외의 구어 자료 초점) ○ 이주 노동자	○ 세종학당 ○ 국외 정규 교육기관 ○ 외국인/교포
수준	○ 초급, 중급, 고급, 최고급	○ 초급, 중급, 고급	○ 초급, 중급, 고급
언어권	○ 중국어, 일본어, 어, 그 외 수집 가능한 언어권	○ 수집 가능한 언어권	○ 다양한 언어권(보완적)

- 온라인 수집 : 온라인 수집 시스템을 설계하여, 학습자의 자발적 참여에 의한 자료 수집을 병행한다.

## 1.2. 주요 쟁점과 계획 수립의 방향

- 연차별 구축 계획은 6개년에 걸쳐 구축하게 될 한국어 학습자 말뭉치의 규모에 대한 타당성 검증과 균형성 확보, 실질적인 자료의 효용성, 자료 수집의 용이성을 파악하기 위한 기초 연구 결과를 바탕으로 수립되었다.

### 1) 전체 규모 산출의 타당성 검증

- 말뭉치의 수집에 있어 상반된 견해를 가지는 부분 중 하나는 규모에 관한 것이다. 하나는 말뭉치가 수억 어절로 구성된 BE(Bank of English)와 CIC(Cambridge International Corpus)와 같이 ‘엄청나게 큰 코퍼스(mega-corpora)’의 형태로 더 커져야 한다는 것이고, 다른 하나는 연구 목적에 맞게 특정 사용역과 장르에 초점을 둔, 더 작고 특화된 말뭉치가 구축되어야 한다는 것이다(Almut Koester, 2009). 본 연구에서는 균형성과 대표성을 갖춘 학습자 말뭉치의 적정 규모를 산출하기 위하여 통계적 접근, 말뭉치 구축 이론, 한국어 학습자의 분포에 관한 자료를 검토하여 300

만 어절을 적정 규모로 보았다.

- **통계적 접근에 기초한 적정 규모의 산출:** 국립국어원(2010)에 따라 1,000만 어절 규모의 균형 말뭉치에서 2-gram의 토큰 수를 산출한 후 그것의 증가 추이를 분석해 본 결과, 300만 어절 수준에서 안정적으로 증가하기 시작하였다. 이로써 학습자의 변인, 사용역, 장르별 연구에서 활발하게 활용되는 학습자 말뭉치는 약 300만 어절 규모에서 신뢰성 있게 작동할 수 있음을 알 수 있다.
- **말뭉치 구축 이론에 기초한 규모 산출:** 이론적 근거에 따라 ‘표본의 크기(평균 100어절)×학습자 말뭉치의 구성 범주(96개)×표본 수(250)’로 최소의 규모를 산출했을 때 약 240만 어절이 균형성을 담보하기 위한 최소 규모로 파악되었다.
- **한국어 학습자 통계에 기초한 수집 가능 규모 산출:** 300만 어절은 약 15,000여 명(1인 평균 100어절, 두 개 이내의 자료 수집 기준)으로부터 수집 가능한 분량으로 한국어 학습자 통계 분석 자료에 따르면 자료 수집에 큰 어려움이 없을 것으로 파악되었다.

#### ☞ 한국어 학습자 말뭉치의 경쟁력

○국외에서도 학습자 말뭉치의 구축이 활발하게 이루어지고 있는데, 현재 사용 가능한 주요 말뭉치들의 규모는 대개 100만 어절 내외인 경우가 많다. ICLE의 경우는 21개의 국가가 공동 참여 형식으로 구축한 말뭉치로 각 국가가 약 20만 어절 내외의 자료를 수집하여 모은 것으로 현재도 계속 구축 진행 중이다. 천만 어절 이상의 규모를 자랑하는 CLC 등은 대규모 영어 능력 평가의 작문 자료를 오랜 기간 동안 수집한 것이다. 전 세계의 영어 학습자 수와 EFL 환경을 고려할 때 한국어 학습자 말뭉치의 규모나 구성(다양성과 균형성) 면에서 매우 체계적으로 설계되었음을 알 수 있다. 이를 지속하기 위해서는 단기간 양적인 목표를 달성하는 데에 그치지 않고 급변하는 한국어 교육 환경과 교육과정, 학습자의 특성에 맞춰 지속적인 구축을 통해 자료를 업데이트해 나아가야 할 것이다.

## 2) 자료의 균형성

- 본 연구에서는 대상별 자료 수집 시 [수집 계획 수립 및 시험 구축-중점 구축-수집 자료의 구축 및 가공, 균형성 확보를 위한 추가 구축]의 세 단계로 수집하여 수집의 체계성과 균형성을 확보하기로 한다. 이에 따라 중점 수집 기간 동안에 현실적으로 수집 가능한 자료들을 최대한 수집하고, [중점 구축] 기간 이후에 각 연차별로 수집된 자료의 통계를 기반으로 상대적으로 부족한 수준과 국적의 자료를 추가 수집하여 균형성을 담보하도록 한다.

## 3) 자료의 효용성

- 자료 구축 후 사용자들이 목적에 따라 자료를 잘 활용할 수 있는가에 관한 것이다. 300어절이 적지 않은 규모이나 자료 구성 시 고려하는 변인이 많아질수록 변인별 규모가 적어진다. 따라서 학습자의 수준이나 국적과 같이 실질적인 활용에서 유의미하고 비중이 큰 일부 변인으로 수집 대상을 통제할 필요가 있다. 아울러 이론적이고 산술적인 균형성보다 학습자의 실제 분포나 연구자의 요구 등을 고려하여 현실적인 균형성을 고려하는 것이 자료의 효용성이나 수집의 용이성을 고려할 때 더 타당하다.
- 국내 학습자 자료의 경우 가장 많은 비중을 차지하고 비교적 수집이 용이하다(2015년 수집 협조가 잘 이루어짐). 고급 단계의 학습자 수가 급격히 감소하기는 하나 1-6급까지 전체 등급의 학습자가 분포한다.
- 국외 자료나 이주여성 자료의 경우 국내 학습자의 비해 전체적인 비중도 낮고 시험 수집에서 살펴본 것처럼 중·고급 단계의 학습자가 절대적으로 적어 자료의 균형성을 맞추기가 어렵다. 따라서 일정 분량의 자료를 수집하기 위해서는 국내보다 더 충분한 수집 기간이 필요하며, 그 외의 부수적인 고려 사항이 많다.
- ☞ 이에 따라 ‘국내 학습자>이주민≥국외 학습자’의 비중으로 자료를 수집하는 것이 자료의 효용성 측면에서 합리적이라고 보았다.

#### 4) 자료 수집의 용이성

○ 자료 수집은 실제 자료를 수집할 수 있는 여건이 되어야 하고, 그에 따른 협조가 이루어져야만 가능한 일이다. 본 연구에서는 2015년의 국내 자료 수집에서 28개 기관의 협조를 얻어 성공적으로 자료를 수집하였으며, 계속 수집의 가능성을 발견하였다. 중장기 계획 수립을 위해 이주민과 국외 자료를 시험 수집하여 자료 수집과 구축의 쟁점들을 파악해 보았다. 또한 설문조사를 통해 국내 기관, 이주민 대상의 기관, 국외의 대학 소속 교사들로부터 자료 수집에 관한 의견을 수렴해 보았다.

☞ 그 결과 이주민, 국외 기관의 자료 수집이 국내에 비해 용이하지 않으며, 그에 따라 수집 가능한 양도 학습자의 수에 비해 많지 않을 것으로 파악되었다. 이에 따라 국내 학습자>이주민=국외 학습자의 비중으로 자료를 수집하되, 2-6년차 사업 기간 동안 자료 수집 설계 및 시험 수집 기간, 중점 수집 기간, 균형성 담보를 위한 추가 수집 기간을 두어 국내 학습자, 이주민, 국외 학습자의 자료를 동시에 수집함으로써 일정 분량 이상의 자료를 균형성 있게 수집할 수 있다는 결론을 도출하였다.

##### (1) 이주민 자료 수집 사례 분석

- 수집 대상: 베트남 여성결혼이민자 7명
- 수집 대상 선정 방식: 현직 교사를 개인 접촉한 후 교사가 학습자 섭외
- 수집 자료 유형: 인터뷰 형식의 구어 발화 자료
- 수집 분량: 1시간의 인터뷰
- 수집의 문제점 및 고려 사항
  - 다문화센터와의 협약이 이루어지면 수집은 가장 용이할 것으로 예상되나 기관 차원의 보상을 직접적으로 요구하는 경우가 있었다.
  - 교사를 직접 접촉하여 개인적으로 섭외할 경우 대규모의 자료 수집이 어려우며 지속적인 자료 수집이 어려울 수 있다. 특히, 수집자 교육이나 관리를 고려하여 연구진의 인력풀을 활용할 경우 그 범위는 더욱 축소될 것으로 예상된다.
  - 지속적인 자료 수집을 위해서는 학습자 외에 가족의 동의가 필요할 수 있다. 특히, 아동 학습자의 경우 부모의 동의가 필요하며, 소속 기관을 접촉할 경우 담임교사, 학교의 동의가 함께 이루어져야 한다.

- 기관이나 학습자 개인의 의지와 무관하게 육아나 가정환경, 가족 등의 영향을 많이 받아서 섭외, 지속적인 자료 수집 등이 어려울 수 있다.
- 정규 교육을 받지 않은 경우가 많기 때문에 학습자의 숙달도 단계를 측정할 수 있는 기준이 없다.
- 정교 교육을 받지 않기 때문에 교육과정에 종속되지 않은 상태에서의 언어 습득 특성을 살필 수 있는 자료를 수집하기에 용이하지만, 자연 습득이기 때문에 환경에 따라 발달 속도가 매우 상이할 수 있다. 체류 기간이나 가족 구성, 가족과의 관계에 따라 한국어의 구사 수준이 매우 달랐다.

## (2) 국외 자료 시험 수집

- 시험 수집 대상: 미국의 에모리 대학과 베트남 다낭 외대에서 한국어 강좌를 수강 중인 학생 중 참여 희망자 150여 명
- 시험 수집 대상 선정 방식: 교사가 교육과정 중 한국어 관련 과목 수강자의 동의를 구함
- 시험 수집 자료 유형: 작문 자료
- 시험 수집 분량: 시험 구축으로 계속 진행 중  
(다낭대의 경우 자율적으로 참여 의사를 밝힌 학생이 52명이었으나 실제 16명이 동의서를 제출하였으며, 총 18부의 작문이 수집됨)
- 시험 수집의 문제점 및 국외 자료 수집 시 고려 사항
  - 시수나 교육 내용 등 교육과정의 차이로 인한 KSL(국내 어학당)의 급 구분과 KFL의 급 구분 기준이 다르다. 미국의 경우 헤리티지 반의 경우 학생들의 수준에 따라 교육 내용이 유동적으로 변경되며, 한 학기 동안의 교수·학습 분량이 많다.
  - 수업 진행상 학습자 자료를 모으기가 어려운 수업이 있다. 드라마 수업, 한자어 수업, 경영 한국어 등은 정식 언어 수업에서 다소 벗어나 있으며, 말하기 수업의 경우 학습자들이 대화문을 외우고 매우 제한된 상황에서 많은 단서가 주어진 상태에서 말하기 발표를 수행하도록 되어 있어 학습자 말뭉치 자료로 적합하지 않다.
  - 교육과정에 따라 문법 중심 또는 회화 위주의 수업이 이루어지며, 자료 수집을 위해 별도의 과제를 부과하기 어렵다.
  - 국내 어학당 기준으로 3급 후반부 내지는 4급 수준의 학습자가 많지 않아 자료를 수집하기 어렵다.

- 자료 수집 및 처리에서 자료의 녹음, 스캔, 전송 등의 어려움이 있다.

### (3) 자료 구축에 관한 설문조사

#### ○ 설문조사의 목적

- 설문 조사를 통해 이주민과 국외 학습자 자료 수집과 구축 시 고려해야 할 사항들을 조사해 보고 이를 계획안에 반영한다. 아울러 학습자 말뭉치 자료 활용에 관한 의견 수렴을 통해 말뭉치 보급 시스템 설계 방향 설정에 참고한다.

#### ○ 설문 조사 대상

- 국내 한국어 교육기관 교사
- 이주민 대상의 교사
- 국외 학습자 대상의 교수자

#### ○ 설문 조사 문항

1. 교육 현장에서 수집 가능한 학습자 대상과 언어 자료에는 무엇이 있는지 자유롭게 써 주십시오.  
(예) [1급] - 작문  
[2급] - 인터뷰 자료/ 자유 회화 녹음 자료
2. 각 교육기관별로 자료를 수집할 수 있는 적절한 시기가 있다면 알려주십시오.  
(예) 수시 가능 / 연중 6월, 7월 적절 등
3. 말하기나 글쓰기 자료를 제공하는 학생이나 그 자료를 수집하는 교사에게 어느 정도의 보상(금전 혹은 물품)이 필요하다고 생각하십니까?
4. 학생들에게서 자료를 수집할 때 특히 고려할 점이 있으면 기술해 주십시오.
5. 학습자(아동 학습자의 경우 학습자 부모)에게 언어 자료 활용 동의서를 받는 것에 어려움을 없을까요?  
※ [동의서]는 학습자의 언어로 번역된 것이 제공되며, 학습자의 이름은 수집되지 않습니다.
6. 국외인 경우 교육 현장에서 교포 학습자와 외국인 학습자의 비율은 어떻게 됩니까?

## 라. 설문조사 결과

### ㄱ. 응답자 기초 정보

- 한국어 교육 분야 종사 유형: 한국어교육 관련 연구자(대학원생 포함) 4명(8.90%), 한국어 교수자 19명(42.20%), 한국어교육 연구자이면서 교수자 19명(42.20%), 기타 3명(6.70%)
- 한국어 교수 경력: 1년 미만 2명(4.20%), 1년 이상 3년 미만 8명(16.70%), 3년 이상 5년 미만 6명(12.50%), 5년 이상 27명(56.30%), 해당 사항 없음 5명(10.4%)
- 대학 부설 교육기관 유형: 대학 부설 교육기관 32명(71.10%), 사설 교육기관 2명(4.40%), 대학원 11명(24.40%), 다문화센터, 국외 대학 등 기타 7명(15.60%)
- 학습자 말뭉치 사용 경험 여부: 있다 14명(29.20%), 없다 34명(70.80%)

### ㄴ. 학습자 말뭉치 구축에 관한 의견

1. 교육 현장에서 수집 가능한 학습자 대상과 언어 자료에는 무엇이 있는지 자유롭게 써 주십시오.

- [어학연수생] 시험 작문, 수업 작문 및 과제 작문자료, 발표 자료, 인터뷰 자료, 중간/기말고사 작문, 일기
- [유치원생] 발표 자료
- [교환 학생] 발표자료, 작문자료
- [학부생] 발표자료, 작문 자료
- [결혼 이주 여성] 발표 자료, 작문 자료
- [고려인/조선족 자녀] 발표 자료, 작문 자료
- [다문화 초등학생] 작문 자료. 일기, 독서 감상문

2. 각 교육기관별로 자료를 수집할 수 있는 적절한 시기가 있다면 알려 주십시오.

- [학부생, 교환 학생] 학기 중, 3월-6월, 9월-12월
- [어학연수생] 봄(3-5월), 여름(6월-7월), 가을(9월-11월), 겨울 학기(11월-1월) (중간고사, 기말고사 후)

- [초등학생] 독서 대회 기간, 방학 과제, 학년말 12월, 수시 평가

3. 말하기나 글쓰기 자료를 제공하는 학생이나 그 자료를 수집하는 교사에게 어느 정도의 보상(금전 혹은 물품)이 필요하다고 생각하십니까?

- 거의 모두가 작은 선물, 기념품, 상품권 등의 보상책을 희망하였으며, 교사뿐 아니라 학생에게도 작은 보상이 주어져야 한다는 의견도 있었다.

4. 학생들에게서 자료를 수집할 때 특히 고려할 점이 있으면 기술해 주십시오.

- 답을 한 31명 중 21명이 자료 수집의 절차적 문제(학습자, 교수자, 학교, 학부모 등의 동의)와 자료 수집의 목적에 대한 보다 쉽고 명확한 설명이 필요하다고 응답하였다. 그 외에 4명은 학생들의 특성(심리적)이나 수준 등 학습자 측면에 대한 고려가 필요하다고 답변하였으며, 또 다른 4명은 자료 수집의 체계성, 명확한 수집 지침에 대하여 언급하였다.

5. 학습자(아동 학습자의 경우 학습자 부모)에게 언어 자료 활용 동의서를 받는 것에 어려움을 없을까요?

- 큰 어려움이 없을 것이라는 의견과 어려움이 있을 것이라는 의견이 반반 정도로 나뉘었으며, 어려움이 없는 경우도 목적에 대한 상세한 설명, 부모의 동의, 개인 정보의 확실한 폐기 등을 전제로 하였다. 어려움이 있을 것이라는 의견에서는 당사자에게 이해를 구하기 힘들 것이라는 의견과 교사 입장에서의 부담, 학부모의 거부감 등을 들었다.
- [다문화 가정 아동] 학습자가 많지만 동의서 받기 힘든 경우가 있습니다. / 학부모의 경우 생계를 위한 활동을 하고 있기 때문에 한국어에 대한 이해가 부족하여 더 어렵다. / 초등학교와 학부모는 어느 정도 거부감이 있어서 상황 설명이 충분히 있어야 할 겁니다.
- [대학 부설 어학원] 개설되는 과목들이 많고 수업 시간도 제각각이라서 모든 과목에서 자료를 수집하는 데에는 한계가 있습니다. 각 과목 선생님들에게 동의서 받는 것을 요청하고 있지만 현장 현실상 다 이뤄지지 않을

때도 많고 매번 동의서를 받는 데에는 현실적으로 어려움이 있는 것은 사실입니다. / 개인 정보의 공개를 원하지 않는 학생들의 경우 동의서 사인을 거절하는 경우가 있습니다. 동아시아권 이외 지역의 학습자들에게서 이런 경향이 두드러집니다. / 교사의 권위에 기대어 원하지 않는데도 불구하고 찬성할 수 있는 가능성이 있다.

- [이주민] 경우에 따라 결혼이주민여성이나 고려인/조선족 자녀들의 경우 원하지 않는 경우도 있을 듯하다."
- [기타] 일일이 학습자들에게 동의서를 받아 취합하는 일이 번거롭다. / 학습자 말뭉치에 대한 이해 부족으로 동의서를 받기가 어렵다는 답변이 있었다. / 그 외에도 응답자가 학습자 말뭉치에 대한 이해가 부족해 답변을 하기 어렵다고 응답한 경우도 있었다.

6. 국외인 경우 교육 현장에서 교포 학습자와 외국인 학습자의 비율은 어떻습니까?

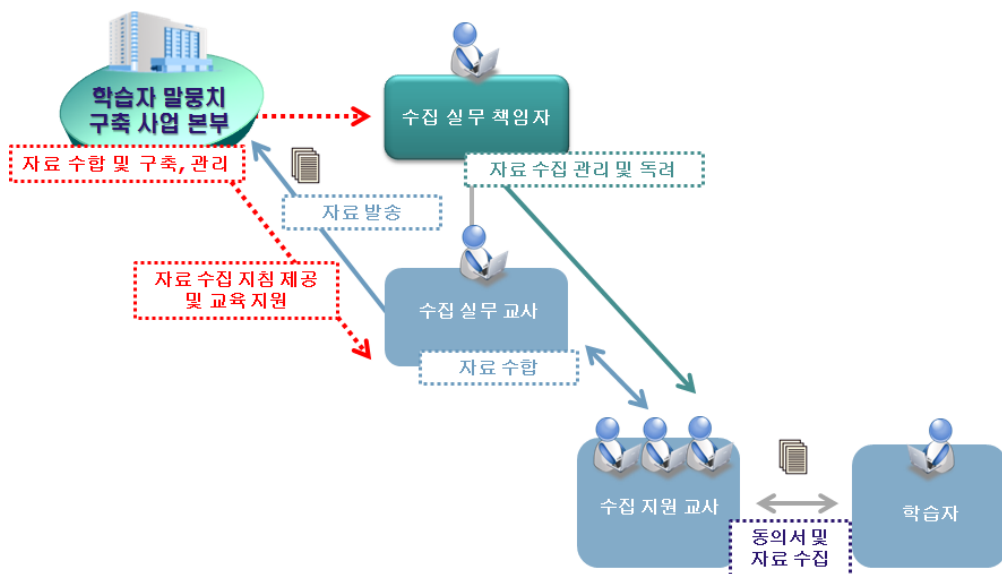
- 유의미한 응답이 없었다.

## 2. 사업 조직의 구성 및 운영

### 2.1. 기본 계획

#### ① 자료 수집을 위한 조직 운영 체계

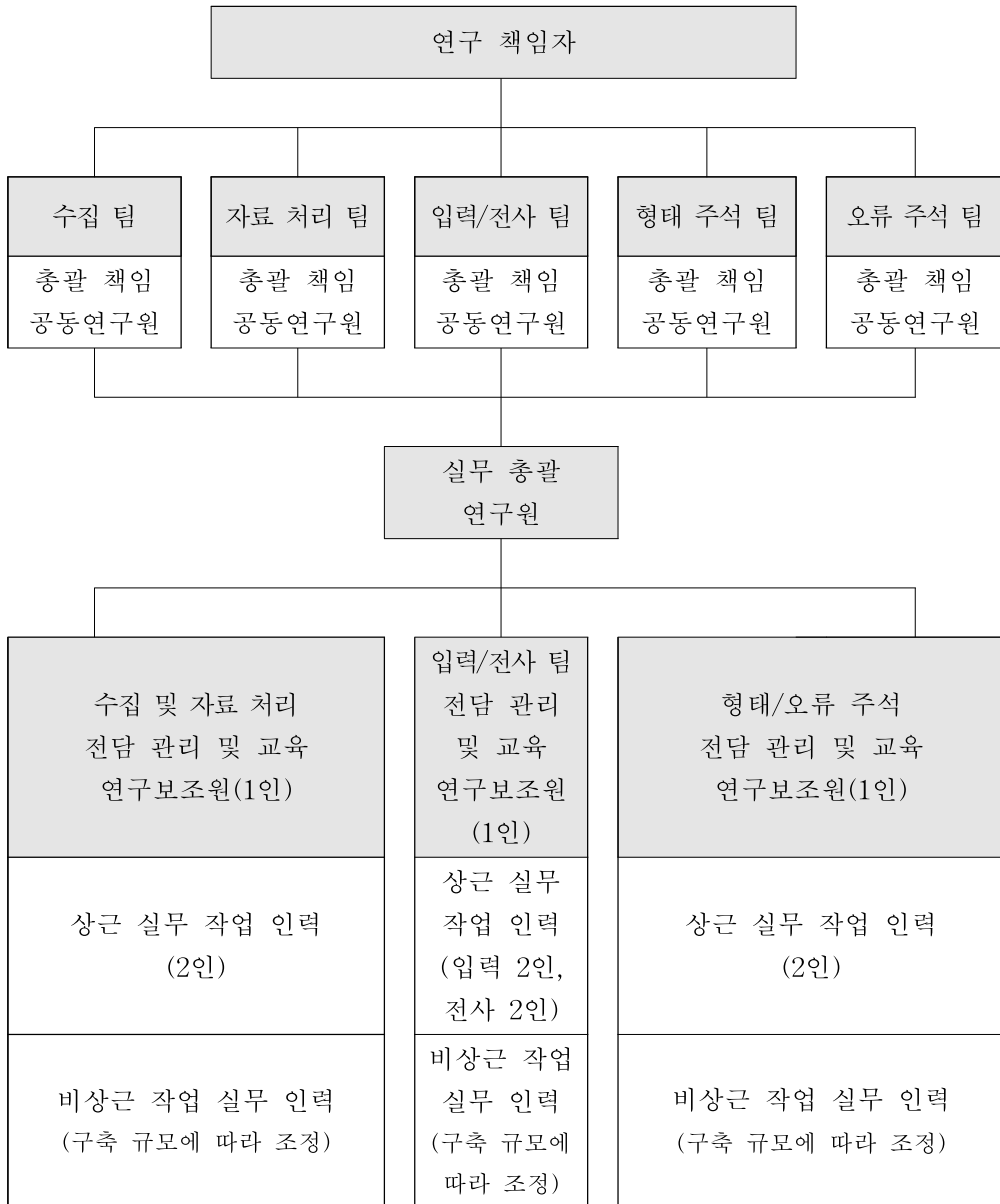
- 각 수집 기관에서 자료를 수집한 후에 구축 본부에 직접 발송하며, 자료 수집과 관리, 구축에 관한 제반 업무를 구축 본부 중심으로 수행한다.



<그림 1> 자료 수집 조직 체계

#### ② 자료 구축을 위한 조직 운영 체계

- 구축 본부를 중심으로 작업을 진행하되, 각 단계별 특성과 구축 규모에 따라 작업 인력을 배치한다. (인력 배치 예시)



<그림 2> 자료 구축 조직 체계

## 2.2. 주요 쟁점과 계획 수립의 기본 방향

### 1) 자료 수집 조직 체계

- 2015년 자료 수집 초기에 7개의 지역별 거점 기관을 두어 수집과 자료 처리(분류, 스캔, 파일명 부여, 파일 등록), 1차 입력/전사까지 전담하는 것을 기본 계획으로 하였으나, 기관에 배당되는 예산 규모의 적정성과 형평성에 관한 수집 기관의 문제 제기가 있었다. 또한 작업 인력 확보, 전문성, 지침 공유와 교육, 관리의 측면에서 실무적으로 어려움이 있었다. 이러한 문제는 각 기관에 배당할 수 있는 충분한 예산의 확보를 통해서 해결 가능하다. 그러나 한정된 예산 내에서 현실적으로 해결할 수 있는 방안은 업무가 집중되는 문제가 있기는 하지만 구축 본부를 중심으로 자료의 수집이 이루어지는 것이라고 본다.

### 2) 자료 구축 조직 체계

- 자료 수집 이후의 자료 처리와 입력, 전사, 가공의 전 단계에서 많은 구축 단계별로 훈련된 전문 인력 확보, 시간과 비용에 대한 고려가 절대적으로 필요하다. 특히, 구축 본부를 중심으로 수집과 구축의 전 과정이 이루어지면서 수집된 자료의 분류, 스캔, 전산화하는 작업에서 많은 인력과 시간, 비용이 발생하였다. 그 외의 구축 단계에서는 학부생과 대학원생을 실무 구축 작업에 대거 투입하여 인력 확보와 비용 절감을 하고자 하였는데, 시험 기간 등 작업의 연속성, 일정한 수준의 작업 결과물 산출을 위한 부차적인 검수 작업과 관리의 어려움이 있었다.

이러한 문제를 해결하기 위하여 향후의 사업에서는 작업 단계별 전문성, 업무의 연속성 여부에 따라 참여 인력의 학위 수준이나 참여 형태(상근/비상근, 전담)를 달리하는 방안을 도입할 필요가 있다.

### 3. 총 예산 측정

#### 3.1. 기본 계획

##### 1) 사업비 총예산

○ 3단계 6년간의 사업비 총 예산은 다음과 같다.

<표 8> 한국어 학습자 말뭉치 총 예산

총 예산	1,989,042,815원
1차 연도	176,000,000원
2차 연도	309,605,013원
3차 연도	423,170,388원
4차 연도	473,412,888원
5차 연도	391,119,138원
6차 연도	215,735,388원

○ 학습자 말뭉치 전체 수집 규모에 따른 총 예산의 규모는 다음과 같다.

<표 9> 한국어 학습자 말뭉치 총 예산과 수집 규모

예산액	자료 유형	자료 수집	원시 말뭉치	형태 주석 말뭉치	오류 주석 말뭉치
1,989,042,815원	문어	300만+a	300만	300만	120만
	구어	70만+a	70만	70만	40만

○ 예산 산출 내역은 다음과 같다.

<표 10> 한국어 학습자 말뭉치 구축 사업비 총예산

구분		총액	산출 내역	비고	
기초 연구 및 교육 관리비	사업 책임 총괄 인건비	36,696,336	764,507원(3,058,029원×0.25)×1 명×8개월×6년=36,696,336원	3단계 6년 사업, 2015년 사업 계 획에 따라 1년 사업 기간을 8 개월로 산정함	
	영역별 책임 총괄비인건 비	157,573,920	468,970원(2,344,854원×0.20)×7 명×8개월×6년=157,573,920원		
	실무 책임 총괄 인건비	75,237,936	1,567,457원×1명×8개월×6년= 75,237,936원		
	구축 단계별 전담 관리 및 교육 인건비	157,999,536	1,097,219원×3명(수집, 구축, 가공)×8개월×6년= 157,999,536원		
소계		427,507,728			
말 뭉 치 구 축 비	수집 1)	문어	300,000,000	100원×3백만 어절=300,000,000원	
		구어	210,000,000	300원×70만 어절=210,000,000원	
	자료 처리	스캔 및 파일 변환	15,875,000	500원×31,750개=15,875,000원	
		파일 정보 등록	15,875,000	500원×31,750개=15,875,000원	
	입력 /전사	문어	30,000,000	10원×3백만 어절=30,000,000원	
		구어	21,000,000	30원×70만 어절=21,000,000원	
	1차 검수	문어	21,000,000	7원×3백만 어절=21,000,000원	구축 비용의 70%
		구어	14,000,000	20원×70만 어절=14,000,000원	
	2차 검수	문어	9,000,000	3원×3백만 어절=9,000,000원	구축 비용의 30%
		구어	7,000,000	10원×70만 어절=7,000,000원	
	형태	문어	90,000,000	30원×3백만 어절=90,000,000원	

구분			총액	산출 내역	비고
	주식	구어	21,000,000	30원×70만 어절=21,000,000원	구축 비용의 70%
	1차 검수	문어	60,000,000	20원×3백만 어절=60,000,000원	
		구어	14,000,000	20원×70만 어절=14,000,000원	
	2차 검수	문어	30,000,000	10원×3백만 어절=30,000,000원	구축 비용의 30%
		구어	7,000,000	10원×70만 어절=7,000,000원	
	오류 주식	문어	60,000,000	50원×1 백20만 어절=60,000,000원	
		구어	20,000,000	50원×40만 어절=20,000,000원	
	1차 검수	문어	42,000,000	35원×1 백20만 어절=42,000,000원	구축 비용의 70%
		구어	14,000,000	35원×40만 어절=14,000,000원	
	2차 검수	문어	18,000,000	15원×1 백20만 어절=18,000,000원	구축 비용의 30%
		구어	6,000,000	15원×40만 어절=6,000,000원	
소계			1,025,750,000		
기구축 말뭉치 수정, 보완비	문어	75,000,000	30원×2백50만 어절=75,000,000원	지침 업데이트 등에 의한 기구축 자료의 정비 비용으로 입력 및 전사, 형태 주식, 오류 주식 작업 비용 20% 내외 책정 (마지막 연도 구축 분량 제외)	
	구어	30,000,000	50원×60만 어절=30,000,000원		
소계			105,000,000		
홍보 및 교육비			50,000,000	학술 워크숍 개최비 1,000만원×1회×5년= 50,000,000원	
사업 부 대	소모품비		12,000,000	2015 예산 계획에 준함	
	유인물비		15,780,000	2015 예산 계획에 준함	
	연구 용역 재료비		12,000,000	2015 예산 계획에 준함	

구분		총액	산출 내역	비고
비용	회의비	40,000,000	2015 예산 계획에 준함	
	자문비	18,000,000	2015 예산 계획에 준함	
	교통통신비	10,077,264	2015 예산 계획에 준함	
	여비	6,000,000	2015 예산 계획에 준함	
소계		158,780,000		
일 반 관 리 비 (5%)		86,105,749		
부가세 (10%)		180,822,074		
총액		1,989,042,815		

## 2) 연차별 예산

### (1) 1차 연도 예산

☞ 1차 연도는 종료된 사업으로 예산 계획에서 제외함

### (2) 2차 연도 예산

○ 한국어 학습자 말뭉치 전체 수집 규모에 따른 2차 연도 예산의 규모는 다음과 같다.

- 1) 수집 대상에 따라 네트워크의 규모(예. 국내의 경우 교사 1인당 평균 10명 내외의 학습자를 담당하게 되어 교사 1,000명, 학습자 8,000여 명이 수집에 관여함), 수집 방법(예. 이주민 자료 수집의 경우 방문 수집이 필요할 수 있으며, 수집 대상자에게 그에 상응하는 경제적 보상을 고려해야 함) 등이 달라질 수 있다. 그러나 이는 향후 실질적인 수집 인력이 (가)확정되고 그에 따라 구체적인 수집 방법이 정해진 후에 확정 가능하다. 따라서 향후 수집 규모와 수집 네트워크, 수집 방법이 확정된 후 연차별 세부 예산을 편성할 수 있도록 1차 연도의 수집 결과를 토대로 어절당 평균 수집 비용을 산출하였다.

<표 11> 한국어 학습자 말뭉치 2차 연도 예산과 수집 규모

예산액	자료 유형	자료 수집	원시 말뭉치	형태 주석 말뭉치	오류 주석 말뭉치
309,605,013원	문어	50만	50만	30만	11만
	구어	15만	15만	8만	4만

○ 예산 산출 내역은 다음과 같다.

<표 12> 한국어 학습자 말뭉치 구축 사업비 2차 연도 예산

구분			총액	산출 내역
기초 연구 및 교육 관리비	사업 책임 총괄 인건비		6,116,056	764,507원 (3,058,029원×0.25)×1 명×8개월=6,116,056원
	영역별 책임 총괄비인건비		26,262,320	468,970원 (2,344,854원×0.20)×7 명×8개월=34,141,072원
	실무 책임 총괄 인건비		12,539,656	1,567,457원×1명×8개월=12,539,6 56원
	구축 단계별 전담 관리 및 교육 인건비		26,333,256	1,097,219원×3명(수집, 구축, 가공)×8개월=52,666,512원
소계			71,251,288	
말 뭉 치 구 축 비	수집	문어	50,000,000	100원×50만 어절=50,000,000원
		구어	45,000,000	300원×15만 어절=45,000,000원
	자료 처리	스캔 및 파일 변환	2,687,500	500원×5,375개=2,687,000원
		파일 정보 등록	2,687,500	500원×5,375개=2,687,000원
	입력 /전사	문어	5,000,000	10원×50만 어절=5,000,000원
		구어	4,500,000	30원×15만 어절=4,500,000원
	1차 검수	문어	3,500,000	7원×50만 어절=3,500,000원
		구어	3,000,000	20원×15만 어절=3,000,000원
	2차 검수	문어	1,500,000	3원×50만 어절=1,500,000원

구분			총액	산출 내역
		구어	1,500,000	10원×15만 어절=1,500,000원
	형태 주석	문어	9,000,000	30원×30만 어절=9,000,000원
		구어	2,400,000	30원×8만 어절=2,400,000원
	1차 검수	문어	6,000,000	20원×30만 어절=6,000,000원
		구어	1,600,000	20원×8만 어절=1,600,000원
	2차 검수	문어	3,000,000	10원×30만 어절=3,000,000원
		구어	800,000	10원×8만 어절=800,000원
	오류 주석	문어	5,500,000	50원×11만 어절=5,500,000원
		구어	2,000,000	50원×4만 어절=2,000,000원
	1차 검수	문어	3,850,000	35원×11만 어절=3,850,000원
		구어	1,400,000	35원×4만 어절=1,400,000원
	2차 검수	문어	1,650,000	15원×11만 어절=1,650,000원
		구어	600,000	15원×4만 어절=600,000원
소계			157,175,000	
기구축 말뭉치 수정, 보완비		문어	9,000,000	30원×30만 어절=9,000,000원
		구어	2,500,000	50원×5만 어절=2,500,000원
소계			11,500,000	
홍보 및 교육비			10,000,000	학술 워크숍 개최비 10,000,000원 (장소대여료,강사료,자료및홍보 물인쇄비,다과준비비등)
사업 부 대 비 용	소모품비		2,000,000	2015 예산 계획에 준함
	유인물비		2,630,000	2015 예산 계획에 준함
	재료비		2,000,000	2015 예산 계획에 준함
	회의비		6,000,000	2015 예산 계획에 준함
	자문비		3,000,000	2015 예산 계획에 준함

구분		총액	산출 내역
	교통통신비	1,500,000	2015 예산 계획에 준함
	여비	1,000,000	2015 예산 계획에 준함
소계		28,130,000	
일 반 관 리 비 (5%)		13,402,814	
부가세 (10%)		28,145,910	
총액		309,605,013	

### (3) 3차 연도 예산

- 한국어 학습자 말뭉치 전체 수집 규모에 따른 3차 연도 예산의 규모는 다음과 같다.

<표 13> 한국어 학습자 말뭉치 3차 연도 예산과 수집 규모

예산액	자료 유형	자료 수집	원시 말뭉치	형태 주식 말뭉치	오류 주식 말뭉치
423,170,388원	문어	80만	80만	40만	25만
	구어	20만	20만	15만	10만

- 예산 산출 내역은 다음과 같다.

<표 14> 한국어 학습자 말뭉치 구축 사업비 3차 연도 예산

구분			총액	산출 내역
기초 연구 및 교육 관리비 (고정 인건비)			71,251,288	2차 연도와 동일
말 뭉 치 구 축 비	수 집	문어	80,000,000	100원×80만 어절=80,000,000원
		구어	60,000,000	300원×20만 어절=60,000,000원
	자료 처리	스캔 및 파일 변환	4,250,000	500원×8,500개=4,250,500원
		파일 정보 등록	4,250,000	500원×8,500개=4,250,500원
	입력 /전사	문어	8,000,000	10원×80만 어절=8,000,000원
		구어	6,000,000	30원×20만 어절=6,000,000원

구분			총액	산출 내역
	1차 검수	문어	5,600,000	7원×80만 어 절=5,600,000원
		구어	4,000,000	20원×20만 어 절=4,000,000원
	2차 검수	문어	2,400,000	3원×80만 어 절=2,400,000원
		구어	2,000,000	10원×20만 어 절=2,000,000원
	형태 주식	문어	12,000,000	30원×40만 어 절=12,000,000원
		구어	4,500,000	30원×15만 어 절=4,500,000원
	1차 검수	문어	8,000,000	20원×40만 어 절=8,000,000원
		구어	3,000,000	20원×15만 어 절=3,000,000원
	2차 검수	문어	4,000,000	10원×40만 어 절=4,000,000원
		구어	1,500,000	10원×15만 어 절=1,500,000원
	오류 주식	문어	12,500,000	50원×25만 어 절=12,500,000원
		구어	5,000,000	50원×10만 어 절=5,000,000원
	1차 검수	문어	8,750,000	35원×25만 어 절=8,750,000원
		구어	3,500,000	35원×10만 어 절=3,500,000원
	2차 검수	문어	3,750,000	15원×25만 어 절=3,750,000원
		구어	1,500,000	15원×10만 어 절=1,500,000원
소계			244,500,000	
기구축 말뭉치 수정, 보완비		문어	15,000,000	30원×50만 어 절=15,000,000원
		구어	7,500,000	50원×15만 어 절=7,500,000원
소계			22,500,000	
홍보 교육비 및 사업 부대 비용			28,130,000	2차 연도와 동일
일 반 관 리 비 (5%)			18,319,064	
부가세 (10%)			38,470,035	
총액			423,170,388	

#### (4) 4차 연도 예산

- 한국어 학습자 말뭉치 전체 수집 규모에 따른 4차 연도 예산의 규모는 다음과 같다.

<표 15> 한국어 학습자 말뭉치 4차 연도 예산과 수집 규모

예산액	자료 유형	자료 수집	원시 말뭉치	형태 주석 말뭉치	오류 주석 말뭉치
473,412,888원	문어	90만	90만	50만	30만
	구어	20만	20만	20만	15만

- 예산 산출 내역은 다음과 같다.

<표 16> 한국어 학습자 말뭉치 구축 사업비 4차 연도 예산

구분			총액	산출 내역
기초 연구 및 교육 관리비 (고정 인건비)			71,251,288	2차 연도와 동일
말 뭉 치 구 축 비	수집	문어	90,000,000	100원×90만 어절=90,000,000원
		구어	60,000,000	300원×20만 어절=60,000,000원
	자료 처리	스캔 및 파일 변환	4,750,000	500원×9,500개=4,750,000원
		파일 정보 등록	4,750,000	500원×9,500개=4,750,000원
	입력 /전사	문어	9,000,000	10원×90만 어절=9,000,000원
		구어	6,000,000	30원×20만 어절=600,000원
	1차 검수	문어	6,300,000	7원×90만 어절=6,300,000원
		구어	4,000,000	20원×20만 어절=4,000,000원
	2차 검수	문어	2,700,000	3원×90만 어절=2,700,000원
		구어	2,000,000	10원×20만 어절=2,000,000원
	형태 주석	문어	15,000,000	30원×50만 어절=15,000,000원
		구어	6,000,000	30원×20만 어절=6,000,000원
	1차	문어	10,000,000	20원×50만 어절=10,000,000원

구분			총액	산출 내역
	검수	구어	4,000,000	20원×20만 어절=4,000,000원
	2차 검수	문어	5,000,000	10원×50만 어절=5,000,000원
		구어	2,000,000	10원×20만 어절=2,000,000원
	오류 주석	문어	15,000,000	50원×30만 어절=15,000,000원
		구어	7,500,000	50원×15만 어절=7,500,000원
	1차 검수	문어	10,500,000	35원×30만 어절=10,500,000원
		구어	5,250,000	35원×15만 어절=5,250,000원
	2차 검수	문어	4,500,000	15원×30만 어절=4,500,000원
		구어	2,250,000	15원×15만 어절=2,250,000원
소계			276,500,000	
기구축 말뭉치 수정, 보완비		문어	24,000,000	30원×80만 어절=24,000,000원
		구어	10,000,000	50원×20만 어절=10,000,000원
소계			34,000,000	
홍보 교육비 및 사업 부대 비용			28,130,000	2차 연도와 동일
일 반 관 리 비 (5%)			20,494,064	
부가세 (10%)			43,037,535	
총액			473,412,888	

## (5) 5차 연도 예산

- 한국어 학습자 말뭉치 전체 수집 규모에 따른 5차 연도 예산의 규모는 다음과 같다.

<표 17> 한국어 학습자 말뭉치5차 연도 예산과 수집 규모

예산액	자료 유형	자료 수집	원시 말뭉치	형태 주석 말뭉치	오류 주석 말뭉치
391,119,138원	문어	50만	50만	90만	25만
	구어	10만	10만	20만	10만

○ 예산 산출 내역은 다음과 같다.

<표 18> 한국어 학습자 말뭉치 구축 사업비 5차 연도 예산

구분			총액	산출 내역
기초 연구 및 교육 관리비 (고정 인건비)			71,251,288	2차 연도와 동일
말 뭉 치 구 축 비	수집	문어	50,000,000	100원×50만 어절=50,000,000원
		구어	30,000,000	300원×10만 어절=30,000,000원
	자료 처리	스캔 및 파일 변환	2,625,000	500원×5,250개=2,625,000원
		파일 정보 등록	2,625,000	500원×5,250개=2,625,000원
	입력 /전사	문어	5,000,000	10원×50만 어절=5,000,000원
		구어	3,000,000	30원×10만 어절=3,000,000원
	1차 검수	문어	3,500,000	7원×50만 어절=3,500,000원
		구어	2,000,000	20원×10만 어절=2,000,000원
	2차 검수	문어	1,500,000	3원×50만 어절=1,500,000원
		구어	1,000,000	10원×10만 어절=1,000,000원
	형태 주석	문어	27,000,000	30원×90만 어절=27,000,000원
		구어	6,000,000	30원×20만 어절=6,000,000원
	1차 검수	문어	18,000,000	20원×90만 어절=18,000,000원
		구어	4,000,000	20원×20만 어절=4,000,000원
	2차 검수	문어	9,000,000	10원×90만 어절=9,000,000원
		구어	2,000,000	10원×20만 어절=2,000,000원

구분			총액	산출 내역
	오류 주석	문어	12,500,000	50원×25만 어절=12,500,000원
		구어	5,000,000	50원×10만 어절=5,000,000원
	1차 검수	문어	8,750,000	35원×25만 어절=8,750,000원
		구어	3,500,000	35원×10만 어절=3,500,000원
	2차 검수	문어	3,750,000	15원×25만 어절=3,750,000원
		구어	1,500,000	15원×10만 어절=1,500,000원
소계			202,250,000	
기구축 말뭉치 수정, 보완비		문어	27,000,000	30원×90만 어절=27,000,000원
		구어	10,000,000	50원×20만 어절=10,000,000원
소계			37,000,000	
홍보 교육비 및 사업 부대 비용			28,130,000	2차 연도와 동일
일 반 관 리 비 (5%)			16,931,564	
부가세 (10%)			35,556,285	
총액			391,119,138	

## (6) 6차 연도 예산

- 한국어 학습자 말뭉치 전체 수집 규모에 따른 6차 연도 예산의 규모는 다음과 같다.

<표 19> 한국어 학습자 말뭉치 6차 연도 예산과 수집 규모

예산액	자료 유형	자료 수집	원시 말뭉치	형태 주석 말뭉치	오류 주석 말뭉치
215,735,388원	문어	-	-	70만	25만
	구어	-	-	5만	-

○ 예산 산출 내역은 다음과 같다.

<표 20> 한국어 학습자 말뭉치 구축 사업비 6차 연도 예산

구분			총액	산출 내역
기초 연구 및 교육 관리비 (고정 인건비)			71,251,288	2차 연도와 동일
말 뭉 치 구 축 비	형태 주석	문어	21,000,000	30원×70만 어절=21,000,000원
		구어	1,500,000	30원×5만 어절=1,500,000원
	1차 검수	문어	14,000,000	20원×70만 어절=14,000,000원
		구어	1,000,000	20원×5만 어절=1,000,000원
	2차 검수	문어	7,000,000	10원×70만 어절=7,000,000원
		구어	500,000	10원×5만 어절=500,000원
	오류 주석	문어	12,500,000	50원×25만 어절=12,500,000원
	1차 검수	문어	8,750,000	35원×25만 어절=8,750,000원
	2차 검수	문어	3,750,000	15원×25만 어절=3,750,000원
	소계		70,000,000	
기구축 말뭉치 수정, 보완비		문어	15,000,000	30원×50만 어절=15,000,000원
		구어	5,000,000	50원×10만 어절=5,000,000원
소계			20,000,000	
홍보 교육비 및 사업 부대 비용			28,130,000	2차 연도와 동일
일 반 관 리 비 (5%)			9,469,064	
부가세 (10%)			19,885,035	
총액			218,735,388	

### 3) 모듈 단위 예산

○ 모듈 단위 예산은 연차별 구축 예산 조정에 활용할 수 있도록 문어 10만 어절, 구어 1만 어절 단위의 예산을 산출한 것이다. 한국어 학습자 말뭉치

예산은 다음과 같이 구성되며, 크게 기본 예산 항목과 변동 예산 항목으로 나뉜다.

인건비 + 자료 구축비/기구축 말뭉치의 수정, 보완 + 홍보 교육비 및 사업 부대 비용 + 일반관리비(5%)+부가세(10%)

- 기본 예산 항목: 사업 운영을 위한 필요한 항목. 인건비, 홍보 교육비 및 사업 부대 비용, 일반관리비, 부가세. 물가 상승에 따른 인건비 조정 등을 제외하고 변동이 없음
- 변동 예산 항목: 자료 구축에 소요되는 비용. 구축 규모에 따라 증감이 됨

○ 다음은 문어 10만 어절 기준의 구축 예산이다.

<표 21> 한국어 학습자 말뭉치 구축 사업비 모듈 단위 예산: 문어 10만 어절

구분			총액	산출 내역	비고
말 뭉 치 구 축 비	수 집		10,000,000	100원×10만 어절=10,000,000원	
	자료 처리	스캔 및 파일 변환	500,000	500원×1,000개=500,000원	
		파일 정보 등록	500,000	500원×1,000개=500,000원	
	입력		1,000,000	10원×10만 어절=1,000,000원	
	1차 검수		700,000	7원×10만 어절=700,000원	구축 비용의 70%
	2차 검수		300,000	3원×10만 어절=300,000원	구축 비용의 30%
	형태 주석		3,000,000	30원×10만 어절=3,000,000원	
	1차 검수		2,000,000	20원×10만 어절=2,000,000원	
	2차 검수		1,000,000	10원×10만 어절=1,000,000원	
	오류 주석		5,000,000	50원×10만 어절=5,000,000원	
	1차 검수		3,500,000	35원×10만 어절=3,500,000원	구축 비용의 70%
	2차 검수		1,500,000	15원×10만 어절=1,500,000원	구축 비용의

구분		총액	산출 내역	비고
				30%
소계		29,000,000		
기구축 말뭉치 수정, 보완비		3,000,000	30원×10만 어절=3,000,000원	
소계		3,000,000		
합계		32,000,000		

○ 다음은 구어 1만 어절 기준의 구축 예산이다.

<표 22> 한국어 학습자 말뭉치 구축 사업비 모듈 단위 예산: 구어 1만 어절

구분			총액	산출 내역	비고
말 뭉 치 구 축 비	수집		3,000,000	300원×1만 어절=3,000,000원	
	자료 처리	스캔 및 파일 변환	12,500	500원×25개=12,500원	
		파일 정보 등록	12,500	500원×25개=12,500원	
	입력 /전사		300,000	30원×1만 어절=300,000원	
	1차 검수		200,000	20원×1만 어절=200,000원	구축 비용의 70%
	2차 검수		100,000	10원×1만 어절=100,000원	구축 비용의 30%
	형태 주석		300,000	30원×1만 어절=300,000원	
	1차 검수		200,000	20원×1만 어절=200,000원	
	2차 검수		100,000	10원×1만 어절=100,000원	
	오류 주석		500,000	50원×1만 어절=500,000원	
	1차 검수		350,000	35원×1만 어절=350,000원	구축 비용의 70%
	2차 검수		150,000	15원×1만 어절=150,000원	구축 비용의 30%
소계			5,225,000		
기구축 말뭉치 수정,			500,000	50원×10만 어절=500,000원	

구분	총액	산출 내역	비고
보완비			
소계	500,000		
합계	5,725,000		

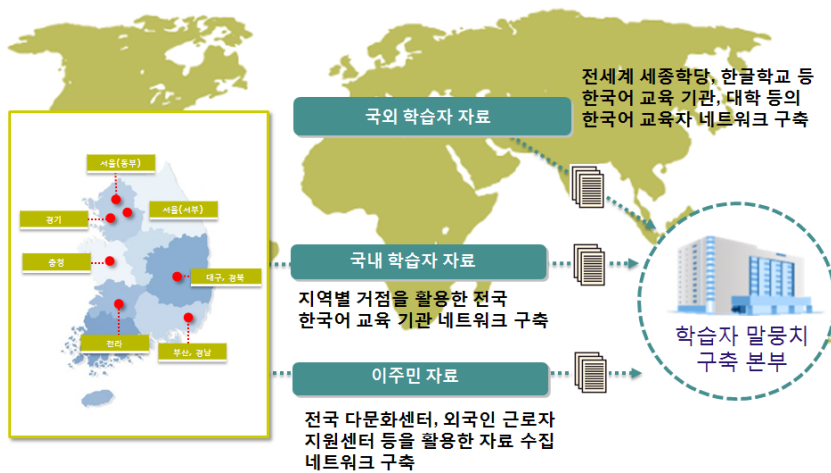
### 3.2. 주요 쟁점과 계획 수립의 기본 방향

- 2015년 학습자 말뭉치 구축 과정에서 예산 수립과 관련하여 다음의 사항에 대한 고려가 필요함을 파악하였다.
  - 자료 처리(분류, 스캔, 파일링) 및 관리 예산의 추가 배정
  - 수집 실무 책임자, 수집 지원 교사, 학습자 지급 비용 책정
  - 자료 수집 규모에 비례하여 수집 실무 교사의 비용을 차등 책정
  - 수집 기관에서 소요되는 수집 작업 비용 책정(예. 복사/스캔비)
- 본 연구에서는 위의 사항을 고려하여 자료 수집 - 자료 처리 - 입력 및 전사 - 형태 주석 - 오류 주석의 전 단계에서 적정 인력과 시간을 투입하고 적절한 보상을 할 수 있도록 단계별로 세분화하여 현실적인 예산을 편성하고자 하였다.
- 아울러 전체 예산과 연차별 예산, 문어 10만 어절/구어 1만 어절 기준의 모듈 단위 예산을 제시하여 사업 실행 계획에 맞게 예산을 효율적으로 산출할 수 있도록 하였다.

## 4. 자료 수집의 방향

### 4.1. 자료 수집 방법

#### 1) 구축 본부-수집 기관의 직접 접촉을 통한 수집



<그림 3> 지역별 거점 기관을 통한 오프라인 방식의 자료 수집

○ 국내, 이주민, 국외 기관의 수집 네트워크를 구축한 후 구축 본부와 수집 실무자와의 직접 접촉을 통해 자료를 수집한다. 우편 및 택배, 전자우편 등을 이용하여 수집 자료를 주고받게 된다.

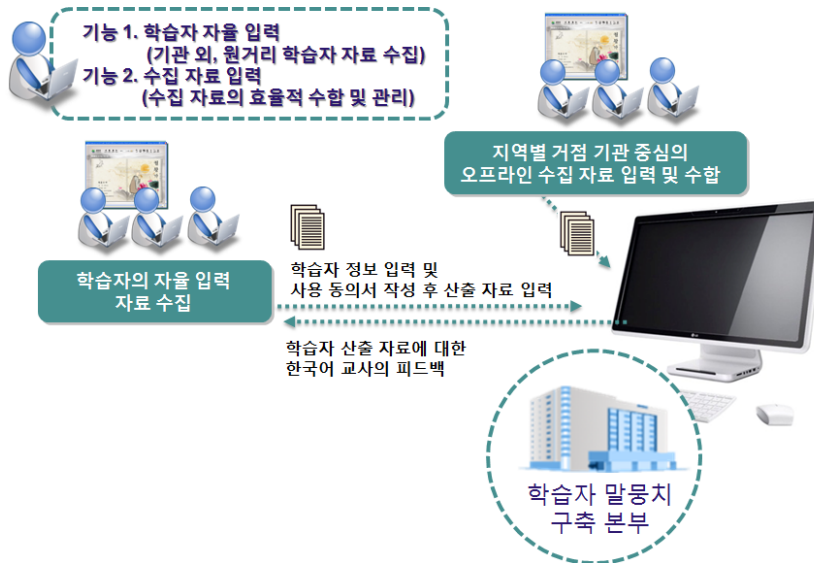
○ 집단별 수집 경로는 다음과 같다.

<표 23> 한국어 학습자 말뭉치의 수집 경로

구분	수집 대상	자료 수집 경로	수집 전략
국내학	어학 연수생	국내 교육기관	구축 본부
	대학(원) 유학생	국내 대학(원)	중심의 기관 관리 및 공조

습 자	기관 외 외국인 (*해당 기관의 협조와 수집 학습자의 자발적 의사가 필요함)	* 온라인 수집 시스템 * 한국어능력시험 자료	체제
	이 주 민	다문화가정 지원센터 217개소의 방문교사, 개인 교수자	공문을 통한 협조 요청, 관계 기관 대표 및 실무자 대상의 설명회 개최
국 외 학 습 자	세종학당	세종학당	한국어교육자 협의회 등의 네트워크 활용
	국외 정규 교육기관의 한국학 전공 및 한국어 학습자	국외 대학 국외 한국어교육자 네트워크	

## 2) 온라인 자료 구축 시스템을 활용한 자율적 수집[제안]



<그림 4> 온라인 자료 수집 시스템을 활용한 수집 체계도

- 온라인 자료 수집 시스템을 활용하여 자료 구축의 효율성을 제고하며, 교육 기관 외의 학습자 산출 자료를 추가적으로 수집하는 데에도 활용할 수 있다. 즉, 교육 기관에서 수집한 말뭉치를 수집 교사가 직접 입력하여 업로드하게 될 경우 자료 관리 및 구축 비용을 절감할 수 있으며, 학습자의 산출물에 대한 교사의 피드백, 광범위한 홍보 활동 등을 통해 기관 외의 학습자가 산출 자료를 자율적으로 올릴 수 있도록 유도하여 보조적인 자료 수단으로 활용할 수 있다. 아울러 시간이 흐름에 따라 학습자의 구성이나 자료의 특성도 달라지므로 사업 기간 내에는 물론 사업이 종료된 후 지속적인 자료를 수집할 수 있다는 장점이 있다.

## 4.2. 수집 범위

- 한국어 학습자 말뭉치는 수집 대상 기관과 과제 설계의 방법, 수집 기간에 따라 교육기관 말뭉치와 기획 말뭉치, 횡적 말뭉치와 종적 말뭉치를 수집 범위로 정한다.

### 1) 수집 기관과 수집 방법, 수집 기간에 따른 말뭉치의 종류

- 말뭉치는 수집 기관과 수집 방법에 따라 교육기관 말뭉치, 기획 말뭉치, 자연 발화 말뭉치로 나뉜다.
  - **교육기관 말뭉치:** 기관의 정규 교육과정 내에서 이루어지는 다양한 활동을 통해 산출되는 자료를 수집하여 구축하는 말뭉치
  - **기획 말뭉치:** 교육 기관 종속도를 낮추기 위한 것으로 텍스트의 장르와 주제를 계획적으로 통제하여 수집하여 구축하는 말뭉치
  - **자연 발화 말뭉치:** 교육기관 말뭉치나 기획 말뭉치와 달리 인위적으로 계획되지 않은 실제 발화 자료를 수집하여 구축하는 말뭉치
- 말뭉치는 수집 기간에 따라 횡적 말뭉치와 종적 말뭉치로도 나뉜다.
  - **횡적 말뭉치:** 수집 기간 내에 산출된 자료를 일괄 수집하여 구축하는 말뭉치로 수준별로 수집된 대규모 자료로부터 학습자의 중간언어를 관찰할 수 있음

- **종적 말뭉치:** 초급에서 고급까지의 과정에서 산출한 자료를 단계별로 모두 수집한 것으로 언어 발달과 습득의 순서를 관찰할 수 있음

○ 한국어 학습자 말뭉치는 다음과 같은 방법에 의해 수집된다.

<표 24> 수집 대상 기관과 수집 방법에 따른 학습자 말뭉치

구분	수집 방법 및 자료 유형	수집 기간에 따른 말뭉치 유형
교육기관 말뭉치	정규과정의 중간 및 기말 쓰기/말하기 시험 자료	횡적 말뭉치
	교과 과정에 포함된 다양한 주제 및 다양한 텍스트 형식의 수업/과제 작문	횡적 말뭉치
	대학, 대학원 재학생의 보고서, 논문/구두 발표 자료	횡적 말뭉치
	대학, 대학원생의 토론 자료	횡적 말뭉치
기획 말뭉치	텍스트의 장르별 기획 말뭉치(논설문, 생활문, 감상문; 대화, 발표 등)	횡적 말뭉치
	백일장, 말하기 대회 자료	횡적 말뭉치
	학습자 언어 습득 고찰을 위한 종적 말뭉치	종적 말뭉치
	교사와 학습자 상호 작용 연구를 위한 교실/수업 자료 말뭉치(*협조 필요)	횡적 말뭉치
	정규 과정의 졸업 좌담회	횡적 말뭉치
	한국어능력시험 작문 자료	횡적 말뭉치
자연 발화 말뭉치	(정기적/비정기적 수집) 일상대화	종적/횡적 말뭉치
	(정기적/비정기적 수집) 계획되지 않은 인터뷰	종적/횡적 말뭉치
	(일정 기간 (정기적/비정기적 수집) 내레이션	종적/횡적 말뭉치

☞ 토론은 다자 발화이기 때문에 음성 녹음 자료만으로는 발화자 구분이 어

렵고 그로 인해 전사를 정확하게 할 수 없다. 그런 이유로 1차 연도에는 토론 자료를 수집하기는 하였으나 시험 구축을 위한 일부 자료 외에는 구축 대상에서 제외하였다. 향후 문제점을 보완하여 토론 자료를 추가 수집하는 방안에 대해서도 고려해 볼 수 있다.

## 2) 대상별 자료 수집 범위

- 국내 학습자, 이주민, 국외 학습자는 학습 환경, 한국어에 노출되는 시간이나 사용 시간 등에 의해 중간언어의 양상, 한국어를 습득하는 순서나 과정에서 차이를 보일 수 있다. 본 연구에서는 수집 대상별 특성이 잘 나타나는 자료들을 집중적으로 수집하고자 한다. 각 자료를 수집하기 위한 세부 지침은 수집 네트워크와 수집 대상이 어느 정도 정해진 후에 그에 맞게 정하기로 한다.

<표 25> 수집 대상 기관과 수집 방법에 따른 학습자 말뭉치

구분	수집 범위 및 자료 유형	국내 학습 자	이주민	국외 학습자
교육기관 말뭉치	정규과정의 중간 및 기말 쓰기/말하기 시험 자료	○		
	교과 과정에 포함된 다양한 주제 및 다양한 텍스트 형식의 수업/과제 작문	○	○	○
	대학, 대학원 재학생의 보고서, 논문/구두 발표 자료	○		○
기획 말뭉치	텍스트의 장르별 기획 말뭉치(논설문, 생활문, 감상문; 대화, 발표 등)	○		○
	백일장, 말하기 대회 자료	○		○
	학습자 언어 습득 고찰을 위한 종적 말뭉치	○		○

	교사와 학습자 상호 작용 연구를 위한 교실/수업 자료 말뭉치(*협조 필요)	○	○	○
	한국어능력시험 작문 자료	○	○	○
자연 발화 말뭉치	(정기적/비정기적 수집)일상대화		○	
	(정기적/비정기적 수집) 계획되지 않은 인터뷰		○	
	(정기적/비정기적 수집)내레이션		○	

## 5. 말뭉치 구축/가공 인력 실무 교육 및 홍보

### 5.1. 교육

- 말뭉치 구축/가공 인력 실무 교육은 체계적이고 일관성 있는 말뭉치 구축을 위한 것으로 실제 말뭉치 구축 작업에 참여하는 실무 작업자들을 대상으로 한다. 따라서 각 구축에 필요한 단계별 지침과 도구 교육이 그 내용을 이룬다.

#### 1) 교육 대상

- 한국어 학습자 말뭉치 구축 단계별 실무 작업자

#### 2) 교육 내용

- 교육 내용은 크게 지침 교육과 도구 사용 교육/실습으로 나뉜다. 각 단계별 업무에 필요한 세부 지침과 도구 사용 교육 이전에 학습자 말뭉치 구축에 관한 제반 절차와 내용에 대한 기본 교육을 실시하여 업무에 대한 이해를 돕고 전문성을 제고한다.

<표 26> 말뭉치 구축/가공 인력 교육 내용

	지침 교육	도구 교육 및 실습
기본 교육	<ul style="list-style-type: none"> <li>○ 한국어 학습자 말뭉치 구축 사업의 개요 및 절차</li> <li>○ 국내외 학습자 말뭉치 구축 현황 및 활용</li> </ul>	<ul style="list-style-type: none"> <li>○ 온라인 구축 도구 전반</li> </ul>
수집	<ul style="list-style-type: none"> <li>○ 구어와 문어 자료 수집을 위한 과제 유형</li> <li>○ 종적 자료 수집 과제 유형 및 자료의 처리</li> <li>○ 학습자 동의서 수집 및 처리</li> <li>○ 수집 자료의 처리와 관리</li> </ul>	<ul style="list-style-type: none"> <li>○ 온라인 구축 도구: 수집 자료 업로드</li> <li>☞ 1차 연도 수집은 구축 본부와 수집 기관의 직접 접촉 방식을 취하였으나 향후 온라인 구축 도구를 통한 수집 기관의 직접 업로드 방식 병행 가능</li> </ul>
자료 처리 및 파일 등록	<ul style="list-style-type: none"> <li>○ 자료의 분류</li> <li>○ 스캔 및 음성 파일 변환</li> <li>○ 파일명 부여 체계</li> <li>○ 학습자 정보 및 파일 정보 등록(헤더 마크업)</li> </ul>	<ul style="list-style-type: none"> <li>○ 온라인 구축 도구: 스캔/음성 원본 파일 업로드 및 파일 등록, 파일명 생성</li> <li>○ 스캐너 사용</li> <li>○ 음성 파일 변환 도구</li> </ul>
입력	<ul style="list-style-type: none"> <li>○ 문어 입력 및 검수 방법, 쟁점</li> </ul>	<ul style="list-style-type: none"> <li>○ 온라인 구축 도구: 파일 입력 및 마크업, 할당 받은 작업 파일 관리 및 작업</li> <li>○ 온라인 구축 도구 외 입력용 텍스트 에디터</li> </ul>
전사	<ul style="list-style-type: none"> <li>○ 구어 전사 및 검수 방법, 쟁점</li> </ul>	<ul style="list-style-type: none"> <li>○ 온라인 구축 도구: 전사 파일 업로드 및 수정, 마크업, 할당 받은 작업 파일 관리 및 작업</li> <li>○ 전사 도구 엘란(ELAN)</li> </ul>
형태 주석	<ul style="list-style-type: none"> <li>○ 형태 분석 방법 및 절차</li> </ul>	<ul style="list-style-type: none"> <li>○ 온라인 구축 도구: 형태</li> </ul>

	<ul style="list-style-type: none"> <li>○ 형태 주석 체계 틀의 구성과 내용</li> <li>○ 형태 분석 자료 검수 및 검수 관련 쟁점</li> </ul>	주석, 할당 받은 작업 파일 관리 및 작업
오류 주석	<ul style="list-style-type: none"> <li>○ 오류 식별, 판정 및 교정의 기준</li> <li>○ 오류 주석 체계 틀의 구성과 내용</li> <li>○ 오류 분석 자료 검수 및 검수 관련 쟁점</li> </ul>	○ 온라인 구축 도구: 오류 주석, 할당 받은 작업 파일 관리 및 작업

### 3) 교육 방법

- 정기/비정기 워크숍: 작업 단계별 지침 및 도구 사용 교육, 실습
- 정기/비정기 세미나: 작업 단계별 쟁점 발표와 토론
- 전자 메일 등을 활용한 수시 교육: 업데이트되는 지침의 즉시 교육

## 5.2. 홍보

- 한국어 학습자 말뭉치를 홍보하는 것은 구축 완료된 말뭉치의 효율적인 보급 및 확산을 위한 것이다. 한국어교육 분야에서 말뭉치를 활용한 연구, 교육 및 학습의 효용성을 알고 있으면서도 실질적으로 사용 가능한 말뭉치의 부재, 말뭉치 활용 방법과 도구 사용 지식의 부족으로 매우 제한적으로 사용되어 왔다. 말뭉치 홍보는 한국어 학습자 말뭉치의 구축을 널리 알리는 것과 함께 (예비) 사용자들의 말뭉치에 대한 이해와 사용 능력을 제고하는 것을 주요한 목표로 한다.

## 1) 말뭉치 공개 및 배포 이전의 사전 홍보 활동

### ○ 한국어 학습자 말뭉치 활용 아카데미 개최

오프라인으로 개최하는 한국어 학습자 말뭉치 활용 아카데미는 주로 국내의 사용자에게 한정된다는 한계가 있다. 따라서 국외 교수자 초청 행사 등과 연계하여 국외 사용자를 대상으로 한 워크숍 방식의 교육 프로그램 운영도 고려해 볼 만하다.

<표 27> 한국어 학습자 말뭉치 활용 아카데미 개최 계획

	방법	유관 기관	주요 대상
국내	학술대회에 ‘학습자 말뭉치’ 관련 독립 분과 운영	한국언어문화학회, 국제한국어교육학회 등 한국어교육 관련 학회	연구자/교사
	국어 정보학 아카데미와의 연계	연세대학교 언어정보연구원의 언어관측소	연구자/교사
	언어교육원 학술대회와의 연계	연세대학교 한국어학당 외	교사/학습자
국외	세계 한국어 교육자 대회	국립국어원	교사
	국외 한국어 교사 초청 교육	국립국어원	교사

### ○ 학습자 말뭉치 구축 및 활용에 관한 학술 발표

### ○ 학습자 말뭉치 구축 협의회를 중심으로 한 기관 차원의 홍보

### ○ 국립국어원, 누리세종학당, 학회 누리집을 통한 홍보 및 교육용 프로그램 공유

## 2) 말뭉치 공개 이후의 사후 홍보 활동

- 학습자 말뭉치를 활용한 연구 프로젝트 사이트의 운영
- 한국어 학습자 말뭉치 사용자 중심의 커뮤니티 결성 및 지원
- 학습자 말뭉치 검색 시스템의 지속적인 업데이트 및 관리
- 한국어교육 연구자 및 교육자 커뮤니티의 배너를 활용한 홍보 및 링크

# Ⅲ. 말뭉치 구축 지침 수립

## 1. 수집 지침

### 1.1. 주요 쟁점

- 한국어 말뭉치 구축의 목적은 연구 및 교수·학습에의 활용에 있다. 따라서 수집과 관련한 쟁점은 주로 자료의 배포와 활용 범위의 문제와 관련된다.

- 자료의 공개와 활용, 학습자의 개인 정보 보호를 위한 IRB 준용

### 1.2. 지침 수립의 기본 방향

#### 1) IRB 준용 문제 검토의 배경

- 미국의 대다수 연방정부에서 적용하는 공용 기본법 벨몬트 보고서에 따라, 사회과학연구의 심리적 위협, 개인의 자율성과 사생활침해, 명예훼손의 위험이 있는 모든 연구에서는 이 보고서의 윤리원칙을 따라 IRB(기관생명윤리심사위원회(Institutional Review Board)) 심의를 받도록 하고 있다.
- 최근 국내에서도 이와 관련한 규정이 확대·강화되고 있는데, 한국어 학습자 말뭉치의 경우 다음과 같은 점에서 IRB 준용에 대한 면밀한 검토가

필요하다.

- 언어 약자인 외국인 학습자의 자료를 수집 대상으로 함
- 학습자가 산출한 자료에 개인 식별 정보(외국인등록번호, 성명, 소속 등) 및 사생활 관련 내용이 포함됨. 특히 종적 말뭉치 수집 대상자의 경우 이러한 문제에 더욱 심각하게 노출될 수 있음
- 자료의 활용을 위해 개인 정보를 수집해야 함

## 2) IRB 준용 문제 검토 결과

- 전문가의 자문을 통해 IRB 준용 문제를 검토한 결과 한국어 학습자 말뭉치 수집과 구축은 생명윤리법 적용 대상이 아니며, ‘기관위원회’ 심의를 면제 받을 수 있는 사업인 것으로 확인되었다.

## 3) 연구 윤리 준수를 위한 말뭉치 수집 및 구축 지침 조정

- 위와 같은 유권해석을 받기는 하였으나 생명윤리 관련하여 최대한 수집물(말뭉치) 산출자의 정보 보호와 자율성을 보장하고자 노력하는 것을 기본 방침으로 하며, 연구 윤리의 준수를 위하여 학습자 말뭉치 수집 지침에 다음과 같은 사항을 포함하였다.

<표 28> 연구 윤리 준수 조건을 고려한 말뭉치 구축 지침

구분	적용 지침
수집	<ul style="list-style-type: none"> <li>○ 수집 실무 담당자에게 IRB 관련 사항을 공지</li> <li>○ 학습자에게 자료 제공 및 이용, 개인 정보 활용에 관한 동의서를 받음</li> <li>○ 개인 정보 수집 항목에 학습자의 성명, 소속 기관의 ID 등의 개인 식별 정보는 일체 포함하지 않음</li> <li>○ 동의서에 연구 목적, 연구 대상자 참여 기간(각 기관의 학기), 정보 보호에 대한 사항, 개인 정보 제공에 관한 사항,</li> </ul>

	<p>철회에 관한 사항 등을 반영함</p> <ul style="list-style-type: none"> <li>○ 일반 수집 대상자와 종적 수집 대상자용 동의서를 구분하며, 종적 수집 대상자용 동의서에 수집 조건 및 내용 등을 자세히 제시함</li> <li>○ 자료 제공 및 이용에 관한 동의서와 개인 정보 조사 자료는 절취하여 각각 관리함으로써 개인을 식별할 수 없도록 함</li> </ul>
구축	<ul style="list-style-type: none"> <li>○ 원문 입력 시, 혹시 정보가 드러날 고유명사(이름)에 해당하는 부분을 모두 익명화하여 전사</li> </ul>
자료의 공개	<ul style="list-style-type: none"> <li>○ [출처 표시 + 상업적 이용 금지 + 변경 금지] 조건으로 자료를 공개함</li> <li>○ 정식 동의서를 받은 종적 학습자 말뭉치만 제시하며 그 외에는 키워드를 중심으로 한 앞뒤 몇 문장으로 제한하는 방안 고려</li> </ul>
기타	<ul style="list-style-type: none"> <li>○ 연구진 및 수집 실무 교사를 대상으로 한 연구 윤리 교육 실시(온라인 강의 수강 후 본부에 확인서 제출)</li> </ul>

#### 4) 학습자 동의서와 개인 정보 수집

- 학습자 개인 정보는 특정 집단의 자료를 선택적으로 추출하여 활용하기 위해 수집된다. 국적, 수준, 제1 언어 등의 학습자 변인에 의해 언어 사용 양상이 달라지기 때문이다.
- 본 연구에서는 이들 항목 중 IRB 규정에 따라 학습자 개인의 인권과 사생활을 보호하면서도, 연구나 교수·학습에의 자료 활용에 필요한 주요 항목을 중심으로 학습자 개인 정보를 수집하기로 하였다.
- IRB 규정에 따라 개발된 학습자 동의서와 개인 정보 수집 양식은 다음과 같다. 동의서는 학습자의 이해를 돕기 위하여 한국어를 포함하여 7개 언어로 작성되었다.

☐ 동의서 양식 1. 횡적 자료(일반) 수집 학습자용

한국어 학습자 말뭉치 구축 사업을 위한 학습자 자료 이용 동의서

국립국어원에서 한국어교육의 질적 향상을 위해 학습자들의 언어 자료(말뭉치)를 수집하여 활용하는 사업(사업 수행: 연세대학교 산학협력단)을 추진하고 있습니다. 여러분이 제공한 자료는 한국어 교수 방법 개선, 한국어 교재 개발, 한국어 교육 분야 및 인접 학문 분야의 연구에 사용됩니다. 이 연구에 참여하는 분들은 경제적 인 손해나 신체적 위험이 없습니다. 만약 참여를 원하지 않을 때에는 참여 의사를 철회할 수 있습니다. 또한 수집하는 개인 정보는 본 사업의 목적 외로는 사용되지 않으며, 비밀 유지를 위하여 식별할 수 없는 형태로 사용될 것입니다. 감사합니다.

문의처: 연세대학교 산학협력단

02-2123-4199

☐ 아래 문항과 2015년 여름/가을 학기의 쓰기/말하기 자료를 제공하며 사용을 허락합니다.

날짜 \_\_\_\_\_

이름 \_\_\_\_\_ (서명)

✂-----

다음은 연구를 위한 자료로 활용될 정보입니다. 개인 신상 정보는 비밀이 보장되며 외부로 유출되지 않습니다. (가능하면 한국어로 응답해 주세요. 필요하면 영어를 사용해도 좋습니다.)

1. 성별: ☐ F ☐ M

2. 나이: \_\_\_\_\_

3. 현재 등급: \_\_\_\_\_

4. 국적: \_\_\_\_\_ ( ※ 교포 여부 ☐ 교포 ☐ 외국인 )

5. 제1 언어: \_\_\_\_\_

6. 한국어 학습 기간(한국어를 얼마 동안 공부했습니까?): \_\_\_\_\_ 년 \_\_\_\_\_ 개월

(예. 1년 3개월)

7. 한국에서의 거주 기간(한국에서 얼마 동안 살았습니까?):      년      개월

(예. 1년 3개월)

8. 한국어 학습 목적

☐ 진학   ☐ 취업   ☐ 거주   ☐ 취미   ☐ 결혼   ☐ 기타 (                      )

9. 직업:

10. 한국어 외의 사용 가능 외국어(잘하는 언어 순서대로 쓰시오):

☐ 동의서 양식 2. 원문/음성 파일 제공에 관한 선택적 동의서(학문 목적)

#### 한국어 학습자 말뭉치 원문/음성 파일 이용에 관한 동의서

본인은 국립국어원의 한국어 학습자 말뭉치 구축 사업의 취지를 이해하고 2015년에 산출한 쓰기/말하기 자료의 제공과 연구 목적의 사용을 허락하였습니다. 이에 덧붙여 보다 광범위한 연구에 이용될 수 있도록 쓰기 원문/음성 녹음 자료 전체의 공개와 사용을 허락합니다. 이 경우에도 비밀 유지를 위하여 개인 정보가 식별할 수 없는 형태로 사용되며 외부로 유출되지 않을 것임을 보장 받았습니다.

☐ 저는 위의 내용을 충분히 이해하였으며 동의합니다.

날짜 \_\_\_\_\_

전공 \_\_\_\_\_

이름 \_\_\_\_\_ (서명)

문의처: 연세대학교 산학협력단 02-2123-4199

☐ 동의서 양식 3. 종적 자료 일반 학습자용

한국어 학습자 말뭉치 구축 사업을 위한 학습자 자료 이용  
동의서(종적)

국립국어원에서 한국어교육의 질적 향상을 위해 학습자들의 언어 자료(말뭉치)를 수집하여 활용하는 사업(사업 수행: 연세대학교 산학협력단)을 추진하고 있습니다. 여러분이 제공한 자료는 한국어 교수 방법 개선, 한국어 교재 개발, 한국어 교육 분야 및 인접 학문 분야의 연구에 사용됩니다.

- 수집 시기: 1-6급 매학기(격주 1회)
- 수집 내용: 수준별 주제에 관한 쓰기 자료 1편, 말하기 자료 1편
- 소요 시간: 50분 내외 소요
- 사례: 20,000원(말하기 1편, 쓰기 1편/1회, 50분 내외 소요)
- 사례 지급 방법: 매 학기별로 계좌 입금

이 연구에 참여하는 분들은 경제적인 손해나 신체적 위험이 없습니다. 만약 참여를 원하지 않을 때에는 참여 의사를 철회할 수 있습니다. 또한 수집하는 개인 정보는 본 사업의 목적 외로는 사용되지 않으며, 비밀 유지를 위하여 식별할 수 없는 형태로 사용될 것입니다. 감사합니다.

문의처: 연세대학교 산학협력단

02-2123-4199

☐ 아래 문항의 정보와 1-6급의 쓰기/말하기 자료를 격주로 제공하며 쓰기 원문/음성 녹음 자료 전체의 공개와 연구 목적의 사용을 허락합니다.

날짜 \_\_\_\_\_

이름 \_\_\_\_\_ (서명)

✂-----

다음은 연구를 위한 자료로 활용될 정보입니다. 개인 신상 정보는 비밀이 보장되며 외부로 유출되지 않습니다. (가능하면 한국어로 응답해 주세요. 필요하다면 영어를 사용해도 좋습니다.)

1. 성별: ☐ F ☐ M

2. 나이: \_\_\_\_\_
3. 현재 등급: \_\_\_\_\_
4. 국적: \_\_\_\_\_ ( ※ 교포 여부 ☐ 교포 ☐ 외국인 )
5. 제1 언어: \_\_\_\_\_
6. 한국어 학습 기간(한국어를 얼마 동안 공부했습니까?): \_\_\_\_\_ 년 \_\_\_\_\_ 개월  
(예. 1년 3개월)
7. 한국에서의 거주 기간(한국에서 얼마 동안 살았습니까?): \_\_\_\_\_ 년 \_\_\_\_\_ 개월  
(예. 1년 3개월)
8. 한국어 학습 목적  
☐ 진학 ☐ 취업 ☐ 거주 ☐ 취미 ☐ 결혼 ☐ 기타 ( \_\_\_\_\_ )
9. 직업: \_\_\_\_\_
10. 소속: \_\_\_\_\_ (전공: \_\_\_\_\_ )
11. 한국어 외의 사용 가능 외국어(잘하는 언어 순서대로 쓰시오): \_\_\_\_\_

## 2. 자료 처리 지침

### 2.1. 주요 쟁점

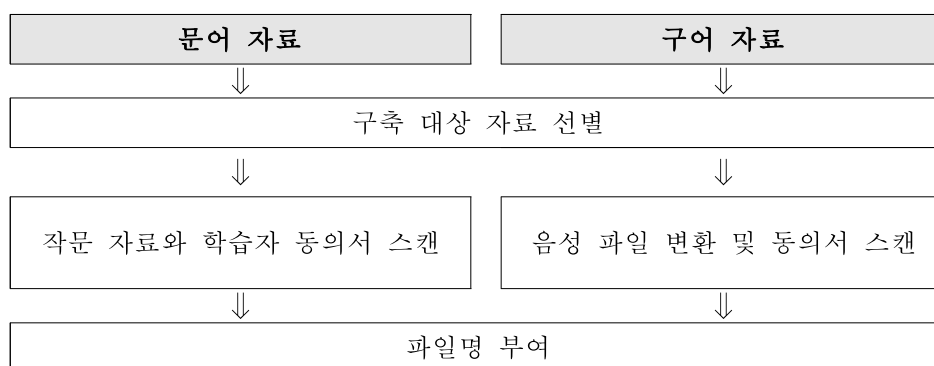
○ 자료 처리는 본격적인 말뭉치 구축을 위한 전 단계로 수집 자료를 분류하고, 전산화한 후, 효율적으로 관리하도록 하는 데에 목적이 있다. 따라서 자료 구축과 관리가 잘 연계되도록 처리 절차를 마련하는 것이 중요하다.

- 자료 처리 절차의 체계화
- 구축 대상 자료 선별 기준
- 자료에 관한 기본 정보를 포함한 파일명 부여 체계

## 2.2. 지침 수립의 기본 방향

### 1) 자료 처리 절차의 체계화

- 자료 처리는 파일을 전산화하여 말뭉치 자료로서 본격적인 구축과 가공 작업을 하기 위한 전처리 단계이다. 따라서 우선적인 구축 자료를 선정하여 자료를 기계가독형의 자료로 변환하여 쉽게 관리할 수 있도록 하는 것이 중요하다. 다음은 본 연구에서의 자료 처리 절차이다.



<그림 5> 한국어 학습자 말뭉치 자료 처리 절차

### 2) 말뭉치 구축 대상 자료 선별 기준

- 말뭉치 구축을 위해서는 IRB 규정에 따라 학습자의 서명이 완료되고 자료의 활용을 위해 필요한 개인 정보가 빠짐없이 입력이 되어야 한다. 그 외에도 다음과 같은 기준으로 우선적으로 구축할 자료를 선정하도록 한다.

<표 29> 말뭉치 구축 대상 자료 선별 지침

문어	구어
<ul style="list-style-type: none"> <li>○ 학습자 동의서에 서명한 자료</li> <li>○ 학습자 동의서의 개인 정보 모두 입력된 항목 선정</li> <li>○ 동일 학습자의 자료 2개 이하로 제한</li> </ul>	

○ 영어권, 일본어권 자료/1, 5, 6급 단계의 자료 우선 선정	
○ 완결된 텍스트 작문 자료 선정	○ 완결된 담화 단위의 발화 자료 선정
○ 텍스트의 길이 평균 100어절 이상의 자료 선정. 단, 숙달도 단계를 고려하여 1, 2급은 50어절 내외의 자료를 포함함	○ 발화 길이 2분 이상의 자료 선정
○ 복사 또는 스캔 파일의 경우 화질이 좋은 자료 선정	○ 음질이 좋은 자료 선정
	○ 교사의 개입이 많지 않고 학습자의 발화가 중심인 자료를 우선 선정

### 3) 파일명 부여

- 자료의 효율적인 관리를 위하여 자료의 유형과 국적, 수집 기관, 수준 등의 정보가 포함된 파일명을 부여한다. 파일 분류 및 파일명 부여 체계는 다음과 같다.

예) 종적\_문어\_중국\_서울대\_1급\_0001\_01.txt

자료 코드		학습자 변인 정보 코드				
종적	문어	중국	서울대	1급	0001	01
자료 및 학습자 유형		국적	자료 번호    페이지 번호			
자료 유형		수준				
수집 기관: 코드화하여 비공개 처리됨						

<그림 6> 한국어 학습자 말뭉치 파일명 부여 체계

<표 30> 한국어 학습자 말뭉치 파일명 코드

구분	범주	설명	항목	코드
자료 코드	자료 및 학습자 유형	학습자의 특성에 따른 분류	일반 학문 목적 종적 이주	일반 학문 종적 이주
	자료 유형	자료의 유형을 구분하는 코드 부여	문어(Written) 구어(Spoken)	문어 구어
학습자 정보 코드	언어권	학습자의 제1 언어를 구분하는 코드 부여	중국어 일본어 베트남어 영어 ...	중국 일본 베트남 영어 ...
	자료 수집 기관	자료 수집 기관명	서울대 경희대 ...	서울대 경희대 ...
	수준	학습자의 수준을 구분하는 코드 부여	1급 2급 3급 4급 5급 6급 최고급(Superior)	1 2 3 4 5 6 7
	학습자 구분 번호	기관의 학습자 구분을 위한 일련번호	0001 0002 ...	0001 0002 ...
자료 번호	자료 번호	동일한 학습자가	01	01

		두 개 이상의 자료를 제공할 경우 자료를 구분하기 위한 일련번호	02 ...	02 ...
--	--	---	-----------	-----------

### 3. 문어 자료 입력 지침

#### 3.1. 주요 쟁점

- 학습자가 산출한 한글 텍스트를 원문 그대로 입력하는 것을 원칙으로 한다. 이때 한글의 체계나 어문 규범에 익숙하지 않은 외국인 학습자는 모어 화자에게는 나타나지 않는 독특한 오류들을 유발하는 경우가 많으며, 이것을 어떻게 처리할 것인가가 문어 입력의 주요한 쟁점이 된다.

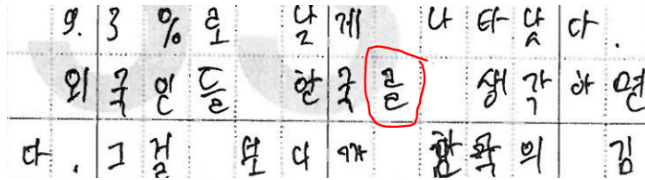
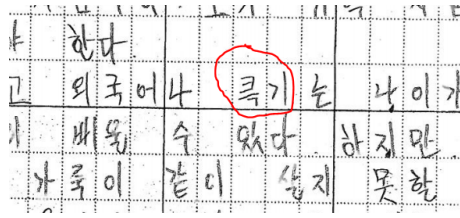
- 한글에 없는 글자의 처리
- 띄어쓰기와 문장부호의 처리

#### 3.2. 지침 수립의 기본 방향

##### 1) 한글에 없는 글자의 처리

###### (1) 처리상의 쟁점

- 외국인 학습자는 한글의 음절 구성 방식에 익숙하지 않기 때문에 한글에 없는 글자를 자주 산출하게 된다. 이러한 현상은 다음의 예와 같이 컴퓨터 자판으로 입력이 되지 않는 글자의 경우 전산화가 불가능하게 된다.



<그림 7> 학습자 자료 예시: 한글에 없는 글자

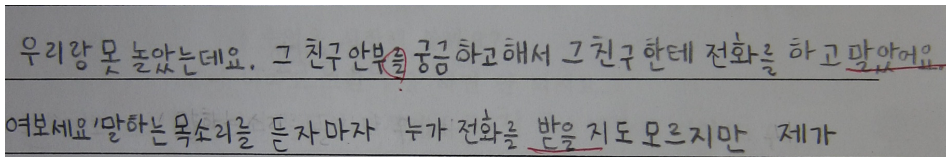
## (2) 본 연구의 처리 방안

- 이러한 경우 해당 자리에 형태 오류임을 나타내는 태그를 사용하여 ‘<MISF>한글에 없는 글자</MISF>’와 같이 주석 처리한다.

## 2) 띄어쓰기 및 문장부호의 처리

### (1) 처리상의 쟁점

- 어문 규범은 한국어 모어 화자도 정확히 알기가 어렵지만, 외국인 학습자의 경우는 더더욱 그러하다. 특히, 띄어쓰기의 경우 여백이 일정하지 않아 띄어쓰기를 하고 있는지도 파악하기 어렵거나 아예 무시하는 경우가 많다. 따라서 일관된 기준에 의해 띄어쓰기를 반영하기가 어렵다.



<그림 8> 학습자 자료 예시: 띄어쓰기 구분이 어려운 경우

- 한편, 가공 말뭉치 구축을 고려할 때 띄어쓰기 오류의 연구를 위해 학습자가 쓴 대로 입력할 경우 어절 단위의 지능 형태소 분석 도구를 사용할 경우 정확성이 현저하게 낮아져 실질적으로 분석이 불가능해지게 된다.

## (2) 본 연구의 처리 방안

- 본 연구에서는 어문 규범에 따라 띄어쓰기를 수정하여 입력한다. 이는 띄어쓰기 오류 연구가 한국어 교육에서 그다지 큰 주목의 대상이 되지 않고 있어 효용성이 크지 않다고 판단했기 때문이다. (스캔본은 그대로) 또한 본 연구에서는 형태 주석 말뭉치를 구축하게 되므로 자료 처리의 효율성을 고려할 필요가 있다.

## 4. 구어 자료 전사 지침

### 4.1. 주요 쟁점

- 구어 전사는 구어 변이형, 준언어적인 요소의 표기 등의 문제로 논의거리가 많다. 특히 한국어의 음운 체계에 익숙하지 않은 외국인 발화 자료라는 점에서 모어 화자의 자료보다 훨씬 더 여러 가지 복잡한 문제에 직면하게 된다. 한국어 학습자 말뭉치 구어 전사 지침은 자료 활용 시의 호환성과 국가 주도 사업 결과물 간의 연계성을 고려하여, 21세기 세종 한국어 균형 말뭉치의 구어 전사 지침에 학습자 성격을 고려해 수정·보완하였다. 이때 고려 사항은 다음과 같다.

- 전사 단위의 문제
  - 억양 단위의 기준과 전사 도구에서의 반영
- 표기의 문제
  - 학습자의 발음 오류 및 한국어의 음운 체계에는 없는 독특한 발음 표기 (중간음 발음, 외국어 발음)
  - 외래어 발음 시 규범 발음과 현실 발음이 다른 경우의 표기
- 한국어 학습자에게 주로 나타나는 독특한 강세, 억양 처리의 문제
- 전사 기호 및 마크업 체계의 문제

## 4.2. 지침 수립의 기본 방향

### 1) <한국어 학습자 말뭉치>와 <21세기 세종 균형 말뭉치> 전사 지침의 비교

○ <한국어 학습자 말뭉치>와 <21세기 세종 균형 말뭉치> 전사 지침은 기본적으로 학습자 자료와 모어 화자 자료의 비교 분석을 위해 기본 틀을 공유하고 있다. 그러나 <한국어 학습자 말뭉치>가 비모어 화자인 외국인 학습자의 구어 자료를 대상으로 하고 있기 때문에 학습자의 발음 특성이나 오류에 의한 독특한 발음의 처리를 위한 추가적인 지침의 마련이 필요하였다. <한국어 학습자 말뭉치>와 <<21세기 세종 한국어 균형 말뭉치>의 구어 전사 지침을 비교해 보면 다음과 같다.

<표 31> <21세기 세종 균형 말뭉치>와 <2015 한국어 학습자 말뭉치> 전사 지침의 비교

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
발화자 정보	발화자 표시	P1	<person id=1 sex=M age=20s> 1:	유지
	분명하지 않을 때	?	P?:	유지

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
	동시 발화	모두/ 나머지/ P2,P3	P2,P3:네.	유지
억양 단위	하강	.	2:네.	유지
	상승	?	2:어디 갈 거예요?	유지
	약한 상승이나 하강	,	1:그래서 그랬는데 이번에,	유지
	활기, 기운찬 어조	!	1:아!	유지
	억양 단위 경계의 처리	스페이스 없이 엔터(enter)	2:어디 갈 거예요? 1:안 아직까지 그냥 계획만 잡아 왔는데, 2:음.	유지
	하나의 억양 단위가 끼어들어 의해 끊어진 경우	-	6:기자가 와서 - 2:응. 6:- 그 사람한테 인터뷰를 시작했어.	유지
	두 억양 단위가 휴지 없이 이어질 경우	&	3:요거는 교수 학습의 개요지, &요 표는, 4:아::,	유지
겹침 현상	겹침 현상	세종: [ ] 학습자: 전사 도구의 시간 표시로 대체	<세종> 1:여의도 거기 [벚꽃]했잖아요. 2:[윤중로] <학습자> 1 03:49.2 03:51.3 네. 다 거짓말이기 때문에. 2 03:50.8 03:52.2 아 왜 거짓말을 하나요? ☞ 발화 겹침이	수정 ☞ 엘란에서 자동으로 시간 표시

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
			있음을 알 수 있음	
	연속적인 겹침 현상	번호로 발화자 구분	2:자~ 이제~ [1이건 그::~ -1] 3:[1무조건 자기들이1] 안 만들었다 그런 식으로 [2어거지를2] 쓰면 안 되죠. 2:[2뒤~ 우리2] 바로 이런 점 때문에	수정 ☞ 엘란에서 자동으로 시간 표시되므로 일반적인 겹침과 동일하게 처리
잘 들리지 않는 부분	잘 들리지 않는 부분	<X X>	<X보통X>	유지
	전혀 들리지 않는 부분	세종: ... 학습자: <X안들림X>	<세종> 1: 거기까지 ... 2: ... 너무나 거 같더라. <학습자> 1: 거기까지 <X안들림X> 2: <X안들림X> 너무한 거 같더라.	수정
	들리지 않는 음절 수만큼	X	2: 근테 그거 진짜 XX해야 되겠더라	유지
전사자의 설명	전사자의 설명	<note></note>	1:응. <note><p>장소 이동으로 인해 잠시 멈춤.</p></note> 2:우리 때는 그런 거 없었잖아.	유지
	상호작용이 없이 나타나는 동시다발적 대화	<note></note>	<note><p>첫 번째 그룹의 대화임</p></note> 1:... 그러니까 이제?	유지

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
			인터넷에도 잘 사용하는데 이미 저장 ... <note><p>첫 번째 그룹의 대화 끝남. 아래는 두 번째 그룹의 대화로, 첫 번째 그룹의 대화와 동시에 진행된 것임.</p></note> 4:이리 와. 5:<vocal desc='개부를때내는 혀차는소리'> 이리 이리 와 봐. 이리 와 이리 와.	
혼잣말		<세종> 없음  <학습자> <monologue ></monolog ue>	<세종> 없음  <학습자> <monologue>미치겠 네.</monologue>	신규 추가
표기 지침	구어의 발음 특성, 개인의 발음 특성, 지역적인 특성 등에 의해 철자법대로 소리 나지 않는 발음(표준 발음이 아닌 경우)	소리 나는 대로 적고, 원래의 형태가 없이 내용을 이해하기 어려울 때에는 ( ) 안에 규범 표기를 밝힘	1: 청구(친구)와 간남(강남)에 갔습니다.(갔습니다)	유지

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
	숫자 표기	발음에 따라 한글로 표기	7:오늘 제 동생이 이케 하나 오백 원이라고 사 가지고 왔더라구.	유지
	외래어 · 외국어 표기	발음에 따라 한글로 표기하고, 원래의 형태가 없이 내용을 이해하기 어려울 때에는 ( ) 안에 규범 표기를 밝힘	1:이거도 오리지날(오리지널) 제주도 감귤이 아니야.	유지
	끊어진 단어(불완전하게 발화된 단어)	=	4:사실 학습 자료랑 학= 형태는 떨어져도 되는데.	유지
	한 어절 발화 도중 다른 억양 단위로 전사될 때 조사나 어미에	=	1:주부 우울증, =이라고 말할 수 있겠습니다.	유지
	띄어쓰기	맞춤법에 따름		유지
	축약형	'(apostrophe, 영문따옴표) 를 사용해서 두 음소를 연결	사귀'어, 바뀌'어, ...	유지
	표현적 장음	::	1:많은 경우에 논문, 저:: 어:: 연구는 네이션,	유지

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
	담화표지	물결표(~)	1:많은 경우에 논문, 그~ 어::~ 연구는 네이션, 국가라는 거하구 직결되는 과정이죠.	유지 ☞ 담화표지 구분이 어려워 실제 전사는 보류함
준음성	웃음	<vocal desc='웃음'>	5:<vocal desc='웃음'> 맞아요.	수정 ☞ 대화 흐름에 영향을 주는 경우에만 반영. 특히, 학습자 개인의 발화 특성으로 습관적으로 반복해서 들이마시는 숨소리나 혀 차는 소리, 헛기침 등은 반영하지 않음
	기침	<vocal desc='기침'>		
	하품	<vocal desc='하품'>		
	재채기	<vocal desc='재채기'>		
	목청 가다듬는 소리(음, 으음)	<vocal desc='목청가 다듬는소리'>		
	들이마시는 숨(쓰)	<vocal desc='들이마 시는숨'>		
	내쉬는 숨(후우)	<vocal desc='내쉬는 숨'>		
	혀 차는 소리(쓰)	<vocal desc='혀차는 소리'>		
	헛기침(에 험)	<vocal desc='헛기침'>		
	한숨	<vocal desc='한숨'>		
	노래	<vocal desc='노래'>		

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
	웃으면서 말하는 부분	<@ @>		
	박수치면서 말하는 부분	<# #>	<#이리 와 이리와.#>	
	노래를 부르는 부분	<M M>		
	박수나 손가락 부딪치는 소리	<kinesics desc=' '>	박수 <kinesics desc='박수'>	유지
	대화 흐름에 영향을 주는 전화벨 소리라든지 기타 음성 아닌 소리	<event desc=' '>	<event desc='전화벨소리'>	유지
2차 전사	구어적 변이형	철자법 보충. 단, 주 등장하거나 쉽게 원래 형태로 이해될 수 있는 것들은 일일이 철자형을 붙여주지 않음	1:브릿지를 한 가닥을 너(넣어) 줬어요.	유지
	발음 오류	한국어 모어 화자의 발음과 음운적으로 구분이 될 정도로 발음에 오류가 있는 경우 ( )	청구(친구)와 간남(강남)에 갔습니다(갈습니다). 가티(같이) 가자. ㄱ와 ㄲ의 중간 발음-요자(여자) 기과 ㄷ의 중간 발음-휘사(회사) ㅎ와 ㄱ의 중간	신규 추가

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
		안에 철자 보충	발음-화반수(과반수)	
		한국어 모어 화자의 발음과 음성 혹은 변이음의 구분이 모호한 경우 ( ) 안에 철자 보충	‘가’의 ㄱ을 유성음으로 발음_가구(가구) ‘구’의 ㄱ을 무성음으로 발음-가구(가구) ‘심’의 ㅁ을 개방음으로 발음- 안녕하십니까(안녕하 십니까)	신규 추가
		음운규칙으로 인해 한국어의 철자대로 발음되지 않으나, 학습자가 이를 철자대로 발음하는 경와 같이 철자 전사를 통해 학습자의 발음 오류를 반영하기 어려운 경우 [ ] 안에 학습자 발음 표기	경음화-무조건[무조 건] 구개음화-같이[가티, 같이] 자음동화-신라[신라] 연음-앞에[앞에] 자음동화-먹는[먹는]	신규 추가
	외국어, 외래어 발음	외국어나 외래어를	인텔뷰(인터뷰) 세너(센터)	신규 추가

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
		원어에 가깝게 발음할 경우 ( ) 안에 철자 보충	팔너(파트너)	
		한국어 모어화자와 다른 규범을 할 경우 [ ] 안에 학습자 발음 표기	카페: 현실 발음 [까페], 학습자 발음 [카페] 버스: 현실 발음 [빠스], 학습자 발음 [버스] → 각각 ‘카페[카페]’, ‘버스[버스]’로 처리함	신규 추가
	방언형 표시	확실한 방언형(대응 하는 표준형 형태소가 없는 것)의 경우 표기	차는 <dia>여일</dia> 있어.	유지 ☞ 지방 거주 학습자의 경우에 나탈 수 있음
	긴 휴지	(1초 이상의 쉽은) 0.1초 단위까지 표시	2:{1.2} 그럴까?	유지
	짧은 휴지	한 어절 안에서의 짧은 쉽은 ‘.’로 표시, 하나의 억양단위 내부에서의 짧은 쉽은 따로 표시	2:아::~~ 그리고 어::~~ 남의 의견을 잘 듣고 수용하고 대화..로 타협해야 된다고 하면서,	유지

구분		21세기 세종 한국어 균형 말뭉치 전사 지침		2015 한국어 학습자 말뭉치 전사 지침
대분류	소분류	기호	예시	
	인용	<Q Q>	<Q이제는 불확실성의 시대는 아니다. 이제는 뭐냐? 다양성의 시대다.Q> 라고 말하죠.	유지
	텔레비전 방송이나 강의 등 텍스트 종류 표현	<R R>	1:그 답에 <R생각과 느낌이 유기적으로 잘 짜여져 조직체를 이룰 때 좋은 글이 될 수 있다R>라고 돼 있어요.	유지
	익명성 보장을 위한 마크업	<name> : 사람 이름, 단체 이름, 학교 이름 등 <social security number> : 주민등록번호 <card-num> : 신용카드 번호 <address> : 주소 <tel-num> : 전화 번호	5: 그게 어찌면 <name1> 선배님이라든지 다른 선배님들 말::들은 걸 생각해 보면,	유지

## 2) 한국어 학습자 발화 자료의 특징과 전사 지침의 수립

### (1) 억양 단위의 기준과 전사 도구에서의 반영

#### ① 처리상의 쟁점

- 억양 단위는 다음과 같은 문제점을 가진다.
  - 명확한 판단 기준이 없어 전사 작업자 간의 차이가 크다. 특히, 전사 도구에서 억양 단위로 음성 자료를 분절(segmentation)하게 되는데, 분절 단위와 억양 단위가 일치하지 않는 경우가 많다. 그 결과 작업자에 따라 매우 짧게 혹은 길게 억양 단위를 나누게 된다.
  - 외국인 학습자의 발화 특성상 비정상적인 억양이나 강세의 사용으로 억양 단위 구분이 어렵다.

#### ② 본 연구의 처리 방안

- 외국인 학습자의 발화는 학습한 것을 그대로 실현하는 경우가 많아 모어 화자에 비해 통사 단위에 의한 발화가 많은 편이어서 모어 화자의 발화와 달리 억양 단위의 효용성이 크지 않은 경우도 있다. 억양 단위의 구분이 모호하며, 학습자 발화의 경우 비교적 통사 단위의 발화가 많은 점을 고려하여 지나치게 짧은 단위로 억양 단위를 자르지 않도록 한다. 일정한 의미를 포함한 절 단위를 최소의 분절 단위로 한다.
- 아울러 전사 도구에서의 분절 단위가 억양 단위가 일치하지 않아도 된다. 전사가 완료된 후 텍스트 파일로 출력이 되므로 도구 내에서 길게 자르더라도 필요한 위치에 억양 구분 기호를 삽입하기만 하면 된다.

### (2) 학습자의 발음 오류 및 한국어의 음운 체계에는 없는 독특한 발음 표기

#### ① 처리상의 쟁점

- 학습자의 발음 오류는 음소, 음절, 음운규칙에 의한 오류 외에도 한국어의 음운 체계에 없는 중간 발음이나 모국어식의 발음에 의한 음성 차원의 오류까지 그 양상이 매우 다양하다. 따라서 이들 전체를 포괄할 수 있는 명확한 지침이 필요하다.

가. 음운규칙을 적용하지 못하여 발생한 오류

예) 앞에[앞에], 무조건[무조건]

나. 불분명한 발음, 중간 소리의 표기

예) 건강 (모음 ㅏ와 ㅓ의 중간 발음으로 ‘건강’과 ‘곤강’의 경계에 있을 때)  
와이파이 (자음 ‘ㅍ’와 ‘ㅃ’의 중간 발음으로 ‘와이파이’와  
‘와이빠이’의 경계에 있을 때)

다. 음절 발음에서의 오류

예) 불파음의 실현: 피라미드[피라민트], 방식[방식크]  
겹자음의 받침이 아닌데도 두 개의 소리를 내는 경우: 할가요[함까요]  
학습자가 한국어의 무성음을 유성음으로 처리하는 경우

라. 전반적으로 발음이 좋지 않아 대부분 오류로 들리는 경우

예) 과반수[화반수?], 회사[괴사?], 능력[능력?]

## ② 본 연구의 처리 방안

- 한국어의 음운 체계에 없는 음운의 발음이나 표기가 어려운 중간 발음, 외국인 학습자에게만 나타나는 독특한 발음을 할 경우 소리 나는 대로 적되, 가장 가까운 소리가 나는 한글 표기로 전사한다. 이때, 원래의 표기가 없이 내용 파악을 하기 어려운 경우 ( ) 안에 규범 표기를 넣는다. 또한 음운적 구분이 모호하거나 특징적인 사항이 있을 때에 메모를 남기도록 한다.

유형 1. 한국어 모어 화자의 발음과 음운적으로 구분이 될 정도로 발음에 오류가 있는 경우

예) 성생닌(선생님)

예) 요\_크와 ㄱ의 중간 발음\_자(여자)

예) 휘\_기과 ㅅ의 중간 발음\_사(회사)

유형 2. 한국어 모어 화자의 발음과 음성 혹은 변이음의 구분이 모호한 경우

예) 가\_ㄱ를 유성음으로 발음\_구(가구)

예) 가구\_ㄱ를 무성음으로 발음\_(가구)

예) 안녕하심\_口을 개방음으로 발음\_니까(안녕하십니까)

유형 3. 단, 음운규칙으로 인해 한국어의 철자대로 발음되지 않으나, 학습자가 이를 철자대로 발음하는 경우

예) 무조건[무조건] - 경음화

예) 같이[가티, 같이] - 구개음화

예) 신라[신라] - 자음동화

- 유형 3의 경우는 철자 전사를 통해 이를 반영하기 어려우므로 원래의 표기를 먼저 적고, 학습자의 실제 발음을 [ ]에 남겨 오류가 있음을 알 수 있도록 한다.

### (3) 외래어의 규범 발음과 현실 발음이 다른 경우의 표기

#### ① 처리상의 쟁점

- 규범 발음과 현실 발음이 다른 경우가 있다. 특히 외국인 학습자 발화의 경우 외래어 또는 외국어 발화 시 원어식의 발음을 하거나 한국어 모어화자의 현실 발음이 아닌 규범 발음을 하여 그러한 현상이 더욱 두드러지게 나타난다. 이는 오류는 아니지만 그 정도의 차가 커서 모어 화자와 다른 발음 특성 중의 하나로 볼 수 있으므로 전사 시 어떻게 반영해야 할 것인지에 대한 고려가 필요하다.

예) 인터뷰: 영어식 발음을 사용하여 [인틸뷰]에 가까운 소리가 남

센터: 영어식 발음을 사용하여 [세너]에 가까운 소리가 남

파트너: 영어식 발음을 사용하여 [팔너]에 가까운 소리가 남

#### ② 본 연구의 처리 방안

- 학습자의 실제 발음으로 전사하되 한국어의 음운 체계로 전사가 불가능한 발음의 경우는 가장 가까운 소리가 나는 한국어 표기로 전사한다. 이때, 원래의 형태 표기가 없이 내용을 이해할 수 없는 경우에 한해서만 규범 표기를 밝히는 것으로 한다.

예) 이너뷰(인터뷰)/ 세너(센터)/ 팔너(파트너)

- 외국인 학습자 발화의 경우 외래어 또는 외국어 발화 시 원어식의 발음을 하거나 한국어 모어화자의 현실 발음이 아닌 규범 발음을 하여 어색하게 들리는 경우가 있다. 이 경우 철자 전사를 통해 반영하기 어려우므로 원래의 표기를 먼저 적고, 학습자의 실제 발음을 [ ]에 남겨 오류가 있음을 알 수 있도록 한다. 예) 버스[버스] / 카페[카페]

#### (4) 한국어 학습자에게 주로 나타나는 독특한 강세, 억양 처리

##### ① 처리상의 쟁점

- 실제 전사를 하다 보면 학습자의 실제 발화는 굉장히 어색한 억양으로 일관되게 발화하고 있지만, 텍스트화 된 전사 자료에서는 한국어를 유창하게 구사하는 것처럼 보이는 문제가 있다. 또한 문법적으로 정확하게 구사하고 있지만, 억양이 전체적으로 모어 화자와 달라 부자연스럽다.

예) 한국어의 일반적인 억양 패턴과 다른 억양이 실현되는 경우

걱정: [걱뽕], [걱..정]

예) 불필요한 강세가 실현되는 경우

##### ② 본 연구의 처리 방안

- 현 지침에 따라 억양 단위 전사를 통해 나타나는 억양만 포함하며 그 외의 독특한 억양이나 강세는 반영하지 않는다. 단, 전사 시 주석자의 설명에 관련 사항을 남겨 사용자가 개방형 주석 시스템을 사용하여 목적에 따라 추가 전사를 할 수 있도록 한다.

#### (5) 전사 기호 및 마크업 체계의 문제

##### ① 처리상의 쟁점

- 2015년 한국어 학습자 말뭉치에서는 앞서 제시한 지침에 따라 전사 작업을 수행하였다. 현행 지침에 포함된 전사 기호는 호환성을 위하여 <21세기 세종 균형 말뭉치>에 따르고 있다. 그런데 형태 주석, 오류 주석 시의 마크업 제거 작업의 효율성, 자료 배포 시 개인 정보 보호를 위한 삭제 처리의 용이성 등에서 향후 사용하게 될 한국어 학습자 말뭉치 구축 도구와의 호환성을 고려할 필요가 있다.

② 본 연구의 처리 방안

- 시스템 팀과의 협의를 통해 ‘마크업 체계(수정안)<표32>’를 마련해 둔 상태이며, 이에 따라 마크업 체계를 변환할 예정이다.<sup>2)</sup>

### 4.3. 2016년 구어 전사 지침 및 기구축 자료 보완 방향

- 구어 전사 지침은 실제 전사 작업에서 추가적으로 논의되는 쟁점들을 정리한 후 실제 예시와 함께 지속적으로 보완해 나아갈 예정이다. 아울러 온라인 구축 시스템 작업에서의 효율성을 고려하여 다음의 ‘마크업 체계(수정안)’에 따라 원시 말뭉치를 수정하고, 지침을 보완한다<sup>3)</sup>.

<표 32> 전사 마크업 체계 수정안

대분류	소분류	현행 기호	수정 기호
발화자 정보 (person)	발화자 표시	1	<PERSON_ID WHO="P1">
	분명하지 않을 때	?	<PERSON_ID WHO="UNKNOWN">
	필요에 따라	모두/나머지/2,3	<PERSON_IDWHO="ALL"> <PERSON_IDWHO="OTHERS"> <PERSON_IDWHO="P2,P3">
역양 단위	하강	.	학생이 왜<FALLING>요</FALLING>.
	상승	?	알고 있어<RISING>요</RISING>?
	약한 상승이나 하강	,	아니<LEVEL>면</LEVEL>
	활기, 기운찬 어조	!	<ANIMATED>아</ANIMATED >
	길게 발음될	::	모두

- 2) 2015년에 구축된 말뭉치는 기술적으로 일괄 변환한 후에 2차적인 검수 작업을 통해 수정이 누락되거나 잘못 처리된 부분을 수정할 예정이다.
- 3) ‘마크업 체계(수정안)’은 온라인 구축 시스템 환경과 학습자 말뭉치 구축 작업의 효율성을 고려하여 시스템 팀에서 구안하여 제안한 것이다.

대분류	소분류	현행 기호	수정 기호
	때		<LENGTHENING>다</LENGTHENING>
	하나의 억양 단위가 끼어들어 의해 끊어진 경우	-	<PERSON_IDWHO="P1"> 그런거<TRUNC_IN/> <PERSON_IDWHO="P2"> 그<RISING>래</RISING>? <PERSON_IDWHO="P1"> <TRUNC_IN/>있었어.
	두 억양 단위가 휴지 없이 이어질 경우	&	한국사람<CONTINUED/>그 다음
	한 어절 발화 도중 다른 억양 단위로 전사될 때	=	학교<SWITCHING/>를
겹침 현상	겹침 현상	표시하지 않음 (ELAN)	
잘 들리지 않는 부분	잘 들리지 않는 부분	<X X>	<UNCERTAIN>거기서</UNCERTAIN>
	전혀 들리지 않는 부분	<note>안 들림</note>	<ABSOLUTELY UNCERTAIN/>
	들리지 않는 음절 수만큼	X	지금 <UNCERTAIN COUNT="2">XX</UNCERTAIN COUNT="2">에서 삽니다.
발음 오류	필수적인 음운 규칙을 지키지 않아 발음이 이상한 경우	( )	<PRONUNCIATION_ERROR> <ORIGINAL>설날래</ORIGINAL> <CORRET>설날에</CORRECT> </PRONUNCIATION_ERROR>
전사자의 설명	전사자의 설명	기호 없이 메모	<COMMENT>학생 모두가 ""네""라고 한다.</COMMENT>
준음성	웃음, 기침, 하품, 재채기, 박수 등과 같은 언어	<vocal desc='웃음'>	<VOCAL DESC="LAUGH"/>

대분류	소분류	현행 기호	수정 기호
	외적 소리		
기타 (각각의 이름)	구어적 변이형, 외국인 특유의 발음, 발음 오류	(    )	<VARIATION> <ORIGINAL>너</ORIGINAL> <CORRET>넣어</CORRECT> </VARIATION>
	끊어진 단어(불완전하 게 발화된 단어)	=	<TRUNC_WORD>이약</TRUNC _WORD>    나누면서
	방언형 표시	<dia></dia>	차는    <DIA>여일</DIA> 있어.
	숨    표시 (pause)	긴휴지{    }	<PAUSE    DUR="6.3"/>
		짧은휴지..	활발한    편인<PAUSE/>다고
	인용	<Q    Q>	<QUOTATION>우물    안 개구리</QUOTATION>라는
	책이나 자료를 보고 읽은 경우	<R    R>	<REFERENCE>좋은    글이 될 수 있다</REFERENCE>라고 되어 있어요.
	익명성 보장을 위한 마크업 (privacy)	<name>:사람이름 ,단체이름등	<PRIVACY> <NAME>제임스</NAME> </PRIVACY> 입니다.
		<id-num>:주민등 록번호,학번등개 인식별번호	<PRIVACY> <ID_NUM>900120</ID_NUM> </PRIVACY>
		<card-num> : 신용카드 번호	<PRIVACY> <CARD_NUM>1234567890</CAR D_NUM> </PRIVACY>
		<address> : 주소	<PRIVACY> <ADDRESS>서울 서대문구 신촌동</ADDRESS> </PRIVACY>

대분류	소분류	현행 기호	수정 기호
		<tel-num> : 전화번호	<PRIVACY> <TEL_NUM>01012345678</TEL_NUM> </PRIVACY>

## 5. 마크업 지침

### 5.1. 주요 쟁점

- 마크업은 컴퓨터가 텍스트 이외의 서식에 관한 정보를 컴퓨터가 인식할 수 있도록 만든 것으로 말뭉치의 경우 주로 다음 사항을 표기하게 된다.
  - 헤더 마크업: 파일 정보, 발화자 정보, 입력 및 전사 관련 기록 등
  - 본문 마크업: 발화자 표시, 문장 및 문단 구문 등의 텍스트의 구조 태그, 입력과 전사 시 표시되는 텍스트 정보 관련 태그
- 이때 기술적으로는 마크업 언어의 선택, 내용적으로는 마크업이 될 항목에 대한 것이 주요한 쟁점이 된다. 한국어 학습자 말뭉치의 경우 구축에서 보급까지의 전 과정이 웹 기반의 온라인 시스템을 통해 이루어지도록 설계되고 있으며, 마크업 언어와 표기 체계는 필요에 따라 다양하게 출력 가능하다<sup>4)</sup>.
- 한편, 헤더 마크업의 경우 자료 활용의 측면에서 어떠한 정보들을 담는가가 매우 중요한데, 선행 연구에서 제안한 항목들 수집하여 검토한 후 IRB 규정에 위배되지 않으며 자료 활용에 필요한 발화자 정보와 발화 자료 정보를 중심으로 구성하였다.

4) 헤더와 본문 마크업을 위해서는 마크업 언어를 선택하게 되는데, 21세기 세종 한국어 균형 말뭉치의 경우 국제표준화기구(ISO)가 표준화하여 보편적으로 사용되어 온 SGML(Standard Generalized Markup Language) 기반의 TEI(The Text Encoding Initiative) 마크업을 채택하고 있다. 한편, 최근에는 SGML의 장점을 최대한 수용하여 표준화 작업이 이루어진 언어로 문서의 내용에 관련된 태그를 사용자가 직접 정의할 수 있으며 호환성이 뛰어난 XML을 선호한다. 한국어 학습자 말뭉치의 경우 향후 XML 언어를 마크업 언어로 채택하는 방향으로 계획하고 있다.

## 5.2. 지침 수립의 기본 방향

### 1) 헤더 마크업을 위한 정보 구성

- 학습자 말뭉치의 헤더 정보 항목은 크게 학습자 변인, 학습 변인, 환경 변인, 과제 변인, 구어 자료의 경우 대화 상대자 변인, 발화 맥락 변인 등 다양한 범주의 항목이 구성될 수 있다.
- 본 연구에서는 다음과 같이 개인을 식별할 수 있는 정보나 사생활과 관련된 항목들을 제외하고, 언어 습득과 발달과 관련된 주요 변인과 언어 사용에 크게 영향을 미치는 과제 활동 관련 변인을 중심으로 헤더 정보를 구성하였다.

<표 33> 한국어 학습자 말뭉치의 헤더 정보 항목

◎: 국내 수집 시 적용 항목 ○: 이주민, 국외 자료 수집 시 추가 적용 항목

구분	세부 항목		조철현 외(2002)	고려 학습자 말뭉치	한국어 학습자 말뭉치	비고
학습자 변인	성별		○	○	◎	
	나이		○	○	◎	
	직업		○	○	◎	
	국적		○	○	◎	
	학습자 모국어		○	○	◎	
	학습자 등급	수준(초중고)	○	○	○	국외 학습자(국 내외 등급 체계가 다르므로 필요함)
		등급(1-6)	○	○	◎	
학습 변인	학교	기관	○	○	◎	
		개인	○		○	온라인 구축 시스템을 통한

구분	세부 항목		조철현 외(2002)	고려 학습자 말뭉치	한국어 학습자 말뭉치	비고
						자율적 자료 제공 학습자
	학습자 외국어	언어		○	◎	
		숙달도		○	○	
	한국어 학습 경험	기간	○	○	◎	
		기관		○	○	
	학습 목적			○	◎	
환경 변인	한국 방문 경험					
	한국 거주 기간				◎	
	한국어 사용 공동체 생활 경험				○	이주 여성
	거주 시기			○		
	거주지					이주 여성 (방언 사용 가능성)
	한국어 대화 상대자 유무				○	국외 학습자 (교포)
	가족과의 사용 언어				○	국외 학습자 (교포)
과제 변인	과제 유형		○	○	◎	
	과제 장르			○	◎	
	과제 활동 주제		○		◎	
	과제 환경				◎	
발화 상대자	화청자 관계				○	자연 발화 자료 수집 시 필요
	친소관계				○	

구분	세부 항목	조철현 외(2002)	고려 학습자 말뭉치	한국어 학습자 말뭉치	비고
변인	권력관계			○	정보
발화	발화 상황			○	
맥락	발화 장소			○	
변인	발화 주제			○	

- 다음은 위의 정보를 기준으로 2015년에 구축된 말뭉치의 헤더 정보를 출력한 예시이다.

```

<?xml version="1.0" encoding="UTF-8"?>
- <Korean_Learners_Corpus>
  - <Header>
    - <file_info>
      <id>1779</id>
      <filename>일반_문어_대만_경희대_1급_0006</filename>
      <agreementFilename>Ag_일반_문어_대만_경희대_1급_0006</agreementFilename>
      <collectTime>2015 여름 학기</collectTime>
      <assignmentType>기획_작문</assignmentType>
      <assignmentGenre>생활문</assignmentGenre>
      <assignmentTheme>2015년에 하고 싶은 것</assignmentTheme>
      <sentenceCount>11</sentenceCount>
      <wordCount>42</wordCount>
      <inputDate>9.21</inputDate>
      <person_input>이지영</person_input>
      <checkDate> </checkDate>
      <person_check>김태환</person_check>
      <source_type> </source_type>
      <learn_env> </learn_env>
      <file_type> </file_type>
    </file_info>
    - <learner_info>
      <gender>여</gender>
      <age>22</age>
      <current_grade>초급1</current_grade>
      <nationality>대만</nationality>
      <motherTongue>중국어</motherTongue>
      <learningPeriod>6</learningPeriod>
      <residencePeriod>1</residencePeriod>
      <learningGoal>취미</learningGoal>
      <job>학생</job>
      <currentAcademy>경희대학교</currentAcademy>
      <otherlang>영어</otherlang>
    </learner_info>
  </Header>

```

<그림 9> 헤더 마크업 예시

## 2) 본문 마크업

- 본문 마크업은 과거 작업자가 수작업으로 하나하나 부착했던 것과 달리 구축 도구에서 직접 제어, 출력이 가능하다.

## 6. 형태 주석 지침

### 6.1. 주요 쟁점

- 학습자 말뭉치는 정규적인 텍스트가 아니다. 즉 일반적인 말뭉치에 비해 띄어쓰기나 맞춤법 등의 오류가 상당수 포함될 뿐 아니라 오류의 유형도 일반적인 모어 화자와 달리 불규칙하고 산발적으로 나타나기 때문에 형태 정보를 주석하는 과정에서 고려해야 할 사항이 많다. 특히 학습자 말뭉치는 형태 정보의 주석에서 머무르는 것이 아니라 이후 오류 주석 과정을 다시 거쳐야 하는 만큼 형태 정보의 주석 과정에서 이후 과정을 미리 염두에 둘 필요가 있다. 이번 사업에서는 학습자 말뭉치 구축을 본격적으로 구축하기 위한 기초 사업으로서 약 20여 만 어절 규모의 말뭉치에 대해 형태 정보를 주석하였다. 이 과정에서 드러난 몇 가지 문제를 간략히 보이면 다음과 같다.

- 자동 형태 분석 성공률이 매우 낮음

일반적인 텍스트의 경우 95%를 상회하는 높은 성공률을 보이는 분석 도구를 사용했음에도 불구하고 학습자 말뭉치의 경우에는 70% 이하의 성공률을 보인다.

- 오류 유형의 비정규성

학습자의 다양한 변인들, 즉 학습자의 등급, 학습자의 모어 등 다양한 이유로 오류 유형이 일관적이지 않고 매우 불규칙하다. 이는 결국 형태 분석 과정에 수작업 비중이 높을 수밖에 없음을 보여주는 것이다.

- 숙련된 분석 전문가의 필요성

그나마 학습자 말뭉치의 텍스트 규모가 크지 않다는 점은 다행스러운 일이다. 숙련된 소수의 분석 전문가가 일관성을 확보하면서 주석 작업을 수행한다면 본 사업에서 궁극적으로 목표로 삼고 있는 300만 어절의 형태 분석이 불가능하지만은 않으리라 본다.

- 형태적인 차원을 넘어서는 오류 분석

형태적인 차원에서는 올바른 분석이라 하더라도 차원을 확장하면 오류인 경우, 오류 분석 과정에서 이를 정밀히 포착하기 위한 방법이 요구된다. 이를테면 ‘먹어 싶다’와 같은 경우는 ‘먹어’와 ‘싶다’라는 어절 차원에서는 형태적으로 문제가 없으나 오류 분석에서는 이에 대한 오류 정보가 주석되어야 할 것이다.

## 6.2. 지침 수립의 기본 방향

### 1) 형태 주석

#### (1) 형태 주석 작업의 문제와 기본 공정

- 형태 분석의 경우 기술적으로 필연적으로 오류를 포함할 수밖에 없는 학습자 언어 자료를 대상으로 형태 분석을 하여 품사 주석을 부착하는 것은 결코 간단한 문제가 아니다. 한국어 모어 화자 자료의 경우도 지능형태소 분석 도구를 이용하여 형태 분석을 할 경우 90% 이상의 정확성을 보임에도 불구하고 분석이 불가하거나 오분석된 자료의 후처리 작업이 반드시 필요한데, 한국어의 음운, 어휘, 구문 체계에 맞지 않는 다양한 형식을 포함한 학습자 언어의 경우 분석의 정확성이 떨어질 수 있다. 따라서 1) 분석 도구의 활용한 기계적 분석 2) 기술적으로 처리 가능한 후처리 작업 3) 오류 주석과 연계한 최종 수정으로 단계화하여 형태 주석 작업을 한다.

## (2) 분석 도구의 활용과 특성

○ 분석 도구: KMAT의 개요와 특성

○ 특성

- 21세기 세종계획의 형태소 분석 표지를 기본적으로 사용
- 일부 표지에 대한 수정이 포함되어 있음
  - XR(어근) 표지 제거
  - 긍정지정사와 부정지정사 중 긍정지정사만 지정사로, 부정지정사는 형용사로 분석
- 부사격조사와 공동격조사의 통합

## (3) 형태소 분석 표지

<표 34> 형태소 분석 표지

대분류	표지 내용	한국어 학습자 말뭉치의 표지	세종 표지
(1) 체언	일반명사	NNG	NNG/XR
	고유명사	NNP	NNP
	의존명사	NNB	NNB
	대명사	NP	NP
	수사	NR	NR
(2) 용언	동사	VV	VV
	형용사	VA	VA
	보조용언	VX	VX
	지정사	VCP	VCP/VCN
(3) 수식언	관형사	MM	MM
	일반부사	MAG	MAG
	접속부사	MAJ	MAJ
(4) 독립언	감탄사	IC	IC
(5) 관계언	주격조사	JKS	JKS

대분류	표지 내용	한국어 학습자 말뭉치의 표지	세종 표지
	보격조사	JKC	JKC
	관형격조사	JKG	JKG
	목적격조사	JKO	JKO
	부사격조사	JKB	JKB/JC
	호격조사	JKV	JKV
	인용격조사	JKQ	JKQ
	보조사	JX	JX
(6) 의존형 태	선어말어미	EP	EP
	어말어미(연결)	EC	EC
	어말어미(종결)	EF	EF
	명사형전성어미	ETN	ETN
	관형형전성어미	ETM	ETM
	명사파생접두사	XPN	XPN
	명사파생접미사	XSN	XSN
	동사파생접미사	XSV	XSV
	형용사파생접미사	XSA	XSA
(7) 기호	마침표, 물음표, 느낌표	SF	SF
	쉼표, 가운데점, 콜론, 빗금, 줄표, 물결	SP	SP
	따옴표, 괄호표	SS	SS
	줄임표	SE	SE
	불임표(숨김, 빠짐)	SO	SO
	외국어	SL	SL
	한자	SH	SH
	기타 기호	SW	SW
	숫자	SN	SN
	분석불능범주	NA	NA

#### (4) 자동 분석 결과 중 분석 불능(NA) 및 오분석 자료 처리

- 분석 불능(NA) 및 오분석 자료는 아래의 예와 같이 형태상 실재하지 않는 형태소라고 하더라도 한국어의 통사 배열 규칙에 따라 그 위치와 기능을 기준으로 품사 주석을 수정하였다. 이는 오류 분석 결과 활용의 측면에서 교정 어절과의 일대일 대응을 통해 특정 품사와 형태소의 오류 양상을 손쉽게 살필 수 있도록 하기 위한 것이다.

##### ① 작문 자료 예시

주말에 시장에 갔어요.  
 저는 시장에서 과일을 샀어요.  
사과과 바나나와 배와 수박을 샀어요.  
 다음에 저는 커피숍에 갔어요.  
 커피숍에서 우리 남편도 왔어요.  
 우리 남편은 커피를 마셔지만 저는 케이크를 먹었어요.  
 아주 좋아요!  
 다음에 우리는 극장에 갔어요.  
 날씨가 춥니까 택시를 탔어요.  
 극장에서 우리는 친구 두 명(~~조사 누락~~→을) 만났어요.  
 남자친구 한 명 그리고 여자친구 한 명.  
 영화는 아주 좋았어요!  
 다음에 우리는도 같이 식당을 갔어요.  
 저는 불고기를 먹고 우리 친구가 냉면을 먹었어요.  
 아이스크림하고 맥주 네 병도 맛있어요.  
 우리(~~조사 누락~~→의) 밤은 아주 좋았어요!  
 길이 복잡하니까 지하철을 탔어요.  
 우리는 사당역에서 내렸어요.

##### ② 분석 불능 및 오분석 예시

<표 35> 형태소 분석 결과 중 분석 불능 및 오분석 예시

교정하지 않은 원자료를 분석한 경우		
사과과	사과과/NA	수정필요

남편은	남편/NA+은/JX	수정필요
마셔지만	마시/VV+어/EC+지/VX+만/EC	NULL
출니까	출/VV+니까/EC	NULL
좋았어요!	c/SL+이/VCP+있/EP+어요/EF+!/SF	NULL
우리는다	우리는다/NA	수정필요
맛있어요.	맛있어/NF+이/VCP+오/EF+./SF	수정필요
우리	우리/NP	NULL

### ③ 처리 방안

<표 36> 오류가 발생한 부분의 형태소 분석 결과와 처리 방안

유형	원어절	형태소 분석 결과	특징 및 후처리 방안
오류이기는 하지만 형태소 분석 결과 자체는 정상적으로 분석된 경우	우리	우리/NP	<ul style="list-style-type: none"> <li>○ 대체로 표기가 정확한 상태에서의 생략 오류, 대치 오류에서 주로 발생</li> <li>○ 형태 분석 결과는 수정하지 않음</li> <li>○ 오류 주석 단계에서 오류 판정을 하기 위한 교정 어절만 생성하면 됨</li> </ul>
	출니까	출/VV+니까/EC	
분석 불능	사과과	사과과/NA,	<ul style="list-style-type: none"> <li>○ 이형태오류, 철차 오류, 첨가 오류 등 일반적인 형태에서 벗어난 경우에 주로 발생</li> <li>○ 원칙적으로 형태 분석 결과의 수정, 오류 주석 단계에서 오류 판정을 하기 위한 교정 어절만 생성이 필요함</li> </ul>
	남편은	남편/NF+은/JX	
	우리는다	우리는다/NA	
	맛있어요.	맛있어/NF+이/VCP+오/EF+./SF	
오분석	좋았어요!	c/SL+이/VCP+있/EP+어요/EF+!/SF	○ 분석 불능과 동일

#### ④ 수정 예시

- 다음과 같이 한국어의 통사 배열 규칙에 따라 그 위치와 기능을 기준으로 품사 주석을 수정한다.

<표 37> 형태소 분석 결과 중 분석 불능 및 오분석 수정 예시

원어절	수정 전	수정 후
사과과	사과과/NA	사과/NNG+과/JKB
남편은	남편/NF+은/JX	남편/NNG+은/JX
마شى지만	마시/VV+어/EC+지/VX+만/EC	마시/VV+어/EC+지만/EC
츨니까	츨/VA+니까/EC	츨/VA+니까/EC
츨었어요!	c/SL+이/VCP+였/EP+어요/EF+!/SF	츨/VA+였/EP+어요/EF+!/SF
우리는도	우리는도/NNA	우리/NP+는/JK+도/JK
맛었어요.	맛았어/NF+이/VCP+오/EF+./SF	맛/VV+였/EP+어요/EF+./SF
우리	우리/NP	우리/NP

## 2) T 단위 주석

### (1) T 단위의 개념과 효용성

- T-unit(minimal terminal unit)라는 개념은 L1 문장성숙도(syntactic maturity)에 대한 연구인 힌트(Hunt, 1965)에서 처음 제시되었다. T-unit은 다음의 예시와 같이 종속절 빈도(subordination ratio)라는 기준에 더해 절 내부의 구조까지를 분석하였다.

There are many different contributions between artists and scientists to society. || First artists contribute to society for entertainment. || Many people need it for relax after hard work. || Artists contribute to society as film artists, singers and so on. || Furthermore artists contribute to society with make new-work fields which are related with kind of activity. || (Schneider and Connor 1990: 415)

T-unit 코딩의 예시(Foster et al., 2000:362에서 재인용)

- T-unit 개념은 이후 L1와 L2 구어 연구에서도 활발히 차용되었는데 영 (Young, 1995:19)은 그 이유로 첫째, T-unit은 절을 기반으로 하고 있으므로 문장이라는 개념에 비해 구어 연구에서 적용하기 쉽다, 둘째, 단어나 말차례(turn) 등의 단위들에 비해 T-unit는 그에 일치되는 지시적 의미를 가지고 있기 때문에 주제를 가진 연속체를 연구하는 데 적용할 수 있다는 점을 들었다. 그러나 하나의 T-unit 단위가 말더듬 등으로 인해 잘못 시작되었을 경우 먼저 발화된 앞부분은 하나의 T-unit 단위로 볼 수 없다고 보았으며 맞장구('음'), 담화 표지('응', '좋아') 등도 T-unit 단위에서 제외했다.<sup>5)</sup>

## (2) T 단위 주석을 위한 분석 단위

- 문어는 기본적인 단위가 문장으로 종결어미로 마무리되며 마침표와 같은 문장부호로 인해 명확하게 그 단위가 구분되어 T 단위를 하나의 통사 단위로 보는 것에 타당성이 있으며 주석 작업도 비교적 간단하다. 이에 반해 구어는 종결어미를 사용하여 발화를 끝내는 경우가 많지 않고 억양이나 휴지 등의 운율적인 요소에 의해 영향을 받는다. 따라서 문장이 아닌 억양단위를 기본 단위로 보는 것이 타당할 수 있으며, 본 제안에서도 그러한 점을 고려하여 억양단위 전사를 할 예정이다.

## (3) T 단위 주석 적용 범위

- 한국어의 문장에서 T 단위는 대등 접속과 문장 종결 위치가 되리라 본다. 대등 접속어미의 형태는 '-지만, -으나, -고' 등에 한정되어 일괄 처리가 가능하기는 하나, 이들 형태는 종속적으로도 사용이 가능하므로 일일이 의미를 변별하지 않고는 형태만으로 대등 접속 여부를 판정하기 어렵다. 따라서 접속에 나타나는 T 단위는 개별 연구자들이 주석을 달 수 있도록 시스템을 제공하는 것으로만 한정하고 (열린 주석 표지 제공), 문장 단위 T 단위만을 주석하기로 한다.

---

5) "The following elements were counted as one T-unit: a single clause, a matrix plus subordinate clause, two or more phrases in apposition, and fragments of clauses produced by ellipsis. Co-ordinate clauses were counted as two t-units. Elements not counted as t-units include back channel cues such as mhm and yeah, and discourse boundary markers such as okay, thanks or good. False starts were integrated into the following t-unit." (Young, 1995:38)

## 6.3. 2016년 형태 주석 지침 및 기구축 자료 보완 방향

- 2015년 구축한 형태 주석 말뭉치는 문어 20만 어절, 구어 2만 어절이다. 지능형 형태 분석 도구를 사용하여 형태 분석을 한 후 수작업으로 1차 검수 작업을 하였다. 그리고 오류 주석 대상이 된 문어 4만 어절, 구어 1만 어절에 한해 2차 검수 작업을 수행하였다.
- 2차 차 검수 작업의 핵심은 향후 보다 실용적으로 자료를 활용하기 위하여 학습자 오류로 인해 분석 불능(NA)된 부분을 처리하는 것이었다. 2016년에는 그러한 내용을 반영하여 형태 주석 지침을 정교화하고, 오류 주석 대상에서 제외되었던 17만 어절에 대한 2차 검수가 필요하다.

## 7. 오류 주석 틀(태그 세트) 및 주석 작업 지침

### 7.1. 주요 쟁점

- 오류 주석은 여타 말뭉치와 비교하여 학습자 말뭉치만이 가지는 가장 변별적인 특성이다. 그러나 오류 판정과 교정, 주석 과정에서 피할 수 없는 분석의 체계성과 일관성, 자의성의 문제가 발생하게 된다. 또한 학자들 간의 이견이 다양한 오류의 범주를 어떻게 체계화하여야 오류 연구에 관한 이론적 토대 위에서 실용적인 정보로서 그 효용 가치를 얻게 될지에 대한 논의가 필요하다. 이러한 문제들은 결국 자료의 활용 목적과 연구 목적, 관점에 따라 달라질 수 있으나, 국가 주도의 공개 자료로서 가장 효과적으로 오류 주석을 부착하기 위해 아래의 쟁점에 대해 지침을 마련한다.

- [형식적 체계] 오류 주석 단위
- [내용적 체계] 오류 주석 틀(태그 세트)의 체계화 및 영역별 쟁점
- [주석 작업의 체계] 오류와 실수의 구분

오류 판정과 교정의 기준

## 7.2. 지침 수립의 기본 방향

### 1) 오류 주석의 단위

- 기존의 오류 주석은 학습자가 산출한 원문을 그대로 유지한 상태에서 어절 단위를 중심으로 이루어져 왔다. 아래의 문장에서의 오류들인 ‘참가’(어휘의 대치 오류), ‘여야 하면’(표현문형의 대치 오류), ‘과’(조사의 대치 오류), ‘포함된’(어휘의 대치 오류)등은 아래와 같은 어절 단위의 주석 방식으로는 오류의 정확한 위치를 포착하기가 어렵다.

예) 앞으로 모든 사람들이 참가해야 하면(√/ 참여하면) 모든 사람과(√/ 사람이) 포함된(√/ 함께하는) 행복한 사회를 만들 수 있다.

- 반면, 형태소 단위를 중심으로 한 주석은 다음과 같이 오류 위치와 형태를 보다 명확하게 파악하여 주석하고 그 결과를 분석할 수 있다는 장점이 있다. 아울러 두 개 이상의 형태소가 결합된 표현문형 오류의 경우, 형태소 단위의 주석은 다양하게 나타나는 이형태와 활용형까지 정확하게 포착할 수 있다. 이러한 점을 고려하여 한국어 학습자 말뭉치는 형태소 단위 중심의 오류 주석 방식을 선택하였다<sup>6)</sup>.

---

6) <그림>의 예시 자료는 온라인 구축 시스템을 본격적으로 사용하기 전에 엑셀을 사용하여 주석을 한 예이다. 엑셀에서는 표현문형과 같이 두 개 이상의 형태소를 하나로 묶기 위해 셀을 병합해야 하는데, 그 경우 자료 처리에 어려움이 생기기 때문에 한 그룹이라는 뜻으로 하나로 묶여야 할 복수의 형태소에 주석을 반복하여 제시하였다. 이것은 온라인 구축 시스템에서 하나로 통합되어 처리가 될 예정이다.

원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태 주석 분석 여부	오류 중위	오류 현상
앞으로	으로/JKB					
모든	모든/MM					
사람들이	사람/NNG					
사람들이	들/XSN					
사람들이	이/JKS					
참가해야하면	참가/NNG		참여해야	참여/NNG	VW(어휘 단어)	REP(대치)
참가해야하면	하/XSV		참여해야	하/XSV		
참가해야하면	아아/EC		참여해야	아아/EC	GF(문법 표현문형)	REP(대치)
참가해야하면	하/VX		참여해야		GF(문법 표현문형)	REP(대치)
참가해야하면	면/EC		참여해야		GF(문법 표현문형)	REP(대치)
모든	모든/MM					
사람과	사람/NNG		사람이	사람/NNG		
사람과	과/JKB		사람이	이/JKS	GPT(문법 조사)	REP(대치)
포함된	포함/NNG		함께하는	함께/NNG	VW(어휘 단어)	REP(대치)
포함된	되/XSV		함께하는	하/XSV	VW(어휘 단어)	REP(대치)
포함된	ㄴ/ETM		함께하는	는/ETM	VW(어휘 단어)	REP(대치)
행복한	행복/NNG					
행복한	하/XSA					
행복한	ㄴ/ETM					
사회를	사회/NNG					
사회를	를/JKO					
만들	만들/VV					
만들	ㄹ/ETM					
수	수/NNB					
있다.	있/VV					
있다.	다/EF					

<그림 10> 형태소 단위 중심의 오류 주석 예시

## 2) 오류 주석 틀의 체계화 및 영역별 쟁점

### (1) 한국어 학습자 말뭉치 오류 주석 체계 및 표지

- 2015년 연구에서는 선행 연구를 바탕으로 한 한국어 학습자 말뭉치 오류 주석 틀과 주석 표지를 마련하여 주석하였다. 그리고 5만 어절의 샘플 주석 작업을 수행한 기초 연구를 통해 몇몇 쟁점을 토대로, 향후 연구에서 적용하게 될 수정안(<표38>의 우측)을 제시한다(지침 수정의 이유는 다음 절 참고). <표38>의 주석 체계는 연구 결과물로 제시하는 최소한의 주석 표지이며, 연구자에 따라 세부 오류 주석의 표지를 확장할 수 있도록 설계한다. 수정안에서 기본 주석이란 모든 오류에 대해 필수적으로 주석되는 것을 의미하며 1:1로 주석됨에 반해, 확장 주석이란 관련 오류가 있는 경우에만 주석되며 한 형태에 2개 이상의 오류가 나타나면 중복 주석이 가능하다.

<표 38> 한국어 학습자 말뭉치의 오류 주석 틀 및 주석 표지

2015년 주석 말뭉치 적용 체계				수정안				
범주	오류 영역		주석 표지	구분	범주	오류 유형		주석 표지
분석 여부	불가능		IMP	기본 주석	분석 여부	불가능 - 전체적 오류 포함		IMP
오류 층위	발음	음절	PS		오류 양상	누락		OM
		음운규칙	PC	첨가		ADD		
	어휘	명사	VNN			대치		REP
		대명사	VNP			오어순		MISO
		수사	VNR		오류 영역	실질어	고유명사	CNNP
		동사	VVV				일반명사	CNNG
		형용사	VVA	의존명사			CNNB	
		지정사	VVC	대명사			CNP	
		관형사	VMM	수사			CNR	
		부사	VMA	동사			CVV	
		감탄사	VIC	형용사			CVA	
	문법	조사	GPT	보조용언			CVX	
		어미	GE	지정사			CVC	
		높임	GH	관형사			CMM	
		시제	GT	일반부사			CMAG	
		사동	GC	접속부사			CMAJ	
		피동	GPS	감탄사			CIC	
		부정	GN	기능어	주격조사	FNP		
		표현문형	GF		관형격조사	FGP		
		문장	GS		목적격조사	FOP		
	담화	지시	DR		부사격조사	FAP		
		접속	DC		호격조사	FVP		
		담화표지	DM		인용격조사	FQP		
		구어/문어	DS		보조사	FXP		
	오류	누락	OM				연결어미	FED

2015년 주식 말뭉치 적용 체계			수정안				
범주	오류 영역	주식 표지	구분	범주	오류 유형		주식 표지
현상	첨가	ADD				종결어미	FFE
	오형태	MISF				선어말어미	FPE
	오어순	MISO				명사형 전성어미	FNE
						관형사형 전성어미	FAE
	대치	REP			구 단위 표현		PHE
			확장 주식	오류 층위	음운	음소[발음/ 철자] <sup>7)</sup>	PP
						음 절[ 발 음/ 철자]	PS
						음운규칙	PC
					형태	단어 형성[합성법]	MCP
						단어 형성[파생법]	MDV
						굴절[곡용]	MDC
						굴절[활용]	MCJ
					통사	높임	SH
						시제	ST
						사동	SC
						피동	SP
						부정	SN
					담화	지시	DR
						접속	DC
						담화표지	DM
						구어/문어 오류	DS

7) 음운 층위에 있는 음소 오류, 음절 오류, 음운규칙 오류 외에도 음성 차원의 변이음 오류가 별도로 설정될 수 있다. 사실상 변이음 오류는 전사 단계에서 음성 차원의 전사를 통

- ① ‘2015년 한국어 학습자 말뭉치 오류 주석 틀’을 사용한 주석 작업의 쟁점
- ‘2015년 한국어 학습자 말뭉치 오류 주석 틀’은 선행 연구에서의 주석 체계에 대한 분석을 토대로 실용적인 체계로 구성이 되었으나 실제 주석 과정에서 다음과 같은 문제점이 있었다.
    - 높임 오류의 경우 조사, 선어말어미, 종결어미, 어휘 등 다양한 방법으로 실현되는데, 각각에 해당하는 주석 영역이 동일한 층위에 있거나 위계에 맞지 않게 배열되어 있음
    - 사동과 피동의 경우 어휘적으로 실현되기도 하고 문법적으로 실현되기도 하는데, 어느 하나를 선택할 경우 사동/피동 오류 또는 어휘 오류 전반을 살피기가 어려움
    - 시제 오류의 경우 관형사형 어미, 선어말어미, 표현문형을 통해 실현되는데, 어느 하나를 선택할 경우 시제 오류 또는 어미 오류, 표현문형 오류 전반을 살피기가 어려움
  - 그 외에도 오류 현상의 오형태 오류가 발음, 어휘, 문법의 모든 영역에서 발생하면서 주석의 체계성을 고민하게 하였고, 발음 오류의 경우 한국어 학습자의 발음 오류 중 가장 빈번하게 발생하는 변이음 오류를 포괄할 수 없었다.
  - 이러한 문제들을 해결하기 위하여 언어학적 층위에 맞게 주석의 틀을 재 배열하고, 한국어 교육 분야에서 중요하게 다루어져야 할 오류 항목들을 추가하는 방식으로 오류 주석 틀을 수정하였다.
- ② ‘한국어 학습자 말뭉치 오류 주석 틀(수정안)’의 특징
- 오류 영역과 오류 층위를 구별
    - ‘2015년 한국어 학습자 말뭉치 오류 주석 틀’은 한국어 교육 연구에서 주요한 관심의 대상이 된 주석 항목들을 고르게 포함하고 있으나, 언어학으로나 작업자의 직관에서 항목 간의 층위가 맞지 않거나 충돌하는 부분이 있었다. 수정안에서는 이러한 부분을 고려하여 다양한 항목들이 이론적, 실용적 측면에서 모두 체계성과 균형성을 갖출 수 있도록 수정하였다.

---

해 학습자의 발음을 정확하게 반영해야 하지만 그러기 위해서는 해당 음을 정확히 식별하고 국제음성기호(IPA)를 사용하여 표기해야 하는 어려움이 있다. 따라서 본 연구에서는 2016년 기구축 자료의 보완 단계에서 그러한 위치를 판단할 수 있도록 약식 표기를 남기는 절충안을 고려하고 있다.

- 수정 전의 오류 주석 체계에서 가장 문제가 된 주석 항목 간의 충돌, 작업자 간의 이견과 혼란은 오류 범주와 각 범주에 속한 하위의 오류 영역들이 언어학적 체계보다는 교육 현장에서 실용적으로 재편한 체계에 따라 구성되었기 때문이라고 볼 수 있다. 이는 오류 분석의 관점에서 오류의 위치와 형태, 오류의 원인을 명확하게 구분하지 못하고 혼재함으로 해서 발생한 문제이기도 하다. 예를 들어, ‘한국에 온 비행기 안에서 남자친구를 만났다.’라는 문장에서 오류가 발생한 위치와 형태를 기준으로 오류를 판정하면 ‘관형사형 어미의 오류’이지만 오류가 발생한 원인을 기준으로 판정하면 ‘시제 오류’가 될 수 있다. 즉, 상대시제의 용법을 정확하게 이해하지 못해 관형사형 어미의 형태를 사용한 것이다.
- 이에 본 연구에서는 이러한 문제점을 보완하기 위하여, [오류 층위]에 있던 [어휘] 영역으로 위치를 이동하였다. 아울러 오류 층위의 [조사, 어미]를 오류 영역으로 이동하였다. 오류 영역은 형태 주석 결과를 참고하여 별도의 층위에서 오류가 발생한 위치에 오류가 발생한 항목의 위치와 형태 정보를 주석하는 것을 의미한다.

#### ○ 오류 현상의 재조정

- [오형태]는 주로 발음이나 규범 표기에 대한 지식이 부족하여 발생하게 되는데, 학습자 오류 중 가장 빈번하게 발생하는 오류 유형 중 하나로 보아 [확장 주석] 범주의 [음운] 영역에서 주석할 수 있도록 하였다. 한편, [오어순]은 [대치], [누락], [첨가]와 다소 층위는 다르지만 구 또는 문장 단위 이상에서 두 개 이상의 형태소가 관여하는 오류로 남겨 두었다.

#### ○ 개방형 주석 체계 제공

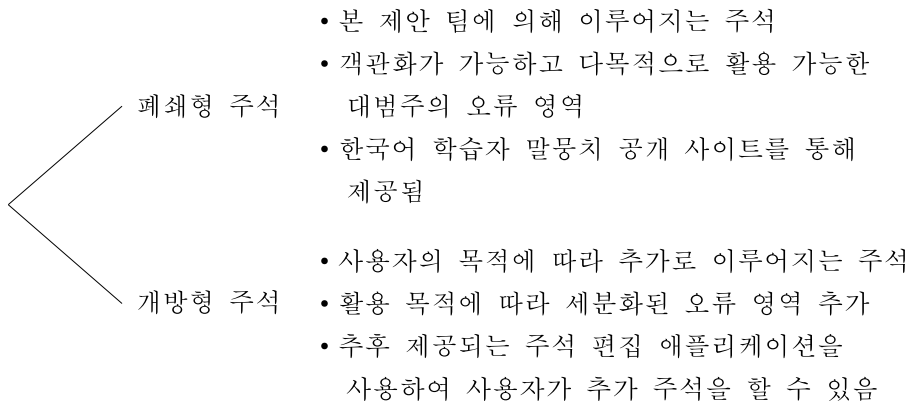
- 한국어 학습자의 언어 사용을 주제로 한 선행 연구 분석을 통해 본 연구를 통해 제공되는 주석과 사용자가 확장해서 사용 가능한 개방형 주석 체계를 제안하였다<sup>8)</sup>. 오류 범주는 체계화의 방식도 다양하지만, 하나의 범주

8) 다음은 오류분석, 중간언어, 언어 습득 연구 등 학습자 언어를 대상으로 한 연구에서 많이 다루어진 주제들로 본 연구에서 한국어 학습자 말뭉치 오류 주석 범주 체계를 정리하는 데에 기초 자료가 되었다.

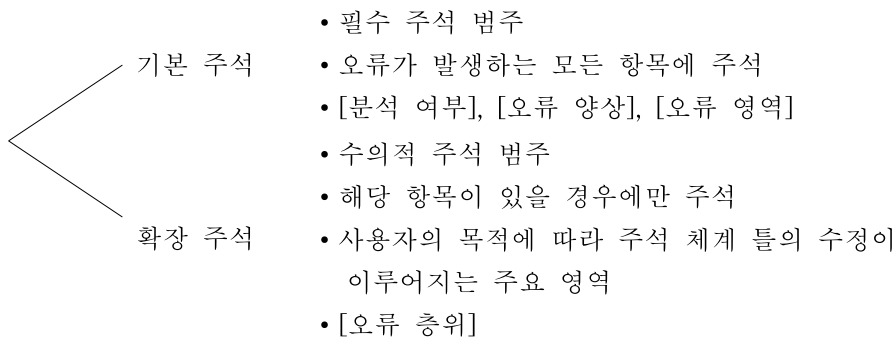
영역 구분		세부 주제
어문 규범		맞춤법, 띄어쓰기 오류
발음	음절	음소 차원 또는 음절 차원에서 발생하는 오류

안에 속하는 하위 범주들도 많아 사용자 모두가 합의할 수 있는 방식으로 체계화하기는 어렵다. 또한 지나치게 다양하고 세분화된 범주를 채택하였을 때에 오히려 활용도를 떨어뜨리게 될 우려가 있다. 본 연구에서는 이러한 점을 고려하여 커다란 체계 내에서 보편적인 범주들로 오류 주석 표지를 체계화하여 제시하고, 사용자가 목적에 따라 주석을 가감하여 추가 주석을 할 수 있도록 열린 주석 체계를 마련하였다.

		예) 평음, 격음, 경음의 구분
	음운규칙	연음규칙, 비음화 등 음운변동에 관한 오류
어휘	단일어	개별 어휘의 형태, 의미, 사용의 차원에서 발생하는 오류
	파생어	접사의 과잉 적용과 같은 파생어 사용 오류
	합성어	합성어 사용 오류
	언어 관계	언어 사용 오류
문법	조사	격조사, 보조사, 접속조사 사용 오류
	어미	연결어미, 선어말어미, 종결어미, 전성어미 사용 오류
	높임	높임법 오류
	시제	시제 및 시상 오류
	사동	사동사, 사동 표현, 사동문 오류
	피동	피동사, 피동 표현, 피동문 오류
	부정	부정 표현 및 부정문 오류
	표현문형	문법적 언어 오류
	문장	문장 성분, 호응 관계 오류
	문법	높임법, 연결어미
담화	자기 수정	-
	주제와 초점	조사 '이/가와 '은/는' 연구
	담화 구조	응집성, 설명 담화, 학술 텍스트
	지시	담화 차원에서의 지시사 사용 오류
	접속	접속 부사 및 접속 표지 오류
	담화표지	문어와 구어의 담화표지 사용 오류



■ **주석 범주를 [기본 주석]과 [확장 주석]으로의 이원화:** 기존의 오류 범주를 아래와 같이 두 개로 나누고 위치를 이동함으로써 오류 범주 간의 중복과 충돌을 피하고 오류 주석 작업의 효율성을 제고할 수 있다.



#### 한국어 학습자 말뭉치 오류 주석 체계 틀의 경쟁력

- **언어학적인 체계성과 일관성, 다양한 목적의 사용자를 고려한 범용성:** 조철현(2002)를 비롯한 선행 연구에서의 주석 체계 틀은 대개 한국어교육 현장에서 실제로 적용되어 온 교육과정이 잘 반영되어 있다. 교육과정을 토대로 한 귀납적 분석 결과에 의해 주석 체계가 만들어졌기 때문에 실용적인 것이 장점이라고 할 수 있다. 그러나 자료에서 발견되지 않았거나 분석 대상에서 제외된 구어 오류의 경우 주

석 체계 내에서 빈칸으로 남아 있고, 오류 주석 항목 간의 체계성이 다소 부족하다는 아쉬움이 있다. 이와 달리 한국어 학습자 말뭉치는 문어와 구어를 모두 포괄하여 언어학적으로 체계적이면서도 각기 다른 목적으로 자료를 활용하게 될 연구자, 교사, 학습자들에게 유용한 주석 항목을 포함하였다.

- **개방형 주석 체계에 의한 확장성과 유연성:** 오류 주석은 목적에 따라 보다 다양한 층위에서 세분화할 수 있다. 한국어 학습자 말뭉치는 사용자가 필요에 따라 주석을 가감하여 확장할 수 있도록 개방형 주석 체계를 채택하였다. 이는 국외의 말뭉치에서도 찾아보기 어려운 새로운 시도 중 하나이다.
- **다층주석을 통한 주석의 체계성과 정보의 유용성:** 한국어 학습자 말뭉치는 다음과 같이 두 가지 측면에서 다층 주석 체계를 채택하였다.

- 주석 체계상[이론적]: 분석 여부, 오류 영역, 오류 현상, 오류 층위
- 언어학적 층위와 주석 처리의 방식[기술적]: 형태소, 구 단위, (문장 단위는 추후 고려)

국외의 말뭉치 중 오류 주석 체계가 체계적으로 문서화되어 공개되고 있는 CLC, the FreeText project, ICLE, NICT의 경우 평균 50개 내외, 최대 100개에 달하는 주석 표지를 가지고 있지만 이들은 주로 품사별 오류를 주석하는 데에 초점을 두고 있으며, 그 밖의 발음, 어휘, 문법, 담화, 화용 오류를 다루기 위한 표지는 찾아보기가 힘들다. 이에 비해 한국어 학습자 말뭉치는 다층 주석을 통해 학습자의 언어 관찰에 유용한 정보들을 고루 포함하고 있다.

<표 39> 국외 학습자 말뭉치의 오류 주석 체계

구분	오류 주석 체계 틀
CLC	<ul style="list-style-type: none"> <li>○ 수정 분류체계: 오형태, 누락, 대치가 필요한 단어나 구, 불필요한 단어나 구, 잘못 파생된 단어</li> <li>○ 기타: 구두법, 가산, (동형어와 다른) 유사 형태-다른</li> </ul>

	<p>의미의 어휘 오류, 일치, 그 밖의 다른 오류들(논항 구조, 부정확한 굴절, 철자 오류, 부정형의 부정확한 형태 등)</p> <p>○ 품사 중심의 언어학적 분류 차원 대명사, 접속사, 관형사, 형용사, 명사, 양화사, 전치사, 동사, 부사</p>
the FreeText project	<p>○ 언어학적 분석 층위: 형식, 형태, 문법, 어휘, 통사, 사용역, 문체, 구두법, 오자</p> <p>○ 오류 범주: 동음이의어, 합성어, 태, 언어 관계(prefab), 어순 등</p> <p>○ 품사 범주와 하위범주: 형용사-비교급 형용사, 관사-부정관사, 명사-고유 명사</p>
ICLE	<p>○ 언어학적 범주: 형태, 구두법, 문법, 어휘-문법, 사용역, 잉여적 단어/단어 누락/어순 그리고 문체</p> <p>○ 하위범주: 품사 정보, 오류 유형(철자, 시제, 비교급/최상급, 셀 수 있는/셀 수 없는 등)</p>
NICT	<p>○ 주요 범주 또는 품사 범주: 명사, 조동사, 형용사, 부사, 전치사, 관사, 대명사, 접속사, 관계대명사, 의문사 외</p> <p>○ 오류 범주: noun case, verb lexis, number of adjective, adverb inflection, complement of preposition 등</p>

## (2) 한국어 학습자 말뭉치 오류 범주 체계 개발을 위한 영역별 쟁점

- 오류 범주에 대한 연구자들의 견해가 상이한 만큼 영역별로 다양한 오류 범주를 분류하는 방식이 다양하다. 따라서 보다 효용성 있는 오류 주석 작업을 위해서는 영역별 오류 범주와 관련된 쟁점에 대한 면밀한 검토가 필요하다<sup>9)</sup>.

9) 오류 범주를 체계화하는 일은 그다지 간단하지 않다. 오류 범주를 체계화하는 것이 이론적으로는 의의가 있지만, 방대하고도 다양한 오류 현상을 포함하고 있는 학습자 자료를 대상으로 한 분석에서는 오히려 연역적인 방법에 의한 체계화보다는 귀납적으로 필요한

### ① 음운

- 음운 층위는 주로 발음 오류를 다루는 영역이다. 구어 의사소통 능력이 중시되면서 주로 대조언어학적 관점에서 학습자의 모국어와 목표 언어 간의 음운 체계 차이로 인한 오류의 규명이 지속적으로 이루어져 왔다. 이에 비해 대규모의 학습자 자료를 기반으로 하여 음소, 음절, 어절, 문장, 담화 등의 층위에서 다양하게 나타날 수 있는 발음 오류를 체계화하거나 전반적으로 다룬 연구는 거의 없다. 이는 그간 이루어진 오류 연구의 경우 주로 문어를 중심으로 하였기 때문이라고 볼 수 있다. 이정희(2002)의 경우 발음의 영향으로 나타난 맞춤법 오류를 초성 오류, 중성 오류, 종성 오류로 구분하여 발음 오류로 다루고 있으나, 문어 자료를 대상으로 하였기 때문에 실제 발음 오류의 특성을 모두 담아내기는 어려웠다.
- 본 연구에서는 발음 오류를 음소[발음/철자], 음절[발음/철자], 음운규칙 오류로 나누어 발음 오류로 인해 구어와 문어에 두루 나타나는 변이음 오류와 음절 층위의 오류를 주석하도록 하였다. 그 외에 강세, 억양 등과 같이 어휘, 구, 문장, 담화 층위에 걸쳐 의사소통에 영향을 주는 오류 표지는 사용자가 원문 음성 파일을 들으면서 추가 주석하여 사용할 수 있다.

### ② 형태

- 형태 층위는 주로 어휘 오류를 다루는 영역이다. 수정된 주석 체계에서는 기본 주석 영역에 [오류 영역]을 두어 오류가 일어난 부분의 품사를 주석하도록 하였다. 형태 층위에서는 그 외에 합성어, 파생어 등의 조어 과정에서 발생하는 오류와 어미의 활용, 조사의 사용에서 나타나는 오류를 주석하도록 하였다. 이는 선행 연구에서 단어 형성과 관련된 오류를 어휘 오류로 처리하고, 어미의 활용이나 조사 사용에 관한 오류를 문법 오류로 처리하였던 것과 달리 형태 층위에 통합하여 언어학적인 체계성을 고려하였다는 점에서 차별화된다.

### ③ 통사

- 통사 층위는 문법 오류를 다루는 영역이다. 높임, 시제, 사동, 피동, 부정 등의 문법 범주에서 발생하는 오류가 그 대상이 된다.

---

항목을 추가해 나가는 것이 더 실용적일지도 모른다. 국외의 오류 주석 체계는 일면 그러한 특성을 반영하고 있는 것으로 보이기도 한다. 언어학적으로 혹은 기존의 이론에 기반하여 오류 범주를 체계화하기보다는 다양한 층위의 항목들이 혼재되어 있는 모습을 보인다.

#### ④ 담화

- 담화 층위는 문장 단위를 넘어서 발생하는 오류를 다루는 영역이다. 최근 담화 능력에 대한 관심이 높아지면서 담화 연구가 꾸준히 증가하고 있다. 그러나 담화 연구의 경우 그 범위가 넓고 어휘와 문법, 발음 영역에서 다양한 현상과 표지를 통해 나타나므로 체계화가 쉽지는 않다. 또한 구어 담화의 경우 문법정보다는 발화 맥락 안에서 적절하고 효과적인 의미 전달에 초점이 주어지기 때문에 오류 판정 기준을 정하기도 쉽지 않다. 이러한 이유로 본 연구에서는 담화표지, 지시, 접속으로 비교적 표지가 분명하고 판정 기준이 명확한 항목을 오류 표지 체계 안에 포함시켰다.

### (3) 개방형 주석 체계 틀의 확장 가능성 및 예시

- 본 연구에서는 개방형 주석 체계를 통해, 공개용으로 제공되는 오류 주석 외에 연구자가 연구 목적에 따라 기존의 주석 영역을 세분화하거나, 새로운 주석 영역을 추가하고, 불필요한 주석 영역을 삭제하는 방식으로 오류 주석 체계를 수정, 보완할 수 있도록 설계하였다. 즉, 사용자는 기본 주석을 활용하여 필요한 주석 영역을 생성하여 확장할 수도 있으며, 오류 층위별로 연구자가 필요한 개방형 주석을 생성하여 확장할 수도 있다. 주석 표지는 사용자의 편의에 따라 다양한 기호 체계를 사용하여 만들 수 있다. 이러한 추가적인 주석 작업의 편의성을 위하여 향후 자료 배포 단계에서 주석 편집 도구를 함께 제공할 계획이다.

#### ① 기본 주석을 활용한 개방형 주석

- 현 오류 주석 체계에서는 오류의 양상과 오류 영역에 대한 기본 주석을 필수적으로 제시하고 있다. 이러한 오류 주석 체계를 활용하여 다음과 같은 개방형 주석이 가능하다.

##### 가. 연어 오류 주석

- 기존 한국어 학습자 말뭉치 연구에서 많은 연구 대상이 되어 온 것 중의 하나는 연어 오류이다. 이러한 연어 오류는 일반적으로 연어 관계에 있는 둘 이상의 어휘 중 하나를 다른 어휘로 대체하여 나타나는 경우가 일반적이다. 따라서 실질어(명사, 동사, 형용사, 관형사, 부사)와 이의 대치 오류를 통해 ‘명사+용언, 부사+용언, 관형사+명사, 명사+명사’ 등 다양한 유형

의 언어 오류를 검색하고 이를 주석할 수 있다.

예) 저는 한국에 와서 새 친구를 만들었습니다(사귀었습니다\_CVV\_REF)

나. 학습자 전략에 의한 오류 주석

- 학습자는 자신이 모르는 어휘를 자신의 모국어로 전환하거나 코드 스위칭을 통해 한국어가 아닌 다른 외국어로 어휘를 대체하여 사용하기도 한다. 이는 어휘와 문장 차원에서 이루어질 수 있으나, 우선 어휘 차원에서 이루어지는 모국어로서의 전환은 실질어와 이의 대체 오류를 통해 해당 어휘를 검색한 후 이에 해당 오류를 주석하여 활용할 수 있다.

예) 저는 큰 dream(꿈\_CNN\_REF)이 있어요.

다. 기타

- 이 외에 각 문법 영역(조사, 어미, 문형)에 따른 오류를 검색하여 의미에 따른 오류, 사용에 따른 오류, 모국어 전이에 따른 오류, 과잉 일반화에 의한 오류 등 다양한 오류 주석을 생성하여 이를 활용할 수 있다.

② 확장 주석의 추가 오류 표지를 활용한 개방형 주석

- 현 오류 주석 체계에서는 음운, 형태, 통사, 담화 층위에서의 오류를 일부 확장하여 주석하고 있다. 실제 학습자 말뭉치에서는 이들 층위에서의 다양한 오류 양상이 가능하여 이에 대한 연구를 위해 오류 주석 체계가 확장될 필요가 있다. 따라서 오류 층위에 따라 다음과 같은 확장형 개방형 주석을 생성하고 활용할 수 있다.

가. 음운

- 학습자의 발음 오류는 주로 모국어와 목표어 간의 음운 체계 차이에 의해 나타나는 경우가 많은데, 이러한 현상을 면밀하게 살펴보기 위해 음소 차원이 아닌 음성 차원, 그리고 운소 차원의 오류 표지를 세분화할 필요가 있다. 가령, ‘-(으)르걸’이라는 어미는 상승 억양일 경우는 ‘추측’, 하강 억양일 경우는 ‘후회’로 그 기능이 달라진다. 그 차이를 변별하지 못해 의사소통이 원활하게 이루어지지 않는다면 억양에 의한 오류로 간주할 수 있는데, 이러한 작업은 연구자가 음성 파일을 직접 들으면서 추가 주석을

하는 것이 효율적이다. 따라서 다음과 같은 오류 주석의 생성과 활용이 가능할 수 있다.

<표 40> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-음운 층위

오류 유형	
음운	음소[발음/철자]
	음절[발음/철자]
	음운규칙
	억양
	강세
	… 사용자가 목적에 따라 추가 주석 범주 생성

#### 나. 형태

- 형태 층위의 오류는 지금까지 주로 문법 범주에서 다루었던 어미, 조사 오류와 어휘 범주에서 다루었던 오류들이 포함된다. 그 중 어휘 오류는 선행 연구에서 체계에 의해 오류를 범주화하여 다루어 왔다.
- 먼저 Corder(1981)에서는 오류의 유형을 아래의 표와 같이 구분하였다.

<표 41> Corder(1981)의 오류 유형

	Graphological Phonological	Grammatical	Lexico-semantic
Omission	…	…	…
Addition	…	…	…
Selection	…	…	…
Ordering	…	…	…

즉, 어휘 오류는 세 번째인 어휘의미론적(Lexico-semantic) 범주에 해당되며 표면적인 오류 형태에 따라서 누락(Omission), 첨가(Addition), 대치(Selection), 어순(Ordering)로 나누고 있다. 누락 오류는 어떤 요소가 나타나야 하는 곳에서 누락된 것을 말하고 첨가 오류는 나타나지 않아야 할 요소가 나타난 오류이며 대치 오류는 맞는 것 대신에 틀린 것을 선택하여

나타난 오류를 말한다. 마지막으로 어순 오류는 순서가 틀린 것을 의미한다. 그러나 표면적인 형태를 기반으로 한 이러한 오류 분석은 체계적인 분석의 시작점에 불과하다. Corder는 학습자의 언어를 제대로 이해하려면 오류의 범주나 유형이 보다 체계적이어야 하며 오류를 판정하는 것 역시 중요하다고 한다.

- Carl James(1998)는 어휘 오류의 중요성을 강조하면서 다양한 연구들을 언급하고 있다. 먼저 Graberg(1971)는 독일어 고급 학습자들의 오류를 분석한 결과 어휘 오류가 53%로 빈번하게 나타났다고 하였으며 Meara(1984) 또한 다른 유형의 오류보다 어휘 오류가 수적으로 우세했다고 한다. 그리하여 James는 어휘 오류를 크게 형태적인 어휘 오류(Formal errors of lexis)와 의미론적인 어휘 오류(Semantic errors in lexis)로 구분하였다. 형태적인 어휘 오류는 잘못된 어휘 선택의 오류(Formal misselection)와 조어 오류(Misformation)와 왜곡된 오류(Distortions)로 세분화하였다. 먼저 잘못된 어휘 선택의 오류(Formal misselection)는 말의 오용을 포함하는 오류 유형으로 주로 잘못된 맞춤법의 오류가 많으며 오류로 사용된 어휘가 실제로 존재하는 어휘라는 것이 큰 특징이다. 아래의 예가 잘못된 어휘 선택의 오류이다.

예) He wanted to \*cancel(√conceal) his guilt.

- 조어 오류(Misformation)는 학습자가 존재하지 않은 어휘를 생산해 내는 오류이다. 이 오류는 다시 모국어 단어를 차용하여 쓴 오류와 목표어 구조에 모국어의 어휘를 가져와서 새로운 단어를 생산한 오류, 모국어의 직역으로 인해 발생한 오류로 나뉘 볼 수 있다. 마지막으로 왜곡된 오류(Distortions)는 제1언어의 영향을 받지 않은 언어 내 오류이며 목표어에는 존재하지 않는 어휘를 학습의 과정 중에 어떤 규칙을 잘못 적용하여 발생한 오류라고 할 수 있다. 누락, 대치, 어순 오류, 혼합 등이 이에 해당한다. 혼합은 두 가지의 어휘를 섞어 사용한 경우로 ‘starps’(√stops+starts)와 같은 오류를 말한다.

의미론적 어휘 오류(Semantic errors in lexis)는 어휘의 관계에 대한 혼동에서 오는 오류(Confusion of sense relations)와 언어 관계의 오류(Collocational errors)로 볼 수 있는데 전자는 다시 상하위어의 오류, 유의어의 오류, 두 가지 종류의 하위어의 오류로 나뉘 볼 수 있다.

예) The flowers had a special \*smell(√scent/√perfume).: 상하위어의 오류

예) She is my \*nephew(√niece).: 잘못된 하위어의 선택 오류

예) ...a \*regretful(√penitent): 가까운 유의어의 잘못된 사용 오류

언어 관계의 오류는 어떤 어휘는 특정한 것과 결합하는데 이 관계를 알지 못하여 발생하는 오류를 말한다.

- 한국어교육에서 살펴보면 이정희(2002)에서 어휘 오류는 문법 형태의 오류보다 훨씬 형태가 다양하고 문법에서 나타나는 정문과 비문의 경계처럼 오류가 명확하지도 않기 때문에 연구하기 어려운 분야라고 하였다. 특히 한자권 학습자들의 경우에는 유의어를 다른 어휘로 대체하는 현상이 많이 나타나고 비한자권 학습자들의 경우에는 생략이나 표기법 오용 등의 현상을 많이 관찰할 수 있다고 하였다. 어휘 오류에는 어휘를 생략하거나 불필요한 말을 첨가하거나 어휘간의 의미를 구별하지 못한 경우들, 어휘 파생과 호응이 잘못 되어 나타난 오류문은 특히 고급 단계의 학습자들에게 많이 나타나고, 지시어 사용의 오류는 모국어 배경이나 학습 단계의 차이 없이 폭넓게 나타났다고 밝혔다, 마찬가지로 접속어의 문제도 있었다. 이외에 초급 단계의 학습자의 경우 코드 전환의 예들도 많이 나타나는데 영어권 학습자의 경우에는 다른 언어권 학습자들보다 더 많이 코드 전환 전략을 사용하였으며 한자어권 학습자들은 한자를 그대로 옮겨 쓰는 모국어 전이 현상을 많이 보여 주었다.
- 그 외에도 형태 층위에서 학습자의 어휘 사용과 발달을 측정하는 데에 중요한 지표가 될 수 있는 어종(고유어, 한자어, 외래어)이나 조어 방식(단일어, 파생어, 합성어 등) 등을 다룰 수 있다. 다음은 그 중 일부를 표지로 추가하여 보인 예이다.

<표 42> 한국어 학습자 말뭉치의 오류 주식 체계 틀의 확장 예시-형태 층위

오류 유형	
형태	단어 형성[합성법]
	단어 형성[파생법]
	굴절[곡용]
	굴절[활용]

오류 유형	
	어종[고유어]
	어종[외래어]
	어종[한자어]
	… 사용자가 목적에 따라 추가 주석 범주 생성

#### 다. 통사

- 선행 연구의 문법 오류를 살펴보면 오류 층위와 현상이 통합되고 있어 다소 복잡해 보이지만 공통적으로 다루고 있는 범주에서는 큰 차이를 보이지 않는다. 다만, 격조사, 보조사, 접속조사, 과거 시제, 현재 시제, 미래 시제와 같이 특정 문법 범주를 세분화한 경우가 있다.
- 그 외에 문장 호응 오류, 문장 성분을 포함한 문장 배열 규칙과 관련된 오류, 오류의 위치와 양상이 문장 전체에 걸쳐 있는 오류의 처리를 위한 추가 주석을 고려해 볼 수 있다. 이를 위해서는 오류 주석 영역을 형태 단위 이상으로 범주화하여야 하므로 이를 기술적으로 처리 가능하도록 하는 방안이 필요하다.

<표 43> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-통사 층위

오류 유형		
통사	높임	
	시제	
	사동	장형
		단형
	피동	장형
		단형
	부정	‘안’ 부정
		‘못’ 부정
		‘말다’ 부정
	… 사용자가 목적에 따라 추가 주석 범주 생성	

라. 담화

- 담화 단위는 다른 언어 단위에 비해 소위 오류와 비오류의 중간지대(middle ground)가 상대적으로 넓다. 사실 오류와 비오류는 엄격한 문법성과 낮은 용인성을 기준으로 이분법적으로 구분되어야 하지만, 담화에는 적절성이나 원형성이 떨어지는 사례들이 많고 이들 예들의 경우 용인성에 다소 문제가 있으나 비오류로 간주될 정도는 아니어서 오류 또는 비오류로의 객관적인 식별이 어렵기 때문이다. 객관적인 주석 말뭉치 자료를 제공하여야 하는 본 프로젝트의 특성으로, 이들 오류와 비오류의 경계에 있는 담화 사례들은 본 연구진의 자의적인 해석을 보류하고 향후 주석 말뭉치를 활용할 개별 연구자의 판단에 따라 분석되어 활용될 수 있을 것이다. 닫힌 체계를 지향하는 본 오류 주석 체계에는 포함하지 않았지만 개방형 주석을 할 경우 추가 가능한 주석은 다음과 같다.

<표 44> 한국어 학습자 말뭉치의 오류 주석 체계 틀의 확장 예시-담화 층위

오류 유형	
담화	지시
	접속
	담화표지
	구어/문어 오류
	생략, 병렬구조, 어휘적 응집장치, 시제 등과 같은 응집성(cohesion)
	상호접근성이나 관련성 등과 같은 응결성(coherence)
	추측, 바람, 능력, 가능성, 당위성, 정도성, 의지, 의도, 시도, 완료, 봉사 등의 양태
	허락, 금지, 제안, 권유, 명령 등의 화행
	사용역, 언어적 공손성 등 언어 변이
	수사 구조, 단락 구성 양태 등 담화 구조
	비선호 대응, 순서교대, 타자 수정 등 대화분석
	... 사용자가 목적에 따라 추가 주석 범주 생성

### 3) 오류와 실수의 구분

- 오류와 실수는 구분하지 않고 규범상의 일탈은 모두 오류 주석의 대상으로 삼기로 하였다. 이는 연구자의 판단 영역으로 자료만으로 학습자의 의도를 파악할 수 없기 때문에 주석 작업자의 자의적인 해석을 막기 위한 것이다.
- 한편, 구어 자료의 경우 학습자가 실수임을 인지하고 자기 수정을 하게 되는데, 수정하기 이전의 일탈도 오류로 간주하여 주석 대상에 포함시킨다.

### 4) 오류 판정과 교정의 기준

- 보편적으로 오류의 식별과 판정 기준은 문법성과 용인 가능성으로 삼는다. 문법성이란 의미적으로나 형태적으로 완성된 형식을 갖추지 못하고 한국어의 문법 체계에 맞지 않는 비문법적 문장을 생성하는 경우를 말한다. 다음은 오류 판정과 교정의 기준이다.

<표 45> 오류 판정과 교정의 기준

쟁점	쟁점의 내용	문제 해결 방안
쟁점 1. 오류 판정의 상대성	일반적으로 오류 판정은 문법성(grammaticality)과 용인가능성(acceptability)을 기준에 따르는데 이에 대한 직관이 사람마다 다름	학습자 언어를 접한 경험이 많고 오류 처방에 관한 전문 지식이 있는 한국어교육 전문가에 의한 오류 판정과 수정
쟁점 2. 오류 교정의 대상	오류 교정의 대상을 어디까지 할 것인지의 문제	학습자 말뭉치에 오류 주석을 부착하여 학습자 오류를 유형화하는 것이 목적이므로 모든 오류를 수정 대상으로 함
쟁점 2. 오류	오류로 판정된 문장을 교정할 때	학습자의 표현 의도

교정의 방식	그것을 문법적으로 완전한 문장으로 바꿀 것인지 용인가능한 수준의 문장으로 바꿀 것인지에 대한 문제	를 반영하여 용인가능한 수준으로 최소한의 수정함
--------	--	----------------------------

### 7.3. 2016년 오류 주식 지침 및 기구축 자료 보완 방향

- 2016년에는 수정된 오류 주식 체계를 지침에 반영하고 이에 따라 기구축된 오류 주식 말뭉치를 수정·보완하는 작업이 필요하다.

## VI. 1차 연도 한국어 학습자 말뭉치 구축의 실제

### 1. 구축 과정 및 절차

- 1차 연도에 구축된 한국어 학습자 말뭉치는 기초 연구를 통해 수립한 구축 지침에 따랐으며, 다음과 같은 단계를 거쳐 구축되었다.

<표 46> 한국어 학습자 말뭉치 구축 과정 및 절차

구축 단계	세부 절차	구축 작업 인력 배치
자료 수집	한국어 학습자 말뭉치 구축 네트워크 구성	28개 기관의 수집 실무 책임자 28명
	자료 수집 지침 교육	구축 본부 실무 인력 1명
	자료 수집	28개 기관의 수집 실무 교사 30명, 지원 교사 1,000명
자료 처리	자료 처리 지침 교육	구축 본부 실무 인력 1명

	수집 자료 분류 및 스캔	자료 처리 전담 인력 3명 (※ 전용 스캐너 1대)
	파일명 부여	
	파일 등록	
원시 말뭉치 구축	입력/전사 자료 선정	구축 본부 실무 인력 5명
	입력/전사 지침 교육	구축 본부 실무 인력 5명
	입력/전사	입력 전담 인력 9명 전사 전담 인력 5명 추가 지원 인력 25명
	검수	입력 자료 검수 인력 5명 전사 자료 검수 인력 9명 추가 지원 인력 25명
형태 주식 말뭉치 구축	형태 주식 자료 선정	구축 본부 실무 인력 1명
	형태 주식 지침 교육	구축 본부 실무 인력 1명
	형태 주식	책임 실무 인력 1명 (공동연구원)
	검수	책임 실무 인력 1명 오류 주식 전담 인력 5명
오류 주식 말뭉치 구축	오류 주식 자료 선정	구축 본부 실무 인력 1명
	오류 주식 지침 교육	구축 본부 실무 인력 1명
	오류 주식	오류 주식 전담 인력 5명
	검수	

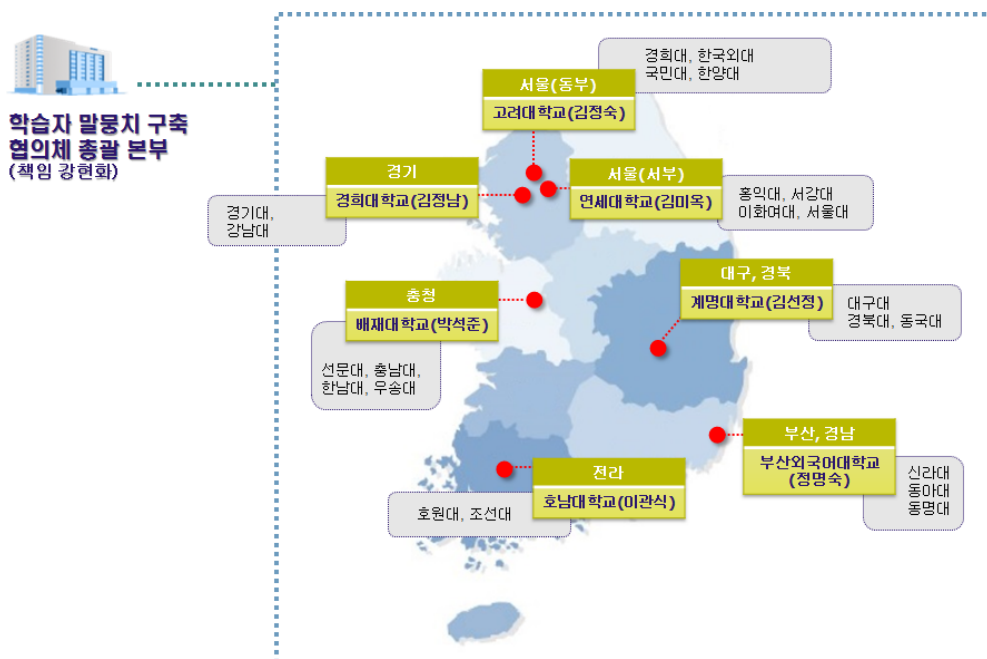
## 2. 단계별 세부 구축 내용 및 결과

### 2.1. 자료 수집

#### 1) 한국어 학습자 말뭉치 구축 네트워크 구성

- 지역 배분을 고려하여 서울(동부/서부), 경기, 충청, 호남, 대구·경북, 부산·경남의 7개 거점 기관을 중심으로 한국어 학습자 말뭉치 협의체를 구성한

후 자료 수집을 위한 총 28개의 수집 네트워크를 마련하였다.



<그림 11> 지역별 거점 기관을 중심으로 한 협의체 구성

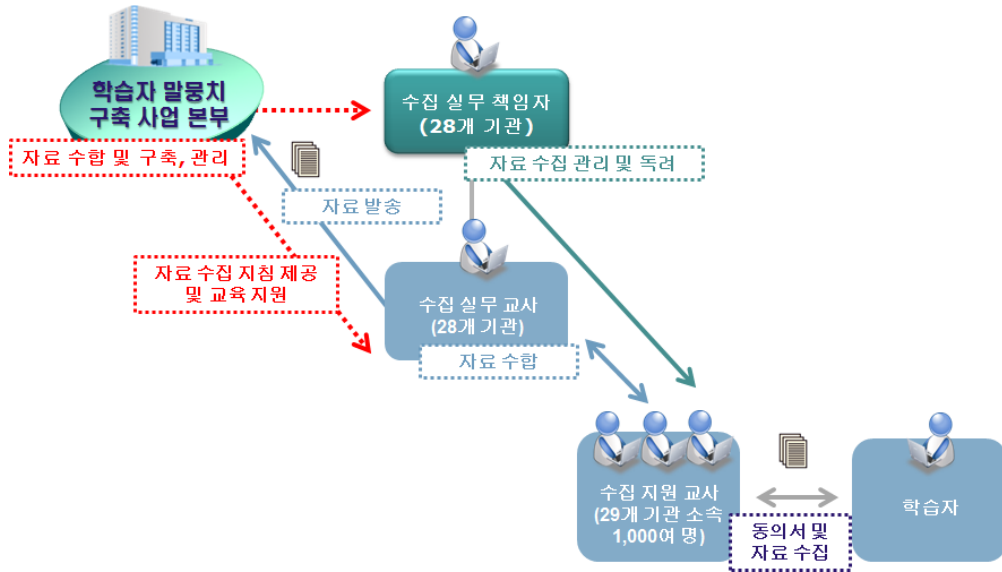
- 수집 기관에는 수집 실무 및 홍보, 교육을 총괄하는 책임자와 실질적인 수집을 담당하는 실무 교사가 배치되었다. 아울러 각 기관에는 학습자들과 대면하여 학습자 동의서를 받고 자료 수집을 담당하는 수집 지원 교사 1,000여 명이 수집 업무에 관여하였다.

<표 47> 수집 기관별 실무 책임자 및 실무 교사 명단

지역	수집 기관	책임자	실무 교사
서울(동부)	고려대학교 한국어교육센터	김정숙	심재경
	경희대학교 국제교육원	이정희	이지영 윤혜리
	한국외국어대학교 한국어문화교육원	허 용	김은정
	한양대학교 국제어학원	이영숙	배소영
	국민대학교 국제교육원 한국어교육센터	이동은	노미연

지역	수집 기관	책임자	실무 교사
서울(서부)	연세대학교 한국어학당	김미옥	단현주
	서울대학교 언어교육원	안경화	신범숙 김폴잎
	서강대학교 한국어교육원	김현정	김현정
	이화여자대학교 언어교육원	김현진	박진철
	홍익대학교 국제언어교육원	이화진	홍연정
경기	경희대학교 언어교육원	김유미	이지민
	경기대학교 언어교육원	배현대	배현대
	강남대학교 한국어교육원	서정숙	서정숙
충청	배재대학교 한국어교육원	박석준	김민우
	충남대학교 언어교육원	김세진	박광진
	한남대학교 한국어학당	임예영	임예영
	우송대학교 한국어교육원	임명옥	임새아미
	선문대학교 한국어교육원	라혜민	유순천
호남	호남대학교 국제교류본부	이관식	추민교
	호원대학교 한국어교육원	진대연	김은영
	조선대학교 언어교육원	강희숙	박수연
대구·경북	계명대학교 국제교육센터	김선정	홍종호
	대구대학교 한국어교육센터	우창현	김현진
	경북대학교 한국어문화원	서효경	안민지
	동국대학교(경주) 국제교류교육원	배현숙	박솔지
부산·경남	부산외국어대학교 한국어문화교육원	정명숙	오상민
	신라대학교 한국어교육센터	조정순	구다혜
	동아대학교 언어교육원	김지혜	김지혜
	동명대학교 언어교육원	김상수	김상수

○ 자료 수집 절차 및 체계는 다음과 같다.



<그림 12> 자료 수집 체계도

○ 자료 수집은 구축 본부와 29개 수집 실무 책임자, 수집 실무 교사, 수집 지원 교사 1,000여 명의 긴밀한 네트워크를 통해 이루어지며 각각의 역할은 다음과 같다.

- 구축 본부: 수집 지침 개발 및 교육 지원
- 수집 실무 책임자: 수집 실무 교사 및 지원 교사 관리 및 수집 업무 독려
- 수집 실무 교사: 수집 지원 교사가 수집한 자료를 수합하여 구축 본부에 제출
- 수집 지원 교사(담임교사): 반별 학습자와 접촉하여 학습자 동의서 및 자료 수집 지원

## 2) 자료 수집 지침 배포 및 교육

○ 자료 수집 지침 교육은 수집 실무 책임자, 수집 실무 교사, 실질적으로 학습자를 접촉하여 자료를 수집하는 지원 교사에게 필요한 내용들을 중심으

로 이루어졌다. 자료 수집 지침 교육에 포함된 내용은 다음과 같다.

- 자료 수집 및 처리 지침: 자료 수집, 수합, 발송에 관한 세부 업무 지침
  - 한국어 학습자 말뭉치 구축 사업을 위한 학습자 자료 이용 동의서 수집 (횡적/일반)
  - 한국어 학습자 말뭉치 원문/음성 파일 이용에 관한 동의서 수집 (학문 목적 학습자에 한하여 선택)
  - 수집 자료 유형과 수집 방법: 수집 대상별 자료 유형, 수집 원칙, 수집 방법. 과제 활동
- 자료 수집 지침은 전자메일을 통해 수집 책임자 및 수집 실무자에게 배포되었다. 아울러 정기, 비정기적인 유선 통화 및 전자메일을 통해 수집 지침에 관해 보충 설명을 하고 실무자들의 궁금증을 해소하는 방식으로 비대면 교육을 실시하였다.
- 자료 수집과 구축의 효율성 제고를 위하여 기관별 수집 실무자의 요구와 1차 자료를 수집하는 과정에서 발생하는 문제점을 반영하여 자료 수집 지침 및 학습자 동의서를 수정·보완하였다.

### 3) 자료 수집 결과

- 수집 계획에 따라 2015년 여름 학기, 가을 학기 자료 28개의 기관에 재학 중인 학습자 약 7,000여 명의 자료가 수집되었다.
- 종적 말뭉치는 구축 본부에서 지정한 주제에 관한 작문과 말하기 자료 각 1편씩을 2주 간격으로 수집하고 있다. 종적 말뭉치 수집 대상자는 44명으로 수집 목표 인원인 30명의 약 1.5배수로 중도에 학업을 포기하여 정규 교육과정을 마치지 못하는 학생이 발생할 수 있음을 감안한 것이다. 실제로 최초에 47명을 섭외하였으나 여름 학기 이후에 7명의 학생이 유급 또는 자진 철회하였으며, 가을 학기에 4명의 학습자를 추가 섭외하였다.

<표 48> 국적별 종적 말뭉치 수집 대상자 분포

수집 기관	국적	학습자 수(명)
경희대	베트남	2
	사우디 아라비아	1
	스웨덴	1
	중국	6
	태국	1
	홍콩	1
고려대	사우디아라비아	1
	수단	1
	중국	4
	콜롬비아	1
국민대	몽골	2
	중국	1
서울대	인도네시아	1
	중국	2
	키르기스스탄	1
연세대	말레이시아	2
	몽골	1
	이탈리아	1
	인도네시아	1
	일본	1
	프랑스	1
경희대(국제)	중국	4
배재대	중국	5
부산외대	중국	4
	몽골	1
합계		47



## 2) 입력 및 전사 지침 교육

- 기초 연구 단계에서 개발된 입력 및 전사 지침, 전사 도구로 사용된 소프트웨어 엘란의 사용 방법을 교육하였다. 본 사업에서 전사 도구로 사용한 엘란은 본래 다층 주석에 최적화된 도구로 음성과 동영상을 동기화하여 음성 언어 외에도 준언어, 몸짓언어 등에 관한 정보를 다수의 층위에서 주석할 수 있는 소프트웨어이다.

## 3) 입력 및 전사

- 작업 지침에 따라 입력과 전사 작업을 진행하였다. 입력 작업은 메모장이나 기타 텍스트 에디터 프로그램에서 하였으며, 전사 작업은 엘란을 사용하여 전사 작업을 한 후에 텍스트 파일 형식으로 자료를 출력하였다.

18-0

010

※ 이 답안지는 연습용 모의 답안지입니다.

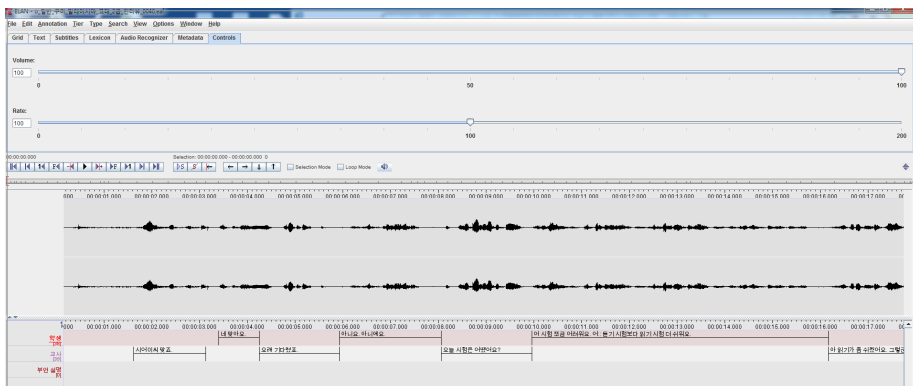
주관식 답안 (Answer sheet for composition)									
46	이래 빈칸에 150자에서 300자 이내로 적음(한글·한자·영어·숫자 포함)								
	저는	주말에	보통	건대	에				
15									
30	가	오.	건대	에서	좋은	친구	를	있	
45	었	어요.	모	름	날	씨가	너	무	더
60	위	서	그	래	서	안	가	요.	기
75	에	서	드	라	마	하	고	영	화
90	게	요.	이	중	에	서	특	히	코
105	화	하	고	공	포	영	화	를	좋
120	그	런	데	공	포	영	화	보	다
135	화	더	재	미	있	어요.	중	국	에
150	말	하	는	한	국	민	니	있	었
165	서	우	리	가	끔	만	나	요.	언
180	쁘	고	싶	어요.	지	난	주	삼	청
195	가	어요.	평	화	날	씨	종	고	경
210	너	무	예	뻔	어요.	만	있	는	윤
225	많	이	있	어요.	우	리	"떡	복	이
240	리	떡	같	다	반	지	를	산	어요.
255	루	에	뽕	꼴	난	후	에	기	속
270	어요.	기	분	이	너	무	종	았	어요.
285	다	음	에	다	시	만	날	까	요.
300									

※ 국어전 영역의 방향을 바워서 읽었을 때 'ㄱ'을 적어줍니다.  
※ Please do not turn the page sideways. No point will be given.

<그림 14> 학습자 작문 원본 스캔 파일 예시

	1	2	3	4	5
▶	1	저는 주말에 보통 전대에 가요.			
	2	전대에서 좋은 친구를 있었어요.			
	3	요즘 날씨가 너무 더워서 그래서 안 가요.			
	4	기숙사에서 드라마하고 영화를 볼게요.			
	5	이 중에서 특히 코미디 영화하고 공포 영화가 좋아해요.			
	6	그런데 공포 영화보다 코미디 영화 더 재미있어요.			
	7	중국에서 아는 한국 언니 있었어요.			
	8	언니가 보고 싶어요.			
	9	지난주 삼청동에 갔어요.			
	10	날씨 좋고 경치가 너무 예뻐했어요.			
	11	맛있는 음식도 많이 있었어요.			
	12	우리 "떡볶이치스"를 먹었어요.			
	13	너무 맛있어요.			
	14	우리 독감다 반지를 샀어요.			
	15	하루에 끝난 후에 기숙사 갔어요.			
	16	기분이 너무 좋았어요.			
	17	다음에 다시 만날까요.			

<그림 15> 학습자 작문 입력 파일 예시



<그림 16> 전사 도구 엘란의 실행 화면

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
▶	1	학생	1	00:00:03.400	00:00:04.252	네 맞아요.																	
	2	학생	1	00:00:05.930	00:00:08.061	아니요. 아니에요.																	
	3	학생	1	00:00:09.961	00:00:16.179	어 시험 조금 어려워요. 어:: 듣기 시험보다 읽기 시험 더 쉬워요.																	
	4	학생	1	00:00:22.729	00:00:25.587	어:: 저는 말레이시아 <note>한글덱</note> 사람이예요.																	
	5	학생	1	00:00:31.207	00:00:36.866	고향 어:: 전에 사진작가였어요. 그렇지만 지금 학생이에요.																	
	6	학생	1	00:00:42.869	00:00:45.581	어:: 특히 모자 사진 자주 찍었어요.																	
	7	학생	1	00:00:46.416	00:00:49.518	아 모범 모범 사진 자주 찍어요.																	
	8	학생	1	00:00:57.459	00:01:05.255	공부 어:: 한국에서 어:: 한국어를 공부해 취직하고 싶어요.																	
	9	학생	1	00:01:12.220	00:01:19.600	어:: 사진작가 되고싶어요.																	
	10	학생	1	00:01:27.130	00:01:31.840	네 제 취미는 사진 찍는 것을 좋아해요.																	
	11	학생	1	00:01:35.306	00:01:43.650	보통 일주일엔 한 두번쯤 찍어요. 그렇지만 요즘 시험 있어서 못 찍어요.																	
	12	학생	1	00:01:47.705	00:01:54.077	보통 공원이나 바다에 자주 사진을 찍어요.																	
	13	학생	1	00:02:00.330	00:02:06.075	사람 개인사진은 개인사진 자주 찍어요.																	
	14	학생	1	00:02:11.162	00:02:20.613	어 어:: 제가 재미= 재미있는 사진 찍을 수 있기 때문에 사진을 자주 찍어요.																	
	15	학생	1	00:02:27.210	00:02:34.950	음 네. 대략 때 어:: 사년전에 대학 때 사진찍을 공부했어요.																	
	16	학생	1	00:02:43.669	00:02:47.128	어 없어요.																	
	17	학생	1	00:03:01.922	00:03:08.852	네 있어요. 특히 어:: 소녀시대 좋아요.																	
	18	학생	1	00:03:11.489	00:03:14.513	소녀시대 중에서 배연 좋아요.																	
	19	학생	1	00:03:18.076	00:03:25.407	눈이= 눈이 크고 다리가 조금 작은 편이에요.																	
	20	학생	1	00:03:29.103	00:03:37.855	얼굴은 음:: 얼굴이 예쁜 편이에요.																	
	21	학생	1	00:03:42.478	00:03:46.866	바로																	
	22	학생	1	00:03:49.120	00:03:56.250	음 이십 삼요.																	
	23	학생	1	00:03:59.018	00:04:04.799	그니까 얼굴형이 둥그란= 둥그란 편이에요.																	

<그림 17> 전사 도구 엘란으로 전사한 후 출력한 파일의 예시

#### 4) 검수

- 입력과 전사와 완료된 파일은 지침을 숙지하고 있는 전담 인력에 의한 검토와 수정 과정을 거쳐 완성되었다.

#### 5) 원시 말뭉치 구축 결과

- 원시 말뭉치는 문어 30만 어절, 구어 10만 어절의 규모로 구축되었다. 구어의 경우 주요한 수집 자료인 말하기 시험 자료의 특성상 교사와의 인터뷰 자료가 많고, 고급 단계의 경우 발화 길이가 길어진다. 이러한 점을 고려하여 학습자 발화의 분량을 확보하고 다양한 학습자의 자료를 확보하기 위하여 당초 목표 규모인 5만 어절의 2배수에 달하는 양을 구축하였다. 원시 말뭉치의 세부 구성은 다음과 같다.

<표 49> 원시 말뭉치의 수준별 자료 규모

구분		1급	2급	3급	4급	5급	6급	합계
문어	어절 수	51,446	55,869	53,839	50,477	51,676	50,274	313,581
	파일 수	723	496	454	389	363	321	2,746
구어	어절 수	6,875	13,751	12,702	15,717	29,328	27,438	105,811
	파일 수	30	17	19	17	30	17	130

<표 50> 원시 말뭉치의 국적별 자료 규모

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	184,881	1,670	31,289	32	216,170	1,702
일본	48,365	398	16,236	19	64,601	417
대만	13,682	102	9,693	7	23,375	109
미국	6,452	58	7,395	10	13,847	68
베트남	7,669	68	1,929	5	9,598	73
카자흐스탄	4,054	34	4,279	3	8,333	37

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
말레이시아	4,800	37	1,929	5	6,729	42
태국	4,391	36	2,098	5	6,489	41
홍콩	6,029	47	151	1	6,180	48
러시아	2,118	19	2,511	3	4,629	22
우즈베키스탄	2,254	21	2,143	2	4,397	23
싱가포르	2,187	17	1,282	2	3,469	19
영국	661	6	2,507	3	3,168	9
독일	1,904	16	1,071	2	2,975	18
이탈리아	1,876	14	1,045	2	2,921	16
프랑스	1,619	13	1,243	3	2,862	16
한국	1,420	12	1,369	2	2,789	14
호주	927	10	1,783	4	2,710	14
스웨덴	1,252	13	1,383	3	2,635	16
몽골	1,650	15	760	2	2,410	17
인도네시아	1,596	14	0	0	1,596	14
싱가폴	1,529	8	0	0	1,529	8
미얀마	351	3	1,127	1	1,478	4
벨기에	280	4	1,131	1	1,411	5
스페인	429	5	975	1	1,404	6
캐나다	1,209	12	181	1	1,390	13
터키	678	6	590	1	1,268	7
사우디아라비아	795	9	445	1	1,240	10
네덜란드	261	3	862	1	1,123	4
이집트	143	1	898	1	1,041	2
코스타리카	149	1	755	1	904	2
콜롬비아	139	1	691	1	830	2
스리랑카	121	1	615	1	736	2
세르비아	76	1	624	1	700	2
아제르바이젠	633	5	0	0	633	5

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
필리핀	481	5	0	0	481	5
핀란드	459	4	0	0	459	4
불가리아	60	1	359	1	419	2
모로코	393	3	0	0	393	3
이중 국적	297	4	0	0	297	4
멕시코	275	2	0	0	275	2
인도	241	3	0	0	241	3
니카라과	238	2	0	0	238	2
이란	236	2	0	0	236	2
브라질	196	2	0	0	196	2
루마니아	185	1	0	0	185	1
헝가리	173	1	0	0	173	1
우크라이나	170	1	0	0	170	1
브루나이	169	1	0	0	169	1
뉴질랜드	163	1	0	0	163	1
도미니카	162	2	0	0	162	2
나이지리아	158	2	0	0	158	2
칠레	158	1	0	0	158	1
아르헨티나	148	1	0	0	148	1
에콰도르	146	1	0	0	146	1
라오	145	1	0	0	145	1
베네수엘라	144	1	0	0	144	1
마카오	141	1	0	0	141	1
트리니다드토바고 공화국	140	1	0	0	140	1
네팔	139	1	0	0	139	1
마다가스카	132	1	0	0	132	1
르완다	130	1	0	0	130	1
중동	128	2	0	0	128	2

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
볼리비아	124	1	0	0	124	1
가나	116	1	0	0	116	1
가봉	115	1	0	0	115	1
방글라데시	109	1	0	0	109	1
엘살바도르	108	1	0	0	108	1
파테말라	104	1	0	0	104	1
파키스탄	97	3	0	0	97	3
이디오피아	93	1	0	0	93	1
노르웨이	92	1	0	0	92	1
오스트리아	85	1	0	0	85	1
에디오피아	84	1	0	0	84	1
폴란드	75	1	0	0	75	1
키르기스스탄	69	1	0	0	69	1
캄보디아	64	1	0	0	64	1
남수단	59	1	0	0	59	1
합계	313,581	2,746	105,811	130	419,392	2,876

## 2.4. 형태 주석 말뭉치 구축

### 1) 형태 주석 자료 선정

- 형태 주석 자료의 선정은 형태 주석 말뭉치로 구축될 자료를 선별해 내는 과정이다. 원시 말뭉치로 구축된 자료 중 일부를 학습자의 수준별로 균형성을 맞추어 선정하였다.

## 2) 형태 주석 지침 교육

- 형태 주석 지침 교육은 형태 주석 검수 작업자를 위한 것으로 기초 연구 단계에서 개발된 형태 주석 지침을 교육하였다. 교육 내용은 형태 주석 작업 및 검수에 필요한 형태 주석 태그 세트의 설명과 학습자의 오류로 인해 분석 불능(NA) 처리된 자료의 처리 방법에 대한 내용이 핵심을 이루었다.

## 3) 형태 주석

- 작업 지침에 따라 리눅스 기반의 지능형 형태 주석 도구 KMAT를 사용하여 형태 주석 작업을 수행하였다. 다음은 형태 주석이 완료된 자료의 예이다.

	1	2	3	4	5
1	제가	제/NP+가/JKS			
2	친구하고	친구/NNG+하고/JKB			
3	영향을	영향/NA+을/JKO			
4	가고	가/VV+고/EC			
5	싫습니다.	싫/VX+습니다/EF+./SF			
6	저는	저/NP+는/JX			
7	한국	한국/NNP			
8	서울만	서울/NNP+만/JX			
9	갑니다.	가/VV+브니다/EF+./SF			
10	친구하고	친구/NNG+하고/JKB			
11	같이	같이/MAG			
12	한국	한국/NNP			
13	영향을	영향/NA+을/JKO			
14	갑니다.	가/VV+브니다/EF+./SF			
15	우리	우리/NP			
16	한국	한국/NNP			
17	음식하고	음식/NNG+하고/JKB			
18	문화들도	문화/NNG+들/JKO+도/JX			
19	좋아합니다.	좋아하/VV+브니다/EF+./SF			
20	하지만	하지만/NA			
21	월요일부터	월요일/NNG+부터/JX			
22	금요일까지	금요일/NNG+까지/JX			
23	수업입니다.	수업/NNG+이/VCP+브니다/EF+./SF			
24	주말에	주말/NNG+에/JKB			
25	맛있지	맛있/VX+지/EC			
26	먹고	먹/VV+고/EC			
27	놀습니다.	놀/VV+습니다/EF+./SF			
28	매주말	매/MM+주말/NNG			
29	나무	나무/NA			
30	재미있습니다.	재미/NNG+일/NA+습니다/EF+./SF			
31	우리	우리/NP			
32	한국	한국/NNP			
33	드라마를	드라마/NNG+를/JKO			
34	피습니다.	피/VV+습니다/EF+./SF			
35	그리고	그리고/MAG			
36	아이스크림하고	아이스크림/NNG+하고/JKB			
37	맛집을	맛집/NNG+을/JKO			
38	먹습니다.	먹/VV+습니다/EF+./SF			
39	주말에	주말/NNG+에/JKB			
40	기쁩니다.	기쁘/VX+브니다/EF+./SF			
41	지금	지금/MAG			
42	방학을	방학/NNG+을/JKO			
43	기다리겠습니다.	기다리/VV+겠/EF+습니다/EF+./SF			

<그림 18> 형태 주석 결과의 예시

#### 4) 검수

- 형태 주석 검수 작업은 두 단계에 걸쳐 이루어졌다. 먼저, 형태 주석을 마친 후 형태 주석 도구에 의해 오분석된 것을 수정하였다. 그리고 나서 오류 주석 작업과 병행하여 학습자 오류로 인해 분석 불능(NA) 처리되거나 오분석된 부분들을 검토하여 수정하였다.
- 형태 주석이 완료된 파일은 온라인 구축 시스템을 통해 다음과 같이 XML 기반의 파일 형식으로 출력하여 활용 가능하다.

```

- <BODY>
  - <T_UNIT>
    <s>제가 친구하고 영행을 가고 싶습니다.</s>
    - <학습자오류>
      <ERROR error_type="PS">영행</ERROR>
      <ERROR error_type="MISF">영행</ERROR>
    </학습자오류>
    - <형태분석>
      - <word>
        <w>제가</w>
        <morph pos="NP">제</morph>
        <morph pos="JKS">가</morph>
        <w>친구하고</w>
        <morph pos="NNG">친구</morph>
        <morph pos="JKB">하고</morph>
        <w>영행을</w>
        <morph pos="NA">영행</morph>
        <morph pos="JKO">을</morph>
        <w>가고</w>
        <morph pos="VV">가</morph>
        <morph pos="EC">고</morph>
        <w>싶습니다.</w>
        <morph pos="VX">싶</morph>
        <morph pos="EF">습니다</morph>
        <morph pos="SF">.</morph>
      </word>
    </형태분석>
  </T_UNIT>
- <T_UNIT>

```

<그림 19> XML 기반의 학습자 말뭉치 출력 파일 예시-형태 주석

## 5) 형태 주식 말뭉치 구축 결과

- 형태 주식 말뭉치는 문어 20만 어절, 구어 3만 어절의 규모로 구축되었다. 구어의 경우 주요한 수집 자료인 말하기 시험 자료의 특성상 교사와의 인터뷰 자료가 많고, 고급 단계의 경우 발화 길이가 길어진다. 이러한 점을 고려하여 학습자 발화의 분량을 확보하고 다양한 학습자의 자료를 확보하기 위하여 인터뷰보다는 학습자 단독의 발표 자료를 조금 더 많이 선정하고, 당초 목표 규모인 2만 어절의 1.5배수에 달하는 양을 구축하였다. 형태 주식 말뭉치의 세부 구성은 다음과 같다.

<표 51> 형태 주식 말뭉치의 수준별 자료 규모

구분		1급	2급	3급	4급	5급	6급	합계
문어	어절 수	34,077	30,109	35,377	35,136	35,231	31,618	201,548
	파일 수	483	271	312	271	248	208	1,793
구어	어절 수	4,251	4,722	4,789	4,725	7,070	4,844	30,401
	파일 수	17	6	7	8	8	4	50

<표 52> 형태 주식 말뭉치의 국적별 자료 규모

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	122,609	1,128	4,885	12	127,494	1,140
일본	28,400	228	4,445	5	32,845	233
대만	7,372	58	986	1	8,358	59
미국	4,291	37	3,279	5	7,570	42
베트남	5,413	50	842	1	6,255	51
카자흐스탄	3,149	26	1,756	2	4,905	28
태국	3,229	28	1,229	3	4,458	31
말레이시아	3,253	26	1,149	2	4,402	28
홍콩	3,718	29	151	1	3,869	30
우즈베키스탄	1,488	14	2,143	2	3,631	16
싱가포르	1,822	14	724	1	2,546	15
독일	1,340	11	829	1	2,169	12

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
러시아	1,285	12	643	1	1,928	13
스웨덴	835	10	1,060	2	1,895	12
몽골	966	10	588	1	1,554	11
영국	352	3	878	1	1,230	4
호주	218	3	862	3	1,080	6
인도네시아	1,064	9	0	0	1,064	9
네덜란드	190	2	862	1	1,052	3
이집트	143	1	898	1	1,041	2
사우디아라비아	594	7	445	1	1,039	8
프랑스	951	7	0	0	951	7
한국	900	7	0	0	900	7
캐나다	744	6	0	0	744	6
스리랑카	121	1	615	1	736	2
터키	678	6	0	0	678	6
이탈리아	567	5	0	0	567	5
필리핀	481	5	0	0	481	5
스페인	355	4	0	0	355	4
멕시코	275	2	0	0	275	2
미얀마	246	2	0	0	246	2
아제르바이젠	229	2	0	0	229	2
이중 국적	212	3	0	0	212	3
핀란드	208	2	0	0	208	2
벨기에	190	3	0	0	190	3
루마니아	185	1	0	0	185	1
헝가리	173	1	0	0	173	1
모로코	172	1	0	0	172	1
브루나이	169	1	0	0	169	1
나이지리아	158	2	0	0	158	2
코스타리카	149	1	0	0	149	1

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
아르헨티나	148	1	0	0	148	1
에콰도르	146	1	0	0	146	1
라오	145	1	0	0	145	1
베네수엘라	144	1	0	0	144	1
트리니다드토바고 공화국	140	1	0	0	140	1
네팔	139	1	0	0	139	1
콜롬비아	139	1	0	0	139	1
마다가스카	132	1	0	0	132	1
도미니카	131	1	0	0	131	1
르완다	130	1	0	0	130	1
중동	128	2	0	0	128	2
볼리비아	124	1	0	0	124	1
가나	116	1	0	0	116	1
브라질	111	1	0	0	111	1
방글라데시	109	1	0	0	109	1
엘살바도르	108	1	0	0	108	1
이디오피아	93	1	0	0	93	1
노르웨이	92	1	0	0	92	1
에디오피아	84	1	0	0	84	1
폴란드	75	1	0	0	75	1
키르기스스탄	69	1	0	0	69	1
캄보디아	64	1	0	0	64	1
남수단	59	1	0	0	59	1
인도	28	1	0	0	28	1
합계	201,548	1,793	30,401	50	231,949	1,843

## 6) 형태 주식 말뭉치 분석 결과

- 형태 주식 결과를 바탕으로 전체 말뭉치, 문어 말뭉치, 구어 말뭉치의 수준별 품사 분포를 분석한 결과는 다음과 같다.

### (1) 전체 말뭉치 수준별 품사 분포

<표 53> 형태 주식 결과 분석: 전체 말뭉치의 수준별 품사 분포

품사		구분	1급	2급	3급	4급	5급	6급	합계
체언	일반명사	빈도	12,502	14,073	14,990	17,725	18,704	17,339	95,333
		백분율	19.27	19.71	20.69	22.19	23.60	24.13	21.68
	고유명사	빈도	2,420	1,663	1,113	588	685	715	7,184
		백분율	3.73	2.33	1.54	0.74	0.86	1.00	1.63
	의존명사	빈도	1,909	1,951	2,448	3,004	2,911	2,793	15,016
		백분율	2.94	2.73	3.38	3.76	3.67	3.89	3.41
	대명사	빈도	1,882	2,229	1,789	1,380	1,088	757	9,125
		백분율	2.90	3.12	2.47	1.73	1.37	1.05	2.08
	수사	빈도	367	135	164	142	260	83	1,151
		백분율	0.57	0.19	0.23	0.18	0.33	0.12	0.26
용언	동사	빈도	5,948	7,184	7,090	7,100	6,566	5,846	39,734
		백분율	9.17	10.06	9.79	8.89	8.28	8.14	9.04
	형용사	빈도	2,045	2,387	2,832	3,219	2,954	2,490	15,927
		백분율	3.15	3.34	3.91	4.03	3.73	3.47	3.62
	보조용언	빈도	1,196	1,305	1,512	1,941	1,719	1,622	9,295
		백분율	1.84	1.83	2.09	2.43	2.17	2.26	2.11
	지정사	빈도	1,040	895	954	1,133	1,412	1,221	6,655
		백분율	1.60	1.25	1.32	1.42	1.78	1.70	1.51
수식언	관형사	빈도	458	824	1,026	1,103	1,418	1,218	6,047
		백분율	0.71	1.15	1.42	1.38	1.79	1.70	1.38
	일반부사	빈도	2,458	3,337	3,425	3,127	2,677	2,386	17,410
		백분율	3.79	4.67	4.73	3.91	3.38	3.32	3.96

품사		구분	1급	2급	3급	4급	5급	6급	합계
	접속	빈도	922	907	774	730	667	569	4,569
	부사	백분율	1.42	1.27	1.07	0.91	0.84	0.79	1.04
독립언	감탄사	빈도	408	813	302	528	475	1,013	3,539
		백분율	0.63	1.14	0.42	0.66	0.60	1.41	0.80
관계언	주격조사	빈도	1,882	2,160	2,757	3,578	3,215	2,641	16,233
		백분율	2.90	3.03	3.81	4.48	4.06	3.68	3.69
	보격조사	빈도	9	128	88	134	205	179	743
		백분율	0.01	0.18	0.12	0.17	0.26	0.25	0.17
	관형격조사	빈도	347	557	625	953	1,096	1,104	4,682
		백분율	0.53	0.78	0.86	1.19	1.38	1.54	1.06
	목적격조사	빈도	3,182	2,912	2,507	2,814	2,757	2,661	16,833
		백분율	4.90	4.08	3.46	3.52	3.48	3.70	3.83
	부사격조사	빈도	5,531	4,521	4,210	3,896	3,960	3,499	25,617
		백분율	8.52	6.33	5.81	4.88	5.00	4.87	5.83
	인용격조사	빈도	5	128	17	9	23	6	188
		백분율	0.01	0.18	0.02	0.01	0.03	0.01	0.04
	보조사	빈도	3,260	3,180	3,077	3,386	3,299	2,907	19,109
		백분율	5.02	4.45	4.25	4.24	4.16	4.05	4.35
의존형태	연결어미	빈도	3,197	4,732	5,700	6,698	6,536	6,284	33,147
		백분율	4.93	6.63	7.87	8.39	8.25	8.75	7.54
	종결어미	빈도	6,768	5,644	5,026	4,575	3,867	3,109	28,989
		백분율	10.43	7.90	6.94	5.73	4.88	4.33	6.59
	선어말어미	빈도	1,701	2,641	1,485	1,131	1,099	893	8,950
		백분율	2.62	3.70	2.05	1.42	1.39	1.24	2.04
	명사형전성어미	빈도	158	361	355	486	615	444	2,419
		백분율	0.24	0.51	0.49	0.61	0.78	0.62	0.55
	관형사형전성어미	빈도	1,452	2,859	3,492	4,487	4,993	4,572	21,855
		백분율	2.24	4.00	4.82	5.62	6.30	6.36	4.97
	형용사과생접미사	빈도	196	385	611	654	701	492	3,039

품사		구분	1급	2급	3급	4급	5급	6급	합계
		백분율	0.30	0.54	0.84	0.82	0.88	0.68	0.69
		빈도	328	351	551	1,393	1,003	871	4,497
	명사파생 접미사	백분율	0.51	0.49	0.76	1.74	1.27	1.21	1.02
		빈도	979	1,349	1,595	2,119	2,629	2,693	11,364
	동사파생 접미사	백분율	1.51	1.89	2.20	2.65	3.32	3.75	2.58
분석 불능		빈도	2,338	1,791	1,940	1,845	1,733	1,438	11,085
		백분율	3.60	2.51	2.68	2.31	2.19	2.00	2.52
합계		빈도	64,888	71,402	72,455	79,878	79,267	71,845	439,735
		백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

## (2) 문어 말뭉치 수준별 품사 분포

<표 54> 형태 주석 결과 분석: 문어 말뭉치의 수준별 품사 분포

품사		구분	1급	2급	3급	4급	5급	6급	합계
체언	일반 명사	빈도	11,613	13,132	14,125	16,549	16,736	15,893	88,048
		백분율	19.43	20.21	21.12	22.69	24.12	25.44	22.21
	고유 명사	빈도	2,238	1,405	988	431	471	597	6,130
		백분율	3.74	2.16	1.48	0.59	0.68	0.96	1.55
	의존 명사	빈도	1,786	1,757	2,303	2,749	2,563	2,482	13,640
		백분율	2.99	2.70	3.44	3.77	3.69	3.97	3.44
	대명사	빈도	1,759	2,088	1,607	1,219	940	541	8,154
		백분율	2.94	3.21	2.40	1.67	1.35	0.87	2.06
	수사	빈도	259	77	134	78	95	55	698
		백분율	0.43	0.12	0.20	0.11	0.14	0.09	0.18
용언	동사	빈도	5,568	6,582	6,562	6,561	5,709	5,099	36,081
		백분율	9.31	10.13	9.81	9.00	8.23	8.16	9.10
	형용사	빈도	1,859	2,075	2,584	2,961	2,747	2,111	14,337
		백분율	3.11	3.19	3.86	4.06	3.96	3.38	3.62
	보조 용언	빈도	1,166	1,220	1,397	1,827	1,576	1,466	8,652
		백분율	1.95	1.88	2.09	2.51	2.27	2.35	2.18

품사		구분	1급	2급	3급	4급	5급	6급	합계
	지정사	빈도	986	822	884	1,009	1,250	1,086	6,037
		백분율	1.65	1.27	1.32	1.38	1.80	1.74	1.52
수식언	관형사	빈도	402	756	895	874	1,019	817	4,763
		백분율	0.67	1.16	1.34	1.20	1.47	1.31	1.20
	일반부사	빈도	2,274	2,960	3,113	2,854	2,398	1,859	15,458
		백분율	3.80	4.56	4.65	3.91	3.46	2.98	3.90
	접속부사	빈도	863	819	702	628	534	435	3,981
		백분율	1.44	1.26	1.05	0.86	0.77	0.70	1.00
독립언	감탄사	빈도	2	7	2	1	3	2	17
		백분율	0.00	0.01	0.00	0.00	0.00	0.00	0.00
관계언	주격조사	빈도	1,746	1,973	2,548	3,396	2,904	2,388	14,955
		백분율	2.92	3.04	3.81	4.66	4.19	3.82	3.77
	보격조사	빈도	7	124	78	130	185	171	695
		백분율	0.01	0.19	0.12	0.18	0.27	0.27	0.18
	관형격조사	빈도	333	544	588	921	1,003	1,030	4,419
		백분율	0.56	0.84	0.88	1.26	1.45	1.65	1.11
	목적격조사	빈도	3,047	2,852	2,371	2,670	2,431	2,477	15,848
		백분율	5.10	4.39	3.55	3.66	3.50	3.97	4.00
	부사격조사	빈도	5,226	4,191	4,013	3,570	3,477	3,194	23,671
		백분율	8.74	6.45	6.00	4.89	5.01	5.11	5.97
	인용격조사	빈도	5	128	16	8	20	4	181
		백분율	0.01	0.20	0.02	0.01	0.03	0.01	0.05
	보조사	빈도	3,002	2,964	2,853	3,110	3,033	2,553	17,515
		백분율	5.02	4.56	4.27	4.26	4.37	4.09	4.42
의존형태	연결어미	빈도	2,999	4,398	5,213	6,221	5,850	5,432	30,113
		백분율	5.02	6.77	7.80	8.53	8.43	8.70	7.60
	종결어미	빈도	6,333	4,996	4,691	4,163	3,439	2,725	26,347
		백분율	10.59	7.69	7.01	5.71	4.96	4.36	6.65
	선어말어미	빈도	1,506	2,493	1,262	970	813	688	7,732

품사		구분	1급	2급	3급	4급	5급	6급	합계
		백분율	2.52	3.84	1.89	1.33	1.17	1.10	1.95
	명사형전 성어미	빈도	144	345	320	472	549	419	2,249
		백분율	0.24	0.53	0.48	0.65	0.79	0.67	0.57
	관형사형 전성어미	빈도	1,352	2,695	3,277	4,169	4,521	4,166	20,180
		백분율	2.26	4.15	4.90	5.72	6.52	6.67	5.09
	형용사파 생접미사	빈도	167	369	591	626	647	444	2,844
		백분율	0.28	0.57	0.88	0.86	0.93	0.71	0.72
	명사와생 접미사	빈도	270	340	496	1,336	843	805	4,090
		백분율	0.45	0.52	0.74	1.83	1.21	1.29	1.03
	동사와생 접미사	빈도	911	1,286	1,497	1,987	2,372	2,479	10,532
백분율		1.52	1.98	2.24	2.72	3.42	3.97	2.66	
분석 불능		빈도	1,956	1,574	1,765	1,443	1,257	1,048	9,043
		백분율	3.27	2.42	2.64	1.98	1.81	1.68	2.28
합계		빈도	59,779	64,972	66,875	72,933	69,385	62,466	396,410
		백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

### (3) 구어 말뭉치 수준별 품사 분포

<표 55> 형태 주석 결과 분석: 구어 말뭉치의 수준별 품사 분포

품사		구분	1급	2급	3급	4급	5급	6급	합계
체언	일반명사	빈도	889	941	865	1,176	1,968	1,446	7,285
		백분율	17.40	14.63	15.50	16.93	19.91	15.42	16.81
	고유명사	빈도	182	258	125	157	214	118	1,054
		백분율	3.56	4.01	2.24	2.26	2.17	1.26	2.43
	의존명사	빈도	123	194	145	255	348	311	1,376
		백분율	2.41	3.02	2.60	3.67	3.52	3.32	3.18
	대명사	빈도	123	141	182	161	148	216	971
		백분율	2.41	2.19	3.26	2.32	1.50	2.30	2.24
	수사	빈도	108	58	30	64	165	28	453
		백분율	2.19	0.91	0.45	0.89	2.38	0.28	1.15

품사		구분	1급	2급	3급	4급	5급	6급	합계
용 언		백분율	2.11	0.90	0.54	0.92	1.67	0.30	1.05
	동사	빈도	380	602	528	539	857	747	3,653
		백분율	7.44	9.36	9.46	7.76	8.67	7.96	8.43
	형용사	빈도	186	312	248	258	207	379	1,590
		백분율	3.64	4.85	4.44	3.71	2.09	4.04	3.67
	보조 용언	빈도	30	85	115	114	143	156	643
		백분율	0.59	1.32	2.06	1.64	1.45	1.66	1.48
	지정사	빈도	54	73	70	124	162	135	618
		백분율	1.06	1.14	1.25	1.79	1.64	1.44	1.43
수 식 언	관형사	빈도	56	68	131	229	399	401	1,284
		백분율	1.10	1.06	2.35	3.30	4.04	4.28	2.96
	일반 부사	빈도	184	377	312	273	279	527	1,952
		백분율	3.60	5.86	5.59	3.93	2.82	5.62	4.51
	접속 부사	빈도	59	88	72	102	133	134	588
		백분율	1.15	1.37	1.29	1.47	1.35	1.43	1.36
독 립 언	감탄사	빈도	406	806	300	527	472	1,011	3,522
		백분율	7.95	12.53	5.38	7.59	4.78	10.78	8.13
관 계 언	주격 조사	빈도	136	187	209	182	311	253	1,278
		백분율	2.66	2.91	3.75	2.62	3.15	2.70	2.95
	보격 조사	빈도	2	4	10	4	20	8	48
		백분율	0.04	0.06	0.18	0.06	0.20	0.09	0.11
	관 형 격 조사	빈도	14	13	37	32	93	74	263
		백분율	0.27	0.20	0.66	0.46	0.94	0.79	0.61
	목 적 격 조사	빈도	135	60	136	144	326	184	985
		백분율	2.64	0.93	2.44	2.07	3.30	1.96	2.27
	부 사 격 조사	빈도	305	330	197	326	483	305	1,946
		백분율	5.97	5.13	3.53	4.69	4.89	3.25	4.49
	인 용 격 조사	빈도			1	1	3	2	7
		백분율	0.00	0.00	0.02	0.01	0.03	0.02	0.02
	보조사	빈도	258	216	224	276	266	354	1,594

품사		구분	1급	2급	3급	4급	5급	6급	합계
의 존 형 태		백분율	5.05	3.36	4.01	3.97	2.69	3.77	3.68
	연결 어미	빈도	198	334	487	477	686	852	3,034
		백분율	3.88	5.19	8.73	6.87	6.94	9.08	7.00
	종결 어미	빈도	435	648	335	412	428	384	2,642
		백분율	8.51	10.08	6.00	5.93	4.33	4.09	6.10
	선어말 어미	빈도	195	148	223	161	286	205	1,218
		백분율	3.82	2.30	4.00	2.32	2.89	2.19	2.81
	명사형전 성어미	빈도	14	16	35	14	66	25	170
		백분율	0.27	0.25	0.63	0.20	0.67	0.27	0.39
	관형사형 전성어미	빈도	100	164	215	318	472	406	1,675
		백분율	1.96	2.55	3.85	4.58	4.78	4.33	3.87
	형용사파 생접미사	빈도	29	16	20	28	54	48	195
		백분율	0.57	0.25	0.36	0.40	0.55	0.51	0.45
	명사파생 접미사	빈도	58	11	55	57	160	66	407
		백분율	1.14	0.17	0.99	0.82	1.62	0.70	0.94
	동사파생 접미사	빈도	68	63	98	132	257	214	832
		백분율	1.33	0.98	1.76	1.90	2.60	2.28	1.92
	분석 불능	빈도	382	217	175	402	476	390	2,042
		백분율	7.48	3.37	3.14	5.79	4.82	4.16	4.71
	합계	빈도	5,109	6,430	5,580	6,945	9,882	9,379	43,325
		백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

## 2.5. 오류 주석 말뭉치 구축

### 1) 오류 주석 자료 선정

- 오류 주석 자료 선정은 오류 주석 말뭉치로 구축될 자료를 선별해 내는 과정이다. 형태 주석 작업이 완료된 자료 중 일부를 수준별 균형성을 맞추어 선정하였다.

## 2) 오류 주석 지침 교육

- 오류 주석 지침 교육은 기초 연구 단계에서 개발된 오류 주석 지침과 엑셀에서 이루어지는 주석 작업 처리 방법을 설명하였다. 엑셀에서의 주석 작업 처리 방법은 온라인 구축 시스템에서의 오류 주석 작업은 필요한 경우 두 개 이상의 형태소를 블록 처리하여 주석이 가능하도록 설계되었는데, 엑셀에서는 그것을 구현할 수 없기 때문에 교정 어절의 생성 시 형태소가 가감할 때 이를 처리하는 방식에 대한 설명이 주를 이루었다.

## 3) 오류 주석

- 작업 지침에 따라 오류 주석 작업을 진행하였다. 오류 주석 작업은 분석 어절을 포함한 원문 문장과 파일명에 있는 국적, 숙달도 정보를 참조하면서 하였다.

	D	E	F	G	H	I	J	K	L
1	파일 번호	원어절	형태 주석	형태 주석 수정	교정 어절	교정 어절의 형태	주석 분석 여부	오류 종류	오류 현상
2	1	<title>"론딩리"의	/SS		<title>"론딩리"을	/SS			
3	2	<title>"론딩리"의	론딩리/NNP		<title>"론딩리"을	론딩리/NNP			
4	3	<title>"론딩리"의	/SS		<title>"론딩리"을	/SS			
5	4	<title>"론딩리"의	의/JKG		<title>"론딩리"을	을/JKO		GPT(문법 조사)	REP(대치)
6	5	소개합니다.</title>	소개/NG						
7	6	소개합니다.</title>	하/XSV						
8	7	소개합니다.</title>	는 니다/EF						
9	8	저는	저/NP						
10	9	저는	는/JX						
11	10	대만	대만/NNP						
12	11	론딩리에	론딩리/NNP						
13	12	론딩리에	에/JKB						
14	13	산지	살/VV						
15	14	산지	니/ETM						
16	15	산지	지/NNB						
17	16	15년	15/SN		15년이	15/SN			
18	17	15년	년/NNB		15년이	년/NNB			

원문	파일명
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01
<title>"론딩리"의 소개합니다.</title> 저는 대만 론딩리에 산지 15년 왔습니다.	일반_문어_대만_이대_2급_0012-01

<그림 20> 오류 주석 작업 화면 예시

#### 4) 검수

- 오류 주석이 1차 완료된 파일은 지침을 숙지하고 있는 전담 인력과 공동 연구원에 의한 검토와 수정 과정을 거쳐 완성되었다.
- 오류 주석이 완료된 파일은 온라인 구축 시스템을 통해 다음과 같이 XML 기반으로 출력하여 파일 형태로 활용 가능하다.

```

- <BODY>
  - <T_UNIT>
    <s>제가 친구하고 영행을 가고 싶습니다.</s>
    - <학습자오류>
      <ERROR error_type="PS">영행</ERROR>
      <ERROR error_type="MISF">영행</ERROR>
    </학습자오류>
    - <형태분석>
      - <word>
        <w>제가</w>
        <morph pos="NP">제</morph>
        <morph pos="JKS">가</morph>
        <w>친구하고</w>
        <morph pos="NNG">친구</morph>
        <morph pos="JKB">하고</morph>
        <w>영행을</w>
        <morph pos="NA">영행</morph>
        <morph pos="JKO">을</morph>
        <w>가고</w>
        <morph pos="VV">가</morph>
        <morph pos="EC">고</morph>
        <w>싶습니다.</w>
        <morph pos="VX">싶</morph>
        <morph pos="EF">습니다</morph>
        <morph pos="SF">.</morph>
      </word>
    </형태분석>
  </T_UNIT>
- <T_UNIT>

```

<그림 21> XML 기반의 학습자 말뭉치 출력 파일 예시-오류 주석

#### 5) 오류 주석 말뭉치 구축 결과

- 원시 말뭉치는 문어 30만 어절, 구어 10만 어절의 규모로 구축되었다. 구어의 경우 주요한 수집 자료인 말하기 시험 자료의 특성상 교사와의 인터뷰 자료가 많고, 고급 단계의 경우 발화 길이가 길어진다. 이러한 점을 고려하여 학습자 발화의 분량을 확보하고 다양한 학습자의 자료를 확보하기

위하여 당초 목표 규모인 5만 어절의 2배수에 달하는 양을 구축하였다.  
원시 말뭉치의 세부 구성은 다음과 같다.

<표 56> 오류 주식 말뭉치의 수준별 자료 규모

구분		1급	2급	3급	4급	5급	6급	합계
문어	어절 수	7,469	7,796	8,269	7,033	7,246	7,188	45,001
	파일 수	103	67	62	50	48	46	376
구어	어절 수	2,333	2,454	3,270	2,083	3,123	2,247	15,510
	파일 수	10	3	5	4	4	2	28

<표 57> 오류 주식 말뭉치의 국적별 자료 규모

국적	문어		구어		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	24,252	212	2,035	7	26,287	219
일본	11,255	90	1,316	1	12,571	91
카자흐스탄	2,392	18	818	1	3,210	19
미국	1,586	12	919	2	2,505	14
대만	962	7	986	1	1,948	8
베트남	445	5	842	1	1,287	6
스웨덴	168	2	1,060	2	1,228	4
러시아	511	4	643	1	1,154	5
영국	268	2	878	1	1,146	3
태국	679	5	330	1	1,009	6
독일	120	1	829	1	949	2
이집트	0	0	898	1	898	1
싱가포르	129	1	724	1	853	2
홍콩	641	5	151	1	792	6
말레이시아	0	0	749	1	749	1
호주	0	0	648	2	648	2
세르비아	0	0	624	1	624	1
스리랑카	0	0	615	1	615	1

사우디아라비아	0	0	445	1	445	1
우즈베키스탄	349	3	0	0	349	3
몽골	213	2	0	0	213	2
프랑스	200	1	0	0	200	1
터키	184	1	0	0	184	1
아르헨티나	148	1	0	0	148	1
인도네시아	129	1	0	0	129	1
미얀마	125	1	0	0	125	1
볼리비아	124	1	0	0	124	1
캐나다	121	1	0	0	121	1
합계	45,001	376	15,510	28	60,511	404

## 6) 오류 주석 말뭉치 분석 결과

### (1) 분석 여부

<표 58> 오류 주석 결과의 분석: 분석 여부

유형	구분	1급	2급	3급	4급	5급	6급	합계
문어	빈도	73	198	168	161	74	74	748
	백분율	0.02	0.05	0.04	0.04	0.02	0.02	0.19
구어	빈도	62	78	6	46	192	4	388
	백분율	0.14	0.18	0.01	0.11	0.44	0.01	0.90
합계	빈도	135	276	174	207	266	78	1,136
	백분율	0.03	0.06	0.04	0.05	0.06	0.02	0.26

## (2) 오류 현상

### ① 전체 말뭉치

<표 59> 오류 주석 결과의 분석: 전체 말뭉치의 오류 현상

유형	구분	1급	2급	3급	4급	5급	6급	합계
대치	빈도	477	750	854	584	571	492	3,728
	백분율	32.47	42.74	45.74	43.13	38.58	42.45	41.04
누락	빈도	170	350	245	75	153	146	1,139
	백분율	11.57	19.94	13.12	5.54	10.34	12.60	12.54
첨가	빈도	79	108	171	90	51	71	570
	백분율	5.38	6.15	9.16	6.65	3.45	6.13	6.27
오어순	빈도	26	17	41	18	21	0	123
	백분율	1.77	0.97	2.20	1.33	1.42	0.00	1.35
오형태	빈도	717	530	556	587	684	450	3,524
	백분율	48.81	30.20	29.78	43.35	46.22	38.83	38.79
합계	빈도	1,469	1,755	1,867	1,354	1,480	1,159	9,084
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

### ② 문어 말뭉치

<표 60> 오류 주석 결과의 분석: 문어 말뭉치의 오류 현상

유형	구분	1급	2급	3급	4급	5급	6급	합계
대치	빈도	423	722	802	540	457	444	3,388
	백분율	36.69	43.84	50.03	46.55	45.88	48.21	45.29
누락	빈도	158	326	237	71	141	136	1,069
	백분율	13.70	19.79	14.78	6.12	14.16	14.77	14.29
첨가	빈도	63	96	159	88	47	59	512
	백분율	5.46	5.83	9.92	7.59	4.72	6.41	6.84
오어순	빈도	26	17	33	18	17	0	111
	백분율	2.25	1.03	2.06	1.55	1.71	0.00	1.48

오 형태	빈도	483	486	372	443	334	282	2,400
	백분율	41.89	29.51	23.21	38.19	33.53	30.62	32.09
합계	빈도	1,153	1,647	1,603	1,160	996	921	7,480
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

### ③ 구어 말뭉치

<표 61> 오류 주석 결과의 분석: 구어 말뭉치의 오류 현상

유형	구분	1급	2급	3급	4급	5급	6급	합계
대치	빈도	54	28	52	44	114	48	340
	백분율	17.09	25.93	19.70	22.68	23.55	20.17	21.20
누락	빈도	12	24	8	4	12	10	70
	백분율	3.80	22.22	3.03	2.06	2.48	4.20	4.36
첨가	빈도	16	12	12	2	4	12	58
	백분율	5.06	11.11	4.55	1.03	0.83	5.04	3.62
오어순	빈도	0	0	8	0	4	0	12
	백분율	0.00	0.00	3.03	0.00	0.83	0.00	0.75
오 형태	빈도	234	44	184	144	350	168	1,124
	백분율	74.05	40.74	69.70	74.23	72.31	70.59	70.07
합계	빈도	316	108	264	194	484	238	1,604
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

## (3) 오류 총위

### ① 오류 총위별 분포

#### ○ 전체 말뭉치

<표 62> 오류 주석 결과의 분석: 전체 말뭉치의 오류 총위

유형	구분	1급	2급	3급	4급	5급	6급	합계
말음	빈도	236	38	176	134	333	167	1,084
	백분율	16.07	2.17	9.43	9.90	22.50	14.41	11.93
어휘	빈도	507	598	557	485	489	433	3,069
	백분율	34.51	34.07	29.83	35.82	33.04	37.36	33.78

문법	빈도	717	1,081	1,115	716	645	552	4,826
	백분율	48.81	61.60	59.72	52.88	43.58	47.63	53.13
담화	빈도	9	38	19	19	13	7	105
	백분율	0.61	2.17	1.02	1.40	0.88	0.60	1.16
합계	빈도	1,568	1,853	1,966	1,453	1,579	1,258	9,183
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

## ○ 문어 말뭉치

<표 63> 오류 주석 결과의 분석: 문어 말뭉치의 오류 총위

유형	구분	1급	2급	3급	4급	5급	6급	합계
발음	빈도	0	0	2	0	1	1	4
	백분율	0.00	0.00	0.12	0.00	0.10	0.11	0.05
어휘	빈도	493	580	542	465	423	407	2,910
	백분율	42.76	0.00	33.81	40.09	42.47	44.19	38.90
문법	빈도	651	1,029	1,040	676	559	506	4,461
	백분율	56.46	0.00	64.88	58.28	56.12	54.94	59.64
담화	빈도	9	38	19	19	13	7	105
	백분율	0.78	0.00	1.19	1.64	1.31	0.76	1.40
합계	빈도	1,153	1,647	1,603	1,160	996	921	7,480
	백분율	100.00	0.00	100.00	100.00	100.00	100.00	100.00

## ○ 구어 말뭉치

<표 64> 오류 주석 결과의 분석: 구어 말뭉치의 오류 총위

유형	구분	1급	2급	3급	4급	5급	6급	합계
발음	빈도	236	38	174	134	332	166	1,080
	백분율	74.68	35.19	65.91	69.07	68.60	69.75	67.33
어휘	빈도	14	18	15	20	66	26	159
	백분율	4.43	16.67	5.68	10.31	13.64	10.92	9.91
문법	빈도	66	52	75	40	86	46	365
	백분율	20.89	48.15	28.41	20.62	17.77	19.33	22.76

담화	빈도	0	0	0	0	0	0	0
	백분율	0.00	0.00	0.00	0.00	0.00	0.00	0.00
합계	빈도	316	108	264	194	484	238	1,604
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

② 오류 층위의 발음 영역

○ 전체 말뭉치

<표 65> 오류 주석 결과의 분석: 전체 말뭉치의 오류 층위-발음

유형	구분	1급	2급	3급	4급	5급	6급	합계
음절	빈도	226	36	168	128	330	164	1,052
	백분율	95.76	94.74	95.45	95.52	99.10	98.20	97.05
음운 규칙	빈도	10	2	8	6	3	3	32
	백분율	4.24	5.26	4.55	4.48	0.90	1.80	2.95
소계	빈도	236	38	176	134	333	167	1,084
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

○ 문어 말뭉치

<표 66> 오류 주석 결과의 분석: 문어 말뭉치의 오류 층위-발음

유형	구분	1급	2급	3급	4급	5급	6급	합계
음절	빈도	0	0	0	0	0	0	0
	백분율	0.00	0.00	0.00	0.00	0.00	0.00	0.00
음운 규칙	빈도			2		1	1	4
	백분율	0.00	0.00	100.00	0.00	100.00	100.00	100.00
소계	빈도	0	0	2	0	1	1	4
	백분율	0.00	0.00	100.00	0.00	100.00	0.00	100.00

○ 구어 말뭉치

<표 67> 오류 주석 결과의 분석: 구어 말뭉치의 오류 층위-발음

유형	구분	1급	2급	3급	4급	5급	6급	합계
음절	빈도	226	36	168	128	330	164	1,052
	백분율	95.76	94.74	96.55	95.52	99.40	98.80	97.41
음운 규칙	빈도	10	2	6	6	2	2	28
	백분율	4.24	5.26	3.45	4.48	0.60	1.20	2.59
소계	빈도	236	38	174	134	332	166	1,080
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

③ 오류 층위의 어휘 영역

○ 전체 말뭉치

<표 68> 오류 주석 결과의 분석: 전체 말뭉치의 오류 층위-어휘

유형	구분	1급	2급	3급	4급	5급	6급	합계
명사	빈도	305	251	245	291	274	232	1,598
	백분율	60.16	41.97	43.99	60.00	56.03	53.58	52.07
대명사	빈도	16	34	38	9	15	7	119
	백분율	3.16	5.69	6.82	1.86	3.07	1.62	3.88
수사	빈도	4	2	3	3	7	5	24
	백분율	0.79	0.33	0.54	0.62	1.43	1.15	0.78
동사	빈도	64	148	118	89	83	82	584
	백분율	12.62	24.75	21.18	18.35	16.97	18.94	19.03
형용사	빈도	52	58	34	23	30	26	223
	백분율	10.26	9.70	6.10	4.74	6.13	6.00	7.27
보 용언	빈도	4	16	18	14	13	8	73
	백분율	0.79	2.68	3.23	2.89	2.66	1.85	2.38
지정사	빈도	6	6	17	14	5	16	64
	백분율	1.18	1.00	3.05	2.89	1.02	3.70	2.09
관형사	빈도	10	12	16	3	9	4	54

	백분율	1.97	2.01	2.87	0.62	1.84	0.92	1.76
부사	빈도	37	63	55	27	31	30	243
	백분율	7.30	10.54	9.87	5.57	6.34	6.93	7.92
감탄사	빈도	1	1	0	2	0	0	4
	백분율	0.20	0.17	0.00	0.41	0.00	0.00	0.13
접사	빈도	8	7	13	10	22	23	83
	백분율	1.58	1.17	2.33	2.06	4.50	5.31	2.70
소계	빈도	507	598	557	485	489	433	3,069
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

○ 문어 말뭉치

<표 69> 오류 주석 결과의 분석: 문어 말뭉치의 오류 층위-어휘

유형	구분	1급	2급	3급	4급	5급	6급	합계
명사	빈도	299	243	241	281	238	228	1,530
	백분율	60.65	41.90	44.46	60.43	56.26	56.02	52.58
대명사	빈도	16	32	36	9	15	5	113
	백분율	3.25	5.52	6.64	1.94	3.55	1.23	3.88
수사	빈도	4	2	3	3	7	5	24
	백분율	0.81	0.34	0.55	0.65	1.65	1.23	0.82
동사	빈도	62	142	118	83	67	71	543
	백분율	12.58	24.48	21.77	17.85	15.84	17.44	18.66
형용사	빈도	48	58	32	23	30	26	217
	백분율	9.74	10.00	5.90	4.95	7.09	6.39	7.46
보조 용언	빈도	4	16	18	14	11	6	69
	백분율	0.81	2.76	3.32	3.01	2.60	1.47	2.37
지정사	빈도	6	6	17	12	3	16	60
	백분율	1.22	1.03	3.14	2.58	0.71	3.93	2.06
관형사	빈도	10	12	14	3	7	4	50
	백분율	2.03	2.07	2.58	0.65	1.65	0.98	1.72
부사	빈도	35	61	53	27	29	30	235
	백분율	7.10	10.52	9.78	5.81	6.86	7.37	8.08

감탄사	빈도	1	1					2
	백분율	0.20	0.17	0.00	0.00	0.00	0.00	0.07
접사	빈도	8	7	10	10	16	16	67
	백분율	1.62	1.21	1.85	2.15	3.78	3.93	2.30
소계	빈도	493	580	542	465	423	407	2,910
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

○ 구어 말뭉치

<표 70> 오류 주석 결과의 분석: 구어 말뭉치의 오류 층위-어휘

유형	구분	1급	2급	3급	4급	5급	6급	합계
명사	빈도	6	8	4	10	36	4	68
	백분율	42.86	44.44	26.67	50.00	54.55	15.38	42.77
대명사	빈도	0	2	2	0	0	2	6
	백분율	0.00	11.11	13.33	0.00	0.00	7.69	3.77
수사	빈도	0	0	0	0	0	0	0
	백분율	0.00	0.00	0.00	0.00	0.00	0.00	0.00
동사	빈도	2	6	0	6	16	11	41
	백분율	14.29	33.33	0.00	30.00	24.24	42.31	25.79
형용사	빈도	4	0	2	0	0	0	6
	백분율	28.57	0.00	13.33	0.00	0.00	0.00	3.77
보조 용언	빈도	0	0	0	0	2	2	4
	백분율	0.00	0.00	0.00	0.00	3.03	7.69	2.52
지정사	빈도	0	0	0	2	2	0	4
	백분율	0.00	0.00	0.00	10.00	3.03	0.00	2.52
관형사	빈도	0	0	2	0	2	0	4
	백분율	0.00	0.00	13.33	0.00	3.03	0.00	2.52
부사	빈도	2	2	2	0	2	0	8
	백분율	14.29	11.11	13.33	0.00	3.03	0.00	5.03
감탄사	빈도	0	0	0	2	0	0	2
	백분율	0.00	0.00	0.00	10.00	0.00	0.00	1.26

접사	빈도	0	0	3	0	6	7	16
	백분율	0.00	0.00	20.00	0.00	9.09	26.92	10.06
소계	빈도	14	18	15	20	66	26	159
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

#### ④ 오류 층위의 문법 영역

##### ○ 전체 말뭉치

<표 71> 오류 주석 결과의 분석: 전체 말뭉치의 오류 층위-문법

유형	구분	1급	2급	3급	4급	5급	6급	합계
조사	빈도	421	499	554	321	348	273	2,416
	백분율	58.72	46.16	49.69	44.83	53.95	49.46	50.06
어미	빈도	174	368	388	222	216	215	1,583
	백분율	24.27	34.04	34.80	31.01	33.49	38.95	32.80
높임	빈도	4	2	6	4	1	0	17
	백분율	0.56	0.19	0.54	0.56	0.16	0.00	0.35
시제	빈도	27	119	31	36	13	24	250
	백분율	3.77	11.01	2.78	5.03	2.02	4.35	5.18
사동	빈도	0	0	1	7	0	2	10
	백분율	0.00	0.00	0.09	0.98	0.00	0.36	0.21
피동	빈도	0	6	3	9	1	7	26
	백분율	0.00	0.56	0.27	1.26	0.16	1.27	0.54
부정	빈도	4	15	12	3	2	0	36
	백분율	0.56	1.39	1.08	0.42	0.31	0.00	0.75
표현 문형	빈도	3	9	21	23	12	3	71
	백분율	0.42	0.83	1.88	3.21	1.86	0.54	1.47
문장	빈도	84	63	99	91	52	28	417
	백분율	11.72	5.83	8.88	12.71	8.06	5.07	8.64
소계	빈도	717	1,081	1,115	716	645	552	4,826
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

○ 문어 말뭉치

<표 72> 오류 주석 결과의 분석: 문어 말뭉치의 오류 층위-문법

유형	구분	1급	2급	3급	4급	5급	6급	합계
조사	빈도	375	471	519	299	292	245	2,201
	백분율	57.60	45.77	49.90	44.23	52.24	48.42	49.34
어미	빈도	164	358	364	204	194	199	1,483
	백분율	25.19	34.79	35.00	30.18	34.70	39.33	33.24
높임	빈도	4	2	2	4	1	0	13
	백분율	0.61	0.19	0.19	0.59	0.18	0.00	0.29
시제	빈도	21	109	29	36	13	24	232
	백분율	3.23	10.59	2.79	5.33	2.33	4.74	5.20
사동	빈도	0	0	1	7	0	2	10
	백분율	0.00	0.00	0.10	1.04	0.00	0.40	0.22
피동	빈도	0	4	3	9	1	5	22
	백분율	0.00	0.39	0.29	1.33	0.18	0.99	0.49
부정	빈도	4	15	12	3	2	0	36
	백분율	0.61	1.46	1.15	0.44	0.36	0.00	0.81
표현 문형	빈도	3	9	19	23	8	3	65
	백분율	0.46	0.87	1.83	3.40	1.43	0.59	1.46
문장	빈도	80	61	91	91	48	28	399
	백분율	12.29	5.93	8.75	13.46	8.59	5.53	8.94
소계	빈도	651	1,029	1,040	676	559	506	4,461
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

○ 구어 말뭉치

<표 73> 오류 주석 결과의 분석: 구어 말뭉치의 오류 층위-문법

유형	구분	1급	2급	3급	4급	5급	6급	합계
조사	빈도	46	28	35	22	56	28	215
	백분율	69.70	53.85	46.67	55.00	65.12	60.87	58.90
어미	빈도	10	10	24	18	22	16	100
	백분율	15.15	19.23	32.00	45.00	25.58	34.78	27.40
높임	빈도	0	0	4	0	0	0	4
	백분율	0.00	0.00	5.33	0.00	0.00	0.00	1.10
시제	빈도	6	10	2	0	0	0	18
	백분율	9.09	19.23	2.67	0.00	0.00	0.00	4.93
사동	빈도	0	0	0	0	0	0	0
	백분율	0.00	0.00	0.00	0.00	0.00	0.00	0.00
피동	빈도	0	2	0	0	0	2	4
	백분율	0.00	3.85	0.00	0.00	0.00	4.35	1.10
부정	빈도	0	0	0	0	0	0	0
	백분율	0.00	0.00	0.00	0.00	0.00	0.00	0.00
표현 문형	빈도	0	0	2	0	4	0	6
	백분율	0.00	0.00	2.67	0.00	4.65	0.00	1.64
문장	빈도	4	2	8	0	4	0	18
	백분율	6.06	3.85	10.67	0.00	4.65	0.00	4.93
소계	빈도	66	52	75	40	86	46	365
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

⑤ 오류 층위의 담화 영역

○ 전체 말뭉치

<표 74> 오류 주석 결과의 분석: 전체 말뭉치의 오류 층위-담화

유형	구분	1급	2급	3급	4급	5급	6급	합계
지시	빈도	1	2	0	0	3	1	7
	백분율	11.11	5.26	0.00	0.00	23.08	14.29	6.67
접속	빈도	8	6	7	4	0	1	26
	백분율	88.89	15.79	36.84	21.05	0.00	14.29	24.76
담화 표지	빈도	0	2	0	5	1	2	10
	백분율	0.00	5.26	0.00	26.32	7.69	28.57	9.52
구어/ 문어	빈도	0	28	12	10	9	3	62
	백분율	0.00	73.68	63.16	52.63	69.23	42.86	59.05
소계	빈도	9	38	19	19	13	7	105
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

○ 문어 말뭉치

<표 75> 오류 주석 결과의 분석: 문어 말뭉치의 오류 층위-담화

유형	구분	1급	2급	3급	4급	5급	6급	합계
지시	빈도	1	2	0	0	3	1	7
	백분율	11.11	5.26	0.00	0.00	23.08	14.29	6.67
접속	빈도	8	6	7	4	0	1	26
	백분율	88.89	15.79	36.84	21.05	0.00	14.29	24.76
담화 표지	빈도	0	2	0	5	1	2	10
	백분율	0.00	5.26	0.00	26.32	7.69	28.57	9.52
구어/ 문어	빈도	0	28	12	10	9	3	62
	백분율	0.00	73.68	63.16	52.63	69.23	42.86	59.05
소계	빈도	9	38	19	19	13	7	105
	백분율	100.00	100.00	100.00	100.00	100.00	100.00	100.00

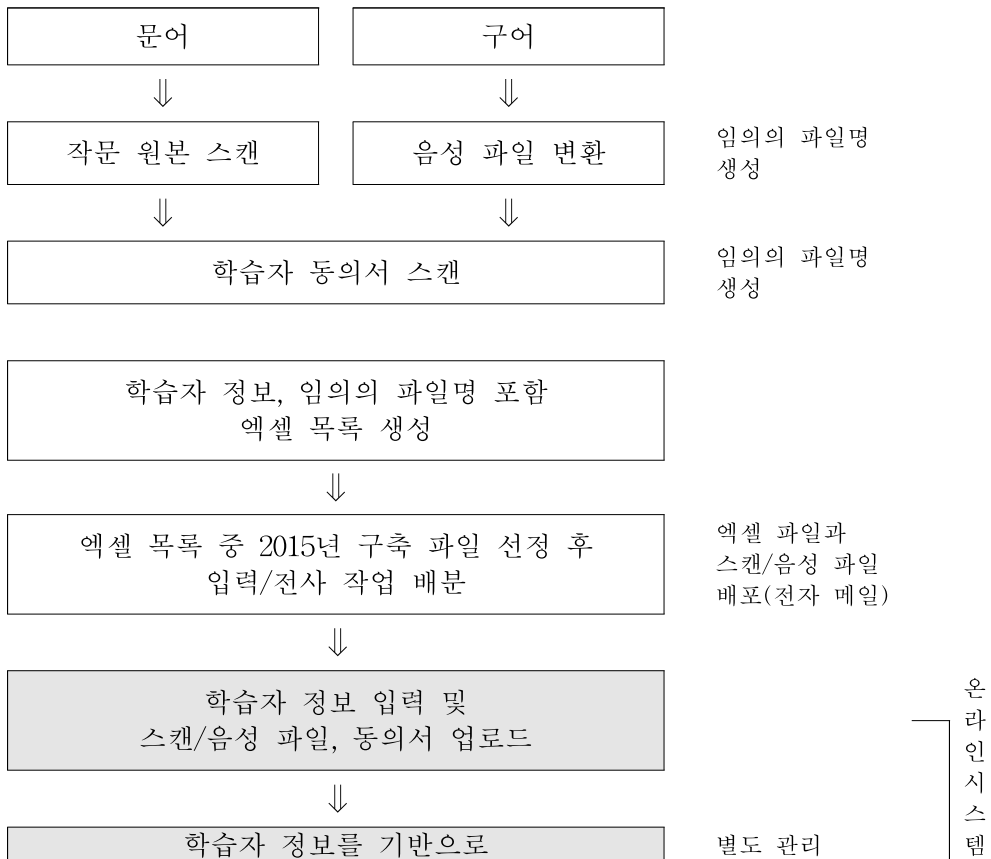
○ 구어 말뭉치

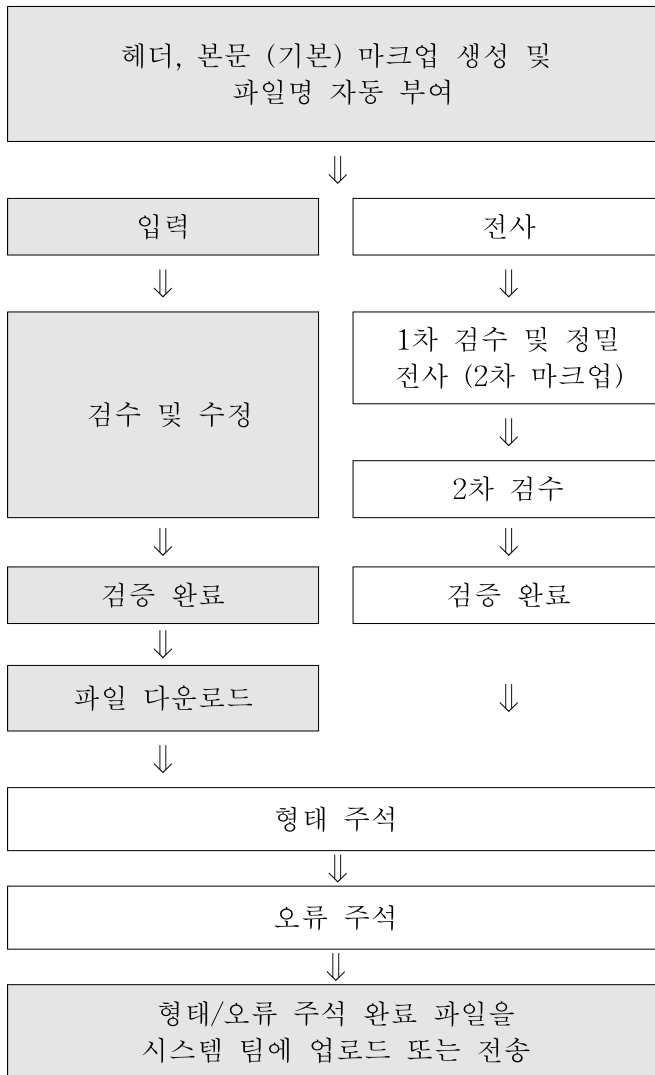
구어 자료의 경우 담화 오류는 나타나지 않았다.

### 3. 온라인 구축 시스템 개발을 위한 협의 작업

#### 3.1. 한국어 학습자 말뭉치 구축과 온라인 구축 시스템의 연계 모형

○ 한국어 학습자 말뭉치는 온라인 구축 시스템을 활용하여 수집, 구축, 가공의 절차를 체계화하는 것을 목표로 하고 있다.





자료로  
프로그램 팀에  
넘기거나  
내부에서 백업

에  
서  
작  
업  
예  
정

○ 온라인 구축 시스템을 활용한 작업의 범위는 다음과 같다.

<표 76> 온라인 구축 시스템을 활용한 작업의 범위

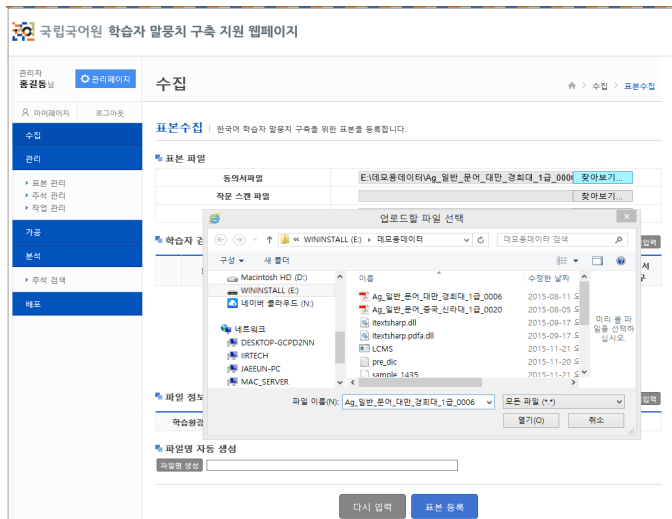
말뭉치 유형	절차	작업 범위 및 도구의 주요 기능
원시 말뭉치	학습자 동의서, 작문 스캔 파일, 음성 파일 업로드	온라인 구축 시스템 (☞ 주요 기능: ① 온라인 구축 시스템 업로드 전에 임 의로 부여된 스캔/음성 파일, 동의서 파일 이름 변환 (rename) 또는 기존 파일명 을 그대로 두고 향후 생성될 관련 파일과 연동 ② 헤더, 본문 (기본)마크업 생성, 파 일명 자동 부여 ③ 검증요청 - 검수(수정) - 관리자 검증으로 단계별 작업 관리 기능
	학습자 정보를 기반으로 헤더, 본문 (기본) 마크업 생성, 파일명 생성	
	문어 자료 입력 및 검수 (※ 구어는 공개용 전사 도구 사용)	
형태 주석 말뭉치	형태소 분석	온라인 구축 시스템 (☞ 주요 기능: 1차 주석 및 검수가 완료된 파일을 시스 템 내에서 형태 주석한 후 형태 주석 태그 세트를 참조 하면서 수정. 오류 주석 세 트와 연동)
오류 주석 말뭉치	오류 주석 및 검수	온라인 구축 시스템 (☞ 주요 기능: 형태 주석 태그 세트와 연동하여 시스 템 내에서 오류 주석 태그 세트를 참조하면서 수정) 개방형 주석 체계를 활용하 여 확장 주석이 가능한 주석 편집 애플리케이션(공개용)

## 3.2. 한국어 학습자 말뭉치 구축 도구의 활용

- 현재 온라인 구축 시스템 개발이 1차 완료되어 구축된 자료를 업로드하였으며, 시험 사용을 통해 문제점들을 보완한 후 2016년에는 본격 가동할 예정이다. 구축 도구는 크게 작업 도구(파일 등록, 입력, 형태 주석, 오류 주석)와 관리 도구(파일 관리, 작업 배분 및 인력 관리 등)로 나눈다. 작업 도구에서의 주요 기능을 중심으로 한 화면을 보이면 다음과 같다<sup>10)</sup>.

### 1) 파일 등록 기능

- 스캔이 된 작문 자료 또는 음성 파일을 업로드하는 기능. 업로드 후 입력, 형태 주석, 오류 주석 등의 작업을 수행할 수 있다.



<그림 22> 파일 업로드 화면 예시

### 2) 파일 정보 입력 및 파일명 생성 기능

- 학습자 정보를 입력하여 파일명을 자동으로 생성한다. 입력된 학습자 정

10) 한국어 학습자 말뭉치 구축 도구에 관한 세부 사항은 <2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구> 보고서 참고.

보는 헤더 마크업 자동 생성에도 활용된다.

<그림 23> 학습자 개인 정보 입력 화면 예시

<그림 24> 파일 정보 입력 화면 예시



### 3.3. 2016년 한국어 학습자 말뭉치 구축 도구 활용 및 추가 기능 제안

- 2015년에는 시험 구축과 함께 본격적인 구축 작업의 편이성과 효율성 제고하기 위한 구축 지원 도구 개발이 병행되었다. 2016년에는 기구축된 자료의 탑재와 함께 구축 도구를 본격적으로 사용할 예정이다. 그 과정에서 고려할 사항은 다음과 같다.
  - 수집, 구축, 가공 단계 간의 연계: 파일 간의 연계 및 통합 관리
  - 구축 지침의 업데이트에 따른 도구의 업데이트 및 버전 관리: 업데이트 내용과 처리 방식의 체계화
  - 지속적인 버그 수정 및 기능 강화를 통한 구축 도구 사용 환경의 안정화
- 아울러 보다 선진화된 통합형의 말뭉치 구축과 실제 구축 과정에서의 효율성 제고를 위해 다음의 도구 포함 여부를 제안한다. 이는 도구의 효용성과 예산, 시스템 개발 팀과의 협의 등을 고려하여 최종적으로 확정되어야 할 것이다.
  - 구어 전사 지원 도구: 음성 인식(개발 가능성과 실효성에 대한 협의 필요), 마크업 등 전사 공정의 반자동화를 위한 소프트웨어
  - 문어 입력 지원 도구: OCR, 띄어쓰기 자동 교정 소프트웨어

## 4. 말뭉치 구축/가공 인력 실무 교육 및 홍보

### 4.1. 단계별 수집 지침 배포 및 온/오프라인 교육, 워크숍

- 본 연구에서는 말뭉치 수집, 입력 및 전사, 가공의 각 단계에서 필요한 지침을 개발하여 실무자에게 배포하였다. 아울러 온라인/오프라인, 정기/비정기적으로 지침을 교육하고 워크숍을 실시하였다.

<표 77> 말뭉치 구축/가공 인력 실무 교육 및 홍보 내용

대상	교육 내용	교육 방법
수집 실무자	○ 수집 및 자료 처리 지침 교육	전자메일, 전화, 온라인 대화
자료 처리 실무자	○ 수집 자료 처리 및 파일 관리 ○ 지침 교육	정기/비정기 워크숍 및 세미나
입력 및 전사 작업자	○ 입력 및 전사 지침 교육 ○ 입력 및 전사 연습 ○ 작업 관련 쟁점 토론 워크숍	정기/비정기 워크숍 및 세미나
오류 주석 실무 작업자	○ 수집 및 자료 처리 지침 교육 ○ 오류 주석 연습 ○ 작업 관련 쟁점 토론 워크숍	정기/비정기 워크숍 및 세미나

## 4.2. 학술대회 발표

- 국제한국어교육학회 제25차 국제학술대회에서 특별 기획 연구로 발표
  - 제목: 한국어 학습자 말뭉치 구축을 위한 기초 연구
  - 발표: 강현화(연구 책임자)
  - 토론: 최은규(서울대), 조남호(명지대)

## V. 말뭉치 활용 방안 연구

### 1. 사용자 집단별 활용 모형

#### 1.1. 연구자를 위한 활용 모형

##### 1) 활용 분야

##### (1) 학습자 언어 연구

- 학습자 언어에 연구는 주로 학습자가 산출한 언어 자료를 바탕으로 한 오류 분석, 중간언어 연구, 언어 발달 및 습득에 관한 연구를 할 수 있다. 다음은 선행 연구에서 학습자 자료를 활용하여 이루어진 연구의 동향을 분석하여 정리한 것이다. 대부분의 선행 연구에서 연구자가 직접 구축한 소규모의 자료를 주로 사용하여 연구 결과를 일반화하거나 신뢰하기가 어려운 측면이 있었는데, 대규모의 균형성을 갖춘 한국어 학습자 말뭉치를 활용하여 다음과 같은 연구를 수행함으로써 그러한 한계들을 보완할 수 있다.

<표 78> 한국어 학습자 말뭉치를 활용한 연구 사례

세부 주석 항목		주석 말뭉치를 활용해서 가능한 연구
철자		맞춤법, 띄어쓰기 오류
발음	음소	음소 인지에 관한 오류, 어두 폐쇄음 발음(평음, 경음, 격음) 오류 종성 발음 오류 분석 단모음 오류, 이중모음 오류 받침 발음 습득, 어중 자음군 폐쇄지속 시간
	음절	발음과 형태 오류 음운규칙 오류
	음운규칙	연음규칙 오류, 비음화 오류, 연음규칙의 습득

세부 주석 항목		주석 말뭉치를 활용해서 가능한 연구
	초분절음소	강세구와 음장의 습득, 초분절 요소에 의한 유창성 평가, 억양(강세구와 문말 억양) 습득
어휘	단일어	한자어 오류, 양태부사 오류, 정도부사 오류, 품사별 어휘 사용 양상, 개별 동사의 의미 사용 양상, 수 분류사의 사용 양상, 어휘의 다양도, 평균 발화 길이(MLU)와 어휘
	합성어	합성어 사용 양상
	파생어	명사+접미사 구성의 어휘 사용 오류
	관용표현	관용표현 사용 양상
	언어 관계	서술성 언어 사용 오류
	상용구	[명사+동사] 상용어구 사용 양상
	속담	속담 오류
문법	격조사	조사(격조사) 오류 조사와 동사 구성 오류, 무정성 주어와 타동사 구문 오류, 조사 중첩 사용 양상,
	보조사	보조사 사용 양상
	접속조사	접속조사 사용 양상
	연결어미	어미(연결어미, 관형사형 어미) 오류, 종속 접속문 구성 오류, 동사의 연결 구성 사용 양상, 숙달도별 연결어미의 정확도 측정, 통사적 숙달도 측정, 내포문 오류, 명사구 보문절 습득, 관계절 습득
	종결어미	해라체 평서문 사용 양상
	선어말어미	선어말어미의 중첩 사용 양상
	높임	높임법 오류
	시제	시제 오류, 시상 오류, 복문에서의 시제 사용, 시상의 담화 기능 사용 양상, 시제 표현 문법 발달 패턴
	사동	사동 오류
	피동	피동 표현 오류
	부정	부정 표현 오류, 부정문 습득
	표현문형	문법적 언어 오류, 양태 표현 사용 양상

세부 주석 항목		주석 말뭉치를 활용해서 가능한 연구
	문장	문장 성분, 호응 관계
담화	지시	지시사 오류, 응결 장치 습득
	접속	접속 부사 및 접속 표지 오류, 인과관계 접속 표현 사용 양상, 응결 장치 습득
	담화표지	학술 논문의 담화표지 사용 양상, 토론 담화표지 습득 양상
	화행	문법 표현의 화용적 오류/실패, 요청 화행의 공손 의미 사용 양상, 요청 화행에서의 어휘적 완화 장치 습득, 요청 화행의 적절성의 문제, 응답 화행 습득
	기타	문체 오류, 이야기 구술 수행 담화의 특징

## (2) 교수요목 설계 및 교재 개발

- 오류 분석과 중간언어 분석 결과를 토대로 학습자의 언어 습득의 순서를 밝히고 그에 따라 어휘 또는 문법 제시 순서를 결정할 수 있다. 또한 교재 개발 시 학습자가 자주 일으키는 오류 정보를 활용하여 문법 및 어휘 정보를 기술할 수 있다.

## (3) 언어권별 오류를 활용한 사전의 어휘 정보 기술

- 학습자 말뭉치는 학습자 사전 편찬이나 사전의 기술에 활용될 수 있다. 개방형지식대사전 구축 사업의 일환으로 개발된 <한국어 기초사전>은 한국어 학습자를 위한 웹 기반의 학습 사전으로 5만여 개의 표제어가 수록되어 있다. 현재는 일반적인 어휘 정보가 중점적으로 기술되어 있는데, 학습자 말뭉치를 활용하여 학습자들이 생산한 오류를 예시할 수 있다. 그럼으로써 사전 사용자들이 그러한 오류를 범하지 않도록 방지하는 데 기여할 것이다. 또한 학습자에게 유의미한 표제어를 선정하고 용례를 제시하기 위한 기초 자료로 활용할 수 있다.

#### (4) 외국인 학습자를 위한 한국어 용법 사전

- 오류 주석 말뭉치를 활용하여 장르별, 주제별 작문 또는 말하기 용법 사전을 편찬할 수 있다. 발음, 어휘, 문법, 담화 층위에서 보편적으로 발생하는 오류(common error)를 제시하고 올바른 글쓰기 또는 말하기를 위한 용법을 제시한다.

#### (5) 영역별 오류 사전

- 음운, 형태, 통사, 담화 층위의 오류를 영역별로 추출하여 다양한 오류 현상을 제시하고, 오류의 처방과 예방을 설명과 연습 활동을 제시하여 학습을 도울 수 있다. 한국어 학습자에게서 보편적으로 발생하는 오류, 언어권별, 숙달도별로 자주 발생하는 오류를 중심으로 기술할 수도 있다.

### 2) 말뭉치의 활용

#### (1) 기본 주석 또는 말뭉치 활용 도구를 사용한 분석

- 빈도 정보  
음소, 형태소, 음절, 어절, 표현문형 등의 다양한 층위의 언어 단위에 나타난 오류 빈도, 오류와 비오류의 비율 등을 분석하여 오류의 양상, 중간언어 발달 정도와 언어 습득 순서 등을 진단할 수 있다. 또한 학습자 변인, 자료 변인 등에 의한 오류 빈도를 비교하여 집단별 언어 사용 양상을 파악할 수 있다.
- 타입(Type)과 토큰(Token) 정보  
학습자가 산출한 자료의 어휘 타입과 토큰을 분석함으로써 어휘의 다양성, 어휘 밀집도를 알 수 있다.
- 키워드 분석  
키워드 분석은 분석 말뭉치에서 특징적으로 많이 사용된 어휘를 추출하

는 것이다. 한국어 모어 화자와 학습자, 국적, 수준, 학습 목적과 같은 변인이 다른 학습자 간, 텍스트의 주제, 장르가 다른 자료 간의 어휘 사용 양상을 분석해 볼 수 있다.

- 연어관계(collocation) 분석

연어관계 분석을 통해 학습자가 어휘 사용 양상을 전반적으로 관찰할 수 있다. 특히, 일정한 범위 내에서 연쇄 또는 비연쇄하여 공기하는 어휘들이 무엇인지를 파악함으로써 어휘 사용 능력을 측정해 볼 수 있다.

- 군집(cluster) 분석

군집 분석을 통해 언어 또는 표현문형 사용 양상을 파악할 수 있다. 특히, 연쇄 구성을 이루는 구 단위 이상의 어휘 또는 문법 표현 사용 능력을 측정할 수 있다.

## (2) 개방형 주석 체계를 활용한 연구자의 추가 분석

- 개방형 주석 도구를 사용하여 연구 목적에 따라 확장 주석을 할 수 있다. 한국어 학습자 말뭉치 배포 시스템에서 제공하는 주석 도구를 활용하여 연구 목적에 따라 주석을 가감한다. 예를 들어, 한국어 학습자 말뭉치에서는 상세한 억양과 강세 등은 주석 대상에서 제외하였는데, 음성 파일을 직접 들으면서 추가 주석할 수 있다.

## 3) 활용 조건

### (1) 자료의 공개 범위

- 작문 및 전사 텍스트 전문
- 음성 파일과 엘란 전사 원본 파일

## (2) 자료 활용을 위한 도구(애플리케이션)의 지원

- 형태 주석 편집 도구
- 확장 주석을 위한 주석 편집 도구와 사용 설명서
- 학습자 말뭉치에서 사용한 엘란 또는 엘란과 호환성을 가진 전사 도구, 사용 설명서

## 1.2. 한국어 교사를 위한 활용 모형

### 1) 활용 분야

#### (1) 오류의 진단과 처방

- 오류 주석 말뭉치를 활용한 오류 통계는 한국어 학습자의 언어 사용 양상에 대한 관찰하도록 해 준다. 직접적으로는 학습자의 오류를 진단하고 다양한 유형의 오류 양상을 바탕으로 합리적인 오류의 예방과 처방을 할 수 있다. 이때 학습자의 숙달도 단계, 국적 등의 변인별 오류 양상을 관찰하여 집단별 언어 사용 양상을 파악하고 각각에 적합한 맞춤형 처방을 내릴 수 있다.

#### (2) 언어권별 교수 방법 개발

- 오류 주석 말뭉치를 활용하여 언어권별 발음, 어휘, 문법, 담화 오류의 양상을 파악하고 보다 효과적인 교수법을 구안하는 데에 활용할 수 있다. 가령, 언어권별 학습자에 따라 차별적 교수 방법을 제시할 수 있으며 조사, 어미와 같은 문법 항목, 어휘 등의 제시 순서를 결정할 수 있다. 뿐만 아니라 언어권별 자료 분석을 통해 모국어의 간섭으로 인한 오류의 양상을 정확하고 이해하기 쉽게 교수할 수 있다.

### (3) 수업 활동 및 자료 개발

- 오류 정보를 활용하여 수업 활동 및 수업 자료를 개발할 수 있다. 학습자가 오류 자료를 통해 오류의 양상을 인식하고 학습 항목의 용법을 스스로 발견하고 주도적으로 배우게 함으로써 학습에서 습득으로의 전환을 용이하게 한다.

### (4) 오류 정보를 활용한 평가 문항 개발

- 성취도 평가, 숙달도 평가 문항 개발 시 학습자가 공통적으로 일으킨 오류들을 활용할 수 있다. 이때 오류를 포함한 문맥 색인을 활용할 수 있다.

## 2) 말뭉치의 활용

- 문맥 색인 검색  
문맥 색인은 특정 키워드를 포함한 용례를 검색하는 것으로 수업 활동이나 수업 자료 개발, 어휘나 문법 정보 기술 등에 활용할 수 있다.
- 빈도 정보
- 타입과 토큰 정보

## 3) 활용 조건

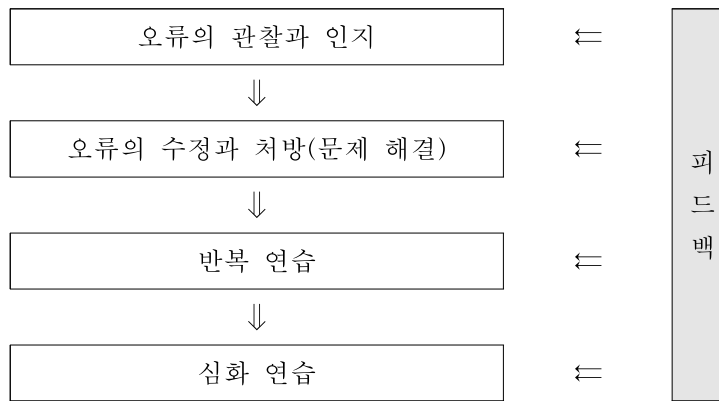
- 한국어 학습자 온라인 구축 시스템의 검색기 제공 정보

## 1.3. 학습자를 위한 활용 모형

### 1) 활용 분야

- 한국어 학습자 말뭉치는 학습자가 오류의 발견과 관찰, 수정 활동을 통한

자기 주도적 학습의 기초 자료로 활용할 수 있다. 컴퓨터 보조 언어학습(Computer Assisted Language Learning)은 컴퓨터 프로그램 또는 컴퓨터 기능을 언어 교수·학습자의 매체로 활용하는 방법이다. 컴퓨터 기술의 발전과 인터넷의 확산으로 최근 주목받는 교수 방법 중 하나로 행동주의와 구성주의 이론을 기반으로 한다. 이에 따라 학습자 말뭉치를 데이터베이스로 한 학습 도구를 사용하여 학습자가 오류를 인지하고, 그 원인을 탐색하여 깨닫는 과정을 반복함으로써 보다 자기주도적이고 효율적인 학습을 할 수 있다. 학습자 말뭉치를 활용한 학습 절차 및 운용 원리는 다음과 같다.



<그림 27> 학습자 말뭉치를 활용한 학습 절차

- **오류의 관찰과 인지:** 학습자가 산출한 자료를 관찰하고 그 안에서 포함된 오류를 인지한다.
- **오류의 수정과 처방(문제 해결):** 오류의 원인과 수정, 또는 오류를 예방하기 위한 문제 해결 방법을 스스로 탐색하여 깨닫는다.
- **반복 연습을 통한 자동화:** 유사한 패턴의 오류를 포함한 자료를 반복적으로 접하고 연습 활동을 함으로써 학습 내용을 완전히 이해한다.
- **심화 연습을 통한 강화:** 난이도가 높은 연습 활동을 통해 학습 내용을 확인하고 강화한다.
- **즉각적인 피드백:** 각 활동의 단계에서 오류/비오류 데이터베이스에 기초한 즉각적인 피드백이 주어진다.

## 2) 말뭉치의 활용

### ○ 문맥 색인

문맥 색인 검색을 통해 오류를 포함한 용례를 검색하여 수준별, 언어권별로 자주 일으키는 오류의 양상을 비교해 본다. 비오류를 포함한 용례와의 비교를 통해 스스로 오류를 수정할 수 있다.

### ○ 빈도

특정 오류의 빈도, 국적이나 수준 등의 변인별 빈도 비교를 통해 오류 양상을 정확히 파악하고 자신의 학습 정도, 발달 단계를 스스로 진단할 수 있다.

### ○ 키워드

키워드 분석을 통해 특정 주제, 특정 단계, 특정 장르에서 자주 사용되는 어휘를 참조할 수 있다. 참조 말뭉치와의 비교를 통해 관찰하고자 하는 텍스트의 특성을 파악함으로써 맥락에 맞는 어휘를 풍부하게 사용하여 텍스트를 구성할 수 있다.

## 3) 활용 조건

- 검색 기능과 함께 반복 학습, 피드백 기능을 갖춘 말뭉치 기반의 학습 도구 개발이 요구됨

## 2. 말뭉치 활용을 위한 검색 기능 설계(안)

### 2.1. 말뭉치 검색 기능 요약

- 한국어 학습자 말뭉치가 사용자 집단별로 활용되기 위해서는 원시 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치가 가진 고유의 속성을 고려한 검색 기능이 잘 설계되어야 한다. 다음은 앞서 제시한 말뭉치의 활용을 위해 한국어 학습자 말뭉치 구축 시스템 개발 팀과 연계하여 개발해야 할 검색 기능이다.

<표 79> 말뭉치 검색 기능 설계(안)

		원시 말뭉치	형태 주석 말뭉치	오류 주석 말뭉치
검색 단위	억양			○
	음소	○		○
	음절	○		○
	형태소		○	○
	어절	○		
	표현 문형		○	○
검색 조건	자료 유형별	○	○	○
	기관별	○	○	○
	언어권별	○	○	○
	수준별	○	○	○
	등급별	○	○	○
검색 범위	오류	○	○	○
	오류+비오류	○	○	○
(통계 기반의) 검색 정보	문맥 색인	○	○	
	연어	○	○	
	군집(cluster)	○	○	
	키워드	○	○	
	사용어휘 목록	○	○	
검색 결과	음소(자소)통계	○	○	○
	음절 통계	○	○	○
	어절 통계	○	○	○
	형태소 통계	○	○	○
	표현문형 통계	○	○	○
	파일별 통계	○	○	○
	변인에 따른 집단별 통계	○	○	○

## 2.2. 검색 기능에 관한 세부 내용

### 1) 검색 단위

- 한국어 학습자 말뭉치는 원시 말뭉치, 형태 말뭉치, 오류 주석 말뭉치가 담고 있는 주요 정보에 그에 따라 주어지는 속성에 의해 검색 단위가 정해진다. 원시 말뭉치는 기본적으로 가공되지 않은 자료로 음소, 음절, 어절 등의 기본적인 언어 단위로 검색이 이루어진다. 형태 주석 말뭉치는 형태소 단위로 주석이 붙여진 자료로 형태 단위의 검색을 기본으로 한다. 여기에 두 개 이상의 형태소가 결합된 표현문형이 제한된 목록 내에서 검색 가능하다. 오류 주석 말뭉치는 오류 주석 체계에 포함된 억양, 음소, 음절, 형태소, 표현문형 단위로 검색이 가능하다.

### 2) 검색 조건

- 한국어 학습자 말뭉치는 학습자 변인, 자료 수집 과제 활동, 파일 이력에 관한 정보가 파일 단위로 기록되어 있다. 이들 정보는 사용자가 특정 집단의 자료를 제한하여 검색하도록 하는 데에 활용된다. 다음은 검색 조건으로 적용 가능한 정보들이다.

<표 80> 조건 검색의 세부 변인

조건 범주	세부 조건
자료 유형별	문어, 구어 / 횡적, 종적 / 과제 유형별 장르별 / 주제별
학습 목적 및 대상	일반, 학문목적, 이주여성 / 교포, 비교포
기관별	국내, 해외 / 지역별 / 기관별(익명)
언어권별	중국어, 일본어, 영어 등의 언어권
수준별	초급, 중급, 고급, (최고급)
등급별	1, 2, 3, 4, 5, 6, 7급 이상
기타	성별, 연령 등 학습자의 개인 정보 학습 기간, 사용 가능한 외국어 등의 학습 변인

### 3) 검색 범위

- 한국어 학습자 말뭉치는 오류 주석 결과를 기반으로 하여 검색 범위를 1) 오류 어절, 2) 오류 어절과 비오류 어절 전체로 지정할 수 있다. 1)은 학습자 오류 분석을 위한 것이고 2)는 중간언어를 살피기 위한 것이다. 원시 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치는 하나의 체계로 연계되어 있어 1) 또는 2)에서 특정 항목을 검색하면 검색 항목을 포함한 원문의 문장을 함께 참조하여 볼 수 있다.

### 4) 검색 정보

- 검색 정보는 음소, 음절, 형태소, 어절, 표현문형 단위의 오류, 오류+비오류 외에도 다음과 같은 정보들을 검색할 수 있다. 이들 정보는 학습자가 산출한 자료에 사용된 다양한 층위의 언어 요소들을 통계적 기법과 통합하여 추출하는 것으로 언어 사용의 유창성과 다양성을 체계적으로 살필 수 있도록 해 준다.

#### (1) 문맥 색인

- 문맥 색인은 특정 어휘나 형태소를 포함한 용례를 검색하는 기능이다. 오류를 포함한 용례를 검색하여 수업 활동이나 수업 자료 개발, 어휘나 문법 정보 기술 등에 활용할 수 있다.

☞ 다음은 글잡이Ⅱ를 사용하여 시험 구축된 한국어 학습자 말뭉치에서 ‘학교’를 포함한 용례를 추출한 결과이다. 학교를 중심으로 왼쪽과 오른쪽에 이어지는 내용을 볼 수 있으며, 특정 용례를 선택하면 상단에 앞뒤의 문맥을 포함한 원문이 보인다.

Geul3

저는 마뽀이예요. 수단에서 왔어요. 그리고 저는 학생이예요.  
학교에서 한국어를 공부해요.  
우리 교실에는 외국 사람이 많아요. 우리 선생님이 한국어를 잘 가르쳐 주세요. 한국어 공부는 재미있지만 어려워요.

번호	왼쪽 문맥	중심 패턴	오른쪽 문맥	원천파일
1		오늘	학교는	J:\부품예한부드...
2		그리고	한국어를 공부해요.	J:\부품예한부드...
3		그리고	학교에	J:\부품예한부드...
4	한국어센터는 어학당 분야에서 최고라고	학교입니다.	와요.	J:\부품예한부드...
5		학교	근처에 완충이 많아요.	J:\부품예한부드...
6		학교	마루에 이야기 많이 했대.	J:\부품예한부드...
7		학교와	집에서 규칙을 어긴 경향이 있다.	J:\부품예한부드...
8	경험</title> 나는 대학교였을 때 집에서	학교까지	멀어서 기숙사에서 살았는데 어느 날 반 전...	J:\부품예한부드...
9	그렇지만 다음 날	학교에서	기숙사에 돌아와서 방안에 불은 메모를 읽...	J:\부품예한부드...
10		우리 동아	있는 컴퓨터 센터 가서 컴퓨터를 이용한다.	J:\부품예한부드...
11	'같이 밖에 나가서	학교	근처에 있는 편의점 갈까?	J:\부품예한부드...
12	우리 뒤에서 물론	학교에	들어간다.	J:\부품예한부드...
13		학교	할난 후에 집에 가는 길에서 계속 우리 어...	J:\부품예한부드...
14	'앞으로	학교	규칙을 어기지 말아요' 말한다.	J:\부품예한부드...
15		학교	규칙을 어기면 느낌 안 좋다.	J:\부품예한부드...
16	그래서 앞으로 재가	학교	규칙을 꼭 잘 지킬 것이다.	J:\부품예한부드...
17		학교에	다녀요.	J:\부품예한부드...
18	그래서 우리는 다음 출퇴근에	학교에	가다가 우리 반 친구의 집에서 피자를 먹고...	J:\부품예한부드...
19	만하지 않다</title> 나는 친구 같이	학교	8급에 규칙을 한 번 어기겠었다.	J:\부품예한부드...
20	8급에 학생들은 쉬는 시간에	학교	근저 area에 가지 않았다.	J:\부품예한부드...
21	그래서	학교	안에서 쉬는 시간을 한다.	J:\부품예한부드...

<그림 28> 문맥 색인 분석 예시

## (2) 연어

- 연어 관계는 하나의 어휘가 다른 어휘와 공기하는 현상을 말한다. 즉, 중심어(node)와 어느 한 방향으로 공기하는 다른 어휘(span) 간의 관계로 일정한 범위 내에서 비연속적으로 발생하는 어휘 간의 공기 관계를 포함한다. 연어관계 분석을 통해 학습자의 어휘 사용 양상을 관찰할 수 있다.

☞ 다음은 Antconc3.4.4를 사용하여 시험 구축된 한국어 학습자 말뭉치에서 ‘학교’와 공기 관계를 이루고 있는 어휘들을 분석한 결과이다. 학교를 중심으로 왼쪽으로 한 어절, 오른쪽으로 두 어절까지의 어휘를 범위로 설정하였다. 함께 참조하여 본 문맥 색인을 보면 ‘학교 규칙’, ‘학교 기숙사 규칙’ 등이 사용되고 있음을 볼 수 있다.

File Global Settings Tool Preferences Help

AntConc 3.4.4w (Windows) 2014

Corpus Files

316W-4159231150

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Total No. of Collocate Types: 592 Total No. of Collocate Tokens: 1101

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	33	1	32	6.70590	규칙을
2	32	1	31	7.66151	근처에
3	18	0	18	3.58401	있는
4	11	0	11	6.55875	어디
5	11	0	11	5.38591	생활이
6	11	0	11	6.74019	기숙사
7	11	1	10	6.15028	앞에
8	9	0	9	7.72867	근처
9	8	0	8	8.97379	폭박이
10	7	0	7	7.44010	생활도
11	7	0	7	6.02982	기숙사에서
12	8	1	7	7.79321	규칙은
13	8	1	7	7.42124	규칙
14	6	0	6	6.44327	폭박
15	6	0	6	6.29571	앞에
16	5	0	5	9.88068	폭박은
17	6	1	5	3.37818	있어서
18	15	10	5	1.76347	있다
19	5	0	5	2.50216	아주
20	5	0	5	4.73500	생활을
21	5	0	5	4.71075	밖에
22	5	0	5	1.80066	때문에
23	5	0	5	9.61764	규치
24	11	7	4	3.07449	후에
25	4	0	4	9.88068	폭박의
26	5	1	4	4.16918	좋아요
27	4	0	4	8.55875	경문에
28	4	0	4	1.39183	잘
29	6	2	4	3.81818	있어요
30	4	0	4	8.71075	읽기

Search Term ☒ Words ☒ Case ☐ Regex

Window Span ☐ Same

From... 1L To... 2R

Min. Collocate Frequency

1

Sort by ☐ Invert Order

Sort by (Freq(R))

Files Processed

Clone Results

### <그림 29> 연어 분석 예시

[illegible]

<그림 30> 연어 분석 결과의 문맥 색인 정보 예시

### (3) 군집

- 군집(cluster)은 언어관계와 마찬가지로 어휘 간의 공기관계를 나타낸다. 다만 중심어(node)와 연쇄하여 나타나는 어휘들을 통계적으로 산출해 낸 것이라는 점에서 언어와 다소 차이가 있다. 군집 분석을 통해 학습자의 언어 또는 표현문형 사용 양상을 파악할 수 있다.

☞ 다음은 Antconc3.4.4를 사용하여 시험 구축된 한국어 학습자 말뭉치에서 3-gram을 추출한 결과이다. 한국어 학습자가 사용한 표현문형 목록을 살펴기 위해서 형태 주석 말뭉치를 분석 자료로 활용하였다. 원시 말뭉치를 분석 대상으로 하면 언어 구성을 살펴볼 수 있다.

Rank	Freq	Range	N-gram
1	544	1	= 수 있
2	330	1	하 수 수
3	230	1	하 있 다
4	226	1	는 것 이
5	194	1	수 있 다
6	178	1	하 는 것
7	169	1	= 것 이
8	160	1	하 있 예요
9	154	1	시행 틀이
10	151	1	다 고 생각 하
11	148	1	하 이야 하
12	147	1	것 이 다
13	144	1	고 실 습니다
14	141	1	하 고 실
15	137	1	기 때문 에
16	134	1	하 고 있
17	128	1	하 있 습니다
18	122	1	= 게 이
19	118	1	생각 하 ㅓ 다
20	115	1	고 실 다
21	114	1	= 수 없
22	113	1	고 있 다
23	110	1	게 이 예요
24	110	1	하 = 태
25	107	1	하 지 않
26	102	1	다 나 는
27	99	1	아 아 하 ㅓ 다
28	96	1	가 고 실
29	89	1	예 가 있

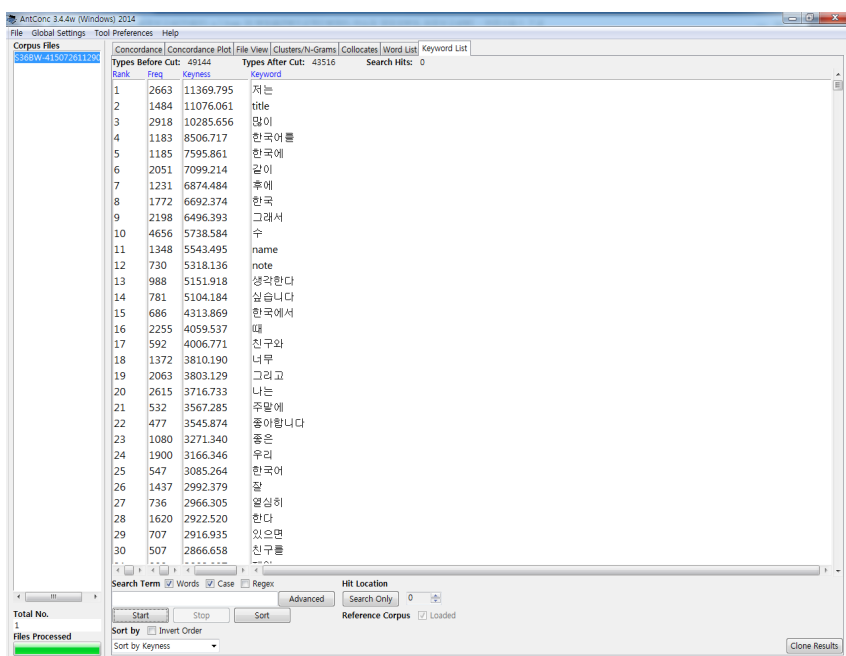
<그림 31> 군집 분석 예시

### (4) 키워드

- 키워드 분석은 학습자 말뭉치와 참조 말뭉치(예. 한국어 모어 화자 말뭉치, 학습 목적이나 대상별 특성이 다른 한국어 학습자 말뭉치)와의 비교

를 통해 분석 말뭉치에서 특징적으로 많이 사용된 어휘를 추출하는 것을 말한다. 키워드 분석을 통해 한국어 모어 화자와 학습자, 국적, 수준, 학습 목적과 같은 변인이 다른 학습자 간, 텍스트의 주제, 장르가 다른 자료 간의 어휘 사용 양상을 분석해 볼 수 있다.

- ☞ 다음은 Antconc3.4.4를 사용하여 한국어 모어 화자 말뭉치를 참조 말뭉치로 비교하여 시험 구축된 한국어 학습자 말뭉치에서 변별적으로 많이 사용된 어휘를 추출한 것이다. 원시 말뭉치를 참조 말뭉치와 분석 말뭉치로 사용하여 특징적으로 사용된 어절 단위의 표현이 분석되었는데, 형태 분석 말뭉치 또는 분석된 어휘 목록으로도 분석이 가능하다.



<그림 32> 키워드 분석 예시

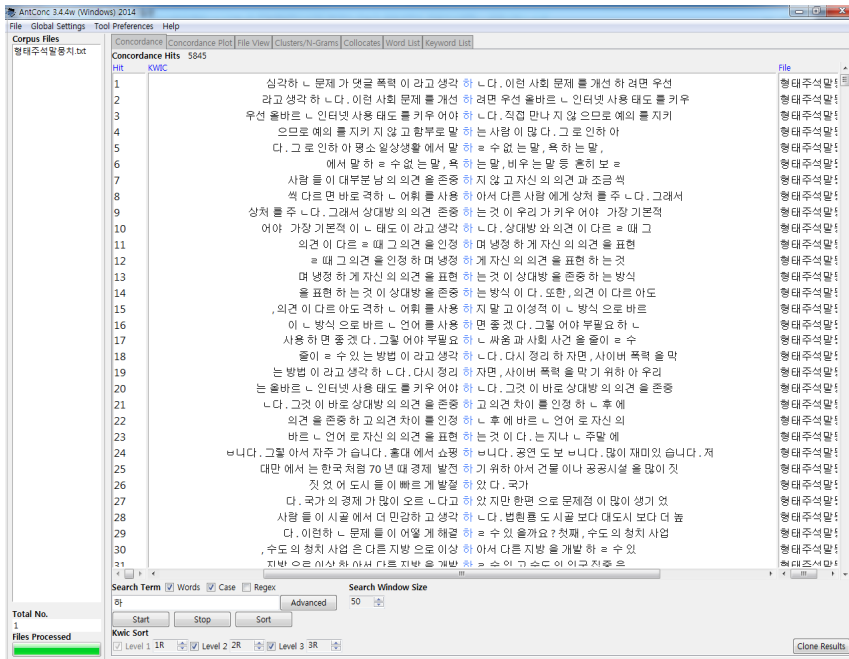
## (5) 사용 어휘 목록

- 사용 어휘 목록은 학습자가 산출한 자료에 사용된 어휘 목록을 빈도순으로 추출한 것이다. 말뭉치 전체, 수준별, 장르별, 주제별 등의 변인별 텍스트에서 고빈도로 사용된 어휘 목록을 분석해 볼 수 있다.

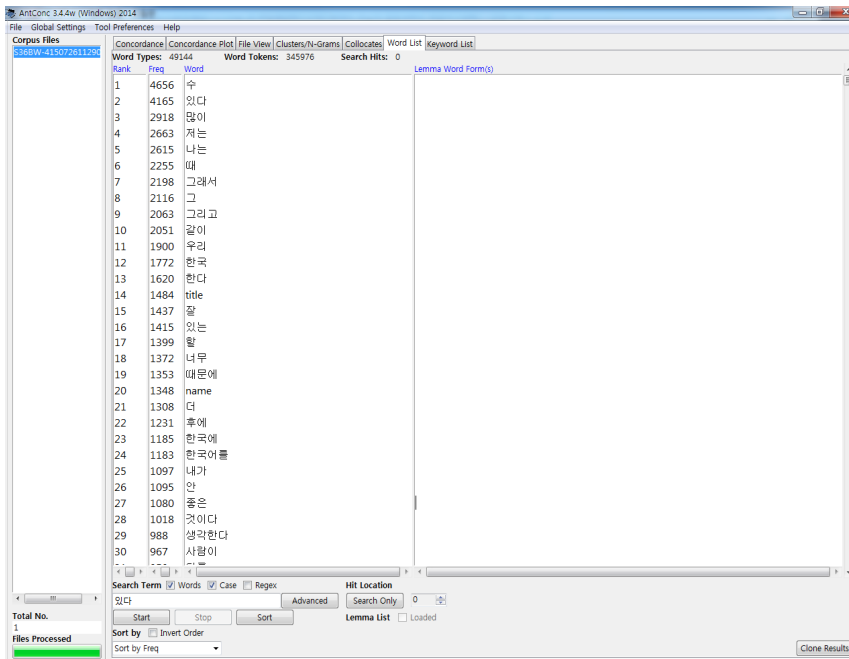
- ☞ 다음은 Antconc3.4.4를 사용하여 시험 구축된 학습자 말뭉치에서 사용된 어휘 목록을 추출한 결과이다. 형태 주석 말뭉치를 분석 대상으로 한 경우, 함께 참조하여 본 문맥 색인 자료를 보면 접사 ‘-하다’ 또는 동사 ‘하다’의 어간 ‘하-’가 가장 많이 사용되었음을 볼 수 있다. 한편, 원시 말뭉치를 분석 대상으로 한 경우, 의존명사 ‘수’, ‘있다’, ‘많이’, ‘저는’ 등이 고빈도로 나타났음을 볼 수 있다.

Rank	Freq	Word	Lemma Word Form(s)
1	5845	하	
2	5120	이	
3	3750	에	
4	3739	는	
5	3103	을	
6	2889	다	
7	2603	가	
8	2261	있	
9	2124	고	
10	1898	를	
11	1748	은	
12	1715	여	
13	1666	나	
14	1618	아	
15	1617	았	
16	1480	요	
17	1406	어요	
18	1361	에서	
19	1304	도	
20	1259	의	
21	1216	었	
22	1148	습니다	
23	1106	를	
24	1043	것	
25	848	나	
26	848	사람	
27	837	수	
28	823	그	
29	768	습니다	
30	765	아서	

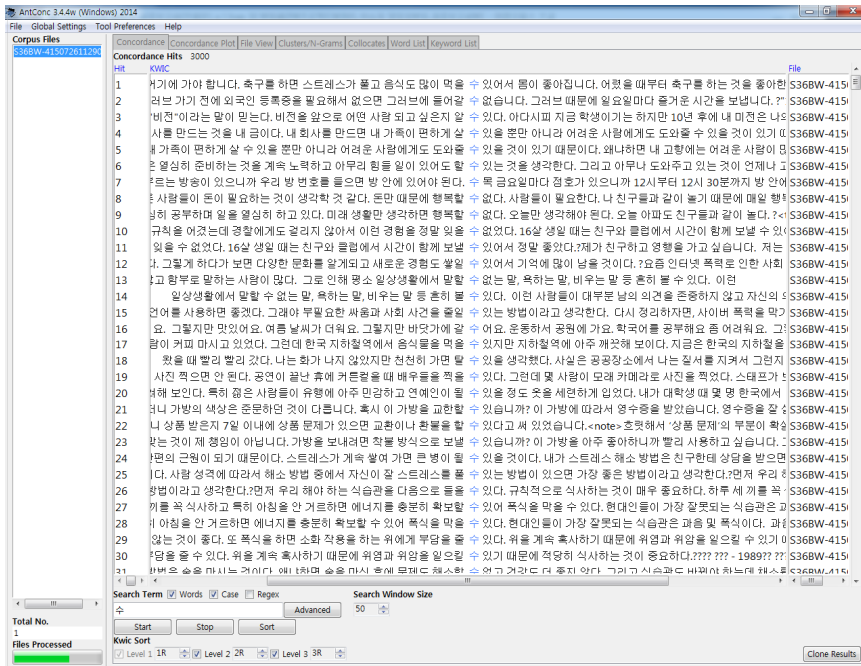
<그림 33> 사용 어휘 목록 예시: 형태 주석 말뭉치의 경우



<그림 34> 사용 어휘 목록의 문맥 색인 정보 예시: 형태 주식 말뭉치의 경우



<그림 35> 사용 어휘 목록 예시: 원시 말뭉치의 경우



<그림 36> 사용 어휘 목록의 문맥 색인 정보 예시: 원시 말뭉치의 경우

## 5) 검색 결과

- 검색 결과는 음소, 음절, 어절, 형태소, 표현문형 등의 검색 단위별, 파일별, 학습자의 수준이나 국적, 자료 유형 등의 변인에 따른 집단별 통계 결과를 빈도와 백분율, 그래프로 제시해 준다.

## VI. 결론 및 제언

### 1. 연구 요약

이 연구는 한국어 학습자 말뭉치 구축을 위한 기본 계획을 수립하고, 기초 연구를 통해 말뭉치 구축 지침을 마련한 후 그에 따라 약 35만 어절 규모의 말뭉치를 시험 구축하는 데에 주요한 목적이 있다. 그리고 말뭉치 구축과 가공에 필요한 실무 인력을 양성하고 향후 말뭉치가 널리 보급되어 실용적으로 활용될 수 있도록 홍보하고 활용 방안을 모색해 보는 데에 목적이 있다.

#### ○ 한국어 학습자 말뭉치 구축 기본 계획 수립

한국어 학습자 말뭉치는 3단계 6년간의 사업을 통해 총 370만 어절 규모의 말뭉치(문어 300만, 구어 70만 어절)를 수집한다. 수집 대상은 국내 학습자, 이주민, 국외 학습자이며, 구축 본부와 수집 기관의 직접 접촉을 통한 수집과 온라인 구축 시스템을 통한 자율적 수집의 두 가지 방법을 병행하도록 한다. 수집 자료는 학습자가 산출한 문어와 구어로 교육기관 말뭉치, 기획 말뭉치, 자연 발화 말뭉치이다. 수집 자료는 다시 수집 기간에 따라 횡적 말뭉치와 종적 말뭉치로 나뉜다.

#### ○ 한국어 학습자 말뭉치 구축 지침 수립

한국어 학습자 말뭉치 구축은 크게 수집, 구축(자료 분류 및 전산화 작업, 입력 및 전사), 가공(형태 주석, 오류 주석)의 단계를 거친다.

수집 지침은 자료의 공개와 활용, 학습자의 개인 정보 보호를 위한 IRB 준용, 교육과정 종속도를 낮추고 균형성을 담보하기 위한 자료 수집 방법을 주로 다루고 있다.

구축 지침은 수집 자료의 분류와 전산화, 관리를 주요한 내용으로 하는 자료 처리 지침과 문어 입력 지침, 구어 전사 지침으로 구분된다. 문어 입력 지침과 구어 전사 지침은 <21세기 세종 한국어 균형 말뭉치>의 지침을 근간으로 하되, 비모어 화자의 쓰기, 발화 특성을 고려하여 수정·보완하였다.

가공 지침은 형태 주석과 오류 주석 지침으로 나뉜다. 형태 주석 지침은 <21세기 세종 한국어 균형 말뭉치>의 지침을 근간으로 하되, 비모어 화자의 오류로 인한 분석 불능이나 오분석 자료 처리에 관한 사항을 추가하였다. 오

류 주석 지침은 주석 체계의 일관성과 활용상의 효용성을 고려하여 기본 주석과 확장 주석으로 나눈 뒤, 한국어 교육 분야에서 중요하게 다루어지는 항목들을 체계적으로 배열하였다.

### ○ 한국어 학습자 말뭉치 실제 수집, 구축, 가공

2015년 한국어 학습자 말뭉치는 앞선 단계에서 마련한 지침을 토대로 원시 말뭉치 35만 어절(문어 30만 어절, 구어 5만 어절), 형태 주석 말뭉치 22만(문어 20만, 구어 2만 어절), 오류 주석 5만 어절(문어 4만, 구어 1만)이 구축되었다. 실제 구축 과정에서 다음과 같은 사항이 추가적인 지침으로 논의되었는데, 이는 2016년 한국어 학습자 말뭉치 구축 사업을 통해 정비가 될 예정이다.

- 원시 말뭉치: 온라인 구축 지원 도구의 주석 방식에 최적화된 마크업 체계로의 변환
- 형태 주석 말뭉치: 학습자 오류로 인한 분석 불능(NA) 자료의 처리
- 오류 주석 말뭉치: 최종 수정된 오류 주석 체계로의 변환 및 추가 주석 항목 분석

### ○ 한국어 학습자 말뭉치 구축/가공 인력 실무 교육 및 홍보

교육은 체계적이고 일관성 있는 말뭉치 구축을 위한 것으로 한국어 학습자 말뭉치 구축 단계별 실무 작업자를 대상으로 한 지침 교육과 도구 교육을 중심으로 이루어진다. 정기/비정기 워크숍과 세미나, 전자 메일 등을 활용한 수시 교육의 방법을 활용할 수 있다.

홍보는 말뭉치의 효율적인 보급과 확산을 위한 것으로 한국어 학습자 말뭉치의 구축 사업 홍보와 (예비) 사용자들의 말뭉치 사용 능력을 제고를 주요한 목표로 한다. 홍보 대상은 연구자, 교사, 학습자로 학술 워크숍, 초청 교육 프로그램에서의 워크숍 분과 운영, 학술 발표, 유관기관의 누리집을 통한 홍보, 사용자 커뮤니티 운영 등의 방법을 활용할 수 있다.

### ○ 한국어 학습자 말뭉치 활용 방안 연구

한국어 학습자 말뭉치 활용 방안은 연구자, 교사, 학습자로 나누어 사용자 집단별 목적을 고려하여 제시하였다. 연구자의 경우 한국어 학습자 언어 연구에 관한 선행 연구 분석을 통해 학습자 말뭉치의 활용 분야와 범위를 보였

다. 교사와 학습자는 한국어 교수·학습 환경에서 보다 효율적인 교수·학습을 위한 교수요목 설계, 교재 및 자료 개발, 교수법 개발, 컴퓨터 보조 학습 도구 개발을 위한 기초 자료로의 활용 범위를 제안하였다. 그리고 각 분야에서의 사용을 위해 제공되어야 할 검색 기능을 제시하였다.

향후 이를 바탕으로 검색 서비스 기능을 포함한 한국어 학습자 말뭉치 구축 시스템이 마련되면, 실제 자료를 기반으로 하여 보다 구체화된 모형 제시가 가능해질 것이다.

## 2. 연구의 의의 및 기대 효과

<한국어 학습자 말뭉치>는 <21세기 세종 한국어 균형 말뭉치> 이후로 구축되는 국가 주도의 말뭉치로서의 위상을 가진다. <한국어 학습자 말뭉치>는 한국어 비모어 화자인 외국인 학습자의 자료를 수집하여 구축한 것으로 다음과 같은 의의가 있다.

### ○ 국가 주도의 한국어 학습자 균형 말뭉치 구축

한국어 학습자 언어 연구를 위해서는 학습자가 산출한 자료가 필요한데, 2002년 문화체육관광부의 주도로 구축된 50만 어절의 말뭉치 외에는 한국어 학습자 말뭉치 구축이 시도된 바 없으며, 문화체육관광부(2002)의 경우 자료 사용에 대한 학습자의 동의 절차를 거치지 못해 배포와 활용이 불가능한 상황이다. 이 연구에서는 자료의 배포와 활용을 고려하여 IRB 규정에 따라 자료 제공 및 사용에 관한 학습자의 동의를 얻어 370만 어절 규모의 말뭉치를 구축하게 된다. 한국어 학습자 말뭉치 구축은 한국어 교육 학계의 오랜 숙원 사업으로 2015년 한국어 교육 기관 및 교사의 적극적인 협조와 지지를 얻어 성공적으로 수행되었다.

### ○ [연구] 학습자 말뭉치 활용 연구를 통한 국제 수준의 학술 교류 기반 조성

외국어 또는 제2 언어 교육 분야에서 학습자 말뭉치를 활용한 연구가 활발하게 이루어지고 있다. 영어 교육 분야의 경우 CLC(Cambridge Learner Corpus), ICLE(International Corpus of Learner English)와 같은 대규모 말뭉치를 비롯한 말뭉치가 다양하게 구축되어 있으며, 학습자 말뭉치 연합회(Learner Corpus Association)와 같은 학술 단체를 중심으로 자료를 공유하고

정기적인 학술 모임 등을 개최하는 등 활발한 학술 교류가 이루어지고 있다. 한국어 학습자 말뭉치 구축은 관련 연구의 활성화, 한국어교육 연구자 간의 학술 교류, 세계 규모의 학술대회 대회 참여 등을 통한 국제 수준의 학술 교류를 활성화에 기여할 수 있다.

#### ○ [교육] 한국어교육 이론의 체계화 및 교육 자료 구축의 기반 조성

기존의 언어 교수법은 교수자의 경험과 직관에 전적으로 의존함으로써, 언어 교육의 효율성 및 교수법의 적합성을 평가하기 어려웠다. 반면, 학습자 말뭉치를 활용한 연구는 외국어로서의 한국어 학습 측면에서 한국어 사용에 대한 실질적이며 구체적인 설명과 예시를 제시할 수 있다. 뿐만 아니라 학습자 말뭉치에 근거하여 사전뿐만 아니라 학습교재 편찬, 다양한 교수 이론을 체계화할 수 있으며 그에 필요한 내용 자료를 구축할 수 있다.

#### ○ [학습] 한국어 학습자의 다양화에 따른 교수·학습 환경의 과학화

세계적으로 말뭉치를 이용한 언어 교육이 활발하게 이루어지고 있으며 이미 상당한 성과를 거두고 있다. 학습자 말뭉치는 학습자의 성별이나 나이, 학습 기간, 모국어, 사용 교재 등과 같은 여러 가지 변인이 언어 학습에 미치는 영향을 비교·분석하는 데에 활용할 수 있다. 그리고 그 결과는 학습자의 수준, 제1 언어, 학습 목적 등의 변인이나 상황에 맞는 맞춤형교육을 위한 자료로 활용할 수 있다. 그 외에도 또한 각광받고 있는 학습자 중심의 컴퓨터 보조 언어 학습(CALL) 도구 개발을 위한 기초 자료로 활용할 수 있다. 그럼으로써 보다 체계적이고 과학적인 환경에서 보다 수준 높은 학습 환경을 제공할 수 있다.

#### ○ 한국어의 세계화 및 국제 경쟁력 강화

한국어 학습자 말뭉치는 한국어 교육 연구와 교수·학습을 위한 기초 자료로 연구자, 교사, 학습자에 의해 광범위하게 활용될 수 있다. 한국어 학습자 말뭉치를 활용한 연구는 제2 언어, 외국어교육 연구의 주축으로서 국제 수준의 학술 교류 기반을 조성하는 데에 기여할 것이다. 아울러 한국어교육 이론을 체계화하고 교육 자료를 구축함으로써 다변화되어 가고 있는 한국어 교육 환경에 맞는 양질의 교육 서비스를 제공할 수 있게 된다. 그 결과 한국어의 세계화와 함께 한국의 국제 경쟁력 강화에 이바지할 것이다.

### 3. 보고서 활용 방안

모어 화자의 자료를 중심으로 한 말뭉치 구축 이론을 소개하는 논저는 비교적 많은 반면, 학습자 말뭉치 구축에 관한 실제적인 내용을 다룬 논의는 그리 많지 않다. 본 보고서에는 한국어 학습자 말뭉치 구축 방법과 절차, 각 단계에서의 지침 수립과 관련한 쟁점들이 체계적으로 기술되어 있는데, 이들 자료는 다음과 같이 활용될 수 있다.

#### ○ 한국어 학습자 말뭉치 구축에 관한 이론적 지침

한국어 학습자 말뭉치는 비모어 화자의 자료를 수집하여 구축한 자료로 특수 말뭉치로 분류할 수 있다. 따라서 말뭉치 구축의 일반적인 체계는 참조할 수 있지만 비모어 화자가 산출한 작문이나 말화 자료를 구축 대상으로 하는 만큼 수집, 구축, 가공의 전 단계에서 세부적인 방법과 절차는 달라질 수밖에 없다. 그러나 그간 한국어 학습자 말뭉치 구축이 활발하게 이루어지지 못하였고, 주로 개인 연구자가 접근이 용이한 자료들을 무작위로 수집한 경우가 많았기 때문에 구축 방법론에 관한 이론이 체계화되지 못하였다. 이 보고서에 담긴 말뭉치 구축의 쟁점과 그것을 풀어나가는 과정은 한국어 학습자 말뭉치 구축에 관한 이론적 지침이 될 수 있다.

#### ○ 한국어 학습자 말뭉치 구축의 실제를 위한 실용적 지침

이 보고서에는 한국어 학습자 말뭉치의 수집, 자료 처리, 입력/전사, 형태주석과 오류 주석의 전 과정이 기록되어 있다. 본 연구에서는 각 단계에서 이론적인 말뭉치 구축 절차를 충실히 따르되, 최신의 자료 처리 기술을 최대한 활용하여 말뭉치 구축의 지난한 절차들을 효율적으로 변경하였다. 가령, 과거에는 자료 입력/전사 단계의 전후에 헤더 마크업이나 본문 마크업을 일일이 입력하는 수고를 해야 했는데, 본 연구에서는 사전에 등록된 정보를 활용하여 소프트웨어를 활용하여 자동 부착되도록 하였다. 보고서에 기술된 구축 방법과 절차는 학습자 말뭉치 구축의 실제 사례로 향후 학습자 말뭉치를 구축하고자 하는 기관이나 연구자들에게 실용적인 지침으로 활용될 수 있다.

## 4. 정책 제언

### ○ 한국어 교육 학계와 사용자들의 요구를 반영하기 위한 지속적 말뭉치 구축과 행정적 지원

한국어 학습자 말뭉치는 연구자, 교사, 학습자 등 다양한 목적의 사용자들이 그 활용 가치를 충분히 인정하고 공감함에도 불구하고 구축 절차에 대한 지식과 경험의 부족, 자료 접근의 어려움 등으로 실제 구축으로 이어지지 못한 측면이 있다. 그런 만큼 본 사업에 대한 학계와 사용자들의 기대와 호응이 크기도 하다.

본 연구에서는 사용자들의 요구를 충분히 반영하기 위하여 선행 연구를 면밀히 분석하여 중장기 계획을 수립하였다. 앞으로 남은 과제는 중장기 계획에 따라 충실히 자료를 구축해 나가는 것이다. 여기에는 많은 인력과 시간, 비용이 소요된다. 2015년의 경우 35만 어절 규모의 자료를 시험 구축하는 데에 수집 단계에서 약 1,000여 명, 구축과 가공 단계에서 30여 명의 인력이 투입되었다. 구축 완료 후 신뢰할 수 있는 자료로서 효용성을 가지려면 구축의 각 단계에서 필요한 현실적인 인력, 시간, 비용이 뒷받침되어야 할 것이다.

### ○ 중장기 구축 계획 달성 후 지속적인 업데이트를 위한 상시 수집 체계 마련

한국어 교육 환경은 다양한 국적, 학습 목적 등의 학습자 변인에 따라 시기각각 변한다. 그러한 흐름에 따라 학습자들의 요구가 달라지며, 그러한 요구에 부응하여 교육과정, 교육 자료, 교수 방법 또한 달라진다. 더불어 학습자들이 산출하는 자료의 내용이나 특성도 달라진다. 따라서 학습자 말뭉치는 이러한 변화를 반영할 수 있도록 6개년의 구축 사업이 종료된 후에도 지속적인 자료 수집이 이루어져야 한다.

본 연구에서는 이를 위해 웹으로 접속 가능한 학습자 말뭉치 구축 시스템을 개발하여 학습자가 자율적인 의지에 따라 시스템상에서 자료를 직접 제공하도록 하는 방안을 부가적으로 제안하고자 한다. 시스템의 실제 운영을 위해서는 시스템 개발, 지속적인 자료 제공을 독려하기 위한 보상(예, 학습자 산출 자료에 대한 피드백)이 전제되어야 하므로 시스템 개발과 유지, 운영에 관한 실행 가능성을 탐색이 필요하다.

### ○ 실용적인 활용을 위한 검색기-활용 도구 자료 배포 시스템 구축

본 사업은 한국어 교육 연구와 교수·학습에서 한국어 학습자 말뭉치가 실용적으로 활용되도록 하는 데에 그 취지가 있다. 그에 따라서 구축 사업과 함께 배포를 위한 시스템 개발 사업을 병행하는 것은 합리적인 판단이라고 하겠다. 여기에 덧붙여 활용 시 효용성을 극대화하기 위해서는 사용자 집단의 속성에 따라 기능이 차별화되어야 한다.

- 연구자: 검색+말뭉치의 확장 구축/추가 주석 도구의 통합: 다양한 검색 기능과 통계 정보의 제공 외에 연구자에게는 추가 구축 또는 추가 주석 시 활용 가능한 구축 도구와 사용 지침을 제공함으로써 활용 범위를 넓힐 수 있다.
- 교사, 학습자: 검색+CALL 기반의 교수·학습 활동 기능의 통합: 한편, 교사나 학습자의 경우 말뭉치를 활용하여 교수·학습 활동으로 연계할 수 있는 컴퓨터 보조 언어 학습 프로그램이 함께 제공될 필요가 있다.

## 참고 문헌

- 강현화(2003), 스페인어권 한국어 학습자의 어미,조사 및 시상, 사동 범주의 오류 분석, 국제한국어교육학회
- 강현화(2010) 한국어 학습자 사전 표제어 선정을 위한 자료 구축 및 선정 방법에 관한 연구, 한국사전학 16 한국사전학회
- 강현화(2011) 한국어 학습자 말뭉치의 자료 구축 방안 대한 기초 연구, 한국사전학 17. 한국사전학회
- 고석주(2002), 학습자 말뭉치에서 조사 오류의 특징, 연세대학교 한국어학당
- 고석주(2004), 오류 유형 주석을 위한 기초 연구, 72-77쪽, 한국 문화사
- 고승연(2013), 아랍어권 한국어 학습자의 발음 오류 분석, 한국어문화교육학회
- 곽수진(2010), 한국어 학습자의 문장 성분 호응 관계 오류 연구, 한국어의미학회
- 권기양(2006), KFL 학습자의 오류에 대하여, 한국언어과학회
- 김경화(2013), 고급단계 한국어학습자의 오류연구, 길림성민족사무위원회
- 김미옥(2002), 학습 단계에 따른 한국어 학습자 오류의 통계적 분석, 연세대학교 한국어학당
- 김미옥(2003), 한국어 학습자의 단계별 언어권별 어휘 오류의 통계적 분석, 국제한국어교육학회
- 김미옥·정희정(2003), “한국어 학습자 작문에 나타난 어휘 오류 분석”, 제3회 한국어 교육 국제 워크숍 발표 요지, 연세대 언어정보연구원 외국어로서의 한국어교육 연구센터, 102-135쪽
- 김아름(2014), 한국어 학습자의 문법 및 화용오류에 대한 인식, 한국국어교육학회
- 김유미(2002), 학습자 말뭉치를 이용한 한국어 학습자 오류 분석 연구, 연세대학교 한국어학당
- 김유미(2006), 학문 목적 한국어 학습자를 위한 문어 학술 말뭉치 구축 -구어 말뭉치 자료 수집과 전사에 대하여-, 한국응용언어학회.
- 김유정(2005), 한국어 학습자 말뭉치 오류 분석의 기준, 한국어 교육 16(1)
- 김정숙(2002), 영어권 한국어 학습자의 조사 사용 오류 분석과 교육 방법, 국제한국어교육학회
- 김정숙(2002), 한국어 학습자 말뭉치 구축을 위한 기초 연구 -개인 정보 표지 체계와 오류 정보 표지 체계를 중심으로-, 이중언어학회
- 김정숙, 김유정(2002), 한국어 학습자 말뭉치 구축을 위한 기초 연구 -개인정보 표지 체계와 오류 정보 표지 체계를 중심으로-, 이중언어학회.
- 김정은(2003), 한국어교육에서의 중간언어와 오류 분석, 한국어 교육 14(1), 29-50쪽, 국제한국어교육학회.
- 김지민, 신승용(2010), 어휘오류 분석의 문제점과 어휘오류 처리 방안 연구

- 김지영(2014), 중국인 유학생의 한국어 사용 오류 분석, 시학과 언어학 26, 7-29쪽, 시학과언어학회.
- 남길임(2007), 학습자 오류 말뭉치를 활용한 한국어 용법 사전의 편찬, 한말연구회
- 남윤주 외(2014), L2로서의 한국어 자연말화 코퍼스의 구축과 활용, 통일인문학논총.
- 노미연(2012), 한국어 학습자의 구어 오류와 후속 상호작용 분석 연구, 동국대학교 박사학위논문, 152-158쪽.
- 민영란(2008), 부정적 전이로 인한 중국어권 학습자의 오류 분석, 한국어 교육 19(1)
- 박수연(2007), 한국어 학습자 오류 말뭉치 구축과 그 문제점에 관한 연구, 연세대학교 언어정보개발원
- 서상규, 유현경, 남윤진(2002), 한국어 학습자 말뭉치와 한국어교육, 국제한국어교육학회.
- 신성철(2002), 호주 한국어 학습자의 어휘 오류 분석 연구, 한국어 교육 13(1), 307-338쪽, 국제한국어교육학회.
- 신성철(2007), 영어권 한국어 학습자의 철자 오류 유형과 패턴, 국제한국어교육학회
- 유석훈(2001), 외국어로서의 한국어 학습자 말뭉치 구축의 필요성과 자료 분석, 국제한국어교육학회.
- 이동은(2007), 한국어 학습자의 철자 오류와 개선 방안 -북미지역 청소년 교포 학습자를 대상으로-, 한국어학회
- 이병운(2011), 베트남인 학습자의 작문 오류 경향 분석, 우리말글학회
- 이승연(2006), 한국어 학습자 말뭉치 오류 표지 방안 재고, 이중언어학회.
- 이승연(2007), 한국어 학습자 오류 판정 및 수정 기준 연구-교사, 비교사 집단간 오류 판별 비교 실험을 바탕으로, 이중언어학회
- 이승연(2007), 한국어교육을 위한 한국어 학습자 말뭉치의 구축과 활용 연구, 고려대 박사학위논문.
- 이유림, 김영주(2013), 교사의 피드백 방법이 한국어 학습자의 작문 내 어휘 오류 감소에 미치는 영향, 외국어로서의 한국어교육 39, 165-191쪽, 연세대학교 언어연구교육원 한국어학당.
- 이정희(2002), 한국어 오류 판정과 분류 방법에 관한 연구, 국제한국어교육학회
- 이정희(2003), 초급 단계 한국어 학습자의 어휘 오류, 이중언어학회
- 이정희(2005), 한국어 학습자의 어휘 오류 분류에 관한 연구, 이중언어학회
- 이정희(2008), 중국어권 한국어 학습자의 어휘 오류 연구 -원인 분석을 중심으로-, 국제한국어교육학회
- 이정희(2009), 중국어권 한국어 학습자의 어휘 오류 연구, 한국어 교육 19(3), 1-23쪽, 국제한국어교육학회.
- 이훈호(2015), 한국어 오류 분석 연구의 동향 분석 연구, 외국어교육연구 29(2), 107-135쪽, 한국외국어대학교 외국어교육연구소.
- 전영옥(2010), 여성결혼이민자의 한국어 어휘 오류 분석, 한말 연구 27

- 조철현 외(2002), 한국어 학습자의 오류 유형 조사 연구, 문화관광부.
- 최원평, 유효려(2010), 중국 대학생 글쓰기에 나타난 어휘 오류 연구, 언어와 문화 6(3), 289-305쪽, 한국언어문화교육학회.
- 한상미(2014), 중급 한국어 학습자의 구어 담화에 나타난 조사 오류 연구, 국제한국어 교육학회
- 한송화(2001), 말뭉치와 학습자 오류를 이용한, 외국인 학습자를 위한 한국어 어휘 사전의 의미 기술, 한국어정보학회
- 한송화(2002), 한국어 학습자의 오류 분석, 연세대학교 한국어학당
- Brock, C , Crookes, C , Day, R., and Long, M. (1986). The differential effects of corrective feedback in native speaker-non-native speaker conversation. In R. Day (Ed.), Talking to learn. Rowley, MA: Newbury House. pp. 229-236.
- Brock, C. (1986). The effects of referential questions on ESL classroom discourse. TESOL Quarterly, 20, 47-59.
- Corder. S. P.(1981), Error Analysis and Interlanguage, Oxford University Press.
- Foster, P. and Skehan, P. (1996) The influence of planning on performance in task-based learning. Studies in Second Language Acquisition 18. pp. 299 - 324.
- Foster, P., Tonkyn, A. and Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. Applied Linguistics 21:3. pp. 354-375.
- Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research report No. 3. Champaign, IL, USA: NCTE. pp. 1467-1770.
- James, C.(1998), Errors in Language Learning and Use. New York : Addison Welsey Longman Inc. 144-154쪽.
- Pica, T., Holliday, L., Lewis, L. and Morgenthaler, L. (1989) Comprthensible Output As An Outcome of Linguistic Demandes On the Learner, Studies in Second Language Acquisition 11:1. pp. 63-90.
- Young, R. (1995). Conversational Styles in Language Proficiency Interviews. Language Learning 45:1. pp. 3 - 42.

<Abstract>

## 2015 Project on Basic Research and Construction of the Korean Learner Corpus

This study aims to establish the fundamental plans to construct the Korean Learner Corpus, and to build a pilot corpus of approximately 350,000-words informed by the basic research.

**Establishment of fundamental plans for constructing the Korean learner corpus:** The Korean learner corpus collected total of 3,700,000-words(3,000,000-words of written, 700,000-words of spoken) through out the three-stage period in six years. The corpus was collected from are domestic learners, immigrant learners and overseas learners of Korean as a foreign/second language.

**Establishment of guidelines for constructing the Korean learner corpus:** The guidelines established for construction of Korean learner corpus include guidelines for data collection/process, written language input, spoken language transcription, morph tagging and error annotation. The guidelines of each stage were guided by <The 21st Century Sejong Project> for consistency and systemicity as a corpus project led by the nation's institution, and to compare with the data produced by the non-native Korean speakers. However, the features of non-native speakers' texts/utterances were taken into consideration, and revised accordingly.

**Collecting, constructing and editing of learners' corpora:** The Korean learner corpus of 2015 is composed of 350,000-words(300,000 written, 50,000 spoken) from raw corpus, 220,000-words(200,000 written, 20,000 spoken) form morph-tagged corpus and 50,000 error

annotated-words(40,000 written, 10,000 spoken) on the basis of the guidelines established in the previous stage.

Moreover, on the job training(OJT) plans for the constructing/editing manpower was established for systematic corpus construction, and public relation plans to better promote the Korean learner corpus to larger population. The public relation includes a tutorial program to improve the user's ability to utilize the corpus. Furthermore, corpus application plan was suggested to enhance the efficiency in utilizing the Korean learner corpus.

<Korean Learner Corpus> carries the status as state-led corpus project following <The 21st Century Sejong Project>. <Korean Learner Corpus> is constructed from the data produced by foreign learners who are non-native Korean speakers, thus the corpus expect to contribute to the globalization of Korean language and to strengthen international competitiveness by influencing relevant research, education and teaching-learning environment as follows.

- [Research] Constructing a foundation for academic exchange at an international level conducting research using Korean learner corpus
- [Education] Constructing foundation for systematization of Korean language education theory and for developing instructional materials
- [Learning] Scientification of teaching-learning environment in accordance with diversification of Korean learners

Keywords: Korean learner corpus; written corpus; spoken corpus; morph tagged corpus; error annotated corpus

연구 책임자: 강현화(연세대학교 국어국문학과 교수)  
 공동 연구원: 김선정(계명대학교 한국문화정보학과 교수)  
                   김일환(고려대학교 민족문화연구원 HK연구교수)  
                   김정숙(고려대학교 국어국문학과 교수)  
                   안경화(서울대학교 언어교육원 대우 부교수)  
                   이동은(국민대학교 국어국문학과 부교수)  
                   이정희(경희대학교 국제교육원 부교수)  
                   한송화(연세대학교 언어정보연구원 HK교수)  
                   황용주(국립국어원 한국어진흥과 학예연구관)  
 연구 보조원: 홍혜란(연세대학교 대학원 박사과정 수료)  
                   강민석(고려대학교 대학원 박사과정)  
                   공나형(연세대학교 대학원 박사과정)  
                   김경아(연세대학교 대학원 박사과정)  
                   김형주(경희대학교 대학원 박사과정)  
                   송지혜(연세대학교 대학원 박사과정)  
                   신범숙(서울대학교 대학원 박사과정 수료)  
                   이민아(국민대학교 대학원 박사과정 수료)  
                   홍종호(계명대학교 대학원 박사과정)  
                   유성희(연세대학교 대학원 석사과정)  
 담당 연구원: 황용주(국립국어원 한국어진흥과 학예연구관)

## 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업

<b>발 행 인</b>	송 철 의
<b>발 행 처</b>	국립국어원 서울시 강서구 금낭화길 148(방화 3동 827) 전화: 02-2669-9743~4    전송: 02-2669-9727
<b>인 쇄 일</b>	2015년 12월 18일
<b>발 행 일</b>	2015년 12월 18일
<b>인 쇄</b>	학위사

※ 이 보고서는 국립국어원 누리집(<http://www.korean.go.kr>)에서도 내려받을 수 있습니다.