

국립국어원 2016-01-29

발간등록번호
11-1371028-000635-01

2016년 한국어 학습자 말뭉치 연구 및 구축 사업

연구 책임자
강 현 화



제 출 문

국립국어원장 귀하

“2016년 한국어 학습자 말뭉치 연구 및 구축 사업”에 관하여
귀 원과 체결한 연구용역 계약에 의하여 연구보고서를 작성하여
제출합니다.

2016년 12월 16일

연구 책임자: 강현화(연세대학교)

연구 기관 연세대학교 산학협력단

연구 책임자 강현화(연세대)

공동 연구원 김선정(계명대), 김일환(고려대),
김정숙(고려대), 김한샘(연세대),
안경화(서울대), 이동은(국민대),
이정희(경희대), 한송화(연세대),
황용주(국립국어원), 김수현(국립국어원)

연구 보조원 홍혜란(연세대), 공나형(연세대),
김보영(연세대), 김선영(연세대),
유소영(연세대), 윤종원(연세대),
허희정(연세대), 하법정(연세대)

2016년 한국어 학습자 말뭉치 연구 및 구축 사업

이 연구는 한국어 학습자 말뭉치 구축 중장기 사업 계획에 따른 제2차 연도 연구이다. 제1차 연도의 <2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업>에서 마련된 학습자 말뭉치 구축 지침을 보다 정교하게 보완하고 그 지침에 따라 2015년에 구축된 35만 어절 규모의 말뭉치를 수정·보완함과 동시에 약 80만 어절 규모의 말뭉치를 새롭게 구축하는 데에 주요한 목적이 있다. 이에 따른 주요 과업과 연구의 성과는 다음과 같다.

한국어 학습자 말뭉치 연구: 한국어 학습자 말뭉치 연구는 체계적인 말뭉치 구축을 위한 구축 지침의 정교화와 말뭉치 활용을 위한 모형 제시로 나뉜다. 실제적인 자료 구축에 따른 주요한 쟁점들을 반영하여 제1차 연도에 마련된 수집 지침, 자료 처리 지침, 문어 입력 지침, 구어 전사 지침, 형태 주석 지침, 오류 주석 지침을 정교화하였다. 또한 이러한 지침에 따라 구축된 말뭉치가 효율적으로 활용될 수 있도록 연구자, 교사, 학습자를 위한 활용 모형을 제시하였다. 사용자 집단별 활용 방안은 향후 웹 사이트를 통해 배포될 검색 기능과 결과를 활용한 실제적인 모형으로 구체화되었다.

한국어 학습자 말뭉치 관련 교육 및 홍보: 한국어 학습자 말뭉치 관련 교육 및 홍보에서는 체계적인 말뭉치 구축을 위해 필요한 수집 자료의 처리, 문어 입력, 구어 전사, 형태 주석, 오류 주석의 각 단계별 지침 교육과 도구 사용 교육이 이루어졌다. 아울러 정기/비정기 워크숍을 통해 작업자 간의 일관성과 통일성을 높임으로써 구축 자료의 신뢰성을 높이고자 하였다. 한편 한국어 학습자 말뭉치 사업을 널리 알리고 사용자들의 활용 능력 제고하기 위하여 2회에 걸쳐 한국어 교육 연구자와 교사를 대상으로 한 워크숍을 개최하였다.

한국어 학습자 말뭉치 수집 및 가공: 한국어 학습자 말뭉치 수집은 제1차 연도에 수집을 시작한 국내 교육 기관 학습자 자료의 집중 수집과 함께 이주민 자료의 시험 수집이 이루어졌다. 이주민 자료의 시험 수집은 2017년 제3

차 연도의 수집 계획을 수립하기 위한 것으로 이주민 자료 수집과 관련한 쟁점들을 파악하는 데에 주력하였다. 한국어 학습자 말뭉치 가공은 2015년 제1차 연도에 구축한 원시 말뭉치 약 35만 어절(문어 30만 어절, 구어 5만 어절), 형태 주석 말뭉치 약 22만 어절(문어 20만 어절, 구어 2만 어절), 오류 주석 말뭉치 약 5만 어절(문어 4만 어절, 구어 1만 어절) 규모의 말뭉치를 새로운 지침에 따라 수정·보완하는 작업이 선행되었다. 그리고 2016년에는 원시 말뭉치 약 80만 어절(문어 70만 어절, 구어 10만 어절), 형태 주석 말뭉치 약 55만(문어 50만, 구어 5만 어절), 오류 주석 약 15만 어절(문어 10만, 구어 5만) 규모의 말뭉치가 새롭게 구축되었다. 그 결과 2015년, 2016년에 구축한 전체 말뭉치의 규모는 원시 말뭉치 약 115만 어절(문어 100만, 구어 15만 어절), 형태 주석 말뭉치 약 77만 어절(문어 70만 어절, 구어 7만 어절), 오류 주석 말뭉치 약 20만 어절(문어 14만 어절, 구어 6만 어절) 규모의 말뭉치가 구축되었다.

<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, 학습에 다음과 같이 활용됨으로써 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, 국내 교육기관의 학습자 자료,
이주민 자료 시험 수집, 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. 학습자 말뭉치 연구	5
1. 구축 지침의 정교화 및 작업 체계의 정비	5
1.1. 문어 입력 지침	5
1.2. 구어 전사	7
1.3. 형태 주석	10
1.4. 오류 주석	14
2. 말뭉치 활용 연구 체계 구축	26
2.1. 연구자를 위한 활용 모형	26
2.2. 교사를 위한 활용 모형	33
2.3. 학습자를 위한 활용 모형	36
III. 학습자 말뭉치 구축 관련 교육 및 홍보	43
1. 말뭉치 구축/가공 인력 실무 교육	43
1.1. 문어 입력 인력 실무 교육	44
1.2. 구어 전사 인력 실무 교육	46
1.3. 형태 주석 인력 실무 교육	47
1.4. 오류 주석 인력 실무 교육	48
2. 한국어 학습자 말뭉치 홍보	50
2.1. 한국어 학습자 말뭉치 아카데미 개최	50
2.2. 학술대회 발표	51

IV. 학습자 말뭉치 수집 및 가공 53

1. 2015년 기구측 학습자 말뭉치의 보완 53
2. 2016년 학습자 말뭉치 수집 및 가공 54
 - 2.1. 수집 54
 - 2.2. 문어 입력 72
 - 2.3. 구어 전사 77
 - 2.4. 형태 주석 84
 - 2.5. 오류 주석 97
3. 온라인 구축 시스템 개발을 위한 협의 작업 117
 - 3.1. 주요 진행 사항 117
 - 3.2. 남은 과제 117

V. 결론 118

1. 연구 요약 118
2. 연구의 의의 및 기대 효과 120
3. 보고서 활용 방안 122
4. 정책 제언 123

참고문헌

부록. MLAT-기반 한국어 적성 평가지

표 차례

<표 1> 연구의 범위와 세부 과업	2
<표 2> 연구 수행 방법	2
<표 3> 연구 추진 경과	3
<표 4> 2015년-2016년 구축 학습자 말뭉치의 누적 규모	4
<표 5> 지침 보완 내용: 문어 입력 지침	5
<표 6> 지침 보완 내용: 구어 전사 지침	8
<표 7> 지침 보완 내용: 구어 전사 지침	9
<표 8> 지침 보완 내용: 형태 주석 지침	11
<표 9> 지침 보완 내용: 오류 주석 체계	15
<표 10> 오류 양상 변경 사항	16
<표 11> 학습자 말뭉치 오류 주석 체계 틀: 기본 주석	17
<표 12> 학습자 말뭉치 오류 주석 체계 틀: 오류 양상	18
<표 13> 학습자 말뭉치 오류 주석 체계 틀: 오류 층위	18
<표 14> 말뭉치 구축/가공 인력 교육의 주요 내용	43
<표 15> 제1차 한국어 학습자 말뭉치 아카데미 프로그램	50
<표 16> 제2차 한국어 학습자 말뭉치 아카데미 프로그램	51
<표 17> 학술대회 발표	51
<표 18> 2015년 기구축 학습자 말뭉치의 규모	53
<표 19> 수집 대상별 자료 수집 범위: 국내 교육기관 학습자	54
<표 20> 국내 교육 기관의 실무 책임자 및 실무 교사 명단	55
<표 21> 숙달도별 자료 수집 현황	57
<표 22> 국적별 자료 수집 현황	57
<표 23> 국적별 자료 수집 현황	59
<표 24> 수집 대상별 자료 수집 범위	60
<표 30> 이주민 자료 수집을 위한 네트워크 마련 절차	61
<표 25> 시험 수집을 위한 이주민 자료 수집 네트워크	61
<표 26> 학습자 정보 수집 항목	62
<표 27> MLAT-기반 한국어 적성 평가 문항 구성	63
<표 28> MLAT-기반 한국어 적성 평가 오답자 수	64
<표 29> 이주민 자료 시험 수집 현황: 숙달도별(파일 수, 개)	66
<표 30> 이주민 자료 시험 수집 현황: 국적별(파일 수, 개)	67

<표 31> 한국어 학습자 말뭉치 연차별 구축 계획	68
<표 32> 문어 말뭉치의 숙달도별 자료 분포	74
<표 33> 문어 말뭉치의 국적별 자료 분포	75
<표 34> 구어 말뭉치의 숙달도별 자료 분포	82
<표 35> 문어 말뭉치의 국적별 자료 분포	83
<표 36> 형태 주석 말뭉치의 숙달도별 자료 분포	87
<표 37> 문어 형태 주석 말뭉치의 국적별 자료 규모	88
<표 38> 구어 형태 주석 말뭉치의 국적별 자료 규모	90
<표 39> 문어 형태 주석 결과_숙달도 단계별	91
<표 40> 문어 형태 주석 결과_국적별	93
<표 41> 구어 형태 주석 결과_숙달도 단계별	94
<표 42> 구어 형태 주석 결과_국적별	95
<표 43> 오류 주석 말뭉치의 숙달도별 자료 분포	101
<표 44> 문어 오류 주석 말뭉치의 국적별 자료 분포	102
<표 45> 구어 오류 주석 말뭉치의 국적별 자료 분포	103
<표 46> 문어 오류 주석 결과_분석 불능 여부: 숙달도 단계별	104
<표 47> 문어 오류 주석 결과_분석 불능 여부: 국적별	105
<표 48> 문어 오류 주석 결과_오류 위치: 숙달도 단계별	105
<표 49> 문어 오류 주석 결과_오류 위치: 국적별	106
<표 50> 문어 오류 주석 결과_오류 층위: 숙달도 단계별	108
<표 51> 문어 오류 주석 결과_오류 층위: 국적별	108
<표 52> 문어 오류 주석 결과_오류 양상: 숙달도 단계별	109
<표 53> 문어 오류 주석 결과_오류 양상: 국적별	110
<표 54> 구어 오류 주석 결과_분석 불능 여부: 숙달도 단계별	110
<표 55> 구어 오류 주석 결과_분석 불능 여부: 국적별	110
<표 56> 구어 오류 주석 결과_오류 위치: 숙달도 단계별	111
<표 57> 구어 오류 주석 결과_오류 위치: 국적별	112
<표 58> 구어 오류 주석 결과_오류 층위: 숙달도 단계별	114
<표 59> 구어 오류 주석 결과_오류 층위: 국적별	115
<표 60> 문어 오류 주석 결과_오류 양상: 숙달도 단계별	116
<표 61> 문어 오류 주석 결과_오류 양상: 국적별	116

그림 차례

<그림 1> 원시 말뭉치 검색 예시 1	26
<그림 2> 원시 말뭉치 검색 결과 예시 1	27
<그림 3> 원시 말뭉치 검색 예시 2	27
<그림 4> 원시 말뭉치 검색 결과 예시 2	28
<그림 5> 형태 주석 말뭉치 검색 예시 1	29
<그림 6> 형태 주석 말뭉치 검색 결과 예시 1	29
<그림 7> 형태 주석 말뭉치 검색 결과 예시 2	30
<그림 8> 형태 주석 말뭉치 검색 결과 예시 3	30
<그림 9> 형태 주석 말뭉치 내려받기 기능을 활용한 직접 주석 예시	30
<그림 10> 오류 주석 말뭉치 검색 예시 1	31
<그림 11> 오류 주석 말뭉치 내려받기 기능을 활용한 직접 주석 예시	32
<그림 12> 오류 주석 말뭉치 검색 예시 2	33
<그림 13> 오류 주석 말뭉치 검색 결과 예시 1	34
<그림 14> 오류 주석 말뭉치 검색 예시 2	35
<그림 15> 오류 주석 말뭉치 검색 결과 예시 3	35
<그림 16> <한국어기초사전>과 <표준국어대사전>의 표제어 제시 방식	36
<그림 17> 오류 주석 학습자 말뭉치 용례 검색 결과: 텔레비전	36
<그림 18> <한국어기초사전> 검색 결과	37
<그림 19> 학습자 말뭉치 오류 형태를 반영한 <한국어기초사전> 검색 결과 모형	37
<그림 20> 학습자 말뭉치 오류 형태 목록을 제시한 <한국어기초사전> 미시구조 모형	38
<그림 21> 학습자 말뭉치에 나타난 표현 문형 오류 포함 문장의 예	38
<그림 22> 오류 주석 학습자 말뭉치 용례 검색 결과: 학국	39
<그림 23> 한국어 학습자 말뭉치 활용 시스템의 검색 상세 조건	40
<그림 24> 한국어 학습자 말뭉치 활용 시스템의 오류 주석 기반 검색 조건	41
<그림 25> 학습자 말뭉치에 기반한 자동 첨삭 시스템 프로세스 도식	41
<그림 26> 구어 전사 인력 교육 절차 및 내용	47
<그림 27> 인터넷 카페를 활용한 작업자 질의응답 예시	47
<그림 28> 오류 주석 작업 단계별 교육 절차 및 내용	48
<그림 29> 자료 수집 체계: 국내 교육기관	56
<그림 30> 표본 정보 및 학습자 정보 등록 화면	72

<그림 31> 원본 스캔 파일 및 동의서 화면	73
<그림 32> 본문 입력 및 자동 주식 화면	73
<그림 33> 마크업 화면	74
<그림 34> 엘란을 활용한 전사 화면	77
<그림 35> 발화자 정보 선택 화면 1	78
<그림 36> 발화자 정보 선택 화면 2	78
<그림 37> 자동 주식 생성 화면 1	79
<그림 38> 자동 주식 생성 화면 2	79
<그림 39> 구어 주식 표지	80
<그림 40> 주식 작업 화면 1	80
<그림 41> 구어 주식 작업 화면 2	81
<그림 42> 완료 화면 1	81
<그림 43> 완료 화면 2	82
<그림 44> 형태 분석 작업 할당 화면	85
<그림 45> 형태 주식 작업 화면	86
<그림 46> 형태 주식 작업 완료 및 오류 주식으로의 연계	87
<그림 47> 오류 주식 작업 할당 화면	97
<그림 48> 개인 오류 작업자들에게 할당 완료된 파일 및 화면	98
<그림 49> 오류 주식 작업 절차	99
<그림 50> 오류 주식 검수 작업 1차: [보류/검토] 내용 확인	100
<그림 51> 오류 주식 검수 작업 2차 : 엑셀 파일 확인	100

I. 연구 개요

1. 연구의 목적 및 필요성

1.1. 연구의 목적

- 본 연구의 목적은 한국어 학습자 말뭉치 구축 사업 중장기 계획에 의해 한국어 학습자의 숙달도별·언어권별 문어·구어 말뭉치 자료를 구축하여 과학적인 한국어교육 연구 환경을 조성하는 데에 있다. 본 연구는 2015년도에 착수한 <한국어 학습자 말뭉치 기초 연구 및 구축 사업>에 이은 제2 차 연도의 사업으로 다음을 세부 목표로 하였다.

- 말뭉치 구축 지침의 정교화 및 말뭉치 활용 체계 구축을 위한 연구
- 학습자 말뭉치 구축 관련 교육 및 홍보
- 말뭉치 수집 및 구축 가공

1.2. 연구의 필요성

- 국가 주도의 한국어 학습자 균형 말뭉치 구축
- 학습자 말뭉치를 활용한 연구 기반 조성
- 한국어 학습자의 다양화에 따른 교수·학습 환경의 과학화
- 학습자 말뭉치 기반 연구를 통한 국제 수준의 학술 교류

2. 연구의 범위

○ 본 연구의 범위와 세부 내용은 다음과 같다.

<표 1> 연구의 범위와 세부 과업

연구의 범위	세부 내용
학습자 말뭉치 연구	○ 말뭉치 구축 지침 정교화 ○ 말뭉치 활용 연구 체계 구축
학습자 말뭉치 구축 관련 교육 및 홍보	○ 말뭉치 구축/가공 인력 실무 교육 ○ 한국어 학습자 말뭉치 아카데미 개최(2회) ○ 학술대회 발표
학습자 말뭉치 수집 및 구축 가공	○ 한국어 학습자 실제 말뭉치 자료 수집 및 구축 ○ 1차 연도 말뭉치 정제 및 2016년 구축된 말뭉치 가공 (형태분석, 오류분석)

3. 연구 방법

○ 본 연구는 말뭉치 구축을 위한 지침 정교화, 말뭉치 활용에 관한 기반 연구와 실제 말뭉치 구축 연구가 순환적으로 이루어졌다. 기반 연구는 문헌 연구, 사례 조사, 현황 조사, 전문가 의견 수렴을 통해 이루어졌으며 이를 바탕으로 한 실제 말뭉치 구축 작업이 수행되었다. 각 연구 방법에 따른 연구 내용과 목적은 다음과 같다.

<표 2> 연구 수행 방법

	내용	목적 및 의의
문헌 연구	○ 한국어 학습자 말뭉치의 활용에 관한 선행 연구	○ 말뭉치 활용 연구 체계 구축의 기초 자료
사례 분석	○ 국내외 학습자 말뭉치 활용 사례	○ 말뭉치 활용 연구 체계 구축의 기초 자료
현황 조사	○ 수집 대상 기관의 교육과정	○ 현실적인 자료 수집 가능성

	및 학습자 분포 등 말뭉치 구축 기초 환경 조사	검토
의견 수렴	<ul style="list-style-type: none"> ○ 한국어 학습자 말뭉치 협의체 구성을 통한 자문 및 의견수렴 ○ 사업 착수 보고회의, 중간 보고회의, 최종보고회의를 통한 정기 자문 및 평가 	<ul style="list-style-type: none"> ○ 학계의 참여 유도 ○ 연구 방향 및 방법, 절차의 타당성 검증

4. 연구 추진 일정

○ 연구 추진 경과는 다음과 같다.

<표 3> 연구 추진 경과

단계	작업 내용	1 개월	2 개월	3 개월	4 개월	5 개월	6 개월	7 개월	8 개월
연구	구축 지침의 정교화	○	○	○	○	○	○	○	○
	말뭉치 활용 연구 체계 구축					○	○	○	○
수집	자료 수집 지침 교육	○	○	○	○	○	○	○	○
	자료 수집	○	○	○	○	○	○	○	○
구축	말뭉치 입력/전사 지침 및 도구 교육	○	○	○	○	○			
	스캔 및 파일 분류, 등록	○	○	○	○	○	○	○	○
	입력 및 전사	○	○	○	○	○			
가공	말뭉치 가공 지침 및 도구 교육	○	○	○	○	○			
	형태 분석	○	○	○	○	○	○	○	○
	오류 주석	○	○	○	○	○	○	○	○

홍보	한국어 학습자 말뭉치 아카데미 개최(2회)				○		○		
	학술대회 발표	○					○		
마무리	1차 연도 사업 마무리 및 2차 연도 기초 계획 수립								○

5. 연구 결과

○ 본 사업의 연구 결과는 다음과 같다.

- 구축 단계별 지침 정교화
- 학습자 말뭉치 활용 모형 제시
- 2015년 기구축 말뭉치의 보완
- 국내 교육기관 자료의 신규 수집 및 구축

<표 4> 2015년-2016년 구축 학습자 말뭉치의 누적 규모

구분		1급	2급	3급	4급	5급	6급	계
원시	문어	163,995 (2,248)	175,555 (1,517)	173,133 (1,296)	190,458 (1,441)	160,478 (1,067)	138,491 (826)	1,002,110 (8,395)
	구어	50,525 (143)	19,133 (48)	17,754 (49)	21,019 (57)	20,760 (34)	24,781 (24)	153,972 (355)
형태 주석	문어	108,838 (1,500)	114,650 (987)	136,439 (1,035)	125,314 (890)	112,162 (741)	104,708 (636)	702,111 (5,789)
	구어	10,190 (38)	10,866 (31)	8,852 (30)	12,847 (40)	10,941 (23)	18,579 (18)	72,275 (180)
오류 주석	문어	21,647 (329)	23,790 (216)	24,351 (188)	23,261 (184)	25,990 (184)	23,847 (141)	142,886 (1,242)
	구어	9,534 (34)	9,069 (29)	8,852 (30)	10,769 (37)	10,941 (23)	14,803 (15)	63,968 (168)

단위: 어절 수, () 안은 파일 수

II. 학습자 말뭉치 연구

1. 구축 지침의 정교화 및 작업 체계의 정비

1.1. 문어 입력 지침

1) 지침 수정 및 보완의 쟁점

○ 문어 입력 지침의 정교화는 1차 연도에 텍스트 에디터를 기반으로 한 입력 작업을 시스템 기반으로 전환하면서 발생하는 문제점들을 해결하고 지침에 반영하는 데에 초점이 주어졌다.

- 입력과 식별이 어려운 문자의 표기 방법
- 일반 문자 외의 기호 처리 방식

2) 수정 및 보완 내용

○ 자판에서 입력이 불가능하거나 원문에서 식별이 어려운 문자는 별도의 기호를 정의하여 입력한 후 시스템에서 마크업 처리하였다. 또한 일반 문자 이외에 기호는 자판에서 그대로 입력한 후 시스템에서 별도의 기호로 마크업 처리하여 일반 텍스트와 구분하였다.

<표 5> 지침 보완 내용: 문어 입력 지침

	2015년 지침	2016년 지침	비고
입력과 식별이 어려운 문자의 표기 방법	<note></note>를 사용하여 작업자 메모를 남김	‘거꾸로 된 물음표(¿)’를 사용해 입력	수정
일반 문자 외의 기호 처리 방식		자판에서 입력한 후 시스템 내에서 별도의 기호로 마크업	수정

(1) 입력과 식별이 어려운 문자의 표기 방법

다양한 외국어 표기와 식별하기 어려운 문자 등 기본 자판에서 입력이나 표기가 어렵다고 판단되는 모든 문자와 기호 형태는 ‘거꾸로 된 물음표(?)’를 사용해 입력한다.

- ‘외국문자’는 영어와 한자 이외의 외국어를 입력할 때 ‘?’ 기호 입력 후 마크업한다.

<예> <EX_Alpha>????</EX_Alpha>

- ‘식별불가’는 원본에서 다양한 이유로 확인이 어렵거나 지워진 문자 등을 표기할 때 ‘?’로 입력 후 마크업한다.

<예> <CNI>??</CNI>

(2) 일반 문자 외 기호처럼 인식되는 경우 처리 방식

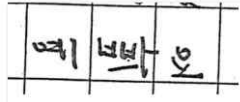
일반적인 한국어 글자 이외에 기호로 인식되는 경우에는 자판에서 직접 입력한 후 마크업으로 일반 문자와 구분하기로 한다.

- ‘기타기호’는 문장 앞에 붙인 블릿 기호나 다른 특수 기호들을 원본 그대로 입력 후 마크업할 때 사용한다.

<예> <EX_Symbol>※★</EX_Symbol>

- ‘한글기호’는 두벌식 한글과 같이 자판에서 하나의 음절로 입력이 불가능한 글자를 입력할 때 사용한다. 초성과 중성 모음, 종성 받침을 한 음절 단위로 입력하고, 입력되지 않은 나머지 중성 모음을 다음 음절 위치에 가로로 적고 하나의 단위로 마크업한다.

<예>



예쁘다 요 : 시스템에 '예쁘다'와 같이 입력 후 '한글기호' 마크업 처리

- 예<NSS>쁘다</NSS>요

3) 문어 입력 지침

☞ 부록 참고

1.2. 구어 전사

1) 지침 수정 및 보완의 쟁점

○ 구어 전사 작업에서 학습자의 발음 오류 및 한국어의 음운 체계와 맞지 않는 발음의 표기 문제 1차 연도 지침 수립 과정에서부터 끊임없는 쟁점이 되어 왔다. 구어 전사 지침의 정교화는 음소, 음절, 음운 규칙에 의한 오류 외에도 한국어의 음운 체계에 없는 중간 발음이나 모국어식의 발음에 의한 음성 차원의 오류 등 다양한 유형의 학습자 오류를 포괄하되, 오류 주석에서 발음 오류 처리의 효율성을 고려하여 전사가 이루어질 수 있도록 하였다.

- 음운 규칙을 적용하지 못하여 발생한 오류

예) 무조건[무조건]

- 불분명한 발음, 중간 소리의 표기

예) 간담(강과 간의 중간 발음으로 날 때)

- 음절 발음에서의 오류

예) 불파음의 실현: 생각[생가ㄱ]

무성음을 유성음으로 실현: 브이아이피(브이의 발음이 'v'에 가까울 때)

- 외래어의 규범 발음과 현실 발음이 다른 경우

예) 인터뷰: 영어식 발음을 사용하여 [인털뷰]에 가까운 소리가 남

센터: 영어식 발음을 사용하여 [세너]에 가까운 소리가 남

파트너: 영어식 발음을 사용하여 [팔너]에 가까운 소리가 남

2) 수정 및 보완 내용

- 2015년 1차 연도의 경우 괄호에 규범 표기를 기입하였다. 하지만 이럴 경우 학습자의 발화가 가시적이지 않다는 단점이 있었다. 따라서 학습자의 발화를 좀 더 가시적으로 드러내고 추후의 오류 주석 과정의 편의성을 위하여 괄호 안에 학습자의 발화를 기입하도록 수정하였다.

<표 6> 지침 보완 내용: 구어 전사 지침

	2015년 지침	2016년 지침	비고
발음 오류 처리	발음 오류는 소리 나는 대로 적되, () 안에 본래의 표기를 넣어 철자법 전사를 보충함	올바른 철자를 먼저 적고 () 안 오류 발음을 적음	수정

- 수정된 지침의 내용은 다음과 같다.

1) 대원칙

정확한 음소, 음절의 발음이나 음운 규칙을 몰라서 일으킨 발음 오류는 올바른 철자 옆에 예와 같이 () 안에 기입한다.

1: 친구(칭구)와 강남(간남)에 갔습니다(갸슨니다).

※전사자 간 괄호의 내용이 상이할 수 있는 오류는 그 내용을 통일하기 위하여 표기 형태를 지정하였음.

㉠ 중간 발음: AB(‘B’의 ‘O’에서 ‘X’와 ‘Y’와 중간 발음)

㉡ 외국어 발음: AB(‘B’의 OO식 발음)

㉔ 음운상의 오류: AB('B'의 'O'의 OOO화)
예) 필요('필'의 'ㄹ'의 권설음화)

- 2015년 전사 작업의 경우 시스템을 전제한 전사라기보다는 학습자 발화의 흐름에 따른 기계적인 전사를 수행한 바 있다. 따라서 세그멘테이션의 경우 구체적인 지침은 마련되어 있지 않았다. 이는 학습자의 발화에 있어 세그멘테이션의 분할은 모어 화자의 직관에 의하여 이루어지기 때문에 구체적으로 규정을 하기가 매우 모호하기 때문이다. 하지만 시스템을 통한 작업이 이루어지다 보니, 세그멘테이션의 분할 단위를 구체화할 필요성이 제기되었다. 따라서 전사 지침에 독립된 세그멘테이션을 보았을 때 맥락을 보지 않고서도 의미가 판단될 수 있는 단위로 분할한다는 내용을 기재하였다. 즉 세그멘테이션은 어절 단위로 분할한다. 1초 이상의 휴지는 {}라는 기호를 사용하여 그 길이를 기재하기 때문에, 이는 어절 단위에 따라 세그멘테이션을 나누어도 휴지를 알아보는 데에는 문제가 없다. 이는 아래와 같이 정리할 수 있다.

<표 7> 지침 보완 내용: 구어 전사 지침

	2015년 지침	2016년 지침	비고
세그멘테이션	-	세그멘테이션의 분할 단위는 어절 단위로 함.	보완

- 보완된 지침의 내용은 다음과 같다.

○ 세그멘테이션은 학습자의 발화의 흐름을 반영하여 전사하되 분할의 단위는 어절 단위로 한다.

3) 구어 전사 지침

☞ 부록 참고

1.3. 형태 주석

1) 지침 수정 및 보완의 쟁점

- 학습자 말뭉치에서의 형태 주석 작업은 학습자 언어의 비정규성을 고려해야 할 뿐만 아니라, 추후에 이루어질 오류 주석 단계 역시 고려해야 하는 작업이다. 이 과정에서 숙련된 분석 전문가에 의한 일관성 있는 작업을 통해 주석의 질을 보장하는 것이 중요하다고 할 수 있겠다. 이번 사업에서는 작년에 구축한 형태 주석 말뭉치 22만 어절에 대한 보완 작업과 더불어 올해 구축분 말뭉치에 대한 형태 주석 작업을 진행하였다. 이 과정에서 형태 주석 작업 지침을 정교화하는 한편, 지침을 이용한 실제 작업 과정의 대원칙을 설정하였다. 또한 지속적인 작업자 교육 및 관리를 통해 작업 일관성을 확보하고자 하였다.
- 형태 주석 지침의 정교화는 두 가지 차원에서 이루어졌다. 먼저, 21세기 세종 계획 현대 문어 형태 주석 말뭉치 분석 지침과의 호환성을 도모하는 한편, 분석의 일관성을 도모하기 위하여 일부 항목을 보완 및 추가하였다. 그 내용은 다음과 같다.
 - 21세기 세종계획과의 호환성
 - ▶ 명사 + ‘하다’ 류의 분리 문제
 - ▶ ‘척하다’ 류의 분석
 - ▶ 지정사의 처리
 - ▶ 부사격조사와 접속조사와 구분
 - ▶ 접두사 목록의 확장
 - ▶ 접미사 목록의 확장
 - ▶ XR(어근) 표지 복원

- 분석의 일관성

- ▶ 유사 접사류 목록의 제시
- ▶ 학습자 오류 어절의 처리

○ 구어 자료에 대한 형태 분석은 1차적으로 문어 분석 지침을 따르되, 끊어진 어절, 억양 단위가 바뀐 발화, 불분명한 어절 등 구어의 특성에 따른 별도의 추가 지침을 마련하여 따르도록 하였다.

2) 수정 및 보완 내용

<표 8> 지침 보완 내용: 형태 분석 지침

	2015년 지침	2016년 지침	비고
명사 + ‘하다’ 류의 분리 문제	분리	분리	유지
‘척하다’ 류의 분석	분리	분리	유지
지정사 처리	‘이다’는 지정사(VCP), ‘아니다’는 형용사(VA) 처리	‘이다’는 VCP, ‘아니다’는 VCN으로 모두 지정사 처리(VCP)	수정
부사격조사와 접속조사의 구분	부사격조사로 통합	부사격조사와 접속조사의 구분	수정
접두사 목록의 확장	1개	33개로 확장	수정
접미사 목록의 확장	15개	53개로 확장	수정
XR(어근) 표지 복원	없음	XR(어근) 표지 복원	추가
유사 접사류 목록의 제시	없음	80개 항목 제시	추가
학습자 오류 어절의 처리	없음	체인 추정 범주의 경우 조사부터 분리, 용언 추정 범주의 경우 어미부터 분리하여 분석함	추가

(1) 명사 + ‘하다’ 류의 분리 문제

- ‘공부하다’, ‘출석하다’, ‘졸업하다’와 같은 ‘명사+하다’ 류에 대해서 오류 주석 작업자들로부터 분리하지 않고 하나의 용언으로 처리하는 것이 오류 주석이 용이하다는 의견이 제시되었으나, 논의 결과 세종 계획 지침을 따라 분리하여 처리하는 것으로 결정하였다. 이로 인해 발생할 수 있는 검색상의 문제의 경우, 오류 주석 차원에서는 접사 관련 표지를 추가하고 주석대상을 통합하여 처리할 수 있게 하고 검색 차원에서는 파생어를 통합하여 검색할 수 있게 하여 해결하도록 하였다.

(2) ‘척하다’ 류의 분석

- ‘척하다’, ‘양하다’, ‘체하다’, ‘뚫하다’와 같은 형태는 보조용언으로 취급하는 경우가 있으나, 본 분석에서는 이들 앞에 항상 관형어가 온다는 분포적 특성을 반영하여 ‘의존명사(NNB)+접사’로 나누어 처리하기로 하였다.

(3) 긍정 지정사와 부정 지정사의 분리

- 2015년 지침에서는 긍정 지정사 ‘이다’만을 지정사로 인정하고 ‘아니다’를 형용사로 처리하였으나, 본 연구에서는 21세기 세종계획 지침에 따라 VCP(긍정 지정사), VCN(부정 지정사)으로 나누어서 처리하는 것으로 변경하였다.

(4) 부사격조사와 접속조사의 구분

- 2015년 지침에서는 부사격조사와 접속조사를 부사격조사(JKB)로 통합하여 처리하였으나, 본 연구의 지침에서는 21세기 세종계획 지침에 따라 분리하여 처리하였다.

(5) 접두사 목록의 확장

- 2015년 지침에서는 접두사로 인정하는 항목이 1개뿐이었으나, 본 연구에

서는 21세기 세종계획 지침에 따라 접두사 목록을 33개로 확장하였다. (자세한 접두사 목록은 지침에 제시함)

(6) 접미사 목록의 확장

- 2015년 지침에서는 접미사로 인정하는 항목이 15개뿐이었으나, 본 연구에서는 21세기 세종계획 지침에 따라 접두사 목록을 53개로 확장하였다. (자세한 접두사 목록은 지침에 제시함)

(7) XR(어근) 표지 복원

- 21세기 세종계획 지침에 따라 XR(어근) 표지를 복원하였다.

(8) 유사 접사류 목록의 제시 및 확정

- 유사 접사류란 접사처럼 사용되는 명사를 총칭하는 것으로, 이들은 사전에 ‘(일부 명사 뒤/앞에 붙어)~의 뜻을 나타내는 말’로 등재되며, 앞뒤에 붙어 쓰이는 명사와 띄어쓰기 없이 함께 쓰인다. 본 연구에서는 이들 유사 접사류 80개에 대한 목록을 제시하고, 이러한 형태가 사용된 경우에는 분리하지 않는 것으로 결정하였다.

(9) 학습자 오류 어절의 처리

- 오류 어절 처리에 있어 이전에는 체언 추정 범주(NF), 용언 추정 범주(NV)로 처리하던 것을 이번 사업에서는 가급적이면 분석하도록 하였다. 체언 추정 범주에 대해서는 형태가 온전한 조사부터 분석하는 것으로, 용언 추정 범주에 대해서는 형태가 온전한 어간부터 분석하는 것을 원칙으로 하였다. 두 가지 접근이 모두 불가능한 경우에 한해서만 분석 불능(NA)으로 처리하였다.

3) 형태 주석 지침

☞ 부록 참고

1.4. 오류 주석

1) 지침 수정 및 보완의 쟁점

○ 2016년 연구에서는 2015년 연구에서 마련한 한국어 학습자 말뭉치 오류 주석 틀과 주석 표지를 형태 주석 체계와 맞물려 처리하고, 오류 주석 작업의 효율성을 위하여 보다 체계적으로 주석 체계 틀을 수정하였다.

- 오류 주석의 세부적인 체계 수정
- 오류 주석 범주인 [오류 영역]을 [오류 위치]로 용어 변경
- 오류 주석 범주인 [오류 양상]의 세부 오류 유형 변경
- 오류 주석 표지 추가

○ 또한 오류 판정과 관련된 오류의 식별, 오류의 교정의 기본 원칙 및 세부 범주별 오류 주석 지침을 구체적으로 추가 보완하였다.

- 오류의 식별 기준 보완
- 오류의 교정 기준 보완
- 시제 오류 처리 기준 보완
- 사동과 피동 오류 처리 기준 보완
- 높임 오류 처리 기준 보완
- 구어 오류 처리 기준 보완

2) 수정 및 보완 내용

(1) 오류 주석의 체계 부분 수정

- 1차 연도 지침에서는 오류 범주 간의 중복과 충돌을 피하고 오류 주석 작업의 효율성을 제고하기 위하여 기존의 오류 범주를 [기본 주석]과 [확장 주석]으로 나누어 이원화 시킨 바 있다. 기본 주석은 필수 주석 범주로 오류가 발생하는 모든 항목에 주석하여 [분석 여부], [오류 양상], [오류 영역]이 이에 해당한다고 하였으나 본 연구에서는 [분석 여부], [오류 위치]가 기본 주석에 해당하는 것으로 세부적인 체계를 변경하였다. [오류 양상]의 경우 모든 오류 형태에 주석되지 않으며, 오형태와 대치와 같이 중복 주석이 가능한 영역이기 때문에 확장 주석 체계로 이동하였다. 따라서 기본 주석에 해당하는 [오류 위치]는 모든 오류에 대해 1:1로 주석하고(분석이 불가능한 오류는 분석 불가능 표시), 해당 항목이 있을 경우에만 주석하는 수의적 주석인 확장 주석에는 [오류 양상]과 [오류 층위]가 두었으며, 한 형태에 대해 2개 이상의 오류가 나타나면 중복 주석을 가능하게 했다.

<표 9> 지침 보완 내용: 오류 주석 체계

	2015년 지침	2016년 지침
기본 주석	[분석 여부], [오류 양상], [오류 영역]	[분석 여부], [오류 위치]
확장 주석	[오류 층위]	[오류 양상], [오류 층위]

- 기존의 오류 주석 범주에서 [오류 영역]으로 지칭하던 것을 [오류 위치]로 변경하였다. [오류 위치]는 오류가 나타난 부분의 품사를 주석하는 기본 주석으로서 오류가 발생한 품사 위치를 가리킨다. 기존의 [오류 영역]은 오류가 나타난 품사 외에 품사가 포함된 전체 영역을 아우르는 것으로 보고, 보다 특정적으로 오류가 나타난 위치를 포착하기 위하여 [오류 위치]로 용어를 변경하였다.
- 2015년도 연구에서 수정안으로 제시한 [오류 양상]에는 누락(OM), 첨가(ADD), 대치(REP), 오어순(MISO)이 세부 오류 유형으로 제시되었다.

2016년도 연구에서는 오어순(MISO)의 경우, [오류 층위]의 통사 영역에서 다루어져야 할 것으로 판단하여 [오류 층위]-통사(어순, WO)로 이동하였다.

- 아울러 주로 규범 표기에 대한 지식 부족으로 인하여 발생하는 오철자 오류와 활용을 잘못하여 발생하는 오류를 누락, 첨가, 대치로 처리하기 어렵다고 판단하여 확장 주석 범주의 음운 영역에서 주석하였던 오형태(MIF)를 [오류 양상]으로 이동하여 추가하였다.¹⁾

<표 10> 오류 양상 변경 사항

2015년 지침		2016년 지침	
오류 양상	주석 표지	오류 양상	주석 표지
누락	OM	누락	OM
첨가	ADD	첨가	ADD
대치	REP	대치	REP
오어순	MISO	오형태	MIF

(2) 오류 주석 표지 보완

- 형태 주석 지침의 변경으로 인하여 오류 주석 표지에서 다음과 같이 표지가 추가되었다. 먼저, [오류 위치]에서는 접두사(CXPN), 명사파생접미사(CXSN), 동사파생접미사(CXSV), 형용사파생접미사(CXSA), 어근(CXR), 접속조사(FJC), 보격조사(FCP)가 새로 추가되었다. 또한 구 단위 표현(PHE) 외에 표현 문형(PE)도 추가되었으며, 표현 문형은 <한국어 문법 사전>(국립국어원, 2005) 목록을 기준으로 한다. 이는 교재(기관)마다 표현 문형으로 제시하고 있는 목록과 수가 다르기 때문에 어느 정도 정제된 목록으로 <한국어 문법 사전>이 적합하다고 판단하였기 때문이다.
- 다음으로 [오류 층위]의 형태 영역에서 품사(POS)를 추가하였다. 이는 오류 주석을 하는 과정에서 동사형이나 형용사형을 사용해야 하는데 명사형을 사용한 경우와 같이 품사를 제대로 인식하지 못하여 발생한 오류들이 나타났기 때문이다.

1) 오류의 양상은 이론적으로는 누락, 첨가, 대치 중 하나이나, 단순 철자 오류나 활용 오류 같은 것들은 이 기준으로 분류하는 것이 무의미하므로, 오형태로 별도 처리하였다.

<예> 일부 사람은 가족이 가장 중용(√중요하다고) 생각했는데 그 이유 점은 돈으로 바뀔 수 없다고 지적을 했다.

☞ ‘중요하다’라는 형용사를 사용해야 하는데 명사형인 ‘중용’만 사용했으므로 명사의 품사 대치 오류로 처리한다. 아울러 ‘중요’도 ‘중용’이라고 사용했기 때문에 오형태 오류도 동시에 주석한다. 이처럼 품사 오류는 동일 의미 단어의 품사를 잘못 사용한 경우에 오류 층위에서 품사 대치 오류로 처리하며, 이때 뒤에 따라오는 활용 형태의 누락 요소들은 품사를 제대로 사용하지 못한 것으로 인한 것이기 때문에 별도의 누락 처리는 하지 않는다.

<표 11> 학습자 말뭉치 오류 주석 체계 틀: 기본 주석

	오류 유형		주석 표지
분석 불가능	전체적 오류 포함		IMP
오류 위치	실질어휘	고유명사	CNNP
		일반명사	CNNG
		의존명사	CNNB
		대명사	CNP
		수사	CNR
		동사	CVV
		형용사	CVA
		보조용언	CVX
		지정사	CVC
		관형사	CMM
		일반부사	CMAG
		접속부사	CMAJ
		감탄사	CIC
		접두사	CXPN
		명사파생접미사	CXSN

	오류 유형		주석 표지
		동사파생접미사	CXSV
		형용사파생접미사	CXSA
		어근	CXR
	기능어휘	주격조사	FNP
		관형격조사	FGP
		목적격조사	FOP
		부사격조사	FAP
		접속조사	FJC
		보격조사	FCP
		호격조사	FVP
		인용격조사	FQP
		보조사	FXP
		연결어미	FED
		종결어미	FFE
		선어말어미	FPE
		명사형 전성어미	FNE
		관형사형 전성어미	FAE
	구 단위 표현		PHE
	표현 문형		PE

<표 12> 학습자 말뭉치 오류 주석 체계 틀: 오류 양상

	오류 유형		주석 표지
오류 양상	누락		OM
	첨가		ADD
	대치		REP
	오형태		MIF

<표 13> 학습자 말뭉치 오류 주석 체계 틀: 오류 층위

오류 층위	오류 유형		주석 표지
	발음	음소	PP
	발음	음절	PS
		음운규칙	PC
		원어식 발음	PN(임시 기호)
		중간 발음(변이음포함)	PA(임시 기호)
	형태	단어 형성[합성법]	MCP
		단어 형성[파생법]	MDV
		굴절[곡용]	MDC
		굴절[활용]	MCJ
		품사	POS
	통사	높임	SH
		시제	ST
		사동	SC
		피동	SP
		부정	SN
		어순	WO
	담화	지시	DR
		접속	DC
		담화표지	DM
		구어/문어 오류	DS

(3) 오류의 식별 기준 보완

- 어휘와 문법 오류가 있을 경우, 기능어 중심으로 문법 오류를 우선시하여 처리하기로 하였다.

<예> 저는 가을이(√/가을을) 좋아해요.

☞ 위의 문장은 ‘가을이 좋아요’ 또는 ‘가을을 좋아해요’라는 두 개의 교정형이 가능하다. 어떠한 교정형을 제시하느냐에 따라서 주격조사의 오류가 될 수도 있고, 동사의 품사 대치 오류가 될 수도 있으나 본 연구에서는 기능어 중심으로 조사를 오류로 처리하는 것을 원칙으로 한다.

(4) 오류의 교정 기준 보완

- 오류의 교정은 정보가 소실되지 않는 차원에서 최소한의 교정을 원칙으로 하였다. 따라서 앞부분의 오류를 수정하는 것으로 인해 뒷부분에까지 영향을 주어 수정하게 되면 앞부분도 수정하지 않도록 처리하였다.

<예> 제가(✓저는) 가족에서 3남매 중 막내로 태어났다(✓태어나다) 보니 성격은 좀 활발하고 어릴 때부터 오빠랑 언니 보고 좀 눈치가 있는 편입니다.

☞ ‘제가 가족에서 3남매 중 막내로 태어났다 보니’를 보다 자연스러운 문장으로 만들기 위해서 ‘가족에서’를 삭제하거나 ‘에서’를 ‘중’으로 교정할 수도 있으나 이를 수정하게 되면 뒤에 오는 문장에까지 영향을 줄 수도 있다. 즉, ‘저는 가족 중 3남매이 막내로 태어났다 보니’로 교정하게 되면 앞부분으로 인하여 뒷부분까지 많은 형태를 교정하게 된다. 따라서 ‘3남매 중 막내로 태어났다’는 문장에 오류가 없는 문장으로 ‘제가’를 ‘저는’으로만 대치하고, 뒷부분에서는 ‘태어나다 보니’로만 교정한다. 교정 어절은 정보가 소실되지 않는 차원에서 최소로 한다.

- 누락 오류에서는 필수적인 성분이 생략되었을 경우에만 교정 어절을 주는 것을 원칙으로 하였다. 수의적인 성분을 교정 어절로 주는 것은 예측 교정 어절로 학습자의 의도와 상관없을 수 있으며, 주석자마다 다르게 처리할 수 있다. 따라서 조사나 어미 누락을 중심으로 필수적인 성분만 누락 오류로 교정 어절을 제시하여 처리하였다.

<예> 저는 여러 까지(√/가지) 능력서가(√/자격증을) 취득하지만(√/했지만) 그 중에서 영어를(√/영어) 능수하는(√/능숙한) 정도입니다.

☞ ‘그 중에서 영어를 가장 능숙한 편입니다’라고 교정하여, ‘가장’이라는 부사를 첨가하고, ‘능숙한 편입니다’라고 교정할 수 있으나, ‘가장’은 필수적인 성분이 아니다. 이처럼 필수적인 성분이 아닌 것을 추가하여 [누락] 오류로 처리해서는 않는다. 또한 최소한의 교정 원칙에 따라 ‘능숙한 정도입니다’는 용인 가능하기 때문에 ‘정도’를 ‘편’으로 교정하지 않는다.

- 오류 영역에서 교정 어절로 인해 조사나 어미가 바뀌는 경우는 오류로 처리하지 않는다. 기본적으로 조사와 용언 오류가 동시에 있을 경우, 조사 오류로 처리하는 것이 원칙이지만 용언을 반드시 교정해야 할 때, 용언 때문에 조사가 바뀌는 경우에는 용언 대치 오류로만 처리하고 조사 오류는 주석하지 않았다.

<예> 나라가(√/나라를) 발전하다(√/발전시키다)

☞ 사동표현 ‘시키다’로 용언을 교정해야 할 경우, 조사도 같이 바뀌게 되어 조사까지 교정해야 한다. 이처럼 용언을 교정하여 격이 바뀔 때에는 조사 ‘를’은 대치로 처리하지 않고 ‘가’ 아래에 교정어절만 써주는 것으로 처리한다. 따라서 [오류 위치-접사], [오류 양상-대치], [오류 층위-사동]오류로 주석한다.

- 시제 오류는 시제 또는 시상을 나타내기 위한 선어말어미 오류 등을 말한다. 따라서 과거시제나 미래시제를 써야하는데 현재시제를 사용하거나 시제를 나타내는 요소를 사용하지 않는 경우 시제 오류로 처리한다.
- 기본적으로 시제 오류의 양상은 대치 오류로 주석한다. 이것은 학습자가 시제를 인식하지 못한 것으로 판단하여 현재와 과거, 현재와 미래 등으로 대치한 것으로 보았다. 그러나 동사의 경우 기본형을 사용한 경우에만 누락으로 처리하고, 그 외에 과거 시제나 미래 시제 자리에 현재 시제를 사

용한 경우는 대치 오류로 주석하였다. 형용사의 경우 기본형과 현재형이 같기 때문에 기본형을 현재형으로 간주하고 대치 오류로 처리하였다.

<예> 앞에 집이 새로 강아지를 산다(√샀다).

☞ 과거 시제 ‘샀다’를 써야할 자리에 ‘산다’라는 현재형을 사용할 경우에 시제 대치 오류로 주석한다.

- 사동과 피동은 사동사/피동사, 사동/피동 표현에서 발생하는 오류를 말한다. 사동사/피동사/, 사동/피동 표현을 사용해야 하는데 사용하지 않은 경우는 기본적으로 대치 오류로 처리함을 원칙으로 하였다.

<예> 왜냐하면 중국 밥물 중에서 노동밥에 따라서 소득 격차 등 불공평 제도를 감소할 수 있다(√감소시킬 수 있다).

☞ 사동으로 표현해야 하는데 ‘감소하다’를 사용하였다. 이를 ‘감소시키다’의 사동 형태로 교정하고, 오류 양상은 대치 오류로 처리한다.

우리 집에 물을 열리면(√열면) 계단을 있다.

☞ 피동표현을 사용하지 말아야 하는데, ‘열리면’으로 피동형을 사용했기 때문에 피동 대치 오류로 처리한다.

기술이 발달해서 멋진 영화나 공연이 많아질수록 전통문화를 점점 잊어버리게 했다.(√잊어버리게 된다).

☞ 사동 표현 ‘-게 하다’와 피동 표현 ‘-게 되다’의 대치 오류로 처리한다.

- 높임 표현 오류는 용인 가능성으로 인하여 그 판정 기준이 명확하지 않을 수 있다. 한국어 모어 화자들도 일상적인 언어생활에서 높임 표현을 잘 지키지 않는 경우가 많으며, 화청자 관계에 따라 적절성과 용인 가능성이 달라지기 때문이다. 따라서 필수적인 높임 표현 오류 판정 기준을 설정하기 어렵고, 주석자에 따라 오류 판정 기준이 달라질 수 있다. 기존의 높임 오류 처리 시 한국어 모어 화자의 일상적인 언어생활에서 용인 가능한 경우에는 오류로 처리하지 않았으나, 주석자 간의 일관성을 고려하여 동일 문

장 내에서 한 부분이라도 높임 요소를 실현시킨 경우에는 호응이 되도록 실현시키지 않은 부분을 높임 오류로 처리한다. 즉, 주격조사 ‘-께서’를 사용한 반면에 서술어에서는 높임 선어말어미 ‘-시-’를 사용하지 않은 경우는 오류로 판정하도록 한다.

<예> 근데 부모님께서 나에게 유학하라고 해서(√하셔서) 부모님의 말(√말씀)대로 한국에 와서 유학했다.
 ☞ ‘부모님께서’라고 앞에서 높임 표현을 실현시켰기 때문에 뒤에 오는 서술어에서도 마찬가지로 높임의 선어말어미 ‘-시-’를 실현시켜야 하나 그렇지 않았으므로 높임 누락 오류로 주석하고, ‘말’의 경우 ‘말씀’의 어휘 높임 대치 오류로 주석한다.

(5) 구어 오류 처리 기준 보완

- 2016년 연구에서는 구어 오류 지침을 보다 정교화 하였다. 기본적으로 구어 오류 주석과 문어 오류 주석의 기본 원칙 및 처리 방법은 동일하다. 그러나 발화 상에서 나타나는 발음 오류의 경우, 기본 주석인 오류 위치와 확장 주석에 해당하는 오류 층위[발음]을 주석하고, 오류 양상(대치, 누락, 첨가, 오형태)은 주석하지 않는다. 발음상의 오류를 오류 양상으로 나타내는 것이 무의미하고, 중요한 것은 음소(PP), 음절(PS), 음운규칙(PC), 원어식 발음(PN), 중간 발음(변이음 포함)(PA) 가운데 어떠한 발음 오류 인지를 살펴보는 것이 중요하다고 판단하였기 때문이다.
- 2015년도 연구에서는 주석자의 자의적인 해석을 막기 위해 ‘오류와 실수를 구분하지 않는다’는 기본 원칙에 따라 구어 자료에서 발화 중에 학습자의 자기 수정이 일어난 경우, 수정하기 이전의 일탈도 오류로 간주하여 주석 대상에 포함시킨다고 한 바 있다. 그러나 2016년 연구에서는 수정 후 발화에 초점을 두고 오류 여부를 판정한다. 즉, 말더듬거림이나 수정 전 앞부분의 발화는 오류로 주석하지 않는다. 다만, 전사 단계에서 특정 표시를 하므로 이를 통해 향후 검색이 가능하도록 하였다.

<예> 친= 친구가 한국 음식이:: = 음식을 좋아해서

☞ ‘친= 친구’와 같은 말더듬거림은 오류로 처리하지 않는다.
 ‘음식이:: = 음식을’과 같이 수정 전 발화에서 조사를 잘못
 사용하였지만, 다시 수정하여 조사를 고쳐 제대로 사용한
 경우에 앞부분 ‘음식이’는 오류로 처리하지 않는다.

- 구어에서 조사의 생략이나 축약된 형태의 사용은 한국어에서 일반적으로 나타나는 현상이다. 또한 구어 오류 주석을 억양 단위로 했을 경우, 조사의 생략은 자연스럽고, 용인 가능성이 문어에 비해 높아지기 때문에 구어에서는 이러한 형태를 용인 가능한 것으로 보고, 용인 가능하지 않은 생략일 때에만 누락으로 처리하였다.

<예> 제 한국 생활(√ 생활은)
 아주 재미있고
 한국도 좋아요
 ☞ 구어에서는 조사 생략이 자연스러운 경우가 있기 때문에
 엄격하게 잡지 않도록 한다. 조사 생략을 용인 가능한 것
 으로 보고 오류로 처리하지 않는다.

- 구어에서는 한국어 모어 화자들의 현실 발음을 고려하여 일반적으로 많이 사용되며, 구어에서 허용되는 형태는 오류로 처리하지 않는다.

<예> [그리구](√그리고)::, 음::
 ☞ ‘그리고’를 [그리구]라고 발음하는 것은 한국어 모어 화자들에게도 많이 나타난다. 이처럼 한국어 모어 화자들의 현실 발음을 고려하여, 쭈, [바래요](바라요) 등 구어에서 허용되는 발음은 오류로 처리하지 않는다.
 [그쵸](√그렇죠)::, 음::
 ☞ ‘그쵸’와 같은 축약된 형태는 구어에서 많이 나타나는 현상으로 오류로 처리하지 않는다.

- 오류 층위[발음]에는 총 5가지가 있는데, 그중 ‘음소, 음절, 음운규칙’ 3가지를 우선적으로 주석하였다. 원어식 발음과 중간 발음은 오류의 원인에

해당하는 문제이기 때문에 외래어(및 모국어 화자의 원어식 발음)의 경우에만 ‘원어식 발음’ 오류로 주석하고, 변이음(음성대치)이 분명하게 식별될 경우에만 ‘중간 발음’ 오류로 주석하였다.

- <예> 예:: [코피](√커피)= 카페에서::, 예:: 공부를, 합니다::.
- ☞ 음소 오류는 음소 단위에서 발화가 잘못 사용된 경우로, [코피]는 모음 ‘아’와 ‘오’를 교체하여 발음하였으므로 음소 오류로 주석한다.
 - 음 [바르표](√발표)::, 하고::,
 - ☞ 음절 오류는 음절 단위에서 발화가 잘못 사용된 경우로 원래 음절보다 더 많게 혹은 더 적게 발화한 경우이다. [바르표]는 2음절인 [발표]를 3음절로 발음하고 있으므로 음절 오류로 주석한다.
 - [학교](√학교)
 - ☞ 음운규칙 오류는 유음화, 연음화하여 발음해야 하는데 글자 그대로 절음화하여 발화한 경우이다. ‘학교’의 경우 [학교]라고 발음해야 하나, 단어 하나씩 발음하여 [학교]라고 부자연스럽게 발음하므로 음운규칙 오류로 주석한다.

3) 오류 주석 지침

☞ 부록 참고

2. 말뭉치 활용 연구 체계 구축

2.1. 연구자를 위한 활용 모형

1) 원시 말뭉치 어휘 검색 활용

- 원시 말뭉치를 활용하여 어휘나 어휘를 포함한 파생어, 합성어 등의 학습자 용례를 검색할 수 있다. 검색 시, 아래에 보는 바와 같이 국적이나 모국어, 학습기관, 등급, 나이, 학습 목적, 학습 환경, 말뭉치 유형 등을 선택하여 원하는 자료만을 검색하는 것도 가능하다. 이를 검색한 결과는 아래에 보는 바와 같이 각 용례에 대한 상세한 정보가 부착된 채로 엑셀로 다운 받을 수 있어 엑셀에서 연구자가 원하는 주석이나 소트가 가능하여, 연구자가 원하는 어휘 오류 연구나 학습자의 어휘의 중간언어 발달 연구에 활용할 수 있다.

원시 말뭉치 검색	
검색 조건에 해당하는 용례를 검색합니다.	
검색조건	
검색어	한국 +
검색어 포함 조건	OR
국적	국적을 선택하세요.
모국어	모국어를 선택하세요.
학습기관	학습기관을 선택하세요.
한국어 등급	<input type="checkbox"/> 1급 <input type="checkbox"/> 2급 <input type="checkbox"/> 3급 <input type="checkbox"/> 4급 <input type="checkbox"/> 5급 <input type="checkbox"/> 6급 <input type="checkbox"/> 6급 이상
나이	<input type="checkbox"/> 10대 <input type="checkbox"/> 20대 <input type="checkbox"/> 30대 <input type="checkbox"/> 40대 <input type="checkbox"/> 50대
학습목적	<input type="checkbox"/> 진학 <input type="checkbox"/> 취업 <input type="checkbox"/> 거주 <input type="checkbox"/> 취미 <input type="checkbox"/> 결혼
말뭉치 유형	<input type="checkbox"/> 문어 <input type="checkbox"/> 구어
학습환경	<input type="checkbox"/> 국내 <input type="checkbox"/> 국외
상세조건 ●	

<그림 1> 원시 말뭉치 검색 예시 1

표본번호	표본	중심어	원시 문맥	국적	모국어	급수	학습환경	자료유형
6012	"####"도	한국어를	공부합니다.	베트남	베트남어	1급	국내	문어
6012	"김철수"는	한국어를	공부합니다.	베트남	베트남어	1급	국내	문어
6048	"김철수 씨"도	한국어를	공부합니다.	프랑스	프랑스어	1급	국내	문어
5560	"아름다운 차 박물관"에서는	한국	차가 많이 있었어요.	일본	일본어	1급	국내	문어
1655	"저는 친구하고 같이	한국에	여행하고 싶어요."	중국	중국어(만다린어)	1급		문어
6820	# 선생님은	한국어를	못하는 우리한테 한국어를 한국어로	일본	일본어	1급	국내	문어
6820	# 선생님은 한국어를 못하는 우리한테	한국어를	한국어로 열심히 가르쳐 주셨습니다.	일본	일본어	1급	국내	문어
6820	# 선생님은 한국어를 못하는 우리한테 한국어를	한국어로	열심히 가르쳐 주셨습니다.	일본	일본어	1급	국내	문어
8310	##### 씨	한국에	살아요.	중국	중국어(만다린어)	1급	국내	구어
8307	##### 씨	한국에서	어디에 가 보고 싶어요?	중국	중국어(만다린어)	1급	국내	구어
8189	##### 씨는	한국에서	어디에 가 보고 싶어요?	대만	중국어(타이완어)	1급	국내	구어
4476	##### 씨는	한국어를	아주 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어
4593	'불고기'를 먹고	한국어를	공부했어요.	중국	중국어(만다린어)	1급	국내	문어
2993	10시간에	한국어	공부하다.	중국	중국어(만다린어)	1급	국내	문어
2941	10시부터 11시까지	한국	드라마를 봅니다.	나이지리아	영어	1급	국내	문어
2962	10시부터 11시까지	한국	드라마를 봅니다.	미얀마	버마어	1급	국내	문어
2951	10시부터 12시까지 도서관에서	한국어를	공부합니다.	말레이시아	중국어(만다린어)	1급	국내	문어

<그림 2> 원시 말뭉치 검색 결과 예시 1

- 예를 들어, 원시말뭉치를 통해 ‘한국’의 추출된 용례를 활용하여 학습자들의 ‘한국’과 관련한 합성어나 파생어 산출 양상이나, ‘한국’과의 공기 관계 어휘, 오류 등 ‘한국’과 관련한 학습자 언어 산출 양상을 학습자 변인이나 자료 변인에 따라 살펴볼 수 있을 것이다.
- 원시말뭉치 검색의 공기관계 검색을 활용하여 학습자의 국적이나 모국어 혹은 등급에 따른 두 어휘 간의 공기율을 살펴볼 수 있으며, 이를 활용하여 학습자의 언어 연구뿐 아니라 문화 적응도 및 문화 인식, 선호도 등 다양한 연구에 활용할 수 있을 것이다. 일례로 ‘한국’과 ‘좋다’, ‘나쁘다’의 공기 관계를 살펴보면 ‘좋다’와의 공기 관계가 훨씬 많음을 알 수 있어, 학습자들의 ‘한국’에 대한 평가의 한 면을 살필 수 있게 된다. 또 일정한 범위 내에서 연쇄하여 공기하는 어휘들도 함께 살펴볼 수 있어 학습자의 어휘 사용 능력을 측정해 볼 수 있다.

원시 말뭉치 검색
검색 조건에 해당하는 용례를 검색합니다.

검색어
한국
+
종
-

검색어 포함 조건
AND

국적
국적을 선택하세요.

모국어
모국어를 선택하세요.

학습기관
학습기관을 선택하세요.

한국어 등급
☒ 1급 ☐ 2급 ☐ 3급 ☐ 4급 ☐ 5급 ☐ 6급 ☐ 6급 이상

나이
☐ 10대 ☐ 20대 ☐ 30대 ☐ 40대 ☐ 50대

학습목적
☐ 진학 ☐ 취업 ☐ 거주 ☐ 취미 ☐ 결혼

말뭉치 유형
☐ 문어 ☐ 구어

☐ 학습환경
☐ 국내 ☐ 국외

상세조건

<그림 3> 원시 말뭉치 검색 예시 2

표본번호문맥	국적	모국어	급수	학습환경	자료유형
4476 #####시는 한국어를 아주 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어
2626 <name>대 한국어학당이 작고 시끄럽지만 밤 친구들과 같이 공부한	홍콩	중국어(광둥어)	1급		문어
2218 저는 한국 가수가 좋아해서 한국에 왔어요.	일본	일본어	1급		문어
2631 그래서 한국에서 재가 좋아하는 장소가 신촌입니다.	홍콩	중국어(만다린어)	1급		문어
4606 그래서 나는 한국에 아주 좋아해요.	중국	중국어(만다린어)	1급	국내	문어
3294 그래서 저는 한국어를 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어
2278 그래서 저는 한국은 좋아해요.	중국	중국어(만다린어)	1급		문어
7207 그래서 한국어를 아주 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어
3268 그래서, 저는 한국을 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어
8328 그러면 어 ##### 한국에서 어떤 것이 좋아요?	타이	타이어	1급	국내	구어
2243 그래서 한국이 좋아요.	중국	중국어(만다린어)	1급		문어
3241 그러지만, 한국이 좋습니다.	브라질	포르투갈어	1급	국내	문어
3129 그런데 저는 한국어 안 좋아해서 한국 예쁜 여자하고 이야기했어서	중국	중국어(만다린어)	1급	국내	문어
2208 그런데 한국에서 친구가 많은 한국 생활이 아주 좋아요.	브라질	포르투갈어	1급		문어
3268 그렇지만, 저는 한국을 아주 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어
2310 그리고 날마다 저는 한국어를 공부하려고 해서 집이 조용하고 밝은	중국	중국어(만다린어)	1급		문어
477 그리고 한국어를 잘 공부 후에 좋아하는 아이들은 유창하게 대화를 하	중국	중국어(만다린어)	1급		문어
3267 그리고 한국을 좋아합니다.	중국	중국어(만다린어)	1급	국내	문어

<그림 4> 원시 말뭉치 검색 결과 예시 2

2) 형태 주석 말뭉치 어휘 검색 활용

- 원시말뭉치의 어절 단위 검색 결과는 어형의 변화가 없는 명사나 부사, 관형사 등의 검색 시 좀 더 규모가 큰 말뭉치에서 용례를 검색할 수 있다는 장점이 있다. 그러나 용언이나 조사 등 어절 단위 검색을 통해 용례를 추출하기 번거롭거나 정확성이 떨어지는 품사의 학습자 용례 검색을 위해 형태소 분석이 되어 있는 형태소 분석 말뭉치를 활용할 수 있다. 형태소 분석 말뭉치를 통한 용례 검색도 위와 동일하며, 형태소 분석 말뭉치를 통해 검색된 각 형태소별 검색 결과를 활용하여 연구자가 자신의 연구 목적에 맞도록 주석함으로써, 학습자들의 각 형태소에 대한 오류나 어휘나 문법 형태의 발달 양상을 살펴볼 수 있을 것이다. 예를 들어 개별 어휘를 형태소 분석 말뭉치를 통해 검색한 결과를 활용하여 학습자들의 개별 어휘의 의미 사용 양상이나 문법 형태의 사용 양상을 관찰할 수 있을 것이다. 그 사례의 하나로 아래와 같이 동사 ‘가다’의 용례를 검색하여 학습자 변인이나 자료 변인에 따라 학습자들의 ‘가다’ 사용 양상을 살필 수 있으며, ‘가다’의 다의적 쓰임의 확대 양상을 관찰할 수 있을 것이다.

검색조건

검색어

☐ 세종학당표지

☐ 채권

☐ 명사

☐ 일반명사

☐ 고유명사

☐ 의존명사

☐ 대명사

☐ 수사

☐ 용언

☐ 지정사

☐ 긍정지정사

☐ 부정지정사

☒ 동사

☐ 형용사

국적

모국어

학습기관

한국어 등급 ☐ 1급 ☐ 2급 ☐ 3급 ☐ 4급 ☐ 5급 ☐ 6급 ☐ 6급 이상

나이 ☐ 10대 ☐ 20대 ☐ 30대 ☐ 40대 ☐ 50대

학습목적 ☐ 진학 ☐ 취업 ☐ 거주 ☐ 취미 ☐ 결혼

말뭉치 유형 ☐ 문어 ☐ 구어

학습환경 ☐ 국내 ☐ 해외

<그림 5> 형태 주석 말뭉치 검색 예시 1

표본번호	앞 문맥	중심어	뒤 문맥	국적	모국어	급수	학습환경	자료유형
69	그리고 오후 1시에 제 아내하고 식당에	가요.		독일	독일어	1급	국내	문어
69	오후 2시부터 4시까지 공부하러 도서관에	가요.		독일	독일어	1급	국내	문어
69	오후 4시에 신문 읽으러 커피숍에	가요.		독일	독일어	1급	국내	문어
69	그다음에 운동하러 운동장에	가요.		독일	독일어	1급	국내	문어
69	보통 오후 7시에 저녁 식사 하러 식당에	가요.		독일	독일어	1급	국내	문어
70	보통 여덟 시 삼십오 분에 학교에	가요.		독일	독일어	1급	국내	문어
70	오후 세 시에 집에	가요.		독일	독일어	1급	국내	문어
512	그래서 다음 달에 고양이	가다.		독일	독일어	3급		문어
512	그래서	갔다	오다.	독일	독일어	3급		문어
553	또, 교통으로 타고	갈	때, 장애인한테 있는 차리가 많지	독일	독일어	4급		문어
689	8급에 학생들은 쉬는 시간에 학교 근처 area에	가지	안 됐다.	독일	독일어	2급		문어
689	'학교 앞 기게에	감시다!'	{우 문맥 없음}	독일	독일어	2급		문어
689	그리고 쉬는 시간 시작자마자 우리는 학교 앞에	갔다.		독일	독일어	2급		문어

<그림 6> 형태 주석 말뭉치 검색 결과 예시 1

- 형태 주석 말뭉치 검색을 활용하여 학습자의 모국어나 국적, 등급, 학습 목적 등 학습자 변인이나 말뭉치 유형 등의 언어 자료 변인에 따른 어휘나 문법 형태소의 사용량을 측정할 수 있으므로 이를 통해 학습자의 언어 발달이나 언어 자료의 매체에 따른 차이를 추론할 수 있을 것이다. 예를 들면 문법 형태 중 ‘-어서’와 ‘-니까’의 사용을 학습자 변인이나 자료 변인별로 살펴봄으로써 사용 양상과 문법 형태의 언어 발달 양상, 그리고 유사 문법 형태의 선호 양상 등을 관찰할 수 있다. 즉 아래와 같이 ‘-어서’와 ‘-니까’의 용례를 검색하여 총 사용빈도뿐 아니라 학습자 변인이나 자료 변인별 사용빈도의 차이, 이들 문법 형태의 다의적 쓰임의 확대 양상 등을 관찰할 수 있으며, 연구자가 이들 용례에 오류 주석을 함으로써 오류 양상이나 언어 발달 양상 등을 관찰할 수 있을 것이다.

표본번호	말 문맥	중심어	덧 문맥	국적	모국어	급수	학습환경	자료유형
5844	그렇지만 최저기온은 영하		2℃쯤니까 서울보다 따뜻해요.	일본	일본어	1급	국내	문어
2613	부산에		가니까 여자 친구에게 프로포즈를 했어요.	캐나다		1급		문어
4424	그러지만 열자친구 못		가니까 시험이 있었어요.	중국	중국어(만다린)	1급	국내	문어
4429	저는 반 년 전에 한국에		가니까 한국 친구가 없었어요.	중국	중국어(만다린)	1급	국내	문어
2241	저는 한국어		공부하니까 한국에 왔어요.	중국	중국어(만다린)	1급		문어
4456	그래서 다음 학기부터 저는 열심히		공부하니까 저는 대학생이 되고 싶어요.	중국	중국어(만다린)	1급	국내	문어
4489	저는 매일 열심히		공부하니까 내년 계획이 많이 있어요.	중국	중국어(만다린)	1급	국내	문어
5552	[좌 문맥 없음]		그러니까 좀 잘 좀니다.	몽골	몽골어	1급	국내	문어
577	[좌 문맥 없음]		그렇니까 시원한 옷들과 우산이 준비하세요.	말레이시아	말레이어	1급		문어
577	[좌 문맥 없음]		그렇니까 선크림과 모자도 가져오세요.	말레이시아	말레이어	1급		문어
1753	[좌 문맥 없음]		그렇니까 저는 봄을 좋아해요.	미국	영어	1급		문어
2288	[좌 문맥 없음]		그렇니까 한국에 왔어요.	중국	중국어(만다린)	1급		문어
3996	[좌 문맥 없음]		그렇니까 , 제 친구(22)등산을 하고 싶어해요.	인도네시아	인도네시아어	1급	국내	문어
5897	[좌 문맥 없음]		그렇니까 서울 야경을 보십시오.	타이	타이어	1급	국내	문어
5897	[좌 문맥 없음]		그렇니까 가게를 구경하고 쇼핑도 하십시오.	타이	타이어	1급	국내	문어
1094	영화가		끝니까 같이 밥을 먹었어요.	중국	중국어(만다린)	1급		문어
2261	저는 월요일부터 금요일까지 학교에서 공부를 하고 힘(1)니까		한국 생활이 즐거워요.	중국	중국어(만다린)	1급		문어

<그림 7> 형태 주석 말뭉치 검색 결과 예시 2

표본번호	말 문맥	중심어	덧 문맥	국적	모국어	급수	학습환경	자료유형
3510	좋은		가게이어서 사람이 많았어요.	대만	중국어(만다린)	1급		문어
352	나는 한국에 오기 전에 한국과 중국		가까워서 예절이 비슷하다고 생각하고 있다.	중국	중국어(만다린)	1급		문어
2368	이 친구 생일하고 제 생일이		가까워서 같이 놀려고 해요.	일본	일본어	1급		문어
4575	아침에 어머니와 같이 백화점에 소		가까워서 지하철도를 타고 갔어요.	중국	중국어(만다린)	1급	국내	문어
3541	그 지하철도에서 KBS가		가까워서 결코 오 분 걸려요.	미국	영어	1급	국내	문어
3355	우리 집은 학교에서 너무		가까워서 좋아요.	미국	영어	1급	국내	문어
2269	전철한 선생님		가르쳐서 기분이 좋아요.	중국	중국어(만다린)	1급		문어
2226	제가 한국에 가는 이유는 부모님께		가봤어서 "저도 가고 싶어요"를 생각했어서 왔	일본	일본어	1급		문어
2226	한국 생활이 모든 것이 많이 있지만 재미있어서		좋아해요.	일본	일본어	1급		문어
2511	방탄소년단은 회사에서 한식집까지 건너서		오보름 걸려요.	대만	중국어(만다린)	1급		문어
117	여덟 시 사십 분에 학교에		걸려서 가요.	이탈리아	이탈리아어	1급	국내	문어
3626	제 일급 교실이 사중예 있는데 날마다		걸려서 교실에 올라옵니다.	홍콩	중국어(광둥어)	1급		문어
3987	8시 50분 교시에		걸려서 갑니다.	조지아	조지아어	1급	국내	문어
396	빙이점 역에서 비행기로 3시간쯤		걸려서 해남도역에 도착했어요.	중국	중국어(만다린)	1급		문어

<그림 8> 형태 주석 말뭉치 검색 결과 예시 3

- 형태 주석 말뭉치 검색을 통한 추출된 용례에 연구자가 스스로 오류를 주석하여 이를 연구에 활용할 수도 있으나, 오류 판단의 주관성으로 인해 연구의 신뢰도를 검증해야 할 필요가 생기게 된다. 그러나 오류 주석 말뭉치 검색을 활용하면 학습자의 오류와 비오류 생산 어휘나 문법 형태를 살펴볼 수 있어 오류 연구나 중간언어 발달 연구에 유용하게 활용할 수 있다. 예를 들어 오류 위치나 오류 유형, 오류 층위만을 선택하여 학습자 언어 사용 중 오류만 추출하여 학습자 언어 사용의 오류 양상에 대한 연구를 수행할 수 있을 것이다. 그 일례로 주격조사 ‘이/가’의 누락 오류만을 추출하면 다음과 같으며, 이를 활용하여 주격조사 ‘이/가’의 누락 환경을 고찰할 수 있다.

표본번호	교정 형태	말 문맥	중심어	덧 문맥	오류 위치	오류 양상	오류 층위	국적	모국어	급수	학습 환경	학습 환경	자료 유형
1011	이	"슈포맨이 들어왔다"에 남자	연예인	49시간 아내 없이 혼자서 아기들과 함께 노는 것은 프로그램이다.	FNF	OM		카자흐스탄	카자흐어	3급			문어
2779	가	이번에는	상성전자	7년 후에 또 다른 로봇이을 만들었어요.	FNF	OM		미국	영어	6급			문어
1979	이	[좌 문맥 없음]	여행	가고 싶으면 아무 데나 갈 수 있으며 직장에서 나보다 더 높은 지위의 사람도 없고 부담을 찾지 않는다.	FNF	OM		마카오	중국어(광둥어)	4급			문어
2013	이	이제 어린이인지 어른인지 상관없이 인터넷을	사용	가능해서 현재 문제를 생겼다.	FNF	OM		러시아	러시아어	5급			문어
7779	이	그래서 경제적인 여유가 생기고 월급 인상 요인이 있을 때도 하루 문제 없이	인상	가능했다고 한다.	FNF	OM		미얀마	버마어	6급	연세	국내	문어

245	이	우리 반	선생님	가르쳐 한국어 좋아요.	FNF	OM		중국	중국어 만다린어	1급				문어
5799	가	동네에 미국 사람, 중국 사람, 이라크인 사람도 있지만 재미	교포	가장 살고 있다.	FNF	OM		미국	영어	2급	이화대	국내		문어
4328	가	엑소는 유명한 남성 12명 있는 아이돌 그룹인데	재	가장 좋아하는 멤버는 찬열이라고 한다.	FNF	OM		타이	타이어	2급	한국대	국내		문어
5802	이	어머니가 어떤	장면	가장 좋아하는 장면이라고 하셨을 때 잘 대답할 수 없었다.	FNF	OM		미국	러시아어	3급	이화대	국내		문어
6499	가	그런데 나는 일을 할수록 그런 조건보다 앞으로 발전 가능성과 정조적으로 일할 수 있는	것인지	가장 중요하다고 생각한다.	FNF	OM		베트남	베트남어	3급	고려대	국내		문어
7890	가	직장을 선택할 때 복지가 어떻게	되는지	가장 중요하다고 생각한다.	FNF	OM		중국	중국어 만다린어	3급	고려대	국내		문어
2799	이	그래서 나는 20년 동안 살아온 이 인생에서	지금	가장 행복하고 살고 있는 것을 살피고 있다.	FNF	OM		일본	일본어	6급				문어
2899	이	XX 유학은 회사에 자유로운 분위기와 XX	유학만	가지고 있는 독특한 색깔이 있다고 생각하고 매력적이었습니다.	FNF	OM		일본	일본어	4급	서강대	국내		문어
3624	이	최근 사회에서 남자 전업주부를 받을 수 없는 것이 대부분	사람들	가지고 있는 생각이다.	FNF	OM		중국	중국어 만다린어	5급	서울대	국내		문어
7908	가	그리고 저출산으로 인해 생산	인구	감소하고 2060년에 생산 인구가 현재에 비해 52% 수준으로 감소할 전망이다.	FNF	OM		중국	중국어 만다린어	6급	고려대	국내		문어

<그림 9> 형태 주석 말뭉치 내려받기 기능을 활용한 직접 주석 예시

3) 오류 주석 말뭉치 어휘 검색 활용

- 오류 주석 말뭉치는 오류 대상 어휘 및 문법 형태를 기준으로 검색할 수 있을 뿐만 아니라 교정 형태를 기준으로 검색할 수 있어, 사용상의 오류나 어휘나 문법 형태의 미적용 오류를 함께 살펴볼 수 있다. 예를 들어 연결어미 ‘-다가’의 오류를 원형태와 교정형태를 기준으로 검색하면 아래와 같이 학습자의 연결어미 ‘-다가’ 사용을 양면에서 고찰할 수 있다.

오류주석 말뭉치 검색

검색 조건에 해당하는 용례를 검색합니다.

복합 주석 검색

자소 검색

문형 검색

검색조건

원 형태

다가

OR

교정 형태

다가

오류 위치

오류 양상

오류 중위

☐ 오류 위치

☐ 실절어휘

☐ CNNP(고유명사)
 ☐ CNNG(일반명사)
 ☐ CNNB(외존명사)
 ☐ CNP(대명사)

☐ 오류 양상

☐ OM(누락)
 ☐ ADD(첨가)
 ☐ REP(대치)
 ☐ MIF(오형태)

☐ 오류 중위

☐ 발음

☐ PP(음소)
 ☐ PS(음절)
 ☐ PC(음운규칙)
 ☐ 원어식 발음

<그림 10> 오류 주석 말뭉치 검색 예시 1

- 오류 주석 말뭉치의 자소 검색을 활용하여 학습자의 맞춤법이나 음소 오류 등을 살필 수 있다. 예를 들어 어두 자음 검색을 통해 어두 폐쇄음 발음(평음, 경음, 격음) 오류 등을 분석할 수 있으며, 어말 자음 검색을 통해

- 31 -

중성 발음 오류를 분석할 수 있다. 그 사례로 어말 자음 ‘ㄹ’을 오류 분석 말뭉치를 통해 검색함으로써 중국인 학습자들의 어말 자음 ‘ㄹ’의 오류를 분석해 볼 수 있을 것이다.

표본 번호	원 형태	교정 형태	앞 문맥	중심어	뒤 문맥	오류 위치	오류 양상	국적	모국어	급수
8276	개	개월	제 아 지난 일	개	동안어공후를 하 하였습니다.	CNN B	REP	중국	중국어(만 다린어)	1급
2579	겨울	겨울	{좌 문맥 없음}	겨울은	날씨가 아주 춥지만 눈이 안 옵니다.	CNN G	MIF	중국	중국어(만 다린어)	1급
2593	겨울	겨울	올해	겨울에	한국에 있었습니다.	CNN G	MIF	중국	중국어(만 다린어)	1급
8304	교시	교실	{좌 문맥 없음}	교시에	가요.			중국	중국어(만 다린어)	1급
329	내엘	내일	{좌 문맥 없음}	내엘에	한국어 수업이 준비합니다.	CNN G	MIF	중국	중국어(만 다린어)	1급
2302	늘	늘	집에서 노래하고	늘어도	됩니다.	CVV	MIF	중국	중국어(만 다린어)	1급
2271	드	들	기숙사의	친구드을	같이 명동에 갔어요.	CXS N	MIF	중국	중국어(만 다린어)	1급
8302	친구 들	들	{좌 문맥 없음}	친구들하 고	같이 공원에			중국	중국어(만 다린어)	1급
1593		ㄹ	내년에 저는 한국어를 잘	겁니다.				중국	중국어(만 다린어)	1급
8304	바다 으	를	{좌 문맥 없음}	바다으				중국	중국어(만 다린어)	1급
8304	신바	신발	우편하고	신바				중국	중국어(만 다린어)	1급
8275	에버	에별	에	에버	에벌랜드 아두			중국	중국어(만 다린어)	1급
8302	우요 일	월요 일	매주	우요일부 터	토요일까지			중국	중국어(만 다린어)	1급
8302	쭈마	정말	글래스 봄을 쭈 어	쭈마	좋아해요.			중국	중국어(만 다린어)	1급
320	체일	제일	저는 이제 고향 친구들이	체일	만나고 싶습니다.	CMA G	MIF	대만	중국어(만 다린어)	1급
1648	주물	주말	저는 한국에서	주물을	집에 있습니다.	CNN G	MIF	중국	중국어(만 다린어)	1급
242	주알	주말	{좌 문맥 없음}	주알에	친구 집 만납니다.	CNN G	MIF	중국	중국어(만 다린어)	1급
2598	치하 절	지하 절	밥을 먹을 후에	치하절으 로	갔습니다.	CNN G	MIF	중국	중국어(만 다린어)	1급
329	참이 술	참이 술	{좌 문맥 없음}	참이술보 다	맥주 더 맛있습니다.	CNN P	MIF	중국	중국어(만 다린어)	1급

<그림 11> 오류 주석 말뭉치 내려받기 기능을 활용한 직접 주석 예시

- 이와 같이 형태 주석 말뭉치와 오류 주석 말뭉치를 활용한 오류 분석이나 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 있어 이를 교수요목 설계나 교재 개발에 활용할 수 있을 것이다.
- 또 학습자 사전 편찬에 활용될 수 있어, 학습자 말뭉치에서의 오류 검색을 통한 결과를 활용하여 학습자들이 생산한 오류를 예시할 수 있으며, 사전 사용자들이 그러한 오류를 범하지 않도록 방지할 수 있도록 사전에 정보를 제공할 수 있을 것이다. 그리고 학습자의 어휘, 문법 사용 빈도를

통해 학습자 사전 표제어 선정에도 기초 자료로 활용할 수 있을 것이다.

2.2. 교사를 위한 활용 모형

- 오류 주석 말뭉치에서의 어휘나 문법 형태의 오류를 검색함으로써 등급이나 언어권에 따른 학습자 언어 사용 양상과 오류 양상을 관찰하고, 이를 통해 학습자의 오류의 진단과 처방, 그리고 합리적인 오류의 예방을 할 수 있다. 특히 학습자의 숙달도와 모어, 학습 상황 등 다양한 변인에 따른 검색이 가능하므로 각 변인별 오류 양상을 관찰하여 집단별 언어 사용 양상을 파악하고 각각에 적합한 맞춤형 처방을 내릴 수 있다. 일례로 1급 중국인 학습자의 ‘-어서’ 오류를 검색함으로써 교실에서 발생할 수 있는 오류를 미리 예측할 수 있는데, 아래 보는 바와 같이 중국인 1급 학습자들이 ‘-어서’를 ‘-는데, -어도, -니까’ 등 학습한 유사 의미의 연결어미와 혼용하는 사례를 볼 수 있으며, 교실에서 이러한 오류가 발생하지 않도록 처방할 수 있을 것이다.

<그림 12> 오류 주석 말뭉치 검색 예시 2

No.	모국어	급수	원쪽 문맥	중심어	교정형태	오른쪽 문맥	오류주석
1	중국어(만다린어)	1급	단풍이	예쁘니까	어서/EC	등산하는 사람이 많...	FED REP
2	중국어(만다린어)	1급	...매일 저는 친구와	같이요	어서/EC	학교에 가요.	FED REP
3	중국어(만다린어)	1급	아름에 드는 치마가	있서	어서/EC	저도 치미를 샀습니...	FED MF MCJ
4	중국어(만다린어)	1급	여기를 많이 사람이	있어서	오니까/EC	웃이 사 보세요.	FED REP
5	중국어(만다린어)	1급	여름에는 정말	더워서	지만/EC	수영을 할 수 있어요.	FED REP
6	중국어(만다린어)	1급	우리 반 선생님	가르쳐	어서/EC	한국어 좋아요.	FED REP
7	중국어(만다린어)	1급	저는 담배를 안	피우는데	어서/EC	담배를 피우면 안...	FED REP
8	중국어(만다린어)	1급	...저는 담배를 안 피	워서	아도/EC	관합니다.	FED REP

<그림 13> 오류 주석 말뭉치 검색 결과 예시 1

- 따라서 오류 주석 말뭉치에서의 검색 결과를 활용하여 각 숙달도나 언어권에 따른 효과적인 교수법을 구안할 수 있으며, 차별화된 교수법을 개발할 수 있을 것이다. 또한 학습자의 오류를 미리 예측할 수 있으므로 이를 교안에 활용하기도 하여 교수에 직접 이용 가능할 것이다.
- 학습자가 오류 자료를 통해 오류의 양상을 인식하고 학습 항목의 용법을 스스로 발견하고 주도적으로 배우게 함으로써 학습에서 습득으로의 전환을 용이하게 한다. 따라서 각 숙달도나 언어권에 적합한 오류문을 추출하여 수업 활동 및 수업 자료를 개발하여 교수에 직접 활용할 수 있다.
- 본 학습자 말뭉치에서는 교사의 말뭉치 활용도를 제고하기 위하여 어휘나 문법 형태에 의한 검색뿐 아니라 소위 문형이라는 문법적 표현을 중심으로도 학습자 언어 사용을 관찰할 수 있도록 하였다. 일례로 학습자들의 ‘-고 싶다’와 관련한 언어 사용 양상을 살피고 싶으면 문형 검색을 활용하여 다음과 같은 용례 추출이 가능하며, 이를 교수 자료나 교실 활동에 활용할 수 있을 것이다.

오류주석 말뭉치 검색 | 검색 조건에 해당하는 용례를 검색합니다.

복합 주석 검색 자소 검색 문형 검색

검색조건

검색 문형 -고 싶다 검색

상세조건

-게 되다	-게 마련이다	-게 만들다	-게 생겼다	-게 하다	-고 나다
-고 들다	-고 밀다	-고 보다	-고 싶다	-고 싶어 하다	-고 있다
-고 해서	-고는 하다	-곤 하다	-기 나뭇이다	-기 때문	-기 마련이다
-기 십상이다	-기 위한	-기 위해(서)	-기 일쑤이다	-기 전에	-기 짝이 없다
-기가 무섭게	-기가 바쁘게	-기가 쉽다	-기나 하다	-기로 들다	-기로 하다
-기만 하다	-기에 따라	-기에 앞서(서)	-나 보다	-나 싶다	-는 가운데
-는 것	-는 것 같다	-는 고사하고	-는 길에	-는 김에	-는 대로

<그림 14> 오류 주석 말뭉치 검색 예시 2

검색결과5901건 급수 오류자수 20개씩 보기

No.	모국어	급수	대표문형	용례
201	러시아어	1급	-고 싶다	가족이 보고 싶습니다 .
202	중국어(만다린어)	1급	-고 싶다	가족하고 밥을 먹고 싶습니다 .
203	중국어(만다린어)	1급	-고 싶다	가족하고 여행을 하고 싶습니다 .
204	중국어(만다린어)	1급	-고 싶다	가족하고 영화도 보고 싶습니다 .
205	중국어(만다린어)	1급	-고 싶다	가족들이 보고 싶어요 .
206	중국어(광둥어)	1급	-고 싶다	가족와 친구를 만나고 싶다 .
207	일본어	1급	-고 싶다	같이 가고 싶어요 .
208	일본어	1급	-고 싶다	개인 주택은 마당도 있는데 아름다운 곳을 볼 수 있어서 살고 싶습니다 .
209	중국어(만다린어)	1급	-고 싶다	거기는 교통이 편리하고 경치가 좋아서 거기에 살고 싶습니다 .
210	일본어	1급	-고 싶다	거기서 많은 친구를 만들고 싶어요 .

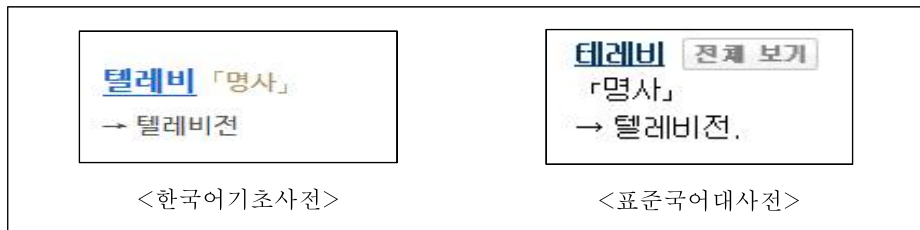
<그림 15> 오류 주석 말뭉치 검색 결과 예시 3

- 형태소 분석 말뭉치에서의 특정 어휘나 문법 형태를 포함한 용례를 검색함으로써 학습자의 실제 사용 용례를 관찰할 수 있으며, 이를 활용하여 교재나 교수 자료에서의 예문에 활용하여 학습자로 하여금 좀 더 친숙한 언어로 새로운 어휘나 문법을 학습하게 할 수 있을 것이다.
- 성취도 평가, 숙달도 평가 문항 개발 시 학습자가 공통적으로 일으킨 오류를 활용하여 문항 선정 및 오답지 개발에 활용하여 평가의 타당성을 제고할 수 있을 것이다.

2.3. 학습자를 위한 활용 모형

1) 학습자 말뭉치 출현 오류형을 반영한 가표제어 추가

- 일반적으로 학습자 사전에서는 학습자가 오류형으로 사전을 검색할 것을 예상하여 사전의 거시구조에 가표제어를 설정하는데, 아래의 예와 같이 모국어 화자를 대상으로 한 사전의 정보를 참고하는 것이 기존의 방식이다. 유용한 정보이기는 하지만 실제 학습자의 오류 양상을 충분히 반영하였다고 볼 수 없다.



<그림 16> <한국어기초사전>과 <표준국어대사전>의 표제어 제시 방식

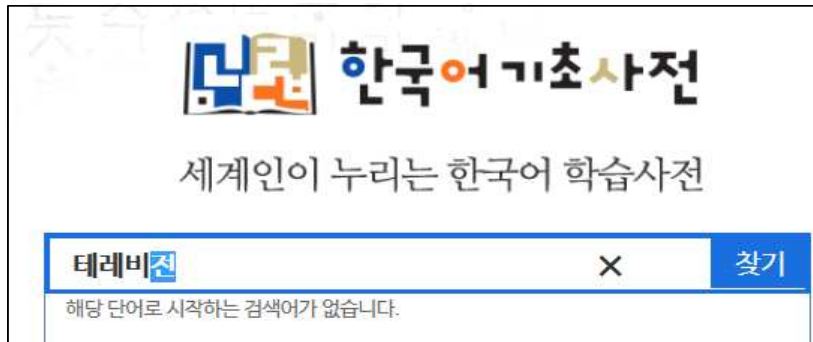
- 학습자 말뭉치의 오류 주석 결과를 활용하면 실제로 학습자들이 검색할 가능성이 있는 오류형을 가표제어로 설정할 수 있게 된다. 학습자 말뭉치 오류 주석 결과를 분석해 보니 ‘텔레비전’의 오류형으로 ‘텔레비정, 텔레비 전, 테레비전’ 등이 출현하였다. 이런 오류형 중 빈도가 높은 것을 가표제어로 설정하면 학습자의 사전 검색 편의성을 제고할 수 있다.

중국어(만다린어)	3급	저도 쉬는 시간에	텔레비정에	텔레비전/NNG	방송을 보는 것이...	CNNG MEF
중국어(만다린어)	3급	...최근 사람들이 쉬는	텔레비정에	텔레비전/NNG	방송을 보는 것이...	CNNG MEF
중국어(만다린어)	1급	나는 같이	텔레비전을	텔레비전/NNG	봅니다.	CNNG MEF
몽골어	1급	제 친구는	테레비전에서	텔레비전/NNG	명중를 보고 비빔밥...	CNNG MEF

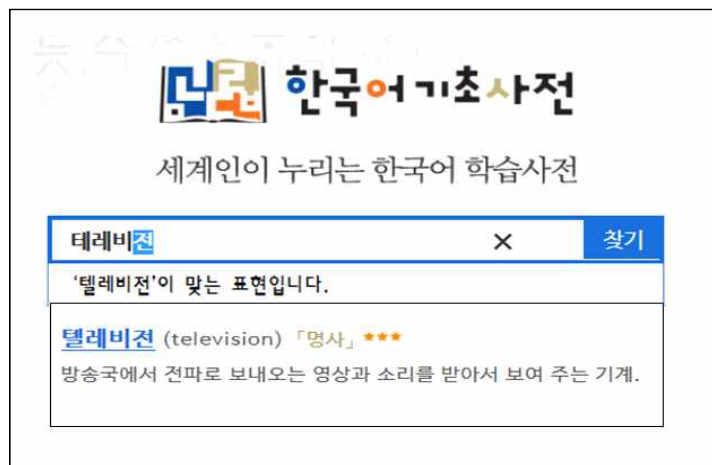
<그림 17> 오류 주석 학습자 말뭉치 용례 검색 결과: 텔레비전

2) 학습자 오류 형태 목록에 기반한 수정 정보 제시

- 학습자 사전의 검색 편의성을 높이기 위해 포털사이트와 같은 방식으로 오류 검색어에 대한 수정 정보를 제시하고 수정된 검색어에 대한 검색 결과를 바로 제시할 수 있다. 학습자 말뭉치의 오류 주석을 통해 오류 형태와 교정 형태의 대응 자료를 구축함으로써 가능한 방식이다.



<그림 18> <한국어기초사전> 검색 결과



<그림 19> 학습자 말뭉치 오류 형태를 반영한
<한국어기초사전> 검색 결과 모형

3) 미시 구조의 오류 형태 목록 제시

- 오류 주석 학습자 말뭉치 분석을 통해 얻게 된 오류 형태의 목록을 학습자 사전 미시 구조의 한 항목으로 제시할 수 있다. 정확한 형태로 검색했다고 하더라도 함께 공부하는 학습자들이 범할 수 있는 오류를 파악하거나 복수 표준어에 대한 정보를 얻을 수 있어 추가를 고려할 만한 항목이다.

텔레비전 (television) 「명사」 ★★★ 방송국에서 전파로 보내오는 영상과 소리를 받아서 보여 주는 기계.	
유의어	<u>티브이</u>
참고어	<u>바보상자</u>
오류 형태	텔레비전(X), 텔래비전(X), 텔레비정(X), 테레비전(X)

<그림 20> 학습자 말뭉치 오류 형태 목록을 제시한
 <한국어기초사전> 미시구조 모형

4) 오류 예문 제시

- 오류 주석 학습자 말뭉치 분석을 통해 얻게 된 학습자의 오류 중에 형태는 바르게 되어 있으나 문법적, 화용적 용법에 어긋나는 표현은 예문으로 미시 구조에 포함시킬 수 있다. 아래 예와 같은 표현 문형이 오류 예문 제시가 필요한 대표적인 사례이다.

중국어(만다린어)	5급	-기 짝이 없다	광고는 지금 거의 유명한 배우들이 출연하기 때문에 광고 비용이 증가하기 짝이 없다.
중국어(만다린어)	3급	-기 짝이 없다	그래서 광고가 관리하기 짝이 없다.

<그림 21> 학습자 말뭉치에 나타난 표현 문형 오류 포함 문장의 예

- 학습자 말뭉치의 용례를 기반으로 예문을 제시할 때에는 해당 표제어 외의 부분에서 오류를 범하지는 않았는지, 너무 단순해서 표제어의 의미를 제대로 전달하지 못하지는 않는지, 맥락을 파악하기 힘들지는 않는지, 너

무 복잡해서 학습자가 이해하기 어렵지는 않은지 등을 살펴서 정제할 필요가 있다. ‘그래서 광고가 편리하기 짝이 없다’는 접속부사 때문에 맥락을 파악하기 힘들고 공기 관계가 어색하여 예문으로 부적절하다.

○ 다음은 학습자의 자기 주도 학습에 활용할 수 있는 방안들이다.

5) 오류의 확인 및 수정

○ 오류 주석 학습자 말뭉치는 활용 시스템의 검색 기능을 통하여 학습자들이 스스로 생성한 오류, 혹은 사용하기에 자신이 없는 표현에 대한 정보를 확인하는 데에 활용될 수 있다. 해당 표현이 오류인지의 여부와 오류에 대한 교정 형태를 같이 제시하기 때문에 오류의 발견과 수정 활동이 동시에 이루어지게 된다. 예를 들어 한국어 학습자들은 ‘한국’의 철자를 ‘학국’으로 틀리게 적는 경우가 많다. 뒤에 오는 음절의 초성과 종성이 모두 ‘ㄱ’인 것에 이끌려 ‘한’의 받침을 ‘ㄱ’으로 잘못 쓰는 것이다. 어떤 학습자가 자신이 알고 있는 ‘학국’이라는 철자에 자신이 없을 때에 사진을 찾으려면 검색 결과를 얻지 못해 당황할 수 있지만, 학습자 말뭉치 활용 시스템에서 오철자를 포함한 ‘학국’을 검색하게 되면 다음과 같은 결과를 통해 ‘학국’이라는 중심어가 형태를 ‘한국’으로 교정해야 할 오류임을 확인하고 바른 표현을 학습하게 된다.

모국어	급수	왼쪽 문맥	중심어	교정형태	오른쪽 문맥	오류주석
영어	1급	[좌 문맥 없음]	학국	한국/NNP	생화가 재미있습니다.	CNNP MIF
중국어(만다린어)	3급	[좌 문맥 없음]	학국	한국/NNP	회사든지 중국 회사...	CNNP MIF
중국어(만다린어)	1급	[좌 문맥 없음]	학국에서	한국/NNP	날씨는 자주 비가...	CNNP MIF
중국어(만다린어)	2급	[좌 문맥 없음]	학국에서	한국/NNP	책도 비싸다.	CNNP MIF
일본어	2급	그 친구도	학국이	한국/NNP	좋아하고 한국어도...	CNNP MIF

<그림 22> 오류 주석 학습자 말뭉치 용례 검색 결과: 학국

6) 분류별 오류 양상 파악

○ 학습자의 메타 정보에 기반한 오류 양상

학습자 말뭉치 구축 시 텍스트를 생산한 학습자의 메타 정보를 함께 구축하므로, 학습자가 직접 모어, 등급, 학습 기관, 학습 환경, 학습 목적 등의 관점에서 자신과 같은 조건을 가진 학습자가 생성한 오류를 부류별로 모아 볼 수 있다. 학습자 말뭉치 활용 시스템의 상세 조건 검색 기능에 구현된 학습자 메타 정보를 보이면 다음과 같다.

모국어	<input type="checkbox"/> 중국어(만다린어) <input type="checkbox"/> 중국어(광둥어) <input type="checkbox"/> 중국어(타이완어) <input type="checkbox"/> 중국어(푸젠어) <input type="checkbox"/> 중국어(민난어) <input type="checkbox"/> 베트남어 <input type="checkbox"/> 태국어 <input type="checkbox"/> 인도네시아어 <input type="checkbox"/> 필리핀어 <input type="checkbox"/> 캄보디아어 <input type="checkbox"/> 미얀마어 <input type="checkbox"/> 라오스어 <input type="checkbox"/> 몽골어 <input type="checkbox"/> 티베트어 <input type="checkbox"/> 네팔어 <input type="checkbox"/> 불어 <input type="checkbox"/> 영어 <input type="checkbox"/> 일본어 <input type="checkbox"/> 한국어(광둥어) <input checked="" type="checkbox"/> 중국어(만다린어) <input type="checkbox"/> 베트남어 <input type="checkbox"/> 태국어 <input type="checkbox"/> 인도네시아어 <input type="checkbox"/> 필리핀어 <input type="checkbox"/> 캄보디아어 <input type="checkbox"/> 미얀마어 <input type="checkbox"/> 라오스어				
	<input type="checkbox"/> 몽골어 <input type="checkbox"/> 티베트어 <input type="checkbox"/> 네팔어 <input type="checkbox"/> 불어 <input type="checkbox"/> 영어 <input type="checkbox"/> 일본어				
학습기관	<input type="checkbox"/> 경복대 <input type="checkbox"/> 경희대(서울) <input type="checkbox"/> 고려대 <input type="checkbox"/> 국민대 <input type="checkbox"/> 에세대 <input type="checkbox"/> 서강대 <input type="checkbox"/> 서울대 <input type="checkbox"/> 신라대 <input checked="" type="checkbox"/> 연세대 <input type="checkbox"/> 이화여대 <input type="checkbox"/> 홍익대 <input type="checkbox"/> 한국외대 <input type="checkbox"/> 한남대 <input type="checkbox"/> 한양대 <input type="checkbox"/> 북일대				
	<input type="checkbox"/> 서울대 <input type="checkbox"/> 연세대 <input type="checkbox"/> 이화여대 <input type="checkbox"/> 홍익대				
한국어 등급	<input checked="" type="checkbox"/> 1급 <input type="checkbox"/> 2급 <input type="checkbox"/> 3급 <input type="checkbox"/> 4급 <input type="checkbox"/> 5급 <input type="checkbox"/> 6급 <input type="checkbox"/> 6급 이상				
나이	<input type="checkbox"/> 10대 <input checked="" type="checkbox"/> 20대 <input type="checkbox"/> 30대 <input type="checkbox"/> 40대 <input type="checkbox"/> 50대				
학습목적	<input checked="" type="checkbox"/> 진학 <input type="checkbox"/> 취업 <input type="checkbox"/> 거주 <input type="checkbox"/> 취미 <input type="checkbox"/> 결혼				
말뭉치 유형	<input checked="" type="checkbox"/> 모어 <input type="checkbox"/> 구어		학습환경 <input checked="" type="checkbox"/> 국내 <input type="checkbox"/> 국외		

<그림 23> 한국어 학습자 말뭉치 활용 시스템의 검색 상세 조건

○ 주석 체계에 기반한 오류 양상

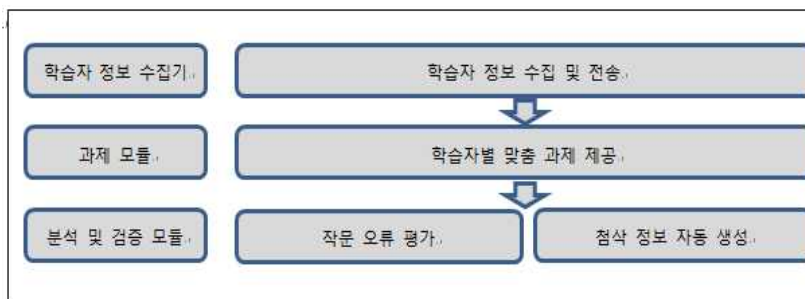
학습자 말뭉치 활용 시스템을 이용하면 학습자가 스스로 취약하다고 생각하는 형태나 문법 범주를 포함하는 자료만 검색하여 오류에 대한 교정 형태를 상세히 살펴보는 맞춤형 학습도 가능하다.

오류 위치	오류 양상	오류 층위
<input type="checkbox"/> 오류 위치 <input type="checkbox"/> 실절어휘 <input type="checkbox"/> CNNP(고유명사) <input checked="" type="checkbox"/> CNNG(일반명사) <input type="checkbox"/> CNNB(의존명사) <input type="checkbox"/> CNP(대명사)	<input type="checkbox"/> 오류 양상 <input type="checkbox"/> OM(누락) <input type="checkbox"/> ADD(첨가) <input checked="" type="checkbox"/> REP(대치) <input type="checkbox"/> MIF(오형태)	<input type="checkbox"/> 통사 <input checked="" type="checkbox"/> SH(높임) <input type="checkbox"/> ST(시제) <input type="checkbox"/> SC(사동) <input type="checkbox"/> SP(피동) <input type="checkbox"/> SN(부정) <input type="checkbox"/> WO(어순)

<그림 24> 한국어 학습자 말뭉치 활용 시스템의 오류 주석 기반 검색 조건

7) 학습자 말뭉치에 기반한 자동 첨삭 시스템 개발

- 학습자들이 교육 기관을 통하지 않고 스스로 언어를 학습하는 방법에는 교재 활용, 이러닝, 애플리케이션 활용 등이 있다. 이러한 매체들은 양방향 커뮤니케이션이 이루어질 수 없기 때문에 학습 내용 테스트에 대한 피드백 제공이 어렵다는 한계점을 지닌다. 하지만 학습자 말뭉치를 데이터 베이스로 하는 자동 첨삭 시스템을 개발한다면 이러한 한계점을 극복할 수 있다.



<그림 25> 학습자 말뭉치에 기반한 자동 첨삭 시스템 프로세스 도식

- 위 그림에서와 같이 학습자의 메타 정보를 수집한 뒤 한국어 구사 수준, 모어, 학습 목적 등을 파악해 해당 학습자에게 맞춤형 과제를 제공하고, 학습자가 과제를 제출하면 도구가 이를 평가하고 첨삭 정보를 자동으로 생성해 피드백을 제공한다. 이러한 도구 개발을 위해서는 한국어 학습자 말뭉치 DB 축적이 우선되어야 하며, 오류 통계 DB, 맞춤 예제 DB 등을 구축하여야 한다.

- 예를 들어, 학습자 말뭉치 오류 통계에서 중국어권 1급 학습자에게 부사격 조사 ‘에’와 ‘에서’ 혼동 오류가 가장 많이 나타났다면, 학습자 정보가 ‘중국어권, 1급’인 학습자에게는 ‘에’와 ‘에서’를 연습하는 예제를 많이 제공하는 것이다.
- 최종 목표는 학습자가 생산한 텍스트의 자동 첨삭 시스템이지만 이런 시스템을 개발하기 위해서는 정교하게 오류 주석이 부착된 대규모의 학습자 말뭉치가 필요하고 오류의 양이 많을 때에 시스템이 필자의 객관적 지식과 주관적 의도를 모두 파악하여 교정하는 것은 쉽지 않다.
- 자동 첨삭 시스템 개발의 방법론은 크게 규칙에 기반한 것과 기계 학습 등 통계에 기반한 것으로 나눌 수 있다. 완전 자동 첨삭의 전단계로는 학습자에게 교정 형태의 후보 중 하나를 고르게 하는 반자동 시스템을, 완전 자동 첨삭에서 한 단계 더 나아간 시스템으로는 쓰기, 말하기 등 언어 표현 기능을 자동적으로 평가하는 시스템을 상정할 수 있다.

Ⅲ. 학습자 말뭉치 구축 관련 교육 및 홍보

1. 말뭉치 구축/가공 인력 실무 교육

- 말뭉치 구축/가공 인력 실무 교육은 실무 작업자에게 말뭉치 구축에 관한 기본 소양과 기술을 익히도록 하여 전문성을 제고하여 체계적인 말뭉치 구축을 위한 것이다. 또한 다양한 변이형을 포함한 학습자 말뭉치의 특성상 작업자의 직관에 의해 달라질 수밖에 없는 세부적인 자료 처리 방식을 최대한 일관성 있게 맞추기 위한 것으로 온라인/오프라인, 정기/비정기 워크숍을 통해 실무자 간에 원활한 소통이 이루어지도록 하였다.
- 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 제고하도록 하였다.

<표 14> 말뭉치 구축/가공 인력 교육의 주요 내용

	지침 교육	도구 교육 및 실습
기본 교육	<ul style="list-style-type: none"> ○ 한국어 학습자 말뭉치 구축 사업의 개요 및 절차 ○ 국내외 학습자 말뭉치 구축 현황 및 활용 	<ul style="list-style-type: none"> ○ 온라인 구축 도구 전반
수집	<ul style="list-style-type: none"> ○ 구어와 문어 자료 수집을 위한 과제 유형 ○ 종적 자료 수집 과제 유형 및 자료의 처리 ○ 학습자 동의서 수집 및 처리 ○ 수집 자료의 처리와 관리 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 수집 자료 업로드 ☞ 1차 연도 수집은 구축 본부와 수집 기관의 직접 접촉 방식을 취하였으나 향후 온라인 구축 도구를 통한 수집 기관의 직접 업로드 방식 병행 가능
자료 처리	<ul style="list-style-type: none"> ○ 자료의 분류 ○ 스캔 및 음성 파일 변환 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 스캔/음성 원본 파일 업로드 및 파일

및 파일 등록	○ 파일명 부여 체계 ○ 학습자 정보 및 파일 정보 등록(헤더 마크업)	등록, 파일명 생성 ○ 스캐너 사용 ○ 음성 파일 변환 도구
입력	○ 문어 입력 및 검수 방법, 쟁점	○ 온라인 구축 도구: 파일 입력 및 마크업, 할당 받은 작업 파일 관리 및 작업 ○ 온라인 구축 도구 외 입력용 텍스트 에디터
전사	○ 구어 전사 및 검수 방법, 쟁점	○ 온라인 구축 도구: 전사 파일 업로드 및 수정, 마크업, 할당 받은 작업 파일 관리 및 작업 ○ 전사 도구 엘란(ELAN)
형태 주석	○ 형태 분석 방법 및 절차 ○ 형태 주석 체계 틀의 구성과 내용 ○ 형태 분석 자료 검수 및 검수 관련 쟁점	○ 온라인 구축 도구: 형태 주석, 할당 받은 작업 파일 관리 및 작업
오류 주석	○ 오류 식별, 판정 및 교정의 기준 ○ 오류 주석 체계 틀의 구성과 내용 ○ 오류 분석 자료 검수 및 검수 관련 쟁점	○ 온라인 구축 도구: 오류 주석, 할당 받은 작업 파일 관리 및 작업

1.1. 문어 입력 인력 실무 교육

- 문어 입력 작업은 스캔 원본의 모든 정보가 표본 등록, 입력과 검수의 단계에서 정확하게 시스템에 옮겨져야 한다. 각 단계별 체크해야 하는 항목과 입력하는 텍스트에는 오타나 실수가 없어야 한다. 이를 위해 각 단계별로 작업자들의 숙달도를 높이고 실수를 줄이기 위한 교육과 자료 점검에서 주요 항목들을 관리해야 한다.

1) 표본 등록 단계

- 체크 항목과 직접 입력하는 항목들이 누락되지 않고 체계적으로 정확하게 입력되도록 교육 및 확인한다.
 - 시스템 체크 항목 : 학습 환경, 수집 기관, 학습자 유형 자료 유형, 작문 유형, 장르, 성별, 현재 등급(TOPIK 등급), 국적(교포 여부), 제1 언어, 한국어 학습 목적, 직업, 기타 언어 등
 - 직접 입력 항목: 수집 시기, 주제, 나이, 한국어 학습 기간, 한국어 거주 기간 등
- 작업자는 '표본 등록' 저장을 누른 후 입력 할당을 요청하기 전에 '작업 목록 보기' 메뉴에서 다시 한 번 자체적으로 입력 정보를 재확인한다.

2) 입력 단계

- 스캔 원본에 학습자가 입력한 텍스트를 정확하게 오타 없이 시스템에 입력해야 한다.
 - 학습자 자료와 교사 첨삭 구분 : 학습자가 입력한 텍스트와 교사가 수정, 첨삭한 내용을 혼동하지 말고 구분하여 학습자 자료만 입력한다.
- 본문 입력 후 자동 주석을 생성한 후 마크업 기능을 숙지해서 적용하도록 한다.
 - '문장' 마크업 : 마침표가 누락된 문장은 작업자가 판단하여 '문장' 마크업을 적용한다.
 - '기호류' 마크업 : 기호나 식별이 어려운 문자를 본문에 입력한 후에 마크업이 누락되지 않도록 주의한다.

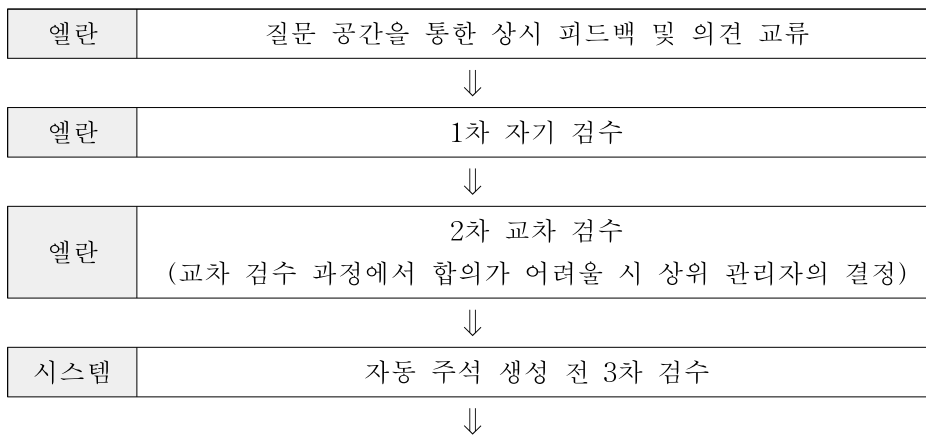
3) 검수 단계

- 표본 정보, 본문 텍스트와 함께 정확한 마크업 적용 여부를 꼼꼼하게 점검한다.
 - 표본 서지 정보 재확인 : 누락된 정보가 없는지 확인한다.

- 본문 입력 내용 : 스캔 원본과 본문 텍스트를 대조하여 오타나 누락된 부분을 확인하고 수정한다.
- 개인 정보 노출 확인 : 학습자 개인 정보가 공개되지 않도록 마크업 여부를 확인한다.
- 본문 주석 : 누락된 주석이나 잘못 적용된 경우를 발견했을 때, 부분적으로 마크업 수정을 한다.

1.2. 구어 전사 인력 실무 교육

- 구어 전사에 있어 가장 문제가 되는 것은 바로 작업자 간 전사 일관성일 것이다. 본 작업에서는 작업자 간 적절한 일관성을 유지하기 위하여 엘란 전사 단계에서 교차 검수를 진행하였다. 교차적으로 검수를 진행하며 합의가 되지 않은 상황에 대해서는 상위 관리자가 판단하여 결정하였다.
- 엘란에서의 검수가 끝난 후 시스템에 업로드를 하며 또 한 차례의 검수가 진행되게 되었다. 이는 엘란에서의 작업이 형태나 오류 작업에 알맞게 전사가 되었는지를 최종적으로 확인하는 작업이다. 만약 엘란에서의 기호가 지침에 조금이라도 맞지 않으면 자동적으로 불러들이는 작업에서 오류가 나기 때문에 자동 주석을 생성하기 이전에 이를 꼼꼼하게 재검토하는 작업이 이루어졌다.



시스템	작업자의 아이디를 교환하여 시스템상에서 4차 검수 합의가 이루어진 최종 파일에 한하여 작업 완료 버튼 누름. (교차 검수 과정에서 합의가 어려울 시 상위 관리자의 결정)
-----	---

<그림 26> 구어 전사 인력 교육 절차 및 내용

1.3. 형태 주석 인력 실무 교육



<그림 27> 인터넷 카페를 활용한 작업자 질의응답 예시

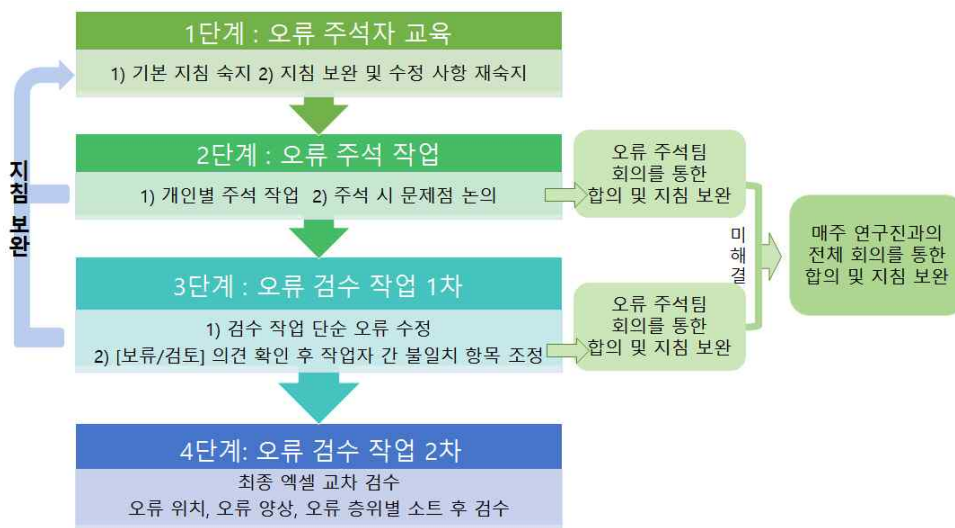
- 작업자들 사이의 불일치를 줄이고, 작업 결과물의 질을 제고하기 위하여 작업자의 선발 및 교육에서부터 실제 작업 과정에서의 검수 절차에 이르기까지 다양한 방식으로 교육 및 관리 작업이 이루어졌다.
- 작업자 선발 과정에서 형태 주석 지침에 대한 교육 및 지침 이해도에 대한 시험을 실시하여 작업자 별 수준을 파악하였다.
- 작업 진행 과정에서 국내 포털 사이트의 카페 서비스를 이용하여 형태 분석 작업 과정에서 발생하는 의문 및 문제점을 상위 관리자 및 작업자 모두와 공유가 가능하도록 하였다. 또한, 형태 주석 작업 과정에서 발생하는 시스템 상의 문제에 대해서도 해당 카페를 통해 제보하도록 하여 빠른 해결이 가능하도록 하였다. 또한, 게시판 운영을 통해 누적된 작업자들의 질

문 사항은 수집 및 DB화 하여 월 2회의 정기 회의에서 작업자들에게 고지하였다.

- 월 2회의 정기 회의를 통하여 작업상의 문제점 및 지침의 변경 사항 등에 대해 공지하였으며, 온라인상에서의 논의로는 해결되지 않는 부분에 대해 토의하고 이를 작업에 반영하도록 하였다.
- 검수자들은 각 작업자의 작업 결과물에 대한 검수 결과를 토대로, 개별 작업자와의 수시 면담을 통해 작업 과정상의 어려움을 파악하고 해결하여 작업물의 질을 제고하도록 하였다.

1.4. 오류 주석 입력 실무 교육

- 오류 주석 작업에 있어서 가장 중요한 것은 오류 주석자 내, 오류 주석자 간의 일관성을 유지하는 것이다. 일관성 있는 오류 주석 작업은 곧 오류 주석의 신뢰도를 높일 수 있다. 이에 오류 주석자들은 다음과 같은 절차를 통해 오류 주석의 일관성 및 신뢰도를 높이기 위해 노력하였다.



<그림 28> 오류 주석 작업 단계별 교육 절차 및 내용

- 1단계는 사전 교육 단계로 기본 오류 주석 지침을 숙지한다. 숙지한 내용

을 바탕으로 오류 주석 작업을 실시한다. 그러나 실제 오류 주석 작업 및 검수 단계에서 나타나는 문제점들이 있기 때문에 매주 오류 주석팀 회의 및 연구진 전체 회의를 통하여 논의하고, 합의한 내용을 바탕으로 계속해서 지침이 보완되고 수정되었다.

- 2단계 개인별 오류 주석시 나타나는 문제점, 처리하기 어려운 부분들은 1차적으로 오류 주석 팀 내 회의를 통해서 논의하였다. 오류 주석 팀 내 회의를 통해서 해결된 사항은 지침에 추가하고, 미해결된 사항에 대해서는 연구진 전체 회의에서 논의하여 해결하였다. 또한 시스템 상에서 오류 주석을 하면서 [보류/검토] 기능을 활용하여, 1차 검수자에게 의견을 남겨두는 방법을 활용하였다.
- 3단계 1차 오류 주석 결과를 검수하였다. 검수 과정에서 잘못된 오류 주석을 수정하고, 주석자가 남겨놓은 [보류/검토] 의견을 확인하여 검수자 선에서 해결할 수 있는 사항을 해결하였다. 그러나 주석자와 검수자 의견이 불일치할 경우 오류 주석팀 내 회의를 통하여 2인 이상 합의 판정하여 결정하였다. 오류 주석 작업 단계와 마찬가지로 오류 주석 팀 내 회의에서 해결하지 못한 사항에 대해서는 연구진 전체 회의에서 논의하여 결정하였다.
- 마지막으로 오류 주석 및 검수가 완료된 파일을 엑셀 파일로 일괄 다운받아 주석자 간 교차 검토를 실시하였다. 이때에는 오류 양상을 중심으로 소트하여 오류 위치, 오류 양상, 오류 층위를 재검수하여 잘못된 주석은 다시 시스템상에서 수정하여 최종 오류 주석 작업을 완료하도록 하였다.
- 이상의 오류 주석 절차를 거쳐 오류 주석 작업이 수행되었다. 오류 주석팀은 매주 모여 공동 오류 주석 작업 및 검수 작업을 실시하면서 논의 사항에 대해서 회의를 진행하였고, 이를 지침에 반영함으로써 오류 주석자간의 일치도와 신뢰도를 높이고자 하였다.

2. 한국어 학습자 말뭉치 홍보

2.1. 한국어 학습자 말뭉치 아카데미 개최

- 한국어 학습자 말뭉치 아카데미는 학습자 말뭉치의 확산을 위한 사전 홍보와 사용자들의 활용 능력 제고를 목표로 시행되었다. 한국어 학습자 말뭉치 구축 단계별 쟁점과 실제, 활용에 관한 내용으로 프로그램을 기획하여 학습자 말뭉치에 대한 이해 위에 실제적인 활용이 가능하도록 하였다.

1) 제1차 한국어 학습자 말뭉치 아카데미

- 일시: 2016년 8월 17일(수) 14:00~17:00
- 장소: 계명대학교 성서 캠퍼스 영암관(인문국제대학) 358호
- 프로그램

<표 15> 제1차 한국어 학습자 말뭉치 아카데미 프로그램

시간	내용	발표
1:30~2:00	등록	
2:00~2:10	개회사	강현화(책임연구원)
2:10~2:20	인사말	황용주(국립국어원)
2:20~2:40	학습자 말뭉치 구축 사업 소개	강현화(연구책임자)
2:40~3:00	학습자 말뭉치 구축의 실제: 문어 자료 입력	윤종원(실무연구원)
3:00~3:20	학습자 말뭉치 구축의 실제: 구어 자료 전사	공나형(실무연구원)
3:20~3:40	학습자 말뭉치 구축의 실제: 형태 주석	허희정(실무연구원)
3:40~3:50	휴식	
3:50~4:10	학습자 말뭉치 구축의 실제: 오류 주석	유소영(실무연구원)
4:10~4:30	학습자 말뭉치 구축 시스템의 활용: 구축에서 검색까지	곽용진 ((주) 이르테크)

4:30~4:50	종합 토론: 참석자와의 자유 토론 및 질의응답	김선정(공동연구원)
4:50~5:00	폐회사	강현화(책임연구원)

2) 제2차 한국어 학습자 말뭉치 아카데미

- 일시: 2016년 10월 15일(토) 13:40-17:00
- 장소: 한국외국어대학교 대학원 1층 전산실습실
- 프로그램(2부로 나누어 동일한 프로그램을 2회 진행)

<표 16> 제2차 한국어 학습자 말뭉치 아카데미 프로그램

진행 순서	내용	발표
1	학습자 말뭉치 구어 전사	공나형(실무연구원)
2	학습자 말뭉치 오류 주석	유소영(실무연구원) 한송화(공동연구원)
3	학습자 말뭉치 검색	곽용진 ((주) 이르테크) 김한샘(공동연구원)

2.2. 학술대회 발표

- 학술대회 발표는 학습자 말뭉치 구축 방법론 및 쟁점에 관한 학술적 교류와 향후 학습자 말뭉치 구축 결과물을 널리 확산시키기 위한 홍보를 목표로 이루어졌다. 학습자 말뭉치 구축 사업을 소개하고 실제 구축 과정에 활용에 관한 다양한 쟁점들을 발표하여 그간 많은 연구가 이루어지지 못한 학습자 말뭉치 구축에 관한 논의를 불러일으키고, 구축 방향에 대한 예비 사용자들의 의견을 수렴하여 보다 실용적이고 활용도가 높은 말뭉치를 구축하고자 하였다.

<표 17> 학술대회 발표

일자	발표 내용	학회명
2016. 4. 16(토)	<기획 워크숍> 한국어 학습자 말뭉치의 주석 과정과 활용 방법 (김일환)	국제한국어교육 학회
2016. 5. 7(토)	학습자 말뭉치에서의 구어 전사와 오류 주석의 쟁점과 실제(한송화, 강현화)	한국언어문화교 육학회
	차세대 학습자 말뭉치 통합 관리 시스템 개발(김한샘, 박용진)	
	학습자 말뭉치 구축과 음성 인식 활용(이화진, 이지연)	

IV. 학습자 말뭉치 수집 및 가공

1. 2015년 기구축 학습자 말뭉치의 보완

- 2015년에는 학습자 말뭉치 구축을 위한 중장기 계획의 수립과 함께 국내 교육기관 자료의 시험 구축이 다음과 같이 이루어졌다²⁾.

<표 18> 2015년 기구축 학습자 말뭉치의 규모

구분		1급	2급	3급	4급	5급	6급	계
원시	문어	49,359 (697)	50,208 (451)	50,704 (422)	49,977 (385)	50,195 (345)	50,554 (316)	300,997 (2,616)
	구어	7,498 (30)	10,679 (19)	8,441 (17)	9,123 (14)	10,162 (13)	7,991 (7)	53,894 (100)
형태 주석	문어	31,664 (456)	30,423 (269)	35,947 (311)	37,425 (292)	34,154 (235)	31,381 (203)	200,994 (1,766)
	구어	3,036 (16)	3,687 (5)	2,032 (5)	3,995 (7)	3,511 (5)	4,985 (4)	21,246 (42)
오류 주석	문어	6,804 (96)	7,652 (64)	6,604 (48)	7,141 (51)	7,400 (48)	7,285 (45)	42,886 (352)
	구어	1,674 (9)	1,890 (3)	2,032 (5)	1,917 (4)	1,978 (3)	2,286 (2)	11,777 (26)

- 2015년에는 문어 입력과 구어 전사, 형태 주석, 오류 주석 지침의 개발과 함께 학습자 말뭉치 구축 도구 개발 작업이 동시에 이루어졌다. 이에 따라 2015년 학습자 말뭉치는 문어는 텍스트 에디터, 구어는 엘란, 형태 주석은 리눅스 기반의 KMAT, 오류 주석은 엑셀을 사용하여 수행되었으며 그 결과물이 최종 단계에서 학습자 말뭉치 구축 도구에 탑재되었다.
- 2016년에는 2015년의 구축 경험과 구축 결과물의 재검토를 통해 지침을 정비하고 이를 2015년 기구축 말뭉치에 반영하여 2016년에 새롭게 구축되는 학습자 말뭉치와의 통일성을 갖추고 질적 완성도를 제고하고자 하

2) 2015년 기구축 학습자 말뭉치는 <2015년 한국어 학습자 말뭉치 기초 연구 및 구축> 최종보고서에 제시된 규모와 약간의 오차가 있다. 이는 2015년 구축한 자료를 한국어 학습자 말뭉치 구축 도구에 업로드하고 질적 보완을 위한 검수 과정에서 발생한 것으로 목표 규모를 상회함에는 변함이 없다.

였다. 각 말뭉치 유형별 보완 사항은 다음과 같다.

- 원시 말뭉치(문어): 구축 도구의 체제에 맞게 본문 주석 수정
- 원시 말뭉치(구어): 수정 지침에 따른 전사 자료의 수정과 검수, 학습자 말뭉치 구축 시스템에의 탑재 및 구축 도구의 체제에 맞게 본문 주석 수정
- 형태 주석 말뭉치: 세종 21세기 한국어 균형 말뭉치와의 호환성, 학습자 자료의 특성을 반영한 지침의 보완과 그에 따른 주석 수정
- 오류 주석 말뭉치: 수정 지침에 따른 오류 주석 수정

2. 2016년 학습자 말뭉치 수집 및 가공

2.1. 수집

1) 집중 수집: 국내 교육기관 자료

(1) 수집 대상

- 국내 한국어 교육 기관 정규 교육과정 수강 학습자 및 학문 목적 학습자

(2) 수집 범위

- 한국어 교육 기관의 정규 교육과정 소속 학습자를 대상으로 한 자료의 수집 범위는 다음과 같다.

<표 19> 수집 대상별 자료 수집 범위: 국내 교육기관 학습자

구분	수집 범위 및 자료 유형
교육기관 말뭉치	정규/비정규 교육과정의 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

(3) 국내 교육기관 말뭉치 수집 네트워크

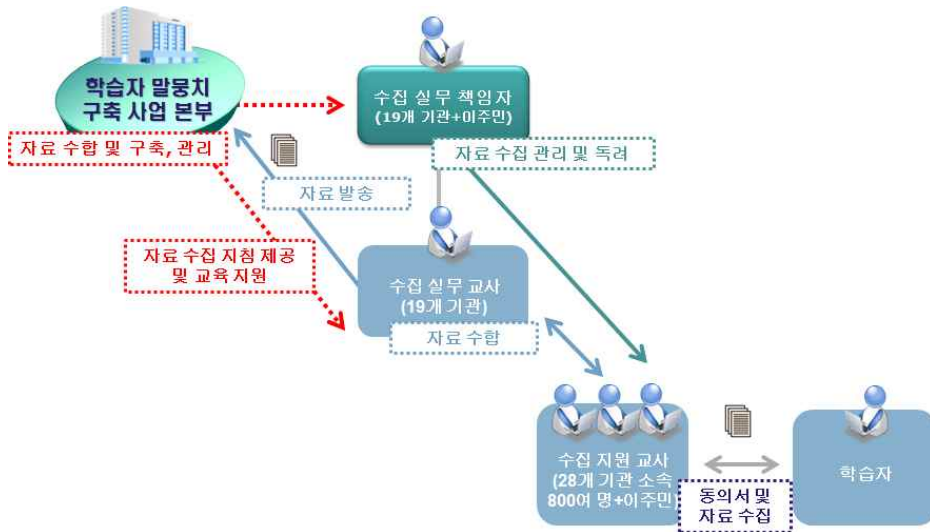
- 2015에는 지역 배분을 고려하여 전국의 한국어 교육기관 중 28개 기관을 대상으로 자료를 수집하였으나, 수집 기관의 자율적인 의사를 반영하여 수집 참여 철회 기관을 제외하고 교육과정상의 등급 체계가 상이한 기관을 제외하여 다음의 18개 기관으로 집약하여 자료를 수집하였다.

<표 20> 국내 교육 기관의 실무 책임자 및 실무 교사 명단

지역	수집 기관	책임자	실무 교사
서울 (동부)	고려대학교 한국어교육센터	김정숙	심재경
	경희대학교 국제교육원	이정희	이지영 윤혜리
	국민대학교 한국어교육센터	이동은	노미연
서울 (서부)	연세대학교 한국어학당	김미옥	단현주
	서울대학교 언어교육원	안경화	김폴잎
	서강대학교 한국어교육원	김현정	김현정
	이화여자대학교 언어교육원	김현진	박진철
	홍익대학교 국제언어교육원	이화진	홍연정
경기	경희대학교(국제) 언어교육원	김유미	이지민
	강남대학교 한국어교육원	서정숙	서정숙
충청	배재대학교 한국어교육원	박석준	김민우
	선문대학교 한국어교육원	라혜민	유순천
호남	호원대학교 한국어교육원	진대연	김은영
대구·경북	계명대학교 국제교육센터	김선정	홍종호
	경북대학교 한국어문화원	서효경	안민지
	동국대학교(경주) 국제교류교육원	배현숙	오누리
부산·경남	부산외국어대학교 한국어문화교육원	정명숙	오상민
	신라대학교 한국어교육센터	조정순	구다혜

(4) 자료 수집 및 구축 체계

- 국내 교육기관 학습자의 자료 수집은 2015년도와 동일하게 구축 본부와 수집 실무자와의 직접 접촉을 통해 진행하였다. 자료의 유형, 수집량에 따라 우편 및 택배, 전자우편 등을 이용하여 수집 자료를 주고받았다.



<그림 29> 자료 수집 체계: 국내 교육기관

(5) 수집 결과

① 일반 자료

- 국내 교육기관 학습자의 자료는 문어 19,938개, 구어 1,476개의 파일을 수집하였다. 문어의 경우 파일 하나당 평균 어절 수를 100어절로 추산하였을 때 약 1,993,800어절 규모의 자료이며 구어의 경우 파일 하나당 평균 어절 수를 350어절로 추산하였을 때 516,600어절로 문어의 경우 수집 목표량을 300%, 구어의 경우는 500% 이상을 초과 달성하였다.³⁾

☞ 위의 수집 규모는 현재까지 수집된 자료의 단순 누적량이며, 실제 구축

3) 자료 분류 과정에서 수집 대상에서 제외되는 자료가 있을 수 있으므로 실제 구축 규모는 달라질 수 있음

대상이 될 자료를 선별하는 질적 검사 과정에서 구축 불가로 판정되는 자료들이 생김으로 해서 변동이 생기게 된다. 2015년도 구축 경험을 기준으로 하였을 때 분량 미달, 미완성 작문 또는 발화, 복사 자료나 음질 불량 등의 이유로 구축 불가 판정되는 자료는 전체 자료의 약 30% 이상으로 질적 검사와 함께 지속적인 수집이 필요하다.

가. 숙달도별

○ 숙달도별 자료 수집 현황은 다음과 같다.

<표 21> 숙달도별 자료 수집 현황

문어		구어	
수준	자료 수(개)	수준	자료 수(개)
1급	4,688	1급	296
2급	4,646	2급	386
3급	4,921	3급	416
4급	3,101	4급	256
5급	1,092	5급	107
6급	471	6급	15
6급 이상, 기타	1,019	6급 이상	28
계	19,938	계	1,504

나. 국적별

○ 수집된 자료의 국적별 분포는 다음과 같다.

ㄱ. 문어 자료

<표 22> 국적별 자료 수집 현황

국적	자료 수(개)	국적	자료 수(개)
중국	11,846	마카오	29
베트남	1,734	캄보디아	28
일본	1,532	콩고	25

홍콩	619	호주	24
대만	503	가나	23
미국	317	이란	23
태국	252	페루	23
말레이시아	251	라오스	22
러시아	194	케냐	22
스웨덴	176	키르기스스탄	21
우즈베키스탄	156	모로코	20
인도네시아	122	방글라데시	20
한국	117	나이지리아	19
카자흐스탄	111	벨기에	19
싱가포르	98	네덜란드	18
몽골	92	사우디아라비아	18
프랑스	92	파키스탄	18
필리핀	73	스위스	17
스리랑카	68	네팔	16
브라질	62	우간다	14
영국	61	콜롬비아	14
인도	59	칠레	13
이탈리아	57	베네수엘라	12
독일	54	우크라이나	12
캐나다	47	르완다	11
아랍어	45	에티오피아	11
노르웨이	40	파나마	11
스페인	38	가봉	10
미얀마	36	에콰도르	10
멕시코	35	코트디부아르	10
터키	34	기타	484

ㄴ. 구어 자료

<표 23> 국적별 자료 수집 현황

국적	자료 수(개)	국적	자료 수(개)
중국	912	독일	4
베트남	211	말레이시아	4
일본	156	몽골	4
대만	49	우즈벡	4
러시아	14	키르기즈스탄	4
영국	9	태국	3
미국	7	터키	3
캄보디아	7	한국	3
인도네시아	6	기타	70
프랑스	6		

나. 종적 자료

- 종적 자료의 경우는 1차 연도에 55명으로 수집을 시작하였으나 유급, 진학, 귀국, 중도 철회 등의 사유로 점점 수집 대상 학습자의 수가 감소하여 50주차까지의 자료가 수집된 학습자 수가 6명이었으며, 60주차까지 수집을 완결된 학습자는 4명에 그쳤다.

2) 시험 수집: 국내 이주민 자료

(1) 수집 대상

- 결혼이민자, 중도입국청소년, 이주노동자

(2) 수집 범위

- 이주민 자료의 수집 범위는 다음과 같다. 이주민의 경우에는 국내 교육기

관 소속 학습자에 비해 상대적으로 자연스러운 습득의 환경에 있으므로 가능하면 교육과정에 종속되지 않은 구어에 비중을 두어 수집하는 것을 목표로 하였다.

<표 24> 수집 대상별 자료 수집 범위

구분	수집 범위 및 자료 유형
교육기관 말뭉치	정규/비정규 교육과정의 중간 및 기말 쓰기/말하기 시험 자료
	교과 과정에 포함된 다양한 주제 및 다양한 텍스트 형식의 수업/과제 작문
기획 말뭉치	교사와 학습자 상호 작용 연구를 위한 교실/수업 자료 말뭉치(*협조 필요)
자연 발화 말뭉치	(정기적/비정기적 수집)일상대화
	(정기적/비정기적 수집) 계획되지 않은 인터뷰
	(정기적/비정기적 수집)내레이션

(3) 수집 네트워크

- 시험 수집을 위한 수집 네트워크는 수집 대상자의 특성을 고려하여 다문화센터, 사회통합프로그램운영 기관 등을 통한 대규모 자료 수집과 개별 교사를 인력풀로 한 수집의 두 가지 방법을 병행하였다. 전자의 경우 이주민을 대상으로 한 교육과정에 속한 학습자의 자료를 대규모로 수집하기 위한 것이며, 후자는 교육기관에 종속되지 않은 상태에서 한국어를 습득해 가는 학습자의 자료를 습득하기 위한 것이다. 수집 네트워크는 다음의 과정을 통해 진행되었다.



<표 30> 이주민 자료 수집을 위한 네트워크 마련 절차

- 위의 과정을 거쳐 다문화가정지원센터 3개소, 중도입국청소년 교육 기관 2개소, 결혼이민자와 이주노동자를 가르치는 교사 2인을 시험 수집을 위한 네트워크로 확정하였다. 그 외에도 소량이지만 자원봉사 등으로 결혼이민자와 이주노동자의 자료 수집을 함께해 준 교사가 다수 있었다.

<표 25> 시험 수집을 위한 이주민 자료 수집 네트워크

구분	수집 경로	수집 대상	수집 대상자 수	비고
교육 기관을 통한 수집	대구 성주 다문화센터	결혼이민자	4명	
	대구 달서구 다문화센터	결혼이민자	74명	
	아산시 다문화센터	결혼이민자	45명	
	인천 한누리학교	중도입국 청소년(중고)	25명	
	서울 대동초등학교	중도입국 청소년(초등)	63명	
개별 교사 인력풀을 통한 수집	김성경	결혼이민자	80명	대구 지역 4개 주민센터

	이상훈	이주노동자	28명	대구 지역 공단
	기타	결혼이민자 이주노동자	6명	서울 지역

(4) 수집 시 고려 사항

① 수집 과제

- 각 기관의 교육과정에서 산출되는 자료를 중심으로 수집하되, 필요 시 국내 교육기관 자료와의 균질성을 고려하여 기획 수집 과제를 활용하여 수집하도록 하였다. 문어 학습자의 필요성이 상대적으로 적고 작문 연습을 기피하는 학습자들의 특성상 문어 자료보다 구어 자료의 수집량이 상대적으로 많았다. 구어 자료의 경우 가능하면 내러티브 형식의 자료를 수집할 수 있도록 동일한 주제를 사용하더라도 과제 수행 방식을 변형하도록 하였다.

② 학습자 동의서 수정 및 추가 항목

- 이주민의 특성을 고려하여 학습자 동의서의 학습자 정보 수집 문항을 다음과 같이 수정하였다.

<표 26> 학습자 정보 수집 항목

일반 학습자 동의서	이주민 동의서	비고
연령	생년월	수정
거주 기간	입국년월	수정
-	주로 사용하는 언어	신규 추가
-	한국어로 대화하는 상대	신규 추가
-	한국어로 말하고 듣는 시간	신규 추가

③ 한국어 수준(등급) 측정을 대체한 한국어 적성 평가

- 이주민의 경우 표준교육과정에 따르는 다문화센터나 사회통합프로그램 운영 기관을 제외하고 학습 수준을 측정하기 어려우며 측정하여도 균질하지 않다. 아울러 대부분의 교육기관에서 중급 이후의 교육과정을 운영하지 않는 경우가 많으며 교육과정에 속하지 않은 채 자연 습득 과정에 있는 학습자의 경우 수준 측정이 불가능하다. 이에 따라 대안으로 MLAT-기반 한국어 적성 평가지⁴⁾를 마련하여 시험 시행하였다.

가. 문항 구성

- 문항 구성은 다음과 같이 총 46개로 구성되었다(☞ 평가지는 부록 참고).

<표 27> MLAT-기반 한국어 적성 평가 문항 구성

영역	문항 유형	세부 문항	문항 수
Part I	Number Learning		
	수의 조합에 대한 청각 테스트	숫자를 듣고 10초 안에 쓰기	10
Part II	Phonetic Script 소리나 음성기호의 대응	16개의 단어를 단어를 듣고 앞에서 들은 단어면 O 아니면 X 표 하기	8
		듣고 맞는 단어를 고르기	4
Part III	Spelling Clues	제공되는 자음과 모음을 사용하여 단어 완성하기	4
Part IV	Words in Sentences	알맞은 단어와 문법을	8

4) MLAT는 Modern Language Aptitude Test의 약자로 Carroll & Sapon(1959)에 의해 개발된 언어 적성 평가지로 학습자들의 외국어 학습을 얼마나 성공적이고 쉽게 할 수 있는지 그 가능성을 평가하기 위해 개발된 것으로 적성 평가 문항에서의 구인은 다음과 같다.

- ① 음성적 부호화(phonetic coding): 청각적 부문으로 소리를 분절하고 구분하는 능력 =(LLAMA의 음성 분석 능력과 유사)
- ② 문법적 민감성(grammatical sensitivity): 문장에서 어휘나 언어 구조의 문법적 기능을 인식하는 능력 =(LLAMA의 문법적 민감도와 유사)
- ③ 암기식 학습 능력(rota learning ability): 일반적 기억의 일종이나 외국어 학습 상황에서는 개인 편차를 보인다 (Carroll, 1990). =(LLAMA의 단순 반복 학습과 유사)
- ④ 연역적 언어 학습 능력(inductive language learning ability): 언어 사용을 주관하는 규칙의 적용 능력으로 역시 일반적 상황과 외국어 학습 상황은 사뭇 다르다.

	문법 구조에 대한 지식	이용해서 문장 완성하기	
Part V	Paired Associates 주어진 언어의 어휘 목록을 익혀 의미 암기	2분 동안 제시된 단어를 외우고 다음 2분 동안 단어 의미를 모국어로 쓰기	12
계			46

나. 시험 시행 대상

○ 다문화센터 3단계 수준의 결혼이주여성 20명

☞ 시험 수집 대상 선정에 앞서 초급, 중급, 고급의 결혼이주여성을 대상으로 평가지를 풀어 보게 한 결과 학습 입문 단계에 있는 초급 학습자의 경우는 일부 문항을 제외하고 풀기가 어려웠고, 고급 단계의 결혼이주여성에게는 매우 쉬워 시험의 의미가 없다고 판단되었다. 이에 따라 학습 과도기에 있으며 여러 가지 환경 요건에 따라 다양한 언어 발달 양상을 보이는 3급 수준의 이주여성 20명을 시험 시행 대상으로 선정하였다.

다. 시험 시행 결과

○ 시행 결과 20명 중 8명은 전체 문항에 대하여 정답을 제시하였고 12명이 1-2개의 오답을 제시하였다. 12명의 학습자가 오답을 보인 문항은 다음과 같다.

<표 28> MLAT-기반 한국어 적성 평가 오답자 수

영역	문항 유형	세부 문항	오답자 수
Part I	Number Learning 수의 조합에 대한 청각 테스트	숫자를 듣고 10초 안에 쓰기	1
Part II	Phonetic Script 소리나 음성기호의 대응	16개의 단어를 듣고 앞에서 들은 단어면 O 아니면 X 표 하기	5
		듣고 맞는 단어를 고르기	1
Part III	Spelling Clues	제공되는 자음과 모음을 사용하여 단어 완성하기	2

Part IV	Words in Sentences 문법 구조에 대한 지식	알맞은 단어와 문법을 이용해서 문장 완성하기	4
---------	------------------------------------	-----------------------------	---

- Part I에서는 네 자리 숫자 듣고 적기에서 1명의 오답자가 있었다.
 - Part II에서는 5명의 오답자가 있었는데 대부분이 단어 ‘노력’을 듣고 적는 데에서 오답이 발생하였다. 나머지 1명은 변이음 ‘탁/닥/딱’을 변별하여 듣는 문제에서 오답이 발생하였다.
 - Part III에서 자음과 모음을 조합하여 ‘청소’를 찾은 후 옮겨 적는 과정에서 발음 습관이 반영되어 ‘정소’로 표기된 것으로 사실상 해당 문항의 평가 문항에서 벗어난 오답 유형이다.
 - Part IV에서는 동사 ‘부르다’의 활용형을 ‘불려요’, ‘부러요’, ‘불어요’로 적어서 오답이 된 경우와 ‘날씨가 추운’의 ‘추운’을 옮겨 적는 과정에서 발음 습관에 의해 나타난 것으로 파악되는 표기 오류에 의한 오답이 있었다.
- 이상의 결과를 종합해 볼 때 시험 시행용으로 만든 MLAT-기반 한국어 적성 평가지는 학습자의 수준을 등급화하기 위한 변별 정보를 제공하지 못하였다. 그럼에도 중간보고회의에서 이를 수집 대상자별로 세분화하여 보다 정교하게 만들어야 할 필요성이 제기되었으며, 시험 수집 과정에서 다양한 등급 체계 혹은 학년 외에 학습자의 수준을 판별할 수 있는 등급 체계가 존재하지 않는 경우가 있어 2017년 본격 수집을 앞두고 이에 대한 심도 깊은 논의와 대안 마련이 필요할 것으로 판단된다. 그 외에도 자료 수집 외에 이와 같은 평가 도구를 현장에서 시행할 수 있는지 등에 대한 사전 조사와 협의가 필요할 것이다.

(5) 수집 결과

- 이주민 자료는 구어 213개 파일(약 10만 어절⁵⁾), 문어 622개(약 6만 2천 어절)을 수집하였으며 자료의 숙달도별, 국적별 분포는 다음과 같다.

가. 숙달도별

- 5) 이주민 자료의 경우 발화 길이가 일반 학습자 자료에 비해 매우 긴 편임. 이에 따라 파일 1개의 평균 어절을 500어절로 추산했을 때 현재 수집한 자료는 10만 어절을 상회함

- 이주민 자료의 숙달도별 분포는 다음과 같다. 자료에서 보는 바와 같이 학습자 대상별 소속 기관에 따라 등급 기준이 상이함을 볼 수 있다.

<표 29> 이주민 자료 시험 수집 현황: 숙달도별(파일 수, 개)

결혼이주여성			이주노동자			중도입국청소년		
수준	구어	문어	수준	구어	문어	수준	구어	문어
2급	2	2	3급		1	초2	8	8
4급	2	2	4급		1	초3	11	11
5급	1	1	5급		1	초4	11	11
1단계	10	23	기초	12	10	초5	2	2
2단계	24	81	초급	8	8	초6	5	5
3단계	38	83	초급1	3	4	중1		5
4단계	4	24	초급2	3	3	중2		3
고급	8	8	토픽 2급		1	중3		10
중급	1	1	토픽 4급		2	고1		7
초급	1	1				고2		4
토픽 3급	7	14				고3		3
토픽 4급	10	11						
토픽 5급	7	9						
토픽 6급	3	3						
토픽반	5	5						
관별불가	12	23						
계	136	296	계	37	41	계	40	72

나. 국적별

- 이주민 자료의 국적별 분포는 다음과 같다. 결혼이주여성의 경우 베트남이 가장 많고 캄보디아, 중국, 일본, 파키스탄, 우즈베키스탄, 필리핀 등이 주를 이루었다. 이주노동자의 경우는 베트남, 네팔, 인도네시아가 많았으

며, 중도입국청소년의 경우 중국이 가장 많았다.

<표 30> 이주민 자료 시험 수집 현황: 국적별(파일 수, 개)

국적	결혼이주여성		이주노동자		중도입국청소년	
	구어	문어	구어	문어	구어	문어
네팔	2	5	6	6		
대만		8				
러시아	1	6				9
미국	1	4				
방글라데시	2	2	2	2		1
베트남	61	103	18	23		1
수단						1
스리랑카				2		
시리아						1
아프가니스탄	1	3				
영국	1	1				
우즈베키스탄	5	15	1	1		2
인도네시아		4	7	4		
일본	17	22				
중국	18	38			32	45
카자흐스탄						2
캄보디아	7	41	1	2		1
키르기스스탄	1	1				
태국	1	3	1	1		
파키스탄	10	22				
프랑스		1				
필리핀	7	15				1
한국	1	2			7	7
국적미상			1		1	1
계	136	296	37	41	40	72

(6) 2017년도 이주민 자료 본격 수집을 위한 계획 수립 방향

① 수집 대상의 선택 또는 대상별 자료 수집 비율의 조정

- 이번 시험 수집의 대상은 결혼이주여성, 이주노동자, 중도입국청소년의 세 집단이었다. 결혼이주여성의 경우 세 집단 중 가장 수가 많고 접촉 및 자료 수집이 용이한 반면, 이주노동자는 이주노동자 지원센터나 자원봉사를 하는 교사를 통해 수집을 해야 하는데 그 수가 많지 않고 불법 체류 등으로 인해 자료 수집에 소극적인 경우가 많았다. 한편, 중도입국청소년의 경우 초·중·고등학교 또는 이들을 위한 대안학교 등의 교육기관과 접촉해야 하는데 기관장, 책임교사, 담임교사, 학부모 등 자료 수집을 위한 허가 및 동의 절차가 매우 까다로운 편이다. 이상과 같은 자료 수집의 용이성이나 자료의 가치 등을 고려하여 선택적으로 자료를 수집하거나 수집 비율을 조정할 필요가 있다.
- <2015년 한국어 학습자 말뭉치 구축 및 연구>에 제시된 연차별 구축 계획에 따르면 이주민 자료의 경우 문어 70만 어절, 구어 15만 어절 규모의 말뭉치를 수집하여 구축하는 것으로 설계되어 있다. 실제 학습자의 수가 가장 많고 분포가 다양하여 자료의 활용도가 가장 높은 결혼이주여성의 자료의 구축 비율을 높이고 자료 수집의 어렵고 활용도가 상대적으로 낮은 이주노동자의 자료의 구축 비율을 낮추는 방안이 현실적일 것으로 보인다. 이와 같은 방식에 따라 100을 기준으로 50(결혼이주여성 자료):30(중도입국청소년):20(이주노동자)의 비율로 구축 비율을 조정할 수 있다.

<표 31> 한국어 학습자 말뭉치 연차별 구축 계획

	1단계		2단계		3단계		계
	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)
--	--	--	--	-------------------	---------	--

② 수집 자료의 유형

- 이주민의 경우 표준 교육과정이나 평가에 ‘긴 글 쓰기’가 포함되지 않으며 실생활에서의 요구도 말하기에 비해 적은 편이다. 따라서 교사가 수업 활동으로 제시하거나 과제로 부과하여도 학습자들에게 동기 부여가 되지 않아 자료의 회수율이 매우 낮고 작문의 질이 좋지 않은 경우가 많았다. 그 결과 시험 수집 과정에서 문어 자료의 수집이 쉽지 않았다. 이에 비해 구어는 수집 대상자를 정하고 실제 수집을 위한 과제의 기획이나 교사 교육에는 다소 어려움이 있었지만 수집된 자료의 질은 문어에 비해 우수하였다. 따라서 이주민 자료의 경우 기획 말뭉치의 형태로 양질의 구어 자료를 수집하는 데에 주력하는 것이 효율적일 것으로 보인다.

③ 수집 네트워크 마련 방안

- 이주민 자료 수집을 위한 네트워크는 다문화센터나 사회통합 프로그램을 운영하는 기관, 이주민 지원센터, 중도입국청소년이 있는 초·중·고등 교육 기관, 그리고 학습자를 개인적으로 접촉하여 자료를 수집하는 교사 인력 풀로 구성할 수 있다. 기관의 경우 여성가족부, 교육과학기술부, 법무부 등의 부처 간 업무 협조를 통한 기관 단위의 수집 가능성을 충분히 검토해 보아야 할 것이다. 또한 기관의 협조가 이루어진다고 하여도 학습자에게 자료 수집을 강제하지 않을 경우 실제 수집이 가능한 자료의 규모는 그다지 크지 않은 것이 현실이다. 따라서 학습자의 수가 많은 지역별 거점 기관과의 협약을 통해 중점적으로 자료를 수집하는 것이 효율적일 것이다.

④ 수집 과제 설계

- 이주민의 경우 국내 교육 기관의 어학연수생들과 달리 제2언어로서 실생활에서의 언어 사용과 사회 통합을 목적으로 한국어를 학습하게 된다. 따라서 시험 수집 과정에서 수집 교사들을 통해 한국어능력시험의 기출문

제를 기반으로 추출한 국내 교육기관용 기획 과제는 이주민들에게 다소 거리감이 있다는 지적이 있었다. 따라서 이주민 학습자 대상별 교육과정을 분석하여 충분한 발화를 산출할 수 있는 수집 과제의 설계가 필요하다.

⑤ 학습자 수준 판별 및 등급화 방안

- 시험 수집 결과에서 학습자의 등급 표기 방식이 결혼이주민이나 이주노동자는 1-6급, 1-4단계, 초·중·고급으로 표기되며, 중도입국청소년은 학년 이외에 학습자의 수준을 판별할 수 있는 등급 체계 자체가 존재하지 않음을 알 수 있었다. 학습자의 수준이 학습자 말뭉치의 주요한 변인 중 하나임을 고려할 때 본격 수집에 앞서 이를 측정할 수 있는 기준이나 등급을 체계화할 수 있는 방안에 대한 마련이 가장 중요한 쟁점이 될 수 있을 것으로 보인다. 사업 최종보고회의에서의 자문 결과에 따르면 정부 부처에서 개발한 표준 교육과정에 준한 이주민의 등급 체계를 수용하며, 표준 교재(『여성결혼이민자와 함께하는 한국어』, 『초등학생을 위한 표준 한국어』, 『중학생을 위한 표준 한국어』, 『고등학생을 위한 표준 한국어』, 공통 교재를 사용하고 있는 다문화센터와 각 급의 공교육 기관의 자료를 집중 수집하고 학습 중인 교재를 등급으로 표기하는 방안이 가장 합리적이다. 다만, 한 가지 고려해야 할 문제는 여성가족부나 법무부의 다문화센터에서 운영 중인 한국어 교육과정이 대부분 중급 단계까지여서 고급 학습자의 자료 수집이 어려워지게 된다. 이는 자원봉사 활동 등을 통해 학습자와 개별 접촉이 가능한 교사 인력풀을 통해 기관의 교육과정을 이수한 학습자들의 자료를 추가 수집하는 방식으로 보완하는 방안을 고려할 수 있다.

⑥ 미성년 중도입국청소년의 동의서 수합

- 미성년 학습자의 경우 학부모의 동의가 필요하다. 시험 수집에서 알림장과 함께 학부모의 동의를 받아 자료 수집을 하였으며 그 과정에서 사업의 취지를 정확히 이해하지 못하거나 자료 처리를 위한 학습자 개인 정보 제공에 대한 부담 등으로 동의를 하지 않는 비율이 매우 높았다. 또한 동의를 한 경우에도 수집 정보에 누락 항목이 많아 이를 보충하기 위한 후처리 작업이 필요하거나 그것이 불가할 경우 자료를 사용하지 못하게 되는 경우가 있었다. 따라서 학교 기관장 또는 실무 담당 교사와의 협의를 통해 이를 효율적으로 해결할 수 있는 방안이 마련되어야 한다.

⑦ 수집 실무자 교육 방안

- 시험 수집 과정에서 다문화센터와 같은 기관의 경우 상근직 행정 담당자가 수집 책임자나 실무자가 되고 교사가 수집 지원을 담당하였다. 이와 같은 경우 행정 담당자는 주로 자료를 수합하는 역할을 하고 실질적인 수집은 교사가 담당하게 되는데 본격 수집을 위해서는 이들을 효율적으로 교육할 수 있는 방안에 대한 고민이 필요하다. 유무선 연락망과 이메일 등을 통한 지속적인 교육과 관리, 소통 외에 수집 초기 단계에 말뭉치 구축 본부에서의 집합 교육을 하거나 지역별 거점 기관에 방문하여 교육을 할 수 있다.

⑧ 수집자 보상 방안

- 시험 수집 과정에서 다문화센터나 개인 접촉을 통해 자료를 수집하는 교사 모두 수집 교사나 학습자의 보상 방안에 대해 매우 관심이 많으며 직접적인 요구도 많음을 알 수 있었다. 양질의 자료를 대규모로 수집하기 위해서는 기관이나 개인 교사, 학습자의 특성을 고려한 적절한 보상 방안에 대한 고민이 필요할 것으로 보인다. 경제적인 보상 외에 각 기관의 수집 자료를 공유하고 활용 방안을 교육하여 해당 기관의 교수·학습 자료로 활용하게 하거나, 교사들 간의 학습 커뮤니티 지원 등의 방안을 추가적으로 고려해 볼 수 있다.

⑨ 자료 구축 관련

- 결혼이주여성을 비롯한 이주민은 거주 지역에서 통용되는 지역 방언에 노출되어 그것을 그대로 습득을 한 경우가 대부분이다. 따라서 구어 자료의 경우 각 지역의 방언이 필연적으로 포함되게 되는데 방언의 특성을 고려한 구어 전사 지침이나 전사 작업자 훈련, 경우에 따라서는 방언 부분에 대한 감수가 필요할 것이다.

2.2. 문어 입력

1) 작업 체계 및 절차

- 문어 입력은 표본 등록, 입력, 검수까지의 작업을 완료하여 형태 주석 이전 단계의 원시 말뭉치를 만들어 내는 과정이다. 시스템 팀과의 협업을 통해 작업자들이 시스템 내에서 각 작업의 단계를 수행하면서 발생하는 버그들을 보고하여 수정하고 작업의 효율성을 제고하기 위한 세부 기능을 요청하여 시스템 환경을 개선하는 작업이 지속적으로 이루어진다.

(1) 표본 정보 및 학습자 정보 등록

- 수집 자료, 학습자 관련 변인 정보와 함께 표본을 등록한다.
 - 표본 정보 : 학습 환경, 수집 기관, 수집 시기, 학습자 유형, 자료 유형, 작문 유형, 장르
 - 학습자 정보 : 성별, 나이, 한국어 수준(등급), 국적, 제1 언어, 한국어 학습 시간, 한국 거주 기간, 학습 목적, 직업, 기타 언어

<그림 30> 표본 정보 및 학습자 정보 등록 화면

(2) 파일 업로드 및 저장

- 원본 스캔 파일과 동의서 파일을 각각 저장하여 수집 정보 등록과 결합하여 하나의 표본으로 저장된다.

<그림 31> 원본 스캔 파일 및 동의서 화면

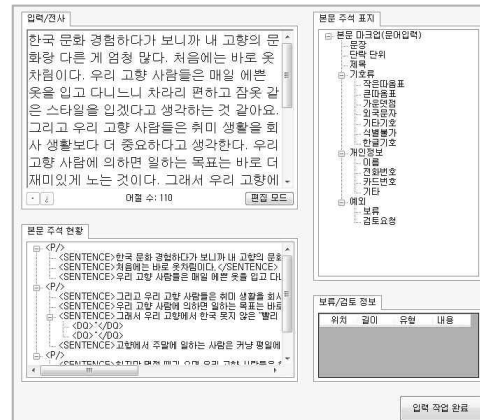
(3) 본문 입력 및 자동 주석

- 스캔 원본을 보며 본문 내용을 입력한 후 자동 주석을 생성한다.

<그림 32> 본문 입력 및 자동 주석 화면

(4) 마크업 단계

- 자동으로 생성된 주식 현황 외에 추가 주식 표시를 위해 마크업 작업 단계를 진행한다. 본문 입력창에서 해당 부분을 드래그한 후에 ‘문장, 단락 단위, 제목, 기호류(가운뎃점, 외국문자, 기타기호, 식별불가, 한글기호 등)’와 ‘개인정보(이름, 전화번호, 카드번호, 기타)’를 각각 마크업한다.



<그림 33> 마크업 화면

2) 문어 입력 결과

- 2016년 문어 원시 말뭉치는 701,113어절이 구축되었다. 그 결과 2015년 기 구축한 333,997만 어절과 합하여 누적 합계 1,002,110어절 규모의 문어 원시 말뭉치가 구축되었다.

(1) 문어 말뭉치의 속달도별 자료 분포

- 문어 말뭉치는 1급에서 6급까지의 자료가 구축되었다. 다음은 각 속달도별 구축 규모를 집계한 것이다.

<표 32> 문어 말뭉치의 숙달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	계
2015	어절 수	49,359	50,208	50,704	49,977	50,195	50,554	300,997
	파일 수	697	451	422	385	345	316	2,616
2016	어절 수	114,636	125,347	122,429	140,481	110,283	87,937	701,113
	파일 수	1,551	1,066	874	1,056	722	510	5,779
계	어절 수	163,995	175,555	173,133	190,458	160,478	138,491	1,002,110
	파일 수	2,248	1,517	1,296	1,441	1,067	826	8,395

(2) 문어 말뭉치의 국적별 자료 분포

- 문어 말뭉치는 단일 국적을 기준으로 총 114개국의 자료가 구축되었으며, 국적별 분포는 다음과 같다.

<표 33> 문어 말뭉치의 국적별 자료 분포

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	177,550	1,596	289,786	2,439	467,336	4,035
일본	46,909	382	111,748	834	158,657	1,216
대만	13,033	96	38,032	303	51,065	399
베트남	6,921	61	35,924	303	42,845	364
미국	5,070	45	28,752	237	33,822	282
홍콩	6,424	49	29,549	226	35,973	275
러시아	1,788	16	15,587	140	17,375	156
태국	4,022	33	16,568	136	20,590	169
말레이시아	4,625	35	14,654	104	19,279	139
카자흐스탄	4,011	34	10,311	80	14,322	114
몽골	1,563	14	10,709	91	12,272	105
프랑스	1,276	11	9,011	71	10,287	82
싱가포르	1,950	15	6,831	46	8,781	61

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
우즈베키스탄	2,168	20	4,810	37	6,978	57
한국	1,242	11	5,367	41	6,609	52
인도네시아	1,577	14	4,858	46	6,435	60
호주	831	9	4,250	28	5,081	37
독일	1,345	11	4,277	43	5,622	54
영국	681	6	3,527	32	4,208	38
이탈리아	1,915	14	2,349	24	4,264	38
캐나다	1,322	12	3,593	34	4,915	46
스웨덴	1,276	13	3,382	50	4,658	63
스페인	442	5	2,616	22	3,058	27
사우디아라비아	817	9	2,972	35	3,789	44
스리랑카	122	1	2,903	22	3,025	23
터키	671	6	1,673	12	2,344	18
미얀마	351	3	1,425	12	1,776	15
필리핀	483	5	1,589	18	2,072	23
키르기스스탄	0	0	2,322	16	2,322	16
네덜란드	276	3	1,045	9	1,321	12
기타 ⁶⁾	10,336	87	30,693	288	41,029	375
계	300,997	2,616	701,113	5,779	1,002,110	8,395

6) 기타에는 튀니지, 콜롬비아, 스위스, 멕시코, 캄보디아, 브라질, 인도, 벨기에, 노르웨이, 헝가리, 에콰도르, 이집트, 폴란드, 벨라루스, 이란, 세르비아, 마카오, 파키스탄, 우크라이나, 엘살바도르, 불가리아, 아제르바이잔, 코스타리카, 뉴질랜드, 모로코, 아르헨티나, 칠레, 핀란드, 알제리, 파라과이, 르완다, 방글라데시, 네팔, 나이지리아, 루마니아, 볼리비아, 페루, 룩셈부르크, 브루나이, 베네수엘라, 도미니카 공화국, 가나, 라오스, 에티오피아, 가봉, 아프가니스탄, 남수단, 포르투갈, 오스트리아, 트리니다드 토바고, 잠비아, 체코, 니카라과, 탄자니아, 모잠비크, 우루과이, 슬로바키아, 이라크, 아르메니아, 티모르, 케냐, 마다가스카르, 세네갈, 동티모르, 예멘, 우간다, 파나마, 쿠웨이트, 이스라엘, 온두라스, 과테말라, 타지키스탄, 쿠바, 시리아, 자메이카, 콩고, 아일랜드, 리비아, 콩고 민주 공화국, 리투아니아, 슬로베니아, 카메룬, 그루지아, 투르크메니스탄이 포함된다.

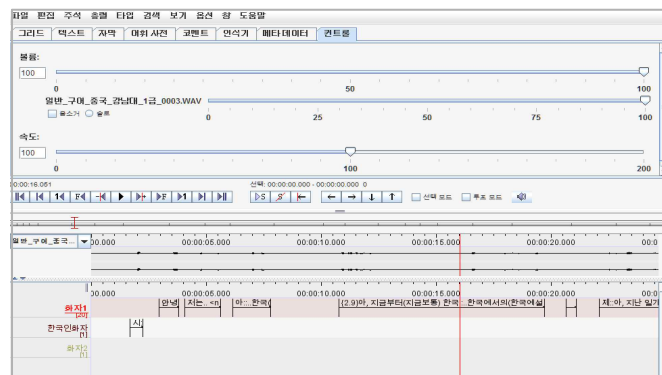
2.3 구어 전사

1) 작업 체계 및 절차

- 2015년 1차 연도에는 전사 도구 ‘엘란(ELAN)’에서 전사하고 엑셀을 활용하여 오류 주석을 실시하였다. 2차 연도 역시 기존의 방법과 동일하게 엘란에서 전사를 하되, 학습자 말뭉치 구축 시스템과 연계하여 작업을 할 수 있도록 시스템 기능을 구현하고 본격적인 작업을 앞두고 세부적인 환경을 개선하고 있다. 이에 따라 엘란은 텍스트 전사를 위한 도구로 활용이 되고, 텍스트 이외의 정보들을 시스템 내에서 2차로 주석하는 방식으로 작업이 이루어지게 된다.

(1) 전사하기

- 엘란을 사용하여 음성 발화를 전사하여 텍스트 파일로 변환한다.



<그림 34> 엘란을 활용한 전사 화면

(2) 시스템에 파일 업로드 후 마크업 작업하기

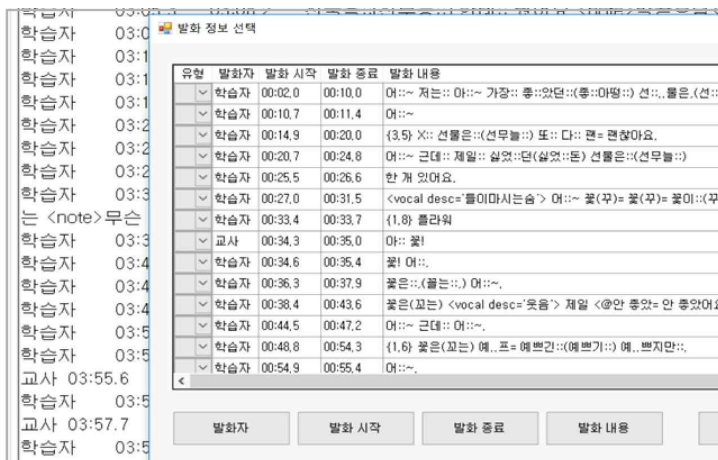
- 엘란에 음성 파일을 모두 전사를 하는 작업이 끝나면 말뭉치 구축 시스템과 연계하여 이를 반영하여 작업하는 과정을 거친다. 그 과정을 나타내면 아래와 같다.

① 발화자 정보 선택

- 주석 작업을 시작하기 전 먼저 발화자 정보를 선택한다.



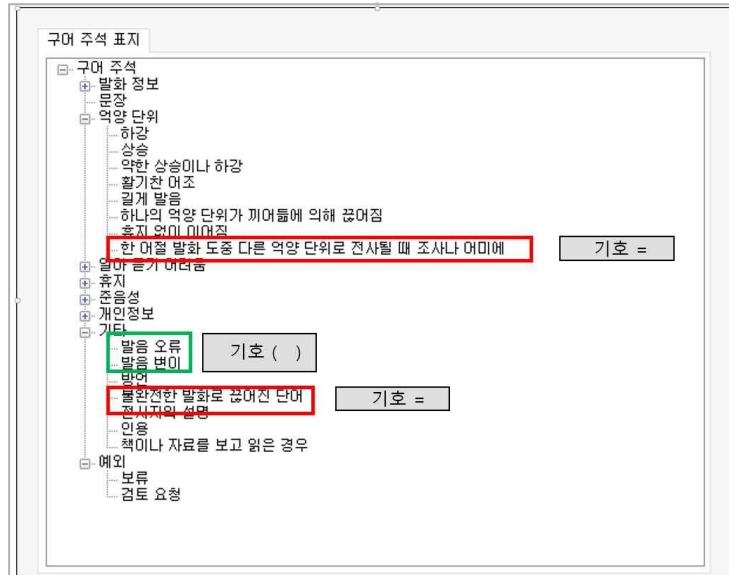
<그림 35> 발화자 정보 선택 화면 1



<그림 36> 발화자 정보 선택 화면 2

② 자동 주석 생성

- 전사 텍스트를 불러들인 후 ‘주석 자동 생성’ 버튼을 눌러 ‘발화 단위’ 및 ‘문장 단위’ 화면에 자동 주석을 생성한다.

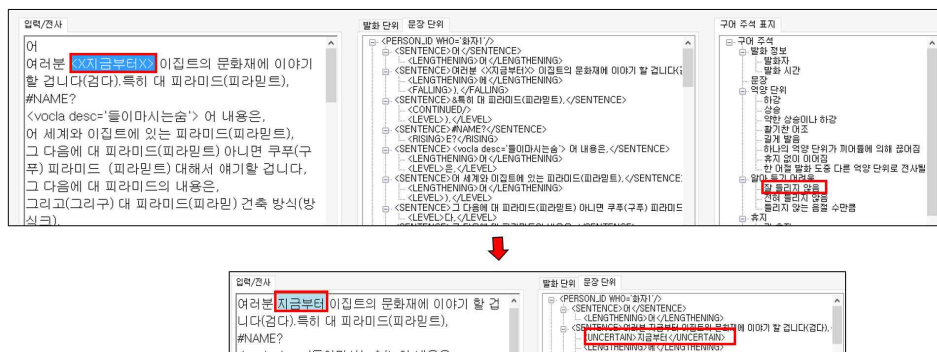


<그림 39> 구어 주석 표지

☞ 자동 주석 창에서 지원되지 않는 마크업은 수동으로 진행하여야 한다.

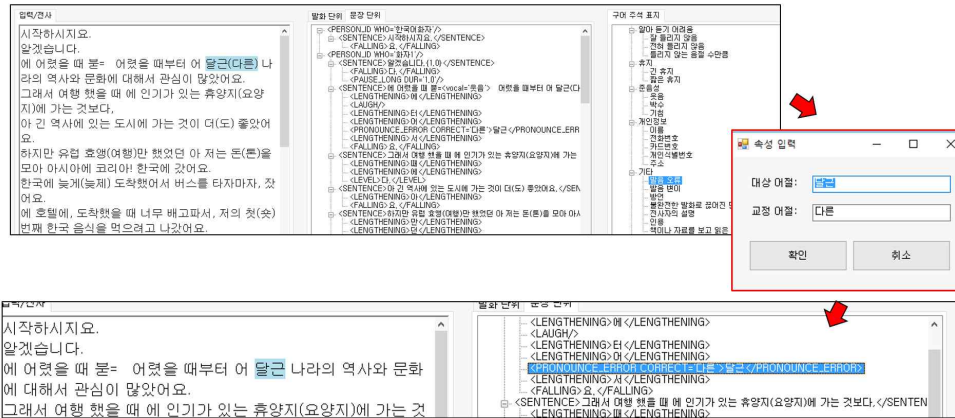
③ 주석 작업하기

- 전사 창에서 문자와 기호를 함께 드래그하고 ‘구어 주석 표지’에서 해당하는 주석을 선택하여 더블클릭을 하는 식으로 주석을 태깅하게 된다.



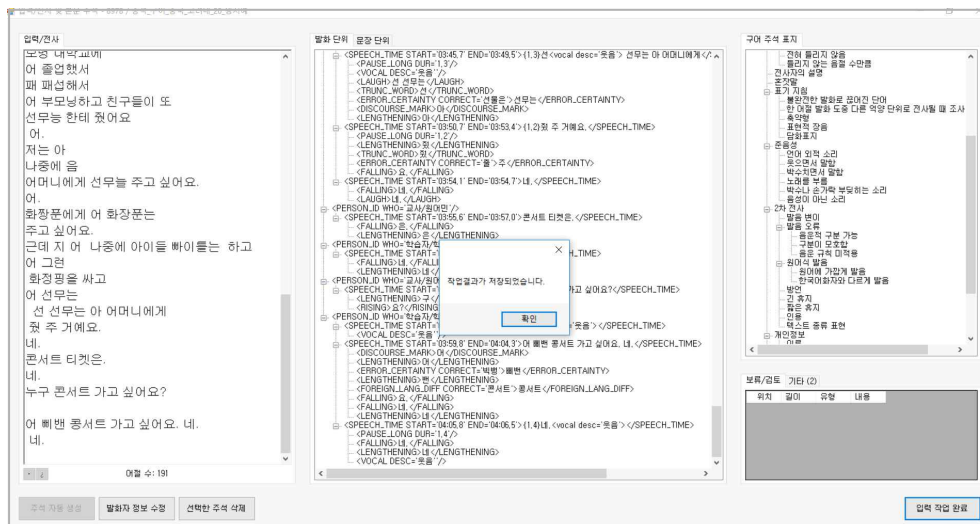
<그림 40> 주석 작업 화면 1

- 발음 오류 및 발음 변이 그리고 긴 쉼과 같이 내용을 입력해야 하는 것은 속성 값을 입력하는 식으로 주석이 이루어지게 된다.



<그림 41> 구어 주석 작업 화면 2

- 완료 화면은 다음과 같다.



<그림 42> 완료 화면 1



<그림 43> 완료 화면 2

3) 구어 전사 결과

- 2016년 구어 원시 말뭉치는 100,078어절이 구축되었다. 그 결과 2015년 기 구축한 약 53,894어절과 합하여 누적 합계 153,972어절 규모의 구어 원시 말뭉치가 구축되었다.

(1) 구어 말뭉치의 속달도별 자료 분포

- 구어 말뭉치는 1급에서 6급 학습자의 자료가 구축되었다. 다음은 각 속달도별 구축 규모를 집계한 것이다.

<표 34> 구어 말뭉치의 속달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	계
2015	어절 수	7,498	10,679	8,441	9,123	10,162	7,991	53,894
	파일 수	30	19	17	14	13	7	100

2016	어절 수	43,027	8,454	9,313	11,896	10,598	16,790	100,078
	파일 수	113	29	32	43	21	17	255
계	어절 수	50,525	19,133	17,754	21,019	20,760	24,781	153,972
	파일 수	143	48	49	57	34	24	355

(2) 구어 말뭉치의 국적별 자료 분포

○ 구어 말뭉치는 단일 국적을 기준으로 총 45개국의 자료가 구축되었다.

<표 35> 문어 말뭉치의 국적별 자료 분포

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	6,428	17	52,721	134	59,149	151
일본	9,206	16	14,501	33	23,707	49
대만	4,044	5	6,782	15	10,826	20
베트남	1,815	5	8,374	22	10,189	27
미국	3,938	6	3,184	5	7,122	11
홍콩	161	1	0	0	161	1
러시아	4,357	4	1,167	4	5,524	8
태국	1,578	4	618	2	2,196	6
말레이시아	1,208	5	1,061	3	2,269	8
카자흐스탄	2,492	3	504	2	2,996	5
몽골	763	2	383	1	1,146	3
프랑스	1,256	3	467	2	1,723	5
싱가포르	658	2	230	1	888	3
우즈베키스탄	1,828	2	478	3	2,306	5
한국	729	1	232	1	961	2
인도네시아	0	0	830	3	830	3
호주	910	3	0	0	910	3
독일	871	2	0	0	871	2

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
영국	2,258	3	0	0	2,258	3
이탈리아	0	0	1,719	3	1,719	3
캐나다	164	1	492	1	656	2
스웨덴	835	3	0	0	835	3
스페인	955	1	796	3	1,751	4
사우디아라비아	434	1	361	1	795	2
스리랑카	691	1	282	1	973	2
터키	582	1	0	0	582	1
미얀마	641	1	448	1	1,089	2
필리핀	0	0	331	1	331	1
네덜란드	863	1	0	0	863	1
기타 ⁷⁾	4,229	6	4,117	13	8,346	19
계	53,894	100	100,078	255	153,972	355

2.4. 형태 주석

1) 작업 체계 및 절차

- 1차 연도에는 입력된 학습자 자료에 대해 리눅스 기반의 지능형 형태 주석 도구 KMAT을 이용하여 자동 형태 주석을 실시하였다. 그러나 본 연구에서는 ‘입력 - 자동 형태 분석 - 형태 검수 과정’을 모두 학습자 맞춤형 구축 시스템에서 하도록 시스템 환경을 구축하여 작업을 수행하였다. 이에 따라 자동 형태 분석은 구축 도구에 내장된 형태 주석기를 사용하게 되었으며, 작업 과정에서 형태 주석 작업자들의 요구 사항을 반영하여 지침에 따라 자동 형태 주석 사전을 지침에 맞추어 보완하는 등 시스템 작업 환경을 개선하였다.

7) 기타에는 콜롬비아, 스위스, 멕시코, 캄보디아, 벨기에, 헝가리, 이집트, 벨라루스, 세르비아, 불가리아, 아제르바이잔, 코스타리카, 아르헨티나, 알제리, 파라과이, 네팔이 포함된다.

- 자동 형태 주석이 완료된 결과물에 대한 1차 및 2차 검수를 진행하는 과정에서, 지침 적용의 대원칙을 설정하였다. 1차적으로 학습자 말뭉치 형태 주석 지침을 기초로 하되, 지침을 통해서 해결이 불가능한 어휘에 대해서는 <표준국어대사전>을 참조하도록 하였다. 사전으로 확인이 불가능한 경우, 세종 말뭉치 검색 결과를 참조하도록 하였다. 단, 조사 및 어미 결합형에 대해서는 사전에 등재된 형태일지라도 따로 분석하는 것을 원칙으로 하였다.
- 구어 자료의 경우, 자동 형태 주석 과정에서 구어 전사자가 전사 과정에서 작성한 규범 표기를 적극적으로 활용하였다. 이는 실질적으로 학습자의 구어 자료에 나타난 오류 발화에 대한 교정 어절이라고 볼 수 있다. 전사자가 부여한 규범 표기를 적극적으로 활용함으로써, 정규적인 텍스트가 아닌 학습자 말뭉치의 특성을 어느 정도 극복하는 반면, 전사자가 부여한 규범 표기의 형태소 분석 결과를 오류 발화의 형태소 분석에 역으로 활용할 수 있었다.

(1) 작업 할당

- 입력 및 전사가 완료된 자료는 [형태 분석 변환] 기능을 이용하여 자동 형태 분석이 된 후 형태 주석 작업자들에게 할당된다.

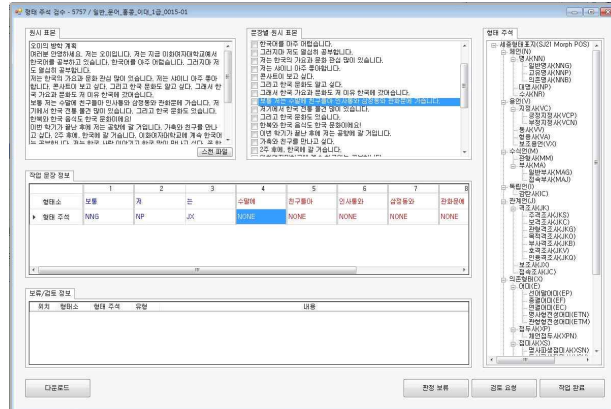
표준번호	원문	분석결과	작업자	작업일자	작업시간	작업량	비고
4000	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4001	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4002	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4003	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4004	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4005	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4006	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4007	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4008	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4009	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4010	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4011	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4012	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4013	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4014	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4015	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4016	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4017	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4018	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4019	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4020	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...
4021	원문: ...	분석결과: ...	작업자: ...	작업일자: ...	작업시간: ...	작업량: ...	비고: ...

<그림 44> 형태 분석 작업 할당 화면

(2) 1차 형태 주석 검수

- 시스템에 내장된 도구를 사용한 기계 분석 과정에서 오분석된 부분이나 학습자 오류에 의한 분석 불능 처리된 부분을 검토하여 수정하는 단계이

다. 중의성을 가질 수 있는 형태소는 파란색으로, 자동 형태소 분석기로 분석이 어려운 형태에 대해서는 빨간색으로 형태소가 표시되도록 하여 작업자의 편의 및 정확한 분석을 돕도록 시스템 환경을 구현하였다.



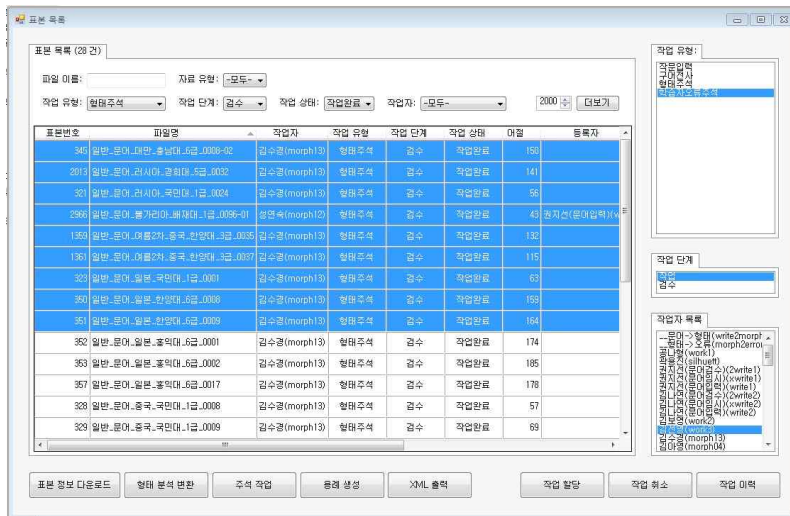
<그림 45> 형태 주석 작업 화면

(3) 2차 형태 주식 검수

- 1차 형태 주석 검수를 마친 후 숙련된 상위의 작업자가 검수를 하여 질적 완성도를 제고하는 단계이다. 1차 형태 주석 검수 자료에 대한 전수 검수 작업이 이루어지며 검수 의견을 1차 작업자에게 전달하여 검수와 교육이 동시에 진행된다.

(4) 오류 주석으로의 연계

- 2단계의 검수 작업을 마친 후 작업 [완료] 처리를 하면 오류 주석 작업을 할 수 있도록 작업 단계가 넘어간다.



<그림 46> 형태 주석 작업 완료 및 오류 주석으로의 연계

2) 형태 주석 결과

- 2016년 형태 주석 말뭉치는 문어 501,117어절, 구어 51,029어절이 구축되었다. 2015년 기구축한 문어 200,994어절, 구어 21,246어절과 합산한 형태 주석 말뭉치의 누적 합계는 문어 702,111어절, 구어 72,275어절이다.

(1) 형태 주석 말뭉치의 속달도별 자료 분포

- 형태 주석 말뭉치의 속달도별 자료 분포는 다음과 같다.

<표 36> 형태 주석 말뭉치의 속달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	계
2015	문어	어절 수	31,664	30,423	35,947	37,425	34,154	200,994
		파일 수	456	269	311	292	235	1,766
	구어	어절 수	3,036	3,687	2,032	3,995	3,511	21,246
		파일 수	16	5	5	7	5	42
2016	문어	어절 수	77,174	84,227	100,492	87,889	78,008	501,117

	구어	파일 수	1,044	718	724	598	506	433	4,023
		어절 수	7,154	7,179	6,820	8,852	7,430	13,594	51,029
		파일 수	22	26	25	33	18	14	138
계	문어	어절 수	108,838	114,650	136,439	125,314	112,162	104,708	702,111
		파일 수	1,500	987	1,035	890	741	636	5,789
	구어	어절 수	10,190	10,866	8,852	12,847	10,941	18,579	72,275
		파일 수	38	31	30	40	23	18	180

(2) 형태 주석 말뭉치의 국적별 자료 분포

① 문어

- 문어 형태 주석 말뭉치는 단일 국적 기준 100개국의 자료가 구축되었으며 국적별 자료 분포는 다음과 같다.

<표 37> 문어 형태 주석 말뭉치의 국적별 자료 규모

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	122,104	1,110	193,840	1,609	315,944	2,719
일본	28,197	225	109,312	810	137,509	1,035
베트남	5,500	50	29,176	222	34,676	272
대만	7,636	59	23,326	173	30,962	232
미국	3,543	29	23,707	199	27,250	228
홍콩	4,118	31	17,224	121	21,342	152
러시아	995	9	11,672	100	12,667	109
카자흐스탄	3,214	27	7,968	65	11,182	92
태국	2,960	26	8,140	63	11,100	89
말레이시아	3,025	24	5,711	43	8,736	67
몽골	964	10	8,139	62	9,103	72
프랑스	751	6	6,855	52	7,606	58
싱가포르	1,743	13	4,832	32	6,575	45

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
우즈베키스탄	1,499	14	3,466	31	4,965	45
호주	218	3	3,185	25	3,403	28
영국	367	3	3,089	28	3,456	31
인도네시아	1,034	9	2,107	22	3,141	31
독일	1,345	11	1,941	25	3,286	36
한국	891	7	2,750	22	3,641	29
이탈리아	575	5	2,560	22	3,135	27
스리랑카	122	1	2,752	20	2,874	21
캐나다	765	6	2,530	27	3,295	33
사우디아라비아	612	7	1,525	13	2,137	20
스페인	364	4	1,747	15	2,111	19
스웨덴	853	10	1,049	13	1,902	23
미얀마	245	2	1,449	12	1,694	14
키르기스스탄	0	0	1,480	10	1,480	10
터키	671	6	789	5	1,460	11
네덜란드	203	2	374	3	577	5
필리핀	483	5	385	5	868	10
기타 ⁸⁾	5,997	52	18,037	174	24,034	226
계	200,994	1,766	501,117	4,023	702,111	5,789

8) 기타에는 벨기에, 멕시코, 폴란드, 세르비아, 이집트, 벨라루스, 우크라이나, 캄보디아, 파키스탄, 인도, 아르헨티나, 마카오, 헝가리, 알제리, 핀란드, 아제르바이잔, 뉴질랜드, 칠레, 네팔, 모로코, 이란, 노르웨이, 브라질, 가봉, 볼리비아, 에티오피아, 루마니아, 포르투갈, 르완다, 오스트리아, 엘살바도르, 나이지리아, 방글라데시, 도미니카 공화국, 가나, 잠비아, 콜롬비아, 아프가니스탄, 니카라과, 룩셈부르크, 에콰도르, 라오스, 브루나이, 페루, 슬로바키아, 아르메니아, 체코, 코스타리카, 케냐, 마다가스카르, 베네수엘라, 트리니다드 토바고, 튀니지, 쿠웨이트, 이스라엘, 이라크, 과테말라, 타지키스탄, 불가리아, 세네갈, 콩고, 탄자니아, 스위스, 콩고 민주 공화국, 남수단, 리투아니아, 슬로베니아, 카메룬, 그루지아, 트르크메니스탄이 포함된다.

② 구어

- 구어 형태 주식 말뭉치는 단일 국적 기준 40개국의 자료가 구축되었으며 국적별 자료 분포는 다음과 같다.

<표 38> 구어 형태 주식 말뭉치의 국적별 자료 규모

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	3,913	10	15,696	41	19,609	51
일본	3,150	4	9,972	24	13,122	28
베트남	520	1	5,783	15	6,303	16
대만	989	1	4,967	12	5,956	13
미국	1,567	3	1,871	3	3,438	6
홍콩	161	1	0	0	161	1
러시아	568	1	1,167	4	1,735	5
카자흐스탄	1,583	2	504	2	2,087	4
태국	356	1	1,177	3	1,533	4
말레이시아	405	2	554	2	959	4
몽골	575	1	383	1	958	2
프랑스	0	0	467	2	467	2
싱가포르	220	1	230	1	450	2
우즈베키스탄	841	1	478	3	1,319	4
호주	910	3	0	0	910	3
영국	660	1	0	0	660	1
인도네시아	0	0	830	3	830	3
독일	631	1	0	0	631	1
한국	0	0	232	1	232	1
이탈리아	0	0	347	1	347	1
스리랑카	691	1	282	1	973	2
캐나다	0	0	492	1	492	1
사우디아라비아	434	1	361	1	795	2
스페인	0	0	796	3	796	3

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
스웨덴	551	2	0	0	551	2
미얀마	0	0	448	1	448	1
네덜란드	863	1	0	0	863	1
필리핀	0	0	331	1	331	1
기타 ⁹⁾	1,658	3	3,661	12	5,319	15
계	21,246	42	51,029	138	72,275	180

3) 형태 주석 결과의 분석

(1) 문어

① 속달도 단계별

<표 39> 문어 형태 주석 결과_속달도 단계별

	1급	2급	3급	4급	5급	6급	계
일반명사	41,495	43,117	53,492	55,568	53,569	51,992	299,233
고유명사	10,001	7,663	4,971	2,220	2,158	2,764	29,777
의존명사	7,023	6,593	9,296	8,822	8,101	8,025	47,860
대명사	7,711	6,814	6,851	4,945	2,998	2,782	32,101
수사	1,122	416	447	349	339	277	2,950
동사	18,640	20,946	26,110	23,626	20,099	17,861	127,282
형용사	6,234	8,963	10,031	8,708	7,459	5,424	46,819
보조용언	2,839	3,608	5,366	5,416	4,603	4,275	26,107
부정지정사	29	130	251	340	503	508	1,761
긍정지정사	3,539	2,613	3,476	3,528	3,764	3,476	20,396
관형사	1,405	2,065	3,412	3,294	3,075	2,848	16,099

9) 기타에는 벨기에, 멕시코, 세르비아, 이집트, 벨라루스, 캄보디아, 아르헨티나, 헝가리, 알제리, 아제르바이잔, 네팔, 파라과이가 포함된다.

	1급	2급	3급	4급	5급	6급	계
일반부사	8,581	11,281	13,042	10,227	7,554	6,687	57,372
접속조사	3,395	2,907	3,093	2,129	1,521	1,265	14,310
감탄사	16	48	24	10	11	8	117
주격조사	5,734	7,623	10,035	9,473	8,503	7,231	48,599
보격조사	47	235	333	550	691	668	2,524
관형격조사	1,650	1,879	2,139	2,781	3,225	3,315	14,989
목적격조사	9,213	7,976	8,800	9,487	8,611	8,396	52,483
부사격조사	15,821	13,109	13,537	11,344	10,178	9,591	73,580
공동격조사	1,464	1,086	1,186	1,041	1,235	1,076	7,088
호격조사	2	20	6	2	2	0	32
인용격조사	11	171	108	44	73	119	526
보조사	11,318	10,669	11,556	10,614	9,758	8,379	62,294
선어말어미	5,463	6,246	6,900	4,356	2,886	3,409	29,260
연결어미	8,594	14,786	22,063	21,529	18,951	17,486	103,409
종결어미	22,261	17,223	17,435	13,427	10,469	8,776	89,591
명사형 어미	281	1,280	1,542	1,706	1,970	1,359	8,138
관형사형 어미	4,692	9,070	13,908	14,805	14,942	13,838	71,255
체언접두사	16	12	33	323	269	313	966
명사파생접미사	1,371	1,997	2,502	3,995	4,356	4,665	18,886
동사파생접미사	3,338	4,431	6,873	7,279	7,533	7,941	37,395
형용사파생접미사	1,039	1,657	2,896	2,593	2,402	1,957	12,544
어근	493	547	681	738	821	709	3,989
분석불능	338	323	342	200	253	184	1,640
계	205,176	217,504	262,737	245,469	222,882	207,604	1,361,372

② 국적별

<표 40> 문어 형태 주석 결과_국적별

	중국	일본	베트남	대만	미국	기타	계
일반명사	134,149	58,062	15,154	14,336	11,290	66,242	299,233
고유명사	12,709	5,944	1,156	1,104	1,504	7,360	29,777
의존명사	20,450	10,467	2,229	2,088	1,779	10,847	47,860
대명사	15,844	4,924	1,843	1,043	1,273	7,174	32,101
수사	1,350	344	160	96	152	848	2,950
동사	55,312	26,828	6,048	5,417	5,082	28,595	127,282
형용사	22,077	9,387	2,327	1,907	1,691	9,430	46,819
보조용언	10,961	6,221	1,313	1,095	883	5,634	26,107
부정지정사	629	393	100	119	64	456	1,761
긍정지정사	8,569	4,659	1,046	909	759	4,454	20,396
관형사	6,875	3,040	724	750	712	3,998	16,099
일반부사	28,338	9,609	2,877	2,577	2,067	11,904	57,372
접속조사	7,034	2,356	692	509	572	3,147	14,310
감탄사	68	13	3	3	10	20	117
주격조사	22,668	9,446	2,335	2,269	1,939	9,942	48,599
보격조사	1,019	675	96	152	64	518	2,524
관형격조사	6,226	2,940	807	807	612	3,597	14,989
목적격조사	23,545	10,095	2,697	2,220	2,017	11,909	52,483
부사격조사	32,339	14,102	3,450	3,096	3,101	17,492	73,580
공동격조사	2,568	1,645	401	320	373	1,781	7,088
호격조사	22	4	1	0	1	4	32
인용격조사	172	141	10	25	38	140	526
보조사	27,153	14,263	3,165	2,464	2,102	13,147	62,294
선어말어미	12,383	5,735	1,487	1,161	1,519	6,975	29,260
연결어미	42,445	23,918	5,461	4,771	3,723	23,091	103,409
종결어미	44,513	14,567	4,329	3,337	3,574	19,271	89,591
명사형 어미	3,265	1,902	468	348	319	1,836	8,138

	중국	일본	베트남	대만	미국	기타	계
관형사형 어미	29,307	17,257	3,488	3,466	2,589	15,148	71,255
체언접두사	471	182	16	81	22	194	966
명사파생접미사	8,053	3,389	842	1,051	949	4,602	18,886
동사파생접미사	16,060	7,831	2,191	1,929	1,297	8,087	37,395
형용사파생접미사	5,985	2,252	722	529	424	2,632	12,544
어근	1,691	822	269	175	154	878	3,989
분석불능	932	109	81	34	88	396	1,640
계	605,182	273,522	67,988	60,188	52,743	301,749	1,361,372

(2) 구어

① 숙달도 단계별

<표 41> 구어 형태 주석 결과_숙달도 단계별

	1급	2급	3급	4급	5급	6급	계
일반명사	3,135	3,086	2,813	4,457	4,390	6,251	24,132
고유명사	889	763	589	435	551	423	3,650
의존명사	498	606	506	753	625	1,273	4,261
대명사	603	516	456	691	334	674	3,274
수사	224	148	89	107	195	111	874
동사	1,767	1,880	1,556	2,125	1,765	2,827	11,920
형용사	482	831	586	760	426	1,277	4,362
보조용언	284	252	308	361	224	595	2,024
부정지정사	15	25	39	44	37	113	273
긍정지정사	273	257	213	361	321	532	1,957
관형사	182	258	435	563	467	954	2,859
일반부사	744	1,022	998	1,247	702	1,963	6,676
접속조사	275	346	278	407	219	529	2,054
감탄사	2,995	2,661	1,427	2,777	1,601	3,383	14,844
주격조사	413	499	579	683	608	1,119	3,901

	1급	2급	3급	4급	5급	6급	계
보격조사	3	11	26	29	32	70	171
관형격조사	38	74	80	134	248	278	852
목적격조사	344	357	393	601	576	770	3,041
부사격조사	966	892	705	891	1,076	1,134	5,664
공동격조사	139	56	65	73	119	127	579
호격조사	0	7	0	0	0	0	7
인용격조사	0	6	7	6	18	16	53
보조사	591	629	602	973	669	1,354	4,818
선어말어미	467	442	640	537	673	759	3,518
연결어미	642	1,064	1,489	1,805	1,349	2,929	9,278
종결어미	2,135	1,823	983	1,309	962	1,661	8,873
명사형 어미	19	67	67	78	117	143	491
관형사형 어미	280	584	693	1,147	942	1,692	5,338
체언접두사	1	2	1	5	20	83	112
명사파생접미사	131	129	173	295	274	608	1,610
동사파생접미사	278	312	447	629	547	919	3,132
형용사파생접미사	104	86	123	159	143	247	862
어근	44	30	57	63	38	117	349
분석불능	414	386	225	476	301	453	2,255
계	19,375	20,107	17,648	24,981	20,569	35,384	138,064

② 국적별

<표 42> 구어 형태 주석 결과_국적별

	중국	일본	베트남	대만	미국	기타	계
일반명사	6,582	4,673	2,118	2,380	1,063	7,316	24,132
고유명사	881	605	478	237	180	1,269	3,650
의존명사	1,085	886	383	420	191	1,296	4,261
대명사	814	541	413	304	100	1,102	3,274
수사	211	134	65	53	25	386	874

	중국	일본	베트남	대만	미국	기타	계
동사	3,105	2,330	1,222	1,074	497	3,692	11,920
형용사	1,124	919	416	409	210	1,284	4,362
보조용언	524	443	209	217	85	546	2,024
부정지정사	82	79	9	27	8	68	273
긍정지정사	485	519	150	171	60	572	1,957
관형사	721	662	224	338	115	799	2,859
일반부사	1,749	1,249	684	719	351	1,924	6,676
접속조사	455	396	221	246	64	672	2,054
감탄사	3,784	2,575	2,182	1,176	486	4,641	14,844
주격조사	957	853	291	403	188	1,209	3,901
보격조사	46	52	3	22	3	45	171
관형격조사	209	196	45	108	42	252	852
목적격조사	851	596	181	270	131	1,012	3,041
부사격조사	1,394	1,063	589	504	280	1,834	5,664
공동격조사	152	133	43	64	24	163	579
호격조사	1	0	2	0	0	4	7
인용격조사	11	23	0	2	13	4	53
보조사	1,365	1,091	380	460	200	1,322	4,818
선어말어미	806	758	356	259	199	1,140	3,518
연결어미	2,235	2,338	673	956	428	2,648	9,278
종결어미	2,553	1,298	1,217	723	325	2,757	8,873
명사형 어미	102	130	20	48	25	166	491
관형사형 어미	1,349	1,316	358	535	223	1,557	5,338
체언접두사	47	21	0	29	0	15	112
명사파생접미사	483	353	45	150	76	503	1,610
동사파생접미사	827	713	284	321	135	852	3,132
형용사파생접미사	274	150	72	88	33	245	862
어근	117	73	24	35	6	94	349
분석불능	697	355	292	179	62	670	2,255
계	36,078	27,523	13,649	12,927	5,828	42,059	138,064

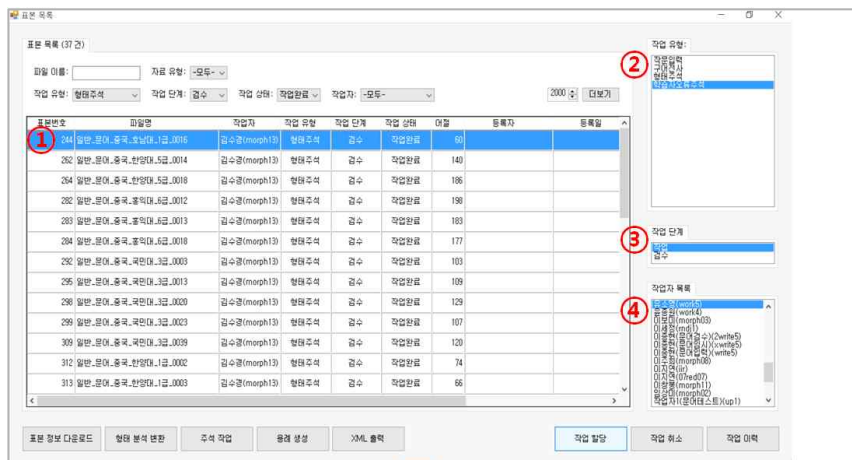
2.5. 오류 주석

1) 작업 체계 및 절차

- 오류 주석 작업은 크게 작업 할당, 오류 주석 작업, 오류 주석 검수 세 단계로 진행된다.

(1) 오류 주석 작업 할당

- 오류 주석 작업은 형태 주석에서 검수를 완료한 파일을 오류 주석자들에게 할당하는 것으로부터 시작된다. 따라서 아래의 그림과 같이 시스템 상에서 ① 형태 주석에서 검수 완료된 표본들을 다시 ② 작업 유형에서 ‘학습자오류주석’, ③ 작업 단계에서 ‘작업’을 선택하여 ④ 오류 주석 작업자들에게 표본을 할당하면 해당 작업자들에게 표본 목록이 형성된다.



<그림 47> 오류 주석 작업 할당 화면

작업 관리자

파일 이름:

작업 유형: [오류] 작업 단계: [오류] 작업 상태: [오류] 분류/종류: [오류]

표본번호	파일명	작업자	작업유형	작업단계	작업상태	대량	보류	완료	작업할당일	작업종료일
1	9747_일반_구어_비트남_통역_309_202_004X201_방01	유소영(work5)	학습자오류주석	작업	작업할당	207	0	0	2016-12-10 오전 1:46	
2	8332_일반_구어_참가자_최대_4급_0037	유소영(work5)	학습자오류주석	작업	작업할당	234	0	0	2016-12-10 오전 1:46	
3	8331_일반_구어_한국_외대_4급_0020	유소영(work5)	학습자오류주석	작업	작업할당	232	0	0	2016-12-10 오전 1:46	
4	8281_일반_구어_중국_광남대_2급_0006	유소영(work5)	학습자오류주석	작업	작업할당	96	0	0	2016-12-10 오전 1:46	
5	8248_일반_구어_일본_고대_5급_0005	유소영(work5)	학습자오류주석	작업	작업할당	358	0	0	2016-12-10 오전 1:46	
6	8247_일반_구어_일본_광남대_2급_0007	유소영(work5)	학습자오류주석	작업	작업할당	116	0	0	2016-12-10 오전 1:46	
7	8283_일반_구어_중국_광남대_2급_0011	유소영(work5)	학습자오류주석	작업	작업중	100	0	0	2016-12-08 오후 3:15	
8	8239_일반_구어_아랍어권_광남대_4급_0007	유소영(work5)	학습자오류주석	작업	작업할당	189	0	0	2016-12-08 오후 3:15	
9	8237_일반_구어_스페인_고대_4급_0007	유소영(work5)	학습자오류주석	작업	작업할당	245	0	0	2016-12-08 오후 3:15	
10	8236_일반_구어_사우디아라비아_고대_5급_0002	유소영(work5)	학습자오류주석	작업	작업할당	361	0	0	2016-12-08 오후 3:15	
11	8233_일반_구어_비트남_최대_4급_0053	유소영(work5)	학습자오류주석	작업	작업할당	212	0	0	2016-12-08 오후 3:14	
12	8184_일반_구어_아랍_고대_5급_0013	유소영(work5)	학습자오류주석	작업	작업할당	188	0	0	2016-12-08 오후 3:14	
13	1576_일반_문어_일본_광대_2급_0028	유소영(work5)	학습자오류주석	작업	작업완료	76	0	0	2016-12-06 오전 9:53	2016-12-06 오전 9:54
14	7192_일반_문어_러시아_공북대_1급_0031-01	유소영(work5)	학습자오류주석	작업	작업중	85	1	0	2016-11-26 오후 5:14	
15	2804_일반_문어_일본_연세대_4급_0017-01	유소영(work5)	학습자오류주석	작업	작업중	180	0	0	2016-11-26 오후 5:13	

오류 주석 다운로드 판정 보류 다운로드 검토 요청 다운로드 검토 요청 검색 작업 이력 닫기

<그림 48> 개인 오류 작업자들에게 할당 완료된 파일 및 화면

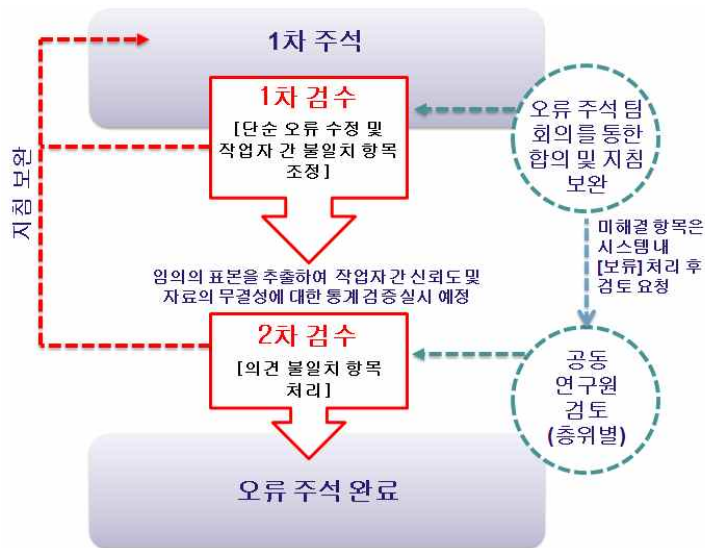
- 위의 그림처럼 개인 오류 작업자들에게 파일을 할당하면 작업상태에서 ‘작업할당’으로 파일 목록이 생성된다. 이 파일을 클릭하여 작업을 시작하면 작업상태는 ‘작업중’으로 바뀌게 되고, 오류 주석을 완료하여 ‘작업완료’를 누르면 작업상태는 ‘작업완료’로 바뀌게 된다.

(2) 오류 주석 작업

- 오류 주석 작업은 할당된 표본 목록을 선택하여 실시한다. 표본을 선택하면 크게 5부분으로 나눌 수 있는 창이 나타난다. ① 원시 표본, ② 문장별 원시 표본, ③ 작업 문장 정보, ④ 학습자오류주석표지/형태주석표지, ⑤ 보류 검토/기타 창이 나타난다.
- ① 원시 표본은 문어 자료와 구어 자료를 전사한 문장들이 나타나며, ② 문장별 원시 표본에서는 문장별로 클릭하게 되면, ③ 작업 문장 정보에 해당 문장을 형태 주석한 것이 표시된다. 오류 주석은 형태 주석한 문장을 보면서 이 창에서 오류 주석 작업을 수행한다.
- 오류가 있는 부분에 교정어절을 직접 써주고, 교정주석에는 교정어절의 형태 주석을 하는 것으로 ④ 학습자오류주석표지/형태주석표지에서 형태 주석 표지를 누르고, 해당 형태 주석 표지를 클릭하면 자동으로 주석이 반영된다. 다음으로 분석 불가능한 오류일 경우에는 분석불가능(IMP) 표지

를 주고, 가능할 경우에 해당 오류 위치와 오류 양상, 오류 층위를 각각
④ 학습자오류주석표지에서 선택하면 태그가 저절로 부착된다. 이렇게 문
장별로 오류 주석 작업을 진행하도록 한다.

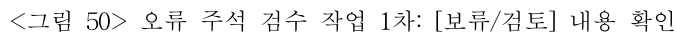
- 오류 주석을 하면서 문어 전사자의 실수로 보이는 경우에는 ① 원시 표
본 창의 **스캔파일**을 클릭하여 실제 학습자의 원본 파일을 확인하여, 학습
자가 쓴 것과 다를 때는 수정한다.
- 오류 주석을 하면서 여러 교정을 줄 수 있는 경우, 또는 오류 양상, 층위
등을 주석하는 데 논의가 필요한 경우는 ⑤ 오류 검토/기타 창에 써서 검
수 단계에서 확인하여 결정하도록 한다.
- 모든 문장에 대해 오류 주석을 완료하면 아래의 **작업완료**를 눌러 작업을
완료한다.



<그림 49> 오류 주석 작업 절차

- 1차적으로 개별 주석자가 작업 완료한 오류 주석에 대하여 교차 검수를 실시한다. 오류 주석 작업 할당과 마찬가지로 오류 주석이 완료된 표본들
을 다시 작업 유형에서 ‘학습자오류주석’ → 작업 단계 ‘검수’를 선택하여
검수 파일들을 검수자들에게 할당한다.
- 할당된 검수 표본 목록을 선택하여 오류 주석 작업과 동일하게 주석된 오
류들을 검수한다. 잘못된 오류 주석을 수정하고, 오류 주석 단계에서 검토

○ 최종적으로 오류 주식 검수를 완료한 파일은 엑셀로 다운로드 받아 오류 위치, 오류 양상, 오류 총위를 소트하여 다시 한 번 검수한다.



<그림 51> 오류 주석 검수 작업 2차 : 엑셀 파일 확인

2) 오류 주석 결과

- 2016년 오류 주석 말뭉치는 문어 100,000어절, 구어 52,191어절이 구축되었다. 2015년 기구축한 문어 42,886어절, 구어 11,777어절과 합산한 오류 주석 말뭉치의 누적 합계는 문어 142,886어절, 구어 63,968어절이다.

(1) 오류 주석 말뭉치의 숙달도별 자료 분포

- 오류 주석 말뭉치의 숙달도별 자료 분포는 다음과 같다.

<표 43> 오류 주석 말뭉치의 숙달도별 자료 분포

구분			1급	2급	3급	4급	5급	6급	계
2015	문어	어절 수	6,804	7,652	6,604	7,141	7,400	7,285	42,886
		파일 수	96	64	48	51	48	45	352
	구어	어절 수	1,674	1,890	2,032	1,917	1,978	2,286	11,777
		파일 수	9	3	5	4	3	2	26
2016	문어	어절 수	14,843	16,138	17,747	16,120	18,590	16,562	100,000
		파일 수	233	152	140	133	136	96	890
	구어	어절 수	7,860	7,179	6,820	8,852	8,963	12,517	52,191
		파일 수	25	26	25	33	20	13	142
계	문어	어절 수	21,647	23,790	24,351	23,261	25,990	23,847	142,886
		파일 수	329	216	188	184	184	141	1,242
	구어	어절 수	9,534	9,069	8,852	10,769	10,941	14,803	63,968
		파일 수	34	29	30	37	23	15	168

(2) 오류 주석 말뭉치의 국적별 자료 분포

① 문어

- 문어 오류 주석 말뭉치는 총 68개 국적의 자료가 구축되었으며 국적별 자료 분포는 다음과 같다.

<표 44> 문어 오류 주석 말뭉치의 국적별 자료 분포

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	22,592	193	32,128	288	54,720	481
일본	10,986	87	24,358	197	35,344	284
미국	1,410	10	8,786	78	10,196	88
베트남	456	5	4,130	33	4,586	38
대만	1,111	8	3,624	36	4,735	44
러시아	379	3	3,085	29	3,464	32
카자흐스탄	2,381	18	665	6	3,046	24
태국	556	4	1,361	12	1,917	16
프랑스	199	1	2,496	18	2,695	19
말레이시아	0	0	2,117	18	2,117	18
홍콩	645	5	1,567	14	2,212	19
독일	124	1	1,173	16	1,297	17
인도네시아	77	1	1,007	10	1,084	11
몽골	213	2	1,288	12	1,501	14
영국	287	2	824	8	1,111	10
우즈베키스탄	348	3	819	7	1,167	10
호주	0	0	674	7	674	7
싱가포르	135	1	781	7	916	8
스웨덴	161	2	575	5	736	7
미얀마	124	1	623	4	747	5
사우디아라비아	0	0	389	4	389	4
캐나다	116	1	494	6	610	7
스페인	0	0	289	3	289	3
이탈리아	0	0	529	8	529	8
세르비아	0	0	77	1	77	1
멕시코	0	0	137	2	137	2
이집트	0	0	131	1	131	1
아르헨티나	152	1	154	1	306	2

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
기타 ¹⁰⁾	434	3	5,719	59	6,153	62
계	42,886	352	100,000	890	142,886	1,242

② 구어

- 구어 오류 주석 말뭉치는 총 39개 국적의 자료가 구축되었으며 국적별 자료 분포는 다음과 같다.

<표 45> 구어 오류 주석 말뭉치의 국적별 자료 분포

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국	924	5	16,225	43	17,149	48
일본	1,297	1	9,972	24	11,269	25
미국	927	2	1,871	3	2,798	5
베트남	520	1	5,783	15	6,303	16
대만	989	1	4,967	12	5,956	13
러시아	568	1	1,167	4	1,735	5
카자흐스탄	627	1	504	2	1,131	3
태국	356	1	1,177	3	1,533	4
프랑스	0	0	467	2	467	2
말레이시아	286	1	554	2	840	3
홍콩	161	1	0	0	161	1
독일	631	1	0	0	631	1
인도네시아	0	0	830	3	830	3
몽골	0	0	383	1	383	1
영국	660	1	0	0	660	1

- 10) 기타에는 한국, 헝가리, 필리핀, 네덜란드, 아제르바이잔, 네팔, 캄보디아, 모로코, 키르기스스탄, 우크라이나, 터키, 인도, 마카오, 니카라과, 포르투갈, 폴란드, 이란, 슬로바키아, 가나, 마다가스카르, 볼리비아, 루마니아, 도미니카 공화국, 가봉, 콜롬비아, 오스트리아, 과테말라, 노르웨이, 핀란드, 스위스, 불가리아, 나이지리아, 파키스탄, 르완다, 페루, 라오스, 세네갈, 에티오피아가 포함된다.

- 11) 기타에는 한국, 알제리, 벨라루스, 헝가리, 필리핀, 파라과이, 아제르바이잔, 네팔, 캄보

국적	2015		2016		계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
우즈베키스탄	0	0	478	3	478	3
호주	648	2	262	1	910	3
싱가포르	220	1	230	1	450	2
스웨덴	551	2	0	0	551	2
미얀마	0	0	448	1	448	1
사우디아라비아	434	1	361	1	795	2
캐나다	0	0	492	1	492	1
스페인	0	0	796	3	796	3
스리랑카	691	1	282	1	973	2
이탈리아	0	0	347	1	347	1
세르비아	763	1	0	0	763	1
벨기에	0	0	740	1	740	1
멕시코	0	0	533	3	533	3
이집트	524	1	0	0	524	1
아르헨티나	0	0	341	1	341	1
기타 ¹¹⁾	0	0	2,981	10	2,981	10
계	11,777	26	52,191	142	63,968	168

3) 오류 주석 결과의 분석¹²⁾

(1) 문어

① 분석 불가능 여부

가. 숙달도 단계별

다아가 포함된다.

12) 오류 주석 결과 분석 자료는 2016년 12월 16일 시스템 통계를 기준으로 한 것으로 향후 이루어질 검수 과정에서 변경될 수 있음

<표 46> 문어 오류 주석 결과_분석 불능 여부: 숙달도 단계별

	1급	2급	3급	4급	5급	6급	계
분석 불가능	86	121	189	247	112	56	811

나. 국적별

<표 47> 문어 오류 주석 결과_분석 불능 여부: 국적별

	중국	미국	일본	베트남	카자흐스탄	기타	계
분석 불가능	544	54	38	35	27	113	811

② 오류 위치

가. 숙달도 단계별

<표 48> 문어 오류 주석 결과_오류 위치: 숙달도 단계별

	1급	2급	3급	4급	5급	6급	계
고유명사	152	66	32	29	5	23	307
일반명사	640	559	601	706	625	628	3,759
의존명사	77	106	143	133	87	80	626
대명사	64	99	113	62	58	35	431
수사	17	8	8	6	12	11	62
동사	232	326	342	366	282	227	1,775
형용사	131	115	108	80	65	52	551
보조용언	18	29	33	28	26	27	161
지정사	18	39	33	31	23	34	178
관형사	18	26	33	35	28	21	161
일반부사	186	193	138	143	99	98	857
접속부사	59	49	39	33	16	7	203
감탄사	1	0	0	0	1	0	2
체언접두사	0	1	3	4	0	6	14
명사파생접미사	13	11	12	28	22	25	111

	1급	2급	3급	4급	5급	6급	계
동사파생접미사	19	33	47	54	50	64	267
형용사파생접미사	9	6	7	9	11	10	52
어근	12	7	27	8	16	18	88
주격조사	310	358	368	390	281	228	1,935
관형사형전성어미	54	128	183	181	139	145	830
목적격조사	313	294	315	352	231	192	1,697
부사격조사	636	480	415	411	275	238	2,455
접속조사	128	64	44	40	27	19	322
보격조사	0	6	12	14	9	3	44
인용격조사	1	3	7	3	3	6	23
보조사	235	316	248	265	199	173	1,436
연결어미	182	335	398	349	287	242	1,793
종결어미	111	216	219	156	71	53	826
선어말어미	130	268	207	150	65	94	914
명사형전성어미	6	31	24	17	21	12	111
관형격조사	43	55	65	77	62	72	374
구 단위 표현	13	14	40	41	48	26	182
표현문형	121	219	253	289	181	132	1,195
계	3,949	4,460	4,517	4,490	3,325	3,001	23,742

나. 국적별

<표 49> 문어 오류 주석 결과_오류 위치: 국적별

	중국	일본	미국	대만	베트남	기타	계
고유명사	86	56	61	15	8	81	307
일반명사	1,694	661	256	124	106	918	3,759
의존명사	284	116	40	24	20	142	626
대명사	184	63	29	12	21	122	431
수사	26	8	5	0	1	22	62
동사	785	310	137	68	37	438	1,775

	중국	일본	미국	대만	베트남	기타	계
형용사	275	93	43	8	8	124	551
보조용언	74	30	11	5	7	34	161
지정사	73	37	9	8	5	46	178
관형사	72	24	16	4	1	44	161
일반부사	364	164	87	27	26	189	857
접속부사	87	33	13	7	10	53	203
감탄사	0	0	0	0	0	2	2
체인접두사	11	1	0	2	0	0	14
명사파생접미사	44	17	10	5	4	31	111
동사파생접미사	129	50	20	5	10	53	267
형용사파생접미사	20	8	4	2	1	17	52
어근	43	9	8	6	2	20	88
주격조사	1,057	165	150	67	58	438	1,935
관형사형전성어미	390	180	60	36	23	141	830
목적격조사	859	149	133	71	55	430	1,697
부사격조사	1,093	372	188	82	84	636	2,455
접속조사	135	49	23	15	11	89	322
보격조사	17	5	2	3	0	17	44
인용격조사	10	3	3	1	0	6	23
보조사	727	149	124	38	49	349	1,436
연결어미	808	344	107	51	54	429	1,793
종결어미	477	79	59	25	16	170	826
선어말어미	488	106	65	27	29	199	914
명사형전성어미	57	22	6	1	8	17	111
관형격조사	186	58	19	14	10	87	374
구 단 위 표 현	92	19	17	3	5	46	182
표현문형	581	173	89	24	38	290	1,195
계	11,228	3,553	1,794	780	707	5,680	23,742

③ 오류 층위

가. 숙달도 단계별

<표 50> 문어 오류 주석 결과_오류 층위: 숙달도 단계별

		1급	2급	3급	4급	5급	6급	계
발음	음절	0	0	0	0	0	1	1
형태	단어 형성[합성법]	1	1	5	0	3	1	11
	단어 형성[파생법]	4	12	15	12	14	13	70
	굴절[곡용]	80	46	27	30	23	18	224
	굴절[활용]	165	281	271	216	111	97	1,141
	품사	52	45	37	20	18	24	196
통사	높임	72	59	89	51	38	18	327
	시제	54	213	204	131	76	104	782
	사동	2	9	13	15	40	20	99
	피동	8	25	62	51	44	56	246
	부정	7	6	16	3	6	6	44
	어순	88	72	37	62	17	44	320
담화	지시	1	1	0	0	0	0	2
	접속	2	3	1	5	0	2	13
	구어/문어 오류	241	190	175	150	80	64	900
계		777	963	952	746	470	468	4,376

나. 국적별

<표 51> 문어 오류 주석 결과_오류 층위: 국적별

		중국	일본	미국	대만	베트남	기타	계
발음	음절	0	0	0	0	0	1	1
형태	단어 형성[합성법]	7	0	0	1	1	11	11

	단어 형성[파생법]	36	10	5	1	1	70	70
	굴절[곡용]	102	54	4	8	9	224	224
	굴절[활용]	614	163	75	40	20	1,141	1,141
	품사	103	27	12	6	5	196	196
통사	높임	157	35	27	3	16	327	327
	시제	426	126	48	29	13	782	782
	사동	39	13	5	4	7	99	99
	피동	94	70	23	6	7	246	246
	부정	13	12	3	3	3	44	44
	어순	122	55	38	6	13	320	320
담화	지시	0	1	0	0	0	2	2
	접속	4	5	0	0	1	13	13
	구어/문어 오류	362	157	63	45	40	900	900
계		2,079	729	303	152	136	4,376	4,376

④ 오류 양상

가. 숙달도 단계별

<표 52> 문어 오류 주석 결과_오류 양상: 숙달도 단계별

	1급	2급	3급	4급	5급	6급	계
누락	645	780	712	714	448	375	3,674
첨가	186	262	264	367	232	222	1,533
오 형태	1,527	1,418	1,311	1,247	900	827	7,230
대치	1,636	2,039	2,325	2,211	1,841	1,630	11,682
계	3,994	4,499	4,612	4,539	3,421	3,054	24,119

나. 국적별

<표 53> 문어 오류 주석 결과_오류 양상: 국적별

	중국	일본	미국	대만	베트남	기타	계
누락	1,936	306	308	147	93	884	3,674
첨가	659	227	137	39	54	417	1,533
오형태	3,366	1,206	537	265	179	1,677	7,230
대치	5,430	1,880	832	348	397	2,795	11,682
계	11,391	3,619	1,814	799	723	5,773	24,119

(2) 구어

① 분석 불가능 여부

가. 속달도 단계별

<표 54> 구어 오류 주석 결과_분석 불능 여부: 속달도 단계별

	1급	2급	3급	4급	5급	6급	계
분석 불가능	22	26	10	30	48	13	149

나. 국적별

<표 55> 구어 오류 주석 결과_분석 불능 여부: 국적별

	중국	일본	대만	베트남	미국	기타	계
분석 불가능	45	45	12	12	6	29	149

② 오류 위치

가. 숙달도 단계별

<표 56> 구어 오류 주석 결과_오류 위치: 숙달도 단계별

	1급	2급	3급	4급	5급	6급	계
고유명사	62	17	50	12	40	34	215
일반명사	307	298	278	311	542	530	2,266
의존명사	24	20	19	15	32	17	127
대명사	6	8	14	25	6	13	72
수사	26	11	2	3	36	12	90
동사	35	48	56	98	83	92	412
형용사	20	27	36	30	31	27	171
보조용언	5	4	6	2	5	6	28
지정사	2	1	2	6	8	4	23
관형사	15	25	11	22	22	28	123
일반부사	40	43	75	49	73	81	361
접속부사	26	9	12	11	15	16	89
감탄사	1	1	0	3	2	2	9
체인접두사	0	0	0	0	0	4	4
명사파생접미사	12	7	11	11	10	17	68
동사파생접미사	1	2	3	8	12	9	35
형용사파생접미사	0	0	1	2	3	2	8
어근	9	5	8	8	9	9	48
주격조사	32	22	49	44	43	52	242
관형사형전성어미	19	23	37	45	50	32	206
목적격조사	29	54	30	37	46	65	261
부사격조사	58	57	43	49	75	56	338
접속조사	13	7	5	3	4	3	35
보격조사	0	0	0	3	0	0	3
호격조사	0	1	0	0	0	0	1
인용격조사	0	0	0	2	0	0	2

	1급	2급	3급	4급	5급	6급	계
보조사	38	27	33	54	37	45	234
연결어미	25	32	79	71	44	71	322
종결어미	21	44	5	21	13	17	121
선어말어미	14	16	15	17	15	11	88
명사형전성어미	0	5	4	1	1	2	13
관형격조사	2	2	4	2	15	7	32
구 단위 표현	0	0	3	2	0	0	5
표현문형	11	12	13	21	16	39	112
계	853	828	904	988	1,288	1,303	6,164

나. 국적별

<표 57> 구어 오류 주석 결과_오류 위치: 국적별

	중국	일본	대만	베트남	러시아	기타	계
고유명사	65	54	21	24	4	47	215
일반명사	723	514	217	187	47	578	2,266
의존명사	24	34	11	23	2	33	127
대명사	29	12	6	4	2	19	72
수사	23	18	7	10	4	28	90
동사	112	74	44	35	16	131	412
형용사	48	30	19	18	7	49	171
보조용언	8	3	4	3	0	10	28
지정사	7	4	4	0	0	8	23
관형사	48	20	5	18	2	30	123
일반부사	106	71	43	44	7	90	361
접속부사	21	7	21	5	2	33	89
감탄사	1	0	1	4	0	3	9
체언접두사	2	1	1	0	0	0	4
명사파생접미사	22	12	7	7	2	18	68

	중국	일본	대만	베트남	러시아	기타	계
동사파생접미사	12	6	9	4	1	3	35
형용사파생접미사	3	1	1	2	1	0	8
어근	20	11	3	1	0	13	48
주격조사	74	23	36	25	5	79	242
관형사형전성어미	41	43	33	53	1	35	206
목적격조사	85	58	22	22	1	73	261
부사격조사	92	33	50	53	9	101	338
접속조사	9	3	12	1	1	9	35
보격조사	1	0	0	0	0	2	3
호격조사	0	0	0	1	0	0	1
인용격조사	1	0	0	1	0	0	2
보조사	78	28	34	42	2	50	234
연결어미	106	64	40	25	6	81	322
종결어미	36	33	9	13	5	25	121
선어말어미	21	7	12	18	2	28	88
명사형전성어미	5	1	0	1	3	3	13
관형격조사	8	4	6	3	0	11	32
구 단위 표현	0	0	0	0	0	5	5
표현문형	41	12	20	13	0	26	112
계	1,872	1,181	698	660	132	1,621	6,164

③ 오류 층위

가. 숙달도 단계별

<표 58> 구어 오류 주석 결과_오류 층위: 숙달도 단계별

		1급	2급	3급	4급	5급	6급	계
발음	음소	643	468	596	566	883	859	4,015
	음절	22	77	29	25	35	19	207
	음운규칙	19	13	4	11	12	23	82
	원어식 발음	4	10	5	3	14	4	40
형태	단어 형성 [합성법]	0	0	3	0	0	0	3
	단어 형성 [파생법]	1	0	0	3	2	0	6
	굴절[곡용]	5	8	4	7	12	3	39
	굴절[활용]	5	5	9	15	10	11	55
통사	품사	4	7	4	7	5	5	32
	높임	8	9	2	20	5	13	57
	시제	13	16	13	11	22	12	87
	사동	0	0	0	1	1	3	5
	피동	0	2	1	7	7	15	32
	부정	1	1	0	0	1	2	5
	어순	0	3	10	3	7	7	30
담화	지시	0	0	0	1	0	0	1
	접속	1	0	0	3	2	1	7
계		726	619	680	683	1,018	977	4,703

나. 국적별

<표 59> 구어 오류 주석 결과_오류 층위: 국적별

		중국	일본	베트남	대만	러시아	기타	계
발음	음소	1,247	874	416	436	99	1,033	4,015
	음절	38	79	16	6	1	70	207
	음운규칙	19	26	8	4	5	20	82
	원어식 발음	7	3	1	4	0	26	40
형태	단어 형성 [합성법]	1	0	0	0	2	0	3
	단어 형성 [파생법]	3	0	1	0	0	2	6
	굴절[곡용]	11	10	2	5	0	12	39
	굴절[활용]	23	8	5	3	1	15	55
통사	품사	9	0	7	5	0	12	32
	높임	30	1	10	10	0	6	57
	시제	11	9	28	13	1	25	87
	사동	4	0	0	0	0	1	5
	피동	3	13	2	3	0	11	32
	부정	2	2	0	0	0	1	5
	어순	5	1	6	5	4	9	30
담화	지시	0	0	0	0	0	1	1
	접속	0	0	4	1	0	2	7
계		1,413	1,026	506	495	113	1,246	4,703

④ 오류 양상
가. 숙달도 단계별

<표 60> 문어 오류 주석 결과_오류 양상: 숙달도 단계별

	1급	2급	3급	4급	5급	6급	계
누락	40	52	74	71	64	61	362
첨가	16	17	22	40	28	37	160
오 형태	32	49	44	54	47	58	284
대치	105	162	137	235	207	257	1,103
계	193	280	277	400	346	413	1,909

나. 국적별

<표 61> 문어 오류 주석 결과_오류 양상: 국적별

	중국	베트남	대만	일본	말레이시아	기타	계
누락	98	48	49	28	23	116	362
첨가	44	18	24	18	4	52	160
오 형태	90	35	34	48	6	71	284
대치	351	147	139	129	31	306	1,103
계	583	248	246	223	64	545	1,909

3. 온라인 구축 시스템 개발을 위한 협의 작업

3.1. 주요 진행 사항

- 2015년 기구축 학습자 말뭉치의 보완 작업에서는 말뭉치 구축 도구의 본격 사용으로 인해 구축 도구의 체제와 사용 환경에 맞춰 주식 체계 등을 수정·보완하여 반영하는 작업이 병행되었다. 이 과정에서 말뭉치 구축 도구의 버그 수정, 표본 등록에서 입력/전사, 형태 주식, 오류 주식의 각 단계의 지침을 반영한 기능 추가와 사용 환경 개선 작업이 지속적으로 이루어졌다. 이로 인해 실질적인 구축 작업 외에 시스템 개발 팀과의 협의와 시스템 환경 안정화 작업에 많은 시간과 노력이 소요되었다.

3.2. 남은 과제

- 2016년도 사업에서 학습자 말뭉치 구축에 필요한 구축 도구의 기능을 상세화하고 사용 환경을 안정화하였지만 보다 효율적인 말뭉치 구축을 위하여 다음과 같은 후속 과제를 남겨 두고 있다.
- 학문 목적 학습자 자료 구축을 위한 주식 체계 추가
- 시스템 내에서 구어 음성 파일을 동기화하여 작업하고 확인할 수 있는 기능
- 기구축 결과물에 기반한 형태 주식 학습 사전 업데이트
- 기구축 결과물에 기반한 오류 주식 결과 검색 및 참조 기능

V. 결론

1. 연구 요약

이 연구는 한국어 학습자 말뭉치 구축 중장기 사업 계획에 따른 제2차 연도 연구이다. 제1차 연도의 <2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업>에서 마련된 학습자 말뭉치 구축 지침을 보다 정교하게 보완하고 그 지침에 따라 2015년에 구축된 35만 어절 규모의 말뭉치를 수정·보완함과 동시에 약 80만 어절 규모의 말뭉치를 새롭게 구축하는 데에 주요한 목적이 있다. 이에 따른 주요 과업과 연구의 성과는 다음과 같다.

○ 한국어 학습자 말뭉치 연구

한국어 학습자 말뭉치 연구는 체계적인 말뭉치 구축을 위한 구축 지침의 정교화와 말뭉치 활용을 위한 모형 제시를 위한 기초 연구의 성격을 지닌다. 1차 연도에 마련된 학습자 말뭉치 구축 지침을 토대로 하여 실제적인 자료를 구축하고 가공해 나가는 과정에서 대두되는 세부적인 쟁점들을 반영하여 수집과 자료 처리, 문어 입력, 구어 전사, 형태 주석, 오류 주석의 각 단계별 지침을 정교화하였다. 또한 이러한 지침에 따라 구축된 말뭉치가 효율적으로 활용될 수 있도록 연구자, 교사, 학습자를 위한 활용 모형을 제시하였다. 사용자 집단별 활용 방안은 향후 웹 사이트를 통해 배포될 검색 기능과 결과를 활용한 실제적인 모형으로 학습자의 국적, 숙달도 단계, 학습 목적 등의 학습자 변인이나 자료 변인에 따른 언어 사용 양상이나 중간언어, 언어 발달 연구에의 활용, 이러한 정보를 기반으로 한 숙달도별, 언어권별 맞춤형 교수법 개발, 학습자 사전 편찬, 자기 주도적 학습을 위한 자동 첨삭 시스템 개발에의 활용 모형으로 구체화되었다.

○ 한국어 학습자 말뭉치 관련 교육 및 홍보

한국어 학습자 말뭉치 관련 교육 및 홍보에서는 체계적인 말뭉치 구축을 위해 필요한 수집 자료의 처리, 문어 입력, 구어 전사, 형태 주석, 오류 주석의 각 단계별 지침 교육과 도구 사용 교육이 이루어졌다. 아울러 주례회의 및 월례회의, 온라인 매체를 활용한 정기/비정기 워크숍을 통해 작업자 간의 일관성과 통일성을 높임으로써 구축 자료의 신뢰성을 높이려고 하였다. 한편

한국어 학습자 말뭉치 사업을 널리 알리고 사용자들의 활용 능력 제고하기 위하여 여름과 가을 두 차례에 걸쳐 한국어 교육 연구자와 교사를 대상으로 한 워크숍을 개최하였다. 제1회 워크숍은 홍보에 중점을 두어 한국어 학습자 말뭉치 사업 소개와 실제 구축 도구를 사용한 입력 및 전사, 형태 주석, 오류 주석 작업 시연을 하였고, 제2회 워크숍에서는 예비 사용자를 위한 교육에 중점을 두어 워크숍 참가자들이 전사와 오류 주석을 해 보고 웹 기반의 검색 도구를 사용하여 구축된 자료를 검색해 보도록 하였다. 이를 통해 한국어 학습자 말뭉치에 대한 홍보와 교육은 물론 한국어 학습자 말뭉치에 대한 사용자의 요구를 확인하고 의견을 수렴할 수 있었다.

○ 한국어 학습자 말뭉치 수집 및 가공

한국어 학습자 말뭉치 수집은 제1차 연도에 수집을 시작한 국내 교육 기관 학습자 자료의 집중 수집과 함께 이주민 자료의 시험 수집이 이루어졌다. 이주민 자료의 시험 수집은 2017년 제3차 연도의 수집 계획을 수립하기 위한 것으로 결혼이주여성, 이주노동자, 중도입국청소년의 자료를 수집해 보고, 2017년부터 시작 예정인 이주민 자료의 집중 수집을 앞두고 수집 과정에서 발생하는 다양한 쟁점들을 파악하는 데에 주력하였다. 그 결과 2017년과 2018년, 제3차 연도와 제4차 연도의 이주민 자료 집중 수집 시에 수집 대상의 분포를 고려한 수집 자료의 비율 조정, 학습자의 실제 언어 사용 환경을 고려한 구어 자료 수집 비중의 확대, 여성가족부와 법무부, 교육부와와의 공조 체제를 통한 기관 중심의 자료 수집 네트워크 마련 등이 필요함을 제안하였다.

한편, 한국어 학습자 말뭉치 가공은 2015년 제1차 연도에 구축한 원시 말뭉치 약 35만 어절(문어 30만 어절, 구어 5만 어절), 형태 주석 말뭉치 약 22만 어절(문어 20만 어절, 구어 2만 어절), 오류 주석 말뭉치 약 5만 어절(문어 4만 어절, 구어 1만 어절) 규모의 말뭉치를 새로운 지침에 따라 수정·보완하는 작업이 선행되었다. 그리고 2016년에는 원시 말뭉치 약 80만 어절(문어 70만 어절, 구어 10만 어절), 형태 주석 말뭉치 약 55만(문어 50만, 구어 5만 어절), 오류 주석 약 15만 어절(문어 10만, 구어 5만) 규모의 말뭉치가 새롭게 구축되었다. 그 결과 2015년, 2016년에 구축한 전체 말뭉치의 규모는 원시 말뭉치 약 115만 어절(문어 100만, 구어 15만 어절), 형태 주석 말뭉치 약 77만 어절(문어 70만 어절, 구어 7만 어절), 오류 주석 말뭉치 약 20만 어절(문어 14만 어절, 구어 6만 어절) 규모의 말뭉치가 구축되었다.

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 2015년과 2016년 두 해 동안 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 말뭉치는 개인 연구자가 접근하기 어려운 자료를 대규모로 구축하여 제공함으로써 한국어 교육 관련 연구를 활성화하고 말뭉치 연구에의 요구가 큼에도 불구하고 사용 기반이 마련되지 않아 제한적으로 활용되었던 말뭉치 연구의 확산이 기대된다. 이는 한국어교육 연구자 간의 학술 교류, 더 나아가 세계 규모의 학술대회 대회 참여 등을 통한 국제 수준의 학술 교류를 활성화에 기여할 수 있다. 아울러 그간 개인 연구자에 의해 구축된 소규모 자료를 기반으로 한 연구에서 담보하기 어려웠던 대표성과 균형성의 문제를 해결함으로써 연구 결과를 실제적인 교수·학습에 보다 효율적으로 적용할 수

있게 될 것이다.

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

기존의 언어 교수법은 교수자의 경험과 직관에 상당 부분을 의존함으로써 언어 교육의 효율성 및 교수법의 적합성을 평가하기 어려웠다. 반면, 학습자 말뭉치를 활용한 연구는 외국어로서 혹은 제2 언어로서의 한국어 학습의 측면에서 한국어 사용에 대한 실질적이며 구체적인 설명과 예시를 제시할 수 있기 때문에 이러한 문제를 체계적으로 보완할 수 있게 된다. 학습자의 모국어, 국적, 숙달도 단계 등의 변인에 따른 언어 사용 양상을 분석하여 점점 다양해져 가는 학습자의 요구에 부응한 과학적인 교수 환경을 제공할 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, 학습기간, 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

본 연구는 한국어 학습자 말뭉치 구축을 위한 연구와 실제 구축 작업 수행을 핵심 과업으로 하였으며, 본 보고서에는 그러한 과정과 결과, 그와 관련된 쟁점들이 체계적으로 기술되어 있다. 한국어 학습자 말뭉치 구축을 위한 연구의 성과는 한국어 교육 연구자를 포함한 예비 사용자들에게 한국어 학습자 말뭉치 구축의 방법과 절차, 각 단계에서의 지침 수립과 관련한 쟁점을 통해 한국어 학습자 말뭉치 구축에 관한 이론과 실제 구축을 위한 가이드라인을 제공해 준다. 또한 학습자 말뭉치의 효율적인 활용을 위한 방안들을 제시해 준다. 그 외에도 실제 구축이나 시험 수집 결과 등은 중장기 사업으로 계획되어 있는 학습자 말뭉치 구축 사업을 계획하고 수행해 나아갈 해당 부처와 차기 연구진들에게 제3차 연도 사업을 구체화하기 위한 가이드라인을 제공해 준다. 본 보고서는 구체적으로 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 발화 자료를 구축 대상으로 하는 만큼 수집, 구축, 가공의 전 단계에서 고려해야 할 사항과 쟁점이 매우 많다. 그러나 일반적인 말뭉치 구축 이론을 소개하는 논저는 비교적 많은 반면, 학습자 말뭉치 구축에 관한 실제적인 내용을 다룬 논의는 그리 많지 않고 그간 한국어 학습자 말뭉치 구축이 활발하게 이루어지지 못하였기 때문에 구축 방법에 관한 이론이 체계화되지 못하였다. 본 보고서에는 한국어 학습자 말뭉치 구축의 쟁점과 그것을 풀어가는 과정이 체계적으로 기술되어 있는데, 이는 한국어 학습자 말뭉치 구축에 관한 이론적 지침이 될 수 있다.

○ 한국어 학습자 말뭉치 구축의 실제를 위한 실용적 지침

본 보고서에는 한국어 학습자 말뭉치의 수집, 자료 처리, 입력과 전사, 형태 주석, 오류 주석의 방법과 절차가 기록되어 있다. 본 연구의 주요한 특징 중 하나는 학습자 말뭉치 구축을 위한 표본 등록에서 입력/전사, 형태 주석, 오류 주석의 전 과정을 통합하여 작업할 수 있는 말뭉치 구축 도구를 활용하였다는 것이다. 본 연구에서 사용한 말뭉치 구축 도구는 작업자의 편의성이나 말뭉치 구축과 가공, 자료 관리의 효율성을 고려한 것으로 선진화되어 있다.

본 보고서에서는 한국어 학습자 말뭉치 구축에 관한 이론적인 지침 외에도 말뭉치 구축 도구를 활용한 실제적인 구축 방법과 절차가 충실하게 기술되어 있다. 한국어 학습자 말뭉치 구축 도구는 향후 구축 결과물과 함께 애플리케이션으로 개발하여 배포하여 연구자들이 학습자 말뭉치를 폭넓게 활용할 수 있는 기반을 제공할 예정이다. 본 보고서에 기술된 실제적인 구축 방법과 절차는 향후 학습자 말뭉치를 구축하고자 하는 기관이나 연구자들에게 실용적인 지침으로 활용될 수 있다.

4. 정책 제언

○ 지속적인 자료의 업데이트 및 다양화를 위한 온라인 수집 체계 마련

첫째, 지속적인 온라인 수집 시스템의 구축은 수집 자료의 다양성과 양적 규모를 추가적으로 확보할 수 있다. 본 연구는 2015년에서 2020년까지 6년간의 중장기 계획에 따라 수행되며, 3단계 6년에 걸쳐 국내, 이주민, 국외 학습자의 자료를 집중 수집하여 구축하도록 설계되어 있다. 이러한 계획에 따르면 중장기 계획이 끝나는 2020년 이후에는 더 이상의 자료 수집이 이루어지지 않는다. 그러나 한국어 교육 환경은 다양한 국적, 학습 목적 등의 학습자 변인에 따라 시시각각 변한다. 그리고 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 이루어질 수 있는 환경을 마련할 필요가 있다. 뿐만 아니라 현재 수집되는 자료는 주로 국내 교육 기관 학습자 자료의 확보에는 용이하지만 그 밖의 환경에서 한국어를 배우고 사용하는 학습자들의 언어 특성을 관찰하기 어렵다는 한계가 있다. 따라서 장기적으로 자료의 다양성 측면에서 다양한 환경에서 공부하는 학습자들의 자료를 폭넓게 수집하여 구축할 필요가 있다.

둘째, 온라인 수집 방법은 자료 제공자인 학습자의 동의를 전제로 하므로, IRB 규정에 따른 자료 수집의 제한점을 보다 쉽게 해결할 수 있는 장점이 있다. 시스템의 실제 운영을 위해서는 시스템 개발과 함께 학습자들의 자료 제공을 독려하기 위한 보상 방안이 전제되어야 한다. 즉, 학습자들은 제공한 자료(말하기, 작문)에 대한 한국어 교육 전문가의 피드백을 받는 것을 전제로

자료를 제공할 가능성이 높으므로, 학습자 자료에 대한 피드백 체제를 갖추기 위해서는 전문 인력의 확보와 이를 운영할 시스템 개발과 유지, 운영을 위한 예산 배정이 수반되어야 한다.

○ 이주민 자료의 효율적 수집을 위한 부처 간의 공조 체제 마련

한국어 학습자 말뭉치 구축 중장기 계획에 따르면 2017년은 이주민 자료의 집중 수집과 구축이 시작되는 시기이다. 본 연구에서는 이에 앞서 실제적인 수집과 관련한 쟁점의 파악을 위해 결혼이주여성, 이주노동자, 중도입국청소년을 대상으로 한 시험 수집을 하였다. 수집 경로로는 결혼이주여성의 한국어 교육을 담당하는 여성가족부 건강가정지원센터와 법무부 사회통합 프로그램을 운영하는 다문화센터, 이주노동자들의 한국어 교육을 담당하는 이주노동자 지원센터와 지역 노동조합, 중도입국청소년의 한국어 교육을 담당하는 초등학교와 대안학교, 그리고 자원 봉사 등의 형태로 활용하는 개인 교사 인력풀이 활용되었다. 두 가지 경로를 통해 자료를 수집하는 과정에서 정부 부처에서 개발한 표준 교재를 활용하는 기관에 소속하지 않은 학습자의 경우 특정한 교육과정에 종속되지 않아 자연적인 언어 습득 양상을 살필 수 있는 장점이 있지만 자료 활용의 측면에서 가장 중요한 변인 중 하나인 숙달도 단계를 측정하기 어렵다는 문제점이 발견되었다. 이에 따라 자료 수집은 표준 교재를 통해 학습자의 숙달도 단계를 등급화할 수 있는 각 지역의 다문화센터와 각 급의 공교육 기관을 통한 대규모 수집이 바람직하다는 잠정적 결론을 도출해 낼 수 있었다. 그러나 제1단계(2015-2016년) 연구에서 자료 수집을 위해 학계의 공통적인 관심과 요구에 의한 공조 체제를 이끌어낼 수 있었던 것과 달리, 이들 기관의 경우 자료 수집의 필요성에 대한 공감을 이끌어 내기가 쉽지 않을 뿐만 아니라 자료 수집에 수반되는 보상에 대한 요구가 매우 명시적이고 구체적인 경향이 있다. 따라서 실제적인 수집 계획을 수립하는 과정에서 유관 기관인 여성가족부(건강가족지원센터 다문화센터), 법무부(사회통합 프로그램 운영 기관), 교육부(각 급의 학교)와의 공조 체제를 마련하는 것이 필수적이다.

○ 학습자 말뭉치의 확산과 효율적 활용을 위한 통합형 배포 시스템 구축

본 연구는 한국어 교육 연구와 교수·학습에서 한국어 학습자 말뭉치가 실용적으로 활용되도록 하는 데에 궁극적인 목적이 있다. 그에 따라서 구축 도구 지원과 함께 구축 설계를 반영한 배포 시스템 개발을 병행하는 것은 합리

적인 판단이라고 하겠다. 학습자 말뭉치 활용의 차원에서 효용성을 극대화하기 위해서는 사용자 집단의 속성에 따라 기능이 차별화되어야 한다. 현재 부분적인 공개를 앞두고 시험 운영 중인 배포 시스템에서는 다양한 조건에 따른 자료의 검색과 통계 정보를 제공하는 기능이 잘 구현되어 있다. 여기에 사용자의 사용 목적과 요구를 고려하여 다음과 같은 기능을 추가함으로써 활용도를 높일 수 있을 것이다.

- **구어 음성 파일의 동기화:** 구어 자료의 경우 음성 녹음 파일이 원본이며 이것을 듣고 전사한 후 일정한 체계에 따라 주석을 붙여 음성적 특성을 표시한 텍스트 파일을 구축하였다. 음성 자료의 경우 발화에 포함된 어휘나 문법, 담화적 특성 외에 학습자의 발음이나 억양, 휴지 등의 음운 정보가 주요한 연구, 교수·학습의 대상이 될 수 있다. 따라서 필요한 경우 사용자가 전사 단위에 따라 분절된 부분의 음성을 들을 수 있도록 음성 파일이 동기화되어 자료로 제공될 필요가 있다.
- **말뭉치의 확장 구축과 추가 주석을 위한 도구의 제공:** 다양한 검색 기능과 통계 정보의 제공 외에 연구자에게는 추가 구축 또는 추가 주석 시 활용 가능한 구축 도구와 사용 지침을 제공함으로써 연구자들의 활용 범위를 넓힐 수 있다.
- **CALL 기반의 교수·학습 활동 기능의 통합:** 교사나 학습자의 경우 말뭉치를 활용하여 교수·학습 활동으로 연계할 수 있는 컴퓨터 보조 언어 학습 프로그램이 함께 제공될 필요가 있다. 형태 주석이나 오류 주석 정보를 활용하여 학습자의 자기 주도적 학습을 도울 수 있는 첨삭 프로그램 등이 부수적으로 개발된다면 교수·학습의 측면에서 학습자 말뭉치를 매우 효율적으로 활용할 수 있을 것이다.

○ 한국어 교육 학계와 사용자들의 요구를 반영하기 위한 지속적 말뭉치 구축, 이를 위한 행정적 지원과 현실적인 예산 편성

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 매우 크다. 이는 본 사업에서 개최한 두 차례의 워크숍을 통해서도 확인된 바 있다.

말뭉치 구축 작업은 노동집약적인 작업이기 때문에 많은 인력과 시간, 비용과 노력이 소요된다. 학습자 말뭉치의 경우는 한국어를 비모어로 하는 화자들의 발화 산출 자료를 기반으로 한 특수 말뭉치로 그것이 배가된다. 제1단계(2015-2016년) 사업의 경우 약 110만 어절 규모의 자료를 시험 구축하는데에 수집 단계에서 1,000여 명, 구축과 가공 단계에서 50여 명의 인력이 투입되었다. 제1단계의 수집 대상은 국내 교육 기관 학습자의 자료로 수집과 구축, 가공의 전 단계에서 학습자 말뭉치에 대한 학계의 요구와 염원에 의한 공조로 많은 시간과 노력이 소요되는 작업을 적은 비용으로 수행할 수 있었다. 그러나 장기적으로 그러한 이해관계가 부족한 이주민 자료와 국외 학습자 자료의 수집과 양질의 자료를 구축하기 위해서는 구축의 각 단계에서 필요한 현실적인 인력, 시간, 비용이 뒷받침되거나 합리적으로 목표 구축 규모를 조정할 필요가 있다.

참고 문헌

- 강현화 외(2010), 한국어 학습자 말뭉치 구축 설계, 국립국어원.
- 서상규 외(2015), 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업, 국립국어원.
- 강현화(2010) 한국어 학습자 사전 표제어 선정에 관한 자료 구축 및 선정 방법에 관한 연구, 한국사전학 16, 한국사전학회.
- 강현화(2011) 한국어 학습자 말뭉치의 자료 구축 방안 대한 기초 연구, 한국사전학 17, 한국사전학회.
- 강현화·조민정(2003), 스페인어권 한국어 학습자의 어미, 조사 및 시상, 사동 범주의 오류 분석, 한국어교육 14(2), 국제한국어교육학회.
- 고석주(2002), 학습자 말뭉치에서 조사 오류의 특징, 외국어로서의 한국어교육 27(1), 연세대학교 한국어학당.
- 고석주(2004), 오류 유형 주석을 위한 기초 연구, 한국 문화사.
- 고승연(2013), 아랍어권 한국어 학습자의 발음 오류 분석, 한국어문화교육 7(1), 한국어문화교육학회.
- 곽수진(2010), 한국어 학습자의 문장 성분 호응 관계 오류 연구, 한국어 의미학 32, 한국어의미학회.
- 권기양(2006), KFL 학습자의 오류에 대하여: 중국인 학습자 중심으로, 언어과학 13(3), 한국언어과학회.
- 김경화(2013), 고급단계 한국어학습자의 오류연구, 중국조선어문 188, 길림성민족사무위원회.
- 김미옥(2002), 학습 단계에 따른 한국어 학습자 오류의 통계적 분석, 외국어로서의 한국어 교육 27(1), 연세대학교 한국어학당.
- 김미옥(2003), 한국어 학습자의 단계별 언어권별 어휘 오류의 통계적 분석, 한국어 교육 14(3), 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, 새국어교육 100, 한국국어교육학회.
- 김유미(2002), 학습자 말뭉치를 이용한 한국어 학습자 오류 분석 연구, 외국어로서의 한국어교육 27, 연세대학교 한국어학당.
- 김유미(2006), 학문 목적 한국어 학습자를 위한 문어 학술 말뭉치 구축 -구어 말뭉치 자료 수집과 전사에 대하여-, 한국언어문화교육학회 학술대회, 한국응용언어학회.
- 김유정(2005), 한국어 학습자 말뭉치 오류 분석의 기준, 한국어 교육 16(1), 국제한국

어교육학회.

김일환(2016), 한국어 학습자 말뭉치의 주석 과정과 활용 방법, 국제한국어교육학회
춘계학술발표논문집, 국제한국어교육학회.

김정숙(2002), 영어권 한국어 학습자의 조사 사용 오류 분석과 교육 방법, 한국어교육
13(1), 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-, 이중언어학 21, 이중언어학회.

김정은(2003), 한국어교육에서의 중간언어와 오류 분석, 한국어 교육 14(1), 국제한국
어교육학회.

김지민, 신승용(2010), 어휘오류 분석의 문제점과 어휘오류 처리 방안 연구, 언어와 문
화 6(2), 한국언어문화교육학회.

김지영(2014), 중국인 유학생의 한국어 사용 오류 분석, 시학과 언어학 26, 시학과언어
학회.

김한샘·곽용진(2016), 차세대 학습자 말뭉치 통합 관리 시스템 개발, 한국언어문화교
육학회 학술대회 발표 자료집, 한국언어문화교육학회.

남길임(2007), 학습자 오류 말뭉치를 활용한 한국어 용법 사전의 편찬, 한말연구회.

남윤주 외(2014), L2로서의 한국어 자연발화 코퍼스의 구축과 활용, 통일인문학논총.

노미연(2012), 한국어 학습자의 구어 오류와 후속 상호작용 분석 연구, 동국대학교 박
사학위논문.

민영란(2008), 부정적 전이로 인한 중국어권 학습자의 오류 분석, 한국어 교육 19(1),
국제한국어교육학회.

박수연(2007), 한국어 학습자 오류 말뭉치 구축과 그 문제점에 관한 연구, 언어 사실
과 관점 17, 연세대학교 언어정보연구원.

서상규, 유현경, 남윤진(2002), 한국어 학습자 말뭉치와 한국어교육, 한국어교육 13(1),
국제한국어교육학회.

신성철(2002), 호주 한국어 학습자의 어휘 오류 분석 연구, 한국어 교육 13(1), 국제한
국어교육학회.

신성철(2007), 영어권 한국어 학습자의 철자 오류 유형과 패턴, 한국어 교육 18(3), 국
제한국어교육학회.

유석훈(2001), 외국어로서의 한국어 학습자 말뭉치 구축의 필요성과 자료 분석, 한국
어교육 12(1), 국제한국어교육학회.

이동은(2007), 한국어 학습자의 철자 오류와 개선 방안 -북미지역 청소년 교포 학습
자를 대상으로-, 한국어학 35, 한국어학회.

이병운(2011), 베트남인 학습자의 작문 오류 경향 분석: 조사·어미를 중심으로, 우리말
글 52, 우리말글학회.

- 이승연(2006), 한국어 학습자 말뭉치 오류 표지 방안 제고, 이중언어학 31, 이중언어학회.
- 이승연(2007), 한국어 학습자 오류 판정 및 수정 기준 연구-교사, 비교사 집단간 오류 판별 비교 실험을 바탕으로, 이중언어학 33, 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, 연세대학교 언어연구교육원 한국어학당.
- 이정희(2002), 한국어 오류 판정과 분류 방법에 관한 연구, 한국어교육 13(1), 국제한국어교육학회.
- 이정희(2003), 초급 단계 한국어 학습자의 어휘 오류, 이중언어학 22, 이중언어학회.
- 이정희(2009), 중국어권 한국어 학습자의 어휘 오류 연구, 한국어 교육 19(3), 1-23쪽, 국제한국어교육학회.
- 이화진·이지연(2016), 학습자 말뭉치 구축과 음성 인식 활용, 한국언어문화교육학회 학술대회 발표 자료집, 한국언어문화교육학회.
- 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, 한말연구학회.
- 조철현 외(2002), 한국어 학습자의 오류 유형 조사 연구, 문화관광부.
- 최원평, 유효려(2010), 중국 대학생 글쓰기에 나타난 어휘 오류 연구, 언어와 문화 6(3), 한국언어문화교육학회.
- 한상미(2014), 중급 한국어 학습자의 구어 담화에 나타난 조사 오류 연구, 한국어교육 25(3), 국제한국어교육학회.
- 한송화(2001), 말뭉치와 학습자 오류를 이용한, 외국인 학습자를 위한 한국어 어휘 사전의 의미 기술, 한국어정보학 4, 한국어정보학회.
- 한송화·강현화(2016), 학습자 말뭉치에서의 구어 전사와 오류 주석의 쟁점과 실제, 한국언어문화교육학회 학술대회 발표 자료집, 한국언어문화교육학회.
- 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, pp. 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. pp. 144-154.
- Pica, T., Holliday, L., Lewis, L. and Morgenthaler, L. (1989) Comprhensible 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.

부록. MLAT-기반 한국어 적성 평가지

■ 적성 평가 문항에서의 유형

1. Part I, Number Learning: 수의 조합에 대한 청각 테스트(3분)

1 유형

이제부터 여러분은 숫자 조합에 대한 문제를 풀 것입니다.

모두 10 문항입니다. 들은 숫자를 쓰세요. 단, 들은 후 10초 안에 쓰셔야 합니다.

1 문항

보기					
*	2	이	*	50	오십
1	10	(script) 십	2	24	(script) 이십사
3	62	(script) 육십이	4	93	(script) 구십삼
5	101	(script) 백일	6	212	(script) 이백십이
7	432	(script) 사백삼백이	8	691	(script) 육백구십일
9	1030	천삼십	10	12140	만이천백사십

2. Part II, Phonetic Script¹³⁾ (소리나 음성기호의 대응)

2-1 유형(4분)

먼저 16개의 단어를 들으세요.

다음에 하나 씩 불러주는 단어를 듣고 앞에서 들은 단어면 O 아니면 X 표 하세요.

초급 단어(8개): 학교, 나라, 노래, 가족, 취미, 바다, 날씨, 약속,

중급 단어(8개): 외모, 후회, 능력, 발전, 여유, 취소, 수술, 절약

2-1 문항

1	바다	(O X)	2	노력	(O X)
3	날씨	(O X)	4	사진	(O X)
5	책임	(O X)	6	수술	(O X)
7	약속)	(O X)	8	얼굴	(O X)

2-2 유형(2분)

다음을 듣고 맞는 단어를 고르세요.

< 5. 모음 변별 6. 받침 변별 7. 경, 평, 격음 변별 8. 자음 변별 >

2-2 문항

5	거수	① 가수 ② 고수 ③ 거수	6	밤	① 밤 ② 방 ③ 밥
7	탁	① 탁 ② 닻 ③ 딱	8	자주	① 사수 ② 자주 ③ 차조

3. Part III, Spelling Clues(3분)

: 표준 표기 규약이 아닌 발음 위주의 표기 읽고, 의미가 가장 가까운 숨겨진 단어 찾기

3-1 유형 (주제: 집안일)

지금 우리는 집안일과 관련된 단어에 대해 말할 것입니다.

이 단어들은 무작위로 섞여 있는 철자들로 되어 있습니다. 그 단어가 어떤 것인지 알려 주기 위해 단서가 제공됩니다. 제공되는 단서를 보고 철자를 바르게 재배열하여 단어를 완성해서 쓰세요.

3-1 문항

	Spelling	Clues	Answer
1	뵤ㄹ / ㅏ ㅏ / ㄹ	(script) 옷이 더러워졌을 때 이것을 합니다.	빨래
2	ㅏ ㅏ / ㅏ ㅏ / ㅇ	(script) 손님이 오기 전에 집을 깨끗하게 하는 것을 말합니다.	청소
3	ㄱㅏㅏ / ㅏ ㅏ / ㄹ	(script) 음식을 먹고 나서 그릇이나 접시를 씻 는 것입니다.	설거지
4	ㄱㅏㅏ / ㅏ ㅏ ㅏ / ㅇ	(script) 음식이 상하지 않도록 차갑게 보관하기 위해서 여기에 넣어 둡니다.	냉장고

4. Part IV, Words in Sentences¹⁴⁾ : 문법 구조에 대한 지식(8분)

4-1 유형

다음 문장을 읽고 알맞은 단어와 문법을 이용해서 문장을 완성하세요.

4-1 문항

	문 항	평가 요소	급
1	저는 남자가 아닙니다. 여자입니다. 저는 한국 사람이 아닙니다. _____ 입니다.	어휘의 성격, 품사 파악하기	1급
2	저는 한국 노래를 잘 몰라요. (모르다) 저는 시간이 있을 때마다 고향 노래를 _____. (부르다)	불규칙	1급
3	할머니께서 신문을 읽으세요. 할아버지 _____ (산책하다)	높임 표현	1급

4	작년에 은행에서 돈을 빌려서 차를 샀어요. 내년에는 돈을 열심히 모아서 집을 _____ <으/ㄹ 거예요>	시제	1급
5	어제는 날씨가 _____ <은/ㄴ/는> 데다가 눈도 많이 왔어요. (출다)	시제/ 형태 교체	2급
6	가: 주말에 함께 등산할까요? 나: 아니요. 좀 피곤해서 등산 안 하려고요. 가: 다음 달에 같이 김장할래요? 나: 아니요. 김치가 아직 많이 남아서 _____ _____.	부정	3급
7	가: 지금 뭐 하세요? 나: (읽히고/ 책을/ 저는/ 있어요/ 아이에게)	어순	4급
8	휴대전화가 고장 나서 전화가 잘 안 걸려요. 창문이 고장 나서 잘 안 _____. (열다)	피동	4급

5. Part V, Paired Associates¹⁵⁾: 주어진 언어의 어휘 목록을 익혀 의미 암기(5분)

5-1 유형

이제부터 한국어 단어 12개가 제시될 것입니다. 이 단어들은 모국어 의미도 함께
연결되어 제시됩니다. 여러분은 2분 동안 다음의 단어들을 외울 수 있습니다. 다
음에 2분 동안 단어 의미를 모국어로 쓰세요.

합격 치료 신문 명절
행복하다 실망스럽다 실용적이다 긍정적이다
허락하다 구경하다 처리하다 회복하다

5-1 문항

1	합격		2	명절	
3	치료		4	신문	
5	행복하다		6	긍정적이다	
7	실용적이다		8	실망스럽다	
9	회복하다		10	허락하다	
11	처리하다		12	구경하다	

-
- 13) LLAMA의 음성 분석 능력에서는 20개의 구어 단어를 들은 후 다시 한 개씩 들고 들은 단어인지를 구별하는 문항으로 되어 있음. 이에 들은 단어 변별하는 문항과 한국어의 자음과 모음의 음운을 변별하는 문항으로 구성하였음.
- 14) 한국어 문장에서 어휘나 언어 구조의 문법적 기능을 인식하는 능력. 어휘(품사 혹은 어휘 의미), 문장 구조, 조사, 시제, 부정, 높임, 형태 교체 등 각 문항 별로 학습할 내용을 선정, 규칙을 설명하고 유사한 문제로 측정함.
- 15) 기존의 문항들은 인공 언어를 만들어서 대상과 연결시켜 학습시킨 후 대상과 인공언어를 연결하는 문항으로 구성되어 있음. (2분 동안, 20개) 초, 중급 학습자를 대상으로 하는 문제이니 고급의 단어를 주는 것이 암기 능력을 평가하는 데 유용할 것 같아서 고급 단어를 제시함. 제시 단어와 모국어 연결시켜 2분 동안 학습시키고 해당 어휘에 모국어 의미를 쓰도록 하는 문항.

연구 책임자: 강현화(연세대학교 국어국문학과 교수)
 공동 연구원: 김선정(계명대학교 한국문화정보학과 교수)
 김일환(고려대학교 민족문화연구원 HK연구교수)
 김정숙(고려대학교 국어국문학과 교수)
 김한샘(연세대학교 언어정보연구원 HK교수)
 안경화(서울대학교 언어교육원 대우 부교수)
 이동은(국민대학교 국어국문학과 부교수)
 이정희(경희대학교 국제교육원 부교수)
 한송화(연세대학교 언어정보연구원 HK교수)
 황용주(국립국어원 특수언어진흥과 학예연구관)
 김수현(국립국어원 한국어진흥과 학예연구사)
 연구 보조원: 홍혜란(연세대학교 언어정보연구원 전문연구원)
 공나형(연세대학교 대학원 박사과정)
 김보영(연세대학교 대학원 박사과정)
 김선영(연세대학교 대학원 박사과정)
 유소영(연세대학교 대학원 박사과정)
 윤종원(연세대학교 대학원 박사과정)
 허희정(연세대학교 대학원 박사과정)
 하법정(연세대학교 대학원 석사과정)
 담당 연구원: 김수현(국립국어원 한국어진흥과 학예연구사)

2016년 한국어 학습자 말뭉치 연구 및 구축 사업

발행인	송철의
발행처	국립국어원 서울시 강서구 금낭화길 148(방화 3동 827) 전화: 02-2669-9743~4 전송: 02-2669-9727
인쇄일	2016년 12월 16일
발행일	2016년 12월 16일
인쇄	학위사

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