

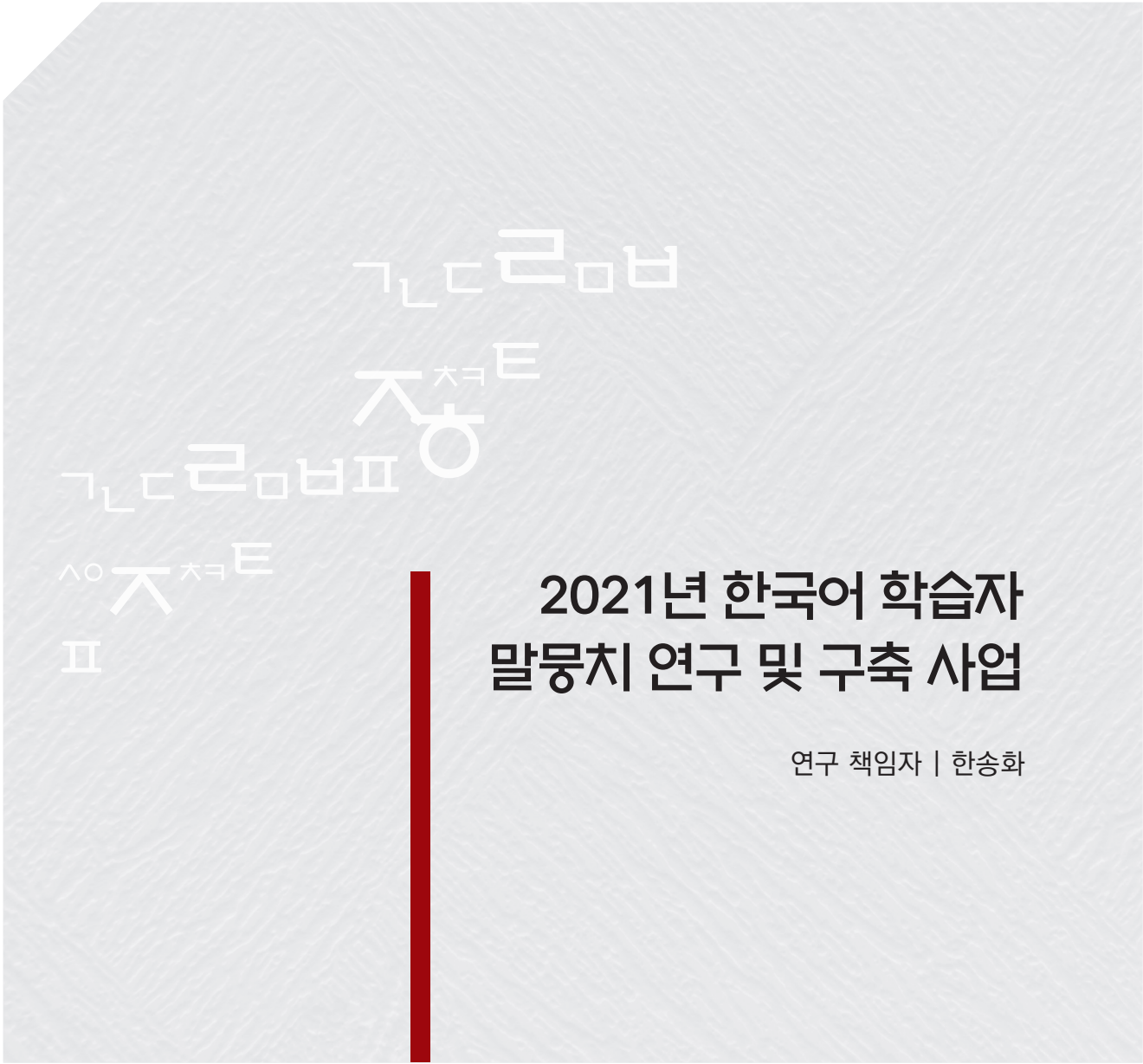
국립국어원 2021-01-17

2021년 한국어 학습자 말뭉치 연구 및 구축 사업

국립국어원



NATIONAL INSTITUTE OF KOREAN LANGUAGE



2021년 한국어 학습자 말뭉치 연구 및 구축 사업

연구 책임자 | 한송화



발 간 등 록 번 호

11-1371028-000746-10

2021년 한국어 학습자 말뭉치 연구 및 구축 사업

연구 책임자: 한 승 화



국립국어원

제 출 문

국립국어원장 귀하

“2021년 한국어 학습자 말뭉치 연구 및 구축 사업”에 관하여 위원과
체결한 연구용역 계약에 의하여 연구보고서를 작성하여 제출합니다.

2021년 12월 10일

연구 책임자: 한송화(연세대학교)

연구 기관	연세대학교 산학협력단
연구 책임자	한송화(연세대)
공동 연구원	김선정(계명대), 김인철(상명대), 김일환(성신여대), 김한샘(연세대), 장석배(미국 밴더빌트대), 홍혜란(연세대), 박미영(국립국어원), 조은(국립국어원)
연구 보조원	김동은(연세대), 김미선(연세대), 김선영(연세대), 송지혜(연세대), 유소영(연세대), 허희정(연세대)

2021년 한국어 학습자 말뭉치 연구 및 구축 사업

이 연구는 1차 중장기 계획에 따른 <2015-2020 한국어 학습자 말뭉치 연구 및 구축> 사업에 이어 2차 중장기 계획을 수립하고 그 계획에 따라 실제 말뭉치를 구축하는 것을 목적으로 하였다. 이에 따른 주요 과업과 연구의 성과는 다음과 같다.

학습자 말뭉치 2차 중장기 계획 수립: 학습자 말뭉치 2차 중장기 계획은 학습자 말뭉치의 구축과 활용에 영향을 미치는 언어 자원 구축 관련 정책과 법·제도에 대한 분석, 학계와 민간 분야를 포함한 다양한 사용자 집단의 요구 분석, 선진 사례 분석, 기구축 말뭉치 분석 등의 기초 연구 결과를 바탕으로 하여 2021년에서 2025년까지 총 5개년 계획으로 수립되었다. 5년간의 사업을 통해 2020년까지 구축한 말뭉치와 합하여 원시 말뭉치 1,000만 어절(문어 600만, 구어 400만 어절), 형태 주석 말뭉치 1,000만 어절(문어 600만, 구어 400만 어절), 오류 주석 말뭉치 500만 어절(문어 300만 어절, 구어 200만 어절) 규모의 말뭉치를 구축하는 것을 목표로 설정하였다. 아울러 말뭉치의 활용도 제고를 위하여 참조 말뭉치를 구축하며, 연구자가 개별적으로 구축한 말뭉치 또는 타 기관 제공 말뭉치와의 통합 구축을 병행할 것을 제안하였다.

한국어 학습자 말뭉치 수집 및 구축·가공: 2021년 한국어 학습자 말뭉치는 2015-2020년에 상대적으로 부족한 장르와 주제의 자료를 집중적으로 수집할 수 있도록 과제를 기획하여 수집하였다.

말뭉치 구축은 원시 말뭉치 831,142어절(문어 419,371어절, 구어 411,771어절), 형태 주석 말뭉치 200,981어절(문어 100,781어절, 구어 100,200어절), 오류 주석 151,906어절(문어 104,314어절, 구어 47,592어절) 규모의 말뭉치가 새롭게 구축되었다. 그 결과 2015-2021년에 구축한 전체 말뭉치의 규모는 원시 말뭉치 5,220,564어절(문어 3,697,952어절, 구어 1,522,612어절), 형태 주석 말뭉치 3,704,586어절(문어 2,602,914어절, 구어 1,101,672어절), 오류 주석 말뭉치 1,153,848어절(문어 590,548어절, 구어 563,300어절)이 되었다.

말뭉치 구축 지원 도구 검증: 한국어 학습자 말뭉치는 표본 등록에서 말뭉치 주석 가공까지 전체 작업 공정을 관리하고 수행할 수 있는 말뭉치 구축 지원 도구를 활용하여 구축해 오고 있다. 본 연구에서는 작업자들에게 안정적인 구축 환경을 제공하기 위하여 구축 실무 연구원들을 중심으로 성능 피드백팀을 구성하여 지속적인 모니터링을 통해 학습자 말뭉치 구축 지원 도구의 성능 개선과 안정화를 위한 피드백을 제공하였다. 또한 향후 수행될 대규모 구축 사업의 효율성 제고를 위하여 지원 도구에 내장된 형태 주석기의 성능을 객관적으로 평가하고 개선 사항을 도출하여 세부적인 기능을 고도화하기 위한 방향성을 제시하였다.

구축 말뭉치 검수 정교화: 구축 말뭉치 검증은 구축된 말뭉치 자료의 질적 제고를 위한 것으로, 문어 입력과 구어 전사, 형태 주석, 오류 주석의 각 작업 단계별로 3단계 작업 및 검수 체제에 따라 작업 공정을 진행하였다. 그 외에도 시스템 기반의 데이터 검증을 통한 오조작 데이터와 이상 데이터 검수를 상호보완적으로 적용하였다. 또한 전체 표본 목록 대조를 통한 중복 표본 추출, 데이터 통계의 정확성을 높이기 위하여 표본 정보를 전 사업 기간에 걸쳐 검수하였다.

학습자 말뭉치 관련 교육 및 홍보: 한국어 학습자 말뭉치 관련 교육은 구축 실무 작업자와 사용자를 대상으로 하여 이루어졌다. 지침 교육과 도구 사용 교육 외에도 구축 과정에서 발생하는 다양한 문제를 해결하기 위한 즉각적 피드백 시스템을 운영하고 정기 워크숍을 통해 말뭉치 구축에 관한 여러 가지 쟁점들을 공유하였다. 사용자를 대상으로 한 교육은 총 5회의 학습자 말뭉치 아카데미 개최를 통해 진행되었다. 각 회차별로 사용자 또는 프로그램을 차별화하여 기초 과정에서 심화 과정까지 다양한 내용을 다루었으며, 각 분야의 전문가를 패널로 한 집담회도 개최하였다. 또한 학술대회 발표를 통해 학습자 말뭉치를 활용한 연구 성과를 학계에 보고하였다.

<한국어 학습자 말뭉치>는 한국어 교육 연구, 교수, 학습에 광범위하게 활용됨으로써 한국어의 세계화 및 국제 경쟁력 강화에 이바지할 것이다.

주요어: 한국어 학습자 말뭉치, 2차 중장기 계획 수립, 문어 말뭉치, 구어 말뭉치, 원시 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	4
3. 연구 방법	5
3.1. 요구분석	6
3.2. 사례조사	7
3.3. 문헌 연구	7
3.4. 전문가 자문	8
4. 연구 수행 기간 및 추진 경과	9
5. 연구 결과	10
 II. 학습자 말뭉치 구축 중장기 계획 수립	 12
1. 말뭉치 관련 정책 환경 및 법·제도 분석, 학계 등 다양한 민간의 요구분석	12
1.1. 말뭉치 관련 정책 환경 및 법·제도 분석	12
1.2. 학계와 민간 분야 등 다양한 사용자 집단의 요구분석	66
2. 학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례 분석	100
2.1. 기본 방향	100
2.2. 연구 내용	101
2.3. 종합 및 적용	123

3. 2015-2020년 한국어 학습자 말뭉치 연구 및 구축 성과 검토	126
3.1. 기본 방향	126
3.2. 연구 내용	127
3.3. 종합 및 적용	140
4. 말뭉치 구축·활용에 관한 저작권 확보를 위한 세부 실행 방법 마련 ...	141
4.1. 저작권 이용 허락 동의	142
4.2. 개인정보 수집 및 이용, 관리	148
5. 한국어 학습자 말뭉치 수집·구축·활용의 중장기 목표 및 단계별 세부 전략 수립	150
5.1. 중장기 계획 수립을 위한 기본 방향 및 목표	150
5.2. 목표 구축 규모	152
5.3. 단계별 구축 계획	153
5.4. 말뭉치 수집 전략	157
5.5. 말뭉치 구축 및 가공 전략	161

Ⅲ. 말뭉치 수집 및 구축·가공 169

1. 말뭉치 수집	169
1.1. 수집 대상	169
1.2. 수집 네트워크	169
1.3. 수집 과제	169
1.4. 자료 수집 현황	171
2. 구축 및 가공	175
2.1. 원시 말뭉치	175
2.2. 형태 주석 말뭉치	180

2.3. 오류 주석	184
IV. 말뭉치 구축 지원 도구 검증	187
1. 구축 지원 도구 성능 평가팀 운영을 통한 피드백	187
2. 대규모 구축을 위한 형태 주석기 성능 평가	188
V. 구축 말뭉치 검수 정교화	193
1. 작업 공정에서의 3단계 검수 체계	193
2. 개별 표본의 표본 정보 검수 체계 강화	193
3. 작업 중 생성된 오조작 데이터 검증	194
VI. 학습자 말뭉치 교육 및 홍보	195
1. 말뭉치 구축/가공 인력 실무 교육	195
1.1. 교육 대상	195
1.2. 교육 방법	195
1.3. 교육 내용	196
1.4 참여 인력	197
2. 한국어 학습자 말뭉치 아카데미 개최	199
3. 학술대회 발표	200
4. 말뭉치 소개·활용 자료 제작, 한국어교수학습센터 게재 및 아카데미 배포 ..	201

VII. 결론	202
1. 연구 요약	202
2. 연구의 의의 및 기대 효과	205
3. 보고서 활용 방안	207
4. 정책 제언	208

참고 자료

부록 1. 기초 연구 자료

부록 2. 2021년 한국어 학습자 말뭉치 구축 지침

표 차례

<표 1> 연구의 범위와 세부 연구 내용	4
<표 2> 연구 방법 및 절차	5
<표 3> 요구분석 방법 및 내용	7
<표 4> 전문가 자문단 구성	8
<표 5> 연구 추진 경과	9
<표 6> 2015-2020년 학습자 말뭉치의 구축 규모	10
<표 7> 공공데이터 이용 가이드의 공공데이터 소개	14
<표 8> 일반적 공공데이터와 학습자 말뭉치 비교	15
<표 9> 언어 자원 관련 법령	16
<표 10> 지능정보화 기본법 조항별 핵심 내용	16
<표 11> 지능정보화 기본법 학습자 말뭉치 관련 조항과 핵심 내용	17
<표 12> 공공데이터법(약칭) 조항별 핵심 내용	18
<표 13> 공공데이터법(약칭) 학습자 말뭉치 관련 조항과 핵심 내용	19
<표 14> 정보공개법(약칭) 조항별 핵심 내용	20
<표 15> 정보공개법(약칭) 학습자 말뭉치 관련 조항과 핵심 내용	20
<표 16> 저작권법의 용어와 내용	23
<표 17> 저작권 관련 판례	24
<표 18> 저작권법에서의 이용 및 양도에 관한 내용	25
<표 19> 저작권법에서의 이용 및 양도에 관한 내용	26
<표 20> 국외 말뭉치에서의 저작권과 이용 내용	26
<표 21> ICLE의 메타 정보	31
<표 22> Data Collection for Learner Corpus of Latvian(LaVA) 동의서 세부 내용(Inga Kaija, Ilze A. Auzina, 2020:43-45)	32

<표 23> CEDEL2(Corpus Escrito del Español L2) 동의서의 세부 정보	35
<표 24> 개인정보 보호법의 내용	36
<표 25> 국외 말뭉치 말뭉치에서 수집된 학습자 변인	40
<표 26> 국외 말뭉치 수집 과정에서 개인정보 보호를 위한 처리 예시	41
<표 27> 말뭉치에서의 개인정보 침해 사례	45
<표 28> 국외 말뭉치 자료에서의 개인정보 보호를 위한 처리 예시	46
<표 29> 국외 학습자 말뭉치의 메타데이터	50
<표 30> IRB의 내용과 기능	55
<표 31> 보건복지부령의 인간 대상 연구의 범위	56
<표 32> 인간 대상 연구의 유형	57
<표 33> IRB의 심의를 면제할 수 있는 인간 대상 연구	59
<표 34> 저작권재산권 양도와 이용허락의 차이	62
<표 35> 국외 말뭉치의 수집 및 이용에 관한 사항	63
<표 36> 설문조사 문항 구성	67
<표 37> 응답자 국적 빈도	68
<표 38> 응답자 직업 빈도	69
<표 39> 응답자 소속기관 빈도	70
<표 40> 응답자 소속기관 소재지 분포	71
<표 41> 학습자 말뭉치 이용 시기 빈도	72
<표 42> 학습자 말뭉치를 알게 된 경로 빈도	73
<표 43> 학습자 말뭉치 자료 다운로드 경로 빈도	74
<표 44> 학습자 말뭉치 이용 목적 빈도	75
<표 45> 한국어 및 한국어 교육 연구 활용에서의 연구 목적 빈도	76
<표 46> 한국어 및 한국어 교육 연구 활용에서의 연구 영역 빈도	77

<표 47> 연구 영역별 세부 연구 주제	77
<표 48> 교수학습 자료 개발에 활용에서의 연구 목적 분포	79
<표 49> 사용 말뭉치 유형 빈도	80
<표 50> 사용 말뭉치 자료 유형 빈도	81
<표 51> 사용 말뭉치 학습자 수준 빈도	82
<표 52> 사용 말뭉치 언어권 빈도	83
<표 53> 사용 말뭉치 자료 형식 빈도	84
<표 54> 사용 말뭉치에 대한 만족도 빈도	85
<표 55> 말뭉치 구성 관련 개선 사항 및 제안 사항	85
<표 56> 말뭉치 제공 형식 관련 개선 사항 및 제안 사항	86
<표 57> 말뭉치 가공 사용 관련 개선 사항 및 제안 사항	86
<표 58> 한국어 학습자 말뭉치 전반에 대한 개선 사항 및 제안 사항	87
<표 59> 한국어 학습자 말뭉치 나눔터 이용 횟수 빈도	89
<표 60> 한국어 학습자 말뭉치 나눔터에서의 주 이용 빈도	90
<표 61> 한국어 학습자 말뭉치 나눔터 만족도 빈도	91
<표 62> 자료의 접근성 및 UI의 편의성에 관한 개선 사항 및 제안 사항	94
<표 63> 검색 조건 추가에 관한 제안 사항	94
<표 64> 검색 메뉴 사용에 관한 제안 사항	95
<표 65> 민간 분야 전문가 집담회 패널 구성	97
<표 66> CZESL의 과제 유형	103
<표 67> SECCL의 과제 유형	105
<표 68> Corpus Escrito del Español L2의 작문 주제	110
<표 69> ICLE의 에세이 주제	112
<표 70> 국외의 학습자 말뭉치 구축 개요: 200만 어절 규모 이상, 구글 인용 빈도 100 이상 학습자 말뭉치	115

<표 71> 2015-2020년 한국어 학습자 말뭉치 유형별 구축 통계	127
<표 72> 2015-2020년 한국어 학습자 말뭉치 대상별·수준별 통계: 형태 주석 말뭉치 ..	128
<표 73> 2015-2020년 한국어 학습자 말뭉치 대상별·수준별 통계: 오류 주석 말뭉치 ...	129
<표 74> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 원시 문어 말뭉치	129
<표 75> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 형태 문어 말뭉치	130
<표 76> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 오류 문어 말뭉치	131
<표 77> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 원시 구어 말뭉치	132
<표 78> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 형태 구어 말뭉치	132
<표 79> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 오류 구어 말뭉치	133
<표 80> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 원시 문어 말뭉치	134
<표 81> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 형태 문어 말뭉치	134
<표 82> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 오류 문어 말뭉치	135
<표 83> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 원시 구어 말뭉치	136
<표 84> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 형태 구어 말뭉치	136
<표 85> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 오류 구어 말뭉치	137
<표 86> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 원시 말뭉치	138
<표 87> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 형태 주석 말뭉치	139
<표 88> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 오류 주석 말뭉치	140
<표 89> 저작권 이용 허락 동의서 시범 적용 후 의견수렴 문항 구성	144
<표 90> 개인정보 수집·이용에 대한 동의 항목	149
<표 91> 고유식별정보 처리에 대한 동의 항목	149
<표 92> 개인정보의 제3자 제공에 대한 동의	150
<표 93> 1차 중장기 계획과 2차 중장기 계획의 기본 방향 비교	151
<표 94> 2021-2025년 학습자 말뭉치 수집 및 구축·가공 목표	153

<표 95> 단계별 구축의 방향	154
<표 96> 2021-2025년 단계별 구축 규모	156
<표 97> 2021-2025년 단계별 구축 규모	156
<표 98> 수집 네트워크의 다변화	158
<표 99> 균형성 보완을 위한 기획 수집 전략	160
<표 100> 2021-2022년 학습자 말뭉치 형태 주식 체계	162
<표 101> 2015-2020년 학습자 말뭉치 오류 주식 체계: 기본 주식	164
<표 102> 2015-2020년 학습자 말뭉치 오류 주식 체계 틀: 오류 층위	165
<표 103> 2015-2020년 학습자 말뭉치 오류 주식 체계: 오류 양상	166
<표 104> 기획 과제: 문어	170
<표 105> 기획 과제: 구어	171
<표 106> 문어 자료 수집 현황 및 분포	171
<표 107> 구어 자료 수집 현황 및 분포	173
<표 108> 문어 원시 말뭉치의 숙달도별 자료 분포	176
<표 109> 문어 원시 말뭉치의 언어권별 자료 분포	176
<표 110> 구어 원시 말뭉치의 숙달도별 자료 분포	178
<표 111> 구어 원시 말뭉치의 언어권별 자료 분포	179
<표 112> 문어 형태 주식 말뭉치의 숙달도별 자료 분포	180
<표 113> 문어 형태 주식 말뭉치의 언어권별 자료 분포	181
<표 114> 구어 형태 주식 말뭉치의 숙달도별 자료 분포	182
<표 115> 구어 형태 주식 말뭉치의 언어권별 자료 분포	183
<표 116> 문어 오류 주식 말뭉치의 숙달도별 자료 분포	184
<표 117> 문어 오류 주식 말뭉치의 언어권별 자료 분포	185
<표 118> 구어 오류 주식 말뭉치의 숙달도별 자료 분포	186

<표 119> 구어 오류 주석 말뭉치의 언어권별 자료 분포	186
<표 120> 말뭉치 구축/가공 인력 교육 내용	196
<표 121> 학습자 말뭉치 활용 아카데미 개최	199
<표 123> ‘학습자 말뭉치 활용 매뉴얼’의 구성	201

그림 차례

<그림 1> ICLE(International Corpus of Learner English) 학습자 동의서 예시	29
<그림 2> 동의서 및 설문 양식(Inga Kaija, Ilze A. Auzina(2020))	32
<그림 3> CEDEL2(Corpus Escrito del Español L2) 동의서	34
<그림 4> 학습자 말뭉치의 변인(Granger, 2008:4)	40
<그림 5> 개인정보의 범주 구분(대한상공회의소, 2018:5)	44
<그림 6> 응답자 국적 분포	68
<그림 7> 응답자 직업 분포	69
<그림 8> 응답자 소속기관 분포	70
<그림 9> 응답자 소속기관 소재지 분포	71
<그림 10> 학습자 말뭉치 이용 시기 분포	72
<그림 11> 학습자 말뭉치를 알게 된 경로 분포	73
<그림 12> 학습자 말뭉치 자료 다운로드 경로 분포	74
<그림 13> 학습자 말뭉치 이용 목적 분포	75
<그림 14> 한국어 및 한국어 교육 연구 활용에서의 연구 목적 분포	76
<그림 15> 한국어 및 한국어 교육 연구 활용에서의 연구 영역 분포	77
<그림 16> 교수학습 자료 개발에 활용에서의 연구 목적 분포	78
<그림 17> 사용 말뭉치 유형 분포	79
<그림 18> 사용 말뭉치 자료 유형 분포	80
<그림 19> 사용 말뭉치 학습자 수준 분포	81
<그림 20> 사용 말뭉치 언어권 분포	82
<그림 21> 사용 말뭉치 자료 형식 분포	83
<그림 22> 사용 말뭉치에 대한 만족도 분포	84
<그림 23> 한국어 학습자 말뭉치 나눔터 이용 횟수 분포	88

<그림 24> 한국어 학습자 말뭉치 나눔터에서의 주 이용 메뉴 분포	89
<그림 25> 한국어 학습자 말뭉치 나눔터 만족도 분포	90
<그림 26> 한국어 학습자 말뭉치 나눔터의 자료 접근 성 및 UI의 편의성 만족도 분포	92
<그림 27> 한국어 학습자 말뭉치 나눔터의 검색 기능 만족도 분포	93
<그림 28> 한국어 학습자 말뭉치 나눔터의 통계 정보 만족도 분포	93
<그림 29> 저작권 이용 허락 동의 관련 방침 수립 절차	142
<그림 30> 설문 결과: 자료 수집 참여 시 사용 가능한 시간	145
<그림 31> 설문 결과: 자료 수집에 대한 보상 방법	146
<그림 32> 2차 중장기 계획에서의 학습자 말뭉치 구축 방향	151
<그림 33> 2021-2025년 학습자 말뭉치 구축 방향	154
<그림 34> 2021-2025년 학습자 말뭉치 수집 전략	158
<그림 35> 2021-2025년 학습자 말뭉치 구축 및 가공 전략	161
<그림 36> 오류 주석 가공 전략: 주석 체계의 이원화	167
<그림 37> 외부 검수단 운영을 통한 검수 체계 강화 모형	168
<그림 38> 형태 분할 정확성에 관한 형태소 주석기의 성능 비교	190
<그림 39> 분석의 정확성에 관한 형태소 주석기의 성능 비교	191

I. 연구 개요

1. 연구의 목적 및 필요성

1.1. 연구의 목적

○ 본 연구는 1차 중장기 계획에 따른 <2015-2020 한국어 학습자 말뭉치 연구 및 구축> 사업에 이은 2차 중장기 계획을 수립하고 그 계획에 따라 실제 말뭉치를 구축하는 데에 주요한 목적이 있다. 지난 1차 중장기 계획이 주로 한국어 교육 연구 및 교수 현장에서의 활용에 초점을 두었던 것에서 더 나아가 2차 중장기 계획에서는 그 범위를 확대하여 연구와 교육뿐만 아니라 민간 분야에까지 활용 범위를 확대할 수 있는지 가능성을 모색해 보고자 하였다. 다음은 2021년 <한국어 학습자 말뭉치 연구 및 구축 사업>의 세부 목표이다.

- 말뭉치 구축 중장기 계획 수립
- 말뭉치 수집 및 구축 가공
- 말뭉치 구축 지원 도구 검증
- 구축 말뭉치 검수 정교화
- 말뭉치 교육 및 홍보

1.2. 연구의 필요성

○ 공공 언어 자원으로서 국가 주도의 한국어 학습자 말뭉치 구축

본 연구는 <2015-2020년 한국어 학습자 말뭉치 연구 및 구축 사업>에 이어 새로운 중장기 계획을 수립하는 것을 주요한 과업으로 하는 사업이다. 한국어 학습자 말뭉치는 2015년에서 수립한 중장기 계획에 따라 국내 교육기관의 유학생과 이주민, 국외의 한국어 학습자 자료를 대규모로 수집하여 약 440만 어절 규모의 균형 말뭉치를 구축하였다. 본 연구는 2002년 문화체육관광부의 주도로 수행된 50만 어절의 규모의 한국어 학습자 말뭉치 구축 사업 이후 13년 만에 새롭게 국립국어원에서 시작된 국가 주도의 사업으로 한국어

의 세계화를 위한 지식 기반 구축 사업으로서 높은 평가를 받았다. 본 연구는 그 후속 사업으로 공공 언어 자원으로서는 대규모의 국가 주도 한국어 학습자 말뭉치를 구축한다는 데에 의의가 있다.

○ 2015-2020년 학습자 말뭉치 보완을 통한 균형성 확보

<2015-2020년 한국어 학습자 말뭉치 연구 및 구축 사업>에서는 1단계 국내 학습자, 2단계 이주민, 3단계 국외 학습자의 자료를 집중 구축 대상으로 하여 원시 말뭉치를 기준으로 문어 자료 약 330만 어절, 구어 자료 110만 어절을 구축하였다. 이는 귀납적인 구축 결과로 학습자의 수준, 제1언어, 자료의 장르, 주제 등의 변인을 고려한 균형 말뭉치를 구축하기까지 지속적인 보완이 이루어져야 한다. 본 연구는 <2015-2020년 한국어 학습자 말뭉치 연구 및 구축 사업> 성과에 대한 분석을 바탕으로 특정 변인에 편중된 자료를 보완함으로써 균형성을 확보하여 국가 주도 학습자 말뭉치의 질을 제고하고자 하였다는 점에서 실효성이 크다고 하겠다.

○ 1,000만 어절 규모의 대규모 학습자 말뭉치로의 확장

학습자 말뭉치 구축 및 활용에 관한 연구는 외국어 또는 제2 언어 교육 분야의 핵심적인 트렌드 중 하나라고 할 수 있다. 국외에서는 민간 기관, 대학, 개인 연구자에 의해 약 183종의 학습자 말뭉치가 구축되어 왔으며, 영어 교육 분야의 경우 1,000만 어절 이상의 대규모 말뭉치를 구축한 경우도 많다. 2020년을 기준으로 CLC(Cambridge Learner Corpus)는 약 5천만 어절, The Hong Kong University of Science & Technology(HKUST)의 Learner Corpus는 약 2천 5백만 어절, The Longman Learners' Corpus는 1천만 어절 규모에 이르며, 모어 화자의 자료를 포함한 The Uppsala WordReference Corpus의 경우는 약 1억 3천만 어절에 달한다. 말뭉치 구축의 역사가 상대적으로 긴 국외에서 규모가 5백만 어절 이상인 말뭉치의 수가 7개, 100만 어절 이상인 말뭉치의 수가 37개에 불과함을 고려할 때 <2015-2020년 한국어 학습자 말뭉치 연구 및 구축 사업>에서 440만 어절 규모의 말뭉치를 구축하였다는 것은 대단한 성과가 아닐 수 없다. 본 연구는 여기에서 더 나아가 1,000만 어절을 목표로 학습자 말뭉치의 규모를 확장함으로써 세계적인 수준의 학습자 말뭉치를 구축하고, 한국어 교육 및 연구뿐만 아니라 민간 분야까지 사용의 폭을 확대할 수 있다는 점에서 의미가 있다.

○ 대규모의 언어 자원으로써 한국어 학습자 말뭉치 규모 확대

최근 빅데이터가 다양한 분야에서 광범위하게 활용되면서 중요한 국가 자원으로써 주목받고 있다. 이에 따라 국립국어원에서는 2018년에 ‘21세기 세종 한국어 균형 말뭉치’를 넘어서는 ‘국어 빅데이터’ 구축 사업에 착수하였고 2019년부터 본격화될 예정이다. 그 밖에도 국어기본법에서 공용어로서의 지위를 인정한 수어 말뭉치를 구축하고 있다. 이는 거시적으로 국어 빅데이터의 일부로서 다양한 유형의 국어 자료를 구축한다는 점에서 매우 가치 있는 일이라고 하겠다. 한국어 학습자 말뭉치는 국어 빅데이터의 일부로서 구축 초기부터 ‘21세기 세종 한국어 균형 말뭉치’와의 호환성을 고려하며 동일한 체계에 따라 구축 지침을 마련한 바 있다. 한국 언어·문화의 세계화라는 측면에서 비모어 화자가 산출한 언어 자료인 한국어 학습자 말뭉치가 국어 빅데이터 구축 사업과 함께 진행되고 있다는 것은 더욱 의미 있는 일이라고 하겠다.

○ 학습자 말뭉치를 활용한 연구 방법론과 도구 사용 교육에 대한 요구

한국어 학습자 말뭉치에 대한 연구자와 교수자들의 관심이 날로 커지고 있다. 그럼에도 불구하고 일정한 크기와 균형성, 대표성을 기본 요건으로 하는 말뭉치의 특성상 연구자 개개인이 자료를 구축하기가 쉽지 않다는 것이 뛰어넘기 어려운 벽이 되어 왔다. 이러한 점에서 국가 수준의 학습자 말뭉치를 구축한다는 것은 국가 자원으로써 광범위하게 활용 가능한 자료를 구축하여 다양한 목적의 사용자들이 공유할 수 있다는 점에서 큰 의미가 있다. 한편, 말뭉치에 관한 또 하나의 벽은 말뭉치 활용법에 대해 사용자들이 느끼는 어려움과 거리감이라고 할 수 있다. 본 연구에서는 이러한 사용자들을 위해 ‘한국어 학습자 말뭉치 나눔터’를 통해 제공되는 학습자 말뭉치 활용 방법을 소개하고, 더 나아가 더욱 광범위하게 활용할 수 있도록 자료 처리 도구 사용법을 교육하게 된다.

2. 연구의 범위

○ 본 연구의 범위와 세부 내용은 다음과 같다.

<표 1> 연구의 범위와 세부 연구 내용

연구의 범위	세부 연구의 내용
학습자 말뭉치 중장기 계획 수립	<ul style="list-style-type: none"> ○ 말뭉치 관련 정책 환경 및 법·제도 분석, 학계 등 다양한 민간의 요구분석 ○ 학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례 분석 ○ 2015-2020 말뭉치 연구 및 구축 성과 검토 ○ 저작권 확보를 위한 세부 실행 방안 마련 ○ 중장기 목표, 단계별 세부 전략 수립
말뭉치 수집 및 구축·가공	<ul style="list-style-type: none"> ○ 80만 어절의 말뭉치 구축 (문어 40만 어절, 구어 40만 어절) ○ 형태 주석 및 오류 주석 가공 (형태 주석 20만 어절, 오류 주석 15만 어절)
말뭉치 구축 지원 도구 검증	<ul style="list-style-type: none"> ○ 구축 지원 도구 성능 피드백 ○ 구축 지원 도구 개발 업체와의 정기 회의(격월 1회)
구축 말뭉치 검수 정교화	<ul style="list-style-type: none"> ○ 문어, 구어 말뭉치의 입력/전사 작업 및 검수 시스템 정교화
말뭉치 교육 및 홍보	<ul style="list-style-type: none"> ○ 말뭉치 구축/가공 인력 실무 교육 ○ 한국어 학습자 말뭉치 아카데미 개최(4회 이상) ○ 학술대회 발표(상반기, 하반기 각 1회) ○ 말뭉치 소개·활용 자료집 제작, 한국어교수학습샘터 게재 및 아카데미 배포

3. 연구 방법

- 본 연구에서는 과업의 유형을 크게 연구, 말뭉치 구축 및 가공, 교육 및 홍보의 세 가지로 보고 그에 적합한 연구 방법을 적용하여 연구를 수행하였다.

<표 2> 연구 방법 및 절차

과업 내용		연구 방법	과업 유형
학습자 말뭉치 중장기 계획 수립	말뭉치 관련 정책 환경 및 법·제도 분석, 학계 등 다양한 민간의 요구분석	○ 요구분석(설문조사, 사용자 간담회)	연구
	학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례 분석	○ 사례조사 ○ 문헌 연구	연구
	2015-2020 말뭉치 연구 및 구축 성과 검토	○ 구축 성과에 대한 계량적 분석, 구축 단계별 쟁점 및 대안 모색을 위한 연구 성과의 질적 분석 병행	연구
	저작권 확보를 위한 세부 실행 방안 마련	○ 문헌 연구 ○ 사례조사 ○ 전문가 자문	연구
	중장기 목표, 단계별 세부 전략 수립	○ 문헌 연구 ○ 전문가 자문 ○ 기초 연구 성과 종합	연구
말뭉치 수집 및 구축·가공		○ 말뭉치 구축 및 가공	구축 및 가공

과업 내용		연구 방법	과업 유형
말뭉치 구축 지원 도구 검증		○ 구축 도구 성능 평가팀 운영, ○ 개발업체·구축팀·수요 기관의 정기회의(격월)	구축 및 가공
구축 말뭉치 검수 정교화		○ 실제 말뭉치 검수	구축 및 가공
말뭉치 교육 및 홍보	말뭉치 구축 인력 실무 교육	○ 작업 실무자 정기회의(격주) ○ 작업 실무자 교육 및 워크숍(월 1회)	교육
	학습자 말뭉치 아카데미 개최	-	교육 및 홍보
	학술대회 발표	-	홍보
	학습자 말뭉치 소개 및 활용에 관한 자료집 제작, 배포	-	교육 및 홍보

3.1. 요구분석

- 요구분석은 학습자 말뭉치의 활용도 제고를 위해 중장기 계획 수립의 기초 연구로서 한국어 교육 연구자 및 교원, 민간 분야의 사용자 집단의 의견을 수렴하는 과정이다. 본 연구에서는 다음과 같이 집단별 요구분석의 내용과 방법을 차별화하여 사용자의 의견을 폭넓게 수렴하고자 하였다.

<표 3> 요구분석 방법 및 내용

조사 대상	조사 목적	조사 방법	인원 수
한국어 교육 연구자 및 교원	한국어 학습자 말뭉치 활용 범위, 학습자 말뭉치 나눴 터 사용 만족도 및 개선 방 향	설문조사	약 148명 (학술지 논문 및 학위논문 발표자, 학습자 말뭉치 아 카데미 참가자)
민간 분야	민간 분야에서의 학습자 말 뭉치 활용 범위 및 활용을 위한 구축 방향	집담회	약 5명 (인공지능 개발, 포털사이트, 에듀테크 분야 종사자)

3.2. 사례조사

- 사례조사는 학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례, 저작권과 관련한 국내의 영어교육 또는 국외의 학습자 말뭉치 관련 정책을 조사하기 위한 것이다. 본 연구에서는 국외의 대학, 민간 기관 또는 연구자들이 구축한 학습자 말뭉치를 대상으로 하여 사례조사를 하여 중장기 계획의 방향을 설정하기 위한 기초 자료로 활용하였다.

3.3. 문헌 연구

- 문헌 연구는 학습자 말뭉치 구축에 관한 이론 연구와 선행연구 검토의 두 가지를 포함한다. 이론 연구는 학습자 말뭉치를 포함해 말뭉치 설계와 구축, 가공, 배포에 관한 이론적 체계를 검토하여 말뭉치 구축의 방향 수립과 구축 지침에 참고하기 위한 것으로 본 연구에서는 특히, 저작권에 관한 법령 검토에 초점을 두어 진행하였다. 한편, 선행연구 검토는 학습자 말뭉치 구축 및 활용에 관한 선진 사례 및 연구 성과를 검토하기 위한 것으로 사례조사 내용을 보완하기 위한 목적으로 이루어졌다.

3.4. 전문가 자문

- 전문가 자문은 각 분야의 전문가의 의견을 수렴하여 중장기 계획 수립과 전략에 대한 타당성을 확보하고 한국어 학습자 말뭉치 활용도를 제고할 수 있는, 실효성 있는 계획을 마련하기 위한 것이다. 본 연구에서는 법률 전문가, 전 세계 한국어 학습자의 한국어 교육을 담당하고 있는 세종학당재단의 기관장, 한국어 교육 학계의 교수, 민간 분야 실무자의 전문성과 경험에 관한 의견을 수렴하였다. 이는 한국어 학습자 말뭉치 활용의 폭을 극대화하기 위한 것이다.

<표 4> 전문가 자문단 구성

분야		자문위원
저작권 관련 법률 검토		우원상 법률사무소
한국어 학습자 말뭉치 구축 및 배포, 활용	한국어 교육 유관 기관	강현화(세종학당재단 이사장)
	한국어 교육	김정숙(고려대학교 국어국문학과 교수)
	민간 분야	곽용진(주) 이르테크 대표) 박진규(ETRI 복합지능연구실 실장) 이기황(바이브컴퍼니, 지식&인사이트랩 이사) 이형구(옐로우 크리에이티브 대표)

4. 연구 수행 기간 및 추진 경과

○ 연구 수행 기간: 2021년 4월 20일-2021년 12월 10일

○ 연구 추진 경과

<표 5> 연구 추진 경과

세부 연구 내용			4월	5월	6월	7월	8월	9월	10월	11월	12월
착수	계약 및 착수보고회의		○	○							
기초 연구	법·제도 분석, 저작권 확보를 위한 세부 실행 방안 마련			○	○	○	○	○	○		
	정책 환경 및 학계 등 다양한 민간의 요구분석			○	○	○	○				
	학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례 분석			○	○	○	○				
	2015-2020 말뭉치 연구 및 구축 성과 검토			○							
	중장기 목표, 단계별 세부 전략 수립						○	○	○	○	
구축 및 가공	말뭉치 수집 및 구축가공	자료 수집						○	○	○	○
		구축가공						○	○	○	○
	말뭉치 구축 지원 도구 검증							○	○	○	
	구축 말뭉치 검수 정교화					○	○	○	○	○	○
교육 및 홍보	말뭉치 구축 인력 실무 교육						○	○	○	○	○
	학습자 말뭉치 아카데미 개최(5회)					○	○			○	○
	학술대회 발표(2회)									○	
	학습자 말뭉치 소개 및 활용에 관한 자료집 제작, 배포								○	○	○
마무리	최종 보고서 작성										○

5. 연구 결과

○ 본 사업의 연구 결과는 다음과 같다.

- 학습자 말뭉치 중장기 계획 및 세부 전략 수립
- 한국어 학습자 말뭉치 아카데미 개최 및 학회 발표
- 말뭉치 구축 지원 도구 검증
- 구축 말뭉치 검수 정교화
- 말뭉치 소개·활용 자료집 제작 및 배포
- 말뭉치 수집 및 구축·가공

<표 6> 2015-2021년 학습자 말뭉치의 구축 규모

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
원시 말뭉치									
2015 - 2020	문어	385,317 (5,732)	535,903 (5,483)	626,589 (5,239)	603,099 (4,626)	575,274 (3,682)	409,248 (2,483)	143,151 (155)	3,278,581 (27,400)
	구어	205,149 (685)	225,004 (506)	227,925 (502)	194,128 (412)	118,922 (212)	86,720 (138)	52,993 (86)	1,110,841 (2,541)
2021	문어	53,953 (958)	59,943 (649)	59,807 (649)	47,991 (357)	146,242 (972)	49,933 (322)	1,502 (8)	419,371 (3,802)
	구어	53,621 (114)	98,243 (160)	149,089 (148)	49,607 (101)	30,593 (44)	19,475 (25)	11,143 (9)	411,771 (601)
합계	문어	439,270 (6,690)	595,846 (6,132)	686,396 (5,775)	651,090 (4,983)	721,516 (4,654)	459,181 (2,805)	144,653 (163)	3,697,952 (31,202)
	구어	258,770 (799)	323,247 (666)	377,014 (650)	243,735 (513)	149,515 (253)	106,195 (163)	64,136 (95)	1,522,612 (3,142)
형태 주석 말뭉치									
2015 - 2020	문어	356,901 (5,179)	428,486 (4,330)	444,985 (3,709)	421,797 (3,300)	408,938 (2,731)	369,900 (2,309)	71,126 (83)	2,502,133 (21,641)
	구어	196,790 (651)	196,798 (431)	205,309 (451)	177,406 (366)	108,999 (196)	86,473 (137)	29,697 (33)	1,001,472 (2,265)

구분		1급	2급	3급	4급	5급	6급	6급 이 상	합계
2021	문어	23,687 (387)	5,225 (71)	0 (0)	725 (5)	34,093 (254)	37,051 (283)	0 (0)	100,781 (1,000)
	구어	15,030 (44)	8,691 (11)	1,889 (4)	31,608 (63)	27,643 (43)	15,339 (21)	0 (0)	100,200 (186)
합계	문어	380,588 (5,566)	433,711 (4,401)	444,985 (3,709)	422,522 (3,305)	443,031 (2,985)	406,951 (2,592)	71,126 (83)	2,602,914 (22,641)
	구어	211,820 (695)	205,489 (442)	207,198 (455)	209,014 (429)	136,642 (239)	101,812 (158)	29,697 (33)	1,101,672 (2,451)

오류 주식 말뭉치

2015 - 2020	문어	83,469 (1,214)	90,420 (928)	88,502 (767)	82,766 (686)	77,161 (548)	75,099 (481)	3,693 (20)	501,110 (4,644)
	구어	89,604 (300)	97,163 (211)	90,832 (224)	91,655 (214)	69,554 (120)	54,956 (75)	7,068 (12)	500,832 (1,156)
2021	문어	7,264 (100)	13,904 (155)	16,874 (159)	3,603 (31)	32,447 (226)	30,222 (211)	0 (0)	104,314 (882)
	구어	4,223 (9)	7,332 (9)	10,932 (27)	18,237 (20)	4,564 (5)	2,304 (3)	0 (0)	47,592 (73)
합계	문어	89,438 (1,298)	102,422 (1,056)	104,095 (911)	85,942 (713)	101,518 (718)	103,440 (680)	3,693 (20)	590,548 (5,396)
	구어	95,122 (325)	106,397 (247)	103,045 (266)	110,319 (238)	82,208 (181)	59,141 (90)	7,068 (12)	563,300 (1,359)

- 단위: 어절 수, () 안은 표본 수

II. 학습자 말뭉치 구축 중장기 계획 수립

1. 말뭉치 관련 정책 환경 및 법·제도 분석, 학계 등 다양한 민간의 요구분석

- 본 연구는 대규모의 한국어 학습자 말뭉치를 구축하여 체계적이고 과학적인 한국어 교육을 위한 기초 연구, 4차 산업혁명 시대의 최첨단 기술을 적용한 인공지능 기술 개발 등 교육 및 연구 분야에서 폭넓게 활용하도록 하는 데에 목적이 있다. 이에 따라 본 연구에서는 말뭉치 관련 정책 환경과 법, 제도에 대한 면밀한 분석을 통해 국가 언어 자원으로서의 법률적·기술적 쟁점들을 찾아내 합리적인 해결 방안을 마련하고자 하였다. 아울러 학계와 민간 분야의 다양한 사용자 집단의 요구를 분석하여 중장기 계획안 마련의 기초 자료로 활용함으로써 학습자 말뭉치의 활용도를 제고하고자 하였다.

1.1. 말뭉치 관련 정책 환경 및 법·제도 분석

1.1.1. 기본 방향

- 학습자 말뭉치와 관련되는 언어 자원 관련 정책 및 법률에 무엇이 있는지 살펴보고, 학습자 말뭉치에 직접적으로 영향을 미치는 세부 조항을 살펴보는 것은 학습자 말뭉치의 구축과 자료 활용의 원칙을 세우는 데에 중요한 토대가 된다. 이에 따라 본 연구에서는 언어 자원 구축과 관련된 정책 환경, 공공데이터로서의 언어 자원 구축에 관한 법과 제도, 특히 저작권법과 개인정보 보호법, IRB 규정을 살펴보고 국외 학습자 말뭉치에서의 적용 사례를 검토함으로써 학습자 말뭉치에의 적용을 위한 기초 자료로 삼고자 하였다.

1.1.2. 연구 내용

(1) 언어 자원 구축에 관한 정책 환경

○ 4차 산업혁명 시대의 도래와 함께 국가 경쟁력 제고를 위해 D.N.A.(데이터, 네트워크, 인공지능) 등의 지능정보기술 개발의 원천이 되는 빅데이터 구축과 활용의 중요성에 대한 인식의 전환과 함께 구체적인 방안이 정책적 차원에서 논의되고 있다. 이에 따라 정부에서는 2017년 9월 ‘대통령직속 4차산업혁명위원회’의 발족과 함께 4차 산업혁명과 관련한 다양한 국가 전략과 주요 정책을 수립하고 구체적인 실행 계획을 내놓고 있다. 또한 한국인공지능윤리협회와 같은 민간단체를 통해 인공지능 기술의 개발과 사용에 관한 윤리 문제에 관한 연구가 활발하게 이루어지고 있다. 이들 기관의 주요 역할과 기조는 다음과 같다.

▪ 4차산업혁명위원회

대통령직속 기관으로 『4차산업혁명위원회의 설치 및 운영에 관한 규정』에 의해서 설립되었으며, 다음의 사항을 심의·조정하고 있다.

- 4차 산업혁명에 대한 종합적인 국가전략
- 4차 산업혁명 관련 각 부처별 실행계획과 주요 정책
- 4차 산업혁명의 근간이 되는 과학기술 발전 지원, 인공지능·ICT 등 핵심기술 확보 및 기술혁신형 연구개발 성과창출 강화에 관한 사항
- 전 산업의 지능화 추진을 통한 신산업·신서비스 육성에 관한 사항 등

▪ 한국인공지능윤리협회

2019년 10월 23일 ‘제1장 인간과 인공지능의 관계, 제2장 선하고 안전한 인공지능, 제3장 인공지능 개발자의 윤리, 제4장 인공지능 소비자의 윤리, 제5장 공동의 책임과 이익의 공유’ 총 5개의 장, 40개 조항으로 한국인공지능윤리헌장을 제정·공포하였으며, 2장에 다음과 같이 빅데이터와 관련된 조항들이 포함되어 있다.

한국인공지능윤리현장 제2장 선하고 안전한 인공지능
제11조. 인공지능 기술과 알고리즘은 기록과 문서화에 의해 투명하게 관리되어야 한다.
제12조. 인공지능 기술, 알고리즘, 데이터들은 외부 공격에 방어할 수 있는 강력한 보안체계를 유지해야 한다.
제13조. 인공지능에 학습되는 빅데이터는 신뢰할 수 있고, 편향적이지 않으며, 합법적이어야 한다.
제14조. 빅데이터 수집 시 합법적인 절차에 따라야 하며 개인의 프라이버시를 침해하지 않아야 한다.

(2) 공공데이터로서의 학습자 말뭉치

- 공공데이터에 대해 정부의 공공데이터 포털(www.data.go.kr)에서는 공공데이터법 제1조와 제3조에 근거하여 공공데이터 포털을 통한 데이터의 편리한 이용 및 이용권의 보편적 확대를 위해 노력하고, 각 공공기관이 보유한 공공데이터 목록과 데이터를 통합 관리 및 제공하고 있다. 공공데이터 포털에서 제공하는 공공데이터 이용 가이드의 내용은 다음과 같다.

<표 7> 공공데이터 이용 가이드의 공공데이터 소개

구분		내용
공공데이터	공공데이터 개요	<ul style="list-style-type: none"> ○ 공공데이터란 공공기관이 만들어내는 모든 자료나 정보, 국민 모두의 소통과 협력을 이끌어내는 공적인 정보를 말함. ○ 공공데이터 포털은 각 공공기관이 보유하고 있는 공공데이터를 하나로 통합 관리하는 창구 역할을 함.
	공공데이터 목록	<ul style="list-style-type: none"> ○ 공공데이터 목록에서 데이터 명칭, 키워드, 설명 등을 포함하여 검색하거나, 데이터 유형별, 기관별로 자유로운 검색이 가능함. ○ 공공데이터를 활용한 창업을 위해, 공공데이터 목록을 통해 개방 가능한 공공데이터를 쉽게 찾을 수 있음.
제공 신청	공공데이터	<ul style="list-style-type: none"> ○ 개방된 공공데이터 외에 목록에 포함되지 않은

	제공 신청	공공데이터가 필요한 경우, 별도로 신청 가능함 ○ 제공 신청이 반려되면 분쟁조정 신청으로 한 번 더 요청 가능함.
분쟁 조정	공공데이터 제공 분쟁 조정	○ 공공데이터의 제공거부 또는 제공 중단과 관련하여, 복잡한 행정소송을 거치지 않고 간단한 분쟁 조정 절차만으로 데이터를 이용할 수 있도록 지원하는 제도 ○ 공공데이터 제공 분쟁조정 위원회 홈페이지, 이메일, FAX를 통해 신청 가능함.

- 학습자 말뭉치는 국립국어원이 업무상 작성하여 공표한 저작물로 공공데이터에 해당되지만, 공공데이터 포털에서 제공하는 일반적인 공공데이터와의 차이가 있다. 학습자 말뭉치는 학습자가 작성한 저작물로, 국립국어원에서 저작권 전체를 보유하는 경우가 아니라면, 영리 목적에 따른 이용을 허용할 수 없어 자유 이용의 범위에 대한 제한이 있으며 실제로 국립국어원의 정책 목적 또한 상업적 이용이 아닌 교육 및 연구에의 활용에 있다. 또한 학습자 말뭉치의 특성상, 해당 데이터에는 학습자의 개인 정보 또한 포함되므로, 국민에게 제공할 의무를 가지는 일반적인 공공데이터와 다르게 공개 내용에도 제한이 있을 수밖에 없다.

<표 8> 일반적 공공데이터와 학습자 말뭉치 비교

차이점	일반적 공공데이터	학습자 말뭉치
데이터 구성	○ 인구수, 수익률, 시설 현황, 사업 현황, 사고 통계와 같은 객관적 수치 및 그래프	○ 학습자가 작성한 작문 및 음성발화를 전산화하여 구성 ○ 메타 정보는 학습자의 개인 정보로 구성됨.
데이터 활용 목적	○ 상업적 이용 허용	○ 상업적 이용 제한 ○ 교육, 연구 목적의 경우 자유 이용
공개 제한	○ 웹페이지를 통해 데이터 이용자가 전체 데이터를 다운로드 사용 ○ 목록에 없는 경우 요청 시 추가 제공	○ 웹페이지에서 단어 검색, 예문 등 부분적인 이용은 허용함. ○ 전체 데이터는 이용자 서약 후 제공

(3) 적용 법령

- 언어 자원과 관련되는 법령 중 학습자 말뭉치와 관련되는 주요 법령은 다음과 같다.

<표 9> 언어 자원 관련 법령

구분	관련 법령
지능정보화 기본법	<ul style="list-style-type: none"> ○ 저작권법 ○ 공공데이터법(약칭) ○ 정보공개법(약칭)
공공데이터법(약칭)	<ul style="list-style-type: none"> ○ 저작권법 ○ 정보공개법(약칭)
정보공개법(약칭)	<ul style="list-style-type: none"> ○ 개인정보 보호법

① 지능정보화 기본법

- ‘지능정보화 기본법’([시행 2021. 6. 10.] [법률 제17344호, 2020. 6. 9., 전부 개정])은 언어 자원 관련 정책 및 제도와 관련한 법 규정이다. ‘지능정보화 기본법’에서는 데이터의 유통·활용을 비롯하여 공공지능정보화와 관련된 법적·제도적·윤리적 고려 사항에 대해서 다루고 있다.

<표 10> 지능정보화 기본법 조항별 핵심 내용

관련 조항	내용
제2장 지능정보사회 정책의 수립 및 추진체계	<ul style="list-style-type: none"> ○ 지능정보사회 계획의 수립 ○ 지능정보화 정책 등의 조정
제3장 분야별 지능정보화의 추진	<ul style="list-style-type: none"> ○ 공공지능정보화의 추진 ○ 지식재산 및 지식재산권의 보호
제4장 지능정보기술의 고도화 및 지능정보서비스의 이용촉진	<ul style="list-style-type: none"> ○ 지능정보기술의 개발 ○ 지능정보서비스의 이용촉진
제5장 지능정보화의 기반 구축	<ul style="list-style-type: none"> ○ 데이터 관련 시책의 마련 ○ 데이터의 유통·활용
제6장 지능정보사회의 기반 조성	<ul style="list-style-type: none"> ○ 정보문화의 창달과 확산 ○ 정보격차 해소 시책의 마련 ○ 지능정보사회윤리

- 학습자 말뭉치는 문자 및 음성으로 표현된 자료로, ‘데이터’에 해당되는 자료에 해당하고, ‘지능정보화 기본법’의 적용 대상에 포함된다. 그 외에도 ‘지능정보화 기본법’에서는 공공데이터의 시책과 유통·활용에 대해 규정하고 있다. 다음은 ‘지능정보화 기본법’ 조항 중 학습자 말뭉치와 관련되는 데이터의 정의와 지능정보화 진행 규정을 정리한 것이다.

<표 11> 지능정보화 기본법 학습자 말뭉치 관련 조항과 핵심 내용

조항	내용
제1장 총칙, 제2조(정의)	<ul style="list-style-type: none"> ○ “정보”란 광(光) 또는 전자적 방식으로 처리되는 부호, 문자, 음성, 음향 및 영상 등으로 표현된 모든 종류의 자료 또는 지식을 말한다. ○ “지능정보기술”이란 다음 각 목의 어느 하나에 해당하는 기술 또는 그 결합 및 활용 기술을 말한다. <ul style="list-style-type: none"> 가. 전자적 방법으로 학습·추론·판단 등을 구현하는 기술 나. 데이터(부호, 문자, 음성, 음향 및 영상 등으로 표현된 모든 종류의 자료 또는 지식을 말한다)를 전자적 방법으로 수집·분석·가공 등 처리하는 기술 다. 물건 상호간 또는 사람과 물건 사이에 데이터를 처리하거나 물건을 이용·제어 또는 관리할 수 있도록 하는 기술 라. 「클라우드컴퓨팅 발전 및 이용자 보호에 관한 법률」 제2조제2호에 따른 클라우드컴퓨팅기술 마. 무선 또는 유·무선이 결합된 초연결지능정보통신 기반 기술 바. 그 밖에 대통령령으로 정하는 기술
제1장 총칙, 제3조(지능정보사회 기본원칙)	<ul style="list-style-type: none"> ○ 국가와 지방자치단체는 지능정보사회 구현시책의 추진 과정에서 민간과의 협력을 강화하고, 민간의 자유와 창의를 존중하고 지원한다.
제3장 분야별 지능정보화의 추진, 제19조(지식재산 및 지식재산권의 보호)	<ul style="list-style-type: none"> ○ 국가기관 등은 지능정보화를 추진할 때 「지식재산 기본법」 제3조제3호에 따른 지식재산권이 합리적으로 보호될 수 있도록 해야 함 ○ 국가기관 등은 공공지능정보화를 추진할 때 지능정보 서비스를 제공하는 자 등의 「지식재산 기본법」 제

조항	내용
	<p>3조제1호에 따른 지식재산에 관한 권리 또는 이익을 침해해서는 안 됨.</p> <ul style="list-style-type: none"> ○ 권리 또는 이익을 침해받거나 침해받을 우려가 있는 자는 「정보통신 진흥 및 융합 활성화 등에 관한 특별법」 제7조제5항에 따른 실무위원회에 진정을 제기할 수 있음. ○ 다만, 저작권과 관련된 분쟁은 「저작권법」에 따른 한국저작권위원회가 조정함.
제5장 지능정보화의 기반 구축 제42조(데이터 관련 시책의 마련)	<ul style="list-style-type: none"> ○ 정부는 지능정보화의 효율적 추진과 지능정보서비스의 제공·이용 활성화에 필요한 데이터의 생산·수집 및 유통·활용 등을 촉진하기 위하여 필요한 정책을 추진해야 함. ○ 다만, 공공데이터에 관한 사항은 「공공데이터의 제공 및 이용 활성화에 관한 법률」에 따름.
제5장 지능정보화의 기반구축 제43조(데이터의 유통·활용)	<ul style="list-style-type: none"> ○ 정부는 데이터의 효율적인 생산·수집·관리와 원활한 유통·활용을 위하여 국가기관 등, 법인, 기관 및 단체와의 협력체계를 구축하고, 이를 위한 지원이 가능함. ○ 다만, 공공데이터에 관한 사항은 「공공데이터의 제공 및 이용 활성화에 관한 법률」에 따름.

② 공공데이터법(약칭)

- 데이터의 유통·활용 규정에서 언급된 ‘공공데이터의 제공 및 이용 활성화에 관한 법률’(약칭 : 공공데이터법)([시행 2020. 12. 10.] [법률 제17344호, 2020. 6. 9., 타법개정])이란, 공공기관이 보유·관리하는 ‘공공데이터’에 대한 제공 및 이용 활성화에 대한 법 규정이다. ‘공공데이터법(약칭)’은 공공데이터의 범위, 제공 절차 등에 대해 다루고 있다.

<표 12> 공공데이터법(약칭) 조항별 핵심 내용

관련 조항	내용
제2장 공공데이터 정책의 수립 등	<ul style="list-style-type: none"> ○ 공공데이터전략위원회 ○ 공공데이터 이용 활성화
제3장 공공데이터 등록 등 제공기	<ul style="list-style-type: none"> ○ 제공대상 공공데이터의 범위

반 조성	○ 공공데이터 목록의 제외
제4장 공공데이터의 제공절차 등	○ 공공데이터의 제공 ○ 분쟁의 조정

- 학습자 말뭉치는 국립국어원의 정책 과정에서 생성되어 관리되는 자료 또는 정보로 ‘공공데이터’에 해당되며, 학습자 말뭉치의 구축 및 가공을 통해, 전산화된 자료를 공유하는 것은 해당 법령에서 규정하는 ‘제공’에 해당된다. 이러한 공공데이터의 제공의 범위와 관련하여, ‘공공데이터법(약칭)’은 다음과 같이 규정하고 있다.

<표 13> 공공데이터법(약칭) 학습자 말뭉치 관련 조항과 핵심 내용

조항	내용
제1장 총칙, 제2조(정의)	<ul style="list-style-type: none"> ○ “공공기관”이란 국가기관, 지방자치단체 및 「지능정보화 기본법」 제2조제16호에 따른 공공기관을 말한다. ○ “공공데이터”란 데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보로서 다음 각 목의 어느 하나에 해당하는 것을 말한다. <ul style="list-style-type: none"> ○ “기계 판독이 가능한 형태”란 소프트웨어로 데이터의 개별내용 또는 내부구조를 확인하거나 수정, 변환, 추출 등 가공할 수 있는 상태를 말한다. ○ “제공”이란 공공기관이 이용자로 하여금 기계 판독이 가능한 형태의 공공데이터에 접근할 수 있게 하거나 이를 다양한 방식으로 전달하는 것을 말한다.
제3장 공공데이터 등 목록 등 제공기반 조성, 제17조(제공대상 공공데이터의 범위)	<ul style="list-style-type: none"> ○ 공공기관의 장은 해당 공공기관이 보유·관리하는 공공데이터를 국민에게 제공하여야 한다. 다만, 다음 각 호의 어느 하나에 해당하는 정보를 포함하고 있는 경우에는 그러하지 아니한다. <ol style="list-style-type: none"> 1. 「공공기관의 정보공개에 관한 법률」 제9조에 따른 비공개대상정보 2. 「저작권법」 및 그 밖의 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 것으로 해당 법령에 따른 정당한 이용허락을 받지 아니한 정보

③ 정보공개법(약칭)

- 공공데이터 공개 범위 규정에서 언급된 ‘공공기관의 정보공개에 관한 법률’(약칭: 정보공개법)([시행 2021. 6. 23.] [법률 제17690호, 2020. 12. 22., 일부개정])은 공공기관이 보유·관리하는 정보에 대한 제공 및 공개 의무에 대해 규정하는 법령이다. ‘정보공개법(약칭)’은 공공기관의 정보 공개 의무를 명시하는 동시에 비공개 대상 정보의 범위에 대해서 규정하고 있다.

<표 14> 정보공개법(약칭) 조항별 핵심 내용

관련 조항	내용
제2장 정보공개 청구권자와 공공기관의 의무	○ 정보공개 청구권자 ○ 공공기관의 의무
제3장 정보공개절차	○ 비공개 대상 정보 ○ 정보공개절차 청구방법
제4장 불복 구제 절차	○ 이의신청 ○ 제3자의 비공개 요청 등
제5장 정보공개위원회 등	○ 정보공개위원회의 설치 ○ 위원회의 구성 등

- 학습자 말뭉치의 제공은, 국립국어원이라는 공공기관에 의한 정보 제공이므로 ‘정보공개법(약칭)’에서 규정하는 ‘공개’에 해당되고, 관련 조항의 적용을 받는다. 다음은 학습자 말뭉치와 관련되는 ‘정보공개법(약칭)’의 조항을 발췌한 것이다.

<표 15> 정보공개법(약칭) 학습자 말뭉치 관련 조항과 핵심 내용

조항	내용
제1장 총칙, 제1조(목적)	○ “공개”란 공공기관이 이 법에 따라 정보를 열람하게 하거나 그 사본·복제물을 제공하는 것 또는 「전자정부법」 제2조제10호에 따른 정보통신망(이하 “정보통신망”이라 한다)을 통하여 정보를 제공하는 것 등을 말한다.
제3장 정보공개절차,	○ 공공기관이 보유·관리하는 정보는 공개 대상이

<p>제9조(비공개 대상 정보)</p>	<p>된다. 다만, 다음 각 호의 어느 하나에 해당하는 정보는 공개하지 아니할 수 있다. <개정 2020. 12. 22.></p> <ol style="list-style-type: none"> 1. 다른 법률 또는 법률에서 위임한 명령(국회규칙·대법원규칙·헌법재판소규칙·중앙선거관리위원회규칙·대통령령 및 조례로 한정한다)에 따라 비밀이나 비공개 사항으로 규정된 정보 2. 국가안전보장·국방·통일·외교관계 등에 관한 사항으로서 공개될 경우 국가의 중대한 이익을 현저히 해칠 우려가 있다고 인정되는 정보 3. 공개될 경우 국민의 생명·신체 및 재산의 보호에 현저한 지장을 초래할 우려가 있다고 인정되는 정보 4. 진행 중인 재판에 관련된 정보와 범죄의 예방, 수사, 공소의 제기 및 유지, 형의 집행, 교정(矯正), 보안처분에 관한 사항으로서 공개될 경우 그 직무수행을 현저히 곤란하게 하거나 형사피고인의 공정한 재판을 받을 권리를 침해한다고 인정할 만한 상당한 이유가 있는 정보 5. 감사·감독·검사·시험·규제·입찰계약·기술개발·인사관리에 관한 사항이나 의사결정과정 또는 내부검토 과정에 있는 사항 등으로서 공개될 경우 업무의 공정한 수행이나 연구·개발에 현저한 지장을 초래한다고 인정할 만한 상당한 이유가 있는 정보. 다만, 의사결정 과정 또는 내부검토 과정을 이유로 비공개할 경우에는 제13조제5항에 따라 통지를 할 때 의사결정 과정 또는 내부검토 과정의 단계 및 종료 예정일을 함께 안내하여야 하며, 의사결정 과정 및 내부검토 과정이 종료되면 제10조에 따른 청구인에게 이를 통지하여야 한다. 6. 해당 정보에 포함되어 있는 성명·주민등록번호
-----------------------	---

	호 등 「개인정보 보호법」 제2조제1호에 따른 개인정보로서 공개될 경우 사생활의 비밀 또는 자유를 침해할 우려가 있다고 인정되는 정보. 다만, 다음 각 목에 열거한 사항은 제외한다.
제3장 공공데이터 등록 등 제공기반 조성, 제17조(제공대상 공공데이터의 범위)	<ul style="list-style-type: none"> ○ 공공기관의 장은 해당 공공기관이 보유·관리하는 공공데이터를 국민에게 제공해야 함. ○ 다만, 「공공기관의 정보공개에 관한 법률」 제9조에 따른 비공개대상정보, 「저작권법」 및 그 밖의 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 것으로 해당 법령에 따른 정당한 이용허락을 받지 아니한 정보에 해당하는 정보를 포함하고 있는 경우에는 제공 대상에 포함하지 않음.

(4) 저작권 보호법과 학습자 말뭉치

○ 최근 언어 자원의 구축과 활용에서 저작권은 그 어떤 문제보다 핵심적인 쟁점이 되고 있다. 그런 점에서 저작권 보호 법률에서 저작권의 범위와 권리 및 보호에 관한 규정을 검토함으로써 학습자 말뭉치의 구축과 활용에 저작권 개념이 적용될 수 있는지를 살펴보는 것은 중요한 의미를 지닌다. 한국어 학습 과정에서 학습자들이 산출한 자료가 저작물로 인정된다면, 학습자 말뭉치의 구축 과정에서부터 저작물 계약 유형은 물론, 배포와 이용까지 고려하여 이용자의 범위, 자료 이용의 목적 등에 제한이 생길 수 있기 때문이다. 저작권 보호 관련 법률 조항 및 해외 학습자 말뭉치 사례에서의 저작권 보호와 관련한 사항을 검토한 결과 주요한 쟁점으로 정리된 사항은 다음과 같다.

- 학습자 자료의 저작물 인정 여부
- 자료 활용 목적에 따른 저작권 양도 방식
- 국외 학습자 말뭉치의 저작권 처리 방식 검토

① 학습자 자료의 저작물 인정 여부

○ 학습자 말뭉치는 학습자가 산출한 발화와 작문을 수집하여 모아 놓은 자

료이다. 따라서 학습자가 생산한 구어와 문어의 자료가 저작물로 인정되는 경우에는 국내의 <저작권법>에 준용하여 자료를 처리해야 할 것이다. 국내에서는 <저작권법> ([시행 2021. 6. 9.] [법률 제17588호, 2020. 12. 8., 일부개정])을 통해 저작물의 범위와 권리를 규정하고 있다. 다음은 국내의 <저작권법>에서 정의하는 저작권과 관련한 핵심 용어 및 내용을 발췌하여 정리한 것이다.

<표 16> 저작권법의 용어와 내용

규정	내용
제1장 총칙, 제2조 (정의)	<ul style="list-style-type: none"> ○ “저작물”은 인간의 사상 또는 감정을 표현한 창작물을 말한다. ○ “저작자”는 저작물을 창작한 자를 말한다. ○ “편집물”은 저작물이나 부호·문자·음·영상 그 밖의 형태의 자료(이하 “소재”라 한다)의 집합물을 말하며, 데이터베이스를 포함한다. ○ “편집저작물”은 편집물로서 그 소재의 선택·배열 또는 구성에 창작성이 있는 것을 말한다. ○ “데이터베이스”는 소재를 체계적으로 배열 또는 구성한 편집물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것을 말한다. ○ “배포”는 저작물등의 원본 또는 그 복제물을 공중에게 대가를 받거나 받지 아니하고 양도 또는 대여하는 것을 말한다. ○ “공표”는 저작물을 공연, 공중송신 또는 전시 그 밖의 방법으로 공중에게 공개하는 경우와 저작물을 발행하는 경우를 말한다.
제1장 제3조	<ul style="list-style-type: none"> ○ 외국인의 저작물은 대한민국이 가입 또는 체결한 조약에 따라 보호된다. ○ 대한민국 내에 상시 거주하는 외국인(무국적자 및 대한민국 내에 주된 사무소가 있는 외국법인을 포함한다)”에 해당되어 저작물을 저작권법에 의해 보호받을 수 있다.
제2장 제10조	<ul style="list-style-type: none"> ○ ‘저작권자’는 ‘저작권법’ 제2장 제10조의 규정에 따라, 저작물에 공표와 동일성 유지에 대한 ‘저작인격권’, 저작물의 양도, 배포, 복제 등의 ‘저작재산권’을 갖는다.

- 저작권법에 따르면 저작권 보호 대상이 되는 ‘저작물’의 여러 유형 중 어문저작물이 포함되어 있다. 학습자가 산출한 자료가 어문저작물의 대표적인 예인 시, 소설 등의 문학 작품과 같이 저작물로 간주될 수 있는지가 관건이 된다. 한국어 학습 과정에서 산출한 자료의 경우 일정 부분은 배운 문법과 어휘를 사용하여 통제된 양식의 글을 모방하고 또 나머지 부분은 학습자의 사고를 반영한 창작물이 되기도 하기 때문이다. 한편, 학습자 산출 자료 중 초급 학습자의 경우, 발화나 텍스트의 양이 크지 않은데, 이러한 자료를 저작물로 인정할 수 있는지 검토하기 위하여 텍스트 길이와 관련된 저작물 판결 사례를 소개한다.

<표 17> 저작권 관련 판례

판결 사례	내용
서울남부지방법원 2013년 5월 9일 선고 2012고정4449판결	○ 피고인이 이외수의 트위터 글을 무단 복제하여 “이외수 어록 24억짜리 언어의 연금술”이라는 제목의 전자책 파일을 제작한 사건에서 법원은 “일반적으로 트윗 글은 140자 이내라는 제한이 있고 신변잡기적인 일상적 표현도 많으며, 문제된 이 사건 트윗 글 중에도 문구가 짧고 의미가 단순한 것이 있기는 하지만, 이외수의 그러한 트위터 글조차도 짧은 글귀 속에서 삶의 본질을 꿰뚫는 촌철살인의 표현이나 시대와 현실을 풍자하고 약자들의 아픔을 해학으로 풀어내는 독창적인 표현형식이 포함되어 있는 것이 대부분이고, 각 글귀마다 이외수 특유의 함축적이면서도 역설적인 문체가 사용되어 그의 개성을 드러내기에 충분한 사실을 인정할 수 있다”고 판시하였다.
서울중앙지법 2018년 9월 4일 민사1003단독	○ “저작자의 개성이 창작행위에 나타나 있는지를 판단할 때에는 용어의 선택이나 전체 구성, 표현방식 등을 종합적으로 검토해야 한다”고 전제한 후에 “해당 문장은 사상과 표현, 용어 선택에 있어서 독창적인 표현 형식이 포함되었으므로, 창작성이 인정된다”고 하면서 “난 우리가 좀더 청춘에 집중했으면 좋겠어”에 대한 어문 저작물을 인정하여 이 문구를 홍보 수단으로 이용한 현대백화점에 손해배상 300만원 허락하였다.

- 위의 판례에 따라 연구진에 속한 저작권법 전문가는 학습자 자료 또한 학습자의 독창적인 글이 포함되므로 길이와 무관하게 저작물로 인정하여야 한다는 의견을 제시하였다.

② 자료 활용 목적에 따른 저작권 양도 방식

- 학습자 자료가 저작권으로 인정되는 경우, 저작물을 이용하는 데 있어서 저작권 이용 허락 및 양도에 관한 사항이 또다시 중요한 쟁점이 된다. 다음은 <저작권법>에서 이용과 관련한 핵심 내용을 발췌하여 정리한 것이다.

<표 18> 저작권법에서의 이용 및 양도에 관한 내용

규정	내용
제24조의2 제1항	○ “국가 또는 지방자치단체가 업무상 작성하여 공표한 저작물이나 계약에 따라 저작재산권의 전부를 보유한 저작물은 허락 없이 이용할 수 있다.
제24조의2 제2항	○ “국가는 「공공기관의 운영에 관한 법률」 제4조에 따른 공공기관이 업무상 작성하여 공표한 저작물이나 계약에 따라 저작재산권의 전부를 보유한 저작물의 이용을 활성화하기 위하여 대통령령으로 정하는 바에 따라 공공저작물 이용활성화 시책을 수립·시행할 수 있다.
제45조 (저작재산권의 양도)	○ 저작재산권은 전부 또는 일부를 양도할 수 있다. ○ 저작재산권의 전부를 양도하는 경우에 특약이 없는 때에는 제22조에 따른 2차적저작물을 작성하여 이용할 권리는 포함되지 아니한 것으로 추정한다. 다만, 프로그램의 경우 특약이 없는 한 2차적저작물작성권도 함께 양도된 것으로 추정한다.
제46조 (저작물의 이용허락)	○ 저작재산권자는 다른 사람에게 그 저작물의 이용을 허락할 수 있다. ○ 제1항의 규정에 따라 허락을 받은 자는 허락받은 이용 방법 및 조건의 범위 안에서 그 저작물을 이용할 수 있다. ○ 제1항의 규정에 따른 허락에 의하여 저작물을 이용할

	수 있는 권리는 저작권재산권자의 동의 없이 제3자에게 이를 양도할 수 없다.
--	--

- 학습자 말뭉치와 관련하여 저작물 이용 허락과 양도 계약 방식에서 다음의 내용을 고려할 수 있다.

<표 19> 저작권법에서의 이용 및 양도에 관한 내용

계약 형태	내용
저작물 이용허락	<ul style="list-style-type: none"> ○ 저작물 이용 허락은 비독점적 이용을 허락하는 형태로, 저작권은 학습자에게 유지되며, 이용 허락을 받은 기관이 자유롭게 이용이 가능함. ○ 공공누리 조건에 맞지 않음. ○ 제3자에게 제공하는 권한이 없다는 조항을 추가했지만 법적 효용은 검토해야 함(제3자에게 제공하는 경우 저작물 양도는 받는 것이 일반적).
저작물 양도	<ul style="list-style-type: none"> ○ 국립국어원에서 권리를 양도받게 되어 이용, 변형, 권한 재양도 등이 모두 가능함. ○ 공공누리 자료에 포함 가능함. ○ 자료 제공자(학습자) 본인이 본인의 작문을 이용할 수 없게 됨. 이러한 문제로 공모전 등에서 논란이 되는 경우가 있어 이용허락 쪽으로 계약서를 쓰는 추세임. ○ 양도 계약의 경우, 수집비(인건비)와 저작권료에 대한 비용 처리가 각각 이루어져야 함.

③ 국외 학습자 말뭉치의 저작권 처리 방식 검토

- 국외 학습자 말뭉치의 경우도 학습자의 저작권과 관련한 고려가 이루어지고 있다. 다음은 학습자의 저작권 및 자료의 이용과 관련한 사항을 명시하고 있는 말뭉치의 웹사이트를 참고하여 그 내용을 정리한 것이다.

<표 20> 국외 말뭉치에서의 저작권과 이용 내용

말뭉치	내용
International Corpus of Learner(ICLE)	<ul style="list-style-type: none"> ○ 자료 수집 과정에서 학습자 프로필(learner profile) 마지막 부분에 자료 사용 허락(본인의 에세이가 연구 목적으로 사용되는 것에 동의합니다.)을 서명 받음. ○ 학습자 본인의 작문이어야 하고 제3자로부터 도움을 받으면 안 됨. ○ 연구를 포함한 비영리 교육 목적으로만 자료 이용을 허락함.
The Longman Learners' Corpus	<ul style="list-style-type: none"> ○ 구체적인 수집 절차를 밝히고 있지는 않으나 웹사이트에 “Longman Learners' Corpus를 성장시키기 위해서는 더 많은 학생 스크립트가 필요합니다. 이 중요한 연구 프로젝트에 참여하려면 세션이 끝날 때 학생들의 논문을 보내주십시오. 100건의 제출물마다 Longman에서 강의실용 Longman 사전 사본을 보내드립니다”라는 사항을 명시하고 있음.
British Academic Written English corpus(BAWE)	<ul style="list-style-type: none"> ○ 대학교 사이트(University of Warwick)를 통해 학습자들이 직접 과제를 첨부하는 온라인 시스템을 구축함. 참여자가 제출한 과제와 정보가 온라인으로 제출된 다음 각 과제에 대한 저작권 고지 양식(copyright disclaimer form)을 자동으로 작성하는 화면으로 넘어가도록 함. 효율성을 위해, 학생들은 제출한 모든 과제에 대해 하나의 정보 페이지만 작성하도록 함. ○ 따라서 과제에 따라 달라지는 모든 필드(학년 등)는 과제가 익명화 될 때 연구 조교가 완료하거나 수동으로 작성함. ○ 제출 페이지에 첨부된 과제에는 작성자와 과제 레퍼런스가 포함되었고 '서명 대기 중'으로 저장됨. 그리고 학생들은 저작권 양식에 서명(to sign the copyright forms)하고 지불금을 받으라는 이메일을 받음. 연구팀은 연구 예산에서 참여에 대한 인센티브로 소액을 제공함.
Corpus Escrito del	<ul style="list-style-type: none"> ○ 문어는 온라인(http://learnercorpora.com/)으로 수집

말뭉치	내용
Español L2(CEDEL2)	<p>하여 전 세계 어디서나 참여할 수 있도록 함.</p> <ul style="list-style-type: none"> ○ 모든 참가자가 양식과 지침을 올바르게 이해했는지 확인하기 위해 양식은 모국어로 작성함. ○ 시험 자료의 성격(삽화를 보고 스페인어로 스토리를 적는 과제와 간단한 어휘 및 문법 테스트)을 가지므로 저작권에 대한 내용을 언급하지 않음. 자료의 이용에 대한 내용만 명시됨. ○ 데이터의 이용에 대해 “귀하의 언어 데이터는 전자 데이터베이스(=말뭉치)의 일부가 될 것이며 언어 연구 목적으로 사용됩니다. 코퍼스는 (예상) 10년 이상 동안 연구 커뮤니티에서 온라인으로 무료로 사용할 수 있습니다.”라고 동의서에 안내함.
Data Collection for Learner Corpus of Latvian(LaVA)	<ul style="list-style-type: none"> ○ a Creative Commons Attribution 4.0 International Licence 저작권 표시 규정에 따라 저작권 문제를 기술함(작문 사용 목적과 규정을 명확히 기술하는 데에 도움을 받음). ○ 라트비아 법에 근거한 저작권 규정을 설계함. ○ 학습 과정에서 작성된 작문은 저작권에 의해 보호받음 ○ 혹은 학습자와 교육기관 사이의 학습 동의에 명시된 것을 따름. ○ 학습자 작문 (혹은 일부) 배포는, 학습자의 서명 동의하에 이루어짐. ○ 학습자는 저자를 표명할 권리를 가지고, 언제든 몇 번이든 본인의 저작물을 이용할 수 있음. ○ 학생들은 언제든지 프로젝트에서 철회할 수 있으며 참여에 대한 인센티브는 제공되지 않음.
University of Pittsburgh English Language Institute Corpus (PELIC)	<ul style="list-style-type: none"> ○ 수집과 구축 과정에서 윤리적, 법적 문제를 해결하기 위하여 PITTSBURGH 대학의 기관 검토위원회 (IRB)를 통해 데이터 수집 프로세스의 모든 측면을 검토 후 승인함. ○ 학생들은 L1 (하위 레벨 학생) 또는 영어 (상위 레벨 학생)로 사전 동의에 서명하여 텍스트 사용을 허용함. ○ 모든 학생들이 프로젝트에 참여할 수 있었고 제외

말뭉치	내용
	<p>기준은 없었음.</p> <p>○ 학생들은 언제든지 프로젝트에서 철회할 수 있으며 참여에 대한 인센티브는 제공되지 않음.</p> <p>○ 코퍼스가 공개된 후 학생들에게 별도의 연락은 하지 않았으나, 프로젝트 시작 시 데이터가 공개될 것이라는 것을 공지 받음.</p>

- 저작권과 자료 이용에 관한 사항을 학습자에게 고지하는 계약서나 동의서의 양식은 대부분 공개되지 않았는데, 일부 공개되고 있는 경우를 살펴보면 다음과 같다.¹⁾

가. ICLE의 사례

<그림 1> ICLE(International Corpus of Learner English) 학습자 동의서 예시

LEARNER PROFILE

LEARNER PROFILE

=====

Text code : (do not fill in)

Essay :

Title :

Approximate length required : -500 words +500 words

Conditions : timed untimed

Examination : yes no

Reference tools : yes no

What reference tools ?

 Bilingual dictionary :

 English monolingual dictionary :

 Grammar :

1) ICLE(International Corpus of Learner English) 학습자 프로파일은 웹상에서 확인할 수 있으며, CEDEL2(Corpus Escrito del Español L2) 동의서는 참가자들이 Google Docs에서 직접 작성하여 제출할 수 있도록 하고 있다.

Other(s) :

=====

Surname :

First names :

Age :

Male

Female

Nationality :

Native language :

Father's mother tongue :

Mother's mother tongue :

Language(s) spoken at home : (if more than one, please give the average % use of each)

Education :

Primary school - medium of instruction :

Secondary school - medium of instruction :

Current studies :

Current year of study :

Institution :

Medium of instruction :

English only

Other language(s) (specify)

Both

=====

Years of English at school :

Years of English at university :

Stay in an English-speaking country :

Where ?

When ?

How long ?

=====

Other foreign languages in decreasing order of proficiency :

=====

I hereby give permission for my essay to be used for research

purposes.

Date :

Signature : ...



- ICLE의 경우는 웹사이트를 통해 학습자들이 직접 자료를 제공할 수 있으며, 이때 자료 이용 허락에 대한 동의와 함께 다음과 같은 내용의 자료 및 학습자에 관한 메타 정보를 입력하도록 하고 있다.

<표 21> ICLE의 메타 정보

구분	세부 항목
자료 정보	<ul style="list-style-type: none"> ○ 텍스트 코드, 에세이, 제목, 대략적인 분량(500자 내외) ○ 시간제한 여부, 시험 여부, 참고 자료 여부, 어떤 자료를 사용했는가?(외국어 사전, 영어 사전, 문법책, 기타)
신상 정보	<ul style="list-style-type: none"> ○ 이름, 나이, 성별 ○ 국적, 모국어, 아버지의 모국어, 어머니의 모국어, 가정에서 사용하는 언어(하나 이상인 경우, 평균 %를 표시) ○ 현재 학습 상태: 학습 기간, 교육 기관, 기관 언어, 영어만 사용, 다른 언어 (구체적으로 기입), 둘다 사용 ○ 학력: 초등교육, 중등교육
목표어 학습 경험	학교에서 영어를 배운 기간, 대학교에서 영어를 배운 기간, 영어 사용 국가 거주 기간: 장소, 시기, 기간
기타 외국어 능숙도	잘하는 것부터 내림차순으로 응답

나. Data Collection for Learner Corpus of Latvian(LaVA)의 사례

<그림 2> 동의서 및 설문 양식(Inga Kaija, Ilze A. Auzina(2020))

<p>Information letter of the project researcher group for Latvian learners</p> <p>Dear student,</p> <p>The project <i>Development of Learner Corpus of Latvian: methods, tools and applications</i> (Project No. lzp-2018/1-0527) is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The goal of the project is to create an error-annotated Latvian language learner corpus and develop corpus-based teaching materials.</p> <p>The project is financed by Latvian Council of Science; the project leader is senior researcher of IMCS UL Dr. philol. Ilze Auzina (e-mail: ilze.auzina@lumii.lv).</p> <p>What do you have to do?</p> <p>Please read carefully and sign the Permission that you agree to allow the text written during your Latvian language studies to be included in the Latvian learner corpus. Complete the questionnaire and provide the necessary information for the further use of the text in research. On the other side of the page, write an essay on the topic that the lecturer has assigned to you.</p> <p>Data storage and privacy</p> <p>Collected data will be stored at the IMCS UL on the password protected server. The data stored will be completely anonymous. A unique identifier will be assigned to each data provider.</p> <p>After the end of the project the <i>Learner Corpus of Latvian</i> will be publicly available on the corpora website of IMCS UL.</p> <p>Participation</p> <p>Participation is voluntary. Over the course of the project, you may request that texts written by you are removed from the database and refuse to participate without specifying the reason. This should be done by informing the group of researchers. In case of refusal, all materials collected will be deleted.</p> <p>On behalf of the project team of researchers, Ilze Auzina, IMCS UL senior researcher</p> <div style="display: flex; justify-content: space-around; align-items: center;">  <div style="text-align: center;">  <small>Latvijas Zinātnes padome</small> </div> </div>	<p style="text-align: center;">PERMISSION</p> <p>I agree that this text, written in 2019, can be included in the <i>Learner Corpus of Latvian</i> and, as a part of the corpus, can be made publicly available in various forms, fully or partly, with such conditions:</p> <ul style="list-style-type: none"> • I agree that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus. • I confirm that none of the data in this text can lead to identification of any existing people. • I agree that the text is anonymous and my name is not mentioned anywhere on the corpus website or its public documentation. Each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author. • The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms. • The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched unlimited amount of times. • All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.). • I will have the right to withdraw my consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. I am aware of this opportunity as a data provider. <p style="text-align: center;">INFORMATION ABOUT THE AUTHOR</p> <p>Age: _____</p> <p>Gender: _____</p> <p>Mother tongue (-s): _____</p> <p>Other languages you speak: _____</p> <p>How long have you been living in Latvia? _____</p> <p>For how many semesters have you been learning Latvian language?</p> <p><input type="checkbox"/> This is the first semester.</p> <p><input type="checkbox"/> This is the second semester.</p> <p><input type="checkbox"/> Other (please specify): _____</p> <div style="display: flex; justify-content: space-between; margin-top: 10px;"> <p>_____ Data</p> <p>_____ Signature</p> <p>_____ Name, surname</p> </div> <p style="text-align: center; margin-top: 10px;">THANK YOU!</p>
--	---

- Lava의 경우는 다음과 같은 내용으로 말뭉치 구축 프로젝트에 대한 정보와 자료 이용 허락에 대한 동의서, 학습자의 목표 언어 생산에 영향을 미칠 수 있는 정보를 중심으로 한 메타 정보를 기재하도록 하고 있다.

<표 22> Data Collection for Learner Corpus of Latvian(LaVA) 동의서 세부 내용(Inga Kaija, Ilze A. Auzina, 2020:43-45)

항목	내용
프로젝트 정보 안내	<ul style="list-style-type: none"> ○ 프로젝트에 대한 기본 정보, 수행 기관, 연락처 ○ 학습자를 위한 간략한 지침 ○ 말뭉치와 개인정보보호를 위한 서버 보안 안내 ○ 프로젝트 참여 의지 표현에 대한 설명(예. 학습자 본인이 자신의 작문을 말뭉치에서 제외하고 싶은 경우 어떻게 해야 하는지)

항목	내용
허가서/이용 동의	<ul style="list-style-type: none"> ○ 아래 7개 항목에 모두 동의하는 경우에 서명함. ○ 코퍼스가 무료로 제공되고 학문적, 교육적 목적으로 사용되는 것에 동의하고 코퍼스 구축에 참여하는 것에 대한 금전적 보상을 받지 않는 것에 동의함. ○ 본인의 작문에 실존 인물을 특정하는 정보가 없는 것을 확인함. ○ 작문이 익명화되어 제공되고 본인의 이름이 코퍼스 사이트나 공개문서 어디에도 공개되지 않을 것에 동의함. ○ 학습자 개개인에게 익명화 코드를 부여하고 저자를 드러내지 않는 대신 동일 학습자의 작문인 것을 표시하는 것에 동의함. ○ 코퍼스로 구축된 자료는 교육 자료, 논문 등 다양한 작업에 이용될 수 있다는 것에 동의함. ○ 코퍼스와 이에 포함된 모든 자료들은 무기한 공개적으로 접근가능하며, 연구자들 또한 무한정으로 열람할 수 있다는 것에 동의함. ○ 학습자는 자신의 동의를 언제든지 철회할 권리를 가지며, 동意的 철회는 철회 전 동의에 근거한 합법성 과정에 영향을 주지 않으며, 데이터 제공자로서의 기회에 대해 숙지한다는 것에 동의함. ※ 저자를 명시하는 것으로 저작권 문제를 해결하는 경우는 자주 있지만, 익명화하는 것을 기존 방침으로 삼음. ※ 학습자가 나중에 동의를 철회하는 경우, 해당 텍스트를 찾아서 삭제
메타 정보 설문	<ul style="list-style-type: none"> ○ 나이, 성별, 모국어, 외국어, 라트비아 거주 기간, 라트비아어 학습 기간 ○ 날짜, 서명, 이름

다. CEDEL2(Corpus Escrito del Español L2)의 사례

<그림 3> CEDEL2(Corpus Escrito del Español L2) 동의서

Thank you for your interest in CEDEL2. There are four parts:

- (1) basic anonymous questions about yourself (age, etc),
- (2) questions about your linguistic background,
- (3) a short text in Spanish, and
- (4) a short placement test to estimate your level of Spanish.

Before starting, please read the consent details below:

WHAT IS OUR AIM?

We are investigating how people learn a foreign language. In particular, we are interested in the way people learn certain grammatical properties and why they may be difficult to learn.

HOW WILL DATA BE COLLECTED?

In this study, you will fill in the aforementioned online forms. It will not take you more than an hour.

WHAT WILL WE DO WITH YOUR DATA?

Your linguistic data will be part of an electronic database (=corpus) and they will be used for linguistic research purposes. The corpus will be freely available online to the research community for (foreseeably) ten years or longer. For additional info on data protection, please visit the following link from the Data Protection Office of the University of Granada:
https://secretariageneral.ugr.es/pages/proteccion_datos/leyendas-informativas/_img/informacionadicionalproduccioninvestigadora/%21.

GUARANTEES FOR PARTICIPANTS:

1. Your participation is entirely voluntary.
2. You can freely withdraw from the study at any time without consequence.
3. You can ask us to withdraw your data from the study at any time in the future.
4. All the information you will provide is anonymous and at no stage are personal details or the names of the participants requested.
5. As a result of the above, you will not be able to be identified in any way by future users/researchers of the corpus.

6. You will not be subject to any degree of emotional, physical, psychological, or other harm as a result of this research.

7. You will not suffer harm or discrimination in case you decide not to participate.

More info:
 If you have any queries about this project, please get in touch with the Principal Investigator:
 Dr Cristóbal Lozano (Universidad de Granada), cristoballozano@ugr.es

If you agree with the terms above, please click on 'I agree'. If not, click on 'I do not agree' and leave this form. *
☐ I agree
☐ I do not agree

- CEDEL2의 경우 학습자가 자료를 제공하기 위해 (1) 자신에 대한 기본적인 익명 질문(나이 등)과 (2) 언어 배경에 대한 질문에 응답한 후, (3) 스페인어로 된 짧은 텍스트를 쓰고, (4) 스페인어 수준을 측정하기 위해 간단한 배치 평가에 응하는 절차를 거치게 된다. 다음은 자료 제공에 관한 서명에 앞서 학습자들이 읽어 보도록 제공되는 세부 정보이다.

<표 23> CEDEL2(Corpus Escrito del Español L2) 동의서의 세부 정보

항목	내용
목표	우리는 사람들이 외국어를 배우는 방법을 조사하고 있습니다. 특히 우리는 사람들이 특정 문법적 특성을 배우는 방식과 그들이 배우기 어려운 이유에 관심이 있습니다.
자료 수집 방법	이 연구에서는 앞서 언급한 온라인 양식을 작성하게 됩니다. 한 시간 이상 걸리지 않습니다.
자료 활용 범위	귀하의 언어 데이터는 전자 데이터베이스(=말뭉치)의 일부가 될 것이며 언어 연구 목적으로 사용됩니다. 말뭉치는 (예상) 10년 이상 동안 연구 커뮤니티에서 온라인으로 무료로 사용할 수 있습니다. 데이터 보호에 대한 추가 정보를 보려면 그라나다 대학교 데이터 보호 사무소에서 다음 링크를 방문하십시오.
참가자에 대한 보증	1. 귀하의 참여는 전적으로 자발적입니다. 2. 귀하의 결과 없이 언제든지 자유롭게 연구를 철회할 수 있습니다.

	<p>3. 귀하는 향후 언제든지 연구에서 귀하의 데이터를 철회하도록 요청할 수 있습니다.</p> <p>4. 귀하가 제공할 모든 정보는 익명이며 어떠한 단계에서도 개인 정보나 요청한 참가자의 이름이 없습니다.</p> <p>5. 위의 결과로 향후 코퍼스의 사용자/연구원이 귀하를 식별할 수 없게 됩니다.</p> <p>6. 귀하는 이 연구의 결과로 정서적, 신체적, 심리적 또는 기타 피해를 보지 않을 것입니다.</p> <p>7. 참여하지 않기로 결정하더라도 피해나 차별을 받지 않습니다.</p>
--	--

(5) 개인 정보 보호 관련 법률 규정과 학습자 말뭉치

○ 학습자 말뭉치 구축 및 활용 연구에서는 공개된 자료의 활용을 위해 학습자의 개인 정보가 수집된다. 아울러 학습자가 산출한 발화 및 텍스트 자료에도 개인을 식별할 수 있는 정보 및 사생활과 관련한 내용이 포함될 수 있다. 따라서 학습자의 개인 정보가 노출되어 개인의 인권 및 사생활에 피해가 가지 않도록 학습자의 개인 정보 보호에 각별한 주의가 필요하다. 이에, 개인 정보 보호 관련 법률 규정을 검토하고, 해외 학습자 말뭉치 사례에서 개인 정보 보호를 위한 조치 등을 살펴본 결과, 주요한 쟁점으로 정리된 사항은 다음과 같다.

- 개인 정보 보호 규정과 학습자 말뭉치의 개인 정보 유형
- 학습자 말뭉치 구축 과정에서의 개인 정보 보호 조치
- 자료 공개 시의 메타데이터 항목

① 개인 정보 보호 규정과 학습자 말뭉치의 개인 정보 유형

가. 개인 정보 보호법

○ 개인 정보의 처리 및 보호에 관한 사항은 법률 ‘개인 정보 보호법’([시행 2020. 8. 5] [법률 제16930호, 2020. 2. 4, 일부개정])으로 제정되어 있다. 다음은 ‘개인 정보 보호법’에서 규정하고 있는 개인 정보의 정의 및 수집, 이용, 제공 등과 관련한 핵심 사항을 발췌하여 정리한 것이다.

<표 24> 개인정보 보호법의 내용

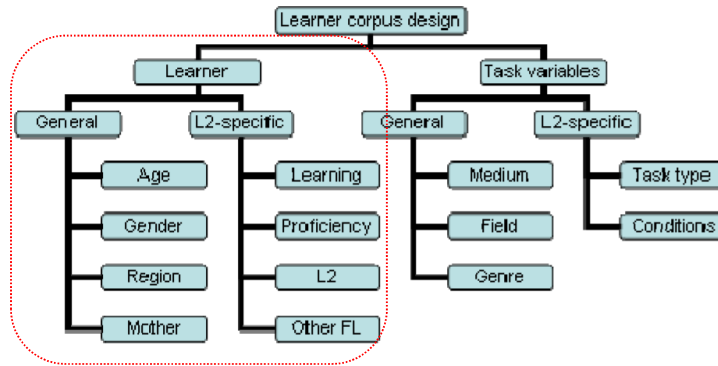
규정	내 용
제1장 총칙, 제2조 (정의)	<ul style="list-style-type: none"> ○ “개인 정보”란 살아 있는 개인에 관한 정보로서 다음 각 목의 어느 하나에 해당하는 정보를 말한다. <ol style="list-style-type: none"> 1. 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아 볼 수 있는 정보 2. 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보. 이 경우 쉽게 결합할 수 있는지 여부는 다른 정보의 입 수 가능성 등 개인을 알아보는 데 소요되는 시간, 비 용, 기술 등을 합리적으로 고려하여야 한다. 3. 가목 또는 나목을 제1호의2에 따라 가명 처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용 · 결합 없이는 특정 개인을 알아볼 수 없는 정보(이하 “가명 정보”라 한다)
제1장 제3조 (개인 정보 보호 원칙)	<ul style="list-style-type: none"> ○ 개인 정보 처리자가 개인 정보의 처리 목적을 명확하게 하고, 그 목적에 필요한 범위에서 최소한의 개인 정보만을 적법하고 정당하게 수집하여야만 한다. ○ 또한, 개인 정보의 처리 목적에 필요한 범위에서 적합하게 개인 정보를 처리하여야 하며, 그 목적 외의 용도로 활용하지 않도록 한다.
제3장 개인 정보의 처리 제15조(개인 정보의 수집 · 이용)	<ul style="list-style-type: none"> ○ 개인 정보 처리자는 다음 각 호의 어느 하나에 해당하는 경우에는 개인 정보를 수집할 수 있으며 그 수집 목적의 범위에서 이용할 수 있다. <ol style="list-style-type: none"> 1. 정보 주체의 동의를 받은 경우 2. 법률에 특별한 규정이 있거나 법령상 의무를 준수하기 위하여 불가피한 경우 3. 공공기관이 법령 등에서 정하는 소관 업무의 수행을 위하여 불가피한 경우 4. 정보 주체와의 계약의 체결 및 이행을 위하여 불가피하게 필요한 경우 5. 정보 주체 또는 그 법정대리인이 의사 표시를 할 수 없는 상태에 있거나 주소불명 등으로 사전 동의를 받을 수 없는 경우로서 명백히 정보 주체 또는 제3

규정	내 용
	<p>자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우</p> <p>6. 개인 정보 처리자의 정당한 이익을 달성하기 위하여 필요한 경우로서 명백하게 정보 주체의 권리보다 우선하는 경우. 이 경우 개인 정보 처리자의 정당한 이익과 상당한 관련이 있고 합리적인 범위를 초과하지 아니하는 경우에 한한다.</p> <p>○ 개인 정보 처리자는 제1항 제1호에 따른 동의를 받을 때에는 다음 각호의 사항을 정보 주체에게 알려야 한다. 다음 각호의 어느 하나의 사항을 변경하는 경우에도 이를 알리고 동의를 받아야 한다.</p> <ol style="list-style-type: none"> 1. 개인 정보의 수집·이용 목적 2. 수집하려는 개인 정보의 항목 3. 개인 정보의 보유 및 이용 기간 4. 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용 <p>○ 개인 정보 처리자는 당초 수집 목적과 합리적으로 관련된 범위에서 정보 주체에게 불이익이 발생하지 여부, 암호화 등 안전성 확보에 필요한 조치를 하였는지 여부 등을 고려하여 대통령령으로 정하는 바에 따라 정보 주체의 동의 없이 개인 정보를 이용할 수 있다.</p>
<p>제3장 개인 정보의 처리</p> <p>제16조(개인 정보의 수집 제한)</p>	<p>○ 개인 정보 처리자는 제15조 제1항 각호의 어느 하나에 해당하여 개인 정보를 수집하는 경우에는 그 목적에 필요한 최소한의 개인 정보를 수집하여야 한다. 이 경우 최소한의 개인 정보 수집이라는 입증 책임은 개인 정보 처리자가 부담한다.</p> <p>○ 개인 정보 처리자는 정보 주체의 동의를 받아 개인 정보를 수집하는 경우 필요한 최소한의 정보 외의 개인 정보 수집에는 동의하지 아니할 수 있다는 사실을 구체적으로 알리고 개인 정보를 수집하여야 한다.</p> <p>○ 개인 정보 처리자는 정보 주체가 필요한 최소한의 정보 외의 개인 정보 수집에 동의하지 아니한다는 이유로 정보 주체에게 재화 또는 서비스의 제공을 거부하여서는 아니된다.</p>

규정	내 용
제3장 개인 정보의 처리 제17조(개인 정보의 제공)	<ul style="list-style-type: none"> ○ 개인 정보 처리자는 다음 각호의 어느 하나에 해당되는 경우에는 정보 주체의 개인 정보를 제3자에게 제공(공유를 포함한다. 이하 같다)할 수 있다. <ol style="list-style-type: none"> 1. 정보 주체의 동의를 받은 경우 2. 제15조 제1항 제2호·제3호·제5호 및 제39조의 3 제2항 제2호·제3호에 따라 개인 정보를 수집한 목적 범위에서 개인 정보를 제공하는 경우 ○ 개인 정보 처리자는 제1항 제1호에 따른 동의를 받을 때에는 다음 각호의 사항을 정보 주체에게 알려야 한다. 다음 각호의 어느 하나의 사항을 변경하는 경우에도 이를 알리고 동의를 받아야 한다. <ol style="list-style-type: none"> 1. 개인 정보를 제공받는 자 2. 개인 정보를 제공받는 자의 개인 정보 이용 목적 3. 제공하는 개인 정보의 항목 4. 개인 정보를 제공받는 자의 개인 정보 보유 및 이용 기간 5. 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용

나. 학습자 말뭉치의 개인 정보 유형

- 학습자 말뭉치를 구축하는 과정에서 개인 정보를 수집하는 것은 자료의 활용을 고려한 것이라고 할 수 있다. Granger(2008)에서는 학습자 말뭉치 설계 시 고려해야 할 변인 정보로 다음과 같은 것들을 제안한 바 있다. 이들 정보는 크게 학습자를 특징짓는 학습자 변인과 언어 상황에 속하는 작업 변인으로 구분된다. 그리고 각 범주에는 일반(general) 변인, 즉 모든 말뭉치를 구축할 때 적용되는 변인과 학습자 말뭉치에 적합한 L2 특정(L2-specific) 변인이 포함된다. 이 중 학습자 변인은 4개의 일반 변인(연령, 성별, 지역, 모어 배경)과 L2 특정 변인(학습 맥락, 숙달도, L2 노출량 및 기타 외국어)으로 구성되어 있다.



<그림 4> 학습자 말뭉치의 변인(Granger, 2008:4)

- 이 중 과제 변인을 제외한 학습자 변인이 주로 개인 정보에 해당되며 실제 말뭉치 구축에서 다음과 같이 적용되고 있다.

<표 25> 국외 말뭉치 말뭉치에서 수집된 학습자 변인

말뭉치	내 용
International Corpus of Learner (ICLE)	<ul style="list-style-type: none"> ○ 이름, 나이, 성별, 국적, 모국어(아버지의 모국어, 어머니의 모국어, 가정에서 사용하는 언어) ○ 교육 수준, 현재 학습 상태(학습 기간, 교육 기관, 기관 언어, 영어만 사용, 다른 언어), 학교에서 영어를 배운 기간, 대학교에서 영어를 배운 기간, 영어 사용 국가 거주 기간(장소, 시기, 기간) ○ 기타 외국어 능숙도
Cambridge Learner Corpus (CLC)	<ul style="list-style-type: none"> ○ 각각의 시험 작문에 대해 참여자들의 정보가 제공됨. 이 정보에는 모국어, 나이, 성별, 영어 학습 정보와 기간, 시험 레벨 등이 포함된다.
Corpus Escrito del Español L2 (CEDEL2)	<ul style="list-style-type: none"> ○ 자신에 대한 기본적인 익명 질문(나이 등) ○ 언어 배경에 대한 질문
Data Collection for Learner Corpus of Latvian (LaVA)	<ul style="list-style-type: none"> ○ 나이, 성별, 모국어, 외국어, 거주 기간, 라트비아 고등 교육 시설에서의 학습 기간

LANGSNAP 3.0	○ 국적, 모국어, 직업, ID(식별자), 성명, 학습 목적, 수집 장소
The University of Pittsburgh English Language Institute Corpus (PELIC)	○ 성별, 연령, 모국어, 외국어, 학습자 수준, 언어 환경(집에서 사용하는 언어) ○ 학습 기간, 거주 기간, 기관 학습 기간, 학습 기관명, 전공 등
The Spoken and Written English Corpus of Chinese Learners (SWECCCL)	○ 성별, 연령, 모국어(언어 배경), 학습 단계, 목표 언어에 대한 노출 등에 대한 정보
Corpus and Repository of Writing (Crow)	○ 성별, 국적, 연령, 학습자 수준 ○ 기관 학습 기간, 학습 기관명, 전공
The AKCES/CZESL	○ 성별, 모국어, 나이, 직업

② 학습자 말뭉치 구축 과정에서의 개인 정보 보호 조치

가. 수집 과정에서의 개인 정보 보호 조치

- 다음은 국외의 학습자 말뭉치 구축을 위한 자료 수집 과정에서 학습자의 개인 정보 보호를 위해 어떠한 절차를 거쳤는지 조사한 내용을 정리한 것이다. 대다수의 말뭉치에서는 수집 동의서에 개인 정보와 관련한 사항을 명시하고 있다.

<표 26> 국외 말뭉치 수집 과정에서 개인 정보 보호를 위한 처리 예시

말뭉치	내 용
International Corpus of Learner(ICLE)	○ 자료 수집 과정에서 학습자 프로필(learner profile) 마지막 부분에 자료 사용 허락(본인의 에세이가 연구 목적으로 사용되는 것에 동의합니다.)을 서명 받음. ○ 학습자 본인의 작문이어야 하고 제3자로부터 도움을 받으면 안 됨. ○ 연구를 포함한 비영리 교육 목적으로만 자료 이용을 허락함.

말뭉치	내 용
COLT 말뭉치	<ul style="list-style-type: none"> ○ 참여 동의서에서 ‘귀하 및 귀하가 녹음한 내용에 포함된 사람들의 익명성은 충분히 보장된다는 점을 보증합니다’라는 기술을 함. ○ 이후 ‘충분한 익명성(full anonymity)’이라는 용어가 무엇을 함의하는지에 대해 논의한 결과, 모든 성과 주소를 삭제하는 쪽으로 동의가 이루어졌으나 성을 제외한 이름 부분은 변경하지 않은 채로 둬. 그러나 이후, 이 연구에서는 충분한 설명을 통한 동의라는 관점에서 볼 때 이 방침이 윤리적으로나 법적으로 문제가 있다고 판단하여 참여자들에게 제시되는 문서의 내용을 바꾸어 말뭉치로 공개될 때 성을 제외한 이름은 그대로 남을 것이라는 점을 명확히 하기로 결정함(Hasund 1998:24-5).
British Academic Written English (BAWE) corpus	<ul style="list-style-type: none"> ○ 대학교 사이트(University of Warwick)를 통해 학습자들이 직접 과제를 첨부하는 온라인 시스템을 구축할 필요한 상황별 정보의 대부분은 드롭다운 메뉴를 통해 제공되었으며 학생들은 프로젝트와 프로젝트의 목적에 관한 정보를 알 수 있었음. ○ 연구팀에서는 특정 상황별 정보를 필요로 했는데, 예를 들어, 영어 이외의 언어를 사용하는 참여자들은 그들의 모국어와 영국에서의 학습 기간을 명시해야 했음. 이 정보는 파일 헤더로 전송되어 말뭉치 사용자가 원할 경우 할당을 필터링하는 데 사용할 수 있게 됨. ○ 문맥 정보를 확인해야 하는 경우를 대비해 수집 과정에서 참여자의 연락처와 이름을 수집했으나 프로젝트 종료 시에 폐기됨.
Corpus Escrito del Español L2(CEDEL2)	<ul style="list-style-type: none"> ○ 동의서의 내용에서 ‘GUARANTEES FOR PARTICIPANTS’에 다음과 같은 사항을 명시함. <ol style="list-style-type: none"> 1. 귀하의 참여는 전적으로 자발적입니다. 2. 귀하는 결과 없이 언제든지 자유롭게 연구를 철회할 수 있습니다. 3. 귀하는 향후 언제든지 연구에서 귀하의 데이터를 철회하도록 요청할 수 있습니다. 4. 귀하가 제공할 모든 정보는 익명이며 어떠한 단계

말뭉치	내 용
	<p>에서도 개인 정보나 요청한 참가자의 이름이 없습니다.</p> <p>5. 위의 결과로 향후 코퍼스의 사용자/연구원이 귀하를 식별할 수 없게 됩니다.</p> <p>6. 귀하는 이 연구의 결과로 정서적, 신체적, 심리적 또는 기타 피해를 보지 않을 것입니다.</p> <p>7. 참여하지 않기로 결정하더라도 피해나 차별을 받지 않습니다.</p>
Data Collection for Learner Corpus of Latvian (LaVA)	<ul style="list-style-type: none"> ○ 다음 두 가지 법령을 적용하여 처리함. the European Union's new General Data Protection Regulation(GDPR, 2016), the Personal Data Processing Law(FPDAL, 2018). 두 법 모두 개인식별 정보에 따른 특정 가능성을 강조함. ○ GDPR은 개인 정보는 “개인(데이터 주체)를 식별하거나 식별가능성 있는 모든 정보”로 규정함. ○ 식별 가능한 개인이란 직접적, 혹은 간접적으로 식별 가능한 사람으로, 특히 이름, 고유 번호, 주거지, 온라인 식별자, 혹은 육체적, 심리적, 유전적, 정신적, 경제적, 문화적, 혹은 사회적 정체성을 드러내는 하나 이상의 요소와 같은 식별자에 대한 것임. ○ 개인 정보가 포함된 구축 데이터를 익명화하거나, 일체의 개인적 자료 공개를 피하는 방법으로만 가능함. ○ 정보, 이용 동의, 메타 정보 설문을 하나의 서류로 통합하여 한 면의 종이로 출력하여 서명하게 함. 이용 동의서에 개인 정보와 관련한 사항을 명시하여 이에 동의하는 경우에만 서명하게 함. ○ 본인의 작문에 실존 인물을 특정하는 정보가 없는 것을 확인함. ○ 작문이 익명화되어 제공되고 본인의 이름이 코퍼스 사이트나 공개문서 어디에도 공개되지 않을 것에 동의함. ○ 학습자 개개인에게 익명화 코드를 부여하고 저자를 드러내지 않는 대신 동일 학습자의 작문인 것을 표시하는 것에 동의함. ○ 학습자는 자신의 동의를 언제든지 철회할 권리를 가지며,

화 처리를 들 수 있다. 학습자 말뭉치 구축 과정에서 익명화 처리 대상이 될 수 있는 개인 정보는 자료 수집 과정에서 메타 정보로 수집하는 학습자 당사자에 관한 정보와 학습자가 쓴 글이나 말화에서 가족 등과 같은 사적 주제에서 드러나게 되는 정보를 조합하여 개인을 유추할 수 있게 되는 경우와 자료에 등장하는 제3자의 개인 정보 및 사생활에 대한 민감한 내용으로 상당히 포괄적이다. 이와 관련된 문제로 지적된 사례를 살펴보면 다음과 같다.³⁾

<표 27> 말뭉치에서의 개인 정보 침해 사례
(McEnery, Tony·Hardie, Andrew 지, 최재웅 역 2018:116-130)

말뭉치	사 례
BNC	<ul style="list-style-type: none"> ○ Sampson(2000:4.1절)에 따르면, BNC에서 대화 참여자들이 다른 사람의 욕, 비난을 하는 경우가 있었음. 일례로 이름이 드러난 어떤 미국 여배우의 성적 도덕성에 대해 부정적인 말을 화자들이 하고 있음. ○ 대화의 대상이 되는 사람의 경우, 화자가 아니기 때문에 이 사람은 말뭉치 구축자들에게 배포 동의와 관련한 어떤 서류에도 서명해 준 적이 없다. 따라서 언급된 내용이 무엇이냐에 따라 익명화 처리가 필요함을 알 수 있음. ○ BNC 말뭉치 파일 중 일부는 익명화가 표면적으로만 이루어진 결과, 아직도 그 익명화를 무력화시킬 만한 충분한 정보가 텍스트 내에 남아 있는 경우가 있음. 익명화 절차로 사람 이름의 성이 지워졌으나, 언급된 사람이 누구인지를 드러내기에 충분하다 할 수 있는 맥락이 제시되어 있는 경우가 있음. ○ 구어 말뭉치는 익명화를 다루는 방식에 있어서 어느 정도 무계획적인 면이 드러남. 이름은 익명화가 되어 있지만 내용 중 사생활 노출 문제에 영향을 줄 만한 다른 특질들이 익명화되지 못함.

3) McEnery, Tony • Hardie, Andrew(최재웅 역, 2018: 116-122)에서는 국외 말뭉치에서 윤리적으로 문제의 소지가 있는 내용을 소개하면서 말뭉치 구축 시에 사생활 문제를 심각하게 고려하지 못하게 될 경우 생길 수 있는 심각한 결과를 사전에 방지해야 함을 강조한 바 있다. 사례는 McEnery, Tony • Hardie, Andrew/최재웅 역(2018:116~130)을 참조하여 정리하였다.

	○ BNC의 녹음 내용은 대중에게 공개되어 있음에도 익명화가 전혀 되어 있지 않으므로 녹음 내용을 들으면 텍스트에서 익명화된 부분을 풀어낼 수 있다는 문제가 있음.
Speech Act Annotated Corpus	○ 익명화되었어야 할 신용카드 정보가 포함됨.
Lancaster Corpus of Children's Writing	○ 아이들의 사적 정보가 포함됨.

- 위와 같은 사례를 방지하기 위해서는 수집된 자료에서 노출될 수 있는 개인 정보 및 사생활 관련 내용을 익명화하는 작업이 필수임을 알 수 있다. 국외 학습자 말뭉치 구축에서 이러한 사항을 고려한 처리 예시를 살펴보면 다음과 같다.

<표 28> 국외 말뭉치 자료에서의 개인 정보 보호를 위한 처리 예시

말뭉치	개인 정보 처리 및 익명 처리 내용
University of Pittsburgh English Language Institute Corpus(PELIC) ⁴⁾	<ul style="list-style-type: none"> ○ 모든 데이터는 공개 전에 익명화되었지만 학습자 응답의 내용은 검토되거나 검열되지 않음. ○ 학습자에게 고유한 식별자(identifier)가 할당되고 모든 개인 정보가 공개 파일에서 제거되는 과정을 거침. ○ 텍스트 자체에서 학생과 교사의 개인 이름을 비식별화(ANON_NAME_0 문자열로 대체)함. 이 식별 절차는 모든 학생 및 교사의 이름과 성 목록을 사용하여 자동으로 이루어짐. 여러 개의 다른 개인 이름을 포함하는 텍스트의 경우 텍스트 일관성 및 참조를 유지하기 위해 ANON_NAME_1, ANON_NAME_2와 같은 추가 익명 식별자가 사용됨. ○ 웹 사이트 URL(ANON_URLPAGE)과 이메일 주소(ANON_EMAIL)도 변환함. 이때, 모든 익명 식별자를 ANON_로 시작하도록 표준화함으로써 학생이 생성한 텍스트로 오인되지 않고 텍스트 분석에서 쉽게 분리할 수 있게 함.
LANGSNAP 3.0 ⁵⁾	○ 참가자의 신원을 보호하기 위해 인터뷰를 익명화함.

말뭉치	개인 정보 처리 및 익명 처리 내용
	<ul style="list-style-type: none"> ○ ‘Name’, ‘City’ 의 간단한 코드로 이름, 장소, 기관 등을 익명화함.
Data Collection for Learner Corpus of Latvian (LaVA) ⁶⁾	<ul style="list-style-type: none"> ○ 개인 정보 보호와 관련하여 작문의 주제로 인하여 개인 정보가 드러날 가능성과 학습자가 쓴 글의 필체로 인하여 개인을 유추할 가능성을 고려함. ○ 작문의 주제가 막대한 개인 정보를 작성하도록 요구하는 경우가 많고, 익명화는 부담이 큰 과제가 되기 쉽기 때문에 학습자들에게 일체의 개인 정보를 포함하지 않고, 개인 및 사생활 관련 주제일 경우에는 가상의 인물에 대한 것으로 대체하도록 함. ○ 텍스트를 수집한 교사는 프로젝트에 사용된 저작권 및 개인 데이터 보호 시스템에 대해 학생들에게 지시하고 특히 주제에 관계없이 텍스트에 실제 개인 정보가 포함되어서는 안 된다는 것을 상기시킴. 주제가 이 아이디어와 모순되는 경우(예: “내 친구와 가족”), 학생들은 가상의 인물에 대해 글을 쓰거나 실제 정보를 거짓 정보로 바꾸라는 지시를 받음. ○ 실존 인물에 대한 조건은 학생들에게 안내를 하면서 특히 강조했으며, 개인 정보가 무엇인지 분명하게 이해하지 못하는 학생이 있으므로 프로젝트에 참여한 교사가 해당 상황을 설명하고 어떤 정보로 대체할 것인지 도와주도록 함. ○ 필체의 문제는 말뭉치 공개에서 작문 스캔 파일을 포함하는 경우, 학습자 필체로 인한 개인이 노출될 가능성에 대해서 고려함. ○ 비슷한 필체의 작문을 쓴 학습자가 매우 많으며, 해당 코퍼스는 1000건 이상의 작문으로 구축되었고, 한 학습자는 한 학기에 최대 두 학기까지만 작문을 제출하였으므로 모든 학습자가 두 번씩 참여했다 하더라도, 최소 500명의 저자가 존재함. 따라서 필체만으로 누군가를 특정하기는 매우 어려움. ○ 학습자에 대해 제공되는 실제 정보는 성별, 외국어, 작문 당시 대략적 연도의 메타 정보뿐이기 때문에 학습자 필체와 메타 정보를 결합하여 특정인을 도출할 가

말뭉치	개인 정보 처리 및 익명 처리 내용
	<p>능성이 충분하지 못하다고 판단함. 그러할 가능성을 최소화하기 위하여 익명성을 유지하고 개인 정보 요소를 보호함.</p> <ul style="list-style-type: none"> ○ 학습자가 제출한 텍스트 원본은 디지털화한 후 교사에게 반환되고, 텍스트를 다시 학생들에게 전달하여, 개인 정보를 실수로 포함하였는지 등을 확인하고 필요한 경우 권한 취소 가능성에 대해 다시 한번 상기시키도록 함. ○ 코퍼스는 위의 그림과 같이 업로드, 디지털화, 주석 및 검색을 위한 단일 인터페이스를 제공하는 통합 다기능 플랫폼(그림 참고, Inga Kaija, Ilze A. Auzina(2020:46))에 구축됨. ○ 웹 플랫폼은 에세이의 스캔 사본을 저장하고, 두 개의 독립적인 디지털 텍스트를 입력하고 수정한 텍스트를 비교한 후, 자동으로 형태 주석 및 오류 주석이 달린 텍스트를 편집하고, 주석 간 일치를 만드는 데 사용할 수도 있음. ○ 메타데이터가 포함된 수집된 에세이는 손으로 쓴 것으로, 추가 데이터 처리 단계를 위해 디지털화함. <div data-bbox="532 1093 1153 1450"> <p>The screenshot shows the 'Report index' page of the LaVA system. It displays a table with 5 columns: 'Id', 'Image names', 'Actions', 'Original text', 'Corrected text', and 'Error annotations'. The table lists three reports with IDs 235, 234, and 233. Each report row has buttons for 'Add meta', 'First', 'Second', and 'Final' under the 'Original text' and 'Corrected text' columns. Below the table, four arrows point from the columns to descriptive labels: 'Image names' points to 'Scanned texts, Metadata agreements'; 'Original text' points to 'Digitization of the texts'; 'Corrected text' points to 'Correction of the texts'; and 'Error annotations' points to 'Error annotation'.</p> </div> <p>디지털화는 3단계로 진행됨: (1) 과제 및 에세이 스캔; (2) 메타데이터 입력; (3) 디지털 형식으로 텍스트 다시 쓰기. 과제의 스캔 이미지는 문제가 발생할 경우 데이터 정확성을 검증하는 데 도움이 됨. 메타데이터</p>

말뭉치	개인 정보 처리 및 익명 처리 내용
	는 수동으로 입력되며, 익명성을 유지하기 위해 저자의 이름은 포함되지 않음.

③ 자료 공개 시의 메타데이터 항목

- 자료의 공개와 활용을 위해 이용자들에게 학습자의 메타데이터가 제공되게 된다. 메타데이터는 학습자의 개인 정보와도 밀접한 관련이 있기 때문에 보호 측면을 강화하여 이를 최소화하는 것이 바람직하지만 제한 범위에 따라 활용 측면에서 데이터의 유용성에 대한 문제가 제기될 수 있다. 다음은 국외 학습자 말뭉치 공개, 제공되고 있는 학습자의 메타데이터를 정리한 것이다.

-
- 4) Naismith, B., Han, N.-R., & Juffs, A. *University of Pittsburgh English Language Institute Corpus (PELIC)* (in Press)
- 5) Tracy-Ventura, N., & Huensch, A. (2018). The potential of longitudinal learner corpora in SLA research. In A. Gudmestad & A. Edmonds (Eds.), *Critical Reflections on data in second language acquisition*. pp.16-18.
- 6) Inga Kaija, Ilze A. Auzina. (2020). *Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection*. Political Science, pp.42-46.

<표 29> 국외 학습자 말뭉치의 메타데이터

말뭉치	수준	국적	성별	모국어	L3 ⁷⁾	교포여부	언어환경 ⁸⁾	나이	수집년도	출생년도	한글기간	거주기간	현재거주지역	기관한글기간	한글기관명	교육수준	직업	전국	원국	메타정보 작성일자
스페인어 학습자 코퍼스	○	○	○	○				○	○					○						○ (모두 가능)
러시아학습자 코퍼스	○	○	○	○	○	○	○	○						○				○		○
ICLE(International Corpus of Learner English)		○	○	○	○		○	○			○	○		○	○					○ (모두 가능)
The Jinan Chinese Learner Corpus (JCLC)	○	○	○	○	○	○		○	○		○					○				

7) Other Acquired Languages

8) 집에서 사용하는 언어, 부모님 L1, 중고등학교 언어 등

말뭉치	수준	구분	성별	모국어	L3 ⁷⁾	교과영역	언어환경 ⁸⁾	나이	수집연도	출생연도	학습기간	거주기간	현재거주지역	기관학습기간	학습기관명	교육수준	직업	전국	원문	메타정보신원부
The AKCES/CZESL (Acquisition corpora of Czech/Czech as a second language) corpus				○				○	○										○	
Corpus and Repository of Writing (Crow)	○	○							○					○				○		○
The University of Pittsburgh English Language Institute Corpus (PELIC)	○			○	○		○	○		○		○						○		

말뭉치	수준	구분	성별	머무는 국가	L3 기	교과서	언어 환경 ⁸⁾	나이	수집된 출생연도	학습기간	거주기간	현재거주지역	기관 학습기간	학습기관명	교육수준	직업	전국	문헌 제	메타 정보 신원 부 지 명
The Bilingual Corpus of Chinese English Learners (BICCEL)	○	○	○														○		
The Spoken and Written English Corpus of Chinese Learners (SWECCCL)	○	○	○	○													○		
The Taiwanese Corpus of Learner English (TLCE)	○	○		○				○											

말뭉치	수준	구분	정렬	머문어	L3	교과영과	언어활용 ⁸⁾	나이	수집연도	출생연도	학습기간	거주기간	현재거주지역	기관학습기간	학습기관명	교육수준	직업	전공	원문제	메타정보신원부
The TELEC Secondary Learner Corpus (TSLC)		○																		
The Hong Kong University of Science & Technology (HKUST) learner corpus		○																		
The Japanese Learner English Corpus (NICT JLE)	○	○	○						○			○					○			
The Gachon Learner Corpus	○	○	○	○	○	○				○	○	○						○		

말뭉치	수준	구분	정렬	머문어	L3 ⁷⁾	교과영과	언어활용 ⁸⁾	나이	수집연도	출생연도	학습기간	거주기간	현재거주지역	기관학습기간	학습기관명	교육수준	직업	전국	문권	메타정보신원부여
The Michigan Corpus of Upperlevel Student Papers (MICUSP)	○		○	○				○										○		○
The Cambridge Learner Corpus (CLC)	○			○				○												
The Longman Learners' Corpus	○	○																		
Data Collection for Learner Corpus of Latvian (LaVA)			○	○	○			○				○		○						
LANGSNAP 3.0		○																		

(6) IRB 관련 규정과 학습자 말뭉치

- 국외뿐 아니라 국내에서도 IRB와 관련한 규정이 확대·강화됨에 따라, 인간을 대상으로 한 연구에서는 IRB(Institutional Review Board)의 심의를 받은 후, 규정에 따라 연구를 진행해야 한다. 국내의 경우 「생명윤리 및 안전에 관한 법률」에 의거하여 IRB 심의를 받도록 요구되고 있다. 학습자 말뭉치 구축 및 활용 연구의 경우, 학습자의 자료를 수집 대상으로 하며, 수집 과정에서 활용을 위한 개인 정보 수집이 이루어진다는 점에서 IRB 준용 문제가 쟁점이 된다. 이에 따라, 학습자 말뭉치 구축 및 활용 연구의 성격을 고려한 IRB 준용에 대한 면밀한 검토가 필요하다.

① IRB의 내용과 기능

- IRB(Institutional Review Board, 기관생명윤리심사위원회/임상시험심사위원회)의 내용과 기능을 정리하면 다음과 같다.

<표 30> IRB의 내용과 기능

	내 용
정의	○ 인간 피험자가 참여하는 생물의학 연구에 대해 검토하고 시작을 승인하며, 주기적으로 확인하기 위해 시험기관이 공식적으로 지정한 심사위원회나 단체
내용	○ IRB는 인간 피험자 보호를 위한 주요 보호 수단의 하나로 피험자 동의서와 함께 실험 피험자에 대한 중요한 보호 수단이 된다. 따라서 임상시험을 비롯하여 인간을 대상으로 하는 연구에서는 시험 기관에 속해 있는 기관의 IRB 승인 후 연구를 시작할 수 있다.
기능	○ 연구의 윤리적·과학적 타당성을 심의 - 기관위원회의 주 업무는 연구 시작 전에 연구자가 작성한 연구계획서를 심의(연구 결과를 심의하는 것이 아님) - 연구계획서의 윤리적·과학적 타당성을 심의 - 과학적 타당성은 연구대상자의 보호와 연관된 윤리적 측면에 중점을 두고 심사 ○ 연구대상자 등으로부터 적법한 절차에 따라 동의를 받았는지 여부 심의

	<ul style="list-style-type: none"> - 동의서 양식이 적절하고 적법한지 여부 확인 - 동의서나 다른 설명서 등을 통해 해당 연구에 참여하는 연구 대상자에게 연구에 대한 충분한 설명이 이뤄질 수 있는지 여부 확인 - 위력이나 권위로 피험자를 모집하는지 여부 등
○ 연구대상자 등의 안전에 관한 사항	<ul style="list-style-type: none"> - 위기 및 돌발 상황 발생 시 대처 계획 수립 여부 등 확인
○ 연구대상자 등의 안전에 관한 사항	<ul style="list-style-type: none"> - 익명화, 암호화 등 개인정보보호 대책 수립 여부 확인 - 연구종료 후 개인 정보 처리 방안 등 검토

② IRB 심의 대상

- 연구의 유형에 따라서 IRB의 심의 대상과 심의 면제 대상으로 구분된다.⁹⁾ 먼저 IRB 심의 대상은 크게 인간 대상 연구와 인체유래물¹⁰⁾ 연구로 나뉜다. 인간 대상 연구는 “사람을 대상으로 물리적으로 개입하거나 의사소통, 대인 접촉 등의 상호작용을 통하여 수행하는 연구, 또는 개인을 식별할 수 있는 정보를 이용하는 연구로서 보건복지부령으로 정하는 연구”(「생명윤리 및 안전에 관한 법률」 제2조 제1호에 근거)를 말한다. 보건복지부령의 인간 대상 연구의 범위는 다음과 같다.

<표 31> 보건복지부령의 인간 대상 연구의 범위

법령	내 용
시행규칙 제2조(인간 대상 연구의 범위)	<p>① 「생명윤리 및 안전에 관한 법률」(이하 “법”이라 한다) 제2조 제1호에서 “보건복지부령으로 정하는 연구”란 다음 각호의 연구를 말한다.</p> <p>1. 사람을 대상으로 물리적으로 개입하는 연구: 연구대상자를 직접 조작하거나 연구대상자의 환경을 조작하여 자료를 얻는 연구</p>

9) IRB 심의는 시험기관에 속해 있는 기관의 승인을 받게 되므로 이 연구에서는 연세대학교 IRB(<https://irb.yonsei.ac.kr/>)의 내용을 참고하여 정리함.

10) ‘인체유래물’이란 인체로부터 수집하거나 채취한 조직·세포·혈액·체액 등 인체구성물 또는 이들로부터 분리된 혈청, 혈장, 염색체, DNA, RNA, 단백질 등(「생명윤리 및 안전에 관한 법률」 제2조제11호)을 일컬으며, 이와 같은 인체유래물을 직접 조사·분석하는 연구를 ‘인체유래물 연구’라 한다(법 제2조제12호). 학습자 맞춤형 연구는 인체유래물 연구에 포함되지 않으므로 이에 대해서는 다루지 않도록 한다.

	<p>2. 의사소통, 대인 접촉 등의 상호작용을 통하여 수행하는 연구: 연구대상자를 직접 조작하거나 연구대상자의 환경을 조작하여 자료를 얻는 연구</p> <p>3. 개인을 식별할 수 있는 정보를 이용하는 연구: 연구대상자를 직접·간접적으로 식별할 수 있는 정보를 이용하는 연구</p> <p>② 제1항에도 불구하고 다음 각호의 연구는 제1항 각호의 연구에 포함되지 아니한다.</p> <p>1. 국가나 지방자치단체가 공공복지나 서비스 프로그램을 검토·평가하기 위해 직접 또는 위탁하여 수행하는 연구</p> <p>2. 「초·중등 교육법」 제2조 및 「고등교육법」 제2조에 따른 학교와 보건복지부장관이 정하여 고시하는 교육기관에서 통상적인 교육실무와 관련하여 하는 연구</p> <p>③ 제2항 각호의 연구를 하는 연구자는 필요하다고 판단되는 경우 법 제10조 제3항 제1호 각 목의 사항에 대하여 다음 각호의 위원회에 심의를 요청할 수 있다.</p> <p>1. 법 제10조에 따른 기관생명윤리위원회(이하 "기관위원회"라 한다).</p> <p>2. 법 제12조에 따른 공용기관생명윤리위원회(이하 "공용위원회"라 한다).</p>
--	---

③ 인간 대상 연구의 유형

- 인간 대상 연구는 중재(intervention) 연구, 상호작용(interaction)을 통한 연구, 개인 식별 정보를 포함한 연구, 복합적으로 수행되는 연구 유형으로 나눌 수 있는데, 각 유형별 연구 내용을 살펴보면 다음과 같다.

<표 32> 인간 대상 연구의 유형

연구의 유형	내 용
중재(intervention) 연구	<p>○ 사람을 대상으로 물리적으로 개입하는 연구란, “연구”를 위해 연구대상자에게 어떤 침습적 행위(식품, 의약품 등의 섭취, 혈액채취 등)를 하거나 연구대상자의 환경을 조작(시각, 청각 등에 자극 또는 스트레스 유발) 물리적 개입이 포함된 연구를 수행하고 그 결과를 얻어 연구에 이용하는 것으로써 다음의 연구들이 해당할 수 있다.</p>

	<ul style="list-style-type: none"> ① 「약사법 시행규칙」 내지 「의료기기법 시행규칙」에 따라 승인된 임상시험계획서에 따라 수행되는 의약품 또는 의료기기를 이용한 임상시험 ② 식품의약품안전청 고시 「생물학적 동등성시험 관리기준」에 따라 시험기관에서 수행되는 생물학적 동등성시험 ③ 그 밖에 화장품·건강기능식품·생의약품·생물학적 제제 등에 대한 안전성·효능·효과를 보기 위해 해당 물질을 직접 연구대상자에게 적용한 후 그로부터 얻은 정보를 이용하는 연구 ④ 그 밖에 소음, 물리적 자극 등으로 연구대상자의 환경을 조작하여 얻은 정보를 이용하는 연구 등 실험적 연구
상호작용(interaction)을 통한 연구	<ul style="list-style-type: none"> ○ 의사소통, 대인접촉 등의 상호작용을 통하여 수행하는 연구란, “연구”를 위해 연구대상자를 선정하고 연구대상자의 대면을 통한 설문조사나 행동관찰 등으로 자료를 얻어 그 정보를 이용하는 연구로 다음과 같은 연구의 유형을 말한다. ① 연구를 위해 연구대상자의 행동관찰 등을 수행하여 자료를 얻는 연구 ② 연구를 위해 연구대상자를 대면하며 설문조사 등을 통해 자료를 얻는 연구 ③ 그 밖에 연구를 위해 연구대상자를 접촉하고 조사 및 관찰 등을 수행하는 연구
개인 식별 정보를 포함한 연구	<ul style="list-style-type: none"> ○ 개인 식별 정보를 포함한 연구란, 위의 두 연구 유형처럼 연구대상자를 직접 대면하거나 연구대상자로부터 정보를 직접 수집하지는 않지만, 연구대상자를 직·간접적으로 식별할 수 있는 자료를 이용하는 연구를 말한다.

- 한편, 생명윤리법상 인간 대상 연구 또는 인체유래물 연구라 할지라도, 기관위원회의 심의를 면제할 수 있는 경우가 있다. 아는 「생명윤리 및 안전에 관한 법률」 제13조에 규정되어 있으며, 연세대학교 IRB에서는 제15조 제2항 내지 제36조 제2항에 따라 “연구대상자(또는 인체유래물기증자) 및 공공에 미치는 위험이 미미한 경우”에 한하여 “국가위원회 심의를 거

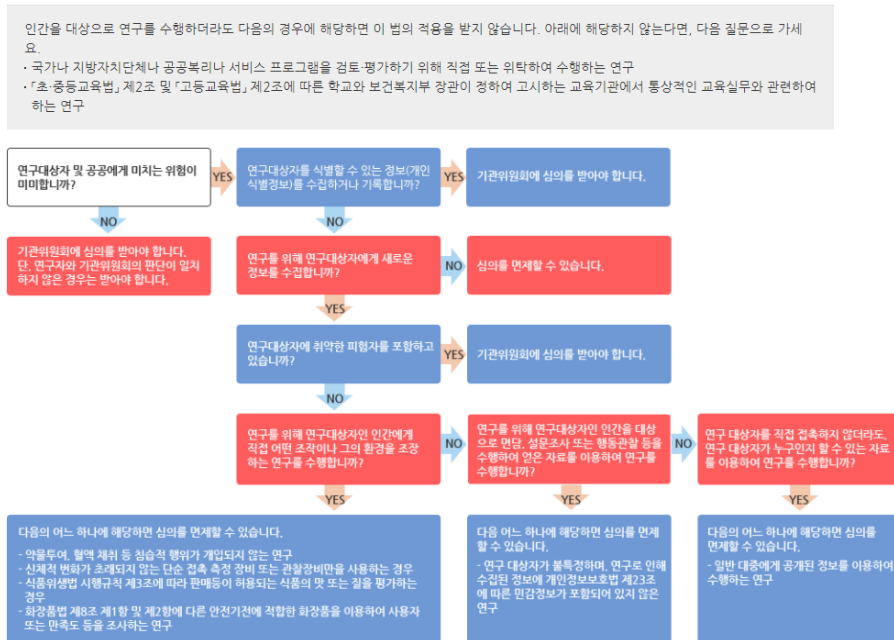
처 보건복지부령으로 정하는 연구”는 심의를 면제할 수 있다는 점을 명시하고 있다.¹¹⁾

<표 33> IRB의 심의를 면제할 수 있는 인간 대상 연구

법령	내 용
제13조	<p>① 법 제15조 제2항에서 “보건복지부령으로 정한 기준에 맞는 연구”란 일반 대중에게 공개된 정보를 이용하는 연구 또는 개인 식별정보를 수집·기록하지 않는 연구로서 다음 각호의 어느 하나에 해당하는 연구를 말한다.</p> <p>1. 연구대상자를 직접 조작하거나 그 환경을 조작하는 연구 중 다음 각 목의 어느 하나에 해당하는 연구</p> <p>가. 약물 투여, 혈액채취 등 침습적(侵襲的) 행위를 하지 않는 연구</p> <p>나. 신체적 변화가 따르지 않는 단순 접촉 측정 장비 또는 관찰 장비만을 사용하는 연구</p>

11) IRB에서는 인간을 대상으로 한 연구 중 IRB 심의 면제가 되는 연구인지의 여부를 확인할 수 있도록 다음과 같은 체크 리스트를 두고 있다.

● 인간대상연구 심의면제 CHECKLIST



법령	내 용
	<p>다. 「식품위생법 시행규칙」 제3조에 따라 판매 등이 허용되는 식품 또는 식품첨가물을 이용하여 맛이나 질을 평가하는 연구</p> <p>라. 「화장품법」 제8조에 따른 안전기준에 맞는 화장품을 이용하여 사용감 또는 만족도 등을 조사하는 연구</p> <p>2. 연구대상자 등을 직접 대면하더라도 연구대상자 등이 특정되지 않고 「개인 정보 보호법」 제23조에 따른 민감 정보를 수집하거나 기록하지 않는 연구</p> <p>3. 연구대상자 등에 대한 기존의 자료나 문서를 이용하는 연구</p> <p>② 제1항에도 불구하고 제1항 제1호 및 제2호의 연구 중 「약사법 시행규칙」 별표 3의2 제2호 더목에 따른 취약한 환경에 있는 피험자(Vulnerable Subjects)를 대상으로 하는 연구는 기관위원회의 심의를 받아야 한다.</p>
법 제15조 제2항 내지 시행규칙 제13조	<p>기관위원회의 심의를 면제할 수 있는 인간 대상 연구는 일반대중에게 공개된 정보를 이용하는 연구이거나, 연구대상자에 대한 개인식별정보를 수집하거나 기록하지 않는 연구로서 다음에 해당하는 연구를 말한다.</p> <p>① 연구대상자로 "취약한 환경의 피험자"를 포함하지 않는 연구로서 연구대상자를 직접 조작하거나 연구대상자의 환경을 조작하여 얻은 자료(data)를 이용하는 연구라 할지라도,</p> <p>1. 약물 투여나 혈액채취 등의 침습적 행위가 개입되지 않은 연구</p> <p>2. 신체적 변화가 초래되지 않는 단순 접촉 측정 장비 또는 관찰 장비만을 사용하는 연구</p> <p>3. 「식품위생법」 시행규칙 제3조에 따라 판매 등이 허용되는 식품을 이용하여 맛 또는 질을 평가하는 연구</p> <p>4. 「화장품법」 제8조 제1항 및 제2항에 따른 안전기준에 적합한 화장품을 이용하여 사용감 또는 만족도 등을 조사하는 연구는 심의를 면제한다.</p> <p>② 연구대상자로 "취약한 환경의 피험자"를 포함하지 않는 연구로서 의사소통이나 대인 접촉 등의 상호작용 즉, 연구대상자 대면을 통한 설문조사나, 연구대상자의 행동관찰 등을 통해 얻은</p>

법령	내 용
	<p>자료(data)를 이용하는 연구라 할지라도, 그 연구대상자가 불특정하며, 연구대상자로부터 "민감정보"를 수집하거나 기록하지 않는 연구는 심의를 면제한다.</p> <p>③ ①과 ②에 해당하지 않더라도, 연구대상자를 직접 또는 간접적으로 식별할 수 있는 정보를 포함하고 있는 정보(information)를 이용하는 연구로서 이때 연구대상자 등에 관한 정보가 이미 생성된 기존의 자료나 문서를 이용하는 연구는 심의를 면제한다.</p>

1.1.3. 종합 및 적용

- 저작권 관련 정책 및 제도, 법률 검토 결과는 2차 중장기 계획 수립의 측면에서 크게 저작권, 개인 정보 보호, IRB로 구분하여 정리해 볼 수 있다.

① 저작권

- 국내 법령과 판결 사례를 기반으로 살펴보았을 때, 학습자가 산출한 자료와 이를 토대로 구축된 학습자 말뭉치 또한 저작물로 인정받을 여지가 있다. 국외의 학습자 말뭉치의 경우도 이에 대한 문제의식을 가지고는 있으나 합의에 이르지 못하고 있는 것으로 보인다. 이는 LaVA와 같이 구체적으로 저작권법에 근거하여 저작권에 의해 학습자의 자료가 보호된다는 사항을 엄격하게 명시하고 있는 경우와 자료의 제공과 이용에 관한 동의를 구하는 방식을 취하는 경우, 이에 대한 언급조차 없는 경우가 다양하게 있음을 통해 드러난다. 장기적으로 국가 언어 자원으로써 자료의 활용도 제고라는 측면에서 전자의 방식이 자료 수집과 이용에 관한 문제를 합법적으로 해결할 수 있다는 점에서 저작권법에 따라 자료를 수집하는 것이 원칙이 되어야 하겠지만, 자료 수집과 구축의 효율성을 고려하여 후자의 방식을 취하는 것이 현실적이라고 판단된다.
- 본 연구에서는 저작권법의 해석 결과에 따라 학습자 말뭉치 자료를 저작물로서 간주할 수 있는 가능성이 있다는 견해에는 동의하나, 자료 수집과 구축에 관한 현실적 문제와 함께 학습자가 산출한 자료가 대개 한국어 학습 과정에서 산출한 작문/말하기 자료, 또는 본 연구에서 기획한 과제에

따라 산출한 작문/말하기 자료이며, 학습자가 산출한 자료 자체는 대개 교육과정 중에 연습 또는 평가 등 학습 과정에서 생산되는 결과물로서 학습자에게 재산상의 이익을 주지 않는다는 점을 근거로 하여 지적재산권을 가진 저작물로서 간주하는 것은 유보하는 입장을 취하기로 하였다.

- 그럼에도 원칙에 따라 학습자가 산출한 자료를 저작물로 간주하고 저작권 보호의 대상으로 처리한다면 그 저작권을 양도하는 방식과 이용 허락의 방식으로 계약을 체결할 수 있다. 두 가지 유형에 따른 차이를 정리해 보면 다음과 같다.

<표 34> 저작재산권 양도와 이용허락의 차이

구분	이용 허락	양도
관련 법령	<p>저작권법 제46조(저작물의 이용허락)</p> <p>① 저작재산권자는 다른 사람에게 그 저작물의 이용을 허락할 수 있다.</p> <p>② 제1항의 규정에 따라 허락을 받은 자는 허락받은 이용 방법 및 조건의 범위 안에서 그 저작물을 이용할 수 있다.</p> <p>③ 제1항의 규정에 따른 허락에 의하여 저작물을 이용할 수 있는 권리는 저작재산권자의 동의 없이 제3자에게 이를 양도할 수 없다.</p>	<p>저작권법 제45조(저작재산권의 양도)</p> <p>① 저작재산권은 전부 또는 일부를 양도할 수 있다.</p> <p>① 저작재산권의 전부를 양도하는 경우에 특약이 없는 때에는 제22조에 따른 2차적 저작물을 작성하여 이용할 권리는 포함되지 아니한 것으로 추정한다. 다만, 프로그램의 경우 특약이 없는 한 2차적 저작물 작성권도 함께 양도된 것으로 추정한다.</p>
저작재산권 보유	학습자 본인	구축 기관
이용 범위	교육 및 연구 목적까지의 이용 가능	이용에 제한이 없음
권한 재양도	불가능(필요 시 별도 양도)	가능
공공누리 등재	불가능	가능
본인 작문	가능	불가능(학습자 본인에게 이용

구분	이용 허락	양도
이용		권한 없음)
문제점	공공누리 등재가 제한됨	자료 제공자의 권리가 제한됨

- 국외 학습자 말뭉치의 경우, 연구 및 교육 목적을 위한 비영리 자료로만 그 이용을 허용하고 있는데, 이는 어떤 방식을 취하든 내용상으로는 저작권 양도가 아닌 이용 허락을 전제로 함을 의미한다. 이 경우 저작권권이 학습자 당사자에게 있기 때문에 말뭉치의 이용 범위가 제한될 수 있다. 본 연구에서는 향후 저작권법을 적용한 학습자 자료 수집 및 구축으로의 전환을 전제로 양도의 방식으로 자료 제공 및 이용에 관한 계약서를 작성하되, 계약서라는 서식이 주는 학습자의 심리적 부담을 고려하여 계약서의 제목이나 내용을 완화하여 기술하기로 하였다. 아울러 자료 이용 과정에서 불가피하게 발생할 수 있는 저작권 침해의 문제를 최소화하기 위하여 이용자 서약서에 관련 사항을 명시하였다.
- 다음은 실제 적용을 위해 국외의 학습자 말뭉치 수집 과정에서 수집 동의 및 저작권과 관련된 방침을 정리한 것이다.

<표 35> 국외 말뭉치의 수집 및 이용에 관한 사항

말뭉치	수집 및 이용 동의	이용 및 배포 사항
International Corpus of Learner(ICLE)	O	연구를 포함한 비영리 교육 목적으로만 자료 이용 허락
British Academic Written English (BAWE) corpus	O (온라인 수집)	-
CEDEL2(Corpus Escrito del Español L2)	O (온라인 수집)	연구 커뮤니티에서 온라인으로 무료 공개라고 명시함
Data Collection for Learner Corpus of Latvian (LaVA)	O	교육자료, 논문 등 다양한 작업에 이용될 수 있다는 것을 명시하고 이에 동의하게 함
University of Pittsburgh English	O	학습자들은 프로젝트 시작 시 데이터가 공개될 것이라는

Language Institute Corpus (PELIC)		것을 공지 받음
--------------------------------------	--	----------

② 개인 정보

- ‘개인정보 보호법’에서는 개인 정보의 개념에서부터 개인 정보 보호 원칙, 수집 및 이용, 제공 등에 관한 규정을 다루고 있으므로 학습자 말뭉치 수집 동의서에 개인 정보와 관련한 항목을 명시할 시에 법령을 참조, 준용할 수 있다.
- 개인 정보는 ‘식별되거나 식별 가능한 개인에 관한 모든 정보’를 의미하므로, 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것 또한 개인 정보의 정의에 포함된다. 따라서 학습자가 산출한 작문이나 발화에서는 학습자의 인적 사항, 가족관계 및 가족 구성원 정보 등이 드러나게 되는 경우가 있으므로 개인 정보법에 저촉되지 않도록 학습자 자료 수집과 말뭉치 구축 과정에서 개인 정보 수집 및 처리에 대한 사항을 유의할 필요가 있다.
- 학습자 말뭉치의 경우 특히, 성별, 국적, 연령, 모국어와 같은 학습자 변인이 말뭉치 설계 과정에서 고려되며, 이에 대한 정보를 수집하게 되므로 학습자 변인이 개인 정보와 밀접한 관련이 있다는 것을 알 수 있었다. 국외의 학습자 말뭉치 수집 및 이용 동의서에서도 개인 정보 수집에 대한 사항을 명시하고 있었으며, 목적 외의 용도로 활용하지 않고 동의를 거부할 권리가 있으며, 동의 거부에 따른 불이익이 없음을 밝히고 있었다.
- 학습자가 산출한 자료에서 개인 정보 및 사생활과 관련한 내용이 포함될 수 있기 때문에, 말뭉치 구축 과정에서 모든 사람의 사생활이 익명화를 통해 보호되어야 하며, 학습자가 사적 주제에 대해 글을 쓰거나 발화할 때, 지나친 사생활이 공개되는 경우 구축 과정에서 검토가 필요해 보인다. 구축 과정에서 개인 정보와 관련한 부분을 편집하고 익명화하는 것은 중요한 쟁점이 된다.
- 또한 말뭉치가 공개, 배포될 때, 스캔 및 녹음 파일의 원본 자료에 대한 익명화 처리도 중요한 쟁점이 될 수 있음을 해외 사례를 통해 알 수 있었다. 자료 내에서 개인 정보 및 민감한 사생활에 대한 익명화가 이루어지지 않았을 경우에는 녹음 자료 공개 및 접근에 제약을 가하여 윤리적인 문제가 발생하지 않도록 미연에 방지해야 할 것이다.
- 학습자 말뭉치 구축 단계에서는 개인 정보의 유출을 방지하기 위하여 학

습자 말뭉치에서 드러나는 개인 정보에 대하여 비식별화 처리를 하게 되는데, 이때 어떠한 정보를 비식별화하여 처리할 것인지, 즉 비식별화의 범위에 대한 논의가 필요함을 알 수 있었다. 국외의 경우, 이름, 전화번호, 이메일, 지역과 같은 정보가 주요 비식별화 대상으로 처리되었다.

- 학습자 말뭉치에서 수집되는 개인 정보와 공개 시에 제공되는 메타데이터는 구분된다. 학습자의 개인 정보를 최대한 보호하고자 한다면 메타데이터를 최소한으로 제한하는 방안을 고려해 볼 수 있으나, 사용자의 관점에서 사용 목적에 따라 메타데이터는 중요한 변인 정보로 활용될 수 있기 때문에 사용자의 편의를 고려한다면 보다 다양한 메타데이터를 제공할 필요성도 있다. 국외에서 공개되고 있는 학습자 말뭉치의 메타데이터를 검토한 결과, 학습자의 국적, 성별, 모국어, 숙달도 정보가 가장 공통된 항목으로, 이를 근거로 공개되는 메타데이터를 제한하는 방안을 고려해 볼 수 있다.
- 아울러 개인정보 보호법에 근거하여 개인 정보 수집 및 이용에 관한 사항을 자료 제공자인 학습자에게 명확히 제시하고 규정을 준수하는 것이 중요하다. 이에 본 연구에서는 개인 정보 수집 및 이용에 관한 동의서를 새롭게 작성하기로 하였다.

③ IRB

- IRB의 심의 대상은 인간을 대상으로 하는 연구로, 중재 연구, 상호작용을 통한 연구, 개인식별정보를 포함한 연구가 이에 해당한다. 이 중 학습자 말뭉치는 외국인 학습자의 자료를 수집 대상으로 하며, 종적 말뭉치 수집의 경우, 자료 수집 방식에 있어서 대상자와의 대면을 통해 자료 수집이 이루어지기 때문에 ‘상호작용을 통한 연구’에 포함될 수 있다. 또한 자료의 수집 과정에서 자료의 활용을 위해 개인 정보가 수집될 뿐만 아니라 학습자가 산출한 자료에도 개인 식별 정보와 사생활 관련 내용이 포함될 수 있다는 점에서 ‘개인식별정보를 포함한 연구’에 해당된다.
- 그러나 학습자 말뭉치의 경우 자료 수집이 사회적 약자인 외국인 학습자를 대상으로 한다고 할지라도, 보건복지부령 생명윤리법상 법 제15조 제2항 내지 제36조 제2항 “연구대상자(또는 인체유래물기증자) 및 공공에 미치는 위험이 미미한 경우”에 해당되며, 이 경우 심의를 면제할 수 있다는 조항에 의거하여 IRB의 심의를 면제받을 수 있을 것으로 보인다.
- 아울러 학습자 말뭉치 구축의 경우 연구의 범위가 언어 자원으로서 자료

를 데이터화하는 데에 그치며, 실질적인 자료 이용자가 아니므로 IRB 심의 규정은 자료 이용자가 받는 것이 타당하다는 결론에 도달하였다. 이에 따라 IRB 심의의 문제는 논외로 하였다.

1.2. 학계와 민간 분야 등 다양한 사용자 집단의 요구분석

1.2.1. 기본 방향

- 국가 언어 자원으로서 말뭉치 구축의 목적은 연구와 교육, 더 나아가 산업계에서의 폭넓은 활용을 통해 국가 경쟁력을 제고하는 것이다. 따라서 다양한 사용자 집단의 요구를 토대로 활용도 높은 말뭉치를 구축하는 것이 핵심 과제라고 할 수 있다. 이에 본 연구에서는 학계의 사용자 집단을 대상으로 한 설문조사를 통해 기구축 한국어 학습자 말뭉치를 비판적으로 검토하여 개선점을 도출해 내고, 민간 분야의 전문가 집담회를 통해 대규모 언어 자원으로서 4차 산업혁명 시대에 부합하는 말뭉치 구축 방향을 모색하고자 하였다.

1.2.2. 연구 내용

(1) 설문조사

① 설문조사 개요

- 설문조사는 한국어 교육 연구자 및 교원 등 학계의 학습자 말뭉치 사용자를 대상으로 학습자 말뭉치 활용 목적과 사용 경험, 만족도를 조사하여 기구축 말뭉치에 관한 개선 사항을 점검하고 2차 중장기 계획을 통해 구축해 나갈 말뭉치의 보완 방향을 모색하는 데에 목적이 있다. 조사 개요는 다음과 같다.

- 조사내용: 한국어 학습자 말뭉치 사용자 요구조사
- 조사 기간: 2021년 6월 28일~7월 9일
- 조사대상: 한국어 학습자 말뭉치 사용 경험자 148명¹²⁾

- 설문조사 문항은 1) 기본 정보, 2) 한국어 학습자 말뭉치 활용 경험과 평가, 3) 한국어 학습자 말뭉치 배포 시스템인 ‘한국어 학습자 말뭉치 나눔터’ 이용 경험과 평가 세 개 영역으로 구성되었으며, 세부 문항 구성은 다음과 같다.

<표 36> 설문조사 문항 구성

영역	문항 구성
I. 기본 정보	1. 국적 2. 직업 3. 소속 기관 4. 소속 기관 소재지
II. 한국어 학습자 말뭉치 활용 경험과 평가	1. 이용 시기 2. 알게 된 경로 3. 자료 이용 경로 4. 이용 목적 <ul style="list-style-type: none"> 4-1. 연구 목적 4-2. 연구 영역 4-3. 연구 주제 4-4. 교수·학습 자료 개발 세부 영역 5. 사용한 학습자 말뭉치의 유형 <ul style="list-style-type: none"> 5-1. 학습자 말뭉치의 유형 5-2. 자료 유형 5-3. 수준 5-4. 언어권 5-5. 자료 형식 6. 만족도 7. 제안 사항
III. 한국어 학습자 말뭉치 나눔터 이용 경험과 평가	1. 이용 빈도 2. 이용 메뉴 3. 만족도 4. 향후 계속 이용 의사 5. 추천 의사 6. 제안 사항

- 12) 설문은 국립국어원에 학습자 말뭉치 자료를 요청하여 제공받은 사용자와 학습자 말뭉치 아카데미 참여자 980명을 대상으로 배포하였으며, 그중 148명이 응답하였다.

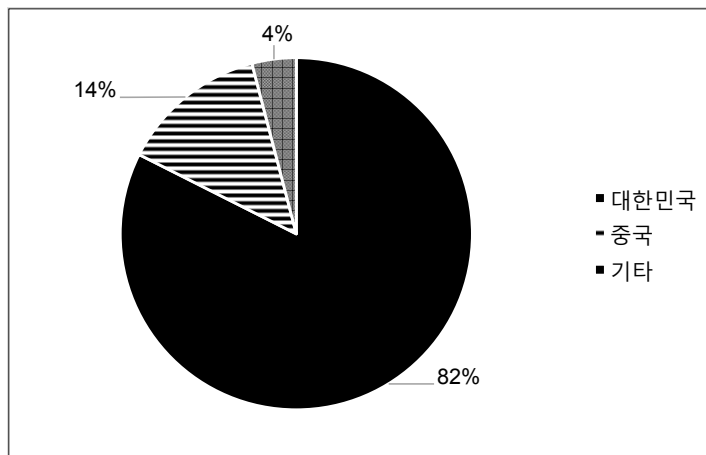
② 설문조사 결과의 분석

○ 다음은 총 148명의 응답 자료를 처리한 결과이다.

기본 정보

1. 국적

○ 응답자의 국적은 <그림 6>과 같이 전체의 82%가 대한민국으로 가장 많았으며, 13%가 중국, 그 외 미국, 대만, 몽골, 베트남, 네덜란드가 각 1%로 구성되어 있음을 알 수 있다.



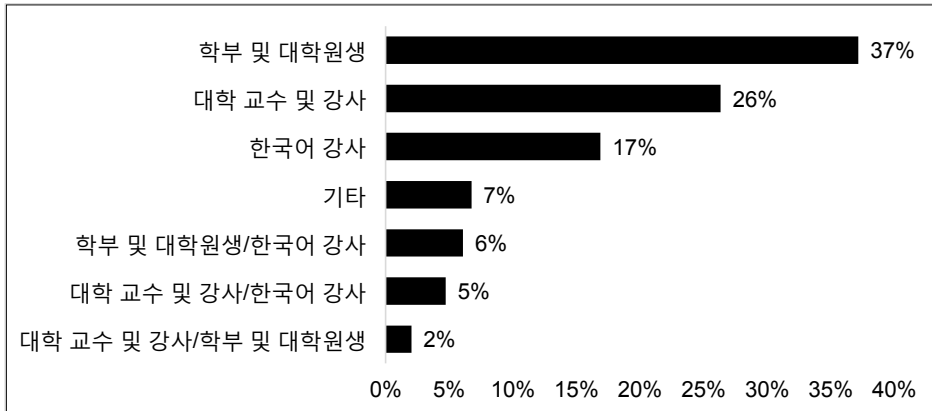
<그림 6> 응답자 국적 분포

<표 37> 응답자 국적 빈도

국적	빈도
대한민국	122
중국	20
미국	2
대만	1
몽골	1
베트남	1
네덜란드	1
합계	148

2. 응답자의 직업

- 응답자의 직업은 <그림 7>와 같이 전체 응답 중에 학부 및 대학원생이 37%로 가장 많았으며, 이어서 대학 교수 및 강사 26%, 한국어 강사가 17%의 순으로 나타났다.¹³⁾ 기타 응답으로는 기업의 한국어 교육 관련 부서의 직장인이나 고등학교 교사, 인공지능 엔지니어 등이 있었다.



<그림 7> 응답자 직업 분포

<표 38> 응답자 직업 빈도

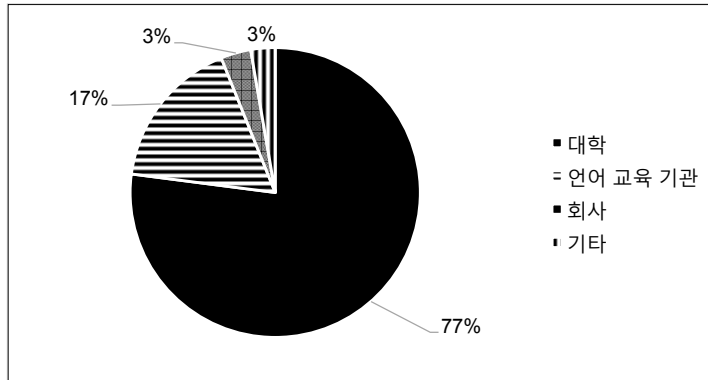
직업	빈도
학부 및 대학원생	55
대학 교수 및 강사	39
한국어 강사	25
기타	10
학부 및 대학원생/한국어 강사	9
대학 교수 및 강사/한국어 강사	7
대학 교수 및 강사/학부 및 대학원생	3
합계	148

3. 응답자의 소속 기관

- 응답자의 소속 기관은 <그림 8>과 같이 전체 응답 중에 대학이 77%로

13) 이 문항은 복수 선택이 가능한 문항으로 응답 수의 합이 148과 일치하지 않는다.

가장 많았고 두 번째로는 언어 교육 기관이 17%였다. 그 외에 회사나 초
중고등학교에서 근무하는 사람도 있었다.



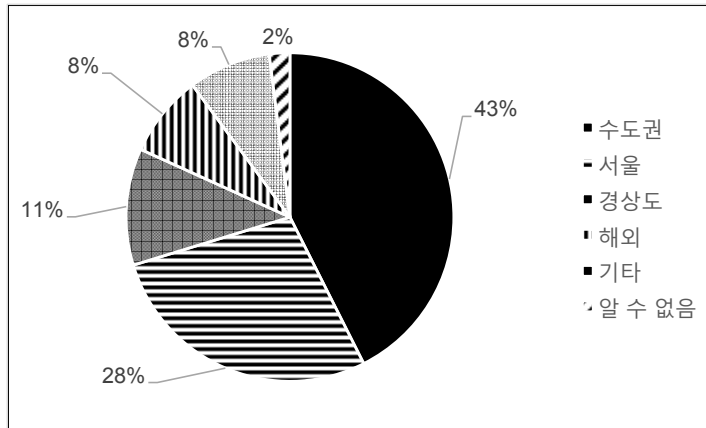
<그림 8> 응답자 소속기관 분포

<표 39> 응답자 소속기관 빈도

소속기관	빈도
대학	114
언어 교육 기관	25
회사	5
초중고등학교	3
개인 교습	1
합계	148

4. 응답자의 소속 기관 소재지

- 응답자의 소속 기관 소재지는 <그림 9>와 같이 수도권과 서울이 43%, 28%로 절반 이상을 차지했다. 두 지역을 제외한 곳에서는 경상도가 11%로 가장 많았다. 비수도권 지역에서의 학습자 말뭉치 사용을 늘리기 위하여 말뭉치 아카데미를 지방에서 개최하거나 온라인으로 개최하여 비수도권과 해외에서의 사용을 함께 늘릴 수 있는 방안이 필요한 것으로 보인다. 해외 지역에서는 일본, 미국, 중국, 캐나다, 대만, 몽골이 있었다.



<그림 9> 응답자 소속기관 소재지 분포

<표 40> 응답자 소속기관 소재지 분포

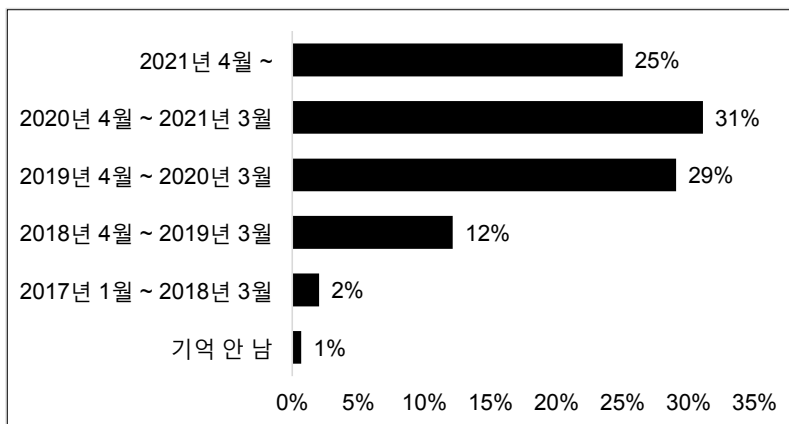
소속기관 소재지	빈도
수도권	63
서울	41
경상도	17
해외	12
전라도	5
충청도	4
강원도	2
제주도	1
알 수 없음 ¹⁴⁾	3
합계	148

14) 광역시, 비수도권, 지방의 응답이 있었다.

한국어 학습자 말뭉치 활용 경험과 평가

1. 자료 이용 시기

- 국립국어원에 한국어 학습자 말뭉치를 요청하거나 다운로드받은 시기를 조사하였다. 응답 결과를 보통 전년도의 구축 자료가 나눔터에 업데이트 되는 시기를 기준으로 나누어 분류한 결과 2020년 4월부터 2021년 3월까지의 사용률이 가장 많았고, 2021년 4월부터 조사가 진행된 7월 초까지 25%의 사용률이 나타난 것으로 보아 앞으로의 사용 빈도가 점점 더 증가할 것으로 보인다.



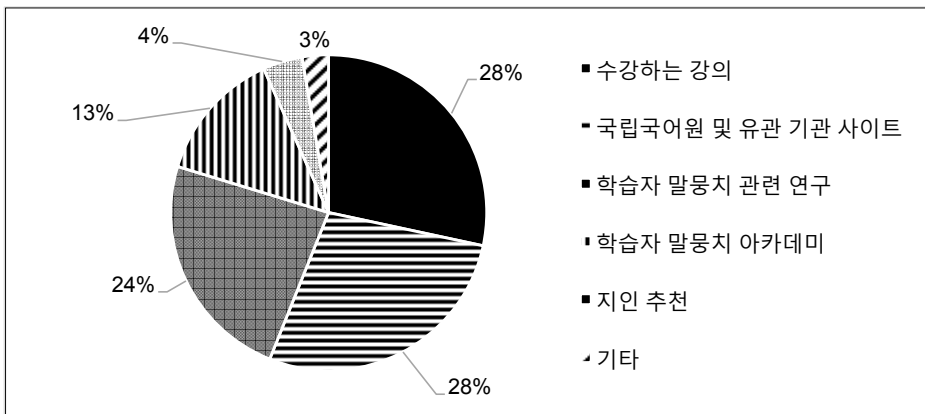
<그림 10> 학습자 말뭉치 이용 시기 분포

<표 41> 학습자 말뭉치 이용 시기 빈도

사용 시기	빈도
2017년 1월 ~ 2018년 3월	3
2018년 4월 ~ 2019년 3월	18
2019년 4월 ~ 2020년 3월	43
2020년 4월 ~ 2021년 3월	46
2021년 4월 ~	37
기억 안 남	1
합계	148

2. 학습자 말뭉치를 알게 된 경로

- 한국어 학습자 말뭉치를 알게 된 경로를 조사한 결과 <그림 11>에서 알 수 있듯이 대학이나 대학원에서 수강하는 강의나 국립국어원 및 유관 기관 사이트의 홍보를 통해 알게 되었다는 응답이 각 28%로 가장 많았고, 그다음으로 학습자 말뭉치를 활용한 연구를 통해 알게 되었다는 응답이 24%로 두 번째로 많았다. 학습자 말뭉치 아카데미를 통해 알게 되었다는 응답은 13%였고, 그 밖에 지인의 추천이나 학회, 검색을 통해서 알게 되었다는 응답이 있었다.



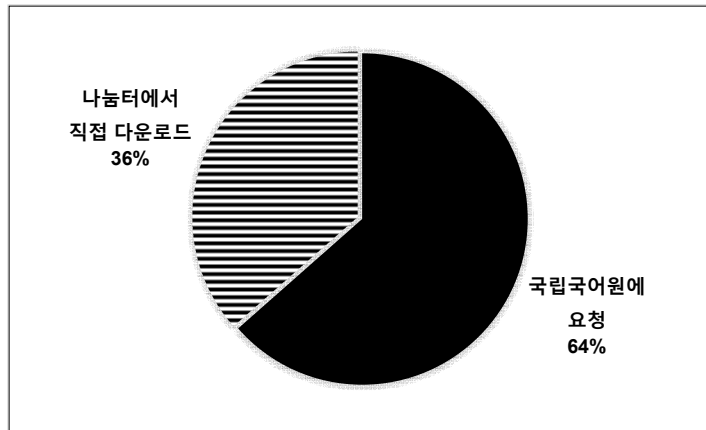
<그림 11> 학습자 말뭉치를 알게 된 경로 분포

<표 42> 학습자 말뭉치를 알게 된 경로 빈도

알게 된 경로	빈도
수강하는 강의	42
국립국어원 및 유관 기관 사이트	41
학습자 말뭉치 관련 연구	35
학습자 말뭉치 아카데미	20
지인 추천	6
기타	4
합계	148

3. 학습자 말뭉치를 제공받은 경로

- 한국어 학습자 말뭉치를 얻은 경로로는 국립국어원에 요청하여 얻은 경우가 64%로 절반 이상을 차지하였고 36%의 응답자가 나눔터에서 검색한 자료를 직접 다운로드하여 사용했다고 답하였다.



<그림 12> 학습자 말뭉치 자료 다운로드 경로 분포

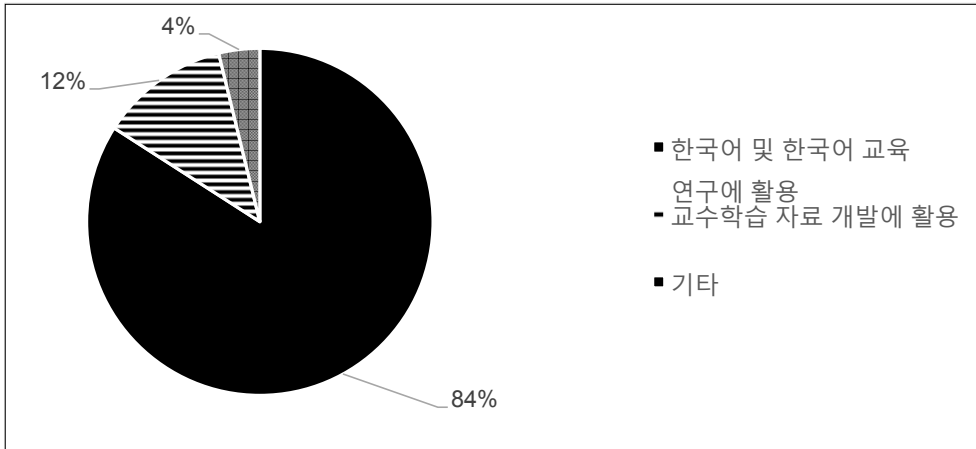
<표 43> 학습자 말뭉치 자료 다운로드 경로 빈도

자료를 얻은 경로	빈도
국립국어원에 요청	94
나눔터에서 직접 다운로드	54
합계	148

4. 학습자 말뭉치 이용 목적

- 한국어 학습자 말뭉치를 이용하는 목적으로는 한국어 및 한국어 교육 연구에의 활용이라는 응답이 84%로 대부분을 차지했으며, 교수학습 자료 개발에 활용하기 위함이라는 응답이 12%로 나타났다. 기타 응답으로는 수업 시간에 실습 자료로 활용하거나 자연어처리 모델 개발 데이터로 활용한다는 응답 등이 있었다.¹⁵⁾

15) 여기에서 ‘한국어 및 한국어 교육 연구에 활용’에 응답한 사람들은 4-1, 4-2, 4-3 문항에서 상세 질문에 대한 응답을 하게 하였으며, ‘교수학습 자료 개발에 활용’을 선택한 응답자들은 4-4에서 상세 질문에 응답하도록 하였다.



<그림 13> 학습자 말뭉치 이용 목적 분포

<표 44> 학습자 말뭉치 이용 목적 빈도

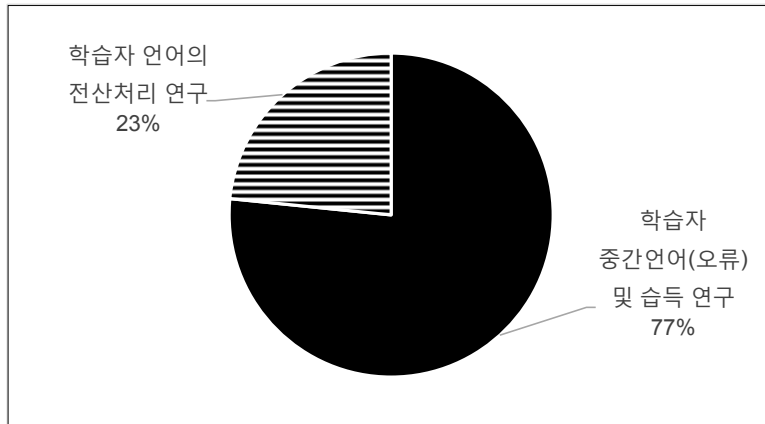
이용 목적	빈도
한국어 및 한국어 교육 연구에 활용	132
교수학습 자료 개발에 활용	19
기타	6
합계	157 ¹⁶⁾

4-1. 연구를 위한 활용의 세부 목적

- 한국어 학습자 말뭉치를 이용하는 목적을 묻는 질문에서 한국어 학습자 말뭉치를 한국어 및 한국어 교육 연구에 활용한다고 응답한 사람에게 연구의 목적을 조사한 결과 학습자 중간언어(오류) 및 습득에 관한 연구를 위해 사용했다고 한 응답이 77%로 절반 이상을 차지하였고, 학습자 언어의 전산처리 연구를 위해 사용했다고 한 응답은 23%로 나타났다.¹⁷⁾

16) 복수 응답 가능

17) 기타에서 딥러닝 자연어 처리나 AI대화문이라고 응답한 것은 학습자 언어의 전산처리 연구에 포함하였고, 대조·오류 분석 연구, 화용적 사용 양상 연구라고 응답한 것은 학습자 중간언어(오류) 및 습득 연구에 포함하였다.



<그림 14> 한국어 및 한국어 교육 연구 활용에서의 연구 목적 분포

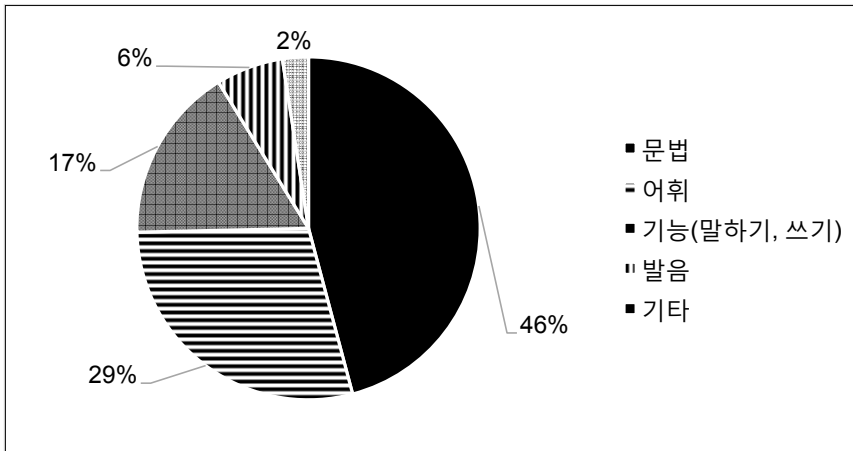
<표 45> 한국어 및 한국어 교육 연구 활용에서의 연구 목적 빈도

연구 목적	빈도
학습자 중간언어(오류) 및 습득 연구	118
학습자 언어의 전산처리 연구	36
합계	154 ¹⁸⁾

4-2. 연구 영역

- 한국어 학습자 말뭉치를 한국어 및 한국어 교육 연구에 활용한다고 응답한 사람의 연구 영역을 조사한 결과 문법을 연구하는 사람이 46%로 가장 많았고 어휘 연구가 29%, 말하기 쓰기와 같은 기능을 연구했다는 응답이 17%로 나타났으며, 발음 연구는 6%에 불과했다. 기타 응답에서는 담화 및 화용 영역이나 학습자의 전략, 교육 방안 등과 같은 응답이 있었다.

18) 복수 응답 가능



<그림 15> 한국어 및 한국어 교육 연구 활용에서의 연구 영역 분포

<표 46> 한국어 및 한국어 교육 연구 활용에서의 연구 영역 빈도

연구 영역	빈도
문법	94
어휘	59
기능(말하기, 쓰기)	34
발음	13
기타	5
합계	205 ¹⁹⁾

4-3. 연구 주제

- 4-2에서 응답한 각 연구 영역의 구체적인 연구 주제에 대해서 응답한 결과를 정리하면 다음과 같다.

<표 47> 연구 영역별 세부 연구 주제

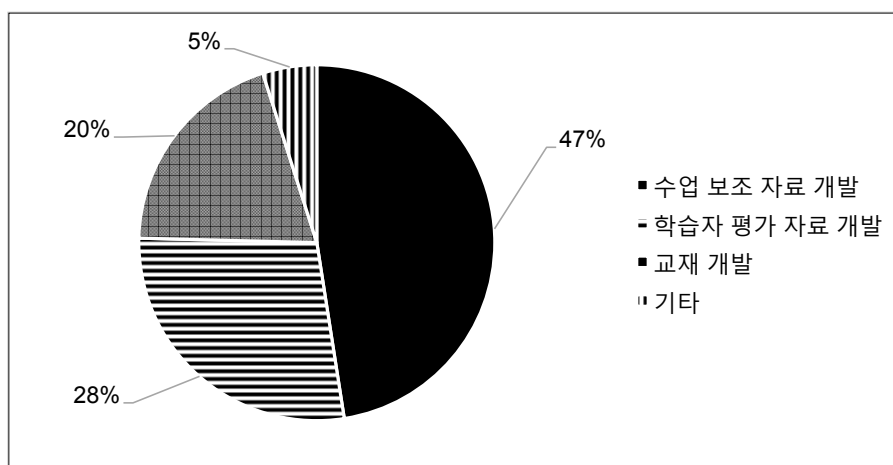
연구 영역	세부 연구 주제
1) 문법	○ 조사(주격 조사/주격 보조사/부사격 조사), 어미(연결 어미, 관형사형 전성어미, 선어말어미), 피동, 부정, 시제, 높임법

19) 복수 응답 가능

2) 어휘	○ 구어체 어휘 오류, 어휘 사용 양상(풍요도/유의어/다의어/개별 품사 어휘), 어휘 사용 빈도, 연어
3) 기능(말하기, 쓰기)	○ 말하기 교육, 쓰기(이메일 쓰기, 쓰기 점수와 자질)
4) 발음	○ 발음, 억양
5) 기타	○ 담화 표지, 화행(감탄), 의사소통 전략, 표현 문형 ²⁰⁾

4-4. 교수·학습 자료 개발을 위한 활용의 세부 목적

- 한국어 학습자 말뭉치를 이용하는 목적을 묻는 질문에서 교수학습 자료 개발에 활용에 응답한 사람에게 구체적인 목적을 조사한 결과 수업 보조 자료 개발을 위함이 47%로 가장 많았고, 학습자 평가 자료 개발은 28%, 교재 개발을 위한 사용은 20%로 나타났다. 기타 응답으로는 기말 보고서 제출이나 한국어 학습자에 대한 교육 내용에 적용을 위해 사용했다는 응답이 있었다.



<그림 16> 교수·학습 자료 개발에 활용에서의 연구 목적 분포

20) 헤지 표현, 정형 표현 등의 용어 포함

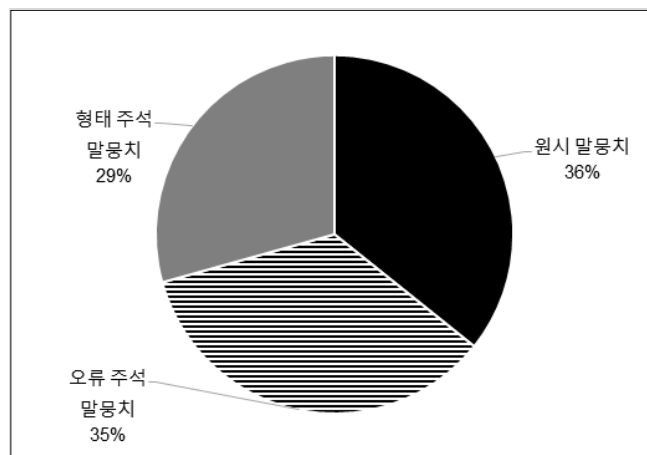
<표 48> 교수·학습 자료 개발에 활용에서의 연구 목적 분포

자료 유형	빈도
수업 보조 자료 개발	29
학습자 평가 자료 개발	17
교재 개발	12
기타	3
합계	61

5. 주로 이용한 한국어 학습자 말뭉치 유형

5-1. 이용한 말뭉치의 유형

- 원시 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치 중에서 어떤 유형의 말뭉치를 사용했는가에 대한 질문에 세 말뭉치가 큰 차이 없이 골고루 이용되고 있음을 알 수 있었다. 특히 원시 말뭉치와 오류 주석 말뭉치는 36%와 35%로 사용 빈도가 거의 같게 나타났다.



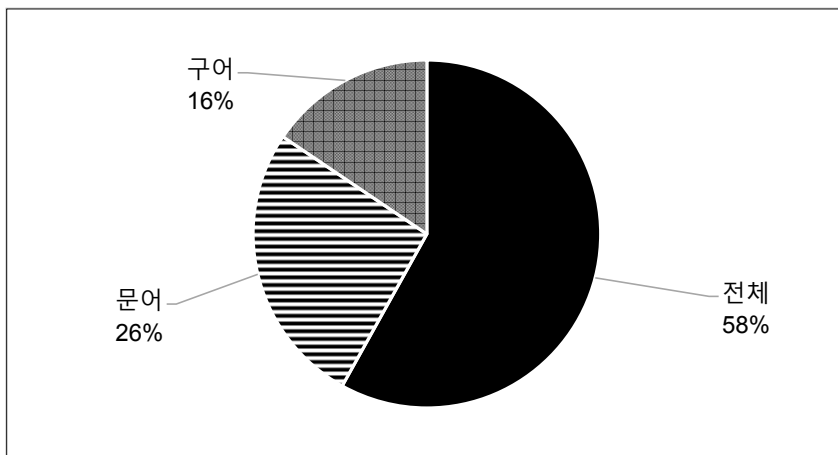
<그림 17> 사용 말뭉치 유형 분포

<표 49> 사용 말뭉치 유형 빈도

말뭉치 유형	빈도
원시 말뭉치	89
오류 주석 말뭉치	87
형태 주석 말뭉치	73
합계	249 ²¹⁾

5-2. 이용한 자료의 유형

- 문어 자료와 구어 자료 중에 어느 것을 사용했는가에 대한 질문에는 문어 자료와 구어 자료를 모두 사용했다는 응답이 58%로 가장 많았고, 문어 자료를 사용했다는 응답은 26%, 구어 자료를 사용했다는 응답은 16%로 문어 자료가 더 많이 이용되고 있음을 알 수 있었다.



<그림 18> 사용 말뭉치 자료 유형 분포

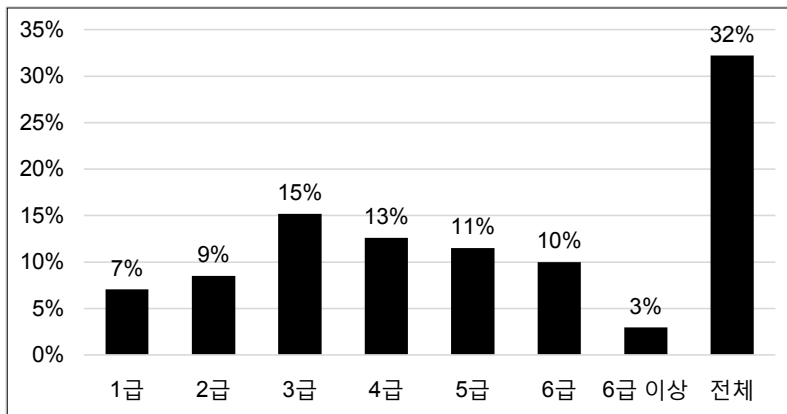
21) 복수 응답 가능

<표 50> 사용 말뭉치 자료 유형 빈도

자료 유형	빈도
전체	86
문어	39
구어	23
합계	148

5-3. 이용한 말뭉치의 등급

- 이용한 말뭉치의 등급에 대한 질문에 대해서는 1급부터 6급 이상까지 전체의 수준을 포함한 자료를 사용하였다는 응답이 32%로 가장 많았고, 개별 수준에서는 3급이 15%, 4급이 13%, 5급이 11%, 6급이 10%, 2급이 9%, 1급이 7%, 6급 이상이 3%의 순으로 나타났다. 이를 보면 중급 수준의 학습자 말뭉치가 가장 많이 사용된다는 것을 알 수 있고 그 뒤로 고급, 초급의 순으로 사용되고 있음을 알 수 있다.



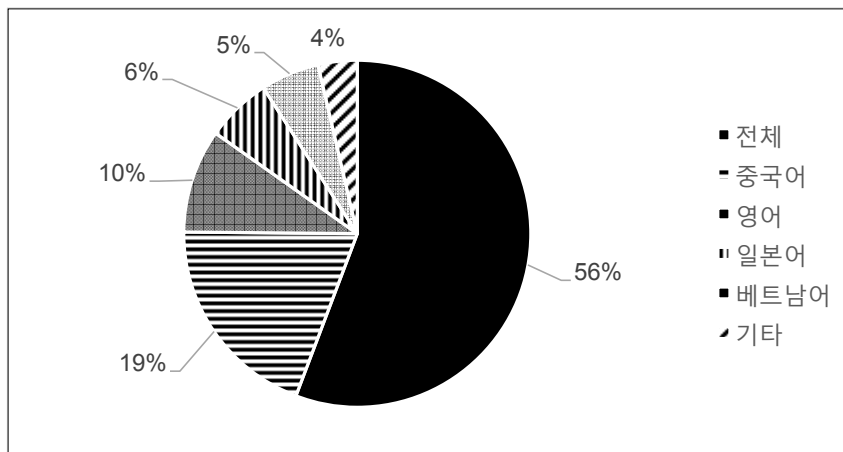
<그림 19> 사용 말뭉치 학습자 수준 분포

<표 51> 사용 말뭉치 학습자 수준 빈도

학습자 수준	빈도
1급	19
2급	23
3급	41
4급	34
5급	31
6급	27
6급 이상	8
전체	87
합계	270 ²²⁾

5-4. 이용한 말뭉치의 언어권

- 이용한 말뭉치의 언어권에 대한 조사에서는 특정 언어권을 선택하지 않고 전체 언어권을 대상으로 사용했다는 응답이 56%로 절반을 차지했다. 개별 언어권에서는 중국어권이 19%, 영어권이 10%, 일본어권 6%, 베트남어권 5%의 순으로 나타났다. 기타 언어권에서는 대만어, 스페인어, 독일어, 몽골어, 태국어, 프랑스어가 있었다.



<그림 20> 사용 말뭉치 언어권 분포

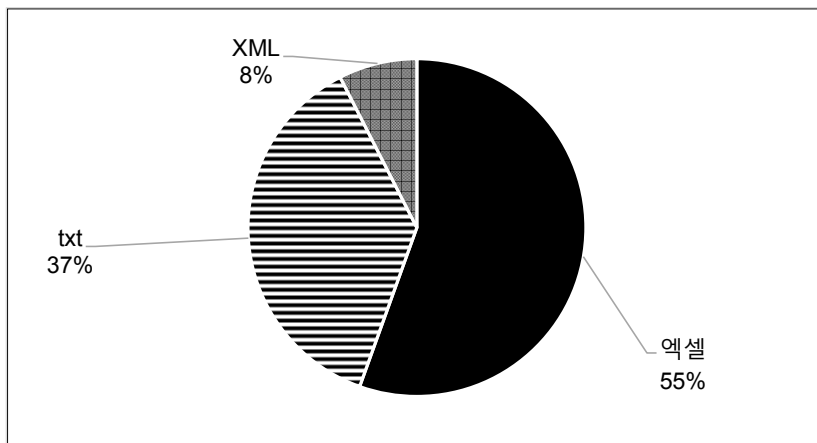
22) 복수 응답 가능

<표 52> 사용 말뭉치 언어권 빈도

언어권	빈도
전체	92
중국어	32
영어	16
일본어	10
베트남어	9
기타	6
합계	165 ²³⁾

5-5. 이용한 말뭉치의 자료 형식

- 이용한 말뭉치의 자료 형식을 묻는 질문에서는 엑셀 형식의 자료를 사용했다는 응답이 55%로 절반을 차지했고, txt 형식은 37%, XML 형식의 자료를 사용했다는 응답은 8%에 불과했다. XML 형식의 자료는 사용하는 데에 있어 엑셀이나 txt 형식보다 좀 더 전문적인 지식을 요구하는 형식의 자료이기 때문에 사용률이 많이 높지 않은 것으로 보인다.



<그림 21> 사용 말뭉치 자료 형식 분포

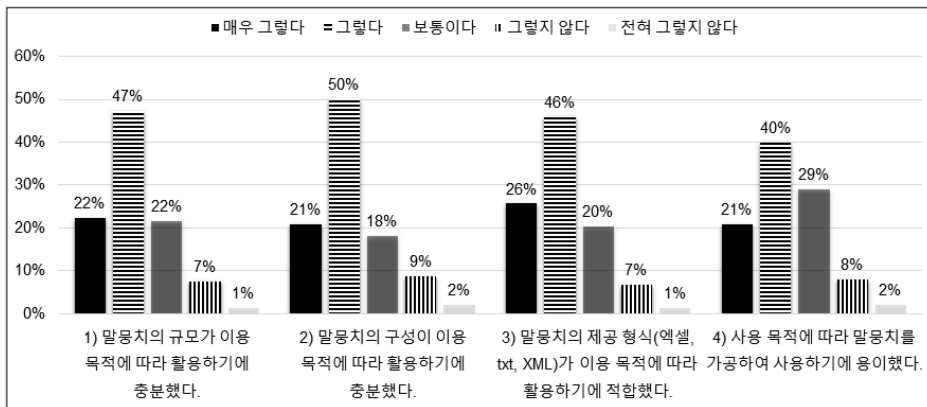
23) 복수 응답 가능

<표 53> 사용 말뭉치 자료 형식 빈도

자료 형식	빈도
엑셀	111
txt	74
XML	15
합계	200 ²⁴⁾

6. 한국어 학습자 말뭉치에 대한 만족도

- 이용한 말뭉치에 대한 만족도를 조사하기 위하여 네 가지 항목에 대해 만족도를 물어본 결과를 정리하면 <그림 22>와 같다. ‘매우 그렇다’와 ‘그렇다’의 대답을 긍정적인 응답으로 봤을 때 모든 항목에서 대체적으로 만족하고 있다는 것을 알 수 있다. 그러나 네 번째 문항인 ‘사용 목적에 따라 말뭉치를 가공하여 사용하기에 용이했다’의 항목에서는 다른 문항보다 ‘보통이다’의 응답이 29%로 가장 많았다. ‘보통이다’와 ‘그렇지 않다’, ‘전혀 그렇지 않다’의 세 문항을 부정적인 응답으로 봤을 때 세 문항에 대한 응답을 합산하여 살펴보면 역시 네 번째 문항이 39%로 가장 만족도가 가장 떨어지는 부분이라는 것을 알 수 있었다.



<그림 22> 사용 말뭉치에 대한 만족도 분포

24) 복수 응답 가능

<표 54> 사용 말뭉치에 대한 만족도 빈도

문항	매우 그렇다	그렇다	보통 이다	그렇지 않다	전혀 그렇지 않다
1) 말뭉치의 규모가 이용 목적에 따라 활용하기에 충분했다.	33	70	32	11	2
2) 말뭉치의 구성이 이용 목적에 따라 활용하기에 충분했다.	31	74	27	13	3
3) 말뭉치의 제공 형식(엑셀, txt, XML)이 이용 목적에 따라 활용하기에 적합했다.	38	68	30	10	2
4) 사용 목적에 따라 말뭉치를 가공하여 사용하기에 용이했다.	31	59	43	12	3

6-1. 말뭉치 구성에 관한 개선점과 제안 사항

- 말뭉치 구성과 관련하여 개선할 점이나 제안 사항에 대한 의견으로는 다음과 같은 것들이 있었다.

<표 55> 말뭉치 구성 관련 개선 사항 및 제안 사항

항목	세부 내용
1) 자료의 주제	○ 주제별 자료의 양적 확대 ○ 구어/문어 자료의 주제 다양화 ○ 주제별 말뭉치 구성
2) 말뭉치 자료 유형	○ 구어 말뭉치의 양적 확대(16명)
3) 학습자 유형	○ 국외 학습자 자료 및 특정 언어권 학습자 자료 양적 확대(10명) ○ 특정 언어권에 편향되지 않았으면 좋겠다는 의견이 많았고, 응답자 개인이 원하는 특정 언어권의 양적 확대를 원하는 의견이 많았음
4) 학습자 수준	○ 중/고급 학습자 자료 확대(6명)
5) 말뭉치 유형	○ 오류 주석 말뭉치의 양적 확대(4명)
6) 기타	○ 문어 자료 전문/음성 DB 제공(2명) ○ 파일 구성을 국제 표준인 M2 파일로 변경 요청(1명)

6-2. 말뭉치 제공 형식에 관한 개선점과 제안 사항

- 말뭉치 제공 형식(엑셀, txt, XML)과 관련하여 개선할 점이나 제안 사항에 대한 의견으로는 다음과 같은 것들이 있었다.

<표 56> 말뭉치 제공 형식 관련 개선 사항 및 제안 사항

항목	세부 내용
1) 전체적인 의견	<ul style="list-style-type: none"> ○ 웹에서 엑셀로 내려받기를 할 때 모든 페이지가 한 번에 다운로드 되도록 개선 필요(4명) ○ 제공 형식별 활용 가이드 제공이나 상세한 교육 필요(2명) ○ 산출된 자료의 원문까지 찾아볼 수 있는 검색기 사용까지 연결되기 바람(1명) ○ 텍스트 파일에는 정보가 부족한 반면 엑셀 파일은 너무 복잡하게 구성되어 있음(1명)
2) 자료 형식별 의견	<ul style="list-style-type: none"> ○ 엑셀(2명) <ul style="list-style-type: none"> - 표본의 총 어절 수가 제공되었으면 함 - 엑셀 한 칸에 한 어절이 모두 들어간 자료(예: '적/VA + 은/ETM'이 한 칸 안에)도 필요할 때도 있음 ○ txt(2명) <ul style="list-style-type: none"> - 용례 검색 프로그램에서 활용할 수 있는 확장자로 제공되기 바람 - 형태 주석/오류 주석 말뭉치도 txt 형태로 제공되기 바람 ○ XML(1명) <ul style="list-style-type: none"> - 별도로 프로그래밍을 하지 않으면 원하는 정보를 추출하기 어려움

6-3. 말뭉치 가공 사용에 관한 개선점과 제안 사항

- 사용자가 말뭉치를 추가로 가공하여 사용하는 것과 관련하여 개선할 점이나 제안 사항에 대한 의견으로는 다음과 같은 것들이 있었다.

<표 57> 말뭉치 가공 사용 관련 개선 사항 및 제안 사항

항목	세부 내용
1) 개선 사항	<ul style="list-style-type: none"> ○ 주석에 오류가 많음(4명) ○ 말뭉치 요청 시 서약서 제출 간소화(2명) ○ 말뭉치의 지속적인 수정 보완(1명)
2) 제안 사항	<ul style="list-style-type: none"> ○ 웹상에서 다양한 방식으로 가공할 수 있게 하거나 가공 방법 관련 온라인 교육 제공(3명) ○ 더 편리하고 유용한 검색 도구 제공 필요(3명) ○ 조건에 따라 분류된 말뭉치 필요(수준/언어/주제 등)(2명) ○ 원시 말뭉치에서 더 많은 정보 제공 필요(구어 원시말뭉치 내에서 화자 구분)(1명) ○ 모든 정보가 한 파일에 모두 들어간 자료 필요(1명) ○ 앞뒤 문맥에서 2차 검색 가능(1명)

7. 기타 제안 사항

- 한국어 학습자 말뭉치의 효율적 활용을 위해 한국어 학습자 말뭉치 구축 연구팀에 제한하고 싶은 사항을 자유롭게 적는 문항에서는 다음과 같은 의견들이 있었다.

<표 58> 한국어 학습자 말뭉치 전반에 대한 개선 사항 및 제안 사항

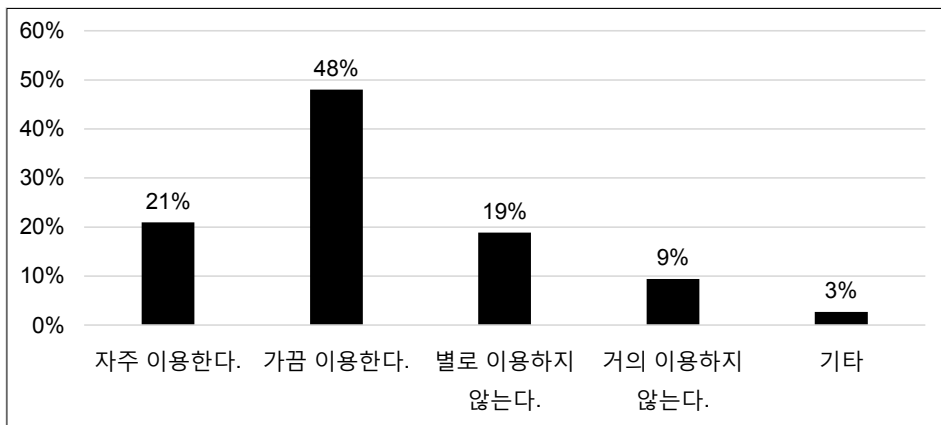
항목	세부 내용
1) 말뭉치 사용 매뉴얼 강의 및 말뭉치 아카데미 활성화	<ul style="list-style-type: none"> ○ 해외에 있거나 아카데미 불참석자를 위한 동영상 업로드 요청(8명)
2) 말뭉치에 대한 요구	<ul style="list-style-type: none"> ○ 오류 주석 말뭉치의 정확성 및 양적 확대(4명) ○ 중복 자료에 대한 처리 필요(3명) ○ 말뭉치 자체의 주제 다양화와 양적 확대(3명) ○ 다양한 목적을 가진 학습자 자료의 균형(학문목적/이주 여성 등)(3명) ○ 구어 자료의 양적 확대(1명) ○ 원시 말뭉치의 양적 확대(1명) ○ 한 학습자의 문/구어 대조 말뭉치 구축 필요(1명)

3) 나눔터 사용에 대한 요구	<ul style="list-style-type: none"> ○ 용례 검색의 어려움(5명) ○ 자료 다운로드의 절차 간소화(3명) ○ 자료 업데이트 일자 공지(1명) - 모집단 수가 바뀌어 연구 진행 중 수정 과정이 필요했음 ○ 해외 거주 외국인을 위하여 편리한 회원 가입 방법 제공(1명)
4) 말뭉치 활용 도구 개발 및 보급	- 1명

한국어 학습자 말뭉치 나눔터 이용 경험과 평가

1. 한국어 학습자 말뭉치 나눔터 이용 빈도

- 한국어 학습자 말뭉치 나눔터를 얼마나 자주 이용하는지에 대한 조사에서는 ‘가끔 이용한다’는 응답이 48%로 나타났고 그 뒤로 ‘자주 이용한다’는 응답이 21%, ‘별로 이용하지 않는다’는 응답이 19%, ‘거의 이용하지 않는다’는 응답이 9%로 나타났다. 기타 응답으로는 예전에는 자주 이용했지만 최근에는 거의 이용하지 않는다는 응답과 사용하지 않는다는 응답이 있었다.



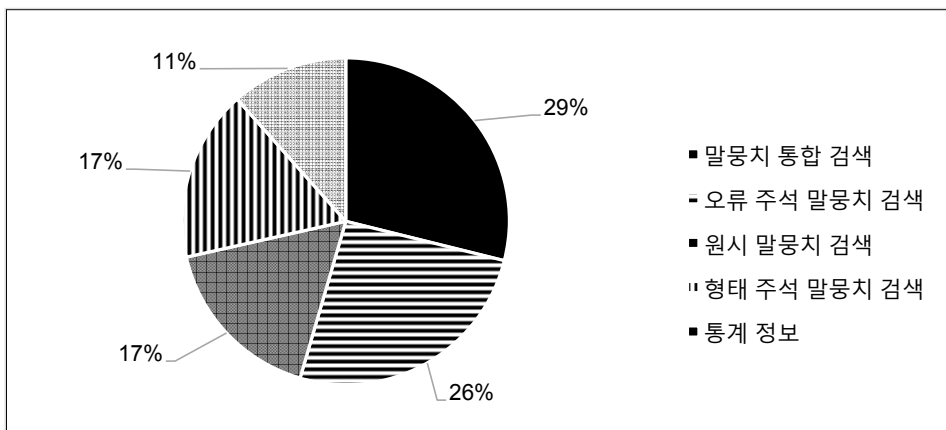
<그림 23> 한국어 학습자 말뭉치 나눔터 이용 횟수 분포

<표 59> 한국어 학습자 말뭉치 나눔터 이용 횟수 빈도

이용 빈도	빈도
자주 이용한다.	31
가끔 이용한다.	71
별로 이용하지 않는다.	28
거의 이용하지 않는다.	14
기타	4
합계	148

2. 주요 이용 메뉴

- 한국어 학습자 말뭉치 나눔터에서 주로 이용하는 메뉴로는 29%를 차지한 말뭉치 통합 검색 메뉴를 가장 많이 사용하는 것으로 나타났으며, 말뭉치 유형별 검색 메뉴에서는 오류 주석 말뭉치 검색이 26%로 가장 많았다. 원시 말뭉치 검색 메뉴와 형태 주석 말뭉치 검색 메뉴는 각 17%로 동일한 결과가 나왔고, 통계 정보 메뉴 사용에 대한 응답은 11%를 차지했다.



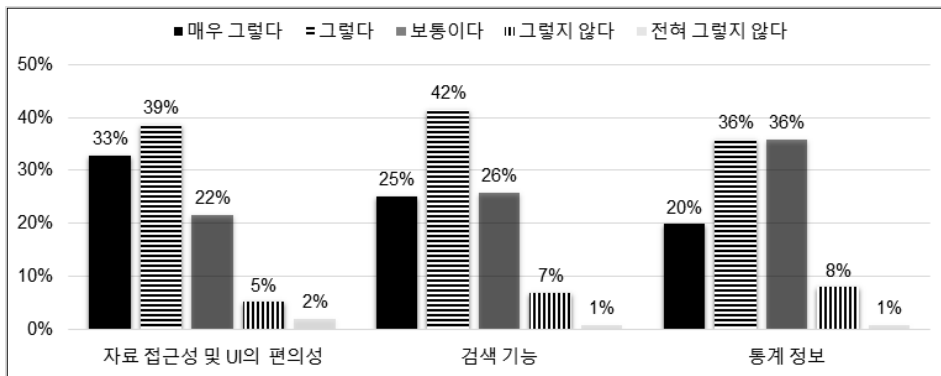
<그림 24> 한국어 학습자 말뭉치 나눔터에서의 주 이용 메뉴 분포

<표 60> 한국어 학습자 말뭉치 나눔터에서의 주 이용 빈도

나눔터 메뉴	빈도
말뭉치 통합 검색	80
오류주석 말뭉치 검색	71
원시 말뭉치 검색	47
형태 주석 말뭉치 검색	47
통계 정보	32
합계	277 ²⁵⁾

3. 한국어 학습자 말뭉치 나눔터에 대한 만족도

- 한국어 학습자 말뭉치 나눔터에 대한 만족도를 조사하기 위하여 크게 ‘자료 접근성 및 UI의 편의성’, ‘검색 기능’, ‘통계 정보’의 세 가지 영역으로 나누고 각 영역에서 좀 더 세부적인 항목에 대해 만족도를 조사하였다. 먼저 세 영역별 만족도 결과를 비교해 보면 <그림 25>와 같다. 세 영역 중에 ‘자료 접근성 및 UI의 편의성’과 ‘검색 기능’ 영역에 대한 만족도가 가장 높은 것으로 나타났고, ‘통계 정보’에 대한 만족도가 비교적 낮은 것을 알 수 있었다.



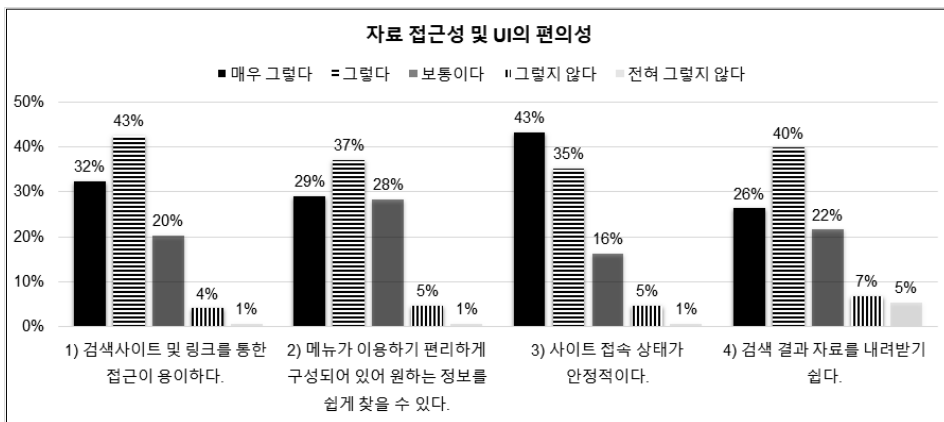
<그림 25> 한국어 학습자 말뭉치 나눔터 만족도 분포

25) 복수 응답 가능

<표 61> 한국어 학습자 말뭉치 나눔터 만족도 빈도

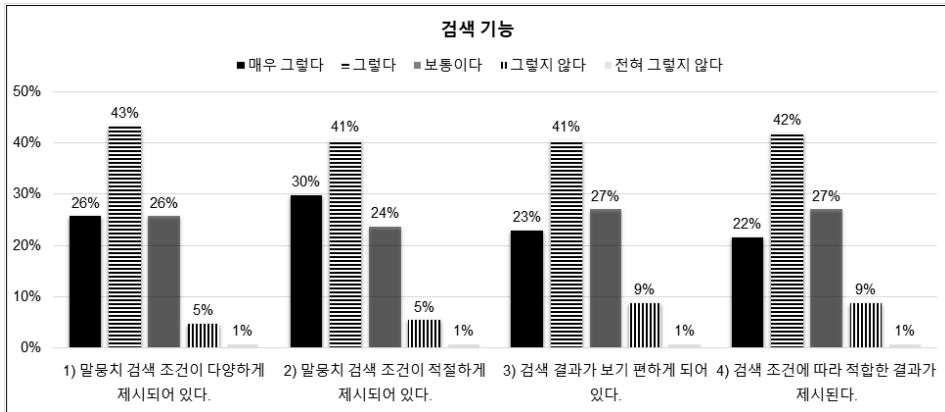
구분	문항	매우 그렇다	그렇다	보통 이다	그렇지 않다	전혀 그렇지 않다
자료 접근성 및 UI의 편의성	1) 검색 사이트 및 링크를 통한 접근이 용이하다.	48	63	30	6	1
	2) 메뉴가 이용하기 편리하게 구성되어 있어 원하는 정보를 쉽게 찾을 수 있다.	43	55	42	7	1
	3) 사이트 접속 상태가 안정적이다.	64	52	24	7	1
	4) 검색 결과 자료를 내려받기 쉽다.	39	59	32	10	8
검색 기능	1) 말뭉치 검색 조건이 다양하게 제시되어 있다.	38	64	38	7	1
	2) 말뭉치 검색 조건이 적절하게 제시되어 있다.	44	60	35	8	1
	3) 검색 결과가 보기 편하게 되어 있다.	34	60	40	13	1
	4) 검색 조건에 따라 적합한 결과가 제시된다.	32	62	40	13	1
통계 정보	1) 원하는 통계 결과를 제공하고 있다.	32	54	53	8	1
	2) 통계 자료를 확인하는 데 불편함이 없다.	30	51	51	15	1
	3) 통계 수치와 검색 결과가 일치한다.	26)	54	55	12	1

- 다음으로 각 영역별 항목에 대한 만족도를 살펴보도록 하겠다. 먼저 ‘자료 접근성 및 UI의 편의성’ 영역의 만족도 응답 결과는 <그림 26>과 같다. 검색 사이트 및 링크를 통한 접근의 용이성과 사이트 접속 상태의 안정성에 대해서는 대체로 만족하고 있는 것으로 보이나 메뉴 이용의 편리성이나 검색 결과 자료 내려받기의 용이성에 대해서는 다른 두 항목에 비해 만족도가 낮게 나타났기 때문에 향후 이 부분에 대한 보완이 필요한 것으로 보인다.



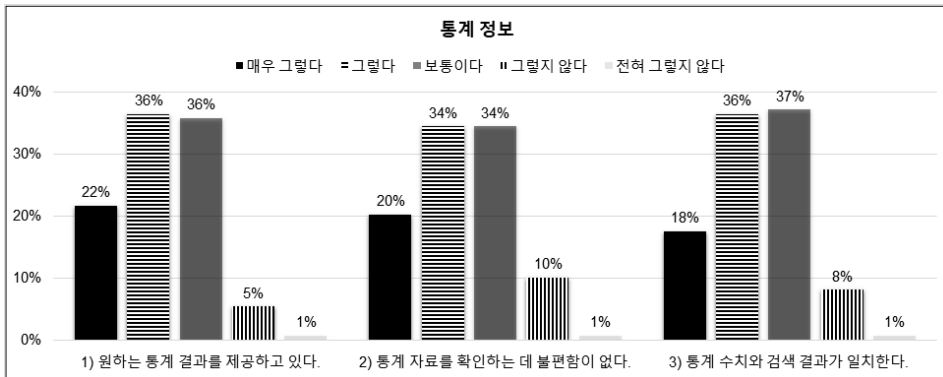
<그림 26> 한국어 학습자 말뭉치 나눔터의 자료 접근성 및 UI의 편의성 만족도 분포

- 다음으로 ‘검색 기능’ 영역의 만족도 응답 결과는 <그림 27>과 같다. 이 영역에서는 네 항목에 대한 만족도가 비슷하게 나타났으며, 대체적으로 만족한다는 응답이 많은 것으로 보인다. 그러나 ‘매우 그렇다’는 의견보다는 ‘보통이다’라는 응답이 좀 더 많은 비중을 차지하고 있기 때문에 이 영역에 대해서도 좀 더 보완을 해야 할 필요가 있을 것으로 보인다.



<그림 27> 한국어 학습자 말뭉치 나눔터의 검색 기능 만족도 분포

- 마지막으로 ‘통계 정보’ 영역의 만족도 응답 결과는 <그림 28>과 같다. 앞의 분석 결과에서 이 영역은 세 영역 중에 비교적 만족도가 낮은 영역이었다. 하위 세 항목에서도 ‘그렇다’는 응답과 ‘보통이다’라는 응답이 비슷한 비중으로 나타나고 있는 것을 볼 수 있다. 앞으로 통계 결과 조회에 대한 편의성과 통계 결과의 신뢰도를 높일 수 있는 방안을 강구해야 할 것으로 보인다.



<그림 28> 한국어 학습자 말뭉치 나눔터의 통계 정보 만족도 분포

3-1. 자료의 접근성 및 UI의 편의성에 대한 개선점

- 자료의 접근성 및 UI의 편의성에서 개선되었으면 하는 점에 대한 질문에 대한 응답은 다음과 같다.

<표 62> 자료의 접근성 및 UI의 편의성에 관한 개선 사항 및 제안 사항

항목	세부 내용
1) 검색 기능 개선(7명)	<ul style="list-style-type: none"> ○ 검색 메뉴 상세화 ○ 검색 조건의 복잡성 ○ 키워드 검색 외 상세 검색 추가
2) 검색 결과 다운로드의 불편함 개선(6명)	
3) 다양한 형태의 통계 수치 제공 필요(3명)	○ 원하는 조건에 맞는 통계 수치 요구
4) 나눔터 사용 가이드 필요(2명)	
5) 검색 결과의 신뢰성 개선(1명)	○ 동일 조건의 검색 결과가 다르게 나올 때가 많음.

3-2. 말뭉치 자료 검색 및 활용에 대한 개선점

- 말뭉치 자료 검색 및 활용에 있어 현재 검색 메뉴의 기본/상세 검색 조건 외에 추가하는 검색 조건에 대한 응답으로는 다음과 같은 것들이 있다.

<표 63> 검색 조건 추가에 관한 제안 사항

항목	세부 내용
1) 학습자 검색 조건 세부화(4명)	○ 성별/연령/전공 등
2) 복수 검색 조건 지정(1명)	○ 원형태/교정형태를 하나 이상 지정 검색
3) 앞뒤 문맥에서 2차 검색 기능 추가(1명)	
4) 형태 주석 말뭉치에서 문장 성분별 검색 조건 추가(1명)	
5) 오류 주석 말뭉치 내에서 형태 주석 검색 기능 추가(1명)	

3-3. 검색 메뉴 사용에 관한 개선점

- 각 검색 메뉴 사용에 있어 학습자 말뭉치 나눔터에서 제공하는 이용 정보만으로 검색 기능을 이해하거나 이용하는 데에 어떤 어려움이 있느냐는 질문에 대한 응답은 다음과 같다.

<표 64> 검색 메뉴 사용에 관한 제안 사항

항목	세부 내용
1) 검색 기능에 대한 교육 필요/메뉴얼 필요(8명)	<ul style="list-style-type: none"> ○ 원하는 정보를 어디에서 어떻게 찾아야 하는지 모름 ○ 검색어 선정에 어려움이 있음(검색어 사용 예시가 없어 불편함) ○ 검색 조건에 대한 상세한 설명 필요
2) 오류 주석 말뭉치에서 어절 단위 검색이 어려움(1명)	
3) 검색 결과를 한 번에 다운로드하기가 어려움(1명)	

3-4. 검색 결과에 관한 개선점

- 검색 결과로 제공되는 정보 중 개선이 필요하거나 추가되었으면 하는 점에 대해서는 다음과 같은 의견이 있었다.

- 중복 검색 제거(4명)
- 검색 결과 한 번에 다운로드(3명)
- 검색 결과의 정확성(2명)
- 오류 기준이 홈페이지에도 간단히 명시되었으면 함(1명)
- 긴 문장 전체 보기(1명)
- 앞뒤 문맥 제공(1명)
- 오류 주석 말뭉치의 양적 확대(1명)
- 구어 말뭉치 내에서의 화자 정보(교원/학습자)(1명)

3-5. 기타 검색 기능에 관한 개선점

○ 그 밖에 검색 기능에서 개선되었으면 하는 점에 대해서는 다음과 같은 응답이 있었다.

- 사용자 매뉴얼(2명)
- 원문 제공(1명)
- 검색어 제안(1명)
- 검색 결과 한 번에 다운로드(1명)
- 호응 관계 검색 기능(1명)
- 앞뒤 문맥 검색 조건 지정 기능(1명)

3-6. 통계 정보에 관한 개선점

○ 통계 정보에서 개선되었으면 하는 점에 대해서는 다음과 같은 응답이 있었다.

- 장르별/국적별/수준별/언어권별 등의 오류 통계 등의 세부 통계 정보 추가(6명)
- 전체 통계와 개별 통계 결과의 불일치 개선(1명)
- 사용자가 지정한 조건에 맞는 통계 정보 제공(1명)
- 오류 분포 및 양상 통계 정보 제공(1명)

4. 학습자 말뭉치 추천 의사와 이유

○ 앞으로 한국어 학습자 말뭉치 나눔터를 계속 이용하거나 다른 사람에게 추천할 의사가 있는가 하는 질문에 79명이 “있다”, 2명이 “없다”고 응답하였으며, 그 이유로 다음과 같은 사항을 들었다. 추천할 의사가 없다고 응답한 경우는 이유를 밝히지 않았다.

- 학습자 언어 사용 양상의 비교 분석 용이함
- 연구 목적으로 활용할 수 있는 공적 자료임
- 개인 말뭉치에 비해 높은 신뢰성과 대표성을 지님
- 습득 연구나 교재 개발, 수업 준비 등에 활용 가능함
- 대규모 자료임
- 학습자 실물 자료를 접할 수 없는 사람들에게 유용함

5. 기타 제안 사항

- 그 밖의 학습자 말뭉치 활용의 효율성과 편리함을 위해 한국어 학습자 말뭉치 나눔터 개선과 관련하여 제안하고 싶은 사항에 대해서 다음과 같은 의견이 있었다.

- 지속적인 자료의 구축과 업데이트(4명)
- 활용 방안 예시 제시 또는 말뭉치 활용에 대한 특강(3명)
- 다양한 자료의 구축(2명)
- 사용 방법 안내/XLM 구조에 대한 매뉴얼(2명)
- 검색 결과 한 번에 다운로드(2명)
- 구어 자료에서 발음 관련 오류 확인의 어려움으로 원자료에서 발음 부분 직접 확인했으면 좋겠음(1명)

(2) 민간 분야 전문가의 집담회

- 민간 분야 전문가와의 집담회는 빅데이터, 인공지능, 에듀테크 등의 기술과 접목한 폭넓은 학습자 말뭉치 활용 가능성 탐색을 목적으로 이루어졌다. 각 분야의 전문가 4명을 초청하고 학계 및 민간 분야의 참가자를 청중으로 하여 진행되었다.

① 패널 구성

<표 65> 민간 분야 전문가 집담회 패널 구성

전문 분야	패널
인공지능	박전규(ETRI 복합지능연구실 실장)
빅데이터	이기황(바이브컴퍼니, 지식&인사이트랩 이사)
에듀테크	이형구(옐로우 크리에이티브 대표)
학습자 말뭉치	곽용진(주) 이르테크 대표)

② 논의 사항

- 4차 산업혁명의 시작이 많은 분야에서 큰 변화를 일으키고 있다. 언어 처리 기술 분야에서 가장 주목할 만한 변화는 무엇인가?
- 현재 에듀테크 기술은 어디까지 와 있는가? 외국어 교육 분야에서 적용

가능한 기술과 개발 사례

- 한국어 교육 또는 외국어 교육 분야에서 인공지능 기술이 어떻게 활용되고 있는가? 일반적인 현황과 구체적인 사례
- 지금까지 한국어 학습자 말뭉치는 주로 한국어 교육 연구와 교육을 위한 자료로 활용되어 왔다. 활용 범위를 기업을 포함한 민간 분야까지 확장한다면 어떻게 활용될 수 있는가?
- 한국어 학습자 말뭉치는 원시 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치의 세 가지 형태로 구축하고 있다. 자료의 활용의 측면에서 원시 말뭉치와 품사 주석 또는 오류 주석을 부착한 말뭉치의 효용성은 어느 정도인가? 민간 분야에서의 활용도를 극대화하기 위한 구축 및 가공 형태는 무엇인가?
- 말뭉치는 많을수록 좋지만 학습자 말뭉치의 특성상 구축 가능한 규모가 한정적일 수밖에 없다. 민간에서의 활용을 위해 필요한 최소 규모는 어느 정도인가?
- 학습자 말뭉치의 양적 확대를 위해서는 효율적인 수집 방법을 모색하는 것이 큰 과제이다. 최신 기술을 적용한 수집 가능성, 방안, 효과
- 자료의 활용을 위해서는 저작권 및 이용에 관한 동의가 필수적으로 요구된다. 최근 민간에서 다양한 유형의 빅데이터를 수집하고 있는데, 이와 관련된 문제를 어떻게 처리하고 있는가?
- 한국어 학습자 말뭉치의 활용 범위를 학계에서 민간까지 확대하기 위해서 구축에서 고려해야 할 사항은 무엇인가?

1.2.3. 종합 및 적용

- 학계 및 민간 분야 등 다양한 집단을 대상으로 한 의견수렴 결과는 말뭉치 구축 및 가공, 배포 및 활용, 교육 및 홍보의 측면에서 정리해 볼 수 있다.

① 구축 및 가공

- 한국어 학습자 말뭉치의 양적 확대와 다양성 확보 필요

설문조사 결과에서 한국어 학습자 말뭉치는 문어와 구어를 아우르는 전체 영역에서 그 사용이 두드러짐을 확인할 수 있었다. 문어 또는 특정 자료의

사용은 전체 사용 빈도에 비해 현저히 낮게 나타났으며, 그중에서도 구어의 사용이 낮게 나타났다. 학습자 수준이나 언어권에서도 특정 변인의 자료보다는 전체 자료를 사용하는 비중이 높았다. 이는 연구 목적에 따른 것일 수도 있지만 또 다른 한편으로는 특정 유형 또는 특정 변인의 자료만을 연구 대상으로 삼기에는 자료의 양이 부족하기 때문일 수도 있다. 이는 서술 답변에서 확인되었다. 이로부터 구어 말뭉치의 양적 확대와 학습자 수준, 언어권별, 장르별, 주제별 자료의 양적 확대 및 다양성 확보가 필요함을 알 수 있다.

○ 형태 주석 및 오류 주석 말뭉치의 확대 필요

형태 주석 및 오류 주석 말뭉치는 자료 검색을 용이하게 해 준다. 교수 현장에서는 이를 이용해 특정한 형태소 및 오류를 검색할 수 있고, 연구자들은 이러한 정보를 통해 특정한 영역의 중간언어 특성이나 언어 발달 과정을 연구할 수 있다. 최근 언어 연구에서도 빅데이터 및 인공지능을 기반으로 한 연구가 확대됨에 따라 대규모의 형태 주석 말뭉치나 학습자가 산출한 발화에 대한 교정문이 포함된 오류 주석 말뭉치에 대한 요구도 커지고 있다. 전문가 집담회에서도 메타버스, 인공지능을 기반으로 한 자동 평가, 맞춤형 교육과정 제공 등의 에듀테크 기술 개발에서의 학습자 말뭉치 활용 가능성이 언급되었는데, 이는 대규모의 자료가 전제가 된다. 이러한 사실들을 통해 형태 주석 및 오류 주석 말뭉치의 비중 확대가 필요함을 확인할 수 있다.

○ 대규모 말뭉치 구축을 위한 수집 방식의 변화

말뭉치의 기본 요건 중 하나는 규모이다. 1차 중장기 계획을 통해 구축한 말뭉치는 원시 말뭉치를 기준으로 약 440만 어절에 이르며, 수집이나 구축 방식을 고려할 때 적은 규모는 아니나 다양한 변인의 자료라는 특성을 고려할 때보다 광범위한 활용을 위해 규모를 지속적으로 확대해 나가야 한다. 이를 위해 교육 기관 및 인력풀을 중심으로 한 기존의 수집 방식에서 용이한 데이터의 공유와 확산을 장점으로 하는 AI나 앱과 같은 도구를 기반으로 한 수집으로의 수집 방식 변화가 필수적이다. 가장 접근이 용이한 방법으로, 국외 학습자 말뭉치의 경우 웹사이트를 통해 학습자의 자율적 참여를 유도하는 사례가 있었는데, 이를 벤치마킹해 볼 수 있을 것이다.

② 배포 및 활용

○ 한국어 학습자 말뭉치 검색 결과 다운로드 방식에 대한 개선 필요

한국어 학습자 말뭉치 자료를 다운로드하여 사용하는 경우 한국어 학습자 말뭉치 나눔터에서 직접 다운로드하기보다는 국립국어원에 직접 요청하여 자료를 받아 사용하는 경우가 더 많은 것으로 나타났다. 구체적인 서술 답변을 통하여 나눔터에서 직접 다운로드를 하는 경우 검색 결과를 한 번에 다운로드하지 못하는 번거로움과 학습자 말뭉치 나눔터에서의 검색 기능 이용에 어려움을 느끼는 이용자가 많기 때문임을 알 수 있었다.

③ 이용자 교육

○ 한국어 학습자 말뭉치와 한국어 학습자 말뭉치 나눔터 사용에 대한 가이드 제공 필요

사용자가 원하는 자료를 찾기 위하여 나눔터를 이용할 때 정보를 확인하거나 검색을 하는 방법을 잘 알지 못하여 이용 시 많은 번거로움을 겪고 있었다. 이에 대해 나눔터에서 상세 가이드를 제공해야 할 필요성이 제기되었다. 또한 주석 정보 제공 연구 목적에 맞게 말뭉치를 가공하는 데에서 어려움을 느끼는 이용자가 많았는데, 이에 대해서 말뭉치 가공에 대한 도구 및 방법에 대한 안내가 필요하다는 의견이 많았다. 이를 통하여 한국어 학습자 말뭉치 검색 결과 다운로드 방식에 대한 개선과 말뭉치 가공 도구 및 방법에 대한 안내가 어느 정도 요구된다는 것을 알 수 있다.

○ 한국어 학습자 말뭉치 아카데미 개최의 다양화 필요

설문조사 결과를 살펴보았을 때 아직까지는 내국인과 연구자 및 서울과 수도권권을 중심으로 한국어 학습자 말뭉치 사용이 이루어지고 있음을 알 수 있었다. 이러한 조사 결과를 통하여 비수도권 지역이나 해외에서의 한국어 학습자 말뭉치 사용 확대를 위하여 한국어 학습자 말뭉치 아카데미 개최의 다양화가 필요함을 확인할 수 있다. 이를 위하여 국내에서의 개최 지역 확대 및 해외에 거주하는 잠재적 이용자를 위하여 온라인 개최 또는 온라인 교육 자료의 배포 등이 이루어질 수 있을 것이다.

2. 학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례 분석

2.1. 기본 방향

- 학습자 말뭉치 구축을 위해서는 설계와 자료 수집, 구축과 가공, 정비, 배포, 활용까지 여러 단계의 작업 절차를 거치게 된다. 말뭉치의 대표성과 균형형성은 설계와 자료 수집 단계에서뿐만 아니라 말뭉치 구축이 완료되기까지 가장 중점적인 쟁점이 되는 문제이다. 그 외에도 말뭉치 구축과 가공을 위한 체계적인 지침의 마련과 효율성을 고려한 작업 공정, 말뭉치의 정확성과 완성도를 높이기 위한 정비 체계와 지침 마련, 효율적인 배포, 활용도 제고를 위한 홍보와 사용자 교육, 활용 모형 개발과 사용자 지침 개발 등 각 단계별로 해결해 나가야 할 다양한 과제가 있다. 본 연구에서는 국외에서 구축한 학습자 말뭉치의 설계, 구축, 배포 및 활용과 관련한 사항에 대한 분석을 통해 한국어 학습자 말뭉치의 향후 개선 방향을 모색해 보고자 하였다.

2.2. 연구 내용

2.2.1. 전 세계의 학습자 말뭉치

- 루베인 대학교의 영어 말뭉치 언어학 센터(Centre for English Corpus Linguistics)에서는 전 세계에서 구축한 186개의 학습자 말뭉치(Learner corpora around the world) 목록을 제공하고 있다. 다음은 개별 사례 분석에 앞서 목록과 함께 제공되는 정보 중 학습자 말뭉치의 기초 정보라고 할 수 있는 이들 자료의 목표 언어, 수집 매체, 숙달도 분포를 개략적으로 정리한 것이다.

- 목표 언어 : 단일어 말뭉치로는 영어가 가장 많으며(101개) 스페인어(14개), 독일어(13개) 등이 그 뒤를 이었다. 또한 2개 이상 언어(Multilingual)에는 병렬말뭉치의 언어도 포함되어 있다.
- 수집 매체 (도구) : 전체 말뭉치 중 문어는 120개, 구어는 42개로 문어

수집이 구어보다 많이 이루어졌다. 문/구어 모두 수집한 말뭉치는 20개이며 수업 녹화본이나 컴퓨터 매개 의사소통(CMC:Computer-Mediated Communication)을 포함한 멀티미디어 말뭉치는 3개였다.

- 숙달도 분포 : 초/중/고급 학습자를 대상으로 한 말뭉치가 70개로 가장 많으며 고급 학습자 대상 33개, 중/고급 학습자 대상 23개로 그 뒤를 이었다. 기타 분류에는 대상자의 숙달도가 구분되지 않은 경우(중등 교육 과정, 수준을 알 수 없는 대학생 등)과 종적 말뭉치를 포함하고 있다.

2.2.2. 주요 학습자 말뭉치의 구축 사례

- 여기에서는 루베인 대학교의 영어 말뭉치 언어학 센터(Centre for English Corpus Linguistics)에서 정리한 전 세계의 학습자 말뭉치(Learner corpora around the world) 186개 중 2백만 어절 이상의 규모, 구글 학술 검색 인용 횟수 100편 이상인 22건의 말뭉치에 대해 살펴보기로 하겠다.²⁶⁾

○ The Jinan Chinese Learner Corpus (JCLC)

JCLC는 중국어 학습자 말뭉치로 L2 학습자 텍스트의 말뭉치를 개발하는 것에 목표를 두었다. 말뭉치 크기는 약 600만의 토큰(token)으로 59 개국의 학습자가 생산한 8,739개 텍스트를 포함하고 있다. 숙달도 수준은 학습 기간에 따라 분류되었고(초급: 1년 이하, 중급: 2-3년, 고급: 3년 이상) 텍스트의 57%는 과제(assignments), 43%는 시험(exams)으로 구성되었다. 공개적으로 배포되지는 않았으나 연구자에게 요청하면 자료를 받을 수 있다.

본 말뭉치의 대부분은 중국 내의 다양한 대학 기관의 어학당에서, 몇몇 데이터는 중국 밖의 대학교에서 수집되었다. 텍스트를 선택할 때 말뭉치의 대표성을 충분히 고려하였으나 학생의 인구통계학적 분포의 불균형성은 현재 학습자 말뭉치에서 지속적인 문제이므로 이러한 제약을 고려해 코퍼스를 설계하였다. 본 말뭉치는 비교적 큰 규모의 코퍼스인 TOEFL11의 숙달도 분포와 비슷하게 중급 학습자가 절반 이상을 차지하는 모습²⁷⁾을 보였다.

이 프로젝트의 목적은 중국어 학습자의 작문 자료 모음을 개발하는 것이며,

26) 수집 및 배포, 메타데이터에 대한 자세한 정보가 필요한 경우 해당 말뭉치 구축 프로젝트의 책임자에게 이메일로 정보를 요청하였다.

27) 반면 국립국어원의 한국어 학습자 말뭉치는 초급 41.4%, 중급 36%, 고급 21.7% 로 중급보다 초급의 숙달도가 상대적으로 높은 분포를 보이고 있다.

연구자는 이 코퍼스를 활용해 언어학자와 다른 연구자들이 중국어 학습자 다국어 분석, 제2 언어 습득(SLA) 또는 교육 자료 개발을 포함한 다양한 주제를 조사하는 데 도움이 될 수 있다고 언급하였다.

○ The AKCES/CZESL (Acquisition corpora of Czech/Czech as a second language) corpus

CZESL은 ‘제2언어로서의 체코어 학습자 말뭉치’로 카를로바 대학교(Charles University)에서 시작한 AKCES(체코어 습득 말뭉치) 프로젝트의 일부로 구성되었다. 2009년-2012년에 주로 학교, 즉 공식적인 환경에서 텍스트를 수집하고 관련 기관 및 개인의 동의를 받아 말뭉치에 포함시켰으며 에세이와 학교 필기시험 원고를 수집하고 스캔하여 전자 형식으로 입력하였다. 원시 말뭉치(plain)는 1,315개의 에세이와 732편의 석사학위 논문으로 약 230만 개의 토큰으로 구성되었다. 말뭉치는 아래와 같이 세 부분으로 구성되어 있으며 각 말뭉치에 따라 수집된 과제의 종류는 다르다.

<표 66> CZESL의 과제 유형

과제 유형	과제 수행 방식
ciz(체코어 학습자가 작성한 텍스트)	다양한 유형과 수준의 언어 교육 수업에서 작성한 에세이
kval(체코어 학습자가 작성한 학술 텍스트)	체코 대학의 석/박사 과정에서 공부하는 학습자들의 학술 자료
rom(루마니아어 배경을 가진 체코 학생들이 작성한 텍스트)	사회적 배제에 의해 위협에 처한 지역 사회에서 루마니아어 배경을 가진 학생이 학교에서 작성한 텍스트

이 말뭉치는 프로젝트를 통해 구축한 말뭉치로 원시 말뭉치를 기본으로 하는 하위 코퍼스들이 많으며(업데이트된 코퍼스 포함 11개) 하위 코퍼스는 주로 원시 말뭉치를 바탕으로 주석을 달거나 확장한 말뭉치이다. 또한 한 홈페이지²⁸⁾에 정리가 되어 있고 말뭉치뿐만 아니라 주석 편집기, 교정기 등의 도구들도 제공하고 있다.

28) <http://utkl.ff.cuni.cz/dokuwiki/doku.php?id=czesl:czesl>

○ Corpus and Repository of Writing (Crow)

Crow는 일리노이 대학교의 인문학 연구를 위한 일리노이 프로그램의 지원을 받아 여러 대학의 연구원들이 협력하여 구축한 영어 학습자 말뭉치이다. 2009년 가을 학기부터 2019년 봄 학기까지의 다양한 국적의 학부생이 입학 후 첫 해에 글쓰기 강좌에서 산출한 문어 말뭉치로 홈페이지에 계속 데이터가 추가되고 있다(2021년 6월 기준, 10,911개의 텍스트 자료 및 100만 단어 이상 포함). 토플 점수 80-105점 사이의 학생들의 과제이며 점수로 숙달도가 구분된다. 본 말뭉치에는 다양한 장르의 텍스트가 있으나 주로 논설문, 연구 제안서, 서사 내러티브(Literacy Narrative) 등의 장르가 주를 이루고 있다. 홈페이지²⁹⁾에서는 글쓰기 모델, 인용 표현(signal phrases)에 대한 수업 계획, 연구 및 분석, 말뭉치 구축 등 이 말뭉치를 다른 분야에 어떻게 적용하여 활용할 수 있는지 예시를 보여 준다.

○ The University of Pittsburgh English Language Institute Corpus (PELIC)

PELIC는 피츠버그 대학교 어학당의 집중 영어 프로그램에서 7년에 걸쳐(2005년~2012년) 수집한 말뭉치로 다양한 언어 배경과 능숙도 수준을 가진 1100명 이상의 학생들이 참여하였다. 현재 PELIC는 깃허브(GitHub)에서 공개적으로 사용 가능한 420만 단어의 학습자 문어 말뭉치를 제공하고 있다. 숙달도는 CEFR의 기준을 적용하여 중/고급 학습자를 네 단계(Pre-Intermediate, Intermediate, Upper-Intermediate, Advanced)로 구분하였다. 수집 과제는 다섯 가지의 수업 유형(문법, 듣기, 읽기, 말하기, 쓰기)과 10개의 질문 유형(단락 쓰기, 짧은 답변, 객관식, 에세이, 빈칸 채우기, 문장 완성하기, 워드 뱅크, 차트, 단어 선택, 오디오 녹음)을 기준으로 분류하였다.³⁰⁾ 텍스트 유형은 주로 단답형(49%) 및 단락 응답(37%)이며 쓰기 수업 제출의 22%는 에세이(총 말뭉치 텍스트의 8%)로 구성되었다.

PELIC은 종단적이며 자연스러운 교실 환경에서 학습자의 언어 개발을 추적하고자 모든 데이터를 자연스러운 교육 환경에서 수집하고자 하였다. 학습자는 1학기~4학기 수업에 참여했으며 평균 학습 기간은 1.8학기로 각 학습자는 평균 39.3개의 총 텍스트를 제출하였다.

29) <https://crow.corporaproject.org/>

30) 수업 유형은 쓰기 수업(63%), 문법(22%) 및 읽기 (14%)이 대부분이며 듣기 및 말하기는 토큰의 3%만을 차지하였다.

학생들은 웹 브라우저 인터페이스를 통해 SQL 데이터베이스의 필드에 텍스트를 입력하고 이 인터페이스는 학생이 등록된 수업을 기반으로 각 과제를 관련 메타데이터(수업 유형, 학생 식별자 등)와 자동으로 연결해 준다. 수집된 데이터는 Perceptron Tagger을 이용한 형태 주석이 이루어졌다. 데이터는 원래 비공개로 저장되었지만 2020년에는 보다 현대적인 원칙에 따라 다른 접근 방식이 채택되어 공개되었다. 또한 음성 공개를 위해 음성 PELIC 데이터를 준비하는 초기 단계에 있다.

○ The Bilingual Corpus of Chinese English Learners(BICCEL)

BICCEL은 중국어와 영어의 이중언어 코퍼스로 영어를 전공으로 하는 4학년 학생들의 국가 시험(말하기 시험)에서 수집된 말뭉치이다. 2001년부터 2005년까지 1,100여명의 학습자가 산출한 말뭉치이며 교실 수업에서의 과제 작문(문어) 또한 수집되었다. Kolesnikova& González-González(2016)에서 정리된 내용 외의 공개 정보는 없다.

○ The Spoken and Written English Corpus of Chinese Learners (SWECCCL)

SWECCCL은 두 개의 하위 말뭉치로 각각 100만 단어 이상 구축되었으며 모두 형태 주석이 되어 있다.

- SECCL: 중국인 학습자들의 영어 구어 말뭉치로 1996년부터 2002년까지 총 7년간 영어 전공 국가 4급 구술시험의 음성 자료 1,148개와 필사본 1,148개를 포함한다. 학습자는 영어 전공 2학년으로 구술 시험은 어학실에서 진행하였다. 시험은 3문항으로 구성되어 있는데 과제 유형은 아래와 같다.

<표 67> SECCL의 과제 유형

과제 유형	과제 수행 방식
이야기 듣고 다시 말하기	응시자는 준비 없이 녹음을 두 번 듣고 3분 이내로 개인적인 경험을 말해야 함
지정된 주제에 따라 말하기	과거 경험과 관련 있는 주제로 말하기. 준비 및 발표 시간은 3분임
대화하기	응시자는 조건을 읽고 A와 B의 역할을 각각 준비해 4분 동안 대화하기

- WECCL: 중국인 학습자들의 영어 문어 말뭉치로 WECCL의 설계는 SECCL과 거의 동일하다. 작문 말뭉치는 주로 9개 대학의 영어 전공 1-4학년 학생들을 대상으로 수집되어 선정된 말뭉치는 대표성의 폭이 넓다고 할 수 있다. 말뭉치 내용은 다양한 주제에 대한 영어 작문이다. 과제의 종류는 논설문, 서사적 글쓰기, 설명문으로 구성되었다(논설문 3,059개, 서사적 글 529개, 설명적 글 90개를 포함하여 총 1,186,215 단어 총 3,578개의 에세이).

○ The Taiwanese Corpus of Learner English (TLCE)

TLCE는 2000년 기준으로 대만에서 가장 큰 주석이 달린 영어 학습자 말뭉치로 알려졌다. 1999년 전후로 수집 및 구축되었고 영어를 전공하는 대만 대학생들의 영어 작문 2105 편 (약 73 만 단어)를 포함하고 있다. 이메일을 통해 수집한 파일과, 학습자가 프린트한 인쇄물, 수기 작성 등으로 수집되었다. 본 말뭉치에는 TOSCA-ICLE 태거(tagger/lemmatizer)[Aarts, Barkema, & Oostdijk, 1997]를 사용하여 다양한 언어적 특징에 대해 주석이 있으며, 각 단어에 해당 보조 정리와 형태론적, 구문론적 및 의미론적 정보가 추가되었다. 학습자의 과제는 크게 두 가지 유형으로 하루 또는 한 주에 가장 걱정되는 되는 일을 기록한 ‘개인 일기(29.4%)’와 ‘작문(70.6%)’으로 구성된다. 작문은 주로 서술문(description), 서사적 글쓰기(narration), 설명문(exposition), 논설문(argumentation)과 같은 다양한 작문 전략을 기반으로 하는 에세이로 구성된다³¹⁾. 기타 유형은 자전적 글 autobiographical writings, 편지, 상상력이 풍부하고 창의적인 글(imaginative writings or creative writings) 등이 있다.

○ The TELEC Secondary Learner Corpus (TSLC)

TSLC는 1994년에 시작한 말뭉치로 홍콩 교사 연합회 TELEC (Teachers of English Language Education Centre) 에서 수집하였다. 2백만 어절 이상의 초등(primary school) 및 중등학교(secondary school) 학생들의 작문 및 구어 발화를 포함하고 있다.

개발 초기 단계에서 교사들에게 어떤 종류의 글을 수집하고 싶은지 말하기보다 모든 것을 가져와서 어떤 패턴이 나타나는지 확인하기로 방식으로 수집하였다.

현재 말뭉치에는 개인적 편지, 공식/비즈니스 편지, 편집자에게 보내는 편

31) 학생들은 보통 대학 첫해에 서술문과 서사적 글쓰기를 배우며, 설명문과 논설문은 두 번째 해와 세 번째 해에 배우게 된다.

지, 신문 또는 잡지 사설, 특집 기사, 연설, 보고서 및 자유 작문과 같은 텍스트 유형을 대표하는 학생 작문이 포함되어 있다. 위의 텍스트 유형 내에서 서사적 글(narratives), 경험적 글(recounts), 서술문(descriptions), 설명문(explanations), 논설문(arguments)과 같은 다섯 가지의 장르로 학습자의 과제를 구분할 수 있다.

TSLC는 TeleTeach용 교육 파일을 개발하는 데 사용되었다. 말뭉치는 학생들의 작문 및 발화 교정 및 수정을 처리하는 기술을 갖추도록 설계된 파일을 개발하는 데 특히 유용한 것으로 입증되었는데, 이러한 조사는 과용, 과소 사용(또는 회피), 오류(어휘, 언어 및 구문), 올바른 사용과 같은 흥미롭고 유용한 사용 패턴을 보여 주었다. 또한 영어 교사들을 위한 지원 사이트(Telenex, <http://www.telenex.hku.hk>)인 TeleNex 홈페이지에서 문법 및 사용 측면에 대한 교사들의 질문에 답변할 때 답변자가 활용할 수 있는 자원이 되기도 하였다.

○ The Japanese Learner English Corpus (NICT JLEC)

JLEC는 일본 국립연구개발법인 정보통신연구기구(NICT)가 주도하여 2004년에 구축 시작한 구어 말뭉치로, 영어 말하기 유창성 시험인 ACTFL-ALC SST의 전사물로 구성되어 있다. 말뭉치는 총 1,281개 샘플로 구성되어 있으며, 이는 120만 어절, 300시간 분량에 해당하는 양이다. 말뭉치에 포함된 샘플은 SST의 숙달도 분류에 따라 9개의 숙달도 그룹으로 나누어져 있다. 본 말뭉치를 구성하는 모든 파일은 인터뷰 참여자 정보, 인터뷰 구조 표지, 말 차례 표지 등의 내용이 주석되어 있으며, 이 중 일부는 문법적, 어휘적 오류에 한해 오류 태깅이 되어 있다. 말뭉치의 배포는 공식 홈페이지를 통한 다운로드의 형태로 이루어지고 있다.

○ The Gachon Learner Corpus

GLC는 2012 ~ 2014년에 걸쳐 가천대 소속의 연구자가 구축한 말뭉치로, 2500명의 참가자가 작성한 25,073건의 텍스트로, 이는 총 250만 어절에 해당한다. 수집 도구로는 구글 폼(Google Forms)을 이용하였으며, 이를 통해 학습자 프로필, 학습자 텍스트를 수집하였다. 수집 대상 정보는 성별, 생년, 학년, 전공, 모어, 부모의 모어, 가정에서 사용하는 언어, 중/고등학교에서 주로 사용한 언어, 영어를 공부한 횟수, 영어권 국가 연수 경험, 토익/토플/아이엘츠 점수, 영어 외의 외국어, 학습자의 영작 능력에 대한 자기 평가, 영어 학

습 동기 등으로 구성되어 있다. 수집된 학습자 자료는 20 여개의 질문에 대한 100 - 150 어절 내외의 답변으로 구성되어 있다. 본 말뭉치는 말뭉치를 구축한 연구자의 개인 블로그에서 배포 중이며, EFL, 언어학 연구에 한정하여 사용이 가능하다는 점을 밝히고 있다.

○ The Michigan Corpus of Upper-level Student Papers (MICUSP)

MICUSP는 Michigan's English Language Institute / Corpus Research Group 에 의해 구축된 영어 학문 목적 글쓰기 말뭉치이다. 본 말뭉치의 총 규모는 260만 어절로, 4개 학문 분야의 16개 전공에 걸친 학부 4학년부터 대학원 3년차까지의 자료로 구성되어 있으며, 원어민과 학습자 자료가 모두 포함되어 있는 것이 특징이다. 말뭉치 전체의 배포 및 접근 관련한 정보는 별도로 공개되어 있지 않으나, MICUSP SIMPLE 이라는 사이트를 통해 말뭉치의 일부를 이용할 수 있다. 해당 사이트에서는 전체 말뭉치 자료 중 A학점을 받은 800여 건의 작문에 대한 색인 검색을 제공하고 있다.

○ The BATMAT Corpus

이 말뭉치는 핀란드의 Åbo Akademi 대학에서 'Advances in Applied Linguistics' 프로젝트의 일환으로 2012년~2017년에 걸쳐 구축한 초고급 학습자 말뭉치이다. 이는 Åbo Akademi 대학의 언어학, 문학, 사회학 전공의 핀란드어/스웨덴어 모어 영어 화자가 작성한 120건의 석사, 박사 논문으로 구성되어 있다. 이 중 석사 논문은 2002 - 2016년 사이의 자료로 구성되어 있으며, 박사논문은 1972-2016년 자료로 구성되어 있다. 말뭉치의 총 규모는 300만 어절 내외로 알려져 있으며, 말뭉치의 배포 및 이용과 관련한 정보는 공개되어 있지 않다. 2021년 현재, 본 말뭉치를 이용하여 작성된 3편의 논문이 발표되어 있다.

○ Cambridge Learner Corpus

CLC는 캠브리지 대학 출판부에서 Cambridge Assessment English와 협업하여 구축한 것으로, Cambridge English 시험 작문을 이용하여 구축되었다. 2021년 현재 공개된 자료에 따르면 총 5000만 어절 규모로, 140여 개 언어권 자료로 구성되어 있으며, 초급 - 고급 학습자 자료를 아우르고 있다. 본 말뭉치의 학습자 자료는 자체 오류 주석 체계에 의해 주석되어 있다. 이 오류 주석 체계의 초기 버전은 공개되어 있으나, 현재 오류 주석에 사용되는 버전은 공개되어 있지 않다. 이 말뭉치는 일반에게 공개되어 있지 않으나, 전체 말뭉치의 일부

를 SketchEngine이라는 사이트를 통해 이용할 수 있다. SketchEngine은 2003년에 설립된 Lexica Computing Limited사에서 개발한 말뭉치 분석/관리 도구로, 이 사이트에서는 주석되지 않은 학습자 말뭉치인 The Open Cambridge Learner Corpus를 공개하고 있다. 이 사이트를 통해 이용할 수 있는 말뭉치는 총 290만 어절 규모로, 간단한 POS 주석과 레마 분석(lemmatizing)이 되어 있다. 본 말뭉치는 캠브리지 대학 출판부의 영어 교재 개발에 적극적으로 활용되고 있다. 또한, CEFR에 상응하는 영어 학습자 숙달도 평가 루브릭을 개발하고자 하는 프로젝트인 English Profile에도 이용되고 있다.

○ The Chinese Learner English Corpus

이 말뭉치는 중국 우한대 영문학과 3학년 학생 263명이 작성한 560개의 에세이(382, 256단어)로 2005년~2007년에 수집되었다. 개인 연구자(Wang)의 박사 논문을 위해 수집된 말뭉치로 Uppsala Student English Corpus(미국)와 직접 비교하기 위해 구축하였다. 다양한 변수(성별, 나이, 언어 배경, 학습 단계, 대상 언어에 대한 노출)에 대한 정보를 수집하고 이후에 CCLE 데이터베이스에 기록하였다. 본 말뭉치와 비교 대상 말뭉치를 통해 중국어와 스웨덴어 학습자의 영어에 대한 대조 연구 가능하다.

○ Corpus Escrito del Español L2

이 말뭉치는 스페인의 그라나다 대학교(University of Granada) 연구팀에서 구축하고 있는 스페인어 학습자 말뭉치이다. 2004년에 설계하여 2006년에 온라인 수집을 시작해 첫 번째 버전은 2017년에 무료로 공개되었고 두 번째 버전은 2020년에 무료로 공개되었다. 현재 가장 큰 스페인어 학습자 말뭉치로 알려져 있으며 4,399명의 참가자와 1,105,936개의 단어를 포함하고 있다. L1 배경이 다양한 학습자³²⁾들의 비교군으로 원어민의 서브 말뭉치도 수집하였다(스페인어권, 영어권, 일본, 포르투갈, 그리스, 아랍계 원어민). 학습자 숙달도는 크게 두 가지 방법을 사용하여 측정하였다. 첫 번째는 위스콘신 대학교(University of Wisconsin, 1998)에서 개발한 43점의 표준화된 대학 수준의 배치시험을 실시하여 6가지 수준으로 분류하였다(Lower beginner, Upper beginner, Lower intermediate, Upper intermediate, Lower advanced, Upper advanced). 두 번째는 학습자가 스스로 측정하여 숙달도를 입력하는 것이다.

32) 영어, 독일어, 네덜란드어, 프랑스어, 포르투갈어, 이탈리아어, 그리스어, 러시아어, 일본어, 중국 아랍어 등

6점 척도에 따라 말하기, 듣기, 읽기, 쓰기 각각의 기술에 대한 스페인어 숙달도를 자체 평가한다. 그런 다음 4개의 평가 점수의 평균을 구해 ‘숙달도 자기 평가(Proficiency self-assessment)’라는 변수를 생성한다. 이밖에도 학습자는 보유하고 있는 스페인어 공인 인증 시험 점수를 입력해 숙달도에 대한 변수를 다양화하였다.

학습자가 작문을 해야 하는 과제 질문은 다음과 같다.

<표 68> Corpus Escrito del Español L2의 작문 주제

번호	제목	과제 설명
1	거주 지역	귀하가 살고 있는 지역은 어떤 지역입니까?
2	유명한 사람	유명한 사람에 대해 이야기하십시오.
3	영화	최근에 본 영화를 요약하십시오.
4	작년 공휴일	지난 여름 휴가 동안 무엇을 했습니까?
5	향후 계획	앞으로의 계획은 무엇입니까?
6	최근 여행	최근에 한 여행에 대해 설명하십시오.
7	경험	최근에 겪은 경험에 대해 이야기하십시오.
8	테러	세계에 테러리즘 문제에 대해 이야기하십시오.
9	금연법	새로운 금연법에 대해 어떻게 생각하십니까?
10	동성 커플	동성 커플이 결혼하고 아이를 입양할 권리가 있어야 한다고 생각하십니까?
11	마리화나 합법화	마리화나가 합법이어야 한다고 생각하십니까?
12	이민	이민에 관한 주요 측면을 분석하세요.
13	개구리	다음 그림을 보고 이야기를 다시 말해 보세요. (지시문) 그림에 나타난 이야기를 들려주세요. 이야기에 새로운 측면을 추가하거나 그림의 일부 측면을 무시할 수 있습니다. 텍스트는 "어느 날..."로 시작해야 합니다.
14	채플린	다음 채플린 비디오 클립을 보고 이야기를 다시 들려주세요. (지시문) 다음 채플린 비디오 클립을 시청하십시오(4분). 이야기를 요약합니다. 비디오 클립은 두 번 이상 볼 수

		있습니다.(https://www.youtube.com/watch?v=4QkTNJFhu-g)
--	--	--

수집과 배포에 관한 내용은 홈페이지(<http://cedel2.learnercorpora.com/>)를 통해 공개되어 있으며 학습자의 언어학적 배경 변수, 과제 변수 등이 자세하여 다양한 변수에 따른 연구가 가능하다. 또한 홈페이지에서 검색 기능을 제공하여 원하는 정보를 검색할 수 있고 다운 받을 수 있으며 무료로 배포하고 있다.

○ French Interlanguage Database

FID는 1998년 Frida 프로젝트의 일환으로 구축되었으며 Sylviane Granger 교수의 ‘영어 말뭉치 언어학 센터’에서 외국어로서의 프랑스어 코퍼스를 통합하기 위해 시작되었다. 현재 20만 단어가 구축되었으며 45만 단어 구축을 목표로 하고 있다. 중급 학습자의 서술문, 논설문, 신문 등의 장르를 포함하고 있다.

이 말뭉치의 구축 목적은 Frida에서 관찰된 학습자 오류를 기반으로 프랑스어 학습을 위한 멀티미디어 도구를 구축하는 것으로 궁극적으로 프랑스어를 사용하지 않는 사용자에게 적합한 자동 교정기의 설계 연구로 이어질 수 있다. 영어 사용자가 작성한 텍스트, 네덜란드어 사용자가 작성한 텍스트, 다양한 모국어 배경을 가진 학습자가 작성한 텍스트의 세 개의 하위 말뭉치로 이루어져 있다. 원시 말뭉치 외에도 오류 주석이 달린 말뭉치를 사용할 수 있다. 오류 주석은 오류 영역(문법, 어휘, 철자법 등), 오류 범주(성별, 숫자 등), 문법 범주(명사, 형용사 등) 세 가지 부분으로 주석되었다.

○ The Tswana Learner English Corpus(TLEC)

TLEC는 2000년 말 수집을 시작하여 2006년 국제 학습자 영어 코퍼스(ICLE)의 배포에 포함된 말뭉치이다. 20만 단어로 구성되었으며 남아프리카 공화국의 다양한 영어의 첫 번째 형태 주석된 말뭉치로 츠와나어를 사용하는 영어 학습자의 논설문으로 구성되었다. 남아프리카에서 현지 인간 언어 기술 산업 육성을 위해 남아프리카 공화국의 영어 및 기타 10개 토착 언어의 더 많은 말뭉치를 수집하고 형태 주석을 목표로 하고 있다.

○ The International Corpus of Learner English (ICLE)

본 말뭉치는 성인(대학교 학부생), 고급 학습자를 대상으로 하는 550만 어절 규모의 영어 학습자 말뭉치로, 제2 언어(ESL)가 아닌 외국어로서의 영어(EFL) 학습자를 수집 대상으로 한다. 2002년의 버전 1을 시작으로, 2009년의 버전 2를 거쳐, 최근에 25개 모어 배경 학습자의 자료를 포함하는 버전 3을 공개하였다. 본 말뭉치를 구성하는 자료는 주로 논쟁적 글쓰기, 문헌 검토 보고서 등으로 이루어져 있으며, 각 언어권별 하위 말뭉치는 200,000어절 내외로 구성되어 있다. 본 말뭉치에서 수집하고 있는 논쟁적 글쓰기의 경우, 응집성, 응결성, 논리 구조 등의 면에서 담화에 집중된 작문을 수집할 수 있다는 점에서 유의미하다고 할 수 있다. 텍스트 유형은 논설문이 94.08%로 대부분을 차지하며, 다음으로는 문학 에세이가 3.82%, 기타 2.1%로 구성되어 있다. 주제 면에서는 루베인 대학에서 제공하는 주제 목록에 포함된 주제가 주로 다루어지고 있으며, 이 에세이들은 보통 시간 제한이 없고(60%) 시험 조건에서 작성되지 않았으며(62%) 에세이의 절반 미만(44%)이 사전 등의 참조 도구의 지원으로 작성된 것이다.

<표 69> ICLE의 에세이 주제

번호	에세이 주제
1	몇몇 사람들은 현대 사회가 과학과 기술, 산업에 둘러싸여 있으며 꿈을 꾸고 상상하는 곳은 더이상 없다고 말합니다. 당신의 의견은 무엇입니까?
2	대부분의 대학 학위는 이론적이며 우리/학생들이 현실 세계를 사는 데 준비되지 못하게 합니다.
3	마르크스는 종교가 대중의 아편이라고 말한 적이 있습니다. 21세기 초에 그가 살아 있었다면 그는 종교를 텔레비전으로 대체했을 것입니다.
4	교도소 시스템은 구식입니다. 문명국가는 범죄자를 처벌해서는 안 됩니다. 범죄자를 갱생시켜야 합니다.
5	옛말: "돈은 모든 악의 뿌리"입니다.
6	빈곤은 아프리카의 HIV/AIDS 전염병의 원인입니다.
7	페미니스트들은 좋은 것보다 여성의 대의에 해를 끼쳤습니다.
8	19세기 빅토르 위고는 "자연이 부르짖는데 인류는 귀를 기울이지 않는다고 생각하는 것은 얼마나 슬픈 일인가"라고 말했습니다. 오늘날에도 여전히 사실이라고 생각합니까?

9	식당에서 흡연을 금지하는 것의 장단점에 대해 토론하십시오.
10	조지 오웰은 그의 소설 "동물농장"에서 "모든 사람은 평등하지만 어떤 사람은 다른 사람보다 더 평등하다"고 썼습니다. 이것이 오늘날 얼마나 사실입니까?

이 말뭉치에 포함된 자료 수집은 협력 관계에 있는 각국 대학을 통해 이루어지고 있으며, 구축이 완료된 말뭉치는 공식 웹사이트를 통해 공개하고 있다. 말뭉치 이용을 위해서는 이용 권한을 별도로 구매해야 한다.

○ The Louvain International Database of Spoken English Interlanguage

이 말뭉치는 International Corpus of Learner English의 구어 대응물을 만드는 것을 목적으로, 1995년부터 구축이 시작된 말뭉치이다. 지정 주제 토론, 자유 대화, 사진 묘사의 3개 과제로 이루어진 50여 건의 인터뷰로 구성되어 있으며, 중고급 - 고급 학습자의 발화로 이루어져 있다. 총 규모는 약 100만 어절이며, 본 말뭉치에 포함되어 있는 학습자 모어로는 아랍어 (사우디 아라비아, 레바논), 바스크 어, 브라질 포르투갈어, 불가리아 어, 중국어, 크로아티아어, 체코어, 네덜란드어, 에스토니아 어, 핀란드어, 불어, 독어, 그리스어, 이란 어, 이탈리아어, 일본어, 리투아니아 어, 노르웨이어, 폴란드어, 스페인어, 스웨덴어, 대만 중국어, 터키어 등이 있다. 2021년 현재, 홈페이지 상에 공개된 정보에 따르면, 본 말뭉치는 CD-ROM으로 배포되고 있다.

○ The Interphonologie du Français Contemporain corpus

이 말뭉치는 불어를 외국어/제2 언어로 구사하는 학습자의 구어를 대상으로 하는 음성 말뭉치로, 2008년에 구축 작업이 시작되었다. 말뭉치를 구성하는 자료는 주어진 단어 읽기, 문장 듣고 반복하기, 짧은 문장 읽기 등의 과제로 구성되어 있으며, 20여 개 언어권의 불어 학습자 자료로 구성된 하위 말뭉치로 이루어져 있다. 각 언어권 하위 말뭉치를 구성하는 학습자의 숙달도 분포나 각 하위 말뭉치의 정량적인 규모에 대해서는 별도의 정보를 제공하지 않고 있다.

공식 웹사이트(<http://cblle.tufts.ac.jp/ipfc/ipfcsearch/>)에서 전체 말뭉치를 열람할 수 있다.

○ English Students' Oral Corpus in Chile

이 말뭉치는 칠레 UCN 대학교의 영어교육학과 학생들을 대상으로 구축한 구어 말뭉치로, 3세대 이상 칠레에 거주하였으며, 영어권 국가에 방문한 적이 없는 학습자 32명을 대상으로 한다. 이들은 대학교 1-4학년으로 구성되어 있으며, 각 학년은 CEFR의 B1 ~ C1에 상응한다. 각 학년별로 남녀 4명씩으로 구성되어 있다. 말뭉치는 15분 길이의 인터뷰 자료로 구성되어 있으며, MS 워드를 이용하여 인터뷰 내용을 전사하되, 별도의 주석 작업은 되어 있지 않다. 말뭉치의 전체 규모는 약 73000어절로, 2015년~2018년에 구축된 분량에 이어 2019~2023에 걸쳐 2차 구축 작업이 진행 중이다. 특히, 구어 말뭉치이기 때문에 구장 구조에 문제가 없으며 조음 관련 장애가 없는 학습자를 대상으로 한다는 점을 명시적으로 밝히고 있다. 현재 구축된 자료를 공개적으로 배포하고 있지는 않으나, 연구 책임자를 통해 자료를 받을 수 있다.

○ The LeaP (Learning Prosody in a Foreign Language) Corpus

이 말뭉치는 독어/영어 학습자의 의미 운율 (semantic prosody) 습득을 연구하고자 하는 목적으로, University of Bielefeld에서 2001년에서 2003년에 걸쳐 구축한 말뭉치로, 약 12시간 분량의 독어 학습자, 영어 학습자, 독어 원어민, 영어 원어민 발화로 구성되어 있다. 구체적으로는 영어 학습자 녹음 183건, 독어 학습자 녹음 176건, 독어 원어민 녹음 10건, 영어 원어민 녹음 8건으로 구성되어 있으며, 원어민 녹음은 학습자와의 대조를 목적으로 포함되었다.

이 말뭉치의 자료 수집 대상은 해외 거주 경험, 발음 관련 수업 수강 경험 등을 고려하여 다양하게 구성되었다. 본 말뭉치 자료는 자유 발화, 주어진 동화 읽기, 읽은 내용 다시 말하기, 존재하지 않는 단어 목록 읽기 등의 과제로 구성되어 있으며, 녹취 자료는 ESPS, Praat를 이용하여 구 - 단어 - 음절 - 세그먼트 - 톤 - 피치 - POS - lemma 의 8개 층위로 나누어 주석되었다. 주석된 말뭉치 자료는 <https://sourceforge.net/projects/leapcorpus/> 을 통해 이용이 가능하다.

루베인 대학교의 영어 말뭉치 언어학 센터(Centre for English Corpus Linguistics)의 전 세계의 학습자 말뭉치(Learner corpora around the world) 186개 중 구축 사례를 통해 상세히 소개한 200만 어절 규모 이상, 구글 인용 빈도 100 이상의 학습자 말뭉치 개요는 다음과 같다.

<표 70> 국외의 학습자 말뭉치 구축 개요: 200만 어절 규모 이상, 구글 인용 빈도 100 이상 학습자 말뭉치

말뭉치명	목표 언어	수집 시기	규모	문/구어	태깅 정보	숙달도 분포	자료 유형	수집 및 배포 특징
The Jinan Chinese Learner Corpus (JCLC)	중국어	2006~	600만	문어	없음(원시 말뭉치)	초/중/고(학습 기간에 따른 분류)	과제 작문, 시험 작문	연구자에게 연락하면 자료를 받을 수 있음.
The AKCES/CZESL (Acquisition corpora of Czech/Czech as a second language) corpus	체코어	2009~2012	230백만	문어	형태 주석/오류 주석/구문 주석	CEFR	에세이, 석사학위 논문	배포는 2012년~2020년에 이루어짐. 원시말뭉치를 기준으로 다양한 주석 말뭉치들이 하위 코퍼스를 이룸. 부분적으로 온라인 검색 및 전체 다운로드가 가능함.
Corpus and Repository of Writing (Crow)	영어	2009~2019	100만+	문어	없음(원시 말뭉치)	중-고급(토플 점수 80점 이상)	과제 작문	아카이브 형식으로 홈페이지에(https://crow.corporaproject.org) 계속 데이터가 추가됨

말뭉치명	목표 언어	수집 시기	규모	문/구어	태깅 정보	숙달도 분포	자료 유형	수집 및 배포 특징
The University of Pittsburgh English Language Institute Corpus (PELIC)	영어	2005~2012	420백만	주로 문어	형태 주석	CEFR(중-고급)	교실 작문	깃허브에서 무료 배포 중 (https://github.com/ELI-Data-Mining-Group/PELIC-dataset)
The Bilingual Corpus of Chinese English Learners (BICCEL)	영어	2001~2005		주로 구어	-	고급(영어 전공 4학년 학생)	시험작문, 과제작문	중국어와 영어의 이중 코퍼스이며 공개 정보가 부족함
The Spoken and Written English Corpus of Chinese Learners (SWECCCL)	영어	1996~2002	200만	문/구어	형태 주석	고급(영어 전공 학습자의 말하기 시험)	시험작문	구축은 2003년부터 함.
The Taiwanese Corpus of Learner English (TLCE)	영어	1999년 전후	73만	문어	형태 주석 구문주석 의미주석	고급(영어 전공 대학생)	개인일기, 과제작문	수집/배포 관련 정보 X
The TELEC Secondary Learner Corpus (TSLC)	영어	1994~??	200만	문/구어		초등/중등학교 학생	교실작문, 과제작문	

말뭉치명	목표 언어	수집 시기	규모	문/구어	태깅 정보	숙달도 분포	자료 유형	수집 및 배포 특징
The Japanese Learner English Corpus (NICT JLE)	영어	2004 - ??	120만	구어	구어주석, 오류주석	초급 - 고급	ACTFL-ALC SST 테스트	홈페이지에서 다운로드 가능 학습자 언어와 비교하기 위한 원어민 하위 말뭉치를 포함함
The Gachon Learner Corpus	영어	2012-2014	250만	문어	-	-	자유작문	연구자 개인 블로그에서 배포
The Michigan Corpus of Upper-level Student Papers (MICUSP)	영어		260만	문어	-	중급-고급 (학부 4학년 ~대학원)	보고서	전체 자료의 일부를 웹사이트를 통해 공개
The BATMAT Corpus	영어	2012-2017	300만	문어	-	고급(대학원)	학위논문	
The Cambridge Learner Corpus (CLC)	영어		5000만 +	문어	형태 주석/오류 주석	-	시험작문	말뭉치 전체는 공개 X SketchEngine이라는 웹사이트에서 주석되지 않은 버전을 일부 공개

말뭉치명	목표 언어	수집 시기	규모	문/구어	태깅 정보	숙달도 분포	자료 유형	수집 및 배포 특징
The Chinese Learner English Corpus	영어	2005~2007	38만	문어	-	고급(영문과 3학년 학생)	과제작문	
Corpus Escrito del Español L2	스페인어	2006~	100만+	문어	-	다양하게 수집	지정작문(그림 보고 이야기 완성하기, 문법 문제 등)	온라인(구글 설문)을 통해 다양한 학습자의 작문을 수집하고 있으며 홈페이지(http://cedel2.lea.mercorpora.com/)에서 무료로 공개 중(검색 기능 제공)
French Interlanguage Database	프랑스어	1998~?	20만	문어	형태 주석/오류 주석	중급	지정작문(서술문, 논설문 등)	
The Tswana Learner English Corpus	츠와나어	2000~2006	20만	문어	형태 주석	-	지정작문(논설문)	
The International Corpus of Learner English	영어	2002~현재	550만	문어	형태 주석	중고급-고급	시험작문	웹사이트 통해서 공개 이용 권한을 판매함

말뭉치명	목표 언어	수집 시기	규모	문/구어	태깅 정보	숙달도 분포	자료 유형	수집 및 배포 특징
The Louvain International Database of Spoken English Interlanguage	영어	1995-	100만	구어	-	중고급-고급	지정주제토론, 자유 대화, 사진 묘사	
The Interphonologie du Français Contemporain corpus	프랑스어	2008 -		구어	-	-	반복, 읽기, 구어 생산	사이트 개설 준비 중
English Students' Oral Corpus in Chile	영어	2014-	7만	구어	-	중급-고급	인터뷰	2019-2023에 걸쳐 추가 구축 중
The LeaP Corpus	영어, 독어	2001-2003	12시간 분량	구어	형태 주석	중급-고급	자유 발화, 읽기, 읽은 내용 다시 말하기 등	홈페이지에서 전체 공개

2.2.3. 학습자 말뭉치의 활용 사례

○ Wang(2015)에서는 학습자 말뭉치를 이용해 7가지의 활용 방법이 있다고 제안한 바 있다. 이 외에도 여러 연구에서 학습자 말뭉치 데이터를 다양한 분야에서 적용하려는 시도가 계속되고 있으며 이는 다양한 연구 성과로 축적되고 있다. 본 절에서는 Wang(2015)과 그 밖의 선행연구 등을 통해 국외 학습자 말뭉치 활용에 관한 사례를 소개하고자 한다.

○ 자동 에세이 채점

자동 에세이 채점은 대규모 학습자 데이터와 NLP 방법(Yannakoudakis et al., 2011)을 사용하여 숙달도 수준 간의 차이를 조사하는 데 의존하는 활발한 연구 영역이다. 숙달도 데이터가 포함된 경우 JCLC를 사용하여 현재까지 수행되지 않은 자동 채점 기술을 중국어로 확장할 수 있다.

○ 오류 감지 및 수정

학습자 말뭉치 데이터에 대해 훈련된 오류 감지 및 교정 시스템을 구축하는 연구가 증가하고 있다(Dahmeier and Ng, 2011; Han 등, 2010). 최근에는 학습자 텍스트를 이용한 한자어순 오류 감지 및 수정(Cheng et al., 2014)까지 확대됐다. 큰 규모의 JCLC는 오류 주석을 추가하여 활용할 수 있다.

○ 학습자 모국어(L1) 식별

다른 언어로 작성된 글을 바탕으로 학습자의 모국어를 추론하는 작업이다(Malmasi and Dras, 2015). 이 작업은 주로 학습자 말뭉치에 의존하며 JCLC는 여기에 직접 적용될 수 있다. 최근 NLI 공유 작업(Tetreault et al., 2013)에 대한 검토에 좋은 개요가 나와 있다. NLI 방법은 이미 아랍어와 핀란드어를 포함한 다른 언어에서 검증되었다(Malmasi and Dras, 2014a; Malmasi and Dras, 2014b).

○ 전이 가설 연구

연구원들은 최근에 머신 러닝 및 NLP와 결합된 데이터 기반 기술을 사용하여 학습자 말뭉치에서 언어 전이 가설을 추출하는 방법을 조사했다(Swanson and Charniak, 2014).

○ 제2언어 습득 연구

JCLC에는 다양한 L1이 존재하여 서로 다른 모국어 간의 대조적인 언어 간 분석이 가능하다. 서로 다른 L1-L2 조합의 대규모 데이터를 이용할 수 있으므로 다른 학습자를 추론 가능하게 하는 광범위한 언어 습득 연구가 가능하다.

○ 교육학적 자료 개발

학습자 말뭉치는 어려운 영역을 식별하고 자료 개발자가 다른 그룹에 속한 학생들의 강점과 약점을 고려한 자료를 만들 수 있도록 하는 데 사용되었다 (McEnery and Xiao, 2011). 이것은 또한 코퍼스에서 파생된 지식을 사용하여 설계 프로세스를 안내할 수 있는 강의 계획서 개발로 확장될 수 있다. 언어 전이 분석과 결합된 학습자 데이터는 필요 기반(need-based) 및 데이터 주도(data-driven) 접근 방식 내에서 교육 자료 개발을 지원하는 데 사용할 수 있다. 언어 사용 패턴이 밝혀지면 교육 가능성을 평가하고 모국어별 맞춤 연습 및 교육 자료를 만드는 데 사용이 가능하다.

○ 자동 평가 생성

위에서 언급한 오류 감지 및 언어 전이 추출 방법과 결합된 이 데이터는 테스트 자료(예: Cloze 테스트)를 자동으로 생성하는 데 사용할 수 있다. 이러한 접근 방식에 따라 Sakaguchi et al. (2013)은 대규모 영어 학습자 데이터를 사용하여 언어 학습자를 위한 빈칸 채우기 퀴즈 항목을 만들었다. 이 분야의 이전 연구에서도 언어 테스트를 위한 객관식 문제의 자동 생성을 고려했지만(Hoshino and Nakagawa, 2005), 학습자 데이터는 없었다. 자연적으로 생성된 오류를 포함하는 학습자 말뭉치의 사용은 훨씬 더 좋은 시너지를 가지고 관사, 전치사 및 동의어를 넘어 더 복잡한 언어 오류를 평가할 수 있다. 현재 오류에 대한 추가 주석으로 JCLC를 이러한 작업에 사용할 수 있다.

○ 학습자 오류 사전 편찬

다음으로는 말뭉치를 이용한 학습자 오류 사전 편찬을 제안한 연구가 있다. CLC(The Cambridge Learner Corpus)는 1600만 단어로 구성되었으며 86개의 다른 모국어 학습자의 영어 시험 자료를 수집해 모든 오류를 주석하였다. 8개의 EFL 시험으로 구성되며 일반 영어와 비즈니스 영어를 포함하고 있다. 총 600만 단어에 오류 주석이 되어 있으며, 주석 체계는 두 글자

(two-letter) 시스템으로 첫 번째 글자는 오류 양상(ex.철자 오류, 생략), 두 번째 글자는 품사 분류를 나타낸다.

말뭉치는 두 명의 연구자에 의해 수동으로 주석되었으며, 한 연구자가 두 번째 작업을 감독하여 주석의 일관성 문제를 최소한으로 유지하였다. 주석을 통해 얻은 정보는 영어 학습자를 위한 도구 개발에 사용하기 위해 심사 기관, 교사, 사전 편찬자, 연구원 및 ELT 저자에게 전달된다. 주석은 그 자체가 목적이 아니라 오류가 반복적으로 발생하는 문맥에 대한 책갈피 역할을 한다. 또한 가능한 한 오류와 함께 추가되는 교정형의 추가 기능이 매우 중요하다. 오류를 '해석'하거나 다른 말로 바꾸어 표현하지 않도록 주의하고, 상대적으로 확실하고 명확하게 대체가 가능한 경우에만 교정형을 추가한다.

CUP에서 코딩된 말뭉치에서 수집된 데이터는 사전 프로젝트를 수행하는 사전 편찬자와 연구자가 학습자에게 특히 문제가 되는 단어와 구성을 식별하는 데 사용된다. 수집된 말뭉치 정보를 바탕으로 학습자에게 그러한 단어나 구성의 올바른 사용을 지시하는 최선의 방법에 대한 결정이 내려진다. 영향을 받는 언어 그룹에 대한 정보도 관련이 있으며 특정 오류가 발생하는 시험 수준도 주목할 필요가 있다. 이 검색 옵션은 예를 들어 어떤 오류가 일반적으로 기본 수준 오류인지 확인하고 여러 숙달도에서도 여전히 지속되는 보다 고집스러운 오류를 식별할 수 있게 해주기 때문에 특히 유용하다.

○ 자동 숙달도 판별

Hasan et al(2008)에서는 기계 학습 알고리즘을 통해 자동 숙달도 검사를 제안하였다. 본 연구에서 사용한 코퍼스는 일본어 학습자의 영어 코퍼스(NICT JLE)이며 120만 단어로 구성된 구어 말뭉치이다. 46개의 오류 태그와 11개의 유창성 관련 기능을 이용하여 가장 효과적인 유창성 및 오류 기능 17개³³⁾를 추출하였다. 이러한 기능을 사용한 알고리즘으로 초급-중급-고급 숙달도를 판별한 결과 전체 예측 정확도는 약 90%를 보였다. 한편 고급 수준에 대한 예측 정확도는 다른 그룹에 비해 낮게 나왔는데(80%) 이는 고급 수준을 판별하는 것은 어휘와 정교하고 복잡한 문장 구조 등에 의해 결정되기 때문이라고 해석할 수 있다.

33) 각 인터뷰의 단어 수, 문장 당 단어 수, 각 인터뷰의 문장 수, 일본어 발화 수, 관사 오류 수, 긴 휴지 횟수, 동사 일치 오류의 수, 명사 오류의 수, 동사 상 오류의 수, 양적 형용사 오류의 수, 보어 오류의 수, 전치사 오류의 수, 컷오프 수, 형용사 오류의 수 (긍정, 비교, 최상급 형용사), 자가 수정 횟수, 형용사 오류의 수, 명사 오류의 수

○ 학습자 언어 프로필(profile)³⁴ 분석 및 평가 루브릭(Rubric) 개발

영어 프로필 단어 목록(English Profile Wordlists) 프로젝트는 캠브리지 대학 출판부(Cambridge University Press)에서 말뭉치를 활용해 2007년~2012년 동안 A1 수준부터 C2 수준의 단어를 목록화한 연구이다. 주로 교사, 교사 트레이너, 시험 준비자, 자료 작성자 및 강의 계획서 작성자를 위한 온라인 어휘 자료로 단어, 구, 구동사 및 관용구의 CEFR(Common European Framework of Reference) 수준에 대한 정보를 제공하며 매년 추가되어 현재 약 15,000개의 표제어를 포함하고 있다. 이 프로젝트는 학습자가 ‘알아야 할 것’을 목록화하기 보다는 학습자가 ‘실제로 알고 있는 것’에 초점을 맞춰 수집되었다. 이를 위해 학습자 말뭉치인 캠브리지 학습자 말뭉치(Cambridge Learner Corpus, 5천만 단어)와 모국어 말뭉치인 캠브리지 영어 말뭉치(Cambridge English Corpus, 1.2 - 10억 단어)를 활용하였다. 이 프로젝트는 사전 편찬 연구에서 파생하였기 때문에 단어를 수집할 때 사전 편찬자가 주어진 의미의 발생 횟수에 따라 E, I, A(Essential, Improver, Advanced)의 세 가지 상대 빈도 수준 중 하나를 할당하는 방식으로 수집되었다. 공식 홈페이지(<https://www.englishprofile.org/>)에서 어휘와 문법 프로필을 검색할 수 있다.

2.3. 종합 및 적용

- 학습자 말뭉치 구축·정비, 배포·활용 관련 선진 사례 분석 결과는 학습자 말뭉치의 균형성 확보를 위한 과제 유형과 숙달도 분포, 배포와 활용으로 나누어 정리해 볼 수 있다.

(1) 과제 유형

- 전반적으로 살펴보았을 때, 상대적으로 말뭉치 수집 기준을 명확하게 밝혀놓은 ICLE를 제외하면 말뭉치에 포함된 샘플의 주제, 각 샘플이 작성된 환경, 말뭉치 수집 대상자에게 주어진 과제 등에 대해서 명확하게 밝혀놓은 경우가 그리 많지 않음을 알 수 있었다. 반면에, 말뭉치가 어떤 장르의

34) 텍스트에서 발견되는 각 단어, 구, 관용구 및 언어의 수준을 포함하는 자료

작문으로 구성되었는지에 대해서는 상대적으로 명확하게 밝혀놓은 경우가 많았다. 예를 들어 PELIC의 경우 자연스러운 교실 환경에서 인위적으로 제한되지 않은 데이터를 수집하였는데 결과적으로 이러한 데이터는 과제 유형, 텍스트 길이 등 여러 가지 면에서 균질하지 못하다고 볼 수 있다. 이를 통해, 문어 말뭉치의 경우 수집 과정에서 학습자에게 주어지는 과제 자체의 설계보다는 수집 대상 작문의 주제, 장르와 같은 담화적 측면에 좀 더 중점을 두고 수집이 이루어진다는 것을 알 수 있었다.

- 흥미로운 현상 중 하나는 특정한 종류의 작문으로만 이루어진 말뭉치가 무척 드물다는 것이다. 대부분의 학습자 문어 말뭉치가 시험 작문과 과제 작문의 혼합으로 이루어져 있으며, 장르 면에서도 논설문, 서술문 등 다양한 장르로 구성되어 있음을 알 수 있었다. 특히, 사례 분석 대상 말뭉치 중 시험 작문으로만 이루어진 말뭉치는 Cambridge Learner Corpus (이하 CLC) 하나뿐이었는데, CLC는 어학 능력 평가를 목적으로 하는 공인 시험의 작문 자료를 모은 말뭉치라는 점에서 시험 작문을 포함하고 있는 다른 말뭉치와 그 성격이 다르다고 볼 수 있다.
- 이와 더불어, 각 말뭉치를 구성하고 있는 과제 작문이 모두 동질적인 성격을 지닌 것은 아니라는 점에 주의를 기울일 필요가 있을 것으로 보인다. 이를테면, 수업 중에 작성한 과제 작문이더라도 언어 능력을 테스트하기 위해 주어진 과제 작문과, 목표어를 통해 학습을 수행하는 학습자가 전공 과목 등에서 작성한 과제 작문은 그 성격이 다를 것으로 추정할 수 있다.
- 한편, 학위 논문, 학술 논문, 수업 페이퍼의 단일 장르로 이루어진 말뭉치가 복수 존재하는 것 역시 흥미로운 현상이라고 할 수 있을 것이다. 이는 학술적 글쓰기, 특수 목적 외국어 교육이라는 특별한 수요의 존재를 나타내는 동시에, 학습자 문어 연구에서 담화 영역에 대한 연구자들의 관심을 반영한다고 볼 수 있다.
- 구어 말뭉치의 경우, 문어 말뭉치에 비해 그 규모는 작지만 과제 구성의 측면에서 볼 때 좀 더 정교하게 구성되어 있었다. 구어 말뭉치 구축 과정에서 주어진 과제는 크게 두 가지로 나눌 수 있었는데, 첫째, 인터뷰, 지정 주제 토론, 사진 묘사, 자유 대화와 같이 학습자 언어의 사용 양상 자체에 중점을 두고 구성된 한 부류와, 반대로 학습자 구어에서 나타나는 음성학적 현상 자체에 관심을 두고 이를 포착하기 위해 단어 목록 읽기, 문장 반복하기와 같이 미리 설계된 과제를 이용하여 자료 수집을 진행한 한 부류로 나누어 볼 수 있었다. 후자의 경우, 말뭉치 수집에 사용한 읽기 자

료, 단어 목록 등을 말뭉치와 함께 공개하는 경향이 있었다.

(2) 숙달도 분포

- 분석 대상 학습자 말뭉치의 숙달도 기준은 말뭉치마다 상이하여 정량적인 비교를 하기에 적절하지 않다고 할 수 있다. 그럼에도 불구하고, 분석 대상 학습자 말뭉치의 숙달도 기준을 비교해 보았을 때 일반적인 초급-중급-고급의 평가 단위를 사용한 경우와, 학습 기간을 기준으로 한 경우, 학령/학년을 기준으로 숙달도를 판별한 경우로 나누어 볼 수 있었다.
- 일반적인 초급-중급-고급의 평가 단위를 사용하여 숙달도를 기재한 말뭉치의 경우, 이러한 숙달도 판별이 어떤 기준에 의한 것인지 명확하게 밝힌 사례가 드물었다. 반면, 유럽에서 구축된 학습자 말뭉치의 경우 CEFR(Common European Framework of Reference, 유럽 공통 언어 기준)을 기준으로 숙달도를 기재한 경우가 있었다.(PELIC, The AKCES/CZESL corpus)
- 그러나 같은 숙달도 레벨의 자료를 모은 말뭉치라고 하여도, 이들이 실제로 같은 수준의 학습자 자료를 나타낸다고 판단하기는 어렵다. 예를 들어, ICLE의 경우, 해당 말뭉치 수집의 목적이 고급 학습자 자료를 모으는 것에 있었기 때문에 따라서 대학교 3, 4학년 내지는 대학원 첫 학기에 재학 중인 학생을 대상으로 자료를 수집하였다. 일반적인 숙달도 기준에 따르면 중고급 (upper intermediate) - 고급(advanced)로 보는 것이 적절할 것으로 예상 되었으나, 각 언어권 자료 중 임의로 추출한 20건에 대해 CEFR 기준에 맞추어 채점 시도, 언어권별로 숙달도 분포가 상이하게 나타나는 경향을 보였다.
- 대학의 과제 작문, 시험 작문으로 구성된 말뭉치의 경우 목표어 전공 여부, 대학에서의 학년, 토플 점수 등을 기준으로 숙달도를 기재하는 경우가 있었다 (Corpus and Repository of Writing, The Chinese Learner English Corpus, The Taiwanese Corpus of Learner English, The Michigan Corpus of Upper-level Student Papers, The BATMAT corpus 등)
- 반면에 초/중등 교육 과정에 재학 중인 학습자 자료로 구성된 학습자 말뭉치는 매우 적다는 것을 알 수 있었다. 본 연구의 분석 대상 말뭉치 중에는 The TELEC Secondary Learner Corpus가 유일한 사례에 해당한다.

(3) 배포와 활용

- 본 연구에서 살펴본 해외 학습자 말뭉치의 배포 방식을 살펴보면 크게 일반에 공개하지 않은 경우, CLC와 같이 말뭉치의 사용 권한에 대한 비용 지불이 필요하거나 접근 권한을 제한하는 경우, Github, SourceForge, 블로그 등의 매체를 이용하여 무료로 배포하는 경우, 말뭉치 배포를 위한 별도의 웹사이트/웹 도구를 이용하는 경우로 나누어 볼 수 있었다.
- Github, SourceForge, 클라우드 서비스, 블로그 등을 이용하거나 말뭉치 구축 연구자 및 책임자와의 연락을 통해 말뭉치를 배포하는 경우는 대개 연구자 개인 또는 소수의 연구자 그룹이 구축한 말뭉치의 대부분이 이 경우로 말뭉치의 유통을 관리하기 어려우며, 해당 말뭉치를 사용하고자 하는 연구자/교사가 말뭉치 관련 툴의 사용에 익숙하지 않으면 활용하기가 어렵다는 문제가 있다.
- 이와 달리 별도의 웹사이트 및 도구를 이용하는 경우는 말뭉치가 유통되는 과정에서 손상되는 일을 막을 수 있으나 사용자의 편의를 위해 웹사이트의 유지와 보수, 관리가 지속적으로 잘 이루어져야 한다는 어려움이 있다. 한국어 학습자 말뭉치의 경우 학습자 말뭉치 나눔터를 통해 자료를 검색하고 내려받을 수 있도록 하고 있는데, 자료의 업데이트 및 업데이트 정보 관리, 지속적인 모니터링과 의견수렴을 통한 사용상의 편의성 및 효율성 제고가 이루어져야 할 것이다.

3. 2015-2020년 한국어 학습자 말뭉치 연구 및 구축 성과 검토

3.1. 기본 방향

- 2015-2020년 국어원 한국어 학습자 말뭉치 연구 및 구축 성과는 한국어 학습자 말뭉치 구축을 위한 2차 중장기 계획 수립을 위한 방향성을 모색하기 위한 기초 자료로 활용할 수 있다. 또한 그 계획을 효율적으로 실행하기 위한 전략 수립의 틀을 제공해 준다. 이에 본 연구에서는 2015-2020년 국어원 한국어 학습자 말뭉치 연구 및 구축 성과에 대한 계량적 분석을 바탕으로 말뭉치 구축 이론과 실제 활용의 측면에서 비판적으로 검토하였다.

3.2. 연구 내용

3.2.1. 말뭉치 유형별 구축 비율

- 2015-2020년 국어원 한국어 학습자 말뭉치는 원시 말뭉치 4,400,369어절, 형태 주석 말뭉치 3,503,729어절, 오류 주석 말뭉치 1,002,025어절이 구축되었다. 이 중 형태 주석 말뭉치는 원시 말뭉치의 79.6%(문어 76.3%, 구어 89.3%), 오류 주석 말뭉치는 22.8%(문어 15.3%, 구어 44.6%)를 차지한다.

<표 71> 2015-2020년 한국어 학습자 말뭉치 유형별 구축 통계

구분	자료 유형	1급	2급	3급	4급	5급	6급	6급 이상	합계
원시	문어	385,178	536,162	629,223	600,711	576,786	406,679	143,861	3,278,600
	구어	206,596	232,282	226,433	194,674	115,057	87,179	59,548	1,121,769
	합계	591,774	768,444	855,656	795,385	691,843	493,858	203,409	4,400,369
형태주석	문어	357,025	428,296	446,631	420,466	409,840	368,164	71,836	2,502,258
	구어	197,396	195,761	202,717	176,599	110,986	86,932	31,080	1,001,471
	합계	554,421	624,057	649,348	597,065	520,826	455,096	102,916	3,503,729
오류주석	문어	83,301	90,745	88,172	83,233	77,167	74,802	3,773	501,193
	구어	89,895	96,733	87,967	92,820	71,541	54,808	7,068	500,832
	합계	173,196	187,478	176,139	176,053	148,708	129,610	10,841	1,002,025

3.2.2. 대상별 · 수준별 말뭉치의 구축 비율

(1) 원시 말뭉치

- 원시 말뭉치는 국내 교육기관의 학습자 자료 3,162,760어절, 이주민 학습자 자료 397,812어절, 국외 학습자 자료 240,428어절이 구축되었다. 구축 비율로 환산하면 국내 학습자 자료는 85.5%(문어 96.5%, 구어 53.4%), 이주민 자료는 9%(문어 9.4%, 구어 7.9%), 국외 자료는 5.5%(문어 19.0%,

구어 5.5%)를 차지하였다.

(2) 형태 주석 말뭉치

- 형태 주석 말뭉치는 국내 교육기관의 학습자 자료 2,890,840어절, 이주민 학습자 자료 389,167어절, 국외 학습자 자료 196,712어절이 구축되었다. 구축 비율로 환산하면 국내 학습자 자료는 82.5%(문어 95.4%, 구어 50.4%), 이주민 자료는 11.1%(문어 3.5%, 구어 30%), 국외 자료는 6.4%(문어 1.1%, 구어 19.6%)를 차지하였다.

<표 72> 2015-2020년 한국어 학습자 말뭉치 대상별·수준별 통계: 형태 주석 말뭉치

자료 유형	수집 대상	1급	2급	3급	4급	5급	6급	6급 이상	합계
국내	문어	333,941	397,708	425,114	397,234	399,783	364,989	67,748	2,386,517
	구어	64,714	81,345	93,296	90,359	83,453	68,984	22,172	504,323
	합계	398,655	479,053	518,410	487,593	483,236	433,973	89,920	2,890,840
이주민	문어	15,701	18,440	19,456	21,902	10,057	3,175	0	88,731
	구어	56,516	76,388	67,247	67,378	19,093	13,814	0	300,436
	합계	72,217	94,828	86,703	89,280	29,150	16,989	0	389,167
국외	문어	7,383	12,148	2,061	1,330	0	0	4,088	27,010
	구어	76,166	38,028	42,174	18,862	8,440	4,134	8,908	196,712
	합계	83,549	50,176	44,235	20,192	8,440	4,134	12,996	223,722
전체 합계		554,421	624,057	649,348	597,065	520,826	455,096	102,916	3,503,729

(3) 오류 주석 말뭉치

- 오류 주석 말뭉치는 국내 교육기관의 학습자 자료 776,527어절, 이주민 학습자 자료 95,105어절, 국외 학습자 자료 130,393어절이 구축되었다. 구축 비율로 환산하면 국내 학습자 자료는 77.5%(문어 87.2%, 구어 67.8%), 이주민 자료는 9.5%(문어 7.6%, 구어 11.4%), 국외 자료는 13.0%(문어

5.2%, 구어 20.8%)를 차지하였다.

<표 73> 2015-2020년 한국어 학습자 말뭉치 대상별·수준별 통계: 오류 주석 말뭉치

자료 유형	수집 대상	1급	2급	3급	4급	5급	6급	6급 이상	합계
국내	문어	71,045	71,820	76,118	70,333	73,272	74,310	80	436,978
	구어	43,047	55,231	60,106	65,013	61,238	50,022	4,892	339,549
	합계	114,092	127,051	136,224	135,346	134,510	124,332	4,972	776,527
이주 민	문어	5,109	6,818	10,101	11,570	3,895	492	0	37,985
	구어	9,074	15,042	10,081	16,267	3,699	2,957	0	57,120
	합계	14,183	21,860	20,182	27,837	7,594	3,449	0	95,105
국외	문어	7,147	12,107	1,953	1,330	0	0	3,693	26,230
	구어	37,774	26,460	17,780	11,540	6,604	1,829	2,176	104,163
	합계	44,921	38,567	19,733	12,870	6,604	1,829	5,869	130,393
구어 합계		89,895	96,733	87,967	92,820	71,541	54,808	7,068	500,832
전체 합계		173,196	187,478	176,139	176,053	148,708	129,610	10,841	1,002,025

3.2.3. 언어권별 말뭉치의 구축 비율

(1) 문어

① 원시 말뭉치

- 문어 원시 말뭉치는 총 3,278,600어절 중 언어권별로 중국어권 학습자 자료가 1,565,873어절(47.8%)로 가장 많은 비중을 차지하며, 이어서 일본어권 432,954어절(13.2%), 베트남어권 223,758어절(6.8%), 영어권 217,151어절(6.6%)을 차지하였다.

<표 74> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 원시 문어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	167,312	225,979	272,159	282,542	305,522	204,926	107,433	1,565,873

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
일본어	38,141	68,767	88,574	91,033	83,011	62,892	536	432,954
베트남어	36,905	43,769	49,158	39,083	34,259	14,588	5,996	223,758
영어	29,853	42,928	42,610	38,199	29,731	30,847	2,983	217,151
광둥어	11,385	24,974	31,460	33,497	27,360	29,499	0	158,175
러시아어	10,873	16,042	22,059	18,197	21,377	9,121	5,145	102,814
태국어	11,235	16,957	13,206	12,111	8,555	6,637	321	69,022
몽골어	9,909	12,290	15,045	12,339	11,679	5,483	511	67,256
인도네시아어	6,591	7,813	7,135	10,204	6,857	4,939	1,097	44,636
프랑스어	9,993	8,143	10,166	4,682	4,984	3,533	933	42,434
기타	52,981	68,500	77,651	58,824	43,451	34,214	18,906	354,527
합계	385,178	536,162	629,223	600,711	576,786	406,679	143,861	3,278,600

② 형태 주석 말뭉치

- 문어 형태 주석 말뭉치는 총 2,502,258어절 중 언어권별로 중국어권 학습자 가료가 939,452어절(37.5%)로 가장 많은 비중을 차지하며, 이어서 일본어권 423,278어절(16.9%), 베트남어권 228,882어절(8.9%), 영어권 212,280어절(8.5%)을 차지하였다.

<표 75> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 형태 문어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	141,952	140,974	141,338	139,062	161,978	173,775	40,372	939,451
일본어	38,141	68,512	88,406	90,862	76,417	60,860	80	423,278
베트남어	36,905	43,568	49,004	38,705	33,116	14,588	5,996	221,882
영어	29,661	42,509	42,392	37,379	27,878	30,700	1,761	212,280
러시아어	10,873	15,396	21,488	16,747	20,514	9,121	4,602	98,741
광둥어	9,955	5,940	6,096	14,239	19,596	25,975	0	81,801
태국어	11,235	16,957	13,111	11,727	7,928	6,637	321	67,916
몽골어	9,866	11,984	14,817	12,119	10,946	5,483	511	65,726
인도네시아어	6,387	7,813	6,971	9,175	6,655	4,939	1,097	43,037
스페인어	6,874	9,976	9,970	5,743	5,016	1,825	293	39,697

기타	55176	64667	53038	44708	39796	34261	16,803	308449
합계	357,025	428,296	446,631	420,466	409,840	368,164	71,836	2,502,258

③ 오류 주석 말뭉치

- 문어 형태 주석 말뭉치는 총 501,193어절 중 언어권별로 일본어권 학습자 자료가 112,826어절(22.5%), 영어권 112,576어절(22.5%), 중국어권 110,372어절(22.0%)로 거의 비슷한 비중으로 전체 자료의 67%를 차지하며, 베트남어권 57,360어절(11.4%), 러시아어권 자료 31,040어절(6.2%)을 차지하였다.

<표 76> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 오류 문어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
일본어	18,387	18,651	18,354	19,143	19,586	18,625	38,211	112,826
영어	20,903	21,046	17,693	17,538	17,799	17,597	35,396	112,576
중국어	18,333	19,821	18,933	17,890	16,473	16,607	33,080	110,372
베트남어	8,450	7,265	13,452	10,829	7,025	10,339	17,364	57,360
러시아어	3,864	5,520	6,366	5,765	6,253	2,660	8,913	31,040
태국어	4,408	6,432	3,268	3,542	3,844	3,034	6,878	24,528
스페인어	2,142	1,752	1,420	1,247	152	317	469	7,030
카자흐어	247	436	1,520	855	552	541	1,093	4,151
아랍어	90	1,436	697	583	102	1,205	1,307	4,113
광둥어	256	826		297	922	919	1,841	3,220
기타	6,221	7,560	6,469	5,544	4,459	2,958	7,417	33,977
합계	83,301	90,745	88,172	83,233	77,167	74,802	151,969	501,193

(2) 구어

① 원시 말뭉치

- 구어 원시 말뭉치는 총 1,121,769어절 중 언어권별로 중국어권 학습자 자료가 269,931어절(24.1%)로 가장 많은 비중을 차지하며, 이어서 베트남어권 188,589어절(16.1%), 일본어권 120,900어절(10.8%), 태국어권 120,728어

절(10.8%)을 차지하였다.

<표 77> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 원시 구어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	38,942	43,188	37,885	40,123	33,333	41,886	34,574	269,931
베트남어	43,854	40,969	39,880	39,536	16,313	5,500	2,537	188,589
일본어	11,241	26,310	23,575	20,999	20,228	18,127	420	120,900
태국어	39,603	19,472	33,473	12,227	4,672	4,376	6,905	120,728
러시아어	11,512	16,881	21,037	15,698	3,767	3,682	4,565	77,142
인도네시아어	12,216	12,031	11,576	7,613	11,224	4,361	1,143	60,164
영어	9,069	14,298	12,725	6,205	6,136	560	864	49,857
타갈로그어	11,948	11,615	11,635	6,632	1,682	1,372	0	44,884
스페인어	8,014	12,615	6,066	6,731	2,222	511	707	36,866
싱할라어	6,204	5,306	6,539	5,740	3,726	2,990	0	30,505
기타	13,993	29,597	22,042	33,170	11,754	3,814	7,833	122,203
합계	206,596	232,282	226,433	194,674	115,057	87,179	59,548	1,121,769

② 형태 주석 말뭉치

- 구어 형태 주석 말뭉치는 총 1,001,471어절 중 언어권별로 중국어권 학습자 가료가 224,904어절(22.5%)로 가장 많은 비중을 차지하며, 이어서 베트남어권 158,862어절(15.9%), 태국어권 115,258어절(11.5%), 일본어권 114,760어절(11.5%)을 차지하였다.

<표 78> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 형태 구어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	37,320	32,481	28,577	31,506	31,280	41,886	21,854	224,904
베트남어	39,692	30,085	29,882	38,737	14,295	5,500	671	158,862
태국어	39,603	18,950	33,473	12,227	4,672	4,376	1,957	115,258
일본어	10,494	22,909	23,121	19,881	20,228	18,127	0	114,760
러시아어	10,654	13,467	19,979	15,368	3,767	3,435	441	67,111

인도네시아어	11,715	9,980	11,576	7,613	11,224	4,361	1,143	57,612
영어	9,069	13,679	10,909	5,093	6,136	560	0	45,446
타갈로그어	11,948	10,987	11,635	6,632	1,682	1,372	0	44,256
스페인어	8,014	11,997	6,066	5,555	2,222	511	707	35,072
싱할라어	6,204	5,306	6,539	5,740	3,726	2,990	0	30,505
기타	12,683	25,920	20,960	28,247	11,754	3,814	4,307	107,685
합계	197,396	195,761	202,717	176,599	110,986	86,932	31,080	1,001,471

③ 오류 주석 말뭉치

- 구어 오류 주석 말뭉치는 총 500,832어절 중 언어권별로 일본어권 학습자
자료가 111,765어절(22.3%), 중국어권 109,862어절(21.9%), 베트남어권
106,517어절(21.3%)로 비슷한 비중으로 전체 자료의 약 67%를 차지하며,
이어서 영어권 자료가 43,081어절(8.6%), 인도네시아어권 자료가 25,597어
절(5.8%)을 차지하였다.

<표 79> 2015-2020년 한국어 학습자 말뭉치 언어권별 통계: 오류 구어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
일본어	10,078	22,909	22,251	19,881	18,519	18,127	36,646	111,765
중국어	19,505	13,635	12,358	13,251	18,664	27,338	46,002	109,862
베트남어	27,477	18,022	15,006	28,158	12,354	5,500	17,854	106,517
영어	9,069	11,314	10,909	5,093	6,136	560	6,696	43,081
인도네시아어	9,302	4,137	5,319	3,949	5,677	629	6,306	29,013
스페인어	6,874	8,309	4,154	4,402	1,347	511	1,858	25,597
태국어	3,378	4,332	5,863	3,590	850	628	1,478	20,598
러시아어	1,458	2,146	7,038	2,741	1,271	686	1,957	15,340
아랍어		2,855	560	3,854	512	829	1,341	8,610
키르기스어		1,830		2,254	549		549	4,633
기타	2,754	7,244	4,509	5,647	5,662	0	5,662	25,816
합계	89,895	96,733	87,967	92,820	71,541	54,808	126,349	500,832

3.2.4. 장르별 말뭉치의 구축 비율

(1) 문어

① 원시 말뭉치

- 문어 원시 말뭉치는 생활문 1,450,081어절(44.2%), 논설문 1,016,836어절(31.0%)로 두 장르가 주를 이루었고, 보고서 238,892어절(7.3%), 설명문 225,318어절(6.9%) 어절을 차지하였다.

<표 80> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 원시 문어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
생활문	343,099	392,677	435,047	203,206	38,838	33,856	3,358	1,450,081
논설문	263	1,971	58,260	286,907	413,552	250,808	5,075	1,016,836
보고서	0	120	6,731	29,236	38,622	31,328	132,855	238,892
설명문	12,031	62,157	57,266	39,720	42,714	9,955	1,475	225,318
수필	5,047	9,855	13,444	18,839	33,394	23,745	630	104,954
기행문	14,014	64,480	12,905	4,603	706	5,517	330	102,555
감상문	0	114	41,043	5,838	2,532	1,789	138	51,454
기사문	0	142	0	12,037	6,169	28,077	0	46,425
전기문	0	0	0	0	259	21,604	0	21,863
편지글	10,724	4,646	4,527	325	0	0	0	20,222
합계	385,178	536,162	629,223	600,711	576,786	406,679	143,861	3,278,600

② 형태 주석 말뭉치

- 문어 형태 주석 말뭉치는 생활문 1,144,638어절(45.7%), 논설문 785,348어절(31.4%)로 두 장르가 주를 이루었고, 이어서 설명문 170,967어절(6.8%), 보고서 102,576어절(4.1%)을 차지하였다.

<표 81> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 형태 문어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
생활문	319,518	306,317	301,852	151,125	31,207	31,780	2,839	1,144,638

논설문	263	1,971	52,984	200,994	294,205	231,768	3,163	785,348
설명문	10,837	57,593	34,490	28,357	29,416	9,388	886	170,967
보고서	0	120	1,388	8,267	14,047	14,766	63,988	102,576
수필	3,373	6,458	9,257	17,380	32,811	23,745	630	93,654
기행문	12,421	51,014	9,795	3,165	524	5,517	330	82,766
감상문	0	114	33,913	5,408	2,532	1,789	0	43,756
기사문	0	142	0	5,445	4,839	27,884	0	38,310
전기문	0	0	0	0	259	21,527	0	21,786
편지글	10,613	4,567	2,952	325	0	0	0	18,457
합계	357,025	428,296	446,631	420,466	409,840	368,164	71,836	2,502,258

③ 오류 주석 말뭉치

- 문어 오류 주석 말뭉치는 생활문 260,296어절(51.9%)로 가장 높은 비중을 차지하였고, 이어서 논설문 144,798어절(28.9%), 설명문 43,677어절(8.7%)을 차지하였다

<표 82> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 오류 문어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
생활문	77,344	70,578	60,271	35,240	8,481	7,041	1,341	260,296
논설문	53	652	10,686	30,015	54,055	47,235	2,102	144,798
설명문	3,012	9,409	8,186	10,617	9,090	3,363	0	43,677
수필	101	166	334	3,242	3,213	6,942	0	13,998
기행문	442	8,232	2,383	1,099	307	169	330	12,962
기사문				679	1,350	9,344	0	11,373
감상문		114	5,251	2,341	671		0	8,377
편지글	2,349	1,594	1,061				0	5,004
전기문						708	0	708
합계	83,301	90,745	88,172	83,233	77,167	74,802	3,773	501,193

(2) 구어

① 원시 말뭉치

- 구어 원시 말뭉치는 인터뷰가 836,148어절(74.5%)로 가장 높은 비중을 차지하였고, 이어서 발표 212,152어절(18.9%), 내러티브 46,449어절(4.1%), 자유 대화 27,020어절(2.4%)을 차지하였다.

<표 83> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 원시 구어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
인터뷰	179,401	189,122	168,465	134,926	62,673	63,516	38,045	836,148
발표	17,551	24,843	48,945	49,628	44,620	18,320	8,245	212,152
내러티브	7,666	10,212	8,521	8,235	7,764	4,051	0	46,449
자유 대화	1,978	8,105	502	1,885		1,292	13,258	27,020
합계	206,596	232,282	226,433	194,674	115,057	87,179	59,548	1,121,769

② 형태 주석 말뭉치

- 구어 형태 주석 말뭉치는 인터뷰가 754,334어절(75.3%)로 가장 높은 비중을 차지하였고, 이어서 발표 175,531어절(17.5%), 내러티브 46,449어절(4.6%), 자유 대화 25,157어절(2.5%)을 차지하였다.

<표 84> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 형태 구어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
인터뷰	173,566	164,169	151,756	131,356	58,602	63,516	11,369	754,334
발표	15,313	14,011	41,938	35,123	44,620	18,073	6,453	175,531
내러티브	7,666	10,212	8,521	8,235	7,764	4,051	0	46,449
자유 대화	851	7,369	502	1,885		1,292	13,258	25,157
합계	197,396	195,761	202,717	176,599	110,986	86,932	31,080	1,001,471

③ 오류 주석 말뭉치

- 구어 오류 주석 말뭉치는 인터뷰가 367,878어절(73.5%)로 가장 높은 비중을 차지하였고, 이어서 발표 110,305어절(22.0%), 자유 대화 13,719어절(2.7%), 내러티브 8,930어절(1.8%)을 차지하였다.

<표 85> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 오류 구어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
인터뷰	78,985	82,722	63,930	62,668	33,493	44,156	1,924	367,878
발표	10,139	6,930	22,164	25,299	35,174	8,278	2,321	110,305
자유 대화	325	6,892	502	1,885		1,292	2,823	13,719
내러티브	446	189	1,371	2,968	2,874	1,082	0	8,930
합계	197,396	195,761	202,717	176,599	110,986	86,932	31,080	1,001,471

3.2.5. 주제별 말뭉치의 구축 비율

- 본 연구에서는 2015-2020년 국어원 한국어 학습자 말뭉치의 주제를 ‘국제 통용 한국어 교육과정’에서 제시한 17개의 주제 범주에 따라 재범주화하여 분석해 보고자 한다. 이는 2015-2020년 기구축 말뭉치의 주제가 표본별로 세부 주제를 담고 있어 주제별 말뭉치의 비중을 조금 더 명확하고 용이하게 파악하기 위한 것이다. 다음 자료는 본 제안을 위해 1차 분류한 결과이며, 본격적인 연구에서 검토 과정을 통해 보다 정밀하게 분석이 이루어질 것이다.

(1) 원시 말뭉치

- 원시 말뭉치는 문어의 경우 ‘사회’가 855,054어절(26.1%)로 가장 높은 비중을 차지하였고, 이어서 ‘일상생활’이 680,816어절(20.8%), ‘개인 신상’이 488,681어절(14.9%)을 차지하였다. 구어의 경우는 ‘개인 신상’이 380,087어절(33.9%), ‘일상생활’이 343,795어절(30.6%)로 64%에 이르렀다.

<표 86> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 원시 말뭉치

주제 범주	문어	구어	합계
개인 신상	488,681	380,087	868,768
건강	70,990	19,023	90,013
공공 서비스	402	1,696	2,098
교육	144,295	52,805	197,100
교통	8,169	1,970	10,139
기후	38,376	25,705	64,081
대인 관계	257,340	50,005	307,345
사회	855,054	62,404	917,458
쇼핑	40,380	5,061	45,441
식음료	17,688	19,344	37,032
여가와 오락	325,187	92,755	417,942
여행	158,193	23,239	181,432
예술	1,829	5,000	6,829
일과 직업	91,559	17,234	108,793
일상생활	680,816	343,795	1,024,611
전문 분야	52,483	4,668	57,151
주거 환경	47,158	16,978	64,136
합계	3,278,600	1,121,769	4,400,369

(2) 형태 주석 말뭉치

- 형태 주석 말뭉치는 문어의 경우 ‘사회’가 626,439어절(25.0%)로 가장 높은 비중을 차지하였고, 이어서 ‘일상생활’이 546,152어절(21.8%), ‘개인 신상’이 367,194어절(14.7%)을 차지하였다. 구어의 경우는 ‘개인 신상’이 319,815어절(31.9%), ‘일상생활’이 318,870어절(31.8%)로 63%에 이르렀다.

<표 87> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 형태 주석 말뭉치

주제 범주	문어	구어	합계
개인 신상	367,194	319,815	687,009
건강	47,514	14,641	62,155
공공 서비스	245	1,696	1,941
교육	90,539	47,582	138,121
교통	6,484	1,351	7,835
기후	31,437	24,817	56,254
대인 관계	204,492	48,566	253,058
사회	626,439	55,625	682,064
쇼핑	21,850	3,222	25,072
식음료	13,408	19,344	32,752
여가와 오락	266,886	84,209	351,095
여행	127,895	18,361	146,256
예술	1,309	4,679	5,988
일과 직업	72,722	17,234	89,956
일상생활	546,152	318,870	865,022
전문 분야	35,248	4,668	39,916
주거 환경	42,444	16,791	59,235
합계	2,502,258	1,001,471	3,503,729

(3) 오류 주석 말뭉치

- 오류 주석 말뭉치는 문어의 경우 ‘일상생활’이 128,786어절(25.7%), 이어서 ‘사회’가 112,796어절(22.5%)로 약 46%에 이르렀고, 이어서 ‘개인 신상’이 75,628어절(15.1%), ‘여가와 오락’이 57,250어절(11.4%)를 차지하였다. 구어의 경우는 ‘개인 신상’이 116,509어절(33.2%), ‘일상생활’이 145,196어절(29.0%)로 약 63%에 이르렀다.

<표 88> 2015-2020년 한국어 학습자 말뭉치 장르별 통계: 오류 주석 말뭉치

주제 범주	문어	구어	합계
개인 신상	166,509	75,628	242,137
건강	7,157	7,345	14,502
교육	32,086	14,612	46,698
교통	745	412	1,157
기후	6,730	4,925	11,655
대인 관계	22,282	40,043	62,325
사회	29,675	112,796	142,471
쇼핑	3,068	5,889	8,957
식음료	12,353	4,680	17,033
여가와 오락	39,017	57,250	96,267
여행	11,187	25,386	36,573
예술	4,679	58	4,737
일과 직업	11,003	11,632	22,635
일상생활	145,196	128,786	273,982
전문 분야	1,622	1,705	3,327
주거 환경	7,523	10,046	17,569
합계	500,832	501,193	1,002,025

3.3. 종합 및 적용

- 2015-2020년에 구축한 한국어 학습자 말뭉치는 원시 말뭉치 440만 어절, 형태 주석 말뭉치 350만 어절, 오류 주석 말뭉치 100만 어절을 구축하였다. 말뭉치 유형별 비중에서 형태 주석 말뭉치가 원시 말뭉치의 약 79.5%, 오류 주석 말뭉치가 22.7%로 오류 주석 말뭉치의 비중이 매우 적음을 확인할 수 있다. 의견수렴 결과에서도 확인된 바와 같이 자료 활용의 폭을 넓히기 위해 형태 주석 및 오류 주석 말뭉치의 확대가 요구된다.
- 이와 함께 자료의 변인별 균형성 확보를 위한 노력도 지속되어야 한다. 원시 말뭉치를 기준으로 살펴보면 총 4,400,369어절 중 수집 대상별 말뭉치는 국내 학습자의 자료가 85.5%, 이주민 자료가 9.0%, 국외 학습자 자

료의 경우 5.5%로 이주민 자료와 국외 학습자 자료의 비중이 현저히 낮음을 확인할 수 있다.

- 수준별로는 1급이 13.4%, 2급이 17.5%, 3급이 19.4%, 4급이 18.1%, 5급이 15.7%, 6급이 11.2%, 6급 이상이 4.6%로 3급이 가장 많고, 1급과 6급 이상이 다른 등급에 비해 현저하게 적다. 이는 1급의 경우 산출 발화의 길이가 다른 등급에 비해 적고, 고급 단계로 갈수록 교육과정에서 이탈하는 학습자의 수가 증가하면서 수집 자료의 양도 다른 등급에 비해 적어서 비롯된 결과이다.
- 자료 유형별로는 구어의 비중이 25.5%로 문어에 비해 적으며, 장르별로는 문어의 경우 생활문이 44.2%, 논설문이 31.0%로 총 9가지 장르 중 두 개의 장르에 편중되어 있으며, 구어의 경우 인터뷰, 발표, 내러티브, 자유 대화의 4가지 장르 중 교사와의 인터뷰가 74.5%로 주로 한 가지 장르에 편중되어 있다.
- 주제별로는 문어의 경우 ‘사회’(26.1%), ‘일상생활’(20.8%), ‘개인 신상’(14.9%) 순으로 <국제통용 한국어 교육 표준 모형>에서 제시하고 있는 17개의 주제 중 세 가지 주제에 편중되어 있으며, 구어의 경우 ‘개인 신상’(33.9%), ‘일상생활’(30.6%), ‘여가와 오락’(8.3%) 순으로 역시 두 가지 주제에 편중되어 있다.
- 그 외에도 종적 자료와 일부 기획 자료를 제외한 대부분의 자료가 수업 중 과제, 성취도 평가의 산출물에 편중된 경향이 있었다. 이는 1차 중장기 계획에서 선 구축 후 균형의 방식으로 균형성을 확보하는 과정에서 현실적인 학습자 분포의 특성이 그대로 반영된 결과이다. 중장기 계획에서는 이러한 분포 특성을 참고하여 상대적으로 부족한 자료를 보충하기 위한 기획 수집을 확대하는 전략이 필요하다.

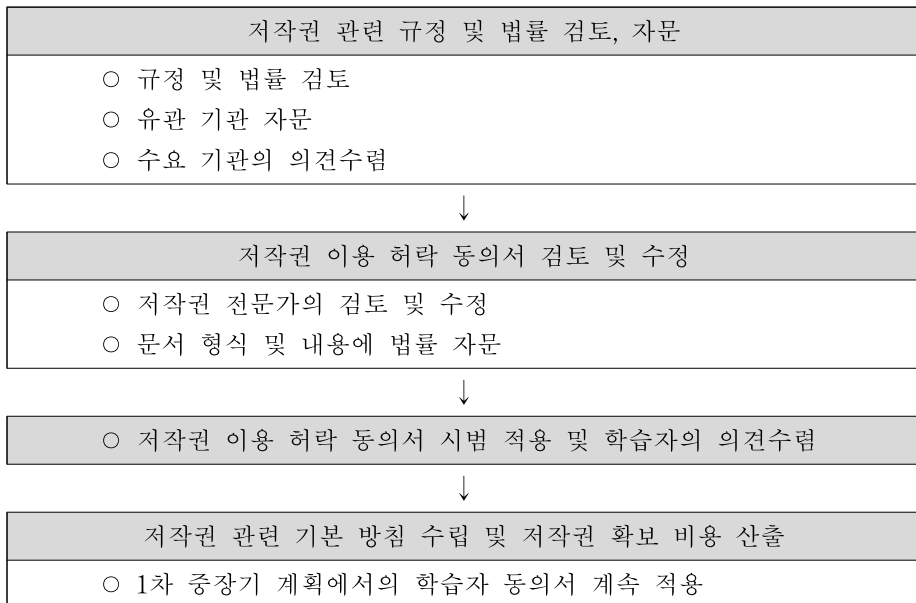
4. 말뭉치 구축·활용에 관한 저작권 확보를 위한 세부 실행 방법 마련

- 말뭉치는 디지털 언어 자원으로 교육과 연구에서의 활용도 제고를 위해 자료 제공 및 이용 허락에 관한 학습자의 동의를 구하는 것이 구축의 첫 단계이다. 이에 따라 1차 중장기 계획에서는 학습자 동의서를 통해 그러

한 문제를 해결하였다. 본 연구에서는 2차 중장기 계획 수립에 앞서 국가 언어 자원으로서 학습자 말뭉치 활용의 범위를 학계뿐만 아니라 민간 분야의 활용 가능성을 염두에 두고 학습자 말뭉치의 저작권 확보의 타당성과 실현 가능성에 관한 법률적 검토 및 전문가 자문을 실시하였다. 그리고 그 결과를 반영하여 저작권 이용 허락 동의, 개인 정보 수집 및 이용에 관한 세부 실행 계획을 마련하였다.

4.1. 저작권 이용 허락 동의

○ 저작권에 관한 논의는 다음의 절차에 따라 진행되었다.



<그림 29> 저작권 이용 허락 동의 관련 방침 수립 절차

(1) 저작권 관련 규정 및 법률 검토

○ 저작권 관련 규정 및 법률 검토(☞ 1.1 참고) 및 세종학당재단의 한국어 학습용 앱 ‘AI 한국어 선생님’을 통해 수집되는 자료 처리 방침, 수요 기관인 국립국어원과의 협의의 절차를 통해 이루어졌다. 그 결과, 저작권법의 해석과 판례에 따라 학습자가 산출한 자료가 저작물로 처리되는 것이

원칙이 되어야 하나 상당 부분 교육과정에서 산출되는 통제적이고 모방적인 산출 자료이거나 구축 팀의 기획 과제에 의해 산출되는 자료로 경제적인 이익을 취할 수 있는 자료가 아니라는 점에서 저작물보다는 학습자 말뭉치 구축을 위한 데이터로서 간주하기로 하였다. 자문 과정에서, 수집 목적에는 다소 차이가 있으나 앱을 활용해 한국어를 연습하는 과정에서 기록되는 학습자 자료를 수집하는 세종학당재단의 경우도 법률 자문을 통해 이와 같은 방식으로 처리 지침을 따르고 있음을 확인하였다.

(2) 저작권 이용 허락 동의서 검토 및 수정

- 저작권 이용 허락 동의서는 국립국어원의 ‘국가 언어 자원(학습자 말뭉치) 이용 약정서’에 대한 법률적 검토를 통해 저작권 확보를 위한 조항을 추가하여 수정·보완하는 방식으로 작성되었다. 검토 및 수정 과정에서의 쟁점은 학습자 말뭉치 활용의 측면에서 복제권, 전송권, 배포권, 2차적 저작물 작성권과 같은 동의의 대상, 이용 허락 기간, 권리자와 이용자의 권리와 의무에 관한 세부 사항을 조정하는 것이었다. 이러한 쟁점을 반영한 동의서는 서문과 함께 다음의 15개 조항으로 이루어져 있다(☞ 동의서는 부록 참고).

- 서문
- 제1조 (동의의 목적)
- 제2조 (정의)
- 제3조 (동의의 대상)
- 제4조 (이용허락 기간)
- 제5조 (권리자의 의무)
- 제6조 (이용자의 권리 및 의무)
- 제7조 (확인 및 보증)
- 제8조 (동의내용의 변경)
- 제9조 (동의의 해지)
- 제10조 (손해배상)
- 제11조 (비용의 부담)
- 제12조 (분쟁해결)
- 제13조 (비밀유지)

- 제13조 (기타부속합의)
- 제14조 (동의의 해석 및 보완)
- 제15조 (동의 효력 발생일)

(3) 저작권 이용 허락 동의서 시범 적용 및 학습자의 의견수렴

- 저작권 이용 허락 동의서는 학습자의 심리적 부담을 경감하기 위하여 동의서라는 명칭으로 완화된 표현을 쓰기는 하였으나 저작권법에 기초하여 관련 사항을 명시하였기 때문에 저작권법에 대한 이해가 없는 일반인에게 어려운 내용이며 한국어를 제2 언어 또는 외국어로서 학습하는 학습자들에게는 더욱 그러하다. 이에 본 연구에서는 저작권 이용 허락 동의서를 영어, 일본어, 중국어 세 개 언어로 번역을 한 후 내용에 대한 이해를 돕기 위한 동영상 제작하여 세 개 언어를 제1 언어로 하는 30여 명의 학습자를 대상으로 시범 적용을 하였다. 그리고 자료 수집 참여 및 동의 절차에 대한 의견을 수렴하였다. 참여 학습자의 국적은 미국이 10명(33.3%), 일본이 14명(46.7%), 중국이 6명(20.0)이었으며, 한국어 수준은 초급이 7명(23.3%), 중급이 23명(76.7%)이었다.
- 의견수렴을 위한 설문 문항은 다음과 같이 6개의 문항으로 구분하였으며, 첫 번째 문항에는 연관 질문이 2개 추가되었다(☞ 설문지는 부록 참고).

<표 89> 저작권 이용 허락 동의서 시범 적용 후 의견수렴 문항 구성

번호	질문
1	서명 전 저작권 이용 허락 동의서의 내용 이해 여부 1-1. 동영상 설명 자료의 유용성 여부 1-2. 서명 시 느낀 감정
2	서명 전 개인 정보 수집, 이용 및 제3자 제공 동의서의 내용 이해 여부
3	자료 수집 참여 시 사용 가능한 시간
4	적절한 보상 방법
5	향후 자료 수집 참여 의사
6	자료 수집 참여 추천 의사

① 서명 전 저작권 이용 허락 동의서의 내용 이해 여부

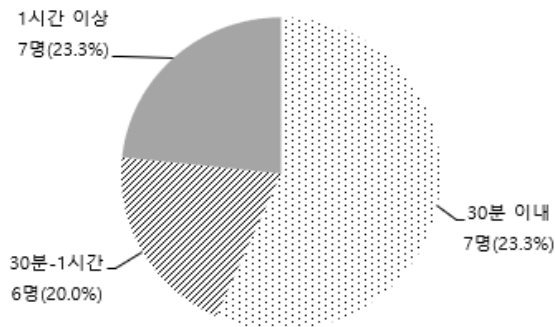
- 서명 전에 저작권 이용 허락 동의서의 내용을 충분히 이해하였는지를 묻는 문항에는 30명 전원이 그렇다고 응답하였다.
- 30명 모두 동영상 설명이 도움이 되었으며, 동의서에 서명을 할 때 ‘특별한 감정을 느끼지 않았다’고 응답한 학습자가 20명(66.7%), ‘부담스러웠다’ 응답한 학습자가 10명(33.3%)였다. 부담스러웠다고 응답한 학습자 중 1명은 ‘기타’에 여권번호를 제출하는 것에 대한 거부감을 언급하였다.

② 서명 전 개인 정보 수집, 이용 및 제3자 제공 동의서의 내용 이해 여부

- 서명 전에 개인 정보 수집, 이용 및 제3자 제공 동의서의 내용을 충분히 이해하였는지를 묻는 문항에는 30명 전원이 그렇다고 응답하였다.

③ 자료 수집 참여 시 사용 가능한 시간

- 자료 수집에 사용할 수 있는 시간은 ‘30분 이내’가 17명(56.7%), ‘30분-1시간’이 6명(20.0%), ‘1시간 이상’이 7명(23.3%)으로 응답하여 학습자에게 1시간 미만의 시간이 과제 수행에 적절한 시간임을 알 수 있었다.

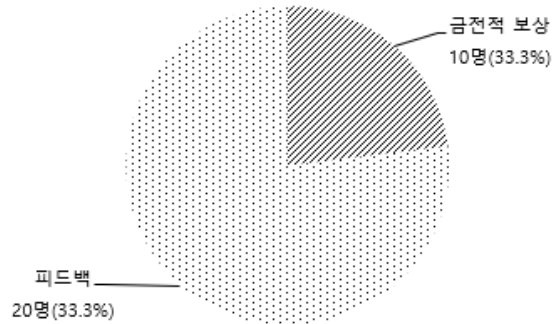


<그림 30> 설문 결과: 자료 수집 참여 시 사용 가능한 시간

④ 적절한 보상 방법

- 적절한 보상 방법으로 ‘작문/말하기 자료에 대한 피드백’이라고 응답한 학습자가 20명(66.7%), ‘금전적 보상’이라고 응답한 학습자가 10명(33.3%)으로, 피드백에 대한 요구가 큼을 확인할 수 있었다. ‘금전적 보상’의 경우, 대부분의 학습자가 2-3만 원이라고 응답하였으며, 적게는 1만 원, 많게는 5만 원이 적정 금액이라고 응답한 학습자도 있었다. 또한 ‘금전적 보상’

대신 ‘한국어 교재’를 제공해 주면 좋겠다고 응답한 학습자도 있었다.



<그림 31> 설문 결과: 자료 수집에 대한 보상 방법

⑤ 향후 자료 수집 참여 의사

- 향후 자료 수집 참여 의사를 묻는 문항에 대해서는 25명(83.3%)이 ‘네’, 5명(16.7%)이 ‘아니오’라고 응답하였다. ‘네’라고 응답한 학습자의 경우는 주된 이유로 ‘(다소 부담스럽지만) 자신의 한국어를 향상시키는 데 도움이 되어서’, ‘미래의 한국어 학습자에게 도움이 되어서’를 꼽았고, ‘아니오’라고 응답한 학습자는 ‘시간이 걸려서 힘들기 때문에’, ‘귀찮기 때문에’라고 설명하였다. 이러한 응답을 4번의 응답과 연관 지어 보면 금전적 보상보다는 학습 효과를 느끼도록 할 수 있는 피드백이 학습자에게 더 매력이 있는 보상이 될 수 있을 것이다.

⑥ 자료 수집 참여 추천 의사

- 친구에게 자료 수집에 참여하도록 추천할 의사가 있는지 묻는 문항에 대해서는 18명(60.0%)이 ‘네’, 12명(40.0%)이 ‘아니오’라고 응답하였다. ‘네’라고 응답한 학습자는 그 이유를 ‘공부가 되기 때문에’, ‘언어 기술을 향상시키는 데 도움이 되기 때문에’, ‘한국어 학습자에게 도움을 주기 위해’, ‘자료 수집이 의미가 있는 일이기 때문에’, ‘보람이 있기 때문에’, ‘어려지 않은 일이고 배운 것을 적용해 볼 수 있는 기회가 되기 때문에’, ‘보수를 받을 수 있기 때문에’, ‘많은 사람이 참여해야 더 좋은 자료가 되며, 이 자료가 언어 습득 연구의 발달과 한국어 학습 자료 구축에 이용되기 때문에’라고 답하였다. 반면, ‘아니오’라고 응답한 학습자는 그 이유로 ‘개인의 선택이기 때문에’, ‘친구가 한국어 모어 화자이거나 다른 언어 사용자라서

도움이 되지 않기 때문에’, ‘시간이 걸리기 때문에’, ‘친구가 부담스럽게 느낄 수 있기 때문에’를 들었다. 이로부터 수집 단계에서 자료 수집의 목적과 의미를 학습자에게 충분히 설명하는 것이 학습자의 참여를 독려하는 데에 도움이 됨을 알 수 있다.

(4) 저작권 관련 기본 방침 수립

- 본 연구에서는 앞선 단계의 연구를 통해 자료 수집의 효율성 제고, 더 나아가 온라인 수집 등 수집 방식의 변화가 요구되는 시점에서 내용 이해를 위해 문서 외의 보조적 수단이 필요한 저작권 이용 허락 동의서를 사용한 수집 방식이 적절하지 않다는 결론에 이르렀다. 이에 저작권 이용 허락 동의서를 사용한 수집을 유보하고 1차 중장기 계획에서와 같이 학습자 동의서를 이용해 자료를 수집하기로 하였다.
- 학습자 동의서는 법률 계약서의 형식을 갖추어 저작권에 관한 사항을 직접적으로 언급하고 있지는 않았으나 학습자의 권리를 보호하기 위한 자료 수집 목적, 수집 대상, 이용 범위, 개인 정보 보호, 철회 의사 표시 등에 관한 핵심 조항을 모두 포함하고 있다. 따라서 사용자가 자료를 이용하는 과정에서 윤리적 문제에 주의한다면 큰 분쟁의 소지가 거의 없을 것으로 보이며, 이는 국외 학습자 말뭉치에서도 보편적으로 적용하고 있는 방식이기도 하다.
- 이에 따라 본 연구에서는 자료 이용의 범위, 자료 이용 방법, 적절하지 못한 사용으로 인한 분쟁의 책임 소재와 이용 제한 등에 관한 조항을 추가하여 ‘국가 언어 자원(학습자 말뭉치) 이용 약정서’를 수정·보완하였다(☞ 약정서는 부록 참고).

(5) 저작권 확보를 위한 비용 산출

- 1차 중장기 계획에서는, 2019-2020년 기획 자료 수집 시 자료 제공에 대한 사례로 학습자에게 문어 자료 5,000원, 구어 자료 10,000원에 상당하는 상품권을 지급한 바 있으며, 2차 중장기 계획 수립을 위한 2021년 기획 자료 수집에서는 문어 자료 5,000원, 구어 자료 15,000원에 상당하는 상품권을 지급하였다. 이는 수집 교사들의 의견수렴을 통해 자료 수집 참여를 독려하는 데에 일정 정도 영향력을 미칠 수 있는 적정 금액으로 합의된

금액이다.

- 본 연구에서는 학습자의 산출 자료가 경제적 이득을 취할 수 있는 자료가 아니며, 무엇보다도 수집 과정에서 자료 수집, 구축 목적과 의의에 공감대가 형성된 상태에서 자료를 제공하는 경우가 많다는 점, 현실적인 예산 규모 등을 고려하여, 구축 팀에서 제안한 과제를 수행하여 자료를 제공하는 경우에 한해 2020년의 지급 기준에 따라 문어 자료 5,000원, 구어 자료 10,000원 또는 그에 상당하는 유가증권을 지급할 것을 제안하는 바이다. 그 밖의 교육과정에서 산출되는 성취도 평가의 경우는 기관 간의 협약을 통해 제공받되 소정의 기념품을 지급할 수 있다. 또한 향후 온라인 수집을 하게 될 경우 자율적으로 참여하는 학습자를 위한 보상 방안으로서 산출 자료에 대한 평가나 피드백 등을 고려해 볼 수 있다. 이를 위해서는 기구축 자료를 활용한 자동 평가와 한국어 교원의 풀을 활용한 질적 평가 방안, 피드백 교사 고용을 위한 예산 확보가 되어야 할 것이다.

4.2. 개인정보 수집 및 이용, 관리

- 학습자 말뭉치 수집에서 수집, 이용되는 개인 정보는 학습자 당사자의 개인 정보와 학습자가 산출한 자료에 등장하는 제3자의 정보가 포함될 수 있다. 1차 중장기 계획에서는 학습자의 개인 정보 보호를 위해 동의서와 함께 수집되는 메타 정보의 경우 개인을 특정할 수 없는 국적, 언어권, 한국어 수준에 한정하여 공개하고, 자료에 포함된 다양한 유형의 개인 정보의 경우 비식별화 처리를 하였다. 본 연구에서는 2차 중장기 계획의 수립에 앞서 개인 정보 보호 관련 규정 및 법률 검토와 전문가 자문을 통해 개인 정보 보호에 관한 그 밖의 고려 사항을 점검하였다. 이에 따라 1차 중장기 계획에서의 방식을 유지하되, ‘개인 정보 수집·이용 및 제3자 제공 동의서’를 새롭게 작성하였다.
- ‘개인 정보 수집·이용 및 제3자 제공 동의서’는 개인 정보 수집·이용에 대한 동의, 고유식별정보 처리에 대한 동의, 개인 정보의 제3자 제공에 대한 동의의 세 개 항목으로 구성되며 세부 내용은 다음과 같다(☞ ‘개인 정보 수집·이용 및 제3자 제공 동의서’는 부록 참고).

(1) 개인정보 수집·이용에 대한 동의

<표 90> 개인정보 수집·이용에 대한 동의 항목

항목	내용
수집·이용 목적	<ul style="list-style-type: none"> ○ 한국어 교육의 질적 향상을 위해 학습자들의 언어 자료를 수집하여 교육 및 연구, 민간에서 활용 가능한 말뭉치로 구축 ○ 신원 확인 및 민원 사항 처리 ○ 참여자의 수당 및 저작권료 정산 ○ 제3자(특정 필요) 이용 목적 <ul style="list-style-type: none"> - 감사 및 실사, 정밀 정산 등 연구 종료 후의 관리 자료
수집·이용할 항목	<ul style="list-style-type: none"> ○ 성명, 국적, 출생연도, 제1 언어, 한국어 학습 기간, 한국에서의 거주 기간, 연락처, 직업, 사용하는 외국어, 학력, 산출한 음성 및 텍스트 자료 ○ 은행 계좌 정보, 주민등록번호, 외국인등록번호, 여권번호
보유·이용 기간	○ 동의한 시점으로부터 5년(또는 동의한 시점부터 사업 완료일 이후 5년)

(2) 고유식별정보 처리에 대한 동의

<표 91> 고유식별정보 처리에 대한 동의 항목

항목	내용
수집하는 고유식별정보 항목	○ 주민등록번호, 외국인 등록번호, 여권번호
고유식별정보의 수집 및 이용 목적	<ul style="list-style-type: none"> ○ 국립국어원이 발주한 용역 사업 수행자(연세대학교 산학협력단)의 연구과제 수행 및 본 연구의 저작권료 정산, 세금 처리 ○ 제3자(특정 필요) 이용 목적 <ul style="list-style-type: none"> - 감사 및 실사, 정밀 정산 등 연구 종료 후의 관리 자료
고유식별정보의 보유 및 이용 기간	○ 동의한 시점으로부터 5년(또는 동의한 시점부터 사업 완료일 이후 5년)

(3) 개인 정보의 제3자 제공에 대한 동의

<표 92> 개인 정보의 제3자 제공에 대한 동의

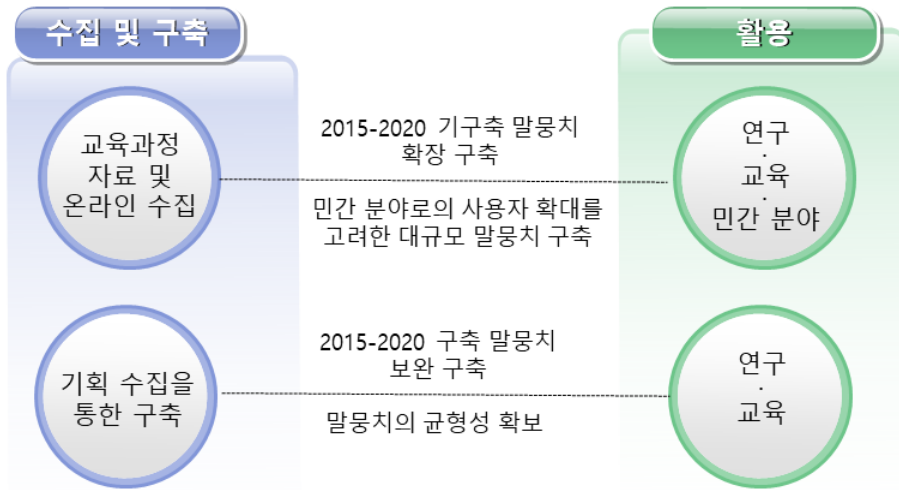
항목	내용
제공 목적	○ 국립국어원(용역 사업 수행자: 연세대학교 산학협력단) 연구 및 이후 저작물 관리와 민원 해결
제공 항목	○ 성명, 주민등록번호, 외국인 등록번호, 여권번호
제공받는 자	○ 국립국어원, 용역 사업 수행자(연세대학교 산학협력단), 제3자 적시
제3자 보유·이용 기간	○ 동의한 시점으로부터 5년(또는 동의한 시점부터 사업 완료일 이후 5년)

5. 한국어 학습자 말뭉치 수집·구축·활용의 중장기 목표 및 단계별 세부 전략 수립

- 본 연구에서는 앞선 단계에서 수행한 기초 연구와 전문가 자문 결과를 종합·분석하여 <2015-2020년 한국어 학습자 말뭉치 연구 및 구축> 사업에서 구축한 말뭉치를 보완하고, 더 나아가 학계뿐만 아니라 민간 분야에서의 활용 가능성을 고려하여 다양한 사용자 집단이 목적에 따라 실용적으로 활용할 수 있는 말뭉치를 구축해 나가기 위한 중장기 계획을 수립하였다.

5.1. 중장기 계획 수립을 위한 기본 방향 및 목표

- 본 연구에서는 기초 연구를 통해 한국어 교육 연구 및 교육, 민간 분야로의 활용 범위 확대 가능성을 전제로 하여 대규모 말뭉치 구축을 위한 양적 확대와 균형성을 확보하기 위한 질적 보완이 필요함을 확인하였다. 이에 따라 중장기 계획 수립을 위한 기본 방향을 다음과 같이 설정하였다.



<그림 32> 2차 중장기 계획에서의 학습자 말뭉치 구축 방향

○ 다음은 1차 중장기 계획과 2차 중장기 계획을 비교하여 제시한 것이다.

<표 93> 1차 중장기 계획과 2차 중장기 계획의 기본 방향 비교

구분		1차 중장기 계획(2015-2020년)	2차 중장기 계획(2021-2025년)
정책 환경		○ 한류 확산과 함께 한국어, 한국 문화를 세계화하기 위한 국가적 관심과 그에 따른 정책 확대	○ 4차 산업혁명 시대의 도래에 따른 대규모 언어 자원에 대한 관심과 관련 정책 확대
사업 목표		○ 한국어 교수·학습을 위한 기초 연구 자료로 국가 주도의 학습자 말뭉치 구축	○ 한국어 교육을 위한 기초 연구, 기술 개발 등을 위한 대규모 말뭉치 구축
사업 기간		○ 2015-2020년(6년)	○ 2021-2025년(5년)
말뭉치 구축의 방향	구축 목표	○ 학습자 변인(대상, 수준)을 고려한 균형 말뭉치 구축	○ 2015-2020년 기구축 말뭉치 보완을 위한 기획 말뭉치 구축 ○ 학계, 교육계 외에도 민간 분야에서의 기술 연구를 위한 대규모 말뭉치 구축
	수집	○ 우편, 이메일을 통한	○ 온라인 수집, 온라인 학습

		연구진의 직접 수집 ○ 학계와 교육기관의 참여 유도를 통한 수집 ○ 수업 과제, 성취도 평가를 중심으로 한 교육과정 산출 자료 수집	매체를 기반으로 한 자동 수집 ○ 학계와 교육기관의 인력 풀을 활용한 개별화된 자료 수집 ○ 유관 기관의 협약을 통한 집중 수집 ○ 기획 과제, 숙달도 평가 자료를 중심으로 한 교육과정 외의 자료 수집
	구축 및 가공	○ 구축 지원 도구 개발을 통한 구축 방법과 절차의 체계화, 선진화 ○ 유형별(원시 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치) 구축을 통한 활용성 제고	○ 구축 지원 도구의 고도화, 인공지능 기술을 활용한 원시 말뭉치 확대 ○ 간단 주석과 정밀 주석으로의 이원화를 통한 오류 주석 말뭉치 확대
활용		○ 한국어 교육 연구 및 교육적 활용	○ 한국어 교육 및 교육적 활용 ○ 민간 분야에서의 활용

5.2. 목표 구축 규모

- 2015-2020년 중장기 계획에서는 원시 말뭉치 440만 어절, 형태 주석 말뭉치 350만 어절, 오류 주석 말뭉치 100만 어절을 구축하였다. 본 연구에서는 2015-2020년의 구축 성과와 학습자 말뭉치의 활용 범위에 대한 검토를 토대로 목표 규모의 타당성을 검증해 보고 이상적인 안으로, 2025년까지의 전체 누적 규모를 기준으로 원시 말뭉치와 형태 주석 말뭉치 각 1,000만 어절, 오류 주석 말뭉치 500만 어절로 구축 목표를 설정하였다. 오류 주석 말뭉치의 경우는 작업과 비용, 활용상의 효용성을 고려하여 전체 500만 어절 중 200만 어절만 2015-2020년의 주석 체계를 적용하고(정밀 주석), 나머지 300만 어절은 최소한의 정보만 주석하는(간단 주석) 방식으로 주석을 이원화하는 안을 제안하였다. 이는 예산 현황이나 현실적인 구

축 가능성을 고려하여 조정될 수 있다.

<표 94> 2021-2025년 학습자 말뭉치 수집 및 구축·가공 목표

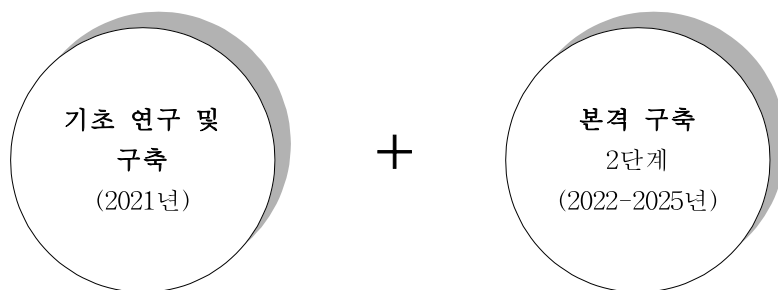
구분	문어			구어			합계		
	2015-20	2021-25	합계	2015-20	2021-25	합계	2015-20	2021-25	합계
자료 수집	328만 +a	372만 +a	600만 +a	112만 +a	288만 +a	400만 +a	440만 +a	560만 +a	1,000만 +a
원시	328만	272만	600만	112만	288만	400만	440만	560만	1,000만
형태 주석	250만	350만	600만	100만	300만	400만	350만	650만	1,000만
오류 주석	50만 (50만)	250만 (50만)	300만 (100만)	50만 (50만)	150만 (50만)	200만 (100만)	100만 (100만)	400만 (200만)	500만 (200만)

※ 오류 주석의 () 수치는 전체 구축 자료 중 정밀 주석이 적용되는 말뭉치의 규모임.

5.3. 단계별 구축 계획

(1) 단계별 구축 방향

- 2015-2020년 중장기 계획에서는 3단계 총 6개년 계획에 따라 1단계에서는 국내 학습자 자료, 2단계에서는 이주민 자료, 3단계에서는 국외 학습자를 집중 구축하는 동시에 말뭉치의 균형성 확보를 이전 단계의 자료 구축을 지속하였다. 본 연구에서는 다음과 같이 기초 연구 1년(2021년), 본격 구축 2단계(2022-2025년)로 총 5개년의 계획을 학습자 말뭉치 구축을 위한 2차 중장기 계획안으로 제안하고자 한다.



<그림 33> 2021-2025년 학습자 말뭉치 구축 방향

- 다음은 단계별 구축의 방향이다. 본 연구에서는 2015-2020년 중장기 계획에 따른 말뭉치 구축 성과에 대한 분석과 선진 사례 분석, 요구분석, 전문가 자문을 토대로 교육기관의 네트워크를 활용한 교육과정 자료 수집과 온라인 수집을 통한 대규모 구축과 기획 수집을 통한 균형성 보완 구축을 두 개의 큰 축으로 제안하고자 한다. 대규모 구축은 사용자 및 활용 목적의 확장을 위한 것으로 2015-2021년에 구축 대상으로 삼은 국내 학습자, 이주민, 국외 학습자의 자료를 광범위하게 수집하여 구축하는 것으로 한다. 한편, 균형성 보완 구축은 2015-2021년의 성과 분석을 통해 균형성 확보의 측면에서 보완이 필요한 특정 변인의 자료를 수집하여 구축하는 것이다. 이러한 변인에는 학습자 대상, 언어권(학습자의 제1 언어), 자료의 매체와 장르가 포함되며 단계별로 집중 구축 대상을 정하여 수집과 구축을 진행해 나간다.
- 아울러 학습자 언어 연구 자료로서의 활용도 제고를 위해 한국어 모어 화자 말뭉치를 참조 말뭉치로 추가 구축하고자 한다. 또한 자료의 다양성 확보를 위한 방안으로 학습자 말뭉치를 이용하는 연구자 또는 타 기관에서 구축한 말뭉치를 제공받아 특수 말뭉치(sub-corpus)로 구축할 것을 제안한다. 이 중 연구자 또는 타 기관 제공 말뭉치는 구축 규모를 한정하지 않으며, 자료의 호환성과 참여 독려를 위해 Stand-alone 방식의 말뭉치 구축 도구를 제공하고 사용 방법을 교육하는 방안을 고려해 볼 수 있다.

<표 95> 단계별 구축의 방향

구분	기초 연구	구축 1단계		구축 2단계	
	2021년	2022년	2023년	2024년	2025년
대규모 구축	온라인 수집 자료 (국내 학습자 + 이주민 + 국외 학습자)				
균형성 확보 구축	대상의 초점화 (학문목적, 이주민, 국외 학습자)				
		언어권별 균형성 확보 (중국어, 일본어, 영어, 베트남어, 태국어, 러시아어, 스페인어권)			
				장르·주제별 초점화	
참조 말뭉치 구축		활용도 제고를 위한 한국어 모어 화자 말뭉치 구축			
연구자 제공 말뭉치		자료의 다양성 확보 및 대규모 말뭉치 구축을 위한 연구자 및 타 기관 제공 말뭉치 통합 구축			

(2) 연차별 구축 규모

○ 다음은 단계별 구축 계획에 따른 연차별 구축 규모이다.

① 학습자 말뭉치

○ 연차별 구축 목표에는 연구자 및 타 기관 제공 말뭉치의 분량은 고려하지 않았다. 이는 전체 말뭉치에 포함되는 특수 말뭉치(sub-corpus)로서 자율적인 의사에 따라 많은 연구자와 기관이 참여할 수 있도록 독려하여 지속적으로 자료를 확보해 나가는 것이 합당하기 때문이다.

<표 96> 2021-2025년 단계별 구축 규모

구분		기초 연구	구축 1단계		구축 2단계		합계
		2021년	2022년	2023년	2024년	2025년	
자료	문어	40만+a	60만+a	60만+a	60만+a	52만+a	272만+a
수집	구어	40만+a	60만+a	60만+a	60만+a	68만+a	288만+a
원시	문어	40만	60만	60만	60만	52만	272만
구축	구어	40만	60만	60만	60만	68만	288만
형태	문어	10만	80만	90만	90만	80만	350만
주석	구어	10만	70만	70만	80만	70만	300만
오류	문어	10만	50만 (10만)	70만 (15만)	70만 (15만)	50만 (10만)	250만
주석	구어	5만	25만 (10만)	40만 (15만)	40만 (15만)	40만 (10만)	150만

※ 오류 주석의 () 수치는 전체 구축 자료 중 정밀 주석이 적용되는 말뭉치의 규모임.

② 참조 말뭉치

- 참조 말뭉치는 장기 계획에 따라 향후 단계적으로 규모를 확대해 나가되, 2021-2025년 중장기 계획에서는 문어와 구어 각 100만 어절씩 200만 어절을 목표 규모로 제안한다.

<표 97> 2021-2025년 단계별 구축 규모

구분		기초 연구	구축 1단계		구축 2단계		합계
		2021년	2022년	2023년	2024년	2025년	
자료	문어	10만	20만+a	25만+a	25만+a	20만+a	100만+a
수집	구어	10만	20만	25만	25만	20만	100만
원시	문어	10만	20만	25만	25만	20만	100만
구축	구어	10만	20만	25만	25만	20만	100만
형태	문어	10만	20만	25만	25만	20만	100만
주석	구어	10만	20만	25만	25만	20만	100만

5.4. 말뭉치 수집 전략

(1) 기본 설계

가. 국외 자료와 이주민 자료의 비중 확대를 통한 대표성 확보

- 2015-2020년 학습자 말뭉치는 국내 학습자 자료의 비중이 매우 높다. 2021-2025년 중장기 계획에서는 이러한 편중성을 해소하기 위하여 국외 자료와 이주민 자료의 비중을 확대하여 한국어 학습자 말뭉치로서의 대표성을 갖추도록 하는 것을 목표로 한다.

나. 수준별, 언어권별, 자료 변인별 자료의 균형성 확보

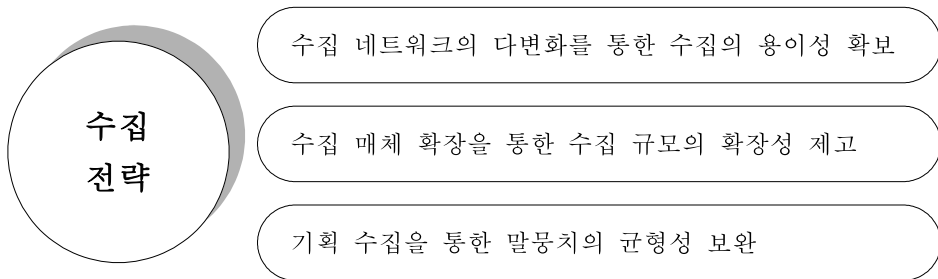
- 2015-2020년 자료는 수준, 언어권, 자료 변인에 따른 분포가 다소 불균형하다. 2021-2025년 중장기 계획에서는 자료의 균형성 확보를 위해 기구축 말뭉치 분석 결과를 토대로 하여 다음의 방향에 따라 자료를 구축하여 말뭉치를 구축한다.
 - 수준별: 고급(5, 6급), 6급 이상의 자료 집중 확대
 - 언어권별: 중국어권, 영어권, 일본어권, 베트남어권, 태국어권, 러시아어권, 스페인어권의 자료를 중점 구축하되, 2015-2020년 자료 중 비중이 가장 높은 중국어권을 제외한 6개 언어권의 자료를 집중 확대
 - 매체별: 구어 자료의 비중 확대
 - 장르별: 구어는 자유 대화, 내러티브, 문어는 설명문, 논설문 확대

말뭉치 구축 외의 추가 연구 제안
<ul style="list-style-type: none"> ○ 학습자의 등급 표준화 방안 <ul style="list-style-type: none"> - 이주민, 국외 자료의 확대와 함께 학습 대상, 학습 환경을 고려한 준속달도 기준 및 측정 도구 개발을 위한 추가적인 연구 필요 ○ 균형성의 문제 <ul style="list-style-type: none"> - 학습자 말뭉치는 동적 말뭉치로 균형성의 문제보다는 대표성의 문제가 중요하므로 다양한 자료를 수집하되, 주요 변인인 수준별, 7개 언어권별 자료는 각 변인별로 최소 규모를 정하여 자료의 활용도를 제고하는 것이 바람직함.

- 최소 규모의 타당성은 자료 활용 목적에 따라 달라질 것이므로 이를 설정하고 검증하기 위한 추가적인 연구 필요

(2) 수집 전략

- 2021-2025년 중장기 계획에서는 2015-2020년 말뭉치 수집 및 구축 결과에 대한 분석을 토대로 수집의 용이성 제고와 자료의 다양성 확보를 위해 자료를 수집하며, 이를 위한 세부 전략은 다음과 같다.



<그림 34> 2021-2025년 학습자 말뭉치 수집 전략

① 수집 네트워크의 다변화를 통한 수집의 용이성 확보

- 2021-2025년에는 효율적인 자료 수집을 위해 2015-2020년의 자료 수집 네트워크를 기반으로 하되, 특히 국외 자료 집중 구축을 위해 국외의 대학에 교원을 파견 중인 한국국제교류재단과의 협약을 통해 네트워크를 확장하는 것을 목표로 한다.

<표 98> 수집 네트워크의 다변화

네트워크	수집 기관	수집 대상	수집 변인
한국어 교육기관	한국어학당	국내 자료	국내 자료 중 불균형한 수준, 국적, 장르, 주제
	다문화센터, 사회통합 프로그램 운영 기관, KSL 교육 기관	이주민 자료	전체

학계	국내 대학(원)	국내 자료 참조 말뭉치	학문 목적 학습자 자료
	국외 대학(원)	국외 자료	전체
유관 기관	한국국제교류재단	국외 자료	전체
	세종학당재단	국외 자료	전체

- 학계의 네트워크: 각 대학 교양학부의 글쓰기 수업, 외국인 전용 학부 학생을 대상으로 한 학문 목적 학습자 자료, 한국어 모어 화자의 참조 말뭉치
- 한국어교육 기관: 구축 팀이 제공하는 기획 자료(☞ 교육과정 자료는 수집 대상으로 하지 않음.)
- 세종학당재단: 숙달도 평가 자료
: Cambridge Learner Corpus는 현재 약 5천 만 어절에 이르는 대규모 영어 학습자 말뭉치로 Cambridge사의 숙달도 평가 자료를 지속적으로 수집함으로 해서 시험에 응시하는 전 세계의 영어 학습자 자료를 대규모로 수집하고 있다. 이러한 방식의 장점은 대규모 수집뿐만 아니라 장르, 주제 등 과제의 균질성, 학습자 집단의 다양성, 평가 등급에 따른 말뭉치 자료의 등급 체계 표준화 등을 들 수 있다.
- 한국국제교류재단: 국외 학습자 자료

② 수집 매체 확장을 통한 수집 규모의 확장성 제고

- 본 연구에서는 대규모 말뭉치로의 확장을 효율적으로 수행해 나가기 위한 방안으로 세종학당재단의 <세종학당 AI 선생님>, 구축 지원 도구 개발 업체인 (주)이르테크의 <KOKOA> 앱을 활용한 자료 수집 가능성을 탐색하였다. 그 결과, <세종학당 AI 선생님>은 학습된 시나리오를 기반으로 한 말하기 연습 앱으로 대화의 주제나 상황이 한정적이어서 자료의 활용도 면에서 말뭉치에 적절한 자료의 유형이 아니라는 결론에 이르렀다. 또한 <KOKOA>는 수집이 아닌 학습용 앱으로 설계되었기 때문에 자율적인 수집 참여가 어렵고, 사용 조건상의 제약이 있어 대규모 수집 도구로서 적절하지 않은 것으로 확인되었다.
- 이에 따라 2021-2025년 중장기 계획에서는 자료 수집을 위한 웹사이트를 구축하여 자료를 수집하는 방안을 제안한다. 웹사이트의 구축과 관리는 구축 연구팀에서 하며, 국립국어원의 학습자 말뭉치 나눔터와 누리집 배너를 통해 홍보와 안내를 하도록 한다.

- 아울러 비대면 수집 방법을 활용한다. 최근 코로나-19 감염병 사태로 비대면 접촉이 일상화되어 있으며, 줌(Zoom)을 비롯한 매체가 업무, 학업 환경 외에도 일상 속에서 광범위하게 활용되고 있다. 이에 따라 비대면 회의 도구를 사용해 자료 수집의 용이성을 확보한다.

③ 기획 수집을 통한 기구축 말뭉치의 균형성 보완

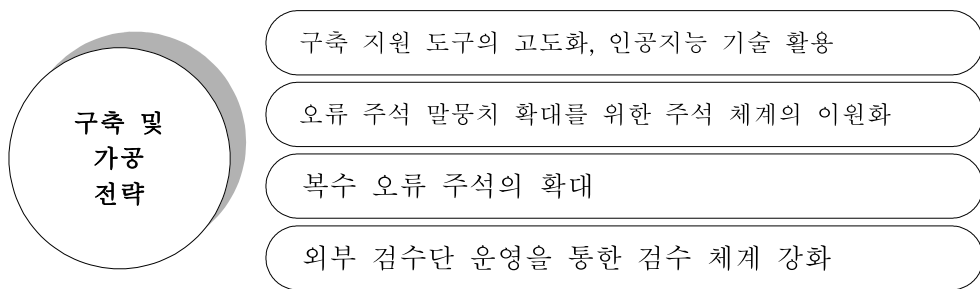
- 기획 수집은 2015-2020년 사업에서 구축한 말뭉치의 균형성을 보완하는 것을 주목적으로 하는 수집 방식이다. 2015-2020년 말뭉치 성과 분석을 바탕으로 설정한 2021-2025년 기획 수집의 방향은 다음과 같다.

<표 99> 균형성 보완을 위한 기획 수집 전략

구분		2015-2020 구축 현황	2021-2025 구축 전략
학습자 변인	대상	○ 이주민, 국외 학습자 자료의 비중이 매우 적음	○ 국내 학습자 중 학문 목적 학습자 자료 확대 ○ 이주민, 국외 학습자 자료의 비중 확대
	제1언어	○ 중국어권 자료의 비중이 가장 높고, 그 뒤를 잇는 일본어권, 영어권, 베트남어권, 태국어권 등은 그에 비해 매우 적음	○ 일본어권, 영어권, 베트남어권, 태국어권, 러시아어권, 스페인어권, 학습자의 자료 비중 확대
	수준	○ 1급, 5급, 6급 이상의 비중이 매우 적음	○ 1급, 5급, 6급의 자료 비중 확대
자료 변인	유형	○ 구어의 비중이 적음	○ 구어 비중 확대
	장르	○ 문어: 생활문에 집중됨 ○ 구어: 인터뷰에 집중됨	○ 문어: 논설문, 설명문을 중심으로 한 장르의 다양화 ○ 구어: 내러티브, 자연 발화 자료 확대
	주제	○ 사회, 일상생활, 개인 신상에 편중	○ 17개의 다양한 주제 중 수준별 집중 구축 주제를 2-3개 내외로 수집

5.5. 말뭉치 구축 및 가공 전략

- 2015-2020년 사업에서는 원시 말뭉치, 형태 주석 말뭉치, 오류 주석 말뭉치로 세 가지 유형의 말뭉치를 구축하였다. 이 중 구어는 전사에 소요되는 시간과 비용, 노력이 문어에 비해 매우 커 전체 자료의 25.5%로 그 비중이 매우 적다. 이와 마찬가지로 오류 주석 가공 말뭉치도 원시 말뭉치의 22.8%로 활용도를 고려할 때 매우 부족하다. 이에 본 연구에서는 기구축 말뭉치의 성과 분석, 학계 및 민간 분야의 사용자와 전문가 의견을 토대로 원시 말뭉치 중 구어 말뭉치의 비중 확대, 형태 주석 말뭉치와 오류 주석 말뭉치의 규모 확대를 제안하고자 한다. 이를 수행하기 위한 2021-2025년 전략은 다음과 같다.



<그림 35> 2021-2025년 학습자 말뭉치 구축 및 가공 전략

(1) 구축 지원 도구의 고도화, 인공지능 기술을 활용한 원시 말뭉치 확대

- 말뭉치 구축은 수집에서 구축, 가공의 전 공정에 소요되는 시간과 노력, 비용이 크며, 많은 전문 인력을 필요로 하는 노동 집약적인 특성이 있다. 학습자 말뭉치는 비정형의 언어를 포함한 비모어 화자의 언어 자원이라는 점에서 그러한 특성이 더욱 강하다. 이에 2015-2020년 말뭉치 구축에서는 문어 입력과 구어 전사 작업의 전 과정에 연구 인력이 투입되었다. 2021-2025년 연구에서는 구어 말뭉치의 비중이 대폭 확대되는 만큼 충분한 시간을 가지고 구어 전사에서 STT(speech to text) 기반의 전사 도구의 활용 가능성을 지속적으로 탐색하고 검증할 것을 제안한다.

(2) 오류 주석 말뭉치 확대를 위한 주석 체계의 이원화

○ 학습자 말뭉치 주석은 형태 주석과 오류 주석으로 나뉜다.

① 형태 주석 체계

- 2015-2020년 사업에서 형태 주석은 <21세기 세종 한국어 균형 말뭉치>와의 호환성을 고려하여 세종 말뭉치의 형태 주석을 수용하되, 학습자 말뭉치가 비정형 언어 자료로 이루어진다는 점을 고려하여 몇 가지 세부 처리 지침을 수정해 주석 체계를 마련하고 이에 따라 주석을 하였다.
- 형태 주석은 현재 350만 어절이 구축되었으며, 그 과정에서 주석 체계에 관한 쟁점이 특별히 발견되지 않았다. 이에 따라 1차 연도 중장기 계획에 따른 기구축 말뭉치와의 호환성을 고려하여 다음과 같이 이전의 체계를 그대로 수용할 것을 제안한다.

<표 100> 2015-2020년 학습자 말뭉치 형태 주석 체계

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

대분류	형태 주식 내용	기호	세종 표지
	목적격조사	JKO	JKO
	부사격조사	JKB	JKB
	호격조사	JKV	JKV
	인용격조사	JKQ	JKQ
	보조사	JX	JX
	접속조사	JC	JC
(6) 의존형태	선어말어미	EP	EP
	어말어미(연결)	EC	EC
	어말어미(종결)	EF	EF
	명사형전성어미	ETN	ETN
	관형형전성어미	ETM	ETM
	채언접두사	XPN	XPN
	명사파생접미사	XSN	XSN
	동사파생접미사	XSV	XSV
	형용사파생접미사	XSA	XSA
	어근	XR	XR
(7) 기호	마침표, 물음표, 느낌표	SF	SF
	쉼표, 가운뎃점, 콜론, 빗금, 줄표, 물결	SP	SP
	따옴표, 괄호표	SS	SS
	줄임표	SE	SE
	불임표 (숨김, 빼짐)	SO	SO
	외국어	SL	SL
	한자	SH	SH
	기타 기호	SW	SW
	숫자	SN	SN
	분석불능범주	NA	NA

② 오류 주석 체계

- 2020-2015년 사업에서 적용한 오류 주석 체계는 기본 주석과 확장 주석으로 나뉜다. 기본 주석은 모든 형태에 주석을 부착하는 필수 주석으로 분석 가능 여부를 판정하는 ‘분석 불가능’과 ‘오류 위치’로 구분된다. 확장 주석은 해당 항목이 있는 경우에 한 해 주석을 하는 수의적 주석으로 ‘오류 층위’와 ‘오류 양상’으로 구분된다. 다음은 기본 주석이다.

<표 101> 2015-2020년 학습자 말뭉치 오류 주석 체계: 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
		동사파생접미사	CXSV
		형용사파생접미사	CXSA
		어근	CXR
	기능어휘	주격조사	FNP
		관형격조사	FGP
		목적격조사	FOP

	오류 유형		주석 표지
		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

○ 다음은 확장 주석 중 오류 층위의 주석 체계이다.

<표 102> 2015-2020년 학습자 말뭉치 오류 주석 체계 틀: 오류 층위

	오류 유형		주석 표지
오류 층위	발음	음소	PP
		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[공용]	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

○ 다음은 확장 주석 중 오류 양상의 주석 체계이다.

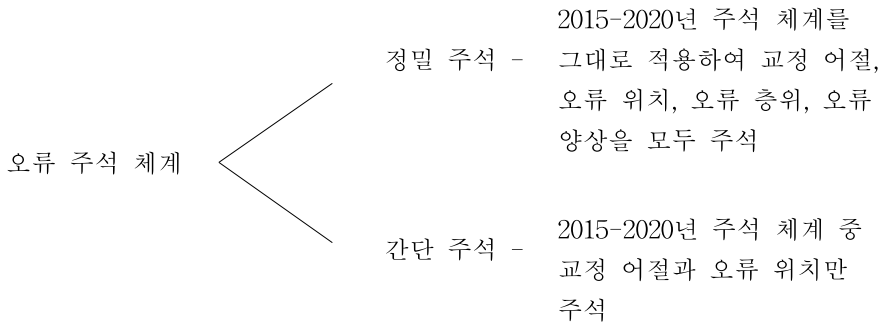
<표 103> 2015-2020년 학습자 말뭉치 오류 주석 체계: 오류 양상

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

○ 말뭉치 가공 단계 중 오류 주석은 오류의 식별과 판정, 교정 어절 제시, 오류 위치, 오류 층위, 오류 현상 등으로 다차원 주석을 하는 가장 복잡한 공정이다. 이 과정에서 오류의 식별과 판정, 주석에 대한 작업자의 자의성으로 인한 ‘주석의 일관성’ 문제는 세부 지침을 통한 교육과 합의 절차를 지속함에도 불구하고 어느 정도 감수할 수밖에 없는 과제 중 하나였다. 아울러 복잡한 공정으로 인해 소요되는 많은 시간과 노력, 비용의 문제는 원시 말뭉치나 형태 주석 말뭉치에 비해 현저하게 적은 구축 규모라는 결과로 이어졌으며, 이는 본 연구에서의 기구축 말뭉치에 대한 성과 분석을 통해 분명하게 확인되었다. 또한 다양한 분야의 사용자를 대상으로 한 의견수렴, 전문가 자문 등을 통해서도 오류 주석 말뭉치의 확대에 대한 요구가 큼을 확인할 수 있었다. 이에 따라 본 연구에서는 오류 주석 말뭉치의 규모를 현재 원시 말뭉치의 22.8%에서 50% 수준으로 확대하여 500만 어절로 설정하였다. 이는 2015-2020년에 구축한 약 100만 어절의 5배에 달하는 규모로 2022년부터 2025년까지 연간 100만 어절의

오류 주식 가공을 해야 목표 규모를 달성할 수 있다는 결론에 이른다.

- 이에 따라 본 연구에서는 활용의 측면에서 학계와 민간 분야의 요구를 수용하여 다음과 같이 간단 주식과 정밀 주식으로 주식 체계를 이원화할 것을 제안한다.



<그림 36> 오류 주식 가공 전략: 주식 체계의 이원화

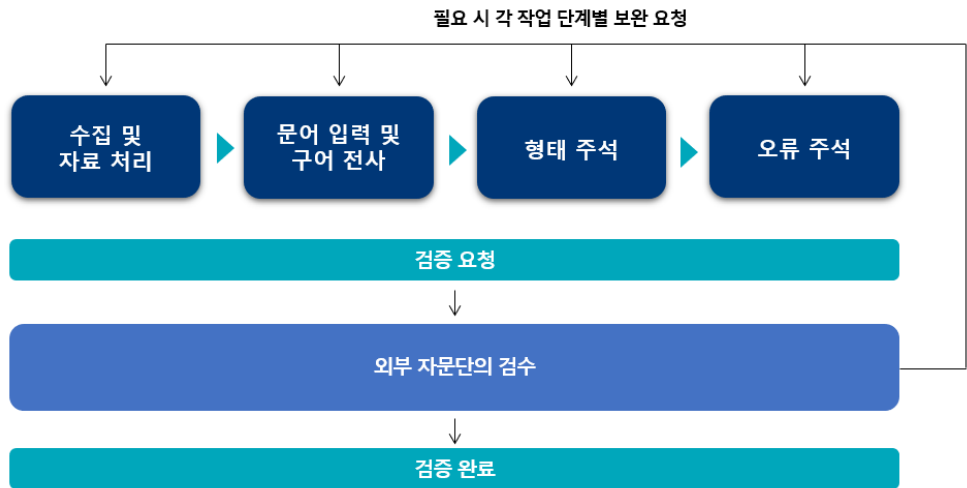
(3) 복수 오류 주식의 확대

- 정밀 주식 시 오류 층위는 주식이 중복되어 주식자에 따라 판정이 달라지는 경우가 있다. 예를 들면, 학습자가 ‘결과었다’라고 썼을 경우 이를 형태 층위의 지정사 오류로 처리할 것인지, 통사 층위의 시제 오류로 처리할 것인지가 쟁점이 된다. 이런 경우 주식의 정확성과 일관성 제고를 위해 복수 주식을 허용해 왔는데, 이를 확대할 필요가 있다.

(4) 외부 검수단 운영을 통한 검수 체계 강화

- 2015-2020년 사업에서 말뭉치 구축 및 가공 단계에서 작업과 별개로 3단계의 검수 체계를 유지해 왔다. 이는 구축 팀 내의 검수 체계로 입력, 전사, 형태 주식, 오류 주식의 작업 단계별 팀 내의 검수, 상위 작업자의 검수, 작업 단계 간 연계를 통한 검수의 절차를 거쳐 왔다. 그 밖에도 그림에도 구축 지원 도구를 통한 이상 표본 검증, 중복 표본 검증, 표본 등록 정보 검수 등의 절차가 별도로 이루어져 왔다. 그림에도 많은 작업자가 참여하여 구축하는 대규모의 말뭉치라는 특성상 사용 과정에서 발생하는 오류가 보고되고 있다. 대개는 사소한 오류이거나 관점에 따라 오류 여부

가 달라질 수 있는 것이지만 사용자의 입장에서 말뭉치의 신뢰성을 떨어뜨릴 수 있고, 국가 언어 자원으로로서의 신뢰성 확보는 중요한 문제이므로 검수 체계를 한층 강화할 필요가 있다. 이에 구축 팀 외에 수요 기관 또는 별도의 검수 팀을 운영하여 검수 체계를 강화할 것을 제안한다. 이를 통해 보다 체계적이고 효율적으로 말뭉치의 질을 높일 수 있을 것이다. 다음은 외부 검수 팀의 검수 단계를 추가했을 때의 검수 체계를 나타낸 것이다.



<그림 37> 외부 검수단 운영을 통한 검수 체계 강화 모형

Ⅲ. 말뭉치 수집 및 구축·가공

1. 말뭉치 수집

1.1. 수집 대상

- 2021년 학습자 말뭉치 구축을 위한 수집은 기획 수집으로 한정되었으며, 국내의 학습자 중 대학(원)의 학문 목적 학습자와 4급 이상의 중·고급 학습자를 대상으로 한 중점 수집과 그 밖의 학습자를 대상으로 한 수집이 함께 이루어졌다.

1.2. 수집 네트워크

- 2021년 자료 수집 네트워크는 국내 대학(원)의 학문 목적 학습자를 가르치고 있는 교강사와 대학 부설 한국어 교육 기관이 중심이 되었다.
 - 학계의 네트워크와 인력풀을 활용한 참여 독려
: 국내 대학 교양학부, 외국인 전용 학부 및 대학원 교강사
 - 대학 부설 한국어 교육 기관

1.3. 수집 과제

- 수집 과제는 2015-2020년에 상대적으로 부족한 장르와 주제의 자료를 집중 수집할 수 있도록 설계하였으며, 초급과 중·고급으로 이원화하였다.

(1) 문어

- 문어는 초급은 생활문 중 비교하기, 중·고급은 논설문으로 각각 다음과 같이 두 가지 주제를 제시하고 그중 하나를 선택하여 과제를 수행하도록 하였다.

<표 104> 기획 과제: 문어

수준	주제	장르
초급	주제1. 자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, ……) 주제2. 내가 가장 좋아하는 것과 싫어하는 것 (일, 행동, 말, 사람, 물건, ……)	생활문
중·고급	주제 1. 과학 기술의 발전이 인간의 생활에 미치는 영향 (인터넷, 로봇, 인공지능(AI), ……) 주제 2. 인구 문제와 미래 사회 (저출산, 고령화, 인구 절벽(급격한 인구 감소), 1인 가구, ……)	논설문

(2) 구어

- 구어는 기존의 자료가 주로 성취도 평가 자료로 교사의 질문에 학습자가 답하는 인터뷰 자료의 비중이 높음을 참고하여 학습자의 자연스러운 발화가 중심이 되는 내러티브 과제를 설계하였다. 자료를 수집하는 과정에서 학습자가 미숙한 경우 교사가 수집 준비와 진행을 돕는 진행자로서 참여할 수 있으나, 학습자 스스로 과제 수행 방식과 내용이 포함된 PPT를 보면서 다음의 흐름에 따라 말하는 것을 원칙으로 하였다.

<표 105> 기획 과제: 구어

발화 구성	세부 발화 내용
자기소개	간단한 자기소개
과거	태어난 곳, 고향 소개 어릴 때 성격 기억나는 친구 기억나는 일 어릴 때 꿈
현재	사는 곳 성격 좋아하는 것과 싫어하는 것

	하는 일 꿈, 진로
미래	앞으로 10년 후의 내 모습 노후의 삶 (65세 이후 어떻게 살고 싶은가) 죽기 전에 꼭 하고 싶은 일

1.4. 자료 수집 현황

- 자료 수집은 전량 기획 수집으로 진행되었으며, 한 명의 학습자가 문어 과제와 구어 과제를 수행하여 자료를 제출하거나 학습자의 의사에 따라 둘 중 하나에만 참여하기도 하였다.

(1) 문어

- 문어 자료는 총 52개 국가의 자료 579개가 수집되었다.

<표 106> 문어 자료 수집 현황 및 분포

국적	1급	2급	3급	4급	5급	6급	합계
베트남	2	9	42	53	22	17	145
일본	8	3	4	58	42	25	140
중국	14	21	28	20	6	3	92
몽골		11	5	10	7	7	40
독일	10	3			2	1	16
프랑스	7	4	1	2	2		16
인도네시아			7	2	1	3	13
미얀마			4	3	3	1	11
대만	1	1		4	1	3	10
스웨덴	5	3				2	10
스페인	5	2					7
필리핀			2		5		7

국적	1급	2급	3급	4급	5급	6급	합계
러시아			2	2		2	6
이탈리아	2	2		2			6
우즈베키스탄		1	1		3		5
말레이시아	1		1		2		4
미국	2				2		4
네덜란드	1					1	2
도미니카공화국		1	1				2
멕시코	1		1				2
스리랑카			1		1		2
스위스	1			1			2
에티오피아			2				2
인도		2					2
카자흐스탄						2	2
캄보디아		1	1				2
타지키스탄		1			1		2
태국				1	1		2
파키스탄		1			1		2
나이지리아		1					1
네덜란드, 모로코	1						1
루마니아					1		1
바레인				1			1
베트남			1				1
벨기에	1						1
벨라루스				1			1
부룬디		1					1
브라질				1			1
세르비아					1		1
싱가포르		1					1
아르메니아			1				1

국적	1급	2급	3급	4급	5급	6급	합계
아제르바이잔		1					1
영국	1						1
오스트리아		1					1
우간다		1					1
우루과이			1				1
우크라이나					1		1
이집트					1		1
짐바브웨						1	1
케냐				1			1
콜롬비아	1						1
폴란드	1						1
합계	65	72	106	162	106	68	579

(2) 구어

○ 구어 자료는 총 48개 국가의 자료 360개가 수집되었다.³⁵⁾

<표 107> 구어 자료 수집 현황 및 분포

국적	1급	2급	3급	4급	5급	6급	합계
일본	8	3	3	52	28	16	110
중국	7	10	12	15	5	1	50
베트남	1	3	10	13	11	10	48
몽골		4	2	6	3	3	18
프랑스	7	4	1	2	1		15
독일	9	3			1	1	14
인도네시아			6	2	1	2	11
스웨덴	5	3				1	9

35) 구어 자료는 보고서 작성 시점을 기준으로 수집이 진행 중이어서 360개 외의 자료가 추가될 예정이다.

국적	1급	2급	3급	4급	5급	6급	합계
미얀마			1	3	3	1	8
스페인	5	2					7
필리핀			2		5		7
대만	1			2	1	2	6
러시아			2	2		1	5
이탈리아	2	1		2			5
말레이시아	1		1		2		4
미국	1	1			1		3
도미니카공화국		1	1				2
멕시코	1		1				2
브라질				2			2
스위스	1			1			2
우즈베키스탄		1			1		2
인도		2					2
캄보디아		1	1				2
타지키스탄		1			1		2
파키스탄		1			1		2
나이지리아		1					1
네덜란드						1	1
네덜란드, 모로코	1						1
루마니아					1		1
바레인				1			1
벨기에	1						1
벨라루스				1			1
세르비아					1		1
싱가포르		1					1
아르메니아			1				1
아제르바이잔		1					1
에티오피아			1				1

국적	1급	2급	3급	4급	5급	6급	합계
영국	1						1
오스트리아		1					1
우간다		1					1
우루과이			1				1
우크라이나					1		1
이집트					1		1
짐바브웨						1	1
콜롬비아	1						1
태국					1		1
폴란드	1						1
합계	54	46	46	104	70	40	360

2. 구축 및 가공

2.1. 원시 말뭉치

(1) 문어

- 2021년 문어 원시 말뭉치는 419,371어절이 구축되었다. 그 결과 2015년부터 2020년까지 구축한 말뭉치와 합산하여 누적 합계 3,697,952어절 규모의 문어 원시 말뭉치가 구축되었다.

① 숙달도별 자료 분포

- 문어 원시 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 숙달도별 구축 규모를 집계한 것이다.

<표 108> 문어 원시 말뭉치의 속달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015-2020	어절 수	385,317	535,903	626,589	603,099	575,274	409,248	143,151	3,278,581
	파일 수	5,732	5,483	5,239	4,626	3,682	2,483	155	27,400
2021	어절 수	53,953	59,943	59,807	47,991	146,242	49,933	1,502	419,371
	파일 수	958	649	536	357	972	322	8	3,802
합계	어절 수	439,270	595,846	686,396	651,090	721,516	459,181	144,653	3,697,952
	파일 수	6,690	6,132	5,775	4,983	4,654	2,805	163	31,202

② 언어권별 자료 분포

- 문어 원시 말뭉치는 138개국³⁶⁾ 92개 언어권의 자료가 구축되었다. 다음은 문어 원시 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 109> 문어 원시 말뭉치의 언어권별 자료 분포

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	1,565,848	12,390	155,494	1,455	1,721,342	13,845
일본어	432,960	3,535	82,438	777	515,398	4,312
베트남어	223,757	2,143	28,998	263	252,755	2,406

36) 138개 국가에는 중국, 일본, 베트남, 홍콩, 대만, 미국, 러시아, 태국, 몽골, 말레이시아, 인도네시아, 싱가포르, 우즈베키스탄, 카자흐스탄, 프랑스, 스웨덴, 한국, 필리핀, 영국, 캐나다, 이탈리아, 독일, 호주, 스페인, 키르기스스탄, 스리랑카, 캄보디아, 멕시코, 터키, 브라질, 인도, 미얀마, 방글라데시, 네팔, 마카오, 사우디아라비아, 노르웨이, 이란, 파키스탄, 콜롬비아, 우루과이, 에콰도르, 네덜란드, 우크라이나, 포르투갈, 페루, 모로코, 파나마, 아제르바이잔, 아랍에미리트연합, 벨기에, 우간다, 나이지리아, 폴란드, 뉴질랜드, 투르크메니스탄, 베네수엘라, 이디오피아, 튀니지, 아프리카니스탄, 이집트, 가나, 스위스, 라오스, 볼리비아, 엘살바도르, 핀란드, 케냐, 헝가리, 가봉, 칠레, 르완다, 브루나이, 아르헨티나, 동티모, 세네갈, 벨라루스, 콩고, 덴마크, 탄자니아, 시리아, 도미니카 공화국, 루마니아, 타지키스탄, 불가리아, 요르단, 이라크, 이스라엘, 체코, 아르메니아, 알제리, 파라과이, 오스트리아, 에스토니아, 슬로바키아, 카메룬, 코트디부아르, 예멘, 남아프리카, 수단, 코스타리카, 세르비아, 리투아니아, 니카라과, 마다가스카르, 팔레스타인, 슬로베니아, 앙골라, 남수단, 과테말라, 룩셈부르크, 쿠웨이트, 몰도바, 자메이카, 레바논, 쿠바, 잠비아, 그리스, 트리니다드 토바고, 리비아, 보츠와나, 카타르, 라트비아, 모잠비크, 라이베리아, 말레이, 도미니카 연방, 베닝, 온두라스, 북한, 알바니아, 바베이도스, 피지, 아일랜드, 그루지아, 콩고 민주 공화국, 저지, 아이슬란드 포함된다.

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
영어	217,155	1,925	23,963	237	241,118	2,162
광둥어	158,170	1,304	26,013	240	184,183	1,544
러시아어	102,817	947	26,396	254	129,213	1,201
태국어	69,024	619	28,157	139	97,181	758
몽골어	67,256	629	5,912	66	73,168	695
스페인어	40,075	369	16,941	167	57,016	536
인도네시아어	44,636	368	1,862	17	46,498	385
프랑스어	42,434	384	959	11	43,393	395
말레이어	27,582	183	932	8	28,514	191
스웨덴어	25,210	284	820	5	26,030	289
카자흐어	21,789	179	1,089	7	22,878	186
아랍어	19,312	204	1,336	11	20,648	215
이탈리아어	18,969	135	870	8	19,839	143
싱할라어	16,798	101	878	5	17,676	106
우즈베크어	14,264	131	2,400	27	16,664	158
한국어	15,052	56	1,293	8	16,345	64
독일어	14,842	142	1,472	11	16,314	153
포르투갈어	11,960	111	1,965	15	13,925	126
타갈로그어	12,719	168	424	2	13,143	170
터키어	9,672	85	1,857	12	11,529	97
키르기스어	10,122	79	462	5	10,584	84
기타 ³⁷⁾	96,158	929	6,440	52	102,598	981
합계	3,278,581	27,400	419,371	3,802	3,697,952	31,202

37) 기타에는 중국어, 일본어, 베트남어, 영어, 광둥어, 러시아어, 태국어, 몽골어, 스페인어, 인도네시아어, 프랑스어, 말레이어, 스웨덴어, 카자흐어, 아랍어, 이탈리아어, 싱할라어, 우즈베크어, 한국어, 독일어, 포르투갈어, 타갈로그어, 터키어, 키르기스어, 기타, 합계, , 버마어, 크메르어, 페르시아어, 네팔어, 벵골어, 노르웨이어, 네덜란드어, 라오어, 투르크멘어, 힌디어, 헝가리어, 우크라이나어, 아제르바이잔어, 우르두어, 타밀어, 폴란드어, 스와힐리어, 핀란드어, 암하라어, 루마니아어, 불가리아어, 르완다어, 쿠르드어, 테툼어, 덴마크어, 타지크어, 아르메니아어, 간다어, 슬로바키아어, 위구르어, 이그보어, 카탈루냐어, 히브리어, 체코어, 중국어(만다린어), 자바어, 룩셈부르크어, 세르비아어, 칸나다어, 슬로베니아어, 노르웨이어 (뉘노르스크), 에스토니아어, 마다가스카르어, 리투아니아어,

(2) 구어

- 2021년 구어 원시 말뭉치는 411,771어절이 구축되었다. 그 결과 2015년부터 2020년까지 구축한 말뭉치와 합산하여 누적 합계 1,522,612어절 규모의 구어 원시 말뭉치가 구축되었다.

① 숙달도별 자료 분포

- 구어 원시 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 숙달도별 구축 규모를 집계한 것이다.

<표 110> 구어 원시 말뭉치의 숙달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 - 2020	어절 수	205,149	225,004	227,925	194,128	118,922	86,720	52,993	1,110,841
	파일 수	685	506	502	412	212	138	86	2,541
2021	어절 수	53,621	98,243	149,089	49,607	30,593	19,475	11,143	411,771
	파일 수	114	160	148	101	44	25	9	601
합계	어절 수	258,770	323,247	377,014	243,735	149,515	106,195	64,136	1,522,612
	파일 수	799	666	650	513	256	163	95	3,142

② 언어권별 자료 분포

- 구어 원시 말뭉치는 85개국³⁸⁾ 49개 언어권의 자료가 구축되었다. 다음은

핀란드어, 그리스어, 웨일스어, 조지아어, 마라티어, 티베트어, 말라얄람어, 텔루구어, 벨라루스어, 덩카어, 티그리냐어, 트위어, 마오리어, 알바니아어, 구자라트어, 피지어, 월로프어, 세부아노어, 라트비아어, 파슈토어, 아이슬란드어, 판테어(Fanti), 츠와나어, 아프리카칸스어권이 포함된다.

- 38) 85개 국가에는 중국, 베트남, 태국, 일본, 인도네시아, 우즈베키스탄, 필리핀, 러시아, 대만, 스리랑카, 몽골, 우루과이, 키르기스스탄, 미국, 카자흐스탄, 말레이시아, 에콰도르, 캄보디아, 미얀마, 멕시코, 스페인, 이탈리아, 영국, 한국, 파나마, 콜롬비아, 포르투갈, 베네수엘라, 싱가포르, 페루, 프랑스, 캐나다, 네팔, 아랍에미리트연합, 코스타리카, 호주, 아르헨티나, 홍콩, 터키, 이집트, 볼리비아, 스웨덴, 독일, 브라질, 사우디아라비아, 이디오피아, 르완다, 방글라데시, 네덜란드, 폴란드, 가나, 엘살바도르, 케냐, 벨라루스, 도미니카 공화국, 정보 없음, 파키스탄, 모로코, 아제르바이잔, 벨기에, 우간다, 나이지리아, 뉴질랜드, 튀니지, 아프카니스탄, 스위스, 헝가리, 동티모, 루마니아, 타지키스탄, 불가리아, 요르단, 아르메니아, 알제리, 파라과이, 에스토니아, 코트디부아르, 예멘, 세르비아, 과테말라, 보츠와나, 미국령 사모아, 모리타니, 오만, 소말리아가 포함된다.

구어 원시 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 111> 구어 원시 말뭉치의 언어권별 자료 분포

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	261,572	667	83,594	204	345,166	871
태국어	120,728	308	173,581	135	294,309	443
베트남어	186,671	377	41,090	79	227,761	456
일본어	120,906	260	21,618	28	142,524	288
스페인어	36,866	75	58,505	106	95,371	181
러시아어	77,139	209	1,940	5	79,079	214
인도네시아어	60,164	138	0	0	60,164	138
영어	49,868	100	6,461	9	56,329	109
타갈로그어	44,884	89	339	1	45,223	90
싱할라어	30,505	52	0	0	30,505	52
버마어	20,326	24	1,726	2	22,052	26
키르기스어	19,478	38	0	0	19,478	38
몽골어	9,585	37	5,754	9	15,339	46
우즈베크어	14,210	27	220	1	14,430	28
기타 ³⁹⁾	57,939	140	16,943	22	74,882	162
합계	1,110,841	2,541	411,771	601	1,522,612	3,142

39) 기타에는 아랍어, 카자흐어, 한국어, 크메르어, 프랑스어, 포르투갈어, 네팔어, 세부아노어, 광둥어, 이탈리아어, 독일어, 터키어, 네덜란드어, 암하라어, 타타르어, 타지크어, 말레이어, 스웨덴어, 폴란드어, 스와힐리어, 세르비아어, 루마니아어, 벵골어, 아르메니아어, 불가리아어, 에스토니아어, 자바어, 페르시아어, 르완다어, 아제르바이잔어, 테툼어, 폴라어, 헝가리어, 우르두어, 라틴어권이 포함된다.

2.2. 형태 주식 말뭉치

(1) 문어

- 2021년 문어 형태 주식 말뭉치는 100,781어절이 구축되었다. 그 결과 2015년부터 2020년까지 구축한 말뭉치와 합산하여 누적 합계 2,602,914어절 규모의 문어 형태 주식 말뭉치가 구축되었다.

① 숙달도별 자료 분포

- 문어 형태 주식 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 숙달도별 구축 규모를 집계한 것이다.

<표 112> 문어 형태 주식 말뭉치의 숙달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 - 2020	어절 수	356,901	428,486	444,985	421,797	408,938	369,900	71,126	2,502,133
	파일 수	5,179	4,330	3,709	3,300	2,731	2,309	83	21,641
2021	어절 수	23,687	5,225	0	725	34,093	37,051	0	100,781
	파일 수	387	71	0	5	254	283	0	1,000
합계	어절 수	380,588	433,711	444,985	422,522	443,031	406,951	71,126	2,602,914
	파일 수	5,566	4,401	3,709	3,305	2,985	2,592	83	22,641

② 언어권별 자료 분포

- 문어 형태 주식 말뭉치는 133개국⁴⁰⁾ 88개 언어권의 자료가 구축되었다.

40) 133개 국가에는 중국, 일본, 베트남, 미국, 홍콩, 태국, 대만, 러시아, 몽골, 말레이시아, 인도네시아, 싱가포르, 우즈베키스탄, 카자흐스탄, 프랑스, 스웨덴, 한국, 필리핀, 영국, 캐나다, 이탈리아, 호주, 독일, 키르기즈스탄, 캄보디아, 스페인, 스리랑카, 터키, 인도, 미얀마, 멕시코, 방글라데시, 사우디아라비아, 네팔, 브라질, 노르웨이, 파키스탄, 이란, 콜롬비아, 마카오, 우루과이, 포르투갈, 네덜란드, 우크라이나, 에콰도르, 아제르바이잔, 모로코, 나이지리아, 투르크메니스탄, 우간다, 뉴질랜드, 벨기에, 튀니지, 아랍에미리트연합, 폴란드, 가나, 스위스, 파나마, 페루, 가봉, 이디오피아, 이집트, 아프가니스탄, 베네수엘라, 케냐, 세네갈, 엘살바도르, 벨라루스, 라오스, 덴마크, 핀란드, 볼리비아, 콩고, 아르헨티나, 브루나이, 탄자니아, 칠레, 르완다, 루마니아, 도미니카 공화국, 동티모, 불가리아, 헝가리, 이라크, 타지키스탄, 아르메니아, 시리아, 요르단, 카메룬, 예멘, 알제리, 에스토니아, 남아프리카, 체코, 세르비아, 코트디부아르, 이스라엘, 니카라과, 수단,

다음은 문어 형태 주식 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 113> 문어 형태 주식 말뭉치의 언어권별 자료 분포

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	939,328	7,880	20,620	177	959,948	8,057
일본어	423,276	3,478	25,933	205	449,209	3,683
베트남어	221,882	2,126	10,753	119	232,635	2,245
영어	212,280	1,891	15,166	160	227,446	2,051
러시아어	98,741	910	8,267	81	107,008	991
광둥어	81,801	643	3,293	68	85,094	711
태국어	67,916	610	6,844	50	74,760	660
몽골어	65,726	612	1,587	21	67,313	633
인도네시아어	43,037	359	915	15	43,952	374
스페인어	39,697	365	2,585	36	42,282	401
프랑스어	39,580	362	518	11	40,098	373
말레이어	24,418	164	192	5	24,610	169
카자흐어	21,114	175	149	1	21,263	176
이탈리아어	17,772	121	174	4	17,946	125
아랍어	16,663	177	0	0	16,663	177
스웨덴어	16,193	205	35	1	16,228	206
싱갈라어	16,037	95	28	1	16,065	96
우즈베크어	13,999	129	236	5	14,235	134
한국어	12,990	47	0	0	12,990	47
독일어	12,342	119	327	4	12,669	123
타갈로그어	11,181	151	35	1	11,216	152
기타 ⁴¹⁾	106,160	1,022	3,124	35	109,284	1,057
합계	2,502,133	21,641	100,781	1,000	2,602,914	22,641

오스트리아, 파라과이, 코스타리카, 마다가스카르, 슬로바키아, 남수단, 슬로베니아, 자메이카, 잠비아, 룩셈부르크, 트리니다드 토바고, 리투아니아, 쿠바, 쿠웨이트, 과테말라, 보츠와나, 라이베리아, 앙골라, 리비아, 바베이도스, 콩고 민주 공화국, 아일랜드, 바레인, 도미니카 연방, 카타르, 팔레스타인, 온두라스, 저지, 그리스, 알바니아, 그루지아, 아이슬란드, 몰도바, 모잠비크가 포함된다.

(2) 구어

- 2021년 구어 형태 주식 말뭉치는 100,200어절이 구축되었다. 그 결과 2015년부터 2020년까지 구축한 말뭉치와 합산하여 누적 합계 1,101,672어절 규모의 구어 형태 주식 말뭉치가 구축되었다.

① 숙달도별 자료 분포

- 구어 형태 주식 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 숙달도별 구축 규모를 집계한 것이다.

<표 114> 구어 형태 주식 말뭉치의 숙달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 - 2020	어절 수	196,790	196,798	205,309	177,406	108,999	86,473	29,697	1,001,472
	파일 수	651	431	451	366	196	137	33	2,265
2021	어절 수	15,030	8,691	1,889	31,608	27,643	15,339	0	100,200
	파일 수	44	11	4	63	43	21	0	186
합계	어절 수	211,820	205,489	207,198	209,014	136,642	101,812	29,697	1,101,672
	파일 수	695	442	455	429	239	158	33	2,451

- 41) 기타에는 포르투갈어, 키르기스어, 터키어, 버마어, 크메르어, 벵골어, 네팔어, 페르시아어, 노르웨이어, 네덜란드어, 투르크멘어, 아제르바이잔어, 힌디어, 우크라이나어, 라오어, 우르두어, 폴란드어, 타밀어, 헝가리어, 루마니아어, 스와힐리어, 쿠르드어, 덴마크어, 암하라어, 불가리아어, 핀란드어, 타지크어, 르완다어, 아르메니아어, 카탈루냐어, 태툼어, 이그보어, 간다어, 룩셈부르크어, 세르비아어, 체코어, 슬로바키아어, 자바어, 슬로베니아어, 위구르어, 노르웨이어 (뉘노르스크), 히브리어, 마다가스카르어, 칸나다어, 핀자브어, 조지아어, 중국어(만다린어),馬拉티어, 티베트어, 말라얄람어, 에스토니아어, 텔루구어, 그리스어, 벨라루스어, 덩카어, 티그리냐어, 리투아니아어, 마오리어, 알바니아어, 구자라트어, 윌로프어, 세부아노어, 파슈토어, 아이슬란드어, 판테어(Fanti), 츠와나어, 아프리카ンス어권이 포함된다.

② 언어권별 자료 분포

- 구어 형태 말뭉치는 69개국⁴²⁾ 37개 언어권의 자료가 구축되었다. 다음은 구어 형태 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 115> 구어 형태 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	224,905	550	15,773	24	240,678	574
베트남어	158,862	313	30,410	67	189,272	380
일본어	114,760	250	21,678	29	136,438	279
태국어	115,258	292	1,032	1	116,290	293
러시아어	67,111	191	2,149	6	69,260	197
인도네시아어	57,612	134	0	0	57,612	134
영어	45,446	91	10,019	17	55,465	108
스페인어	35,072	71	15,585	33	50,657	104
타갈로그어	44,256	88	339	1	44,595	89
싱할라어	30,505	52	0	0	30,505	52
버마어	19,399	22	0	0	19,399	22
키르기스어	18,372	36	0	0	18,372	36
우즈베크어	13,445	24	665	2	14,110	26
기타 ⁴³⁾	56,469	151	2,550	6	59,019	157
합계	1,001,472	2,265	100,200	186	1,101,672	2,451

42) 69개 국가에는 중국, 베트남, 태국, 일본, 인도네시아, 우즈베키스탄, 필리핀, 러시아, 스리랑카, 미국, 카자흐스탄, 우루과이, 대만, 키르기즈스탄, 몽골, 말레이시아, 캄보디아, 미얀마, 이탈리아, 스페인, 멕시코, 영국, 한국, 포르투갈, 싱가포르, 에콰도르, 캐나다, 프랑스, 호주, 아랍에미리트연합, 베네수엘라, 이집트, 스웨덴, 독일, 페루, 터키, 사우디 아라비아, 네팔, 브라질, 가나, 아르헨티나, 정보 없음, 홍콩, 방글라데시, 파키스탄, 콜롬비아, 네덜란드, 아제르바이잔, 모로코, 나이지리아, 우간다, 뉴질랜드, 벨기에, 폴란드, 스위스, 벨라루스, 르완다, 불가리아, 헝가리, 아르메니아, 요르단, 예멘, 알제리, 세르비아, 파라과이, 코스타리카, 보츠와나, 소말리아, 오만 포함된다.

43) 기타에는 아랍어, 몽골어, 카자흐어, 크메르어, 포르투갈어, 세부아노어, 이탈리아어, 프랑스어, 독일어, 광둥어, 네팔어, 스웨덴어, 네덜란드어, 세르비아어, 터키어, 벵골어, 아르메니아어, 불가리아어, 아제르바이잔어, 헝가리어, 우르두어, 말레이어, 폴란드어, 라틴어권이 포함된다.

2.3. 오류 주석

(1) 문어

- 2021년 문어 오류 주석 말뭉치는 104,314어절이 구축되었다. 그 결과 2015년부터 2020년까지 구축한 말뭉치와 합산하여 누적 함께 590,548어절 규모의 문어 오류 주석 말뭉치가 구축되었다.

① 속달도별 자료 분포

- 문어 오류 주석 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 속달도별 구축 규모를 집계한 것이다.

<표 116> 문어 오류 주석 말뭉치의 속달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 - 2020	어절 수	83,469	90,420	88,502	82,766	77,161	75,099	3,693	501,110
	파일 수	1,214	928	767	686	548	481	20	4,644
2021	어절 수	7,264	13,904	16,874	3,603	32,447	30,222	0	104,314
	파일 수	100	155	159	31	226	211	0	882
합계	어절 수	89,438	102,422	104,095	85,942	101,518	103,440	3,693	590,548
	파일 수	1,298	1,056	911	713	718	680	20	5,396

② 언어권별 자료 분포

- 문어 오류 주석 말뭉치는 85개국⁴⁴⁾ 55개 언어권의 자료가 구축되었다. 다

44) 85개 국가에는 일본, 미국, 중국, 베트남, 태국, 싱가포르, 러시아, 대만, 캐나다, 말레이시아, 호주, 영국, 카자흐스탄, 우즈베키스탄, 필리핀, 한국, 캄보디아, 네팔, 스페인, 키르기스스탄, 인도네시아, 우루과이, 포르투갈, 프랑스, 홍콩, 몽골, 독일, 파키스탄, 방글라데시, 사우디아라비아, 뉴질랜드, 미얀마, 이탈리아, 아랍에미리트연합, 인도, 우크라이나, 스웨덴, 가나, 나이지리아, 이집트, 벨라루스, 네덜란드, 터키, 헝가리, 모로코, 아프가니스탄, 도미니카 공화국, 멕시코, 스리랑카, 아르헨티나, 아제르바이잔, 폴란드, 이란, 우간다, 니카라과, 마카오, 케냐, 자메이카, 잠비아, 남아프리카, 루마니아, 세르비아, 벨기에, 요르단, 불가리아, 콜롬비아, 리투아니아, 르완다, 마다가스카르, 아일랜드, 시리아, 오스트리아, 라오스, 슬로바키아, 노르웨이, 스위스, 페루, 세네갈, 남수단, 핀란드, 볼리비아, 과테말라, 브루나이, 가봉, 이디오피아가 포함된다.

음은 문어 오류 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 117> 문어 오류 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
영어	112,576	1,018	42,526	400	155,102	1,418
일본어	112,743	1,046	41,989	301	154,732	1,347
중국어	110,372	982	3,300	33	113,672	1,015
베트남어	57,360	564	9,097	61	66,457	625
러시아어	31,040	298	201	3	31,241	301
태국어	24,528	199	339	3	24,867	202
스페인어	7,030	78	1,083	12	8,113	90
아랍어	4,113	33	55	1	4,168	34
기타 ⁴⁵⁾	41,348	426	5,724	68	47,072	494
합계	501,110	4,644	104,314	882	605,424	5,526

(2) 구어

- 2021년 구어 오류 주석 말뭉치는 47,592어절이 구축되었다. 그 결과 2015년부터 2020년까지 구축한 말뭉치와 합산하여 누적 합계 563,300어절 규모의 구어 원시 말뭉치가 구축되었다.

① 숙달도별 자료 분포

- 구어 오류 주석 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 숙달도별 구축 규모를 집계한 것이다.

45) 기타에는 카자흐어, 프랑스어, 크메르어, 광둥어, 타갈로그어, 네팔어, 포르투갈어, 한국어, 우즈베크어, 몽골어, 인도네시아어, 말레이어, 독일어, 키르기스어, 우르두어, 버마어, 벵골어, 이탈리아어, 스웨덴어, 네덜란드어, 터키어, 우크라이나어, 힌디어, 루마니아어, 아제르바이잔어, 헝가리어, 폴란드어, 페르시아어, 슬로바키아어, 자바어, 세르비아어, 마다가스카르어, 싱할라어, 히브리어, 노르웨이어, 핀란드어, 카탈루냐어, 이그보어, 텔루구어, 불가리아어, 르완다어, 콰슈토어, 세부아노어, 리투아니아어, 라오어, 암하라어가 포함된다.

<표 118> 구어 오류 주석 말뭉치의 속달도별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015-2020	어절 수	89,604	97,163	90,832	91,655	69,554	54,956	7,068	500,832
	파일 수	300	211	224	214	120	75	12	1,156
2021	어절 수	4,223	7,332	10,932	18,237	4,564	2,304	0	47,592
	파일 수	9	9	27	20	5	3	0	73
합계	어절 수	95,122	106,397	103,045	110,319	82,208	59,141	7,068	563,300
	파일 수	325	247	266	238	181	90	12	1,359

② 언어권별 자료 분포

- 구어 오류 주석 말뭉치는 57개국⁴⁶⁾ 35개 언어권의 자료가 구축되었다. 다음은 구어 오류 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 119> 구어 오류 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	109,862	278	17,385	23	127,247	301
일본어	111,765	245	1,709	2	113,474	247
베트남어	106,517	221	2,402	4	108,919	225
영어	43,081	88	0	0	43,081	88
인도네시아어	29,013	66	3,664	4	32,677	70
스페인어	25,597	61	2,703	3	28,300	64
태국어	20,598	60	1,764	6	22,362	66
러시아어	15,340	36	2,418	3	17,758	39

46) 57개 국가에는 중국, 일본, 베트남, 인도네시아, 태국, 우루과이, 미국, 대만, 말레이시아, 러시아, 이탈리아, 필리핀, 키르기스스탄, 카자흐스탄, 스페인, 영국, 우즈베키스탄, 포르투갈, 몽골, 한국, 캐나다, 싱가포르, 호주, 캄보디아, 아랍에미리트연합, 프랑스, 이집트, 미얀마, 스웨덴, 멕시코, 네팔, 독일, 사우디아라비아, 터키, 스리랑카, 정보 없음, 홍콩, 파키스탄, 방글라데시, 벨라루스, 네덜란드, 헝가리, 모로코, 아르헨티나, 아제르바이잔, 폴란드, 세르비아, 벨기에, 요르단, 불가리아, 콜롬비아, 예멘, 오만, 알제리, 파라과이, 소말리아, 코스타리카가 포함된다.

47) 기타에는 아랍어, 타갈로그어, 카자흐어, 키르기스어, 포르투갈어, 이탈리아어, 몽골어,

언어권	2015-2020		2021		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
기타 ⁴⁷⁾	39,059	101	15,547	28	54,606	129
합계	500,832	1,156	47,592	73	548,424	1,229

IV. 말뭉치 구축 지원 도구 검증

- 본 연구에서는 2021년 말뭉치 구축 및 가공 작업을 위한 안정적인 구축 환경을 유지하고 2022-2025년까지의 대규모 구축의 효율성을 높이기 위하여 지속적인 모니터링을 통해 학습자 말뭉치 구축 지원 도구의 안정성을 검증하였다. 또한 구축 지원 도구의 성능 평가를 통해 개선 사항을 도출해냄으로써 세부적인 기능을 고도화하기 위한 방향성을 제시하였다.

1. 구축 지원 도구 성능 평가팀 운영을 통한 피드백

- 본 연구에서는 표본 등록, 입력, 전사, 형태 주석 및 오류 주석 작업 과정에서 구축 도구를 실질적으로 사용하는 팀장급 관리자들을 중심으로 구축 지원 도구 성능 평가 팀을 구성하여 구축 도구를 모니터링하고 각 팀원들의 사용 의견을 수렴하여 구축 지원 도구 관리 업체에 정보를 제공하도록 하였다.
- 구축 지원 도구의 성능에 대한 피드백은 안정적인 환경에서 말뭉치 구축 작업을 수행할 수 있도록 오작동과 다양한 유형의 버그를 관리하고 세부적인 기능을 모니터링하는 데에 초점이 두어졌다.

프랑스어, 우즈베크어, 버마어, 광둥어, 네팔어, 크메르어, 싱할라어, 독일어, 스웨덴어, 네덜란드어, 세르비아어, 터키어, 벵골어, 불가리아어, 아제르바이잔어, 헝가리어, 우르두어, 말레이어, 폴란드어, 라틴어권이 포함된다.

2. 대규모 구축을 위한 형태 주석기 성능 평가

1) 성능 평가 방법

○ 본 연구에서는 형태 주석기 성능 평가를 위해 한국어 학습자 말뭉치 형태 주석 팀의 주석/검수 과정을 거친 최종 결과물과 LCMS 내장 형태소 주석기, 현재 무료 배포 중인 형태소 주석기의 오류율을 비교하였다. 분석의 초점은 다음의 두 가지로 나눌 수 있다.

- 주어진 문자열을 한국어의 형태 단위에 맞게 적절하게 분할하였는가.
- 분할된 형태에 적절한 형태 태그를 부여하였는가.

2) 성능 평가 분석 대상

○ 꼬꼬마(Kkma)

- 서울대학교 IDS (Intelligent Data Systems) 연구실에서 수행한, 자연어 처리를 하기 위한 다양한 모듈 및 자료를 구축하기 위한 과제인 ‘꼬꼬마 프로젝트’의 일환으로 개발됨.
- 홈페이지(<http://kkma.snu.ac.kr/documents/>)를 통해 배포함.
- 한국어 형태소 분석을 위한 Python 패키지인 KoNLPy에 포함되어 있음.
- KoNLPy에 포함된 버전의 경우, 세종 형태 주석 말뭉치보다 더 세분화된 형태 주석 체계를 채택함. 구체적으로는 의존명사, 보조용언, 어미류의 분석이 세종계획 형태 주석 말뭉치보다 상세하게 분석되어 있으며, 접속부사와 일부 부사, 어미류의 태그가 세종과 상이함.

○ 코모란(Komoran)

- Shineware에서 2013년부터 개발
- 한국어 형태소 분석을 위한 파이썬 패키지인 KoNLPy에 포함되어 있음.
- 세종 형태 주석 말뭉치와 동일한 형태 주석 체계를 사용함.

○ 에트리(ETRI) 형태소분석 API

- OPEN API 형태로 제공함(https://aiopen.etri.re.kr/service_api.php).
- 언어 분석 서비스를 문어, 구어로 나누어서 제공함.

- 세종 형태 주석 말뭉치와 거의 동일한 형태 주석 체계를 사용하되, 관형사 범주만 수관형사, 성상관형사, 지시관형사로 세분화 되어 있음.

○ 카이(Khائي)

- 2016년부터 다음카카오에서 개발한, 21세기 세종계획 결과물을 기반으로 학습한 딥러닝 기반 형태소 분석기(<https://github.com/kakao/khائي>)
- 세종 형태 주석 말뭉치와 거의 동일한 형태 주석 체계를 사용하되, 일부 오형태 집단을 위한 별도의 범주가 있음.

○ 키위(KIWI)

- github 페이지(<https://github.com/bab2min/Kiwi>)를 통해 최신 버전 배포 중임.
- 세종 형태 주석 말뭉치에 기반한 형태 주석 체계를 사용하되, 웹 텍스트 관련 태그가 일부 추가되어 있음.

○ 은전한닢(a.k.a. mecab)

- 일본 교토 대학교에서 개발한 일본어 형태소 분석기 mecab를 한국어에서 사용할 수 있도록 수정/개조한 형태소 분석기
- 한국어 형태소 분석을 위한 Python 패키지인 KoNLPy에 포함되어 있음.
- KoNLPy에 포함된 버전의 경우, 세종 형태 주석 말뭉치와 거의 동일한 형태 주석 체계를 사용하나, 의존명사의 분석 정도에 차이가 있음.

○ 유태거(Utagger)

- 울산대 한국어처리연구실에서 개발한 형태소 분석기
- 홈페이지(<http://nlplab.ulsan.ac.kr/doku.php?id=start>)를 통해 분석 프로그램 및 업데이트 된 사전 파일을 배포함.
- 이 외에 KoalaNLP 등의 패키지를 통해 이용 가능함.
- 세종 형태 주석 말뭉치와 동일한 형태 주석 체계를 사용함.

3) 형태소 분석 샘플 선정

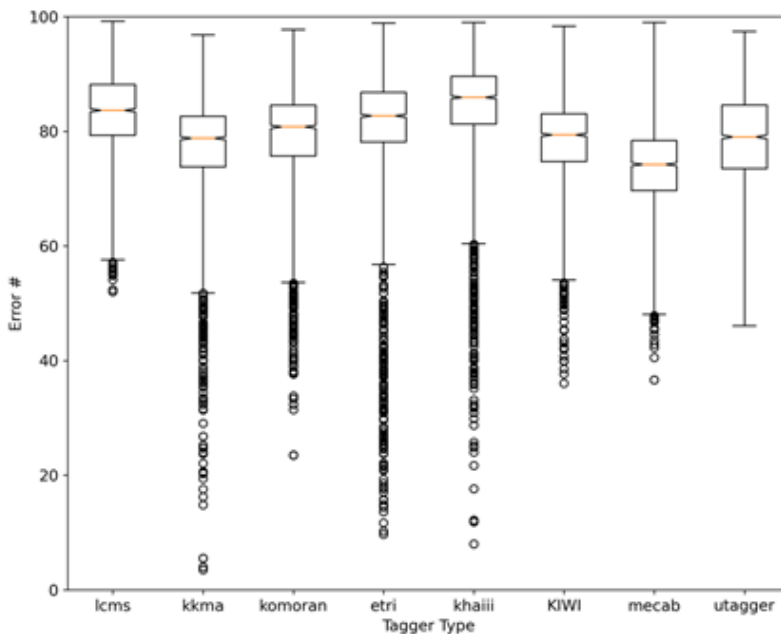
- 형태소 분석 샘플은 총 470,628어절 규모의 학습자 말뭉치로 다음과 같은 원칙에 따라 선정되었다.
- 한국어 학습자 말뭉치의 형태 주석 부분의 언어권/급수 간 균형성을 최대

한 반영함.

- 기본적으로 각 언어권 - 급수 집단의 전체 어절 수의 10%를 포함하는 것을 원칙으로 함.
- 해당 언어권 - 급수의 집단의 샘플이 1개밖에 없을 경우에는 해당 샘플을 포함시킴.

4) 분석 결과

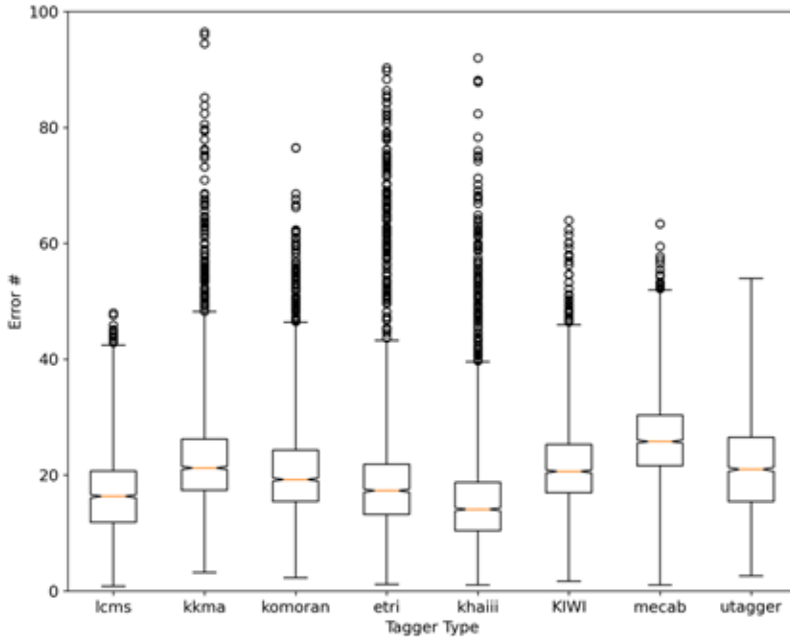
① 형태 분할 정확성



<그림 38> 형태 분할 정확성에 관한 형태소 주석기의 성능 비교

- 대부분의 태거가 정답 시퀀스에 비해 분할을 덜 한 것으로 나타났다.
- 꼬꼬마, 에트리, 카이의 경우 아웃라이어의 개수가 다른 형태소 주석기에 비해 많은 것으로 나타났다.
- 각 형태소 분석기의 분할 결과를 비교해 본 결과, 주어진 텍스트를 정답 시퀀스보다 덜 나눈 사례는 카이를 사용하였을 때 1건이 발견되었다.

나. 분석의 정확성



<그림 39> 분석의 정확성에 관한 형태소 주석기의 성능 비교

- LCMS의 경우 오류율의 중간값이 가장 낮게 형성되어 있으며, 아웃라이어의 수가 가장 적을 것을 확인할 수 있었다. 또한, 가장 많이 벗어난 아웃라이어의 오류율이 50% 내외에 머무르고 있다는 점에서 다른 형태소 주석기와 구분되는 경향을 보였다.
- 숙달도별로는 대부분의 형태소 주석기가 숙달도가 올라갈수록 오류율이 떨어지는 경향을 보이는 데 반해, LCMS는 오히려 숙달도가 올라갈수록 오류율이 올라가는 경향을 보였다.
- 학습자 언어 처리에 있어 LCMS의 한계를 확인하기 위해 숙달도 집단에서 공통적으로 아웃라이어로 분류된 샘플들, 즉 분석기가 정답을 내놓는 데 크게 실패한 파일들을 살펴본 결과, LCMS의 경우, 주어진 텍스트의 형태 분할을 비교적 잘 수행하는 것으로 나타났으나, 분할된 형태에 대한 주석이 잘못된 경우가 빈번하게 나타나는 것을 볼 수 있었다. 특히, 분할된 형태를 분석 불능(UNKNOWN)으로 처리한 경우가 가장 많이 나타났다.

- 학습자의 오철자 오류뿐만 아니라 옳은 형태를 분석 불능(UNKNOWN)으로 처리한 사례가 중급 학습자 자료에서 발견되었다. 전체 실험 말뭉치의 오류 유형 빈도를 분석한 결과, 분할된 형태를 분석 불능(UNKNOWN)으로 처리한 유형은 LCMS에서 가장 많이 발견되었다.
- (2급)
 - 십일원/NNG_십일원/UNKNOWN
 - 케어/NNG_케어/UNKNOWN
 - 태꼭/NNP_태꼭/UNKNOWN
 - 파타야/NNP_파타야/UNKNOWN
- (4급)
 - 이천십육/NR_이천십육/UNKNOWN
 - 감자기/MAG_감자기/UNKNOWN
 - 깃뺨/VV_깃뺨/UNKNOWN
 - 식업/NNG_식업/UNKNOWN
 - 심은경/NNP_심은경/UNKNOWN

5) 형태소 주석기 개선 방향

- 학습자 말뭉치 구축 지원 도구인 LCMS의 형태 주석기는 종합적으로 다른 형태 주석기에 비해 형태소 분할과 주석의 정확성이 상대적으로 높거나 평균을 웃도는 수준인 것으로 확인되었다.
- 다만, 다른 주석기와 비교해 주석의 정확성에서 분석 불능(UNKNOWN)으로 주석되는 경우가 많은 것으로 나타났다. 이러한 결과는 작업자들이 해당 주석을 일일이 확인하여 다시 붙임으로써 수정이 되어야 한다. 이에 따라 작업자들의 작업 피로도가 상승하며 작업의 정확성을 저하시키는 원인이 된다.
- 이에 따라 LCMS의 형태 주석기에서 분석 불능(UNKNOWN) 처리되는 데이터를 모아 형태 주석기를 개선하고 다양한 형태 주석기의 장점을 취해 형태 주석기의 성능을 높일 것을 구축 지원 도구 업체에 제안하였다.

V. 구축 말뭉치 검수 정교화

- 본 연구에서는 2015-2020년 검수 체계와 지침을 준용하여 작업 공정에서의 3단계 검수 체계, 개별 표본의 표본 정보 검수 체계 강화, 작업 중 생성된 오조작 데이터 검증의 세 가지 방식을 통해 구축 말뭉치의 검수를 정교화하였다.

1. 작업 공정에서의 3단계 검수 체계

- 한국어 학습자 말뭉치는 구어 전사, 문어 입력의 각 단계에서 최소 두세 차례의 전수 검수 과정을 거친다. 구어 전사와 문어 입력 자료의 경우, 작업자 간 교차 검수(1차 검수), 상위 작업자의 검수(2차 검수)를 거쳐 다음의 작업 단계인 형태 주석과 오류 주석 단계에서 두 차례의 검수에서 미처 수정하지 못한 오류들을 다시 수정하여 최종적으로 3차 검수를 하였다. 형태 주석 및 오류 주석 말뭉치의 경우 작업자 간의 교차 검수, 상위 작업자 검수를 기본으로 하되, 형태 주석과 오류 주석 간의 연계 작업을 통해 작업 단계 간 검수와 공동 연구원의 검수를 통해 3단계 검수를 하여 자료의 질적 제고를 도모하였다.

2. 개별 표본의 표본 정보 검수 체계 강화

- 문어 입력과 구어 전사 작업의 첫 단계는 표본 등록에서 시작된다. 표본 등록 작업에서는 다양한 학습자 변인과 자료 변인을 입력하게 되는데, 이 정보는 말뭉치 자료에 관한 메타 정보이자 조건별 자료 검색을 위한 색인 정보로서 중요한 기능을 한다. 따라서 이러한 정보들은 정확성과 일관성 제고의 차원에서 지속적으로 관리하며 수정·보완하는 것이 중요하다. 본 연구에서는 표본별로 부여되는 표본 정보의 검수 체계를 강화하여 자료의 내적 질을 높이고자 하였다.

3. 작업 중 생성된 오조작 데이터 검증

- 작업 중 생성된 오조작 데이터 검증은 작업 로그나 표본 정보, 주식 정보, 구축 시기 등에 관한 통계 정보를 추출한 후 전체 데이터의 구조나 작업 공정상 논리적으로 타당하지 않는 통계 정보를 확인하여 해당 표본을 집중 검수하는 것이다.
- **작업 진행 이상 표본 검증:** 작업 로그(log)를 검색하여 작업 진행상의 이상이 의심되는 표본을 검증한다. 작업 이력이 남아 있지 않거나 작업이나 검수에 소요된 시간이 30초 미만인 표본을 검수하여 자료의 무결성을 높이고자 하였다.
- **파일의 표본 정보와 파일 정보 간 이상 표본 검증:** 파일의 표본 정보와 파일 정보 간 이상 항목이 의심되는 표본을 추출하여 검증하는 것을 목적으로 한다. 표본 등록 과정에서 해당 정보의 등록이나 저장이 정상적으로 이루어지지 않은 경우로 파일 정보가 있으나 파일 표본 정보가 없는 경우, 파일 정보가 없으나 파일 표본 정보가 있는 경우가 검수 대상이 되었다.

VI. 학습자 말뭉치 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 하고 각 구축 단계별 작업자로서의 전문성을 제고하여 체계적인 말뭉치를 구축해 나가기 위한 것이다. 이 교육은 또한 자료 처리 방식의 일관성 확보를 위해서도 필요하다. 다양한 변이형을 포함한 학습자 말뭉치의 특성상 작업자의 직관에 의해 세부적인 자료 처리 방식이 달라질 우려가 있기 때문이다. 이에 따라 본 사업에서는 작업자 교육 방침에 따라 온라인/오프라인, 정기/비정기 워크숍을 통해 실무자 간에 원활한 소통이 이루어지도록 하였다.

1.1. 교육 대상

- 한국어 학습자 말뭉치 수집, 구축, 가공 실무 작업자

1.2. 교육 방법

- 수집, 입력, 전사, 형태 주석, 오류 주석 팀별 정기 워크숍(주 1회)
: 작업 관련 쟁점에 대한 토론 및 지침 교육
- 온라인 카페, 채팅 프로그램을 활용한 수시 질의응답
: 작업 과정에서 발생하는 문제점이나 궁금증을 실시간으로 해결
- 학습자 말뭉치 구축 시스템을 활용한 피드백 제공
: 미해결 주석 항목을 검토 요청 항목으로 남겨 전체 회의를 통해 해결하고 지침에 반영함

1.3. 교육 내용

- 교육 내용은 크게 지침 교육과 도구 사용 교육/실습으로 나뉜다.

<표 120> 말뭉치 구축/가공 인력 교육 내용

	지침 교육	도구 교육 및 실습
수집	<ul style="list-style-type: none"> ○ 자료 수집 과제 유형 및 수집 방법 ○ 학습자 동의서 수집 및 처리 ○ 수집 자료의 처리와 관리 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 수집 표본 등록 및 표본 관리
자료 처리 및 파일 등록	<ul style="list-style-type: none"> ○ 자료의 분류 ○ 스캔 및 음성 파일 변환 ○ 파일명 부여 체계 ○ 학습자 정보 및 파일 정보 등록(헤더 마크업) 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 스캔/음성 원본 파일 업로드 및 파일 등록, 파일명 생성 ○ 스캐너 사용 ○ 음성 파일 변환
입력	<ul style="list-style-type: none"> ○ 문어 입력 및 검수 방법, 쟁점 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 파일 입력 및 마크업, 할당 받은 작업 파일 관리 및 작업
전사	<ul style="list-style-type: none"> ○ 구어 전사 및 검수 방법, 쟁점 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 도구 내 전사 및 마크업, 할당 받은 작업 파일 관리 및 작업
형태 주석	<ul style="list-style-type: none"> ○ 형태 분석 방법 및 절차 ○ 형태 주석 체계 ○ 형태 분석 자료 검수 및 검수 관련 쟁점 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 형태 주석, 할당 받은 작업 파일 관리 및 작업
오류 주석	<ul style="list-style-type: none"> ○ 오류 식별, 판정 및 교정의 기준 ○ 오류 주석 체계 ○ 오류 분석 자료 검수 및 검수 관련 쟁점 	<ul style="list-style-type: none"> ○ 온라인 구축 도구: 오류 주석, 할당 받은 작업 파일 관리 및 작업

1.4 참여 인력

(1) 자료 스캔 및 처리

이주아(연세대학교 석사졸업)

(2) 문어 입력 작업 및 검수

이주아(연세대학교 석사졸업/문어 검수)

박현진(연세대학교 국어국문학과 석사과정/문어 검수)

강수진(언어정보학협동과정 박사과정)

강연주(한국학협동과정 박사과정)

박예슬(한국학협동과정 박사과정)

신여진(한국학협동과정 석사과정)

오혜수(한국학협동과정 석사과정)

정하은(한국학협동과정 석사과정)

(3) 구어 전사 작업 및 검수

<작업자>

김한근(연세대학교 국어국문학과 박사과정, 구어 전사 작업)

최진수(연세대학교 언어정보학협동과정 석사과정, 구어 전사 작업)

강유진(연세대학교 국어국문학과 석사과정, 구어 전사 작업)

염수원(연세대학교 국어국문학과 석사과정, 구어 전사 작업)

장수진(연세대학교 국어국문학과 박사과정, 구어 전사 작업)

천성호(연세대학교 국어국문학과 박사과정, 구어 전사 작업)

<검수자>

김미선(연세대학교 국어국문학과 박사과정, 구어 검수)

윤현애(연세대학교 국어국문학과 박사수료, 구어 작업 관리 및 검수)

김정현(연세대학교 국어국문학과 박사수료, 구어 검수)

오세원(연세대학교 국어국문학과 석사졸업, 구어 검수)

이유미(연세대학교 국어국문학과 석사졸업, 구어 전사 작업)

강재연(연세대학교 국어국문학과 석사졸업, 구어 전사 작업)

(4) 형태 주석 작업 및 검수

서지혜(연세대학교 국어국문학과 박사졸업/형태 주석 작업 관리 및 작업·검수)

배미연(연세대학교 국어국문학과 박사과정/형태 주석 작업 관리 및 작업·검수)

김동은(연세대학교 국문과 박사수료/형태 주석 작업 및 검수)

강수진(연세대학교 언정협 박사과정/형태 주석 작업 및 검수)

박나래(연세대학교 국문과 박사과정/형태 주석 작업 및 검수)

김나영(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

김재희(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

남소현(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

박은현(연세대학교 언어정보학협동과정 박사수료/형태 주석 작업 및 검수)

배승희(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

윤나영(연세대학교 국문과 박사과정/형태 주석 작업 및 검수)

이창봉(연세대학교 국문과 박사수료/형태 주석 작업 및 검수)

이혜진(연세대학교 언어정보학협동과정 박사과정/형태 주석 작업 및 검수)

이효진(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

조아라(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

최민경(연세대학교 국문과 석사과정/형태 주석 작업 및 검수)

하태현(연세대학교 국문과 박사과정/형태 주석 작업 및 검수)

(5) 오류 주석 작업 및 검수

손연정(연세대학교 국어국문학과 박사졸업/오류 주석 작업 관리, 작업 및 검수)

유소영(연세대학교 국어국문학과 박사수료/오류 주석 작업 관리, 작업 및 검수)

김진희(연세대학교 국어국문학과 박사졸업/오류 주석 작업 및 검수)

박소영(연세대학교 국어국문학과 박사과정/오류 주석 작업 및 검수)

송지혜(연세대학교 국어국문학과 박사과정/오류 주석 작업 및 검수)

오지연(연세대학교 국어국문학과 박사수료/오류 주석 작업 및 검수)

최영룡(연세대학교 국어국문학과 박사수료/오류 주석 작업 및 검수)

허희정(연세대학교 국어국문학과 박사과정/오류 주석 작업 및 검수)

2. 한국어 학습자 말뭉치 아카데미 개최

- 한국어 학습자 말뭉치 아카데미는 학습자 말뭉치의 폭넓은 활용을 위해 학습자 말뭉치를 홍보하고 사용자 교육을 하는 데에 목적이 있다.
- 본 연구에서는 총 5회의 학습자 말뭉치 아카데미를 개최하였다. 2021년 아카데미는 각 회차별로 사용자 또는 프로그램을 차별화하여 기초 과정에 서 심화 과정까지 다양하게 다루었다. 또한 강의와 실습형의 프로그램 외에도 각 분야의 전문가를 패널로 한 집담회 형식의 아카데미를 개최하여 학습자 말뭉치의 활용 범위와 나아가야 할 방향에 대해서도 논의를 이끌어 냈다. 개최 방법에 있어서도 실시간 줌(Zoom)과 함께 유튜브 생중계, 비대면과 대면의 하이브리드 방식 등으로 두 가지 형식을 병행하여 학습자의 여건에 따라 참가 플랫폼을 선택할 수 있도록 하였다.

<표 121> 학습자 말뭉치 활용 아카데미 개최

일시	프로그램	강사	참가자
1차 (7.20)	○ 한국어 학습자 말뭉치 구축과 활용 현황 ○ 한국어 학습자 말뭉치 활용의 실제: 한국어 학습자의 조사 사용 양상 분석	장석배 (공동연구원)	Zoom 30명
2차 (8.24)	○ 한국어 학습자 말뭉치 구축과 활용 현황 ○ 민간 분야에서의 학습자 말뭉치 활용 가능성과 구축 방향 ○ 청중과의 토론 및 질의응답	이기황 (바이브 컴퍼니), 곽용진 (주) 이르테크), 박전규(ETRI 복합지능연구실), 이형구(엘로우 크리에이티브)	Zoom 100명 Youtube 55명
3차(11.12)	○ 한국어 학습자 말뭉치 구축과 활용 현황 ○ 한국어 학습자 말뭉치 기반 연구를 위한 자료 처리의 기초 - 텍스트 에디터를 활용	남신혜 (연세대)	Zoom 50명 Youtube 83명

	한 전처리 작업과 엑셀을 활용한 기술 통계		
4차(11.17)	한국어 학습자 말뭉치 구축과 활용 현황 학습자 말뭉치를 활용한 한국어 어휘 습득 연구의 실제	이승연 (삼육대)	Zoom 50명 Youtube 78명
5차 (12.4)	한국어 학습자 말뭉치 구축과 활용 현황 한국어 학습자 오류 주석 말뭉치의 주석 체계 이해와 활용의 실제	유소영 (구축실무연구원)	Zoom 55명 대면 20명

3. 학술대회 발표

- 학술대회 발표는 학습자 말뭉치 구축 및 활용에 관한 쟁점을 논의하는 학술 교류의 장을 마련하고 학습자 말뭉치 구축 성과를 확산시키는 것을 목표로 한다. 본 연구에서는 다음과 같이 학술대회 발표와 학술지 논문 게재를 통해 한국어 학습자 말뭉치 구축의 쟁점을 논의하고 활용 모형을 제시하였다.

1) 학술대회 발표

- 일시: 2021년 11월 25일(목) 한국어문화교육학회
- 제목: 국외 학습자 말뭉치 사례 분석 - 구축 및 활용을 중심으로

2) 학술지 논문 게재

- 한국어 학습자 오류 사전 개발을 위한 표제어 선정과 배열에 관한 연구 (사전학)
- 한국어 학습자 오류 사전 개발 모형 연구-사례 분석과 사전의 미시구조설계를 중심으로(언어와 문화)

4. 말뭉치 소개·활용 자료 제작, 한국어교수학습센터 게재 및 아카데미 배포

- 학습자 말뭉치는 2015-2020년의 1차 중장기 계획의 성공적 수행으로 국가 주도의 공공 언어 자원으로서 학습자 말뭉치라는 성과물을 산출해 내면서 한국어 교육 연구의 지평을 넓혀 왔다. 2021-2025년 2차 중장기 계획 수립을 위한 본 연구에서는 한국어 학습자 말뭉치를 보다 손쉽게 이용할 수 있도록 하기 위해 다음의 구성으로 ‘학습자 말뭉치 활용 매뉴얼’을 제작하였다.

<표 123> ‘학습자 말뭉치 활용 매뉴얼’의 구성

구성	세부 내용
I. 학습자 말뭉치 알아보기	1. 학습자 말뭉치의 정의 2. 학습자 말뭉치의 개념
II. 학습자 말뭉치 나눔터 활용하기	1. 학습자 말뭉치 나눔터 소개 2. 학습자 말뭉치 나눔터 주요 메뉴 3. 통계 자료 활용하기 4. 검색하기 1) 원시 말뭉치 2) 형태 주석 말뭉치 3) 오류 주석 말뭉치 4) 말뭉치의 상세 검색 5) 검색 결과 내려받기 5. 말뭉치 신청하기

Ⅶ. 결론

1. 연구 요약

본 연구는 한국어 교육 및 연구, 민간 분야에서 학습자 말뭉치를 광범위하게 활용하는 것을 목적으로, 1차 중장기 계획에 따른 <2015-2020 한국어 학습자 말뭉치 연구 및 구축> 사업에 이어 2021-2025년에 진행될 2차 중장기 계획을 수립하고 2021년 목표에 따라 실제 말뭉치를 구축하였다. 이에 따른 주요 과업과 연구의 성과는 다음과 같다.

○ 학습자 말뭉치 중장기 계획 수립

학습자 말뭉치 중장기 계획은 학습자 말뭉치의 구축과 활용에 영향을 미치는 언어 자원 구축 관련 정책과 법·제도에 대한 분석, 학계와 민간 분야를 포함한 다양한 사용자 집단의 요구분석, 선진 사례 분석, 기구축 말뭉치 분석 등의 기초 연구 결과를 바탕으로 하여 2021년에서 2025년까지 총 5개년 계획으로 수립되었다. 기초 연구에서는 공공언어 자원으로서 한국어 학습자 말뭉치의 위상을 점검해 보고, 공공데이터법, 저작권법, 개인정보 보호법, IRB 규정 등의 관련 법령과 규정을 기반으로 말뭉치 구축과 활용에 관한 세부 사항들을 점검하였다. 또한 148명의 사용자를 대상으로 설문조사를 하여 기구축 말뭉치와 학습자 말뭉치 나뉘터 이용에 관한 의견을 수렴하였다. 그 외에도 기구축 말뭉치의 성과 분석, 2백만 어절 이상의 규모, 구글 학술 검색 인용 횟수 100 이상인 22건의 국외 학습자 말뭉치 구축 사례에 대한 심층 분석, 자연언어 처리·인공지능·에듀테크 분야 종사자와의 집담회, 3회의 한국어 교육 전문가 자문을 통해 기구축 말뭉치의 보완 방향을 모색하였다.

그 결과 한국어 학습자 말뭉치는 기초 연구 단계(2021년)와 2단계의 본격 구축(2022-2025년)으로 총 5년간의 사업을 통해 2015년부터 2025년까지 누적 규모를 기준으로 원시 말뭉치 1,000만 어절(문어 600만, 구어 400만 어절), 형태 주석 말뭉치 1,000만 어절(문어 600만, 구어 400만 어절), 오류 주석 말뭉치 500만 어절(문어 300만 어절, 구어 200만 어절) 규모의 말뭉치를 구축하는 것을 목표로 설정하였다. 2021-2025년 중장기 계획에서는 말뭉치의 양적 확대와 함께 대상별, 언어권별 균형성을 보완하는 것을 목표로 하며, 말뭉치의 활용도 제고를 위하여 참조 말뭉치 구축, 연구자가 개별적으로 구축한 말뭉

치 또는 타 기관 제공 말뭉치와의 통합 구축을 병행할 것을 제안하였다. 아울러 대규모 말뭉치로의 양적 확대를 위해 기존의 수집 방식과 함께 온라인을 기반으로 한 수집 시스템을 구축하는 방안을 제시하였다. 균형성 확보를 위한 집중 구축 대상은 학습자 대상별로는 학문 목적 학습자와 이주민, 국외 학습자, 언어권별로는 일본어, 영어, 베트남어, 태국어, 러시아어, 스페인어권이며, 그 외에도 기획 과제를 통해 장르나 주제를 고려하도록 한다.

○ 한국어 학습자 말뭉치 수집 및 구축·가공

2021년 한국어 학습자 말뭉치는 2015-2020년에 상대적으로 부족한 장르와 주제의 자료를 집중적으로 수집할 수 있도록 과제를 기획하여 수집하였다. 문어는 생활문(초급)과 논설문(중·고급) 쓰기, 구어는 내러티브 과제를 설계하였으며, 국내 대학(원)의 학문 목적 학습자와 중·고급 학습자를 대상으로 한 중점 수집과 그 밖의 학습자를 대상으로 한 수집이 함께 진행되었다.

말뭉치 구축은 원시 말뭉치 831,142어절(문어 419,371어절, 구어 411,771어절), 형태 주석 말뭉치 200,981어절(문어 100,781어절, 구어 100,200어절), 오류 주석 151,906어절(문어 104,314어절, 구어 47,592어절) 규모의 말뭉치가 새롭게 구축되었다. 그 결과 2015-2021년에 구축한 전체 말뭉치의 규모는 원시 말뭉치 5,220,564어절(문어 3,697,952어절, 구어 1,522,612어절), 형태 주석 말뭉치 3,704,586어절(문어 2,602,914어절, 구어 1,101,672어절), 오류 주석 말뭉치 1,153,848어절(문어 590,548어절, 구어 563,300어절)이 되었다.

○ 말뭉치 구축 지원 도구 검증

한국어 학습자 말뭉치는 표본 등록에서 말뭉치 주석 가공까지 전체 작업 공정을 관리하고 수행할 수 있는 말뭉치 구축 지원 도구를 활용하여 구축해 오고 있다. 본 연구에서는 작업자들에게 안정적인 구축 환경을 제공하기 위하여 구축 실무 연구원들을 중심으로 성능 피드백팀을 구성하여 지속적인 모니터링을 통해 학습자 말뭉치 구축 지원 도구의 성능 개선과 안정화를 위한 피드백을 제공하였다. 또한 향후 수행될 대규모 구축 사업의 효율성 제고를 위하여 지원 도구에 내장된 형태 주석기와 꼬꼬마(Kkma), 코모란(Komorán), 에트리(ETRI) 형태소 분석 API, 카이(Khائي), 키위(KIWI), 은전한닢(a.k.a. mecab), 유태거(Utagger)와의 성능 비교를 통해 현재 사용 중인 형태소 주석기의 성능을 객관적으로 평가하고 개선 사항을 도출하여 세부적인 기능을 고도화하기 위한 방향성을 제시하였다.

○ 구축 말뭉치 검수 정교화

구축 말뭉치 검증은 구축된 말뭉치 자료의 질적 제고를 위한 것으로, 문어 입력과 구어 전사, 형태 주석, 오류 주석의 각 작업 단계별로 3단계 작업 및 검수 체제에 따라 작업 공정을 진행하였다. 그 외에도 시스템 기반의 데이터 검증을 통한 오조작 데이터와 이상 데이터 검수를 상호보완적으로 적용하였다. 또한 전체 표본 목록 대조를 통한 중복 표본 추출, 데이터 통계의 정확성을 높이기 위하여 표본 정보를 전 사업 기간에 걸쳐 검수하였다.

○ 학습자 말뭉치 관련 교육 및 홍보

한국어 학습자 말뭉치 관련 교육은 구축 실무 작업자와 사용자를 대상으로 하여 이루어졌다. 구축 실무자 작업자 교육은 효율적이고 체계적인 말뭉치 구축을 위해 2015년 이후 지속해 오고 있는 것으로, 지침 교육과 도구 사용 교육 외에도 구축 과정에서 발생하는 다양한 문제를 해결하기 위한 즉각적 피드백 시스템을 운영하고 정기 워크숍을 통해 말뭉치 구축에 관한 여러 가지 쟁점들을 공유하였다. 그럼으로써 작업자의 전문성을 제고하고 결속력을 강화할 수 있었다.

사용자를 대상으로 한 교육은 학습자 말뭉치 아카데미를 통해 이루어졌다. 2021년에는 총 5회의 학습자 말뭉치 아카데미를 개최하였다. 각 회차별로 사용자 또는 프로그램을 차별화하여 기초 과정에서 심화 과정까지 다양한 내용을 다루었다. 또한 강의와 실습형의 프로그램 외에도 각 분야의 전문가를 패널로 한 집담회 형식의 아카데미를 새롭게 시도하여 학습자 말뭉치의 활용 범위와 나아가야 할 방향에 대해서도 논의를 이끌어 냈다. 개최 방법에 있어서도 실시간 줌(Zoom)과 동시에 유튜브 생중계, 비대면과 대면의 하이브리드 방식 등으로 두 가지 형식을 병행하여 학습자의 여건에 따라 참가 플랫폼을 선택할 수 있도록 하여 참여 기회를 넓히고자 하였다.

2. 연구의 의의 및 기대 효과

한국어 학습자 말뭉치는 외국어 또는 제2 언어로서 한국어를 학습하는 비모어 화자가 산출한 언어 자료를 수집하여 구축한 국가 주도의 공공 언어 자원으로서 과학적이고 선진화된 한국어 교육 기반을 마련하기 위한 기초 자료이자 빅데이터로서 인공지능 기술 개발을 위한 원천 자료로 활용도가 높고 유용한 자료이다. <한국어 학습자 말뭉치 연구 및 구축> 사업의 의의 및 기대 효과를 구체적으로 설명하면 다음과 같다.

○ 국가 주도의 대규모 공공언어 자원으로서의 한국어 학습자 말뭉치 구축

한국어 학습자 말뭉치 연구는 2010년 기초 연구에서 시작되어 2015년 본격적인 구축을 위한 1차 중장기 계획 수립 이후 현재까지 이어져 오고 있다. 본 연구는 2020년까지의 1차 중장기 계획에 따른 목표를 달성한 후 이어진 후속 사업으로 학습자 말뭉치에 대한 사용자의 기대와 활용도를 제고하여 2025년까지 1,000만 어절 규모의 대규모 균형 말뭉치로의 확장 구축을 제안하였다. 국외 학습자 말뭉치의 경우, 2020년을 기준으로 CLC(Cambridge Learner Corpus), The Hong Kong University of Science & Technology(HKUST), The Longman Learners' Corpus를 제외하고는 1,000만 어절을 밑도는 규모이거나 100만 어절 미만의 소규모 말뭉치가 대부분이다. 대규모 말뭉치를 구축한 주체도 대규모 공인 평가 시행 기관이나 대학 기관으로 한국어 학습자 말뭉치처럼 국가 주도인 경우는 드물다. 이러한 점에서 국가 주도의 대규모 공공언어 자원으로서의 한국어 학습자 말뭉치 구축을 위한 본 연구의 의미가 크다고 하겠다.

○ 한국어 학습자 균형 말뭉치 구축

학습자 말뭉치는 한국어 학습자의 언어를 관찰하기 위한 자료로 대표성과 함께 균형성이 중요하다. 모어 화자 자료와 달리 학습자의 제1언어, 숙달도 단계, 학습 환경, 경험 등의 다양한 변인이 존재하는 만큼 이를 얼마나 균형적으로 안배하느냐가 자료의 대표성에도 영향을 미치게 된다. 본 연구에서 제안한 중장기 계획에서는 2015-2020년 구축 말뭉치의 성과를 점검하고 다양한 분야의 사용자와 전문가의 의견을 수렴하여 자료의 균형성을 제고하기 위한 계획을 수립하였다. 이에 따르면 다양한 변인 중 특히 대상별, 언어권별, 수준별, 장르·주제별 변인을 중심으로 하여 말뭉치의 균형성을 확보하되, 언

어권의 경우 학습자의 현실적인 분포를 고려하여 현재 가장 많은 비중을 차지하는 중국어권을 포함하여 7개 언어권의 자료가 균형적으로 구축되도록 설계하였다. 이러한 과정에서 ‘균형성’이라는 개념이 적어도 학습자 말뭉치에서는 산술적인 의미의 균형성이 아닌, 자료의 활용도 측면에서 학습자의 실제 분포를 고려한 귀납적인 결과로서의 현실적인 균형성임을 확인할 수 있었다. 본 연구는 이처럼 한국어 학습자 말뭉치의 다양한 특성을 바탕으로 하여 그에 최적화된 한국어 학습자 균형 말뭉치를 구축하기 위한 초석으로서 의미가 있다고 하겠다.

○ 한국어 교육 이론의 체계화 및 교육 자료 구축의 기반 조성

한국어 학습자 말뭉치는 학습자의 언어 발달과 습득의 지표가 되는 중간언어 특성을 실증적으로 보여 주며, 이는 연구와 함께 교육 자료와 교수법, 평가 도구 등을 개발하기 위한 기초 자료라는 점에서 가치가 크다. 한국어 교육 학계와 교수 현장에서는 이미 오래전부터 이와 같은 학습자 말뭉치의 의의와 중요성, 필요성에 대해 상당한 공감대가 형성되어 왔다. 그러나 이러한 연구를 위한 자료에 대한 접근이 쉽지 않고, 연구 방법에 대한 심리적인 진입장벽이 높아 매우 제한된 범위에서 연구가 이루어지는 경향이 있었다. 또한 개인이 구축한 소규모 자료의 경우 대표성과 균형을 갖추기가 어려워 연구 결과를 일반화하기 어렵다는 한계가 있었다. 한국어 학습자 말뭉치는 이러한 문제를 해소함으로써 말뭉치를 활용한 연구를 활성화하여 한국어 교육 이론을 체계화하고 교육 자료를 구축해 나갈 수 있는 기반을 제공하였다. 이러한 관심은 그간 한국어 학습자 말뭉치 아카데미에 대한 관심과 호응, 학습자 말뭉치 나눔터(<https://kcorpus.korean.go.kr>)의 접속자 수와 자료 배포 요청 수의 지속적인 증가를 통해서 알 수 있다. 본 연구에서 제안한 중장기 계획에 따라 2025년까지 구축될 1,000만 어절 규모의 한국어 학습자 균형 말뭉치는 이를 더욱 가속화할 수 있을 것으로 기대된다.

○ 국어 빅데이터로서 민간 분야에서의 활용성 제고

4차 산업혁명 시대의 도래와 함께 최근 빅데이터가 다양한 분야에서 광범위하게 활용되면서 중요한 국가 자원으로서 주목받고 있다. 이에 따라 국립국어원에서는 2018년부터 ‘국어 빅데이터’ 구축 사업에 착수하여 현재 본격 구축 작업이 진행되고 있다. 한국어 학습자 말뭉치는 한국어 교육 연구를 위한 자료라는 독립된 위상과 지위를 가짐과 동시에 한국어 비모어 화자에 의

해 산출된 국어 빅데이터의 일부로서 자연언어 처리, 인공지능, 에듀테크 기술 개발의 원천 자료로 활용될 수 있다.

○ 한국어의 세계화 및 국제 경쟁력 강화

한국 언어·문화에 대한 세계인의 관심이 나날이 커지고 있다. 이는 한국어 학습자의 양적인 증가 외에도 점점 확대되고 있는 지역 분포, 다양해지는 학습 목적 등을 통해 어렵지 않게 확인할 수 있다. 한국어 학습자 말뭉치는 이처럼 서로 다른 환경, 다양한 목적으로 한국어를 배우고자 하는 한국어 학습자들의 요구에 맞는 체계적이고 질 높은 교육을 제공하기 위한 기초 자료로 활용할 수 있다. 그럼으로써 일차적으로는 한국 언어·문화를 세계화하고, 더 나아가 한국 언어·문화를 널리 알리고 정치·경제·사회의 각 분야에서 폭넓게 활동할 수 있는 국제 인력을 양성함으로써 해서 한국의 국제 경쟁력 강화에 기여할 수 있을 것이다.

3. 보고서 활용 방안

본 연구는 한국어 학습자 말뭉치 중장기 계획 수립을 위한 기초 연구와 실제 말뭉치 구축을 핵심 과업으로 하였다. 보고서에는 과업 수행을 위한 방법과 절차, 관련 쟁점과 결과가 상세하게 기술되어 있으며 말뭉치 구축의 단계별 지침이 부록으로 첨부되어 있다. 이는 다양한 목적의 사용자들에게 한국어 학습자 말뭉치 구축에 관한 이론과 적용, 그리고 말뭉치의 활용을 위한 지침으로 활용될 수 있다.

○ 한국어 학습자 말뭉치 구축에 관한 이론적 지침

한국어 학습자 말뭉치는 한국어를 모국어로 하지 않는 외국인 한국어 학습자의 자료를 수집하여 구축한 자료로 비정형의 발화를 포함하고 있으며 매우 다양한 변인이 전제된다는 점에서 특수 말뭉치로 분류할 수 있다. 따라서 말뭉치를 구축함에 있어 범용 말뭉치 구축에 관한 이론적 체계를 따르되 학습자 자료의 특성에 따른 다양한 쟁점에 대한 세심한 고려가 필요하다. 본 연구에서는 학습자 말뭉치 구축에 관한 이론적 논의를 찾아보기 힘든 상황에서 많은 시행착오를 거치며 2015년에서 2021년 7년간의 연구를 수행해 오면서 학습자 말뭉치 설계와 구축, 가공에 관한 쟁점과 해결 방안을 체계적으로 기

술하였다. 그러한 점에서 본 보고서는 지금까지 체계화되지 못하였던 한국어 학습자 말뭉치 구축에 관한 실제적인 모형이자 이론적 지침으로 활용될 수 있다고 하겠다.

○ 한국어 학습자 말뭉치 구축을 위한 실용적 지침

본 보고서는 자료 수집 및 처리, 입력과 전사, 형태 주석, 오류 주석 지침이 포함되어 있다. 이들 지침은 자료의 호환성을 위해 <21세기 세종 한국어 균형 말뭉치>의 구축 지침을 기반으로 하되 비모어 화자 자료인 학습자 말뭉치의 특성을 반영하여 초안을 작성한 후 지난 7년간의 구축 과정에서 대두되는 수많은 쟁점들을 실례와 함께 기술하여 정교화한 것이다. 따라서 학습자 말뭉치 구축 이론과 실재를 모두 포괄하는 학술 자료로는 물론 향후 학습자 말뭉치를 구축하고자 하는 기관이나 연구자들에게 실용적인 지침으로 활용될 수 있다.

4. 정책 제언

○ 안정적인 중장기 계획 수행을 위한 예산 지원

본 연구에서는 2021년부터 2025년까지 총 5년간의 중장기 계획을 수립하고 그 계획에 따라 원시 말뭉치 1,000만 어절, 형태 주석 말뭉치 1,000만 어절, 오류 주석 말뭉치 500만 어절 규모의 말뭉치 구축을 제안하였다. 아울러 말뭉치의 활용도를 높이기 위하여 형태 주석 말뭉치와 오류 주석 말뭉치의 비중을 확대하고, 구어 말뭉치의 규모를 확대하는 방안을 세부적인 안으로 제안하였다. 이는 1차 중장기 계획에서 2015-2020년까지 5년간 구축한 원시 말뭉치 440만 어절, 형태 주석 말뭉치 350만 어절, 오류 주석 말뭉치 100만 어절을 훨씬 넘어서는 규모로 2025년까지 안정적으로 목표를 달성해 나가기 위해서는 물가 상승분을 반영한 예산 지원이 필요하다.

○ 참조 말뭉치 구축을 통한 활용성 제고

한국어 학습자 말뭉치는 비모어 화자의 자료로 비정형성을 특징으로 한다. 즉, 모어 화자들에서는 발견되지 않는 학습자들만의 독특한 언어 사용 체계인 중간언어를 포함한 자료이다. 이에 따라 학습자 말뭉치는 학습자의 이러한 언어 사용 특성을 통해 언어 발달과 습득 과정을 관찰하는 데에 활용되어 왔다. 참조 말뭉

치는 이러한 특성을 더욱 변별적으로 포착하여 분석할 수 있는 비교 자료로, 한국어 모어 화자들의 자료가 해당된다. 한국어 학습자와 동일한 과제를 사용하여 한국어 모어 화자가 산출한 자료를 수집하여 참조 말뭉치로 구축하면 두 집단의 언어 사용 양상을 보다 체계적으로 살필 수 있게 된다. 이는 한국어 교육 연구뿐만 아니라 민간 분야에서의 활용에서도 유용한 자료가 될 수 있다는 점에서 중장기 계획에 새롭게 더해지는 과업으로 실행해 볼 만한 가치가 있다고 판단된다.

○ 한국어 학습자 말뭉치를 활용한 교수·학습 자료 개발 사업으로의 연계

한국어 학습자 말뭉치는 2021년까지 520만 어절의 원시 말뭉치와 370만 어절의 형태 주석 말뭉치, 115만 어절의 오류 주석 말뭉치가 구축되었다. 구축된 말뭉치는 한국어 학습자 말뭉치 나눔터를 통해 배포되어 한국어 교육 연구와 교수·학습에 활용되고 있으며, 이는 성과물의 활용이라는 점에서 의미가 있다. 그러나 한국어 교수·학습 현장의 교원과 학습자에게 파급력을 가지기 위해서는 이러한 성과가 축적되고 통합되기까지 상당한 시간이 필요해 보인다. 이러한 시간 차이를 극복하고 학습자 말뭉치의 활용 범위를 교원, 학습자까지 넓히기 위해 말뭉치 자료를 활용한 국가 주도의 교수·학습 자료 개발 사업으로 연계할 필요가 있다. 이는 한국어 교육의 과학화, 선진화라는 목표를 촉진하는 데에도 기여할 것이다.

○ 효율적인 자료 수집을 위한 정부 부처, 유관 기관과의 공조 체계 강화

한국어 학습자 말뭉치 구축 사업의 성패는 자료 수집이 얼마나 효율적으로 이루어지는가와 직결된다고 할 수 있다. 자료 수집은 말뭉치 구축의 첫 단계로, 2015-2021년 사업에서는 한국어 교육 학계를 중심으로 한 수집 네트워크를 기반으로 수집을 진행하였다. 그 외에 세종학당재단과의 업무 협조를 통해 국외의 세종학당 학습자 자료를 수집하기도 하였다. 2차 중장기 계획에서 집중적으로 구축하고자 하는 국외 학습자, 이주민 자료를 성공적으로 수집하기 위해서는 한국어능력시험을 주관하고 있는 국립국제교육원, 전 세계의 세종학당 학습자 네트워크를 보유하고 있으며, 한국어 숙달도 평가를 개발 중인 세종학당재단, 한국학을 개설하고 있는 국외 대학에 교원을 파견하고 있는 한국국제교류재단, 이주민을 대상으로 법무부의 사회통합프로그램을 운영하고 있는 기관 등의 관계 기관의 협조가 절실히 요구된다.

○ 연구자 또는 타 기관 구축 말뭉치 통합을 위한 기반 마련

학습자 말뭉치를 활용한 연구는 오랫동안 많은 연구자들이 주목해 온 만큼 개인 연구자들이 구축한 자료가 많이 축적되어 있을 것으로 추정된다. 또한 실제로 연구 목적에 따라 한국어 학습자 말뭉치 나눔터를 통해 제공받은 말뭉치와 함께 다른 경로를 통해 구축된 말뭉치를 통합하여 사용하는 경우도 적지 않은 것으로 보고되고 있다. 이 경우 불가피하게 자료 간의 호환성에 문제가 생기기 마련이며, 이로 인해 그간 한국어 학습자 말뭉치 구축 도구 배포와 교육에 대한 요구도 있어 왔다. 이를 적극적으로 수용하여 호환성의 문제를 해소하도록 하고, 자율적인 의사에 따라 연구에 활용한 자료를 학습자 말뭉치에 제공하도록 하여 서브 말뭉치로 구축하는 방안을 고려해 볼 필요가 있다. 이는 자료의 다양화와 함께 균형성 확대, 규모 확대의 측면에서 긍정적으로 평가된다.

○ 말뭉치 자료의 질적 제고를 위한 감수단 운영

한국어 학습자 말뭉치는 연구진의 작업 공정에 따라 입력, 전사, 형태 주석, 오류 주석의 각 작업 단계별로 3단계 작업 및 검수 체계에 따라 작업을 진행하고 있다. 이러한 작업 공정과 함께 작업자의 전문성 제고를 위한 교육을 통해 구축 결과물의 일관성과 정확성을 확보하기 위한 노력을 지속하고 있으나, 다수의 작업자가 투입되어 대규모 자료를 구축하는 작업의 특성상 오류가 발생하는 것을 피하기가 어렵다. 이에 따라 구축팀 내부 또는 발주기관을 중심으로 한 감수단을 운영하여 현재의 작업 공정 외에 최종 검증 단계를 둬으로써 말뭉치 구축 결과물의 정확성을 한층 더 높일 수 있다.

참고 자료

1. 논문 및 저서

- 강현화 외(2015), 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업 보고서, 국립국어원.
- 강현화 외(2016), 2016년 한국어 학습자 말뭉치 연구 및 구축 사업 보고서, 국립국어원.
- 강현화 외(2017), 2017년 한국어 학습자 말뭉치 연구 및 구축 사업 보고서, 국립국어원.
- 한송화 외(2018), 2018년 한국어 학습자 말뭉치 연구 및 구축 사업 보고서, 국립국어원.
- 한송화 외(2019), 2019년 한국어 학습자 말뭉치 연구 및 구축 사업 보고서, 국립국어원.
- 한송화 외(2020), 2019-2020년 한국어 학습자 말뭉치 연구 및 구축 사업 보고서, 국립국어원.
- 강현화(2010) 한국어 학습자 사전 표제어 선정을 위한 자료 구축 및 선정 방법에 관한 연구, 한국사전학 16, 한국사전학회.
- 강현화(2011) 한국어 학습자 말뭉치의 자료 구축 방안 대한 기초 연구, 한국사전학 17, 한국사전학회.
- 강현화(2017), 중국인 한국어 학습자 말뭉치에 나타난 중간언어 분석 연구, 언어사실과 관점 41, 연세대학교 언어정보연구원, 5-47.
- 강현화(2017), 학습자 말뭉치의 구축과 활용, 소통.
- 강현화·조민정(2003), 스페인어권 한국어 학습자의 어미, 조사 및 시상, 사동 범주의 오류 분석, 한국어 교육 14(2), 국제한국어 교육학회.
- 고석주(2002), 학습자 말뭉치에서 조사 오류의 특징, 외국어로서의 한국어 교육 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.
- 권기양(2006), KFL 학습자의 오류에 대하여: 중국인 학습자 중심으로, 언어과학 13(3), 한국언어과학회.
- 권영일, 김진철, 김성현. (2017). 빅데이터 정책의 현황과 미래. 한국지능정보시스템학회 학술대회논문집, 90-91.
- 김경화(2013), 고급단계 한국어 학습자의 오류연구, 중국조선어문 188, 길림성민족사부위원회.

- 김동국, 이혁. (2015). 빅데이터 기반의 개인정보 비식별화 동향. 인터넷정보학회지, 16(2), 15-22.
- 김미경·강현화(2017), 중·고급 중국어권 한국어 학습자의 조사 '가'와 '는' 선택 요인 연구, 외국어로서의 한국어 교육 47, 25-52.
- 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), 국제한국어 교육학회.
- 김미옥·정희정(2003), 한국어 학습자 작문에 나타난 어휘 오류 분석, 제3회 한국어 교육 국제 워크숍 발표 요지, 연세대 언어정보연구원 외국어로서의 한국어 교육연구센터, 102-135쪽.
- 김배현. (2014). 해외 주요국가의 빅데이터 정책 비교 분석. 한국콘텐츠학회지, 12(1), 38-40.
- 김병일, 신현철, 안창원. (2017). 빅데이터 분석과 데이터 마이닝을 위한 저작권 제한. 계간 저작권, 30(1): 29-61
- 김병철. (2014). 개인정보보호법에 기반한 빅데이터 활용 방안 연구. 디지털융복합연구, 12(12), 87-92.
- 김선남, 이환수. (2014). 빅데이터 시대의 개인정보보호 방안 : 빅데이터 가이드라인을 중심으로. 2014년 한국경영정보학회 추계학술대회, 343-348.
- 김아름(2014), 한국어 학습자의 문법 및 화용오류에 대한 인식, 새국어교육 100, 한국국어교육학회.
- 김유미(2002), 학습자 말뭉치를 이용한 한국어 학습자 오류 분석 연구, 외국어로서의 한국어 교육 27, 연세대학교 한국어학당.
- 김유정(2005), 한국어 학습자 말뭉치 오류 분석의 기준, 한국어 교육 16(1), 국제한국어 교육학회.
- 김일환(2016), 한국어 학습자 말뭉치의 주석 과정과 활용 방법, 국제한국어 교육학회 춘계학술발표논문집, 국제한국어 교육학회.
- 김정숙(2002), 영어권 한국어 학습자의 조사 사용 오류 분석과 교육 방법, 한국어 교육 13(1), 국제한국어 교육학회.
- 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), 한국어 교육에서의 중간언어와 오류 분석, 한국어 교육 14(1), 국제한국어 교육학회.
- 김정현. (2016). 유럽연합의 빅데이터 관련 법제에 관한 고찰. 법제처.
- 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, 시학과언어학회.

김한샘, 배미연(2017), 학문 목적 학습자의 객관화 전략 사용 양상 연구 - 중국인 학습자의 학술 텍스트를 중심으로, 언어사실과 관점 41, 연세대학교 언어정보연구원, 5-47.

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

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

남영준. (1997). 디지털자료에 대한 저작권적 해석에 관한 연구 - 코퍼스를 중심으로. 정보관리학회지, 14(1), 161-181.

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

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

대한상공회의소. (2018), 개인정보보호제도 개선방안 연구 보고서.

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

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

서상규 외(2010), 한국어 학습자 말뭉치 구축 설계, 국립국어원.

서상규, 유현경, 남윤진(2002), 한국어 학습자 말뭉치와 한국어 교육, 한국어 교육 13(1), 국제한국어 교육학회.

성욱준. (2016). 공공부문 빅데이터 정책 활성화 연구. 한국정책학회보, 25(2), 125-150.

성지은, 박기량. (2014). 빅데이터를 활용한 정책 사례 분석과 시사점. 과학기술정책, 24(2), 94-106.

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

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

양관석. (2019), 인공지능의 빅데이터 활용을 위한 법적 연구 : 저작물과 개인정보를 포함한 빅데이터를 중심으로, 단국대학교 박사학위 논문.

엄수현, 이인경, 이우기. (2018). 빅데이터 기반 개인정보 비식별화 동향. 정보화연구 (구 정보기술아키텍처연구), 15(4), 545-552.

유석훈(2001), 외국어로서의 한국어 학습자 말뭉치 구축의 필요성과 자료 분석, 한국어 교육 12(1), 국제한국어 교육학회.

유영성, 빈미영, 옥진아, 최조순, 천영석. (2014). 지자체의 공공 빅데이터 정책 사례연

- 구. 정책연구, 1-50.
- 이강민, 김성보, 김응모. (2018). 빅데이터와 저작권. 한국정보기술학회
종합학술발표논문집, 566-569.
- 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 재고, 이중언어학 31, 이중언어학
회.
- 이승연(2007), 한국어 교육을 위한 한국어 학습자 말뭉치의 구축과 활용 연구, 고려대
박사학위논문.
- 이승연(2007), 한국어 학습자 오류 판정 및 수정 기준 연구-교사, 비교사 집단간 오류
판별 비교 실험을 바탕으로, 이중언어학 33, 이중언어학회.
- 이유림, 김영주(2013), 교사의 피드백 방법이 한국어 학습자의 작문 내 어휘 오류 감
소에 미치는 영향, 외국어로서의 한국어 교육 39, 연세대학교 언어연구교육원
한국어학당.
- 이은서(2017), 중국어권 학습자의 접사 사용 연구, 연세대학교 대학원 석사학위 논문.
- 이정희(2002), 한국어 오류 판정과 분류 방법에 관한 연구, 한국어 교육 13(1), 국제한
국어 교육학회.
- 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,
국제한국어 교육학회.
- 이진태. (2013). 빅데이터 활성화와 저작권 문제- 하둡(Hadoop)을 중심으로 -. 계간
저작권, 26(2): 136-173
- 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.
- 정교일. (2012). 빅데이터와 정보보안. 한국정보기술학회.
- 조진만, 김석현, 최대선, 진승헌. (2013). 자동 구축된 코퍼스를 이용한 비정형 개인정
보 탐지 기법. 한국정보과학회 학술발표논문집, 772-774.
- 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.
- 최재웅 역(2018), 코퍼스 언어학 방법·이론·실제, 고려대학교 출판문화원. McEnery,
T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*.

- Cambridge University Press.
- 한상미(2014), 중급 한국어 학습자의 구어 담화에 나타난 조사 오류 연구, 한국어 교육 25(3), 국제한국어 교육학회.
- 한송화 외(2019), 2019년 한국어 학습자 말뭉치 연구 및 구축 사업 보고서, 국립국어원.
- 한송화(2001), 말뭉치와 학습자 오류를 이용한, 외국인 학습자를 위한 한국어 어휘 사전의 의미 기술, 한국어정보학 4, 한국어정보학회.
- 한송화(2018), 한국어 학습자의 종결어미 사용 양상과 오류 연구, 문법교육 33, 한국문법교육학회, 166-210.
- 한송화, 원미진(2017), 모어 화자와 한국어 학습자 말뭉치에서의 ‘은/는’과 ‘이/가’의 분포와 조사 선택 요인 분석, 언어사실과 관점 41, 연세대학교 언어정보연구원, 5-47.
- 한송화·강현화(2016), 학습자 말뭉치에서의 구어 전사와 오류 주석의 쟁점과 실제, 한국언어문화교육학회 학술대회 발표 자료집, 한국언어문화교육학회.
- Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1), 71-83.
- 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.
- Granger S. (2008). Learner corpora. In Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics. An International Handbook*. Volume 1. Berlin & New York: Walter de Gruyter, 259-275.
- Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.
- Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research report No. 3. Champaign, IL, USA: NCTE. pp. 1467-1770.
- Inga Kaija, Ilze A. Auzina. (2020). *Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection*. Political Science.
- James, C.(1998), *Errors in Language Learning and Use*. New York : Addison

- Welsey Longman Inc. pp. 144-154.
- Kaija, I., & Auziņa, I. (2019). Data collection for learner corpus of Latvian: copyright and personal data protection. *CLARIN*, 2019(02.10).
- Nagata, R., Whittaker, E., & Sheinman, V. (2011, June). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1210-1219).
- Pica, T., Holliday, L., Lewis, L. and Morgenthaler, L. (1989) Comprehensible Output As An Outcome of Linguistic Demands On the Learner, *Studies in Second Language Acquisition* 11:1. pp. 63-90.
- Ryo Nagata, Edward Whittaker, Vera Sheinman. (2011). *Creating a manually error-tagged and shallow-parsed learner corpus*. Association for Computational Linguistics
- Tracy-Ventura, N., & Huensch, A. (2018). The potential of longitudinal learner corpora in SLA research. In A. Gudmestad & A. Edmonds (Eds.), *Critical Reflections on data in second language acquisition*. Amsterdam: John Benjamins.
- Young, R. (1995). Conversational Styles in Language Proficiency Interviews. *Language Learning* 45:1. pp. 3 - 42.

2. 관련 정책 및 법령 자료

- 4차 산업혁명 대정부 권고안(4차산업혁명위원회)
- 4차 산업혁명 해외 정책 자료집(4차산업혁명위원회)
- 개인정보 보호법(법률)(제16930호)(20200805)
- 공공기관의 개인정보보호에 관한 법률(법률)(제08871호)(20080229)
- 공공기관의 정보공개에 관한 법률 (약칭: 정보공개법)(법률)(제17690호)(20201222)
- 공공데이터의 제공 및 이용 활성화에 관한 법률 (약칭: 공공데이터법)(법률)(제17344호)(20201210)
- 글로벌 AI 정책동향(4차산업혁명위원회)
- 저작권법(법률)(제17588호)(20210623)
- 지능정보화 기본법(법률)(제17344호)(20210610)

3. 웹사이트

British Academic Written English (BAWE) corpus

<https://www.abo.fi/en/english-language-and-literature-research/the-batmat-corpus/>

Corpus and Repository of Writing (Crow) <https://writecrow.org/>

Data Collection for Learner Corpus of Latvian (LaVA) <http://lava.korpuss.lv/en/>

ICLE(International Corpus of Learner English)

<https://corpora.uclouvain.be/cecl/icle/trial/>

L2 Spanish Written Corpus(CEDEL2) <http://cedel2.learnercorpora.com/>

LANGSNAP 3.0 <http://scholarcommons.usf.edu/langsnap/>

The AKCES/CZESL (Acquisition corpora of Czech/Czech as a second language) corpus

<https://www.clarin.eu/resource-families/L2-corpora>

The Cambridge Learner Corpus (CLC)

<https://www.sketchengine.eu/cambridge-learner-corpus/>

The Gachon Learner Corpus

<http://koreanlearnercorpusblog.blogspot.com/p/corpus.html>

The Japanese Learner English Corpus (NICT JLE)

https://alaginrc.nict.go.jp/nict_jle/index_E.html

The Jinan Chinese Learner Corpus (JCLC)

<http://web.science.mq.edu.au/~smalmasi/resources/jclc>

The Longman Learners' Corpus

<http://www.pearsonlongman.com/dictionaries/corpus/learners.html>

The Michigan Corpus of Upper-level Student Papers (MICUSP)

<https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple>

The University of Pittsburgh English Language Institute Corpus (PELIC)

<https://github.com/ELI-Data-Mining-Group/PELIC-dataset>

러시아학습자 코퍼스 <http://web-corpora.net/RussianLearnerCorpus/search/>

스페인어 학습자 코퍼스

https://www.cervantes.es/lengua_y_ensenanza/tecnologia_espanol/caes.htm

연세대학교 IRB <https://irb.yonsei.ac.kr/>

부록1. 기초 연구 자료

차 례

국가 언어 자원(학습자 말뭉치) 구축 및 활용 저작권 이용 허락 동의서	1
개인정보 수집·이용 및 제3자 제공 동의서	7
국가 언어 자원(학습자 말뭉치) 이용 약정서(개인용)	10
국가 언어 자원(학습자 말뭉치) 이용 약정서(기관용)	13
한국어 학습자 말뭉치 사용자 요구조사 설문지	16
자료 수집에 대한 학습자의 의견 조사 설문지	23

국가 언어 자원(학습자 말뭉치) 구축 및 활용 저작권 이용 허락 동의서

저작자 및 저작권 이용허락자 _____성명_____ (이하 “권리자”이
함)와 저작권 이용자 국립국어원(이하 “이용자”이라 함)은 아래 저작물에
관한 저작권재산권 이용허락과 관련하여 다음과 같이 상호 합의하여 동의서
를 작성하기로 한다.

제1조 (동의의 목적)

본 동의는 국가 언어 자원(말뭉치) 구축 및 활용을 위해 권리자가 제공하
는 저작물에 대하여 말뭉치의 이용허락과 관련하여 권리자와 이용자 사이
의 저작물 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (동의의 대상)

본 동의의 이용허락 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)
에 대한 저작권재산권 중 당사자가 합의한 다음의 권리로 한다.

저작물: 한국어 학습자가 산출한 글 원문과 음성 원본

저작자 정보: (성명)_____

(여권번호, 외국인등록번호 또는 주민등록번호)

종별: ☐ 글쓰기 자료 ☐ 말하기 음성 녹음

권리: ☒ 복제권 ☒ 전송권 ☒ 배포권 ☒ 2차적저작물작성권

※저작권 이용허락 대상 권리의 내용

1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저
작물의 글과 전사한 텍스트 및 대상저작물의 음성과 청취·전사한
텍스트를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모,

음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물의 음성을 청취·전사하여 텍스트 및 파일로 변형하고, 그 텍스트 및 파일을 복제·변형·번역(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등)하는 일

3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물의 글과 전사한 텍스트, 대상저작물의 음성과 청취·전사한 텍스트 및 파일과 그 복제·변형·번역물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
4. 대상저작물의 글과 전사한 텍스트, 대상저작물의 음성과 청취·전사한 텍스트 및 파일과 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물의 글과 전사 텍스트, 대상저작물의 음성을 청취·전사한 텍스트 및 파일과 그 복제·변형·번역물을 분석 및 처리하여 사용하는 것을 허락하는 일
5. 기타 국립국어원 또는 그로부터 위탁을 받은 자가 국가 언어 자원(말뭉치) 구축 및 활용을 위하여 필요한 일(단, 권리자의 이익을 지나치게 해하지 않는 범위로 한정됨)

제3조 (이용허락 기간)

대상저작물의 이용허락 기간은 동의를 한 날로부터 2051년 12월 31일까지로 하며, 권리자가 이용허락을 중지하고자 하는 의사를 밝히지 아니하면 이용허락이 5년 단위로 자동 갱신된다. 권리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락을 중지하여야 하며, 그렇지 아니하면 이용허락 내용이 유지된다.

제4조 (권리자의 의무)

(1) 권리자는 이용자에게 대상저작물에 관하여 본 동의서 제2조에 따른 저작권재산을 이용할 권리를 제3조의 기간 동안 비독점적으로 허락한다. 다만, 이용자가 국가 언어 자원(말뭉치) 구축 및 활용 사업을 종료할 때까지의 기간 동안은 독점적으로 허락한다.

(2) 권리자는 이용자에게 동의 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다.

(3) 권리자는 대상저작물의 저작권산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 하고, 본 연구 및 연구를 기초로 만들어지는 이차적저작물 이용을 어렵게 하거나 불가능하게 하는 동의를 하여서는 아니된다. 권리자는 대상저작물에 질권 등 대상저작물의 비독점적 영구적 이용에 방해되는 권리가 존재하는 경우, 이용자에게 그 사실을 즉시 알려야 하고, 이용에 방해되는 권리를 말소하는 등 해소하여야 한다.

(4) 권리자는 이용자가 저작자의 저작물을 이용하는 경우에 성명을 표시하지 않거나, 동일성을 유지하지 않는다고 하더라도 본질적인 변경이 아니거나, 명예훼손에 해당하지 않는 경우에는 문제를 제기하지 않는다.

제5조 (이용자의 권리 및 의무)

(1) 이용자는 대상저작물을 제3조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다. 다만, 이용자가 국가 언어 자원(말뭉치) 구축 및 활용 사업을 종료할 때까지의 기간 동안은 독점적으로 자유롭게 이용할 수 있다.

(2) 이용료는 설정하지 아니한다.

(3) 이용자는 대상저작물의 이용권한을 제3자에게 양도하지 않고, 대상저작물 또는 대상저작물의 이차적저작물에 대하여 질권을 설정하고자 하는 경우 권리자에게서 서면으로 동의를 받아야 한다.

(4) 이용자는 관례적으로 저작자의 성명 등 표시를 하는 방식으로 대상저작물을 이용하는 경우에는 그 저작자 성명 등을 표시하도록 노력하여야 한다. 다만 관례적으로 성명표시하지 않는 경우에는 성명표시하지 않을 수 있다.

(5) 이용자는 대상저작물을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만 제2조에 따른 목적에 한하여 제2조에 따른 변형을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 권리자에게 그 사실을 사전에 고지한 후 사소한 수정 및 편집을 할 수 있다.

제6조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

- ① 대상저작물의 저작물이용허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
- ② 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
- ③ 대상저작물에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.

- ① 대상저작물을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것
- ② 대상저작물을 저작자의 명예를 훼손하는 방법으로 이용하지 아니할 것

제7조 (동의내용의 변경)

본 동의 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.

제8조 (동의의 해지)

- (1) 당사자는 천재지변 또는 기타 불가항력으로 동의를 유지할 수 없는 경우를 제외하고 본 동의를 해지하지 않는다.
- (2) 당사자는 상대방이 정당한 이유 없이 본 동의를 위반하는 경우에 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 동의를 해지할 수 있다.
- (3) 본 동의에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

제9조 (손해배상)

당사자가 정당한 이유 없이 본 동의를 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 제1항의 사유로 본 동의를 이행하지 못한 경우에는 손해배상책임을 면한다.

제10조 (비용의 부담)

동의서의 작성에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁해결)

(1) 본 동의에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하고, 분쟁이 원만히 해결되지 않는 경우에는 소제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

(2) 제1항에 따라 해결되지 아니할 때에는 서울남부지방법원을 제1심 소송의 전속관할로 정하여 소송에 의해 해결토록 한다.

제12조 (비밀유지)

양 당사자는 본 동의서의 작성 및 이행과정에서 알게 된 상대방에 관한 정보, 본 동의의 내용 및 대상저작물의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다. 다만, 동의의 내용을 저작자에게 알리는 경우, 법원의 명령에 의한 경우, 법률의 규정에 의한 경우는 예외로 한다.

제13조 (기타부속합의)

(1) 권리자와 이용자는 본 동의의 내용을 보충하거나, 이 동의에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 동의의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (동의의 해석 및 보완)

본 동의서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (동의 효력 발생일)

본 동의의 효력은 동의 체결일로부터 발생한다.

관리자

성명: (인)

생년월일:

주소:

이용자

성명: 국립국어원장 (인)

주소: 서울특별시 강서구
금남화로 154

개인정보 수집·이용 및 제3자 제공 동의서

국립국어원(용역 사업 수행자: 연세대학교 산학협력단)은 한국어교육의 질적 향상을 위해 학습자들의 언어 자료(말뭉치)를 수집하여 활용하는 연구용역과 관련하여 개인정보를 수집, 이용하거나 제3자에게 제공하고 자 하는 경우에는 「개인정보보호법」 제15조, 제17조, 제23조, 제24조에 따라 아래와 같이 귀하의 개인정보 수집, 이용 및 제3자 제공에 대한 동의를 얻고자 합니다.

만약 참여를 원하지 않을 때에는 참여 의사를 철회할 수 있습니다. 또한 수집하는 개인 정보는 아래의 목적으로만 사용될 것입니다. 감사합니다.

[개인정보 수집·이용에 대한 동의]

수집 · 이용목적	<ul style="list-style-type: none"> ○ 한국어교육의 질적 향상을 위해 학습자들의 언어 자료를 수집하여 교육 및 연구, 민간에서 활용 가능한 말뭉치로 구축 ○ 신원 확인 및 민원 사항 처리 ○ 참여자의 수당 및 저작권료 정산 ○ 제3자(특정 필요) 이용 목적 <ul style="list-style-type: none"> - 감사 및 실사, 정밀 정산 등 연구 종료 후의 관리 자료
수집·이용할 항목	<ul style="list-style-type: none"> ○ 성명, 국적, 출생연도, 제1언어, 한국어 학습기간, 한국에서의 거주기간, 연락처, 직업, 사용하는 외국어, 학력, 산출한 음성 및 텍스트 자료 ○ 은행 계좌 정보, 주민등록번호, 외국인등록번호, 여권번호
보유·이용 기간	동의를 한 시점으로부터 5년(또는 동의를 한 시점부터 사업 완료일 이후 5년)
귀하는 이에 대한 동의를 거부할 수 있습니다. 다만, 동의가 없을 경우	

당 기관의 연구에 참여할 수 없음을 알려드립니다.
위의 사항을 숙지하였으며 (<input type="checkbox"/> 동의함, <input type="checkbox"/> 동의하지 않음) ※ 동의 여부를 <input type="checkbox"/> 에 <input checked="" type="checkbox"/> 표기

[고유식별정보 처리에 대한 동의]

수집하는 고유식별정보 항목	주민등록번호, 외국인 등록번호, 여권번호
고유식별정보의 수집 및 이용목적	○ 국립국어원이 발주한 용역 사업 수행자(연세대학교 산학협력단)의 연구과제 수행 및 본 연구의 저작권료 정산, 세금 처리 ○ 제3자(특정 필요) 이용 목적 - 감사 및 실사, 정밀 정산 등 연구 종료 후의 관리 자료
고유식별정보의 보유 및 이용기간	동의한 시점으로부터 5년(또는 동의한 시점부터 사업 완료일 이후 5년)
귀하는 이에 대한 동의를 거부할 수 있습니다. 다만, 동의가 없을 경우 당 기관의 연구에 참여할 수 없음을 알려드립니다.	
위의 사항을 숙지하였으며 (<input type="checkbox"/> 동의함, <input type="checkbox"/> 동의하지 않음) ※ 동의 여부를 <input type="checkbox"/> 에 <input checked="" type="checkbox"/> 표기	

[개인정보의 제3자 제공에 대한 동의]

제공 목적	국립국어원(용역 사업 수행자: 연세대학교 산학협력단) 연구 및 이후 저작물 관리와 민원 해결
제공 항목	성명, 주민등록번호, 외국인 등록번호, 여권번호
제공받는 자	국립국어원, 용역 사업 수행자(연세대학교 산학협력단), 제3자 적시
제3자 보유·이용 기간	동의한 시점으로부터 5년(또는 동의한 시점부터 사업 완료일 이후 5년)

<p>귀하는 이에 대한 동의를 거부할 수 있습니다. 다만, 동의가 없을 경우 당 기관의 연구에 참여할 수 없음을 알려드립니다.</p>
<p>위의 사항을 숙지하였으며 (<input type="checkbox"/> 동의함, <input type="checkbox"/> 동의하지 않음) ※ 동의 여부를 <input type="checkbox"/>에 <input checked="" type="checkbox"/> 표기</p>

※ 개인정보 제공자가 동의한 내용 외의 다른 목적으로 활용하지 않으며, 제공된 개인정보의 이용을 거부하고자 할 때에는 개인정보관리책임자를 통해 열람, 정정, 삭제를 요구할 수 있음.

「개인정보보호법」 등 관련 법규에 의거하여 상기 본인은 위와 같이 개인정보 수집 및 활용에 동의함.

년 월 일

이 름 _____ (서명)

국립국어원·용역사업수행자(연세대학교 산학협력단) 귀하

국가 언어 자원(학습자 말뭉치) 이용 약정서(개인용)

국립국어원으로부터 본 약정서에 따른 자료를 제공받는 _____은
(는) 다음의 사항을 약정한다.

제1조

국립국어원은 한국어 정보 처리 연구와 국어 연구에 필요한 다음의 자료를 제공한다.

[제공 자료] 2015~2020년 한국어 학습자 말뭉치 구축, 가공 성과물
원시 말뭉치(TXT), 형태 주석 말뭉치(엑셀, XML), 오류 주석 말뭉치(엑셀,
XML) 일괄 배포

※ 표본 원본은 제공하지 않음

제2조

_____은(는) 국립국어원으로부터 제1조에 기재한 형태로 제공받
은 자료를 비영리적/비상업적으로 아래의 주제와 목적으로만 이용하고, 제3자
에게 어떤 이유로든 제공하지 않는다.

연구 개발 목적	논문용/석·박사 학위 논문용/연구 테스트 등
연구 주제	

제3조

_____은(는) 국립국어원으로부터 제공받은 제1조에 기재된 자료
가 어떠한 이유에서든지 유출되어 발생하는 문제에 대한 모든 형사 책임뿐만
아니라, 국립국어원에 발생하는 소송비용과 변호사 비용을 포함하는 모든 손
해에 대하여 책임을 부담한다.

제4조

_____은(는) 국립국어원으로부터 제공받은 제1조에 기재된 자료를 활용하여 새로운 결과물을 산출했을 경우, 이용 상황에 따라 합리적이라고 인정되는 방법으로 제공받은 자료 중 활용한 자료의 종류와 출처를 명시하여야 한다.

제5조

국립국어원이 제1조에 기재한 형태로 제공한 자료의 이용중지를 서면으로 요청받은 이용자 _____는 즉시 이용을 중지하여야 한다. 이용 중지 요청에도 불구하고 계속적인 이용으로 발생하는 모든 법적 문제에 대하여 이용자가 소송비용과 변호사 비용을 포함하는 모든 민사적 책임을 부담한다.

제6조

국립국어원과 이용자 사이에 발생한 분쟁은 콘텐츠분쟁조정위원회의 조정의 결정에 따른다.

부칙

이 계약의 효력은 국립국어원과 약정자의 명의를 변경되더라도 실체가 바뀌지 않으면 존속한다.

약정자	이름	(수기 서명)
	생년월일	
	소속기관	
	email	
	전화번호	
	주소	

국립국어원장 귀하

※ 약정서 2부를 작성하여 서명하신 후 아래의 전자우편 주소로 보내
주십시오(스캔본).

전자우편: klang@korea.kr 전화: 02-2669-9670 파일명: 신청연월일, 소속, 이름
모두 기재

국가 언어 자원(학습자 말뭉치) 이용 약정서(기관용)

국립국어원으로부터 본 약정서에 따른 자료를 제공받는 _____은
(는) 다음의 사항을 약정한다.

제1조

국립국어원은 한국어 정보 처리 연구와 국어 연구에 필요한 다음의 자료를 제공한다.

[제공 자료] 2015~2020년 한국어 학습자 말뭉치 구축, 가공 성과물
원시 말뭉치(TXT), 형태 주석 말뭉치(엑셀, XML), 오류 주석 말뭉치(엑셀,
XML) 일괄 배포

※ 표본 원본은 제공하지 않음

제2조

_____은(는) 국립국어원으로부터 제1조에 기재한 형태로 제공받
은 자료를 비영리적/비상업적으로 아래의 주제와 목적으로만 이용하고, 제3자
에게 어떤 이유로든 제공하지 않는다.

연구 개발 목적	논문용/석·박사 학위 논문용/연구 테스트 등
연구 주제	

제3조

_____은(는) 국립국어원으로부터 제공받은 제1조에 기재된 자료
가 어떠한 이유에서든지 유출되어 발생하는 문제에 대한 모든 형사 책임뿐만
아니라, 국립국어원에 발생하는 소송비용과 변호사 비용을 포함하는 모든 손
해에 대하여 책임을 부담한다.

제4조

_____은(는) 국립국어원으로부터 제공받은 제1조에 기재된 자료를 활용하여 새로운 결과물을 산출했을 경우, 이용 상황에 따라 합리적이라고 인정되는 방법으로 제공받은 자료 중 활용한 자료의 종류와 출처를 명시하여야 한다.

제5조

국립국어원이 제1조에 기재한 형태로 제공한 자료의 이용중지를 서면으로 요청 받은 이용자 _____은(는) 즉시 이용을 중지하여야 한다. 이용 중지 요청에도 불구하고 계속적인 이용으로 발생하는 모든 법적 문제에 대하여 이용자가 소송비용과 변호사 비용을 포함하는 모든 민사적 책임을 부담한다.

제6조

국립국어원과 이용자 사이에 발생한 분쟁은 콘텐츠분쟁조정위원회의 조정의 결정에 따른다.

부칙

이 계약의 효력은 국립국어원과 약정자의 명의를 변경되더라도 실체가 바뀌지 않으면 존속한다.

약정 기관	업체명	
	사업자등록번호	
	대표자	
	주소	
	전화번호	
	email	
	약정자 이름	(서명)

국립국어원장 귀하

※ 약정서 2부를 작성하여 서명하신 후 아래의 전자우편 주소로 보내
주십시오(스캔본).

전자우편: klang@korea.kr 전화: 02-2669-9670 파일명: 신청연월일, 소속, 이름
모두 기재

한국어 학습자 말뭉치 사용자 요구조사

안녕하십니까. 이 설문조사는 국립국어원 「한국어 학습자 말뭉치 연구 및 구축」 사업의 일환으로 한국어교육 연구와 교육에 효율적으로 활용 가능한 말뭉치를 구축하기 위하여 사용자의 의견을 수렴하는 데에 목적이 있습니다. 학습자 말뭉치 이용 경험을 토대로 아래의 문항에 답해 주시기 바랍니다. 설문조사 결과는 다른 목적으로는 사용되지 않을 것임을 약속드립니다. 설문에 응해 주셔서 감사합니다.

사업 주관 기관: 국립국어원

사업 수행 기관: 연세대학교 산학협력단

연구 책임자: 한송화

문의: 홍혜란(anna98kr@yonsei.ac.kr)

I. 기본 정보

1. 국적

- ① 대한민국
- ② 그 외()

2. 직업(복수 응답 가능)

- ① 대학 교수 및 강사
- ② 학부 및 대학원생
- ③ 한국어 강사
- ④ 기타()

3. 소속 기관

- ① 대학
- ② 언어 교육 기관
- ③ 기타()

① 국내의 경우: () (예: 수도권)
② 국외의 경우: () (예: 중국)

1. 한국어 학습자 말뭉치 자료를 언제 이용하였습니까? 국립국어원에 자료를 요청하거나 다운로드 받은 시기를 기준으로 ‘월’을 포함해 응답해 주세요.
(년 월) (예: 2019년 2월)

- ① 학습자 말뭉치 아카데미
- ② 학습자 말뭉치 관련 연구
- ③ 국립국어원 및 유관 기관 사이트
- ④ 수강하는 강의
- ⑤ 기타()

- ① 나눔터에서 직접 다운로드
- ② 국립국어원에 요청

- ① 한국어 및 한국어 교육 연구에 활용(4-1로 가십시오.)
- ② 교수학습 자료 개발에 활용(4-4으로 가십시오.)
- ③ 기타()

- ① 학습자 중간언어(오류) 및 습득 연구
- ② 학습자 언어의 전산처리 연구
- ③ 기타()

4-2. 연구 영역

- ① 어휘
- ② 문법
- ③ 기능(말하기, 쓰기)
- ④ 발음
- ⑤ 기타()

4-3. 구체적인 연구의 주제는 무엇입니까?

()

4-4. 교수학습 자료 개발에의 활용

- ① 교재 개발
- ② 수업 보조 자료 개발
- ③ 학습자 평가 자료 개발
- ④ 기타()

5. 주로 이용한 한국어 학습자 말뭉치 유형

5-1. 사용한 말뭉치의 유형은 무엇입니까? (복수 선택 가능)

- ① 원시 말뭉치
- ② 형태 주석 말뭉치
- ③ 오류 주석 말뭉치

5-2. 사용한 말뭉치의 자료 유형은 무엇입니까?

- ① 문어
- ② 구어
- ③ 전체

5-3. 사용한 말뭉치의 학습자 수준은 무엇입니까? (복수 선택 가능)

- ① 1급
- ② 2급
- ③ 3급
- ④ 4급

- ⑤ 5급
- ⑥ 6급
- ⑦ 6급 이상
- ⑧ 전체

5-4. 사용한 말뭉치의 언어권은 무엇입니까?

- ① 전체
- ② 특정 언어권인 경우(사용한 언어권을 모두 써 주십시오.):

5-5. 사용한 말뭉치의 자료 형식은 무엇입니까? (복수 선택 가능)

- ① TXT
- ② 엑셀
- ③ XML

6. 사용한 말뭉치에 대한 만족도 (※ 해당 항목에 표시해 주세요.)

문항	매우 그렇다	그렇다	보통이 다	그렇지 않다	전혀 그렇지 않다
1) 말뭉치의 규모가 이용 목적에 따라 활용하기에 충분했다.					
2) 말뭉치의 구성이 이용 목적에 따라 활용하기에 충분했다.					
3) 말뭉치의 제공 형식(엑셀, txt, XML)가 이용 목적에 따라 활용하기에 적합했다.					
4) 사용 목적에 따라 말뭉치를 가공하여 사용하기에 용이했다.					

6-1. 말뭉치의 구성과 관련해 개선할 점이나 제안 사항이 있으면 자유롭게 써 주세요. (예. 영어권 6급 학습자의 자료가 더 구축되었으면 좋겠음, 논설문 텍스트가 더 구축되었으면 좋겠음, ……)

6-2. 말뭉치의 제공 형식(엑셀, txt, XML)과 관련해 개선할 점이나 제안 사항이 있으면 자유롭게 써 주세요.

6-3. 말뭉치 가공 사용과 관련해 개선할 점이나 제안 사항이 있으면 자유롭게 써 주세요.

7. 한국어 학습자 말뭉치의 효율적 활용을 위해 한국어 학습자 말뭉치 구축 연구팀에 제안하고 싶은 사항이 있으면 자유롭게 써 주세요.

III. 한국어 학습자 말뭉치 나눔터 이용 경험과 평가

1. 한국어 학습자 말뭉치 나눔터를 얼마나 자주 이용하십니까?

- ① 자주 이용한다.
- ② 가끔 이용한다.
- ③ 별로 이용하지 않는다.
- ④ 거의 이용하지 않는다.
- ⑤ 기타()

2. 주로 이용하는 메뉴는 무엇입니까?(복수 응답 가능)

- ① 말뭉치 통합 검색
- ② 원시 말뭉치 검색
- ③ 형태 주석 말뭉치 검색
- ④ 오류 주석 말뭉치 검색
- ⑤ 통계 정보

3. 나눔터 이용에 대한 만족도 (※해당 항목에 표시해 주세요.)

구분	문항	매우 그렇 다	그렇 다	보통 이다	그렇 지 않다	전혀 그렇 지 않다
자료 접근성	1) 검색 사이트 및 링크를 통한 접근이 용이하다.					

및 UI의 편의성	2) 메뉴가 이용하기 편리하게 구성되어 있어 원하는 정보를 쉽게 찾을 수 있다.					
	3) 사이트 접속 상태가 안정적이다.					
	4) 검색 결과 자료를 내려받기 쉽다.					
검색 기능	1) 말뭉치 검색 조건이 다양하게 제시되어 있다.					
	2) 말뭉치 검색 조건이 적절하게 제시되어 있다.					
	3) 검색 결과가 보기 편하게 되어 있다.					
	4) 검색 조건에 따라 적합한 결과가 제시된다.					
통계 정보	1) 원하는 통계 결과를 제공하고 있다.					
	2) 통계 자료를 확인하는 데 불편함이 없다.					
	3) 통계 수치와 검색 결과가 일치한다.					

3-1. 자료의 접근성 및 UI의 편의성에서 개선되었으면 하는 점은 무엇입니까?

3-2. 검색 기능에서 개선되었으면 하는 점은 무엇입니까?

3-3. 통계 정보에서 개선되었으면 하는 점은 무엇입니까?

4. 앞으로 한국어 학습자 말뭉치 나눔터를 계속 이용하실 계획입니까? 그 이유가 무엇인지 구체적으로 써 주세요.

5. 한국어 학습자 말뭉치 나눔터를 다른 사람에게 추천할 의사가 있습니까?
그 이유가 무엇인지 구체적으로 써 주세요.

6. 그 밖에 학습자 말뭉치 활용의 효율성과 편리함을 위해 한국어 학습자 말뭉치 나눔터 개선과 관련해 제안하고 싶은 사항이 있으시면 자유롭게 기술해 주십시오.

자료 수집에 대한 학습자의 의견 조사

1. 저작권 이용 허락 동의서 및 이용 허락서의 내용을 충분히 이해하고 서명하였습니까?

- ① 네 ② 아니오

1-1. 저작권 이용 허락 동의서 및 이용 허락서의 내용을 이해하는 데에 동영상 설명 자료가 도움이 되었습니까?

1-2. 저작권 이용 허락 동의서 및 이용 허락서에 서명하면서 어떤 감정을 느꼈습니까?

- ① 특별한 감정을 느끼지 않았다
② 부담스러웠다
③ 기타: _____

2. 개인정보 수집, 이용 및 제3자 제공 동의서의 내용을 충분히 이해하고 서명하였습니까?

3. 자료 수집을 위해 사용할 수 있는 시간은 얼마나 됩니까?

- ① 30분 이내
② 30분-1시간
③ 1시간 이상

4. 자료 수집자에게 제공하는 적절한 보상 방법은 무엇이라고 생각합니까?

- ① 작문/말하기 자료에 대한 피드백
② 금전적 보상(적정 비용: _____ 달러)
③ 기타: _____

5. 향후에도 자료 수집에 참여하고 싶습니까?

① 네 ② 아니오

이유:

6. 친구에게 자료 수집을 추천하고 싶습니까?

① 네 ② 아니오

이유:

부록2. 2021년 한국어 학습자 말뭉치 구축 지침

차 례

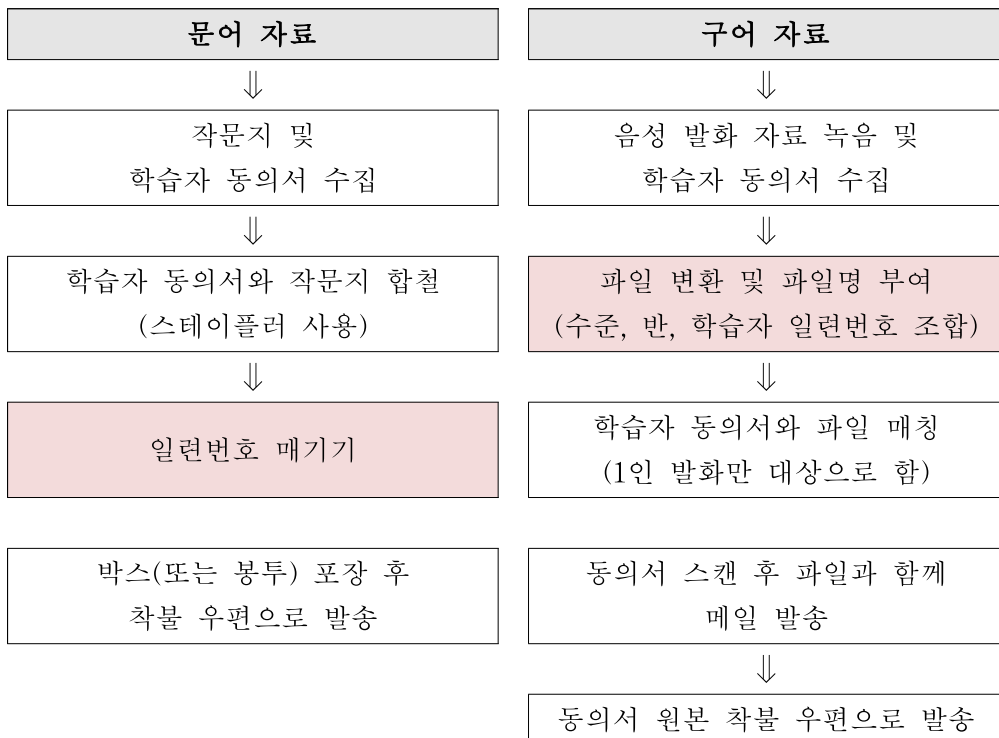
한국어 학습자 말뭉치 수집 지침	1
한국어 학습자 말뭉치 자료 처리 지침	45
한국어 학습자 말뭉치 문어 입력 지침	50
한국어 학습자 말뭉치 구어 전사 지침	57
한국어 학습자 말뭉치 형태 주석 지침	79
한국어 학습자 말뭉치 오류 주석 지침	145

한국어 학습자 말뭉치 자료 수집 지침

1. 자료 수집 대상 및 수집 자료

- ▶ 대상: 국내 한국어 교육 기관의 학습자
- ▶ 자료: 학습자가 산출한 작문과 말하기 자료
- ▶ 수집 시기: 여름 학기와 가을 학기의 각 중간, 기말의 2회(총 4회)를 원칙으로 한다. (추가 가능)

2. 자료 수집 절차



1) 문어

- ① 학습자가 손으로 쓴 작문지나 시험 답안지의 원본(사본도 가능)을 수집한다.
- ② 사본의 경우 복사가 흐릿하여 텍스트를 알아보기 힘든 경우 자료의 활용이 불가능하므로 유의한다.
- ③ 해당 자료의 출처를 파악할 수 있도록 수집된 자료는 반드시 동의서와 함께 수합하여 합철을 한다.
- ④ 합철이 된 파일에 아래와 같이 급별로 네 자리의 일련번호(0001, 0002, ...)를 붙인다.

0001

建立韩语学习者语料库的个人资料使用同意书

韩国国立语院为了韩语教育的发展，正在建设语料库（语料库）以促进韩语教学方法的改进，韩国语教材的开发，韩国语教育领域及韩语学习。为提供语料库建设并促进韩语教学方法的改进，韩国语教材的开发，韩国语教育领域及韩语学习。参加语料库建设的各位没有经济损失和人身危险，如果愿意参加以上语料库建设时请参加同意书。另外，收集的个人资料中用于除本语料库以外的其他目的，方法，为了研究使用，所收集的个人资料不能识别的情况下进行使用，请同意。

寄附处：延世大学 产学合作组 02-2123-4199

同意书提供并使用时以下信息至2015年度/秋季学期论文，会议资料。

日期：7/9

姓名：김민우 (0001)

모바우

下面是语料库所应用的个人信息，信息资料保管不对外提供。

1. 性别： ☒ 男 ☐ 女
2. 年龄： 29
3. 现在韩国留学原因： ☒ 留学 ☐ 研修 ☐ 其他
4. 国籍： ☒ 韩国 ☐ 中国 ☐ 日本 ☐ 美国 ☐ 其他
5. 母语： ☒ 韩语 ☐ 英语 ☐ 其他
6. 韩国语的学历： 0 年 2 月
7. 在韩国语的学习目的： 0 年 2 月
8. 学习韩国语的用途： ☒ 留学 ☐ 研修 ☐ 其他 ☐ 其他
9. 职业： ☒ 学生 ☐ 教师 ☐ 其他
10. 除韩国语以外可以使用的语言(按韩语的翻译来填写)： Chinese English Japanese Korean

001

홍남대학교 국문학과교수인 박민우

제1 주위에 주위를 살펴보니까 어디에 갔나? 누구를 만났습니까? 무엇을 먹었습니까? 여행했습니까?
 주위로, 아래에 표를 모두 사용해서 써주세요. (150~300자) [30점]

이름: _____

주	말	에	한	국	어	를	공	분	했	습	요	돈	을	만	에
취	구	로	의	화	는	배	물	었	습	요	마	며	니	를	
여	구	에	서	만	났	습	니	다	-	저	는	백	화	점	에
소	스	씨	배	를	불	공	그	하	고	먹	었	습	니	요	
저	는	시	장	에	세	곳	빠	나	나	시	과	기	회		
추	려	요	모	도	7	0	0	0	0	원	임	니	다		
인	요	만	에	저	는	국	장	에	세	엄	화	를	읽	었	았
문	표														

- ④ 급별로 번호를 붙이며 각각 0001로 시작한다.
- ④ 한 명의 학습자 자료가 두 개 이상일 경우 0001-01, 0001-02, ... 와 같이 앞자리 수는 동의서와 동일하게 맞추고 뒤에 - 01, -02,...를 붙인다.
- ⑤ 착불 우편을 이용하여 아래의 주소로 발송한다. 이때 인터넷 우체국 택배를 이용하여 신청한 후, 이메일(2016klcorpus@gmail.com)로 등기번호를

통지한다(착불 결제용).

120-749 서울시 서대문구 연세로 50 연세대학교 연세우유사무소
언어정보연구원 한국어 학습자 말뭉치 연구실
홍혜란 (전화 010-8727-9024)

2) 구어

- ① 학습자가 산출한 대화, 발표, 토론 등의 원음을 수집한다.
- ② 녹음을 할 때에는 양질의 음성 자료 확보를 위하여 가능하면 보이스 레코더와 같은 녹음기기를 사용한다.
- ③ 하나의 파일에 한 명의 학습자 자료가 녹음되도록 한다. 하나의 파일에 여러 명의 파일을 연이어 녹음한 경우는 학습자별로 파일을 분리한다. 만약, 파일을 분리하지 못할 경우 발화자를 알 수 있도록 녹음된 순서에 맞춰 학습자의 정보를 정리한 후 동의서와 합철한다.
- ④ 파일명은 다음과 같이 국적, 기관명, 수준, 파일 구분을 위한 번호(0001, 0002,...)를 조합하여 부여하고 언더바(_)를 사용하여 순서대로 이어 붙인다.
예) 대만_한국대_1급_0001.wav, 대만_한국대_1급_0002.wav,...
- ⑤ 발표와 같은 1인 발화에 한해서 파일명을 학습자 동의서에 적어 학습자 정보를 파악할 수 있도록 한다.
- ⑥ 파일을 이메일(2016klcorpus@gmail.com)로 발송한 후 동의서 원본은 착불 우편으로 발송한다.

3. 학습자 동의서 수집

- ▶ 모든 자료는 자료 제공과 사용에 관한 학습자의 동의서를 받은 후 수집한다.
- ▶ 동의서는 같은 학기 중의 동일한 학습자라도 자료 수집 시마다 매번 받는 것을 원칙으로 한다. 예를 들어 한 학습자가 여름 학기 과제 작문 한 편, 기말 쓰기 시험의 작문 한 편을 제공할 경우에도 2번의 동의서를 각각 받

도록 한다. 다만, 자료 수집의 효율성이나 기관 내 사정 등으로 인해 매번 받는 것이 어려울 경우 처음 수집할 때 받은 동의서와 짝을 맞출 수 있도록 학습자의 이름, 수준, 학급(반) 정보를 시험지에 적는다. 구어 자료는 학습자 정보와 파일명을 함께 기록한다.

- ▶ 동의서를 수합한 후 누락된 항목이 있는지 확인한다. 국적 정보와 같이 수집 교사가 확인 가능한 항목이 누락된 경우 적어 넣는다.

- [주의] 1. 동의서는 학습자의 모국어 또는 학습자가 가장 이해하기 쉬운 언어로 번역된 것을 배부하여 자료 수집 목적과 개인 정보 제공 등에 관한 사항을 충분히 이해할 수 있도록 한다. 그 밖의 학습자가 추가적으로 궁금해 하는 사항이 있을 경우에는 설명해 준다.
2. 학습자가 수기로 적고 사인하도록 할 수 있도록 출력하여 배포한다. 동의서와 개인 정보는 학습자의 개인 정보 보호를 위하여 자료 분류 후 절취하여 따로 보관하게 된다.
3. 구어 자료 수집 시 2인 이상의 대화 자료를 녹음할 경우 참여 학습자 각각에게 동의서를 받는다.

[참고] 한국어 학습자 말뭉치 자료의 유형 및 수집 방법

1. 횡적 말뭉치(국내 대학 및 이주민 교육 기관)

1) 문어

(1) 수집 원칙

- 수업 활동 또는 수업 과제, 시험에서 작성한 쓰기 자료를 수집한다.
- 하나의 완결된 글이 되도록 한다.
- 모어화자(가족, 교사 포함) 혹은 동료의 피드백이 이루어지지 않은 글이어야 한다.
- 사전 사용이 배제된 작문을 원칙으로 한다.
- 보기 글을 그대로 베껴 쓰거나 주어진 다량의 어휘를 기반으로 한 작문은 되도록 배제한다.
- 구축 본부에서 제시한 기획 과제를 활용할 경우 학습자의 수준에 맞추어 제시된 글의 종류와 주제로 작문을 하게 하여 이를 수집한다(수집 과제는 요청 시에 별도 제공).

(2) 수집 방법

① 교육과정 내 과제 작문 수집

- 각 교육 기관의 교육과정 실러버스에 이미 포함되어 있는 작문을 활용하여 이를 수집함. 글의 종류 및 주제는 각 기관의 교육과정에 따름

② 성취도 평가 수집

- 각 교육 기관의 성취도 평가(중간 및 기말) 쓰기 시험에 포함된 작문을 활용하여 이를 수집함. 글의 종류 및 주제는 각 기관의 성취도 평가에 따름

③ 교육과정 외 프로젝트를 위한 기획 작문 수집(수집 가능 기관)

- 각 등급에 맞추어 수업 시간(1시간) 내에 다음과 같은 글의 종류와 주제

로 작문을 하게 하여 이를 수집함. 세부 주제는 종적 말뭉치의 과제 활동 자료를 참고함

수준	추천 글의 종류	기타	주제
초급	체험적 글(생활문)	일기, 편지, 이메일 등	소개(자신, 가족 등), 취미, 한국생활, 주말, 계절, 좋아하는 음식, 학교생활, 여행, 일상사 등
중급	체험적 글(생활문) 설명적 글(설명문)	안내문, 감상문 등	소개(가족, 문화, 풍습 등), 취미, 여행, 여가생활, 한국생활, 추억, 영화, 만남, 직업, 후회, 사회문제(환경문제 등), 등
고급	설명적 글(설명문) 논리적 글(논설문)	기사문, 게시문 등	사회문제, 경제문제, 문화, 예술, 봉사, 갈등 등

2) 구어

(1) 수집 원칙

- 발화를 유도하기 위해 유인물 등을 기반으로 할 수는 있으나 그대로 읽는 것은 배제하며 읽은 후 이야기를 할 때에는 되도록 보지 않고 발화하도록 한다.
- 해당 등급의 중반 혹은 그 이후에 발화된 것을 녹음하는 것을 원칙으로 한다.
- 교사는 되도록 자신의 발화를 통제하고, 학생이 자신의 발화를 유지할 수 있도록 안내자 정도의 역할을 하도록 한다.
- 학습자가 단어나 구를 활용한 단답형의 대답만 하지 않도록 하며 과제의 주제 또는 교사의 질문과 관련된 내용을 충분히 발화할 수 있도록 기다려 준다(☞수집 과제는 요청 시에 별도 제공).

(2) 수집 방법

① 교육과정 내 담화 수집

- 학습자와 학습자 간의 역할극 혹은 간단한 토론 등과 같이 학습자와 학습

자 간의 2인 발화의 경우에는 각 학습자가 녹음하여 이를 교사에게 전송하게 하여 이를 수집함

- 토론 등 다인 발화의 경우에는 발화자의 정보를 확인 가능하도록 비디오로 녹화하거나 녹음 및 전사자를 일치시킬 것을 권유함
- 이의 담화 유형과 주제, 시간은 각 기관의 교육과정에 따름

② 성취도 평가 수집

- 교사와 학생, 혹은 학생과 학생 간에 이루어지는 성취도 평가의 담화를 수집함
- 이의 담화 유형과 주제, 시간은 각 기관의 성취도 평가에 따름

③ 졸업좌담회, 말하기 대회 등의 자료 수집

- 졸업좌담회, 말하기 대회 등 공식적인 구어 담화를 수집함. 비디오 녹화를 권유함

④ 교육과정 외 담화 자료로 본 프로젝트를 위한 기획 발화 수집(수집 가능 기관)

- 각 등급에 맞추어 수업 시간 내 또는 수업 시간 외에 다음과 같은 주제로 발표 또는 인터뷰 활동을 통해 자료를 수집함. 발화 시간은 5-10분 이내로 함. 세부 주제는 종적 말뭉치의 과제 활동 자료를 참고함

수준	담화 유형	주제
초급	발표, 인터뷰	소개(자신, 가족 등), 취미, 한국생활, 주말, 계절, 좋아하는 음식, 학교생활, 여행, 일상사 등
중급	발표, 인터뷰	소개(가족, 문화, 풍습 등), 취미, 여행, 여가생활, 한국생활, 추억, 영화, 만남, 직업, 후회, 사회문제(환경문제 등), 등
고급	발표, 인터뷰	사회문제, 경제문제, 문화, 예술, 봉사, 갈등 등

2. 종적 말뭉치 (해당 기관)

1) 문어

(1) 수집 원칙

- 학습자들이 작문을 시작하기 전에 주제와 글의 장르를 충분히 이해한 후 글을 쓸 수 있도록 설명하며, 필요한 경우 쓰기 전 활동처럼 관련 질문들을 하시면서 잠시 이야기를 나눌 수 있음
- 초급 단계의 경우 10문장 이상 쓰도록 지도함(중급 15-20문장, 고급 20문장 이상)
- 완성되지 않은 작문 자료의 경우 말뭉치로 구축하기가 어려우므로 주제에 관해 완결된 글을 쓰도록 함
- 작문은 사전이나 교재 등의 자료를 참고하지 않고 쓸 수 있도록 하며, 가능하다면 숙제로 주지 않고 함께 모여서 쓸 수 있도록 함

(2) 수집 방법

- 각 등급에 맞추어 수업 시간(1시간) 외에 다음과 같은 글의 종류와 주제로 작문을 하게 하여 이를 수집함

2) 구어

(1) 수집 원칙

- 자료 수집을 시작하기 전에 발화를 유도하기 위한 도입 질문 등을 통해 학습자가 발화 주제에 대해 충분히 생각한 후 이야기할 수 있도록 유도함
- 학습자가 발화를 충분히 할 수 있도록 시간적 여유를 줌
- 학습자가 발화를 이어가지 못할 경우 간단한 유도 발화를 해서 발화를 이어갈 수 있도록 도움을 줄 수 있음
- 모든 발화에 대하여 과도하게 맞장구를 치거나 학습자가 말하는 도중에 끼어들지 않도록 함

- 학습자가 오류를 범하더라도 일일이 교정해 주지 않음
- 학습자가 발화를 이어가기 위해 특정 어휘나 표현을 생각하느라고 머뭇거리거나 다소 긴 휴지가 지속될 경우 교사가 먼저 말해 주지 않고 학습자가 스스로 발화를 이어가도록 기다려 줌

(2) 수집 방법

- 교육과정 외 담화(본 프로젝트를 위한 기획 발화) 수집
 - 각 등급에 맞추어 수업 시간 내에 다음과 같은 주제로 발표를 하게 하여 이를 수집함. 발표는 5-10분 이내로 함
 - 각 등급에 맞추어 다음과 같은 주제로 교사가 인터뷰를 하여 이를 수집함. 인터뷰는 5-10분 이내로 함. 교사는 되도록 자신의 발화를 통제하고, 학생이 자신의 발화를 유지할 수 있도록 안내자 정도의 역할을 함

학습자 말뭉치 종적 자료 수집 과제(일반)

1. 문어 수집

수집 시기	문제	수준
02주차	자기소개를 해보십시오. 이름이 무엇입니까? 어느 나라 사람입니까? 무엇을 합니까? 무엇을 좋아합니까?	초급
04주차	여러분의 가족에 대해 쓰십시오. 누가 있습니까? 무슨 일을 합니까? 무엇을 좋아합니까?	
06주차	여러분은 토요일이나 일요일에 무엇을 합니까? 어디에 갑니까? 누구를 만납니까? 여러분의 주말 이야기를 쓰십시오.	
08주차	여러분은 어떤 선물을 받고 싶습니까? 왜 그 선물을 받고 싶습니까? 선물에 대한 글을 쓰십시오.	
10주차	어느 계절을 좋아합니까? 왜 그 계절을 좋아합니까? 그 계절에 특별히 무엇을 합니까? 좋아하는 계절에 대해서 글을 쓰십시오.	
12주차	여러분은 뭐 하는 것을 좋아합니까? 왜 그것을 좋아합니까? 그것을 얼마나 자주 합니까? 여러분의 취미에 대해서 쓰십시오.	
14주차	여러분이 가장 좋아하는 친구는 누구입니까? 그 친구는 무엇을 합니까? 왜 그 친구를 좋아합니까? 여러분이 가장 좋아하는 친구를 소개해 보십시오.	
16주차	여러분은 어디에 자주 갑니까? 왜 그곳에 자주 갑니까? 거기에서 무엇을 합니까? 여러분이 자주 가는 장소에 대해서 쓰십시오.	
18주차	여러분은 올해 무엇을 하고 싶습니까? 왜 그것을 하고 싶습니까? 2011년에 하고 싶은 것에 대해 쓰십시오.	
20주차	여러분은 어디에 여행을 가 봤습니까? 그것에서 무엇을 했습니까? 어땠습니까? 여러분의 여행 경험에 대해 쓰십시오.	
22주차	여러분은 10년 후에 어떻게 살고 싶습니까? 그 이유는 무엇입니까? '10년 후의 나의 계획'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시한 내용이 모두 포함되어야 합니다. ○ 10년 후에 어떻게 살고 싶은가? ○ 그 이유는 무엇인가? ○ 무엇을 준비해야 하는가?	중급

수집 시기	문제	수준
24주차	<p>여러분이 소중하게 생각해서 사랑하는 물건은 무엇입니까? ‘내가 가장 아끼는 물건’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 가장 아끼는 물건은 무엇인가? ○ 왜 그 물건을 아끼는가? ○ 어떻게 그 물건을 가지게 되었는가? 	
26주차	<p>여러분은 취미로 무엇을 배우고 싶습니까? ‘내가 취미로 배우고 싶은 것’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 취미로 배우고 싶은 것은 무엇인가? <p>(※ 한국어를 배우고 싶다는 내용은 쓰지 마십시오.)</p> <ul style="list-style-type: none"> ○ 왜 그것을 배우고 싶은가? ○ 그것을 배운 후에 무엇을 하고 싶은가? 	
28주차	<p>여러분은 늦잠을 자거나 누워서 책을 보는 것과 같은 고치고 싶은 생활 습관이 있습니까? ‘고치고 싶은 나의 생활 습관’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 나의 나쁜 생활 습관 ○ 습관 때문에 생기는 불편하거나 안 좋은 점 ○ 습관을 고치기 위해 해야 할 일 	
30주차	<p>잊지 못할 추억’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 어떤 추억인가요? ○ 왜 지금까지 기억에 남아 있는가? ○ 언제 그 추억이 떠오르는가? 	
32주차	<p>갖고 싶은 직업’이라는 제목으로 글을 써 보십시오. 단 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 직업명, 하는 일, 그 일을 하려는 이유, 그 일에 필요한 조건 	
34주차	<p>나의 성격’이라는 제목으로 글을 써 보십시오. 단 아래에 제시된 내용에 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 성격의 특징, 장점과 단점, 고치고 싶은 부분과 그 이유 	
36주차	<p>내가 생각하는 행복’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 행복을 위해서 어떤 노력을 하는가? 	

수집 시기	문제	수준
	<ul style="list-style-type: none"> ○ 언제 행복하다고 느끼는가? ○ 행복은 무엇이라고 생각하는가? 	
38주차	<p>‘내가 좋아하는 책’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 가장 좋아하는 책은 무엇인가? ○ 그 책은 어떤 내용인가? ○ 그 책을 좋아하는 이유는 무엇인가? 	
40주차	<p>여러분은 어떤 사람처럼 되고 싶습니까? 왜 그 사람처럼 되고 싶습니까? ‘내가 닮고 싶은 사람’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 닮고 싶은 사람은 누구인가? ○ 왜 그 사람처럼 되고 싶은가? ○ 그 사람처럼 되기 위해서 어떻게 해야 하는가? 	
42주차	<p>1)~4)의 내용은 ‘피로를 예방하려면 네 가지를 실천하라’는 글의 소재입니다. 이 소재를 이용하여 글을 쓰십시오.</p> <p>‘피로를 예방하려면 네 가지를 실천하라’</p> <ul style="list-style-type: none"> ○ 체질에 맞는 음식 ○ 수면의 질 ○ 적당한 운동 ○ 긍정적인 사고 	
44주차	<p>올바른 인터넷 사용 태도’에 대한 자신의 견해를 서술하십시오. 단, 아래 제시한 <올바른 인터넷 사용 태도의 예> 중에서 세 가지를 선택하여 쓰되, 각각의 태도를 지키지 않았을 경우에 나타나는 부작용의 예를 포함해야 합니다.</p> <p><올바른 인터넷 사용 태도의 예></p> <ul style="list-style-type: none"> ○ 상대방의 의견 존중하기 ○ 타인의 사생활 보호하기 ○ 의견 차이 인정하기 ○ 바른 언어 사용하기 ○ 정확한 정보 올리기 	고급
46주차	<p>다음 글을 읽고, ‘현대 사회에서 바람직한 신문의 기능’에 대한 자신의 견해를 서술하십시오. 단 아래에 제시한 기능 중에서 두 가지 이상을 선택하여 쓰되, 그 기능이 현대 사회에 중요하다고 생각하는 이유를 포함해야 합니다.</p> <p><신문의 기능></p>	

수집 시기	문 제	수준								
	<div>○ 사건 보도</div> <div>○ 여론 조성</div> <div>○ 정보 제공</div> <div>○ 소통의 분위기 조성</div>									
48주차	<div>다음 글을 읽고 '감시 카메라 설치 확대'에 대한 자신의 견해를 서술하십시오. (찬성하거나 반대하는 입장 중 하나를 선택하여 서술할 것. 단 아래 제시된 각 입장의 논거 중 두 개 이상을 제시할 것.)</div> <div><div>최근 들어 각종 범죄가 급증하면서 감시 카메라 설치가 사회적 문제로 대두되고 있다. 지금까지 감시 카메라는 은행이나 지하 주차장 등에 주로 설치되어 있었으나 이제는 설치 장소를 대폭 확대하자는 것이다. 이러한 감시 카메라 설치 확대에 어떻게 생각하는가?</div><table><tr><th>찬성</th><th>반대</th></tr><tr><td>사회 안전 유지</td><td>개인의 사생활 침해</td></tr><tr><td>범죄 예방</td><td>범죄 예방 효과 불분명</td></tr><tr><td>인권보다 공인이 우선</td><td>가해자의 인권 보호</td></tr></table></div>	찬성	반대	사회 안전 유지	개인의 사생활 침해	범죄 예방	범죄 예방 효과 불분명	인권보다 공인이 우선	가해자의 인권 보호	
찬성	반대									
사회 안전 유지	개인의 사생활 침해									
범죄 예방	범죄 예방 효과 불분명									
인권보다 공인이 우선	가해자의 인권 보호									
50주차	<div>여러분은 성공이 무엇이라고 생각하십니까? 그리고 그러한 성공을 이루기 위해 필요한 것이 무엇이라고 생각하십니까? 이와 관련된 자신의 견해를 서술하십시오. 단, 아래에 제시한 내용이 모두 포함되어야 합니다.</div> <div>○ 내가 생각하는 성공이란 무엇인가?</div> <div>○ 그것을 이루기 위해 필요한 것은 무엇인가</div> <div>○ 그 이유는 무엇인가?</div>									
52주차	<div>여러분은 무엇이 선의의 거짓말이라고 생각하니까? 어떤 경우에 그런 거짓말을 할 수 있다고 생각하니까? 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</div> <div>< 선의의 거짓말이란 ></div> <div>○ 선의의 거짓말이란 무엇인가?</div> <div>○ 선의의 거짓말은 언제 필요한가?</div>									

수집 시기	문제	수준
	○ 선의의 거짓말이 가질 수 있는 문제점은 무엇인가?	
54주차	<p>학교에서는 음악이나 미술과 같은 예술 교육이 이루어지고 있습니다. 이러한 예술 교육이 왜 필요하다고 생각합니까? 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시한 내용이 모두 포함되어야 합니다.</p> <p>< 예술 교육의 필요성 ></p> <p>○ 예술 교육이 왜 필요한가?</p> <p>○ 예술 교육을 통해 얻을 수 있는 효과는 무엇인가?</p>	
56주차	<p>자연을 그대로 보존해야 한다는 주장과 인간을 위해 자연을 개발해야 한다는 주장이 있습니다. 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <p><자연 보존과 자연 개발></p> <p>○ 자연 보존과 자연 개발 중 어느 것이 더 중요하다고 생각하는가?</p> <p>○ 그렇게 생각하는 이유는 무엇인가? (2가지 이상 쓰시오.)</p>	
58주차	<div style="border: 1px solid black; padding: 10px;"> <p>오늘날 직업에 대한 생각은 크게 두 가지로 나뉘는 것 같다. 하나는 여러 방면으로 사회에 도움을 주거나, 공헌할 수 있는 직업을 택해 봉사하는 마음으로 일하고, 그것을 통해 얻어지는 대가로 자신과 가정을 꾸려 나가는 것이다. 다른 하나는 사회에 대한 봉사나 공헌보다는 일에 대한 자기만족과, 욕구 충족, 충분한 대가에 더 큰 비중을 두는 경우이다.</p> <p>전자의 경우, 일이 힘들거나 보수가 적다 하더라도 일에 대한 보람과 긍지 때문에 쉽게 그 일을 그만두거나 직업을 바꾸려 생각은 하지 않는다. 하지만 후자의 경우는 일에 대한 즐거움이나 자기 만족, 충분한 보상이 뒤따르지 않는다고 판단될 때는 언제라도 직장을 옮길 마음의 준비가 되어 있다. 전자의 경우에는 사회를 안정시키는 데에 기여를 하지만 보수적 경향으로 사회적 분위기를 다소 침체시킬 수도 있다. 후자의 경우에는 생동감은 있으나 급격한 변화로 안정감을 잃어버릴 위험이 많고, 이런 변화 속</p> </div>	

수집 시기	문제	수준
	<div> <p>에 적응하지 못하는 이들은 사회 변화의 뒷전으로 밀려날 수밖에 없게 된다.</p> </div> <p>위의 글에 나타난 두 가지 유형의 직업관 중 자신의 생각은 어느 쪽인지 말하고, 그 이유를 설득력 있게 글로 나타내시오.</p>	
60주차	<p>현대 사회는 빠르게 세계화·전문화되고 있습니다. 이러한 현대 사회의 특성을 참고하여, ‘현대 사회에서 필요한 인재’에 대해 아래의 내용을 중심으로 자신의 생각을 쓰십시오.</p> <p>○ 현대 사회에서 필요한 인재는 어떤 사람입니까?</p> <p>○ 이러한 인재가 되기 위해서 어떤 노력이 필요합니까?</p>	

2. 구어 수집

수집 시기	글의 종류	주제	수준
02주차	인터뷰	소개(자신, 가족 등)	초급
04주차	발표	취미	
06주차	인터뷰	주말	
08주차	발표	한국 생활	
10주차	인터뷰	계절	
12주차	발표	좋아하는 음식	
14주차	인터뷰	학교생활	
16주차	발표	여행	
18주차	인터뷰	일상(사)	
20주차	발표	선물	
22주차	인터뷰	소개(고향, 문화, 풍습 등)	중급
24주차	발표	스트레스	
26주차	인터뷰	여가생활	
28주차	발표	추억	
30주차	인터뷰	명절	
32주차	발표	영화	

수집 시기	글의 종류	주제	수준
34주차	인터뷰	만남	
36주차	발표	진로와 직업	
38주차	인터뷰	후회	
40주차	발표	환경 문제	
42주차	인터뷰	성공적인 삶	고급
44주차	발표	경제문제	
46주차	인터뷰	문화	
48주차	발표	갈등	
50주차	인터뷰	예술	
52주차	발표	학교 교육	
54주차	인터뷰	봉사	
56주차	발표	현대인의 생활	
58주차	인터뷰	결혼	
60주차	발표	남성과 여성	

학습자 말뭉치 이주민 자료 수집 과제 (결혼 이주민, 이주 노동자)

1. 문어

종적 자료 수집 시기	문제	수준
02주차	자기소개를 해 보십시오. 이름이 무엇입니까? 어느 나라 사람입니까? 무엇을 합니까? 무엇을 좋아합니까?	초급
04주차	여러분의 가족에 대해 쓰십시오. 누가 있습니까? 무슨 일을 합니까? 무엇을 좋아합니까?	
06주차	여러분은 토요일이나 일요일에 무엇을 합니까? 어디에 갑니까? 누구를 만납니까? 여러분의 주말 이야기를 쓰십시오.	
08주차	여러분은 어떤 선물을 받고 싶습니까? 왜 그 선물을 받고 싶습니까? 쓰십시오.	
10주차	어느 계절을 좋아합니까? 왜 그 계절을 좋아합니까? 그 계절에 특별히 무엇을 합니까?	
12주차	여러분은 뭐 하는 것을 좋아합니까? 왜 그것을 좋아합니까? 그것을 얼마나 자주 합니까? 여러분의 취미에 대해서 쓰십시오.	
14주차	여러분이 가장 좋아하는 친구는 누구입니까? 그 친구는 무엇을 합니까? 왜 그 친구를 좋아합니까? 여러분이 가장 좋아하는 친구를 소개해 보십시오.	
16주차	여러분은 어디에 자주 갑니까? 왜 그곳에 자주 갑니까? 거기에서 무엇을 합니까? 여러분이 자주 가는 장소에 대해서 쓰십시오.	
18주차	여러분은 올해 무엇을 하고 싶습니까? 왜 그것을 하고 싶습니까?	
20주차	여러분의 고향은 어디입니까? 고향을 소개하는 글을 쓰십시오.	중급
22주차	여러분은 10년 후에 어떻게 살고 싶습니까? 그 이유는 무엇입니까? '10년 후의 나의 계획'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시한 내용이 모두 포함되어야 합니다.	

종적 자료 수집 시기	문 제	수준
	다. ○ 10년 후에 어떻게 살고 싶은가? ○ 그 이유는 무엇인가? ○ 무엇을 준비해야 하는가?	
24주차	여러분이 소중하게 생각해서 사랑하는 물건은 무엇입니까? ‘내가 가장 아끼는 물건’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 가장 아끼는 물건은 무엇인가? ○ 왜 그 물건을 아끼는가? ○ 어떻게 그 물건을 가지게 되었는가?	
26주차	‘한국의 첫인상’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 한국에 언제 왔는가? ○ 시내, 길거리는 어떤 모습이었는가? ○ 한국 사람들은 어땠는가? ○ 한국 음식은 어땠는가? ○ 가장 기억에 남는 것은 무엇인가?	
28주차	여러분은 고치고 싶은 생활 습관이 있습니까? ‘고치고 싶은 나의 생활 습관’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 나의 나쁜 생활 습관 ○ 습관 때문에 생기는 불편하거나 안 좋은 점 ○ 습관을 고치기 위해 해야 할 일	
30주차	‘잊지 못할 추억’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 어떤 추억인가요? ○ 왜 지금까지 기억에 남아 있는가? ○ 언제 그 추억이 떠오르는가?	
32주차	‘나의 한국 생활’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 한국에 온 지 얼마나 되었는가? ○ 왜 한국에 오게 되었는가? ○ 한국에서 가장 재미있었던 일이 무엇인가? ○ 한국에서 가장 힘들었던 일이 무엇인가?	

종적 자료 수집 시기	문제	수준
34주차	<p>‘살고 싶은 집’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 어디에 살고 싶은가? ○ 어떤 집에 살고 싶은가? 왜 그런가? ○ 집은 어떻게 꾸미고 싶은가? ○ 집에서 누구와 무엇을 하고 싶은가? 	
36주차	<p>고향의 음식을 소개하는 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.</p> <ul style="list-style-type: none"> ○ 음식 이름 ○ 주로 언제 먹는 음식인가? ○ 어떻게 만드는가? ○ 한국 음식과 비슷한 음식이 있는가? 어떤 점이 비슷하고 어떤 점이 다른가? 	
38주차	<p>여러분은 어떤 사람처럼 되고 싶습니까? 왜 그 사람처럼 되고 싶습니까? ‘내가 닮고 싶은 사람’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 닮고 싶은 사람은 누구인가? ○ 왜 그 사람처럼 되고 싶은가? ○ 그 사람처럼 되기 위해서 어떻게 해야 하는가? 	
40주차	<p>취업을 하려고 합니다. 무슨 일을 하고 싶은지 생각해 보고 자기소개서를 쓰십시오.</p> <ul style="list-style-type: none"> ○ 살아온 과정 ○ 성격의 장단점 ○ 지금까지의 경험 또는 경력 ○ 앞으로의 계획 	
42주차	<p>‘내가 생각하는 행복’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <ul style="list-style-type: none"> ○ 행복을 위해서 어떤 노력을 하는가? ○ 언제 행복하다고 느끼는가? ○ 행복은 무엇이라고 생각하는가? 	고급
44주차	<p>‘절약과 저축’이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.</p> <ul style="list-style-type: none"> ○ 절약하기 위해 무엇을 하는가? ○ 저축을 하고 있는가? 왜 그런가? 	

종적 자료 수집 시기	문제	수준
	○ 돈을 모으면 무엇을 하고 싶은가?	
46주차	<p>‘텔레비전이 우리 생활에 미치는 영향’이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.</p> <p>○ 텔레비전을 자주 보는가? 왜 그런가?</p> <p>○ 무슨 프로그램을 자주 보는가? 왜 그런가?</p> <p>○ 텔레비전이 우리 생활에 미치는 긍정적인 영향은 무엇인가? 부정적인 영향은 무엇인가?</p>	
48주차	<p>‘효과적인 자녀 교육법’이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.</p> <p>○ 현재 자녀가 있는가?</p> <p>○ 자녀가 말을 듣지 않을 때 어떻게 하는가? (현재 자녀가 없는 경우, 부모의 말을 듣지 않는 아이를 어떻게 하면 좋을까?)</p> <p>○ 자녀와 대화를 잘하려면 어떻게 해야 하는가?</p> <p>○ 어떤 부모가 되고 싶은가?</p> <p>○ 자녀 교육을 어떻게 하고 싶은가?</p>	
50주차	<p>여러분은 성공이 무엇이라고 생각하십니까? 그리고 그러한 성공을 이루기 위해 필요한 것이 무엇이라고 생각하십니까? 이와 관련된 자신의 견해를 서술하십시오. 단, 아래에 제시한 내용이 모두 포함되어야 합니다.</p> <p>○ 내가 생각하는 성공이란 무엇인가?</p> <p>○ 그것을 이루기 위해 필요한 것은 무엇인가?</p> <p>○ 그 이유는 무엇인가?</p>	
52주차	<p>여러분은 무엇이 선의의 거짓말(좋은 거짓말)이라고 생각하십니까? 어떤 경우에 그런 거짓말을 할 수 있다고 생각하십니까? 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.</p> <p>< 선의의 거짓말이란 ></p> <p>○ 선의의 거짓말이란 무엇인가?</p> <p>○ 선의의 거짓말은 언제 필요한가?</p> <p>○ 선의의 거짓말이 가질 수 있는 문제점은 무엇인가?</p>	
54주차	<p>‘노후 준비’라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.</p>	

종적 자료 수집 시기	문제	수준
	<ul style="list-style-type: none"> ○ 노후 준비가 왜 필요한가? ○ 노후를 위해 무엇을 준비해야 하는가? ○ 여러분은 노후를 어떻게 준비하고 있는가? 	
56주차	<p>‘칭찬은 고래도 춤추게 한다’는 말처럼 칭찬에는 강한 힘이 있습니다. 그러나 칭찬이 항상 긍정적인 영향을 주는 것은 아닙니다. 아래의 내용을 중심으로 칭찬에 대한 자신의 생각을 쓰십시오.</p> <ul style="list-style-type: none"> ○ 칭찬이 미치는 긍정적인 영향은 무엇입니까? ○ 부정적인 영향은 무엇입니까? ○ 효과적인 칭찬의 방법은 무엇입니까? 	
58주차	<p>‘여성의 사회적 지위와 역할’이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.</p> <ul style="list-style-type: none"> ○ 여성이 일을 해야 한다고 생각하는가? 왜 그런가? ○ 한국에서 여성의 사회적 지위는 어떻다고 생각하는가? 고향과 비교해서 높은 편인가? 낮은 편인가? ○ 여성이어서 좋은 점 혹은 좋지 않은 점이 있다고 생각하는가? 	
60주차	<div style="border: 1px solid black; padding: 10px;"> <p>오늘날 직업에 대한 생각은 크게 두 가지로 나뉘는 것 같다. 하나는 여러 방면으로 사회에 도움을 주거나, 공헌할 수 있는 직업을 택해 봉사하는 마음으로 일하고, 그것을 통해 얻어지는 대가로 자신과 가정을 꾸려 나가는 것이다. 다른 하나는 사회에 대한 봉사나 공헌보다는 일에 대한 자기만족과, 욕구 충족, 충분한 대가에 더 큰 비중을 두는 경우이다.</p> <p>전자의 경우, 일이 힘들거나 보수가 적다고 하더라도 일에 대한 보람과 긍지 때문에 쉽게 그 일을 그만두거나 직업을 바꾸려 생각은 하지 않는다. 하지만 후자의 경우는 일에 대한 즐거움이나 자기만족, 충분한 보상이 뒤따르지 않는다고 판단될 때는 언제라도 직장을 옮길 마음의 준비가 되어 있다. 전자의 경우에는 사회를 안정시키는 데에 기여를 하지만 보수적 경향으로 사회적 분위기를 다소 침체시킬 수도 있다. 후자의 경우에는 생동감은 있으나 급격한 변화로 안</p> </div>	

종적 자료 수집 시기	문제	수준
	<div>정감을 잃어버릴 위험이 많고, 이런 변화 속에 적응하지 못하는 이들은 사회 변화의 뒷전으로 밀려날 수밖에 없게 된다.</div> <p>위의 글에 나타난 두 가지 유형의 직업관 중 자신의 생각은 어느 쪽인지 말하고, 그 이유를 설득력 있게 글로 나타내시오. (200자 내외)</p>	

2. 구어

수집 시기	발화 유형	주제	수준
02주차	인터뷰	자기소개	초급
04주차	발표	가족	
06주차	인터뷰	주말	
08주차	발표	선물(받은 선물, 준 선물, 받고 싶은 선물, 주고 싶은 선물 등)	
10주차	인터뷰	계절	
12주차	발표	취미	
14주차	인터뷰	친구	
16주차	발표	자주 가는 장소	
18주차	인터뷰	올해 계획	
20주차	발표	고향	
22주차	인터뷰	나의 꿈과 미래 계획	중급
24주차	발표	소중한 것들	
26주차	인터뷰	한국(첫인상, 한국에 대한 여러 가지 생각 등)	
28주차	발표	습관	
30주차	인터뷰	추억(어린 시절, 학창 시절 등)	
32주차	발표	나의 한국 생활	
34주차	인터뷰	살고 싶은 집	
36주차	발표	음식(고향 음식, 한국 음식, 좋아하는 음식,	

수집 시기	발화 유형	주제	수준
		싫어하는 음식 등)	
38주차	인터뷰	존경하는 인물	
40주차	발표	나의 삶 (성격, 경험 및 경력, 앞으로의 계획)	
42주차	인터뷰	내가 생각하는 행복	고급
44주차	발표	경제문제	
46주차	인터뷰	텔레비전	
48주차	발표	자녀 교육	
50주차	인터뷰	성공적인 삶	
52주차	발표	거짓말	
54주차	인터뷰	노후	
56주차	발표	칭찬	
58주차	인터뷰	남성과 여성	
60주차	발표	직업	

학습자 말뭉치 이주민 자료 수집 과제(중도입국청소년)

1. 문어

- 20주, 40주, 60주차에는 제시된 그림을 보면서 이야기를 만들어 쓰도록 한다. 학생들의 수준에 따라 이야기를 만들어 쓴 후 이야기 내용과 관련된 학생들의 생각이나 경험담을 함께 쓰도록 할 수 있다. 말하기에서도 동일한 자료를 활용하므로 학습자의 수준을 고려하여 말하기 또는 쓰기를 먼저 하고 관련 내용을 확장함으로써 작문과 발화를 최대한 많이 할 수 있도록 유도한다.

종적 자료 수집 시기	문제	수준
02주차	이름이 뭐예요? 어느 나라 사람이예요? 몇 학년이에요? 무엇을 좋아해요? 자기소개를 해 보세요.	초급
04주차	누가 있어요? 무슨 일을 해요? 무엇을 좋아해요? 여러분의 가족에 대해 쓰세요.	
06주차	여러분의 하루 일과에 대해서 쓰세요. 아침에 몇 시에 일어나요? 그리고 무엇을 해요?	
08주차	여러분은 토요일이나 일요일에 무엇을 해요? 어디에 가요? 누구를 만나요? 여러분의 주말 이야기를 쓰세요.	
10주차	여러분은 뭐 하는 것을 좋아해요? 왜 그것을 좋아해요? 그것을 얼마나 자주 해요? 여러분의 취미에 대해서 쓰세요.	
12주차	어느 계절을 좋아해요? 왜 그 계절을 좋아해요? 그 계절에 특별히 무엇을 해요? 좋아하는 계절에 대해서 글을 쓰세요.	
14주차	여러분이 가장 좋아하는 친구는 누구예요? 여러분이 가장 좋아하는 친구에게 편지를 쓰세요.	
16주차	여러분은 어떤 선물을 받고 싶어요? 왜 그 선물을 받고 싶어요? 지금까지 받은 선물 중에 가장 좋은 선물이 뭐예요? 선물에 대해서 글을 쓰세요.	
18주차	여러분의 고향은 어디예요? 고향에서 무엇이 유명해요? 고향을 소개하는 글을 쓰세요.	
20주차	그림을 보고 이야기를 순서대로 써 보세요.	

종적 자료 수집 시기	문제	수준
22주차	여러분은 무슨 음식을 좋아해요? 무슨 음식을 좋아하지 않아요? '나의 식생활'이라는 제목으로 글을 쓰세요.	중급
24주차	여러분이 소중하게 생각하는 물건은 뭐예요? 왜 그 물건이 소중해요? 그 물건을 어떻게 가지게 되었어요? '내가 가장 아끼는 물건'이라는 제목으로 글을 쓰세요.	
26주차	무슨 과목을 좋아해요? 왜 그래요? 여러분이 알고 있는 좋은 공부 방법이 있어요? '나의 공부 방법'이라는 제목으로 글을 쓰세요.	
28주차	여러분은 고치고 싶은 생활 습관이 있어요? 습관 때문에 생기는 불편한 점이 있어요? '고치고 싶은 나의 생활 습관'이라는 제목으로 글을 쓰세요.	
30주차	20년 후에 여러분은 어디에 있을까요? 무엇을 하고 있을까요? 20년 후 자신의 모습을 상상해 보고 '자신의 미래 모습'이라는 제목으로 글을 쓰세요.	
32주차	여러분은 어디에 여행을 가 봤어요? 누구하고 갔어요? 거기에서 무엇을 했어요? 어땠어요? 여러분의 여행 경험에 대해 쓰세요. (가족 여행, 수학여행, 체험 학습 등)	
34주차	한국에 언제 왔어요? 한국에서 가장 재미있는 일은 뭐예요? 한국에서 가장 힘든 일은 뭐예요? '나의 한국 생활'이라는 제목으로 글을 쓰세요.	
36주차	여러분은 어떤 사람처럼 되고 싶어요? 왜 그 사람처럼 되고 싶어요? '내가 닮고 싶은 사람'이라는 제목으로 글을 쓰세요.	
38주차	음식 이름이 뭐예요? 주로 언제 먹는 음식이에요? 어떻게 만들어요? 한국 음식과 비슷한 음식이 있어요? 어떤 점이 비슷하고 어떤 점이 달라요? 고향 음식을 소개하는 글을 쓰세요.	
40주차	그림을 보고 이야기를 순서대로 써 보세요.	고급
42주차	<div> <p>저는 심각한 고민이 하나 있어요. 저는 3학년인데 키가 140cm이고 몸무게는 455kg예요. 저는 키도 작은 것 같고 뚱뚱한 것 같아요. 저도 가수나 탤런트처럼 더 날씬하고 키도 크고 싶어요. 그래서 요즘 다이어트를 하고</p> </div>	

종적 자료 수집 시기	문 제	수준
	<div style="border: 1px solid black; padding: 10px; margin-bottom: 10px;"> <p>있어요. 그리고 저는 눈이 작고 쌍꺼풀이 없어요. 그래서 성형 수술을 하고 싶어요.</p> </div> <p>여러분도 자신의 외모에 대해 고민을 해 봤어요? 여러분은 자신의 아름다움을 위해 어떤 노력을 하고 있어요? 글에서 읽은 친구의 고민에 대한 여러분의 생각을 써 보세요.</p>	
44주차	여러분은 스트레스를 잘 받는 편이에요? 여러분이 지금 받고 있거나 예전에 받았던 스트레스는 뭐예요? 스트레스를 받았을 때 어떻게 해결했어요? ‘스트레스’라는 제목으로 글을 써 보세요.	
46주차	여러분은 텔레비전을 자주 봐요? 무슨 프로그램을 자주 봐요? 텔레비전이 우리 생활에 미치는 긍정적인 영향은 될까요? 부정적인 영향은 될까요? ‘텔레비전’이라는 제목으로 글을 쓰세요.	
48주차	여러분은 거짓말을 한 적이 있어요? 무슨 거짓말을 했어요? 거짓말을 한 후 어떤 일이 일어났어요? ‘거짓말’이라는 제목으로 글을 쓰세요.	
50주차	여러분은 언제 행복해요? 그리고 언제 슬퍼요? ‘행복한 일과 슬픈 일’이라는 제목으로 글을 쓰세요.	
52주차	학교에서 일어난 큰 실수나 사고를 생각해 보세요. 기억에 남는 일을 쓰세요.	
54주차	여러분은 무슨 놀이(게임, 수업 활동 등)를 좋아해요? 어떻게 해요? 여러분들이 좋아하는 놀이 방법을 소개하는 글을 쓰세요.	
56주차	선생님과 부모님께 칭찬을 받아 본 적이 있지요? 언제 칭찬을 받았어요? 무슨 일로 칭찬을 받았어요? 기분이 어땠어요? 칭찬 받은 일에 대해서 글을 써 보세요.	
58주차	지금까지 읽은 책 중에 가장 재미있었던 책이 뭐예요? 무슨 내용이에요? 읽고 무슨 생각을 했어요? ‘내가 가장 좋아하는 책’이라는 제목으로 글을 써 보세요.	
60주차	그림을 보고 이야기를 순서대로 써 보세요.	

2. 구어

- 2주차, 12주차, 22주차, 32주차, 42주차, 52주차에는 제시한 읽기 텍스트를 큰소리로 낭독하도록 한 후 텍스트와 관련된 질문을 하는 방법으로 학생들과 주제에 관한 대화를 간단히 나눈 후에 학생들에 관한 이야기로 대화를 확장해 나간다.
- 20주, 40주, 60주차에는 제시된 그림을 보면서 이야기를 만들어 보도록 한다. 이야기를 다 만들고 난 후에는 교사가 이야기와 관련된 질문을 하여 이야기에 관한 학생의 의견이나 경험 등을 자유롭게 말하도록 한다.

수집 시기	발화 유형	주제	수준
02주차	인터뷰	자기소개	초급
04주차	발표	가족	
06주차	인터뷰	하루 일과	
08주차	발표	주말	
10주차	인터뷰	취미	
12주차	발표	★ 계절 관련 텍스트 읽기 ★ 계절	
14주차	인터뷰	친구	
16주차	발표	선물(받은 선물, 준 선물, 받고 싶은 선물, 주고 싶은 선물 등)	
18주차	인터뷰	고향	
20주차	발표	★ 그림 보고 이야기 만들기 ★ 이야기와 관련한 자유 대화	
22주차	인터뷰	★ 식사 관련 텍스트 읽기 ★ 나의 식사 생활	중급
24주차	발표	소중한 것들	
26주차	인터뷰	공부와 시험	
28주차	발표	생활 습관	
30주차	인터뷰	미래	
32주차	발표	★ 여행 관련 텍스트 읽기 ★ 여행	
34주차	인터뷰	한국	
36주차	발표	존경하는 인물	
38주차	인터뷰	한국 음식과 고향 음식	

수집 시기	발화 유형	주제	수준
40주차	발표	★ 그림 보고 이야기 만들기 ★ 이야기와 관련한 자유 대화	고급
42주차	인터뷰	★ 중등 외모 관련 텍스트 읽기 ★ 성격과 외모	
44주차	발표	스트레스	
46주차	인터뷰	텔레비전	
48주차	발표	거짓말	
50주차	인터뷰	감정	
52주차	발표	★ 실수 관련 텍스트 읽기 ★ 실수	
54주차	인터뷰	놀이	
56주차	발표	칭찬	
58주차	인터뷰	독서	
60주차	발표	★ 그림 보고 이야기 만들기 ★ 이야기와 관련한 자유 대화	

<2주차 읽기 자료>1)

안녕하세요? 저는 송안나입니다. 대한초등학교 학생입니다. 저는 1학년 5반입니다. 우즈베키스탄에서 왔어요. 지금은 대림동에 살아요. 우리 집은 학교 근처에 있어요. 저는 컴퓨터 게임을 좋아해요. 만나서 반갑습니다.

<12주차 읽기 자료>

한국에는 봄, 여름, 가을, 겨울 사계절이 있습니다.

봄에는 날씨가 따뜻합니다. 산과 들에 예쁜 꽃이 많이 핍니다. 사람들은 꽃놀이를 갑니다.

여름에는 날씨가 더워집니다. 비도 많이 옵니다. 사람들은 넓은 바다로 여행을 갑니다. 우리는 수영을 하고 물총 싸움도 합니다. 시원한 팔빙수도 먹습니다.

가을에는 날씨가 선선해집니다. 산에 가서 단풍 구경을 합니다. 빨간 단풍이 참 아름답습니다. 맛있는 과일도 많이 먹을 수 있습니다.

겨울에는 날씨가 추워집니다. 바람도 많이 불고 눈도 옵니다. 겨울에는 따뜻한 옷을 입고 장갑도 갑니다.

저는 사계절 중에서 추운 겨울을 제일 좋아합니다. 눈사람도 만들도 눈싸움도 할 수 있습니다.

1) 중도입국청소년 자료 수집을 위한 과제는 『초/중/고등학생을 위한 표준한국어 교재』(국립국어원), 『KSL 교육과정 진단도구』(국가평생교육원)의 자료를 발췌하거나 개작하였다. 따라서 과제와 함께 제시되는 텍스트와 그림 자료도 두 자료에서 발췌한 것이다.

<20주차 그림 자료>



<22주차 읽기 자료>

우리 엄마는 항상 ‘무엇을 요리할까?’ 하고 고민하십니다. 왜냐하면 나는 햄이나 고기반찬을 좋아해서 김치나 채소 반찬을 잘 안 먹기 때문입니다.

나는 매일 아침 바빠서 아침을 안 먹고 학교에 갑니다. 점심시간에는 제가 좋아하는 반찬이 없으면 점심을 안 먹고 빵이나 과자를 사 먹으러 매점에 갑니다. 그리고 저녁에는 배가 고파서 한꺼번에 많이 먹습니다.

또 나는 밥보다 햄버거나 치킨, 빵, 과자를 좋아하고 물보다 음료수를 더 좋아합니다. 매일 이렇게 내가 좋아하는 음식을 먹고 싶습니다. 그런데 엄마는 “그런 음식만 먹으면 건강에 안 좋아! 아침을 꼭 먹고 반찬을 골고루 먹어 봐!”라고 말씀하십니다.

나는 왜 내가 좋아하는 음식만 먹으면 안 될까요?

<32주차 읽기 자료>

지난주 토요일에 공주에 갔다 왔다. 오전에 도착해서 먼저 간 곳은 공산성이었다. 공산성은 옛날에 전쟁을 할 때 지은 성이다. 이곳은 경치가 매우 좋고 공주 시내도 잘 보였다.

공산성에서 내려오니 12시였다. 배가 너무 고파서 내려오자마자 공원에서 도시락을 먹었다. 그리고 시청에서 무료로 빌려주는 자전거를 타고 송산리 고분군으로 갔다. 그곳에는 벽화들이 많았다. 옛날 사람들은 무덤에 벽화도 그려 놓고 여러 가지 물건도 넣었다. 신기했다.

2시에 무령왕릉도 갔다. 무령왕릉은 생각보다 정말 컸다. 안에 들어갈 수 없어서 아쉬웠다. 주변에서 사진도 찍고 놀다 보니 오후 3시였다. 정문 옆에는 제기차기와 윷놀이를 할 수 있는 곳도 있었다. 거기서 친구들과 제기차기를 하면서 놀았다. 오늘은 사회 시간에 배웠던 곳에 가서 백제 시대 역사 공부를 할 수 있어서 정말 좋았다.

<40주차 그림 자료>2)



2) 그림은 『엄마와 함께 읽어요. 지식 쏙쏙 만화』 (한국간행물윤리위원회, 2010)에서 발췌함.

<42주차 읽기 자료>

저와 제 동생 마리는 쌍둥이 자매입니다. 우리는 머리가 금색이고 피부가 아주 하얗습니다. 키는 다른 친구들에 비해서 아주 큰 편입니다.

우리는 얼굴은 똑같이 생겼지만 성격은 아주 다릅니다. 저는 조용한 성격이라서 나가서 노는 것보다 집에서 책을 읽거나 엄마 일을 돕는 것을 좋아합니다. 그리고 성격이 좀 느린 편이라서 어떤 일을 할 때 천천히 꼼꼼하게 합니다.

그런데 동생은 활발해서 친구들과 같이 운동장에서 노는 것을 좋아합니다. 또한 성격이 급한 편이라서 무슨 일이든지 빨리 하기 때문에 실수를 자주 합니다. 호기심도 많아서 궁금한 것이 있으면 꼭 물어봅니다.

우리는 이렇게 같은 점도 있고 다른 점도 있지만 기쁠 때나 슬플 때나 늘 함께하는 사이좋은 자매입니다.

<52주차 읽기 자료>

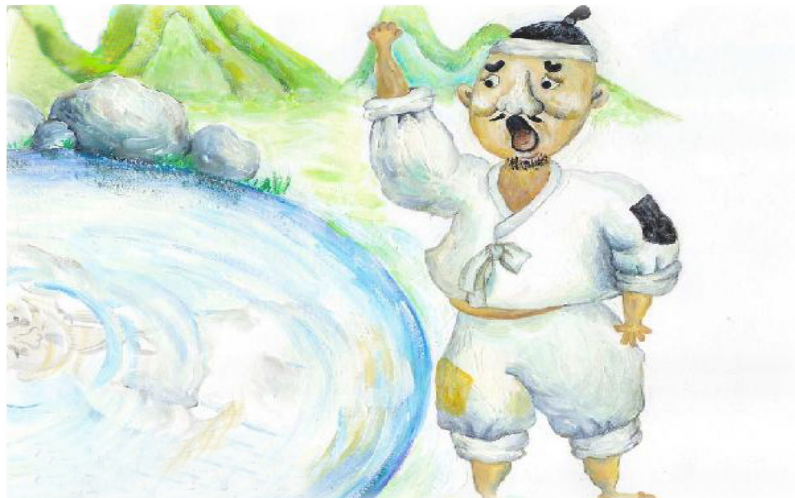
오늘 좀 속상했다. 가장 친한 친구 라몬과 싸웠기 때문이다. 오늘 낮에 라몬과 농구를 하다가 내 실수로 라몬이 넘어졌다. 나는 일부러 한 게 아니라서 미안하다는 말을 안 했는데 그것 때문에 화가 많이 났나 보다. 라몬은 농구공을 던져 버리고, 나한테 소리를 질렀다. 그래서 나도 너무 화가 나서 소리를 질렀다. 그리고 나는 라몬과 크게 싸우게 될까 봐 혼자 집으로 와 버렸다. 그런데 오면서 생각해 보니 내가 일부러 그런 것은 아니지만 나도 넘어지면 기분이 나쁠 것 같다. 어떻게 할까 고민하고 있는데 저녁에 라몬에게서 전화가 왔다. 그리고 나에게 먼저 사과를 했다. 그때 나는 라몬에게 너무 미안했다. 내가 먼저 사과할걸 그랬다. 내일 라몬을 만나면 라몬이 좋아하는 과자를 주면서 다시 한번 사과를 해야겠다.

<60주차 그림 자료>³⁾

※ 다음 그림을 보고 이야기를 만들어 보세요.



1



2

3) 그림은 '키즈짱 잼쟁동화-줍어지는 샘물' 동영상의 주요 장면을 캡처하여 편집한 것임
(<https://www.youtube.com/watch?v=-rHeP6eJKSM>)



3



4



5



6

한국어 학습자 말뭉치 구축을 위한 기획 자료 수집 과제1

1. 문어

- 각 수준별로 다음의 주제와 장르에 따라 글을 쓰도록 하여 수집합니다.
- 다음의 주제를 제시하되 풍부한 글쓰기를 위해 관련 내용을 자유롭게 확장할 수 있도록 합니다.
- 완결된 글이 되도록 하되 글의 길이가 너무 짧지 않은지 확인합니다.
(현재까지의 수집 결과에 따르면 1급 최소 50어절[7~10문장] 이상, 2-3급 100어절 이상[15~20문장], 4급 이상 150어절 이상이 평균 길이임)
- 가족이나 친구, 동료, 한국어 모어 화자의 피드백이 이루어지지 않은 글이 되도록 합니다.

수준	주제	장르
1급	가족, 친구, 취미, 성격, 좋아하는 것과 싫어하는 것, 꿈	생활문
2급	나의 가족, 나의 친구, 나의 이웃	생활문
3급	기억에 남는 여행, 여행 경험	기행문
4급	추천하고 싶은 여행지	설명문
5급	내가 생각하는 성공적인 삶(성공적인 삶이란 무엇인가?)	논설문
6급	결혼에 대한 나의 생각(결혼을 해야 할까? 아니면 혼자 사는 게 좋을까? 한다면 어떤 사람하고 해야 할까? 국제 결혼은 어떨까?)	논설문

2. 구어

- 말하기는 특정 주제에 한정하지 않고 다음과 같은 흐름으로 발화를 진행하도록 합니다. 이때 전개 부분에서는 [나의 현재-나의 과거-나의 미래]의 순으로 이야기를 진행할 수 있도록 적절한 때에 교사가 관련 질문을 해 준다.

단계	발화 내용	비고
도입	▪ 자기소개(이름, 국적 등) - 교사와의 일상적 대화	초급 학습자의 경우 최대한 가능한 주제까지 대화를 이어가도록 함
전개	▪ 나의 현재: 취미, 성격, 좋아하는 것, 싫어하는 것, 가족, 친구 ▪ 나의 과거: 어린 시절, 학창 시절 중 기억에 남는 일 고향 소개, 추천하고 싶은 장소 ▪ 나의 미래: 꿈, 내가 생각하는 성공, 결혼 계획, 결혼에 대한 생각, 죽기 전에 꼭 하고 싶은 일	
마무리	▪ 학습자 격려를 위한 피드백 및 감사 인사	

- 발화 시간은 10분 내외로 합니다. 다만, 1급의 경우 어휘량이 상대적으로 부족한 시기이므로 5분 이상 10분 내외로 시간을 조정할 수 있습니다.
- 초반에는 이름, 국적 등에 대한 질문, 간단한 일상 대화를 통해 학습자의 긴장을 풀어 주면서 발화를 시작하도록 합니다.
- 발화가 시작된 후 교사는 주로 청취자로서 간단한 맞장구를 하며 학습자가 스스로 발화를 이어갈 수 있도록 합니다. 다만, 학습자가 발화를 이어가지 못하거나 도중에 끊길 경우 자연스럽게 발화를 이어갈 수 있도록 관련 주제로 확장을 위한 질문을 해 줍니다.
- 이때 교사의 질문이 지나치게 상세하고 길어지면 학습자가 단답형의 대답을 하게 되는 경우가 많으므로 학습자가 발화를 이어가는 데에 필요한 단서를 제공하는 정도의 질문을 합니다.

한국어 학습자 말뭉치 구축을 위한 기획 자료 수집 과제2

- 수준별로 다음의 과제를 차례대로 진행하도록 합니다.

	초급	중급	고급
준비하기(공통)	자기소개		
과제 1. 제시 자료 보고 말하기	1. 두 장의 제시 그림을 보고 공통점과 차이 점을 비교해서 말하기 2. 두 장소 중 좋 아하는 곳, 그 이유 말하기	1. 한국 전래 동 화 그림 보고 스토리텔링 하 기 2. 이어지는 내용 상상해서 말하 기	1. 비디오 자료 보 고 스토리텔링하 기 2. 이어지는 내용 상상해서 말하기
과제 2. 대화하기	친구와 여행 계획 세우기 - 서로 가고 싶은 여행지 얘기하 기 - 자신이 말한 장 소가 좋은 이유 를 설명하고 여 행지 결정하기 (설명하기, 설득 하기) - 교통편, 숙박, 여행지에서 할 일, 음식 등에 대한 계획 세우 기(제안하기, 의 사 결정하기)	좋아하는 드라마/ 영화에 대해서 이 야기하기 - 좋아하는 드라 마/영화, 최근 에 재미있게 본 드라마, - 출연 배우, 줄 거리, 기억에 남는 장면, - 친구에게 드라 마/영화 추천 하기	결혼에 대한 찬반 의견 나누기

- 교사가 수집 준비, 수집 진행을 돕는 진행자로서 참여할 것을 권장하나 여
건에 따라 학습자가 PPT를 보면서 단독으로 과제를 수행해도 무방합니다.
- 학습자의 자연스러운 언어 사용을 관찰하는 것이 목표이므로 연습을 하거
나 발화할 내용을 메모하여 보면서 말하지 않도록 합니다.

한국어 학습자 말뭉치 구축을 위한 기획 자료 수집 과제3

1. 문어

- 각 수준별로 다음의 주제와 장르에 따라 글을 쓰도록 하여 수집합니다.
- 다음의 주제를 제시하되 풍부한 글쓰기를 위해 관련 내용을 자유롭게 확장할 수 있도록 합니다.
- 완결된 글이 되도록 하되 글의 길이가 너무 짧지 않은지 확인합니다.
(현재까지의 수집 결과에 따르면 1급 최소 50어절[7~10문장] 이상, 2-3급 100어절 이상[15~20문장], 4급 이상 150어절 이상이 평균 길이임)
- 가족이나 친구, 동료, 한국어 모어 화자의 피드백이 이루어지지 않은 글이 되도록 합니다.

수준	주제	장르
초급	주제1. 자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, ……) 주제2. 내가 가장 좋아하는 것과 싫어하는 것 (일, 행동, 말, 사람, 물건, ……)	생활문
중·고급	주제 1. 과학 기술의 발전이 인간의 생활에 미치는 영향 (인터넷, 로봇, 인공지능(AI), ……) 주제 2. 인구 문제와 미래 사회 (저출산, 고령화, 인구 절벽(급격한 인구 감소), 1인 가구, ……)	논설문

2. 구어

- 교사가 수집 준비, 수집 진행을 돕는 진행자로서 참여할 수 있으나, 학습자 스스로 PPT를 보면서 다음과 같은 흐름으로 독백 형식의 발화를 진행하는 것을 권장합니다.

발화 구성	세부 발화 내용
자기소개	간단한 자기소개
과거	태어난 곳, 고향 소개 어릴 때 성격 기억나는 친구 기억나는 일 어릴 때 꿈
현재	사는 곳 성격 좋아하는 것과 싫어하는 것 하는 일 꿈, 진로
미래	앞으로 10년 후의 내 모습 노후의 삶 (65세 이후 어떻게 살고 싶은가) 죽기 전에 꼭 하고 싶은 일

- 발화 시간은 15분 내외로 합니다. 다만, 1급의 경우 어휘량이 상대적으로 부족한 시기이므로 5분 이상 10분 내외로 시간을 조정할 수 있습니다.
- 말하기 전에 관련 내용을 미리 써서 읽지 않으며, 즉흥적으로 발화를 하도록 합니다.

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

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

문의처: 연세대학교 산학협력단 02-2123-4199

☐ 저는 위의 내용을 충분히 이해하였으며 다음의 정보와 말하기/쓰기 자료를 제공하고, 쓰기 원문/말하기 음성 녹음 자료 전체의 공개와 연구 목적의 사용을 허락합니다.

날짜 _____
이름 _____ (서명)

✂-----

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

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

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

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

문의처: 연세대학교 산학협력단 02-2123-4199

☐ 저는 위의 내용을 충분히 이해하였으며 다음의 정보와 말하기/쓰기 자료를 제공하고, 쓰기 원문/말하기 음성 녹음 자료 전체의 공개와 연구 목적의 사용을 허락합니다.

날짜 _____
이름 _____ (서명)
(학습자와의 관계 _____)

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

학교명: _____ 학교 _____ 학년
입학/편입 학년: _____ 학년
(☞ 해당 사항이 없는 경우 쓰지 않아도 됩니다.)

1. 성별: ☐ F ☐ M
2. 출생년: _____ 년(예. 1989년)
3. 현재 등급: _____ (TOPIK: _____)
4. 국적: _____ (※ 교포 여부 ☐ 교포 ☐ 외국인)
5. 제1 언어: _____
6. 한국어 학습 기간(한국어를 얼마 동안 공부했습니까?): _____ 년 _____ 개월
(예. 1년 3개월)
- 6-1. 학습 기관명: _____
- 6-2. 사용 교재명: _____

7. 7-1. 입국년월: _____년_____월(예. 2015년 2월)

7-2. 한국에서 얼마 동안 살았습니까?: _____년_____월(예. 1년 3개월)

8. 한국어 학습 목적

☐ 진학 ☐ 취업 ☐ 거주 ☐ 취미 ☐ 결혼 ☐ 기타 ()

9. 직업: _____

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

11. 평상시에 가장 많이 사용하는 언어는 무엇입니까? _____

12. 한국어로 대화하는 상대는 누구입니까?

☐ 부모님 ☐ 시부모님 ☐ 남편 ☐ 친척
☐ 이웃 ☐ 친구 ☐ 선생님 ☐ 직장 동료 ☐ 기타 ()

13. 한국어로 말하는 시간은 얼마나 됩니까?

☐ 거의 없음 ☐ 하루 1시간~하루 3시간 ☐ 하루 3시간~ 5시간 ☐ 하루 5시간 이상

14. 한국어로 듣는 시간은 얼마나 됩니까?

☐ 거의 없음 ☐ 하루 1시간~하루 3시간 ☐ 하루 3시간~ 5시간 ☐ 하루 5시간 이상

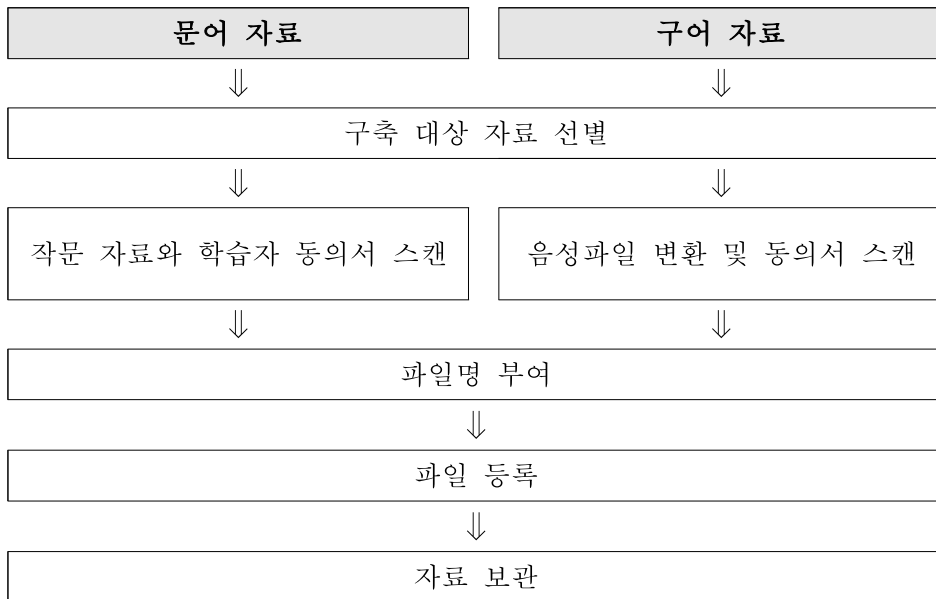
15. 한국어로 나오는 방송 매체(TV, 라디오, 인터넷 동영상)를 보는 시간은 얼마나 됩니까?

☐ 거의 없음 ☐ 하루 1시간~하루 3시간 ☐ 하루 3시간~ 5시간 ☐ 하루 5시간 이상

한국어 학습자 말뭉치 자료 처리 지침

1. 자료 처리 절차

- 자료 처리는 파일을 전산화하여 말뭉치 자료로서 본격적인 구축과 가공 작업을 하기 위한 전처리 단계로 다음과 같은 절차에 따라 처리한다.



2. 단계별 자료 처리 지침

1) 말뭉치 구축 대상 자료 선별

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

문어	구어
<ul style="list-style-type: none"> ○ 학습자 동의서에 서명한 자료 ○ 학습자 동의서의 개인 정보 모두 입력된 항목 선정 ○ 동일 학습자의 자료 2개 이하로 제한 ○ 영어권, 일본어권 자료/1, 5, 6급 단계의 자료 우선 선정 	
<ul style="list-style-type: none"> ○ 완결된 텍스트 작문 자료 선정 ○ 텍스트의 길이 평균 100어절 이상의 자료 선정. 단, 숙달도 단계를 고려하여 1, 2급은 50어절 내외의 자료를 포함함 ○ 복사 또는 스캔 파일의 경우 화질이 좋은 자료 선정 	<ul style="list-style-type: none"> ○ 완결된 담화 단위의 발화 자료 선정 ○ 발화 길이 2분 이상의 자료 선정 ○ 음질이 좋은 자료 선정 ○ 교사의 개입이 많지 않고 학습자의 발화가 중심인 자료를 우선 선정

2) 학습자 동의서 확인 및 스캔

- 학습자 동의서와 작문 자료가 제대로 짝을 이루고 있는지 확인한다. 학습자 동의서나 작문 자료 어느 한쪽이라도 누락된 자료는 구축 불가 자료로 분리하여 따로 모은다.
- 자료를 확인하는 과정에서 수집 기관에 문의가 필요한 사항이나 수집 시 주의해야 할 사항이 있을 경우 학습자 동의서 및 자료 관련 특이 사항에 메모를 남긴다.

접수 날짜	기관명	자료 유형	자료 내용	자료 수(수준별)						합계	자료 관련 메모
				1급	2급	3급	4급	5급	6급 이상		
2015.07.14	동국대학교(경주)	일반	여름 학기 중간고사 쓰기 자료							105	
2015.07.21	한남대학교	일반	여름 학기 중간고사 쓰기 자료	13	12	23	8			56	
2015.07.22	한양대학교	일반	여름 학기 중간고사 쓰기 자료	39	71	78	70	43	9	310	
2015.07.23	호남대학교	일반	여름 학기 기말고사 쓰기 자료							124	
2015.07.24	충남대학교	일반	여름 학기 중간고사 쓰기 자료							466	

동의서 급수와 시험 자료의 급수가 맞지 않아 수집 기관에 확인이 필요함

동의서 급수와 시험 자료의 급수가 맞지 않아 수집 기관에 확인이 필요함

3) 일련번호 부여

- 학습자 동의서와 작문 자료에 일련번호를 부여한 후 스캔한다. 동일한 학습자가 작성한 학습자 동의서와 작문 자료는 같은 일련번호를 부여한다. 이때 학습자 동의서가 두 장으로 분리된 경우는 각각을 '0001-앞, 0001-뒤'로 처리하고, 작문 자료가 두 장 이상일 경우는 '0001-01, 0001-02, 0001-03……'으로 처리한다.

4) 파일명 부여

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

예) 국내일반종적_문어_중국_○○대_1급_0001_01.txt

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

구분	범주	설명	항목	코드
자료 코드	자료 및 학습자 유형	학습자의 특성에 따른 분류	일반 종적 이주 학문 목적 (2017년 현재 학문목적은 일반 최고급으로 분류)	국내일반횡적 국내일반종적 국내일반기획 결혼이주횡적 결혼이주종적 결혼이주기획 이주노동횡적 이주노동종적 이주노동기획 중도입국횡적 중도입국종적 중도입국기획 국외횡적 국외종적 국외기획
	자료 유형	자료의 유형을 구분하는 코드 부여	문어(Written) 구어(Spoken)	문어 구어
학습자 정보 코드	언어권	학습자의 제1 언어를 구분하는 코드 부여	중국어 일본어 베트남어 영어 ...	중국 일본 베트남 영어 ...
	자료 수집 기관	자료 수집 기관명	서울대 경희대 ...	서울대 경희대 ...
	수준	학습자의 수준을 구분하는 코드 부여	1급 2급 3급 4급 5급 6급 최고급	1 2 3 4 5 6 7
	학습자 구분 번호	기관의 학습자 구분을 위한 일련번호	0001 0002 ...	0001 0002 ...

자료 번호	자료 번호	동일한 학습자가 두 개 이상의 자료를 제공할 경우 자료를 구분하기 위한 일련번호	01 02 ...	01 02 ...
----------	-------	---	-----------------	-----------------

한국어 학습자 말뭉치 문어 입력 지침

1. 전체적인 형식 원칙

- 기본적으로 온라인 입력/전사 시스템의 입력 창에서 입력한다.
- 자료를 입력하기 전 표본 정보와 학습자의 개인 정보를 입력한다.
(☞ ‘수집 정보 등록/검증’ 메뉴)
- 학습자가 글 하나를 스스로 완성하였을 경우에만 입력하는 것을 원칙으로 한다. 중간에 채 완성하지 못한 문장은 입력하지 않는다.

나는 영화를 TV에서 방송할 때 특정 장면을 삭제하는 것을 반대하는
입장이다.

첫 번째 이유는 방송 심의의 기준이 모호하는데 방송 회사에 따라서
방송 심의가 다르다. 똑같은 장면인데 이 채널에서 못 보지만
다른 채널에서 볼 수 있기 때문에 삭제하는 것이 효과가 없다고
생각한다.

두 번째는 특정 장면보다 영화 전체의 메시지가 더 중요한데
만약에 그 장면을 삭제하면 사람들이 그 영화를 제대로 볼 수
없다. 단는 폭력이나 야한 장면이 있는 영화 보통 모두 표시구에
네

- 필적을 알아보기 어려운 것은 일단 가장 가까운 상태로 입력한다.
- 단락을 구분하여, 문장 단위로 입력한다. 단락은 자판의 엔터키로 구분하
고, 들여쓰기는 반영되지 않는다.
- 전체 본문 입력이 끝나면 ‘주석 자동 생성’을 클릭하여 본문 주석을 확인
하고 이후 개별 마크업을 진행한다.

2. 입력 지침

- 원본의 텍스트를 그대로 입력하는 것을 원칙으로 한다. 철자 오류가 있더라도 원본 그대로 입력한다.

<예> 특히 말할 때 춘대말을 한다는 것이 자주 반말을 말한다.
→ 수정 안 함.

- 원문의 영어와 한자는 모두 유지한다. 한자는 시스템 입력창에서 글자를 선택 후 마우스 오른쪽을 클릭하여 입력한다.
- 띄어쓰기는 어문 규범과 <표준국어대사전>의 표제어에 맞춰 수정하여 입력한다. 원활한 형태소 분석 작업을 위해 띄어쓰기를 정확히 적용한다.
- 분수 표시는 다음과 같이 입력한다.

<예> 1/2, 3/4

- 영문자, 한글 자모, 괄호 문자 등은 자판을 사용하여 입력한다.

<예> ㄱ ㄴ ㄷ ㄹ, (1) (2) (3)

- 외국어를 함께 쓴 경우 다음과 같이 원문에 따라 병기한다. 단, 입력과 해석의 용이성을 고려하여 영어와 한자에 한정한다.

<예> 아래의 경우 '바프라이(BARFLY)'로 입력한다.

우리는 술을 마시고 싶으면 ^(BARFLY) '바프라이' 술집에 ~~가려고~~ 가요.

- 숫자와 한글 표기를 함께 쓴 경우 원문에 따라 병기한다.

<예> 아래의 경우 '3(세) 달 전'으로 입력한다. 이때 '3달(세 달)'과 같이 동일한 표기가 두 번 이상 입력되도록 하지 않는다.

^(세) 3달 전에 미국에서 한국까지

- 학습자가 작문 중간에 교정 기호를 사용하거나 교정에 관한 문구를 적어 넣은 경우 이를 반영해서 수정 입력한다. 단, 학습자의 답안에 교사가 같

은 색으로 수정 또는 채점을 한 경우, 학습자가 작성하면서 스스로 수정한 것인지 교사가 수정한 것인지 선별해야 한다.

<예> 반 친구도 노래를 잘 볼 수 있어요.
그래서 노래방도 자주 가요.
우리는 함께 때 좋은 기본이 왔는데요.
어떻게 가는지 알아요? 서울까지 피행기를 타야 해요.

- 단락 구분은 하지 않으며 한 행에 한 문장을 입력하는 것을 원칙으로 한다.

3. 문장 부호 및 기호류 마크업

- 문장 부호는 원본 그대로 입력하는 것을 원칙으로 한다.
- 문장 부호 및 기호류는 기본적으로 자판 문자(기호)를 입력하며, 한글 워드 프로그램 등에서 사용하는 전각 기호나 반각 기호를 사용하여 입력하지 않도록 한다.
- 문장 부호는 학습자가 적어 넣은 대로 입력한다. 즉, 문장 부호의 누락이나 생략, 중복 등을 그대로 반영한다. 단, 학습자가 부적절한 위치에 습관적으로 찍은 온점은 문장 부호로 보기 어려우므로 반영하지 않는다.
- 입력이 어려운 문자는 거꾸로 된 물음표(?) 기호를 사용하여 입력한다. 거꾸로 된 물음표(?) 기호는 키보드에 없는 문자, 식별되지 않는 문자 등 기본 자판에서 입력 불가능한 모든 문자와 기호 형태를 의미한다.
 - ‘외국 문자’는 영어와 한자 이외의 외국어를 입력할 때 ¿ 기호 입력 후 마크업할 때 사용한다.

<예> <EX_Alpha>ㄱㄴㄷㄹ</EX_Alpha>

- ‘식별 불가’는 원본에서 다양한 이유로 확인이 어려운 문자나 기호에 대해 ㄱ로 입력 후 마크업한다.

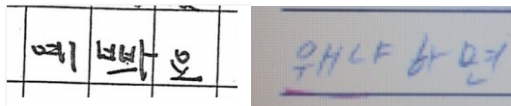
<예> <CNI>ㄱ</CNI>

- ‘기타 기호’는 문장 앞에 붙인 블릿 기호나 다른 특수 기호들을 원본 그대로 입력 후 마크업할 때 사용한다. 키보드에서 한글 자음을 입력 후 ‘한자’키를 눌러서 선택하여 입력한다. (‘기타 기호’는 원본 그대로 입력하므로 ㄱ기호를 입력하지 않도록 주의한다.)

<예> 1) <EX_Symbol>『』「」《》</EX_Symbol>
: [괄호기호] ‘ㄴ’ 입력 후 ‘한자키’ 눌러서 선택
2) <EX_Symbol>※★</EX_Symbol>
: [일반기호] ‘ㄴ’ 입력 후 ‘한자키’ 눌러서 선택
3) <EX_Symbol>m’ kg kcal</EX_Symbol>
:[단위기호] ‘ㄴ’ 입력 후 ‘한자키’ 눌러서 선택

- 두벌식 한글과 같이 자판에서 하나의 음절이나 글자로 입력이 불가능한, 우리말에 없는 글자를 입력할 때에는 해당 글자의 위치에 거꾸로 된 물음표(?)를 입력한 후, 구축 도구 내의 ‘한글 기호’ 주석을 사용하여 마크업한다.

<예>



- 예쁘ㅏ요 : 시스템 ‘예ㄱ요’로 입력 후 ‘한글기호’ 마크업 처리
- 예<NSS>ㄱ</NSS>요
- 우애나하면 : 시스템 ‘ㄱ나하면’으로 입력 후 ‘한글기호’ 마크업 처리

- <NSS>ㄱ</NSS>냐하면
- 3) 좌우대칭된 ㅏ이 포함된 ‘가’
- <NSS>ㄱ</NSS>

- 기호류 중 자주 사용되는 ‘가운뎃점’은 별도의 마크업 없이 입력/전사 창 아래에서 바로 클릭하여 입력한다.

<예> <MP> • </MP>

4. 익명성 보장을 위한 개인 정보의 처리

- 학습자들의 이름, 외국인 등록번호, 카드 번호, 전화번호 등은 신분 보장을 위해 실제 입력 정보에 ‘개인 정보’ 태그로 마크업한다. 이렇게 마크업이 된 정보들은 기호로 자동 처리되어 공개되지 않는다.
- 다음은 마크업 과정에서 각각의 정보를 대신하는 태그들이다.

- 이름 : 사람 이름, 단체 이름, 학교 이름 등 ⇨ <Privacy_Name> 태그

<예> 저는 태국에서 온 <Privacy_Name>사일롬</Privacy_Name>입니다.

- 전화번호 : 학습자의 휴대폰 번호 등 ⇨ <Privacy_PhoneNum> 태그
- 카드 번호 : 학습자의 개인 신용카드 번호 등 ⇨ <Privacy_CardNum> 태그
- 기타 : 개인 식별 번호(주민등록번호, 외국인등록번호, 학번 등), 주소 등 ⇨ <Privacy_Etc> 태그

<예> 저는 서대문구 신촌동 <Privacy_Etc>135</Privacy_Etc> 번지에 삽니다.

5. 기타

- 스캔 과정에서 일부분이 잘린 경우, 잘린 부분이 한두 글자, 또는 한두 단

어 이내로 누가 봐도 추정 가능한 내용일 경우에는 해당 내용을 적어 입력한다. 그 외에는 입력 대상에서 제외한다.

6. 최고급 자료 마크업

- 최고급 자료의 입력은 기존 지침을 동일하게 적용한다.
- 기존의 마크업과 더불어 형식과 내용을 구분하기 위해 아래의 마크업을 사용한다.

	주석	내용	주석 표시
형식 구분	보고서 제목	전체 보고서의 제목	<head>
	본문앞	앞부분의 부속물	<front>
	본문	여러 개의 장절 제목과 본문	<body>
		장절 제목	<title> (기존 주석)
	본문뒤	뒷부분의 부속물	<back>
내용 구분	국문 초록	한글로 된 초록 및 주제어	<Korads>
	외국어 초록	외국어로 된 초록 및 주제어	<Forabs>
	각주 미주	주석 내용	<ft>
	예문 인용	단락이 구분되어 제시된 인용 구절과 예시문	<q>
기타	그림 그래프 도표 설명	문어 입력 과정에서 표, 그림, 그래프 수식 등의 생략을 나타내 주는 표시	<gap reason>

- 각주 미주: 본문과 각주 내용에 각각 각주 표시 1), 2)를 남기고 해당 각주

- 는 본문 뒤, 참고문헌 앞으로 이동 후 <ft> 태그
- 예문 인용: 본문 내에서 문단 구분되어 하나의 단락으로 삽입된 부분을 <q>태그

<예>

중국인 학자인 劉爲는 당시 조선국내와 대청무역에서 유통하는 銀에 대해서 아래와 같이 설명한다.

"조선은 일본 白銀을 萊銀이라고 불렀는데 그 은의 순도가 80% 이상이다. 그 외에 조선에서 유통하는 백은은 또한 순도가 90% 이상의 청나라산 天銀이 있고 순도가 70%-80% 정도의 丁銀이 있다."

그러나 위의 인용문에는 틀린 부분이 있다. 첫 번째는 天銀이란 것은 淸國產 은이 아닌 朝鮮產의 순도가 높은 은인 것이다. 1766년에 북경에 다녀온 홍대용과 1783년에 심양에 다녀온

중국인 학자인 劉爲는 당시 조선국내와 대청무역에서 유통하는 銀에 대해서 아래와 같이 설명한다.

<q>

<sentence>"조선은 일본 白銀을 萊銀이라고 불렀는데 그 은의 순도가 80% 이상이다.</sentence>

<sentence>그 외에 조선에서 유통하는 백은은 또한 순도가 90% 이상의 청나라산 天銀이 있고 순도가 70%-80% 정도의 丁銀이 있다."</sentence>

</q>

그러나 위의 인용문에는 틀린 부분이 있다.

첫 번째는 天銀이란 것은 淸國產 은이 아닌 ……

- 보고서의 장절 제목은 <title>처리한다.

<예>

<body>

<title>1. 서론 </title>

<sentence>본 연구의 목적은…

- 문장 중간에 나타나는 계산식이나 그림은 작업자 메모를 달아 준다.

한국어 학습자 말뭉치 구어 전사 지침

I. 구어 전사 기호 체계

대분류	소분류	기호	예시
억양 단위	하강	.	2:네.
	상승	?	2:어디 갈 거예요?
	약한 상승이나 하강	,	1:그래서 그랬는데 이번에,
	활기, 기운찬 어조	!	1:아!
	하나의 억양 단위가 끼어들 에 의해 끊어진 경우	-	6:기자가 와서 - 2:응. 6:- 그 사람한테 인터뷰를 시작했어.
	두 억양 단위가 휴지 없이 이어질 경우	&	3:요거는 교수 학습의 개요지, &요 표는, 4:아::,
겹침 현상	겹침 현상	엘란에서 자동 표시	1 03:49.2 03:51.3 네. 다 거짓말이기 때문에. 2 03:50.8 03:52.2 아 왜 거짓말을 하나요? ☞ 발화 겹침이 있음을 알 수 있음
잘 들리지 않는 부분	잘 들리지 않는 부분	<X X>	<X보통X>
	전혀 들리지 않는 부분	<note>안들림</note>	1: 거기까지 <note>안들림</note> 2:<note>안들림</note> 너무한 거 같더라.
	들리지 않는 음절수만큼	X	2: 근데 그거 진짜 XX해야 되겠더라
전사자의 설명	-	<note>여기에 전사자의 설명을	1:응.<note>장소 이동으로 인해 잠시 멈춤</note>

대분류	소분류	기호	예시
		입력하세요</note>	2:우리 때는 그런 거 없었잖아.
혼잣말	-	<monologue></monologue>	<monologue>미치겠네.</monologue>
표기 지침	구어의 발음 특성, 개인의 발음 특성, 지역적인 특성 등에 의해 철자법대로 소리 나지 않는 발음(표준 발음이 아닌 경우)	소리 나는 대로 적고, 원래의 형태 없이는 내용을 이해하기 어려울 때는 () 안에 학습자의 발화를, 괄호 밖에는 규범 표기를 밝힘	1: 친구와(칭구와) 강남에(간남에) 갔습니다.(갔슨니다.)
	숫자 표기	발음에 따라 한글로 표기, 만 단위 이상도 모두 붙여 씀	7:오늘 제 동생이 이케 하나 오백 원이라고 사 가지고 왔더라구.
	외래어·외국어 표기	발음 규범에 따라 한글로 표기	2: 이거 크림 장난 아니야. 1:이거도 오리지날 제주도 감귤이 아니야.
	끊어진 단어(불완전하게 발화된 단어)	=	4:사실 학습 자료랑 학= 형태는 떨어져도 되는데.
	한 어절 발화 도중 다른 억양 단위로 전사될 때 조사나 어미에,	=	1:주부 우울증,=이라고 말할 수 있겠습니다.
	띄어쓰기	맞춤법에 따름	
	축약형	'(apostrophe, 영문따옴표)를 사용해서 두 음소를 연결	사귀'어, 바뀌'어, ...
	표현적 장음	::	1:많은 경우에:: 논문, 저::~ 어::~ 연구는 네이션,

대분류	소분류	기호	예시
	담화표지	~	1:그::~~ 음::~~
준음성	웃음	<vocal desc='웃음'>	6:어우 야<vocal desc='웃음'>
	기침	<vocal desc='기침'>	-
	하품	<vocal desc='하품'>	-
	재채기	<vocal desc='재채기'>	-
	목청 가다듬는 소리(음, 으음)	<vocal desc='목청가다듬 는소리'>	-
	들이마시는 숨(쓰)	<vocal desc='들이마시는 숨'>	-
	내쉬는 숨(후우)	<vocal desc='내쉬는숨'>	-
	혀 차는 소리(쓰)	<vocal desc='혀차는소리' >	-
	헛기침(에 험)	<vocal desc='헛기침'>	-
	한숨	<vocal	-

대분류	소분류		기호	예시
			desc='한숨'>	
	노래		<vocal desc='노래'>	-
	웃으면서 말하는 부분		<@ @>	2:<vocal desc='웃음'>너무 좀 <@오버한다.@>
	박수치면서 말하는 부분		<# #>	5:우우 <vocal desc='박수'> <#이리 와 이리 와.#>
	노래를 부르는 부분		<M M>	<M동해물과 백두산이 마르고 닳도록M>
	박수나 손가락 부딪치는 소리		<kinesics desc=' '>	<kinesics desc='박수'>
	대화 흐름에 영향을 주는 전화벨 소리라든지 기타 음성 아닌 소리		<event desc=' '>	<event desc='전화벨소리'>
2차 전사	구어적 변이형	() 안에는 학습자의 발 화를, () 밖 에는 철자 보 충. 단, 주 등 장하거나 쉽 게 원래 형태 로 이해될 수 있는 것들은 일일이 철자 형을 붙여주 지 않음	()	1:책상 위에 놔 뒀. → 놓아(놔) 로 전사하지 않고 변이형 그대로 전사해도 무방함
	발음 오류	한국어 모어 화자의 발음과 음운적으로 구분이 될 정도로 발음에 오류가 있는	()	저는 어제 밥을 먹었어요.(먹어요.)

대분류	소분류	기호	예시
	경우		
	한국어 모어 화자의 발음과 음성 혹은 변이음의 구분이 모호한 경우	()	가구(가구)<note>‘가’의 ㄱ에서 유성음으로 발음</note> 밥을(밥을)<note>ㄹ 권설음</note>
	음운규칙으로 인해 한국어의 철자대로 발음되지 않으나, 학습자가 이를 철자대로 발음하는 경우와 같이 철자 전사를 통해 학습자의 발음 오류를 반영하기 어려운 경우	()	무조건(무조건)<note>경음화 안됨</note> 같이(같이)<note>구개음화 안됨</note>
	외국어, 외래어 발음	()	인터뷰(인틸뷰)<note>원어식 발음</note> 외래어표기법을 우선으로 하여 교정어절을 제시하고, 학습자의 원어를 살려 적는 것을 원칙으로 한다.
			카페: 현실 발음 [까페], 학습자 발음 [카페] 버스: 현실 발음 [빠스], 학습자 발음 [버스]

대분류	소분류		기호	예시
				→ 각각 ‘카페(카페)’, ‘버스(버스)’로 처리함 카페(카페)<note>현실발음 과 다름</note>
	방언형 표시		확실한 방언형(대응하는 표준형 형태소가 없는 것)의 경우 태그 부착	차는 <dia>여일</dia> 있어.
	긴 휴지		(1초 이상의 쉼은) 0.1초 단위까지 표시	2:{1.2}그럴까?
	짧은 휴지		한 어절 안에서의 짧은 쉼은 ‘.’로 표시	2:아::~ 그리고 어::~ 남의 의견을 잘 듣고 수용하고 대화..로 타협해야 된다고 하면서,
	인용		<Q Q>	후자들은 현대 사회에 대하여 <Q불확실성의 시대는 아니다.Q> 라고 말하죠.
	텔레비전 방송이나 강의 등 텍스트 종류 표현		<R R>	1:그다음에 <R생각과 느낌이 유기적으로 잘 짜여져 조직체를 이룰 때 좋은 글이 될 수 있다R>라고 돼 있어요.
	익명성 보장을 위한 마크업		<name> : 사람 이름, 단체 이름, 학교 이름 등 <social security number> : 주민등록번호 <card-num> :	5: 그게 어찌면 <name1> 선배님이라든지 다른 선배님들 말:: 들은 걸 생각해 보면,

대분류	소분류	기호	예시
		신용카드 번호 <address> : 주소 <tel-num> : 전화 번호	
세그멘테이션	분할 단위는 어절 단위로 한다		내{1.2}/가 (X) 내{1.2}가 (O) *빗금은 세그멘테이션 분할을 나타내는 표시이다.

II. 항목별 세부 설명

- 한국어 학습자 말뭉치의 구어 전사 지침은 <21세기 세종 한국어 균형 말뭉치>의 전사 지침을 기초로 하되, 비모어 화자로서 한국어 학습자의 구어 자료에서 나타날 수 있는 여러 가지 발음 표기에 대한 지침 등을 보강하여 수정·보완한 것이다.

1. 전체적인 형식과 규정

1) 발화자 표시

2) 억양 단위

- 구어 자료는 억양 단위 전사를 한다. 다만, 학습에 의한 발화로 모어 화자와 달리 문장 단위 발화가 많고, 발화 길이가 길지 않다. 이러한 특성을 반영하여 통사 구조에 따른 절 단위 혹은 문장 단위의 전사와 억양 단위 전사를 절충하도록 한다.

가. 억양 단위의 개념

- 구어는 문어와는 달리 정보의 흐름이 통사적인 단위로 이루어지지 않는다. 즉, 문어의 기본 단위인 문장은 종결어미로 마무리되고 마침표라는 문장 부호로 인해 명확하게 그 단위를 설정할 수 있지만, 구어는 종결어미를 사용하여 발화를 끝내는 경우가 많지 않고, 억양이나 휴지 등의 운율적인 요소에 영향을 받기 때문에 기본 단위를 운율적인 단위 곧 억양 단위로 본다.
- 억양 단위는 하나의 통일된 억양 윤곽에서 나타난 발화의 연속 단위이다. 단위의 시작에서 기본적인 높이(pitch)로 시작하고, 쉼이 나타나며, 빠른 음절의 연쇄가 나타나는 특징이 있고, 단위의 끝에서는 음절이 길어진다.
- 억양 단위의 구분은 다음과 같이 문장 부호를 사용한다.

하강 억양 .
상승 억양 ?
약한 상승이나 하강 억양 ,
활기에 넘치는 기운찬 어조(감탄의 끝) !

나. 끊어진 억양 단위(붙임표의 사용)

- 계속해서 말을 할 의향이 있는데, 끼어들음을 당해서(혹은 적극적인 호응에 의해서) 말끝이 잘린 경우는 다음의 예와 같이 붙임표(-)를 사용한다(단위의 끝에서는 앞쪽에만, 단위의 시작에서는 뒤쪽에만 스페이스 있음). 시간적 순서에 의해 표현된 발화를 억양 단위로 묶을 수 있게 된다.

<예>	6:그래서 세계의 매스컴에 다 집중이 되면서 기자가 와서 -
	2:응.
	6:- 그 사람한테 인터뷰를 시작을 했어.

- 한 명이 말을 하는 도중에 말을 끊은 것이 아니라 지속적으로 반응을 하는 경우 그 수가 많더라도 모두 반영한다.

<예> 1:어제 인사동에 갔는데 -
 2:네.
 1:- 길에서 공연을 하고 있어서 -
 2:네.
 1:- 보다가 -
 2:네.
 1:- 배고파서 호떡을 사 먹었어요.

다. 억양 단위의 연속성

- 두 억양 단위가 휴지 없이 빨리 이어지는 경우 뒤의 발화 앞에 띄어쓰기 없이 & 기호를 붙인다.

<예> 3:요거는 교수 학습의 개요지, &요 표는,
 4:아:.,

3) 겹침 현상

- 겹침 발화의 표시는 엘란에서 자동으로 처리된다.

4) 잘 들리지 않는 부분

- 잘 들리지 않는 부분은 <X X>안에 전사한다.

<예> 그때도,
 <X보통X> 그런 자만심이 있었다.

- 화자의 발화 내용이 전혀 들리지 않는 부분은 <note>안들림</note>으로 전사한다. 이때는 억양을 확인할 수 없으므로 문장 부호를 넣지 않는다.

<예> 근데 그거 <note>안들림</note>

- 들리지 않는 음절은 그 음절의 수만큼 X를 붙인다.

<예> 근데 그거 진짜 XX해야 되겠더라.

5) 전사 기호의 중복

- 전사 기호가 서로 중복이 될 경우에는 기본적으로 문장 부호를 우선하고 (즉, 어절의 가장 가까운 곳에 붙이고) 웃음이나 박수 등 준음성의 표현을 2차로 한다.

<예> 2:<@길섭이가 그거를 딱 보더니,
허 <Q맛이 없어Q> 그러는 거 알어?@>
7:1<vocal desc='웃음'> 그까 -
2:놀래 애가 놀래 가지고 <Q맛이 없어,Q>
7:- 그게 주식이에요 언니.

6) 전사자의 설명

- 전사자가 전사의 어려움에 대한 것이나 특별한 설명을 붙일 필요가 있을 때는, <note> </note> 태그를 사용하여 표현한다. 녹음 상태, 잡음, 동시 다발대화 설명, 특이한 발음 상태 등의 설명이 덧붙을 수 있다. 주석 태그는 다음과 같이 붙여야 한다.

<예> 1:응.<note>장소 이동으로 인해 잠시 멈춤</note>
2:우리 때는 그런 거 없었잖아,

7) 혼잣말

- 혼잣말은 반영하여 전사하되 혼잣말임을 구분하기 위해 <monologue> </monologue> 태그를 사용하여 표현한다.

<예> <monologue>미치겠네.</monologue>

2. 표기 지침

1) 대원칙

- 발화 내용은 기본적으로 철자법 수준의 전사를 한다. 다만, 구어의 발음 특성, 외국인 학습자의 발음 특성이나 오류, 지역적인 특성 등에 의해 철자법대로 소리 나지 않는 발음(표준 발음이 아닌 경우)에 대해서는 발음 나는 대로 적는다. 이 경우 학습자의 발화는 괄호 안에 밝혀 주되 규범 표기는 괄호 밖에 전사하여 준다.

<예> 1: 친구와(칭구와) 강남에(간남에) 갔습니다(갔습니다)

- ☞ 한국어 학습자 발음치는 일반 발음치와 달리 여러 가지 유형의 발음을 포함하기 때문에 철자법 전사의 수용 범위에 대한 고려가 필요하다. 가령, 위의 문장의 경우 한국어 모어 화자의 음성 자료를 들으면서 전사를 한다면 ‘친구’로 전사하겠지만, 외국인 학습자의 발음이므로 ‘친구’와 적을 것인지 ‘칭구’로 적을 것인지 결정해야 하는 문제가 발생한다. 이 경우 외국인 학습자의 발음에 익숙한 한국어 교사라면 발음 오류를 비교적 쉽게 판단하여 표기에 반영할 수 있지만 일반인의 경우는 어려움이 따른다. 따라서 한국어 모어 화자에게서도 필수적으로 나타나는 음운 변화는 철자대로 전사하고 그 밖의 수의적인 발음은 위의 예와 같이 소리 나는 대로 1차 전사를 한다. 그리고 소리 나는 대로 전사한 표기 형태만으로 그 의미를 파악하기 어려운 경우대로 ()의 밖에 원래의 표기를 보충하여 넣도록 한다. (보충적 표기 관련 지침은 ‘4. 2차 전사/철자법 보충’ 참고)

2) 숫자 표기

- 숫자는 아래의 예에서처럼 발음에 따라 한글로 적는다.

<예> 오늘 제 동생이 이렇게(이케) 하나 오백 원이라고 사 가지고 왔더라고(왔더라구)
: 500원으로 적지 않는다.

3) 외래어·외국어 표기

- 외래어나 외국어는 아래의 예에서처럼 발음에 따라 한글로 적는다.

<예> 2:어떻게 이거 크립 장난 아니야.
1:이거도 오리지널 제주도 감귤이 아니야.

4) 끊어진 단어(불완전하게 발화된 단어)

- 발화된 대로 그대로 전사하고, ‘=’를 붙여 정상적인 단어와 구별할 수 있게 한다. 발화의 수정 등으로 인하여 한 어절이 완전하게 발화되지 못하고 불완전하게 발화된 경우 불완전하게 발화된 어절에 ‘=’를 붙인다.

<예> 4:사실 학습 자료랑 학= 형태는 떨어져도 되는데,

- 하나의 형태소가 완전히 반복되는 발화라 하더라도 학습자가 자가 수정을 하는 경우 ‘=’를 붙인다.

<예> 1:좋아= 좋았어요.

- 발화가 끝나지 않았는데, 말끝을 흐릴 경우 메모를 남긴다.

<예> 1:제주도에 가고 싶지만 돈이.<note>말끝흐림</note>

5) 띄어쓰기

- 띄어쓰기의 경우 맞춤법에 맞게 한다.
- 의존명사는 띄어 쓴다. 단, 특정 시점이나 순서를 나타내는 수사와 함께 사용될 경우는 띄어쓰기를 하지 않는다.(예, 일학년, 일층 등)
- 판단하기 어려운 경우에는 회의를 거쳐 최종 판단하고 향후 유사 사례를 일관되게 처리한다. (예, 오십대, 일 대 이, 등)
- 본용언과 보조 용언도 띄어 쓴다.

6) 축약형의 표기

- 구어에서는 발음의 축약 현상이 많이 나타나는데, 두 음절이 한 음절 사잇소리가 된다거나, 두 음절이 한 음절 겹핥소리가 되는 것 등이다. 구어 말뭉치에서는 발음되는 음절 수와 표기상의 음절 수를 맞추는 것이 원칙이므로 축약형의 경우 모두 표기에 반영한다. 그런데 모음의 축약형의 경

우 대부분 현재 국어의 모음 체계상 표기할 글자가 존재하지만, 반홀소리 된 /ㄱ/, /ㄴ/의 표기가 문제가 된다. /ㄱ/, /ㄴ/가 반홀소리가 되어 /ㄱ/, /ㄴ/와 축약되는 현상은 구어에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 구어 전사에서는 '(apostrophe, 영문따옴표)를 사용해서 두 음소를 연결해 준다.

<예> 사귀'어, 바뀌'어, ...

7) 장음 처리

- 표기와 관련해서 문제가 되는 간투사는 감정을 나타내는 감탄사와 특별한 의미 없이 말버릇 및 머뭇거림의 표지(담화표지)로 사용되는 간투사 유형이다. 이러한 발화는 그 특성상 원 음절보다 길게 발음되는데, 이 경우에는 “::”를 같이 사용해서 표기한다. 참고로 쉼표 표시는 담화표지를 나타냄으로써 후술한다.
- 마지막 음소를 길게 발음하는 경우 역시 “::”로 표기하여 준다.

<예> 1:많은 경우에 논문,
저::~ 어::~ 연구는 네이션,
국가라는 거하고(거하구) 직결되는:: 과정이죠.

- 발화자의 여러 가지 감정을 나타내는 소리들은 실제 구어 전사에서 다양한 형태로 나타난다. ‘오, 허, 응, 어, 어우, 와, 예, 이, 어휴’ 등의 형태를 기본으로 억양이나 길이 등이 달라지면서 놀람, 기쁨, 유감의 감정을 나타내게 된다. 이는 사전에 없는 유형들일 경우가 많은데, 가능한 한 실제 발화에 가깝게 전사하는 것을 원칙으로 한다.

<예> 오, 허, 응, 어, 어우, 와, 와우, 예, 앵, 이, 어휴, 아이, 치, 씨,
헤, 에이...

8) 담화표지

- “이, 그, 저, 아, 어” 등 동일한 형태로 기존 품사의 의미 및 기능을 가지지 않으며 시간을 끌기 위한 주변적인 말일 때 이를 담화표지로 보고 물

결표(~, 숫자1 key 옆에 있음)를 이용하여 표시한다(주로 머뭇거림의 표지로 사용되는 이::~, 그::~, 저::~, 어::~, 아::~ 등이 해당됨).

- 억양과 운율에 의해서만 구분이 가능할 경우는 반드시 전사 단계에서 표시해 준다.
- 이때 담화표지는 대부분 장음을 동반하는 경우가 대부분인데, 반드시 장음 표시와 함께 기재하여 줌을 원칙으로 한다.

<예> 1:많은 경우에 논문,
그::~ 어::~ 연구는 네이션,
국가라는 거하고(거하구) 직결되는 과정이죠.

3. 기타

1) 준음성과 기타 소리들

- 음소가 아닌 요소 즉, 웃음, 기침, 하품, 재채기, 박수와 같은 언어 외적 소리, 전화벨 소리와 같이 사람의 음성 아닌 소리는 대화 흐름에 영향을 주는 경우에만 표기한다. 가령, 발화자가 말하는 도중에 전화벨 소리가 울려서 발화를 멈추고 전화를 받거나, 발화 도중에 웃음소리가 끼어들 경우 발화는 자연스럽게 끊어진다. 이런 경우 전화벨 소리, 웃음을 표기한다. 반면, 발화를 하면서 책장을 넘기거나 볼펜소리를 낼 경우는 말소리와 동시에 소리가 나게 되는데, 이 경우 대화 상대자가 그 소리에 대해 언급하는 등 대화의 내용이나 흐름에 영향을 미치지 않는다면 표기하지 않는다.

- 웃음 <vocal desc='웃음'>
- 기침 <vocal desc='기침'>
- 하품 <vocal desc='하품'>
- 재채기 <vocal desc='재채기'>
- 목청 가다듬는 소리(음, 으음) <vocal desc='목청가다듬는소리'>
- 들이마시는 숨(쓰) <vocal desc='들이마시는숨'>
- 내쉬는 숨(후우) <vocal desc='내쉬는숨'>
- 혀 차는 소리(쯔) <vocal desc='혀차는소리'>
- 헛기침(에헬) <vocal desc='헛기침'>

- 한숨 <vocal desc='한숨'>
- 노래 <vocal desc='노래'>

- 학습자 개인의 발화 특성으로 습관적으로 반복해서 들이마시는 숨소리나 혀 차는 소리, 헛기침 등은 반영하지 않는다.
- 웃으면서 말하는 부분, 박수 치면서 말하는 부분 등도 표시한다.

<예> 5:우우 <vocal desc='박수'>
 <#이리 와 이리 와.#>
 위는 박수만을 치는 경우이고 아래의 경우는 박수를 치며 발화를 하는 경우를 표현한다.

- | | | |
|----------------|----|----|
| - 웃으면서 말하는 부분 | <@ | @> |
| - 박수치면서 말하는 부분 | <# | #> |
| - 노래를 부르는 부분 | <M | M> |

4. 2차 전사

- 2차 전사의 경우 1차 전사 지침을 참고하여 전사된 자료를 검토하고, 아래의 항목에 대해 추가로 작업한다.

1) 철자법 보충

- 1차 전사 작업에서 발음대로 적은 것 가운데, 구어의 발음 특성, 외국인 학습자의 발음 특성 등에 의해 철자대로 소리 나지 않는 발음(표준 발음이 아닌 경우), 음운 규칙이나 정확한 음절 발음을 몰라 일으킨 발음 오류는 () 안에 표기하고 ()밖에는 본래의 표기를 함께 적어 준다. 철자법에 맞는 것을 함께 적어주지 않으면 내용 이해에 어려움이 있을 수 있기 때문이다. 이는 작업자에 따라 1차 전사 과정에서 할 수도 있다.

<예> 친구와(칭구와) 강남에(간남에) 갔습니다(갔슨니다)

<예> 같이(같이)<note>구개음화 안됨</note> 가자.

- 억양단위 맨 끝에 억양기호와 함께 나타나는 경우에는 기호도 함께 붙여준다.

<예> 2:몇 살인데,
그 광은.
광희초등학교 간 사람은?
1:서른 몇 살이나 될 거야.
2:음 젊네 다?(다이?)

- 그러나 구어적 변이형이라 할지라도 자주 등장하거나 쉽게 원래 형태로 이해될 수 있는 것들은 일일이 철자형을 붙여 주지 않는다.

<예> 책상 위에 놉 뒤.

- 소유격 조사 ‘의’의 경우 한글 맞춤법 통일안에서는 [의]와 [에]를 모두 표준발음으로 인정하고 있다. 즉 〈표준 발음법〉 제5 항에서는 단어의 첫음절 이외의 ‘의’는 [이]로, 조사 ‘의’는 [에]로 발음함도 허용하고 있다. 그러나 실제 발화에서 소유격 조사 ‘의’를 [의]라고 발화하는 모어 화자는 매우 드물기 때문에 ‘의’를 [에]로 발음한 경우는 ‘의’ 그대로 전사하고, ‘의’를 [의]로 발화한 경우에는 ‘의(의)’로 전사한다.

<예> [민족에]로 발화하였을 경우

9: 나는,
자랑스런 태극기 앞에,
조국과 민족의,
무궁한 영광을 위하여,

<예> [민족의]로 발화하였을 경우

9: 나는, 자랑스런 태극기 앞에,
조국과 민족의,(민족의,)<note>‘의’로 발음</note>
무궁한 영광을 위하여,

- 외국인 학습자의 발화는 한국어의 음운 체계에 없는 음운의 발음이나 표기가 어려운 중간 발음, 외국인 학습자에게만 나타나는 독특한 발음이 자주 등장한다. 이 경우 () 밖에 규범 표기를 넣어 철자법을 보충하는 것을 기본으로 하나, <예>의 유형 2, 3과 같이 철자 전사를 통해 이를 반영하기 어려우므로 원래의 표기를 먼저 적고, 학습자의 실제 발음을 ()에 남겨 표기는 동일하나 해당 발화에 오류가 있음을 알 수 있도록 한다. 모든 경우 음운적 구분이 모호하거나 특징적인 사항이 있을 때에 메모를 남기도록 한다.

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

선생님(성생닌)

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

가구(가구)<note>‘가’의 ㄱ을 유성음으로 발음</note>

가구(가구)<note>‘구’의 ㄱ을 무성음으로 발음</note>

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

무조건(무조건)<note>경음화 안 됨</note>

같이(같이)<note>구개음화 안 됨</note>

신라(신라)<note>유음화 안 됨</note>

먹는(먹는)<note>자음동화 안 됨</note>

종고(종고)<note>유기음화 안 됨</note>

앞에(앞에)<note>연음 안 됨</note>

- 외국인 학습자의 발화에서 외국어나 외래어 발음 시 원어에 가까운 소리로 발음을 하는 경우가 자주 발생한다. 이는 한국어 모어 화자에게서도 일어나는 현상이기는 하나 외국인 학습자에게서 그 빈도가 더 잦고, 발음 또한 모어 화자의 그것과 많이 다르다. 따라서 이 경우에도 표기 원칙에 맞춰 한글로 적되, 원래의 형태를 파악하기 어렵다고 판단되는 경우에는 () 안에 원어 표기를 적어 밝힌다. 이때 한국어의 음운 체계로 전사가 불가능한 발음은 표기에 반영하지 않는다.

<예> 인터뷰: 영어식 발음으로 [인털류]에 가까운 소리가 남
센터: 영어식 발음으로 [세너]에 가까운 소리가 남
파트너: 영어식 발음으로 [팔너]에 가까운 소리가 남

→ 각각 ‘인터뷰(인털류)’, ‘센터(센너)’, ‘파트너(팔너)’로 전사한 후 <note>원어식 발음</note>을 적는다.

- 외국인 학습자 발화의 경우 외래어 또는 외국어 발화 시 원어식의 발음을 하거나 한국어 모어 화자의 현실 발음이 아닌 규범 발음을 하여 어색하게 들리는 경우가 있다. 이 경우 철자 전사를 통해 반영하기 어려우나 원래의 표기를 먼저 적고, 학습자의 실제 발음을 ()에 남겨 발음에 차이가 있음을 알 수 있도록 한다.

<예> 카페: 현실 발음 [까페], 학습자 발음 [카페]
버스: 현실 발음 [빠스], 학습자 발음 [버스]

→ 각각 ‘카페(카페)’, ‘버스(버스)’로 전사한 후 <note>현실발음과 다름</note>을 적는다.

2) 방언형 표시

- 확실한 방언형(대응하는 표준형 형태소가 없는 것)의 경우는 다음의 예와 같은 태그를 붙인다.

<예> 2:저기여.
 선거 저기 성화 차가 오는 게,
 오 분마다 있어.
 차는 <dia>여일</dia> 있어.

3) 씬

- (1초 이상의 씬은) 0.1초 단위까지 표시한다(전사 도구의 시간 정보를 이용한다). 씬은 발화와 발화 사이의 씬이기 때문에 다음 발화의 시작 전에 표시한다. 만약 씬이 누구의 것인지 불분명할 때는 한 줄에 표시한다.

<예> 1:이거 올라가면서 먹을까?
 2:{1.2}그럴까?
 {4.3}
 ... 하게 먹는다.

- 한 어절 안에서의 짧은 씬은 ‘..’로 표시한다. 하나의 억양 단위 내부에서의 짧은 씬은 따로 표시하지 않는다.

<예> 2:아:: 그리고 어:: 남의 의견을 잘 듣고 수용하고 대화..로 타협해야 된다고 하면서,

4) 인용

- 인용된 부분은 <Q Q>를 사용하여 표시한다. 여러 억양단위에 걸쳐 인용된 경우는 처음과 끝에만 표시를 한다.

<예> 1:근데 요즘 사회학자들은 또는 철학자들은 그렇게 얘기하지 않아요.
 <Q현대사회는 다양성의 시대다.Q>
 라고 말하죠.

5) 텍스트 종류 표현을 위한 전사 부호

- 책이나 자료 등을 보고 읽은 경우는 <R R>를 사용하여 표시한다.

<예> 1:그 답에 <R생각과 느낌이 유기적으로 잘 짜여져 조직체를 이
를 때 좋은 글이 될 수 있다R>라고 돼 있어요.

6) 익명성 보장을 위한 마크업

- 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화번호 등은 노출
되지 않도록 태그로 대신한다. 다음은 마크업 과정에서 각각의 정보를 대신하는
태그들이다.

<name> : 사람 이름, 단체 이름, 학교 이름 등

<id-num> : 주민등록번호, 외국인등록번호, 학번 등 개인 식별 번호

<card-num> : 신용카드 번호

<address> : 주소

<tel-num> : 전화번호

- 여러 사람의 이름이 나올 때는 <name1>, <name2> 등으로 일련번호를
붙여 준다.

<예> 5:네.
거 어떻게 보면 가장 실망..스런 일 중 하난데요,
헤럴드 쪽에서도 그다지 뽕족한,
대안을 가지고 있는 것 같지는 않더라구요,
그게 어찌면 <name1> 선배님이라든지 다른 선배님들 말::들
은 걸 생각해 보면,
<name2> 사장이 <name2> 회장이 있으니까 보안이 있어도
눈치 보여서 얘기를 못한다,

- 사람 이름 뒤에 붙는 접사 ‘-이’는 <name> 태그 시, 함께 포함시킨다. 구
어 전사 후에 수행되는 형태소 분석 시 발생할 수 있는 오류를 최소화시
키기 위하여 구어 전사 과정에서 접사 ‘-이’를 포함시켜서 <name> 태그
를 붙이기로 한다.

<예> 학습자의 이름이 ‘민영’일 때
 학습자 발화: 민영이가, 민영이는, 민영이도, 민영이를
 → 접사 ‘-이’를 포함시켜서 각각 ‘<name>가’, ‘<name>는’,
 ‘<name>도’, ‘<name>를’로 태그를 붙인다.

- 대학교 이름은 ‘<name>대학교’로 붙여서 전사한다. 형태소 분석 시 해당 어절을 하나의 고유명사로 처리하기 때문에 팀 간의 작업상 편의를 위해 구어 전사 과정에서 ‘<name>대학교’로 붙이기로 한다. 만약 ‘대학교’에 발화 오류가 발생하면 () 안에 <name>태그를 넣어서 마크업 처리를 한다.

<예> 학습자가 ‘대학교’를 [대학교]로 발음한 경우
 <name1>대학교(<name1>대학교)

7) 세그멘테이션 분할

- 2015년 전사 작업의 경우 시스템을 전제한 전사라기보다는 학습자 발화의 흐름에 따른 기계적인 전사였다. 따라서 세그멘테이션의 경우 구체적인 지침은 마련하지 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, 발화 단위의 끊김이 빈번하게 일어날 경우 이를 직관적으로 발화의 의미를 판단하지 못하는 경우가 많았다. 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.

2016년 세그멘테이션 지침 추가 부분
○ 세그멘테이션은 학습자의 발화의 흐름을 반영하여 전사하되 분할의 단위는 어절 단위로 한다.

<예> 5:네.
내{1.2}가 학교에,/

빗금은 세그멘테이션의 분할을 나타낸다. 아래와 같이 전사
하지 않는다.

5:네.
내/가 학교에,/

한국어 학습자 말뭉치 형태 주석 지침

※ 본 지침은 21세기 세종 계획 현대문어 형태분석 말뭉치 구축 지침을 기본으로 한다.

I. 학습자 말뭉치의 형태 분석 표지4)

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

4) **【수정】** 기존 세종 지침에 있었던 NF(명사추정범주), NV(용언추정범주)를 삭제하고 대
부분 추정하여 해당 표지로 분석하거나 NA(분석불능범주)로 분석함

	목적격조사	JKO	JKO
	부사격조사	JKB	JKB
	호격조사	JKV	JKV
	인용격조사	JKQ	JKQ
	보조사	JX	JX
	접속조사	JC	JC
(6) 의존형태	선어말어미	EP	EP
	어말어미(연결)	EC	EC
	어말어미(종결)	EF	EF
	명사형 전성어미	ETN	ETN
	관형형 전성어미	ETM	ETM
	체언접두사	XPN	XPN
	명사파생접미사	XSN	XSN
	동사파생접미사	XSV	XSV
	형용사파생접미사	XSA	XSA
	어근	XR	XR
(7) 기호	마침표, 물음표, 느낌표	SF	SF
	쉼표, 가운뎃점, 콜론, 빗금, 줄표, 물결	SP	SP
	따옴표, 괄호표	SS	SS
	줄임표	SE	SE
	불임표(숨김, 빠짐)	SO	SO
	외국어	SL	SL
	한자	SH	SH
	기타 기호	SW	SW
	숫자	SN	SN
	분석불능범주	NA	NA

Ⅱ. 기본 원칙

가. 분석 대상 : 형태분석은 하나의 어절을 분석 대상으로 한다.

나. 분석 원리 : 본 분석은 ‘형태소’ 차원이 아닌 ‘형태’ 차원의 분석이므로 이형태를 최대한 반영한다.

다. 분석 원칙

- 형태 분석은 분석 대상인 원시 말뭉치를 가급적 훼손하지 않는다.
- [보완] 띄어쓰기는 어문 규범의 한글맞춤법을 기본으로 하며 ‘허용’ 규정도 인정한다.

라. [학습자] 분석 기준

- 대원칙 : “21세기 세종 계획 형태 분석 말뭉치 구축 지침”

- 1) “지침”을 보고 분석이 가능하면 지침으로 해결한다.
- 2) “지침”을 통해 해결이 불가능한 어휘에 대한 분석은,
가) 표준국어대사전을 따른다.

■■ 중요하다 [중요/NNG+하/XSA+다/EF]
- 중요 표준국어대사전: 중요/NNG ✓

※ 단, 조사나 어미의 결합형은 지침에 목록으로 제시된 것을 제외하고는 표준국어대사전에 등재되어 있어도 각각 분리해서 분석한다.

■■ 먹었으니까는 [먹/VV+었/EP+으니까/EC+는/JX]
■■ 가고는 [가/VV+고/EC+는/JX]
■■ 좋으니만큼 [좋/VA+으니/EC+만큼/JKB]

- 나) 사전으로 해결이 불가능한 경우 세종 말뭉치를 확인하고 분석한다.

○ 분석 기준 :

1) 뜻(어휘적 의미+기능적 의미)은 알지만 정확한 형태는 모르는 경우 → 원래 품사로 분석한다. <u>교정 어절</u> 이 취할 표지를 준다.	
■■ 문제를 <u>쉬게</u> 풀어요.	[쉬/VA+게/EC]
■■ 너 <u>때문내</u> 죽겠어.	[때문/NNB+내/JKB]
■■ <u>여러까지</u> 문제가 생겼다.	[여러/MM Ⅱ 까지/NNB]
■■ 강에 <u>패수르</u> 버렸다.	[패수/NNG+르/JKO]
2) 형태와 뜻(어휘적 의미+기능적 의미)을 혼동한 경우 → 오류 유형에서는 대치에 해당하는 경우로 <u>보이는 대로</u> 분석한다. <u>오류 어절</u> 만 고려해서 분석한다.	
■■ 내가 <u>고기</u> 가 먹어요.	[고기/NNG+가/JKS]
■■ 입학하자마자 교과서를 <u>팔려고</u> 서점에 갔어요.	[팔/VV+려고/EC]

마. [학습자] 분석 기준의 적용

1) 학습자 말뭉치에서는 종종 문맥에서 전혀 의미를 유추할 수 없는 경우가 나타난다. 이 경우는 주석자가 교정어절을 상정하는 것이 불가능하며, 특정한 형태의 오형태로 추측하기도 어렵다. 이처럼 교정 어절을 상정하기 어려운 경우는 ‘분석불가능(NA)’으로 처리한다.

■■ 그러므로 <u>부스르르</u> 광고는 물가를 인상한다.	[부스르르/NA]
■■ 그리고 경전철 타로 <u>필어</u> 50분쯤 경삭턱역 있습니다.	[필어/NA]
■■ 이번 방학에 저는 친구와 같이 <u>순열전 수열고</u> 싶어요.	[순열전/NA]
	[순열/NA+고/EC]
■■ 그 꿈을 <u>아구할</u> 수 없을 것 같다.	[아구하/NA+ㄹ/ETM]

→ 표현 문형의 구성과 인접한 경우 교정어절을 상정하기 어렵다 해도 표현 문형 구성에 포함되는 형태까지는 분석을 한다.⁵⁾

2) 학습자의 오류 어절에서 문맥적 의미의 추측이 가능할 때는, 최소 교정을 원칙으로 교정 어절을 상정해 형태 분석을 한다.

5) 표현 문형의 경우 <국립국어원2>의 표현 문형 목록을 기준으로 한다.(부록 참고)

가) 상정한 교정어절에 없는 형태가 추가된 경우의 형태 분석은 다음과 같다.

(1) 체언과 조사 결합의 경우는 체언 또는 조사의 오형태로 형태 분석한다.

* 교정어절: √ 학교를

■■ 나는 OO <u>학</u> 곡을 갔다.	[학곡/NNG+을/JKO]
■■ 나는 OO <u>학</u> 교를 갔다.	[학교/NNG+를/JKO]
■■ 나는 OO <u>학</u> 교 <u>고</u> 를 갔다.	[학교/NNG+고/NA+를/JKO]

→ 마지막 예시와 같이 정확한 형태의 체언과 조사가 분리되는데 어절 내에 정체를 알기 힘든 형태가 있을 경우는 해당 형태를 분리해 분석불가능(NA)으로 처리한다.

(2) 용언과 어미 결합의 경우는 용언의 어간 혹은 어근과 어미를 먼저 확보해 분리한 후 잉여적 요소에 대해서는 분석불가능(NA)으로 처리한다.

* 교정어절: √ 많다

■■ 교실에 학생이 <u>만</u> 다.	[만/VA+다/EF]
■■ 내일은 행사가 더 <u> 많</u> 타.	[많/VA+타/EF]
■■ 오늘은 수업이 <u> 많</u> 나 <u>다</u> .	[많/VA+나/NA+다/EF]

■■ 내 꿈을 <u>이</u> 뤄진 <u>고</u> 나 이루기 위해	[이뤄지/VV+ㄴ/NA+고나/EC]
■■ 꼭 잘해야 <u>되</u> 겠 <u>다</u> .	[되/VV+ㄴ/NA+겠/EP+다/EF]

■■ 교통이 꼭 <u>편</u> 립 <u>습</u> 니다.	[편리/NNG+ㅁ/NA+습니다/EF]
■■ 나는 선생님이 <u>되</u> 겠 <u>다</u> .	[되/VV+ㅁ/NA+겠/EP+다/EF]

■■ 공부를 <u>해</u> 고	[하/VV+아/NA+고/EC]
■■ 7시 30분까지 <u>운</u> 동 <u>했</u> 습 <u>니</u> 다.	[운동/NNG+하/XSV+아/NA+ㅁ니다/EF]
■■ 꽃 가게 주인 <u>돼</u> 면	[되/VV+어/NA+면/EC]
■■ 식사가 <u>준</u> 비 <u>되</u> 고 있었다.	[준비/NNG+되/XSV+어/NA+고/EC]

→ 이때, 교정 어절을 기준으로 했을 때 잉여적인 요소가 추가된 오류와 기존의 다른 문법 요소와 혼동한 오류의 경우를 구분해서 주석해야 한다.

■■ 교실에 사람이 <u> 많</u> 은 <u>다</u> .	[많/VA+은/NA+다/EF]
■■ 교실에 사람이 <u> 많</u> 는 <u>다</u> .	[많/VA+는다/EF]

나) 용언의 활용과 관련한 오류의 형태 분석은 다음과 같다.

(1) 학습자의 오류가 활용 규칙과 직접적인 관련이 없는 경우는, 기존 형태 분석 지침대로 어간을 복원해 형태를 분석한다.

- | | |
|-----------------------------|---------------------|
| ■■ 한국말이 너무 <u>아렵다</u> . | [아렵/VA+다/EF] |
| ■■ 저녁식사도 준비하기가 <u>번거워워서</u> | [번거/XR+럽/XSA+어서/EC] |
| ■■ 공간에 <u>해로원데</u> | [해롭/VA+어/NA+ㄴ데/EC] |
| ■■ 아 <u>크다름</u> 꿈이네요. | [크다름/VA+ㄴ/ETM] |

(2) 학습자의 오류가 활용 규칙과 직접적인 관련이 있는 경우, 기존의 형태 분석과 달리 어간을 복원하지 않고 오류 형태를 그대로 살려 분석한다.

- | | |
|----------------------------------|----------------|
| ■■ 한국말이 너무 <u>어려우다</u> . | [어려우/VA+다/EF] |
| ■■ 친구들과 같이 <u>즐거우게</u> 칠 수 있으면 | [즐거우/VA+게/EC] |
| ■■ 다른 사람이 다시 저에게 <u>도울</u> 수 있다. | [도오/VV+ㄹ/ETM] |
| ■■ 정말 <u>추운</u> 경험이었다. | [추으/VA+ㄴ/ETM] |
| ■■ 내가 <u>게으르서</u> | [게으러/VV+어서/EC] |
| ■■ 우리 나라하고 <u>다라서</u> 싫어지만 | [다라/VA+아서/EC] |

(3) 학습자가 활용 규칙을 몰라서 어미를 제대로 선택하지 못한 경우, 어간을 복원하지 않고 잘못된 어간 형태를 살려 분석한다.

- | | |
|-------------------|---------------|
| ■■ 공부가 <u>힘드는</u> | [힘드/VA+는/ETM] |
|-------------------|---------------|

(4) 학습자가 활용 규칙은 알지만 형태소 결합 규칙을 몰라서 잘못된 어절을 만든 경우에는 어간을 복원해서 분석한다.

- | | |
|-----------------------|---------------------------|
| ■■ 공부가 <u>힘든지만</u> | [힘들/VA+ㄴ/NA+지만/EC] |
| ■■ 공부가 <u>힘든입니다</u> . | [힘들/VA+ㄴ/NA+이/VCP+ㅂ니다/EF] |

(5) 다음과 같은 활용 오류형이 나타날 경우 어미를 누락한 오류로 보고 어간의 형태만 분석한다.

- | | |
|--------------------------------|----------|
| ■■ 친구들과 <u>가벼운</u> 장난을 하는 것은 | [가벼우/VA] |
| ■■ 경제력이 부족하거나 <u>힘들</u> 생활을 겪고 | [힘들/VA] |

(6) 학습자가 용언의 활용에서 동일한 형태(혹은 이형태)를 중복해서 사용한

경우는 같은 형태 표지로 분석한다.

■■ 풍선이 <u>부풀</u> 는 것이	[부풀/VV+ㄴ/ETM+는/ETM]
■■ <u>다른</u> 는 게 되게 많이	[다르/VA+ㄴ/ETM+는/ETM]
■■ <u>불평등한</u> 는 생각도 많아요.	[불/XPN+평등/NNG+하/XSA+ㄴ/ETM+는/ETM]
■■ 이권엔 <u>지나지</u> 지 않았습니 다	[지나/VV+지/EC+지/EC]

다) 형태소 경계를 분할하기 어려운 오류 어절의 형태 분석은 다음과 같다.

(1) 상정한 교정 어절의 형태소 음절에 따라 앞에서부터 형태를 분할한 후 형태 표지를 부여한다.

■■ 내 유학생활을 <u>아프로</u> 미래에게	[아/NNG+프로/JKB]
■■ 다언에도 소개해 <u>주게서</u> 요.	[주/VX+게/EP+서요/EF]
■■ <u>어려</u> 슬 데 가을에 좋은 기억이	[어리/VA+어/EP+슬/ETM]
■■ <u>자시</u> 느이 마음대로 했다.	[자시/NNG+느이/JKG]

(2) 오류의 형태가 형태 단계에서도 표시될 수 있도록 가능한 경우 자모 단위로도 형태를 분할한다.

■■ 정말 신기하다고 <u>생각</u> 했다.	[생각/NNG+하/XSV+ㄴ/EP+다/EF]
■■ 어제 영화를 <u>봔</u> 는데	[보/VV+ㄴ/EP+는데/EC]

(3) 기본적으로는 분석한 형태의 결합이 원 어절이 되도록 분할해야 하지만 그 경계 분할이 어려운 오류 어절의 경우, 어간에 학습자가 잘못 쓴 오형태를 주며 어미는 해당 오류 어절이 취해야 하는 어미의 형태를 부여한다.

■■ 쓰레기를 <u>버려</u> 도 되면 좋겠습니다.	[버러/VV+어도/EC]
■■ 많은 친구를 <u>사귀</u> 서 재밌었어요.	[사구/VV+어서/EC]
■■ 소치에 2시간 <u>걸</u> 레요.	[걸레/VV+어요/EF]
■■ 점점 <u>심</u> 해질 것이다.	[심해/VA+어/EC+지/VX+ㄹ/ETM]
■■ A의 주장에 <u>반</u> 에	[반에/VV+어/EC]
■■ 제 꿈을 <u>위</u> 에	[위에/VV+어/EC]
■■ 제 꿈을 <u>위</u> 에서	[위에/VV+어서/EC]

라) 다음의 축약형에서 나타나는 오류형은 이후에 이뤄질 오류 주석을 고려해 형태를 분할하지 않고 오형태로만 분석한다.

■■ 내일은 네 생일이라서 소포를 받았 다 .	[네/NP]
--------------------------------------	--------

■ ■ 재 장소 중에서 제일 좋아하는 곳은

[재/NP]

■ ■ 세 아버지는 키가 큼니다.

[세/NP]

■ ■ 또 궁금한 개 있으면

[개/NNB]

■ ■ 저는 OO어학당 6급 학생이에요.

[학생/NGG+이/VCP+예요/EF]

■ ■ 나는 학생이었다.

[학생/NGG+이/VCP+였/EP+다/EF]

Ⅲ. 표지별 분류 기준 및 세부 지침

가. 체언

- 체언은 명사, 대명사, 수사를 포괄하는 대범주로서, 조사와 결합하거나 그 자체로 다른 체언이나 용언과 어울려 하나의 문장성분이 될 수 있다.

1) 명사(NN)

- 명사는 사물의 이름을 나타내는 품사이다. 본 표지에서는 명사를 일반명사, 고유명사, 의존명사로 세분한다.

가) 일반명사(NNG)

- 사물의 이름을 나타내는 단어로서 표준국어대사전에 명사로 등재된 표제어(고유명사와 의존명사를 제외한 모든 명사)와 독립된 음절(한자어), 약어, 고사성어 등 사전 표제어는 아니나 다른 품사로 분석될 수 없는 단위들을 포함한다.

(1) 일반명사로 분석할 수 있는 단어

(가) 표준국어대사전의 명사 표제어

■■ 국어/NNG, 연구/NNG

(나) 1음절 한자어가 독립된 단위로 사용되는 경우

■■ 서울초등학교 졸 [졸/NNG]

※ 기타

■■ 나는 환경에 '환'자도 모르는 [/'SS+환/NA+'/'SS+자/NNG+도/JX]

(다) 한자성어

■■ 백척간두(百尺竿頭) [백척간두/NNG+(/'SS+百尺竿頭/SH+)/SS]

(라) 외국어를 음차한 경우

■■ 아이 러브 유(I love you) [아이/NNG]

(마) ‘명사 + (분석 목록에 없는) 접사’는 전체를 통합하여 명사로 분석한다.

■■ 2년간 [2/SN+년간/NNB]

■■ 4호선 [4/SN+호선/NNB]

■■ 상상력 [상상력/NNG]

■■ 중국식 [중국식/NNG]

(2) 명사 상당어의 분석

(가) 동사의 활용형이 따옴표 없이 문장 속에서 명사처럼 기능하는 경우는 원래 품사대로 분석한다.

■■ 어디 가느냐가 그의 물음이었다. [가/VV+느냐/EF+가/JKS]⁶⁾

(나) 따옴표를 가진 성분이나 요소도 명사처럼 기능할 수 있으나, 원래 품사대로 분석한다.

■■ 그것은 “는”이 아니라 “를”이다. ["/SS+는/JX+"/SS+이/JKC]

(다) 부사 뒤에 격조사가 쓰이는 것도 의미론적인 따옴의 효과에 의하여 부사가 명사적인 용법을 가지는 것이므로 분석은 ‘부사’로 한다.

■■ 가족을 멀리에 보냈다. [멀리/MAG+에/JKB]

(라) 학습자의 특성상 접사를 명사적 기능으로 사용한 경우 분석하는 접사 목록에 없더라도 원래 품사대로 접사로 분석한다.

■■ 제주도에는 한국의 여명이 도예요. [도/XSN+이/VCP+예요/EF+./SF]

(3) 학생, 학교

- 대학교, 고등학교, 중학교, 대학생, 고등학생, 중학생은 모두 일반명사로 분석한다.

■■ 대학교 [대학교/NNG]

■■ 고등학교 [고등학교/NNG]

6) [수정] ‘21세기세종계획’ 지침에는 ‘느냐/EC’로 되어 있지만 오류이므로 수정함.

■■ 중학교	[중학교/NNG]
■■ 대학생	[대학생/NNG]
■■ 고등학생	[고등학생/NNG]
■■ 중학생	[중학생/NNG]

나) 고유명사(NNP)

- 고유 명사는 특정한 사물에 붙여진 이름으로, 기본적으로 최하의어에 속하는 대상을 서로 변별하기 위하여 붙인 이름이며, 원칙적으로 지시 대상만 가질 뿐 의미 내용은 가지지 않는다. 고유명사의 분석 기준은 매우 다양하므로, 본 지침에서는 다음에 제시하는 것만을 고유명사로 인정한다. 또한, 본 지침은 띄어쓰기 단위의 분석을 원칙으로 하고 있으므로, 한 단어 이상으로 구성된 고유명사(‘바람과 함께 사라지다’)와 같은 경우의 분석을 위해 전체를 아우르는 단위를 설정하지는 않는다.

(1) 인명, 종족명

- (가) ‘씨(氏), 공(公), 군(君), 양(嬢), 웅(翁)’ 등 성 또는 이름 뒤에 같이 쓰이는 호칭어나 직책명은 분리해서 분석한다.

■■ 남수/NNP||군/NNB, 김/NNP||씨/NNB, 최치원/NNP||옹/NNB,
케네디/NNP||씨/NNB⁷⁾, 정/NNP||과장/NNG, 최/NNP||선생/NNG

- (나) 성과 이름, 호가 함께 쓰이면 하나의 단위로 분석한다.

■■ 김철수/NNP, 이태백/NNP

- (다) ‘씨, 군’ 등과 달리 ‘가(哥)’는 접미사이므로, ‘김가(金哥), 이가(李哥)’는 파생어이다.

■■ 김/NNP+가/XSN

- (라) 사람 이름의 뒤에 접사 ‘-이’가 붙는 경우는 이름과 함께 하나의 단위로 분석한다.

7) [수정] 지침 전체적으로 띄어 써야 할 부분이 +기호로 연결되어 있어 || 기호로 수정함.

■ ■ 진현이/NNP + 가/JKS

(마) 특정한 종족의 이름은 고유명사가 된다.

■ ■ 알타이족/NNP, 피그미족/NNP, 돌궐족/NNP, 한족/NNP

(2) 지명

(가) 내륙, 바다, 강, 산, 산맥, 호수, 섬, 만, 계곡, 늪, 주 등의 이름

■ ■ 카스피해/NNP, 템즈강/NNP, 태백산맥/NNP, 미시시피호/NNP, 네바다주/NNP

■ ■ 한강/NNP, 한라산/NNP, 남이섬/NNP, 남극/NNP, 북극/NNP

(나) 주소를 나타내는 도(道), 시(市), 읍(邑), 면(面), 리(里), 군(郡), 구(區), 동(洞), 골, 촌, 로 등의 이름은 그 구역의 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

■ ■ 서울특별시/NNP, 성북구/NNP, 강진군/NNP, 인천동/NNP, 빨래골/NNP, 해방촌/NNP

■ ■ 연세로/NNP, 세검정로/NNP, 상동로/NNP, 테헤란로/NNP

■ ■ 신촌/NNP, 여의도/NNP, 광화문/NNP, 명동/NNP

(3) 국가명 또는 왕조명

(가) 국가의 명칭, 또는 왕조의 명칭은 고유명사로 분석한다.

■ ■ 대한민국/NNP, 조선/NNP

(나) 다른 형태가 붙어 국가나 왕조의 존립 기간을 나타내는 경우 일반명사로 분석한다.

■ ■ 대한제국기/NNG, 조선조/NNG

(다) '남, 북, 남북'은 방향을 가리키는 일반명사와 '남한'과 '북한'을 의미하는 고유명사를 구별한다. 남한을 뜻하는 '남'과 북한을 뜻하는 '북'을 고유명사로 분석한다.

■ ■ 남/NNP+과/JC || 북/NNP+의/JKG || 의견/NNG || 차이/NNG

■ ■ 남북/NNP || 적십자회담/NNG

■ ■ 북/NNP+미/NNPⅡ회담/NNG

(라) 어떤 국가의 국민을 나타내는 ‘국가+인’은 통합하여 일반명사로 분석한다.

■ ■ 이집트인/NNG, 아제르바이젠인/NNG, 이스라엘인/NNG, 조선인/NNG

(마) 어떤 국가의 군대를 나타내는 ‘국가+군’은 통합하여 일반명사로 분석한다.

■ ■ 미군/NNG, 북한군/NNG, 영국군/NNG, 일본군/NNG

(바) 국가명의 약어는 고유명사로 분석한다.

■ ■ 한/NNP+중/NNP+일/NNP

(4) 건축물이나 시설물 혹은 구조물의 이름

(가) 도로, 항만, 철도, 전철, 지하철 및 그 명칭과 함께 쓰이는 부대시설은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

■ ■ 부산항/NNP, 대전역/NNP, 서울지하철/NNP, 인천공항/NNP

■ ■ 홍대입구역/NNP, 홍대입구/NNP(준말)

(나) 빌딩, 박물관, 극장 등 건물명은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

■ ■ 서울역사/NNP, 세종문화회관/NNP, 개나리유치원/NNP, 연세대학교/NNP

■ ■ 국립중앙박물관/NNP, 국립민속박물관/NNP, 루브르박물관/NNP

■ ■ 신라호텔/NNP, 현대백화점/NNP, 동궁예식장/NNP, 명보극장/NNP, 세브란스병원/NNP

(다) 알파벳이나 숫자, 기호를 포함한 경우 전체가 고유명사가 된다.

■ ■ N서울타워/NNP, N-서울타워/NNP, 63빌딩/NNP

※ 기타

■ ■ 남대문/NNPⅡ시장/NNG, 한강/NNPⅡ공원/NNG

(5) 회사, 학교, 정당, 기관이나 단체의 이름

(가) 특정 회사나 학교, 정당 등의 이름은 고유명사로 분석한다. 단, 특정 회사의 상품명은 고유명사가 아닌 일반명사로 취급한다.

- 삼성/NNP, 연세대학교/NNP, 새누리당/NNP, 자유민주주의연합/NNP
- 초코하임/NNG, 한메타자교실/NNG

(나) 정부기관의 명칭은 모두 일반명사로 처리한다. 그러나 거기에 인명, 지명 등의 고유명사가 포함된 경우 그 통합형을 고유명사로 처리한다.

- 헌법/NNG||재판소/NNG, 대/XPN+법원/NNG, 고등/NNG||법원/NNG, 재정/NNG||경제원/NNG
- 서울고등법원/NNP, 서울시경찰서/NNP, 서대문구치소/NNP

(다) 특정 기관이나 단체, 연구소 등의 경우에는 분석하는 것을 원칙으로 한다. 그러나 거기에 인명, 지명 등의 고유명이나 ‘전국’, ‘국제’, ‘세계’ 등이 포함되면 그 통합형을 고유명사로 처리한다.

- 대한축구협회/NNP, 전국은행협회/NNP, 한국전자통신연구소/NNP
- 생활/NNG||체육/NNG||연구소/NNG, 입주자/NNG||대표자/NNG||협의회/NNG

(라) 약어나 준말의 처리

- 고유명사가 축약된 형태(준말)로 쓰일 경우 본디말과 함께 준말도 인정하여 축약된 형태 그대로를 고유명사로 분석한다. 그리고 일반명사로 분석하는 기관명의 약자는 일반명사로 분석한다.

- 육사/NNP, 연대/NNP, 자민련/NNP, 서울고법/NNP
- 정보통신위/NNG (정보/NNG||통신/NNG||위원회/NNG)

(6) 아이돌 등의 그룹명은 (6) 창작물의 제목과 같게 처리한다.

- 소녀시대/NNP, 걸스데이/NNP, 방탄소년단/NNP
- 제국/NNG+의/JKG||아이/NNG+들/XSN, 서태지/NNP+와/JC||아이/NNG+들/XSN

※ 기타

- EXID/SL, YG/SL+Family/SL
- B1A4/NNP, 2NE1/NNP

(7) 책, 연극, 영화, 드라마, TV 프로그램 등의 창작물의 제목

■■ 삼국사기/NNP, 손자병법/NNP, 고래사냥/NNP

■■ 슈키라(슈퍼주니어의 키스 더 라디오) 슈키라/NNG(준말)

어절 미분리 (NN 구성 포함)	사전 등재	전체 NNP	(책) 삼국사기/NNP, 손자병법/NNP
	사전 미등재	전체 NNP	(드라마) 전원일기/NNP, 가을동화/NNP (영화) 어벤저스/NNP, 쿵푸팬더3/NNP (TV프로그램) 런닝맨/NNP, 가족오락관/NNP
어절 분리	사전 등재	나누어 분석	(책) 안네/NNP+의/JKG 일기/NG
	사전 미등재	나누어 분석	(드라마) 서울/NNP+의/JKG 달/NG (영화) 비밀/NG+은/JX 없/VA+다/EF (TV프로그램) 남자/NG+의/JKG 자격/NG

(8) 언어명

- 언어명의 경우 ‘-어’의 형태만을 통합하여 고유명사로 인정한다.

■■ 한국어/NNP, 일본어/NNP, 영어/NNP, 알태국어/NNP, 네덜란드어/NNP

■■ 한국말/NNG, 러시아/NNP||말/NNG, 일본/NNP||말/NNG

■■ 한글/NNG, 알파벳/NNG, 한자/NG

(9) 웹사이트, SNS, APP

- 웹사이트, SNS, APP의 이름은 모두 고유명사로 처리한다.

■■ 네이버/NNP, 다음/NNP, 구글/NNP

■■ 인스타그램/NNP, 페이스북/NNP, 카카오톡/NNP, 트위터/NNP

■■ 직방/NNP, 카카오퍼스/NNP

(10) 캐릭터의 이름

■■ 미키마우스/NNP, 호돌이/NNP, 알라딘/NNP, 키티/NNP, 라이언/NNP

다) 의존명사(NNB)

- 의존명사는 자립해서 쓰일 수 없는 명사로, 수식 성분을 반드시 동반해야 한다. 의존명사는 비단위성 의존명사와 단위성 의존명사로 나눌 수 있으

나, 본 분석에서는 이를 세분해하지 않는다. 또한 의존명사가 일반명사와 같이 독립적으로 쓰일 때는 일반명사로 분석한다. 의존명사와 일반명사의 구분은 표준국어대사전의 등재 여부에 따른다.

(1) 의존명사이지만, 일반명사처럼 쓰이는 경우

(가) “연대, 연도, 연차”는 “년대, 년도, 년차”와 달리 모두 일반명사로 분석한다.

- | | |
|----------------|----------|
| ■■ 연도별로 정리된 자료 | [연도/NNG] |
| ■■ 몇 년도에 일어난 일 | [년도/NNB] |

(나) “월, 연, 일, 주, 달러, 원” 등은 본래 의존명사이지만, 독립되어 쓰일 경우 모두 일반명사의 자격을 가지므로 일반명사로 분석해야 한다.

- | | |
|--------------------|----------|
| ■■ 나는 월 30만원을 받는다. | [월/NNG] |
| ■■ 달러의 가치는 | [달러/NNG] |

(2) 단위를 나타내는 표현

(가) 길이, 무게, 수효, 시간 따위의 수량을 수치로 나타내는 단위들 중 “미터, 그램, 리터” 등은 의존명사(NNB)로, 외국어로 된 “m, g, l” 등은 기호(SW)로 분석한다.

(나) 일반명사가 단위적인 용법으로 쓰인 경우에는 의존명사가 아니므로 주의한다.

- | | |
|--------------------|----------------|
| ■■ 사람, 시간, 그릇, ... | |
| ■■ 한 사람이 교실로 들어왔다. | [사람/NNG+이/JKS] |
| ■■ 자장면 한 그릇만 주세요. | [그릇/NNG+만/JX] |

(3) ‘것’과 구어형 ‘거’의 분석

- ‘거’의 형태를 그대로 인정하여 분석한다. 그러나 다른 형태와의 결합에서 ‘거’의 형태가 유지되지 않는다면 그 때에는 ‘것’으로 복원하여 분석한다.

- | | |
|-------------------|---------------|
| ■■ 공부할 거를 준비해 왔니? | [거/NNB+를/JKO] |
| ■■ 공부할 걸 가져왔니? | [것/NNB+ㄹ/JKO] |

■■ 연습할 건 있니?

[것/NNB+ㄴ/JX]

■■ 먹을 게 모자르다

[것/NNB+이/JKS]

※ [보완] 학습자의 오류로 인해 ‘거’의 형태가 유지되지 않는 경우는, ‘것’으로 복원하지 않는다.

■■ 밥을 먹을 건다.

[거/NNB+이/VCP+ㄴ 다/EF]

※ 학습자의 오류로 인해 축약된 ‘게’의 형태가 유지되지 않는 경우는, 분할하지 않고 분석한다.

■■ 또 궁금한 게 있으면

[개/NNB]

2) 대명사(NP)

- 대명사는 그 자체로는 자신의 본유적 지시물을 가지지 않은 채, 다만 사람이나 사물 등 어떤 대상을 간접적으로 지시하는 품사이다. 단, 동일한 대명사가 방언이나 고어의 이형태를 가진 경우에는 이들도 대명사로 같이 분석한다.

(1) 1인칭 대명사

(가) 1인칭 대명사

■■ 나, 내, 우리, 저, 제, 저희

(나) 2인칭 대명사

■■ 너, 네, 그대, 당신, 댁, 어르신

(다) 기타 대명사

■■ 이이, 이분, 그이, 그분, 저이, 저분, 아무, 아무개, 누구, 무엇, 뭐, 어디, 언제, 자기, 개, 재, 애, 이것, 저것, 그것, 이거, 저거, 그거, 여기, 저기, 거기, 이곳, 그곳, 저곳, 어디, 모(某), 모모(某某)

※ ‘자기’는 대명사로 분석한다.

※ ‘자신’, ‘아무것’은 일반명사로 분석한다.

※ ‘우리나라’는 한국인이 사용하는 경우 ‘우리 한민족이 세운 나라를 스스로 이르는 말.’의 뜻의 일반명사로 분석하지만, 외국 학생들이 사용하는 경우 ‘우리/NP || 나라/NNG’로 분석해야 한다. 학습자 말뭉치의 경우 외국 학생들의 작문이나 구어 전사 텍스트이므로 ‘우리나라(우리 나라)’가 등장하는 경우 모두 ‘우리/NP || 나라/NNG’으로 분석한다.

(2) 대명사와 관형사의 두 가지 분석이 가능한 단어

(가) ‘모(某)’는 관형사와 대명사로 분석될 수 있으므로 주의를 요한다.

■■ 모 기업체	[모/MM]
■■ 김 모 씨	[모/NP 씨/NNB]

(나) ‘모모(某某)’도 위와 같이 분석될 수 있다.

■■ <u>모모</u> 가 말했다	[모모/NP+가/JKS]
■■ <u>모모</u> 기관의 조사를 마쳤다	[모모/MM]

(3) 대명사의 이형태 분석

(가) ‘이것, 그것, 저것; 이거, 그거, 저거’는 분석하지 않고 대명사로 인정한다. ‘~거’의 경우, 다른 형태와의 결합에서 ‘~거’의 형태가 유지되지 않는다면 그 때에도 ‘~것’으로 복원한다.

■■ 난 <u>저거</u> 를 먹을래.	[저거/NP+를/JKO]
■■ 나는 여태 <u>그걸</u> 믿어 왔단다.	[그것/NP+ㄹ/JKO]

(나) 다음과 같이 원형을 밝힐 수 있는 대명사는 원형대로 분석한다.

■■ 내	이제부터는 <u>내</u> 명령을 따라라.	[나/NP+의/JKG]
■■ 내게	<u>내게</u> 전자우편으로 알려 다오.	[나/NP+에게/JKB]
■■ 네게	어제 <u>네게</u> 보낸 선물이 잘못되었다.	[너/NP+에게/JKB]
■■ 제게	문제가 있다면 <u>제게</u> 말씀해 주세요.	[저/NP+에게/JKB]
■■ 누가	<u>누가</u> 전화를 하는 지 보고해라.	[누구/NP+가/JKS]
■■ 뉘	<u>뉘</u> 집 애기가 울고 있는 거야?	[누구/NP+의/JKG]
■■ 뭐가	도대체 <u>뭐가</u> 문제라는 거야?	[뭐/NP+가/JKS]

※ [참고] ‘내가’는 모두 ‘내/NP+가/JKS’로 분석한다.

■■ 내가 내가 살던 집 [내/NP+가/JKS]

(다) ‘뭐’는 ‘무엇’과 대등할 정도로 자주 사용되므로 그 형태 자체를 인정해 준다. 다만, 다음과 같이 조사와 축약되었을 경우에만 원형으로 복원해 준다.

■■ 앞으로 우리가 뭘 하자는 얘inya? [무엇/NP+ㄹ/JKO]

(라) ‘제’의 경우, ‘제/NP+가/JKS’를 제외하고는 모두 ‘저/NP+의/JKG’로 분석한다.

■■ 제가 갈 것입니다. [제/NP+가/JKS]

■■ 철수는 제 잘못을 안다. [저/NP+의/JKG]

■■ 제 무게를 못 견디다. [저/NP+의/JKG]

※ 학습자가 대명사 뒤에서 조사를 누락해서 쓴 경우와 형태적 유사함이 있기 때문에 분석에 주의해야 한다. 이 경우는 ‘저/NP+의/JKG’ 또는 ‘나/NP+의/JKG’로 분석하지 않는다

■■ 제 먹었습니다. [제/NP]

■■ 내 활짝 웃었다. [내/NP]

※ 다음의 대명사와 조사의 축약형에서 나타나는 오류의 경우 원형을 밝혀 분석하지 않고 오형태로 분석한다.

■■ 내일은 네 생일이라서 소포를 받았다. [네/NP]

■■ 재 장소 중에서 제일 좋아하는 곳은 [재/NP]

■■ 세 아버지는 키가 큼니다. [세/NP]

3) 수사(NR)

- 수사는 사물의 수량이나 차례를 나타내는 품사를 말한다.

(1) 수사의 종류

(가) 양수사

- ■ 하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉, 열, 스물, 서른, 마흔, 쉰, 예순, 일흔, 여든, 아흔, 백한들, 두서넛, 서넛, 너댓, 네다섯, 네댓, 대여섯, 예닐곱, 일여덟, 일고여덟, 열두서넛, 열대여섯, 열일고여덟, 스물두서넛
- ■ 일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 십, 백, 천, 만, 억, 조
- ■ 기십, 기백, 기천,
- ■ 수십, 수백, 수천, 수만, 수억, 수십만, 수백만, 수천만

(나) 서수사

- ■ 첫째, 둘째, 셋째, 넷째, ..., 열째, 열한째,..., 스물한째,...
- ■ 아흔아홉째, 백째, 백한째, ...

※ ‘째’는 분석하는 접미사에 해당하지만 서수사에서 쓰인 경우 분석하지 않는다.

- ■ 첫째 [첫째/NR]
- ■ 첫 번째 [첫/MMⅡ번/NNB+째/XSN]

<주의사항>

(가) 복수의 수사가 한 어절 내에 나타날 때에는 전체를 통합해서 분석한다.

- ■ 백만오천삼십사 [백만오천삼십사/NR]

(나) ‘하나’는 표준국어대사전에 그 품사가 명사와 수사로 되어 있지만 본 지침에서는 **수사**로 분석한다.

- ■ 광에 가서 물건 하나만 가져오렴. [하나/NR+만/JX]
- ■ 우리는 하나로 뭉쳤다. [하나/NR+로/JKB]

(다) 때로 수사와 수관형사의 구별이 애매한 경우가 있다. 이 분석에서는 임흥빈(1998)의 견해에 따라, 다음과 같은 특이한 형식을 가진 예만을 수관형사로 취급하고, 그 밖의 것들은 모두 수사로 분석한다.

- 한, 한두, 한두어, 두, 두어, 두세, 두서너, 세, 석, 서, 서너, 네, 너, 넉
- 열한, 스물두, 서른세 등

→ 수관형사로 취급하는 특이한 형식으로 끝나는 경우는 모두 수관형사로 취급한다.

(라) ‘제일, 제이’ 등은 접두사 ‘제-’와 수사의 결합으로 분석한다.

- 제일, 제이, 제삼, 제사, 제오, ..., 제구십구, 제백, ... [제/XPN+일/NR],
[제/XPN+이/NR], ...

나. 용언

- 용언은 동사, 형용사, 지정사를 가리킨다. 용언 범주에서는 분석 대상이 본용언일 경우에만 동사와 형용사로 구분하여 표시하고, 보조용언의 경우에는 보조동사와 보조형용사를 구분하지 않고 ‘VX’라는 하나의 표지만을 준다. 또한 학교 문법에서 서술격조사로 다루는 ‘이다’는 조사의 범주에 넣지 않고 ‘지정사’라는 용언의 하위범주에 넣기로 한다. 지정사는 다시 긍정 지정사(VCP)와 부정 지정사(VCN)로 세분된다.

1) 동사(VV)

- 동사는 사물의 움직임이나 작용을 나타내는 용언을 말한다. 동사는 일반적으로 목적어의 필요성 여부에 따라 자동사, 타동사로 나누기도 하지만, 본 분석에서는 그것을 위한 별도의 표지를 세분하지 않고 모두 ‘VV’로 표시한다.

※ ‘있다’는 모두 **동사**로 처리한다. (세종 말뭉치 기준)

※ ‘감사하다’는 모두 **동사**로 보고 ‘-하-’는 모두 동사파생접미사로 처리한다. (세종 말뭉치 기준)

※ ‘명사/어근/부사 + (분석 목록에 없는) 동사파생접미사’는 전체를 통합하여 동사로 분석한다.

- 말씀드리다 [말씀드리/VV+다/EF]

■■ 반짝거리다

[반짝거리/VV+다/EF]

2) 형용사(VA)

- 형용사는 사물의 성질이나 상태를 나타내는 용언을 가리킨다.

※ [보완] ‘명사/어근/부사 + (분석 목록에 없는) 형용사파생접미사’는 전체를 통합하여 형용사로 분석한다.

■■ 나다

[별나/VA+다/EF]

■■ 맞다

[능글맞/VA+다/EF]

3) 보조용언(VX)

1. 사전 등재

예) 가늘어지다 가늘어지/VV+다/EF

 좋아하다 좋아하/VV+다/EF

2. 사전 미등재

예) 심해지다 심하/VA+아/EC+지/VX+다/EF

 초조해하다 초조/NNG+하/XSA+아/EC+하/VX+다/EF

→ 이 분석에서는 보조용언을 보조동사와 보조형용사로 하위 구분하지 않는다.

(1) 보조용언 분석 원칙

(가) 보조용언의 후보는 표준국어대사전에 그 쓰임이 제시되어 있어야 한다.

(나) 보조용언 앞에는 반드시 다른 용언이 위치해 있어야 한다.

(다) 보조용언이 동시에 두 개 이상이 연결되어 나타날 수도 있다.

(2) 보조용언의 예와 주의사항

- 보조용언의 목록은 다음과 같다. 이 목록은 표준국어대사전을 참고한 것이다.

■ ■ 가다	세월이 흘러 가는 대로 떠도는 나그네	가/VX+는/ETM
■ ■ 가지다	그렇게 해 가지고는 기일을 맞출 수 없다.	가지/VX+고/EC+는/JX
■ ■ 계시다	손님께서 와 계십니다.	계시/VX+ㅂ니다/EF+./SF
■ ■ 나가다	추진해 나가는 과정에서 문제가 생겼다.	나가/VX+는/ETM
■ ■ 나다	아침에 깨어 나 보니 그가 없어졌다.	나/VX+아/EC
■ ■ 내다	힘들겠지만 잘 견뎌 내야 한다.	내/VX+아야/EC
■ ■ 놓다	약속을 잡아 놓고 출장을 가다니	놓/VX+고/EC
■ ■ 달다	이번 시험 문제의 정답을 알려 다오.	달/VX+오/EF+./SF
■ ■ 대다	자꾸 졸라 대는 통에 허락해 주고 말았다.	대/VX+는/ETM
■ ■ 두다	남겨 둔 쌀도 이제 바닥이 났다.	두/VX+ㄴ/ETM
■ ■ 드리다	염려를 끼쳐 드려 송구하옵니다.	드리/VX+어/EC
■ ■ 들다	도무지 내 말은 믿으려 들지 않는다.	들/VX+지/EC
■ ■ 말다	어렵더라도 희망을 잃지 말아야 한다.	말/VX+아야/EC
■ ■ 먹다	나는 오늘도 수업을 빼 먹었다.	먹/VX+었/EP+다/EF+./SF
■ ■ 못하다	그 참상을 차마 보지는 못할 것이다.	못하/VX+ㄹ/ETM
■ ■ 버리다	음식이 다 타 버렸다.	버리/VX+었/EP+다/EF+./SF
■ ■ 보다	이제는 새벽이 오는가 보다.	보/VX+다/EF+./SF
■ ■ 빠지다	썩어 빠진 생선을 사오다니	빠지/VX+ㄴ/ETM
■ ■ 싶다	너를 보고 싶다.	싶/VX+다/EF+./SF
■ ■ 쌓다	꼬치꼬치 물어 쌓는 통에 정신이 없었다.	쌓/VX+는/ETM
■ ■ 아니하다	일이 순리대로 풀리지 아니했다.	아니하/VX+았/EP+다/EF+./SF
■ ■ 앓다	시간이 지나도 기차는 오지 않았다.	앓/VX+았/EP+다/EF+./SF
■ ■ 오다	고향을 떠나 온 지 10년이 지났다.	오/VX+ㄴ/ETM
■ ■ 있다	그녀는 검정 옷을 입고 있었다.	있/VX+었/EP+다/EF+./SF
■ ■ 주다	아버지는 아기에게 동화책을 읽어 주었다.	주/VX+었/EP+다/EF+./SF
■ ■ 지다	한 번 넘어 진 아이는 일어나는 법을 안다.	지/VX+ㄴ/ETM
■ ■ 치우다	다섯 명이 10인분의 식사를 먹어 치웠다.	치우/VX+었/EP+다/EF+./SF

■■■ 터지다	끓인 지 오래 되어서 라면이 불어 터졌다.	터지/VX+었/EP+다/EF+./SF
■■■ 하다	나귀를 쉬게 하는 것이 좋겠다.	하/VX+는/ETM

- ① 다음과 같은 어절은 보조용언으로 취급되기도 하나, 여기서는 ‘의존명사+접사’로 분석한다. 이들 앞에는 항상 관형어가 온다는 분포적인 특성을 중시한 것이다.

■■■ 양하다/채하다/척하다/듯하다/법하다/뻔하다	[양/NNB+하/XSA+다/EF]
■■■ 듯싶다	[듯싶/VX+다/EF]

- ※ 표준국어대사전에 따라, 기존에 접미사로 분석하던 ‘만하’의 지침을 변경하여, ‘만’을 보조사로, ‘하’를 동사로 분석한다. ‘만하’는 ‘만/NNB+하/XSA’로 분석되는 경우도 있으므로 주의해야 한다.⁸⁾

■■■ <u>철수만 한</u> 인재가 없다	[철수/NNP+만/JX 하/VV+ㄴ/ETM]
■■■ 이 음식은 먹을 <u>만하다</u> .	[만/NNB+하/XSA+다/EF+./SF]

- ② ‘버릇하다’의 경우에는 선행 성분으로 관형형이 오는 것은 아니지만, 일반명사 ‘버릇’과 크게 구별되지 않으므로 ‘버릇’은 명사로 분석한다.

■■■ 자꾸 울어 <u>버릇하다</u> .	[버릇/NNG+하/XSV+다/EF+./SF]
-------------------------	--------------------------

- ※ ‘-도록 하다’는 형용사 일부 어간에만 사용되는 등 ‘-게 하다’와 분포가 다르므로 이때의 ‘하다’는 본용언으로 분석한다.

■■■ 열심히 <u>공부하도록</u> 하자.	[공부/NNG+하/XSV+도록/EC 하/VV+자/EF]
--------------------------	-----------------------------------

4) 지정사(VC)

- 지정사는 학교 문법의 서술격 조사에 대응되는 것인데, 용언과 같이 활용한다는 특성을 중시한 술어이다. 여기서는 학교 문법의 ‘이다’를 긍정 지정사로, ‘아니다’를 부정 지정사로 하위 구분한다. 일반적으로 ‘아니다’는 형용사로 다루어지기도 하나, 여기서는 ‘아니다’가 ‘이다’의 부정형이라는 점을 중시하여 ‘부정지정사’로 다룬다.

8) ‘바. 3) 다) 형용사파생접미사’의 주의사항의 내용 이동함.

- 철수는 매우 우수한 학생이다. [학생/NNG+이/VCP+다/EF+./SF]
 ■■ 철수는 모범적인 학생이 아니다. [아니/VCN+다/EF+./SF]

※ [참고] 지정사 ‘이/VCP’를 복원해야 하는 경우

① 체언에 어미가 직접 연결된 경우

- 철수는 훌륭한 교사다. [교사/NNG+이/VCP+다/EF+./SF]

② 조사에 어미가 직접 연결된 경우

- 우리가 그를 본 것은 서울에서다. [서울/NNP+에서/JKB+이/VCP+다/EF+./SF]

③ ‘~였다’

- 그 당시 나는 아이였다. [아이/NNG+이/VCP+였/EP+다/EF+./SF]

④ 어미 ‘-라고, -라는, -라도, -라며, -라면서, -라서’

- 나는 그에게 절교라고 말했다. [절교/NNG+이/VCP+라고/EC]
 ■■ 나는 친구라는 말이 좋다. [친구/NNG+이/VCP+라는/ETM]
 ■■ 거지라도 존중해 주어야 한다. [거지/NNG+이/VCP+라도/EC]
 ■■ 그는 최고라며 나를 추켜 주었다. [최고/NNG+이/VCP+라며/EC]
 ■■ 그는 실수라면서 얼버무렸다. [실수/NNG+이/VCP+라면서/EC]
 ■■ 너는 부자라서 우릴 이해하지 못할 것이다. [부자/NNG+이/VCP+라서/EC]

⑤ 인용문 뒤에 오는 “~며” 는 지정사를 복원하지 않는다.

- 얼마나 친절하냐?며 [친절/NNG + 하/XSA + 나/EF + ?/SF + "/SS + 며/EC]

⑥ [보완] ‘아서/어서’에 종결어미가 결합된 경우 (세종 말뭉치)

- 없어진 것을 확인하기 위해서다. [위하/VV+아서/EC+이/VCP+다/EF+./SF]
 ■■ 그때 그 시절의 사람들이 생각나서다. [생각나/VV+아서/EC+이/VCP+다/EF+./SF]
 ■■ 내가 개를 좋아하는 건 개가 착해서야. [착하/VA+아서/EC+이/VCP+야/EF+./SF]

<주의사항>

(가) 학습자가 지정사 ‘이’를 몰라서 누락한 경우는 ‘이/VCP’를 복원하지 않는다.

■■ 방법은 한 가지예요. [가지/NNB+예요/EF]

■■ 이것은 책상라며 나를 가르쳤다. [책상/NNG+라며/EC]

(나) 학습자가 ‘예요’를 써야하는 부분에서 ‘예요’로 쓴 경우는 ‘이/VCP+예요/EF’로 분석하지 않고 종결어미의 오형태로 분석한다.

■■ 저는 OO여학당 6급 학생이예요. [학생/NNG+이/VCP+예요/EF]

※ ‘아니다’는 부정 지정사(VCN)으로 분석한다.

다. 수식언

1) 관형사(MM)

- 관형사는 체언 앞에서 그것을 꾸미는 품사를 말한다. 관형사는 지시관형사, 수관형사, 성상관형사로 세분될 수 있는데, 본 분석에서는 이를 세분하여 분석하지 않는다.

■■ 각(各) 각 가정 [각/MM]

■■ 그까짓 그까짓 일 [그까짓/MM]

■■ 전(全) 전 국민 [전/MM]

■■ 현(現) 현 정권 [현/MM]

<주의사항>

(가) 관형사는 때로 문맥에 따라 다른 품사로 분석될 가능성이 있으니 문맥을 잘 살펴서 분석해야 한다.

① 관형사, 명사 통용

- 올 예산이 다 바닥이 났다. [올/MM]
■■ 올 들어 물가가 많이 올랐다. [올/NNG]

② 관형사, 부사 통용

- 단 세 명에서 그 일을 꾸몄다. [단/MM]
■■ 단, 그 일은 해서는 안 된다. [단/MAJ]

③ 관형사, 명사, 부사 통용

- 이내 마음을 어찌 알리오. [이내/MM]
■■ 아침 들판에 이내가 끼었다. [이내/NNG]
■■ 그는 이내 떠나갔다. [이내/MAG]

(나) 수사가 명사를 단독으로 수식하는 경우 그것을 관형사로 분석하기 쉬우나, ‘수’를 나타내는 말 가운데서 앞서 언급한 수관형사를 제외하고는 수사는 오로지 수사로만 분석한다. 즉, 수사의 관형사적 쓰임을 인정하지 않는 것이다. 따라서 다음과 같이 ‘다섯’은 모든 환경에서 중의성 없이 ‘수사’로만 분석된다. (1.3 수사 [2]주의사항 참고)

- 다섯이 먹기에 충분하다. [다섯/NR+이/JKS]
■■ 다섯 명이 앉아 있었다. [다섯/NR]

(다) 접미사 ‘-적(的)’이 붙는 경우는 조사와의 결합여부와 관계없이 모두 명사로 분석한다.

- 명사의 부사적인 용법 [부사/NNG+적/XSN+이/VCP+L/ETM]
■■ 명사의 부사적 용법 [부사/NNG+적/XSN]

2) 부사(MA)

- 부사는 주로 용언을 꾸며서 그 뜻을 더 세밀하고 분명하게 해 주는 품사를 말한다. 여기서는 부사를 세분하지 않고, 접속부사와 일반부사로만 나누기로 한다.

가) 접속부사(MAJ)

<주의사항>

① 접속부사는 종종 용언의 활용형으로도 쓰일 수 있으므로 주의한다.

- 그래서 마지막에는 조심하라고 했지? [그래서/MAJ]
■■ 영희가 그래서 결석을 했구나. [그렇/VA + 어서/EC]

② '그리고나서', '그래도'의 분석

- 그리고 나서 [그리/MAG+하/XSV+고/EC || 나/VX+서/EC]
■■ 그래도 [그러/VV+어도/EC]

※ 접속부사는 《표준국어대사전》에 접속부사로 뜻풀이된 것만 인정한다.
아래는 《표준국어대사전》의 접속부사 목록이다.

건테, 고로01 「2」, 그래서, 그러나, 그러니까, 그러면, 그러므로, 그런데, 그럼01 「1」, 그렇지마는, 그렇지만, 그리고, 그리하여, 근테01, 단06, 따라서, 연이나, 연중에, 연즉, 이리하여, 하건만, 하기는, 하기가, 하긴, 하물며, 하지만, 한테03

※ 용언의 활용형

- 그래, 그래도, 그래야, 그러니, 그러다가, 그러매, 그러면서, 그러자, 그렇다면, 그렇잖아도, 그리한 즉

※ 일반 부사

- 게다가, 곧, 다만, 또, 또는, 또한, 및, 예컨대, 요컨대, 왜냐하면, 이러
테면, 한편, 혹시, 혹은

※ [보완] 사전에 등재되지 않은 부사의 약어는 본딴말과 같은 표지로 분석한다.

- 그니까(그러니까), 글고(그리고)/MAJ
■■ 왜냐면(왜냐하면)/MAG

나) 일반부사(MAG)

<주의사항>

- ① 일반부사는 종종 일반명사와 동일한 형태로 구분이 어려운 경우가 있다. 이들은 뒤에 조사가 결합하느냐 여부와, 문맥에서 후행 명사를 수식하느냐의 여부에 따라 부사와 명사로 분석될 수 있다.

■■ 너의 <u>진짜</u> 속셈이 무엇인지 말해 봐라.	[진짜/NNG]
■■ 그 수학 문제는 <u>진짜</u> 어려웠다.	[진짜/MAG]
■■ <u>지금</u> 이 공부하기 딱 좋은 때이다.	[지금/NNG+이/JKS]
■■ 나는 <u>지금</u> 막 집에 도착했다.	[지금/MAG]

- ② 부사적인 용법을 가졌음에도 불구하고 일반부사가 아닌 일반명사로만 표준국어대사전에 등재되어 있는 단어는 오로지 일반명사로만 분석한다.

■■ 구석구석, 무작정, 여기저기, 오랫동안, 이곳저곳, 정작, 좌우간, 처음, 최근, 한때

- ③ 일반부사로 분석하기 쉬운 활용상의 불완전동사인 ‘덜달아, 더불어’는 모두 동사로 옳게 분석해야 함에 주의한다.

■■ 너는 <u>덜달아</u> 왜 난리니?	[덜달/VV+아/EC]
■■ 우리 함께 <u>더불어</u> 살아가자.	[더불/VV+어/EC]

- ④ ‘명사+없이’는 원칙적으로 ‘일반명사+없이/MAG’로 분석하지만, 아래와 같이 하나의 단어로 굳어져 사전에 등재된 경우는 ‘없이’ 통합형 자체를 하나의 일반부사로 분석한다.

■■ 관계없이, 그지없이, 꾸밈없이, 꾸밈없이, 난데없이, 남김없이 등

라. 독립언: 감탄사(IC)

- 감탄사는 화자의 부름이나 느낌, 놀람이나 대답을 직접적으로 나타내는 품사를 말한다.

■■■ 그럼(요), 야호, 어머, 앓, 아, 예, 그래(요), 아니(요), 글썄, 참, 아이구, 와아, 오호, 세상에

<주의사항>

- ① 사람이 입으로 직접 내는 소리를 대상으로 하되, 흉내를 내는 의도가 없는 것과 본능적인 놀람이나 느낌을 나타내는 것을 대상으로 한다. 또한 감탄사와 혼동되는 부사로서 음성상징어류의 부사어가 있는데, 이는 감탄사가 아닌 일반부사로 분석한다.

■■■ 야호! 드디어 정상이다.

[야호/IC+!/SF]

■■■ 쿨럭쿨럭 기침을 했다.

[쿨럭쿨럭/MAG]

- ② 동물의 울음소리 등은 감탄사가 아니라 일반부사로 분석한다.

■■■ 검둥이는 멍멍 짖으며 수풀 속으로 뛰어 갔다. [멍멍/MAG]

- ③ 욕이나 욕설을 나타내는 말은 전체를 감탄사로 분석한다.

■■■ 빌어먹을!

[빌어먹을/IC+!/SF]

- ④ ‘뭐’는 문맥에 따라 대명사와 감탄사의 두 가지 쓰임이 있다.

■■■ 원지도 모른 채

[뭐/NP+이/VCP+ㄴ 지/EF+도/JX]

■■■ 신문에 뭐 대단한 특종이라도 실렸습니까?

[뭐/IC]

- ⑤ 한 어절이 비정상적으로 늘어나거나 다른 기호가 개입되었을 경우 분석불능 범주(NA)로 분석한다.

■■■ 그러어엄/NA, 으~어~이/NA

마. 관계언⁹⁾

- 조사는 주로 체언과 결합하여 다른 말과의 문법적 관계를 나타내거나, 특별한 뜻을 더해 주는 품사를 말한다. 조사는 크게 격조사, 보조사, 접속조사로 나눈다. 한국어는 조사가 중첩하는 경우가 많은데, 이러한 경우 조사의 결합형은 분리해서 분석함을 원칙으로 한다.

■■ 부산에서도 대형 사고가 있었다.	[부산/NNP+에서/JKB+도/JX]
■■ 그녀의 약속이 갑자기 잡혔다.	[그녀/NP+와/JKB+의/JKG]

1) 격조사(JK)

- 이는 체언과 다른 성분 간의 일정한 문법 관계를 나타내는 조사이다.

가) 주격조사(JKS)

- 선행 체언으로 하여금 주어가 되게 하는 조사이다.

■■ 이/가	책이 보인다.	[책/NNG+이/JKS]
	나무가 보인다.	[나무/NNG+가/JKS]
■■ 께서	선생님께서 오신다.	[선생/NNG+님/XSN+께서/JKS]
■■ 서/이서	둘이서 그 일을 꾸몄다고?	[둘/NR+이서/JKS]
	혼자서 그 일을 꾸몄다고?	[혼자/NNG+서/JKS]
■■ 께오서	부대장님께오서	[부대장/NNG+님/XSN+께오서/JKS]
■■ 께옵서	황제께옵서 드나드신다.	[황제/NNG+께옵서/JKS]

※ 다음과 같이 체언 뒤에서 ‘이’가 첨가되어 나타나는 경우, 이때 ‘이’는 모두 주격 조사로 분석한다.

■■ 닭이가 울었다.	[닭/NNG+이/JKS+가/JKS]
■■ 책상이가 있다.	[책상/NNG+이/JKS+가/JKS]

9) 지침에 제시된 조사 목록에서 빠진 이형태와 예시를 추가함

■■ 친구들이 6월에 일이를 찾아있었다.	[일/NNG+이/JKS+를/JKO]
■■ 좋아하는 거는 옷입니다.	[옷/NNG+이/JKS+이/VCP+ㅂ니다/EF]
■■ <u>어른이들</u> 이 수많은 노력을	[어른/NNG+이/JKS+들/XSN+이/JKS]

나) 보격조사(JKC)

- 선행 체언으로 하여금 서술어 ‘되다, 아니다’의 보어가 되게 하는 조사이다. ‘되다, 아니다’ 앞의 조사 ‘이, 가’는 모두 보격조사로 분석한다.

■■ 이/가	얼음이 물이 되었다.	[물/NNG+이/JKC]
	씨앗이 열매가 되었다.	[열매/NNG+가/JKC]
	철수는 범인이 아니다.	[범인/NNG+이/JKC]
	범인은 남자가 아니다.	[남자/NNG+가/JKC]

다) 목적격조사(JKO)

- 선행 체언으로 하여금 목적어가 되게 하는 조사이다.

■■ ㄹ/을/를	수지가 널 좋아해.	[너/NP+ㄹ/JKO]
	민수는 음식을 많이 먹는다.	[음식/NNG+을/JKO]
	너는 바람 소리를 들었다.	[소리/NNG+를/JKO]

라) 관형격조사(JKG)

- 선행 체언으로 하여금 관형어가 되게 하는 조사이다.

■■ 의	나의 친구는 너 하나뿐이다.	[나/NP+의/JKG]
------	-----------------	--------------

마) 부사격조사(JKB)

- 선행 체언으로 하여금 부사어가 되게 하는 조사이다.

■■ 로/으로	망치로 못을 박아야지.	[망치/NNG+로/JKB]
	음식으로 장난치지 마.	[음식/NNG+으로/JKB]

■■ 로서/으로서	교사로서 책임을 다해야 한다.	[교사/NNG+로서/JKB]
■■ 로써/으로써	장관으로써 책임을 다해야 한다.	[장관/NNG+으로써/JKB]
	돌로써 지붕을 만든다고?	[돌/NNG+로써/JKB]
■■ 같이	콩으로써 메주를 쏜다고 해도	[콩/NNG+으로써/JKB]
■■ 더러	바보같이 웃고 다닌다.	[바보/NNG+같이/JKB]
■■ 랑/이랑	나더러 이것도 하라고 한다.	[나/NP+더러/JKB]
	너랑 많이 닮았다.	[너/NP+랑/JKB]
	오늘 동생이랑 싸웠다.	[동생/NNG+이랑/JKB]
■■로부터/ 으로부터	TV로부터 받는 영향력이	[TV/SL+로부터/JKB]
■■ 마냥	시험으로부터 해방되다	[시험/NNG+으로부터/JKB]
■■ 마따나	기영이마냥 놀 수만은 없다.	[기영이/NNP+마냥/JKB]
■■ 만큼	네 말마따나 나도 그래야 한다.	[말/NNG+마따나/JKB]
■■ 보고	눈물만큼 콧물도 흐른다니까.	[눈물/NNG+만큼/JKB]
■■ 보다	영자보고 놀자고 좀 해라.	[영자/NNP+보고/JKB]
■■ 에	직관보다는 논리가 동원돼야 한다.	[직관/NNG+보다/JKB+는/JX]
■■ 에게	나는 너에 대해 아무것도 모른다.	[너/NP+에/JKB]
■■ 에게서	너에게 말하기 싫다.	[너/NP+에게/JKB]
■■ 에서	나는 철수에게서 그 말을 들었다.	[철수/NNP+에게서/JKB]
■■ 에서부터	집에서 학교까지 너무 멀다.	[집/NNG+에서/JKB]
■■ 와/과	연구소에서부터 가게까지는	[연구소/NNG+에서부터/JKB]
	경미와 함께 다닌다면,	[경미/NNP+와/JKB]
	동생과 함께 다닌다면,	[동생/NNG+과/JKB]
■■ 처럼	사람처럼 행동하는 동물이 있다.	[사람/NNG+처럼/JKB]
■■ 하고	그 일하고 관련된 사람은	[일/NNG+하고/JKB]
■■ 한테	그 일은 경비한테 부탁해라	[경비/NNG+한테/JKB]

바) 호격조사(JKV)

- 주로 사람을 가리키는 체언 뒤에 연결되어 그것으로 하여금 부름의 대상이 되게 하는 조사이다.

■■ 아/야	호동아! 이제 그만 일어나거라	[호동/NNP+아/JKV+!/SF]
	철수야! 밥 먹어라	[철수/NNP+야/JKV+!/SF]
■■ 여/이여	주여, 우리에게 힘을 주소서	[주/NNG+여/JKV]
	슬픔이여, 안녕	[슬픔/NNG+이여/JKV]
■■ 시여/이시여	전능자시여 자비를 베풀어 주옵소서	[전능자/NNG+시여/JKV+!/SS]
	신이시여! 우리를 저버리지 마소서	[신/NNG+이시여/JKV+!/SS]

<주의사항>

- 호격조사와 어말어미는 구분해서 분석해야 한다.

■■ 저기 오는 것이 철수야. [철수/NNP+이/|/VCP+야/EF+./SF]

사) 인용격조사(JKQ)

- 인용문이나 인용구를, 동사에 대한 부사적 성분으로 도입하는 조사이다.

■■ 고	그는 "이제 가도 좋다"고 말했다.	[좋/VA+다/EF+ "/SS+고/JKQ]
■■ 라고/이라고	문제가 심각하다"라고 보고했다. 팻말에는 "금지구역"이라고 쓰여 있었다.	[심각/XR+하/XSA+다/EF+ "/SS+라고/JKQ] ["/SS+금지/NNG+구역/NNG+ "/SS+이라고/JKQ]
■■ 하고	영수는 "이제 가자"하고 말문을 닫았다.	[가/VV+자/EF+ "/SS+하고/JKQ]

<주의사항>10)

- ① 인용격조사는 연결어미와 구별하기 어려운 경우가 있으므로 주의한다. 인용 기호가 있을 경우에만 인용격조사로 분석하고, 인용기호가 없는 경우 연결어미로 분석한다.

(1) 인용격조사

- 팻말에는 "금지구역"이라고 쓰여 있었다.
["/SS+금지/NNG+구역/NNG+ "/SS+이라고/JKQ]
- 철수는 "다음 주에 놀러 가도 좋다"고 말하였다.
[좋/VA+다/EF+ "/SS+고/JKQ]
- 먼저 "주민등록증이 있냐?"고 묻는다.
[있/VV+냐/EF+?/SF+ "/SS+고/JKQ]

10) <세종> 분석 결과를 바탕으로 다시 정리함

(2) 연결어미

■■ 철수는 자기가 학생이라고 말했다.

[학생/NNG+이/VCP+라고/EC]

■■ 자장면을 시킨 뒤 집에 가겠다고 우기는 할머니를 달래기 시작했다.

[가/VV+겠/EP+다고/EC]

■■ 내가 안 기쁘냐고 다그쳐 물었을 때,

[기쁘/VA+냐고/EC]

※ [참고]

■■ 시골 아이라고 그것도 모르겠니?

[아이/NNG+라고/JX]

- ② 학습자 말뭉치에서는 생산자가 외국인 학습자이기 때문에 한국어에서 인용 기호로 구현되는 직접 인용, 간접 인용에 대한 지식이 없어 따옴표를 적지 못한 경우가 ‘문어’에서도 많이 발생한다. 이러한 경우는 <세종> 구어에서의 처리와 마찬가지로 인용 기호가 없더라도 직접 인용인 경우 인용격 조사로 분석한다.

■■ 내 일이다라고 말했다.

[일/NNG+이/VCP+다/EF+라고/JKQ]

※ [참고] 다음은 간접 인용의 경우로 보고 분석한다.

■■ 내 일이라고 말했다.

[일/NNG+이/VCP+다고/EC]

- ③ 인용 기호 중 하나인 <“ ”>은 맥락에 따라 인용이 아닌 강조를 위해 사용되기도 한다. 이때는 인용격 조사로 분석하지 않도록 주의한다.

■■ “사랑”이라는 건 뭘까?

[“/SS+사랑/NNG+”/SS+이/VCP+라는/ETM]

■■ 철수는 자기가 “학생”이라고 말했다.

[“/SS+학생/NNG+”/SS+이/VCP+라고/EC]

2) 접속조사(JC)

- 두 단어를 같은 자격으로 이어 주는 구실을 하는 조사를 말한다.

■■ 고/이고 그 사람은 염치고 체면이고가 없어.

[염치/NNG+고/JC]

책이고 책상이고 다 타 버렸다.

[책/NNG+이/고/JC]

■■ 와/과 그 아주머니는 딸기와 사과를 샀다.

[딸기/NNG+와/JC]

그 기계는 사람과 컴퓨터를 구별하지 못한다.

[사람/NNG+과/JC]

■ ■ 나/이나	사과나 배는 모두 몸에 좋은 과일이다.	[사과/NNG+나/JC]
	바자회 물품으로 책이나 옷을 받고 있다.	[책/NNG+이나/JC]
■ ■ 니/이니	시장에는 사과니 배니 과일이 잔뜩 있다.	[사과/NNG+니/JC]
	떡이니 과일이니 잔뜩 먹었다.	[떡/NNG+이니/JC]
■ ■ 다/이다	그는 농구다 축구다 못하는 운동이 없다.	[농구/NNG+다/JC]
	연습이다 레슨이다 시간이 하나도 없다.	[연습/NNG+이다/JC]
■ ■ 랑/이랑	머루랑 다래랑 먹으며 청산에 살고 싶어라.	[머루/NNG+랑/JC]
	떡이랑 과일이랑 많이 먹었다.	[떡/NNG+이랑/JC]
■ ■ 며/이며	잔치상에는 배며 대추며 여러 가지 과일이 차려져 있었다.	[배/NNG+며/JC]
	그림이며 조각이며 미술품으로 가득 찬 화실	[그림/NNG+이며/JC]
■ ■ 에	아버지가 책에, 연필에 많이 사 주셨다.	[책/NNG+에/JC]
■ ■ 하고	이번 준비물로 칼하고 연필을 샀다.	[칼/NNG+하고/JC]

<주의사항>

- ① ‘함께 함’의 뜻을 나타내는 접속조사는 부사격조사와 형태상 동일하므로 주의할 필요가 있다.

■ ■ 철수와 영희가 왔다.	[철수/NNP+와/JC]
■ ■ 철수와 같이 놀았다.	[철수/NNP+와/JKB]
■ ■ 철수랑 영희랑 왔다.	[철수/NNP+랑/JC 영희/NNP+랑/JC]

- ② 표준국어대사전에 조사로 등재(주로 구어체의 경우)된 ‘하며’는 조사로 인정하지 않고 ‘하/VV+며/EC’로 분석한다.

- ③ 접속 조사 중에서 ‘고/이고’, ‘니/이니’, ‘다/이다’, ‘며/이며’, ‘에’의 경우는 주로 ‘-고 -고’, ‘-니 -니’와 같은 구성에서 쓰인다. 이들 접속 조사는 연결어미와 동일한 형태인 경우가 있으므로 주의할 필요가 있다.

■ ■ 슬픔이고 기쁨이고 느끼지 못한다.	[슬픔/NNG+이고/JC]
■ ■ 그 옷은 개성적이고 색다른 현대 감각을 보여준다며,	[개성/NNG+적/XSN+이/VCP+고/EC]

- 옷이며 신이며 흠어져 있었다. [옷/NNG+이며/JC]
 ■ 내부는 어지러운 공간이며, [공간/NNG+이/VCP+며/EC]
 같은 건물 안에 반드시 식당가가 있다.

④ [보완] 학습자의 오류로 인해 두 단어를 이어주는 병렬 구조가 제시되지 않더라도 의미상 접속 조사로 쓰인 경우에는 접속 조사로 분석한다.

- 친구에게 줄 꽃과 샀어요. [꽃/NNG+과/JC]
 ■ 나에게 준 배려심이나 사람을 얼마나 많은지
 어떻게 계산하는지 이제 마음속에는 다 알게 되었다. [배려심/NNG+이나/JC]

3) 보조사(JX)

- 체언이나 부사 또는 용언의 연결 어미나 종결 어미의 뒤에 쓰여 특별한 뜻을 더해 주는 조사를 말한다.

■ 그러/그래	좋습니다그러.	[좋/VA+습니다/EF+그러/JX+./SF]
■ 까지(꺼정/까장)	걸어서 하늘까지	[하늘/NNG+까지/JX]
■ 깨나	힘깨나 쓰게 생겼다.	[힘/NNG+깨나/JX]
■ 나/이나	너나 가라!	[너/NP+나/JX]
	그것이나 가져라.	[그것/NP+이나/JX]
■ 나마/이나마	네 덕에 늦게나마 일을 마쳤다.	[늦/VA+게/EC+나마/JX]
	빵이나마 먹어라.	[빵/NNG+이나마/JX]
■ ㄴ/은/는	난 학생이다.	[나/NP+ㄴ/JX]
	오늘은 금요일이다.	[오늘/NNG+은/JX]
	이 종이는 어제 사 온 것이다.	[종이/NNG+는/JX]
■ ㄴ커녕/은커녕/는커녕	빨린커녕 천천히도 못 건졌다	[빨리/MAG+ㄴ커녕/JX]
	돈은커녕 먹을 쌀도 없다.	[돈/NNG+은커녕/JX]
	돕기는커녕 방해할 생각만 했다.	[돕/VV+기/ETN+는커녕/JX]
■ 다	물건을 거기다 놓아라.	[거기/NP+다/JX]
	그 물건을 거기에다 놓아라.	[거기/NP+에/JKB+다/JX]
■ 다가	책상을 어디다가 둘까요?	[어디/NP+다가/JX]
	집에다가 놓아 두어라.	[집/NNG+에/JKB+다가/JX]
■ 대로	철수는 철수대로 고민이 있다.	[철수/NNP+대로/JX]
■ 따라	오늘따라 버스도 안 온다.	[오늘/NNG+따라/JX]
■ 도/두	강아지도 주인은 알아본다.	[강아지/NNG+도/JX]

■■ 란/이란	코알라란 호주에 사는 초식동물이다. 사람이란 분수를 지킬 줄 알아야 한다.	[코알라/NNG+란/JX] [사람/NNG+이란/JX]
■■ ㄹ랑/일랑	강엘랑 가지 마라. 그 일에 대해선 걱정일랑 하지 말아라.	[강/NNG+에/JKB+ㄹ랑/JX] [걱정/NNG+일랑/JX]
■■ 마다	꽃마다 독특한 향기가 있다.	[꽃/NNG+마다/JX]
■■ 마저	장미마저 시들고 말았다.	[장미/NNG+마저/JX]
■■ 만	사람은 뽕만으로 살 수 없다.	[뽕/NNG+만/JX+으로/JKB]
■■ 밖에	이제는 떠날 수밖에 없다.	[수/NNB+밖에/JX]
■■ 부터	우선 노약자부터 태워야 한다.	[노약자/NNG+부터/JX]
■■ 뿐	가진 것은 집 한 채뿐이다.	[채/NNB+뿐/JX+이/VCP+다/EF]
■■ 서꺼	국물이나 동치미서꺼 아무 거나	[동치미/NNG+서꺼/JX]
■■ 사/이사	내사 그걸 이미 했지. 남이사 무슨 상관이야.	[내/NP+사/JX] [남/NNG+이사/JX]
■■ 야/이야	그야 그렇지. 그가 인간성이야 그만이지.	[그/NP+야/JX] [인간성/NNG+이야/JX]
■■ 야말로/이야말로	사과야말로 가을의 과일이다. 통일이야말로 최대의 과업이지.	[사과/NNG+야말로/JX] [통일/NNG+이야말로/JX]
■■ 요	나는 그림을요 잘 그립니다.	[그림/NNG+을/JKO+요/JX]
■■ 조차	이젠 봄조차 빼앗기는구나.	[봄/NNG+조차/JX]
■■ 치고	값싼 물건치고 쓸 만한 게 없지.	[물건/NNG+치고/JX]

(1) 보조사 분석 기준

- 보조사는 ‘이다’의 활용어미와 구분하기 어려운 경우가 있다. 흔히 보조사로 간주되던 몇몇 형태들은 연결어미와 의미상의 차이가 없으며, 분포상으로도 구별되지 않기 때문에 이런 대상들은 보조사로 분석하지 않는다.

[기준 1] 대상 형태가 용언의 어미로 사용되는가.

[기준 2] 대상 형태가 체언에 후행할 때 서술어의 자격을 가지고 사용되는가.

(가) [기준 1, 2]에 부합하는 다음의 형태들은 모두 ‘연결어미’로 분석한다.

■■ (이)ㄴ들, (이)ㄴ즉, (이)든, (이)든지, (이)라도, (이)라서, (이)라야

(나) [기준 1, 2]에 부합하지 않는 다음의 형태들은 ‘보조사’가 된다.

■■ (이)나마, (이)야, (이)ㄹ랑, (이)야말로, (이)란

(다) [기준 1]에 부합하지 않으나, [기준 2]에는 부합하는 형태는 ‘중의성’을 가진다.

■■ (이)나, (이)요

(라) 다음의 형태는 서술격조사 ‘이다’의 활용형과는 관계가 없으므로 모두 보조사가 된다.¹¹⁾

■■ 까지, 깨나, 는(은/ㄴ), 대로, 도, 따라, 마다, 마저, 만, 밖에, 부터, 뿐, 조차, 치고, ㄴ 커녕

※ [참고] ‘만’, ‘뿐’은 의존 명사로도 분석될 수 있음.

(마) 종결어미 뒤에 나타나는 ‘든지, 든가, 거나’ 등의 경우는 보조사로 분석한다.

■■ 공부를 잘한다든지 운동을 잘한다든지 [잘/MAG+하/XSV+ㄴ 다/EF+든지/JX]

■■ 시기라든가 질투라든가 하는 데에까지 [시기/NNG+이/VCP+라/EF+든가/JX]

■■ 그녀는 예쁘다거나 귀엽다거나 하는 [예쁘/VA+다/EF+거나/JX]

<주의사항>

(가) 다음의 형태들은 분석 결과에 중의성이 생기므로, 이들을 분석할 때는 특히 주의해야 한다.

■■ (이)란	코알라란 동물은 호주에 주로 서식한다.	[코알라/NNG+이/VCP+란/ETM]
	코알라란 매우 귀여운 동물이다.	[코알라/NNG+란/JX]
■■ (이)나	밥이나 빵을 먹도록 해라.	[밥/NNG+이나/JC]
	그가 비록 열심히 하나 능력은 부족하다.	[하/VV+나/EC]
	어제 내가 술을 마셨나?	[마시/VV+었/EP+나/EF+?/SF]

11) [삭제] 말고

→ ‘말고’는 ‘표준국어대사전’에 보조사로 등재되어 있지 않으며, 세종 말뭉치에서도 보조사로 분석하지 않았으므로 목록에서 삭제함.

■■ (이)야	철수야 그 일을 할 수 있지.	[철수/NNP+야/JX]
	내가 좋아하는 것은 철수야.	[철수/NNP+이/VCP+야/EF+./SF]
	철수야! 부르는 소리	[철수/NNP+야/JKV]
■■ (이)요	밥을 먹다가요	[먹/VV+다가/EC+요/JX]
	밥이요, 빵이요.	[밥/NNG+이/VCP+요/EC]

(나) ‘중결어미+요(보조사)’는 중결어미로 통합하여 분석한다.

■■ 말씀대로 했는걸요. [하/VV+았/EP+는걸요/EF+./SF]

(다) ‘비중결어미+요(보조사)’는 통합하지 않고 각각 분석해 준다.

■■ 제가 몸이 좀 아파서요 지각을 했어요. [아프/VA+아서/EC+요/JX]

■■ 내가요, 왜요? [내/NP+가/JKS+요/JX]

[왜/MAG+요/JX+?/SF]

(라) [보완] 보조사 ‘요’의 분석

(1) A: 선생님이 집에 오셨었어.

B: 선생님이요? [선생/NNG+님/XSN+이/JKS+요/JX]

A: 커서 선생님이 되는 게 어떠니?

B: 선생님이요? [선생/NNG+님/XSN+이/JKC+요/JX]

(2)¹²⁾ A: 선생님에 대해 알고 있니?

B: 선생님이요? [선생/NNG+님/XSN+이요/JX]

A: 가장 좋아하는 명절이 언제예요?

B: 설날이요 [설날/NNG+이요/JX]

(마) ‘말고’는 용언 ‘말다’의 활용형으로 처리한다.

■■ 돈말고 지혜가 필요하다. [돈/NNG+말/VV+고/EC]

12) 본 지침에서는 세종 현대 구어 말뭉치의 분석 결과를 따라 (2)와 같은 경우를 ‘이요/JX’로 분석한다.

바. 의존형태

1) 어미¹³⁾

가) 선어말어미(EP)

- 용언이 활용할 때, 어간과 어말 어미 사이에 나타나는 것으로 높임법이나 시제, 양태를 나타내는 문법적인 요소이다. 선어말어미의 목록은 연구자에 따라 다를 수 있으나 이 분석에서는 아래의 것만을 선어말어미로 인정한다.

■■ -겠-	그 일은 내일 처리하겠다.	[처리/NGG+하/XSV+겠/EP+다/EF]
■■ -(으)시-	선생님께서 손수 만드신	[만들/VV+시/EP+ㄴ/ETM]
	삼촌은 형님이 있으시다.	[있/VV+으시/EP+다/EF]
■■ -오/으오/ 옵/으옵-	어머님께 선물을 받치오니	[받치/VV+오/EP+니/EC]
	책을 읽으오니	[읽/VV+으오/EP+니/EC]
	어머님께 선물을 받치옵고	[받치/VV+옵/EP+고/EF]
	책을 읽으옵고	[읽/VV+으옵/EP+고/EC]
■■ -았/었-	그는 집에 갔다.	[가/VV+았/EP+다/EF+./SF]
	우리가 먹었던 음식이 잘못됐다.	[먹/VV+었/EP+던/ETM]
■■ -았었/었었-	거기는 전에 갔었던 곳이다.	[가/VV+았었/EP+던/ETM]
	우리가 먹었던 음식에 문제가 있다.	[먹/VV+었었/EP+던/ETM]

<주의사항>

- ① 선어말어미가 한 음절로 통합된 경우에는 각각 분리해서 분석한다.

■■ -셨- 그 일은 어머니께서 하셨다. [하/VV+시/EP+었/EP+다/EF+./SF]

- ② 다음의 선어말어미는 그 어간이 생략되었을 경우에 어간을 복원해 준다.

■■ -겠- 이것은 그대로 두어야겠다. [두/VV+어야/EC+하/VX+겠/EP+다/EF+./SF]

■■ -았/었- 철수가 그것을 가져오랬다. [가져오/VV+라/EF+하/VV+았/EP+다/EF+./SF]

13) 지침에 제시된 어미 목록에서 빠진 이형태와 예시를 추가함

■■ -(으)시- 선생님께서 가자시오. [가/VV+자/EF+하/VV+시/EP+오/EF+./SF]

③ 위의 선어말어미가 포함되지 않은 어미 형태는 그대로 연결어미로 분석한다.

■■ -랄까-, -대야-, -래야-

④ [보완] ‘-여’나 ‘-였-’은 ‘-아’나 ‘-았-’으로 수정한 후 분석한다.

■■ 공부를 하였다. [하/VV+았/EP+다/EF]

■■ 공부를 열심히 하여 시험을 잘 보았다. [하/VV+아/EC]

나) 종결 어미(EF)

- 용언의 어간이나 선어말 어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 한 문장을 끝맺는 역할을 하는 어미이다.

■■ -거든	나는 이것이 좋거든!	[좋/VA+거든/EF+!/SF]
■■ -게	그만한 돈이 있으면 좋게.	[좋/VA+게/EF+./SF]
■■ -구나/는구나	넌 정말 멋지구나!	[멋지/VA+구나/EF+?/SF]
	앞이 잘 안 보이는구나.	[보이/VV+는구나/EF+./SF]
■■ -구려/는구려	당신도 가시겠구려.	[가/VV+시/EP+겠/EP+구려/EF+./SF]
	잘도 먹는구려.	[먹/VV+는구려/EF+./SF]
■■ -구면/는구면	학교가 참 크구면.	[크/VA+구면/EF+./SF]
	공부를 잘하는구면.	[잘/MAG+하/XSV+는구면/EF+./SF]
	이것이 무엇인가?	[무엇/NP+이/VCP+ㄴ가/EF+?/SF]
■■ -ㄴ가/은가/는가	그것이 좋은가?	[좋/VA+은가/EF+?/SF]
	그가 집에 있는가?	[있/VV+는가/EF+?/SF]
	이제 시작인걸.	[시작/NNG+이/VCP+ㄴ걸/EF+./SF]
■■ -ㄴ걸/은걸/는걸	그 책은 벌써 다 읽은걸.	[읽/VV+은걸/EF+./SF]
	그는 벌써 갔는걸.	[가/VV+았/EP+는걸/EF+./SF]
■■ -나	자네 그리로 가나?	[가/VV+나/EF+?/SF]
	키가 얼마나 크냐?	[크/VA+냐/EF+./SF]
■■ -냐/으냐/느냐	물이 얼마나 깊으냐?	[깊/VA+으냐/EF+?/SF]
	그것보다 이것이 낫느냐?	[낫/VA+느냐/EF+?/SF]
■■ -냐고/으냐고 /느냐고	그가 누구냐고?	[누구/NP+이/VCP+냐고/EF+?/SF]
	그렇게 싫어? 싫으냐고?	[싫/VA+으냐고/EF+?/SF]
	너 뭐 해? 뭐 하느냐고?	[하/VV+느냐고/EF+?/SF]
■■ -네	정말 큰일 났네!	[나/VV+았/EP+네/EF+!/SF]
■■ -니	그게 없니?	[없/VA+니/EF+?/SF]

■■ -다/ㄴ다/는다	그게 사실이다. 이건 말도 안 된다. 아이가 글을 잘 읽는다.	[사실/NNG+이/VCP+다/EF+./SF] [되/VV+ㄴ다/EF+./SF] [읽/VV+는다/EF+./SF]
■■ -다구/ㄴ다구 /는다구	돈이 많다구? 너도 간다구? 소설책을 읽는다구?	[많/VA+다구/EF+?/SF] [가/VV+ㄴ다구/EF+?/SF] [읽/VV+는다구/EF+?/SF]
■■ -다나/ㄴ다나 /는다나	그도 가겠다나. 나를 잘 안다나. 건강한 여자를 찾는다나.	[가/VV+겠/EP+다나/EF+./SF] [알/VV+ㄴ다나/EF+./SF] [찾/VV+는다나/EF+./SF]
■■ -다네/ㄴ다네 /는다네	일을 망쳤다네 우리네 짧은 인생도 간다네. 평소에도 한복을 잘 입는다네.	[망치/VV+었/EP+다네/EF+./SF] [가/VV+ㄴ다네/EF+./SF] [입/VV+는다네/EF+./SF]
■■ -다니까/ㄴ다니까 /는다니까	돈이 없단니까! 어머니가 오늘은 꼭 오신다니까. 내 말을 믿지를 앓는다니까.	[없/VA+다니까/EF+!/SF] [오/VV+시/EP+ㄴ다니까/EF+!/SF] [앓/VX+는다니까/EF+./SF]
■■ -다니/ㄴ다니 /는다니	서울이 이렇게 변화하다니. 이 긴 시를 어떻게 외운다니? 이 많은 책을 언제 읽는다니?	[변화/XR+하/XSV+다니/EF+?/SF] [외우/VV+ㄴ다니/EF+?/SF] [읽/VV+는다니/EF+?/SF]
■■ -다면서/ㄴ다면서 /는다면서	술은 싫다면서? 니가 축구를 잘한다면서? 달팽이도 먹는다면서?	[싫/VA+다면서/EF+?/SF] [잘/MAG+하/XSV+ㄴ다면서/EF+?/SF] [먹/VV+는다면서/EF+?/SF]
■■ -다오/ㄴ다오 /는다오	그가 가지고 있다오. 꽃은 이른 봄에 핀다오. 이 나무는 열매를 많이 맺는다오.	[있/VX+다오/EF+./SF] [피/VV+ㄴ다오/EF+./SF] [맺/VV+는다오/EF+./SF]
■■ -단다/ㄴ단다 /는다단다	나도 슬프단다. 선생님께서 공부를 가르쳐 주신단다. 누에는 뽕잎을 먹는단다.	[슬프/VA+단다/EF+./SF] [주/VX+시/EP+ㄴ단다/EF+./SF] [먹/VV+는다단다/EF+./SF]
■■ -도다/는다다	꽃이 아름답도다. 짐이 조서를 내리는도다.	[아름답/VA+도다/EF+./SF] [내리/VV+는다다/EF+./SF]
■■ -ㄹ걸/을걸	모른다고 할걸. 생각만큼 쉽지 않을걸.	[하/VV+ㄹ걸/EF+./SF] [않/VX+을걸/EF+./SF]
■■ -ㄹ게/을게	그렇게 할게. 남은 밥은 내가 먹을게.	[하/VV+ㄹ게/EF+./SF] [먹/VV+을게/EF+./SF]
■■ -ㄹ까/을까	이제 밥을 할까? 이 과자는 내가 먹을까?	[하/VV+ㄹ까/EF+?/SF] [먹/VV+을까/EF+?/SF]
■■ -렴/으렴	맘대로 해 보렴. 이것 좀 먹으렴.	[보/VX+렴/EF+./SF] [먹/VV+으렴/EF+./SF]
■■ -려무나/으려무나	더 놀다 가려무나. 책이나 읽으려무나.	[가/VV+려무나/EF+./SF] [읽/VV+으려무나/EF+./SF]
■■ -라니까/으라니까	그 사람이 아니라니까.	[아니/VCN+라니까/EF+./SF]

	가만히 있으라니까.	[있/VV+으라니까/EF+./SF]
■■ -ㅁ세/음세	그날 꼭 음세.	[오/VV+ㅁ세/EF+./SF]
	곧 밥을 먹음세.	[먹/VV+음세/EF+./SF]
■■ -ㅂ니까/습니까	이제야 옵니까?	[오/VV+ㅂ니까/EF+?/SF]
	그래도 되겠습니까?	[되/VV+겠/EP+습니까/EF+?/SF]
■■ -ㅂ니다/습니다	이렇게 합니다.	[하/VV+ㅂ니다/EF+./SF]
	정말 재미있습니다.	[재미있/VA+습니다/EF+./SF]
■■ -ㅂ시다/읍시다	다시 만납시다.	[만나/VV+ㅂ시다/EF+./SF]
	여기 앉읍시다.	[앉/VV+읍시다/EF+./SF]
■■ -ㅂ시오/읍시오	서둘러 주십시오.	[주/VX+시/EP+ㅂ시오/EF+./SF]
	여기 앉읍시오.	[앉/VV+읍시오/EF+./SF]
■■ -ㅂ디까/습디까	신부가 예쁘디까?	[예쁘/VA+ㅂ디까/EF+?/SF]
	보기에 좋습디까?	[좋/VA+습디까/EF+?/SF]
■■ -ㅂ디다/습디다	참 좋은 곳입디다.	[곳/NNB+이/VCP+ㅂ디다/EF+./SF]
	덕수궁에 사람이 많습디다	[많/VA+습디다/EF+./SF]
■■ -세/으세	제대로 좀 하세.	[하/VV+세/EF+./SF]
	이 책을 우리 함께 읽으세.	[읽/VV+으세/EF+./SF]
	함께 가.	[가/VV+아/EF+./SF]
■■ -아/어/여	밥 먹어!	[먹/VV+어/EF+!/SF]
	같이 해.	[하/VV+아/EF+./SF]
■■ -야	그건 사실이 아니야.	[아니/VCN+야/EF]
■■ -아라/어라	웃기지 말아라.	[말/VX+아라/EF+./SF]
	천천히 먹어라.	[먹/VV+어라/EF+./SF]
	물이 깨끗하오.	[깨끗/XR+하/XSA+오/EF+./SF]
■■ -오/으오/소	나는 요즘 논어를 읽으오.	[읽/VV+으오/EF+./SF]
	그 곳에는 내가 가겠소.	[가/VV+겠/EP+소/EF+./SF]
■■ -자	잠이나 자자.	[자/VV+자/EF+./SF]
■■ -자꾸나	약속을 좀 미루자꾸나.	[미루/VV+자꾸나/EF+./SF]
■■ -자니까	그만 따지자니까.	[따지/VV+자니까/EF+./SF]
■■ -지	그가 언제 오지?	[오/VV+지/EF+?/SF]

<주의사항>

(가) ‘종결어미+요’는 통합해서 종결어미로 분석한다.

■■ 말씀대로 <u>했는걸요</u> .	[하/VV+았/EP+는걸요/EF+./SF]
■■ 뭐 <u>먹었는데요?</u>	[먹/VV+었/EP+는데요/EF+?/SF]

※ [참고] ‘비종결어미+요’는 통합해서 분석하지 않는다.

■■■ 그 애는 노래는 잘 부르는데요. [부르/VV+는데/EC+요/JX+./SF]

춤은 잘 못 춰요.

■■■ 어제 비가 많이 와서요. 지각을 했어요. [오/VV+아서/EC+요/JX+./SF]

(나) ‘-세요’는 다음과 같이 선어말어미까지 분석한다.

■■■ 어서 출근하세요. [출근/NNG+하/XSV+시/EP+어요/EF+./SF]

(다) ‘-죠’는 축약형을 그대로 분석한다.

■■■ 어서 출근하죠. [출근/NNG+하/XSV+죠/EF+./SF]

※ [참고] 다음의 경우는 <표준>을 따라 종결어미로 분석한다.

■■■ 아픈데 밥을 먹을까 싶다. [먹/VV+을까/EF]

■■■ 진짜 부자유친이 아닐까 생각합니다. [아니/VCN+ㄹ까/EF]

■■■ 돈을 어떻게 쓰느냐에 따라 [쓰/VV+느냐/EF+에/JKB]

■■■ 무슨 일이 있었는가 했다. [있/VV+었/EP+는가/EF]

■■■ 보통 사용할 때는 뭔가 게임을 할 때 [뭐/NP+이/VCP+ㄴ가/EF]

■■■ 언제 고향에 갈지 잘 모르겠습니다. [가/VV+ㄹ지/EF]

■■■ 왜 좋았는지 알아요. [좋/VA+았/EP+ㄴ지/EF]

■■■ 자기 적성에 맞는지 안 맞는지 고려하지 않습니다.

[맞/VV+는지/EF || 안/MAG || 맞/VV+는지/EF]

다) 연결 어미(EC)

- 용언의 어간이나 선어말 어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 문장을 종결시키지 못하고 뒤에 오는 절을 연결시켜 주는 어미를 말한다.

■■■ -거나	누가 오거나 알은 체 할 것 없다.	[오/VV+거나/EC]
■■■ -거니	비가 오겠거니 생각했다.	[오/VV+겠/EP+거니/EC]
■■■ -거늘	이미 늦었거늘 어찌 빨리 가는가?	[늦/VV+었/EP+거늘/EC]
■■■ -거든	가거든 말해라.	[가/VV+거든/EC]
■■■ -건대	내가 보건대, 네 말이 옳다.	[보/VV+건대/EC]
■■■ -건마는	말렸건마는 아직도 축축하다.	[말리/VV+었/EP+건마는/EC]
■■■ -게	개를 굵게 하지 마라.	[굵/VV+게/EC]
■■■ -고	일단 먹고 보자.	[먹/VV+고/EC]
	일을 하고 밥을 먹자.	[하/VV+고/EC]

■■ -곤	종종 지각하곤 했다.	[지각/NNG+하/XSV+곤/EC]
■■ -고자	병을 낫고자 몸부림쳤다.	[낫/VV+고자/EC]
■■ -기에	실수했기에 용서해 주었다.	[실수/NNG+하/XSV+았/EP+기에/EC]
■■ -ㄴ데/은데/ 는데	예쁜데 미워한다. 방이 좁은데 가구는 많다. 눈이 오는데 차를 가져가지 말까?	[예쁘/VV+ㄴ데/EC] [좁/VV+은데/EC] [오/VV+는데/EC]
■■ -ㄴ들/는들	간다 한들 아주 같까? 그걸 먹는들 뭐가 달라지겠나.	[하/VV+ㄴ들/EC] [먹/VV+는들/EC]
■■ -ㄴ즉/은즉	배가 고평즉 속이 쓰리다. 물이 맑은즉 고기가 많기는 어렵소.	[고프/VV+ㄴ즉/EC] [맑/VV+은즉/EC]
■■ -ㄴ지라/은지라/ 는지라	눈이 온지라 길이 미끄럽다. 기분이 좋은지라 다정하다. 선생님께서 고집을 굽히지 않으시는지라	[오/VV+ㄴ지라/EC] [좋/VV+은지라/EC] [않/VX+으시/EP+는지라/EC]
■■ -나/으나	눈이 오나 비가 오나 밥을 먹으나 마나이다.	[오/VV+나/EC] [먹/VV+으나/EC]
■■ -나니	멀리 보이나니 넓은 들이로다.	[보이/VV+나니/EC]
■■ -나마/으나마	도와주지는 못하나마 방해를 해서는 맛은 없으나마 많이 드세요.	[못하/VX+나마/EC] [없/VV+으나마/EC]
■■ -노니	묻노니, 포부가 무엇이냐?	[묻/VV+노니/EC+,,/SP]
■■ -니/으니	밥을 다 먹고 보니 배가 불렀다. 이 옷은 작으니 큰 것으로 바꿔 주세요.	[보/VX+니/EC] [작/VV+으니/EC]
■■ -느니	앉아서 걱정하느니 나가서 하겠다.	[걱정/NNG+하/XSV+느니/EC]
■■ -니까/으니까	웃기니까 좋다. 약속을 했으니까 만나야 한다.	[웃기/VV+니까/EC] [하/VV+았/EP+으니까/EC]
■■ -다가	자랑하다가 망신당했다.	[자랑/NNG+하/XSV+다가/EC]
■■ -다기에/ㄴ다기에 /는다기에	그녀가 예쁘다기에 보러 왔소. 앞으로 잘 한다기에 승낙했다. 빵을 먹는다기에 주었다.	[예쁘/VV+다기에/EC] [하/VV+ㄴ다기에/EC] [먹/VV+는다기에/EC]
■■ -다손/ㄴ다손 /는다손	밑다손 치더라도 구박하지 말자. 그가 제시간에 온다손 하더라도 내 앞의 음식은 다 먹는다손 치더라도	[밑/VV+다손/EC] [오/VV+ㄴ다손/EC] [먹/VV+는다손/EC]
■■ -대도/ㄴ대도 /는대도	시간이 있대도 만나 주질 않는다. 늦으면 큰일 난대도 서두르질 않아요. 떠들면 야단맞는대도 계속 떠들었다.	[있/VV+대도/EC] [나/VV+ㄴ대도/EC] [야단맞/VV+는대도/EC]
■■ -더라도	가더라도 꼭 돌아와라.	[가/VV+더라도/EC]
■■ -던들	진작 알았던들 방법을 취했지.	[알/VV+았/EP+던들/EC]
■■ -도록	미치도록 일했다.	[미치/VV+도록/EC]
■■ -든지	외모가 어떠하든지 무슨 상관인가?	[어떠/XR+하/XSA+든지/EC]

■■ -되	싸우되 꼭 지도록 해라.	[싸우/VV+되/EC]
■■ -ㄹ뿐더러/ 을뿐더러	비가 올뿐더러 바람도 분다. 그는 재산이 많을뿐더러 재능도 많다	[오/VV+ㄹ뿐더러/EC] [많/VV+을뿐더러/EC]
■■ -ㄹ수록/ 을수록	갈수록 태산이다. 이 책은 읽을수록 감동을 준다. 비가 얼마나 올지 천둥이 다 친다.	[가/VV+ㄹ수록/EC] [읽/VV+을수록/EC] [오/VV+ㄹ지/EC]
■■ -ㄹ지/올지	내일은 얼마나 날씨가 좋을지 오늘 밤하늘에 별이 유난히 빛난다.	[좋/VV+을지/EC]
■■ -ㄹ지라도/ 을지라도	이길지라도 명예롭지는 않다.	[이기/VV+ㄹ지라도/EC]
■■ -ㄹ지언정/ 을지언정	마음에 걱정이 있을지라도 내색하지 마라. 그것은 무모한 행동일지언정 죽을지언정 그 일은 못하겠다.	[있/VV+을지라도/EC] [행동/NNG+이/VCP+ㄹ지언정/EC] [죽/VV+을지언정/EC]
■■ -라고	바보라고 생각한다.	[바보/NNG+이/VCP+라고/EC]
■■ -락	오르락 내리락	[오르/VV+락/EC]
■■ -랍시고	그는 반장이랍시고 행패만 부린다.	[반장/NNG+이/VCP+랍시고/EC]
■■ -러/으러	청소하러 가자. 점심 먹으러 집에 간다.	[청소/NNG+하/XSV+러/EC] [먹/VV+으러/EC]
■■ -려/으려	학교에 가려 한다. 웃으려 한다.	[가/VV+려/EC] [웃/VV+으려/EC]
■■ -려니와/ 으려니와	비용도 문제려니와 일꾼도 문제다. 이 마을은 경치도 좋으려니와	[문제/NNG+이/VCP+려니와/EC] [좋/VV+으려니와/EC]
■■ -련마는/ 으련마는	보면 반가우련마는 볼 수가 없네. 벌써 제 잘못을 알았으련마는	[반갑/VV+련마는/EC] [알/VV+았/EP+으련마는/EC]
■■ -며/으며	노래하며 춤을 춘다. 강물이 맑으며 깊다.	[노래/NNG+하/XSV+며/EC] [맑/VV+으며/EC]
■■ -면/으면	지옥이 존재하면 만원일 것이다. 내일 날씨가 좋으면 소풍을 가겠다.	[존재/NNG+하/XSV+면/EC] [좋/VV+으면/EC]
■■ -면서/으면서	푸르면서 검은 물빛 밥을 먹으면서 신문을 본다.	[푸르/VV+면서/EC] [먹/VV+으면서/EC]
■■ -므로/으므로	비가 오므로 가지 않겠다. 강이 깊으므로 배 없이 건널 수 없다.	[오/VV+므로/EC] [깊/VV+으므로/EC]
■■ -아/어	입을 막아 버렸다. 밥을 먹어 버렸다.	[막/VV+아/EC] [먹/VV+어/EC]
■■ -아도/어도	암만 봐도 모르겠다. 나는 부자가 아니어도 행복하다.	[보/VV+아도/EC] [아니/VCN+어도/EC]
■■ -아서/어서	땀을 놓아서 췌을 잡았다. 그는 걸어서 학교에 갔다.	[놓/VV+아서/EC] [걸/VV+어서/EC]
■■ -아야/어야	이 일은 잘해야 한다. 사람은 먹어야 산다.	[잘/MAG+하/XSV+아야/EC] [먹/VV+어야/EC]

■■ -자마자	오자마자 당했다.	[오/VV+자마자/EC]
■■ -지	우기지 못해 버렸다.	[우기/VV+지/EC]
■■ -지마는	비가 오지마는 가야 한다.	[오/VV+지마는/EC]

<주의사항>

(가) 어미에 따라서는 분석의 중의성이 생길 수 있으므로 문맥 확인을 통해 형태분석을 결정한다.

■■ 너는 내가 왔는데 기쁘지도 않니?	[오/VV+았/EP+는데/EC]
■■ 내가 지금 있는 <u>데</u> 가 어디지?	[있/VV+는/ETMⅡ데/NNB+가/JKS]
■■ 다들 <u>만족</u> 하는지 아무런 불평이 없다.	[만족/NNG+하/XSV+는지/EC]
■■ 너를 <u>만난</u> <u>지도</u> 꽤 오래구나.	[만나/VV+ㄴ/ETMⅡ지/NNB+도/JX]

(나) ‘-음직’은 “음직/EC”로 분석한다. 그러나 ‘바람직, 먹음직’ 등은 그 자체가 하나의 어근이므로 더 이상 분석할 수 없다는 것에 유의한다.

■■ 어른답고 믿음직하게 행동해라.	[믿/VV+음직/EC+하/VX+게/EC]
■■ 그것 참 먹음직스럽다.	[먹음직/XR+스럽/XSA+다/EF+./SF]
■■ 그것은 매우 바람직한 일이다.	[바람직/XR+하/XSA+ㄴ/ETM]

라) 명사형 전성 어미(ETN)

- 한 문장의 성격을 임시로 바꾸어 다른 문장 속에서 명사적인 역할을 하게 하는 어미를 말한다.

■■ -기	그 일은 정말 중요하기 때문이다.	[중요/NNG+하/XSA+기/ETN]
■■ -ㅁ/-음	학생 신분임을 밝히다.	[신분/NNG+이/VCP+ㅁ/ETN]
	장사는 신용을 얻음이 제일이다.	[얻/VV+음/ETN+이/JKS]

<주의사항>

(가) 불규칙 용언 어간에 명사형 전성 어미가 붙어 있을 경우 ‘-음’이 아닌 ‘-ㅁ’으로 분석한다.

■■ 김철수 지음

[짓/VV+ㅁ/ETN]

(나) “음, 기”가 붙은 말이 단순히 명사형이나 아니면 굳어진 명사이냐 하는 것은 물론 문맥에 따라 결정되어야 하지만 먼저 그것이 “사전”에 등재되어 있느냐의 여부를 살펴보아야 한다.

■■ 책을 읽기가 어렵다.

[읽/VV+기/ETN+가/JKS]

■■ 읽기 교육이 문제가 된다.

[읽기/NNG]

마) 관형사형 전성 어미(ETM)

- 용언의 성격을 임시로 바꾸어 다른 문장 속에서 관형사적인 역할을 하게 하는 어미이다.

■■ -ㄴ/은	어제 떠난 사람	[떠나/VV+ㄴ/ETM]
	어제 먹은 빵에 이상이 있었다.	[먹/VV+은/ETM]
■■ -는	잃어버린 물건을 찾는 일은 어렵다.	[찾/VV+는/ETM]
■■ -던	이제까지 미루던 일을 오늘 해치웠다.	[미루/VV+던/ETM]
■■ -ㄹ/을	나에게는 아직 처리할 일이 있다.	[처리/NNG+하/XSV+ㄹ/ETM]
	물이 깊을 것이다.	[깊/VA+을/ETM]
■■ -런	어제런 듯하다.	[어제/NNG+이/VCP+런/ETM]

<주의사항>

(가) 불규칙 용언 어간에 관형사형 전성 어미가 있을 경우 ‘-은, -을’이 아닌 ‘-ㄴ, -ㄹ’로 분석한다.

■■ 그녀의 고운 얼굴

[곱/VA+ㄴ/ETM]

■■ 그녀는 매우 아름다울 것이다.

[아름답/VA+ㄹ/ETM]

(나) 종결 어미에 이어서 전성 어미가 올 경우 통합해서 전성어미로 처리한다.

■■ 어느 쪽에 더 비중을 두느냐는 것이

[두/VV+느냐는/ETM]

2) 체언 접두사(XPN)

- 접두사는 명사와 수사에 결합하는 접사류를 묶어서 체언접두사만을 설정하기로 한다.
- 명사 접두사에는 한자어계 접두사와 고유어계 접두사가 있는데, 그 목록의 풍부함에 비해 대개가 생산성이 그리 높지 않다. 일단 여기서는 비교적 생산성이 높다고 인정되는 접두사와, 접두사를 분리했을 경우 단일한 표제어로 등재될 수 있는 경우에 한해서 접두사 분석을하기로 한다.

가(假)-가건물, 고(高)-고물가, 과(過)-과보호, 구(舊)-구소련, 날-날음식, 노(老)-노부부, 대(大)-대선배, 만-만아들, 맨-맨몸, 무(無)-무의식, 미(未)-미완성, 반(反)-반독재, 범(汎)-범세계, 부(不)-부도덕, 불(不)-불합리, 비(非)-비논리, 생(生)-생김치, 소(小)-소강당, 신(新)-신정당, 왕(王)-왕족발, 재(再)-재충전, 저(低)-저임금, 제(第)-제13차, 준(準)-준전시, 초(超)-초만원, 최(最)-최고급, 친(親)-친러시아, 탈(脫)-탈냉전시대, 폐(廢)-폐광산, 풋-풋살구, 피(被)-피고소인, 한-한가운데, 헛-헛고생

※ [보완] 단, 예외적으로 ‘대부분, 대다수, 무조건’의 경우는 체언 접두사를 분리하지 않는다.

3) 접미사(XS)

- 파생 접미사에는 어기의 품사를 바꾸는 것과 그렇지 않은 것이 있는데, 이들을 별도로 구별하여 표지를 부여하지는 않는다.

가) 명사파생접미사(XSN)

- 명사파생접미사는 명사나 다른 어근에 후행하여 그것이 명사의 기능을 수행할 수 있도록 만들어 주는 의존 형태이다. 그러나 명사파생접미사는 연구자에 따라 그 목록이 다르며, 실제로도 구분이 애매한 경우가 많다. 본 분석에서는 접미사의 생산성과 접미사를 제외한 형태의 독립성을 기준으로 다음과 같이 목록을 마련하였다.

가(價)-매매가, 가(哥)-김가, 경(頃)-두 시경, 계(系)-몽고계, 계(界)-교육계, 광(狂)-메모광, 권(圈)-운동권, 권(權)-참정권, 당(當)-한 사람당, 대(臺)-억대, 덕(宅)-청주 덕, 론(論)-비평론, 별(別)-가구별, 여(餘)-삼십여, 류(類)-자연류, 률, 율(率)-경쟁 률, 리(裡)-비밀리, 분(分) 분량-일인분, 분(分)-3분의, 산(産)-중국산, 상(上)-역사 상, 생1(生)갑자생, 생2(生)견습생, 성(性)-인간성, 시(視)-영웅시, 용(用)-전쟁용, 적(的)-사상적, 형(型)-기본형, 형(形)-도시형, 제(制)-봉건제, 층(層)-선수층, 치(值)-보름치, 풍(風)-복고풍, 화(化)-도구화, 기-기름기, 께-10분께, 꿀-십 원꿀, 끼리-전우끼리, 궂-노름궂, 네-동이네, 님-선생님, 들-우리들, 들이-1ㄹ들이, 배기-열 살배기, 뽕-조카뽕, 씹-만원씹, 장이-간판장이, 쟁이-심술쟁이, 쯤 -내일쯤, 질-서방질, 짜리-백 원짜리, 째1 -이틀째, 째2-옹기째, 치레-인사치레, 투성이-먼지 투성이

<주의사항>

(가) 명사파생접미사인 ‘-들’은 그 분포가 매우 다양하여 일부에서는 이를 보조사와 접미사로 나누어 분석하기도 한다. 그러나, 본 분석에서는 이들을 모두 명사파생접미사로 처리한다. ‘먹고들’의 ‘-들’도 선행성분이 어미이긴 하나, 일치하는 대상은 선행하는 명사로 해석할 수도 있기 때문이다.

■■ 사람들이 우리 집에 왔다.

[사람/NNG+들/XSN]

■■ 그들은 밥을 먹고들 싶었다.

[먹/VV+고/EC+들/XSN]

(나) ‘-님’은 다음과 같이 세 가지의 분석 중의성을 가지므로 주의해서 분석한다.

① ‘임’의 의미로 쓰인 경우: 보통명사

■■ 님과 이별하다.

[님/NNG+과/JKB]

② 사람의 ‘이름’이나 ‘성’ 뒤에서 쓰인 경우: 의존명사

■■ 김철수 님께서 오셨습니다.

[김철수/NNP || 님/NNB+께서/JKS]

③ 그 밖의 경우: 명사파생접미사

■■ 과장님이 부르십니다.

[과장/NNG+님/XSN+이/JKS]

(다) 목록에 있는 접미사라도 사전에 등재되지 않은 명사나 어근과 함께 사용됐다면 전체를 명사로 분석한다.

■■ 획기적

[획기적/NNG]

나) 동사파생접미사(XSV) → ‘명사/부사/어근+동사파생접미사’로 분석한다.

- 동사파생접미사는 어기 또는 어근에 붙어서 그것을 동사로 만들어 주는 기능을 갖는 접미사이다.

※ 여기서는 그러한 접미사 중 생산성이 높은 아래의 넷만 동사파생접미사로 인정하여 분석한다.

■■ 당하	아군이 공격당하는 데에는 이유가 있다.	[공격/NNG+당하/XSV+는/ETM]
■■ 되	아침식사가 이미 준비되어 있었다.	[준비/NNG+되/XSV+어/EC]
■■ 시키	강아지를 운동시키려고 공원에 나갔다.	[운동/NNG+시키/XSV+려고/EC]
■■ 하	외국에서 공부하는 일이 쉬운 것은 아니다.	[공부/NNG+하/XSV+는/ETM]

<주의사항>

(가) ‘-하’ 접사는 생산성이 높기 때문에 모든 ‘N하다’가 표제어로 등재되어 있지 않다. ‘N 하다’와 같이 구로 보는 것은 의미적으로 명사를 수식하는 요소가 선행하는 것이 명확한 경우로만 한정하고 그 이외의 경우는 구로 보지 않고 ‘-하’를 접사로 처리한다.

■■ 외국에서 공부하는 것은 힘들다. [공부/NNG+하/XSV+는/ETM]

■■ 외국에서 공부 하는 것은 힘들다. [공부/NNG+하/XSV+는/ETM]

■■ 카페에서 벉락치기 공부 하는 것을 [공부/NNG || 하/VV+는/ETM]

■■ 카페에서 벉락치기 공부하는 것을 [공부/NNG || 하/VV+는/ETM]

(나) 학습자가 잘못 접미사를 사용한 경우 교정어절을 상정했을 때 교정어절의 품사가 동사일 때는 동사파생접미사, 교정어절의 품사가 형용사일 때

는 형용사파생접미사로 분석한다.

■■ 음식을 먹하다.	[먹/VV+하/XSV+다/EF+./SF]
■■ 마음이 <u>아프</u> 한 아주머니가 집에 돌아왔다.	[아프/VA+하/XSA+ㄴ/ETM]
■■ 그렇지 <u>않</u> 하다면	[않/VX+하/XSA+ㄴ 다면/EC]

다) 형용사파생접미사(XSA) → ‘명사/부사/어근+형용사파생접미사’로 분석한다.

- 형용사파생접미사는 어기나 어근에 붙어서 그것을 형용사로 파생시키는 접미사이다.

※ 여기서는 그러한 접미사 중 생산성이 높은 아래의 다섯만 형용사파생접미사로 인정하여 분석한다.

■■ 답	사람이 사람답게 행동해야 사람이지	[사람/NNG+답/XSA+게/EC]
■■ 되	자식된 도리로 어떻게 그런 짓을..	[자식/NNG+되/XSA+ㄴ/ETM]
■■ 롭	어려운 일일수록 슬기롭게 대처하라.	[슬기/NNG+롭/XSA+게/EC]
■■ 스럽	그녀의 사랑스러운 표정을 보거라.	[사랑/NNG+스럽/XSA+ㄴ/ETM]
■■ 하	멍청한 표정을 짓지 말아라.	[멍청/XR+하/XSA+ㄴ/ETM]

4) 어근(XR)

※ [보완] 표준국어대사전에 등재된 2음절 이상의 어근만 어근으로 인정하여 분석한다.

■■ 따듯도 하다	[따뜻/XR+도/JX] [하/VV+다/EF]
■■ 이러하다	[이러/XR+하/XSA+다/EF]

<주의사항>

- 어근의 분석 대상은 표준국어대사전의 표제어 중 2음절 이상의 어휘이다. ‘하다’가 결합한 어휘 중 ‘하다’에 선행하는 음절이 1음절일 경우에는 어근 분리 현상이 매우 제한적이므로 이 경우에는 통합형으로 분석한다.

■■ 듣직하다	[듣직/XR+하/XSA+다/EF]
---------	--------------------

■■ 취하다 [취하/VV+다/EF] '취'는 어근
(위하다, 반하다, 강하다, 약하다, 중하다, 대하다, 의하다, 통하다 등)

■■ 밥하다 [밥/NNG+하/XSV+다/EF] '밥'은 명사
(절하다, 인하다, 비하다, 한하다 등)

■■ 잘되다 [잘/MAG+되/XSV+다/EF] '잘'은 부사

※ “못하다”의 경우 ‘못하/VV, 못하/VA, 못하/VX’의 세 가지 경우가 존재하는데, 이때 ‘-하’를 분석할 경우 본용언의 분석과 보조용언의 분석이 동형이 되기 때문에 예외로 취급해서 ‘-하’를 분석하지 않는다.

■■ 노래를 못한다. [못하/VV+ㄴ 다/EF]

■■ 음식 맛이 저번보다 못하다. [못하/VA+다/EF]

■■ 밥을 먹지 못한다. [못하/VX+ㄴ 다/EF]

※ 참고

■■ 숙제를 못 했다. [못/MAG || 하/VV+았/EP+다/EF]

사. 기타

1) 기호

- 영문이나 한자, 기호 등이 어절 중간에 개입하여 올바른 분석이 불가능한 경우에는 각각의 요소를 분리하여 분석한다. 이 경우 표지를 줄 수 없는 불완전한 형태가 생길 수 있다.

■■ 마이크로소프트(microsoft)사 [마이크로소프트/NNP+(/SS+microsoft/SL+)/SS+사/NNG]

■■ 농·수산물 [농/NNG+·/SP+수산물/NNG]

■■ 초·중·고 [초/NNG+·/SP+중/NNG+·/SP+고/NNG]

■■ 위, 아래 집 [위/NNG+/,SP+아랫집/NNG]

cf. ■■ 대~박 [대/NA+~/SS+박/NA]

2) 준말

- 준말은, 그것이 본딤말과 대등하게 사용되고 분석결과가 동일한 어절 단위를 형성할 경우에 한해서만 복원한다. 그러나 다음에서처럼, 본딤말로 복원할 경우 어절 수에 변화가 생길 뿐 아니라 본딤말로 복원하는 정도가 일관성을 띠지 않게 되는 경우는 굳이 복원하지 않는다. 그러나, 이러한 원칙이 모든 경우에 일관적으로 적용될 수 있는 것은 아니다. 결국 준말의 처리는 해당 어절에 따라 임의적일 수 있다.

■■ 라는	[라는/ETM]	(○)
	[라고/JKQ 하/VV+는/ETM]	(×)
■■ 려는	[려는/ETM]	(○)
	[려고/EC 하/VX+는/ETM]	(×)

3) 분석불능범주(NA)

- ※ 그 자체가 사전에 등재되어 있지도 않으면서, 축약의 정도가 심하거나 분석하기 어려운 방언형의 경우 분석불능범주로 처리한다.

■■ 담배가 <u>쪼매턴게</u> 하마 자라서 빠나?	[쪼매턴게/NA]
■■ 친구한테 전화를 <u>적긴</u> 일이었다.	[적긴/NA]
■■ “부산국제영화” <u>제가니와</u>	[제가니와/NA]
■■ <u>있잖아</u> 요	[있/VV+잖/NA+아요/EF]
■■ ㅋㅋ	[ㅋㅋ/NA]
■■ ππ	[ππ/NA]
■■ ○ㅋ○ㅋ	[○ㅋ○ㅋ/NA]
■■ ^^	[^^/NA]

4) 합성어

- 합성어는 표준국어대사전에 등재되어 있는 것만을 인정한다.

구성		띄어쓰기 상태(학습자)	분석 방법
N+N 구성	사전 등재	1. 국어사전(‘-’로 등재)	국어사전/NNG
		2. 국어 사전(‘-’로 등재)	국어사전/NNG
		3. 국어 교육(‘^’로 등재)	국어/NNG 교육/NNG
		4. 국어교육(‘^’로 등재)	국어/NNG + 교육/NNG
	사전 미등재	1. 국어연구	국어/NNG 연구/NNG
		2. 국어 연구	국어/NNG 연구/NNG
본 용언 + 보조 용언 구성	사전 등재	1. 좋아하다	좋아하/VV+다/EF
		2. 좋아 하다	좋아하/VV+다/EF
	사전 미등재	1. 가보다	가/VV+아/EC+보/VX+다/EF
		2. 가 보다	가/VV+아/EC 보/VX+다/EF

<주의사항>

(가) 표제어가 사전의 표제어로 등록되어 있는 경우는 그대로 분석한다.

■■ 정치권력 (사전: 정치-권력) [정치권력/NNG]

(나) 합성어로 등재되어 있되 띄어쓰기를 허용한 합성어는 세분하여 분석하는 것을 원칙으로 한다.

■■ 학생운동 (사전표기: 학생^운동) [학생/NNG+운동/NNG]
[학생/NNG || 운동/NNG]

(다) 합성어로 등록되어 있지 않은 표제어는 분리해서 분석하되, 사전 표제어로 등록되어 있는 최대한 많은 음절수의 단어를 생성하도록 나눈다.

(라) 3음절 어휘와 같이 어느 쪽으로 나뉘어도 음절수가 같고, 양쪽 분석이 모두 사전 표제어라면 뒤쪽을 먼저 분석한다.

■■ 차창밖 [차/NNG+창밖/NNG]

■■ 이등품 [이/NR+등품/NNG]

5) 접사처럼 쓰이는 ‘명사’의 처리

- 일부 명사는 사전에 ‘(일부 명사 뒤/앞에 붙어)~의 뜻을 나타내는 말.’로 등재되며, 이들은 앞뒤에 함께 쓰인 명사와 합쳐서 명사로 분석한다. 이들의 목록은 다음과 같다.

가01 「04」	(일부 명사 뒤에 붙어) ‘주변4’의 뜻을 나타내는 말.	강가 //넷가 //우물가.
감03 「02」	(옷을 뜻하는 명사 뒤에 붙어) ‘옷을 만드는 재료’의 뜻을 나타내는 말.	한복감//양복감.
감03 「04」	(일부 명사 뒤에 붙어) ‘자격을 갖춘 사람’의 뜻을 나타내는 말.	신랑감//머느릿감//사윗감//장군감.
감03 「05」	(일부 명사 뒤에 붙어) 대상이 되는 도구, 사물, 사람, 재료의 뜻을 나타내는 말.	구경감 //놀림감 //떨감 //양념감 //안춧감 //장난감//웃음감//사형감//노벨상감//마느질감.
값 「07」	(일부 명사 뒤에 붙어) ‘가격’, ‘대금’, ‘비용’의 뜻을 나타내는 말.	기름값 //물값 //물건값 //부식값 //신문값 //우옷값 //음식값.
값 「08」	(일부 명사 뒤에 붙어) ‘수치’의 뜻을 나타내는 말.	변숫값//분석값//위상값//저항값
과04 「02」	(일부 명사 뒤에 붙어) 학과나 전문 분야를 나타내는 말.	국어과 //마취과 //물리학과.
구15 「03」	(일부 명사 뒤에 붙어) ‘법령 집행을 위하여 정한 구획’의 뜻을 나타내는 말.	선거구 //투표구.
구이01 「02」	(일부 명사 뒤에 붙어) 구운 음식의 뜻을 나타내는 말.	갈비구이//생선구이//참새구이.
군03 「02」	(일부 명사 뒤에 붙어) 왕자군을 뜻하는 말.	경녕군 //복성군.
군05 「02」	(일부 명사 뒤에 붙어) ‘군대3’의 뜻을 나타내는 말.	시민군//예비군//유엔군//진압군.
극04 「02」	(일부 명사 뒤에 붙어) ‘연극’, ‘드라마’ 따위의 뜻을 나타내는 말.	고발극//사극//실험극//특집극
금06 「04」	(일부 명사 앞에 붙어) ‘금색1’, ‘금제1’의 뜻을 나타내는 말.	금두꺼비 //금목걸이 //금수저.
급04 「05」	(직급 따위를 나타내는 일부 명사 뒤에 붙어) ‘그 직급’의 뜻을 나타내는 말.	과장급 //부장급 //간부급.
길01 「10」	(일부 명사 뒤에 붙어) ‘과정’, ‘도중’, ‘중간’의 뜻을 나타내는 말.	산책길//시장길
꽃01 「07」	(일부 명사 뒤에 붙어) ‘그 꽃’의 뜻을 나타내는 말.	도라지꽃//무궁화꽃//목련꽃//민들레꽃//사과꽃//유채꽃.
난05 「02」	(고유어와 외래어 명사 뒤에 붙어) ‘구분된 지면’의 뜻을 나타내는 말.	어린이난//가십난//컴퓨터난//해외 토픽난.
놀이01 「04」	(일부 명사 뒤에 붙어) ‘모방4’, ‘장난’, ‘흥내’의 뜻을 나타내는 말.	시장놀이//병원놀이//엄마놀이//학교놀이.
대15 「03」	(일부 명사 뒤에 붙어) 받침이 되는 시설이나 이용물의 뜻을 나타내는 말.	급수대 //조희대 //독서대.
택01 「04」	(일부 명사 뒤에 붙어) ‘택호’를 나타내는 말.	윤 판서택

덩어리 「03」	(일부 명사 뒤에 붙어) [같은 말] 덩이(3. 그러한 성질을 가지거나 그런 일을 일으키는 사람이나 사물을 나타내는 말).	꿀칫덩어리 //심술덩어리 //애꿎덩어리//제주덩어리.
덩이 「03」	(일부 명사 뒤에 붙어) 그러한 성질을 가지거나 그런 일을 일으키는 사람이나 사물을 나타내는 말. [비슷한 말] 덩어리.	꿀칫덩이 //심술덩이.
란01	(한자어 명사 뒤에 붙어) ‘알’의 뜻을 나타내는 말.	수정란//무정란.
란02	(한자어 명사 뒤에 붙어) ‘구분된 지면’의 뜻을 나타내는 말. ‘칸01’으로 순화.	광고란//독자란//투고란.
란03 「01」	(한자어 뒤에 붙어) ‘난초’의 뜻을 나타내는 말.	금자란//문주란//은란.
량05	(한자어 명사 뒤에 붙어) 분량이나 수량의 뜻을 나타내는 말.	가사량//노동량//작업량.
례01 「01」	(일부 명사 뒤에 붙어) ‘본보기’의 뜻을 나타내는 말.	인용례//판결례
마님 「02」	(일부 명사 뒤에 붙어) 상전(上典)을 높여 이르는 말.	대감마님 //영감마님.
마마 「04」	(임금 및 그의 가족과 관련된 명사 뒤에 붙어) ‘존대’의 뜻을 나타내는 말.	대비마마//대왕마마.
망09 「02」	(일부 명사 뒤에 붙어) 그물처럼 얽혀 있는 조직이나 짜임새의 뜻을 나타내는 말.	교통망 //연락망 //점포망//유통망//판매망.
명02 「02」	(일부 명사 뒤에 붙어) ‘이름’의 뜻을 나타내는 말.	곡명//작품명//저자명
모12 「03」	(일부 명사 앞에 붙어) 어떠한 것에서 갈려 나오거나 생겨난 것의 근본이 됨의 뜻을 나타내는 말.	모기업 //모은행.
무침 「02」	(일부 명사 뒤에 붙어) ‘양념을 해서 무친 반찬’의 뜻을 나타내는 말.	시금치무침//복어무침//꿀 뱅이무침//파래무침.
문06 「01」	(일부 명사 뒤에 붙어) 학술 전문의 종류를 나타내는 말.	어학문 //법학문.
문06 「02」	(일부 명사 뒤에 붙어) 씨족에 따른 집안을 나타내는 말.	강씨문(姜氏門) //이씨문(李氏門).
미14 「02」	(일부 명사 앞 또는 뒤에 붙어) ‘아름다움’의 뜻을 나타내는 말.	미소년 //송고미 //우아미//각선미//교양미// 백치미//미남자.
반10 「03」	(일부 명사 뒤에 붙어) ‘작은 집단’의 뜻을 나타내는 말.	단속반//작업반
밭01 「05」	(일부 명사 뒤에 붙어) 그 식물이나 자연물, 수산물 따위가 많이 나는 곳.	고추밭 //대나무밭 //흙밭 //파래밭.
병03 「02」	(일부 명사 뒤에 붙어) ‘병사2’의 뜻을 나타내는 말.	운전병//탈영병.
병04 「02」	(일부 명사 뒤에 붙어) ‘질병2’의 뜻을 나타내는 말.	간질병 //심장병.
병05 「03」	(일부 명사 뒤에 붙어) ‘용기’를 나타내는 말.	농약병 //링거병 //요구르트병 //참기름병 //플라스틱병.
볶음 「02」	(일부 명사 뒤에 붙어) 볶아서 만든 음식의 뜻을 나타내는 말.	쇠고기볶음 //야채볶음.
불09 「02」	(일부 명사 뒤에 붙어) ‘부처1’의 뜻을 나타내는 말.	무량수불 //아미타불.

	나타내는 말.	
비05 「03」	(일부 명사 뒤에 붙어) ‘비율2’의 뜻을 나타내는 말.	농도비 //혼합비.
비19 「03」	(일부 명사 뒤에 붙어) 기념하여 세운 물건의 뜻을 나타내는 말.	문학비 //문인비.
빛 「07」	(일부 명사 뒤에 붙어) ‘빛깔’의 뜻을 나타내는 말.	능금빛 //산빛.
상04 「03」	(일부 명사 뒤에 붙어) ‘상차림’을 나타내는 말.	다과상 //생신상 //차례상.
상23 「02」	(일부 명사 뒤에 붙어) 조각이나 그림을 나타내는 말.	성당의 성모 마리아상.
상23 「03」	(일부 명사 뒤에 붙어) ‘모범2’, ‘본보기’의 뜻을 나타내는 말.	교사상 //어머니상.
상25 「02」	(일부 명사 뒤에 붙어) ‘상장10’, ‘상패4’, ‘상품4’ 따위의 뜻을 나타내는 말.	감독상 //봉사상 //선행상 //작품상 //효행상.
색03 「05」	(일부 명사 뒤에 붙어) ‘색깔’의 뜻을 나타내는 말.	딸기색 //바이올렛색.
선14 「07」	(일부 명사 뒤에 붙어) ‘광선1’의 뜻을 나타내는 말.	감마선 //엑스선.
식04 「04」	(일부 명사 뒤에 붙어) ‘수법’, ‘수식’을 나타내는 말.	곱셈식 //덧셈식 //나눗셈식 //뺄셈식.
쌍02 「03」	(일부 명사 앞에 붙어) ‘두 짝으로 이루어짐.’의 뜻을 나타내는 말.	쌍가락지 //쌍가마 //쌍권총.
씨01 「05」	(일부 식물이나 동물을 나타내는 명사 뒤에 붙어) 그 식물이나 동물의 씨를 나타내는 말.	배추씨//살구씨//굴씨//조개씨.
안04 「04」	(일부 명사 뒤에 붙어) ‘안건’의 뜻을 나타내는 말.	개정안 //채택안 //협상안.
알01 「09」	(일부 식물이나 동물을 나타내는 명사 뒤에 붙어) 그 식물이나 동물의 알을 나타내는 말.	머루알//은행알//타조알.
액03 「02」	(일부 명사 뒤에 붙어) ‘액체’의 뜻을 나타내는 말.	냉각액 //링거액 //수정액.
양20 「02」	(고유어와 외래어 명사 뒤에 붙어) 분량이나 수량을 나타내는 말.	구름양//알칼리양.
옥03 「02」	(일부 명사 앞에 붙어) ‘옥색1’, ‘옥제2’의 뜻을 나타내는 말.	옥제털이 //옥매트 //옥침대.
왜03 「03」	(일부 명사 앞에 붙어) ‘일본식의’, ‘일본의’의 뜻을 나타내는 말.	왜간장 //왜모시.
은04 「02」	(일부 명사 앞에 붙어) ‘은색’, ‘은제3’의 뜻을 나타내는 말.	은갈치 //은귀고리 //은목걸이 //은찰잔.
자08 「03」	(일부 명사 앞에 붙어) 모체에 딸려 있음을 나타내는 말.	자회사.
잡이01 「04」	(일부 명사 뒤에 붙어) 민속놀이나 전통 음악에서 기술이나 재주, 장단 따위를 이르는 말.	
재비01	(일부 명사 뒤에 붙어) 국악에서, 악기를 연주하거나 노래를 부르거나 춤을 추는 기능자를 이르는 말.	가야금재비 //춤재비 //노래재비.
조15 「03」	(일부 명사 뒤에 붙어) 특정한 임무나 역할을 맡아 수행하기 위하여 조직하는 작은 집단을 나타내는 말.	작업조 //폭파조.

조림01 「02」	(일부 명사 뒤에 붙어) 조리 음식의 뜻을 나타내는 말.	고등어조림 //연근조림.
주24 「05」	(일부 명사 뒤에 붙어) ‘주식’의 뜻을 나타내는 말.	우량주//전환주.
주머니 「03」	(일부 명사 뒤에 붙어) 무엇이 유난히 많은 사람을 비유적으로 이르는 말.	고생주머니 //병주머니 //피주머니 //근심주머니
즙 「02」	(먹을 것을 나타내는 일부 명사 뒤에 붙어) ‘농축액’을 나타내는 말.	미나리즙 //석류즙 //배즙 //양파즙 //쥬즙.
직06 「04」	(일부 명사 뒤에 붙어) ‘직무’, ‘직분’, ‘직업’, ‘직위’의 뜻을 나타내는 말.	사제직//사도직.
집01 「09」	(일부 명사 뒤에 붙어) 물건을 팔거나 영업을 하는 가게를 나타내는 말.	갈빗집 //고깃집 //꽃집 //피자집.
집01 「10」	(일부 명사 뒤에 붙어) ‘택호’를 나타내는 말.	“그럼, 이 집 택호는 영월집이라고 합시다. 알기 쉽게…….”
찜01 「02」	(일부 명사 뒤에 붙어) 찜 음식의 뜻을 나타내는 말.	갈비찜 //아귀찜.
책01 「04」	(일부 명사 뒤에 붙어) ‘서적’임을 나타내는 말.	국어책//소설책//요리책.
터01 「04」	(일부 명사 뒤에 붙어) ‘자리1’나 ‘장소5’의 뜻을 나타내는 말.	낚시터 //놀이터 //일터 //휴터.
튀김01 「02」	(일부 명사 뒤에 붙어) 튀긴 음식의 뜻을 나타내는 말.	새우튀김 //오징어튀김
티02 「02」	(일부 명사 뒤에 붙어) ‘어떤 태도나 기색’의 뜻을 나타내는 말.	막내티 //소녀티 //중년티 //춘티.
표05 「07」	(일부 명사 뒤에 붙어) ‘그 사람이 만든 물건’의 뜻을 더하는 말.	엄마표 //아빠표 //신랑표 //주부표.
후08 「03」	(일부 명사 앞에 붙어) ‘뒤나 다음’의 뜻을 나타내는 말.	후더침 //후보름 //후서방.

아. 구어 형태 분석 말뭉치

※ 구어 전사 말뭉치의 특성

- 구어 말뭉치에서 마침표는 하나의 문장이 끝났음을 나타내는 것이 아니라 억양 단위를 나타내는 기호이므로 주석할 때 주의해야 한다.

■■ 친구랑 같이 여행 왔어요 음

[오/VV+았/EP+어요/EF]

→ 구어 말뭉치에서 문장 기호는 억양 단위를 의미하기 때문에 위와 같은 예시에서

‘왔어요’ 뒤에 임의로 마침표를 추가하지 않도록 한다.

→ 또한, 문장 기호가 없어서 자동 주석에서는 종결어미를 대부분 연결어미로 분석하는데, 이 경우 종결어미로 분석해야 한다.

■■ 그냥 매일매일 <u>쉬고</u> .	[쉬/VV+고/EC]
북경하고 고향에 갔다 왔다 갔다 <u>왔다</u> .	[오/VV+았/EP+다/EC]
<u>했어요</u> .	[하/VV+았/EP+어요/EF]
■■ 어~ 그렇게 어 가고 싶지 <u>않았어요</u>	[않/VX+았/EP+어요/EF]
돈이 <u>없어서</u> .	[없/VA+어서/EC]
■■ 아까 드렸던 종이를 한번 <u>살펴보고요</u> ,	[살펴보/VV+고/EC+요/JX]
다음으로 넘어갈게요.	

→ 구어 전사는 문장 단위가 아니라 억양 단위로 전사가 되기 때문에 하나의 발화가 여러 개의 억양 단위로 나뉘어져 제시될 수 있다. 따라서 연결어미도 종결부에 위치할 수 있어 분석을 할 때에 주의해야 한다.

→ 억양 단위를 나타내는 문장 기호 역시 위의 예시처럼 연결어미 뒤에서도 나타날 수 있기 때문에, 맥락에 따라 어미를 구분해 분석해야 한다.

- 구어 형태 분석은 문어 형태 분석 지침을 따르지만 불완전하게 발화되거나 자기 수정을 하는 등의 끊어진 발화나 억양 단위 발화와 같이 구어 말뭉치의 특성을 드러내는 경우 아래와 같이 분석한다.

1) 완전한 어절

- 기본적으로 발화가 완전히 이뤄진 어절은 문어 형태 분석 지침을 따라 분석한다.

■■ 저희가 <u>하여</u> 하고	[하/VV+아/EC]
■■ 충고를 해 <u>준</u> <u>쥘</u> 줘는데	[주/VX+L/ETM 주/VX+어/EC]

2) 끊어진 어절

- 끊어진 어절은 어절의 일부만 발화된 경우나 불분명한 경우이다. 끊어진

채로 발화된 어절은 형태 단위가 온전히 발화되었을 때만 분석하고, 형태 단위가 온전히 발화되지 못해 형태소가 확보되지 못한 경우는 분석불가능 (NA)으로 처리한다.

가) 어절의 일부만 발화되었지만 분석이 가능한 경우

■■ 어 손의 관절에 통증= 통증이	[통증/NNG]
■■ 미국과 같= 같은	[같/VA]
■■ 정말 아름다우= 아름다운 곳이에요.	[아름다우/VA]
■■ 그랬= 그랬어요.	[그러/VV+었/EP]
■■ 비결인=인 것 같은데.	[이/VCP+ㄴ/ETM]

나) 어절의 일부만 발화되어 분석이 불가능하거나 중의성이 발생하는 경우

■■ 음 플라스 될 수 아이다, 마이너= 아, 플라스는	[마이너/NA]
■■ 어 운동할 시= 힘도 부족해서	[시/NA]
■■ 어제 약= 약= 약국에 갔어요	[약/NA 약/NA]
■■ 한국인과 가= 같은	[가/NA]
■■ 의사 선생님도 곧 나= 나올 수 있다고	[나/NA]
■■ 경복궁에 가 봐= 봤다.	[보/VV+아/NA]
■■ 스페인에서 있어= 있어요.	[있/VV+어/NA]
■■ 슬퍼=퍼 가지구	[퍼/NA]

<주의사항>

- 합성어와 파생어 앞에서 이들 어휘의 일부가 끊어진 채로 발화된 경우 형태 단위가 온전해 분석이 가능한 경우는 분석하고 그렇지 못한 경우는 NA 표지를 부여한다.

■■ 많= 많이 주세요	[많/VA]
■■ 일본인 학습= 학습자들이	[학습/NNG]
■■ 일상= 일상생활에서 주로 공부를 해요	[일상/NNG]
■■ 한국 음식을 좋아= 좋아하지만	[좋/VA+아/EC]
한국 음식을 좋아하= 좋아하지만	[좋아하/VV]
■■ 한국어 공부가 힘= 힘들었지만	[힘/NNG]

다) 어절의 일부를 더듬으며 반복하는 경우 (용언의 경우)

- 결혼하 하기 한 [결혼/NNG+하/XSV||하/XSV+기/ETN||하/XSV+ㄴ/ETM]
■■ 그런 게 제일 그 비결이라 라고 하면 [비결/NNG+이/VCP+라/EF||라고/EC]

라) 어절 중간에 간투사 따위가 들어가는 경우

- 심약 어 하다 [심약/NNG || 어/IC || 하/VV+다/EF]
■■ 좋아 어 하다 [좋/VV+아/EC || 어/IC || 하/VX+다/EF]

3) 억양 단위가 바뀐 어절

가) 억양 단위가 형태소 경계로 바뀐 경우

- 발화자가 불완전하게 발화한 것은 아니지만 한 어절을 발화하는 도중에 억양 단위가 바뀌어서 조사나 어미 등 문법 형태소가 실질 형태소와 다른 억양 단위로 전사될 때, 억양 단위를 통합하지 않고 경계를 살려 형태 주석한다. 하지만 주석은 통합했을 때의 표지를 부여한다.

- 캐 유 두 미어 페이버? [페이버/NNG]
가 무슨 뜻? [가/JKS]
■■ 좋아. [좋/VV+아/EF]
라고 대답했지. [라고/JKQ]
■■ 주부 우울증. [우울증/NNG]
이라고 말할 수 있겠습니까. [이/VCP+라고/EC]
■■ 공부. [공부/NNG]
한다고 [하/XSV+ㄴ 다고/EC]

나) 억양 단위가 형태소를 가르는 경우

- 형태소 중간에 억양 단위가 바뀌어서 다른 억양 단위로 전사될 때, 각각 분석불가능(NA)으로 처리한다.

- 어. [어/NA]
제는 별일 없었어. [제/NA+는/JX]

■■ 선두주
자가 도착했다.

[선두/NNG || 주/NA]
[자/NA+가/JKS]

4) 불분명한 어절 (X)

- 잘 들리지 않아 추측하여 전사한 어절은 최대한 분석하고 분석이 불가능한 경우에는 분석불가능(NA)으로 처리한다.

■■ 소리 중에 XXX 이게	[XXX/NA]
■■ 교육 개방 XX안이	[XX안/NA+이/JKS]
■■ XX에 제출돼	[XX/NA+에/JKB]
■■ XX스의 이론을	[XX스/NA+의/JKG]
■■ 신발을 X다	[X다/NA]

5) 간투사의 처리

가) 그, 저

- 그, 저 : 조사가 붙어 있다면 ‘대명사’, 조사 없이 확실하게 뒤의 명사를 수식할 때는 ‘관형사’, 위의 경우가 아니거나 확실하게 감탄사로 사용된 경우에는 ‘감탄사’로 처리한다. (※ 구분이 애매한 경우 감탄사로 분석한다.)

■■ 그는 참으로 좋은 사람이다.	[그/NP+는/JX]
■■ 그 책 좀 이리 쥘 봐.	[그/MM 책/NNG]
■■ 그 무엇인가를 알아내고자 했지만	[그/MM 무엇/NP+이/VCP+ㄴ가/EF+를/JKO]
■■ 그 왜 있잖아요.	[그/IC]
■■ 이도 저도 다 싫다.	[저/NP+도/JX]
■■ 저 둘 중에 하나를 선택해라.	[저/MM 둘/NR 중/NNB+에/JKB]
■■ 저, 뭐라더라..	[저/IC]
■■ 저 말씀 중에 잠시 실례하겠습니다..	[저/IC]

나) 아니

- 아니 : 대답이나 감탄일 때는 ‘감탄사’, 부정이나 반대의 뜻을 나타낼 때나

명사와 명사 또는 문장과 문장 사이에서 강조할 때는 ‘부사’로 처리한다.

- ■ A : 자니?
B : 아니, 안 자. [아니/IC]
B' : 아니요, 안 자요. [아니요/IC]
■ ■ 아니, 그럴 수가 있니? [아니/IC]
■ ■ 아침까지만 해도, 아니 점심 먹을 때만 해도... [아니/MAG]

다) 그래

- 그래 : 대답이나 감탄, 놀라움, 담화 표지로 쓰였을 경우는 ‘감탄사’, 서술어의 대응으로 쓰였을 경우에는 용언의 활용형으로 분석한다.

- ■ A : 점심에 같이 밥 먹을까?
B : 그래, 알겠어. [그래/IC]
■ ■ A : 점심에 같이 밥 먹을까요?
B : 그래요. 뭐 먹을까요? [그래/IC+요/JX+?/SF]
■ ■ 왜 그래요? [그렇/V+어요/EF+?SF]
[그러/VV+어요/EF+?SF]

<주의사항>

(가) [보완] 맥락에 따라 감탄사로 쓰였는지 판단이 어려운 경우가 있다. 이때 그 형태가 선·후행 형태소와 같을 때는 선·후행 형태소를 반복한 것으로 분석하고, 그렇지 않은 경우에는 감탄사로 처리한다.

- ■ 학교에 에 [에/JKB]
가서 에 반 친구를 만났어요 [에/IC]
■ ■ 그 그 사람은 제 친구예요. [그/NP||그/NP||사람/NNG+은/JX]
어학당에서 그 처음 만났어요. [그/IC]

(나) [보완] 감탄사가 반복되는 경우에는 구어 전사에서 구분한 어절 경계에 따라 형태 주석한다.

- ■ 네네 맞아요. [네네/IC]
■ ■ 네 네 네 그래서 [네/IC||네/IC||네/IC]

6) 구어형의 분석

- 세종 문어 형태 분석 지침에는 구어형 분석에 대한 기술이 자세하지 않다. 따라서 기본적으로는 문어 형태 분석 지침을 중심으로 여기서도 분석을 하지만, 일부 해결할 수 없는 경우에 한해서는 세종 구어의 형태 분석을 따른다.

■■ 뭘로

[무엇/NP+으로/JKB]

■■ 걸로

[것/NNB+으로/JKB]

- 문어 지침에서 대명사 ‘뭐’와 의존 명사 ‘거’가 그 형태가 유지되지 않고 조사와 축약되어 나타나는 경우에는 각각 ‘무엇’과 ‘것’으로 복원하고 있다. 구어에서도 위와 같은 문어 지침을 따라 원형을 복원해준다.

■■ 그쵸, 그쵸

[그쵸/IC], [그쵸/IC]

■■ 이케 하면 되나요?

[이케/MAG]

■■ 여따 집어넣어

[여따/MAG]

- 음운적 축약이 일어나 형태 분석이 불가능한 경우는 해당 축약형 전체가 가지는 기능을 고려해 형태 표지를 할당한다.

한국어 학습자 말뭉치 오류 주석 지침

I. 학습자 말뭉치 오류 주석 체계 틀

1. 기본 주석

- 오류 위치는 오류가 나타난 부분의 품사를 주석한다. 오류 위치는 기본 주석으로 형태소 분석에 기대어 모든 오류에 대해 오류가 발생한 품사에 전수 주석한다. 오류는 오류 위치 검색으로 찾을 수 있다.¹⁴⁾

	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

14) 구 단위 주석과 표현 문형 주석은 구 전체와 구 구성요소에 각각 주석함을 원칙으로 한다.

	오류 유형		주석 표지
		접두사	CXPN
		명사파생접미사	CXSN
		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	
	표현 문형		PE

2. 확장 주석

- 확장 주석은 한국어 교육의 선행 오류 연구에서 유의미한 주석에 초점을 두어 교수자의 활용에 초점을 둔 주석이다. 연구자들은 필요한 주석을 추가하여 스스로 주석할 수 있다. 교정 어절에 대한 형태 주석에 기대어 주석한다.

2.1. 오류 양상

○ 어휘나 문법의 층위에서 발생하는 오류 양상만을 주석한다.¹⁵⁾

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

2.2 오류 층위

○ 교정 어절에 대한 형태 주석에 기대어 주석한다. ‘발음’은 구어 자료에 한하여 주석한다.

	오류 유형		주석 표지
오류 층위	발음	음소	PP
		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

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

	오류 유형		주석 표지
		사동	SC
		피동	SP
		부정	SN
		어순	WO
	담화	지시	DR
		접속	DC
		담화표지	DM
		구어/문어 오류	DS

II. 오류 판정 및 수정 지침

1. 기본 원칙

1) 오류의 식별

- 오류의 식별은 오류 여부를 식별하는 것으로부터 시작된다. 교정 어절을 만들거나 교정 어절(때로는 어절을 넘는 단위)을 만들 수 있는 가능성이 있는 경우를 오류로 본다.
- 오류의 판단은 문법성을 기준으로 삼는다. 문법성이란 의미적으로나 형태적으로 완성된 형식을 갖추지 못하고 한국어의 문법 체계에 맞지 않는 비문법적 문장을 생성하는 경우를 말한다. 즉, 문법성을 기준으로 어문 규범에 어긋나며, 용인하기 어려운 일탈은 모두 오류 판정과 주석의 대상으로 삼는다.

<예> 우리는 술을 마시고 싶으면 ‘바프라이’(BARFLY)(√ ‘바이프라이’라고 하는) 술집에 가요.
 ☞ 초급 학습자가 생성한 문장으로 ‘라고 하는’을 포함한 문장이 초급보다 높은 수준이지만 정확한 문장 생성에 실패

하였으므로 오류로 주석한다.

- 외국어로 표기된 것은 오류로 본다.

<예> 그리고 제 new(√새로운) 친구들은 많이 만나고 싶습니다.
☞ 'new'라고 영어를 그대로 표기한 것은 한국어와 외국어의
대치 오류로 주석한다.

- 오류의 판단에는 용인 가능성도 고려될 수 있는데, 이는 오류 주석자에 따라 달라질 수 있으므로 복수의 주석자가 지침을 통해 합의하여 판정한다. 용인 가능성이라는 기준은 '엄격하게' 적용하여 '일관되게' 처리하도록 한다.
- 어휘 혹은 문법 오류로 동시에 판정할 수 있는 경우, 기능어 중심으로 문법 오류를 우선시하여 처리한다.
- 오류의 판단은 문장 단위에서 이루어진다. 오류 판정 시 문제가 될 때에는 앞뒤 문장까지는 살펴볼 수 있지만, 주석의 일관성을 위해 담화 단위로 보지 않고 기본적으로 문장 단위에서만 처리하도록 한다. 단, 오류 층위에서 담화 오류에 해당하는 지시(DR), 접속(DC)의 경우, 선행문과 후행문과의 의미적 연결을 고려해야 오류 판단이 가능하기 때문에 앞, 뒤 문장을 고려하여 오류를 판단한다.
- 구어 자료의 경우, 문장으로 파악하지 않고 억양 단위로 끊어서 각 단위를 기준으로 오류를 식별하고 판정한다.

<예> 무슨 파티하면
우리 학생들이.
열심히 공부한=
연세대학교 열심히 공부해서
조금 피곤한,
=것이에요.
☞ 이 경우 억양 단위로 끊어서 보면 크게 문제가 되지 않지만, 문장 단위로 보면 여러 가지 층위에서 오류 처리가 가능하며 일관된 기준에 의한 처리가 어렵다. 구어 자료는

문장 단위가 아닌 억양 단위를 기준으로 하여 오류를 식별하고 판정한다.

2) 오류와 실수의 구분

- 오류와 실수는 구분하지 않는다. 즉, 실수인지 오류인지의 여부와 상관없이 규범상의 일탈은 모두 오류로 간주한다. 이는 연구자의 판단 영역으로 자료만으로 학습자의 의도를 파악할 수 없기 때문에 주석 작업자의 자의적인 해석을 막기 위한 것이다.
- 구어 자료에서 발화 중에 학습자의 자기 수정이 일어난 경우, 수정하기 이전의 일탈은 오류로 간주하지 않는다. 학습자 스스로 오류임을 인지하고 수정을 하였으므로 수정 후 발화에 초점을 두고 오류 여부를 판정한다. 수정 후 발화에서도 오류가 나타난 경우는 이전의 일탈도 모두 오류로 주석하고, 수정 후 발화가 제대로 되었을 경우에는 이전의 일탈은 오류로 주석하지는 않으나 교정어절은 써주도록 한다.

3) 오류의 교정(교정 어절 원칙)

- 오류의 교정은 오류로 식별된 부분을 올바르게 고치는 것을 말한다. 따라서 오류로 식별된 것은 교정의 대상이 된다.
- 오류의 교정은 학습자의 표현 의도를 고려하여 최소한의 교정을 원칙으로 한다. 즉, 학습자의 표현 의도나 의미를 자의적으로 유추하여 교정 어절을 생성하지 않으며, 학습자가 산출한 형태를 가능한 한 훼손하지 않고 최대한 원문을 유지할 수 있는 형태로 수정한다.

<예> 현대(✓ 현재) 세계적으로 환경 문제가 대두되고 있다.
 ☞ ‘현대’를 ‘현재’로 수정

- 오류로 판정된 문장을 교정할 때 그것을 문법적으로 완전한 문장으로 바꿀 것인지 용인가능한 수준의 문장으로 바꿀 것인지와 관련하여서는 학습자의 표현 의도를 반영하여 용인 가능한 수준으로 최소한의 교정을 하며, 한국어 모어 화자의 보편적인 언어 사용 방식에 따라 교정한다.

- 또한 오류의 교정은 정보가 소실되지 않는 차원에서 최소한의 교정을 원칙으로 한다. 즉, 앞부분의 오류를 수정하는 것으로 인해 뒷부분에까지 영향을 미쳐 뒷부분까지 교정이 필요한 경우, 학습자의 의도에서 벗어날 수 있으며 주석자의 자의적인 해석이 지나치게 반영될 수 있기 때문에 앞부분에서 최소한의 교정만 하며, 전면 교정이 필요할 때에는 분석불가능(IMP)으로 주석한다.
- 오류 영역에서 교정 어절로 인해 조사나 어미가 바뀌는 경우, 교정 어절의 영향을 받아 바뀐 조사와 어미는 오류로 처리하지 않는다.
- 기본적으로 맥락을 살펴 되도록 내용어보다 기능어를 우선 교정하는 것을 원칙으로 하므로, 뒤의 용언을 바꾸지 않는 방향에서 조사 오류로 처리하는 원칙이 우선이지만 용언을 반드시 교정해야 할 경우, 용언이 대치되면서 용언 때문에 조사가 바뀌는 경우에는 용언 대치 오류로만 처리하고 조사 오류로는 처리하지 않는다. 단, 사동과 피동 오류에 한하여 사동사와 피동사로 바뀌면서 조사가 바뀔 때에는 사동/피동과 관련한 오류라는 것을 표시해주는 차원에서 조사도 오류로 주석하며, 오류 층위에 사동과 피동을 주석한다.

<예> 아파트가 평형이(√/평수가) 많으면(√/넓으면, √/크면) 친구들을 부를 수 있다.

☞ ‘평형’을 ‘평수’로 교정함에 따라 조사 ‘이’가 ‘가’로 바뀌게 되었으므로 조사는 오류로 주석하지 않는다. 이때에는 ‘평형’과 ‘많다’만 대치 오류로 처리하고, 조사 ‘이’는 오류로 주석하지 않는다.

나라가(√/나라를) 발전하다(√/발전시키다)

☞ 사동 ‘시키다’로 교정해야 할 경우, 조사까지 교정해야 한다. 이 경우에는 조사 ‘가’와 ‘를’도 대치로 처리한다. 따라서 [오류 위치-주격조사], [오류 양상-대치], [오류 층위-사동]과 [오류 위치-동사파생접미사], [오류 양상-대치], [오류 층위-사동]오류로 주석한다.

4) 오류 판정의 대상

- 오류 판정은 오류로 식별되어 교정된 부분이 어떤 범주의 오류인지를 판정하는 것을 말한다.
- 오류 판정은 오류에 대한 주석이므로 교정 어절이 아닌 오류 어절(원어절)을 기준으로 한다. 즉, 학습자가 산출한 언어 형태와 오류 발생 위치를 기준으로 오류를 판정한다.¹⁶⁾

<예> 가끔 술을 마시지 않아서(√않을 때는) 영화를 보러 영화극장에 갈 거예요.
☞ 이 경우 오류 어절인 ‘않아서’의 ‘아서’를 기준으로 하여 [어미 오류]로 판정한다.
호주는 어디든지(√어디인지? 어디에 있는지?) 알아요?
☞ 오류 어절을 기준으로 하면 ‘든지’의 오류로 보아 [조사 오류]로 주석한다.

- 오류 주석은 형태소 단위를 기본으로 한다. 따라서 구 단위 이상의 어휘나 표현은 구성 요소를 형태로 나누어 분석한다. 다만, 교정할 위치가 어절이나 구 단위 차원에서 처리하여야만 용이하거나 구 단위로만 교정이 가능한 한 경우는 구 단위로 처리할 수 있다. 또한 표현 문형의 경우는 <국립국어원 2> 목록을 확인하여 해당 표현이 목록에 있을 경우는 표현 문형도 함께 주석한다. 즉, 기본적으로 형태 단위로 분석하여 주석하되, 구 단위와 표현 문형도 중복 주석할 수 있다. (☞ 3. 범주별 세부 오류 유형의 처리 예시-2) 오류 위치-(4) 표현 문형(PE) 참고)

<예> 내일은 비가 온 것(√올 것) 같아요.
☞ ‘(으)ㄴ 것’, ‘(으)ㄹ 것’은 표현 문형 목록에 포함되어 있기 때문에 이 경우에는 오류가 발생한 ‘온 것’을 하나의 덩어리 표현으로 처리하는 동시에 구성 요소인 관형사형 전성어미로도 분석하여 오류가 나타난 위치를 중복 주석한다.

16) 이후 웹사이트에서는 교정어절을 중심으로 한 검색도 가능하게 하여, 미사용으로 인한 오류를 파악할 수 있게 한다.

즉, [오류 위치-표현 문형, 관형사형 전성어미], [오류 양상-대치] 오류로 주석한다.

은행에 저축한(✓저축할) 겨우에는(✓경우에는) 얼마정도 이익을 얻을지 미리 알아서 더 편할 것 같아요.

☞ ‘-(으)ㄴ 경우에는’을 한 덩어리로 처리할 수도 있으나, 본 연구에서 참고 목록으로 삼고 있는 <국립국어원 2> 목록에 표현 문형으로 제시되어 있지 않기 때문에 [오류 위치-관형사형 전성어미], [오류 양상-대치]와 [오류 위치-명사], [오류 양상-오형태] 오류로 각각 처리한다.

5) 기타

- 문장부호 사용에 관한 오류는 주석 대상에서 제외한다. 즉, 학습자가 생략 또는 누락한 문장부호가 있다고 하더라도 오류로 판정하지 않는다.

<예> 예) 광고가 주는 정보가 모두 진실이 아니라는 예방적인 생각도 필요해요(✓ 온점 누락)

2. 오류의 범주

- 본 연구에서는 오류의 범주를 오류 위치와 오류 양상, 오류 층위 세 가지로 설정한다. 그리고 이를 다시 기본 주석과 확장 주석으로 이원화 하여 오류를 주석한다.
- 기본 주석은 오류 위치가 해당되며, 모든 오류에 대해 1:1로 주석하는 필수 주석이다(분석이 불가능한 오류는 분석 불가능[IMP] 표지로 주석하며, 표현 문형의 경우는 각 형태소와 표현 문형[PE] 표지가 중복 주석된다). 확장 주석은 오류 양상과 오류 층위가 해당되며, 이는 관련 오류가 있는 경우에만 주석하는 수의적 주석이다. 확장 주석의 경우, 한 형태에 2개 이상의 오류가 나타나면 중복 주석이 가능하다.

1) 분석 여부

- ‘분석 여부’의 판단은 오류로 식별된 형태에 대해 교정이 가능한지 여부와 특정 범주의 오류로 판정 가능한지를 파악하는 것을 말한다. 따라서 부적절한 표현이 연속되거나 문장 구조의 이상으로 학습자의 표현 의도를 파악하기 어려운 경우 ‘분석 불가능[IMP]’으로 판정할 수 있다.

영역	주석 표지	포함 범위	예시
분석 불가능	IMP	문맥 내에서 해석이 불가능한 경우	한국여자 좋춤하고(√/좋고? 조용하고?, IMP) 예쁘기 대문에(때문에, MIF) 결혼(√/결혼, MIF)하고 싫어요.

2) 오류 위치

- 오류가 발생한 위치 표지로서 [오류 위치]를 주석한다. 오류 위치는 오류가 일어난 부분, 즉 오류가 발생한 위치의 품사(형태소)에 대해 주석한다.¹⁷⁾
- 오류 위치는 기본적으로 형태 주석에 따라 처리한다. 형태 주석에서 <표준국어대사전>에 근거하여 형태소 분석을 하기 때문에, 오류 주석은 이에 입각하여 오류 위치를 주석한다. 다음은 오류로 식별된 부분의 품사 위치를 표시한 주석 표지이다.

위치		주석 표지	포함 범위	예시
실질 어휘	고유 명사	CNNP	고유 명사 어휘의 사용에서 나타난 오류	그리고 저는 독요(√/도쿄, CNNP, MIF)에 가고 싶어요.
	일반 명사	CNNG	일반 명사 어휘의 사용에서 나	애기와(√/아기와, CNNG,

17) 단, 누락 오류의 경우에는 원 어절이 없으므로 교정 어절에 따라 주석한다.

위치	주석 표지	포함 범위	예시
		타난 오류	MIF) 노인들한테 건강이 나빠졌다.
의존 명사	CNNB	의존 명사 어휘의 사용에서 나타난 오류	그럼데 아쉬운 건(√것, CNNB, MIF)도 있다.
대명사	CNP	대명사 어휘의 사용에서 나타난 오류	내(√우리, CNP, REP) 아버지가 남편하고 친하게 되면 좋겠다.
수사	CNR	수사 어휘의 사용에서 나타난 오류	이 셋(√세, CNR, MIF) 가지 단어의 뜻에 따라 이 외모지상주의라는 말을 충분히 이해할 수 있다.
동사	CVV	동사 어휘의 사용에서 나타난 오류	그로 인해 평소 일상생활에서 말할 수 없는 말, 욕하는 말, 비우는(√비웃는, CVV, MIF) 말 등 흔히 볼 수 있다.
형용사	CVA	형용사 어휘의 사용에서 나타난 오류	불고기 먹기 때문에 기분이 기쁩니다(√좋습니다, CVA, REP).
보조 용언	CVX	보조용언 어휘의 사용에서 나타난 오류	다른 사람에게 아픈다운 모습을 보여 싶기(√주기, CVX, REP) 위하여 노력하세요.
지정사	CVC	지정사 어휘의 사용에서 나타난 오류	그러니까 저는 외모지상주의가 위험이라고(√위험하다고, CVC, REP)생각한다.
관형사	CMM	관형사 어휘의 사용에서 나타난 오류	그렇게 되면 어느(√어떤, CMM, REP) 사람은

위치	주석 표지	포함 범위	예시
		오류	돈이나 개인 정보를 잃어버릴 수도 있다.
일반 부사	CMAG	일반부사 어휘의 사용에서 나타난 오류	내 남편은 꼭(✓정말, CMAG, REP) 멋있게 생겼다.
접속 부사	CMAJ	접속부사 어휘의 사용에서 나타난 오류	그런데(✓그런데, CMAJ, MIF) 제 가격을 정말 보고 싶습니다.
감탄사	CIC	감탄사 어휘의 사용에서 나타난 오류	‘네(✓네, CIC, MIF), 알겠습니다’라고 대답했다.
접두사	CXPN	접두사 어휘의 사용에서 나타난 오류	최초 임금을 실행하면 처소득층(✓저소득층, CXPN, MIF) 사람의 살기가 보증할 수 있다.
명사파생 접미사	CXSN	명사 파생 접미사 어휘의 사용에서 나타난 오류	두 번째(✓번제, CXSN, MIF)에 간 곳이는 경주였습니다.
동사파생 접미사	CXSV	동사 파생 접미사 어휘의 사용에서 나타난 오류	그 꿈을 위해서 매일 운동해다(✓운동한다, CXSV, MIF).
형용사파생 접미사	CXSA	형용사 파생 접미사 어휘의 사용에서 나타난 오류	이러한 사회에서 자신이 하고 싶은 직업을 할 수 있으면 너무 행복안(✓행복한, CXSA, MIF) 것이다.
어근	CXR	어근 어휘의 사용에서 나타난 오류	시원(✓시원한, CXR, REP) 옷을 준비하세요.

위치		주석 표지	포함 범위	예시
기능 어휘	주격 조사	FNP	주격조사의 사용 에서 나타난 오 류	그리고 여행이(√을, FNP, REP) 너무 좋아합니다. 고시원에서 많이 학생(√이, FNP, OM) 살았다.
	관형격 조사	FGP	관형격조사의 사 용에서 나타난 오류	‘론딩리’의(√를, FGP, REP) 소개합니다.
	목적격 조사	FOP	목적격조사의 사 용에서 나타난 오류	그래서 저는 한국 사람하고 다른 외국 사람을(√과, FOP, MIF/REP) 교류하고 싶습니다.
	부사격 조사	FAP	부사격조사의 사 용에서 나타난 오류	차 안에(√에서, FAP, REP) 잤어요.
	접속 조사	FJC	접속조사의 사용 에서 나타난 오 류	한국어 문법와(√과, FJC, MIF) 중국어 문법이 비슷하지 않았다.
	호격 조사	FVP	호격조사의 사용 에서 나타난 오 류	친구아(√친구야, FVP MIF), 같이 가자.
	인용격 조사	FQP	인용격조사의 사 용에서 나타난 오류	내가 감사하다고 말한다는 게 ‘고맙다’다고(√라고, FQP, REP) 했다.
	보조사	FXP	보조사의 사용에 서 나타난 오류	론딩리는 맛있는 음식은(√이, FXP, REP) 많습니다.
	연결 어미	FED	연결어미의 사용 에서 나타난 오 류	그리고 친구들과 같이 노래방 가고(√가서, FED, REP) 노래를 부르고 싶습니다.
	종결	FFE	종결어미의 사용	특히 아랫목에 정말

위치		주석 표지	포함 범위	예시
	어미		에서 나타난 오류	따뜻한다(√ 따뜻하다, FFE, MIF).
	선어말 어미	FPE	선어말어미의 사용에서 나타난 오류	내일부터 수업이 시작됐어요(√ 시작돼요, FPE, REP).
	명사형 전성 어미	FNE	명사형 전성어미의 사용에서 나타난 오류	우리 계획은 저녁을 먹기(√ 먹은, FNE, REP) 후에 우리 만든 신분증을 가지고 갈 겁니다.
	관형사형 전성어미	FAE	관형사형 전성어미의 사용에서 나타난 오류	중요하는(√ 중요한, FAE, MIF) 것은 사람마다 다르다는 것을 인정하는 것이다.
구 단위 표현		PHE	구 단위 표현 사용에서 나타난 오류	가족들이 관심하면(√ 관심을 가지면, PHE, REP) 생존이 행복해지겠다.
표현 문형		PE	보조 용언이나 여러 요소의 결합 구성으로 이루어진 표현 문형의 사용에서 나타난 오류(제시 목록 참고)	신세대는 기상세대와 가치관이 달라서 세대 차이가 생기기 마련이다(생기기 마련이다, PE, MIF).

3) 오류 양상

- 오류 양상은 표면적으로 드러난 오류의 모습으로, 누락, 첨가, 대체, 오형태 4가지로 설정한다.
- 오류 양상은 확장 주석으로 수의적 주석에 해당한다. 따라서 누락, 첨가,

대치, 오형태 오류로 보기 어려운 오류 양상은 주석하지 않는다.

양상	주석 표지	포함 범위	예시
누락	OM	완전한 문장/발화에서 나타나야 할 형태가 빠져 있는 경우	저는 여덟 시부터 여덟 시 삼십분까지 저녁(√을, FOP, OM) 먹어요.
첨가	ADD	완전한 문장/발화에서 나타나지 말아야 할 형태가 쓰인 경우나 중복된 형태를 반복해서 사용한 경우	한국말은 동경에 있었을 때, 일년간 동안(√일년 간, CNNG, ADD) 한국 YMCA에서 공부했습니다.
대치	REP	다른 의미의 어휘를 사용하거나 적절한 품사를 사용하지 못한 경우	용서를 줄(√할, CVV, REP) 수 있게
		한국어에 없는 표현이나 한국어가 아닌 다른 언어를 사용한 경우	나는 cousin(√사촌, CNNG, REP)한테 이야기했어요.
오형태	MIF	오형태 오류: 한국어에 존재하지 않는 어휘를 만들어 내거나 조사와 어미의 활용 형태가 잘못된 경우 즉, 활용 또는 곡용을 잘못하여 다른 이형태를 사용한 경우	이 시간은 별로 덤지 않고 시원해서(√시원해서, FED, MIF, MCJ) 숙제하기에 좋습니다.
		맞춤법 오류: 철자를 잘못 사용한 경우	우리는 피간했어요(√피곤했어요, CNNG, MIF).

4) 오류 층위

- 오류 층위는 오류로 식별된 부분을 언어학적 층위에 따라 나눈 오류의 범주이다. 즉, 언어학적 측면에서 어느 영역의 오류인지를 판단하는 것으로, 본 연구에서는 오류 층위를 교수자나 학습자들이 자주 활용할 일부 영역(발음, 형태, 통사, 담화)에 한정하여 주석하였다.
- 오류 층위는 오류 어절(원어절)과 교정 어절 모두를 고려하여 해당 층위에 맞게 주석한다.
- 발음 층위는 구어에서의 발음 오류를 다루는 영역이다. 발음 층위에서는 음소, 음절, 음운 규칙에서 발생하는 오류와 학습자의 원어식 발음, 변이음을 포함한 중간 발음에서 나타나는 오류를 주석한다.

층위		주석 표지	포함 범위	예시
발음	음소	PP	음소 차원에서 발생하는 오류 예) 평음, 격음, 경음의 구분	[구어] 케이키도(✓ 케이크도, CNNG, PP) 있고 생일파티 주인공이도 있어요.
	음절	PS	음절 차원에서 발생하는 오류. 음절의 발음을 정확하게 하지 못한 경우로 원래 음절보다 더 적게 혹은 많게 발음한 경우와 축약해야 하는데 축약하지 않고 발음한 경우 또는 그 반대의 경우	[구어] 우리::= 우리나라도:: 마야크(✓ 마약, CNNG, PS),... 어:: 판매::, 될 수 있..=있긴 한데::
	음운 규칙	PC	음운변동에 관한 오류로 구어에서	[구어] 한국계(✓ 한국에, CNNG,

층위		주석 표지	포함 범위	예시
			필수적 음운 규칙의 일탈 또는 음운 규칙을 적용하지 않고 절음화하여 발음한 경우 예) 연음규칙, 비음화 유음화, 구개음화, 경음화 등	PC) 가요.
	원어식 발음	PN	원어식 발음으로 발생하는 외국어 오류	[구어] 예를 들면 자기 계발, 재미, 수업 등 그래서 아래 그래프(√그래프, CNNG, PN)를 보며는,
	중 간 발음(변이음 포함)	PA	변이음을 포함한 중간 발음으로 발생하는 오류.	[구어] 전공(√전공, CNNG, PA) ('저'와 '조'의 중간발음)

- 형태 층위는 어휘 오류를 다루는 영역이다. 형태 층위에서는 합성어, 파생어 등의 조어 과정에서 발생하는 오류와 어미의 활용, 조사의 사용 등에서 나타나는 오류를 주석한다.

층위		주석 표지	포함 범위	예시
형태	단 어 형성[합성법]	MCP	단어 합성에서 나타나는 오류	해물고기(√물고기, CNNG, MCP)가 많았어요.
	단 어 형성[파생법]	MDV	단어 파생에서 나타나는 오류	작년 방학 때는 LG 전자에서 통역사로 일한 경험이 있고 한국에 대한

층위	주석 표지	포함 범위	예시
			사이트에서 번역사(✓번역가, CNNG, POS/MDV)로 일한 경험도 있기 때문에 <name>에서 일할 수 있는 자신을 가지고 있다.
굴절 [곡용]	MDC	조사 이형태 선택에서 나타나는 오류	론딩리는 지하철와(✓과, FAP, MIF, MDC) 가까워서 편리합니다.
굴절 [활용]	MCJ	용언과 어미의 활용에서 나타나는 오류	내가 10년 후에 좋하고 행복 살으면(✓살면, FED, MIF, MCJ) 좋겠다
품사	POS	동일 의미의 품사 선택에서 나타나는 오류	주말에 친구하고 같이 유명의(✓유명한, CNNG, REP, POS/MDV) 곳이 가고 싶습니다.

- 통사 층위는 문법 오류를 다루는 영역이다. 통사 층위는 높임, 시제, 사동, 피동, 부정 등의 문법 범주와 관련되어 해당 문법범주를 제대로 사용하지 못했을 경우 주석한다.

층위	주석 표지	포함 범위	예시
통사	높임	SH	조사, 선어말어미, 종결어미 등 높임 관련 문법 형태소와 높임 어휘의 오류
	시제	ST	시제를 나타내는 문법 형태소의 오류
			할머니께서 우유를 마시십니다(✓드십니다, CVV, REP, SH).
			어제부터 항상 시제를 확인하기로 한다(✓했다, FPE, REP, ST).

층위	주석 표지	포함 범위	예시
사동	SC	사동사, 사동 표현의 오류	갈릴레이는 새로 발명된 망원경을 사용하여 연구를 깊었다. (✓ 깊게 하였다, CVA, MIF/REP, SC)
피동	SP	피동사, 피동 표현의 오류	교실 문이 닫아(✓ 닫혀, CVV, REP, SP) 있었습니다.
부정	SN	부정 표현의 오류	한국에 온 후에 한 문장도 못(✓Ø, CMAG, ADD, SN) 들을 수 없었다.
어순	WO	한국어의 통사 구조에 맞지 않는 방식으로 문장이 배열된 오류	사람이 상태에 위독한(✓ 위독한 상태에, CVA, WO) 빠집니다.

- 담화 층위는 문장 단위를 넘어서 발생하는 오류를 다루는 영역이다. 담화 연구의 경우 그 범위가 넓고 어휘와 문법, 발음 영역에서 다양한 현상과 표지를 통해 나타나므로 체계화가 쉽지는 않다. 또한 구어 담화의 경우 문법정보보다는 발화 맥락 안에서 적절하고 효과적인 의미 전달에 초점이 주어지기 때문에 오류 판정 기준을 정하기도 쉽지 않다. 이러한 이유로 본 연구에서는 담화표지, 지시, 접속으로 비교적 표지가 분명하고 판정 기준이 명확한 항목만을 주석의 대상으로 포함시켰다.
- 본 연구는 문장 내에서의 오류 판단이 기본 원칙이기 때문에 문장 단위를 넘어서는 담화 오류는 최소한으로 제한하여 주석한다. 지시(DR)와 접속(DC)에 한해서 선행문과 후행문과의 의미적 연결을 고려해 오류 여부를 판단한다.

층위		주석 표지	포함 범위	예시
담화	지시	DR	부적절한 지시사의 선택으로 선행문과 후행문의 관계를 결속성 있게 나타내지 못한 경우	나는 롯데월드 아이스링크에 자주 가요. 여기(✓거기, CNP, REP, DR)에 가면 스트레스가 풀려요.
	접속	DC	선행문과 후행문의 의미 관계를 나타내는 데에 부적절한 접속 부사 및 접속 표지를 사용한 경우	나는 이런 남자를 만나면 경혼하고 싶습니다. 그래서(✓그러면, CMAJ, REP, DC) 기분이 좋을 거예요.
	담화 표지	DM	부적절한 담화 표지를 선택하거나 누락한 경우	우리 하숙집에서 현대백화점까지 그냥(CMAG, ADD, DM) 10분쯤 걸렸어요..
	구어/문어	DS	구어체/문어체, 격식체/비격식체의 혼용으로 인한 오류	근데(✓그런데, CMAJ, REP, DS) 어떤 사람을 평가할 때 외모만 보면 그거도 안 된다..

3. 범주별 세부 오류 유형의 처리 예시

1) 분석 여부

- ‘분석 여부’는 오류를 특정 범주의 오류로 판정 가능한지를 파악하는 것을 말한다. 부적절한 표현이 연속되거나 문장 구조의 이상으로 학습자의 표현 의도를 파악하기 어려운 경우 ‘분석 불가능(IMP)’로 판정할 수 있다.

<예> 나도 한번도 많고, 기 사람하고, 밥그릇, 노래했어요.(√IMP)
난 졸업만 뜬다면 드디오 내가 기다리는 시간이라고 생각하
고 귀가 아주 밝다(√IMP).

- 분석 여부는 기본 값이 ‘분석 가능’으로 설정되어 있으므로, 오류임에 분명 하지만 교정어절을 주기 어려워 오류의 판정이 불가능한 경우에만 주석을 한다.

2) 오류 위치

(1) 실절어휘

① 고유명사(CNNP)

- 고유명사의 형태, 의미, 사용 오류를 말한다.

<예> 중구(√중국) 요리를 맛있었습니다.
하지만 독요(√도쿄) 쇼핑은 조금 비싸요.

② 일반명사(CNNG)

- 일반명사의 형태, 의미, 사용 오류를 말한다.

<예> 애기와(√아기와) 노인들한테 건강이 나빠졌다.
현대(√현재) 세계적으로 환경문제가 대두되고 있다.

③ 의존명사(CNNB)

- 의존명사의 형태, 의미, 사용 오류를 말한다.

<예> 희망 10명(√년) 후에 자기 가 수 있다.
내가 한국에 온 지 7개월(√개월)이 되었다.

④ 대명사(CNP)

- 대명사의 형태, 의미, 사용 오류를 말한다.

<예> 내(√우리) 아버지가 남편하고 친하게 되면 좋겠다.
저(√나)는 유학생으로 온 외국인이다.

⑤ 수사(CNR)

- 수사의 형태, 의미, 사용 오류를 말한다.

<예> 오후 일곱(√일곱) 시에 홍콩 친구하고 저녁 식사를 했어요.
요리를 배우기가 열(√십) 년 전이 시작했습니다.

⑥ 동사(CVV)

- 동사의 형태, 의미, 사용 오류를 말한다.

<예> 그런데, 기숙사에서 술을 미실(√마시는) 것 안 된다.
그것을 어쩔 수 없는 것이고 누군가가 그 변화를 세우는(√
멈추는) 것이 못한다.

⑦ 형용사(CVA)

- 형용사의 형태, 의미, 사용 오류를 말한다.

<예> 불고기 먹기 때문에 기분이 기쁩니다(√좋습니다).
인심 약박한(√야박한) 시대 속에서 법이 사람의 부합리적인
행동을 제약할 수 있는 효과적인 방법이다.

⑧ 보조용언(CVX)

- 보조용언의 형태, 의미, 사용 오류를 말한다.

<예> 태하교에 가고 싶습니다(√ 싶습니다).
 인터넷 쇼핑을 하고 싶으면 조심한다(√ 조심해야 한다).

- 오류 주석 시, 보조용언과 결합된 구성이 표현 문형 목록에 있을 경우에는 표현 문형(PE)위치로도 중복 주석한다.

<예> 세대 차이를 극복하기 위해서 신세대와 가상세대는 자주 이야기를 해 뉘야 된다(√ 된다)
 ☞ ‘뉘다’는 보조용언 ‘된다’를 잘못 사용한 것이기 때문에 [오류 위치-보조용언]으로 처리하는 동시에, ‘-어/아야 되다’가 표현 문형의 목록에 있기 때문에 표현 문형 위치도 중복 주석한다. 즉, [오류 위치-표현 문형, 보조용언], [오류 양상-오형태] 오류로 처리한다.

⑨ 지정사(CVC)

- 지정사의 형태, 의미, 사용 오류를 말한다.

<예> 학생예요(√ 학생이예요)
 ☞ 지정사 ‘이예요’와 ‘예요’를 잘못 사용한 것이기 때문에 지정사 오류로 처리한다.

- 지정사와 연결어미/종결어미가 결합할 때, 지정사를 누락시키거나 첨가했을 경우는 지정사 누락, 첨가 오류로 처리한다. 단, 축약을 잘못된 경우는 오철자 오류로 처리한다.

<예> 다른 사람들에게 도와주고 싶기 때문에 한국어를 열심히 공부할 것다(√ 것이다).
 ☞ 지정사 ‘이다’가 생략되었으므로 [오류 위치-지정사], [오류 양상-누락]으로 처리한다.
 남자이예요(√ 남자예요)

☞ 받침이 없는 명사 뒤에서 ‘이에요’를 ‘예요’로 줄여서 쓰는 것이 일반적이거나, 필수적인 표준 규범은 아니므로 오류로 처리하지 않는다.

학생예요(√ 학생이에요)

☞ 받침 유무에 따라 ‘이에요’와 ‘예요’를 선택하여 사용하나, 이 경우 학습자가 ‘이에요’와 ‘예요’를 잘못 선택해서 사용한 것인지, 지정사 ‘이’를 누락한 채 종결어미를 잘못 쓴 것인지 판단이 어렵다. 본 연구에서는 지정사의 경우, 써야할 자리에 쓰지 않은 경우는 누락으로 보고, ‘예요’와 ‘예요’는 종결어미의 오철자 오류로 처리한다. [오류 위치-지정사], [오류 양상-누락], [오류 위치-종결어미], [오류 양상-오형태] 오류로 주석한다.

학생예요(√ 학생이에요)

☞ [오류 위치-지정사], [오류 양상-누락]

학생이에요(√ 학생이에요)

☞ [오류 위치-종결어미], [오류 양상-오형태]

아니예요(√ 아니예요)

☞ [오류 위치-종결어미], [오류 양상-오형태]

학생이여서(√ 학생이어서)

☞ [오류 위치-연결어미], [오류 양상-오형태]

학생이었어요(√ 학생이었어요)

☞ [오류 위치-선어말어미], [오류 양상-오형태]

공부를 할 거예요(√ 거예요)

☞ [오류 위치-지정사], [오류 양상-누락]

☞ 지정사와 관련된 오류에서 지정사를 쓰고 어미와 축약하지 않거나 잘못 축약을 시킨 경우는 오철자 오류로 처리한다. 오류 위치는 오류가 발생한 위치에 따라 주석한다.

- 문어에서 지정사를 축약해서 사용한 경우에는 오류로 보기 어려운 측면이 있으나 모어 화자의 보편적인 언어 사용 방식에 있어서 어색한 것으로 보고 지정사 오류로 처리한다. 이때에는 오류 양상은 주석하지 않고, [오류 위치-지정사], [오류 층위-문어/구어] 오류로 처리한다.

<예> 여기는 우리 학곤데(√학교인데) 정말 아름답다.
 ☞ 문어(격식체)에서 축약형으로 사용하는 것은 어색하기 때문에 [문어/구어] 오류로 처리한다. 단, 이때에는 오류 양상은 주석하지 않는다.

⑩ 관형사(CMM)

- 관형사의 형태, 의미, 사용 오류를 말한다.

<예> 그렇게 되면 어느(√어떤) 사람은 돈이나 개인 정보를 잃어버릴 수도 있다.
 두(√이) 년 한국에 있을 겁니다.

⑪ 일반부사(CMAG)

- 부사의 형태, 의미, 사용 오류를 말한다.

<예> 내 남편은 꼭(√정말) 멋있게 생겼다.
 어히려(√오히려) 남정보다 여성의 힘이 더 강하는 경우도 있는 정도다.

⑫ 접속부사(CMAJ)

- 접속부사의 형태, 의미, 사용 오류를 말한다.

<예> 그래서(√그래서) 피자하고 맥주를 먹고 많이 얘기했다.
 그러니까(√그래서) 안목이 높아지거니와 다양한 문화의 향유하고 새로운 것들을 깨닫기도 한다.

⑬ 감탄사(CIC)

- 감탄사의 형태, 의미, 사용 오류를 말한다.

<예> 아침 6시에 일어나서 하는 출근 준비, 이제 안녕(√안녕)~
응(√네). 선생님.

⑭ 접두사(CXPN)

- 접두사의 형태, 의미, 사용 오류를 말한다.

<예> 인심 약박한 시대 속에서 법이 사람의 부합리적인(√불합리적인) 행동을 제약할 수 있는 효과적인 방법이다.
산세대(√신세대) 사람들이 부모님 입장에 많이 생학하고 부모님도 신세대 입장 색학하면 세대 차이를 줄일 수 있다.

⑮ 명사파생접미사(CXSN)

- 명사파생접미사의 형태, 의미, 사용 오류를 말한다.

<예> 저의 첫 번째(√번째) 고민은 어떻게 시간을 잘 지킨 좋은 습관은 될 수 있는 것이다.
여성들의 사회 진출에 따라서 이혼률(√이혼율)이 높아진 것이 큰 원인이라고 할 수 있다.

⑯ 동사파생접미사(CXSV)

- 동사파생접미사의 형태, 의미, 사용 오류를 말한다.

<예> 제 한국어를 좋아합니다(√좋아합니다).
이에 따라서 노동사의 인권이 보장하게(√보장되게) 되어 안정한 생활을 할 수 있게 되었다.

⑰ 형용사파생접미사(CXSA)

- 형용사파생접미사의 형태, 의미, 사용 오류를 말한다.

<예> 고독스러운(√고독한) 중학생

이러한 사회에서 자신이 하고 싶은 직업을 할 수 있으면 너무 행복안(√행복한) 것이다.

⑱ 어근(CXR)

- 어근의 형태, 의미, 사용 오류를 말한다.

<예> 주말에 날씨가 따뽕(√따뜻)하니까 산을 갔어요.
밤에도 시원(√시원한) 옷을 준비하세요.

(2) 기능어휘

① 주격조사(FNP)

- 주격조사의 형태, 의미, 사용 오류를 말한다.

<예> 책가(√책이) 재미있어요.
그리고 우리 친구와 만나고 같이 시장에 고기와 사과와 오렌지가(√오렌지를) 샀어요.

- ‘나는’과 ‘내가’가 상호 교정 어절이 될 때에는 조사 오류로 한 번만 처리한다. 즉, 주격조사나 보조사의 대치로 인하여 대명사의 형태가 바뀌는 경우에는 오류로 처리하지 않고 교정 어절만 써준다.

<예> 그래서 제가(√나는) 지금 열심히 공부하고 있다.
☞ [오류 위치-주격조사], [오류 양상-대치] 오류로 처리한다.

저는(√제가) 공부할 때도 일할 때도 늘 새로운 아이디어를 가지고 있는 것을 보면 친구와 동료는 저를 창의적이라고 많이 하였습니다.
☞ [오류 위치-보조사], [오류 양상-대치] 오류로 처리한다.

② 관형격조사(FGP)

- 관형격조사의 형태, 의미, 사용 오류를 말한다.

<예> 10년 후의(√에) 아버지 같은 성공한 사람이 되고 싶다.
기숙사의(√기숙사에) 규칙을 있다.

③ 목적격조사(FOP)

- 목적격조사의 형태, 의미, 사용 오류를 말한다.

<예> 집에서 포도를(√포도를) 먹었습니다.
미래에 나를 사랑하는 남편하고 귀여운 아기를(√아기가) 있으면 좋겠다.

④ 부사격조사(FAP)

- 부사격조사의 형태, 의미, 사용 오류를 말한다.

<예> 8급에(√의) 학생들은 쉬는 시간에 학교 근처 area에 가지 안됐다.
미국에(√에서) 영어 제일 중요하다.

⑤ 접속조사(FJC)

- 접속조사의 형태, 의미, 사용 오류를 말한다.

<예> 미국고(√과) 일본이 두 나라에 가고 싶다.
외모과(√와) 노력이 다 중요하다.

⑥ 호격조사(FVP)

- 호격조사의 형태, 의미, 사용 오류를 말한다.

<예> 친구아(√친구야), 어 내가 영화를 보고 싶은데.
철수아(√철수야), 어디 가니.

⑦ 인용격조사(FQP)

- 인용격조사의 형태, 의미, 사용 오류를 말한다.

<예> 선생님이 내가 읽은 책이 봐서 나한테 "수험 후에 사무실에 와요"이라고(√라고) 말했다.
나는 원래 "생선을 먹었어요"(√라고) 말했어야 했는데 김장 해서 "선생을 먹었어요"라고 말했다.

⑧ 보격조사(FCP)

- 보격조사의 형태, 의미, 사용 오류를 말한다.

<예> 10년 후에 30살(√이) 될 것이다.
이것은 제일 큰 문제이(√문제가) 되는 이유가 무엇일까?

⑨ 보조사(FXP)

- 보조사의 형태, 의미, 사용 오류를 말한다.

<예> 네 번째 부모님은 내가 좋은 미래는(√미래를) 기대하고 있다.
나는(√내가) 고등학교 때 우리 엄마가 고등학교 교장이었다.

⑩ 연결어미(FED)

- 연결어미의 형태, 의미, 사용과 관련된 오류를 말한다.

<예> 그리고 우리 친구와 만나고(√만나서) 같이 시장에 고기와 사과와 오렌지가 샀어요.
나는 시계를 보면(√보고) 잠깐 놀랐다.

- 용언의 받침 유무에 따라 어미의 이형태 선택이 달라지는 경우는 연결어미의 활용 오류로 처리한다.

<예> 이렇게 살으면(√살면) 정말 행복할 수 있다.
그리고 학국 음식을 먹려고(√먹으려고) 해요.

⑪ 종결어미(FFE)

- 종결어미의 형태, 의미, 사용과 관련된 오류를 말한다.
- 종결어미를 활용하지 않고 기본형을 사용한 경우(-다)는 [오류 양상-오형태], [오류 층위-활용] 오류로 주석한다.
- 종결어미 이형태 활용 오류는 [오류 양상-오형태], [오류 층위-활용] 오류로 주석한다.

<예> 그때부터 시간을 지키다(✓지킨다).
그래서 일본에서 웃어른은 노약자석에서 꼭 앉는다(✓앉는다).

⑫ 선어말어미(FPE)

- 선어말어미의 형태, 의미, 사용과 관련된 오류를 말한다.

<예> 인종 차별 때문에 많은 사람들이 죽었으니까 앞으로 인종 차별이 없었으면 좋게다(✓좋겠다).
그래서 너무 배가 고파다(✓고팠다).

⑬ 명사형 전성어미(FNE)

- 명사형 전성어미의 형태, 의미, 사용과 관련된 오류를 말한다.

<예> 과학자들의 의식에 의하면 아이들은 어른들보다 배우기(✓배우는) 능력이 6배 뛰어나다고 한다.
언어를 배울 때는 언어만 말고 그 나라의 문화도 공부하기(✓공부하는 것) 중요하다.

⑭ 관형사형 전성어미(FAE)

- 관형사형 전성어미의 형태, 의미, 사용과 관련된 오류를 말한다.
- 관형사형 전성어미는 시제와도 관련되므로 시제와 관련한 오류의 경우에는 오류 층위에서 ‘시제(ST)’ 오류도 함께 주석한다.

<예> 과식이나 하지 말고 여러 가지 음식을 골고루 먹을(✓먹는) 것이 중요해요.

☞ 관형사형 전성어미 대치 오류로 처리한다.
 한국에 온(√오는) 비행기에서 친구를 만났어요.
 ☞ 관형사형 전성어미 대치 오류로 처리하며, 시제와 관련된
 어 오류 층위에 시제(ST)도 주석한다.

(3) 구 단위 표현(PHE)

- 구 단위 표현은 어절 단위로 이루어진 표현을 잘못 사용한 경우를 말한다. 교정 어절 주석은 구 단위 표현 단위를 묶어서 처리한다.

<예> 남자의 아내가 더 이상 돈 없는 힘든 인생을 살고 싶지 않아
 오랜 고민과 망설임을 한 나머지(√ 끝애) 남편과 5살 어린 아
 이를 두고 더 좋은 인생을 찾으려 다른 도시로 이사를 간다.

- 구 단위 표현은 오류의 교정이 어절을 넘어서는 구 단위로만 처리해야 할 때 주석한다. 따라서 형태소 차원에서 교정이 가능한 경우는 형태소 단위를 오류 위치로 주석한다.

<예> 이상한 날씨로 악화가 나타났다(√악화가 되었다).
 ☞ ‘악화가 나타났다’라는 구 단위 오류로 처리할 것인지, ‘나
 타나다’ 동사의 대치 오류로만 처리할 것인지 문제가 될 수
 있다. 이 경우 가능한 분리하여 ‘나타났다’를 ‘되다’로 교정
 하여 [오류 위치-동사], [오류 양상-대치]로 처리한다.

- 본 연구에서는 ‘연어 오류’를 별도로 설정하지 않았기 때문에 구 단위 표현에 연어 오류가 있을 경우, 별도 처리하지 않고 해당 품사의 대치로 처리한다.

<예> 태도를 키워야 한다 (√ 길러야 한다)
 ☞ ‘태도를 키워야 한다’를 ‘태도를 길러야 한다’로 교정할 때,
 이를 연어 오류로 볼 수 있다. 그러나 연어의 정의가 판단

자마다 다를 수 있다는 문제가 있다. 따라서 연어 오류를 주석하기 위해서는 연어 목록이 선행되어야 하므로 본 연구에서는 연어 오류를 별도로 주석 하지 않고 동사 대치 오류로 처리한다.

이사를 옮기다(✓ 이사를 가다)

☞ ‘이사를 옮기다’의 경우, ‘짐을 옮기다’, ‘이사를 가다’ 2가지로 교정이 가능하다. 본 연구에서는 동사 교정을 우선으로 하여 [오류 위치-동사], [오류 양상-대치]로 주석한다.

(4) 표현 문형(PE)

- 보조 용언 구성, 여러 요소의 결합 구성으로 이루어진 표현 문형을 잘못 사용한 경우를 말한다.
- 표현 문형의 목록은 관점에 따라 상이할 수 있으므로, 외국인을 위한 <한국어 문법 사전>(국립국어원, 2005)에 표현으로 제시된 항목 중 두 개 이상의 요소로 이루어진 결합 구성에 한정하여 처리한다.¹⁸⁾
- 오류 주석은 형태 단위를 기본으로 하기 때문에 표현 문형 오류의 경우, 형태 단위로도 분석하여 오류 위치를 주석하는 동시에 표현 문형 오류로도 중복 주석한다.

<예> 왜냐하면 환경오염이 심해지면 건강이 나빠지기 십상이다(✓ 십상이기 때문이다).

☞ 앞에서 ‘왜냐하면’을 사용했기 때문에 서술어에서 ‘-기 때문이다’를 호응해서 사용해야 한다. 이러한 경우, ‘-기 때문’에 해당하는 각각의 형태소인 명사형 전성어미와 의존명사의 누락 오류로 처리하는 동시에, ‘-기 때문’이 표현 문형 목록에 있기 때문에 표현 문형도 오류 위치에 중복 주석한다.

18) 표현 문형 목록 기준을 모든 표현 문형의 합집합으로 할 경우, 다양한 형태들이 표현 문형 안으로 들어오기 때문에 본 연구에서는 어느 정도 정제된 목록으로서 <한국어 문법 사전>(국립국어원, 2005)을 기준으로 정한다. 표현 문형 목록은 <부록>으로 첨부하였다.

따라서 [오류 위치-명사형 전성어미, 의존명사/표현 문형],
[오류 양상-누락]으로 주석한다.

3) 오류 양상

(1) 누락(OM)

- [정의] 누락 오류는 완전한 문장 또는 발화에서 나타나야 할 형태가 빠져 있는 경우를 말한다.
- [주석 방식] 누락 오류의 경우, 오류 위치는 교정 어절이 기준이 된다. 따라서 교정 어절을 입력하고, 누락된 품사(교정 어절)를 오류 위치로 주석한 후, 오류 양상을 누락(OM)으로 주석한다.
- [처리 기준] 누락 오류는 조사나 어미 누락을 우선적으로 주석한다.

<예> 돈이 많은 사람들(√은) 투자 할 수도 있어요.
☞ 문어에서 조사의 생략은 모어 화자의 언어생활에서도 일반적이지 않다. 따라서 조사나 어미 누락을 중심으로 누락 오류를 판단하며, 이때에는 누락된 보조사 ‘은’을 오류 위치로 주석한다. [오류 위치-보조사], [오류 양상-누락(OM)]으로 주석한다.

- 누락 오류는 필수적인 성분이 생략됐을 경우에만 처리한다. 따라서 필수적인 성분이 아닌 주의적이거나 없는 정보를 더 추가해주지 않도록 한다..

<예> 저는 여러 까지 능력서가 취득하지만 그 중에서 영어를 (√가장, OM?) 능숙하는 정도입니다.
※ ‘그 중에서 영어를 가장 능숙한 편입니다’라고 교정하여, ‘가장’이라는 부사를 첨가하고, ‘능숙한 편입니다’라고 교정할 수 있는가?
☞ 필수적인 성분이 아닌 것을 추가하여 [누락] 오류로 처리해서는 안 된다.
☞ 최소한의 교정 원칙에 따라 ‘능숙한 정도입니다’를 교정

어절로 삼는다. 즉, 수의적인 것은 [누락]으로 처리하지 않고, 필수적인 성분이 생략됐을 경우에만 [누락] 오류로 처리한다.

- 누락 오류는 하나의 유의미한 교정 어절에만 누락 오류로 주석하고, 뒤따라 오는 요소들은 누락 오류로 처리하지 않는다. 즉, 용언이나 체언(실질어휘)의 누락으로 인해서 뒤따라오는 어미나 조사(기능어휘)는 교정어절만 써주고 누락 오류로 주석하지 않도록 주의한다.

<예> 우리는 함께 (√있을) 때 좋은 기분이 왔는데요.

☞ ‘있을’이 누락되었는데, 형태 주석을 기본단위로 오류주석을 할 때, 동사 ‘있’과 관형사형 전성어미 ‘을’을 각각 누락 오류로 처리할 수 있다. 그러나 동사 뒤에 오는 관형사형 어미는 동사에 의해 따라오는 것으로 판단하여, 동사 ‘있’ 하나만을 누락 오류로 주석한다. 즉, [오류 위치-동사], [오류 양상-누락]으로 주석한다.

- [주석 예시] 누락 오류의 오류 위치별 일부 예시는 다음과 같다.
- 명사 누락은 다음과 같은 것들이 해당된다.

<예> (√환경을) 개발한 탓에 산과 나무 점점 없졌다.

☞ 문장의 필수 성분인 목적어가 누락된 명사 누락 오류로 처리한다. 이때 ‘환경’으로 인해 따라오는 ‘을’은 누락 오류로 주석하지 않고 교정어절만 써주도록 한다. [오류 위치-명사], [오류 양상-누락] 오류로 주석한다.

- 동사 누락은 다음과 같은 것들이 해당된다.

<예> 한번 (√먹어) 보야 한다.

☞ 본동사 ‘먹다’가 누락되었으므로 [오류 위치-동사], [오류 양상-누락]으로 주석한다. 뒤따라오는 연결어미는 교정어절만 써 주고 누락으로 주석하지 않는다.

- 조사 누락은 다음과 같은 것들이 해당된다.

<예> 그로 인해 평소 일상생활에서 말할 수 없는 말 욕하는 말 등 비우는 말 등(√ 을) 흔히 볼 수 있다.

☞ ‘등’을 써줬기 때문에 목적격 조사 ‘을’을 사용하지 않아도 된다고 용인할 수도 있으나, 조사 누락은 엄격하게 적용하여 누락 오류로 처리한다. 이 경우는 ‘등’ 맨 마지막에만 [오류 위치-목적격조사], [오류 양상-누락] 오류로 처리한다.

- 관형사형 전성어미 누락은 다음과 같은 것들이 해당된다.

<예> 10년 후에 내가 가(√ 갈) 수 있다.

☞ 관형사형 전성어미를 누락한 오류로, 관형사형 전성어미 ‘-(으)ㄴ’의 [누락]으로 주석한다. 경우에 따라서 받침을 제대로 쓰지 못한 오철자 오류로 볼 수도 있으나, 주석의 일관성을 위해 누락 오류로 주석한다. 단, 구어에서는 받침 발음을 실현시키지 못한 것인지, 문법 요소를 누락시킨 것인지에 대한 구분이 필요하므로 주의하여 처리한다.

- 주의: 한 단어에서 단순 철자(음소)가 누락된 경우는 오철자 오류로 ‘누락’이 아닌 ‘오형태’로 처리한다.

<예> 사라(√ 사람)마다 님비 현상이 다 있을 것이다.

☞ 사람의 종성 ‘ㄹ’이 누락되었으나, 문어에서 단어 내에서의 음소 생략은 오형태 오류로 처리한다.

아침에 친구를 만나서(√ 만나서) 혼자 청소합니다.

☞ 동사 ‘만나서’에서 받침 ‘ㄴ’이 생략된 형태는 누락이 아닌 철자의 오류로 처리하여 오형태 오류로 처리한다.

(2) 첨가(ADD)

- [정의] 첨가 오류는 완전한 발화에서 나타나지 말아야 할 형태가 쓰인 경우나 중복된 형태를 사용한 경우를 말한다.

- [주석 방식] 첨가된 부분은 해당 부분의 교정 어절 없이 첨가된 위치를 오류 위치로 주석하고, 오류 양상을 첨가(ADD)로 주석한다.

<예> 종일에(√종일) 반 친구와 나는 만나서 언제나 재미있는 시간을 하고 있어요.
 ☞ 부사 ‘종일’에 불필요하게 부사격 조사 ‘에’를 첨가한 오류로, [오류 위치-부사격 조사], [오류 양상-첨가(ADD)]로 주석한다.

- [처리 기준] 첨가 오류는 하나의 유의미한 교정 어절에만 주석하고, 첨가된 요소로 인해 뒤 따라 오는 요소들은 첨가 오류로 처리하지 않는다. 즉, 용언이나 체언(실질어휘)의 첨가로 인해서 뒤따라오는 어미나 조사(기능어휘)는 첨가 오류로 주석하지 않도록 주의한다. 이때 시스템상에서의 처리 방식은 조사나 어미로 분석된 형태소를 체언 또는 용언에 결합시켜 하나의 첨가 오류로 주석한다.

<예> 머지않은 장래에 장래에(√첨가) 인간 복제도 가능하게 될 것이다.
 ☞ 문어에서 ‘장래에’를 두 번 중복하여 썼기 때문에 두 번째 ‘장래에’를 첨가 오류로 주석한다. 이때 부사격 조사 ‘에’는 명사 ‘장래’로 인해 따라온 요소로서 ‘장래’와 ‘에’ 각각을 첨가 오류로 주석하지 않고, 명사 첨가 오류로만 주석한다. 따라서 [오류 위치-명사], [오류 양상-첨가]로 주석한다.

- [처리 예시] 첨가 오류의 오류 위치별 일부 예시는 다음과 같다.
- 조사 첨가는 다음과 같은 것들이 해당된다.

<예> 대만의 수도가(√수도) 타이베이
 ☞ 주격조사 ‘가’가 첨가된 오류로 주석한다.
 ※참고: 이때 ‘수도가’를 ‘수도인’으로 교정할 경우에는, 주격 조사 ‘가’와 서술격조사 ‘이’의 대치 오류로도 볼 수 있으나 본 연구에서는 ‘최소 수정 원칙’으로 주격조사 첨가 오

류로 처리한다.

- 표현 문형 첨가는 다음과 같은 것들이 해당된다.

<예> 저는 <name> 한국어센터에서 공부하고 있는 동안 선생님한테서 도움이 많이 받아 공부하고 있는(√ 공부하는) 시간이 아주 즐거웠습니다.

☞ ‘-고 있’이 첨가된 오류로 연결어미와 보조용언 각각을 오류 위치로 주석하며, ‘-고 있다’는 표현 문형 목록에도 있기 때문에 표현 문형도 중복 주석한다. 따라서 [오류 위치-연결어미, 보조용언/표현 문형], [오류 양상-첨가] 오류로 주석한다.

(3) 대치(REP)

- [정의] 대치 오류는 의미적 오류로 서로 다른 의미의 어휘를 바꾸어 쓴 경우를 말한다. 즉, 학습자가 어휘의 의미나 용법을 잘못 이해하여 다른 어휘를 사용한 경우이다.
- [주석 방식] 대치되어야 할 형태소(품사)를 오류 위치로 주석하고, 오류 양상을 대치(REP)로 주석한다.

<예> 직접 비판을 받을 때보다 상처가 더 많은(√큰) 것이다.

☞ 맥락상 ‘상처가 많다’보다 ‘상처가 크다’가 더 적절한 표현으로, 형용사 ‘많다’와 ‘크다’의 대치 오류로 처리한다. 따라서 [오류 위치-형용사], [오류 양상-대치(REP)]로 주석한다.

전통의 아름다움이 사람들에게 알려주는 것도 전통을 보존하려고(√ 보존하려면) 해야 할 일이다.

☞ 연결어미 ‘려고’와 ‘려면’의 대치 오류로, [오류 위치-연결어미], [오류 양상-대치(REP)]로 주석한다.

- [처리 기준] 대치 오류는 한국어에 없는 표현이나 학습자의 모국어를 사용한 경우도 포함한다.

<예> 그런데 요즘 부모님들이 자식이 2살부터 play group(√유치원)에 보내는데 놀면서 유치원에 입학하기 위해 준비한다.
 ☞ 영어 단어를 그대로 사용한 경우, 오류 위치를 해당 품사로, 오류 양상을 대치 오류로 주석한다. 즉, [오류 위치-명사], [오류 양상-대치]로 주석한다.

- 일상적인 언어생활에서 보편적으로 쓰지 않는 것으로 판단되는 외국어를 사용한 경우도 대치 오류에 포함된다. 외래어인지 외국어인지 판단하기 어려운 경우는 <표준국어대사전> 등재 여부를 참고하여 판단한다.

<예> 이메일 에드레스(√주소) 다 있어요.
 ☞ ‘이메일’은 <표준>에도 등재되어 있고 일상적으로도 많이 쓰이는 어휘이므로 오류로 처리하지 않으나, ‘어드레스’는 <표준>에 등재되어 있지만 전산 분야와 같은 특수 분야에서 한정적으로 사용되는 의미로 등재되어 있어 ‘외국어’ 사용으로 간주하여 대치 오류로 주석한다.

- 피동과 사동은 한 단위로 보고 대치 오류로 처리한다. 즉, ‘-어지다’, ‘-이/히/리/기/우/구/추(접사)’, 사동 표현 ‘-게 하다’, 피동 표현 ‘-게 되다’ 등은 대치 오류로 처리한다.

<예> 누가 돈이 없다면 행복할 수 없다고 생학했는데 한편에 일부 사람은 가족이 가장 중용 생각했는데 그 이유 점은 돈으로 바뀔(√바꿀) 수 없다고 지적을 했다.
 ☞ ‘바꾸다’를 써야 하는 자리에 피동사 ‘바뀌다’를 사용하였기 때문에 [오류 위치-동사], [오류 양상-대치], [오류 층위-피동] 오류로 주석한다.

- 대치 오류의 경우 대치된 요소로 인해 뒤 따라 바뀌는 요소들은 대치 오류로 처리하지 않는다. 즉, 용언이나 체언(실질어휘)의 대치로 인해서 바

뛰는 어머니 조사(기능어휘)는 교정어절만 써주고 대치 오류로 주석하지 않도록 주의한다.

<예> 그들은 환경문제보다 자기가(✓자신의) 먹고 살기가 더 중요하다고 생각한다.

☞ 대명사 ‘자기’보다 ‘자신’이 더 자연스럽다. 따라서 대명사의 [대치] 오류로 처리한다. 단, 주격조사 ‘가’도 관형격 조사 ‘의’로 대치되지만, 이는 앞의 대명사 대치로 인한 것이기 때문에 뒤의 조사의 경우는 대치 오류로 주석하지 않고 교정어절만 써주도록 한다.

○ [처리 예시] 대치 오류의 오류 위치별 일부 예시는 다음과 같다.

○ 조사 대치는 다음과 같은 것들이 해당된다.

<예> 한국 영화를(✓영화는) 재미있습니다.

☞ 보조사 ‘는’ 자리에 목적격 조사 ‘를’을 잘못 사용한 경우이므로 [오류 위치-목적격 조사], [오류 양상-대치] 오류로 주석한다.

○ 명사 대치는 다음과 같은 것들이 해당된다.

<예> 보고 결정하는 이야기가(✓일이) 많이 있다.

☞ 명사 ‘이야기’를 ‘일’로 대치하므로 [오류 위치-명사], [오류 양상-대치] 오류로 주석한다. 명사가 바뀔므로 해서 뒤따라 바뀌는 조사 ‘가’와 ‘이’는 대치로 주석하지 않는다.

○ 연결어미 대치는 다음과 같은 것들이 해당된다.

<예> 저는 지난 주말에 날씨가 좋니까(✓좋아서) 기숙사 친구하고 같이 한강공원에 갔습니다.

☞ ‘-아서/어서’를 사용해야 할 곳에 ‘-니까’를 잘못 사용하였으므로 [오류 위치-연결어미], [오류 양상-대치, 오형태(‘-니까’의 활용형도 잘못 사용)], [오류 층위-활용] 오류로 주석한다.

- 주의: 대치와 오형태 오류 판단 시, 의미적 대치인지, 오형태 오류인지 혼동되는 경우가 있다. 다시 말해서, 오류의 원인이 의미의 문제인지 형태의 문제인지를 고려해야 하는데, 이럴 경우 문맥을 고려하여 오류를 판단하며, 학습자들이 얼마나 이러한 오류를 보일 수 있는가를 고려한다. 예를 들어, 한국어 교육 과정에서 다루고 있는 어휘인지, 숙달도를 고려했을 때 사용할 수 있는 어휘인지, 해당 맥락에서 출현할 수 있는 어휘인지 등을 고려하여 처리할 수 있다. 아울러 교정은 최소 수정 원칙을 기본으로 하되, 모어 화자에게 자연스러운 용인 가능한 교정 어절로 수정하도록 한다.
- 문맥에 따른 유추를 통한 대치 오류 판단의 예는 다음과 같다.

<예> 긴정한(✓진정한) 아름다움이란 착한 말씀(✓마음씨)이다.
 ≡ 문맥을 통해 ‘말씀’은 ‘마음씨’라고 유추해볼 수 있다. 이처럼 전체 맥락을 통해 ‘마음씨’라는 교정 어절을 추정을 해볼 수 있다면 [대치] 오류로 주석한다.

<예> 저녁에 커피를 마시면서 간간한(✓간단한) 책을 읽고 싶다.
 ≡ ‘간간하다’라는 어휘가 존재하나, 학습자의 수준 및 의도를 고려했을 때, ‘간간하다’를 사용했다고 보기 어렵다. 따라서 ‘간간하다’와 ‘간단하다’의 어휘 대치 오류로 판단하지 않고, ‘간단하다’의 오형태(오철자) 오류로 주석한다. 아울러 ‘간단한’ 책이라고 하면 엄밀한 의미에서 정확한 표현이 아니라고 판단될 수도 있으나, 본 연구에서는 최소 교정을 원칙으로 하며, ‘가벼운 또는 단순한’의 의미로 모어 화자들도 사용할 수 있는 표현으로 보고 이와 같은 경우 ‘간단한’의 오철자 오류, 즉[오형태] 오류로 주석한다.

(4) 오형태(MIF)

- [정의] 오형태 오류는 어휘나 문법의 조합 양상과 활용 형태가 잘못된 형태로 제시된 경우를 말한다. 즉, 단어 내 도치나 이형태 사용 등을 사용한 경우와 의미적으로 전혀 관련이 없는 항목이 선택된 경우, 어휘나 문법을 사용함에 있어서 다른 어휘나 문법으로 대체하여 이해할 가능성이 없는 경우로 형태가 잘못 사용된 경우를 말한다.
- [주석 방식] 오형태 오류는 음소 단위 형태를 잘못 쓴 오철자 오류와 용

언 활용, 조사 이형태 곡용, 어미 활용 등 형태를 잘못 활용한 경우를 포함한다. 따라서 오철자 및 잘못된 활용이 나타난 부분을 오류 위치로 주석하고, 오류 양상을 오형태(MIF)로 주석한다. 단, 오철자 오류는 오류 층위에 활용(MCJ)을 주석하지 않도록 주의하고, 조사 이형태를 잘못 사용한 경우는 오류 층위에서 굴절(곡용)(MDC)으로 주석하고, 용언의 규칙/불규칙 활용과 어미 활용을 잘못된 경우는 굴절(활용)(MCJ)으로 주석한다.

<예> 우리나라에서 과일들하고 야채들도 많아서 과일와(√과) 야채도 다른 나라에 팔아요.
 ☞ 접속조사의 이형태를 잘못 사용한 경우로, [오류 위치-접속조사], [오류 양상-오형태(MIF)], [오류 층위-굴절(곡용)(MDC)]로 주석한다.
 지금 친구 같이 등산에 가시다(√갑니다).
 ☞ 종결어미 ‘니다’에 대한 철자 오류로, [오류 위치-종결어미], [오류 양상-오형태(MIF)] 오류로 주석한다.

- [처리 예시] 오형태 오류의 오류 위치별 일부 예시는 다음과 같다.
- 명사 오형태는 다음과 같은 것들이 해당된다.

<예> 우리 집에 물(√문)을 열리면 계단을 있다.
 ☞ 맥락상 ‘문’을 써야하는데 ‘물’을 쓴 경우, ‘물’이라는 단어가 존재하기 때문에 의미적인 단어와 단어 간의 대치로 볼 수 있으나, ‘문’과 유사한 형태를 잘못 쓴 오철자 오류로 판단한다. 따라서 [오류 위치-명사], [오류 양상-오형태] 오류로 주석한다.

- 조사 오형태는 다음과 같은 것들이 해당된다.

<예> 20, 30대 남녀는 친구을(√를) 중요하게 생각하는 사람들이 많았다.
 ☞ 받침으로 끝날 때 목적격 조사 ‘를’을 써야하는데, ‘을’을

썼기 때문에 조사를 잘못 활용하여 쓴 것으로 [오류 위치-목적격조사], [오류 양상-오형태], [오류 층위-곡용] 오류로 주석한다.

○ 선어말어미 오형태는 다음과 같은 것들이 해당된다.

<예> 제주 친구하고 옥등산에서 등산을 가세요(√갔어요).
 ☞ 과거 시제를 나타내는 선어말 어미 ‘-았-’이 생략된 형태이다. 그러나 ‘가어요’로 쓰지 않고 ‘가세요’로 썼기 때문에 이것은 과거를 인식하고 있다고 보고 오형태 오류로 처리한다. 즉, ‘해습니다’, ‘마셔지만’, ‘와지만’처럼 과거 시제 선어말 어미 ‘었’에서 ‘쓰’를 누락시킨 경우는 선어말 어미 오형태(오철자) 오류로 처리한다. [오류 위치-선어말 어미], [오류 양상-오형태] 오류로 처리한다.

○ 연결어미 오형태는 다음과 같은 것들이 해당된다.

<예> 갈 수 있는다면(√있다면) 언제까지도 기다린다”고 해서 희망자들이 속출하고 있다.
 ☞ ‘있다면’을 써야할 자리에 ‘있는다면’으로 연결어미를 잘못 활용하여 쓴 것이기 때문에 [오류 위치-연결어미], [오류 양상-오형태], [오류 층위-활용] 오류로 주석한다.

○ 종결어미 오형태는 다음과 같은 것들이 해당된다.

<예> 그 이유는 제가 우라 아내보다 한국에 돈을 잘 못 벌읍니다(√법니다).
 ☞ 동사 ‘벌다’를 활용하여 ‘법니다’로 써야하는데 ‘벌’을 그대로 사용하고 있기 때문에 [오류 위치-동사, 종결어미], [오류 양상-오형태], [오류 층위-활용] 오류로 처리한다.

아름답니다(√아름답습니다), 시끄럽니다(√시끄럽습니다)

재미있입니다(√재미있습니다), 맛있입니다(√맛있습니다)
 ☞ ‘비니다/습니다/니다/입니다’는 종결어미 오형태 활용 오류로 처리한다.

4) 오류 층위

(1) 발음

① 음소(PP)

- [정의] 음소 오류는 음소 단위에서 발화가 잘못 사용된 경우를 말한다.
- [주석 방식] 잘못 발음된 품사에 오류 위치를 주석하며, 오류 양상은 주석하지 않고, 오류 층위에 음소(PP)를 주석한다.
- [처리 기준] 단어 내에서 명확하게 다른 음소로 발음한 경우나 음소를 발음하지 못한 경우, 음소 오류로 주석한다.
- 구어에서 받침과 같은 특정 음소를 실현시키지 못하였을 때에는 오류 양상에 누락을 주석하지 않도록 주의한다. 문법적 요소를 누락시킨 경우에는 누락을 주석할 수 있지만, 학습자가 발음하지 제대로 하지 못하여 실현되지 않은 음소에 대해서는 오류 층위에서 음소만을 주석한다.

<예> 팔이(√빨리), 말을::, 즐= 아:: 잘해서::,
 ☞ 자음 ‘ㅁ’을 ‘ㄹ’로 발음하여, 음소 오류로 주석한다. [오류 위치-일반부사], [오류 양상-없음], [오류 층위-음소]
 예:: 코피(√커피)= 카페에서::, 예:: 공부를, 합니다::.
 ☞ 모음 ‘아’와 ‘오’를 교체하여 발음하므로 음소 오류로 주석한다. [오류 위치-명사], [오류 양상-없음], [오류 층위-음소]

부모니(√부모님)

☞ 한 단어 안에서 받침을 발음하지 못한 경우도 마찬가지로 음소 오류로 주석한다. [오류 위치-명사], [오류 양상-없음]

② 음절(PS)

- [정의] 음절 오류는 음절 단위에서 발화가 잘못 사용된 경우를 말한다. 음절 오류는 원래 음절보다 적게 또는 더 많게 발화한 경우가 해당된다.
- [주석 방식] 음절 오류가 발생한 품사에 오류 위치를 주석하며, 오류 양상은 주석하지 않고, 오류 층위에 음절(PS)을 주석한다.
- [처리 기준] 원래의 음절수보다 늘어 발음하거나 축약이 불가능한 단어를 축약하여 발음한 경우 음절 오류로 주석한다.

<예> 예:: 매이루(√매일) 노무::, 즐겁..습니다::.

☞ 2음절인 ‘매일’에 모음을 삽입하여 3음절로 발음하고 있으므로 음절 오류로 주석한다. [오류 위치-일반부사], [오류 양상-없음], [오류 층위-음절]로 주석한다.

제가, 한국에 와서.. 사 개워르(개월) 정도:: 살았습니다::

☞ 2음절인 ‘개월’을 받침 ‘ㄱ’과 모음 ‘으’를 결합하여 3음절로 발음하고 있으므로 음절 오류로 주석한다. [오류 위치-의존명사], [오류 양상-없음], [오류 층위-음절]로 주석한다.

- 학습자가 단어를 완전히 발화하지 못하였을 때에는 음절이 줄어든 것처럼 보이거나 이는 형태를 제대로 구현시키지 못한 오류로 음절이 아닌 오형태 오류로 처리한다.

<예> 길가에 아 음 쓰레기통을 쓰레(√쓰레기) 버리고

☞ ‘쓰레기’의 음절이 줄어든 것처럼 보일지라도 이는 발음 차원의 문제로 보기 힘들기 때문에 형태 오류로 처리한다.

③ 음운규칙(PC)

- [정의] 음운규칙 오류는 구어 발화에 나타난 필수적 음운 규칙의 일탈을

말한다. 유음화, 연음화하여 발음해야 하는데, 글자 그대로 절음화하여 발음한 경우가 해당된다.

- [주석 방식] 음운규칙을 적용하지 못한 품사를 오류 위치로 주석하며, 오류 양상은 주석하지 않고, 오류 층위에 음운규칙(PC)을 주석한다.
- [처리 기준] 음운규칙 오류는 음운규칙을 적용하지 않은 경우와 적용했으나 잘못 발음한 경우 두 가지로 나눌 수 있다.
- 첫째, 음운규칙 미적용 오류는 음운규칙으로 인해 철자와 다르게 발음해야 하나, 학습자가 철자대로 절음화 하여 발음한 경우다. 학습자가 철자대로 발음했기 때문에 원어절과 교정어절의 형태는 동일하다.

<예> 설날(√설날)

☞ 학습자가 발화 시 유음화 규칙에 따라 [설랄]로 발음하지 않고 [설]과 [날]을 각각 끊어서 개별 음절 발음에 충실하였다면 ‘음운규칙’ 오류로 처리한다. 이밖에 ‘육학년’을 [유강년]으로 발음하지 않고 글자 그대로 [육학년]으로 발음한 경우나 ‘학교’를 [학교]로 발음하지 않고 글자 그대로 [학교]라고 발음한 경우를 음운규칙 오류로 처리한다.

무조건(√무조건) ☞ 경음화 미적용

같이(√같이) ☞ 구개음화 미적용

신라(√신라) ☞ 유음화 미적용

앞에(√앞에) ☞ 연음 미적용

먹는(√먹는) ☞ 비음화 미적용

☞ 위의 예들은 음운규칙을 적용하지 않고, 철자 그대로 발음한 경우다. 따라서 음운규칙이 적용되어야 하는 위치에 오류 위치를 주석하고, 오류 층위에는 음운규칙을 주석한다.

- 둘째, 음운규칙 미적용 외에 음운규칙을 적용시켜야 하는 단어에서 잘못 적용한 경우도 음운규칙 오류로 주석한다. 그러나 이때에는 음운규칙과 음소 오류를 중복 주석한다. 그러나 음운규칙을 적용시켜야 하는 단어이나 음운규칙과 상관없는 위치에서 다른 음소로 발음한 경우는 ‘음소’ 오류로만 주석한다.

<예> 학교(√ 학교)

☞ [학교]로 발음해야 하는데 [학교]로 발음했을 경우, 음운규칙과 음소 오류를 중복 주석한다. 그러나 음운규칙이 적용되지 않는 위치에서 다른 음소로 발음한 경우는 음소(PP) 오류로만 처리한다(예 핵교(√ 학교)).

- 학습자가 음운규칙을 몰라서 발생시킨 오류인지 음소를 변별하지 못하여 발생시킨 오류인지를 판단하여 오류 층위에 음운규칙과 음소를 구분하여 주석한다.

<예> 한국어(√ 한국어)

☞ 음운규칙 미적용 오류. 오류 위치: 명사, 층위: 음운규칙(PC)

한국거(√ 한국어)

☞ 오류 위치: 명사, 층위: 음운규칙(PC)

한구꺼(√ 한국어)

☞ 음운규칙은 적용하였으나 음소 변별 실패
오류 위치: 명사, 층위: 음소(PP)

한국꺼(√ 한국어)

☞ 오류 위치: 명사, 층위: 음운규칙(PC), 음소(PP)

- 한 단어 내에서 음운 규칙 또는 음소와 관련한 오류가 나타났을 때에는 그 단어의 품사를 주석하면 되지만, 체언과 조사, 용언의 어간과 어미 사이에서 음운 규칙이 적용되면서 나타난 오류의 경우, 오류 위치를 무엇으로 삼을 것인가가 문제가 된다. 이때에는 오류가 나타난 발음이 어느 부분의 영향을 받았는지를 고려하여 처리한다. 예를 들어, 모음으로 시작되는 조사나 어미의 경우 본래 음가를 가지고 있지 않았으나, 선행하는 음절의 받침 발음에 영향을 받아 초성에서 그 발음이 실현된다. 따라서 조사나 어미의 초성 위치에서 오류가 나타나더라도 오류 위치를 영향을 미친 체언과 용언의 어간으로 주석한다.

<예> 한국계(√ 한국에) 갔어요.

☞ 오류 위치: 고유명사, 오류 층위: 음운규칙(PC)

한국계(√한국에) 갔어요.

☞ 오류 위치: 고유명사, 오류 층위: 음운규칙, 음소(종성이 그대로 실현됨과 동시에 [ㄱ]와 [ㄱᄃ] 음소를 변별하지 못한 오류로 해석한다.)

네 말이 맞쵸요(√맞아요).

☞ 오류 위치: 형용사, 오류 층위: 음운규칙(PC), 음소(PP)
음식글(√음식을) 먹어요.

☞ 오류 위치: 명사, 오류 층위: 음운규칙(PC)

- 연음이 적용되는 단어에서 연음은 시켰지만 음소를 잘못 발음하였을 때에는 음소 오류로 처리한다.

<예> 네 말이 마차요(√맞아요).

☞ 오류 위치: 형용사, 오류 층위: 음소(PP)

음시글(√음식을) 먹어요.

☞ 오류 위치: 명사, 오류 층위: 음소(PP)

- 종성이 다음 음절의 초성으로 넘어가서 발음되는 연음과 달리, 경음화나 격음화 오류는 오류의 양상에 따라, 발음된 위치를 고려하여 오류 위치를 주석한다.

<예> 물건을 찰코(√찰고찰꼬).

☞ 오류 위치: 연결어미, 오류 층위: 음소(PP)

[찰꼬]로 발음되어야 하는데, 연결어미 ‘고’를 ‘코’로 발음하였다. 이것이 앞의 종성 [ㄷ]의 영향을 받은 것인지, 학습자가 초성 위치에서 [ㄱ]와 [ㄱᄃ]를 변별하지 못한 것인지를 판별하기는 어렵다. 본래 음가를 가지고 있지 않은 어미가 연음 규칙에 따라 발음을 잘못하였을 때에는 선행하는 어간의 품사를 오류 위치로 주석한다. 그러나 ‘고’처럼 본래 음가를 가지고 있던 위치에서 음소를 변별하지 못하였을 때에는 해당 위치에 오류 위치를 주석한다. 따라서 이러한 경우 오류 위치:

연결어미, 오류 층위: 음소(PP)로 주석한다.

응답뻘(√응답한) 결과

☞ 오류 위치: 명사, 오류 층위: 음운규칙(PC)

☞ 오류 위치: 동사파생접시마, 오류 층위: 음소(PP)

격음화가 적용되어 [응다판]으로 발음해야 하지만, [응답]을 그대로 발음하는 동시에 격음으로 발음해야 하는 [판]을 [뻘]으로 발음하였다. 따라서 ‘응답’과 동사파생접시마 ‘하’를 형태 분리하므로 ‘응답’에는 음운규칙을 주석하고, ‘하’에는 음소를 주석한다.

④ 원어식 발음(PN)

- [정의] 원어식 발음은 학습자의 외국어 발음으로 인한 발화 오류를 말한다. 즉, 외국어나 외래어 발음 시, 원어에 가까운 소리로 발음하는 경우다. 이는 한국어 모어 화자에게서도 일어나는 현상이기는 하나 외국인 학습자에게서 그 빈도가 더 잦고, 발음 또한 모어 화자의 그것과 많이 다르다. 따라서 원어식 발음은 외래어 표기법과 불일치하므로, 이를 표시해 주는 차원에서 외국어와 외래어의 경우, 한국어와 다르게 발음했을 때 ‘원어식 발음’ 오류로 주석한다.
- [주석 방식] 원어식 발음으로 발음한 품사에 오류 위치를 주석하며, 오류 양상은 주석하지 않고, 오류 층위에 원어식 발음(PN)을 주석한다.
- [처리 기준] 외국어나 외래어에서 한국어의 표준 발음과 다르게 발음한 경우 원어식 발음 오류로 주석한다. 원어식 발음(PN)의 경우, 발음의 차이로 인해 한국어의 외래어 표기와 다르게 음절이 줄거나 늘어날 수 있다. 이때는 음절 오류가 아닌 원어식 발음 오류로 주석한다.

<예> 인텔뷰(√인터뷰)

세너(√센터)

☞ 한국어 외래어 표기법과 다르게 원어식 발음으로 발화한 경우 원어식 발음 오류로 처리한다.

팔너(√파트너)

그대 처음에 같 뻔남(√베트남)에서

☞ 원어식 발음의 차이로 인해 한국어의 외래어 표기와 다르게 음절이 줄어들었다. 이때는 음절 오류가 아닌 원어식 발음 오류로 주석한다.

○ 학습자 모국어의 외래어 발음도 포함한다.

<예> [요한스버그](√요하네스버그)

이.. 기자는:: 한국에서:: 이:: 마약, 없는::, 아:: [이미지])(√이미지) 줌:: 없어,지고:: 있습니다 지금.

☞ 외국어 발음 오류로 외래어 표기법과 ‘다르다’는 차원에서 [오류 층위-원어식 발음] 오류로 처리한다.

⑤ 중간 발음(변이음포함) (PA)

- [정의] 중간 발음은 변이음을 포함하여 학습자의 외국어와 한국어의 중간 발음으로 인한 발화 오류를 말한다.
- [주석 방식] 변이음으로 발음한 품사에 오류 위치를 주석하며, 오류 양상은 주석하지 않고, 오류 층위에 중간 발음(PA)를 주석한다.
- [처리 기준] 중간발음은 한국어 모어 화자와는 다른 발음, 즉 변이음과 관련된 오류와 음소와 음소 간의 중간 발음 두 가지가 포함된다.
- 첫째, 음성과 관련된 변이음은 구어 전사 과정에서 유성음, 무성음으로 표기해준 경우에 근거해 변이음 오류로 처리한다. 이때는 음성과 관련된 문제로, 원어절과 교정어절의 형태는 동일하다. 다만, 변이음은 구어 전사에서 전사자 특성에 따라 다르게 식별될 수 있는 문제가 있다. 따라서 구어 전사 과정에서 변이음이 분명하게 식별된 메모에 근거하여 주석하도록 한다.

<예> 가구(√가구)

☞ ‘가’의 ㄱ을 유성음으로 발음

‘구’의 ㄱ을 무성음으로 발음

‘구’에서 ㄱ과 ㄱ의 중간 발음

파란색(√ 파란색)

☞ ‘파’에서 f로 발음됨

☞ 구어 전사에서 위와 같이 기술한 메모에 근거해 중간 발음(PA) 오류로 주석한다.

- 둘째, 학습자가 원어절과 교정어절 사이의 발음, 즉 음소 간의 중간 형태로 발음한 경우도 중간발음 오류로 처리한다.

<예> 여자가<note>‘여’를 ‘으’와 ‘유’의 중간 발음으로 발음</note>

☞ 구어 전사 시, 음소와 음소 간의 중간 발음으로 들릴 때 괄호 안에 (‘X’와 ‘Y’와 중간 발음)으로 표기한다. 이를 바탕으로 하여 중간 발음으로 표기된 경우, [오류 위치-명사], [오류 양상-없음], [오류 층위-중간 발음]으로 주석한다.

화반수(√ 과반수)

☞ ‘ㅎ’과 ‘ㄱ’의 중간 발음

☞ 구어 전사에서 위와 같이 ‘음소’와 ‘음소’의 중간 발음으로 기술한 메모에 근거해 중간 발음(PA) 오류로 주석한다.

(2) 형태

① 단어 형성[합성법](MCP)

- [정의] 단어 형성[합성법] 오류는 조어 과정에서 발생하는 오류를 말한다. 즉, 학습자가 존재하지 않는 어휘를 생산해 내는 오류가 포함된다.
- [주석 방식] 학습자가 조어 과정에서 형태를 잘못 만들어 낸 경우, 오형태 오류로 볼 수도 있다. 그러나 오형태 오류가 오철자 오류와 이형태 활용 오류에 해당하는 오류 양상이라고 할 때, 합성과 파생 관련한 오류를 형태 오류에 포함시킬 수 있는가가 문제가 된다. 본 연구에서는 파생과 합성 오류의 경우 오형태 오류로 보기 어렵고, 오류 양상은 수의적 주석이므로 오

류 양상을 필수적으로 주석하지 않고 오류 층위에서 파생(MDV)과 합성(MCP)만을 주석하도록 한다.

- 따라서 학습자가 생산해 낸 형태가 한국어에는 없는 합성어일 경우, 오류 양상은 주석하지 않고 오류 층위에서 합성(MCP)으로 주석한다.

<예> 해물 고기(√물고기)가 많 많았어요.

☞ ‘물고기’를 ‘해물’과 ‘고기’로 잘못 합성하였으므로 [오류 위치-명사], [오류 양상-없음], [오류 층위-합성법]으로 처리한다.

② 단어 형성[파생법](MDV)

- [정의] 단어 형성[파생법] 오류는 조어 과정에서 접사를 잘못 사용한 오류를 말한다.
- [주석 방식] 학습자가 파생접사(유사파생접사 포함)를 사용하여 생산해 낸 형태가 한국어에는 없는 파생어일 경우, 오류 양상은 주석하지 않고 오류 층위에서 파생(MDV)으로 주석한다.
- 단, 접사와 접사의 대치의 경우나 접사의 불필요한 첨가 또는 생략은 오류 양상에 대치, 첨가, 생략으로 주석한다.

<예> 친구와 그 사람을 사귀하면(√사귀면) 제일 좋은 일 그 사람이 멋있습니다.

☞ 동사 ‘사귀다’에 다시 동사파생접미사 ‘-하다’를 붙여 한국어에는 없는 형태를 만들어 낸 것으로 [오류 위치-동사], [오류 양상-없음], [오류 층위-파생법] 오류로 주석한다.

그 다음에 여름에는 수영을 하다든가 성풍기를 사용하다든가 해서 건강적인(√건강한) 감온 방법이 선택하면 좋다.

☞ 접사 ‘하다’ 대신 ‘적’을 사용해 형용사를 파생시킨 경우로, 이때에는 접사 ‘적’과 ‘하다’ 대치 오류로 주석한다. [오류 위치-접미사], [오류 양상-대치], [오류 층위-파생법]으로 주석한다.

이런 데다가 의사 선생님께 의하면 균형 깨진 영양성(√영양) 바람에 났던 여드름이 더 날 계속한다고 걱정했는데도 그렇지 않습니다.

☞ 접미사 ‘-성’을 과잉적용한 오류로 [오류 위치-접미사], [오류 양상-첨가], [오류 층위-파생법]으로 주석한다.

- [처리 기준] 동사 어간에 ‘하다’를 붙여 한국어에는 없는 동사를 만들어냈을 경우는 파생 오류로 처리한다. 이는 합성 오류로도 볼 수 있지만 형태주석에서 ‘-하다’를 파생접미사로 처리하고 있어 처리의 연계성과 일관성을 고려하여 파생 오류로 처리한다.

<예> 그날 수업 후 집에 도착하자마자 어머니가 나한테 혼했다(혼냈다).

☞ 이는 ‘혼내다’를 명사 ‘혼’에 ‘하다’를 붙여 ‘혼+하다’로 한국어에는 없는 어휘를 생산한 것이다. 이 경우 ‘하다’를 동사로 볼 수도 있고 파생접사로도 볼 수 있다. 형태주석에서는 이를 동사파생접미사로 주석하기 때문에 오류주석에서도 파생 오류로 판단하도록 한다. [오류 위치-동사], [오류 양상-없음], [오류 층위-파생법]으로 처리한다.

- 접사는 문법범주에서 논의하는 존대, 피동/사동, 복수 표지 중 피동/사동만 대치 오류로 처리한다. 접사 중 문법적인 성격 강한 존대(님), 복수 표지(들) 등은 형태주석에서 접사로 따로 분리하여 처리하기 때문에 형태주석에서 분리하는 접사들은 ‘누락/첨가’로 처리하고, 피동/사동은 어휘 대치로 처리한다. 또한 어휘적 의미를 더해주는 접사의 경우, 예를 들어 ‘사과’를 ‘곶사과’로 썼을 때에는 형태분석에서 ‘곶’을 분리하지 않기 때문에 어휘적 의미를 더해주는 접사가 덧붙여진 단어는 대치 오류로 처리한다.
- 그밖에 형태주석에서 접사로 분리하지는 않지만 유사파생접사로 볼 수 있는 형태들을 사용하여 어휘를 파생시킨 경우는 오류 양상은 주석하지 않고, 오류 층위에 파생으로 주석한다.

<예> 그리고 나는 계속 매일 지각했을 때 나는 번금도 내고 반성서(√반성문)도 썼다.

☞ 반성서에서 ‘서’는 형태 주석에서 분리하여 처리하지 않고, 반성서를 하나의 명사로 주석한다. 이를 오류 주석에서는 반성문을 한국어 어휘에는 없는 ‘반성+서’로 파생시킨 것으로 보고 [오류 위치-명사], [오류 양상-없음], [오류 층위-파생법]으로 주석한다.

행운하면 다음 학기는 <name>대학교 어학관(√어학원)에서 4급 공부할 거야.

☞ ‘관’과 ‘원’ 모두 형태 주석에서 접사로 따로 분리하여 처리하지 않는다. 그러나 이는 유사파생접사로 볼 수 있기 때문에 ‘어학원’ 명사를 잘못 파생시킨 오류로 보고 오류 층위에 파생법으로 주석한다.

③ 굴절[곡용](MDC)

- [정의] 굴절[곡용] 오류는 조사의 이형태를 잘못 사용한 경우를 말한다.
- [주석 방식] 굴절[곡용] 오류의 오류 양상은 기본적으로 오형태(MIF) 오류로 주석한다.

<예> 지금 가족가(√가족이) 너무 보고 싶습니다.

☞ 받침 뒤에서 주격조사 ‘가’로 잘못 사용하였으므로 [오류 위치-주격조사], [오류 양상-오형태], [오류 층위-곡용] 오류로 처리한다.

- 굴절[곡용] 오류 주석 시, 체언의 오류로 인한 조사 오류는 오류로 주석하지 않도록 주의한다.

<예> 종일(√종이)을(√를) 낭비할 게 아니라 절약할 것이다.

☞ ‘종일을’은 ‘종이를’로 교정되나, 이때는 체언을 잘못 사용한 것으로 인해 목적격 조사 ‘을’을 썼다고 보고, ‘을’은 곡용 오류로 주석하지 않는다. 따라서 이는 명사 오형태

오류로만 주석하고, 목적격 조사 ‘을’에는 교정어절 ‘를’만 써준다.

④ 굴절[활용](MCJ)

- [정의] 굴절[활용] 오류는 용언의 활용과 어미의 활용을 잘못된 경우를 말한다. 즉, 용언의 규칙 및 불규칙 활용, 어미의 이형태 오류가 포함된다.
- [주석 방식] 활용 양상에 따라 오류 위치를 판단하여 주석하며, 오류 양상은 오형태(MIF)로 주석하고, 오류 층위에 굴절[활용](MCJ)을 주석한다.
- [처리 기준] 용언의 규칙/불규칙 활용과 어미 이형태에 따라 활용 오류를 처리한다. 단순 오철자 오류의 경우 오형태(MIF)만 처리하며, 오류 층위에 활용(MCJ)을 주석하지 않는다.

<예> 미국에 영어 제일 중요하다.(√중요하다).

☞ ‘중요하다’를 ‘중요한다’로 종결 어미를 잘못 활용하였으므로 [오류 위치-종결어미], [오류 양상-오형태], [오류 층위-활용] 오류로 주석한다.

이렇게 살으면(√살면) 정말 행복할 수 있다.

☞ 받침 ‘ㄷ’ 뒤에서 ‘-으면’의 형태로 잘못 활용하였으므로 오형태 활용 오류로 주석한다.

- [쟁점] 활용 오류는 한국어 학습자가 생산하는 형태에 따라 오류 위치를 명확하기 판단하기 어려운 경우들이 있다. 따라서 활용의 양상에 따라 오류 위치를 용언의 어간에만 주거나 어미에만 줄 수도 있으며, 어간과 어미 양쪽에 줄 수도 있다. 각각의 예시는 다음과 같다.
- [용언의 어간 활용 오류 1] 용언의 불규칙 활용을 시키지 않거나 과잉 적용시킨 경우 용언의 어간 활용 오류로 처리한다.

<예> 한국의 여름 날씨는 더웁니다.(√덥습니다.)

☞ ‘ㄷ불규칙 활용’ 형용사인 ‘덥다’를 종결어미 ‘ㄷ니다’ 앞에서 ‘더우’의 형태로 과잉 적용한 경우다. 이처럼 불규칙 활용을 잘못 적용시킨 경우 [오류 위치-형용사/동사]로 주석하고 [오류 양상-오형태], [오류 층위-활용]으로 주

석한다.

- [용언의 어간 활용 오류 2] 한국어 학습자들은 이형태가 없는 어미 앞에서 ‘아/어’나 ‘으’와 같은 매개 요소를 사용하는 경우가 많은데, 본고에서는 이를 학습자가 하나의 어간으로 재구성하고 있는 중간언어로 보고 용언 어간의 활용 오류로 처리한다.

<예> 2015년에는 친구하고 같이 많이 놀아고(√놀고) 싶습니다.
저는 좋아하는 프로그램은 많아지만(√많지만) 다른 방송에
비해서 동물에 대한 다큐멘터리는 프로그램이 제일 좋아한
다.
우리 미래 길에 꼭 잘 조심해고(√조심하고),
☞ 이형태가 없는 어미 ‘고, 지만’ 앞에 ‘아/어’ 또는 ‘으’가
첨가된 경우, 어간 활용과 어미 활용의 구분이 어렵다.
학습자들이 동사 어간에 ‘아/어’를 첨가해 하나의 어간으
로 재구성했다고 볼 수도 있고, ‘아고, 아지만’의 형태로
어미를 잘못 활용한 것으로도 볼 수 있다. 본 연구에서는
이와 같은 경우 ‘놀아’, ‘많아’, ‘해’를 하나의 어간으로 재
구성한 중간언어로 보고 용언의 활용 오류로 처리한다.
이는 학습자들이 연결어미(예: 해고), 종결어미(예: 됩니
다), 관형사형 전성어미(예: 해는) 앞에서 동일한 형태를
생산해 내는 것으로 보아, 학습자들이 ‘해’를 하나의 단위
로 인식하고 활용을 잘못 적용하였다고 판단했기 때문이
다. 따라서 이와 같은 형태들은 해당 품사(용언 어간)를
오류 위치로 주석하고, [오류 양상-오형태], [오류 층위-
활용] 오류로 주석한다.

- ‘하다/되다’ 활용 오류에서 ‘해(√하), 돼(√되)’로 잘못 쓴 경우는 용언의
활용 오류로 처리하는 반면에 ‘하(√해), 되(√돼)’의 경우는 용언의 어간과
어미 활용 오류로 처리함에 주의한다.

<예> 우리 미래 길에 꼭 잘 조심해다(√조심하다),

특정 장면은 항상 중요 메시지가 있어서 삭제되면(√삭제되면)

☞ ‘하다, 되다’에서, ‘해, 돼’로 잘못 쓴 경우는 어간의 활용 오류로 처리한다. 위의 경우 형태소 분석에 따라, ‘해’와 ‘돼’는 동사파생 접미사의 활용 오류로 주석한다.

그리고 한국어 말하야 해요(√말해야 해요)

아무지 친하도(√친해도) 존댓말로 써야 한다.

그래서 시청자 왕따 당할까봐 걱정이 되서(√돼서) 그 물건을 사게 된다.

☞ ‘하다/되다’ 용언 어간+‘어/아’ 계열 어미‘에서 ‘어/아’를 누락시킨 경우는 어간과 어미에 모두 오형태 활용 오류로 처리한다. ‘하다/되다’의 경우 어미와 결합할 때, ‘하/해’, ‘되/돼’로 형태를 바꾸기 때문에 오형태 활용 오류로 볼 수 있다. 그러나 오류 위치를 어간 어미 중 무엇으로 처리하느냐가 문제가 된다. 이 경우, 학습자가 용언 어간, 어미 둘 중 어느 곳을 잘못 사용하였는지 정확히 분리하기 어려워 어간과 어미 양쪽에 오형태 활용 오류를 주석한다.

- [용언의 어간 활용 오류 3] 이밖에 어미 앞에서 ‘ㄴ, ㄹ, ㅂ’ 등의 불필요한 요소를 첨가했을 경우도 용언 어간의 활용 오류로 처리한다.

<예> 왜냐하면 어렸을 때부터 커피숍이나 호텔의 사장님 될고(√되고) 싶어 하기 때문이다.

☞ 연결어미 앞에 ‘ㄹ’ 요소가 첨가된 경우, [오류 위치-동사], [오류 양상-오형태], [오류 층위-활용] 오류로 주석한다.

달른(√다른)

알랐다(√알았다)

☞ ‘ㄹ’ 앞에서 ‘ㄹ’이 첨가된 경우는 [III] 발음의 영향으로 인

한 오철자 오류로 볼 수도 있다. 그러나 본 연구에서는 앞에서 ‘ㄴ, ㄹ, ㅂ’ 요소들이 첨가된 경우 오형태 활용 오류로 보아, 이 역시 오형태 활용 오류로 처리한다.

- [어미 활용 오류 1] 용언의 받침 유무에 따라 어미의 이형태 선택이 달라지는 경우는 활용 오류로 처리한다.

<예> 이렇게 살으면(√살면) 정말 행복할 수 있다.
 ➡ 받침의 유무에 따라 연결어미의 이형태를 선택하여 사용해야 하는데, 잘못 사용한 경우로 [오류 위치-연결어], [오류 양상-오형태], [오류 층위-활용 오류]로 주석한다.

- [어미 활용 오류 2] 이형태의 선택 뿐 아니라 이형태가 있는 어미에서 오류가 나타난 경우 오형태 활용 오류로 처리한다. 따라서 ‘아서/어서’에서 ‘서’만 쓰거나 ‘아도/어도’에서 ‘도’만 쓴 경우, ‘(으)니, (으)면’ 등에서 ‘으’를 쓰지 않은 경우 연결어미 활용 오류로 처리한다. 마찬가지로 종결어미에서도 ‘아요/어요’에서 ‘요’만 쓴 경우 종결어미 활용 오류로 처리한다.
- [어미 활용 오류 3] 또한, 용언 어간과 어미의 축약상의 오류는 활용 오류로 처리한다. 어미 이형태 활용 오류 외에 필수적으로 탈락시켜야 하는데 시키지 않은 경우 또는 과도하게 축약을 시켜버린 경우 모두 어미의 활용을 제대로 모르는 것으로 판단하여 활용 오류에 포함한다.

<예> 가아서(√가서), 가아도(√가도), 가아요(√가요)
 한 시간 쉬요(√쉬어요)
 ➡ ‘가아서’처럼 축약을 시키야 하는데 축약을 하지 않은 경우와 ‘쉬요’처럼 ‘쉬어요’를 과도하게 축약시킨 경우는 [오류 위치-연결어미/종결어미], [오류 양상-오형태], [오류 층위-활용 오류]로 주석한다.

- [어미 활용 오류 4] 종결어미에서 ‘ㅂ니다/습니다’ 외에 ‘니다’ 또는 ‘입니다’를 잘못 사용한 경우도 종결어미 오형태 활용 오류로 처리한다.

<예> 힘들습니다(√ 힘듭니다)

- ☞ 형용사 ‘힘들’과 종결어미 ‘습니다’ 양쪽 모두 활용을 잘못한 것으로, [오류 위치-형용사, 종결어미], [오류 양상-오형태], [오류 층위-활용]으로 주석한다.

아름답니다(√ 아름답습니다)

시끄럽니다(√ 시끄럽습니다)

- ☞ ‘ㅂ’ 받침으로 끝나는 용언 어간의 경우, 용언 어간 ‘아름다’와 종결어미 ‘ㅂ니다’로 결합시킨 것인지, ‘아름답’과 ‘니다’의 형태로 결합한 것인지 불분명하다. 이때, 용언어간 ‘아름답’과 ‘니다’로 결합한 것으로 일괄 처리하고, ‘ㅂ니다/습니다’와 같이 [오류 위치-종결어미], [오류 양상-오형태], [오류 층위-활용] 오류로 주석한다.

재미있입니다(√ 재미있습니다)

맛있입니다(√ 맛있습니다)

- ☞ 형용사에 ‘입니다’를 결합시킨 경우도 종결어미의 활용 오류로 주석한다.

- [용언의 어간 + 어미 활용 오류] 어간 활용과 어미 활용 모두 실패한 경우에는 오류 위치에 해당 용언의 품사와 어미를 모두 주석한다.

<예> 그 이유는 제가 우라 아내보다 한국에 돈을 잘 못 벌읍니다(√ 법니다).

- ☞ 동사 ‘벌다’를 활용하여 ‘법니다’로 써야하는데 ‘벌’을 그대로 사용하고 있기 때문에 [오류 위치-동사, 종결어미], [오류 양상-오형태], [오류 층위-활용] 오류로 처리한다.

한국 CD를 듣면서(√ 들으면서) 한국어를 말하다.

- ☞ ‘ㄷ 불규칙’을 적용시키지 못하였으며, 연결어미 ‘으면서/면서’의 이형태 또한 제대로 사용하지 못한 오류로 동사의 어간과 연결어미 양쪽 모두의 활용 오류로 처리한다.

[오류 위치-동사, 연결어미], [오류 양상-오형태], [오류
층위-활용]으로 주석한다.

⑤ 품사(POS)

- [정의] 품사 오류는 동일 의미 단어의 품사를 잘못 사용한 경우를 말한다.
즉, 같은 의미인 단어의 품사를 제대로 인식하지 못하여 명사를 동사로 사용하거나, 부사를 형용사로 잘못 사용한 경우를 말한다.
- [주석 방식] 품사 오류의 오류 양상은 기본적으로 대치(REP) 오류로 주석하고 층위에 품사(POS)를 주석한다.
- [처리 기준] 오류 층위에서 품사에 해당하는 오류는 품사에서 나타난 오류(품사가 달라진 경우)와 품사를 몰라서 생겨난 오류(품사 혼동으로 인한 오류)로 구분할 수 있다. 이때, 본고에서는 품사가 달라진 것보다 품사를 모르고 있는 것을 우선적으로 적용하여, 품사에 대한 인식이 없어서 생겨난 오류만 품사 오류로 주석한다. 따라서 문맥에 따라 교정 어절이 바뀌면서 품사가 달라진 경우(단순히 원어절과 교정어절의 품사가 상이한 경우, 다시 말해서 의미가 다른 품사)는 품사 오류로 처리하지 않는다.
- 즉, 원어절과 교정 어절이 의미를 공유하면서 품사를 제대로 사용하지 못한 경우, 품사에 대한 인식이 부재한 것으로 간주하고 이를 품사 오류로 처리한다.

<예> 그래서 우리는 빠른(√빨리) 우리 집에 다가했다.
 ☞ 부사 ‘빨리’를 쓸 자리에 형용사 ‘빠르다’를 사용하였다.
 이는 학습자가 동일한 의미의 단어에서 부사 품사를 모르기 때문에 형용사를 관형형으로 사용한 것으로 보고 품사 대치 오류로 처리한다.

- 품사 오류에서 ‘파생/합성’과 관련된 오류는 오류 층위에서 품사(POS)와 단어 형성[합성법](MCP) 또는 단어 형성[파생법](MDV)을 중복 주석한다. 이는 표면상 원어절과 교정어절에서 드러나는 차이에 주목하여 품사 오류로 처리하는 동시에 파생과 합성을 하면서 품사가 바뀌는 경우인데, 품사를 바꾸는 접사에 대한 인식이 없다고 판단한 것이다.
- 따라서 ‘N+하다, 되다, 시키다, 있다, 없다, 나다(화나다, 겁나다, 불나다,

열나다 등)’에서 어간만 사용한 경우는 오류 위치[명사 또는 어근] - 오류 양상[대치] - 오류 층위[품사, 파생/합성(파생접사가 아닌 경우)] 중복 주석 처리한다.

<예> 이 문제들을 방지하기 위해 인터넷 사용(√/사용하는) 사람들이 몇 가지 태도를 갖춰야 한다는 것을 청구한다.

☞ 동사 ‘사용하다’를 쓸 자리에 명사 ‘사용’만 사용하였다. 이는 명사와 동사 품사를 모르고 있다고 보고 품사 오류로 처리한다. 동시에 파생접미사 ‘-하다’를 붙여서 동사를 파생시키지 못한 오류로 보고 파생 오류도 중복 처리한다.

- 내 생가기는 한국 사람들 다른 나라 사람들보다 친절(√/친절한) 것 같다.
- 그리고 한국 친구를 인사고(√/인사하고) 싶습니다.

☞ 위의 예시들은 크게 3가지로 해석해 볼 수 있다.

- 1) 동사와 명사를 구분하지 못하여 동사를 써야할 자리에 명사를 쓴 경우로 학습자들이 품사를 제대로 인식하지 못해서 발생한 오류. 즉, (특히 중국인 학습자의 경우) 명사가 동사적 기능을 한다고 인식하여 명사를 쓴 경우
- 2) ‘명사’에 ‘하다’ 접사를 붙여서 동사를 파생시키는 것을 모르기 때문에 ‘하’를 누락시킨 것으로 파생어를 만드는 방법을 모르는 경우
- 3) ‘N+하다’ 동사는 알고 있지만 뒤의 연결어미와 결합시키면서 기본형 ‘다’외에 ‘하다’를 같이 생략하여 활용했다고 볼 수도 있다.

내용적으로 학습자의 오류 원인을 예측해봤을 때, 크게 위의 3가지로 해석할 수 있다. 그러나 우선 원어절과 교정어절상 표면적으로 드러나는 것은 동사를 써야하는데 명사를 썼기 때문에 품사 대치 오류로 우선 처리한다. 즉, 원어절과 교정어절에서 드러나는 차이에 주목하여 품사 오류로 처리한다. 그러나 한편으로 이러한 품사 오류는

조어법과도 긴밀하게 관련된다. 파생과 합성을 하면서 품사가 바뀌게 되기 때문에 파생/합성에는 품사의 의미도 포함되어 있다고 볼 수 있다. 따라서 이러한 오류에 대해서는 품사를 바꾸는 접사에 대한 인식이 없다고 보고 ‘품사 대치’ 오류로 우선 주석한 후, ‘파생/합성’ 오류도 함께 주석한다.

오류 양상[대치] - 오류 층위[품사, 파생/합성] 중복 주석 처리한다.

- 한국 공부가 너무 재미(√재미있)기는 하지만, 단어 위우가기 힘들다.

☞ ‘재미있다’에서 ‘재미’만을 사용하여 활용했다면, 형용사와 동사 품사 대치 오류인 동시에, ‘재미’와 ‘있’을 합성시키지 못한 것으로 보고 오류 층위에서 품사와 합성법을 중복 주석한다. 오류 위치[명사] - 오류 양상[대치] - 오류 층위[품사, 합성법]

- 마찬가지로 지정사 ‘이다’와 파생접미사 ‘하다’가 대치된 경우, 즉 ‘명사(어근)+하다’ 동사를 쓸 자리에 ‘명사+이다’를 쓴 경우, ‘동사/형용사’와 명사 품사 혼동으로 보고 품사 대치 오류로 처리한다. 동시에 접사 ‘하다’를 사용하여 동사를 파생시키는 것을 모른다고 판단하여 오류 층위에 품사와 파생을 중복 주석한다.

<예> 이번 축제에 제수님 추구하고 부활을 위미입니다(√의미합니다).

☞ 동사를 쓸 자리에 ‘체언+이다’의 꼴로 나타냈다. 이러한 경우 품사 오류로 처리하는 동시에 ‘하다’를 사용하여 동사를 파생시키는 것을 모르는 경우라고 해석하여 파생오류도 중복 주석한다.

- ‘외국어’에 파생접사 ‘하다’를 결합하여 동사/형용사를 만들 수 있는데, 외국어 명사만 사용한 경우도 동일하게 품사 오류와 파생 오류로 중복 주석

한다.

<예> 그리고 장애인의 사회참여를 스마트(√스마트한) 사람이 필요합니다.

☞ 외국어의 명사형만 사용한 경우도 오류 양상[대치] - 오류 층위[품사, 파생] 중복 주석 처리한다.

- 품사 대치 오류로 처리할 경우, 원래 뒤에서 연결되어 있던 요소들을 첨가 처리하거나 품사 대치로 인해 따라오는 요소들을 누락으로 처리하지 않는다. 이는 품사를 제대로 인식하지 못한 것이므로, 품사 오류가 발생한 위치와 함께 뒤의 요소는 한 덩어리로 묶어서 처리하고 첨가 또는 누락으로 주석하지 않도록 한다.

<예> 직장을 선택의(√선택하는) 조건들이 많이 있고 사람마다 다른 생각하고 의미도 있는 것 같다.

☞ 동사 ‘선택하다’를 명사 ‘선택’만을 썼을 때, 뒤의 관형격 조사 ‘의’를 쓴 것은 문법적으로 틀리지 않다. 따라서 관형격 조사 ‘의’를 추가로 첨가처리 하지 않는다.

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

☞ 동사 ‘중요하다’를 써야할 자리에 명사 ‘중요’의 오형태 ‘중용’만을 썼기 때문에 명사의 품사 대치 오류로 처리한다. 아울러 ‘중요’도 ‘중용’이라고 썼기 때문에 오형태도 동시 주석한다. [오류 위치-명사], [오류 양상-대치, 오형태], [오류 층위-품사] 오류로 처리한다. 이때, 품사 대치의 경우 뒤에 따라오는 요소는 누락으로 처리하지 않는다. 즉, ‘중요하다’ 뒤에 오는 연결어미 ‘-고’는 누락 처리하지 않는다. 이는 품사를 제대로 사용하지 못한 것으로 인한 것이기 때문에 뒤에 따라오는 활용 형태들, ‘-한, -하게, -다 고, -하니까’ 등 관형사형 전성어미나 연결어미 등은 ‘누

락’ 처리하지 않는다.

(3) 통사

① 높임(SH)

- [정의] 높임 오류는 조사, 선어말어미, 종결어미 부분의 높임 관련 문법 형태소, 높임 어휘의 사용이 잘못된 경우를 말한다.
- [주석 방식] 높임 요소를 잘못 사용하였을 때(예를 들어 주격 조사 ‘가’와 ‘께서’의 교체)에는 오류 양상을 대치(REP)로 주석한다. 높임 요소를 사용해야 하는데 사용하지 않은 경우는 누락(OM)을, 불필요한 높임 요소를 사용했을 때에는 첨가(ADD)로 오류 양상을 주석한다. 총위는 모두 높임(SH)으로 주석한다.
- [처리 기준] 높임법은 주체 높임법과 객체 높임법, 상대 높임법이 있으며, 이는 다시 문법적 높임과 어휘적 높임으로 나뉜다. 이 중 문법적 높임 오류는 주격조사와 서술에서의 높임 호응 관계 불일치, 주체 높임을 나타내는 선어말어미 ‘-시-’의 잘못된 사용, 상대 높임을 나타내는 대명사와 종결어미의 호응 관계 불일치, 객체 높임을 나타내는 부사격조사 ‘께’의 잘못된 사용이 해당되며, 어휘적 높임 오류는 ‘게시다, 드리다, 모시다, 잡수시다, 주무시다’ 등의 특수한 높임말을 써야할 자리에 쓰지 않은 경우 또는 그 반대의 경우가 해당된다.
- 그러나 특수 어휘에 의해서 표현되는 높임법의 경우 존대와 겸양을 나타내는 특수 어휘가 다양하다. 높여야 할 대상인물을 직접적으로 높이는 어휘, 대상과 관계있는 것을 간접적으로 높이는 어휘, 객체를 높이는 동사 어휘, 접미사나 접두사가 붙어 존대나 겸양을 나타내는 어휘 등 다양하기 때문에 높임과 관계된 모든 어휘를 오류 주석 대상으로 삼기 어렵다. 이러한 이유로 본 연구에서는 어휘적 높임 오류보다 문법적 높임 오류를 우선 주석한다.
- 따라서 주된 높임 오류의 대상이 되는 품사는 대명사, 조사(주격조사 ‘께서’, 부사격 조사 ‘께’), 선어말어미, 종결어미이다.
- 높임의 오류에는 높임법을 써야 하는 환경에서 낮춤말을 쓴 경우와 반대로 낮춤말을 써야 하는 환경에서 높임말을 사용한 경우를 모두 포함한다.

<예> 할아버지께서 사 주었어요(√/주셨어요).
 근데 부모님께서 나에게 유학하라고 해서(√/하셔서) 부모님의 말(√/말씀)대로 한국에 와서 유학했다.
 ☞ 주체 높임을 실현시키기 위해 ‘께서’와 ‘-시-’를 실현시켜야 하는데 높임 선어말어미는 실현시키지 않았으므로 높임 오류로 판정한다. 명사 ‘말’의 경우 ‘말씀’의 어휘 높임 대치 오류로 주석한다.

- 한국어의 높임법은 주체 높임과 상대 높임, 특수어휘에 의한 객체 높임 모두 담화층위에서 화청자 관계에 따라 실현되는 것으로 높임 표현 오류의 적절성과 용인 가능성이 다르게 적용될 수 있다. 한국어 모어 화자들도 일상적인 언어생활에서 높임 표현을 잘 지키지 않는 경우가 많으며, 그 사용이 일관적이지 않다. 즉, 높임 표현 오류는 용인 가능성으로 인하여 문법성이나 적합성 기준을 일관적으로 적용하기에 어려움이 따른다. 이러한 점을 고려해 일관된 높임 오류 주석을 위하여 동일 문장 내 높임 표현의 호응을 필수적인 높임 표현의 오류 판정 기준으로 삼았다. 예를 들어, 주격조사 ‘께서’를 사용하였지만 서술어에서는 높임 선어말어미 ‘-시-’를 사용하지 않은 경우와 같이, 문장 내 한 부분이라도 높임 요소를 실현시킨 경우에는 실현시키지 않은 부분을 높임 오류로 판정한다.

<예> 근데 부모님께서 나에게 유학하라고 해서(√/하셔서) 부모님의 말(√/말씀)대로 한국에 와서 유학했다.
 ☞ 주체 높임을 실현시키기 위해 ‘께서’와 ‘-시-’를 실현시켜야 하는데 높임 선어말어미는 실현시키지 않았으므로 높임 오류로 판정한다. 명사 ‘말’의 경우 ‘말씀’의 어휘 높임 대치 오류로 주석한다.

- 상대높임법에서 대명사와 종결어미 양쪽 모두 교정이 가능할 때에는 종결어미를 기준으로 대명사 오류로 일괄 처리한다. 즉, 종결어미에 따라 대명사 ‘나’와 ‘저’의 대치 오류로 우선 처리한다.

<예> 저는(√나는) 10년 후에 생활이 부유하고 싶다.

☞ 상대높임법 체계에서 해라체를 사용했는데, 대명사는 자신을 낮추는 ‘저’를 잘못 사용하였다. 따라서 상대높임법 체계에 맞추어 ‘나’로 교정한다. 반대로, 합쇼체를 사용하고, 대명사 ‘나’를 사용한 경우도 마찬가지로 [오류 위치- 대명사], [오류 양상-대치], [오류 층위-높임]으로 주석한다.

- 학습자가 글이나 발화에서 ‘-다/ㄴ다’와 ‘-ㅂ니다/습니다’ 등 종결어미를 혼용하여 사용한 경우에는 학습자의 종결어미 선택에 일관성이 부족한 것으로 판단하여 더 자주 쓴 종결어미로 교정하고 대치 오류로 주석한다. 그런데 두 종결어미가 대치될 때 이를 높임 오류로 볼 수 있는가에 대해서는 문어와 구어에서 다르게 접근할 수 있다. 문어의 경우 구어에 비해서 학습자가 높임 범주를 인식하지 못하였다고 일괄적으로 판단하기 어려운 측면이 있는 데 반해, 구어에서는 발화 상대를 고려하였을 때 높임에 대한 인식을 고려해야 한다고 보고 말뭉치의 유형에 따라 달리 처리한다.
- 높임을 나타내는 접사 ‘님’의 경우, 생산성이 강한 접사로 형태 주석에서 접미사로 따로 분석 처리한다. 이와 연계하여 오류 주석에서는 높임을 나타내는 접사 ‘님’의 과잉사용 또는 미사용의 경우 ‘첨가’와 ‘누락’ 오류로 처리한다.

<예> 의사님(√의사), 책상님(√책상), 선생(√선생님)

☞ ‘선생님’을 ‘선생’으로 쓴 경우, 어휘 대치로 볼 수도 있으나, 접사 ‘님’ 누락으로 처리한다. 왜냐하면 ‘의사님’, ‘책상님’ 등으로 쓴 경우는 존대를 과잉생산한 것으로서 ‘첨가’로 처리해야 해야 하는데, 동일 요소에 대하여 다르게 처리하게 되는 문제가 있다. 따라서 ‘님’과 같이 존대를 나타내는 접사는 일괄 ‘누락/첨가’로 처리한다. [오류 위치- 접사], [오류 양상-누락/첨가], [오류 층위-높임]으로 주석한다.

② 시제(ST)

- [정의] 시제 오류는 시제 또는 시상을 나타내기 위한 선어말어미, 관형사

형 전성어미, ‘-(으)ㄴ 것’ 등의 오류를 말한다.

<예> 옛날에 한국의 정통 난방법을 온돌이다(√온돌이었다).
 때로는 한국말을 공부할 때 끝이 없는 것 같은데(√같았는데)
 데) 오느날 갑자기 끝에 왔다.

- [주석 방식] 시제 오류는 오류 양상을 두 가지로 처리한다. 하나는 시제 요소를 사용하지 않고 기본형을 사용했을 경우, 시제를 사용하지 않았다고 판단하여 누락(OM) 오류로 주석한다. 다른 하나는 시제를 제대로 인식하지 못하여 과거 시제나 미래 시제 자리에 현재 시제를 사용한 경우, 그 반대의 경우 등은 대치(REP) 오류로 처리한다. 즉, 시제를 사용했으나 현재와 과거, 과거와 미래처럼 잘못 사용한 경우는 시제 간 대치로 처리한다. 양상은 달라질 수 있지만 층위에는 모두 시제(ST)를 주석한다.
- 이때, ‘이다/아니다’, 형용사, 연결어미 앞에서의 용언은 기본형이 현재를 나타내기 때문에 기본형을 사용했을 경우, 현재 시제로 인식하고 대치(REP)로 주석한다.

<예> 먹다(√먹었다) ⇨ 누락
 먹다(√먹는다) ⇨ 종결어미 오형태 활용
 먹는다(√먹었다) ⇨ 대치
 습니다(√었습니다) ⇨ 대치
 있다(√있었다) ⇨ 대치
 예쁘다(√예뻤다) ⇨ 대치
 이다(√이었다) ⇨ 대치
 아니다(√아니었다) ⇨ 대치

⇨ 형용사와 ‘이다/아니다’는 기본형과 현재가 같으므로 기본형을 현재형으로 간주하고 대치로 처리한다. 단, 예쁘다(√예뻤다)와 같은 경우는 오형태 활용 오류에 해당한다.

밥을 먹지만(√먹었지만),
 밥을 먹고(√먹었고)

⇨ 연결어미에서도 현재형으로 보고, 앞에 선어말어미 ‘-았-’이 와야 하는데 사용하지 않은 경우 [오류 위치-선어말어미], [오류 양상-대치], [오류 층위-시제] 오류로 처리한다.

- 시제 선어말 어미 ‘-었-’과 ‘-겠-’이 생략된 경우, 기본형을 제외하고는 대치 오류로 처리한다. 그러나 시제 선어말 어미의 문법적 제약이 있는 연결어미 앞에서 ‘-었-’과 ‘-겠-’을 사용한 경우는 첨가(ADD) 오류로 처리한다.

<예> 아침에 밥을 먹을 때 해물을 먹다(√먹었다).

☞ 기본형 ‘먹다’를 사용해서 선어말 어미 ‘-었-’ 누락 오류로 처리한다.

고향에서 향주까지 4시간 걸렸서(√걸려서) 좀 피곤했다.

몇 년 전에 영국에서 임신분에게 동물 실험을 했던 약을 주었다 보니(√주다 보니)

☞ 연결어미 ‘어서’ 또는 ‘-다 보니’ 앞에 과거시제 선어말 어미 ‘-었-’이 올 수 없으나 ‘-었-’을 사용했다. 이처럼 ‘-었-’과 ‘-겠-’을 사용할 수 없는 문법적 제약이 있는 자리에 사용한 경우는 선어말 어미 첨가(ADD) 처리한다.

- ‘-겠-’의 경우, 미래보다는 추측이나 의지와 같은 양태 의미를 나타내는 경우가 많다. 본 연구에서는 ‘-겠-’이 양태 의미를 나타내는 경우에는 시제를 주석하지 않는다.

<예> 앞으로 학교 규칙을 안 어긴다(√어기겠다).

☞ 의지를 나타내야 하는데 현재형 ‘ㄴ다’로 잘못 사용한 것으로, 선어말 어미 ‘-겠-’ 대치 오류로 처리한다.

사람들이 항상 물건을 어떻게 선택할지 모르겠다(√모른다).

☞ ‘모르겠다’의 경우 ‘-겠-’이 시제의 의미를 나타낸다고 보기 어렵다. 따라서 이때의 ‘-겠-’은 오류 층위에 시제를 처리하지 않는다.

- 관형사형 전성어미 ‘-(으)ㄴ, 는, (으)ㄹ’의 경우, 뒤에 오는 (의존)명사에 따라 시제를 나타내는 경우가 있고 그렇지 않은 경우가 있다. 시제를 나타내지 않을 경우와 시제로 인한 오류일 경우를 구분하여 처리한다. 예를

들어, 시제를 나타내지 않는 경우로 ‘-(으)ㄴ 때, -(으)ㄴ 따름이다, -(으)ㄴ/는 편이다, -(으)ㄴ 후’의 구성 등이 있다. 이때의 관형사형 전성어미는 특정한 시제의 의미가 없기 때문에 시제 오류로 처리하지 않도록 주의한다.

<예> 가(√가는/√갈) 사람

☞ ‘가는 사람’ 또는 ‘갈 사람’을 써야하는데, ‘가 사람’으로 쓴 경우는 관형사형 전성어미 [누락]으로 처리한다. 그러나 ‘가는 사람’을 ‘갈 사람’으로 썼을 경우는 [오류 위치-관형사형 전성어미], [오류 양상-대치], [오류 층위-시제] 오류로 처리한다.

과식이나 하지 말고 여러까지 음식을 골고루 먹을(√먹는) 것이 중요해요.

☞ 이때의 ‘-(으)ㄴ 것’에서 관형사형 전성어미는 시제를 나타낸다고 보기 어렵다. 따라서 관형사형 전성어미 대치 오류로 처리하고 오류 층위에 시제는 주석하지 않는다.

- 관형사형 전성어미에서 시제 대치 오류 판단에 어려움이 있을 경우, 현재(‘-는’)와 미래(‘-(으)ㄴ’)가 둘 다 가능할 때에는 용인 가능한 것으로 판단하여 오류로 처리하지 않고, 명확하게 과거형을 써야 하는데 쓰지 않은 경우나 반대의 경우 시제 오류로 처리한다.
- 연결어미 ‘-(으)ㄴ지/는지/-(으)ㄴ지’는 시제 대치 오류로 처리하지 않도록 주의한다.

<예> 10년 후에 어느 나라에 살고 있는지(√있을지) 잘 모르는데 그때는 좋은 일이 있었으면 좋겠다.

☞ 연결어미 ‘-(으)ㄴ지/는지/-(으)ㄴ지’의 대치의 경우, 오류 층위에 시제를 주석하지 않도록 주의한다.

③ 사동(SC)

- [정의] 사동 오류는 사동사, 사동 표현의 사용, 사동문 생성에서 발생한 오류를 말한다.

- [주석 방식] 사동사, 사동 표현을 사용해야 하는데 사용하지 않은 경우는 기본적으로 오류 양상을 대치(REP) 오류로 주석하고, 층위에는 사동(SC)을 주석한다.
- [처리 기준] 사동 표현은 접미사 ‘-이/히/리/기/우/구/추-’에 의한 사동, ‘-게 하다’에 의한 통사적 사동, ‘-내다, 만들다, 시키다’ 등 어휘적 사동으로 나타낼 수 있다. 이 연구에서는 접미사 ‘-이/히/리/기/우/구/추-’에 의한 사동사와 ‘-게 하다’, ‘시키다’ 사동 표현에 의한 사동으로 제한한다. ‘내다, 만들다, 시키다’ 중 ‘시키다’는 한국어 교육에서 ‘사동’을 나타내는 표현으로 교수하고 있는 상황을 고려하여 포함하나 나머지 어휘에 의한 사동은 맥락에 따라 다르게 처리될 수 있기 때문에 주석자 간 일관성을 유지하기 위해 제외한다.
- ‘형용사 - 게 하다’의 경우, ‘사동’으로 처리하지 않도록 주의한다. 예) 방을 깨끗하게 해야 한다.

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

☞ ‘감소하다’와 ‘감소시키다’의 사동 대치로 처리한다. 또한 ‘-게 하다’의 사동 표현도 대치 오류로 처리한다.

전통의 아름다움이 사람들에게 알려주는(√알려주는) 것도 전통을 보존하려고 해야 할 일이다.

☞ ‘알려주는’의 경우, 사동접미사 ‘리’를 인식하였으나 형태를 잘못 사용한 것으로 보고, 이러한 경우에는 오형태 오류로 처리한다.

- 사동사, 사동 표현에서 나타난 오류를 모두 볼 수 있도록 오류 양상에 관계없이, 철자를 잘못 사용한 오형태 오류도 오류 층위에서 사동으로 주석한다.
- 원어절에서 사동사, 사동 표현을 사용한 경우와 교정 어절이 사동사, 사동 표현이어야 하는 경우 모두 오류 층위에서 사동으로 주석한다.
- 사동을 쓸 자리에 피동을 썼거나 반대의 경우는 사동과 피동으로 중복 주석한다. 오류 위치와 오류 양상은 원어절 기준으로 주석하지만 오류 층위는 원어절과 교정어절 양쪽에서 주석함에 따라 사동과 피동을 중복 주석

한다.

- 사동 표현 ‘-게 하다’와 일반 사동사가 대치된 경우, 오류 위치는 형태 주석에 따라 일관되게 처리한다. 동사 또는 ‘연결어미+보조용언’으로 분리되어 처리되었을 경우는 각각의 품사로 오류 위치를 주석하며, 사동 표현의 경우 표현 문형 목록에 해당되기 때문에 표현문형(PE)도 중복 주석한다.

<예> 더 간단하게 하려고 생각하면 에어컨의 온도를 조금만 높게 하는(√높이는) 것만 한 방법이 없다.

☞ ‘-게 하다’가 사동사로 대치된 경우로 오류 위치는 ‘연결어미, 보조용언, 표현 문형’으로 주석하며, 오류 양상은 대치로 주석한다.

④ 피동(SP)

- [정의] 피동사, 피동 표현의 사용, 피동문 생성에서 발생한 오류를 말한다.
- [주석 방식] 피동사, 피동 표현을 사용해야 하는데 사용하지 않은 경우는 기본적으로 대치(REP) 오류로 주석하고, 층위에는 피동(SP)을 주석한다.
- [처리 기준] 피동 표현은 접미사 ‘-이/히/리/기-’에 의한 피동, ‘-아/어지다, -게 되다’에 의한 통사적 피동, ‘-되다, 받다, 당하다’ 등 어휘적 피동으로 나타낼 수 있다. 이 연구에서는 형태를 중심으로 접미사 ‘-이/히/리/기-’에 의한 피동사와 ‘-어지다’ 피동 표현에 의한 피동으로 제한한다.
- 단, 통사적 피동 ‘-아/어지다’의 경우, ‘형용사+아/어지다’는 피동보다는 상태변화의 의미를 나타낸다고 보고 ‘피동’으로 주석하지 않는다. 상태변화가 일어나게 된 요인이 타의에 의해 발생하여 피동의 의미가 내포되어 있더라도 한국어 교재 및 학습 기관에서 피동과 상태변화를 분리하여 교수하고 있으며, 맥락에 따라 피동과 상태변화를 구분하여 주석할 경우, 주석자 간 일관성이 떨어질 수 있기 때문에 ‘형용사+아/어지다’는 일괄적으로 오류 층위에서 피동으로 주석하지 않는다.
- ‘-게 되다’의 경우도 변화의 의미를 나타내는 경우가 많으며, 학교문법에서 피동에 포함시키지 않는 논의에 근거해 본 연구에서도 제외하였다.
- 즉, ‘형용사+아/어지다’와 ‘-게 되다’는 기본 의미를 변화로 보고, 피동으로 다루지 않으며, 맥락에 따라 다르게 판단할 수 있는 어휘적 피동도 제외한다.

<예> 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. 오류 층위는 원어절과 교정 어절 양쪽 모두를 기준으로 하기 때문에 이때에는 오류 층위에 사동(SC)과 피동(SP)을 중복 주석한다.

- 조사와 용언 교정이 모두 가능한 경우, 격조사 오류를 우선적으로 처리하나, 문맥에 따라 양쪽을 모두 바꿔야 하는 경우는 양쪽 모두 오류 주석한다. 특히, 피동문에서 용언을 교정하여 바뀌게 되는 조사의 경우, 교정어절만 써주고 오류로 처리하지 않았으나 피동/사동 구조를 모르고 있다는 측면에서 오류 층위에서 사동/피동을 주석하는 것으로 수정하여 처리한다.

<예> 둘째, 의학 기술의 발전에 따라 수명을(√수명이) 연장하지만(√연장되지만) 노인층 증가도 할 수 있다.
 ☞ ‘되다’의 피동으로 용언을 교정함에 따라 조사도 바뀌게 된다. 이 경우는 양쪽 모두 오류로 처리하고 조사에도 오류 층위에 ‘피동’을 주석한다.

- 피동 오류를 처리하는 데 있어, ‘동사+아/어지다’의 경우, <표준국어대사전>에 한 단어로 등재되어 있는 동사가 있는 반면, 등재되지 않은 단어가 있다. 이 경우, 형태소 분석에서 사전에 등재되어 있는 단어는 동사로 주석하고, 그렇지 않은 경우는 ‘연결어미+보조 용언’으로 분리하여 주석한다. 따라서 오류 주석에서 오류 위치는 형태 분석에 따라 처리하므로, 동사로 분석했을 때에는 그 품사를 따르고, 연결어미, 보조 용언으로 분리하였을 때에는 분리된 형태대로 오류 위치를 주석한다.
- 이중피동을 사용한 경우, 첨가 오류로 처리한다.

<예> 환경 오염이 심해지게 되고(√심해지고) 있지만 더 이상 심하기 전에 여러분의 도움이 필요하다고 생각한다.

☞ ‘심해지다’에 ‘-게 되다’까지 첨가된 이중 피동표현으로 ‘-게 되다’의 ‘연결어미+보조용언’, 표현문형(PE)의 첨가 오류로 처리한다.

⑤ 부정(SN)

- [정의] 부정 표현의 사용, 부정문의 생성에서 발생한 오류를 말한다.

<예> 한국에서 혼자서 살다가보니 외로울 때가 많이 있으니까 그냥 혼자 있지 말고(√않고) 친구들이랑 같이 공부를 해요.

- [주석 방식] 일반적으로 부사 ‘아니(안), 못’이나 부정의 의미를 가진 용언 ‘아니다, 아니하다(않다), 못하다, 말다’를 써서 부정문을 만드는 방법에 근거하여 부정 부사를 잘못 사용하거나 해당 용언에서 오류가 났을 경우, 오류 층위에 부정(SN)을 주석한다.
- [처리 기준] ‘없다, 모르다’, 부정 의미의 접두사는 부정 오류에 포함하지 않는다.
- 장형부정인 ‘-지 않다’, ‘-지 못하다’, ‘-지 말다’의 경우, 표현문형 목록에 해당되기 때문에 오류 위치는 보조용언과 표현문형을 중복 주석한다. (※ ‘-고(야) 말다’는 부정의 의미를 나타내는 것이 아니기 때문에 부정으로 처리하지 않도록 주의한다.)
- 장형부정이 더 자연스럽지만 단형부정을 썼을 때 용인가능하기도 하다. 따라서 단형부정을 장형부정으로 반드시 바꿔야하는 경우 기준 마련이 필요한데, 합성어나 파생어의 경우 단형부정문을 만들지 않으며, 용언의 음절이 긴 경우에도 단형부정을 허용하지 않기 때문에 이에 해당하는 용언의 경우는 장형부정으로 교정하고, 나머지의 경우 단형부정의 용인가능성을 인정하도록 한다. 단, 단형부정의 용인가능성의 경계가 분명하지 않으므로 주석자의 판단에 따라 적절하지 않다고 판단하여 장형부정으로 교정했을 경우는 적절성의 오류도 포함한 것으로 한다.

<예> 하지만 한국어는 안(√ 못) 잘합니다(√ 합니다). 그래서 한국 친구가 아직 없습니다.

☞ 능력을 부정하는 경우, 부정부사 ‘못’을 사용해야 하는데, ‘안’을 썼으므로 ‘안’을 ‘못’으로 교정하고 대치 오류로 처리한다. [오류 위치-일반부사], [오류 양상-대치], [오류 층위-부정]. 또한 ‘안’과 ‘못’ 부정의 경우, 서술어 ‘잘합니다’도 ‘합니다’로 교정하고 대치 오류로 처리한다.

- ‘-하다’ 파생동사들의 경우는 체언과 ‘-하다’가 분리될 때 ‘하다’ 앞에 아니(안)를 넣어 단형부정문을 만들 수 있다. 따라서 ‘-하다’ 파생동사 앞에 부정부사를 쓴 경우는 오류로 처리하고, 이때는 어순 오류와 부정 오류로 중복 주석한다.

<예> 스페인어 안 사용해서(√ 사용 안 해서) 스페인어만 말하기 저금 어렵습니다.

☞ ‘N+하다’ 파생동사 앞에 부정부사 ‘안’을 사용한 경우로, 이때에는 ‘사용 안 해서’와 ‘사용하지 않아서’ 두 가지로 교정이 가능하다. 그러나 이 경우, 최소 수정 원칙에 의해서 단형부정을 장형부정으로 바꾸는 것보다 단형부정의 위치를 잘못 사용한 것으로 보고 오류 층위에서 어순 오류와 부정 오류로 중복 주석한다.

⑥ 어순(WO)

- [정의] 어순 오류는 한국어의 통사 구조에 맞지 않는 방식으로 문장 전체 또는 일부가 배열된 경우를 말한다.

<예> 그래서 잘 아직까지(√ 아직까지 잘) 몰라요.
저녁까지 많이 이야기도(√ 이야기도 많이) 합니다.

- [주석 방식] 어순 오류는 오류 위치와 오류 층위만 주석하고, 오류 양상은 주석하지 않는다. 또한 교정어절을 줄 필요가 없어 교정된 어순을 반영하여 앞이나 뒤에 추가하지 않는다. 아울러, 2개의 문장 성분이 상호 교체될

때에는 2개 모두 대치 어순(WO) 오류로 주석한다.

<예> 저는 많이 여행을 (√ 많이) 가고 싶습니다.
저는 한국 여행에서 자주 서울만 (√ 자주) 갔습니다.
☞ 일반적으로 성분 부사는 서술어 앞에서 수식해야 하는데,
이처럼 ‘많이’, ‘자주’가 명사 앞에 온 경우, [오류 위치-일반부사], [오류 양상-없음(빈칸)], [오류 층위-어순] 오류로 처리한다.

- [쟁점] 한국어 어순의 특징 중 하나는 문장성분의 자리 이동이 비교적 자유롭다는 것이다. 그렇기 때문에 그만큼 용인가능성이 크다고 할 수 있다. 이에 따라 주석자 간의 일치도도 다르게 나타날 수 있어, 어순 오류의 경우 최소 수정의 기준을 마련하여 처리한다.
- [처리 기준] 문장 부사는 자리 이동이 자유롭지만 성분 부사의 경우는 제한되기 때 특정한 성분을 수식해야 하는 성분부사의 위치를 잘못 사용했을 때는 오류로 주석한다.
- 관형사의 경우, ‘지시관형사-수관형사-성상관형사’ 순의 기준을 적용하여 처리한다.
- 어순 오류는 오류 위치와 오류 층위만 주석하고, 오류 양상은 주석하지 않는다. 또한 교정어절을 줄 필요가 없어 교정된 어순을 반영하여 앞이나 뒤에 추가하지 않는다.

<예> 저는 많이 여행을 (√ 많이) 가고 싶습니다.
저는 한국 여행에서 자주 서울만 (√ 자주) 갔습니다.
☞ 일반적으로 성분 부사는 서술어 앞에서 수식해야 하는데,
이처럼 ‘많이’, ‘자주’가 명사 앞에 온 경우, [오류 위치-일반부사], [오류 양상-없음(빈칸)], [오류 층위-어순] 오류로 처리한다.

- 시간을 나타내는 표현의 배열이 잘못되었을 경우, 어순 오류로 처리한다. 시간을 나타내는 표현은 ‘년도-월-일-오전/오후/밤/낮/아침/점심/저녁-시-분-초’의 순서로 배열되는 것이 일반적으로 이를 기준으로 시간의 배열 어순 오류를 판단하여 처리한다.

<예> 8반 시(√8시 반)에 학교에 가서 가요

☞ 시간표현에서 시보다 분을 먼저 배열했기 때문에, ‘반’과 ‘시’의 어순 대치 오류로 주석한다. 이 때, 어순이 상호교체되는 것으로 명사 ‘반’과 의존명사 ‘시’ 모두를 대치 어순 오류로 주석한다.

○ 용인 가능성을 적용하여 다음은 어순 오류로 처리하지 않는다.

<예> 저는 우즈베키스탄에서 한국에 2014년 7월에 왔어요.

☞ 시간부사와 장소부사가 함께 올 때 한국어에서는 보통 시간 부사를 먼저 쓴다. 그리하여 ‘저는 2014년 7월에 우즈베키니스탄에서 한국에 왔어요.’가 더 자연스러울 수 있지만 문법적으로는 용인 가능한 것으로 보고 어순 오류로 처리하지 않는다.

○ 조사의 경우, 어순 오류로 처리하지 않고 조사 첨가 또는 누락 오류로 처리한다.

<예> 그래서 한국말을(√한국말) 공부(√공부를) 참 좋아했습니다.

☞ ‘한국말을’에서의 목적격 조사 ‘을’을 ‘공부’ 뒤로 보내는 어순 조정으로도 교정이 가능하다. 그러나 이때는 조사의 배열 문제라기보다는 조사를 잘못 사용한 것으로 판단하여 어순 오류로 처리하지 않고 앞의 ‘을’ 첨가, 뒤의 ‘을’ 누락 오류로 주석한다.

(4) 담화

① 지시(DR)

○ [정의] 지시 오류는 부적절한 지시사의 선택으로 선행문과 후행문의 관계를 결속성 있게 나타내지 못한 경우를 말한다.

○ [주석 방식] 담화 층위에서의 오류는 의미 대치 오류를 중심으로 처리함

을 원칙으로 한다. 이에 따라, 지시 표현에서 의미 간 대치(REP) 오류를 중심으로 하여 오류 층위에 지시(DR)를 주석한다. 지시 표현에서 나타난 단순 오철자 오류는 오류 양상에 오형태(MIF)만 주석하고, 오류 층위에서 지시를 주석하지 않도록 주의한다.

- [처리 기준] 지시 오류는 앞 뒤 문장과 연결, 상황 맥락을 통해서 오류 판단이 가능하기 때문에 문장 단위를 기본원칙으로 삼으나 지시 오류의 경우는 문장 이상의 단위를 고려해 오류를 판단한다.

<예> 저기에(✓거기에) 가면 좋을 것 같아요.

☞ 맥락상 ‘저기’보다는 ‘거기’가 더 적절한 표현으로, 대명사 대치 오류로 주석한다. 아울러 이는 지시 표현에 해당되므로 오류 층위에서 지시(DR)도 함께 주석하도록 한다. [오류 위치-대명사], [오류 양상-대치], [오류 층위-지시(DR)]로 주석한다.

② 접속(DC)

- [정의] 접속 오류는 선행문과 후행문의 의미 관계를 나타내는 데에 부적절한 접속사를 사용한 경우를 말한다. 접속 부사 및 접속 표지의 오류가 포함된다.
- [주석 방식] 담화 층위에서의 오류는 의미 대치 오류를 중심으로 처리함을 원칙으로 한다. 이에 따라, 접속 표현에서 의미 간 대치 오류를 중심으로 하여, 오류 위치에 접속 부사(CMAJ), 오류 양상에 대치(REP), 오류 층위에 접속(DC)을 주석한다. 접속 표현에서 나타난 단순 오철자 오류는 오류 양상에 오형태(MIF)만 주석하고, 오류 층위에서 접속을 주석하지 않도록 주의한다.
- [처리 기준] 접속 오류는 앞 뒤 문장과 의미적 연결을 통해서 오류 판단이 가능하기 때문에 문장 단위를 기본원칙으로 삼으나 접속 오류의 경우는 문장 이상의 단위를 고려해 오류를 판단한다.

<예> 그래서(✓그러면) 어떻게 해야 전통을 보존할 수 있을까요?

그래서(✓그러니까) 전통을 보존하기 위해 더 많이 노력을 해서 전통은 없어지지 않도록 하세요.

☞ 접속부사 ‘그래서’를 과잉 사용하고 있는 양상으로, 앞뒤 문장을 고려했을 때, 각각 ‘그러면’과 ‘그러니까’가 더 적절하다. 따라서 접속부사의 대치 오류로 주석하여 [오류 위치-접속부사], [오류 양상-대치], [오류 층위-접속(DC)]로 주석한다.

③ 담화표지(DM)

- [정의] 담화표지 오류는 담화표지와 간투사의 오류로, 부적절한 담화 표지를 선택하거나, 잘못된 형태로 이들을 사용한 경우를 말한다.
- [주석 방식] 담화표지에 해당하는 품사의 위치를 오류 위치로 주석하고, 층위에 담화표지((DM)를 주석한다.
- [처리 기준] 담화표지는 미시 담화표지와 거시 담화표지로 나눌 수 있으나 연구자마다 그 정의가 다르고, 해당 형태도 다르기 때문에 담화표지의 목록을 마련하기 쉽지 않다. 이러한 이유로 미시 담화표지에 초점을 두고 오류를 판단하도록 한다. 구어 발화에서 학습자가 L1의 영향으로 인한 간투사 사용과 모어 화자와는 다른 위치에서 담화표지를 사용한 경우를 오류로 주석한다.

<예> 아~ 그럼 제가 음~ 오늘 밤에, 데~(✓에~) 잊어버리지 않으면 추대할게요.

☞ ‘에~’는 간투사로 볼 수 있는데, 이를 ‘데~’로 잘못 발음하고 있어 오형태 오류로 주석하고, 오류 층위에서 담화표지 오류로 처리한다.

- 학습자가 L1을 간투사처럼 사용한 경우 담화 표지를 주석한다. L1을 사용하는 것이 오류라고 할 수는 없지만, 학습자가 구어 발화에서 L1을 어떻게 사용하고 있는지, 발화 전략 등을 파악할 수 있다는 차원에서 이에 대해 담화 표지를 부착한다.

<예> 아노(✓담화 표지)... 음..

☞ 일본인 학습자가 한국어로 발화하기 전, L1의 간투사를 그대로 사용한 경우이다. 감탄사(IC)로 형태 주석하므로 [오류

위치-감탄사], [오류 층위-담화표지]를 주석한다.

④ 구어/문어 오류(DS)

- [정의] 구어체(구어성)/문어체(문어성), 격식체/비격식체의 혼용에 의해 담화 맥락에서의 일관성이 떨어지는 경우를 말한다.
- [주석 방식] 문어에서 구어성이 강한 어휘나 구어에서 문어성이 강한 어휘를 사용한 경우 해당하는 품사의 위치를 오류 위치로 주석하고, 오류 양상은 대치(REP), 오류 층위에 구어/문어 오류(DS)를 주석한다.
- [처리 기준] 구어/문어 오류는 상황에 따라 용인 가능성을 적용할 수 있기 때문에 엄격하게 그 기준을 적용하기가 어렵다. 이에 구어체(구어성)/문어체(문어성)를 판단하는 기준은 <표준국어대사전>으로 삼는다. <표준국어대사전>에서 ‘문어적 표현’이라고 기술되어 있을 경우 ‘문어체’로 보고, ‘구어적 표현’이라고 기술되어 있을 경우 ‘구어체’로 판단한다. 따라서 문어에서 ‘구어체’를 사용한 경우, 구어에서 ‘문어체’를 사용한 경우에는 담화층위에서 구어/문어 오류로 주석한다.
- 단, <표준국어대사전>에는 기술되지 않았지만 문어에서 구어성이 강한 표현이거나 구어에서 문어성이 강한 표현일 경우에는 ‘용인가능성’ 기준을 적용하여 주석자간 논의 후 처리하고, 처리한 것을 검토하여 다시 목록화하는 방향으로 오류를 주석한다.

<예> 근데(√그런데) 특별한 명절이 있다.

☞ ‘근데’는 일반적으로 구어에서 자주 사용하는 접속부사로, 문어에서는 ‘그런데’를 사용하는 것이 더 자연스럽다. ‘근데’는 구어체라고 보고 문어에서 사용했을 경우, 구어/문어 오류로 처리한다.

이거(√이것은) 내 꿈이다.

☞ 해라체를 사용한 문어 텍스트에서 조사를 동반하지 않은 구어형 ‘이거’가 사용되었으므로 담화 층위에서 다른 문장과 어울리지 않으므로 구어/문어 오류로 처리한다.

- ‘하고’, ‘한테’ 등을 구어적 표현으로 보고, 문어에서 사용했을 경우 오류로

주석한다. 구어/문어 오류에 해당하는 목록은 다음과 같다.

<예> 한테(√ 에게)
 하고(√ 와/과)
 거/계(√ 것/것이)
 아무거(√ 아무것)
 근데(√ 그런데)
☞ 위의 예시들을 문어에서 사용한 경우, 오류 위치에 해당 품사를 주석하고, [오류 양상-대치], [오류 층위-구어/문어 오류]로 처리한다.

5. 구어 오류 주석

- 구어 자료의 경우, 문장으로 파악하지 않고 억양 단위로 끊어서 각 단위를 기준으로 오류를 식별하고 판정한다.

<예> 무슨 파티하면
 우리 학생들이.
 열심히 공부한=
 연세대학교 열심히 공부해서
 조금 피곤한,
 =것이에요.
☞ 이 경우 억양 단위로 끊어서 보면 크게 문제가 되지 않지만, 문장 단위로 보면 여러 가지 층위에서 오류 처리가 가능하며 일관된 기준에 의한 처리가 어렵다. 구어 자료는 문장 단위가 아닌 억양 단위를 기준으로 하여 오류를 식별하고 판정한다.

- 말더듬거림은 오류로 처리하지 않는다. 다만, 전사 단계에서 특정 표시를 하므로 이를 통해 향후 검색이 가능하게 한다.
- 자기 수정 발화의 경우, 수정 전 앞부분의 발화는 오류로 주석하지 않는다

다. 수정 후 발화에 초점을 두고 오류 여부를 판정한다.

<예> 친= 친구가 한국 음식이:: = 음식을 좋아해서

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

- [주석 방식] 구어 오류 주석과 문어 오류 주석의 기본 원칙 및 처리 방법
은 동일하다. 그러나 발화 상에서 나타나는 발음 오류의 경우에는 오류
양상(대치, 누락, 첨가, 오형태)을 주석하지 않고, 오류 위치와 오류 층위
[발음]만 주석한다.
- [처리 기준] 구어에서는 발음 오류와 어휘 및 문법 오류의 구분이 명확하
지 않을 수 있다. 즉, 학습자가 어휘와 문법을 잘못 사용한 것인지 단순히
발음을 잘못된 것으로 인해 나타난 오류인지 판별하는 데 어려움이 있다.
구어에서는 발음의 영향과 함께 어휘 및 문법 오류를 표시해주는 차원에
서 오류 층위에서 중복 주석을 한다. 그러나 조사의 경우, 문법 오류를 우
선 처리하도록 하고, 관형사형 전성어미를 사용해야 할 자리에 사용하지
못했을 경우, 문법 오류로 우선 처리한다. 문어와 마찬가지로 문법 오류를
우선 처리하도록 하되, 둘을 구분해야 할 때에는 학습자의 발음 양상을
살핀다. 만약, 학습자가 반복적으로 특정 음소를 다른 음소로 발음하거나
받침을 실현시키지 못하는 경우는 문법 누락 오류와 발음 오류를 구분하
여 처리한다.

<예> 가(√갈) 수 있는 방법으(√을) 모라서(√몰라서)

☞ 원칙적으로 관형사형 전성어미를 사용해야 할 자리에 사
용하지 못했을 경우, 문법 오류로 보고 누락 오류로 처리
한다. 그러나 예시처럼 학습자가 관형사형 전성어미 ‘-ㄴ’
뿐만 아니라 목적격 조사 ‘을’과 동사 ‘몰라서’에서 반복적
으로 ‘ㄴ’ 받침을 제대로 발음하지 못하는 양상을 보이면
누락이 아닌, 음소 오류로 처리한다. 이는 발음 층위의 문
제이기 때문이다.

- 발음 오류는 문법뿐 아니라 어휘 오류와의 구분도 모호할 때가 있다. 구어 발음 오류와 어휘 형태 오류의 구분이 어려울 때에는 학습자가 형태를 제대로 모르는 것인지 발음의 문제인지를 판단한다. 구어에서 발음 차원이 아닌, 형태를 잘못 발화한 경우는 오형태(MIF) 오류로 처리한다.

<예> 사잉(√/사건)

보석필(√/보살핌)

그 부모들이 그 고절(√/걱정)이 많이 되어서

아무리 보다도 요즘 여성들이 사회 화풍 더 많이 참가하고

그 자아시선(√/자아실현) 이런 거도 어 요구 많이 생겨서

☞ 발음의 위치를 고려해 보면 동일 조음 기관이나 유사한 부분에서 발음되었다고 보기 힘들다. 이러한 형태들은 음소의 변별보다는 단어의 형태를 잘못 알고 있거나 형태를 잘못 만들어낸 것이다. 유사 발음과도 떨어져 음소 오류로 볼 수 없고, 한국어에 없는 형태들을 발음했다고 판단하여 오형태 오류로 처리한다.

- 구어 오류 주석에서의 또 다른 쟁점은 구어의 특성으로 볼 수 있는 현실 발음과 준말을 오류로 처리해야 하는가이다. 예를 들어, ‘김밥[김밥]’으로 발음했을 때 ‘적절성’을 기준으로 하여 ‘한국어 모어 화자’와 다르게 발음한다는 차원에서 오류로 볼 수도 있을 것이다. 그러나 현실 발음은 그 기준을 확정하기가 쉽지 않다. 한국어 화자도 표준 발음으로 발음하는 경우가 있고, 현실 발음을 어느 범위까지 인정해야 하는지도 문제가 되기 때문에 본 연구에서는 현실 발음에 어긋난다고 해서 오류로 처리하지는 않는다. 반대로 현실 발음을 인정해 일반적으로 한국어 모어 화자에서도 많이 나타나는 발음일 경우에 오류로 처리하지 않는다.
- 즉, 구어에서는 한국어 모어 화자들의 현실 발음을 고려하여 일반적으로 많이 사용되며, 구어에서 허용되는 형태는 오류로 처리하지 않는다.

<예> 할려고(√/하려고)

☞ ‘ㄹ’로 시작하는 단어 앞에 받침 ‘ㄹ’을 첨가하여 발음하는

것은 한국인 모어 화자에게서도 많이 나타나는 현상이다.
현실 발음을 고려하여 이러한 경우는 오류로 처리하지 않는다.

[그리구](√그리고)::, 음::

☞ ‘그리고’를 [그리구]라고 발음하는 것은 한국어 모어 화자들에게도 많이 나타난다. 이처럼 한국어 모어 화자들의 현실 발음을 고려하여, 쫌, [바래요](바라요) 등 구어에서 허용되는 발음은 오류로 처리하지 않는다.

○ 구어의 특성이나 표현 의도에 의한 발음 특성은 오류로 처리하지 않는다.

<예> 표현 의도에 의한 수의적 경음화: 쪼금
구어에 의한 발음 특성: ~먹었구요 / ~했어여

○ 구어에서는 준말이 용인가능하기 때문에 오류로 보기 어려운 측면이 있다. 그러나 모든 준말을 허용할 수 있는 것은 아니기 때문에 구어에서 준말의 오류 판단 기준이 필요하다. 이에 따라 본 연구에서는 <표준국어대사전>에서 ‘~의 준말’로 등재되어 있는 것을 기준으로 삼는다. <표준>을 기준으로 ‘준말’로 등재되어 있는 형태는 오류로 처리하지 않고, 등재되지 않은 형태는 오류로 처리한다.

<예> 그래서 맘 먹고 여기 왔어요.

☞ ‘맘’의 경우, <표준>에 ‘마음의 준말’로 등재되어 있다. 따라서 오류로 처리하지 않는다.

[그쵸](√그렇죠)::, 음::

[글구](√그리고)

☞ 그러나 <표준>에 등재되어 있지 않더라도, ‘그쵸’와 같이 구어에서 축약된 형태로 많이 나타나는 용례들은 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, 엄격하게 처리하지 않도록 한다.

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

- 구어 전사 시, 분명하게 들리지 않아서 <X X>로 처리한 부분은 분석불능(IMP)으로 주석한다.

<예> 그:: 마약::, <X청국::죄::X>라는 의미는, 한국에:: 마약, 없는, 뜻입니다. 한국에서:: 그:: 마약은 불법이라서,
그리고::, 어:: 그 뒤,에는 그 <X흔들이(흔들)X>라는 단어도.. 있어서, 그::
☞ <X X> 부분은 전사가 정확하게 듣지 못한 부분을 전사한 것으로 판단이 불분명하기 때문에 분석불능으로 처리한다.

- 구어 오류는 오류 층위에서 발음 오류와 가장 밀접하다. 오류 층위 [발음]에는 음소(PP), 음절(PS), 음운규칙(PC), 원어식 발음(PN), 중간 발음(변이음 포함(PA) 총 5가지가 있는데, 그중 ‘음소, 음절, 음운규칙’ 3가지를 우선적으로 주석한다. 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 주석한다. 변이음 식별이 분명하지 않은 경우에는 주석하지 않는다.(☞ 세부 처리 방법은 ‘3. 범주별 세부 오류 유형의 처리, 4) 오류 층위, (1) 발음’을 참고한다.)

<예> [보롱](√복용),하는 뜻이느(√뜻은)::, 그 마약을 쓰는 아니면
마약을, 하,느:: 것입니다.

☞ 복용을 [보공]으로 발음해야 하나 [보롱]으로 발음하였으
므로 음소 오류로 처리한다. 오류 층위[발음]에 해당하는
오류이기 때문에 오류 양상은 주석하지 않고, 오류 위치
와 오류 층위만 주석한다. 따라서 [오류 위치-명사], [오
류 양상-없음(빈칸)], [오류 층위-음소]로 주석한다.
또한 ‘뜻이느’은 ‘뜻은’에 조사 ‘이’를 첨가한 것이기 때문
에 발음의 오류가 아닌, 문어와 마찬가지로 조사 첨가 오
류로 처리한다. 이 경우, [오류 위치-주격조사], [오류 양
상-첨가]로 주석한다.

<부록> 표현 문형 목록

표제어	형태 정보	대표형
-게 되다		-게 되다
-게 마련이다		-게 마련이다
-게 만들다		-게 만들다
-게 생겼다		-게 생겼다
-게 하다		-게 하다
-고 나다		-고 나다
-고 들다		-고 들다
-고 말다		-고 말다
-고 보다		-고 보다
-고 싶다		-고 싶다
-고 싶어 하다		-고 싶어 하다
-고 있다		-고 있다
-고 해서		-고 해서
-고는 하다		-고는 하다
-곤 하다		-곤 하다
-기 나름이다		-기 나름이다
-기 때문		-기 때문
-기 마련이다		-기 마련이다
-기 십상이다		-기 십상이다
-기 위한		-기 위한
-기 위해(서)		-기 위해(서)
-기 일쑤이다		-기 일쑤이다
-기 전에		-기 전에
-기 짝이 없다		-기 짝이 없다
-기가 무섭게		-기가 무섭게
-기가 바쁘게		-기가 바쁘게
-기가 쉽다		-기가 쉽다
-기나 하다		-기나 하다
-기로 들다		-기로 들다
-기로 하다		-기로 하다

표제어	형태 정보	대표형
-기만 하다		-기만 하다
-기에 따라		-기에 따라
-기에 앞서(서)		-기에 앞서(서)
-ㄴ 것		
-ㄴ 것 같다		
-ㄴ 결과		
-ㄴ 김에		
-ㄴ 나머지		
-ㄴ 대로1		
-ㄴ 대로2		
-ㄴ 대신에		
-ㄴ 데요		
-ㄴ 듯		
-ㄴ 듯하다		
-ㄴ 마당에		
-ㄴ 모양이다		
-ㄴ 법이다		
-ㄴ 이상		
-ㄴ 줄		
-ㄴ 지2		
-ㄴ 채로		
-ㄴ 척하다		
-ㄴ 탓		
-ㄴ 편이다		
-ㄴ 후에		
-ㄴ가 보다		
-ㄴ다는 것이		
-ㄴ 데도 불구하고		
-나 보다		-나 보다
-나 싶다		-나 싶다
-는 가운데		-는 가운데

표제어	형태 정보	대표형
-는 것		-는 것
-는 것 같다		-는 것 같다
는 고사하고	은 고사하고	
-는 길에		-는 길에
-는 김에		-는 김에
-는 대로	-은 대로, -ㄴ 대로	
-는 대신에	-은 대신에, -ㄴ 대신에	
-는 덕분에/이다		
-는 데다가	-은 데다가, -ㄴ 데다가	
-는 도중에		
-는 동시에		-는 동시에
-는 동안		-는 동안
-는 등 마는 등		-는 등 마는 등
-는 듯	-은 듯, -ㄴ 듯	
-는 듯하다	-은 듯하다, -ㄴ 듯하다	
-는 마당에	-은 마당에, -ㄴ 마당에	
-는 만큼	-은 만큼, -ㄴ 만큼	
는 말할 것도 없고		
-는 모양이다	-은 모양이다, -ㄴ 모양이다	
는 물론	-은 물론	
-는 바람에		-는 바람에
-는 반면에	-은 반면에, -ㄴ 반면에	
-는 법이다	-은 법이다, -ㄴ 법이다	
-는 사이		-는 사이
-는 수밖에 없다		
-는 이상	-은 이상, -ㄴ 이상	
-는 적이 있다/없다		-는 적이 있다/없다

표제어	형태 정보	대표형
-는 줄	-은 줄, -ㄴ 줄	
-는 중이다		-는 중이다
-는 척하다	-은 척하다, -ㄴ 척하다	
-는 채하다		
-는 탓	-은 탓, -ㄴ 탓	
-는 통에		-는 통에
-는 편이다	-은 편이다, -ㄴ 편이다	
-는 한		-는 한
-는 한이 있어도/있더라도		-는 한이 있어도/있더라도
-는 한편		-는 한편
-는가 보다	-은가 보다, -ㄴ가 보다	
-는다는 것이	-ㄴ다는 것이	
-는데도 불구하고	-은데도 불구하고, -ㄴ데도 불구하고	
-도록 하다		-도록 하다
-ㄴ 것 같다		
-ㄴ 것1		
-ㄴ 것2		
-ㄴ 것이 아니라		
-ㄴ 대로		
-ㄴ 듯		
-ㄴ 듯하다		
-ㄴ 따름이다		
-ㄴ 때		
-ㄴ 리가 없다		
-ㄴ 만큼		
-ㄴ 만하다		
-ㄴ 모양이다		
-ㄴ 바에		
-ㄴ 법하다		

표제어	형태 정보	대표형
-르 뻔하다		
-르 뿐만 아니라		
-르 수밖에 없다		
-르 줄		
-르 테고		
-르 테냐		
-르 테니		
-르 테다		
-르 테면		
-르 테야		
-르 테지만		
-르 텐데		
-르까 보다		
-르락 말락 하다		
-려고 하다		
-려나 보다		
로 인하다	으로 인하다	
를 가지고	을 가지고	
를 막론하고	을 막론하고	
를 불문하고		
를 위해(서)	을 위해(서)	
만 같아도		만 같아도
만 아니면		만 아니면
-면 되다		
-면 몰라도		
-면 안 되다		
-면 좋겠다		
-아 가다	-어 가다, -여 가다	
-아 가지고	-어 가지고, -여 가지고	
-아 계시다		
-아 내다	-어 내다, -여 내다	

표제어	형태 정보	대표형
-아 놓다	-어 놓다, -여 놓다	
-아 대다	-어 대다, -여 대다	
-아 두다	-어 두다, -여 두다	
-아 드리다	-어 드리다, -여 드리다	
-아 버리다	-어 버리다, -여 버리다	
-아 보다	-어 보다, -여 보다	
-아 보이다	-어 보이다, -여 보이다	
-아 오다	-어 오다, -여 오다	
-아 있다	-어 있다, -여 있다	
-아 주다	-어 주다, -여 주다	
-아 치우다	-어 치우다, -여 치우다	
-아도 되다	-어도 되다, -여도 되다	
-아서는 안 되다		
-아야 되다	-어야 되다, -여야 되다	
-아야 하다	-어야 하다, -여야 하다	
-어 가다		
-어 가지고		
-어 내다		
-어 놓다		
-어 대다		
-어 두다		
-어 드리다		
-어 버리다		
-어 보다		
-어 보이다		
-어 오다		
-어 있다		

표제어	형태 정보	대표형
-어 주다		
-어 치우다		
-어도 되다		
-어야 되다		
-어야 하다		
에 관하여		에 관하여
에 관한		에 관한
에 대하여		
에 대한		
에 따라		에 따라
에 따르면		에 따르면
에 비하여		에 비하여
에 의하면		에 의하면
에 의하여		에 의하여
에도 불구하고		에도 불구하고
-여 가다		
-여 가지고		
-여 내다		
-여 놓다		
-여 대다		
-여 두다		
-여 드리다		
-여 버리다		
-여 보다		
-여 보이다		
-여 오다		
-여 있다		
-여 주다		
-여 치우다		
-여도 되다		
-여야 되다		

표제어	형태 정보	대표형
-여야 하다		
-으려고 하다	-려고 하다	
-으려나 보다	-려나 보다	
으로 인하다	로 인하다	
-으면 되다	-면 되다	
-으면 몰라도	-면 몰라도	
-으면 안 되다	-면 안 된다	
-으면 좋겠다	-면 좋겠다	
-은 가운데		-은 가운데
-은 것	-ㄴ 것	
-은 것 같다	-ㄴ 것 같다	
-은 결과	-ㄴ 결과	
은 고사하고		
-은 김에	-ㄴ 김에	
-은 나머지	-ㄴ 나머지	
-은 다음에	-ㄴ 다음에	
-은 다음에야	-ㄴ 다음에야	
-은 대로1	-ㄴ 대로, -는 대로	
-은 대로2	-ㄴ 대로, -는 대로	
-은 대신에	-ㄴ 대신에, -는 대신에	
-은 데다가1	-ㄴ 데다가	
-은 데다가2	-ㄴ 데다가, -는 데다가	
-은 뒤에		
-은 듯	-ㄴ 듯, -는 듯	
-은 듯하다	-ㄴ 듯하다, -는 듯하다	
-은 마당에	-ㄴ 마당에, -는 마당에	
-은 만큼	-ㄴ 만큼, -는 만큼	
-은 모양이다	-ㄴ 모양이다, -는 모양이다	
은 물론	는 물론	

표제어	형태 정보	대표형
-은 반면에	-ㄴ 반면에, -는 반면에	
-은 법이다	-ㄴ 법이다, -는 법이다	
-은 이상	-ㄴ 이상, -는 이상	
-은 줄	-ㄴ 줄, -는 줄	
-은 지2	-ㄴ 지	
-은 채로	-ㄴ 채로	
-은 척하다		-은 척하다
-은 체하다		
-은 탓	-ㄴ 탓, -는 탓	
-은 편이다	-ㄴ 편이다, -는 편이다	
-은 후에	-ㄴ 후에	
-은가 보다	-ㄴ가 보다, -는가 보다	
-은데도 불구하고	-ㄴ 데도 불구하고, -는데도 불구하고	
을 가지고	-ㄴ 가지고	
-을 것 같다	-ㄴ 것 같다	
-을 것1	-ㄴ 것	
-을 것2	-ㄴ 것	
-을 것이 아니라	-ㄴ 것이 아니라	
-을 나름이다		
-을 대로	-ㄴ 대로	
-을 듯	-ㄴ 듯	
-을 듯하다	-ㄴ 듯하다	
-을 따름이다	-ㄴ 따름이다	
-을 때	-ㄴ 때	
-을 리가 없다	-ㄴ 리가 없다	
-을 리가 있다		
을 막론하고	를 막론하고	
-을 만큼	-ㄴ 만큼	
-을 만하다	-ㄴ 만하다	

표제어	형태 정보	대표형
-을 모양이다	-르 모양이다	
-을 바에	-르 바에	
-을 법하다	-르 법하다	
을 불문하고		
-을 뻔하다	-르 뻔하다	
-을 뿐만 아니라	-르 뿐이다	
-을 뿐이다		
-을 수 없다		
-을 수 있다		
-을 수밖에 없다	-르 수밖에 없다	
을 위해(서)	를 위해(서)	
-을 줄	-르 줄	
-을 테고	-르 테고	
-을 테냐	-르 테냐	
-을 테니	-르 테니	
-을 테니까		
-을 테다	-르 테다	
-을 테면	-르 테면	
-을 테야	-르 테야	
-을 테지만	-르 테지만	
-을 텐데	-르 텐데	
-을까 보다	-르까 보다	
-을락 말락 하다	-르락 말락 하다	
-지 말다		-지 말다
-지 못하다		-지 못하다
-지 않다		-지 않다

2021 Project on the Research and Construction of the Korean Language Learner Corpus

This project was the <2015-2020 Project on the Research and Construction of the Korean Language Learner Corpus> project according to the first mid- to long-term plan. Then, the second mid- to long-term plan was established with the purpose of building an actual corpus according to the plan. The major tasks and goals of this project are as follows.

2nd mid- to long-term plan establishment: The mid- to long-term plan for the learner corpus is an analysis of policies, laws, and systems related to language resource construction that affect the construction and use of the learner corpus, analysis of the needs of various user groups including academia and the private sector, analysis of precedent cases, and analysis of existing established corpora based on the results of basic research analysis, a five-year plan was established for 2021 to 2025. The goal was to build a The goal was to build up accumulated knowledge through the corpus over five years of business up to 2020 with the original corpus of 10 million items (written items: 6 million, spoken items: 4 million), 10 million morph-tagged corpora (written items: 6 million, spoken items: 4 million), and 5 million error annotation corpora (written items: 3 million, spoken items: 2 million). At the same time, it was proposed that a reference corpus be constructed to enhance the utilization of the original corpus, and a corpus individually constructed by the researcher be built or integrated with a corpus provided by other institutions.

Korean language learner corpus collection, revision, and construction: The 2021 Korean language learners' corpus planned and collected tasks to intensively collect data on genres and topics that were

relatively lacking in 2015–2020. The team constructed new items in the form of 831,142 raw corpora (written items: 419,371, spoken items: 411,771), 200,981 morph-tagged corpora (written items: 100,781, spoken items: 100,200), and 151,906 error-annotated corpora (written items: 104,314, spoken items: 47,592). In total, from 2015 to 2021 5,220,564 raw corpora (written items: 3,697,952, spoken items: 1,522,612), 3,704,586 morph-tagged corpora (written items: 2,602,914, spoken items: 1,101,672), and 1,153,848 error-annotated corpora (written items: 590,548, spoken items: 563,300) were produced.

Verification of corpora construction support tool: The Korean language learner corpus has been built using a corpus-building support tool to manage the entire process from sample registration to corpus annotation processing. To provide a stable construction environment to workers, a performance feedback team was formed centered on construction research staff and continuous monitoring in this study, feedback was provided for performance improvement and stabilization of the learner corpus construction support tool. In addition, to improve the efficiency of the large-scale construction project to be carried out in the future, the performance of the form annotator built into the support tool was objectively evaluated and improvements were made to suggest a direction for upgrading detailed functions.

Elaboration of construction corpus inspection: Verification of the construction corpus was to improve the quality of the constructed corpus data, and the process was carried out according with the three-steps work and inspection system for each work step of written input: spoken language transcription, morph-tagged annotation, and error annotation. Otherwise, complementary techniques of erroneous data and abnormal data inspection through system-based data verification were applied. In addition, sample information was inspected throughout the entire project period to increase the accuracy of statistics and duplicate sampling through comparison of the entire sample list.

Learner corpus education and promotion: Korean language learner corpus-related education was provided for construction practical staff and users. In addition to instruction and tool use education, an immediate feedback system was operated to solve various problems that occurred during the construction process, and various issues related to corpus construction were shared through regular workshops. Training for users was conducted a total of five times through learner corpus academies. By differentiating users or programs for each session, various content from basic to advanced courses was covered, and a panel meeting was held with experts in each field. In addition, the research results using the learner corpus were reported to the academic community through presentations at an academic conference.

This follow-up stage of the <Korean Language Learner Corpus> Project can be applied broadly by education researchers, Korean language instructors, and Korean language learners to strengthen the systemization of the Korean language and its international competitiveness.

Keywords: Korean language learner corpus, 2nd mid- to long-term plan establishment, written language corpus, spoken language corpus, raw corpus, morph-tagged corpus, error-annotated corpus

<기획·연구>

국립국어원 박미영 학예연구사

국립국어원 조은 연구원

<연구 참여자>

연구 책임자 한송화(연세대학교)

공동 연구원 김선정(계명대학교)

김인철(상명대학교)

김일환(성신여자대학교)

김한샘(연세대학교)

장석배(미국 밴더빌트대)

홍혜란(연세대학교)

박미영(국립국어원)

조 은(국립국어원)

연구 보조원 김동은(연세대학교)

김미선(연세대학교)

김선영(연세대학교)

송지혜(연세대학교)

유소영(연세대학교)

허희정(연세대학교)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2021년 12월 10일

발행일: 2021년 12월 10일

인 쇄: 학위사

※ “이 책은 국립국어원의 용역비로 수행한 ‘한국어 학습자 말뭉치 연구 및 구축’ 사업의 결과물을 발간한 것입니다.”