국립국어원 2022-01-17

발 간 등 록 번 호
11-1371028-000746-10

2022년 한국어 학습자 말뭉치 연구 및 구축 사업

연구 책임자: 한 송 화

국립국어원

제 출 문

국립국어원장 귀하

"2022년 한국어 학습자 말뭉치 연구 및 구축" 사업에 관하여 귀 원과 체결한 연구용역 계약에 의하여 연구보고서를 작성하여 제출합니다.

2022년 12월 9일

연구 책임자: 한송화(연세대학교)

연구 기관	연세대학교 산학협력단
연구 책임자	한송화(연세대)
공동 연구원	강현화(연세대), 김선정(계명대), 김일환(성신여대), 김한샘(연세대), 장석배(미국 밴더빌트대), 홍혜란(연세대), 박미영(국립국어원), 홍혜진(국립국어원)
연구 보조원	김동은(연세대), 김미선(연세대), 서지혜(연세대), 손연정(연세대)

<국문 초록>

2022년 한국어 학습자 말뭉치 연구 및 구축 사업

본 연구는 <2021년 한국어 학습자 말뭉치 연구 및 구축>에서 수립한 제2차 중장기 계획에 따라 100만 어절의 원시 말뭉치를 수집하여 구축하고, 말뭉치의 균형성을 고려하여 30만 어절의 형태 주석 말뭉치와 20만 어절의 오류 주석 말뭉치를 구축·가공하는 것을 목표로 하였다. 이에 따른 주요 과업과 연구 성과는 다음과 같다.

한국어 학습자 말뭉치 수집 및 구축·가공: 2022년 학습자 말뭉치는 대상별·수준별·언어권별·자료 변인별 분포 특성을 분석하여 상대적으로 비중이 적은 자료를 집중 수집하여 구축하는 것을 목표로 하였다.

원시 말뭉치 1,008,315어절(문어 708,096어절, 구어 300,219어절), 형태 주석 말뭉치 308,790어절(문어 157,313어절, 구어 151,477어절), 오류 주석 203,490어절(문어 73,829어절, 구어 129,661어절) 규모의 말뭉치가 새롭게 구축되었다. 그 결과 전체 말뭉치의 규모는 원시 말뭉치 6,230,590어절(문어 4,407,583어절, 구어 1,823,007어절), 형태 주석 말뭉치 4,013,233어절(문어 2,760,085어절, 구어 1,253,148어절), 오류 주석 말뭉치 1,346,015어절(문어 674,636어절, 구어 671,379어절)이 되었다.

아울러 학습자 말뭉치의 활용도 제고를 위해 30명의 한국어 모어 화자를 대상으로 자료를 수집하여 총 문어 10,440어절, 구어 23,912어절의 참조 말뭉치를 시범 구축하였다.

학습자 말뭉치의 검수 정교화 및 품질 관리: 학습자 말뭉치의 검수 정교화 및 품질 관리는 양적 확대와 더불어 언어 자원으로서의 질적 제고를 위한 것으로 데이터의 무결성을 확보하는 것을 목적으로 한다. 이에 따라 문어 입력과 구어 전사, 형태 주석, 오류 주석의 각 작업 단계별로 3단계 작업 및 검수 체계에 따라 작업 공정을 진행하였으며, 공동 연구원을 중심으로 내부 검수단을 운영하여 무작위로 자료를 검수하는 절차를 두어 검수 체계를 강화하였다. 그 외에도 시스템 기반의 데이터 검증을 통한 오조작 데이터와 이상 데이터 검수를 상호보완적으로 적용하였으며, 최종 단계에서 전체 표본 정보

검수를 통한 메타정보 검증 작업과 중복 표본 검수 작업을 수행하였다. 또한 구축 말뭉치의 통계 정보의 정확성 제고를 목적으로 기구축 말뭉치의 통계 정보를 검토하고 이를 토대로 LCMS의 작업 할당 취소 표본의 이력 관리 방식 개선, <국립국어원 한국어 학습자 말뭉치 나눔터>의 이용자 검색 통계 제시 방식 개선을 제안하였다.

학습자 말뭉치 교육 및 홍보: 한국어 학습자 말뭉치 교육은 실무 작업자와 사용자를 대상으로 하여 이루어졌다. 실무 작업자 교육은 작업자에게 말뭉치 구축에 관한 기본 소양과 기술을 익히도록 하고 각 구축 단계별 작업자로서의 전문성을 제고하여 체계적인 말뭉치를 구축해 나가기 위한 것으로, 지침 교육과 도구 사용 교육을 기본으로 한다. 그리고 구축 과정에서 발생하는 다양한 문제를 해결하기 위한 즉각적 소통과 피드백 시스템을 운영하고 정기 워크숍을 통해 말뭉치 구축에 관한 쟁점과 대응 방안을 공유하였다.

사용자를 대상으로 한 교육은 기초 과정에서 심화 과정까지 차별화된 총 4회의 학습자 말뭉치 아카데미를 통해 이루어졌다. 기초 과정으로는 학습자 말뭉치 기반 연구를 위한 자료 처리를 주제로 하여 2회가 이루어졌으며, 심화 과정으로는 인공지능 기술을 기반으로 한 한국어 학습자 말뭉치 활용 사례, 학습자 말뭉치를 활용한 연구 주제 탐색과 적용이라는 주제로 하여 2회가 이루어졌다. 또한 한국어 학습자 말뭉치를 소개하고 활용 방안을 설명하는 안내 자료와 동영상을 제작하여 국립국어원 누리집과 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 배포하였다.

한국어 학습자 말뭉치는 한국어 교육 연구, 교수, 학습에 광범위하게 활용됨으로써 한국어의 세계화 및 국제 경쟁력 강화에 이바지할 것이다.

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

차 례

표 차례

그림 차례

Ⅰ. 연구 개요

1. 연구의 목적 및 필요성

1.1. 연구의 목적

○ 본 연구는 2021년에 수립한 한국어 학습자 말뭉치 제2차 중장기 계획에 따라 국가 언어 자원으로서 한국어 학습자 말뭉치의 규모를 확대하고 균형성을 확보하는 것을 목표로 한다. 이에 따라 규모 확대의 측면에서 2025년까지 누적 규모 1,000만 어절(원시 말뭉치 기준)의 학습자 말뭉치 구축을 목표로, 기구축된 520만 어절 규모의 말뭉치에 100만 어절의 말뭉치를 구축하여 더함으로써 누적 규모 620만 어절의 말뭉치가 구축되었다. 아울러 균형성 확보를 위하여 기구축 말뭉치의 성과 분석을 통해 대상, 수준, 언어권, 주제, 장르 등의 변인별 자료가 균형성을 갖출 수 있도록 말뭉치 수집 및 구축·가공 작업을 수행하였다. 이는 궁극적으로 한국어 교육과 연구, 산업계 기술 연구 분야에서의 활용을 활성화하고, 그럼으로써 체계적이고 과학적인 한국어 교육의 기반을 마련하기 위한 것이다. 다음은 본 연구의 핵심 과업이다.

- 한국어 학습자 말뭉치 수집 및 구축·가공
- 구축 학습자 말뭉치 검수 정교화 및 품질 관리
- 한국어 학습자 말뭉치 교육 및 홍보

1.2. 연구의 필요성

○ 공공 언어 자원으로서 국가 주도의 한국어 학습자 말뭉치 구축

한국어 학습자 말뭉치는 2015년에 수립한 제1차 중장기 계획에 따라 국내 교육 기관의 유학생과 이주민, 국외의 한국어 학습자 자료를 대규모로 수집하여 약 520만 어절 규모의 균형 말뭉치를 구축하였다. 이러한 성과는 2002년 문화체육관광부의 주도로 수행된 50만 어절 규모의 한국어 학습자 말뭉치 구축 사업 이후 13년 만에 새롭게 시작된 국가 주도의 사업으로 한국어의 세

계화를 위한 지식 기반 구축 사업으로서 높은 평가를 받았다. 본 연구는 그 후속 사업으로 균형성을 갖춘 대규모 한국어 학습자 말뭉치를 구축한다는 데에 의의가 있다.

○ 1,000만 어절 규모의 대규모 학습자 말뭉치로의 확장

학습자 말뭉치 기반 연구는 최근 외국어 또는 제2언어 교육 분야의 핵심적인 경향 중 하나라고 할 수 있다. 국외에서는 민간 기관, 대학, 개인 연구자에 의해 약 183종의 학습자 말뭉치가 구축되어 왔으며, 그 중에는 1,000만 어절 이상의 대규모 말뭉치도 다수 포함되어 있다. 대표적인 영어 학습자 말뭉치인 CLC(Cambridge Learner Corpus)는 약 5천만 어절, The Hong Kong University of Science & Technology(HKUST) 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. 본 연구에서는 2015년부터 2021년까지 구축한 520만 어절 규모의 한국어 학습자 말뭉치에 100만 어절의 원시 말뭉치를 추가로 구축하여 누적 규모 620만 어절의 말뭉치를 구축하였다. 제2차 중장기 계획에 따르면 한국어 학습자 말뭉치는 2025년까지 누적 1,000만 어절 규모로 구축될 예정으로, 이는 세계적인 수준의 학습자 말뭉치를 구축한다는 점에서 큰 의미가 있다. 아울러 그 활용 범위를 한국어 교육 및 연구뿐만 아니라 산업계, 일반인 등 민간까지 확대한다는 점에서 의의가 있다.

○ 2015-2021년 학습자 말뭉치 보완을 통한 균형성 확보

2015년부터 2020년까지의 한국어 학습자 말뭉치 연구 및 구축 사업에서는 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. 2021년에는 제2차 중장기 계획 수립을 위한 기초 연구와 함께 문어와 구어 각 40만 어절씩 총 80만 어절 규모의 말뭉치를 추가로 구축하였다. 제1차 중장기 계획에서는 선구축 후균형의 방식으로 한국어 학습자의 실제 분포 특성을 반영하여 귀납적으로 균형을 맞추고자 하였기 때문에 학습자의 수준, 제1언어, 자료의 장르, 주제 등의 변인이 다소 불균형하다는 한계가 있다. 본 연구에서는 2015년부터 2021년까지의 한국어 학습자 말뭉치 연구 및 구축 사업의 성과에 대한 분석을 바탕으로 대상, 수준, 언어권, 주제·장르의 변인 측면에서 특정 변인에 편중된 자료를 보완함으로써 균

형성을 확보하여 국가 주도 학습자 말뭉치의 질을 제고한다는 점에서 실효성이 크다고 하겠다.

○ 국가 언어 자원으로서의 품질 제고

학습자 말뭉치는 대규모 언어 자원이라는 특성상 오랜 시간과 다수의 작업자들의 노력이 소요되는 노동집약적인 작업을 통해 구축된다. 비모어 화자가 산출한 언어 자료이기 때문에 입력과 전사, 형태 주석과 오류 주석의 전 과정에서 자료 처리에 관한 다양한 쟁점과 이견이 발생하며, 이를 일관성 있게 체계적으로 처리하는 것이 매우 중요하다. 이러한 특성으로 인해 모어 화자 말뭉치보다 한층 복잡한 작업 공정이 요구되며, 지속적인 검증과 보완이 요구된다. 본 연구에서는 2022년 신규 구축 외에도 2015년부터 2021년까지 구축된 결과물을 지속적으로 분석하고 모니터링하였다. 이 과정에서 남아 있는 오류들을 수정하고 이전 말뭉치와 신규 구축 말뭉치와의 체계를 맞추어 감으로써 말뭉치의 질적 제고를 도모할 수 있다. 이는 국가 언어 자원으로서의 신뢰성 확보 차원에서 매우 중요한 일이라고 할 수 있다.

○ 학습자 말뭉치 기반 연구 방법론과 사용 교육에 대한 요구

한국어 학습자 말뭉치에 대한 연구자와 교수자들의 관심이 날로 커지고 있다. 그럼에도 불구하고 자료의 접근이 어려워 개인이 말뭉치를 구축하기는 쉽지 않았다. 이러한 점에서 국가 수준의 학습자 말뭉치를 구축한다는 것은 국가 자원으로서 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 학습자 말뭉치 소개 및 활용 자료집을 현행화하여 제공하고, 학습자 말뭉치 아카데미를 통해 자료 처리 및 활용 방법을 교육하였다. 이는 학습자 말뭉치의 활용도 제고에 기여할 것이다.

2. 연구의 범위

○ 본 연구의 범위와 세부 내용은 다음과 같다.

<표 1> 연구의 범위와 세부 과업

연구의 범위	세부 내용
한국어 학습자 말뭉치 수집 및 구축·가공	○ 100만 어절의 한국어 학습자 원시 말뭉치 자료 (문어, 구어) 수집 및 구축·가공 ○ 2015-2021년 기구축 말뭉치 보완을 위한 기획 구축·가공 ○ 한국어 학습자 말뭉치의 활용도 제고를 위한 한국어 모어 화자 참조 말뭉치 수집·구축 ○ 수집·구축 대상 자료의 한국어 학습자 이용 허락 확보
구축 학습자 말뭉치 검수 정교화 및 품질 관리	○ 구어·문어 입력 검수 등 검수 정교화 및 언어 자원 품질 확보 ○ 구축 말뭉치 통계 정보의 정확성 제고
한국어 학습자 말뭉치 교육 및 홍보	○ 말뭉치 구축 인력 실무 교육 ○ 한국어 학습자 말뭉치 이용자를 위한 아카데미 개최 (상시, 최소 연 4회) ○ 한국어 학습자 말뭉치 소개·활용 자료집 현행화 등, 국립국어원 관련 누리집(한국어 학습자 말뭉치 나눔터 등) 게재 및 아카데미 배포

3. 연구 방법

○ 본 연구는 학습자 말뭉치 수집 및 구축·가공을 핵심 과업으로 한다. 이를 체계적으로 수행하기 위한 수행 방법은 다음과 같다.

<표 2> 과업 내용과 연구 수행 방법

과업 내용		연구 방법	과업 유형
한국어 학습자 말뭉치 수집 및 구축·가공	100만 어절의 한국어 학습자 원시 말뭉치 자료 (문어, 구어) 수집 및 구축·가공	○ 기구축 자료 분석 ○ 말뭉치 수집 및 구축 계획 설계 ○ 실제 말뭉치 구축	수집, 구축·가공
	2015-2021년 기구축 말뭉치 보완을 위한 기획 구축·가공	○ 기구축 자료 분석 ○ 말뭉치 수집 및 구축 계획 설계 ○ 실제 말뭉치 구축	수집, 구축·가공
	한국어 학습자 말뭉치의 활용도 제고를 위한 한국어 모어 화자 참조 말뭉치 수집·구축	○ 기구축 자료 분석 ○ 말뭉치 수집 및 구축 계획 설계 ○ 실제 말뭉치 구축	수집, 구축·가공
	수집·구축 대상 자료의 한국어 학습자 이용 허락 확보	○ 기구축 자료 분석 ○ 말뭉치 수집 및 구축 계획 설계 ○ 실제 말뭉치 구축	수집
구축 학습자 말뭉치 검수 정교화 및 품질 관리	구어·문어 입력 검수 등 검수 정교화 및 언어 자원 품질 확보	○ 3차 검수 체계 운영 ○ 내부 검수단 운영	자료 검수 및 검증
	구축 말뭉치 통계 정보의 정확성 제고	○ 학습자 말뭉치 나눔터 및 LCMS 통계 분석 및 검증	자료 검수 및 검증

과업 내용		연구 방법	과업 유형
한국어 학습자 말뭉치 교육 및 홍보	말뭉치 구축 인력 실무 교육	○ 작업 실무자 정기회의(격주) ○ 작업 실무자 교육 및 워크숍(월 1회)	교육 및 홍보
	한국어 학습자 말뭉치 이용자를 위한 아카데미 개최	-	교육 및 홍보
	한국어 학습자 말뭉치 소개·활용 자료집 현행화 등, 국립국어원 관련 누리집(한국어 학습자 말뭉치 나눔터 등) 게재 및 아카데미 배포	-	교육 및 홍보

4. 연구 추진 일정

○ 연구 추진 일정은 다음과 같다.

<표 3> 연구 추진 일정

과업 내용			3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
사업 계획 수립 및 사업 착수			●									
한국어 학습자 말뭉치 수집 및 구축·가공	100만 어절의 한국어 학습자 원시 말뭉치 자료 (문어, 구어) 수집 및 구축·가공	수집		●	●	●	●	●	●	●	●	●
		구축		●	●	●	●	●	●	●	●	●
		가공			●	●	●	●	●	●	●	
	2015-2021년 기구축 말뭉치 보완을 위한 기획 구축·가공	수집		●	●	●	●	●	●	●	●	●
		구축			●	●	●	●				
		가공			●	●	●	●	●	●		
	한국어 학습자 말뭉치의 활용도 제고를 위한 한국어 모어 화자	수집							●			
		구축										●

과업 내용		3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
	참조 말뭉치 수집·구축										
	수집·구축 대상 자료의 한국어 학습자 이용 허락 확보		●	●	●	●	●	●	●	●	
구축 학습자 말뭉치 검수 정교화 및 품질 관리	구어·문어 입력 검수 등 검수 정교화 및 언어 자원 품질 확보		●	●	●	●	●	●	●	●	●
	구축 말뭉치 통계 정보의 정확성 제고	●	●	●	●	●	●	●	●	●	
한국어 학습자 말뭉치 교육 및 홍보	말뭉치 구축 인력 실무 교육	●	●	●	●	●	●	●	●	●	●
	한국어 학습자 말뭉치 이용자를 위한 아카데미 개최			●			●		●	●	
	한국어 학습자 말뭉치 소개·활용 자료집 현행화 등, 국립국어원 관련 누리집(한국어 학습자 말뭉치 나눔터 등) 게재 및 아카데미 배포			●			●		●	●	

과업 내용	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
최종보고회의 및 사업 마무리										●

5. 연구 결과

○ 본 사업의 연구 결과는 다음과 같다.

- 100만 어절의 한국어 학습자 원시 말뭉치 자료(문어, 구어) 수집 및 구축·가공, 2015-2021년 기구축 말뭉치 보완을 위한 기획 구축·가공

<표 4> 2015-2022년 학습자 말뭉치의 구축 규모

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
원시 말뭉치									
2015 - 2021	문어	439,612 (6,690)	596,563 (6,134)	690,248 (5,796)	650,251 (4,977)	722,627 (4,662)	459,546 (2,802)	140,640 (141)	3,699,487 (31,202)
	구어	277,908 (833)	313,967 (668)	406,673 (690)	246,126 (519)	139,498 (245)	107,877 (166)	30,739 (21)	1,522,788 (3,142)
	합계	717,520 (7,523)	910,530 (6,802)	1,096,921 (6,486)	896,377 (5,496)	862,125 (4,907)	567,423 (2,968)	171,379 (162)	5,222,275 (34,344)
2022	문어	155,083 (2,060)	145,112 (1,455)	136,285 (987)	159,864 (1,016)	66,767 (365)	42,309 (205)	2,676 (10)	708,096 (6,098)
	구어	19,569 (75)	34,875 (91)	55,805 (103)	81,511 (154)	77,199 (115)	29,140 (38)	2,120 (2)	300,219 (578)
	합계	174,652 (2,135)	179,987 (1,546)	192,090 (1,090)	241,375 (1,170)	143,966 (480)	71,449 (243)	4,796 (12)	1,008,315 (6,676)

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
합계	문어	594,695 (8,750)	741,675 (7,589)	826,533 (6,783)	810,115 (5,993)	789,394 (5,027)	501,855 (3,007)	143,316 (151)	4,407,583 (37,300)
	구어	297,477 (908)	348,842 (759)	462,478 (793)	327,637 (673)	216,697 (360)	137,017 (204)	32,859 (23)	1,823,007 (3,720)
	합계	892,172 (9,658)	1,090,517 (8,348)	1,289,011 (7,576)	1,137,752 (6,666)	1,006,091 (5,387)	638,872 (3,211)	176,175 (174)	6,230,590 (41,020)
형태 주석 말뭉치									
2015 – 2021	문어	381,081 (5,569)	433,472 (4,399)	447,834 (3,725)	421,654 (3,300)	443,722 (2,991)	407,971 (2,594)	67,038 (62)	2,602,772 (22,640)
	구어	214,522 (702)	205,527 (443)	210,319 (460)	210,050 (431)	136,266 (240)	103,875 (160)	21,112 (15)	1,101,671 (2,451)
	합계	595,603 (6,271)	638,999 (4,842)	658,153 (4,185)	631,704 (3,731)	579,988 (3,231)	511,846 (2,754)	88,150 (77)	3,704,443 (25,091)
2022	문어	–	47,075 (491	52,975 (467)	41,103 (313)	13,557 (59)	2,603 (5)	–	157,313 (1,335)
	구어	21,318 (53)	53,933 (113)	52,051 (111)	23,453 (59)	–	722 (1)	–	151,477 (337)
	합계	21,318 (53)	101,008 (604)	105,026 (578)	64,556 (372)	13,557 (59)	3,325 (6)	–	308,790 (1,672)
합계	문어	381,081 (5,569)	480,547 (4,890)	500,809 (4,192)	462,757 (3,613)	457,279 (3,050)	410,574 (2,599)	67,038 (62)	2,760,085 (23,975)
	구어	235,840 (755)	259,460 (556)	262,370 (571)	233,503 (490)	136,266 (240)	104,597 (161)	21,112 (15)	1,253,148 (2,788)
	합계	616,921 (6,324)	740,007 (5,446)	763,179 (4,763)	696,260 (4,103)	593,545 (3,290)	515,171 (2,760)	88,150 (77)	4,013,233 (26,763)
오류 주석 말뭉치									

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 – 2021	문어	89,925 (1,302)	104,084 (1,080)	104,832 (924)	86,977 (719)	109,714 (775)	105,275 (690)	–	600,807 (5,490)
	구어	93,897 (310)	103,304 (220)	104,126 (255)	107,738 (229)	71,964 (122)	55,797 (74)	4,892 (5)	541,718 (1,215)
	합계	183,822 (1,612)	207,388 (1,300)	208,958 (1,179)	194,715 (948)	181,678 (897)	161,072 (764)	4,892 (5)	1,142,525 (6,705)
2022	문어	9,242 (85)	9,202 (105)	11,859 (130)	21,280 (212)	10,185 (102)	12,061 (118)	–	73,829 (752)
	구어	42,471 (133)	19,861 (30)	19,365 (46)	21,269 (42)	16,592 (27)	10,103 (14)	–	129,661 (292)
	합계	51,713 (218)	29,063 (135)	31,224 (176)	42,549 (254)	26,777 (129)	22,164 (132)	-	203,490 (1,044)
합계	문어	99,167 (1,387)	113,286 (1,185)	116,691 (1,054)	108,257 (931)	119,899 (877)	117,336 (808)	–	674,636 (6,242)
	구어	136,368 (443)	123,165 (250)	123,491 (301)	129,007 (271)	88,556 (149)	65,900 (88)	4,892 (5)	671,379 (1,507)
	합계	235,535 (1,830)	236,451 (1,435)	240,182 (1,355)	237,264 (1,202)	208,455 (1,026)	183,236 (896)	4,892 (5)	1,346,015 (7,749)

- 한국어 모어 화자 참조 말뭉치 수집·구축
- 구어·문어 입력 검수 등 검수 정교화 및 언어 자원 품질 확보
- 구축 말뭉치 통계 정보의 정확성 제고
- 한국어 학습자 말뭉치 이용자를 위한 아카데미 개최
- 한국어 학습자 말뭉치 소개·활용 자료집 현행화, 국립국어원 관련 누리집 게재 및 아카데미 배포

Ⅱ. 한국어 학습자 말뭉치 수집 및 구축·가공

1. 100만 어절의 한국어 학습자 원시 말뭉치 자료(문어, 구어) 수집 및 구축·가공

○ 본 연구에서는 100만 어절의 한국어 학습자 원시 말뭉치(문어, 구어) 수집 및 구축·가공에 앞서 2015-2021년까지의 구축 말뭉치의 유형별 분포 특성을 분석하였다. 그리고 그 결과를 바탕으로 2022년 말뭉치 구축 목표를 설정하고 그에 따른 세부 사항을 설계하였다.

1.1. 말뭉치 구축 설계

1) 목표 규모 설정

○ 본 연구에서는 수요 기관과의 협의를 통해 문어 70만 어절, 구어 30만 어절로 총 100만 어절 규모의 원시 말뭉치를 구축하는 것으로 목표를 설정하였다.

<표 5> 원시 말뭉치 구축 목표

문어			구어			합계		
기구축	2022년	누적 합계	기구축	2022년	누적 합계	기구축	2022년	누적 합계
369만	70만	439만	152만	30만	182만	521만	100만	621만

2) 변인별 구축 목표 설정

(1) 변인별 말뭉치 구축을 위한 기획 구축의 비중 설정

○ 본 연구에서는 2차 중장기 계획에 따라 대규모 구축(수집 네트워크 및 학습자의 자율적 참여 기반)을 병행하되, 기구축 말뭉치 분석 결과를 토대로

대상별·수준별·언어권별·자료 변인별 자료를 집중적으로 구축해 나갈 수 있도록 목표 규모를 설정하였다. 이는 2025년까지 점진적으로 말뭉치의 균형성을 확대해 나가기 위한 것으로, 원시 말뭉치의 경우 대규모 구축과 기획 구축의 비중을 각각 50%로 설정하기로 하였다.

<표 6> 변인별 말뭉치 구축을 위한 기획 구축의 비중

문어			구어			합계		
대규모 구축 (전체)	기획 구축 (변인)	누적 합계 (전체)	대규모 구축 (전체)	기획 구축 (변인)	누적 합계	대규모 구축 (전체)	기획 구축 (변인)	누적 합계
26만 (50%)	26만 (50%)	52만 (100%)	24만 (50%)	24만 (50%)	48만 (100%)	50만 (50%)	50만 (50%)	100만 (100%)

○ 위와 같은 목표에 따라 수집과 구축을 진행하되, 특히 수집의 경우는 현실적인 수집 결과에 따라 구축 비중이 귀납적으로 조정될 수 있음을 전제로 하였다. 이는 이어지는 변인별 목표에도 동일하게 적용되었다.

(2) 변인별 구축 목표 설정

○ 본 연구에서는 기구축 말뭉치의 분포 특성을 고려하여 100만 어절 규모의 원시 말뭉치 구축을 위한 대상별, 수준별, 언어권별 수집 및 구축 목표를 설정하였다.

① 대상별

○ 대상에 따른 기구축 말뭉치의 구축 비중을 분석한 결과, 전체 자료 중 국외 학습자와 이주민 자료의 비중이 현저하게 낮아 2025년까지 두 집단의 자료를 지속적으로 구축할 필요가 있음을 확인하였다. 아울러 국내 학습자 자료의 경우, 최근 유학생의 수가 점점 증가하고 있음을 고려할 때 학문 목적 학습자의 자료 비중 또한 매우 적음을 확인할 수 있다.

<표 7> 2015-2021년 말뭉치의 구축 규모: 대상별

대상	문어	구어
국내 학습자 (학문 목적)	3,559,627 (209,309)	820,036 (94,260)
이주민	88,819	308,982
국외 학습자	51,041	393,770
합계	3,699,487	1,522,788

○ 이에 따라 본 연구에서는 국내 학습자 중 학문 목적 학습자의 자료와 국외 자료를 집중적으로 수집하여 구축하는 것을 목표로 설정하였다.[1)]

② 수준별

○ 학습자의 수준에 따른 기구축 말뭉치의 구축 비중을 분석한 결과, 문어의 경우 1급과 2급, 6급, 6급 이상의 자료 비중이 상대적으로 낮고, 구어의 경우, 특히 5급과 6급, 6급 이상의 자료의 비중이 매우 낮음을 확인할 수 있었다.

<표 8> 2015-2021년 말뭉치의 구축 규모: 수준별

수준	문어	구어
1급	439,612	277,908
2급	596,563	313,967
3급	690,248	406,673
4급	650,251	246,126
5급	722,627	139,498
6급	459,546	107,877
6급 이상	140,640	30,739
합계	3,699,487	1,522,788

1) 당초 계획에서는 이주민 자료를 포함하여 세 집단의 자료를 고르게 수집하여 구축하는 것을 목표로 설정하였으나, 자료 수집의 효율성을 고려할 때 집중 수집·구축을 하는 것이 더 적절하다고 판단이 되었다. 이에 따라 수집 목표를 수정하였다.

○ 이에 따라 본 연구에서는 구어와 문어의 분포 특성을 고려하여 1급과 2급, 5급과 6급, 대학(원)에 진학하여 학습하고 있는 6급 이상의 학습자 자료를 집중적으로 수집하여 구축하는 것을 목표로 설정하였다.

③ 언어권별

○ 학습자의 수준에 따른 기구축 말뭉치의 구축 비중을 분석한 결과, 언어권별 자료는 구축량 순위를 기준으로 전체 자료의 약 50%에 이르는 중국어권 자료를 제외한 상위 10위의 언어권 중 일본어권, 영어권, 베트남어권, 타이어권, 스페인어권, 러시아어권의 비중을 높일 필요가 있음을 확인하였다.

<표 9> 2015-2021년 말뭉치의 구축 규모: 언어권별

언어권	문어	구어
일본어권	516,104	142,524
영어권	241,269	56,329
베트남어권	252,924	227,830
타이어권	97,309	294,310
스페인어권	57,047	95,395
러시아어권	129,235	79,083

○ 이에 따라 본 연구에서는 이들 언어권의 자료를 집중적으로 구축하되, 특히 기구축 규모의 말뭉치가 현저히 적은 영어권, 스페인어권, 러시아권을 집중적으로 구축하는 것을 목표로 설정하였다.

(3) 장르·주제별

○ 학습자의 수준에 따른 기구축 말뭉치의 구축 비중을 분석한 결과, 장르별 자료는 문어의 경우, 생활문과 논설문의 비중이 압도적으로 높고, 구어의 경우 인터뷰 자료의 비중이 두드러지게 높아 그 밖의 장르의 비중을 높일 필요가 있음을 확인할 수 있었다. 또한 주제의 경우 구어와 문어 모두 '개인 신상, 일상생활, 여가와 오락, 사회, 교육, 대인 관계, 여행 등'과 같은 일상적인 주제에 집중되고 있어 사회적인 주제나 전문적인 주제로의 확대가 필요함을 알 수 있었다.

<표 10> 2015-2021년 말뭉치의 구축 규모: 장르별

문어		구어	
장르	구축 규모	장르	구축 규모
생활문	1,650,609	인터뷰	978,900
논설문	1,207,050	발표	280,981
설명문	251,033	내러티브	156,844
보고서	239,247	자유 대화	106,063
기행문	110,326		
수필	104,868		
감상문	53,853		
기사문	38,078		
전기문	21,863		
편지글	21,655		
합계	3,699,487	합계	1,522,788

○ 이에 따라 본 연구에서는 장르별·주제별로 다음과 같은 목표를 설정하였다.

- 문어의 경우 전체 자료 중 장르별 비중이 낮은 설명문과 보고서를 집중적으로 구축한다. 이 중 설명은 주로 5급과 6급의 학습자를 대상으로 하며, 보고서는 6급 이상의 학습자로 국내 대학(원)에 재학 중인 학습자를 대상으로 하여 수집한다. 주제는 수집 비중이 낮은 '교육', '일과 직업', '건강', '전문 분야', '기후' 등을 집중적으로 구축한다.
- 구어의 경우는 내러티브, 자유 대화, 발표 자료를 집중적으로 구축한다. 이 중 발표 자료는 6급 이상의 학문 목적 학습자를 대상으로 한다. 주제는 수집 비중이 낮은 주제 중 '여가와 오락', '대인 관계', '여행' 등과 같이 일상적인 주제와 '사회', '교육', '일과 직업' 등의 사회적인 주제의 자료를 집중적으로 구축한다.

1.2. 실제 수집 및 원시 말뭉치 구축

○ 본 연구에서는 앞선 단계에서 설정한 구축 목표에 따라 자료 수집, 입력 및 전사 작업을 수행하고 있다.

1) 자료 수집

○ 2022년의 자료 수집은 대규모 구축을 위한 수집과 균형성 확보를 위한 수집의 두 가지에 초점이 있다. 대규모 수집은 2025년까지 1,000만 어절 규모의 말뭉치를 효율적으로 구축해 나가고 그 이후에도 지속적인 자료의 업데이트를 위한 환경 설정을 위한 것이다. 한편, 균형성 확보를 위한 수집은 대상, 수준, 언어권, 장르, 주제 등의 변인별 균형성 확보를 위해 전체 말뭉치에서 비중이 적은 특정 변인의 자료를 집중적으로 수집하기 위해 수집 대상을 특정하거나 과제를 기획하여 수집하는 방식이다. 이에 따른 자료 수집 세부 설계와 진행 현황은 다음과 같다.

(1) 수집 대상 및 수집 네트워크

○ 100만 어절의 원시 말뭉치 구축을 위한 자료 수집은 학습자 변인과 자료 변인을 특정하지 않은 대규모 수집과 기획 수집으로 구분된다.

<표 11> 자료 수집 대상과 경로

구분	대규모 수집	기획 수집
수집 대상	○ 변인을 특정하지 않은 모든 자료	○ 특정 대상, 수준, 언어권, 장르, 주제의 자료
수집 경로	○ 온라인 수집을 통한 자율적 참여 ○ 국내외의 한국어 교육 기관 및 학계, 유관 기관 등의 자율적 참여	○ 국내외의 한국어 교육 기관 및 학계, 유관 기관 등의 수집 네트워크

○ 다음은 이 중 기획 수집을 위한 수집의 대상이 되는 학습자 유형과 수집 경로이다.

<표 12> 한국어 학습자 말뭉치의 수집 경로

구분	수집 대상	자료 수집 경로
국내 학문 목적 학습자	○ 대학(원) 유학생	○ 국내 대학(원)의 교양학부 글쓰기 수업, 외국인 학부 대학의 수업
국외 학습자	○ 대학(원)의 한국학 전공 및 한국어 학습자 ○ 그 외의 한국어 학습자	○ 한국국제교류재단 ○ 국외 한국어 교육자 네트워크 ○ 세종학당재단

○ 현재 수집에 참여하고 있는 수집 기관은 다음과 같다. 이 중 한국어학당은 대규모 수집과 기획 수집에 모두 참여하고 있으며, 그 외의 기관은 학습자 변인, 과제 변인에 따른 기획 수집에 참여하고 있다.[2)]

<표 13> 수집 네트워크: 수집 참여 기관

구분	수집 네트워크 유형	수집 기관
국내	한국어학당	○ 한양대 국제교육원 ○ 건국대학교 언어교육원 ○ 계명대학교 한국어학당 ○ 한국외국어대학교 한국어문화교육원
	대학(원)	○ 성신여자대학교 ○ 계명대학교

2) 수집 네트워크의 확대를 위하여 국립국어원-한국국제교류재단이 업무 협약을 체결하고 학습자 말뭉치 자료 수집을 적극 지원하기로 합의하였다. 이에 따라 교원을 파견하는 국외 대학에 공문과 함께 수집 안내문을 발송하였으며, 아직 많지는 않으나 수집 참여 의사를 밝혀 오거나 문의를 해 오는 교원들이 있었다. 그 외에도 국립국어원에서 2022년 8월에 열린 2022 EAKLE(유럽한국어교육자협회) 참가자와, 벨기에 브뤼셀 한글학교와 세종학당, 룩셈부르크 세종학당 교원들을 대상으로 학습자 말뭉치를 소개하고, 수집 참여 의사가 있는 교원 28명의 명단을 구축팀에 제공해 주었다. 구축팀에서는 수집 참여에 대한 상세한 내용을 안내하고 수집 협조를 요청하였다.

		○ 연세대학교 ○ 가톨릭대학교 ○ 한국학중앙연구원
	기타	○ 국내 대학(원) 유학생 모임 ○ 연구진의 인적 네트워크(한국어 교원) ○ 포털사이트의 한국어 교원 커뮤니티
국외	한국국제교류재단	○ 아프리카 코트 이브와르 펠릭스우푸에 부와니 대학교 ○ 슬로바키아 코메니우스 대학교 한국학과 ○ 인도네시아 가자마다 대학교 한국언어문화학과 ○ 불가리아 소피아 대학교 ○ 슬로베니아 류블랴나 대학교
	대학	○ 태국 부라파대학 ○ 이탈리아 나폴리 오리엔탈레 국립대학교 ○ 이탈리아 시에나 외국인대학
	세종학당	○ 미국 오번 세종학당

○ 다음은 국내 포털 사이트의 한국어 (예비) 교원 커뮤니티로, 대상 커뮤니트 목록을 수집한 후 회원들의 활동 현황을 파악하여 활동이 활발한 기관을 대상으로 학습자 말뭉치에 대해 소개하고 자율적인 수집 참여를 독려하였다. 이는 즉각적인 수집 효과로 이어지지는 못하고 있으나 대규모 수집을 위해 다양한 참여자들의 자율적인 수집을 유도하려는 노력이 필요하다고 판단된다.

<표 14> 수집 네트워크: 한국어 (예비) 교원 커뮤니티

소속/기관	커뮤니티명	회원 수	성격
네이버 카페	국제한국어교육자협회	9,245	현직 한국어 교사들이 가장 많이 가입하고 양질의 정보 공유 가능.
	대학원 입학을 준비하는 사람들의 모임	201,421	대학원 진학을 원하는 사람들의 카페. 한국어 교육 분야로 진학하려는 사람들도 많이 가입되어 있음.

	한국어교원양성과정	42,711	양성과정이나 대학원 진학을 목표로 한 학생들이 많음.
	한도사-한국어교원자격증/한국어교사강사/양성과정/TOPIK	5,532	한국어교원자격증 취득을 원하는 이들이 주로 가입, 졸업 후 전망이나 취업 정보도 활발히 공유되고 있음.
	학은모-학점은행제	210,220	학점은행제로 한국어 교원자격증 취득하려는 이들도 많이 보임.
	연세 한국어교사 양성과정	1,602	연세어학당에서 교사 양성과정을 마친 사람들의 카페.
	한교사 - 한국어교원/한국어강사/한국어교사/채용정보/자격증	1,633	한국어 교사가 되려는 이들을 위한 카페. 활동이 활발하지는 않은 상태.
다음 카페	한국어 교실	17,038	한국어를 배우고 가르치는 사람들을 위한 공간. 활동은 미미한 편이지만 관리자가 꾸준히 관리함.(홍보 가능)
	한국어와 한국어 교원	2,121	활동이 저조한 편이지만 관리자가 한국어 관련 뉴스를 꾸준히 업데이트함.

(2) 수집 방법

○ 자료 수집 방법은 직접 수집과 온라인 기반 수집의 두 가지 방식으로 진행되고 있다.

- 직접 수집: 이메일, 클라우드를 기반으로 수집 교사와 직접 소통
- 온라인 기반 수집: 누리집을 통한 수집 안내 및 자료 제출

○ 이 중 온라인 기반 수집은 2022년 본 연구에서 처음 시도되는 것으로 향후 대규모 수집을 위해 학습자의 자율적인 참여를 유도하기 위해 시범 운영을 하고 있다(한국어 학습자 말뭉치 수집 누리집 https://klcorpus.com/).

<그림 1> 학습자 말뭉치 수집 누리집 첫 화면

○ ‘한국어 학습자 말뭉치 수집 누리집’의 페이지 구성 및 주요 기능은 다음과 같다.

<표 15> ‘학습자 말뭉치 수집 누리집’의 구성과 주요 기능

구성	페이지 수	기능
학습자 말뭉치 소개	1	○ 학습자 말뭉치 소개
어떻게 참여할까요?	1	○ 자료 수집 참여 방법 안내
무엇에 대해 쓰고 말할까요?	1	○ 자료 수집 과제 소개 ○ 과제 파일 다운로드 링크 제공
제출하기	1	○ 리스트 게시판: 메시지, 파일 업로드(학습자 동의서, 수집 자료 파일을 각각 업로드하도록 구성) ○ 비밀보기 기능: 게시 자료 보기와 다운로드 권한 제한(글쓴이, 관리자)
	1	○ 피드백하기: 메시지, 파일 업로드 ○ 게시판 구조와 비밀보기 기능은 [제출하기]와 동일

공지 사항	1	○ 리스트 게시판

(3) 수집 과제

○ 수집 과제는 대규모 수집 과제와 기획 수집 과제로 구분된다. 대규모 수집 과제는 구축팀에서 제안하지 않은 장르와 주제의 자료로 수집 기관의 교육과정이나 수업 활동과 관련된 과제, 시험, 말하기/쓰기 대회 등을 말한다. 이와 달리 기획 수집 과제는 앞선 단계에서 분석한 2015-2021년 기 구축 말뭉치의 장르와 주제를 토대로 구축팀에서 설정한 장르와 주제의 과제를 말한다.

① 문어

○ 기획 말뭉치의 문어 수집 장르와 주제는 다음과 같다.

<표 16> 문어 기획 말뭉치 수집 장르와 주제

장르	주제
보고서	○ 대학(원) 수업에서의 산출 자료
설명문	○ 1급: 가족 소개, 친구 소개, 고향 소개(계절, 날씨), 하루 일과, 여행 장소 소개 ○ 5급: 혈액형별 성격의 특징, 건강 관리법, 고향의 명소, 고향의 역사적 사건, 도시 소개, 문화 차이, 생활 속 과학, 자국의 문화와 풍습, 정책, 교육관, 한국과 자국의 언어 차이, 한국과 자국의 대중문화 비교 ○ 6급: 문화유산 소개, 주거 형태, 사회적 사건, 대중매체의 기능, 여성의 경제 활동이 사회에 미치는 영향, 교육 문제, 지구 온난화의 원인
논설문	○ 5급, 6급: 인터넷, 로봇, 인공지능, 저출산, 고령화, 1인 가구

○ 세부 과제는 수집 기관의 학습자 수준과 특성을 반영하여 위의 주제들을 활용할 수 있도록 하였다. 다음은 기획 과제의 기본 모형으로 2021년부터

수집 과제로 활용해 오고 있다.

<표 17> 기획 말뭉치 수집 과제: 문어

수준	주제	장르
초급	주제 1. 자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, ……) 주제 2. 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

○ 기획 말뭉치의 구어 수집 장르와 주제는 다음과 같다.

<표 18> 구어 기획 말뭉치 수집 장르와 주제

장르	주제
발표	○ 대학(원) 수업에서의 산출 자료
자유 대화	○ 일상생활과 관련된 자유 주제의 대화
내러티브	○ 간단한 자기소개 ○ 과거 이야기: 태어난 곳, 고향 소개, 어릴 때 성격, 기억나는 친구, 기억나는 일, 어릴 때 꿈 ○ 현재 이야기: 사는 곳, 성격, 좋아하는 것과 싫어하는 것, 하는 일, 꿈, 진로 ○ 미래 이야기: 앞으로 10년 후의 내 모습, 노후의 삶, 죽기 전에 꼭 하고 싶은 일

○ 이 중 내러티브 수집 과제의 모형을 보이면 다음과 같다.

<표 19> 기획 말뭉치 수집 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

2) 원시 말뭉치 구축

○ 원시 말뭉치 구축은 문어 입력과 구어 전사 작업을 중심으로 이루어진다. 문어 입력 및 구어 전사 말뭉치 구축 지원 도구인 LCMS에 학습자 동의서와 자료의 원본을 등록한 후 지침에 따라 작업자가 입력/전사 작업을 한 후 3단계의 검수를 거쳐 구축이 완료된다. 문어 입력 및 구어 전사 현황은 다음과 같다.

(1) 문어

○ 2022년 원시 문어 말뭉치는 708,096어절이 구축되었다. 그 결과 2015년부터 2021년까지 구축한 말뭉치와 합산하여 누적 합계 4,407,583어절 규모의 문어 원시 말뭉치가 구축되었다.

① 수준별 자료 분포

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

<표 20> 문어 원시 말뭉치의 수준별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 – 2021	어절 수	439,612	596,563	690,248	650,251	722,627	459,546	140,640	3,699,487
	파일 수	6,690	6,134	5,796	4,977	4,662	2,802	141	31,202
2022	어절 수	155,083	145,112	136,285	159,864	66,767	42,309	2,676	708,096
	파일 수	2,060	1,455	987	1,016	365	205	10	6,098
합계	어절 수	594,695	741,675	826,533	810,115	789,394	501,855	143,316	4,407,583
	파일 수	8,750	7,589	6,783	5,993	5,027	3,007	151	37,300

② 언어권별 자료 분포

○ 문어 원시 말뭉치는 141개국[3] 98개 언어권의 자료가 구축되었다. 다음은 문어 원시 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

---

3) 141개국에는 가나, 가봉, 과테말라, 그루지아, 그리스, 나이지리아, 남수단, 남아프리카, 네덜란드, 네팔, 노르웨이, 뉴질랜드, 니카라과, 대만, 덴마크, 도미니카 공화국, 도미니카 연방, 독일, 동티모, 라오스, 라이베리아, 라트비아, 러시아, 레바논, 루마니아, 룩셈부르그, 르완다, 리비아, 리투아니아, 마다가스카르, 마카오, 말레이시아, 멕시코, 모로코, 모잠비크, 몰도바, 몽골, 미국, 미얀마, 바레인, 바베이도스, 방글라데시, 베냉, 베네수엘라, 베트남, 벨기에, 벨라루스, 보츠와나, 볼리비아, 부룬디, 북한, 불가리아, 브라질, 브루나이, 사우디아라비아, 세네갈, 세르비아, 수단, 스리랑카, 스웨덴, 스위스, 스페인, 슬로바키아, 슬로베니아, 시리아, 싱가포르, 아랍에미리트연합, 아르메니아, 아르헨티나, 아이슬란드, 아일랜드, 아제르바이잔, 아프카니스탄, 알바니아, 알제리, 앙골라, 에스토니아, 에콰도르, 엘살바도르, 영국, 예멘, 오만, 오스트리아, 온두라스, 요르단, 우간다, 우루과이, 우즈베키스탄, 우크라이나, 이디오피아, 이라크, 이란, 이스라엘, 이집트, 이탈리아, 인도, 인도네시아, 일본, 자메이카, 잠비아, 저지, 중국, 짐바브웨, 체코, 칠레, 카메룬, 카자흐스탄, 카타르, 캄보디아, 캐나다, 케냐, 코스타리카, 코트디부아르, 콜롬비아, 콩고, 콩고 민주 공화국, 쿠바, 쿠웨이트, 키르기스스탄, 타지키스탄, 탄자니아, 태국, 터키, 튀니지아, 트르크메니스탄, 트리니다드 토바고, 파나마, 파라과이, 파키스탄, 팔레스타인, 페루, 포르투칼, 폴란드, 프랑스, 피지, 핀란드, 필리핀, 한국, 헝가리, 호주, 홍콩이 포함되어 있다.

4) 기타에는 크메르어, 페르시아어, 네팔어, 벵골어, 노르웨이어, 네덜란드어, 라오어, 힌디어, 투르크멘어, 헝가리어, 우르두어, 폴란드어, 아제르바이잔어, 우크라이나어, 타밀어,

<표 21> 문어 원시 말뭉치의 언어권별 자료 분포

언어권	2015-2021		2022		합계	
	어절 수	파일수	어절 수	파일수	어절 수	파일수
중국어	1,721,599	13,845	86,439	671	1,808,038	14,516
일본어	516,104	4,312	133,113	1,032	649,217	5,344
베트남어	252,924	2,406	267,274	2,643	520,198	5,049
영어	241,269	2,162	43,396	402	284,665	2,564
광둥어	184,229	1,544	50,139	352	234,368	1,896
러시아어	129,235	1,201	19,667	169	148,902	1,370
타이어	97,309	758	26,132	221	123,441	979
몽골어	73,168	695	22,098	187	95,266	882
스페인어	57,047	536	19,843	162	76,890	698
인도네시아어	46,551	385	2,275	12	48,826	397
프랑스어	43,393	395	3,981	32	47,374	427
말레이어	28,516	191	1,340	8	29,856	199
스웨덴어	25,993	289	1,724	11	27,717	300
카자흐어	22,878	186	341	2	23,219	188
아랍어	20,648	215	750	5	21,398	220
이탈리아어	19,839	143	1,469	8	21,308	151
독일어	16,311	153	3,276	23	19,587	176
버마어	8,962	78	10,009	55	18,971	133
우즈베크어	16,664	158	2,186	14	18,850	172
싱할라어	17,670	106	586	3	18,256	109
한국어	16,361	64	318	2	16,679	66
포르투갈어	13,919	126	2,506	24	16,425	150
타갈로그어	13,140	170	1,880	12	15,020	182
터키어	11,529	97			11,529	97
키르기스어	10,584	84	542	3	11,126	87
기타[4]	93,645	903	6,812	45	100,457	948
합계	3,699,487	31,202	708,096	6,098	4,407,583	37,300

스와힐리어, 루마니아어, 암하라어, 핀란드어, 타지크어, 불가리아어, 르완다어, 덴마크어,

③ 수준별·언어권별 자료 분포

가. 2022년 신규 구축

○ 2022년 구축된 수준별·모국어별 분포는 다음과 같다.

<표 22> 2022년 원시 말뭉치 신규 구축 현황: 문어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
베트남어	어절 수	74,706	49,020	52,500	67,379	17,858	5,811	-	267,274
	파일 수	1,117	551	418	427	103	27	-	2,643
일본어	어절 수	11,190	33,010	22,044	39,418	18,238	8,590	623	133,113
	파일 수	143	335	150	258	100	43	3	1,032
중국어	어절 수	27,602	11,751	10,748	10,873	6,056	18,001	1,408	86,439
	파일 수	316	100	66	65	33	88	3	671
광둥어	어절 수	3,903	10,422	14,181	10,554	7,962	3,117	-	50,139
	파일 수	52	96	92	59	39	14	-	352
영어	어절 수	8,893	12,168	11,303	7,871	2,773	388	-	43,396
	파일 수	112	127	89	57	15	2	-	402
타이어	어절 수	9,459	6,585	5,131	3,601	1,116	240	-	26,132
	파일 수	101	57	32	24	6	1	-	221
몽골어	어절 수	6,551	4,052	3,418	3,776	3,065	969	267	22,098
	파일 수	80	31	27	24	18	5	2	187
스페인어	어절 수	4,181	5,206	6,105	2,835	1,516	-	-	19,843
	파일 수	47	51	38	16	10	-	-	162
러시아어	어절 수	3,582	3,847	4,181	5,324	1,438	1,062	233	19,667
	파일 수	48	42	34	32	7	5	1	169

쿠르드어, 테툼어, 아르메니아어, 위구르어, 카탈루냐어, 간다어, 슬로바키아어, 이그보어, 세르비아어, 히브리어, 체코어, 중국어(만다린어), 자바어, 룩셈부르크어, 칸나다어, 슬로베니아어, 노르웨이어(뉘노르스크), 에스토니아어, 니안콜어, 마다가스카르어, 리투아니아어, 펀자브어, 그리스어, 웨일스어, 조지아어, 쇼나어, 마라티어, 하우사어, 티베트어, 말라얄람어, 텔루구어, 벨라루스어, 딩카어, 티그리냐어, 바슈키르어, 트위어, 보스니아어, 마오리어, 알바니아어, 구자라트어, 피지어, 월로프어, 룬디어, 라트비아어, 세부아노어, 파슈토어, 아이슬란드어, 츠와나어, 판티어, 아프리칸스어가 포함되어 있다.

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
버마어	어절 수	619	1,288	2,586	3,069	1,530	917	-	10,009
	파일 수	5	8	13	16	8	5	-	55
기타	어절 수	4,397	7,763	4,088	5,164	5,215	3,214	145	29,986
	파일 수	39	57	28	38	26	15	1	204
합계	어절 수	155,083	145,112	136,285	159,864	66,767	42,309	2,676	708,096
	파일 수	2,060	1,455	987	1,016	365	205	10	6,098

나. 2015-2022년 누적 구축

○ 2015년에서 2022년까지 누적 구축된 수준별·모국어별 분포는 다음과 같다.

<표 23> 2015-2022년 원시 말뭉치 누적 구축 현황: 문어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	어절 수	227,783	241,784	285,454	298,121	396,905	251,004	106,987	1,808,038
	파일 수	3,430	2,406	2,299	2,234	2,526	1,529	92	14,516
일본어	어절 수	51,275	119,464	132,526	147,825	118,864	78,184	1,079	649,217
	파일 수	703	1,238	1,080	1,093	738	486	6	5,344
베트남어	어절 수	115,089	98,755	105,413	112,872	58,643	23,430	5,996	520,198
	파일 수	1,742	1,126	902	781	362	133	3	5,049
영어	어절 수	40,576	59,824	59,823	49,969	38,174	32,785	3,514	284,665
	파일 수	597	622	515	395	249	177	9	2,564
광둥어	어절 수	17,227	39,128	49,989	48,046	42,047	37,931	-	234,368
	파일 수	262	379	400	354	267	234	-	1,896
러시아어	어절 수	16,589	25,152	32,740	30,441	27,418	11,795	4,767	148,902
	파일 수	260	306	296	244	184	71	9	1,370
타이어	어절 수	25,188	35,560	27,371	16,187	10,725	8,089	321	123,441
	파일 수	289	275	178	114	74	47	2	979
몽골어	어절 수	17,437	17,631	19,311	17,233	15,857	7,484	313	95,266
	파일 수	241	186	165	128	109	50	3	882
스페인어	어절 수	13,313	20,220	21,343	11,937	7,632	2,152	293	76,890

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
	파일 수	179	203	165	84	52	14	1	698
인도네시아어	어절 수	7,012	8,751	7,234	10,511	8,487	5,734	1,097	48,826
	파일 수	103	85	54	64	54	36	1	397
프랑스어	어절 수	11,235	9,501	10,371	5,490	5,893	3,951	933	47,374
	파일 수	150	89	81	42	38	26	1	427
말레이어	어절 수	1,816	4,309	9,217	7,692	3,263	1,083	2,476	29,856
	파일 수	31	44	60	40	14	7	3	199
스웨덴어	어절 수	5,947	6,832	8,270	3,674	1,779	1,215	-	27,717
	파일 수	103	73	71	32	12	9	-	300
카자흐어	어절 수	1,984	3,841	5,243	4,850	4,564	2,737	-	23,219
	파일 수	32	35	39	38	30	14	-	188
아랍어	어절 수	3,603	4,216	4,554	3,278	3,347	1,882	518	21,398
	파일 수	59	46	44	28	25	15	3	220
이탈리아어	어절 수	3,118	2,231	2,093	2,564	3,759	2,802	4,741	21,308
	파일 수	51	21	21	22	19	16	1	151
독일어	어절 수	4,475	2,755	2,863	3,965	3,705	1,679	145	19,587
	파일 수	53	27	26	31	26	12	1	176
버마어	어절 수	1,687	2,797	3,466	3,770	4,302	2,949	-	18,971
	파일 수	22	24	23	24	26	14	-	133
우즈베크어	어절 수	3,590	2,360	3,818	3,479	3,942	1,493	168	18,850
	파일 수	52	29	35	26	22	7	1	172
싱할라어	어절 수	1,710	3,795	3,206	2,432	2,626	852	3,635	18,256
	파일 수	20	27	24	17	15	5	1	109
한국어	어절 수	496	198	753	787	2,737	9,976	1,732	16,679
	파일 수	9	2	6	5	19	23	2	66
포르투갈어	어절 수	2,291	4,209	2,605	2,788	2,831	1,701	-	16,425
	파일 수	34	37	25	23	20	11	-	150
타갈로그어	어절 수	3,080	4,200	3,527	2,731	1,029	389	64	15,020
	파일 수	46	61	39	28	6	1	1	182

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
터키어	어절 수	862	1,638	2,436	1,654	3,105	1,431	403	11,529
	파일 수	14	18	22	12	22	8	1	97
키르기스어	어절 수	787	2,625	1,054	2,152	3,170	942	396	11,126
	파일 수	12	23	6	15	22	7	2	87
기타	어절 수	16,525	19,899	21,853	15,667	14,590	8,185	3,738	100,457
	파일 수	256	207	207	119	96	55	8	948
합계	어절 수	594,695	741,675	826,533	810,115	789,394	501,855	143,316	4,407,583
	파일 수	8,750	7,589	6,783	5,993	5,027	3,007	151	37,300

(2) 구어

○ 2022년 원시 구어 말뭉치는 300,219어절이 구축되었다. 그 결과 2015년부터 2021년까지 구축한 말뭉치와 합산하여 누적 합계 1,823,007어절 규모의 구어 원시 말뭉치가 구축되었다.

① 수준별 자료 분포

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

<표 24> 구어 원시 말뭉치의 수준별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 – 2021	어절 수	277,908	313,967	406,673	246,126	139,498	107,877	30,739	1,522,788
	파일 수	833	668	690	519	245	166	21	3,142
2022	어절 수	19,569	34,875	55,805	81,511	77,199	29,140	2,120	300,219
	파일 수	75	91	103	154	115	38	2	578
합계	어절 수	297,477	348,842	462,478	327,637	216,697	137,017	32,859	1,823,007
	파일 수	908	759	793	673	360	204	23	3,720

② 언어권별 자료 분포

○ 구어 원시 말뭉치는 90개국[5] 51개 언어권의 자료가 구축되었다. 다음은 구어 원시 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 25> 구어 원시 말뭉치의 언어권별 자료 분포

언어권	2015-2021		2022		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	345,244	871	45,060	88	390,304	959
베트남어	227,830	456	116,225	200	344,055	656
타이어	294,310	443	2,648	4	296,958	447
일본어	142,524	288	76,269	153	218,793	441
스페인어	95,395	181	6,795	16	102,190	197
러시아어	79,083	214	5,024	9	84,107	223
영어	56,329	109	9,173	19	65,502	128
인도네시아어	60,164	138	901	2	61,065	140
타갈로그어	45,223	90	-	-	45,223	90
싱할라어	30,505	52	393	1	30,898	53
몽골어	15,339	46	8,779	14	24,118	60
버마어	22,052	26	751	1	22,803	27
키르기스어	19,478	38	877	1	20,355	39
우즈베크어	14,430	28	2,600	6	17,030	34
프랑스어	4,896	11	5,685	17	10,581	28
기타[6]	69,986	151	19,039	47	89,025	198

5) 90개 국가에는 가나, 과테말라, 나이지리아, 남수단, 네덜란드, 네팔, 뉴질랜드, 대만, 도미니카 공화국, 독일, 동티모, 러시아, 루마니아, 르완다, 말레이시아, 말리, 멕시코, 모로코, 모리타니, 몽골, 미국, 미국령 사모아, 미얀마, 방글라데시, 베네수엘라, 베트남, 벨기에, 벨라루스, 보츠와나, 볼리비아, 불가리아, 브라질, 사우디아라비아, 세르비아, 소말리아, 수단, 스리랑카, 스웨덴, 스위스, 스페인, 싱가포르, 아랍에미리트연합, 아르메니아, 아르헨티나, 아제르바이잔, 아프카니스탄, 알제리, 에스토니아, 에콰도르, 엘살바도르, 영국, 예멘, 오만, 오스트리아, 요르단, 우간다, 우루과이, 우즈베키스탄, 우크라이나, 이디오피아, 이집트, 이탈리아, 인도네시아, 일본, 중국, 칠레, 카자흐스탄, 캄보디아, 캐나다, 케냐, 코스타리카, 코트디부아르, 콜롬비아, 키르기스스탄, 타지키스탄, 태국, 터키, 튀니지아, 파나마, 파라과이, 파키스탄, 페루, 포르투칼, 폴란드, 프랑스, 필리핀, 한국, 헝가리, 호주, 홍콩이 포함되어 있다.

6) 기타에는 독일어, 프랑스어, 러시아어, 스웨덴어, 타이어, 우즈베크어, 이탈리아어, 광둥

언어권	2015-2021		2022		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
합계	1,522,788	3,142	300,219	578	1,823,007	3,720

③ 수준별·모국어별 자료 분포

가. 2022년 신규 구축

○ 2022년 구축된 수준별·모국어별 분포는 다음과 같다.

<표 26> 2022년 원시 말뭉치 신규 구축 현황: 구어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
베트남어	어절 수	67	16,792	25,777	26,483	45,059	2,047	-	116,225
	파일 수	1	40	44	44	68	3	-	200
일본어	어절 수	3,602	3,028	10,361	26,076	14,841	17,495	866	76,269
	파일 수	17	11	19	57	24	24	1	153
중국어	어절 수	3,217	5,459	12,997	14,937	5,866	2,584	-	45,060
	파일 수	12	13	24	28	8	3	-	88
영어	어절 수	1,440	1,131	2,286	544	3,772	-	-	9,173
	파일 수	6	2	6	1	4	-	-	19
몽골어	어절 수	-	318	202	4,134	1,782	2,343	-	8,779
	파일 수	-	1	1	7	3	2	-	14
스페인어	어절 수	3,132	913	1,273	-	-	1,477	-	6,795
	파일 수	8	3	3	-	-	2	-	16
기타	어절 수	8,111	7,234	2,909	9,337	5,879	3,194	1,254	37,918
	파일 수	31	21	6	17	8	4	1	88
합계	어절 수	19,569	34,875	55,805	81,511	77,199	29,140	2,120	300,219
	파일 수	75	91	103	154	115	38	2	578

어, 타지크어, 인도네시아어, 키르기스어, 네팔어, 버마어, 말레이어, 딩카어, 포르투갈어, 우르두어, 폴란드어, 싱할라어, 아랍어, 우크라이나어가 포함되어 있다.

나. 2015-2022년 누적 구축

○ 2015년에서 2022년까지 누적 구축된 수준별·모국어별 분포는 다음과 같다.

<표 27> 2015-2022년 원시 말뭉치 누적 구축 현황: 구어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	어절 수	53,156	73,225	74,052	68,349	45,418	54,992	21,112	390,304
	파일 수	249	221	167	156	79	72	15	959
베트남어	어절 수	50,236	64,064	74,706	76,856	66,914	10,502	777	344,055
	파일 수	120	133	139	136	109	17	2	656
타이어	어절 수	69,918	47,932	153,426	14,899	6,040	4,743	-	296,958
	파일 수	164	71	166	28	9	9	-	447
일본어	어절 수	14,844	30,499	34,578	52,863	44,115	40,193	1,701	218,793
	파일 수	60	58	77	118	70	56	2	441
스페인어	어절 수	29,334	23,538	23,346	19,165	4,225	1,988	594	102,190
	파일 수	64	43	38	40	8	3	1	197
러시아어	어절 수	14,186	19,038	22,708	18,426	5,330	4,419	-	84,107
	파일 수	45	51	57	44	14	12	-	223
영어	어절 수	10,644	19,630	15,084	7,002	10,860	2,282	-	65,502
	파일 수	32	29	34	17	13	3	-	128
인도네시아어	어절 수	11,302	13,237	15,122	7,881	8,307	5,216	-	61,065
	파일 수	38	26	29	17	18	12	-	140
타갈로그어	어절 수	11,948	11,615	11,635	6,632	1,682	1,711	-	45,223
	파일 수	31	29	13	12	2	3	-	90
싱할라어	어절 수	6,204	5,306	6,539	6,133	3,726	2,990	-	30,898
	파일 수	10	11	10	11	6	5	-	53
몽골어	어절 수	2,046	3,407	5,031	7,229	4,062	2,343	-	24,118
	파일 수	10	11	13	16	8	2	-	60
버마어	어절 수	2,633	6,469	6,653	6,297	751	-	-	22,803
	파일 수	5	7	6	8	1	-	-	27
키르기스어	어절 수	1,469	4,525	5,440	5,063	3,858	-	-	20,355
	파일 수	5	9	8	10	7	-	-	39

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
우즈베크어	어절 수	1,946	7,706	2,637	4,075	666	-	-	17,030
	파일 수	6	14	5	8	1	-	-	34
프랑스어	어절 수	2,716	2,871	2,057	2,021	548	368	-	10,581
	파일 수	11	6	4	5	1	1	-	28
기타	어절 수	14,895	15,780	9,464	24,746	10,195	5,270	8,675	89,025
	파일 수	58	40	27	47	14	9	3	198
합계	어절 수	297,477	348,842	462,478	327,637	216,697	137,017	32,859	1,823,007
	파일 수	908	759	793	673	360	204	23	3,720

2. 2015-2021년 기구축 말뭉치 보완을 위한 기획 구축·가공

○ 한국어 학습자 말뭉치의 균형성 확보는 활용도 제고의 측면에서 2015년 사업이 시작된 이후 가장 중요한 쟁점이자 목표가 되어 왔다. 앞에서 제시한 분포 분석 결과에서 2015-2021년 구축 말뭉치는 대상별로는 국내의 학문 목적 학습자와 이주민, 국외 학습자 자료, 한국어 숙달도 수준에서는 1급과 2급, 5급, 6급, 6급 이상의 자료, 언어권별로는 일본어권, 영어권, 베트남어권, 타이어권, 스페인어권, 러시아어권 자료의 비중이 현저히 낮음을 확인하였다. 이에 따라 본 연구에서는 2022년 구축 말뭉치 중 일정 비율을 균형성 확보를 위한 기획 말뭉치에 배정하여 이들 자료 중 일부를 집중적으로 수집하였다.

2.1. 목표 규모 설정

○ 본 연구에서는 수요 기관과의 협의를 통해 형태 주석 말뭉치 30만 어절(문어 15만 어절, 구어 각 15만 어절), 오류 주석 말뭉치 20만 어절(문어 10만 어절, 구어 10만 어절)을 목표 규모로 설정하였다.

<표 28> 말뭉치 유형별 구축 목표

구분	문어			구어			합계		
	기구축	2022년	누적 합계	기구축	2022년	누적 합계	기구축	2022년	누적 합계
형태 주석	260만	15만	275만	110만	15만	125만	370만	30만	400만
오류 주석	55만	10만 (5만)	75만	60만	10만	70만	115만 (15만)	20만 (5만)	135만 (20만)

○ 이 중 오류 주석 말뭉치는 간편 주석(오류 위치, 교정 어절)을 기본으로 하되, 20만 어절 중 5만 어절에 한하여 정밀 주석을 하는 것으로 하였다. 정밀 주석은 교육 현장에서의 활용을 목적으로 한 것으로 2025년까지 수준별, 언어권별 균형을 최대한 맞추는 것을 목표로 한다.

2.2. 변인별 구축 목표 설정

○ 본 연구에서는 2차 중장기 계획에 따라 대규모 구축(수집 네트워크 및 학습자의 자율적 참여 기반)을 병행하되, 기구축 말뭉치 분석 결과를 토대로 대상별·수준별·언어권별·자료 변인별 자료를 집중적으로 구축해 나갈 수 있도록 구축 방향을 설정하였다. 이는 2025년까지 점진적으로 말뭉치의 균형성을 확대해 나가기 위한 것으로, 형태 주석 말뭉치와 오류 주석 말뭉치의 경우 100% 기획 구축을 하기로 하였다.

<표 29> 변인별 말뭉치 구축을 위한 기획 구축의 비중

구분	문어			구어			합계		
	대규모 구축 (전체)	기획 구축 (변인)	누적 합계 (전체)	대규모 구축 (전체)	기획 구축 (변인)	누적 합계	대규모 구축 (전체)	기획 구축 (변인)	누적 합계
형태 주석	-	15만 (100%)	15만 (100%)	-	15만 (100%)	15만 (100%)	-	40만 (100%)	40만 (100%)
오류 주석	-	10만 (100%)	10만 (100%)	-	10만 (100%)	10만 (100%)	-	20만 (100%)	20만 (100%)

○ 형태 주석 말뭉치와 오류 주석 말뭉치의 대상별, 수준별, 언어권별 변인에 따른 기본 구축 방향은 원시 말뭉치 구축을 위한 구축 방향에 따르되, 2015-2021년 기구축 말뭉치의 세부 구축 현황을 검토하여 가공 대상 말뭉치를 선정하였다.

2.3. 실제 구축·가공

○ 본 연구에서는 형태 주석 체계와 오류 주석 체계는 2015-2021년의 지침을 따르되, 오류 주석의 경우 정밀 주석과 간단 주석으로 이원화하여 진행하였다.

- 정밀 주석: 교정어절, 오류위치, 오류층위, 오류양상
- 간단 주석: 교정어절, 오류위치

1) 형태 주석 말뭉치 가공 현황

(1) 문어

○ 2022년 형태 주석 문어 말뭉치는 157,313어절이 구축되었다. 그 결과 2015년부터 2021년까지 구축한 말뭉치와 합산하여 누적 합계 2,760,085어절 규모의 문어 형태 주석 말뭉치가 구축되었다.

① 수준별 자료 분포

○ 문어 형태 주석 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되었다. 다음은 수준별 구축 규모를 집계한 것이다.

<표 30> 문어 형태 주석 말뭉치의 수준별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 – 2021	어절 수	381,081	433,472	447,834	421,654	443,722	407,971	67,038	2,602,772
	파일 수	5,569	4,399	3,725	3,300	2,991	2,594	62	22,640
2022	어절 수	-	47,075	52,975	41,103	13,557	2,603	-	157,313
	파일 수	-	491	467	313	59	5	-	1,335
합계	어절 수	381,081	480,547	500,809	462,757	457,279	410,574	67,038	2,760,085
	파일 수	5,569	4,890	4,192	3,613	3,050	2,599	62	23,975

② 언어권별 자료 분포

○ 문어 형태 주석 말뭉치는 133개국[7] 88개 언어권의 자료가 구축되었다. 다음은 문어 형태 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 31> 문어 형태 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2021		2022		합계	
	어절 수	파일 수	어절 수	파일 수	어절 수	파일 수
중국어	959,834	8,056	212	2	960,046	8,058
일본어	449,202	3,683	65,878	623	515,080	4,306
베트남어	232,624	2,245	20,153	160	252,777	2,405

7) 133개국에는 가나, 가봉, 과테말라, 그루지아, 그리스, 나이지리아, 남수단, 남아프리카, 네덜란드, 네팔, 노르웨이, 뉴질랜드, 니카라과, 대만, 덴마크, 도미니카 공화국, 도미니카 연방, 독일, 동티모, 라오스, 라이베리아, 러시아, 루마니아, 룩셈부르그, 르완다, 리비아, 리투아니아, 마다가스카르, 마카오, 말레이시아, 멕시코, 모로코, 모잠비크, 몰도바, 몽골, 미국, 미얀마, 바레인, 바베이도스, 방글라데시, 베네수엘라, 베트남, 벨기에, 벨라루스, 보츠와나, 볼리비아, 불가리아, 브라질, 브루나이, 사우디아라비아, 세네갈, 세르비아, 수단, 스리랑카, 스웨덴, 스위스, 스페인, 슬로바키아, 슬로베니아, 시리아, 싱가포르, 아랍에미리트연합, 아르메니아, 아르헨티나, 아이슬란드, 아일랜드, 아제르바이잔, 아프카니스탄, 알바니아, 알제리, 앙골라, 에스토니아, 에콰도르, 엘살바도르, 영국, 예멘, 오스트리아, 온두라스, 요르단, 우간다, 우루과이, 우즈베키스탄, 우크라이나, 이디오피아, 이라크, 이란, 이스라엘, 이집트, 이탈리아, 인도, 인도네시아, 일본, 자메이카, 잠비아, 저지, 중국, 체코, 칠레, 카메룬, 카자흐스탄, 카타르, 캄보디아, 캐나다, 케냐, 코스타리카, 코트디부아르, 콜롬비아, 콩고, 콩고 민주 공화국, 쿠바, 쿠웨이트, 키르기스스탄, 타지키스탄, 탄자니아, 태국, 터키, 튀니지아, 트르크메니스탄, 트리니다드 토바고, 파나마, 파라과이, 파키스탄, 팔레스타인, 페루, 포르투칼, 폴란드, 프랑스, 핀란드, 필리핀, 한국, 헝가리, 호주, 홍콩이 포함되어 있다.

영어	227,447	2,051	11,870	107	239,317	2,158
러시아어	107,008	991	21,683	209	128,691	1,200
타이어	74,760	660	22,549	98	97,309	758
광둥어	85,094	711	203	1	85,297	712
몽골어	67,313	633	-	-	67,313	633
스페인어	42,282	401	14,765	135	57,047	536
인도네시아어	43,952	374	-	-	43,952	374
프랑스어	40,098	373	-	-	40,098	373
말레이어	24,610	169	-	-	24,610	169
카자흐어	21,263	176	-	-	21,263	176
이탈리아어	17,946	125	-	-	17,946	125
아랍어	16,663	177	-	-	16,663	177
스웨덴어	16,228	206	-	-	16,228	206
싱할라어	16,065	96	-	-	16,065	96
우즈베크어	14,235	134	-	-	14,235	134
한국어	12,990	47	-	-	12,990	47
독일어	12,669	123	-	-	12,669	123
타갈로그어	11,216	152	-	-	11,216	152
기타[8]	109,273	1,057	-	-	109,273	1,057
합계	2,602,772	22,640	157,313	1,335	2,760,085	23,975

8) 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, 테툼어, 이그보어, 간다어, 룩셈부르크어, 세르비아어, 체코어, 슬로바키아어, 자바어, 슬로베니아어, 위구르어, 노르웨이어(뉘노르스크), 히브리어, 마다가스카르어, 칸나다어, 펀자브어, 조지아어, 중국어(만다린어), 마라티어, 티베트어, 말라얄람어, 에스토니아어, 그리스어, 텔루구어, 벨라루스어, 딩카어, 티그리냐어, 리투아니아어, 마오리어, 알바니아어, 구자라트어, 월로프어, 세부아노어, 파슈토어, 아이슬란드어, 츠와나어, 판테어, 아프리칸스어가 포함되어 있다.

③ 수준별·언어권별 분포

가. 2022년 신규 구축·가공

○ 2022년 구축된 수준별·모국어별 분포는 다음과 같다.

<표 32> 2022년 형태 주석 말뭉치 신규 작업 현황: 문어

모국어	구분	1급	2급	3급	4급	5급	6급	합계
일본어	어절 수	-	17,943	22,082	17,545	6,589	1,719	65,878
	표본 수	-	231	224	141	25	2	623
타이어	어절 수	-	12,018	9,129	859	213	330	22,549
	표본 수	-	54	37	5	1	1	98
러시아어	어절 수	-	5,909	6,459	8,370	945	-	21,683
	표본 수	-	76	63	66	4	-	209
베트남어	어절 수	-	6,167	3,909	6,641	3,085	351	20,153
	표본 수	-	72	36	38	13	1	160
스페인어	어절 수	-	5,038	5,268	3,359	1,100	-	14,765
	표본 수	-	58	45	24	8	-	135
영어	어절 수	-	-	6,128	4,117	1,625	-	11,870
	표본 수	-	-	62	37	8	-	107
중국어	어절 수	-	-	-	212	-	-	212
	표본 수	-	-	-	2	-	-	2
광둥어	어절 수	-	-	-	-	-	203	203
	표본 수	-	-	-	-	-	1	1
합계	어절 수	-	47,075	52,975	41,103	13,557	2,603	157,313
	표본 수	-	491	467	313	59	5	1,335

나. 2015-2022년 누적 구축·가공

○ 2015년에서 2022년까지 누적 구축된 수준별·모국어별 분포는 다음과 같다.

<표 33> 2015-2022년 형태 주석 말뭉치 누적 구축 현황: 문어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	어절 수	142,191	140,978	141,935	139,572	161,979	195,729	37,662	960,046
	표본 수	2,003	1,353	1,131	1,112	1,132	1,304	23	8,058
일본어	어절 수	40,085	86,454	110,482	108,407	100,522	69,130	-	515,080
	표본 수	560	903	930	835	637	441	-	4,306
베트남어	어절 수	40,383	49,735	52,913	45,346	40,785	17,619	5,996	252,777
	표본 수	625	575	484	353	259	106	3	2,405
영어	어절 수	31,683	47,656	48,520	42,098	35,202	32,397	1,761	239,317
	표본 수	485	495	426	338	233	175	6	2,158
러시아어	어절 수	13,007	21,305	28,559	25,117	25,980	10,733	3,990	128,691
	표본 수	212	264	262	212	177	66	7	1,200
타이어	어절 수	15,729	28,975	22,240	12,586	9,609	7,849	321	97,309
	표본 수	188	218	146	90	68	46	2	758
광둥어	어절 수	13,248	5,940	6,096	14,239	19,596	26,178	-	85,297
	표본 수	209	49	49	114	130	161	-	712
몽골어	어절 수	10,886	11,984	14,817	12,119	10,946	6,515	46	67,313
	표본 수	161	132	126	92	76	45	1	633
스페인어	어절 수	9,132	15,014	15,238	9,102	6,116	2,152	293	57,047
	표본 수	132	152	127	68	42	14	1	536
인도네시아어	어절 수	7,012	7,813	6,971	9,175	6,655	5,229	1,097	43,952
	표본 수	103	80	52	61	43	34	1	374
프랑스어	어절 수	10,344	8,074	9,660	4,267	4,220	3,533	-	40,098
	표본 수	141	76	75	31	26	24	-	373
말레이어	어절 수	1,639	4,309	7,772	5,815	1,907	880	2,288	24,610
	표본 수	29	44	51	31	6	6	2	169
카자흐어	어절 수	1,984	3,841	5,243	4,519	3,280	2,396	-	21,263
	표본 수	32	35	39	36	22	12	-	176

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
이탈리아어	어절 수	2,923	1,550	1,443	1,809	2,988	2,492	4,741	17,946
	표본 수	49	17	14	16	13	15	1	125
아랍어	어절 수	3,446	3,984	2,439	2,680	1,714	1,882	518	16,663
	표본 수	57	44	23	23	12	15	3	177
스웨덴어	어절 수	5,446	6,240	1,281	1,545	744	972	-	16,228
	표본 수	98	69	12	13	6	8	-	206
싱할라어	어절 수	1,710	3,554	2,799	2,016	1,499	852	3,635	16,065
	표본 수	20	26	21	14	9	5	1	96
우즈베크어	어절 수	3,590	1,644	2,851	2,375	2,114	1,493	168	14,235
	표본 수	52	20	25	19	10	7	1	134
한국어	어절 수	112	198	542	142	2,130	8,134	1,732	12,990
	표본 수	2	2	4	2	15	20	2	47
독일어	어절 수	2,776	2,028	1,414	2,935	2,213	1,303	-	12,669
	표본 수	41	20	13	22	17	10	-	123
타갈로그어	어절 수	2,816	3,685	1,643	2,525	483	-	64	11,216
	표본 수	43	55	24	26	3	-	1	152
기타	어절 수	20,939	25,586	15,951	14,368	16,597	13,106	2,726	109,273
	표본 수	327	261	158	105	114	85	7	1,057
합계	어절 수	381,081	480,547	500,809	462,757	457,279	410,574	67,038	2,760,085
	표본 수	5,569	4,890	4,192	3,613	3,050	2,599	62	23,975

(2) 구어

○ 2022년 형태 주석 구어 말뭉치는 151,477어절이 구축되었다. 그 결과 2015년부터 2021년까지 구축한 말뭉치와 합산하여 누적 합계 1,253,148어절 규모의 구어 형태 주석 말뭉치가 구축되었다.

① 수준별 자료 분포

○ 구어 형태 주석 말뭉치는 1급에서 6급, 그리고 6급 이상의 자료가 구축되

었다. 다음은 수준별 구축 규모를 집계한 것이다.

<표 34> 구어 형태 주석 말뭉치의 수준별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 – 2021	어절 수	214,522	205,527	210,319	210,050	136,266	103,875	21,112	1,101,671
	표본 수	702	443	460	431	240	160	15	2,451
2022	어절 수	21,318	53,933	52,051	23,453	-	722	-	151,477
	표본 수	53	113	111	59	-	1	-	337
합계	어절 수	235,840	259,460	262,370	233,503	136,266	104,597	21,112	1,253,148
	표본 수	755	556	571	490	240	161	15	2,788

② 언어권별 자료 분포

○ 구어 형태 말뭉치는 73개국[9] 37개 언어권의 자료가 구축되었다. 다음은 구어 형태 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 35> 구어 형태 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2021		2022		합계	
	어절 수	표본 수	어절 수	표본 수	어절 수	표본 수
중국어	240,677	574	77,374	189	318,051	763
베트남어	189,272	380	34,416	68	223,688	448
일본어	136,438	279	1,096	2	137,534	281
타이어	116,290	293	1,733	3	118,023	296
스페인어	50,657	104	30,567	64	81,224	168
러시아어	69,260	197	6,291	11	75,551	208

9) 73개국에는 가나, 과테말라, 나이지리아, 네덜란드, 네팔, 뉴질랜드, 대만, 도미니카 공화국, 독일, 러시아, 르완다, 말레이시아, 멕시코, 모로코, 몽골, 미국, 미얀마, 방글라데시, 베네수엘라, 베트남, 벨기에, 벨라루스, 보츠와나, 볼리비아, 불가리아, 브라질, 사우디아라비아, 세르비아, 소말리아, 스리랑카, 스웨덴, 스위스, 스페인, 싱가포르, 아랍에미리트 연합, 아르메니아, 아르헨티나, 아제르바이잔, 알제리, 에콰도르, 엘살바도르, 영국, 예멘, 오만, 요르단, 우간다, 우루과이, 우즈베키스탄, 이집트, 이탈리아, 인도네시아, 일본, 중국, 카자흐스탄, 캄보디아, 캐나다, 코스타리카, 콜롬비아, 키르기스스탄, 태국, 터키, 파나마, 파라과이, 파키스탄, 페루, 포르투갈, 폴란드, 프랑스, 필리핀, 한국, 헝가리, 호주, 홍콩이 포함되어 있다.

언어권	2015-2021		2022		합계	
	어절 수	표본 수	어절 수	표본 수	어절 수	표본 수
인도네시아어	57,612	134	-	-	57,612	134
영어	55,465	108	-	-	55,465	108
타갈로그어	44,595	89	-	-	44,595	89
싱할라어	30,505	52	-	-	30,505	52
버마어	19,399	22	-	-	19,399	22
키르기스어	18,372	36	-	-	18,372	36
우즈베크어	14,110	26	-	-	14,110	26
기타[10]	59,019	157	-	-	59,019	157
합계	1,101,671	2,451	151,477	337	1,253,148	2,788

③ 수준별·언어권별 자료 분포

가. 2022년 신규 구축·가공

○ 2022년 구축된 수준별·모국어별 분포는 다음과 같다.

<표 36> 2022년 형태 주석 말뭉치 신규 작업 현황: 구어

모국어	구분	1급	2급	3급	4급	5급	6급	합계
중국어	어절 수	9,778	23,031	25,722	18,121	-	722	77,374
	표본 수	33	54	55	46	-	1	189
베트남어	어절 수	921	17,093	15,907	495	-	-	34,416
	표본 수	1	33	33	1	-	-	68
스페인어	어절 수	-	-	1,096	-	-	-	1,096
	표본 수	-	-	2	-	-	-	2
러시아어	어절 수	-	1,733	-	-	-	-	1,733
	표본 수	-	3	-	-	-	-	3
타이어	어절 수	10,189	8,066	7,957	4,355	-	-	30,567
	표본 수	18	17	18	11	-	-	64

10) 기타는 아랍어, 몽골어, 카자흐어, 크메르어, 포르투갈어, 세부아노어, 이탈리아어, 프랑스어, 독일어, 광둥어, 네팔어, 스웨덴어, 네덜란드어, 세르비아어, 터키어, 벵골어, 아르메니아어, 불가리아어, 아제르바이잔어, 헝가리어, 우르두어, 말레이어, 폴란드어, 라틴어로 24개 언어권이 포함되어 있다.

일본어	어절 수	430	4,010	1,369	482	-	-	6,291
	표본 수	1	6	3	1	-	-	11
합계	어절 수	21,318	53,933	52,051	23,453	-	722	151,477
	표본 수	53	113	111	59	-	1	337

나. 2015-2022년 누적 구축·가공

○ 2015년에서 2022년까지 누적 구축된 수준별·모국어별 분포는 다음과 같다.

<표 37> 2015-2022년 형태 주석 말뭉치 누적 구축 현황: 구어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	어절 수	47,097	55,870	54,161	49,628	38,422	51,761	21,112	318,051
	표본 수	211	145	133	122	69	68	15	763
베트남어	어절 수	50,169	47,178	46,173	49,858	21,855	8,455	-	223,688
	표본 수	119	92	91	91	41	14	-	448
일본어	어절 수	11,242	27,394	24,217	26,787	27,462	20,432	-	137,534
	표본 수	43	46	58	61	44	29	-	281
타이어	어절 수	39,931	20,683	34,687	13,674	4,672	4,376	-	118,023
	표본 수	129	45	82	25	7	8	-	296
스페인어	어절 수	23,820	19,148	14,730	18,790	4,225	511	-	81,224
	표본 수	54	37	29	39	8	1	-	168
러시아어	어절 수	11,951	17,477	21,348	16,483	4,610	3,682	-	75,551
	표본 수	41	48	54	42	12	11	-	208
인도네시아어	어절 수	10,801	11,186	15,122	7,613	8,307	4,583	-	57,612
	표본 수	37	23	29	16	18	11	-	134
영어	어절 수	9,204	17,635	12,798	6,458	7,088	2,282	-	55,465
	표본 수	26	26	28	16	9	3	-	108
타갈로그어	어절 수	11,948	10,987	11,635	6,632	1,682	1,711	-	44,595
	표본 수	31	28	13	12	2	3	-	89
싱할라어	어절 수	6,204	5,306	6,539	5,740	3,726	2,990	-	30,505
	표본 수	10	11	10	10	6	5	-	52
버마어	어절 수	2,633	4,889	6,653	5,224	-	-	-	19,399

	표본 수	5	5	6	6	-	-	-	22
키르기스어	어절 수	1,469	3,648	4,334	5,063	3,858	-	-	18,372
	표본 수	5	8	6	10	7	-	-	36
우즈베크어	어절 수	1,470	7,345	1,434	3,195	666	-	-	14,110
	표본 수	4	13	2	6	1	-	-	26
기타	어절 수	7,901	10,714	8,539	18,358	9,693	3,814	-	59,019
	표본 수	40	29	30	34	16	8	-	157
합계	어절 수	235,840	259,460	262,370	233,503	136,266	104,597	21,112	1,253,148
	표본 수	755	556	571	490	240	161	15	2,788

2) 오류 주석 말뭉치 가공 현황

(1) 문어

○ 2022년 오류 주석 문어 말뭉치는 73,829어절이 구축되었다.[11] 그 결과 2015년부터 2021년까지 구축한 말뭉치와 합산하여 누적 합계 674,636어절 규모의 문어 오류 주석 말뭉치가 구축되었다.

① 수준별 자료 분포

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

<표 38> 문어 오류 주석 말뭉치의 수준별 자료 분포

구분		1급	2급	3급	4급	5급	6급	합계
2015 ~ 2021	어절 수	89,925	104,084	104,832	86,977	109,714	105,275	600,807
	표본 수	1,302	1,080	924	719	775	690	5,490
2022	어절 수	9,242	9,202	11,859	21,280	10,185	12,061	73,829

11) 이는 설계 단계에서 설정한 100,000어절에 약간 못 미치는 규모로 형태 주석이 완료된 문어 말뭉치 중 기획 구축 대상에 포함되는 수준 또는 언어권 자료가 부족함에 따른 것이다. 문어 말뭉치에서 구축하지 못한 분량은 구어 말뭉치에서 추가로 구축하였다.

	표본 수	85	105	130	212	102	118	752
합계	어절 수	99,167	113,286	116,691	108,257	119,899	117,336	674,636
	표본 수	1,387	1,185	1,054	931	877	808	6,242

② 언어권별 자료 분포

○ 문어 오류 주석 말뭉치는 85개국[12] 54개 언어권의 자료가 구축되었다. 다음은 문어 오류 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 39> 문어 오류 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2021		2022		합계	
	어절 수	표본 수	어절 수	표본 수	어절 수	표본 수
일본어	153,885	1,337	20,129	205	174,014	1,542
영어	153,254	1,408	8,771	85	162,025	1,493
중국어	113,259	1,012	44,612	460	157,871	1,472
베트남어	65,869	622	-	-	65,869	622
러시아어	31,162	300	-	-	31,162	300
타이어	24,734	201	-	-	24,734	201
스페인어	8,113	90	-	-	8,113	90
기타[13]	50,531	520	317	2	50,848	522
합계	600,807	5,490	73,829	752	674,636	6,242

12) 85개국에는 가나, 가봉, 과테말라, 나이지리아, 남수단, 남아프리카, 네덜란드, 네팔, 노르웨이, 뉴질랜드, 니카라과, 대만, 도미니카 공화국, 독일, 라오스, 러시아, 루마니아, 르완다, 리투아니아, 마다가스카르, 마카오, 말레이시아, 멕시코, 모로코, 몽골, 미국, 미얀마, 방글라데시, 베트남, 벨기에, 벨라루스, 볼리비아, 불가리아, 브루나이, 사우디아라비아, 세네갈, 세르비아, 스리랑카, 스웨덴, 스위스, 스페인, 슬로바키아, 시리아, 싱가포르, 아랍에미리트연합, 아르헨티나, 아일랜드, 아제르바이잔, 아프카니스탄, 에티오피아, 영국, 오스트리아, 요르단, 우간다, 우루과이, 우즈베키스탄, 우크라이나, 이란, 이집트, 이탈리아, 인도, 인도네시아, 일본, 자메이카, 잠비아, 중국, 카자흐스탄, 캄보디아, 캐나다, 케냐, 콜롬비아, 키르기스스탄, 태국, 터키, 파키스탄, 페루, 포르투갈, 폴란드, 프랑스, 핀란드, 필리핀, 한국, 헝가리, 호주, 홍콩이 포함된다.

13) 기타는 아랍어, 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, 세부아노어, 파슈토어, 리투아니아어, 라오어, 암하라어

③ 수준별·언어권별 자료 분포

가. 2022년 신규 구축·가공

○ 2022년 구축된 수준별·모국어별 분포는 다음과 같다.

<표 40> 2022년 오류 주석 말뭉치 신규 작업 현황: 문어

모국어	구분	1급	2급	3급	4급	5급	6급	합계
중국어	표본 수	4,641	5,102	6,483	6,457	10,185	11,744	44,612
	어절 수	50	55	69	68	102	116	460
일본어	표본 수	4,601	4,100	5,376	6,052	-	-	20,129
	어절 수	35	50	61	59	-	-	205
영어	표본 수	-	-	-	8,771	-	-	8,771
	어절 수	-	-	-	85	-	-	85
인도네시아어	표본 수	-	-	-	-	-	182	182
	어절 수	-	-	-	-	-	1	1
몽골어	표본 수	-	-	-	-	-	135	135
	어절 수	-	-	-	-	-	1	1
합계	표본 수	9,242	9,202	11,859	21,280	10,185	12,061	73,829
	어절 수	85	105	130	212	102	118	752

나. 2015-2022년 누적 구축·가공

○ 2015년에서 2022년까지 누적 구축된 수준별·모국어별 분포는 다음과 같다.

<표 41> 2015-2022년 오류 주석 말뭉치 누적 구축 현황: 문어

모국어	구분	1급	2급	3급	4급	5급	6급	합계
일본어	어절 수	22,969	22,654	23,634	25,219	39,619	39,919	174,014
	표본 수	272	244	225	225	288	288	1,542
영어	어절 수	25,189	31,816	30,834	26,512	25,651	22,023	162,025
	표본 수	365	326	265	221	171	145	1,493
중국어	어절 수	23,550	25,934	26,449	25,668	26,658	29,612	157,871
	표본 수	337	251	221	224	218	221	1,472

로 47개 언어권이 해당된다.

모국어	구분	1급	2급	3급	4급	5급	6급	합계
베트남어	어절 수	8,450	7,265	13,753	11,360	11,587	13,454	65,869
	표본 수	135	89	131	103	81	83	622
러시아어	어절 수	3,864	5,599	6,978	5,808	6,253	2,660	31,162
	표본 수	59	73	58	50	46	14	300
타이어	어절 수	4,408	6,502	3,268	3,678	3,844	3,034	24,734
	표본 수	58	52	23	25	26	17	201
스페인어	어절 수	2,348	2,513	1,536	1,247	152	317	8,113
	표본 수	33	31	15	8	1	2	90
기타	어절 수	8,389	11,003	10,239	8,765	6,135	6,317	50,848
	표본 수	128	119	116	75	46	38	522
합계	어절 수	99,167	113,286	116,691	108,257	119,899	117,336	674,636
	표본 수	1,387	1,185	1,054	931	877	808	6,242

○ 정밀 주석은 향후의 활용도를 고려하여 가장 규모가 많은 중국어권과 일본어권 학습자의 자료를 대상으로 하였으며, 2015부터 2021년까지 정밀 주석을 한 말뭉치와의 누적 규모를 토대로 균형성을 고려하여 선정하였다. 수준별·모국어별 분포는 다음과 같다.

<표 42> 2022년 오류 주석 말뭉치 정밀 주석 작업 현황

모국어	구분	1급	2급	3급	4급	5급	6급	합계
중국어	어절 수	4,641	4,946	6,574	6,457	10,075	11,114	43,807
	표본 수	50	54	70	68	101	112	455
일본어	어절 수	213	-	-	6,160	-	-	6,373
	표본 수	2	-	-	60	-	-	62
합계	어절 수	4,854	4,946	6,574	12,617	10,075	11,114	50,180
	표본 수	52	54	70	128	101	112	517

(2) 구어

○ 2022년 오류 주석 구어 말뭉치는 129,661어절이 구축되었다.[14] 그 결과 2015년부터 2021년까지 구축한 말뭉치와 합산하여 누적 합계 671,379어절 규모의 구어 오류 주석 말뭉치가 구축되었다.

① 수준별 자료 분포

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

<표 43> 구어 오류 주석 말뭉치의 수준별 자료 분포

구분		1급	2급	3급	4급	5급	6급	6급 이상	합계
2015 – 2021	어절 수	93,897	103,304	104,126	107,738	71,964	55,797	4,892	541,718
	표본 수	310	220	255	229	122	74	5	1,215
2022	어절 수	42,471	19,861	19,365	21,269	16,592	10,103	-	129,661
	표본 수	133	30	46	42	27	14	-	292
합계	어절 수	136,368	123,165	123,491	129,007	88,556	65,900	4,892	671,379
	표본 수	443	250	301	271	149	88	5	1,507

② 언어권별 자료 분포

○ 구어 오류 주석 말뭉치는 65개국[15] 35개 언어권[16]의 자료가 구축되었다.

14) 이는 설계 단계에서 설정한 100,000어절을 상회하는 규모로 당초 계획했던 문어 말뭉치의 비중이 줄면서 상대적으로 늘어난 것이다.

15) 65개국에는 가나, 나이지리아, 네덜란드, 네팔, 뉴질랜드, 대만, 독일, 러시아, 르완다, 말레이시아, 멕시코, 모로코, 몽골, 미국, 미얀마, 방글라데시, 베네수엘라, 베트남, 벨기에, 벨라루스, 보츠와나, 불가리아, 사우디아라비아, 세르비아, 소말리아, 스리랑카, 스웨덴, 스페인, 싱가포르, 아랍에미리트연합, 아르헨티나, 아제르바이잔, 알제리, 에콰도르, 영국, 예멘, 오만, 요르단, 우간다, 우루과이, 우즈베키스탄, 이집트, 이탈리아, 인도네시아, 일본, 중국, 카자흐스탄, 캄보디아, 캐나다, 코스타리카, 콜롬비아, 키르기스스탄, 태국, 터키, 파라과이, 파키스탄, 페루, 포르투갈, 폴란드, 프랑스, 필리핀, 한국, 헝가리, 호주, 홍콩이 포함된다.

16) 기타는 아랍어, 타갈로그어, 카자흐어, 키르기스어, 포르투갈어, 이탈리아어, 몽골어, 프랑스어, 우즈베크어, 버마어, 광둥어, 네팔어, 크메르어, 싱할라어, 독일어, 스웨덴어, 네

다음은 구어 오류 주석 말뭉치의 언어권별 자료 분포를 나타낸 것이다.

<표 44> 구어 오류 주석 말뭉치의 언어권별 자료 분포

언어권	2015-2021		2022		합계	
	어절 수	표본 수	어절 수	표본 수	어절 수	표본 수
중국어	127,247	301	27,796	57	155,043	358
일본어	112,550	246	16,862	25	129,412	271
베트남어	107,562	222	9,615	18	117,177	240
타이어	20,694	64	43,930	135	64,624	199
영어	43,081	88	7,930	14	51,011	102
스페인어	28,300	64	22,357	40	50,657	104
인도네시아어	30,606	65	-	-	30,606	65
러시아어	17,072	36	755	2	17,827	38
기타	54,606	129	416	1	55,022	130
합계	541,718	1,215	129,661	292	671,379	1,507

③ 수준별·언어권별 자료 분포

가. 2022년 신규 구축·가공

○ 2022년 구축된 수준별·모국어별 분포는 다음과 같다.

<표 45> 2022년 오류 주석 말뭉치 신규 작업 현황: 구어

모국어	구분	1급	2급	3급	4급	5급	6급	합계
타이어	어절 수	27,558	-	15,340	1,032	-	-	43,930
	표본 수	95	-	39	1	-	-	135
중국어	어절 수	4,473	3,571	-	4,936	7,018	7,798	27,796
	표본 수	19	7	-	8	12	11	57
스페인어	어절 수	5,842	3,688	1,069	8,880	2,878	-	22,357
	표본 수	10	4	1	20	5	-	40
일본어	어절 수	1,079	4,485	-	4,311	4,682	2,305	16,862

덜란드어, 세르비아어, 터키어, 벵골어, 불가리아어, 아제르바이잔어, 헝가리어, 우르두어, 말레이어, 폴란드어, 라틴어로 27개 언어권이 해당된다.

모국어	구분	1급	2급	3급	4급	5급	6급	합계
	표본 수	3	6	-	6	7	3	25
베트남어	어절 수	2,958	3,506	1,137	-	2,014	-	9,615
	표본 수	4	8	3		3	-	18
영어	어절 수	135	4,611	1,819	1,365	-	-	7,930
	표본 수	1	5	3	5	-	-	14
러시아어	어절 수	426	-	-	329	-	-	755
	표본 수	1	-	-	1	-	-	2
광둥어	어절 수	-	-	-	416	-	-	416
	표본 수	-	-	-	1	-	-	1
합계	어절 수	42,471	19,861	19,365	21,269	16,592	10,103	129,661
	표본 수	133	30	46	42	27	14	292

나. 2015-2022년 누적 구축·가공

○ 2015년에서 2022년까지 누적 구축된 수준별·모국어별 분포는 다음과 같다.

<표 46> 2015-2022년 오류 주석 말뭉치 누적 구축 현황: 구어

모국어	구분	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	어절 수	24,033	22,639	13,901	25,437	26,701	37,440	4,892	155,043
	표본 수	104	59	50	56	41	43	5	358
일본어	어절 수	11,157	27,394	22,251	24,192	23,986	20,432	-	129,412
	표본 수	42	46	55	58	41	29	-	271
베트남어	어절 수	30,435	21,528	17,007	28,639	14,068	5,500	-	117,177
	표본 수	68	43	38	54	27	10	-	240
타이어	어절 수	30,399	3,529	24,107	5,111	850	628	-	64,624
	표본 수	106	13	63	13	2	2	-	199
영어	어절 수	9,204	15,925	12,728	6,458	6,136	560	-	51,011
	표본 수	26	24	27	16	8	1	-	102
스페인어	어절 수	13,631	11,082	6,773	14,435	4,225	511	-	50,657
	표본 수	36	20	11	28	8	1	-	104
인도네시	어절 수	8,096	5,343	8,865	3,706	4,596	-	-	30,606

아어	표본 수	25	10	16	6	8	-	-	65
러시아어	어절 수	2,246	2,146	7,972	4,192	1,271	-	-	17,827
	표본 수	6	6	16	8	2	-	-	38
기타	어절 수	7,167	13,579	9,887	16,837	6,723	829	0	55,022
	표본 수	30	29	25	32	12	2	0	130
합계	어절 수	136,368	123,165	123,491	129,007	88,556	65,900	4,892	671,379
	표본 수	443	250	301	271	149	88	5	1,507

3. 한국어 학습자 말뭉치의 활용도 제고를 위한 한국어 모어 화자 참조 말뭉치 수집·구축

○ 한국어 모어 화자 참조 말뭉치는 학습자가 산출한 자료로부터 언어 사용의 특성을 포착하기 위한 참조 자료이다. 본 연구에서는 기구축 말뭉치의 수집 장르, 주제와 동일한 한국어 모어 화자의 발화 자료 3만 어절(30명, 표본 1개당 1,000어절 이하)을 수집하여 참조 말뭉치로서 시범 구축을 진행하였다. 이는 향후의 사업에서 참조 말뭉치 구축 확대 가능성을 탐색하기 위한 것이다.

3.1. 수집 과제 설계

○ 한국어 모어 화자 참조 말뭉치의 수집 과제는 한국어 학습자 말뭉치의 자료에 포함된 장르와 주제의 자료를 대상으로 하여 동질성을 갖추도록 해야 한다. 이에 본 연구에서는 2022년 기획 과제 중 설명문(문어)과 내러티브(구어)를 하나의 과제로 제시하여 참조 말뭉치를 수집하였다.

<표 47> 참조 말뭉치 수집 과제 및 수집 규모

자료 유형	표본 수
문어(설명문)	30개
구어(내러티브)	30개

3.2. 수집 네크워크

○ 한국어 모어 화자 참조 말뭉치는 표본 수와 규모가 많지 않기 때문에 연구진의 소속 기관인 연세대학교, 성신여자대학교, 계명대학교의 학부생 각 10명씩 총 30명을 대상으로 수집하였다.

3.3. 실제 수집 및 구축

○ 수집된 모어 화자 참조 말뭉치는 원시 말뭉치의 형태로 구축이 되었으며 문어 10,440어절, 구어 23,912어절이 수집되었다. 문어와 구어 30명의 평균 발화 수는 각각 348.0어절, 797.1어절로 개인별 편차가 매우 컸다. 수집 참여자별로 산출한 어절 수는 다음과 같다.

<표 48> 한국어 모어 화자 참조 말뭉치

화자 구분	문어	구어
모어 화자01	179	541
모어 화자02	354	729
모어 화자03	354	745
모어 화자04	228	673
모어 화자05	182	702
모어 화자06	356	1,418
모어 화자07	179	590
모어 화자08	195	708
모어 화자09	300	526
모어 화자10	536	790
모어 화자11	420	1,024
모어 화자12	340	1,361
모어 화자13	289	1,303
모어 화자14	286	956

모어 화자15	474	875
모어 화자16	451	1,592
모어 화자17	486	1,337
모어 화자18	344	942
모어 화자19	331	1,033
모어 화자20	393	474
모어 화자21	353	462
모어 화자22	411	374
모어 화자23	306	505
모어 화자24	384	538
모어 화자25	454	440
모어 화자26	465	498
모어 화자27	331	576
모어 화자28	428	898
모어 화자29	453	478
모어 화자30	178	824
합계	10,440	23,912

4. 수집·구축 대상 자료의 한국어 학습자 이용 허락 확보

○ 학습자 말뭉치는 학습자가 한국어를 학습하는 과정에서 산출한 자료이지만 작문이나 발화를 산출하는 과정에 학습자의 창의적인 활동이 포함된다는 점에서 저작물로서 인정될 여지가 있다. 그러나 창작을 목적으로 한 것이 아닌, 학습 과정에서의 수행 결과이고 그것이 학습자에게 경제적인 이익을 주지 않는다는 점에서 저작권 보호의 대상이 되는 완전한 저작물이라고 보기는 어려운 점이 있다. 이에 따라 2021년 기초 연구에서는 학습자 자료를 언어 자원인 말뭉치 구축을 위한 데이터로 간주하여 2015-2020년까지 사용한 학습자 말뭉치 동의서를 통해 자료 제공 및 이용 허락에 대한 학습자의 동의를 얻어 자료를 수집한 바 있다.

○ 본 연구에서는 2021년의 연구 결과에 따라 2015-2020년에 사용한 학습자 동의서를 통해 수집 및 구축 대상이 되는 자료에 대한 이용 허락을 확보하고 있다. 다음은 학습자 동의서에 포함된 조항으로, 간소화된 형식을 취하고 있지만 학습자의 이용 허락을 위해 명시해야 할 조항들이 모두 포함되어 있음을 확인할 수 있다.

<표 49> 한국어 학습자 말뭉치 동의서의 내용

구분	동의서 내용
사업 배경	국립국어원에서 한국어교육의 질적 향상을 위해 학습자들의 언어 자료(말뭉치)를 수집하여 활용하는 사업(사업 수행: 연세대학교 산학협력단)을 추진하고 있습니다.
이용 목적	여러분이 제공한 자료는 한국어 교수 방법 개선, 한국어 교재 개발, 한국어 교육 분야 및 인접 학문 분야의 연구에 사용됩니다.
참여자의 안전	이 연구에 참여하는 분들은 경제적인 손해나 신체적 위험이 없습니다.
철회 가능 여부	만약 참여를 원하지 않을 때에는 참여 의사를 철회할 수 있습니다.
개인 정보 보호 및 비밀 유지	또한 수집하는 개인 정보는 본 사업의 목적 외로는 사용되지 않으며, 비밀 유지를 위하여 식별할 수 없는 형태로 사용될 것입니다.
자료 제공 및 이용 허락	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 포함)
개인 정보 보호 및 비밀 유지	다음은 연구를 위한 자료로 활용될 정보입니다. 개인 신상 정보는 비밀이 보장되며 외부로 유출되지 않습니다. (가능하면 한국어로 응답해 주세요. 필요하면 영어를 사용해도 좋습니다.)
학습자 정보	성별, 출생년, 현재 등급, 국적, 제1 언어, 한국어 학습 기간, 한국에서의 거주 기간, 한국어 학습 목적, 직업, 한국어 외의 사용 가능 외국어

Ⅲ. 구축 학습자 말뭉치 검수 정교화 및 품질 관리

1. 구어·문어 입력 검수 등 검수 정교화 및 언어 자원 품질 확보

○ 한국어 학습자 말뭉치는 비정형성을 특징으로 하는 비모어 화자의 자료로, 광학식 문자 판독 장치(OCR: optical character reader/recognition)이나 음성을 인식하여 텍스트로 변환해 주는 STT(Speech to Text) 기술의 적용이 어렵다. 이에 따라서 40여 명에 이르는 연구 인력이 자료를 일일이 입력하고 전사를 한 후, 주석을 붙이는 작업을 수행하여 만들어지는 결과물이다. 이에 따라 문어 입력과 구어 전사, 형태소 주석, 오류 주석의 단계별로 심층 검수의 작업 공정을 마련하여 두고 있으나 미처 수정되지 못한 오류가 있을 수밖에 없다. 이에 본 연구에서는 자료의 정확성을 높이고 국가 자료로서의 신뢰성을 확보하기 위해 기구축 말뭉치의 작업 공정과 검수 체계를 살펴보고 한 단계 더 검수를 정교화하고 말뭉치의 품질을 제고하기 위한 방안을 마련하여 적용하였다.

1.1. 작업 공정에서의 3단계 검수 체계 유지와 내부 검수단 운영

○ 한국어 학습자 말뭉치는 문어 입력, 구어 전사, 형태 주석, 오류 주석의 각 구축 단계별로 작업, 검수, 심층 검수의 3차 검수를 거친다. 본 연구에서는 이러한 체계를 유지하되, 공동 연구원을 중심으로 한 작업자 외의 내부 검수단을 운영하고 있다. 내부 검수단은 주기적으로 표본을 임의로 추출하여 각 단계별 구축 자료의 무결성을 검증하고 수정 항목이 발견될 경우 구축 작업팀에 보고하여 수정 작업을 하도록 하였다.

1.2. 개별 표본의 표본정보 검수 체계 강화

○ 자료 구축의 첫 단계는 표본 등록에서 시작된다. 표본 등록 작업에서는 다양한 학습자 변인과 자료 변인을 기반으로 한 메타정보를 입력하게 되는데, 이 정보는 체계적인 말뭉치 구축을 위한 자료의 유형 정보는 물론 조건별 자료 검색을 위한 색인 기능을 하므로 가장 기본적이면서도 중요한 정보하고 할 수 있다. 본 연구에서는 자료의 내적 질을 높이고, 통계 정보의 정확성과 함께 표본 관리의 효율성을 높이기 위하여 각 표현의 표본 정보 검수 체계를 강화하였다. 표본 정보는 작업자가 1차 등록을 하며 검수자가 항목별 정보를 확인하는 과정을 거쳐 검수가 이루어지고 있다. 본 연구에서는 이러한 절차에 더하여 내부 검수단에서 표본을 임의로 추출하여 학습자 동의서와 표본 정보를 대조하여 표본 정보의 무결성을 검증하고 수정 항목이 발견될 경우 구축 작업팀에 보고하여 수정 작업을 하였다.

1.3. 작업 중 생성된 오조작 데이터 검증

○ 작업 중 생성된 오조작 데이터 검증은 작업 로그나 표본 정보, 주석 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. 본 연구에서는 다음과 같이 작업 진행 과정에서 이상이 발견되는 표본, 파일의 표본 정보와 파일 정보가 불일치하거나 기타 문제가 발견되는 표본들을 상시 추출하여 검수 작업을 진행하였다.

- 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 표본은 우선적인 검수 대상이 된다.
- 파일의 표본 정보와 파일 정보 간 이상 표본 검증: 파일의 표본 정보와 파일 정보 간 이상이 의심되는 표본을 추출하여 검증한다. 표본 등록 과정에서 해당 정보의 등록이나 저장이 정상적으로 이루어지지 않은 경우로 파일 정보가 있으나 파일 표본 정보가 없는 경우, 파일 정보가 없으나 파

일 표본 정보가 있는 경우가 검수 대상이 된다.

2. 구축 학습자 말뭉치 검수 정교화 및 품질 관리

2.1. 언어 자원 품질 확보를 위한 중복 표본 검수

○ 학습자 말뭉치는 2022년 10월을 기준으로 약 41,062개의 표본이 구축되어 있다. 다수의 작업자가 많은 수의 표본을 등록하는 과정에서 간혹 중복 표본이 발견되는 경우가 있는데, 이러한 문제를 해소하기 위하여 2021년 연구에서는 다음의 방법으로 중복 표본을 필터링한 바 있다.

- 표본별 파일명 검토를 통해 중복 가능성이 있는 표본 전수 조사
- 통계적으로 텍스트 내의 거리를 측정하여 텍스트의 내용이 비슷한 표본 전수 조사

○ 본 연구에서는 구축이 완료되는 시점에 위와 동일한 방법으로 2022년에 새롭게 구축되는 표본과 기구축 표본의 중복 여부를 확인하고, 중복 표본이 발견될 시 해당 표본을 삭제함으로써 국가 언어 자원으로서 학습자 말뭉치 자료의 무결성을 확보하고자 하였다.

2.2. 구축 말뭉치 통계 정보의 정확성 제고

1) 기구축 말뭉치의 통계 정보 검토

○ 국립국어원 한국어 학습자 말뭉치 나눔터와 LCMS의 통계는 크게 구축 통계, 자료 검색 통계로 구분할 수 있다. 이 중 구축 통계에는 구축이 완료된 말뭉치에 대한 통계와 구축 작업 진행 현황 통계가 포함되어 있다. 본 연구에서는 구축 도구 지원팀과의 협업을 통해 다음 통계 항목의 정확성을 검증하였다. 통계 정보 검증은 표본 정보 기반의 통계 정보와

LCMS 통계 정보 간의 일관성, 개별 표본 항목의 통계 정보 이상 유무 확인, 작업 통계와 구축 통계의 일치 여부 등을 중심으로 이루어졌다.

<표 50> 국립국어원 한국어 학습자 말뭉치 나눔터와 LCMS의 통계 정보

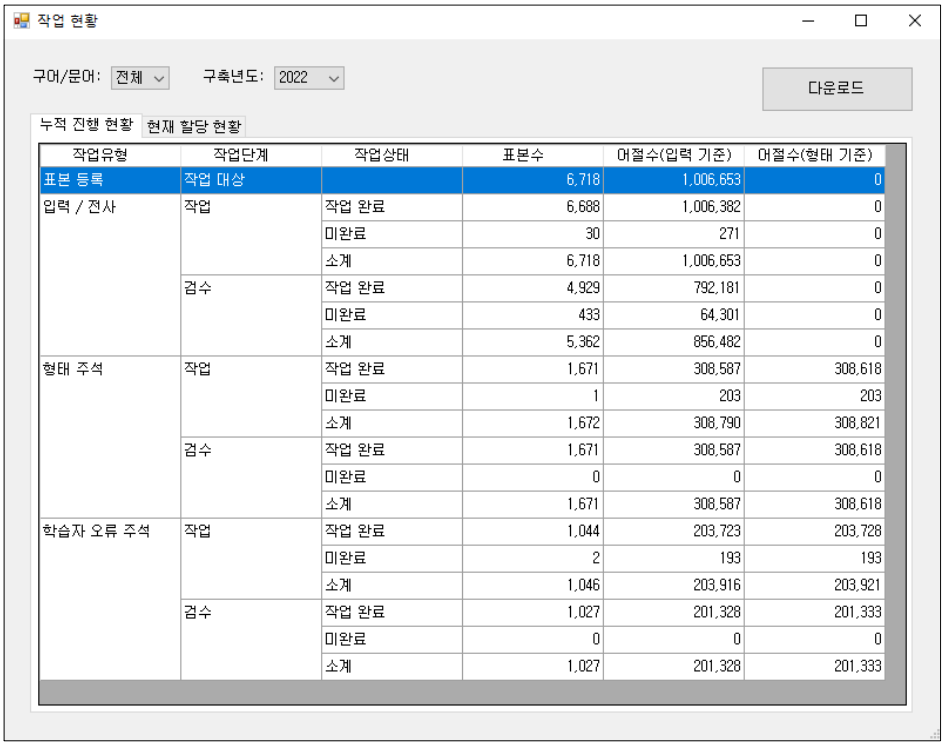
구분	통계 정보 유형	세부 내용	학습자 말뭉치 나눔터	LCMS
구축 통계	말뭉치 유형별	원시, 형태 주석, 오류 주석 말뭉치의 표본 수와 어절 수	○	○
	변인별 구축 현황	국적별	○	○
		언어권별	○	○
		수준별	○	○
		자료 유형별	○	○
	주석 말뭉치 표지별 통계	형태 주석이 된 형태소	○	○
		오류 위치가 주석이 된 형태소	○	○
		오류 양상이 주석된 형태소	○	○
		오류 층위가 주석이 된 형태소	○	○
	오류 주석 통계	오류 주석 빈도	○	○
		오류 위치 빈도	○	○
		오류 양상 주석	○	○
		오류 층위 주석	○	○
	작업 현황 통계	작업 배분 및 검수 진행 표본 수 및 어절 수		○
자료 검색 통계	검색어 통계	말뭉치 유형별 검색어의 검색 수	○	
	이용자 검색 시 제공되는 통계 정보	검색된 표본의 수와 비율 예) 검색된 표본 8,502개(28.4%), 검색된 어절의 수와 비율 예) 검색된 어절: 18,248건(0.45%)	○	

2) 구축 말뭉치 통계 정보의 정확성 제고 방안 마련 및 개선

○ 본 연구에서는 국립국어원 한국어 학습자 말뭉치 나눔터와 LCMS의 통계 정보 검토 결과를 바탕으로 다음의 두 가지 측면에서 구축 말뭉치 통계 정확성을 제고하기 위한 개선 방안이 필요함을 확인할 수 있었다.

(1) LCMS의 작업 할당 취소 표본의 이력 관리 방식 개선

○ 말뭉치 구축 도구인 LCMS에서는 다음과 같은 작업 현황 정보를 확인할 수 있다. 이 정보는 [작업]에서 [검수] 단계로 이어지는 학습자 말뭉치의 구축 공정을 그대로 반영한 것으로 [미완료]에는 [작업 할당]-[작업 접수]-[작업 진행]의 세부 단계가 포함된다.



작업유형	작업단계	작업상태	표본수	어절수(입력 기준)	어절수(형태 기준)
표본 등록	작업 대상		6,718	1,006,653	0
입력 / 전사	작업	작업 완료	6,688	1,006,382	0
		미완료	30	271	0
		소계	6,718	1,006,653	0
	검수	작업 완료	4,929	792,181	0
		미완료	433	64,301	0
		소계	5,362	856,482	0
형태 주석	작업	작업 완료	1,671	308,587	308,618
		미완료	1	203	203
		소계	1,672	308,790	308,821
	검수	작업 완료	1,671	308,587	308,618
		미완료	0	0	0
		소계	1,671	308,587	308,618
학습자 오류 주석	작업	작업 완료	1,044	203,723	203,728
		미완료	2	193	193
		소계	1,046	203,916	203,921
	검수	작업 완료	1,027	201,328	201,333
		미완료	0	0	0
		소계	1,027	201,328	201,333

<그림 2> LCMS 작업 현황 화면 예시

○ 이 과정에서 [작업 할당]이 잘못된 표본에 대한 [할당 취소]가 있을 수 있

는데, [할당 취소]가 이루어진 표본이 [미완료], 즉 작업 진행 중인 표본에 포함되어 통계가 집계되는 문제가 있다. 이에 따라 구축 도구 지원팀에 2022년에는 [할당 취소] 표본의 이력 삭제를 요청하여 조정을 한 바 있었다. 이러한 문제는 정확한 구축 통계 파악을 위해 개선이 필요한 부분이며, 이에 대해 구축 도구 지원팀에도 의견을 제시한 상태이다.

(2) <국립국어원 한국어 학습자 말뭉치 나눔터>의 이용자 검색 통계

○ <국립국어원 한국어 학습자 말뭉치 나눔터>에서 검색어를 사용하여 말뭉치를 검색하면 아래와 같이 '검색된 표본'과 '검색된 어절'의 통계 정보가 함께 제시된다. 이 정보는 표본 수와 어절 수, 그리고 둘의 비율이 함께 제시된다. 이때 제시되는 비율은 전체 말뭉치의 표본 수, 전체 말뭉치의 어절 수 대비 검색된 표본 수와 검색된 어절 수의 비율을 산출한 것이다.

<그림 3> 국립국어원 한국어 학습자 말뭉치 나눔터의 검색 통계 예시: 형태 주석(위), 오류 주석(아래)

○ 이러한 비율 외에 실제 검색된 표본의 총 어절 수 대비 검색된 어절 수, 그리고 그 비율이 추가로 제시될 필요가 있다. 이는 오류 말뭉치에서 학습자의 오류율을 나타내 주는 지표이며, 사용자들이 오류 주석 말뭉치를 통해 실질적으로 필요로 하는 정보이기 때문이다.

Ⅳ. 한국어 학습자 말뭉치 교육 및 홍보

1. 말뭉치 구축 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. 아울러 다양한 변이형을 포함한 학습자 말뭉치의 특성상 작업자의 직관에 의해 세부적인 자료 처리 방식이 달라질 우려가 있으므로 자료 처리 방식의 일관성 확보를 위해서도 매우 중요하다. 이에 따라 본 연구에서는 온라인/오프라인, 정기/비정기 워크숍을 통해 실무자 간에 원활한 소통이 이루어지도록 하고 있다.

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 51> 말뭉치 구축/가공 인력 교육 내용

	지침 교육	도구 교육 및 실습
수집	○ 자료 수집 과제 유형 및 수집 방법 ○ 학습자 동의서 수집 및 처리 ○ 수집 자료의 처리와 관리	○ 온라인 구축 도구: 수집 표본 등록 및 표본 관리
자료 처리 및 파일 등록	○ 자료의 분류 ○ 스캔 및 음성 파일 변환 ○ 파일명 부여 체계 ○ 학습자 정보 및 파일 정보 등록(헤더 마크업)	○ 온라인 구축 도구: 스캔/음성 원본 파일 업로드 및 파일 등록, 파일명 생성 ○ 스캐너 사용 ○ 음성 파일 변환
입력	○ 문어 입력 및 검수 방법, 쟁점	○ 온라인 구축 도구: 파일 입력 및 마크업, 할당 받은 작업 파일 관리 및 작업
전사	○ 구어 전사 및 검수 방법, 쟁점	○ 온라인 구축 도구: 도구 내 전사 및 마크업, 할당 받은 작업 파일 관리 및 작업
형태 주석	○ 형태 분석 방법 및 절차 ○ 형태 주석 체계 ○ 형태 분석 자료 검수 및 검수 관련 쟁점	○ 온라인 구축 도구: 형태 주석, 할당 받은 작업 파일 관리 및 작업
오류 주석	○ 오류 식별, 판정 및 교정의 기준 ○ 오류 주석 체계 ○ 오류 분석 자료 검수 및 검수 관련 쟁점	○ 온라인 구축 도구: 오류 주석, 할당 받은 작업 파일 관리 및 작업

2. 한국어 학습자 말뭉치 이용자를 위한 아카데미 개최

○ 2015-2021년 사업에서 매해 2회-5회까지 학습자 말뭉치 아카데미를 개최해 왔음에도 불구하고 여전히 자료 처리 및 분석 방법에 대한 지식과 경험의 부족으로 사용자들의 접근이 용이하지 않은 분야이다. 이는 사용자들이 지금까지 개최한 학습자 말뭉치 아카데미에 보여 온 호응을 통해 확인할 수 있다. 본 연구에서는 대학원생 및 한국어 교육 연구자, 산업계 연구자 등 다양한 사용자를 대상으로 하여 한국어 교육 연구와 기술 개발을 위한 연구에의 폭넓은 활용을 돕기 위해 참가자의 수준, 대상, 목적 등을 고려하여 학습자 말뭉치 아카데미를 기획하였다.

<표 52> 학습자 말뭉치 활용 아카데미 개최

일시	프로그램	강사	참가자
1차 (6. 30)	○ 학습자 말뭉치 기반 연구를 위한 자료 처리(기초) - 에디터를 활용한 학습자 말뭉치 자료 처리	장채린 (명지대학교)	Zoom 71명
2차 (8. 18)	○ 인공지능 기술을 기반으로 한 한국어 학습자 말뭉치 활용 사례 - 한국어 학습자 원어민성 측정, 한국어 학습자 모어 판별, 한국어 교육용 챗봇 설계를 중심으로 ○ 한국어 학습자 원어민성 측정을 위한 주성분 분석(PCA) 실습	이진 (연세대학교)	Zoom 178명
3차 (10. 28)	○ 학습자 말뭉치를 활용한 연구 주제 탐색과 적용 - 학습자 말뭉치를 활용한 연어 연구의 실제	유소영 (연세대학교)	대면 20명 Zoom 100명
4차 (11. 25)	○ 학습자 말뭉치 기반 연구를 위한 자료 처리의 실제: - 학습자 말뭉치의 구조 알아보기 - 학습자 말뭉치 가공하기	최정도 (계명대학교)	Zoom 76명 유튜브 106명

	: 텍스트 에디터, 엑셀 등 - 학습자 말뭉치 검색하기 : 학습자 말뭉치 나눔터, 앤트콘크(AntConc) 등 - 빈도 산출하기 : 엑셀, 앤트콘크(AntConc) 등 - 연구에 실제 적용하기		

3. 한국어 학습자 말뭉치 소개·활용 자료집 현행화 등, 국립국어원 관련 누리집 게재 및 아카데미 배포

3.1. ‘한국어 학습자 말뭉치 활용 매뉴얼’ 배포

○ <2021년 한국어 학습자 말뭉치 연구 및 구축> 사업에서는 사용자들이 한국어 학습자 말뭉치를 보다 손쉽게 이용할 수 있도록 하기 위해 ‘한국어 학습자 말뭉치 활용 매뉴얼’을 제작하였으며, 현재 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 배포되고 있다. 본 연구에서는 ‘한국어 학습자 말뭉치 아카데미’ 참가자들에게 자료집을 배포하여 사용자들의 학습자 말뭉치의 활용 능력을 제고하고자 하였다. ‘한국어 학습자 말뭉치 활용 매뉴얼’의 구성은 다음과 같다.

<표 53> ‘한국어 학습자 말뭉치 활용 매뉴얼’의 구성

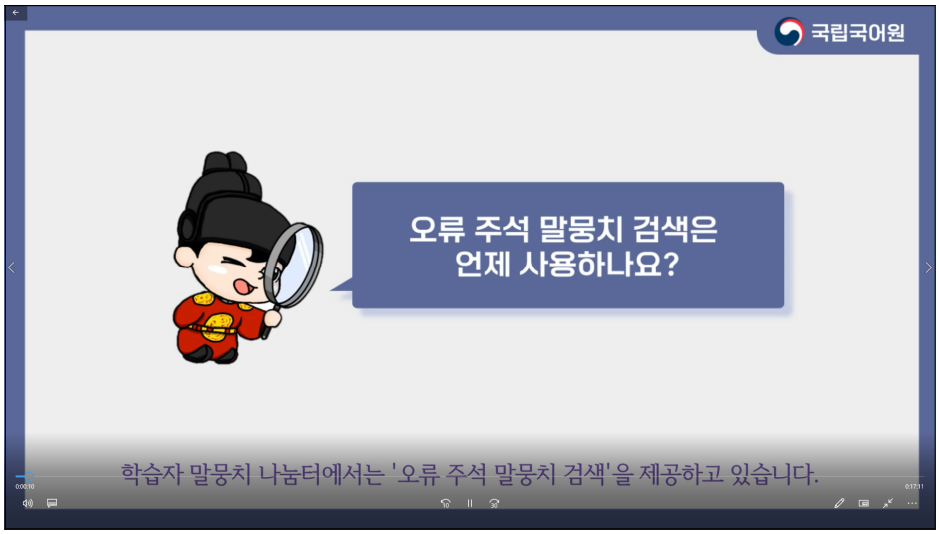
구성	세부 내용
Ⅰ. 학습자 말뭉치 알아보기	1. 학습자 말뭉치란? 2. 학습자 말뭉치 유형 3. 학습자 말뭉치 구축 현황 4. 학습자 말뭉치의 구성
Ⅱ. 학습자 말뭉치 나눔터 알아보기	1. 학습자 말뭉치 나눔터: 주요 메뉴 2. 검색 3. 검색 이용 안내 4. 통계

	5. 통계 자료 예시
Ⅲ. 학습자 말뭉치 나눔터 활용하기	1. 통합 말뭉치 검색 2. 원시 말뭉치 검색 3. 원시 말뭉치의 활용 4. 형태 주석 말뭉치 검색 5. 형태 주석 말뭉치의 활용 6. 오류 주석 말뭉치의 검색 7. 오류 주석 말뭉치의 활용 8. 말뭉치 상세 검색
Ⅳ. 학습자 말뭉치 기반 연구를 위한 활용 도구 더 알아보기	1. 엑셀 2. 텍스트 에디터 3. 앤트콩크(AntConc)
Q&A	

3.2. ‘한국어 학습자 말뭉치 활용’ 동영상 제작

○ 본 연구에서는 <2021년 한국어 학습자 말뭉치 연구 및 구축> 사업에서 제작한 ‘한국어 학습자 말뭉치 활용 매뉴얼’을 바탕으로 동영상을 제작하였다. 동영상은 사용자가 한국어 학습자 말뭉치를 보다 쉽게 활용하도록 하기 위한 것이다. 동영상은 <국립국어원 한국어 학습자 말뭉치 나눔터> 게시판에 게시되어 기초 과정의 한국어 학습자 말뭉치 활용에 관한 상시 아카데미 운영의 효과를 가진다.

<그림 4> '학습자 말뭉치 활용' 동영상 예시 화면 1



<그림 5> '학습자 말뭉치 활용' 동영상 예시 화면 2

Ⅴ. 결론

1. 연구 요약

본 연구는 한국어 교육 및 연구, 민간 분야에서의 활용을 목적으로, <2021년 한국어 학습자 말뭉치 연구 및 구축>에서 수립한 제2차 중장기 계획에 따라 한국어 학습자 말뭉치 말뭉치를 구축하였다. 이에 따른 주요 과업과 연구 성과는 다음과 같다.

○ 한국어 학습자 말뭉치 수집 및 구축·가공

본 연구에서는 2015-2021년까지 구축한 말뭉치의 대상별·수준별·언어권별·자료 변인별 분포 특성을 분석하고, 그 결과를 바탕으로 2022년 구축 말뭉치를 설계하였다. 이에 따라 기구축 말뭉치에서 상대적으로 비중이 적은 국내 대학(원)에 재학 중인 학문 목적 학습자의 자료와 국외 자료, 1급과 2급, 5급과 6급, 6급 이상, 영어권, 스페인어권, 러시아권 자료를 집중적으로 구축하기로 하였다. 아울러 기획 과제를 통해 문어의 경우 설명문이나 논설문, 구어의 경우 내러티브나 자유 발화를 집중적으로 수집하였다.

학습자 말뭉치 구축은 원시 말뭉치 1,008,315어절(문어 708,096어절, 구어 300,219어절), 형태 주석 말뭉치 308,790어절(문어 157,313어절, 구어 151,477어절), 오류 주석 203,490어절(문어 73,829어절, 구어 129,661어절) 규모의 말뭉치가 새롭게 구축되었다. 그 결과 전체 말뭉치의 규모는 원시 말뭉치 6,230,590어절(문어 4,407,583어절, 구어 1,823,007어절), 형태 주석 말뭉치 4,013,233어절(문어 2,760,085어절, 구어 1,253,148어절), 오류 주석 말뭉치 1,346,015어절(문어 674,636어절, 구어 671,379어절)이 되었다.

그리고 학습자 말뭉치의 활용도 제고를 위해 30여 명의 한국어 모어 화자를 대상으로 자료를 수집하여 총 문어 10,440어절, 구어 23,912어절의 참조 말뭉치를 원시 말뭉치의 형태로 시범 구축하였다.

○ 학습자 말뭉치의 검수 정교화 및 품질 관리

한국어 학습자 말뭉치는 2015년 구축이 시작된 이후 매년 일정 규모의 자료를 누적해 가며 양적으로 확대되어 가고 있다. 학습자 말뭉치의 검수 정교화 및 품질 관리는 양적 확대와 더불어 언어 자원으로서의 질적 제고를 위한

것으로 데이터의 무결성을 확보하는 것을 목적으로 한다. 이에 따라 문어 입력과 구어 전사, 형태 주석, 오류 주석의 각 작업 단계별로 3단계 작업 및 검수 체제에 따라 작업 공정을 진행하였으며, 공동 연구원을 중심으로 내부 검수단을 운영하여 무작위로 자료를 검수하는 절차를 두어 검수 체계를 강화하였다. 그 외에도 시스템 기반의 데이터 검증을 통한 오조작 데이터와 이상 데이터 검수를 상호보완적으로 적용하였으며, 최종 단계에서 전체 표본 정보 검수를 통한 메타정보 검증 작업과 중복 표본 검수 작업을 수행하였다. 또한 구축 말뭉치의 통계 정보의 정확성 제고를 목적으로 기구축 말뭉치의 통계 정보를 검토하고 이를 토대로 LCMS의 작업 할당 취소 표본의 이력 관리 방식 개선, <국립국어원 한국어 학습자 말뭉치 나눔터>의 이용자 검색 통계 제시 방식 개선을 제안하였다.

○ 학습자 말뭉치 교육 및 홍보

한국어 학습자 말뭉치 교육은 실무 작업자와 사용자를 대상으로 하여 이루어졌다. 실무자 작업자 교육은 작업자에게 말뭉치 구축에 관한 기본 소양과 기술을 익히도록 하고 각 구축 단계별 작업자로서의 전문성을 제고하여 체계적인 말뭉치를 구축해 나가기 위한 것으로, 지침 교육과 도구 사용 교육을 기본으로 한다. 그리고 구축 과정에서 발생하는 다양한 문제를 해결하기 위한 즉각적 소통과 피드백 시스템을 운영하고 정기 워크숍을 통해 말뭉치 구축에 관한 쟁점과 대응 방안을 공유하였다.

사용자를 대상으로 한 교육은 기초 과정에서 심화 과정까지 차별화된 총 4회의 학습자 말뭉치 아카데미를 통해 이루어졌다. 기초 과정으로는 학습자 말뭉치 기반 연구를 위한 자료 처리를 주제로 하여 2회가 이루어졌으며, 심화 과정으로는 인공지능 기술을 기반으로 한 한국어 학습자 말뭉치 활용 사례, 학습자 말뭉치를 활용한 연구 주제 탐색과 적용이라는 주제로 하여 2회가 이루어졌다.

이들 아카데미는 실시간 줌(Zoom)을 기본으로 하되, 학습자의 여건에 따라 참가 플랫폼을 선택할 수 있도록 하기 위하여 회차에 따라 유튜브 생중계를 병행하기도 하고, 대면 방식을 병행하기도 하였다. 또한 한국어 학습자 말뭉치를 소개하고 활용 방안을 설명하는 안내 자료와 동영상을 제작하여 국립국어원 누리집과 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 배포하였다.

2. 연구의 의의 및 기대 효과

한국어 학습자 말뭉치는 한국어 학습자가 산출한 언어 자료를 수집하여 구축한 국가 주도의 공공 언어 자원으로서 한국어 교육과 연구를 위한 기초 자료이자 빅데이터로서 인공지능 기술 개발을 위한 원천 자료로써 활용도가 높은 자료이다. <2022년 한국어 학습자 말뭉치 연구 및 구축> 사업의 의의 및 기대 효과는 다음과 같다.

○ 국가 주도의 대규모 공공언어 자원으로서의 한국어 학습자 말뭉치 구축

한국어 학습자 말뭉치 연구는 2010년 기초 연구에서 시작되어 2015년 본격적인 구축을 위한 제1차 중장기 계획 수립 이후 현재까지 이어져 오고 있다. 본 연구는 2025년까지 1,000만 어절 규모의 균형 말뭉치로의 확장을 목표로 하는 제2차 중장기 계획에 따른 2차 연도의 연구로 100만 어절 규모의 원시 말뭉치 구축을 목표로 하였다. 국외 학습자 말뭉치의 경우, 전 세계의 영어 학습자를 대상으로 한 자료 또는 공인 숙달도 평가 자료를 누적적으로 구축하고 있는 일부 말뭉치를 제외하고 대부분의 말뭉치가 1,000만 어절을 밑도는 규모이거나 100만 어절 미만의 소규모 말뭉치가 대부분임을 고려할 때 특수 말뭉치로서 상당히 의미 있는 규모라고 할 수 있다.

○ 균형 말뭉치 구축을 통한 한국어 학습자 말뭉치의 질적 제고

학습자 말뭉치는 한국어 학습자의 언어를 관찰하기 위한 자료로, 학습자의 제1언어, 숙달도 단계, 학습 환경, 경험 등의 다양한 변인이 존재하는 만큼 대표성과 함께 균형성이 중요하다. 본 연구에서 제안한 중장기 계획에서는 2015-2021년 기구축 말뭉치의 성과를 점검하고 이를 바탕으로 대상별, 언어권별, 수준별, 장르·주제별 변인을 중심으로 하여 말뭉치의 균형성을 확보하는 것을 목표로 하여 2022년 말뭉치 구축 목표를 설계하였다. 이는 2025년까지 대규모 언어 자원으로서의 완성도를 점진적으로 높여 나가고, 한국어 교육과 연구에서의 활용도를 제고하기 위한 것이다.

○ 한국어 교육 이론의 체계화 및 선진화된 교육 자료 구축의 기반 조성

한국어 학습자 말뭉치는 학습자의 언어 발달과 습득의 지표가 되는 중간언어 특성을 실증적으로 보여 준다. 이는 한국어 교육 연구를 위한 기반 자료이자 교수 자료, 교수법, 평가 도구 등을 개발하기 위한 기초 자료라는 점에

서 매우 유용한 자료이다. 한국어 학습자 말뭉치는 국립국어원 한국어 학습자 말뭉치 나눔터(https://kcorpus.korean.go.kr)를 통해 한국어 교육 연구자와 교수자, 학습자는 물론 한국어 교육 콘텐츠 개발, 산업 자원 개발을 목표로 하는 민간 기관에서 자유롭게 활용 가능하도록 하고 있다. 또한 학습자 말뭉치 아카데미를 통해 말뭉치 활용 방법에 관한 사용자 교육을 하고 있다. 이는 그간 사용자들이 가졌던 말뭉치 구축과 활용에 대한 어려움을 해소하는 데에 기여하였으며, 말뭉치를 활용한 연구를 활성화하여 한국어 교육의 이론을 체계화하고 체계화된 교육 자료를 구축해 나갈 수 있는 기반을 제공하였다.

○ 한국 언어·문화의 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 한국 언어·문화를 세계화하고, 더 나아가 한국 언어·문화를 널리 알리고 정치·경제·사회의 각 분야에서 폭넓게 활동할 수 있는 국제 인력을 양성함으로 해서 한국의 국제 경쟁력 강화에 기여할 수 있을 것이다.

3. 보고서 활용 방안

본 연구는 실제 말뭉치 구축을 주요한 과업으로 하였다. 보고서에는 과업 수행의 방법과 절차, 결과가 상세하게 기술되어 있으며 말뭉치 구축의 단계별 지침이 부록으로 첨부되어 있다. 이는 다양한 목적의 사용자들에게 한국어 학습자 말뭉치 구축의 이론과 실제, 그리고 말뭉치의 활용을 위한 지침으로 활용될 수 있다.

○ 한국어 학습자 말뭉치 구축과 활용에 관한 이론적 지침

한국어 학습자 말뭉치는 비모어 화자의 자료를 구축한 자료로 한국어 모어 화자에게서는 나타나지 않는 비정형의 발화를 포함하고 있으며 다양한 변인

에 대한 고려가 필요하다는 점에서 특수 말뭉치로 분류할 수 있다. 따라서 말뭉치를 구축함에 있어 일반적인 말뭉치 구축에 관한 이론과 지침을 준수하되 특수 말뭉치로서 학습자 자료의 특성에 대한 세심한 고려가 필요하다. 본 연구에서는 학습자 말뭉치 구축에 관한 이론적 논의를 찾아보기 힘든 상황에서 2010년 기초 연구에서 시작하여 2015년에서 2022년까지의 연구와 구축 작업을 수행해 오면서 대두되는 학습자 말뭉치 설계와 구축, 가공에 관한 쟁점과 해결 방안을 체계적으로 기술하였다. 본 보고서는 특히 2022년 100만 어절의 원시 말뭉치와 30만 어절의 형태 주석 말뭉치, 20만 어절의 오류 주석 말뭉치를 구축·가공하는 과정에서의 연구 방법과 절차, 결과를 수록하고 있다. 그러한 점에서 본 보고서는 한국어 학습자 말뭉치 구축에 관한 실제적인 모형이자 이론적 지침으로 활용 가능하다.

○ 한국어 학습자 말뭉치 구축과 활용을 위한 실용적 지침

본 보고서에는 부록으로 자료 수집 및 처리, 입력과 전사, 형태 주석, 오류 주석 지침을 수록하고 있다. 이들 지침은 자료의 호환성을 위해 <21세기 세종 한국어 균형 말뭉치>의 구축 지침을 기반으로 하되 비모어 화자 자료인 학습자 말뭉치의 특성을 반영하여 작성된 후, 실제 자료를 구축해 오면서 대두되는 수많은 쟁점과 해결 방안들을 실례와 함께 수록하는 과정을 통해 정교화된 것이다. 따라서 학습자 말뭉치 구축 이론과 실제를 모두 포괄하는 학술적 자료로 활용 가능하며, 학습자 말뭉치를 구축하고자 하는 기관이나 연구자들, 또는 본 연구에서 구축한 학습자 자료를 연구 목적에 맞게 추가 가공하고자 하는 사용자들에게 실용적인 지침으로 활용될 수 있다.

4. 정책 제안

○ 효율적인 자료 구축을 위한 수집 체계와 방식의 개편

한국어 학습자 말뭉치 구축의 첫 단계는 자료 수집으로, 그간의 사업에서는 한국어 교육 학계를 중심으로 한 수집 네트워크를 기반으로 수집을 진행해 왔다. 아울러 세종학당재단, 국제교류재단 등의 기관과의 업무 협약, 온라인 수집, 한국어 교사를 중심으로 한 수집단 운영 등 다양한 방식으로 수집의 효율성 제고를 위한 방안을 모색하고 시도해 왔다. 그러나 본 사업의 취지에 대한 이해와 공감대를 어느 정도 형성하고 있음에도 불구하고, 실질적

인 자료 수집으로 이어지지 않는 경우가 많았다. 이에 양질의 자료를 지속적으로 구축해 나가기 위해서는 자료 수집 체계와 방식의 개편이 절실하게 요구된다. 외국어 말뭉치의 경우, 대규모의 공인 평가와 연계하여 수험자에게 연구 또는 교육 자료로서의 자료 이용에 대한 허락을 받고 자료를 누적적으로 수집하거나 전 세계의 기관들이 연합하여 구성된 학습자 말뭉치 협의체를 통해 조직적으로 자료를 수집해 나가는 사례를 볼 수 있다. 한국어 학습자 말뭉치에서도 이를 벤치마킹하여 한국어능력시험을 주관하고 있는 국립국제교육원, 세종학당의 한국어 숙달도 평가 자료를 연계하거나 국내의 어학당과 이주민 교육 기관, 국외의 세종학당, 한국국제교류재단의 객원 교수 파견 대학 등과의 업무 협약을 통해 조직적인 자료 수집을 해 나갈 필요가 있다.

○ 한국어 교육 연구 용역과의 연계를 통한 국가 언어 자원으로서의 활용도 제고

한국어 학습자 말뭉치는 2022년 사업 성과까지 포함하여 현재 약 620만 어절의 원시 말뭉치와 400만 어절의 형태 주석 말뭉치, 134만 어절의 오류 주석 말뭉치가 구축되었다. 구축된 말뭉치는 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 배포되어 한국어 교육 연구와 교수·학습에 활용되고 있으며, 학위논문과 학술지 논문 등의 연구 성과 산출로 이어지고 있다. 이는 학습자 말뭉치의 활용이라는 측면에서 상당히 의미 있는 일이나, 대규모 예산이 투입되어 구축된 공공 언어 자원인 만큼 보다 적극적인 활용 방안을 모색할 필요가 있다. 학습자 말뭉치는 한국어 학습자의 중간언어를 포함하고 있는 자료로 이는 한국어 학습자의 언어 습득과 발달을 관찰하기 위한 연구 자료로는 물론 교수·학습 자료로의 활용 가치가 매우 높다. 학습자 말뭉치 자료를 활용한 국가 주도의 교수·학습 자료 개발 사업으로 연계하거나 <한국어기초사전>과 같이 국립국어원에서 제공하고 있는 서비스와의 연계를 통해 학습자 말뭉치의 활용 범위를 교원, 학습자까지 넓힐 수 있으며, 그것이 곧 학습자 말뭉치 구축 사업의 목적이자 의의가 될 것이다.

○ 대상과 사용 목적을 고려한 국립국어원 한국어 학습자 말뭉치 나눔터 서비스의 정교화

한국어 학습자 말뭉치는 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 구축 자료를 사용자에게 배포하고 있다. 이를 통해 현재까지 많은 사용자들이 자료를 검색하거나 내려받아 연구 또는 교육 자료로 광범위하게 활용해 오고 있다. 그러나 사용자가 주로 연구자로 현장에서의 교수자나 학습자로

확대되지 못하고 있다는 한계가 있다. 따라서 교수 현장에서 수업이나 교수 자료로의 활용을 목적으로 하는 교수자, 학습 자료로의 활용을 목적으로 하는 학습자에게 보다 친숙한 방식으로 자료의 구조나 검색 방식을 차별화하여 서비스를 제공할 필요가 있다. 예를 들어, 학습자 오류 정보를 기반으로 한 오류 사전을 개발하여 서비스할 수 있는데, 이는 앞서 제기하였던 한국어 교육 연구 용역과의 연계를 통해 학습자 말뭉치를 기반으로 한 오류 사전을 기획하여 개발하고 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 검색하도록 할 수 있다.

○ 자료 활용의 효율성과 사용자의 편의성 제고를 위한 자료 배포 방식 개선

한국어 학습자 말뭉치는 <국립국어원 한국어 학습자 말뭉치 나눔터>를 통해 검색하고 내려받는 방식 외에 국립국어원에 요청하여 제공받는 방식을 통해 사용자에게 배포되고 있다. 이 중 후자의 경우 연구를 목적으로 하는 한국어 교육 연구자나 민간 기관에서 주로 이용하는 방식인데, 자료의 디렉토리 구조가 복잡하여 사용자에게 친숙한 구조로의 개선이 요구된다. 현재 제공되고 있는 방식은 '구축 연도>말뭉치 유형>개별 표본'의 3단계로 계층화된 디렉토리 구조를 통해 제공하고 있으며, 각 표본 폴더 안에 텍스트(TXT), 엑셀(Excel), 엑스엠엘(XML) 세 가지 형식의 파일 저장하여 제공하고 있다. 이에 따라 사용자들이 필요로 하는 자료를 추출하기 위하여 복잡한 전처리 과정을 거치게 되는데, '말뭉치 유형>파일 형식'으로 불필요한 디렉토리를 제거하고 단순화하여 편의성을 제고할 수 있다.

참고 자료

강현화(2010), 한국어 학습자 사전 표제어 선정을 위한 자료 구축 및 선정 방법에 관한 연구, 한국사전학 16, 한국사전학회.

강현화(2017), 학습자 말뭉치의 구축과 활용, 소통.

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・조민정(2003), 스페인어권 한국어 학습자의 어미, 조사 및 시상, 사동 범주의 오류 분석, 한국어교육 14(2), 국제한국어교육학회.

고석주(2002), 학습자 말뭉치에서 조사 오류의 특징, 외국어로서의 한국어교육 27(1), 연세대학교 한국어학당.

고석주(2004), 오류 유형 주석을 위한 기초 연구, 한국 문화사.

고승연(2013), 아랍어권 한국어 학습자의 발음 오류 분석, 한국어문화교육 7(1), 한국어문화교육학회.

국립국어원(2010), 한국어 학습자 말뭉치 구축 설계, 국립국어원 연구용역 결과보고서(연구책임: 강현화).

국립국어원(2015), 2015년 한국어 학습자 말뭉치 기초 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 서상규).

국립국어원(2016), 2016년 한국어 학습자 말뭉치 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 강현화).

국립국어원(2017), 2017년 한국어 학습자 말뭉치 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 강현화).

국립국어원(2018), 2018년 한국어 학습자 말뭉치 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 한송화).

국립국어원(2019), 2019년 한국어 학습자 말뭉치 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 한송화).

국립국어원(2020), 2019-2020년 한국어 학습자 말뭉치 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 한송화).

국립국어원(2021), 2021년 한국어 학습자 말뭉치 연구 및 구축, 국립국어원 연구용역 결과보고서(연구책임: 한송화).

권기양(2006), KFL 학습자의 오류에 대하여: 중국인 학습자 중심으로, 언어과학 13(3), 한국언어과학회.

김경화(2013), 고급단계 한국어학습자의 오류연구, 중국조선어문 188, 길림성민족사무위원회.

김미경·강현화(2017), 중·고급 중국어권 한국어 학습자의 조사 '가'와 '는' 선택 요인 연구, 외국어로서의한국어교육 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), 국제한국어교육학회.
김미옥・정희정(2003), 한국어 학습자 작문에 나타난 어휘 오류 분석, 제3회 한국어 교육 국제 워크숍 발표 요지, 연세대 언어정보연구원 외국어로서의 한국어교육 연구센터, 102-135쪽.
김아름(2014), 한국어 학습자의 문법 및 화용오류에 대한 인식, 새국어교육 100, 한국국어교육학회.
김유미(2002), 학습자 말뭉치를 이용한 한국어 학습자 오류 분석 연구, 외국어로서의 한국어교육 27, 연세대학교 한국어학당.
김유정(2005), 한국어 학습자 말뭉치 오류 분석의 기준, 한국어 교육 16(1), 국제한국어교육학회.
김일환(2016), 한국어 학습자 말뭉치의 주석 과정과 활용 방법, 국제한국어교육학회 춘계학술발표논문집, 국제한국어교육학회.
김정숙(2002), 영어권 한국어 학습자의 조사 사용 오류 분석과 교육 방법, 한국어교육 13(1), 국제한국어교육학회.
김정숙(2002), 한국어 학습자 말뭉치 구축을 위한 기초 연구 -개인 정보 표지 체계와 오류 정보 표지 체계를 중심으로-, 이중언어학회.
김정숙, 김유정(2002), 한국어 학습자 말뭉치 구축을 위한 기초 연구 -개인정보 표지 체계와 오류 정보 표지 체계를 중심으로-, 이중언어학 21, 이중언어학회.
김정은(2003), 한국어교육에서의 중간언어와 오류 분석, 한국어 교육 14(1), 국제한국어교육학회.
김지민, 신승용(2010), 어휘오류 분석의 문제점과 어휘오류 처치 방안 연구, 언어와 문화 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.
남윤주 외(2014), L2로서의 한국어 자연발화 코퍼스의 구축과 활용, 통일인문학논총.
노미연(2012), 한국어 학습자의 구어 오류와 후속 상호작용 분석 연구, 동국대학교 박

사학위논문.
민영란(2008), 부정적 전이로 인한 중국어권 학습자의 오류 분석, 한국어 교육 19(1), 국제한국어교육학회.
박수연(2007), 한국어 학습자 오류 말뭉치 구축과 그 문제점에 관한 연구, 언어 사실과 관점 17, 연세대학교 언어정보연구원.
서상규, 유현경, 남윤진(2002), 한국어 학습자 말뭉치와 한국어교육, 한국어교육 13(1), 국제한국어교육학회.
신성철(2002), 호주 한국어 학습자의 어휘 오류 분석 연구, 한국어 교육 13(1), 국제한국어교육학회.
신성철(2007), 영어권 한국어 학습자의 철자 오류 유형과 패턴, 한국어 교육 18(3), 국제한국어교육학회.
유석훈(2001), 외국어로서의 한국어 학습자 말뭉치 구축의 필요성과 자료 분석, 한국어교육 12(1), 국제한국어교육학회.
이동은(2007), 한국어 학습자의 철자 오류와 개선 방안 -북미지역 청소년 교포 학습자를 대상으로-, 한국어학 35, 한국어학회.
이병운(2011), 베트남인 학습자의 작문 오류 경향 분석: 조사·어미를 중심으로, 우리말글 52, 우리말글학회.
이승연(2006), 한국어 학습자 말뭉치 오류 표지 방안 재고, 이중언어학 31, 이중언어학회.
이승연(2007), 한국어 학습자 오류 판정 및 수정 기준 연구-교사, 비교사 집단간 오류 판별 비교 실험을 바탕으로, 이중언어학 33, 이중언어학회.
이승연(2007), 한국어교육을 위한 한국어 학습자 말뭉치의 구축과 활용 연구, 고려대 박사학위논문.
이유림, 김영주(2013), 교사의 피드백 방법이 한국어 학습자의 작문 내 어휘 오류 감소에 미치는 영향, 외국어로서의 한국어교육 39, 연세대학교 언어연구교육원 한국어학당.
이은서(2017), 중국어권 학습자의 접사 사용 연구, 연세대학교 대학원 석사학위 논문.
이정희(2002), 한국어 오류 판정과 분류 방법에 관한 연구, 한국어교육 13(1), 국제한국어교육학회.
이정희(2003), 초급 단계 한국어 학습자의 어휘 오류, 이중언어학 22, 이중언어학회.
이정희(2009), 중국어권 한국어 학습자의 어휘 오류 연구, 한국어 교육 19(3), 1-23쪽, 국제한국어교육학회.
이화진·이지연(2016), 학습자 말뭉치 구축과 음성 인식 활용, 한국언어문화교육학회 학술대회 발표 자료집, 한국언어문화교육학회.
이훈호(2015), 한국어 오류 분석 연구의 동향 분석 연구, 외국어교육연구 29(2), 107-135쪽, 한국외국어대학교 외국어교육연구소.
전영옥(2010), 여성결혼이민자의 한국어 어휘 오류 분석, 한말 연구 27, 한말연구학회.

조철현 외(2002), 한국어 학습자의 오류 유형 조사 연구, 문화관광부.
최원평, 유효려(2010), 중국 대학생 글쓰기에 나타난 어휘 오류 연구, 언어와 문화 6(3), 한국언어문화교육학회.
한상미(2014), 중급 한국어 학습자의 구어 담화에 나타난 조사 오류 연구, 한국어교육 25(3), 국제한국어교육학회.
한송화(2001), 말뭉치와 학습자 오류를 이용한, 외국인 학습자를 위한 한국어 어휘 사전의 의미 기술, 한국어정보학 4, 한국어정보학회.
한송화, 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.
Brock, C , Crookes, C , Day, R., and Long, M. (1986). The differential effects of corrective feedback in native speaker－non-native speaker conversation. In R. Day (Ed.), Talking to learn. Rowley, MA: Newbury House. pp. 229－236.
Brock, C. (1986). The effects of referential questions on ESL classroom discourse. TESOL Quarterly, 20, pp. 47-59.
Corder. S. P.(1981), Error Analysis and Interlanguage, Oxford University Press.
Foster, P. and Skehan, P. (1996) The influence of planning on performance in task-based learning. Studies in Second Language Acquisition 18. pp. 299 - 324.
Foster, P., Tonkyn, A. and Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. Applied Linguistics 21:3. pp. 354-375.
Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research report No. 3. Champaign, IL, USA: NCTE. pp. 1467-1770.
James, C.(1998), Errors in Language Learning and Use. New York : Addison Welsey Longman Inc. pp. 144-154.
Pica, T., Holliday, L., Lewis, L. and Morgenthaler, L. (1989) Comptrhensible Output As An Outcome of Linguistic Demandes On the Learner, Studies in Second Language Acquisition 11:1. pp. 63-90.
Young, R. (1995). Conversational Styles in Language Proficiency Interviews. Language Learning 45:1. pp. 3 - 42.

부록 1. 2015-2021년 한국어 학습자 말뭉치 분석 결과

1. 말뭉치 유형별 구축 비율

○ 2015-2021년 국립국어원 한국어 학습자 말뭉치는 원시 말뭉치 5,220,275어절, 형태 주석 말뭉치 3,704,443어절, 오류 주석 말뭉치 1,142,525어절이 구축되었다. 이 중 형태 주석 말뭉치는 원시 말뭉치의 70.9%(문어 70.4%, 구어 72.3%), 오류 주석 말뭉치는 21.9%(문어 16.2%%, 구어 35.6%)를 차지한다.

<표 1> 2015-2021년 국립국어원 한국어 학습자 말뭉치 유형별 구축 통계

구분	자료 유형	1급	2급	3급	4급	5급	6급	6급 이상	합계
원시	문어	439,612	596,563	690,248	650,251	722,627	459,546	140,640	3,699,487
	구어	277,908	313,967	406,673	246,126	139,498	107,877	30,739	1,522,788
	합계	717,520	910,530	1,096,921	896,377	862,125	567,423	171,379	5,222,275
형태 주석	문어	381,081	433,472	447,834	421,654	443,722	407,971	67,038	2,602,772
	구어	214,522	205,527	210,319	210,050	136,266	103,875	21,112	1,101,671
	합계	595,603	638,999	658,153	631,704	579,988	511,846	88,150	3,704,443
오류 주석	문어	89,925	104,084	104,832	86,977	109,714	105,275	-	600,807
	구어	93,897	103,304	104,126	107,738	71,964	55,797	4,892	541,718
	합계	183,822	207,388	208,958	194,715	181,678	161,072	4,892	1,142,525

2. 대상별·수준별 말뭉치의 구축 비율

1) 원시 말뭉치

○ 원시 말뭉치는 국내 교육기관의 학습자 자료 4,379,663어절, 이주민 학습자 자료 397,801어절, 국외 학습자 자료 444,811어절이 구축되었다. 이주민 학습자의 자료와 국외 학습자 자료가 국내 교육기관 자료에 비해 현저하게 적어 확대가 필요함을 확인할 수 있다.

<표 2> 2015-2021년 국립국어원 한국어 학습자 말뭉치 대상별·수준별 통계: 원시 말뭉치

수집 대상	자료 유형	1급	2급	3급	4급	5급	6급	6급 이상	합계
국내	문어	412,751	554,837	659,001	625,881	712,302	454,215	140,640	3,559,627
	구어	109,225	166,076	165,282	154,684	111,908	88,921	23,940	820,036
	합계	521,976	720,913	824,283	780,565	824,210	543,136	164,580	4,379,663
이주민	문어	15,647	18,470	19,500	21,977	10,050	3,175	-	88,819
	구어	58,179	79,802	69,084	69,010	19,093	13,814	-	308,982
	합계	73,826	98,272	88,584	90,987	29,143	16,989	-	397,801
국외	문어	11,214	23,256	11,747	2,393	275	2,156	-	51,041
	구어	110,504	68,089	172,307	22,432	8,497	5,142	6,799	393,770
	합계	121,718	91,345	184,054	24,825	8,772	7,298	6,799	444,811

2) 형태 주석 말뭉치

○ 형태 주석 말뭉치는 국내 교육기관의 학습자 자료 3,086,817어절, 이주민 학습자 자료 389,157어절, 국외 학습자 자료 228,469어절이 구축되었다. 원시 말뭉치와 마찬가지로 이주민 학습자의 자료와 국외 학습자 자료가 국내 교육기관 자료에 비해 현저하게 적으며, 특히 문어 자료의 비중이 낮아 이들 자료를 집중적으로 확대할 필요가 있음을 확인할 수 있다.

<표 3> 2015-2021년 국립국어원 한국어 학습자 말뭉치 대상별·수준별 통계: 형태 주석 말뭉치

자료 유형	수집 대상	1급	2급	3급	4급	5급	6급	6급 이상	합계
국내	문어	354,246	402,854	425,108	397,713	433,397	402,970	67,038	2,483,326
	구어	79,994	90,462	95,569	120,935	109,713	85,706	21,112	603,491
	합계	434,240	493,316	520,677	518,648	543,110	488,676	88,150	3,086,817
이주	문어	15,621	18,470	19,500	21,904	10,050	3,175	-	88,720

민	구어	56,516	76,388	67,248	67,378	19,093	13,814	-	300,437
	합계	72,137	94,858	86,748	89,282	29,143	16,989	-	389,157
국외	문어	11,214	12,148	3,226	2,037	275	1,826	-	30,726
	구어	78,012	38,677	47,502	21,737	7,460	4,355	-	197,743
	합계	89,226	50,825	50,728	23,774	7,735	6,181	-	228,469

3) 오류 주석 말뭉치

○ 오류 주석 말뭉치는 국내 교육기관의 학습자 자료 869,304어절, 이주민 학습자 자료 126,343어절, 국외 학습자 자료 146,878어절이 구축되었다. 국내 교육기관 학습자 자료의 비중이 높고, 이주민 학습자와 국외 학습자 자료의 비중이 낮아 이들을 확대해야 함을 알 수 있다.

<표 4> 2015-2021년 국립국어원 한국어 학습자 말뭉치 대상별·수준별 통계: 오류 주석 말뭉치

자료 유형	수집 대상	1급	2급	3급	4급	5급	6급	6급 이상	합계
국내	문어	75,822	84,352	89,981	70,445	105,719	103,145	-	529,464
	구어	43,102	55,231	59,557	65,013	62,023	50,022	4,892	339,840
	합계	118,924	139,583	149,538	135,458	167,742	153,167	4,892	869,304
이주민	문어	6,666	7,584	11,625	14,495	3,895	699	-	44,964
	구어	12,984	20,964	17,295	23,447	4,418	2,271	-	81,379
	합계	19,650	28,548	28,920	37,942	8,313	2,970	-	126,343
국외	문어	7,437	12,148	3,226	2,037	100	1,431	-	26,379
	구어	37,811	27,109	27,274	19,278	5,523	3,504	-	120,499
	합계	45,248	39,257	30,500	21,315	5,623	4,935	-	146,878

3. 언어권별 말뭉치의 구축 비율

1) 문어

(1) 원시 말뭉치

○ 문어 원시 말뭉치는 총 3,699,487어절 중 중국어권 학습자 자료가 1,721,599어절로 가장 많은 비중을 차지하며, 이어서 일본어권 516,104어절, 베트남어권 252,924어절, 영어권 241,269어절을 차지하였다.

<표 5> 2015-2021년 국립국어원 한국어 학습자 말뭉치 언어권별 통계: 원시 문어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	200,181	230,033	274,706	287,248	390,849	233,003	105,579	1,721,599
일본어	40,085	86,454	110,482	108,407	100,626	69,594	456	516,104
베트남어	40,383	49,735	52,913	45,493	40,785	17,619	5,996	252,924
영어	31,683	47,656	48,520	42,098	35,401	32,397	3,514	241,269
광둥어	13,324	28,706	35,808	37,492	34,085	34,814	-	184,229
러시아어	13,007	21,305	28,559	25,117	25,980	10,733	4,534	129,235
타이어	15,729	28,975	22,240	12,586	9,609	7,849	321	97,309
몽골어	10,886	13,579	15,893	13,457	12,792	6,515	46	73,168
스페인어	9,132	15,014	15,238	9,102	6,116	2,152	293	57,047
인도네시아어	7,012	7,813	7,135	10,257	8,008	5,229	1,097	46,551
기타	58,190	67,293	78,754	58,994	58,376	39,641	18,804	380,052
합계	439,612	596,563	690,248	650,251	722,627	459,546	140,640	3,699,487

(2) 형태 주석 말뭉치

○ 문어 형태 주석 말뭉치는 총 2,602,772어절 중 중국어권 학습자 자료가 959,834어절로 가장 많은 비중을 차지하며, 이어서 일본어권 449,202어절, 베트남어권 232,624어절, 영어권 227,447어절을 차지하였다.

<표 6> 2015-2021년 국립국어원 한국어 학습자 말뭉치 언어권별 통계: 형태 문어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	142,191	140,978	141,935	139,360	161,979	195,729	37,662	959,834
일본어	40,085	68,511	88,400	90,862	93,933	67,411	-	449,202
베트남어	40,383	43,568	49,004	38,705	37,700	17,268	5,996	232,624
영어	31,683	47,656	42,392	37,981	33,577	32,397	1,761	227,447
러시아어	13,007	15,396	22,100	16,747	25,035	10,733	3,990	107,008
광둥어	13,248	5,940	6,096	14,239	19,596	25,975	-	85,094
타이어	15,729	16,957	13,111	11,727	9,396	7,519	321	74,760
몽골어	10,886	11,984	14,817	12,119	10,946	6,515	46	67,313
인도네시아어	7,012	7,813	6,971	9,175	6,655	5,229	1,097	43,952
스페인어	9,132	9,976	9,970	5,743	5,016	2,152	293	42,282
기타	57,725	64,693	53,038	44,996	39,889	37,043	15,872	313,256
합계	381,081	433,472	447,834	421,654	443,722	407,971	67,038	2,602,772

(3) 오류 주석 말뭉치

○ 문어 오류 주석 말뭉치는 총 600,807어절 중 일본어권 153,885어절, 영어권 153,254어절로 거의 비슷한 비중을 차지하였으며, 중국어권 113,259어절, 베트남어권 65,869어절, 러시아어권 자료 31,162어절을 차지하였다.

<표 7> 2015-2021년 국립국어원 한국어 학습자 말뭉치 언어권별 통계:
오류 문어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	합계
일본어	18,368	18,554	18,258	19,167	39,619	39,919	153,885
영어	25,189	31,816	30,834	17,741	25,651	22,023	153,254
중국어	18,909	20,832	19,966	19,211	16,473	17,868	113,259
베트남어	8,450	7,265	13,753	11,360	11,587	13,454	65,869
러시아어	3,864	5,599	6,978	5,808	6,253	2,660	31,162
타이어	4,408	6,502	3,268	3,678	3,844	3,034	24,734
스페인어	2,348	2,513	1,536	1,247	152	317	8,113
아랍어	145	1,436	697	583	102	1,205	4,168
카자흐어	247	436	1,520	855	552	541	4,151
프랑스어	832	715	407	116	844	717	3,631
기타	7,165	8,416	7,615	7,211	4,637	3,537	38,581
합계	89,925	104,084	104,832	86,977	109,714	105,275	600,807

2) 구어

(1) 원시 말뭉치

○ 구어 원시 말뭉치는 총 1,522,788어절 중 중국어권 학습자 자료가 345,244어절로 가장 많은 비중을 차지하며, 이어서 타이어권 294,310어절, 베트남어권 227,830어절, 일본어권 142,524어절을 차지하였다.

<표 8> 2015-2021년 국립국어원 한국어 학습자 말뭉치 언어권별 통계:
원시 구어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	49,939	67,766	61,055	53,412	39,552	52,408	21,112	345,244
타이어	69,918	46,892	153,426	14,369	4,962	4,743	-	294,310
베트남어	50,169	47,272	48,929	50,373	21,855	8,455	777	227,830
일본어	11,242	27,471	24,217	26,787	29,274	22,698	835	142,524

스페인어	26,202	22,625	22,073	19,165	4,225	511	594	95,395
러시아어	13,136	18,736	22,436	16,483	4,610	3,682	-	79,083
인도네시아어	11,302	13,237	15,122	7,613	8,307	4,583	-	60,164
영어	9,204	18,499	12,798	6,458	7,088	2,282	-	56,329
타갈로그어	11,948	11,615	11,635	6,632	1,682	1,711	-	45,223
싱할라어	6,204	5,306	6,539	5,740	3,726	2,990	-	30,505
기타	18,644	34,548	28,443	39,094	14,217	3,814	7,421	146,181
합계	277,908	313,967	406,673	246,126	139,498	107,877	30,739	1,522,788

(2) 형태 주석 말뭉치

○ 구어 형태 주석 말뭉치는 총 1,101,671어절 중 중국어권 학습자 자료가 240,677어절로 가장 많은 비중을 차지하며, 이어서 베트남어권 189,271어절, 일본어권 136,38어절, 타이어권 116,290어절을 차지하였다.

<표 9> 2015-2021년 국립국어원 한국어 학습자 말뭉치 언어권별 통계: 형태 구어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	37,319	32,839	28,439	31,507	38,422	51,039	21,112	240,677
베트남어	49,248	30,085	30,266	49,363	21,855	8,455	-	189,272
일본어	11,242	27,394	23,121	26,787	27,462	20,432	-	136,438
타이어	39,931	18,950	34,687	13,674	4,672	4,376	-	116,290
러시아어	11,521	13,467	19,979	16,001	4,610	3,682	-	69,260
인도네시아어	10,801	11,186	15,122	7,613	8,307	4,583	-	57,612
영어	9,204	17,635	12,798	6,458	7,088	2,282	-	55,465
스페인어	13,631	11,082	6,773	14,435	4,225	511	-	50,657
타갈로그어	11,948	10,987	11,635	6,632	1,682	1,711	-	44,595
싱할라어	6,204	5,306	6,539	5,740	3,726	2,990	-	30,505
기타	13,473	26,596	20,960	31,840	14,217	3,814	-	110,900
합계	214,522	205,527	210,319	210,050	136,266	103,875	21,112	1,101,671

(3) 오류 주석 말뭉치

○ 구어 오류 주석 말뭉치는 총 541,718어절 중 중국어권 학습자 자료가 127,247어절, 일본어권 112,550어절, 베트남어권 107,562어절로 가장 높은 비중을 차지하며, 이어서 영어권 자료가 43,081어절, 인도네시아어권 자료가 30,606어절을 차지하였다.

<표 10> 2015-2021년 국립국어원 한국어 학습자 말뭉치 언어권별 통계: 오류 구어 말뭉치

제1언어	1급	2급	3급	4급	5급	6급	6급 이상	합계
중국어	19,560	19,068	13,901	20,501	19,683	29,642	4,892	127,247
일본어	10,078	22,909	22,251	19,881	19,304	18,127	-	112,550
베트남어	27,477	18,022	15,870	28,639	12,054	5,500	-	107,562
영어	9,069	11,314	10,909	5,093	6,136	560	-	43,081
인도네시아어	8,096	5,343	8,865	3,706	4,596	-	-	30,606
스페인어	7,789	7,394	5,704	5,555	1,347	511	-	28,300
타이어	2,841	3,529	8,767	4,079	850	628	-	20,694
러시아어	1,820	2,146	7,972	3,863	1,271	-	-	17,072
아랍어	-	2,855	560	3,854	512	829	-	8,610
타갈로그어	3,996	1,650	767	428	-	-	-	6,841
기타	3,171	9,074	8,560	12,139	6,211	-	-	39,155
합계	93,897	103,304	104,126	107,738	71,964	55,797	4,892	541,718

4. 장르별 말뭉치의 구축 비율

1) 문어

(1) 원시 말뭉치

○ 문어 원시 말뭉치는 총 3,699,487어절 중 생활문이 1,651,527, 논설문이 1,207,050어절로 두 장르가 주를 이루었으며, 그 외에 설명문 251,033어절, 보고서 239,247어절로 다른 장르에 비해 상대적으로 높은 비중을 차지하였다.

<표 11> 2015-2021년 국립국어원 한국어 학습자 말뭉치 장르별 통계: 원시 문어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
생활문	395,484	433,751	480,924	227,428	59,967	51,867	2,106	1,651,527
논설문	263	8,240	70,098	314,213	525,077	285,211	3,948	1,207,050
설명문	13,138	67,785	58,574	40,269	57,402	12,390	1,475	251,033
보고서	-	233	6,732	29,289	38,653	31,367	132,973	239,247
기행문	14,956	70,815	13,362	4,603	706	5,884	-	110,326
수필	5,047	9,854	13,444	18,753	33,394	24,375	-	104,867
감상문	-	114	42,454	6,026	3,332	1,789	138	53,853
기사문	-	142	-	9,345	3,520	25,059	-	38,066
전기문	-	-	-	-	259	21,604	-	21,863
편지글	10,724	5,629	4,660	325	317	-	-	21,655
합계	439,612	596,563	690,248	650,251	722,627	459,546	140,640	3,699,487

(2) 형태 주석 말뭉치

○ 문어 형태 주석 말뭉치는 총 2,602,772어절 중 생활문이 1,194,329어절로 가장 높은 비중을 차지하였으며, 논설문이 839,381어절로 그 뒤를 이었다. 그 외에 설명문이 174,526어절로 생활문이나 논설문에 비해서는 적지만 다른 장르에 비해 상대적으로 높은 비중을 차지하였다.

<표 12> 2015-2021년 국립국어원 한국어 학습자 말뭉치 장르별 통계: 형태 문어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
생활문	342,170	311,981	302,772	154,740	37,742	43,426	1,498	1,194,329
논설문	263	1,971	53,266	201,465	320,498	261,252	666	839,381
설명문	11,548	56,795	34,230	28,357	32,450	10,260	886	174,526
보고서	-	120	1,388	8,267	14,047	14,766	63,988	102,576
수필	3,412	6,458	9,257	17,294	32,811	24,375	-	93,607
기행문	13,075	51,324	10,056	3,165	524	5,517	-	83,661
감상문	-	114	33,913	5,408	2,691	1,789	-	43,915
기사문	-	142	-	2,633	2,609	25,059	-	30,443
전기문	-	-	-	-	259	21,527	-	21,786
편지글	10,613	4,567	2,952	325	91	-	-	18,548
합계	381,081	433,472	447,834	421,654	443,722	407,971	67,038	2,602,772

(3) 오류 주석 말뭉치

○ 문어 오류 주석 말뭉치는 총 600,807어절 중 생활문이 294,682어절로 높은 비중을 차지하였고, 이어서 논설문 195,444어절, 설명문 50,247어절을 차지하였다.

<표 13> 2015-2021년 국립국어원 한국어 학습자 말뭉치 장르별 통계:
오류 문어 말뭉치

장르	1급	2급	3급	4급	5급	6급	합계
생활문	83,679	80,634	71,887	37,228	10,987	10,267	294,682
논설문	53	987	12,201	30,583	79,770	71,850	195,444
설명문	3,182	10,408	10,095	10,617	12,403	3,542	50,247
기행문	442	10,039	3,194	2,785	307	1,978	18,745
수필	-	166	592	3,242	4,764	7,056	15,820
기사문	-	-	-	181	812	9,071	10,064
감상문	-	114	5,802	2,341	671	666	9,594
편지글	2,569	1,736	1,061	-	-	-	5,366
전기문	-	-	-	-	-	845	845
합계	89,925	104,084	104,832	86,977	109,714	105,275	600,807

2) 구어

(1) 원시 말뭉치

○ 구어 원시 말뭉치는 인터뷰가 총 1,522,788어절 중 978,900어절로 가장 높은 비중을 차지하였고, 이어서 발표 280,981어절, 내러티브 156,844어절, 자유 대화 106,063어절이 구축되었다.

<표 14> 2015-2021년 국립국어원 한국어 학습자 말뭉치 장르별 통계:
원시 구어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
인터뷰	218,618	227,246	256,427	140,746	67,207	64,524	4,132	978,900
발표	30,166	46,690	63,854	70,082	45,806	19,042	5,341	280,981
내러티브	25,255	26,700	22,182	31,994	26,485	23,019	1,209	156,844
자유 대화	3,869	13,331	64,210	3,304	-	1,292	20,057	106,063
합계	277,908	313,967	406,673	246,126	139,498	107,877	30,739	1,522,788

(2) 형태 주석 말뭉치

○ 구어 형태 주석 말뭉치는 총 1,101,671어절 중 인터뷰가 779,765어절로 가장 높은 비중을 차지하였고, 이어서 발표 194,338어절, 내러티브 101,408어절, 자유 대화 26,160어절이 구축되었다.

<표 15> 2015-2021년 국립국어원 한국어 학습자 말뭉치 장르별 통계: 형태 구어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
인터뷰	183,007	171,144	158,515	134,369	66,077	63,737	2,916	779,765
발표	18,068	16,132	42,780	48,584	45,516	18,320	4,938	194,338
내러티브	11,901	10,882	8,522	24,904	24,673	20,526	-	101,408
자유 대화	1,546	7,369	502	2,193	-	1,292	13,258	26,160
합계	214,522	205,527	210,319	210,050	136,266	103,875	21,112	1,101,671

(3) 오류 주석 말뭉치

○ 구어 오류 주석 말뭉치는 총 548,718어절 중 인터뷰가 409,389어절로 가장 높은 비중을 차지하였고, 이어서 발표 111,978어절, 자유 대화 13,719어절, 내러티브 6,632어절이 구축되었다.

<표 16> 2015-2021년 국립국어원 한국어 학습자 말뭉치 장르별 통계: 오류 구어 말뭉치

장르	1급	2급	3급	4급	5급	6급	6급 이상	합계
인터뷰	82,604	89,201	78,066	78,592	34,216	45,831	879	409,389
발표	10,522	6,930	23,861	26,023	35,174	8,278	1,190	111,978
자유 대화	325	6,892	502	1,885	-	1,292	2,823	13,719
내러티브	446	281	1,697	1,238	2,574	396	-	6,632
합계	93,897	103,304	104,126	107,738	71,964	55,797	4,892	541,718

5. 주제별 말뭉치의 구축 비율

1) 문어

(1) 원시 말뭉치

○ 원시 말뭉치는 문어는 총 3,669,487어절 중 '사회'가 1,014,454어절로 가장 높은 비중을 차지하였고, 이어서 '일상생활'이 821,617어절, '개인 신상'이 546,410어절 구축되었다.

<표 17> 2015-2021년 국립국어원 한국어 학습자 말뭉치 주제별 통계: 원시 문어 말뭉치

주제 범주	1급	2급	3급	4급	5급	6급	6급 이상	합계
사회	16,742	26,094	52,572	270,952	345,734	242,389	59,971	1,014,454
일상생활	132,210	199,266	213,232	127,863	83,094	59,784	6,168	821,617
개인 신상	71,255	110,373	152,819	111,753	68,891	25,373	5,946	546,410
여가와 오락	87,983	37,930	139,112	19,877	30,988	15,728	5,102	336,720
대인 관계	34,883	79,820	32,943	46,727	14,734	52,142	1,162	262,411
여행	33,493	104,605	22,506	10,504	4,819	7,718	-	183,645
교육	5,073	2,632	13,955	10,351	56,415	29,025	28,464	145,915
일과 직업	983	1,502	4,796	23,276	54,531	11,266	1,665	98,019
건강	2,046	6,689	35,020	13,717	18,032	8,742	4,088	88,334
전문 분야	1,033	1,632	2,085	1,361	35,867	4,275	23,876	70,129
기후	30,083	9,204	1,047	-	315	260	-	40,909
쇼핑	12,063	9,005	6,385	4,085	8,771	303	-	40,612
주거 환경	3,457	1,438	12,180	2,927	123	433	1,281	21,839
식음료	3,322	4,842	402	6,438	183	632	500	16,319
교통	4,844	1,155	243	-	-	-	2,417	8,659
예술	-	254	58	420	-	1,351	-	2,083
공공 서비스	142	122	893	-	130	125	-	1,412

합계	439,612	596,563	690,248	650,251	722,627	459,546	140,640	3,699,487

(2) 형태 주석 말뭉치

○ 형태 주석 말뭉치는 문어는 총 2,602,772어절 중 '사회'가 660,484어절로 가장 높은 비중을 차지하였고, 이어서 '일상생활'이 597,590어절, '개인 신상'이 397,551어절 구축되었다.

<표 18> 2015-2021년 국립국어원 한국어 학습자 말뭉치 주제별 통계: 형태 문어 말뭉치

주제 범주	1급	2급	3급	4급	5급	6급	6급 이상	합계
사회	16,742	19,825	36,852	158,778	194,110	206,828	27,349	660,484
일상생활	106,827	140,919	136,164	91,799	65,819	55,323	739	597,590
개인 신상	64,477	88,699	101,100	67,730	47,966	24,334	3,245	397,551
여가와 오락	78,899	32,996	90,954	16,716	22,704	14,573	-	256,842
대인 관계	30,134	50,981	19,477	32,944	6,987	51,294	429	192,246
여행	29,751	72,080	18,521	8,960	4,470	7,351	-	141,133
교육	5,073	2,231	9,540	6,635	28,342	24,206	9,771	85,798
일과 직업	983	1,376	4,268	19,459	37,975	11,201	-	75,262
건강	2,046	6,612	15,275	11,178	13,700	6,669	2,222	57,702
전문 분야	1,033	1,632	2,085	1,225	17,431	3,608	19,763	46,777
기후	25,391	5,318	1,130	-	315	260	-	32,414
쇼핑	9,385	4,403	4,301	623	3,650	303	-	22,665
주거 환경	3,347	1,269	7,391	2,215	123	433	1,103	15,881
식음료	3,322	3,890	402	2,972	-	632	-	11,218
교통	3,529	865	97	-	-	-	2,417	6,908
예술	-	254	58	420		831	-	1,563
공공 서비스	142	122	302	-	130	125	-	821
합계	381,081	433,472	447,834	421,654	443,722	407,971	67,038	2,602,772

(3) 오류 주석 말뭉치

○ 오류 주석 말뭉치 문어는 총 600,807어절 중 '일상생활'이 152,845어절, 이어서 '사회'가 151,187어절로 높은 비중을 차지하였고, 이어서 '개인 신상'이 80,274어절, '여가와 오락'이 61,209어절 구축되었다.

<표 19> 2015-2021년 국립국어원 한국어 학습자 말뭉치 주제별 통계: 오류 문어 말뭉치

주제 범주	1급	2급	3급	4급	5급	6급	합계
사회	8,626	8,632	14,707	26,540	47,863	46,477	152,845
일상생활	23,503	32,966	32,491	22,794	16,760	22,673	151,187
개인 신상	17,056	19,356	14,375	10,290	10,275	8,922	80,274
여가와 오락	13,384	7,890	21,463	5,703	7,288	5,481	61,209
대인 관계	9,352	11,442	4,477	8,138	1,819	8,747	43,975
여행	4,514	14,019	5,887	7,031	2,398	2,610	36,459
일과 직업	494	576	3,385	2,192	8,890	2,987	18,524
교육	2,044	716	502	1,240	7,179	3,970	15,651
건강	1,315	3,448	2,515	1,162	1,931	2,088	12,459
전문 분야	612	800	634	1,028	4,751	308	8,133
쇼핑	2,214	745	2,526	290	560	-	6,335
기후	4,329	1,011	382	-	-	125	5,847
식음료	1,756	1,759	88	371	-	96	4,070
주거 환경	379	493	1,232	198	-	125	2,427
예술	-	-	58	-	-	666	724
교통	347	231	97	-	-	-	675
공공 서비스	-	96	-	-	-	-	96
합계	89,925	104,084	104,832	86,977	109,714	105,275	600,807

2) 구어

(1) 원시 말뭉치

○ 원시 말뭉치 구어는 총 1,522,788어절 중 '개인 신상'이 650,220어절로 가장 높은 비중을 차지하였고, 이어서 '일상생활'이 444,601어절, '여가와 오락'이 101,758어절로 다른 주제에 비해 상대적으로 많이 구축되었다.

<표 20> 2015-2021년 국립국어원 한국어 학습자 말뭉치 주제별 통계: 원시 구어 말뭉치

주제 범주	1급	2급	3급	4급	5급	6급	6급 이상	합계
개인 신상	132,458	155,349	183,609	93,016	43,527	39,653	2,608	650,220
일상생활	70,154	91,009	140,540	62,783	42,483	21,542	16,090	444,601
여가와 오락	35,805	21,692	18,867	11,844	4,974	8,576	-	101,758
사회	1,978	2,779	13,562	19,570	17,157	9,623	1,545	66,214
교육	1,018	910	6,054	10,750	7,729	21,934	5,235	53,630
대인 관계	2,484	15,745	6,106	13,830	8,409	2,506	1,465	50,545
여행	4,768	9,762	12,581	4,671	514	-	-	32,296
기후	20,211	4,279	1,516	505	572	-	-	27,083
건강	2,153	921	11,708	4,208	2,490	647	-	22,127
식음료	2,785	5,575	4,234	3,429	2,732	589	-	19,344
일과 직업	136	-	4,335	5,970	4,886	2,201	-	17,528
주거 환경	454	3,542	1,604	10,117	1,261	-	-	16,978
예술	413	766	205	4,771	231	-	-	6,386
전문 분야	-	236	324	196	814	-	3,796	5,366
쇼핑	1,272	517	1,428	466	1,362	-	-	5,045
교통	123	885	-	-	357	606	-	1,971
공공 서비스	1,696	-	-	-	-	-	-	1,696
합계	277,908	313,967	406,673	246,126	139,498	107,877	30,739	1,522,788

(2) 형태 주석 말뭉치

○ 형태 주석 말뭉치 구어는 총 1,101,671어절 중 '개인 신상'이 393,578어절로 가장 높은 비중을 차지하였고, 이어서 '일상생활'이 330,820어절로 높은 비중을 차지하였다. 그 외의 주제는 '여가와 오락' 86,040어절, '사회' 58,208어절 등으로 두 가지 주제에 비해 비중이 현저하게 낮아졌다.

<표 21> 2015-2021년 국립국어원 한국어 학습자 말뭉치 주제별 통계: 형태 구어 말뭉치

주제 범주	1급	2급	3급	4급	5급	6급	6급 이상	합계
개인 신상	84,686	81,270	74,817	74,383	40,585	36,438	1,399	393,578
일상생활	61,828	66,584	77,519	53,156	42,483	21,175	8,075	330,820
여가와 오락	31,915	17,792	11,864	10,919	4,974	8,576	-	86,040
사회	1,878	614	11,384	17,121	16,867	9,203	1,142	58,209
교육	1,018	910	2,394	9,879	7,729	21,934	5,235	49,099
대인 관계	2,484	14,753	5,494	13,455	8,409	2,506	1,465	48,566
기후	19,900	3,851	1,516	505	572	-	-	26,344
여행	2,211	8,749	6,307	3,406	514	-	-	21,187
식음료	2,785	5,575	4,234	3,429	2,732	589	-	19,344
일과 직업	136	-	4,335	5,970	4,886	2,201	-	17,528
주거 환경	454	3,542	1,604	10,117	1,261	-	-	16,978
건강	2,153	-	7,218	3,346	2,490	647	-	15,854
예술	413	766	205	3,702	231	-	-	5,317
전문 분야	-	236	-	196	814	-	3,796	5,042
쇼핑	842	-	1,428	466	1,362	-	-	4,098
교통	123	885	-	-	357	606	-	1,971
공공 서비스	1,696	-	-	-	-	-	-	1,696
합계	214,522	205,527	210,319	210,050	136,266	103,875	21,112	1,101,671

(3) 오류 주석 말뭉치

○ 오류 주석 말뭉치 구어는 총 541,718어절 중 '개인 신상'이 201,408어절, '일상생활'이 148,403어절로 높은 비중을 차지하였다. 그 외에는 '여가와 오락'이 39,684어절, '교육'이 31,980어절, '사회'가 31,260어절로 두 가지 주제에 비해 현저하게 적었다.

<표 22> 2015-2021년 국립국어원 한국어 학습자 말뭉치 주제별 통계: 오류 구어 말뭉치

주제 범주	1급	2급	3급	4급	5급	6급	6급 이상	합계
개인 신상	44,160	48,390	40,394	39,325	14,420	14,719	-	201,408
일상생활	28,322	31,465	32,086	25,764	19,392	8,607	2,767	148,403
여가와 오락	12,048	5,638	5,991	5,860	1,571	8,576	-	39,684
교육	330	492	707	7,586	5,624	16,306	935	31,980
사회	119	614	1,901	10,742	12,951	4,933	-	31,260
대인 관계	449	4,782	5,494	4,210	6,509	838	-	22,282
식음료	2,700	1,870	4,234	228	2,732	589	-	12,353
일과 직업	136	-	1,541	5,256	3,875	582	-	11,390
여행	948	1,931	5,063	2,731	514	-	-	11,187
건강	-	-	3,632	1,263	1,615	647	-	7,157
주거 환경	454	3,542	1,604	271	1,261	-	-	7,132
기후	3,279	3,313	-	138	-	-	-	6,730
예술	413	766	205	3,702	231	-	-	5,317
쇼핑	416	-	1,274	466	912	-	-	3,068
전문 분야	-	236	-	196	-	-	1,190	1,622
교통	123	265	-	-	357	-		745
합계	93,897	103,304	104,126	107,738	71,964	55,797	4,892	541,718

부록 2. 2022년 한국어 학습자 말뭉치 구축 지침

차 례

한국어 학습자 말뭉치 자료 수집 지침

1. 자료 수집 대상 및 수집 자료

▶ 대상: 한국어 교육기관의 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

문어 자료	구어 자료
⇓	⇓
작문지 및 학습자 동의서 수집	음성 발화 자료 녹음 및 학습자 동의서 수집
⇓	⇓
학습자 동의서와 작문지 합철 (스테이플러 사용)	파일 변환 및 파일명 부여 (수준, 반, 학습자 일련번호 조합)
⇓	⇓
일련번호 매기기	학습자 동의서와 파일 매칭 (1인 발화만 대상으로 함)
박스(또는 봉투) 포장 후 착불 우편으로 발송	동의서 스캔 후 파일과 함께 메일 발송
	⇓
	동의서 원본 착불 우편으로 발송

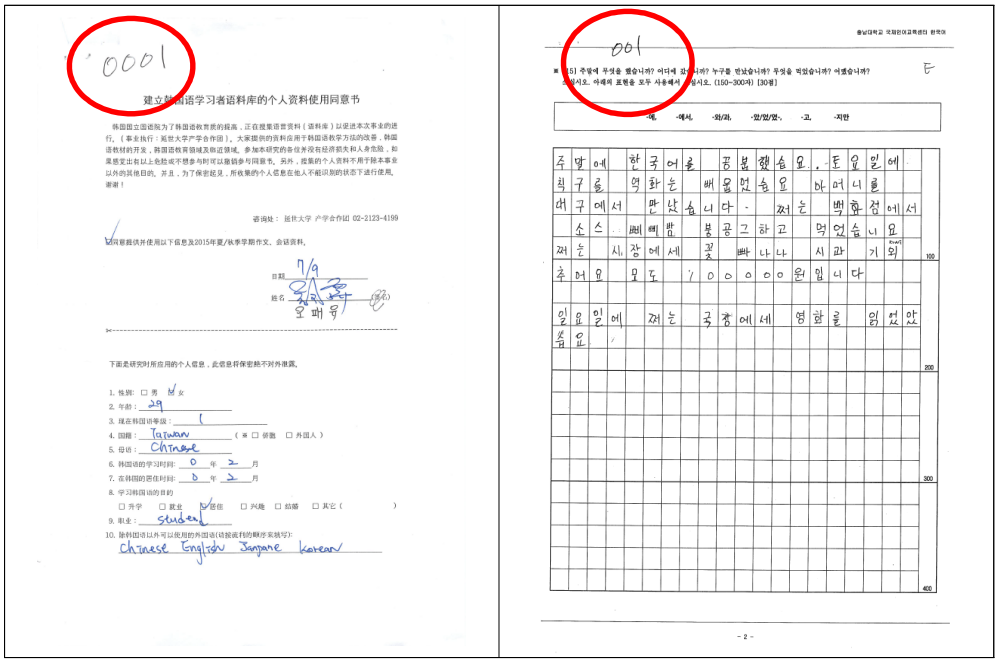
1) 문어

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.



☞ 급별로 번호를 붙이며 각각 0001로 시작한다.

☞ 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 교육기관의 교육과정 실러버스에 이미 포함되어 있는 작문을 활용하여 이를 수집함. 글의 종류 및 주제는 각 기관의 교육과정에 따름

② 성취도 평가 수집
○ 각 교육기관의 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 주제로 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 학습자 간의 2인 발화의 경우에는 각 학습자가 녹음하여 이를 교사에게 전송

하게 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, 가능하면 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	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	여러분은 취미로 무엇을 배우고 싶습니까? '내가 취미로 배우고 싶은 것'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 취미로 배우고 싶은 것은 무엇인가? (※ 한국어를 배우고 싶다는 내용은 쓰지 마십시오.) ○ 왜 그것을 배우고 싶은가? ○ 그것을 배운 후에 무엇을 하고 싶은가?	
28주차	여러분은 늦잠을 자거나 누워서 책을 보는 것과 같은 고치고 싶은 생활 습관이 있습니까? '고치고 싶은 나의 생화 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	갖고 싶은 직업'이라는 제목으로 글을 써 보십시오. 단 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 직업명, 하는 일, 그 일을 하려는 이유, 그 일에 필요한 조건	
34주차	나의 성격'이라는 제목으로 글을 써 보십시오. 단 아래에 제시된 내용에 모두 포함되어야 합니다. ○ 성격의 특징, 장점과 단점, 고치고 싶은 부분과 그 이유	
36주차	내가 생각하는 행복'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 행복을 위해서 어떤 노력을 하는가?	

수집 시기	문제	수준
	○ 언제 행복하다고 느끼는가? ○ 행복은 무엇이라고 생각하는가?	
38주차	'내가 좋아하는 책'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 가장 좋아하는 책은 무엇인가? ○ 그 책은 어떤 내용인가? ○ 그 책을 좋아하는 이유는 무엇인가?	
40주차	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. ○ 닮고 싶은 사람은 누구인가? ○ 왜 그 사람처럼 되고 싶은가? ○ 그 사람처럼 되기 위해서 어떻게 해야 하는가?	
42주차	1)~4)의 내용은 '피로를 예방하려면 네 가지를 실천하라'는 글의 소재입니다. 이 소재를 이용하여 글을 쓰십시오. '피로를 예방하려면 네 가지를 실천하라' ○ 체질에 맞는 음식 ○ 수면의 질 ○ 적당한 운동 ○ 긍정적인 사고	고급
44주차	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. <올바른 인터넷 사용 태도의 예> ○ 상대방의 의견 존중하기 ○ 타인의 사생활 보호하기 ○ 의견 차이 인정하기 ○ 바른 언어 사용하기 ○ 정확한 정보 올리기	
46주차	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 포함하야 합니다. <신문의 기능>	

<table>
<tr><th>수집
시기</th><th>문제</th><th>수준</th></tr>
<tr><td></td><td>○ 사건 보도
○ 여론 조성
○ 정보 제공
○ 소통의 분위기 조성</td><td></td></tr>
<tr><td>48주차</td><td>다음 글을 읽고 '감시 카메라 설치 확대'에 대한 자신의 견해를 서술하십시오. (찬성하거나 반대하는 입장중 하나를 선택하여 서술 할 것. 단 아래 제시된 각 입장의 논거 중 두 개 이상을 제시할 것.)

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?

<table><tr><th>찬성</th><th>반대</th></tr><tr><td>사회 안전 유지
범죄 예방
인권보다 공인이 우선</td><td>개인의 사상활 침해
범죄 예방 효과 불분명
가해자의 인권 보호</td></tr></table></td><td></td></tr>
<tr><td>50주차</td><td>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.
○ 내가 생각하는 성공이란 무엇인가?
○ 그것을 이루기 위해 필요한 것은 무엇인가
○ 그 이유는 무엇인가?</td><td></td></tr>
<tr><td>52주차</td><td>여러분은 무엇이 선의의 거짓말이라고 생각합니까? 어떤 경우에 그런 거짓말을 할 수 있다고 생각합니까? 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다.

< 선의의 거짓말이란 >
○ 선의의 거짓말이란 무엇인가?
○ 선의의 거짓말은 언제 필요한가?</td><td></td></tr>
</table>

수집 시기	문제	수준
	○ 선의의 거짓말이 가질 수 있는 문제점은 무엇인가?	
54주차	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. < 예술 교육의 필요성 > ○ 예술 교육이 왜 필요한가? ○ 예술 교육을 통해 얻을 수 있는 효과는 무엇인가?	
56주차	자연을 그대로 보존해야 한다는 주장과 인간을 위해 자연을 개발해야 한다는 주장이 있습니다. 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. <자연 보존과 자연 개발> ○ 자연 보존과 자연 개발 중 어느 것이 더 중요하다고 생각하는가? ○ 그렇게 생각하는 이유는 무엇인가? (2가지 이상 쓰시오.)	
58주차	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. 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	

수집 시기	문제	수준
	에 적응하지 못하는 이들은 사회 변화의 뒷전으로 밀려날 수밖에 없게 된다. 위의 글에 나타난 두 가지 유형의 직업관 중 자신의 생각은 어느 쪽인지 말하고, 그 이유를 설득력 있게 글로 나타내시오.	
60주차	현대 사회는 빠르게 세계화·전문화되고 있습니다. 이러한 현대 사회의 특성을 참고하여, '현대 사회에서 필요한 인재'에 대해 아래의 내용을 중심으로 자신의 생각을 쓰십시오. ○ 현대 사회에서 필요한 인재는 어떤 사람입니까? ○ 이러한 인재가 되기 위해서 어떤 노력이 필요합니까?	

2. 구어 수집

수집 시기	글의 종류	주제	수준
02주차	인터뷰	소개(자신, 가족 등)	초급
04주차	발표	취미	
06주차	인터뷰	주말	
08주차	발표	한국생활	
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
(결혼이주민, 이주노동자)

1. 문어

종적 자료 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? 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.	
22주차	여러분은 10년 후에 어떻게 살고 싶습니까? 그 이유는 무엇입니까? '10년 후의 나의 계획'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시한 내용이 모두 포함되어야 합니	중급

종적 자료 수집 시기	문제	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	‘살고 싶은 집’이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 어디에 살고 싶은가? ○ 어떤 집에 살고 싶은가? 왜 그런가? ○ 집은 어떻게 꾸미고 싶은가? ○ 집에서 누구와 무엇을 하고 싶은가?	
36주차	고향의 음식을 소개하는 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다. ○ 음식 이름 ○ 주로 언제 먹는 음식인가? ○ 어떻게 만드는가? ○ 한국 음식과 비슷한 음식이 있는가? 어떤 점이 비슷하고 어떤 점이 다른가?	
38주차	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. ○ 닮고 싶은 사람은 누구인가? ○ 왜 그 사람처럼 되고 싶은가? ○ 그 사람처럼 되기 위해서 어떻게 해야 하는가?	
40주차	취업을 하려고 합니다. 무슨 일을 하고 싶은지 생각해 보고 자기 소개서를 쓰십시오. ○ 살아온 과정 ○ 성격의 장단점 ○ 지금까지의 경험 또는 경력 ○ 앞으로의 계획	
42주차	‘내가 생각하는 행복'이라는 제목으로 글을 쓰십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. ○ 행복을 위해서 어떤 노력을 하는가? ○ 언제 행복하다고 느끼는가? ○ 행복은 무엇이라고 생각하는가?	고급
44주차	‘절약과 저축’이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다. ○ 절약하기 위해 무엇을 하는가? ○ 저축을 하고 있는가? 왜 그런가?	

종적 자료 수집 시기	문제	수준
	○ 돈을 모으면 무엇을 하고 싶은가?	
46주차	'텔레비전이 우리 생활에 미치는 영향'이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다. ○ 텔레비전을 자주 보는가? 왜 그런가? ○ 무슨 프로그램을 자주 보는가? 왜 그런가? ○ 텔레비전이 우리 생활에 미치는 긍정적인 영향은 무엇인가? 부정적인 영향은 무엇인가?	
48주차	'효과적인 자녀 교육법'이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다. ○ 현재 자녀가 있는가? ○ 자녀가 말을 듣지 않을 때 어떻게 하는가? (현재 자녀가 없는 경우, 부모의 말을 듣지 않는 아이를 어떻게 하면 좋을까?) ○ 자녀와 대화를 잘하려면 어떻게 해야 하는가? ○ 어떤 부모가 되고 싶은가? ○ 자녀 교육을 어떻게 하고 싶은가?	
50주차	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. ○ 내가 생각하는 성공이란 무엇인가? ○ 그것을 이루기 위해 필요한 것은 무엇인가 ○ 그 이유는 무엇인가?	
52주차	여러분은 무엇이 선의의 거짓말(좋은 거짓말)이라고 생각합니까? 어떤 경우에 그런 거짓말을 할 수 있다고 생각합니까? 이에 대한 자신의 견해를 서술하십시오. 단, 아래에 제시된 내용이 모두 포함되어야 합니다. < 선의의 거짓말이란 > ○ 선의의 거짓말이란 무엇인가? ○ 선의의 거짓말은 언제 필요한가? ○ 선의의 거짓말이 가질 수 있는 문제점은 무엇인가?	
54주차	'노후 준비'라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다.	

종적 자료 수집 시기	문제	수준
	○ 노후 준비가 왜 필요한가? ○ 노후를 위해 무엇을 준비해야 하는가? ○ 여러분은 노후를 어떻게 준비하고 있는가?	
56주차	'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. ○ 칭찬이 미치는 긍정적인 영향은 무엇입니까? ○ 부정적인 영향은 무엇입니까? ○ 효과적인 칭찬의 방법은 무엇입니까?	
58주차	'여성의 사회적 지위와 역할'이라는 제목으로 글을 쓰십시오. 단, 다음의 내용이 모두 포함되어 있어야 합니다. ○ 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	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. 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 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 안	

종적 자료 수집 시기	문제	수준
	정감을 잃어버릴 위험이 많고, 이런 변화 속에 적응하지 못하는 이들은 사회 변화의 뒷전으로 밀려날 수밖에 없게 된다. 위의 글에 나타난 두 가지 유형의 직업관 중 자신의 생각은 어느 쪽인지 말하고, 그 이유를 설득력 있게 글로 나타내시오. (200자 내외)	

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 사람이에요? 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.	
42주차	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	고급

종적 자료 수집 시기	문제	수준
	있어요. 그리고 저는 눈이 작고 쌍커풀이 없어요. 그래서 성형 수술을 하고 싶어요. 여러분도 자신의 외모에 대해 고민을 해 봤어요? 여러분은 자신의 아름다움을 위해 어떤 노력을 하고 있어요? 글에서 읽은 친구의 고민에 대한 여러분의 생각을 써 보세요.	
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 뭘까요? 부정적인 영향은 뭘까요? '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	
42주차	인터뷰	★ 중등 외모 관련 텍스트 읽기 ★ 성격과 외모	고급
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>[17]

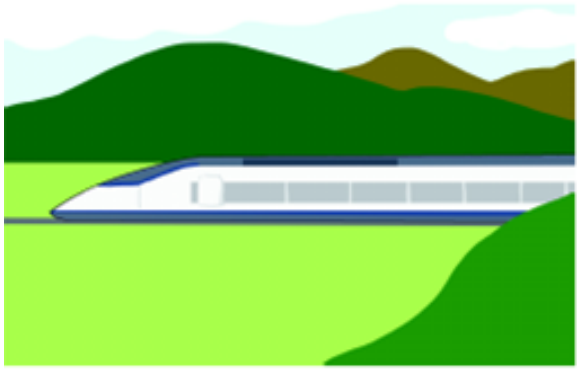
안녕하세요? 저는 송안나입니다. 대한초등학교 학생입니다. 저는 1학년 5반입니다. 우즈베키스탄에서 왔어요. 지금은 대림동에 살아요. 우리 집은 학교 근처에 있어요. 저는 컴퓨터 게임을 좋아해요. 만나서 반갑습니다.

<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 팥빙수도 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 낍니다.
저는 사계절 중에서 추운 겨울을 제일 좋아합니다. 눈사람도 만들도 눈싸움도 할 수 있습니다.

17) 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 먹어봐!"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>[18]

18) 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 나빴을 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 한 번 사과를 해야겠다.

<60주차 그림 자료>[19]

※ 다음 그림을 보고 이야기를 만들어 보세요.

1

2

19) 그림은 '키즈짱 잼잼동화-젊어지는 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 과제(1차)

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? 국제결혼은 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 주제로의 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 구축 사업을 위한 학습자 자료 이용 동의서(일반)

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

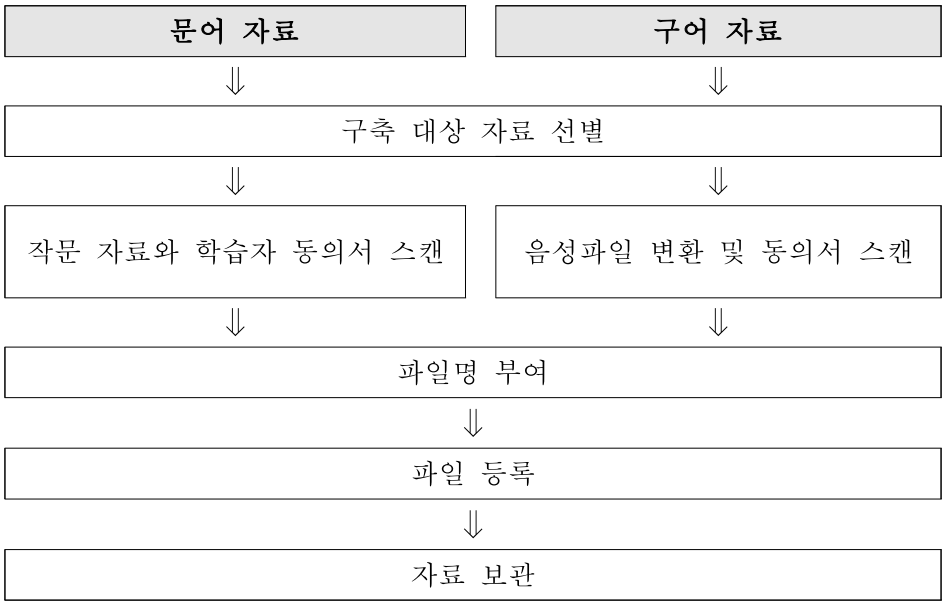
15. 한국어로 나오는 방송 매체(TV, 라디오, 인터넷 동영상)를 보는 시간은 얼마나 됩니까?

☐ 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

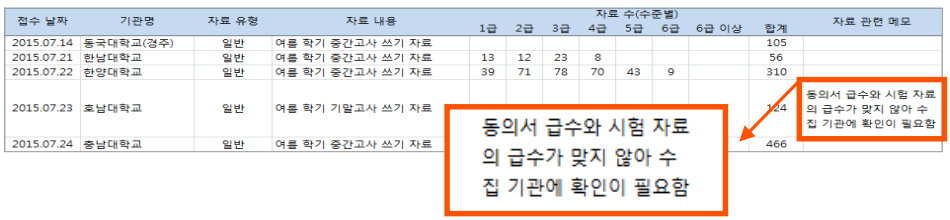
○ 말뭉치 구축을 위해서는 IRB 규정에 따라 학습자의 서명이 완료되고 자료의 활용을 위해 필요한 개인 정보가 빠짐없이 입력이 되어야 한다. 그 외에도 다음과 같은 기준으로 우선적으로 구축할 자료를 선정하도록 한다.

문어	구어
○ 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	
○ 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	○ 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.



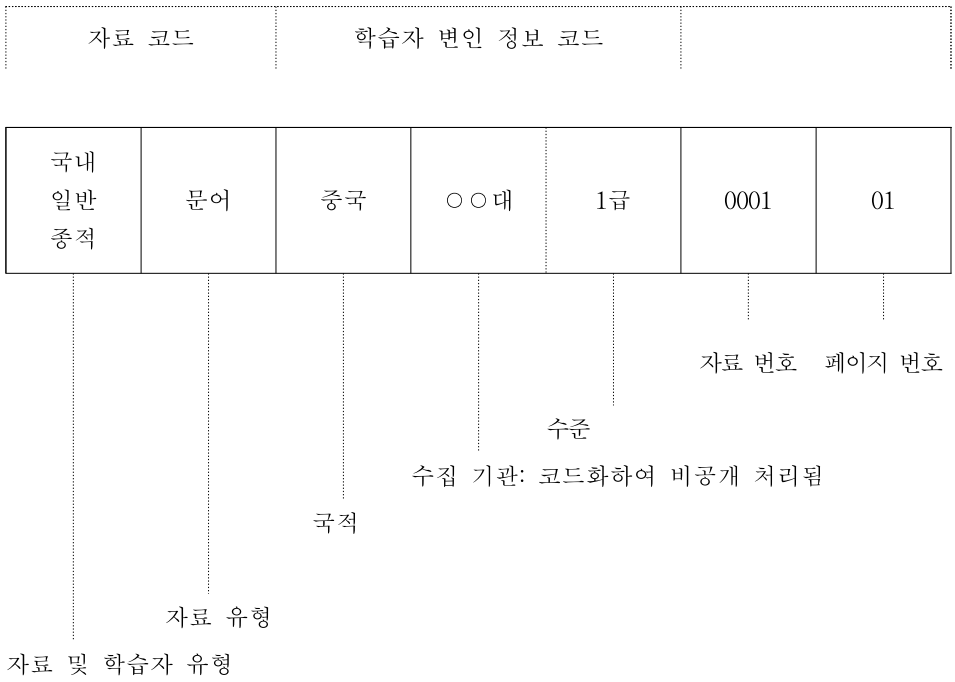
3) 일련번호 부여

○ 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



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.
○ 필적을 알아보기 어려운 것은 일단 가장 가까운 상태로 입력한다.
○ 단락을 구분하여, 문장 단위로 입력한다. 단락은 자판의 엔터키로 구분하고, 들여쓰기는 반영되지 않는다.
○ 전체 본문 입력이 끝나면 '주석 자동 생성'을 클릭하여 본문 주석을 확인하고 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.

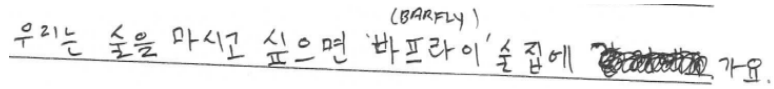
<예> 1/2, 3/4

○ 영문자, 한글 자모, 괄호문자 등은 자판을 사용하여 입력한다.

<예> ㄱㄴㄷㄹ, (1) (2) (3)

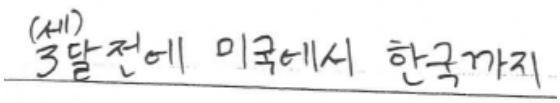
○ 외국어를 함께 쓴 경우 다음과 같이 원문에 따라 병기한다. 단, 입력과 해석의 용이성을 고려하여 영어와 한자에 한정한다.

<예> 아래의 경우 '바프라이(BARFLY)'로 입력한다.



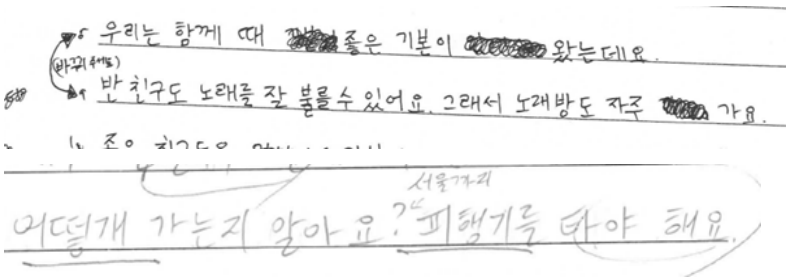
○ 숫자와 한글 표기를 함께 쓴 경우 원문에 따라 병기한다.

<예> 아래의 경우 '3(세) 달 전'으로 입력한다. 이때 '3달(세 달)'과 같이 동일한 표기가 두 번 이상 입력되도록 하지 않는다.



○ 학습자가 작문 중간에 교정 기호를 사용하거나 교정에 관한 문구를 적어 넣은 경우 이를 반영해서 수정 입력한다. 단, 학습자의 답안에 교사가 같은 색으로 수정 또는 채점을 한 경우, 학습자가 작성하면서 스스로 수정한 것인지 교사가 수정한 것인지 선별해야 한다.

<예> 반 친구도 노래를 잘 불를 수 있어요.
그래서 노래방도 자주 가요.
우리는 함께 때 좋은 기본이 왔는데요.
어떻개 가는지 알아요? 서울까지 피행기를 타야 해요.



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.

○ 문장부호는 학습자가 적어 넣은 대로 입력한다. 즉, 문장부호의 누락이나 생략, 중복 등을 그대로 반영한다.

○ 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.

- '외국문자'는 영어와 한자 이외의 외국어를 입력할 때 ¿ 기호 입력 후 마크업할 때 사용한다.

<예> <EX_Alpha>¿¿¿¿¿</EX_Alpha>

- '식별불가'는 원본에서 다양한 이유로 확인이 어려운 문자나 기호에 대해 ¿로 입력 후 마크업한다.

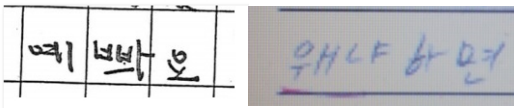
<예> <CNI>¿¿</CNI>

○ '기타기호'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. ('기타기호'는 원본 그대로 입력하므로 ¿기호를 입력하지 않도록 주의한다.)

<예> 1) <EX_Symbol>『』 「」 《》 </EX_Symbol> : [괄호기호]'ㄴ' 입력 후 '한자키' 눌러서 선택 2) <EX_Symbol>※★</EX_Symbol> : [일반기호]'ㅁ' 입력 후 '한자키' 눌러서 선택 3) <EX_Symbol>㎡ ㎏ ㎉</EX_Symbol> :[단위기호] 'ㄹ' 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 '한글기호' 주석을 사용하여 마크업한다.

<예>



1) 예**쁘ㅏ**요 : 시스템 '예¿요'로 입력 후 '한글기호' 마크업 처리

- 예<NSS>¿</NSS>요

2) **우ㅐ**냐하면 : 시스템 '¿냐하면'으로 입력 후 '한글기호' 마크업 처리

- <NSS>¿</NSS>냐하면

3) 좌우대칭된 ㄱ이 포함된 '가'

- <NSS>¿</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, 전화 번호 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
- 카드번호 : 학습자의 개인 신용카드 번호 등 ☞ <Privacy_CardNum> 태그
- 기타 : 개인식별 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 단어 이내로 누가 봐도 추정 가능한 내용일 경우에는 해당 내용을 적어 입력한다. 그 외에는 입력 대상에서 제외한다.

6. 최고급 자료 마크업

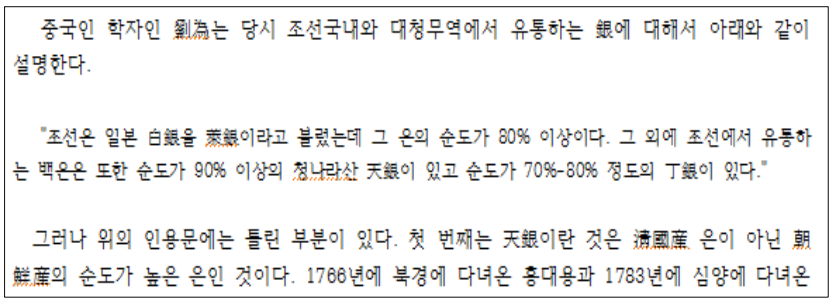
○ 최고급 자료의 입력은 기존 지침을 동일하게 적용한다.
○ 기존의 마크업과 더불어 형식과 내용을 구분하기 위해 아래의 마크업을 사용한다.

	주석	내용	주석 표지
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	<Forabs>

	주석	내용	주석 표지
		및 주제어	
	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 표지	<gap reason>

- 각주 미주: 본문과 각주 내용에 각각 각주표시 1),2)를 남기고 해당 각주는 본문뒤, 참고문헌 앞으로 이동 후 <ft> 태그
- 예문 인용: 본문 내에서 문단 구분되어 하나의 단락으로 삽입된 부분을 <q>태그

<예>



중국인 학자인 劉為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 清國産 은이 아닌 朝鮮産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 淸國産 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
발화자 정보	화자 1의 표시	1	<person id=1 sex=M age=20s> 1:
	화자가 불분명할 때	?	P?:
	동시 발화	모두/ 나머지/ 2,3	2,3:네.
억양 단위	하강	.	2:네.
	상승	?	2:어디 갈 거예요?
	약한 상승이나 하강	,	1:그래서 그랬는데 이번에,
	활기, 기운찬 어조	!	1:아!
	억양 단위 경계의 처리	엘란에서 발화 단위를 분할하면 줄 바꿈 처리됨	2:어디 갈 거예요? 1:안 아직까지 그냥 계획만 잡아 놨는데, 2:음.
	하나의 억양 단위가 끼어듦에 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

대분류	소분류	기호	예시
			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>연음되지 않음</note>	1:응. 2:
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)	소리나는 대로 적고, 원래의 형태가 없이 내용을 이해하기 어려울 때에는 ( ) 안에 학습자의 발화를, 괄호 밖에는 규범 표기를 밝힘	1: 친구(칭구)와 강남(간남)에 갔습니다(갔슨니다)
	숫자 표기	발음에 따라 한글로 표기	7:오늘 제 동생이 이케 하나 오백 원이라고 사 가지고 왔더라구.
	외래어・외국어 표기	발음에 따라 한글로 표기	2:어떻게 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	

대분류	소분류	기호	예시
	축약형	'(apostrophe, 영문따옴표)를 사용해서 두 음소를 연결	사귀'어, 바뀌'어, …
	표현적 장음	::	1:많은 경우에:: 논문, 저::~ 어::~ 연구는 네이션,
	담화표지	~	1:많은 경우에 논문, 그::~ 어::~ 연구는 네이션,
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='목청가다듬는 소리'>	-
	들이마시는 숨(쓰)	<vocal desc='들이마시는숨'>	-
	내쉬는 숨(후우)	<vocal desc='내쉬는숨'>	-
	혀 차는 소리(쯔)	<vocal desc='혀차는소리'>	-
	헛기침(에헴)	<vocal desc='헛기침'>	-
	한숨	<vocal desc='한숨'>	-
	노래	<vocal desc='노래'>	-
	웃으면서 말하는 부분	<@ @>	2:[<vocal desc='웃음'>] 너무 좀 <@오버한다.@>

대분류	소분류		기호	예시
	박수 치면서 말하는 부분		<# #>	5:[우우 <vocal desc='박수'>] <#이리 와 이리 와.#>
	노래를 부르는 부분		<M M>	-
	박수나 손가락 부딪치는 소리		<kinesics desc=' '>	박수 <kinesics desc='박수'>
	대화 흐름에 영향을 주는 전화벨 소리라든지 기타 음성 아닌 소리		<event desc=' '>	<event desc='전화벨소리'>
2차 전사	구어적 변이형	( ) 안에는 학습자의 발화를, ( ) 밖에는 철자 보충. 단, 주 등장하거나 쉽게 원래 형태로 이해될 수 있는 것들은 일일이 철자형을 붙여주지 않음	( )	1:브릿지를 한 가닥을 넣어(너) 줬어요.
	발음 오류	한국어 모어 화자의 발음과 음운적으로 구분이 될 정도로 발음에 오류가 있는 경우	( )	친구(칭구)와 강남(간남)에 갔습니다(갔슨니다). 같이(가티) 가자. 과반수(화반수)
		한국어 모어 화자의 발음과 음성 혹은 변이음의	( )	가구(가구)<note>'가'의 ㄱ을 유성음으로 발음</note> 가구(가구)<note>'구'의 ㄱ을 무성음으로 발음

대분류	소분류		기호	예시
		구분이 모호한 경우		</note>
		음운규칙으로 인해 한국어의 철자대로 발음되지 않으나, 학습자가 이를 철자대로 발음하는 경와 같이 철자 전사를 통해 학습자의 발음 오류를 반영하기 어려운 경우	( )	무조건(무조건)<note>경음화되지 않음</note> 같이(가티)<note>구개음화되지 않음</note> 신라(신라)<note>자음동화되지 않음</note> 앞에(앞에)<note>연음되지 않음</note> 먹는(먹는)<note>자음동화되지 않음</note>
	외국어, 외래어 발음	외국어나 외래어를 원어에 가깝게 발음할 경우	( )	인터뷰(이너뷰)
		한국어 모어 화자와 다른 규범을 할 경우	( )	카페: 현실 발음 [까페], 학습자 발음 [카페] 버스: 현실 발음 [뻐스], 학습자 발음 [버스] → 각각 '카페(카페)', '버스(버스)'로 처리함
	방언형 표시		확실한 방언형(대응하는 표준형 형태소가 없는 것)의 경우 태그 부착	차는 <dia>여일</dia> 있여.
	긴 휴지		(1초 이상의 쉼은)	2:{1.2} 그럴까?

대분류	소분류	기호	예시
		0.1초 단위까지 표시	
	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>	혹자들은 현대 사회에 대하여 <Q불확실성의 시대는 아니다.Q> 라고 말하죠.
	텔레비전 방송이나 강의 등 텍스트 종류 표현	<R R>	1:그 다음에 <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> : 신용카드 번호 <address> : 주소 <tel-num> : 전화번호	5: 그게 어쩌면 <name1> 선배님이라든지 다른 선배님들 말:: 들은 걸 생각해 보면,
발화 단위 분할(segmentation)	분할 단위는 어절 단위로 한다		내{1.2}/가 (X) 내{1.2}가 (O) 빗금은 분할 단위를 나타내는 단위이다.

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

○ 모든 발화자에 관한 정보는 시스템 등록 시 주석 입력 창에서 발화자의 기본 정보를 선택하여 입력한다.

<예> 1:학습자 2:원어민

○ 본문 전사에서 발화자 정보와 발화자 표시는 반드시 일치해야 하고 발화자가 분명하지 않을 경우에는 필요에 따라서는 '모두'나 '나머지' 등의 지칭을 사용할 수 있고, 화자 2와 화자 3이 동시에 말하는 경우는 '2,3'으로 표시하기도 한다.

<예> 1:어~ 물건을 바꾸고 싶은데요, 2:네,

○ 발화자 표시에 스페이스를 넣지 않는다.

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 기본단위인 문장은 종결어미로 마무리되고 마침표라는 문장부호로 인해 명확하게 그 단위를 설정할 수 있지만, 구어는 종결어미를 사용하여 발화를 끝내는 경우가 많지 않고, 억양이나 휴지 등의 운율적인 요소에 영향을 받기 때문에 기본단위를 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 문장부호를 사용한다.

하강 억양	.
상승 억양	?
약한 상승이나 하강 억양	,
활기에 넘치는 기운찬 어조(감탄의 끝)	!

○ 하나의 억양단위 경계에 스페이스 없이 엔터(enter)를 친다.

- 전사 도구에서는 세그멘테이션을 분할하는 것으로 대체한다.

<예> 2:어디 갈 거예요?
1:안 아직까지 계획이 없는데,
2:음.

나. 끊어진 억양단위(붙임표의 사용)

○ 계속해서 말을 할 의향이 있는데, 끼어듦을 당해서(혹은 적극적인 호응에 의해서) 말끝이 잘린 경우는 다음의 예와 같이 마침표를 쓰지 않고 붙임표(-)를 사용한다(단위의 끝에서는 앞쪽에만, 단위의 시작에서는 뒤쪽에만 스페이스 있음). 시간적 순서에 의해 표현된 발화를 억양단위로 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. 억양단위의 연속성

○ 두 억양단위가 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 겹친 부분을 따로 표시하지 않고 전사 도구를 통해 표시되는 발화 시간을 함께 제시한다[20].

4) 잘 들리지 않는 부분

○ 잘 들리지 않는 부분은 <X X>안에 전사한다. 문장부호 다음에 붙인다.

<예> 그때도,
<X보통X> 그런 자만심이 있었다.

20) 21세기 세종계획 말뭉치 구어 전사 지침에서는 연속적인 겹침과 비연속적인 겸칩, 동시다발적 발화 등 다양한 겹침 상황에 따른 세부 전사 지침을 제시하고 있다. 한국어 학습자 말뭉치에서 이러한 지침을 포함하지 않은 것은 전사 도구를 통해 제시되는 발화 시간 정보가 겹침에 관한 정보들을 대체하기 때문이다.

○ 화자의 발화 내용이 전혀 들리지 않는 부분은 <안들림>으로 전사한다. 이때는 억양을 확인할 수 없으므로 문장부호를 넣지 않는다. 억양단위의 끝 부분에 말줄임표가 올 때는 앞에만 스페이스를 두고, 억양단위의 시작 부분에 올 말줄임표가 올 때는 뒤에만 스페이스를 두며, 중간에 올 때는 양쪽에 스페이스를 둔다.

<예> 4:<vocal desc='웃음'> 1:거기까지 <note>안들림</note> 2:근데 그거 진짜 신고해야 되겠더라. 6:해야 돼. 2:<note>안들림</note> 너무한 거 같더라.

○ 들리지 않는 음절은 그 음절의 수만큼 X를 붙인다.

<예> 근데 그거 진짜 XX해야 되겠더라.

5) 전사 기호의 중복

○ 전사 기호가 서로 중복될 경우에는 기호의 유형을 고려하여 우선순위에 따라 기입한다. 기호의 유형은 다음과 같이 네 가지로 나눌 수 있다.

< >, ( ), 문장부호(온점, 느낌표, 물음표/쉼표), 기타

○ 전사 도구를 통한 전사 시 기호가 중복되어 사용되는 경우 시스템상의 오류가 발생할 수 있으므로 < > 기호 안에는 하나의 문장 단위만 입력한다. (억양을 나타내는 문장부호와 동시에 사용 불가하므로 분리하여 입력한다.)

○ 억양 기호는 바로 앞의 음절의 억양을 표시하므로 다른 기호보다 우선하여 사용한다. 웃음이나 박수 등의 준음성 표현이 있을 경우는 그 다음에 붙인다.

6) 전사자의 설명

○ 전사자가 특정 발화 구간에 대한 설명을 붙일 필요가 있을 경우 <note></note> 태그를 사용하여 표현한다. 녹음 상태, 잡음, 동시다발적인 대화,

특이한 발음 상태 등의 설명이 덧붙을 수 있다. 주석 태그는 다음과 같이 붙여야 하고 앞뒤로 엔터를 쳐 준다.

<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: 친구(칭구)와 강남(간남)에 갔습니다(갔슨니다).

☞ 한국어 학습자 말뭉치는 일반 말뭉치와 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 전

사하고 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 크림 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,

○ 발화자가 불완전하게 발화한 것은 아니지만 한 어절을 발화하는 도중에 억양 단위가 바뀌어서 조사나 어미 등 문법 형태소가 실질 형태소와 다른 억양 단위로 전사될 때, 조사나 어미에 '='를 붙인다.

<예> 1:주부 우울증, =이라고 말할 수 있겠습니다.

○ 그러나 완전한 어절이 완전 반복되는 발화의 경우는 '=' 표시를 하지 않고

전사만 한다.

○ 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)

○ 수를 적을 때는 만 단위로 띄어 쓴다.(예, 십이억 삼천백만 팔백구 불 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 겹홀소리가 되는 것 등이다. 구어 말뭉치에서는 발음되는 음절 수와 표기상의 음절 수를 맞추는 것이 원칙이므로 축약형의 경우 모두 표기에 반영한다. 그런데 모음의 축약형의 경우 대부분 현재 국어의 모음 체계상 표기할 글자가 존재하지만, 반홀소리된 /ㅟ/, /ㅚ/의 표기가 문제가 된다. /ㅟ/, /ㅚ/가 반홀소리가 되어 /ㅝ/, /ㅘ/와 축약되는 현상은 구어에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 구어 전사에서는 ’(apostrophe, 영문따옴표)를 사용해서 두 음소가 연결됨을 표시하고 구축 도구 내에서 마크업 처리한다.

<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 나타냄

으로써 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, 에, 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, 에, 엥, 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 물결표(~, 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(에헴) <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> 친구(칭구)와 강남(간남)에 갔습니다(같슨니다).

<예> 같이(가티) 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 붙여주지 않는다.

<예> 책상 위에 놔 둬.

○ 소유격 조사 '의'의 경우 한글 맞춤법 통일안에서는 [의]와 [에]를 모두 표준발음으로 인정하고 있다. 즉 〈표준 발음법〉 제5 항에서는 단어의 첫 음절 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.) 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 3과 같이 철자 전사를 통해 이를 반영하기 어려우나 원래의 표기를 먼저 적고, 학습자의 실제 발음을 ( )에 남겨 표기는 동일하나 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)
여자(요자)
회사(휘사)
과반수(화반수)

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>경음화되지 않음</note>
같이(가티)<note>구개음화되지 않음</note>
신라(신라)<note>자음동화되지 않음</note>
앞에(앞에)<note>연음되지 않음</note>
먹는(먹는)<note>자음동화되지 않음</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 학습자의 원어 발음을 최대한 원어에 가깝게 ( ) 안에 적어 밝힌다. 이때 한국어의 음운 체계로 전사가 불가능한 발음은 표기에 반영하지 않는다.

<예>	인터뷰: 영어식 발음으로 [인털뷰]에 가까운 소리가 남 센터: 영어식 발음으로 [세너]에 가까운 소리가 남 파트너: 영어식 발음으로 [팥너]에 가까운 소리가 남 → 각각 '인터뷰(인털뷰)', '센터(세너)', '파트너(팥너)'로 전사함

○ 외국인 학습자 발화의 경우 외래어 또는 외국어 발화 시 원어식의 발음을 하거나 한국어 모어화자의 현실 발음이 아닌 규범 발음을 하여 어색하게 들리는 경우가 있다. 이 경우 철자 전사를 통해 반영하기 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 [뻐스], 학습자 발음 [버스] → 각각 '카페(카페)', '버스(버스)'로 전사함.

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> 있여.

3) 쉼

○ (1초 이상의 쉼은) 0.1초 단위까지 표시한다(전사 도구의 시간 정보를 이용한다). 쉼은 발화와 발화 사이의 쉼이기 때문에 다음 발화의 시작 전에 표시한다. 만약 쉼이 누구의 것인지 불분명할 때는 한 줄에 표시한다.

<예>	1:이거 올라가면서 먹을까? 2:{1.2} 그럴까? {4.3} … 하게 먹는다.

○ 한 어절 안에서의 짧은 쉼은 '..'로 표시한다. 하나의 억양 단위 내부에서의 짧은 쉼은 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 읽기 인용

○ 책이나 자료 등을 보고 읽은 경우는 <R R>를 사용하여 표시한다.

<예>	1:그 담에 <R생각과 느낌이 유기적으로 잘 짜여져 조직체를 이룰 때 좋은 글이 될 수 있다R>라고 돼 있어요.

6) 익명성 보장을 위한 마크업

○ 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등은 노출되지 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> : 전화 번호

○ 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 뾰족한,
대안을 가지고 있는 것 같지는 않더라구요,
그게 어쩌면 <name1> 선배님이라든지 다른 선배님들 말::들은 걸 생각해 보면,
<name2> 사장이 <name2> 회장이 있으니까 보안이 있어도 눈치 보여서 얘기를 못한다,

한국어 학습자 말뭉치 형태 주석 지침

※ 본 지침은 21세기 세종 계획 현대문어 형태분석 말뭉치 구축 지침을 기본으로 한다.

Ⅰ. 학습자 말뭉치의 형태 분석 표지[21]

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
	목적격조사	JKO	JKO

21) [수정] 기존 세종 지침에 있었던 NF(명사추정범주), NV(용언추정범주)를 삭제하고 대부분 추정하여 해당 표지로 분석하거나 NA(분석불능범주)로 분석함.

	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
	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
	붙임표(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) **표준국어대사전**을 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/JX]

   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. **교정 어절**이 취할 표지를 준다.

■■ 문제를 쉬게 풀어요. [쉬/VA+게/EC]
■■ 너 때문내 죽겠어. [때문/NNB+내/JKB]
■■ 여러까지 문제가 생겼다. [여러/MM ‖ 까지/NNB]
■■ 강에 패수른 버렸다. [패수/NNG+른/JKO]

2) 형태와 뜻(어휘적 의미+기능적 의미)을 혼동한 경우
→ 보이는 대로 분석한다. **오류 어절**만 고려해서 분석한다.

■■ 내가 고기가 먹어요. [고기/NNG+가/JKS]
■■ 입학하자마자 교과서를 팔려고 서점에 갔어요. [팔/VV+려고/EC]

마. [학습자] 분석 기준의 적용

- 학습자 언어에서 나타나는 오류의 유형에 따라 다음과 같이 분석한다.

1) 오형태가 나타난 경우 최소 교정을 원칙으로 교정 어절을 상정해 형태 분석을 한다. 교정 어절을 상정하기 어려운 경우는 **분석불가능(NA)**으로 처리한다.

■■ 그러므로 부스르른 광고는 물가를 인상한다. [부스르른/NA]
■■ 그리고 경전철 타로 필어 50분쯤 경삭턱역 있습니다. [필어/NA]
■■ 이번 방학에 저는 친주와 같이 순열전 수열고 싶어요. [순열전/NA]
[순열/NA+고/EC]
■■ 그 꿈을 아구할 수 없을 것 같다. [아구하/NA+ㄹ/ETM]

→ 분석 불가능은 문맥에서 전혀 의미를 유추할 수 없어 오형태로도 보기 어려운 경우이다.

→ 표현 문형의 구성과 인접한 경우 교정어절을 상정하기 어렵다 해도 표현 문형 구성에 포함되는 형태까지는 분석을 한다.[22)]

2) 교정 어절의 상정이 가능한 경우 교정 어절이 취할 형태 표지를 부여한다.

가) 형태 분할이 가능한 것은 교정 어절을 기준으로 최대한 분할하는 것을 원칙으로 한다.

---

22) 표현 문형의 경우 <국립국어원2>의 표현 문형 목록을 기준으로 한다.(부록 참고)

■■ 아무도 물라다.(√몰랐다) [무르/VV+아/EP+다/EF]
■■ 저는 중국 사람은데(√사람인데) [사람/NNG+으/VCP+ㄴ데/EC]

나) 상정한 교정어절에 없는 형태가 추가된 경우의 분석은 다음과 같다.

(1) 체언과 조사 곡용의 경우는 체언 또는 조사의 오형태로 분석한다.

■■ 학곡을 갔다. [학곡/NNG+을/JKO]
■■ 학굔을 갔다. [학굔/NNG+을/JKO]
■■ 학교으를 갔다. [학교/NNG+으를/JKO]

(2) 용언과 어미 활용의 경우는 용언의 어간 혹은 어근과 어미를 확보해 분리한 후 잉여적 요소에 대해서는 NA로 처리한다.

■■ 내 꿈을 이뤄진고나 이루기 위해 [이뤄지/VV+ㄴ/NA+고나/EC]
■■ 하숙집인기 대문에 사람이 많았다. [하숙집+이/VCP+ㄴ/NA+기/ETN]
■■ 꼭 잘해야 된겠다. [되/VV+ㄴ/NA+겠/EP+다/EF]
■■ 힘이 강한하지 [강하/VA+ㄴ/NA+하/XSA+지/EC]
■■ 교실에 사람이 많단. [많/VA+다/EF+ㄴ/NA]

■■ 교통이 꼭 편립습니다. [편리/NNG+ㅂ/NA+습니다/EF]
■■ 나는 선생님이 됩겠다. [되/VV+ㅂ/NA+겠/EP+다/EF]

■■ 공부를 해고 [하/VV+아/NA+고/EC]
■■ 7시 30분까지 운동햅니다. [운동/NNG+하/XSV+아/NA+ㅂ니다/EF]
■■ 교실에 사람이 많아다. [많/VA+아/NA+다/EF]

■■ 꽃 가계 주인 돼면 [되/VV+어/NA+면/EC]
■■ 식사가 준비돼어 있었다. [준비/NNG+되/XSV+어/NA+어/EC]
■■ 할 수 있는 일이 많아져셨는데 [많아지/VV+어/NA+시/EP+었/EP+는데/EC]

→ 이때, 교정 어절을 기준으로 했을 때 잉여적인 요소가 추가된 오류와 기존의 다른 문법 요소와 혼동한 오류의 경우를 구분해서 주석해야 한다.

■■ 교실에 사람이 많은다. [많/VA+은/NA+다/EF]
■■ 교실에 사람이 많는다. [많/VA+는다/EF]

다) 불규칙 용언의 활용과 관련한 오류의 형태 분석은 다음과 같다.

(1) 불규칙 용언의 형태 분석에서, 학습자가 불규칙 활용의 오류로 어간을 잘 못 쓴 경우에는 어간을 복원하지 않고 오류가 발생한 어간의 형태를 그대로 살려 분석한다.

■■ 한국말이 너무 어려우다. [어려우/VA+다/EF]
■■ 친구들과 같이 즐거우게 칠 수 있으면 [즐거우/VA+게/EC]
■■ 다른 사람이 다시 저에게 도올 수 있다. [도오/VV+ㄹ/ETM]
■■ 정말 추은 경험이었다. [추으/VA+ㄴ/ETM]

→ 하지만 활용 오류가 용언의 불규칙 활용과 직접적인 관련이 없거나 어미의 오류로 보이는 경우는 기존의 형태 분석대로 어간을 살려 형태를 부여한다.

■■ 한국말이 너무 아렵다. [아렵/VA+다/EF]
■■ 저녁식사도 준비하기가 번거러워서 [번거/XR+럽/XSA+어서/EC]

■■ 공부가 힘드는 [힘들/VA+는/ETM]
■■ 풍선이 부푼는 [부풀/VV+ㄴ/NA+는/ETM]
■■ 공부가 힘든지만 [힘들/VA+ㄴ/NA+지만/EC]
■■ 공부가 힘든입니다. [힘들/VA+ㄴ/NA+이/VCP+ㅂ니다/EF]

(2) 또한 다음과 같이 어미가 누락한 것으로 보이는 경우는 오류가 발생한 어간 형태를 그대로 살려 분석한다.

■■ 친구들과 가벼우 장난을 하는 것은 [가벼우/VA]
■■ 경제력이 부족하거나 힘들 생활을 겪고 [힘들/VA]

라) 형태소 경계를 분할하기 어려운 오류의 형태 분석은 다음과 같다.

(1) 상정한 교정 어절의 형태소 음절에 따라 앞에서부터 형태를 분할한 후 형태 표지를 부여한다.

■■ 자시느이 마음대로 했다. [자시/NNG+느이/JKG]
■■ 내 유학생활을 아프로 미래에게 [아/NNG+프로/JKB]
■■ 다언에도 소개해 주게서요. [주/VX+게/EP+서요/EF]
■■ 어려슬 데 가을에 좋은 기역이 [어리/VA+어/EP+슬/ETM]
■■ 그것을 마가려고 하는 것들 중에 [마/VV+가려고/EC]

(2) 오류의 형태가 형태 단계에서도 표시될 수 있도록 가능한 경우 자모 단위로도 형태를 분할한다.

■■ 정말 신기하다고 생각했다. [생각/NNG+하/XSV+ㅆ/EP+다/EF]
■■ 어제 영화를 봤는데 [보/VV+ㅆ/EP+는데/EC]

→ 일부 오류의 경우는 분석한 형태의 결합이 원 어절의 형태가 되지 않더라도 기본적으로는 분할을 원칙으로 다음과 같이 처리한다.

■■ 쓰래기를 버러도 되면 좋겠습니다. [버러/VV+어도/EC]
■■ 많은 친구를 사궈서 재밌었어요. [사구/VV+어서/EC]
■■ 소치에 2시간 걸레요. [걸레/VV+어요/EF]
■■ 점점 심헤질 것이다. [심헤/VA+어/EC+지/VX+ㄹ/ETM]

마) 다음의 경우는 형태 분할을 하지 않고 오형태로 분석한다.

(1) 다음의 축약형에서 나타나는 오류 유형은 분할하지 않고 오형태로 분석해 형태 표지를 할당한다.

■■ 내일은 네 생일이라서 소포를 받았다. [네/NP]
■■ 재 장소 중에서 제일 좋아하는 곳은 [재/NP]
■■ 세 아버지는 키가 큽니다. [세/NP]

■■ 또 궁금한 개 있으면 [개/NNB]

■■ 저는 OO어학당 6급 학생이예요. [학생/NNG+이/VCP+예요/EF]
■■ 나는 학생이였다. [학생/NNG+이/VCP+였/EP+다/EF]

(2) 복합어의 구성 요소인 어근이나 접사가 누락된 경우 단어의 오형태로 보고 복합어 전체의 품사로 분석한다.

■■ 눈이 많(√많이) 왔다. [많/MAG]
■■ 줄임말은 젊은들이(√젊은이) 많이 사용하는 [젊은/NNG+들/XSN+이/JKS]

→ 하지만 용언의 어간이 어미와 결합한 활용형에서 어미가 누락된 경우는 용언의 어간으로 분석한다.

■■ 많(√많은) 음식이 있었어요. [많/VA]

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]

※ [보완]

■■ 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) [보완] '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][23)]

(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) [보완] 학습자의 특성상 접사를 명사적 기능으로 사용한 경우 분석하는 접사 목록에 없더라도 원래 품사대로 접사로 분석한다.

■■ 제주도에는 한국의 여명이 도예요. [도/XSN+이/VCP+에요/EF+./SF]

(3) [보완] 학생, 학교

- 대학교, 고등학교, 중학교, 대학생, 고등학생, 중학생은 모두 일반명사로 분석한다.

---

23) [수정] '21세기세종계획' 지침에는 '느냐/EC'로 되어 있지만 오류이므로 수정함.

■■ 대학교 [대학교/NNG]
■■ 고등학교 [고등학교/NNG]
■■ 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(孃), 옹(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/NNB[24], 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

24) [수정] 지침 전체적으로 띄어 써야 할 부분이 +기호로 연결되어 있어 + 기호로 수정함.

분석한다.

■■ 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, 인창동/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/NNP‖회담/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) [보완] 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

※ [보완]

■■ 남대문/NNP‖시장/NNG, 한강/NNP‖공원/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

■■ 쵸코하임/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) [보완] 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

※ [보완]

■■ EXID/SL, YG/SL+Family/SL

■■ B1A4/NNP, 2NE1/NNP

(7) [보완] 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/NNG
	사전 미등재	나누어 분석	(드라마) 서울/NNP+의/JKG ‖ 달/NNG (영화) 비밀/NNG+은/JX ‖ 없/VA+다/EF (TV프로그램)남자/NNG+의/JKG ‖ 자격/NNG

(8) 언어명

- 언어명의 경우 '-어'의 형태만을 통합하여 고유명사로 인정한다.

■■ 한국어/NNP, 일본어/NNP, 영어/NNP, 알타이어/NNP, 네덜란드어/NNP

■■ 한국말/NNG, 러시아/NNP‖말/NNG, 일본/NNP‖말/NNG

■■ 한글/NNG, 알파벳/NNG, 한자/NNG

(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, 카카오버스/NNP

(10) [보완] 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 나뉠 수 있으나, 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+ㄴ다/EF]

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, 걔, 쟤, 얘, 이것, 저것, 그것, 이거, 저거, 그거, 여기, 저기, 거기, 이곳, 그곳, 저곳, 어디, 모(某), 모모(某某)

※ [보완] '자기'는 대명사로 분석한다.

※ [보완] '자신', '아무것'은 일반명사로 분석한다.

※ [보완] '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.

■■ 모모가 말했다 [모모/NP+가/JKS]
■■ 모모 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 저거를 먹을래. [저거/NP+를/JKO]
■■ 나는 여태 그걸 믿어 왔단다. [그것/NP+ㄹ/JKO]

(나) 다음과 같이 원형을 밝힐 수 있는 대명사는 원형대로 분석한다.

■■ 내 이제부터는 내 명령을 따라라. [나/NP+의/JKG]
■■ 내게 내게 전자우편으로 알려 다오. [나/NP+에게/JKB]
■■ 네게 어제 네게 보낸 선물이 잘못되었다. [너/NP+에게/JKB]
■■ 제게 문제가 있다면 제게 말씀해 주세요. [저/NP+에게/JKB]
■■ 누가 누가 전화를 하는 지 보고해라. [누구/NP+가/JKS]
■■ 뉘 뇌 집 애기가 울고 있는 거야? [누구/NP+의/JKG]
■■ 뭐가 도대체 뭐가 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 얘기냐? [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]

※ [보완] 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]

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, 백한둘, 두서넛, 서넛, 너덧, 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, ...

※ [보완] '째'는 분석하는 접미사에 해당하지만 서수사에서 쓰인 경우 분석하지 않는다.

■■ 첫째 [첫째/NR]
■■ 첫 번째 [첫/MM∥번/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) [보완] 때로 수사와 수관형사의 구별이 애매한 경우가 있다. 이 분석에서는 임홍빈(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.

※ [보완] '있다'는 모두 **동사**로 처리한다. (세종 말뭉치 기준)

※ [보완] '감사하다'는 모두 **동사**로 보고 '-하-'는 모두 동사파생접미사로 처리한다. (세종 말뭉치 기준)

※ [보완] '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.	아 니 하 / V X + 았 / E P + 다 /EF+./SF
■■ 않다	시간이 지나도 기차는 오지 않았다.	않/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/체하다/척하다/듯하다/법하다/뻔하다 [양/NNB+하/XSA+다/EF]
■■ 듯싶다 [듯싶/VX+다/EF]

※ 표준국어대사전에 따라, 기존에 접미사로 분석하던 '만하'의 지침을 변경

하여, '만'을 보조사로, '하'를 동사로 분석한다. '만하'는 '만/NNB+하/XSA'로 분석되는 경우도 있으므로 주의해야 한다.[25]

■■ 철수만 한 인재가 없다 [철수/NNP+만/JX ‖ 하/VV+ㄴ/ETM]
■■ 이 음식은 먹을 만하다. [만/NNB+하/XSA+다/EF+./SF]

② '버릇하다'의 경우에는 선행 성분으로 관형형이 오는 것은 아니지만, 일반 명사 '버릇'과 크게 구별되지 않으므로 '버릇'은 명사로 분석한다.

■■ 자꾸 울어 버릇하다. [버릇/NNG+하/XSV+다/EF+./SF]

※ [보완] '-도록 하다'는 형용사 일부 어간에만 사용되는 등 '-게 하다'와 분포가 다르므로 이때의 '하다'는 본용언으로 분석한다.

■■ 열심히 공부하도록 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.

■■ 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]

---

25) '바. 3) 다) 형용사파생접미사'의 주의사항의 내용 이동함.

③ '~였다'

■■ 그 당시 나는 아이였다. [아이/NNG+이/VCP+었/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 친절하나?"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 걔를 좋아하는 건 걔가 착해서야. [착하/VA+아서/EC+이/VCP+야/EF+./SF]

<주의사항>

(가) [보완] 학습자가 지정사 '이'를 몰라서 누락한 경우는 '이/VCP'를 복원하지 않는다.

■■ 방법은 한 가지에요. [가지/NNB+에요/EF]
■■ 이것은 책상라며 나를 가르쳤다. [책상/NNG+라며/EC]

(나) [보완] 학습자가 '에요'를 써야하는 부분에서 '예요'로 쓴 경우는 '이/VCP+에요/EF'로 분석하지 않고 종결어미의 오형태로 분석한다.

■■ 저는 OO어학당 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 명이서 그 일을 꾸몄다.	[단/MM]
■■ 단, 그 일은 해서는 안 된다.	[단/MAJ]

③ 관형사, 명사, 부사 통용

■■ 이내 마음을 어찌 알리요.	[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+ㄴ/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]

※ [보완] 접속부사는 《표준국어대사전》에 접속부사로 뜻풀이된 것만 인정한다. 아래는 《표준국어대사전》의 접속부사 목록이다.

건데, 고로01「2」, 그래서, 그러나, 그러니까, 그러면, 그러므로, 그런데, 그럼01「1」, 그렇지마는, 그렇지만, 그리고, 그리하여, 근데01, 단06, 따라서, 연이나, 연중에, 연즉, 이리하여, 하건만, 하기는, 하기야, 하긴, 하물며, 하지만, 한데03

※ 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, 이를테면, 한편, 혹시, 혹은

※ [보완] 사전에 등재되지 않은 부사의 약어는 본딧말과 같은 표지로 분석한다.

■■ 그니까(그러니까), 글고(그리고)/MAJ
■■ 왜냐면(왜냐하면)/MAG

나) 일반부사(MAG)

<주의사항>

① 일반부사는 종종 일반명사와 동일형태를 띠고 있어 구분이 어려운 경우가 있다. 이들은 뒤에 조사가 결합하느냐의 여부와, 문맥에서 후행 명사를 수식하느냐의 여부에 따라 부사와 명사로 분석될 수 있다.

■■ 너의 진짜 속셈이 무엇인지 말해 봐라. [진짜/NNG]
■■ 그 수학 문제는 진짜 어려웠다. [진짜/MAG]
■■ 지금이 공부하기 딱 좋은 때이다. [지금/NNG+이/JKS]
■■ 나는 지금 막 집에 도착했다. [지금/MAG]
■■ 오늘은 내 생일이 아니다. [오늘/NNG+은/JX]
■■ 오늘은 그가 왔다. [오늘/MAG+은/JX]

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 '덩달아, 더불어'는 모두 동사로 옳게 분석해야 함에 주의한다.

■■ 너는 덩달아 왜 난리니? [덩달/VV+아/EC]
■■ 우리 함께 더불어 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, 끊임없이, 난데없이, 남김없이 등

라. 독립언

1) 감탄사(IC)

- 감탄사는 화자의 부름이나 느낌, 놀람이나 대답을 직접적으로 나타내는 품사를 말한다.

■■ 그럼(요), 야호, 어머, 앗, 아, 예, 그래(요), 아니(요), 글쎄, 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.

■■ 뭔지도 모른 채 [뭐/NP+이/VCP+ㄴ지/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. 관계언[26]

- 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]
■■ 그녀와의 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)

26) 지침에 제시된 조사 목록에서 빠진 이형태와 예시를 추가함.

- 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]

※ [보완] 다음과 같이 체언 뒤에서 '이'가 첨가되어 나타나는 오류의 경우, 이때 '이'는 모두 주격 조사로 분석한다.

■■ 닭이가 울었다. [닭/NNG+이/JKS+가/JKS]
■■ 책상이가 있다. [책상/NNG+이/JKS+가/JKS]
■■ 친구들이 6월에 일이를 찾아있어다. [일/NNG+이/JKS+를/JKO]
■■ 좋아하는 거는 옷이입니다. [옷/NNG+이/JKS+이/VCP+ㅂ니다/EF]
■■ 어른이들이 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]
	장관으로서 책임을 다해야 한다.	[장관/NNG+으로서/JKB]
■■ 로써/으로써	돌로써 지붕을 만든다고?	[돌/NNG+로써/JKB]
	콩으로써 메주를 쑨다고 해도	[콩/NNG+으로써/JKB]
■■ 같이	바보같이 웃고 다닌다.	[바보/NNG+같이/JKB]
■■ 더러	나더러 이것도 하라고 한다.	[나/NP+더러/JKB]
■■ 랑/이랑	너랑 많이 닮았다.	[너/NP+랑/JKB]
	오늘 동생이랑 싸웠다.	[동생/NNG+이랑/JKB]
■■ 로부터/	TV로부터 받는 영향력이	[TV/SL+로부터/JKB]
으로부터	시험으로부터 해방되다	[시험/NNG+으로부터/JKB]
■■ 마냥	기영이마냥 놀 수만은 없다.	[기영이/NNP+마냥/JKB]
■■ 마따나	네 말마따나 나도 그래야 한다.	[말/NNG+마따나/JKB]
■■ 만큼	눈물만큼 콧물도 흐른다니까.	[눈물/NNG+만큼/JKB]
■■ 보고	영자보고 놀자고 좀 해라.	[영자/NNP+보고/JKB]
■■ 보다	직관보다는 논리가 동원돼야 한다.	[직관/NNG+보다/JKB+는/JX]
■■ 에	나는 너에 대해 아무것도 모른다.	[너/NP+에/JKB]
■■ 에게	너에게 말하기 싫다.	[너/NP+에게/JKB]
■■ 에게서	나는 철수에게서 그 말을 들었다.	[철수/NNP+에게서/JKB]
■■ 에서	집에서 학교까지 너무 멀다.	[집/NNG+에서/JKB]
■■ 에서부터	연구소에서부터 가게까지는	[연구소/NNG+에서부터/JKB]
■■ 와/과	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.	[심각/XR+하/XSA+다/EF+"/SS+라고/JKQ]
	팻말에는 "금지구역"이라고 쓰여 있었다.	["/SS+금지/NNG+구역/NNG+"/SS+이라고/JKQ]
■■ 하고	영수는 "이제 가자"하고 말문을 닫았다.	[가/VV+자/EF+"/SS+하고/JKQ]

<주의사항>[27]

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]

(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]

② [보완] 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 일이다고 말했다. [일/NNG+이/VCP+다고/EC]

③ 인용 기호 중 하나인 <“ ”>은 맥락에 따라 인용이 아닌 강조를 위해 사

27) <세종> 분석 결과를 바탕으로 다시 정리함.

용되기도 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]

※ [참고]

■■ 시골 아이라고 그것도 모르겠니? [아이/NNG+라고/JX]

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+이고/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	잔칫상에는 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]
■■ 나는 철수랑 영희랑 같이 놀았다. [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 [보완] 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 [개성/NNG+적/XSN+이/VCP+고/EC]
색다른 현대 감각을 보여준다며,

■■ 옷이며 신이며 흩어져 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.

■■ 그려/그래	좋습니다그려.	[좋/VA+습니다/EF+그려/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 걷겠다	[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.	[코알라/NNG+란/JX]
	사람이란 분수를 지킬 줄 알아야 한다.	[사람/NNG+이란/JX]
■■ ㄹ랑/일랑	강엘랑 가지 마라.	[강/NNG+에/JKB+ㄹ랑/JX]
	그 일에 대해선 걱정일랑 하지 말아라.	[걱정/NNG+일랑/JX]
■■ 마다	꽃마다 독특한 향기가 있다.	[꽃/NNG+마다/JX]
■■ 마저	장미마저 시들고 말았다.	[장미/NNG+마저/JX]
■■ 만	사람은 빵만으로 살 수 없다.	[빵/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]
■■ 서껀	국물이나 동치미서껀 아무 거나	[동치미/NNG+서껀/JX]
■■ 사/이사	내사 그걸 이미 했지.	[내/NP+사/JX]
	남이사 무슨 상관이야.	[남/NNG+이사/JX]
■■ 야/이야	그야 그렇지.	[그/NP+야/JX]
	그가 인간성이야 그만이지.	[인간성/NNG+이야/JX]
■■ 야말로/이야말로	사과야말로 가을의 과일이다.	[사과/NNG+야말로/JX]
	통일이야말로 최대의 과업이지.	[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.[28]

■■ 까지, 깨나, 는(은/ㄴ), 대로, 도, 따라, 마다, 마저, 만, 밖에, 부터, 뿐, 조차, 치고, ㄴ커녕

※ [참고] '만', '뿐'은 의존 명사로도 분석될 수 있음.

---

28) [삭제] 말고,
→ '말고'는 '표준국어대사전'에 보조사로 등재되어 있지 않으며, 세종 말뭉치에서도 보조사로 분석하지 않았으므로 목록에서 삭제함.

(마) [보완] 종결어미 뒤에 나타나는 '든지, 든가, 거나' 등의 경우는 보조사로 분석한다.

■■ 공부를 잘한다든지 운동을 잘한다든지 [잘/MAG+하/XSV+ㄴ다/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]
■■ (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, 빵이오.	[밥/NNG+이/VCP+요/EC]

(나) '종결어미+요(보조사)'는 종결어미로 통합하여 분석한다.

■■ 말씀대로 했는걸요. [하/VV+았/EP+는걸요/EF+./SF]

(다) '비종결어미+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) A: 선생님에 대해 알고 있니?
B: 선생님이요? ('오'의 오류) [선생/NNG+님/XSN+이/VCP+요/EF]

(마) '말고'는 용언 '말다'의 활용형으로 처리한다.

■■ 돈말고 지혜가 필요하다. [돈/NNG‖말/VV+고/EC]

바. 의존형태

1) 어미[29]

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/NNG+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 먹었었던 음식에 문제가 있다.	[먹/VV+었었/EP+던/ETM]

29) 지침에 제시된 어미 목록에서 빠진 이형태와 예시를 추가함.

<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]
■■ -(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.

■■ -랠까-, -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]
■■ -구먼/는구먼	학교가 참 크구먼.	[크/VA+구먼/EF+./SF]
	공부를 잘하는구먼.	[잘/MAG+하/XSV+는구먼/EF+./SF]
■■ -ㄴ가/은가/는가	이것이 무엇인가?	[무엇/NP+이/VCP+ㄴ가/EF+?/SF]
	그것이 좋은가?	[좋/VA+은가/EF+?/SF]
	그가 집에 있는가?	[있/VV+는가/EF+?/SF]
■■ -ㄴ걸/은걸/는걸	이제 시작인걸.	[시작/NNG+이/VCP+ㄴ걸/EF+./SF]

	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]
	그가 누구냐고?	[누구/NP+이/VCP+냐고/EF+?/SF]
■■ -냐고/으냐고 /느냐고	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]
	그게 사실이다.	[사실/NNG+이/VCP+다/EF+./SF]
■■ -다/ㄴ다/는다	이건 말도 안 된다.	[되/VV+ㄴ다/EF+./SF]
	아이가 글을 잘 읽는다.	[읽/VV+는다/EF+./SF]
	돈이 많다구?	[많/VA+다구/EF+?/SF]
■■ -다구/ㄴ다구 /는다구	너도 간다구?	[가/VV+ㄴ다구/EF+?/SF]
	소설책을 읽는다구?	[읽/VV+는다구/EF+?/SF]
	그도 가겠다나.	[가/VV+겠/EP+다나/EF+./SF]
■■ -다나/ㄴ다나 /는다나	나를 잘 안다나.	[알/VV+ㄴ다나/EF+./SF]
	건강한 여자를 찾는다나.	[찾/VV+는다나/EF+./SF]
	일을 망쳤다네	[망치/VV+었/EP+다네/EF+./SF]
■■ -다네/ㄴ다네 /는다네	우리네 짧은 인생도 간다네.	[가/VV+ㄴ다네/EF+./SF]
	평소에도 한복을 잘 입는다네.	[입/VV+는다네/EF+./SF]
	돈이 없다니까!	[없/VA+다니까/EF+!/SF]
■■ -다니까/ㄴ다니까 /는다니까	어머니가 오늘은 꼭 오신다니까.	[오/VV+시/EP+ㄴ다니까/EF+!/SF]
	내 말을 믿지를 않는다니까.	[않/VX+는다니까/EF+./SF]
	서울이 이렇게 번화하다니.	[번화/XR+하/XSV+다니/EF+?/SF]
■■ -다니/ㄴ다니 /는다니	이 긴 시를 어떻게 외운다니?	[외우/VV+ㄴ다니/EF+?/SF]
	이 많은 책을 언제 읽는다니?	[읽/VV+는다니/EF+?/SF]
	술은 싫다면서?	[싫/VA+다면서/EF+?/SF]
■■ -다면서/ㄴ다면서 /는다면서	니가 축구를 잘한다면서?	[잘/MAG+하/XSV+ㄴ다면서/EF+?/SF]
	달팽이도 먹는다면서?	[먹/VV+는다면서/EF+?/SF]
	그가 가지고 있다오.	[있/VX+다오/EF+./SF]
■■ -다오/ㄴ다오 /는다오	꽃은 이른 봄에 핀다오.	[피/VV+ㄴ다오/EF+./SF]
	이 나무는 열매를 많이 맺는다오.	[맺/VV+는다오/EF+./SF]
	나도 슬프단다.	[슬프/VA+단다/EF+./SF]
■■ -단다/ㄴ단다 /는단다	선생님께서 공부를 가르쳐 주신단다.	[주/VX+시/EP+ㄴ단다/EF+./SF]
	누에는 뽕잎을 먹는단다.	[먹/VV+는단다/EF+./SF]
	꽃이 아름답도다.	[아름답/VA+도다/EF+./SF]
■■ -도다/는도다	짐이 조서를 내리는도다.	[내리/VV+는도다/EF+./SF]

■■ -ㄹ걸/을걸	모른다고 할걸.	[하/VV+ㄹ걸/EF+./SF]
	생각만큼 쉽지 않을걸.	[않/VX+을걸/EF+./SF]
■■ -ㄹ게/을게	그렇게 할게.	[하/VV+ㄹ게/EF+./SF]
	남은 밥은 내가 먹을게.	[먹/VV+을게/EF+./SF]
■■ -ㄹ까/을까	이제 밥을 할까?	[하/VV+ㄹ까/EF+?/SF]
	이 과자는 내가 먹을까?	[먹/VV+을까/EF+?/SF]
■■ -렴/으렴	맘대로 해 보렴.	[보/VX+렴/EF+./SF]
	이것 좀 먹으렴.	[먹/VV+으렴/EF+./SF]
■■ -려무나/으려무나	더 놀다 가려무나.	[가/VV+려무나/EF+./SF]
	책이나 읽으려무나.	[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 옴세.	[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 예쁩디까?	[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]
■■ -아/어/여	함께 가.	[가/VV+아/EF+./SF]
	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]
■■ -오/으오/소	물이 깨끗하오.	[깨끗/XR+하/XSA+오/EF+./SF]
	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 했는걸요. [하/VV+았/EP+는걸요/EF+./SF]
■■ 뭐 먹었는데요? [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 경우도 <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 어

미로서 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 굶게 하지 마라.	[굶/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.	[예쁘/VA+ㄴ데/EC]
	방이 좁은데 가구는 많다.	[좁/VA+은데/EC]
	눈이 오는데 차를 가져가지 말까?	[오/VV+는데/EC]
■■ -ㄴ들/는들	간다 한들 아주 갈까?	[하/VV+ㄴ들/EC]
	그걸 먹는들 뭐가 달라지겠나.	[먹/VV+는들/EC]
■■ -ㄴ즉/은즉	배가 고픈즉 속이 쓰리다.	[고프/VA+ㄴ즉/EC]
	물이 맑은즉 고기가 많기는 어렵소.	[맑/VA+은즉/EC]
■■ -ㄴ지라/은지라/는지라	눈이 온지라 길이 미끄럽다.	[오/VV+ㄴ지라/EC]
	기분이 좋은지라 다정하다.	[좋/VA+은지라/EC]
	선생님께서 고집을 굽히지 않으시는지라	[않/VX+으시/EP+는지라/EC]
■■ -나/으나	눈이 오나 비가 오나	[오/VV+나/EC]
	밥을 먹으나 마나이다.	[먹/VV+으나/EC]
■■ -나니	멀리 보이나니 넓은 들이로다.	[보이/VV+나니/EC]
■■ -나마/으나마	도와주지는 못하나마 방해를 해서는	[못하/VX+나마/EC]
	맛은 없으나마 많이 드세요.	[없/VA+으나마/EC]
■■ -노니	묻노니, 포부가 무엇이냐?	[묻/VV+노니/EC+,/SP]
■■ -니/으니	밥을 다 먹고 보니 배가 불렀다.	[보/VX+니/EC]
	이 옷은 작으니 큰 것으로 바꿔 주세요.	[작/VA+으니/EC]
■■ -느니	앉아서 걱정하느니 나가서 하겠다.	[걱정/NNG+하/XSV+느니/EC]
■■ -니까/으니까	웃기니까 좋다.	[웃기/VV+니까/EC]
	약속을 했으니까 만나야 한다.	[하/VV+았/EP+으니까/EC]
■■ -다가	자랑하다가 망신당했다.	[자랑/NNG+하/XSV+다가/EC]
■■ -다기에/ㄴ다기에/는다기에	그녀가 예쁘다기에 보러 왔소.	[예쁘/VA+다기에/EC]
	앞으로 잘 한다기에 승낙했다.	[하/VV+ㄴ다기에/EC]

	빵을 먹는다기에 주었다.	[먹/VV+는다기에/EC]
■■ -다손/ㄴ다손 /는다손	밉다손 치더라도 구박하지 말자.	[밉/VA+다손/EC]
	그가 제시간에 온다손 하더라도	[오/VV+ㄴ다손/EC]
	내 앞의 음식은 다 먹는다손 치더라도	[먹/VV+는다손/EC]
■■ -대도/ㄴ대도 /는대도	시간이 있대도 만나 주질 않는다.	[있/VV+대도/EC]
	늦으면 큰일 난대도 서두르질 않아요.	[나/VV+ㄴ대도/EC]
	떠들면 야단맞는대도 계속 떠들었다.	[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/	비가 올뿐더러 바람도 분다.	[오/VV+ㄹ뿐더러/EC]
을뿐더러	그는 재산이 많을뿐더러 재능도 많다	[많/VA+을뿐더러/EC]
■■ -ㄹ수록/	갈수록 태산이다.	[가/VV+ㄹ수록/EC]
을수록	이 책은 읽을수록 감동을 준다.	[읽/VV+을수록/EC]
	비가 얼마나 올지 천둥이 다 친다.	[오/VV+ㄹ지/EC]
■■ -ㄹ지/을지	내일은 얼마나 날씨가 좋을지 오늘 밤하늘에 별이 유난히 빛난다.	[좋/VA+을지/EC]
■■ -ㄹ지라도/	이길지라도 명예롭지는 않다.	[이기/VV+ㄹ지라도/EC]
을지라도	마음에 걱정이 있을지라도 내색하지 마라.	[있/VV+을지라도/EC]
■■ -ㄹ지언정/	그것은 무모한 행동일지언정	[행동/NNG+이/VCP+ㄹ지언정/EC]
을지언정	죽을지언정 그 일은 못하겠다.	[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.	[청소/NNG+하/XSV+러/EC]
	점심 먹으러 집에 간다.	[먹/VV+으러/EC]
■■ -려/으려	학교에 가려 한다.	[가/VV+려/EC]
	웃으려 한다.	[웃/VV+으려/EC]
■■ -려니와/	비용도 문제려니와 일꾼도 문제다.	[문제/NNG+이/VCP+려니와/EC]
으려니와	이 마을은 경치도 좋으려니와	[좋/VA+으려니와/EC]
■■ -련마는/	보면 반가우련마는 볼 수가 없네.	[반갑/VA+련마는/EC]
으련마는	벌써 제 잘못을 알았으련마는	[알/VV+았/EP+으련마는/EC]
■■ -며/으며	노래하며 춤을 춘다.	[노래/NNG+하/XSV+며/EC]
	강물이 맑으며 깊다.	[맑/VA+으며/EC]
■■ -면/으면	지옥이 존재하면 만원일 것이다.	[존재/NNG+하/XSV+면/EC]
	내일 날씨가 좋으면 소풍을 가겠다.	[좋/VA+으면/EC]
■■ -면서/으면서	푸르면서 검은 물빛	[푸르/VA+면서/EC]

	밥을 먹으면서 신문을 본다.	[먹/VV+으면서/EC]
■■ -므로/으므로	비가 오므로 가지 않겠다.	[오/VV+므로/EC]
	강이 깊으므로 배 없이 건널 수 없다.	[깊/VA+으므로/EC]
■■ -아/어	입을 막아 버렸다.	[막/VV+아/EC]
	밥을 먹어 버렸다.	[먹/VV+어/EC]
■■ -아도/어도	암만 봐도 모르겠다.	[보/VV+아도/EC]
	나는 부자가 아니어도 행복하다.	[아니/VCN+어도/EC]
■■ -아서/어서	덫을 놓아서 꿩을 잡았다.	[놓/VV+아서/EC]
	그는 걸어서 학교에 갔다.	[걷/VV+어서/EC]
■■ -아야/어야	이 일은 잘해야 한다.	[잘/MAG+하/XSV+아야/EC]
	사람은 먹어야 산다.	[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 <u>왔는데</u> 기쁘지도 않니? [오/VV+았/EP+는데/EC]
■■ 내가 지금 <u>있는 데가</u> 어디지? [있/VV+는/ETM‖데/NNB+가/JKS]
■■ 다들 <u>만족하는지</u> 아무런 불평이 없다. [만족/NNG+하/XSV+는지/EC]
■■ 너를 <u>만난 지도</u> 꽤 오래구나. [만나/VV+ㄴ/ETM‖지/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 <u>믿음직하게</u> 행동해라. [믿/VV+음직/EC+하/VX+게/EC]
■■ 그것 참 <u>먹음직스럽다.</u> [먹음직/XR+스럽/XSA+다/EF+./SF]
■■ 그것은 매우 <u>바람직한</u> 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 명사형이냐 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) 접두사(XP)

- 접두사는 명사와 수사에 결합하는 접사류를 묶어서 체언접두사만을 설정하기로 한다.

가) 체언 접두사(XPN)

- 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 분석을 하기로 한다.

가(假)-가건물, 고(高)-고물가, 과(過)-과보호, 구(舊)-구소련, 날-날음식, 노(老)-노부부, 대(大)-대선배, 맏-맏아들, 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, 댁(宅)-청주댁, 론(論)-비평론, 별(別)-가구별, 여(餘)-삼십여, 류(類)-자연류, 률, 율(率)-경쟁률, 리(裡)-비밀리, 분(分) 분량-일인분, 분(分)-3분의, 산(産)-중국산, 상(上)-역사상, 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(値)-보름치, 풍(風)-복고풍, 화(化)-도구화, 기-기름기, 께-10분께, 꼴-십 원꼴, 끼리-전우끼리, 꾼-노름꾼, 네-동이네, 님-선생님, 들-우리들, 들이-1ℓ들이, 배기-열 살배기, 뻘-조카뻘, 씩-만원씩, 장이-간판장이, 쟁이-심술쟁이, 쯤 -내일쯤, 질-서방질, 짜리-백 원짜리, 째1 -이틀째, 째2-옹기째, 치레-인사치레, 투성이-먼지투성이

<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.

※ [보완] 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) [보완] '-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 시험 공부 하는 것을 [공부/NNG || 하/VV+는/ETM]
■■ 카페에서 시험 공부하는 것을 [공부/NNG || 하/VV+는/ETM]

(나) [보완] 학습자가 잘못 접미사를 사용한 경우 교정어절을 상정했을 때 교정어절의 품사가 동사일 때는 동사파생접미사, 교정어절의 품사가 형용사일 때는 형용사파생접미사로 분석한다.

■■ 음식을 먹하다. [먹/VV+하/XSV+다/EF+./SF]
■■ 마음이 아프한 아주머니가 집에 돌아왔다. [아프/VA+하/XSA+ㄴ/ETM]
■■ 그렇지 않한다면 [않/VX+하/XSA+ㄴ다면/EC]
■■ 한국어 공부를 열심 하다. [열심/NNG+하/XSV+다/EF]

(다) [보완] 'NA+하다'의 접사 '-하'는 XSV로 분석한다.

다) 형용사파생접미사(XSA) → '명사/부사/어근+형용사파생접미사'로 분석한다.

- 형용사파생접미사는 어기나 어근에 붙어서 그것을 형용사로 파생시키는 접미사이다.

※ [보완] 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.

■■ 따뜻도 하다 [따뜻/XR+도/JX] [하/VV+다/EF]
■■ 이러하다 [이러/XR+하/XSA+다/EF]

<주의사항>(예를 보완함.)

- 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.

■■ 듬직하다 [듬직/XR+하/XSV+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

※ [보완] "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+ㄴ다/EF]
■■ 음식 맛이 저번보다 못하다. [못하/VA+다/EF]

■■ 밥을 먹지 못한다. [못하/VX+ㄴ다/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 본딧말과 대등하게 사용되고 분석결과가 동일한 어절 단위를 형성할 경우에 한해서만 복원한다. 그러나 다음에서처럼, 본딧말로 복원할 경우 어절 수에 변화가 생길 뿐 아니라 본딧말로 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)

※ [보완] 그 자체가 사전에 등재되어 있지도 않으면서, 축약의 정도가 심하거나 분석하기 어려운 방언형의 경우 분석불능범주로 처리한다.

■■ 담배가 쪼매턴게 하마 자라서 빼나?	[쪼매턴게/NA]
■■ 친구한테 전화를 적긴 일이었다.	[적긴/NA]
■■ "부산국제영화"제가니와	[제가니와/NA]
■■ 있잖아요	[있/VV+잖/NA+아요/EF]
■■ ㅋㅋ	[ㅋㅋ/NA]
■■ ㅠㅠ	[ㅠㅠ/NA]
■■ ㅇㅋㅇㅋ	[ㅇㅋㅇㅋ/NA]
■■ ^^	[^^/NA]

4) 합성어

- 합성어는 표준국어대사전에 등재되어 있는 것만을 인정한다.

		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

[보완] <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 잇는 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) [보완] 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 //냇가 //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//며느릿감//사윗감//장군감.
감03「05」	(일부 명사 뒤에 붙어) 대상이 되는 도구, 사물, 사람, 재료의 뜻을 나타내는 말.	구경감 //놀림감 //땔감 //양념감 //안줏감 //장난감//웃음감//사형감//노벨상감//바느질감.
값「07」	(일부 명사 뒤에 붙어) '가격', '대금', '비용'의 뜻을 나타내는 말.	기름값 //물값 //물건값 //부식값 //신문값 //우윳값 //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//사기극//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', '흉내'의 뜻을 나타내는 말.	시장놀이//병원놀이//엄마놀이//학교놀이.
대15「03」	(일부 명사 뒤에 붙어) 받침이 되는 시설이나 이용물의 뜻을 나타내는 말.	급수대 //조회대 //독서대.
댁01「04」	(일부 명사 뒤에 붙어) '택호'를 나타내는 말.	윤 판서댁
덩어리「03」	(일부 명사 뒤에 붙어) [같은 말] 덩이(3. 그러한 성질을 가지거나 그런 일을 일으키는 사람이나 사물을 나타내는 말).	골칫덩어리 //심술덩어리//애꿋덩어리//채주덩어리.
덩이「03」	(일부 명사 뒤에 붙어) 그러한 성질을 가지거나 그런 일을 일으키는 사람이나 사물을 나타내는 말. [비슷한 말] 덩어리.	골칫덩이 //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//북어무침//골뱅이무침//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 //숭고미 //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.
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.	옥재떨이 //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 //은찻잔.
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) 조린 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 //꾀주머니 //근심주머니
즙「02」	(먹을 것을 나타내는 일부 명사 뒤에 붙어) '농축액'을 나타내는 말.	미나리즙 //석류즙 //배즙 //양파즙 //칡즙.
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) 찐 음식의 뜻을 나타내는 말.	갈비찜 //아귀찜.
책01「04」	(일부 명사 뒤에 붙어) '서적2'임을 나타내는 말.	국어책//소설책//요리책.
터01「04」	(일부 명사 뒤에 붙어) '자리1'나 '장소5'의 뜻을 나타내는 말.	낚시터 //놀이터 //일터 //흉터.
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 //촌티.
표05「07」	(일부 명사 뒤에 붙어) '그 사람이 만든	엄마표 //아빠표 //신랑표 //주부표.

	물건'의 뜻을 더하는 말.	
후08「03」	(일부 명사 앞에 붙어) '뒤나 다음'의 뜻을 나타내는 말.	후더침 //후보름 //후서방.

아. [보완] 구어 형태 분석 말뭉치 (추가)

※ 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 '어요'를 대부분 연결어미로 분석하는데, 이 경우 종결어미로 분석해야 한다.

■■ 그냥 매일매일 쉬고. [쉬/VV+고/EC]
북경하고 고향에 갔다 왔다 갔다 왔다. [오/VV+았/EP+다/EC]
했어요. [하/VV+았/EP+어요/EF]
■■ 어~ 그렇게 어 가고 싶지 않았어요 [않/VX+았/EP+어요/EF]
돈이 없어서. [없/VA+어서/EC]
■■ 아까 드렸던 종이를 한번 살펴보고요, [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 주의하여야 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 말뭉치

의 특성을 드러내는 경우 아래와 같이 분석한다.

1) 완전한 어절

- 기본적으로 발화가 완전히 이뤄진 어절은 문어 형태 분석 지침을 따라 분석한다.

■■ 저희가 하여 하고 [하/VV+아/EC]
■■ 충고를 해 준 줘 줄 [주/VX+ㄴ/ETM ‖ 주/VX+어/EC ‖ 주/VX+ㄹ/ETM]

2) 끊어진 어절

- 끊어진 어절은 어절의 일부만 발화된 경우나 불분명한 경우인데, 앞/뒤에 완성된 발화가 나타나는 경우 끊어진 어절은 분석불가능(NA)로 처리한다.[30]

가) 어절의 일부만 발화되어 분석하지 않는 경우

■■ 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]
■■ 한국어 공부가 힘= 힘들었지만 [힘/NA]
■■ 의사 선생님도 곧 나= 나을 수 있다고 [나/NA]
■■ 경복궁에 가 봐= 봤다. [봐/NA]
■■ 그랬= 그랬어요. [그랬/NA]
■■ 스페인에서 있= 있어= 있어요. [있/NA ‖ 있어/NA]
■■ 의사가= 가 [가/NA]
■■ 슬퍼= 퍼 가지구 [퍼/NA]

나) 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]

다) 어절 중간에 간투사 따위가 들어가는 경우

30) 세종 구어에서는 끊어진 어절과 불분명한 어절을 UNT, UNC 등으로 세분하였지만, 학습자 자료의 특성과 작업자들의 혼동을 고려하여 통합하여 NA로 처리한다.

■■ 심약 어 하다 [심약/NNG ‖ 어/IC ‖ 하/XSV+다/EF]
■■ 좋아 어 하다 [좋/VA+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.

■■ 캔 유 두 미어 페이버? [페이버/NNG]
가 무슨 뜻? [가/JKS]
■■ 좋아. [좋/VA+아/EF]
라고 대답했지. [라고/JKQ]
■■ 주부 우울증. [우울증/NNG]
이라고 말할 수 있겠습니다. [이/VCP+라고/EC]
■■ 공부. [공부/NNG]
한다고 [하/XSV+ㄴ다고/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, [선두/NNG ‖ 주/NA]
자가 도착했다. [자/NA+가/JKS]

4) 불분명한 어절

- 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 줘 봐. [그/MM ㅣ 책/NNG]
■■ 그 무엇인가를 알아내고자 했지만 [그/MMㅣ무엇/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 ㅣ 둘/NR ㅣ 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 대용으로 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/VA+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 따

른다.

■■ 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, 그죠 [그쵸/IC], [그죠/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.

자. [보완] 구어 형태 분석 말뭉치 (시스템 관련)

1) 교정 어절이 있는 경우 형태 주석

- 구어 말뭉치의 경우 구어 팀에서 학습자 오류 어절에 대해 일차적인 '교정 어절'을 주고 있다. 따라서 교정 어절이 없는 문어와 달리 구어에서는 형태 주석 단계에서 아래와 같이 '교정 어절'과 '교정 주석'이 함께 나타나는데, 교정 어절에 대해서도 형태 주석을 수정해 준다.

가) 통합이나 분할이 필요 없이 수정이 가능한 경우

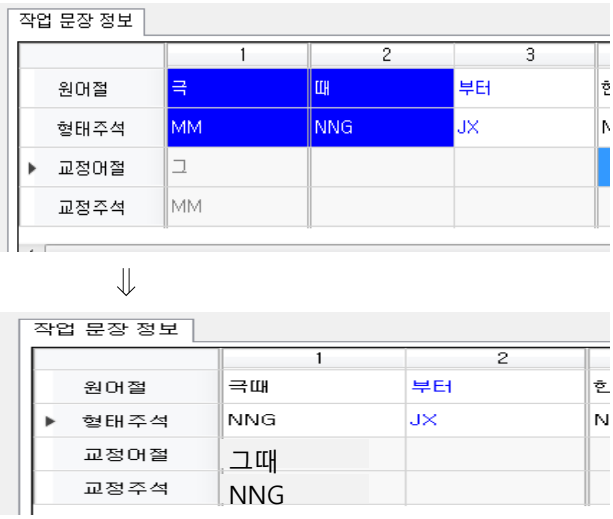
작업 문장 정보

	4	15	16	17	
원어절		많이	목꼬		문화
형태주석		MAG	VV	EC	NNG
교정어절			먹	고	
교정주석			VV	EC	

⇓

나) 통합이든 분할이든 수정했을 때 분석이 가능한 경우

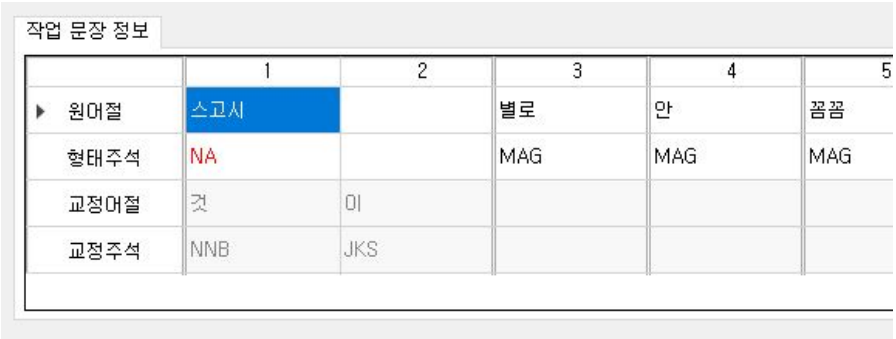
- ‘그때’는 한 단어로 NNG로 분석해야 한다. 먼저 원어절에서 어절 경계가 분할된 1번과 2번의 어절을 통합해서 1번을 ‘극때’로 수정하고 NNG로 분석을 한 후, 2번을 삭제한다. 이때 아래 교정 어절에 ‘그/MM’만 남게 되는데 ‘그때’로 수정하고 NNG로 교정 주석을 수정한다.



다) 통합이든 분할이든 수정할 수 없으며, 원어절의 형태주석을 NA 처리해야 하는 경우

- 구어 팀에서 제시한 교정어절대로 형태 분석하기에 무리가 있다고 판단되는 경우는 NA로 분석한다. 아래 제시된 경우처럼 ‘스고시’는 ‘것이’의 오류로 ‘것+이’로 분리하여 분석하기에는 무리가 있으므로 NA로 처리해야

한다. 이 경우 교정어절에 '것' 위의 형태주석 1번 칸의 '스고시'는 NA로 수정하고, 2번 칸의 형태 주석은 시스템 왼쪽 아래의 '주석 삭제' 기능을 이용하여 삭제하고 빈칸으로 남겨둔다.



- 위와 같이 구어에서 제시한 교정어절대로 형태 분석을 하기 어려운 경우는, 메모를 남겨 다음 단계 작업에서 확인할 수 있도록 한다.

<메모 형식> 구어 팀에서 제시한 교정어절대로 '○○'을 '△△'으로 교정할 가능성이 낮아 보임

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. 모든-오류는 오류 위치 검색으로 찾을 수 있다.[31]

	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

31) 구 단위 주석과 표현 문형 주석은 구 전체와 구 구성요소에 각각 주석함을 원칙으로 한다.

	오류 유형		주석 표지
		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		PHE
	표현 문형		PE

2. 확장 주석

○ 확장 주석은 한국어교육의 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 발생되는 오류 양상만을 주석한다.[32]

	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

32) 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
쪼금 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.[33]

> <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. 단, 표현 문형의 경우는 <국립국어원 2> 목록을 확인하여 해당 표현이 목록에 있을 경우는 표현 문형도 함께 주석한다. 따라서 형태 단위로 분석하여 오류로 처리하는 동시에 '표현 문형' 오류도 중복 주석한다. (☞ 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. 즉, [오류 위치-표현 문형, 관형사형 전성어미], [오류 양상-대치] 오류로 주석한다.

33) 이후 웹사이트에서는 교정어절을 중심으로 한 검색도 가능하게 하여, 미사용으로 인한 오류를 파악할 수 있게 한다.

은행에 저축한(√저축할) 겨우에는(√경우에는) 얼마정도 이익을 얻을지 미리 알아서 더 편할 것 같아요.

☞ '-(으)ㄹ 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 아니다는 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 좋좀하고(√좋고? 조용하고?, IMP) 예쁘기 대문에(때문에, MIF) 결훈(√결혼, MIF)하고 싶어요.

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.[34)]

○ 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,

34) 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 아른다운 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	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	그런대(√그런데, CMAJ, MIF) 제 가적을 정말 보고 싶습니다.
	감탄사	CIC	감탄사 어휘의 사용에서 나타난 오류	'녜(√네, CIC, MIF), 알겠습니다'라고 대딥했다.
	접두사	CXPN	접두사 어휘의 사용에서 나타난 오류	최조 임금을 실행하면 처소득층(√저소득층, CXPN, MIF) 사람의 살기가 보증할 수 있다.
	명사파생 접미사	CXSN	명사 파생 접미사 어휘의 사용에서 나타난 오류	두 번제(√번째, CXSN, MIF)에 간 곳이는 경주였습니다.
	동사파생 접미사	CXSV	동사 파생 접미사 어휘의 사용에서 나타난 오류	그 꿈을 위해서 매일 운동해다(√운동한다, CXSV, MIF).
	형용사파생접미사	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 사용에서 나타난 오류	그리고 여행이(√을, FNP, REP) 너무 좋아합니다. 고시원에서 많이 학생(√이, FNP, OM) 살았다.
	관형격 조사	FGP	관형격조사의 사용에서 나타난 오류	'론딩리'의(√를, FGP, REP) 소개합니다.
	목적격 조사	FOP	목적격조사의 사용에서 나타난 오류	그래서 저는 한국 사람하고 다른 외국 시람을(√과, FOP, MIF/REP) 교류하고 싶습니다.
	부사격 조사	FAP	부사격조사의 사용에서 나타난 오류	차 안에(√에서, FAP, REP) 잤어요.
	접속 조사	FJC	접속조사의 사용에서 나타난 오류	한국어 문법와(√과, FJC, MIF) 중국어 문법이 비슷하지 않았다.
	호격 조사	FVP	호격조사의 사용에서 나타난 오류	친구아(√친구야, FVP MIF), 같이 가자.
	인용격 조사	FQP	인용격조사의 사용에서 나타난 오류	내가 감사하다고 말한다는 게 '고맙다'다고(√라고, FQP, REP) 했다.
	보조사	FXP	보조사의 사용에서 나타난 오류	론딩리는 맛있는 음식은(√이, FXP, REP) 많습니다.
	연결 어미	FED	연결어미의 사용에서 나타난 오류	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 겁다.
	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	앞으로는 얼마나 어려운 일이 생기는 때(√생겨, PHE, REP)도 포기하지 않고 사소한 일에도 최선을 다하겠다.
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 만련이다(생기기 마련이다, PE, MIF).

3) 오류 양상

○ 오류 양상은 표면적으로 드러난 오류의 모습으로, 누락, 첨가, 대치, 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 시간는 별로 덥지 않고 시원하서(√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] 한국게(√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 그라프(√그래프, CNNG, PN)를 보며는,
	중간 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	단어 형성[합성법]	MCP	단어 합성에서 나타나는 오류	해물고기(√물고기, CNNG, MCP)가 많았어요.
	단어 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	할머니께서 우유를 마시십니다(√드십니다, CVV, REP, SH).
	시제	ST	시제를 나타내는 문법 형태소의 오류	어제부터 항상 시계를 확인하기로 한다(√했다, FPE, REP, ST).

층위		주석 표지	포함 범위	예시
	사동	SC	사동사, 사동 표현의 오류	갈릴레이는 새로 발명된 망원경을 사용하여 연구를 깊였다. (√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 문법성보다는 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 남자을 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 줄업만 뜬다면 드디오 내가 기다리는 시간이라고 생각하고 귀가 아주 밝다(√IMP).

○ 분석 여부는 기본 값이 ‘분석 가능’으로 설정되어 있으므로, 오류임에 분명하지만 교정어절을 주기 어려워 오류의 판정이 불가능한 경우에만 주석을 한다.

2) 오류 위치

(1) 실질어휘

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 10멷(√년) 후에 자기 가 수 있다. 내가 한국에 온 지 7개원(√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 일홉(√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 십습니다(√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 둬야 돤다(√된다)
☞ '돤다'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(√학생이에요)
☞ 지정사 '이에요'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 잘못한 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.
남자이에요(√남자예요)

☞ 받침이 없는 명사 뒤에서 ‘이에요’를 ‘예요’로 줄여서 쓰는 것이 일반적이나, 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 ‘에요’는 종결어미의 오철자 오류로 처리한다. [오류 위치-지정사], [오류 양상-누락], [오류 위치-종결어미], [오류 양상-오형태] 오류로 주석한다.

학생에요(√학생이에요)
☞ [오류 위치-지정사], [오류 양상-누락]
학생이예요(√학생이에요)
☞ [오류 위치-종결어미], [오류 양상-오형태]
아니예요(√아니에요)
☞ [오류 위치-종결어미], [오류 양상-오형태]
학생이여서(√학생이어서)
☞ [오류 위치-연결어미], [오류 양상-오형태]
학생이였어요(√학생이었어요)
☞ [오류 위치-선어말어미], [오류 양상-오형태]
공부를 할 거에요(√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 학꾄데(√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) 멌있게 생겼다.
어히려(√오히려) 남성보다 여성의 힘이 더 강하는 경우도 있는 정도다.

⑫ 접속부사(CMAJ)

○ 접속부사의 형태, 의미, 사용 오류를 말한다.

<예> 그레서(√그래서) 피자하고 맥주를 먹고 많이 애기했다.
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 안넝(√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 색학히면 세대 차이를 줄일 수 있다.

⑮ 명사파생접미사(CXSN)

○ 명사파생접미사의 형태, 의미, 사용 오류를 말한다.

<예>	저의 첫 번제(√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 좋아햅니다(√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 따뜽(√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.

기속사의(√기숙사에) 규칙을 있다.

③ 목적격조사(FOP)

○ 목적격조사의 형태, 의미, 사용 오류를 말한다.

<예> 집에서 포도을(√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 끈저 area에 가지 안됐다. 미국에(√에서) 영어 제일 중요한다.

⑤ 접속조사(FJC)

○ 접속조사의 형태, 의미, 사용 오류를 말한다.

<예> 미국고(√과) 일본이 두 나라에 가고 싶다. 외모과(√와) 노력이 다 중요한다.

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 나한데 "수협 후에 사무실에

와요"이라고(√라고) 말았다. 나는 원래 "생선을 먹었어요"(√라고) 말했어야 했는데 김장해서 "선생을 먹었어요"라고 말았다.

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) 짬깐 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(-ㄴ다/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 앉은다(√앉는다).

⑫ 선어말어미(FPE)

○ 선어말어미의 형태, 의미, 사용과 관련된 오류를 말한다.

<예> 인중 차별 때문에 많은 사람들이 죽었으니까 앞으로 인중 차별이 없었으면 좋게다(√좋겠다).
그래서 너무 배가 고펐다(√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) 중요한다.

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 까지 음식을 골고루 먹을(√먹는) 것이 중요해요.

☞ 관형사형 전성어미 대치 오류로 처리한다.
한국에 온(√오는) 비행기에서 친구를 만났어요.
☞ 관형사형 전성어미 대치 오류로 처리하며, 시제와 관련되어 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(√ 끝에) 남편과 5살 어린 아이를 두고 더 좋은 인생을 찾으러 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, '나타나다' 동사의 대치 오류로만 처리할 것인지 문제가 될 수 있다. 이 경우 가능한 분리하여 '나타났다'를 '되다'로 교정하여 [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 동사 대치로 처리한다(본 연구는 연어 오류를 주석하지 않는다).

<예> 태도를 키워야 한다 (√ 길러야 한다)
☞ '태도를 키워야 한다'를 '태도를 길러야 한다'로 교정할 때, 이를 연어 오류로 볼 수 있다. 그러나 구 단위 또는 연어

표현에서 용인 가능성의 기준이 판단자마다 다를 수 있다는 문제가 있다. 따라서 연어 오류를 주석하기 위해서는 연어 목록이 선행되어야 하고, 연어에 대한 판단이 정해져야 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.[35)]

○ 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 목록

35) 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.

에 있기 때문에 표현 문형도 오류 위치에 중복 주석한다. 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 수의적이거나 없는 정보를 더 추가해주지 않도록 한다..

<예> 저는 여러 까지 능력서가 취득하지만 그 중에서 영어를 (√가장, OM?) 능수하는 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 기본이 왔는데요.
☞ '있을'이 누락되었는데, 형태주석을 기본단위로 오류주석을 할 때, 동사 '있'과 관형사형 전성어미 '을'을 각각 누락오류로 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 누락]은 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.

○ [조사 누락]은 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 누락]은 다음과 같은 것들이 해당된다.

<예> 10년 후에 내가 가(√갈) 수 있다. ☞ 관형사형 전성어미를 누락한 오류로, 관형사형 전성어미 '-(으)ㄹ'의 [누락]으로 주석한다. 경우에 따라서 받침을 제대로 쓰지 못한 오철자 오류로 볼 수도 있으나, 주석의 일관성을 위해 누락 오류로 주석한다.

※ 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 'ㅁ'이 누락되었으나, 문어에서 단어 내에서의 음소 생략은 오형태 오류로 처리한다. 아침에 친구를 마나서(√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 첨가]는 다음과 같은 것들이 해당된다.

<예> 대만의 수도가(√수도) 타이베이는

☞ 주격조사 ‘가’가 첨가된 오류로 주석한다.

※참고: 이때 ‘수도가’를 ‘수도인’으로 교정할 경우에는, 주격조사 ‘가’와 서술격조사 ‘이’의 대치 오류로도 볼 수 있으나 본 연구에서는 ‘최소 수정 원칙’으로 주격조사 첨가 오류로 처리한다.

○ [표현 문형 첨가]는 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 알여주는 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 품사

로, 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> 누구가 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 바뀌는 어미나 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 대치]는 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.

○ [연결어미 대치]는 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. 이때에는 문맥을 고려하여 오류를 판단하며, 학습자들이 얼마나 이러한 오류를 보일 수 있는가를 고려한다. 다시 말해서 수준별로 사용할 수 있는 어휘를 고려하여 오류를 처리한다. 아울러 최소 수정 원칙을 기본으로 하되, 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 해 볼 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 용언 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 잘못한 경우는 굴절(활용)(MCJ)으로 주석한다.

<예> 우리나라에서 과일들하고 야채들도 많아서 과일와(√과) 야

채도 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 'ㅂ니다'에 대한 철자 오류로, [오류 위치-종결어미], [오류 양상-오형태(MIF)] 오류로 주석한다.

○ [처리 예시] 오형태 오류의 오류 위치별 일부 예시는 다음과 같다.

○ [명사 오형태]는 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 오형태]는 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 오형태]는 다음과 같은 것들이 해당된다.

<예> 제주 친구하고 옥등산에서 등산을 가서요(√갔어요).

☞ 과거 시제를 나타내는 선어말 어미 '-았-'이 생략된 형태이다. 그러나 '가어요'로 쓰지 않고 '가서요'로 썼기 때문

에 이것은 과거를 인식하고 있다고 보고 오형태 오류로 처리한다. 즉, '해습니다', '마셔지만', '와지만'처럼 과거 시제 선어말 어미 '었'에서 '씨'을 누락시킨 경우는 선어말 어미 오형태(오철자) 오류로 처리한다. [오류 위치-선어말 어미], [오류 양상-오형태] 오류로 처리한다.

○ [연결어미 오형태]는 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 오형태]는 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)

☞ 'ㅂ니다/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.

<예> 팔이(√빨리), 말을::, 즐= 아:: 잘해서::,

☞ 자음 'ㅃ'을 'ㅍ'로 발음하여, 음소 오류로 주석한다. [오류 위치-일반부사], [오류 양상-없음], [오류 층위-음소]

에:: 코피(√커피)= 카페에서::, 에:: 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], [오류 층위-음소]

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 늘여 발음하거나 축약이 불가능한 단어를 축약하여 발음한 경우 음절 오류로 주석한다.

<예> 에:: 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 'ㄹ'과 모음 'ㅇ'를 결합하여 3음절로 발음하고 있으므로 음절 오류로 주석한다. [오류 위치-의존명사], [오류 양상-없음], [오류 층위-음절]로 주석한다.

③ 음운규칙(PC)

○ [정의] 음운규칙 오류는 구어 발화에 나타난 필수적 음운 규칙의 일탈을 말한다. 유음화, 연음화하여 발음해야 하는데, 글자 그대로 절음화하여 발화한 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 [유캉년]으로 발음하지 않고 글자 그대로 [육학년]으로 발음한 경우나 '학교'를 [학꾜]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> 학쿄(√학교)
☞ [학꾜]로 발음해야 하는데 [학쿄]로 발음했을 경우, 음운규칙과 음소 오류를 중복 주석한다. 그러나 음운규칙이 적용되지 않는 위치에서 다른 음소로 발음한 경우는 음소(PP) 오류로만 처리한다(예 핵교(√학교).

○ 구어의 특성이나 표현 의도에 의한 발음 특성은 오류로 처리하지 않는다.

<예> 표현 의도에 의한 수의적 경음화: 쪼금
구어에 의한 발음 특성: ~먹었구요 / ~했어여

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> 인털뷰(√인터뷰)
세너(√센터)
☞ 한국어 외래어 표기법과 다르게 원어식 발음으로 발화한 경우 원어식 발음 오류로 처리한다.

팥너(√파트너)
그대 처음메 같 뻳남(√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:: [이메지)](√이미지) 좀:: 없어,지고:: 있습니다 지금.
☞ 외국어 발음 오류로 외래어 표기법과 '다르다'는 차원에서 [오류 양상-오형태] 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 ㅋ과 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)

○ 단어 형성[합성법] 오류는 조어 과정에서 발생하는 오류를 말한다. 즉, 학습자가 존재하지 않는 어휘를 생산해 내는 오류가 포함된다.

○ 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 오류 양상을 필수적으로 주석하지 않고 오류 층위에서 파생과 합성만을 주석하도록 한다.

○ 따라서 학습자가 생산해 낸 형태가 한국어에는 없는 합성어일 경우, 오류 양상은 주석하지 않고 오류 층위에서 합성으로 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)

○ 단어 형성[파생법] 오류는 조어 과정에서 접사를 잘못 사용한 오류를 말한다.

○ 학습자가 파생접사(유사파생접사 포함)를 사용하여 생산해 낸 형태가 한국어에는 없는 파생어일 경우, 오류 양상은 주석하지 않고 오류 층위에서 파생으로 주석한다.

○ 단, 접사와 접사의 대치의 경우나 접사의 불필요한 첨가 또는 생략은 오

류 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 선생님에 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.

○ 동사 어간에 '하다'를 붙여 한국어에는 없는 동사를 만들어냈을 경우는 파생 오류로 처리한다. 이는 합성 오류로도 볼 수 있지만 형태 주석에서 '-하다'를 파생접미사로 처리하고 있어 처리의 연계성과 일관성을 고려하여 파생 오류로 처리한다.

<예> 그날 수업 후 집에 도차자마자 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. 형태 주석에서는 이를 동사파생접미사로 주석하기 때문에 오류

주석에서도 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 형태 주석에서 접사로 따로 분리하여 처리하기 때문에 형태 주석에서 분리하는 접사들은 '누락/첨가'로 처리하고, 피동/사동은 어휘 대치로 처리한다. 또한 어휘적 의미를 더해주는 접사의 경우, 예를 들어 '사과'를 '풋사과'로 썼을 때에는 형태 분석에서 '풋'을 분리하지 않기 때문에 어휘적 의미를 더해주는 접사가 덧붙여진 단어는 대치 오류로 처리한다.

○ 그밖에 형태 주석에서 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 형태주석에서 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)

○ 굴절[곡용] 오류는 조사의 이형태를 잘못 사용한 경우를 말한다.

○ 굴절[곡용] 오류의 오류 양상은 기본적으로 오형태 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 잘못한 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)를 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 중요한다(√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 'ㄹ' 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.)

☞ 'ㅂ불규칙 활용' 형용사인 '덥다'를 종결어미 'ㅂ니다' 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 주석한다.

○ [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 좋아한다.

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 어간으로 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 재구성한 중간언어로 보고 용언의 활용 오류로 처리한다. 이는 학습자들이 연결어미(예: 해고), 종결어미(예: 햅니

다), 관형사형 전성어미(예: 해는) 앞에서 동일한 형태를 생산해 내는 것으로 보아, 학습자들이 '해'를 하나의 단위로 인식하고 활용을 잘못 적용하였다고 판단했기 때문이다. 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 삭제돼면(√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 [ll] 발음의 영향으로 인한 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 시켜야 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)

○ 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.

○ 오류 층위에서 품사에 해당하는 오류는 품사에서 나타난 오류(품사가 달라진 경우)와 품사를 몰라서 생겨난 오류(품사 혼동으로 의한 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.

○ 품사 오류의 오류 양상은 기본적으로 대치 오류로 주석한다.

<예> 그래서 우리는 빠른(√빨리) 우리 집에 다가했다.

☞ 부사 '빨리'를 쓸 자리에 형용사 '빠르다'를 사용하였다. 이는 학습자가 동일한 의미의 단어에서 부사 품사를 모

르기 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 생각기는 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 위우기 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 추구와 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 주석한다.

<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 긁어서 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, -다고, -하니까’ 등 관형사형 전성어미나 연결어미 등은 ‘누락’ 처리하지 않는다.

(3) 통사

① 높임(SH)

○ 높임 오류는 조사, 선어말어미, 종결어미 부분의 높임 관련 문법 형태소, 높임 어휘의 사용이 잘못된 경우를 말한다.

○ 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 ‘계시다, 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.

○ 문장 이상의 단위인 담화 상에서 높임말과 반말을 혼용하여 사용하고 있는 경우에는 문장 단위로 오류를 판단한다는 원칙하에 한 문장 내에서 높임 표현에 문제가 없으면 오류로 주석하지 않는다.

○ 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.

○ 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)

○ 시제 오류는 시제 또는 시상을 나타내기 위한 선어말어미, 관형사형 전성어미, '-(으)ㄹ 것' 등의 오류를 말한다.

<예> 옛날에 한국의 정통 난방법을 온돌이다(√온돌이었다).
때로는 한국말을 공부할 때 끝이 없는 것 같은데(√같았는데) 오느날 갑자기 끝에 왔다.

○ 시제 오류는 오류 양상을 두 가지로 처리한다. 하나는 시제 요소를 사용하지 않고 기본형을 사용했을 경우, 시제를 사용하지 않았다고 판단하여 누락 오류로 주석한다. 다른 하나는 시제를 제대로 인식하지 못하여 과거 시제나 미래 시제 자리에 현재 시제를 사용한 경우, 그 반대의 경우 등은 대치 오류로 처리한다. 즉, 시제를 사용했으나 현재와 과거, 과거와 미래처럼 잘못 사용한 경우는 시제 간 대치로 처리한다.

○ 이때, '이다/아니다', 형용사, 연결어미 앞에서의 용언은 기본형이 현재를 나타내기 때문에 기본형을 사용했을 경우, 현재 시제로 인식하고 대치로 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, 예쁜다(√예쁘다)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 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.

○ 미래시제를 나타내는 '-겠-'의 경우, 추측이나 의지와 같은 양태 의미와도 관련되는데, 본 연구에서는 추측을 제외한 경우 오류 층위에 시제를 주석한다.

<예> 앞으로 학교 규칙을 안 어긴다(√어기겠다).
☞ 미래의 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 경우는 대치 또는 오형태 활용 오류로 처리하고, 시제를 나타낼 때에는 관형사형 전성어미의 시제 대치 오류로 처리한다.

○ 시제를 나타내지 않는 경우로 '-(으)ㄹ 때, -(으)ㄹ 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)ㄹ 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)ㄹ')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)ㄹ지'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)ㄹ지'의 대치의 경우, 오류 층위에 시제를 주석하지 않도록 주의한다.

③ 사동(SC)

○ 사동 오류는 사동사, 사동 표현의 사용, 사동문 생성에서 발생한 오류를 말한다.

○ 사동사, 사동 표현을 사용해야 하는데 사용하지 않은 경우는 기본적으로 오류 양상을 대치 오류로 주석한다.

○ 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 밥률 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 알여주는(√알려주는) 것도 전통을 보존하려고 해야 할 일이다.

☞ '알여주는'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 주석한다.

○ 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 에어큰의 온도를 조금만 높게 하는(√높이는) 것만 한 방법이 없다.
☞ '-게 하다'가 사동사로 대치된 경우로 오류 위치는 '연결어미, 보조용언, 표현 문형으로 주석하며, 오류 양상은 대치로 주석한다.

④ 피동(SP)

○ 피동사, 피동 표현의 사용, 피동문 생성에서 발생한 오류를 말한다.

○ 피동사, 피동 표현을 사용해야 하는데 사용하지 않은 경우는 기본적으로 대치 오류로 처리한다.

○ 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 주석에 따라 일괄 처리하여, 동사로 분석했을 때는 그 품사를 따르고, 연결어미, 보조 용언으로 분리하였을 경우 해당 품사를 오류 위치로 주석한다.

○ 이중피동을 사용한 경우, 첨가 오류로 처리한다.

<예> 환경 오염이 심해지게 되고(√심해지고) 있지만 더 이상 심하

기 전에 여러분의 도움이 필요한다고 생각하다.
☞ '심해지다'에 '-게 되다'까지 첨가된 이중 피동표현으로 '-게 되다'의 '연결어미+보조용언', 표현문형(PE)의 첨가 오류로 처리한다.

⑤ 부정(SN)

○ 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.

○ 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.

○ '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)

○ 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.

○ 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.

○ 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.

○ 2개의 문장 성분이 상호 교체될 때에는 2개 모두 대치 어순 오류로 주석한다.

<예> 8반 시(√8시 반)에 학교에 거러서 가요
☞ 시간표현에서 시보다 분을 먼저 배열했기 때문에, '반'과 '시'의 어순 대치 오류로 주석한다. 이때, 어순이 상호교체되는 것으로 명사 '반'과 의존명사 '시' 모두를 대치 어순 오류로 주석한다.

○ 조사의 경우, 어순 오류로 처리하지 않고 조사 첨가 또는 누락 오류로 처리한다.

<예> 그래서 한국말을(√한국말) 공부(√공부를) 참 좋아했니다.

☞ ‘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 처리함을 원칙으로 한다. 이에 따라, 지시 표현에서 의미 간 대치 오류를 중심으로 하여 오류 층위에 지시(DR)를 주석한다. 지시 표현에서 나타난 단순 오철자 오류는 오류 양상에 오형태만 주석하고, 오류 층위에서 지시를 주석하지 않도록 주의한다.

○ [처리 기준] 지시 오류는 앞 뒤 문장과의 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 오형태만 주석하고, 오류 층위에서 접속을 주석하지 않도록 주의한다.

○ [처리 기준] 접속 오류는 앞 뒤 문장과의 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 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. 미시 담화표지의 경우, 학습자의 L1의 영향으로 인한 간투사 사용과 모어화자와는 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.

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 문장

과 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/게(√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

1) 구어 오류 주석 기본 원칙

○ 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
쪼금 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. 수정 후 발화에 초점을 두고 오류 여부를 판정한다.

<예> 친= 친구가 한국 음식이:: = 음식을 좋아해서

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

○ 구어 오류 주석과 문어 오류 주석의 기본 원칙 및 처리 방법은 동일하다. 그러나 발화 상에서 나타나는 발음 오류의 경우에는 오류 양상(대치, 누락, 첨가, 오형태)을 주석하지 않고, 오류 위치와 오류 층위 [발음]만 주석한다.
○ 구어에서는 발음 오류와 어휘 및 문법 오류의 구분이 명확하지 않을 수 있다. 즉, 학습자가 어휘와 문법을 잘못 사용한 것인지 단순히 발음을 잘못한 것으로 인해 나타난 오류인지 판별하는 데 어려움이 있다. 구어에서는 발음의 영향과 함께 어휘 및 문법 오류를 표시해주는 차원에서 오류 층위에서 중복 주석을 하도록 한다. 그러나 조사의 경우, 문법 오류를 우선 처리하도록 하고, 관형사형 전성어미를 사용해야 할 자리에 사용하지 못했을 경우도 받침을 발음하지 못한 음소 오류보다는 누락 오류로 문법 오류를 우선 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 경우

가 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](좀), [바래요](바라요) 등 구어에서 허용되는 발음은 오류로 처리하지 않는다.

○ 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::

☞ 그러나 <표준>에 등재되어 있지 않더라도, '그쵸'와 같이

구어에서 축약된 형태로 많이 나타나는 용례들은 오류로 처리하지 않는다.

○ 구어에서 발음의 차원이 아닌, 형태를 잘못 발화한 경우는 오형태(MIF) 오류로 처리한다.

<예> 사잉(√사건)
보석필(√보살핌)
☞ 이와 같은 예시들은 발음을 잘못했다기보다 형태를 잘못 만들어낸 것이다. 유사 발음과도 멀어져 음소 오류로 볼 수 없고, 한국어에 없는 형태들을 발음했다고 판단하여 오형태 오류로 처리한다.

○ 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 전사가가 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> [보쿙](√복용),하는 뜻이는(√뜻은)::, 그 마약을 쓰는 아니면 마약을, 하,는:: 것입니다.

☞ 복용을 [보굥]으로 발음해야 하나 [보쿙]으로 발음하였으므로 음소 오류로 처리한다. 오류 층위[발음]에 해당하는 오류이기 때문에 오류 양상은 주석하지 않고, 오류 위치와 오류 층위만 주석한다. 따라서 [오류 위치-명사], [오류 양상-없음(빈칸)], [오류 층위-음소]로 주석한다.

또한 '뜻이는'은 '뜻은'에 조사 '이'를 첨가한 것이기 때문에 발음의 오류가 아닌, 문어와 마찬가지로 조사 첨가 오류로 처리한다. 이 경우, [오류 위치-주격조사], [오류 양상-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		
-ㄴ데도 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, ###, ㄴ대로	
-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 둥 마는 둥		###는 둥 마는 둥
-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 체하다		
-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
-ㄹ 것 같다		
-ㄹ 것1		
-ㄹ 것2		
-ㄹ 것이 아니라		
-ㄹ 대로		
-ㄹ 듯		
-ㄹ 듯하다		
-ㄹ 따름이다		
-ㄹ 때		
-ㄹ 리가 없다		
-ㄹ 만큼		
-ㄹ 만하다		
-ㄹ 모양이다		
-ㄹ 바에		
-ㄹ 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	###ㄴ데도 불구하고, ###는데도 불구하고	
을 가지고	###ㄹ 가지고	
-을 것 같다	###ㄹ 것 같다	
-을 것1	###ㄹ 것	
-을 것2	###ㄹ 것	
-을 것이 아니라	###ㄹ 것이 아니라	
-을 나름이다		
-을 대로	###ㄹ 대로	
-을 듯	###ㄹ 듯	
-을 듯하다	###ㄹ 듯하다	
-을 따름이다	###ㄹ 따름이다	
-을 때	###ㄹ 때	
-을 리가 없다	###ㄹ 리가 없다	
-을 리가 있다		
을 막론하고	를 막론하고	
-을 만큼	###ㄹ 만큼	
-을 만하다	###ㄹ 만하다	

표제어	형태 정보	대표형
-을 모양이다	###ㄹ 모양이다	
-을 바에	###ㄹ 바에	
-을 법하다	###ㄹ 법하다	
을 불문하고		
-을 뻔하다	###ㄹ 뻔하다	
-을 뿐만 아니라	###ㄹ 뿐이다	
-을 뿐이다		
-을 수 없다		
-을 수 있다		
-을 수밖에 없다	###ㄹ 수밖에 없다	
을 위해(서)	###를 위해(서)	
-을 줄	###ㄹ 줄	
-을 테고	###ㄹ 테고	
-을 테냐	###ㄹ 테냐	
-을 테니	###ㄹ 테니	
-을 테니까		
-을 테다	###ㄹ 테다	
-을 테면	###ㄹ 테면	
-을 테야	###ㄹ 테야	
-을 테지만	###ㄹ 테지만	
-을 텐데	###ㄹ 텐데	
-을까 보다	###ㄹ까 보다	
-을락 말락 하다	###ㄹ락 말락 하다	
-지 말다		###지 말다
-지 못하다		###지 못하다
-지 않다		###지 않다

<Abstract>

2022 Project on the Research and Construction of the Korean Language Learner Corpus

This study aimed to collect and construct a primitive corpus of 1 million words, according to the second mid- to long-term plan established in <2021 Korean Learner Corpus Research and Construction>. Further, considering the balance of the corpus, it also aimed to build and process a morphological annotation corpus of 300,000 words, and an error annotation corpus of 200,000 words. The major tasks and goals of this project are as follows.

Korean language learner corpus collection, revision, and construction: The 2022 learner corpus aimed to intensively collect and build relatively small data, by analyzing the distribution characteristics based on target, level, language, and data variability.

The team constructed new items in the form of 1,008,315 raw corpora (written items: 708,096, spoken items: 300,219), 308,790 morph-tagged corpora (written items: 157,313, spoken items: 151,477), and 203,490 error-annotated corpora (written items: 73,829, spoken items: 129,661). In total, 6,230,590 raw corpora (written items: 4,407,583, spoken items: 1,823,007), 4,013,233 morph-tagged corpora (written items: 2,760,085, spoken items: 1,253,148), and 1,346,015 error-annotated corpora (written items: 674,636, spoken items: 671,379) were produced.

In addition, to improve the utilization of the learner corpus, data were collected from approximately 30 native Korean speakers, constructing a reference corpus of 10,440 written words and 23,912 spoken words as a pilot project.

Elaboration of a learner corpus inspection and quality management: The inspection, elaboration, and quality control of the

learner corpus qualitatively improve and quantitatively expand it as a language resource, where the purpose is to secure data integrity. Accordingly, the work process was carried out based on the three-step work and inspection system, for each work step of written input, spoken word transcription, morphological annotation, and error annotation. Hence, an internal inspection team focusing on co-researchers and randomly reviewing data strengthened the inspection system.

Furthermore, this was complemented with an inspection for erroneous data and abnormal data through a system-based data verification; in the final stage, meta information was verified through a total sample information verification along with the duplicate samples. In addition, to improve the accuracy of the statistical information of the construction corpus, the statistical information of the instrumental axis corpus was reviewed. Based on these results, it was proposed to improve the history management method of the LCMS work assignment cancellation sample as well as the presentation method of the user search statistics of the "Korean Language Learner Corpus Search Engine".

Learner corpus education and promotion: Korean language learner corpus-related education was provided for construction practical staff and users. Training practitioner workers helps building a systematic corpus, by acquainting those with basic knowledge and skills related to corpus construction, and enhancing their expertise as a worker at each construction stage. It is based on instructional training and tools usage training. In addition, an immediate communication and feedback system was operated to solve various problems stemming from the construction process, whereas issues related to corpus construction and countermeasures were shared through regular workshops.

Training for users was conducted a total of four times through learner corpus academies from the basic course to the advanced course. The basic course had two sessions on data processing for research based on learner corpus. Additionally, the advanced course had two sessions on the usage

of the Korean learner corpus based on artificial intelligence technology and the search and application of research topics using the learner corpus. In addition, guide materials and videos introducing the Korean learner corpus and explaining how to use it were produced and distributed through the National Institute of Korean Language website and the “Korean Language Learner Corpus Search Engine”.

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, written language corpus, spoken language corpus, raw corpus, morph-tagged corpus, error-annotated corpus

<기획·연구>
국립국어원 홍혜진 학예연구관
국립국어원 박미영 학예연구사
국립국어원 유상미 연구원
국립국어원 조 은 연구원

<연구 참여자>
연구 책임자 한송화(연세대학교)
공동 연구원 강현화(연세대학교)
김선정(계명대학교)
김일환(성신여자대학교)
김한샘(연세대학교)
장석배(미국 밴더빌트대)
홍혜란(연세대학교)
홍혜진(국립국어원)
박미영(국립국어원)
연구 보조원 김동은(연세대학교)
김미선(연세대학교)
서지혜(연세대학교)
손연정(연세대학교)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2022년 12월 9일
발행일: 2022년 12월 9일
인 쇄: 학위사

※ “이 책은 국립국어원의 용역비로 수행한 ‘2022년 한국어 학습자 말뭉치 연구 및 구축’ 사업의 결과물을 발간한 것입니다.”