국립국어원 2024-01-08

발 간 등 록 번 호
11-1371028-001001-01

2023년 한국어 학습자 말뭉치 구축

연구 책임자
곽 용 진

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구 용역 계약에 따라 ‘2023년 한국어 학습자 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 7월～2024년 5월

2024년 5월 2일

연구 책임자: 곽용진((주)이르테크)

연구 기관	㈜이르테크, 연세대학교 산학협력단
연구 책임자	곽용진(㈜이르테크)
공동 연구원	한송화(연세대), 이지연(㈜이르테크), 강현화(연세대), 김선정(계명대), 김한샘(연세대), 홍혜란(연세대)
연구 보조원	이근영(㈜이르테크), 최용기(㈜이르테크), 이순미(㈜이르테크), 정해영(㈜이르테크), 홍은기(㈜이르테크), 오누리(㈜이르테크), 김윤희(㈜이르테크), 박정혁(㈜이르테크), 박하영(㈜이르테크), 박우정(㈜이르테크), 임수용(㈜이르테크), 박상원(㈜이르테크), 오수연(㈜이르테크), 조현수(㈜이르테크), 윤하정(㈜이르테크), 안수빈(㈜이르테크), 최영롱(연세대), 신유진(연세대), 하태현(연세대), 오지연(연세대)

<국문 초록>

2023 한국어 학습자 말뭉치 구축

본 사업은 한국어 학습자 말뭉치 제2차 중장기 계획의 일환으로 국가 언어 자원으로서의 말뭉치 규모를 확대하고, 언어 자원 활용 활성화 기반을 마련하는 것을 목표로 하였다. 이에 따른 주요 과업과 성과는 다음과 같다.

○ 한국어 학습자 말뭉치 구축

한국어 학습자 쓰기 자료 15,779개, 총 2,085,629어절을 문어 원시 말뭉치로 구축하였다. 이는 한국어 학습자 말뭉치 중장기 계획의 구축 목표인 1,000만 어절 중 약 20%에 해당하는 구축량으로 양적으로 큰 성과를 이루었다고 할 수 있다. 특히 기획 수집을 진행하여 쓰기 장르의 균형성을 보완하였다는 점, 전체 구축량의 53%를 국외 자료로 구축함으로써 기존에 1.2%였던 국외 수집량을 17.8%로 확대하였다는 점에 의의가 있다.

○ 한국어 학습자 쓰기 평가 자료 구축

주제와 장르가 서로 다른 2종의 쓰기 자료 각 1,500건씩 총 3,000건의 작문을 대상으로 한국어 학습자 쓰기 평가 자료를 구축하였다. 이를 위해 등급 체계 및 평가 구인, 평가 기준 설정의 바탕이 되는 기초 연구를 수행하고, 내용 조직, 글의 구조, 언어 사용의 세 가지 구인을 설정하고 각 구인별로 척도와 점수를 부여하는 분석적 채점을 실시하였다. 그 결과 등급별 점수(100점 만점), 1-6급의 누적 점수(600점 만점), 최종 점수(100점 만점)로 3종의 점수와 함께, ‘국제 통용 한국어 교육 표준 모형’을 기준으로 한 등급, 한국어교육 전문가의 총평이 부여되었다. 이러한 성과는 기구축 말뭉치에서 쟁점이 되었던 등급 표준화의 기반을 마련하고, 자동 채점, 자동 첨삭 등의 기술 개발을 위한 기반 자료를 마련하여 공공 데이터로서 연구 및 교육, 산업계에서의 활용도를 제고하였다는 점에서 의미가 크다.

○ **한국어 학습자 말뭉치 교육 및 홍보**

자료 수집과 구축을 위한 실무자 교육을 실시하고, 한국어 학습자 말뭉치 이용자를 대상으로 총 6회의 ‘한국어 학습자 말뭉치 교실’을 개최하였다. ‘한국어 학습자 말뭉치 교실’은 한국어교육 학계 및 교육 분야 외에도 민간 분야로의 저변 확대를 위하여 교육 대상, 진행 방식, 프로그램을 다양하게 구성하여 운영하였다. 또한 보다 나은 교육 프로그램을 개발해 나갈 수 있도록 2023년 한국어 학습자 말뭉치 교실 참가자를 대상으로 교육 프로그램에 대한 의견을 수렴하였다. 그 외에도 <학습자 말뭉치 나눔터>에 게시된 ‘학습자 말뭉치 활용 매뉴얼’을 매 회차의 학습자 말뭉치 교실 참가자들에게 홍보하고 배포하여 이용을 활성화하였다. 이는 한국어 학습자 말뭉치의 체계적인 구축을 위한 기반을 마련하고, 사업의 성과를 확산하고 활용도를 하였다는 점에서 의의가 있다.

최근 인공 지능 기술 발전에 따라 한국어교육 분야에서도 언어 자원의 수요가 증대하고 있다. 한국어 학습자 말뭉치는 자연언어 처리, 에듀테크 기술 개발 등의 원천 자료로 사용되어 여러 연구 사업에 이바지할 것으로 기대된다.

주요어 : 한국어 학습자 말뭉치, 문어 말뭉치, 원시 말뭉치, 학습자 쓰기 평가, 한국어 학습자 말뭉치 교육

차 례

제1장 사업 개요

제2장 사업 수행 내용

차 례

제3장 결론

차 례

표 차례

차 례

그림 차례

차 례

제 1 장

사업 개요

1. 사업의 목적 및 범위

1.1. 사업의 목적

본 사업은 한국어 학습자 말뭉치 제2차 중장기 계획('15~'25년 총 1,000만 어절 구축)의 일환으로 국가 언어 자원으로서의 말뭉치 규모를 확대하고, 언어 자원 활용 활성화 기반을 마련하는 것을 목표로 하였다. 이를 위해 기구축된 6,264,557어절의 말뭉치에 2,085,629어절의 원시 말뭉치를 추가 구축하여 누적 규모 8,350,186어절의 원시 말뭉치를 구축하였다. 또한 한국어교육 전문가의 쓰기 자료 평가를 통해 3,000건 이상의 한국어 학습자 쓰기 평가 자료를 구축하고, 말뭉치 구축자와 이용자를 위한 한국어 학습자 말뭉치 교육 및 홍보를 실시하였다. 다음은 본 사업의 목표 달성을 위해 수행한 핵심 과업이다.

- 한국어 학습자 말뭉치 구축
- 한국어 학습자 쓰기 평가 자료 구축
- 한국어 학습자 말뭉치 교육 및 홍보

1.2. 사업의 범위

본 사업의 범위와 세부 과업 내용은 다음과 같다.

<표 1> 사업의 범위와 세부 과업 내용

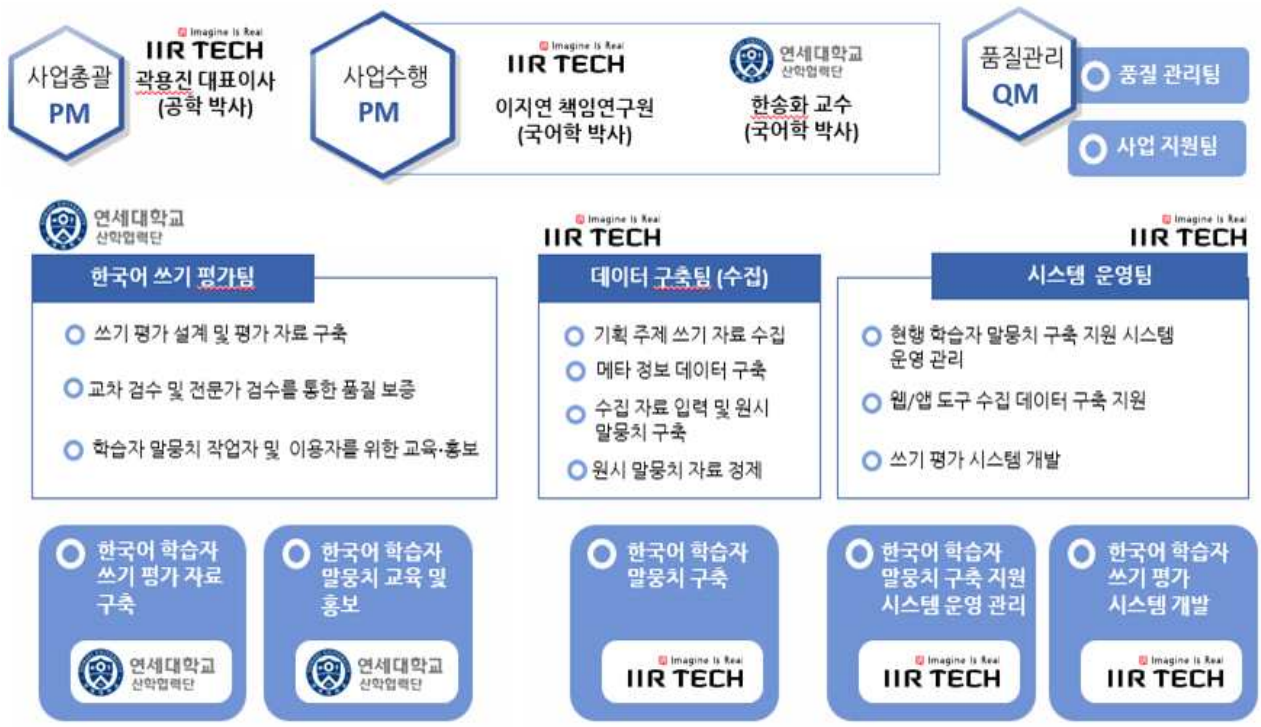
사업의 범위	세부 과업 내용
한국어 학습자 말뭉치 구축	ㅇ 한국어 학습자 쓰기 자료 15,000편 이상 수집 ㅇ 구축 대상 자료에 대한 메타 정보 구축 ㅇ 쓰기 자료 입력 및 원시 말뭉치 구축 ㅇ 개인정보 표지 부착 등 자료 정제 ㅇ 자료 검수 및 품질 관리 등

한국어 학습자 쓰기 평가 자료 구축	ㅇ 한국어교육 전문가가 평가한 쓰기 평가 자료(평가 점수, 평가 의견 등 포함) 3,000건 이상 구축
한국어 학습자 말뭉치 교육 및 홍보	ㅇ 한국어 학습자 말뭉치 이용자를 위한 아카데미 5회 이상 개최 ㅇ 한국어 학습자 말뭉치 구축 및 검수, 정비 인력 실무 교육 ㅇ 한국어 학습자 말뭉치 소개·활용 자료집 현행화 및 배포

2. 사업 추진 체계

2.1. 총괄 추진 체계

본 사업은 아래와 같은 수행 체계로 추진되었다.



<그림 1> 사업 추진 체계

2.2. 사업 추진 절차

본 사업은 아래와 같은 절차를 통해 추진되었다. 한국어 학습자 말뭉치 구축을 위해서는 수집, 구축, 가공의 단계를 거치는데, 각 단계마다 검수 과정이 포함된다. 2023년 사업에서는 문어 원시 말뭉치를 집중적으로 구축하였다. 쓰기 평가 자료 구축은 기초 연구, 평가 계획 수립, 평가 자료 구축의 단계로 진행되었다.

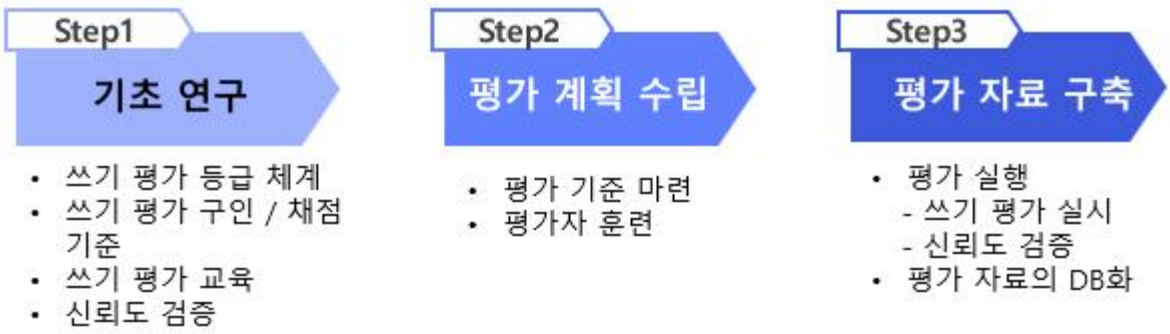
□ 총괄 추진 절차

<그림 2> 사업 추진 절차

□ 한국어 학습자 말뭉치 구축 절차

<그림 3> 한국어 학습자 말뭉치 구축 절차

□ 한국어 학습자 쓰기 평가 자료 구축 절차



<그림 4> 한국어 학습자 쓰기 평가 자료 구축 절차

3. 사업 추진 일정

구분		M	M+1	M+2	M+3	M+4	M+5	M+6	M+7	M+8	M+9
사전 분석 / 기획	기존 사업 분석	—									
	담당자/사용자 요구 사항 수집	—									
	일정 및 세부 내용 협의	—									
학습자 말뭉치 구축	업무 및 수집 절차 개선 사항 분석	—									
	학습자 선정 기준 설계	—									
	주제·제시 자료 구성 설계 및 확정	—									
	수집처 섭외		—	—	—	—	—	—	—		
	실무 작업자 교육		—	—							
	자료 수집 및 원시 말뭉치 구축		—	—	—	—	—	—	—	—	
	수집 진행 사항 관리		—	—	—	—	—	—	—	—	
	수집 데이터 기본 검증		—	—	—	—	—	—	—	—	
	데이터 가공 / 구축		—	—	—	—	—	—	—	—	
	메타 데이터 구축 / 정제		—	—	—	—	—	—	—	—	
	구축 데이터 검증 / 확인		—	—	—	—	—	—	—	—	
	원시 말뭉치 1·2차 납품 및 검수			▲			▲				
	검증 완료 데이터 평가 요청		—	—	—	—	—	—	—	—	
쓰기 평가 자료 구축	기초 연구	—	—								
	쓰기 평가자 모집	—									
	쓰기 평가 체계 마련		—								
	쓰기 평가자 훈련		—								
	쓰기 평가 도구 개발		—	—	—						
	쓰기 평가 수행 및 데이터 구축		—	—	—	—	—	—	—	—	
	쓰기 평가 자료 1·2차 납품 및 검수					▲			▲		
교육 및 홍보	교육 및 홍보 프로그램 기획	—									
	말뭉치 구축 실무자 교육		—								
	학습자 말뭉치 교실 개최 및 동영상 강좌 상시 운영			▲		▲	▲		▲		▲
	교육 자료 현행화 및 배포				▲		▲		▲		▲
사업 관리	산출물 관리			—	—	—	—	—	—	—	—
	착수, 중간, 완료 및 수시 보고	▲					▲				▲
	최종 납품 및 검수										▲

제 2 장

사업 수행 내용

1. 한국어 학습자 말뭉치 구축

본 사업에서는 특정 주제·제시 자료(5개 내외)에 대해 한국어 학습자가 작성한 한국어 쓰기 자료 15,000편 이상(표본 한 편당 100어절 이상, 약 150만 어절 이상)을 수집하였다. 특정 주제 또는 제시 자료에 편중되지 않도록 수집하였으며, 한 학습자당 최대 제출 가능 편수는 5편으로 제한하였다.

기구축 한국어 학습자 말뭉치의 분포를 고려하여 학습자 유형별, 수준별, 언어권별, 자료 변인별 균형성을 확보하는 것을 기본 방향으로 하였으며, 수집 시에는 한국어 학습자 대상 개인정보 수집·이용 동의, 쓰기 자료 대상 저작권 이용 허락(동의)을 필수적으로 체결하였다.

한국어 학습자 말뭉치 수집을 위한 방법 및 절차는 다음과 같다.

<그림 5> 한국어 학습자 말뭉치 수집 방법 및 절차

1.1. 설계

1.1.1. 수집 방향 설계

한국어 학습자 말뭉치의 균형성 확보를 위해 기구축 말뭉치의 수집 대상, 수준, 언어권, 자료 장르 등을 분석하여 2023년 사업의 자료 수집 및 구축 방향을 설계

하였다.

(1) 기구축 말뭉치 분석

2015년-2022년 사업에서 구축이 완료된 말뭉치 중 문어 말뭉치는 약 440만 어절, 37,320개의 표본으로 구성되어 있다. 본 사업에서는 말뭉치의 균형성 확보를 위해 기구축된 학습자 말뭉치의 변인별 분포를 분석하여 학습자 유형별, 수준별, 언어권별, 자료 변인별 말뭉치의 세부적인 목표 규모를 설정하기 위한 기초 자료로 활용하였다.

① 학습자 유형별

2015년-2022년 사업에서 문어 원시 말뭉치는 국내 교육 기관의 학습자 자료 35,739개(4,275,949어절), 이주민 학습자 자료 1,202개(90,648어절), 국외 학습자 자료 379개(51,041어절)로 구축되었다. 이주민 및 국외 학습자 자료의 비중이 현저하게 낮아 두 대상의 자료를 확대할 필요가 있음을 확인할 수 있다.

<표 2> 2015-2022년 한국어 학습자 말뭉치 학습자 유형별 통계 (문어. 원시)

수집 대상	표본 수	어절 수
일반	35,739	4,275,949
이주	1,202	90,648
국외	379	51,041
총합	**37,320**	**4,417,638**

② 수준별

수준별로는 1급-5급까지 비슷한 분포를 보인다. 어절 수로만 보면 1급의 어절 수가 부족하다고도 할 수 있으나, 표본 수로 보면 모든 급수 가운데 1급의 표본이 가장 많이 수집된 것을 확인할 수 있다. 이는 1급의 경우, 학습자의 어휘 산출량이 상위 등급에 비해 현저히 적기 때문이다. 수준별 통계를 보면 6급, 6급 이상의

자료 비중이 상대적으로 적으므로 이들 수준의 자료를 확대할 필요가 있다. 다만, 6급, 6급 이상의 고급 학습자는 다른 수준의 학습자에 비해 수적으로 적게 분포하기 때문에 학습자 수준은 전체적으로 고르게 수집하는 것을 목표로 하였다.

<표 3> 2015-2022년 한국어 학습자 말뭉치 수준별 통계 (문어. 원시)

수준(급수)	표본 수	어절 수
1급	8,725	593,191
2급	7,569	739,829
3급	6,779	825,358
4급	6,018	813,461
5급	5,024	788,991
6급	2,994	498,449
6급이상	117	135,335
정보없음	94	23,024
총합	**37,320**	**4,417,638**

③ 언어권별

언어권별로는 중국어권 학습자의 자료 비중이 가장 높으며 다음으로 일본어, 베트남어, 영어, 러시아어, 타이어, 몽골어, 스페인어 등의 순으로 자료가 분포되어 있었다. 이는 한국어 학습자의 구성비가 자료에 반영된 것이지만 학습자 말뭉치의 균형성과 자료의 활용도 제고의 측면에서 중국어권 외의 자료 비중을 확대해 나갈 필요가 있다.

<표 4> 2015-2022년 한국어 학습자 말뭉치 언어권별 통계 (문어. 원시)

모국어	표본 수	모국어	어절 수
중국어	16,403	중국어	2,040,734
일본어	5,347	일본어	650,143
베트남어	5,036	베트남어	520,031
영어	2,564	영어	284,665
러시아어	1,368	러시아어	148,775
타이어	978	타이어	123,330
몽골어	895	몽골어	96,332
스페인어	698	스페인어	76,890
프랑스어	427	인도네시아어	48,826
인도네시아어	397	프랑스어	47,374
스웨덴어	300	말레이어	29,856
아랍어	220	스웨덴어	27,717
말레이어	199	카자흐어	23,219
카자흐어	188	아랍어	21,398
타갈로그어	182	이탈리아어	21,308
독일어	176	독일어	19,587
우즈베크어	172	버마어	18,971
이탈리아어	151	우즈베크어	18,850
포르투갈어	150	싱할라어	18,256
기타	1,469	기타	181,376
총합	**37,320**	**총합**	**4,417,638**

④ 장르별

장르별로는 생활문의 비중이 현저하게 높으며, 다음으로 논설문의 비중이 높았다. 이는 한국어 교육 과정 및 일상적 문어 의사소통의 특성이 반영된 결과로 단순히 비중이 적은 장르의 구축 비율을 확대하기보다는 수준별 특성을 고려하여 이들의 비중을 균형 있게 확대해 나갈 필요가 있다.

<표 5> 2015-2022년 한국어 학습자 말뭉치 장르별 통계 (문어. 원시)

장르	표본 수	장르	어절 수
생활문	21,758	생활문	2,066,719
논설문	9,804	논설문	1,408,256
설명문	2,168	설명문	274,612
기행문	1,330	보고서	239,247
수필	582	기행문	140,613
편지글	497	수필	100,046
감상문	427	감상문	61,333
보고서	216	편지글	44,125
기사문	193	기사문	30,598
평전	109	평전	16,111
신문기사	87	투고문	12,146
투고문	61	신문기사	11,175
전기문	41	전기문	5,752
시	18	에세이	2,960
에세이	13	시	1,861
홍보문	10	홍보문	1,181
건의문	2	건의문	436
신문칼럼	2	신문칼럼	309
기타	1	기타	86
안내문	1	안내문	72
총합	**37,320**	**총합**	**4,417,638**

(2) 2023년 수집 및 구축 설계

① 수집 및 구축 목표

본 사업은 15,000개 표본, 약 150만 어절 규모(표본 1개당 평균 100어절 기준)의 학습자 쓰기 말뭉치를 수집하여 원시 말뭉치로 구축하는 것을 목표로 하였다.

<표 6> 2023년 문어 원시 말뭉치 구축 목표

기구축		**신규 구축**		합계	
표본 수	어절 수	**표본 수**	**어절 수**	표본 수	어절 수
37,320	약 440만	15,000	150만+α	52,320	590만+α

② 균형성 보완을 위한 자료 수집 및 구축 방향

2015년-2022년 사업의 말뭉치 구축 현황 분석을 토대로 말뭉치의 균형성 확보를 위한 자료 수집 및 구축의 방향은 다음과 같다.

<표 7> 2023년 문어 원시 말뭉치 구축 방향

구분		2015-2022 구축 현황	**2023 구축 전략**
학습자 변인	대상	○ 이주민, 국외 학습자 자료의 비중이 매우 적음	○ 이주민, 국외 학습자 자료의 비중 확대
	수준	○ 6급, 6급 이상의 비중이 매우 적음	○ 6급, 6급 이상의 자료 비중을 확대하되 전체적으로 모든 수준의 자료를 고르게 수집
	제1언어	○ 중국어권 자료의 비중이 가장 높고, 그 뒤를 잇는 일본어권, 베트남어권, 영어권, 러시아어권, 타이어권 등은 그에 비해 매우 적음	○ 중국어권 외 학습자의 자료 비중 집중 확대
자료 변인	장르	○ 생활문에 집중됨	○ 생활문 이외의 논설문, 설명문 등의 비중 확대

1.1.2. 주제 선정

2023년 사업에서는 특정 주제·제시 자료에 대해 한국어 학습자가 작성한 쓰기 자료를 수집하였다. 특정 주제에 대한 기획 수집으로 한정하여 자유 주제 수집은 이루어지지 않았으며, 각 교육 기관에 사전에 주제를 전달하고 해당 주제의 쓰기 자료를 전달받는 방식으로 수집을 진행하였다. 쓰기 주제는 수준별로 5개씩으로 제한하였으며, ①쓰기 평가를 위한 주제 선정 ②수준별 장르 분배 ③장르에 따른 주제 선정의 과정을 거쳤다.

우선 쓰기 평가 자료 구축을 위해 모든 수준의 학습자에게 공통되는 주제 2가지를 선정하였다. 쓰기 평가 자료의 대상이 되는 주제는 '자신의 나라와 한국 비교'와 '기후 위기와 해결 방안'으로 초·중·고급 학습자가 모두 해당 주제로 쓰기를 진행하였다. '자신의 나라와 한국 비교'는 모든 수준의 학습자가 쓰기에 적절한 주제

로, 수준에 따라 쓰기의 내용 및 구성, 어휘와 문법 등이 다양하게 출력될 수 있다. ‘기후 위기와 해결 방안’은 미디어를 활용하여 글쓰기 활동을 진행하는 주제로, 이를 통해 제시 자료의 다각화를 모색하였다. 학습자 수준에 따라 미디어 자료를 각각 다르게 제시하고, 학습자가 본인의 수준에 맞는 자료를 감상한 뒤 쓰기를 작성하도록 하였다.[1)]

기존 쓰기 자료의 장르가 생활문에 집중되어 있었으므로 올해 사업에서는 설명문과 논설문의 비중을 확대하였다. 쓰기 장르는 생활문, 설명문, 논설문으로 구성하되 학습자의 수준을 고려하여 수준별 개수를 조정하였다.

장르에 따른 주제를 선정할 때는 국내 대학 기관의 한국어 교재 및 기존 수집 주제를 분석하여 기획 수집에 적합한 주제로 선정하였다. 이렇게 선정된 최종 주제는 아래 표와 같다.

<표 8> 2023년 문어 원시 말뭉치 구축 주제 및 장르

수준	주제	장르
공통	자신의 나라와 한국 비교	설명문
	기후 위기와 해결 방안 (미디어 활용)	논설문
초급	내가 가장 좋아하는 것과 싫어하는 것	설명문
	내가 가 본 여행지	설명문
	올해의 계획	생활문
중급	과학 기술의 발전이 인간의 생활에 미치는 영향	논설문
	10년 후의 나의 계획	생활문
	효율적인 시간 관리 방법	설명문
고급	내가 생각하는 성공	논설문
	인구 문제	논설문
	역사적 사건이나 인물 소개	설명문

1) 다양한 자료를 제시할 수 있는 주제로 공통 주제를 선정하였으나 ‘기후 위기와 해결 방안’은 초급 학습자가 작성하기에는 다소 난도가 높은 주제였다. 향후 공통 주제를 선정할 시에는 이에 유념하여 모든 수준의 학습자가 작성 가능한가를 보다 우선적인 기준으로 고려해야 할 것이다. 제시 자료에 대한 구체적인 내용은 1.2.2. 수집 방식을 참고할 것.

1.1.3. 동의서 작성

글쓰기 자료를 수집하면서 학습자의 동의서 수집도 함께 이루어졌는데 동의서의 내용은 학습자 개인정보 수집·이용에 대한 동의, 저작권 이용 허락 계약 체결에 대한 것이었다. 2023년 사업에서는 학습자 동의서에 대한 법률 자문을 실시하고, 동의서의 내용을 보완하여 개인정보 보호 및 활용에 대한 효력을 보다 강화하고자 하였다.

개인정보 보호법에 따라 수집 목적, 수집 근거, 수집 항목, 보유 기간, 동의 여부, 제3자 제공, 활용 방안 등에 대하여 수집·이용 및 제3자 제공을 구분하고 항목별로 별도 동의를 받는 것으로 적법한 동의서 양식을 제작하였다. 그리고 해당 내용에 대한 학습자의 명확한 이해를 돕기 위해 한국어, 영어, 중국어, 일본어, 베트남어, 태국어, 스페인어, 러시아어, 몽골어, 인도네시아어 총 10개 언어로 동의서를 제공하였다.

학습자는 각 항목에 대한 동의 여부를 체크하고, 날짜를 기입한 뒤 서명을 하였다. 그리고 다음 페이지에 제시된 수집 항목에 대한 답변을 작성하였다. 만일 글쓰기 수집 과정에서 학습자가 개인정보 및 자료의 제공에 동의하지 않는 경우, 해당 쓰기 자료는 폐기하였다.

<표 9> 학습자 말뭉치 동의서의 조항

구분	동의서 내용
동의서 이름	한국어 학습자 말뭉치 구축·활용을 위한 학습자 언어 자료 이용 허락서, 개인정보 수집·이용 및 제3자 제공 동의서
사업 배경	국립국어원에서는 한국어교육의 질적 향상을 위해 한국어 학습자의 언어 자료(말뭉치)를 수집하여 활용하는 사업을 추진하고 있습니다.
이용 목적	여러분이 제공한 자료는 한국어 교수 방법 개선, 한국어 교재 개발, 한국어 교육 및 인공지능 모델 개발 등 다양한 분야의 연구·개발에 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 개인정보는 본 사업의 목적 외로는 사용되지 않으며, 비밀 유지를 위하여 식별할 수 없는 형태로 사용될 것입니다.
학습자 자료 이용 허락	한국어 학습자 말뭉치 구축·활용을 위해 쓰기 자료를 제공하고, 쓰기 원문 자료 전체의 공개와 연구·개발 목적의 이용을 허락합니다.
개인정보 수집·이용 내역	• 항목: 성별, 출생년, 현재 (TOPIK) 등급, 국적, 제1언어, 한국어 학습 기간, 한국 거주 기간, 한국어 학습 목적, 직업, 한국어 외 사용 가능 외국어 • 수집·이용 목적: 국립국어원 한국어 학습자 말뭉치 구축·활용 • 보유·이용 기간: 개인정보 수집·이용 동의일로부터 30년
개인정보 제3자 제공 내역	• 제공받는 자: 국립국어원 한국어 학습자 말뭉치 이용 약정자 • 제공 목적: 한국어 정보 처리 연구·개발, 국어 연구·개발 • 제공 항목: 성별, 출생년, 현재 (TOPIK) 등급, 국적, 제1언어, 한국어 학습 기간, 한국 거주 기간, 한국어 학습 목적, 직업, 한국 거주 기간, 한국어 학습 목적, 직업, 한국어 외 사용 가능 외국어 • 보유·이용기간: 개인정보 수집·이용 동의일로부터 30년
정보 폐기 안내	* 단, 개인정보 폐기 요청 시 즉시 삭제
동의 거부 안내	위의 개인정보 수집·이용에 대한 동의를 거부할 권리가 있습니다. 그러나 동의를 거부할 경우 한국어 학습자 말뭉치 구축·활용 사업 참여에 제한을 받을 수 있습니다.
개인정보 취급 업무 위탁 내역(고지 사항)	• 위탁 받는 자(수탁업체): ㈜이르테크·연세대학교 산학협력단 • 업무 내용: 2023년 한국어 학습자 말뭉치 구축

1.2. 수집

1.2.1. 수집 기관

본 사업에서는 5개 내외의 주제에 대해 작성한 쓰기 자료를 수집하여 기획 수집 말뭉치를 구축하였다. 교육 과정 외의 자료를 수집해야 한다는 점에서 수집 기관의 긴밀한 협조가 요구되므로 학계와 교육 기관의 수집 네트워크를 최대한 활용하였다.

<그림 6> 학습자 말뭉치 수집 네트워크

수집 대상은 국내 학습자, 이주민, 국외 학습자로 다양하게 구성하되 기존 말뭉치 구축량이 저조했던 이주민과 국외 학습자의 수집 비율을 확대하고자 하였다. 학습자 말뭉치 수집 대상과 수집 경로는 다음과 같다.

<표 10> 학습자 말뭉치 수집 경로

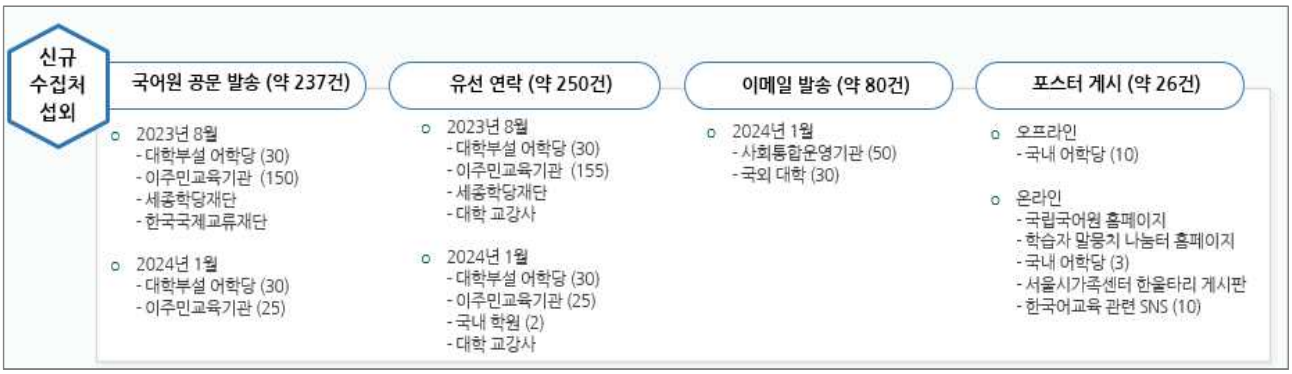
구분	수집 대상	자료 수집 경로
국내 학습자	어학 연수생	대학 부설 한국어교육 기관
	대학, 대학원 진학 유학생	국내 대학, 대학원
	기관 외 외국인	학습자 자율 기술

이주민	결혼 이민자	전국 다문화가족 지원센터
국외 학습자	세종학당 학생	세종학당
	국외 대학 학생	한국국제교류재단 교수 파견 대학 국외 한국어 교육자 네트워크
	기타 학습자	EPS-TOPIK 교육 기관 등

쓰기 자료 수집을 위해 대학 부설 한국어교육 기관, 다문화가족 지원센터, 세종학당재단, 한국국제교류재단(KF)에 협조 요청 공문을 발송하고, 유선 연락이 가능한 곳은 유선으로 한 번 더 협조를 요청하였다. 그리고 학습자가 개인적으로도 수집에 참여할 수 있도록 학습자 말뭉치 수집용 포스터를 제작하여 온라인과 오프라인을 통해 게시하였다.

<그림 7> 학습자 말뭉치 수집용 포스터

150만 어절 이상의 원시 말뭉치 구축을 위해서는 기존 수집처 이외 신규 수집처 섭외가 필수적으로 요구되었다. 신규 수집처 섭외 과정을 정리하면 아래와 같다.



<그림 8> 신규 수집처 섭외 과정

국내에서는 대학(원), 대학 부설 한국어 교육 기관, 이주민 교육 기관을 대상으로 수집을 진행하고, 국외에서는 한국국제교류재단 파견 교수가 재직 중인 대학(원), 그 외 대학(원), 세종학당, EPS-TOPIK 교육 기관 등을 대상으로 글쓰기 자료 수집을 진행하였다.

<표 11> 학습자 말뭉치 수집 참여 기관

구분		수집 참여 기관
국내 대학 부설 한국어 교육 기관 / 국내 대학(원)	기관 수집 (10곳)	건국대학교 언어교육원 경희대학교 국제교육원 계명대학교 한국어학당 배재대학교 한국어교육원 부산외국어대학교 한국어문화교육원 성균관대학교 한국어학당 한국외국어대학교 한국어문화교육원 한양대학교 국제교육원 호남대학교 한국어교육원 홍익대학교 국제언어교육원
	개별 수집 (11곳)	국민대학교 목원대학교 한국어교육원

구분		수집 참여 기관
		부산외국어대학교 삼육대학교 서울과학기술대학교 순천향대학교 연세대학교 우송대학교 충남대학교 한밭대학교 호서대학교
국내 이주민 교육 기관 (5곳)		경기 안산시 다문화가족지원센터 대구 서구 다문화가족지원센터 울산 남구 다문화가족지원센터 인천 동구 다문화가족지원센터 전남 여수시 다문화가족지원센터
국외 한국어 교육 기관	대학(원) (10곳)	슬로베니아 류블랴나대학교 이탈리아 시에나외국어대학교 중국 북경외국어대학교 중국 서안번역대학교 중국 연변대학교 중국 장춘인문대학교 중국 통화사범대학교 중국 흥의민족사범대학 태국 부라파대학교 폴란드 포즈난 아담미츠키에비츠 대학교
	한국국제 교류재단 (KF) (6곳)	베트남 외교아카데미 베트남 하노이국립외국어대학교 베트남 하노이산업대학 베트남 하노이인문사회과학대학교 슬로바키아 코메니우스대학교 인도네시아 가자마다대학교
	세종학당 (5곳)	세종학당 양성과정 방글라데시 세종학당 베트남 하노이1 세종학당

구분		수집 참여 기관
		이집트 세종학당 이탈리아 세종학당
	EPS-TOPIK 교육 기관 (2곳)	방글라데시 BKTTC (방글라데시 한국기술교육센터) 방글라데시 BOESL (방글라데시 해외고용서비스 공사)

국내에서 26곳, 국외에서 23곳, 총 49기관이 2023년 한국어 학습자 말뭉치 구축 사업에 참여하였다.

1.2.2. 수집 방식

2023년 사업에서는 한국어 학습자 말뭉치를 구축함에 있어 오프라인 수집과 온라인 수집을 병행하였다. 오프라인 수집은 기존의 수집 방식으로 교사가 학생들에게 글쓰기를 안내하고, 학생들의 쓰기 자료와 동의서를 전달받는 방식으로 진행되었다.

온라인 수집은 수행사 ㈜이르테크가 개발한 KOKOA 애플리케이션을 활용한 수집 방식으로, 학습자가 애플리케이션을 설치한 후 직접 글쓰기 자료를 입력하는 방식이다. 온라인 수집은 중간 경로(교사) 없이 학습자의 글쓰기 수집이 가능하며, 비대면으로 이루어진다는 점에서 국외 학습자 및 개인 학습자의 자료 수집에 용이하다. 이는 오프라인 수집의 한계점을 극복한 것으로 온라인 수집을 통해 말뭉치 규모의 확장성을 제고할 수 있다. 또한 학습자가 도구를 활용해 자료를 산출하는 동시에 자료가 저장되므로 별도의 입력 작업이 필요하지 않다는 점에서 작업을 간소화시키고, 데이터 관리의 효율성을 향상시킨다. 별도의 입력 작업이 필요 없으므로 입력 단계에서 발생할 수 있는 작업자 오류도 방지할 수 있다. 온라인 수집은 멀티미디어 자료를 제공할 수 있다는 점에서 다양한 자료를 통한 쓰기 자료 수집에도 기여한다.

2023년 사업에서는 '기후 위기와 해결 방안' 주제를 온라인으로 수집하였으며, 학습자를 위한 멀티미디어 자료로 초급은 자체 제작 이미지, 중급은 자체 제작 영상, 고급은 뉴스 영상을 제공하였다. 학습자는 이미지 또는 영상을 감상한 뒤 휴대폰의 자판을 이용해 직접 글을 입력하였다. 다만, 초급 학습자는 한글 입력이

어려울 수 있으므로 초급 학습자에 한해, 종이에 글을 쓰고 사진을 찍은 뒤 사진을 업로드할 수 있도록 하였다.

온라인 글쓰기 진입 및 제출 과정

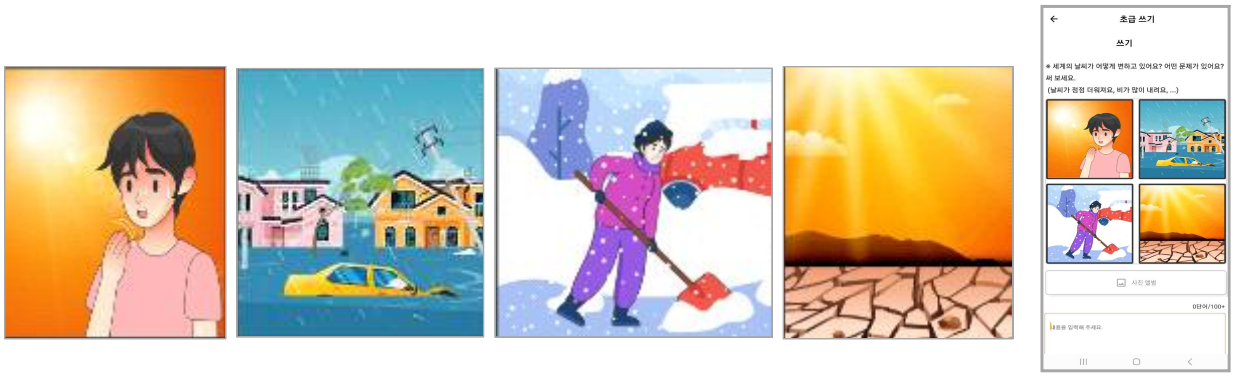
<그림 9> 온라인 글쓰기 제출 과정

초급 학습자는 제시된 그림을 보고 글을 작성할 수 있도록 기후 위기로 인해 세계 각국에서 발생하는 이상 기후 현상을 여러 그림으로 제시하였다. 그리고 세계

날씨 변화에 대한 몇 가지 예문을 함께 제공하여 학습자의 이해를 돕고자 하였다. 중급 학습자는 영상을 확인한 뒤 기후 위기를 이해하고 글을 쓸 수 있도록 하였는데 영상은 학습자의 수준에 알맞게 수행사에서 자체적으로 제작하고 자막을 삽입하였다. 고급 학습자는 기후 위기에 관한 뉴스를 보고[2] 글을 쓸 수 있도록 하였다. 온라인 수집에서 사용된 수준별 멀티미디어 자료는 다음과 같다.

2) 뉴스 영상 활용을 위해 법률 자문을 실시하고 자문 내용에 따라 영상을 제공하였다. 저작권에 문제가 되지 않도록 영상에 대한 무단 편집을 하지 않고, 출처 및 목적을 밝힌 뒤 뉴스 영상을 이용하였다.

초급 (자제 제작 이미지)

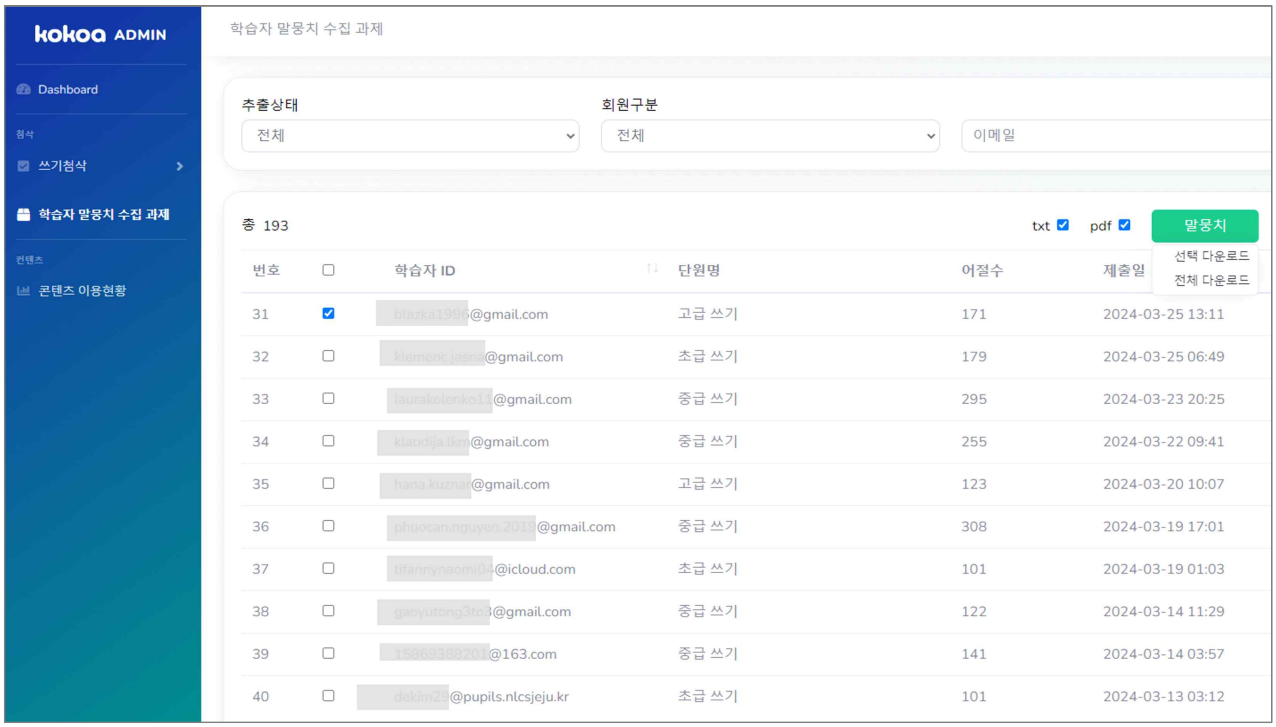


중급 (자체 제작 영상)

고급 (뉴스 영상)

<그림 10> 온라인 수집에 사용된 수준별 제시 자료

수행사는 온라인 수집을 위해 학습자들이 앱에서 제출한 쓰기 자료를 확인할 수 있는 전용 웹페이지를 구축하였다. 웹페이지에서는 학습자 ID, 학습자 수준, 제출 일시, 어절 수 등의 정보를 한눈에 확인할 수 있으며, 자료를 선택하여 txt 또는 pdf 파일로 다운로드 받을 수 있다.



<그림 11> 온라인 수집을 위한 웹 화면

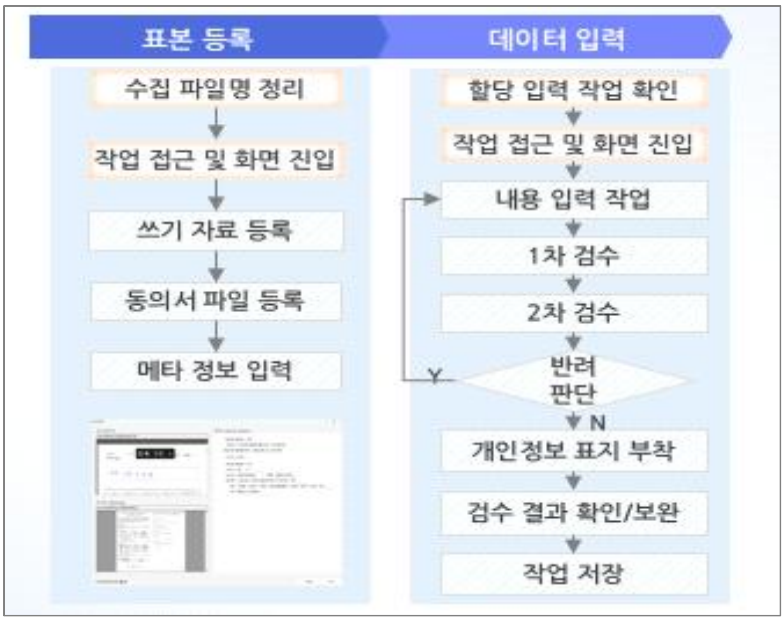
<그림 12> 온라인으로 수집한 쓰기 자료 샘플

온라인으로 수집한 쓰기 자료는 오프라인 수집 자료와 마찬가지로 한국어 학습자 말뭉치 구축 지원 시스템(LCMS)에 업로드하여 말뭉치로 구축하였다. 온라인 수집의 경우, 동의서는 이메일로 별도 제출하도록 하고 이메일 주소를 기준으로 쓰기 자료-동의서 짝을 맞춰 관리하였다.[3)]

1.3. 구축

1.3.1. 말뭉치 구축

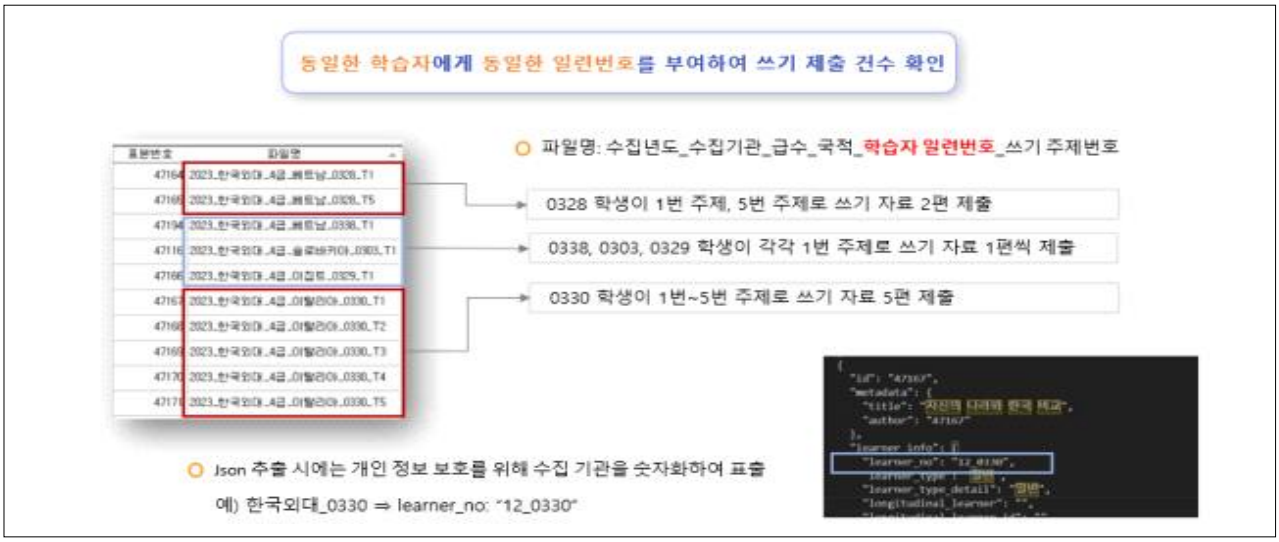
수집한 쓰기 자료는 한국어 학습자 말뭉치 구축 지원 시스템(LCMS)을 이용하여 메타 정보를 입력하고 표지를 부착하여 원시 말뭉치 형태로 구축하였다. 학습자 말뭉치 구축 작업 절차를 도식으로 나타내면 아래 그림과 같다.



<그림 13> 학습자 말뭉치 구축 작업 절차

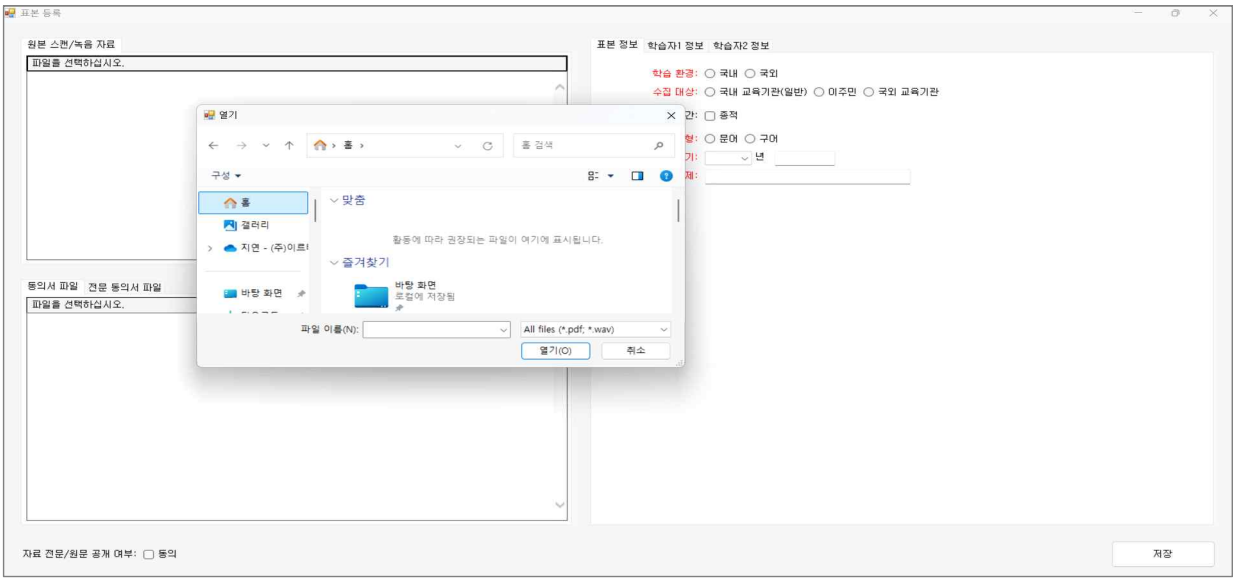
3) 온라인으로 수집된 쓰기 자료는 오프라인으로 수집된 자료에 비해 수집량이 현저히 저조하였다. 그 원인은 크게 세 가지 정도로 추정되는데 첫째, 기관의 애플리케이션 소개 기피, 둘째, 외국인 학습자의 한글 타이핑 미숙, 셋째 쓰기 자료와 동의서의 동시 제출 불가이다. 향후 동의서와 쓰기 자료 제출이 일괄 가능하도록 페이지를 구성하고, 중고급 학습자를 대상으로 수집 기관의 적극적 협조 아래 온라인 수집을 진행한다면 온라인 수집량을 확대할 수 있을 것이다. 또한 온라인으로 수집된 자료는 주제 무관, 어절 수 미준수, 한국어 이외 텍스트 작성 등 다양한 불량 요인을 포함하고 있었으므로 온라인 제출이 가능하도록 하는 제약 조건을 보다 강화해야 할 것이다.

수집한 쓰기 자료를 원시 말뭉치로 구축하기 위해서는 표본 등록과 입력, 검수 과정을 거친다. 표본 등록을 위해서는 가장 먼저 쓰기 자료와 동의서의 짝을 맞춰 파일명을 정리하는 작업이 필요하다. 한 학습자당 최대 제출 가능 편수를 5편으로 제한하였으므로 파일명 입력 시 동일한 학습자에게 동일한 일련번호를 부여하여 쓰기 제출 건수를 한눈에 확인할 수 있도록 하였다.



<그림 14> 학습자별 동일한 일련번호 부여

파일 및 파일명 정리가 완료되면 한국어 학습자 말뭉치 구축 지원 시스템(LCMS)에 메타 정보와 함께 표본을 등록한다.

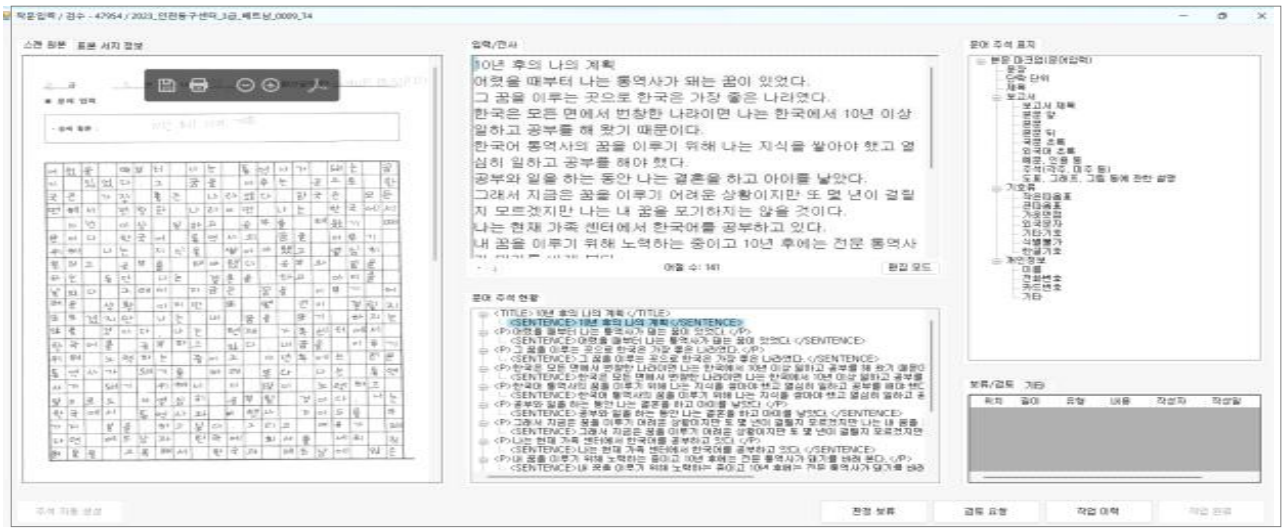


<그림 15> LCMS의 표본 등록 화면

메타 정보에는 '자료 유형, 장르, 주제, 수집 시기, 수집 기관' 등의 표본 정보와 '학습자 유형, 국적, 성별, 나이, 한국어 수준, TOPIK 등급, 재외동포 여부, 한국어 학습 기간, 한국 거주 기간, 학습 목표, 직업, 모국어, 기타 구사 언어' 등의 학습자 정보가 포함된다.

<그림 16> LCMS의 표본 메타 정보 및 학습자 메타 정보

표본을 등록한 후에는 학습자의 쓰기 자료를 보고 내용을 입력한 뒤 주석 작업을 진행하였다. 문장 단위를 구분하고 각종 기호 등에 주석 처리를 하였는데 학습자의 이름, 외국인 등록번호, 주소, 전화번호 등의 개인정보에는 개인정보 표지를 부착하여 비식별 처리하였다.

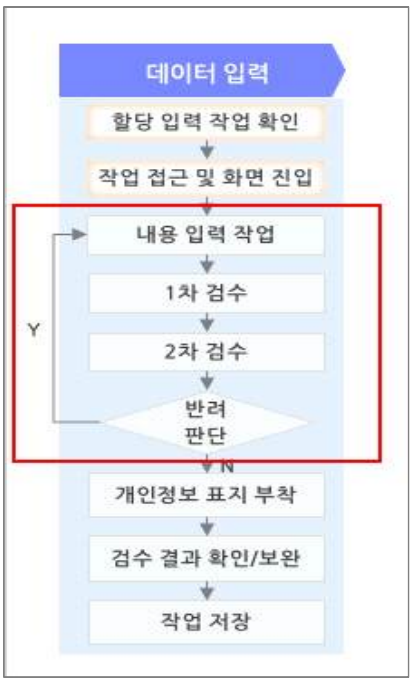


<그림 17> LCMS의 입력 작업 화면

1.3.2. 말뭉치 검수

학습자 말뭉치 데이터의 정확성 검증을 위해 내용 검수, 표본 정보 검수, 자동 주석 검수를 진행하였다.

내용 검수는 입력 데이터 및 주석 내용에 오류가 없는지를 살피는 수동 검수로 작업자, 검수자 간 최소 2단계의 100% 검수를 진행하였다. 초기 검수 후에는 작업자에게 오류 내용에 대한 피드백을 전달하여 동일 오류 발생을 방지하고자 하였으며, 지속적으로 오류를 발생하는 경우 작업을 반려하고 별도의 작업자 교육을 실시하였다.



<그림 18> 내용 검수 단계

<그림 19> 작업자의 오류 피드백 교육 자료

입력자 혹은 검수자가 표본 입력/검수 시, 중복 표본임이 의심되는 표본은 즉각 관리자에게 알리도록 하였다. 그 결과, 동일 기관에서 중복 표본으로 의심되는 사례가 몇 차례 발견되어 해당 기관의 구축 표본을 대상으로 중복 표본 검사를 실시하고, 중복 의심 표본 41건을 찾아 표본 목록에서 삭제하였다.[4)]

표본 정보 검수는 표본명과 동의서명의 일치 여부, 쓰기 자료의 메타 정보 등을 검수하는 과정이다. 표본명과 동의서명이 불일치하는 경우, 자료의 세트가 맞는지 확인하고 파일 수정 혹은 파일명 수정을 통해 쓰기 자료와 동의서 간 이상 표본이 없도록 하였다. 2023년 사업에서는 문어 자료를 정해진 주제로 수집하였으므로 자료 유형, 작문 유형, 장르, 주제가 모두 일관성을 갖추고 있다. 따라서 자료 유형이 문어인지, 작문 유형이 기획 작문인지 확인하고, 장르와 주제가 수준별로 제시된 다섯 가지 주제에 부합하는지 검수하여 데이터 일관성을 제고하였다. 이 밖에도 수집 기관 정보를 전수 확인하여 기관명이 다르거나 미수집 기관으로 제시되어 있는 경우, 해당 수집 기관으로 기관명을 수정하였다.

데이터 정합성을 높이기 위해 주석 작업의 수동 검수에 앞서 자동화 검수를 먼저 진행하였다. 문장 주석 생성이 안 된 표본, <p>와 <sentence> 주석이 빈 값이 표본, <p> 아래 <sentence>가 복수로 존재하는 표본 등이 검수 대상이 되었다.

<그림 20> 데이터 정합성을 위한 자동화 검수 진행

4) 해당 기관은 방글라데시의 BOESL(방글라데시 해외고용서비스 공사)로 근로자들을 대상으로 EPS-TOPIK(고용허가제 한국어능력시험)을 교육하는 기관이다. 2024년 3월까지 구축된 쓰기 자료 618건을 대상으로 90% 이상 일치도 검사를 진행하였으며, 검사 결과 41건이 중복 의심 표본으로 확인되어 해당 41건을 목록에서 삭제하였다. 국내 인구 감소 문제로 외국인 고용허가제(EPS)가 지속적으로 확대되고 있어 EPS-TOPIK 학습자의 규모가 점점 증가하고 있으므로 보다 적극적인 자료 수집 전략이 필요하다. 향후 사업에서는 EPS-TOPIK 학습자 유형에 대한 메타 정보를 좀 더 구체화하여 학습자별 자료 특성을 구분할 필요가 있다.

오류가 발견되면 수동으로 해당 내용을 수정한 이후, 다시 한 번 자동 검수를 실시하여 오류가 없는지 재확인하였다.

올해 사업에서는 원시말뭉치 구축만 진행하고 형태 주석과 오류 주석 작업은 진행하지 않았기 때문에 작업자의 실수로 형태 주석 및 오류 주석으로 작업이 할당되어 이력이 남은 경우, 시스템에서 해당 이력을 삭제하여 통계 결과 오류를 방지하였다.

1.3.3. 구축 결과

2023년 사업에서는 문어 원시 말뭉치 15,000편 이상, 150만 어절 이상을 구축하는 것을 목표로 하였다. 이에 수행사는 총 15,779개 표본, 2,085,629어절의 문어 원시 말뭉치를 구축하였다. 표본 하나당 평균 어절 수는 132어절이다.[5] 2015년-2022년 구축된 문어 원시 말뭉치가 총 4,417,638어절이므로 기존에 구축된 말뭉치 규모의 약 47%에 달하는 말뭉치를 2023년 사업에서 수집·구축한 것이라고 볼 수 있다. 2022년까지의 사업에서는 원시 말뭉치뿐 아니라 형태 주석 말뭉치, 오류 주석 말뭉치를 함께 구축하였으나 2023년 사업에서는 원시 말뭉치만 구축 대상으로 하였으므로 구축 목표량이 상당히 증가하였다.

2015년부터 2023년까지 구축된 문어 원시 말뭉치를 모두 합산하면 총 53,099개 표본, 6,503,267어절이다.

다음에서는 학습자 변인(유형, 수준, 언어권)과 자료 변인(장르, 주제)을 기준으로 (1) 2023년에 구축된 원시 말뭉치가 1.1.에서 제시한 설계 방향에 맞게 구축되었는지 말뭉치의 통계를 분석하고, (2) 2015년-2023년에 구축된 문어 원시 말뭉치의 전체 통계를 살피도록 하겠다.

---

5) 본 사업은 100어절 이상의 표본을 수집하며, 표본 1편의 어절 수가 500어절 이상일 경우 표본 2편으로 산정하는 원칙을 준수하였다. 하지만 수집·구축 과정에서, 초급 학습자의 경우 산출할 수 있는 어휘가 한정적이라는 점에서 100어절 이상의 글을 쓰기에는 무리가 있다는 의견이 기관 전문가에 의해 제시되었다. 이에 과업 변경을 통해 초급 학습자가 100어절 미만의 글을 제출했을 경우, 어절 수를 합산하여 100어절당 1편으로 편수를 산정하고, 중고급 학습자가 100어절 미만의 글을 작성했을 경우, 70어절 이상 표본의 어절 수만 합산하여 140어절당 1편으로 편수를 산정하도록 하였다. 결국 총 15,779편의 쓰기 자료를 구축하여 납품하였지만, 이러한 산정 결과에 따라 15,021편을 납품한 것으로 산정되었다. 본 보고서의 통계는 LCMS에 업로드된 표본 수에 따라 결과가 산출되므로 실제 구축된 15,779편을 대상으로 통계 결과를 보고하겠다.

1) 학습자 유형

기구축 문어 원시 말뭉치에서는 이주민과 국외 학습자의 비중이 현저히 낮았으므로 수집 과정에서 이들의 자료 비중을 확대하여 대상별 균형성이 확보될 수 있도록 하였다. 특히 국외 학습자의 자료 비중을 집중적으로 확대하였다.

<표 12> 2015년-2022년 말뭉치 구축 현황 - 학습자 유형

학습자 유형	어절 수	비율	표본 수	비율
일반	4,275,949	96.8	35,739	95.8
이주	90,648	2.1	1,202	3.2
국외	**51,041**	**1.2**	**379**	**1.0**
총합	4,417,638	100.0	37,320	100.0

<표 13> 2023년 말뭉치 구축 현황 - 학습자 유형

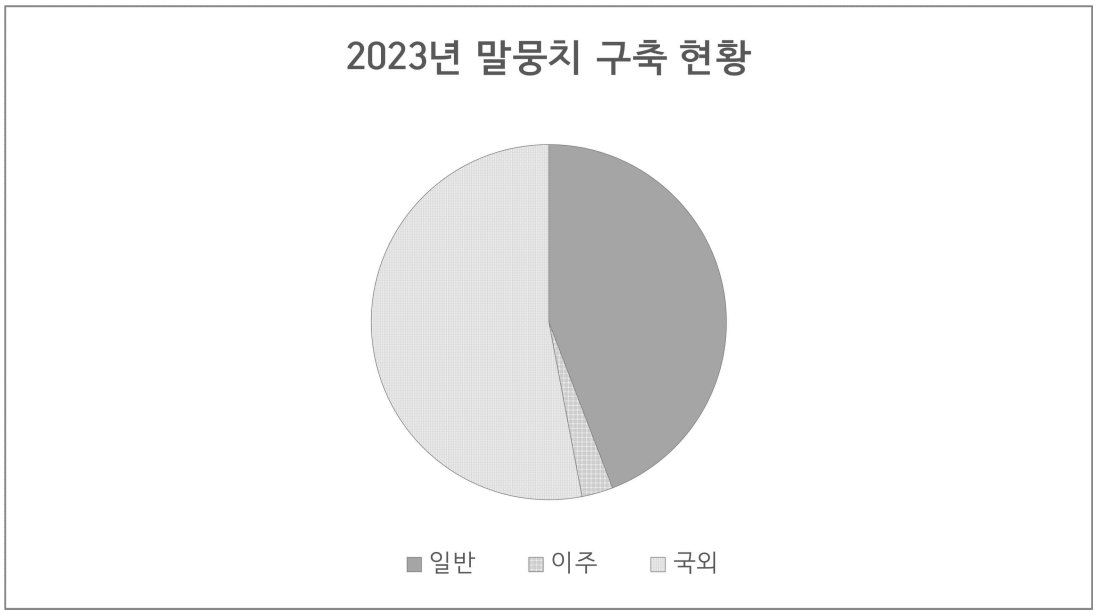
학습자 유형	어절 수	비율	표본 수	비율
일반	921,106	44.2	7,643	48.4
이주	59,335	2.8	398	2.5
국외	**1,105,188**	**53.0**	**7,738**	**49.0**
총합	2,085,629	100.0	15,779	100.0

<표12>의 2015년-2022년 말뭉치 구축 현황을 보면 일반 학습자의 비중이 전체 말뭉치 비중의 96.8%를 차지하고 있다. 이에 2023년 말뭉치에서는 일반 학습자 44.2%, 국외 학습자 53%로 국외 학습자의 비중을 대폭 확대하였다. 이주민 자료는 2.8%로 그동안 수집했던 말뭉치와 비교했을 때 그 비중은 비슷하나 어절 수로만 보면 2015년-2022년에 수집했던 말뭉치 규모의 65.5%를 차지하는 양이므로 결코 적다고는 할 수 없다.

말뭉치 사업을 설계하고 시작할 당시 국내 기관이 우선적으로 수집 대상이 되었으므로 국내에서 국외로의 수집 기관 이동은 자연스러운 현상이라고 할 수 있다. 2023년의 국외 자료 수집량은 기존 국외 수집량의 약 2배 가까운 수치이다. 구축

된 표본 수를 비교하면 일반 학습자와 국외 학습자가 각각 48.5%, 49%로 거의 차이가 없으나, 국내 일반 초급 학습자가 100어절 미만의 작문을 산출하는 경우가 다수 존재하여 어절 수에서는 국외 학습자와 9%가량 차이가 발생하였다.[6)]

학습자 말뭉치의 구축 현황을 참고하면 2023년 사업에서 국외 학습자의 쓰기 자료 비중을 좀 더 확대해야 좋을 것이나, 2023년 사업에서는 학습자 말뭉치 구축뿐 아니라 구축된 말뭉치를 대상으로 학습자의 쓰기 자료를 평가하는 과업을 함께 진행하였으므로 원활한 평가 자료 구축을 위해 국외 학습자와 국내 학습자의 쓰기 자료를 유사한 비중으로 수집하였다.



<그림 21> 2023년 말뭉치의 학습자 유형별 비중 (어절 수 기준)

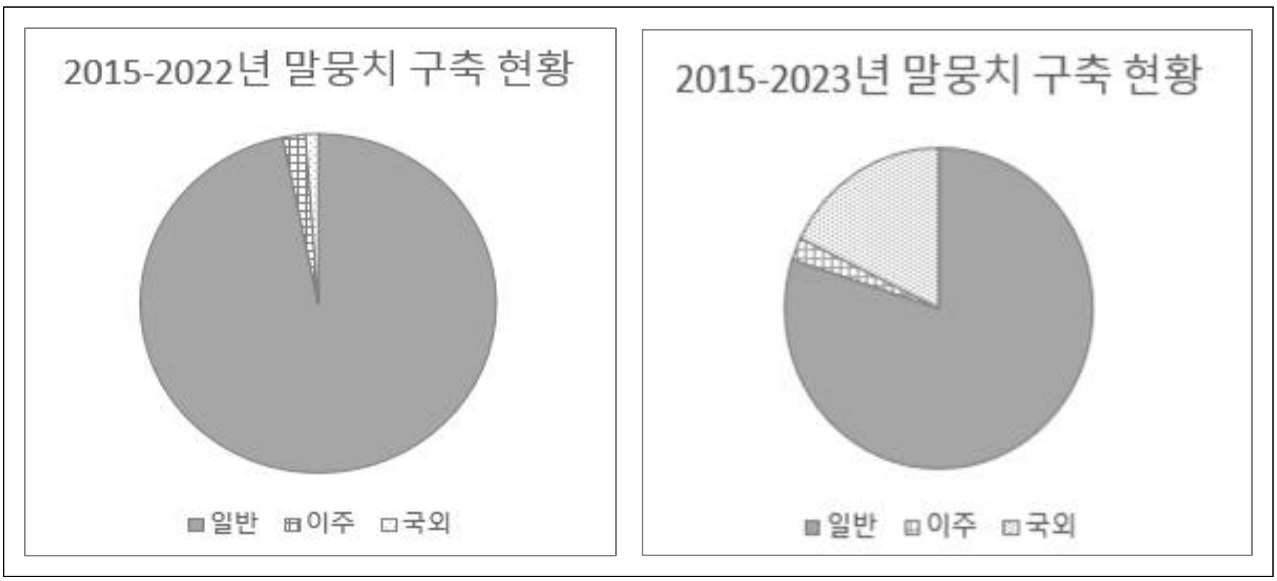
2023년에 국외 학습자의 말뭉치 구축량이 대폭 증가하면서 2015년-2023년 전체 말뭉치에서 학습자 유형별 현황은 아래와 같이 변화하였다.

6) 현재 LCMS의 학습자 수준 분류 체계는 국제 통용 한국어 표준 교육과정에서 제시하는 1급-6급까지의 분류 체계를 따르고 있다. 국내 한국어교육 기관에서 학습하는 학습자의 경우, 현재 소속되어 있는 반의 수준에 따라 급수 체계를 비교적 명확하게 제공할 수 있으나 국외 학습 기관에서 학습하는 학습자의 경우, 이러한 급수 체계를 명확하게 제공할 수 없는 경우가 적지 않았다. 이 경우, 담당 선생님이 학습자의 수준을 파악한 뒤 적절한 급수를 지정해 주거나 학년에 따라 급수를 지정해서 학습자 메타 정보를 입력하였다. 하지만 국외 초급 학습자의 경우, 국내 초급 학습자보다 100어절 이상 글 쓰기를 더 잘 해냈다는 점을 미루어 짐작할 때, 국내 학습자 1급과 국외 학습자 1급이 동일한 수준인지에 대한 추가 연구가 필요하다고 판단된다. 국외 학습자의 참여 비중이 높아진 만큼 향후 사업에서는 메타 정보 추가를 통해 국내와 국외 학습자의 수준 구분, 학습목적 학습자와 비학습목적 학습자의 구분 등의 메타 정보 확대를 검토할 필요가 있다.

<표 14> 2015년-2023년 말뭉치 구축 현황 - 학습자 유형

학습자 유형	어절 수	비율	표본 수	비율
일반	5,197,055	79.9	43,382	81.7
이주	149,983	2.3	1,600	3.0
국외	**1,156,229**	**17.8**	**8,117**	**15.3**
총합	6,503,267	100.0	53,099	100.0

구축된 말뭉치에서 학습자 유형별 비중이 변화하는 양상을 한눈에 비교할 수 있도록 도표로 나타내면 다음과 같다.



<그림 22> 학습자 유형별 비중의 변화 양상 (어절 수 기준)

2) 학습자 수준

기구축 문어 원시 말뭉치에서는 6급, 6급 이상의 고급 학습자의 자료 비중이 상대적으로 적었으나 이들 고급 학습자는 다른 수준의 학습자에 비해 수적으로 적게 분포하기 때문에 학습자 수준은 전체적으로 고르게 수집하는 것을 목표로 하였다.

<표 15> 2015년-2022년 말뭉치 구축 현황 - 학습자 수준

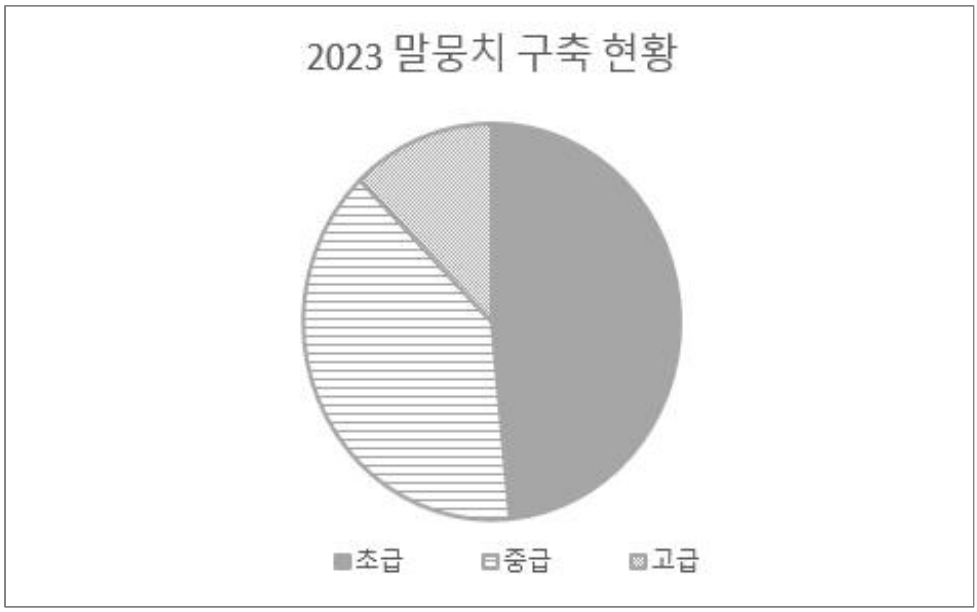
수준(급수)	어절 수	비율	표본 수	비율
1급	593,191	13.4	8,725	23.4
2급	739,829	16.7	7,569	20.3
3급	825,358	18.7	6,779	18.2
4급	813,461	18.4	6,018	16.1
5급	788,991	17.9	5,024	13.5
6급	498,449	11.3	2,994	8.0
6급이상	135,335	3.1	117	0.3
정보없음	23,024	0.5	94	0.3
총합	4,417,638	100.0	37,320	100.0

<표 16> 2023년 말뭉치 구축 현황 - 학습자 수준

수준(급수)	어절 수	비율	표본 수	비율
1급	405,732	19.5	3,491	22.1
2급	595,638	28.6	4,874	30.9
3급	491,687	23.6	3,542	22.4
4급	318,674	15.3	2,117	13.4
5급	166,445	8.0	1,074	6.8
6급	88,110	4.2	542	3.4
정보없음	19,343	0.9	139	0.9
총합	2,085,629	100.0	15,779	100.0

<표15>의 2015년-2022년 말뭉치 구축 현황에서 어절 수를 기준으로 보면 학습자 수준별로 거의 고른 비율을 나타내고 있다. 실제 구축된 표본 수를 보면 1급 표본 개수와 6급 표본 개수에서 15.4%의 차이를 보이지만, 어절 수를 기준으로 하면 2.1%밖에 차이가 나지 않는다. 이것은 고급 학습자일수록 산출 가능한 어휘 수가 증가하기 때문에 적은 표본이라도 많은 어절 수를 포함하기 때문이다.

2023년에 구축된 말뭉치는 초급 학습자의 비중이 가장 높고, 고급 학습자의 비중이 낮다.



<그림 23> 2023년 말뭉치의 학습자 수준별 비중 (어절 수 기준)

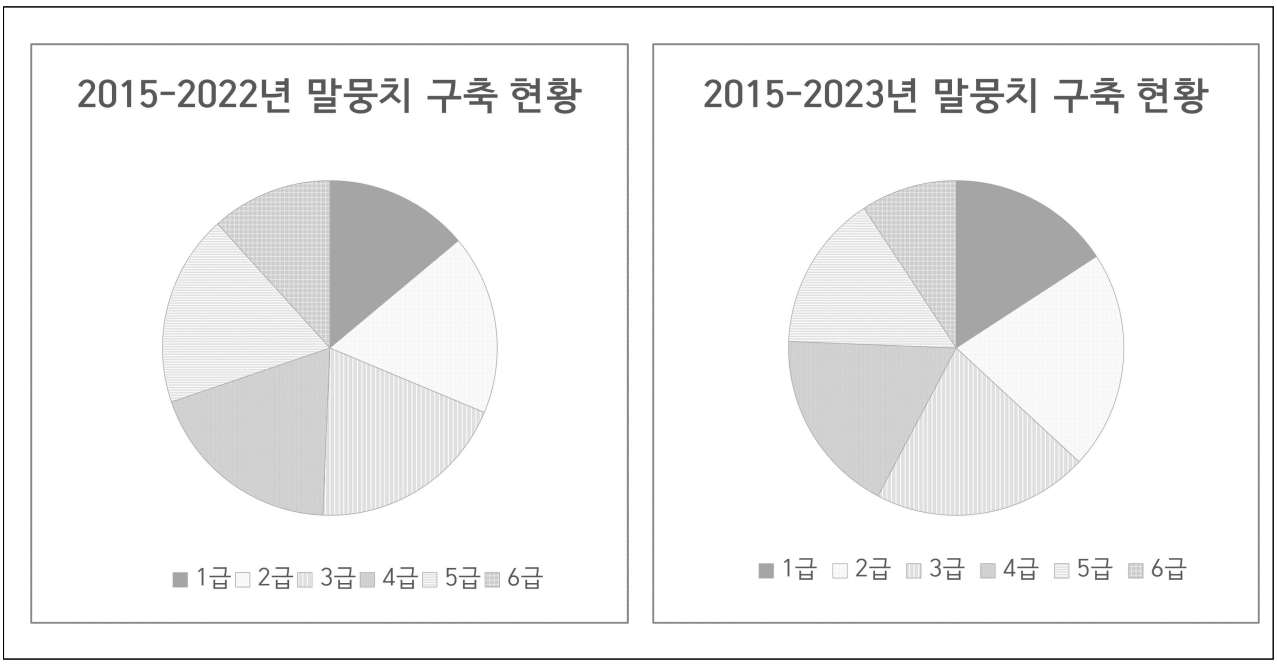
2023년 구축 말뭉치에서 고급 학습자의 비중이 현저히 줄어든 것은 국외 학습자의 비중이 증가했기 때문이다. 국외 학습자는 주로 초중급 학습자로 구성되어 있어 고급 학습자의 자료를 수집하기에 어려움이 있었다. 국내 학습자의 경우, 대학 학부생이나 대학원생의 제출 보고서를 통해 고급 학습자의 글쓰기 자료 수집 및 어절 수 확보가 가능할 것으로 기대했으나, 챗GPT 등 거대 언어 모델의 등장으로 보고서 제출 과제가 시험이나 발표 등으로 대체됨에 따라 글쓰기 자료 수집이 원만히 진행되지 않았다.

2023년 구축 말뭉치에 국외 학습자의 비중을 늘린 결과 학습자의 수준별 자료가 고르게 수집되지는 못했으나, 2015년-2023년 전체 말뭉치에서 학습자 수준별 현황은 전년 대비 크게 달라지지 않았다.

<표 17> 2015년-2023년 말뭉치 구축 현황 - 학습자 수준

수준(급수)	어절 수	비율	표본 수	비율
1급	998,923	15.4	12,216	23.0
2급	1,335,467	20.5	12,443	23.4
3급	1,317,045	20.3	10,321	19.4
4급	1,132,135	17.4	8,135	15.3
5급	955,436	14.7	6,098	11.5
6급	586,559	9.0	3,536	6.7
6급이상	135,335	2.1	117	0.2
정보없음	42,367	0.7	233	0.4
총합	6,503,267	100.0	53,099	100.0

구축된 말뭉치에서 학습자 수준별 비중이 변화하는 양상을 한눈에 비교할 수 있도록 도표로 나타내면 다음과 같다. 초급 학습자의 수가 늘고 고급 학습자의 수가 줄었지만 전체적으로 큰 폭의 변화는 없었다.



<그림 24> 학습자 수준별 비중의 변화 양상 (어절 수 기준)

3) 학습자 언어권

기구축 문어 원시 말뭉치에서는 중국어권 학습자의 비중이 월등히 높았으므로 중국어권 이외의 자료 비중을 확대 수집하는 것을 목표로 하였다.

<표 18> 2015년-2022년 말뭉치 구축 현황 - 학습자 언어권

순위	모국어	어절 수	비율	표본 수	비율
1	중국어	2,040,734	46.2	16,403	44.0
2	일본어	650,143	14.7	5,347	14.3
3	베트남어	520,031	11.8	5,036	13.5
4	영어	284,665	6.4	2,564	6.9
5	러시아어	148,775	3.4	1,368	3.7
6	타이어	123,330	2.8	978	2.6
7	몽골어	96,332	2.2	895	2.4
8	스페인어	76,890	1.7	698	1.9
9	인도네시아어	48,826	1.1	397	1.1
10	프랑스어	47,374	1.1	427	1.1
11	말레이어	29,856	0.7	199	0.5
12	스웨덴어	27,717	0.6	300	0.8
13	카자흐어	23,219	0.5	188	0.5
14	아랍어	21,398	0.5	220	0.6
15	이탈리아어	21,308	0.5	151	0.4
16	독일어	19,587	0.4	176	0.5
17	버마어	18,971	0.4	133	0.4
18	우즈베크어	18,850	0.4	172	0.5
19	싱할라어	18,256	0.4	109	0.3
20	기타	181,376	4.1	1,559	4.2
총합		4,417,638	100.0	37,320	100.0

<표 19> 2023년 말뭉치 구축 현황 - 학습자 언어권

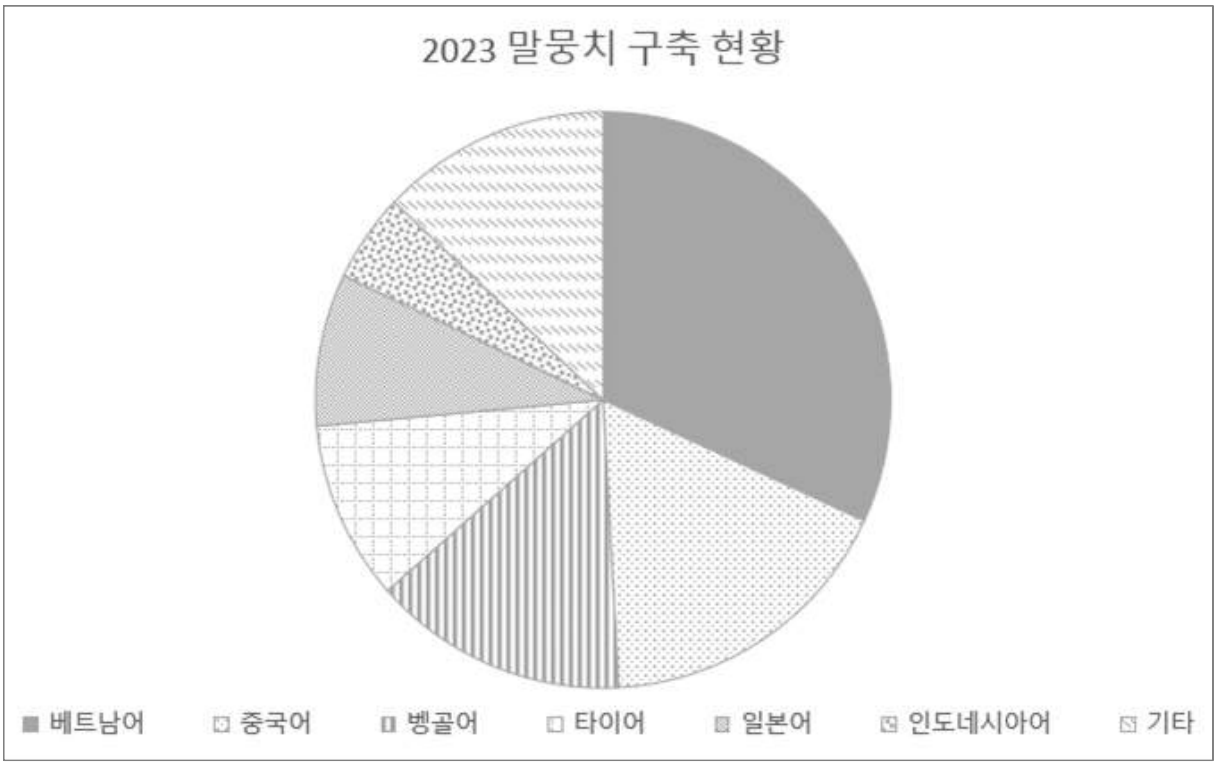
순위	모국어	어절 수	비율	표본 수	비율
1	베트남어	665,147	31.89	5,279	33.46
2	중국어	359,887	17.26	3,035	19.23
3	벵골어	298,914	14.33	2,455	15.56

4	타이어	210,070	10.07	908	5.75
5	일본어	178,526	8.56	1,416	8.97
6	인도네시아어	103,311	4.95	628	3.98
7	몽골어	39,434	1.89	301	1.91
8	러시아어	29,753	1.43	216	1.37
9	영어	22,284	1.07	181	1.15
10	자바어	16,286	0.78	92	0.58
11	버마어	15,633	0.75	109	0.69
12	프랑스어	15,006	0.72	130	0.82
13	아랍어	13,417	0.64	120	0.76
14	이탈리아어	12,879	0.62	120	0.76
15	키르기스어	11,066	0.53	78	0.49
16	스페인어	10,187	0.49	85	0.54
17	슬로베니아어	10,039	0.48	58	0.37
18	우즈베크어	7,919	0.38	57	0.36
19	슬로바키아어	5,908	0.28	47	0.30
20	기타	59,963	2.88	464	2.94
총합		2,085,629	100.0	15,779	100.0

<표18>의 2015년-2022년 말뭉치 구축 현황을 보면 중국어권 학습자의 비중이 46.2%로 거의 절반가량을 차지하고, 일본어 > 베트남어 > 영어 > 러시아어 > 타이어의 순서로 그 뒤를 잇고 있다.

이에 2023년 말뭉치 구축에서는 중국어권 이외의 자료 비중을 확대하고자 하였으며 그 결과, 베트남어권 학습자의 자료가 약 32%로 가장 높은 비중을 차지하였다. 그 뒤로는 중국어 약 17%, 벵골어 약 14%, 타이어 약 10%, 일본어 약 9%, 인도네시아어 약 5%로, 한 언어권이 절대적인 비중을 차지하지 않도록 자료를 수집 및 구축하였다.[7)]

7) 해당 순위는 어절 수를 기준으로 한 것으로, 표본 수를 기준으로 한 것과는 다소 차이가 있다.



<그림 25> 2023년 말뭉치의 학습자 언어권별 비중 (어절 수 기준)

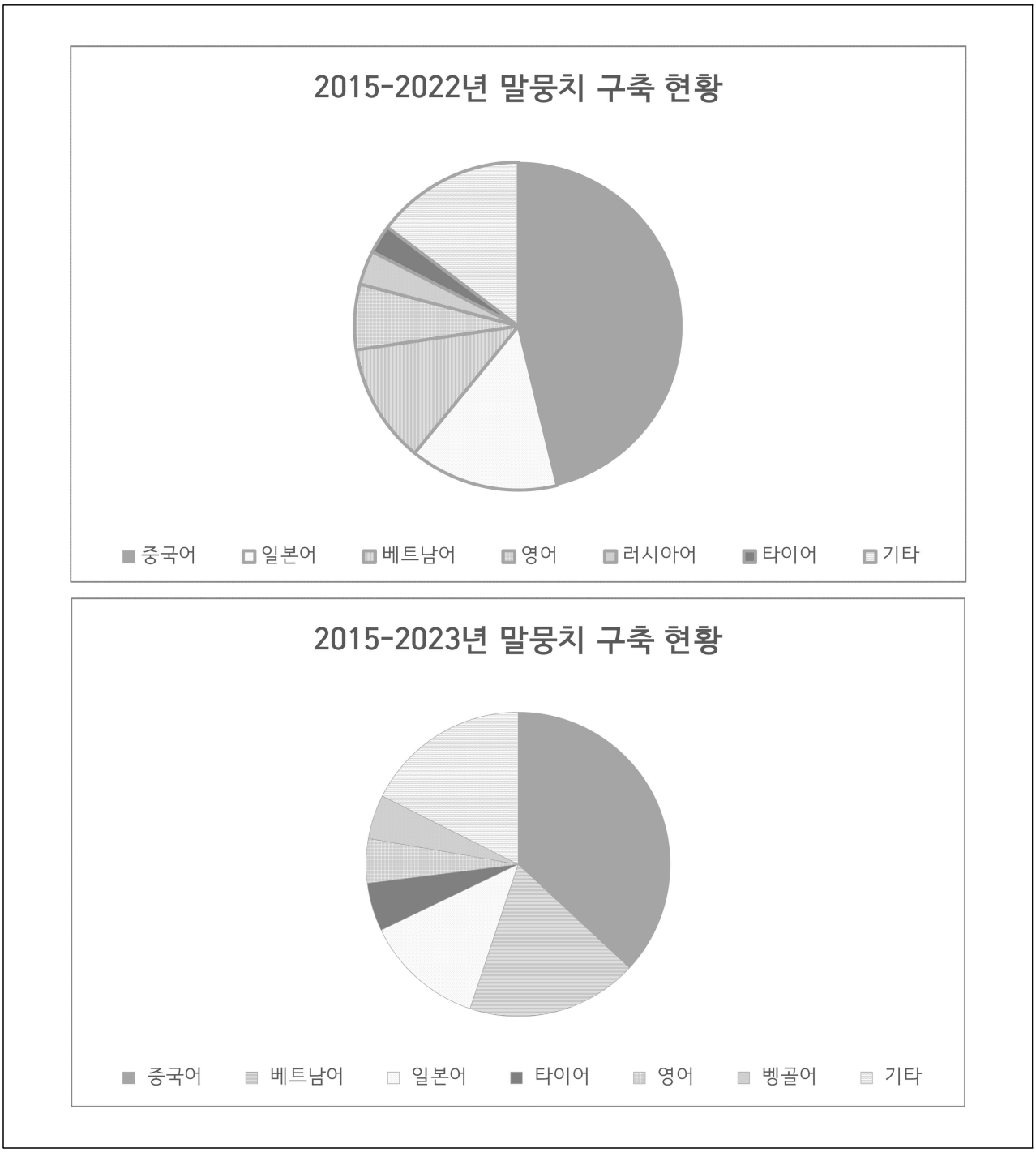
기존 학습자 말뭉치는 주로 국내 학습자를 대상으로 자료를 수집했기 때문에 국내 학습자의 구성 비율에 영향을 받았지만, 2023년 사업에서는 국외 학습자의 자료 비중을 대폭 확대하면서 중국어권 외 베트남어권, 벵골어권, 타이어권, 일본어권 등의 자료에 대해서도 적절한 비중으로 수집이 가능했다.

2023년 사업에서 중국어권 이외 학습자의 비중을 확대했지만 2015년-2023년 전체 말뭉치에서는 여전히 중국어권 학습자가 가장 높은 비중을 차지하였다. 다만 그 비율은 감소하였으며, 중국어권 이후 일본어 > 베트남어 > 영어 > 러시아어 > 타이어권이었던 순서는 베트남어 > 일본어 > 타이어 > 영어 > 벵골어권의 순서로 변경되었다.

<표 20> 2015년-2023년 말뭉치 구축 현황 - 학습자 언어권

순위	모국어	어절 수	비율	표본 수	비율
1	중국어	2,400,621	36.9	19,438	36.6
2	베트남어	1,185,178	18.2	10,315	19.4
3	일본어	828,669	12.7	6,763	12.7
4	타이어	333,400	5.1	1,886	3.6
5	영어	306,949	4.7	2,745	5.2
6	벵골어	305,074	4.7	2,531	4.8
7	러시아어	178,528	2.7	1,584	3.0
8	인도네시아어	152,137	2.3	1,025	1.9
9	몽골어	135,766	2.1	1,196	2.3
10	스페인어	87,077	1.3	783	1.5
11	프랑스어	62,380	1.0	557	1.0
12	아랍어	34,815	0.5	340	0.6
13	버마어	34,604	0.5	242	0.5
14	이탈리아어	34,187	0.5	271	0.5
15	말레이어	34,150	0.5	232	0.4
16	스웨덴어	29,186	0.4	311	0.6
17	우즈베크어	26,769	0.4	229	0.4
18	카자흐어	25,315	0.4	205	0.4
19	독일어	24,899	0.4	218	0.4
20	기타	283,563	4.4	2,228	4.2
총합		6,503,267	100.0	53,099	100.0

구축된 말뭉치에서 학습자 언어권별 비중이 변화하는 양상을 한눈에 비교할 수 있도록 도표로 나타내면 다음과 같다.



<그림 26> 학습자 언어권별 비중의 변화 양상 (어절 수 기준)

4) 장르

기구축 문어 원시 말뭉치에서는 생활문의 비중이 현저히 높긴 했으나, 인위적으로 비중이 적은 장르의 구축 비율을 확대하기보다는 한국어 교육 과정 및 일상적

의사소통의 특성을 반영하여 장르의 비중을 균형 있게 확대하는 것을 목표로 하였다. 특히 2023년 말뭉치 구축은 주제를 정한 뒤 해당 주제의 작문 자료를 수집하는 기획 수집으로 진행되었기 때문에 1.1.2.에서 언급한 바와 같이 생활문 2개, 논설문 4개, 설명문 5개로 장르를 구성하였다.[8] 학습자의 수준을 고려하되 설명문과 논설문을 비중을 확대하는 방식으로 수집을 진행하였다.

<표 21> 2015년-2022년 말뭉치 구축 현황 - 쓰기 장르

장르	어절 수	비율	표본 수	비율
생활문	2,066,719	46.8	21,758	58
논설문	1,408,256	31.9	9,804	26
설명문	274,612	6.2	2,168	6
보고서	239,247	5.4	216	1
기행문	140,613	3.2	1,330	3.6
수필	100,046	2.3	582	1.6
감상문	61,333	1.4	427	1.1
편지글	44,125	1.0	497	1.3
기사문	30,598	0.7	193	0.5
평전	16,111	0.4	109	0.3
기타	35,978	0.8	236	0.6
총합	4,417,638	100.0	37,320	100.0

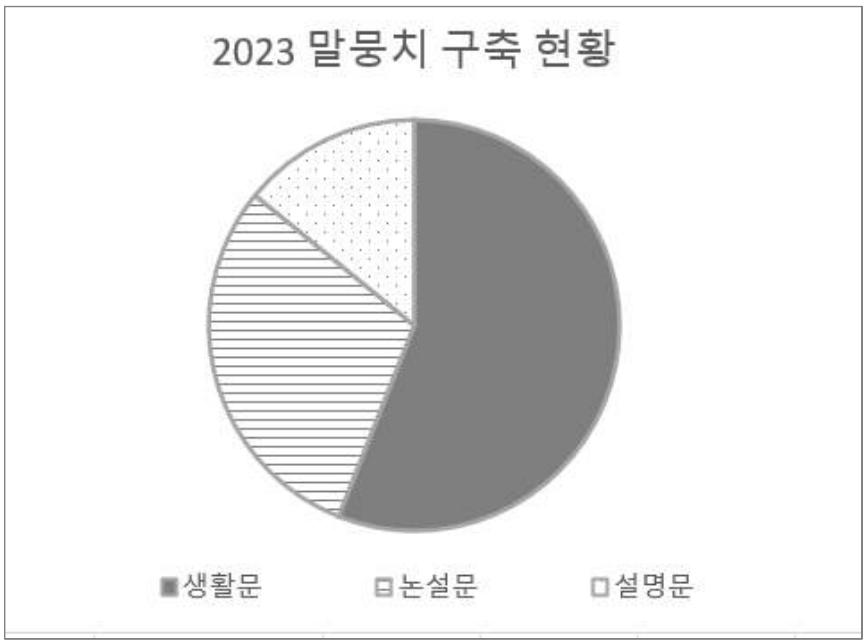
<표 22> 2023년 말뭉치 구축 현황 - 쓰기 장르

장르	어절 수	비율	표본 수	비율
생활문	1,168,203	56.0	9,384	59.5
논설문	623,450	29.9	4,102	26.0
설명문	293,976	14.1	2,293	14.5
총합	2,085,629	100.0	15,779	100.0

<표22>에서 확인할 수 있듯이 2023년 사업에서는 생활문, 논설문, 설명문의 세

8) 수준과 관계없는 공통 주제 2개는 각각 설명문과 논설문으로 구성하고, 초급은 생활문 1개와 설명문 2개, 중급은 생활문 1개, 설명문 1개, 논설문 1개, 고급은 설명문 1개와 논설문 2개로 구성하였다.

가지 장르에 대해서 수집 및 구축을 진행하였다. 논설문과 설명문의 비중을 높이고자 해당 장르에 각각 4개, 5개의 주제를 할당했지만 쓰기 자료 평가를 위해 학습자 수준과 무관하게 대량 수집을 진행한 주제가 생활문과 논설문이었기 때문에 상대적으로 생활문과 논설문이 높은 비중으로 수집되었다.



<그림 27> 2023년 말뭉치의 쓰기 장르별 비중 (어절 수 기준)

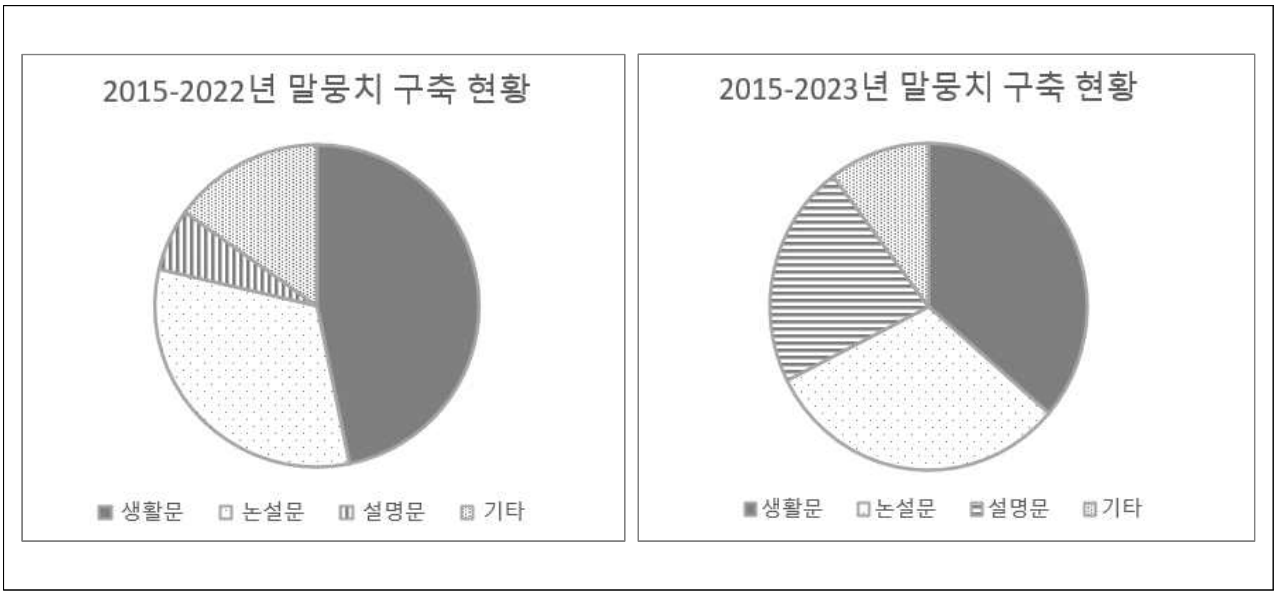
2023년 구축 말뭉치에서도 여전히 생활문이 가장 높은 비중으로 구축되기는 하였으나, 2015년-2023년 전체 말뭉치를 살펴보면 생활문의 비중이 줄고, 설명문의 비중이 증가한 것을 확인할 수 있다.

<표 23> 2015년-2023년 말뭉치 구축 현황 - 쓰기 장르

장르	어절 수	비율	표본 수	비율
생활문	2,360,695	36.3	24,051	45
논설문	2,031,706	31.2	13,906	26
설명문	1,442,815	22.2	11,552	22
보고서	239,247	3.7	216	0.4
기행문	140,613	2.2	1,330	2.5
수필	100,046	1.5	582	1.1
감상문	61,333	0.9	427	0.8
편지글	44,125	0.7	497	0.9

기사문	30,598	0.5	193	0.4
평전	16,111	0.2	109	0.2
기타	35,978	0.6	236	0.4
총합	6,503,267	100.0	53,099	100.0

구축된 말뭉치에서 생활문, 논설문, 설명문의 비중이 변화하는 양상을 한눈에 비교할 수 있도록 도표로 나타내면 다음과 같다.



<그림 28> 쓰기 장르별 비중의 변화 양상 (어절 수 기준)

5) 주제

2023년 학습자 말뭉치 수집은 기획 수집으로 진행되었기 때문에 최종 산출물에 특정 주제의 자료가 편중되지 않도록 고른 분포로 자료를 수집, 구축하고자 하였다. 다만, 쓰기 평가의 원활한 진행을 위해 평가 주제를 우선적으로 수집하도록 수집 참여 기관에 요청하였다.

쓰기 주제는 초급, 중급, 고급별로 각각 다섯 가지로(수준 무관 평가 주제 포함) 제시하였으며 수준별 쓰기 주제는 다음과 같다.

<표 24> 2023년 문어 원시 말뭉치 수준별 쓰기 주제

수준	주제
초급	자신의 나라와 한국 비교
	기후 위기와 해결 방안
	내가 가장 좋아하는 것과 싫어하는 것
	내가 가 본 여행지
	올해의 계획
중급	자신의 나라와 한국 비교
	기후 위기와 해결 방안
	과학 기술의 발전이 인간의 생활에 미치는 영향
	10년 후의 나의 계획
	효율적인 시간 관리 방법
고급	자신의 나라와 한국 비교
	기후 위기와 해결 방안
	내가 생각하는 성공
	인구 문제
	역사적 사건이나 인물 소개

말뭉치 분석 결과 초급, 중급, 고급의 모든 학습자들의 자료에서 쓰기 평가 주제인 '자신의 나라와 한국 비교', '기후 위기와 해결 방안'을 주제로 하는 쓰기가 가장 높은 비중을 차지했으며, 나머지 주제의 쓰기 자료는 비교적 고른 분포를 나타냈다.

초급 학습자들의 쓰기 자료에서는 '자신의 나라와 한국 비교'를 주제로 쓴 글이 24%로 가장 높은 비중을 차지하였다. 하지만 다른 평가 주제인 '기후 위기와 해결 방안'은 초급 학습자들에게는 난도가 높은 주제였으므로 14.1%의 가장 낮은 비율로 수집되었다. 나머지 주제에 대한 쓰기 자료는 약 21%, 20%, 17%의 비중을 차지하였다,

<표 25> 2023년 말뭉치 구축 현황 - 쓰기 주제 (초급)

주제	어절 수	비율	표본 수	비율
자신의 나라와 한국 비교	240,486	24.0	2,058	24.6
내가 가장 좋아하는 것과 싫어하는 것	213,443	21.3	1,803	21.6
내가 가 본 여행지	203,834	20.4	1,850	22.1
올해의 계획	166,661	16.6	1,429	17.1
기후 위기와 해결 방안	141,283	14.1	1,035	12.4
기타	35,663	3.6	190	2.3
총합	1,001,370	100.0	8,365	100.0

<그림 29> 2023년 말뭉치의 주제별 비중 - 초급 (어절 수 기준)

중급 학습자들의 쓰기 자료에서 역시 쓰기 평가 주제인 '자신의 나라와 한국 비교', '기후 위기와 해결 방안'이 각각 30.4%, 16.8% 비율로 1위, 2위로 나타났다. 나머지 주제에 대해서는 약 14%, 13%, 12%의 비율로 쓰기 자료가 구축되었다.

<표 26> 2023년 말뭉치 구축 현황 - 쓰기 주제 (중급)

주제	어절 수	비율	표본 수	비율
자신의 나라와 한국 비교	246,488	30.4	1,789	31.6
기후 위기와 해결 방안	136,350	16.8	913	16.1
과학 기술의 발전이 인간의 생활에 미치는 영향	111,385	13.7	781	13.8
효율적인 시간 관리 방법	108,876	13.4	860	15.2
10년 후의 나의 계획	95,028	11.7	674	11.9
기타	112,234	13.8	642	11.3
총합	810,361	100.0	5,659	100.0

<그림 30> 2023년 말뭉치의 주제별 비중 - 중급 (어절 수 기준)

고급 학습자들의 쓰기 자료에서도 쓰기 평가 주제인 '자신의 나라와 한국 비교', '기후 위기와 해결 방안'이 가장 높은 비중으로 구축되었으며 각각 24.9%, 19.3%의 비중을 차지하였다. 나머지 세 가지 주제는 약 17%, 13%, 13%의 비중으로 구축되었다.

<표 27> 2023년 말뭉치 구축 현황 - 쓰기 주제 (고급)

주제	어절 수	비율	표본 수	비율
자신의 나라와 한국 비교	63,299	24.9	398	24.6
기후 위기와 해결 방안	49,252	19.3	302	18.7
인구 문제	42,358	16.6	259	16.0
역사적 사건이나 인물 소개	33,939	13.3	252	15.6
내가 생각하는 성공	33,594	13.2	222	13.7
기타	32,113	12.6	183	11.3
총합	254,555	100.0	1,616	100.0

<그림 31> 2023년 말뭉치의 주제별 비중 - 고급 (어절 수 기준)

학습자가 동의서에 본인의 급수를 표시하지 않은 경우, '급수없음'으로 자료를 구축하였는데 해당 자료는 약 139편, 19,343어절로 집계되었다. '급수없음'으로 구축된 쓰기 주제는 다음과 같다.

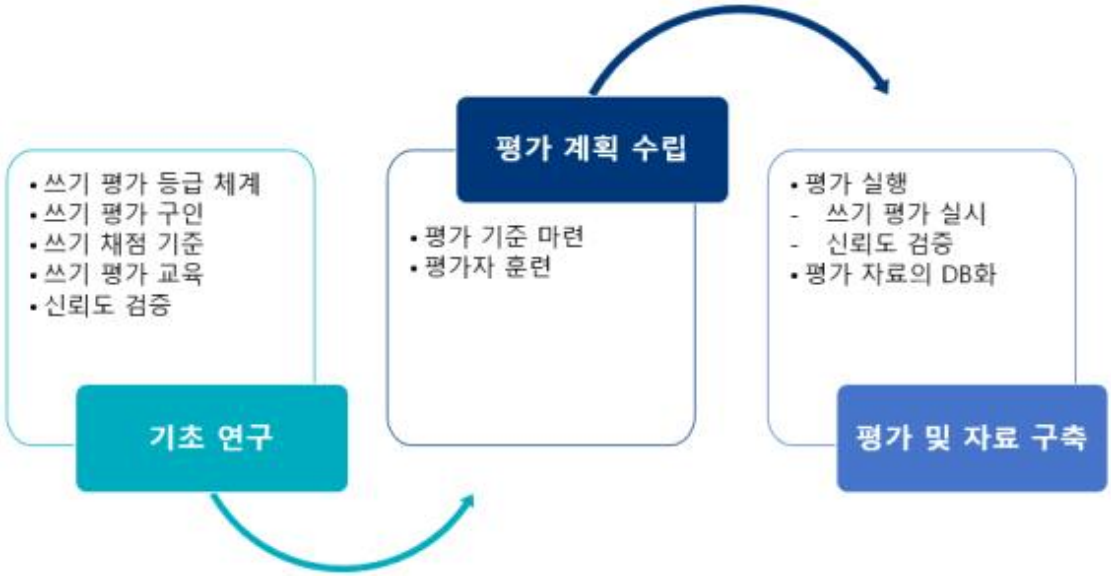
<표 28> 2023년 말뭉치 구축 현황 - 쓰기 주제 (급수없음)

주제	어절 수	비율	표본 수	비율
자신의 나라와 한국 비교	6,339	32.8	42	30.2
기후 위기와 해결 방안	4,738	24.5	36	25.9
내가 생각하는 성공	2,174	11.2	16	11.5
내가 가장 좋아하는 것과 싫어하는 것	1,937	10.0	14	10.1
내가 가 본 여행지	1,277	6.6	9	6.5
올해의 계획	1,214	6.3	9	6.5
10년 후의 나의 계획	483	2.5	4	2.9
효율적인 시간 관리 방법	483	2.5	4	2.9
과학 기술의 발전이 인간의 생활에 미치는 영향	383	2.0	3	2.2
인구 문제	164	0.8	1	0.7
역사적 사건이나 인물 소개	151	0.8	1	0.7
총합	19,343	100.0	139	100.0

2023년 한국어 학습자 말뭉치 구축 내용을 분석한 결과, 1.1.에서 설계한 내용에서 크게 벗어나지 않는 범위에서 말뭉치가 적절하게 구축되었다는 것을 확인할 수 있었다. 특히 2023년 사업에서는 국외 학습자의 비중을 확대한 것이 기존 말뭉치와의 가장 극명한 차이점이라고 하겠다.

2. 한국어 학습자 쓰기 평가 자료 구축

한국어 학습자 쓰기 평가 자료 구축은 자동 채점, 자동 첨삭 등의 기술 개발을 위한 기반 자료를 마련하여 공공 데이터로서 연구 및 교육, 산업계에서의 활용도를 제고하는 것을 목적으로 한다. 2종의 주제 작문 각 1,500편씩 총 3,000건의 작문을 대상으로 한국어교육 전문가의 평가 점수, 평가 의견을 제시하였다.



<그림 32> 한국어 학습자 쓰기 평가 자료 구축 연구 방법 및 절차

2.1. 기초 연구

1) 평가 등급

평가 구인 및 기준, 점수 부여 체계 등을 포함한 평가 계획을 수립하기 위해서는 한국어 숙달도 등급 체계에 대한 검토가 선행되어야 한다. 본 사업에서는 연구의 성과가 인공지능 학습용 데이터 구축을 위한 기초 연구로 한국어 표준 교육과정의 등급 체계에 따르는 것이 타당하다고 보았다. 이에 따라 국내외의 한국어 교육과정은 물론 한국어능력시험(TOPIK), 세종한국어평가(SKA) 등의 공인 평가의 근간이 되고 있는 '국제 통용 한국어교육 표준 모형'의 등급 체계에 따라 6등급 체계로 평가 체계를 마련하였다. 다음은 '국제 통용 한국어교육 표준 모형'의 쓰기 영역의 총괄 목표를 제시한 것이다.

<표 29> ‘국제 통용 한국어교육 표준 모형’의 쓰기 총괄 목표(국립국어원, 2017)

등급	세부 내용
1급	기본적인 맞춤법에 맞게 글자를 정확하게 쓸 수 있으며 일상생활에 관한 간단한 글을 쓸 수 있다.
2급	일상생활에서 경험한 일이나 친숙한 인물에 관한 글을 쓸 수 있다.
3급	친숙한 사회적·추상적 주제로 된 글을 간단한 구조로 쓸 수 있다.
4급	친숙한 사회적·추상적 주제로 된 글을 정확하게 쓸 수 있으며 설명문, 논설문, 쉽고 짧은 감상문을 쓸 수 있다.
5급	친숙하지 않은 사회적·추석적 주제나 자신의 전문 분야에 관한 글을 구조에 맞게 쓸 수 있다.
6급	친숙하지 않은 사회적·추상적 주제나 자신의 전문 분야에 관한 글을 논리적인 구조로 쓸 수 있으며 다양한 장르의 특성을 고려한 글을 쓸 수 있다.

다음은 총괄 목표에 따른 등급별 목표 및 내용 기술을 제시한 것이다.

<표 30> ‘국제 통용 한국어교육 표준 모형’의 쓰기 등급별 목표 및 내용(국립국어원, 2017)

등급	목표/내용	세부 내용
1급	목표	1. 소리를 듣고 쓸 수 있다. 2. 간단한 메모(목록 작성)를 할 수 있다. 3. 구, 절 단위 혹은 짧은 문장 단위의 일상적인 글을 쓸 수 있다.
	내용	1. 한글 자음과 모음, 받침을 보고 쓴다. 2. 구두로 익힌 어휘를 따라 쓴다. 3. 짧은 문장을 듣고 쓴다.
2급	목표	1. 일상적인 주제와 관련된 짧고 간단한 글을 쓸 수 있다. 2. 친교 목적의 글(이메일, 편지 등)을 쓸 수 있다.
	내용	1. 자신이 경험한 것을 글로 쓴다. 2. 친숙한 인물, 사물, 장소 등을 간단하게 소개하는 글을 쓴다. 3. 친숙한 인물, 사물, 장소 등을 구체적으로 소개하는 글을 쓴다. 4. 미래 계획을 글로 쓴다. 5. 일상생활 및 여가 생활과 관련된 글을 쓴다. 6. 시간 순서대로 글을 쓴다. 7. 그림을 보고 상황을 추측하여 문장을 만들어 쓴다.
3급	목표	1. 개인적이고 친숙한 내용의 글을 익숙하게 쓸 수 있다. 2. 간단하게 설명하는 글을 쓸 수 있다.

등급	목표/내용	세부 내용
	내용	1. 문어와 구어의 특징을 이해하여 글을 쓴다. 2. 단락의 구조를 이해한 글을 쓴다. 3. 중심문장과 뒷받침문장으로 이루어진 글을 쓴다. 4. 간단히 설명하는 글을 쓴다. 5. 일상생활과 관련된 대화나 생활문을 구체적으로 정확하게 쓴다. 6. 개인적 관계를 유지하는 데 필요한 글을 쓴다.
4급	목표	1. 다양한 기법(열거, 비교 등)을 활용하여 설명하는 글을 쓸 수 있다. 2. 친숙한 사회적 소재(직업, 사랑, 결혼 등)에 대한 간단한 글을 쓸 수 있다.
	내용	1. 간단한 논리적인 구조의 글을 쓴다. 2. 하나의 의미를 다양한 방법으로 표현하는 글을 쓴다. 3. 자신과 친숙한 사회적 소재를 글을 쓴다. 4. 글 전체의 구성을 염두에 두고 글을 쓴다. 5. 일반적인 주제에 관한 말이나 대화를 듣고 요약하여 쓴다. 6. 수집한 정보나 표현 등을 활용하여 글을 확장하여 쓴다. 7. 특정한 입장에 대해 찬반논거를 제시하는 글을 쓴다. 8. 편지를 쓸 수 있고, 그 편지에서 사건과 경험이 나에게 주는 의미를 분명하게 표현한다. 9. 감상문, 수필 등을 쓴다. 10. 다양한 주제에 관한 글을 읽고 요약하여 쓴다. 11. 다양한 주제에 대해 자신의 생각이나 느낌을 쓴다. 12. 자신의 견해나 태도를 나타내는 글을 쓴다.
5급	목표	1. 문어와 구어의 차이를 알고 문어의 격식에 맞는 글을 쓸 수 있다. 2. 논증의 형식에 맞추어 주장하는 글을 쓸 수 있다.
	내용	1. 논리적 구조에 맞게 글을 쓴다. 2. 다양한 담화 표지를 활용한 글을 쓴다. 3. 다양한 기법을 활용해 설명하는 글을 쓴다. 4. 문어적 표현을 적절하게 활용하여 글을 쓴다. 5. 자신의 생활과 관련이 적은 사회적, 추상적 소재에 대해 간단한 글을 쓴다. 6. 설명문, 보고서, 논설문 등을 요약해서 쓴다. 7. 감상문, 수필 등을 요약해서 쓴다. 8. 일반적인 주제에 관한 친숙한 내용의 글을 읽고 첨삭하여 다시 쓴다.

등급	목표/내용	세부 내용
		9. 정보를 이용하거나 인용해서 글을 쓴다. 10. 설명문, 보고서, 논설문 등을 쓴다. 11. 다양한 주제들에 대해 분명하고 상세한 텍스트를 쓴다. 12. 다양한 표현법 중 적절한 표현을 선태갧 글을 쓴다. 13. 다양한 분야와 관련된 서식을 작성하거나 통신문을 쓴다. 14. 다양한 주제에 관한 자신의 생각을 조리 있게 글로 표현한다.
6급	목표	1. 다양한 기법(예시, 비유 등)을 활용하여 설명하거나 논증하는 글을 쓸 수 있다. 2. 자신의 전문 분야(직업적, 학문적 영역 등)에 관한 글을 쓸 수 있다.
	내용	1. 업무적, 학문적 영역에서 요구되는 글을 쓴다. 2. 격식에 맞게 어휘, 문체 등을 선택하여 글을 쓴다. 3. 다양한 주제의 글을 읽고 적절한 결론을 추론하여 글을 완성한다. 4. 다양한 주제에 관한 글을 읽고 이에 대한 자신의 생각을 논리적으로 쓴다.
7급	목표	1. 업무 보고서, 사업계획서 등의 공적이고 전문적인 글을 작성할 수 있다. 2. 논문, 보고서 등을 격식에 맞게 쓸 수 있다. 3. 전반적인 사회적, 추상적 주제에 대해 자신의 의견을 글로 표현할 수 있다.
	내용	1. 전문 분야에서 요구되는 글을 쓴다. 2. 정확하고 격조 있는 글을 쓴다. 3. 상호텍스트성을 활용한 지식 구성의 글을 쓴다. 4. 한국어 텍스트 구조와 해당 분야의 텍스트 특성을 살려 글을 쓴다.

2) 평가 구인

평가에서 구인은 “시험 점수를 통해서 추론할 수 있는 수험자의 언어능력에 대한 이론적 근거(지현숙, 2014:10)”가 되는 것으로 평가 대상으로 삼고자 하는 언어능력을 이루는 구성 요소로, 구인 설정은 실제 언어 사용 시 어떤 요소들이 관련되는가와 측정하고자 하는 요소가 무엇인가를 결정하는 것을 의미한다(Weigle, S. C., 2001). 구인을 설정할 때에는 의사소통 능력의 근간이 되는 문법적 능력, 담화적 능력, 사회언어학적 능력, 전략적 능력을 기반으로 하며, 평가 영역별 특성

에 따라 직접적인 평가 구인이 되는 요소들을 변별적으로 설정할 수 있다. 쓰기 평가의 구인은 실제 한국어 쓰기 교육의 내용이 어떻게 구성되며, 그중 무엇을 측정하고자 하는지에 따라 달라질 수 있다. 일반적으로 쓰기 평가의 구인은 다음과 같이 쓰기 지식과 관련된 것으로 평가자가 관찰 및 측정 가능한 것으로 규정하는 것이 일반적이다(강현화 외, 2023:203-205).

- 철자, 구두점, 띄어쓰기: 어문규범에 맞는 철자, 구두점, 띄어쓰기 능력
- 어휘력: 어휘를 맥락과 상황에 맞게 정확하게 잘 사용하는가에 관한 능력
- 문법 사용 능력: 문법 사용은 다양한 문법 형태를 정확하게 상황과 맥락에 맞게 잘 사용하는가에 관한 능력
- 텍스트 구성 능력: 텍스트 구성 능력은 문장 단위를 넘어 텍스트 차원에서의 구성 능력
- 정확성: 정확성은 어휘, 문법, 표현 등의 언어적 요소를 얼마나 정확하게 사용하는가와 관련이 되어 있는 항목
- 다양성: 다양성이란 의미 전달을 위하여 사용하는 어휘나 문법의 범위가 얼마나 넓은가의 정도
- 적절성: 적절성은 글의 종류, 독자, 맥락, 상황 등에 따라 언어를 잘 선택하여 표현하는 능력
- 비판적 사고와 창의성: 쓰기 주제에 대한 비판적 사고를 통한 창의적인 내용 구성 및 글쓰기 능력
- 주제 지식: 쓰기에 어떤 내용을 쓰는가와 관련된 요소로 주어진 쓰기 과제에서 요구하는 내용을 알맞은 방향 쓰는 능력

한국어 공인 숙달도 평가인 한국어능력시험(TOPIK)에서는 이러한 항목들을 토대로 작문형 문항에 대한 평가 구인을 다음과 같이 설정하고 있다.

<표 31> ‘한국어능력시험(TOPIK)’의 쓰기 평가 구인

평가 범주	평가 내용
내용 및 과제 수행	• 주어진 과제를 충실히 수행하였는가? • 주제에 관련된 내용으로 구성하였는가? • 주어진 내용을 풍부하고 다양하게 표현하였는가?
글의 전개 구조	• 글의 구성이 명확하고 논리적인가? • 글의 내용에 따라 단락 구성이 잘 이루어졌는가? • 논리 전개에 도움이 되는 담화 표지를 적절하게 사용하여 조직적으로 연결하였는가?
언어 사용	• 문법과 어휘를 다양하고 풍부하고 사용하며 적절한 문법과 어휘를 선택하여 사용하였는가? • 문법, 어휘, 맞춤법 등의 사용이 정확한가? • 글의 목적과 기능에 따라 격식에 맞게 글을 썼는가?

본 사업에서 한국어 학습자의 쓰기 평가 자료를 구축하기 위해 이들을 참고하되, 본 사업에 제시한 과제가 숙달도 평가 문항과는 목적이 다름을 고려하여 다음과 같이 세부 평가 내용을 조정하여 평가 구인을 설정하였다.

<표 32> 본 사업의 쓰기 평가 구인

평가 범주	평가 내용
내용 조직	• 글의 목적과 기능에 맞는 적절한 형식으로 글을 썼는가 • 주제에 관련된 내용으로 구성하였는가 • 주어진 내용을 풍부하게 표현하였는가
글의 구조	• 글의 구성이 명확하고 논리적인가 • 글의 내용에 따라 단락 구성이 잘 이루어졌는가 • 문장 간의 연결이 자연스럽고 짜임새 있는가 • 글의 내용이 긴밀하게 구성되었는가
언어 사용	• 문법과 어휘를 정확하게 사용하였는가 • 문법과 어휘를 다양하게 사용하였는가 • 문법과 어휘를 적절하게 사용하였는가

3) 평가 방식과 기준

쓰기 평가의 방식에는 종합적 채점과 분석적 채점이 있다. 종합적 채점은 인상적 평가를 통해 쓰기 결과물에 단일한 점수를 부여하는 방식이다. 채점자들이 쓰기 채점을 빠르게 진행할 수 있으나 채점자의 역량에 따라 채점의 신뢰도가 달라질 수 있다는 한계가 있다. 이와 달리 분석적 채점은 평가 구인별로 점수를 부여하는 방식으로 채점자 간 차이를 차이를 최소화하여 채점의 신뢰도를 높일 수 있다.

본 사업에서는 채점의 신뢰도를 제고하기 위하여 분석적 채점 방식을 채택하였으며, 정교한 채점을 위하여 앞서 제시한 구인을 토대로 하여 다음과 같이 5간 척도로 구분하여 평가 기준을 마련하였다.

<표 33> 본 사업의 쓰기 평가 기준

구인 / 척도	내용 조직	글의 구조	언어 사용
5	• 주제와 관련된 내용을 풍부하게 구성함 • 글의 내용이 장르의 특성에 맞게 잘 구성됨	• 글의 시작과 마무리, 단락이 짜임새 있게 구성되어 글의 완성도가 높음 • 글의 내용이 담화적으로 긴밀하게 구성됨	• 글의 주제와 관련된 어휘와 문법을 풍부하고 다양하게 사용함 • 어휘와 문법을 글의 특성에 맞게 적절하게 사용함 • 해당 수준의 어휘와 문법 사용에서 오류가 거의 없음
4	• 주제와 관련된 내용을 비교적 풍부하게 구성함 • 글의 내용이 장르의 특성에 맞게 비교적 잘 구성됨	• 글의 시작과 마무리, 단락이 어느 정도 짜임새 있게 구성되어 글의 완성도가 비교적 높음 • 글의 내용이 담화적으로 비교적 긴밀하게 구성되어 글의 흐름을 방해하지 않음	• 글의 주제와 관련된 어휘와 문법을 비교적 풍부하고 다양하게 사용함 • 글의 특성을 고려하여 어휘와 문법을 비교적 적절하게 사용함 • 어휘와 문법 사용에서 오류가 약간 있으나 글

			의 내용을 이해하는 데에 문제가 없음
3	• 주제와 관련된 내용으로 글을 구성하였으나 내용이 다소 빈약함 • 간혹 부적절한 부분이 있으나 글의 내용이 장르의 특성에 어느 정도 맞게 구성됨	• 글의 시작과 마무리, 단락이 짜임새 있게 구성되지 못하여 글의 완성도가 부족함 • 글의 내용이 담화적으로 어느 정도 긴밀하게 구성되었으나 간혹 글의 흐름을 방해하는 부분이 있음	• 제한적이지만 글의 주제와 관련된 어휘와 문법을 어느 정도 풍부하고 다양하게 사용함 • 간혹 부적절한 부분도 있으나 글의 특성을 고려하여 어휘와 문법을 어느 정도 적절하게 사용함 • 어휘와 문법 사용에서 오류가 많아 글의 내용을 이해하는 데에 다소 문제가 있음
2	• 간혹 주제와 관련이 없는 내용이 글에 포함되거나 내용이 빈약함 • 장르의 특성에 맞게 글의 내용을 구성하지 못함	• 글의 시작과 마무리, 단락 구성에서 짜임새가 없고 각각의 내용을 나열함 • 글의 내용이 담화적으로 긴밀하지 못하여 글의 흐름이 매끄럽지 못함	• 글의 주제와 관련된 어휘를 다양하게 사용하지 못하고 쉽고 단순한 어휘와 문법을 반복적으로 사용함 • 글의 특성을 고려하여 어휘와 문법을 적절하게 사용하지 못함 • 어휘와 문법 사용에서 오류가 너무 많아 글의 내용을 이해하기가 어려움
1	• 주제와 관련된 내용으로 글을 구성하지 못함	• 완결된 구조의 담화를 구성하지 못함	• 어휘와 문법 사용에 오류가 많아 글의 내용을 이해할 수 없음

아울러 다음과 같이 내용 조직 30점, 글의 구조 30점, 언어 사용 40점으로 총점을 100점으로 설정하고, 척도별 점수 구간을 제시하였다. 채점 시에는 채점자가 각 구인별로 적정 점수를 부여하도록 하고, 입력된 점수를 토대로 척도 값이 자동

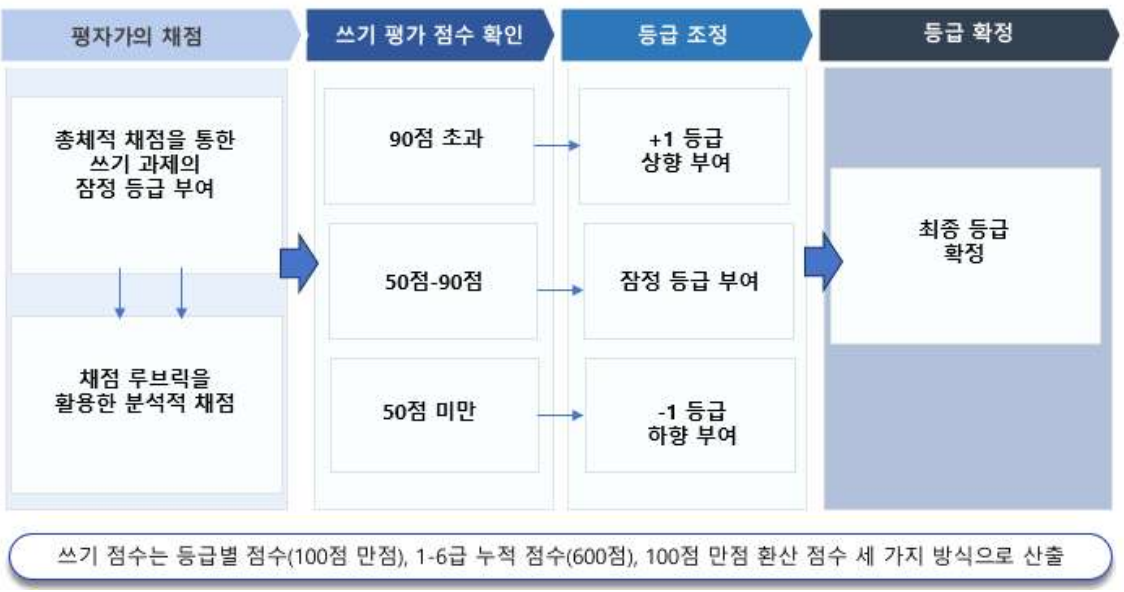
반환되도록 하였다.

<표 34> 척도별 점수 구간

구인 / 척도	내용 조직(30점)	글의 구조(30점)	언어 사용(40점)
5	28점이상	28점이상	37점이상
4	24-27점	24-27점	32-36점
3	19-23점	19-23점	26-31점
2	15-18점	15-18점	21-25점
1	14점 이하	14점 이하	20점이하

4) 평가 점수 및 등급 부여 체계

쓰기 점수는 평가 결과 부여된 점수(각 등급별 100점 만점)를 기반으로 하여 다음과 같이 등급을 부여하였다. 즉, 평가자가 분석적인 평가에 앞서 작문과 함께 제공된 메타정보, 경험적 직관을 토대로 잠정 등급을 부여하고, 평가 기준에 따라 분석적 채점을 한 후 산출된 점수가 50-90점이면 잠정적으로 부여한 등급을 유지하고, 90점을 초과하면 등급을 +1 등급 상향 조정하고, 50점 미만이면 -1 등급 하향 조정하였다.



<그림 33> 평가 점수 및 등급 부여 체계

쓰기 점수는 각 등급별로 부여한 100점 만점의 점수, 1-6급까지 각 등급별로 100점씩 누적하여 산출한 600점 만점의 점수, 이를 다시 100점 만점으로 환산한 점수 세 가지로 제시하였다. 다음은 세 가지 유형의 점수 환산표 일부를 보인 것이다.

<표 35> 점수 환산표

등급	등급별 점수	1-6급 누적 점수 (600점 만점)	환산 점수 (100점 만점)
1급	50	50	8
1급	75	75	13
1급	100	100	17
2급	50	150	25
2급	75	175	29
2급	100	200	33
3급	50	250	42
3급	75	275	46
3급	100	300	50
4급	50	350	58
4급	75	375	63
4급	100	400	67
5급	50	450	75
5급	75	475	79
5급	100	500	83
6급	50	550	92
6급	75	575	96
6급	100	600	100

2.2. 채점자 훈련

본 사업에서는 본격적인 채점에 앞서 다음과 같은 절차에 따라 채점자 훈련을 통해 채점자 간 신뢰도와 채점자 내의 신뢰도를 확보하였다.

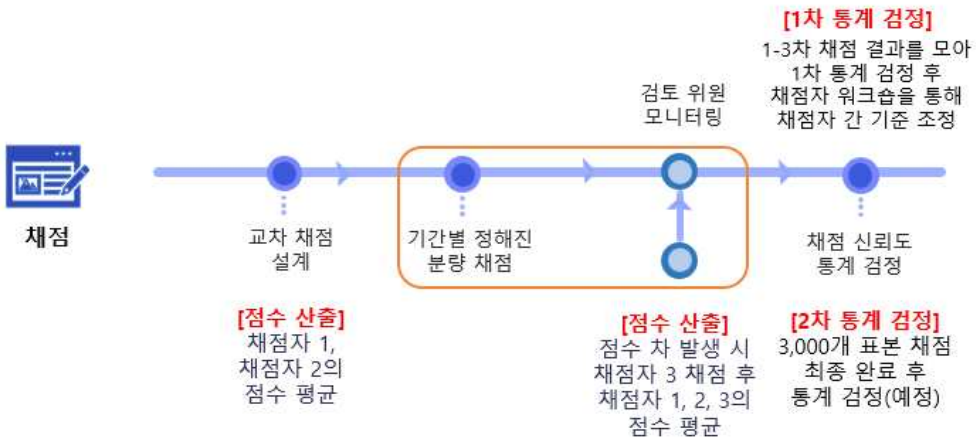
<그림 34> 채점자 교육 절차

사업을 수행하는 동안 총 4회의 채점자 워크숍을 실시하였으며, 워크숍 외에도 격주로 정기 실무회의와 수시회의를 진행하면서 회차별 채점 결과를 검토하였다.

<표 36> 채점자 워크숍 개최 일시 및 내용

목표	구분	일시	내용
채점 기준 설명 및 적용연습, 조정	1차	2023년 8월 28일(월) 10:00-12:00	평가 목적, 등급 기술 및 채점 기준 설명, 샘플 채점
	2차	2023년 10월 12일(목) 16:00-18:00	1차 채점 결과 검토 및 채점자 간 기준 조정
	3차	2023년 12월 15일(금) 16:00-18:00	2-3차 채점 결과 검토 및 채점자 간 기준 조정
점수 체계 확정	4차	2023년 3월 14일(목) 17:00-19:00	점수 체계 및 채점 관련 쟁점 논의

아울러 채점자 교육의 일환으로 초기 300개 표본의 채점 결과에 대한 통계 검정을 실시하고 그 결과를 채점자들에게 피드백함으로써 채점자 간, 그리고 채점자 내적 신뢰도를 확보하고자 하였다.



<그림 35> 채점자 신뢰도 확보 방안

1) 채점자 간 신뢰도

채점 신뢰도는 두 평가자 간의 점수 차이를 측정하는 방법으로 확인할 수 있다. 둘 이상의 평가자 간의 채점 점수 사이의 신뢰도를 측정하는 방법은 두 점수 사이의 상관관계에 기초하여 피어슨의 상관계수를 구하는 것으로 .6 이상이면 어느 정도 신뢰성이 있고, .7 이상이 신뢰성이 높다고 할 수 있다. 초기 300건의 채점 결과에 대한 채점자 간 신뢰도는 등급 판정이 .973으로 매우 높은 것으로 나타난 반면, 척도 판정이나 점수 부여에 있어서는 언어 척도 .700, 언어 점수 .732를 제외하고 높지 않았다. 이를 통해 본 사업에서는 채점자 간 신뢰도를 높이기 위한 노력이 필요함을 확인하고 워크숍과 함께 실무회의를 통해 지속적인 조정과 협의를 하면서 채점을 진행하였다.

<표 37> 채점자 간 신뢰도 검정 결과

항목	Reliability	ICC
내용 조직 척도	.350	.517
글의 구조 척도	.426	.568
언어 사용 척도	.538	.700
내용 조직 점수	.345	.513
글의 구조 점수	.374	.545
언어 사용 점수	.579	.732
등급	.948	.973

2) 채점자 내 신뢰도

본 사업에서는 다국면 라쉬모형을 바탕으로 채점자 내 신뢰도를 분석하였다. 1차 채점 신뢰도 분석 결과, 채점자 내 신뢰도에서 과적합과 부적합 평가자가 발견되었다. 분석 기준은 내적합 제곱평균값 0.5~1.5, 내적합 표준화값 -2~+2로 하였고, 위의 기준에서 벗어나 있을 경우에 과적합, 부적합으로 해석하였다.

- 채점자 1, 3 : 과적합 (채점자 1 표준화값 -2.8, 채점자 3 표준화값 -4.1)
- 채점자 2, 4 : 부적합 (채점자 2 표준화값 3.4, 채점자4 표준화값 2.1)l

채점자 간 신뢰도 산출 공식에 따르면 채점자 간 신뢰도는 [(실제일치도-기대일치도)/(100-기대일치도)]=(40.7-45.2)/(100-45.2)=-0.08이다. 실제 일치도와 기대일치도가 같았을 때 0로짓을 갖게 된다. 일치도가 음수의 값을 보일 경우 모형에 의해서 예측된 채점자 간 일치도에 비해 실제 채점자 사이에 채점 불일치가 더 클 수 있음을 의미한다.

채점 경향성 분석 결과, 채점자 1이 가장 후한 채점자로 평균 3. 57을 주었고, 채점자 3이 가장 엄한 채점자로 평균 3.07점을 주었다. 이는 채점자 내의 신뢰도가 낮음으로 해서 채점자 간의 신뢰도를 낮추고 있는 것으로 파악된다.

1차 채점 신뢰도 검정 결과는 평가 초기에 채점자 내 신뢰도를 높이고 채점자 간 채점 기준을 조정하는 것을 목적으로 실시하였다. 신뢰도를 높이기 위한 조치로 분석 결과를 채점자에게 공개하고 워크숍을 통해 채점자 훈련을 실시하여 채점 기준을 맞추고자 하였다. 그 결과 3,000건의 채점에 대해서는 높은 신뢰도를 확보할 수 있었다(평가 결과에 대한 신뢰도는 2.3의 기술 참고).

2.3. 평가 결과

1) 학습자 쓰기 평가 자료 분포

본 사업에서는 주제 1(한국과 자신의 나라 비교) 1,500건, 주제 2(기후 문제와 대응 방안) 1,500건으로 총 3,000건의 학습자 작문을 평가하였다. 평가 대상이 된 작문은 총 81개국 학습자가 산출한 자료로 베트남 887건으로 가장 많았고, 다음으로 중국 535건, 일본 417건, 태국 230건, 방글라데시 127건이 그 뒤를 이었다. 다음은 국적별 자료의 분포를 나타낸 것이다.

<표 38> 학습자 쓰기 평가 자료의 국적 분포

국적	주제1	주제2	합계
베트남	256	631	887
중국	366	169	535
일본	288	129	417

태국	51	179	230
방글라데시	2	125	127
몽골	50	31	81
대만	58	16	74
인도네시아	21	42	63
러시아	39	11	50
미얀마	22	27	49
홍콩	36	9	45
프랑스	31	4	35
우즈베키스탄	22	8	30
이탈리아	14	14	28
슬로바키아	10	15	25
독일	17	4	21
말레이시아	16	5	21
이집트	9	9	18
키르키즈스탄	14	3	17
영국	12	2	14
미국	12	1	13
카자흐스탄	4	9	13
스페인	8	4	12
스리랑카	9	1	10
슬로베니아	1	9	10
기타	132	43	175
합계	1,500	1,500	3,000

2) 학습자 쓰기 평가 결과

3,000건의 표본에 대한 평가 결과, 주제 1은 다음에서 보는 바와 같이 1급 207개, 2급 368개, 3급 333개, 4급 259개, 5급 241개, 6급 92개로 최종 등급이 부여되었다. 주제 2는 1급 75개, 2급 220개, 3급 484개, 4급 459개, 5급 189개, 6급 73개로 최종 등급이 부여되었다. 각 표본에 대해서는 등급 외에 등급별 점수(100점 만점), 1-6급 누적 점수(600점 만점), 최종 점수(100점 만점)와 함께 총평을 제시하였다.

<표 39> 학습자 쓰기 평가 결과

주제	1급	2	3	4	5	6	합계
주제1	207	368	333	259	241	92	1,500
주제2	75	220	484	459	189	73	1,500
합계	282	588	817	718	430	165	3,000

3) 학습자 쓰기 평가 결과에 대한 신뢰도 검정

본 사업에서는 피어슨 상관계수와 급내 상관계수(intra-class correlation coefficient: ICC)를 통해 3,000개 표본의 채점 결과에 대해 채점자 간 신뢰도 검정을 실시하였다. 피어슨의 상관계수는 두 채점자가 동일한 집단의 피험자에게 얼마나 유사하게 점수를 부여하고 있는가를 분석하기 위해 사용되는 것으로 상관계수가 0.6 이상이면 신뢰성이 있고 0.7 이상이면 신뢰성이 높다고 본다. 반면, 급내 상관계수는 평가의 반복성과 재현성, 채점자 간 신뢰도를 평가하기 위해 사용되는 또 다른 신뢰도 계수로 0부터 1사이의 값을 가지며 0.8 이상이면 신뢰도가 높다고 할 수 있다.

분석 결과 주제1의 경우, 언어 점수가 .832로 가장 높은 신뢰도를 보이며 구조 점수가 .674로 다른 두 구인에 비해 상대적으로 낮으나 상관계수나 ICC 모두 기준점 이상의 높은 점수를 보이는 것으로 나타났다.

〈표 40〉 주제 1 채점의 신뢰도

항목	상관계수	ICC
내용 조직 점수	.708	.829
글의 구조 점수	.674	.806
언어 사용 점수	.832	.908
총점	.787	.881

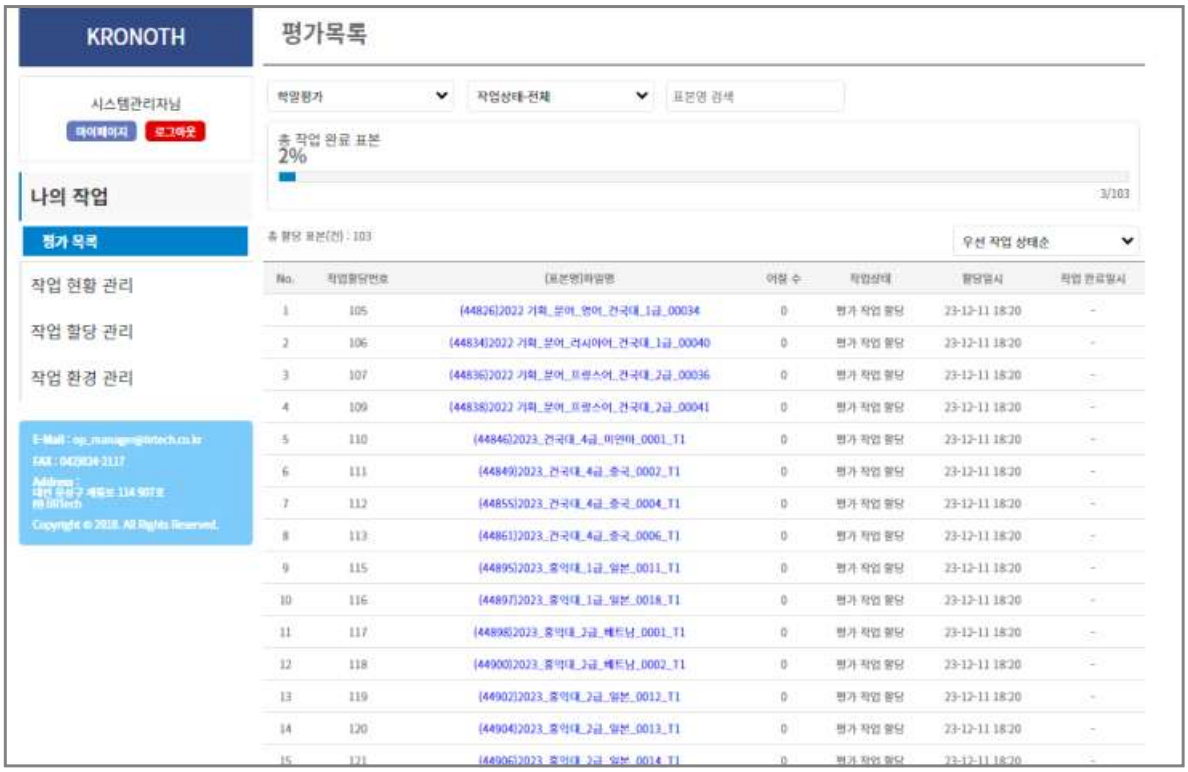
주제2의 경우, 상관계수가 언어 사용 .939, 내용 조직 .902, 글의 구조 .895로 주제1에 비해 모든 구인에서 높은 신뢰도를 보이고 있으며 ICC 또한 높아 매우 우수한 것으로 나타났다.

<표 41> 주제 2 채점의 신뢰도

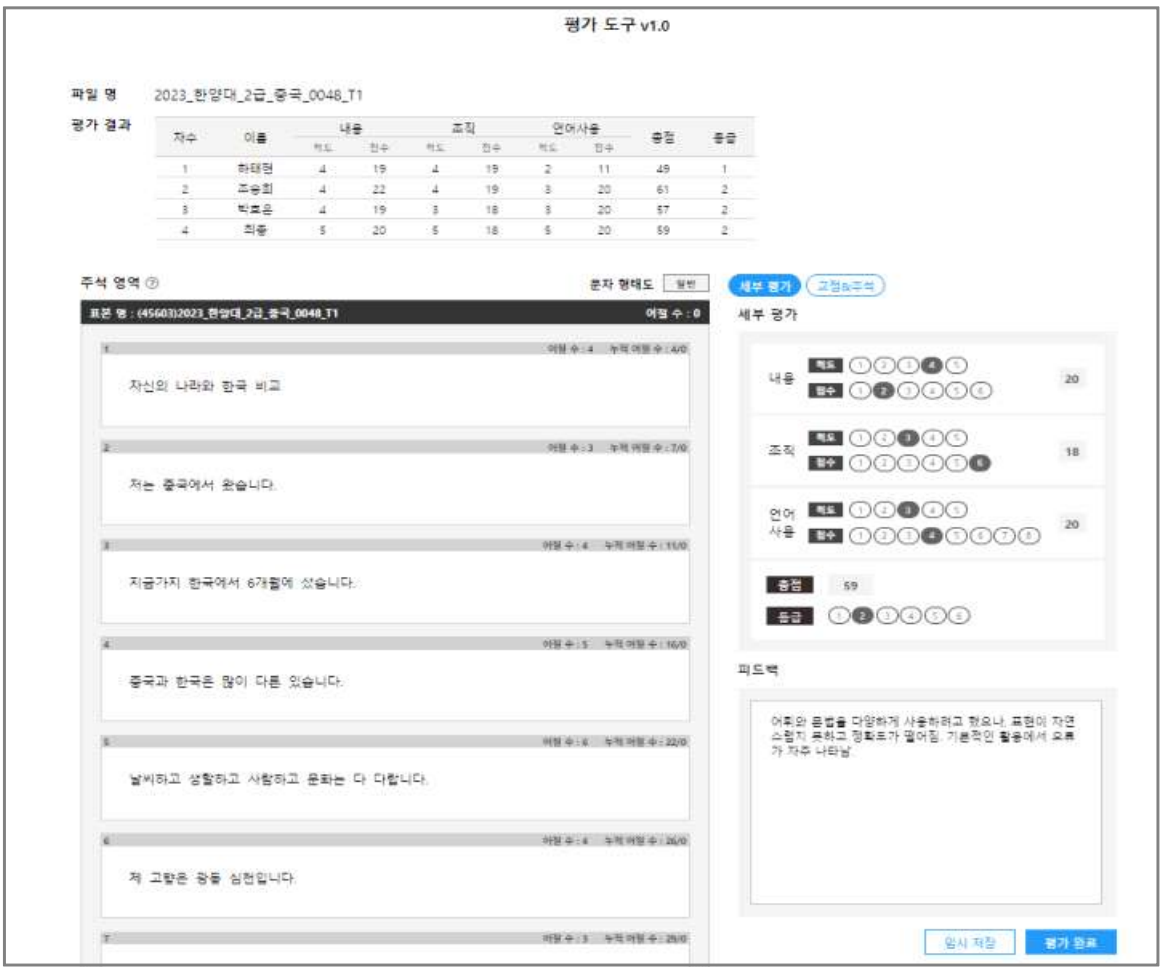
항목	Reliability	ICC
내용 조직 점수	.902	.949
글의 구조 점수	.895	.945
언어 사용 점수	.939	.969
총점	.919	.958

2.4. 평가 시스템 개발

쓰기 평가를 위해 ㈜이르테크가 개발한 평가 시스템 KRONOTH 3.0을 최적화하여 사용하였다. 평가 시스템은 평가 결과를 조회하는 용도로 활용하였는데 쓰기 원문, 세부 평가 결과, 피드백을 동시에 확인할 수 있다.



<그림 36> 쓰기 평가 도구 화면 - 평가 자료 목록



<그림 37> 쓰기 평가 도구 화면 - 평가 내용 조회

3. 한국어 학습자 말뭉치 교육 및 홍보

3.1. 말뭉치 구축 실무자 교육

원활한 DB 구축을 위해 말뭉치 수집, 자료 처리, 입력, 구축 도구 사용 방법 등 말뭉치 구축 및 정비 지침 등에 대한 실무 교안을 마련하고 교육 실행 계획을 수립하여 사업 참여자 전원 대상 말뭉치 구축 및 정비 실무 교육을 실시하였다. 참여 인력 교육 지원 및 기술 지원을 위한 사업 수행 개요는 다음과 같다.

○ 일시: 2023년 9월 6일(수) 오후 2시
○ 개최 방식: 강의와 실습, 비대면 Zoom
○ 교육 담당: 김동은(연세대학교, 전 문어 구축 실무 담당자)
○ 교육 대상: (주)이르테크 구축 실무자, 국립국어원 사업 관리팀
○ 교육 프로그램

<표 42> 말뭉치 구축 실무자 교육 프로그램

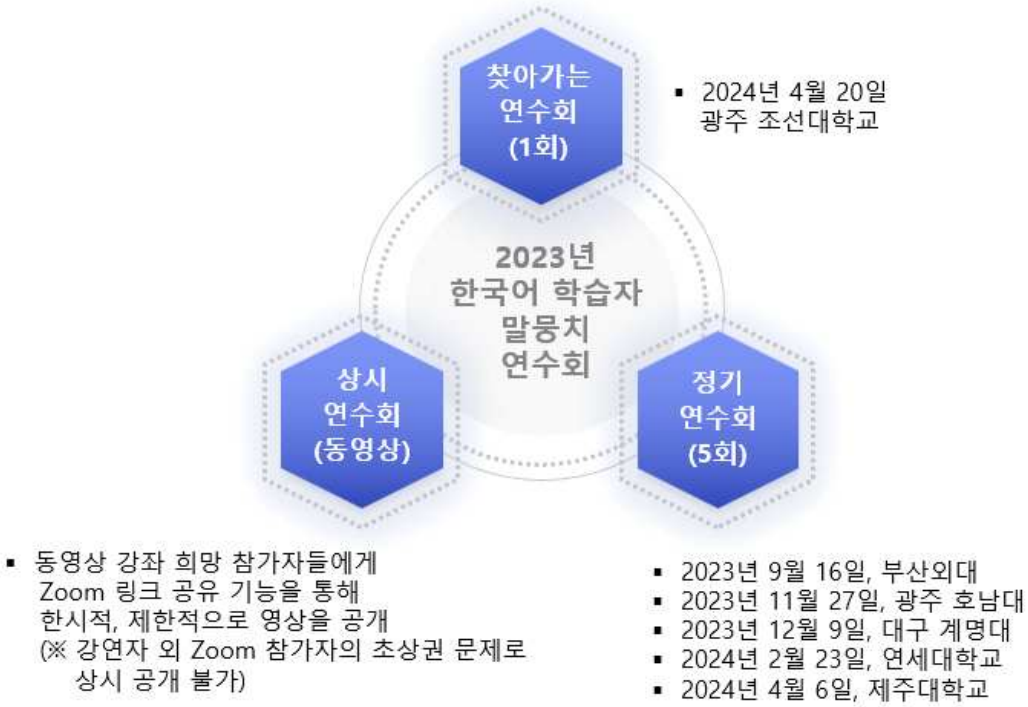
구분	지침 교육	구축 지원 도구 교육 및 실습
기본 교육	▪ 한국어 학습자 말뭉치의 이해 ▪ 한국어 학습자 말뭉치 구축 • 사업의 개요 및 절차 ▪ 개인정보보호, DB 구축 보안 교육	▪ 구축 지원 도구의 체계 및 사용 방법 전반 ▪ 작업 할당 및 접수 등의 절차 ▪ 이력 관리
수집	▪ 자료 수집을 위한 과제 유형 ▪ 학습자 동의서, 메타정보 수집 ▪ 수집 자료 처리 및 관리 방안	
자료 처리	▪ 자료의 분류 ▪ 스캔 파일 변환 ▪ 파일명 부여 체계 ▪ 학습자와 파일에 관한 메타 정보 등록	▪ 스캔 원문 파일 등록, 파일명 생성 등 표본 등록 방법 및 절차

입력	▪ 작문 입력 지침 및 쟁점 ▪ 작문 검수 지침 및 쟁점	▪ 작문 입력 방법 및 절차 ▪ 작문 검수 방법 및 절차
평가	▪ 작문 평가 기준 및 지침 ▪ 작문 평가 검수 방법 및 쟁점	▪ 작문 평가 도구 사용법 ▪ 작문 평가 도구 제약 사항

3.2. 말뭉치 이용자를 위한 연수회 개최

3.2.1. 개최 내용

본 사업에서는 공공 데이터로서 한국어 학습자 말뭉치의 확산과 이용자의 활용 능력을 제고하기 위해 한국어 학습자 말뭉치 이용자를 대상으로 총 6회의 '한국어 학습자 말뭉치 교실'을 개최하였다. 개최 개요는 다음과 같다.



<그림 38> 한국어 학습자 말뭉치 교실 개최 개요

한국어교육 학계 및 교육 분야 외에도 민간 분야에서도 폭넓게 활용할 수 있도록 한국어 학습자 말뭉치 이용자의 저변을 확대하기 위하여 대상, 진행 방식, 프

로그램을 다양하게 구성하여 운영하였다. 회차별 세부 프로그램은 다음과 같다.

<표 43> 한국어 학습자 말뭉치 교실 개최 프로그램

회차	일시 및 장소	프로그램	대상	진행 방식	참가 인원
1차	2023년 9월 16일(토) 15:00-17:00 부산외대	○ 한국어 학습자 말뭉치 구축 현황 소개(한송화, 연구책임자) ○ 말뭉치 기반 연구를 위한 응용 프로그램 활용의 실제 - 학습자 말뭉치, 모어 말뭉치를 대상으로(최정도, 계명대) - 제1부. 강의 학습자 말뭉치, 모어 말뭉치 비교·분석을 위한 연구 방법과 절차, 응용 프로그램 소개 - 제2부. 자료 처리 실습 1. json형식의 파일 변환하기 2. Utagger를 활용한 형태소 분석하기 3. 텍스트 에디터와 정규 표현식을 활용한 말뭉치 자료 전처리하기 4. 안트콘크, 엑셀을 활용한 형태소, 단어 등의 목록, 빈도 추출과 분포 특성 비교·분석하기	한국어교육 연구자 및 교사	대면+비대면(강의 + 실습)	대면 30명 비대면 240명
2차	2023년 11월 27일(금) 15:00-17:00 광주 호남대	○ 한국어 학습자 말뭉치 구축 현황 소개 (한송화, 연구책임자) ○ 한국어 교수·학습을 위한 학습자 말뭉치의 활용 - 앤트콘크를활용한 자료 검색과 추출의 기초(한승규, 호남대) - 제1부. 강의 1. 앤트콘크의주요 기능 2. 한국어 학습자 말뭉치와 앤트콘	한국어교육 연구자 및 교사	대면+비대면(강의 + 실습)	대면 20명 비대면160명

회차	일시 및 장소	프로그램	대상	진행 방식	참가 인원
		크: 교수·학습에서의 활용 - 제2부. 활용 실습 1. 앤트콘크활용을 위한 자료 전처리하기 2. 앤트콘크를활용한 자료 검색과 분석하기 1) 문맥 색인 검색하기 2) 어휘 목록 추출하기 3) 연어 추출하기 4) 문법 표현 추출하기 5) 키워드 추출하기 3. 엑셀을 활용한 형태소, 단어 등의 빈도 추출과 분포 특성 비교·분석하기			
3차	2023년 12월 9일(토) 11:00-13:00 대구 계명대	○ 한국어 학습자 말뭉치 구축 현황 소개 (한송화, 연구책임자) ○ 학습자 말뭉치를 활용한 한국어교육 연구의 설계와 실천: 한국어 학습자의 표현문형 사용 양상 분석을 중심으로 (강범일, 연세대) - 제1부 강의 1. 연구 설계: 연구의 흐름 2. 연구 방법론 설계 - 제2부 실습 1. 표현문형 추출하기 2. 수준별 분포 분석하기 1) 빈도 2) 다양도 3. 수준 표현문형 분석하기	한국어교육 연구자 및 교사	대면+비대면 (강의 + 실습)	대면 30명 비대면140명
4차	2024년 2월 23일(금) 15:00-17:30	○ 한국어 학습자 말뭉치 구축 현황 소개 (한송화, 연구책임자)	한국어교육 연구자 및	대면+비대면 (강의 +	대면 30명 비대

회차	일시 및 장소	프로그램	대상	진행 방식	참가 인원
	연세대학교	○ 언어 교육 서비스 개발과 학습자 말뭉치의 활용 - 사회: 김한샘(공동연구원) - 토론: 박전규(튜터러스랩스대표) 주하돈(케이토픽대표) 곽용진(이르테크대표) 함영균(테디썸대표)	산업계 관계자	실습)	면140명
5차	2024년 4월 6일(토) 10:00-14:00 제주대학교 교육대학 사라캠퍼스	○ 한국어 학습자 말뭉치 활용 현황 (홍혜란, 공동연구원) ○ 한국어 교수자와 연구자를 위한 학습자 말뭉치의 검색과 활용 (김한샘, 공동연구원) - 제1부. <학습자 말뭉치 나눔터> ㅣ 활용하기 1. <학습자 말뭉치 나눔터>의 자료 활용을 위한 자료 구축 과정 및 자료 구조의 이해 2. <학습자 말뭉치 나눔터>의 검색 메뉴를 활용한 자료 검색과 활용 - 제2부. 검색 도구 활용하기 1. 앤트콘크(Ant Conc)를 활용한 자료 검색과 활용 2. 텍스트 에디터를 활용한 자료 검색과 활용	한국어교육 연구자 및 교사	대면+비대면 (강의 + 실습)	대면 25명 비대면130명
6차	2024년 4월 20일(토) 10:00-11:50 광주 조선대학교	○ 한국어 학습자 말뭉치 자료를 활용한 교육 자료 개발 : 학습자 오류 말뭉치를 활용한 어휘/문법 교수 자료 만들기(한송화, 연구책임자) 1. 한국어 학습자 말뭉치 자료를 활용한 교육 자료 만들기 1) <한국어 학습자 말뭉치 나눔터> 검색 메뉴와 자료 활용 방법 2) <한국어 학습자 말뭉치 나눔터>	한국어 교수자및 학습자	대면+비대면 (강의 + 실습)	대면 20명 비대면120명

회차	일시 및 장소	프로그램	대상	진행 방식	참가 인원
		오류 말뭉치를 활용한 교육 자료 만들기 2. 교육 자료 개발을 위한 <한국어 학습자 말뭉치> 검색 심화 연습하기 1) <한국어 학습자 말뭉치 나눔터>의 검색 메뉴를 활용한 자료 검색과 정렬, 내려받기 2) 엑셀을 사용하여 검색 파일 통합하기 3) 모어화자와학습자의 언어 사용 양상 비교를 위한 모어화자말뭉치 활용하기			

3.2.2. 참가자 의견수렴

본 사업에서는 한국어 학습자 말뭉치 이용자의 요구를 반영하여 보다 나은 교육 프로그램을 개발해 나갈 수 있도록 2023년 한국어 학습자 말뭉치 교실 참가자를 대상으로 교육 프로그램에 대한 만족도 조사와 함께 앞으로 듣고 싶은 프로그램에 관한 의견을 수렴하였다. 응답자는 총 147명이었으며, 조사 대상 및 항목, 응답 결과는 다음과 같다.

1) 조사 대상 및 항목

① 조사 대상: 2023년 한국어 학습자 말뭉치 교실 참가자 전원

② 조사 항목

○ 소속, 직위, 참석한 연수회

○ 프로그램 만족도

- 프로그램 운영 방식(대면+비대면)의 효율성

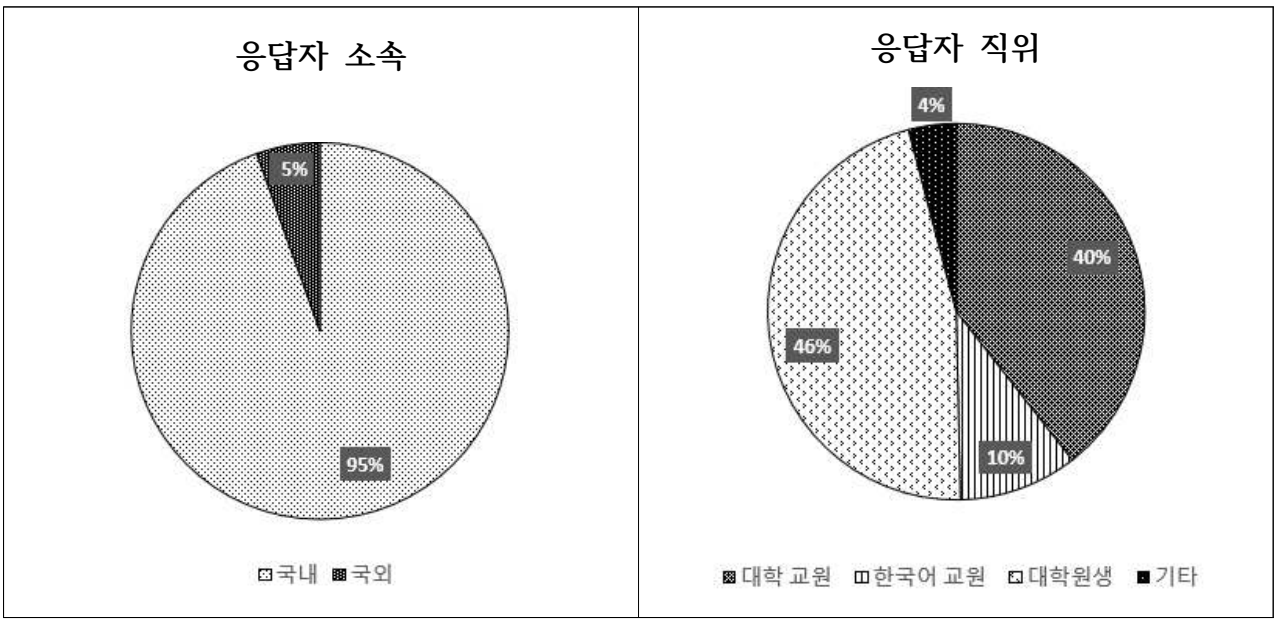
- 강의 방식(강의+실습)의 적절성
- 강의 시간의 적절성
- 강의 난이도
- 실습 난이도
- 참가 후 말뭉치 이해 능력 향상도
- 참가 후 말뭉치 활용 능력 향상도

③ 향후 학습자 말뭉치 교실에서 듣고 싶은 강의 주제

④ 학습자 말뭉치 교실에 대한 자유 의견

2) 조사 결과

① 소속 및 직위



<그림 39> 한국어 학습자 말뭉치 교실 참가자 의견수렴 결과: 소속 및 직위

응답자는 총 147명으로 국내 기관 소속이 139명, 국외 기관 소속이 8명이었다. 직위는 대학 교원이 58명, 한국어 교원 15명, 대학원생 68명, 기업체를 포함한 기타가 6명이었다.

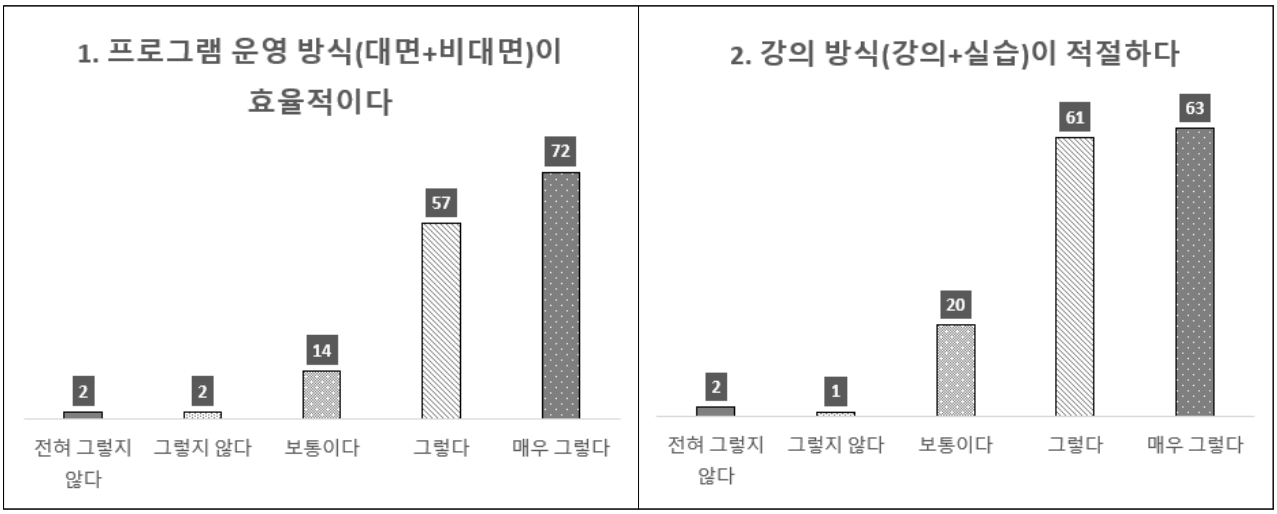
② 참석한 연수회

응답자가 참석한 연수회는 제1차가 38명, 제2차 38명, 제3차 25명, 제4차 35

명, 제5차 53명, 제6차 41명이었다.

③ 프로그램 만족도

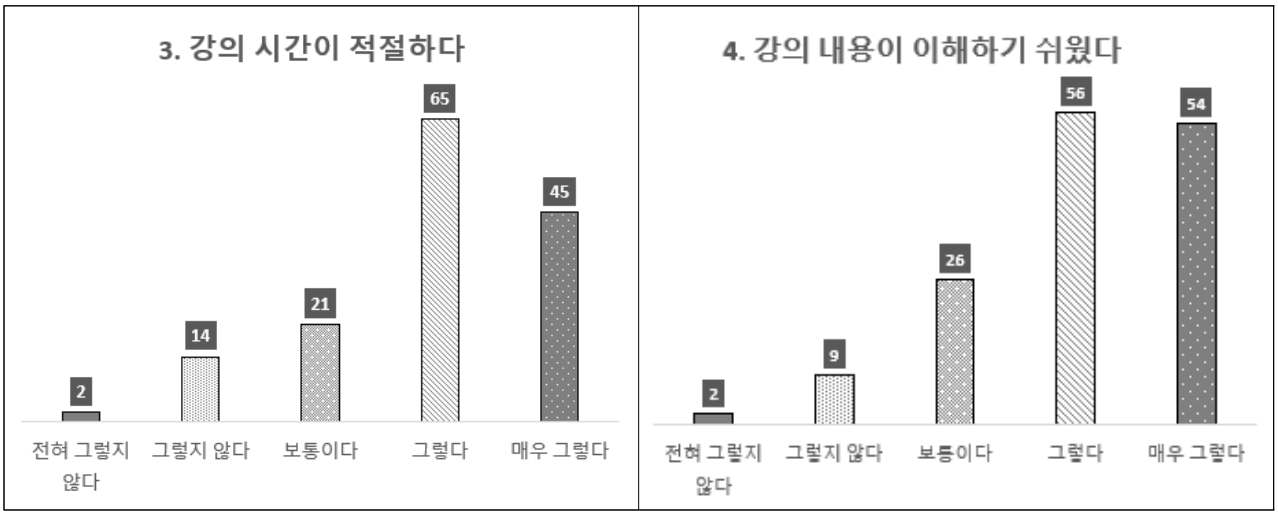
프로그램 만족도 조사는 총 7개 항목에 대하여 조사가 진행되었는데, 모든 항목에 대하여 보통 수준 이상으로 만족하는 것으로 나타났다.



<그림 40> 한국어 학습자 말뭉치 교실 참가자 의견수렴 자료 1(단위: 명)

첫 번째, 프로그램 운영 방식(대면+비대면)의 효율성을 묻는 질문에 대해서는 '매우 그렇다' 72명(49.0%), '그렇다' 57명(38.8%), '보통이다' 14명(9.5%), '그렇지 않다' 2명(1.4%), '전혀 그렇지 않다' 2명(1.4%)로 나타났다. 이를 통해 현행 운영 방식이 참가자들의 상황에 따라 참가 방식을 선택하게 함으로써 참가 기회를 확대할 수 있는 효율적인 방식임을 확인할 수 있었다.

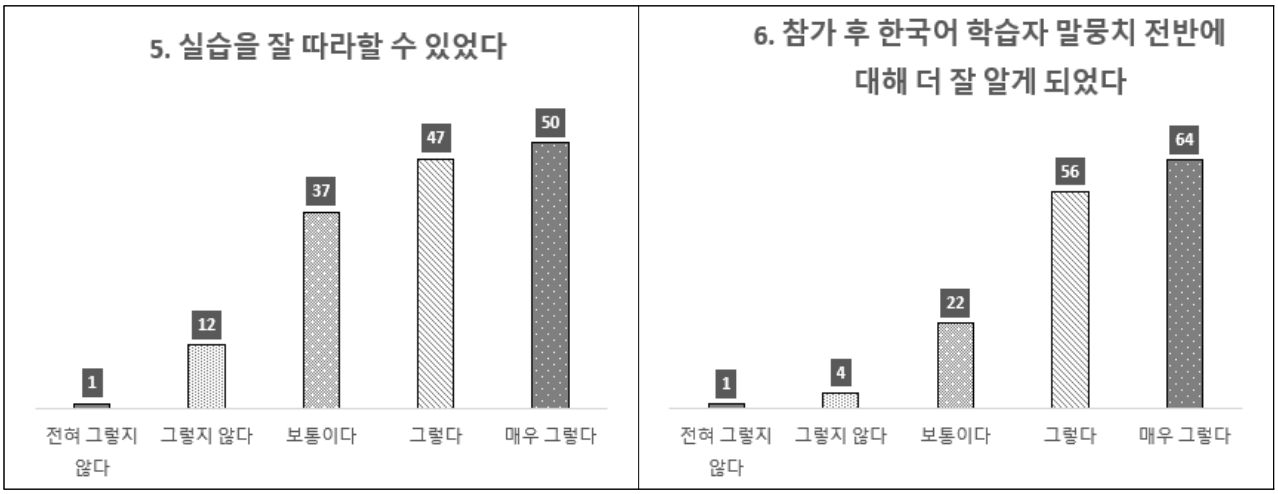
두 번째, '강의 방식(강의+실습)의 적절성을 묻는 질문에 대해서는 '매우 그렇다' 63명(42.9%), '그렇다' 61명(41.5%), '보통이다' 20명(13.6%), '그렇지 않다' 1명(0.7%), '전혀 그렇지 않다' 2명(1.4%)으로 나타나, 강의와 실습을 병행하는 강의 방식 또한 한국어 학습자 말뭉치 교실 운영에 적절한 것으로 확인되었다.



<그림 41> 한국어 학습자 말뭉치 교실 참가자 의견수렴 자료 2(단위: 명)

세 번째, 강의 시간의 적절성을 묻는 문항에 대해서는 '매우 그렇다' 45명(30.6%), '그렇다' 65명(44.2%), '보통이다' 21명(14.3%), '그렇지 않다' 14명(9.5%), '전혀 그렇지 않다' 2명(1.4%)으로 나타났다. 이는 다른 항목에 비해 근소하게 만족도가 낮았는데, 통상 2시간 내외의 시간에 이론 강의와 실습을 병행하는 것이 프로그램에 따라서는 촉박한 경우가 있었기 때문인 것으로 보인다. 이러한 결과를 통해 향후에는 프로그램의 내용이나 난이도를 고려하여 교육 시간을 유연하게 조정할 필요가 있음을 알 수 있었다

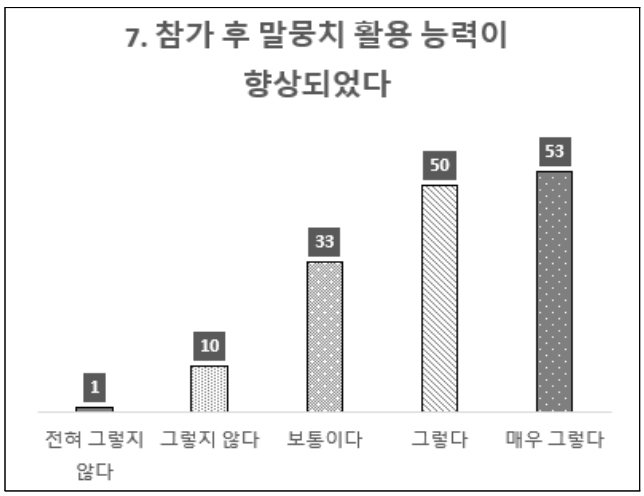
네 번째, 강의 내용의 난이도를 묻는 질문에 대해서는 '매우 그렇다' 54명(36.7%), '그렇다' 56명(38.1%), '보통이다' 26명(17.7%), '그렇지 않다' 9명(6.1%), '전혀 그렇지 않다' 2명(1.4%)으로 나타나, 참가자에 따라 느끼는 난이도에 다소 차이가 있음을 알 수 있었다. 이는 참가자들의 컴퓨터 활용 능력이나 말뭉치 활용 경험이 영향을 미치는 부분으로, 현행 프로그램에서와 같이 난이도에 따라 기초 과정과 심화 과정을 구분하여 운영하는 것이 지속될 필요가 있음을 알 수 있었다.



<그림 42> 한국어 학습자 말뭉치 교실 참가자 의견수렴 자료 3(단위: 명)

다섯 번째, 실습 난이도를 묻는 질문에 대해서는 '매우 그렇다' 50명(34.0%), '그렇다' 47명(32.0%), '보통이다' 37명(25.2%), '그렇지 않다' 12명(8.2%), '전혀 그렇지 않다' 1명(0.7%)으로 나타나, 강의 내용의 난이도에 대한 응답과 비슷한 양상을 보였다.

여섯 번째, 참가 후 한국어 학습자 말뭉치의 이해도를 묻는 질문에 대해서는 '매우 그렇다' 64명(43.5%), '그렇다' 56명(38.1%), '보통이다' 22명(15.0%), '그렇지 않다' 4명(2.7%), '전혀 그렇지 않다' 1명(0.7%)으로 나타났다. 이를 통해 한국어 학습자 말뭉치 이용자의 저변 확대 및 활용도 제고를 위하여 교육 프로그램이 지속적으로 운영되어야 함을 확인할 수 있었다.



<그림 43> 한국어 학습자 말뭉치 교실 참가자 의견수렴 자료 4(단위: 명)

마지막으로, 교육 프로그램 참가 후 말뭉치 활용 능력 향상을 묻는 질문에 대해서는 ‘매우 그렇다’ 53명(36.1%), ‘그렇다’ 50명(34.0%), ‘보통이다’ 33명(22.4%), ‘그렇지 않다’ 10명(6.8%), ‘전혀 그렇지 않다’ 1명(0.7%)으로 나타났다. 일회적인 참가로 말뭉치 활용 능력이 향상되기 어려움을 생각할 때, 이러한 결과는 향후 참가자들의 의견을 지속적으로 수렴하여 다양한 요구를 반영한 프로그램이 운영될 필요가 있음을 보여 준다고 하겠다.

④ 향후 한국어 학습자 말뭉치 교실에서 듣고 싶은 강의 주제

향후 한국어 학습자 말뭉치 교실에서 듣고 싶은 강의 주제는 주로 말뭉치를 활용한 연구 방법론과 함께, 실질적인 활용에 필요한 도구 사용 방법, 활용 사례와 방법 등에 대한 요구가 큼을 알 수 있었다. 다음은 참가자들의 응답을 유형화한 것이다.

<표 44> 한국어 학습자 말뭉치 교실에서 듣고 싶은 강의 주제

번호	듣고 싶은 강의 주제	응답자 수
1	말뭉치를 활용한 연구 방법론	33
2	Python, R, AntConc, Text Editor, Excel 등 다양한 프로그램 및 툴 사용법	29
3	말뭉치 실제 활용 사례	18
4	말뭉치 활용 방법	12
5	교수 자료 제작 방법	10
6	이전 강의를 다시 듣고 싶음	7
7	학습자 오류 분석 연구 방법	6
8	개인 말뭉치 구축 및 주석 방법	5
9	주제 상관없이 다 듣고 싶음	5
10	구어말뭉치 활용 강의	3
11	드라마 대본 등의 자료를 활용한 말뭉치 분석	2
12	잘 모르겠음	2
13	옛 자료를 활용한 말뭉치 처리	1

⑤ 자유 의견

그 밖에 자유 의견으로 ‘프로그램이 유익함’, ‘실습을 위한 강의 시간 연장의

필요성', '녹화본 및 강의 자료 공유', '초·중·고급 수준별 운영 필요성', '말뭉치 교실 개최 횟수 확대' 등에 대한 의견이 있었다. 다음은 참가자들의 응답을 유형화한 것이다.

<표 45> 자유 의견

번호	의견	응답자 수
1	유익한 시간이었음	23
2	실습을 위한 강의 시간 연장 필요	18
3	녹화본 및 강의 자료 공유 필요	10
4	초·중·고급 수준별 운영 필요	9
5	말뭉치 교실 개최 횟수 확대	7
6	비대면 방식이 효율적임(비대면 전용 특강 마련 필요)	9
7	쉽고 친절한 설명이 좋았음	7
8	프로그램을 직접 활용하고 응용해 볼 수 있어서 좋았음	6
9	대면과 비대면 강의 구분 필요	5
10	말뭉치 자료의 활용 및 접근성 확대 필요	3
11	소규모로 진행되면 좋겠음	3
12	속도가 빨라서 따라가기 어려움	3
13	연수나 워크숍같은 집중 교육 필요	3
14	개최 시간 변경 필요	2

3.3. 말뭉치 소개 · 활용 자료집 현행화 및 배포

본 사업에서는 <학습자 말뭉치 나눔터>에 게시된 '학습자 말뭉치 활용 매뉴얼'을 매 회차의 학습자 말뭉치 교실 참가자들에게 홍보하고 배포하여 이용을 활성화하였다. 아울러 총 6회의 2023년 학습자 말뭉치 교실 강의 자료를 참가자들에게 배포하여 교육 이후에도 참가자 스스로 활용할 수 있도록 하였다. 이들 자료는 선별과 재가공을 거쳐 향후 학습자 말뭉치 활용 심화 자료를 개발하는 데에 활용할 수 있을 것이다.

제 3 장

결론

1. 사업 결과 요약

본 사업은 한국어 학습자 말뭉치 제2차 중장기 계획의 일환으로 국가 언어 자원으로서의 말뭉치 규모를 확대하고, 언어 자원 활용 활성화 기반을 마련하는 것을 목표로 추진되었다. 본 사업의 주요 과업과 결과는 다음과 같다.

○ 한국어 학습자 말뭉치 구축

한국어 학습자 쓰기 자료 15,779개를 문어 원시 말뭉치로 구축하였다. 어절 수로 환산하면 2,085,629어절로, 이는 한국어 학습자 말뭉치 중장기 계획의 구축 목표인 1,000만 어절 중 약 20%에 해당하는 구축량으로 양적으로 큰 성과를 이루었다고 할 수 있다. 특히 기획 수집을 진행하여 쓰기 장르의 균형성을 보완하였다는 점, 전체 구축량의 53%를 국외 자료로 구축함으로써 기존에 1.2%였던 국외 수집량을 17.8%로 확대하였다는 점에서 2023년 학습자 말뭉치 구축의 의의가 있다.

말뭉치 사업을 설계하고 시작할 당시 국내 기관이 우선적으로 수집 대상이 되었으므로 사업 후반부의 국내에서 국외로의 수집 기관 이동은 자연스러운 현상이라고 할 수 있다. 국외 학습자의 쓰기 자료 규모를 확대함에 따라 학습자의 수준과 언어권 통계에도 변화가 있었다. 기존 학습자 말뭉치는 주로 국내 학습자를 대상으로 자료를 수집했기 때문에 국내 학습자의 구성 비율에 영향을 받았다. 그 결과 중국어권 학습자의 자료 비율이 월등히 높았으므로 2023년 사업에서는 중국어권 외 다른 언어권의 자료를 수집하고자 베트남, 방글라데시, 태국 등의 국가에서 쓰기 자료를 확대 수집하였다. 이에 따라 전체 말뭉치에서 중국어권 자료가 차지하는 비중은 기존 46.2%에서 약 9% 감소한 36.9%로 조정되고 다른 언어권 자료의 비중이 증가하였다.

하지만 국외 한국어 학습자는 대부분 초중급 수준이라는 점에서 학습자 수준별 자료를 고르게 구축하기에는 어려움이 있었다. 전체 수집 자료의 47%는 국내 학습자의 쓰기 자료로 구축되었지만 국내 기관의 학습자 구성에서 고급 학습자가 차지하는 비중이 적기 때문에 2023년에 구축된 고급 학습자의 자료는 12.2%에 그쳤다. 향후 말뭉치 구축 사업에서는 고급 학습자의 쓰기 자료를 확대 구축할 수

있는 방안이 고려되어야 할 것이다. 또한 국외 학습자의 자료 비중이 높아진 만큼 국외 학습자의 세부 유형과 수준에 대한 상세 메타 정보를 추가한다면 말뭉치를 보다 정교하게 구축할 수 있을 것이다.

○ 한국어 학습자 쓰기 평가 자료 구축

주제와 장르가 서로 다른 2종의 쓰기 자료 각 1,500건씩 총 3,000건의 작문을 대상으로 한국어 학습자 쓰기 평가 자료를 구축하였다. 이를 위해 등급 체계 및 평가 구인, 평가 기준 설정의 바탕이 되는 기초 연구를 수행하고, 내용 조직, 글의 구조, 언어 사용의 세 가지 구인을 설정하고 각 구인별로 척도와 점수를 부여하는 분석적 채점을 실시하였다. 그 결과 등급별 점수(100점 만점), 1-6급의 누적 점수(600점 만점), 최종 점수(100점 만점)로 3종의 점수와 함께, '국제 통용 한국어 교육 표준 모형'을 기준으로 한 등급, 한국어교육 전문가의 총평이 부여되었다. 이러한 성과는 말뭉치 구축의 측면에서 기구축 말뭉치에서 쟁점이 되었던 등급 표준화의 기반을 마련하였다는 점에서도 의의가 있다. 아울러 무엇보다도 자료 활용의 측면에서 최근 학계는 물론 민간 기업 등에서 주목받고 있는 자동 채점, 자동 첨삭 등의 기술 개발을 위한 기반 자료를 마련하여 공공 데이터로서 연구 및 교육, 산업계에서의 활용도를 제고하였다는 점에서 의미가 크다. 이처럼 말뭉치 구축과 활용의 두 가지 축에서 큰 의미가 있는 작업이니만큼 활용도 제고를 위해 평가 자료의 규모를 확대할 필요가 있다. 또한 학습자 자료의 경우 주제, 장르와 같은 자료 변인, 숙달도 수준, 국적 등과 같은 학습자 변인 등에 따라 다양한 특성을 지닌 자료가 산출될 수 있다는 점을 고려하여 주제, 장르 등을 세분화해야 할 것이다.

○ 한국어 학습자 말뭉치 교육 및 홍보

한국어 학습자 말뭉치 교육 및 홍보는 한국어 학습자 말뭉치의 체계적인 구축, 성과의 확산과 활용도 제고를 목적으로 한다. 본 사업에서는 자료 수집과 구축을 위한 실무자 교육을 실시하고, 한국어 학습자 말뭉치 이용자를 대상으로 총 6회의 '한국어 학습자 말뭉치 교실'을 개최하였다. '한국어 학습자 말뭉치 교실'은 한국

어교육 학계 및 교육 분야 외에도 민간 분야로의 저변 확대를 위하여 교육 대상, 진행 방식, 프로그램을 다양하게 구성하여 운영하였다. 또한 보다 나은 교육 프로그램을 개발해 나갈 수 있도록 2023년 한국어 학습자 말뭉치 교실 참가자를 대상으로 교육 프로그램에 대한 의견을 수렴하였다. 그 결과 2023년에 개최한 '한국어 학습자 말뭉치 교실'에 대한 만족도와 함께 개선 방향, 새로운 프로그램에 대한 아이디어를 얻을 수 있었다. 그 외에도 <학습자 말뭉치 나눔터>에 게시된 '학습자 말뭉치 활용 매뉴얼'을 매 회차의 학습자 말뭉치 교실 참가자들에게 홍보하고 배포하여 이용을 활성화하였다. '한국어 학습자 말뭉치 교실'은 2015년 한국어 학습자 말뭉치 구축 사업 착수 이후 현재까지 매해 진행해 오면서 학계의 주목을 받으며 한국어교육 연구자 및 교원들의 말뭉치 활용 역량 강화, 학술지 논문 및 석박사 학위논문을 통한 한국어 학습자 말뭉치 활용 연구 결과물의 축적이라는 성과를 이끌어 냈다. 그간 참가자들의 관심과 호응, 금번 실시한 의견수렴을 통해 한국어 학습자 말뭉치 교실 확대에 대한 요구가 확인된 바, 앞으로 더욱 세분화된 참가자의 요구를 반영한 다양한 프로그램을 지속적으로 만들어 가야 할 필요가 있다. 아울러 국내뿐만 아니라 일본, 중국, 베트남 등 한국어교육이 활성화된 국외 지역에서의 요구에 부응하여 개최 지역을 확장해 나갈 필요가 있다.

2. 사업의 의의

언어 자원 활용의 활성화 기반을 마련하기 위해 국가 주도의 한국어 학습자 말뭉치 제2차 중장기 계획이 수립되었고 2015년부터 2025년에 걸쳐 총 1,000만 어절의 한국어 학습자 말뭉치 자료 구축을 목표로 하고 있다. 그 결과 2023년 사업까지 약 835만 어절의 학습자 말뭉치가 구축되었다.

한국어 학습자의 언어를 관찰하기 위해서는 체계적으로 설계되고 구축된 학습자 산출 자료가 적정 규모만큼 필요하다. 한국어 학습자 말뭉치는 국가 주도의 공공 자료로서 자유로운 배포와 폭넓은 활용이 가능한 한국어 학습자 말뭉치이다. 이를 위해 학습자의 개인정보 수집·이용 동의 및 저작권 이용 허락 동의를 체결하고, 말뭉치의 균형성을 보완함으로써 실제적인 활용도를 높이고자 하였다. 한국어 학

습자 말뭉치는 한국어 학습자 말뭉치 나눔터(https://kcorpus.korean.go.kr/)를 통해 누구나 접근이 가능하다.

최근 인공 지능 기술 발전에 따라 한국어 교육 분야에서도 언어 자원의 수요가 증대하고 있다. 한국어 학습자 말뭉치는 빅데이터로의 일부로 자연언어 처리, 에듀테크 기술 개발 등의 원천 자료로 사용되어 여러 연구 사업에 이바지할 것으로 기대된다.

3. 제언

○ 다양한 언어권 학습자의 원시 말뭉치 확보

한국어 학습자 말뭉치는 연구와 교육, 산업계 등에서 폭넓게 활용된다. 특히 언어권별 자료 산출 결과를 분석할 수 있다는 점에서 학습자의 중간 언어 관찰에 용이하며, 분석 결과는 해당 언어권의 한국어 교재 개발이나 연구, 보조 자료 개발 등에 활용될 수 있다.

사업의 초반부에는 학습자의 실제 분포나 연구자의 요구 등을 검토하고, 자료의 효용성 및 수집의 용이성 등을 고려하여 국내 학습자의 자료를 우선적으로 수집하였다. 그 결과 자연히 국내 학습자의 구성 비율에 영향을 받게 되었고, 특정 언어권 학습자의 비중이 높게 집계되었다. 향후 사업에서는 국외 학습자의 수집 비중을 늘리고, 이에 따라 다양한 언어권 학습자의 원시 말뭉치를 확보하려는 노력이 필요하다. 다만 국외 학습자의 경우, 고급 학습자의 비중이 적다는 점을 감안해야 하므로 언어권별, 수준별 균형성을 함께 고려하여 수집 설계를 해야 할 것이다.

○ 학습자 말뭉치의 특성을 고려한 다양한 성격의 하위 말뭉치 구성

2023년 사업에서는 EPS-TOPIK 학습자 및 국외 학습자의 원시 말뭉치 수집 비중을 확대하였다. 국내 인구 감소 문제로 외국인 고용허가제가 지속적으로 확대되면서 EPS-TOPIK 학습자의 규모 또한 확대되고 있다. 한편, 국외 대학의 학습자의 경우, 학년으로 반이 나뉘어져 있고, 한국어능력시험(TOPIK)을 치르지 않은 학

습자가 많아 국제 통용 한국어 표준 교육과정에서 제시하는 1급-6급의 분류 체계를 명확하게 제공하기 어려운 경우가 적지 않았다.

국내의 일반 목적이나 학습 목적 학습자 이외, 기타 학습자들의 말뭉치 참여 비중이 높아진 만큼 향후 사업에서는 이들의 특성을 고려한 다양한 성격의 하위 말뭉치를 구성하여 사용자가 학습자 특성에 맞는 연구 및 개발을 할 수 있도록 하는 전문적인 연구가 필요하다.

○ 이용자 편의를 위한 형태 분석·오류 분석 말뭉치 확대

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

<부록>

부록1. 학습자 동의서 양식
부록2. 2023년 한국어 학습자 말뭉치 수집 지침
부록3. 2023년 한국어 학습자 말뭉치 구축 지침

부록1. 학습자 동의서 양식

한국어 학습자 말뭉치 구축·활용을 위한 학습자 언어 자료 이용 허락서, 개인정보 수집·이용 및 제3자 제공 동의서

국립국어원에서는 한국어교육의 질적 향상을 위해 한국어 학습자의 언어 자료(말뭉치)를 수집하여 활용하는 사업을 추진하고 있습니다. 여러분이 제공한 자료는 한국어 교수 방법 개선, 한국어 교재 개발, 한국어 교육 및 인공지능 모델 개발 등 다양한 분야의 연구·개발에 사용됩니다. 이 연구에 참여하는 분들은 경제적인 손해나 신체적 위험이 없습니다. 만약 참여를 원하지 않을 때에는 참여 의사를 철회할 수 있습니다. 또한 수집하는 개인정보는 본 사업의 목적 외로는 사용되지 않으며, 비밀 유지를 위하여 식별할 수 없는 형태로 사용될 것입니다. 내용을 자세히 읽으신 후 동의 여부를 결정하여 주세요.

■ 학습자 자료 이용 허락

한국어 학습자 말뭉치 구축·활용을 위해 쓰기 자료를 제공하고, 쓰기 원문 자료 전체의 공개와 연구·개발 목적의 이용을 허락합니다.

☞ **위와 같이 학습자 언어 자료의 이용 허락에 동의하십니까?** □**동의** □**동의하지 않음**

■ 개인정보 수집·이용 내역

항목	수집.이용 목적	보유·이용기간
성별, 출생년, 현재 TOPIK 등급, 국적, 제1언어, 한국어 학습기간, 한국 거주기간, 한국어 학습목적, 직업, 한국어 외 사용 가능 외국어	국립국어원 한국어 학습자 말뭉치 구축·활용	**개인정보 수집·이용 동의일로부터 30년** * 단, 개인정보 폐기 요청 시 즉시 삭제

※ 위의 개인정보 수집·이용에 대한 동의를 거부할 권리가 있습니다. 그러나 동의를 거부할 경우 한국어 학습자 말뭉치 구축·활용 사업 참여에 제한을 받을 수 있습니다.

☞ **위와 같이 개인정보를 수집·이용하는 데 동의하십니까?** □**동의** □**동의하지 않음**

■ 개인정보 제3자 제공 내역

제공받는 자	제공목적	제공항목	보유, 이용기간
국립국어원 한국어 학습자 말뭉치 이용 약정자	**한국어 정보처리 연구·개발, 국어 연구·개발**	성별, 출생년, 현재 TOPIK 등급, 국적, 제1언어, 한국어 학습기간, 한국 거주기간, 한국어 학습목적, 직업, 한국어 외 사용 가능 외국어	**개인정보 수집·이용 동의일로부터 30년** * 단, 개인정보 폐기 요청 시 즉시 삭제

※ 위의 개인정보 수집·이용에 대한 동의를 거부할 권리가 있습니다. 그러나 동의를 거부할 경우 한국어 학습자 말뭉치 구축·활용 사업 참여에 제한을 받을 수 있습니다.

☞ **위와 같이 개인정보를 제3자 제공하는 데 동의하십니까?** □**동의** □**동의하지 않음**

■ 개인정보 취급 업무 위탁 내역(고지사항)

위탁 받는 자(수탁업체)	업무내용
㈜이르테크·연세대학교 산학협력단	2023년 한국어 학습자 말뭉치 구축

년 월 일

본인 성명 (서명 또는 인)

국립국어원 귀중

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. 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)
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)

Letter of permission to use learners' language materials for constructing and utilizing Korean language learner corpus / Consent to collection and use of personal information and provision to third parties

The National Institute of Korean Language promotes a project that collects language data (corpus) from Korean learners and uses this set to improve the quality of Korean education. The materials one provides are used for research and development in various fields such as improving Korean language teaching methods, developing Korean language textbooks, Korean language education, and AI model development. Participants of this project experience no economic or physical harm. One may choose to withdraw their participation if required. The personal information of participants will not be used for any other purpose, and any identifiable information will remain confidential. Please read the contents carefully and decide whether you agree.

■ **Permission to use learner's materials**

Writing materials are provided to build and utilize the Korean language learner corpus, and the disclosure of all original writing materials and their use for research and development purposes is permitted.

☞ **Do you agree to the permission to use learners' language materials as above?**

☐**Agree** ☐**Disagree**

■ **Details of collection and use of personal information**

Items	Purpose of collection and use	Retention/use period
Gender, Year of Birth, Current Level of Korean, Nationality, Native Language, Time of Studying Korean Language, Time of Residence in Korea, Purpose of Studying Korean Language, Occupation, Any Other Foreign Language That You Can Speak	National Institute of Korean Language Korean learner corpus construction and utilization	**30 years from the date of consent to the collection and use of personal information** * However, if personal information is requested to be disposed of, it will be deleted immediately.

※ You have the right to refuse consent to the collection and use of personal information above. However, if you refuse to provide consent, you may be restricted from participating in the Korean learner corpus construction/utilization project.

☞ **Do you agree to the collection and use of personal information as above?**

☐**Agree** ☐**Disagree**

■ **Details of provision of personal information to third parties**

Recipient	Purpose of provision	Provided items	Retention/use period
Contractor for use the Korean language learner corpus of National Institute of Korean Language	**Korean language information processing research and development, Korean language research and development**	Gender, Year of Birth, Current Level of Korean, Nationality, Native Language, Time of Studying Korean Language, Time of Residence in Korea, Purpose of Studying Korean Language, Occupation, Any Other Foreign Language That You Can Speak	**30 years from the date of consent to the collection and use of personal information** * However, if personal information is requested to be disposed of, it will be deleted immediately.

※ You have the right to refuse consent to the collection and use of personal information above. However, if you refuse to provide consent, you may be restricted from participating in the Korean learner corpus construction/utilization project.

☞ **Do you agree to provide personal information to third parties as above?**

☐**Agree** ☐**Disagree**

■ **Details of consignment of personal information handling (Notice)**

Consignee (consignment company)	Business details
IIR TECH Inc. / Yonsei University-Industry Foundation	Construction of the Korean learner corpus in 2023

Date: _____

Name: _____ (Signature)

To the National Institute of Korean Language

The questionnaire includes your personal information that would help us conduct the research analysis. Strict confidentiality will be maintained at all times. (Please fill out the form in Korean if you can. This is also recommended in English.)

1. Gender: ☐ F ☐ M

2. Year of Birth: _____ (eg. 1989)

3. Current Level of Korean: _____ (TOPIK:)

4. Nationality : _____

(※ I am a/an Overseas Korean ☐ Foreigner ☐)

5. Native Language : _____

6. Time of Studying Korean Language (How long you've been studying?) :

_____ years _____ months (eg. 1year 8months)

7. Time of Residence in Korea How long you've been staying in Korea?) :

_____ years _____ months (eg. 1year 10months)

8. Purpose of Studying Korean Language :

☐ for admission to the universities in Korea ☐ for a job

☐ for residence in Korea ☐ for a hobby ☐ for marriage

☐ others ( )

9. Occupation : _____

10. Any Other Foreign Languages That You Can Speak

(Note: If there are multiple languages, please list them in the order of fluency) :

_____

(영어)

为构建和利用韩国语学习者语料库的学习者语言资料使用许可书，
个人信息收集、使用和第三方提供同意书

为了提高韩国语教育的质量，国立国语院正在推进收集和利用韩国语学习者语言资料(语料库)的项目。您提供的资料将用于改善韩国语教学方法、开发韩国语教材、韩国语教育以及人工智能模型等多个领域的研究和开发。参与此研究的人员不会遭受经济损失或人身伤害。如果您不愿意参与，可以随时撤回参与意愿。此外，收集的个人信息将不会用于本项目目的以外的用途，并将以无法识别的形式使用以便保密。请仔细阅读内容后决定是否同意。

■ 学习者资料的使用许可

为了韩国语学习者语料库的建构和利用，提供写作资料，并允许公开完整的写作原文资料并以研究开发为目的加以利用。

☞ **您是否同意如上所述使用学习者语言资料的使用许可？　□同意　□不同意**

■ 个人信息收集及使用详情

项目	收集和使用的目的	保留/使用期限
性别、出生年份、现TOPIK等级、国籍、第一语言、韩国语学习时间、韩国居住时间、韩国语学习目的、职业、韩国语以外其他可用外语	韩国国立国语院韩国语学习者语料库的构建与利用	自同意收集和使用个人信息之日起 30 年 *但要求销毁个人信息时，将立即删除

※　您有权拒绝上述个人信息收集和使用的同意。但如果拒绝同意，您可能会被限制参与韩国语学习者语料库的构建与利用项目。

☞ **您是否同意如上所述个人信息的收集和使用？　□同意　□不同意**

■ 向第三方提供个人信息的详情

接受者	提供目的	提供项目	保留/使用期限
韩国国立国语院韩国语学习者语料库使用承包商	韩国语信息处理研究开发， 韩国语研究开发	性别、出生年份、现TOPIK等级、国籍、第一语言、韩国语学习时间、韩国居住时间、韩国语学习目的、职业、韩国语以外其他可用外语	自同意收集和使用个人信息之日起 30 年 *但要求销毁个人信息时，将立即删除

※　您有权拒绝上述个人信息收集和使用的同意。但如果拒绝同意，您可能会被限制参与韩国语学习者语料库的构建与利用项目。

☞ **您是否同意如上所述向第三方提供个人信息？　□同意　□不同意**

■ 个人信息处理业务委托详情(通知事项)

受托方（受托企业）	业务内容
IIR TECH Inc. / 延世大学产学合作团	2023年韩国语学习者语料库构建

年 _____月_____日_____

姓名 _____(签名或印)

致韩国国立国语院

以下信息将作为研究资料使用，个人信息将被保密，不会外泄。（请尽可能用韩国语回答，必要时可以使用英语。）

1. 性别: □ 女 □ 男

2. 出生年份： _____年(例如 ：1989年)

3. 现韩国语等级：_____(TOPIK:)

4. 国籍：_____（※ 侨胞与否 □ 侨胞 □ 外国人 ）

5. 第一语言：_____

6. 韩国语学习时间（学习韩国语多久了？）：_____年 _____个月（例如：一年零三个月）

7. 在韩国的居住时间（在韩国居住多久了？）：_____年 _____个月（例如：一年零三个月）

8. 学习韩国语的目的 ：

□ 升学 □ 就业 □ 居住 □ 兴趣 □ 结婚 □ 其它（ ）

9. 职业：_____

10. 韩国语以外其他可用外语（按照熟练程度从高到低罗列）：

_____

(중국어)

韓国語学習者言語資料構築・活用のための学習者言語資料利用承諾書
個人情報収集・利用及び第三者提供同意書

国立国語院では韓国語教育の質的向上のため学習者様の言語資料（コーパス）を収集・活用する事業を推進しております。みなさまよりご提供いただいた資料は韓国語教授方法の改善、韓国語教材開発、韓国語教育及び人工知能モデルの開発等、さまざまな分野の研究・開発に使用されます。この研究に参加することによる経済的な損害や身体的危険はありません。ご参加をご希望されない場合は撤回することも可能です。又、収集された個人情報は本事業の目的のみに使用し保護のため識別できない形態で使用されます。以下の内容をご確認のうえご署名ください。

■ 学習者資料の利用許可

韓国語学習者言語資料構築・活用のために作文資料を提供し、作文原文資料全体の公開と研究・開発目的の利用を承諾します。

☞ **上記の学習者の言語資料の利用許可に同意しますか？　　□同意　　□同意しない**

■ 個人情報収集・利用内訳

項目	収集・利用目的	保有・利用期間
性別、出生年、級、国籍、母国語、韓國語学習期間、韓国内居住期間、韓国語の学習目的、職業、使用可能な外國語	国立国語院 韓国語学習者の言葉の構築・活用	個人情報収集・利用同意日から30年 *ただし、個人情報の廃棄を要請した場合は直ちに破棄

※　上記の個人情報の収集・利用に同意しない権利があります。　但し、同意しない場合、韓国語学習者言語資料構築・活用事業への参加が制限されることがあります。

☞ **上記の個人情報を収集・利用することに同意しますか？　　□同意　　□同意しない**

■ 個人情報第三者提供内訳

提供される者	提供目的	提供項目	保有・利用期間
国立国語院韓国語学習者の言葉利用規約者	韓国語情報処理研究・開発、国語研究・開発	性別、出生年、級、国籍、母国語、韓国語学習期間、韓国内居住期間、韓国語の学習目的、職業、使用可能な外国語	個人情報収集・利用同意日から30年 *ただし、個人情報の廃棄を要請した場合は直ちに破棄

※　上記の個人情報の収集・利用に同意しない権利があります。　但し、同意しない場合、韓国語学習者言語資料構築・活用事業への参加が制限されることがあります。

☞ **上記の個人情報を第三者に提供することに同意しますか？　　□同意　　□同意しない**

■ 個人情報取扱業務委託内訳（告知事項）

委託される者（受託業者）	業務内容
IIR TECH Inc.・延世大学産学協力団	2023年の韓国語学習者言語資料の構築

日付 _____

氏名 _____**(署名)**

国立国語院御中

以下は研究のための資料として活用される情報です。個人情報は、秘密が保証され、外部に流出されることはございません。(出来れば韓国語でお答えください。英語でも構いません。)

1. 性別: □ 女 □ 男
2. 出生年: _____年(例. 1989年)
3. 級: _____(TOPIK:)
4. 国籍: _____(* 在日韓国人の場合は、☑をご記入してください。□)
5. 母国語: _____
6. 韓国語学習期間(どれくらい勉強しましたか?) _____年 _____ヵ月 (例. 1年 3ヵ月)
7. 韓国內居住期間(どれくらい住んでいましたか?) _____年 _____ヵ月 (例. 1年 3ヵ月)
8. 韓国語の学習の目的:

   □ 進学 □ 就職 □ 居住 □ 趣味 □ 結婚 □ その他 ( )
9. 職業: _____
10. 使用可能な外国語(使える言葉順にお書きください):

    _____

(일본어)

Giấy cho phép sử dụng tài liệu tiếng Hàn của người học để xây dựng và sử dụng kho ngữ liệu tiếng Hàn /
Đồng ý thu thập, sử dụng thông tin cá nhân và cung cấp cho bên thứ ba

Nhằm nâng cao chất lượng giáo dục tiếng Hàn, Viện quốc ngữ quốc gia Hàn Quốc đang tiến hành dự án thu thập và sử dụng tài liệu ngôn ngữ của học viên. Tài liệu học viên cung cấp được sử dụng để nghiên cứu và phát triển trong nhiều lĩnh vực khác nhau, chẳng hạn như cải tiến phương pháp giảng dạy tiếng Hàn, phát triển sách giáo khoa tiếng Hàn, giáo dục tiếng Hàn và phát triển mô hình trí tuệ nhân tạo. Những học viên tham gia vào việc nghiên cứu này sẽ hoàn toàn không có gì nguy hại đến tính mạng hay tổn thất về mặt kinh tế. Tuy nhiên trong trường hợp học viên không muốn tham gia thì có thể huỷ việc tham gia. Ngoài ra, những thông tin cá nhân mà học viên cung cấp, chúng tôi cam kết không dùng vào bất kỳ mục đích nào khác ngoài dự án này, và để không làm tiết lộ thông tin cá nhân của học viên tài liệu được thu thập dưới dạng không thể xác định được. Vui lòng đọc kỹ nội dung và quyết định có đồng ý hay không.

■ **Đồng ý sử dụng tài liệu của người học**

Tôi đồng ý cung cấp tài liệu viết để xây dựng và sử dụng kho tài liệu học tiếng Hàn, đồng ý việc công khai toàn bộ tài liệu viết gốc và việc sử dụng với mục đích nghiên cứu và phát triển.

☞ **Bạn có đồng ý cho phép sử dụng tài liệu ngôn ngữ của người học như trên không?**

☐Đồng ý ☐Không đồng ý

■ **Nội dung chi tiết thu thập và sử dụng thông tin cá nhân**

Hạng mục	Mục đích thu thập và sử dụng	Thời gian lưu giữ/sử dụng
Giới tính, ngày tháng năm sinh, cấp Topik hiện tại, Quốc tịch, Tiếng mẹ đẻ, thời gian học tiếng Hàn, thời gian cư trú tại Hàn, mục đích của việc học tiếng Hàn, nghề nghiệp, có thể sử dụng ngôn ngữ khác ngoài tiếng Hàn.	Xây dựng và sử dụng kho ngữ liệu của người học tiếng Hàn tại Viện Ngôn ngữ Quốc gia	30 năm kể từ ngày đồng ý thu thập và sử dụng thông tin cá nhân *Tuy nhiên, nếu thông tin cá nhân được yêu cầu xóa, nó sẽ bị xóa ngay lập tức.

※ Học viên có quyền từ chối sự đồng ý cho việc thu thập và sử dụng thông tin cá nhân ở trên. Tuy nhiên, nếu từ chối đồng ý, học viên có thể bị hạn chế tham gia vào dự án xây dựng/sử dụng kho ngữ liệu dành cho người học tiếng Hàn.

☞ **Bạn có đồng ý việc thu thập và sử dụng thông tin cá nhân như trên không?**

☐Đồng ý ☐Không đồng ý

■ **Nội dung chi tiết cung cấp thông tin cá nhân cho bên thứ ba**

Người tiếp nhận	Mục đích cung cấp	Hạng mục cung cấp	Thời gian lưu giữ/sử dụng
Người giao dịch sử dụng kho dữ liệu tiếng Hàn của người học của Viện quốc ngữ quốc gia Hàn Quốc	Nghiên cứu và phát triển xử lý thông tin tiếng Hàn, Nghiên cứu và phát triển ngôn ngữ Hàn Quốc	Giới tính, ngày tháng năm sinh, cấp Topik hiện tại, Quốc tịch, Tiếng mẹ đẻ, thời gian học tiếng Hàn, thời gian cư trú tại Hàn, mục đích của việc học tiếng Hàn, nghề nghiệp, có thể sử dụng ngôn ngữ khác ngoài tiếng Hàn.	30 năm kể từ ngày đồng ý thu thập và sử dụng thông tin cá nhân * Tuy nhiên, nếu thông tin cá nhân được yêu cầu xóa, nó sẽ bị xóa ngay lập tức.

※ Học viên có quyền từ chối sự đồng ý cho việc thu thập và sử dụng thông tin cá nhân ở trên. Tuy nhiên, nếu từ chối đồng ý, học viên có thể bị hạn chế tham gia vào dự án xây dựng/sử dụng kho ngữ liệu dành cho người học tiếng Hàn.

☞ **Bạn có đồng ý cung cấp thông tin cá nhân cho bên thứ ba như trên không?**

☐Đồng ý ☐Không đồng ý

■ **Nội dung chi tiết ủy thác công việc xử lý thông tin cá nhân (Thông báo)**

Người được ủy thác (công ty ủy thác)	Nội dung công việc
Quỹ Hợp tác giáo dục và doang nghiệp ㈜Irtech và trường đại học Yonsei	Xây dựng kho ngữ liệu của người học tiếng Hàn năm 2023

Ngày tháng năm

Họ và tên (ký tên)

Kính gửi Viện Ngôn ngữ Quốc gia Hàn Quốc

Sau đây là những thông tin được sử dụng làm tài liệu nghiên cứu. Thông tin cá nhân sẽ được đảm bảo bí mật và không bị tiết lộ ra bên ngoài. (Nếu có thể, vui lòng trả lời bằng tiếng Hàn hoặc khi cần có thể trả lời bằng tiếng Anh)

1. Giới tính : ☐ F ☐ M

2. Năm sinh : _____năm (Ví dụ: năm 1989)

3. Cấp học hiện tại : _____(TOPIK:)

4. Quốc tịch : _____

(※ Có phải là Hàn kiều hay không ☐Hàn kiều ☐ Người nước ngoài))

5. Tiếng mẹ đẻ (ví dụ: tiếng Việt, tiếng Trung) : _____

6. Thời gian học tiếng Hàn (Bạn đã học tiếng Hàn được bao lâu rồi?): _____năm _____tháng
(Ví dụ: 1 năm 3 tháng)

7. Thời gian cư trú tại Hàn Quốc (Bạn đã sống ở Hàn Quốc được bao lâu rồi?) : _____năm _____tháng
(Ví dụ: 1 năm 3 tháng)

8. Mục đích của việc học tiếng Hàn :

☐ Học lên cao ☐ Đi làm ☐ Sống ở Hàn ☐ Giải trí

☐ Kết hôn với người Hàn ☐ Nội dung khác _____

9. Nghề nghiệp : _____

10. Có thể sử dụng ngôn ngữ khác ngoài tiếng Hàn

(Vui lòng ghi theo trình tự từ ngoại ngữ giỏi đến kém hơn) :

_____

(베트남어)

หนังสืออนุญาตให้ใช้ข้อมูลทางภาษาของผู้เรียน เพื่อการจัดทำและการใช้งานคลังข้อมูลผู้เรียนภาษาเกาหลี โดยแสดงความยินยอมให้รวบรวม ใช้ข้อมูลส่วนบุคคล และมอบข้อมูลให้กับบุคคลที่ 3

สถาบันภาษาเกาหลีแห่งชาติส่งเสริมโครงการรวบรวมข้อมูลทางภาษาของผู้เรียนภาษาเกาหลี (คลังข้อมูล) เพื่อการยก ระดับคุณภาพของการศึกษาภาษาเกาหลี โดยข้อมูลที่ได้รับจากทุกท่านจะถูกนำไปใช้ในการวิจัยและพัฒนาในสาขาต่าง ๆ เช่น การพัฒนาวิธีการสอนภาษาเกาหลี การพัฒนาหนังสือเรียนภาษาเกาหลี การพัฒนารูปแบบปัญญาประดิษฐ์ หรือการเรียนการสอน ภาษาเกาหลี เป็นต้น ซึ่งผู้เข้าร่วมโครงการวิจัยนี้จะไม่ได้รับความเสียหายทางเศรษฐกิจหรืออันตรายทางด้านกายภาพใด ๆ ทั้งสิ้น และหากท่านไม่ต้องการเข้าร่วมโครงการวิจัยนี้ ท่านสามารถขอยกเลิกการเข้าร่วมได้ สำหรับข้อมูลส่วนบุคคลที่รวบรวมจะไม่ถูกนำไปใช้เพื่อวัตถุประสงค์อื่นใดนอกเหนือจากวัตถุประสงค์ของโครงการวิจัยนี้ และข้อมูลส่วนบุคคลจะถูกใช้ในรูปแบบที่ไม่สามารถระบุตัวตนได้เพื่อเป็นการรักษาความลับ ทั้งนี้ กรุณาอ่านเนื้อหาให้ละเอียดและเลือกว่าเห็นด้วยหรือไม่

■ **การอนุญาตให้ใช้ข้อมูลทางภาษาของผู้เรียน**

ข้าพเจ้ายินดีสนับสนุนข้อมูลงานเขียนภาษาเกาหลีเพื่อการจัดทำและการใช้งานคลังข้อมูลผู้เรียนภาษาเกาหลี อนุญาตให้ เปิดเผยข้อมูลของต้นฉบับงานเขียนภาษาเกาหลีทั้งหมด และอนุญาตให้ใช้ข้อมูลฯ ในวัตถุประสงค์เพื่อการวิจัยหรือพัฒนา

☞ **ท่านอนุญาตและยินยอมให้ใช้ข้อมูลทางภาษาของผู้เรียนดังรายละเอียดข้างต้นหรือไม่** ☐**ยินยอม** ☐**ไม่ยินยอม**

■ **รายละเอียดการรวบรวมและการนำข้อมูลส่วนบุคคลไปใช้**

รายการข้อมูล	วัตถุประสงค์การรวบรวมและการนำไปใช้	ระยะเวลาการเก็บข้อมูล และการนำไปใช้
เพศ, ปีเกิด, ระดับ TOPIK ณ ปัจจุบัน, สัญชาติ, ภาษาที่ 1, ระยะเวลาที่เรียนภาษาเกาหลี, ระยะเวลาที่อยู่ในประเทศเกาหลี, วัตถุประสงค์ของการเรียนภาษาเกาหลี, อาชีพ, ภาษาต่างประเทศอื่น ๆ ที่ท่านสามารถใช้ได้ นอกเหนือจากภาษาเกาหลี	เพื่อการจัดทำและการใช้งานคลังข้อมูลผู้เรียนภาษาเกาหลี ในสถาบันภาษาเกาหลีแห่งชาติ	30 ปี นับจากวันที่ได้รับการยินยอมให้รวบรวมและใช้ข้อมูลส่วนบุคคล * ทั้งนี้ ข้อมูลส่วนบุคคลจะถูกลบในทันทีที่มีการแจ้งขอให้ทำลายข้อมูลดังกล่าว

※ ท่านมีสิทธิ์ในการปฏิเสธให้ความยินยอมเกี่ยวกับการรวบรวมและการใช้ข้อมูลส่วนบุคคลด้านบน แต่กรณีที่ปฏิเสธการยินยอม ดังกล่าว ท่านอาจได้รับการจำกัดการเข้าร่วมโครงการจัดทำและการใช้งานคลังข้อมูลผู้เรียนภาษาเกาหลี

☞ **ท่านอนุญาตและยินยอมให้รวบรวมและใช้ข้อมูลส่วนบุคคลดังรายละเอียดข้างต้นหรือไม่** ☐**ยินยอม** ☐**ไม่ยินยอม**

■ **รายละเอียดการมอบข้อมูลส่วนบุคคลให้กับบุคคลที่ 3**

ผู้รับข้อมูล	วัตถุประสงค์ในการมอบข้อมูล	รายการข้อมูลที่มอบ	ระยะเวลาการเก็บข้อมูลและการนำไปใช้
ผู้ทำสัญญาการใช้คลังข้อมูลผู้เรียนภาษาเกาหลี โดยสถาบันภาษาเกาหลีแห่งชาติ	เพื่อการวิจัยและพัฒนาการจัดการข้อมูลทางด้านภาษาเกาหลี, การวิจัยและพัฒนาภาษาเกาหลี	เพศ, ปีเกิด, ระดับ TOPIK ณ ปัจจุบัน, สัญชาติ, ภาษาที่ 1, ระยะเวลาที่เรียนภาษาเกาหลี, ระยะเวลาที่อยู่ในประเทศเกาหลี, วัตถุประสงค์ของการเรียนภาษาเกาหลี, อาชีพ, ภาษาต่างประเทศอื่น ๆ ที่ท่านสามารถใช้ได้ นอกเหนือจากภาษาเกาหลี	30 ปี นับจากวันที่ได้รับการยินยอมให้รวบรวมและใช้ข้อมูลส่วนบุคคล * ทั้งนี้ ข้อมูลส่วนบุคคลจะถูกลบในทันทีที่มีการแจ้งขอให้ทำลายข้อมูลดังกล่าว

※ ท่านมีสิทธิ์ในการปฏิเสธให้ความยินยอมเกี่ยวกับการรวบรวมและการใช้ข้อมูลส่วนบุคคลด้านบน แต่กรณีที่ปฏิเสธการยินยอม ดังกล่าว ท่านอาจได้รับการจำกัดการเข้าร่วมโครงการจัดทำและการใช้งานคลังข้อมูลผู้เรียนภาษาเกาหลี

☞ **ท่านอนุญาตและยินยอมให้มอบข้อมูลส่วนบุคคลให้แก่บุคคลที่ 3 ดังรายละเอียดข้างต้นหรือไม่** ☐**ยินยอม** ☐**ไม่ยินยอม**

■ **รายละเอียดการมอบหมายให้จัดการข้อมูลส่วนบุคคล(ประกาศ)**

ผู้รับมอบ(บริษัทที่รับมอบการจัดการข้อมูล)	รายละเอียดหน้าที่ที่รับมอบ
IIR TECH Inc. และ University Industry Foundation, Yonsei University	จัดทำคลังข้อมูลผู้เรียนภาษาเกาหลี ปี 2023

วันที่ เดือน ปี

ชื่อ-นามสกุล (ลายเซ็น)

ขอมอบเอกสารฉบับนี้ให้กับสถาบันภาษาเกาหลีแห่งชาติ

เนื้อหาด้านล่างนี้เป็นข้อมูลที่จะถูกนำไปใช้เพื่อการทำวิจัย โดยข้อมูลส่วนบุคคลจะถูกรักษาเป็นความลับและไม่ถูกนำไป เผยแพร่สู่ภายนอก (หากเป็นไปได้กรุณาตอบด้วยภาษาเกาหลี หรือสามารถใช้ภาษาอังกฤษในการตอบได้เมื่อจำเป็น)

1. เพศ : ☐ F ☐ M

2. ปีเกิด : ปี ค.ศ. _____(ตัวอย่าง ปี 1989)

3. ระดับภาษาเกาหลี ณ ปัจจุบัน : _____(TOPIK:)

4. สัญชาติ : _____

(※ เป็นชาวเกาหลีโพ้นทะเลหรือไม่ ☐ ชาวเกาหลีโพ้นทะเล ☐ ชาวต่างชาติ)

5. ภาษาที่ 1 : _____

6. ระยะเวลาที่เรียนภาษาเกาหลี(ท่านเรียนภาษาเกาหลีเป็นระยะเวลานานเท่าใด) : _____ปี _____เดือน
(ตัวอย่าง 1 ปี 3 เดือน)

7. ระยะเวลาที่อยู่ในประเทศเกาหลี(ท่านพำนักอาศัยอยู่ในประเทศเกาหลีเป็นระยะเวลานานเท่าใด)
: _____ปี _____เดือน
(ตัวอย่าง 1 ปี 3 เดือน)

8. วัตถุประสงค์ของการเรียนภาษาเกาหลี

☐ ศึกษาต่อ ☐ ประกอบอาชีพ ☐ พำนักอาศัย ☐ งานอดิเรก ☐ แต่งงาน ☐ อื่น ๆ ( )

9. อาชีพ : _____

10. ภาษาต่างประเทศอื่น ๆ ที่ท่านสามารถใช้ได้ นอกเหนือจากภาษาเกาหลี

(กรุณาเรียงลำดับตามภาษาที่ใช้ได้ดี) : _____

(태국어)

Carta de permiso para utilizar materiales lingüísticos de los estudiantes para construir y utilizar corpus de estudiantes del idioma coreano / Consentimiento para la recopilación y el uso de información personal y su suministro a terceros

El Instituto Nacional de la Lengua Coreana promueve el proyecto, cual recoge datos lingüísticos (Corpus) de los estudiantes de coreano y utiliza el conjunto con el fin de mejorar la calidad de la educación del idioma coreano. Los materiales que tu proporcionas se utilizarán para la investigación y el desarrollo en diversos campos, como la mejora de los métodos de enseñanza del idioma coreano, el desarrollo de libros de texto en idioma coreano, la educación del idioma coreano y el desarrollo de modelos de inteligencia artificial. Los que participan en este proyecto no tendrá ningún daño económico ni daño físico. Cuando no deseas participar más, puedes optar por retirar tu participación. Tu información personal no será utilizada para ningún otro propósito y cualquier información que pueda identificarle permanecerá confidencial. Lee atentamente el contenido y decida si estás de acuerdo.

■ **Permiso para utilizar los materiales del alumno**

Se proporcionan materiales de escritura para crear y utilizar corpus de estudiantes del idioma coreano, y se permite la divulgación de todos los materiales de escritura originales y su uso con fines de investigación y desarrollo.

☞ **¿Estás de acuerdo con el permiso para utilizar los materiales lingüísticos de los estudiantes como se indica arriba?**

☐**De acuerdo** ☐**En desacuerdo**

■ **Detalles de recopilación y uso de información personal**

Elementos	Propósito de recopilación y uso	Período de conservación/uso
Género, Año de Nacimiento, Nivel Actual de Coreano, Nacionalidad, Lengua Materna, Tiempo Estudiando el Idioma Coreano, Tiempo de Residencia en Corea, Propósito de Estudiar el Idioma Coreano, Ocupación, Cualquier Otro Idioma Extranjero Que Puedes Hablar	Construcción y utilización de corpus de estudiantes del coreanos por el Instituto Nacional del Idioma Coreano	30 años desde la fecha del consentimiento para la recopilación y uso de información personal * Sin embargo, si se solicita la eliminación de información personal, se eliminará inmediatamente.

※ Tiene derecho a rechazar el consentimiento para la recopilación y el uso de la información personal anterior. Sin embargo, si se niega a dar tu consentimiento, es posible que se le restrinja la participación en el proyecto de construcción/utilización del corpus de estudiantes coreanos.

☞ **¿Estás de acuerdo con la recopilación y el uso de información personal como se indica arriba?**

☐**De acuerdo** ☐**En desacuerdo**

■ **Detalles del suministro de información personal a terceros**

Destinatario	Propósito de la provisión	Artículos proporcionados	Período de conservación/uso
Contratista para utilizar el corpus de estudiantes de idioma coreano del Instituto Nacional del Idioma Coreano	Investigación y desarrollo del procesamiento de información en idioma coreano, Investigación y desarrollo del idioma coreano	Género, Año de Nacimiento, Nivel Actual de Coreano, Nacionalidad, Lengua Materna, Tiempo Estudiando el Idioma Coreano, Tiempo de Residencia en Corea, Propósito de Estudiar el Idioma Coreano, Ocupación, Cualquier Otro Idioma Extranjero Que Puedes Hablar	30 años desde la fecha del consentimiento para la recopilación y uso de información personal * Sin embargo, si se solicita la eliminación de información personal, se eliminará inmediatamente.

※ Tiene derecho a rechazar el consentimiento para la recopilación y el uso de la información personal anterior. Sin embargo, si se niega a dar tu consentimiento, es posible que se le restrinja la participación en el proyecto de construcción/utilización del corpus de estudiantes coreanos.

☞ **¿Aceptas proporcionar información personal a terceros como se indica arriba?**

☐**De acuerdo** ☐**En desacuerdo**

■ **Información personal Manejo de detalles del envío (Aviso)**

Destinatario (empresa consignadora)	Detalles del negocio
IIR TECH Inc. / Fundación Universidad-Industria de Yonsei	Construcción del corpus de estudiantes de coreano en 2023

Fecha: _____

Nombre: _____ **(Firma)**

Al Instituto Nacional del Idioma Coreana

El siguiente cuestionario incluye tu información personal que nos ayudará a hacer análisis de investigación científica. Estricta confidencialidad se mantendrá en todo momento. (Si es posible, llena el formulario en coreano o inglés.)

1. Género: ☐ M ☐ H
2. Año de Nacimiento: _____(p.ej. 1989)
3. Nivel actual de coreano: _____(TOPIK:)
4. Nacionalidad: _____

(※ Yo soy un coreano que ha vivido en el extranjero ☐ Extranjero ☐)

5. Lengua Materna: _____
6. Tiempo Estudiando el Idioma Coreano(¿Durante cuánto tiempo has estado aprendiendo?) :

_____años _____meses (p.ej. 1 año 8 meses)

7. Tiempo de Residencia en Corea(¿Por cuánto tiempo has estado en Corea?) :

_____años _____meses (p.ej. 1 año 10 meses)

8. Propósito de estudiar el idioma coreano:

☐ para ingresar a una universidad en Corea ☐ para trabajar en Corea

☐ para vivir en Corea ☐ para un pasatiempo

☐ para el matrimonio

☐ otro ( )

9. Ocupación: _____
10. Cualquier otro idioma extranjero que puedes hablar

(Nota: Si hay varios idiomas, por favor escríbelos en orden de fluidez) :

_____

(스페인어)

Письмо-разрешение на использование языковых материалов учащихся для создания и использования лингвистического корпуса изучающих корейский язык
Согласие на сбор и использование персональной информации и её предоставление третьим лицам

Государственный институт корейского языка в целях улучшения преподавания корейского языка продвигает проект, в ра мках которого осуществляется сбор языковых данных (лингвистический корпус) изучающих корейский язык. Предоставляемые вами материалы используются для проведения исследований и разработок в различных областях, таких как совершенствован ие методов преподавания корейского языка, разработка учебников корейского языка, обучение корейскому языку и создание моделей искусственного интеллекта. Данный проект не несёт никакого финансового ущерба и физического вреда для прини мающих в нём участие. Вы можете отказаться от участия в проекте при отсутствии соответствующего желания. Ваша персона льная информация не будет использоваться для каких-либо других целей, и любая информация, которая может идентифицир овать вас, останется конфиденциальной. Пожалуйста, внимательно прочтите указанное ниже содержание и примите решение об участии.

■ **Разрешение на использование материалов учащегося**

Предоставляю письменные материалы для создания и использования корпуса изучающих корейский язык. Даю разрешен ие на раскрытие всех оригинальных письменных материалов и их использование в целях исследований и разработок.

☞ **Вы согласны на использование языковых материалов учащихся, как указано выше?**

☐**Да** ☐**Нет**

■ **Описание собираемой и используемой персональной информации**

Вид	Цель сбора и использования	Период хранения / использования
Пол, дата рождения, текущий уровень TOPIK, гра жданство, первый (родной) язык, период изучения ко рейского языка, период проживания в Республике Ко рея, цель изучения корейского языка, профессия, вла дение другими иностранными языками кроме корейс кого.	Создание и использов ание корпуса учащихся ко рейского языка в Государс твенном институте корейск ого языка	**30 лет со дня согласия на сбор и использование персональной инф ормации** * Личная информация может быть немед ленно удалена по запросу.

※ Вы имеете право отказать в согласии на сбор и использование персональной информации, указанной выше. В этом слу чае вы можете быть ограничены в участии в проекте по созданию/использованию корейского корпуса учащихся.

☞ **Вы согласны на сбор и использование персональной информации, как указано выше?**

☐**Да** ☐**Нет**

■ **Описание личной информации, предоставляемой третьим лицам**

Получатель	Цель предоставления	Предоставляемая информация	Срок хранения и использования
Сторона, предо ставившая согласие на использование к орпуса учащихся к орейского языка Го сударственного инс титута корейского я зыка	**Исследования и ра зработки в области об работки информации н а корейском языке, исследования и разр аботки корейского язы ка**	Пол, дата рождения, текущий уровень TOPIK, гражданство, первый (родной) язык, период изучения корейского языка, период проживания в Республике Корея, цель изуч ения корейского языка, профессия, владени е другими иностранными языками кроме к орейского.	**30 лет со дня согл асия на сбор и использ ование персональной и нформации** * Личная информация мо жет быть немедленно удален а по запросу.

※ Вы имеете право отказать в согласии на сбор и использование персональной информации, указанной выше. В этом слу чае вы можете быть ограничены в участии в проекте по созданию/использованию корейского корпуса учащихся.

☞ **Вы согласны на предоставление персональной информации третьим лицам, как указано выше?**

☐**Да** ☐**Нет**

■ **Подробности передачи обязанностей по обработке персональной информации (уведомление)**

Принимающая субподряд сторона (субподрядчик)	Описание задач
IIR TECH Inc., Промышленный фонд Университета Ёнсе	Создание корпуса учащихся корейского языка в 2023 году

Дата _____

ФИО _____ (подпись)

Для предоставления в Государственный институт корейского языка

Ниже требуется заполнить личные данные, необходимые для исследования. Их сохранность и защита от огласки гарантируется (Желательно заполнять на корейском языке. При необходимости можно заполнить н а английском).

1. Пол: ☐ Ж ☐ М

2. Год рождения: _____г. (напр. 1989 г.)

3. Текущий уровень: _____(TOPIK:)

4. Гражданство: _____(※ Зарубежный соотечественник: ☐ Да ☐ Нет)

5. Первый (родной) язык: _____

6. Срок изучения корейского языка: _____г. _____мес. (Напр.: 1 г. 3 мес.)

7. Срок проживания в Республике Корея: _____г. _____мес. (Напр.: 1 г. 3 мес.)

8. Цель изучения корейского языка:

☐ Поступление в ВУЗ ☐ Трудоустройство ☐ Проживание в Республике Корея

☐ Увлечение ☐ Брак ☐ Иное ( )

9. Профессия: _____

10. Владение другими иностранными языками кроме корейского

(перечислить в порядке убывания уровня мастерства владения):

_____

(러시아어)

Солонгос хэл суралцагчийн корпус санг бҮтээх, ашиглахад солонгос хэл суралцагчдын материалыг ашиглахыг зөвшөөрөх хуудас / Хувийн мэдээлэл цуглуулах, ашиглах, гуравдагч этгээдэд дамжуулахыг зөвшөөрөх хуудас

Үндэсний хэл шинжлэлийн төвөөс солонгос хэлний бодит сургалтыг сайжруулахын тулд солонгос хэл суралцагчдын хэлний материалыг цуглуулан ашиглах төсөл хэрэгжҮҮлж байна. Таны бидэнд хураалгасан материалыг солонгос хэл заах арга зҮйг сайжруулах, солонгос хэлний сурах бичиг боловсруулах, солонгос хэлний боловсрол, хиймэл оюун ухааны загвар боловсруулах гэх мэт төрөл бҮрийн чиглэлээр судалгаа, боловсруулалт хийхэд ашиглана. Тус судалгаанд оролцон материалаа хураалгаж буй танд санхҮҮгийн болон бие эрхтний хохирол ямар нэгэн байдлаар учрахгҮй. Хэрэв уг төсөлд хамрагдахыг хҮсэхгҮй тохиолдолд оролцохгҮй байх бҮрэн боломжтой. Мөн таны хураалгасан материалыг дээр дурдсанаас өөр зорилгоор ашиглагдахгҮй, судалгаанд ашиглах Үед хувь хҮний мэдээлийг нууцлах ямар нэгэн байдлаар танигдуулахгҮйгээр ашиглана. Агуулгыг анхааралтай уншиж, судалгаанд оролцохыг зөвшөөрөх эсэхээ шийднэ ҮҮ.

■ Суралцагчийн материалыг ашиглах зөвшөөрөл

Бичих материалыг солонгос хэл суралцагчийн корпус санд ҮҮсгэх, ашиглахад зориулалтаар хураалгасан бөгөөд бҮх эх бичвэрийг ил болгох, судалгааны болон хөгжҮҮлэх зорилгоор ашиглахыг зөвшөөрч байна.

☞ **Та суралцагчдын материалыг дараах байдлаар ашиглахыг зөвшөөрч байна уу?**

☐ **Зөвшөөрч байна** ☐ **ЗөвшөөрөхгҮй**

■ Хувийн мэдээллийг цуглуулах, ашиглахтай холбоотой мэдээллийн дэлгэрэнгҮй агуулга

Хамаарах мэдээлэл	Цуглуулах, ашиглах зорилго	Хадгалах/ашиглах хугацаа
ХҮйс, төрсөн он, солонгос хэлний тҮвшин одоогийн байдлаар, иргэншил, Эх хэл, Солонгос хэл сурал цсан хугацааг жил сараар бичнэ ҮҮ, Солонгост ний т амьдарсан хугацаа, Солонгос хэл сурч буй зорил го, Мэргэжил, Солонгос хэлнээс гадна эзэмшсэн хэ лээ дэс дарааллын дагуу бичнэ ҮҮ.	Солонгос хэлний Үндэсний хҮрээлэн Солонгос хэл суралцагчийн корпус санд ҮҮсгэх, ашиглах	**Хувийн мэдээлэл цуглуулах, ашиглахыг зөвшөөрсөн өдрөөс хойш 30 жил** * Хэрэв хувийн мэдээллийг устгах хҮсэлт гаргавал тэр даруй устгана.

※ Танд дээрх хувийн мэдээллийг цуглуулах, ашиглахыг зөвшөөрөхөөс татгалзах эрх бий. Хэрэв та мэдээллээ ашиглуулах аас татгалзвал та солонгос хэл суралцагчдын корпус санг ҮҮсгэх, ашиглах эрхээ хязгаарлуулж болзошгҮйг анхаарна уу .

☞ **Та дээрх хувийн мэдээллийг цуглуулах, ашиглахыг зөвшөөрч байна уу?**

☐ **Зөвшөөрч байна** ☐ **ЗөвшөөрөхгҮй**

■ Гуравдагч этгээдэд дамжуулах хувийн дэлгэрэнгҮй мэдээллийн дэлгэрэнгҮй агуулга

ХҮлээн авагч	Хангамжийн зорилго	Хамаарах мэдээлэл	Хадгалах/ашиглах хугацаа
Солонгос хэлний Үндэсний хҮрээлэнгийн солонгос хэл суралцагчийн корпусыг ашиглах гэрээт гҮйцэтгэгч	**Солонгос хэлний мэдээлэл боловсруулах судалгаа, хөгжҮҮлэлт, Солонгос хэлний судалгаа, хөгжҮҮлэлт**	ХҮйс, төрсөн он, солонгос хэлний тҮвшин одоогийн байдлаар, иргэншил, Эх хэл, Солонгос хэл суралцсан хугацааг жил сараар бичнэ ҮҮ, Солонгост нийт амьдарсан хугацаа, Солонгос хэл сурч буй зорилго, Мэргэжил, Солонгос хэлнээс гадна эзэмшсэн хэлээ дэс дараалын дагуу бичнэ ҮҮ.	**Хувийн мэдээллийг цуглуулах, ашиглахыг зөвшөөрсөн өдрөөс хойш 30 жил** * Хэрэв хувийн мэдээллийг устгах хҮсэлт гаргавал тэр даруй устгана.

※ Танд дээрх хувийн мэдээллийг цуглуулах, ашиглахыг зөвшөөрөхөөс татгалзах эрх бий. Хэрэв та мэдээллээ ашиглуулахаас татгалзвал таны солонгос хэл суралцагчдын корпус санг ҮҮсгэх, ашиглах эрхийг хязгаарлаж болзошгҮйг анхаарна уу .

☞ **Та дээр дурдсаны дагуу гуравдагч этгээдэд хувийн мэдээллээ өгөхийг зөвшөөрөх ҮҮ?**

☐ **Зөвшөөрч байна** ☐ **ЗөвшөөрөхгҮй**

■ Хувийн мэдээлийг хариуцах ажлын дэлгэрэнгҮй мэдээлэл (Мэдэгдэл)

ГҮйцэтгэгч (гҮйцэтгэч компани)	Бизнесийн дэлгэрэнгҮй мэдээлэл
IIR TECH Inc. / Йонсей их сургууль-аж Үйлдвэр академийн хамтын ажилгаагааны зөвлөл	2023 он солонгос хэлний сургалтын корпус санд ҮҮсгэх

Он сар өдөр _____

Нэр _____ **/гарын Үсэг/**

Солонгос хэлний Үндэсний хҮрээлэнд

Дараах нь судалгаатай холбоотой мэдээллийн чанартай асуултууд юм. Таны хувийн мэдээлэл нууцлагдах б өгөөд гадагш ямар нэгэн байдлаар ашиглагдахгүй. (Аль болох солонгосоор бөглөнө үү. Солонгосоор бөглөх боломжгүй бол англиар бөглөж болно.)

1. Хүйс : _____

2. Төрсөн он : _____

3. Таны солонгос хэлний түвшин одоогийн байдлаар : _____(TOPIK:)

4. Иргэншил : _____

5. Эх хэл : _____

6. Солонгос хэл суралцсан хугацааг жил, сараар нь бичнэ үү :

_____ жил _____ сар (Жишээ нь: 1жил 3сар)

7. Солонгост нийт амьдарсан хугацаа :

_____ жил _____ сар (Жишээ нь: 1жил 3сар)

8. Солонгос хэл сурч буй зорилго :

☐ Их дээд сургуульд элсэн сурах ☐ Ажилд орох ☐ Оршин суух

☐ Сонирхол ☐ Гэрлэх ☐ Бусад

9. Мэргэжил : _____

10. Солонгос хэлнээс гадна эзэмшсэн хэлээ (сайн эзэмшсэн дарааллын дагуу) бичнэ үү :

_____

(몽골어)

Surat Pernyataan Persetujuan Penggunaan Materi Bahasa Pembelajar untuk Pembuatan dan Pemanfaatan Korpus Pembelajaran Bahasa Korea/ Persetujuan untuk Pengumpulan dan Penggunaan Informasi Pribadi dan Penyediaan kepada Pihak Ketiga.

Untuk meningkatkan kualitas pendidikan bahasa Korea, Institut Nasional Bahasa Korea mengadakan proyek untuk mengumpulkan dan memanfaatkan materi bahasa (korpus) pembelajar bahasa Korea. Materi yang Anda berikan akan digunakan untuk penelitian dan pengembangan di berbagai bidang seperti penyempurnaan metode pengajaran bahasa Korea, pengembangan buku teks bahasa Korea, pendidikan bahasa Korea, dan pengembangan model kecerdasan buatan (Artificial Intelligence). Tidak ada kerugian finansial dan bahaya fisik yang berisiko terjadi pada partisipan dalam penelitian ini. Jika Anda tidak ingin berpartisipasi, Anda dapat mengundurkan diri dari penelitian ini. Selain itu, informasi pribadi yang dikumpulkan tidak akan digunakan untuk tujuan apa pun selain untuk proyek ini dan akan digunakan dalam bentuk yang tidak dapat diidentifikasi untuk menjaga kerahasiaan. Silakan membaca isi dari surat pernyataan ini dengan teliti dan memutuskan apakah Anda setuju atau tidak setuju dengan pernyataan di bawah ini.

■ **Persetujuan Penggunaan Materi Pembelajar**

Menyediakan materi tulis untuk pembuatan dan pemanfaatan korpus pembelajar bahasa Korea, dan setuju bahwa seluruh materi tulisan asli akan ditampilkan dan digunakan untuk tujuan penelitian dan pengembangan.

☞ **Apakah Anda setuju dengan penggunaan materi bahasa pembelajar sesuai pernyataan di atas?**

☐Setuju ☐Tidak Setuju

■ **Perihal pengumpulan dan penggunaan informasi pribadi**

Jenis informasi pribadi yang dikumpulkan dan digunakan	Tujuan pengumpulan dan penggunaan	Durasi penyimpanan/penggunaan
Jenis kelamin, tahun kelahiran, level TOPIK yang masih berlaku, kewarganegaraan, bahasa ibu, lama belajar bahasa Korea, lama tinggal di Korea, tujuan belajar bahasa Korea, pekerjaan, bahasa asing lain yang dikuasai selain bahasa Korea.	Pembuatan dan Pemanfaatan Korpus Pembelajar Bahasa Korea oleh Institut Nasional Bahasa Korea.	**30 tahun sejak tanggal persetujuan pengumpulan dan penggunaan informasi pribadi** * Informasi pribadi akan segera dihapus jika ada permintaan penghapusan.

※ Anda berhak untuk menolak memberikan persetujuan untuk pengumpulan dan penggunaan informasi pribadi di atas. Jika Anda menolak untuk memberikan persetujuan, Anda tidak diperbolehkan berpartisipasi dalam proyek pembuatan/pemanfaatan korpus pelajar bahasa Korea.

☞ **Apakah Anda setuju dengan pengumpulan dan penggunaan informasi pribadi sesuai pernyataan di atas?**

☐Setuju ☐Tidak Setuju

■ **Perihal penyediaan informasi pribadi kepada pihak ketiga**

Penerima	Tujuan penyediaan	Jenis informasi pribadi yang disediakan	Durasi penyimpanan/penggunaan
Kontraktor untuk penggunaan korpus pembelajar bahasa Korea dari Institut Nasional Bahasa Korea	**Penelitian dan Pengembangan Pengolahan Informasi Korea, Penelitian dan Pengembangan Bahasa Korea**	Jenis kelamin, tahun kelahiran, level TOPIK yang masih berlaku, kewarganegaraan, bahasa ibu, lama belajar bahasa Korea, lama tinggal di Korea, tujuan belajar bahasa Korea, pekerjaan, bahasa asing lain yang dikuasai selain bahasa Korea.	**30 tahun sejak tanggal persetujuan pengumpulan dan penggunaan informasi pribadi** * Informasi pribadi akan segera dihapus jika ada permintaan penghapusan.

※ Anda berhak menolak memberikan persetujuan untuk pengumpulan dan penggunaan informasi pribadi di atas. Namun, jika Anda menolak untuk memberikan persetujuan, Anda tidak diperbolehkan untuk berpartisipasi dalam proyek pembuatan/pemanfaatan korpus pelajar bahasa Korea.

☞ **Apakah Anda setuju untuk memberikan informasi pribadi kepada pihak ketiga sesuai pernyataan di atas?**

☐Setuju ☐Tidak Setuju

■ **Perihal pendelegasian penanganan informasi pribadi (pemberitahuan)**

Penerima delegasi (Perusahaan konsinyasi)	Detail
IIR TECH Inc. / Yayasan Industri-Universitas Yonsei	Pembuatan korpus pembelajar bahasa Korea tahun 2023

Tanggal:

Nama: **(Tanda tangan)**

kepada Institut Nasional Bahasa Korea

Informasi pribadi sebagaimana tercantum di bawah akan digunakan sebagai data untuk penelitian. Informasi pribadi yang Anda sediakan bersifat rahasia dan akan dijaga agar tidak tersebar ke luar. (Jika memungkinkan, harap menjawab dengan bahasa Korea atau bila perlu, Anda juga bisa menggunakan bahasa Inggris)

1. Jenis kelamin : ☐ Perempuan ☐ Laki-laki

2. Tahun kelahiran : Tahun _____ (Contoh: Tahun 1989)

3. Level sekarang : _____(TOPIK: _____)

4. Kewarganegaraan : _____

(※ Perihal negara asal ☐ orang Korea yang tinggal di luar negeri ☐ orang asing)

5. Bahasa ibu : _____

6. Lama belajar bahasa Korea (Berapa lama belajar bahasa Korea?): selama ____ tahun ____ bulan
(contoh: 1 tahun 3 bulan)

7. Lama tinggal di Korea (Berapa lama tinggal di Korea?): selama ____ tahun ____ bulan
(contoh: 1 tahun 3 bulan)

8. Tujuan belajar bahasa Korea:

☐ pendidikan (masuk perguruan tinggi) ☐ pekerjaan ☐ menetap di Korea

☐ hobi ☐ menikah ☐ lain (_____)

9. Pekerjaan: _____

10. Bahasa asing lain yang dikuasai, selain bahasa Korea. (Urutkan bahasa sesuai tingkat kelancaran):

_____

(인도네시아어)

부록2. 2023년 한국어 학습자 말뭉치 수집 지침

한국어 학습자 말뭉치 자료 수집 지침

※ 국립국어원에서 주관하는 '2023년 한국어 학습자 말뭉치 구축' 사업 수행을 위해 학습자의 쓰기 자료 수집을 요청드립니다. 협조에 응해 주셔서 진심으로 감사드립니다.

1. 자료 수집 대상 및 수집 자료

▶ 대상: 한국어 교육 기관의 학습자
▶ 자료: 학습자가 산출한 작문 자료 (문어)
  - 학습자 1인당 최대 5편 제출 가능
  - 각 작문 자료는 100어절 이상 작성
▶ 수집 유형: 특정 주제에 대한 기획 자료 수집
▶ 수집 시기: 2023년 8월 - 2024년 4월

2. 작문 자료 주제

- 각 수준별로 다음의 주제에 따라 글을 쓰도록 하여 수집합니다.
- 학습자 1인당 최대 5편의 주제에 대해 글을 쓸 수 있습니다.
- 다음의 주제를 제시하되 풍부한 글쓰기를 위해 관련 내용을 자유롭게 확장할 수 있도록 합니다. (세부 질문을 넣은 예시 과제를 별첨해 드립니다. 필요 시 사용해 주십시오.)
- 하나의 완결된 글로, 글의 길이가 100어절이 넘어야 합니다.
- 가족이나 친구, 동료, 한국어 모어 화자의 피드백이 이루어지지 않은 글이 되도록 합니다.

※ 1번과 2번은 모든 수준의 공통 주제입니다.(★표) 많은 수집을 부탁드립니다.

※ 2번은 애플리케이션을 통해 그림 또는 영상을 보고 타이핑하여 제출하는 온라인 과제입니다. (초급은 종이에 내용을 작성한 뒤 사진 촬영하여 제출 가능합니다.)

<수준별 수집 주제>

수준	주제 번호	주제
초급	1	★자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, …)
	2	★날씨 변화와 해결 방법 (더워지는 날씨, 비가 많이 내리는 날씨, …)
	3	내가 가장 좋아하는 것과 싫어하는 것 (일, 행동, 사람, 물건, …)
	4	내가 가 본 여행지 (위치, 교통편, 날씨, 음식, 간 곳, …)
	5	올해의 계획 (하고 싶은 일, 이유)
중급	1	★자신의 나라와 한국 비교 (비슷한 점, 다른 점)
	2	★기후 문제와 해결 방법
	3	과학 기술의 발전이 인간의 생활에 미치는 영향
	4	10년 후의 나의 계획
	5	효율적인 시간 관리 방법
고급	1	★자신의 나라와 한국 비교 (유사점, 차이점)
	2	★기후 위기와 해결 방안
	3	내가 생각하는 성공
	4	인구 문제의 원인과 해결 방법
	5	역사적 사건이나 인물 소개

3. 자료 수집 절차 (오프라인)

학습자 작문 자료 및 동의서 수집

⇓

100어절 이상 자료 선별

⇓

스캔하여 PDF 변환
/ 사진 촬영하여 이미지 파일 형식으로 저장

⇓

파일명 부여

① 학습자가 손으로 쓴 작문지와 동의서를 수집합니다.

② 수집된 자료가 100어절 이상인지 확인합니다.

③ 자료를 스캔하거나 사진 촬영하여 PDF 파일로 변환합니다. (PDF 변환이 어려울 경우, 이미지 파일 형식(jpg, png, bmp)으로 변환) 이때 화면이 흐릿하여 글자가 잘 보이지 않는 경우가 있으니 유의합니다.

④ 해당 자료의 출처를 파악할 수 있도록 동의서와 작문 자료에 다음과 같은 파일명을 부여합니다. (주제번호는 2p 참조)

자료	파일명	예시
동의서	Ag_급수_국적_이름	Ag_3급_대한민국_김서준
작문 자료	급수_국적_이름_T주제번호	3급_대한민국_김서준_T1

※ 이름은 한글/영문 무관함

4. 자료 수집 절차 (온라인)

학습자에게 온라인 제출 매뉴얼 전달

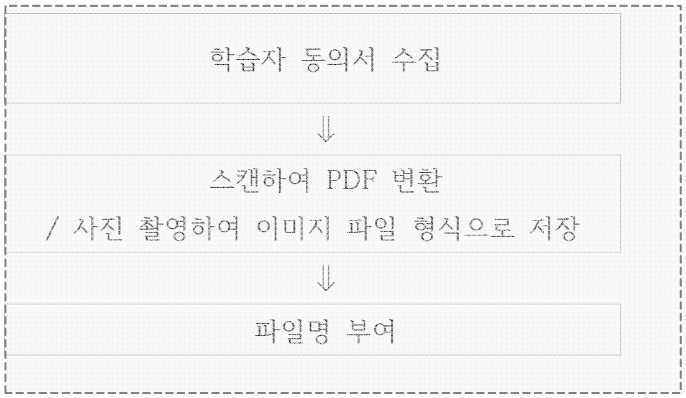
⇓

학습자의 ID 전달받아 별첨4.에 기입

⇓

별첨4.에 학습자 과제 제출 여부 표시

⇓



※학습자가 온라인 과제만 제출하는 경우 동의서 수집 필요

① 학습자에게 온라인 제출 매뉴얼을 전달 또는 설명합니다.
② 학습자의 이름과 ID를 전달받아 <별첨 파일4>에 기입합니다.
(ID는 KOKOA 앱에 회원 가입 시 생성됩니다.)
③ <별첨 파일4>에 학습자의 과제 제출 여부를 표시합니다.
④ 학습자가 온라인 과제만 제출할 경우(오프라인 과제 제출이 없어서 동의서를 미작성한 경우), 해당 학습자의 동의서를 수집합니다.
⑤ 동의서를 스캔하거나 사진 촬영하여 PDF 파일로 변환합니다. (PDF 변환이 어려울 경우, 이미지 파일 형식(jpg, png, bmp)으로 변환) 이때 화면이 흐릿하여 글자가 잘 보이지 않는 경우가 있으니 유의합니다.

5. 학습자 동의서 수집

▶ 모든 자료는 자료 제공과 사용에 관한 학습자의 동의서를 받은 후 수집합니다.
▶ 동의서를 수합한 후 누락된 항목이 있는지 확인합니다. 국적 정보와 같이 수집 교사가 확인 가능한 항목이 누락된 경우 적어 넣습니다.

[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 합니다. 그 밖의 학습자가 추가적으로 궁금해 하는 사항이 있을 경우에는 설명해 줍니다.
2. 학습자가 수기로 적고 사인하도록 할 수 있도록 출력하여 배포합니다.
3. 학습자 한 명당 동의서는 한 번만 받습니다. 처음 수집할 때 받은 동의서와 짝을 맞출 수 있도록 쓰기 자료의 파일명에 들어가는 학습자 이름과 동의서 파일명에 들어가는 학습자 이름을 반드시 동일하게 적어 주십시오.
4. 사업 종료 이후, 동의서 원본은 아래 주소로 착불 발송해 주십시오.
착불 발송 시, 이메일로 등기 번호 또는 송장 번호를 통지해 주시기 바랍니다.
(2023corpus@gmail.com)

(주)이르테크
주소: (34188) 대전 유성구 계룡로 114, 907호(봉명동, 대전유성BYC빌딩)
전화번호: 042-824-1108

6. 학습자에게 전달할 자료

① 학습자 개인 정보 수집·이용 및 저작권 이용 동의서
- 자료 수집 동의서입니다. 총 9개 언어로 번역되어 나갑니다.
(영어, 중국어, 일본어, 베트남어, 태국어, 러시아어, 스페인어, 인도네시아어, 몽골어)

② 학습자 매뉴얼 (한국어 학습자 말뭉치 수집 참여 가이드)
- 쓰기 참여 가이드입니다. 주제 및 쓰기 방법 등이 제시되어 있습니다.
- 직접 제출, 온라인 제출 매뉴얼이 각각 제시되어 있습니다.
- 초급, 중급, 고급 각 급수별로 다르게 제작됩니다.
- 초급에 한해 위의 9개 언어로 번역되어 나갑니다.

⇒ *학습자 매뉴얼은 직접 전달하셔도 되고, 설명으로 대체하셔도 됩니다.*

아래의 내용을 별첨해 드립니다.
수집에 참고 부탁드립니다.
별첨4.는 학생 쓰기 자료와 함께 제출해 주십시오.

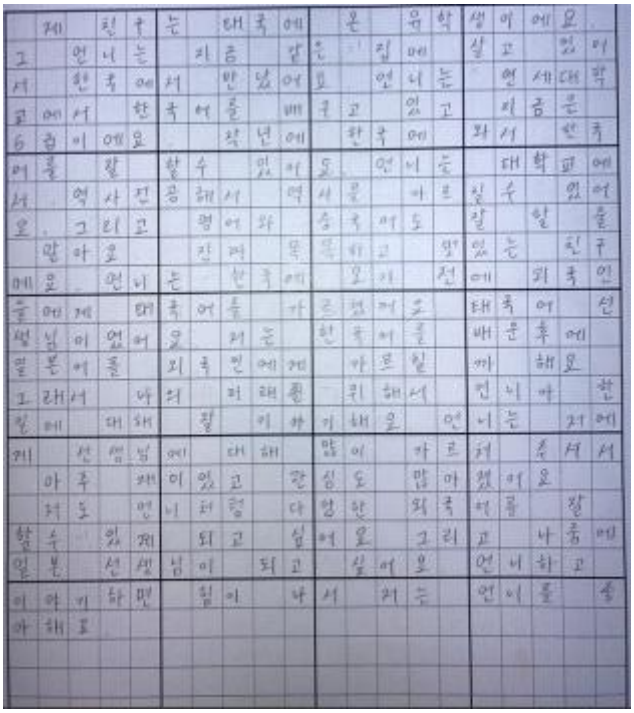
별첨1. 수준별 쓰기 과제 예시 (필요하신 경우에 사용하십시오.)
별첨2. 100어절 쓰기 예시
별첨3. 온라인 과제 제출 매뉴얼
별첨4. 온라인 과제 ID 작성 양식
별첨5. 100어절 쓰기 양식 (필요하신 경우에 사용하십시오.)

별첨1. 수준별 쓰기 과제 예시

수준	번호	주제
초급	1	**★자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, ...)**
	2	**★날씨 변화와 해결 방법 (더워지는 날씨, 비가 많이 내리는 날씨, ...)**
	3	**내가 가장 좋아하는 것과 싫어하는 것 (일, 행동, 말, 사람, 물건, ...)**
	4	**내가 가 본 여행지 (위치, 교통편, 날씨, 음식, 간 곳, ...)**
	5	**올해의 계획 (하고 싶은 일, 이유)**
중급	1	**★자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, ...)** ○ 여러분 나라와 한국은 어떤 점이 비슷합니까? ○ 여러분 나라와 한국은 어떤 점이 다릅니까?
	2	**★기후 문제와 해결 방법 (폭염, 폭우, 지구 온난화, ...)** ○ 전 세계적으로 어떤 기후 변화가 문제가 되고 있습니까? ○ 기후가 변하면서 우리 생활은 어떻게 바뀌었습니까? ○ 기후 문제를 해결할 수 있는 방법은 무엇입니까?
	3	**과학 기술의 발전이 인간의 생활에 미치는 영향 (스마트폰, 인터넷, 로봇, 인공지능, ...)** ○ 과학 기술의 발전이 미치는 긍정적인 영향은 무엇입니까? ○ 과학 기술의 발전이 미치는 부정적인 영향은 무엇입니까? ○ 과학 기술의 발전에 따른 문제 해결 방안은 무엇입니까?
	4	**10년 후의 나의 계획** ○ 여러분은 10년 후에 어떻게 살고 싶습니까? ○ 그 이유는 무엇입니까? ○ 무엇을 준비해야 합니까?
	5	**효율적인 시간 관리 방법** ○ 언제 시간 관리가 필요합니까? ○ 효율적으로 시간을 쓸 수 있는 방법은 무엇입니까?
고급	1	**★자신의 나라와 한국 비교 (날씨, 생활, 사람, 문화, ...)** ○ 여러분 나라와 한국은 어떤 유사점이 있습니까? ○ 여러분 나라와 한국은 어떤 차이점이 있습니까?
	2	**★기후 위기와 해결 방안 (폭염, 폭우, 지구 온난화, 해수면 상승, ...)** ○ 기후 위기가 우리 생활에 미치는 영향은 무엇입니까? ○ 기후 위기의 원인은 무엇입니까? ○ 기후 위기를 해결할 수 있는 방안은 무엇입니까?
	3	**내가 생각하는 성공** ○ 여러분은 성공이 무엇이라고 생각합니까? ○ 성공하기 위해 필요한 것이 무엇이라고 생각합니까? ○ 그 이유는 무엇입니까?
	4	**인구 문제** ○ 현재 무엇이 가장 심각한 인구 문제라고 생각합니까?(저출산, 고령화, 1인 가구, ...) ○ 그 문제의 원인은 무엇입니까? ○ 그 문제의 해결 방법은 무엇입니까?
	5	**역사적 사건이나 인물 소개** ○ 여러분 나라 또는 다른 나라의 역사적인 사건이나 인물을 소개해 주십시오. ○ 그 사건이 갖는 역사적 의미는 무엇입니까? / 그 인물을 통해 얻을 수 있는 교훈이 있습니까?

별첨2. 100어절 쓰기 예시

<원고지> 105어절



<A4 용지> 111어절

부모가 육아서를 읽는 이유는 크게 두 가지가 있습니다. 첫째는 책을 통해 위안을 얻고자 함이고, 둘째는 책을 통해 계몽을 얻고자 함입니다. 위안은 곧, '내가 그래도 아이를 잘 키우고 있구나', '우리 아이만 이런 것은 아니구나'하는 것입니다. 이런 내용은 매우 일반적인 이야기, 가령, '부정어를 사용하지 말라', '결과만 보고 아이를 혼내지 말라'하는 것 등의 책의 내용과 관련됩니다. 계몽은 이와는 반대로 내가 알고 싶었던 이야기, 내가 알고 있던 것과는 다른 이야기와 관련됩니다. '아이가 자만해도 내버려 둬라', '아이가 적극적이지 않아도 걱정하지 말라' 등의 내용입니다. 본 책은 이 두 가지 내용 중 어느 한 가지에 치중하지 않고 적절히 내용을 잘 조합하고 있어 흥미롭게 읽을 수 있었습니다.

결론적으로 이 책은 아이를 키우는 부모에게 도움이 될 수 있는 책이라고 생각합니다.

별첨3. 온라인 과제 제출 매뉴얼

※ 초급에 한해 종이에 글을 쓰고 사진 촬영하여 자료를 업로드할 수 있습니다.

<table>
<tr><td>1</td><td>휴대폰이나 태블릿 PC에서
'KOKOA' 앱 다운로드</td><td>2</td><td>회원 가입</td></tr>
<tr><td colspan="2">iOS - Appstore
 
Android - Playstore
  </td><td colspan="2">  </td></tr>
<tr><td>3</td><td>교재 선택 → 단원 선택</td><td>4</td><td>미디어 확인 → 과제 작성</td></tr>
<tr><td colspan="2">교재명: Writing Assignments
단원: 초급, 중급, 고급 중 선택
  </td><td colspan="2">초급: 그림 / 중급, 고급: 영상
  </td></tr>
</table>

별첨4. 온라인 과제 ID 작성 양식

※ 학생 이름은 **동의서 파일명**에 적은 이름과 동일한 이름으로 적어 주십시오.

no.	학생 이름	KOKOA ID	과제 제출 여부	비고
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

※ KOKOA ID는 **KOKOA 앱**을 다운받은 후, 회원 가입하면 생성됩니다.

※ 문제 입력

- 상세 질문 1
- 상세 질문 2
- 상세 질문 3

여기
까지
↓

부록3. 2023년 한국어 학습자 말뭉치 구축 지침

한국어 학습자 말뭉치 문어 입력 지침

※ 자료 처리 절차

○ 자료 처리는 파일을 전산화하여 말뭉치 자료로서 본격적인 구축과 가공 작업을 하기 위한 전처리 단계로 다음과 같은 절차에 따라 처리한다.

문어 자료

⇓

구축 대상 자료 선별

⇓

작문 자료와 학습자 동의서 스캔

⇓

파일명 부여

⇓

파일 등록

⇓

자료 보관

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

문어
○ 학습자 동의서에 서명한 자료 ○ 학습자 동의서의 개인 정보 모두 입력된 항목 선정
○ 완결된 텍스트 작문 자료 선정 ○ 텍스트의 길이 100어절 이상의 자료 선정 ○ 복사 또는 스캔 파일의 경우 화질이 좋은 자료 선정

1. 전체적인 형식 원칙

○ LCMS의 입력 창에서 입력한다.
○ 자료를 입력하기 전 표본 정보와 학습자의 개인 정보를 입력한다.
○ 학습자가 글 하나를 스스로 완성하였을 경우에만 입력하는 것을 원칙으로 한다.
○ 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 <u>춘대말</u>을 한다는 것이 자주 반말을 말한다. → 수정 안 함.

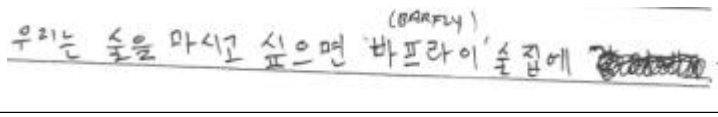
○ 원문의 영어와 한자는 모두 유지한다. 한자는 시스템 입력창에서 글자를 선택 후 마우스 오른쪽을 클릭하여 입력한다.
○ 띄어쓰기는 어문 규범과 <표준국어대사전>의 표제어에 맞춰 수정하여 입력한다. 원활한 형태소 분석 작업을 위해 띄어쓰기를 정확히 적용한다.
○ 분수 표시는 다음과 같이 입력한다.

<예> 1/2, 3/4

○ 영문자, 한글 자모, 괄호문자 등은 자판을 사용하여 입력한다.

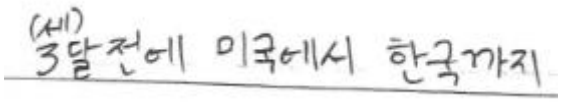
<예> ㄱㄴㄷㄹ, (1) (2) (3)

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

<예> 아래의 경우 '바프라이(BARFLY)'로 입력한다. 

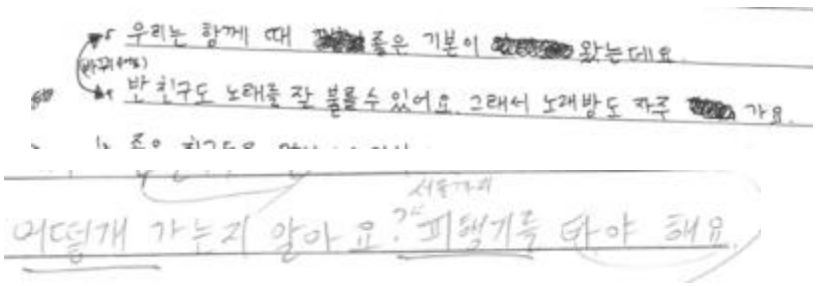
○ 숫자와 한글 표기를 함께 쓴 경우 원문에 따라 병기한다.

<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 마크업

한다.

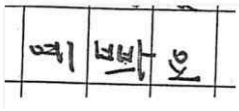
<예>	<CNI>¿¿</CNI>

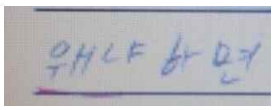
○ ‘기타기호’는 문장 앞에 붙인 블릿 기호나 다른 특수 기호들을 원본 그대로 입력 후 마크업 할 때 사용한다. 키보드에서 한글 자음을 입력 후 ‘한자’키를 눌러서 선택하여 입력한다. (‘기타기호’는 원본 그대로 입력하므로 ¿기호를 입력하지 않도록 주의한다.)

<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**ㅃ**ㅏ요 : 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.

<기획·연구>

국립국어원 홍혜진 학예연구관

국립국어원 김초록 연구원

<연구 참여자>

연구 책임자 곽용진 (㈜이르테크)

공동 연구원 한송화(연세대학교)

이지연(㈜이르테크)

강현화(연세대학교)

김선정(계명대학교)

김한샘(연세대학교)

홍혜란(연세대학교)

연구 보조원 이근영(㈜이르테크)

최용기(㈜이르테크)

이순미(㈜이르테크)

정해영(㈜이르테크)

홍은기(㈜이르테크)

오누리(㈜이르테크)

김윤희(㈜이르테크)

박정혁(㈜이르테크)

박하영(㈜이르테크)

박우정(㈜이르테크)

임수용(㈜이르테크)

박상원(㈜이르테크)

오수연(㈜이르테크)

조현수(㈜이르테크)

윤하정(㈜이르테크)

안수빈(㈜이르테크)
최영롱(연세대학교)
신유진(연세대학교)
하태현(연세대학교)
오지연(연세대학교)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775
전송 02-2669-9727
인쇄일: 2024년 5월 2일
발행일: 2024년 5월 2일
인 쇄: 세종기획

※ 이 책은 국립국어원의 용역비로 수행한 '2023년 한국어 학습자 말뭉치 구축' 사업의 결과물을 발간한 것입니다.