

국립국어원 2023-01-62

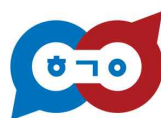
발간등록번호
11-1371028-000987-01

# 2023년 한국어-외국어 병렬 말뭉치 구축

연구 책임자  
이 정 희



국립국어원



한국어-외국어  
병렬 말뭉치 구축 사업단



## 제 출 문

국립국어원장 귀하

국립국어원의 국고 보조금으로 수행한 ‘2023년 한국어-외국어 병렬  
말뭉치 구축’ 사업의 결과 보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 5월 ~ 2023년 12월

2023년 12월 29일

연구 책임자: 이정희(경희대학교)

연구 기관: 2023년 한국어-외국어 병렬 말뭉치 구축 사업단  
(☎국제한국어교육학회, ㈜플리토)

연구 책임자: 이정희

연구 참여자: 김일환, 김종민, 박진욱, 이동규, 이동은, 이수미,  
이영준, 임채훈, 조남호, 최문석, 최홍열, 김연희,  
김영근, 문진숙, 박지민, 윤세윤, 이두용, 전지연,  
정성호, 지화숙, 한재민, 국혜민, 김한별, 박광길,  
서유리, 이상후, 이요셉, 이혜민, 최예린, 이정수,  
강동한, 김진구, 최승미, 김재훈, 이제영, 김이주





<연구 수행자> 2023년 한국어-외국어 병렬 말뭉치 구축 사업단

연구 책임자	이정희(경희대학교)	
연구 참여자	김일환(성신여자대학교)	김종민(경희대학교)
	박진욱(대구가톨릭대학교)	이동규(고려대학교)
	이동은(국민대학교)	이수미(성균관대학교)
	이영준(한국학중앙연구원)	임채훈(송실대학교)
	조남호(명지대학교)	최문석(경희대학교)
	최홍열(강원대학교)	
	김연희(고려대학교)	김영근(경희대학교)
	문진숙(경희대학교)	박지민(경희대학교)
	윤세운(경희대학교)	이두용(고려대학교)
	전지연(강원대학교)	정성호(고려대학교)
	지화숙(경희대학교)	한재민(경희대학교)
	국혜민(송실대학교)	김한별(대구가톨릭대학교)
	박광길(강원대학교)	서유리(경희대학교)
	이상후(고려대학교)	이요셉(경희대학교)
	이혜민(경희대학교)	최예린(경희대학교)
	이정수(㈜플리토)	강동한(㈜플리토)
	김진구(㈜플리토)	최승미(㈜플리토)
	김재훈(㈜플리토)	이제영(㈜플리토)
	김이주(㈜플리토)	



## <국문 초록>

### 2023년 한국어-외국어 병렬 말뭉치 구축

이 사업은 ‘2021년 한국어-외국어 병렬 말뭉치 구축 사업’과 ‘2022년 한국어-외국어 병렬 말뭉치 구축 사업’에 이어 대한민국의 새로운 교류 협력 관계로 주목받고 있는 국가의 언어들을 대상으로 한국어-외국어 병렬 말뭉치를 구축하고, 구축된 병렬 말뭉치의 활용 방안을 수립하는 것이 목표이다. 병렬 말뭉치 구축 대상은 아세안-인도 지역의 6개 언어(베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어)와 유라시아 지역의 2개 언어(러시아어, 우즈베크어)이다.

병렬 말뭉치 구축은 크게 ‘수집’, ‘구축’, ‘산출’의 세 단계로 진행되었다. ‘수집’ 단계에서는 한국어 문어·구어 원시 데이터를 수집하고, 수집한 데이터를 검수·정제하였다. 검수·정제가 완료된 데이터를 대상으로 ‘구축’ 단계에서 번역이 이루어졌으며, 번역문은 교정 작업자의 교정과 (주)플리토 내부 언어 전문가(linguist)의 1차 검수(5% 표본 검수)를 거쳤다. 이후 (사)국제한국어교육학회 번역 검수원이 번역문을 2차 전수 검수하였으며, 2차 검수가 완료된 문장은 검수팀장이 3차 검수(20% 표본 검수)를, 감수자가 감수(10% 표본 검수)를 진행하였다. 이상의 단계를 거쳐 품질이 검증된 번역문은 ‘산출’ 단계에서 메타 정보와 함께 최종 데이터로 산출되었다.

최종적으로 한국어-외국어 병렬 말뭉치 총 11,045,120어절(한국어 원문 기준)과 추가 제안 사항이었던 한국어-영어 병렬 말뭉치 1,380,640어절(한국어 원문 기준)을 구축하였다.

다음으로 병렬 말뭉치의 활용 방안으로서 병렬 말뭉치 용례 검색기의 프로토타입을 제안하였다. 병렬 말뭉치를 활용하여 웹 기반 용례 검색기 프로토타입을 개발함으로써 말뭉치 이용의 진입 장벽을 낮추고 언어·외국어 교육 및 연구 분야에서의 활용도를 높이하고자 하였다.

또한 병렬 말뭉치와 인공지능 기술에 관련된 지식을 공유하고 사업의 성과를 대외적으로 확산시키기 위하여 국제 심포지엄을 개최하였다. 그리고 해외 출장(필리핀)을 통해 한국어-외국어 병렬 말뭉치 구축에 대한 관심이 높다는 것을 확인하고 기관 간 협력 네트워크의 기반을 마련하였다. 아울러 사업 수행 과정에서 연구

한 결과물을 학술 대회에서 발표하고 학술지에 게재함으로써 관련 연구의 활성화를 도모하였다.

이 사업의 기대 효과는 다음과 같다.

첫째, 대규모·고품질의 병렬 말뭉치 구축을 통해 언어 데이터 산업의 기초를 마련할 수 있을 것이다.

둘째, 다국어 언어 처리 및 인공지능(AI) 기반 통·번역 모델의 품질을 지속적으로 향상시킬 수 있을 것이다.

셋째, 국가 정책 및 관련 업계의 수요를 충족시킬 수 있을 것이다.

이 사업을 통해 구축된 한국어-외국어 병렬 말뭉치는 그 자체로서 가치가 있을 뿐만 아니라 다양한 분야에 활용되는 기초 자료로 그 역할을 할 수 있을 것으로 기대된다.

**주요어:** 한국어-외국어 병렬 말뭉치, 병렬 말뭉치 용례 검색기, 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어

# 차 례

## 제1장 사업 개요

1. 사업의 목표 .....	1
2. 사업의 범위 .....	1
3. 사업 수행 절차 및 체계 .....	2
3.1. 사업 수행 절차 .....	2
3.2. 사업 수행 체계 .....	4
4. 사업 수행 내용 .....	8

## 제2장 사업 수행

1. 원문 수집 및 정제 .....	13
1.1. 원문 수집 .....	13
1.2. 원문 정제 .....	19
1.3. 개인 정보 비식별화 .....	24
2. 번역 지침 수정·보완 .....	28
2.1. 지침 및 원문 검토 .....	29
2.2. 원문 데이터 및 초기 번역 데이터 검토 .....	29
2.3. 지침 수정 및 보완 .....	29
2.4. 지침 최종본 완성 및 적용 .....	30
3. 번역 .....	31
3.1. 번역 작업자 선정 .....	31
3.2. 번역 인력 교육 .....	32
3.3. 번역 절차 .....	32
3.4. 번역 데이터 구축 환경 .....	33
3.5. 번역 품질 향상 활동 .....	35
4. 검수 및 감수 .....	37
4.1. 번역 검수팀 .....	37
4.2. 감수자 .....	39
4.3. 번역 검수팀 및 감수자 교육 .....	39

# 차 례

4.4. 검수 품질 향상 활동 .....	47
<b>5. 용례 검색기 .....</b>	<b>53</b>
5.1. 프로토타입 개발 .....	53
5.2. 시범 사용 및 자문 의견 .....	54
5.3. 프로토타입 보완 .....	60
<b>6. 병렬 말뭉치 구축 및 메타 정보 .....</b>	<b>62</b>
6.1. JSON 포맷 .....	62
6.2. 최종 데이터 구조 .....	63
6.3. JSON 예시 .....	63
<b>7. 보안 .....</b>	<b>65</b>
7.1. 보안 관리 대상 및 담당자 .....	65
7.2. 보안 관리 방법 .....	67

## 제3장 사업 수행 결과

<b>1. 말뭉치 데이터 구축 결과 .....</b>	<b>79</b>
1.1. 최종 구축 데이터 .....	79
1.2. 번역 품질 비교 .....	79
1.3. 데이터 품질 관리 결과 .....	95
<b>2. 대외 활동 .....</b>	<b>96</b>
2.1. 사업단 국제 심포지엄 개최 .....	96
2.2. 학술 논문 게재 .....	107
2.3. 국외 출장 .....	109
2.4. 대외 홍보 .....	114
<b>3. 활용 방안 및 기대 효과 .....</b>	<b>116</b>
3.1. 병렬 말뭉치의 활용 방안 .....	116
3.2. 사업의 기대 효과 .....	119
3.3. 제언 .....	120

## 표 차례

<표 1> 사업 범위 .....	1
<표 2> 사업 수행 절차별 세부 내용 .....	3
<표 3> 사업단 직책별 역할 .....	5
<표 4> 월별 사업 추진 경과 .....	8
<표 5> 원문 수집 목표 .....	13
<표 6> 원문 수집 수량 .....	13
<표 7> 일상 대화 말뭉치 수집 진행 방식 .....	15
<표 8> 구어체 원문 정보 .....	15
<표 9> 문어체 원문 정보 .....	16
<표 10> 원문 예시 .....	18
<표 11> 기계적 정제 기준 .....	20
<표 12> 항목별 원문 정제 예시 .....	22
<표 13> 띄어쓰기 정제 예시 .....	23
<표 14> 인용문 정제 예시 .....	24
<표 15> 번역 지침 집필 과정 .....	28
<표 16> 번역 및 교정 절차 .....	33
<표 17> 번역 및 교정 작업자 주요 업무 .....	33
<표 18> 1차 검수 작업자 주요 업무 .....	34
<표 19> 번역 품질 향상을 위한 기계적 정제 내용 .....	36
<표 20> 번역 품질 향상을 위한 구조적 검증 기준 .....	36
<표 21> 영어 데이터 품질 향상 활동 내용 .....	37
<표 22> 언어별 번역 검수원 .....	38
<표 23> 언어별 감수자 .....	39
<표 24> 검수·감수 인력 보안 교육 내용 .....	40
<표 25> 지침 공통 교육 내용 .....	41
<표 26> 언어별 지침 내용 .....	41
<표 27> 검수 유형별 검수 방법 .....	43
<표 28> 오류 유형 기준 .....	43
<표 29> 번역문 오류 유형 .....	51
<표 30> 언어별 번역 오류 양상 및 원인 .....	52
<표 31> 용례 검색기 대상 자료 및 환경 .....	54
<표 32> 용례 검색기 스키마 예시 .....	55

## 표 차례

<표 33> 용례 검색기 자문 및 보완 내용 .....	56
<표 34> 용례 검색기 프로토타입 기능 사항 .....	57
<표 35> 유사 단어 검색 예시 .....	61
<표 36> 메타데이터 JSON 형식 예시 .....	63
<표 37> 병렬 말뭉치 JSON 형식 예시 .....	64
<표 38> 보안 관리 사항 .....	68
<표 39> 보안 점검 항목 및 분류 .....	68
<표 40> 최종 산출물 저장 공간 .....	69
<표 41> 시스템 접근 통제 관련 법률 및 규정 .....	71
<표 42> 사업 참여자 보안 교육 .....	72
<표 43> ‘사이버 보안 점검의 날’ 시행 개요 .....	73
<표 44> PC 보안 점검 항목 .....	73
<표 45> 최종 구축 데이터 수량 .....	79
<표 46> 기계 번역과 병렬 말뭉치의 번역 비교(베트남어) .....	80
<표 47> 기계 번역과 병렬 말뭉치의 번역 비교(인도네시아어) .....	82
<표 48> 기계 번역과 병렬 말뭉치의 번역 비교(태국어) .....	84
<표 49> 기계 번역과 병렬 말뭉치의 번역 비교(인도 힌디어) .....	86
<표 50> 기계 번역과 병렬 말뭉치의 번역 비교(캄보디아 크메르어) .....	87
<표 51> 기계 번역과 병렬 말뭉치의 번역 비교(필리핀 타갈로그어) .....	89
<표 52> 기계 번역과 병렬 말뭉치의 번역 비교(러시아어) .....	91
<표 53> 기계 번역과 병렬 말뭉치의 번역 비교(우즈베크어) .....	93
<표 54> 데이터 품질 감리 결과 .....	95
<표 55> 국제 심포지엄 오전 식순 .....	96
<표 56> 국제 심포지엄 주제 발표 내용 .....	97
<표 57> 국제 심포지엄 오후 식순 .....	99
<표 58> 국제 심포지엄 패널 토의 및 개인 발표 내용 .....	100
<표 59> 국제 심포지엄 기사 게재 언론 .....	115
<표 60> A사 번역기와 병렬 말뭉치 번역기의 BLEU 점수 비교(한→우) .....	116
<표 61> A사 번역기와 병렬 말뭉치 번역기의 BLEU 점수 비교(우→한) .....	117
<표 62> 병렬 말뭉치의 교육 분야 활용 방안 .....	119



# 그림 차례

[그림 1] 사업 수행 절차 요약도 .....	2
[그림 2] 사업 수행 체계 및 인력 구성 .....	5
[그림 3] 유튜브 채널 ‘전성기 TV’ .....	14
[그림 4] 검수 플랫폼의 원문 신고 버튼 .....	21
[그림 5] 원문 정제 지침 교육 예시 .....	22
[그림 6] 신조어 목록 예시 .....	24
[그림 7] 전산 프로그램을 이용한 기계적 개인 정보 처리 과정 예시 .....	25
[그림 8] 번역 검수 세부 지침 수정 사항 예시(우즈베크어) .....	30
[그림 9] 번역 검수 세부 지침 보완 사항 예시(인도네시아어) .....	30
[그림 10] 번역 검수 지침 최종본 목록 .....	31
[그림 11] 작업자 선정을 위한 품질 평가 점수 예시 .....	32
[그림 12] 번역 및 작업 지침 교육 내용 예시 .....	32
[그림 13] 플리토 아케이드 시스템의 번역·교정 환경 .....	34
[그림 14] 플리토 아케이드 시스템의 표본 검수 환경 .....	35
[그림 15] 검수·감수 교육 절차 .....	40
[그림 16] 검수·감수 보안 교육 예시 .....	40
[그림 17] 검수 지침 교육 예시 .....	42
[그림 18] 검수 플랫폼 작업 환경 예시 .....	42
[그림 19] 검수 플랫폼 사용 교육 예시 .....	44
[그림 20] 감수 작업 환경 예시 .....	44
[그림 21] 감수 방법 교육 예시 .....	45
[그림 22] 해외 거주자 성희롱·성폭력 예방 교육 예시 .....	46
[그림 23] 한국양성평등교육진흥원 교육 이수증 예시 .....	46
[그림 24] 성희롱·성폭력 예방 교육 이수 현황 관리 대장 .....	47
[그림 25] 3차 검수 예시 .....	48
[그림 26] 번역 검수원 대상 검수 재교육 예시 .....	48
[그림 27] 로마자 표기 및 문장 기호 오류 확인 예시 .....	49
[그림 28] 원문 이해 지원 예시 .....	50
[그림 29] 번역 품질 개선을 위한 의견 작성 및 전달 예시 .....	50
[그림 30] 용례 검색기 암호키 설정 .....	55
[그림 31] Search Documents by Query 검색 예시 .....	58
[그림 32] 검색 결과 표 예시 .....	58

## 그림 차례

[그림 33] ‘*가’ 검색 시 결과값 .....	59
[그림 34] ‘?가?’ 검색 시 결과값 .....	59
[그림 35] 검색 결과 문장 수, 다운로드 예시 .....	59
[그림 36] 엑셀 다운로드 예시 .....	60
[그림 37] 온라인 세종 한일 병렬 말뭉치 형태소 검색 예시 .....	60
[그림 38] 스토리지 이중화 .....	61
[그림 39] metadata JSON 형식 .....	62
[그림 40] document_paragraph JSON 형식 .....	62
[그림 41] 병렬 말뭉치 JSON 형식 .....	63
[그림 42] 구축 데이터 샘플 Excel 예시 .....	63
[그림 43] 보안 담당자 역할 체계 .....	66
[그림 44] 사무실 출입문 전경(좌: (사)국제한국어교육학회, 우: (주)플리토) ·	67
[그림 45] 보안 관리 대장 예시 .....	68
[그림 46] 저장 관리 예시(NAS) .....	69
[그림 47] 저장 관리 예시(Cloud) .....	70
[그림 48] 자가 보안 점검 결과 예시 .....	74
[그림 49] 보안 교육 예시 .....	75
[그림 50] 국제 심포지엄 행사 사진 .....	104
[그림 51] 학술지 게재 논문 .....	108
[그림 52] 필리핀 언론 보도 기사 일부 발췌 .....	113
[그림 53] 필리핀 출장 일정별 사진 .....	113
[그림 54] 국제 심포지엄 언론 보도 예시(뉴시스) .....	115



# 제 1 장

## 사업 개요



## 1. 사업의 목표

본 사업은 대한민국의 새로운 교류 협력 관계로 주목받고 있는 국가 언어들을 대상으로 한국어-외국어 병렬 말뭉치를 구축하는 것이 목표이다. 병렬 말뭉치 구축의 대상이 되는 8개 언어는 아세안-인도 지역의 6개 언어(베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어)와 유라시아 지역의 2개 언어(러시아어, 우즈베크어)이다.

본 사업에서는 다국어 언어 처리 및 인공지능 기반 번역 기술의 향상뿐만 아니라 한국어와 신한류 콘텐츠의 전 세계적인 확산에 기여할 수 있는 고품질의 한국어-외국어 병렬 말뭉치를 구축함으로써 4차 산업 혁명 시대에 자동 번역 등 언어문화 산업의 성장 동력을 마련하고자 한다.

## 2. 사업의 범위

본 사업의 범위는 한국어와 아세안-인도 6개국, 유라시아어 2개국 언어 총 8개 언어의 병렬 말뭉치를 구축하는 것이다. 먼저 한국어-외국어 병렬 말뭉치 구축을 위한 검증 체계를 수립하고 이를 기준으로 언어별로 138만 어절, 총 1,104만 어절을 구축한다. 8개 언어에는 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어가 해당한다. 그리고 말뭉치 구축 과정에서 원시 데이터의 저작권을 해결하고 구축한 데이터를 활용하는 방안을 제시하는 것도 이 사업의 범위이다. 본 사업에서는 활용 방안의 일환으로 병렬 말뭉치 용례 검색기의 프로토타입을 제안하였다. 이 외에 한국어-외국어 병렬 말뭉치를 소개하고 활용 사례를 공유하는 워크숍을 개최하는 과업도 포함한다. 본 사업에서는 범위를 확장하여 국제 심포지엄을 개최하였다.

<표 1> 사업 범위

- 한국어-외국어 병렬 말뭉치 구축 및 검증 체계 수립
- 한국어-외국어 병렬 말뭉치 총 1,104만 어절 구축
  - 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어 각 138만 어절
- 한국어-외국어 병렬 말뭉치 저작권 해결 및 활용 방안 수립
- 한국어-외국어 병렬 말뭉치 소개 및 활용 사례 공유 워크숍 개최
  - 한국어-외국어 병렬 말뭉치 워크숍 홍보 계획 수립 및 홍보

### 3. 사업 수행 절차 및 체계

#### 3.1. 사업 수행 절차

본 사업은 ‘2021년 한국어-외국어 병렬 말뭉치 구축(이하 2021년 사업)’과 ‘2022년 한국어-외국어 병렬 말뭉치 구축(이하 2022년 사업)’의 연속 사업으로서 지난 두 사업의 수행 절차를 기본 바탕으로 하되, 일부 단계들을 보완하여 사업 수행의 효율성을 제고하였다. 사업 수행 절차는 전체적으로 데이터 수집과 정제, 번역, 검수, 감수, 결과물 제출로 이루어진다.

구체적으로 살펴보면, 먼저 원시 데이터 수집 단계에서는 국립국어원으로부터 문어체·구어체 원시 데이터를 제공받고 유튜브 대본을 추가 구매하였다. 원시 데이터 수집을 완료한 후에는 기계적 정제와 전문가 정제로 이루어진 총 2단계의 정제 절차를 수행하여 높은 수준의 원문 정제를 실시하였다.

다음으로 정제된 원문은 번역 과정을 거치고 번역문을 대상으로 (주)플리토에서 1차 검수 작업을 수행하였다. 그리고 1차 검수 단계를 통과한 데이터는 도착어 기준 원어민 검수원의 2차 검수(전수)와 검수팀장 주도의 3차 검수(20%), 해외 원어민 교수의 감수(10%) 작업을 진행하였다.

이러한 과정을 통해서 최종적으로 총 1,104만 어절 이상의 병렬 데이터를 산출하고 국립국어원에 제출하였다. 사업 수행 절차 요약도와 절차별 세부 내용은 다음과 같다.



[그림 1] 사업 수행 절차 요약도

<표 2> 사업 수행 절차별 세부 내용

사업 수행 절차		업무 주체	세부 내용
수집	원시 데이터 수집	(주)플리토	<ul style="list-style-type: none"> <li>- 국립국어원에서 제공한 문어체·구어체 데이터 활용</li> <li>- 구어체 원시 데이터의 추가 확보를 위해 유튜브 대본 구매</li> <li>- 원시 데이터 저작권 확보</li> <li>· 유튜브 대본 구매 시 국가 언어 자원 구축(말뭉치) 및 활용 저작권 이용 허락 계약서를 작성하여 저작권 확보</li> </ul>
	원문 정제	(사)국제 한국어 교육학회 및 (주)플리토	<ul style="list-style-type: none"> <li>- 구축한 원시 데이터 검수 및 정제</li> <li>· 1차 정제: 기준에 따른 데이터 기계 처리</li> <li>· 2차 정제: 원문 정제팀 운영으로 원문에 대한 지속적이며 집중적인 정제 실시, 국어학·한국어학 전공의 전문가 활용</li> <li>- 데이터 관리 번호 및 분류·주제·소주제 등의 메타 정보 및 헤더 부착</li> </ul>
구축	번역	(주)플리토	<ul style="list-style-type: none"> <li>- 번역에 관한 저작권 확보</li> <li>· 번역 플랫폼 가입 절차 활용 및 별도의 저작권 이용 허락 계약 체결을 통한 저작권 확보</li> </ul>
		(사)국제 한국어 교육학회	<ul style="list-style-type: none"> <li>- 번역 지침 제공</li> </ul>
		(주)플리토	<ul style="list-style-type: none"> <li>- 자체 플랫폼에서의 번역</li> <li>· 플랫폼 환경에서 클라우드소싱, 현지 번역 업체 및 전문 번역사를 활용한 번역 수행</li> <li>- 자체 플랫폼에서의 번역문 전수 교정</li> <li>· 플랫폼 환경에서 번역물 전수 상호 교정</li> </ul>
	1차 검수	(주)플리토	<ul style="list-style-type: none"> <li>- 1차 표본 검수</li> <li>· (주)플리토 내부 언어 전문가가 플랫폼 환경에서 번역문의 5%를 무작위로 표본 검수하여 1차 검수 수행</li> <li>· 1차 검수에서 품질 기준 미달 문장은 재번역 실시</li> </ul>
	2차 검수	(사)국제한국어교육	<ul style="list-style-type: none"> <li>- 2차 전수 검수</li> <li>· 1회 이상 검수 교육 실시</li> </ul>

		학회	<ul style="list-style-type: none"> <li>· 학회의 전문 인력을 활용하여 전체 번역문에 대한 2차 검수 수행</li> <li>- 번역 품질 관련 회의 실시</li> <li>· 번역 품질 향상을 위한 검수 애로 사항 및 품질 관련 회의 실시</li> <li>- 자문 위원단 자문 요청 및 회의 실시</li> <li>· 외국어 자문 위원단을 구성하여 자문 회의 참석 및 자문 의견 전달</li> </ul>
	3차 검수		<ul style="list-style-type: none"> <li>- 3차 표본 검수</li> <li>· 언어별 검수팀장이 검수문의 20%를 표본 검수 수행</li> <li>· 검수 시 자주 발생하는 오류 정리</li> <li>· 검수원 평가 및 오류율이 높은 검수원 재교육 실시</li> </ul>
	감수		<ul style="list-style-type: none"> <li>- 표본 감수</li> <li>· 한국어 관련 전공 교수 및 원어민 언어 전문가의 10% 표본 감수 수행</li> <li>· 검수의 품질 향상을 위하여 수시로 감수 의견 전달</li> </ul>
산출	결과물 제출	(사)국제 한국어 교육학회 및 (주)플리토	<ul style="list-style-type: none"> <li>- 데이터 최종 점검</li> <li>- 최종 데이터 산출 및 제출</li> <li>· (주)플리토: 구축 데이터 최종 산출</li> <li>· (사)국제한국어교육학회: 최종 보고서 작성 및 보고</li> </ul>

### 3.2. 사업 수행 체계

본 사업을 수행하는 조직 체계는 크게 총괄팀, 자문 위원단, 수집팀, 원문 정제팀, 구축팀, 번역 검수팀, 감수팀으로 나눌 수 있다.

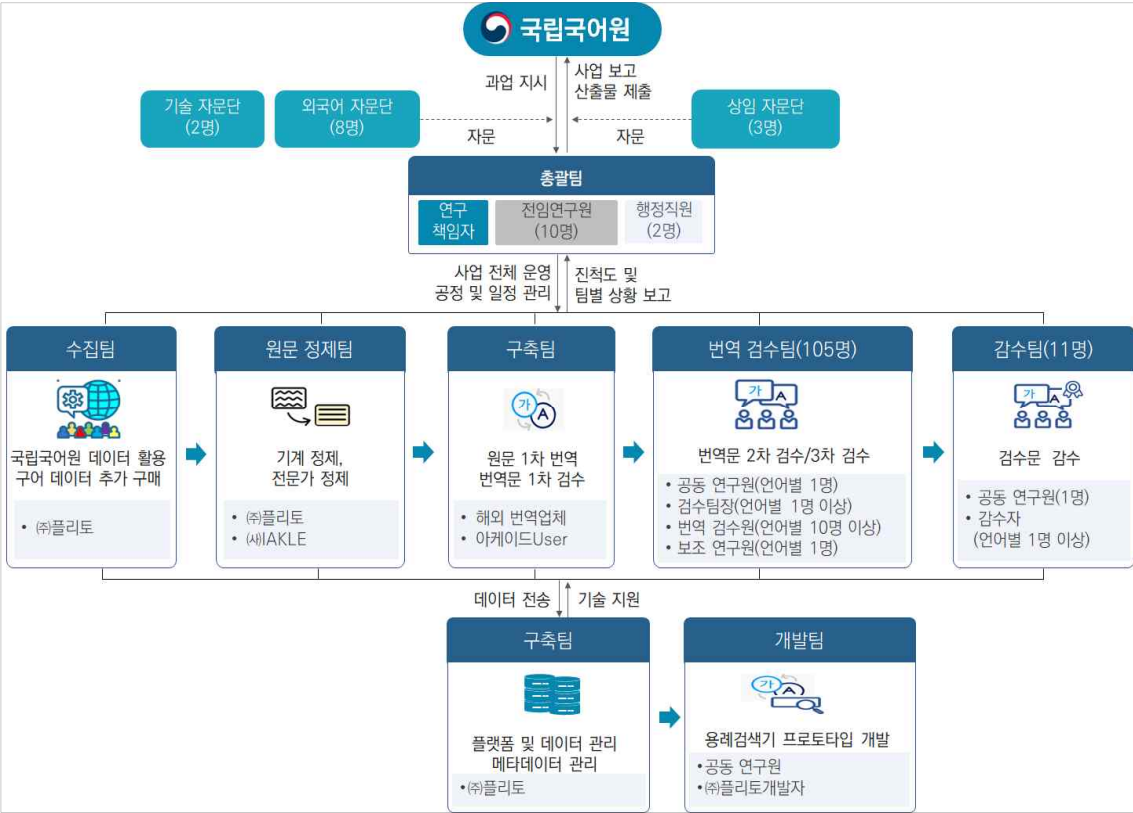
총괄팀은 (사)국제한국어교육학회 소속의 연구 책임자 1명과 전임 연구원 10명, 행정 직원 2명으로 구성하여 사업 계획 및 관리 등 전반적인 업무를 수행하였다. 자문 위원단은 다시 세 분야로 나누어 국제한국어교육학회 전임 회장들로 구성된 상임 자문단, 8개 언어와 관련한 외국어 자문단, 말뭉치 구축 기술 관련한 기술 자문단을 정기적으로 운영하여 사업 수행에 대한 자문을 구하였다.

다음으로 수집팀과 원문 정제팀, 구축팀, 번역 검수팀, 감수팀은 병렬 말뭉치 데



이터에 직접적으로 관여한다. (주)플리토에서는 수집팀, 구축팀을 이루어 원시 데이터를 수집·정제·번역하였다. 번역 전에 전문가 수준의 정제를 수행하기 위하여 (사)국제한국어교육학회에서 원문 정제팀을 운영하였다. 번역 검수팀과 감수팀도 (사)국제한국어교육학회에서 운영하여 번역 데이터를 검수·감수함으로써 최종 산출물 데이터의 품질을 제고하는 작업을 수행하였다.

본 사업에서는 말뭉치 구축 외에 병렬 말뭉치를 활용한 웹 기반 용례 검색기의 개발 가능성을 타진하고자 하였다. 이를 전담하는 개발팀을 새롭게 조직하고 용례 검색기 프로토타입을 개발하였다. 사업단의 수행 체계와 직책별 역할에 대한 상세한 내용은 다음과 같다.



[그림 2] 사업 수행 체계 및 인력 구성

<표 3> 사업단 직책별 역할

팀명	직책	업무 내용
총괄팀	연구 책임자	<ul style="list-style-type: none"> <li>- 사업 계획 전반의 수립</li> <li>- 세부 사업 과제의 조정 및 총괄</li> <li>- 사업 인력 배치 및 관리·감독</li> <li>- 사업비의 집행 전반에 대한 관리·감독</li> </ul>

		<ul style="list-style-type: none"> <li>- 사업 중간 및 최종 결과 보고</li> <li>- 국립국어원 및 컨소시엄사 상시 소통</li> </ul>
	전임 연구원	<ul style="list-style-type: none"> <li>- 사업 세부 계획의 수립 및 진행 상황 점검·조율</li> <li>- 언어별 검수 및 감수 인력 및 진척률 관리</li> <li>- 산출물 품질 및 사업 위험 관리</li> <li>- 번역 지침 및 교육 자료 보완</li> <li>- 각종 회의 주관 및 운영</li> <li>- 사업 중간 및 최종 결과 보고 지원</li> <li>- 기타 사업 수행 지원</li> </ul>
	행정 직원	<ul style="list-style-type: none"> <li>- 사업 관련 행정 사무 및 관리</li> <li>- 사업비 집행 및 정산, 증빙 자료 관리</li> <li>- 기타 사업단 운영 지원</li> </ul>
번역 검수팀	공동 연구원	<ul style="list-style-type: none"> <li>- 검수 진행 일정 점검 및 조율</li> <li>- 번역 지침 보완 지원</li> <li>- 검수에 대한 국어학적 지원</li> <li>- 검수 전반에 대해 총괄팀과 의사소통</li> <li>- 각종 회의 참석</li> </ul>
	검수팀장	<ul style="list-style-type: none"> <li>- 2차 검수(전수) 및 3차 검수(20%) 실시</li> <li>- 번역 검수 지침 보완</li> <li>- 검수 전반의 오류 파악 및 분석</li> <li>- 번역 검수원 대상 재교육 실시</li> <li>- 각종 회의 참석</li> </ul>
	번역 검수원	<ul style="list-style-type: none"> <li>- 2차 검수(전수) 실시</li> <li>- 검수 시 논의 필요 사항 보고</li> <li>- 번역 지침에 대한 피드백 제공</li> <li>- 각종 회의 참석</li> </ul>
	보조 연구원	<ul style="list-style-type: none"> <li>- 팀 내 행정 업무 보조</li> <li>- 팀 회의 운영 지원 및 회의록 작성</li> <li>- 한국어 원문 이해 지원</li> <li>- 한국어 원문 정제 실시</li> </ul>
감수팀	공동 연구원	<ul style="list-style-type: none"> <li>- 감수 진행 일정 점검 및 관리</li> <li>- 번역 지침 보완 지원</li> <li>- 감수 전반에 대해 총괄팀과 의사소통</li> <li>- 각종 회의 참석</li> </ul>
	감수자	<ul style="list-style-type: none"> <li>- 감수(10%) 실시</li> </ul>

		<ul style="list-style-type: none"> <li>- 번역 지침 보완 지원</li> <li>- 감수 상황 및 논의 필요 사항 보고</li> <li>- 각종 회의 참석</li> </ul>
원문 정제팀	보조 연구원	<ul style="list-style-type: none"> <li>- 한국어 원문 정제 실시</li> <li>- 원문 정제 관련 회의 참석</li> </ul>
개발팀	공동 연구원	<ul style="list-style-type: none"> <li>- 용례 검색기 프로토타입 개발 요구 분석 및 설계</li> <li>- 개발 진행 일정 점검 및 관리</li> <li>- 개발 상황 및 논의 필요 사항 보고</li> <li>- 각종 회의 참석</li> </ul>
자문 위원단	상임 자문위원	<ul style="list-style-type: none"> <li>- 사업 추진 방향 자문</li> <li>- 병렬 말뭉치에 대한 국어학적 자문</li> <li>- 말뭉치 구축 과정과 체계에 대한 자문</li> <li>- 병렬 말뭉치 활용 방안에 대한 자문</li> </ul>
	외국어 자문위원	<ul style="list-style-type: none"> <li>- 언어별 번역 및 감수 지침에 대한 자문</li> <li>- 번역 및 감수 과정에서 발생하는 문제에 대한 자문</li> <li>- 병렬 말뭉치 활용 방안에 대한 자문</li> </ul>
	기술 자문위원	<ul style="list-style-type: none"> <li>- 병렬 말뭉치 설계 및 구축 자문</li> <li>- 말뭉치 저작 도구에 대한 자문</li> <li>- 병렬 말뭉치 활용 방안에 대한 자문</li> <li>- 용례 검색기 개발에 대한 자문</li> </ul>
(주) 플리토	총괄 책임자	<ul style="list-style-type: none"> <li>- 컨소시엄 사업 계획 전반 수립</li> </ul>
	실무 책임자	<ul style="list-style-type: none"> <li>- 컨소시엄 세부 사업 과제의 조정 및 총괄</li> <li>- 컨소시엄 사업 인력 배치 및 관리·감독</li> <li>- 컨소시엄 사업비 집행 전반에 대한 관리·감독</li> <li>- 데이터 보안 관리·감독</li> </ul>
	사업 운영 담당자	<ul style="list-style-type: none"> <li>- 컨소시엄 사업 관련 행정 사무 및 관리</li> <li>- 데이터 보안 관리</li> <li>- 원시 데이터 수집 및 관리</li> <li>- 번역 인력 및 진척 관리</li> <li>- 플리토 내부 언어 전문가 관리</li> <li>- 최종 데이터 산출 및 관리</li> </ul>
	번역 인력	<ul style="list-style-type: none"> <li>- 한국어 원문 번역</li> </ul>
	언어 전문가	<ul style="list-style-type: none"> <li>- 번역문 1차 감수(5%)</li> <li>- 번역문 품질 관리</li> </ul>

#### 4. 사업 수행 내용

<표 4> 월별 사업 추진 경과

월 추진 과업	2023년															
	5월		6월		7월		8월		9월		10월		11월		12월	
주요 일정			착수보고				요구정의 감리		중간보고 설계감리		중간감사				최종보고 최종감리 최종감사	
사업 계획 수립																
각종 지침 보완																
참여 인력 교육																
원문 수집 및 정제																
번역 및 1차 검수(5%)																
2차 검수(전수)																
3차 검수(20%)																
감수(10%)																
최종 점검 및 산출물 납품																

##### 1) 사업 계획 수립 및 착수 보고

○ 사업 계획 수립: 2023년 5월 31일 완료

○ 착수 보고회: 2023년 6월 13일 실시

##### 2) 원문 구축 및 정제

○ 원시 데이터 수집

- 한국어 원시 데이터: 2023년 8월 9일 완료

- 저작권 확보: 2023년 12월 29일 완료

○ 원시 데이터 검수 및 정제: 2023년 9월 11일 완료

### 3) 각종 지침 보완 및 참여 인력 교육

○ 지침 보완: 2023년 6월 20일 완료

○ 지침 교육: 2023년 8월 14일 완료

- 이후 인력 총원에 따라 수시로 지침 교육 실시

### 4) 번역 및 1차 검수

○ 번역 업체 선정 및 계약: 2023년 6월 19일 완료

○ 번역 및 1차 검수(5%): 2023년 11월 3일 완료

### 5) 2·3차 검수 및 감수

○ 번역 데이터 검수: 2023년 12월 29일 완료

- 2차 검수(전수), 3차 검수(20%)

○ 번역 데이터 감수(10%): 2023년 12월 29일 완료

### 6) 의사소통 및 자문 의견 수렴

○ 단계별 보고

- 착수 보고회: 2023년 6월 13일 실시

- 중간 보고회: 2023년 9월 18일 실시

- 최종 보고회: 2023년 12월 27일 실시

○ 발주처 보고 및 컨소시엄사 회의

- 발주처 격주 보고 실시

- 컨소시엄사 격주 회의 진행

○ 자문 의견 수렴

- 상임 자문 회의 실시(4회)
- 외국어 자문 회의 실시(2회)
- 기술 자문 회의 실시(2회)

## 7) 사업 효과 확산

### ○ 국제 심포지엄 개최: 2023년 12월 8일 실시

- 주제: 한국어-외국어 병렬 말뭉치 구축의 활용과 응용
- 장소: 대한상공회의소 의원회의실 / 온·오프라인 병행
- 프로그램 구성
  - 병렬 말뭉치 활용 연구 및 웹 기반 용례 검색기 관련 주제 발표
  - 병렬 말뭉치 구축의 활용과 응용에 대해 심도 있는 패널 토의
  - <한국어-외국어 병렬 말뭉치 2021>의 산업계 활용 및 언어별 학술 연구 사례 발표

### ○ 논문 게재

- 한국어-외국어 병렬 말뭉치 관련 연구 결과를 KCI 등재 학술지에 투고하여 3편이 게재됨.

### ○ 해외 출장

- 필리핀: 2023년 9월 3일~7일



## 제 2 장

## 사업 수행







## 1. 원문 수집 및 정제

### 1.1. 원문 수집

#### 1) 수집 목표 및 수집 결과

본 사업단에서는 한국어와 외국어의 병렬 말뭉치를 구축하기 위해 먼저 한국어 원문을 수집하였다. 고품질의 원천 데이터 확보를 위한 검수·교정 과정에서 사용 불가 문장, 저품질 문장 등이 제거·교정되며 수량의 차이가 생길 것을 고려해 전반적으로 20% 상향하여 수집하였다. 수집 목표와 실제 수집 수량은 다음과 같다.

<표 5> 원문 수집 목표

구분	목표 기준	구매/수집 목표 (20% 상향)	획득 방식	세부 내용
	수량(어절)	수량(어절)		
구어체	828,000	844,560	국립국어원 제공	국립국어원 일상 대화 말뭉치 2022 활용
		149,040	무상 기부	저작권 계약(무상 기부) 후 활용
문어체	552,000	662,400	국립국어원 제공	국립국어원 신문 말뭉치 2023 활용

<표 6> 원문 수집 수량

구분	목표 기준	실제 구매/수집량 (20% 상향)	획득 방식	출처
	수량(어절)	수량(어절)		
구어체	828,000	1,191,862 (약 41% 상향 수집)	국립국어원 제공	국립국어원 일상 대화 말뭉치 2022
		179,560 (약 20% 상향 수집)	무상 기부	라이나 전성기 재단
문어체	552,000	665,643 (약 20% 상향 수집)	국립국어원 제공	국립국어원 신문 말뭉치 2023

## 2) 원문 선정 과정

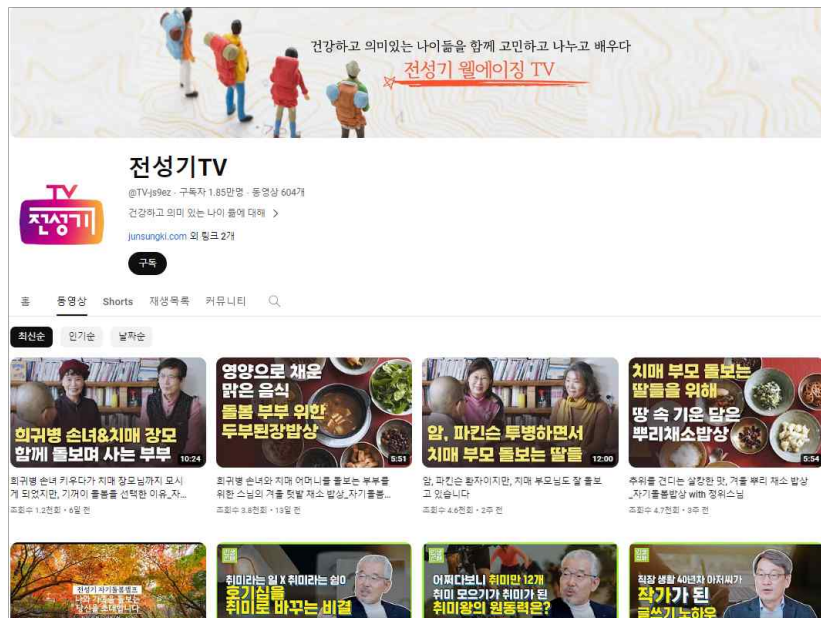
말뭉치에는 언중들의 사고방식 및 의식 구조가 반영되어 있어 특정한 영역에만 국한되지 않고 다양한 주제와 정보를 포함하여 원문을 활용하는 것이 중요하다.

이에 본 사업단에서는 고품질의 데이터 수집을 위해 기구축된 국립국어원의 신문 말뭉치 2023과 일상 대화 말뭉치 2022를 문어체와 구어체로 활용하였고, 자료 수집의 용이성, 자료 활용성, 시장의 규모, 현재 언어의 경향 등을 종합적으로 판단하여 구어체로 활용 가능한 가장 적합한 원시 데이터로 유튜브 대본을 선택하였다.

또한, 말뭉치를 활용할 때 보다 효과적으로 다양한 목적에 사용될 수 있도록 원문 선정 과정에서 주제, 성별, 연령대 등의 다양한 메타 정보를 골고루 포함하여 말뭉치의 편향성을 사전에 방지하였다.

신문 말뭉치의 경우 국내 뉴스 기사 자료를 토대로 현재의 정치, 사회, 문화 등의 다양한 이슈를 모두 다루고 있어 여러 산업에 영향력을 미칠 수 있으며, 글쓰기의 토대인 육하원칙을 바탕으로 작성되어 있어 분명한 의사 표현을 담고 있으므로 매우 적절한 자료로 판단하였다.

또한 유튜브 대본은 중·장년을 대상으로 촬영된 유튜브 채널 ‘전성기 TV’에서 추출한 대본 및 출판물 ‘전성기 웰에이징 시리즈’의 인터뷰 원고를 확보하여 부족한 고령층의 데이터를 확보하고, 중장년 화자들의 관점, 언어 사용 환경과 의사 표현이 반영되어 있기에 본 과제에 활용하기에 적절한 자료로 판단하였다.



[그림 3] 유튜브 채널 ‘전성기 TV’

### 3) 원문 수집 절차

#### (1) 구어체

구어체 원문은 유튜브 대본, 일상 대화 말뭉치로 구분하여 수집하였다. 유튜브 대본은 재단법인 라이나전성기재단과의 직접 계약을 통해 중·장년과 밀접한 주제를 대상으로 촬영된 영상의 대본을 구매하였으며 일상 대화 말뭉치는 국립국어원의 일상 대화 말뭉치 2022 자료를 제공 받아 원천 자료로 활용하였다.

<표 7> 일상 대화 말뭉치 수집 진행 방식

		순서	확인 방식
<pre> 1 { 2   "id": "SDRW2200000001", 3   "metadata": { 4     "title": "국립국어원 구어 말뭉치 SDRW2200000001", 5     "creator": "국립국어원", 6     "distributor": "국립국어원", 7     "year": "2022", 8     "category": "구어 &gt; 사적대화 &gt; 일상대화", 9     "annotation_level": [ 10      "원시" 11    ], 12    "sampling": "본문 전체" 13  }, 14  "document": [ 15    { 16      "id": "SDRW2200000001.1", 17      "metadata": { 18        "title": "3인 일상 대화", 19        "author": "개인 발화자", 20        "publisher": "개인 발화 녹음", 21        "date": "20220824", 22        "topic": "회사/학교 &gt; 학교 생활 및 전공 이야기", 23        "environment": "", 24        "speaker": [ 25          { 26            "id": "SD22000005", 27            "age": "20대", 28            "occupation": "학생", 29            "sex": "남성", 30            "birthplace": "서울", 31            "principal_residence": "서울", 32            "current_residence": "서울", 33            "education": "대재" 34          } 35        ] 36      } 37    } 38  ] 39 } </pre>		1	공식 절차를 통해 자료 요청
		2	JSON 변환
		3	Metadata 분류
		4	사업 특성에 맞는 Metadata, form 추출

<표 8> 구어체 원문 정보

번호	분야	출처	세부 분야
1	일상 대화 말뭉치 2022 (2인 일상대화)	국립국어원	가족/관혼상제
			건강/다이어트
			경제/재테크
			기타
			대중교통
			먹거리

			반려동물
			방송/연예
			생활/주거환경
			쇼핑
			스포츠/레저/취미
			우정
			음악
			취직
			회사/학교
			휴가
2	유튜브 대본 (2인 인터뷰)	라이나전성기재단	건강/다이어트
			기타
			반려동물
			생활/주거환경
			스포츠/레저/취미
			여행일반
			은퇴/노후
			음악
			책/독서

## (2) 문어체

문어체 원문은 국립국어원의 신문 말뭉치 2023 자료를 제공받아 신문 말뭉치에 포함된 문화·연예, 사회, 정치 등 다양한 분야의 신문 기사를 활용하였으며, 세부 내용은 다음과 같다.

<표 9> 문어체 원문 정보

번호	분야	출처	세부 분야
1	IT·과학	국립국어원	IT·과학일반

			과학
			모바일
			보안
			인터넷·SNS
			콘텐츠
2	정치		선거
			정치일반
			청와대
			행정·자치
3	사회		교육·시험
			교육·입시
			날씨
			노동·복지
			여성
			의료·건강
			장애인
			환경
4	경제		경제일반
			부동산
			취업·창업
5	문화		대중음악
			문화일반
			미술·건축
			방송·연예
			생활
			연극·클래식
			영화
			음악

			전시·공연
			책
			출판
			학술·문화재
6	지역		강원/경기/경남/경북/광주/대구/ 대전/부산/울산/전남/전북/제주/ 충남/충북/지역일반
7	국제		국제일반
			러시아
			미국·북미
			아시아
			유럽·EU
			일본
			중국
			중남미

### (3) 원문 예시

위 절차로 수집된 구어체와 문어체 원문의 예시는 다음과 같다.

<표 10> 원문 예시

구분	출처	수집 원문
문어체	국립국어원 신문 말뭉치 2023	주거지역의 중심에 위치한 투표소로 다양한 연령대의 시민들을 볼 수 있었다.
		당진제철소에서 사망 사고가 발생한 지 불과 사흘만에 다시 사망자가 발생했다.
		선관위의 5일 발표에 따르면 4~5일 이틀 동안 유권자 4419만 7692명 중 1632만3602명이 사전투표에 참여했다.
		일반시민과 기업인들에게는 지역 예술작가들의 작품을 구매해 소장할 수 있는 기회의 장이 된다.

		인천에 거주하는 30대 A씨는 코로나19 증상을 보이자 한 이비인후과를 찾아 진료를 받고 처방전을 요구했지만, 병원 측은 'PCR 검사를 한 뒤에 처방전을 주겠다'고 안내했다.
구어체	국립국어원 (일상 대화 말뭉치)	나는 솔직히 말하면 애들한테 좀 미안한 마음이 많아서 '나처럼 살라'고 하고 싶진 않아.
		마사지도 받고 맛있는 것도 많이 먹고 잘 쉬다 온 거 같아.
		나도 그 영화 봤는데 그래서 핀란드에 꼭 가 보고 싶어졌던 거 같아.
		그러니까 name2, name3만 전날 안 가면 같이 가면 되는데, 다음 날 갈 사람 중에 같이 갈 사람들이 없어.
		혹시 그러면 저, 사장님은 친구들끼리 가시나요?
	라이나전성기재단	그런데 내가 직장을 그만두고 딱, 나오잖아요.
		등산 가서도 저를 아시는 분이 만나서 "어? name1 아니세요?" 하면 깜짝깜짝 경기 일으켜요.
		요즘에는 신혼 주부분들이 "결혼해서 처음으로 요리 시작하는데 배우게 되어서 너무 즐겁고 행복하다.", 어느 시청자분들은 "머느리 되고 싶다." 그런 댓글도 있더라고요.
		40대 이상 되면 잇몸병으로 멀쩡한 치아를 잃게 되는 일이 많거든요.
		퇴직 후 불면증이 생겼다면 낮 동안 계속 움직이세요.

## 1.2. 원문 정제

이상의 과정을 통해 수집한 원문은 번역에 용이하도록 정제 과정을 거쳤다. 본 사업단에서는 최소 3단계에 걸쳐 정제를 진행하였으며 이를 통해 원문 품질을 최대한 높일 수 있도록 하였다. 정제 단계별 세부 내용은 다음과 같다.

### 1) 원문 정제 절차

#### (1) 1차 정제(기계적 정제)

수집한 한국어 원문은 가장 먼저 기계적인 정제 과정을 통해 어절 수, 특수 문자(이모티콘), 개인 정보, 유사·일치 문장의 삭제 등 문장을 1차적으로 다듬는 작업을 진행하였다.

<표 11> 기계적 정제 기준

정제 기준	내용
어절 수	구어체: 평균 10어절, 문어체: 평균 15어절
부적합 내용	부적절하거나 자연스럽지 않은 내용, 불필요한 발화, 오류가 있는 문장(오·탈자 포함), 이모티콘 등 삭제
개인 정보	전화번호, 이메일 등의 개인 정보 비식별화 및 삭제
중복 내용	완전 일치 문장 및 유사 일치 문장 삭제
문장 형태 정제	형태소, 발화자 정보를 분석해 비구조화된 문장 정제

## (2) 2차 정제(전문가 정제)

본 사업의 결과물은 국립국어원이 구축한 공공 데이터이며, 8개 언어 병렬 말뭉치의 기준이 될 수 있는 기초 자료이므로 규범성과 표준성을 지향해야 한다. 따라서 표현의 다양성은 유지하되, 원칙과 허용이 있는 항목에는 원칙을 적용하여 일관성을 확보할 필요가 있다고 보았다.

또한 공공 데이터로서 공공성과 윤리성을 최대한 확보하되 언어의 실제성도 고려해야 한다. 이는 일반 언중 수준에서는 적용하기가 어려우며 언어학이나 한국어학 관련 전공자들이 전공 지식을 활용하여 고민이 필요한 수준의 작업이다.





게다가 고맥락적인 한국어의 특성상 텍스트에서 하나의 문장만을 떼어 놓으면 의미를 파악하는 것이 쉽지 않을 때가 많으며, 한국어는 문장 성분의 생략과 어순변경이 자유로운 편이다. 이러한 한국어의 특성으로 인해 언어에 따라 번역이 어려울 수 있으므로 이를 고려한 정제 역시 필요하다고 보았다.

이에 본 사업에서는 공공 데이터로서 병렬 말뭉치의 규범성과 적절성을 제고하기 위하여 2021년·2022년 사업의 원문 정제 지침을 재점검하였으며, 전문가 수준의 원문 정제팀을 별도로 구성·운영하였다. 원문 정제팀은 담당 전임 연구원 2명, 공동 연구원 8명, 한국어학 및 언어학 분야의 학사·석사 이상의 정제원 15명으로 구성하였다.

## (3) 3차 정제(보완 정제)

2차 정제가 완료된 원문은 번역 후 검수를 진행하였다. 검수 시 번역 검수원들은 원문을 확인하여 원문에 오류(단순 오타자, 의미 불분명 등)가 있을 경우 ‘원문 신고’ 기능을 활용하여 원문의 오류를 보고하였다.



OriginText	Translation	QC
<p>통창이고, 어쨌든 <span style="border: 1px solid red; padding: 2px;">충수</span>도 높고, 뷰도 좀 더 괜찮고. </p>	<p>Cửa sổ lớn này, dù gì thì cũng thuộc dạng cao tầng, hướng nhìn cũng đẹp nữa.</p>	<div></div> <div></div> <div></div>

[그림 4] 검수 플랫폼의 원문 신고 버튼

이때 번역 검수원은 바로 원문을 신고하지 않고, 먼저 해당 언어를 담당하는 공동·전임·보조 연구원과 논의한 후 원문에 오류가 있는 것이 확실한 경우에만 신고하도록 하였다. 신고된 원문은 전임 연구원들이 재검토하여 의미의 변화가 없는 수준에서 적절한 문장으로 수정하거나 이것이 어려운 경우 삭제하였으며, 재수정된 원문을 대상으로 재번역 및 검수를 실시하였다.

또한 검수와 감수 작업이 진행되는 동안 원문 정제팀에서는 다시 한번 원문을 전수 검수하였다. 이와 같이 원문 종류에 따라 최소 4차례 이상 전수 검수하는 과정을 통해 원문의 품질을 최대한 제고하고자 하였다.

#### (4) 원문 정제 지침 교육

아무리 높은 수준의 정제자라고 하더라도, 여러 정제자가 동시에 작업을 하므로 원문 정제의 통일성을 확보할 필요가 있었다. 이를 위해 원문 정제를 실시하는 작업자 전원을 대상으로 정제 지침 교육을 실시하였으며 문어체와 구어체의 원문 성격이 다른 만큼 2차례에 걸쳐 실시하였다.

원문 정제 지침	목차
제1장 목적 및 방향	
제1절 목적	
제2절 정제 기본 방향	
제2장 정제 절차	제3장 세부 정제 기준
제1절 정제 절차(기계 정제)	제1절 한국어 어문 규범에 따른 정제
제2절 2차 정제(인적 보완 정제)	제2절 의미/내용/맥락에 따른 정제
제3절 3차 정제(전문가 정제)	제3절 개인 정보 및 비윤리적 내용 처리
제4절 보완 정제	제4절 원문 명료화 및 외국어 번역을 고려한 조치
	제5절 기타
	제6절 참고 자료

[그림 5] 원문 정제 지침 교육 예시

#### (5) 원문 정제 예시

항목별 원문 정제의 예시는 다음과 같다.

<표 12> 항목별 원문 정제 예시

구분	원문 및 정제 내용
오자 및 비표준어	<ul style="list-style-type: none"> <li>새우젓 안 짜게 할라고(→하러고) 막 거기에다가 양파 썰어 넣고 마늘 썰어 넣고 고춧가루 섞고 막 고추 다져서 섞고 그러잖아.</li> </ul>
외래어와 로마자 및 한자	<ul style="list-style-type: none"> <li>근데 빔 프로젝트(→프로젝터)도 그 화질에 따라서 가격이 천 차만별이어서 좋은 거 안 사면 그냥 그렇다고 하더라고.</li> </ul>
띄어쓰기	<ul style="list-style-type: none"> <li>그러면 나는 홈쇼핑(→홈 쇼핑)을 틀어.</li> </ul>
비윤리적 내용	<ul style="list-style-type: none"> <li>내가 이렇게 경상도 사람으로서 경상도 사람이 싫은 이유가 신라 때문에 내가 경상도 사람이 싫어.</li> </ul> <p>→ 문장 삭제</p>
원문 명료화	<ul style="list-style-type: none"> <li>나도 어 아무래도 언어의 장벽이 조금 크더라고 생각보다 나가 보니까.</li> </ul> <p>→ 나도 어, 아무래도 언어의 장벽이 생각보다 조금 크더라고,</p>

	나가 보니까.
문장 부호와 특수 문자	<ul style="list-style-type: none"> <li>그냥 계속 누워만 있어야 되고(→되고,) 맞아.</li> </ul>
인용문과 따옴표	<ul style="list-style-type: none"> <li>그러니까 맛있는 재료라 할지라도 아 이걸 어떻게 먹어야지 맛있지라는 생각을 하지.</li> <li>→ 그러니까 맛있는 재료라 할지라도 '아, 이걸 어떻게 먹어야지 맛있지?'라는 생각을 하지.</li> </ul>

띄어쓰기는 원칙과 허용이 있으나, 2022년 사업에서 국립국어원과의 협의에 따라 일관성을 확보하기 위해 원칙을 따르는 것으로 결정하였으며 본 사업에서도 이를 따랐다. 전문 용어와 고유 명사의 띄어쓰기는 단어별로 띄어 쓰는 것을 원칙으로 하되 구어의 경우 언어 사용의 실제성을 고려하여 붙여 쓰는 것을 허용하였다. 단, 이 경우 한 텍스트 안에서는 띄어쓰기 여부가 통일되도록 하였다.

<표 13> 띄어쓰기 정제 예시

구분	원문 및 정제 내용
합성어	<ul style="list-style-type: none"> <li>그 다음날(→그다음날)이 조금 피곤하지 않겠어?</li> </ul>
의존 명사	<ul style="list-style-type: none"> <li>당진 제철소에서 사망 사고가 발생한 지 불과 사흘만(→사흘 만)에 다시 사망자가 발생했다.</li> </ul>
접사	<ul style="list-style-type: none"> <li>산책 식(→산책식)으로 그런 여행도 참 괜찮을 거 같아요.</li> </ul>
보조 용언	<ul style="list-style-type: none"> <li>하나 먹어보니까(→먹어 보니까) 되게 맛있더라.</li> </ul>
전문 용어	<ul style="list-style-type: none"> <li>상·하반기 각각 6개월의 급여가 있는 유급 과정으로, 재학하고 있는 학교나 교육청, 직속 기관과 같은 교육기관(→교육 기관)에서 청소, 급식, 사무 분야의 보조 업무를 맡게 된다.</li> </ul>
고유 명사	<ul style="list-style-type: none"> <li>아무래도 에펠탑(→에펠 탑)은 너무나 유명한 관광지잖아.</li> </ul>

본 사업에서 사용한 문어체 원문은 신문 기사이다. 기사문의 장르적 특성상 직접 인용 표현과 간접 인용 표현이 혼재되어 있는 경우가 많아 한국어 어문 규범에 맞게 기호와 인용 표지를 수정하였다. 간접 인용문에 작은따옴표를 사용해야 한다는 규범은 없으나 ‘문장 내용 중에서 주의가 미쳐야 할 곳이나 중요한 부분을 특별히 드러내 보일 때에도 작은따옴표를 쓸 수 있다’는 한국어 어문 규범의 해설에 따라, 문장에서 인용된 부분을 표시하여 번역에 도움이 될 수 있도록 하는 경우 작은따옴표 사용을 허용하였다.

<표 14> 인용문 정제 예시

구분	원문 및 정제 내용
직접 → 간접	<ul style="list-style-type: none"> <li>"생일 같다"고 말할 정도로 많은 선물도 받았다.</li> </ul> → '생일 같다'고 말할 정도로 많은 선물도 받았다.
간접 → 직접	<ul style="list-style-type: none"> <li>'이분도 별거 없네'라고 덧붙였다.</li> </ul> → "이분도 별거 없네."라고 덧붙였다.

구어체 원문의 경우 일상생활을 주제로 한 대화 및 유튜브 대본이라는 원문의 특성상 줄임말, 외래어 등을 포함한 신조어가 많았다. 이에 번역 용이성 제고를 위해 별도로 공유 문서를 활용하여 번역·검수·감수에 활용할 수 있도록 하였다.

신조어							
원문 또는 정제 데이터	≡ 신조어 ≡	유형	≡ 비교 ≡	≡ 우리말샘 등재 ≡	≡ 우리말샘 미감수/미등재 ≡	≡ 조사 ≡	≡ 뜻풀이 ≡
나는 초코아이스크림이 더 좋는데, 근데 왜 맞팔 안 했어?	초코아이스크림	합성어	초코+아이스크림	O		명사	초콜릿이 들어간 아이스크림
이제부터 우리 인천인 거네.	맞팔	줄임말	맞팔로우(follow)	O		명사	누리 소통망 서비스에서 사용자가 서로 팔로 관계를 맺음. 또는 그런 것.
혹시 페이스북이나 트위터도 해?	인친	줄임말	인스타(Instagram) 친구		O	명사	누리 소통망 서비스인 '인스타그램(Instagram)'상의 '친구'를 줄여 이르는 말.
초록볼에도 자 조심하면서 건너야 하잖아.	페이스북	줄임말	페이스북(facebook)		O	명사	누리 소통망 서비스인 '페이스북(facebook)'을 줄여 이르는 말.
그럼 난 새깅 패션으로 한 뒤에 길을 건너야겠다.	초록볼	합성어	초록+볼	O		명사	'홍신호'를 일상적으로 이르는 말.
그래, 여기는 이렇게 그냥 눈치 게임으로 건너나 봐.	새깅	외래어	sagging		O	명사	바지나 청바지의 윗부분이 허리 아래로 말하도록 처지는 방식으로, 때때로 착용자의 속바지 대부분이 드러나는 것을 말한다.
호돌 가면 우리 노쇼로 못 들어가는 건 아니었지?	눈치 게임			O		명사 구	어떤 일에 나서기 전 눈치를 보아 가며 경장함을 비유적으로 이르는 말.
와, 지금 기상 악화로 할주로 다 섰다운이네.	노쇼	외래어	no-show	O		명사	오기로 한 사람이 예약이나 약속을 취소하지 않고 나타나지 않는 일. 또는 그런 사람.
나는 새우튀김이랑 오징어튀김이 맛있더라.	셋다운	외래어	shutdown	O		명사	전원 공급의 중단이나 사고, 기타 오류 따위의 이유로 컴퓨터 시스템의 작동이 중지되는 일.
직원을 피싱하는 화요일, 수요일에 사람이 적대	오징어튀김	합성어	오징어+튀김	O		명사	먹기 좋게 자른 오징어에 말가루를 뿌워서 기름에 튀긴 음식.
와들어랑 망고주스도 파나 봐!	피싱	줄임말	오피셜(official)	O		명사	특정한 사람이나 단체 따위에서 발언하거나 발표한 이야기를 이르는 말.
저기 리어카에서 탕후루를 팔고 있네.	망고주스	외래어	mango juice	O		명사	망고의 즙을 내어 단맛을 더한 음료.
나는 달달한 씨앗호떡이 먹고 싶어	탕후루	외래어	tanghulu(糖葫蘆)	O		명사	중국 베이징의 전통 음식. 산사자, 해당화 열매 따위를 꼬챙이에 꿰 다음 물엿 등을 발라 굳혀 만든다.
같이 버스킹 보는 것도 재밌을 거 같아.	씨앗호떡	합성어	씨앗+호떡	O		명사	호박, 해바라기 따위의 씨앗으로 만든 소를 넣은 호떡.
	버스킹	외래어	busking	O		명사	사람들이 많이 다니는 길거리에서 여는 공연.

[그림 6] 신조어 목록 예시

### 1.3. 개인 정보 비식별화

본 사업에서 구축하는 병렬 말뭉치 데이터는 공개 데이터이만큼 개인 정보가 노출되지 않도록 정제 단계에서 개인 정보를 비식별화하는 과정을 거쳤다.

기계 정제 단계에서 수집된 원문에 개인 정보가 포함되어 있는지 전산 프로그램의 기계 정규식을 통해 확인 후 제거할 수 있도록 조치하였다.

이후 전문가 정제 단계에서도 기업명·상표명 등이 상업적 홍보 목적으로 표현되거나 유명인의 이름이 비윤리적 표현이나 민감한 내용으로 이어지는 경우 삭제하거나 비식별화 처리하였다.

```

def make_personal_info_col(df, text_col, result_col):
    df = df.with_columns(
        pl.when(
            pl.col(text_col).apply(
                lambda x: re.findall(r"[0-9]{2,4}-[0-9]{3,4}-[0-9]{3,4}", x) != []
            )
        ).then("Y")
        .when(
            pl.col(text_col).apply(
                lambda x: re.findall(
                    r"[a-zA-Z0-9\.\@\-\_\/&]{2,}\.[a-zA-Z0-9\.\@\-\_\/&]{2,}", x
                ) != []
            )
        ).then("Y")
        .when(pl.col(text_col).apply(lambda x: x.startswith("기획 ")))
        .then("Y")
        .otherwise("")
        .alias(result_col)
    )
    return df

```

[그림 7] 전산 프로그램을 이용한 기계적 개인 정보 처리 과정 예시

#### 1.4. 저작권 확보

수집 대상으로 선정된 자료에 대한 저작권 확보와 번역물에 대한 자유 배포가 가능하도록 양도 계약서 및 저작권 이용 허락 계약서를 체결해 저작권 문제를 해소하였다.

구어체의 경우 저작권의 귀속 주체인 라이나전성기재단과 계약을 진행하였고, 구축 완료된 저작물의 경우 양수인 국립국어원, 양도인 사단법인 국제한국어교육학회와 주식회사 플리토의 양도 계약을 별도로 체결하였다.

구분	계약서
<p>국가 언어 자원 구축(말뭉치) 및 활용 저작권 이용 허락 계약서 / 라이나전성기재단</p>	<p style="text-align: center;"><b>국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서</b></p> <p>저작자 및 저작권 이용허락자 <u>라이나전성기재단</u> (이하 "권리자"이라 함)과 저작권 이용자 국립국어원(이하 "이용자"이라 함)은 아래 저작물에 관한 저작권재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.</p> <p><b>제1조 (계약의 목적)</b> 본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위해 권리자가 무상으로 기증하는 저작물의 저작권재산권 이용허락과 관련하여 권리자와 이용자 사이의 저작물 권리관계와 이용허락 범위를 명확히 하는 것을 목적으로 한다.</p> <p><b>제2조 (정의)</b> 본 계약에서 사용하는 용어의 뜻은 다음과 같다. (1) '수집 원문'이라 함은 국립국어원 및 국립국어원이 보조한 보조사업의 수행자(2차 보조 또는 용역 사업자를 포함한다.)(이하 "과업수행자")가 2023년 한국어-외국어 병렬 말뭉치 구축 사업(사업기간: 2023.5.~2023.12.)의 번역 대상이 되는 한국어 원문 또는 외국어 원문을 수집한 것을 말한다. (2) '대상저작물'이라 함은 '수집 원문' 중에서 국립국어원 및 과업수행자가 병렬 말뭉치 번역 대상으로 선정한 한국어 원문 또는 외국어 원문을 말한다. (3) '복제·변형·번역물'이라 함은 국립국어원 및 과업수행자가 '대상저작물'에 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착, 원문 교정, 원문 수정, 번역 등의 처리를 더한 결과물을 말한다.</p> <p><b>제3조 (계약의 대상)</b> 본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상저작물")에 대한 저작권재산권 중 본 조에 명시한 이용 허락 범위로 한다.</p> <p>저작물: 저작자가 2023년 한국어-외국어 병렬 말뭉치 구축 사업을 위해 유튜브 채널 "전성기 TV"에서 추출한 발화 텍스트 데이터 및 출판물 "전성기 웹에이징 시리즈"의 인터뷰 원고 텍스트 데이터에서 추출하여 제공한 12,535문장 (대본 문장 내용은 첨부1의 별지로 한다) 저작자: <u>라이나전성기재단</u></p> <div style="border: 1px solid black; padding: 5px;"> <p>※저작권 이용허락 대상 권리의 내용</p> <ol style="list-style-type: none"> <li>1. 국립국어원 및 국립국어원이 보조한 보조사업의 수행자(2차 보조 또는 용역 사업자를 포함한다). 국립국어원이 발주한 용역 사업 수행자가 '수집 원문', '대상저작물', '복제·변형·번역물'을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일</li> <li>2. 국립국어원 및 국립국어원이 보조한 보조사업의 수행자(2차 보조 또는 용역 사업자를 포함</li> </ol> </div>
	<p>출처: 국가 언어 자원 구축(말뭉치) 및 활용 저작권 이용 허락 계약서_라이나전성기재단.pdf</p>

구어체의 경우 문어체와 마찬가지로 저작권 귀속 주체와 계약을 통하여 저작권 문제를 해소할 수 있도록 하였다.

구분	계약서
<p>저작 재산권 전부에 대한 양도 계약서 / (주)플리토</p>	<p style="text-align: center;"><b>저작재산권 전부에 대한 양도계약서</b></p> <p>저작권자 및 저작권 양도인 <u>주식회사 플리토</u>(이하 “양도인”이라 함)와 저작권 양수인 <u>국립국어원</u>(이하 “양수인”이라 함)은 아래 저작물 (1) 2023년 한국어-외국어 병렬 말뭉치 구축 사업을 통해 수집된 원문 자료의 수정 작업 저작물 등 산출물, (2) 2023년 한국어-외국어 병렬 말뭉치 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필리핀 타갈로그어, 태국어, 인도 힌디어, 캄보디아 크메르어, 러시아어, 우즈베크어) 번역 저작물 등 산출물에 관한 저작재산권(이하 “저작재산권”이라 함)과 관련하여 다음과 같이 계약을 체결한다.</p> <p style="text-align: center;">다 음</p> <p><b>제1조 (계약의 목적)</b> 본 계약은 저작재산권 이전과 관련하여 양도인과 양수인 사이의 권리관계를 명확히 하는 것을 목적으로 한다.</p> <p><b>제2조 (계약의 대상)</b> 본 계약의 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”이라 함)에 대한 저작재산권으로 한다.</p> <p>저작물: (1) 국립국어원의 보조사업자 (사)국제한국어교육학회와 주식회사 플리토가 수행한 2023년 한국어-외국어 병렬 말뭉치 구축 사업을 통해 수집된 원문 자료의 수정 작업 저작물 등 산출물 (2) 국립국어원의 보조사업자 (사)국제한국어교육학회와 주식회사 플리토가 수행한 2023년 한국어-외국어 병렬 말뭉치 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필리핀 타갈로그어, 태국어, 인도 힌디어, 캄보디아 크메르어, 러시아어, 우즈베크어) 번역 저작물 등 산출물</p> <p>저작권자: <u>주식회사 플리토</u> 종별: 어문저작물 권리: 저작재산권 전부 <input type="checkbox"/> 복제권, 공연권, 공중송신권(방송권, 전송권, 디지털음성송신권), 전시권, 배포권, 대여권, 2차적저작물 작성권(번역 등)</p> <p><b>제3조 (저작재산권 양도범위)</b> 본 계약에 의한 저작재산권 양도범위는 제2조에서 정한 복제권 등 저작재산권 일체를 의</p>
	<p style="text-align: center;">출처: 저작 재산권 전부에 대한 양도 계약서_(주)플리토.pdf</p>
<p>저작 재산권 전부에 대한 양도 계약서 / (사)국제한국어 교육학회</p>	<p style="text-align: center;"><b>저작재산권 전부에 대한 양도계약서</b></p> <p>저작권자 및 저작권 양도인 (사)국제한국어교육학회(이하 “양도인”이라 함)와 저작권 양수인 <u>국립국어원</u>(이하 “양수인”이라 함)은 아래 저작물 (1) 2023년 한국어-외국어 병렬 말뭉치 구축 사업을 통해 수집된 원문 자료의 수정 작업 저작물 등 산출물, (2) 2023년 한국어-외국어 병렬 말뭉치 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필리핀 타갈로그어, 태국어, 인도 힌디어, 캄보디아 크메르어, 러시아어, 우즈베크어) 번역 저작물 등 산출물의 <u>감수 저작물 등 산출물에 관한 저작재산권</u>(이하 “저작재산권”이라 함)과 관련하여 다음과 같이 계약을 체결한다.</p> <p style="text-align: center;">다 음</p> <p><b>제1조 (계약의 목적)</b> 본 계약은 저작재산권 이전과 관련하여 양도인과 양수인 사이의 권리관계를 명확히 하는 것을 목적으로 한다.</p> <p><b>제2조 (계약의 대상)</b> 본 계약의 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”이라 함)에 대한 저작재산권으로 한다.</p> <p>저작물: (1) 국립국어원의 보조사업자 (사)국제한국어교육학회와 주식회사 플리토가 수행한 2023년 한국어-외국어 병렬 말뭉치 구축 사업을 통해 수집된 원문 자료의 수정 작업 저작물 등 산출물 (2) 국립국어원의 보조사업자 (사)국제한국어교육학회와 주식회사 플리토가 수행한 2023년 한국어-외국어 병렬 말뭉치 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필리핀 타갈로그어, 태국어, 인도 힌디어, 캄보디아 크메르어, 러시아어, 우즈베크어) 번역 저작물 등 산출물의 감수 저작물 등 산출물</p> <p>저작권자: (사)국제한국어교육학회 종별 : 어문저작물 권리 : 저작재산권 전부 <input type="checkbox"/> 복제권, 공연권, 공중송신권(방송권, 전송권, 디지털음성송신권), 전시권, 배포권, 대여권, 2차적저작물 작성권(번역 등)</p>
	<p style="text-align: center;">출처: 저작 재산권 전부에 대한 양도 계약서_(사)국제한국어교육학회.pdf</p>

## 2. 번역 지침 수정·보완

2023년 한국어-외국어 병렬 말뭉치 구축을 위해 한국어를 8개 언어로 번역, 검수, 감수할 때 일관된 기준을 적용하기 위해 지난 사업 지침의 총칙과 공통 지침 및 8개 언어별 세부 지침을 수정·보완하였다.

번역 지침을 수정·보완하기 위해 먼저 지난 2021년 사업과 2022년 사업에서 마련한 번역 지침을 검토하고, 각 언어의 최신 번역 용례나 관련 규정과 자료를 참조하였다. 또한 이번 2023년 사업에서 번역될 한국어 원문의 내용과 장르 등의 특징을 검토하였다. 이후에는 실제 구어 및 문어 데이터의 초기 번역 결과를 검수해 보고, 기존 지침에서 수정·보완해야 할 사항들을 검수팀장과 번역 검수원들이 종합적으로 논의하여 반영하였다.

이렇게 수정된 지침은 플리토를 통해 번역사들에게 전달하여 번역 시 적용하도록 요청하였다. 이후 검수, 감수 진행 과정에서 언어별로 수정·보완 사항이 발견되면 개별 사항에 따라 지침을 수정·보완하여 플리토와 공유하였다. 2023년 지침의 경우, 지난 사업에서 여러 번 검토와 자문을 거친 지침을 기반으로 하였으며, 사업 간 연계성과 일관성을 유지하고자 수정·보완 사항은 제한적으로 이뤄졌다.

<표 15> 번역 지침 집필 과정

순서	작업	주요 내용
1	기존 지침 및 관련 자료 검토	<ul style="list-style-type: none"> <li>2021년 사업 및 2022년 사업에서 사용한 번역 지침 검토</li> <li>언어별 번역 용례나 규정 등 관련 자료 참고</li> </ul>
↓		
2	원문 및 초기 번역 데이터 검토	<ul style="list-style-type: none"> <li>2023년 사업의 한국어 원문 내용 검토</li> <li>구어와 문어 각 원문의 초기 번역 데이터를 검수한 후 검수원 간 논의</li> <li>기존 지침에서 추가로 수정·보완해야 할 사항이 있는지 논의</li> </ul>
↓		
3	지침 수정·보완	<ul style="list-style-type: none"> <li>검수팀장과 검수원이 종합 검토하여 수정·보완</li> <li>공동 연구원과 전임 연구원이 한국어 설명 검토</li> </ul>
↓		



4	최종본 완성 및 적용	<ul style="list-style-type: none"> <li>• 지침 최종본을 플리토를 통해 번역사에게 공유</li> <li>• 완성된 지침으로 번역, 검수, 감수에 본격적으로 적용</li> <li>• 기타 개별적 수정 사항 발생 시 검수원 간 논의 후 언어별 지침을 수정하고 재공유</li> </ul>
---	-------------	--

## 2.1. 지침 및 원문 검토

2023년 사업의 데이터에 적용할 지침을 마련하기 위해 가장 먼저 지난 2021년 사업 및 2022년 사업에서 사용됐던 기존의 총칙, 공통 지침 및 8개 언어별 세부 지침을 검토하였다. 2021년 사업의 지침은 문화체육관광부 훈령 ‘공공 용어의 외국어 번역 및 표기 지침’과 플리토의 번역 가이드라인을 기반으로 작성되었고, 국내 어문 규정 및 8개 언어 사용 지역의 관련 규정, 언어별 특징, 번역 관례, 그리고 실제 번역 데이터 검수 결과 등을 종합적으로 고려하여 작성되었다. 2022년 사업의 지침은 2021년 사업의 지침을 바탕으로 검수팀장, 검수원, 감수자, 외국어 자문 위원 등 여러 참여자의 추가 검토를 진행한 뒤 종합적으로 논의하여 수정, 보완하였다. 이번 2023년 사업에서는 지난 두 사업에서 사용된 지침의 원리와 규정을 대부분 유지하여 사업 간 연계성과 일관성을 최대한 높이하고자 하였다.

## 2.2 원문 데이터 및 초기 번역 데이터 검토

2023년 사업에서 번역해야 할 한국어 원문 데이터의 장르와 주제 분야를 검토하여 기존 번역 지침에서 수정하거나 보완해야 할 사항들을 논의하였다. 특히 이번 사업에서는 구어 데이터의 실제성을 높이기 위해 구어적 표현(생략, 준말, 반복, 도치 등)과 담화 표지들을 정제 과정에서도 유지하였고, 이러한 특성을 고려한 번역, 검수, 감수가 될 수 있도록 하였다.

또한 한국어 문어 데이터와 구어 데이터가 8개 언어로 번역되어 플랫폼에 올라온 초기에 시범적으로 검수하면서 최신 지침이 잘 적용되었는지 점검하고, 지침에 추가하거나 수정해야 할 사항이 있는지 확인하였다. 이러한 과정을 통해 필요한 경우 지침에 업데이트하여 번역 작업에 반영되도록 하였다.

## 2.3. 지침 수정 및 보완

지난 사업에 사용된 기존 지침과 관련 자료를 검토하고, 원문과 초기 번역 데이터를 검수한 결과들을 종합하여 검수팀장과 번역 검수원이 논의하여 수정·보완 사

항을 도출한 뒤 구체적인 지침 항목이나 용례를 수정·보완하였다. 지난 사업에서 이미 수차례 검토된 지침이었기에 수정 사항은 적고, 주로 새로운 한국어 원문 데이터의 특성을 고려하여 지침을 보완하는 방향으로 작업을 진행하였다. 한편 공동 연구원과 전임 연구원들은 지침 내 한국어 설명 부분이나 지침의 체계 등을 검토하고 검수팀장과 함께 논의하여 필요에 따라 수정·보완하였다. 이상의 과정을 통해 수정·보완된 지침은 즉각 플리토를 통해 번역사들에게 공유하여 번역 시 적용할 수 있도록 하였다. 이를 통해 번역 작업의 효율성과 일관성을 유지하면서 품질을 향상시키고자 하였다.

<p><b>제2항 대명사</b></p> <ol style="list-style-type: none"> <li>1. 한국어의 대명사는 우즈베크어의 문법적 특징에 따라 적절하게 번역한다.</li> <li>2. 주어나 목적어에 위치한 인칭 대명사나 지시 대명사가 생략된 경우, 우즈베크어의 특징에 따라 생략 또는 복원할 수 있다.</li> </ol> <p><b>제3항 수사와 단위 명사</b></p> <ol style="list-style-type: none"> <li>1. 숫자(기수/서수), 날짜(연/월/일/요일), 시간 등은 아라비아 숫자와 국제 통용 단위로 쓰는 것을 원칙으로 하되, 우즈베크어권에서 사용하는 숫자 표기를 준용한다.</li> </ol> <p>(한) 204,000 → (우) Ikki yuz-u to'rt ming (한) 564 → (우) Besh yuz-u oltmish to'rt (한) 2021년 10월 13일 → (우) O'n uchinchi oktabr ikki ming yigirma</p>	→	<p><b>제2항 대명사</b></p> <ol style="list-style-type: none"> <li>1. 한국어의 대명사는 우즈베크어의 문법적 특징에 따라 적절하게 번역한다.</li> <li>2. 주어나 목적어에 위치한 인칭 대명사나 지시 대명사가 생략된 경우, 우즈베크어의 특징에 따라 생략 또는 복원할 수 있다.</li> </ol> <p><b>제3항 수사와 단위 명사</b></p> <ol style="list-style-type: none"> <li>1. 숫자(기수/서수), 날짜(연/월/일/요일), 시간 등은 아라비아 숫자와 국제 통용 단위로 쓰는 것을 원칙으로 하되, 우즈베크어권에서 사용하는 숫자 표기를 준용한다.</li> </ol> <p>(한) 204,000 → (우) 204ming (한) 564 → (우) 564 (한) 2021년 10월 13일 → (우) 13-oktabr 2021-yl</p>
---	---	--

[그림 8] 번역 검수 세부 지침 수정 사항 예시(우즈베크어)

<p>5. 반복의 경우 인도네시아어에서도 반복 현상이 있을 때 유사하게 번역한다. 다만 반복한 번역의 결과가 어색해질 경우 반복된 일부를 생략할 수 있다.</p> <p>예) (한) 어제 내가 시장에 갔는데 사람들이 <u>진짜 진짜 많았어</u>. (인) Kemarin saya pergi ke pasar dan orang <u>benar-benar banyak</u>.</p>
--

[그림 9] 번역 검수 세부 지침 보완 사항 예시(인도네시아어)

## 2.4. 지침 최종본 완성 및 적용

검수팀장을 중심으로 검수원들의 다양한 의견을 종합하여 지침의 최종본을 완성하고 플리토와 공유한 후 이를 본격적으로 번역, 검수, 감수에 적용하였다. 검수와 감수 과정에서 새로운 쟁점이나 주의 사항이 발견되면 해당 사항을 검수원들이 논의하여 필요하다고 판단되었을 때 개별적으로 지침을 수정·보완하였고 다시 플리토

와 공유하였다. 이러한 과정을 통해 번역, 검수, 감수 과정의 효율성과 일관성을 유지하고자 하였다.

	00. 한국어-외국어 번역 공통 지침_230616.pdf
	01. 한국어-베트남어 번역 세부 지침_230620.pdf
	02. 한국어-인도네시아어 번역 세부 지침_230620.pdf
	03. 한국어-태국어 번역 세부 지침_230616.pdf
	04. 한국어-인도 힌디어 번역 세부 지침_230620.pdf
	05. 한국어-캄보디아어 번역 세부 지침_230620.pdf
	06. 한국어-필리핀 타갈로그어 번역 세부 지침_230620.pdf
	07. 한국어-러시아어 번역 세부 지침_230619.pdf
	08. 한국어-우즈베크어 번역 세부 지침_231018.pdf

[그림 10] 번역 검수 지침 최종본 목록

### 3. 번역

#### 3.1. 번역 작업자 선정

##### 1) 번역 업체 및 프리랜서 전문 번역사의 선정

번역 데이터의 품질 향상을 위해 본 사업에 대한 이해도가 높고 2021·2022년 사업에 참여하였던 현지 번역 업체 및 프리랜서 전문 번역사(이하 ‘작업자’) 중 결과물의 품질이 우수한 작업자를 선정하였다.

또한 모국어를 100% 활용할 수 있는 작업자를 발굴하였다. 고품질 번역을 제공할 수 있는 작업자를 발굴하고자 샘플 테스트를 진행하였으며, 번역 지침을 얼마나 잘 준수하여 번역하였는지를 평가하고 그 결과를 바탕으로 신규 작업자를 선정하였다.

[illegible]

### 3.2. 번역 인력 교육

[illegible]

### 3.3. 번역 절차

<표 16> 번역 및 교정 절차

단계	번역	교정
주체	• 도착어 기준 현지 번역 업체 및 프리랜서 전문 번역사	• 도착어 기준 현지 번역 업체 및 프리랜서 전문 검수자
방식	• 제공된 번역 지침에 따라, 원문-번역문 문장 쌍 생성	• 제공된 지침에 따라, 플리토 검수문 형식에 맞게 교정 수행
내용	• 표기 방식, 어휘 및 문법 채택 방식 등을 종합 고려하여 번역 수행	• 자연스러운 표현, 어색한 직역투 등을 수정
납품	• 실시간 번역 작업 후 플리토 아케이드 시스템 혹은 매주 엑셀 형식으로 제출	• 실시간 교정 작업 후 플리토 아케이드 시스템 혹은 매주 엑셀 형식으로 제출
수정·보완 의무	• 플리토 작업 관리자의 수정·보완 지시에 따라 의무 수행	

### 3.4. 번역 데이터 구축 환경

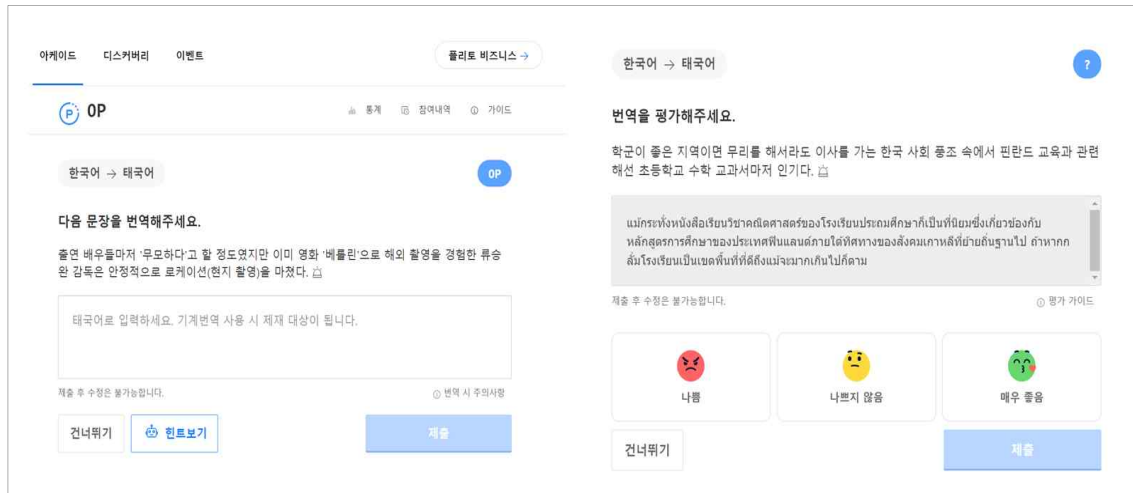
#### 1) 번역·교정

본 사업은 번역 품질 향상을 위하여, 현지의 번역 업체와 전문 번역사들이 현지 업체 사무실 또는 해당 과제 진행을 위한 작업 환경에서 플리토 아케이드 시스템 (혹은 엑셀 파일 형식<sup>1)</sup>)을 통해 번역 및 교정 작업을 수행하였다. 교정에 참여하는 작업자들의 평균적인 언어 수준은 해당 언어 국가의 학위 소지 또는 출생 및 거주 경험을 보유하고 있으며 해당 언어의 통번역 작업 경험이 다수 있는 경력자들로, 번역에 참여하는 작업자들에 비해 높은 수준을 필수 조건으로 하여 작업을 수행하도록 하였으며, 작업자들의 주요 업무는 아래와 같다.

<표 17> 번역 및 교정 작업자 주요 업무

<ul style="list-style-type: none"> <li>• 해당 언어별 번역 또는 교정(작업자별 개별 계정 부여)</li> <li>• 플리토 아케이드 시스템을 통해 문장 단위의 데이터를 실시간 제출 혹은 매주 할당된 엑셀 파일 제출</li> <li>• 완료 산출물에 대한 보완 지시 사항 대응</li> </ul>
--

1) 일부 국가의 경우 현지 인터넷 연결이 원활하지 않거나 아케이드 시스템을 활용하는 데 어려움이 있었는데, 이 경우 엑셀 파일 형식을 통해 작업을 진행하였다.



[그림 13] 플리토 아케이드 시스템의 번역·교정 환경

## 2) 1차 검수(5% 표본 검수)

현지 번역 업체와 전문 번역사들의 번역·교정 결과물의 품질 보증을 위해, 채용 대행업체를 통해 모국어 화자와 도착어(외국어) 구사가 가능한 한국인 번역사를 파견 형식으로 채용하여 작업 관리자로 활용하였다. 작업 관리자는 플리토 아케이드 시스템(혹은 엑셀 파일 형식)을 통해 실시간으로 구축되는 산출물을 검수하여 작업자들에 대한 작업량·정확도 등을 관리하였다. 작업 관리자의 주요 업무는 아래와 같다.

### <표 18> 1차 검수 작업자 주요 업무

- 일 단위 번역·검수 완료 산출물에 대한 1차 검수(5% 표본 검수)
- 플리토 아케이드 시스템(혹은 엑셀 파일)을 통해 작업자별 작업 달성도 추적·관리
- 완료 산출물에 대한 개선 필요 사항 전달 및 보완 지시



사를 작업 관리자로 채용하여 활용하였다. 한국인 번역사는 모국어 화자와 함께 작업을 하며 한국어 원문 및 한국어에 대한 이해도를 높여 보다 정확한 작업이 가능하도록 하였다. 작업 관리자는 구축 프로세스를 운영하기 위한 조직 체계를 마련하여 작업자들에 대한 작업량·정확도 등을 관리하며 작업 진행을 위한 교육 및 훈련을 실시하였다.

이외에 번역 지침 및 언어 특성 이해 등이 떨어지는 작업자에 대해서는 수정·보완을 지시하며, 3번 이상의 수정·보완 작업에도 품질이 향상되지 않는 경우 번역 및 교정 작업을 중단하였다.

번역문의 사소한 오류는 기계적 정제를 거쳐 기호 종류를 통일하고 불필요한 문자 등을 제거하였으며, 기계적 정제 내용은 다음과 같다.

<표 19> 번역 품질 향상을 위한 기계적 정제 내용

정제 내용
줄 바꿈, 탭, 공백이 여러 칸 있는 경우 공백 한 칸으로 변환
전각 기호는 반각 기호로 변환
특수 기호의 통일화
마침표가 없이 끝난 문장은 언어 특성에 맞게 추가

플리토 아케이드 시스템과 엑셀 형식으로 진행한 번역 파일에 대해 구조적인 검증을 거쳤으며, 검증을 통과하지 못한 문장은 작업자가 다시 번역 및 교정하도록 하였다. 구조적 검증의 기준과 불합격의 경우는 다음과 같다.

<표 20> 번역 품질 향상을 위한 구조적 검증 기준

검증 기준	불합격
중복	원문과 번역문, 서로 다른 번역문이 중복되는 경우
텍스트	원문 혹은 번역문이 누락된 경우, 번역문이 숫자 및 특수 문자로만 구성된 경우
비율	원문과 번역문의 비율이 극히 차이 나는 경우
언어	도착어 외의 언어가 들어간 경우
MT 유사도	번역문과 기계 번역기의 번역문의 유사도가 90% 이상인 경우 ※ 단문의 경우, 전문 검수자가 재검증



### 3) (추가 제안) 영어 데이터 품질 향상 활동

추가로 제안한 한국어-영어 병렬 말뭉치는 플리토에서 자체적으로 품질 향상 활동을 진행하였으며, 그 내용은 다음과 같다.

<표 21> 영어 데이터 품질 향상 활동 내용

단계	수행 내용	세부 사항
1	번역문에 대한 전수 교정(proofreading)	<ul style="list-style-type: none"> <li>- 번역문 전체 대상 문법 오류, 오역 등을 교정</li> <li>- 부적합 시 번역 업체 및 전문 번역사에게 재수정 요청</li> </ul>
2	MT 유사도 및 기계적 검증	<ul style="list-style-type: none"> <li>- 번역문과 번역기의 유사도 확인</li> <li>- 부적합 내용, 중복 내용 등 기계적 처리 가능 부분 처리</li> <li>- 부적합 시 번역 업체 및 전문 번역사에게 재수정 요청</li> </ul>
3	내부 언어 전문가 1차 5% 표본 검수	<ul style="list-style-type: none"> <li>- 전수 검수와 기계적 검증을 통과한 데이터에서 일부를 추출하여 도착어 기준 모국어 화자인 플리토 내부 언어 전문가가 표본 검수 진행</li> <li>- 부적합 시 번역 업체 및 전문 번역사에게 재수정 요청</li> </ul>
4	내부 언어 전문가 2차 전수 검수	<ul style="list-style-type: none"> <li>- 어색한 표현, 오타자, 문장 구조 등을 교정</li> <li>- 국내에서 가장 많이 사용되는 미국식 영어 표현을 위주로 수정·보완하여 번역문에 통일성을 부여</li> </ul>

## 4. 검수 및 감수

### 4.1. 번역 검수팀

#### 1) 구성

본 사업의 번역 검수 인력들은 모국어와 한국어가 능숙한 이중 언어 화자로서 한국어를 정확하게 이해하고 모국어로 번역할 수 있는 충분한 능력을 갖추었다. 외국어 번역 과정에서 모국어 화자 수준의 언어 숙달도가 아니면 판단하기 어려운 미세한 오류들이 발생할 수 있으나 본 사업의 번역 검수 인력들은 이중 언어 화자 수준으로 번역의 오류를 방지할 수 있었다. 구체적으로는 번역 검수팀은 국내 대학의 석·박사 대학원생, 해외 대학의 한국어 관련 전공 석·박사 대학원생 수준의

전문 인력으로 구성되었으며, 이들은 해당 언어 가능자 혹은 관련 지역 교육 경험 이 풍부한 한국어 교수, 석·박사 과정 이상의 인력으로 문법상의 오류뿐만 아니라 맥락상의 자연스러움까지도 검수를 진행하였다.

번역 검수팀은 팀 내 총괄적인 관리·감독과 원문의 내용을 세밀하게 분석하고 전달하는 공동 연구원, 검수 진행과 동시에 번역 검수원의 교육·관리를 담당하는 검수팀장, 전수 검수를 담당하는 번역 검수원, 정확한 한국어 원문 이해를 도와주는 보조 연구원으로 구성되었다.

## 2) 직책별 역할

번역 검수원은 문법, 어휘, 표기법 등의 단순한 번역 오류의 결과를 번역 검수 지침에 따라 수정하였다. 또한 한국어 원문이 포함하고 있는 구체적인 맥락과 상황을 고려하여 정확하게 번역하고 자연스러운 표현으로 다듬는 역할을 수행하였다.

검수팀장은 번역 검수원들의 검수 데이터를 20% 표본 검수(3차 검수)하여 검수원별 오류 유형과 오류율을 분석하고 보고하였다. 이후 번역 검수원과 샘플 점검 결과를 공유하여 검수 오류율이 높은 작업자를 대상으로 재교육을 실시하였고, 그럼에도 시정되지 않을 경우 해당 번역 검수원과의 계약을 해지하였다.

번역 검수원의 전수 검수와 검수팀장의 3차 검수가 동시에 진행되어 번역 검수 초반 데이터의 품질을 높일 수 있었고, 지침을 명확히 숙지하지 않은 신규 검수원의 검수 품질을 높일 수 있었다.

또한 번역 검수 과정상에서 검출된 오류를 분석하고 유형화함으로써 검수 과정의 효율성을 높이고 추후 오류 재발 방지 교육 및 수정 지침의 근거로 삼았다.

번역 검수 인력은 2021년 및 2022년 사업에서 번역 검수를 한 경험이 있는 검수원으로 우선 선발하여 진행하였으며, 언어별 검수 진행 상황에 따라 번역 및 번역 검수 테스트를 거쳐 번역 검수원을 추가 선발하였다.

<표 22> 언어별 번역 검수원

언어	베트남어	인도네시아어	태국어	인도 힌디어	캄보디아 크메르어	필리핀 타갈로그어	러시아어	우즈베크어	계
기존 인원	10	22	13	13	19	14	17	12	120
추가 인원	-	1	1	3	1	5	-	2	13
총 참여 인원 <sup>2)</sup>	10	23	14	16	20	19	17	14	133

2) ‘기존 인원’은 2021년 사업부터 참여한 번역 검수원이며, ‘총 참여 인원’은 기존 인원을 포

## 4.2. 감수자

(사)국제한국어교육학회는 인적 네트워크를 통해 국외 여러 국가의 언어 전문가 풀을 구축하고 긴밀한 협력 관계를 유지하고 있다. 이를 활용하여 도착어를 모국어로 사용하고 한국어 실력이 뛰어난 해외 대학 교수 수준의 언어 전문가로 감수 전문팀을 구성하였다.

감수자는 전체 구축 수량의 10%를 무작위 샘플 추출하여 감수를 진행하였다. 주요 검수 항목을 중심으로 감수를 실시하였으며, 이를 통해 번역의 완성도를 제고하였다.

감수자로는 베트남의 호찌민시국립인문사회과학대학교, 베트남외교아카데미, 인도네시아의 인도네시아국립대학교, 가자마다대학교, 태국의 부라파대학교, 송클라대학교, 인도의 자와할랄네루대학교, 캄보디아의 왕립프놈펜대학교, 필리핀의 필리핀국립대학교, 러시아의 모스크바국립외국어대학교, 우크라이나의 우신스키국립사범대학교, 우즈베키스탄의 타슈켄트국립동방대학교 등의 교수가 참여하였다.

<표 23> 언어별 감수자

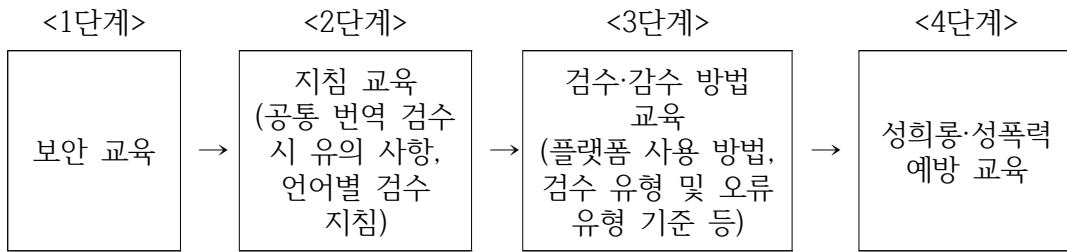
언어	베트남어	인도네시아어	태국어	인도 힌디어	캄보디아 크메르어	필리핀 타갈로그어	러시아어	우즈베크어	계
기존 인원	2	2	2	2	1	1	2	2	14
추가 인원	-	-	-	-	1	1	-	-	2
총 참여 인원	2	2	2	2	2	2	2	2	16

## 4.3. 번역 검수팀 및 감수자 교육

검수와 감수를 진행하는 번역 검수원, 검수팀장, 감수자를 대상으로 다음과 같은 4단계 교육을 실시하였다.

---

함하여 사업 기간 내 검수에 참여한 번역 검수원 전체를 의미한다. 이는 감수자에도 해당된다.



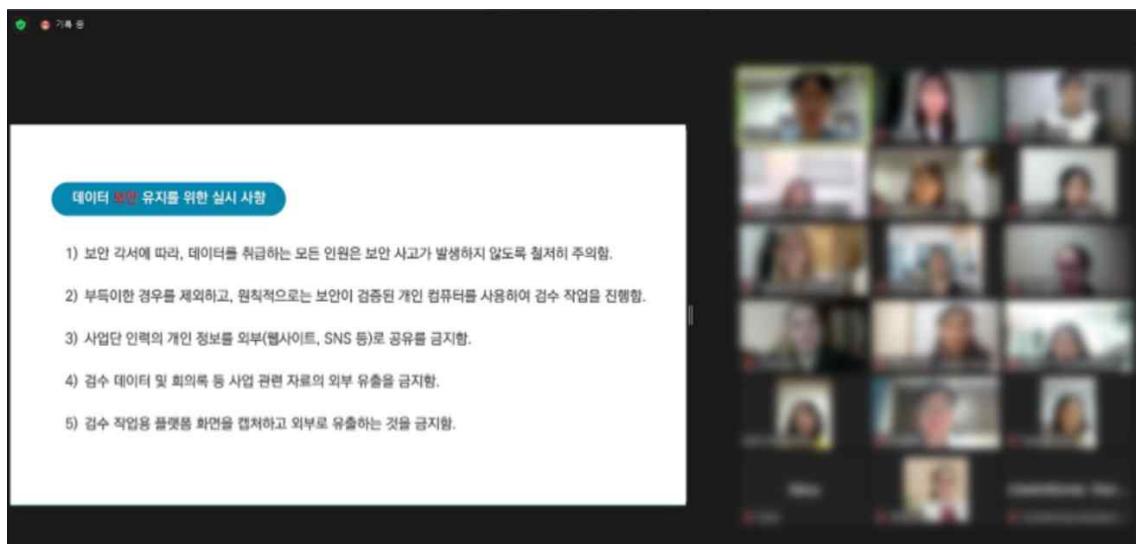
[그림 15] 검수·감수 교육 절차

## 1) 보안 교육

먼저 검수 및 감수 인력 전원을 대상으로 보안 교육을 실시하여 데이터 보안에 각별한 주의를 기울이도록 하여 보안 사고를 미연에 방지하였다. 보안 교육 내용과 예시는 다음과 같다.

<표 24> 검수·감수 인력 보안 교육 내용

- 보안 각서에 따라, 사업 과정에서 산출되는 모든 데이터를 복제하거나 외부에 배포하는 것을 금지한다.
- 원칙적으로는 보안이 검증된 개인 컴퓨터를 사용하여 검수 및 감수 작업을 실시한다.
- 검수 및 감수용 플랫폼의 개인 계정 정보의 보안에 주의하며 타인과 플랫폼 계정을 공유하는 것을 금지한다.
- 검수 및 감수용 플랫폼의 화면을 외부인에게 공유하거나 SNS 등에 게시하는 것을 금지한다.
- 검수 및 감수 데이터, 회의록 등 사업 관련 자료의 외부 유출을 금지한다.





[그림 16] 검수·감수 보안 교육 예시

## 2) 지침 교육

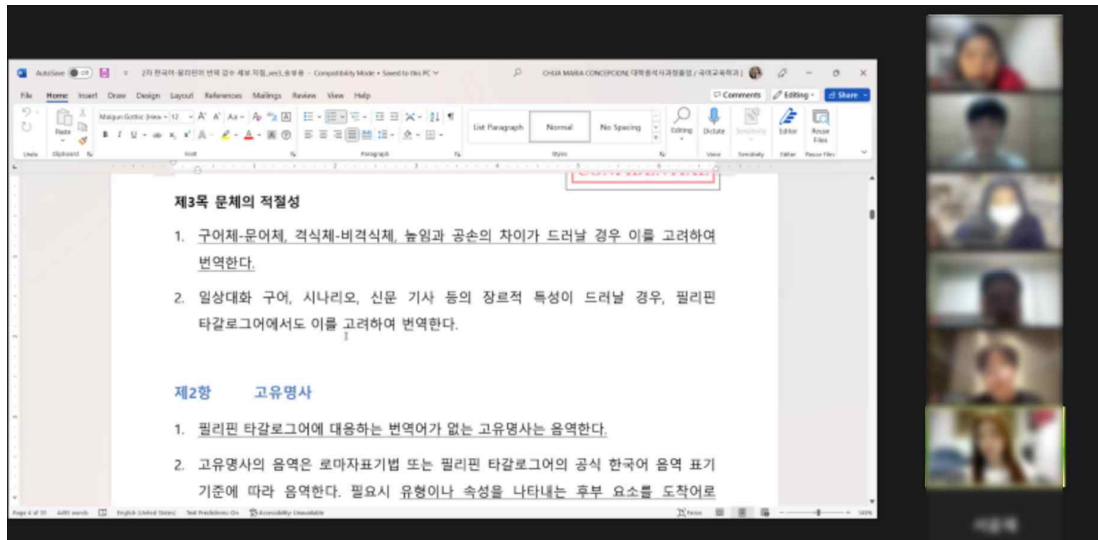
보안 교육을 실시한 후에는 사업 목적과 목표를 성공적으로 달성하기 위하여 지침 교육을 실시하고 방법론, 절차적 지식, 기술을 공유함으로써 사업 수행 인력의 전문성을 함양하였으며, 사전 설계된 수행 지침에 대한 교육을 통하여 효율적이고 체계적인 절차를 준수함으로써 일관되고 오류가 적은, 신뢰할 만한 말뭉치를 구축하였다. 또한 구체적인 단계별 검수 기준 및 검수 방법 교육을 통해 정확하고 자연스러운 언어 자료를 마련하고자 하였다.

<표 25> 지침 공통 교육 내용

항목	교육 내용
사업 소개	<ul style="list-style-type: none"> <li>- 사업 개요</li> <li>- 데이터 구축</li> <li>- 사업의 특징점</li> </ul> 
한국어 원문의 특징 및 번역 검수 시 유의 사항	<ul style="list-style-type: none"> <li>- 한국어 원문의 특징</li> <li>- 문어 번역 검수 시 유의 사항</li> <li>- 구어 번역 검수 시 유의 사항</li> </ul> 

<표 26> 언어별 지침 내용

항목	지침 내용
내용 검수	<ul style="list-style-type: none"> <li>- 내용 일치 여부 검수</li> <li>- 오역 및 누락 내용 검수</li> <li>- 문장 구조 및 문법·관용구·어휘 일치 검수</li> <li>- 언어별 번역 특이 사항</li> </ul>
표기 검수	<ul style="list-style-type: none"> <li>- 문장 부호 및 특수 문자 일치 검수</li> <li>- 오타자 및 띄어쓰기 오류 검수</li> <li>- 언어별 번역 특이 사항</li> </ul>



[그림 17] 검수 지침 교육 예시

### 3) 검수·감수 방법 교육

#### (1) 검수 플랫폼

모든 검수 인력을 대상으로 개인 계정을 지급하여 본 사업에 최적화된 플랫폼에서 검수를 실시하였다.

SID	OriginText	Translation	QC	QC Reason	QC Translation	Submit
1024 2803	가족에게 헌신하는 성공한 건축가 에반(키아누 리브스)은 휴일을 맞아 여행을 떠난 가족들을 뒤로하고 홀로 집에 남는다. 🏠	Arsitek Evan (Keanu Reeves) yang sukses mendedikasikan diri untuk keluarganya, pada hari libur keluarganya pergi wisata dan ia ditinggal sendirian di rumah.	👍 👎 👉	<input type="checkbox"/> 어휘 <input type="checkbox"/> 문법 <input type="checkbox"/> 내용 <input type="checkbox"/> 표기법 <input type="checkbox"/> 기타	Arsitek Evan (Keanu Reeves) yang sukses mendedikasikan diri untuk keluarganya, pada hari libur keluarganya pergi wisata dan ia ditinggal sendirian di rumah.	<input type="button" value="Submit"/> <input type="button" value="Pending"/>

[그림 18] 검수 플랫폼 작업 환경 예시

검수 플랫폼은 문장 번호(SID), 출발어 원문(Origin Text), 도착어 번역문(Translation), 검수(QC), 오류 유형(QC Reason), 검수문(QC Translation), 제출(Submit), 보류(Pending)로 구성되어 있다.

번역 검수원들은 플랫폼에서 출발어 원문(Origin Text)과 도착어 번역문(Translation)을 보고 검수(QC) 유형을 판단한 후에 그에 따라 검수 작업을 수행하였다.

검수(QC) 유형은 좋음, 부분 수정, 재번역으로 구성되며, 별도의 수정이 필요 없는 경우 좋음, 사소한 오류로 수정이 가능한 경우 부분 수정, 오류가 많아 재번역

또한 보류(Pending) 기능은 원문 또는 검수에 대한 질의가 있을 경우, 잠시 보류할 수 있는 것으로 이를 통해 검수에 편의를 높였다.

## 검수 유형

## 조음



- 부분



- 수정

## 재번역



- 번역문에 오류가 있을 시(부분 수정, 재번역) 선택하는 오류 유형(QC Reason)의 기준은 어휘, 문법, 내용, 표기법, 기타로 구분하였다.

오류 유형

## 어휘

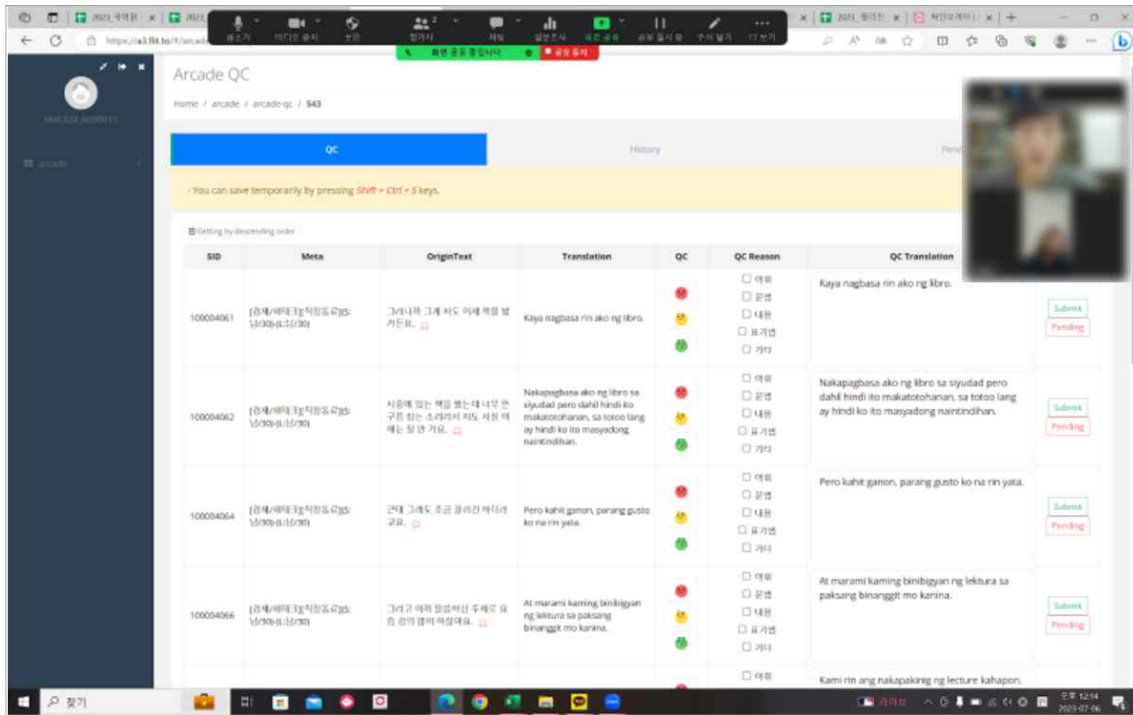
- 문법

- 내용

- ## 표기법

- 기타

- 문법적으로 오류가 없으나 지나친 직역으로 인해 어색한 표현인 경우



[그림 19] 검수 플랫폼 사용 교육 예시

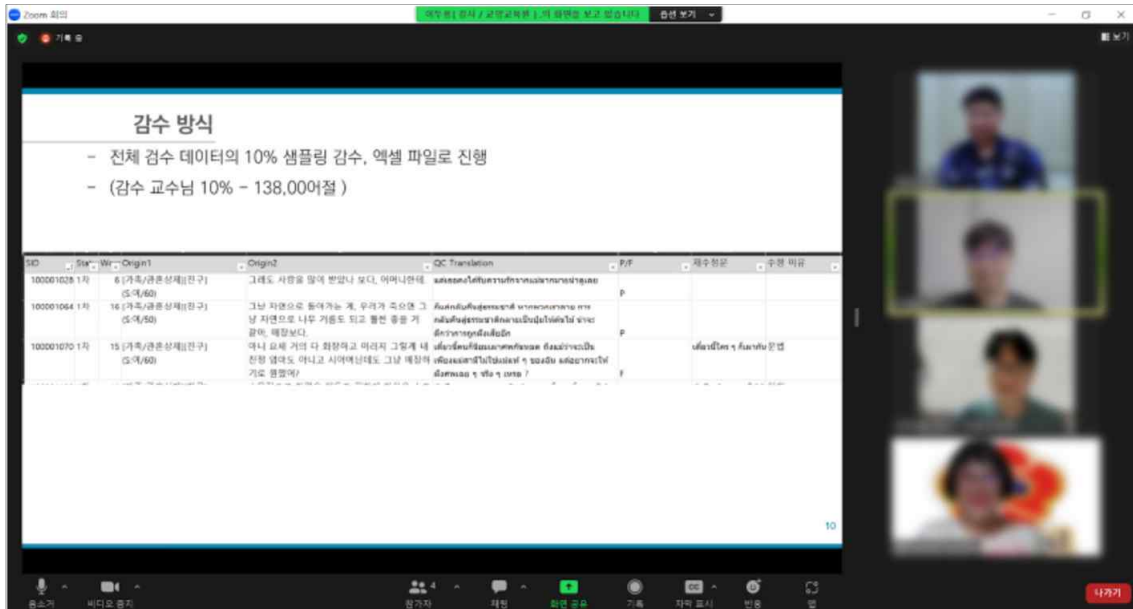
## (2) 감수 플랫폼

감수는 엑셀 파일로 실시하였다. 원문(Origin Text)과 검수문(QC user result)을 보고 감수 결과(Assess)를 결정하였는데, 검수문에 오류가 없으면 P, 오류가 있으면 F를 선택하며 감수 결과가 F일 경우 검수문을 감수자가 직접 수정하고 의견(comment)란에 수정 이유를 작성하였다.

ID	SID	META	ORIGIN	이탈 수	Translation	QC user result	Result	QC Date	Assess	comment
IAKLE23_koi000003	100148180		[소평]기타가죽(S:5/30)-L:5/30)반죽이 기적과 불 건 증거 적었어.	6	Aku usak ngiti menakaninya karena mada.	Aku usak ngiti menakaninya karena mada.	Perfect	2023/07/10 23:30:47 +09	P	
IAKLE23_koi000003	100148182		[소평]기타가죽(S:5/30)-L:5/30)다 아라서 불 나가 업알하야지 잘거	7	Aku hanya rajin memakannya kalau lebih	Aku hanya rajin memakannya kalau lebih	Perfect	2023/07/11 00:16:16 +09	P	
IAKLE23_koi000003	100148207		[소평]기타가죽(S:5/30)-L:5/30)무슨 얘기, 하려 할지?	4	Aku mau mengatakan apa, ya?	Aku mau mengatakan apa, ya?	Perfect	2023/07/11 00:16:16 +09	P	
IAKLE23_koi000003	100148219		[소평]기타가죽(S:5/30)-L:5/30)이런 거 5만 원 정도 막단 거 같은 데.	8	Sepertinya yang begini sekitar 50.000 won.	Sepertinya yang begini sekitar 50.000 won.	Perfect	2023/07/10 23:35:03 +09	P	
IAKLE23_koi000003	100148031		[소평]기타가죽(S:5/30)-L:5/30)아름, 할라니 집 가서 뭐 하는 거야, 그	8	Hem, jadi, pergi ke rumah nenek dan	Hem, jadi, pergi ke rumah nenek dan	So So	2023/07/11 23:27:49 +09	P	
IAKLE23_koi000003	100148032		[소평]기타가죽(S:5/30)-L:5/30)아, 할라니 집 가서 일하곤 종돈 팔	11	Oh, jadi kamu mendapatkan uang dengan	Oh, jadi kamu mendapatkan uang dengan	So So	2023/07/11 23:26:03 +09	P	
IAKLE23_koi000003	100148034		[소평]기타가죽(S:5/30)-L:5/30)아, 그럴 어떤 일해서 종돈 팔아?	6	bekerja apa?	bekerja apa?	So So	2023/07/11 23:24:24 +09	P	
IAKLE23_koi000003	100148039		[소평]기타가죽(S:5/30)-L:5/30)요즘은 다들씩 관심이 있어 가지고	8	Aku sedang menabung karena tertarik dengan	Aku sedang menabung karena tertarik dengan	So So	2023/07/11 22:51:36 +09	F	akhir-akhir ini aku tertarik dengan Dior, jadi aku mengumpulkan uang.
IAKLE23_koi000003	100148041		[소평]기타가죽(S:5/30)-L:5/30)아, 다들 어떤 거, 옷?	5	Ya, Dior yang bagaimana, baju?	Hem, Dior yang mana, pakaian?	So So	2023/07/11 22:50:31 +09	P	
IAKLE23_koi000003	100148043		[소평]기타가죽(S:5/30)-L:5/30)신발 관심 있거든, 지금.	4	Sekarang, aku tertarik pada sepatu, lo.	Sekarang, aku tertarik pada sepatu, lo.	So So	2023/07/11 22:49:55 +09	P	
IAKLE23_koi000003	100148053		[소평]기타가죽(S:5/30)-L:5/30)연구들이 신발이 좋 버퍼다 보니까,	11	memiliki semacam barang-barang bemerek.	memiliki semacam barang-barang bemerek.	So So	2023/07/11 22:55:32 +09	P	
IAKLE23_koi000003	100148143		[소평]기타가죽(S:5/30)-L:5/30)근데 아들 막는 굴이 안올아도	10	canany?	canany?	So So	2023/07/11 22:47:46 +09	F	Namun, kamu pasti cantik waktu masih kecil bahkan jika kamu tidak memiliki barang bemerek.

[그림 20] 감수 작업 환경 예시

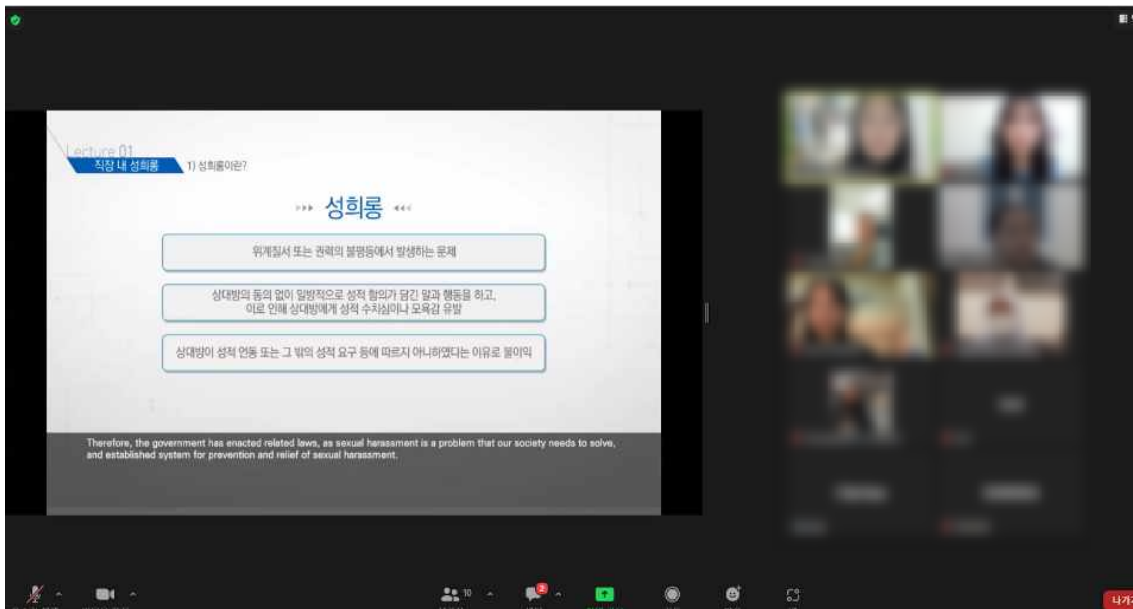




[그림 21] 감수 방법 교육 예시

#### 4) 성희롱·성폭력 예방 교육

사업에 참여하는 모든 인력을 대상으로 성희롱 성폭력 예방 교육을 실시하였다. 국내 거주자는 한국양성평등교육진흥원의 성희롱·성폭력 예방 통합 교육 수료 후 교육 이수증을 발급받았으며, 해외 거주자는 한국양성평등교육진흥원 교육 이수가 불가하여 온라인 회의 플랫폼에서 성희롱 성폭력 예방 교육 영상을 같이 시청하였다. 해외 거주자는 교육 후에 이수 확인서를 발급함으로써 교육 이수 현황을 관리하였다.



[그림 22] 해외 거주자 성희롱·성폭력 예방 교육 예시

## 교 육 이 수 증

---

소속기관	한국어 외국어 병렬 말뭉치 구축 사업단		
교육기간	2023.05.12~2023.08.02		
교육시간	2시간		
과 정 명	공감 더하기, 폭력예방교육 (공공기관) [성희롱·성폭력 예방]		
과정내역	과정	과목명	시간
	성희롱·성폭력예방통합(1/2)	공감 더하기, 성희롱·성폭력 예방교육 (1) (공공기관)	1시간
	성희롱·성폭력예방통합(2/2)	공감 더하기, 성희롱·성폭력 예방교육 (2) (공공기관)	1시간

2023.08.02

위와 같이 사이버교육 과정을 이수하였음을 확인합니다.

한국양성평등교육진흥원

[그림 23] 한국양성평등교육진흥원 교육 이수증 예시

팀	역할	성함	성희롱·성폭력 예방 교육 이수 확인서
인도네시아어 검수팀	번역 검수원		○
인도네시아어 검수팀	번역 검수원		○
태국어 검수팀	검수 관리팀장		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
태국어 검수팀	번역 검수원		○
인도 힌디어 검수팀	검수 관리팀장		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○
인도 힌디어 검수팀	번역 검수원		○

[그림 24] 성희롱·성폭력 예방 교육 이수 현황 관리 대장

#### 4.4. 검수 품질 향상 활동

##### 1) 3차 검수

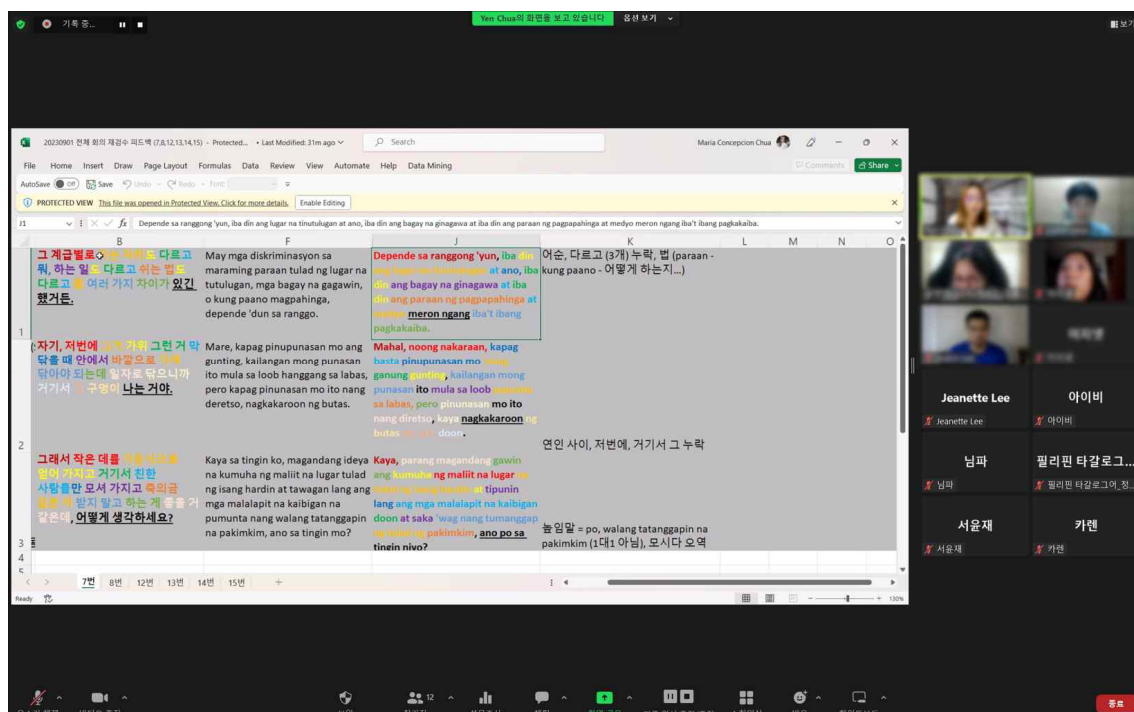
본 사업에서는 번역문을 3차례에 걸쳐 검수하였으며, 이 중 3차 검수 단계에서는 전체 검수 데이터의 20%를 검수팀장이 표본 검수하였다.

SID	Origin	QC Translation	P/F	수정문	재검수 의견
20026201	(여)무심기도 하겠지만 너무 황홀할 거 같아.	Akan menakutkan juga, tetapi sepertinya akan sangat menakjubkan.	P		
20027217	(여)세트로 사면 더 저렴하겠지?	Akan lebih murah jika beli set, kan?	F		
	(남)5지 선다형과 주관식 문제로 나누어서 출제하는 것이 좋	Akan lebih baik membagi pertanyaan dengan 5			
20032016	겠습니다.	pertanyaan pilihan ganda dan pertanyaan subjektif.	F		
20032118	(여)다시 한번 찾아 보겠습니다.	Akan saya coba cari lagi.	P		
		Akan lebih baik jika ada tempat yang memberikan kupon			
20036713	(여)알인 쿠폰 주는 곳이면 좋는데 그냥 아무 데서나 시켜.	diskon, tetapi pesanlah di mana saja.	P		
20099618	(여)이것저것 하면 좀 많이 들어요.	Akan lebih mahal jika melakukan ini dan itu.	P		
		Saya akan periksa apakah ada satu ukuran yang lebih			
20102511	(여)제가 한 사이즈 큰 거 재고 있는지 봐 드릴게요.	besar.	F		
		Akhir-akhir ini mungkin karena menua, aku tidak seperti			
20004903	(여)요즘 나이 먹어서 그런가, 예전 같지 않네.	dahulu lagi.	P		
		Aku hanya menunggu hari ini karena akhir-akhir ini tidak			
20012004	(여)요즘 삶에 낙이 없어서 이날만 기다렸어.	ada kebahagiaan dalam hidup.	P		
	(남)최근에는 확실히 애견 동반해서 여행오는 사람들도 많더	Akhir-akhir ini pasti banyak orang yang berpergian dengan			
20014719	리코.	anjingnya.	P		
		Akhir-akhir ini yang menjadi tren sepakbola tentu saja Son			
20023912	(남)요즘 축구는 손흥민이 대세이지.	Heung Min.	F		
		Akhir-akhir ini seperti akting itu wajib untuk ido jugall			
20027316	(여)요즘 아이돌도 연기는 필수인가 봐!	Akhir-akhir ini gelombang dinginya terasa semakin	F		
		panjang.			
20034907	(남)요즘 들어 한파가 더 길어졌 느낌이야.	Akhir-akhir ini pilates sangat populer, jadi silakan	P		
		mencobanya.			
20037813	(여)요즘 필라테스가 유행인데 한번 해보세요.	Akhir-akhir ini pose model sedang populer.	F		
20039504	(여)요즘 모델 포즈 유행이잖아.		P		

[그림 25] 3차 검수 예시

또한 검수팀장과 번역 검수원 간 3차 검수 결과에 대한 의견을 자유롭게 공유함으로써 검수 능력의 상호 증진을 도모하였다.

3차 검수 결과를 바탕으로 오류율이 높은 번역 검수원을 대상으로 재교육을 실시하였으며, 번역 검수원의 오류 원인에 따라 번역 검수 지침 재교육, 검수 샘플 평가 및 피드백 등을 실시하였다.



[그림 26] 번역 검수원 대상 검수 재교육 예시

## 2) 검수문 표기 오류 추가 점검

불필요한 기호 삽입, 번역문 내 한국어 표기 등의 오류를 추출하여 해당 문장을 검수한 인력을 대상으로 확인을 요청하고 수정하였다. 베트남어, 인도네시아어 등과 같이 원어를 살려 로마자로 표기해야 하는 경우 번역문에 표기 오류가 나타나는 경우가 있는데, 이 경우 전임 연구원과 보조 연구원이 정확한 표기를 확인하고, 표기 오류로 보이는 경우 수정하였다.

QC Translation	재검수 의견	수정 후 문장
Không phải chứ, trước tiên là bạn nữ trông hơi kích động nên cực kỳ buồn cười mà tớ thì là đứa thường hay xem <b>youtube</b> như thế này chỉ để giết thời gian mà.	YouTube	Không phải chứ, trước tiên là bạn nữ trông hơi kích động nên cực kỳ buồn cười mà tớ thì là đứa thường hay xem YouTube như thế này chỉ để giết thời gian mà.
Ừ, nhưng với mình cái việc mà mọi người nói quá trên <b>Youtube</b> thì không hợp với mình lắm.	YouTube	Ừ, nhưng với mình cái việc mà mọi người nói quá trên YouTube thì không hợp với mình lắm.
Nhưng mà việc biên tập đang ngày càng trở nên theo hướng gây kích thích hơn, cũng đâu còn cách nào khác vì nếu các chủ đề <b>Youtube</b> không gây tò mò như thế mà không chuyển màn hình thì khó mà tập trung được.	YouTube	Nhưng mà việc biên tập đang ngày càng trở nên theo hướng gây kích thích hơn, cũng đâu còn cách nào khác vì nếu các chủ đề YouTube không gây tò mò như thế mà không chuyển màn hình thì khó mà tập trung được.
Nhưng mà vì cún con và mèo con đã có nhiều rồi nên thông qua <b>Youtube</b> anh muốn xem động vật nào mà nó đặc biệt hơn một chút.	YouTube	Nhưng mà vì cún con và mèo con đã có nhiều rồi nên thông qua YouTube anh muốn xem động vật nào mà nó đặc biệt hơn một chút.
Ừm, có rất nhiều loài bò sát, nhưng thực tế các <b>Youtuber</b> không chỉ làm về loài bò sát đâu, mà có nhiều trường hợp họ vừa tự vận hành Pet Shop, vừa quay video cùng nên có nhiều loại lắm.	YouTube	Ừm, có rất nhiều loài bò sát, nhưng thực tế các YouTuber không chỉ làm về loài bò sát đâu, mà có nhiều trường hợp họ vừa tự vận hành Pet Shop, vừa quay video cùng nên có nhiều loại lắm.

[그림 27] 로마자 표기 및 문장 기호 오류 확인 예시

## 3) 한국어 원문 이해 지원

복잡한 상황 맥락이나 전문 용어, 신조어 등으로 인해 번역 검수원이 원문을 이해하는 데에 어려움을 겪을 수 있다. 번역 검수원이 원문에 대해 질문할 경우 팀 내 한국인 공동 연구원 및 보조 연구원, 전임 연구원이 원문의 상황 맥락과 의미를 파악하여 설명하였다. 이를 다른 언어 검수팀과 내용을 공유하기 위해 구글 스프레드시트에 질문과 답변을 정리하였다.

질문	의견1
김포 후쿠오카 노선은 못 봤고 보통 인천 후쿠오카가 많아. -> '후쿠오카'는 장소명인가요? 표기는?	일본 지역명 फुकुओका
왜냐하면 양가 다 만이었거든요. -> 양가 양가죽인가요? 아님 양 쪽 신부 신랑?	양쪽 신부신랑
딱 팔국수예다가 '실비감차'를 얻어서 먹으면 얼마나 맛있게. -> 얼마나 맛있게?	얼마나 맛있게= 얼마나 맛있는지 몰라. 매우 맛있어 라는 의미
막 양치 교육 같은 걸 시키려고 이 한 번 만지고 간식 주고 이 한 번 만지고 간식 주고 이런 식으로 시키거든. -> 어색한 문장?	막 양치 교육 같은 걸 시키려고 이 한 번 만지고 간식 주고, 이 한 번 만지고 간식 주고, 이런 식으로 시키거든. -> 반라동을 양치에 관한 문장으로 해당 심료가 있다고 생각하고 검수해야 할 듯
너네 '콜라' 피 검사 안 했어? -> 여기서 '콜라'가 무슨 뜻이에요?	반라동을 이름으로 고유명사로 심료 처리가 되어 있는 것입니다. 음역해 주시면 됩니다.
खरीद (buy) nuqta lagana h?	hअन / हँ
화요일, 목요일이었을 좋겠다. 학교 안 가게. -> 도치인가요?	네. '학교 안 가게 화요일, 목요일이었을 좋겠다.'의 도치 문장입니다.
रेस्टोरी= रेस्तोरी ? kaun sa saahi h?	रेस्टोरेट
जनबु =जोबु= 전주 kaun sa sahi hoga?	जनबु दीर्घ वाला कोई नहीं है
내가 좋아했던 시그널이랑 독전 둘 다 주인공으로 나오셨는데 나는 조건중 배우처럼 되고 싶은데 그게 잘 안 되더라고. -> 복문인가요?	복문은 아닙니다. 앞의 -는대를 전제를 나타내고 뒤의 -는대는 대립으로 번역이 되어야 합니다.
그것은 선생님의 통격지일 수도 있습니다. -> 통격지의 다른 의미가 있나요?	앞이 가려져 제대로 보지 못한다는 비유적인 의미예요. 필하는 좋게 보인다는 의미. 보통 이성애 대해 말할 때 그 의미로 사용됩니다.

[그림 28] 원문 이해 지원 예시

#### 4) 번역 품질 개선을 위한 의견 취합 및 전달

번역문에서 자주 발견되는 오류, 원문 의미 오역 등의 예시를 정리하여 플리토에 의견을 전달하여 번역이 개선될 수 있도록 하였다. 이를 향후 번역 작업에 반영되도록 하여 번역의 정확도와 검수의 효율성을 높이고자 하였다.

순번	SID	한국어 원문	러시아어 번역문	의견
1	100004160	[경제/재테크][직장동료](S:남/30)-(L:남/30)가 급적 중 젊은 분들 거는 잘 안 읽는 경향이 있어요.	По возможности я не читаю то, что для предназначено для молодежи.	젊은 작가들이 쓴 책이 아니라 젊은 사람들을 위한 책을 안 읽는 것으로 번역함.
2	100004162	[경제/재테크][직장동료](S:남/30)-(L:남/30)가 약간 인제 연루이 있으신 그런 성공하신 분들을 인제 책을 보고서 참고를 많이 하시는 거예요?	А Вы часто используете книги успешных людей чуть старшего возраста в качестве примера?	이 문장에서의 '참고한다'는 말은 청자가 본인 인생에 있어 누구누구의 책을 참고한다는 것으로 이해되는데 번역문은 누구누구의 책을 예시로 활용한다는 뜻으로 처리. 또한 '약간 연루 있으신'이라는 표현은 (젊은 사람들보다는) 나이 있으신 분들을 이야기하는 것인데 '살짝 나이 더 많은 사람들'로 번역되어 있음.
3	100004165	[경제/재테크][직장동료](S:남/30)-(L:남/30)실적이 없고 사기 치는 경우를 워낙 많이 봐서 일단 검증이 된 사람 위주로 보려고 해요.	Я слышал много историй про тех кто ничего не смог получить или попались на обман, поэтому для начала просто хочу понаблюдать за проверенными людьми.	책 쓴 이들이 실제로 실적도 없으면서 책을 쓴 거를 사기친다고 표현하는 것 같음. 그래서 어느 정도 검증된 사람들의 책 위주로 본다고 이야기하는데, 번역문은 실적 없는 사람들에 대한 이야기와 사기 당한 사람들에 대한 이야기를 많이 들어서 검증된 사람들을 우선 관찰하겠다는 뜻으로 번역함.
4	100004177	[경제/재테크][직장동료](S:남/30)-(L:남/30)아예 실체가 없는 줄 알고.	Думал, что имеет особого смысла.	오역일뿐만 아니라 문장 자체도 비문이라서 모국어 화자 아닌 것으로 보임.
5	100005260	[먹거리][모임-동아리지인](S:여/60)-(L:여/50)오하러 옛날 서울 사람들은 싱겁게 먹거든.	А раньше сеульцы наоборот ели пресную еду.	나이가 많은 서울 사람들이 싱겁게 먹는다는 의미가 아니라 예전에는 서울 사람들이 싱겁게 먹었다고 오역함.

[그림 29] 번역 품질 개선을 위한 의견 작성 및 전달 예시

#### 5) 번역 오류 유형 정리 및 공유

검수 과정에서 나타난 오류 유형을 범주별로 정리하고, 원인을 분석하여 번역 품질을 높이고자 하였다. 검수 과정에서 나타난 번역문의 오류 유형은 다음과 같다<sup>3)</sup>.

3) 검수 오류 유형 정리 및 공유 과정은 사업 기간 동안 수시로 진행되었으며, 여기에 제시된 표는 전수 검수 종료 후 모든 오류 유형을 정리한 것이다.



<표 29> 번역문 오류 유형

언어	어휘	문법	내용	표기법	기타
베트남어	50.6%	5.9%	9.1%	8.4%	26.1%
인도네시아어	36.5%	10.6%	7.4%	16.3%	29.3%
태국어	34.0%	6.2%	10.0%	32.5%	17.4%
인도 힌디어	27.7%	23.7%	36.5%	4.5%	7.6%
캄보디아 크메르어	33.0%	11.8%	21.7%	18.5%	15.1%
필리핀 타갈로그어	42.8%	14.0%	32.7%	8.0%	2.5%
러시아어	31.5%	9.6%	8.6%	12.6%	37.6%
우즈베크어	40.4%	21.2%	16.6%	19.9%	1.9%
<b>평균</b>	<b>37.1%</b>	<b>12.9%</b>	<b>17.8%</b>	<b>15.1%</b>	<b>17.2%</b>

전반적으로 어휘 오류의 비율이 가장 높고 내용 오류, 기타 오류, 문법 오류, 표기법 오류의 순으로 높았다.

언어마다 오류 유형의 비율이 다양한 양상을 보였지만 어휘 오류의 비율이 가장 높은 것이 모든 언어의 공통적인 특징이었다. 어휘 오류는 인명, 지명, 기관명 등의 고유 명사나 전문 용어, 구어 담화 표지의 사용, 특정 어휘의 사용역이 부적절한 경우들이 많았다.

문법 오류의 경우, 어순과 시제, 능·피동, 성·수 일치, 개별 문법 항목(조사, 접속사 등)에서 오류가 많이 발생하였다.

내용 오류는 문장 구조가 복잡하거나 원문에서 한국의 경제, 사회, 문화적 배경의 지식을 많이 요하는 경우에 원문의 의미가 훼손되거나 일부 내용이 누락되는 경우가 많았다.

표기법 오류는 주로 고유 명사의 음역 표기나 쉼표 및 따옴표 등의 문장 부호에서 많이 발생하였으며, 단순 철자 오류도 다수 있었다.

기타 오류는 베트남어, 인도네시아어, 러시아어에서 비율이 높은 편으로 나타났는데 이는 모두 지나친 직역으로 인해 나타난 것이었다.

언어별 번역 오류 양상과 주요 원인은 다음과 같다.

<표 30> 언어별 번역 오류 양상 및 원인

언어	오류 양상 및 주요 원인
베트남어	<ul style="list-style-type: none"> <li>• 어휘 오류가 50.6%로 과반을 차지했다. 인명, 작품명, 기관명, 직책명 등의 표기 오류가 많아 이를 수정하였다. 또한 어휘의 오역이나 누락이 다수 발생했으며 원문에 없는 어휘를 추가하여 번역한 경우도 있었다.</li> <li>• 문법 오류의 경우에는 어순이 맞지 않는 경우가 많았으며 내용 오류의 경우는 원문의 의미를 잘못 해석하여 원문과 다른 내용으로 번역한 경우가 있었다.</li> <li>• 이 외에 맞춤법 오류, 의역이 필요함에도 직역을 한 경우(예: '빨간 날') 등이 있었다.</li> </ul>
인도네시아어	<ul style="list-style-type: none"> <li>• 가장 높게 나타난 오류는 어휘 오류로 영어식 표현을 인도네시아어 표현으로 변경한 경우와 문어와 구어에 적합한 어휘로 수정한 경우가 많았다.</li> <li>• 기타를 제외하고 비율이 높은 표기법 오류는 쉼표, 따옴표 등의 문장 부호를 수정한 것이 다수를 차지하였다. 문장 부호의 사용은 인도네시아어 대사전(Kamus Besar Bahasa Indonesia, KBBI)과 제5차 완성 철자법(EYD V)의 기준에 따라 수정하였다.</li> <li>• 문법 오류는 시제 및 피사동의 부적합한 사용이 많았으며, 내용 오류는 한국어 원문의 일부가 누락된 경우와 한국어 원문의 일부 단어를 잘못 번역한 경우 등이 있었다.</li> <li>• 이 외에 기타 오류는 지나친 직역으로 인해 내용 이해는 가능하나 인도네시아어 문장이 어색한 경우가 다수 나타났다.</li> </ul>
태국어	<ul style="list-style-type: none"> <li>• 가장 많은 비중을 차지하는 어휘에 대한 오류는 다양한 원인이 있다. 문맥에 따라 다른 대응으로 번역해야 할 단어를 부적절한 단어나 상위어로 번역한 경우, 의미역으로 번역하는 것이 더 적절한데 음역한 경우, 기관명의 약어나 전문 용어를 오역한 경우, 성별이나 직위를 확인하여 태국어답게 번역해야 할 지칭어를 오역한 경우, 연어나 문맥을 고려하지 않고 번역한 경우, 담화 표지 번역이 어색한 경우 등이 있다.</li> <li>• 표기법에 대한 오류에는 인명, 행정 구역의 지명, 음식 등과 같은 고유 명사 음역 방식이 태국의 왕립학술원 규정과 다르게 번역한 경우 등이 있다.</li> <li>• 기타 오류에는 띄어쓰기나 문장 부호(“, :, ~ 등)가 잘못 표기된 경우, 영어 차용어의 성조 부호를 포함한 사소한 오타자, 문어체의 시간 표기, 한국어 발음 표기법, 영어 표기가 더 적절한 고유 명사, 기관명의 약어 병기를 삭제할 필요가 있는 경우 등이 있다.</li> <li>• 내용에 관한 오류는 담화적인 의미가 포함되어 있어서 정확한 내포 의미를 이해하고 맥락에 맞는 적절한 해석이 선행되어야 할 문장들이 대부분이다. 그리고 문어체에서 내용이 길고 문장 구조가 조금 복잡한 문장에서 오류가 있었고, 구어체는 신조어나 관용 표현이 포함된 문장 등에서 원문의 의미가 왜곡된 경우가 많았다.</li> <li>• 문법 오류에는 어순을 바꾸고 잘못 번역한 경우를 예로 들 수 있으며, 그 외에도 태국어 문법상 전치사, 수동태·능동태, 시상, 관계대명사 등을 잘못 사용하여 번역한 경우 등이 있다.</li> </ul>



인도 힌디어	<ul style="list-style-type: none"> <li>• 내용 오류의 비율이 가장 높게 나타났는데 원문 이해 부족에서 나타난 의미 불일치(축약어 등)와 내용 누락이 많았다.</li> <li>• 다음으로 어휘 오류의 비율이 높았는데 부적합한 어휘 번역이 많이 보였다.</li> <li>• 문법 오류의 경우 어순 오류가 가장 많았는데 한국어 어순 그대로 번역이 가능한데 어순을 바꾸어서 오류로 처리된 경우가 많았다.</li> <li>• 표기법 오류는 고유 명사 음역에서 많이 나타났다.</li> </ul>
캄보디아 크메르어	<ul style="list-style-type: none"> <li>• 오류 유형 중 어휘 오류가 전체 오류의 3분의 1을 차지했다. 한자 문화권이 아닌 캄보디아 크메르어의 특성상, 동음이의어에 대한 이해 부족으로 대응어를 잘못 적용한 오류(예: ‘개인 차’의 ‘차’를 ‘차이’가 아닌 ‘자동차’로 번역)와 정치 및 법률과 관련한 전문 용어에 대한 오류가 많았다.</li> <li>• 내용 오류는 구어체의 경우 고맥락과 도치로 인해 내용 파악의 어려움으로 인한 오류가 많았고 문어체의 경우 한국의 정치, 경제, 사회적 이슈에 대한 이해도가 낮은 것에서 기인한 오류가 많았다. 또한 번역문에서 원문의 내용을 상당 부분 누락하고 번역해 놓은 문장도 다수 발견되었다.</li> <li>• 표기법의 오류도 상당수 발견되었는데 인명, 지역명의 음역에서 나타난 오류가 많았다.</li> <li>• 문법 오류는 접사, 접속사 등의 오류가 많았다.</li> <li>• 기타 오류에는 마침표, 물음표, 띄어쓰기 오류가 있었다.</li> </ul>
필리핀 타갈로그어	<ul style="list-style-type: none"> <li>• 가장 높은 오류는 어휘 오류다. 고유 명사의 경우 인명, 작품명, 기관명, 지역명에 지침대로 일관성 있게 번역해야 하는데 이 부분에서 수정이 많았다. 또 적절한 어휘 대응이 안 되는 경우 조금 더 정확하게 한국어의 의미를 반영할 수 있는 어휘(주로 동사)로 수정하였다.</li> <li>• 다음으로 내용 오류가 높게 나왔는데 이는 누락, 첨가의 원인이 가장 많았고 한국어 원문의 이해 부족도 원인이었다. 누락은 한국어의 구어적 특징인 간투사 미반영, 접속부사 미반영, 문장 내 삽입 구를 전체 미반영하는 경우도 있었다. 첨가는 공식 명칭을 찾기 어렵거나 적절한 대응어가 없을 때, 관용적 표현의 번역 능력이 부족할 때 1:1 번역보다는 번역사의 이해를 바탕으로 의미역하면서 발생하였다.</li> <li>• 문법적 오류는 po 높임 표현의 부적절한 번역, 격조사('-에'를 '-을/를'에 해당하는 ng으로 번역)의 오류, 피사동의 부정확성, 과거 시제의 미반영으로 인한 오류 등이 있었다.</li> </ul>
러시아어	<ul style="list-style-type: none"> <li>• 기타를 제외하고 어휘 오류의 비율이 가장 높았는데 원문의 의미를 정확하고 자연스럽게 전달하기 위해 수정한 경우가 많았다.</li> <li>• 다음으로 비율이 높은 표기법 오류의 경우, 음역 시 콘체비치 체계 미준수, 쉼표 및 따옴표 등 문장 부호 사용의 오류, 단순 철자 오류 등이 나타났다. 특히 러시아어에서는 쉼표의 사용 규칙이 엄격하여 문장 부호의 오류가 높은 편이다.</li> <li>• 문법 오류는 문장 성분의 성·수 불일치, 주절과 종속절의 주어 불일치, 시제 및 어순의 오류 등이 있었다.</li> <li>• 이 외에 내용 오류로는 한국어 원문의 의미가 일부 누락되거나 불필요한 문장 성분들이 추가된 경우 등이 있었다.</li> </ul>

우즈베크어	<ul style="list-style-type: none"> <li>• 어휘 오류의 비율이 가장 높게 나타났는데 원문의 의미에 맞는 적합한 어휘 선택, 인명, 지역명, 기관명 어휘의 수정 과정에서 어휘 오류의 빈도가 높게 나타났다.</li> <li>• 다음으로 높게 나타난 오류는 문법 오류로 시제 및 어순의 오류, 조사 오류 등이 있었다.</li> <li>• 표기 오류는 우즈베크어에 적용되어야 하는 하이픈(-) 누락, 숫자 표기, 특히 h와 X 표기의 혼동, 단순 철자 오류 등이 다수 나타났다.</li> <li>• 그 외 내용 오류로서 원문의 의미와 불일치, 내용 누락, 내용 첨가 등이 나타났다.</li> </ul>
-------	---

## 5. 용례 검색기

### 5.1. 프로토타입 개발

본 사업의 결과물인 병렬 말뭉치 배포 형식의 경우 JSON 형식으로, 관련 지식이 없는 경우 접근 및 활용에 어려움을 겪을 수 있다. 이에 결과물을 쉽게 활용할 수 있도록 용례 검색기 프로토타입 개발을 추가로 제안하였다.

용례 검색기는 일반 사람들도 손쉽게 이용할 수 있는 도구로, 구축된 병렬 말뭉치를 활용하여 웹 기반 용례 검색기 프로토타입을 개발함으로써 말뭉치 이용의 진입 장벽을 낮추고 언어·외국어 교육 및 연구 분야에서의 활용도를 높이하고자 하였다.

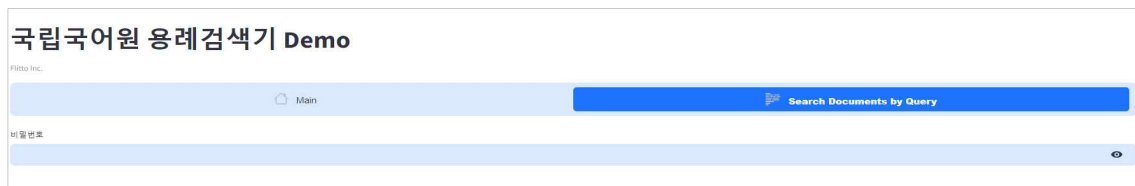
2021년 사업에서 구축한 병렬 말뭉치를 활용하여 웹 기반 용례 검색기의 성능을 검증하고, 시범 사용을 통해 발견된 문제점을 개선하여 용례 검색기의 활용도를 높이하고자 하였다. 구체적인 자료와 개발 환경은 다음과 같다.

<표 31> 용례 검색기 대상 자료 및 환경

대상 자료	<ul style="list-style-type: none"> <li>• 2021년 한국어-외국어 병렬 말뭉치</li> <li>- 문어체 29,987문장(424,358어절)</li> <li>- 구어체 102,057문장(635,764어절)</li> </ul>
개발 환경	<ul style="list-style-type: none"> <li>• 서버 환경</li> <li>- OS : Ubuntu 20.04</li> <li>- Framework : streamlit</li> <li>- Database : duckDB</li> <li>- 개발 언어 : Python</li> <li>• 클라이언트 환경</li> </ul>

	<ul style="list-style-type: none"> <li>- 운영 체제 : 아래 환경에서 테스트 완료</li> <li>- Windows 계열, Linux 계열, MacOS</li> </ul>
--	---

용례 검색기 내 말뭉치는 2021년 사업에서 자체 구축 및 구매로 저작권 이용 허락을 체결한 데이터를 사용하였으며, 사용자 무단 사용 방지를 위하여 인증된 사용자만 접근이 가능하도록 인증 암호기를 설정하였다.



[그림 30] 용례 검색기 암호기 설정

## 5.2. 시범 사용 및 자문 의견

국내외 기존 용례 검색기는 단일 언어 위주라는 점을 고려하여, 본 사업의 8개 언어가 모두 검색될 수 있는 다국어 말뭉치 데이터에 최적화된 기능, 사용 가치를 확인할 수 있는 용례 검색기 프로토타입을 구현하는 것을 목표로 하였다.

8개 언어를 용례 검색기에 저장할 스키마를 설계할 때 모든 언어 쌍의 확장 가능성을 고려해야 한다. 예를 들어 한국어 말뭉치 데이터를 이용해 베트남어, 러시아어의 말뭉치 데이터를 생성하려면 한국어-베트남어 병렬 말뭉치 외에 베트남어-러시아어 병렬 말뭉치도 자동적으로 생성될 수 있도록 해야 한다.

<표 32> 용례 검색기 스키마 예시

<pre>{   [     group_id : 1,     texts : [       {         "text_id" : 1,         "lang_code" : "ko",         "text" : "나는 매일 아침 학교에 간다."       },       {         "text_id" : 1,         "lang_code" : "vi",</pre>
---

```

        "text" : "Tôi đến trường mỗi sáng."
    },
    ...
    {
        "text_id" : 1,
        "lang_code" : "ru",
        "text" : "Каждое утро я хожу в школу."
    }
]
],
[
    "group_id" : 2,
    ...
],
...
}

```

위와 같은 방식으로 원문인 한국어 기준으로 그룹화하여 8개 언어 쌍에 대해서 병렬 말뭉치가 만들어지므로 효율적인 설계가 가능하다.

위 데이터 처리 기술을 바탕으로 용례 검색기를 개발하는 과정에서 기술 자문위원을 구성하여 필수적인 용례 검색기 기능에 대한 자문 의견을 받았다. 용례 검색기를 개발하기 위해서는 대표적으로 검색 속도, 검색어의 기준, 검색 결과의 편의성을 고려해야 한다는 의견을 반영하여 개발의 방향성을 조정하였다.

<표 33> 용례 검색기 자문 및 보완 내용

기존	보완	보완 사유
검색 단어가 포함된 문장, 포함되지 않았지만 유사한 문장 모두 결과에 포함	검색 단어가 포함된 문장만 결과에 포함	검색 결과의 신뢰성 문제, 검색어 기준 구분의 모호함
문장 내 검색 단어에 대한 구분 없음	문장 내 검색 단어에 대해 하이라이트 표시	검색 결과 확인의 편의성
한정적인 검색 기능	와일드카드 검색 기능 추가	검색 결과 확인의 다양성, 편의성

이외 검색 속도를 개선하기 위해 제한된 서버의 성능 내에서 8개 언어를 한꺼번

에 처리해도 성능에 문제가 없도록 검색 알고리즘을 잘 설계하는 것이 필요하다. 검색 시 일치하는 단어 검색 알고리즘을 구성할 때 관계형 데이터베이스(RDBMS : Relational Database Management System)는 많은 데이터들의 관계를 정의하면서 저장하는 데 유리하다. 반면 말뭉치는 텍스트 타입으로 저장하는데 관계형 데이터베이스에서는 문자열에 포함된 특정 단어를 검색하는 데 많은 시간이 걸린다.

따라서 기존 관계형 데이터베이스보다는 필요한 정보만 저장하면서 검색에 최적화된 Elasticsearch 등을 고려하였다. Elasticsearch는 Lucene이라는 프레임워크 기반으로 데이터를 저장하면서 단어별로 색인을 생성한다. 단어와 색인이 정의된 표만 검색하면 실제 문서의 번호를 찾을 수 있으므로 기존 데이터베이스보다 훨씬 빠르게 검색할 수 있도록 개선하였다.

이와 같은 개선 과정을 거쳐 용례 검색기 프로토타입을 개발하였으며, 상세한 기능 사항은 다음과 같다.

<표 34> 용례 검색기 프로토타입 기능 사항

화면 구성	<p><b>1. 메인, 용례 검색으로 구성</b></p> <p><b>가. 메인</b></p> <ul style="list-style-type: none"> <li>- 개요</li> <li>- 기능 소개</li> <li>- 주의 사항</li> <li>- 사용 안내서 다운로드</li> </ul> <p><b>나. 접근 방식</b></p> <ul style="list-style-type: none"> <li>- 인증 암호키 입력 후, [Search Documents by Query] 화면으로 접근</li> </ul> <p><b>다. Search Documents by Query</b></p> <ul style="list-style-type: none"> <li>- 검색 단위: 어절</li> <li>- 검색 분류: 전체, 문어체, 구어체</li> <li>- 검색어</li> <li>- 언어(쌍) (복수 선택 가능)</li> <li>- 다운로드 (xlsx)</li> </ul>
검색 기능	

[그림 31] Search Documents by Query 검색 예시

### 1. 검색어

- 어절 기준으로 검색

### 2. 상세 검색

- 검색 형태 선택 (전체, 문어체, 구어체)
- 검색어 입력 (어절, 와일드카드)

### 3. 언어 (쌍)

- 검색 언어 설정 (복수 선택 가능)
- 한국어(ko), 베트남어(vi), 인도네시아어(id), 태국어(th), 인도 힌디어(hi), 캄보디아 크메르어(km), 필리핀 타갈로그어(tl), 러시아어(ru), 우즈베크어(uz) 지원

### 4. 와일드카드 검색 (한국어, 외국어)

- \*(별표): 개수에 상관없이 모든 문자 대체를 의미
- \*가: '가'로 끝나는 모든 글자
- 가\*: '가'로 시작하는 모든 글자
- ?(물음표): 하나의 문자로 대체를 의미
- ?가: '가'로 끝나는 2음절어
- ?가?: 중간에 '가'가 들어가는 3음절어
- 가??: '가'로 시작하는 3음절어

### 검색 결과

#### 1. 검색 내용 확인

- 검색 결과 문장 수 확인
- 검색어가 반드시 포함된 문장만 추출

#### 2. 검색 결과 표 확인

- 결괏값: 문장 번호, 검색한 언어 쌍 순서대로 배치
- 한 화면당 문장 50개씩 추출

#### 3. 검색 단어 하이라이트

- 검색한 문자는 빨간색으로 구분

#### 4. 와일드카드 검색 결과 예시

- \*가 와 ?가? 검색 결과

No.	ko	th	th
1	제가 저 프로그램으로 <b>자동차</b> 를 샀어요.	ආයුරුමෙහිදී මම මෝටර් රථයක් ගත්තේ.	В этот же компьютер я купил автомобиль.
2	마이, 당신 <b>자동차</b> 를 그렸어?	ඔබේ මෝටර් රථය ගැන?	О, вы нарисовали машину самому?
3	<b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
4	<b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
5	그 <b>자동차</b> 를 누가 디자인했어? 정말 좋아할 거예요?	මේ <b>자동차</b> කවුරු විද්‍යා කළා? ඔබට ඉඩක් තියෙනවා?	Кто это проект сделал, чтобы очень понравился?
6	네, 저는 나랑 <b>자동차</b> 가 최고예요!	ඔබගේ <b>자동차</b> විද්‍යා කළාට ඔබට ඉඩක් තියෙනවා.	Да, я люблю и думаю, что это самое лучшее!
7	<b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
8	나와 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
9	그리고, 저가 디자인한 저 <b>자동차</b> 를 <b>자동차</b> 를 사주세요.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Именно, если дизайн удачи, то можно купить машину.
10	수많은 <b>자동차</b> 를 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
11	바람이 불면 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
12	<b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
13	그리고, <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
14	수많은 <b>자동차</b> 를 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
15	바람이 불면 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
16	<b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
17	나와 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
18	수많은 <b>자동차</b> 를 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.
19	바람이 불면 <b>자동차</b> 를 사고 싶고 정말 <b>자동차</b> 가 좋아.	මම මෝටර් රථයක් ගන්නවාට ඉඩක් තියෙනවා.	Машину очень хочется и люблю.

[그림 32] 검색 결과 표 예시

No.	ko	id
1	여, 여자가 그 상지르기요?	Ah, rupanya ini adalah Saengji yang itu!
2	그게서 그 용원 <b>한가</b> 가 뭐예요?	Jadi apa kalimat terakhirnya?
3	나 <b>한가</b> 가 뭐예요? 정말 좋아할 거예요?	Karena aku cenderung penuh kerentanan, jadi aku pasti menaruhkan O seperti ini!
4	한가 <b>한가</b> 가 뭐예요?	Apa kamu tahu asal-mula 'hopping'?
5	<b>한가</b> 그건 왜 있어요?	Ada sesuatu seperti itu.
6	<b>한가</b> 더 한 건지 하잖아?	Aku pun kan peka.
7	<b>한가</b> 더 한 건지 하잖아?	Diperjeng-jing yang direbuskan oleh mami!
8	두 <b>한가</b> 가 있는 건지 아예 안은 맞아요?	Ada dua hal, ini benar penjasannya, kan?
9	<b>한가</b> 좀 이해해 줘.	Kuharap kamu mengerti.
10	그런 <b>한가</b> 좀 설명해 줄게.	Kalau begitu, aku akan mendengarkan lalu menjelaskan juga!
11	아이고, <b>한가</b> 그걸 뭐예요?	Aduh, kamu memang seperti itu!
12	한가 <b>한가</b> 는 어떤 <b>한가</b> 가 있어요?	Akuwa Tinghua apa yang ada di within karaker?
13	재발 <b>한가</b> 로만 <b>한가</b> 보지.	Semoga berjalan seperti ini saja.

[그림 33] ‘\*가’ 검색 시 결괏값

No.	ko	id
1	그런 <b>한가</b> 스 문장은 거지?	Unda, Tonkasi bo'ladiku-a?
2	나츠로 <b>한가</b> 할 수 있지 않나?	Nachosi ham go'ha olmasimkimi?
3	그렇게 너무 교배하는 <b>한가</b> ?	Shuni ayt, haddan oshirib yabordinmi?
4	내 것도 그런 <b>한가</b> ?	Manik ham shundaymikan?
5	여가다가 되면 되는 <b>한가</b> ?	Bu yengir tekizam bo'ladimi?
6	그렇게, 저렇게 가도 되는 <b>한가</b> ?	Shuda, mana shunday ketu ham bo'lamikan?
7	당지도 직접 읽는 <b>한가</b> ?	Kimchi ham o'zimiz olamizmi?
8	그렇게, 조금만 사한 <b>한가</b> ?	Shunaqaku, biraz tuzlanganimmi?
9	도 스토리이크리니 이렇 <b>한가</b> ?	Yana zarbami, ajblanar!
10	몇기 전에 할 일이라 쓰러지는 <b>한가</b> ?	Kalatk yeyilshidn oldin yaxshilicha tushirib ol deganimmi bu?
11	이렇게 하면 되는 <b>한가</b> ?	Shunday qilam bo'ladimi?
12	네, 저렇게 적지 <b>한가</b> 는 과정을 미개 알려 주십시오?	Ha, menga tashit izanish jaryonini endi o'ngatmagil!

[그림 34] ‘?가?’ 검색 시 결괏값

## 5. 다운로드

- 검색 결과 오른쪽 상단에 다운로드 기능 박스
- xlsx 다운로드 지원

다운로드

count

410

NIKL\_result.xlsx

Download

[그림 35] 검색 결과 문장 수, 다운로드 예시

A		B	C
1	No.	ko	ru
2	1	저기 저 조그만 초록색 자동차 말이야.	Я про ту маленькую машину зеленого цвета.
3	2	자동차 끌고 오길 정말 잘한 거 같아.	Как хорошо, что мы приехали на машине.
4	3	자동차 극장도 다음에 가자!	Давай потом и в автокинотеатр пойдём!
5	4	바람 많이 불면 자동차 주인이 알아서 나오지 않을까?	Если ветер будет сильно дуть, не выйдет ли хозяин машины сам?
6	5	자동차 사고는 정말 조심해야 돼.	Надо быть очень осторожными с автомобильными авариями.
7	6	그러게, 자동차 없었으면 정말 고생했겠다.	Да, без неё нам бы пришлось непросто.
8	7	검은색 반소매 차림을 한 경호원들은 차량이 유턴하는 동안 왕복 6차선인 국제 보상으로 달리는 자동차 사이로 끼어들어 운행을 막아섰다.	Охранники, одетые в черную одежду с короткими рукавами, во время движения транспорта по шестиполосной улице Кукепосанно с двусторонним движением влезали между машин и блокировали движение.
9	8	이 밖에 '자동차 관리법' 개정으로 자동차 결함에 대한 자동차 제작사 등의 처벌과 관리를 강화한다.	Кроме того, пересмотр "закона о контроле автомобилей" усилит наказание и контроль над производителями автомобилей и другими лицами в случае дефектов автомобилей.
10	9	한편 '분노의 질주9'는 지상과 상공을 넘나들며 자동차 추격 액션을 펼쳐 영화 애호가들에게 박수받았다.	тречен любителями кино, которые по достоинству оценили экшен, насыщенный сценами и погони на автомобилях на земле и в воздухе.
11	10	지하 주차장에 주차해 둔 자동차 조수석에서 혼자 낮잠을 잔 적이 있어.	Однажды я одна задремала на переднем сидении в припаркованной на подземной парковке машине.
12	11	저자는 자신의 아파트 구매기, 자동차 차종 선택 경험 등을 소개하면서 경제 현상을 풀어 설명한다.	Автор объясняет экономические явления, рассказывая о покупке собственной квартиры, опыте выбора модели автомобиля и т.п.

[그림 36] 엑셀 다운로드 예시

### 5.3. 프로토타입 보완

용례 검색기 프로토타입은 용례 검색기의 필수적인 기능을 구현하고 다국어 병렬 말뭉치의 활용성을 확인하였지만, 공공에 개방하여 언어·외국어 교육 및 연구 분야에 폭넓게 사용되기 위해서는 추가적인 기능이 필요할 것으로 보인다.

사용성 향상을 위해 검증된 오픈소스를 통해 한국어 혹은 외국어 형태소의 분석이나 단어 의미의 해석을 연동하여 사용할 수 있는 기능을 추가하는 것을 고려해 볼 수 있다. 하지만 다국어 말뭉치라는 점과 영어, 일본어와 달리 소수 언어인 점을 고려하여 기능 구현의 현실적인 부분을 해결하는 것이 필요하다.

그 다음날도 같은 일을 반복했다.
그[그.NP] 다음날도[다음날/NNG+도/X] 같은[같은/VA+은/ETM] 일을[일/NGG+을/JKQ] 반복했다.[반복/NGG+하/XSV+다/EP+다/EF+/SE]
その次の日もまた同じ事を繰り返した。
その[その/GS] 次[次/NG] の[の/PCS] 日[日/NDEADP] に[に/PJKQ] も[も/JL/PR] また[また/CJ] 同[同/J/CS] 事[事/NDEG] を[を/PJKQ] 繰[繰/J/CS] り返[繰り返す/VIN] し[し/JAU] 。[。/SYE]

[그림 37] 온라인 세종 한일 병렬 말뭉치 형태소 검색 예시

검색 결과의 다양성을 위해 프로토타입의 필수 검색 기능을 넘어서 공기어 분석, 관련어 분석 등 유사한 단어를 보여줄 수 있는 추가 기능을 구현하는 것을 고려해 볼 수 있다. 국립국어원에서 구축한 말뭉치는 문어체, 구어체 등 다양하므로 일상

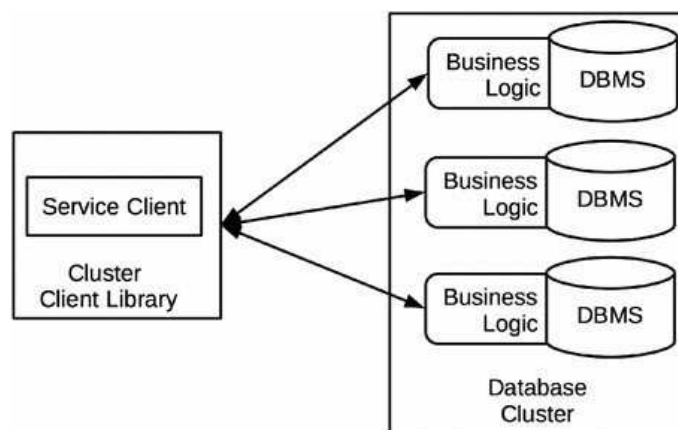


생활에서 쓰이는 단어 대부분이 말뭉치에 포함되어 있지만 용례 검색기를 통해 검색했을 때 단어가 포함된 문장이 없을 수도 있다. 이러한 경우 형태가 일치하지는 않지만 의미상 유사한 단어로 검색해서 제공한다면 사용자에게 도움이 될 수 있을 것이다.

<표 35> 유사 단어 검색 예시

1. 일치하는 단어만 검색하는 경우:
  - 검색어: 책방
  - 검색 결과 없음
2. 유사한 단어를 추가로 검색하는 경우
  - 검색어: 책방
  - 검색 결과
    - 한국어: 나는 매주 토요일마다 동네에 있는 서점에 간다.
    - 베트남어: Tôi đi đến hiệu sách trong khu phố vào mỗi thứ bảy hàng tuần.
    - 러시아어: Каждую субботу я хожу в местный книжный магазин.
    - ...

또한 이후 공공에 개방하여 누구나 자유롭게 사용하기 위해서는 말뭉치를 안정적으로 저장할 수 있는 스토리지 이중화가 필수적이다. 클라이언트 저장소와 데이터베이스 저장소를 따로 두어 서비스 중단이 발생하지 않도록 대상 데이터를 안정적으로 저장해야 한다.



[그림 38] 스토리지 이중화

## 6. 병렬 말뭉치 구축 및 메타 정보

## 6.1. JSON 포맷

국립국어원 말뭉치 metadata JSON 형식						
1수준	2수준	3수준	4수준	5수준	타입	설명
id					str	말뭉치 파일 ID
metadata					obj	파일의 메타 정보
	title				str	파일 제목
	creator				str	생성자
	distributor				str	배포자
	year				str	구축 연도
	category				str	분류
	annotation_level				arr(str)	분석 층위
	sampling				str	샘플링 방식
document					arr(obj)	
	id				str	문서 ID
	metadata				arr(obj)	문서의 메타 정보
		title			str	문서 제목
		author			str	저자
		publisher			str	발간자
		date			str	생성일
		topic			str	국어원 설정 주제
		original_topic			str	타 기관 설정 주제
		speaker			arr(obj)	
			id		str	발화자 ID
			age		str	발화자 나이
			sex		str	발화자 성별

[그림 39] metadata JSON 형식

국립국어원 말뭉치 document_paragraph JSON 형식						
1수준	2수준	3수준	4수준	5수준	타입	설명
document_paragraph					arr(obj)	
	id				str	문서 ID
	paragraph				arr(obj)	
		id			str	문장 ID
		form			str	정제된 문장
		original_form			str	정제 이전 문장

[그림 40] document\_paragraph JSON 형식

1수준	2수준	3수준	4수준	5수준	타입	설명
language_info					obj	
	source_language				str	출발어
	target_language				str	도착어
parallel					arr(obj)	
	id				str	문장 ID
	source				str	정제문
	target				str	최종 번역문
	revision				arr(obj)	
		revision1			str	1차 검수 문장
		revision2			str	2차 검수 문장

[그림 41] 병렬 말뭉치 JSON 형식

## 6.2. 최종 데이터 구조

[illegible]

[그림 42] 구축 데이터 샘플 Excel 예시

### 6.3. JSON 예시

<표 36> 메타데이터 JSON 형식 예시

```
{
  "id": "NIOR2302402070",
  "metadata": {
    "title": "국립국어원 병렬말뭉치 원천 자료 문어 NIOR2302402070",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2023",
```

```

        "category": "신문>전국 종합지",
        "annotation_level": [
            "원시"
        ],
        "sampling": ""
    },
    "document": [
        {
            "id": "NIOR2302402070.1",
            "metadata": {
                "title": "",
                "author": "",
                "publisher": "국립국어원",
                "date": "",
                "topic": "정치>선거",
                "original_topic": "정치",
                "speaker": {
                    "id": "",
                    "age": "",
                    "sex": ""
                }
            }
        },
        {
            "id": "NIOR2302402070.2",
            "metadata": {
                "title": "",
                "author": "",
                "publisher": "국립국어원",
                "date": "",
                "topic": "정치>정치일반",
                "original_topic": "정치",
                "speaker": {
                    "id": "",
                    "age": "",
                    "sex": ""
                }
            }
        }
    ]
}

```

<표 37> 병렬 말뭉치 JSON 형식 예시

```

{
    "language_info": {
        "source_language": "ko",

```

```

        "target_language": "id"
    },
    "parallel": [
        {
            "id": "NIOR2302402070.1.1",
            "source": "무술년을 관통하는 핵심 키워드 중 하나는 '58년 정치인'이다.",
            "target": "Salah satu kata kunci inti yang dilalui di tahun 2018 adalah '58 tahun berpolitik'.",
            "revision": {
                "revision1": "Salah satu kata kunci inti yang dilalui di Tahun Anjing adalah '58 tahun berpolitik'.",
                "revision2": "Salah satu kata kunci inti yang dilalui di tahun anjing adalah '58 tahun berpolitik'."
            }
        },
        {
            "id": "NIOR2302302070.1.2",
            "source": "87년 체제 이후 소위 386세대보다 존재감이 약했던 이유다.",
            "target": "Ini alasan mengapa setelah sistem tahun 1987, keberadaannya disebut lebih lemah daripada generasi 386..",
            "revision": {
                "revision1": "Inilah alasan mengapa sistem setelah tahun 1987 keberadaannya disebut lebih lemah daripada generasi 386.",
                "revision2": "Ini alasan mengapa setelah sistem tahun 1987, keberadaannya disebut lebih lemah daripada generasi 386."
            }
        }
    ]
}

```

## 7. 보안

본 사업단에서는 사업 수행 과정에서의 참여 인력, 장비, 자료 등에 대한 물리적·기술적 관리를 통해 안전하고 체계적으로 사업을 수행함을 목적으로 보안 관리 대상과 보안 관리 담당자를 지정하였으며 이에 따라 보안 사항을 준수하여 사업을 수행하였다.

### 7.1. 보안 관리 대상 및 담당자

#### 1) 보안 관리 대상

본 사업에서는 사업장 및 과업 수행에 따른 관계 시설 보안, 정보 보안, PC 보안, 저장 매체 보안, 산출물 보안 등 물리적·기술적 사항과 참여 인력을 보안 관리 대상으로 하였으며, 기타 사업의 유형에 의해 필요하다고 판단되는 대상을 규정하였다.

## 2) 담당자 지정 체계

### (1) 보안 관리 총괄 책임자

본 사업의 보안 관리 총괄 책임은 본 사업 보조 사업자 중 최종 데이터 산출을 전담하는 ㈜플리토의 총괄 책임자에게 있으며, 총괄 책임자는 사업 전반의 보안에 대한 책임을 가지고 보안 관리자를 지정하여 운영하였다. 총괄 책임자는 본 사업과 관련된 인원·장비·자료에 대한 보안 관리 전반을 총괄하였으며 보안 관리 업무를 일부 위임할 보안 관리자를 지정하고 감독하였다.

### (2) 보안 관리자

보안 관리 총괄 책임자는 보안 관리 활동의 수행 및 통제를 위하여, 사업 관련 보안 관리와 업무상 관계가 있는 자를 보안 관리자로 지정하였다. 이에 따라 보안 관리자는 사업 관리자가 수행하였으며, 보안 관리자는 본 사업과 관련된 기술적, 물리적 보안에 대한 수행을 책임지고 사업 수행 중 전반적인 보안 활동을 수행, 감시, 통제하였다.



[그림 43] 보안 담당자 역할 체계

## 7.2. 보안 관리 방법

### 1) 물리적 보안

#### (1) 사무실·주요 장비 설치 장소에 대한 출입 보안

사무 공간은 CCTV, 잠금 장치 등 항상 출입이 통제·관리된 상태에서 인가된 인원만 출입이 가능하도록 운영하며, 공사나 시설 이동 등 비인가자 출입이 필요한 경우에는 반드시 보안 관리 총괄 책임자의 사전 승인을 거친 후에 진행하도록 하였다.

외부인의 방문이 있는 경우, 사업단 또는 플리토의 담당자가 외부인을 응대하도록 하며 지정된 장소 또는 공간 내에서만 회의 및 협의를 진행하도록 하였다.



[그림 44] 사무실 출입문 전경(좌: (사)국제한국어교육학회, 우: (주)플리토)

데이터 가공·검수 및 저장의 주요 공간인 서버가 보관된 서버실에 출입이 필요한 자는 보안 관리 총괄 책임자의 사전 승인을 받은 후 출입할 수 있도록 하였다. 사전 승인을 받아 출입한 자 및 출입 일시 등은 ‘서버실 출입 관리 대장’으로 별도 관리하였다.

#### (2) PC 및 보조 기억 장치 반출·반입 통제

데이터 가공, 검수 및 추출 등에 활용되는 모든 전산 장비(PC, 노트북 등)는 전산 장비 대장에 등록하여 관리하도록 하며, 인가되지 않은 장비의 사용은 원천적

으로 금지하였다. 또한 무선랜, 외장형 HDD 및 USB 등 비인가 전산(통신) 장비 무단 반입·사용을 금지하였으며, 관리 대상 전산 장비의 반·출입이 필요한 경우 자료 무단 방출 여부를 확인하고 대장에 기록하도록 하였다.

PC, 노트북 등은 아래 보안 관리 사항을 주기적으로 점검하고, 전체 점검 결과는 보안 관리자가 별도로 관리하도록 하였고 전산 장비 대장을 마련하여 참여 인원의 전산 장비 현황을 별도로 관리하였다. 또한 보조 기억 장치는 보조 기억 매체 관리 대장을 통해 별도로 관리하였다.

<표 38> 보안 관리 사항

- 자리 비움 시 화면 보호 기능 사용
- 백신 프로그램 및 보안 업데이트 상시 수행

### (3) 보안 관리 대장

보안 관리는 매월 다음과 같은 5개 항목, 7개 분류에 대해 점검을 진행하였으며, 2023년 5월 3주 차부터 사업 종료 시까지 보안 점검을 실시하였다.

<표 39> 보안 점검 항목 및 분류

- 보안 점검 항목 및 분류 현황
- 서버 보안: 플리토 개인 정보 암호화
- 네트워크 보안: 정보 처리 시스템 접근 통제, 시스템 접근 권한 관리
- 사업 수행 인력 관리: 전산 장비 보안
- 근무 환경 보안: 소프트웨어 패치 관리, 서버실 출입 관리
- 산출물 보안 관리

보안 관리 대장							
프로젝트 명: 2023 한-외 병렬 말뚝치 구축 사업				보안관리 책임자(정): 이정수 (플리토 CEO)			
문서번호: KFPC-007		V2.0		보안관리 책임자(부): 이재영 (플리토 B2B팀 매니저)			
				※ 보안 관리 대장은 KFPC-008(2023년 한국어·외국어 병렬 말뚝치 구축사업 보안관리계획서) 문서를 바탕으로 함			
				※ 보안 관리 대장은 월 1회 업데이트 됨(주관기관에서 요청한 보안관리 기간으로 설정함)			
NO	보안 항목 대분류	보안 항목 세부분류	참고	2023년 5월	2023년 6월	2023년 7월	2023년 8월
1	서버 보안	플리토 개인정보 암호화	KFPC-008 2.2 나. 개인정보 보호	확인	확인	확인	확인
2	네트워크 보안	정보처리시스템 접근통제	KFPC-008 2.1 시스템관리 보안	확인	확인	확인	확인
3	네트워크 보안	시스템 접근권한 관리	KFPC-008 부원3 계정 관리 대장	확인	확인	확인	확인
4	사업 수행 인력 관리	전산 장비 보안	보안 프로그램 설치 유무 확인 및 점검	확인	확인	확인	확인
5	근무 환경 보안	소프트웨어 패치 관리	KFPC-008 2.2 라 소프트웨어 패치	확인	확인	확인	확인
6	근무 환경 보안	서버실 출입 관리	KFPC-008 부원1 서버실 출입 관리대장	확인	확인	확인	확인
7	산출물 보안 관리	산출물 보안 관리	데이터 파일 암호화	확인	확인	확인	확인
NO	보안 항목 대분류	보안 항목 세부분류	참고	2023년 9월	2023년 10월	2023년 11월	2023년 12월
1	서버 보안	플리토 개인정보 암호화	KFPC-008 2.2 나. 개인정보 보호	확인	확인	확인	확인
2	네트워크 보안	정보처리시스템 접근통제	KFPC-008 2.1 시스템관리 보안	확인	확인	확인	확인
3	네트워크 보안	시스템 접근권한 관리	KFPC-008 부원3 계정 관리 대장	확인	확인	확인	확인
4	사업 수행 인력 관리	전산 장비 보안	보안 프로그램 설치 유무 확인 및 점검	확인	확인	확인	확인
5	근무 환경 보안	소프트웨어 패치 관리	KFPC-008 2.2 라 소프트웨어 패치	확인	확인	확인	확인
6	근무 환경 보안	서버실 출입 관리	KFPC-008 부원1 서버실 출입 관리대장	확인	확인	확인	확인
7	산출물 보안 관리	산출물 보안 관리	데이터 파일 암호화	확인	확인	확인	확인

[그림 45] 보안 관리 대장 예시

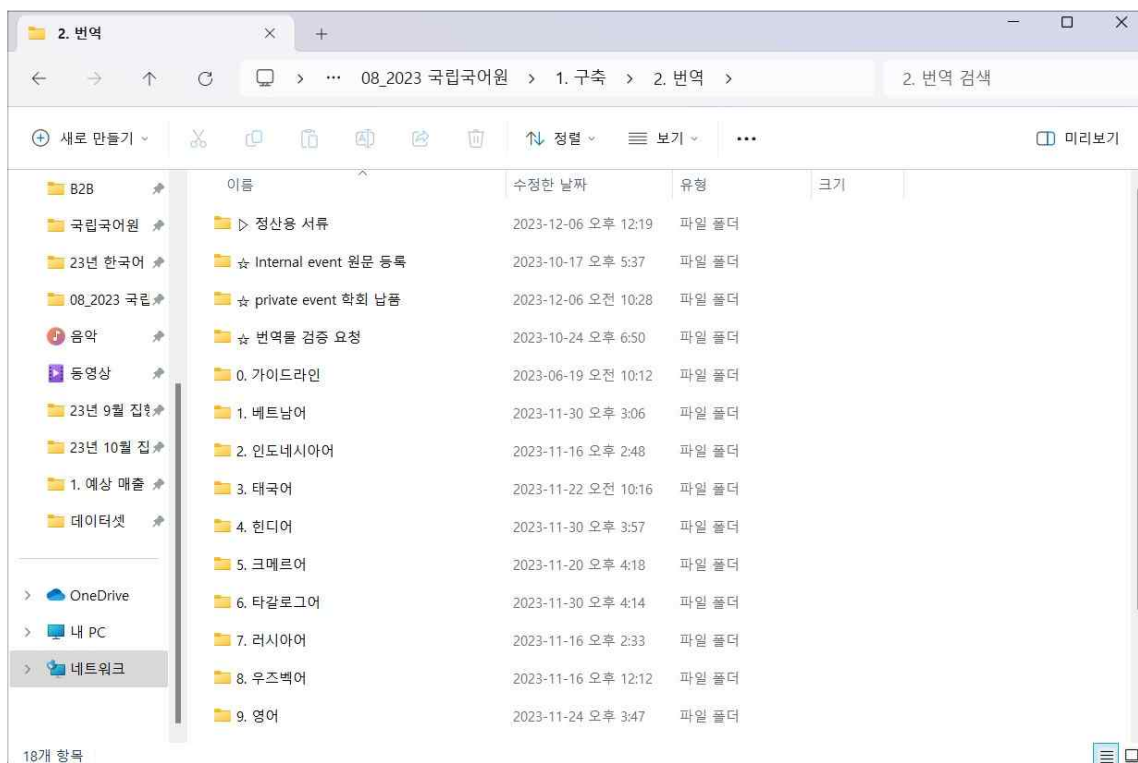


#### (4) 사업 수행 산출물 관리

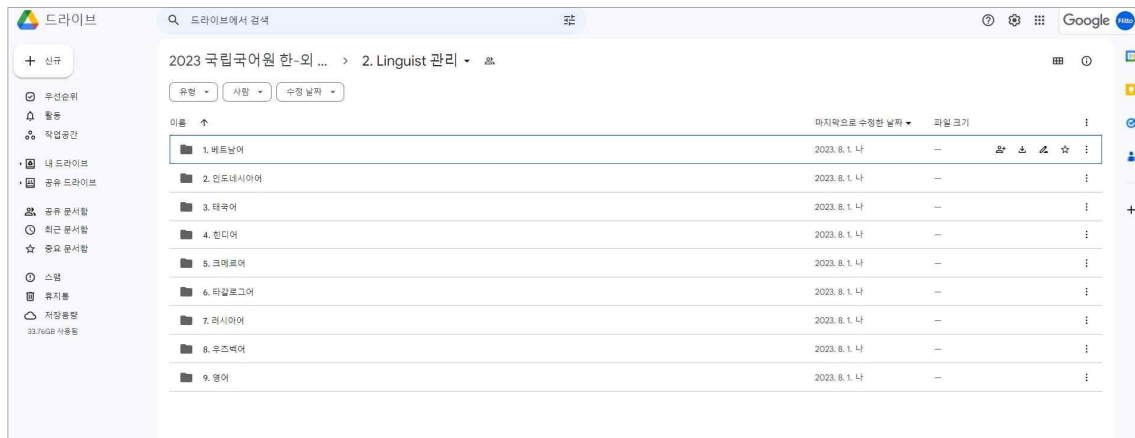
최종 산출물 도출을 위한 과정상의 결과물들은 개별 담당자들의 PC 또는 서버에 저장하였다. 최종 산출물은 사업 수행 계획에 따라 3개의 공간에 각각 동일하게 보관하며, 내·외부 자료 요청은 보안 관리자를 통해 제공하였다.

<표 40> 최종 산출물 저장 공간

1차 저장 공간	데이터셋 담당자(보안관리자) PC
2차 저장 공간	NAS (Network Attached Storage)
3차 저장 공간	Cloud (Google Drive)



[그림 46] 저장 관리 예시(NAS)



[그림 47] 저장 관리 예시(Cloud)

국립국어원과 사업단 간 자료 전송이 필요한 경우, 국립국어원 및 사업단의 전자 우편을 이용하여 송·수신하고 메일을 통해 전달되는 자료는 암호화하였다. 사업 완료 후 생산된 최종 산출물은 보안 조치를 통해 사업 관련 산출물(중간 산출물 포함)을 모두 삭제하였으며, 참여 인원 전원이 자료 미보유 확인서를 국립국어원에 제출하였다.

## 2) 기술적 보안

### (1) 시스템 관리 보안

서버를 포함한 정보 시스템에 대한 접근은 사전 인가된 사용자에게 한하여 허용하였으며 정보 시스템 접근 인가는 보안 관리자의 대면 승인을 받고, 할당받은 단말기 등을 통해서만 접근을 허용하였다.

관리자 권한으로 접속이 가능한 단말기 등을 운영하지 않으며, IP를 지정하여 운영하였고 휴대형 무선 모뎀이나 스마트폰 무선 모뎀 등 전반적으로 무선 인터넷 활용은 통제하였다.

또한 비인가자의 시스템 접속은 관련 법률 및 회사의 관련 규정에 따라 관리하였다.

<표 41> 시스템 접근 통제 관련 법률 및 규정

<b>관련 법률</b>	소프트웨어 진흥법 (시행일 2021.12.30.) 통신 비밀 보호법 (시행일 2022.10.20.) 전기 통신 기본법 (시행일 2019.06.25.) 정보 통신 산업 진흥법 (시행일 2020.12.10.) 지능 정보화 기본법 (시행일 2022.07.21.)
<b>회사 관련 규정</b>	주식회사 플리토 전산 관리 규정 (제정 2019.10.24.)

서버에 접속 가능한 모든 사용자에게 대해서 계정으로 관리하며, 사용 목적에 따라 사용자 계정을 분류하였으며 서버에 접속 가능한 계정은 계정 관리 대장을 통해 관리하였다.

(주)플리토에서 개발한 클라우드소싱 저작 도구 ‘아케이드(Arcade)’에서는 사용자를 대상으로 개별 ID를 부여함으로써 접속 경로를 폐쇄적으로 관리하고, 정보 유출의 가능성을 최소화하였다. 계정별 비밀번호는 최대 60일 이내 변경 주기를 두었고 비밀번호는 공유가 불가하며 비밀을 유지하도록 관리하였다.

또한 보안 관리 책임자는 외부 업체 등에 기술 지원을 의뢰하는 경우에도, 온라인에 의한 원격 기술 지원 작업을 허용해서는 안 됨을 원칙으로 하였다. 다만 부득이한 경우, 필요한 보안 대책을 강구한 후 원격 기술 지원 작업을 허용할 수 있으며 해당 내용에 대해서는 반드시 기록을 남겼다.

## (2) 시스템 구축 보안

시스템 구축에 관한 보안 정책은 (주)플리토의 기존 개발 및 운영 환경 보안 정책을 따랐다.

클라이언트와 웹서버 간 전송 시 암호화는 SSL 방식으로 하고 개인 정보를 처리하고 관리하는 개인 정보 처리 시스템은 DB에 저장된 개인 정보를 암호화하여 저장함으로써 개인 정보의 유출, 위·변조, 훼손 등을 방지하도록 하였다.

DB 접속 권한은 플리토의 기존 정책에 동일한 정책에 따라 운영하되, 개발자별 접속 권한을 구분하여 관리하였다.

또한 보안 취약점 개선을 위하여 운영 체제·응용 프로그램에 대해 보안 패치를 반영하도록 하였다. 보안 패치는 사업 참여 인력의 전산 장비를 대상으로 관리하며, 새로운 패치 및 업그레이드 발표 시 파일을 다운로드받아 보안 조치를 수행하여 파일을 최신 상태로 유지하였다.

### 3) 참여 인력 보안

#### (1) 사업 수행 인력 관리

본 사업 참여자 전체를 대상으로 보안 서약서를 작성 및 제출하였고 사업 착수 시, 전 투입 인력에 대한 보안 서약서를 제출한 경우 이를 대체할 수 있도록 하였다. 또한 사업 종료 시 참여자 전체를 대상으로 보안 점검 결과서를 작성 및 제출하였다.

#### (2) 보안 교육 실시

사업 참여자 전체를 대상으로 다음과 같이 보안 교육을 실시하였다.

<표 42> 사업 참여자 보안 교육

구분	내용
교육 방법	<ul style="list-style-type: none"><li>• 보안 교육은 사업 기간 내 정기 교육으로 1회 이상 실시함.</li><li>• 교육 방식은 보안 교육 자료를 활용하여 세미나 형식으로 진행함.</li></ul>
교육 대상	<ul style="list-style-type: none"><li>• 본 사업 참여자 전체</li><li>• 정기 교육 미이수자는 교육일로부터 90일 이내에 보안 관리자에 의해 전달 교육함.</li></ul>
교육 내용	<ul style="list-style-type: none"><li>• 사업 수행에 대한 보안 및 개인 정보 보호 등 전반적인 사항</li><li>• 비밀 유지 의무 준수 및 위반 시 처벌 내용 등 본 사업에 필요하다고 요구되는 내용</li></ul>
교육 기록	<ul style="list-style-type: none"><li>• 교육 이후 보안 교육 참석자 명단으로 관리</li></ul>

### 4) 사이버 보안 점검의 날 운영

(주)플리토에서 시행하는 보안 관리와 별도로 (사)국제한국어교육학회에서는 매달 첫째, 셋째 주 수요일을 ‘사이버 보안 점검의 날’로 지정하여 각 PC의 보안 사항을 점검하고 보안 교육을 실시하였다.

<표 43> ‘사이버 보안 점검의 날’ 시행 개요

날짜	PC 보안 점검 결과	보안 교육 내용
23.06.07.	양호	정보보호 실전 수칙 1-1
23.06.22.	양호	정보보호 실전 수칙 1-2
23.07.05.	양호	정보보호 실전 수칙 1-3
23.07.21.	양호	정보보호 실전 수칙 2-1
23.08.02.	양호	정보보호 실전 수칙 2-2
23.08.18.	양호	정보보호 실전 수칙 2-3
23.08.31.	양호	정보보호 실전 수칙 3
23.09.13.	양호	정보보안 보안대책 쉽게 이해하기 1
23.09.25.	양호	정보보안 보안대책 쉽게 이해하기 2
23.10.11.	양호	4차 산업혁명 시대 모바일·사이버 보안 한 번에 이해하기
23.10.25.	양호	개인정보와 개인정보보호의 이해
23.11.08.	양호	개인정보 유출방지 예방법
23.11.22.	양호	기업에서의 개인정보보호 원칙
23.12.06.	양호	기업의 정보유출 경로 및 피해사례
23.12.20.	양호	랜섬웨어란 무엇인가요?

PC 보안은 ‘AhnLab Office Security Assessment’ 프로그램을 활용하여 자가 점검을 실시하고 미흡 사항을 조치하였으며, 보안 점검 항목과 점검 예시는 다음과 같다.

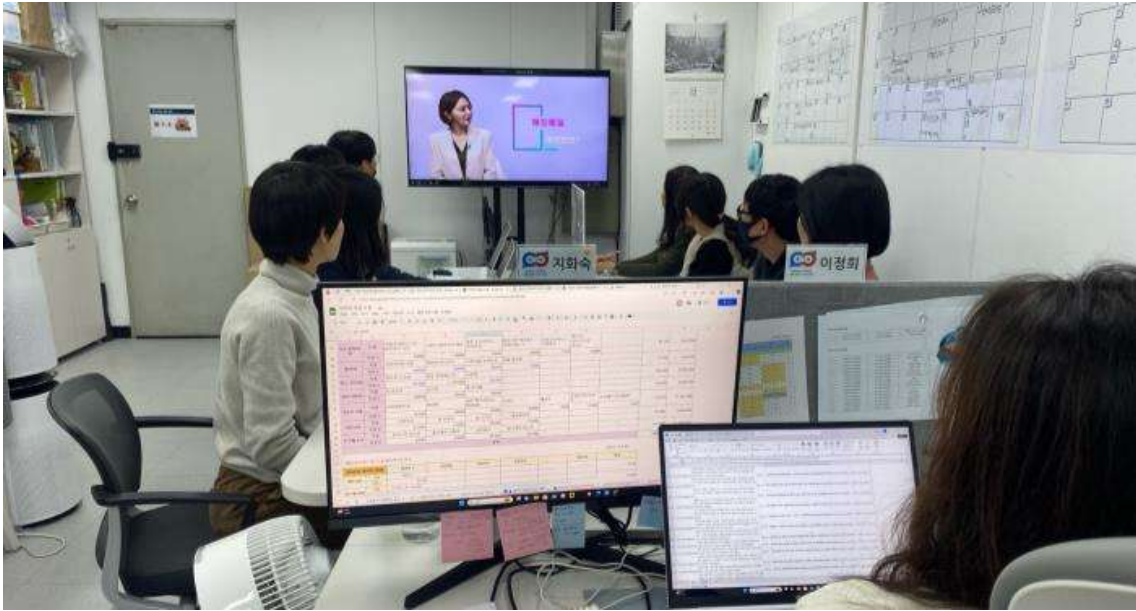
<표 44> PC 보안 점검 항목

- 악성코드 백신 설치 및 실행, 최신 보안 패치 점검
- OS 및 MS Office, 한글 프로그램 최신 보안 패치 점검
- 로그인 패스워드 안정성 및 사용 기간 점검
- 화면 보호기 설정 점검
- 사용자 공유 폴더 설정 점검
- USB 자동 실행 설정 점검
- 미사용 ActiveX 프로그램 점검

보안 교육은 본 사업의 특성에 맞게 개인 정보 보호, 데이터 보안 등에 관한 공공기관용·공개용 강의 영상을 통해 실시하였다.

연결	기기	보안 취약 건 수	보안 점검 점수
✓ 연결	[기타정보 삭제됨] [기타정보 삭제됨] [기타정보 삭제됨]	0	100
✓ 연결	[기타정보 삭제됨] [기타정보 삭제됨] [기타정보 삭제됨]	0	100
✓ 연결	[기타정보 삭제됨] [기타정보 삭제됨] [기타정보 삭제됨]	0	100
✓ 연결	[기타정보 삭제됨] [기타정보 삭제됨] ( [기타정보 삭제됨] )	0	100
✓ 연결	[기타정보 삭제됨] [기타정보 삭제됨] [기타정보 삭제됨]	0	100

[그림 48] 자가 보안 점검 결과 예시



[그림 49] 보안 교육 예시







## 제 3 장

### 사업 수행 결과





## 1. 말뭉치 데이터 구축 결과

### 1.1. 최종 구축 데이터

본 사업에서는 언어별 1,380,000어절, 총 1,104만 어절의 한국어-외국어 병렬 말뭉치 구축을 목표로 하였다. 단계별로 살펴보면, 번역은 총 11,045,120어절, 검수 역시 총 11,045,120어절로 목표 대비 100.05%의 수량을 달성하였다. 추가 제안 사항이었던 한국어-영어 병렬 말뭉치도 총 1,380,640어절로 100.05% 구축을 완료하였다. 이로써 본 사업의 목표 수량이 모두 달성되었음을 알 수 있으며 최종 구축 데이터의 상세한 수량은 다음과 같다.

<표 45> 최종 구축 데이터 수량

구분		번역		검수		감수	
언어	목표	어절	비율	어절	비율	어절	비율
베트남어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
인도네시아어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
태국어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
인도 힌디어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
캄보디아 크메르어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
필리핀 타갈로그어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
러시아어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
우즈베크어	1,380,000	1,380,640	100.05%	1,380,640	100.05%	138,000	100%
<b>합계</b>	<b>11,040,000</b>	<b>11,045,120</b>	<b>100.05%</b>	<b>11,045,120</b>	<b>100.05%</b>	<b>1,104,000</b>	<b>100%</b>
영어 (추가 제안)	1,380,000	1,380,640	100.05%	-	-	-	-

### 1.2. 번역 품질 비교

최근 인공지능 기술의 발전으로 기계 번역의 성능도 지속적으로 향상되고 있는 추세이다. 그러나 관용 표현이나 동음이의어 다의어, 생략된 문장 성분 등은 원문의 맥락을 파악해야만 정확히 번역할 수 있기 때문에 여전히 오역이 자주 발생하는 항목들이다. 더욱이 데이터의 규모가 크고 질이 높을수록 정확한 맥락 파악이 가능하므로 저자원 언어일수록 그 정확성은 더 낮아질 수밖에 없다.

본 사업에서는 고품질의 병렬 말뭉치 구축에 목표를 두고 여러 절차를 통해 번역 품질을 확보하였다. 현재 대중들이 많이 사용하고 있는 두 개의 기계 번역 플랫폼과 본 사업의 번역을 비교하는 차원에서 언어별로 사례들을 문어 2개, 구어 2개씩 살펴봄으로써 품질 활동의 결과를 확인하고자 한다.

기계 번역에서는 문맥을 제대로 파악하지 못하고 동음이의어 및 다의어를 잘못 번역하거나 관용·비유 표현을 축자적 의미로 번역한 경우가 있었다. 예를 들어 기계 번역에서 휴일을 뜻하는 ‘빨간날’을 문자 그대로 ‘빨간색 날’ 혹은 ‘빨간 검’으로 번역하는 오류를 보였다. 또한 ‘팔랑귀’는 ‘귀가 얇다’라는 관용 표현이지만, 기계 번역에서는 본 뜻을 파악하지 못하고 ‘나쁜 사람, 공감이 없는 사람’의 의미로 번역하여 의미 전달이 잘못되었다.

또한, 구어체에서는 특히 한국의 사회문화적 이해가 수반되어야 하는 줄임말이나 신조어의 경우에도 오역이 발생하였다. 예를 들어 ‘강아지 상’은 ‘강아지와 비슷한 생김새나 얼굴 모습’을 뜻하는 의미로 번역해야 하는데 기계 번역에서는 문자 그대로 ‘강아지 동상’의 의미로 번역하여 의미 전달이 잘못되었다.

이외에 기계 번역에서 문장 성분의 생략이 빈번한 한국어의 특성으로 인해 주어나 목적어를 잘못 번역하여 원문의 의미가 훼손되었거나, 복잡한 문장 구조를 이해하지 못하고 과도하게 누락하여 번역한 경우도 있었다.

반면에 본 사업에서 구축한 병렬 말뭉치에서는 앞서 제시한 어휘나 표현들이 한국어 원문의 맥락을 적절히 전달할 수 있도록 번역되어 있다. 언어별로 기계 번역과 병렬 말뭉치의 번역을 비교한 내용을 상세히 제시하면 다음과 같다.

<표 46> 기계 번역과 병렬 말뭉치의 번역 비교(베트남어)

원문	문어1	시민 단체가 선정한 지점에서는 '지정 폐기물'에 해당하는 pH 12.9에 이르는 폐알칼리 수준의 물질이 나왔다.
번역	기계 번역①	Tại điểm được nhóm dân sự lựa chọn, vật liệu kiềm thải có độ pH 12,9, tương ứng với 'chất thải được chỉ định', đã được sản xuất. (At the point chosen by the civic group, waste alkaline material with a pH of 12.9, corresponding to 'designated waste', was produced.)
	기계 번역②	Tại thời điểm được lựa chọn bởi nhóm công dân, một mức độ kiềm thải của chất thải có độ pH là 12,9, tương ứng với "chất thải được chỉ định", đã được tìm thấy. (At the time selected by the group of citizens, a waste alkalinity level of waste with a pH of 12.9, corresponding to the "specified waste", was found.)

	<b>병렬 말뭉치</b>	Tại điểm được lựa chọn bởi các tổ chức dân sự, xuất hiện chất có nồng độ kiềm đạt ngưỡng pH 12,9, tương ứng với "chất thải chỉ định". (At the point selected by civic organizations, there was a substance with alkali concentration reaching the pH threshold of 12.9, which corresponds to "designated waste.")
	<b>비교</b>	기계 번역에서는 pH 12.9에 이르는 물질이 아닌, pH 12.9의 물질이 나왔다는 의미로 오역함. 병렬 말뭉치에서는 문맥에 맞는 의미로 번역함.
<b>원문</b>	<b>문어2</b>	각종 여론 조사에서 초박빙으로 흘러갔던 만큼 여야 유력 후보들의 지지층이 결집됐을 것으로 보이기 때문이다.
<b>번역</b>	<b>기계 번역①</b>	Trong khảo sát phản hồi, cơ sở hỗ trợ có ảnh hưởng sẽ chảy vào lớp băng siêu mỏng sẽ đóng vai trò quyết định. (In the feedback survey, the influential support base that will flow into the ultra-thin ice layer will play a decisive role.)
	<b>기계 번역②</b>	Điều này là do dường như cơ sở ủng hộ của các ứng cử viên đối lập hàng đầu đã được củng cố vì nó đã cực kỳ chặt chẽ trong các cuộc thăm dò dư luận khác nhau. (This is because it seems that the support base of the leading opposition candidates has been strengthened as it has been extremely tight in various opinion polls.)
	<b>병렬 말뭉치</b>	Điều này là do có thể thấy trong nhiều cuộc thăm dò dư luận khác nhau, cục diện sát sao đã diễn ra nên tầng lớp các cử tri ủng hộ của những ứng cử viên có tiềm năng thuộc chính đảng và phe đối lập đã hợp nhất lại. (This is because it can be seen in various public opinion polls that a close-up has taken place, the constituency of potential candidates of political and opposition parties has united.)
	<b>비교</b>	‘여야’는 ‘여당’, ‘야당’의 줄임말이지만, 기계 번역에서는 한 의미를 누락하여 번역함. 또한 지지층이 결집되었다는 의미를 지지층이 강화 혹은 결정적인 역할을 할 예정이라는 의미로 오역함. 병렬 말뭉치에서는 두 오류 없이 의미를 모두 살려 번역함.
<b>원문</b>	<b>구어1</b>	근데 이제 결혼 거의 막바지쯤에 가구 고르러 갈 때 그때 진짜 많이 싸웠어.
<b>번역</b>	<b>기계 번역①</b>	Nhưng bây giờ, khi cuộc hôn nhân gần kết thúc, chúng tôi đã phải đấu tranh rất gay gắt khi đi chọn đồ nội thất. (But now, as our marriage nears its end, we have struggled mightily when it comes to choosing furniture.)
	<b>기계 번역②</b>	Nhưng bây giờ, gần cuối đám cưới, khi chúng tôi đi lấy đồ đạc, chúng tôi đã cãi nhau rất nhiều. (But now, near the end of the wedding, when we went to get

		our things, we argued a lot.)
	<b>병렬 말뭉치</b>	Nhưng mà đoạn gần cuối chuẩn bị kết hôn ý, lúc đi chọn đồ nội thất lúc đấy thật sự cãi nhau rất nhiều luôn. (We were getting married at the end, and we had a lot of fights when we picked out furniture.)
	<b>비교</b>	‘결혼 거의 막바지’는 결혼을 준비하는 마지막 과정을 의미함. 하지만 기계 번역에서는 ‘결혼을 준비하는 막바지’의 의미를 ‘결혼 상태가 끝나 갈 때’로 오역하거나 ‘가구’의 의미를 누락하여 번역함.
<b>원문</b>	<b>구어2</b>	아니면 그때는 기차를 내일로 티켓을 끊어서 가도 괜찮지 않을까?
<b>번역</b>	<b>기계 번역①</b>	Hay là mua vé tàu ngày mai cũng được nhỉ? (Or maybe I can take the train until tomorrow and get my ticket?)
	<b>기계 번역②</b>	Hoặc có lẽ tôi có thể đi tàu đến ngày mai và lấy vé? (Or is it okay to buy a train ticket tomorrow?)
	<b>병렬 말뭉치</b>	Nếu không thì lúc đó mình mua vé tàu Naeillo đi cũng không sao mà nhỉ? (Or wouldn't it be okay to buy train ticket for Naeilo?)
	<b>비교</b>	본 문장의 ‘내일로 티켓’은 내일 티켓을 끊겠다는 의미가 아닌 관광 상품 ‘내일로’를 일컫는 말임. 기계 번역에서는 내일 티켓을 끊겠다는 의미로 오역하였으나, 병렬 말뭉치에서는 고유 명사 내일로의 의미를 살려 번역함.

<표 47> 기계 번역과 병렬 말뭉치의 번역 비교(인도네시아어)

<b>원문</b>	<b>문어1</b>	당시는 이육사 시인이 대표작 '청포도'를 발표한 때다.
<b>번역</b>	<b>기계 번역①</b>	Pada saat itu, penyair Lee Yuk-sa menerbitkan mahakaryanya 'Green Grapes'. (At that time, poet Lee Yuk-sa published his masterpiece 'Green Grapes'.)
	<b>기계 번역②</b>	Saat itu, penyair Lee Yuk-sa menerbitkan karya representatifnya 'Grapes Hijau'.. (At that time, the poet Lee Yuk-sa published his representative work 'Green Grapes'.)
	<b>병렬 말뭉치</b>	Saat itu merupakan saat di mana penyair Lee Yuk Sa menerbitkan 'Anggur Hijau', karya representatifnya. (This was the time when poet Lee Yuk Sa published 'Green Grapes', his representative work.)
	<b>비교</b>	기계 번역에서는 ‘청포도’를 인도네시아어로 번역하지 못하고 영어로 표기함. 또한 동사의 의미는 ‘발표했다’라는 뜻으로 오역 표기되어

		있지만, 병렬 말뭉치에서는 ‘발표한 때’로 의미를 올바르게 반영함.
원문	문어2	'소년 사건의 핵심은 속도전'이라며 '3분짜리 처분'을 남발하기도 한다.
번역	기계 번역①	"Inti dari kasus remaja adalah pertarungan kecepatan," katanya, seraya menambahkan bahwa dia juga menggunakan "disposisi tiga menit." (“The essence of juvenile cases is a battle of speed,” he said, adding that he also uses a “three-minute disposition.”)
	기계 번역②	Dia mengatakan, "Inti dari kasus anak laki-laki adalah perang kecepatan," dan "pembuangan tiga menit" berlebihan. (He said, "The essence of the boys' case was a war of speed," and the "three-minute waste" was excessive.)
	병렬 말뭉치	Beberapa orang terlalu sering menggunakan 'disposisi tiga menit', dengan mengatakan, "Kunci dari kasus remaja itu adalah kecepatan." (Some people use the 'three-minute disposition' too often, saying, "The key to a juvenile case is speed.”)
비교		기계 번역의 문장 구조는 ‘누군가가 남발한다’라기 보다는 ‘남발했다고 보인다, 인정했다’의 의미로 쓰임. 또한 소년의 뜻이 청소년이 아닌 남자의 의미로 쓰인 기계 번역 결과가 있으며, 처분을 ‘무언가를 버리거나 처리하다’의 의미로 오역함. 병렬 말뭉치에서는 문장 구조를 원문과 동일하게 유지하고 의미 표현 역시 정확하게 번역함.
원문	구어1	근데 면은 솔직히 조금 그래.
번역	기계 번역①	Tapi sejujurnya, mienya agak perih. (But to be honest, the noodles were a bit sore.)
	기계 번역②	Tapi mienya agak begitu. (But the noodles are a bit like that.)
	병렬 말뭉치	Akan tetapi, sejujurnya mi agak seperti itu. (However, to be honest, noodles are a bit like that.)
비교		본 문장의 조금 그렇다는 말은 거절 혹은 선호하지 않는다는 의미임. 하지만 기계 번역에서는 아프다는 의미로 오역하고, ‘솔직히 말해서 ~다’라는 의견 표현을 누락함. 병렬 말뭉치에서는 의견 표현과 문맥의 의미를 잘 살려 번역함.
원문	구어2	나는 백현 제일 좋아했어, 강아지 상이여 가지고.
번역	기계 번역①	Aku paling menyukai Baekhyun, dengan patung anjingnya. (I like Baekhyun the most, with his dog statue.)
	기계 번역②	Saya paling menyukai Baekhyun, saya membuat anak anjing itu bernyanyi. (I like Baekhyun the most, I made the puppy sing.)

병렬 말뭉치	Aku paling suka 'Baekhyun', karena parasnya mirip anjing. (I like Baekhyun the most, because he looks like a dog.)
비교	본 문장의 '강아지 상'은 '강아지와 비슷한 생김새나 얼굴 모습.'을 나타내는 신조어임. 기계 번역에서는 '강아지 상'을 '강아지 동상' 또는 '강아지 노래'와 같이 오역하여 문장 자체의 의미가 다르게 번역됨. 병렬 말뭉치에서는 문장 전체의 의미와 '강아지 상'의 의미를 잘 살려 번역함.

<표 48> 기계 번역과 병렬 말뭉치의 번역 비교(태국어)

원문	문어1	한 치 앞을 내다볼 수 없는 판세 속에서 최근 코로나 19 확진자까지 폭증하자 한 명의 유권자라도 더 투표장으로 이끌어 내기 위해서다.
번역	기계 번역①	ตอนนี้เพื่อดึงดูดผู้ลงคะแนนเสียงเพิ่มอย่างน้อยหนึ่งคนไปยังคูหาลงคะแนน เนื่องจากจำนวนผู้ป่วยโรคโควิด-19 ที่ได้รับการยืนยันได้เพิ่มสูงขึ้นเมื่อเร็วๆ นี้ ในสถานการณ์ที่ไม่อาจคาดการณ์อนาคตได้ (This is to attract at least one more voter to the voting booth. Due to the number of COVID-19 patients Confirmed cases have increased recently in a situation where the future cannot be predicted.)
	기계 번역②	ด้วยจำนวนผู้ติดเชื้อ COVID-19 ที่เพิ่มขึ้นเมื่อเร็ว ๆ นี้ท่ามกลางสถานการณ์ที่คาดเดาไม่ได้ (With the recent increase in COVID-19 cases amid an unpredictable situation.)
	병렬 말뭉치	ท่ามกลางสถานการณ์ที่คาดเดาไม่ได้ ก็เพื่อดึงดูดผู้มีสิทธิเลือกตั้งให้มายังสถานที่เลือกตั้งอย่างน้อยก็หนึ่งคน หลังจากมีการเพิ่มขึ้นอย่างรวดเร็วของผู้ติดเชื้อโควิด 19 เมื่อไม่นานมานี้ (Amid unpredictable situations, it is to attract at least one voter to the polling place after a recent surge in COVID 19 infected people.)
비교		기계 번역에서는 원문과 다르게 번역문을 두 문장으로 구성하거나 과도하게 누락됨. 길게 이어지는 한 문장의 복잡한 구조를 이해하지 못하고, 기계 번역문 모두 “코로나 19 확진자가 증가했다.”라는 내용으로 결론을 지었음. 병렬 말뭉치에서는 원문의 의미를 모두 살리며 구조에 맞게 정확히 번역함.
원문	문어2	배우자에게 전하고 싶은 200자 이내의 편지를 공모하면 캘리그라피 손 편지로 제작해 5월 21일 부부의 날까지 배우자에게 우편으로 전달한다.
번역	기계 번역①	หากคุณส่งจดหมายความยาว 200 ตัวอักษรหรือน้อยกว่าที่คุณต้องการส่งให้คู่สมรสของคุณ คุณจะต้องสร้างจดหมายคัดลายมือที่เขียนด้วยลายมือและส่งให้คู่สมรสของคุณภายในวันที่ 21 พฤษภาคม ซึ่งเป็นวันคู่รัก (If you are sending a letter of 200 characters or less that you would like to send to your spouse. You will need to



		create a handwritten letter and send it to your spouse by May 21, Couples Day.)
	기계 번역②	หากคุณส่งจดหมายถึงคู่สมรสของคุณภายใน 200 ตัวอักษรคุณจะได้รับจดหมายมือประดิษฐ์ตัวอักษรและส่งไปยังคู่สมรสของคุณภายในวันที่ 21 พฤษภาคม (If you send a letter to your spouse within 200 characters, you will receive an artificial letter and send it to your spouse by May 21.)
	병렬 말뭉치	หากประกาศรับจดหมายที่มีความยาวไม่เกิน 200 ตัวอักษรที่ต้องการส่งต่อให้กับคู่สมรสแล้ว จะจัดทำเป็นจดหมายเขียนมืออักษรวิจิตร และจะส่งไปรษณีย์ไปให้คู่สมรสจนถึงวันที่ 21 พฤษภาคม ซึ่งเป็นวันสามีภรรยา (If the notice of receipt of a letter up to 200 characters in length that you want to send to your spouse, it will be prepared as a handwritten letter and will be mailed to your spouse until May 21, which is husband and wife's day)
	비교	기계 번역에서는 ‘캘리그래피 손 편지 제작’과 ‘우편 전달’을 본인이 직접 하는 의미로 오역하였고, 구어체 느낌의 문장 구조를 띄고 있음. 병렬 말뭉치에서는 공모를 하면 누군가가 제작해 보내 준다는 의미를 잘 살려서 번역함.
원문	구어1	빨간날 꺼 가지고 갔었을 때일걸.
번역	기계 번역①	คงเป็นตอนที่ฉันเอาดาบสีแดงติดตัวไปด้วย (It must have been when I took the red sword with me.)
	기계 번역②	ฉันคิดว่ามันเป็นตอนที่ฉันเอาวันสีแดงไปด้วย (I think it was when I took the red day with me.)
	병렬 말뭉치	น่าจะเป็นอย่างตอนที่ไปตอนที่ติดกับวันหยุดนะ (It's supposed to be the holiday season.)
	비교	본 문장의 '빨간날'을 기계 번역에서는 ‘붉은 검’ 또는 ‘빨간색 날’로 오역하여 ‘휴일’의 의미를 살리지 못함. 그러나 병렬 말뭉치에서는 한국어 의미에 맞게 ‘휴일’로 올바르게 번역함.
원문	구어2	사실 나도 팔랑귀라서 사람들 말에 ‘아, 그렇구나.’ 하고 동조하는 것도 없지 않아.
번역	기계 번역①	จริงๆ แล้ว ฉันก็เป็นคนเลวเหมือนกัน ดังนั้นมีคนพูดว่า ‘โ้ เข้าใจแล้ว’ และไม่มีอะไรที่ฉันเห็นด้วย (Actually, I'm a bad person too. So when people say, 'Oh, I get it,' and there's nothing I agree with.)
	기계 번역②	ในความเป็นจริงฉันไม่มีความเห็นอกเห็นใจกับคนที่พูดว่า ‘โ้ใช่’ (In fact, I have no sympathy for people who say 'oh yes')
	병렬 말뭉치	จริง ๆ แล้วฉันเองก็เป็นคนเอออย่าง ใครพูดอะไรก็จะพูดว่า "อ้อ อย่างนี้เองสินะ" แต่มักจะไม่ได้เห็นพ้องอะไรก็ตาม (Actually, I'm an easy-going person, so anyone who says something will say, "Oh, that's it," even if I don't agree with anything.)

<b>비교</b>	기계 번역에서는 ‘팔랑귀’의 의미를 오역하여 나쁜 사람, 공감에 없는 사람으로 표현함. 팔랑귀의 결과인 의견을 동조한다는 의미 역시 영향을 받아 누락되거나 오역으로 번역함. 병렬 말뭉치에서는 ‘팔랑귀’를 의미 번역하여 의미를 잘 살려 전체적으로 올바르게 번역함.
-----------	--

<표 49> 기계 번역과 병렬 말뭉치의 번역 비교(인도 힌디어)

<b>원문</b>	<b>문어1</b>	사실상 마의 7년에 접어들며 해체 수순을 밟게 된 것.
<b>번역</b>	<b>기계 번역①</b>	वास्तव में, जैसे ही हम मा के 7वें वर्ष में प्रवेश करते हैं, हम विघटन की प्रक्रिया में हैं (In fact, as we enter the 7th year of Ma, we are in a process of disintegration.)
	<b>기계 번역②</b>	वास्तव में, सात साल की आयु में, यह अभी भी काम करना शुरू कर देता है। (In fact, at the age of seven, it still begins to work.)
	<b>병렬 말뭉치</b>	दरअसल यह शापित 7वें साल में प्रवेश करके भंग करने की प्रक्रिया शुरू हो गई है। (In fact, the process of entering the cursed 7th year has begun to dissolve.)
<b>비교</b>		기계 번역에서는 ‘일이 잘되지 아니하게 해살을 부리는 요사스러운 장애물’을 의미하는 ‘마’를 말 그대로 ‘मा(Ma)’로 번역하거나 또는 누락하였음. 병렬 말뭉치에서는 단어의 의미를 살려 번역함.
<b>원문</b>	<b>문어2</b>	윤공주는 '화려함 속에 감춰진 깊이감을 느낄 수 있을 것'이라며 '이번 시즌은 또 다른 아이다가 탄생하지 않을까' 싶다고 말했다.
<b>번역</b>	<b>기계 번역①</b>	राजकुमारी यून ने कहा, "आप वैभव में छिपी गहराई को महसूस कर पाएंगे, और मुझे लगता है कि इस मौसम में एक और ऐडा पैदा होगी। ("You'll be able to feel the hidden depths in the splendor, and I think another Ada will be born this season," said Princess Yoon.)
	<b>기계 번역②</b>	राजकुमारी यून ने कहा, "मुझे लगता है कि मैं वैभव में छिपी गहराई को महसूस कर पाऊंगी," उन्होंने आगे कहा, "मुझे लगता है कि यह सीज़न एक और बच्चा होगा।" मैंने कहा '।' (Princess Yoon said, "I think I will be able to feel the depth hidden in the splendor," adding, "I think this season will be another child." I said '.')
	<b>병렬 말뭉치</b>	युन गोंग -जु ने बोला, 'आप ग्लैमर में छिपी हुई गहराई को महसूस कर सकेंगे, मुझे लगता है कि इस सीज़न में एक और आइडा पैदा होगा।'

		(You will feel the hidden depth in the glamour, I think there will be another idea this season, said Yun Gong-ju)
<b>비교</b>		기계 번역은 사람 이름인 ‘윤공주’를 동음이의어인 ‘공주’로 인식하여 ‘राजकुमारी यून (Princess Yoon)’으로 번역하였으나 병렬 말뭉치에서는 사람 이름으로 번역함.
<b>원문</b>	<b>구어1</b>	생각보다는 첫 단추는 꿔다고 봐야지.
<b>번역</b>	<b>기계 번역①</b>	सोचने के बजाय , मुझे लगता है कि पहला कदम इसे आसान बनाना है। (Instead of thinking, I think the first step is to take it easy.)
	<b>기계 번역②</b>	सोचने की बजाय पहला कदम उठाएं। (Take the first step instead of thinking.)
	<b>병렬 말뭉치</b>	ऐसा समझ सकते हैं कि सोच के परे , पहली शुरुआत तो कर ली है। (You can understand that beyond thinking, the first start has already been made.)
<b>비교</b>		기계 번역에서는 ‘생각보다는’을 ‘생각하기보다는’의 의미로 오역함. 또한 ‘첫 단추는 꿔다’라는 의미를 ‘첫 시작이 쉽다, 첫 시작을 하자’의 의미로 원문과 다른 의미로 표현되어 있지만, 병렬 말뭉치에서는 관용 의미를 살려 ‘첫 시작을 했다’라는 의미로 잘 번역함.
<b>원문</b>	<b>구어2</b>	그냥 개복숭아처럼 생겨 갖고 그걸 한 자루 따 왔어.
<b>번역</b>	<b>기계 번역①</b>	यह सिर्फ एक कुत्ते आड़ू की तरह दिखता है और मैं इसका एक बैग उठाया। (It just looks like a dog peach and I picked up a bag of it.)
	<b>기계 번역②</b>	मैं इसे अभी उठाया क्योंकि यह कुत्ते के आड़ू जैसा दिखता है। (I just picked it up because it looks like a dog peach.)
	<b>병렬 말뭉치</b>	यह एक चीनी जंगली आड़ू जैसा दिख रहा था इसलिए मैं एक टोकरी तोड़ लाई। (It looked like a Chinese wild goose so I broke a basket.)
<b>비교</b>		‘한국과 중국의 산간 지역에서 자생하는 야생 복숭아나무의 열매’인 ‘개복숭아’를 말 그대로 개와 복숭아를 붙여 오역하였으나, 병렬 말뭉치에서는 열매 품종의 이름을 알맞게 번역함.

<표 50> 기계 번역과 병렬 말뭉치의 번역 비교(캄보디아 크메르어)

<b>원문</b>	<b>문어1</b>	20여 미터에 이르는 긴 줄을 선 시민들은 체온 체크를 하고 신분증 확인을 한 뒤 소중한 한 표를 행사했다.
<b>번역</b>	<b>기계</b>	បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ បង្អួច ៣០ម៉ែត្រ

	번역①	រួចបា ឆ្នោតឱ្យ មា ឆ្នោត (People stand in a line about 20 meters long, check the temperature, check the ID card and vote for the value.)
	기계 번역②	ពលរដ្ឋ បា នត្រង់ ផ្លូវ ឆ្នោត ប្រវែង ២០ ម៉ែត្រ ឆ្លើយ ត្រិះរិះ មើល សីតុណ្ហភាព ព របស់ ពួក គេ គឺ IDs របស់ ពួក គេ បា ឆ្នោត ត្រិះរិះ ហើយ កា បា ឆ្នោត របស់ ពួក គេ បា ឆ្នោត (Citizens lined up in a 20-meter line to check their temperature, their IDs checked and their votes counted.)
	병렬 말뭉치	ប្រជា ពលរដ្ឋ ត្រង់ ប្រវែង ២០ ម៉ែត្រ ហើយត្រា ត្រិះរិះ កម្រិត សីតុណ្ហភាព អត្តសញ្ញាណ ប័ណ្ណ បន្ទាប់មកបា ឆ្នោតមួយសន្លឹកមា នៃ មធ្យម ប្រសិទ្ធភាព (People line up for about 20 meters and check the temperature and check their ID cards, then vote for one of their most valuable ballots.)
비교		기계 번역에서는 ‘소중한 표’의 의미를 ‘가치를 위한 투표’, ‘투표’의 의미로 누락, 오역하여 표현함. 병렬 말뭉치에서는 선거를 위해 ‘소중한 한 표를 행사함’의 의미를 살려 번역함.
원문	문어2	또 다른 변수는 코로나19 확진자 투표다.
번역	기계 번역①	អថេរមួយទៀត គឺ បា ឆ្នោតនៃអ្នកជំងឺ COVID-19 (Another variable is the election of the specified case on COVID-19.)
	기계 번역②	អថេរមួយទៀត គឺ បា ឆ្នោតអ្នកជំងឺ COVID-19 (Another variable is the COVID-19 case vote.)
	병렬 말뭉치	ល លើកនេះ ក្បាអថេរមួយទៀត គឺ បា ឆ្នោតរបស់អ្នកជំងឺ COVID-19 (Another variant factor is the election of COVID-19 patients.)
비교		본 문장의 의미는 코로나19 확진자들의 투표가 변수로 작용할 수 있다는 의미임. 기계 번역에서는 코로나19에 관한 선거의 의미로 오역하여 확진자의 의미를 누락함. 병렬 말뭉치에서는 확진자의 선거표라는 의미로 올바르게 번역됨.
원문	구어1	자세가 바로 서면서 마음가짐까지 당당해졌거든요.
번역	기계 번역①	ពេល ឆ្នាំ យល់ ខ្លួន មា គឺ ទៀង ក្នុង អា ស្បៀង យល់ ខ្លួន កា នែ ឆ្នាំ ឆ្នាំ (When my demeanor is honest, my demeanor becomes more confident.)
	기계 번역②	ឆ្នាំ យល់ ខ្លួន គឺ ទៀង ក្នុង ហើយ ឆ្នាំ យល់ ខ្លួន មា ឆ្នាំ ឆ្នាំ (My behaviour is honest and my mind is confident.)
	병렬 말뭉치	ព ឆ្នាំ យល់ ខ្លួន ពេល នែ ល យល់ ខ្លួន យល់ ខ្លួន ឆ្នាំ យល់ ខ្លួន (Because while standing upright, even my mental attitude is more confident.)
비교		자세를 바로 세운다는 내용을 기계 번역에서는 ‘태도가 정직하다’라는 표현으로 오역함. 병렬 말뭉치에서는 자세를 곧게 한다는 의미를 반영하여 번역함. 또한 인과관계가 드러나는 원문과 달리 기계 번역에서는 나열식으로 번역함. 병렬 말뭉치에서는 인과관계를 살려 올바르게

		르게 번역함.
원문	구어2	건강 없는 100세 인생 의미 없다.
번역	기계 번역①	ဗၼ် ၁၀၀၀၀၀၀ ကျန်းမာမှု မရှိ ကျန်းမာမှု (Life for 100 people without health is meaningless.)
	기계 번역②	ဗၼ် ၁၀၀ ဗၼ် ၁၀၀ ကျန်းမာမှု မရှိ ကျန်းမာမှု (A 100-year-age life is unhealthy, meaningful.)
	병렬 말뭉치	ဗၼ် ၁၀၀ ဗၼ် ၁၀၀ ကျန်းမာမှု မရှိ ကျန်းမာမှု (100 years of unhealthy life is meaningless.)
비교		기계 번역에서는 ‘100세’를 ‘100명’으로 잘못 번역함. 병렬 말뭉치에서는 ‘100세 인생’이라는 표현을 적절하게 번역함. 또한 기계 번역에서는 ‘의미 없다’라는 표현을 ‘의미 있다’라는 뜻으로 오역함. 병렬 말뭉치에서는 ‘의미 없다’라는 표현으로 올바르게 번역함.

<표 51> 기계 번역과 병렬 말뭉치의 번역 비교(필리핀 타갈로그어)

원문	문어1	다양한 장기 자랑을 선보이고 따뜻한 음색으로 노래를 부르는 아이들의 모습에 '동요 들으면서도 눈물을 흘릴 수가 있다'며 '감동적'이라는 시청자 반응이 이어졌다.
번역	기계 번역①	Ang mga bata, na nagpakita ng iba't ibang mga organo at kumanta ng mga kanta na may mainit na tono, ay sinundan ng isang tugon mula sa mga manonood, na nagsabi, "Maaari akong lumuha habang nakikinig sa mga rhymes ng nursery," at "ito ay napaka nakakaantig." (The children, who showed different organs and sang songs with warm tones, were followed by a response from the audience, who said, "I can cry while listening to nursery rhymes," and "it's very touching.")
	기계 번역②	Ang hitsura ng mga bata na umaawit sa isang mainit na tono, na nagpapakita ng iba't ibang pangmatagalang pagmamalaki, ay humantong sa isang "nakakaakit" na tugon ng madla, na sinasabi na maaari silang " umiyak habang nakikinig sa mga awit ng nursery." (The appearance of children singing in a warm tone, showing a variety of lasting pride, led to an "appealing" audience response, saying they could " cry while listening to nursery rhymes.")
	병렬 말뭉치	Nagpatuloy ang reaksiyon ng mga manonood na sinasabi nilang "nakakaantig", at "maaaari raw maiyak kahit na habang nakikinig pa sa pambatang awitin", sa mga anyo ng mga batang kumakanta sa madamdaming boses at nagpapakita ng iba't ibang mga talento. (Viewers continued to react to what they said was "touching," and "able to cry even while listening to children's

		songs," to the forms of children singing in passionate voices and showing different talents.)
<b>비교</b>		기계 번역에서는 ‘장기 자랑’을 ‘mga organo(organs)’ 또는 ‘pangmatagalang pagmamalaki(lasting pride)’로 오역함. 병렬 말뭉치에서는 ‘mga talento(talents)’로 문맥에 맞게 번역함.
<b>원문</b>	<b>문어2</b>	치열한 경쟁과 완고한 성적 지상주의는 학생들을 억누른다.
<b>번역</b>	<b>기계 번역①</b>	Ang mabangis na kumpetisyon at matigas ang ulo seksual na pangigibabaw ay pumipigil sa mga mag aaral. (The fierce competition and stubborn sexual dominance are holding back learners.)
	<b>기계 번역②</b>	Ang mabangis na kumpetisyon at matigas ang ulo na seksual na groundism ay pumipighati sa mga mag-aaral. (Fierce competition and stubborn sexual groundism oppress students.)
	<b>병렬 말뭉치</b>	Sumasakal sa mga estudyante ang malupit na kumpetisyon at mahigpit na hirarkiya sa akademiko. (Students are suffocated by fierce competition and strict academic hierarchies.)
<b>비교</b>		기계 번역에서는 ‘성적 지상주의’의 ‘성적’을 ‘sekswal(sexual)’로 오역하여 아예 다른 표현을 만들어냄. 병렬 말뭉치에서는 ‘akademiko(academic)’으로 적절하게 번역함.
<b>원문</b>	<b>구어1</b>	제 보폭이 넓어진 만큼 저와 제 주변에 대한 이해의 폭과 시야도 넓어진 것 같아요.
<b>번역</b>	<b>기계 번역①</b>	Habang lumalawak ang aking paghakbang, sa palagay ko ay lumawak ang aking pag unawa sa aking sarili at sa aking kapaligiran. (As I have grown, I have grown in my understanding of myself and my environment.)
	<b>기계 번역②</b>	Sa pagpapalawak ng aking bilis, ang lawak at perspektibo ng pag-unawa patungo sa akin at ang aking paligid ay lumalawak din. (As my speed expands, the breadth and perspective of understanding towards me and my surroundings also expands.)
	<b>병렬 말뭉치</b>	Habang lumalaki ang aking hakbang, parang ganun din po ang aking pang-unawa sa mga paligid ko. (As my step grew, so did my understanding of those around me.)
<b>비교</b>		물리적인 보폭이 넓어졌다는 표현을 기계 번역에서는 의역이 되어 ‘성장한 만큼’이나 ‘속도가 는 만큼’으로 의역함. 병렬 말뭉치에서는 ‘보폭이 넓어진 만큼’으로 올바르게 번역함.
<b>원문</b>	<b>구어2</b>	딸이 끼가 많아서 어릴 때부터 가수를 하고 싶다고 했는데, 그럴 때 마다 제가 강력하게 말렸거든요.

번역	기계 번역①	Sinabi sa akin ng anak ko na gusto niyang maging singer mula sa murang edad dahil marami siyang ngipin, ngunit sa tuwing gagawin ko iyon, mariin ko siyang pinanghihinaan ng loob. (My son told me that he wanted to be a singer from a young age because he had so many teeth, but every time I did that, I would strongly discourage him.)
	기계 번역②	Ang anak ko ay talagang talentado. lagi niyang sinabi na gusto niyang maging mang-aawit mula pa noong bata, pero napaka-hihirap ko siya ngayon. (My son is really talented. She used to say that she wanted to be a singer ever since she was a kid, but I've had a hard time with her now.)
	병렬 말뭉치	Dahil matalent po ang anak kong babae, kaya sinabi nyang gusto nya maging mang-aawit noong bata pa, kaso tuwing nagyayari ang ganun mariin ko po syang pinigilan. (Because my daughter is talented, she said she wanted to be a singer when she was young, but whenever that happens, I firmly stop her.)
비교		기계 번역에서는 ‘끼’를 ‘이(치아)’로 잘못 인식하여 번역함. 병렬 말뭉치에서는 끼가 많다는 것을 재능이 많다는 표현으로 적절하게 번역함. 또한 기계 번역에서는 ‘딸’을 ‘아들’로 잘못 번역하여 뒷 문장의 대명사 성별과 일치하지 않음. 병렬 말뭉치에서는 ‘딸’로 올바르게 번역함.

<표 52> 기계 번역과 병렬 말뭉치의 번역 비교(러시아어)

원문	문어1	이민기, 김지원, 손석구, 이엘은 '나의 해방일지'를 통해 '인생캐'를 다시 썼다.
번역	기계 번역①	Ли Мин Ги, Ким Джи Вон, Сон Сок-Гу и Иль переписали «жизненную катушку» через «Дневник моего освобождения». (Lee Min-gi, Kim Ji-won, Song Sok-gu, and Il Il rewrote the "lifeline" through My Liberation Diary.)
	기계 번역②	Ли Мин Ги, Ким Джи Вон, Сон Сук Гу и Эл переписали «Историю жизни» в «Дневнике моего освобождения». (Lee Min-gi, Kim Ji-won, Song Suk-gu, and El rewrote the "Life Story" in The Diary of My Liberation.)
	병렬 말뭉치	С помощью «Моего дневника освобождения» Ли Минги, Ким Дживон, Сон Сокку и Ли Эль переписали историю «своих лучших ролей». (With the help of "My Diary of Liberation", Lee Mingi, Kim Jiwon, Son Sokka, and Lee El rewrote the story of "their best roles".)

<b>비교</b>		‘인생캐’는 ‘인생 캐릭터’를 줄여서 말하는 신조어임. 기계 번역에서는 해당 단어를 ‘жизненную катушку(lifeline)’ 또는 ‘Историю жизни(Life Story)’로 오역하여 의미를 제대로 전달하지 못했지만, 병렬 말뭉치에서는 단어의 의미가 잘 전달되도록 번역함.
<b>원문</b>	<b>문어2</b>	특히 김시덕은 "빚투'라는 단어가 나오기 전부터 방송국에 모르는 사람들이 찾아와 돈을 갚으라고 요구했다'고 했다.
<b>번역</b>	<b>기계 번역①</b>	В частности, Ким Сидок заявил: «Еще до того, как появилось слово «долг», на радиостанцию приходили незнакомцы и просили вернуть деньги». (Specifically, Kim Si-deok stated, "Even before the word 'debt' existed, strangers would come to the radio station and ask for money back.")
	<b>기계 번역②</b>	В частности, Ким Си Дук сказал: «Еще до того, как прозвучало слово «долг», в участок пришли незнакомые люди и потребовали, чтобы я вернул деньги». (In particular, Kim Si-duk said, "Even before the word 'debt' was uttered, strangers came to the station and demanded that I return the money.")
	<b>병렬 말뭉치</b>	В частности, Ким Сидок сказал: «Еще до того, как появился термин "инвестиции путем занимания денег", в телевизионную компанию приходили незнакомые мне люди и требовали вернуть им деньги». (In particular, Kim Sidock said, "Even before the term "investment through money" appeared, strangers came to the television company and demanded their money back.")
<b>비교</b>		기계 번역에서는 빚내서 투자한다는 의미의 신조어 '빚투'를 'долг(debt)'으로 번역하여 정확한 의미가 누락되었음. 반면에 병렬 말뭉치에서는 내용을 살려 번역함.
<b>원문</b>	<b>구어1</b>	나 매운 거 먹으면 바로 직방이잖아.
<b>번역</b>	<b>기계 번역①</b>	Когда я ем острую пищу, это сразу к делу. (When I eat spicy food, it gets straight to the point.)
	<b>기계 번역②</b>	Если я ем острую пищу, то это зигбанг. (If I eat spicy food, it's a zigbang.)
	<b>병렬 말뭉치</b>	У меня же сразу будет реакция, как только съем острое. (I will have a reaction as soon as we eat spicy.)
<b>비교</b>		원문의 ‘직방’의 의미는 반응이 곧바로 온다는 뜻임. 기계 번역에서는 해당 단어를 그대로 음역하거나, 본론으로 간다는 의미로 오역하였으나, 병렬 말뭉치에서는 ‘어떤 결과나 효과가 지체 없이 곧바로 나타나는 일’의 의미로 잘 번역함.
<b>원문</b>	<b>구어2</b>	아, 근데 맨날 팬들이 데뷔하라고 그렇게 난리를 쳤는데 약간 조금 뒷북으로 데뷔하는 거 아니야?
<b>번역</b>	<b>기계 번역①</b>	Ах, но фанаты всегда так суетятся по поводу дебюта, но он не дебютирует ли он немного за кулисами?



		(Ah, but fans are always so fussed about the debut, but isn't he debuting a little behind the scenes?)
	<b>기계 번역②</b>	Но все время фанаты так рьяно пытались дебютировать, что-то вроде того, чтобы дебютировать немного позже. (But all the time, the fans were trying so hard to debut, something like debut a little later.)
	<b>병렬 말뭉치</b>	А, фанаты же вроде так требовали дебюта, получается, они дебютировали немного с опозданием? (And, fans seem so demanded a debut, it turns out, they debuted a little late?)
<b>비교</b>		기계 번역에서는 ‘팬이 데뷔하려고 노력했다’라는 의미로 전체 내용을 오역하여 상이한 내용으로 번역하거나, 늦었다는 의미의 ‘뒷북’을 위치상의 ‘뒤쪽’으로 오역함. 병렬 말뭉치에서는 원문의 맥락을 고려하여 ‘어떤 일이나 사태가 끝난 다음에 태도나 행동을 취함’의 의미로 자연스럽게 번역함.

<표 53> 기계 번역과 병렬 말뭉치의 번역 비교(우즈베크어)

<b>원문</b>	<b>문어1</b>	공주에 대한 고정 관념을 뛰어넘어 필 공주와 일곱 기사의 뜨거운 우정을 다룬다.
<b>번역</b>	<b>기계 번역①</b>	Bu malika haqidagi stereotiplardan oshib ketadi va malika Fil va yetti qahramon o'rtasidagi ehtirosli do'stlik bilan shug'ullanadi. (It goes beyond the stereotypes about the Queen and engages in a passionate friendship between the Queen Elephant and the Seven Heroes.)
	<b>기계 번역②</b>	Malika stereotipidan tashqariga chiqib, men Phil Princess bilan ettita maqolaning qizg'in do'stligi bilan shug'ullanaman. (Going beyond the princess stereotype, I'm going to have a passionate friendship of seven articles with Phil Princess.)
	<b>병렬 말뭉치</b>	Phil malikalar haqidagi stereotiplardan o'tib, malika va etti ritsar o'rtasidagi ehtirosli do'stlik haqida gapiradi. (Phil goes beyond the stereotypes about princesses and talks about the passionate friendship between the princess and the seven knights.)
<b>비교</b>		기계 번역에서는 ‘공주’를 ‘여왕’, ‘Queen Elephant’ 등으로 오역함. 병렬 말뭉치에서는 ‘공주’로 올바르게 번역함. 또한 기계 번역은 ‘기사’를 ‘knights’가 아닌 ‘articles’로 잘못 번역하여 문장의 의미가 전반적으로 잘못되었으나, 병렬 말뭉치에서는 적절하게 번역됨.
<b>원문</b>	<b>문어2</b>	야망은 크지만 능력은 부족하고, 자신감이 지나쳐 주변의 웃음을 유발하는 캐릭터는 이광수의 장기이기도 하다.
<b>번역</b>	<b>기계 번역①</b>	Katta orzu-umidga ega bo'lgan, ammo qobiliyati yo'q, o'ziga bo'lgan ishonchi haddan tashqari ko'pligi tufayli uning

		atrofida kulgiga sabab bo'ladigan personaj ham Lee Kvang-so'ning organi hisoblanadi. (Lee Kwang-so's body is also a character with great ambition but no ability, causing laughter around him due to his excessive self-confidence.)
	<b>기계 번역②</b>	Uning ambitsiyalari katta, ammo qobiliyati yo'q va o'ziga bo'lgan ishonchi tugagan qahramon ham atrofida kulgini qo'zg'atadi, Li Guangsooning organi. (His ambitions are high, but the character who lacks ability and lacks self-confidence also provokes laughter around him, Li Guangsoo's organ.)
	<b>병렬 말뭉치</b>	Katta ambitsiyaga ega, lekin qobiliyati oz va o'ziga haddan tashqari ishonadigan va atrofdagilarni kuldiradigan personajlar ham Li Kvang Suning eng yaxshi iste'dod va qobiliyati hisoblanadi. (Characters who have great ambition but little ability and are overconfident and make others laugh are also Lee Kwang Soo's best talents and abilities.)
<b>비교</b>		기계 번역에서는 '장기'를 '신체의 기관'으로 오역함. 병렬 말뭉치에서는 '가장 잘하는 재주'라는 문맥에 맞는 의미로 올바르게 번역함.
<b>원문</b>	<b>구어1</b>	17년도는 월드컵 결승에서 졌어.
<b>번역</b>	<b>기계 번역①</b>	'17-yilda jahon chempionati finalini yo'qotdik. (We lost the World Cup final in '17.)
	<b>기계 번역②</b>	2017 yilda jahon chempionati finalida mag'lub bo'lgandik. (In 2017, we lost in the final of the World Cup.)
	<b>병렬 말뭉치</b>	2017-yilda 'Afsonalar ligasi Jahon Chempionati'ning final o'yinida yutqazdi. (In 2017, he lost in the final game of the League of Legends World Championship.)
<b>비교</b>		기계 번역에서는 '월드컵'을 월드컵으로 오역함. 병렬 말뭉치에서는 게임 LoL(League of Legends)+World cup의 합성어인 '월드컵'의 의미를 잘 살려 번역함.
<b>원문</b>	<b>구어2</b>	여의도에서 여섯 시에 퇴근하면 진짜 장난 아니지 않음?
<b>번역</b>	<b>기계 번역①</b>	Yeouidoda soat 6 da ishdan ketish hazil emasmi? (Isn't it a joke to leave work at 6 in Yeouido?)
	<b>기계 번역②</b>	Yeouidoda soat oltida ishdan ketsangiz, haqiqiy hazil emasmi? (Isn't it a real joke if you leave work at six o'clock in Yeouido?)
	<b>병렬 말뭉치</b>	Yoidoda soat oltida ishdan chiqsa, dahshat emasmi-a? (Isn't it terrible if Yoido leaves work at six o'clock?)
<b>비교</b>		기계 번역에서는 '장난 아니지 않음?'을 'ketish hazil

	emasmi?(isn't it a joke to leave?)' 또는 'haqiqiy hazil emasmi?(isn't it a real joke?)'로 번역하여 '힘든 일이다'라는 의미를 살리지 못했지만, 병렬 말뭉치에서는 이 의미가 잘 전달되도록 번역함,
--	---

### 1.3. 데이터 품질 감리 결과

사업 수행 과정에서 감리 업체가 한국어-외국어 병렬 말뭉치의 품질을 점검하였다. 품질 점검은 1차와 2차로 나누어 총 2회를 실시하였다. 1차는 재검수한 데이터에서 언어별 6,000어절(총 48,000어절, 3,390문장) 점검을 실시하였고, 2차는 감수한 데이터에서 언어별 4,200어절(총 33,600어절, 2,504문장) 점검을 실시해 총 81,600어절(5,894문장) 품질 감리가 진행되었다. 감리 결과가 언어별 데이터의 적합률이 99.45~100%인 것으로 나타나 '고품질 언어 데이터 구축'이라는 사업 목표를 달성한 것을 확인할 수 있었다. 1차 품질 감리에서 적합률이 99.73%가 나왔으며, 2차에서는 99.77%가 나와 구축한 말뭉치의 품질이 적합 판정을 받았으며, 이를 합산하면 5,894문장 중 5,879문장이 적합 판정을 받아 전체적으로 99.75%의 적합률을 보였다. 데이터 품질 감리 결과의 상세한 내용은 다음과 같다.

<표 54> 데이터 품질 감리 결과

구분	1차			2차			전체	
언어	문장	적합	적합률	문장	적합	적합률	적합	적합률
베트남어	383	381	99.48%	262	262	100.00%	643	99.97%
인도네시아어	486	486	100.00%	342	341	99.71%	827	100.00%
태국어	451	449	99.56%	300	299	99.67%	748	99.91%
인도 힌디어	358	358	100.00%	268	268	100.00%	626	100.00%
캄보디아 크메르어	363	363	100.00%	366	366	100.00%	729	99.94%
필리핀 타갈로그어	365	363	99.45%	244	243	99.59%	606	99.84%
러시아어	438	436	99.54%	367	365	99.46%	801	99.91%
우즈베크어	546	545	99.82%	355	354	99.72%	899	99.97%
합계	3,390	3,381	99.73%	2,504	2,498	99.77%	5,879	99.75%

## 2. 대외 활동

### 2.1. 사업단 국제 심포지엄 개최

본 사업단은 ‘2023 한국어-외국어 병렬 말뭉치 구축 사업’의 일환으로 말뭉치 구축 관련 국내외 전문가들을 초청하여 국제 심포지엄을 개최하였다. 이번 국제 심포지엄은 ‘국립국어원 한국어-외국어 병렬 말뭉치의 활용과 응용’이라는 주제로 한국어-외국어 병렬 말뭉치 구축 사업을 대외에 널리 홍보하고 국내외 말뭉치의 활용 및 응용 사례를 공유함으로써 산업계 및 학계의 활용을 촉진하는 계기를 마련하고자 하였다.

#### 1) 오전 일정

국제 심포지엄은 장소원 원장(국립국어원)이 사업의 의의를 설명하고 국제 심포지엄을 시작하는 개회사로 시작되었다. 이후 인공지능 병렬 말뭉치 분야의 학계 전문가인 박진호 교수(서울대), 이도길 교수(고려대)가 각각 ‘한중 병렬 말뭉치를 이용한 대조문법 연구 시론: 중국어의 대응 표현이 한국어에 대해 시사하는 바를 찾아서’, ‘한국어 말뭉치 용례 검색기의 개발과 활용’이라는 주제로 주제 발표를 하였다. 그리고 정주연 연구사(국립국어원)가 ‘한국어-외국어 병렬 말뭉치 사업 소개 및 활용 현황’이라는 주제로 병렬 말뭉치 사업의 추진 배경과 목적, 말뭉치 구축량, 말뭉치 구축 절차, 말뭉치 활용 현황 및 방법 등 사업에 대한 전반적인 내용을 소개하였다.

<표 55> 국제 심포지엄 오전 식순

사회: 이정희 교수(경희대, 연구 책임자)	
09:00-09:10	개회사 장소원 원장 [국립국어원]
09:10-10:10	[주제 발표 ①] 한중 병렬 말뭉치를 이용한 대조문법 연구 시론: 중국어의 대응 표현이 한국어에 대해 시사하는 바를 찾아서 발표   박진호 교수 [서울대학교]
10:10-11:10	[주제 발표 ②] 말뭉치 용례 검색기의 개발과 활용 발표   이도길 교수 [고려대학교]
11:10-11:30	한국어-외국어 병렬 말뭉치 사업 소개 및 활용 현황 발표   정주연 연구사 [국립국어원]

11:30-13:00	점심시간
-------------	------

<표 56> 국제 심포지엄 주제 발표 내용

발표자	제목	내용
박진호 교수 (서울대)	한중 병렬 말뭉치를 이용한 대조문법 연구 시론: 중국어의 대응 표현이 한국어에 대해 시사하는 바를 찾아서	한국어는 풍부한 종결어미를 가지고 있는데, 이들 중 상당수는 양태, 증거성, 의외성 등 풍부한 의미성분을 지니고 있으며, 이들의 의미 차이를 명쾌하게 기술하기가 쉽지 않고, 학습자도 그 사용 양상을 체득하기가 어렵다. 한편 중국어에는 여러 문말어기사가 있는데, 이들 역시 미세한 의미 차이를 보여서 연구나 학습에서 중요한 난관이 된다. 이 둘을 따로 연구하기보다, 병렬 말뭉치를 이용하여 둘 사이의 대응 양상을 살펴보면, 이 둘의 의미 기술 및 의미 차이 파악에 유용한 단서들을 발견할 수 있다. 이러한 대응 양상을 살펴봄으로써, 한국어의 종결어미와 중국어의 문말어기사에 대한 새로운 통찰을 얻을 수 있다.
이도길 교수 (고려대)	한국어 말뭉치 용례 검색기의 개발과 활용	이 발표는 2014년부터 일반에 공개되고 최근 개선된 웹 기반의 말뭉치 분석 도구 (corpus.korea.ac.kr; 이하 분석 도구)에서 제공하는 기능 중 용례 검색기에 대한 개발과 활용 방법에 대해 소개한다. 현재 분석 도구에 수록된 말뭉치는 3종으로서 2000년대 신문 텍스트로 구성된 물결 21 코퍼스, 동아일보 코퍼스, 한국 근대 잡지 코퍼스이다. 이 발표에서는 분석 도구에 대한 간략한 소개와 더불어 단어 빈도, 공기어 분석 등의 세부 도구와 용례 검색기의 연동 방식, 용례 검색기에서 활용할 수 있는 검색 기능과 화면 구성, 원문 확인 등의 기능을 알아보고자 한다. 또한 용례 검색기의 개발을 위한 말뭉치 처리 과정과 용례 검색기에서 빠른 검색 결과를 얻기 위한 내부 색인 구조를 소개한다.

정주연 연구사 (국립국어원)	한국어-외국어 병렬 말뭉치 구축 사업 및 활용 현황 소개	한국어-외국어 병렬 말뭉치 구축 사업은 8개 언어(베트남어, 인도네시아어, 태국어, 필리핀 타갈로그어, 캄보디아 크메르어, 인도 힌디어, 러시아어, 우즈베크어)를 대상으로 한국어-외국어 병렬 말뭉치를 구축하는 것이다. 이 사업의 목표는 언어 연구 및 자동 통·번역 기술 등 기술 개발에 필요한 기반 자료를 확보하고 언어 자원의 불균형을 해소하며, 해당 국가와 한국 간 정치·경제·문화 교류를 확대하는 것이다. 2021년 9백만 어절, 2022년 10백만 어절을 구축하였고 2023년 11백만 어절을 구축 중에 있다. 2021년 1차 사업 데이터 공개 후 말뭉치 데이터 신청 건수는 426건이었는데 가장 신청이 많은 언어는 베트남어로 79건이었다. 신청 주체별로는 중소기업 143건, 대학교 138건, 개인 사용자 105건 순이었고 이외 대기업, 공공기관, 기타가 있었다.
--------------------	--	---

## 2) 오후 일정

국제 심포지엄 오후 일정은 패널 토의 및 산업계 활용 사례 발표와 언어별 병렬 말뭉치 활용 연구 발표로 이루어졌다. 먼저 이정희 교수(연구 책임자, 경희대)를 좌장으로 산·학 전문가인 김일환 교수(성신여대), 이정수 대표(㈜플리토), 김영택 부사장(㈜솔트룩스이노베이션), 김유석 대표(㈜시스트란)를 초청하여 ‘한국어-외국어 병렬 말뭉치의 활용과 응용’이라는 주제로 패널 토의를 진행하였다. 이번 패널 토의에서는 저자원 언어 데이터 부족으로 학계나 산업계에서 겪고 있는 문제와 이를 극복하기 위한 여러 노력들과 향후 병렬 말뭉치 구축 사업 시 고려 사항 및 병렬 말뭉치 활용의 제약 사항, 병렬 말뭉치 기반 용례 검색기의 활용 분야 및 개발 시 유의 사항 등에 대해 다각도로 논의하였다. 아울러 병렬 말뭉치를 활용한 AI 기술 개발 및 학계 연구의 동향 및 전망에 관한 산업계와 학계의 의견도 청취하였다.

산업계 활용 사례 발표는 이정수 대표(㈜플리토)와 김윤기 엔지니어(㈜업스테이지)가 각각 ‘21년 한·외 병렬 말뭉치 사업 데이터를 활용한 LLM 성능 향상’과 ‘업스테이지의 LLM 데이터 활용 사례’라는 주제로 발표를 하였다. 이정수 대표는 발표에서 초거대 언어 모델(Language Model, LLM)의 발전은 특히 저자원 언어에 있어 모델의 정확성을 향상시키기 위해서는 데이터 품질이 중요함을 강조하였다. 김윤기 엔지니어는 발표를 통해 (주)업스테이지의 LLM 데이터를 수집과 저장의 과

정 및 LLM 데이터의 품질을 높이기 위한 프로젝트 등에 대해 소개하였다.

언어별 병렬 말뭉치 활용 연구는 태국어, 인도 힌디어, 우즈베크어, 러시아어, 필리핀 타갈로그어 5개 언어의 국내외 학자들의 연구 발표와 토론으로 이루어졌다. 카첸 탄시리(Kachen Tansiri) 태국 쯔랄롱꼰대 시린톤태국어연구소 이사와 박경은 한국외대 교수는 ‘Linguistic Insights: Unveiling and Addressing Common Errors in Korean-Thai Translations - A Guided Approach to Post-editing Excellence’라는 주제로 한국어-태국어 번역에서의 오류의 유형을 분류하고 이를 개선하기 위한 접근 방식을 논의하였다. 쿠마르 스리잔(Kumar Srijan) 부산외대 교수와 뒤웨디 아난드 뿌라까쉬 샤르마(Dwivedi Anand Prakash Sharma) 델리대 교수는 ‘고유 명사 음역 기준의 중요성 및 한계점: 한국어-힌디어 병렬 말뭉치 번역에 나타난 고유 명사 음역을 중심으로’라는 주제로 한국어-힌디어 병렬 말뭉치 구축을 위해 마련된 한국어-힌디어(데바나가리 표기) 표기 지침의 타당성과 중요성 및 한계점에 대해 논의하였다. 갈라노바 딜노자(Kalanova Dilnoza) 호남대 교수는 ‘병렬 말뭉치를 활용한 한국어-우즈베크어 번역 양상 연구: 고유 명사를 중심으로’라는 주제로 2021년에 구축된 국립국어원 한국어-우즈베크어 병렬 말뭉치 데이터를 활용하여 고유 명사 번역을 분석함으로써 고유 명사 번역 방안을 제시하였다. 또한, 모졸 따지아나(Mozol Tatiana) 모스크바국립외대 교수와 마블레예바 다리아(Mavleeva Darya) 모스크바국립외대 교수는 ‘Discourse Markers of Korean and Russian Languages as Means of Mitigation’이라는 주제로 2021년에 구축된 국립국어원 한국어-러시아어 병렬 말뭉치를 활용하여 한국어 담화 표지 ‘혹시’, ‘좀’과 대응하는 러시아어 표현을 분석하였다. 알드린 리(Aldrin P. Lee) 필리핀국립대 교수는 ‘A Typology of Translation Errors in the Korean-Tagalog Parallel Corpus’라는 주제로 한국어-타갈로그어 병렬 말뭉치에서의 다양한 번역 오류 유형을 분석하였다.

<표 57> 국제 심포지엄 오후 식순

사회: 이정희 교수(경희대, 연구 책임자)	
13:00-14:00	[패널 토의] 한국어-외국어 병렬 말뭉치의 활용과 응용 좌장   이정희 교수 [경희대, 연구 책임자] 토론   김일환 교수 [성신여대], 이정수 대표 [(주)플리토], 김영택 부사장 [(주)솔트룩스이노베이션], 김유석 대표 [(주)시스트란]
14:00-14:30	21년 한-외 병렬 말뭉치 사업 데이터를 활용한 LLM 성능 향상 발표   이정수 대표 [(주)플리토]
14:30-15:00	업스테이지의 LLM 데이터 활용 사례

	발표   김윤기 [㈜업스테이지]
	사회: 이두용(고려대, 전임 연구원)
15:00-15:30	Linguistic Insights: Unveiling and Addressing Common Errors in Korean-Thai Translations - A Guided Approach to Post-editing Excellence 발표   카첸 탄시리(Kachen Tansiri) 이사 [쫄랄롱꼰대 시린톤태국어 연구소], 박경은 교수 [한국외대] 토론   시무앙 케와린(Simuang Kewalin) 교수 [한국외대]
15:30-16:00	고유 명사 음역 기준의 중요성 및 한계점: 한국어-힌디어 병렬 말뭉치 번역에 나타난 고유 명사 음역을 중심으로 발표   꾸마르 스리잔(Kumar Srijan) 교수 [부산외대], 뒤웨디 아난드 뿌라까쉬 샤르마(Dwivedi Anand Prakash Sharma) 교수 [델리대] 토론   이지현 교수 [한국외대]
16:00-16:30	병렬 말뭉치를 활용한 한국어-우즈베크어 번역 양상 연구: 고유 명사를 중심으로 발표   갈라노바 딜노자(Kalanova Dilnoza) 교수 [호남대] 토론   이지은 교수 [한국외대]
16:30-17:00	Discourse Markers of Korean and Russian Languages as Means of Mitigation 발표   모졸 따지아나(Mozol Tatiana) 교수 [모스크바국립외대], 마블레예바 다리아(Mavleeva Darya) 교수 [모스크바국립외대] 토론   박 카밀라(Pak Kamilla) 교수 [수원대]
17:00-17:30	A Typology of Translation Errors in the Korean-Tagalog Parallel Corpus 발표   알드린 리(Aldrin P. Lee) 교수 [필리핀국립대] 토론   쉐레이 디타(Shirley Dita) 교수 [라살대]

<표 58> 국제 심포지엄 패널 토의 및 개인 발표 내용

발표자	제목	내용
이정희 교수 (경희대, 연구 책임자), 김일환 교수 (성신여대), 이정수	한국어-외 국어 병렬 말뭉치의 활용과 응용	1) 저자원 언어 데이터 부족으로 학계나 산업계에서 겪고 있는 문제와 이를 극복하기 위한 노력 저자원 번역에 대한 수요는 증가하고 있지만 언어 데이터 부족으로 기계 번역 엔진을 만드는 데 어려움이 있다. 이를 극복하기 위해 유사 어군에 속하는 언어들을 대상으로 그룹 평러닝을 시도하거나 글로벌 네트워크를 통해 데이터 스와핑을 진행하고 있다. 2) 향후 병렬 말뭉치 구축 사업 시 고려 사항 산업계에서는 병렬 말뭉치 구축 사업이 오랫동안 지속되



<p>대표 (주)플리토), 김영택 부사장 (주)솔트룩 스이노베이 션), 김유석 대표 (주)시스템 란)</p>		<p>기를 바라지만 예산의 한계가 있어서 결국 효율성을 따져야 한다. 이를 위해 적정 구축 수량의 기준을 세우는 것이 중요한데 사용자(기업)가 사용할 만한 성능을 가진 기계 번역 엔진을 만들기 위해서 대략 200~500만 문장이 필요한 것으로 추산한다. 그리고 산업계의 수요 조사를 통해서 구축 언어를 선정하는 방법도 있다. 또한 원문의 품질이 번역 데이터 품질에 영향을 크게 미친다는 점과 도메인 특화된 말뭉치 구축이라든지 LLM 인공지능 성능에 영향을 크게 미치는 프롬프트 데이터 세트 구성도 고려해 볼 만하다.</p> <p>3) 병렬 말뭉치 활용의 제약 사항</p> <p>국가 공공 데이터로 구축된 병렬 말뭉치를 해외에서 무단으로 배포하거나 판매할 위험이 있어서 접근 방식의 제한은 필수적일 수밖에 없다. 해외의 무단 사용은 오히려 국내 산업계의 발전을 저하할 수 있는 위험이 있다. 학계에서는 JSON 파일로 배포되는 병렬 말뭉치를 활용하기 쉽지 않은데 활용도를 높이기 위해서는 웹 기반 용례 검색기 서비스를 제공해야 한다.</p> <p>4) 병렬 말뭉치 기반 용례 검색기의 활용 분야 및 개발 시 유의 사항</p> <p>다양한 검색 기능이 포함된 웹 기반 검색 도구가 필요하며, 검색 속도의 측면에서 DB 방식이 아닌 검색엔진을 사용해야 한다. 그리고 별도의 사용자 교육이 필요하지 않을 정도로 사용하기 쉬운 검색기를 만들었으면 좋겠다. 메타태깅의 정확성이 LLM 성능에 영향을 크게 미친다는 결과가 있는데 제대로 된 용례 검색기를 개발하기 위해서는 메타태깅이 중요하다.</p> <p>5) 병렬 말뭉치를 활용한 AI 기술 개발 및 학계 연구의 동향 및 전망</p> <p>산업계에서는 다양한 기계 번역 엔진을 통해 얻은 데이터를 대상으로 규칙 기반 방식을 통해 데이터의 우선순위를 정하고 이 데이터는 학습에 다시 사용되어 엔진 성능 개선에 기여하고 있다. 학계의 경우, 아직 병렬 말뭉치를 활용한 연구 성과가 많지 않지만 다른 연구자들의 참여를 유도하기 위해 소규모 연구 프로젝트를 활성화하는 노력도 필요하다.</p>
이정수	21년 한-외	초거대 언어 모델(Large Language Model, LLM)의 발전

대표 (㈜플리토)	병렬 말뭉치 사업 데이터를 활용한 LLM 성능 향상	은 특히 낮은 자원 언어에 있어 모델의 정확성을 향상시키기 위해 데이터 품질의 중요성을 강조한다. 고품질 데이터는 모델의 정확성 향상에 중요하며, 특히 낮은 자원 언어에서는 더욱 그 중요성이 부각된다. 예를 들어, 플리토의 우즈베크어 LLM 학습은 언어 처리 기술을 크게 향상시켰다. 전이 학습과 주기적인 평가는 모델 훈련에서 중요하며 특히 다국어 전략에서 효과적이다. 이러한 접근 방식은 우즈베크어에서 성능이 30% 향상된 것을 확인하였다. 그러나 영어 기반 LLM의 우세함은 영어 이외의 언어에 대한 자원 부족을 초래하므로 다양한 언어를 지원하는 고품질 다국어 데이터가 필요하다. 이는 한국어 LLM의 글로벌 확장 및 소규모 언어 모델의 개선에 큰 도움이 될 것이다.
김윤기 엔지니어 (㈜업스테이지)	업스테이지 의 LLM 데이터 활용 사례	업스테이지에서는 Data-Centric LLM 파트를 신설하였고 LLM 데이터를 수집하고 있다. 수집된 LLM 데이터를 업스테이지 내부 보안 규칙에 맞도록 분산하여 저장하고, 사용에 용이하도록 표준화하여 저장하고 있다. 또한, LLM 데이터의 품질을 높이기 위한 ETL 파이프라인 오픈소스 프로젝트도 진행 중에 있다. 해당 ETL 파이프라인을 통해 LLM 데이터를 손쉽게, 빠르게 전처리할 수 있을 뿐만 아니라, 다양한 종류의 데이터를 한눈에 볼 수 있도록 관리하는 데이터 현황 대시보드도 운영 중에 있다. 이러한 과정을 통해 얻은 노하우를 모두에게 공유함으로써 LLM 생태계를 구축하고자 한다.
Kachen Tansiri 이사 (쫄랄롱꼰 대 시린톤태국 어연구소), 박경은 교수 (한국외대)	Linguistic Insights: Unveiling and Addressing Common Errors in Korean- Thai Translation s - A Guided Approach to Post-editin g Excellence	이 연구는 한태 병렬 언어 말뭉치 자료의 분석을 통하여 한태 번역에서 빈번하게 관찰되는 다양한 오류의 특성을 Costa et al. (2015)이 제안한 분류 체계를 적용하여 ①맞춤법(철자 오류, 구두점 사용 오류 등) ②어휘(생략, 추가, 미번역 등) ③문법(문법 요소의 선택 오류와 어순 오류 등) ④의미(부적절한 단어 선택, 의미 혼동, 연어 및 관용어 사용 오류 등) ⑤담화(스타일, 다양성, 장르 등)의 다섯 가지 수준의 오류로 분류하고, 이를 바탕으로 포스트에디팅에서 참고할 수 있는 가이드라인을 제시하는 두 가지의 큰 목적을 가지고 있다. 연구의 결과는 향후 포스트에디팅의 속도와 효율성을 제고하고 체계성과 일관성을 갖춘 감수 시스템 구축에 기여할 수 있으며 나아가 AI 기반 번역 시스템의 발전에 일조할 수 있을 것으로 기대한다.

<p>꾸마르 스리잔 교수 (부산외대), Dwivedi Anand Prakash Sharma 교수 (델리대)</p>	<p>고유 명사 음역 기준의 중요성 및 한계점: 한국어-힌 디어 병렬 말뭉치 번역에 나타난 고유 명사 음역을 중심으로</p>	<p>이 연구의 목적은 한국어-힌디어 병렬 말뭉치 구축을 위해 마련된 한국어-힌디어(데바나가리 표기) 표기 지침의 타당성과 중요성 및 한계점에 대해 논의하는 것이다. 언어에 따라 다른 외국어와 비슷한 소리가 있는 반면에 다른 언어에서는 볼 수 없는 고유의 소리가 있다. 따라서 한 언어의 모든 소리를 다른 언어로 음역하는 데는 한계가 있다. 이러한 문제는 한국어와 외국어의 자모 체계의 음운적 특성 및 대조를 고려해 만든 한국어-외국어 표기법으로 다소 해결이 가능하다. 물론 힌디어에 없는 한국어 음소 표기는 한계점이 있지만 한국어-힌디어 표준 표기법이 없는 상황에서 본 연구에서 제시한 한국어-힌디어 표기 지침이 향후 한국어 고유 명사의 힌디어 표기법 및 번역 전략의 표준화를 위한 지침의 발판이 되기를 기대한다.</p>
<p>갈라노바 딜노자 교수 (호남대)</p>	<p>병렬 말뭉치를 활용한 한국어-우 즈베크어 번역 양상 연구: 고유 명사를 중심으로</p>	<p>이 연구에서는 2021년에 구축된 한국어-우즈베크어 병렬 말뭉치 데이터를 활용하여 고유 명사 번역을 분석함으로써 고유 명사 번역 방안을 제시하였다. 먼저 고유 명사 중에서 가장 많이 쓰이는 인명, 기관명과 상품명, 행정 구역명 등으로 분리하여 분석하였다. 인명 번역은 한국어와 우즈베크어의 공통적인 인명 규칙, 축 성과 이름의 순서에 근거하고 우즈베크어 고유 명사 음역 규범 중에 원래 언어 음역 규범을 따라 하고 있다. 기관명과 상품명 번역의 경우에는 공식 영어와 우즈베크어 관용 표기가 적용된다. 그리고 ‘서울’을 제외한 모든 행정구역명은 원래 언어 규범으로 번역된다. 본 연구는 한국어-우즈베크어의 고유 명사 번역 사전을 만드는 데 기초 자료가 될 것으로 기대된다.</p>
<p>Mozol Tatiana 교수 (모스크바 국립외대), Mavleeva Darya 교수 (모스크바 국립외대),</p>	<p>Discourse Markers of Korean and Russian Languages as Means of Mitigation</p>	<p>이 연구에서는 &lt;국립국어원 한국어-러시아어 병렬 말뭉치 2021&gt;을 활용하여 한국어 담화 표지 ‘혹시’, ‘좀’과 대응하는 러시아어 표현을 분석하였다. 그 결과, 러시아어보다 한국어에서 담화 표지(‘혹시’, ‘좀’)가 언어적 완화 장치로서 더 많이 사용되었으며 모든 유형의 화행에서 간접 전략의 사용도 더 빈번한 것으로 나타났다. 이로써 한국어와 러시아어에서 공손성을 표현하는 방식이 다르다는 사실을 확인하였다. 이러한 현상은 양 언어의 의사소통 문화가 가지는 성향(집단주의와 개인주의)과 담화 참여자 간 관계의 대칭성 등에서 나타나는 차이에 기인한 것으로 보인다.</p>



개회사(장소원 국립국어원장)



주제 발표 ①(박진호 교수)



주제 발표 ②(이도길 교수)



사업 소개(정주연 연구사)



패널 토의(이정희 교수, 김일환 교수, 이정수 대표, 김영택 대표, 김유석 대표)



청중석 전경



산업계 활용 사례 1(이정수 대표)



산업계 활용 사례 2(김윤기 엔지니어)



언어별 말뭉치 활용 연구 1 발표  
(Kachen Tansiri 이사)



언어별 말뭉치 활용 연구 1 토론  
(Simuang Kewalin 교수)



언어별 말뭉치 활용 연구 2 발표  
(Kumar Srijan 교수, D. A. P. Sharma 교수)



언어별 말뭉치 활용 연구 2 토론  
(이지현 교수)

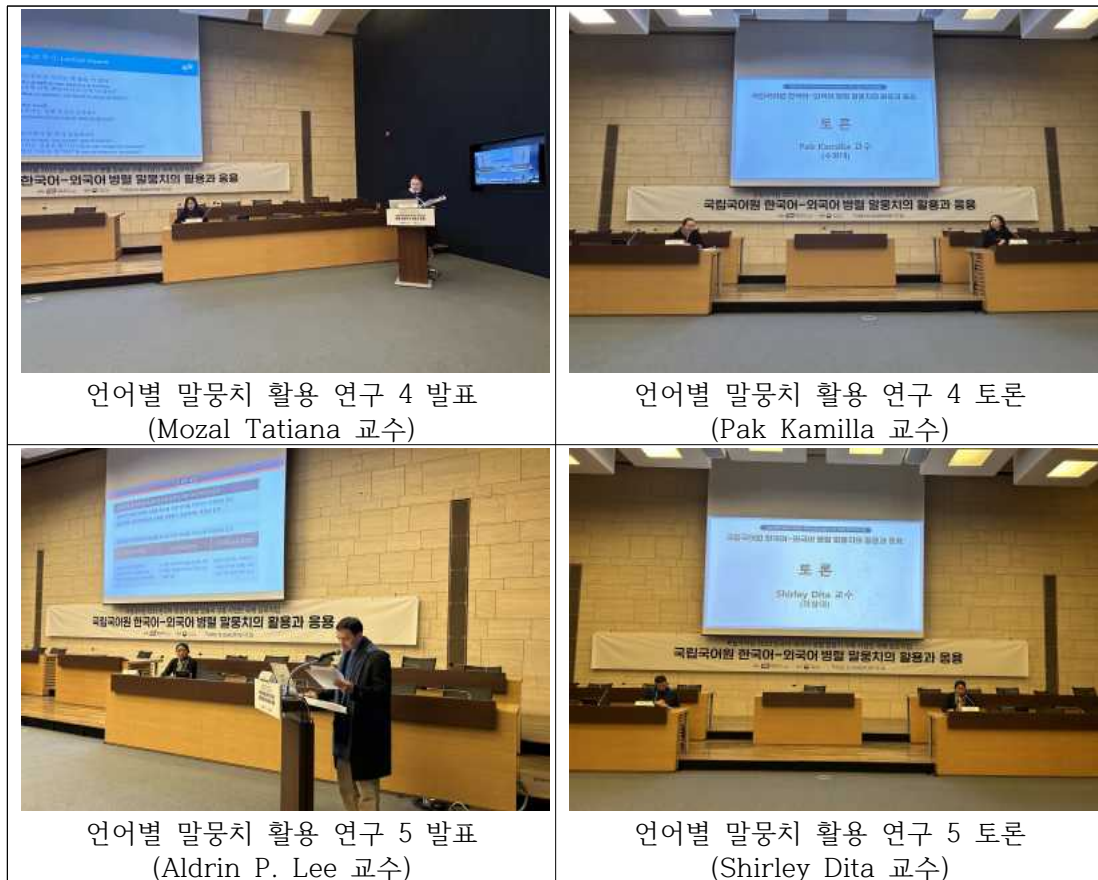


언어별 말뭉치 활용 연구 3 발표  
(Kalanova Dilnoza 교수)



언어별 말뭉치 활용 연구 3 토론  
(이지은 교수)





[그림 50] 국제 심포지엄 행사 사진

### 3) 성과

사업단은 국립국어원과 긴밀히 협력하고 여러 언론을 통해 국제 심포지엄을 적극적으로 홍보하였으며, 체계적인 기획과 운영을 바탕으로 국제 심포지엄을 성공적으로 개최하였다.

이로써 2021년부터 2022년, 2023년 3차에 걸쳐 구축한 고품질의 한국어-외국어 병렬 말뭉치를 국제 심포지엄을 통해 효과적으로 홍보하였다.

특히 국내외 전문가들을 초청하여 심포지엄을 개최함으로써 관련 분야의 해외 전문가들 간 교류의 기회를 마련함. 이를 통해 국제적인 협력과 지식 교환을 촉진하며 국립국어원의 사업 성과를 국제적으로 공유하는 계기가 되었다.

병렬 말뭉치를 이용한 대조문법 연구와 한국어 말뭉치 용례 검색기를 주제로 한 발표를 통해 관련 지식을 대중과 공유하고, 병렬 말뭉치를 학술 연구계의 활용 등을 촉진하는 계기를 마련함. 그리고 패널 토의 시간을 준비하여 병렬 말뭉치의 활용과 응용 방안에 대해 각계 전문가들이 진솔하고 심도 있게 논의하였다. 이를 통해 초거대 언어 모델(LLM) 시대에서 현재 구축 중인 병렬 말뭉치가 가지는 의의와

향후 발전 가능성을 확인할 수 있었다.

또 본 사업의 구축 언어 중 5개 언어의 언어별 활용 연구 발표와 토론을 통해 학술적인 교류와 지식을 공유하였다. 이를 통해 학술 연구에서의 병렬 말뭉치의 활용 가능성을 모색하고 학술 연구의 사례를 제시함으로써 다른 연구자들의 병렬 말뭉치 활용 연구를 유도하는 계기가 되었다.

주제 발표와 패널 토의를 통해 웹 기반 용례 검색기의 개발이 병렬 말뭉치의 활용도를 높이는 방안임을 확인하였으며, 만족도 조사 결과에서도 청중들의 관심과 요구가 많다는 사실을 파악하였다. 이로써 병렬 말뭉치를 활용한 용례 검색기 개발의 필요성에 대해 전문가와 대중의 공감대를 조성하는 기회가 되었다.

참가자들 간의 교류와 네트워킹을 통해 국내외에서의 다양한 협업 기회를 모색하고 이를 통해 심포지엄 이후에도 국제적인 협력의 가능성을 확장할 수 있을 것으로 기대한다.

## 2.2. 학술 논문 게재



한국어-외국어 병렬 말뭉치 구축 사업의 일환으로 관련 연구 결과를 학술지에 게재하였다. 특별히 이번 학술 성과로 연구 책임자 단독 저자로 국제한국어교육학회가 발행하는 KCI 등재지 『한국어교육』 34권 4호에 「한국어·외국어 병렬 말뭉치 구축의 쟁점-원문 정제를 중심으로」 제목의 논문을 게재하였다. 한국어·외국어 병렬 말뭉치의 구축 과정의 전반적인 개요와 고품질의 번역을 위해 병렬 말뭉치 구축 과정에서 실시했던 원문 정제의 원칙을 살펴보고 실제 예문들을 통해 어떻게 원문 정제가 이루어졌는지, 해결해야 할 문제는 무엇인지에 대해 논의하였다. 구체적으로 병렬 말뭉치 구축 대상 8개 언어인 베트남어, 인도네시아어, 태국어, 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어의 번역 품질을 향상하기 위해 기계적 정제, 비전문가 그룹 수준에서의 정제, 전문가 그룹 수준에서의 정제 등 세 단계의 정제 작업을 소개하였으며, 고품질 번역의 말뭉치를 위해 원문의 정확성과 규범성, 공공성과 윤리, 정보의 명확성, 번역의 용이성 기준을 기반한 원문 정제의 필요성을 주장하였다.

다음으로 본 사업의 참여자인 전임 연구원과 감수 교수가 주저자와 공동 저자로, 연구 책임자가 교신 저자로 참여하여 2편의 학술 논문을 게재하였다.

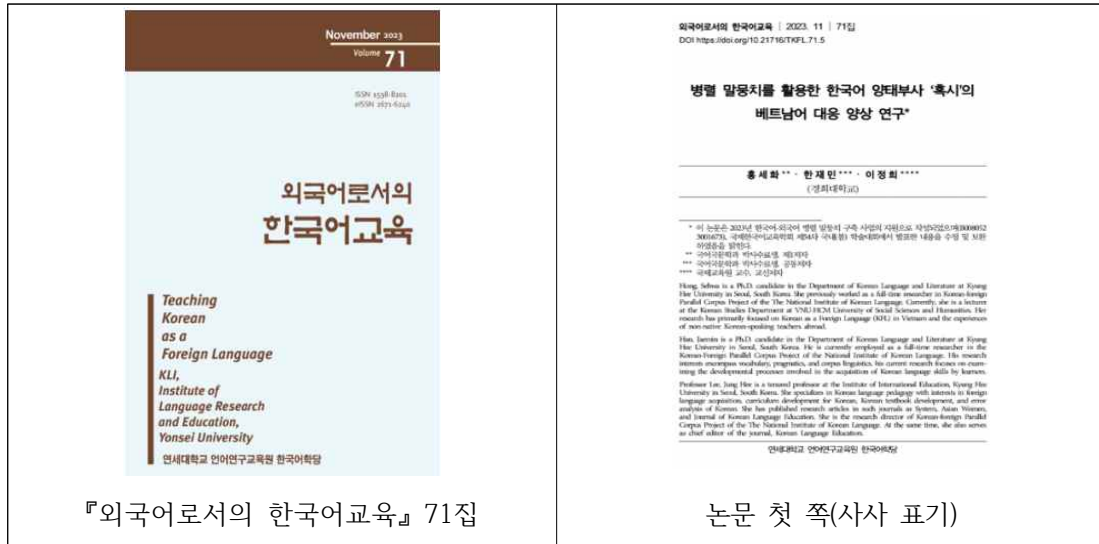
먼저, 「A Parallel Corpus-Based Comparative Analysis of Korean and Tagalog Negation」 제목으로 필리핀 감수 교수가 주저자로 참여한 논문이 『한국어교육』 34권 3호에 게재가 되었다. 한국어-외국어 병렬 말뭉치 구축 사업단의 1차 병렬 말뭉치 데이터를 기반으로 한국어와 필리핀 타갈로그어의 부정 표현의 대

조 연구에서 두 언어의 특정 문법을 비교하였는데, 언어 유형론적인 측면에서 의미 있는 연구이다. 또한 한국어 또는 타갈로그어를 외국어로 가르치는 데 있어서 교육적 문법 자료로 활용하는 데 객관적 정보를 제공해 줄 수 있다.

두 번째로, 연세대학교 언어교육원 한국어학당에서 발행하는 KCI 등재지 『외국어로서의 한국어교육』 71집에 논문을 게재하였다. 논문 제목은 「병렬 말뭉치를 활용한 한국어 양태부사 ‘혹시’의 베트남어 대응 양상 연구」이며, 이 연구 또한 한국어-외국어 병렬 말뭉치 구축 사업단의 1차 말뭉치 데이터를 사용하여 한국어의 ‘혹시’가 가지는 다양한 의미에 해당하는 베트남어 표현의 패턴을 분석하였다. 이 연구는 그동안 다루지 않았던 한국어 ‘혹시’와 베트남어 표현의 대응 양상을 병렬 말뭉치를 활용하여 연구 방법의 객관성을 확보하였다는 점과 베트남인 대상 한국어교육의 기초 자료로 활용될 수 있다는 데 연구의 의의를 찾아볼 수 있다.

 <p>『한국어교육』 34권 4호</p>	<p>한국어·외국어 병렬 말뭉치 구축의 쟁점 -원문 정제를 중심으로-</p> <p>이 경 희 (영희대학교)</p> <p>Lee, Jung Hye. 2023. An Issue of Korean-foreign Language Parallel Corpus Construction Focusing on Original Text Refinement. <i>Journal of Korean Language Education</i> 34-4, 313-337. This study introduces the project of the Korean-foreign language parallel corpus conducted by the National Institute of the Korean Language and examines the contents related to original text refinement as a process for advanced quality in corpus construction. The process of original text refinement is especially important in the process of constructing a parallel corpus on the premise of foreign language translation. Since the translator who must perform the translation is not a Korean but a native speaker of the target language, errors or mistakes in the original text are too difficult to guess and understand the context. Therefore, establishing the principles for original text refinement and having consistency can be a great help in enhancing the translation quality of the corpus. The target languages for parallel corpus construction are eight languages (Vietnamese, Indonesian, Thai, Hindi, Khmer, Tagalog, Russian, and Uzbek). In this research, three stages of refinement such as mechanical refinement, refinement at a non-expert group level and expert group level refinement were conducted to advance quality in translation. And the original texts were modified in terms of accuracy and normativity, in terms of politeness and ethics and in terms of clarity of information and ease of translation while the minimum scope of modification was made to avoid the damage of the meaning. In spoken Korean, however, it needs to be considered</p> <p>* 이 논문은 2023년 한국언어교육원 발행 발행의 구축 사업의 지원을 받아 발행된 연구이며 (ISSN0254-2687), 국립한국어교육원 제34권 4호를 국립한국어교육원 학회지인 『한국어교육』 34권 4호로 표제합니다.</p> <p>www.kci.go.kr</p> <p>논문 첫 쪽(사사 표기)</p>
 <p>『한국어교육』 34권 3호</p>	<p>A Parallel Corpus-Based Comparative Analysis of Korean and Tagalog Negation*</p> <p>Aldrin P. Lee · Ji Yeon Jeon · Jung Hye Lee<sup>†</sup> (University of the Philippines Diliman · Kangwon National University · Kyung Hee University)</p> <p>Lee, Aldrin P.-Jeon, Ji Yeon Jeon, Jung Hye. 2023. A Parallel Corpus-Based Comparative Analysis of Korean and Tagalog Negation. <i>Journal of Korean Language Education</i> 34-3, 201-237. The goal of this research is to compare the negative expressions in Korean and Tagalog based on the corpus data of the Korean-foreign Language Parallel Corpus Building Project Phase 1. First, the negative expressions are classified into different types based on previous research analyses on negative expressions in the two languages. Second, the special grammatical properties of the contextual negation in the two languages are compared and any restrictions, conditions and the like that could be attributed to any of these properties are discussed. Finally, preliminary research results on non-conventional negation or alternative methods of negative expressions that can be inferred from parallel corpus data are also shared with the intention of expanding and modifying the existing types of negative expressions in the two languages. Through this understanding, a more in-depth explanation becomes possible through the analysis of negative expressions based on</p> <p>* It is hereby disclosed that this article was written with the support of the Korean-foreign Language Parallel Corpus Construction Project in 2023, and that its contents are the revised and extended version of the manuscript presented at the 34th Local (Spring) Conference of the International Association for Korean Language Education.</p> <p><sup>†</sup> Aldrin P. Lee(lead author), Ji Yeon Jeon(senior author), Jung Hye Lee(corresponding author).</p> <p>논문 첫 쪽(사사 표기)</p>





[그림 51] 학술지 게재 논문

## 2.3. 국외 출장

한국어-외국어 병렬 말뭉치 구축 사업의 홍보와 국가 간 협력 도모를 위해 필리핀에 출장을 다녀왔다.

필리핀의 말뭉치 구축 사업 유관 기관 및 대학과 공공 기관을 방문하여 사업 홍보 및 상호 협력을 도모하고 필리핀 내 한국어 교육과 연구 및 한국어-필리핀 타갈로그어 번역 활용 산업 현황을 파악하고자 2023년 9월 3일부터 9월 7일까지 3박 5일의 일정으로 필리핀 마닐라에 출장을 다녀왔다.

### 2.3.1. 필리핀국립대학교 언어학과와 심포지엄 개최

필리핀국립대학교를 방문하여 언어학과와 공동으로 심포지엄을 개최하였다. 이정희 교수는 1차 사업의 데이터가 공개되었으며 관심 있는 사람들은 연구의 목적으로 다운받을 수 있음을 안내해 주었다. 본 사업단에서 구축하고 있는 말뭉치 대상 언어가 대부분 특수 외국어이기 때문에 필리핀 타갈로그어와 마찬가지로 자동 번역을 위한 데이터 양이 적음을 설명하였다. 특히 필리핀 타갈로그어의 번역과 검수 진행이 어려운 상황을 설명하였다.

본 사업의 한국어-필리핀 타갈로그어 감수 교수인 ALDRIN P. LEE(필리핀국립대학교 언어학과 교수)는 1차 병렬 말뭉치를 토대로 연구한 ‘한국어-필리핀 타갈로그어 부정 표현 대조 연구’를 발표하였다(2023.8. KCI 한국어 교육 34권 3호 게재).

필리핀국립대학교 학생을 대상으로 플리토에서 번역사를 선발하는 ‘번역 챌린지’에 선발된 필리핀국립대학교 재학생 및 졸업생 3명이 현재 플리토에서 활동하고 있다는 것을 소개하고, 한국어를 필리핀 타갈로그어로 퀄리티 높은 번역문으로 말뭉치를 구축하는 과정에서 한국어에 대한 이해 수준이 높은 인재가 필요함을 강조하였다.

또한 필리핀 국립대학교 학생들을 대상으로 말뭉치 구축에 참여할 수 있는 방법, 말뭉치를 다운받아 연구에 활용할 수 있는 실례를 보여 주었다.

### 2.3.2. 필리핀국립대학교 한국학 연구소 및 SWF 프로젝트 팀 면담

필리핀국립대학교 한국학 연구소 및 SWF(Sentro ng Wikang Filipino) 프로젝트 팀과 만나 각자 구축하고 있는 말뭉치의 구축 방식과 목표 등을 공유하였다. 필리핀국립대학교의 언어학과는 교수진 플랜 A, 플랜 B로 나뉘어 구성되어 있다. 플랜 A는 필리핀 언어에 집중해 다양한 필리핀어를 연구하는 교수진이고 플랜 B는 동남아시아 국가-한국, 일본, 중국, 태국, 인도네시아 언어를 모국어로 하면서 필리핀어와 연관시켜 가르칠 수 있는 교수진으로 구성되어 있다. 이와 같이 필리핀국립대학은 필리핀 내에서 다양한 언어들을 연구하고 보존하고자 하는 노력을 하는 동시에 필리핀과 밀접한 관계에 있는 나라의 언어 연구, 전공에도 매우 큰 관심과 노력을 기울이고 있다.

이어 SWF 프로젝트 참여 교수들을 만나 해당 사업에 대한 설명을 들었다. SWF 프로젝트는 음성 데이터를 전사하여 말뭉치를 구축한 것으로 ICE(International Corpus of English)를 기반으로 하였다. 프로젝트는 현재 진행 중이며, 해당 말뭉치와 관련하여 설명 자료를 온라인으로 공개하고 있다. SWF의 Director인 Jayson Petras는 본 사업의 원문 구축 방식, 주제 구성, 필리핀 타갈로그어로의 번역, 검수, 감수의 체계적인 방법에 대해 매우 고무적인 일이라고 하였으며 이후 지침 교환 등 협업을 할 수 있는 기회가 생기기를 기대한다고 하였다.

### 2.3.3. 필리핀국립대학교 총장 면담

필리핀국립대학교의 총장(Edgardo Carlo L. Vistan II(Chancellor University of the Philippines Diliman))을 만나 사업의 주관 기관인 국립국어원 소개 영상(영문)을 활용하여 국어원과 병렬 말뭉치 구축 사업을 소개하였다. 필리핀국립대학의 교수들이 본 사업에 참여하고 있으며 앞으로도 우수한 인재를 육성하기 위해 필리핀국립대학의 협조를 부탁했다. 2021년 사업부터 언어학과 알드리 리 교수(감

수 교수), 배경민 교수(검수팀장), 마리아 콘셉손 추아(검수팀장) 선생이 본 사업에 참여하여 큰 역할을 하고 있다는 것과 데이터 번역 품질 향상을 위해서는 두 언어에 능통한 전문가의 검수가 매우 중요하므로 이중 언어 능력을 갖춘 우수한 인재가 필요하다는 것을 강조하였다.

필리핀 내 한국어-필리핀 타갈로그어 번역 인력 양성 방안에 대해서도 논의하였다. 본 사업의 컨소시엄사인 플리토에서 필리핀국립대학교 학생들을 대상으로 ‘한국어 번역 챌린지’를 진행하였다. 참여했던 학생들 중 3명의 학생 James Dominic R. Manrique(학부 졸업생, 동 대학원 언어학과 재학 중), Sarah Eve Perlawan(학부 졸업생, 동 대학원 언어학과 재학 중), Darla Lorraine Abrera(학부 재학생)가 현재 플리토에서 번역에 참여하고 있음을 설명하였다.

Edgardo Carlo L.Vistan II 총장은 인력 양성 방안에 앞서 필리핀국립대학교 교환 학생 프로그램에 참여하는 한국 학생이 거의 없다는 것과 한국 내에 필리핀 타갈로그어를 가르치는 대학이나 단기 프로그램조차도 없다는 것을 지적하였다. 본 사업단에서 우수한 인재 양성의 중요성을 언급하는데 한국도 말뭉치 구축뿐만 아니라 필리핀 타갈로그어를 가르치는 기관이나 연구소, 프로그램 등을 운영한다면 한국에서 더 수월하게 우수한 인재를 영입할 수 있을 것이라고 하였다. 하지만 한국어-필리핀 타갈로그어 병렬 말뭉치를 구축하는 것은 매우 고무적인 일이고 이를 위해 필리핀국립대학에서는 학위나 비학위 과정으로 한국어-필리핀 타갈로그어, 필리핀 타갈로그어-한국어 번역 인재 양성 프로그램을 고민해 보겠다는 희망적인 메시지를 전하였다.

#### 2.3.4. KWF 원장 면담

Komisyon sa Wikang Filipino 원장(Arthur P. Casanova, PhD)을 만나 국립국어원을 소개하였고, 현재 KWF에서 진행 중인 말뭉치 연구원들을 대상으로 한-필 감수자 ALDRIN P. LEE 교수가 한국어-외국어 병렬 말뭉치 사업의 목적과 취지, 절차 등을 소개하였다. 필리핀어위원회(Komisyon sa Wikang Filipino, KWF)는 필리핀 언어정책에 관련된 전반적인 일을 하며, 필리핀의 철자를 통일하고, 어휘와 문법적인 속성을 정리하는 일을 해 왔다. 국립국어원의 말뭉치 프로젝트와 유사한 코퍼스 구축 프로젝트를 진행하고 있었다.<sup>4)</sup> ‘필리핀 타갈로그어’라는 표현에 대해 매우 난색을 표하고 현재 필리핀에서 표방하고 있는 언어 정책과 반하는 것으로 본 사업단에서도 필리핀 타갈로그어가 아닌, 필리핀어(FILIPINO)로 정식 명칭을 변경해 주기를 강력히 요청하였다. 이정희 교수는 한국어-외국어 병렬 말뭉

4) <https://kwf.gov.ph/gabay-ng-mamamayan/>

치 1차 데이터가 공개되었으며 사전 신청 후 승인을 받으면 무료로 데이터를 이용할 수 있다는 것을 설명하였고, 알렉산드라 원장도 필리핀어위원회가 현재 필리핀어 말뭉치를 5만 어절 구축하였으며 아직 공개는 하지 않았으나 진행 중이라고 설명하였다.

KWF 원장은 말뭉치 구축을 효율적으로 진행하기 위해 국립국어원과 필리핀어위원회 간 공식 협약이 가능할 것이라고 하였고 이를 통해 공식적으로 말뭉치 데이터 교환도 가능할 것이라고 하였다. 알렉산드라 원장은 KWF에서 진행하고 있는 말뭉치의 지침을 공유할 의향이 있으며 양국 말뭉치 구축에서 필요시 협조하겠다고 하였다.

### 2.3.5. 라살대 응용언어학과에서 진행한 말뭉치 활용 설명회

필리핀 라살대 응용언어학과 학과장(SHIRLEY N. DITA, Ph.D.(Chairperson))과 교수 및 학생, 필리핀 언어학회 임원들을 만나 본 사업을 설명하였다. 공개된 말뭉치를 다운받아 연구에 활용할 수 있다는 것을 알려 주고 다운받는 방법까지 안내하였다. 고품질 번역 데이터의 구축은 기계 번역 수준을 향상시켜 정치, 경제, 사회, 문화 각 분야의 국가 간 교류 협력에 필수적인 번역 업무를 효율화할 것이고 1차 사업 데이터가 공개되었으며 관심 있는 연구자는 다운로드가 가능하며, 본 말뭉치가 연구뿐만 아니라 다양한 분야에서 활용되기를 기대한다고 설명하였다.

사업이 계속해서 진행되고 있는 만큼 산업계에서의 활용 외에 학계에서의 선도적 활용이 매우 중요한 역할을 할 것이며, 이미 공개된 1차 한국어-필리핀 타갈로그어 병렬 말뭉치를 활용하여 연구하고 필리핀 타갈로그어나 관련 학술지에 투고할 수 있다는 것을 설명하였다. 필리핀언어학회 교수들이 사업단 유관 학술 대회에서 발표하는 것도 연구 활성화에 도움이 될 것이라고 설명하였고 실제로 알드리 교수와 전지연 전임 연구원은 KCI 저널에 두 편의 논문을 게재하였으며, 알드리 교수는 2023년 12월 8일 ‘3차 말뭉치 워크숍’에서 세 번째 연구를 발표할 예정이라고 설명하였다.

ALDRIN P. LEE 교수의 왕성한 연구 활동으로 응용언어학과 교수진과 필리핀언어학회 임원들은 이미 본 사업단의 말뭉치 구축에 대해서 사전 지식을 가지고 있었다. 한국에서 먼저 말뭉치 구축을 시작한 것에 대해 양국 관계 교류에 도움이 될 것이라고 하였고 본인들이 구축하고 있는 말뭉치와 비교·대조해 볼 수 있는 기회가 될 것이라고 하였다. 또한 말뭉치를 외부에 공개해 연구에 활용하기를 권장하는 분위기를 매우 반겼으며 학생들도 말뭉치 다운방법에 대해 문의하는 적극적인 모습을 보였다.

### 2.3.6. 현지 언론 인터뷰

마지막으로 <The Philippine News Agency>와 현지 언론사 2곳이 인터뷰를 진행하였다. 이정희 교수, 알드린 리 감수 교수, 배경민 팀장은 한국어-외국어 병렬 말뭉치 사업의 목적과 취지, 절차 등을 소개하며 필리핀 방문의 목적과 방문 기관 및 협력을 요하는 사항들에 대해 설명하였고 본 사업에 대한 기자의 질의에 답하였다. 현지 언론 3곳에서 총 4개의 기사(PNA<sup>5)</sup>, 스포츠 서울<sup>6)</sup>, 파이낸셜<sup>7)</sup>, 파이낸셜<sup>8)</sup>)가 보도되었다.



[그림 52] 필리핀 언론 보도 기사 일부 발췌



UP 심포지엄에서 발표하는 이정희 교수



UP 심포지엄 참석자들과 기념 촬영

- 5) <https://www.pna.gov.ph/articles/1209298>
- 6) <https://www.sportsseoul.com/news/read/1346699>
- 7) <https://www.thefinancialdistrict.com.ph/post/bridging-two-countries-through-korean-filipino-languages>
- 8) <https://www.thefinancialdistrict.com.ph/post/korean-government-funds-groundbreaking-language-project>





[그림 53] 필리핀 출장 일정별 사진

## 2.4. 대외 홍보

‘국립국어원 한국어-외국어 병렬 말뭉치의 활용과 응용’이라는 주제로 열린 사업단 국제 심포지엄과 관련하여 언론 홍보를 하였다. 언론 홍보를 통해 본 사업에 관한 내용을 외부에 널리 알리고 공유하였는데 그 결과 뉴시스, 매일경제 등 8개 언론사에 국제 심포지엄과 관련하여 기사가 게재되었다.

<표 59> 국제 심포지엄 기사 게재 언론

언론명(게재일)	제목
뉴시스 (23.12.4.)	'한국어-외국어 병렬 말뭉치의 구축 사업단' 심포지엄 개최
매일경제 (23.12.5.)	국립국어원 AI에 뛰어든 까닭...“병렬말뭉치 3000만 어절 구축 도전”
우리문화신문 (23.12.5.)	한국어 잘하는 K-챗지피티, 한국어 저자원 언어 기반
한국강사신문 (23.12.7.)	국립국어원 2023 한국어-외국어 병렬 말뭉치 구축 사업단 '국제 심포지엄' 개최
디지털데일리 (23.12.7.)	'한국어-외국어 병렬 말뭉치 구축 사업단 국제 심포지엄' 8일 개최
시타임스 (23.12.7.)	국립국어원, 한국어-외국어 말뭉치 구축 사업단 심포지엄 개최
뉴스와이어 (23.12.7.)	플리토 '한국어-외국어 병렬 말뭉치 구축 사업단 국제 심포지엄' 개최
메트로신문 (23.12.7.)	플리토, 한-외국어 병렬 말뭉치 구축 사업단 국제 심포지엄 개최



[그림 54] 국제 심포지엄 언론 보도 예시(뉴시스)

### 3. 활용 방안 및 기대 효과

#### 3.1. 병렬 말뭉치의 활용 방안

인공 지능 기술은 자연어 처리 기술의 수준과 함께 성장한다고 볼 때 한국어를 기반으로 하는 병렬 말뭉치 구축은 매우 중요한 사업이므로 향후 지속적인 양질의 병렬 말뭉치 구축이 필요하다. 일정 규모의 병렬 말뭉치가 구축될 때 산업, 학계, 교육 분야에 활용할 수 있는 방안은 다음과 같다.

##### 1) 고성능 언어 모델을 기반으로 기계 번역 시스템 개발

가장 대표적으로 병렬 말뭉치 구축을 통해 언어 모델을 만들어 문장을 번역하는데 활용할 수 있다. 특히, 딥러닝과 같은 인공 지능 기술을 이용하여 기계 번역 시스템을 개발하는 경우, 많은 양의 품질 좋은 병렬 말뭉치가 필수적이다.

실제로 2021년·2022년·2023년 사업에서 구축한 한국어-우즈베크어 병렬 말뭉치를 학습시켜 생성한 번역기(이하 ‘병렬 말뭉치 번역기’)의 성능을 국외 AI 기업의 번역기(이하 ‘A사 번역기’)와 대조하였다. 그 결과 아래와 같이 병렬 말뭉치 번역기의 BLEU 점수<sup>9)</sup>가 A사 번역기보다 높게 나와 한국어-외국어 병렬 말뭉치를 통해 기계 번역의 성능을 향상시킬 수 있다는 사실을 확인하였다.

<표 60> A사 번역기와 병렬 말뭉치 번역기의 BLEU 점수 비교(한→우)

구분	문장	BLEU 점수
한국어 ①	나도 우리 아빠의 아들인데 나도 나중에 이러면 어떡하나 하는 두려움도 있습니다.	
우즈베크어	Men ham dadamning o'g'liman, men ham keyinchalik shunday bo'lsam, nima qilaman, degan qo'rquv ham bor.	
A사 번역기	Men ham otamning o'g'liman, agar buni keyinroq qilsam nima qilishimdan qo'rqaman.	25.3
병렬 말뭉치 번역기	Men ham o'g'limning o'g'liman, lekin men ham keyinchalik bunday qilsam, nima bo'ladi, degan xavotir ham bor.	62.5

9) BLEU(Bilingual Evaluation Understudy)는 기계 번역기의 성능을 나타내는 대표적인 방법이다. BLEU 점수를 산출하는 방법은 기계 번역 문장과 인간 번역의 문장 간에 일치하는 단어의 개수를 계산하고 짧은 문장에 대한 페널티(Brevity Penalty), n-gram 조정 등을 통해 더 정확한 값이 나오도록 한다.



한국어 ②	이전에 모아 둔 돈도 거의 없습니다.	
우즈베크어	Bundan avval yeg'ib qo'ygan pulim ham deyarli yo'q.	
A사 번역기	Ilgari ozgina pul tejalgan.	0
병렬 말뭉치 번역기	Ilgari yig'ib qo'ygan pul ham deyarli yo'q.	28.9
한국어 ③	내 입장을 개가 이해할 수 있도록!	
우즈베크어	Mening ko'nglimni u tushunishi uchun!	
A사 번역기	U mening pozitsiyamni tushunishi uchun!	25.0
병렬 말뭉치 번역기	Mening ko'nglimni o'sha o'zi tushunishi uchun!	40.0
한국어 ④	그래서 이 음료를 완전 추천드려요!	
우즈베크어	Shuning uchun, bu ichimlikni juda maslahat beraman!	
A사 번역기	Shunday qilib, men ushbu ichimlikni juda tavsiya qilaman!	14.3
병렬 말뭉치 번역기	Shuning uchun mana bu ichimlikni juda tavsiya qilaman!	28.6

<표 61> A사 번역기와 병렬 말뭉치 번역기의 BLEU 점수 비교(우→한)

구분	문장	BLEU 점수
우즈베크어 ①	Qilishni istagan narsasi nimaligini o'rganib boradigan jarayon ko'proq kerak deb o'ylayman.	
한국어	하고 싶은 게 뭔지 배워 가는 과정이 더 필요하다고 생각해.	
A사 번역기	무엇을 하고 싶은지 알아가는 과정이라고 생각합니다.	29.4
병렬 말뭉치 번역기	마음에 드는 게 뭔지 배워 가는 과정이 더 필요하다고 생각해.	76.5
우즈베크어 ②	Yana kelgan mijozlar uchun bugun ham ish boshlaylik!	
한국어	또 와 주신 손님들을 위해 오늘도 영업 시작하자!	
A사 번역기	반복 고객을 위해 오늘 시작하겠습니다!	20.2
병렬 말뭉치 번역기	또 온 손님을 위해 오늘도 일을 시작하자!	42.9
우즈베크어 ③	Chindan ham hammasini tashlab, aylanishga yoki sayohatga borishni xohlayman.	

한국어	진짜 그냥 다 던지고 놀거나 여행 떠나 버리고 싶어.	
A사 번역기	나는 정말로 모든 것을 버리고 스페인이나 여행을 가고 싶습니다.	12.5
병렬 말뭉치 번역기	난 진짜 다 버리고 놀거나 여행 가고 싶어.	54.0

이러한 고성능의 기계 번역 시스템은 다국어 화상 회의, 번역 자동 평가 등 다양한 분야에 적용할 수 있을 것이다.

## 2) 기업 실무 및 비즈니스 운영에 적용

기업에서는 특정 문서에서 정보를 추출하는 기술인 정보 추출 모델을 개발하는데 병렬 말뭉치를 활용할 수 있다. 다양한 언어로 작성된 문서에서 정보를 추출할 수 있어 업무 자동화를 통한 생산성과 효율성을 높일 수 있다. 그리고 진출하고자 하는 국가의 언어로 된 병렬 말뭉치를 활용하여 오피니언 마이닝(opinion mining)이 가능한 감성 분석 모델을 개발함으로써 브랜드 인식 및 고객 반응을 파악할 수 있을 것이다.

## 3) 언어 및 외국어 번역 연구와 교육에 적용

교육 분야에서는 병렬 말뭉치 활용을 위한 다국어 용례 검색기 설계 및 개발하여 활용할 수 있다. 그리고 언어 및 외국어 번역 연구와 교육을 위한 기초 자료로 활용할 수 있으며 상세한 내용은 다음과 같다.

<표 62> 병렬 말뭉치의 교육 분야 활용 방안

언어 연구 및 교육	<ul style="list-style-type: none"> <li>- 한국어 학습용 챗봇, 모바일 학습 응용 프로그램, 용례 검색기 등 개발</li> <li>- 한국어 교재 개발(특히 현지 맞춤 교재)을 위한 언어 자료</li> <li>- 해외에서 한국어 수업·자가 학습용 보조재 제작을 위한 자료</li> <li>- 한국어와 외국어의 대조 연구를 위한 자료</li> <li>- 한국어 학습자의 중간 언어 및 오류 분석을 위한 분석 자료</li> <li>- 이민자 및 근로자 등의 사회 통합 교육을 위한 기초 자료</li> <li>- 이중 언어 사전 편찬에서 용례, 언어, 대역어 선정을 위한 자료</li> </ul>
외국어 번역 연구 및 교육	<ul style="list-style-type: none"> <li>- 통번역 전문가 양성을 위한 실무 교육 자료</li> <li>- 다양한 실무 현장에서 정확한 번역을 위한 참고 자료</li> </ul>

	<ul style="list-style-type: none"> <li>- 번역학에서 번역 텍스트의 특성 분석을 위한 자료</li> <li>- 자동 번역 등 기계 번역 분야 연구 자료</li> </ul>
--	--

### 3.2. 사업의 기대 효과

#### 1) 대규모·고품질의 병렬 말뭉치 구축을 통한 언어 데이터 산업의 기초 마련

국가 기관 주도로 이루어진 병렬 말뭉치 구축은 정부 기관 및 산업계, 연구 및 교육계 등 다양한 분야에서 활용될 수 있다. 예를 들어 공공 기관에서 활용 가능한 데이터베이스 구축으로 이어져 지역 관광 관련 신규 서비스 개발에 기여할 수 있다. 또한 전 세계적인 한국어 학습 수요 증대에 맞는 학습 자료 생성에도 활용할 수 있다.

#### 2) 다국어 언어 처리 및 인공지능(AI) 기반 통·번역 모델의 지속적인 품질 향상

AI 학습용 데이터 구축으로 인해 AI 기술이 발전되면 기계 번역의 정확도도 향상된다. 기계 번역의 높은 정확성으로 인해 학습용 데이터 구축 과정에서 데이터 번역의 자동화 비율을 늘릴 수 있다. 즉, AI 기술 발전이 그 발전에 토대가 되는 학습용 데이터 구축을 가속화하는 선순환 구조를 만들 수 있다.

AI 기술의 발전은 여행, 비즈니스 상담 등 일상생활에서 원활한 소통을 촉진하며 OTT(OTT) 및 케이팝 콘텐츠의 보급에 더해 신한류 콘텐츠의 확산에 기여할 것이다. 또한, 번역 품질 자동 예측 서비스와 결합하여 전문 번역사가 확인하지 않아도 품질이 우수한 번역 문장을 자동으로 필터링하고, 그 외 번역이 필요한 문장만 번역하도록 하는 사업화 모델에 적용할 수 있다.

### 3) 국가 정책 및 관련 업계의 수요 충족

국립국어원의 한국어-외국어 병렬 말뭉치 구축 사업은 영어나 중국어, 일본어 등에 비해 그간 미진했던 아세안-인도 6개 언어 및 유라시아 2개 언어를 대상으로 데이터를 구축했다는 점에서 그 의미가 크다. 이들 지역과 관련한 국가 정책 및 산업계·연구계의 수요에 효과적으로 대응할 수 있으며, 국가 간 교류를 활성화함으로써 상호 이해 증진 및 미래 지향적 상생의 경제 협력 기반을 조성할 것이다.

### 3.3. 제언

#### 1) 한국어-외국어 병렬 말뭉치의 데이터 구축 지속

전 세계적으로 초거대 언어 모델(LLM)과 챗GPT의 열풍이 불고 있는 상황에서 한국형 챗GPT 개발과 성능 향상 등 국내 인공지능 기술의 경쟁력 확보를 위해서는 대규모의 고품질 언어 데이터가 필요하다. 2021년과 2022년 사업에 이어 이번 사업에서도 다른 언어에 비해 큰 주목을 받지 못했던 8개 언어를 대상으로 대규모 데이터를 구축하였다. 하지만 인공지능 기반 기계 번역 시스템의 성능 향상에 유의미한 결과를 얻기 위해서는 한 언어당 최소 100~500만 문장(1,000~5,000만 어절, 평균 10어절로 계산)의 번역 쌍이 요구된다. 2023년 사업까지 산출한 데이터의 수량이 이 기준에 미치지 못하므로 후속 사업을 통해 데이터를 계속 구축해 나가야 한다.

#### 2) 원시 데이터로서 국가 주도로 구축한 언어 데이터의 지속 활용

고품질의 병렬 말뭉치를 구축하기 위해서는 번역뿐만 아니라 한국어 원시 데이터의 품질도 중요하다. 하지만 저작권 문제가 완전히 해결된 대규모의 원시 데이터를 수집하는 일은 쉽지 않다. 이러한 문제를 해결하는 방법으로는 국가 기관에서 구축한 단일 언어 데이터를 병렬 말뭉치 구축에 활용하는 것이다. 지난 두 차례의 사업과 달리 2023년에는 국립국어원에서 구축한 문·구어 말뭉치를 원시 데이터로 활용함으로써 수집 시간 단축과 비용 절감, 원문 데이터의 품질 보장이라는 효과를 얻었다. 이러한 원시 데이터 수집 방식은 향후 사업에서도 계속 이어져야 할 것이다.

### 3) 웹 기반 병렬 말뭉치 용례 검색기 개발

현재 한국어-외국어 병렬 말뭉치는 기계 처리 및 분석에 용이한 JSON 형식으로 ‘모두의 말뭉치’에서 배포되고 있다. JSON은 이용자가 관련 지식이 없다면 접근하고 활용하는 데에 상당한 어려움이 있다. 기존에 국립국어원에서 구축한 <21세기 세종계획>의 ‘세종 말뭉치’나 ‘한국어 학습자 말뭉치’는 웹 기반 용례 검색 서비스를 제공하여 일반인들도 쉽게 접근할 수 있었다. 이번 ‘사업에서 추가 제안 사항으로 웹 기반 병렬 말뭉치 용례 검색기의 프로토타입을 개발하여 용례 검색기의 개발 가능성과 효용성을 확인하였다. 그리고 국제 심포지엄을 통하여 사용자 친화적 이면서 검색 속도가 빠른 웹 기반 용례 검색기의 필요성을 파악하였다. 본 사업에서 구축한 병렬 말뭉치를 활용하여 웹 기반 용례 검색기를 본격적으로 개발하고 운영한다면 교육 및 연구 분야 내 병렬 말뭉치의 활용도를 올릴 수 있을 것이다.

### 4) 말뭉치 교육 워크숍 운영

웹 기반 용례 검색기 개발과 더불어 말뭉치 활용 교육을 위한 워크숍 운영을 통해서도 본 사업에서 구축한 데이터의 활용도를 높일 수 있다. 앞서 언급한 바와 같이 일반 이용자가 JSON 형식을 자신의 목적에 맞게 가공하여 활용하기는 쉽지 않다. 따라서 말뭉치와 관련하여 이용자들의 요구 사항을 사전에 조사하고 이를 바탕으로 체계적인 교육 프로그램을 구성한다면 이용자들이 말뭉치를 활용할 때 겪는 불편을 해소할 수 있다. 또한, 온·오프라인 실습 워크숍이나 단기간 집중 코스 등 워크숍 형태를 다양화함으로써 이용자의 요구를 다각도로 충족시키고 교육 효과도 제고할 수 있을 것이다.



## <Abstract>

# 2023 Korean-Foreign Language Parallel Corpus Construction

Following the “2021 Korean-Foreign Language Parallel Corpus Construction Project” and “2022 Korean-Foreign Language Parallel Corpus Construction Project,” this project aims to construct Korean-foreign language parallel corpus with languages of countries drawing attention as Korea's new exchange partners and formulate ways to utilize the constructed corpus. The target languages for parallel corpus constructions are six languages in the ASEAN-India region (Vietnamese, Indonesian, Thai, Hindi, Khmer, and Tagalog) and two languages in the Eurasian region (Russian and Uzbek).

The construction of parallel corpus was largely carried out in three stages: “collection,” “construction,” and “production.” In the “collection” stage, historical Korean literary and colloquial texts were collected, and the collected data was proofread and refined. The collected and proofread data was translated in the “construction” stage, and the translation was edited by an editor and proofread primarily by an internal language expert(linguist) of Flitto Inc.(5%). The International Association for Korean Language Education then carried out a secondary proofreading(100%). The sentences that went through the secondary proofreading were then put through a third proofreading conducted by the proofreading leader(20%) and supervision conducted by the final supervisor(10%). The translation that verified its quality through the above steps, along with meta-data, was finalized as the final data in the “production” stage.

As a result, a total of 11,045,120 words (based on the original Korean) of the Korean-foreign language parallel corpus and the additional requested 1,380,640 words (based on the original Korean) of Korean-English parallel corpus were constructed.

Next, a parallel corpus example search engine prototype was suggested as a way to utilize the parallel corpus. By developing a web-based example search engine prototype using parallel corpus, the project aimed to lower the entry barrier for corpus usage and increase its availability in language and foreign language education and research.

Furthermore, an international symposium was held to share knowledge about parallel corpus and artificial intelligence (AI) technology and spread project's results externally. Additionally, through overseas trips, high interest in the construction of Korean-foreign language parallel corpus was confirmed, and the foundation for inter-agency cooperation networks was established. Moreover, the results from the research were presented at an academic conference and written in an academic journal to promote the revitalization of related research.

The expected effects of this project are as follows:

First, through the construction of large-scale and high-quality parallel corpus, the foundation of the language data industry can be established.

Second, multilingual language processing and artificial intelligence (AI) based interpretation/translation models will continuously improve their quality.

Third, the demands of national policies and related industries can be met.

The Korean-foreign language parallel corpus constructed from this project is not only valuable as is but is also expected to serve as the basic data that can be used in various fields.

**Keywords:** Korean-foreign Parallel Corpus, Parallel Corpus Example Search Engine, Vietnamese, Indonesian, Thai, Hindi, Khmer, Tagalog, Russian, Uzbek



<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 정주연 학예연구사

국립국어원 강정미 연구원

<사업 참여자>

연구 기관 2023년 한국어-외국어 병렬 말뭉치 구축 사업단  
(사)국제한국어교육학회, (주)플리토

연구 책임자 이정희

연구 참여자 김일환, 김종민, 박진욱, 이동규, 이동은, 이수미,  
이영준, 임채훈, 조남호, 최문석, 최홍열, 김연희,  
김영근, 문진숙, 박지민, 윤세윤, 이두용, 전지연,  
정성호, 지화숙, 한재민, 국혜민, 김한별, 박광길,  
서유리, 이상후, 이요셉, 이혜민, 최예린, 이정수,  
강동한, 김진구, 최승미, 김재훈, 이제영, 김이주

---

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금낭화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2023년 12월 29일

발행일: 2023년 12월 29일

인 쇄: 해림복사

---

※ 이 보고서는 국립국어원의 국고 보조금으로 수행한 ‘2023년 한국어-외국어 병렬 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.