

국립국어원 2022-01-06

발간등록번호

11-1371028-000896-01

2021년 한국어-외국어 병렬 말뭉치 구축 사업

연구 책임자
이 정 희

제 출 문

국립국어원장 귀하

국립국어원의 국고 보조금으로 수행한 ‘2021년 한국어-외국어 병렬 말뭉치 구축 사업’의
결과 보고서를 작성하여 제출합니다.

■ 사업 기간: 2021년 8월 ~ 2022년 5월

2022년 5월 26일

연구 책임자: 이정희(경희대학교)

연구 기관	21 한국어-외국어 병렬 말뭉치 구축 사업단 (사)국제한국어교육학회, (주)플리토)
연구 책임자	이정희
연구 참여자	이동은, 김종민, 박진욱, 손혜진, 이수미, 이영준, 임채훈, 최문석, 최홍열, 권상희, 김연희, 문진숙, 박지민, 이두용, 전지연, 정성호, 지화숙, 한재민, 홍세화, 국혜민, 김수희, 김인후, 김한별, 박민선, 박소영, 박정모, 박종우, 서유리, 심하은, 오승혜, 이규민, 이상후, 이요셉, 이재림, 이종현, 임예찬, 조영민, 주우찬, 최지선, 이정수, 유동현, 강동한, 김진구, 이서현, 최선재

연구 책임자	이정희(경희대학교)
	이동은(국민대학교)
	김종민(경희대학교)
	박진욱(대구가톨릭대학교)
	손혜진(고려대학교)
	이수미(성균관대학교)
	이영준(한국학중앙연구원)
	임채훈(숭실대학교)
	최문석(경희대학교)
	최홍열(강원대학교)
	권상희(홍익대학교)
	김연희(경희대학교)
	문진숙(경희대학교)
	연구 참여자
연구 참여자	박지민(경희대학교)
	이두용(고려대학교)
	전지연(강원대학교)
	정성호(경희대학교)
	지화숙(경희대학교)
	한재민(경희대학교)
	홍세화(경희대학교)
	국혜민(숭실대학교)
	김수희(경북대학교)
	김인후(경희대학교)
	김한별(대구가톨릭대학교)
	박민선(한국외국어대학교)
	박소영(성신여자대학교)

	박정모(강원대학교)
	박종우(성신여자대학교)
	서유리(경희대학교)
	심하은(성균관대학교)
	오승혜(국민대학교)
	이규민(고려대학교)
	이상후(고려대학교)
	이요셉(경희대학교)
	이재림(국민대학교)
	이종현(강원대학교)
	임예찬(경희대학교)
	조영민(경희대학교)
	주우찬(고려대학교)
	최지선(성신여자대학교)
	이정수(㈜플리토)
	유동현(㈜플리토)
	강동한(㈜플리토)
	김진구(㈜플리토)
	이서현(㈜플리토)
	최선재(㈜플리토)

2021년 한국어-외국어 병렬 말뭉치 구축 사업

이 사업은 대한민국의 새로운 협력 국가로 주목 받고 있는 8개국과의 문화 교류를 위한 첫걸음으로서 한국어와 8개 외국어의 병렬 말뭉치를 구축하는 것을 목적으로 하였다. 병렬 말뭉치 구축 대상 언어는 아세안-인도 지역의 6개 언어(베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어)와 유라시아 지역의 2개 언어(러시아어, 우즈베크어)이며, 한국어-외국어 병렬 말뭉치 이외에 시범적으로 외국어(베트남어)-한국어 병렬 말뭉치를 구축하였다.

병렬 말뭉치 구축은 크게 ‘수집’, ‘구축’, ‘정제’, ‘산출’의 네 단계로 진행되었다. ‘수집’ 단계에서는 한국어와 베트남어 문·구어를 수집하고, 수집한 데이터를 검수·정제하였다. 이렇게 정제된 데이터는 ‘구축’ 단계에서 사업단에서 작성한 번역 지침에 따라 1차 번역이 이루어졌으며, 1차 번역문은 현지 상위 검수자를 통한 1차 전수 검수, (주)플리토 내부 언어 전문가(linguist)를 통한 2차 검수를 거쳤다. 2차에 걸친 검수가 끝난 번역문은 ‘정제’ 단계를 통해 전수 검수의 과정을 거쳤다. 검수를 거친 번역문은 다시 최종 검수자가 10% 검수를 진행하였다. 이상의 단계를 거쳐 품질이 검증된 번역문은 ‘산출’ 단계에서 최종 데이터로서 산출되었다.

이를 통해 한국어-외국어 병렬 말뭉치 총 8,400,096어절, 외국어(베트남어)-한국어 병렬 말뭉치 52,500문장, 추가로 한국어-영어 병렬 말뭉치 1,050,000어절을 구축하였다.

이 사업을 통해 구축된 병렬 말뭉치 구축 사업의 기대 효과는 다음과 같다.

첫째, 관련 산업계뿐만 아니라 연구계, 교육계 등 다양한 분야에서 활용될 수 있으며, 특히 그간 미진했던 8개 언어 관련 연구의 기초 자료로서 활용될 수 있을 것이다.

둘째, 기계번역 분야에서 비교적 학습이 덜 되었던 아세안-인도, 유라시아 언어에 대한 학습이 이루어질 수 있으며 다양한 유형의 언어 학습을 통해 기계 번역의 기술과 품질이 향상될 수 있을 것이다.

셋째, 한국과 해당 지역의 교류를 활성화함으로써 상호 이해 증진 및 미래지향적 상생의 경제 협력 기반을 조성할 수 있을 것이다.

넷째, 학습 및 번역 소프트웨어 개발 등을 통한 언어 학습 및 통·번역 관련 분야를 활성화시킬 수 있을 것이다.

또한 본 사업단에서는 사업의 성과를 대외적으로 확산시키기 위해 번역·검수 워크숍을 실시하였으며 언론 홍보를 진행하였다. 특히 번역·검수 워크숍에는 200명 이상이 참가하여 한국어-외국어 병렬 말뭉치 구축에 대한 높은 관심을 확인하고 인적 네트워크를 구축하였다. 아울러 사업 수행 과정에서 연구된 결과물을 학술대회에 발표하여 사업의 중요성을 알리고 관련 연구의 활성화를 도모하였다.

이러한 목표와 과정으로 구축된 한국어-외국어 병렬 말뭉치는 그 자체로서 가치가 있을 뿐만 아니라 다양한 분야에 활용될 수 있는 기초 도구로서의 역할을 할 수 있을 것으로 기대된다.

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

차 례

제1장 사업 개요

1. 사업의 목표 및 중요성	2
1.1. 사업의 목표	2
1.2. 사업의 중요성	3
2. 사업의 범위	6
3. 사업 수행 절차 및 체계	7
3.1. 사업 수행 절차	7
3.2. 사업 수행 체계	9
4. 사업 수행 내용	12
4.1. 주요 진행 사항 및 조치 사항	12
4.2. 주요 변경 사항	15

제2장 사업 수행

1. 원문 구축 및 정제	18
1.1. 원문 구축	18
1.2. 원문 정제	42
1.3. 개인 정보 비식별화	62

차 례

2. 번역 지침 및 교육	65
2.1. 번역 지침 제작 과정	65
2.2. 교육 영상 제작	81
3. 번역	91
3.1. 번역 업체 선정 과정	91
3.2. 번역 인력 교육	95
3.3. 번역 절차	97
3.4. 품질 향상 활동	99
4. 번역 검수 및 감수	102
4.1. 번역 검수 인력 관련	102
4.2. 최종 감수 인력 관련	104
4.3. 번역 검수·감수 인력 교육	105
4.4. 품질 향상 활동 예시	115
5. 병렬 말뭉치 구축 및 메타 정보	119
5.1. JSON 포맷	119
5.2. 최종 데이터 구조	122
5.3. JSON 예시	123
6. 보안 관리	129
6.1. 보안 관리 대상	129
6.2. 담당자 지정 체계	129
6.3. 담당자 역할 체계	129
6.4. 보안 관리 방법	130

차 례

제3장 사업 수행 결과

1. 말뚝치 데이터 구축 결과	137
1.1. 최종 구축 데이터	137
1.2. 데이터 품질 감리 결과	139
2. 대외 활동	140
2.1. 워크숍	140
2.2. 학술대회 발표	148
2.3. 대외 홍보	150
3. 활용 방안 및 기대 효과	153
3.1. 병렬 말뚝치의 활용 방안	153
3.2. 사업의 기대 효과	154
3.3. 제언	155
참고자료	157
부록	158
영문 초록	175

표 차 례

<표 1> 언어별 병렬 말뭉치 구축의 중요성	3
<표 2> 사업 수행 절차 상세 내용	7
<표 3> 참여 인력 업무 분장표	9
<표 4> 원문 구축 및 수집 목표 수량	18
<표 5> 원문 구축 및 수집 목표 수량	19
<표 6> 원문 선정 이유	20
<표 7> 원문 매체의 구성	20
<표 8> 번역과 감수에 관한 저작권 양도계약서-(사)국제한국어교육학회 ..	22
<표 9> 번역과 감수에 관한 저작권 양도계약서 - (주)플리토	22
<표 10> 저작권 이용 허락 계약서 - (주)아주뉴스코퍼레이션	23
<표 11> 저작권 이용 허락 계약서 - (주)웨이커	23
<표 12> 저작권 이용 허락 계약서 - (주)플리토	24
<표 13> 저작권 이용 허락 계약서 - 시나리오 작가	24
<표 14> 저작권 확보한 최종 데이터 수량	25
<표 15> 시나리오 구매 목록	32
<표 16> 한국어 신문 기사 구매 목록	33
<표 17> 베트남어 신문 기사 구매 목록	38
<표 18> 한국어 원문 예시	40
<표 19> 베트남어 원문 예시	41
<표 20> 한국어 원문별 정제 절차	42
<표 21> 한국어 구어체 기계적 정제 내용	43
<표 22> 2차 일반인 수준 한국어 구어체 원문 수정 예시	48
<표 23> 자주 발생하는 한국어 정제 오류 예시	49
<표 24> 전문가 수준 한국어 구어체 원문 정제 예시	49
<표 25> 한국어 문어체 기계적 정제 내용	50
<표 26> 2차 일반인 수준 한국어 문어체 원문 수정 예시	53
<표 27> 전문가 수준 한국어 문어체 원문 정제 예시	54
<표 28> 베트남어 원문별 정제 절차	57
<표 29> 베트남어 구어체 기계적 정제 내용	57
<표 30> 개인 정보 비식별화 기준	62
<표 31> 개인 정보 비식별화 예시 문장	63
<표 32> 개인 정보 비식별화 이후 재정제 예시	64
<표 33> 언어별 특징 및 번역상 오류 예상 요소 및 주의사항	68

표 차례

<표 34> 지침 교육 영상 내용 및 담당자	83
<표 35> 번역·검수 작업자의 주요 업무 목록	92
<표 36> 작업 관리자의 주요 업무 목록	100
<표 37> 기계적 검수 정제 내용	100
<표 38> 구조적 검증 기준과 불합격(Fail) 상황	101
<표 39> 영어 데이터 품질 향상 진행 내용	101
<표 40> 해외 거주자 대상 해외 송금 처리 과정	102
<표 41> 번역 검수 인력	103
<표 42> 최종 감수 인력	104
<표 43> 공통 교육 내용	106
<표 44> 언어별 번역 검수 세부 지침 내용	107
<표 45> 품질 관리(QC) 유형별 검수 방법	108
<표 46> 오류 유형의 기준	109
<표 47> 언어별 번역 검수 오류 유형 결과	109
<표 48> 언어별 오류 원인	110
<표 49> 메타데이터(Metadata)의 JSON 형식	119
<표 50> document_paragraph JSON 형식	120
<표 51> 병렬 말뭉치 JSON 형식	121
<표 52> 산출물 대상 3차례의 저장공간	132
<표 53> 시스템 보안 관련 법률 및 회사 규정	134
<표 54> 최종 구축 데이터 수량(한국어-외국어)	137
<표 55> 최종 구축 데이터 수량(외국어-한국어)	138
<표 56> 데이터 품질 감리 결과(한국어-외국어)	139
<표 57> 데이터 품질 감리 결과(외국어-한국어)	139
<표 58> 워크숍 세부 순서	140
<표 59> 워크숍 특강 내용	141
<표 60> 워크숍 오후 프로그램 세부 순서	143
<표 61> 언어별 워크숍 내용 및 활동 사진	143

그림 차례

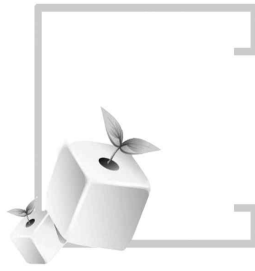
<그림 1> 사업의 범위	6
<그림 2> 사업 수행 절차 요약도	7
<그림 3> 사업 수행 체계 및 인력 구성	9
<그림 4> 사업 수행 주요 일정	12
<그림 5> 플리토 저작권 관련 이용약관 화면	21
<그림 6> 저작권 이용 허락 기간/㈜웨이커	26
<그림 7> 저작권 이용 허락 기간/㈜아주뉴스코퍼레이션	26
<그림 8> 저작권 이용 허락 기간/㈜웨이커	27
<그림 9> 저작권 이용 허락 기간/㈜플리토	27
<그림 10> 플리토 회원 가입	28
<그림 11> 플리토 이용약관 및 개인정보처리방침	29
<그림 12> 이용자별 성별, 참여 언어 설정 후 데이터 공급에 대한 동의	29
<그림 13> 실제 원문 구축 페이지인 아케이드 접속	30
<그림 14> 원문 구축 및 검수 참여 가이드 파악	30
<그림 15> 한국어 구어체 원문 구축 페이지	31
<그림 16> 플리토 내부 서버 내에 구축 및	31
<그림 17> 시나리오 '스톤피쉬' 예시	32
<그림 18> 신문 기사 게재 지면인 더구루(THEGURU) 사이트 화면 캡처	33
<그림 19> 플리토 회원 가입	34
<그림 20> 서비스 이용약관 및 개인정보처리방침 동의	35
<그림 21> 이용자별 성별, 참여 언어 설정 후 데이터 공급에 대한 동의	35
<그림 22> 실제 원문 구축 페이지인 아케이드 접속	36
<그림 23> 원문 구축 및 검수 참여 가이드 파악	36
<그림 24> 베트남어 구어체 원문 구축 페이지	37
<그림 25> 구축된 베트남어 문장의 검수 페이지	37
<그림 26> 플리토 내부 서버 내에 구축 및	38
<그림 27> 아주경제 베트남 사이트(Tiếng Việt AJU)	39
<그림 28> 아케이드 내 한국어 구어 데이터 검수	43
<그림 29> 일반인 수준 1차 한국어 구어체 원문 정제 가이드라인	45
<그림 30> 한국어 원문 정제 지침 교육 캡처본	46
<그림 31> 일반인 수준 2차 정제 한국어 구어체 지침	47
<그림 32> 일반인 수준, 전문가 수준 원문 정제 관련 질의 예시	48
<그림 33> 한국어 문어체 원문 정제 가이드라인	50

그림 차례

<그림 34> 일반인 수준 2차 정제 한국어 문어체 지침 예시	52
<그림 35> 플랫폼의 원문 신고 기능	55
<그림 36> 검수 후 발견된 원문 오류 공유 시트	56
<그림 37> 아케이드 내 베트남어 구어 데이터 검수	58
<그림 38> 베트남어 구어체 원문 정제 가이드라인	59
<그림 39> 베트남어 문어체 원문 정제 가이드라인	60
<그림 40> 베트남어 원문 전문가 수준 정제 예시	61
<그림 41> 플리토 아케이드 원문 구축 참여 가이드	62
<그림 42> 모국어 화자에게 제공되는 원문 정제 가이드라인	63
<그림 43> 전산 프로그램을 이용한 기계적 처리 과정	63
<그림 44> 번역 지침 제작 과정	65
<그림 45> 공공 용어의 외국어 번역 및 표기 지침(예시)	66
<그림 46> 플리토의 기존 번역 지침(예시)	67
<그림 47> (예시) 샘플 데이터 번역 및 논의	72
<그림 48> 한국어-외국어 번역 검수 공통 지침 Ver. 1	74
<그림 49> (예시) 한국어-외국어 번역 검수 세부 지침 Ver. 1	75
<그림 50> (예시) 최종 감수자의 자문 의견서	76
<그림 51> (예시) 언어별 외국어 자문위원의 자문 의견서	77
<그림 52> (예시) 한국어-외국어 번역 검수 공통 지침 Ver. 2	78
<그림 53> (예시) 한국어-외국어 번역 검수 세부 지침 Ver. 2	79
<그림 54> (예시) 언어별 세부 지침의 수시 개정	80
<그림 55> 한국어-외국어 번역 검수 지침 최종본 목록	80
<그림 56> 교육 영상 제작 개요	81
<그림 57> 영상 제작 업체 사전 회의	83
<그림 58> 영상 촬영을 위한 PPT 초안	84
<그림 59> 영상 1팀 촬영 장면	85
<그림 60> 촬영용 PPT 예시(태국어, 인도 힌디어)	85
<그림 61> 1팀 촬영 영상 피드백 반영 영상	86
<그림 62> 검수 관리팀장 촬영 장면(태국어, 인도 힌디어)	87
<그림 63> 자막 삽입 영상(태국어, 인도 힌디어)	88
<그림 64> 언어별 최종 지침 교육 영상	89
<그림 65> 비공개 YouTube 계정 및 시청 관리 페이지	90
<그림 66> 번역 수급·관리 업체 입찰 공고안	91

그림 차례

<그림 67> 업체 선정 및 계약 진행	91
<그림 68> 샘플 테스트 진행 예시	93
<그림 69> 샘플 테스트의 번역 지침 준수 여부 검수 예시	93
<그림 70> 샘플 테스트 품질 결과 점수 예시	94
<그림 71> 번역 및 작업 지침 예시	96
<그림 72> 번역 작업 예시	97
<그림 73> 검수 작업 예시	98
<그림 74> 기계적 검수 작업 예시	98
<그림 75> 검수·감수 지침 교육 절차	105
<그림 76> 검수용 플랫폼 문·구어 선택 예시	107
<그림 77> 검수용 플랫폼 작업 환경	108
<그림 78> 성희롱 성폭력 예방 교육 실시	114
<그림 79> 검수팀장의 10% 샘플링 재검수 예시	115
<그림 80> 검수 결과에 대한 검수 관리팀장과 검수원의 의견 교환 예시 ...	116
<그림 81> 번역 검수원 대상 검수 재교육 실시	116
<그림 82> 두 문장 이상 번역문 점검 및 수정 예시	117
<그림 83> 한국어 원문에 대한 질문 및 답변 정리 예시	118
<그림 84> 번역 품질 개선을 위한 의견 작성 및 전달 예시	118
<그림 85> 국립국어원 메타데이터(Metadata) JSON 형식	120
<그림 86> 국립국어원 document_paragraph JSON 형식	121
<그림 87> 국립국어원 병렬 말뭉치 JSON 형식	121
<그림 88> 한국어-외국어 구축 데이터 샘플	122
<그림 89> 원천 데이터 JSON 예시	123
<그림 90> 분석 데이터 JSON 예시	125
<그림 91> 보안 관리 담당자 역할 분포	129
<그림 92> 병렬 말뭉치 사업단 및 플리토 출입문 전경	130
<그림 93> 보안 관리 대장 예시	132
<그림 94> 저장 공간 예시 NAS(Network Attached Storage)	133
<그림 95> 저장 공간 예시 Cloud(Google Drive)	133
<그림 96> 워크숍 특강 현장	142
<그림 97> 학술대회 논문 발표 사진	149
<그림 98> 내일신문 기사	150
<그림 99> 인공지능신문 기사	151
<그림 100> 한-아세안센터 뉴스레터 기사 일부 발췌	152



제 1 장

사업 개요



제1장 사업 개요

1. 사업의 목표 및 중요성

1.1. 사업의 목표

- 본 사업의 목표는 대한민국의 새로운 교류 협력 관계로 주목받고 있는 국가들의 언어를 대상으로 한국어-외국어 간 병렬 말뭉치 데이터를 구축하는 것이다.
 - 병렬 말뭉치 대상 8개 언어는 아세안-인도 지역의 6개 언어(베트남, 인도네시아, 태국, 필리핀, 캄보디아, 인도)와 유라시아 지역의 2개(러시아, 우즈베키스탄) 언어이다.
- 자동 번역 등 언어문화산업의 성장 동력이 되는 고품질의 한국어-외국어 병렬 말뭉치를 구축하고자 한다.
 - 번역된 말뭉치 데이터를 이용해서 자동 번역 등 4차 산업 혁명 시대에 언어 문화 산업의 동력이 될 고품질의 병렬 말뭉치를 구축하고자 한다.
 - 언어 산업에 활용 가능한 빅데이터를 구축하고 활용함으로써 관련 산업 발전 분야를 마련하는 것이다.
 - 아직까지 많이 구축되지 않은 다국어 언어 처리 및 인공지능망 번역 기술을 향상시키는 데 이바지할 것이다.
 - 전 세계적인 한국어 학습 수요의 증대와 신한류 콘텐츠 확산에 기여하게 될 것이다.
- 산업계·연구계의 수요에 대응하여 한국과 아세안-인도 지역과 유라시아 지역의 교류를 활성화함으로써 상호 이해를 증진하고 미래 지향적 상생의 경제 협력 기반을 조성하는 기초가 될 것이다.

1.2. 사업의 중요성

○ 특수외국어 병렬 말뭉치 구축의 중요성

- 2018년에 제정된 '특수외국어 교육 진흥에 관한 법률'에 따라 국가 발전을 위해 전략적으로 필요한 외국어로 53개 언어가 지정되었다. 한국과의 외교, 국방, 무역, 경제, 관광, 문화, 교육 등의 교류 협력을 강화하기 위해서는 특수외국어에 관한 정보 수집과 전문 인력 양성이 필요하고 이를 위해서는 한국어와 특수외국어 간의 빅데이터 구축이 필요하다.
- 2020년 대외경제정책연구원의 '특수외국어진흥계획 마련을 위한 실태조사' 연구에 따르면 정부 지원 우선 대상 특수외국어로 선정된 언어 중에는 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 우즈베크어 등이 있었다. 선정 기준은 정량평가(발전 잠재력, 한국과의 정치·경제적 협력 가능성, 학문적 수요, 타 언어의 대체성, 산업별 전략적 중요성) 및 관련 기관의 정성평가를 종합 평가한 것인데, 그중 정량평가에서도 높은 점수를 받은 필리핀 타갈로그어 역시 잠재성과 수요에 비해 관련 데이터가 부족하므로 이 또한 병렬 말뭉치 구축 대상이 될 필요가 있다.
- 또한 특수외국어는 아니지만 국제적으로 영향력이 강하고 한국과 외교적, 경제적, 문화적, 전략적 중요성이 큰 러시아어 역시 언어 데이터가 부족한 만큼 병렬 말뭉치 구축이 필요하다.

○ 언어별 병렬 말뭉치 구축의 중요성

〈표 1〉 언어별 병렬 말뭉치 구축의 중요성

언어	목표
베트남어	- 인도차이나 반도의 중심 국가로 동아시아와 동남아시아를 연결하는 교통의 요충지로 정치, 경제적으로 주목받고 있다. - 지난 10년간 연평균 6%대의 경제 성장률을 기록하였고 한국 기업 약 8천여개가 진출하여 한국과 베트남의 경제 협력이 지속적으로 강화되고 있으며, 인적 교류도 확산되고 있다. 그리고 아세안 국가 중 가장 많은 한국인이 방문하는 국가이며, 양국 국민의 상호 방문은 계속 증가하고 있다. - 다양한 분야에 걸쳐 베트남어 시장 진출을 위한 수요가 지속적으로 증가하고 있으며, 외교 및 산업 현장에서 활동할 수 있는 베트남어 통번역 전문가와 지역 전문가가 필요하다.

언어	목표
인도네시아어	- 인도네시아어는 대외경제정책연구원의 보고서에 따르면 베트남어, 아랍어에 이어 중점 언어 평가 3위를 차지하였다. 동남아 최대 소비시장으로 인구가 2.7억 명에 달하며 경제 규모(GDP) 또한 약 1조 달러로 ASEAN 10개국의 37%를 차지하고 있어 동남아 최대 국가로 발전 잠재력이 매우 높은 것으로 평가되고 있다. - 한국 최초 해외 투자 대상국이었으며, 첫 번째 플랜트 수출대상국이다. 이후 '전략적 동반자 관계' 및 한-인도네시아 포괄적 경제동반자협정(CEPA)이 타결되며 한국과 긴밀한 관계를 유지하고 있다. - 한국 내 인도네시아어학과 및 인도네시아 내 한국어 관련 학과가 개설되어 있어, 언어 간의 교류 또한 활성화되어 있다. - 외교적, 경제적, 교육적 등의 다방면에서 한국에 중요한 언어이다.
태국어	- 태국어는 외교적, 경제적 협력 가능성이 높은 언어로서 대외경제정책연구원 자료에서 중점 언어 평가 결과 6위를 차지한다. - 한국관광공사의 외래 관광객 조사에 따르면, 최근 수년간 태국인 입국자 수는 6위로 관광 분야의 잠재력이 크다. - 한국국제교류재단 자료에 따르면, 2021년 태국 내 한류 동호회인 수가 약 1,480만 명에 달할 정도로 태국 내 한류의 영향력이 나날이 증가하고 있다. - 현재 전 세계에서 한국어를 배우는 중·고등학생 수가 가장 많은 나라로서 한국어교육 측면에서의 중요성이 큰 언어이다.
인도 힌디어	- 2019년 기준 세계 5위의 GDP 규모로 인구는 13억 명을 넘어 연평균 6~7%의 경제 성장률을 보이고 있고 중국에 이어 세계 최대 시장으로 급부상 중이다. - 힌디어는 인도뿐만 아니라 네팔, 파키스탄, 피지, 모리셔스 등에서도 지역 공용어로 사용되며, 전 세계적으로 4번째로 많이 사용되고 있는 언어이다. - 인도 정부는 힌디어를 국어(National Language), 나아가 UN 공용어 지정을 위해 노력하고 있으며, 국외적으로 힌디어의 위상이 높아지고 있으나 국내적 관심은 아직 부족한 상황이다. - 한국과 상품교역, 서비스교역, 경제 협력뿐만 아니라 문화 및 인력 교류도 활발하게 진행되고 있어 미래의 지속적인 발전을 위하여 언어 빅데이터의 구축이 중요하다.
캄보디아 크메르어	- 캄보디아 내 한류 열풍으로 한국의 문화예술, K-POP의 확산에 따라 한국어와 한국에 대한 관심이 증가하고 있다. 그리고 한국어를 배우는 학생들의 증가로 캄보디아 대학 내 한국어 및 한국학 강좌 개설이 확대되고 학생들의 교류가 늘었다. - 인공지능 기술 분야에서 열악한 환경의 캄보디아를 지원하여 한국어-크메르어 인공지능 기반 기계 번역의 발전을 도모할 수 있다.
필리핀	- 필리핀은 국내에서 언어 자원이 매우 적은 언어이다. 그러나 아세안 국가별 방

언어	목표
타갈로그어	문자 수는 베트남 다음으로 높으며, 국내 체류자(결혼 이민자, 근로자 포함)도 태국, 베트남에 이어 세 번째로 많고, 2020년 무역량도 급부상하고 있다. - 필리핀 내에서 한국어 학과 및 교양·어학원이 개설되어 있는 학교가 13개 학교로 필리핀 내 한국어교육이 매우 활성화 되어 있다. 그러나 한국에서는 필리핀 타갈로그어 교육이 전무한 상황이므로 인재들이 한국으로의 진출보다는 필리핀 내의 한국어교육과 관련 없는 회사로 흡수되고 한다. - 필리핀 정부에서 국내에 필리핀 관련 학과 개설을 위해 매우 적극적으로 접촉하고 있으며, 국내에서 소외 받는 필리핀인들과 필리핀 대학교에서 한국어를 공부하고 있는 인재들을 활용하기 위해 필리핀 타갈로그어의 언어 자원 구축이 필요하다. - 필리핀 타갈로그어는 영어로 대체될 수 없는 언어의 고유성이 존재(접요소, 높임 표현 발달, 영어와 다른 어순 등은 필리핀어의 독자성을 보여주는 요소)하며, 필리핀이 지니는 시장성에 주목한다면 지속적으로 말뭉치 구축이 필요하다.
러시아어	- 러시아어를 공용어로 하는 사용자 수는 세계 8위인 약 1억 5,000만 명이며 제2언어로 포함하면 약 2억 6천만 명으로 독립국가연합(CIS) 및 유엔 등의 국제기구에서 강한 위상을 지닌 언어이다. - 한류가 드라마와 K-POP에서 시작하여 최근에 한국어, 영화, 식품, 뷰티 등의 다양한 영역으로 확장되고 있어, 경제·사회·문화적 측면에서 향후 러시아 및 중앙아시아의 관계 발전을 위한 언어 자원이 필수적이다.
우즈베크어	- 2022년 한국과의 수교 30년을 맞아 경제 교류는 물론 문화 교류(한국 드라마, K-POP)가 활발하게 이루어지고 있다. - 현재 한국 내 거주 중인 우즈베크인은 2018년 기준 약 6만 6천 명으로 중앙아시아 국가 중에서 가장 많은 수치를 차지하고 있다. - 공공 부문에서는 한국을 벤치마킹하기 위해 찾는 공무원, 학자들을 대상으로 통번역 수요가 계속 증가하고 있으며 민간부문에서는 기업들의 우즈베키스탄 진출로 인해 현지인과의 소통을 위한 우즈베크어 교육이 이루어지고 있다. - 교육 부문에서는 매년 증가하고 있는 우즈베키스탄 유학생들로(2017년 2,921명, 2019년 7,426명) 인해 한국어교육의 수요가 증가함에 따라 한국어교육과 한국문화교육에서 한국어-우즈베크어 말뭉치의 중요성이 크다.

2. 사업의 범위

- 본 사업의 범위는 아세안-인도 지역 6개 언어와 유라시아 지역 2개 언어를 번역하여 말뭉치를 구축하는 것이다.
 - 먼저, 이를 위해서 한국어-외국어 병렬 말뭉치 구축을 위한 검증 체계를 수립하고 한국어와 외국어 병렬 말뭉치 총 800만 어절을 구축하였다.
 - 해당 언어에는 베트남어, 인도네시아어, 태국어, 인도-힌디어, 캄보디아-크메르어, 필리핀- 타갈로그어, 러시아어, 우즈베크어가 포함된다.
 - 또한 외국어-한국어 병렬 말뭉치를 시범 구축하여 이번 첫 사업에서 베트남어를 총 50만 어절을 구축하였다.
- 말뭉치 구축 과정에서 관련 저작권을 해결하고 이 해결된 저작권을 통해서 데이터를 어떻게 활용할 수 있을지에 대한 활용 방안 수립이 이 사업의 범위라 할 수 있다.

- 한국어-외국어 병렬 말뭉치 구축 및 검증 체계 수립
 - 구축 체계 및 방법론, 검증 체계 수립
- 한국어-외국어 병렬 말뭉치 총 800만 어절 구축
 - 아세안-인도 6개 언어: 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아어 크메르어, 필리핀 타갈로그어
 - 유라시아 2개 언어: 러시아어, 우즈베크어
- 외국어-한국어 병렬 말뭉치 시범 구축
 - 베트남어 총 50만 어절 구축
- 한국어-외국어 병렬 말뭉치 저작권 해결 및 활용 방안 수립

〈그림 1〉사업의 범위

3. 사업 수행 절차 및 체계

3.1. 사업 수행 절차

- 데이터 수집·구축은 (주)플리토(Flitto)를 통해서 이뤄졌고, 번역 내용 검수와 최종 감수는 (사)국제한국어교육학회(IAKLE)에서 진행하였다. 번역 내용 검수와 최종 감수는 1단계가 아니라 순환적이고 교차적으로 진행하였다.
- 수행 절차를 구체적으로 살펴보면, 먼저 (주)플리토에서 원문 데이터를 수집하고 수집된 원문 데이터의 검수와 정제를 진행하였다. 이 과정에서 (사)국제한국어교육학회도 정제 작업에 참여하였고 데이터 품질 향상에 기여하였다.
 - 다음으로 데이터를 가공한 후에 번역을 하는 과정을 거치게 되는데 번역된 데이터는 플리토에서 검수 과정을 1차적으로 거친 다음 정제 단계를 거쳐 학회에 넘어오면 데이터에 대한 검수와 교정 그리고 최종 감수를 진행하였다. 검수는 국내에서 유학을 한 경험이 있는 석사 학위 이상을 소지한 전문가들이 검수와 교정을 진행하였고, 각 대학에서 한국어 전공을 담당하고 있는 해외 대학 교수들이 최종 감수를 진행하였다.
 - 이러한 과정을 통해 총 850만 어절 이상의 병렬 데이터가 최종적으로 산출되었다.



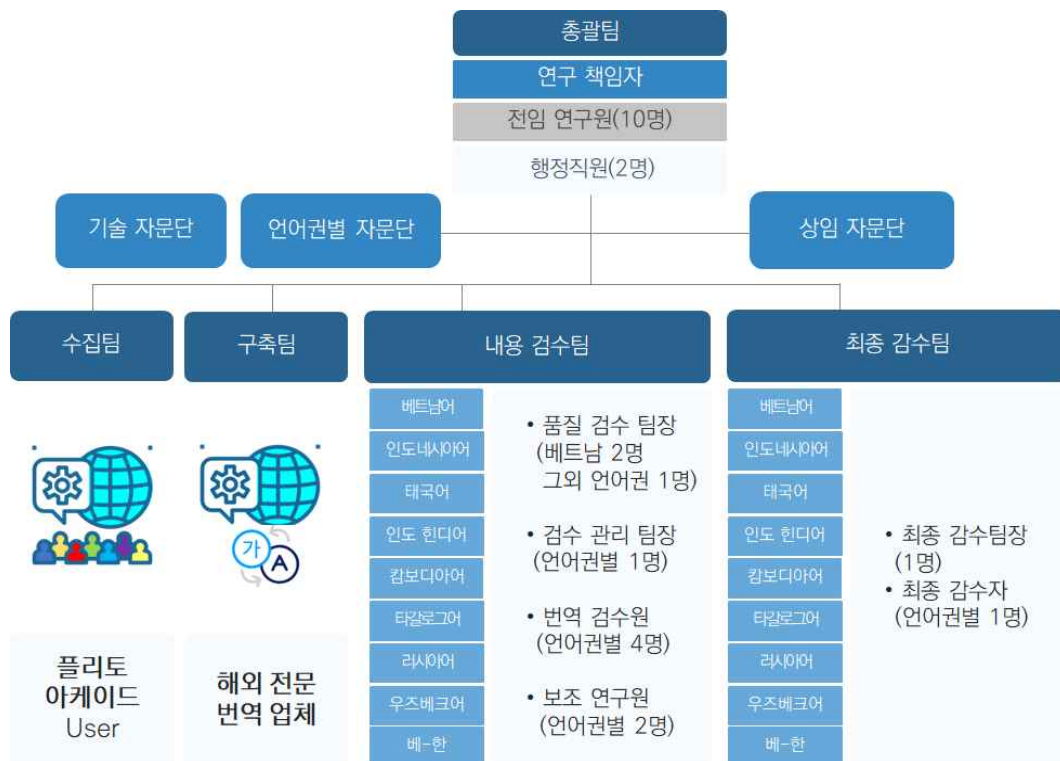
〈그림 2〉 사업 수행 절차 요약도

〈표 2〉 사업 수행 절차 상세 내용

구축 절차		업무 주체	상세 내용
수집	원문 데이터 수집	플리토	- 원문 수집팀: 공공데이터 및 뉴스 기사 및 원문 데이터 저작권 문제 해결 - 유저(User): 자체 아케이드를 통한 구어 데이터 구축
	원문 데이터 검수 및 정제		- 원문 검수팀: 정제 기준에 따른 데이터 기계 처리 및 플랫폼 업로드 - 유저(User): 자체 아케이드를 통한 무의미한 데이터 및 부적절한 내용 등 제외
구축	데이터 가공 번역	플리토	- 번역 지침 제공 및 번역 수행: 번역 지침 작성 및 배포 - 메타 정보 및 헤더 부착: 구축팀에서 데이터 관리 번호 및 분류·주제·소주제 등의 메타 정보 및 헤더 부착
	데이터 플랫폼 내 검수		- 도착어 번역자를 통한 1차 전수 검수 수행 - 플리토 내부 언어 전문가의 2차 검수 수행: 2차 번역 결과물에 대하여 플리토 소속 언어 전문가들이 5% 샘플링 검수 실시
정제	데이터 정제 검수 및 교정	(사) 국제한국어 교육학회	- 학회 전문 인력을 활용한 검수 교육 및 전수 검수 수행 - 연구 책임자: 검수 인력 배치 및 분장, 전체 관리 감독 - 검수 관리팀장 및 번역 검수원: 전수 검수 수행 - 품질·검수 관리팀장 및 전임 연구원: 지침 작성 및 배포 - 전임 연구원 및 검수 관리팀장: 검수 교육 실시 - 검수 품질 관련 수시 및 정기 회의 실시 - 연구 책임자: 검수 품질 최종 검토 및 회의 주최 - 품질·검수 관리팀장 및 전임 연구원: 검수 일정 공유 및 주요 사항에 대한 공유 - 보조 연구원: 회의록 작성 및 회의 운영 지원 - 언어 및 기술 관련 등 자문위원단 운영 - 연구 책임자 및 총괄팀: 자문 회의 주최 및 자문 내용 공정 반영 - 자문위원단: 정기 자문 회의 참석 및 자문 의견 전달
	데이터 정제 교수 감수		- 학회 소속 한국어 관련 전공 및 원어민 언어 전문가의 최종 감수 수행 - 연구 책임자 및 총괄팀: 감수 인력 관리 및 품질 최종 검토 - 최종 감수자: 10% 샘플링 감수 및 논의 사항 보고 - 최종 감수팀장: 감수 지침 검토 및 감수 일정 관리
산출	결과물 제출	(사) 국제한국어 교육학회, (주)플리토	- 최종 데이터 산출 및 제출 - 구축 데이터 최종 산출: 플리토 구축팀 - 연구 책임자 및 총괄팀: 최종 보고서 작성 및 보고

3.2. 사업 수행 체계

- 본 사업은 연구 책임자 1인과 전임 연구원 10명 그리고 행정 직원 2인으로 구성되었다. 자문단을 3개 분야에서 운영하였는데 말뚝치 구축에 필요한 기술 자문단, 그리고 8개 언어와 관련된 국내 언어별 외국어 자문단, 그리고 국제한국어교육학회 전임 회장들로 구성된 상임 자문단을 상시 운영하였다.
- 최종 번역 산출물은 3단계를 통해 이뤄졌는데, 먼저 1차적으로 플리토에서 원문 데이터를 수집 및 구축하였으며 1차적으로 번역한 데이터를 내부적으로 검수하였고 다음 단계로 내부 검수된 번역 데이터를 (사) 국제한국어교육학회에서 석사 이상으로 구성된 언어별 검수원들이 번역 데이터의 내용을 검수 교정하였다. 마지막으로 품질 향상을 위해 언어별 검수팀장이 재검수 과정을 수행하였으며 최종 해외 검수자들이 10% 검수를 진행하여 번역 데이터 품질을 높이는 과정을 수행하였다.



〈그림 3〉 사업 수행 체계 및 인력 구성

〈표 3〉 참여 인력 업무 분장표

팀명	직책	업무 내용
총괄팀	연구 책임자	• 사업 전반 계획의 수립 • 세부 사업 과제의 조정 및 총괄 감독 • 사업 인력 배치 및 관리, 감독 • 사업비의 집행 전반에 대한 관리 및 감독 • 사업 중간 및 최종 결과 보고
	전임 연구원	• 사업 세부 계획의 수립 및 진행 상황 점검, 조율 • 언어별 검수 및 감수 자료 분배 및 취합, 관리 • 검수 및 감수 지침 작성 및 교육 자료 제작 • 산출물 품질 관리 및 사업 위험 관리 • 사업단 전반의 의사소통 체계 구축 및 관리 • 정기 전체 회의 및 자문 회의 운영 지원 • 사업 중간 및 최종 결과 보고 지원 • 기타 사업 수행 지원
	행정 직원	• 사업 관련 행정 사무 및 관리 • 사업비 집행 및 정산, 증빙 자료 관리 • 사업단 기자재 관리 • 기타 사업단 운영 지원
내용 검수팀	공동 연구원 (품질 관리팀장)	• 검수 진행 일정 점검 및 조율 • 검수 지침 작성 및 검토 지원 • 검수에 대한 국어학적 지원 • 검수의 신뢰도 및 타당성 점검 • 팀 내 수시 회의 운영 • 검수 전반에 대한 총괄팀과의 협의 • 정기 전체 회의 및 자문 회의 참석
	검수 관리팀장	• 검수 진행 일정 점검 및 관리 • 검수 지침 작성 및 검토 지원 • 번역 검수원 대상 교육 및 교육 결과 점검 • 검수 전반의 오류 파악 및 분석 • 팀 내 수시 회의 참석 • 검수 전반에 대한 품질 관리팀장과의 협의 • 2차 번역에 대한 검수
	번역 검수원	• 2차 번역에 대한 검수 • 검수 시 논의 필요 사항 보고 • 검수 지침에 대한 피드백 • 팀 내 수시 회의 참석
	보조 연구원(행정)	• 팀 내 행정 업무 전반에 대한 진행 및 보고 • 팀 내 검수 계획 취합 및 보고 • 팀 내 수시 회의 운영 지원 및 회의록 작성

팀명	직책	업무 내용
		- 검수 기초 자료 수집 및 관리 - 검수원 요청 시 원문 자료에 대한 해설과 자료 제공 - 검수 관련 논의 사항 정리 및 보고
	보조 연구원(기술)	- 검수 시스템에 대한 관리 - 태깅 누락 및 오류 사항에 대한 전수 점검 - 검수 상황 점검 및 보고 - 팀 내 수시 회의 운영 지원 및 회의록 작성 - 검수 기초 자료 수집 및 관리
최종 감수팀	최종 감수팀장	- 최종 감수 진행 일정 점검 및 관리 - 최종 감수 지침 작성 및 검토 지원 - 최종 감수의 신뢰도 및 타당성 점검 - 최종 감수 전반에 대한 총괄팀과 협의 - 팀 내 수시 회의 참석
	최종 감수자	- 데이터 최종 감수(샘플링 10%) - 번역의 품질을 높이기 위한 최종 감수 진행 - 최종 감수 상황 및 논의 필요 사항 보고 - 팀 내 수시 회의 참석
자문 위원단	상임 자문위원	- 구축 말뭉치에 대한 국어학적 자문 - 사업 추진 방향 자문 - 언어 자료로서의 가치 검증 - 말뭉치 구축 과정과 체계에 대한 자문
	외국어 자문위원	- 언어별 검수 및 감수 지침에 대한 자문 - 언어별 검수 및 감수자 대상 교육 자료에 대한 자문 - 작업자별 검수 결과물에 대한 샘플링을 통해 산출물의 정확성 및 검수 능력 대한 검증(언어별 월간 자문 회의)
	기술 자문위원	- 말뭉치 설계와 구축 지침 수립에 관한 자문 - 검수와 감수 플랫폼 구축 및 관리에 대한 자문 - 메타 데이터의 오류 유무, 오류 원인 등에 대한 분석 및 자문 - 말뭉치 활용 방향성 및 계획 자문
수집 및 구축팀	총괄 책임자	수집 및 구축 총괄 책임
	실무 책임자	수집 및 구축 사업 실무 관리
	원문 데이터 수집	원문 데이터 획득 및 정제
	1차 번역/검수	1차 번역/검수 데이터 관리
	2차 검수/교정	2차 검수/교정 데이터 관리
	라벨링/어노테이션	최종 데이터 산출

4. 사업 수행 내용

4.1. 주요 진행 사항 및 조치 사항

월 추진과업	'21년8월	'21년9월	'21년10월	'21년11월	'21년12월	'22년1월	'22년2월	'22년3월	'22년4월	'22년5월
주요 일정		- 착수 보고			- 중간감사 - 중간 보고	- 설계 감리				- 최종감리 - 최종감사 - 최종보고
원문 데이터수집										
저작권 확보										
원문 데이터정제										
번역 검수 지침										
데이터 번역(한-외)										
데이터 번역(외-한)										
데이터 검수(한-외)										
데이터 검수(외-한)										
데이터 검수(한-외)										
데이터 검수(외-한)										
최종 납품										

〈그림 4〉 사업 수행 주요 일정

가. 사업 계획 수립 및 착수

- 사업 계획 수립: 2021년 8월 26일 완료
- 착수 보고회: 2021년 9월 9일 실시

나. 원문 구축 및 정제

- 원문 데이터 수집
 - 한국어-외국어, 외국어-한국어: 2022년 1월 27일 완료
 - 저작권 확보: 2022년 2월 22일 완료
- 원문 데이터 검수 및 정제
 - 한국어-외국어, 외국어-한국어: 2022년 2월 10일 완료

다. 번역

○ 번역 업체 선정

- 번역 인력의 안정적 확보를 위해 해외 번역 수급·관리 업체 입찰 공고 후 선정 및 계약: 2021년 10월 28일 완료

○ 원문 데이터 번역

- 한국어-외국어: 2022년 3월 27일 완료
- 외국어-한국어: 2022년 3월 28일 완료

라. 검수 및 감수

○ 번역 검수·감수 지침

- 번역 검수원 전체 교육(1차): 2021년 8월 27일 완료
- 검수·감수 지침 작성: 2021년 11월 12일 완료
- 검수·감수 지침 교육(1차): 2021년 12월 15일 완료
- 검수·감수 지침 교육(추가): 2022년 3월 3일 완료
- 이후 언어별 인력 충원에 따라 수시로 지침 교육 실시

○ 번역 데이터 검수

- 한국어-외국어: 2022년 5월 17일 완료
- 외국어-한국어: 2022년 5월 15일 완료
- 검수 품질 향상을 위한 10% 재검수: 2022년 5월 20일 완료 예정

○ 번역 데이터 감수

- 한국어-외국어: 2022년 5월 11일 완료
- 외국어-한국어: 2022년 5월 11일 완료

마. 인력 관리

○ 행정 직원 및 전임 연구원 선발

- 행정 직원: 2021년 9월 14일 완료
- 전임 연구원: 2021년 9월 30일 완료
- 감리 단계 추가에 따른 업무량 증가로 전임 연구원 충원: 12월 13일 완료

○ 접수·감수 인력 선발 및 관리

- 개인 사정, 접수량 저조, 감수 품질이 불량한 번역 감수원 해촉 및 신규 번역 감수원 선발
- 사업 기간 연장으로 인해 일부 번역 감수원 계약 연장, 최종 감수자 전원 계약 연장, 접수량 저조로 인한 최종 감수자 추가 계약

바. 의사소통 및 의견 수렴

○ 단계별 보고

- 착수 보고회: 2021년 9월 9일 실시
- 중간 보고회: 2022년 2월 10일 실시
- 최종 보고회: 2022년 5월 18일 실시

○ 발주처 및 컨소시엄사 정기 보고 및 회의

- 발주처 격주 보고 실시
- 발주처 월간 보고 실시
- 컨소시엄사 격주 회의 진행

○ 자문 의견 수렴

- 상임 자문 회의 실시(8회)
- 외국어 자문 회의 실시(4회)
- 기술 자문 회의 실시(2회)

사. 사업 효과 확산

○ 워크숍 실시

- 온·오프라인으로 ‘인공지능(AI) 병렬 말뭉치 구축의 실제’ 워크숍 실시: 2022년 2월 26일
- 인문학에서의 빅데이터 활용 및 병렬 말뭉치 구축 관련 특강 실시
- 언어별 번역 및 감수 워크숍 실시

○ 학술대회 발표

- 국제한국어교육학회(IAKLE) 제53차 국내(봄) 학술대회에 참가: 2022년 5월 14일 실시
- 구축 대상 8개 언어 및 병렬 말뭉치 관련 주제로 3건 발표

4.2. 주요 변경 사항

가. 원문 구축 및 정제

○ 원문 데이터 구축 방안 변경 및 재변경

- 사업계획서상 ‘자체 구축’으로 제안한 분야 중 국립국어원 데이터 연계를 목적으로 ‘모두의 말뭉치’ 또는 ‘한국어 교재’를 사용하여 문어체 원문 데이터로 활용
- ‘모두의 말뭉치’ 2차 저작권 활용이 불가하여 자체 구축으로 재변경

○ 2단계 원문 정제 단계 추가

- 원문 품질 향상을 위해 1차 원문 정제(특수외국어) 이후 2차 원문 정제(학회)를 진행

○ 구어체 원문 재구축

- 플리토 자체 구축 구어체 원문의 저작권 문제로 사용 불가
- 약관 변경, 동의 문구 변경 등을 통해 ‘저작권 사항이 재정비된’ 플리토 아카이브 구축 원문으로 변경

나. 번역

○ 번역 방식 변경

- 보조금 특성상 예산 집행의 절차를 위해 클라우드소싱 방식에서 번역 업체를 보조 사업자로 활용하는 방식으로 변경
- 번역 방식의 변경으로 번역 작업 및 언어 전문가(Linguist) 검수 시 엑셀 시트 활용

다. 병렬 말뭉치 구축

○ 메타데이터 검수 단계 축소

- 메타데이터 부착 자동화로 부착 전 샘플 검수로 수정

○ JSON 포맷 변경

- 국어원의 요청 사항을 반영하여 JSON 포맷 변경

라. 기타 변경 사항

○ 점수비·감수비 지급 기준

- 문장 기준에서 어절 기준으로 계산 방식 변경

○ 베트남어-한국어 데이터 수량 기준 확정

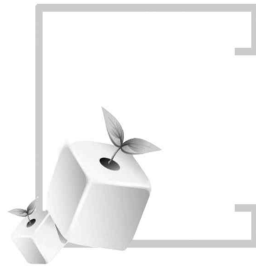
- 제안 요청서에 혼재되어 있던 어절·문장의 데이터 수량 기준을 문장으로 확정

○ 한국어-영어 병렬 말뭉치 구축 수행 방안 확정

- 번역 데이터 구축까지로 범위 확정

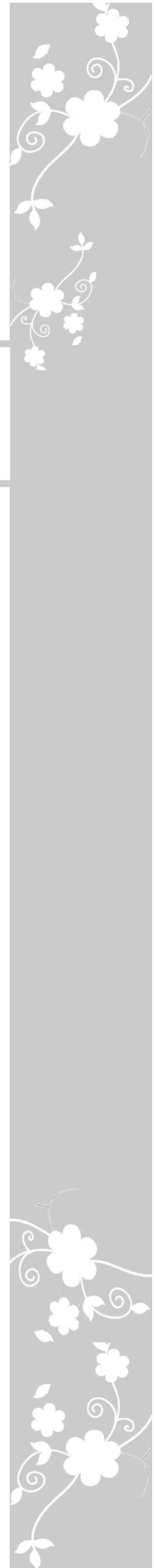
○ 사업 기간 연장

- 원문 재구축으로 인해 기존 사업 기간 연장
- 기존 4월 26일 종료에서 5월 26일 종료로 한 달 연장



제 2 장

사업 수행



제2장 사업 수행

1. 원문 구축 및 정제

1.1. 원문 구축

가. 원문 구축 목표 및 수집 수량

〈표 4〉 원문 구축 및 수집 목표 수량

구분	어체	목표 기준		수집 목표 (20%↑)	
		수량 (어절)	수량 (문장)	수량 (어절)	수량 (문장)
한-외 (한국어)	구어체	630,000	126,000	736,500	147,300
				19,500	2,500
	문어체	420,000	28,000	464,400	35,723
				39,600	4,500
베-한 (베트남어)	구어체	-	31,500	-	38,000
	문어체	-	21,000	-	25,000

- 원문 데이터의 품질을 높이기 위한 검수 및 교정 과정에서 사용 불가 문장, 저품질 문장 등이 제거 및 교정되며 수량의 차이가 생길 것을 고려해 전반적으로 20% 상향하여 수집하였다.

〈표 5〉 원문 구축 및 수집 목표 수량

구분	어체	목표 기준		수집 목표 (20%↑)		실제 수집 수량 (20%↑)
		수량 (어절)	수량 (문장)	수량 (어절)	수량 (문장)	
한-외 (한국어)	구어체	630,000	126,000	736,500	147,300	770,000어절
				19,500	2,500	20,000어절
	문어체	420,000	28,000	464,400	35,723	502,000어절
				39,600	4,500	45,000어절
베-한 (베트남어)	구어체	-	31,500	-	38,000	38,000문장
	문어체	-	21,000	-	25,000	25,000문장

나. 원문 선정 과정

1) 원문 선정

- 본 과제로 도출된 고품질의 언어 데이터가 한국어-외국어 대조 언어 연구나 해외의 한국어 학습 교육을 위한 기초 자료로의 활용 가능성을 고려하였다.
- 언중들의 사고방식 및 의식 구조를 반영해 구축되어 온 말뭉치뿐 아니라, 한국의 경제·사회·문화 등 기타 다양한 요소를 전파할 수 있도록 각계 분야에서 수집된 원문 활용의 중요성을 고려하였다.
- 이에 저작권 분쟁 요소를 사전에 제거한 국내외 뉴스 기사, 시나리오 등의 저작물을 활용하여 고품질의 데이터를 구축하기로 하였다.
- 본 사업 수행에 있어 자체 구축하는 구어체 대화문 데이터 이외에 추가적인 원문 데이터가 필요하며, 자료 수집의 용이성, 자료 활용성, 시장의 규모, 현재 언어의 흐름 등을 종합적으로 판단했을 때 문어체로 활용 가능한 가장 적합한 원문 데이터로써 신문 기사와 시나리오를 채택하였다.

〈표 6〉 원문 선정 이유

구분	선정 이유
신문 기사	시간 흐름별 정치, 경제, 사회, 문화 등의 다양한 이슈를 모두 다루고 있어 여러 산업에 걸쳐 영향력을 구사할 수 있으며, 모든 글이 글쓰기의 토대인 육하원칙을 바탕으로 분명한 의사 표현을 담아 작성된 데이터로서 본 과제에 활용하기에 매우 적절한 자료로 판단하였다.
시나리오	자모, 음절, 어휘, 어절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야의 응용 발전에 기여할 수 있는 기본적인 정보가 포함되어 있고, 또한 언어적 및 비언어적 정보 포함으로 한국문화 정서가 반영되어 있어, 향후 타 국가와의 언어·문화 교류에 있어 가장 적절한 예시로 활용될 가능성을 내포하고 있다고 판단하였다.

○ 이와 같이 선정된 원문 매체를 기반으로 설정한 세부 매체의 구성은 아래와 같다.

〈표 7〉 원문 매체의 구성

구분	어체	매체	세부 내용
한-외 (한국어)	구어체	자체 구축	플리토 아케이드로 자체 구축
		시나리오	저작권 계약(구매) 후 활용 (대화문, 작품 9개 내외)
	문어체	분야별 뉴스	경제 / 글로벌 / 사회 / 산업 / IT 과학 5종의 분야별 뉴스 기사 저작권 계약(구매) 후 활용
		시나리오	저작권 계약(구매) 후 활용 (지문, 작품 9개 내외)
베-한 (베트남어)	구어체	자체 구축	플리토 아케이드로 자체 구축
	문어체	분야별 뉴스	경제 / 글로벌 / 사회 / 산업 / IT 과학 5종의 분야별 뉴스 기사 저작권 계약(구매) 후 활용

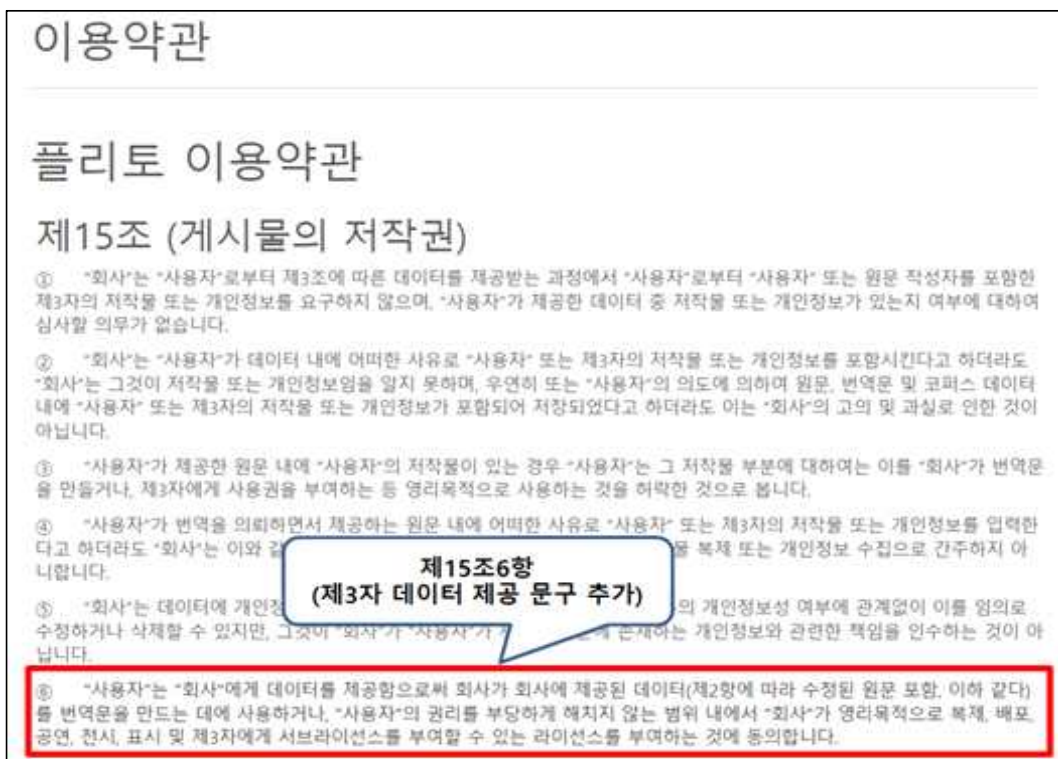
2) 저작권 확보 및 관리 방안

(1) 저작권 확보

○ 사전에 마련한 기준을 근거로 수집한 자료들의 저작권을 확보하여 본 과제에서 활용하고, 번역물에 대한 자유 배포가 가능하도록 양도계약서 및 저작권 이용 허락 계약서를 체결하여 저작권 이슈를 해소하였다.

○ 구어체

- 플리토에서 자체적으로 활용하는 플랫폼의 이용약관을 수정하여 저작권 이슈를 해소하였다.



〈그림 5〉 플리토 저작권 관련 이용약관 화면

○ 문어체

- (사)국제한국어교육학회, (주)플리토, (주)아주뉴스코퍼레이션 등 분야별 저작권 귀속 주체와의 계약을 통해 저작권 이슈를 해소하였다.

〈표 8〉 번역과 감수에 관한 저작재산권 양도계약서 - (사)국제한국어교육학회

구분	계약서	
	첫 페이지	날인 페이지
번역과 감수에 관한 저작재산권 전부에 대한 양도계약서 / (사)국제한국 어교육학회	저작재산권 전부에 대한 양도계약서 저작권과 및 저작권 양도인 (사)국제한국어교육학회(이하 “양도인” 이하 함)와 저작 권 양수인 국립국어원(이하 “양수인” 이하 함)은 아래 저작물 2021년 한국어-외국어 별명 발문지 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필리핀 타갈로그어, 태국어, 인도 네시아어, 캄보디아 크메리어, 러시아어, 우즈베크어)와 한국어(베트남어) 번역 저작물 등 산출물의 감수 저작물 등 산출물에 관한 저작재산권(이하 “저작재산권” 이하 함)과 관련하여 다음과 같이 계약을 체결한다. 다 음 제1조 (계약의 목적) 본 계약은 저작재산권 이전과 관련하여 양도인과 양수인 사이의 권리관계를 명확히 하 는 것을 목적으로 한다. 제2조 (계약의 대상) 본 계약의 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)에 대한 저작재산 권으로 한다. 저작물: 국립국어원의 보조사업자 (주)플리토와 (사)국제한국어교육학회가 수행한 2021 년 한국어-외국어 별명 발문지 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필 리핀 타갈로그어, 태국어, 인도네시아어, 캄보디아 크메리어, 러시아어, 우즈베크어)와 한국어 (베트남어) 번역 저작물 등 산출물의 감수 저작물 등 산출물 저작권자: (사)국제한국어교육학회 종별: 이론저작물 권리: 저작재산권 전부 ☐ 특제권, 공표권, 공중송신권(방송권, 전송권, 디지털방송송신권), 전시권, 배포권, 대여 권, 2차저작물 작성권 제3조 (저작재산권 양도범위) 본 계약에 의한 저작재산권 양도 범위는 제2조에서 정한 복제권 등 저작재산권 일체를 의미한다. 제4조 (양도인의 의무) (1) 양도인은 양수인에게 제3조에 의한 대상저작물의 저작재산권을 양도한다. (2) 양도인은 양수인에게 2022년 5월26일까지 대상저작물을 제공하여야 하며 양수인 의 수정보완 요청이 있을 경우 이에 응해야 한다. 수정보완 요청의 기간은 대상저작물 - 1 -	제16조 (계약 효력 발생일) 본 계약의 효력은 계약 체결일로부터 발생한다. 2022년 2월 22일 양도인 : 성 명: 사단법인 국제한국어교육학회 (인) 주 소: 서울시 동대문구 경희대로 26, 교수회관 310호 양수인 : 성 명: 국립국어원 (인) 주 소: 서울시 강서구 금남로 154 • 양도인 : (사) 국제한국어교육학회 • 양수인 : 국립국어원
	* 출처 : 001_번역 및 감수에 관한 저작재산권 전부에 대한 양도계약서 (사)국제한국어교육학회.pdf	

〈표 9〉 번역과 감수에 관한 저작재산권 양도계약서 - (주)플리토

구분	계약서	
	첫 페이지	날인 페이지
번역과 감수에 관한 저작재산권 전부에 대한 양도계약서 / (주)플리토	저작재산권 전부에 대한 양도계약서 저작권과 및 저작권 양도인 (주)플리토(이하 “양도인” 이하 함)와 저작권 양수인 국립 국어원(이하 “양수인” 이하 함)은 아래 저작물 (1) 2021년 한국어-외국어 별명 발문 지 구축 사업을 통해 수집된 원문 자료의 수정·작업 저작물 등 산출물, (2) 2021년 한 국어-외국어 별명 발문지 구축 사업의 외국어(영어, 인도네시아어, 베트남어, 필리핀 타갈로그어, 태국어, 인도네시아어, 캄보디아 크메리어, 러시아어, 우즈베크어)와 한국어(베트남어) 번역 저작물 등 산출물에 관한 저작재산권(이하 “저작재산권” 이하 함)과 관련하여 다음과 같이 계약 을 체결한다. 다 음 제1조 (계약의 목적) 본 계약은 저작재산권 이전과 관련하여 양도인과 양수인 사이의 권리관계를 명확히 하 는 것을 목적으로 한다. 제2조 (계약의 대상) 본 계약의 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)에 대한 저작재산 권으로 한다. 저작물: (1) 국립국어원의 보조사업자 (주)플리토와 (사)국제한국어교육학회가 수행한 2021년 한국어-외국어 별명 발문지 구축 사업을 통해 수집된 원문 자료의 수 정·작업 저작물 등 산출물 (2) 국립국어원의 보조사업자 (주)플리토와 (사)국제한국어교육학회가 수행한 2021년 한국어-외국어 별명 발문지 구축 사업의 외국어(영어, 인도네시아어, 베트남 어, 필리핀 타갈로그어, 태국어, 인도네시아어, 캄보디아 크메리어, 러시아어, 우즈베크어)와 한 국어(베트남어) 번역 저작물 등 산출물 저작권자: (주)플리토 종별: 이론저작물 권리: 저작재산권 전부 ☐ 특제권, 공표권, 공중송신권(방송권, 전송권, 디지털방송송신권), 전시권, 배포권, 대여 권, 2차저작물 작성권 제3조 (저작재산권 양도범위) 본 계약에 의한 저작재산권 양도 범위는 제2조에서 정한 복제권 등 저작재산권 일체를 의미한다.	제14조 (계약의 해석 및 보완) 본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민 법 등을 준용하고 사회 통념과 조리에 맞게 해결한다. 제16조 (계약 효력 발생일) 본 계약의 효력은 계약 체결일로부터 발생한다. 2022년 2월 22일 양도인 : 성 명: 주식회사 플리토 (인) 주 소: 서울특별시 강남구 테헤란로 211, 9층 양수인 : 성 명: 국립국어원 (인) 주 소: 서울시 강서구 금남로 154 • 양도인 : (주) 플리토 • 양수인 : 국립국어원
	* 출처 : 002_번역 및 감수에 관한 저작재산권 전부에 대한 양도계약서 (주)플리토.pdf	

〈표 10〉 저작권 이용 허락 계약서 - (주)아주뉴스코퍼레이션

구분	계약서	
	첫 페이지	날인 페이지
국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 / (주)아주뉴스코퍼레이션	국가 언어 자원 구축(말뭉치) 및 활용 저작권 이용 허락 계약서 저작자 및 저작권 이용허락자 (주)아주뉴스코퍼레이션 (이하 "권리자" 이 함)과 저작권 이용자 국립국어원(이하 "이용자" 이 함)은 아래 저작물에 관한 저작권산권 이용허락과 관련하여 다음과 같이 계약을 체결한다. 제1조 (계약의 목적) 본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위해 권리가 제공되는 저작물의 저작권산권 이용허락과 관련하여 권리와 이용자 사이의 저작물 권리관계를 명확히 하는 것을 목적으로 한다. 제2조 (정의) 본 계약에서 사용하는 용어의 뜻은 다음과 같다. (1) "수집 원문"이라 함은 국립국어원 및 국립국어원이 보존한 보조사업의 수행자(2차) 보조 또는 용역 사업자를 포함한다. (이하 "과업수행자"가 2021년 한국어-외국어 병렬 말뭉치 구축 사업(사업기간: 2021.8.-2022.4.제정)의 번역 대상이 되는 한국어 원문 또는 외국어 원문을 수집한 것을 말한다.) (2) "대상지작물"이라 함은 "수집 원문" 중에서 국립국어원 및 과업수행자가 병렬 말뭉치 번역 대상으로 선정된 한국어 원문 또는 외국어 원문을 말한다. (3) "복제·변형·번역물"이라 함은 국립국어원 및 과업수행자가 "대상지작물"에 편집 및 삭제, 언어 단위별 분리, 언어적 비언어적 정보 부각, 원문 교정, 원문 수정, 번역 등의 처리를 다한 결과물을 말한다. 제3조 (계약의 대상) 본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상지작물")에 대한 저작권산권 중 본 조에 명시한 이용 허락 범위로 한다. 저작물: 저작자가 2021년 한국어-외국어 병렬 말뭉치 구축 사업을 위해 "Xinh 16 All"의 4개년 (2018, 2019, 2020, 2021) 기사 전체에서 추출하여 제공한 신문기사 21,000문장 저작자: (주)아주뉴스코퍼레이션 ※저작권 이용허락 대상 권리의 내용 1. 국립국어원 및 국립국어원이 보존한 보조사업의 수행자(2차) 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 "수집 원문", "대상지작물", "복제·변형·번역물"을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일 2. 국립국어원 및 국립국어원이 보존한 보조사업의 수행자(2차) 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 자료, 음질, 어휘, 어절, 구절, 문장 및 텍스트	이 제3자에게 공개하여서는 아니된다. (2) 권리자가 제1항을 준수하지 않을 경우, 이용자 또는 과업수행자는 본 계약을 해지할 수 있으며, 권리자는 이로 인해 발생하는 모든 손해를 배상하여야 한다. 2022년 2월 22일 권리자 : (주)아주뉴스코퍼레이션 대표이사 (인) 사용자 : (주)아주뉴스코퍼레이션 대표이사 (인) 주소 서울특별시 중로구 종로1길 42/1호 주소 서울특별시 강서구 금남로154/1호 마일딩 11층 과업수행자 : (주)아주뉴스코퍼레이션 대표이사 (인) 과업수행자 : (주)아주뉴스코퍼레이션 대표이사 (인) 성명 사단법인 국제한국어교육학회 (인) 성명 주식회사 플리토 (인) 주소 서울특별시 동대문구 서울시립대로 163 주소 서울특별시 강남구 테헤란로21/1호 서울시립대학교 인문학관 5318호 • 권리자 : (주)아주뉴스코퍼레이션 • 이용자 : 국립국어원 • 과업수행자 : (사)국제한국어교육학회, (주)플리토
	* 출처 : 003_국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서_(주)아주뉴스코퍼레이션.pdf	

〈표 11〉 저작권 이용 허락 계약서 - (주)웨이커

구분	계약서	
	첫 페이지	날인 페이지
국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 / (주)웨이커	국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 저작권 이용허락자 (주)웨이커 (이하 "권리자" 이 함)와 저작권 이용자 국립국어원(이하 "이용자" 이 함)은 아래 저작물에 관한 저작권산권 이용허락과 관련하여 다음과 같이 계약을 체결한다. 제1조 (계약의 목적) 본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위해 권리가 제공되는 저작물의 저작권산권 이용허락과 관련하여 권리와 이용자 사이의 저작물 권리관계를 명확히 하는 것을 목적으로 한다. 제2조 (정의) 본 계약에서 사용하는 용어의 뜻은 다음과 같다. (1) "수집 원문"이라 함은 국립국어원 및 국립국어원이 보존한 보조사업의 수행자(2차) 보조 또는 용역 사업자를 포함한다. (이하 "과업수행자"가 2021년 한국어-외국어 병렬 말뭉치 구축 사업(사업기간: 2021.8.-2022.4.제정)의 번역 대상이 되는 한국어 원문 또는 외국어 원문을 수집한 것을 말한다.) (2) "대상지작물"이라 함은 "수집 원문" 중에서 국립국어원 및 과업수행자가 병렬 말뭉치 번역 대상으로 선정된 한국어 원문 또는 외국어 원문을 말한다. (3) "복제·변형·번역물"이라 함은 국립국어원 및 과업수행자가 "대상지작물"에 편집 및 삭제, 언어 단위별 분리, 언어적 비언어적 정보 부각, 원문 교정, 원문 수정, 번역 등의 처리를 다한 결과물을 말한다. 제3조 (계약의 대상) 본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상지작물")에 대한 저작권산권 중 본 조에 명시한 이용 허락 범위로 한다. 저작물: 2018년 01월 01일 ~ 2021년 12월 31일까지(1년 간)의 다구류(THE GURU) 기사 전체에서 권리자가 저작권 또는 저작권 제약을 하여 제공한 신문 기사 25,000문장(기사 내용: 원문, 어휘, 어절, 구절, 문장 및 텍스트) 저작자: 다구류(THE GURU) ※저작권 이용허락 대상 권리의 내용 1. 국립국어원 및 국립국어원이 보존한 보조사업의 수행자(2차) 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 "수집 원문", "대상지작물", "복제·변형·번역물"을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일 2. 국립국어원 및 국립국어원이 보존한 보조사업의 수행자(2차) 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 자료, 음질, 어휘, 어절, 구절, 문장 및 텍스트	(2) 권리자가 제1항을 준수하지 않을 경우, 이용자 또는 과업수행자는 본 계약을 해지할 수 있으며, 권리자는 이로 인해 발생하는 모든 손해를 배상하여야 한다. 2022년 2월 22일 권리자 : (주)웨이커 대표이사 (인) 사용자 : (주)웨이커 대표이사 (인) 주소 서울특별시 강남구 선릉로 66/1호 주소 서울특별시 강서구 금남로154/1호 이즈스케이, (주)웨이커 과업수행자 : (주)웨이커 대표이사 (인) 과업수행자 : (주)웨이커 대표이사 (인) 성명 사단법인 국제한국어교육학회 (인) 성명 주식회사 플리토 (인) 주소 서울특별시 동대문구 서울시립대로 163 주소 서울특별시 강남구 테헤란로21/1호 서울시립대학교 인문학관 5318호 • 권리자 : (주)웨이커 • 이용자 : 국립국어원 • 과업수행자 : (사)국제한국어교육학회, (주)플리토
	* 출처 : 004_국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서_(주)웨이커.pdf	

〈표 12〉 저작권 이용 허락 계약서 - (주)플리토

구분	계약서	
	첫 페이지	날인 페이지
국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 / (주)플리토	국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 저작권자 및 저작권 이용허락자 (플리토(이하 "권리자" 이 함)와 저작권 이용자 국립국어원(이하 "이용자" 이라 함)은 아래 저작물에 관한 저작권권 이용허락과 관련하여 다음과 같이 계약을 체결한다. 제1조 (계약의 목적) 본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위해 권리자가 제공하는 저작물의 저작권재산권 이용허락과 관련하여 권리자와 이용자 사이의 저작물 권리관계를 명확히 하는 것을 목적으로 한다. 제2조 (정의) 본 계약에서 사용하는 용어의 뜻은 다음과 같다. (1) "수집 원문"이라 함은 국립국어원 및 국립국어원이 보조한 보조사업의 수행과정에서 보조 또는 용역 사업자를 포함한다.(이하 "과업수행자"가 2021년 한국어-외국어 병행 말뭉치 구축 사업(사업기간: 2021.8.-2022.4.예정)의 번역 대상이 되는 한국어 원문 또는 외국어 원문을 수집한 것을 말한다. (2) "대상지작물"이라 함은 "수집 원문" 중에서 국립국어원 및 과업수행자가 병행 말뭉치 번역 대상으로 선정된 한국어 원문 또는 외국어 원문을 말한다. (3) "복제·변형·번역물"이라 함은 국립국어원 및 과업수행자가 "대상지작물"에 편집 및 삭제, 언어 단위별 분리, 언어적 비언어적 정보 부재, 원문 교정, 원문 수정, 번역 등의 처리를 다한 결과물을 말한다. 제3조 (계약의 대상) 본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상지작물"에 대한 저작권재산권 중 본 조에 명시한 이용 허락 범위로 한다. 저작물: 플리토 아카데미를 통해 저작자가 2021년 한국어-외국어 병행 말뭉치 구축 사업을 위해 제공한 대화 자료 (한국어 100,556문장 / 베트남어 32,000문장) (대화 자료 내용은 첨부1의 별지로 한다) 저작권자: ㈜ 플리토 ※저작권 이용허락 대상 권리의 내용 1. 국립국어원 및 국립국어원이 보조한 보조사업의 수행과정에서 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 "수집 원문", "대상지작물", "복제·변형·번역물"을 일정한 형식으로 전자적 기록 매체에 담기 보존하는 일 2. 국립국어원 및 국립국어원이 보조한 보조사업의 수행과정에서 보조 또는 용역 사업자를 포함	는 예외로 한다. 제13조 (계약의 해석 및 보완) 본 계약에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다. 제14조 (계약 효력 발생일) 본 계약의 효력은 동의 체결일로부터 발생한다. 2022년 2월 22일 권리자 성명: 주식회사 플리토 주소: 서울특별시 강남구 테헤란로211, 5층 이용자 (인) 성명: 국립국어원장 주소: 서울특별시 강서구 금남화로 154 · 권리자: 주식회사 플리토 · 이용자: 국립국어원
	* 출처 : 005_국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서_(주)플리토.pdf	

〈표 13〉 저작권 이용 허락 계약서 - 시나리오 작가

구분	계약서	
	첫 페이지	날인 페이지
국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 / 시나리오 작가(황수환)	국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서 저작자 및 저작권 이용허락자 황수환(이하 "권리자" 이 함)과 저작권 이용자 국립국어원(이하 "이용자" 이라 함)은 아래 저작물에 관한 저작권재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다. 제1조 (계약의 목적) 본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위해 권리자가 제공하는 저작물의 저작권재산권 이용허락과 관련하여 권리자와 이용자 사이의 저작물 권리관계를 명확히 하는 것을 목적으로 한다. 제2조 (정의) 본 계약에서 사용하는 용어의 뜻은 다음과 같다. (1) "수집 원문"이라 함은 국립국어원 및 국립국어원이 보조한 보조사업의 수행과정에서 보조 또는 용역 사업자를 포함한다.(이하 "과업수행자"가 2021년 한국어-외국어 병행 말뭉치 구축 사업(사업기간: 2021.8.-2022.4.예정)의 번역 대상이 되는 한국어 원문 또는 외국어 원문을 수집한 것을 말한다. (2) "대상지작물"이라 함은 "수집 원문" 중에서 국립국어원 및 과업수행자가 병행 말뭉치 번역 대상으로 선정된 한국어 원문 또는 외국어 원문을 말한다. (3) "복제·변형·번역물"이라 함은 국립국어원 및 과업수행자가 "대상지작물"에 편집 및 삭제, 언어 단위별 분리, 언어적 비언어적 정보 부재, 원문 교정, 원문 수정, 번역 등의 처리를 다한 결과물을 말한다. 제3조 (계약의 대상) 본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상지작물"에 대한 저작권재산권 중 본 조에 명시한 이용 허락 범위로 한다. 저작물: 본 계약서의 부속합의서 첨부1에 목록화한 대상지작물 저작자: 황수환 ※저작권 이용허락 대상 권리의 내용 1. 국립국어원 및 국립국어원이 보조한 보조사업의 수행과정에서 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 "수집 원문", "대상지작물", "복제·변형·번역물"을 일정한 형식으로 전자적 기록 매체에 담기 보존하는 일 2. 국립국어원 및 국립국어원이 보조한 보조사업의 수행과정에서 보조 또는 용역 사업자를 포함한다. 국립국어원이 발주한 용역 사업 수행자가 지도, 음질, 어휘, 어절, 구절, 문장 및 텍스트 단위와 국어, 연규, 언어, 처리 분야에 융통하기 위해 대상지작물을 복제·변형	이 제3자에게 공개하여서는 아니된다. (2) 권리자가 제1항을 준수하지 않을 경우, 이용자 또는 과업수행자는 본 계약을 해지할 수 있으며, 권리자는 이로 인해 발생하는 모든 손해를 배상하여야 한다. 2022년 2월 22일 권리자 : 성명 황수환 주소 경기도 성남시 분당구 이용자 : 성명 국립국어원장 주소 서울특별시 강서구 금남화로 154 과업수행자 : 성명 사단법인 국제한국어교육학회 (인) 성명 주식회사 플리토 주소 서울특별시 동대문구 서울시장대로 153, 3층 주소 서울특별시 강남구 테헤란로211, 5층 서울시립대학교 인문학관 5318호 · 권리자 : 황수환 · 이용자 : 국립국어원 · 과업수행자 : (사)국제한국어교육학회, (주)플리토
	* 출처 : 006_국가 언어 자원구축(말뭉치) 및 활용 저작권 이용 허락 계약서_황수환.pdf	

○ 이와 같이 저작권을 확보한 데이터의 수량은 아래 표와 같다.

〈표 14〉 저작권 확보한 최종 데이터 수량

구분	구축 수량	어체	목표 기준		실제 구매/구축량 (20%↑)	획득방식	출처
			수량 (어절)	수량 (문장)	수량		
한-외 (한국어)	1,050,000 어절	구어	630,000	126,000	770,000 어절	자체 구축	플리토 아케이드
					20,000 어절	시나리오	작가 황수환
		문어	420,000	28,000	502,000 어절	분야별 뉴스	더구루 (THEGURU)
					45,000 어절	시나리오	작가 황수환
베-한 (베트남어)	52,500 문장	구어		31,500	38,000 문장	자체 구축	플리토 아케이드
		문어		21,000	25,000 문장	분야별 뉴스	아주뉴스코퍼레이션 (Kinh tế Aju)

(2) 관리 방안

○ 구매 대상인 신문 기사, 시나리오 원문의 저작권을 '저작권 이용 허락 계약서' 체결 단계에서 이용 허락 계약 체결일로부터 최소 2051년 12월 31일까지 활용할 수 있게 아래와 같이 계약 조항에 포함하였다. 또한 계약의 유효 기간이 경과해도 권리자의 허락 중단 의사가 없다면 5년 단위로 자동 갱신될 수 있게 하여 향후 발생할 수 있는 저작권 이용에 관한 이슈를 사전에 방지하였다.

○ 플리토 아케이드를 통해 직접 수집한 구어체 원문의 경우, 참여자들이 회원 가입 단계의 '플리토 이용약관'에 동의함으로써 플리토에 귀속시켰다. 그리고 '저작권 이용 허락 계약서' 조항에 이용 허락 기간을 정하지 않음으로써 데이터의 이용 제한 가능성을 미연 방지하였다.

제4조 (이용허락 기간)

- (1) 대상저작물의 이용허락 최소 기간은 계약 체결일부터 2051년 12월 31일까지로 하며, 최소 기간 만료 후 관리자가 이용허락을 중지하고자 하는 의사를 밝히지 아니하면 이용허락이 5년 단위로 자동 갱신되며, 관리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락을 중지하여야 하며, 그렇지 아니하면 이용허락 내용이 유지된다.
- (2) 이용자의 명칭, 구조, 권한, 사업명이 변경되는 경우에도 본 계약의 효력이 미치는 것으로 한다.

제5조 (관리자의 의무)

- (1) 관리자는 이용자에게 대상저작물에 관하여 본 계약서 제3조에 따른 저작권권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.
- (2) 관리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다.
- (3) 관리자는 대상저작물에 본 계약 이행에 지장을 주는 저작권권 전부 또는 일부에 대한 제3자 이용허락권, 질권 등이 존재하는 경우, 사전에 이용자에게 이 사실을 알려야 한다.

〈그림 6〉 저작권 이용 허락 기간/(주)웨이커

제4조 (이용허락 기간)

- (1) 대상저작물의 이용허락 최소 기간은 계약 체결일부터 2051년 12월 31일까지로 하며, 최소 기간 만료 후 관리자가 이용허락을 중지하고자 하는 의사를 밝히지 아니하면 이용허락이 5년 단위로 자동 갱신되며, 관리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락을 중지하여야 하며, 그렇지 아니하면 이용허락 내용이 유지된다.
- (2) 이용자의 명칭, 구조, 권한, 사업명이 변경되는 경우에도 본 계약의 효력이 미치는 것으로 한다.

제5조 (관리자의 의무)

- (1) 관리자는 이용자에게 대상저작물에 관하여 본 계약서 제3조에 따른 저작권권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.
- (2) 관리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다.
- (3) 관리자는 대상저작물에 본 계약 이행에 지장을 주는 저작권권 전부 또는 일부에 대한 제3자 이용허락권, 질권 등이 존재하는 경우, 사전에 이용자에게 이 사실을 알려야 한다.

〈그림 7〉 저작권 이용 허락 기간/(주)아주뉴스코퍼레이션

제4조 (이용허락 기간)

- (1) 대상저작물의 이용허락 최소 기간은 계약 체결일부터 2051년 12월 31일까지로 하며, 최소 기간 만료 후 권리자가 이용허락을 중지하고자 하는 의사를 밝히지 아니하면 이용허락이 5년 단위로 자동 갱신되며, 권리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락을 중지하여야 하며, 그렇지 아니하면 이용허락 내용이 유지된다.
- (2) 이용자의 명칭, 구조, 권한, 사업명이 변경되는 경우에도 본 계약의 효력이 미치는 것으로 한다.

제5조 (권리자의 의무)

- (1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제3조에 따른 저작재산권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.
- (2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다.
- (3) 권리자는 대상저작물에 본 계약 이행에 지장을 주는 저작재산권 전부 또는 일부에 대한 제3자 이용허락권, 질권 등이 존재하는 경우, 사전에 이용자에게 이 사실을 알려야 한다.

〈그림 8〉 저작권 이용 허락 기간/(주)웨이커

제4조 (이용허락 기간)

- (1) 대상저작물의 이용허락 기간은 기한을 정하지 않는다.
- (2) 이용자의 명칭, 구조, 권한, 사업명이 변경되는 경우에도 본 계약의 효력이 미치는 것으로 한다.

제5조 (권리자의 의무)

- (1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제3조에 따른 저작재산권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.
- (2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다.
- (3) 권리자는 대상저작물에 본 계약 이행에 지장을 주는 저작재산권 전부 또는 일부에 대한 제3자 이용허락권, 질권 등이 존재하는 경우, 사전에 이용자에게 이 사실을 알려야 한다.
- (4) 권리자는 이용자가 저작자의 저작물을 이용하는 경우에 성명을 표시하지 않거나, 동일성을 유지하지 않는 경우라도 명예훼손에 해당하지 않는 경우에는 문제를 제기하지 않는다.

〈그림 9〉 저작권 이용 허락 기간/(주)플리토

다. 원문 수집 절차

1) 한국어 구어체

○ 자체 구축

- (주)플리토에서 개발한 클라우드소싱(Crowd-Sourcing) 저작 도구 ‘아케이드(Arcade)’를 활용하였다.
- 플리토 아케이드(Flitto Arcade) 회원 가입 시, 이용약관에 대한 동의를 통해 저작권 이슈를 해소하였다.
- AWS(Amazon Web Service)에서 구축된 본 플랫폼은 이중화 서버로 서비스 중단 없이 진행 가능하며, 구축에 참여하는 사용자(user)들이 본인의 PC나 Mobile 기기를 활용하여 원하는 시간·장소에서 손쉽게 번역·검수·교정을 할 수 있는 환경을 구축하였다.
- 한국어 구어체 데이터의 구축 절차는 아래와 같다.

〈그림 10〉 플리토 회원 가입

디렉토리

서비스 이용약관

1:1 전문 번역 이용약관

개인정보처리방침

위치정보 이용약관

GDPR 준수를 위한 노력

플리토 이용약관

제1조 (목적)

① 주식회사 플리토(이하 “회사”라 합니다)는 플리토(이하 플리토 웹사이트와 어플리케이션, 이를 통한 서비스 등을 통칭하여 “서비스”라 합니다)를 통해 “서비스” 사용자들이 번역을 의뢰하거나, 번역을 수행하거나, 혹은 번역된 콘텐츠를 향유할 수 있도록 하는 번역 플랫폼입니다.

② 본 이용약관은 플리토 및 플리토 관련 제반 서비스의 이용과 관련하여 “회사”와 “사용자”의 권리, 의무 및 책임사항, 기타 필요한 사항을 규정함을 목적으로 합니다.

제2조 (정의)

① 본 이용약관에서 사용하는 용어의 정의는 다음과 같습니다.

- “사용자”라 함은 본 이용약관에 동의하고 사용자 가입을 완료한 “서비스” 이용자를 말합니다.
- “사용자명”이라 함은 “사용자”의 식별과 “서비스” 이용을 위하여 “사용자”가 정하고 “회사”가 승인하는 문자와 숫자, 특수 기호의 조합을 의미합니다.
- “비밀번호”라 함은 “사용자”가 “서비스”를 이용하는 과정에서 “사용자명”을 사용할 권한이 있는 “사용자”임을 확인하고, “사용자”의 “서비스” 내 비밀을 보호하기 위하여 “사용자” 자신이 정한 문자 또는 숫자, 특수기호의 조합을 의미합니다.
- “계정”이라 함은 “사용자명”과 “비밀번호”에 기반한 로그인 계정을 의미합니다.
- “포인트”라 함은 “사용자”가 플리토에서 번역 활동을 통해 적립하거나 결제를 통하여 구입하고, “서비스” 이용 시 “회사”가 정한 기준에 따라 결제수단으로 사용할 수 있는, “서비스” 상의 가상 데이터를 의미합니다.
- “요청”이라 함은 “사용자”가 번역 또는 교정의 결과물을 얻기 위하여 플리토에 제공한 글, 사진, 음성, 음향 및 각종 파일과 링크 등 모든 형태의 원문을 제공하는 행위를 의미합니다.
- “번역”이라 함은 “사용자”가 번역 또는 교정의 수행을 위하여 플리토에 제공한 글, 사진, 음성, 음향 및 각종 파일과 링크 등 모든 형태의 번역문 또는 교정문을 작성하거나 제공하는 행위를 의미합니다.
- “번역가”라 함은 본 이용약관에 따라 요청 원문에 대한 번역 또는 교정 서비스를 제공하는 자를 말합니다.

〈그림 11〉 플리토 이용약관 및 개인정보처리방침

이벤트

아케이드에서
즐고 포인트를
아케이드란?

참여하기

기본 설정

아케이드의 다양한 학습활동을 적절하게 추천드리기 위해 다음 정보의 입력을 부탁드립니다.

성별

☐ 남성
☒ 여성

연령

20대

참여 언어

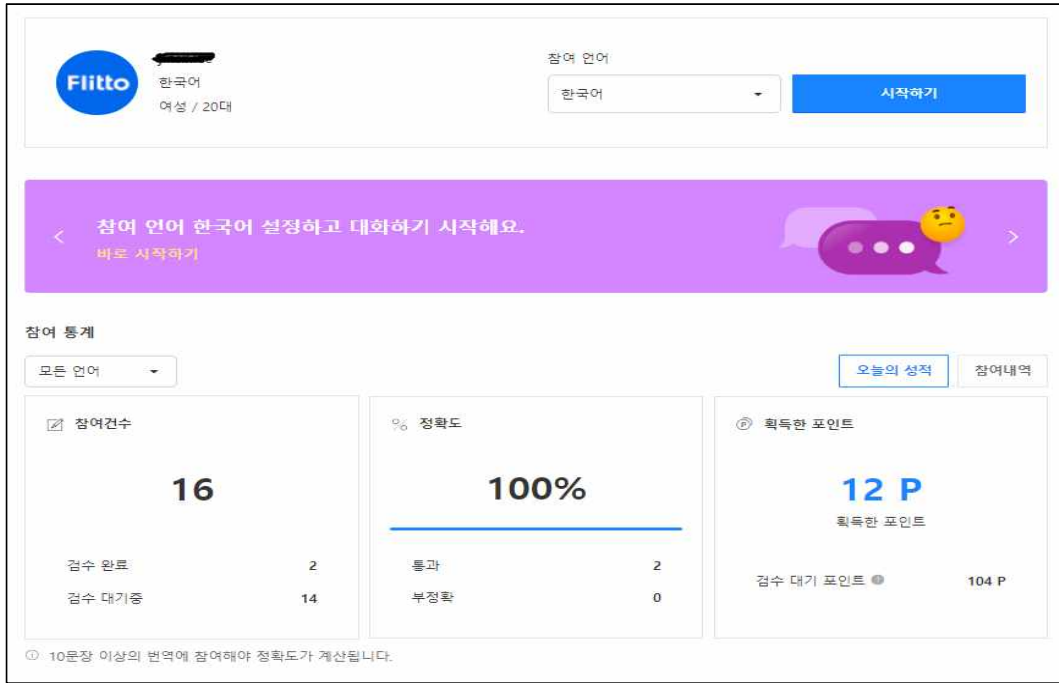
한국어

☐ 아케이드 참가 시 입력하신 번역문은 플리토 데이터로 저장되며 제3자에게 공급될 수 있습니다. (필수)

취소

완료

〈그림 12〉 이용자별 성별, 참여 언어 설정 후 데이터 공급에 대한 동의



〈그림 13〉 실제 원문 구축 페이지인 아케이드 접속

참여 가이드

- 제시된 주제와 상황에 맞게 대화를 이어가 주세요.**
주제에 맞는 대화를 제출해 주세요.
(예: 무역거래, 날씨, 일상생활, 해외 영업 등)
상황에 맞는 대화를 제출해 주세요.
(예: 친구와의 대화, 거래처와의 대화, 가게 점원과 고객과의 대화 등)
- 텍스트입력창의 최소, 최대 글자수 가이드를 확인해 주세요.**
가이드에서 제시된 최소, 최대 글자수를 만족해야 제출이 가능합니다.
띄어쓰기는 글자수에 포함되지 않습니다.
- 제출된 대화문은 다른 사용자들에게 적합 문장으로 평가를 받아야 통과되며 포인트를 획득할 수 있습니다.**
다음은 적합한 문장의 기준입니다.
주제와 연관성이 있는 문장
문맥상 적절한 문장
- 다른 사용자들에게 부적합 문장으로 평가 받은 글은 포인트를 받을 수 없습니다.**
다음은 부적합한 문장의 기준입니다.
개인정보 포함, 무의미한 내용(오탈자 포함), 갈등 조장 내용(비속어 포함)의 포함 유무는 다른 사용자들이 판단하며, 플리토는 이러한 결과를 변경할 수 없음을 알려드립니다.
개인정보 포함 : 주민번호, 전화번호 등의 개인을 식별할 수 있는 정보가 포함된 글

〈그림 14〉 원문 구축 및 검수 참여 가이드 파악

OP
 통계 참여내역 가이드

한국어
 30P

우리 회사는 선사입니다. 고객사와 채팅하는 상황에서 대화를 이어가 주세요.

B/L을 내일까지 발급 받아야 하는데 가능한가요?

아직 선적이 안 돼서 내일까지는 힘들어요.

선적기일은 오늘까지 아닌가요?

오늘 밤 늦게 선적 될 예정입니다.

제출 후 수정은 불가능합니다. 14 / 40

건너뛰기
 제출

〈그림 15〉 한국어 구어체 원문 구축 페이지

Arcade Review
 Home / arcade / arcade-review / chat-events / 53

Search Filters
 Q Search

1,616,175 results found.

Topic ID	Room ID (SID)	Order num	Content	Status	User	Points	Created At	Updated At
1	1 (1)	1	영수증 가지고 일주일 이내에 오시면 됩니다.	Passed	nose82	20	6/16/21, 4:03 PM	6/16/21, 4:38 PM
1	1 (1)	2	문자로 주소 보내주시겠어요?	Passed	jiyeon4	20	6/16/21, 4:38 PM	6/16/21, 5:14 PM
1	1 (1)	3	저는 고객인데요 방문 수거해주시나요?	Passed	TopSeven	20	6/16/21, 5:14 PM	6/16/21, 5:38 PM
1	1 (1)	4	방문 수거도 가능합니다. 고객님의에서 가능한 시간을 지정해주세요.	Passed	yeajin.kim	20	6/16/21, 5:38 PM	6/16/21, 6:07 PM
1	1 (1)	5	내일 오전에 가능할까요?	Passed	hyeyoon.jin	20	6/16/21, 6:07 PM	6/16/21, 9:05 PM
1	1 (1)	6	내일 오전 10시쯤 찾아뵙겠습니다.	Passed	dongho.won	20	6/16/21, 9:05 PM	6/16/21, 10:07 PM
1	1 (1)	7	내일 오전에 반쯤할 물건을 포장하여 현관앞에 두겠습니다.	Passed	MongolianBBQ	20	6/16/21, 10:07 PM	6/17/21, 1:31 AM

〈그림 16〉 플리토 내부 서버 내에 구축 및
검수 완료된 한국어 구어체 문장 저장

○ 시나리오 구매

- 해외의 한국어 학습자들을 대상으로 한국어 및 한국문화 등에 대한 교육으로 활용되는 데이터를 구축하기 위해 시나리오 작가(황수환)와의 직접 계약을 통해 저작권이 해소된 시나리오 원문을 구매하였다.

〈표 15〉 시나리오 구매 목록

작품 유형	작품명	저작자
단막	AtomicWinter	황수환
	BOTIS	
	Cognition	
	Welcome to my애미	
	맘스랜드	
	시간을 현상하다	
미니8부작	死의 형제들	
미니16부작	72, 마신	
시리즈	스톤피쉬	

로그 라인 <i>바다, 바람, 모래, 파도 그리고 그 모든 걸 자유롭게 타고 있는 너와 나의 이야기</i> 기획 의도 <i>“바다의 끝자락을 뚫지 못한 햇빛이 낮게 깔려 아직 해안에 닿지 못했다. 1.5m까지 솟구치는 파도가 투명하게 낮게 깔린 빛이 새벽임을 말해줬다. 스웰이 올라오며, 두껍던 파도의 끝자락이 말려 올라오며, 양쪽으로 밀리며 말린다. 가운데가 검푸른 색을 띠며, 흰색 커튼을 드리운 듯한 거대한 세트가 생긴다. 이를 바라보던 이들은 흥분을 감추지 못했다. 반면, 파도의 인사를 처음 보는 사람에게는 두려움이 가득했다.</i> <i>군함처럼 밀려오던 파도는 해안에 다가오며, 유리 조각처럼 산산이 깨지며, 하얗게 거품을 낸다. 깨진 파도는 해안의 모래를 안고 다시 바다로 사라졌다. 야인의 모습을 한 서퍼들이 하나둘 서프보드를 들고 바다로 향한다.”</i> 인스타에서 근육질의 몸매를 보이며, 미식축구 선수 복장을 한 인플루언서를 본 적 있다. 그는 단 한 번도 미식축구를 해본 적이 없는 남자였다. 그는 연기하고 있었다. 도시의 시멘트 상자 안에 갇혀, 가공된 육체미를 뽐내기 위해 운동을 하던 내 남성
--

〈그림 17〉 시나리오 ‘스톤피쉬’ 예시

2) 한국어 문어체

○ 신문 기사 구매

- 다양한 분야에서 복수의 집필자가 작성한 문어체 원문을 수집하기 위해 (주)웨이크와 데이터 원문 계약을 진행하여 저작권이 해소된 경제/글로벌/사회/산업/IT 과학 분야의 신문 기사를 구매하였다.

〈표 16〉 한국어 신문 기사 구매 목록

번호	분야	세부 분야	계약 주체	게재 지면	게재 연도
1	IT 과학	IT	(주)웨이크	더구루 (THEGURU)	2018~2021 (4개년)
2	경제	경제			
		생활경제			
		금융·증권			
3	산업	산업			
		부동산			
4	사회	사회			
5	글로벌	글로벌			



〈그림 18〉 신문 기사 게재 지면인 더구루(THEGURU) 사이트 화면 캡처

○ 시나리오 구매

- 시나리오 작가(황수환)와 직접 계약하여 저작권이 해소된 원문을 구매하였다.

3) 베트남어 구어체

○ 자체 구축

- (주)플리토에서 개발한 클라우드소싱(Crowd-Sourcing) 저작 도구 ‘아케이드(Arcade)’를 활용하였다.
- 플리토 아케이드(Flitto Arcade) 회원 가입 시, 이용약관에 대한 동의를 통해 저작권 이슈를 해소하였다.
- AWS(Amazon Web Service)에서 구축된 본 플랫폼은 이중화 서버로 서비스 중단 없이 진행 가능하며, 구축에 참여하는 사용자들이 본인의 PC나 Mobile 기기를 활용하여 원하는 시간·장소에서 손쉽게 번역·검수·교정을 할 수 있는 환경을 구축하였다.
- 베트남어 구어체 데이터의 구축 절차는 아래와 같다.

Sign up with Email

E-mail

Please input your email address.

ID

Password

Ngôn ngữ mẹ đẻ

Tiếng Việt

① Ngôn ngữ mẹ đẻ không thể thay đổi nên vui lòng lựa chọn cẩn trọng.

☐ Agree to all

☐ Agree to Terms of Service [Xem thêm](#)

☐ Agree to Privacy Policy [Xem thêm](#)

☐ Agree to receive promotional notice (optional)

Sign up

Already registered? [Log in](#)

〈그림 19〉 플리토 회원 가입

Directory
Terms and Conditions of Service
Terms and Conditions of 1:1 Pro Translation Service
Privacy Policy
Terms and Conditions of Location Information Service
Efforts for GDPR Compliance
Menu Translation - Terms
Menu Translation - Privacy Policy

Flitto Terms and Conditions

Article 1 (Purpose)

① Flitto Co., Ltd. (hereinafter referred to as "Company") makes users of "Service" to request translations, to translate, or to enjoy the translated contents through "Flitto" (hereinafter referred to as "Service", collectively referred to as the "Flitto Website, Applications, and Services").

② These Terms and Conditions are intended to prescribe the rights, obligations, and responsibilities of "Company" and "User", and other necessary matters related to the use of Flitto and all the services related to Flitto.

Article 2 (Definition)

① The terms used in these terms and conditions are defined below.

- "User" means a user of "Service" who has agreed to these Terms of Use and has completed the User Registration.
- "User Name" means a combination of letters, numbers, and special symbols that are designated by the "User" and approved by the "Company" to identify the "User" and use the "Service".
- "Password" means a combination of letters, numbers and special symbols that "User" set to confirm that "User" is the "User" who is authorized to use "User Name" in the process of using "Service", and to protect the secrets of "User" in "Service".
- "Account" means a login account based on "User Name" and "Password".
- "Points" means virtual data in "Services" that "Users" can earn through translation activities or buy through purchase in Flitto, and can be used as a payment method under the criteria set by the Company when using the Services.
- "Request" means an act of "User" providing all forms of original text, including texts, photographs, voices, sounds, and various files and links, to Flitto for the results of translation or proofreading.
- "Translation" means an act of "User" creating or providing all forms of translation or proofreading, including texts, photographs, voices, sounds, and various files and links to Flitto.
- "Translator" means any individual providing services for translating or proofreading the original text in accordance with this Terms of Use.

② Any part of the terms used in these Terms and Conditions not defined in this Article is subject to the relevant laws and general practices.

〈그림 20〉 서비스 이용약관 및 개인정보처리방침 동의

Discovery

User settings

Please fill in the following information for proper recommendations of Arcade learning activities.

Sex
☐ Nam
☐ Nữ

Age
20s

Ngôn ngữ tham gia
Select language

☐ Kết quả đã nhập khi tham gia Acarde sẽ được Flitto lưu lại và có thể cung cấp cho bên thứ 3 (bao gồm Viện ngôn ngữ quốc gia và bên thứ 3 được cung cấp thông qua Viện ngôn ngữ quốc gia). (bắt buộc)

Hủy

Hoàn tất

〈그림 21〉 이용자별 성별, 참여 언어 설정 후 데이터 공급에 대한 동의



Tiếng Việt

Nữ / 20s

Ngôn ngữ tham gia

Tiếng Việt

Bắt đầu

Arcade là gì? Xem hướng dẫn

Xem thêm



Trạng thái hoạt động

Tất cả ngôn ngữ

Hoạt động hôm nay

Hoạt động của tôi

☒ Số lượng tham gia

0

Đánh giá hoàn tất

0

Đang chờ kiểm tra

0

☐ Độ chính xác

0%

Qua

0

Không chính xác

0

☐ Điểm đã nhận

0 P

Điểm đã nhận

Điểm bạn sẽ nhận

0 P

① Độ chính xác được tính sau 10 câu đầu tiên.

〈그림 22〉 실제 원문 구축 페이지인 아케이드 접속

Participation Guide

Arcade là gì?

Nội dung lưu ý

Hướng dẫn đánh giá

Kiểu hội thoại

1. Hãy viết tiếp cuộc hội thoại sao cho phù hợp với tình huống, chủ đề cho trước.

Hãy viết câu hội thoại phù hợp với chủ đề.

(Ví dụ: Giao dịch thương mại, thời tiết, cuộc sống thường nhật, Sales Overseas, v.v)

Hãy viết câu hội thoại phù hợp với tình huống.

(Ví dụ: Cuộc đối thoại với bạn, cuộc đối thoại với đối tác giao dịch, cuộc đối thoại giữa nhân viên cửa hàng và khách hàng, v.v)

2. Hãy kiểm tra số từ tối đa và tối thiểu có thể nhập vào hộp viết lời thoại.

Bạn chỉ có thể nộp bài khi thực hiện đúng yêu cầu về số ký tự tối đa và tối thiểu như hướng dẫn.

Đấu cách không được bao hàm trong tổng ký tự.

3. Câu hội thoại bạn gửi lên phải được người dùng khác đánh giá là câu phù hợp thì mới có thể được thông qua và nhận điểm.

Tiêu chuẩn của câu văn phù hợp như sau:

Là câu có liên quan tới chủ đề

Là câu có mạch văn trôi chảy

〈그림 23〉 원문 구축 및 검수 참여 가이드 파악

- 36 -

Arcade Review

Home / arcade / arcade-review / chat-events / 116

Search Filters Q Search

23,202 results found.

Topic ID	Room ID (SID)	Order num	Content	Status	User	Points	Created At	Updated At
6395	216776 (1)	11	minh có, bạn đặt giúp mình nhé. mình cảm ơn bạn nhiều.	Passed		10	1/15/22, 2:48 AM	1/15/22, 12:31 PM
6395	216776 (1)	12	vậy bây giờ mình sẽ thuê hai xe để đủ nhiều người nhé	Passed		10	1/15/22, 12:31 PM	1/15/22, 12:40 PM
6395	216776 (1)	13	không có loại xe nào đủ chỗ cho 12 người luôn hả bạn	Passed		10	1/15/22, 12:40 PM	1/16/22, 3:57 PM
6395	216776 (1)	14	dạ vâng hiện tại xe mướn sáu chỗ bên em đã có lịch trình mất rồi ạ	Passed		10	1/16/22, 3:57 PM	1/16/22, 3:58 PM
6395	216776 (1)	15	Vậy thôi vậy, đi 2 xe riêng thì có nhiều bất tiện lắm. Đi xe chung với đoàn thì cũng được rồi.	Passed		10	1/16/22, 3:58 PM	1/16/22, 6:35 PM
6395	216777 (2)	1	hôm nay chúng ta sẽ đi đến những địa điểm du lịch nào vậy?	Passed		10	1/12/22, 3:01 PM	1/16/22, 6:02 PM
6395	216777 (2)	2	buổi sáng chúng ta sẽ đi tham quan đồi chè cầu đất nhé	Passed		10	1/16/22, 6:02 PM	1/16/22, 7:29 PM
6395	216777 (2)	3	Ồ thế à. Trước khi đi thì chúng ta ăn sáng ở đâu vậy.	Passed		10	1/16/22, 7:29 PM	1/16/22, 10:02 PM

〈그림 26〉 플리토 내부 서버 내에 구축 및
검수 완료된 베트남어 구어체 문장 저장

4) 베트남어 문어체

○ 신문 기사 구매

- 다양한 분야에서 복수의 집필자가 작성한 문어체 원문을 수집하기 위해 (주)아주뉴스코퍼레이션과 데이터 원문 계약을 진행하였으며, 저작권이 해소된 경제/글로벌/사회/산업/IT 과학 분야의 신문 기사를 구매하였다.

〈표 17〉 베트남어 신문 기사 구매 목록

번호	분야	세부 분야	계약 주체	게재 지면	게재 연도
1	IT 과학	IT	(주)아주뉴스코퍼레이션	Kinh tế Aju	2018~2021 (4개년)
2	경제	경제			
		생활경제			
		금융·증권			
3	산업	산업			
		부동산			
4	사회	사회			
5	글로벌	글로벌			

KOREAN

CHINESE

ENGLISH

JAPANESE

THỨ BA, 17. 05. 2022

Peace in Ukraine

KINH TẾ AJU

Tìm kiếm tin tức

Q

KINH TẾ CHÍNH TRỊ

ĐỜI SỐNG - XÃ HỘI

GIẢI TRÍ

DU LỊCH

VIỆT NAM

THẾ GIỚI

billboard

MUSIC AWARDS

Năm thứ 6 liên tiếp BTS giành giải tại Billboard Music Awards

NETFLIX

사이버지옥

(사이버지옥)

Netfix ra mắt phim tài liệu 'Cyber Hell' dựa trên vụ án "phòng chat thứ N"

Phim tài liệu 'Cyber Hell: Exposing an Internet Horror', phỏng theo vụ án "phòng chat thứ N" của Hàn Quốc, sẽ có mặt trên Netflix vào ngày 18/5 tới. Đạo diễn 'Cyber Hell: Exposing an Internet Horror' đã thực hiện Hoàng Phương Ly phóng viên lyhoang215@ajunews.com

17:08 20-04-2022

TIN TỔNG HỢP

KINH TẾ CHÍNH TRỊ

Tổng thống Yoon Suk-yeol nhậm chức

KINH TẾ CHÍNH TRỊ

Tổng thống Moon Jae-in phát biểu kết thúc nhiệm kỳ

KINH TẾ CHÍNH TRỊ

Chính phủ mới công bố 110 nhiệm vụ quốc gia

KINH TẾ CHÍNH TRỊ

Hội nghị thượng đỉnh Hàn-Mỹ sẽ diễn ra vào tháng 5

〈그림 27〉 아주경제 베트남 사이트(Tiếng Việt AJU)

- 39 -

라. 원문 예시

1) 한국어

〈표 18〉 한국어 원문 예시

구분	출처	수집 원문
문어체	(주)웨이커	26일 관련 업계에 따르면 베이징현대는 중장기 목표로 2023~2025년 전체 매출에서 신에너지 제품의 비중을 20% 달성한다는 목표를 설정한 것으로 알려졌다.
		28일 관련 업계에 따르면 아모레퍼시픽 럭셔리 뷰티브랜드 설화수는 티몰이 개최하는 슈퍼 브랜드 대회 이벤트 '스노우쇼'에 참가, 안티에이징 제품을 선보였다.
		한편, 싱가포르 관세당국은 창이공항 면세사업 입찰 결과를 연말께 발표할 예정이다.
		툼 디드리오씨는 이날 "주민들의 건강과 복지를 희생시키고 대기업의 이익만을 보호하려 한다."며 알렘하림푸드가 과거 식수 오염을 불러온 사례들을 언급했다.
		지난 2년간 공급받은 아무르스틸 제품의 우수성을 확인해 계약을 연장하기 위한 조치에서다.
	황수환 작가	로비에는 아무도 없고, 벽에 걸린 TV에선 신호가 안 잡히고, 전자음만 들린다.
		핸드폰을 보다가 바닥을 보다가도 주현을 보고는 물길 갈라지듯이 사람들이 주현을 피한다.
		거울을 한번 보고는 옷 몇 벌을 집어 나간다.
		사람의 흔적이 전혀 보이지 않는다.
		뉘엿뉘엿 해가 지고 주변에 그림자가 길게 늘어져 있다.
구어체	플리토 아케이드	잠시면요 ~ 색상은 검정 흰색 회색이 있습니다.
		그럼 내일 저녁먹고 치러가자!
		장어 초밥 좋아하시는 것 같은데 하나 먹어도 되나요?
		난 비행중엔 소화가 잘 안되더라구.
		집근처 카페에서 커피 한잔 사서 가자.
	황수환 작가	이제 임부장이 골든포레스트 전담해서 관리 해야 하니까 내가 직급도 거기 면이 서라고 올려 준거야!
		오부장은 계약서 검토 다시 하고, 돌아오면 볼 수 있게 해줘.
		나도 서핑하려고 요즘 보드 보고 있잖아!
		너 경상도 함양 출신이라며, 넌 왜 여기 있냐?
		이정도는 알려줘도 되는 거 아니에요?

2) 베트남어

〈표 19〉 베트남어 원문 예시

구분	출처	예시
문어체	(주)아주뉴 스코퍼레 이션 (Kinh tế Aju)	Trong tình huống 52 giờ một tuần được giới thiệu và dân số có thể sản xuất đang giảm, cải thiện năng suất lao động được coi là một nhiệm vụ cần giải quyết.
		Trước đó, ngày 16/4/2019, cảnh sát đã khám nhà riêng và xe hơi của Park ở thành phố Hanam , đồng thời tiến hành khám xét thân thể của người này.
		Hầu hết các nhân viên Việt Nam làm việc tại các công ty Hàn Quốc ở Việt Nam trả lời rằng họ hài lòng với môi trường làm việc, nhưng cho rằng văn hóa tổ chức là một lĩnh vực cần được cải thiện.
		CTA dự đoán rằng thời đại kỹ thuật số, bắt đầu từ những năm 2000, sẽ bước qua kỷ nguyên kết nối vào những năm 2010 và bước vào kỷ nguyên dữ liệu vào năm 2020.
		"Tuy nhiên, báo cáo cũng phản ánh những thách thức đầy khó khăn liên quan đến sự lệch pha trong tốc độ hồi phục kinh tế tại các nước và khắp toàn cầu và nguy cơ tổn thất kinh tế kéo dài từ khủng hoảng", nhà kinh tế Gopinath bổ sung.
구어체	플리토 아케이드	Dạo này tôi bận bịu quá mà không có thời gian chăm sóc tóc nữa.
		Tôi muốn tổ chức tiệc vào buổi tối cuối tuần này.
		Vậy em qua một lát thôi nhé.
		Đang đói thế này tớ sợ không chờ được mất.
		Họ yêu nhau vì tính cách chứ không phải ngoại hình nên có quan trọng gì đâu.

1.2. 원문 정제

가. 한국어 원문 정제 절차

- 한국어 원문 정제는 원문의 특성에 따라 최대 4단계 원문 정제를 실시하였다.
- 아케이드 내 구축 구어체의 경우 아케이드 사용자들이 구축한 것으로 오타자 및 내용 중복이 많기 때문에 4단계 정제를 실시하였으며, 시나리오 및 신문의 경우 아케이드 내 검수를 제외한 3차 정제를 실시하였다.
- 각 단계는 별도로 진행되는 것이 아니라 순환식으로 진행되었으며, 번역 및 검수, 감수 완료될 때까지 지속적으로 진행되었다.

〈표 20〉 한국어 원문별 정제 절차

어체	매체	1단계	2단계	3단계	4단계
구어체	자체 구축	기계적 정제	아케이드 내 정제	일반인 수준 정제	전문가 수준 정제
	시나리오	기계적 정제	일반인 수준 정제	전문가 수준 정제	-
문어체	신문	기계적 정제	일반인 수준 정제	전문가 수준 정제	-

나. 한국어 구어체 원문 정제

1) 기계적 정제

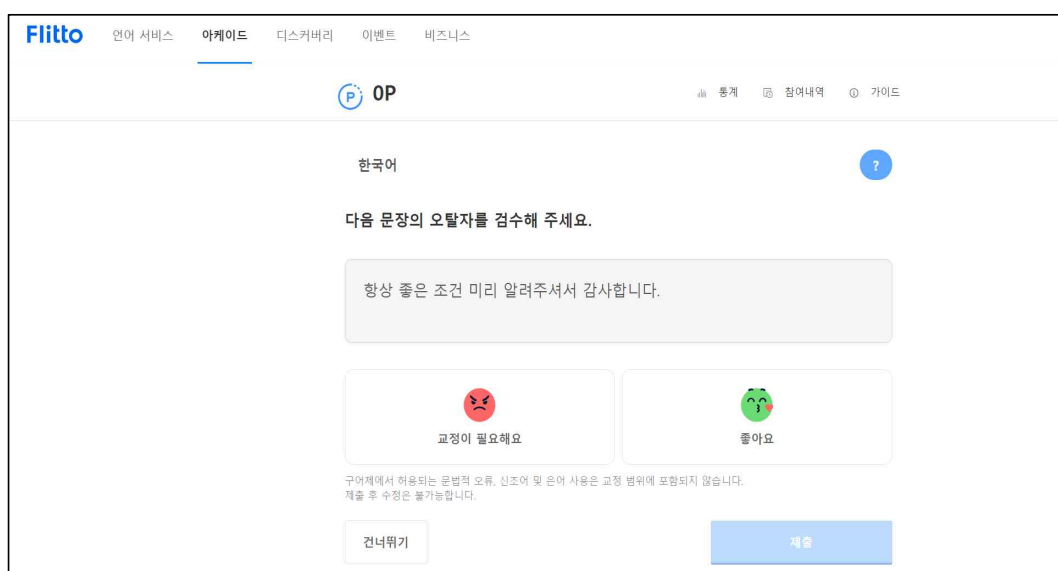
- 수집 및 구매를 통해 구축된 원문 데이터는 인력 정제 이전에 어절 수, 부적합 내용, 개인 정보와 같이 기계적으로 정제 가능 요소들을 선별하여 작업을 진행하였다.

〈표 21〉 한국어 구어체 기계적 정제 내용

정제 기준	내용
어절 수	문장당 평균 약 5어절(최솟값 3 ~ 최댓값 20)
부적합 내용	부적절한 내용, 문장 내 자연스럽지 않은 어투, 불필요한 발화, 이모티콘 포함, 오류가 있는 문장(오타자 포함)
개인 정보	인명, 전화번호, 이메일 주소, 상표명 등의 정보 비식별화 및 삭제
중복 내용	완전 일치 문장의 제거
특수문자	구두점 확인 및 특수 문자 제거

2) 아케이드 내 정제

- 기계적 정제를 마친 데이터는 1차적으로 플리토에서 자체 구축하고 운영 중인 플랫폼인 ‘아케이드’에서 검수를 진행하였다.
- ‘아케이드’에는 구축된 구어체 원문 데이터와 정제를 위한 가이드라인이 등록되어 있으며, 플랫폼 유저들(검수자들)이 확인하여 구축된 문장의 적합 및 부적합 여부를 선택할 수 있는 환경을 조성하였다.
- 오타자, 부적합 내용, 중복 문장 등 향후 문장 번역 진행 시에 문제가 될 수 있는 요소들을 사전에 확인하고 제거함으로써 보다 고품질의 말뭉치 데이터를 구축할 수 있는 기반을 마련하였다.



〈그림 28〉 아케이드 내 한국어 구어 데이터 검수

- 본 단계는 자체 구축 구어체에 해당하는 것으로 시나리오 구어체의 경우 본 단계를 거치지 않았다.

3) 일반인 수준 정제

- 일반인 수준 정제는 총 2차로 진행되었다.
- 1차 일반인 수준 정제는 클라우드소싱을 활용한 아케이드 내 정제가 종료된 이후, (주)플리토에서 별도 고용한 검수자들(일반 수준의 모국어 화자: 프리랜서)을 활용하여 전체 데이터 검수와 교정을 수행하였다.
- 검수자들은 엑셀 형식으로 작업 할당량을 지정 받았으며, 플리토는 검수·교정 업무 수행을 위해 아래 예시와 같은 지침(가이드라인)을 전달하였다.

2. 문장 검수 가이드

※ 문장에 오류가 있는 경우, 문장을 수정 및 삭제해 주셔야 합니다.

0. 수정 사항이 없는 올바른 문장

→ 수정하지 않고 G열 수정 여부 / H열 삭제 여부에 0가 유지될 수 있도록 그대로 두시면 됩니다.

A	B	C	D	E	F	G	H	I
sid	대분류	중분류	소분류	source	sentence	수정 여부	삭제 여부	비고
165201	일상	관광	쇼핑	먹는 길에 유기농 주스로 할까?	먹는 길에 유기농 주스로 할까?	0	0	

1. 문장 검수 시 수정되어야 하는 문장

→ 해당 상황에 맞게 수정 후 G열 수정여부에 1을, 수정된 문장은 F열에 기입해주시면 됩니다.

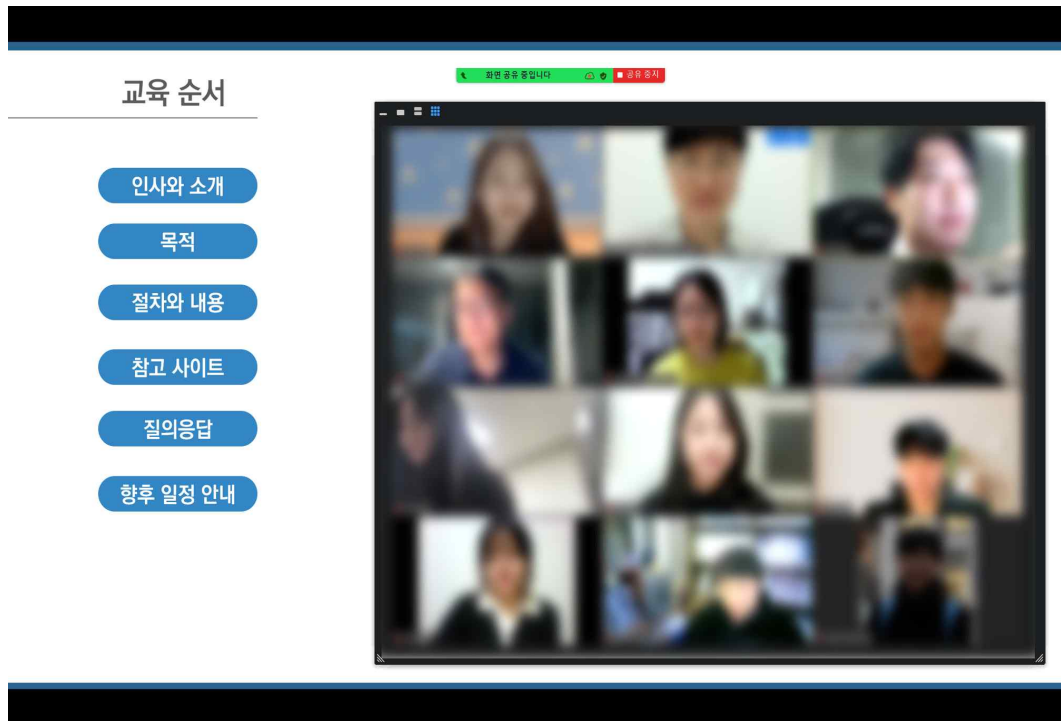
A	B	C	D	E	F	G	H	I
sid	대분류	중분류	소분류	source	sentence	수정 여부	삭제 여부	비고
165207	일상	일상	의료	평소에 드시는 것보다 조금만 더 섭취하시면 되요.	평소에 드시는 것보다 조금만 더 섭취하시면 <u>돼요</u> .	1	0	

- 수정이 필요한 상황

상황	수정 방법	예시	주의
한 셀에 문장이 2개 이상 있는 경우	1. 한 문장으로 수정 or 불가한 경우 2. 한 문장만 놔두고 나머지 부분은 삭제	(전) 응응 곧 경기시작하겠다 얼른 앉아서 보자. (후) 곧 경기 시작할거 같으니까 얼른 앉아서 보자. (전) 동양공이 제일 많이 나가요 대표음식중에 하나예요. (후) 대표 음식 중 하나인 동양공이 제일 많이 나가요. (전) 워낙 인기상품이라 제가 재고 파악해드릴게요 잠시만요. (후) 워낙 인기 상품이라 제가 재고 파악해드릴게요.	한 셀에는 반드시 한 문장이 있어야 합니다.
맞춤법 오류가 있는 경우 (띄어쓰기, 오타 포함)	한국어 맞춤법에 맞도록 수정	(전) 맞아 요새 두번씩 하게 된당. (후) 요새 두 번씩 하게 된다. (전) 휴가때 많이 먹고 잤나게 놀자구! (후) 휴가때 많이 먹고 재미나게 놀자고!	자주 틀리는 맞춤법 가이드 참고

〈그림 29〉 일반인 수준 1차 한국어 구어체 원문 정제 가이드라인

- 기계적, 플랫폼 총 2차례에 걸쳐 정제된 데이터는 본 검수 과정에서 부적절한 표현, 오타자, 개인 정보 포함, 비속어, 띄어쓰기, 기계학습에 부적합한 문장, 기준을 벗어난 어절 수, 비문 등이 추가로 정제되었으며 플리토는 총 3번의 정제 작업을 통해 원문 데이터 품질을 향상시켰다.
- 2차 일반인 수준 정제는 플리토에서의 정제가 종료된 후, (사)국제한국어교육학회의 보조연구원(한국어학 및 언어학 학·석사)이 전수 검토하였다.
- 원문의 품질 향상을 위해 (사)국제한국어교육학회에서는 원문 정제 원칙을 수립하고, 2회 이상의 원문 정제 교육을 실시하여 원문 정제의 품질을 균등하게 하였다.



〈그림 30〉 한국어 원문 정제 지침 교육 캡처본

○ 원문 정제 지침의 주요 내용은 아래와 같다.

- 원문 정제의 목적은 원문의 실제성을 최대한 보존하면서, 원어민 번역사가 원문의 의미를 파악하고 번역하기에 적합한 형태로 데이터를 교정 또는 처리하는 것이다.
- 한글맞춤법에 따라 탈자 및 오자, 띄어쓰기를 수정하는 것이 원칙이나, 구어의 경우 입말의 특성을 살리는 방향으로 수정하였다. 예를 들어 한국어 모어 화자들의 구어에서 주로 쓰이는 ‘바래요’, ‘할려고’ 등과 같은 표현은 입말 특성을 반영하는 측면에서 교정하지 않았다.
- 그 외에 비문, 사실과 다른 내용이 적시된 경우, 부적합 내용 및 개인 정보가 포함되어 있는 경우, 기준 어절 수 미달의 문장 등은 내용을 수정하거나 삭제 처리를 하였다.
- 구어의 경우 소주제 분류를 확인하여 주제 분류의 오류가 없는지도 점검하였다.

2021 한국어-외국어 병렬 말뭉치 구축 사업을 위한 한국어 구어체 원문 정제 지침

○ 목적

- 원문의 실제성을 최대한 보존하면서, 원어민 번역사가 원문의 의미를 파악하고 번역하기에 적합한 형태로 데이터를 교정 또는 처리한다.

○ 세부 지침

1. 원문 수집 업체(플리토)의 1차 정제

정제 기준	내용
어절 수	<구어체> 문장당 평균 5어절
	<문어체> 문장당 평균 15어절
부적합 내용	부적절한 내용, 문장 내 자연스럽게 읽히지 않는 어투, 불필요한 발화, 이모티콘 등의 처리, 오류가 있는 문장(오타자 포함)
개인 정보	인명, 전화번호, 이메일주소, 상표명 등의 정보 비식별화 및 삭제
중복 내용	완전 일치 문장의 제거

2. 내용 검수팀의 2차 정제

1) 한글맞춤법에 따른 교정

① 탈자: 맥락을 고려하여 탈자가 존재한다고 판단되는 경우 교정함.

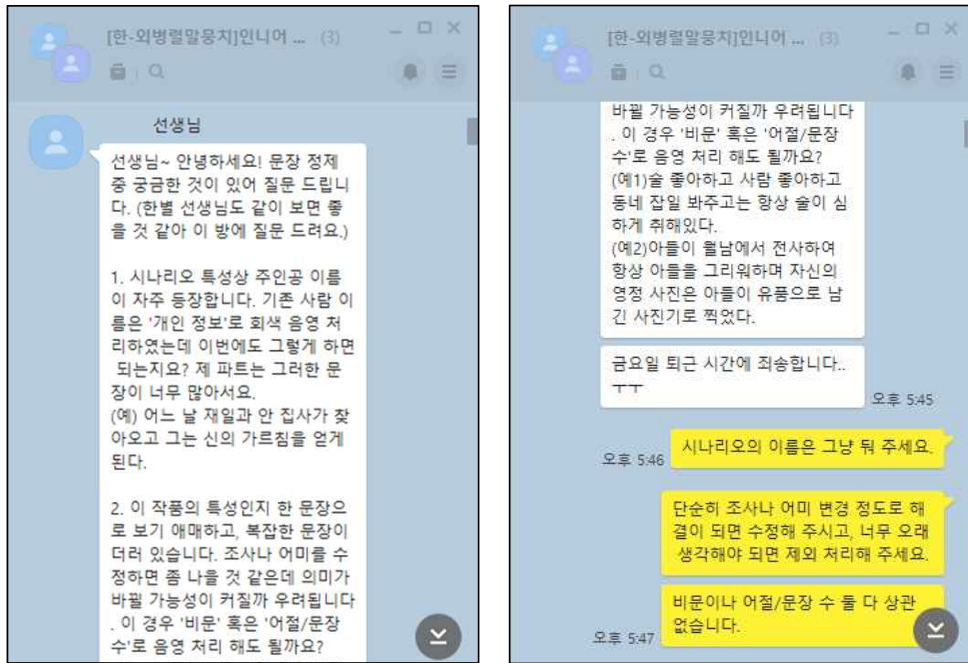
- 구어 원문: 구어는 문장 성분의 생략이 잦은 특징을 고려하여 탈자 여부를 판단해야 함.

② 오자: 문맥상 오자가 발견되면, 한글맞춤법에 따라 교정함.

- 한글맞춤법에서는 표준어를 소리대로 적되 어법에 맞도록 적도록 함. 즉, 표준어를 대상으로 그 형태소의 본모양을 밝혀 적는 것을 말함. 예컨대 '할께요'처럼 한글맞

〈그림 31〉 일반인 수준 2차 정제 한국어 구어체 지침

- 이때 전문가 수준의 전임 연구원(국어학, 한국어학, 번역학 박사 수료 이상)이 일반인 수준의 보조연구원과 팀을 이루어 실시간으로 원문 정제에 대한 질의가 가능하도록 하였다.



〈그림 32〉 일반인 수준, 전문가 수준 원문 정제 관련 질의 예시

〈표 22〉 2차 일반인 수준 한국어 구어체 원문 수정 예시

원문	수정 원문	비고
○○이 벌써부터 학원에 다니고 있어?	벌써부터 학원에 다니고 있어?	개인 정보 비식별화 이후 비정제된 부분 수정
245는 클 거 같은데 처분 색상으로 240 주문 가능할까요?	245는 클 거 같은데 처음 본 색상으로 240 주문 가능할까요?	단순 오타자 및 비문 수정
7시면 일찍 서둘러야 겠네 요.	7시면 일찍 서둘러야겠네 요.	띄어쓰기 수정

4) 전문가 수준 정제

- 두 차례로 실시한 일반인 수준의 원문 전수 정제 이후 국어학, 한국어학, 번역학 박사 수료 이상의 전문가 수준 전임 연구원이 전수 원문 검토를 실시하였다. 이때 오류율이 높은 보조연구원(일반인 수준)을 대상으로 원문 지침에 대한 재교육을 실시하였으며, 자주 발생하는 오류를 정리하여 재안내하였다.

〈표 23〉 자주 발생하는 한국어 정제 오류 예시

유형	예시																															
맞춤법	* 자주 틀리는 맞춤법 예시 <table><tr><td>X</td><td>O</td></tr><tr><td>- 되요</td><td>돼요</td></tr><tr><td>- 아니예요</td><td>아니예요</td></tr><tr><td>- '명사'이에요</td><td>'명사'이에요</td></tr><tr><td>- '명사'예요</td><td>'명사'예요</td></tr><tr><td>- (으)ㄹ 거예요</td><td>(으)ㄹ 거예요</td></tr><tr><td>- 드릴께여</td><td>드릴게요</td></tr><tr><td>- (으)ㄹ꺼 같네요</td><td>-(으)ㄹ 거 같네요</td></tr><tr><td>- -(으)ㄴ데요</td><td>-(으)ㄴ 대요(간접 인용)</td></tr><tr><td>- -는게</td><td>-는 게</td></tr></table>		X	O	- 되요	돼요	- 아니예요	아니예요	- '명사'이에요	'명사'이에요	- '명사'예요	'명사'예요	- (으)ㄹ 거예요	(으)ㄹ 거예요	- 드릴께여	드릴게요	- (으)ㄹ꺼 같네요	-(으)ㄹ 거 같네요	- -(으)ㄴ데요	-(으)ㄴ 대요(간접 인용)	- -는게	-는 게										
X	O																															
- 되요	돼요																															
- 아니예요	아니예요																															
- '명사'이에요	'명사'이에요																															
- '명사'예요	'명사'예요																															
- (으)ㄹ 거예요	(으)ㄹ 거예요																															
- 드릴께여	드릴게요																															
- (으)ㄹ꺼 같네요	-(으)ㄹ 거 같네요																															
- -(으)ㄴ데요	-(으)ㄴ 대요(간접 인용)																															
- -는게	-는 게																															
띄어쓰기	<table><tr><td>- '명사' 이에요</td><td>'명사'이에요</td></tr><tr><td>- '명사'입니다</td><td>'명사'입니다</td></tr><tr><td>- 지난 주</td><td>지난주, 지난번, 지난날, 지난달, 지난해, 지난밤</td></tr><tr><td>- 이번주</td><td>이번 주</td></tr><tr><td>- 다음주</td><td>다음 주</td></tr><tr><td>- 다음 날</td><td>다음날, 다음번</td></tr><tr><td>- 이 때</td><td>이때, 그때, 이날, 그날, 이거, 그거, 저거, 이것저것, 여기저기</td></tr><tr><td>- 첫 날</td><td>첫날</td></tr><tr><td>- 김씨</td><td>김 씨</td></tr><tr><td>- 백원</td><td>백 원, 100원</td></tr><tr><td>- 아무 것</td><td>아무것, 아무거</td></tr><tr><td>- 가져다 주다</td><td>가져다주다, 가져다드리다</td></tr><tr><td>- 문제 없다</td><td>문제없다, 상관없다, 관계없다, 다시없다</td></tr><tr><td>- 알아 보다</td><td>알아보다, 물어보다, 살펴보다, 지켜보다, 찾아보다, 찾아오다</td></tr><tr><td>- 오래 가다</td><td>오래가다</td></tr></table>		- '명사' 이에요	'명사'이에요	- '명사'입니다	'명사'입니다	- 지난 주	지난주, 지난번, 지난날, 지난달, 지난해, 지난밤	- 이번주	이번 주	- 다음주	다음 주	- 다음 날	다음날, 다음번	- 이 때	이때, 그때, 이날, 그날, 이거, 그거, 저거, 이것저것, 여기저기	- 첫 날	첫날	- 김씨	김 씨	- 백원	백 원, 100원	- 아무 것	아무것, 아무거	- 가져다 주다	가져다주다, 가져다드리다	- 문제 없다	문제없다, 상관없다, 관계없다, 다시없다	- 알아 보다	알아보다, 물어보다, 살펴보다, 지켜보다, 찾아보다, 찾아오다	- 오래 가다	오래가다
- '명사' 이에요	'명사'이에요																															
- '명사'입니다	'명사'입니다																															
- 지난 주	지난주, 지난번, 지난날, 지난달, 지난해, 지난밤																															
- 이번주	이번 주																															
- 다음주	다음 주																															
- 다음 날	다음날, 다음번																															
- 이 때	이때, 그때, 이날, 그날, 이거, 그거, 저거, 이것저것, 여기저기																															
- 첫 날	첫날																															
- 김씨	김 씨																															
- 백원	백 원, 100원																															
- 아무 것	아무것, 아무거																															
- 가져다 주다	가져다주다, 가져다드리다																															
- 문제 없다	문제없다, 상관없다, 관계없다, 다시없다																															
- 알아 보다	알아보다, 물어보다, 살펴보다, 지켜보다, 찾아보다, 찾아오다																															
- 오래 가다	오래가다																															

- 2차 전수 검사 시에는 원문의 형식적 측면뿐만 아니라 내용적 측면도 검토하여 원문의 의미를 명확하게 하였으며, 의미가 유사한 문장이 다수 있는 경우 내용을 수정하여 중복되는 문장이 없도록 하였다.

〈표 24〉 전문가 수준 한국어 구어체 원문 정제 예시

원문	수정 원문	비고
벌써부터 학원에 다니고 있어?	다섯 살짜리가 벌써 학원에 다녀요?	1차 수정 이후 원문을 명확히 수정
7시면 일찍 서둘러야겠네요.	탐승 시간이 7시면 일찍 서둘러야겠네요.	1차 수정 이후 원문을 명확히 수정
난 여기 기다리고 있을게.	난 여기에서 기다리고 있을 테니까 다녀와.	유사 문장 수정
그럼 난 여기 기다리고 있을게.	그럼 난 여기서 기다리고 있을게.	
그러면 나는 여기서 기다리고 있을게.	그러면 나는 저기서 구경하고 있을게.	

다. 한국어 문어체 원문 정제

1) 기계적 정제

- 구매를 통해 구축된 원문 데이터는 인력 정제 이전에 어절 수, 부적합 내용, 개인 정보와 같이 기계적으로 정제 가능한 요소를 선별하여 진행하였다.

〈표 25〉 한국어 문어체 기계적 정제 내용

정제 기준	내용
어절 수	〈문어체〉 문장당 평균 약 15어절 (최솟값 10 ~ 최댓값 30)
부적합 내용	부적절한 내용, 문장 내 자연스럽지 않은 어투, 불필요 발화, 이모티콘 등의 처리, 오류가 있는 문장(오타자 포함)
개인 정보	인명, 전화번호, 이메일주소, 상표명 등의 정보 비식별화 및 삭제
중복 내용	완전 일치 문장의 제거

2) 일반인 수준 정제

- 일반인 수준의 정제는 구어와 동일하게 2차로 진행되었다.
- 기계적 정제가 종료된 후, 전체 데이터를 플리토가 별도 고용한 검수자들(일반 수준의 모국어 화자 : 프리랜서)을 통해 엑셀 형식으로 작업 할당량을 지정한 후 지침에 따라 검수·교정을 지시하였다. 일반인 수준 1차 정제 작업을 수행하는 작업자에게 제공되는 검수 가이드는 아래의 예시와 같다.

검수 방법 ● 업무 : 한국어 원문을 확인한 후 문장 수정 및 삭제 ● 삭제 : J열의 드롭박스 내용을 선택한 후 C열 데이터 삭제 1/ 특수문자가 포함된 문장 (e.g. 일본 노선의 경우 인천발 △도쿄 △오사카 △후쿠오카 △미야자키 등을 다음달 말 까지 운휴한다. 2/ 미완성 문장 (문장 끊김) 3/ 부적절한 내용 (차별, 성적인 표현 등) ● 수정 : J열의 드롭박스 내용을 선택한 후, C열 데이터 수정 및 셀 색상 변경 1/ 띄어쓰기 오류 2/ 맞춤법 오류 3/ 구두점 누락
--

〈그림 33〉 한국어 문어체 원문 정제 가이드라인

- 본 검수에서는 데이터 확보 과정에서 발생하는 오류 표기 및 기계학습에 부적절한 문장 등 수정·제거하였다.

- 2차 일반인 수준 정제는 플리토에서의 정제가 종료된 후, (사)국제한국어교육학회의 보조연구원(한국어학 및 언어학 학·석사)이 전수 검토하였다.
- 원문의 품질 향상을 위해서 (사)국제한국어교육학회에서는 원문 정제 원칙을 수립하고, 2회 이상의 원문 정제 교육을 실시하여 원문 정제의 품질을 균등하게 하였다.
- 원문 정제 지침의 주요 내용은 아래와 같다.
 - 원문 정제의 목적은 원문의 실제성을 최대한 보존하면서, 원어민 번역사가 원문의 의미를 파악하고 번역하기에 적합한 형태로 데이터를 교정 또는 처리하는 것이다.
 - 구어 정제와 달리 한글맞춤법에 따라 탈자 및 오자, 띄어쓰기 등을 엄격히 정제한다. 단어의 띄어쓰기는 우리말샘 사전을 기준으로 한다.
 - 그 외에 비문, 사실과 다른 내용이 적시된 경우, 부적합 내용 및 개인 정보가 포함되어 있는 경우, 기준 어절 수 미달의 문장 등은 내용을 수정하거나 삭제 처리를 하였다.

2021 한국어-외국어 병렬 말뭉치 구축 사업을 위한 한국어 원문 정제 지침

○ 목적

- 원문의 실제성을 최대한 보존하면서, 원어인 번역사가 원문의 의미를 파악하고 번역하기에 적합한 형태로 데이터를 교정 또는 처리한다.

○ 세부 지침

1. 원문 수집 업체(플리토)의 1차 정제

정제 기준	내용
어절 수	<구어체> 문장당 평균 5어절 <문어체> 문장당 평균 15어절
부적합 내용	부적절한 내용, 문장 내 자연스럽지 않은 어투, 불필요한 발화, 이모티콘 등의 처리, 오류가 있는 문장(오탈자 포함)
개인 정보	인명, 전화번호, 이메일주소, 상표명 등의 정보 비식별화 및 삭제
중복 내용	완전 일치 문장의 제거

2. 내용 검수팀의 2차 정제

1) 한글맞춤법에 따른 교정

- ① 탈자: 맥락을 고려하여 탈자가 존재한다고 판단되는 경우 교정함.
- 문어 원문: 문어는 내용이 많고 구조가 복잡하므로 번역 편의를 위해 적극적으로 탈자를 보완할 필요가 있음.
 - * 문단의 단어 일부가 편집상의 오류로 누락된 경우에는 원문을 검색하여 추가함.
예) (원문) 부는 국산 터빈 사용을 권장하고 있으나 한전은 지멘스 등 외산도 고려하고 있어 입찰 방식이 정해지지 않아서다.
→ (수정) 정부는 국산 터빈 사용을 권장하고 있으나 한전은 지멘스 등 외산도 고려하고 있어 입찰 방식이 정해지지 않아서다.
- ② 오자: 문맥상 오자가 발견되면, 한글맞춤법에 따라 교정함.
- 한글맞춤법에서는 표준어를 소리대로 적되 어법에 맞도록 적도록 함. 즉, 표준어를 대상으로 그 형태소의 본모양을 밝혀 적는 것을 말함. 예컨대 '할게요'처럼 한글맞춤법에 어긋나게 적은 오자를 '할게요'로 교정하는 작업을 말함.
 - 구어 원문: 구어의 특성을 살리는 방향으로 교정함. 예를 들어 한국어 모어 화자들의 구어에서 주로 쓰이는 '바래요', '할려고' 등과 같은 표현은 입말의 특성을 반영

〈그림 34〉 일반인 수준 2차 정제 한국어 문어체 지침 예시

- 이때 전문가 수준의 전임 연구원(국어학, 한국어학, 번역학 박사 수료 이상)이 일반인 수준의 보조연구원과 팀을 이루어 실시간으로 원문 정제에 대한 질의가 가능하도록 하였으며, 기사 내용의 사실 확인이 필요할 경우 해당 기사문을 검색하여 원문 확인 후 정제를 실시하였다

〈표 26〉 2차 일반인 수준 한국어 문어체 원문 수정 예시

원문	수정 원문	비고
<u>무리한 사업 확대에 따른 당연한 적자</u> 라는 지적과 '성장 동력 확보를 위한 일시적 적자'라는 의견 등 실적 부진 배경을 놓고 반응이 엇갈리고 있다.	<u>'무리한 사업 확대에 따른 당연한 적자'</u> 라는 지적과 '성장 동력 확보를 위한 일시적 적자'라는 의견 등 실적 부진 배경을 놓고 반응이 엇갈리고 있다.	문장 부호 오류 수정
이어 그는 " <u>다당</u> 지역의 새로운 일자리 창출과 지역사회공헌에도 적극 나설 방침"이라고 덧붙였다.	이어 그는 " <u>다낭</u> 지역의 새로운 일자리 창출과 지역 사회 공헌에도 적극 나설 방침"이라고 덧붙였다.	단순 오타자 수정
강진이 고개를 숙인 <u>체</u> 눈물 <u>이</u> 똑 똑 떨어진다.	강진이 고개를 숙인 <u>채</u> 눈물 <u>을</u> 똑 똑 흘린다.	단순 오타자 및 비문 수정

4) 전문가 수준 정제

- 두 차례로 실시한 일반인 수준의 원문 전수 정제 이후 국어학, 한국어학, 번역학 박사 수료 이상의 전문가 수준 전임 연구원이 전수 원문 검토를 실시하였다. 이때 오류율이 높은 보조연구원(일반인 수준)을 대상으로 원문 지침에 대한 재교육을 실시하였으며, 자주 발생하는 오류를 정리하여 재안내하였다.
- 2차 전수 검사 시에는 원문의 형식적 측면뿐만 아니라 내용적 측면도 검토하여 원문의 의미를 명확하게 하였으며, 신문의 경우 기사 내용의 진위 여부를 파악하여 정제하였다.

〈표 27〉 전문가 수준 한국어 문어체 원문 정제 예시

원문	수정 원문	비고
<u>자이카산은</u> 이번에 조달한 자금으로 인력 확충과 기술 개선에 투자하고 진출 지역을 확대할 계획이다.	<u>자이카산은</u> 이번에 조달한 자금으로 인력 확충과 기술 개선에 투자하고 진출 지역을 확대할 계획이다.	회사명 수정
이재용 삼성전자 부회장 파기환송심 재판부 " <u>총수가 무서워할 정도의 실효적인 준법감시제도가 있었다면 이번 사건은 일어나지 않았다.</u> "	이재용 삼성전자 부회장 파기환송심 재판부 " <u>총수가 무서워할 정도의 실효적인 준법감시제도가 있었다면 이번 사건은 일어나지 않았다.</u> "고 말했다.	문장 부호 및 비문 수정
1일(현지시간) 미래에셋대우 베트남법인은 자본금 규모를 기존 4조 3000억 동(약 2,200억 원)에서 <u>5455억 5000만 동</u> (약 2,840억 원)으로 확대한다고 밝혔다.	1일(현지 시간) 미래에셋대우 베트남법인은 자본금 규모를 기존 4조 3000억 동(약 2,200억 원)에서 <u>4조 5455억 5000만 동</u> (약 2,840억 원)으로 확대한다고 밝혔다.	내용 진위 여부 파악 후 수정
<u>재홍은 오 부장</u> 고개를 들고 올라가는 엘리베이터 숫자만을 바라본다.	<u>재홍과 오 부장은</u> 고개를 들고 올라가는 엘리베이터 숫자만을 바라본다.	내용을 명확하게 수정

라. 추가 정제

- 3차 또는 4차에 걸쳐 전수 검사가 완료된 데이터는 번역 및 검수가 되었으며, 검수 시 번역 검수원들은 원문을 확인하여 원문에 오류가 있는 경우 ‘원문 신고’ 기능을 활용하여 원문 신고를 하였다.

OriginText	Translation	QC
단무지 빼고 참치 많이 맛있게 싸주세요. 🍱️	Tolong bungkus yang banyak tuna yang enak tanpa acar lobak	  

〈그림 35〉 플랫폼의 원문 신고 기능

- 이때 신고된 원문 오류의 유형은 단순 오타자가 포함된 원문, 의미가 불분명한 원문 등이 있었다.
- 신고된 원문은 전문가 수준의 전임 연구원이 재검토하여 적절한 문장으로 수정하였으며, 재수정된 문장을 기준으로 재번역 및 검수를 실시하였다. 이때 기존의 원문과 구별하여 플랫폼에 업로드하여 추가 원문 수정이 용이하도록 하였다.
- 이와 동시에 보조 연구원 및 전임 연구원이 검수된 데이터의 원문을 다시 검토하며 원문 오류가 없는지 교차 검토하였으며, 검수 및 감수가 진행된 데이터 중 원문 오류가 발견된 경우, 이를 수정하여 재검수 및 감수를 진행하였다.
 - 검수된 데이터의 원문 검토 시 전체 언어권 담당자들이 볼 수 있는 구글 스프레드시트를 활용하여 원문 오류에 대해 사업 참여자 모두가 공유할 수 있도록 하였다.

원문 오류				
No.	일시	원문 오류	원문	내용
316	2022.02.04	원문 오류	러시아 국민 라면'으로 불리는 '팔도 도시락'이 유해성분 검출 논란에도 뜨거운 사랑을 받고 있다.	따옴표 없음
317	2022.02.04	원문 오류	파우더 루주 틴트'는 고밀도 피그먼트 가죽처럼 빈틈없이 채워주는 무광 틴트로, 부드럽고 탄력적인 텍스처로 각질은 눌러주면서 파우더리하게 피니시 되는 제품이다.	따옴표 없음
318	2022.02.04	원문 오류	빈티지 필터 예디션'은 최근 Z세대 사이에서 인기가 재현되는 필름 사진만의 부드러운 색상과 빈티지 느낌에서 영감을 얻어 만들어진 메이크업 컬렉션이다.	따옴표 없음
319	2022.02.04	원문 오류	그래서 내가 이렇게 현소장한테 도와 달라고 왔잖아.	현 소장
320	2022.02.04	원문 오류	나 오늘 사실 현소장 만나러 왔어요.	현 소장
321	2022.02.04	원문 오류	아니 고사장 어떤 일로 이른 아침부터 가게를 다 여시오?	고 사장
322	2022.02.04	원문 오류	너도 알아서 오른 선택을 하길 바란다.	옳은
323	2022.02.04	원문 오류	마누라가 밥 먹으라고 계속 전화와.	전화 와
324	2022.02.04	원문 오류	내가 속이다 몽크려 저서 영혼이 보러 가는 거.	몽크려저서
325	2022.02.04	원문 오류	아무도 방문하지 않는 숙소의 성을 찾아가는 팩 보그를 통해 '에이본 레이디'의 직업의식을 담고 있다.	에이본
326	2022.02.04	원문 오류	미국 리서치 컨설팅 기업' 프로스트 앤 설리번은 미국 뇌전증 시장이 2024년까지 약 41억 달러(약 5조280억 원)까지 성장할 것으로 전망했다.	따옴표 위치??
327	2022.02.04	원문 오류	현대중공업과 대우조선이 최종 인수합병에 성공하려면 한국 외 해외 경쟁 당국의 기업결합심사를 통과해야만 한다.	기업결합심사
328	2022.02.04	원문 오류	미진의 아버지가는 한참 뜬다가 표정이 굳는다.	가 삭제
329	2022.02.05	원문 오류	남성이 왼팔을 들어 정수를 저지하려 할 때 성의의 왼 다리가 남성을 막는다.	남성을
330	2022.02.05	원문 오류	부검 실 밖 밤 웅성거리는 소리와 기자들이 물려든다.	부검실 물려쓰기
331	2022.02.05	원문 오류	2008년 베트남 호찌민에 최초 멀티플렉스 세운 후 지난 3월 기준 33개 극장, 150개 스크린을 운영해왔다.	멀티플렉스를
332	2022.02.05	원문 오류	롯데엔터테인먼트는 1,400만 관객을 모은 '신과 함께'와 '조선 명탐정: 흙혈 괴마의 비밀' 등을 올 베트남 현지에서 개봉했다.	을' 삭제
333	2022.02.05	원문 오류	얼티문 파워 인튜징 컨센트레이트 세럼'은 도시 유해환경과 피부 스트레스로부터 피부 장벽을 탄탄하게	따옴표 없음

〈그림 36〉 검수 후 발견된 원문 오류 공유 시트

- 추가로 재수정하여 플랫폼에 업로드하였으나, 다시 원문 오류 신고가 된 경우 전임 연구원의 검토 및 각 언어별 품질 관리팀장(국어학 및 한국어학, 해당 언어 박사)의 자문을 거쳐 문장을 재정제하였다.

마. 베트남어 원문 정제

- 베트남어 원문 또한 원문의 특성에 따라 최대 4단계 원문 정제를 실시하였다.
- 아케이드 내에서 구축한 베트남어 구어체의 경우 아케이드 사용자들이 구축한 것으로 오타자 및 내용 중복이 많기 때문에 4단계 정제를 실시하였다.
- 각 단계는 별도로 진행되는 것이 아니라 순환식으로 진행되었으며, 번역 및 검수, 감수 완료까지 지속적으로 진행되었다.

〈표 28〉 베트남어 원문별 정제 절차

어체	매체	1단계	2단계	3단계	4단계
구어체	자체 구축	기계적 정제	아케이드 내 정제	내부 언어 전문가 정제	전문가 수준 정제
문어체	신문	기계적 정제	내부 언어 전문가 정제	전문가 수준 정제	-

- 베트남어 구어체 및 문어체는 1단계 정제 시 한국어 원문과 동일하게 인력 정제 이전에 개인 정보, 중복 내용과 같이 기계적으로 정제 가능 요소들을 선별하여 작업을 진행하였다.

〈표 29〉 베트남어 구어체 기계적 정제 내용

정제 기준	내용
개인 정보	인명, 전화번호, 이메일 주소, 상표명 등의 정보 비식별화 및 삭제
중복 내용	완전 일치 문장의 제거

- 기계적 정제를 마친 베트남어 구어체 또한 한국어 구어체와 동일하게 2단계에서 아케이드 내 정제 작업을 실시하였다.
- '아케이드'에는 구축된 구어체 원문 데이터와 정제를 위한 가이드라인이 등록되어 있으며, 플랫폼 유저들(검수자들)이 확인하여 구축된 문장의 적합 및 부적합 여부를 선택할 수 있는 환경을 조성하였다.

- 오타자, 부적합 내용, 중복 문장 등 향후 문장 번역 진행 시에 문제가 될 수 있는 요소들을 사전에 확인하고 제거함으로써 보다 고품질의 말뭉치 데이터를 구축할 수 있는 기반을 마련하였다.

OP

số liệu thống kê
 Lịch sử tham gia
 hướng dẫn

Tiếng Việt

5P

Hãy đánh giá câu trả lời.

Chủ đề: Tiếp tục cuộc trò chuyện giữa nhân viên trang web đại lý thu mua nước ngoài và khách hàng.

☞ Phí đại lý mua hàng ở mỗi quốc gia có khác nhau không?

không! Mỗi sản phẩm là khác nhau.

không giống nhau

sự thích hợp

Nếu câu chứa thông tin cá nhân, nội dung vô nghĩa (bao gồm lỗi chính tả) hoặc khuyến khích xung đột (bao gồm cả ngôn từ tục tĩu), vui lòng xếp hạng câu đó là không phù hợp.
Không thể thực hiện chỉnh sửa sau khi gửi.

nhảy

nộp hồ sơ

〈그림 37〉 아케이드 내 베트남어 구어 데이터 검수

- 베트남어 구어체 3단계, 문어체 2단계 정제 단계에서는 (주)플리토 내 언어 전문가 중 베트남어 모어 화자가 구어체 및 문어체 가이드라인을 숙지한 후 정제를 실시하였다.

4. Câu hỏi thường gặp

1. Hiện tại tôi đang làm việc tại công ty khác, vậy tôi có thể đồng thời làm phiên dịch/dánh giá cho cùng dự án này không?

- Flitto không thấy phiền khi người đang làm việc tại công ty khác tham gia vào dự án, nhưng vui lòng hỏi ý kiến công ty trực thuộc về việc có thể làm đồng thời dưới hình thức tự do hay không.

2. Những câu chứa tên người có phải chỉnh sửa vì bao gồm thông tin cá nhân hay không?

- Không. Thông tin cá nhân là những nội dung liên quan đến sự tồn tại của một cá nhân, có thể nhận biết cá nhân đó dựa vào họ tên, số chứng minh thư, nếu chỉ đơn thuần chứa tên thì không cần chỉnh sửa.

(Dù không thể nhận biết cá nhân nào đó chỉ dựa vào thông tin đã đề cập, nhưng nếu có thể nhận biết nhờ kết hợp với các thông tin khác thì vẫn được coi là thông tin cá nhân)

Ví dụ 1: Jingu của chúng ta giỏi cả chạy bộ nữa! → Không cần sửa

Ví dụ 2: Tôi là Choi Seonjae sống tại quận Yeoksam. → Tôi là *** sống tại quận Yeoksam

Ví dụ 3: Địa chỉ email là btog@flitto.com. → Địa chỉ email là ****@*****.***.

3. Có thể tùy ý chỉnh sửa câu không?

- Có thể. Trường hợp những câu lược bỏ tân ngữ hay chủ ngữ, vị ngữ gây khó hiểu, có thể xóa hoặc sửa câu gốc thành câu hoàn chỉnh mà không làm thay đổi câu quá nhiều.

Ví dụ 1: Tôi vốn làm vừa đúng vừa nhanh! → Có thể sửa thành câu hoàn chỉnh “Tôi vốn làm bài tập vừa đúng vừa nhanh!”

Ví dụ 2: Bạn định đánh kiểu đấy à? → Có thể sửa thành nhiều trường hợp, chẳng hạn như “Bạn định đánh Piano kiểu đấy à?”

Nếu có thắc mắc nào khác ngoài các trường hợp trên đây, vui lòng liên hệ qua email của người phụ trách (soyeun.park@flitto.com).

〈그림 38〉 베트남어 구어체 원문 정제 가이드라인

0. 수정 사항이 없는 올바른 문장

→ 수정하지 않고 C열은 빈칸으로 둡니다.

A	B	C	D	E
SID	sentence	수정	이유	2nd sentence
2258	VSAK cũng đã tổ chức thành công các đợt quyền góp giúp đồng bào trong nước gặp nạn do thiên tai; cũng như vận động đóng góp cho biển đảo Tổ quốc; tổ chức điều hành ở n hóa phân đối sự bành trướng phi pháp của Trung Quốc tại Biển Đông.	빈칸		

1. 문장 검수 시 수정되어야 하는 문장

→ 해당 경우에 맞게 C열에 수정된 문장을 입력 후 D열에 수정된 이유를 쓰면 됩니다.

① 문장에 오타, 띄어쓰기 오류 등 맞춤법 오류가 있는 경우 → 맞춤법에 맞게 수정

A	B	C	D	E
SID	sentence	수정	이유	2nd sentence
12057/28/2	Theo Yonhap, Tổng thống Hàn Quốc Moon Jae-in hôm nay đã chủ trì cuộc họp lần đầu tiên trong 9 tháng qua với Hội đồng An ninh Quốc gia để thảo luận các bước tiếp theo sau khi thượng đỉnh Mỹ - Triều kết thúc tại Hà Nội hôm 28/2 mà không đạt được <u>thỏa thuận</u> .	Theo Yonhap, Tổng thống Hàn Quốc Moon Jae-in hôm nay đã chủ trì cuộc họp lần đầu tiên trong 9 tháng qua với Hội đồng An ninh Quốc gia để thảo luận các bước tiếp theo sau khi thượng đỉnh Mỹ - Triều kết thúc tại Hà Nội hôm 28/2 mà không đạt được <u>thỏa thuận</u> .	오타자 있음 (thảo thuận --> thỏa thuận)	

② 문장에 불필요한 정보나 낱파, 특수문자 등이 포함된 경우 → 필요 없는 부분만 삭제

(단, 문장에는 영향이 가지 않아야 함)

A	B	C	D	E
SID	sentence	수정	이유	2nd sentence
11915	<u>前</u> 시 히 vọng Tổng thống Mỹ Donald Trump và Chủ tịch Triều Tiên Kim Jong-un sẽ nhanh chóng đưa ra thông cáo thống nhất, chấm dứt chiến tranh, mang lại hòa bình cho thế giới cũng như Bán đảo Triều Tiên.	Tôi hi vọng Tổng thống Mỹ Donald Trump và Chủ tịch Triều Tiên Kim Jong-un sẽ nhanh chóng đưa ra thông cáo thống nhất, chấm dứt chiến tranh, mang lại hòa bình cho thế giới cũng như Bán đảo Triều Tiên.	필요없는 특수문자 삭제 (단, 문장에 영향이 가지 않아야 함)	
21866	<u>WASHINGTON</u> - Hôm Chủ nhật, Tạp chí phố Wall trích lời những người trong ngành rằng Chính quyền Donald Trump đang xem xét việc yêu cầu thiết bị di động thế hệ thứ năm được sử dụng ở Hoa Kỳ phải được sản xuất bên ngoài Trung Quốc.	Hôm chủ nhật, tạp chí phố Wall trích lời những người trong ngành rằng Chính quyền Donald Trump đang xem xét việc yêu cầu thiết bị di động thế hệ thứ năm được sử dụng ở Hoa Kỳ phải được sản xuất bên ngoài Trung Quốc.	필요없는 내용 삭제 (단, 문장에 영향이 가지 않아야 함)	

〈그림 39〉 베트남어 문어체 원문 정제 가이드라인

○ 이후 구어체 4단계, 문어체 3단계 전문가 수준 원문 정제 단계에서는 한국어와 베트남어에 모두 유창한 베트남 현지 교수가 원문 정제를 실시하였다.

○ 이때 검수 시 신고된 데이터에 대하여 원문 정제를 실시하였으며, 검수와 원문 정제가 동시에 이루어져 원문 오류에 대해 신속하게 대응할 수 있도록 하였다.

SII	베트남어_원문데이터	오류 내용
2039	Nhà nghiên cứu Lee nhận định, trong tài liệu có nhiều giai thoại cho thấy rằng chủ tịch Hồ Chí Minh đã tương tác rất chặt chẽ với các quan chức Hàn Quốc: "Hwang, một người Hàn Quốc, đã trao đổi thân mật với chủ tịch Hồ Chí Minh bằng tiếng Anh"	- 부적절한 어휘 선택: 'tương tác rất chặt chẽ'
2597	Các chuyên gia về phân phối lưu thông của Hyundai Glovis cũng sẽ tham gia tư vấn định kì cho dự án.	- 오타: 'vua'
2687	Ngoài ra, ngày 27 tại phố đi bộ Shinchon cũng sẽ diễn ra chương trình lễ hội với nhiều nội dung hấp dẫn như show thời trang, nhảy B-boys, DJ.	
2688	Lần lượt vào lúc 2 giờ và 4 giờ chiều ngày 27, 'Bữa tiệc thời trang sinh viên Seoul 365' lấy bối cảnh phố đi bộ Shinchon sẽ được diễn ra với 42 bộ trang phục mang phong cách tươi sáng và thoải mái, phù hợp với đặc điểm của các sinh viên.	- 불필요한 어휘: 'lần lượt', 'các'
2739	IZ*ONE là nhóm nữ debut với đội hình 12 thành viên được Mnet công bố tên ngay tại đêm chung kết của chương trình tuyển chọn thực tập sinh nổi tiếng Produce 48.	
2745	Vì thế, người hâm mộ càng thêm háo hức mong chờ màn debut của nhóm nữ được lựa chọn bởi chính họ.	
2746	Đặc biệt hơn, Wanna One - hiện tượng của Kpop mới đây cũng đã thông báo sẽ comeback vào tháng 11.	
2885	Mặc dù nhận được nhiều giải quan trọng nhưng "Thử Thách Thần Chết: Giữa Hai Thế Giới" không phải là tác phẩm nhận được giải thưởng Daesang.	- 어휘 중복: 'nhận được'
2918	Huawei đang dẫn đầu công nghệ mạng 5G nhờ là một công ty đa ngành nghề và từng có thâm niên sản xuất nhiều thiết bị viễn thông cho cả thế giới.	
2940	Hí vọng rằng IZ*ONE sẽ có đợt debut thành công rực rỡ và lập thêm nhiều thành tích hơn nữa trong tương lai.	
3086	Tại lần lần tổ chức thứ 39 năm nay, tác phẩm giành được nhiều đề cử nhất chính là "1987: When the Day Comes".	- 단어 중복: 'lần'
3089	Trước áp lực mạnh mẽ từ phía truyền thông, Bộ trưởng Bộ Cảnh sát Park Cheo-won buộc phải tổ chức họp báo và tuyên bố cậu sinh viên bị chết vì nhồi máu cơ tim do quá sợ hãi.	- 어휘 선택 및 문법: 'cậu sinh viên bị chết vì nhồi máu cơ tim do quá sợ hãi.'
3145	Hiệp hội Đầu bếp Hàn Quốc chi nhánh Kwangju-Jeonnam sẽ trao các giải thưởng giá trị cho tổng cộng 43 đội đ ược đánh giá cao trong cuộc thi lần này.	- 불필요한 어휘: 'tổng cộng'
3180	Hãy cùng chờ xem những sản phẩm âm nhạc trong tương lai của IU sẽ đạt được những thành tích như thế nào nhé!	- 주어 생략
3227	Trường Đại học Jeonnam sẽ dùng số tiền 100 triệu won do cụ bà Kim trao tặng để trao học bổng cho các sinh viên có thành tích xuất sắc trong học tập.	- 오타: 'đề'
3293	Cũng trong ngày hôm qua, Phát ngôn viên Nhà Xanh, bà Kim Eui-kyeom cho biết, các máy bay vận tải chở quýt đầu tiên đã rời đảo Jeju để đến Bình Nhưỡng.	

〈그림 40〉 베트남어 원문 전문가 수준 정제 예시

1.3. 개인 정보 비식별화

가. 개인 정보 비식별화 기준

〈표 30〉 개인 정보 비식별화 기준

구분	내용	예외
구어체	보통명사화되지 않은 기업명이나 상표명, 그리고 특정 개인 정보(연예인/개인 이름 등 표기)가 포함된 문장	- 보통명사화된 상품명 예: 불닭, 초코파이, 타이레놀 등 - 유명한(대기업, 글로벌기업) 회사명 예: 삼성, 엘지, 애플 등
문어체	심각한 경우가 아닌 경우 수정하지 않음	

나. 단계별 개인 정보 비식별화

1) 구축 단계 개인 정보 비식별화(아케이드를 통한 구어체 구축만 해당)

○ 참여 가이드 제공을 통한 개인 정보 비식별화

다음은 부적합한 문장의 기준입니다.

개인정보 포함, 무의미한 내용(오타자 포함), 갈등 조장 내용(비속어 포함)의 포함 유무는 다른 사용자들이 판단하며, 플리토는 이러한 결과를 변경할 수 없음을 알려드립니다.

개인정보 포함 : 주민번호, 전화번호 등의 개인을 식별할 수 있는 정보가 포함된 글

(예: 010-XXXX-XXXX, 서울시 강남구 XXX, 910327-XXXXXXX 등)

무의미한 내용(오타자 포함) : 무의미한 어휘 반복 등 대화가 이어지지 않는 글, 오타자가 포함된 글

〈그림 41〉 플리토 아케이드 원문 구축 참여 가이드

2) 정제 단계 개인 정보 비식별화

- 원문 데이터의 모국어 화자를 통한 정제 시, 가이드라인 제공을 통한 개인 정보 비식별화

문장에 개인정보(이름, 전화번호, 이메일 주소 등)가 포함되어 있는 경우	**를 사용하여 비식별화 (기존 문자수와 동일하게 작성)	(전) 제 이메일 주소는 btog@flitto.com 입니다. (후) 제 이메일 주소는 ****@*****.*** 입니다.
--	---------------------------------	---

〈그림 42〉 모국어 화자에게 제공되는 원문 정제 가이드라인

3) 전산 프로그램을 통한 개인 정보 비식별화

- 전산 프로그램을 이용한 기계적 처리를 통해 개인 정보 비식별화

```
function get_refine_Aa(refine_options)
  re_tel_corp = s-1[0-9]{3}-[1-9]{0-9}[3]*
  re_tel_land = s-0102-01[0-9]{3}-11-[0-9]{3,4}-[0-9]{4}*
  re_tel_mo = s-01[0-9]{3}-[0-9]{3,4}-[0-9]{4}*
  replace_phone_numbers_A(text) = replace(text, re_tel_corp => ".TEL.", re_tel_land => ".TEL.", re_tel_mo => ".TEL.")
  re_email = s-[a-zA-Z0-9]{1,64}@([a-zA-Z0-9]{1,64}-[a-zA-Z0-9]{1,64})\.[a-zA-Z]{2,6}
  replace_email_A(text) = replace(text, re_email => ".EMAIL.")

  clean_firstchar_A = function(text)
    match_1 = findfirst(x -> isletter(x) || isdigit(x) || x ∈ ['"', '\'', '(', ')', '-', '.', ':', ';', '!', '@', ' ', '(', ')', '[', ']', '0', '#', '0'], text)
    isnothing(match_1) ? text : text[match_1:end]
  end

  refine_Aa = []
  push!(refine_Aa, text -> ismissing(text) ? "" : text)
  push!(refine_Aa, text -> is(text, String) ? string(text) : string(text))
  push!(refine_Aa, remove_wrong_encoding_A)
  push!(refine_Aa, text -> isconv_r2h(text, kana=false, ascii=true, digit=true))
  push!(refine_Aa, text -> replace(text, ["\n", "\r", "\t"] -> " "))
  push!(refine_Aa, text -> join(split(text), " "))
  push!(refine_Aa, text -> replace(text, punct_max_map...))
  push!(refine_Aa, normalize_html_special_chars)
  push!(refine_Aa, clean_firstchar_A)
  if refine_options.is_upper_first
    push!(refine_Aa, uppercasefirst)
  end
  if refine_options.is_mask pii
    push!(refine_Aa, replace_email_A)
    push!(refine_Aa, text -> textacy_preprocessing.replace_url(text, "_URL_"))
    push!(refine_Aa, replace_phone_numbers_A)
  end
  if refine_options.is_remove_emoji
    push!(refine_Aa, text -> textacy_preprocessing.replace_emoji(text, ""))
  end
  push!(refine_Aa, textacy_preprocessing.normalize_whitespace)

  reduce(1, refine_Aa)
end
```

〈그림 43〉 전산 프로그램을 이용한 기계적 처리 과정

다. 개인 정보 비식별화 예시 문장

〈표 31〉 개인 정보 비식별화 예시 문장

구분	출처	예시
국어체	(주)플리토	못 본 사이 **이가 많이 컸네요.
	(주)플리토	**이, 벌써부터 학원에 다니고 있어?
	(주)플리토	저는 뉴욕에서 온 ****입니다.
	(주)플리토	저는 경기도에 사는 ***이라고 합니다.
	(주)플리토	제 핸드폰 번호는 ***** 이예요.

* 신문 기사, 시나리오에는 개인 정보 없음

라. 개인 정보 비식별화 이후 재정제

- 개인 정보 비식별화 이후 해당 문장을 번역에 용이하게끔 보조 연구원(일반인 수준) 및 전임 연구원(전문가 수준)이 수정하였다.
- 수정 시 원문 데이터의 의미는 유지하며 개인 정보가 비식별화되어 있는 부분을 일반적인 내용으로 수정하였다.

〈표 32〉 개인 정보 비식별화 이후 재정제 예시

정제 전	정제 후
못 본 사이 **이가 많이 컸네요.	못 본 사이 아드님이 많이 컸네요.
**이, 벌써부터 학원에 다니고 있어?	다섯 살짜리가 벌써부터 학원에 다니고 있어?
저는 뉴욕에서 온 ****입니다.	저는 뉴욕에서 살다가 왔어요.
저는 경기도에 사는 ***이라고 합니다.	저는 경기도에 살고 있어요.
제 핸드폰 번호는 ***** 이에요.	제 핸드폰 번호 여기 적어둘게요.

2. 번역 지침 및 교육

2.1. 번역 지침 제작 과정

○ 번역 지침 제작 과정 개요

- 본 사업의 번역과 검수 과정에서 일관된 기준을 제시하기 위해 아래와 같은 절차에 따라 공통 지침 및 언어별 세부 지침을 제작하였다.
- 먼저 문헌 연구를 통해 관련 자료를 정리한 뒤, 플리토에서 제공한 샘플 데이터를 번역 검수원들이 번역하여 지침에 반영할 사항을 추출하였다.
- 해당 작업들을 바탕으로 지침 초안을 작성한 뒤, 언어별 외국어 자문위원과 최종 감수자의 자문 검토 결과를 반영하여 지침을 완성하였다.
- 완성된 지침을 플리토에 공유하였고, 실제 번역과 검수 작업에 적용하면서 발견되는 문제들에 대해 언어별 검수원들이 논의하고 수정 보완해야 할 사항은 언어별 지침에 반영하여 개정하였다.
- 언어별 개정 사항을 정리하여 최종본을 확정하고 플리토에 공유하였다.

순서	작업명	주요 내용	비고
1	문헌 연구	◦ 법령 내 번역 관련 지침 및 어문 규범 참고 ◦ 플리토의 기존 지침 참고 ◦ 언어별 특징 및 번역상 오류 예상 요소 연구	
2	샘플 데이터 번역	◦ 샘플 데이터 번역 및 번역 검수원 간 논의 ◦ 번역 검수원 회의에서 지침에 대해 상세히 논의	
3	초안 작성	◦ 검수 관리팀장 및 전임 연구원 초안 작성 ◦ 품질 관리팀장 및 전임 연구원 간 검토 ◦ 플리토를 통해 1차 번역사에게 공유	지침 Ver 1.
4	자문 검토	◦ 최종 감수팀 검토 ◦ 언어별 외국어 자문위원 검토	
5	지침 완성	◦ 검수팀장 및 전임 연구원 수정 및 검토 ◦ 플리토를 통해 1차 번역사에게 공유	지침 Ver 2.
6	적용 및 수정	◦ 실제 번역 및 검수에 적용 ◦ 수정 사항 논의 및 개정	언어권마다 개별적으로 개정
7	최종본 완성	◦ 최종본 완성	

〈그림 44〉 번역 지침 제작 과정

가. 문헌 연구

○ 법령 내 번역 관련 지침 및 어문 규범 참고

- 문화체육관광부의 ‘공공 용어의 외국어 번역 및 표기 지침’ 및 ‘국어의 로마자 표기법’ 등을 참고하여 지침의 틀을 작성하였다.
- 한국어의 공공 용어에 대한 영어, 중국어, 일본어 번역 및 표기의 기본 원칙을 제시한 ‘공공 용어의 외국어 번역 및 표기 지침’은 한국어를 다른 외국어로 번역 및 검수하는 본 사업에 의미 있는 기준을 제시하기에 이를 준용하였다. 다만 각 언어의 특징과 맥락이 다르므로 기본 틀로 잡되 다른 자료들을 함께 참고하여 지침을 제작하였다.
- 한국어를 음역하는 기준은 ‘국어의 로마자 표기법’을 준용하거나 참고하여 새 기준을 작성하도록 하였다. 태국어처럼 이미 공식적인 음역 기준이 있는 경우 이를 따르도록 하였다.

공공 용어의 외국어 번역 및 표기 지침	
제정 2020. 7.15 문화체육관광부훈령 제427호 개정 2021. 7.22 문화체육관광부훈령 제448호	
제1장 총 칙	
제1조(목적) 이 지침은 국어의 공공 용어에 대한 영어, 중국어, 일본어 번역 및 표기의 기본 원칙을 제시함으로써 국가 기관과 지방 자치 단체의 공공 용어 번역의 효율성을 꾀하고 표준화된 번역 용어가 사용되도록 하는 것이 목적이다.	
제2조(용어의 정의) 이 지침에서 쓰인 용어의 뜻은 다음과 같다. 1. ‘음역(音譯)’은 한국어의 발음을 그대로 살려 번역하는 것을 뜻한다. 2. ‘의미역(意味譯)’은 용어의 의미를 살려 번역하는 것을 뜻한다. 3. ‘한자역(漢字譯)’은 한국어의 한자어를 그대로 살려 번역하는 것을 뜻한다. 4. ‘전부(前部) 요소’는 해당 용어에서 유래와 특징 등 개별성을 나타내는 앞부분을 뜻한다. 5. ‘후부(後部) 요소’는 해당 용어에서 유형과 실체 등 보편성을 나타내는 뒷부분을 뜻한다. 후부 요소의 영어, 중국어, 일본어 번역어는 [별표 1]을 참고한다.	
제3조(적용 범위) 이 지침은 용어 통일이 필요한 공공 용어의 영어, 중국어, 일본어 번역 및 표기에 적용한다.	
제2장 기본 지침	
제4조(언어권별 번역 및 표기 원칙) 언어권별 세부 번역 및 표기 규칙은 다음과 같다.	

〈그림 45〉 공공 용어의 외국어 번역 및 표기 지침(예시)

○ 플리토의 지침 반영

- 컨소시엄사인 (주)플리토는 수년간 다국어 번역 데이터를 처리해 왔고, 그 과정에서 적용하던 기존 번역 지침을 참고하여 지침에 반영토록 하였다.
- 예를 들면, 원문과 번역문이 의미상 1대 1 대응을 이루고 원문에 없는 불필요한 설명을 넣지 않으며, 원문에 충실하되 실제 사용하는 표현 위주로 자연스럽게 번역하도록 하는 등의 기준을 반영하였다.

플리토 번역 가이드 라인

1. 원문 문장에서 의미 누락/추가 없도록 주의

- ① 원문과 번역문이 의미상으로 1:1 대응을 이루어야 합니다.
- ② 원문에 충실한 번역을 하되 자연스러움을 위한, 의미의 누락/추가 없는 선에서의 의역은 허용합니다.
- ③ 번역문에는 0 [] / 등의 문장부호를 이용한 불필요한 설명을 넣으면 안 됩니다.
- ④ 원문 끝에 구두점이 있을 경우, 번역문에도 반드시 구두점을 표시합니다.
- ⑤ 원문이 한 문장인 경우, 번역문도 한 문장이어야 합니다.
단, 자연스러운 번역을 위해 심표의 추가 및 삭제가 가능합니다.
- ⑥ 한국어 원문에 주어가 생략되었거나 성별을 알 수 없는 경우, 하나의 주어를 정하여 번역합니다.
(예시) (한) 아까 이 카드로 결제했어요. → (영) I paid with this card earlier. (o)
→ (영) I/He/She paid with this card earlier. (x)

2. 번역 스타일

- ① 어색한 직역투 지양, 자연스러운 문장 구사 부탁드립니다.
- ② 숙어, 관용구 사용 시 실제 사용하는 표현만 사용해주세요.
- ③ 숫자 및 단위는 원문에 맞춰 번역합니다.
시간: (한) 3시 → (영) 3 p.m. or 3 a.m. / (한) 세 시 → (영) three o'clock
단위: (한) 킬로그램 → (영) Kilogram / (한) kg → (영) kg
기타: (한) 서른 명 → (영) thirty people / (한) 30명 → (영) 30 people
- ④ 괄호 사용 시 한 칸 띄고 써주세요. e.g, Sikjangan Mountain (623.5m)
- ⑤ 원문의 작은따옴표(강조, 인용 등의 용도)는 번역시 큰따옴표로 바꿔 번역합니다. (작은따옴표 x)
- ⑥ 책, 영화, 기타 제목 등의 앞 글자는 대문자로 표기합니다. (작은따옴표, 큰따옴표 x)

3. 기타 스타일 가이드

〈그림 46〉 플리토의 기존 번역 지침(예시)

○ 언어별 특징 및 예상 오류 정리

- 사업 준비 과정에서 관련 논문을 참고하고 한국어를 전공한 언어별 모국어 화자로부터 언어별 특징 및 번역상 오류 예상 요소와 주의사항을 정리한 자료를 지침 초안에 반영하였다.

〈표 33〉 언어별 특징 및 번역상 오류 예상 요소 및 주의사항

언어	항목	내용
베트남어	특징	• 베트남어에는 한국어의 시제 선어말어미가 없고 시제 어휘를 동사 앞에 사용함. • 한국어의 높임 선어말어미와 종결어미로 실현되는 높임법이 어휘로 나타남. • 친족 호칭어, 친족 명사가 인칭대명사를 대신하는 경우가 많음. • 한국어와 유사하게 한자어가 많은데 발음은 유사하나 의미와 쓰임이 다른 어휘가 많음.
	주의 사항	• 한국어 선어말어미와 시제 어휘의 대응 적절성 확인해야 함. • 높임 표현의 대응 적절성 확인해야 함. • 공식적 호칭어, 지칭어가 친족어로 교체되는 과정에서 공적 관계가 잘 드러나도록 주의해야 함. • 한자어의 의미를 정확하게 파악하고 그에 대응하는 베트남어로 변경해야 함.
	사례	5 LG전자는 신 가전을 중심으로 내년에도 수익성을 확보해 나간다는 전략이다. LG Electronics đang có kế hoạch đảm bảo tính lợi nhuận vào cả năm tới dựa vào các thiết bị điện gia dụng mới. • ⑤번 문장에서 ‘신 가전’은 ‘새로운 형식의 가전제품(điện tử gia dụng thể hệ mới)’을 의미하는데 이를 한자어로 직역하여 ‘신모델(thiết bị điện gia dụng)’로 번역하여 오류가 나타남.
인도네시아어	특징	• 파생어를 만드는 방법이 다양함. • 한국어에 비해 주어 생략이 매우 드물. • 높임 표현이 없어 한국어에 비해 수평적 관계가 강조됨. • 호칭 표현이 발달한 한국어에 비해 호칭, 지칭 표현을 일일이 번역하는 것이 오히려 번역 투로 여겨짐. • 소유격 표현에서 소유 주체에 대한 표현이 명확함. • 조사로 여러 의미를 드러내는 한국어에 비해 인도네시아어는 조사 체계가 없음.
	주의 사항	• 파생어를 번역하는 과정에서 복합 접사 사용에 주의가 필요함. • 맥락을 고려하여 생략된 주어를 번역 시에 꼭 추가해야 함. • 높임 표현이 없기는 하나 대상에 따라 어투가 달라지므로 주의 필요함. • 호칭 지칭에서 현지 감각에 맞게 생략해야 함. • ‘우리’와 같은 소유격 표현에서 소유 주체에 대한 개념을 명확히

언어	항목	내용		
		하여 번역해야 함. 조사에 함축된 의미에 대응하는 어휘를 번역 시에 추가해야 함.		
	사례	<table border="1"> <tr> <td>1</td><td>홈페이지에 가격이 안 나와있어서 전화드린 거예요.</td><td>Saya menelepon karena harga tidak ditampilkan di halaman web.</td></tr> </table> ①번 문장에서 홈페이지에 어떤 ‘가격(Harga)’이 제시되지 않았는지 ‘Harga+물건’과 같이 구체적으로 번역되어야 함.	1	홈페이지에 가격이 안 나와있어서 전화드린 거예요.
1	홈페이지에 가격이 안 나와있어서 전화드린 거예요.	Saya menelepon karena harga tidak ditampilkan di halaman web.		
태국어	특징	- 높임, 시제 표현에 따른 어형 변화 없음. - 수식 관계 어순이 한국어와 반대임. - 복합문 구성 시 의도-행위로 구성된 한국어와 달리 태국어는 시간 순서대로 동사를 사용해야 함. - 명사 화소의 위치와 사용 방법이 상이함.		
	주의 사항	- 높임과 시제가 어형에 포함되어 있는 경우 이에 대응하는 어휘, 문법 표현을 정확히 대치시켜야 함. - 한국어의 수식어와 피수식어의 관계를 명확히 파악하고 순서를 재배치해야 함. - 복합문의 구성에서 의미 관계를 명확히 분석한 후 시간 순서에 따라 재배치해야 함. - 명사화하기 위한 접사들의 쓰임에 주의하여 현지 감각에 따라 적절한 명사화 방안을 채택해야 함.		
	사례	<table border="1"> <tr> <td>3</td><td>그럼 다음 주 초에는 가능하세요?</td><td>เป็นไปได้ไหมสำหรับต้นสัปดาห์หน้า?</td></tr> </table> ③번 문장의 ‘주(week, สัปดาห์)’는 문어체에서 주로 사용하는 단어로 구어체에서는 ‘อาทิตย์’가 적합함. 구어체와 문어체의 혼동으로 번역 오류가 나타남.	3	그럼 다음 주 초에는 가능하세요?
3	그럼 다음 주 초에는 가능하세요?	เป็นไปได้ไหมสำหรับต้นสัปดาห์หน้า?		
인도 힌디어	특징	- 주어가 빈번하게 생략되는 한국어에 비해 힌디어는 주어의 성별과 수에 따라 동사의 형태가 결정됨. - 한국어에서는 시제 분류가 모호하고 힌디어에 비해 부사나 다른 문법 요소로 표현되는 경우가 많음.		
	주의 사항	- 문맥에 맞게 주어를 설정하고 구체적 정보가 드러나도록 번역해야 함. - 맥락 정보를 반영하고 생략되어 있는 문법 항목들을 활성화시켜야 함.		
	사례	<table border="1"> <tr> <td>1</td><td>홈페이지에 가격이 안 나와있어서 전화드린 거예요.</td><td>मैं कॉल कर रहा हूँ क्योंकि वेबपेज पर कोई मूल्य नहीं दिखाया गया है।</td></tr> </table> ①번 문장에서 ‘전화드린 거예요’의 완료형 표현이 ‘전화드리고 있다(कर रहा हूँ)’로 번역되어 시제 사용의 오류가 나타남.	1	홈페이지에 가격이 안 나와있어서 전화드린 거예요.
1	홈페이지에 가격이 안 나와있어서 전화드린 거예요.	मैं कॉल कर रहा हूँ क्योंकि वेबपेज पर कोई मूल्य नहीं दिखाया गया है।		
캄보디아 크메르어	특징	- 어순이 영어와 유사하며, 수식어-피수식어의 어순이 한국어와 상이함. - 명사의 경우, 성별 개념이 포함되지 않지만 여성이나 남성 화자가 사용하는 어휘가 상이한 경우 있음. - 구어에서는 복수의 개념을 드러내기 위해 명사 뒤에 수사를 붙이		

언어	항목	내용		
		거나 형용사를 반복하여 사용하는 경우가 있음. - 대명사가 아주 복잡하고 높임말이 다양함. 나이, 성별, 관계에 따라 서로 다른 대명사가 사용되며 화자의 성별에 따라 표현이 달라짐. - 어휘 경어법을 주로 사용하는데, 높임의 대상에 따라 상이한 어휘가 쓰임. - 시제, 상 표현이 한국어에 비해 비교적 단순함.		
	주의 사항	- 한국어의 어순을 크메르어의 어순에 맞게 재배치해야 함. - 화자-청자에 따른 어휘 자체가 변화하는 경우가 있기 때문에 말뭉치의 균형성을 위해 화자 청자의 성비를 고려하여 수집해야 함. - 한국어의 다양한 높임법과 이에 상응하는 대명사와 높임법을 적절하게 찾아야 함. 이때 문장의 주어, 목적어뿐 아니라 화자와 청자와의 관계까지 고려해야 함. - 어휘 경어법을 위해 높임 어휘를 선택하기 위해 한국어의 생략된 요소들을 정확히 표현하는 것이 중요함. - 한국어에 있는 시제, 상의 의미가 지나치게 단순화되지 않고 제대로 전달될 수 있도록 주의해야 함.		
	사례	<table border="1"> <tr> <td>1</td><td>홈페이지에 가격이 안 나와있어서 전화드린 거예요.</td><td>ផ្ទះទំនិញនៅអ៊ីនធឺណិតមិនមានតម្លៃទាក់ទងនៅលើគេហទំព័រ។</td></tr> </table> 한국어로 직역하면 캄보디아어로 전화하는 사람은 ‘~에게 전화를 주다’라고 하며, 전화를 받는 사람은 ‘~에게 전화가 온다’임. 이와 같이 화자와 청자에 따라 다른 어휘를 사용하는데 ①번 문장은 전화를 하는 상황이지만 ‘전화가 온다’로 어휘 선택에 오류가 있음.	1	홈페이지에 가격이 안 나와있어서 전화드린 거예요.
1	홈페이지에 가격이 안 나와있어서 전화드린 거예요.	ផ្ទះទំនិញនៅអ៊ីនធឺណិតមិនមានតម្លៃទាក់ទងនៅលើគេហទំព័រ។		
필리핀 타갈로그어	특징	- 타갈로그어는 한국어처럼 비교적 어순이 자유로우나 ‘동사-주어’ 형식이 더 선호되는 ‘동사-주어-목적어’ 유형의 언어임. - 한국어와 유사하게 교착어로서의 특성을 보이나 하나의 접사가 접두사, 접요소, 접미사로 두루 기능하며 접요사의 쓰임이 빈번함. - 한국어에서처럼 격표지가 발달하였으나 그 배치 양상에 있어서는 선행 명사에 붙어 후치사로 실현되는 한국어와 달리 후행 명사와 따로 떨어져 전치사로 실현됨.		
	주의 사항	- 어순을 다른 언어들과 유사하게 ‘SOV’ 형태로 제시하는 것도 가능하나 ‘SOV’ 형태를 유지하려면 연결사가 필요하므로, 보다 자연스러운 번역을 위해 ‘동사-주어-목적어’의 어순 번역도 고려할 필요가 있음. - 한국어에서 접두사, 접미사로 추가된 의미에 대응하는 적절한 접요사를 찾는 것이 중요함. - 한국어의 격표지를 반영하는 데 있어서 주어부 명사의 속성에 따라 전치사와 문법적 요소가 달라지므로 생략된 주어의 명사를 찾		

언어	항목	내용
		고 그 명사의 속성을 파악하는 것이 중요함.
	사례	4 100만 명이 넘는 칠레 사상 최대 규모의 시위 인파가 이날 산티아고에서 개혁을 요구하는 시위를 벌였다고 언론들이 전했다. Ayon sa mga ulat sa media, ang pinakamalaking protesta sa kasaysayan ng Chile ay ginanap kasama ang higit sa isang milyong kataong humihingi ng reporma sa Santiago. ④번 문장에서 ‘산티아고에서 개혁을 요구하는’의 번역인 ‘humihingi ng reporma sa Santiago’의 어순이 어색함.
러시아어	특징	- 러시아어는 동일한 대상이 반복될 경우 다른 어휘로 표현함. - 편의상 약어를 많이 사용하는 경향 있음. - 비교적 어순의 이동이 자유로움. - 능동·피동 형동사는 형용사의 성격을 지니기 때문에 긴 문장의 경우 자신이 수식하는 명사에만 국한하여 직역되는 경향이 있음.
	주의 사항	- 문장 단위 번역이나 기계번역 시, 같은 대상을 서로 다른 어휘로 표현하는 경우에 어휘가 서로 다른 대상을 지칭하는 것으로 보일 수 있으므로 통일된 어휘를 사용해야 함. - 양방향 말뭉치로의 확장을 고려할 때 양방향 모두 약어는 구체적으로 상술해야 함. - 수식 관계를 명확히 드러내기 위해 되도록 전형적인 어순을 지켜서 번역해야 함. - 능동·피동 형동사가 안고 있는 절이 길 경우에 관계대명사를 활용하는 구문으로 변경할 수 있음.
	사례	4 100만 명이 넘는 칠레 사상 최대 규모의 시위 인파가 이날 산티아고에서 개혁을 요구하는 시위를 벌였다고 언론들이 전했다. По сообщениям СМИ, более 1 миллиона чилийских демонстрантов устроили акцию протеста в Сантьяго, требуя реформы. ④번 문장에서 ‘시위(демонстрантов)’를 강조하기 위해 두 번 사용하였으나, 시위 인파는 ‘데모 인파, 시위 운동자’로 번역할 수 있음.
우즈베크어	특징	- 한국어에서는 연결어미로 접속문을 만들지만 우즈베크어에서는 어미뿐만 아니라 접속사나 쉼표(,), 줄표(-) 등을 활용해서도 두 문장 연결이 가능함. - 명사절, 관형사절을 구성하는 방식이 상이함. 절로 변환하는 과정에서 한국어가 보다 복합적 정보를 담는 경향이 있음.
	주의 사항	- 문장 연결 시 현지인들의 감각에 맞는 방법을 선택해야 함. - 한국어의 관형사형 표현에 들어있는 시제, 용언의 종류 등의 정보가 문장 전체에 잘 드러나도록 해야 함.
	사례	1 홈페이지에 가격이 안 나와있어서 전화드린 거예요. Saytning bosh sahifasida narxlarning keltirilmaganligi sababli sizga qo'ng'iroq qilgan edim. ①번 문장의 ‘가격이 안 나와 있어서’가 문어체로 번역되어 있음.

나. 샘플 데이터 번역 및 번역 검수원 회의

○ 샘플 데이터 번역 및 논의

- 플리토에서 제공한 샘플 원문을 대상으로 언어별 번역 검수원들이 직접 번역하고 지침 제작에 반영할 만한 사항을 추출하였다.

연번2	구분	유형	원문	번역문	비고	구분	유형	원문	번역문	비고
1	A	문어	국내 영화, 드라마에서 자주 등장하는 치맥(치킨+맥주), 떡볶이, 라면 먹는 장면을 통해 해외에서의 K-푸드 소비가 늘어나고 있다.	การบริโภคอาหารเกาหลีในต่างประเทศเห็นมากขึ้น จากจากการกินซี่หมัก(ไกทอด+เบียร์), ผัดกบปักษ์ และขนมปังที่ใส่ไส้รูปที่มักปรากฏในภาพยนตร์และละคร	1. 라면이란 단어를 태국어로 번역할 수 있는데 외래어로 사	B	문어	국내 영화, 드라마에서 자주 등장하는 치맥(치킨+맥주), 떡볶이, 라면 먹는 통해 해외	จากการกินซี่หมัก(ไกทอด+เบียร์) ผัดกบปักษ์และขนมปังที่ปรากฏในภาพยนตร์และละครในภาพยนตร์และละครเรื่อง	본문에 괄호가 없을 경우, 태국어 번역에는 괄호 내용을 추가해야
2	A	문어	이들은 광장 주변 발렌시아 대성당과 분수 등을 배경으로 기념사진을 찍거나 광장에 모여 앉아 이중해의 따뜻한 햇볕을 즐겼다.	พวกเขาถ่ายรูปที่ระลึกโดยมีฉากพื้นหลังเป็นมหาวิหารวาเลนเซียและน้ำพุรอบๆ จัตุรัสหรือนั่งรวมกันในจัตุรัสและเพลิดเพลินกับแสงแดดอันอบอุ่นของทะเลเมดิเตอร์เรเนียน		B	문어	이들은 광장 주변 발렌시아 대성당과 분수 등을 배경으로 기념사진을 찍거나 광장에 모여 앉	พวกเขาถ่ายรูปที่ระลึกโดยมีฉากพื้นหลังเป็นมหาวิหารวาเลนเซียและน้ำพุรอบๆ จัตุรัสหรือนั่งรวมกันในจัตุรัสและเพลิดเพลินกับแสงแดดอันอบอุ่นของทะเลเมดิเตอร์เรเนียน	
3	A	문어	문화부에 따르면 메트로폴리탄 미술관은 김정숙 여사와 방탄소년단(BTS)이 방문한 덕에 미술관 SNS 계정에 사상 최대의 인파가 몰렸다고 큰 감사를 표하며 우리 측에 호의적인 입장을	ตามที่กระทรวงวัฒนธรรม ทักษิณภัณฑศิลป์และหอศิลป์แห่งชาติกล่าวว่าการมาเยือนของคุณคิมจองซอกและบีทีเอสทำให้ส่งชนเข้ามามวลตัวกันในปีของ SNS ของพิพิธภัณฑ์เป็นจำนวนมากที่สุดในประวัติศาสตร์ และได้แสดงความขอบคุณ	UNICEF는 괜찮지만, 태국 모를 수 있는 SNS는 의미	B	문어	문화부에 따르면 메트로폴리탄 미술관은 김정숙 여사와 방탄소년단(BTS)이 방문한 덕에 미술관 SNS 계정에 사상	ตามที่กระทรวงวัฒนธรรม ทักษิณภัณฑศิลป์และหอศิลป์แห่งชาติกล่าวว่าการมาเยือนของคุณคิมจองซอกและบีทีเอสทำให้ส่งชนเข้ามามวลตัวกันในปีของ SNS ของพิพิธภัณฑ์เป็นจำนวนมากที่สุดในประวัติศาสตร์ และได้แสดงความขอบคุณ	김정숙 여자(영부인) 추가하면 안 됨.
4	A	문어	100만 명이 넘는 칠레 사상 최대 규모의 시위 인파가 이날 산티아고에서 개혁을 요구하는 시위를 벌였다고 언론들이 전했다.	สื่อรายงานว่าวันนี้ฝูงชนมากกว่าหนึ่งล้านคนซึ่งเป็นจำนวนมากที่สุดในประวัติศาสตร์ของประเทศเมื่อมาประท้วงขบวนปฏิรูปในย่านติอาโก		B	문어	100만 명이 넘는 칠레 사상 최대 규모의 시위 인파가 이날 산티아고에서 개혁을 요구하는 시	สื่อต่างๆ รายงานว่าประชาชนมากกว่าหนึ่งล้านคนออกมาประท้วงขบวนปฏิรูปเรียกร้อง	
5	A	문어	LG전자는 최근 건조기 이슈에 대해 사과하면서 신 가전을 중심으로 내년에도 수익성을 확보해 나간다는 전략이라고 밝혔다.	เครื่องใช้ไฟฟ้า LG กล่าวขอโทษเกี่ยวกับประเด็นเครื่องอบแห้งเมื่อเร็วๆ นี้ และประกาศถึงกลยุทธ์ว่าในปีหน้าก็จะรักษาผลกำไรโดยมุ่งเน้นไปที่เครื่องใช้ไฟฟ้าแบบใหม่		B	문어	LG전자는 최근 건조기 이슈에 대해 사과하면서 신 가전을 중심으로 내년에도 수익성을 확보해	บริษัท LG Electronics ขอโทษสำหรับปัญหาของเครื่องอบแห้งเมื่อเร็วๆ นี้ และประกาศถึงกลยุทธ์ว่าในปีหน้าก็จะรักษาผลกำไรโดยมุ่งเน้นไปที่เครื่องใช้ไฟฟ้าแบบใหม่	
6	A	구어	홈페이지에 가격이 안 나와있어서 전화드린 거예요.	โทรมาที่เพราะว่าไม่เร้าคาบออกในเว็บไซต์/ครับ	โทรมา 방향, 웹 같은 외래어표기는 왕립 공식 규정 참고	B	구어	홈페이지에 가격이 안 나와있어서 전화드린 거예요.	โทรไปเพราะว่าหน้าเว็บไม่แจ้งราคาไว้	โทรไป X
7	A	구어	네 배려해 주셔서 감사합니다.	ขอบคุณที่เอาใจใส่ครับ/ครับ	문화적 표현	B	구어	네 배려해 주셔서 감사합니다.	ขอขอบคุณสำหรับการช่วยเหลือของคุณ	문화적 표현
8	A	구어	그럼 다음 주 초에는 가능하세요?	ถ้าอย่างนั้น คนสัปดาห์หน้าเป็นไปได้ไหมครับ/ครับ	O	B	구어	그럼 다음 주 초에는 가능하세요?	ถ้าอย่างนั้นสัปดาห์หน้าเป็นไปได้ไหมครับ/ครับ	맞췌 성조 표기 오류
9	A	구어	더위가 요즘 심하다가 지금 장마도 엄청 심하고 굉장히 힘들잖아요.	ช่วงนี้อากาศร้อนจัด ตอนที่ฝนก็ตกหนักอีกลำบากมากครับ/ครับ		B	구어	더위가 요즘 심하다가 지금 장마도 엄청 심하	ช่วงนี้อากาศร้อนจัด แล้วตอนนี้ฝนก็ตกหนักด้วยลำบากมาก	
10	A	구어	그니깐 그거 땀에 일로 을 수가 없는 거죠.	นั่นแหละ เพราะเหตุนี้มันก็เลยมาทางนี้ไม่ได้ด้วยครับ/ครับ	X	B	구어	그니깐 그거 땀에 일로 을 수가 없는 거죠.	เพราะฉะนั้นมันจึงไม่ได้มาทำงาน	X

〈그림 47〉 (예시) 샘플 데이터 번역 및 논의

○ 내용 검수팀 정례 회의를 통한 지침 상세화

- 기존 문헌 자료 결과 및 샘플 번역 결과를 바탕으로 언어별 내용 검수팀 정례 회의에서 지침으로 규정할 주요 사항 및 쟁점에 대해 한국인과 원어민 참여자 모두 상세히 논의함.
- 예를 들어 한국어를 인도 힌디어로 번역할 때, 한국어에서 복수를 단수처럼 표현 하더라도 도착어의 맥락에 따라 복수로 번역하도록 하였다. 그리고 태국어의 경우 남성과 여성을 구별하는 대명사나 어조사를 빗금(/)으로 동시 표기하지 않고 맥락에 따라 어느 하나로만 번역하도록 하였다.

다. 초안 작성

○ 검수 관리팀장 및 전임 연구원이 지침 초안 작성

- 지침 담당 전임 연구원이 기존 자료들을 바탕으로 공통 지침 초안을 작성하였다.
- 언어별 검수 관리팀장과 언어별 담당 전임 연구원이 언어별 특징에 따라 공통 지침을 준용하여 언어별 세부 지침의 초안을 작성하였다.

○ 품질 관리팀장 및 전임 연구원 간 검토

- 언어별 세부 지침에 대해 품질 관리팀장 및 타언어 담당 전임 연구원 간 검토를 통해 세부 지침 초안(Ver. 1)을 완성하였다.

한국어-외국어 번역 검수 공통 지침

제 1 장 총칙

제 1 절 목적

- 이 지침의 목적은 한국어와 외국어의 병렬 말뭉치 구축을 위한 다국어의 번역과 검수에 관한 기본 규칙을 제시함으로써 타당성, 효율성을 도모하고 일관된 번역이 이뤄지도록 하는 것이다.

제 2 절 용어의 정의

1. 출발어와 도착어

- 한국어-외국어 병렬 말뭉치 구축 사업에서 '출발어(Source Language)'는 한국어이고, '도착어(Target Language)'는 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어이다.
- 외국어-한국어 병렬 말뭉치의 경우, 출발어는 베트남어이고 도착어는 한국어이다.

2. '음역(音譯)'은 한국어의 발음을 그대로 살려 번역하는 것을 뜻한다. 한국의 로마자표기법 또는 도착어의 한글 음역 표기 기준에 따라 일관되게 표기하는 것이 원칙이다.

예) (한) 백제 → (영) Baekje

제1항 기본 원칙

제1목 의미의 정확성

1. 원문과 번역문이 의미 및 형식상으로 1:1 대응, 즉 등가성(equivalence)을 이루어야 한다. 즉, 원문의 맥락적 의미와 기능을 고려하여 정확하게 번역하며 원문이 한 문장일 경우 번역문도 한 문장으로 번역한다. 단, 한 문장으로 번역이 어려운 경우 심표로 처리할 수 있으며, 1:1로 번역하였을 때 번역문이 어색할 경우 문장을 일부 수정할 수 있다.
2. 의미상 불필요한 첨가와 누락, 내용의 불일치에 주의한다. 다만, 인도네시아어의 문법적 특징에 따라 생략된 한국어의 주어나 목적어 등을 복원하여 추가할 수 있다.
3. 중의적 표현의 경우, 맥락을 파악하여 정확한 의미로 번역한다. 특히 수식어가 수식하는 대상, 부정 표현이 부정하는 대상, 한국어 동음어와 다의어의 의미 파악, 번역 가능한 유의어의 적절한 선택을 고려하여 번역한다.

한국어	인도네시아어	적합 여부
세종대왕은 1443년에 훈민정음을 창제하였다.	Raja Sejong menciptakan Hunminjeongeum pada tahun 1443.	O
	Raja Sejong menciptakan Hunminjeongeum pada tahun 1443 <u>dan menyebarkannya pada tahun 1446.</u>	X

제2목 맥락의 실제성

1. 원문의 목적, 용법, 요구, 맥락 등을 고려하여 적절하게 번역해야 한다.
2. 직역을 원칙으로 하되, 직역이 어색한 경우 원문의 의미에 충실하고 현지에서 실제로 사용하는 자연스러운 표현으로 번역한다.
예) (한) 꼭 한번 가 보고 싶네요.
(인) Saya sangat ingin mengunjunginya suatu hari nanti.
3. 존댓말/반말, 격식체/비격식체의 경우, 원문의 성격과 맥락을 파악하여 번역한

라. 자문 검토

○ 최종 감수팀 검토

- 최종 감수팀장에게 공통 지침에 대해 자문을 의뢰하였다. 자문 결과는 전임 연구원들 논의에 따라 공통 지침에 반영하였다.
- 언어별 최종 감수자에게 세부 지침에 대해 자문을 의뢰하였으며 자문 결과는 언어별 감수 관리팀장과의 논의에 따라 세부 지침에 반영하였다.

자문의견서	
■ 사업(과제명)	2021 한국어-외국어 병렬 말뭉치 구축사업
■ 자 문 내 용	태국어 번역 감수 세부 지침에 대한 자문
■ 자 문 일 자	2021.11.09.
■ 자 문 인	소 속: 부라파대학교
	성 명: 카노크완 사로즈나
지침안에 대한 구체적이고 세부적인 의견은 실제 자료의 문장들을 보면서, 작업을 하면서 판단할 수 있을 것 같습니다. 하지만 일반적인 수준에서는 유용한 준거가 될 수 있는 번역 지침이라고 생각합니다. '한국어-태국어 병연 지침_초안'에서 예시로 올리신 태국어 몇 개가 조심스러워 주석을 달아 봤습니다.	
현 지침 태국왕립학술원의 한태 용차 기준 (http://www.ratchakitcha.soc.go.th/DATA/PDF/2555/E/112/7.PDF) 에 따라 표기한다.	자문 의견 이 자료는 검색이 안 됩니다. 아마도 아래 자료로 업데이트된 것 같습니다. 주소를 올려 놓았으니 확인 부탁드립니다. http://www.ratchakitcha.soc.go.th/DATA/PDF/2555/E/112/7.PDF
어린이집 - ศูนย์รับเลี้ยงเด็ก	'어린이집'이라면 태국어로 'ศูนย์พัฒนาเด็กเล็ก'이 더 일반적입니다. 재확인을 부탁드립니다.
요금소 - ประตูเก็บค่าผ่านทาง	태국어로는 "ด่านเก็บค่าผ่านทาง"이라는 말을 더 널리 사용됩니다. 재확인 부탁드립니다.
보건소 - สถานีอนามัย	태국에서는 현재 "สถานีอนามัย"라는 명칭을 사용하지 않습니다. 현재는 "โรงพยาบาลส่งเสริมสุขภาพตำบล"을 사용하고 있습니다. 확인 부탁드립니다.
지하상가 - อาร์ค데ใต้ดิน	"ศูนย์การค้าใต้ดิน"이 아닌가 싶습니다.
감자탕 - ต้มกระดูกหมู	예시된 태국어를 한국어로 번역하면 '뼈다귀탕'이 됩니다. '뼈다귀탕'과 '감자탕'은 별개의 음식이 아닌지요????
산적 - เลียบไม้โป่ง	지금 '산적'과 '꼬치'를 같은 번역어로 예시하고 있습니다. 구별을 하자면 '산적'은 태국어로 "สะต๊ะ" 이나 "nám"을 사용하는 것이 더 적절할듯 싶습니다.
공공 임대 주택 - บ้านให้เช่าของสาธารณะ	너무 직역을 한 예시라는 생각이 듭니다. 건축이나 부동산 분야의 전문용어로 간주할 수 있다면 '공공 임대 주택'은 태국에서 "บ้านเช่าของรัฐ"라는 말로 사용되고 있습니다. 확인 부탁드립니다.
보건소 - สถานีอนามัย	공공의료시설을 이야기하는 것이라면 "โรงพยาบาลส่งเสริมสุขภาพตำบล"이 더 적절할듯 싶습니다.
생명형 - โภชนาการชีวิต	โภชนาการชีวิต
자유형 - โภชนาการอิสระภาพ	โภชนาการ
재산형 - โภชนาการทรัพย์	โภชนาการ
자격형 - โภชนาการสถานภาพ	โภชนาการหรือจำกัดสิทธิต่างๆ

〈그림 50〉 (예시) 최종 감수자의 자문 의견서

○ 언어별 외국어 자문위원 검토

- 언어별 외국어 자문위원에게 세부 지침에 대해 자문을 의뢰하였고 그 결과를 지침에 반영하였다. 이후 언어별 자문 회의에서 논의된 결과 중 지침에 반영해야 할 사항들은 수정 반영하였다.

자문의견서	
■ 사업(과제명)	2021 한국어-외국어 병렬 말뭉치 구축사업
■ 자 문 내 용	필리핀 타갈로그어 번역 검수 세부 지침에 대한 자문
■ 자 문 일 자	2021.12.26.
■ 자 문 인	소 속: 부산외국어대학교
	성 명: 김동엽
1) <필리핀 타갈로그어> 명칭에 대한 의견 => 필리핀 헌법 제14조 6항에 따르면, "The national language of the Philippines is Filipino"라고 되어 있음. <Tagalog, 타갈로그어>는 필리핀에 존재하는 수많은 지방언어 중의 루손섬 중부 지역에서 사용하는 언어로서 필리핀 정부가 국어로 채택하여 전국적으로 사용하고 있는 언어임. 따라서 <필리핀 타갈로그어>라고 쓰면, 필리핀의 많은 지방 언어들 중 하나를 사용하는 것처럼 보일 수 있음. 필리핀 국어의 명칭이 Filipino로 되어 있으므로 한국어 표기로 <필리핀어>라고 하는 것이 적절하다고 사료됨. 이미 본문에서도 여러 군데 <필리핀어>라는 용어가 사용되고 있음: p. 19. 한국어와 필리핀어에서 대응되는... p. 24. 필리핀어위원회(Commission on the Filipino Language)라고 표현되어 있음. p. 27. 영어-필리핀어 유의어 사전 2) <첨부 6>. 한국어-필리핀어 번역 검수 세부사항 - 관련 의견 p. 6 - (제3목 인공지명) 3. 표 안에 있는 (저수지, 대교) 필리핀어로 번역 요망 P. 7 - (제5목 도로명). 도로명(대로, 로, 길)을 음역하기보다는 도착어 환경에 맞추어 영어(Boulevard (Blvd), Avenue(Ave), Street(St.))로 표현하는 것이 의미 전달에 좋을 듯. 필리핀에서 도로명의 약어를 흔히 사용하고 있음. p. 12 (제1목 보통명사) 2. 보통명사 사주(saju)의 경우, 단지 음역만으로는 의미 전달이 어려움. 의미역을 사용하는 것이 좋겠음. p. 18 (제10목 축약어) 3. 표 안에 필리핀어에서 EU도 적합하다고 봄. p. 21 (제6목 높임법과 공손성) 2 & 3. 출발어와 도착어 순으로 예문을 표기하는 것이 좋겠음. => 한국어를 필리핀어로 번역할 때, (필리핀인)번역자의 출신 지역에 따라서 일부 단어를 지방어로 번역하는 경우도 있음. => 애써 현실에서 사용하지 않는 필리핀어 용어를 찾아서 번역하기보다는 일상에서 흔히 사용하는 대체 영어 표현을 적절히 사용하여 번역하는 것이 의미 전달에 유용함.	

<그림 51> (예시) 언어별 외국어 자문위원의 자문 의견서

마. 지침 완성

○ 검수 관리팀장 및 전임 연구원 수정 및 검토

- 검수 관리팀장과 전임 연구원이 여러 자문 의견과 번역 검수원의 피드백을 바탕으로 공통 및 세부 지침을 수정·보완하여 지침(Ver. 2)을 완성하였다.

○ 플리토를 통해 1차 번역사에게 공유

- 완성된 지침은 플리토에 공유하여 본격적인 번역에 적용토록 요청하였다.

CONFIDENTIAL	
21 한국어-외국어 번역 검수 지침 Ver. 2	
목차	
제 1 장 총칙	4
제 1 절 목적	4
제 2 절 용어의 정의	4
제 3 절 적용 범위 및 저작권 보호	5
제 4 절 지침 적용 절차	5
제 2 장 공통 지침	6
제 1 절 기본 원칙	6
제 1 항 의미의 정확성	6
제 2 항 맥락의 실제성	6
제 3 항 문제의 적절성	6
제 2 절 고유명사	7
제 1 항 인명	7
제 2 항 기관명과 상품명	8
제 3 항 인공 지명	8
제 4 항 행정 구역명	9
제 5 항 도로명	10
제 6 항 자연 지명	10
제 7 항 문화재명 및 예술 작품명	11
제 8 항 음식명	11
제 3 절 어휘 및 표현	13
제 1 항 보통명사	13
제 2 항 대명사	13
제 3 항 숫자와 단위	13

1

이 지침은 저작권법의 보호를 받으므로 본 사업단과의 별도 협의 없이 어떠한 형태로도 복제 및 배포하는 것을 금지한다.

〈그림 52〉 (예시) 한국어-외국어 번역 검수 공통 지침 Ver. 2

러시아어 번역/검수 세부 지침

제 1 항 기본 원칙

제1목 의미의 정확성

1. 원문과 번역문이 의미상으로 1:1 대응, 즉 동가성(equivalence)을 이루어야 한다. 즉, 원문의 맥락적 의미와 기능을 고려하여 정확하게 번역한다.
2. 의미상 불필요한 첨가와 누락, 내용의 불일치에 주의한다. 다만, 러시아어의 문법적 특징에 따라 생략된 한국어의 주어나 목적어 등을 복원하여 추가할 수 있다.
3. 중의적 표현의 경우, 맥락을 파악하여 정확한 의미로 번역한다. 특히 수식어가 수식하는 대상, 부정 표현이 부정하는 대상, 한국어 동음어와 다의어의 의미 파악, 번역 가능한 유의어의 적절한 선택을 고려하여 번역한다.

예)

한국어	러시아어	적합 여부
세종대왕은 1443년에 훈민정음을 창제하였다.	Король Седжон создал Хунминчоным в 1443 году.	O
	Король Седжон создал корейскую письменность, которая была завершена в 1443 году, а опубликована спустя несколько лет в 1446 году.	X

제2목 맥락의 실제성

1. 원문의 목적, 용법, 요구, 맥락 등을 고려하여 적절하게 번역해야 한다.
2. 직역을 원칙으로 하되, 직역이 어색한 경우 원문의 의미에 충실하고 현지에서 실제로 사용하는 자연스러운 표현으로 번역한다.

바. 지침 적용 및 개정

○ 실제 번역과 검수 및 감수에 지침 적용

- 본격적인 번역과 검수 및 감수에 지침을 적용하였다.

○ 보완 사항 논의 및 개정

- 번역과 검수 과정에서 발견된 쟁점 및 주의사항 등에 대해서 언어별 정례 회의에서 수시로 논의하였다. 언어별 검수 상황에 따라 수정 보완 사항이 있을 경우, 세부 지침을 개정하고 개정 사항을 취합하여 최종본을 완성하였다. 지침은 개정될 때마다 플리토와 공유하여 번역에 반영되도록 하였다.

	1. 한국어-베트남어 번역 검수 세부 지침_ver.2.1_211217.pdf
	2. 한국어-인도네시아어 번역 검수 세부 지침_ver.2.1_211217.pdf
	3. 한국어-태국어 번역 검수 세부 지침_ver.2.1_2201221_송부용.pdf

〈그림 54〉 (예시) 언어별 세부 지침의 수시 개정

	0. 한국어-외국어 번역 검수 공통 지침_ver.2.1_220501_최종본_송부용.pdf
	1. 한국어-베트남어 번역 검수 세부 지침_ver.2.3_220506.pdf
	2. 한국어-인도네시아어 번역 검수 세부 지침_ver.2_220506_최종본_송부용.pdf
	3. 한국어-태국어 번역 검수 세부 지침_ver.2.3_220505_최종본_송부용.pdf
	4. 한국어-인도 힌디어 번역 검수 세부 지침_ver.2.3_220507_최종본_송부용 (3).pdf
	5. 한국어-캄보디아 크메르어 번역 검수 세부 지침_Version2.3_2220506_최종본_송부용.pdf
	6. 한국어-필리핀 타갈로그어 번역 검수 세부 지침_ver.2.1_220508_최종본_송부용.pdf
	7. 한국어-러시아어 번역 검수 세부 지침_ver.2.2_220506_최종본_송부용.pdf
	8. 한국어-우즈베크어 번역 검수 세부 지침_ver.2.3_220506.pdf
	9. 베트남어-한국어 번역 검수 세부 지침_ver.2.4_220507_최종본_송부용.pdf

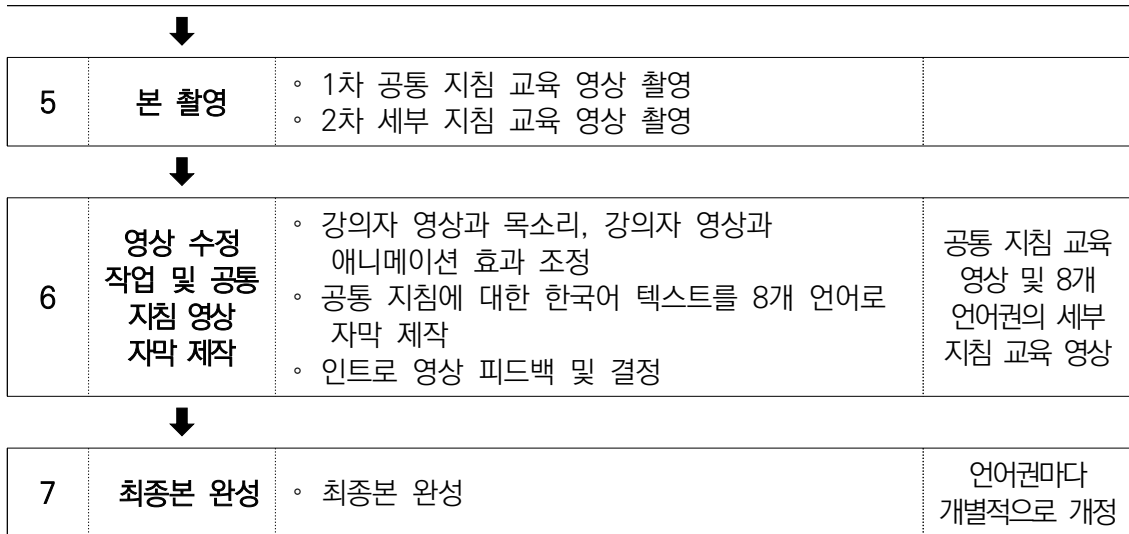
〈그림 55〉 한국어-외국어 번역 검수 지침 최종본 목록

2.2. 교육 영상 제작

○ 교육 영상 제작 과정 개요

- 공통 참조 영상에는 책임 연구원이 본 사업단 소개 및 과업 수행에 대한 개요를 설명하였고, 지침 담당 연구원이 공통 지침에 대한 내용을 설명하였다. 공통 참조 영상에는 8개 언어 자막을 제작하여 삽입함으로써 각 언어권 참여자들이 공통 참조 영상을 용이하게 이해할 수 있도록 하였다.
- 세부 지침 교육 영상에는 언어별 검수 관리팀장과 함께 세부 지침 내용을 바탕으로 교육 영상을 제작하였다. 영상에는 번역 및 검수 시 반영해야 하는 주요 지침 사항을 중심으로 관련 예시와 함께 설명하였다.
- 영상 제작 단계마다 전임 연구원과 검수 관리팀장이 협업하여 언어별 특징을 효율성 있게 잘 반영했는지, 해당 언어로 표현 시 오탈자는 없는지 등을 사전 점검하여 영상에 수시로 수정하고 반영하였다.
- 영상은 비공개 유튜브에 탑재하여 사업단에 참여하는 번역사 및 검수원들만 시청할 수 있게 하여 보안을 유지하면서 교육이 이루어지도록 하였다.

순서	작업명	주요 내용	비고
1	교육 영상 제작 사전 논의	- 공통 지침 영상과 세부 지침 교육 영상 제작 결정 - 연구 책임자, 지침 담당 연구원, 언어별 검수 관리팀장 참여 결정	
↓			
2	검수 관리팀장 섭외 및 사전 준비 안내	- 공통 지침 영상 촬영용 PPT 제작 및 대본 작성 - 세부 지침 영상의 필요성 및 담당 업무 설명 - 언어별 전임 연구원은 지침안 토대로 PPT 작성	
↓			
3	촬영용 PPT 제작	- 작성된 PPT를 토대로 언어권 촬영용 PPT 제작 의뢰 - 언어별 형식 통일, 재생 시간, 언어적 특징을 감안한 피드백 적용 - 촬영용 PPT 완성	
↓			
4	대본 작성	- 제작된 촬영용 PPT를 토대로 검수 관리팀장 대본 작성 - 촬영 시, 대본 노출 방식 논의 - 본 촬영 전 ZOOM 사전 연습 진행	



〈그림 56〉 교육 영상 제작 개요

가. 교육 영상 제작 사전 논의

○ 교육 영상 제작의 필요성

- 언어별 번역 및 검수 지침안이 상세한 만큼 효율적인 교육을 위해서는 핵심 사항에 대해 영상으로 제작하여 반복적으로 시청하도록 하는 것이 필요하다는 의견이 대두되었다.
- 사업의 취지와 성격에 대한 전반적인 소개와 8개 언어권을 아우르는 공통 지침 영상, 그리고 언어별 세부 지침 영상을 제작하기로 결정하였다.

○ 교육 영상 제작 업체 및 담당자 결정

- 교육 영상 제작 업체 선정 및 회의를 진행하였다.



〈그림 57〉 영상 제작 업체 사전 회의

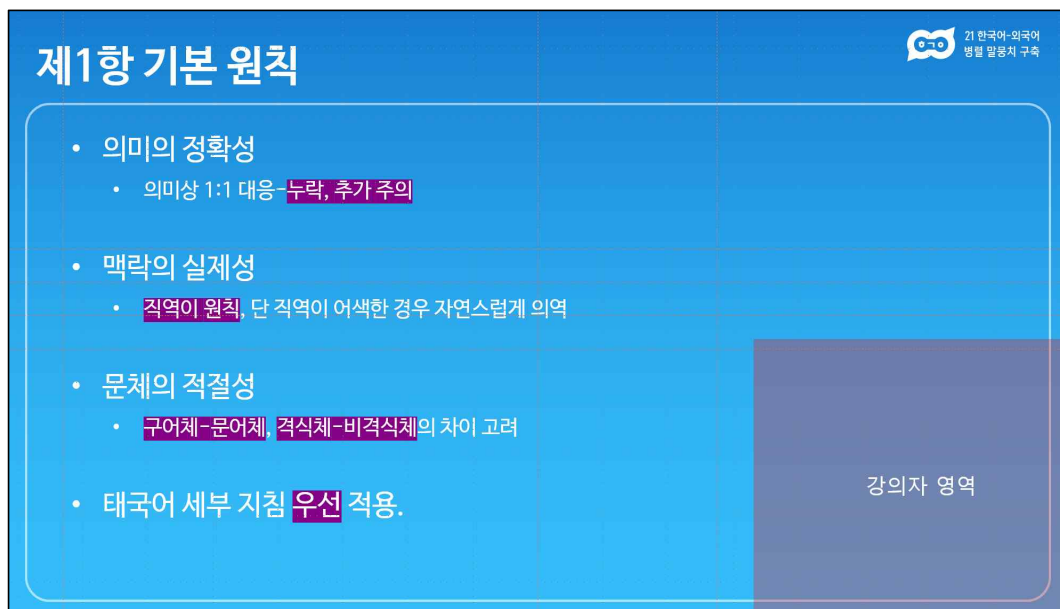
- 2021 한국어-외국어 병렬 영상 제작 업체 사전 회의 말뭉치 구축 사업 소개 및 공통 지침 교육에 대해서는 연구 책임자와 지침 담당 연구원이 영상에서 설명하기로 결정하였다.
- 8개 언어 세부 지침 교육은 각 언어권 검수 관리팀장이 영상에서 설명하고, 언어별 담당 전임 연구원이 보조하여 진행하기로 결정하였다.

〈표 34〉 지침 교육 영상 내용 및 담당자

구분	영상 내용	영상 담당
공통	한국어-외국어 말뭉치 구축 사업 소개	이정희 연구 책임자
	한국어-외국어 번역 검수 공통 지침	이두용 전임 연구원
세부	한국어-베트남어 번역 검수 세부 지침	웬빈뜨엉 검수 관리팀장
	한국어-인도네시아어 번역 검수 세부 지침	플로리안 카롤루스 검수 관리팀장
	한국어-태국어 번역 검수 세부 지침	케와린 시무앙 검수 관리팀장
	한국어-인도 힌디어 번역 검수 세부 지침	스리잔 꾸마르 검수 관리팀장
	한국어-캄보디아 크메르어 번역 검수 세부 지침	망점란 검수 관리팀장
	한국어-필리핀 타갈로그어 번역 검수 세부 지침	츄아 마리아 콘셉손 검수원
	한국어-러시아어 번역 검수 세부 지침	김빅토리아 검수 관리팀장
	한국어-우즈베키스탄어 번역 검수 세부 지침	파르비나 사토로바 검수 관리팀장

나. 검수 관리팀장 섭외 및 사전 준비 안내

- 공통 지침을 담당하는 영상 1팀과 세부 지침을 담당하는 영상 2팀으로 구성하여 작업을 진행하였다(이하-영상 1팀, 영상 2팀).
- 공통 지침 영상 1팀
 - 기제작된 번역 및 검수 지침을 토대로 공통 지침 담당 전임 연구원이 촬영용 PPT를 작성하였다.
 - 업체에서 완성된 PPT를 받아 내용을 수정·보완하였다.
 - 영상 재생 시간에 맞춰 촬영 전 사전 연습을 진행하였다.
- 세부 지침 영상 2팀
 - 영상 제작 담당 연구원이 검수 관리팀장들에게 세부 지침 영상의 필요성 및 담당 업무를 설명하였다.
 - 검수 관리팀장들에게 작업 절차(세부 지침안을 토대로 촬영용 PPT 제작→대본 작성→사전 리허설 진행→본 촬영→영상 후작업→공통 지침에 들어갈 자막 번역→최종 영상 피드백→영상 완성 후, 언어별 번역 검수원 교육 및 공유)를 안내하였다.



〈그림 58〉 영상 촬영을 위한 PPT 초안

다. 촬영용 PPT 제작

○ 공통 지침 영상 1팀

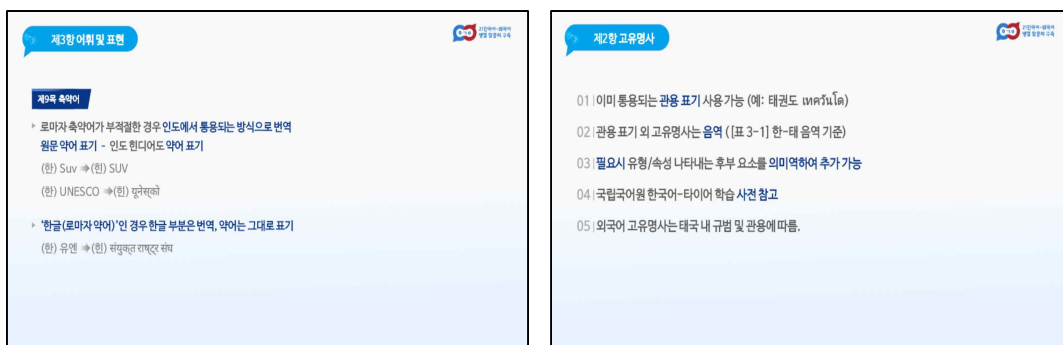
- 촬영을 위한 사전 리허설을 하고 촬영을 진행하였다.



<그림 59> 영상 1팀 촬영 장면

○ 세부 지침 영상 2팀

- 기제작된 번역 및 검수 지침을 토대로 해당 언어권의 전임 연구원들이 촬영용 PPT를 작성하였다.
- 업체에서 완성된 PPT를 받아 검수 관리팀장들과 내용을 수정·보완하였다.

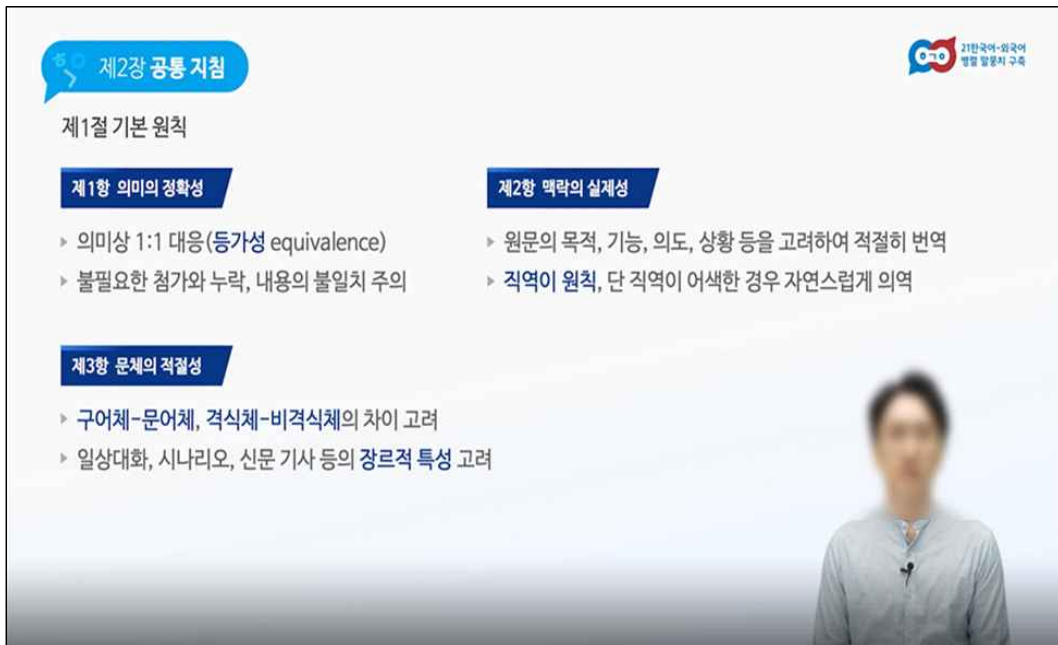


<그림 60> 촬영용 PPT 예시(태국어, 인도 힌디어)

라. 검수 관리팀장 대본 작성

○ 영상 1팀

- 촬영 영상을 보고 수정해야 할 의견들을 취합하였다. 영상의 PPT 텍스트양의 적절성, 강의자의 화면 내 크기 및 강의 속도의 적절성, 애니메이션 효과 등을 논의하였다. 이를 토대로 영상 후작업을 진행하였다.



<그림 61> 1팀 촬영 영상 피드백 반영 영상

○ 영상 2팀

- 제작된 촬영용 PPT(디자인 완성본)를 토대로 검수 관리팀장과 전임 연구원이 대본을 작성하였다.
- 촬영 시, 프롬프터 지원이 되는지 재차 확인(태국어, 인도 힌디어, 캄보디아 크메르어) 후 지원이 안 될 경우의 대본 자막 노출 방식을 논의하였다.
- 태국어, 인도 힌디어, 캄보디아 크메르어의 문자 특수성으로 인해 글자가 깨지지 않는 선에서 프롬프터를 활용할지 워드로 진행할지 논의 및 결정하고 두 가지 방법으로 모두 사전 준비를 하였다.
- 본 촬영 전 검수 관리팀장들은 ZOOM으로 사전 리허설을 진행하였으며 피드백 받은 내용을 숙지하여 촬영 전까지 개인 연습을 진행하였다.
- 모든 검수 관리팀장들의 촬영 일정을 확정(단, 필리핀 타갈로그어는 촬영 담당자가 현지에서 촬영한 영상을 받아 후작업으로 진행)하고 촬영 전 준비, 촬영 시간, 촬영 의상, 촬영 장소를 안내하였다.

마. 본 촬영

○ 1차 영상 후작업

- 1차 영상은 모든 촬영과 피드백을 완료하였으며, 피드백을 토대로 후작업을 진행하였다.

○ 2차 영상 본 촬영

- 8개 언어권의 검수 관리팀장은 각자 자신의 스케줄에 맞춰 스튜디오에 도착해 촬영 준비를 마치고 리허설을 진행한 후 본 촬영에 들어갔다.
- 필리핀 타갈로그어는 현지에서 촬영 후 영상을 받아 형식을 통일하는 방식으로 제작하였다.



〈그림 62〉 검수 관리팀장 촬영 장면(태국어, 인도 힌디어)

바. 영상 수정 작업 및 자막 제작 의뢰

○ 영상 후작업

- 검수 관리팀장 및 전임 연구원은 촬영한 영상을 가지고 언어별로 설명과 화면의 시점이 일치하는지 검토하고 수정해야 할 상황들을 정리하였다.
- 강의자의 설명과 영상의 시점이 불일치하는 부분을 조정하였고, 애니메이션 효과를 통일감 있게 조정하였다.

○ 공통 지침 자막 번역

- 연구 책임자와 검수 관리팀장들은 줌(ZOOM)에서 자막의 방식에 대해 논의하였고 언어별 자막을 영상에 링크하여 삽입하는 방식으로 결정하였다. 또한 한국어로

제공한 자막을 해당 언어로 검수 관리팀장들이 번역하기로 결정하였다.

- 연구 책임자의 사업단 소개 영상과 지침 담당 연구원의 공통 지침 영상에 한국어 자막과 8개 언어권의 자막을 선택해서 시청할 수 있게 제작하였다.

2021년 한국어-외국어 병렬 말뭉치 구축 사업

연구책임 이정희

วันนี้จะขออนุญาตแนะนำโครงการนี้โดยสังเขปค่ะ

제2장 공통 지침

제2절 고유명사

제3항 인공지명

▶ 전부 요소는 음역, 후부 요소는 의미역

(한) 보라매초등학교 ➡ (영) Boramae Elementary School

제4항 행정 구역명 및 제5항 도로명

▶ 전부 요소는 음역, 후부 요소는 음역 또는 의미역 ➡ 세부 지침 참고

(한) 서울시 ➡ (영) Seoul / Seoul City

(한) 강변북로 ➡ (영) Gangbyeonbuk-ro

यह नियम प्रत्येक भाषा के लिए थोड़ा अलग है इसलिए कृपया विस्तृत दिशा-निर्देश को देखें।

〈그림 63〉 자막 삽입 영상(태국어, 인도 힌디어)

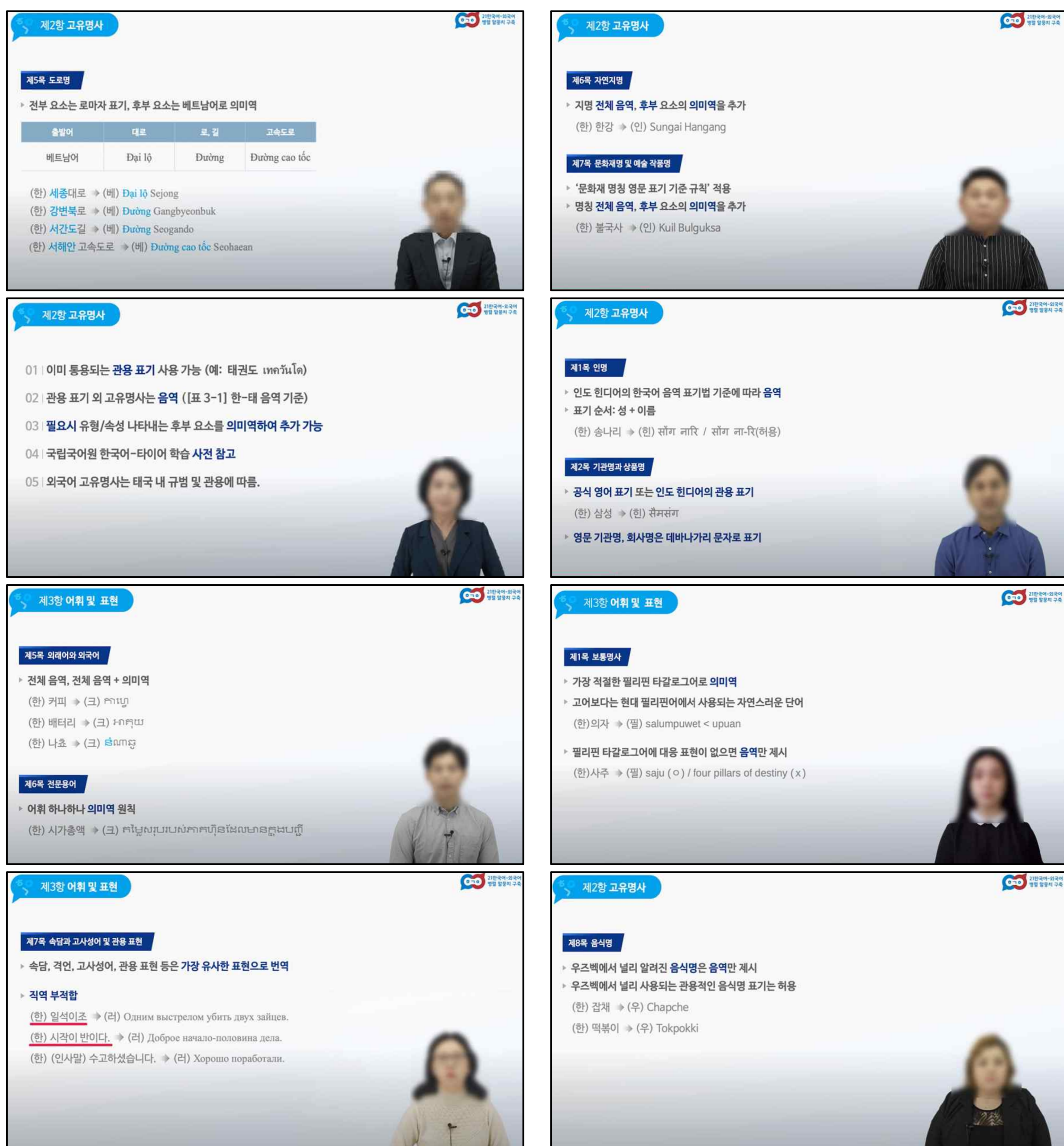
사. 최종본 완성

○ 인트로 영상 결정

- 본 사업단 교육 영상의 특징을 반영한 인트로 영상을 의뢰하였다.
- 제작된 인트로 영상에 대한 전임 연구원들의 피드백을 종합해 논의한 결과를 영상에 반영하였다.

○ 최종본 산출물

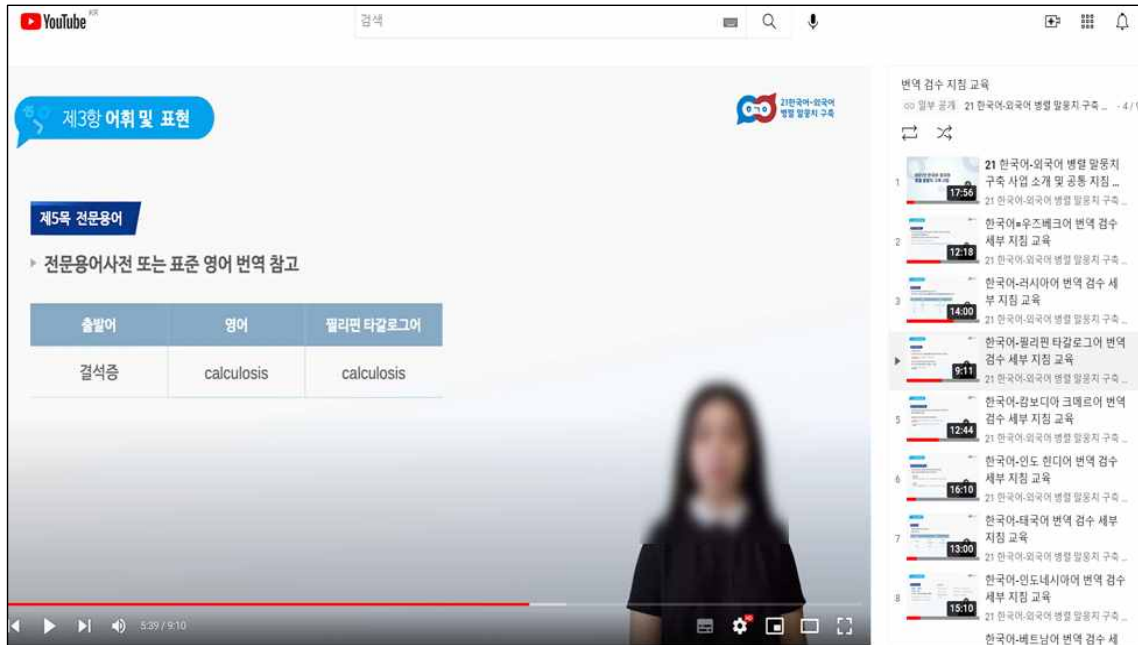
- ' 21 한-외 병렬 말뭉치 구축 사업 소개 및 공통 지침 영상을 완성하였다.
- 한국어-외국어 세부 지침 교육 영상 8편을 완성하였다.



<그림 64> 언어별 최종 지침 교육 영상

○ 영상 이용 방법 및 관리

- 완성된 영상을 비공개 YouTube에 탑재하였으며 시청 권한을 허용한 검수원에 한해서 시청 가능하도록 관리하였다.



<그림 65> 비공개 YouTube 계정 및 시청 관리 페이지

3. 번역

3.1. 번역 업체 선정 과정

○ 본 과제 수행에 있어 안정적인 번역 인력 확보의 중요성이 높아, 플리토는 해외 번역 수급·관리 업체 입찰 공고를 게재하였다. 2021년 10월 28일에 업체 선정 및 계약을 완료하였다.

「2021 한국어-외국어 병렬 말뭉치 구축 사업」을 위한 번역인력/기관 수급·관리 용역 제 안 요 청 서

2021. 9.

주식회사 플리토

담당	주식회사 플리토 628호 임영후 매니저	연락처	Tel : 010-6402-7190 Email : younghoo.lm@plto.com
----	--------------------------	-----	---

1 용역 개요

본 용역은 문화체육관광부 국립국어원이 관리하는 「2021년 한국어-외국어 병렬 말뭉치 구축 사업」의 사업 수행주체인 본회사인 주식회사 플리토가 외국어 번역/검수를 위하여 발주하는 용역입니다.

1. 용역 개요

- 용역 명 : 「2021년 한국어-외국어 병렬 말뭉치 구축 사업」을 위한 번역인력/기관 수급·관리 용역
- 용역기간 : 계약 체결일로부터 2022.4.26.(화)까지
- 용역예산 : 일금 일십사억원(W1,400,000,000원 / 부가가치세 포함)

2 용역 목적

- (한국어-다국어) 번역인력/기관 수급
 - 한국어(출판) → 다국어(도착) 번역/검수를 위한 전문 번역사/기관의 “제용” 및 “발굴”을 통해 한국어-다국어 병렬 데이터(코퍼스) 구축
 - 다국어(도착) : 총 8개(베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아어(크메르), 필리핀 타갈로그어, 러시아어, 우즈베크어)
 - 베트남어의 경우 출판어를 베트남어(도착)을 한국어로 하는 525,000어절 추가
- <전문 번역사/기관의 기준> 도착어8개 다국어 기준이며, 아래의 요건에 부합해야함
 - 전문 번역사 : 도착어의 도착어 화자 또는 전문성을 보유한 인력으로 인어별 10인 내지 전문 번역가 - 한국어-다국어(도착어)의 번역이 가능한 인력을 인어당 10인 이상 보유
- 전문 번역사/기관은 사업 개시 후, 발주처에서도 일부 제공이 가능하며 용역기관은 발주처에서 제공한 인력/기관을 활용하여 용역을 운영함
- (한국어-다국어) 용역 수행 관리
 - 본 사업의 목표인 고품질의 “한국어-다국어 간 병렬 코퍼스” 구축을 위해 번역품질을 보장할 수 있도록 채용 또는 발굴한 번역인력/기관 인력 관리
 - 발주처가 요청하는 납품일정에 맞춰 번역인력/기관별 납기일을 설정하고 이와 관련한 준수를 위한 번역인력/기관 관리

〈그림 66〉 번역 수급·관리 업체 입찰 공고안

「2021년 한국어-외국어 병렬 말뭉치 구축 사업을 위한 번역인력·기관 수급·관리 용역」 제안서 평가결과

□ 평가개요

용역명	2021년 한국어-외국어 병렬 말뭉치 구축 사업을 위한 번역인력·기관 수급·관리 용역	제안업체수	3개
평가위원수	5명		

□ 평가위원 명단

성명	성명	성명	성명	성명
(인하)				

□ 평가 결과

순	위원명	평가항목	일일평가 점수			비고
1	OOO	인력구성	7	8	8	
		사업수행직접성	23	25	25	
		수행계획투수성	24	25	27	
		재안정·과활용성	6	8	7	
		합계	60	66	67	
2	OOO	인력구성	7	9	8	
		사업수행직접성	20	28	23	
		수행계획투수성	20	27	23	
		재안정·과활용성	6	9	8	
		합계	53	73	62	
3	OOO	인력구성	10	10	10	
		사업수행직접성	25	26	28	
		수행계획투수성	21	27	18	
		재안정·과활용성	8	10	10	
		합계	64	73	66	
4	OOO	인력구성	7	9	6	
		사업수행직접성	25	28	23	
		수행계획투수성	25	29	20	
		재안정·과활용성	8	9	6	
		합계	65	75	55	
5	OOO	인력구성	7	9	9	
		사업수행직접성	22	28	25	
		수행계획투수성	22	26	26	
		재안정·과활용성	5	10	8	
		합계	56	73	68	

표준계약서

발주처	주식회사 플리토 대표이사	
계약자	- 양호명 : 티씨티솔루션솔루션(TCW Solution) - 주소 : 서울특별시 중구구로로51, 18층 105호(종로타워) - 전화번호 : 02-3281-3800 - 사업자등록번호 : 779-50-00402 - 대표자 : 서양원	
	계약번호	플리토-코퍼스사업 계약 제2021-002호
	계약명	「2021년 한국어-외국어 병렬 말뭉치 구축사업」을 위한 번역인력/기관 수급·관리 용역
	계약기간	2021년 10월 28일 ~ 2022년 04월 26일
계약내용	계약금액	일금 일십사억원정 (W 1,400,000,000 / VAT 별도) - 월별 번역/양품 내역 등 의사소통에 따라 총액내에서 지급
	주식회사 플리토 대표이사(주) 계약상대자는 본문의 계약문서에 의하여 위와 같이 계약을 체결하고 신의에 따라 성실히 계약상의 의무를 이행할 것을 확인하며, 이를 증명하기 위하여 계약서를 작성하여 당사자가 기명날인한 후 2021년 10월 28일 2021년 10월 28일 붙임 1. 계약서 2. 사업수행계획서 1부 3. 산출내역서 1부	
발주처	기관명	주식회사 플리토
계약상대자	대표자	이정주
	기관명	티씨티솔루션솔루션
	대표자	서양원

〈그림 67〉 업체 선정 및 계약 진행

가. 데이터 구축(번역·검수 데이터) 작업 환경

- 고품질의 병렬 말뭉치 구축을 위하여 현지 번역 업체와 전문 번역가(이하 작업자)들이 현지 각 업체의 사무실 또는 해당 과제 수행을 위한 작업 환경을 마련하여 전문 번역 및 검수 작업을 수행하였다.
- 품질 향상을 위해 검수에 참여하는 작업자들의 평균적인 언어 수준은 번역에 참여하는 작업자들에 비해 높은 수준을 필수 조건으로 하였으며, 작업자들은 아래 업무를 주요 업무로 수행하였다.

〈표 35〉 번역·검수 작업자의 주요 업무 목록

작업자 주요 업무	
1	해당 언어별 번역 또는 검수
2	주 단위 번역·검수 완료 산출물 제출
3	완료 산출물에 대한 보완 지시 사항 대응

나. 번역 업체 및 프리랜서 전문 번역가 선정

- 본 사업에서 도출되는 고품질의 번역 데이터를 구축하기 위해 모국어 화자를 100% 구인해 활용할 수 있는 현지 번역 업체와 프리랜서 전문 번역가를 발굴하였고, 아래 절차를 거쳐 선별된 업체/인원만 작업에 활용하였다.

1) (원문 번역) 샘플 테스트 요청

- 품질 기준에 부합한 병렬 말뭉치 구축을 위해, 번역 업체와 전문 번역가의 수준 확인용 샘플 테스트를 진행하였다.

sid	구분	대분류	어절 수	Source	Translation
7174	문어체	신문기사	11	이후 양국은 원전 산업에서 포괄적으로 협업하기로 하며 협력 관계를 한층 끌어올렸다.	So'ngra ikki davlat atom energetikasi sohasida har tomonlama hamkorlik qilishga kelishib, hamkorlik aloqalarini yana bir pog'ona yuqoriga ko'tardilar.
166492	문어체	신문기사	14	프로페이스는 네볼라 지노믹스가 보유한 유전자 분석 서비스를 자사의 임상 진단 시험 서비스와 통합할 방침이다.	Proface Nebula Genomics korxonasiga tegishli genlarni tahlil qilish xizmatini bir korxonaning klinik diagnostik tekshiruv xizmatlari bilan birlashtirish rejalashtirilmoqda.
166404	문어체	신문기사	10	이번 프로모션이 출시 전 신제품을 맛볼 수 있는 기회인 셈이다.	Bu safargi aksiya yangi mahsulotni taqdim etishdan oldin uni tatib ko'rish imkoniyatini beradi
166388	문어체	신문기사	12	현재 이사회 승인을 모두 마친 상태이며 인수 작업은 내달 말에 완료될 예정이다.	Hozirda direktorlar kengashi to'liq tasdiqlangan bo'lib, qabul qilish jarayoni keyingi oy oxirigacha yakunlanishi kutilmoqda.
12	문어체	신문기사	13	이로 인해 경유 없이 비행시간 12시간 50분이면 일본에서 이탈리아 밀라노에 도착할 수 있다.	Shu sababli, Yaponiyadan Italiyaning Milan shahriga transfersiz 12 soat 50 daqiqada yetib borishingiz mumkin.
164754	문어체	신문기사	21	개발 중인 기술의 잠재적인 이익과 환경적 이점을 극대화하기 위해서는 배터리가 자동차 외 여러 다른 형태의 운송 수단에 쓰이기 적합하다고 판단한 것이다.	Ishlab chiqilyotgan texnologiyaning potensial foydalari va ekologik afzalliklarini maksimal darajada oshirish uchun, batareya avtomobillardan tashqari har xil turdagi transport vositalarida foydalanishga mos deya baho berildi.
70052	구어체	대화하기	7	골프가 은근히 쉬운 건데 많이들 부담스러워 하시더라고요.	Golf juda oson, lekin ko'pchilik buni qiyin deb o'ylaydilar.
70054	구어체	대화하기	6	카레랑 돈가스랑 같이 먹으면 정말 맛있겠는걸?	Karri va cho'chqa go'shti kotletini birgalikda yesa juda mazali bo'lsa kerak-a?
15029	구어체	대화하기	7	온 김에 우리 미친 듯이 먹고 놀자구!	Kelganimizga yarasha, kelinglar mazza qilib ovqatlanib o'ynaylik
15032	구어체	대화하기	5	원래 탈색하면 머리 많이 상하죠.	Umuman olganda sochlarni oqartirganda ular juda yomonlashadi.
15004	구어체	대화하기	11	잠을 못 자도 몇십 분 동안 눈 감고 쉬는 게 중요해요.	Uxlay olmasangiz ham, bir necha daqiqa ko'zlarni yumib dam olish muhimdir.

〈그림 68〉 샘플 테스트 진행 예시

2) 결과물 확인 : 번역 지침과 대조

○ 샘플 테스트에서 제공된 번역 지침을 준수하였는지를 확인하기 위한 검수 작업을 진행하였다.

sid	Category	Category	어절 수	SOURCE	TRANSLATED	PROOFREAD	NOTES
166492	문어체	신문기사	14	프로페이스는 네볼라 지노믹스가 보유한 유전자 분석 서비스를 자사의 임상 진단 시험 서비스와 통합할 방침이다.	Proface berencana untuk mengintegrasikan layanan analisis genetik Nebula Genomics dengan layanan uji diagnostik klinisnya sendiri.		
166404	문어체	신문기사	10	이번 프로모션이 출시 전 신제품을 맛볼 수 있는 기회인 셈이다.	Promosi ini merupakan kesempatan untuk mencicipi produk baru sebelum diluncurkan.		
166388	문어체	신문기사	12	현재 이사회 승인을 모두 마친 상태이며 인수 작업은 내달 말에 완료될 예정이다.	Saat ini, semua persetujuan dewan telah selesai, dan akuisisi diharapkan akan selesai pada akhir bulan depan.	Saat ini, semua persetujuan dewan telah selesai, dan akuisisi rencananya akan selesai pada akhir bulan depan.	예정 = rencana
12	문어체	신문기사	13	이로 인해 경유 없이 비행시간 12시간 50분이면 일본에서 이탈리아 밀라노에 도착할 수 있다.	Karena itu, waktu penerbangan 12 jam 50 menit tanpa transit dapat ditempuh dari Jepang ke Milan, Italia.		
164754	문어체	신문기사	21	개발 중인 기술의 잠재적인 이익과 환경적 이점을 극대화하기 위해서는 배터리가 자동차 외 여러 다른 형태의 운송 수단에 쓰이기 적합하다고 판단한 것이다.	Untuk memaksimalkan potensi dan manfaat lingkungan dari teknologi yang sedang dikembangkan, diputuskan bahwa baterai akan cocok untuk digunakan dalam bentuk transportasi lain selain mobil.	Untuk memaksimalkan potensi keuntungan dan manfaat lingkungan dari teknologi yang sedang dikembangkan, diputuskan bahwa baterai akan cocok untuk digunakan dalam bentuk transportasi lain selain mobil.	이익 빠짐
70052	구어체	대화하기	7	골프가 은근히 쉬운 건데 많이들 부담스러워 하시더라고요.	Bermain golf itu mudah, tapi banyak orang menganggapnya sulit.		
70054	구어체	대화하기	6	카레랑 돈가스랑 같이 먹으면 정말 맛있겠는걸?	Makan kare dengan kutlet daging pasti sangat enak.		
15029	구어체	대화하기	7	온 김에 우리 미친 듯이 먹고 놀자구!	Ayo kita makan dan bermain sampai puas!	Kita sudah datang, Ayo kita makan dan bermain sampai puas!	온 김에 빠짐
15032	구어체	대화하기	5	원래 탈색하면 머리 많이 상하죠.	Biasanya rambut akan rusak jika diwarnai.	Kalau Bleaching rambut memang rambut akan sangat rusak.	탈색 말고 염색으로 원래라는 말 빠짐
15004	구어체	대화하기	11	잠을 못 자도 몇십 분 동안 눈 감고 쉬는 게 중요해요.	Walaupun kamu tidak bisa tidur, sekedar memejamkan mata dan beristirahat beberapa menit saja itu penting.	Walaupun kamu tidak bisa tidur, sekedar memejamkan mata dan beristirahat beberapa puluh menit saja itu penting.	몇십 분 = puluh me
65042	구어체	대화하기	4	오디오 불러와 같이 가보자!	Ayo kita pergi untuk meminjam audio.		
14946	구어체	대화하기	8	엑스트라니까 그렇게 간편하게 오디오 보진 않을 것 같아.	Karena ini tambahan, kurasa audisinya tidak akan terlalu ketat.		

〈그림 69〉 샘플 테스트의 번역 지침 준수 여부 검수 예시

3) 품질 결과 점수 확정

○ 플리토 내부 인력을 활용하여 검수 완료된 샘플 테스트에 평가 점수를 반영하여 선별을 위한 지표로 활용하였다.

sid	Category	Category	어절 수	SOURCE	TRANSLATED	QC SCORE
70054	구어체	대화하기	6	카레랑 돈가스랑 같이 먹으면 정말 맛있겠는걸?	Makan kare dengan kutlet daging pasti sangat enak.	5
15029	구어체	대화하기	7	온 김에 우리 미친 듯이 먹고 놀자구!	Ayo kita makan dan bermain sampai puas!	4
15032	구어체	대화하기	5	원래 탈색하면 머리 많이 상하죠.	Biasanya rambut akan rusak jika diwarnai.	3
15004	구어체	대화하기	11	잠을 못 자도 몇십 분 동안 눈 감고 쉬는 게 중요해요.	Walaupun kamu tidak bisa tidur, sekedar memejamkan mata dan beristirahat beberapa menit saja itu penting.	4
65042	구어체	대화하기	4	오디오 빌리러 같이 가보자!	Ayo kita pergi untuk meminjam audio.	5
14946	구어체	대화하기	8	엑스트라니까 그렇게 깐깐하게 오디오 보진 않을 것 같아.	Karena ini tambahan, kurasa audisinya tidak akan terlalu ketat.	5
14898	구어체	대화하기	6	다음 번에는 다른 식당들도 한번 다녀와보자.	Lain kali ayo kita coba pergi ke restoran lain.	5
14873	구어체	대화하기	11	근데 너는 여기 와서 먹은 음식 중 가장 맛있었던 것 뭐야?	Ngomong-ngomong, apa makanan paling enak yang pernah kamu makan di sini?	5
14856	구어체	대화하기	4	여기는 뭐가 맛있는 집인데?	Masakan apa yang terkenal di restoran ini?	4
14798	구어체	대화하기	6	혼자 하게 되면 뽀뽀할 것 같아.	Sepertinya akan terasa canggung jika pergi ke sini sendirian.	5
14790	구어체	대화하기	4	천천히 하신다더니 너무 빠르네요.	Kamu bilang akan pelan-pelan, tapi cepat juga.	4
14785	구어체	대화하기	7	제가 집안에 바쁜 일이 많아서 못 나왔어요.	Aku tidak bisa keluar karena sibuk dengan pekerjaan rumah yang banyak.	5
14763	구어체	대화하기	7	택시 타면 아마 10분 정도 걸릴 걸?	Kurasa akan memakan waktu 10 menit jika naik taksi.	5
14141	구어체	대화하기	5	두통이 아침 저녁으로 심한 편이에요.	Sakit kepala lebih terasa di pagi dan malam hari.	3
Average						4.629

〈그림 70〉 샘플 테스트 품질 결과 점수 예시

4) 위 3)의 결과에 따른 번역 업체 선정

- 언어별 최소 20개 이상의 번역 업체와 프리랜서 전문 번역가를 대상으로 샘플 테스트를 진행하였으며, 5점 만점 중 4.5점 이상의 평가를 받은 업체와 번역가를 선별하였다.

3.2. 번역 인력 교육

가. 인력 교육의 목표

- 본 사업의 목적과 목표를 성공적으로 달성하기 위하여 선정된 번역 업체와 번역가 대상으로 지침의 전달과 함께 본 사업에서 요구하는 수준으로 번역 품질을 준수할 수 있도록 작업과 관련된 교육을 진행하여 방법론, 절차적 지식, 기술을 공유해 수행 인력의 전문성을 제고하였다.
- 사전에 구축된 수행 지침에 대한 교육을 진행함으로써 운영의 효율성을 높이고, 체계적인 절차로 작업을 진행함으로써 일관성을 높이고 오류율을 낮추어 신뢰도 높은 말뭉치 데이터를 구축하였다.
- 저작권과 개인 정보 보호 교육을 통해 국가 언어 공공재로 사용됨에 있어 제한 없는 말뭉치를 구축하였다.
- 작업자 대상으로 구체적인 단계별 검수 기준과 방법을 교육하여 정확하고 자연스러운 언어 데이터를 구축하였다.

나. 구축 및 검수 지침 교육 절차

- 구축 공정 시행 전에 말뭉치 구축 목표와 조직 구성 체계, 공정 절차 및 저작권, 개인 정보 보호 등 공통 사항에 대한 교육을 진행하였다.
- 구축 공정 시행 전 말뭉치 수집, 가공, 검수 공정별 목표와 절차, 방법 및 세부 사항에 대한 수행 지침 교육을 진행하였다.

다. 작업 지침

1) 한국어-외국어 번역 검수 공통 지침 및 언어별 세부 지침

- 학회에서 제작한 번역 검수 지침을 활용하여 번역 업체와 프리랜서 전문 번역가 대상 교육을 진행하였다.

2) 지침 요약본

- 작업자의 지침에 대한 이해를 높이기 위해 번역 검수 공통 지침에 대한 요약본을 제공하였다.

3) 작업 지침

- 말뭉치 번역 작업에 대한 이해를 돕기 위해 사업 안내 및 번역 작업에 수반되는 사항에 대한 교육을 진행하였다.

CONFIDENTIAL

제2절 어휘 및 표현 번역 지침

- 보통명사는 도적이 중 가장 적절하게 대응하는 단어 및 표현으로 의미역한다.
- 보통명사 중 한국 고유어 명칭을 영어로 번역할 때, 대용어를 찾기 어려운 경우, 한국의 로마자표기법 또는 도적이의 공식 한글 음역 표기 기준에 따라 음역한다. 즉, 의미역하거나 해설을 추가하지 않는다.

한글	영어	적합 여부
사주	sa-ju	O
sa-ju (four pillars of destiny)		X

- 고유명사는 로마자표기법 또는 도적이의 공식적인 한국어 음역 표기 기준에 따라 음역한다. 필요시 속칭을 나타내는 후부 요소를 도적이로 의미역하여 추가 표기할 수 있다.
- 도적이에 공식 한국어 음역 표기 기준이 없을 경우, 한국의 로마자표기법과 외래어 표기법 및 어원 지침을 준용하여 번역해 세부 지침에서 설정한 후, 그 기준에 따라 번역 및 표기한다.

제1항 인명

- 한국어 인명은 성명 전체를 한국의 로마자표기법 또는 도적이의 한글 음역 표기 기준에 따라 표기하는 것을 원칙으로 한다.
- 한국어 인명은 성과 이름의 순서로 띄어 쓴다. 이름은 붙여 쓰는 것을 원칙으로 하고, 음절 사이에 붙임표(.)를 쓰는 것을 허용한다. 이름에서 일어나는 음운 변화는 표기에 반영하지 않는다.

1. 고유명사 Proper Noun

분류	수용 방법	예시	참고
인명	도적이의 표기법으로 번역	한복산 → Han Bok-san / Han Boek-san	7 page
기관명 / 기업명	도적이의 표기법으로 번역	한전전력공사 → Korea Electric Power Corporation 한도 → Hando	7 page
지명	도적이의 표기법으로 번역	한도 → Hando 한도 → Hando	10 page
지명 / 문화유산	도적이의 표기법으로 번역	한도 → Hando 한도 → Hando	7 ~ 8 page
문화유산	도적이의 표기법으로 번역	한도 → Hando	10 page
문화유산 / 문화유산	도적이의 표기법으로 번역	한도 → Hando	8 ~ 9 page
문화유산	도적이의 표기법으로 번역	한도 → Hando	11 page

단, 도적이에서의 관행 및 관용적 표현이 있는 경우 이를 따라주시길 바랍니다.
(오든 번역은 도적이를 기준으로 자연스러운 문장이 되어야합니다.)

However, if there are any customs and idiomatic expressions in the arrival language, you can follow them.
(All translation should be a natural sentence based on the arrival language.)

2. 단어 / 용어 Words / Terms

분류	수용 방법	예시	참고
문화유산 / 문화유산	도적이의 표기법으로 번역	한도 → Hando	12 page
문화유산	도적이의 표기법으로 번역	한도 → Hando	12 page
문화유산 / 문화유산	도적이의 표기법으로 번역	한도 → Hando	12 page
문화유산	도적이의 표기법으로 번역	한도 → Hando	12 ~ 14 page
문화유산 / 문화유산	도적이의 표기법으로 번역	한도 → Hando	14 page
문화유산 / 문화유산	도적이의 표기법으로 번역	한도 → Hando	14 page

- **SID**: Original number of the sentence. Do not edit this column.
- **Category**: Type of sentence. Do not edit this column.
- **No of words**: Number of words in a sentence. Do not edit this column.
- **Note**: Characteristics of the source sentence for reference. Do not edit this column.
(e.g. Shortened form, slang, etc.)
- **Source**: Source sentence that needs translation. Do not edit this column.
- **Human-translated**: Write the translated sentence in the column.
- **Proofread**: Write the proofread sentence (or the human translated sentence) if it is correct in the column.

3. Request for Modification or Rework

- If the submitted file has the errors below, you could be asked to modify or rework it.
 - 1) Errors from non-adherence to the Guideline that Fitts has provided.
 - 2) Caught using machine translation.
 - 3) Changed/Omitted/Added a meaning of the source sentence, types, and other quality issues.
- We will make the request within 3 weeks after submission.
- Submit the modified file within 3 business days upon request.
- Specific guidelines will be given when requested to revise and/or redo the task.
- Modification and/or rework of the task may be requested several times if necessary.
- (In a case where the Fitts's internal inspection is not passed despite THREE times of modification and/or rework, it could be the reason for termination of the contract.)
- The project will be finished after all the modifications and/or rework is done.
- Payment will be made once the submitted file meets the requirement.

4. Delivery Date

- 1) The file will be delivered to you every Wednesday.
- 2) The file must be submitted by EOD next Tuesday to us.

Ex) File delivery date: 12/11/2021 File submission date: 23/11/2021 EOD

All Timeline is based on Korea Standard Time (GMT +9)

Breaking Delivery Date over THREE times could be the reason for termination of the contract.

〈그림 71〉 번역 및 작업 지침 예시

3.3. 번역 절차

가. 번역

- 번역 업무 수행에 앞서, 샘플 테스트로 선별된 작업자들(도착어 기준 현지 번역 업체와 프리랜서 전문 번역가)대상으로 지침 교육을 진행하였다.
- 작업자들은 제공 받은 번역 지침에 따라 파일 형식에 맞게 원문-번역문 문장쌍을 생성하였으며, 진행 시 표기 방식, 어휘 및 문법 채택 방식 등을 종합적으로 고려하여 수행하였다.
- 주 단위로 작업자들이 번역 할당량을 달성하면, 작업물은 플리토로 송부되고 플리토 내부 작업 관리자의 수정·보완 요청에 따라 추가 작업을 하였다.

SID	구분	어절수	원문	번역문
29020	구어체	7	더 필요한 것이 있으면 언제든지 말씀해 주세요.	Silakan memberi tahu saya kapan saja jika ada yang Anda perlukan lagi!
29021	구어체	4	마음에 드는 것으로 골라보세요.	Silakan memilih yang Anda sukai!
29022	구어체	5	밖으로 나가셔서 오른쪽으로 가시면 됩니다.	Anda cukup perlu ke luar kemudian belok ke kanan.
29023	구어체	4	아까 이 카드로 결제했어요.	Tadi saya sudah membayar menggunakan kartu ini.
29024	구어체	4	아이스 아메리카노 하나 주세요.	Saya pesan satu es amerikano.
29025	구어체	4	여기 주문하신 음식 나왔습니다.	Ini makanan yang Anda pesan.
29026	구어체	4	예약자 성함이 어떻게 되시나요?	Siapakah nama pemesan?
29027	구어체	4	인생은 짧고 예술은 길다.	Hidup ini singkat, sedangkan karya seni berumur panjang.
29028	구어체	4	저는 아무거나 잘 먹어요.	Saya bisa makan apapun.
29029	구어체	4	저도 같은 걸로 하겠습니다.	Saya juga akan memesan yang sama.
29030	구어체	4	주문은 조금 이따 할게요.	Saya akan memesan agak nanti.
29031	구어체	4	즐거운 쇼핑 시간 보내세요.	Selamat menikmati waktu berbelanja yang menyenangkan!
29032	구어체	4	즐거운 식사 되시길 바랍니다.	Semoga Anda menikmati makanannya.
29033	구어체	4	최대한 빨리 좀 부탁해요.	Mohon dilakukan sesegera mungkin!
29034	구어체	4	충분히 그러실 수 있습니다.	Anda bisa melakukannya dengan cukup baik.
			몇 호로 가면 되나요?	Saya harus pergi dengan kendaraan nomor

〈그림 72〉 번역 작업 예시

나. 검수

- 번역(원문-번역문 문장쌍을 생성하는 작업) 작업자보다 높은 수준의 작업자를 선별하여 번역 지침을 전달하였으며, 작업자는 지침에 따라 번역문 전수 검수를 수행하였다.
- 자연스러운 표현, 어색한 직역투, 오타자 등을 주로 교정하였으며 번역과 동일하게 주 단위 할당량을 달성하게 되면, 완료된 파일은 플리토로 송부되고 내부 작업 관리자의 수정·보완 요청에 따라 추가 작업을 진행하였다.

Sid	Category	No of words	Source	Human-translated	Proofread
9128	문어체	17	하박로이드의 선대 교체와 경쟁력 강화를 위해 올해 안에 확정수량 6척의 발주 계약을 체결할 것이라는 게 업계 분석이다.	Analisis industri mengatakan bahwa akan menandatangani kontrak pesanan 6 kapal dalam tahun ini untuk mengganti kapal Hapag-Lloyd dan memperkuat daya saing.	Analisis industri mengatakan bahwa akan menandatangani kontrak pesanan 6 kapal dalam tahun ini untuk mengganti kapal Hapag-Lloyd dan memperkuat daya saing.
9129	문어체	19	이날 건설사가 A 씨 접대를 위해 사용한 비용은 골프비 81만 원, 식사비 26만 원 등 모두 107만 원에 달했다.	Biaya yang digunakan oleh perusahaan konstruksi untuk melayani A pada hari itu mencapai 1.07 juta won, termasuk 810.000 won untuk golf dan 260.000 won untuk makan.	Biaya yang digunakan oleh perusahaan konstruksi untuk melayani A pada hari itu mencapai 1.07 juta won, termasuk 810.000 won untuk golf dan 260.000 won untuk makan.
9130	문어체	12	A 씨는 휴무일도 아닌 재택 근무일에 친구까지 불러 직무관련자로부터 환영을 받은 것이다.	A bahkan memanggil temannya pada hari kerja di rumah, bukan hari libur, dan mendapatkan sambutan dari seseorang yang berhubungan dengan pekerjaannya.	Seseorang berinisial A bahkan memanggil temannya pada hari kerja untuk datang ke rumah dan mendapatkan sambutan dari seseorang yang berhubungan dengan pekerjaannya.
9131	문어체	10	부득이한 사정으로 직무관련자와 골프를 같이 한다면, 사전에 소속부서장에 보고하도록 규정한다.	Jika bermain golf dengan orang yang berhubungan dengan pekerjaan karena keadaan yang tidak bisa anda hindari, Anda harus melaporkannya kepada kepala departemen terlebih dahulu.	Jika bermain golf dengan orang yang berhubungan dengan pekerjaan karena keadaan yang tidak bisa dihindari, maka harus melaporkannya kepada kepala departemen terlebih dahulu.
9132	문어체	11	업계에서는 LH 공사의 갑질 문화가 여전히 개선되지 않고 있다는 지적이 나온다.	Di industri muncul pendapat bahwa budaya semena-mena LH Konstruksi masih belum membaik.	Di dalam suatu industri masih memunculkan suatu tindakan semena-mena yang belum bisa diperbaiki.
9133	문어체	24	현대차는 미국 시장에 내놓은 디젤 차가 한 대도 없고 아직까진 출시 계획도 없는 만큼 현지 대형 SUV나 픽업트럭시장 진입을 위한 중장기 준비일 가능성도 제기된다.	Karena Hyundai Motor tidak memiliki satu mobil diesel di pasar Amerika Serikat dan belum ada rencana peluncuran, ada kemungkinan bahwa itu adalah persiapan jangka menengah panjang untuk memasuki SUV besar atau pasar truk pick-up di Amerika Serikat.	Karena Hyundai Motor tidak memiliki satu mobil diesel di pasar Amerika Serikat dan belum ada rencana peluncuran, ada kemungkinan bahwa itu adalah persiapan jangka menengah panjang untuk memasuki SUV besar atau pasar truk pick-up di Amerika Serikat.

〈그림 73〉 검수 작업 예시

다. 기계적 검수

- 플리토 내부 전문 인력의 5% 샘플링 검수에 앞서, 작업자가 검수 완료한 파일 대상으로 번역의 누락, 중복, 기호 통일, MT(기계번역) 유사도 검증 등 기계적 검수를 진행하였다.
- 번역 기준에 맞지 않는 경우, 번역문 수령 2영업일 이내로 작업자에게 수정·보완을 요청해 즉각적인 품질 개선을 진행하였다.

Sid	Category	No of words	Source	Proofread	dup_lang_2	quotation_m	len_ratio
10212	문어체	21	백소영 CJ제일제당 글로벌 마케팅전략담당은 “코로나19 시대에서 인간관계는 그 어느 때보다 중요하다”며 “업사는 비비고를 나눔 문화 매개체로 자리매김시키는데 역량을 집중하고 있다”고 말했다.	Di era COVID-19, "Hubungan antar manusia lebih penting dari sebelumnya", kata Baek So-young, kepala strategi pemasaran global di CJ CheilJedang.	TRUE	FALSE	FALSE
10558	문어체	22	테슬라 상하이 기가팩토리는 테슬라가 미국 외 지역에 지은 첫 번째 공장으로서 올 1월 모델3 생산을 시작해 올해 총 15만 대를 생산할 계획이다.	Shanghai Gigafactory Tesla adalah pabrik pertama yang dibangun oleh Tesla di luar Amerika Serikat.	TRUE	TRUE	FALSE
10712	문어체	19	이어 “공정위의 결정을 바꾸기 위해 구체적인 대안을 제시할 것”이라며 “이에 따라 기일결합 작업이 1~3개월 더 필요할 것으로 예상된다”라고 전했다.	"Kami akan menghadirkan alternatif spesifik untuk mengubah keputusan Komisi Perdagangan yang Adil", katanya.	TRUE	FALSE	FALSE
10750	문어체	14	한편 PUBG는 인도에서 텐센트를 통해 PUBG 모바일과 PUBG 모바일 라이트 두 게임을 서비스 해왔다.	Pasalnya, jika konfirmasi pemerintah India tentang PUBG Mobile memakan waktu lama, rencana untuk merilisnya di akhir tahun bisa jadi meleset.	FALSE	TRUE	TRUE
10763	문어체	15	삼성SDS 미국 법인인 타다와의 협업으로 넥스레저 유니버설의 적용 영역을 확대하며 블록체인 플랫폼 시장을 공략할 계획이다.	Ini memperluas ruang lingkup penggunaan di bidang keuangan, logistik, dan manufaktur, dan memimpin pasar platform blockchain perusahaan.	FALSE	TRUE	TRUE

〈그림 74〉 기계적 검수 작업 예시

3.4. 품질 향상 활동

가. 번역 지침

- 작업자들 대상으로 주요 공통 및 언어별 번역 지침과 주의사항을 사전에 안내하였다.
- 작업자는 전달 받은 지침과 주의사항을 숙지하여 번역에 착수하였다.
- 작업 진행 중에 발생하는 도착 언어별 특이 사항 등을 고려하여 상시로 지침(가이드라인)을 업데이트하였다.

나. 단계별 품질 향상 활동

1) 번역 업체/프리랜서 전문 번역가 테스트 및 교육 진행

- 번역 작업 시작 전 샘플 테스트와 간략한 교육을 진행하여 테스트 결과가 4.5점(5점 만점) 미만이며, 추가 교육을 진행해도 개선의 여지가 없다고 판단되는 번역 업체와 프리랜서 전문 번역가를 배제함으로써 품질 저하를 미연에 방지하였다.

2) 번역문에 대한 전수 검수

- 1차 번역을 진행한 작업자보다 높은 수준의 작업자를 선정하여 번역문 전수 검수를 진행하였으며, 단어 사용의 오류 및 어색한 직역투에 대한 수정으로 구축하는 말뭉치의 품질 향상을 추진하였다.
- 전수 검수 단계에서 수정률이 높은 경우, 1차 번역 작업자에 대한 재교육을 실시하고 추가 수정·보완을 요청하였다.

3) 플리토 언어 전문가의 5% 샘플링 검수

- 현지 번역 업체와 전문 번역가들의 번역·검수 결과물의 품질 보증을 위해 채용 대행 업체를 통해 모국어 화자를 작업 관리자로 채용하였다.
- 채용된 작업 관리자는 구축 프로세스를 운영하기 위한 조직 체계를 마련하고, 작

업자 대상으로 작업 진행을 위한 교육과 훈련을 진행하며, 작업자들의 작업량·정확도·작업 분류 등을 관리하였다.

- 번역 지침에 대한 이해나 언어 특성에 대한 이해 등이 다른 작업자와 비슷한 양상을 지니지 않은 작업자에 대해서는 별도의 교육과 수정·보완을 거쳤으며, 3번 이상의 수정·보완 작업 진행에도 품질이 향상되지 않는 경우 번역 및 검수 작업을 중단시켰다.

〈표 36〉 작업 관리자의 주요 업무 목록

작업 관리자 주요 업무	
1	주 단위 번역·검수 완료 산출물에 대한 샘플링 검수
2	작업자별 작업량 분배 및 달성도 추적·관리
3	완료 산출물에 대한 보완 지시 및 플리토 내부 서버 업로드

다. 기계적 검수

1) 번역물 정제

- 아래와 같이 번역문의 사소한 오류는 기호 종류를 통일하고, 불필요한 문자 등을 제거하는 기계적 검수 정제 작업을 진행하였다.

〈표 37〉 기계적 검수 정제 내용

정제 내용
줄바꿈, 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

2) 구조적 검증

- 아래와 같이 번역문의 사소한 오류는 기호 종류를 통일하고, 불필요한 문자 등을 제거하는 기계적 검수 정제 작업을 진행하였다.
- 엑셀 형식으로 진행한 번역 파일은 아래 기준에 따른 구조적 검증을 거치며, 검증을 통과하지 못한 문장은 작업자가 재번역 및 재검수를 진행하였다.

〈표 38〉 구조적 검증 기준과 불합격(Fail) 상황

검증 기준	불합격
중복	원문과 번역문, 서로 다른 번역문이 중복되는 경우
텍스트	원문 혹은 번역문이 누락된 경우, 번역문이 숫자 및 특수문자로만 구성된 경우
비율	원문과 번역문의 비율이 극히 차이 나는 경우
언어	목표 번역문 언어 외의 언어가 들어간 경우
MT 유사도	번역문과 구글 번역기의 번역문의 유사도가 90% 이상인 경우

라. 영어 데이터 품질 향상 활동

- 추가로 제안한 한국어-영어 병렬 말뭉치는 플리토에서 자체적으로 아래와 같이 품질 향상 활동을 진행하였다.

〈표 39〉 영어 데이터 품질 향상 진행 내용

단계	수행 내용	세부 사항
1	번역문에 대한 전수 검수 (proofreading)	- 번역문 전체 대상 문법 오류, 오역 등을 교정 - 부적합 시 번역 업체 및 전문 번역가에게 재수정 요청
2	MT 유사도 및 기계적 검증	- 번역문과 번역기의 유사도 확인 - 부적합 내용, 중복 내용 등 기계적 처리 가능 부분 처리 - 부적합 시 번역 업체 및 전문 번역가에게 재수정 요청
3	언어 전문가 1차 5% 샘플링 검수	- 전수 검수와 기계적 검증을 통과한 데이터 대상 일부 추출하여 도착어 기준 모국어 화자인 플리토 언어 전문가가 샘플링 검수 진행 - 부적합 시 번역 업체 및 전문 번역가에게 재수정 요청
4	언어 전문가 2차 전수 검수	- 어색한 표현, 오타자, 문장구조 등을 교정 - 국내에서 가장 많이 사용되는 미국식 영어 표현을 위주로 수정·보완하여 번역문에 통일성을 부여

4. 번역 검수 및 감수

4.1. 번역 검수 인력 관련

- 외국어 번역 과정에서 모어 화자 수준의 언어 숙달도가 아니면 판단하기 어려운 미세한 오류들이 발생할 수밖에 없기 때문에, 학회에 소속된 해외 언어 전문가들이 능숙한 이중 언어 화자로서 언어 데이터의 검수 작업을 성공적으로 수행하였다.
- 8개 언어 관련 석·박사 대학원생 이상으로 구성된 검증 전문 인력을 구성하여 언어별 번역 검수를 진행하였다. 해당 언어를 고급 이상의 수준으로 구사하는 자 혹은 관련 지역 교육 경험이 풍부한 한국어 교수, 석·박사 과정 이상의 인력으로 문법상의 오류뿐만 아니라 맥락상의 자연스러움까지도 검수를 진행하였다.
- 국내에서 번역 전문 인력을 확보하기 어려운 특수외국어의 경우, 해외에 거주하는 한국어 전공 원어민을 통해 검수를 진행하였다. 해외 거주자들의 검수비는 보조사업 관련 담당자에게 사전 검토를 거친 뒤 해외 송금을 진행하였다.

〈표 40〉 해외 거주자 대상 해외 송금 처리 과정

(1)	사전 검토: 기획재정부 국고보조금 담당자로부터 사업 특성상 해외 송금이 필요하다면 가능하다는 답변, 문화체육관광부 보조사업자 담당자로부터 e나라도움 시스템 활용과 증빙만 준비하면 가능하다는 답변, e나라도움 시스템 사용 매뉴얼 94~95쪽의 집행 예시 등을 검토하였다.
(2)	이체: '해외 송금'을 사유로 인건비에 해당하는 원화를 보조사업자 계좌로 이체한 후, 집행 금액을 해외 송금 처리하고 차액을 집행 부분 취소 처리하였다.

- 번역 검수팀은 언어별로 팀 내 총괄적인 관리·감독과 원문의 내용을 세밀하게 분석하고 전달하는 역할을 수행하는 품질 관리팀장, 번역 검수원의 교육·관리를 담당하는 검수 관리팀장, 내용 검수를 담당하는 번역 검수원으로 구성되며, 단순 문법이나 어법의 오류가 아닌 맥락과 상황에 따른 세련된 표현으로 다듬는 역할을 수행하였다.
- 검수 관리팀장은 번역 검수원들과 동일한 번역 검수를 수행하며 번역 검수원들의 검수 자료를 10% 샘플 점검하여 작업자별 오류 유형과 오류율을 분석하였다. 검수 오류율이 높은 작업자를 대상으로 재교육을 실시하고, 이후에도 시정되지 않을 경우 인력을 교체하였다.

- 과정에서 검출된 오류를 분석하고 유형화하여 검수 과정의 효율성을 높이고 추후 오류 재발 방지 교육 및 수정 지침의 근거로 삼았다.
- 번역 검수 인력은 최초에 언어별로 4~5명으로 구성하였으나 정해진 기간 내에 검수를 완료하고자 검수 현황에 따라 적게는 10명, 많게는 20명으로 검수 인력을 확충하였다. 자세한 검수 인력은 아래의 표와 같다.

〈표 41〉 번역 검수 인력

언어	베트남어	인도네시아어	태국어	인도 힌디어	캄보디아 크메르어	필리핀 타갈로그어	러시아어	우즈베크어	베트남어- 한국어	계
계획 인원	5	4	4	4	4	4	4	4	3	36
총 참여 인원	16	16	10	11	14	20	15	11	7	120

4.2. 최종 감수 인력 관련

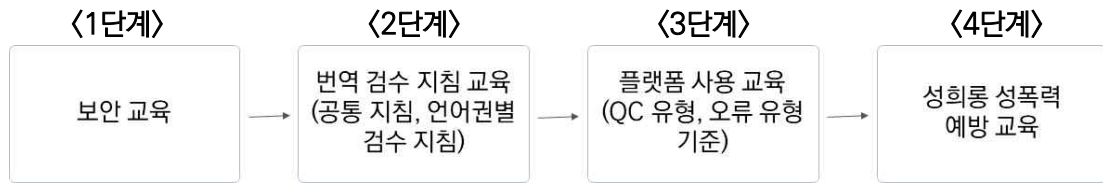
- 학회는 인도네시아 국립대학교, 필리핀 국립대학교, 베트남 호찌민시 국립인문사회과학대학교, 캄보디아 왕립프놈펜대학교, 인도 네루대학교, 태국 부라파대학교, 우즈베키스탄 사마르칸트대학교, 러시아 이르쿠츠크주립대학교 등 인적 네트워크를 통해 다양한 국가의 언어 전문가 풀을 구축하고 긴밀한 협력 관계를 유지하고 있다.
- 해당 지역을 대표하는 수준 높은 대학의 한국어 및 언어 전문가 중 한국어 사용 능력이 뛰어난 전공 교수를 최종 감수 인력으로 우선적으로 섭외하였다.
- 도착어를 모국어로 사용하는 해외 원어민 교수를 최종 감수자로 구성하여 번역 감수의 완성도를 확보하였다.
- 전체 구축 수량의 10%를 무작위로 샘플을 추출하여 내용 검수 항목을 중심으로 최종 감수를 실시하였다.
- 최초의 최종 감수 인원은 각 언어별로 1명이었으며, 구체적인 인원은 아래의 표와 같다. 베트남어권은 최종 감수자의 개인적인 사유로 인원이 충원되었고 러시아어권은 우크라이나 침공으로 인한 특수 상황으로 러시아 현지에서 있는 감수자와의 소통이 어려워 부득이하게 충원되었다.

〈표 42〉 최종 감수 인력

언어	베트남어	인도네시아어	태국어	인도 힌디어	캄보디아 크메르어	필리핀 타갈로그어	러시아어	우즈베크어	베트남어- 한국어	계
계획 인원	1	1	1	1	1	1	1	1	1	9
총 참여 인원	3	1	1	1	1	1	5	1	1	15

4.3. 번역 검수·감수 인력 교육

가. 검수·감수 지침 교육 절차



〈그림 75〉 검수·감수 지침 교육 절차

나. 검수·감수 지침 교육 내용

1) (1단계) 보안 교육

- 보안 교육을 실시함으로써 검수 및 감수 인력 전원이 데이터 보안에 각별한 주의를 기울이도록 하여 보안 사고를 미연에 방지하였다.
- 보안 교육 내용은 아래와 같다.
 - 사업 과정에서 산출되는 모든 데이터를 복제하거나 외부에 배포하는 것을 금지한다.
 - 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.

2) (2단계) 번역 검수 지침 교육

- 검수·감수 지침 교육의 목표
 - 사업 목적과 목표를 성공적으로 달성하기 위하여 검수 지침에 대한 교육을 실시하고 방법론, 절차적 지식, 기술을 공유함으로써 사업 수행 인력의 전문성을 함양한다.
 - 사전 설계된 수행 지침에 대한 교육을 통하여 효율적이고 체계적인 절차를 준수하여 일관되고 오류가 적고 신뢰할 만한 말뭉치를 구축하도록 한다.
 - 저작권과 개인 정보 보호에 대한 교육을 통하여 국가 언어 공공재로 사용하는 것에 제한이 없는 말뭉치를 구축한다.

- 구체적인 단계별 검수 기준 및 검수 방법 교육을 통해 정확하고 자연스러운 언어 자료를 구축한다.

○ 검수·감수 지침 교육 내용

- 지침 교육 내용은 공통 교육과 언어별 검수 지침 교육으로 구성된다.
- 공통 교육은 본 사업에 참여하는 모든 언어권의 번역 검수원들이 공통적으로 받는 교육으로 ‘사업 소개’와 ‘번역 검수 공통 지침’으로 구성된다. ‘사업 소개’는 본 사업의 이해를 높이기 위한 내용으로 사업의 목표 및 범위, 사업 수행 체계, 병렬 말뭉치 구축 절차 등에 대해 소개하고 있다. 그리고 ‘번역 검수 공통 지침’은 번역 검수 지침의 이해를 높이기 위한 내용으로 번역 검수 공통 지침의 목적, 용어의 정의, 적용 범위 및 저작권 보호, 지침 적용 절차 등에 대한 내용으로 구성되어 있다.

〈표 43〉 공통 교육 내용

항목	교육 내용
사업 소개	- 사업의 목표 및 범위 - 사업 수행체계 - 구축 절차 
번역 검수 공통 지침	- 번역 검수 공통 지침의 목적 - 용어의 정의 - 적용 범위 및 저작권 보호 - 지침 적용 절차 

- 언어별 번역 검수 세부 지침 교육은 각 언어의 특징에 따른 교육이다. 각 언어의 특징에 따라 세부 지침을 작성하였으며, 본 사업에 참여하는 모든 번역 검수원들이 정확한 번역 검수를 할 수 있도록 해당 언어권의 번역 검수 세부 지침 교육을 실시하였다. 언어마다 구체적인 내용은 다르게 구성되었으나 대략적인 교육 내용은 기본 원칙, 고유명사, 어휘 및 표현, 문법 및 담화, 표기법, 관련 규범, 사전, 용례 정보 등으로 구성되었다.

〈표 44〉 언어별 번역 검수 세부 지침 내용

항목	지침 내용
언어별 번역 검수 세부 지침	- 기본 원칙 - 고유명사 - 어휘 및 표현 - 문법 및 담화 - 표기법 - 관련 규범, 사전, 용례 정보

3) (3단계) 플랫폼 사용 교육

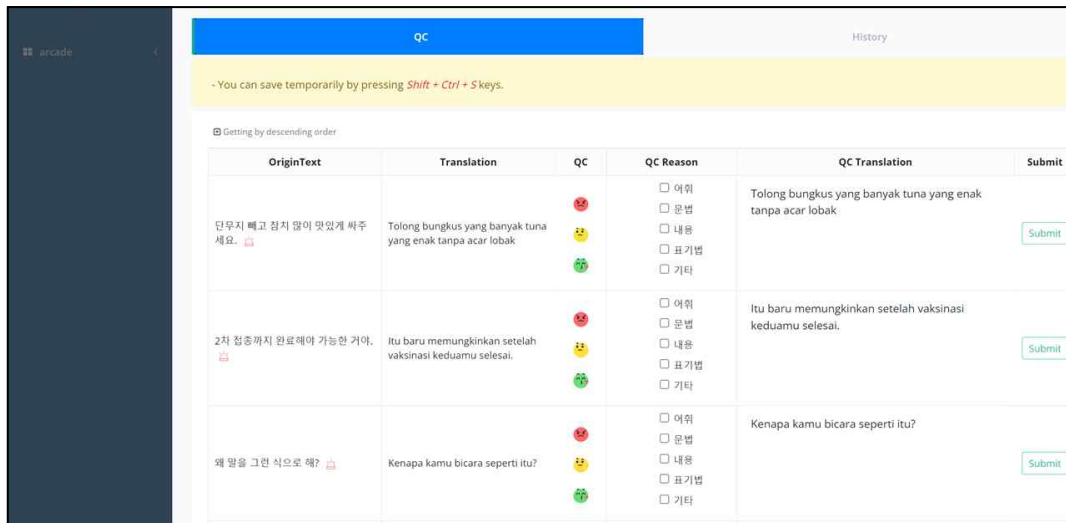
(1) 검수용 플랫폼

- 모든 번역 검수 인력을 대상으로 개인 계정을 지급하여 본 사업에 최적화된 플랫폼에서 번역 검수를 실시하였다.
- 번역 검수원이 검수용 플랫폼에서 로그인한 뒤에 문어체와 구어체를 선택할 수 있도록 플랫폼을 구성하였다.

The screenshot shows a web interface for quality control (QC) tasks. It contains two identical sections. The top section is titled '75. [Private]:[ko-vi] 국립국어원 2021 문어체' and the bottom section is titled '95. [Private]:[ko-vi] 국립국어원 2021 구어체'. Each section has a 'Korean' input field, a right-pointing arrow, a 'Vietnamese' output field, and a blue 'QC Start' button.

<그림 76> 검수용 플랫폼 문·구어 선택 예시

- 출발어 원문(OriginText)과 도착어 번역문(Translation)을 보고 품질 관리(QC) 유형을 판단한 후에 그에 따라 검수 작업을 수행하였다.



〈그림 77〉 검수용 플랫폼 작업 환경

- 품질 관리(QC) 유형은 좋음(Perfect), 보통(So so), 나쁨(Bad)으로 구성되며, 별도의 수정이 필요 없는 경우 좋음(Perfect), 사소한 오류로 수정이 가능한 경우 보통(So so), 오류가 많아 재번역이 필요한 경우 나쁨(Bad)을 선택하도록 하였다.
- 좋음(Perfect)은 번역 검수원이 도착어 번역문이 정확하다고 판단하여 별도의 수정 없이 제출하며, 보통(So so)은 품질 관리 이유(QC Reason: 어휘, 문법, 내용, 표기법, 기타) 중에 한 가지 이상을 선택하고 품질 관리 번역(QC Translation)에 도착어 번역문을 수정해야 한다. 그리고 나쁨(Bad)은 도착어 번역문에 치명적인 오류가 있는 것으로 번역 검수원이 나쁨(Bad)을 선택하면 재번역 처리가 된다. 번역 검수원은 도착어 번역문을 수정하지 않지만 어떤 오류가 있는지 품질 관리 이유(QC Reason) 중에 한 가지 이상을 선택해야 한다.

〈표 45〉 품질 관리(QC) 유형별 검수 방법

QC 유형		내용
Perfect		· 별도 수정 없이 검수문 제출
So so		· 2~3개의 사소한 오류가 있을 시 선택 · 알맞은 QC Reason(오류 유형)을 체크 후 QC Translation을 수정하고 검수문 제출
Bad		· 오류가 많거나 심각한 오류가 있을 시 선택 · QC Reason(오류 유형)만 클릭하고 QC Translation은 수정하지 않고 검수문 제출 · Bad로 선택한 문장은 재번역

- 번역문에 오류가 있을 시(So so, Bad)에 선택하는 품질 관리 이유(QC Reason, 오류 유형)는 어휘, 문법, 내용, 표기법, 기타로 나누어진다.
- ‘어휘’는 두세 개의 어휘를 한국어 원문에 적합하도록 수정하면 되는 경우, ‘문법’은 어순, 시제, 상, 수동태 등 문법적인 요소와 관련된 오류가 있는 경우, ‘내용’은 한국어 원문의 내용을 잘못 이해하여 번역하였거나 원문에 없는 내용이 번역문에 추가되거나 원문에 있는 내용이 번역문에 빠진 경우, ‘표기법’은 오자와 탈자, 문장 부호, 띄어쓰기 등 표기와 관련된 오류가 있는 경우, ‘기타’는 어휘나 문법적인 오류가 없으나 지나친 직역으로 실제 도착어 언중이 사용하지 않는 어색한 표현인 경우를 의미한다.

〈표 46〉 오류 유형의 기준

QC 유형	내용
어휘	· 두세 개의 어휘만 변경하여 오류가 해소될 경우
문법	· 문법(어순, 시제, 상, 수동태 등)과 관련된 오류가 있을 경우
내용	· 원문의 내용을 잘못 번역하였거나 불필요한 내용이 첨가된 경우
표기법	· 오탈자 및 문장 부호 등 표기와 관련된 오류가 있을 경우
기타	· 문법적으로 오류가 없으나 지나친 직역으로 인해 어색한 표현인 경우

- 번역 검수 과정에서 나타난 오류 유형은 전반적으로 어휘 오류의 비율이 가장 높았고 내용 오류, 문법 오류, 표기법 오류, 기타 오류의 순으로 높게 나타났다.

〈표 47〉 언어별 번역 검수 오류 유형 결과

언어	어휘	문법	내용	표기법	기타
베트남어	57.9%	12.5%	16.7%	7.7%	5.1%
인도네시아어	47.3%	5.9%	22.3%	11.4%	13.1%
태국어	46.2%	10.0%	9.4%	29.0%	5.4%
인도 힌디어	37.8%	15.0%	33.1%	6.2%	7.9%
캄보디아 크메르어	36.4%	10.2%	33.0%	11.6%	8.8%
필리핀 타갈로그어	57.8%	23.2%	10.7%	7.9%	0.4%
러시아어	31.7%	25.7%	18.9%	11.9%	11.8%
우즈베크어	26.5%	19.9%	19.0%	17.0%	17.5%
베트남어-한국어	54.3%	5.7%	26.0%	11.2%	2.8%
합계	45.9%	14.8%	19.3%	12.1%	7.8%

- 언어권마다 오류 유형의 비율은 다양한 양상을 보였지만 공통적으로 어휘 오류의 비율이 가장 높은 것을 확인할 수 있었으며, 언어별로 구체적인 내용은 다음과 같이 정리하였다.

〈표 48〉 언어별 오류 원인

언어	목표
베트남어	• 어휘 오류가 가장 높게 나타났는데 베트남어는 호칭어가 다양하게 발달하여 성별, 나이 등에 따라 다르게 사용되는데 문맥에 더 적절한 호칭어로 수정한 경우가 많이 나타났다. • 한국어와 같이 높임말과 반말을 구분하여 사용하는데 한국어 원문에 적합하게 이를 수정하면서 나타난 오류와 의미가 통하는 직역이지만 실제 사용에 더 적합하도록 어휘를 수정한 경우가 있었다. • 내용 오류는 한국어 원문에 적합하게 번역문을 수정한 경우가 많았고, 문법 오류는 번역문의 어순이 바뀌는 경우, 베트남어에 발달한 문장의 마지막에 오는 조사를 수정한 경우가 있었다.
인도네시아어	• 문어와 구어의 어휘 쓰임이 달라 이를 좀 더 적절한 어휘로 변경하며 어휘 오류가 가장 높게 나타났다. • 또한 한자에서 유래한 한국어 어휘의 경우 이를 인도네시아어로 직역하면 어색하기 때문에 관련 어휘나 표현으로 의역해야 하는데, 문맥에 따른 적합한 어휘를 찾기 어려워 어휘 오류 발생이 빈번했다. • 내용 오류와 기타 오류가 많았는데, 이는 문장 성분 생략이 빈번한 한국어를 인도네시아어로 번역하며 나타난 오류로 보이며, 주어나 목적어가 생략될 경우 이를 피동의 형식으로 바꾸어 써야 자연스러운데 이를 직역하며 내용 오류가 발생하거나 뉘앙스 차이가 발생하는 기타 오류가 발생하였다.
태국어	• 태국어는 어휘와 표기법 오류가 많았다. 어휘 오류는 한국어 원문의 의미 정확성이나 문맥적, 관습적 적절성에 대한 번역사의 판단에 따라 재수정한 것이다. 표기법의 경우, 다른 언어에 비해 표기법 오류가 높게 나타났는데 그 이유는 태국의 왕립학술원의 규정이 상세한 편임에 따라 표기법 관련 오류가 많이 나타난 것이다. 다시 말해, 표기법에는 한국어의 음역 방식에 관한 것 외에도 띄어쓰기와 부호(? ! ") 등도 포함되어 있는데 기계 학습에는 큰 영향을 미치지 않을 것으로 보인다.

언어	목표
인도 힌디어	- 가장 큰 비중을 차지한 어휘 오류의 경우 더 나은 어휘로 번역하는 과정에서 오류로 분류된 것이 많았으며, 내용 오류는 문맥을 파악하기 어렵고 문장성분의 생략이 많은 한국어 구어의 특성 때문인 것으로 보인다. - 힌디어와 한국어는 언어 유형론적으로 어순이 동일하나 인용문에서는 어순이 달라지고 긴 문장에서는 관계사가 사용되며 좀 더 복잡한 양상을 보이기 때문에 이런 문법적 차이에서 오는 어순 오류가 있었다. 또한 조사가 누락된 경우가 많았고 특히 주어 뒤의 조사 누락 오류가 다수였다. 이는 인도 힌디어에도 후치사가 다양하게 쓰이나 격조사는 그 사용에서 한국어와 차이가 있는데 이로 인해 발생한 오류로 보인다. - 표기법에서는 잔드라빈두의 사용 오류가 대부분이었다. 잔드라빈두는 비음을 표시하는 표기법의 한 종류로 대부분의 인도 사람들이 이를 제대로 인식하지 않고 잔드라빈두를 빈두로 쓰면서 자주 나타나는 오류 중 하나이다.
캄보디아 크메르어	- 어휘 오류와 내용 오류가 가장 많이 나타났으며, 더 자연스러운 문장으로 수정한 것으로 보인다. - 문법 오류는 한국어와 달리 주어가 반드시 필요한 크메르어에서 적합한 주어로 수정된 경우가 많았으며, 어순과 연결 어미 오류가 있었다.
필리핀 타갈로그어	- 어휘 오류는 어휘적으로 다양한 어원을 가진 어휘가 많으며, 타갈로그어 어휘와 영어를 기반으로 한 어휘 중 실제 사용과 자연스러움에 따른 선택에 어려움이 있기 때문이다. - 문법 오류는 타갈로그어가 교착어로서 접사가 발달했고 어순이 자유롭게 바뀔 수 있으며, 특수 동사가 활용되는 등 문법이 복잡한 편이기 때문인 것으로 보인다. - 타갈로그어의 어순은 VSO로 한국어 SOV의 어순과 다르나, 문어에서는 영어식 어순 SVO를 자연스럽게 쓰고, 구어에서는 VSO 어순을 주로 사용하는 편인데 구어·문어의 주어와 문법적 호응을 세심하게 보지 않아서 생긴 오류가 있었다.
러시아어	- 어휘 오류의 비율이 가장 높았는데, 원문의 맥락의 의미를 더 적합하게 전달하기 위해 수정한 경우가 많았다. - 다음으로 비율이 높은 문법 오류의 경우, 원문에 없는 문장 성분을 러시아어로 번역하는 과정에서 나타나는 성·수의 일치 오류, 한국어와 러시아어의 어순 차이로 인한 오류 등이 많이 나타났다.

언어	목표
	표기 오류로는 러시아어에서 다양한 기능을 수행하는 쉼표 사용의 오류 등이 있었다.
우즈베크어	- 우즈베크어 유형별 오류를 살펴보면 어휘 오류가 26.5%로 가장 높게 나타났다. 한국어 원문의 의미가 왜곡되게 전달되는 부적합한 어휘 번역 유형이 빈번하게 나타났는데, 이런 경우 한국어 원문의 의미를 정확하게 전달하는 가장 적합한 어휘로 수정되었다. 다음으로 문법적인 오류가 19.9%를 차지하였다. - 같은 계통의 교착어 특성을 가진 언어로서 어순이 같다는 공통점이 있지만 주어의 사용, 격조사의 차이, 시제의 차이, 복수 표현의 차이 등의 문법적인 차이가 있으며, 이러한 문법적인 오류들이 나타났다. - 표기 오류는 단순한 철자 오류가 대부분이었고 숫자 표기 및 마침표 누락 등의 오류가 나타났다.

(2) 감수용 플랫폼

- 모든 감수 인력을 대상으로 개인 계정을 지급하여 본 사업에 최적화된 플랫폼에서 감수를 실시하였다.
- 검수용 플랫폼과 동일하게 문어체와 구어체를 먼저 선택하여 감수를 하도록 하였다.
- 번역 감수자는 원문(Origin Text)과 번역문(User Result), 검수문(QC User Result)을 검토하여 감수 결과(Final Assess)를 판정하였다.
- 검수원 아이디(User)를 선택하여 번역 감수를 진행하나 번역 감수자는 어느 번역 검수원인지 등 번역 검수원에 대한 구체적인 정보는 알지 못한 상태로 감수를 진행하였다.
- 검수문에 오류가 없으면 P(Pass), 오류가 있으면 F(Fail)를 클릭하고 검수문을 직접 수정하도록 하였다. 검수문에 오류가 있어 F로 판정할 경우에는 의견 작성(Comment) 부분에 어떤 오류가 있었는지 수정 이유를 반드시 작성하도록 하였다.

The screenshot displays the 'Manual QC User Management' interface. It includes a sidebar with a user profile and a main content area with search filters and a table of users. The table lists users with columns for No., User, Participating Events, Create Date, Status, Update Date, Event Assess Ratio, QC Count, Assessment Count, Assessment Rate, and Assessment Pass Rate. Below the main table, there is a detailed view for a specific user (11581800) showing their assessment results for two events.

No.	User	Participating Events	Create Date	Status	Update Date	Event Assess Ratio	QC Count	Assessment Count	Assessment Rate	Assessment Pass Rate
1	NIKL_koid0004	75 (Private) [koid] 국립국어원 2021	2021-10-25 10:46:29	Activated	2021-12-03 15:06:17	10	11	2	18.2%	50.0%

Origin ID	Origin	QC Status	User Result	QC User Result	Create Date	Update Date	Comment	Init Assess	Final Assess	Update	Done
11581800	이쪽으로 가시면 맛집이 나와요.	👍	Jika lewat sini, kau akan menemukan restoran enak.	Jika lewat sini, kau akan menemukan restoran enak.	2021-12-03 15:06:08			P	P	Update	✓
11581805	니체의 책이 재미있어?	👍	Apakah buku Nietzsche seru?	Apakah buku Nietzsche seru?	2021-12-03 15:06:08			F	P	Update	✓

〈그림 77〉 감수용 플랫폼 작업 환경 예시

4) (4단계) 성희롱 성폭력 예방 교육

- 사업에 참여하는 모든 인력을 대상으로 성희롱 성폭력 예방 교육을 실시하였다.
- COVID-19로 인하여 집체 교육이 불가하므로 온라인 회의 플랫폼에서 성희롱 성폭력 예방 교육 동영상을 같이 시청하였다.
- 성희롱 성폭력 예방 교육은 고용노동부가 제작 및 배포한 동영상으로 유튜브에서 재생할 수 있으며, 약 17분으로 구성되었다.
- 교육 동영상의 내용은 다음과 같다.
 - 직장 내 성희롱이란 무엇이고 어떻게 판단해야 하는가?
 - 직장 내 성희롱 예방 및 대응 관련 사업주의 의무
 - 직장 내 성희롱에 대한 주체별 대응
 - 사내 성희롱 해결과 예방
- 교육 후에 성희롱 성폭력 예방 교육 이수 확인서를 발급하여 교육 이수 현황을 관리하였다.



〈그림 78〉 성희롱 성폭력 예방 교육 실시

4.4. 품질 향상 활동 예시

○ 검수팀장의 10% 샘플링 재검수

- 검수된 번역 데이터의 10%를 샘플링 추출하여 검수 관리팀장이 재검수를 실시하였다. 검수 관리팀장이 특정 번역 검수원에 대한 암묵적인 평가가 이뤄질 수도 있다는 점을 고려하여 번역 검수원의 정보는 공개하지 않고 재검수를 실시하였다.

Origin	QC	QC Reason	QC Translation	재검수 QC	재검수 수정문	재검수 의견
뛰어 올라가는 수환의 옆으로 아파트 벽에 금이 가	So so	내용, 문법	В стороне от вскакивающего Су Хвана трескается	P		
하지만 이번 소니의 EVO 인수로 신뢰 문제를 일으	Perfect		Однако ожидается, что приобретение Sony EVO	P		
데브시스타즈가 서비스하고 있는 모바일 RPG '쿠키	So so	어휘	Мобильная RPG "Cookie Run: Kingdom", обслужи	P		
데국 애플 앱스토어 매출 6위, 구글 플레이 8위를 차	So so	표기법, 어휘	Игра заняла 6-е место по продажам в Apple	P		
터널의 다양한 조명이 보이고 바닥 마찰 노랫소리가	So so	기타, 어휘	В тоннеле видны разноцветные огни, слышен зв	P		
강진은 병주가 들어간 뒤에도 성형외과 건물을 한참	So so	어휘	Даже после того, как Бенджу вошел, Канджин до	F	Даже после того, как Бенджу вошел, Канджин	표기법
테슬라의 투자자인 체이스는 머스크가 규제 기관으	Perfect		Инвестор Tesla Чейз указал, что совет директор	P		
집단 소송은 앞서 개인 투자자인 크리스 리코가 지	So so	기타	Коллективный иск начался 18-го числа (по мест	P		
이스케이프와 링컨 코세어 모델 생산은 다음주 월요	So so	어휘, 문법	Производство моделей Escape и Lincoln Corsair ч	F	Производство моделей Escape и Lincoln	어휘
LG에너지솔루션은 스페인 투자에 대해 '미정'이라는	So so	어휘	LG Energy Solution придерживается своей "неопр	P		
국가 에너지-기후 계획에서는 2030년까지 1900MW 용	Perfect		Национальный план по вопросу энергетики и за	P		
우선 수소 철도 사업이 프랑스 옥시타니와 부르고뉴,	So so	어휘, 문법	В первую очередь проект водородной железной	P		
벨류시스템자산운용은 지난 2009년 설립된 벨류시	So so	내용	Value System Asset Management была основана	P		
또한 아이언맨, 닥터 스트레인지, 블랙 위도우, 캡틴	Perfect		Кроме того, был раскрыт игровой процесс разли	P		
아직 출시 시점은 미정이나 신형 카니발 역시 내년	Perfect		Хотя пока неизвестно, когда он будет выпущен,	P		

〈그림 79〉 검수팀장의 10% 샘플링 재검수 예시

- 검수 관리팀장의 재검수 결과를 해당 번역 검수원에게 공유하여 검수 관리팀장과 번역 검수원이 재검수 결과에 대한 의견을 자유롭게 주고받도록 하였다. 이 과정을 통해 검수 능력의 상호 증진을 도모하였다.

재검수 수정문	재검수 의견	재검수 결과에 대한 담당 검수원 의견	검수원 의견에 대한 검수 관리팀장 최종 의견
F Видно, как красиво раскачиваются волнообразны е фонари. Судя по брошюре, маршрут кажется хорошим. Давайте как следует посмотрим, когда коронавиру с пройдет, и хотя бы покатаемся на машине. Сколько стоит пакет картошки? Не забудь сегодняшний день , и давай выпьем! На это уходит 1/10 часть моей годовой зарплаты. Ведь в последнее время много и таких людей , кот орые специально ищут старые рестораны. По местным правилам вход в воду без купальной одежды запрещён. В недавнем интервью с местными СМИ Чи Дон С оп, генеральный директор отдела по производству аккумуляторов в SK Innovation, упомянул, что "Дж орджия является одним из регионов-кандидатов" в качестве места для строительства совместного за вода BlueOvalSK.	번역문 내용이 맞음 어휘 오류 어휘 기타 어휘 어휘 기타 오타 어휘 빠짐	재검수 결과 동의 재검수 결과 동의 재검수 결과 동의 재검수 결과 동의 재검수 결과 동의 재검수 결과 동의 재검수 결과 동의 그렇다면 много же и таких людей 으 로 수정을 추천합니다. 그래야 -잡아 부분이 번역이 됩니다. 재검수 결과 동의 재검수 결과 동의	- - - - - - - 잡아의 번역은 'ведь' 하나로 충분할 것 같습니다. - -

〈그림 80〉 검수 결과에 대한 검수 관리팀장과 검수원의 의견 교환 예시

- 검수 관리팀장의 재검수를 통해 오류율이 높은 번역 검수원은 재교육을 실시하였
다. 번역 검수원의 번역 오류가 높은 원인에 따라 번역 검수 지침을 재교육하거
나 샘플 문장을 검수하여 이에 대한 피드백을 함으로써 번역 검수 품질을 향상시
켰다.

ID	Original Text	Score	Reviewer Comment
1350	Maari mo itong magawa kahit itang beses.	5 (Perfect)	일치함
1351	Unang beses kong gawin ito, at walang tumulong sa akin.	5 (Perfect)	
1352	Gusto mo talagang pumunta sa swimming pool, hindi ba?	5 (Perfect)	
1353	Mula alas dose ng tanghali hanggang anong oras ito	5 (Perfect)	이름 was
1354	Gusto kong gayahin ang reaksiyon mo kasi nakakatawa	5 (Perfect)	
1355	Baka sakaling hindi magustuhan ito, kaya 'sinisuguradong	5 (Perfect)	The
1356	Pano ko dapat ipahayag ito?	5 (Perfect)	
1357	Siguradong napaka abala po ang araw mo ngayon.	5 (Perfect)	보내었나
1358	Tumagal ng 2 oras para lang makabalit ang chain?	5 (Perfect)	
1359	Sunubukan kong uhawin.	5 (Perfect)	
1360	Gusto kong magkaroon ng mayayn na pag-unay ngayon.	5 (Perfect)	
1361	Gusto kong pumunta sa restawaran ng rin, yung hindi ukat	5 (Perfect)	
1362	Napakasap ng pakiramdam ang paglalakad.	5 (Perfect)	
1363	Salamat sa pag-unawa sa akin bilang kabigay mo.	5 (Perfect)	
1364	Bakit paulit ulit patin mung sinuula ang panghong sulita	5 (Perfect)	The QC
1365	Sa palagay ko ay hindi madaling gawin ito sa unang	5 (Perfect)	
1366	Magkita ang kasama ng alawang lalaki, pero hindi mo	5 (Perfect)	지체됨
1367	Ang bagal na hating naglakad ng ganto na magkasama	5 (Perfect)	일치함

〈그림 81〉 번역 검수원 대상 검수 재교육 실시

- 두 문장 이상이거나 표기 오류가 있는 검수문 추가 점검
 - 전임 연구원과 보조 연구원이 1차로 검수문에서 두 문장 이상이 의심되는 문장을

추출하였다. 두 문장 이상이 의심되는 문장을 검수한 검수 인력을 대상으로 2차 확인을 요청하고 수정하도록 하였다.

- 베트남어, 인도네시아어 등과 같이 원어를 살려 로마자로 표기해야 하는 언어의 경우 번역문에 표기 오류가 나타나는 경우가 있다. 전임 연구원과 보조 연구원이 정확한 원어 표기를 확인하고, 표기 오류로 보이는 경우 해당 번역 검수원이 수정하도록 하였다.

Origin	QC Translation	마침표 수	수정문
레미디엔터테인먼트 관계자는 "텐센트가 게임 산업에 대한 광범위한 전문지식을 보유하고 있다는 것을 알고 있다"라며 "텐센트를 주주로 맞이하게 돼 기쁘다"라고 말했다.	«Мы знаем, что Tencent обладает обширным опытом в игровой индустрии, – сказал представитель Remedy Entertainment. – Мы рады, что Tencent стала нашим акционером».	2	문장 1개
맥쿼리 그룹 측은 "풍력과 태양열 등 250여 개 신재생에너지 관련 프로젝트에 투자하고 있다"라며 "수소는 우리가 주목하는 새 분야며 한국은 굉장히 중요한 시장으로 관심을 가지고 보고 있다"라고 전했다.	Macquarie Group заявила: «Мы инвестируем в более чем 250 проектов, связанных с новыми и возобновляемыми источниками энергии, такими как ветроэнергетика и солнечная энергия. Водород – это новая отрасль, которой мы уделяем внимание, и Корея с интересом рассматривает ее как весьма важный рынок».	2	문장 1개
24일 유코카캐리어스 등에 따르면 지난 21일(현지시간) 현대차 아이오닉 5의 유럽 초도 물량 3000대가 노르웨이 드람멘(DRAMMEN) 항구에 도착했다.	Согласно сообщению EUKOR Car Carriers и др. от 24-го числа, первая партия Hyundai IONIQ 5 в Европе размером в 3 тыс. единиц прибыла в порт Драммен в Норвегии 21-го числа (по местному времени).	3	문장 1개
실제 작년 2만5900 대를 판매해 전년(15만 대) 대비 판매량이 5분의 1 수준으로 줄었다.	Фактически в прошлом году было продано 25,9 тыс. единиц, что является одной пятой от продаж предыдущего года (150 тыс. единиц).	3	문장 1개
업계 관계자는 "빌제이코가 소송을 제기한 텍사스주가 이전부터 특허 보유자에게 유리한 판정을 내리는 것으로 알려져 있다"며 "항후 애플이 어떤 대응을 할지 주목해야 할 것"이라고 말했다.	Представитель отрасли сказал: «Известно, что штат Техас, где BillCo подала иск, ранее выносил решение в пользу патентообладателей. Мы должны обратить внимание на то, как Apple будет реагировать на это в будущем».	2	문장 1개
아울러 현재 연 10만대 전후의 생산량을 2025년 160만대까지 끌어올린다는 공격적인 목표를 세워두고 있다.	Кроме того, в настоящее время он ставит перед собой агрессивную цель по увеличению производства примерно со 100 тыс. единиц в год до 1,6 млн единиц к 2025 году.	2	문장 1개

〈그림 82〉 두 문장 이상 번역문 점검 및 수정 예시

○ 한국어 원문 이해 지원

- 복잡한 상황 맥락이나 전문 용어, 신조어 등으로 인해 번역 검수원이 한국어 원문을 이해하는 데에 어려움을 겪을 수 있다.
- 번역 검수원이 한국어 원문에 대해 질문을 할 경우 팀 내 한국어 모어 화자인 품질 관리팀장 및 보조 연구원, 전임 연구원이 원문의 상황 맥락과 의미를 파악하여 번역 검수원이 한국어 원문을 충분히 이해할 수 있도록 설명하였다.
- 다른 언어권 번역 검수팀과 내용을 공유할 수 있도록 구글 스프레드시트에 날짜, 유형, 한국어 원문, 질문 내용, 답변을 정리하였다.

한국어 원문의 의미 확인				
No.	일시	유형	원문	내용
1	2021.12.14	한국어 의미 설명	이탈리아 디자이너 루도비카 버가가 제작한 무아우아를스는 이번 제로트 내년 봄부터 마이애미, 세인트 바트, 햄튼 등 주요 명품 매장에서도 아마 큐아와 공동 제작한 판츠, 스카프 등을 전시 및 판매할 계획이다.	세인트 바트: St. Barts/St. Barths를 다 허웅 (지역명, 카리브해에 있는 섬) 햄튼: Hamptons(지역명, 미국 버지니아 주의 부촌)
2	2021.12.17	한국어 의미 설명	지난달 4위에 올랐던 타타는 전년 동기 대비 38.80% 감소한 1만 400대를 판매하며 6위로 내려앉았고 토요타도 22.5%가 감소하며 8312대를 판매해 7위를 기록했다.	타타: 인도의 회사 영문표기: TATA
3	2021.12.17	한국어 의미 설명	내년부터 시행되는 국제해사기구(IMO) 환경규제에 따라 저유황 선박 연료유(LSFO) 사용과 관련 선사들의 추가 bunker 비용을 보존하면서 부담을 덜어주기로 한 것이다.	bunker costs: bunker costs, 국가에서 다른 국가로 상품을 운송하는 모든 유형의 해상 운송료
4	2021.12.18	한국어 의미 설명	다미아니 측은 "매장에는 고도로 훈련된 직원들이 고객들과 동반해 다양한 제품을 확인할 수 있게 도와줄 것"이라고 밝혔다.	다미아니: Damiani 이탈리아 명품 주얼리 브랜드 이름
5	2021.12.19	한국어 의미 설명	아기라 유야는 V4의 슬로건인 "싸우거나 공존하거나 선택은 자유"를 표현할 예정이다.	"싸우거나 공존하거나 선택은 자유"는 V4 게임의 슬로건임.
6	2021.12.20	한국어 의미 설명	100억 동(약 4억8000만원)을 투자해 세운 가금류사프자동공급 라인인 하루100t 이상의 사료를 공급할 수 있다.	동은 베트남 화폐단위
7	2021.12.20	한국어 의미 설명	이번 만남은 콜롬비아 정부가 주한 콜롬비아대사관을 통해 '2019 콜롬비아 건설 콩그레스(Columbia Construction Congress)'에 HUG 측 참석을 요청하면서 이뤄졌다.	HUG: 주택도시보증공사의 약자
8	2021.12.22	한국어 의미 설명	온리스크와 캐시카로가 함께 제작한 갤럭시 S11 플러스 컨디션 이미지는 큰 관심을 받으며 많은 보도가 이루어졌는데 IT 정보 유출 분야에서 그들의 인지도를 가지고 있는 빙우주가 반론을 제기해 이목을 끌고 있다.	빙우주: IT 인플루언서, Ice Universe
9	2021.12.31	한국어 의미 설명	또 그를 내 전락통인 이주희 부사장과 정동학 부사장을 각각 전략실 관리총괄과 기획전략부장으로 발탁했다.	부사장보: AVP(assistant vice president)
10	2022.01.03	한국어 의미 설명	이어 "한국에 많은 전통 요리가 있지만, 많은 사람들이 인스턴트 식품인 라면을 즐겨 먹는 다"며 "남비라면"과 일회용 컵 라면을 즐겨 먹을 수 있는 제품도 있어 인기다"라고 덧붙였다.	남비라면: 냄비에 끓여 먹는 봉지에 든 라면

〈그림 83〉 한국어 원문에 대한 질문 및 답변 정리 예시

○ 번역 품질 개선을 위한 의견

- 번역문에서 자주 발견되는 오류, 원문 의미 오역 등의 예시를 정리하여 번역 업체 측에 의견을 전달하였다.
- 번역의 정확도와 검수의 효율성을 위해 품질 개선 의견이 향후 번역 작업에 반영 되도록 하였다.
- 다른 언어권을 담당하는 전임 연구원도 이해할 수 있도록 영어 번역문을 추가하여 내용을 공유하였다.

	원문	번역문/영문	비고
1	여기를 안내해주시는 분 아닌가요?	Разве это не вы привели меня сюда? Isn't it you who brought me here? If you eat all the time or don't eat at all, you will have to go back to the hospital. Where in front, you said?	여기 이 곳을 안내해 주는 사람 맞는지 확인하는 질문인데 화자를 여기에 데리고 온 사람 맞는지 질문
2	점심에만 잠이 오지 않는 걸로 부탁드립니다.	Пожалуйста, только не ложитесь спать во время обеда. Please, just don't go to bed during lunch.	(줄리는 약을 생각하고) 점심에만 잠이 오지 않는 걸로 부탁드립니다 문장 점심시간에 제발 잠 자지 말라는 문장으로 번역되었음
3	꼭 7시 정각에 부탁드립니다.	Прошу не задерживаться и прийти ровно в 19:00. Please do not be late and come at exactly 19:00.	번역문이 해당 문장에 대한 가능한 해석인데 아닐 수도 있기에 무난하게 '꼭 7시까지 부탁드립니다'로 충분히 번역 가능

〈그림 84〉 번역 품질 개선을 위한 의견 작성 및 전달 예시

5. 병렬 말뭉치 구축 및 메타 정보

5.1. JSON 포맷

가. 메타데이터(Metadata) JSON 형식

〈표 49〉 메타데이터(Metadata)의 JSON 형식

No.	속성명	타입	필수 여부	항목 설명
1	id	string	필수	말뭉치 파일 ID
2	metadata	object	필수	파일의 메타 정보
3	title	string	필수	원시 말뭉치: 국립국어원 [말뭉치 유형 구분] [파일 ID] 예) 국립국어원 신문 말뭉치 NWRW1900000001
5	creator	string	필수	생성자: 국립국어원
6	distributor	string	필수	배포자: 국립국어원
7	year	string	필수	말뭉치 구축 연도 예) 2019
8	category	string	필수	분류 예) 신문>전국 종합지
9	annotation_level	string	필수	분석 층위 예) 원시
10	sampling	string	필수	샘플링 방식 (본문 전체 / 부분 추출- 임의 추출 / 부분 추출- 특정 부분 추출 등)
11	document	array		
12	id	string	필수	문서의 ID
13	title	string		기사 제목 없다면 null
14	author	string		문서의 저자
15	publisher	string	필수	발간자
16	date	string		생성일
17	topic	string	필수	소분류
18	original_topic	string	필수	대분류

국립국어원 말뭉치 metadata JSON 형식									
※ 말뭉치 유형: 신문, 문어, 구어, 웹, 메신저									
※ 분석 층위: 원시, 형태분석, 어휘의미 분석, 개체명 분석, 구문 분석, 의미역 분석, 상호참조 해결, 무형 대용어 복원									
※ 요소의 필수 여부는 '필수 말뭉치유형', '필수 분석층위'가 모두 해당하면 필수다. 예를 들어 1수준의 'id'는 전체 말뭉치 유형, 전체 분석층위이므로 무조건 필수다.									
※ 반면, 2수준의 'sentence'는 전체 말뭉치 유형과 분석층위에 필수지만, 원시 말뭉치는 제외한다.									
※ 노란색 바탕의 요소는 동일 타입의 값이 반복되는 배열(array)을 말한다									
※ JSON 타입 약어: object(obj), array(arr), string(str), number(num)									
1수준	2 수준	3 수준	4 수준	5 수준	타입	필수 말뭉치유	필수 분석층위	설명	
id					str	전체	전체	* 말뭉치 파일 ID * 고유 ID로 중복이 없어야 함.	
metadata					obj	전체	전체	* 파일의 메타 정보	
	title				str	전체	전체	* 원시 말뭉치: 국립국어원 [말뭉치 유형 구분] [파일 ID] 예) 국립국어원 신문 말뭉치 NNRW1900000001	
	creator				str	전체	전체	* 생성자: 국립국어원	
	distributor				str	전체	전체	* 배포자: 국립국어원	
	year				str	전체	전체	* 말뭉치 구축 연도 예) 2019	
	category				str	전체	전체	* 분류 예) 신문>전국 종합지	
	annotation_level				arr (str)	전체	전체	* 분석 층위 예) 원시	
	sampling				str	신문, 문어, 구어, 웹	전체	* 샘플링 방식 (분문 전체 / 부분 추출 - 임의 추출 / 부분 추출 - 특정 부분 추출 등)	
	document				arr (obj)				
		id			str	전체	전체	* 문서의 ID * 고유 ID로 중복이 없어야 함.	
		metadata			arr (str)	전체	전체	* 문서의 메타 정보	
			title		str	전체	전체	* 기사 제목 * 구어의 경우 없다면 null	
			author		str	전체	전체	* 문서의 저자 ex) 신문일 경우 기사명, 대본일 경우 대본 저자 등으로 표시	
			publihsr		str	전체	전체	* 발간자	
			date		str	전체	전체	* 생성일	
			topic		str	전체	전체	* 주제 ex) 신문일 경우 기사 주제, 구어일 경우 대화 주제 등으로 표시	
			original_to pic		str	전체	전체	* 신문사에서 설정한 주제	

〈그림 85〉 국립국어원 메타데이터(Metadata) JSON 형식

나. document_paragraph JSON 형식

〈표 50〉 document_paragraph JSON 형식

No.	속성명	타입	필수 여부	항목 설명
1	document_paragraph	array	필수	
2	id	string	필수	문서의 ID
3	paragraph	array	필수	문서의 메타 정보
5	id	string	필수	문장 고유의 ID
6	form	string	필수	정제된 문장
7	original_form	string	필수	정제 이전의 문장

국립국어원 말뭉치 document_paragraph JSON 형식								
※ 말뭉치 유형: 신문, 문어, 구어, 웹, 메신저								
※ 분석 종류: 원시, 형태분석, 어휘의미 분석, 개체명 분석, 구문 분석, 의미역 분석, 상호참조 해결, 무형 대응어 복원								
※ 요소의 필수 여부는 '필수 말뭉치유형', '필수 분석종위'가 모두 해당하면 필수다. 예를 들어 1수준의 'id'는 전체 말뭉치 유형, 전체 분석종위이므로 무조건 필수다.								
※ 반면, 2수준의 'sentence'는 전체 말뭉치 유형과 분석종위에 필수지만, 원시 말뭉치는 제외한다.								
※ 노란색 바탕의 요소는 동일 타입의 값이 반복되는 배열(array)을 말함								
※ JSON 타입 약어: object(obj), array(arr), string(str), number(num)								
1수준	2 수준	3 수준	4 수준	5 수준	타입	필수 말뭉치유형	필수 분석종위	설명
document_paragraph					arr (obj)			
	id				str	전체	전체	* 문서의 ID * 고유 ID로 중복이 없어야 함.
	paragraph				arr (str)	전체	전체	* 문서의 메타 정보
		id			str	전체	전체	* 문장 고유의 ID
		form			str	전체	전체	* 정제된 문장
		original_form			str	전체	전체	* 정제 이전의 문장

〈그림 86〉 국립국어원 document_paragraph JSON 형식

다. 병렬 말뭉치 JSON 형식

〈표 51〉 병렬 말뭉치 JSON 형식

No.	속성명	타입	필수 여부	항목 설명
1	parallel		필수	괄호 안에 원자료, 대상 자료 명시
2	id	string	필수	문장 고유의 ID
3	source	string	필수	정제된 문장
5	target	array	필수	3차 감수된 문장
6	revision1	string	필수	1차 감수된 문장
7	revision2	string	필수	2차 감수된 문장

국립국어원 말뭉치 병렬 말뭉치 JSON 형식								
※ 말뭉치 유형: 신문, 문어, 구어, 웹, 메신저								
※ 분석 종류: 원시, 형태분석, 어휘의미 분석, 개체명 분석, 구문 분석, 의미역 분석, 상호참조 해결, 무형 대응어 복원								
※ 요소의 필수 여부는 '필수 말뭉치유형', '필수 분석종위'가 모두 해당하면 필수다. 예를 들어 1수준의 'id'는 전체 말뭉치 유형, 전체 분석종위이므로 무조건 필수다.								
※ 반면, 2수준의 'sentence'는 전체 말뭉치 유형과 분석종위에 필수지만, 원시 말뭉치는 제외한다.								
※ 노란색 바탕의 요소는 동일 타입의 값이 반복되는 배열(array)을 말함								
※ JSON 타입 약어: object(obj), array(arr), string(str), number(num)								
1수준	2 수준	3 수준	4 수준	5 수준	타입	필수 말뭉치유 형	필수 분석종위	설명
parallel(KOTH)								*괄호 안에 원자료, 대상자료 명시 parallel(KOTH): 병렬 원자료(한국어), 대상자료(태국어)로 구성된 병렬 말뭉치
	id				str	전체	전체	* 문장 고유의 ID
	source				str	전체	전체	* 정제된 문장
	target				arr (obj)	전체	전체	
		revision1			str	전체	전체	* 1차 감수된 문장
		revision2			str	전체	전체	* 2차 감수된 문장

〈그림 87〉 국립국어원 병렬 말뭉치 JSON 형식

5.2. 최종 데이터 구조

○ 최종적으로 아래 샘플과 동일 양식으로 데이터를 구축하였다.

[illegible]

〈그림 88〉 한국어-외국어 구축 데이터 샘플

5.3. JSON 예시

○ 원천 데이터

```
{
  "id": "WXOR2102204280",
  "metadata": {
    "title": "국립국어원 병렬말뭉치 원천 자료 문어 WXOR2102204280",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2021",
    "category": "무작위 구성",
    "annotation_level": [
      "원시"
    ]
  },
  "document": [
    {
      "id": "WXOR2102204270.1",
      "metadata": {
        "title": "",
        "author": "",
        "date": "",
        "publisher": "(주)웨이커",
        "topic": "산업",
        "original_topic": "산업"
      },
      "paragraph": [
        {
          "id": "WXOR2102204270.1.1",
          "original_form": "이로 인해 경유 없이 비행시간 12시간 50분이면  
일본에서 이탈리아 밀라노에 도착할 수 있다.",
          "form": "이로 인해 경유 없이 비행시간 12시간 50분이면  
일본에서 이탈리아 밀라노에 도착할 수 있다."
        },
        {
          "id": "WXOR2102204270.1.2",
          "original_form": "일본 학생들 사이의 인기 여행지 대부분이 유럽  
지역인 것으로 조사됐다.",
          "form": "일본 학생들 사이의 인기 여행지 대부분이 유럽 지역인  
것으로 조사됐다."
        }
      ]
    }
  ]
}
```

```

    }
  ]
},
{
  "id": "WXOR2102204270.2",
  "metadata": {
    "title": "",
    "author": "",
    "date": "",
    "publisher": "(주)웨이커",
    "topic": "산업",
    "original_topic": "산업"
  },
  "paragraph": [
    {
      "id": "WXOR2102204270.2.1",
      "original_form": "삼성전자의 스마트폰 시장 점유율 상승을 이끌고 있는 갤럭시 A 시리즈의 업그레이드 된 모델들이 출시된다.",
      "form": "삼성전자의 스마트폰 시장 점유율 상승을 이끌고 있는 갤럭시 A 시리즈의 업그레이드 된 모델들이 출시된다."
    },
    {
      "id": "WXOR2102204270.2.2",
      "original_form": "22일 삼성전자는 갤럭시 A30과 A50의 업그레이드 모델인 갤럭시 A30s와 A50s를 공개했다.",
      "form": "22일 삼성전자는 갤럭시 A30과 A50의 업그레이드 모델인 갤럭시 A30s와 A50s를 공개했다."
    },
    {
      "id": "WXOR2102204270.2.3",
      "original_form": "갤럭시 A30s의 경우 AP(애플리케이션 프로세서), 배터리, 화면 크기 등에서는 변화가 없었지만 해상도, 카메라, 트림의 추가 등에서 변화있다.",
      "form": "갤럭시 A30s의 경우 AP(애플리케이션 프로세서), 배터리, 화면 크기 등에서는 변화가 없었지만 해상도, 카메라, 트림의 추가 등에서 변화가 있다."
    }
  ]
}
]
}

```

〈그림 89〉 원천 데이터 JSON 예시

○ 분석 데이터

```
{
  "id": "NIXOR2102205260",
  "metadata": {
    "title": "국립국어원 병렬말뭉치 원천 자료 문어 NIXOR2102205260",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2021",
    "category": "인터넷 기반 신문",
    "annotation_level": [
      "월시"
    ]
  },
  "document": [
    {
      "id": "NIOR2102205260.1",
      "metadata": {
        "title": "",
        "author": "",
        "date": "",
        "publisher": "(주)웨이커",
        "topic": "산업",
        "original_topic": "산업"
      },
      "document_paragraph": {
        "id": "NIOR2102205260.1",
        "paragraph": [
          {
            "id": "NIOR2102205260.1.1",
            "original_form": "이로 인해 경유 없이 비행시간 12시간 50분이면 일본에서 이탈리아 밀라노에 도착할 수 있다.",
            "form": "이로 인해 경유 없이 비행시간 12시간 50분이면 일본에서 이탈리아 밀라노에 도착할 수 있다.",
            "target": "Oleh karena itu, dengan waktu penerbannya 12 jam 50 menit tanpa transit, kita bisa tiba di Milan, Italia dari Jepang.",
            "revision": {
              "revision1": "Oleh karena itu, dengan penerbannya 12 jam 50 menit tanpa transit, kita bisa tiba di Milan, Italia dari Jepang.",
            }
          }
        ]
      }
    }
  ]
}
```

```
        "revision2": "Oleh karena itu, dengan waktu pe  
nerbangannya 12 jam 50 menit tanpa transit, kita bisa tiba di Mil  
an, Italia dari Jepang."
```

```
    }
```

```
  },
```

```
  {
```

```
    "id": "NIOR2102205260.1.2",
```

```
    "original_form": "일본 학생들 사이의 인기 여행지 대부  
분이 유럽 지역인 것으로 조사됐다.",
```

```
    "form": "일본 학생들 사이의 인기 여행지 대부분이 유  
럽 지역인 것으로 조사됐다.",
```

```
    "target": "Telah dikaji bahwa sebagian besar tempa  
t wisata populer di antara para pelajar Jepang adalah wilayah Er  
opa.",
```

```
    "revision": {
```

```
      "revision1": "Telah dikaji bahwa sebagian besa  
r tempat wisata populer di antara para pelajar Jepang adalah wil  
ayah Eropa.",
```

```
      "revision2": "Telah dikaji bahwa sebagian besa  
r tempat wisata populer di antara para pelajar Jepang adalah wil  
ayah Eropa."
```

```
    }
```

```
  },
```

```
  {
```

```
    "id": "NIOR2102205260.1.3",
```

```
    "original_form": "통상 스마트폰 한 대에 20g, 전기차  
한 대에 28kg이 들어간다.",
```

```
    "form": "리튬은 통상 스마트폰 한 대에 20g, 전기차 한  
대에 28kg이 들어간다.",
```

```
    "target": "Pada setiap unit ponsel pintar biasanya  
mengandung 20 gram litium dan 28 kilogram pada setiap unit m  
obil listrik.",
```

```
    "revision": {
```

```
      "revision1": "Biasanya dimasukkan 20 gram liti  
um pada setiap unit ponsel pintar dan 28 kilogram litium pada s  
etiap unit mobil listrik.",
```

```
      "revision2": "Pada setiap unit ponsel pintar bia  
sanya mengandung 20 gram litium dan 28 kilogram pada setiap  
unit mobil listrik."
```

```
    }
```

```
  },
```

```

    {
      "id": "NIOR2102205260.1.4",
      "original_form": "2025년에는 공급량이 21만t, 수요량
이 32만6000t으로 그 격차가 더욱 커질 전망이다.",
      "form": "2025년에는 공급량이 21만t, 수요량이 32만 60
00t으로 그 격차가 더욱 커질 전망이다.",
      "target": "Pada tahun 2025, selisih antara jumlah p
asokan 210.000 ton dan jumlah permintaan 326.000 ton tersebut
diperkirakan akan makin besar.",
      "revision": {
        "revision1": "Pada tahun 2025 diperkirakan seli
sih antara jumlah pasokan 210.000 ton dan jumlah permintaan 3
26.000 ton akan makin besar.",
        "revision2": "Pada tahun 2025, selisih antara j
umlah pasokan 210.000 ton dan jumlah permintaan 326.000 ton t
ersebut diperkirakan akan makin besar."
      }
    }
  ]
}
},

```

〈그림 90〉 분석 데이터 JSON 예시

6. 보안 관리

6.1. 보안 관리 대상

- 데이터 구축을 수행하는 사업장과 함께 과업 수행에 참여하는 관계 시설의 보안, 정보 보안, PC 보안, 저장매체 보안, 산출물 보안 등 물리적·기술적 사항과 참여 인력을 보안 관리 대상으로 선정하였다.

6.2. 담당자 지정 체계

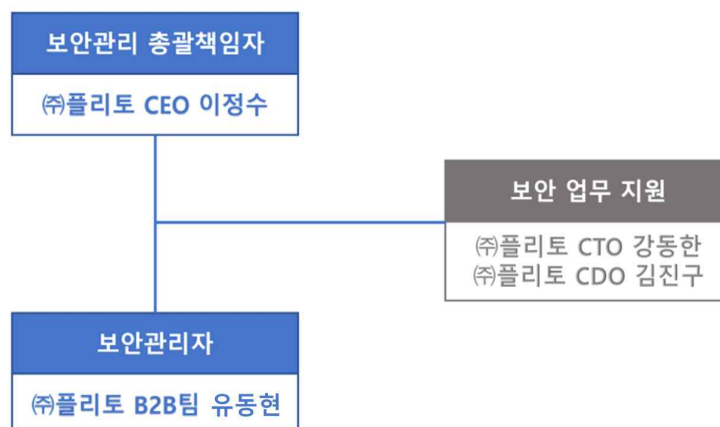
가. 보안 관리 총괄 책임자

- 본 사업과 관련된 인원·장비·자료에 대한 보안 관리 전반을 총괄하였으며, 본 사업과 관련된 보안 관리 업무를 일부 위임할 보안 관리자를 지정하고 감독하였다.

나. 보안 관리자

- 본 사업과 관련된 기술적, 물리적 보안에 대한 수행을 담당하였으며, 사업 수행 과정에서 전반적인 보안 활동을 수행, 감시, 통제하였다.

6.3. 담당자 역할 체계



<그림 91> 보안 관리 담당자 역할 분포

- 보안 관리 총괄 책임자: 본 구축 사업 내 보안 관리 업무 총괄 관리·감독
- 보안 관리자: 본 구축 사업과 관련된 기술적, 물리적 보안의 수행 책임
- 보안 업무 지원: 플리토 서버 및 플랫폼 관리자로 기술적 보안 지원

6.4. 보안 관리 방법

가. 물리적 보안

1) 사무실·주요 장비 설치 장소에 대한 출입 보안

- 사무 공간
 - 사무 공간의 출입문은 항상 닫아 두어야 하며, 공사나 시설 이동으로 인한 일정 시간 출입문 열림이 필요한 경우 보안 관리 총괄 책임자의 사전 승인 이후에 진행하였다.
 - 외부인의 방문이 있는 경우, 사업단 또는 플리토의 담당자가 외부인을 응대하도록 하며 지정된 장소 또는 공간 내에서만 회의 및 협의를 진행하도록 하였다.



〈그림 92〉 병렬 말뭉치 사업단 및 플리토 출입문 전경

- 보호 구역 출입 통제
 - 데이터 가공·검수 및 저장의 주요 공간인 서버가 보관된 서버실에 출입이 필요한 자는 보안 관리 총괄 책임자의 사전 승인 이후 출입하였다.
 - 사전에 출입 허가를 받은 자와 출입한 자의 출입 일시 등은 ‘서버실 출입 관리 대장’으로 하여 별도로 관리하였다.

2) PC 및 보조 기억 장치 반출·입 통제

○ 전산 장비

- 데이터 가공, 검수 및 추출 등에 활용되는 모든 전산 장비(PC, 노트북 등)는 전산 장비 대장에 등록하여 관리하도록 하며, 인가되지 않은 장비의 사용은 원천적으로 금지하였다.
- 무선랜, 외장형 HDD 및 USB 등 비인가 전산(통신)장비 무단 반입·사용을 금지하였다.
- 관리 대상 전산 장비의 반·출입이 필요한 경우 자료 무단 반출 여부를 확인하고 대장에 기록하도록 하였다.
- PC, 노트북 등은 아래 보안 관리 사항을 주기적으로 점검하고, 전체 점검 결과는 보안 관리자가 별도로 관리하도록 하였다.
 - 자리 비움 시 화면 보호기능 사용
 - 백신프로그램 및 보안 업데이트 상시 수행

○ 보조 기억 장치

- 보조 기억 장치는 보조 기억 매체 관리 대장을 통해 별도로 기록하도록 하였다.

3) 보안 관리 대장

○ 보안 관리는 매주 3개 항목 6개 분류에 대한 점검을 진행하였다.

○ 보안 점검 항목 및 분류 현황

- 서버 보안 : 플리토 개인 정보 암호화, 응용프로그램 개발 보안
- 네트워크 보안 : 정보 처리 시스템 접근 통제, 시스템 접근 권한 관리
- 근무 환경 보안 : 작업자 PC 보안, 소프트웨어 패치 관리

○ 보안 점검 실시 현황 : 매주 점검 실시

- 2021년 9월 1주 차부터 과제 종료 시까지 보안 점검을 실시하였다.

보안 관리 대장										
프로젝트명: 2021 한·외 병렬 양방향 구축 사업										
문서번호: KFPC-007				V1.0			보안관리 책임자(장): 이정수 (플리토 CEO) 보안관리 책임자(부): 권가현 (플리토 B2B팀 매니저)			
	보안 항목 대분류	보안 항목 세부분류	9월 1주차 Week 1 (08.30~09.03)	9월 2주차 Week 2 (09.06~09.10)	9월 3주차 Week 3 (09.13~09.17)	9월 4주차 Week 4 (09.20~09.24)	9월 5주차 Week 5 (09.27~10.01)	10월 1주차 Week 6 (10.04~10.08)	10월 2주차 Week 7 (10.11~10.15)	
1	서버 보안	플리토 개인정보 암호화	확인	확인	확인	확인	확인	확인	확인	
2	서버 보안	응용프로그램 개발 보안	확인	확인	확인	확인	확인	확인	확인	
3	네트워크 보안	정보저리 시스템 접근통제	확인	확인	확인	확인	확인	확인	확인	
4	네트워크 보안	시스템 접근제한 관리	확인	확인	확인	확인	확인	확인	확인	
5	근무 환경 보안	직접자 PC 보안	확인	확인	확인	확인	확인	확인	확인	
6	근무 환경 보안	소프트웨어 패치 관리	확인	확인	확인	확인	확인	확인	확인	

〈그림 93〉 보안 관리 대장 예시

4) 사업 수행 산출물 관리

- 최종 산출물 도출을 위한 과정에서 산출되는 단계별 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 52〉 산출물 대상 3차례의 저장공간

1차 저장공간	데이터셋 담당자(보안 관리자) PC
2차 저장공간	NAS (Network Attached Storage)
3차 저장공간	Cloud (Googld Drive)

○ 저장 관리 예시

이름	수정한 날짜	유형	크기
0. 기타	2022-01-20 오전 10:50	파일 폴더	
1. 한 → 영	2022-01-20 오전 11:42	파일 폴더	
2. 한 → 베	2022-01-19 오후 8:00	파일 폴더	
3. 한 → 인니	2022-01-19 오후 8:27	파일 폴더	
4. 한 → 러시아	2022-01-19 오후 5:48	파일 폴더	
5. 한 → 우즈벡	2022-01-19 오후 7:09	파일 폴더	
6. 한 → 크메르	2022-01-19 오후 9:54	파일 폴더	
7. 한 → 타갈로그	2022-01-19 오후 9:54	파일 폴더	
8. 한 → 힌디	2022-01-19 오후 8:27	파일 폴더	
9. 한 → 태국	2022-01-19 오후 9:54	파일 폴더	
10. 베 → 한	2022-01-19 오전 11:18	파일 폴더	

〈그림 94〉 저장 공간 예시 NAS(Network Attached Storage)

내 드라이브 > 2021 국립국어원 Linguists > 3. 번역 파일 > 3. 러시아어 ▾	
이름 ↑	
 1. Hand Out	
 2. Hand Back	
 3. 내부 검수	
 4. 아케이드 검수내역	
 Dinara/Russia/memo	
 러시아어 관리 시트	

〈그림 95〉 저장 공간 예시 Cloud(Google Drive)

나. 기술적 보안

1) 시스템 관리 보안

○ 시스템 접근 통제

- 서버를 포함한 정보 시스템에 대한 접근은 사전 인가된 사용자에게 국한하여 허용하였다.
- 정보 시스템 접근 인가는 보안 관리자의 대면 승인을 받고, 할당된 단말기 등을 통해서만 접근을 허용하였다.
- 관리자 권한으로 접속이 가능한 단말기 등을 운영하지 않으며, IP를 지정하여 운영하였다.
- 비인가자의 시스템 접속은 관련 법률 및 회사의 관련 규정에 따라 징계 처분 대상으로 분류하였다.

〈표 53〉 시스템 보안 관련 법률 및 회사 규정

관련 법률	소프트웨어 진흥법 (시행일 2021.12.30.)
	통신비밀보호법 (시행일 2021.03.16.)
	전기통신기본법 (시행일 2019.06.15.)
	정보통신산업 진흥법 (시행일 2020.12.10.)
	지능정보화 기본법 (시행일 2021.06.10.)
회사 관련 규정	(주)플리토 전산관리규정 (제정 2019.10.24.)

○ 서버 계정 관리

- 서버에 접속 가능한 모든 사용자에게 대해 계정으로 관리하며, 사용 목적에 따라 사용자 계정을 분류하였다.

○ 암호 관리

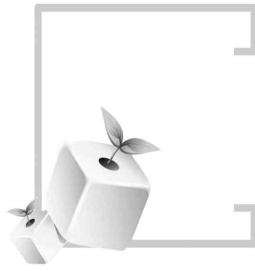
- 계정별 비밀번호는 최대 60일 이내의 변경 주기로 설정하였다.
- 비밀번호는 공유가 불가하며 비밀을 유지하도록 관리하였다.

○ 원격 접속의 금지

- 보안 관리 책임자는 외부 업체 등에 기술지원을 의뢰하는 경우를 포함하여 온라인에 의한 원격 기술 지원 작업을 허용하지 않아야 함을 준수했으며, 다만 부득이한 경우, 필요한 보안 대책을 강구 한 후 원격 기술 지원 작업을 허용할 수 있으며 해당 내용에 대해서는 반드시 기록하는 것을 준수하였다.

2) 시스템 구축 보안

- 시스템 구축에 관한 보안 정책은 (주)플리토의 기존 개발 및 운영 환경 보안 정책을 준수하였다.
- 개인 정보 및 암호화: 플리토의 기존 정책에 따라 운영
- DB 접속 권한: 플리토의 기존 정책에 동일한 정책에 따라 운영
 - 단, 개발자별 접속 권한을 구분하여 관리(DB별 접근 권한 부여 관리)



제 3 장

사업 수행 결과



제3장 사업 수행 결과

1. 말뭉치 데이터 구축 결과

1.1. 최종 구축 데이터

- 한국어-외국어 병렬 말뭉치는 언어별 1,050,000어절, 총 8,400,000만 어절 구축을 목표로 하였다.
- 번역과 검수는 총 8,400,096어절(100%)을 구축하였다.
- 감수는 총 920,424어절(109.6%)로 목표량보다 초과 달성하였다.
- 추가 제안 사항이었던 한국어-영어 병렬 말뭉치도 총 1,050,012어절(100%)을 구축하였다.

〈표 54〉 최종 구축 데이터 수량(한국어-외국어)

구분		번역		검수		감수	
언어	목표	어절	비율	어절	비율	어절	비율
베트남어	1,050,000	1,050,012	100%	1,050,012	100%	124,719	118.8%
인도네시아어	1,050,000	1,050,012	100%	1,050,012	100%	121,819	116.0%
태국어	1,050,000	1,050,012	100%	1,050,012	100%	106,327	101.3%
인도 힌디어	1,050,000	1,050,012	100%	1,050,012	100%	109,248	104.1%
캄보디아 크메르어	1,050,000	1,050,012	100%	1,050,012	100%	122,037	116.2%
필리핀 타갈로그어	1,050,000	1,050,012	100%	1,050,012	100%	106,132	101.1%
러시아어	1,050,000	1,050,012	100%	1,050,012	100%	114,057	108.6%
우즈베크어	1,050,000	1,050,012	100%	1,050,012	100%	116,085	110.6%
소계	8,400,000	8,400,096	100%	8,400,096	100%	920,424	109.6%
영어 (추가 제안)	1,050,000	1,050,012	100%	-	-	-	-

- 외국어(베트남어)-한국어 병렬 말뭉치는 52,500문장 구축을 목표로 하였다.

- 번역과 검수는 총 52,500문장(100%)을 구축하였다.
- 감수는 총 5,902문장(112.4%)로 목표량보다 초과 달성하였다.

〈표 55〉 최종 구축 데이터 수량(외국어-한국어)

구분		번역		검수		감수	
언어	목표	문장	비율	문장	비율	문장	비율
베트남어	52,500	52,500	100%	52,500	100%	5,902	112.4%

1.2. 데이터 품질 감리 결과

- 감리 품질 점검은 감수량의 10%를 대상으로 이루어졌으며, 점검 결과 전체 데이터의 적합률이 98~99% 이상인 것으로 나타나 ‘고품질 언어 데이터 구축’이라는 사업 목표를 달성한 것을 확인하였다.
- 한국어-외국어 병렬 말뭉치 감리 결과, 10,654문장(84,058어절) 중 10,626문장(99.74%)이 적합 판정을 받았다.

〈표 56〉 데이터 품질 감리 결과(한국어-외국어)

구분		문어체		구어체		전체	
언어	문장	적합	적합률	적합	적합률	적합	적합률
베트남어	1,333	295	99.66%	1,038	99.24%	1,324	99.32%
인도네시아어	1,367	278	100.00%	1,089	99.91%	1,366	99.93%
태국어	1,288	327	100.00%	961	100.00%	1,288	100.00%
인도 힌디어	1,349	331	100.00%	1,018	99.80%	1,347	99.85%
캄보디아 크메르어	1,358	309	98.41%	1,049	99.62%	1,349	99.34%
필리핀 타갈로그어	1,325	310	100.00%	1,025	100.00%	1,325	100.00%
러시아어	1,331	304	98.70%	1,027	99.90%	1,326	99.62%
우즈베크어	1,303	292	99.32%	1,011	100.00%	1,301	99.85%
합계	10,654	2,446	99.51%	8,218	99.81%	10,626	99.74%

- 외국어(베트남어)-한국어 병렬 말뭉치 감리 결과, 525문장 중 518문장(98.67%)이 적합 판정을 받았다.

〈표 57〉 데이터 품질 감리 결과(외국어-한국어)

구분		문어체		구어체		전체	
언어	문장	적합	적합률	적합	적합률	적합	적합률
베트남어	525	206	98.10%	312	99.05%	518	98.67%

2. 대외 활동

2.1. 워크숍

- ‘21 한-외 병렬 말뭉치 구축 사업’의 일환으로 ‘인공지능(AI) 병렬 말뭉치 구축의 실제- 신남방·신북방 언어를 중심으로’라는 주제로 사업 기간 중 워크숍을 개최하였다.
- 워크숍 오전 프로그램
 - 오전 프로그램에서는 인공지능 병렬 말뭉치 분야의 학계 교수 및 산업계 전문가인 이찬규 교수(중앙대), 송상헌 교수(고려대), 이정수 대표(㈜플리토)가 “인문학에서의 빅데이터 활용과 전망”, “해외 병렬 말뭉치 구축 현황과 활용 사례”, “산업계 말뭉치 구축의 특성과 난제”를 주제로 특강을 하였다.
 - 이를 통해 병렬 말뭉치 구축에 관한 지식 및 정보를 공유하고, 병렬 말뭉치의 활용 방안과 관련 산업에 미칠 영향력을 전망하는 자리를 가졌다.

〈표 58〉 워크숍 세부 순서

시간	프로그램
09:00~09:10	[인사말] 이정희 교수(경희대학교, 연구 책임자)
09:10~10:10	[기조 강연] 인문학에서의 빅데이터 활용과 전망 이찬규 교수(중앙대학교)
10:10~11:10	[주제 발표] 해외 병렬 말뭉치 구축 현황과 활용 사례 송상헌 교수(고려대학교)
11:10~11:30	휴 식
11:30~12:30	[초청 특강] 산업계 말뭉치 구축의 특성과 난제 이정수 대표(㈜플리토)

〈표 59〉 워크숍 특강 내용

강연자	제목	내용
이찬규 교수 (중앙대학교)	인문학에서의 빅데이터 활용과 전망	인공지능 시대의 가장 중요한 자원은 데이터이다. 인공지능 기계학습 과정에 투입되는 데이터에 따라 산출 결과도 달라지기 때문이다. 인공지능 학습에 필수 요소인 데이터의 유형과 각 유형별 속성을 알아보고, 비정형 데이터에 속하는 언어 데이터를 인공지능 학습에 활용할 수 있는 방안에 대해 살펴보고 국어국문학 분야에서의 데이터 처리와 프로그램 활용, 인공지능 학습에 관한 기초적인 접근을 통해 향후 연구 방향을 모색하고자 하였다.
송상헌 교수 (고려대학교)	해외 병렬 말뭉치 구축 현황과 활용 사례	특수 목적 말뭉치의 하나로 분류되는 병렬 말뭉치는 자연어 처리와 언어 연구의 측면에서 공히 중요성을 가진다. 공학적인 관점에서 병렬 말뭉치는 기계 번역 시스템을 구현하기 위한 학습 및 평가 데이터로 기능을 한다. 언어학적인 관점에서 병렬 말뭉치는 복수의 언어가 어떠한 대응 관계를 가지는가를 유형론적으로 살필 수 있는 실물적 근거가 된다. 이러한 이유에서 영어를 비롯한 여러 언어권에서는 병렬 말뭉치를 오랫동안 다양한 형태로 구축하여 왔다. 본 특강에서는 그 열개를 살펴보고, 한국어 중심의 다국어 병렬 말뭉치 구성에 대한 시사점을 얻고자 하였다.
이정수 대표 (㈜플리토)	산업계 말뭉치 구축의 특성과 난제	메타버스등 공간의 제약을 넘는 플랫폼이 빠르게 성장하면서 이와 더불어 진정한 글로벌 환경을 조성하기 위해서는 언어의 장벽이 반드시 해결되어야 한다. 이로 인하여 다양한 기업들이 언어 인공지능 개발에 뛰어들고 있으며 이러한 인공지능 학습을 위한 언어 데이터 사업도 빠르게 발전하고 있다. 본 특강에서는 그 발전 현황에 대해 설명하고자 하였다.



인사말(이정희 교수)



기조 강연(이찬규 교수)



주제 발표(송상헌 교수)



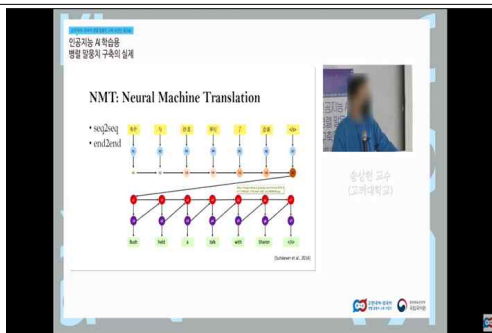
초청 특강(이정수 대표)



단체 사진(오전 특강)



온라인 생중계 화면 ①



온라인 생중계 화면 ②



온라인 생중계 화면 ③

〈그림 96〉 워크숍 특강 현장

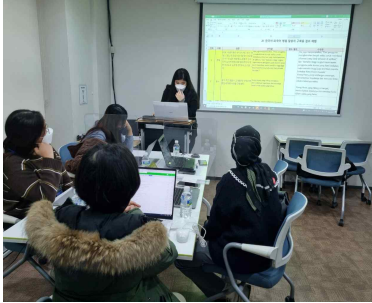
○ 워크숍 오후 프로그램

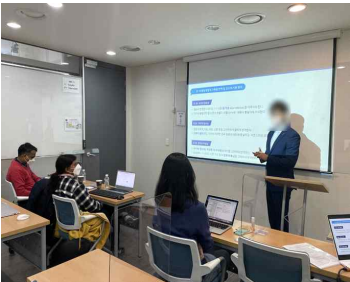
- 오후 프로그램에서는 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어, 8개 언어별로 검수팀장 또는 검수원이 참여하여 번역 및 검수 실습 체험 시간을 마련하였다.
- 이를 통해 참가자들이 번역 및 번역 검수를 직접 해 보고 번역 검수의 어려움, 품질 개선 등에 관해 의견을 서로 교환하고 노하우를 공유하였다.

〈표 60〉 워크숍 오후 프로그램 세부 순서

시간	내용
14:00~14:05	1. 인사 및 간단한 자기소개
14:05~14:15	2. 사업 소개 및 전체 진행 순서 설명
14:15~14:25	3. 번역 및 검수 기본 원칙 설명
14:25~15:05	4. 번역 체험 ① 참여자 대상 한국어 원문 배포 ② 참여자 번역 실시 ③ 번역문 취합 및 화면 공유 ④ 참여자 간 번역 결과에 대한 논의 및 피드백
15:05~15:45	5. 검수 체험 ① 검수용 플랫폼 사용 시현 ② 검수 방법 설명 ③ 참여자 대상 한국어 원문-외국어 번역문 배포 ④ 참여자 검수 실시 ⑤ 검수문 취합 및 화면 공유 ⑥ 참여자 간 번역 결과에 대한 논의 및 피드백
15:45~15:50	6. 최종 산출물 샘플 확인
15:50~15:55	7. 만족도 조사
15:55~16:00	8. 전체 모임 장소 이동
16:00~16:30	9. 수료증 수여 및 기념품 증정

〈표 61〉 언어별 워크숍 내용 및 활동 사진

언어	워크숍 내용	사진
베트남어	오후 워크숍 세부 일정표에 따라 사업 소개 (영상), 번역 및 검수 기본 원칙 설명에 이어 번역 및 검수 실습, 결과에 대한 논의가 각각 40분씩 진행되었다. 시간이 부족할 것으로 판단되어 문어, 구어 각각 5문장을 실습하였다. 출판용 번역과 인공지능 학습용 데이터 번역의 차이에 대해 자세히 알 수 있었고, 앞으로 언어데이터를 활용한 연구에 대해서도 심도 있는 논의를 진행하였다. 그러나 참가자들은 프로그램을 사용하여 직접 말뭉치를 구축하고 분석하는 워크숍이라고 생각하였는데 번역과 검수만 실습하여 아쉽다는 의견이 있었다.	
인도네시아어	오후 워크숍 세부 일정에 따라 사업 소개 영상 시청, 번역 및 검수 기본 원칙 설명에 이어 번역 및 검수 실습, 결과에 대한 교육이 각각 40분씩 진행되었다. 또한 마지막으로 실제 검수 플랫폼에서 검수하는 모습을 시연하였다. 소규모 그룹 워크숍으로 참여자들과 검수원, 전임 연구원 간의 활발한 의견 교환 및 더 적절한 번역을 위한 토론 활동이 매우 긍정적으로 평가되었다.	
태국어	워크숍 일정은 기본 일정대로 진행하였다. 즉, 인사와 자기소개 후 사업 소개 영상을 시청하고 검수팀장의 지침 설명을 들었다. 번역과 검수 실습 역시 예정대로 각각 40분씩 진행하였으나 예상보다 소요 시간이 길어서 충분한 비교 논의는 이뤄지지 못했다. 다만, 참여자 모두 흥미를 가지고 적극적으로 참여하는 태도를 보였다.	

인도 힌디어	워크숍 세부 일정표에 따라 사업 소개(영상), 번역 및 검수 기본 원칙 설명에 이어 번역 및 검수 실습, 결과에 대한 논의를 각 40분씩 진행하였다. 다만, 시간이 부족하여 계획한 문장 수보다 적은 문장 수로 번역 및 검수 실습을 하였다. 번역 및 검수에서 참가자들 모두 적극적으로 의견을 제시하였고 번역을 함에 있어 문체에 따라 어떻게 번역할 것인가에 대해 깊이 있게 논의하였다. 또 검수의 과정을 통해 검수에서 중요 하게 생각해야 하는 것에 대해서도 활 발히 논의하였다. 참가자들 모두 번역 및 검수를 직접 해 봄으로써 번역 시 주의해야 할 점, 검수 시 필요한 원칙 들에 대해 생각해 보는 의미 있는 자리 라고 평가했다. 또한 향후 번역 검수원으로 활동할 의향을 보였다.	
캄보디아 크메르어	오후 워크숍 세부 일정표에 따라 사업 소개(영상), 번역 및 검수 기본 원칙 설명에 이어 번역 및 검수 실습, 결과에 대한 논의가 각각 40분씩 진행되었다. 다만, 시간이 부족한 관계로 취합한 참가자의 번역 및 검수 결과물을 충분히 논의하지 못한 것이 아쉬움으로 남았다. 그럼에도 참가자의 적극적인 의견 제시와 개선된 결과물을 도출해내려는 태도는 매우 긍정적으로 평가되었다. 또한 향후 번역 검수원으로 활동할 의향을 보였다.	
필리핀 타갈로그어	검수 팀장이 한국어 모어 화자로 필리핀 현지 에 계셔서 필리핀 타갈로그어 모어 화자인 라넬 론 검수원이 배경민 팀장과 함께 ZOOM에서 검수 실습자들에게 검수 실습 결과에 대한 피드백을 진행해 주었다. 검수원들이 다소 늦게 도착해 일부만 피드백을 받고 나머지 검수에 대한 피드백은 워크숍이 끝나고 진행하였다. 워크숍 참가자인 김수미 실습자는 검수 교육 진행 후 검수원으로 충원 가능하다는 피	

	드백을 받았으며, 님파씨 참가자는 교육을 한 차례 더 진행 후 테스트를 거쳐 총원 여부를 결정하자고 의견을 주었다. 참가자들은 검수 실습을 하는 과정에서 플랫폼을 이용해 새로운 것을 배우고자 하는 열의가 컸으며, 실습 과정에 대해 만족도가 매우 높았다.	
러시아어	사업 소개 영상 시청과 번역 및 검수 지침 설명에 이어 러시아어 번역 및 검수 체험을 진행하였다. 3명의 모든 참가자들이 오후 워크숍의 전 과정에 적극적으로 참여하였으며, 더불어 평소 번역을 하면서 가지고 있었던 의문과 고민들에 대해서도 논의하는 시간도 가졌다. 다만, 예상보다 시간이 부족하여 사전에 준비한 번역 및 검수 문장 중 일부만 다루어서 아쉬움이 좀 남았다.	
우즈베크어	한-외 병렬 말뭉치 구축 사업에 대한 간단한 소개를 시작으로 오후 워크숍을 진행하였다. 먼저 번역 및 검수의 기본 원칙에 대한 설명으로 1:1번역을 원칙으로 의미의 정확성과 맥락의 실제성, 그리고 문체의 적절성에 대해 예시를 통해 설명하였다. 다음으로 검수 시 수정 범위에 대해 어휘, 문법, 속담 등을 예시로 설명하였다. 예를 들어 다음과 같은 속담인 경우 우즈베키스탄의 문화와 언어에 대응하는 번역이 자연스럽다고 설명하였다. *금강산도 식후경 - Avval taom ba'daz kalom. 마지막으로 검수플랫폼을 통해 실제적인 검수를 간접적으로 체험해봄으로써 워크숍 참가자들의 검수에 대한 관심과 동기부여를 할 수 있는 계기가 되었다. 2시간30분의 시간이 부족하였으나 번역과 검수에 대한 실질적인 이해를 높이는 시간이었다.	

○ 성과

- 오전 프로그램 참가 신청자 171명, 오후 프로그램 참석자 36명으로 총 207명이 참여한 사업단 워크숍을 통해 한-외 병렬 말뭉치 구축에 대한 높은 관심을 확인할 수 있었고, 사업 홍보 및 지식 공유의 기회가 되었다.
- 향후 사업에서 신규 번역 검수원으로 활용할 수 있는 인적 네트워크를 구축하는 계기가 마련되었다.

2.2. 학술대회 발표

- 2022년 5월 14일에 실시된 국제한국어교육학회(IAKLE) 제53차 국내(봄) 학술대회에서 ‘한국어-외국어 병렬 말뭉치 구축’와 관련된 연구 논문 3편을 발표하였다.
- 연구 논문 3편은 연구 책임자와 전임 연구원들이 공동으로 연구를 진행하였으며 주요 내용은 아래와 같다.
 - <한국어-외국어 대조 연구 현황 분석 - 신남방·신북방 지역의 8개 언어를 중심으로 ->의 논문에서는 한국어와 2021년 한국어-외국어 병렬 말뭉치 구축 사업의 대상 언어인 8개 언어 간 대조 연구 동향을 분석하여 향후 연구의 방향성을 제시하였다.
 - <한국 문화 관련 용어의 외국어 번역 및 표기 양상 연구 - 제주관광공사 홈페이지의 베트남어, 인도네시아어, 태국어 버전을 중심으로 ->의 논문에서는 증가하는 아세안 국적의 관광객 및 입국자들을 고려하여 공공 기관 사이트의 아세안 언어(베트남, 태국, 인도네시아어)의 문화 용어 번역 양상을 조사, 분석함으로써 향후 다국어 번역 지침을 위한 기초 자료를 제공하고 관련 지침 마련의 필요성을 제언하였다.
 - <어순 오류 유형 분석 - 힌디어, 크메르어, 타갈로그어 번역 데이터를 중심으로 ->의 논문에서는 이번 사업의 대상 언어 중 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어와 같이 언어 유형론적으로 서로 다른 특성을 지닌 3개 언어의 기본 어순을 대조적인 관점에서 비교 분석하여 외국어 학습 자료로 유용하게 활용될 수 있는 기초 연구를 하였다.
- 학술대회 연구 논문 발표를 통해 병렬 말뭉치 구축이 언어 연구 및 교육을 위한 기초 자료, 외국어 번역 연구 및 교육 자료로서 활용 가치가 높음을 확인하였고 병렬 말뭉치 구축에 대한 학계의 관심을 끌어냈으며 관련 연구의 활성화에 기여하였다.

<제2부> 분과별 발표 (2000/승설대 권리권 306호, 309호)			
제1분과: 한국어 말하기 수업 사례 사회: 최은지(원광대학교)	제2분과: 말하기 과제와 활동 사회: 이정현(한국항공대학교)	제3분과: 한국어-외국어 병렬 학습의 구축 사회: 박나리(한국교통대학교)	제4분과: 워크숍 사례 발표 사회: 강남숙(경인교육대학교)
외국인 학생에 대상 비례한 한국어 토론 수업 사례 연구 발표자: 김지혜(서울과학기술대학교) 토론: 이효정(국민대학교)	그림책 번역 활동의 한국어 말하기 의미 향상 효과 분석 연구 중급 중국어권 학습자를 중심으로 발표자: 최순옥(경희대학교) 토론: 박민옥(대구가톨릭대학교)	한국어-외국어 대조 연구 현황 분석: 신남방-신북방 지역의 8개 언어를 중심으로 발표자: 김연희, 정성호, 한채민 문진숙 이정희(2021 한국어-외국어 병렬 학습지 구축 사업단) 토론: 이경(호남대학교)	한국어 하이플렉스 수업에서의 온라인 학습 도구 활용 사례: 일본 C대학의 초급 한국어 수업을 중심으로 발표자: 위햇님(일본 주교대학교)

목 차

- 1 서론
- 2 연구 대상 및 방법
 - 2.1. 연구 대상
 - 2.2. 연구 방법
- 3 한국어-신남방·신북방 언어 대조 연구 동향 분석
 - 3.1. 연도별 분석
 - 3.2. 연구 목적 및 출판 유형별 분석
 - 3.3. 연구 주제별 분석
- 4 결론

〈그림 97〉 학술대회 논문 발표 사진

2.3. 대외 홍보

- ‘인공지능(AI) 병렬 말뭉치 구축의 실제- 신남방·신북방 언어를 중심으로’라는 주제로 열린 사업단 워크숍과 관련한 언론 홍보를 통해 ‘21 한-외 병렬 말뭉치 구축’ 사업에 관한 내용을 외부에 널리 알리고 공유하였다.
- 2022년 2월 23일 내일신문에 ‘인공지능 병렬 말뭉치 구축 워크숍 열려’라는 제목으로 사업단 워크숍과 관련된 기사가 게재되었다. 워크숍이 열리는 목적 및 워크숍의 주요 강연자와 프로그램이 소개되었다. 더불어 한국어-외국어 병렬 말뭉치 사업단에 대한 소개도 함께 이루어졌다.



〈그림 98〉 내일신문 기사

- 2022년 2월 24일 인공지능신문에 ‘국립국어원, 인공지능 학습용 신남방·신북방 말뭉치 구축에 나서’ 라는 제목으로 기사가 게재되었다. 국립국어원의 인공지능 학습용 신남방·신북방 말뭉치 구축 사업의 의의를 중점적으로 다루었으며 사업의 일환인 워크숍을 소개하였다.



〈그림 99〉 인공지능신문 기사

- 2022년 3월 30일 한-아세안센터 뉴스레터에 ‘언어문화사업 활성화 및 언어문화 교류를 위한 21 한국어-외국어 병렬 말뭉치 구축 사업단 워크숍’이라는 제목으로

로 기사가 게재되었다. 이 기사는 한-아세안센터 블로그 취재 기자가 워크숍을 직접 방문하여 쓴 워크숍 후속 기사로 워크숍의 의의를 비롯해 강연 및 프로그램에 대한 생생하고도 자세한 내용을 담고 있다.



〈그림 100〉 한-아세안센터 뉴스레터 기사 일부 발췌

3. 활용 방안 및 기대 효과

3.1. 병렬 말뭉치의 활용 방안

- 경제·사회 등 다양한 분야에서 활용 가능한 데이터 마련
 - 문어 및 구어 모두 포괄하며 특정 분야에 편중되지 않는 다양한 주제의 데이터를 구축하여 다양한 분야에서 저작권의 문제없이 활용 가능
 - 대규모 병렬 텍스트로 구성된 훈련 데이터 기반으로 자동 번역 시스템 마련, 이를 통해 COVID-19 시대를 극복해 나갈 수 있는 다국어 화상회의, 번역 자동 평가 시스템 등에서 활용 가능
- 병렬 말뭉치 활용을 위한 다국어 용례 검색기 설계 및 개발
 - 용례 검색기는 병렬 말뭉치를 활용하기 위한 가장 기본적인 도구
 - 다국어 용례 검색 도구 개발을 통해 외국어 교육 분야, 특히 여러 언어의 대응 관계를 면밀히 관찰해야 하는 언어 대조 분석에 기여
- 언어 연구 및 교육을 위한 기초 자료로 활용
 - 언어 데이터 기반의 인공지능 학습 기능 향상을 통해 한국어 학습용 챗봇, 모바일 학습 응용 프로그램 등 온라인 학습자료 개발 및 발전에 기여
 - 한국어 교재 개발(특히 현지 맞춤 교재)을 위한 언어 자료
 - 해외 한국어교육 현장에서 수업·자가 학습용 보조재 제작을 위한 자료
 - 한국어와 외국어의 대조 연구를 위한 자료
 - 한국어 학습자의 중간 언어 및 오류분석을 위한 분석 자료
 - 이민자 및 근로자 등의 사회통합 교육을 위한 기초 자료
 - 이중 언어 사전 편찬에서 용례 및 언어, 대역어 등의 선정을 위한 자료
- 외국어 번역 연구 및 교육
 - 통번역 전문가 양성을 위한 실무 교육 및 훈련에 언어 간 대응 관계 분석 자료
 - 다양한 실무 현장에서 정확한 번역을 위한 참고 자료
 - 번역학에서 번역 텍스트의 특성 분석을 위한 자료
 - 통계 기반 번역이나 신경망 자동 번역 등 기계번역 분야 연구 자료

3.2. 사업의 기대 효과

- 첫째, 대규모·고품질의 병렬 말뭉치 구축을 통한 언어 데이터 산업의 기초 마련
 - 국가 기관 주도의 병렬 말뭉치 구축을 통해 약 900만 어절의 고품질 병렬 말뭉치를 구축하였다. 이렇게 구축된 병렬 말뭉치는 관련 산업계뿐만 아니라 연구계, 교육계 등 다양한 분야에서 활용될 수 있다.
 - 특히 그간 미진했던 아세안-인도 6개 언어 및 유라시아 2개 언어 관련 연구의 기초 자료로서 활용될 수 있을 것이다.
- 둘째, 다국어 언어 처리 및 인공지능 기반 번역 기술 향상
 - 병렬 말뭉치를 활용하여 학습이 비교적 덜 진행된 아세안-인도 6개 언어 및 유라시아 2개 언어에 대한 학습이 이루어질 수 있다.
 - 다양한 유형의 언어 학습을 통해 기계번역의 기술과 품질이 향상될 수 있다.
- 셋째, 언어문화 산업 확산
 - 학습 및 번역 소프트웨어 개발 등을 통한 언어 학습 및 통·번역 관련 분야를 활성화할 수 있다.
- 넷째, 국가 정책 및 관련 업계의 수요 충족
 - 아세안-인도, 유라시아 지역 관련 정책에 따라 증대된 산업계·연구계의 수요에 효과적으로 대응할 수 있으며, 이들 지역과의 교류를 활성화함으로써 상호 이해 증진 및 미래지향적 상생의 경제 협력 기반을 조성할 수 있을 것이다.

3.3. 제언

○ 한국어-외국어 병렬 말뭉치의 데이터 규모 확대

- 이번 사업에서 그동안 병렬 말뭉치 구축 대상 언어로 영어나 중국어, 일본어 등에 비해 큰 주목을 받지 못했던 8개 언어를 대상으로 대규모 데이터를 구축할 수 있었다. 그러나 인공지능 기반 기계번역 시스템의 성능 향상에 유의미한 결과를 얻기 위해서는 언어당 최소 200~300만 어절의 번역쌍이 요구된다. 이번 사업을 통해 산출한 데이터의 수량이 앞서 제시한 필요량에 미치지 못하므로 후속 사업에서 구축 규모를 더 확대하여 데이터를 지속적으로 축적하는 것이 필요하다.
- 이번 사업은 아세안-인도 6개 언어와 유라시아 2개 언어를 대상으로 하였다. 하지만 이 외에도 말레이시아어나 미얀마어 등 국제 교류 협력의 측면에서 중요도가 높은 언어들이 남아 있다. 8개 언어의 데이터 구축량이 일정 수준에 도달한 후 현재보다 구축 언어의 외연을 넓힌다면 병렬 말뭉치의 활용성을 한층 더 높일 수 있을 것이다.

○ 번역 작업에서 클라우드소싱 방식의 적용

- 번역은 현지의 번역 업체를 통해 수행되었는데 업체 선정을 비롯하여 높은 수준의 번역 전문가를 확보하기가 쉽지 않았으며 다수의 업체를 관리하고 의사소통하는 데에도 어려움이 다소 있었다. 이러한 문제를 극복할 대안으로는 웹 기반의 플랫폼을 통한 클라우드소싱(crowd sourcing)의 방식이 있다. 이번 사업에서는 한국어 원문 데이터의 수집 절차에만 도입하였는데 번역 작업에도 적용이 가능하다. 번역 능력에 대해 철저한 검증 절차를 거친 플랫폼 사용자들이 기기의 제약 없이 원하는 시간과 장소에서 번역을 한다면 단기간에 번역 데이터를 효율적으로 구축하는 것이 가능해진다. 그리고 플랫폼 사용자 등록 시 저작권 양도 및 개인 정보 이용 동의 등도 자동으로 처리할 수 있어 데이터 저작권 확보 및 개인 정보 보호의 측면에서도 용이하다. 그러므로 후속 사업에서는 고품질의 데이터를 효율적으로 구축하기 위해 클라우드소싱 방식을 적극적으로 활용하는 것이 필요하다.

○ 국가 주도로 구축된 언어 데이터의 원문 데이터로서 활용

- 고품질의 병렬 말뭉치 구축하기 위해서는 도착어 번역만큼이나 한국어 원문 데이터의 수집도 중요하다. 그러나 품질이 좋으면서도 저작권 문제까지 해결된 대규모의 원문 데이터를 확보하는 것은 쉬운 일이 아니다. 이러한 문제의 해결 방법은 단일 언어 말뭉치나 고급 이상 등급의 한국어 교재와 같이 국가 기관의 주도로 구축된 단일 언어 데이터를 병렬 말뭉치 구축에서 활용하는 것이다. 이러한 방식은 비용 절감은 물론 품질이 보증된 데이터를 확보할 수 있다는 장점이 있다. 따라서 앞으로 국가 주도

로 단일 언어 데이터를 구축할 때 병렬 말뭉치의 활용 가능성도 반드시 고려하여 저작권 확보 등에 노력을 기울여야 할 것이다.

○ 고품질의 개체명 병렬 말뭉치 구축

- 최근 몇 년 동안 전 세계적으로 기승을 부렸던 COVID-19가 종식되는 시기를 맞이한 후에는 해외 관광 수요가 폭발적으로 증가할 것이라는 예측이 많다. 이러한 미래 관광 수요의 증가와 현재 한류의 확산세를 고려한다면 한국을 방문할 관광객의 수가 상당할 것이라는 사실을 쉽게 예상할 수 있다. 이를 대비하기 위해 다양한 관광 인프라의 조성이 필요하며 이 중에서 다양한 국가의 방한 관광객의 언어 해소를 위해서는 번역 데이터가 필수적이다. 관광객의 입장에서 음식명이나 지명 등에 대한 번역이 중요한데 원래의 의미를 전달되도록 번역하는 것이 쉽지 않고 번역 방식에 일관성이 없다면 오히려 수요자들에게 혼동을 줄 수 있다. 그러므로 통일된 지침에 따라 전문 번역가의 작업을 통해 고품질의 개체명 병렬 말뭉치 구축이 필요하다. 특히 영어나 일본어, 중국어 외에 특수외국어 등을 중심으로 병렬 말뭉치를 구축한다면 향후 국제 교류 및 산업 경제의 측면에서 다양한 활용 가치를 창출할 수 있을 것이다.

참고자료

1. 논문 및 저서

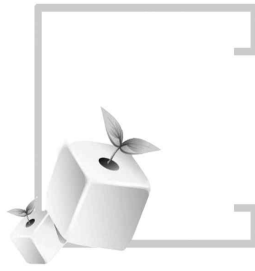
- 대외경제정책연구원(2020), 특수외국어진흥계획 마련을 위한 실태조사, 교육부.
- 문현석 외(2021), 병렬 말뭉치 필터링을 적용한 Filter-mBART 기반 기계번역 연구, 『한국융합학회논문지』 12-5, 한국융합학회, 1-7쪽.
- 민경모(2010), 병렬말뭉치의 개념 및 구조에 관한 몇 문제, 『언어사실과 관점』 25, 연세대학교 언어정보연구원, 41-70쪽.
- 민경모(2019), 병렬 말뭉치와 한국어 교육 연구, 『언어와 문화』 15-1, 한국언어문화교육학회, 97-117쪽.
- 민경모(2020), 다국어 병렬 말뭉치의 구축과 한국어교육 연구에의 활용, 『한국학논집』 78, 187-220쪽.
- 박찬준 외(2020), 병렬 말뭉치의 양과 질이 인공지능기계번역에 미치는 효과, 『제32회 한글 및 한국어 정보처리 학술대회 논문집』, 언어공학연구회, 363-368쪽.
- 신자영(2010), 병렬 코퍼스 및 학습자 코퍼스를 이용한 중간언어 연구 방법론, 『언어사실과 관점』 25, 연세대학교 언어정보연구원, 71-87쪽.
- 안인경 외(2006), 독-한 병렬 코퍼스의 구축과 활용에 관한 연구, 한국연구재단.
- 유현경 · 황은하(2010), 병렬말뭉치 구축과 응용, 『언어사실과 관점』 25, 연세대학교 언어정보연구원, 5-40쪽.

2. 관련 법령

- 공공 용어의 외국어 번역 및 표기 지침(문화체육관광부 훈령 제448호, 2021.07.22. 시행)
- 국어의 로마자 표기법(문화체육관광부 고시 제2014-42호, 2014.12.05. 시행)

3. 웹사이트

- 내일신문 뉴스 기사 http://www.naeil.com/news_view/?id_art=415166
- 인공지능신문 기사 <http://www.aitimes.kr/news/articleView.html?idxno=24360>
- 한-아세안센터 뉴스레터 <https://blog.naver.com/akcsns/222686727087>



부록



[부록 1]

한국어 원문의 오류 검토 예시

No.	일시	원문	오류 내용
1	2021.12.19	그러면서 나라싱가 라오 위언장은 안드라프라데시 주민들을 위해 바스카파트남 철강 보호 차원에서 토지를 사유화한다고 촉구했다.	위언장 → 위원장
2	2021.12.20	베트남 세무 당국의 제재와 세금 비용 증가 등으로 트라파코는 상반기에만 9억 92,000만 동(약 52,000만 원)을 지출했다.	숫자 표기 오류
3	2021.12.21	계약 예정 난도 기준으로 살펴보면 2017년 4,689억 원을 투자해 260량을 구입하고, 지난해 5,373억 원을 들여 336량을 구입할 예정이었다.	난도 → 년도
4	2021.12.21	쿠르드에 발전서와 변전소 등을 건설해주는 대신 5개의 광권을 획득하는 방식이다.	발전서 → 발전소
5	2021.12.22	1일(현지시간) 미래에셋대우 베트남법인은 자본금 규모를 기존 4조 3000억 동(약 2,200억 원)에서 5455억 5000만 동(약 2,840억 원)으로 확대한다고 밝혔다.	5455억 5000만 동 → 4조 5455억 5000만 동
6	2021.12.23	인도네시아 증권업의 성장 잠재력이 크다고 판단하고 브로커리지와 채권 중개뿐만 아니라 기업상장(IPO)을 주관하는 등 업무영역을 확대하고 있다.	판다하고 → 판단하고
7	2021.12.25	데스트 결과 신라면 보다 더 매운맛도 일본 시장에서 충분히 승산이 있다고 판단, 특별한 매운맛으로 일본 내 불고 있는 '신라면' 열풍을 이어간다는 전략이다.	데스트 → 테스트
8	2021.12.26	동풍위에다기아는 신종 코로나바이러스 유행에 따라 점염병 감염 위험이 높아지자 춘절 연휴를 연장해 딜러점 영업을 2월 10일에 재개할 예정이다.	동풍위에다기아 → 동평위에다기아
9	2021.12.27	다만 계약 일정은 즈베즈다 조선소가 건설하는 드라 이도크(육상 도크) 완공 시기에 따라 변경될 수 있다.	드라 이도크 → 드라이 도크
10	2021.12.27	넷마블 EMEA는 오는 21일(현지 시간) 19시부터 23일 19시까지 넷마블 EMEA 이스타불 사무소에서 '하이퍼 잼 2020'을 개최한다.	이스타불 → 이스탄불
11	2021.12.27	이 공로로 이 부사장은 지난 2024년 '자랑스런 삼성인 상'을 받았으며 작년 말 퇴임했다.	연도 오류
12	2021.12.30	홈플러스가 한국을 방문한 캐나다 서스캐처원주 비즈니스 대표단을 다양한 사업 논의를 진행, 향후 행보에 관심이 집중되고 있다.	호응 오류
13	2021.12.3	신라면세점은 창이국제공항 5주년을 기념해 SNS상에서 '신라 슈퍼 스타' 캠페인을 진행한	창이국제공항 5주년 → 창이국제공항점 5주

No.	일시	원문	오류 내용
	0	다.	년
14	2021.12.3 0	금융권 관계자는 "라고액 연봉을 받는 은행원에 대한 곱지 않은 시선과 올해 공무원 임금 인상을 등을 고려하면 올해 임금인상률은 적절한 수준으로 보인다"라고 말했다.	라고액 → 고액
15	2021.12.3 0	작년 매출과 순이익은 각각 84억 3800만 링깃(약 2조 4500억 원), 링깃(약 1285억 원)으로 1년 전보다 9%, 44% 감소했다.	수치 누락
16	2021.12.3 1	블록체인 지갑 선점 경쟁이 치열한 가운데 카카오톡 기반의 암호화폐 지갑 '클립(Klip)'가 본격 시장 경쟁에 뛰어든다.	'클립(Klip)'이
17	2022.01.0 1	최근 일본에서 자연주의 화장품에 대한 소비지 니즈가 높아지면서 비욘드 제품이 주목받자 신제품 출시로 시장 점유율을 높인다는 계획이다.	소비지 → 소비자
18	2022.01.0 3	신도시 내 빌라, 타운 하우스, 상점 등도 친환경 콘트리트, 저탄소 건설 자재를 사용해 친환경적으로 건설된다.	콘트리트 → 콘크리트
19	2022.01.0 3	내년 6월에는 파라마운트 픽처스 실내 테마마크 등 1조 3000억 원 규모의 1-B단계 시설이 2025년 오픈 목표로 착공한다.	테마마크 → 테마파크
20	2022.01.0 3	힌편, CJ CGV는 2011년 베트남에 진출한 이래 현재 78개 극장, 457개 스크린을 보유하고 있다.	힌편 → 한편
21	2022.01.0 4	현대제철이 일본산 고철 입찰 형식을 변경한 건 일본 지역에서 고철 가격을 더 잘 반영하고, 물류 권리 비용을 최소화하기 위한 시도로 풀이된다.	권리 → 관리
22	2022.01.0 4	18일 중국 시장조사기관 가오궁 산업연구원(GGII) 따르면 지난 1월 중국 전기차 배터리 설치용량은 2.32GWh로 전년 동월 대비 53.5% 감소했다.	산업연구원(GGII) 따르면 → 산업연구원(GGII)에 따르면
23	2022.01.0 4	현대모비스인디아(MIN)는 기업의 사회적 책임을 다하기 위해 인도 코로나19 극복을 위해 비정부기구(LNG)에 구호 성금과 구호 물품을 전달했다.	LNG → NGO
24	2022.01.0 4	에이이닥과 현대BS&C가 현대페이 지분을 각각 30%, 20%씩 갖고 있다는 점을 고려하면 사실상 현대페이도 정 사장 개인회사나 마찬가지인 셈이다.	에이이닥 → 에이닥
25	2022.01.0 4	대우건설은 올해 초 국내 금융기관들과 투자 개발 합의를 체결하고 시타레이크시티 내에 주상복합 빌딩을 개발하기로 하였다.	시타레이크시티 → 스타레이크시티
26	2022.01.0 5	두산중공업은 설계, 기자재 제작 및 설치, 시운전(EPC) 공정 등 일괄적으로 맡는 일종의 턴키 방식을 사업을 수행한다.	호응 오류

No.	일시	원문	오류 내용
27	2022.01.05	루퍼트 머독은 회장은 이후에도 38살 연하의 세 번째 부인인 중국계 배구 선수 출신 부웬디 등과 이혼했다.	부웬디 → 웬디
28	2022.01.07	여러 사람과 360도 영상을 주고받는 커뮤니케이션 서비스 '리얼 360', 게임 영상을 최대 5개 화면으로 분할해 초고화질로 중계하는 ' e스포라이브 ' 등 다양한 5G 서비스로 포형을 받았다.	e스포라이브 → e스포츠라이브
29	2022.01.08	현재 후보물질을 발굴한 상태며 향후 LG화학이 전암상 부터 상업화까지 주도적으로 진행할 계획이다.	전암상 → 전임상
30	2022.01.11	레전드 호텔 사이공이 웨딩 박람회를 직접 주관해 개최한 이번이 처음이다.	개최한 → 개최한 것은
31	2022.01.14	이번 가격 인상은 지난해 11월 열린 LA오토쇼에서 연식변경 에 따른 업그레이드 상향 적용에 따른 것으로 보인다.	연식변경 → 연식 변경 상향 → 상향
32	2022.01.14	이외에도 헤드 마운드 디스플레이 광학 시스템, 퀵텀닷 디스플레이 등 디스플레이 관련 특허들도 눈길을 끌었다.	마운드 → 마운트 퀵텀닷 → 쿼텀닷
33	2022.01.14	캠콤 은 21일(현지시간) 열린 '레지던트 이블 쇼케이스'를 통해 2월 2일부터 15일까지 더 디비전 2에 레지던트 이블 아이템이 등장할 것이라고 밝혔다.	캠콤 → 캡콤
34	2022.01.17	앞서 인도 정부는 지난해 초부터 옹광로(고로) 공장을 설립하기 위해 포스코와 현대제철에 합작 일관제철소 설립을 제안했다.	일관제철소 → 일관제철소
35	2022.01.17	한편 이랜드는 중국에서 한때 40여 개 브랜드, 8000여 매장을 중국에서 운영했지만 현재는 20개의 브랜드, 4000여 매장을 갖고 있다.	'중국에서' 반복
36	2022.01.19	또 매년 늘어나는 부채 부담을 줄이기 위해 감자 를 실행하며 강도 높은 경영정상화 조치에 적극 나서왔다.	감자 → 감사
37	2022.01.20	현대백화점 관계자는 " 파로남 과의 이번 협업으로 동영상 기반 이커머스 사업 확대에 역량을 집중하겠다"라고 말했다.	파로남 → 파로님(Paronym)
38	2022.01.22	일일 로그인 보너스를 비롯해 보스 퍼스트 킬 이벤트, 명예의 전당 랭킹 이벤 , 해피 아워 보상 등이 제공된다.	이벤 → 이벤트
39	2022.01.26	세계의 테마마크 기준에서 방문객이 가장 많은 곳은 방문객 1억 5599만 명이 들린 월드 디즈니로 집계됐다.	테마마크 → 테마파크

No.	일시	원문	오류 내용
40	2022.01.26	조합은 수직·수평·별동 증축을 통해 1140가구에서 최대 117가구(일반분양)가 늘어난 1311가구로 짓겠다는 구상이다.	숫자 오류
41	2022.01.28	그러면서 "일반분양 일정이 지연에 따른 부담은 조합과 조합원들에게 있다"라며 조합을 압박했다.	일정이 지연에 → 일정 지연에
42	2022.01.28	육군의 삭제 대상이 되며 최대 서비스 시장 중 하나인 인도에서 위기를 맞게 될 수도 있을 것으로 보인다.	문장 성분 누락
43	2022.01.28	박닌성은 베트남에 중소도시시만 경제 발전은 열 손가락 안에 들고 있다.	중소도시시만 → 중소도시지만
44	2022.01.28	캄보디아는 SD바이오센서를 비롯해 로슈, 애보트, 페이즈 사이언티 등 4개 기업의 제품만 승인했다.	페이즈 사이언티 → 페이즈 사이언티픽
45	2022.01.29	미국 뉴욕 맨해튼에 있는 롯데뉴욕팰리스 호텔이 코로나 19 여파로 아쉬운 9월 보내고 있다.	9월 → 9월을
46	2022.02.03	웨이모는 기존에도 지난 2017년부터 사용한 가사 시뮬레이션 프로그램 '카크래프트'를 가지고 있다.	가사 → 가상
47	2022.02.03	PA는 반도체, LCD나 미세 IT 부품 세척에 쓰는 용제로 대부분 프로플랜에서 추출해 만들어진다.	프로플랜 → 프로필랜
48	2022. 02. 15	이후 주가가 급증하면서 지난달 25일 기준 평가액이 최고 2억 7100만 달러(약 3020억원)까지 치솟았다	급증하면서 → 급등하면서
49	2022.02.03	모두 다부짐 몸을 하고 웃웃을 벗고 달리고 있다.	다부짐 → 다부진
50	2022.02.03	남성이 주변을 두리번거리며 차량으로 다가가 차량 번호판을 땀다.	땀다 → 땀다
51	2022.02.11	CT 찍고 나서 체혈도 할거니까 환복하고 저쪽으로 가시면 돼요.	체혈도 할거니까 → 채혈도 할 거니까
52	2022.02.11	빨라 마트 갔다가 바다 보러 가자.	빨라 → 빨리
53	2022.02.11	근처에 매기 매운탕이 유명한 곳이 있어요.	매기 → 메기

No.	일시	원문	오류 내용
	2		
54	2022.02.1 5	수영과 수영모 추천 좀 해드릴까요?	수영 → 수경
55	2022.02.1 5	오늘 배트민턴 치기 참 좋은 날씨네요.	배트민턴 → 배드민턴
56	2022.02.1 5	친척들 안 본지 너무 오래돼서 보고싶다.	안 본지 → 안 본 지
57	2022.02.1 5	지난 번에 파마한 것도 다들 예쁘다고 했어요	지난 번 → 지난번
58	2022.02.2 3	해물탕에다가 알 추가하고 밥 두 공기도 주세요.	추가하고 → 추가하고
59	2022.02.2 3	영화는 영화관 보다 집에서 많이 봐서 감흥이 없네.	영화관 보다 → 영화관보다
60	2022.02.2 5	지금까지 수많은 사람이 뛰어내렸는데 사고 한번 안 난 거 보면 안전할 거야.	한번 → 한 번
61	2022.02.2 5	중고마켓에서 파는 지 한번 보자.	파는 지 → 파는지

[부록 2]

한국어 원문의 의미 확인 예시

No.	일시	유형	원문	내용
1	2021.12.14	한국어 의미 설명	이탈리아 디자이너 루도비카 버가가 제작한 무아무아돌스는 이번 제휴로 내년 봄부터 마이애미, 세인트 바트, 햄튼 등 주요 명품 매장에서 아마 퓨어와 공동 제작한 판초, 스카프 등을 전시 및 판매할 계획이다.	세인트 바트: St. Barts/St. Barths 둘 다 허용 (지역명, 카리브해에 있는 섬) 햄튼: Hamptons(지역명, 미국 버지니아 주의 부촌)
2	2021.12.17	한국어 의미 설명	지난달 4위에 올랐던 타타 는 전년동기 대비 38.80% 감소한 1만 400대를 판매하며 6위로 내려앉았고 토요타도 22.5%가 감소하며 8312대를 판매해 7위를 기록했다.	타타: 인도의 회사, 영문표기는 TATA
3	2021.12.17	한국어 의미 설명	내년부터 시행되는 국제해사기구(IMO) 환경규제에 따라 저유황 선박 연료유(LSFO) 사용과 관련 선사들의 추가 병커 비용 을 보존하면서 부담을 덜어주기로 한 것이다.	병커 비용: bunker costs, 국가에서 다른 국가로 상품을 운송하는 모든 유형의 해상 운송료
4	2021.12.18	한국어 의미 설명	다미아니 측은 "매장에는 고도로 훈련된 직원들이 고객들과 동반해 다양한 제품을 확인할 수 있게 도와줄 것"이라고 밝혔다.	다미아니: Damiani, 이탈리아 명품 주얼리 브랜드 이름
5	2021.12.19	한국어 의미 설명	야기라 유아는 V4의 슬로건인 ' 싸우거나 공존하거나 선택은 자유 '를 표현할 예정이다.	싸우거나 공존하거나 선택은 자유: V4 게임의 슬로건임.
6	2021.12.20	한국어 의미 설명	이번 만남은 콜롬비아 정부가 주한 콜롬비아대사관을 통해 '2019 콜롬비아 건설 콩그레스 (Columbia Construction Congress)에' HUG 측 참석을 요청하면서 이뤄졌다.	HUG: '주택도시보증공사'의 약자
7	2021.12.22	한국어 의미 설명	온리스크와 캐시카로가 함께 제작한 갤럭시 S11 플러스 렌더링 이미지는 큰 관심을 받으며 많은 보도가 이루어졌	빙우주: IT 인플루언서, Ice Universe

No.	일시	유형	원문	내용
			는데 IT 정보 유출 분야에서 나름의 인지도를 가지고 있는 빙우주 가 반론을 제기해 이목을 끌고 있다.	
8	2021.12.31	한국어 의미 설명	또 그룹 내 전략통인 이주희 부사장보 와 정동혁 부사장보를 각각 전략실 관리총괄과 기획전략본부장으로 발탁했다.	부사장보: AVP(assistant vice president)
9	2022.01.04	한국어 의미 설명	올라용 전기차에는 현대차가 개발한 차세대 전기차 플랫폼 'E-GMP'가 적용될 예정이며, 가격은 100만 루피(약 1660만 원) 이하에 책정될 것으로 알려졌다.	올라: 인도 카헤일링 업체 OLA 올라용 전기차: 올라가 사용할 전기차
10	2022.01.07	한국어 의미 설명	싱가포르 조선소들은 당분간 해양플랜트 부문인 리그선 수주 가뭄에 시달리게 된다.	리그선: 석유시추선
11	2022.01.04	한국어 의미 설명	현대차 전기차 고객들은 바텐폴 공공 충전소를 이용할 수 있을 뿐 아니라 집이나 사업장에도 충전박스를 설치해 이용할 수 있게 된다.	바텐폴: Vattenfall
12	2022.01.22	한국어 의미 설명	이 같은 상황에서 국적 컨테이너 선사 HMM(옛 현대상선) 은 오는 24일 예정대로 2만 4000TEU 초대형 컨테이너선을 투입한다는 방침이다	현대상선: Hyundai Merchant Marine
13	2022.01.22	한국어 의미 설명	그중 가장 돋보이는 전략이 흑미를 이용해 직접 반죽한 맛있는 웰빙도우 와 과일 진액과 토마토홀을 혼합한 특제 소스를 사용한다는 점이다.	웰빙도우: 건강한 식재료로 만든 반죽을 의미
14	2022.01.22	한국어 의미 설명	공사 는 올해 1월 몬테네그로 공항 2곳의 해외공항 투자 개발형(PPP)사업 입찰참가자격 사전심사(PQ)를 통과하며 수주전 참여에 속도를 냈다.	공사: 국가적 사업을 수행하기 위하여 설립된 공공 기업체의 하나 수주전: 물건을 생산하는 업자들이 제품의 주문을 받으려고 서로 극심하게 경쟁하는 일
15	2022.01.2	한국어 의미 설명	원드롭은 지난 3월 미용 의료기기 전문 기업 지티지웰니	1copy™: 코로나19 진단키트 '1 copy™ COVID-19

No.	일시	유형	원문	내용
	2		스와 1copy™의 호주와 일본 등 5개국 독점 판매 협약을 맺고 해외 수출에 매진하고 있다.	qPCR kit'(이하 '1 copy™')
16	2022.01.23	한국어 의미 설명	삼성전자 인니 법인은 "블리블리와 협력해 집콕족들에게 더욱 편리한 쇼핑 경험을 제공할것다"라고 밝혔다.	블리블리: BliBli
17	2022.01.26	한국어 의미 설명	TSMC는 미세 공정에서 초격차 전략을 유지하며 업계 1위로서의 지위를 공고히 한다는 계획이다.	초격차: 더 미세화하고 첨단화하는 과정
18	2022.01.26	한국어 의미 설명	LNG 연료의 주 소비 지역은 유럽이지만, 글로벌 오일 벙커 연료 허브가 된 싱가포르와 말레이시아 시장의 LNG 수요도 증가하고 있다.	글로벌 오일 벙커 연료 허브: 원유, 석유 제품 저장 장치
19	2022.01.26	한국어 의미 설명	넷마블은 킹오파 올스타에 '더 킹 오브 파이터즈 XI'에 등장했던 시온과 B. 제니 캐릭터를 추가했다.	킹오파: the King of Fighters All Stars
20	2022.01.26	한국어 의미 설명	새 지침에는 임상시험 3상 참가자들이 두 번째 백신을 맞은 후 최소 두 달 동안 상태를 추적하는 방안이 담겼다.	3상: 임상 시험의 세 번째 단계. *참고: 임상시험은 기초 연구인 전임상, 첫 번째 단계인 임상 1상, 두 번째 단계인 임상 2상, 세 번째 단계인 임상 3상, 마지막 단계인 임상 4상으로 이루어져 있음.
21	2022.01.26	한국어 의미 설명	시온은 영춘권을 기반의 캐릭터이며 B. 제니는 아랑 MO W에서 등장했으며 스토리에서는 재벌가 본 가문의 외동 딸이자 해적 집단 리린나이츠의 두목으로 표현된다.	시온: Shion, B. 제니: B.Jenet, 아랑 MOW: Garou: Mark of the woules, 리린나이츠: Lillien Knights
22	2022.02.03	한국어 의미 설명	이번에 서부 지역에 호텔을 오픈해 미국 동부와 서부 지역 모두 호텔을 운영함으로써 미국 횡단을 완성하게 됐다.	미국 횡단을 완성하다: 롯데호텔이 미국에 세 개의 호텔을 오픈함. 이것을 비유적으로 미국 횡단을 완성했다고 표현한 것.
23	2022.02.03	한국어 의미 설명	WSJ는 이 회장이 공기 청정 제품이 없는 이유를 물은 후 한 달 만에 '공기 청정 태스크포스'가 꾸려진 사건을 사례로 들었다.	태스크포스: Task Force(TF) 정규 조직과는 다르게, 특정 업무를 해결하거나 사업 목표를 달성하기 위해 전문가 등을 선발하여 '임시로 편성한 조직'을 의미

No.	일시	유형	원문	내용
24	2022.02.03	한국어 의미 설명	경진 대회를 통해 혁신적인 앱 개발을 지원하며 미국의 전방적인 압박 속에 독자 생태계를 강화한다.	독자 생태계: 미국에서 중국 화웨이를 배제하는 정책을 세우고 있음. 그래서 다른 유럽 국가들도 중국 화웨이를 배제한다고 함. 그래서 중국 화웨이도 미국이나 유럽 국가들의 도움 없이도 재능있는 젊은 인재들을 지원하고 있음. = 미국, 유럽의 도움 없이도 스스로 인재를 찾겠다는 의미
25	2022.02.03	한국어 의미 설명	미 11 프로는 미 11과 디스플레이 사양은 동일하지만 1억 800만 화소 카메라 등 쿼드 카메라 및 120W 초고속 충전을 지원한다.	쿼드 카메라: 4개 달린 멀티 카메라. quad camera
26	2022.02.03	한국어 의미 설명	라오어 2는 2013년 최다 고티를 기록한 더 라스트 오브 어스의 7년 만에 후속작으로 지난해 6월 29일 출시됐다.	고티: GotY. Game of the year의 뜻
27	2022.02.03	한국어 의미 설명	해당 패키지는 볼륨 1부터 4까지로 구성되며 낮은 단계의 팩을 구매해야 높은 단계의 팩을 구매할 수 있다.	볼륨: 문맥상으로 보아 '단계'로 해석
28	2022.02.03	한국어 의미 설명	중국 정부는 해외 게임에 대해서 판호를 발급하고 이를 받아야만 서비스할 수 있도록 하고 있다.	판호: 출판물에 대한 국제 표준 번호
29	2022.02.03	한국어 의미 설명	분양 계약자가 시행사에 갖는 손해배상 채권 등을 양수한다.	손해배상 채권 등을 양수한다: 손해배상 채권을 다른 사람에게 주고, 그 대가(돈)를 받는 것.
30	2022.02.03	한국어 의미 설명	할인 외에도 번들에는 구매일부터 30일 동안 뱅골 호랑이 무늬 낙하산, 트라이 하드 그린 마스크맨 백팩, 팝아트 헬멧 등의 아이템이 제공된다.	트라이 하드: 아이템 이름, Tryhard
31	2022.02.03	한국어 의미 설명	상대가 로우킥으로 거리를 두려 하지만 뺨은 다리를 잡혀 균형을 잃고 넘어지고 성오가 재빠르게 마운트를 탄다.	마운트를 탄다: 공격하다
32	2022.02.03	한국어 의미 설명	"돌치가 산호에 숨어 있는 것처럼?" "그럼 이제 돌치가 아닌 뭐 막 광어 이런 게 되는 거지."	돌치: 물고기 이름

No.	일시	유형	원문	내용
33	2022.02.03	한국어 의미 설명	백도는 스완의 알로하 에 환호하는 사람들을 내려다보며 잔에 있던 포트와인을 단번에 마신다.	스완의 알로하: '가수 스완이 노래 알로하를 부른 공연'으로 해석
34	2022.02.03	한국어 의미 설명	리니지 M과 2M은 엔씨소프트의 실적을 견인하고 있는 효자 게임 으로 지난해 매출의 70%를 담당했다.	효자 게임: 큰 매출을 가져다준 게임
35	2022.02.03	한국어 의미 설명	EA는 개발이 지지부진하게 이어지고 있는 '가이아 프로젝트'를 과감히 포기하고 인기 IP, 소위 '될 게임' 에 집중하기로 결정했다.	인기 IP, 될 게임: 이러한 IP게임이 사람들의 인기를 끌기 때문에 '될 게임'이라고 표현한 것
36	2022.02.03	한국어 의미 설명	이번 몬스터는 음과 양의 기운으로 특별한 힘을 사용하는 태생 5성 몬스터 ' 음양사 '와 날카로운 칼날을 휘두르는 태생 4성 몬스터 ' 요괴무사 '다.	음양사: Onmyoji 요괴무사: 의역해야 함.
37	2022.02.03	문장 설명	TfL이 우버의 사업권을 박탈한 동일한 이유로 사업권을 박탈했지만 법원이 우버가 적합한 사업자라는 결정을 내리면서 사업을 재개할 수 있게 됐다.	런던 교통국인 트랜스 포 런던(TfL)'가 올라(인도 택시 운송업 회사)의 런던 운송 사업을 막았다고 함. 이유는 '공공 안전을 위협한다'는 것임.
38	2022.02.08	한국어 의미 설명	네오위즈와 라운드8 스튜디오가 개발하고 있는 MMORPG(다중접속역할수행게임) '블레스 언리쉬드'의 PC 버전 티저 영상이 추가로 공개되며 본격 바람몰이 에 나섰다.	바람몰이: 본격적인 홍보에 나섰다. 많은 사람들의 관심을 끌거나 많은 사람들에게 알리려고 하는 것을 의미함.
39	2022.02.08	한국어 의미 설명	이에 따라 국내 증시 상장사 기준 톱3를 대상으로 CEO·임원과 기술 인력의 연봉 수준을 지면 상·하에 걸쳐 짚어 본다.	수준을 지면 상·하에 걸쳐 짚어 본다: '수준을 전반에 걸쳐 조사한다.' 정도로 이해하면 됨.
40	2022.02.08	한국어 의미 설명	TSMC는 한탕과 판수안 외에도 다른 협력업체들과의 계약을 이르면 시일 내에 완료하고 공장 건설에 착수한다는 방침이다.	한탕: 중국 장시 한탕 시스템 인터그레이션(Jiangxi Hantang System Integration Co., Ltd.이하 한탕) 판수안: fanxuan
41	2022.02.0	한국어 의미 설명	1차 각성명 은 반역을 뜻하는 '리벨리온', 2차 각성명은 피	각성명: 던전앤파이터 게임에서 나오는 말인데 성/명(이

No.	일시	유형	원문	내용
	8		의 복수를 의미하는 '벤데타'로, 특정 세력에게 적의를 품게 된 사연을 암시한다.	름)임. 이 게임에서 '각성'이라는 걸 하면서 캐릭터가 성장한다고 함.
42	2022.02.08	한국어 의미 설명	테슬라 ^향 전기차 배터리에 동박을 납품하며 세계 시장에서 경쟁력을 인정받고 있다.	테슬라 ^향 : 여기서 향은 한자어 向. 테슬라에 전기차 배터리를 납품하며 경쟁력을 인정 받았다는 의미
43	2022.02.08	한국어 의미 설명	완공된 설비는 리비아 '바흐 에살람' 해상 가스전의 생산량을 확장하기 위한 스트러쳐 A&E 프로젝트에 투입된다.	바흐 에살람: dar es salaam
44	2022.02.08	한국어 의미 설명	대만 타이쑈 지방법원은 천정 ^군 을 비롯한 3인에 대해서도 징역형과 벌금형 등을 선고했다.	천정 ^군 : chenzhengkun
45	2022.02.08	한국어 의미 설명	하 ^둠 토벌은 낮은 전투력이 아닌 높은 전투력 유저들을 위한 콘텐츠다.	하 ^둠 토벌: 검은 사막이라는 게임 안에 하나의 단계
46	2022.02.08	한국어 의미 설명	앞서 크라카타우스틸은 지난 7월 포스코와 신일본제철을 상대로 공장 생산시설 인수 제안서를 제출했다.	신일본제철: Nippon Steel
47	2022.02.08	한국어 의미 설명	타이위안 국제공항공사(TIAC)는 지난 18일(현지시간) 제3터미널 토목공사 2단계 입찰에서 삼성물산 [·] 룽공엔지니어링 컨소시엄을 최종 프레젠테이션 참여자로 선정했다.	룽공엔지니어링: RSEA ENGINEERING CORPORATION
48	2022.02.08	한국어 의미 설명	이번 파트너십에 따라 롤리 제휴 사이트에서 삼성전자 TV, 세탁기, 스마트폰, 스마트홈 기기 등을 구매하는 고객은 결제액의 최고 1.5%를 비트코인으로 되돌려 받는다.	롤리: lolli
49	2022.02.08	한국어 의미 설명	재홍은 보드 위에 얹드려 페 ^들 한다.	페 ^들 한다: 패들보트를 타다
50	2022.02.08	한국어 의미 설명	지난 2018년 설립된 스타트업인 코디 ^악 은 캘리포니아와 텍사스에 각각 본사와 테스트 및 운영 시설을 두고 있다.	코디 ^악 : Kodiak Robotics
51	2022.02.08	한국어 의미 설명	또한 게임 내 화폐인 큐 ^나 를 구매할 수 있고 데바 패스라는 배틀 패스 시스템도 도입했다.	큐 ^나 : Quna

No.	일시	유형	원문	내용
52	2022.02.08	한국어 의미 설명	그래도 2차 는 괜찮지 않을까?	2차: 코로나 백신 2차 접종으로 해석 가능
53	2022.02.09	한국어 의미 설명	이에 엔씨를 방문한 매드형 은 환불 혹은 과금한 만큼의 보상을 요구했지만 엔씨측은 이를 "내부 규정상 힘들다"라고 거절했다고 전했다.	매드형: 유튜버 이름
54	2022.02.09	한국어 의미 설명	그럼 저는 후토마끼 랑 우동 주문할게요.	후토마끼: Futomaki
55	2022.02.09	한국어 의미 설명	밸류시스템자산운용은 지난 2009년 설립된 밸류시스템투자자문 이 그 전신으로 2016년 하반기 운용사로 전환했다.	밸류시스템투자자: Valuesystem Financial consulting
56	2022.02.09	한국어 의미 설명	코시국 에 무슨 축제를 한다고.	
57	2022.02.09	한국어 의미 설명	15주년 기념 아트북은 ' 던파 ' 개발 초기부터 현재까지의 스케치, 도트 원화, 일러스트 등을 총망라한 네 번째 공식 아트북이다.	Dungeon & Fighter
58	2022.02.09	한국어 의미 설명	현대차그룹은 최근 시양동핑(向東平) 베이징현대 영업총괄 부총경리를 현대차중국투자유한공사(HMCG) 부사장으로 재배치하고 후임 인사로 ' 현대맨 '을 선임할 계획으로 알려졌다.	현대맨: 현대그룹에 소속한 사람 현대차중국투자유한공사: HMGC
59	2022.02.09	한국어 의미 설명	나한테 귀미테 도 있으니까 그것도 붙이자.	귀미테: 한국에서 파는 멸미약 이름
60	2022.02.11	한국어 의미 설명	그러면 따릉이 를 한번 타볼까?	따릉이: 서울 공용자전거 이름
61	2022.02.11	한국어 의미 설명	진짜 오랜만에 라운딩 이라 더 좋아.	라운딩: 골프 용어

No.	일시	유형	원문	내용
62	2022.02.1 1	문장 설명	나한테 숨겨진 살이 얼마나 많은데!	겉으로는 살이 찢 것 같지 않지만 숨겨져 있는 살이 많다는 말
63	2022.02.1 1	한국어 의미 설명	불명 나 처음 해 보는거라 정말 기대돼.	불명: 불을 보며 아무 생각 없이 가만히 있는 것.
64	2022.02.1 1	한국어 의미 설명	꽃게는 가격 대비 성능비가 별로야 .	성능비가 별로야: 가격이 비싼 것에 비해 양이 적다는 의미로 해석
65	2022.02.1 1	한국어 의미 설명	그런데도 아직 보디 퍼팅 이 정확하게 잘 안 돼서 걱정이예요.	보디 퍼팅: birdie ball putting
66	2022.02.1 1	한국어 의미 설명	조금 비싸서 그렇지 밤 정말 추천해.	밤: BALM, 크림(CREAM) 같은 화장품을 의미
67	2022.02.1 1	한국어 의미 설명	오늘도 수업 있으면 우리 학식 먹을까?	학식: 학생식당의 축약어
68	2022.02.1 1	한국어 의미 설명	오늘 배 에 기름 좀 채워 넣어 보자.	배: 사람의 배
69	2022.02.1 3	한국어 의미 설명	근육이 좀 놀란 거 같은데 평소보다 무리해서 염증이 생긴 건 아닐까 하고요.	근육이 좀 놀란다: 평소보다 과하게 운동하여 근육에 무리가 되었다는 의미
70	2022.02.1 3	문장 설명	얼굴 진짜 소멸할 거 같다.	얼굴이 매우 작다는 의미
71	2022.02.1 3	한국어 의미 설명	산과 내 라는 말이 은근히 매력 있는데.	내: 내((川), 하천
72	2022.02.1 5	한국어 의미 설명	어렸을 때는 진짜 시추 처럼 뿜아서 그렇게 불렀지요.	시추: 개의 종류
73	2022.02.1 5	한국어 의미 설명	아까 바 아이스크림 파는 데도 봤어!	바 아이스크림: 막대 아이스크림
74	2022.02.1	문장 설명	ic칩 부분이 고장난 거 같은데 그냥 뒷면으로 굽어서 해	신용/체크카드에 네모난 칩을 ic칩이라고 하는데 이게 인

No.	일시	유형	원문	내용
	5		주세요.	식이 안 돼서 카드 마그네틱으로 긁어달라는 의미
75	2022.02.23	한국어 의미 설명	주인공은 다양한 마법 카드와 플레이 스타일을 통해 수천 가지가 넘는 조합을 생성, 렐레 가 조종하는 성 안에서 적을 물리치게 된다.	렐레: lele
76	2022.02.23	한국어 의미 설명	그나저나 난 사실 야구알못 이라 그냥 오늘은 먹기만 할게.	야구알못: 야구를 잘 알지 못한다는 의미의 축약어
77	2022.02.23	한국어 의미 설명	너 같은 애 넌리고 깔렸어 .	넌리고 깔렸어: 그 정도로 많다는 의미
78	2022.02.23	한국어 의미 설명	이 정도 성적이면 인 서울 은 가능할 것 같은데요.	인 서울: 서울권에 있는 대학에 간다는 의미
79	2022.02.23	문장 설명	60장은 찍어야 건질까 말까야.	60장을 찍어도 그중에 잘 나온 사진이 1장 있을까 말까 한다는 의미
80	2022.02.23	문장 설명	나는 네가 하면 잘할 거 같아.	‘네가 하면 잘할 거 같다고 나는 생각해.’의 의미로 해석 가능
81	2022.02.23	문장 설명	이 대리님이 피피티 자료 오류 난 거 때문에 부장님께 완전 깨지셨어요 .	깨지셨어요: 크게 혼나다는 의미
82	2022.02.28	한국어 의미 설명	너도 나처럼 열심히 운동해서 얼른 몸 키워 .	몸 키워: 근육질 몸으로 만든다는 의미
83	2022.02.28	한국어 의미 설명	재미있게 짧고 굵게 놀다 갈게요.	짧고 굵게: ‘짧은 시간 동안 아주 재미있게’의 의미
84	2022.02.28	한국어 의미 설명	현재 저희 학교는 모든 교환학생 TO 를 동일하게 두고 있어서요.	TO: Table of Organisation의 약자, 정원을 의미
85	2022.02.28	문장 설명	하늘 아래 같은 마그네틱이 어딴어?	이 세상에 똑같은 마그네틱은 없다는 뜻

No.	일시	유형	원문	내용
86	2022.02.28	한국어 의미 설명	제가 또 운전이라면 한 운전 하거든요.	한 운전: 운전을 아주 잘한다는 뜻
87	2022.02.28	한국어 의미 설명	힙지로 라고 해서 사람들 많이 가던데.	힙지로: Hipjiro, 새롭고 개성 있다는 의미의 ‘힙하다’와 ‘을지로’를 합친 말
88	2022.02.28	한국어 의미 설명	상하이캐리어호 매각에 앞서 지난 3월 인도 폐선소에 재 활용 목적으로 15만 6200DWT 급 하모니 캐리어호 를 매각했다.	하모니 캐리어호: Harmony Carrier
89	2022.03.07	한국어 의미 설명	라면 킬러 는 여전하구만 밥은 적게 먹으면서 라면은 어쩜 그렇게 많이 먹니!	라면 킬러: 라면을 아주 좋아해서 많이 먹는 사람이라는 뜻
90	2022.04.19	한국어 의미 설명	나는 티켓팅만 하면 이선좌던데?	이선좌: 이미 선택된 좌석이라는 뜻의 축약어
91	2022.04.19	문장 설명	예약 취소 건에 있어서는 말씀해주실 수 있으셔야 합니다.	예약 취소의 이유를 말하라는 의미

<Abstract>

2021 Korean-Foreign Language Parallel Corpus Construction Project

The project aims to establish a parallel corpus of Korean and eight foreign languages as the first step in cultural exchanges with eight countries, which are drawing attention as Korea's new partner countries. The target languages for parallel corpus construction are six languages in the ASEAN-India region (Vietnamese, Indonesian, Thai, Hindi, Khmer, and Tagalog) and two languages in the Eurasia region (Russian and Uzbek). In addition to the Korean-foreign language parallel corpus, a foreign language (Vietnamese)-Korean parallel corpus was constructed on a trial basis.

The construction of parallel corpus was largely carried out in four stages: “collection,” “construction,” “refinement,” and “production.” In the “collection” stage, Korean and Vietnamese literary and colloquial texts were collected, and the collected data were proofread and refined. In the “construction” stage, the primary translation of this refined data was carried out according to the translation guidelines established by the project group. The primary translation went through two stages of inspection: first full proofreading by the top local proofreader and second proofreading through the internal language expert (linguist) of Flitto Inc. The proofread translation went through an additional process of full proofreading in the “refinement” stage. The final supervisor then again proceeded with the 10% supervision of the proofread translation. The translation that verified its quality through the above steps was finalized as the final data in the “production” stage.

As a result, a total of 8,400,096 words of Korean-foreign language parallel corpus, 52,500 sentences of foreign language (Vietnamese)-Korean parallel corpus, and additionally 1,050,000 words of Korean-English parallel corpus were constructed.

The expected effects of this parallel corpus construction project are as follows:

First, it can be used not only in related industries but also in various fields such as the research industry and education industry, and in particular, it can be used as basic data for research related to the eight languages that have been insufficient.

Second, learning of ASEAN-India and Eurasian languages, which have been relatively less learned in the field of machine translation, can be made, and the skills and quality of machine translation can be improved through the language-learning of various topics.

Third, by promoting exchange between Korea and the relevant region, enhancing mutual understanding and developing a foundation for economic cooperation on future-oriented coexistence will be possible.

Fourth, it will be able to promote language learning and fields related to interpretation and translation through the development of learning and translation software.

In addition, in order to spread the achievements of the project externally, this project group held a translation and inspection workshop and media promotion. In particular, more than 200 people participated in the translation and inspection workshop, allowing the group to confirm the high interest in building a parallel corpus of Korean-foreign languages and establish a human network. Moreover, the results researched in the process of carrying out the project were presented at an academic conference to inform the importance of the project and promote related research.

The Korean-foreign language parallel corpus constructed with the above aim and process is not only valuable as is but is also expected to serve as a basic tool that can be used in various fields.

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

Project Director: Lee Jung Hee(KyUniversity)

<기획·연구>

국립국어원 강미영 학예연구관

국립국어원 정주연 연구원

국립국어원 유성희 연구원

<연구 참여자>

연구 기관 21 한국어-외국어 병렬 말뭉치 구축 사업단

((사)국제한국어교육학회, (주)플리토)

연구 책임자 이정희(경희대)

연구 참여자 이동은(국민대), 김종민(경희대), 박진욱(대구가톨릭대),
손혜진(고려대), 이수미(성균관대), 이영준(한국학중앙연구원),
임채훈(숭실대), 최문석(경희대), 최홍열(강원대), 권상희(홍익대),
김연희(경희대), 문진숙(경희대), 박지민(경희대), 이두용(고려대),
전지연(강원대), 정성호(경희대), 지화숙(경희대), 한재민(경희대),
홍세화(경희대), 국혜민(숭실대), 김수희(경북대), 김인후(경희대),
김한별(대구가톨릭대), 박민선(한국외대), 박소영(성신여대),
박정모(강원대), 박종우(성신여대), 서유리(경희대),
심하은(성균관대), 오승혜(국민대), 이규민(고려대), 이상후(고려대),
이요셉(경희대), 이재림(국민대), 이종현(강원대), 임예찬(경희대),
조영민(경희대), 주우찬(고려대), 최지선(성신여대),
이정수(주)플리토, 유동현(주)플리토, 강동한(주)플리토,
김진구(주)플리토, 이서현(주)플리토, 최선재(주)플리토

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2022년 5월 26일

발행일: 2022년 5월 26일

인 쇄: 해림복사

※ 이 보고서는 국립국어원의 국고 보조금으로 수행한 ‘2021년 한국어-외국어 병렬 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.