국립국어원 2024-01-05

발간등록번호
11-1371028-000997-01

2023년 한국어-한국수어 병렬 말뭉치 구축

총괄 책임 | 정희찬

2024. 1. 31.

국립국어원

제 출 문

국립국어원장 귀하

국립국어원의 국고 보조금으로 수행한 「2023년 한국어-한국수어 병렬 말뭉치 구축」사업의 결과보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 7월 11일 ~ 2024년 1월 31일

2023년 1월 31일

총괄 책임: 정희찬(한국농아인협회)

사업수행기관	’23년 한국어-한국수어 병렬 말뭉치 구축사업단 ((사)한국농아인협회, (주)케이엘큐브)
총괄 책임	정희찬
실무 책임 및 관리	하윤호, 장철성

<국문 요약>

2023년 한국어-한국수어 병렬 말뭉치 구축

이 사업은 농인 및 청각장애인이 사용하는 수어를 영상 기반으로 인식하여 의사를 전달할 수 있도록 하는 인공지능 기술 및 응용 서비스 개발을 위해 이에 필요한 수어 영상 학습데이터를 구축하는 것을 목적으로 하였다.

한국어-한국수어 병렬 말뭉치 구축은 크게 한국어의 '수집', '정제', '산출' 작업, 한국수어의 '번역', '촬영', '검수' 작업, 인공지능 기술 및 응용 서비스 개발에 필요한 수어 영상 '주석입력', '타입' 작업으로 진행되었다.

한국어 수집 분류는 크게 '의료 분야', '뉴스 분야', '일상생활 분야' 3개의 분야로 구분하고 '의료 분야'는 '시설 안내', '진료 안내', '예약', '입/퇴원', '가정간호', '보건사업', '검진/검사', '진료 상담', '제 증명/자료', '진료비', '보건 행정'의 11가지 항목으로 대분류하였다. 또한 '뉴스 분야'는 '사회', '문화', '국제', '지역'의 4가지 항목으로, '일상생활 분야'는 '예약', '구매', '여행', '교통', '주거', 'A/S'의 6가지 항목으로 대분류하여 수집하였다. 수집한 데이터는 '정제' 지침에 따라 정제 작업을 진행하였으며, 정제된 한국어 문장은 한국어 전문가의 검수를 통해 '산출' 단계에서 한국어 데이터로 산출되었다.

한국수어는 수어 통역 자격을 갖춘 전문가가 번역 지침에 따라 '번역'을 하였으며 상위 검수자와 수어 모델이 공동 검수 과정을 진행한 후 '촬영' 단계를 거쳐 수어 영상 데이터를 산출한 뒤 '(사)한국농아인협회' 수어 전문가의 검수를 거쳐 최종 수어 영상 데이터를 구축하였다.

인공지능 기술 및 응용 서비스 개발에 필요한 인공지능 학습용 데이터로 변환할 수 있도록 촬영된 수어 영상은 전용 프로그램(SLAT-Sign Language Annotation Tool)을 통해 주석 입력 작업을 진행하였다. 주석 입력 작업은 주석의 의미를 갖는 수지 요소와 비수지 요소에 이름을 붙이는 것을 기본으로 하여 한국어에 대응되는 타입(글로스)을 입력하도록 진행하였으며 수어 영상과 수어 주석 데이터는 다시 최종 감수자가 감수를 진행하였다. 이상의 단계를 거쳐 품질을 검증한 뒤 최종 데이터를 산출하였다.

이를 통해 한국어-한국수어 병렬 말뭉치 한국어 1,014,861어절, 수어 영상 105,891건, 인공지능 활용을 위한 JSON 파일 105,891건을 구축하였다. 수집 분야별로는 의료 분야 200,026어절, 뉴스 분야 100,227어절, 일상생활 분야 714,608어절이다.

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

첫째, 인공지능 학습데이터로 변환된 한국수어 데이터는 AI 영상인식, AI 아바타 등을 활용한 융합형 대화 서비스를 통해 농인과 청각장애인이 일상생활에서 원활하게 의사소통할 수 있는 환경을 조성하는 데 이바지할 수 있다.

둘째, 수어는 청각장애인 사회의 주요 언어로 인식됨에 따라, 한국어-한국수어 병렬 말뭉치 구축은 다양한 언어 및 소통 방식을 인정하는 사회적 포용성을 강조하는 데 도움을 줄 수 있다.

셋째, 수어를 통한 소통이 강화되면 청각장애인이 경험하는 사회적 및 직장 내 차별을 줄일 수 있으며, 공공기관이나 기업의 더 포괄적인 정책과 서비스 제공으로 이어질 수 있다.

넷째, 수어를 사용하는 청각장애인들은 자신의 의견을 표현하고 정보에 접근하는데 자립성을 높일 수 있으며. 이는 그들의 삶의 질을 향상시키고 사회적으로 더욱 활발하게 참여할 수 있도록 견인할 수 있다.

본 사업을 통해 한국수어의 인공지능 기술 연구 및 확대에 필요한 중요한 기초 데이터로의 역할을 할 수 있을 것으로 기대된다.

주요어: 한국어, 한국수어, 한국어-한국수어 병렬 말뭉치, 주석, 인공지능 학습데이터

<영문 요약>

2023 Korean-Korean Sign Language Parallel Corpus

This project aimed to build sign language video learning data for the development of artificial intelligence technology and application services that can recognize sign language used by the deaf and hearing impaired based on video and communicate with them.

The construction of the Korean-Korean Sign Language parallel corpus consisted of the ‘collection’, ‘refinement’, and ‘output’ of Korean, ‘translation’, ‘filming’, and ‘inspection’ of Korean Sign Language, and ‘annotation’ and ‘TYPE’ of sign language videos for the development of AI technology and application services.

The classification of the Korean language collection is divided into three areas: ‘medical area’, ‘news area’, and ‘daily life area’. And the ‘medical area’ is divided into eleven major categories: ‘facility information’, ‘medical consultation’, ‘appointment’, ‘admission/discharge’, ‘home care’, ‘health business’, ‘examination/test’, ‘medical consultation’, ‘evidence/data’, ‘medical fee’, and ‘health administration’. The ‘news area’ is divided into ‘society’, ‘news’ is divided into four major categories: ‘Society’, ‘Culture’, ‘International’ and ‘Local’, and ‘Daily Life’ is divided into six major categories: ‘Reservation’, ‘Purchase’, ‘Travel’, ‘Transportation’, ‘Housing’, and ‘After-sales Service’.

The Korean sign language was translated by a qualified sign language interpreter according to the translation guidelines, and after a joint review process between the top reviewer and the sign language model, the sign language video data was produced through the ‘shooting’

stage, and the final sign language video data was built through the re view of the sign language experts of .

In order to convert the captured sign language images into data for AI learning, which is necessary for the development of AI technology and application services, annotation was performed through a dedicate d program(SLAT: Sign Language Annotation Tool). The annotation proc ess was based on naming the resinous and non-resinous elements that have the meaning of the annotation, and the type (gloss) correspondin g to the Korean language was entered, and the sign language video a nd sign language annotation data were again reviewed by the final re viewer. Through the above steps, the quality was verified and calcula ted as the final data.

As a result, we built a Korean-Korean Sign Language parallel corpus of 1,014,861 Korean words, 105,891 sign language videos, and 105,891 JSON files for AI applications. By field of collection, 200,026 words ar e in the medical field, 100,227 words are in the news field, and 714,6 08 words are in the daily life field.

The expected effects of the parallel corpus building plan through thi s project are as follows.

Firstly, the Korean Sign Language data converted into artificial inte lligence training data can contribute to creating an environment wher e the deaf and hearing-impaired people can communicate smoothly in daily life through integrated conversational services utilizing AI video recognition and AI avatars.

Secondly, as sign language is recognized as a major language of the deaf community, the construction of Korean and Korean Sign Languag e Parallel corpus data can help to emphasize social inclusivity by ack nowledging diverse languages and modes of communication.

Thirdly, increased communication through sign languages can reduce social and workplace discrimination experienced by Deaf people, and c an lead to more inclusive policies and service provision by public auth orities and businesses.

Fourthly, deaf people who use sign language can increase their inde pendence in expressing their opinions and accessing information, whic h can improve their quality of life and enable them to participate mo re actively in society.

This project is expected to serve as important basic data for resear ch and expansion of AI technology in Korean Sign Language.

Keywords: Korean, Korean Sign Language, Korean-Korean Sign Lang uage parallel corpus, annotation, AI training data

목 차

Ⅰ. 사업 개요 ····· 1

1. 사업 목표 및 추진 방향 ····· 1

1) 사업 목표 ····· 1

2) 추진 방향 ····· 4

2. 사업 수행 체계 및 절차 ····· 7

1) 사업 수행 체계 및 인력 구성 ····· 7

2) 병렬 말뭉치 구축 절차 ····· 7

3. 사업추진계획 ····· 8

1) 일정별 사업추진계획 ····· 8

2) 세부 사업추진계획 ····· 8

4. 주요 변경 사항 ····· 11

1) 요구사항 ····· 11

2) 작업 및 검수 지침 ····· 11

2) 예산 ····· 11

Ⅱ. 사업 수행 ····· 12

1. 한국어 수집 및 정제 ····· 12

1) 한국어 수집 ····· 12

2) 한국어 정제 ····· 15

3) 한국수어 변환 및 수어 제공 ····· 21

2. 수어 영상 촬영 ····· 23

1) 수어 영상 촬영 기본 원칙 ····· 23

2) 촬영 환경 및 촬영 방식 ····· 23

3) 데이터 정제 방안 ····· 25

3. 주석 입력 ····· 26
1) 주석 기본 원칙 ····· 26
2) 주석 세부 지침 ····· 26
3) 타입 관리 ····· 43
4. JSON 파일 구축 ····· 47
1) 병렬 말뭉치 데이터 구조 ····· 47
2) 데이터 포맷 개요 ····· 48
3) 데이터 백업 관리 ····· 53
5. 병렬 말뭉치 데이터 검수 ····· 54
1) 한국어 데이터 검수 ····· 54
2) 한국수어 영상 검수 ····· 56
3) 주석 검수 ····· 57
6. 병렬 말뭉치 데이터 품질관리 ····· 59
1) 데이터 품질관리 ····· 59
2) 체크리스트 ····· 61
7. 보안 관리 ····· 66
1) 보안 관리 개요 ····· 66
2) 원천 자료 및 구축 자료에 대한 저작권 확보 ····· 66
3) 개인정보보호 등 보안 정책 및 지침 준수 ····· 69
4) 사업 수행을 위한 보안 대책 수립 및 준수 ····· 69
5) 보안 계획 점검 사항 ····· 70
8. 이용자 아카데미(전문가 워크숍) 개최 ····· 71
1) 배경 및 목적 ····· 71
2) 프로그램 및 토의 내용 ····· 71
9. TTA 품질 검증 ····· 75
1) 배경 및 목적 ····· 75
2) 시험규격 및 결과 ····· 75

Ⅲ. 사업 수행 결과 ····· 77
1. 병렬 말뭉치 데이터 구축 결과 ····· 77
1) 최종 구축 데이터 ····· 77

2. 활용 방안 및 기대 효과 ····· 78
1) 병렬 말뭉치의 활용 방안 ····· 78
2) 사업의 기대 효과 ····· 78
3) 제언 ····· 80

참고자료 ····· 81

표 목 차

[표 I-1] 사업 수행 인력 구성안 ····· 7
[표 I-2] 일정별 사업추진 계획 ····· 8
[표 II-1] 수집 대상 세부 분야 ····· 12
[표 II-2] 문어체와 구어체의 비교 ····· 16
[표 II-3] 한국어 문장 작성 기준 ····· 16
[표 II-4] 개인정보 비식별화 지침 ····· 17
[표 II-5] 한국수어 현장 검수 기준 ····· 23
[표 II-6] 촬영 장비 세팅 값 ····· 24
[표 II-7] 촬영 환경 정보 값 ····· 24
[표 II-8] 수어 영상 스튜디오 촬영 환경 ····· 25
[표 II-9] 비수지 표지 층렬 ····· 27
[표 II-10] 일치동사 입력 기준 ····· 35
[표 II-11] 생산적 수어 타입명 ····· 41
[표 II-12] 새 타입 생성 절차 ····· 43
[표 II-13] 수집 대상 세부 분야 ····· 54
[표 II-14] 한국어 문장 검수 기준 ····· 55
[표 II-15] 한국수어 영상 검수 기준 ····· 56
[표 II-16] 주석 검수 기준 ····· 58
[표 II-17] 한국수어 저작도구 검수 기능 ····· 58
[표 II-18] 품질관리 조직 구성 ····· 59
[표 II-19] 품질목표 ····· 60
[표 II-20] 품질점검 일정 ····· 60
[표 II-21] 품질 목표 충족 여부 ····· 60
[표 II-22] 준비성(계획 수립성) 체크리스트 ····· 61
[표 II-23] 준비성(체계 준수성) 체크리스트 ····· 63

[표 Ⅱ-24] 완전성(수집 완전성) 체크리스트 ····· 64
[표 Ⅱ-25] 완전성(정제 완전성) 체크리스트 ····· 65
[표 Ⅱ-26] 완전성(가공 완전성) 체크리스트 ····· 65
[표 Ⅱ-27] 보안 점검 사항 체크리스트 ····· 70
[표 Ⅱ-28] 이용자 아카데미 프로그램 구성 ····· 71
[표 Ⅱ-29] TTA 시험규격서 ····· 75
[표 Ⅲ-1] 최종 구축 데이터 수량 ····· 77

그 림 목 차

[그림 I-1] 수어통역 관련 실태조사 ····· 2
[그림 I-2] 전국 지자체 수어통역센터 운영 현황 ····· 2
[그림 I-3] 사업 추진 필요성 및 배경 ····· 3
[그림 I-4] 애플 SignTime 세션 ····· 5
[그림 I-5] 국내 수어번역 모델 연구 동향 ····· 6
[그림 I-6] 한국어-한국수어 병렬 말뭉치 구축 사업 절차 ····· 7
[그림 Ⅱ-1] 저작도구 문장 중복 확인 ····· 14
[그림 Ⅱ-2] 파일명 코드 부여 지침 ····· 15
[그림 Ⅱ-3] 의료 분야 데이터 파일 고유 코드 구성 ····· 18
[그림 Ⅱ-4] 일상생활 분야 데이터 파일 고유 코드 구성 ····· 18
[그림 Ⅱ-5] 뉴스 분야 데이터 파일 고유 코드 구성 ····· 19
[그림 Ⅱ-6] 저작재산권 이용 허락 계약서 견본 ····· 20
[그림 Ⅱ-7] 수어 표현 범위를 고려한 촬영 화면 구성 ····· 23
[그림 Ⅱ-8] 주석 층렬 1 ····· 26
[그림 Ⅱ-9] 주석 층렬 2 ····· 27
[그림 Ⅱ-10] 주석 분절 기준 ····· 28
[그림 Ⅱ-11] 수어 분절 예시 ····· 28
[그림 Ⅱ-12] 수어 시작점 입력 예시 ····· 28
[그림 Ⅱ-13] 수어 마지막 지점 입력 예시 1 ····· 29
[그림 Ⅱ-14] 수어 마지막 지점 입력 예시 2 ····· 29
[그림 Ⅱ-15] 지화 입력 화면 ····· 30
[그림 Ⅱ-16] 지숫자 입력 화면 ····· 30
[그림 Ⅱ-17] 수표현 입력 화면 ····· 31
[그림 Ⅱ-18] 새 타입 촬영 예시 ····· 31
[그림 Ⅱ-19] 변이형 타입명 예시 ····· 32

[그림 Ⅱ-20] 한 손, 양손 수어 변이형 주석 예시 ····· 33
[그림 Ⅱ-21] 한 손 수어 주석 예시 ····· 33
[그림 Ⅱ-22] 변형 기준에 부합하지 않는 예시 ····· 34
[그림 Ⅱ-23] 기본형과 변형 기준에 부합하지 않는 예시 ····· 34
[그림 Ⅱ-24] 변형 기준에 부합하지 않는 예시 ····· 35
[그림 Ⅱ-25] 일치동사 입력 예시 ····· 35
[그림 Ⅱ-26] 강약 표현 예시 ····· 36
[그림 Ⅱ-27] 수어 이동 변형 타입 예시 ····· 36
[그림 Ⅱ-28] 분류사 타입 생성 예시 ····· 37
[그림 Ⅱ-29] 숫자 분류사 타입 생성 예시 ····· 37
[그림 Ⅱ-30] 수 포합어 입력 예시 ····· 38
[그림 Ⅱ-31] 동시적 결합 구조 예시 ····· 38
[그림 Ⅱ-32] 계기적 결합 구조 예시 ····· 39
[그림 Ⅱ-33] 생산적 수어와 타입명 입력 예시 ····· 42
[그림 Ⅱ-34] 자문 예시 ····· 44
[그림 Ⅱ-35] 타입 업로드 예시 ····· 44
[그림 Ⅱ-36] 지침 불일치 시 타입명 변경 예시 ····· 45
[그림 Ⅱ-37] 외래어 타입명 변경 예시 ····· 45
[그림 Ⅱ-38] 동형이의어 타입명 변경 예시 ····· 45
[그림 Ⅱ-39] 타입 오류 기록 예시 ····· 46
[그림 Ⅱ-40] 한국어-한국수어 병렬 말뭉치 구축 사업 절차 ····· 47
[그림 Ⅱ-41] 데이터 포맷 종류 및 영상 명명 규칙 ····· 48
[그림 Ⅱ-42] 데이터 백업 관리 과정 ····· 53
[그림 Ⅱ-43] 한국수어 영상 파일관리 시스템 화면 ····· 57
[그림 Ⅱ-44] 한국수어 영상 주석 프로그램 검수 화면 ····· 57
[그림 Ⅱ-45] 보안 관리 전략 ····· 66
[그림 Ⅱ-46] 저작권 이용 동의서 견본 ····· 67
[그림 Ⅱ-47] 저작권 법률 검토 ····· 68

[그림 Ⅱ-48] 원시데이터 활용에 따른 데이터 이용 허락 동의서 ····· 68
[그림 Ⅱ-49] 이용자 아카데미 현장 사진 ····· 74
[그림 Ⅱ-50] 제3자 품질 검증 ····· 76
[그림 Ⅲ-1] 수어번역 서비스 기대 효과 ····· 79

Ⅰ. 사업 개요

1. 사업 목표 및 추진 방향

1) 사업 목표

(1) 추진 배경

ㅇ 언어적 다양성 인식

- 수어는 한국 내 청각장애인 사회의 주요 언어 중 하나지만, 수어를 지원하는 자료와 자원이 부족하여 청각장애인들이 자신의 언어로 의사소통하는 데 제약을 겪고 있다.

ㅇ 소통의 어려움

- 청각장애인은 언어적인 소통에 어려움을 겪고 있으며, 기존의 문자 기반 정보는 청각장애인들이 접근하기 어려워 사회적 참여와 정보 획득에 제약받고 있다.

ㅇ 사회적 포용성 강화

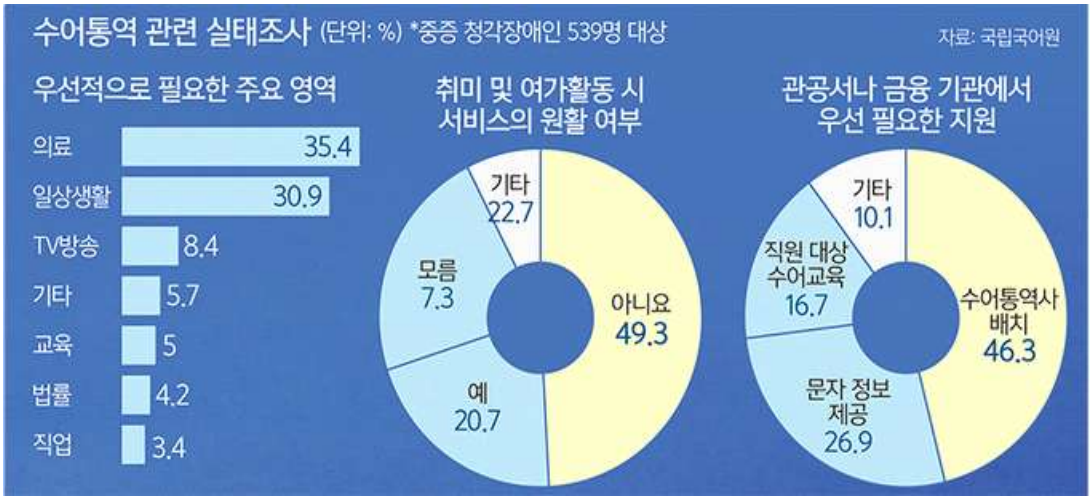
- 한국수어를 포함한 수어는 언어적 소수자에 속하는 청각장애인들의 주요 의사소통 수단으로, 언어적 다양성을 인정하고 지원함으로써 사회적 포용성을 증진할 필요가 있다.

ㅇ 법률 및 정책적 요구

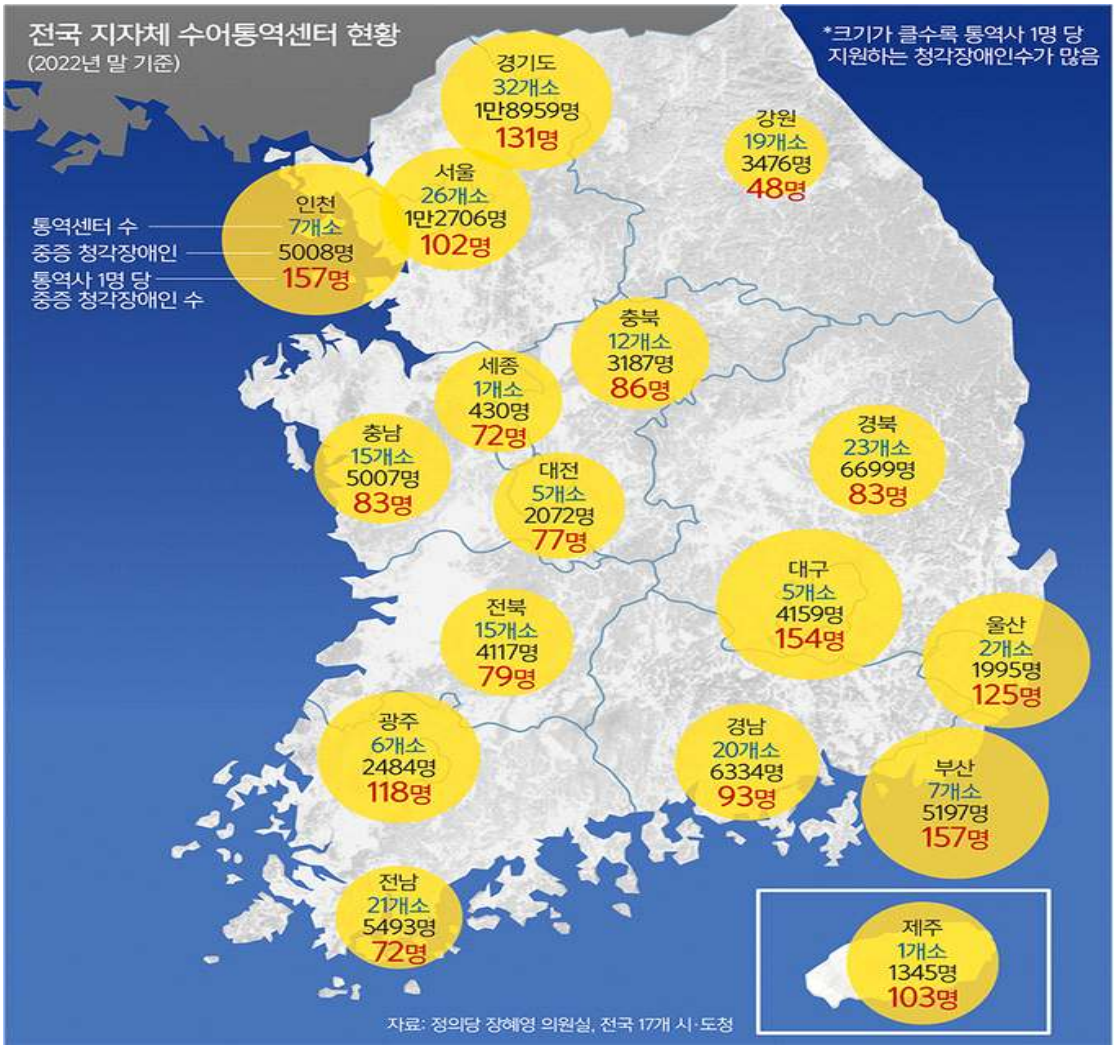
- 많은 국가가 청각장애인의 권리 보장을 위해 수어에 대한 인식과 지원을 강화하고 있으며, 국내에서도 이에 대한 법률적 요구와 정책적 필요성이 증대되고 있다.

ㅇ 기술 발전과 수요 증가

- 인공지능 및 음성 인식 기술의 발전으로 수어 번역 및 인터페이스 기술의 수요가 증가함에 따라 한국어와 한국수어 말뭉치 데이터의 구축이 더욱 절실히 필요해지고 있다.



[그림 I-1] 수어통역 관련 실태조사



[그림 I-2] 전국 지자체 수어통역센터 운영 현황

(2) 사업의 필요성

ㅇ 자동화된 수어 번역 시스템 개발: 한국어-한국수어 병렬 말뭉치는 수어와 한국어 간의 자동 번역 시스템을 개발하는 데 필요한 핵심 데이터로, 이를 통해 인공지능 기반의 수어 번역 시스템을 개발· 향상하는 등 수어 사용자와의 의사소통을 원활하게 돕는 기술을 개발할 수 있다.

ㅇ 자연어 처리 연구: 한국수어의 특성과 구조를 연구하기 위해서는 대량의 한국수어 말뭉치가 필요하고 이를 통해 한국수어의 어휘, 문법, 문장 구조 등을 이해하고 분석함으로써 자연어 처리 기술의 발전에 기여할 수 있다.

ㅇ 자동 수어 인식 및 자막 생성: 한국어-한국수어 병렬 말뭉치를 활용하여 인공지능 모델을 훈련하면, 수어를 자동으로 인식하여 이를 텍스트나 자막으로 변환하는 기술을 개발할 수 있다.

ㅇ 의료 및 장애인 서비스 개선: 의료 분야나 장애인 서비스 분야에서 활용할 수 있으며, 예를 들어 한국수어를 사용하는 환자들에게 필요한 의료 정보를 제공하거나 수어 통역 서비스를 개선하여 장애인들의 의사소통을 원활하게 할 수 있다.

ㅇ 교육 자료 개발 : 한국수어를 사용하는 교육자나 학습자들을 위해 말뭉치를 분석하여 학습 자료나 교재를 개발하고, 수어 교육의 질을 향상하는 등 한국수어 교육 자료 개발에 기여할 수 있다.

[그림 I-3] 사업 추진 필요성 및 배경

2) 추진 방향

(1) 관련 동향

ㅇ 수어 인식 증대

- 최근 몇 년간 청각장애인 사회의 수어 인식이 높아지고 있다.
- 이는 수어를 사용하는 청각장애인들이 사회적으로 활발한 활동을 통해 그들의 언어와 문화를 대중에 알리고자 하는 노력을 보여주고 있기 때문이다.
- 또한, 수어를 배우고자 하는 비청각장애인들의 관심도 증가하고 있으며, 이러한 수어 인식 증대는 청각장애인들의 사회적 참여와 자존감 향상에 도움을 줄 것으로 기대할 수 있다.

ㅇ 기술의 발전

- 최근 급속히 진행되는 기술 발전으로 수어 인식 및 번역 기술 또한 크게 발전하고 있다.
- 특히 인공지능 및 기계 학습 기술을 활용하여 수어를 인식하고 번역하는 기술이 점차 정교해지고 있다.
- 이는 청각장애인들이 수어를 사용하여 더 원활하게 의사소통할 수 있도록 돕는 중요한 기술적 지원을 제공하고 있다.

ㅇ 법률 및 정책의 강화

- 정부는 최근 청각장애인의 권리를 보장하기 위한 법률 및 정책을 강화하고 있다.
- 장애인차별금지법 개정을 통해 총 3개년에 걸쳐 키오스크, 웹, 앱 등의 수어 번역 서비스를 의무화하고 있다.
- 또한 장애물 없는 생활환경 인증제를 법제화함에 따라 어린이·노인·장애인·임산부뿐만 아니라 일시적 장애인 등이 개별시설물·지역을 접근·이용·이동함에 있어 불편을 느끼지 않도록 관련 계획·설계·시공·관리 여부를 공신력 있는 기관이 평가하여 인증하는 법적 환경이 확대되고 있다.

ㅇ 온라인 환경의 변화

- 신종 코로나 바이러스 감염증의 세계적 대유행으로 인해 온라인 환경의 중요성이 더욱 부각되면서, 온라인상의 수어 서비스 및 교육이 활발해지고 있다.
- 이는 청각장애인들이 온라인을 통해 자유로이 의사소통을 이루고

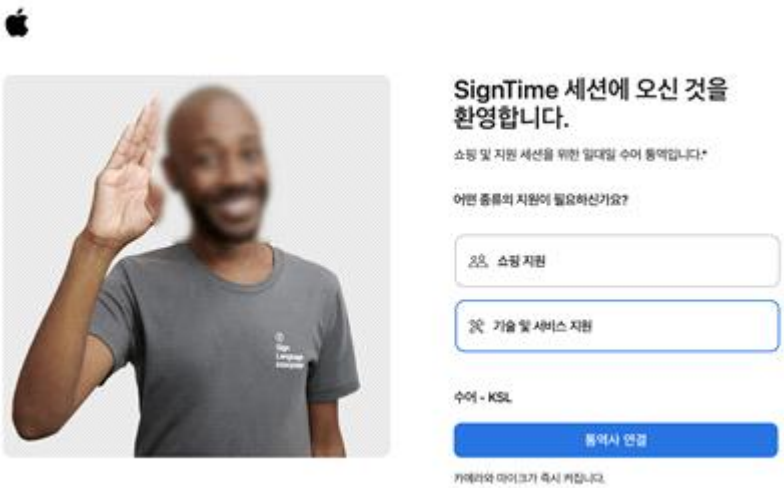
정보를 얻는 데 더욱 편리하고 유용한 환경을 제공해 주고 있다.

ㅇ 사회적 인식의 변화

- 과거에는 청각장애인들이 사회적 소수 집단으로 취급되었지만, 최근에는 그들의 다양성과 인권을 존중하고 보장하려는 인식이 사회적으로 점차 확산하고 있다.
- 사회 구성원이 지닌 다양성과 이에 대한 존중 및 포용성에 대한 인식 확산은 청각장애인들이 더 자신감 있게 사회 활동에 참여할 수 있는 기회를 제공해 주고 있다.

ㅇ 해외 수어 번역 연구 동향

- 글로벌 기업 구글은 유튜브-ASL 수어 데이터 공개, 애플 수어 통역 서비스 제공 등을 통해 대규모 수어 데이터 수집을 진행하고 있다.



[그림 I-4] 애플 SignTime 세션

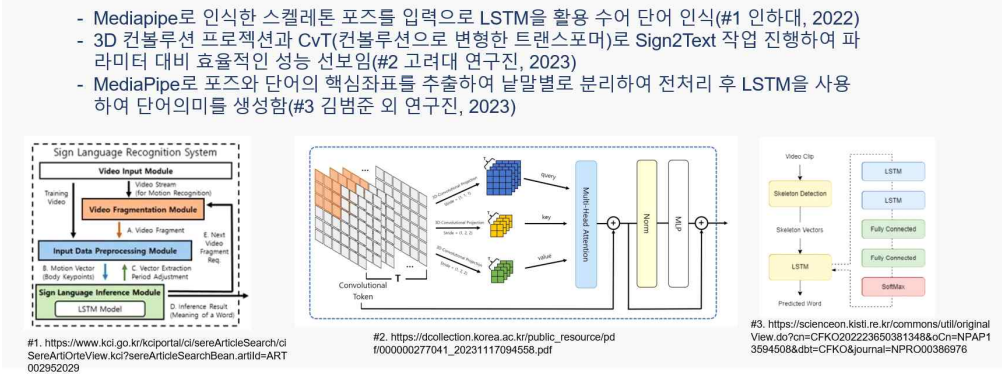
- 딥러닝의 활용: 최근 해외에서는 딥러닝 기술을 활용한 종단 간 수어 연구가 활발하게 진행되고 있으며, 사전 학습 기반의 트랜스포머 모델과 노이즈를 여러 단계에 걸쳐 복원해 나가는 확산 모델이 주로 사용되고 있다.
- 다양한 형태의 수어 데이터셋이 구축되고 있으며, 이를 통해 종단 간 방법 외에도 다양한 접근이 시도되고 있다. 이에 따라 다양한 데이터셋의 활용을 위한 표준화와 모델의 정확한 평가 방법이 중요해지고 있다.

ㅇ 국내 수어 번역 연구 동향

- 초기에는 수어의 구조와 문법을 이해하고 이를 기반으로 하는 수어-한국어 번역 시스템 개발에 초점이 맞춰져 있었으나 최근에는 딥

러닝 기술을 활용한 수어 인식 및 번역 연구가 진행되고 있다.

- 특히, CNN과 RNN을 활용한 모델을 거쳐, 트랜스포머 모델을 이용한 사전학습 모델 연구가 활발하게 이루어지고 있다.



[그림 I-5] 국내 수어번역 모델 연구 동향

(2) 추진 목적

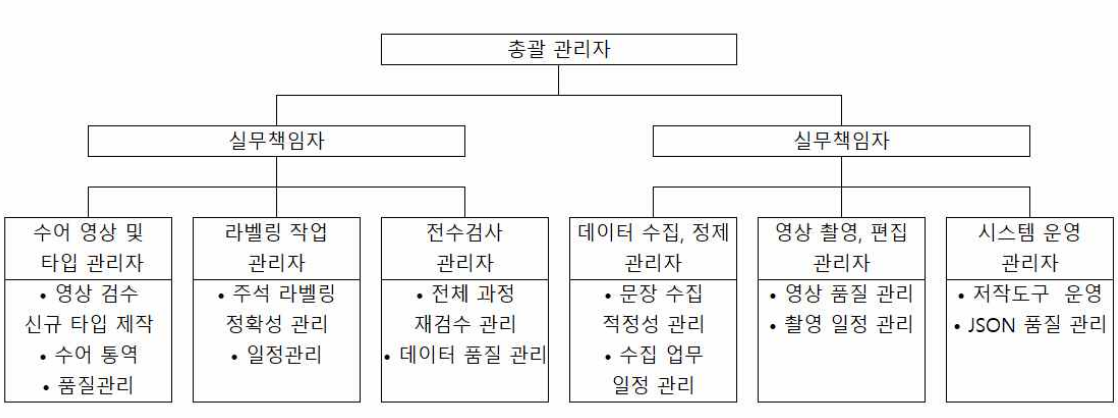
ㅇ 농인들을 위한 수어 데이터 수집

- 농인이 사용하는 수어를 영상 기반으로 인식, 의사 전달을 수행하는 인공 지능 기술 개발을 위해 이에 필요한 수어 영상 학습데이터를 구축하고자 한다.
- 본 사업을 통해 인공지능 학습용 데이터로 ① 한국어 문장, ② 한국수어 영상(문장 단위 분절, 1920*1080p 이상), ③ 형태소(토큰)·타입(단어 사전)·비수지·한국어 대응 정보 등 분석 자료, ④ 분석 자료의 JSON 파일을 구축하고자 한다.

2. 사업 수행 체계 및 절차

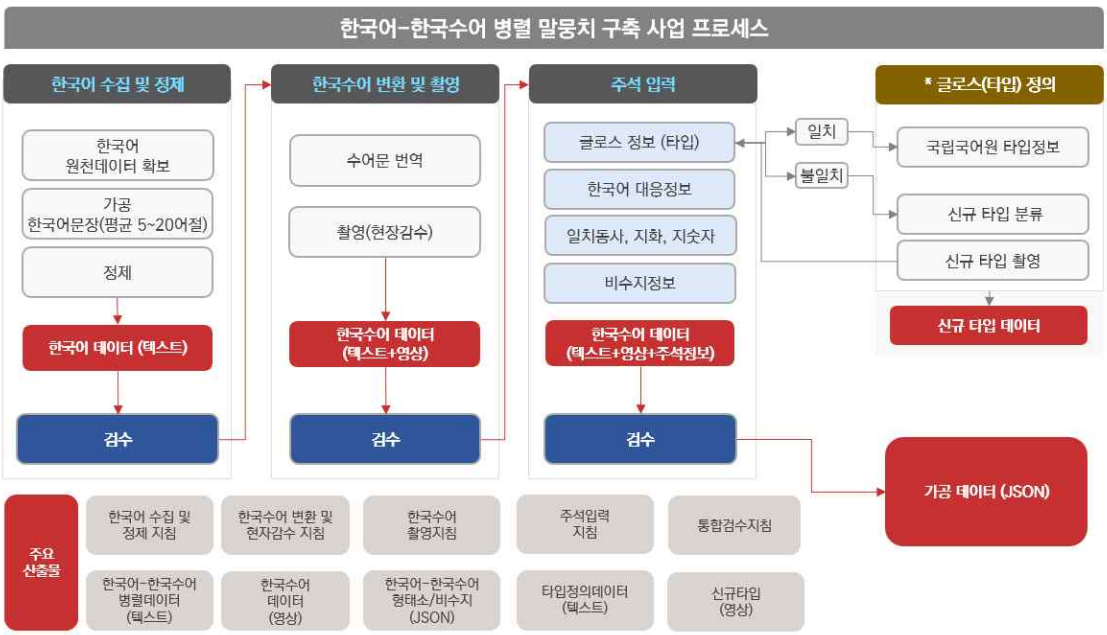
1) 사업 수행 체계 및 인력 구성

○ 본 사업은 총괄 책임자 1인, 실무 책임자 2인, 그리고 실무 관리자 6인으로 구성되었다.



[표 Ⅰ-1] 사업 수행 인력 구성안

2) 병렬 말뭉치 구축 절차



[그림 I-6] 한국어-한국수어 병렬 말뭉치 구축 사업 절차

3. 사업추진계획

1) 일정별 사업 추진 계획

세부 과제명	수행내용	추 진 일 정(월)							비중 (%)
		7	8	9	10	11	12	1	
프로젝트 착수	사업 수행 계획 수립								
	품질 관리 계획 수립								
인프라 구축	영상 촬영 계획 수립								
한국어 수집 및 정제	한국어 수집 카테고리 분류								
	한국어 수집 및 정제								
수어 영상 및 감수	수어 영상 촬영								
	수어 영상 감수								
수어 주석 입력	수어 주석 입력								
타입 분류 및 관리	타입 정의 분류								
	신규 타입 정의 및 촬영								
저작도구 운영	저작도구 운영 및 서버 관리								
이용자 아카데미 및 홍보	전문가 공동연수회 및 홍보 영상								

[표 Ⅰ-2] 일정별 사업추진 계획

2) 세부 사업 추진 계획

(1) 사업 계획 수립 및 착수

- ㅇ 사업 수행 계획 수립: 2023년 7월 30일 완료
- ㅇ 사업 수행 계획서 제출: 2023년 7월 30일 완료

(2) 병렬 말뭉치 데이터 구축

가) 한국어

- ㅇ 데이터 수집 및 정제: 2023년 8월 12일 ~ 2023년 12월 8일(완료)

나) 수어 영상

ㅇ 수어 영상: 2023년 8월 8일 ~ 2024년 1월 19일(완료)

다) 수어 주석

ㅇ 수어 주석 라벨링: 2023년 10월 1일 ~ 2024년 1월 24일(완료)

라) 타입 구분

ㅇ 신규 타입: 2023년 10월 19일 ~ 2024년 1월 24일(완료)

(3) 검수

가) 작업・검수 지침서 작성

ㅇ 한국어 수집 및 정제 지침서 작성: 2023년 10월 10일 완료
ㅇ 한국수어 변환 및 현장 감수 지침서 작성: 2023년 10월 10일 완료
ㅇ 수어 영상 촬영 지침서 작성: 2023년 10월 10일 완료
ㅇ 주석 및 타입 입력 지침서 작성: 2023년 10월 10일 완료
ㅇ 통합 검수 지침서 작성: 2023년 10월 10일 완료

나) 작업・검수 지침 교육

ㅇ 사업 절차 및 품질관리: 2023년 8월 1일 10:00~11:00
ㅇ 한국어 수집 및 정제
- 1차: 2023년 8월 11일 10:00~11:00
- 2차: 2023년 8월 18일 14:00~15:00
ㅇ 수어 영상
- 1차: 2023년 8월 16일 10:00~11:00
- 2차: 2023년 9월 1일 10:00~11:00
- 3차: 2023년 10월 4일 10:00~11:00
ㅇ 주석 입력 및 검수
- 1차: 2023년 9월 12일 14:00~16:00
- 2차: 2023년 9월 13일 10:00~12:00
- 3차: 2023년 10월 4일 10:00~12:00
- 4차: 2023년 11월 1일 14:00~16:00

ㅇ 이후 인력 충원에 따라 수시 실시

(4) 인력 확보

가) 한국어 수집 및 정제: 2023년 11월 10일 완료

나) 수어 모델: 2023년 10월 4일 완료

다) 현장 감수: 2023년 8월 16일 완료

라) 영상 검수: 2023년 8월 16일 완료

바) 주석 입력: 2023년 11월 1일 완료

사) 타입 관리 : 2023년 9월 13일 완료

아) 최종 라벨링 검수: 2024년 1월 31일 완료

(5) 의사소통 및 의견 수렴

가) 단계별 보고

ㅇ 착수 보고: 2023년 7월 25일

나) 정기 보고 및 회의

ㅇ 주간 보고: 매주 화요일 실시

ㅇ 월간 보고: 매월 1주 차 화요일 실시

(6) 이용자 아카데미 및 홍보

가) 이용자 아카데미

ㅇ 전문가 공동연수회: 2023년 11월 21일

나) 홍보 진행

ㅇ 모두의 말뭉치 홍보 영상: 2024년 1월 31일

4. 주요 변경 사항

1) 요구 사항

ㅇ 요구 사항 정의서 일부 변경

2) 작업 및 검수 지침

ㅇ 한국어 수집 및 정제 지침 일부 내용 추가(V1.3)
ㅇ 한국수어 변환 및 현장 감수 지침 일부 내용 추가(V1.2)
ㅇ 수어 영상 촬영 지침 일부 내용 추가(V1.1)
ㅇ 주석 및 타입 입력 지침 일부 내용 추가(V1.4)
ㅇ 통합 검수 지침 신규 작성(V1.0)

Ⅱ. 사업 수행

1. 한국어 수집 및 정제

1) 한국어 수집

(1) 한국어 문장 데이터 수집 분야

ㅇ 한국어 문장의 수집 분류는 구어체의 일상 대화와 문어체로 구분하여 진행하였다. 일상 대화는 기존 구축 병렬 말뭉치(2022년~2023년)의 대화 주제와 중복되지 않도록 하였다.

ㅇ 2023년 한국어-한국수어 병렬 말뭉치 주제는 크게 3가지로 선정하고 대화 주제와 세부 주제 예시를 제시하여 수집을 진행하였다.

분야	대분류	세부 예시 주제	수집 범위
의료	시설 안내	• 오시는 길, 진찰실 안내, 병원 내 시설 안내	20%
	진료 안내	• 증상에 따른 진료 과목 안내, 진료 과목의 설명	
	예약	• 병원 예약 문의, 예약 절차, 예약 확인	
	입/퇴원	• 입원/퇴원의 절차 및 안내	
	가정간호	• 가정간호 제도 안내, 신청 방법	
	보건사업	• 각종 보건 제도 안내, 보건소 이용 안내	
	검진/검사	• 검진 및 검사 방법 안내	
	진료 상담	• 각종 질병에 대한 문의 및 답변	
	제 증명/자료	• 각종 증명서 안내 및 발급 방법	
	진료비	• 진료비 청구, 수납, 건강보험 처리 절차	
	보건 행정	• 보건소 이용 시 행정 처리 안내	
일상생활	예약	• 숙박, 식당, 문화시설 등 예약 상황	70%
	구매	• 백화점, 편의점, 대형마트 등 구매 상황	

	여행	• 유명 관광지 안내, 여행 시 주의사항	
	교통	• 대중교통 이용, 사고 접수 및 처리 방법	
	주거	• 주거복지, 매매, 전월세 계약, 시설 관리	
	A/S	• 전자 제품, 가구 등 A/S 문의 및 답변	
뉴스	사회	• 사건 사고, 인권/복지 등 사회 뉴스	10%
	문화	• 공연/전시, 여행/레저, 패션/뷰티 등 문화 뉴스	
	국제	• 국제 상황, 세계 소식 등 국제 뉴스	
	지역	• 지역 이슈, 행사, 정책 등 지역 뉴스	

[표 II-1] 수집 대상 세부 분야

ㅇ 참여자들에게 위의 대화 주제를 주고, 도움이 될 수 있는 원시 데이터를 참고 자료로 제공하여 자연스러운 대화 및 문장 구성을 유도하였다.

(2) 원시 데이터 획득

ㅇ 한국지능정보사회진흥원(NIA) AI Hub 텍스트 데이터 활용

- 한국지능정보사회진흥원(NIA)의 AI Hub에 등록된 인공지능 학습용 데이터를 활용하여 텍스트 및 음성 데이터를 수집하였다.
- AI Hub에 업로드된 다양한 데이터 중 용도별 목적 대화 데이터와 복지 분야 콜센터 상담 데이터는 본 과업에서 구축될 말뭉치의 주제와 유사성을 보임에 따라 오픈된 데이터를 활용하여 최종 산출 목표인 한국어 문장을 수집하였다.
- 용도별 목적 대화 데이터는 다양한 분야의 고객 상담형 대화, 주문 및 예약형 대화 등 고객 문의와 그에 대한 응대를 위한 목적별 대화 데이터로 각기 다른 용도의 플랫폼에서 수집한 용도별 목적 대화 데이터셋으로 구축하였다.
- 용도별 데이터는 텍스트 데이터 의료 분야 총 1억 어절 및 일상 대화 200만 어절 이상을 확보하여 세부 주제로 분류하였다.
- 의료 분야의 경우 오픈 데이터 활용에 제약이 있으면 크라우드 작업자를 활용한 시나리오 작업을 통해 일상생활 중 의료 서비스 안내 및 방문 진료 등의 상황에 관한 대화 형식의 한국어 문장 데이터를 수집하였다.

○ 원시 데이터 획득 관련 이슈 사항

- 민원 접수, 배송 조회, 서비스 조회 등 일부 주제에 대해서는 개인정보가 포함되어 있어 취급 시 각별한 주의를 기울여 수집하였다.
- 공공 기관 민원/안내의 경우 단순 민원 안내가 대부분이며 기관의 실과별 민원 안내 관련 문서에는 개인정보가 포함되어 있어 취급 시 주의를 기울여 수집하였다.
- 인허가 등 민원에서는 민감 정보에 대한 보안 대책과 가공 대책을 견고하게 수립 및 준수하여 수집하였다.
- 기존에 구축된 텍스트 데이터와 중복성 및 차별성을 비교 및 분석하였다.

○ 원시 데이터 획득 시 적합성 검토 및 원시 데이터 선정

- 원시 데이터 적합성 검토
  - 원시 대화 데이터를 확보하기 위하여 크라우드 소싱을 이용하거나 온라인 민원/안내 게시판을 활용하여 대화 데이터를 구성하였다.
  - 저작권 및 개인정보보호법 등 법적 문제가 발생하지 않도록 데이터 제공 기관과의 업무 협약, 개인정보 삭제를 통한 가명/익명 정보화를 통해 데이터를 확보하고, 크라우드 소싱을 통한 대화 수집을 위해 관련 활용 동의 과정을 사전 수행하였다.
  - 저작도구를 이용한 문장 중복 확인을 통해 원시데이터의 다양성을 확보하였다.

[그림 Ⅱ-1] 저작도구 문장 중복 확인

- 원시 데이터 선정
  - 데이터 품질, 획득 가능성(가능 여부 및 획득량), 획득 비용 및 기술 수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정하였다.

2) 한국어 정제

(1) 원시 데이터 정제

ㅇ 원시 데이터 정제 방식 기준

- 원시 데이터 획득 및 정제 절차를 데이터 획득 분야, 방법별로 아래와 같이 정의하였다.
- 인공지능 기술과 데이터 구축 의도 및 작업 방식 가이드를 제공하여 데이터 관리에 대한 책임, 행정, 제반 절차를 최소화하여 빠른 데이터 획득 및 정제를 지원하였다.
- 수집한 대화 데이터 대부분은 사적인 대화 내용 안에 이름과 연락처, 소속 등 개인의 신원이 노출될 수 있는 다양한 개인정보를 포함하고 있으므로 이에 대한 정제 방침을 수립 및 준수하였다.
- 대화 내용에 포함된 개인정보와 메타정보로 수집하는 성별과 연령, 직업, 출신지 등의 정보가 결합한 형태로 말뭉치로 구축될 경우 개인의 신원이 노출될 우려가 있어 개인정보에 대한 철저한 비식별화를 진행하였다.
- 대화 내용에 비윤리적인 내용이 포함되어 있는 경우 원시 데이터에서 제외하였다.
- 과도한 비식별화로 인하여 대화 내용을 파악하기 불가능하거나, 유효한 개체명(entity)을 추출하기 어려워지는 등 자료의 활용도가 떨어질 가능성이 있으므로 개인정보가 노출되지 않으면서 대화 특성은 잘 반영될 수 있는 비식별화 지침을 마련 및 준수하였다.

<table>
<tr><th>자리</th><th>1</th><th>2 3</th><th>4 5</th><th>6 7</th><th>8 9</th><th>10 11 12</th><th>13 14</th><th>15 16 17 18 19 20 21 22</th></tr>
<tr><td>속성</td><td>유형</td><td>분야</td><td>장르</td><td>병렬</td><td>원자료</td><td>결과자료</td><td>구축연도</td><td>일련번호(8자리)</td></tr>
<tr><td rowspan="4">정의값</td><td rowspan="4">N:문어
S:구어</td><td rowspan="2">LI : 생활</td><td>ME:의료</td><td rowspan="4">PA :병렬</td><td rowspan="4">KO:한국어</td><td rowspan="4">KSL :한국수어</td><td rowspan="4">22</td><td rowspan="4">00000001 ~ 99999999
(8자리 일련번호)</td></tr>
<tr><td>CC:민원행정</td></tr>
<tr><td rowspan="2">CU : 문화</td><td>SH:쇼핑</td></tr>
<tr><td>TO:관광</td></tr>
<tr><td colspan="9">※ 예시 : SLICCPAKOKSL2200019426.json 2022년에 구축한 한국어-한국수어 병렬 말뭉치 파일 (분야: 생활, 장르 : 민원행정)
SCUTOPAKOKSL2200009741.json 2022년도에 구축한 한국어-한국수어 병렬 말뭉치 파일 (분야 : 문화, 장르 : 관광)</td></tr>
</table>

[그림 Ⅱ-2] 파일명 코드 부여 지침

(2) 한국어 데이터 구축 및 정제 지침

ㅇ 한국어 문장은 수집 분야 주제에 맞게 구어체와 문어체로 구분하여 구축하였다.

구분	의미	예시
문어체	일상적인 대화에서 쓰는 말투가 아닌, 글에서 주로 쓰는 말투에 해당하는 문체를 문어체라 한다.	현재는 정말 힘듭니다. 그러나 저는 꼭 가수가 될 것입니다.
구어체	일상생활에서 실제 입으로 발화되는 말투를 그대로 글로 옮긴 것을 말한다.	지금은 진짜 힘들어요. 그렇지만 저는 꼭 가수가 될 거예요

[표 Ⅱ-2] 문어체와 구어체의 비교

ㅇ 한국어 문장은 의미, 내용, 맥락에 따라 적합한 문장을 구성하며 문법 및 수어 번역을 고려하여 작성하였다.

구분	세부 작성 기준
의미에 따른 작성 기준	▪ 정확성: 사용하는 단어와 표현이 의미를 정확히 전달해야 한다. ▪ 명료성: 문장이 간결하고 명확해야 하며 모호성이 없어야 한다. ▪ 적절성: 문장 내용이 적절해야 하며 불필요한 오해를 불러일으키지 않는 내용이어야 한다.
내용에 따른 작성 기준	▪ 완결성: 한 문장이 하나의 완결된 생각을 표현해야 한다. ▪ 연관성: 문장 전후가 내용적으로 연관되어야 한다.
맥락에 따른 작성 기준	▪ 상황 인식: 문장은 맥락상 적절해야 하며, 문화적, 상황적 요소를 고려하여야 한다. ▪ 톤과 레벨: 대화의 상황, 청자 또는 화자의 위치에 따른 적절한 어조와 수준을 고려하여야 한다.
문법적/언어적 측면 작성 기준	▪ 문법 준수: 한국어 문법 규칙에 따라 올바른 문장을 구성해야 하며 비문이나 내용 오류가 없어야 하고 오자, 띄어쓰기 등을 확인한다. ▪ 언어 선택: 문장의 맥락과 목적에 맞는 언어(공손한 말, 반말, 전문 용어 등)를 선택하여야 한다. ▪ 시제 일치: 문장 내 시제가 일관되게 유지되어야 한다.
기타 고려 사항	▪ 윤리 준수: 문장 내용에 비윤리적 내용 및 혐오 표현이 포함되어 있지 않아야 한다.

	▪ 외래어 선택 기준: 외래어 및 로마자, 한자 정제 내용이 포함된 경우 국립국어원 어문 규범에 맞는지 확인한다. ▪ 특수문자: 맥락상 불필요한 특수문자, 구두점(마침표와 쉼표), 이모티콘은 사용하지 않는다.
수어 번역 고려 사항	▪ 작성된 문장은 수어 전문가의 조언을 통해 한국어와 한국수어 간 차이를 고려하여 번역 가능 여부를 파악한다.

[표 Ⅱ-3] 한국어 문장 작성 기준

○ 작성된 문장은 식별자 및 속성자 기준을 참고하여 개인정보 비식별화를 실시하였다.

구분	식별자	속성자
지침	개인의 구분을 위하여 부여된 고유한 값 또는 이름을 비식별화	개인을 특정할 수 있는 상황인지 판단하여 비식별화
항목	▪ 고유 식별 정보(주민등록번호, 운전면허증 번호 등) ▪ 성명(한글, 한문, 영문, 필명 포함) ▪ 상세 주소(구 단위 미만까지 포함) ▪ 이메일, 홈페이지 URL 등 주소 ▪ 생일, 기념일 등 날짜 정보 ▪ 각종 자격증 번호 ▪ 통장 계좌 번호 ▪ 각종 식별 코드(아이디, 사원 번호, 고객 번호 등) ▪ 전화 및 팩스 번호 ▪ 의료 보험, 기록 관련 번호 및 복지 수급자 번호 ▪ 각종 비밀번호, 쿠폰 번호, 파일명	▪ 성별, 연령, 국적, 고향, 우편번호, 병역 여부, 결혼 여부, 종교, 취미, 동호회, 클럽 ▪ 혈액형, 신장, 체중, 허리둘레, 혈압, 눈동자 색깔, 흡연 및 음주 여부, 채식 여부 ▪ 세금 납부액, 신용 등급, 기부금, 건강 보험료 납부액, 소득 분위, 의료 급여자 등 ▪ 학교명, 학과, 학년, 성적, 학력 등 ▪ 경력, 직업, 직종, 직장명, 부서명, 직급

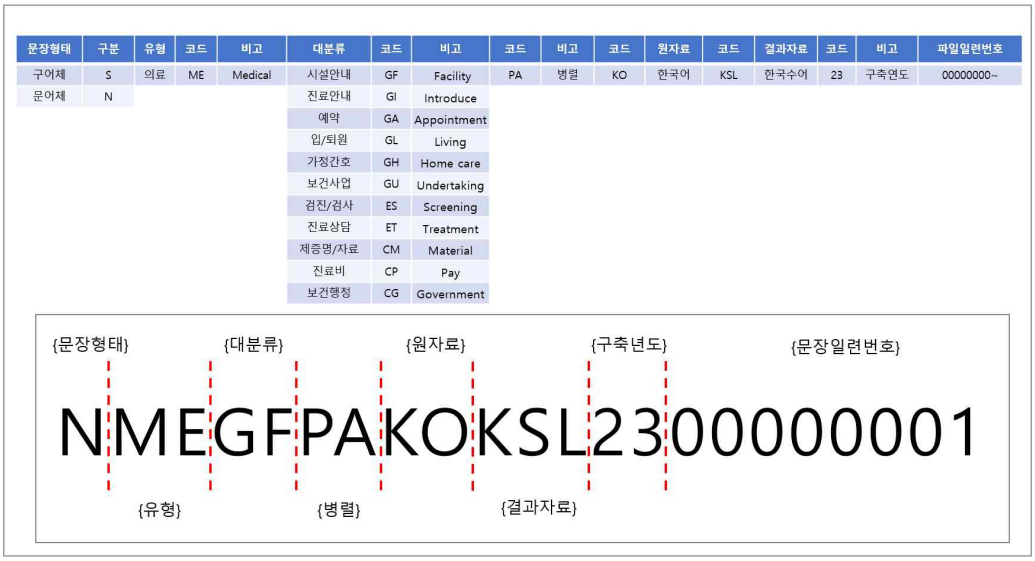
[표 Ⅱ-4] 개인정보 비식별화 지침

(3) 파일 고유 코드 부여

○ 데이터 파일 고유 코드는 검수가 완료된 한국어 문장에 대하여 관리자가 부여하며 이후 촬영한 한국수어 영상도 동일하게 명명하는 것을 원칙으로 하였다.

○ 의료 분야 주제에 대한 데이터 파일 코드는 다음과 같이 구성하며 문장

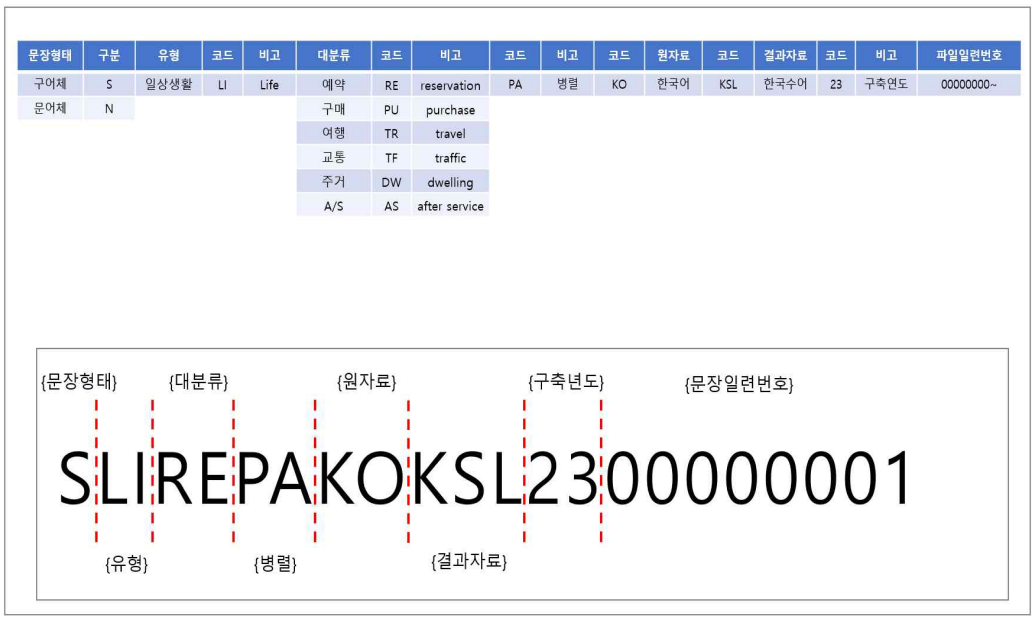
형태 구분은 구어체와 문어체로 나누었고 유형 및 대분류, 병렬 말뭉치, 원데이터, 결과데이터, 구축년도와 일련번호를 구분하여 표시하였다.



문장형태	구분	유형	코드	비고	대분류	코드	비고	코드	비고	코드	원자료	코드	결과자료	코드	비고	파일일련번호
구어체	S	의료	ME	Medical	시설안내	GF	Facility	PA	병렬	KO	한국어	KSL	한국수어	23	구축연도	00000000~
문어체	N				진료안내	GI	Introduce									
					예약	GA	Appointment									
					입/퇴원	GL	Living									
					가정간호	GH	Home care									
					보건사업	GU	Undertaking									
					검진/검사	ES	Screening									
					진료상담	ET	Treatment									
					제증명/자료	CM	Material									
					진료비	CP	Pay									
					보건행정	CG	Government									

[그림 Ⅱ-3] 의료 분야 데이터 파일 고유 코드 구성

○ 일상생활 분야의 주제에 대한 데이터 파일 코드는 다음과 같이 구성하였다.

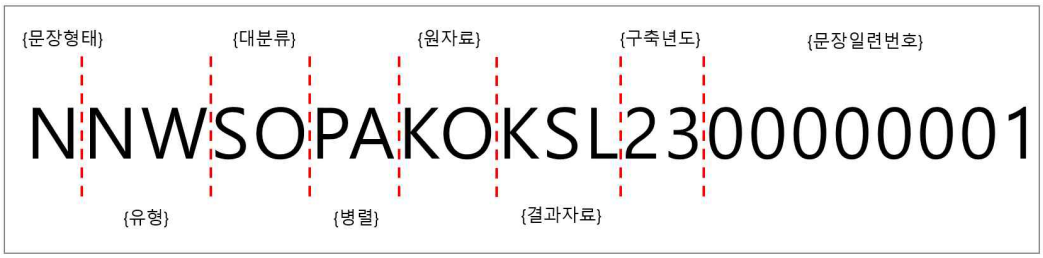


문장형태	구분	유형	코드	비고	대분류	코드	비고	코드	비고	코드	원자료	코드	결과자료	코드	비고	파일일련번호
구어체	S	일상생활	LI	Life	예약	RE	reservation	PA	병렬	KO	한국어	KSL	한국수어	23	구축연도	00000000~
문어체	N				구매	PU	purchase									
					여행	TR	travel									
					교통	TF	traffic									
					주거	DW	dwelling									
					A/S	AS	after service									

[그림 Ⅱ-4] 일상생활 분야 데이터 파일 고유 코드 구성

ㅇ 뉴스 분야 주제에 대한 데이터 파일 코드는 다음과 같이 구성한다.

문장형태	구분	유형	코드	비고	대분류	코드	비고	코드	비고	코드	원자료	코드	결과자료	코드	비고	파일일련번호
구어체	S	뉴스	NW	News	사회	SO	Social	PA	병렬	KO	한국어	KSL	한국수어	23	구축연도	00000000~
문어체	N				문화	CU	Culture									
					국제	IN	International									
					지역	LO	Local									



[그림 Ⅱ-5] 뉴스 분야 데이터 파일 고유 코드 구성

(4) 저작권 검토 및 저작재산권

ㅇ 원천 자료에 대한 저작권 확보 방안

- 한국수어 병렬 말뭉치 데이터 구축을 수행한 데이터 수집 작업자를 대상으로 직접 계약을 체결하며, 데이터 이용 허락 동의를 얻어 저작권을 확보하였다.
- “저작권 이용 허락 동의서” 등을 산출 문서로 제출하였다.

ㅇ 구축 자료에 대한 저작권 확보 방안

- 한국수어 병렬 말뭉치 데이터 구축과 관련한 모든 산출물의 소유권은 국립국어원에 속하며, 저작권에 대한 권리는 국립국어원과 원저작물의 저작자가 공동으로 소유함을 원칙으로 한다.

ㅇ 저작권 관련 법적 검토 확인

- 저작권법에서 보호하고 있는 저작물은 인간의 사상이나 감정을 표현한 창작물을 의미한다(저작권법 제2조 제1호).
- 따라서 한국어 수집 및 가공 참여자들이 창작하는 문장도 그 안에 저작권법상 보호하는 ‘인간의 사상이나 감정’의 ‘창작적 표현’을 저작물로 판단하여 보호할 수 있다.
- 위 저작물성이 인정되는 해당 문장을 사용하고자 할 경우, 저작권자와 저작재산권 이용 허락 또는 양도에 관한 별도의 계약을 체결하여 해

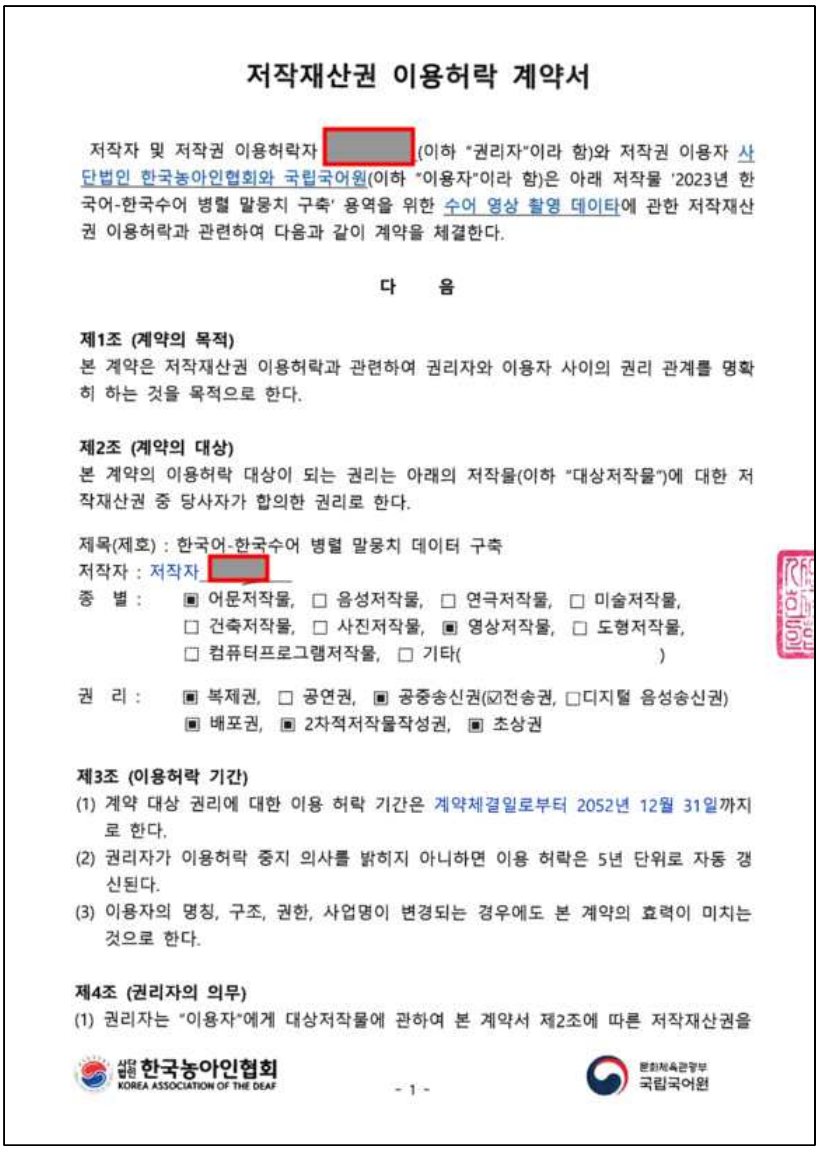
당 저작물 사용에 대한 적법한 권리 또는 권한을 취득하였다.

ㅇ 자유 배포가 가능하도록 이용 허락 계약 체결(비용 처리 포함)

- 한국어 수집 및 가공 참여자를 대상으로 한 이용 약관 동의 절차 및 현금 보상을 통하여 데이터 저작권을 확보하였다.

ㅇ 저작재산권 이용 허락 계약서 체결

- 저작재산권 이용 허락 계약서는 참여자, 수행 기관, 국립국어원의 다자간 계약으로 체결하며 이용 허락 기간은 계약일로부터 30년으로 하며 중지 의사를 밝히지 않으면 5년 단위로 자동 갱신한다.



저작재산권 이용허락 계약서

저작자 및 저작권 이용허락자 [redacted](이하 "권리자"이라 함)와 저작권 이용자 사단법인 한국농아인협회와 국립국어원(이하 "이용자"이라 함)은 아래 저작물 '2023년 한국어-한국수어 병렬 말뭉치 구축' 용역을 위한 수어 영상 촬영 데이타에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

제1조 (계약의 목적)
본 계약은 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리 관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)
본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상저작물")에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

제목(제호) : 한국어-한국수어 병렬 말뭉치 데이터 구축
저작자 : 저작자_[redacted]
종 별 : ▣ 어문저작물, □ 음성저작물, □ 연극저작물, □ 미술저작물, □ 건축저작물, □ 사진저작물, ▣ 영상저작물, □ 도형저작물, □ 컴퓨터프로그램저작물, □ 기타( )

권 리 : ▣ 복제권, □ 공연권, ▣ 공중송신권(☑전송권, □디지털 음성송신권) ▣ 배포권, ▣ 2차적저작물작성권, ▣ 초상권

제3조 (이용허락 기간)
(1) 계약 대상 권리에 대한 이용 허락 기간은 계약체결일로부터 2052년 12월 31일까지로 한다.
(2) 권리자가 이용허락 중지 의사를 밝히지 아니하면 이용 허락은 5년 단위로 자동 갱신된다.
(3) 이용자의 명칭, 구조, 권한, 사업명이 변경되는 경우에도 본 계약의 효력이 미치는 것으로 한다.

제4조 (권리자의 의무)
(1) 권리자는 "이용자"에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을

사단법인 한국농아인협회 KOREA ASSOCIATION OF THE DEAF - 1 - 문화체육관광부 국립국어원

[그림 II-6] 저작재산권 이용 허락 계약서 견본

3) 한국수어 변환 및 수어 제공

(1) 한국수어 변환 기본 원칙

- ㅇ 한국어 문장을 직역하여 의미 왜곡이 발생할 수 있는 수지 한국어 형식의 번역은 피하고 자연스러운 한국수어로 변환한다.
- ㅇ 원문을 정확히 이해하여 충실성과 등가성을 높일 수 있는 내용으로 번역한다.
- ㅇ 정확한 어휘를 선택하여 번역한다.
- ㅇ 한국어와 한국수어의 의미와 용법이 서로 다른 점을 참고하여 번역한다.
- ㅇ 청인과 농인의 문화적 차이를 고려하여 번역한다.

(2) 한국수어 번역 지침

- ㅇ 농인 사회에서 보편적으로 사용하고 직관적으로 이해할 수 있는 수어로 번역한다.
- ㅇ 한국어 단어에 대응하는 한국수어가 없으면 의미상으로 가장 가까운 수어로 번역한다.
- ㅇ 상위어, 전문 용어, 고유 명사 등 해당 수어가 없으면, 지문자로 번역하거나 ‘음차역’을 활용하고, 필요한 경우 추가 정보를 덧붙여 번역할 수 있다.
- ㅇ 의미를 나타내는 제스처를 사용하여 번역할 수 있다.
- ㅇ 한국어의 높임 표현 등은 비수지 기호로 표현하여 번역한다.
- ㅇ 국가명은 한국수어 또는 농인 사회에서 사용되고 있는 해당 국가의 수어로 번역한다.
- ㅇ 여러 개의 수어가 있는 지명은 다양하게 반영하여 번역한다.
- ㅇ 영어 단어의 경우 농인 사회에서 쓰이고 있는 외국 수어를 사용하여 번역할 수 있다.
- ㅇ 합성어의 음운적 변화를 잘 반영하여 번역한다.
- ㅇ 한국어 단어에 대응하는 한국수어가 없거나 의미를 정확히 나타낼 수 없는 경우에는 최소한으로 지화를 사용하되 문장 내 지화의 사용 횟수가 많을 경우 번역을 하지 않을 수 있다.
- ㅇ 수어로 번역하기 어려운 문학적인 표현 등은 번역하지 않는다.
- ㅇ 자격 및 역량이 검증된 농인들을 한국수어 제공자로 선정한다.

(3) 수어 영상 현장 감수 지침

- ㅇ 한국수어 현장 감수자는 한국수어 제공자가 한국수어의 다양한 특성을 반영하여 번역하도록 감수한다.
- ㅇ 현장 감수자가 필요하다고 판단하는 경우 한국어 문장을 일부 수정하여 번역할 수 있다.
- ㅇ 자격 및 역량이 검증된 수어통역사를 감수자로 선정한다.
- ㅇ 한국수어 현장 감수자는 다음과 같은 기준으로 감수를 진행한다.

한국수어 현장 검수 기준	
1	수어 제공자는 배경과 명확히 구분되는 검은색 복장을 준수하였는가?
2	수어 제공자의 머리 모양은 비수지 표현이 명확히 보이도록 관리하였는가?
3	수어의 표현이 카메라 앵글 안에서 모두 표현되었는가?
4	수지 표현이 반대 손 또는 팔꿈치 등에 가려지지 않고 명확하게 촬영되었는가?
5	비수지 표현이 적절하게 이루어졌는가?
6	수어 번역문은 변경, 왜곡되지 않고 표현되었는가?
7	한국수어의 문법이 잘 구현되고 있는가?
8	농인들이 보편적으로 사용하고 직관적으로 이해할 수 있는 수어 표현을 사용하였는가?
9	지시 수어가 명확하게 표현되었는가?
10	필요에 따라 공간동사, 일치동사, 분류사를 활용하였는가?
11	지문자 사용이 적절하게 이루어졌는가?
12	불필요한 동작 또는 비수지 표현이 포함되었는가?
13	발화 속도가 적절하였는가?

[표 II-5] 한국수어 현장 검수 기준

2. 수어 영상 촬영

1) 수어 영상 촬영 기본 원칙

ㅇ 수어 영상은 인공지능 학습데이터의 원시 자료로 다양하게 활용될 수 있으므로 최적의 데이터 획득 환경을 구성하도록 하였다.

ㅇ 수어 제공자의 수화 표현 범위를 특정하고 인체 특징점 좌표 추출을 용이하게 할 수 있도록 획득된 데이터는 잔상 없이 선명하게 확보하였다.

ㅇ 촬영 환경(스튜디오 환경) 및 촬영 장비 세팅 값을 일관되게 유지하여 데이터에 주는 영향을 최소화하였다.

ㅇ 데이터의 일관성을 위해 피사체인 수어 제공자의 복장은 단정하고 일정하게 유지하고, 머리 모양은 비수지 데이터 획득을 위해 얼굴을 가리지 않도록 하였다.

2) 촬영 환경 및 촬영 방식

ㅇ 카메라 앵글은 수어 표현 범위를 고려하여 다음과 같이 구성하였다.

[그림 Ⅱ-7] 수어 표현 범위를 고려한 촬영 화면 구성

ㅇ 촬영 환경 구성

- 촬영 앵글은 수어 제공자의 신장 및 체구를 바탕으로 모델의 키의 -20cm를 기준으로 하되, 모델의 손 길이에 따라 세부 조정을 진행하며 줌 기능은 사용하지 않는다.

- 수어 제공자의 위치는 크로마키 앞 60cm, 카메라 거리는 1.6m로 하고, 총 3개 이상의 지속광 조명을 배치한다.
- 조명은 손 아래 그림자 방지를 위해 모델 정면 1m 하단에 1개, 좌우 그림자 방지를 위해 좌우 측에 각 1개씩 배치한다.
- 좌・우측 조명은 1.9m, 모델의 45도 측면에 배치한다.

ㅇ 촬영 방식

No	속성	세팅 값
1	fps	30
2	셔터스피드	1/60
3	ISO(Gain)	250
4	조리개값	4.0
5	화이트밸런스	5,500k
6	조명밝기	58,000(LUX 1M) x 3

[표 Ⅱ-6] 촬영 장비 세팅 값

ㅇ 촬영 환경 관리 및 통제
- 영상 수집 관리자는 다음과 같이 촬영 환경 정보를 획득하고 기록한다.

No	속성	비고
1	수어 제공자 정보	이름, 지역, 우세 손
2	촬영 날짜	
3	촬영 장소	
4	촬영 담당자	
5	카메라 세팅 값	

[표 Ⅱ-7] 촬영 환경 정보 값

- 영상 수집 관리자는 촬영 전 다음과 같이 환경 구성을 점검한다.

수어 영상 촬영환경 구성	
1	촬영 장비의 fps, 셔터스피드, ISO, 화이트밸런스, 해상도가 사전에 설정된 대로 변경 없이 촬영
2	카메라 높이는 1.4m를 유지
3	수어 제공자는 배경 중앙에서 0.4m 거리를 두고 위치
4	카메라 렌즈와 수어 제공자 거리는 1.6m를 유지
5	지속광 조명을 사용하여 플리커 현상 방지
6	액터에게 그림자가 생기지 않도록 3개의 조명 배치

[표 Ⅱ-8] 수어 영상 스튜디오 촬영 환경

3) 데이터 정제 방안

- ㅇ 인코딩 과정 없이 휴지 구간 편집 기능을 지원하는 소프트웨어(어도비 프리미어 등)를 사용하며 스튜디오별로 영상 편집이 가능한 PC 또는 노트북을 배치하여 영상 촬영 종료 직후 문장 단위로 진행하였다.
- ㅇ 수어 영상은 문장 단위로 분절하고 수어 구현 전후 휴지 구간은 0.5~2초로 일괄 편집하였다.
- ㅇ 해상도는 1920*1080p 이상, Bit-rate는 2M 이상, 확장자는 MP4로 하였다.
- ㅇ 편집이 완료된 수어 영상은 해당 한국어 문장에 부여된 고유 코드와 동일하게 코드를 부여하였다.
- ㅇ 정제 결과물을 실시간으로 확인하여 재촬영 등 사후 처리가 필요한 영상을 정리하였다.

3. 주석 입력

1) 주석 기본 원칙

ㅇ 데이터 구축을 위한 주석은 SLAT(Sign Language Annotation Tool) 시스템을 주석 도구로 사용하였다.

[그림 Ⅱ-8] 주석 층렬 1

ㅇ 본 연구의 주석은 의미를 가지는 수지 요소(manuals)와 비수지 요소(non-manuals)에 이름을 붙이는 것을 기본으로 하였다.

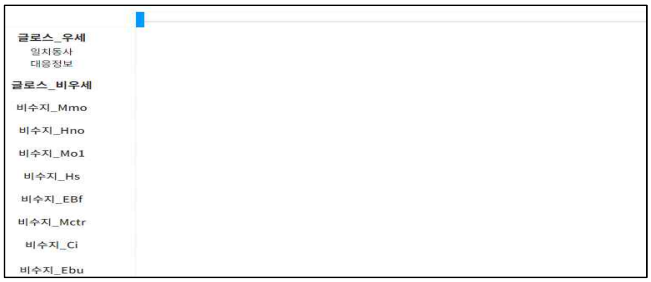
ㅇ 층렬은 크게 수지 층렬과 비수지 층렬로 나뉜다. 주석은 이 층렬에 수어의 시작점과 마지막 지점을 분절하여 토큰을 입력하였다. 이때 토큰은 타입 목록에 있는 타입명을 확인하여 입력하였다.

ㅇ 전사자가 새로운 타입을 발견했을 경우 타입 관리자에게 요청하였고, 타입 관리자는 타입 지침과 SLAT 시스템에 있는 타입 목록을 확인해 생성 여부를 결정하였다.

2) 주석 세부 지침

(1) 주석 층렬

ㅇ 본 연구의 주석 층렬은 다음 [그림 Ⅱ-9]와 같이 구성한다.



[그림 Ⅱ-9] 주석 층렬 2

ㅇ 비수지기호 층렬은 수지 층렬 바로 밑에 두고 동일한 방식으로 분절하였다.

ㅇ ‘글로스_우세’와 ‘글로스 비우세’는 수어 토큰을 입력하였다. 언어 제공자가 오른손잡이일 경우, ‘글로스_우세’에는 오른손 수어를, ‘글로스_비우세’에는 왼손 수어를 입력하였다.

ㅇ 2022년 한국어-한국수어 병렬 말뭉치 연구에서 출현 빈도가 높은 비수지기호를 기준으로 층렬을 나누었으며, 비수지기호 층렬은 [표 Ⅱ-9]와 같다.

순번	약어	라벨(용어 정리)	사용 예시
1	Mo1	입 벌리기(마!, 파!)	끝, 가능, 문장 종결, 부정 등
2	Mmo	마우딩(발음)	지문(숫)자만
3	Mctr	입꼬리가 움직이며 입 꾹 다물기	참다, 기다리다, 열심히 하다, 안녕하세요 등
4	Hno	고개 끄덕임	안녕하세요, 부탁, 필요, 나열, 문장 종결 등
5	Hs	고개를 좌우로 흔들기	불가능, 못하다, 안되다
6	Ebf	눈썹 찌푸리기	심하다, 위험하다, 안 돼
7	Ci	볼 부풀리기	상황 묘사
8	Ebu	눈썹을 위로 올림	의문문, 놀람, 강조

[표 Ⅱ-9] 비수지 표지 층렬

(2) 분절

ㅇ 주석을 입력할 때 다음과 같이 분절한다.

ㅇ 움직임이 없는 수어의 시작점은 수형(손의 모양)이 처음으로 나타나는 때로 한다.

[그림 II-10] 주석 분절 기준

ㅇ 움직임이 있는 수어의 시작점은 손이 움직이기 시작하는 때로 한다.

[그림 II-11] 수어 분절 예시

ㅇ 손의 움직임, 손 모양의 변화 없이 이전 수어에서 새로운 수어가 나타나는 경우는 손의 방향(수향)이 변할 때가 시작점이다.

[그림 II-12] 수어 시작점 입력 예시

○ 모든 수어의 마지막 지점은 손의 모양이 변하기 시작하는 직전으로 한다.

[그림 Ⅱ-13] 수어 마지막 지점 입력 예시 1

○ 이동을 시작하기 직전의 순간이 이전 수어의 마지막 지점이다.
○ 양손을 사용한 수어에서 두 손의 움직임이 동시에 일어나지 않는다면, 각 손을 따로 분절한다.

(4) 글로스

○ 글로스의 기본 원칙은 동일한 형태의 수어에 하나의 이름을 붙이는 것이다. 수어가 여러 의미를 지닐 경우 가장 직관적인 어휘를 사용한다.
※ 예: 통역, 소개, 변호 → [소개1]

○ 다의어인 경우 타입 태그에 해당 의미들을 포함시켜 검색 시 해당 수어를 쉽게 찾을 수 있도록 한다.

[그림 Ⅱ-14] 수어 마지막 지점 입력 예시 2

o 핵심 의미가 같고 형태가 다른 수어를 구분하기 위해 [한글+숫자] 형태를 사용하였다. 기본형을 활용해 다수를 나타내는 의미로 사용하거나, 의미가 확장된 수어를 구분하기 위해 [한글+숫자+#] 글로스 형태로 주석한다.

o 글로스는 고정된 수어, 생산적 수어, 제스처를 생성한다.

o 수지 요소(manuals)가 비수지 요소(non-manuals)와 결합하여 다른 의미로 변할 수 있으나 글로스_우세, 글로스_비우세 층렬에서는 수지 요소 정보만 입력한다. 즉, 비수지기호 변화로 인한 의미 변화는 글로스_우세, 글로스_비우세 층렬에서는 고려되지 않으며 형태 중심으로 주석을 입력한다.

o 한국어와 영어의 지문자 입력은 SLAT 시스템에서 ‘지화’를 선택하고, 띄어쓰기 없이 한글로 입력한다.

[그림 II-15] 지화 입력 화면

o 지숫자 입력은 SLAT 시스템에서 ‘숫자’를 선택해 입력한다.

[그림 II-16] 지숫자 입력 화면

o 숫자 표현은 SLAT 시스템에서 ‘수표현’을 선택하고 시, 시간, 날짜, 날짜(활용), 나이, 층, 번째, 등수, 몇 박 며칠, 몇분의 몇, 더하기 중 하나를 선택하여 입력한다.

[그림 II-17] 수표현 입력 화면

○ 지시(Pointing)에 대한 주석은 언어 제공자의 동작에 따라 다르게 입력하였다. 자신을 지칭할 때는 수형 1지를 [지시1], 9형을 [지시2]로 입력한다.

○ 지시(Pointing)에서 타인이나 다수, 수어를 가리킬 때는 1지를 [지시1#], 9형을 [지시2#]으로 입력한다.

※ 예: '남자 수어를 1지로 가리킴' 글로스_비우세손 [남자1], 글로스_우세손 [지시1#]

○ 지시(Pointing)는 변이형을 구분하지 않고 하나로 전사한다.

※ 예: 자신을 지칭할 때 '육형'으로 가리킨다면 [지시1]로 전사한다.

○ 사람의 신체 부위를 지칭할 때는 지시(Pointing)로 입력하지 않고, [목1], [머리1]로 입력한다.

○ 새로 발견되는 타입의 경우, 아래와 같은 방법으로 촬영하여 SLAT 시스템에 추가해 전사자들이 글로스를 입력할 수 있도록 한다.

- 배경의 색상: 파란색
- 영상화질: 같은 해상도로 통일(1280 x 720)
- 영상포맷: 같은 포맷으로 통일(mp4)
- 촬영 후 편집 시 음성 제거

[그림 II-18] 새 타입 촬영 예시

(1) 고정된 수어

- ㅇ 고정된 수어를 결정하는 기준은 다음과 같다.
- ㅇ 두 명 이상의 언어 제공자가 동일한 의미와 형태로 수어를 사용하면, 그것을 고정된 수어로 판단한다.
- ㅇ 국립국어원 한국수어사전에 수록된 어휘는 고정된 수어로 간주하였다.
- ㅇ 문맥 없이 이해할 수 있는 수어는 고정된 수어이다.
- ㅇ 지역어나 신조어도 위의 기준 중 하나 이상에 해당하면 고정된 수어로 본다.
- ㅇ 지문자어도 사회적으로 형태와 의미가 약속된 경우이므로, 고정된 수어로 본다.
- ㅇ 언어 제공자가 외국 수어를 사용한 경우, 실제 사용하는 어휘인지 확인 후 생성한다.
- ㅇ 도구 사용이나 신체 활동을 모방하는 표현은 도상성을 잘 나타내므로, 단순한 제스처가 아니라 고정된 어휘로 분류하였다. 예를 들어, ‘달리다’를 나타내는 [달리다4] 같은 수어가 이에 해당한다.

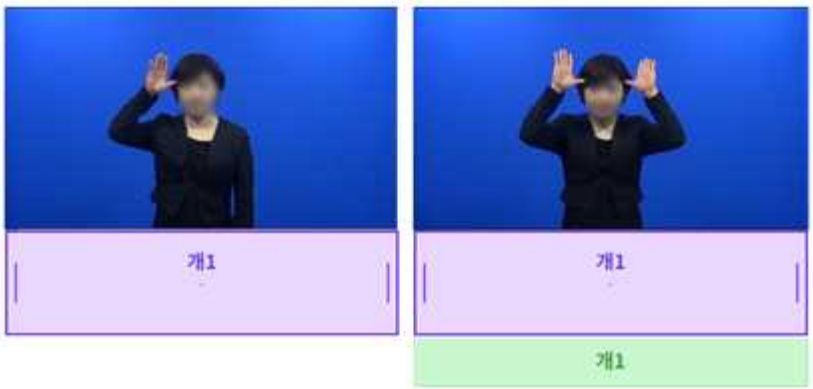
가) 변이형

- ㅇ 변이형이란 의미는 같지만 형태가 다른 수어를 말한다.
- ㅇ 수형(손 모양), 수동(동작), 수위(위치) 변이형은 모두 [한글+숫자] 형태로 전사하였다. 예시는 아래와 같다.

<table>
<tr><th>분류</th><th colspan="2">예시</th></tr>
<tr><td rowspan="3">수형
변이형</td><td>  </td><td>  </td></tr>
<tr><td>[없다1]</td><td>[없다8]</td></tr>
<tr><td colspan="2">손 모양만 다르고 의미가 동일</td></tr>
<tr><td rowspan="3">수동
변이형</td><td>  </td><td>  </td></tr>
<tr><td>[농1]</td><td>[농4]</td></tr>
<tr><td colspan="2">수어 동작만 다르고 의미가 동일</td></tr>
<tr><td rowspan="3">수위
변이형</td><td>  </td><td>  </td></tr>
<tr><td>[창피하다1]</td><td>[창피하다3]</td></tr>
<tr><td colspan="2">수어 위치만 다르고 의미가 동일</td></tr>
</table>

[그림 Ⅱ-19] 변이형 타입명 예시

○ 한 손이나 양손 변이형은 각각 한 층렬과 두 층렬에 전사된다. 이런 변이형을 위해 별도의 타입을 생성하지 않는다.



[그림 Ⅱ-20] 한 손, 양손 수어 변이형 주석 예시

○ 양손 수어에서 같은 우세손을 사용하는 경우, 수어를 다음과 같이 전사한다. [대부분2], [대부분3] 모두 우세손은 동일하나 비우세손이 다르다. [대부분2]에서 비우세손은 접혀진 형태이며, [대부분3]에서는 비우세손이 펴진 형태이다. 이 경우, ‘대부분’ 수어의 전사는 우세손이 동일하고 가장 낮은 번호인 [대부분2]로 전사한다.

[그림 Ⅱ-21] 한 손 수어 주석 예시

○ 한 손 수어와 양손 수어가 의미가 다를 경우, 변이형이 아니므로 별도로 타입을 생성한다.
※ 예: 같다1(양손 수어, A와 B가 같다), 같다4(한 손 수어, 맞다/동의/그렇다)

○ 기저 이미지는 같은 어원을 공유하는 수어를 의미한다. 의미는 같지만 기저 이미지와 형태가 다를 경우, [한글+숫자] 형태로 새로운 타입을 생성한다. 예를 들어, [강원도1]과 [강원도2]는 모두 ‘강원도’를 나타

내지만, [강원도1]은 산과 강 같은 지역적 특성을 나타내고, [강원도2]는 지역적 특산물을 나타낸다. 따라서 이들은 기저 이미지와 형태가 서로 다르므로, 숫자로 구별하여 별도의 타입으로 분류한다.

나) 변형

ㅇ 수어의 변형은 기본형에서 형태, 위치가 변하거나 움직임이 추가되어 의미가 확장되는 것을 말한다. 변형을 판단하는 기준은 다음과 같다.

- 기본형 수어의 핵심 의미가 유지되어야 한다.
- 수어의 형태가 바뀌어도 기본 의미에 추가적인 내용이 포함된 경우에는 이를 변형이 아닌 새로운 의미의 타입으로 간주한다. 예를 들어, ‘가난하다’의 의미가 추가된 [가난3]은 [집1]의 형태가 변형된 것이지만, 새로운 의미를 포함하므로 별도의 타입으로 생성한다.

[가난3]

[그림 Ⅱ-22] 변형 기준에 부합하지 않는 예시

- 수어의 기본 형태가 바뀌고 의미가 반대로 해석되면 이를 변형으로 간주하지 않는다. 예를 들어, [신경2]는 ‘신경’이나 ‘연결’을 나타내지만, 형태가 바뀌면 ‘연결되지 않음’을 의미한다. 이처럼 기본 의미에서 벗어나는 새로운 의미가 추가될 경우 별도의 타입을 생성한다.

	
기본형	독립된 어휘
신경2	[연결안되다1]

[그림 Ⅱ-23] 기본형과 변형 기준에 부합하지 않는 예시

- 수형이 동일하더라도 위치가 변하면서 의미가 달라지는 경우, 이는 변형이 아니라 별도 의미를 가지므로, 각각 다른 타입으로 구분하여 생성한다. 예를 들어, ‘켜다’를 나타내는 [켜다1]은 ‘전등이나 전깃불을 켜다’의 의미를 가지고 있고, [켜다3]은 ‘헤드라이트나 스피커를 켜다’의 의미로 사용된다.

켜다1(전등, 전깃불)

켜다3(헤드라이트)

[그림 Ⅱ-24] 변형 기준에 부합하지 않는 예시

ㅇ 일치동사는 수향이 변하거나 수어가 이동하면서 주어와 목적어가 바뀌는 형태를 의미한다. 이 경우, 수어의 기본 의미는 유지되지만, 수향의 변화에 따라 의미가 확장된다고 볼 수 있다. SLAT 시스템에서는 기본형 수어를 선택한 후, 아래 [표 Ⅱ-10]에 따라 입력하였다.

※ 예: [돕다]는 손바닥의 방향에 따라 의미가 달라질 수 있다.

인칭	주체	예시
1	1인칭(나)	① **내**가 **너**를 도와주다. **1** **2**
2	2인칭(너)	② **네**가 **나**를 도와주다. **2** **1**
3	3인칭(제3자)	③ **그**가 **당신**을 도와주다. **3** **2**

[표 Ⅱ-10] 일치동사 입력 기준

[그림 Ⅱ-25] 일치동사 입력 예시

ㅇ 복수 표현: 동일한 수형을 기준으로 [한글+숫자+#] 형태로 전사한다.

※ 예: [질문1](한 사람), [질문1#](다수)

※ 예: [곳1](한 곳), [곳1#](여러 곳)

ㅇ 강약 표현: 수지 요소(manuals)로 강약을 표현하여 의미가 강화된 경우, [한글+숫자+#] 형태로 생성한다.

※ 예: [비1](일반적), [비1#](강한 비)

[비1]

[비1#]

[그림 Ⅱ-26] 강약 표현 예시

ㅇ 수어 이동 변형: 수어의 위치 및 움직임이 변형되어 의미가 확장된 경우, [한글+숫자+#] 형태로 전사한다.

※ 예: [승용차1](명사), [승용차1#](차가 좌회전함, 언덕길을 가는 중 등)

※ 예: [금이가다1](동사), [금이가다1#](벽에 금이가다 등)

[금이가다1]

[금이가다1#]

[그림 Ⅱ-27] 수어 이동 변형 타입 예시

ㅇ 분류사: 기본적인 수형의 의미가 고정되어 있으며, 추가적인 움직임이나 공간 활용으로 의미가 확장되는 경우를 말한다. 예를 들어, '앉다'의 기본 의미를 가진 [앉다1] 수형에 원을 그리듯이 움직임이 더해지면, 이는 '의자를 원형으로 배치하고 앉다'라는 새로운 의미로 확장된다. 이렇게 움직임을 통해 의미가 확장된 경우, 해당 수형은 [한글+숫자+#] 형태로 생성한다.

- 기본 수형을 기준으로 [한글+숫자+#]을 생성하였다. 예를 들어, 고정된 의미를 가진 [앉다1] 수형을 활용했다면 이를 기준으로 [앉다1#]을 생성한다.

[앉다1]


[앉다1#]

[그림 Ⅱ-28] 분류사 타입 생성 예시

- 숫자와 관련된 분류사는 다양한 숫자 조합으로 인해 상당히 많은 변형이 가능하다. 이러한 다양성으로 인해, 각 숫자를 독립된 수형으로 처리하는 것은 실용적이지 않다. 따라서, 이 경우에는 숫자 자체의 수형보다는 수동을 주요 기준으로 삼아 새로운 타입을 생성한다. 예를 들어, 숫자 '30'과 [더1]의 수동이 결합된 '30분 추가하다'라는 수어가 있을 때, 이는 [더1#]으로 생성한다.


[더1]


[더1#]

[그림 Ⅱ-29] 숫자 분류사 타입 생성 예시

다) 형성

ㅇ 형성은 결합을 통해 새로운 수어 어휘가 형성되는 것을 의미한다. 형성의 유형에는 다음이 포함된다.

ㅇ 수 포합어는 지숫자 수형이 결합하여 형성된 수어를 의미한다. 여기에는 연도, 나이, 날짜, 층, 번째 등이 포함된다. 이런 어휘는 SLAT 시스템에서 '수표현'을 선택하여 입력한다.

※ 주의: 초등학교, 중학교, 고등학교, 대학교 학년 관련 글로스는 변이형이 많기에 별도의 타입으로 주석을 입력하였다.

[그림 II-30] 수 포합어 입력 예시

○ 지문자어: 지문자어는 한글 지문자와 결합된 수어나 활용한 수어를 의미한다. [오늘1]이라는 수어가 'ㅈ'과 결합하여 [지금]으로 사용되거나 한국어 'ㅍ', 'ㅊ'을 차용하여 형성된 [평창1]을 예시로 들 수 있다. 이 경우, 독립된 어휘로 보고 타입을 생성한다.
※ 예시: [카톡1], [고흥1]

라) 복합어

○ 동시적 결합 구조: 두 가지 이상의 요소가 동시에 작용하여 새로운 의미의 합성어를 형성하는 경우를 말한다. 이 구조에서 양손은 동일한 타입으로 취급되어 입력된다.
※ 예: '등산'의 경우, 양손 모두 [등산1]로 입력되는 것이 올바르다(글로스_비우세손 [등산1], 글로스_우세손 [등산1]). 반면, 글로스_비우세손 [산1]과 글로스_우세손 [걷다1]로 분리하여 입력하는 것은 적절하지 않다.

- 한 손이 생산적 수어와 결합된 동시적 결합 구조의 경우, 각 요소를 별도로 전사한다.
※ 예: [그림 31]은 글로스_비우세손 [주먹형1]과 글로스_우세손 [뿌리다1]로 입력해야 한다.

{바람을 주입하다}

[그림 II-31] 동시적 결합 구조 예시

ㅇ 계기적 결합 구조: 두 개 이상의 형태소가 순차적으로 결합하여 새로운 합성어를 형성한다.

- 형태소의 움직임이 줄어들거나 반복되지 않은 경우, 각 형태소를 따로 전사하였다. 예를 들어, [계란1]은 C형으로 손끝을 오므리고 펴는 동작을 하며, [일하다1]은 손바닥을 위로 하고 양손을 옆으로 붙였다 벌리는 동작을 반복한다. 하지만 [아르바이트]와 같이 두 수어가 순차적으로 결합한 경우, [계란1]의 움직임이 반복되지 않는다. 이런 경우에는 [계란1], [일하다1]로 각각 전사한다.

  ※ 예: 공무원은 [공공1+일하다1+사람1]로 전사

- 두 수어가 결합하여 전혀 새로운 의미를 생성하는 경우, 이를 하나의 독립된 타입으로 분류하였다. 이러한 결합이 없으면 새로운 의미가 형성되지 않는다.

  ※ 예: ‘쓰다듬다’와 [덕분1] 수어가 결합해 ‘자기애’라는 새로운 의미를 가지므로, [자기애1]로 타입을 생성한다.

[자기애1]

[그림 Ⅱ-32] 계기적 결합 구조 예시

- 합성어를 구성하는 두 요소 사이에 부드러운 전환 움직임(transitional movements)이 있는 경우, 이를 하나의 타입으로 전사한다. 예를 들어, [알다1]은 가슴을 한 손으로 두 번 쓸어내리는 동작으로 이루어지고, [주다1]은 손바닥을 내밀며 수행된다. 그러나 [알려주다1]과 같은 합성어에서는 이 두 동작이 연속적으로 이루어지기 때문에, 이러한 경우에는 별도로 전사하기 어려우므로 하나의 타입 [알려주다1]로 전사한다.

  ※ 예: ‘보여주다’는 [보다6+주다1]이 아니라 [보여주다1]로 생성하였다.

- 동시 조음 현상은 두 수어가 연결되어 첫 번째 수어의 끝부분과 두 번째 수어의 시작 부분이 구분되기 어려울 정도로 변화하는 현상이

다. 이 현상은 대부분 합성어에서 나타난다. 예를 들어, 국립국어원 타입 [일요일1]은 [빨강1]과 [결석1]의 합성어로, [빨강1]이 뒤이은 '닫다'의 영향을 받아 수형이 같아진다. 이런 경우, [결석1]의 시작점이 명확하지 않기 때문에, 하나의 독립된 타입으로 명명하여 생성한다.

- 나라명, 지명, 기관명 등과 같은 고유 명사는 국립국어원 타입에 따라 하나의 타입으로 생성하였다. 이러한 방식은 통일성을 유지하고, 고유 명사의 특정한 수어 형태를 명확히 하기 위함이다.
  ※ 예: [경상북도1], [소비자보호원1], [국세청1]

ㅇ 계기적 파생어: 어근과 접사가 결합되어 새로운 의미가 형성되는 사례로 이 역시 하나의 타입으로 생성한다.
※ 예: [한적없다1]은 {손털기} 접사가 결합된 수어이기 때문에, 하나의 타입으로 생성하였다.

(2) 생산적 수어

ㅇ 생산적 수어란 관습적인 의미가 없으며 맥락에 따라 해석이 달라지는 수어를 의미한다. 이러한 수어를 판단하는 기준은 아래와 같으며, 다음 중 한 가지 이상 해당되면 '생산적 수어'로 간주한다.

- 농인마다 표현 방식이 다양하며, 고정되어 있지 않다.
- 문장이나 맥락 없이 해당 수어를 단독으로 봤을 때 여러 해석이 가능하다.
- 사전에 등재되지 않았거나 아직 어휘화되지 않은 수어이다.
- 묘사나 제스처에 기반을 두고, 예측 가능한 방식으로 생성된다.
  ※ 예: '파이프관'의 두께를 표현하기 위해 [구형1] 형태로 문장에서 사용한다. 이 수형은 특정 맥락에 따라 다른 의미로 해석될 수 있다.

ㅇ 수형을 기준으로 타입명을 생성하며, 본 연구에서는 수형의 구부림, 접촉 여부, 시작점과 마지막 지점 등을 고려하지 않고 수형에 따라 전사한다.

기존 형태	수형	타입명
주먹형1		주먹형1
일형2, 일형3, 일형3_네모, 일형3_세모, 일형3_원		일형3
이형1, 이형1A		이형1
육형1, 육형1_네모, 육형1_세모, 육형2, 육형2_네모, 육형3		육형1
구형1, 구형1_면, 구형1_선, 구형1_네모, 구형1A_면, 구형1A_선, 구형2, 구형2_원, 구형3, 구형4, 구형4A,		구형1
십형1		십형1
X		삼형1
X		사형1
X		사람형1
X		칠형1
X		여우형1
X		십칠형1
X		아이형1
X		오형1
X		십이형1

[표 Ⅱ-11] 생산적 수어 타입명

	
치아가 밀착되어 있는 형태를 표현	뾰족한 송곳니 표현
우세손 [사형1], 비우세손 [사형1]	우세손 [여우형1]

[그림 Ⅱ-33] 생산적 수어와 타입명 입력 예시

(3) 한국어 동음이의어

ㅇ 한국어의 동음이의어 글로스 형태는 [한글+숫자]이다.

※ 예: 배1(타는 배), 배6(먹는 배)

(4) 기타

ㅇ 지문자, 지숫자의 형태를 가지나, 고정된 수어로 굳어진 경우 타입으로 생성한다.

※ 예: 지문자 비우세손 [무], 우세손 [무] X / 수어 [없다11], [없다11]

ㅇ 청인도 알아볼 수 있는 ‘제스처’도 인공지능 학습을 위해 타입을 만들어 전사한다. 또한 인공지능 학습이 목적이므로, 변이형 등을 고려하지 않는다. 예를 들어, ‘스트레칭’은 언어 제공자마다 손목을 돌려서 표현하거나 어깨를 움직이는 등으로 다양하게 표현한다. 이 동작이 스트레칭의 의미를 가지고 있다는 것을 인공지능에 학습시키는 것이 목표이기에 타입 [스트레칭1]을 만들고 관련 제스처를 하나로 전사한다.

※ 예: [로봇1]

ㅇ [counting hand]는 가족 혹은 숫자를 표현할 때 사용하는 수형이며, 문맥상 의미가 달라지는 생산적 수어인 [구형1]과 구분한다.

ㅇ 외국 수어는 국립국어원의 타입을 확인한 후 그대로 전사한다. 그러나 국립국어원 타입에도 없으면, 온라인 다국어 수어 사전(Spreadthesign)을 확인하거나 언어 제공자 및 자문위원에게 확인해 생성한다.

ㅇ [어림없다1]과 같이 독립적으로 사용되지 않고 일관되게 함께 사용

하는 수어는 하나의 단위로 전사한다. 즉, 의미가 고정적인 조합에서만 형성되는 경우, 이를 하나의 타입으로 생성한다.

※ 예: [어림없다1] O, [어림1][없다] X

※ 예: [~줄알다] O, [~줄][알다] X

- ㅇ 유명인의 수어 이름(Sign Name)의 경우, 타입을 만든다.
- ㅇ 어휘의 수동이 반대가 되며 반의어가 되는 경우, 타입을 분류하여 생성한다.

※ 예: [밝다1], [어둡다1]

※ 예: [온도오르다1], [온도내리다1]

3) 타입 관리

(1) 타입 관리 범위

- ㅇ ‘타입’은 번역이 아닌 수어에 부여된 고유 식별명을 의미한다. 타입 관리는 SLAT 시스템 ‘타입요청 관리’의 타입 목록을 기반으로 한다.
- ㅇ 2023년 한국어-한국수어 병렬 말뭉치 구축을 위해 촬영된 영상에서 발견된 새로운 타입을 생성하고 관리했으며, 국립국어원에서 제공된 타입도 관리하였다.

(2) 새 타입 생성 절차

- ㅇ 새 타입 생성 절차는 [표 II-12]와 같다.

순서	과정	내용
1	타입 문의	전사자 타입 문의
▽		
2	영상 확인	타입 문의한 영상 검토 타입 지침 및 목록 확인 필요시 자문 요청
▽		
3	타입 생성	타입 영상 촬영 및 SLAT 시스템에 업로드
▽		
4	타입명 안내	전사자에게 새 타입명 안내

[표 II-12] 새 타입 생성 절차

- 타입 관리자는 전사자의 타입 문의를 받고, 관련 영상을 검토하여 타입 목록과 지침에 따라 새 타입 생성 필요성을 판단한다.
- 외국 수어 등의 이유로 타입 생성 여부가 모호할 경우, 자문회의를 통해 결정한다.

자문 요청	쿠알라룸푸르 외국 수어가 맞는가?
자문 답변	쿠알라룸푸르 외국 수어가 맞다.

[그림 Ⅱ-34] 자문 예시

- 새로운 타입 생성이 필요할 경우, 지침에 따라 촬영한다.
- SLAT 시스템의 '주석타입 관리'에서 타입명과 영상을 업로드하여 타입 목록에 추가한다.

[그림 Ⅱ-35] 타입 업로드 예시

- 타입 목록에 추가되면, 전사자가 해당 타입명을 입력할 수 있다. 전사자에게 새로운 타입명 생성을 안내하여 일관된 주석 입력을 유도한다.

(3) 국립국어원 타입

ㅇ 국립국어원에서 제공된 3,364개의 타입을 '미사용'으로 설정한다.

ㅇ 전사자가 문의한 타입 중 국립국어원 타입명이 포함된 경우, 검토 후 '사용'으로 변경하여 주석 입력이 가능하도록 하였다. 이 과정

은 언어 제공자의 수어와 국립국어원 타입이 일치하는지, 그리고 동형이의어의 존재 여부를 확인하기 위함이다.

- ‘사용’으로 변경된 국립국어원 타입은 1,949개이다.

(4) 타입명 변경

- 국립국어원 타입명이 주석 타입 지침과 일치하지 않은 경우, 해당 타입명을 변경한다.

수정 전	수정 후
돈주다1	돈1#

[그림 II-36] 지침 불일치 시 타입명 변경 예시

- 타입명을 외래어 표기법에 따라 수정한다.

수정 전	수정 후
할로윈1	핼로윈1

[그림 II-37] 외래어 타입명 변경 예시

- 동형이의어가 발견되었을 경우, 빈도수가 높은 타입명으로 변경한다. 예를 들어, [교토1]로 타입명을 생성했으나 ‘서쪽’이라는 의미로 더 자주 사용되어 [서쪽2]로 변경하였다.

수정 전	수정 후
교토1	서쪽2

[그림 II-38] 동형이의어 타입명 변경 예시

(5) 타입 검토

- ㅇ 타입 목록을 검토하는 과정에서 타입 오류를 발견할 때가 있다. 오류로는 수어의 형태가 같으나 타입명이 두 개인 경우가 있다.
- ㅇ 동일한 형태의 수어에는 단일 타입명이 있어야 하며, 두 개일 경우 더 많이 사용된 타입명으로 통일하였다. 예를 들어, [흐르다2]와 [강1]이 동일한 수어로 확인되었을 때, [흐르다2]로 입력된 수가 더 많아 시스템에서 [강1]로 주석된 것을 [흐르다2]로 일괄 변경하였으며, 주석 입력 관리 '메모'에 이를 기록하였다.

[그림 Ⅱ-39] 타입 오류 기록 예시

4. JSON파일 구축

1) 병렬 말뭉치 데이터 구조

[그림 II-40] 한국어-한국수어 병렬 말뭉치 구축 사업 절차

○ 한국어-한국수어 병렬 말뭉치 사업은 수어를 영상 기반으로 인식하여 한국어로 번역(pose to text)하는 응용 서비스의 개발을 위해 이에 필요한 학습데이터를 구축하는 것을 목표로 하고 있다. 이를 위해 한국어 문장, 수어 영상, 형태소 단위 어노테이션 정보(수어 주석, 비수지 표현) 및 메타정보를 구성하며 각 정보를 동영상 단위 시간 정보(프레임 단위)와 연계하였다.

2) 데이터 포맷 개요

○ 한국어-한국수어 병렬 말뭉치는 약 100만 개의 수어 형태소 및 주석 정보, 10만 개의 학습 데이터셋을 생성, 해당 데이터셋은 영상과 함께 json 파일로 제공되며 포맷 종류 및 영상의 명명 규칙은 다음과 같다.

한국어-한국수어 병렬말뭉치 구축사업
(Json 구조정의서)

```
{
 "id":"SMEETPAKOKSL2300006326",                    작업 아이디 (파일명)
 "opertor":"label2",                               작업자
 "krlgg_sntenc":{
  "koreanText":"정말감사합니다,의사선생님.가능한한빨리건강하게회복하고싶어요.",      한국어문
  "realm":"의료",                                  (분야)
  "thema":"진료상담",                              (대분류)
  "detailThema":"진료상담"                         (세분류)
 },
 "sign_script":{                                   수어주석 정보
  "sign_gestures_strong":[{                        글로스_왼손
   "start":1.199,                                  (시작시간)
   "end":1.377,                                    (종료시간)
   "gloss_id":"맞다2",                             (주석명)
   "express":"s",                                  (주석타입 s:수어 f: 지화 n: 숫자  d: 동적숫자)
                                                   (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있음)
   "direction":{                                   (일치동사)
    "source":"",                                   에서(인칭)
    "target":""                                    으로(인칭)
   },
   "sentence_loc":{                                (대응정보)
    "start":"",                                    (시작)
    "end":""                                       (종료)
   }
  },{
   "start":1.633,                                  (시작시간)
   "end":1.885,                                    (종료시간)
   "gloss_id":"고맙다1",                           (주석명)
   "express":"s",                                  (주석타입 s:수어 f: 지화 n:  숫자  d: 동적숫자)
                                                   (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있음)
   "direction":{                                   (일치동사)
    "source":"",                                   에서(인칭)
    "target":""                                    으로(인칭)
   },
   "sentence_loc":{                                (대응정보)
```

```
      "start":"1",                    (시작)
      "end":"1"                       (종료)
    }
  },{
    "start":2.999,                    (시작)
    "end":3.332,                      (종료)
    "gloss_id":"빨리1",
    "express":"s",                    (주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
                                      (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있음)
    "direction":{                     (일치동사)
      "source":"",
      "target":""                     으로(인칭)
    },
    "sentence_loc":{                  (대응정보)
      "start":"6",                    (시작)
      "end":"6"                       (종료)
    }
  },{
    "start":3.633,                    (시작)
    "end":3.879,                      (종료)
    "gloss_id":"건강1",               (주석명)
    "express":"s",                    (주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
                                      (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어
있음)
    "direction":{                     (일치동사)
      "source":"",                    에서(인칭)
      "target":""                     으로(인칭)
    },
    "sentence_loc":{                  (대응정보)
      "start":"7",                    (시작)
      "end":"7"                       (종료)
    }
  },{
    "start":4.133,                    (시작)
    "end":4.751,                      (종료)
    "gloss_id":"회복1",               (주석명)
    "express":"s",                    (주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
                                      (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어
있음)
    "direction":{                     (일치동사)
      "source":"",                    에서(인칭)
      "target":""                     으로(인칭)
    },
    "sentence_loc":{                  (대응정보)
      "start":"8",                    (시작)
      "end":"8"                       (종료)
    }
  },{
    "start":5.055,                    (시작)
    "end":5.654,                      (종료)
```

```
    "gloss_id":"원하다1",                (주석명)
    "express":"s",                       (주석타입 s:수어 f: 지화 n:  숫자  d: 동적숫자)
                                         (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어
있음)
    "direction":{                        (일치동사)
      "source":"",                       에서(인칭)
      "target":""                        에서(인칭)
    },
    "sentence_loc":{                     (대응정보)
      "start":"9",                       (시작)
      "end":"9"                          (종료)
    }
  },{
    "start":2.266,                       (시작)
    "end":2.605,                         (종료)
    "gloss_id":"부르다1",                (주석명)
    "express":"s",                       (주석타입 s:수어 f: 지화 n:  숫자  d: 동적숫자)
                                         (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어
있음)
    "direction":{                        (일치동사)
      "source":"",                       에서(인칭)
      "target":""                        으로(인칭)
    },
    "sentence_loc":{                     (대응정보)
      "start":"",                        (시작)
      "end":""                           (종료)
    }
  }],
  "sign_gestures_weak":[{                글로스_오른손
    "start":1.633,                       (시작)
    "end":1.885,                         (종료)
    "gloss_id":"고맙다1",                (주석명)
    "express":"s",                       (주석타입 s:수어 f: 지화 n:  숫자  d: 동적숫자)
                                         (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어
있음)
    "direction":null,                    (일치동사)
    "sentence_loc":null                  (대응정보)
  },{
    "start":3.633,                       (시작)
    "end":3.879,                         (종료)
    "gloss_id":"건강1",                  (주석명)
    "express":"s",                       (주석타입 s:수어 f: 지화 n:  숫자  d: 동적숫자)
                                         (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어
있음)
    "direction":null,                    (일치동사)
    "sentence_loc":null                  (대응정보)
  },{
    "start":4.133,                       (시작)
    "end":4.751,                         (종료)
    "gloss_id":"회복1",                  (주석명)
    "express":"s",                       (주석타입 s:수어 f: 지화 n:  숫자  d: 동적숫자)
```

(동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있음)

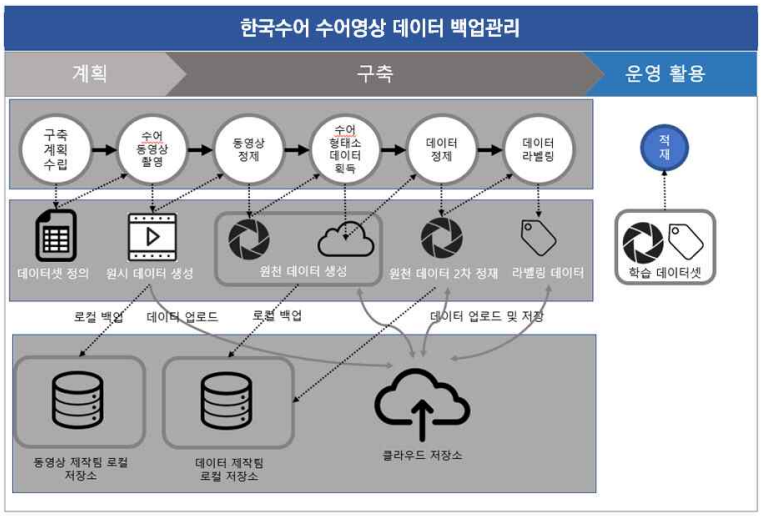
```
    "direction":null,
    "sentence_loc":null
  },{
    "start":5.055,                    (시작)
    "end":5.654,                      (종료)
    "gloss_id":"원하다1",              (주석명)
    "express":"s",                    (주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
                                      (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있음)
    "direction":null,                 (일치동사)
    "sentence_loc":null               (대응정보)
  }]
 },
 "nms_script":{                       비수지
  "Mmo":null,                         Mmo: 단어가 입모양으로 나타나는 것
  "Hno":[{                            Hno: 고개 끄덕임
    "descriptor":"",                  (설명)
    "start":1.633,                    (시작시간)
    "end":1.885                       (종료시간)
  },{
    "descriptor":"",                  (설명)
    "start":5.055,                    (시작시간)
    "end":5.654                       (종료시간)
  }],
  "Mo1":null,                         Mo1: 마우딩과 다르게 특정상황에 입모양이 동반됨
  "Hs":null,                          Hs: 고객를 좌우로 흔드는 것
  "EBf":[{                            Ebf: 눈썹을 찌푸리는 것
    "descriptor":"",
    "start":2.999,                    (시작)
    "end":3.332                       (종료)
  }],
  "Mctr":[{                           Mctr: 입꼬리가 올라가면서 입을 다무는 것
    "descriptor":"",
    "start":1.633,                    (시작)
    "end":1.885                       (종료)
  },{
    "descriptor":"",                  (일치동사)
    "start":4.133,                    (시작)
    "end":5.642                       (종료)
  }],
  "Ci":null,                          Ci: 볼을 부풀리는 것
  "Ebu":null                          Ebu: 눈썹을 위로 올리는 것
 },
 "vido_file_nm":"SMEETPAKOKSL2300006326",  영상파일명
 "potogrf":{                          촬영정보
  "createdTime":"2023-09-1301:45:46",
  "sentence_ID":"SMEETPAKOKSL2300006326",
  "photographer":"JINILGEUN",
```

```
  "location":"Seoul,Korea",
  "devide":"SonyA6400",
  "iris":4.0,
  "gain":"ISO250",
  "whiteBalance":5500,
  "shutterSpeed":"1/60",
  "fps":30.0,
  "width":1920,
  "height":1080,
  "format":"MP4",
  "codec":"XAVCSHD",
  "sl_speaker_id":"유ㅇㅇ",
  "sl_speaker_age":40,
  "sl_speaker_sex":"남",
  "sl_speaker_legion":"ㅇㅇ시",
  "sl_speaker_hand":"오른손"
 },
 "license":"국립국어원"
}
```


[그림 Ⅱ-41] 데이터 포맷 종류 및 영상 명명 규칙

3) 데이터 백업 관리

ㅇ 원천 데이터 및 라벨링 데이터의 훼손 및 멸실을 방지하기 위해 단계별 백업을 진행하였다.



[그림 Ⅱ-42] 데이터 백업 관리 과정

ㅇ 모든 데이터는 제작팀에 의한 1차 로컬 백업을 원칙으로 하였다.

ㅇ 1차 로컬 백업 데이터는 백업 완료 후 프로젝트 종료 시점까지 삭제 불가하며 클라우드 파일 서버 이관 데이터에 문제 발생 시 1차 로컬 백업 파일은 다시 업로드하였다.

ㅇ 데이터 이관은 클라우드 파일 서버를 이용하였다.

ㅇ 모든 데이터는 로컬과 클라우드 서버에 이중 보관하였다.

5. 병렬 말뭉치 구축 데이터 검수

1) 한국어 데이터 검수

○ 2023년 한국어 기준 구축 데이터의 세부 분류는 다음과 같다.

분야	대분류	세부 예시 주제	구축범위
의료	시설 안내	• 오시는 길, 진찰실 안내, 병원 내 시설 안내	20%
	진료 안내	• 증상에 따른 진료 과목 안내, 진료 과목의 설명	
	예약	• 병원 예약 문의, 예약 절차, 예약 확인	
	입/퇴원	• 입원/퇴원의 절차 및 안내	
	가정간호	• 가정간호 제도 안내, 신청 방법	
	보건사업	• 각종 보건 제도 안내, 보건소 이용 안내	
	검진/검사	• 검진 및 검사 방법 안내	
	진료 상담	• 각종 질병에 대한 문의 및 답변	
	제 증명/자료	• 각종 증명서 안내 및 발급 방법	
	진료비	• 진료비 청구, 수납, 건강보험 처리 절차	
	보건 행정	• 보건소 이용 시 행정 처리 안내	
일상생활	예약	• 숙박, 식당, 문화시설 등 예약 상황	70%
	구매	• 백화점, 편의점, 대형마트 등 구매 상황	
	여행	• 유명 관광지 안내, 여행 시 및주의사항	
	교통	• 대중교통 이용, 사고 접수 및 처리 방법	
	주거	• 주거복지, 매매, 전월세 계약, 시설 관리	
	A/S	• 전자 제품, 가구 등 A/S 문의 및 답변	
뉴스	사회	• 사건 사고, 인권/복지 등 사회 뉴스	10%
	문화	• 공연/전시, 여행/레저, 패션/뷰티 등 문화 뉴스	
	국제	• 국제 상황, 세계 소식 등 국제 뉴스	
	지역	• 지역 이슈, 행사, 정책 등 지역 뉴스	

[표 Ⅱ-13] 수집 대상 세부 분야

ㅇ 한국어 검수자는 한국어 전문가로서 한국어 수집 및 정제가 완료된 한국어에 대하여 다음과 같은 기준으로 검수를 진행하였다.

한국어 검수 기준	
1	오타, 띄어쓰기, 문장의 맥락 및 구조 적합성 여부
2	주제 분류 적합성 여부
3	개인정보 비식별화 여부
4	비윤리적 내용 및 혐오 표현 포함 여부
5	외래어 및 로마자, 한자 정제 내용은 어문 규범에 맞는지 확인
6	수어 번역에 적합한 어절로 문장 구성 여부 (4~15어절 사이, 평균 9.5어절)
7	파일 고유 코드 부여 정확성 여부

[표 Ⅱ-14] 한국어 문장 검수 기준

ㅇ 한국어 데이터 검수 방법은 다음과 같다.
- 한국어 정제팀으로부터 한국어 엑셀 데이터 획득
- 한국어 검수 기준(표 Ⅱ-14)에 따라 데이터 검수
- 오류 사항 체크 후 정제팀 피드백
- 오류 사항에 따라 수정 및 삭제 후 재검수

2) 한국수어 영상 검수

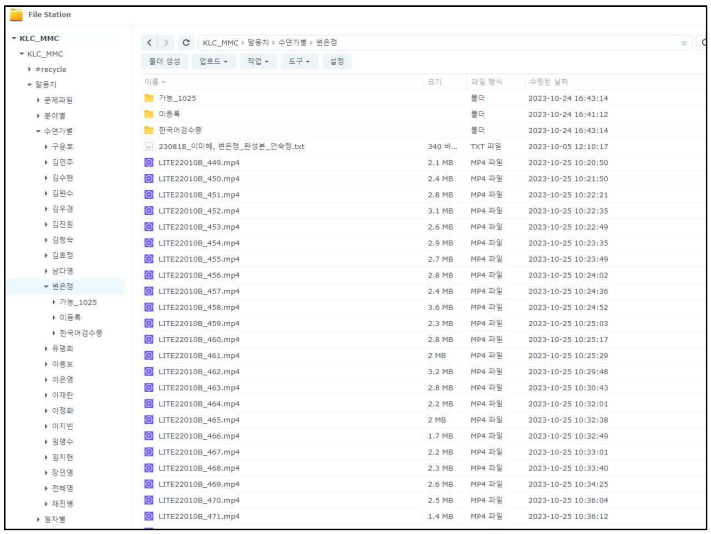
○ 한국수어 영상 검수자는 한국농아인협회에서 자격 및 역량이 검증된 전문가로서 문장 단위로 촬영된 수영 영상에 대하여 다음과 같은 기준으로 검수를 진행하였다.

한국수어 영상 검수 기준	
1	수어 제공자는 배경과 명확히 구분 가능한 검은색 복장을 준수하였는가?
2	수어 제공자는 비수지 표현이 명확히 보이는 머리 모양을 준수하였는가?
3	수어의 표현이 카메라 앵글 안에서 모두 표현되었는가?
4	수지 표현이 반대 손 또는 팔꿈치 등에 가려지지 않고 명확하게 촬영되었는가?
5	비수지 표현이 적절하게 이루어졌는가?
6	수어 번역문은 변경, 왜곡되지 않고 표현되었는가?
7	한국수어의 문법이 잘 구현되고 있는가?
8	농인들이 보편적으로 사용하고 직관적으로 이해할 수 있는 수어 표현을 사용하였는가?
9	지시 수어가 명확하게 표현되었는가?
10	필요에 따라 공간동사, 일치동사, 분류사를 활용하였는가?
11	지문자 사용이 적절하게 이루어졌는가?
12	불필요한 동작 또는 비수지 표현이 포함되었는가?
13	발화 속도가 적절하였는가?

[표 Ⅱ-15] 한국수어 영상 검수 기준

○ 한국수어 영상 데이터 검수 방법은 다음과 같다.

- 파일 관리 프로그램(NAS)에서 최종 편집된 한국수어 영상 및 한국어 엑셀 획득
- 한국수어 영상 검수 기준(표 Ⅱ-15)에 따라 데이터 검수
- 오류 사항 체크 후 현장 감수자에게 피드백
- 오류 사항에 따라 삭제 및 수정 촬영하여 재검수



[그림 II-43] 한국수어 영상 파일관리 시스템 화면

2) 주석 검수

- ㅇ 주석 검수자는 청인일 경우 수어통역사 자격증을, 농인일 경우 수어통역사 자격을 소지한 전문가여야 한다.
- ㅇ 주석 검수는 한국수어 영상 주석 프로그램을 사용하였다.

[그림 II-44] 한국수어 영상 주석 프로그램 검수 화면

- ㅇ 주석 검수자는 주석 파일의 최종 승인 권한을 가지고 있으며 단순히 주석 입력에 대한 평가 이외에도 한국어, 한국수어 영상의 정확성까지 판단하

여 검수하였다.

ㅇ 주석 검수의 기준은 다음과 같다.

순번	내용	세부내용
1	주석 검수	1) 토큰과 분절(2페이지) 2) 글로스 정보 3) 비수지 정확성
2	기타 검수	1) 메모 확인 2) 기타 이외의 발생 오류들

[표 Ⅱ-16] 주석 검수 기준

ㅇ 주석 검수 방법은 다음과 같다.

- 주석 프로그램에서 라벨링이 제출된 상태의 파일을 검수
- 주석 검수 기준(표 Ⅱ-16)에 따라 데이터 검수
- 오류 사항 발생 시 ‘메모’ 기능을 활용하여 오류 내용 기입 후 반려 후 재작업

ㅇ 한국수어 저작도구 검수 시 활용 기능은 다음과 같다.

기능명	기능 내용
보류하기	특정 원인으로 검수를 완료할 수 없는 상황에서 보류 처리를 해놓으면 해당 영상 목록에 남아 있어도 새로운 작업물을 배분받을 수 있다.
저장하기	검수자가 자체적으로 수정한 내용이 있다면 저장하기를 꼭 눌러야 수정한 내용이 저장된다.
승인하기	영상에 입력된 정보에 이상이 없을 경우 승인하기 버튼을 누른다. 모든 과정에 대한 최종 승인 단계이니 신중하게 확인하고 승인한다.
반려하기	작업자가 검수자에게 제출한 주석작업을 다시 작업자에게 되돌려보내는 기능이다. 작업자가 너무 성의 없이 일을 진행했거나 특정 작업자에게서 반복된 실수 등이 나온다면 반려를 진행한다.
영상 최종 반려	영상 자체에서 오류가 발견되어 라벨링 입력, 검수를 정상적으로 진행할 수 없을 경우 판단에 따라 최종 반려하며 최종 반려된 파일은 품질 관리자가 확인 후 수정 또는 삭제한다.

[표 Ⅱ-17] 한국수어 저작도구 검수 기능

6. 병렬 말뭉치 데이터 품질관리

1) 데이터 품질관리

ㅇ 품질관리 조직 구성

조직명		역할
주관기관 (국립국어원)		• 구축사업 진행 과정에서 승인내용 점검 및 의사결정 수행 • 이슈, 긴급문제 발생 시 조정 및 해결 • 업무범위 및 사업 수행 절차에 대한 의사 결정 • 구축사업 품질관리 • 변경 요청사항에 대한 검토 및 승인 • 방법론 적용·변경에 대한 검토 및 승인
수행사 (한국농아인협회 – 케이엘큐브)	PM	• 구축사업 품질 계획 수립 및 수행, 시정조치 지휘 • 진행사항과 의사결정 사항을 주관기관 및 참여자에게 보고 • 감리 사전 점검 및 감리 대응
	품질관리 담당자	• 공정별 위험·이슈 관리 • 공정별 일정 진척 관리 • 구축사업 품질활동 계획 보고 • 각 공정별 기준 및 절차를 설정/이행 확인 • 산출물 품질활동/검토 • 구축사업 품질활동에 대한 시정조치 수행
	구축사업 참여자	• 산출물 품질활동 수행 및 시정조치 수행 • 공정별 품질 검수, 외부 감리 결과 시정조치 수행
기타	외부전문가	• 외부 품질 점검 - 품질점검 전문 업체(TTA) 검토 진행 • 외부 감리 - 프로젝트 관리 및 품질활동의 적절성 확인 - 고객 요구사항의 반영 여부 확인 - 산출물 간 정합성, 일관성, 기술적인 적정성 평가

[표 Ⅱ-18] 품질관리 조직 구성

ㅇ 품질목표

구분	품질지표	개 요	품질목표
요구 사항	요구사항 달성률	정의된 요구사항이 산출물로 제시되고 있는지	95%
	시정조치 달성률	외부 전문가(감리, 품질) 시정조치 이행률	99%
구축공정 품질	준비성	분류/단계별 계획수립성 확인(체크리스트)	95%
	완전성	분류/단계별 완전성 확인(체크리스트)	95%
데이터 품질	적합성	구축 목표 대비 실적(유형별 구축수량)	100%
	정확성	데이터 정확성(한국어문, 영상, 주석) / 오류율	95%

[표 Ⅱ-19] 품질목표

ㅇ 품질점검 일정

구분	품질지표	단계			
		분석	설계	종료	완료
요구 사항	요구사항 달성률	-	-	24.01.11	2024.01.30
	시정조치 달성률	-	11.15	24.01.11	2024.01.30
구축공정 품질	준비성	-	-	24.01.11	2024.01.30
	완전성	-	-	24.01.11	2024.01.30
데이터 품질	적합성	-	-	24.01.11	2024.01.30
	정확성	-	-	24.01.11	2024.01.30

[표 Ⅱ-20] 품질점검 일정

ㅇ 품질 목표 충족 여부

구분	품질지표	실 적 세 부 내 역	품질실적
요구 사항	요구사항 달성률	측정 수식: (19 ÷ 20) × 100 (단위 %) 요구사항(상세ID 수) 20 구현 된 요구사항(상세ID 수) 19	95%
	시정조치 달성률	측정 수식: (22 ÷ 22) × 100 (단위 %) 요구사항(상세ID 수) 22 구현 된 요구사항(상세ID 수) 22	100%
구축 공정 품질	준비성	측정 수식: (41 ÷ 41) × 100 (단위 %) 체크리스트(확인란 수) 41 확인 완료 체크리스트 41	100%
	완전성	측정 수식: (20 ÷ 20) × 100 (단위 %) 체크리스트(확인란 수) 20	100%

		확인 완료 체크리스트 20	
데이터 품질	적합성	측정 수식 120 % = (115+131+120) / 3 의료 = (229,189÷200,000) × 100 뉴스 = (131,244÷100,000) × 100 일상생활 = (839,850÷700,000) × 100	100%
	정확성	측정 수식 97 % = (98+100+93) / 3 한국어문 = (2,646÷125,541) × 100 영상 = (342÷124,529) × 100 주석 = (6,778÷105,299) × 100	97%

[표 Ⅱ-21] 품질 목표 충족 여부

2) 체크리스트

ㅇ 한국어-한국수어 병렬 말뭉치 구축은 구축계획 수립, 데이터 수집, 데이터 정제, 데이터 라벨링 및 데이터 학습의 생애주기를 가지며, 지표와 생애주기 관점에서 각 품질검사 활동을 수행하였다.

ㅇ 준비성(계획수립성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
절차 준비	임무정의	발주기관(수요자)의 요구사항을 분석하였는가?	☑	-사업수행계획서 -요구사항정의서
		한국어-한국수어 병렬 말뭉치 구축 성능지표와 목표를 제시하였는가?	☑	-사업수행계획서
	구축 계획수립	한국어-한국수어 병렬 말뭉치 구축 데이터를 정의하였는가?	☑	-사업수행계획서
		한국어-한국수어 병렬 말뭉치 구축 데이터 분류체계를 정의하였는가?	☑	-사업수행계획서
	데이터 수집/정제	데이터 수집 시 미확보 데이터 수집을 위한 방안을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집/정제 결과에 대한 검수 절차를 마련하였는가?	☑	-통합 검수 지침
	영상 촬영	영상 촬영/편집에 대한 기준절차를 마련하였는가?	☑	-수어 영상 촬영 지침
		영상 촬영/편집에 대한 교육 및 훈련 계획을 수립하였는가?	☑	-수어 영상 촬영 지침 -교육계획서

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
	데이터 라벨링	데이터 라벨링 방법 및 기준을 마련하였는가?	☑	-주석 및 타입 입력 지침
		데이터 라벨링 결과에 대한 검수 절차를 마련하였는가?	☑	-통합 검수 지침
조직 준비	구축 계획수립	한국어-한국수어 병렬 말뭉치 데이터 공정별 관리자를 지정하였는가?	☑	-사업수행계획서
	데이터 수집/정제	데이터 수집/정제을 위한 인력 운영 계획을 수립하였는가?	☑	-사업수행계획서
		데이터 수집/정제를 위한 조직을 구성하고 담당자를 명시하였는가?	☑	-사업수행계획서
	영상 촬영	영상 촬영을 위한 인력 운영 계획을 수립하였는가?	☑	-사업수행계획서
		영상 촬영을 위한 조직을 구성하고 담당자를 명시하였는가?	☑	-사업수행계획서
	데이터 라벨링	데이터 라벨링을 위한 조직의 역할과 책임을 정의하였는가?	☑	-사업수행계획서
		데이터 라벨링을 위한 인력 운영 계획을 수립하였는가?	☑	-사업수행계획서
도구 준비	구축 계획수립	한국어-한국수어 병렬 말뭉치의 요구사항에 맞는 데이터 구축을 제시하고 있는가?	☑	-요구사항정의서
	데이터 수집/정제	데이터 수집/정제를 위한 기준과 훈련 계획을 수립하였는가?	☑	-한국어 수집 및 정제 지침 -교육계획서
		데이터 수집/정제를 위한 템플릿를 정의하였는가?	☑	-한국어 수집 및 정제 지침
	영상 촬영	영상 촬영을 위한 기준안을 정의하였는가?	☑	-수어 영상 촬영 지침
		영상 편집 체계 및 교육안을 정의하였는가?	☑	-수어 영상 촬영 지침
	데이터 라벨링	데이터 라벨링을 위한 작업 도구 교육 및 훈련 계획을 수립하였는가?	☑	-주석 및 타입 입력 지침
위험 관리	구축 계획수립	사업의 위험관리를 위한 계획을 수립하였는가?	☑	-이슈 및 위험관리대장
	데이터 수집/정제	데이터 수집/정제의 위험관리를 위한 계획을 수립하였는가?	☑	-한국어 수집 및 정제 지침
	영상 촬영	영상 촬영 및 편집의 위험관리를 위한 위험 요소 식별을 진행하였는가?	☑	-수어 영상 촬영 지침
	데이터 라벨링	식별된 위험에 대응하기 위한 활동을 수행하고 있는가?	☑	-이슈 및 위험관리대장

[표 Ⅱ-22] 준비성(계획 수립성) 체크리스트

ㅇ 준비성(체계 준수성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
보안 준수	구축 계획수립	한국어-한국수어 병렬 말뭉치 데이터에 대한 보안 관리체계를 마련하였는가? (관리 및 운영 규정, 환경, 접근권한 등)	☑	-위험관리 계획서
		민감정보 보호를 위한 체계를 마련하였는가? (관리 및 운영 규정, 환경, 접근권한 등)	☑	-위험관리 계획서
	데이터 수집/정제	데이터 수집/정제에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경 구성, 접근권한 관리 등)	☑	-위험관리 계획서 -보안서약서
		민감정보 보호를 위한 활동을 수행하고 있는가? (담당자 지정, 로그 관리 등)	☑	-한국어 수집 및 정제 지침 -개인정보 이용 동의서
	영상 촬영	영상 촬영에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경 구성, 접근권한 관리 등)	☑	-수어 영상 촬영 지침 -위험관리 계획서
		영상 편집에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경 구성, 접근권한 관리 등)	☑	-수어 영상 촬영 지침 -위험관리 계획서
	데이터 라벨링	데이터 라벨링에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경 구성, 접근권한 관리 등)	☑	-주석 및 타입 입력 지침 -보안서약서
법·제도 준수	구축 계획수립	한국어-한국수어 병렬 말뭉치 구축을 위한 관련 법·제도적인 검토를 위한 절차 및 해결방안을 제시하고 있는가?	☑	-사업수행계획서
		한국어-한국수어 병렬 말뭉치 구축을 위한 개인정보활용 동의 절차를 마련하고 수행하고 있는가?	☑	-개인정보이용동의서
		한국어-한국수어 병렬 말뭉치 구축을 위한 저작권 활용 동의 절차를 마련하고 수행하고 있는가?	☑	-저작재산권 이용 동의서
	데이터 수집/정제	데이터 수집/정제를 위한 관련 법·제도적인 검토를 위한 절차 및 해결방안을 제시하였는가?	☑	-사업수행계획서
		데이터 수집/정제 시 저작권 보호 대상일 경우 법에 저촉되지 않는 범위 내에서 수집할 수 있는 방안을 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	영상 촬영	영상 촬영을 위한 관련 민감정보 비식별화 조치 등 법·제도적인 검토 절차 및 해결방안을 제시하였는가?	☑	-수어 영상 촬영 지침
	데이터	데이터 라벨링을 위한 법·제도적인 검토 절차 및	☑	-주석 및 타입

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
	라벨링	해결방안을 제시하였는가?		입력 지침

[표 Ⅱ-23] 준비성(체계 준수성) 체크리스트

○ 완전성(수집 완전성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
수집 완전성	데이터 수집/정제	편향성 방지 방안을 마련하였는가?	☑	-한국어 수집 및 정제 지침
		정의된 데이터 수집/정제 방법 및 기준을 적용하고 있는가?	☑	-한국어 수집 및 정제 지침
		데이터 수집/정제 기준 변경에 대한 절차를 마련하였는가?	☑	-변경관리대장
		데이터 수집/정제 방법에 대한 교육 및 훈련을 시행하였는가?	☑	-교육계획서 -교육결과서
		데이터 수집/정제 결과에 대한 검수 절차 및 기준에 따라 수행하고 있는가?	☑	-통합 검수 지침
		데이터 수집/정제 결과에 대한 검수 기준변경 시 절차에 따라 수행하였는가?	☑	-변경관리대장
		데이터 수집/정제 결과에 대한 검수 교육 및 훈련을 시행하였는가?	☑	-교육계획서 -교육결과서

[표 Ⅱ-24] 완전성(수집 완전성) 체크리스트

ㅇ 완전성(정제 완전성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
정제 완전성	영상 촬영	영상 촬영 시 개인 정보 보호 등 비식별화를 실시하였는가?	☑	-수어 영상 촬영 지침
		정의된 영상 편집 방법 및 기준을 적용하고 있는가?	☑	-수어 영상 촬영 지침
		영상 촬영 기준변경에 대한 절차를 마련하였는가?	☑	-변경관리대장
		영상 촬영/편집 방법에 대한 교육 및 훈련을 시행하였는가?	☑	-교육계획서 -교육결과서
		영상 촬영/편집 결과에 대한 검수 절차 및 기준에 따라 수행하고 있는가?	☑	-통합 검수 지침
		영상 촬영/편집 결과에 대한 검수 기준변경 시 절차에 따라 수행하였는가?	☑	-통합 검수 지침 -변경관리대장 -회의록
		영상 촬영/편집 결과에 대한 검수 교육 및 훈련을 시행하였는가?	☑	-교육계획서 -교육결과서

[표 Ⅱ-25] 완전성(정제 완전성) 체크리스트

ㅇ 완전성(가공 완전성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
가공 완전성	데이터 수집	정의된 데이터 라벨링 방법 및 기준을 적용하고 있는가?	☑	-주석 및 타입 입력 지침
		데이터 라벨링 기준변경에 대한 절차를 마련하였는가?	☑	-변경관리대장 -주석 및 타입 입력 지침
		데이터 라벨링 방법에 대한 교육 및 훈련을 시행하였는가?	☑	-교육계획서 -교육결과서
		데이터 라벨링 결과에 대한 검수 절차 및 기준에 따라 수행하고 있는가?	☑	-통합 검수 지침 -검사결과서
		데이터 라벨링 결과에 대한 검수 기준변경 시 절차에 따라 수행하였는가?	☑	-변경관리대장 -통합 검수 지침
		데이터 라벨링 결과에 대한 검수 교육 및 훈련을 시행하였는가?	☑	-교육계획서 -교육결과서

[표 Ⅱ-26] 완전성(가공 완전성) 체크리스트

7. 보안 관리

1) 보안 관리 개요

[그림 Ⅱ-45] 보안 관리 전략

2) 원천 자료 및 구축 자료에 대한 저작권 확보

○ 원천 자료에 대한 저작권 확보

- 한국수어 병렬 말뭉치 데이터 구축을 수행한 크라우드 소싱 및 지자체 인력을 대상으로 직접 계약을 체결하며, 데이터 이용 허락 동의를 얻어 저작권을 확보하였다.
- ‘저작권 이용 허락 동의서’ 등을 별도의 산출 문서로 제출하였다.

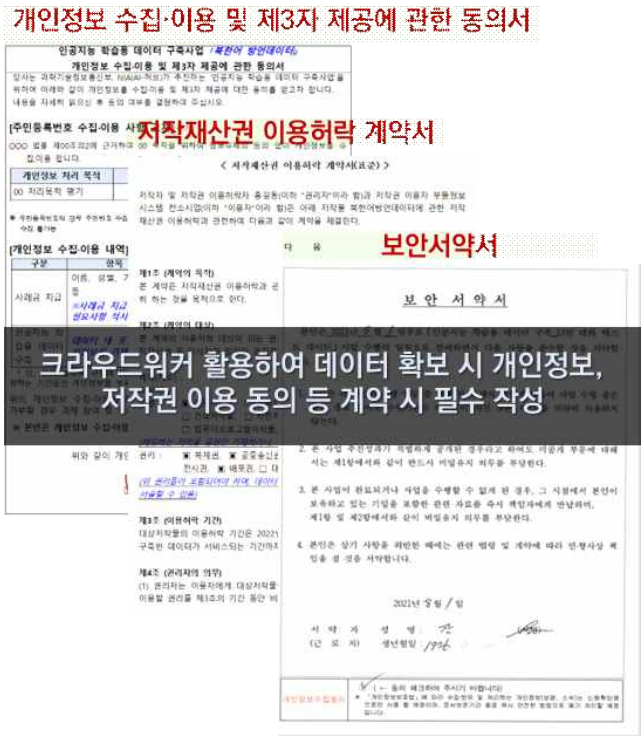
○ 구축 자료에 대한 저작권 확보

- 한국수어 병렬 말뭉치 데이터 구축과 관련한 모든 산출물의 소유권은

국립국어원에 속하며, 저작권에 대한 권리는 국립국어원과 원저작물의 저작자가 공동으로 소유함을 원칙으로 하였다.

ㅇ 자유 배포가 가능하도록 이용 허락 계약 체결(비용 처리 포함)

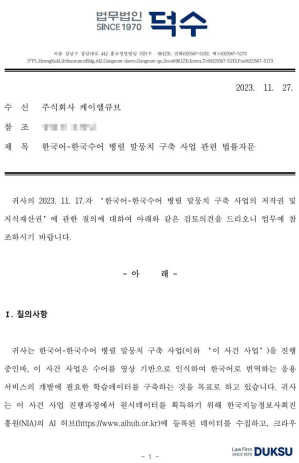
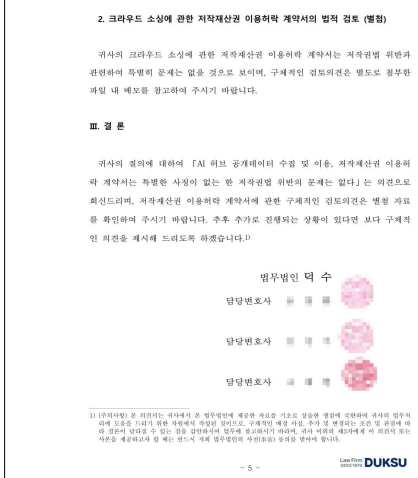
- 크라우드 워커 대상 이용약관 동의 절차 및 현금 보상을 통하여 데이터 저작권을 확보하였다.



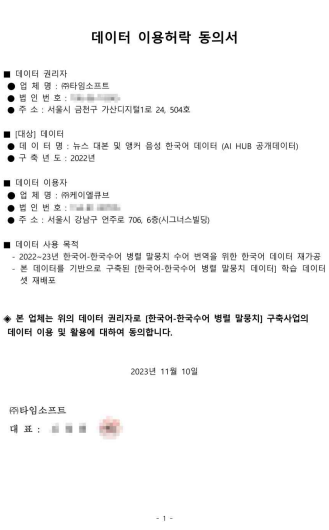
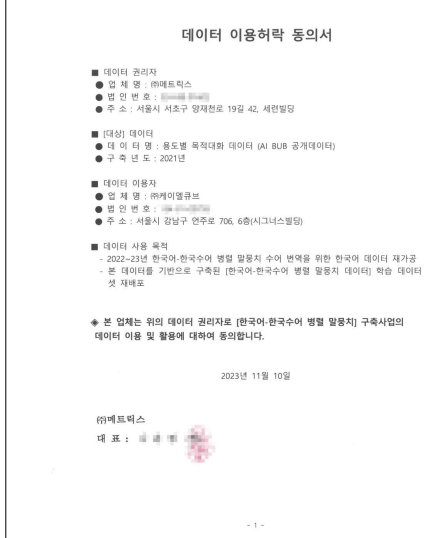
[그림 Ⅱ-46] 저작권 이용 동의서 견본

ㅇ 저작권 관련 법적 검토 확인

- 저작권법에서 보호하고 있는 저작물은 인간의 사상이나 감정을 표현한 창작물을 의미(저작권법 제2조 제1호)
- 따라서 수어 강사들이 창작하는 수어 교육용 문장들의 경우, 그 안에 저작권법상 보호하는 ‘인간의 사상이나 감정’의 ‘창작적 표현’이 포함되어 있다면 저작물로 인정되어 저작권을 보호받는다.
- 위 저작물성이 인정되는 해당 문장을 사용하고자 할 경우, 저작권자와 저작 재산권 이용 허락 또는 양도 관련 별도의 계약을 체결하여 해당 저작물 사용에 대한 적법한 권리 또는 권한을 취득하였다.

구분	저작권 법률 검토 의견서	
	첫 페이지	날인 페이지
저작권에 대한 전문기관 법률검토		
	2023년 한국어-한국수어 병렬 말뭉치 데이터 수집 법률 검토서	

[그림 II-47] 저작권 법률 검토

구분	AI-HUB 데이터 이용 동의서	
	타임소프트	메트릭스
원시데이터 활용에 따른 데이터 이용 허락 동의서		
	2023년 한국어-한국수어 병렬 말뭉치 데이터 수집 법률 검토서	

[그림 II-48] 원시데이터 활용에 따른 데이터 이용 허락 동의서

3) 개인정보보호 등 보안 정책 및 지침 준수

ㅇ 문화체육관광부 개인정보보호지침(훈령)을 준수하였다.

ㅇ 본 사업과 관련하여 개인정보를 처리하거나 제공·연계 시에는 현행 개인정보보호법에서 정하는 의무 조치 사항을 반영하여 제공하였다.

4) 사업 수행을 위한 보안 대책 수립 및 준수

ㅇ 사업 수행 중 정보 유출 등 보안 사고 발생 시 책임 및 보상

- 보조사업자는 본 사업 수행 중 취득한 정보, 원저작물, 산출물 및 관련 자료에 대하여 사업 수행 중은 물론 사업 완료 후에도 비밀 보안을 엄수하며, 이를 위반하여 문제가 발생할 경우 모든 민·형사상 책임을 진다.

ㅇ 사업 계약 단계 보안 대책 수립

- 사업 수행계획서에 자료·인원·장비·네트워크 등에 대한 물리적, 관리적, 기술적 보안 대책 및 누출금지 대상 정보 관리 방안 등 보안 관리 세부 계획을 구체화하여 수립·제공하였다.

ㅇ 참여 인원에 대한 보안 관리

- 참여 인원은 개인의 친필 서명이 들어간 보안서약서를 제출하였다.
- 사업 수행 전 참여 인원에 대해 법적 또는 주관 기관 규정에 따른 비밀 유지 의무 준수 및 위반 시 처벌 내용, 누출 금지 대상 정보 및 정보 누출 등에 대한 보안 교육을 시행했다.

ㅇ 자료에 대한 보안 관리

- 사업 수행에서 생산되는 모든 산출물은 파일 서버 또는 보안 담당관이 지정한 PC에만 저장·관리하고 사업 담당자가 인가하지 않은 비인가자에 대해서는 제공·대여·열람을 금지하였다.

5) 보안 계획 점검 사항

NO.	점 검 항 목	점검결과				비고
		양호	보통	주의	미흡	
1	보안 조직(전담조직)이 구성되어 있는가?	●				
2	보안 내규는 수립하였는가?	●				
3	보안 내규·지침 등이 참여 인력에게 배포되고 시행 및 게시되어 있는가?	●				
4	참여 인원에 대한 보안 교육 및 훈련 계획을 수립하였는가?	●				
5	사업 수행 전 참여 인원에 대한 보안교육을 실시하고 보안서약서를 체결하였는가?	●				
6	보호가 필요한 장비 및 시설에 대하여 보호구역(제한구역, 통제구역)을 지정, 관리하고 있는가?	●				
7	출입문, 회의실 등의 공간을 타 기관·업체 등과 공동으로 사용하지 않고 단독으로 사용하고 있는가?	●				
8	전용 사무실에 특허관련 서류 파기를 위한 문서세단기가 설치되어 있는가?	●				
9	침입차단시스템이 구축되어 있는가?	●				
10	인터넷 검색 PC내 자료 유출차단을 위한 보안통제는 적절하게 수행되고 있는가?	●				
11	주요 데이터에 대한 백업이 적절히 수행되고 있는가?	●				
12	주요 데이터에 대한 접근 권한 설정이 되어 있는가?	●				

[표 Ⅱ-27] 보안 점검 사항 체크리스트

8. 이용자 아카데미(전문가 워크숍) 개최

1) 배경 및 목적

- ㅇ 2023년 한국어-한국수어 병렬 말뭉치 사업 홍보
- ㅇ 전문가와의 협력 체계 구축을 통한 유용성 확보
- ㅇ 한국어-한국수어 병렬 말뭉치 사업을 통한 농인 전문가 양성
- ㅇ 한국어-한국수어 병렬 말뭉치 사업 보완 및 개선 방향 논의

2) 프로그램 및 토의 내용

- ㅇ 프로그램 일정(2023년 11월 21일)

시간	프로그램	비고
13:00~ 13:10	[인사말] 워크숍 소개 및 주제 발표, 참여자 소개	하윤호 PM (한국농아인협회)
13:10~ 13:50	[현황소개] 국/내외 수어 자동 번역 모델 연구 동향	방기덕 연구소장 (케이엘큐브)
14:00~ 14:25	[현황소개] 의료현장 양방향 수어 동시통역 서비스 기술개발	이한규 책임 연구원(한국전자통신연구소)
14:35~ 15:10	[사업소개] '22~'23 한국어-한국수어 병렬 말뭉치 구축 데이터	하윤호 PM (한국농아인협회)
15:20~ 17:20	[토론] 참여 기관 연구 또는 개발 중인 번역 모델의 특성과 필요 학습 데이터 구축의 특성과 난제, 병렬 말뭉치 학습 데이터 요구사항 제시 및 한국어-한국수어 병렬 말뭉치 학습 데이터 유용성 증진 방안 토론	진행: 하윤호 PM 토론: 각 참여 기관

[표 Ⅱ-28] 이용자 아카데미 프로그램 타임테이블구성

ㅇ 국/내외 수어 자동 번역 모델 연구 동향
- 수어 특성 및 자동 번역을 위한 기반 기술
- 국내 및 해외 수어 자동 번역 모델 연구 동향(수어 아바타, FSW)
- 자사 기술개발 방향 및 수어 자동 번역의 미래 전망

ㅇ 의료 현장 양방향 수어 동시통역 서비스 기술개발
- 연구 개요 및 영상(동작 정보) 획득안
- 수어소 및 표제어 개념
- 말뭉치 생성 및 표제어 애니메이션 제작

ㅇ ’22~’23 한국어-한국수어 병렬 말뭉치 구축 데이터
- 사업 프로세스 및 데이터 구축 현황
- 형태소/비수지 및 형태소 개념
- 수어 글로스 개념 정리(기본 정보 및 타입)

ㅇ 전문가 종합 토의 내용
- 국내에서는 기관별 번역 모델 설계 방식이 달라 업체별 기준에 딱 맞는 학습 데이터를 구축하기 어려우므로 다양한 기관들이 활용할 수 있도록 유용성 확장에 대한 기준 마련이 필요하다.
  수어 인식 기술: 연구 초기 단계이지만, 향후 데이터의 활용도를 높이는 차원에서 다방향 수어 영상 데이터 구축이 필요하다, 단, 다방향의 기준을 재정의할 필요가 있다.
  한국어-수어 변환 기술: 참석 기업 기준으로 글로스 기반의 번역 모델이 연구 개발되고 있어 병렬 말뭉치 데이터가 유용성이 있어 보이지만, 학습의 기준이 되는 글로스(타입)의 모호성 때문에 모델 연구에 어려움이 있다. 따라서 글로스(타입)에 대한 표준 체계를 수립하여야 하며 국립국어원에서 기준 또는 표준을 제시할 필요가 있다.
- 국내외적으로 번역 모델에 대해 완벽히 정의된 기준이 없을 뿐 아니라 일반적인 언어 연구보다 어려운 부분이므로 국내에서도 해외와 같이 수요 기관, 전문 기관, 연구 기관, 대학 등이 연계하여 다양한 논의를 통해 수어 번역 기준을 마련하는 수어 연구 생태계 조성이 필요하다.

- 현재 국내에서는 글로스 분석을 통한 번역 모델 연구가 이루어지고 있으나 글로스의 모호성 때문에 연구 개발에 어려움이 있다. 해외의 경우 수어의 의미를 정확히 문자로 표기하는 방법인 Hamnosys나 FSW를 활용하여 번역 기술이 어느 정도 궤도에 오른 것으로 보이며 국제적으로는 ‘SignPudle Online’을 통해 다양한 국가들의 FSW 기반 데이터 구축이 이루어지고 있으나 한국수어는 아직 해당 데이터가 거의 없다.

  브라질에서는 FSW 기반으로 구축된 번역 모델이 상용화 단계까지 이르렀고 SignWriting 표기법을 법제화하여 활용하고 있다. 이 표기법은 시각 언어인 수어를 문자로 가장 정확하게 표현하는 방법으로 국제적으로도 인정받고 있다. 이 같은 국제 동향을 고려할 때 국내에서도 FSW 기반의 수어 데이터를 구축할 필요가 있다.

<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>참석자 소개</td><td>케이엘큐브 발표 1</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>케이엘큐브 발표 2</td><td>한국전자통신연구소 발표</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>농아인협회 발표</td><td>전문가 종합 토의 1</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>전문가 종합 토의 2</td><td>전문가 종합 토의 3</td></tr>
</table>

[그림 Ⅱ-49] 이용자 아카데미 현장 사진

9. TTA 품질 검증

1) 배경 및 목적

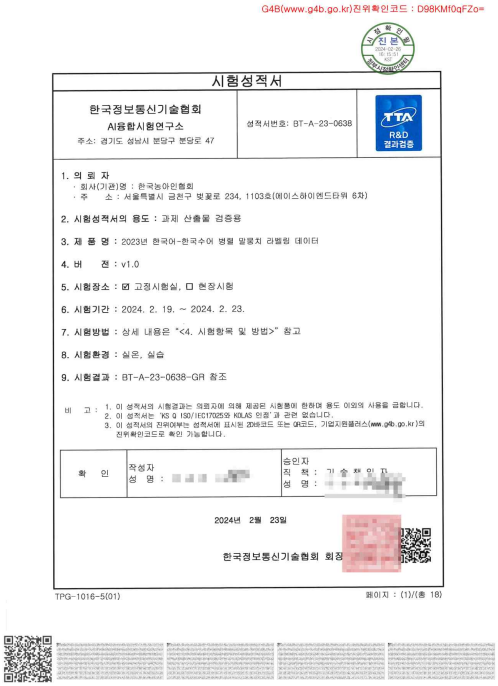
- ㅇ 2023년 한국어-한국수어 병렬 말뭉치 사업의 제3자 품질 검증을 통해 객관적인 데이터의 유용성을 확인한다.
- ㅇ 전문 기관의 품질 검증을 통해 데이터의 충분성, 균등성, 편향성 및 정확도를 확보할 수 있다.

2) 시험규격 및 결과

<table>
<tr><th>ID</th><th>시험 항목</th><th>시험 목표 및 기준</th><th>결과</th></tr>
<tr><td>TC1</td><td>데이터
다양성</td><td><시험목표>
- 데이터 분야별 구축량을 어절 단위로 확인
<시험기준>
ㅇ 기준: 한국어 구축 데이터 분야별 어절 수 기준 초과 여부
<table><tr><th colspan="2">목표 어절수</th></tr><tr><td>의료</td><td>20만</td></tr><tr><td>일상생활</td><td>70만</td></tr><tr><td>뉴스</td><td>10만</td></tr></table></td><td>100%</td></tr>
<tr><td>TC2</td><td>데이터
구조
정확성</td><td><시험목표>
- 라벨링 데이터의 구조정확성 준수도 확인
<시험기준>
ㅇ 기준: 라벨 구성요소 형식 준수 확인
ㅇ 산정식: 준수율(%) = $\frac{(\text{전체속성수} - \text{오류건수}_{\text{구조}})}{\text{전체속성수}} \times 100$</td><td>99.99%</td></tr>
<tr><td>TC3</td><td>데이터
의미
정확성</td><td><시험목표>
- 원천데이터 한국어 스크립트를 한국수어 영상으로 표현한 재현율 확인
<시험기준>
ㅇ 기준: 한국수어 표현의 적정성을 전문가 검사를 통한 측정
<table><tr><th colspan="2">수어 영상 재현율 기준</th></tr><tr><td>Fail</td><td>의미가 다르다</td></tr><tr><td>Pass</td><td>의미가 유사 또는 정확하다</td></tr></table>ㅇ 산정식: 재현율(%) = $\frac{(\text{의미달성건수})}{(\text{샘플링총건수})} \times 100$</td><td>99.05%</td></tr>
</table>

[표 Ⅱ-29] TTA 시험규격서

TTA 데이터 품질검증

시험성적서	결과서
	

*출처: BT-A-23-0638_시험성적서 및 결과서

[그림 Ⅱ-50] 제3자 품질 검증

Ⅲ. 사업 수행 결과

1. 병렬 말뭉치 데이터 구축 결과

1) 최종 구축 데이터

○ 한국어-한국수어 병렬 말뭉치 구축 사업에서 한국어 1,000,000어절에 해당하는 수어 영상/대응 정보(JSON) 구축을 목표로 하였다.

- 한국어는 총 1,014,861어절, 수어 영상/대응 정보(JSON) 105,891개의 데이터를 구축하여 목표량을 달성하였다.

분야 구분		의료	뉴스	일상 생활	합계	목표	달성률
한국어	문장 수	20,826	11,085	73,980	105,891	105,000	100.8%
	어절 수	200,026	100,227	714,608	1,014,861	1,000,000	101.5%
수어 영상	수량	20,826	11,085	73,980	105,891	105,000	100.8%
대응정보 (JSON)	수량	20,826	11,085	73,980	105,891	105,000	100.8%
신규 타입	수량				1,912		

[표 Ⅲ-1] 최종 구축 데이터 수량

2. 활용 방안 및 기대 효과

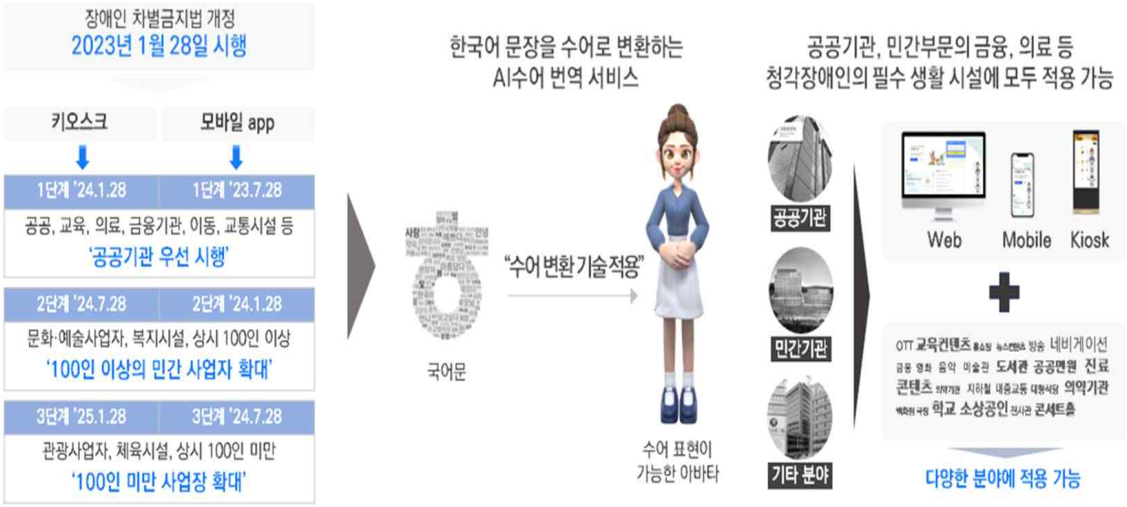
1) 병렬 말뭉치의 활용 방안

- ㅇ 교육 및 학습 자료: 교육기관이나 학교에서 수어를 배우는 학생들에게 유용한 교육 자료 및 학습 자료로 활용될 수 있으며. 이를 통해 청각장애인 학생들의 학습 경험을 향상시킬 수 있다.
- ㅇ 공공서비스: 공공기관의 웹사이트나 애플리케이션에서 한국수어를 지원함으로써 청각 장애인들이 더 쉽고 편리하게 공공 서비스를 접근하고 이용할 수 있다.
- ㅇ 의료 및 보건 서비스: 한국수어를 활용한 의료 및 보건 서비스를 제공하여 청각장애인들이 의료 서비스에 더욱 쉽게 접근할 수 있도록 도움을 준다.
- ㅇ 문화 콘텐츠: 한국수어를 통해 영화, 드라마, 공연 등의 문화 콘텐츠를 제공함으로써 청각장애인들이 문화생활을 즐길 수 있는 기회를 폭넓게 제공할 수 있다.
- ㅇ 연구 및 기술 개발: 한국어와 한국수어 말뭉치 데이터를 활용하여 수어 번역 및 음성 인식 기술 등을 개발하여 청각장애인들의 일상생활을 더 편리하게 만들 수 있다.

2) 사업의 기대 효과

- ㅇ 인공지능 학습데이터로 변환된 한국수어 데이터는 AI 영상인식, AI 아바타 등을 활용한 융합형 대화 서비스를 통해 농인과 청각장애인이 일상생활에서 원활하게 의사소통할 수 있는 환경을 조성하는 데 이바지할 수 있다.
- ㅇ 수어는 청각장애인 사회의 주요 언어로 인식됨에 따라, 한국어와 한국수어 말뭉치 데이터의 구축은 다양한 언어 및 소통 방식을 이해하고 인정하는 사회적 포용성의 출발점이자 기초가 될 수 있다.
- ㅇ 수어를 통한 소통이 강화되면 청각장애인이 경험하는 사회 및 직장 내 어려움 및 차별을 줄일 수 있으며, 공공 기관이나 기업으로부터 더 포괄적인 정책과 서비스 제공을 끌어낼 수 있다.

○ 수어를 사용하는 청각장애인들은 자신의 의견을 표현하고 스스로 정보에 접근하는 등 사회적 자립성을 높일 수 있으며. 이에 따라 삶의 질이 향상되고 당당한 사회 구성원의 일원으로서 더욱 활발하게 참여하고 활동하는 변화를 이룰 수 있다.



[그림 Ⅲ-1] 수어번역 서비스 기대 효과

3) 제언

한국어-한국수어 병렬 말뭉치 구축은 청각장애인이 의료, 교육, 관공서의 민원 행정 등 일상생활 및 다양한 분야에서 의사소통에 필요한 말뭉치를 구축하는 데 목적이 있으므로 장기적 관점에서 계획을 수립하여야 한다. 향후 성공적인 한국어-한국수어 말뭉치 구축 사업을 위해 다음과 같이 제언하고자 한다.

ㅇ 한국어-한국수어 번역을 위한 연구 생태계 조성 필요

- 한국어-한국수어 간 자동 번역을 위한 모델이 활발히 연구되고 있으나 현재까지는 번역 모델에 대해 완벽히 정의된 기준이 없고 한국수어의 특성상 기계적 번역 연구와 언어 연구가 동시에 이루어져야 하므로 수요 기관, 전문 기관, 연구 기관, 대학 등이 연계하여 면밀한 논의와 협력을 통해 수어 번역 기준을 마련하는 수어 연구 생태계 조성이 필요하다.
- 이번 사업은 수어 형태소의 의미(글로스)를 국립국어원 수어 말뭉치 분석 사업의 기준에 준하는 타입(단어사전)으로 정의하여 전사 작업을 진행했으므로 수어의 의미 분석 연구의 활용성을 한층 더 높일 수 있을 것이다.

ㅇ 국가 주도로 구축된 한국어 데이터를 원문 데이터로 활용

- 고품질의 병렬 말뭉치를 구축하기 위해서는 한국수어 번역을 위한 한국어 원문 데이터의 수집도 중요하다. 그러나 품질이 좋으면서 저작권 문제까지 해결된 대규모의 원문 데이터를 확보하는 것은 예산 등 다양한 측면에서 어려움이 따른다. 이에 대한 해결 방법은 국가 기관 주도로 기구축된 한국어 데이터를 병렬 말뭉치 구축에 활용하는 것이다. 이를 통해 저작권 관련 비용을 절감하고 절감된 재원을 한국수어 번역, 검수, 품질점검 등에 활용함으로써 보다 품질이 보증된 데이터를 확보할 수 있는 장점이 있다.
- 이번 사업에서는 AI 허브의 인공지능 학습용 데이터의 일부를 활용하여 원문 데이터를 수집하였다. 오픈 데이터를 활용한 데이터 수집 방식에 따라 이용함에 따라 저작권 및 개인정보보호법 등에 관한 일체의 법적 분쟁의 소지 없이 말뭉치 구축 사업에 적합한 데이터를 확보할 수 있어서 데이터 활용성을 더욱 높일 수 있다.

참고자료

국립국어원(n.d.), 언어정보나눔터 모두의 말뭉치. 검색 일자 2024. 1. 31. https://kli.korean.go.kr/corpus/main/requestMain.do?lang=ko

김범준, 전형기, 이경희(2023), 영상 내의 신체 핵심 좌표 데이터를 활용한 머신러닝 기반 수어 인식 연구, 한국정보통신학회논문지, 27(4), 459-466.

보건복지부・한국보건사회연구원(2020), 2020년 장애인 실태조사. 세종: 한국보건사회연구원.

성호열(2023), 수어번역을 위한 트랜스포머 모델 연구. 고려대학교 석사학위 청구 논문.

세계일보(2023), [단독] 한 달 800건 넘게 수어 통역도… 격무에 이직 빈번 농인만 속앓이 [심층기획-말뿐인 공용어…설 곳 없는 한국수어(手語)] https://www.segye.com/newsView/20230530514742

장애인차별금지법(보건복지부 법률 제18334호, 2023. 1. 28. 시행).

저작권법(문화체육관광부 법률 제18547호, 2021. 12. 8. 시행).

정의손, 조동휘, 박세희, 강현아, 박승보(2022), LSTM을 활용한 수어 단어 인식. 한국컴퓨터정보학회 학술발표논문집, 287-288.

한국수화언어법(문화체육관광부 법률 제18783호, 2022. 7. 19. 시행).

한국전자통신연구원(2020. 6. 3.), ETRI, 장애인 위한 코로나19 지침 아바타 수어 개발[보도자료], 검색 일자 2023. 1. 31. 사이트 주소 https://www.etri.re.kr/kor/bbs/view.etri?b_board_id=ETRI06&b_idx=18224

한국지능정보사회진흥원(n.d.), AI-HUB 데이터 이용정책. 검색 일자 2024. 01. 31. https://www.aihub.or.kr/intrcn/guid/usagepolicy.do?currMenu=151&topMenu=105

Apple(n.d.), signtime. accessed Jan 31. 2024. https://www.signtime.apple/

European Sign Language Center(2018), Spread the sign. accessed Jan 31. https://www.spreadthesign.com/en.us/search/

Signbank(2023), SignPuddle Online. accessed Jan 31. https://www.signbank.org/signpuddle/

Johnston, T., & Schembri, A. (2007). *Australian Sign Language (Auslan) An introduction to sign language linguistics.* Cambridge University Press.

<사업 참여자>

총괄 책임자	정희찬
실무 책임 및 관리자	하윤호, 장철성
담당 연구원	곽정란(주무관), 차예진(연구원)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2024년 1월 31일
발행일: 2024년 1월 31일
인 쇄: ㈜태산인디고

※ 이 보고서는 국립국어원의 국고 보조금으로 수행한 '2023년 한국어-한국수어 병렬 말뭉치 구축' 사업의 결과물을 발간한 것입니다.