

국립국어원 2023-01-19

발간등록번호
11-1371028-000950-01

2022년 한국어-한국수어 병렬 말뭉치 구축

총괄 책임 | 정희찬

2023. 4. 28.



제출문

국립국어원장 귀하

국립국어원의 국고 보조금으로 수행한 「2022년 한국어-한국수어 병렬 말뭉치 구축」사업의 결과보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 8월 31일 ~ 2023년 4월 28일

2023년 4월 28일

총괄 책임 : 정희찬(한국농아인협회)

사업수행기관 ' 22년 한국어-한국수어 병렬 말뭉치 구축사업단
(사)한국농아인협회)

총괄 책임 정희찬

실무 책임 및
관 리 하운호

<국문 요약>

2022년 한국어-한국수어 병렬 말뭉치 구축

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

한국어-한국수어 병렬 말뭉치 구축은 크게 한국어의 ‘수집’, ‘정제’, ‘산출’의 작업, 한국수어의 ‘번역’, ‘촬영’, ‘검수’의 작업, 인공지능 기술 및 응용 서비스 개발에 필요한 수어 영상 ‘키포인트 추출’, ‘키포인트 라벨링’, ‘형태소 라벨링’ 작업으로 진행되었다.

한국어 수집 분류는 크게 ‘생활 분야’, ‘문화 분야’ 2개의 분야로 구분하고 ‘민원/행정’, ‘의료’, ‘쇼핑’, ‘관광’의 4가지 대분류로 나누어 수집하였고, 수집한 데이터는 ‘정제’ 지침에 따라 정제작업이 이루어졌으며, 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 진행 후 ‘촬영’ 단계를 거쳐 수어 영상 데이터를 산출하고 ‘(사)한국농아인협회’ 수어 전문가의 검수를 거쳐 최종 수어 영상 데이터를 구축하였다.

인공지능 기술 및 응용서비스 개발에 필요한 인공지능 학습용 데이

터로 변환할 수 있도록 촬영된 수어 영상은 전용 프로그램(AITOK-키포인트 에디터)을 통해 인체 키포인트 121개를 자동으로 추출하고 저작 도구를 활용하여 라벨링 하였다. 이후 손 움직임과 얼굴 표정(비수지 신호)에 의미를 부여하는 형태소 라벨링 주석 작업을 진행하였으며 수어 영상과 형태소 라벨링 데이터는 다시 최종 감수자가 5% 감수를 진행하였다. 이상의 단계를 거쳐 품질을 검증하여 최종 데이터로서 산출되었다.

이를 통해 한국어-한국수어 병렬 말뭉치 한국어 1,001,087어절, 한국수어 120,295문장, 수어 영상 120,295건, 인공지능 활용을 위한 JSON 파일 120,295건을 구축하였다. 수집 분야별로는 의료서비스 273,663어절, 민원 행정서비스 303,713어절, 쇼핑 서비스 203,370어절, 관광 서비스 220,341어절이다.

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 조성하는데 이바지할 수 있다.

둘째, 한국어-한국수어 말뭉치 데이터를 통해 수어 인식 및 인공지능 번역 분야에 연구 자료로 활용하여 양방향 번역 기술의 발전을 위한 기초 자료로 활용될 수 있다.

셋째, 구축된 한국수어 AI 학습데이터를 다양한 연구 및 개발에 활용하고 공유하여 연구자들과 개발자들이 자유롭게 데이터를 활용할 수

있는 환경을 조성하여 학문적인 발전과 혁신을 촉진할 수 있다.

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

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

<영문 요약>

2022 Korean-Korean Sign Language Parallel Corpus

The purpose of this project was to build sign language video learning data necessary for the development of artificial intelligence technology and application services that enable farmers and hearing-impaired individuals to recognize sign language through videos.

The construction of a parallel corpus between Korean and Korean sign language involved various tasks, including the ‘collection’, ‘refining’, ‘calculation’, ‘translation’, of sign language, ‘shooting’, ‘inspection’, sign language video ‘keypoint extraction’, ‘keypoint labeling’, and ‘morpheme labeling’.

The Korean language collection was primarily classified into two categories: ‘life field’ and ‘cultural field’, with four major subcategories: ‘civil complaint/administration’, ‘medical’, ‘shopping’, and ‘tourism’.

Korean sign language was ‘translated’ by experts qualified in sign language interpretation, following the translation guidelines. After a joint inspection process conducted by high-ranking inspectors and sign language models, the sign language data was obtained through the ‘shooting’ stage.

Sign language images, taken for conversion into artificial intelligence learning data necessary for the development of artificial intelligence(AI) technology and application service, were automatically extracted and labeled using a dedicated program(AITOK – keypoint editor). Morpheme labeling annotation work was carried out, and the sign language images and morpheme labeling data were further reduced by 5%. Through these steps, the quality was verified, and the final data was calculated.

As a result, a total of 1,001,087 Korean words, 120,295 Korean sign language sentences, 120,295 sign language images, and 120,295 JSON files for artificial intelligence usage were constructed. In terms of collection field, there are 273,663 words

for medical services, 303,713 words for civil service administration services, 203,370 words for shopping services, and 220,341 words for tourism services.

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

Firstly, the conversion of Korean sign language data into artificial intelligence learning data can contribute to creating an environment where hearing-impaired individuals can smoothly communicate in their daily lives through convergent conversation services using AI image synthesis and AI avatars.

Secondly, the Korean-Korean sign language corpus data can serve as research data in the fields of sign language recognition and artificial intelligence translation, providing essential groundwork for the development of two-way translation technology.

Lastly, by utilizing and sharing the established Korean sign language AI learning data, an environment can be created for researchers and developers to freely utilize the data, promoting academic development and innovation in various research and development endeavors.

This project is expected to provide vital foundational data for the research and expansion of artificial intelligence technology.

Keywords: Korean, Korean sign language, Korean-Korean sign language parallel corpus, annotation, artificial intelligence learning data

목 차

I. 사업 개요	1
1. 사업 목표 및 추진 방향	1
1) 사업 목표	1
2) 추진 방향	3
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 및 정제	12
2) 한국수어 변환 및 수어 제공	24
3) 수어 영상 촬영	26
2. 키포인트 라벨링 구축	29
1) 키포인트 위치	29
2) 키포인트 라벨링 기본 방법론	32
3) 키포인트 저작도구 사용	33

3. 형태소 라벨링 구축	35
1) 주석 기본 원칙	35
2) 주석 세부 지침	45
4. JSON 파일 구축	47
1) 병렬 말뭉치 데이터 구조	47
2) 데이터 포맷 개요	48
3) 데이터 백업 관리	51
5. 병렬 말뭉치 데이터 검수	52
1) 검수 항목 및 활동 정의	52
2) 문장 데이터 검수	53
3) 수어 영상 촬영데이터 검수	54
4) 한국수어(촬영 영상) 검수	55
5) 형태소 라벨링 데이터 검수	56
6) 키포인트 라벨링 데이터 검수	64
6. 병렬 말뭉치 데이터 품질관리 및 검증	65
1) 데이터 품질관리	65
2) 세부 검증 및 품질관리	75
7. 보안 관리	77
1) 보안 관리 개요	77
2) 원천 자료 및 구축 자료에 대한 저작권 확보	77
3) 개인정보보호 등 보안 정책 및 지침 준수	79
4) 사업 수행을 위한 보안 대책 수립 및 준수	79
5) 보안계획 점검 사항	80
Ⅲ. 사업 수행 결과	81
1. 병렬 말뭉치 데이터 구축 결과	81
1) 최종 구축 데이터	81

2. 활용 방안 및 기대 효과	82
1) 병렬 말뚝치의 활용 방안	82
2) 사업의 기대 효과	82
3) 제언	84
참고자료	85

표 목 차

<표 I -1> 보유 특허	5
<표 I -2> 경쟁력	5
<표 I -3> 제품화 및 활용 분야	6
<표 I -4> 수행기관 인력 구성	7
<표 I -5> 한국어-한국수어 병렬 말뭉치 구축 절차	7
<표 I -6> 일정별 사업 추진 계획	8
<표 II -1> 수집 대상 세부 분야	12
<표 II -2> 수어 통역이 필요한 영역	13
<표 II -3> 수어 통역이 필요한 영역 2	15
<표 II -4> 개인 정보 비식별화 지침	19
<표 II -5> 한국수어 검수 기준	25
<표 II -6> 수어 영상 현장 검수 기준	25
<표 II -7> 수어 영상 2차 검수 기준	26
<표 II -8> 촬영 장비 셋팅 값	28
<표 II -9> 수어 영상 스튜디오 촬영 환경	28
<표 II -10> 키포인트 몸 분류	30
<표 II -11> 키포인트 얼굴/손(왼손) 분류	31
<표 II -12> 키포인트 오른손 분류	32
<표 II -13> 저작 도구 단축키	34
<표 II -14> 일치동사 주석 예시 1	38
<표 II -15> 일치동사 주석 예시 2	38
<표 II -16> 일치동사 주석 예시 3	39
<표 II -17> 일치동사 주석 예시 4	39
<표 II -18> 일치동사 주석 예시 5	40
<표 II -19> 일치동사 주석 예시 6	40

<표 II-20> 생산적 수어 예시	41
<표 II-21> 비수지 신호 용어 정리	42
<표 II-22> 토큰 분절 예시	45
<표 II-23> 양손 주석 예시	46
<표 II-24> 양손 주석 위치와 분절 예시	46
<표 II-25> Json 구조정의서	48
<표 II-26> 검수 항목 및 활동 정의	52
<표 II-27> 개인 고유 식별 정보 비식별화 여부 검수 항목	53
<표 II-28> 개인 특정 가능 정보 비식별화 여부 검수 항목	53
<표 II-29> 수어 영상 촬영데이터 검수 항목	54
<표 II-30> 한국수어 검수 기준	55
<표 II-31> 형태소 라벨링 데이터 검수 내용	56
<표 II-32> 토큰 길이 기준 틀린 주석 예시 1	59
<표 II-33> 토큰 길이 기준 주석 예시 1	59
<표 II-34> 토큰 길이 기준 틀린 주석 예시 2	60
<표 II-35> 토큰 길이 기준 주석 예시 2	60
<표 II-36> 수어 모델 비수지 정보	61
<표 II-37> 비수지 검수 사항 1	62
<표 II-38> 비수지 검수 사항 2	63
<표 II-39> 키포인트 라벨링 데이터 검수 절차	64
<표 II-40> 품질관리 체계	74
<표 II-41> 프로세스 품질 검사 내용 및 일정	75
<표 II-42> 데이터 품질 검사 내용 및 일정	75
<표 II-43> 데이터 품질관리 교육 방안	76
<표 III-1> 최종 구축 데이터 수량	81

그림 목 차

[그림 I -1] 청각장애인에게 높은 사회 장벽 분야	1
[그림 I -2] ICT 기술을 통한 의사소통의 필요성	2
[그림 I -3] 수어 번역 기술의 개념	3
[그림 I -4] 코로나19 방역 지침 AI 음성 수어 서비스	5
[그림 II -1] 민원/행정 원시 데이터 예시	18
[그림 II -2] 파일명 코드 부여 지침	20
[그림 II -3] 한국지능정보사회진흥원 데이터 이용정책	21
[그림 II -4] 지적 재산권 이용 동의 계약 예시	22
[그림 II -5] 동적 숫자 표기법 예시	23
[그림 II -6] 동음이의어 표기법 예시	23
[그림 II -7] 수어 표현범위를 고려한 촬영 화면 구성	27
[그림 II -8] 수어 제공 프롬프트	28
[그림 II -9] COCO Wholebody dataset 키포인트	29
[그림 II -10] AITOK - 키포인트 에디터	33
[그림 II -11] 일치동사 주석 방법 1	36
[그림 II -12] 일치동사 주석 방법 2	37
[그림 II -13] 일치동사 주석 방법 3	37
[그림 II -14] 수어 주석도구 화면	44
[그림 II -15] 수어 토큰 글로스 기준	45
[그림 II -16] 비우세 주석과 문장 흐름	46
[그림 II -17] 영상 단위 병렬 말뭉치 데이터 구조	47
[그림 II -18] 데이터 백업 관리 프로세스	51
[그림 II -19] 수어 토큰 길이 설정	57
[그림 II -20] 수어 토큰 글로스 기준	57
[그림 II -21] 양손 주석 예시	58

[그림 II-22] 문장 흐름에 따른 양손 주석	58
[그림 II-23] 비우세 토큰 주석 예시	58
[그림 II-24] 수어 저작도구 권한	63
[그림 II-25] 품질 요구사항 충족 여부	65
[그림 II-26] 품질 목표 충족 여부	66
[그림 II-27] 준비성(계획 수립성) 체크리스트 1	67
[그림 II-28] 준비성(계획 수립성) 체크리스트 2	68
[그림 II-29] 준비성(체계 준수성) 체크리스트 1	69
[그림 II-30] 준비성(체계 준수성) 체크리스트 2	70
[그림 II-31] 완전성(수집 완전성) 체크리스트	70
[그림 II-32] 완전성(정제 완전성) 체크리스트	71
[그림 II-33] 완전성(가공 완전성) 체크리스트	71
[그림 II-34] 기준 적합성 체크리스트 1	72
[그림 II-35] 기준 적합성 체크리스트 2	73
[그림 II-36] 보안 관리 전략	77
[그림 II-37] 저작권 이용 동의서 샘플	78
[그림 II-38] 보안 점검 사항 체크리스트	80
[그림 III-1] 농인 및 청각장애인이 가지는 기대 효과	83

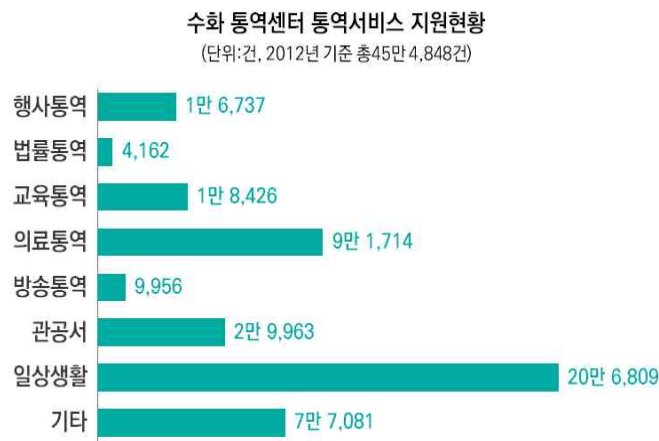
I. 사업 개요

1. 사업 목표 및 추진 방향

1) 사업 목표

(1) 추진 배경

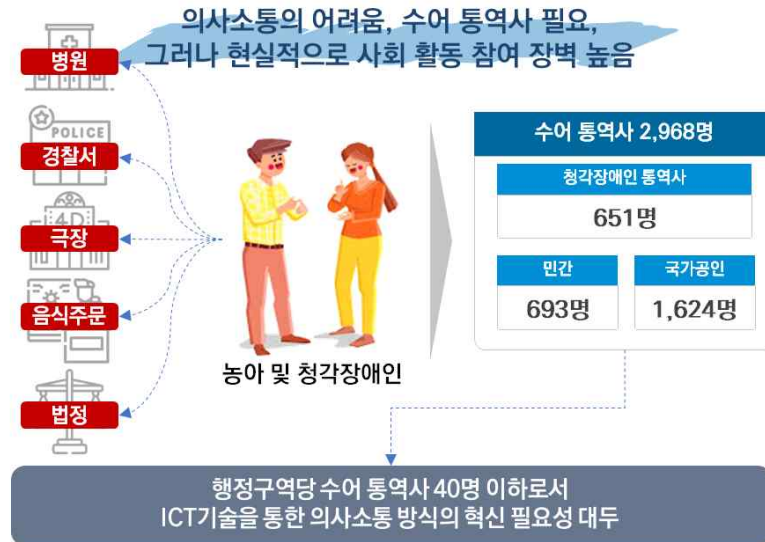
- 2020년 장애인 실태조사 보고에 따르면 등록된 청각장애인은 약 447,948명으로 나타나고 있다(보건복지부·한국보건사회연구원, 2020).
- 농인 및 청각장애인은 구어를 사용하는 데 제한되어 의사소통과 정보 전달에 어려움을 가지고 있다.
- 병원에서의 예약, 안내 및 관공서(동사무소, 구청 등), 경찰서 등에서의 민원 신고의 어려움 등, 일상생활 및 다양한 분야에서 높은 참여 장벽을 경험하고 있다.



[그림 1-1] 청각장애인에게 높은 사회 장벽 분야

- 2016년 한국수화언어법이 제정됨에 따라 한국수어의 필요성과 중요성이 강화되었으며, 국내 수어 통역 서비스를 받은 청각장애인 수도 지속적으로 증가하고 있다.

- 한국수어는 음성언어와 달리 습득과 학습 환경이 매우 열악하여 많은 변이가 나타나므로 다양한 수어 자료를 기록하여 제공할 필요가 있다(국립국어원, 2022).



[그림 1-2] ICT 기술을 통한 의사소통의 필요성

- 농인 및 청각장애인(이하 ‘농인’)이 겪는 사회적 어려움과 차별
 - 전 세계에 농인은 약 4억 명이 있으며, 국내에는 약 44만 명이 있다.
 - 국내 농인들의 주 의사소통 방법은 한국수어이며 농인 10명 중 약 7명이 수어를 제1언어로 사용하고 있다.

(2) 사업의 필요성

- 국내 수어 통역사 자격을 가진 인원은 1,948명으로(한국농아인협회, 2021) 수어 통역 요구에 비해 매우 열악한 상황에 놓여있다.
- 전국의 200여 개 수어 통역센터와 기타 통역중개소에서 농인의 의사소통을 지원하고 있지만, 수요에 비해 공급이 부족하여 농인의 불편함 해소가 미미하고 이에 따라 일상생활에서 수많은 한계와 좌절을 경험하는 것이 현실이다.
- 농인들이 일상 및 다양한 생활 분야에서 겪는 통역의 어려움을 해결하기 위해 인공지능(AI) 기술 활용이 필요하다.
- 인공지능 기술 및 응용 서비스 개발을 위해서는 AI가 인식할 수 있는 수어 영상 학습데이터가 필요하다.

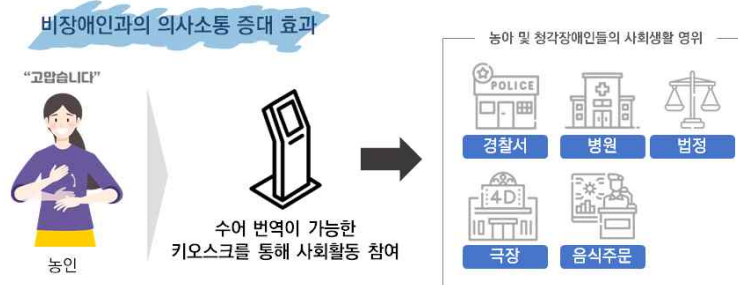
- ICT 기술이 발달하면서 의사소통의 방식도 다양하게 변화하고 있으나, 이를 활용하여 자동 수어 통역 기능을 구현하기 위한 농인 수어 영상 데이터는 부족한 실정이다.
- 한국어-한국수어 병렬 말뭉치 구축 사업을 통해 AI 학습용 수어 영상 학습데이터를 구축하여 사회참여 장벽을 적극적으로 해소할 필요성이 대두되고 있다.
- 농인과의 의사소통의 어려움으로 발생하는 농인들의 문화·생활을 촉진하고 참여 장벽을 해소하기 위해 다양한 형태의 의사소통 방식의 구현이 필요하다.

2) 추진 방향

(1) 관련 동향

가) 수어 번역 서비스 기술

- 기술의 개념
 - 농인과 비장애인 간 대화가 가능하게 하는 수어 통역 서비스를 통해 농인들의 사회생활을 원활하게 한다.



[그림 1-3] 수어 번역 기술의 개념

- 기술의 상세 내용 및 기술 이전 범위
 - 딥러닝 기반 이미지(동영상) 캡셔닝 기술
 - 딥러닝 기반 이미지(동영상) 기반의 행동 인지 기술
 - 멀티모달(이미지, 언어) 데이터 설계 및 구축 기술
 - 자연어 생성 기술, TTS 기술
- 사업화 제약 사항
 - 대규모 수어 통역 학습용 데이터의 신규 구축이 필요
- 국내 기술 동향
 - 자율지능 디지털 동반자 기술을 위한 사용자 의도 및 맥락 인지

기술은 최근 2~3년간 다수의 기업에 의해 개발되고 있으나 개발 기간이 짧고, 투입 인력이 부족할 뿐 아니라 서비스에 연계된 기술 개발 환경이 제대로 이루어져 있지 않아 개별적으로 상황인지 기술을 개발 중이다.

- 수어 영상 데이터 구축 동향

- 수어 인식과 관련해 구축된 데이터의 국내외 주요 사례로는 ‘PHOENIX -2014 : a German Sign Language Dataset for CSLR’ (PHOENIX(2014), ‘Chinese Sign Language Dataset(CSL)’ (USTC), ‘한국전자기술연구원-수어 데이터셋’ (한국전자기술연구원, 2020)이 존재한다.

- 수어 인식 인공지능 기술 동향

- 2018년 9월 한국전자기술연구원(KETI)은 국내 연구 기관 중 최초로 장갑, 마커 등 화자의 신체에 부착하는 별도 장비 없이 한 대의 RGB 카메라만으로도 농인의 수어를 인식할 수 있는 영상 기반 인공지능 수어 인식 시스템을 개발했다. 비수지 요소를 포함해 화자의 몸동작을 특징점(Keypoint)의 위치에서 표현하고, 순환신경망(RNN) 기반의 딥러닝 모델을 이용해 자연어에 가까운 한국어 문장으로 번역하는 데 성공했다(한국전자기술연구원, 2020).
- 구글(Google)은 인공지능 기반 손 모양 인식 모델 개발 현재 Alpha beta 단계로 상용 서비스 단계의 성능은 확보하지 못했으나, 고사양 장비 및 카메라 없이도 손 모양에 대한 인식률을 높였다는 장점을 보유하고 있다.
- 아마존 인공지능 비서 ‘알렉사’를 위한 수어 인식 모듈 탑재 디스플레이 등을 장착해 수어 질의에 대한 답변 제공. 현재까지 낮은 인식률 등으로 상용화되어 있지 않으나, 향후 농인 사용자를 위한 해당 모듈 탑재 계획을 발표했다.



[그림 1-4] 코로나19 방역 지침 AI 음성 수어 서비스

○ 관련 보유 특허

번호	국가	출원번호(출원일)	상태	명칭
1	대한민국	2018-0030386(2018.03.15)	출원 완료	자동수어 인식 방법 및 시스템
2	대한민국	2018-0075133(2018.06.29)	출원 완료	딥러닝 기반 제스처 자동 인식 방법 및 시스템
3	미국	US16/147,962(2018.10.01)	출원 완료	Deep learning-based automatic gesture recognition method and system

[표 1-1] 보유 특허

○ 기술적 경쟁력

경쟁 기술	본 기술의 우수성 및 차별성
디지털 동반자 모델링	- 복합 상황 인지와 플래닝을 통해 개인화된 자율 서비스를 실행 - 멀티모달 입력 데이터를 처리하고 다양한 응용 서비스 실행이 가능한 시스템 모델
복합 상황 인지	사용자를 관찰하여 의도와 맥락 정보를 이해
수어 영상 기반 자연어 문장 생성	영상 수어 정보에 기반한 자연어 문장 생성 시스템

[표 1-2] 경쟁력

나) 국내외 시장 동향 및 전망

○ 국내 시장 동향 및 전망

- 인공지능 관련 국내 시장은 2015년부터 2020년까지 연평균 66.1% 성장하였으며, 북미·유럽 지역 국가에 비해 빠르게 성장하였다.
- 인공지능 관련 시장은 전문가 시스템, 자율로봇, 지능형 가상도우미가 시장을 이끌 것으로 보인다.

○ 해외 시장 동향 및 전망

- 인공지능 기술을 활용한 가상 개인비서 서비스 시장은 2024년 80억 달러 규모로 2016년 대비 900% 이상 성장할 것으로 예상된다.
- 음성인식 시장 규모는 2017년 167억 달러이며, 2025년 268억 달러 규모에 달할 것으로 전망된다.

다) 제품화 및 활용 분야

활용 분야(제품/서비스)	제품 및 활용 분야 세부내용
인공지능 수어 통역 동반자 서비스	시각/청각/언어/상황 등의 여러 지능을 조합하여 장애인과 비장애인 모두를 도와주는 인공지능 서비스

[표 1-3] 제품화 및 활용 분야

(2) 추진 목적

○ 농인들을 위한 수어 데이터 수집

- 농인이 사용하는 수어를 영상 기반으로 인식, 의사 전달을 할 수 있도록 인공 지능 기술 개발에 필요한 수어 영상 학습데이터를 구축하고자 한다.
- 본 사업을 통해 인공지능 학습용 데이터로 ①원천 영상 파일, ②원천 영상 메타 데이터, ③형태소 단위 데이터, ④신체, 손, 얼굴 키포인트 데이터를 구축하고자 한다.

2. 사업 수행 체계 및 절차

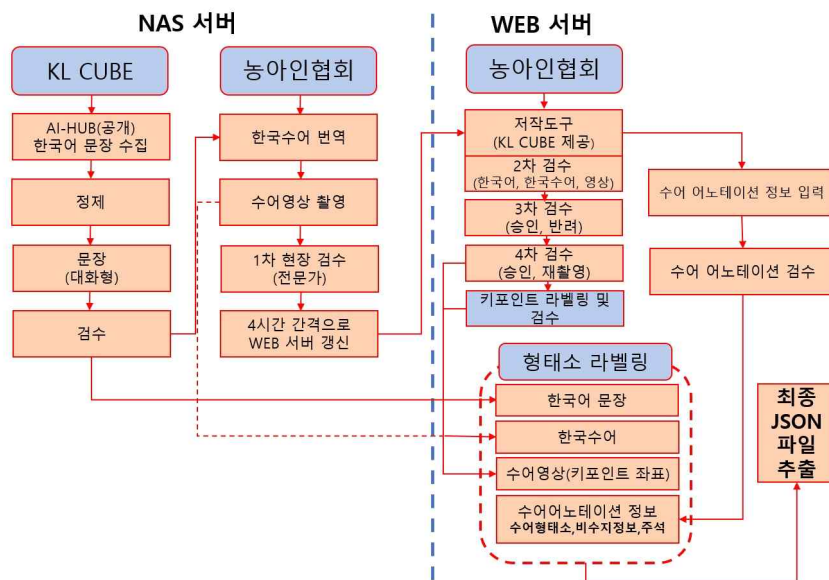
1) 사업 수행 체계 및 인력 구성

- 본 사업은 총괄 책임자 1인, 실무 책임자 1인, 그리고 실무 관리자 4인으로 구성되었다.



[표 1-4] 수행기관 인력 구성

2) 병렬 말뭉치 구축 절차



[표 1-5] 한국어-한국수어 병렬 말뭉치 구축 절차

3. 사업추진계획

1) 일정별 사업추진계획

세부 과제명	수행 내용	추진 일정(월)								비중 (%)
		M	M+1	M+2	M+3	M+4	M+5	M+6	M+7	
공고 및 선정	보조 사업자 선정									2
	사업수행사 공고 및 선정									6
데이터 수집/ 가공/검증	수집 문장 주제 분류									2
	문장 수집									20
	수어 영상 제작 및 메타정보									20
	데이터 검증									10
데이터셋 구축	분류체계 및 표준정의									5
	가공 데이터셋 구축									10
	DB 구축									10
품질 활동	시범 DB 구축									5
	프로세스 품질 검사									5
	데이터 품질 검사									5

[표 1-6] 일정별 사업추진계획

2) 세부 사업추진 계획

(1) 사업 계획 수립 및 착수

- 사업수행 계획 수립 : 2022년 10월 31일 완료
- 사업수행 계획서 제출 : 2022년 10월 31일 완료

(2) 병렬 말뭉치 데이터 구축

가) 한국어

- 데이터 수집 및 정제 : 2022년 10월 26일 ~ 2023년 2월 28일 완료

나) 한국수어

- 한국수어 번역 : 2022년 11월 1일 ~ 2023년 2월 28일 완료
- 한국수어 추가 정보 : 2022년 11월 1일 ~ 2023년 2월 28일 완료

다) 수어 영상

- 수어 영상 : 2022년 11월 1일 ~ 2023년 2월 28일 완료

라) 키포인트 라벨링

- 키포인트 라벨링 : 2022년 12월 15일 ~ 2023년 3월 17일 완료

마) 형태소 라벨링

- 형태소 라벨링 : 2022년 12월 19일 ~ 2023년 3월 30일 완료

(3) 검수

가) 작업·검수 지침서 작성

- 한국어 작업·검수 지침서 작성 : 2022년 11월 30일 완료
- 한국수어, 수어 영상 작업·검수 지침서 작성 : 2022년 11월 30일 완료
- 키포인트 라벨링 작업·검수 지침서 작성 : 2022년 11월 30일 완료
- 형태소 라벨링 작업·검수 지침서 작성 : 2022년 11월 30일 완료

나) 작업·검수 지침 교육

- 사업 절차 및 품질관리 : 2022년 9월 19일 10:00~11:00
- 한국어 수집 및 정제
 - 1차 : 2022년 9월 30일 10:00~11:00
 - 2차 : 2022년 9월 30일 14:00~15:00
- 키포인트 라벨링
 - 1차 : 2022년 11월 8일 18:00~19:00
 - 2차 : 2023년 1월 5일 18:00~19:00
 - 3차 : 2023년 2월 7일 18:00~19:00
 - 4차 : 2023년 3월 3일 18:00~19:00
- 형태소 라벨링
 - 1차 : 2022년 11월 10일 14:00~16:00
 - 2차 : 2022년 11월 30일 14:00~16:00
 - 3차 : 2022년 12월 26일 14:00~16:00

- 4차 : 2023년 1월 13일 14:00~16:00
- 5차 : 2023년 2월 3일 14:00~16:00
- 이후 인력 충원에 따라 수시 실시

(4) 인력 확보

- 가) 한국어 수집 및 정제 : 2023년 3월 10일 완료
- 나) 수어 번역 : 2023년 3월 10일 완료
- 다) 수어 모델 : 2023년 3월 10일 완료
- 라) 현장 감수 : 2023년 3월 10일 완료
- 마) 영상 검수 : 2023년 3월 10일 완료
- 바) 키포인트 라벨링 : 2023년 3월 10일 완료
- 사) 형태소 라벨링 : 2023년 3월 10일 완료
- 아) 최종 라벨링 검수 : 2023년 3월 10일 완료

(5) 의사소통 및 의견 수렴

- 가) 단계별 보고
 - 착수보고 : 2022년 9월 22일
 - 중간보고 : 2022년 12월 22일
- 나) 정기 보고 및 회의
 - 주간보고 : 매주 수요일 실시
 - 월간보고 : 매월 1주차 수요일 실시

4. 주요 변경 사항

1) 요구사항

- 요구사항 정의서 일부 변경
- 요구사항 추적표 일부 변경

2) 작업 및 검수 지침

- 한국어 수집 작업 및 검수 지침 일부 내용 추가(V1.2)
- 한국수어 변환 및 수어 제공 작업 및 검수 지침 단계별 작업 지침 추가(V1.2)
- 수어 영상 촬영 및 검수 지침 일부 내용 추가(V1.2)
- 키포인트 라벨링 작업 및 검수 지침 일부 내용 추가(V1.2)
- 형태소 라벨링 작업 및 검수 지침 일부 내용 추가(V1.2)

3) 예산

- 인건비, 운영비, 보조사업자 인건비 등 비목 내역 변경(1~5차)

II. 사업 수행

1. 한국어 수집 및 정제

1) 한국어 수집 및 정제

(1) 한국어 문장 데이터 수집 분야

- 한국어 문장의 수집 분류는 크게 ‘생활 분야’, ‘문화 분야’ 2개 대분류로 구분하여 수집하였다.
- 수집량은 민원/행정/의료 등이 포함된 ‘생활 분야’에서 50%, 쇼핑/취미/여가생활 등이 포함된 ‘문화 분야’에서 50%를 수집하였다.

분야	대분류	중분류	수집범위
생활 분야	민원/행정	• 주택, 교육, 법률, 직업, 금융, 공공 민원	50%
	의료	• 진료 안내, 병원 이용 안내, 민원	
문화 분야	쇼핑	• 쇼핑	50%
	관광	• 여행 가기, 영화 보기	

[표 II-1] 수집대상 세부 분야

- 2020년 국립국어원 조사에 의하면 수어 통역이 우선으로 필요한 영역에 대한 조사 결과표는 아래와 같다(국립국어원, 2020).

	사례수	의료	일상 생활	TV 방송	교육	법률	직업	유튜브	ATM	키오스크	홈페이지	기타	무응답	계
전체	(539)	35.4	30.9	8.4	5.0	4.2	3.4	1.8	1.2	0.9	0.7	5.7	2.3	100.0
성별														
남자	(287)	33.4	31.2	8.4	4.4	3.5	3.7	2.1	1.3	1.0	0.4	7.1	3.5	100.0
여자	(252)	37.7	30.6	8.4	5.7	4.9	3.2	1.4	1.2	0.8	1.0	4.2	1.0	100.0
연령														
19-29세	(19)	15.4	20.3	6.9	18.6	10.5	15.9	8.2	0.0	4.2	0.0	0.0	0.0	100.0
30대	(24)	29.0	28.0	4.7	19.1	13.2	5.1	0.0	2.8	0.0	0.0	5.5	2.7	100.0
40대	(39)	35.8	29.7	13.2	5.9	4.1	4.3	2.8	1.5	0.0	0.0	2.6	0.0	100.0
50대	(76)	39.8	19.6	8.0	5.6	3.9	9.7	3.3	1.5	1.4	1.6	3.3	2.3	100.0
60대	(375)	35.8	34.8	8.0	3.3	3.4	1.4	1.1	1.1	0.8	0.7	7.0	2.7	100.0
모름	(6)	49.5	23.3	27.3	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	100.0
최종학력														
무학	(121)	32.4	30.6	8.8	6.3	3.6	2.3	2.3	0.0	1.6	0.0	9.9	2.2	100.0
초등학교	(149)	37.4	35.5	8.0	1.4	2.2	2.2	0.0	2.8	0.0	0.0	6.3	4.3	100.0
중학교	(112)	38.2	26.4	10.8	3.9	3.5	3.9	1.0	0.5	1.8	2.2	6.4	1.4	100.0
고등학교	(107)	35.3	34.9	6.7	5.8	5.6	5.9	1.1	1.1	0.0	1.2	1.1	1.4	100.0
대학이상	(49)	31.2	19.3	7.0	13.4	9.6	3.8	1.3	1.3	1.6	0.0	2.6	1.3	100.0
장애 유형														
청각장애	(396)	33.7	31.1	8.9	5.0	4.5	3.5	1.9	1.5	1.1	0.8	5.6	2.5	100.0
언어장애	(113)	42.8	25.9	7.2	5.9	3.4	4.3	1.1	0.5	0.5	0.6	6.1	1.8	100.0
청각기타	(30)	30.3	48.0	6.9	2.1	2.1	0.0	1.7	0.0	0.0	0.0	6.9	2.0	100.0
모름	(1)	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

[표 II-2] 수어 통역이 필요한 영역

- 1순위로 응답한 비율이 높은 영역은 의료가 35.4%로 가장 높게 조사되었으며, 이어 일상생활이 30.9%, 방송 8.4%, 기타 5.7%, 교육 5.0%, 법률 4.2%, 직업 3.4%, 유튜브 등 동영상 1.8%, 현금자동입출금기(ATM) 1.2%, 무인 단말기(키오스크) 0.9%, 홈페이지 0.7% 순으로 조사되었다.
- 성별과의 교차 분석 결과 남녀 모든 성별에서 우선적 필요 영역을 각 37.7%, 33.4%로 의료라고 응답하였고, 다음으로 일상생활, 티브이 방송, 교육 영역 순으로 나타났다.
- 연령과의 교차 분석 결과 20대는 일상생활 영역(20.3%), 교육(18.6%), 직업(15.9%) 순으로 높게 나타났고, 30대는 의료 영역(29.0%), 교육(19.1%), 일상생활 영역(18.0%) 순으로 나타났다.
- 40대 이상의 연령에서는 가장 높은 우선 필요 영역을 의료 영역, 다음으로 일상생활 영역이라고 응답하여 연령이 높아질수록 건강 관련 영역에 대한 수요가 높은 것으로 확인되었다.

- 교육 수준과의 교차 분석 결과 무학부터 대학교 졸업 이상의 교육 수준의 경우 의료 영역에서 수어 통역 서비스가 필요하다는 응답이 높게 조사되었으며, 다음으로 일상생활(19.3%), 티브이 방송(7.0%) 순으로 응답 비율이 높게 조사되었다.
- 장애 유형과의 교차 분석 결과 청각·언어장애와 청각장애는 각 42.8%, 33.7%로 의료 영역을 수어 통역이 가장 필요한 영역이라 응답하였으며, 청각·기타 장애는 일상생활 영역을 수어 통역 서비스가 가장 우선으로 필요한 영역이라고 응답하였다.
- 사용하는 재활 보조 기구와의 교차 분석 결과 보청기를 사용하는 응답자의 34.6%가 일상생활 영역을 수어 통역이 가장 필요한 영역이라 응답하였으며, 재활 보조 기구를 사용하지 않거나 기타 재활 보조 기구를 사용하는 응답자는 가장 필요한 영역을 의료 영역(40.7%)이라 하였다. 한편, 인공 와우를 사용하는 응답자는 티브이 방송(25.3%)을 가장 우선으로 수어 통역이 필요한 영역이라 응답하였고 다음으로 일상생활(18.0%)을 선택하였다.
- 수어를 주된 의사소통 방법으로 사용하는 응답자의 경우 가장 우선으로 필요한 수어 통역 영역에 대해 의료 영역 43.8%, 일상생활 23.0% 순으로 응답하였고, 주된 의사소통 방법으로 수어를 사용하지 않는 응답자의 경우 가장 우선하여 수어 통역이 필요한 영역으로 일상생활(40.3%), 의료(25.5%) 순으로 응답하여 두 집단 간에 차이가 있음을 확인할 수 있었다.

- 수어 통역이 우선으로 필요한 영역으로 다중 응답한 1순위와 2순위의 합을 기준으로 분석한 결과표는 다음과 같다.

	사례수	의료	일상 생활	TV 방송	교육	법률	직업	유튜브	ATM	키오 스크	홈 이지	기타	무응답
전체	(539)	66.0	48.5	17.0	14.7	13.3	8.4	4.2	2.4	1.1	0.8	7.3	2.3
성별													
남자	(287)	60.8	48.1	18.0	14.7	14.5	9.2	4.6	2.5	1.0	0.6	9.2	3.5
여자	(252)	72.0	49.0	15.8	14.7	12.0	7.5	3.7	2.3	1.2	1.0	5.2	1.0
연령													
19-29세	(19)	36.9	32.9	10.3	65.8	13.9	19.9	11.8	0.0	4.2	0.0	0.0	0.0
30대	(24)	72.4	26.0	7.4	32.0	18.5	18.9	0.0	2.8	2.7	0.0	8.3	2.7
40대	(39)	64.5	49.1	18.6	14.2	12.1	17.0	10.1	1.5	0.0	0.0	5.5	0.0
50대	(76)	62.2	38.7	15.7	17.6	13.5	18.0	6.8	4.5	2.1	2.4	4.9	2.3
60대	(375)	67.5	52.6	17.6	10.7	13.3	4.4	2.9	2.3	0.8	0.7	8.4	2.7
모름	(6)	100.0	49.5	50.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
최종학력													
무학	(121)	67.8	53.5	11.1	14.9	5.8	5.4	3.8	0.0	1.6	0.4	11.1	2.2
초등학교	(149)	68.1	51.7	19.6	6.6	11.0	5.7	2.8	3.7	0.0	0.0	8.1	4.3
중학교	(113)	68.0	45.2	20.9	11.5	17.8	6.5	4.1	3.2	1.8	2.2	8.6	1.4
고등학교	(107)	65.9	49.3	16.4	16.4	16.9	13.3	3.9	3.2	1.1	1.2	2.7	1.4
대학이상	(49)	50.9	32.1	16.1	16.1	20.5	17.6	9.9	1.3	1.6	0.0	2.6	1.3
장애 유형													
청각장애	(396)	64.4	47.9	15.7	15.0	14.2	9.8	4.7	2.9	1.2	0.8	7.2	2.5
언어장애	(113)	75.4	47.5	20.2	13.2	12.2	5.8	2.6	1.6	0.5	1.0	6.6	1.8
청각기타	(30)	51.8	59.8	22.2	16.9	6.2	0.0	3.5	0.0	2.1	0.0	11.8	2.0
모름	(1)	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

[표 II-3] 수어 통역이 필요한 영역 2

- 분석 결과 의료(66.0%), 일상생활(48.5%), 티브이 방송(17.0%), 교육(14.7%), 법률(13.3%), 직업(8.4%), 기타(7.3%), 유튜브 등 동영상(4.2%), 현금자동입출금기(2.4%), 무인 단말기(1.1%) 순으로 조사되었다.
- 성별을 기준으로 분석한 결과 여자와 남자 모두 수어 통역 서비스가 필요하다고 응답한 비율이 각 72.0%, 60.8%로 전체 영역 중 의료 영역이 가장 높게 나타났다.
- 연령별로 분석한 결과 20대를 제외한 30대 이상의 연령대에서 의료 영역에 수어 통역 서비스가 필요하다고 응답한 비율이 모두 가장 높게 조사되었으며 특히 30대(72.4%), 60대 이상(67.5%), 40대(64.5%) 순으로 응답하였다.
- 20대는 교육 영역(65.8%)을 수어 통역 서비스가 가장 우선으로 필요한 영역이라고 응답하였으며, 다음으로 의료 영역(36.9%) 비율이 높게 나타났다.

- 최종학력별로 분석한 결과 의료 영역에서 수어 통역이 필요하다고 응답한 비율은 초등학교 졸업(68.1%), 중학교 졸업(68.0%), 무학(67.8%), 고등학교 졸업(65.9%), 대학교 졸업 이상(50.9%) 순으로 높게 조사되었다.
- 장애 유형별로 분석한 결과 청각·언어장애와 청각장애는 각 75.4%, 64.4%로 의료 영역을 수어 통역이 가장 필요한 영역이라 응답하였으며, 다음으로 일상생활 영역, 티브이 방송 영역 순으로 응답률이 높았다.
- 반면, 청각·기타 장애는 일상생활(59.8%) 영역을 가장 우선으로 필요한 영역이라 응답하였고 다음으로 의료 영역(51.8%), 티브이 방송 영역(22.2%)을 선택하였다.
- 사용하는 재활 보조 기구별로 분석한 결과 보청기와 인공 와우를 사용하는 응답자는 일상생활 영역을 각 62.5%, 52.6%로 수어 통역이 가장 필요한 영역이라 응답하였으며, 재활 보조 기구를 사용하지 않거나 기타 재활 보조 기구를 사용하는 응답자도 가장 필요한 영역을 일상생활 영역(69.4%)이라 답하였다.
- 주된 의사소통 방법으로서의 수어 사용 여부와 상관없이 우선으로 필요한 수어 통역 서비스 영역에 대하여 의료 영역, 일상생활 영역 순으로 응답률이 높았다.
- 다만 차이가 있다면 수어를 주된 의사소통 방법으로 사용하는 응답자는 세 번째로 수어 통역이 필요한 영역을 법률 영역(17.1%)이라 답하였으나 수어를 주된 의사소통 방법으로 수어를 사용하지 않는 응답자의 경우 세 번째 필요 영역을 티브이 방송(19.4%)이라 응답하였다(국립국어원, 2020).

(2) 원시 데이터 획득

- 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 목표 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 목적 대화 데이터는 텍스트 데이터 총 4만 6천여 건으로 쇼핑, 민원, 의료, 관광의 주제로 분류하였다.
 - 복지 분야 콜센터 상담 데이터는 콜센터 상담 데이터 수집을 통해 관련 서비스 모델 활용에 적합한 AI 데이터셋 구축을 목적으로 확보된 데이터로, 콜센터 전화망 및 실제 전화 환경에서 수집된 학습용 음성 데이터를 활용하여 사용자의 물음에 적절한 답변을 하는 복지지원 AI 기반 상담 서비스로 활용할 수 있도록 하였다.
- 본 과업에서 구축할 말뭉치 주제와 유사성이 있는 오픈 데이터 활용
- 한국지능정보사회진흥원(NIA)의 AI Hub에 등록된 인공지능 학습용 데이터를 활용하여 생활과 문화 분야에 적용 가능한 문장을 수집하였다.
 - 의료 분야는 오픈 데이터 활용에 제약이 있는 경우 클라우드 작업자를 활용한 시나리오 작업을 통해 농인이 일상생활에서 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 개인 정보가 포함되어 있기에 취급에 있어 각별한 주의를 기울여 수집하였다.
 - 공공 기관 민원/안내의 경우 단순 민원 안내가 대부분이며 기관의 실과별 민원 안내 관련 문서에는 개인 정보가 포함되어 있기에 취급에 주의를 기울여 수집하였다.
 - 인허가 등 민원은 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 클라우드 소싱을 통해서 확보하거나 온라인 민원/안내 게시판을 활용하여 대화 데이터를 구성하였다.
 - 저작권 및 개인정보 보호법 등 법적 문제가 발생하지 않도록 데이터 제공 기관과의 업무협약, 개인 정보 삭제를 통한 가명/익명 정보화를 통해 데이터를 확보하고, 클라우드 소싱을 통한 대화 수집을 위해 활용 동의 과정을 수행하였다.
- 원시 데이터 선정
 - 데이터 품질, 획득 가능성(가능 여부 및 획득량), 획득 비용 및 기술 수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정하였다.
 - 선정된 원시 데이터를 획득하기 위해 필요한 정보 또는 원시 데이터 획득 현황을 파악하기 위한 데이터 명세서를 작성하여 데이터 획득 기준으로 활용하였다.

파일 코드	대화(한국어)
SLICCPAKOKSL2200000001	감사해요 좋은하루 보내세요
SLICCPAKOKSL2200000002	네네 수고하세요 좋은 하루 보내세요
SLICCPAKOKSL2200000003	안녕하세요 어떤 일을 도와드릴까요
SLICCPAKOKSL2200000004	코로나 확진 환자 이동 경로 공개하고 그러잖아요
SLICCPAKOKSL2200000005	네 선생님 확진 환자 관련 문의십니까
SLICCPAKOKSL2200000006	예 그 대상이나 기간 같은 거에 원칙이 있나 해서요
SLICCPAKOKSL2200000007	그 구체적인 내용이 궁금합니다
SLICCPAKOKSL2200000008	화장한 아침이네요 무엇을 도와드릴까요
SLICCPAKOKSL2200000009	안녕하세요 아침부터 죄송합니다
SLICCPAKOKSL2200000010	아닙니다 고객님의 어떤 일 때문에 오셨을까요
SLICCPAKOKSL2200000011	형(법률) 집행 중 발생한 상해나 장애에 대한 치료 보상 절차는 어떻게 돼요
SLICCPAKOKSL2200000012	치료와 관련하여 교정(행동) 시설 내 의무관 또는 필요하다고 인정될 시는요
SLICCPAKOKSL2200000013	외부 의료시설에서 진료를 받게 됩니다
SLICCPAKOKSL2200000014	하지만 수용자의 과실(행동)로 인한 부상은 진료비가 청구 될 수도 있습니다
SLICCPAKOKSL2200000015	지금 말씀 하신게 법으로 되어 있는 거죠

[그림 II-1] 민원/행정 원시데이터 예시

(3) 원시 데이터 정제

- 원시 데이터 정제 방식 기준
 - 원시 데이터 획득 및 정제 절차를 데이터 획득 분야, 방법별로 아래와 같이 정의하였다.
 - 인공지능 기술과 Annotation 기술과 가이드를 제공하여 데이터

관리에 대한 책임, 행정, 제반 절차를 최소화하고 빠른 데이터 획득 및 정제를 지원하였다.

- 수집한 대화 데이터 대부분은 사적인 대화 내용 안에 이름과 연락처, 소속 등을 비롯하여 개인의 신원이 노출될 수 있는 다양한 개인 정보가 포함되어 있으므로 정제 방침을 수립 및 준수하였다.
- 대화 내용에 포함된 개인 정보와 메타정보로 수집하는 성별과 연령, 직업, 출신지 등의 정보가 결합한 형태로 말뭉치로 구축될 경우 개인의 신원이 노출될 우려가 있어 개인 정보에 대한 철저한 비식별화를 진행하였다.
- 대화 내용에 비윤리적인 내용이 포함되어 있는 경우 원시 데이터에서 제외하였다.
- 과도한 비식별화로 인하여 대화 내용을 파악하는 것이 불가능하거나, 유효한 개체명(entity)을 추출하기가 어려워지는 등 자료의 활용도가 떨어질 가능성이 있으므로 개인 정보가 노출되지 않고 대화 특성이 반영될 수 있는 비식별화 지침을 마련 및 준수하였다.

○ 단계별 정제 지침

- 1차 정제는 식별자 및 속성자 기준을 참고하여 개인 정보 비식별화를 실시하였다.

구분	식별자	속성자
지침	개인의 구분을 위하여 부여된 고유한 값 또는 이름을 비식별화	개인을 특정할 수 있는 상황인지 판단하여 비식별화
항목	<ul style="list-style-type: none"> ▪ 고유 식별 정보(주민 등록 번호, 운전 면허증 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 	<ul style="list-style-type: none"> ▪ 성별, 연령, 국적, 고향, 우편 번호, 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, 부서 명, 직급

[표 II-4] 개인 정보 비식별화 지침

- 2차 정제는 대화 내용 중 비윤리적 내용 및 혐오 표현은 삭제하고, 외래어 및 로마자, 한자 정제 내용이 포함되었을 경우 어문 규범에 맞는지 확인하고 맥락상 불필요한 특수 문자(구두점 및 이모티콘 등)를 제거한 후 4~15어절(평균 9.5) 이내로 문장을 재구성하였다.
- 3차 정제는 수어 번역 전문가와 협업할 수 있는 작업 프로세스를 구축하고 한국어와 한국수어 간 차이를 고려하여 번역 가능 여부를 파악하였다.
- 4차 정제는 3차까지 완료된 데이터에 한하여 오타 및 띄어쓰기 등 최종 검수하여 데이터 정제를 완료하고, 파일명에 고유 코드를 부여하였다.

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
속성	유형	분야	장르	병렬	원자료	결과자료	구축연도	일련번호(8자리)														
정의값	N: 문어 S:구어	LI: 생활 CU: 문화	ME:의료 CC:민원행정 SH:쇼핑 TO:관광	PA:병렬	KO:한국어	KSL:한국수어	22	00000001 ~ 99999999 (8자리 일련번호)														
※ 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)																						

[그림 II-2] 파일명 코드 부여 지침

(4) 저작권 검토 및 지적 재산권

○ 원시 데이터 저작권 확보

- 원시 데이터 확보

- 원시 데이터(한국어 문장) 확보는 AI-HUB에 공개된 개방데이터를 사용했으며 한국지능정보사회진흥원의 데이터 이용 정책을 준수하여 저작권을 확보하였다.

데이터 이용정책

AI 허브 개방 데이터

데이터 소개

- AI 허브에서 제공되는 인공지능 학습용 데이터(이하 'AI데이터'라고 함)는 과학기술정보통신부와 한국지능정보사회진흥원의 「지능정보산업 인프라 조성」 사업의 일환으로 구축되었으며, 본 사업의 유·무형적 결과물인 데이터, AI 응용모델 및 데이터 저작도구의 소스, 각종 매뉴얼 등(이하 'AI데이터 등')에 대한 일체의 권리는 AI데이터 등의 구축 수행기관 및 참여기관(이하 '수행기관 등')과 한국지능정보사회진흥원에 있습니다.
- 본 AI데이터 등은 인공지능 기술 및 제품·서비스 발전을 위하여 구축하였으며, 지능형 제품·서비스, 챗봇 등 다양한 분야에서 영리적·비영리적 연구·개발 목적으로 활용할 수 있습니다.

데이터 이용정책

- 본 AI데이터 등을 이용하기 위해서 다음 사항에 동의하며 준수해야 함을 고지합니다.

1. 본 AI데이터 등을 이용할 때에는 반드시 한국지능정보사회진흥원의 사업결과임을 밝혀야 하며, 본 AI데이터 등을 이용한 2차적 저작물에도 동일하게 밝혀야 합니다.
2. 국외에 소재하는 법인, 단체 또는 개인이 AI데이터 등을 이용하기 위해서는 수행기관 등 및 한국지능정보사회진흥원과 별도로 합의가 필요합니다.
3. 본 AI데이터 등의 국외 반출을 위해서는 수행기관 등 및 한국지능정보사회진흥원과 별도로 합의가 필요합니다.
4. 본 AI데이터는 인공지능 학습모델의 학습용으로만 사용할 수 있습니다. 한국지능정보사회진흥원은 AI데이터 등의 이용의 목적이나 방법, 내용 등이 위법하거나 부적합하다고 판단될 경우 제공을 거부할 수 있으며, 이미 제공한 경우 이용의 중지 및 AI 데이터 등의 환수, 폐기 등을 요구할 수 있습니다.
5. 제공 받은 AI데이터 등을 수행기관 등과 한국지능정보사회진흥원의 승인을 받지 않은 다른 법인, 단체 또는 개인에게 열람하게 하거나 제공, 양도, 대여, 판매하여서는 안됩니다.
6. AI데이터 등에 대해서 제 4항에 따른 목적 외 이용, 제5항에 따른 무단 열람, 제공, 양도, 대여, 판매 등의 결과로 인하여 발생하는 모든 민·형사 상의 책임은 AI데이터 등을 이용한 법인, 단체 또는 개인에게 있습니다.
7. 이용자는 AI 허브 제공 데이터셋 내에 개인정보 등이 포함된 것이 발견된 경우, 즉시 AI 허브에 해당 사실을 신고하고 다운로드 받은 데이터셋을 삭제하여야 합니다.
8. AI 허브로부터 제공받은 비식별 정보(재현정보 포함)를 인공지능 서비스 개발 등의 목적으로 안전하게 이용하여야 하며, 이를 이용해서 개인을 재식별하기 위한 어떠한 행위도 하여서는 안됩니다.
9. 향후 한국지능정보사회진흥원에서 활용사례·성과 등에 관한 실태조사할 경우 이에 성실하게 임하여야 합니다.

[그림 II -3] 한국지능정보사회진흥원 데이터 이용정책

○ 원천 자료 및 구축 자료에 대한 저작권 확보 방안

- 원천 자료에 대한 저작권 확보 방안

- 한국수어 병렬 말뭉치 데이터 구축을 수행한 데이터 수집 작업자를 대상으로 직접 계약을 체결하며, 데이터 이용 허락 동의를 얻어 저작권 확보하였다.
- ‘저작권 이용 허락 동의서’ 등은 별도 산출 문서로 제출하였다.

- 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 하였다.

- 자유 배포가 가능하도록 이용 허락 계약 체결(비용 처리 포함)

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

저작권재산권 이용허락 계약서

저작자 및 저작권 이용허락자 저작자_ (이하 "권리자"이라 함)와 저작권 이용자 주식회사 케이엘큐브(이하 "이용자"이라 함)은 아래 저작물 2022년 한국어-한국수어 병렬 말뭉치 구축을 위한 데이터 수집 및 촬영, 라벨링 용역을 위한 주제, 시나리오, 스크립트 데이터에 관한 저작권재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

[그림 II-4] 저작 재산권 이용 동의 계약 예시

- 저작권 관련 법적 검토 확인
 - 저작권법에서 보호하고 있는 저작물은 인간의 사상이나 감정을 표현한 창작물을 의미하였다. (저작권법 제2조 제1호)
 - 따라서 수어 제공자들이 창작하는 수어 교육용 문장들의 경우에도, 그 안에 저작권법상 보호하는 ‘인간의 사상이나 감정’의 ‘창작적 표현’이 이루어졌다면 저작물로 인정되어 저작권을 보호받을 수 있다.
 - 위 저작물성이 인정되는 해당 문장을 사용하고자 할 경우, 저작권자와 저작권재산권 이용 허락 또는 양도 관련 별도의 계약을 체결하여 해당 저작물 사용에 대한 적법한 권리 또는 권한을 취득하였다.

(5) 한국어 문장 추가 정보 기입 방안

- 동적 숫자 표기법
 - 수어의 특성상 수치를 나타내는 표현을 크게 n과 d로 구분하여 단순 숫자, 시, 시간, 날짜, 나이를 명확히 구분하여 별도의 표기법으로 한국어 문장에 표기하였다.
 - 위의 표기는 수집된 한국어 문장을 한국수어로 번역하는 과정에 수연가의 이해를 돕는 목적으로 더욱 정확한 한국수어 제작과 영상 제작을 목적으로 하였다.

표기법	구분	정의	표기 예시
n	숫자	물건의 개수 등 단순 숫자에 적용	방금 주문한 제품을 n:두개 더 구매하고 싶어요
d	시	몇시 몇분 등 특정 시간 표현에 적용	어제 저녁 d:사:일곱시 삼십분에 저희 매장에 방문하셨습니다
	시간	몇 시간 등 일정 시간대 표현에 적용	해당 내용을 전달받기까지 d:시간:세시간 정도 걸린것 같아요
	날짜	몇월 며칠 등 월, 일 표현에 적용	우유의 유통기한이 여기에 적힌 d:날짜:사월이십일 이 맞는건가요
	나이	나이의 표현에 적용	우리 아이가 올해 d:나이:아홉살 인데 이 옷이 맞을까요

[그림 II-5] 동적 숫자 표기법 예시

○ 동음이의어 표기법

- 한국어 문장에는 다양한 동음이의어가 존재하는데 해당 문장에서 표현된 주어나 동사, 목적어 등이 수어로 번역되는 과정에서 오역이 발생할 경우를 대비하여 크라우드 작업자들이 발견하는 다양한 단어에 대해 구분 및 정의하여 해당 문장에 쓰인 단어에 간단한 구분 값을 표기하였다.

※ 동음이의어의 표시는 크라우드 작업자들이 공유하여 일관성을 유지하였다.

- 이를 통해 수어 번역가는 해당 단어에 대한 판단을 최소화할 수 있고 더욱 양질의 한국수어 번역 및 목표 데이터 수량 수집이 가능하다.

단어	구분	정의	표기 예시
배	신체	신체의 배를 뜻함	어제 저녁부터 계속해서 배(신체) 가 아파서요
	과일	과일의 배를 뜻함	과일 대부분 좋아하지만 배(과일) 는 싫어해요
	배수	두배, 세배 등의 배를 뜻함	이번달 전기요금에 지난달의 몇 배(배수) 인가요
	교통	보트, 요트, 유람선 등의 배를 뜻함	이 배(교통) 를 타면 목적지까지 더 빨리 갈 수 있네요
다리	신체	신체의 다리를 뜻함	어서 빨리 다리(신체) 가 나아야갈 수 있을것 같아요
	교량	시설물 다리를 뜻함	이 섬에서 육지로 갈 수 있는 다리(교량) 가 곧 생길 거라고 하네요

[그림 II-6] 동음이의어 표기법 예시

2) 한국수어 변환 및 수어 제공

(1) 한국수어 변환 기본 원칙

- 경제성 : 한국어 문장이 내포한 의미를 변경, 왜곡하지 않는 수준에서 간단명료하게 표현되는 수어의 언어적 특성을 반영하였다.
- 수어 문법 : 문장의 가장 중요한 단어가 마지막에 배치되는 한국수어 어순 등 수어 문법을 정상적으로 적용하였다.
- 보편성 : 모든 농인들이 보편적으로 사용하고 직관으로 이해할 수 있는 수어 단어로 구성하였다.
- 충실성 : 한국어 문장을 직역하여 문장 대응형 문장이 되지 않도록 한국수어의 대화형 한국수어가 되도록 구성하였다.
- 고유명사 작성 세부 지침 : 지역명은 실제 거주민 혹은 지역 수어센터에 문의하여 다양한 표현을 작성하되, 보편성에 어긋나는 지역에 대해서는 지화를 표기하였다. 신조어 혹은 전문 용어와 같이 수어로 명확하게 수립되어 있지 않은 고유명사는 의미적으로 해석한 한국수어로 구성하고 지화 표현을 지양하였다.

(2) 한국수어 번역 지침

가) 1차 번역 지침

- 한국어 문장은 전문 수어 통역사를 통해 경제성, 수어 문법, 보편성을 원칙으로 한국수어를 생성하였다.
- 한국농아인협회에서 자격 및 역량이 검증된 수어 통역사들을 통해 번역을 진행하였다.
- 한국어 문장에 대응하는 적절한 수어 표현이 없는 경우 지문자를 활용하였다.

나) 2차 번역 지침

- 번역된 문장을 수어 통역사 간 교차 검수 진행하였다.
- 경제성, 수어 문법, 보편성 특성이 미반영 되거나 의미상 오역이 있다고 판단되는 경우 재작업을 진행하였다.
- 통역사 간 번역 기준으로 이견 발생 시 품질 관리 실무책임자에게 보고하고, 책임자는 수어 번역 기준을 평가 및 판단한 후 해당 내용을 통역사 및 검수관 품질 교육에 반영하였다.

다) 한국수어 검수 기준

한국수어 검수 기준	
1	문장의 의미가 누락 및 왜곡되지 않도록 번역되었는가?
2	불필요한 수어 표현이 포함되지 않도록 경제적으로 번역되었는가?
3	한국수어 어순 등 한국수어 문법이 정상적으로 적용되었는가?
4	번역문은 일반 농인들이 보편적으로 사용하고 직관적으로 이해할 수 있는 수어 단어로 구성되어있는가?

[표 II-5] 한국수어 검수 기준

(3) 수어 영상 현장 검수자 지침

- 수어 영상은 촬영 현장에서 수어 검수자를 통해 실시간 검수를 진행하였다.
- 검수자는 수어 특성 반영 여부 및 오역 여부를 검수 후 촬영을 진행하였다.
- 촬영 후 해당 한국수어에 대한 수어 제공자의 의견을 수집하여 품질 관리를 진행하였다.

수어 영상 현장 검수 기준	
1	수어 제공자는 배경과 명확히 구분 가능한 검은색 복장 규정을 준수하였는가?
2	수어 제공자는 비수지 표현이 명확히 보이는 머리 모양 규정을 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 정상적으로 표현되었는가?
7	문장의 가장 중요한 단어가 마지막에 배치되는 한국수어 어순 등 수어 문법의 정상적으로 적용되었는가?
8	농인들이 보편적으로 사용하고 직관적으로 이해할 수 있는 수어 표현을 사용하였는가?
9	지시 수어가 명확하게 표현되었는가?
10	필요에 따라 공간동사, 일치동사 분류사를 활용하였는가?
11	고유지명, 명사 표현 등 지문자 사용이 적절하게 이루어졌는가?
12	불필요한 동작 또는 비수지 표현이 포함되었는가?
13	발화 속도가 적절하였는가?

[표 II-6] 수어 영상 현장 검수 기준

(4) 수어 영상 2차 검수자 지침

- 촬영된 수어 영상은 수어 전문 검수관을 통해 2차 검수를 진행하였다.
- 수어 영상 현장 검수자는 번역된 한국수어에 알맞은 수어가 제공되었는지, 데이터 라벨링을 위한 데이터로서 이상 유무를 판단하였다.

수어 영상 2차 검수 기준	
1	영상>한국수어>한국어 문장 순으로 확인 후 상이할 경우 (번역 오류)
2	비수지 누락 및 애매하거나 틀렸을 경우
3	지화 및 숫자가 틀린 경우
4	일치동사가 틀렸을 경우
5	과축약 또는 과해석(한국어 문장에 없는 표현 사용)되었을 경우
6	수어가 이해되지 않거나 표현이 매끄럽지 않은 경우

[표 II-7] 수어 영상 2차 검수 기준

3) 수어 영상 촬영

가) 수어 영상 데이터 획득

- 데이터 획득 환경
 - 키포인트 라벨링 단계 요구사항
 - 추출 모델을 이용한 인체 특징점 좌표 추출을 위해서는 동영상을 구성하는 이미지 프레임이 노이즈 없이 선명하게 하였다.
 - 촬영 환경(스튜디오 환경) 및 기법은 12만 문장 영상 모두 동일하게 유지하여 데이터에 주는 영향을 최소화하였다.
 - 데이터 일관성을 위해 피사체인 수어 제공자의 복장은 단정하고 일정하게 유지되어야 하며 머리 모양은 비수지 정보를 위해 얼굴을 가리지 않도록 하였다.

나) 촬영환경 및 방식

○ 촬영 환경 기준 수립



[그림 II-7] 수어 표현범위를 고려한 촬영 화면 구성

○ 촬영 환경 구성

- 다양한 수어 제공자의 신장 및 체구를 바탕으로 카메라 높이는 모델의 키의 -35cm를 기준으로 하되, 모델의 손 길이에 따라서 세부 조정을 진행했다(줌 기능 사용하지 않음). 수어 제공자 위치는 크로마키 앞 0.6m, 카메라 거리는 1.6m, 총 3개의 지속광 조명을 배치하여 촬영 환경을 구성하고 조명은 손 아래 그림자 방지를 위해 모델 정면 1m 하단에 위치, 좌우 그림자 방지를 위해 좌우 측 1개씩 배치하여 작업하였다. 좌측 조명은 1.9m 높이, 모델 45도 측면, 우측 조명은 1.6m 높이, 모델 기준 45도 측면에 배치하였다.

○ 촬영 방식

- 촬영된 수어 영상은 키포인트 라벨링에 활용되며 일반적으로 키포인트 라벨링 작업은 효율을 높이기 위해 키포인트 추출 모델을 사용한 반자동 라벨링 방식을 취하며 추출된 키포인트의 정확도에 비례하여 키포인트 라벨링 작업 효율이 결정되므로 촬영 세부 요소를 추출, 키포인트 정확도를 기반으로 최적화하였다.

No	속성	세팅 값
1	fps	30
2	셔터스피드	1/2,000
3	iso(gain)	400 (6)
4	조리개값	2.8
5	화이트밸런스	6,500k
6	조명밝기	58,000(LUX 1M) x 3

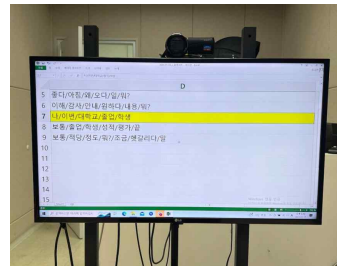
[표 II-8] 촬영 장비 셋팅 값

○ 촬영 환경 관리 및 통제

- 영상 수집 관리자는 촬영 환경 정보(수어 제공자 정보, 촬영날짜, 촬영 시간, 촬영 장소, 촬영 담당자, 카메라 세팅 값 등)를 스튜디오별 촬영일지로 기록/관리하며 모든 영상이 같은 환경에서 촬영되도록 통합 관리, 정제 및 라벨링 프로세스에서 촬영 환경 및 영상 품질 피드백을 받아 오류 발견 시 재촬영을 진행하였다.

○ 한국어 문장의 영향 최소화를 위한 프롬프트 사용 환경 지침

- 수어 모델에게는 한국어 문장의 영향을 최소화하여 수어를 제공할 수 있도록 프롬프트에는 단어만 제공하였다.



[그림 II-8] 수어 제공 프롬프트

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

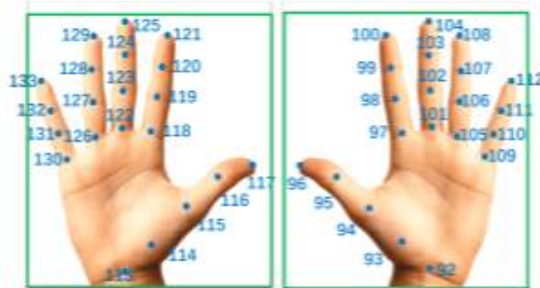
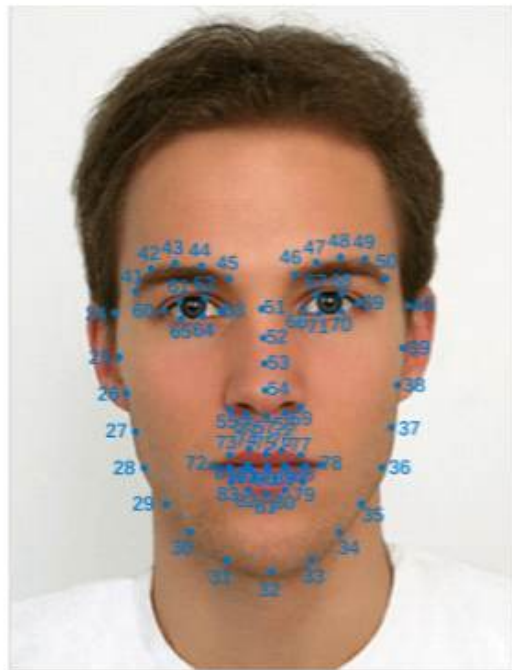
[표 II-9] 수어 영상 스튜디오 촬영 환경

○ 데이터 정제

- 인코딩 과정 없이 휴지 구간 편집 기능을 지원하는 어도비 프리미어를 사용하며 스튜디오별 영상 편집 가능한 PC 또는 노트북을 배치하고 문장 단위로 영상 촬영 종료 직후 진행하였다.
- 수어 구현 전후 휴지 구간은 0.5~2초로 일괄 편집하였다.
- 정제 결과물을 실시간 확인 및 검수하여 재촬영 필요 문장을 정리하였다.

2. 키포인트 라벨링 구축

1) 키포인트 위치



[그림 II-9] COCO Wholebody dataset 키포인트

본 사업은 COCO Wholebody Dataset의 인체 키포인트를 활용하였다(COCO(2017)). 몸 12개, 얼굴 68개, 손 42개로 총 122개가 위와 같은 위치로 구성되며 아래 표는 라벨링해야 하는 키포인트 숫자별 인체 부위 항목 명칭이다.

키포인트 세부 분류 (1/3)		
대분류	세부 분류	인체 위치
몸 (1~12)	1	코끝
	2	오른쪽 눈 동공
	3	왼쪽 눈 동공
	4	왼쪽 귀
	5	오른쪽 귀
	6	왼쪽 어깨
	7	오른쪽 어깨
	8	왼손 팔오금
	9	오른손 팔오금
	10	왼손 손목
	11	오른손 손목
	12	왼쪽 골반

[표 II-10] 키포인트 몸 분류

키포인트 세부 분류 (2/3)		
대분류	세부 분류	인체 위치
얼굴 (24~91)	24~40	턱선
	41~45	오른쪽 눈썹
	46~50	왼쪽 눈썹
	51~54	코대
	55~59	코볼
	60~65	오른쪽 눈
	66~71	왼쪽 눈
	72~78	윗입술 윗선

	79-83	아랫입술 아랫선
	84-88	윗입술 아랫선
	89-91	아랫입술 윗선
왼손 (92~112)	92	손목 주름 중앙
	93	엄지 기부 중앙
	94	엄지 첫 번째 마디
	95	엄지 두 번째 마디
	96	엄지 손끝
	97	검지 첫 번째 마디
	98	검지 두 번째 마디
	99	검지 세 번째 마디
	100	검지 손끝
	101	중지 첫 번째 마디
	102	중지 두 번째 마디
	103	중지 세 번째 마디
	104	중지 손끝
	105	약지 첫 번째 마디
	106	약지 두 번째 마디
	107	약지 세 번째 마디
	108	약지 손끝
	109	소지 첫 번째 마디
	110	소지 두 번째 마디
	111	소지 세 번째 마디
112	소지 손끝	

[표 II-11] 키포인트 얼굴/손(왼손) 분류

키포인트 세부 분류 (3/3)		
대분류	세부 분류	인체 위치
오른손 (113~133)	113	손목 주름 중앙
	114	엄지 기부 중앙
	115	엄지 첫 번째 마디
	116	엄지 두 번째 마디
	117	엄지 손끝
	118	검지 첫 번째 마디
	119	검지 두 번째 마디
	120	검지 세 번째 마디
	121	검지 손끝
	122	중지 첫 번째 마디
	123	중지 두 번째 마디
	124	중지 세 번째 마디
	125	중지 손끝
	126	약지 첫 번째 마디
	127	약지 두 번째 마디
	128	약지 세 번째 마디
	129	약지 손끝
	130	소지 첫 번째 마디
	131	소지 두 번째 마디
	132	소지 세 번째 마디
133	소지 손끝	

[표 II -12] 키포인트 오른손 분류

2) 키포인트 라벨링 기본 방법론

- 국립국어원 한국어-한국수어 병렬 말뭉치 수어 영상 키포인트 라벨링은 키포인트 추출 모델을 활용한 반자동 라벨링 방법을 사용하였다.
- 키포인트 라벨링 작업자는 추출된 키포인트를 영상 프레임별로 대조하여 위치가 일치하지 않는 키포인트를 모두 수정하였다.

- 영상은 주로 상반신만 촬영되며 화면상 등장하지 않는 하체 키포인트, 발바닥 키포인트 등은 라벨링하지 않았다.
- 작업자 공용 폴더에 있는 ‘키포인트 라벨링 가이드 영상’에 따라 작업하며 키포인트 검수 기준의 모든 항목을 충족하도록 작업하였다.
- 영상의 노이즈나 수형의 화면 이탈 등 비정상적인 영상 파일은 바로 보고하여 정상적인 영상을 받아 작업할 수 있도록 하였다.

3) 키포인트 저작도구 사용

- 저작 도구는 ‘AITOK-키포인트 에디터’를 사용하였다.
- 사용자는 케이엘큐브를 통해 NAS 및 저작도구의 ID/PW를 부여받아 사용 권한을 얻었다.
- 작업자는 aitoksmart.iptime.org:5002로 로그인하여 공용 폴더에서 저작 도구를 다운받을 수 있으며, ‘작업대기’, ‘작업완료’ 개인 폴더에는 각각의 작업자에게 할당된 영상들을 업로드하였다.
- 다운받은 저작 도구는 부여받은 전용 ID/PW로 로그인해야 사용이 가능하다.



[그림 II-10] AITOK - 키포인트 에디터

- 작업자는 공용 폴더에 있는 ‘키포인트 라벨링 가이드 영상’을 숙지하고 가이드에 따라 라벨링 작업을 진행 후 생성된 json 파일을 ‘작업완료’ 폴더에 업로드하는 것으로 1차적인 라벨링 작업은 완료된다.
- 업로드된 키포인트 json 파일은 라벨링 검수팀에서 검수한 후 라벨링 오류가 없을 시 최종 완료되며, 오류가 있을 시 작업자에게 재작업을 요청했다. 재작업 시 검수 확인을 받아야 최종 완료가능하며 검수 불합격 상태에서는 작업 보상을 지급하지 않았다.
- 키포인트 에디터는 기본적으로 마우스 드래그 기능을 사용하여 키포인트를 수정하며 작업 효율 향상을 위해 아래와 같은 단축키 기능을 제공하였다.

저작도구 단축키		
분류	단축키	기능
키포인트 라벨링	1-5	왼손 손가락 on/off
	6-0	왼손 손가락 on/off
	A	컨트롤 UI on/off
	S	수정 내용 저장
	U	수정 내용 적용 취소
	C	왼손-오른손 키포인트 변경
	L	왼손 키포인트 이전 프레임 가져오기
	R	오른손 키포인트 이전 프레임 가져오기
영상제어	G	손 그룹 이동
	P	영상 시작/정지
	Z	영상 줌 인/아웃
	<	이전 프레임 이동
>	다음 프레임 이동	

[표 II-13] 저작 도구 단축키

3. 형태소 라벨링 구축

1) 주석 기본 원칙

- 데이터 구축을 위한 주석은 한국수어 영상 주석 프로그램을 사용하였다.
- 본 사업의 주석은 의미를 가지는 손 움직임과 얼굴 표정(비수지)에 이름을 붙이는 것을 기본으로 하였다.
- 본 사업의 주제는 ‘인공지능의 수어 영상 데이터 딥러닝’이며, 심화 단계 이전의 ‘초-중급 수준의 수어와 비수지 습득’을 목표로 하였다.
- 인공지능의 한국수어 학습도가 기초 단계인 것을 고려하여, 학습에 혼동이 없도록 반드시 정해진 규칙에 따라 주석 입력이 이루어지도록 하였다.

(1) 토큰과 분절

- ‘토큰’이란 수어 단어 입력 정보를 의미하며 제시된 한국어 문장을 기준으로 변환된 한국수어를 토큰으로 자동화 시스템에 의해 미리 입력된다.
- ‘분절’이란 토큰의 위치와 길이 등을 의미하며, 수어 발화에 맞춰 조정 작업을 하였다.

(2) 글로스(라벨)

- 글로스는 수어에 라벨을 명명하는 것으로, 수어의 의미나 번역은 아니나 해당 수어의 의미와 연관성이 있다.
- 글로스는 ‘한국어 문장’의 형태를 가진다. [예: 학교 / 끝]
- 전체적으로 수어의 형태가 같은 경우, 수동의 강약이나 공간상 대소 차이가 있더라도 같은 글로스로 주석하였다.
[예: 비/폭우 → 비내리다, 지진/강진 → 지진, 눈/폭설 → 눈내리다]
- 수어의 형태는 다르나 의미가 같은 경우와 한국어 문장 동음이의어의 경우에도 글로스는 같은 한국어 단어의 형태를 가진다.

- 다른 사업자가 찾기 쉽도록 가장 기본적이고 일반적인 단어를 사용하였다.
- 핵심 의미를 최대한 단순하게 정하였다.
- 띄어쓰기를 사용하지 않았다.
- 한국어 문장으로 토큰이 생성된 감탄사(환정, 호응)의 경우 해당 토큰의 길이를 조절하는 방식으로 일반적인 주석 지침과 동일하게 진행했다. 다만, 한국수어 또는 토큰이 생성되지 않은 상태로 모델이 문장의 흐름에 따라 자체적으로 감탄사를 사용한 경우에는 Mmo(마우딩) 비수지를 이용하여 반영하였다.
- 글로스에는 필요시 일치동사의 의미를 표현하기 위해 화자와 청자를 지칭하거나 한정하는 문구가 들어갈 수 있다.

(3) 일치동사

- 일치동사
 - 일치동사에서 나타나는 인칭 정보를 입력하였다.
 - 입력의 기준은 인칭의 기준으로, ‘나(화자)=1’, ‘너(화자 입장에서 상대방)=2’ 로 입력하였다.
 - 내가 너를 도와준다.
1 2
 - 네가 나를 도와준다.
2 1
- 일치동사 주석 방법
 - 토큰 정보 → 일치동사
 - 토큰을 선택하면 일치동사를 입력할 수 있는 메뉴로 이동하였다.

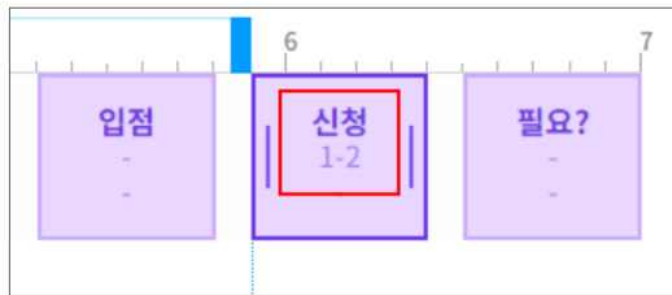


[그림 II-11] 일치동사 주석 방법 1

- 일치동사 선택하기
 - [선택▼]을 클릭하면 일치동사 정보를 선택할 수 있다.

[그림 II-12] 일치동사 주석 방법 2



- 일치동사 입력 확인하기
 - 일치동사를 입력하면 아래 사진과 같이 토큰에 정보가 생성된다.



[그림 II-13] 일치동사 주석 방법 3



- o 일치동사 주석이 필요한 단어들
 - 접수(하다), 신청(하다), 말(하다), 방문(하다), 도움, 알려주다, 참여(하다), 질문, 주문(하다) 등

- 예시1) 돕다(도움, 도와주다 등……)



 <div data-bbox="360 636 676 898" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>돕다 <u>1-2</u> -</p> </div>	 <div data-bbox="924 636 1240 898" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>도움 <u>2-1</u> -</p> </div>
<p>돕다(<u>내가→너를</u>) / [돕다(<u>1→2</u>)]</p>	<p>도움(<u>너가→나를</u>) / [도움(<u>2→1</u>)]</p>

[표 II-14] 일치동사 주석 예시 1

- 예시2) 말하다(말해준다, 대화하다, 말씀 등……)

 <div data-bbox="360 1433 676 1695" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>말(대화)하다 <u>1-2</u> -</p> </div>	 <div data-bbox="924 1433 1240 1695" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>말(대화)하다 <u>2-1</u> -</p> </div>
<p>말하다(<u>내가→너에게</u>) / [말하다(<u>1→2</u>)]</p>	<p>말하다(<u>너가→나에게</u>) / [말하다(<u>2→1</u>)]</p>

[표 II-15] 일치동사 주석 예시 2

 <div data-bbox="352 584 687 835" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>말(대화)해주다 <u>1-2</u> .</p> </div>	 <div data-bbox="911 584 1246 835" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>말(대화)해주다 <u>2-1</u> .</p> </div>
<p>말해주다(<u>내가→너에게</u>) / [말해주다(<u>1→2</u>)] 말해주다(<u>너가→나에게</u>) / [말해주다(<u>2→1</u>)]</p>	

[표 II -16] 일치동사 주석 예시 3

- 예시3) 질문(질문하다, 물어보다, 문의하다 등……)

		<div data-bbox="959 1106 1206 1361" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>질문 <u>1-2</u> .</p> </div>
<p>질문(<u>내가→너에게</u>) / [질문(<u>1→2</u>)]</p>		
		<div data-bbox="959 1449 1206 1704" style="background-color: #e6e6fa; padding: 10px; text-align: center;"> <p>질문 <u>2-1</u> .</p> </div>
<p>질문(<u>너가→나에게</u>) / [질문(<u>2→1</u>)]</p>		

[표 II -17] 일치동사 주석 예시 4

- 예시4) 신청(신청하다, 접수 등……)

신청(내가→너에게) / [신청(1→2)]		
질문(너가→나에게) / [질문(2→1)]		

[표 II -18] 일치동사 주석 예시 5

- 예시5) 방문(내방 등……)

방문(내가→너에게) / [방문(1→2)]		
방문(너가→나에게) / [방문(2→1)]		

[표 II -19] 일치동사 주석 예시 6

(4) 생산적 수어

- 생산적 수어: 일반적인 수어 단어 표현에서 벗어나, 수어의 도상성을 활용하여 특정 상황 및 문장의 맥락 속에서 사용되는 표현을 의미한다.
- 수어의 도상성(圖像性, iconicity): 시각 언어라는 특수성에서 비롯된 것으로, 시각적 표현과 공간을 활용하여 특정 상황을 해설하기 위해 마치 그림처럼 표현하는 것을 의미한다. 의미와 형태가 일정 부분 관습적으로 연결되어 있다는 점에서 비언어적인 시각-제스처와는 구별된다.
- 현재 AI의 학습 능력이 생산적 수어를 인지하기에는 한계가 있어 별도의 추가 작업은 진행하지 않고 생산적 수어의 경우 타입 입력만 진행하였다.
- 고정 수어 사용 후 도상성을 표현한 경우, 고정 수어 시작부터 생산적 수어 끝까지 모두 생산적 수어로 입력하였다. 이때 타입명은 고정 수어를 기본으로 사용하였다.
- 본 사업에서 주석하는 생산적 수어는 다음과 같다.
 - 상황 묘사/형태 변화: 상황을 표현할 때 수동이 다르다는 점에서 차별성을 가진다.
 - 정도 묘사: 수어의 강약 차이, 수어의 크기 차이에 따른 차별성이 있다.
 - 포인팅: 1지 혹은 편 손바닥(9 수형)을 사용하여 특정 대상을 지시하는 경우를 말한다.

분류		문장 예시	글로스 예시
상황 묘사 형태 변화		차가 달리다. 차가 굽은 길을 달리다. 차가 비탈길을 오르다.	자동차1(오른쪽방향 돌면서)
정도 묘사	강약 차이	바람이 분다. 바람이 강하게 분다.	바람(거세게)
	크기 차이	작은 상자 큰 상자	상자(큰 상자)
포인팅		우세: 지시(1지 사용) 비우세: 산 수형	우세: 이것 비우세: 산

[표 II-20] 생산적 수어 예시

(5) 비수지 신호

- 모델이 표현하는 모든 것을 비수지 신호로 입력하는 것이 아니라 수어 문법과 일치하거나 특정한 의미를 내포하고 있는 비수지 신호만 입력하였다. 본 사업에서 주석한 비수지 신호는 다음과 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 꼭 다물기	참다, 기다리다, 열심히 하다, 안녕하세요 등
4	Hno	고개 끄덕임	안녕하세요, 부탁, 필요, 나열, 문장종결 등
5	Hs	고개를 좌우로 흔들기	불가능, 못하다, 안되다
6	Ebf	눈썹 찌푸리기	심하다, 위협하다, 안 돼
7	Ci	볼 부풀리기	상황 묘사
8	Ebu	눈썹을 위로 올림	의문문, 놀람, 강조

[표 II-21] 비수지 신호 용어 정리

- Mo1

- 특정 상황에 짧고 강하게 입을 벌려서 마!, 파! 표현을 하는 것을 의미하였다. 이때 단순히 입을 동그랗게 벌리는 것을 착각해서 기입하지 않도록 주의하였다. 비수지가 나타나는 시간에만 입력한다(“파!”를 표현하고 입을 벌린 채로 유지한다고 해서 Mo1을 길게 기입하는 것이 아니라, 해당 비수지가 나온 순간만을 주석한다는 의미이다)

[예시] 가능(파!), 끝나다(파!), 발생(파!), 무너지다(파!) 등

- Mmo

- 모델이 입을 벌려 발음하는 비수지를 의미한다. 습관적으로 표현하거나 단어마다 말하는 입 모양은 주석 입력하지 않고 지문(숫)자와 함께 사용하는 모양(마우딩)만 기입하였다. 단 문장 부호는 포함하지 않고 모델이 표현하는 입모양에 대응하는 단어만 기입하였다.

- Mctr

- 입꼬리가 움직이며 입을 꼭 다문다. 입을 꼭 다문 채로 미소짓는

입술을 떼올리면 되며 입을 오므리는 것은 Mctr이 아니므로 주의하였다.

[예시] 참다, 기다리다, 열심히 하다. 안녕하세요 등

- Hno

- 고개를 끄덕이는 비수지다. 모델이 단어마다 고개를 끄덕이거나 의미 없이 끄덕이는 건 주석하지 않았다.

[예시] 부탁하다, 필요하다, 나열(A, B, C), 문장종결(~이다.) 등

- Hs

- 머리를 좌우로 흔든다. 보통 부정어에 사용된다.

[예시] 부정어(불가능, 못하다, 안되다) 등

- Ebf

- 눈썹을 찌푸린다. 보통 부정어나 강조에 사용된다.

[예시] 위험하다, 안되다, 심하다 등

- Ci

- 불을 부풀린다.

[예시] 태풍, 무겁다(강조), 차가 밀리다, 살찌다 등

- Ebu

- 눈썹을 위로 올린다. 보통 의문문이나 놀람, 강조에 사용된다.

[예시] 의문문, 놀람 표현, 강조 등

(6) 주석 층렬

○ 본 사업의 주석 층렬은 다음과 같이 구성된다.

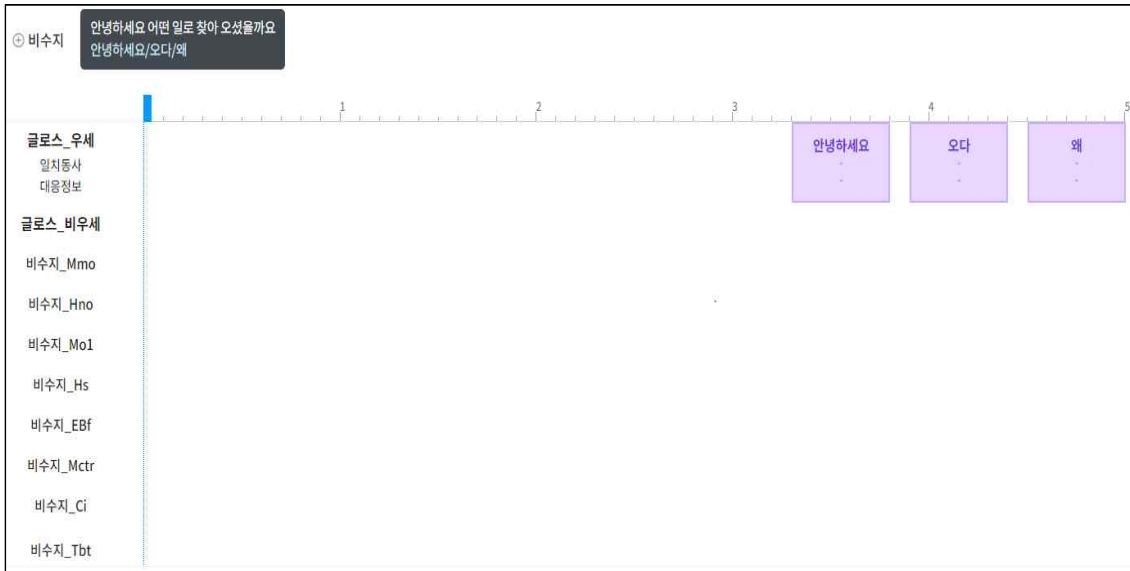
- 한국어 문장 : 제시되는 수어 영상의 한국어 형태이며, 작업 페이지를 커면 자동 생성되어 있다.
- 수어 문장 : 제시된 한국어 문장을 수어로 변환한 형태이며, 수어 문장을 기준으로 토큰이 자동 생성되어 있다.

○ 우세 정보

- 클로스_우세 : 우세손(보통 오른손)의 클로스 정보
- 공간 정보_우세 : 주석 입력 전 좌표 입력을 통한 우세손의 공간 정보 수집

(원: IT기술로 자동 인식하여 우세손의 공간 정보 수집)

- 일치동사_우세 : 일치동사의 주어와 목적어의 인칭
- 대응 정보_우세 : 수어 토큰이 의미하는 한국어 정보



[그림 II-14] 수어 주석 도구 화면

- 비우세정보
 - 글로스_비우세: 비우세손(보통 왼손)의 글로스 정보
 - 공간 정보_비우세: 주석 입력 전 좌표 입력을 통한 비우세손의 공간 정보 수집
(원: IT기술로 자동 인식하여 비우세손의 공간 정보 수집)
- 비수지 신호
 - Mmo : 한국어 문장의 단어가 입 모양으로 나타나는 것을 의미
 - Hno : 고개 끄덕임을 의미
 - Mo1 : 마우딩과 다르게 특정 상황에 입모양이 동반되는 것을 의미
 - Hs : 고개를 좌우로 흔드는 것을 의미
 - EBf : 눈썹을 찌푸리는 것을 의미
 - Mctr : 입꼬리가 움직이며 입을 꼭 다무는 것을 의미
 - Ci : 불을 부풀리는 것을 의미
 - Ebu : 눈썹을 위로 올림

2) 주석 세부 지침

(1) 토큰



준비하기

치기

유지하기

내리기

[그림 II-15] 수어 토큰 글로스 기준

- 한 토큰 안에 두 개 이상의 글로스를 입력할 수 없다.
- ‘치기’와 ‘유지하기’의 부분이 토큰 안에 포함되어야 한다.
- ‘준비하기’와 ‘내리기’는 토큰에 포함되지 않는다.
- ‘준비하기’와 ‘내리기’의 일부가 토큰에 포함되는 오차 범위는 ‘분절’ 항목을 참조하였다.
- 토큰이 미리 입력되어 있으므로 합성어와 같은 단어를 찾아서 입력할 필요 없이 미리 입력된 토큰대로 시간 조절만 하였다.
- 토큰과 토큰을 붙여서 입력하지 않고, 사이를 띄어서 입력하였다.

틀린 예	바른 예

[표 II-22] 토큰 분절 예시

- 지문자, 지숫자, 동적숫자(날짜, 시간, 시 등) 수어의 경우 새로 입력하는 것이 아니라 자동화 시스템에 따라 미리 입력된 토큰대로 시간 조절만 하였다. 이때 양손 주석은 확실하게 기입했다.

<div style="background-color: #d1c4e9; padding: 10px; border: 1px solid black; margin-bottom: 5px;">24 - -</div> <div style="background-color: #e0e0e0; padding: 5px; border: 1px solid black;"> </div>	<div style="background-color: #d1c4e9; padding: 10px; border: 1px solid black; margin-bottom: 5px;">24 - -</div> <div style="background-color: #c8e6c9; padding: 10px; border: 1px solid black; margin-bottom: 5px;">24</div> <div style="background-color: #e0e0e0; padding: 5px; border: 1px solid black;"> </div>
틀린 예	바른 예

[표 II-23] 양손 주석 예시

- 양손 수어인데 간혹 우세나 비우세 토큰이 빠진 경우, 혹은 반대로 입력된 경우가 있으므로 주석 작업을 완료하고 검토하도록 하였다.
- 양손 수어의 경우, 우세 토큰과 비우세 토큰의 위치와 분절이 동일해야 한다. 양손 주석을 통해 입력하도록 하였다.

<div style="background-color: #d1c4e9; padding: 10px; border: 1px solid black; margin-bottom: 5px; width: 80px; margin-left: auto; margin-right: auto;">신청 - -</div> <div style="background-color: #c8e6c9; padding: 10px; border: 1px solid black; margin-bottom: 5px; width: 100px; margin-left: auto; margin-right: auto;"> 신청 </div> <div style="background-color: #e0e0e0; padding: 5px; border: 1px solid black;"> </div>	<div style="background-color: #d1c4e9; padding: 10px; border: 1px solid black; margin-bottom: 5px; width: 80px; margin-left: auto; margin-right: auto;">신청 - -</div> <div style="background-color: #c8e6c9; padding: 10px; border: 1px solid black; margin-bottom: 5px; width: 80px; margin-left: auto; margin-right: auto;">신청</div> <div style="background-color: #e0e0e0; padding: 5px; border: 1px solid black;"> </div>
틀린 예	바른 예

[표 II-24] 양손 주석 위치와 분절 예시

- 양손 수어를 심화해서 보면 특정 수어를 지칭하는 표현이 전후로 나올 때가 있어 비우세 주석을 문장 흐름에 맞게 늘려서 입력하였다.



[그림 II-16] 비우세 주석과 문장 흐름

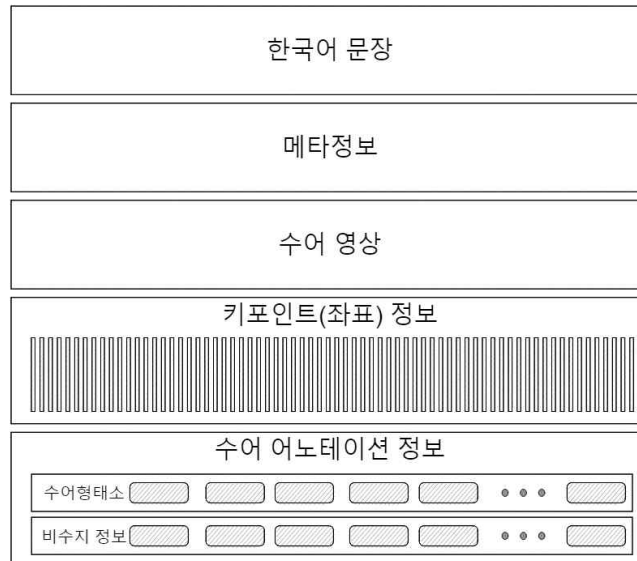
(2) 분절

- 이동이 없는 수어는 최초 수형이 발견된 때가 수어의 시작이다.
- 이동이 있는 수어는 최초 수동이 발견된 때가 수어의 시작이다.
- 수어의 최종 수형과 최종 수동이 최소 한 손에서 사라진 때가 수어의 끝이다.

- 수어가 연속으로 이어지는 경우, 이전 단계에 따라 이전 수어의 최종 수형 혹은 최종 수동이 사라진 이후 다음 수어의 토큰을 입력하였다.
- (3) 대응 정보명 오류
- 토큰에 입력되어 있는 대응 정보명에 오류가 있는 경우 주석 관리자에게 오류 사항을 전달하였다.

4. JSON 파일 구축

1) 병렬 말뭉치 데이터 구조



[그림 II-17] 영상 단위 병렬 말뭉치 데이터 구조

- 한국어-한국수어 병렬 말뭉치 사업은 수어를 영상 기반으로 인식하여 한국어로 번역(voice to text)하는 응용 서비스의 개발에 필요한 학습데이터의 구축을 목표로 하며, 이를 위해 한국어 문장, 수어 영상, 형태소 단위 어노테이션 정보(수어 형태소, 비수지 표현) 및 프레임별 키폰트 데이터, 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, 약 3,600만 개의 키포인트 데이터셋을 포함한 12만 개의 학습데이터셋을 생성, 해당 데이터셋은 영상과 함께 json파일로 제공되면 포맷 종류 및 영상의 명명 규칙은 다음과 같다.

한국어-한국수어 병렬 말뭉치 구축사업 (Json 구조정의서)

```

{
  "id" : "SLICCPAKOKSL2200000006",   작업 아이디 (파일명)
  "opertor" : "cubeworker6", //      작업자
  "krlgg_sntenc" : {
    "koreanText" : "여기서 바로 해결하기는 힘들어 보입니다", 한국어
    "realm" : "생활", (분야)
    "thema" : "민원/행정" (주제)
  },
  "sign_lang_sntenc" : "해결/어렵다", 한국수어
  "sign_lang_trnslator" : "MJ", 한국수어 번역가 정보
  "sign_script" : { 수어주석 정보
    "sign_gestures_strong" : [ { 글로스 우세
      "start" : 2.567, (시작시간)
      "end" : 3.337, (종료시간)
      "gloss_id" : "어렵다", (주석명)
      "express" : "s", (주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
      (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있다.)
      "direction" : { (일치동사)
        "source" : "", 에서(인칭)
        "target" : "" 으로(인칭)
      },
      "sentence_loc" : { (대응정보)
        "start" : "", (시작)
        "end" : "" (종료)
      }
    }
  }, {
    "start" : 1.854, (시작시간)
    "end" : 2.463, (종료시간)
    "gloss_id" : "해결", (주석명)
    "express" : "s", (주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
    (동적숫자는 d:시간, d:시, d:나이, d: 날짜로 구성되어 있다.)
  }
}

```

```

“direction“ : {
    “source“ : ““,
    “target“ : ““
},
“sentence_loc“ : {
    “start“ : ““,
    “end“ : ““
}
} ],
“sign_gestures_weak“ : [ {
    “start“ : 1.854,
    “end“ : 2.463,
    “gloss_id“ : “해결“,
    “express“ : “s“,
    “direction“ : null,
    “sentence_loc“ : null
} ]
},
“nms_script“ : {
    “Mmo“ : null,
    “Hno“ : null,
    “Mol“ : null,
    “Hs“ : null,
    “EBf“ : [ {
        “descriptor“ : ““,
        “start“ : 2.567,
        “end“ : 3.067
    } ],
    “Mctr“ : null,
    “Ci“ : null,
    “Ebu“ : null
},
“vido_file_nm“ : “SLICCPAKOKSL2200000006“,
“potogrf“ : {
    “createdTime“ : “2022-11-11 00:00:00“,
    “sentence_ID“ : “SLICCPAKOKSL2200000006“,
    “photographer“ : “B“,
    “location“ : “b“,
    “devide“ : “FDR_AX700“,
    “iris“ : 2.8,
    “Gain“ : 6,
    “WhiteBalance“ : 6500,
    “ShutterSpeed“ : 2000,
    “fps“ : 30.0,
    “width“ : 1920,
    “height“ : 1080,
    “fileSize“ : 2.97,

```

(일치동사)
에서(인칭)
으로(인칭)

(대응정보)
(시작)
(종료)

글로스 비우세
(시작시간)
(종료시간)
(주석명)
(주석타입 s:수어 f: 지화 n: 숫자 d: 동적숫자)
(동적숫자는 d:시간, d:시,d:나이, d: 날짜로 구성되어 있다.)
(일치동사)
(대응정보)

비수지

Mmo : 단어가 입모양으로 나타나는 것
Hno : 고개 끄덕임
Mol : 마우딩과 다르게 특정상황에 입모양이 동반됨
Hs : 고객을 좌우로 흔드는 것
Ebf : 눈썹을 찌푸리는 것
(설명)
(시작시간)
(종료시간)
Mctr : 입꼬리가 올라가면서 입을 다무는 것
Ci : 볼을 부풀리는 것
Ebu : 눈썹을 위로 올리는 것

영상파일명

촬영정보

```

    "length" : "00:00:04:15",
    "format" : "mp4",
    "codec" : "avc1",
    "sl_speaker_id" : "YJ",
    "sl_speaker_age" : 50,
    "sl_speaker_sex" : "Male",
    "sl_speaker_legion" : "Seoul",
    "sl_speaker_hand" : "right_handed "
  },
  "landmarks" : [ {
    "frame" : 1,
    "predictions" : [ {
      "keypoints" : [ 938.7, 241.37, 0.84, 970.74, 208.0, 0.84, 894.33, 213.17, 0.86, 1009.0
9, 237.72, 0.63, 835.06, 254.36, 0.85, 1103.04, 439.73, 0.84, 780.38, 469.57, 0.85, 1170.44, 6
86.99, 0.83, 731.99, 724.62, 0.83, 1193.14, 930.4, 0.81, 725.62, 957.32, 0.77, 1075.98, 969.6,
0.73, 854.01, 976.36, 0.73, 0.0, -36.17, 0.0, 0.0, -36.17, 0.0, 0.0, -36.17, 0.0, 0.0, -36.17, 0.
0, 0.0, -36.17, 0.0, 0.0, -36.17, 0.0, 0.0, -36.17, 0.0, 0.0, -36.17, 0.0, 0.0, -36.17, 0.0, 0.0,
-36.17, 0.0, 838.68, 228.07, 0.88, 841.71, 248.47, 0.9, 846.65, 269.29, 0.88, 852.81, 289.08,
0.88, 861.88, 307.55, 0.9, 876.15, 322.56, 0.88, 894.42, 334.07, 0.88, 914.06, 340.77, 0.84, 9
35.26, 343.05, 0.87, 955.99, 339.79, 0.87, 975.02, 329.67, 0.87, 988.6, 314.06, 0.88, 997.87,
295.97, 0.87, 1002.12, 275.42, 0.89, 1004.75, 254.01, 0.88, 1005.07, 233.29, 0.89, 1003.19, 2
12.9, 0.87, 863.29, 196.65, 0.89, 874.1, 189.55, 0.87, 886.86, 187.09, 0.89, 899.5, 186.92, 0.8
9, 912.94, 187.78, 0.91, 950.23, 186.26, 0.86, 961.44, 183.64, 0.87, 972.35, 182.11, 0.86, 98
3.16, 182.57, 0.86, 993.27, 188.35, 0.85, 934.61, 211.47, 0.87, 936.4, 222.37, 0.89, 938.29, 2
33.46, 0.85, 940.34, 243.48, 0.85, 920.73, 261.17, 0.88, 929.71, 260.59, 0.89, 938.93, 260.3,
0.89, 947.81, 259.88, 0.87, 955.81, 258.52, 0.86, 877.82, 218.59, 0.88, 888.17, 214.28, 0.85,
899.54, 212.73, 0.88, 910.49, 215.57, 0.88, 899.98, 219.28, 0.89, 888.91, 220.54, 0.88, 955.0
3, 212.63, 0.86, 964.27, 208.79, 0.86, 974.55, 208.32, 0.8, 984.56, 211.38, 0.87, 975.26, 215.
31, 0.84, 965.02, 214.57, 0.85, 908.66, 291.66, 0.84, 920.85, 284.31, 0.87, 934.51, 279.91, 0.
87, 940.92, 279.86, 0.88, 947.13, 279.17, 0.92, 958.59, 281.96, 0.98, 968.39, 287.5, 0.9, 960.
35, 292.83, 0.9, 951.5, 296.49, 0.88, 942.11, 298.21, 0.86, 930.04, 298.05, 0.85, 918.64, 295.
69, 0.86, 911.26, 291.17, 0.88, 926.05, 288.26, 0.87, 941.21, 287.12, 0.89, 953.35, 286.75, 0.
88, 965.41, 287.78, 0.9, 953.18, 287.86, 0.88, 940.97, 288.47, 0.88, 926.09, 289.5, 0.85, 119
8.93, 931.09, 0.71, 1172.45, 952.53, 0.76, 1158.58, 991.61, 0.83, 1154.79, 1025.66, 0.83, 115
1.01, 1049.61, 0.67, 1190.1, 1018.09, 0.78, 1176.23, 1055.92, 0.73, 1159.84, 1063.48, 0.87, 1
144.71, 1062.22, 0.79, 1202.71, 1019.35, 0.69, 1192.62, 1060.96, 0.9, 1169.93, 1062.22, 0.86,
1158.58, 1052.13, 0.53, 1211.53, 1014.31, 0.88, 1200.19, 1049.61, 0.67, 1178.75, 1053.4, 0.6
2, 1169.93, 1039.53, 0.44, 1211.53, 1009.26, 0.56, 1200.19, 1035.74, 0.5, 1188.84, 1039.53,
0.52, 1181.27, 1033.22, 0.41, 722.54, 957.42, 0.54, 744.59, 978.55, 0.63, 759.28, 1015.29, 0.7
9, 766.63, 1055.7, 0.85, 772.14, 1082.34, 0.73, 723.46, 1049.27, 0.72, 744.59, 1091.52, 0.65,
760.2, 1101.62, 0.43, 767.55, 1075.91, 0.19, 715.2, 1049.27, 0.55, 737.24, 1086.01, 0.49, 754.

```

키포인트 정보


```

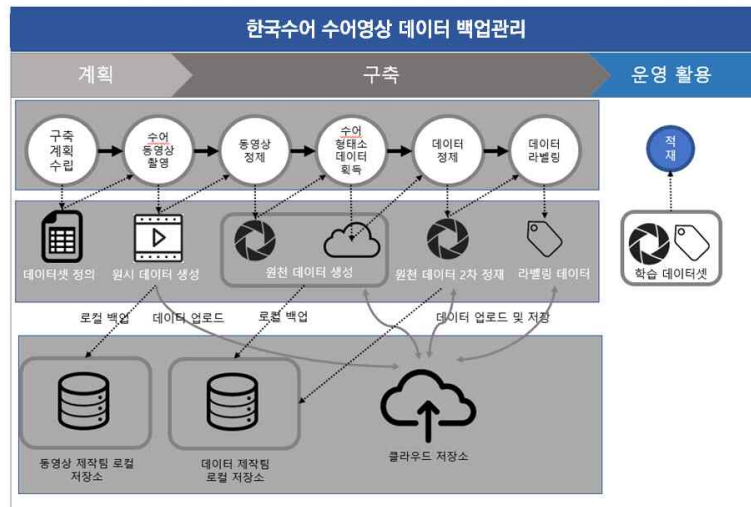
69, 1075.91, 0.31, 760.2, 1061.21, 0.26, 714.28, 1045.6, 0.6, 733.57, 1072.23, 0.57, 751.93, 1
070.4, 0.35, 755.61, 1054.78, 0.22, 716.11, 1040.09, 0.45, 735.4, 1060.29, 0.38, 748.26, 1063.
05, 0.37, 750.1, 1055.7, 0.21 ],
    "hand_pos" : [ 6, 4 ]
  } ]
},
- 이하 생략 -

```

[표 II-25] Json 구조정의서

3) 데이터 백업 관리

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



[그림 II-18] 데이터 백업 관리 프로세스

- 모든 데이터는 제작팀에 의한 1차 로컬 백업을 원칙으로 하였다.
- 1차 로컬 백업 데이터는 백업 완료 후 프로젝트 종료 시점까지 삭제 불가하며 클라우드 파일 서버 이관 데이터에 문제 발생 시 1차 로컬 백업파일은 다시 업로드하였다.
- 데이터 이관은 클라우드 파일 서버를 이용하였다.
- 모든 데이터는 로컬과 클라우드 서버에 이중 보관하였다.

5. 병렬 말뭉치 데이터 검수

1) 검수 항목 및 활동 정의

데이터	검수항목	검수 활동
한국어 문장	데이터 중복 여부	<ul style="list-style-type: none"> 문장 수집 후 문장 간 유사성 및 중복 여부 검수
	편향성 여부	<ul style="list-style-type: none"> 데이터 편향 방지를 위해 수집 세부 카테고리 간 데이터 밸런스 및 분포도 검수
	법·제도 준수	<ul style="list-style-type: none"> 문장 수집 시 저작권 등 관련 법과 규정 준수
수어 영상 및 메타데이터	수어 통역 품질	<ul style="list-style-type: none"> 한국어 문장이 타당한 수어 형태로 번역되었는지 검수 번역된 수어 형태소 시나리오대로 통역사가 촬영되었는지 검수 수어 표현이 영상 범위 안에서 정상적으로 이루어졌는지 검수 수어 통역사의 복장 및 머리 모양 준수 여부 확인
	영상 속성 일치 여부	<ul style="list-style-type: none"> 수집된 메타데이터가 촬영 계획과 일치하는지 검수
	촬영 환경 통제	<ul style="list-style-type: none"> 영상의 fps, 셔터스피드, iso, 화이트 밸런스 설정 일치 여부 검수
	법·제도 준수	<ul style="list-style-type: none"> 수어 통역사의 초상권 동의 여부 확인
형태소, 비수지 정보 데이터	수어 형태소 정보	<ul style="list-style-type: none"> 글로스(우세), 글로스(비우세), 일치동사, 대응 정보 등 형태소 정보 확인
	비수지 정보	<ul style="list-style-type: none"> 입 모양, 고개 움직임, 입꼬리 움직임, 볼 부풀림, 눈썹 움직임 등 비수지 정보 확인
	저작 도구 사용	<ul style="list-style-type: none"> 접속 정보, 문장과 수어 영상 확인, 승인/반려 정보 확인
키폰트 데이터	키폰트별 x,y 좌표 정확성	<ul style="list-style-type: none"> 프레임별 133개 (하반신 제외 123개)의 좌표가 전수 라벨링되어 있는지 검수 라벨링된 좌표가 COCO 데이터 포맷에 설정된 인체 부위와 정확히 일치하는지 검수
	가시성 플래그	<ul style="list-style-type: none"> 키폰트별 가시성 플래그(v값)가 정상적으로 입력되었는지 검수
	저작 도구 사용권	<ul style="list-style-type: none"> 오픈소스 사용권 여부 확인

[표 II -26] 검수 항목 및 활동 정의

2) 문장 데이터 검수

- 수집 대상 목적에 맞게 코드별 분류가 되었는가?
- 개인의 구분을 위하여 부여된 고유한 값 또는 이름을 비식별화하였는가?

항목	<ul style="list-style-type: none"> ▪ 고유 식별 정보(주민 등록 번호, 운전 면허증 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
----	---

[표 II-27] 개인 고유 식별 정보 비식별화 여부 검수 항목

- 개인을 특정할 수 있는 상황인지 판단하여 비식별화하였는지 확인하였다.

항목	<ul style="list-style-type: none"> ▪ 성별, 연령, 국적, 고향, 우편 번호, 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, 부서 명, 직급
----	---

[표 II-28] 개인 특정 가능 정보 비식별화 여부 검수 항목

- 내용에 비윤리적인 내용 또는 혐오 표현이 포함되어 있는지 확인하였다.
- 내용에 외래어 및 로마자, 한자 정제 내용이 포함되었을 경우 어문 규범에 맞는지 확인하였다.
- 오타 및 띄어쓰기, 맞춤법 등을 확인하였다.
- 수집된 문장이 문법적으로 오류가 없는지, 일상용어를 사용하였는지 확인하였다.
- 문장은 4~15어절(평균 9.5) 이내로 구성되었는지 확인하였다.

3) 수어 영상 촬영데이터 검수

○ 수어 영상 촬영데이터(MP4) 검수 절차는 다음과 같다.

수어 영상 촬영데이터 검수	
1	문장 코드에 맞게 파일 분류가 되었는가?
2	문장 코드와 다르거나 누락된 파일은 없는가?
3	문장 앞뒤의 대기시간이 너무 길거나 내용이 잘리지 않게 편집이 되었는가?
4	카메라 앵글은 촬영 환경 구성 지침에 맞게 촬영되었는가?
5	촬영 환경 정보에 대한 누락 또는 오기 표기가 없는가? - 수어 통역사 정보 - 촬영 날짜 - 촬영 시간 - 촬영 장소 - 촬영 담당자 - 카메라 셋팅 값

[표 II -29] 수어 영상 촬영데이터 검수 항목

4) 한국수어(촬영 영상) 검수

- 한국수어 검수는 총 3단계로 진행하며 각 단계별 수어 전문가를 배치하여 아래 한국수어 검수 기준에 따라 중복 검수를 진행하였다.
- 1차 영상 검수자는 ‘확인’ 권한을 가지고 한국수어 검수 기준표에 따라 검수를 진행하며 오류가 발견된 경우 저작 도구 ‘메모’란에 해당 내용을 기재하였다.
- 2차 영상 검수자는 ‘승인’ 과 ‘반려’ 권한을 가지고 1차 영상 검수자가 검수한 문장을 재검수하였다.
- 3차 영상 검수자는 2차 ‘승인’ 과 ‘최종 반려’ 권한을 가지고 2차 영상 검수자가 ‘반려’ 한 문장만 재검토하였다.
- 3차 영상 검수자가 ‘최종 반려’ 한 문장은 수어 촬영 관리자가 확인 후 오류를 수정하여 재촬영 또는 폐기 처리하였다.

수어(촬영 영상) 검수 기준	
1	수어 제공자는 배경과 명확히 구분 가능한 검은색 복장 규정을 준수하였는가?
2	수어 제공자는 비수지 표현이 명확히 보이는 머리 모양 규정을 준수하였는가?
3	수어의 표현이 카메라 앵글 안에서 모두 표현되었는가?
4	수지 표현이 반대 손 또는 팔꿈치 등에 가려지지 않고 명확하게 촬영되었는가 ?
5	수어 번역문이 변경, 왜곡되지 않고 정상적으로 표현되었는가?
6	문장의 가장 중요한 단어가 마지막에 배치되는 한국수어 문법에 맞게 정상적으로 적용되었는가?
7	농인들이 보편적으로 사용하고 직관적으로 이해할 수 있는 수어 표현을 사용하였는가?
8	지시 수어가 명확하게 표현되었는가?
9	필요에 따라 공간동사, 일치동사 분류사를 활용하였는가?
10	고유지명, 명사 표현 등 지문자 사용이 적절하게 이루어졌는가?
11	불필요한 동작 또는 비수지 표현이 포함되었는가?
12	발화 속도가 적절하였는가?

[표 II -30] 한국수어 검수 기준

5) 형태소 라벨링 데이터 검수

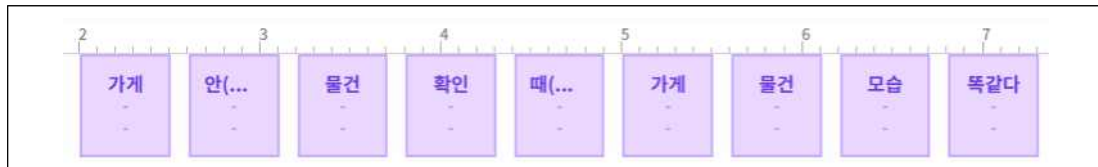
○ 형태소 라벨링 데이터 검수

- 수어 영상별 형태소 라벨링 입력 데이터는 수어 전문가(관리자)가 입력 정보를 일일이 전수 검사하는 방식으로 검수하여 데이터 정확성을 확보했다.
- 영상별 글로스 정보(우세, 비우세)가 정확하게 라벨링되었는지 문장별로 검수를 진행하였다.
- 영상별 일치동사, 대응 정보 등이 정확하게 라벨링되었는지 문장별로 검수를 진행하였다.
- Mmo, Hno, Mol, Hs, Ebf, Mctr 등 다양한 비수지 정보가 정확하게 라벨링되었는지 문장별로 검수를 진행하였다.
- 주석자가 제출한 입력 정보를 관리자가 검수 후 부정확한 정보가 입력된 경우에는 수정 요청(반려), 모두 정확히 입력된 경우에는 승인 처리하였다.

순번	내용	세부내용
1	주석 검수	<ul style="list-style-type: none"> - 토큰 길이와 정확성 - 비수지 정확성 - 일치동사 - 승인 및 반려 - 작업자 질의응답 및 지적
2	모델 수어 오류 확인	<ul style="list-style-type: none"> - 모델이 한국수어대로 수어를 구사하였는지
3	토큰(글로스) 확인	<ul style="list-style-type: none"> - 한국수어대로 글로스가 생성되었는지
4	과생략, 과오역 문장 여부	<ul style="list-style-type: none"> - 한국어 → 한국수어 변환 과정에서 과생략이 있는지 - 한국어 → 한국수어 변환 과정에서 과오역이 있는지
5	기타	<ul style="list-style-type: none"> - 작업자의 메모 확인 - 기타 이외의 발생 오류들

[표 II-31] 형태소 라벨링 데이터 검수 내용

○ 토큰 길이와 정확성



[그림 II-19] 수어 토큰 길이 설정

- 토큰이란 위 사진에 있는 보라색 네모들을 의미한다. 영상 속 모델이 수어를 구사하는 것에 맞춰서 형태소 라벨링 작업자들은 길이 설정을 하였다. 주석 검수자들은 작업자들이 길이 설정을 제대로 했는지를 기본적으로 검수하였다.
- 토큰 길이 설정 기준(아래 사진의 예시 단어 ‘다리’)



준비하기

치기

유지하기

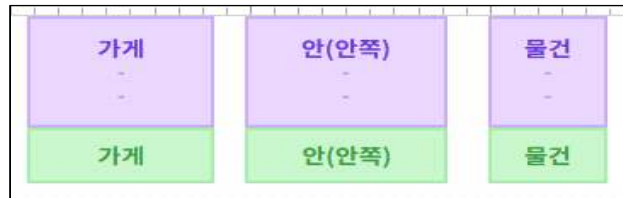
내리기

[그림 II-20] 수어 토큰 글로스 기준

- 한 토큰 안에 두 개 이상의 글로스를 입력할 수 없다.
- ‘치기’와 ‘유지하기’의 부분이 토큰 안에 포함되게 하였다.
- ‘준비하기’와 ‘내리기’는 토큰에 포함되지 않는다.
- 움직임이 있는 수어의 경우 수형이 뚜렷이 보일 때를 기준으로 하였다.
- 양손이 닿음으로써 의미가 완성되는 양손 수어의 경우 양손이 수형을 형성하며 닿는 시점을 기준으로 하였다.
- 움직임이 있는 양손 수어의 경우 양손의 수형이 뚜렷이 보일 때를 기준으로 하며 ‘준비하기’, ‘내리기’의 극히 일부분(주황색 네모)을 포함하는 것도 용인하였다. 단, 이전 또는 다음 수어 수형의 간섭이 있으면 안 된다.
- 양손 수어와 한손 수어가 합쳐져 있는 글로스(합성어 등)의 경우에는 글로스 우세와 비우세의 길이를 동일하게 설정하였다.

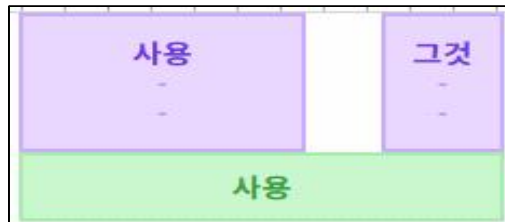
대표적인 예로 수영장 등이 있다.

- 양손 주석의 경우에는 글로스 비우세 주석을 해 줘야 하였다. 비우세 주석 입력은 토큰 위에 마우스 커서를 두고 우클릭을 하여 작업하였다. 특별한 경우가 아니라면 우세, 비우세 토큰의 길이는 같아야 하고 아래 사진에 초록색 네모가 입력된 비우세 토큰이다.



[그림 II-21] 양손 주석 예시

- 특정 수어의 사용이 끝나고 다음 수어를 진행하고 있음에도 불구하고 이전 수어의 수형 일부를 유지하며 지칭하는 표현이 전후로 나온 경우 비우세 토큰을 문장 흐름에 맞게 늘려서 아래와 같이 입력했다.



[그림 II-22] 문장 흐름에 따른 양손주석

- 우세손을 사용하지 않고 비우세 손만 사용한 경우에는 먼저 글로스 우세 토큰을 사용하여 글로스 비우세 토큰 생성 후 글로스 우세 토큰을 삭제한 뒤 글로스 비우세 토큰만을 조절하여 설정하였다.



[그림 II-23] 비우세 토큰 주석 예시

- 토큰 길이 기준의 예시는 아래와 같다.

- 예시1-1) [안녕하세요] → 다음 단어 [서비스] 틀린 주석 예시



[표 II-32] 토큰 길이 기준 틀린 주석 예시 1

- 예시1-2) [안녕하세요] → 다음 단어 [서비스] 올바른 주석 예시



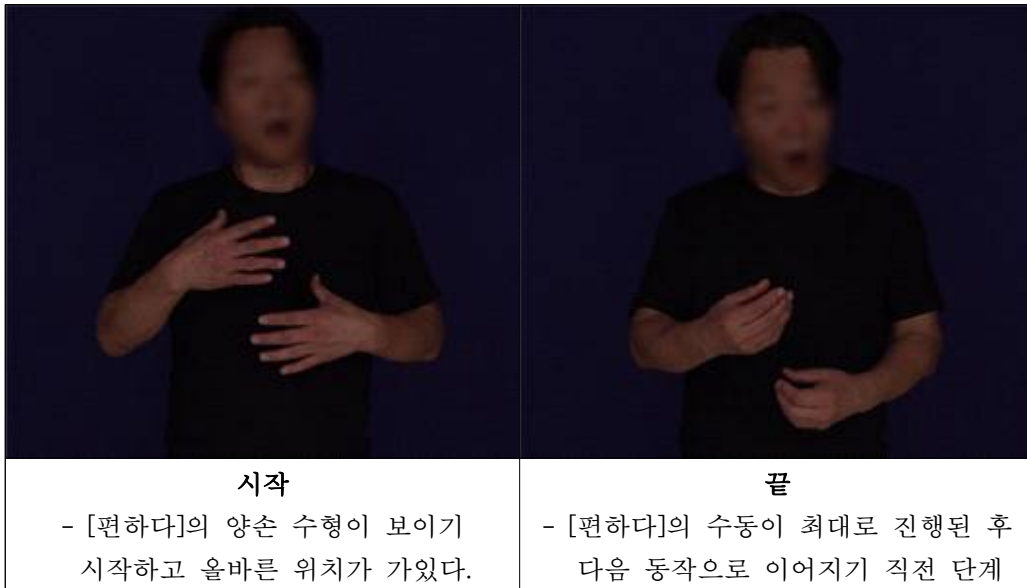
[표 II-33] 토큰 길이 기준 주석 예시

- 예시2-1) [편하다] 틀린 주석 예시



[표 II-34] 토큰 길이 기준 틀린 주석 예시 2

- 예시2-1) [편하다] 올바른 주석 예시



[표 II-35] 토큰 길이 기준 주석 예시 2

○ 비수지 정확성

- 모델이 짓거나 표현하는 모든 것을 비수지 신호로 입력하는 것이 아니라 (중요)수어 문법과 일치하거나 특정한 의미를 내포하고 있는 비수지 신호만 구분하여 입력하였다. 본 연구에서 주석하는 비수지

신호는 다음과 같다.

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

[표 II-36] 수어 모델 비수지 정보

• **Mo1**

특정 상황에 짧고 강하게 입을 벌려서 마!, 파! 표현을 하는 것을 의미한다. 이때 단순히 입을 동그랗게 벌리는 것을 착각해서 기입하지 않도록 주의하여 비수지가 나타나는 시간에만 입력했다.(“파!” 를 표현하고 입을 벌린 채로 유지한다고 해서 Mo1을 길게 기입하는 것이 아니라, 해당 비수지가 나온 순간만을 주석한다는 의미이다)

[예시] 가능(파!), 끝나다(파!), 발생(파!), 무너지다(파!) 등

• **Mmo**

모델이 입을 벌려 발음하는 비수지를 의미한다. 습관적으로 표현하거나 단어마다 말하는 입 모양은 주석을 입력하지 않고 지문자, 지숫자, 동적숫자와 함께 사용하는 입 모양(마우딩)만 기입하였다. 단, 문장부호는 포함하지 않고 모델이 표현하는 입 모양에 대응하는 단어만 기입하였다.

• **Mctr**

입꼬리가 움직이며 입을 꼭 다문다. 입을 꼭 다문 채로 미소 짓는 입술을 떠올리면 되며 입을 오므리는 것은 Mctr이 아니므로 주의하였다.

[예시] 참다, 기다리다, 열심히 하다. 안녕하세요 등

• **Hno**

고개를 끄덕인다. 모델이 단어마다 고개를 끄덕이거나 의미 없이 끄덕이는 건 주석하지 않았다.

[예시] 안녕하세요, 부탁하다, 필요하다, 나열(A, B, C), 문장 종결(~이다.) 등

• **Hs**

머리를 좌우로 흔든다. 보통 부정어에 사용된다.

[예시] 부정어(불가능, 못하다, 안되다) 등

• **Ebf**

눈썹을 찌푸린다. 보통 부정어나 강조에 사용된다.

[예시] 위험하다, 안되다, 심하다 등

• **Ci**

볼을 부풀린다.

[예시] 태풍, 무겁다(강조), 차가 밀리다, 살찌다 등

• **Ebu**

눈썹을 위로 올린다. 보통 의문문이나 놀람, 강조에 사용된다.

[예시] 안녕하세요, 의문문, 놀람 표현, 강조 등

순번	내 용
1	모델이 습관적으로 사용하는 의미 없는 비수지는 모두 주석하지 않는다.
2	문법, 문장 흐름, 문맥에 맞는 비수지만 입력한다. (위에 있는 비수지 정확성 표와 설명 참고)
3	모델이 사용하지 않은 비수지는 입력하지 않는다.
4	작업자가 불필요한 비수지를 입력한 경우에는 모두 삭제 처리한다.

[표 II-37] 비수지 검수 사항 1

- 비수지 검수 추가 중점 사항

- 기존 주석 방침은 모델이 표현하는 모든 동작과 작은 표정 하나까지도 의미가 없어도, 문법에 맞지 않아도 모두 주석 입력하도록 진행했으나, 사업 수행 도중 지침을 변경하여 새로운 검수 기준에 맞게 수정 작업을 진행하였다. 아래 표를 참고하여 작업자들의 작업물을 새로운 지침에 맞추는 작업을 진행하였다.

순번	내 용
1	지문(숫)자와 Mmo를 함께 하는 게 아닌 Mmo는 모두 삭제
2	수어 문법, 문맥에 맞는 비수지 주석이 아닌 모델의 모든 움직임을 주석한 비수지 토큰 삭제
3	모델이 사용하지 않았는데, 문맥상 들어가야 한다고 임의로 넣은 비수지 확인 및 삭제

[표 II -38] 비수지 검수사항 2

○ 승인 및 반려



[그림 II -24] 수어저작도구 권한

- 저장하기: 검수자가 자체적으로 수정한 내용이 있다면 저장하기를 꼭 눌러야 수정한 내용이 저장된다.
- 승인하기: 영상에 입력된 정보에 이상이 없을 경우 승인하기 버튼을 누르면 된다. 모든 과정에 대한 최종 승인 단계이니 신중하게 확인하고 승인하였다.
- 반려하기: 작업자가 검수자에게 제출한 주석 작업을 다시 작업자에게 되돌려보내는 기능이며 작업자가 일을 너무 성의 없게 진행했거나 특정 작업자에게 반복된 실수 등이 나온다면 반려를 진행하였다.
- 영상 최종 반려하기: 영상에 치명적 오류가 발견되어 최종 승인 처리할 수 없는 경우에 사용하며 반려 사유는 저작 도구 메모란에 기재 사항들을 모두 입력한 후 최종 반려하였다.

예시) ① 파일명(예시: LIME0001A)

② 담당자명(예시: cubeworker1)

③ 사유(예시: 글로스가 추가 생성되었음)

○ 모델 수어 오류 확인

- 모델은 한국수어에 기재된 대로 수어를 구사해야 한다. 한국수어에 없는 수어를 사용했거나, 있어야 할 수어가 누락된 영상의 경우 ‘회수 목록(검수자 이름)’ 파일에 기재 사항들을 모두 입력하였다.

6) 키포인트 라벨링 데이터 검수

- 키포인트 검수자는 작업자의 프레임별 수정 키포인트 개수, 작업 시간, 누락 키포인트 개수 등 작업 정보를 확인하며 검수를 진행하였다.
- 검수자는 전수 검사를 통해 5개 이상의 키포인트 오류를 발견 시 작업자에게 재작업을 요청하며 5개 미만의 오류는 검수자가 직접 수정을 진행하였다.
- 키포인트 라벨링 오류율이 지속적으로 높은 작업자는 별도로 분류하여 추가 작업물을 받을 수 없도록 작업자 명단에서 제외하였다.

절차		내용	담당
1 단계	좌표 정확성 검수	<ul style="list-style-type: none"> • 라벨링 결과물 키포인트의 좌표들이 COCO whol ebody 신체 특징과 정확히 일치하는지 검수 • 비가시 키포인트(손으로 가려진 얼굴의 키포인트 등)의 모호성으로 인해 픽셀 특징이 어려울 경우 분류하여 품질 관리 실무 책임자에게 보고 	검수 주석자
2 단계	가시성 플래그 검수	<ul style="list-style-type: none"> • 라벨링 결과물 키포인트의 가시성 플래그가 정확히 설정되어 있는지 검수 • 가시-비가시 판단이 모호한 키포인트는 분류하여 품질 관리 책임자에게 보고 	검수 주석자
3 단계	좌표 및 가시성 보정	<ul style="list-style-type: none"> • 검수된 신체 특징점 및 비가시 플래그 오류값 보정 	검수 주석자
4 단계	데이터 판단	<ul style="list-style-type: none"> • 1, 2단계에서 수집된 모호성 데이터들을 품질 관리 실무협의회를 소집하여 논의 후 해당 데이터 좌표 및 플래그를 확정하고 사례들을 수집하여 주석자 재교육 	품질 관리 실무 책임자
5 단계	수정 데이터 업로드	<ul style="list-style-type: none"> • 검수 완료된 데이터를 최종 결과물 폴더에 업로드 	검수 주석자

[표 II-39] 키포인트 라벨링 데이터 검수 절차

6. 병렬 말뭉치 데이터 품질관리 및 검증

1) 데이터 품질관리

○ 품질 요구 사항 충족 여부

번호	품질 요구사항	요구사항 구분			충족 여부
		구축 공정	원시 데이터	라벨링 데이터	
1	· 원천데이터는 중복이 없어야 한다.		●		확인
2	· 원천데이터의 영상 해상도는 FHD(1920×1080) 이상이어야 한다. · 원천데이터는 MP4 또는 AVI 등을 사용하며 8bit 이상이어야 한다.		●		확인
3	· 원천데이터 영상의 타임라인에 따른 데이터 정보(휴지구간, 수어동작, 수어명, 유사어 등)를 명시해야 한다.			●	확인
4	· 원시데이터 영상은 AI모델이 수어통역사를 정확히 인식할 수 있도록 모션블러 및 노이즈가 없도록 촬영되어야 한다.		●		확인
5	· 학습데이터는 민간 개방 시 사용에 제약이 없도록 개인정보 사용에 따른 동의를 확보해야 한다.	●			확인
6	· 수어 표현의 행동 인지에 대한 판단을 위해 라벨링 시 수어 표현 행동 인지 확인을 위한 전문가를 투입해야 한다.	●			확인
7	· 수어 표현 행동을 위한 각 부분(얼굴, 손, 몸 등)의 데이터를 확인하기 위한 전문가를 투입해야 한다.	●			확인
8	· 라벨링을 통해 생성된 데이터는 사람의 판단을 거쳐 만들어진 real data로 구성해야한다.			●	확인
9	· 라벨링에 필요한 적절한 저작도구를 어노테이터에게 제공해야 한다.	●			확인
10	· 수지표현 외 수어가 표현되기 위한 추가정보들을 데이터로써 제공해야한다			●	확인

[그림 II -25] 품질 요구 사항 충족 여부

○ 품질 목표 충족 여부

품질관리 영역	품질지표	품질목표	품질 목표 달성 기준	달성여부	
구축공정 품질	준비성	95% 이상	준비성 체크리스트 목록의 95% 이상 준수 필요	100%	
	완전성	95% 이상	완전성 체크리스트 목록의 95% 이상 준수 필요	100%	
구축데이터 품질	적합성	기준적합성	95% 이상	기준적합성 체크리스트 목록의 95% 이상 준수 필요	100%
		기술적합성	99% 이상	준수율(%)	99.45%
		통계적다양성	99% 이상	구축 목표 대비 실적	100%
	정확성	의미정확성	99% 이상	정확도, 정밀도, 재현율	99.0%
		구문정확성	99.9% 이상	오류율 (검사대상건수 대비 오류건수)	99.9%

[그림 II -26] 품질 목표 충족 여부

○ 체크리스트

준비성, 완전성, 적합성 지표에 따라 다음과 같이 각 품질 검사 활동을 수행하였다

- 준비성(계획수립성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
절차 준비	임무정의	발주기관(수요자) 요구사항을 분석하였는가?	☑	-사업수행계획서 -요구사항정의서
		모델 성능지표와 목표를 제시하였는가?	☑	-사업수행계획서
	구축 계획수립	학습용 데이터를 정의하였는가?	☑	-사업수행계획서
		학습용 데이터 분류체계를 정의하였는가?	☑	-사업수행계획서
	데이터 수집	데이터 수집 시 미확보 데이터 수집을 위한 방안을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집 결과에 대한 검수 절차를 마련하였는가?	☑	-한국어 수집 및 정제 검수 지침
	데이터 정제	데이터 정제 기준변경에 대한 절차를 마련하였는가?	☑	-한국어 수집 및 정제 지침
		데이터 정제방법에 대한 교육 및 훈련계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 라벨링	데이터 라벨링 방법 및 기준을 마련하였는가?	☑	-키포인트 라벨링 지침 -형태소 라벨링 지침
		데이터 라벨링 결과에 대한 검수 절차를 마련하였는가?	☑	-키포인트 라벨링 검수 지침 -형태소 라벨링 검수지침
조직 준비	구축 계획수립	한국어-한국수어 병렬 말뭉치 데이터 중 시험용 데이터 사용 방식을 위한 관리자를 지정하였는가?	☑	-사업수행계획서
		모델에 적합한지 초기데이터(데이터 규모 10% 이내)를 활용한 학습모델 검토를 위한 조직의 역할과 책임을 정의하였는가?	☑	-사업수행계획서
	데이터 수집	데이터 수집을 위한 인력 운영 계획을 수립하였는가?	☑	-사업수행계획서 -인력관리계획서
		데이터 수집을 위한 관련 법·제도적인 검토 담당 전담 조직을 구성하였는가?	☑	-사업수행계획서 -인력관리계획서
데이터 정제	데이터 정제를 위한 인력 운영 계획을 수립하였는가?	☑	-사업수행계획서 -인력관리계획서	

[그림 II -27] 준비성(계획수립성) 체크리스트 1

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)	
		데이터 정제를 위한 조직을 구성하고 담당자를 명시하였는가?	☑	-사업수행계획서	
	데이터 라벨링	데이터 라벨링을 위한 조직의 역할과 책임을 정의하였는가?	☑	-사업수행계획서 -인력관리계획서	
		데이터 라벨링을 위한 인력 운영 계획을 수립하였는가?	☑	-사업수행계획서 -인력관리계획서	
도구 준비	구축 계획수립	요구사항 변화에 따른 후보 학습모델을 제시하고 있는가?	☑	-사업수행계획서	
		모델 선정에 따른 학습모델이 정의되어 있는가?	☑	-사업수행계획서	
	데이터 수집	데이터 수집을 위한 작업 도구 교육 및 훈련계획을 수립하였는가?	☑	-사업수행계획서	
		데이터 수집을 위한 작업 도구를 정의하였는가?	☑	-사업수행계획서	
	데이터 정제	데이터 정제를 위한 작업 도구 활용 사용자/관리자 매뉴얼이 있는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침 -한국어 수집 및 정제 교육자료	
		데이터 정제를 위한 작업 도구 교육 및 훈련을 시행하였는가?	☑	-사업수행계획서 -말뭉치 작업자 교육결과서	
	데이터 라벨링	데이터 라벨링을 위한 작업 도구 교육 및 훈련계획을 수립하였는가?	☑	-키포인트 라벨링 검수 지침 -형태소 라벨링 검수지침 -키포인트 라벨링 교육자료	
		데이터 라벨링을 위한 작업 도구에 관한 법·제도적(저작권 등) 검토를 하였는가?	☑	-사업수행계획서	
	위험 관리	구축 계획수립	사업의 위험관리를 위한 계획을 수립하였는가?	☑	-사업수행계획서 -이슈 및 위험관리 대장
		데이터 수집	데이터 수집 시 위험관리를 위한 계획을 수립하였는가?	☑	-사업수행계획서 -이슈 및 위험관리 대장
데이터 정제		데이터 정제 시 위험관리를 위한 위험 요소를 식별하였는가?	☑	-사업수행계획서 -이슈 및 위험관리 대장	
데이터 라벨링		식별된 위험 대응을 위한 활동을 수행하고 있는가?	☑	-사업수행계획서 -이슈 및 위험관리 대장	

[그림 II-28] 준비성(계획수립성) 체크리스트 2

- 준비성(체계 준수성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
보안 준수	구축 계획수립	한국어-한국수어 병렬 말뭉치 데이터(훈련용, 검증용, 시험용)에 대한 보안 관리체계를 마련하였는가? (관리 및 운영 규정, 환경, 접근권한 등)	☑	-사업수행계획서
		민감정보 보호를 위한 체계를 마련하였는가? (관리 및 운영 규정, 환경, 접근권한 등)	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 수집	데이터 수집에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경구성, 접근권한 관리 등)	☑	-사업수행계획서 -인력관리계획서 -보안서약서
		민감정보 보호를 위한 활동을 수행하고 있는가? (담당자 지정, 로그 관리 등)	☑	-사업수행계획서 -인력관리계획서 -한국어 수집 및 정제 지침
	데이터 정제	데이터 정제 시 개인정보보호 등 비식별화를 수행하고 있는가?	☑	-한국어 수집 및 정제 지침
		데이터 정제에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경구성, 접근권한 관리 등)	☑	-사업수행계획서
	데이터 라벨링	데이터 라벨링에 대한 보안관리를 수행하고 있는가? (담당자 지정, 운영환경구성, 접근권한 관리 등)	☑	-사업수행계획서 -인력관리계획서 -보안서약서
	법·제 도 준수	구축 계획수립	한국어-한국수어 병렬 말뭉치 구축을 위한 관련 법·제도적인 검토를 위한 절차 및 해결방안을 제시하는가?	☑
한국어-한국수어 병렬 말뭉치 구축을 위한 개인정보활용 동의 절차를 마련하고 수행하고 있는가?			☑	-사업수행계획서 -개인정보이용동의서
한국어-한국수어 병렬 말뭉치 구축을 위한 초상권 활용 동의 절차를 마련하고 수행하고 있는가?			☑	-해당사항 없음
한국어-한국수어 병렬 말뭉치 구축 시 명예훼손 가능성 여부 검토 절차를 마련하고 수행하고 있는가?			☑	-사업수행계획서 -한국어 수집 및 정제 지침
데이터 수집		데이터 수집을 위한 관련 법·제도적인 검토를 위한 절차 및 해결방안을 제시하는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집 시 저작권 보호 대상일 경우 법에 저촉되지 않는 범위 내에서 수집할 수 있는 방안을 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집 시 저작권 보호 대상 저작물 활용에 따른 동의 절차를 마련하고 수행하는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집 시 저작권 보호 대상 저작물 활용에 따른 계약 절차를 마련하고 수행하는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
데이터		데이터 정제를 위한 관련 민감정보 비식별화 조치 등 법·	☑	-사업수행계획서

[그림 II-29] 준비성(체계 준수성) 체크리스트 1

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
	정제	제도적인 검토 절차 및 해결방안을 제시하는가?		-한국어 수집 및 정제 지침
		데이터 정제를 위한 비식별화 기법 적용하는 방안을 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 라벨링	데이터 라벨링을 위한 법·제도적인 검토 절차 및 해결방안을 제시하는가?	☑	-사업수행계획서 -키포인트 라벨링 지침 -형태소 라벨링 지침

[그림 II-30] 준비성(체계 준수성) 체크리스트 2

- 완전성(수집 완전성)

분류	단계	품질검증 체크리스트	확인 (☑)	비고
수집 완전성	데이터 수집	편향성 방지 방안을 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		정의된 데이터 수집 방법 및 기준을 적용하고 있는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집 기준변경에 대한 절차를 마련하였는가?	☑	-변경관리대장
		데이터 수집 방법에 대한 교육 및 훈련을 시행하였는가?	☑	-말뭉치 작업자 교육 결과서
		데이터 수집 결과에 대한 검수 절차 및 기준에 따라 수행하고 있는가?	☑	-한국어 수집 및 정제 검수지침
		데이터 수집 결과에 대한 검수 기준변경 시 절차에 따라 수행하였는가?	☑	-변경관리대장
		데이터 수집 결과에 대한 검수 교육 및 훈련을 시행하였는가?	☑	-말뭉치 작업자 교육 결과서

[그림 II-31] 완전성(수집 완전성) 체크리스트

- 완전성(정제 완전성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
정제 완전성	데이터 정제	데이터 정제 시 개인정보보호 등 비식별화를 실시하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		정의된 데이터 정제 방법 및 기준을 적용하고 있는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 정제 기준변경에 대한 절차를 마련하였는가?	☑	-변경관리대장
		데이터 정제방법에 대한 교육 및 훈련을 시행하였는가?	☑	-말뭉치 작업자 교육 결과서
		데이터 정제결과에 대한 검수절차 및 기준에 따라 수행하고 있는가?	☑	-한국어 수집 및 정제 검수지침
		데이터 정제결과에 대한 검수 기준변경 시 절차에 따라 수행하였는가?	☑	-한국어 수집 및 정제 검수지침 -변경관리대장 -회의록
		데이터 정제결과에 대한 검수 교육 및 훈련을 시행하였는가?	☑	-말뭉치 작업자 교육 결과서

[그림 II -32] 완전성(정제 완전성) 체크리스트

- 완전성(가공 완전성)

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
가공 완전성	데이터 수집	정의된 데이터 라벨링 방법 및 기준을 적용하고 있는가?	☑	-사업수행계획서 -키포인트 라벨링 지침 -형태소 라벨링 지침
		데이터 라벨링 기준변경에 대한 절차를 마련하였는가?	☑	-변경관리대장
		데이터 라벨링 방법에 대한 교육 및 훈련을 시행하였는가?	☑	-말뭉치 작업자 교육 결과서
		데이터 라벨링 결과에 대한 검수 절차 및 기준에 따라 수행하고 있는가?	☑	-키포인트 라벨링 검수 지침 -형태소 라벨링 검수지침
		데이터 라벨링 결과에 대한 검수 기준변경 시 절차에 따라 수행하였는가?	☑	-키포인트 라벨링 검수 지침 -형태소 라벨링 검수지침
		데이터 라벨링 결과에 대한 검수 교육 및 훈련을 시행하였는가?	☑	-말뭉치 작업자 교육 결과서

[그림 II -33] 완전성(가공 완전성) 체크리스트

- 적합성(기준 적합성)

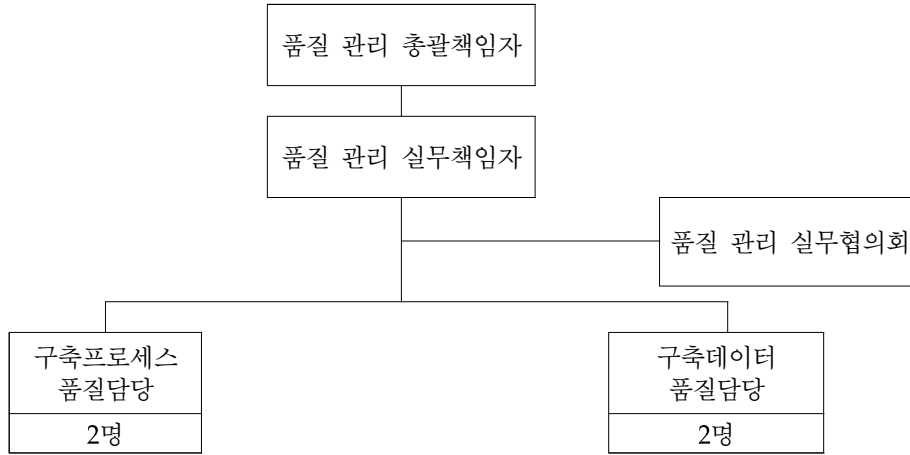
분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
다양성	구축 계획수립	실제 세상의 데이터와 유사한 특성이 데이터에 반영되도록 계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		실제 세상의 데이터와 유사한 변동성을 데이터가 갖도록 계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 수집	실제 세상의 데이터와 유사한 특성이 데이터에 반영되었는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		실제 세상의 데이터와 유사한 변동성을 데이터가 갖고 있는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
신뢰성	구축 계획수립	데이터 수집 시 수집 출처의 객관성 확보를 위한 계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 수집	데이터 수집을 위한 수집 출처에 대한 객관성 확보를 위한 근거를 제시하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
충분성	구축 계획수립	분류체계 및 분류체계별 데이터 수집 최소 수량 결정을 위한 절차를 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		분류체계 및 분류체계별 데이터 수집 최소 수량 결정에 대한 근거를 제시하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 수집	학습모델에 필요한 분류체계 및 분류체계별 데이터 수집 최소 수량을 확보하였는가?	☑	-저작도구 통계 -회의록 -사업수행계획서
균일성	구축 계획수립	분류체계별 데이터 수집 수량에 대한 적합한 비율 결정을 위한 절차를 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침 -회의록 -변경관리대장
		분류체계별 데이터 수집 수량에 대한 적합한 비율 결정을 위한 절차를 마련하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침 -회의록 -변경관리대장
	데이터 수집	분류체계별 데이터 수집 수량에 대한 적합한 비율에 맞게 수량을 확보하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침 -변경관리대장 -회의록
사실성	구축 계획수립	한국어-한국수어 병렬 말뭉치 구축 시 데이터 수집이 인위적인 환경이 경우 실제 환경 및 상황의 특성을 반영하기 위한 계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침

[그림 II -34] 기준 적합성 체크리스트 1

분류	단계	품질검증 체크리스트	확인 (☑)	확인내용 (산출물/비고)
		한국어-한국수어 병렬 말뭉치 구축 시 데이터 수집이 인위적인 환경인 경우 실세계 환경 및 조건이 일관성을 갖도록 계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 수집	데이터 수집이 인위적인 환경인 경우 실제 환경 및 상황의 특성이 반영된 근거를 제시하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
		데이터 수집이 인위적인 환경인 경우 실세계 환경 및 조건이 일관성이 확보된 근거를 제시하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
공평성	구축 계획수립	한국어-한국수어 병렬 말뭉치 구축 시 지역적 편견, 사회적 편견, 인종적 편견 등을 방지하기 위한 계획을 수립하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침
	데이터 수집	한국어-한국수어 병렬 말뭉치 구축 시 지역적 편견, 사회적 편견, 인종적 편견 등의 방지 결과에 대한 근거를 제시하였는가?	☑	-사업수행계획서 -한국어 수집 및 정제 지침

[그림 II -35] 기준 적합성 체크리스트 2

○ 품질 관리 조직 구성



조직 구분	역할과 책임
품질 관리 총괄책임자	<ul style="list-style-type: none"> · 수어 영상 데이터의 품질 관리 총괄 · 원시데이터 수집부터 학습용 데이터 구축 결과에 대한 구축/검사 작업 및 품질 관리 활동을 상호 연계하여 조정, 통제
품질 관리 실무책임자	<ul style="list-style-type: none"> · 수어 영상 데이터의 품질 관리 실무 총괄 · 품질 관리 계획 수립과 품질 검사 결과에 따른 시정 조치 개선 대책 타당성을 검토
품질 관리 실무협의회	<ul style="list-style-type: none"> · 수어 영상 데이터의 품질 관리 주요 계획, 품질 현안 등의 협의 · 원본 영상에 대한 어노테이션 방법 및 저작 도구 적용 방안 협의 · 개인 정보, 지적 재산권 등 관련 법률전문가 자문
구축 프로세스 품질 관리 담당	<ul style="list-style-type: none"> · 수어 영상 데이터의 구축 공정 품질 관리 · 품질 관리 계획에 명시된 품질 관리 활동의 준수 여부를 확인하고 품질 이슈 발생 시 이를 기록하고 개선하는 활동을 수행 · 구축 데이터 특성에 따라 데이터 획득, 정제 공정과 라벨링 공정으로 구분할 수 있으며 통합적으로 점검
구축 데이터 품질 관리 담당	<ul style="list-style-type: none"> · 수어 영상 데이터의 구축 데이터 품질 관리 · 영상 데이터, 라벨링 데이터의 품질을 확보하기 위하여 품질 계획에 수립된 검사 기준과 절차 등에 따라 품질을 검사하고 개선하는 활동을 수행

[표 II-40] 품질 관리 체계

2) 세부 검증 및 품질 관리

○ 프로세스 품질 검사

단계명	세부 업무 내용	일정		담당자
		시작일	종료일	
1. 품질 검사 준비	· 전담 인력 지정 · 협조체계 구성	’ 22.09.01	’ 22.09.20	구축프로세스 품질 담당
	· 검사 데이터 규모 선정 · 검사 데이터 및 검사 도구 준비	’ 22.09.01	’ 22.09.20	구축프로세스 품질 담당
2. 품질 검사 실시	· 검사 실시	’ 22.10.01	’ 22.10.20	구축프로세스 품질 담당
	· 오류 원인 분석 · 개선 기회 도출	’ 22.10.01	’ 22.10.20	구축프로세스 품질 담당
3. 품질 개선 조치	· 개선 계획 수립	’ 22.11.01	’ 22.11.20	구축프로세스 품질 담당
	· 개선 실시	’ 21.11.20	’ 22.12.10	구축프로세스 품질 담당

[표 II -41] 프로세스 품질 검사 내용 및 일정

○ 데이터 품질 검사

단계	세부 업무 내용	일정		담당자
		시작일	종료일	
1. 품질 검사 준비	· 전담 인력 지정 · 검사 데이터 규모 선정 · 검사 데이터 및 검사 도구 준비	’ 22.09.01	’ 22.09.20	구축 데이터 품질 담당
2. 품질 검사 실시	· 검사 실시 · 오류 원인 분석 · 개선기회 도출	’ 22.10.01	’ 22.10.20	구축 데이터 품질 담당
3. 품질 개선 조치	· 개선 계획 수립 · 개선 실시	’ 22.11.01	’ 22.12.20	구축 데이터 품질 담당

[표 II -42] 데이터 품질 검사 내용 및 일정

○ 데이터 품질관리 교육

회차	교육과정	교육내용	교육일정	교육대상
1회차	품질관리 기본교육	- 인공지능 학습용 데이터 품질 관리 가이드 - 인공지능 학습용 데이터 구축 계획 작성 방법 - 품질 관리 도구 사용 방법 등	2022.09.19	구축사업에 참여하는 수행 기관 및 참여 기관 전원
2회차	데이터 수집 및 정제 품질 가이드 교육	- 수집 데이터 품질 점검 기준 설명 - 불필요 영상, 형태소 등 정제 요건 - 영상 파일의 비형태소 제거 및 태깅 오류 유형 확인	1차 2022.09.30. 2차 2022.09.30	데이터 수집 및 정제 작업자
3회차	데이터 라벨링 품질가이드 교육	- 라벨링 데이터에 대한 품질 점검 기준 설명 - 메타데이터 구조 및 라벨링 작업 오류 유형	1차 2022.11.08. 2차 2023.01.05. 3차 2023.02.07. 4차 2023.03.03.	데이터 라벨링 작업자
4회차	데이터 검사 작업 및 품질 가이드 교육	- 라벨링 데이터에 대한 점검항목별 검사 방법 - 저작도구를 이용한 검사 결과 및 결함 등록	1차 2022.11.10. 2차 2022.11.30. 3차 2022.12.26. 4차 2023.01.13. 5차 2023.02.03.	데이터 검사자

[표 II-43] 데이터 품질관리 교육 방안

7. 보안 관리

1) 보안 관리 개요

- 보안체계를 통하여 신뢰성을 향상시키고, 정보의 정확성 및 안정성을 추구하는 것을 보안대책의 목표로 설정
- 국가 정보보안 기본지침(국가정보원), 국가·공공기관 용역업체 보안관리 가이드라인(국가정보원), 문화체육관광부 개인정보보호지침(훈령) 등 보안정책 및 지침을 준수하여 수행
- 사업 단계별 보안 대책 수립 및 준수를 통한 최적의 정보보호 확립



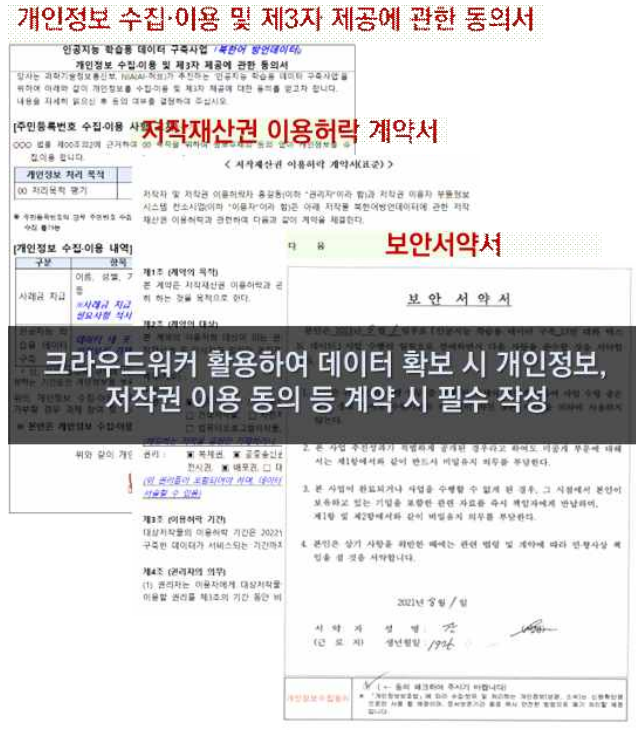
[그림 II-36] 보안 관리 전략

2) 원천 자료 및 구축 자료에 대한 저작권 확보

- 원천 자료에 대한 저작권 확보
 - 한국수어 병렬 말뭉치 데이터 구축을 수행한 클라우드 소싱 및 지자체 인력을 대상으로 직접 계약을 체결하며, 데이터 이용 허락 동의를 얻어 저작권 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 한다.

- 자유 배포가 가능하도록 이용 허락 계약 체결(비용 처리 포함)
 - 클라우드 워커 대상 이용약관 동의 절차 및 현금 보상을 통하여 데이터 저작권을 확보하였다.



[그림 II -37] 저작권 이용 동의서 샘플

- 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 이루어졌다면 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.

3) 개인정보보호 등 보안 정책 및 지침 준수

- 문화체육관광부 개인정보보호지침(훈령)을 준수하였다.
- 본 사업과 관련하여 개인 정보를 처리하거나 제공·연계 시에는 현행 개인정보보호법에서 정하는 의무 조치 사항을 반영하여 제공하였다.

4) 사업 수행을 위한 보안 대책 수립 및 준수

- 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 발생하면 모든 민·형사상 책임을 진다.
- 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.
- 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.
- 자료에 대한 보안 관리
 - 사업 수행에서 생산되는 모든 산출물은 파일 서버 또는 보안 담당관이 지정한 PC에만 저장·관리하고 사업 담당자가 인가하지 않은 비인가자에게 제공·대여·열람을 금지한다.

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 인원 에 대한 보안 교육 및 훈련 계획을 수립하였는가?	●				
5	사업 수행 전 참여 인원 에 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 기관업체 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?	●				

[그림 II-38] 보안 점검 사항 체크리스트

Ⅲ. 사업 수행 결과

1. 병렬 말뭉치 데이터 구축 결과

1) 최종 구축 데이터

- 한국어-한국수어 병렬 말뭉치 구축 사업에서 한국어 1,000,000어절, 한국어/한국수어/수어 영상 120,000문장 구축을 목표로 하였다.
 - 한국어 총 1,001,087어절, 120,295문장을 구축하여 목표량을 달성하였다.
 - 한국수어, 수어 영상은 각 120,295어절을 구축하여 목표량을 달성하였다.

분야 구분		생활(LI)		문화(CU)		합계	목표	달성율
		의료(LIME)	민원행정(LICC)	쇼핑(CUSH)	관광(CUTO)			
한국어	문장수	15,158	43,542	29,052	32,543	120,295	120,000	100.2%
	어절수	273,663	303,713	203,370	220,341	1,001,087	1,000,000	100.1%
한국수어	문장수	15,158	43,542	29,052	32,543	120,295	120,000	100.2%
수어 영상	수량	15,158	43,542	29,052	32,543	120,295	120,000	100.2%

[표 III-1] 최종 구축 데이터 수량

2. 활용 방안 및 기대 효과

1) 병렬 말뭉치의 활용 방안

- 한국수어 병렬 말뭉치 사업을 통하여 수어 말뭉치의 활용 가치를 극대화하며, 수집된 문장 및 어절을 통한 수어 통역 시스템의 핵심이 되는 원천 자료를 획득하는 데 기여할 수 있다.
- 농인들의 수어에 비해 부족한 수어 통역사들의 빈자리를 메꿔줄 AI 수어 번역 서비스 개발을 위한 고품질의 데이터셋을 제공할 수 있다.
- 수어 서비스 제공자들이 말뭉치 자료를 구체적으로 쉽게 이해하고, 말뭉치의 특성과 구조를 고려하여 활용하는 방법과 응용 분야를 체계적이고 단계적으로 접근할 수 있다.

2) 사업의 기대 효과

- 본 사업에서 구축할 문화·생활 관련 분야에 따른 다양한 내용으로 한국수어 병렬 말뭉치 관련 산업 향상에 기여
 - 한국수어 학습용 데이터 구축을 통해 농인들의 의사소통 전달이 원활히 가능하도록 구축된 빅데이터를 제공
 - 농인들과 비장애인의 의사소통 격차를 줄이고 일상생활(병원, 관공서, 법정 등)에서 상호 소통이 가능하도록 기여
 - 본 사업 구축 과정에서 감수 등의 필요 인력으로 농인을 고용하여 관련 일자리 창출에 기여
- 한국수어 병렬 말뭉치 구축을 통한 수어 영상 데이터셋을 확보
 - 문화·생활 분야 수어 영상 데이터 확보 및 품질 향상
 - 구축을 위한 시간 및 비용 절감
 - 다양한 일상생활 분야(민원, 쇼핑 등)의 데이터셋 제공
- 융합 활용 데이터 기반 제공
 - AI 기술 및 아바타 등을 활용한 융합형 대화 서비스 제공 기반 마련
 - 문화 시설 이용, 제품 문의 등의 실시간 대화형 서비스 제공 기반 마련
 - 다양한 산업 분야별 활용 가능

- 중소·벤처기업의 서비스 개발 활성화 기대
 - 신산업 창출 기회 확대 가능
 - 청년 전문 데이터 분석 인력 확대 가능
 - AI 기술 경쟁력 확보 가능
- 포스트 코로나에 대비한 AI 핵심 자원 데이터 확보
 - 비대면 생활 환경 변화에 따른 데이터 활용 방안의 확보
 - 정부 차원의 AI 산업 육성에 따른 데이터의 중요성 증가



농인들의
매우 일상적인 생활을 영위하기 위한 기반마련



사회 활동 참여도 증대



[그림 III-1] 농인 및 청각장애인이 가지는 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.

첫째, 수어 영상의 인공지능 데이터 변환에 대한 작업 방식과 과업 범위에 대한 협의가 필요하다. 한국전자기술연구원(KETI)의 영상 기반 마커리스 수어 인식 기술, 구글(Google)의 인공지능 기반 손 모양 인식 모델, 아마존의 인공지능 비서 ‘알렉사’ 등 다양한 기술이 개발되어 있으나 현재까지는 연구 단계이거나 낮은 인식률 등으로 상용화되지는 않고 있다. 한국뿐만 아니라 전 세계적으로 인공지능 기술은 하루가 다르게 변화하고 있으므로 본 사업에서는 인공지능 데이터로 가공 또는 학습하기 위한 원천 데이터까지만 구축하는 것에 대한 협의가 필요하다.

둘째, 장기적인 관점에서 한국어 수집 분야의 계획과 목표 수립이 필요하다. 본 사업의 목적은 구축된 말뭉치를 통해 인공지능 기술을 접목하여 청각장애인이 일상생활에서 의사소통이 가능한 서비스를 구현하는 데 목적이 있으므로 한국어 수집 분야 및 수집 목표 수립은 본 사업의 성공을 위한 첫걸음일 것이다.

셋째, 수어 영상 자료를 통해 수어 주석을 하는 것은 매우 정교한 기술과 많은 시간이 필요하므로 일관된 주석 기준이 필요하다. 주석 기준은 인공지능 데이터로 변환하고 활용하는 영역에서 절대적 기준이 되기 때문이기도 하다. 따라서 주석의 방식은 다양한 연구 기관 또는 데이터 활용을 목적으로 하는 기관의 논의가 필요할 것으로 보인다.

참고자료

- 국립국어원(2018), 한국수어 문법 기초 연구. 서울: 국립국어원.
- 국립국어원(2019), 2019년 한국수어 말뭉치 연구 및 구축. 서울: 국립국어원.
- 국립국어원(2020), 2020년 한국수어 활용 조사. 서울: 국립국어원.
- 국립국어원(2021), 한국수어 문법. 서울: 국립국어원.
- 국립국어원(2022), 2022년 한국수어 말뭉치 수집 및 분석. 서울: 국립국어원.
- 보건복지부(2013), 청각언어장애인의 의사소통 접근성 강화방안 연구보고서.
세종: 보건복지부.
- 보건복지부·한국보건사회연구원(2020), 2020년 장애인 실태조사. 세종:
한국보건사회연구원.
- 저작권법(문화체육관광부 법률 제18547호, 2021.12.8. 시행).
- 한국수화언어법(문화체육관광부 법률 제18783호, 2022.7.19. 시행).
- 한국전자기술연구원(2020), 영상 기반 마커리스 수어 인식 기술(특허 번호
10-2081854, 10-2098734).
- 한국전자통신연구원(2020. 6. 3.), ETRI, 장애인 위한 코로나19 지침 아바타
수어 개발[보도자료], 검색 일자 2022,10.24. 사이트 주소 [https://www.etri
i.re.kr/kor/bbs/view.etri?b_board_id=ETRI06&b_idx=18224](https://www.etri.re.kr/kor/bbs/view.etri?b_board_id=ETRI06&b_idx=18224)
- 한국지능정보사회진흥원(2022), 2022년 인공지능 학습용 데이터 구축 사업 1-
69 재난안전정보 수어영상 데이터(구축가이드). 검색 일자 2022,10.28. 사
이트 주소 [https://www.aihub.or.kr/file/down.do?fileSn=10219&cnstcPrcuse
FileSn=10219&dataSetSn=636](https://www.aihub.or.kr/file/down.do?fileSn=10219&cnstcPrcuseFileSn=10219&dataSetSn=636)
- 한국지능정보사회진흥원, AI-HUB 데이터 이용정책. 검색 일자 2022.10.24. 사
이트 주소 [https://www.aihub.or.kr/intrcn/guid/usagepolicy.do?currMenu=15
1&topMenu=105](https://www.aihub.or.kr/intrcn/guid/usagepolicy.do?currMenu=151&topMenu=105)
- COCO(2017), COCO-WholeBody Datase. 검색 일자 2022,10.25. 사이트 주소 [h](https://cocodataset.org/)

<https://cocodataset.org/#home>

PHOENIX(2014), Continuous Sign Language Recognition Dataset, 검색 일자 2022,10.27. 사이트 주소 <https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/>

USTC, Chinese Sign Language dataset, 검색 일자 2022,10.25. 사이트 주소 <http://home.ustc.edu.cn/~zhouh156/dataset/csl-daily/>

<사업 참여자>

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

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2023년 4월 28일
발행일: 2023년 4월 28일
인 쇄: (주)어플

※ 이 보고서는 국립국어원의 국고 보조금으로 수행한 ‘2022년 한국어-한국수어 병렬 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.