

국립국어원 2023-01-08

발간등록번호
11-1371028-000939-01

2022년 일상 대화 말뭉치 구축

사업 책임자
황이규



제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 일상 대화 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 07월 05일 ~ 2023년 01월 31일

2023년 1월 31일

사업 책임자: 황 이 규 (주)마인즈랩)

사업 수행자 주식회사 마인즈랩 컨소시엄

사업 책임자 황이규

사업 참여자 안준환, 현영선, 이원문, 남선웅, 박영훈, 이지현, 황주영, 윤기현,
김민수, 김성진, 이재엽, 김미영, 김태권, 김진호, 정현학, 채창운

<사업 수행자> (주)마인즈랩 컨소시엄

사업 책임자	황이규(주식회사 마인즈랩)
사업 참여자	안준환(주식회사 마인즈랩)
	현영선(주식회사 마인즈랩)
	이원문(주식회사 마인즈랩)
	남선웅(주식회사 마인즈랩)
	박영훈((주)나라지식정보)
	이지현((주)나라지식정보)
	황주영((주)나라지식정보)
	윤기현(주식회사 바이칼에이아이)
	김민수(주식회사 바이칼에이아이)
	김성진(주식회사 바이칼에이아이)
	이재엽(주식회사 바이칼에이아이)
	김미영(주식회사 바이칼에이아이)
	김태권(주식회사 스마트미디어테크)
	김진호(주식회사 스마트미디어테크)
	정현학(주식회사 스마트미디어테크)
	채창윤(주식회사 스마트미디어테크)

2022년 일상 대화 말뭉치 구축

본 사업은 2019년부터 이어온 일상 대화 말뭉치를 구축하는 것을 목표로 하며, 2019년 2인 대화 16개 주제 1,000시간, 2020년 2인 대화 15개 주제 500시간, 2021년 다자 대화 1,000시간 구축에 이어 630시간 분량의 일상 대화를 구축하였다. 이에 따른 주요 사업 내용의 성과는 다음과 같다.

음성 녹음 및 정제: 인구 통계학적 분포를 참고하여, 지역별, 성별, 나이별로 다양한 화자를 모집하였다. 전체 2,073명의 화자를 대상으로 일상 대화 및 협력적 대화 말뭉치를 수집하였다. 또한, 기존에 실내의 통제된 환경에서 수집하던 방식에서 벗어나, 실제 환경의 음성을 수집하기 위해 소음이 포함된 비통제 실환경 음성을 20% 정도 포함하였다. 수집되는 대화는 일상 대화와 협력적 대화로 구분하였다. 일상 대화는 기존 대화 주제를 모두 포함하는 16개 주제를 사전에 선정하고, 해당 주제에 관한 신문 기사를 참고하여 대화할 수 있도록 하였다. 협력적 대화는 문화·관광 위주의 10개 주제와 이에 따른 찬성과 반대를 대표하는 키워드, 신문 기사 및 상세 가이드를 참고할 수 있도록 하였다. 각 대화는 최소 2인, 최대 4인의 참가자로 구성되어 있으며, 대화의 평균 시간은 15분 내외로 제한하였다. 참여한 화자 모두 말뭉치 이용 허락 계약서를 작성하였다. 수집 및 정제된 음성 파일의 포맷은 16kHz 표본화, 16bit 양자화 선형 PCM이다.

음성 자료 전사: 기존 일상대화 말뭉치 구축 작업 경험자, 20년 이상 경력의 교정 교열 전문가 및 서지 목록 DB 구축 전문 인력이 전사 및 검증을 담당하였다. 1차 전사한 결과물에 대하여 자연어 처리 기술 및 (반)자동 검증 프로세스를 통해 탐지된 오류 후보를 1차 검수자가 검토하여 수정한 후, 2차 검수자들이 다시 전체 결과를 전수 검사하고 수정하였다.

원시 말뭉치 및 메타 정보 구축: 발화자의 메타 정보와 전사 결과를 이용하여 지침에 맞게 JSON 형식으로 변환하였다. 이는 발화자의 성별, 나이, 주요 성장지 등과 대화 상대방과의 관계, 대화 주제와 대화 형식 등의 내용을 포함하고 있다.

주요어: 일상 대화 말뭉치, 원시 말뭉치, 협력적 대화, 실환경 음성, 음성 자료 전사

차 례

제1장 사업 개요

1. 사업 목적	3
2. 사업 수행 범위	5
3. 사업 수행 절차	7

제2장 사업 수행

1. 대화 주제 및 제시 자료 선정	11
2. 전문가 자문 회의 진행	22
3. 화자 구성 및 모집	23
4. 작업자 선발 및 교육	30
5. 음성 녹음	36
6. 음성 자료 전사	50
7. 음성 정제	61
8. 원시 말뭉치 구축 및 메타 정보 구축	64

제3장 사업 수행 결과

1. 주제별 수집 결과	70
2. 화자 모집 결과	73
3. 정책 제언	94

[붙임 1] 2022년 일상 대화 말뭉치 구축 지침	96
[붙임 2] 개인정보 수집·이용 동의서	110
[붙임 3] 개인정보 제3자 제공 동의서	112
[붙임 4] 국립국어원의 개인정보 제3자 제공(공개) 동의서	114
[붙임 5] 저작권 이용 허락 계약서	116
[붙임 6] 저작권 이용 허락 계약서 미성년자 법정대리인용 동의서	120

표 차례

[표 1] 말뚝치 구축 사업 범위 및 내용	6
[표 2] 일상 대화 주제 및 세부 예시 주제	11
[표 3] 일상 대화 주제 비교(과거 구축 사업)	12
[표 4] 일상 대화 주제별 참고 자료 연결	14
[표 5] 일상 대화 주제별 참고 자료 내용(대중교통)	16
[표 6] 일상 대화 주제별 참고 자료 내용(경제/재테크)	16
[표 7] 협력적 대화 주제	17
[표 8] 협력적 대화 주제별 키워드 및 참고 자료 연결	18
[표 9] 1차, 2차 자문 회의 상세	22
[표 10] 사업 초기 화자 할당표 설계 기준	23
[표 11] 성별 및 나이대별 지역별 모집 목표(단위: 명)	24
[표 12] 통제 대화 성별 및 나이대별 지역별 모집 목표(단위: 명)	25
[표 13] 비통제(실환경) 대화 권역별 데이터 수집 목표 인원 및 시간	25
[표 14] 비통제 대화 카테고리별 데이터 수집 목표 시간	26
[표 15] 통제 대화 발화자 모집 계획 변경안(단위: 명)	27
[표 16] 비통제 대화 발화자 모집 계획 변경안	27
[표 17] 진행 요원 선발 및 운영 방안	30
[표 18] 진행 요원 교육 내용	31
[표 19] 전사 작업자 선발 기준 및 운영	32
[표 20] 전사 작업자 교육	33
[표 21] 개인정보 보호 및 보안 관련 교육	35
[표 22] 코로나-19 집단 감염 방지 화자 관리 방안	37
[표 23] 비통제 환경 녹음 성장지별 수집 목표 시간	39
[표 24] 비통제 환경 녹음 카테고리별 수집 목표 시간	39

표 차례

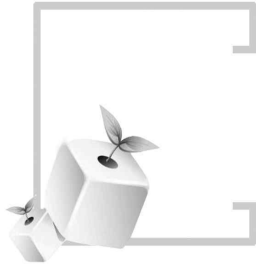
[표 25] 데이터 프로파일링 세부 공정 및 오류 예시	56
[표 26] 오류 어절 후보 선별 기준	59
[표 27] 대화 파일명 부여 방식	64
[표 28] 말뭉치 변환 예시(일부)	64
[표 29] 주제별 대화 수집 결과	71
[표 30] 통제 및 비통제 데이터 전사 및 납품 결과(단위: 명)	74
[표 31] 통제 및 비통제 데이터 전사 및 납품 결과(단위: 시나리오 개수)	75
[표 32] 통제 및 비통제 데이터 전사 및 납품 결과(단위: 시간)	76
[표 33] 통제 데이터 전사 및 납품 결과(단위: 명)	77
[표 34] 통제 데이터 전사 및 납품 결과(단위: 시나리오 개수)	78
[표 35] 통제 데이터 전사 및 납품 결과(단위: 시간)	79
[표 36] 비통제 데이터 수집 카테고리별 전사 및 납품 결과	80
[표 37] 비통제 데이터 발화자 성장지별 전사 및 납품 결과	81
[표 38] 주제별 나이대 분포(단위: 시간)	83
[표 39] 대화 유형 및 인원별 분포(단위: 대화 수량)	85
[표 40] 주제별 성별 분포(단위: 대화 수량)	86
[표 41] 화자 간 관계별 수집 결과(단위: 개)	88
[표 42] 직업별 수집 결과(단위: 명)	89
[표 43] 학력별 수집 결과(단위: 명)	90
[표 44] 출생지별 화자 모집 결과(단위: 명)	91
[표 45] 주 성장지별 화자 모집 결과(단위: 명)	92
[표 46] 현 거주지별 화자 모집 결과(단위: 명)	93

그림 차례

[그림 1] 일상 대화 말뭉치 구축 사업 목적	4
[그림 2] 일상 대화 말뭉치 구축 사업 수행 범위	5
[그림 3] 말뭉치 구축 수행 절차	7
[그림 4] 말뭉치 구축 사업 참여자 모집 카페	28
[그림 5] 말뭉치 구축 사업 참여자 모집 공고	29
[그림 6] 녹음 진행 요원 교육 자료 일부	31
[그림 7] 전사 교육 자료(일부)	34
[그림 8] 통제 대화 수집을 위한 녹음 장비 및 환경	36
[그림 9] 녹음 시작 전 녹음 장소 방역 진행	37
[그림 10] 녹음 장비 및 장비 테스트	38
[그림 11] 외장 마이크 장착	40
[그림 12] 음성 녹음 절차	41
[그림 13] 개인정보 활용 동의서(예시)	42
[그림 14] 저작권 이용 허락 계약서(예시)	43
[그림 15] 음성 자료 수집 일지(예시-1)	43
[그림 16] 음성 자료 수집 일지(예시-2)	44
[그림 17] 통제 및 비통제 녹음 진행	46
[그림 18] 각 카테고리별 비통제 녹음 진행 사진	47
[그림 19] 수집 데이터 검증	48
[그림 20] 공유 시스템 로그인 및 파일 등록(예시)	49
[그림 21] 전사 도구를 이용한 전사 절차	52
[그림 22] 전사 도구에서 전사 캠페인 보기	53
[그림 23] 전사 도구에서 전사 대상 대화 목록 보기	53
[그림 24] 전사 도구에서 전사 수정, 청취 및 결과 보기	54

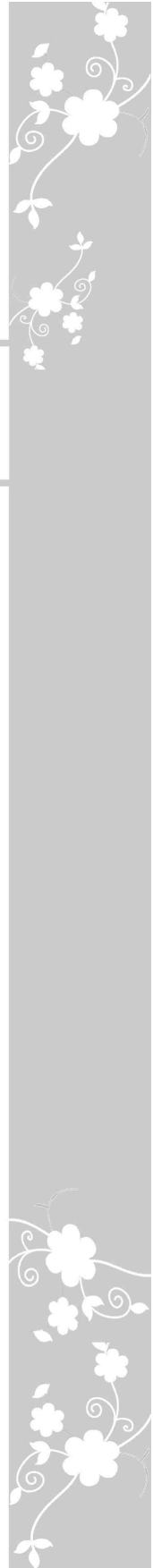
그림 차례

[그림 25] 데이터 프로파일링 절차	55
[그림 26] 전체 철자 전사 중 출현 빈도수 3 이하 어절 목록 예시	58
[그림 27] 오류 어절 후보 선별 예시	59
[그림 28] 오류 어절 후보 목록 예시	60
[그림 29] 발화 도중 휴지로 인한 억양구 분할 예시	62
[그림 30] 개인정보 비식별화(예시)	63
[그림 31] 메타 정보 파일 일부	66
[그림 32] 발화자 메타 정보 일부	66



제1장

사업 개요



1. 사업 목적

인공지능 산업이 발전함에 따라 인공지능 기술을 개발하고 활용하기 위한 대규모 고품질의 한국어 말뭉치 구축의 필요성이 커지고 있다. 문어체나 회의와 같은 공식적인 대화체 말뭉치는 상대적으로 많지만 이에 반해 일상 대화 말뭉치는 부족한 상태이다. 일상 대화 말뭉치는 음성 인식, 자연어 이해, 대화 처리 및 질의응답과 같은 다양한 인공지능 기반 자연어 서비스의 개발을 위한 중요한 데이터이다.

본 사업의 목적은 인공지능 분야 R&D 관련 산업 활성화 및 언어 연구에 기여하기 위해 630시간 이상의 일상 대화 말뭉치를 구축하고 공개하는 것이다. 일상 대화 말뭉치 구축 및 공개를 통해 기초 말뭉치의 양적, 질적 부족으로 인해 개발하지 못했던 기술을 개발하고, 인공지능 기술 개발의 수준을 높일 수 있다. 또한, 국어 자원의 활용도와 가치를 높이기 위해 민간에서 활용 가능한 국가 공공재로서의 말뭉치를 확대 구축하는 것이 본 사업의 목적이다.

2019년 16개 주제(군대, 자동차 등) 일상 대화 원시 말뭉치 1,000시간 구축, 2020년 15개 주제(스포츠/레저, 여행지 등) 일상 대화 원시 말뭉치 500시간 구축, 2021년 15개 주제(휴가, 관혼상제 등) 일상 대화와 8개 주제(AI의 직업에 대해, 공공 공간의 CCTV 설치 등)의 협력적 대화 원시 말뭉치 1,000시간 구축에 이어 본 사업에서는 생활/주거환경 등의 새로운 주제를 포함하여 16개 주제로 하는 2인 대화, 3인 이상의 다자 대화, 대화를 통해 결론을 도출해가는 협력적 대화를 구축하였으며, 비통제 환경에서의 소음을 포함한 일상 대화도 시범적으로 100시간을 구축하여 총 630시간의 원시 말뭉치를 구축하였다. 지난 3년간 구축된 말뭉치에 이어 다양한 화자를 모집하고 다양한 주제에 따라 풍부한 대화 내용을 수집하고, 이를 전사 말뭉치로 구축하였다.

언어 인공지능 산업 발전을 위한 기반 마련 + 국어연구 및 정책 수립 활용

일상 대화 말뭉치(630시간) 기획/수집/정제/가공 및 공개

말뭉치 구축 기획	일상대화 녹음 및 정제	이중 전사 및 원시 말뭉치 구축	메타정보 구축 및 납품
<ul style="list-style-type: none"> 일상대화 수집 상세 기획 철자 및 발음 전사 상세 기획 검수 및 품질관리 상세 기획 	<ul style="list-style-type: none"> 수집 도구 및 환경 준비 발화자 및 오퍼레이터 모집 일상 대화 녹음 (통제/비통제) 녹음 데이터 정제 	<ul style="list-style-type: none"> 전사 도구 및 환경 준비 전사 작업자 모집 및 교육 철자 및 발음 전사 검수 및 품질관리 	<ul style="list-style-type: none"> 메타정보 포함 최종 말뭉치 구축 및 납품 연구보고서 등 문서 산출물 작성 및 제출

국어 연구 및 인공지능 기술·산업 발전을 위한 대규모 말뭉치 필요

고품질 우리말 자원 수요 증대

- 문어체, 회화와 같은 공식적인 대화체 말뭉치는 상대적으로 많으나, 일상 대화 말뭉치는 매우 부족
- 기존 일상 대화 말뭉치 사업의 연장선

인공지능 기술 개발 및 활용

- 음성인식, 자연어 이해, 의도 파악, 대화, 질의응답 등 실제 일상에서의 인공지능 적용을 위한 연구·개발에 활용

인공지능 국가전략 추진

- '19년 12월 인공지능 국가전략 발표
- 데이터태를 포함한 디지털 뉴딜 추진
- 다양한 대량 데이터 수집·정제·공개 추진

[그림 1] 일상 대화 말뭉치 구축 사업 목적

2. 사업 수행 범위

본 사업의 수행 범위는 크게 네 가지로 나눌 수 있다. 첫째, 일상 대화 말뭉치를 구축하기 위한 기획 단계로 데이터 수집을 위한 주제 및 화자 선정, 전사 기준 및 말뭉치의 품질 관리 기준을 수립한다. 둘째, 일상 대화 말뭉치를 수집하기 위한 녹음 환경을 구축하고 대화 데이터를 수집 및 정제한다. 셋째, 녹음된 데이터의 전사 및 검수를 위한 환경을 준비하고 작업자를 모집하여 교육 및 전사를 진행한 후 말뭉치에 대한 품질을 검수한다. 넷째, 말뭉치에 대화 주제, 화자 정보 등의 메타 정보를 부착하여 최종 산출물로 납품할 수 있도록 한다.



[그림 2] 일상 대화 말뭉치 구축 사업 수행 범위

말뭉치의 구축과 관련된 사업의 범위는 다음과 같다.

[표 1] 말뭉치 구축 사업 범위 및 내용

말뭉치 구축의 범위	상세 내용	말뭉치 구축 양
주제 및 제시 자료 선정	- 2019/2020/2021년 구축 주제 및 제안 요청 내용을 고려하여 다양한 주제 선정 - 문화·관광 위주의 협력적 대화 주제 선정	- 일상 대화: 16개 주제 - 협력적 대화: 10개 주제
음성 녹음 및 정제	- 화자별 최대 녹음 시간은 1시간(4개 주제) ¹⁾ 으로 제한 - 음성 개인정보 비식별화 처리	- 2,000명 이상의 화자 참여 - 정제 후 630시간 음성 수집
음성 자료 전사	- 발음 전사와 철자 전사를 병행(이중 전사) - 전사 지침에 맞게 전사하고, 전사 후 수작업 전수 검사 실시	- 최종 제출자료 630시간 이상

성별, 나이, 직업, 지역 등의 비율이 편중되지 않도록 초기에 선정하였으며, 사업 추진 과정에서 주관기관과 협의하여 일부 비율은 조정하였다. 사업의 주요 내용은 다음과 같다.

- 2인~4인이 특정 주제로 일상 대화 또는 협력적 대화 진행
- 대화 내용 녹음 및 정제(정제 후 630시간, 대화 당 12-18분 이내, 평균 15분 내외)
- 해당 녹음 자료에 대한 저작권 이용 허락 계약 체결
- 녹음된 내용 이중 전사(발음 전사/철자 전사)
- 구축된 전사 자료에 대한 메타 정보(화자 정보, 대화 주제, 녹음 일자 등) 구축

1) 사업 초기 2인 대화의 경우 한 화자당 최대 녹음 시간을 30분으로 제한하였으나, 화자 모집에 어려움을 겪으며 사업 진척 지연이 발생함에 따라, 2인 대화 수집 시간을 한 화자당 최대 1시간으로 변경함.

3. 사업 수행 절차

본 사업은 준비 단계, 구축 단계, 검사 단계, 품질 검증의 4단계로 진행되었다. 사업 기간 내에 목표로 한 구축량 달성을 위해 각 단계별 임무를 명확히 하고, 공정별 품질 검증 과정을 두어 최종 말뭉치의 품질에 문제가 없도록 하였다.



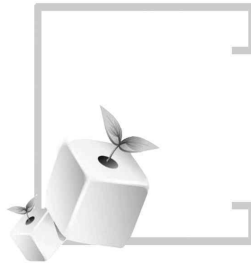
[그림 3] 말뭉치 구축 수행 절차

첫 번째 기획 및 설계 단계에서는 일상 대화 말뭉치를 수집하기 위한 상세 내용을 기획하였다. 수집 방법 및 장소 확보, 대화 주제 선정, 발화자 비율 설정, 일관된 전사를 위한 절차 및 지침서 보완 등 말뭉치를 균등한 비율로 동일한 조건에서 수집할 수 있도록 설계하였다.

두 번째 데이터 수집 및 정제 단계에서는 녹음 작업자인 오퍼레이터와 발화자를 모집하고 수집 환경을 구축하였으며, 음성 데이터 수집 및 정제 후 전사 작업을 진행할 수 있도록 준비하였다. 수집 시 모든 발화자에게 저작권 계약서 및 개인정보 동의서를 받아 추후 관련 이슈가 없도록 하였다. 전사 도구에 음성 파일을 적재하기 위해 음성 파일을 알맞은 형식으로 정제 및 변환하였으며 이와 동시에 음성 데이터 및 발화자 메타데이터를 생성하여 구축하였다.

세 번째 데이터 가공 단계에서는 전사 도구를 활용하여 음성 데이터 전사를 진행하였다. 이에 앞서 음성 전사 및 개인정보 비식별화가 가능하도록 도구 개발 및 고도화를 진행하였으며, 전사 지침서를 활용하여 전사 작업자를 대상으로 교육을 진행하였다. 또한, 전사가 완료된 말뭉치는 대화 주제, 발화자 성별, 나이대, 주 성장지 등의 메타데이터를 생성하고 부착하였다.

네 번째 단계인 데이터 검증 단계에서는 가공이 완료된 말뭉치의 품질을 검증하였다. 전사가 완료된 데이터에 대해 전수 검사를 진행하고, 데이터 프로파일링을 통하여 메타데이터 및 전사에 대한 오류를 탐지 및 수정하여 전체 데이터의 품질에 문제가 없도록 하였다. 또한 각 단계별 검수 과정을 두어 말뭉치의 품질과 구축 공정의 품질을 높이고자 하였다. 총 6단계의 내부 품질 검증 과정을 통해 고품질의 일상 대화 말뭉치 데이터를 구축 및 공개하는 데 문제가 없도록 하였다.



제2장

사업 수행



1. 대화 주제 및 제시 자료 선정

대화의 주제는 일상 대화와 협력적 대화로 구분하여 선정하였다. 일상 대화는 기존 구축 말뭉치(2019년~2021년)의 대화 주제를 참고하였다. 기존 주제와 유사한 주제 및 새로운 대화 주제를 발굴하였으며, 다양한 자유 발화가 수집될 수 있도록 하였다.

일상 대화 말뭉치의 주제는 총 16가지로 선정하였으며, 대화 주제와 세부 주제 예시를 제시하여 발화자가 직접 대화의 주제를 선택할 수 있도록 하였다.

[표 2] 일상 대화 주제 및 세부 제시 주제

번호	주제	세부 주제 예시
1	휴가	여행 시 교통/숙박 선택, 자연/휴양지 소개, 여행 국가/지역 선택, 추천 여행지(국내/해외) 등
2	대중교통	약속 만남 시 교통 선택, 새로운 교통수단 공유 킷보드, 전철, 버스, 기차, 교통 약자석 등
3	음악	대중 음악, 선호 가수 및 노래 추천, 선호하는 음악 장르 등
4	건강/다이어트	가지고 있는 질병/알레르기, 건강 보조제, 건강을 위해 하고 있는 노력, 약 부작용, 다이어트 성공/실패 경험담 등
5	방송/연예	인생 드라마 추천, 예능 프로그램, 선호하는 배우 등
6	스포츠/레저/취미	경기 직접 관람, 등산, 여름/겨울 레저, E-스포츠(게임), 독서(시집, 문학 등), 영화 관람, 만화(웹툰 등), 웹소설 등
7	먹거리	최근 인기있는 음식과 경험, 가장 선호하는 음식, 추천하는 맛집, 배달 음식, 밀키트 등
8	우정	선호하는 것, 성격, 다툼과 화해 등에 대한 친구 사이의 대화, 국경/나이를 초월한 우정 등
9	경제/재테크	부동산, 주식 투자, 비트코인, 고물가, 세금, 인플레이션, 금리 인상 등
10	회사/학교	회사 생활, 회식, 인턴 생활, 진학에 대한 정보와 결정, 입시 경쟁, 전공, 성적 등
11	반려동물	반려동물 입양, 유기 동물 문제, 동물 병원 고비용, 동물 등록제, 반려동물과의 추억 등
12	취직	취업에 대한 정보 공유, 취업에 필요한 자격증/어학 시험, 취준생, 해외 취업 등
13	가족/관혼상제	집안 행사, 연애/결혼, 결혼 준비, 청첩장, 축하, 축의금, 문상 예절, 조의금, 제사 준비 등
14	쇼핑	온라인/오프라인 쇼핑 중 선호하는 쇼핑 방법, 선물을 고르는 상황 등
15	생활/주거환경	이사, 주거 유형, 생활권, 장 보기, 집안일, 계절/날씨 등
16	기타	꿈(목표), 군대 경험 등

기존 구축 말뭉치(2019년~2021년)의 대화 주제와 2022년 구축 말뭉치를 주제별로 비교하면 다음과 같다.

[표 3] 일상 대화 주제 비교(과거 구축 사업)

번호	2019년	2020년	2021년	2022년	세부 주제	비고
1	군대					기타에 포함
2	게임					스포츠/레저/취미에 포함
3	휴일		휴가	휴가	여행 시 교통/숙박 선택, 자연/휴양지 소개, 여행 국가/지역 선택, 추천 여행지(국내/해외) 등	
4	자동차		대중교통	대중교통	약속 만남 시 교통 선택, 새로운 교통수단 공유 킷보드, 전철, 버스, 기차, 교통 약자석 등	
5	만화					방송/연예에 포함
6	영화	영화	음악	음악	대중 음악, 선호 가수 및 노래 추천, 선호하는 음악 장르 등	
7	정치					제외
8	건강/ 다이어트	건강/ 다이어트	건강/ 다이어트	건강/ 다이어트	가지고 있는 질병/알러지, 건강 보조제, 건강을 위해 하고 있는 노력, 약 부작용, 다이어트 성공/실패 경험담 등	
9	방송/ 연예	방송/ 연예	방송/ 연예	방송/ 연예	인생 드라마 추천, 예능 프로그램, 선호하는 배우 등	
10	스포츠/ 레저	스포츠/ 레저	스포츠/ 레저	스포츠/ 레저/ 취미	경기 직접 관람, 등산, 여름/겨울 레저, E-스포츠(게임), 독서(시집, 문학 등), 영화 관람, 만화(웹툰 등), 웹소설 등	
11	먹거리	먹거리	먹거리	먹거리	최근 인기있는 음식과 경험, 가장 선호하는 음식, 추천하는 맛집, 배달 음식, 밀키트 등	
12	자연/ 휴양지					휴가에 포함
13	국가/ 지역					휴가에 포함
14	문학					스포츠/레저/취미에 포함
15	연애/ 결혼	연애/ 결혼	우정	우정	선호하는 것, 성격, 다툼과 화해 등에 대한 친구 사이의 대화, 국경/나이를 초월한 우정 등	연애/결혼-관혼상제에 포함
16	경제/ 재테크		경제/ 재테크	경제/ 재테크	부동산, 주식 투자, 비트코인, 고물가, 세금, 인플레이션, 금리 인상	

					등	
17		여행지 (국내/ 해외)				휴가에 포함
18		계절/ 날씨				기타에 포함
19		회사/ 학교	회사/ 학교	회사/ 학교	회사 생활, 회식, 인턴 생활, 진학에 대한 정보와 결정, 입시 경쟁, 전공, 성적 등	
20		선물				쇼핑에 포함
21		꿈(목표)				기타에 포함
22		반려동물	반려동물	반려동물	반려동물 입양, 유기 동물 문제, 동물 병원 고비용, 동물 등록제, 반려동물과의 추억 등	
23		아르 바이트	취직	취직	취업에 대한 정보 공유, 취업에 필요한 자격증/어학 시험, 취준생, 해외 취업 등	
24		성격				우정에 포함
25		가족	가족			가족/관혼상제와 통합
26			쇼핑	쇼핑	온라인/오프라인 쇼핑 중 선호하는 쇼핑 방법, 선물을 고르는 상황 등	
27			관혼상제	가족/관혼 상제	집안 행사, 연애/결혼, 결혼 준비, 청첩장, 축하, 축의금, 문상 예절, 조의금, 제사 준비 등	
28				생활/주거 환경	이사, 주거 유형, 생활권, 장 보기, 집안일, 계절/날씨 등	
29				기타	꿈(목표), 군대 경험 등	

발화 참여자들에게는 아래의 16가지 대화 주제를 주고, 자유 발화 시 도움이 될 수 있도록 예시 및 신문 기사를 참고 자료로 함께 제시하여 자연스러운 일상 대화를 유도하였다.

[표 4] 일상 대화 주제별 참고 자료 연결

번호	주제	세부 주제	참고 자료 링크
1	휴가	여행 시 교통/숙박 선택, 자연/휴양지 소개, 여행 국가/지역 선택, 추천 여행지(국내/해외) 등	https://news.mt.co.kr/mtview.php?no=2021070609421133987 https://www.fnnews.com/news/202107061409120693 https://www.yna.co.kr/view/AKR20210702157600530?input=1195m
2	대중교통	약속 시간, 장소, 교통, 선택	https://www.chosun.com/national/transport-environment/2021/07/05/AWQIRZ7OVZAPRLZ3DPSELNYIPA/ http://news.tf.co.kr/read/livingculture/1870885.htm https://www.kado.net/news/articleView.html?idxno=1080443
3	음악	대중음악 유행, 선호 가수 및 곡 추천	https://news.joins.com/article/24087649s.joins.com/article/24087649 https://www.ytn.co.kr/_ln/0106_202106200641266124 https://news.mt.co.kr/mtview.php?no=2021070512447232229
4	건강/다 이어트	성인병에 대한 상식, 처방, 대응	https://www.hidoc.co.kr/healthstory/news/C0000612077 https://www.news1.kr/articles/?4361232 http://www.healthinnews.co.kr/news/articleView.html?idxno=24078
5	방송/연 예	드라마, 예능 프로그램 선택	https://www.mk.co.kr/star/hot-issues/view/2021/07/649658/ https://news.joins.com/article/24097781 https://www.hankookilbo.com/News/Read/A2021060908290003684
6	스포츠/ 레저/취 미	직접 운동, 관람, 시청 등에 대한 정보와 결정	https://sports.news.naver.com/news?oid=109&aid=0004437288 https://sports.news.naver.com/news?oid=076&aid=0003751414 https://sports.news.naver.com/news?oid=020&aid=0003368259
7	먹거리	저녁 모임의 음식 종류와 식당 선택	http://www.inews24.com/view/1379663 https://www.mk.co.kr/news/business/view/2021/06/562001/ https://www.sommeliertimes.com/news/articleView.html?idxno=18776
8	우정	선호하는 것, 성격, 다툼과 화해 등에 대한 친구 사이의 대화, 취미	https://www.hankookilbo.com/News/Read/A2021063023330000098 https://news.joins.com/article/24005854 http://weekly.chosun.com/client/news/viw.asp?ctcd=c09&

		토론	nNewsNumb=002655100010
9	경제/재테크	집, 주식 등 투자에 대한 고려와 결정	https://www.hankyung.com/realestate/article/202107058533e https://www.ajunews.com/view/20210705103146300 https://www.fetimes.co.kr/news/articleView.html?idxno=97070
10	회사/학교	취직, 진학에 대한 정보와 결정	http://www.edupress.kr/news/articleView.html?idxno=7625 https://www.hankyung.com/it/article/2021062904701 https://news.kbs.co.kr/news/view.do?ncd=5226521
11	반려동물	개, 고양이의 장단점 비교 및 결정	https://www.hidomin.com/news/articleView.html?idxno=455503 https://www.mk.co.kr/news/culture/view/2021/07/650420/ https://www.msn.com/ko-kr/money/topstories/%EC%A7%91%EC%82%AC%EB%A5%BC-%EC%9C%84%ED%95%9C-%ED%8C%A9%ED%8A%B8%EC%B2%B4%ED%81%AC-%EA%B3%A0%EC%96%91%EC%9D%B4-%EB%88%88-%EA%B9%9C%EB%B0%95%EC%9E%84%EC%9D%80-%EC%98%80%EB%8B%A4/ar-AALMslA
12	취직	대기업/중소기업 취직의 정보 공유와 견해 교환	https://www.mk.co.kr/news/business/view/2021/07/644645/ https://moneys.mt.co.kr/news/mwView.php?no=2021062413118080027 https://www.newswire.co.kr/newsRead.php?no=926613
13	가족/관혼상제	집안 행사에 대한 검토와 결정 결혼, 문상, 제사, 축의금, 참석 등	https://imnews.imbc.com/replay/2021/nwtoday/article/6282583_34943.html http://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0002751181 https://www.hani.co.kr/arti/economy/economy_general/1002116.html
14	쇼핑	핸드폰 구매 시 기종 검토 및 결정	https://www.dailysecu.com/news/articleView.html?idxno=125386 https://biz.chosun.com/it-science/ict/2021/07/06/YTKWM4607NAHDEKEREHWJLFJT4/ https://news.join.com/article/24097527
15	생활/주거 환경	이사, 주거 유형, 생활권, 장 보기, 집안일, 계절/날씨 등	https://imnews.imbc.com/replay/2022/nwdesk/article/6383030_35744.html https://www.hankyung.com/realestate/article/201802275366e https://www.sisamagazine.co.kr/news/articleView.html?idxno=462802
16	기타	성격 유형(mbti)에 관한 이야기, 군대	http://www.ksilbo.co.kr/news/articleView.html?idxno=942781

		이야기, 자신의 목표(꿈)에 대한 이야기, 신화, 역사 이야기, 관심 있는 콘텐츠(자주 보는 유튜브 콘텐츠 등), 자동차(오토바이) 운전에 관한 경험	http://autotimes.hankyung.com/apps/news.sub_view?popup=0&nid=03&c1=03&c2=03&c3=00&nkey=202111261117021 https://www.artinsight.co.kr/news/view.php?no=51879
--	--	--	--

참고한 신문 기사의 상세 내용의 예시는 아래와 같다.

[표 5] 일상 대화 주제별 참고 자료 내용(대중교통)

<p><대중교통-1: 조선일보, 극심하게 붐비는 광역버스 출퇴근시간 운행차량 늘려(21.07.05)></p> <p>출퇴근 시간 극심하게 붐비는 수도권 광역버스 노선에 정부가 전세버스를 추가 투입한다. 4일 국토교통부 대도시권광역교통위원회(대광위)에 따르면 5일부터 경기도 성남 분당구 등과 서울시 사이를 오가는 9000번, 9003번, 9007번, 9300번 버스 노선에 출근 시간 10대, 퇴근 시간 3대의 전세버스가 투입된다. 이달 말까지 시흥의 3200번, 남양주의 M2323번 노선에도 전세버스가 투입된다.</p> <p>정부는 지난해부터 출퇴근 시간에 맞춰 수도권 광역버스 노선에 전세버스를 빌려와 투입·운영하는 사업을 추진하고 있다. 출퇴근 시간 서울 통근 수요가 몰리며 정원 대비 탑승 인원이 178%에 달할 정도로 광역버스가 붐비기 때문이다. 이 때문에 버스가 만석이 돼 정류소에 서지 않고 그냥 통과하거나 입석을 해야 하거나 배차 간격이 길어지는 문제가 발생했다.</p> <p>정부는 지난해 노선 10개에 전세버스를 투입했다. 올해도 이미 15개 노선에 배차를 완료했다. 이번에 노선 6개가 추가되면 지난해부터 총 31개의 노선에 전세버스가 투입되게 된다. 이 31개 노선에서 하루 총 135회 운행이 늘어난다고 국토부는 설명했다. 사업비의 30%는 국비로 지원한다. 지난해와 올해 총 14억4000만원이 투입된다. 국토부 대광위 이광민 광역버스과장은 “전세버스 추가 투입으로 출퇴근 시간대 좌석 수가 평균 31% 늘어나고, 배차 간격이 25% 줄어들 것”이라고 밝혔다.</p>
--

[표 6] 일상 대화 주제별 참고 자료 내용(경제/재테크)

<p><경제/재테크-1: 한경, 눈물도 웃음기도 짝 빠진 부동산 시장(21.07.06)></p> <p>심리·감정보다는 '이성'으로 돌아가는 시장 분위기 믿음 깨진 연인과의 같은 부동산 시장</p>
--

매수자들 '돈 된다'는 곳으로 언제 어디든 몰려

"부산 연초에 계속 청약 안 좋았던 것 같은데, 여기는 분위기가 다르네요."(기자) "부산 사람들도 알거다 알죠. 브랜드에 비싸도 좋은 물건은 무조건 잡지만, 외곽이나 비(非)브랜드는 쳐다도 안 봅니다. 서울에서 '똥똥한 한채', '강남'만 선호하는 것과 같습니다. 예전같이 우르르 가서 계약하고 그런거 안해요."(현지 부동산 업체)

시장에는 '부동산 민심', '부동산은 심리'라는 말이 있었다. 하지만 최근 시장을 보면 정서적인 면은 거의 작용하지 않는 듯하다. 감정은 남아 있지 않고 '냉철한 판단'과 '이성'만 남아있다는 생각이 들 정도다. 오래된 아파트에 사는 집주인의 태도가 "누가 뭐래도 난 우리 동네 좋아"에서 "우리집 빼고 다 올랐다"고 변한 것만 봐도 알 수 있다.

최근 만난 한 전문가는 부동산 시장에 '감성'은 거의 없고 '이성'만 남았으며 이는 곧 믿었던 '정부와의 이별'이 아니겠느냐라고 했다. 처음에는 의아했지만 이별후 심리를 생각해보니 이해가 갔다. 남녀가 이별을 하면 처음에는 상황을 부인하고, 슬프고 우울함에 빠지기도 한다. 하지만 시간이 지나고 일상을 깨닫고 정신을 차리게 된다. 물론 슬픔의 치유과정에서 또다른 만남을 갖기도 하지만, 이는 오래가기 어렵거나 후회하는 경우가 많다.

또한, 자유로운 일상 대화 내용 외에 대화 주제에 대한 논의를 통해 결론을 도출하는 내용의 말뭉치 구축을 위해 문화 관광 위주 10개의 협력적 대화 주제를 추가로 설정하여 제공하였다. 협력적 대화란, 문제 해결 과정에서 문제 해결과 관련된 지식을 공유하는 대화로, 발화자들이 서로 상반된 논리를 가지고 대화하는 논쟁적 상황과 주제에 대해서도 공감하고 상대방의 의견을 수용하는 내용의 협력적 상황을 포함하여 수집하였다. 본 사업에서 선정한 협력적 대화 주제는 아래와 같다.

[표 7] 협력적 대화 주제

번호	협력적 대화 주제	
1	문화	영화/드라마/음악(콘텐츠)
2		연극/뮤지컬/콘서트(공연)
3		전시회/박물관(전시)
4		책/독서
5		스포츠/레저
6		패션/뷰티
7		음식/음료
8		반려동물
9	관광	여행 계획
10		여행 일반

원활한 대화 진행을 위해 협력적 대화 주제와 관련된 세부 주제와 키워드, 그리고 상세 가이드 예시를 함께 제공하여 활발하게 토론이 진행되도록 하였다.

[표 8] 협력적 대화 주제별 키워드 및 참고 자료 연결

번호	주제	세부 주제	상세 가이드 예시
1	영화/드라마/음악(콘텐츠)	영화관, 영화 관람, OTT 플랫폼, 배우, 가수, 대중 음악, 클래식 음악 등	<ul style="list-style-type: none"> ● 영화와 관련된 다양한 개인적 의견 교류 <ul style="list-style-type: none"> - 최근 관람한 영화 및 배우, 줄거리, 상영관에 대한 호불호 토론 - ‘좌석 차등제’ 폐지에 대한 개인 의견 제시 및 찬반 토론 ● 글로벌 OTT 플랫폼의 국내 영화/드라마 콘텐츠에 대한 토론 <ul style="list-style-type: none"> - 인기 요인에 대한 의견을 제시(예 : ‘오징어 게임’, ‘킹덤’, ‘스위트홈’ 등)하고 콘텐츠를 추천 ● 빌보드 차트에 오른 K-pop에 대한 긍정적/부정적 의견 교류 <ul style="list-style-type: none"> - 긍정: 글로벌 차트에서 상위권을 차지하며 K-pop의 위상을 증명 - 부정: K-pop임에도 영어 가사로만 이루어진 노래가 증가, 차트 순위로 음악의 가치가 매겨지는 흐름: 긍정적인 측면을 우세화하여 다양한 K-pop 음악을 국제 무대에 소개할 수 있는 기회로 삼는 결론
2	연극/뮤지컬/콘서트(공연)	뮤지컬, 콘서트, 연극, 클래식, 오페라, 국악, 발레/무용 등	<ul style="list-style-type: none"> ● 공연과 관련된 다양한 개인적 의견 교류 <ul style="list-style-type: none"> - 최근 관람한 공연에 대한 줄거리, 연출, 효과 등 감상평 공유 및 추천, 공연 티켓 불법 중고 거래에 대한 찬반 토론 및 결론 도출 - 개인의 자유 vs 문화 빈부 격차 ● 길거리 공연 <ul style="list-style-type: none"> - 야외무대에서의 공연에 대한 토론 (소음이다 vs 지역 발전에 도움이 된다): 누군가에게는 소음 누군가에게는 아닌 길거리 공연을 금지해야 하는지에 대한 찬반 토론
3	전시회/박물관(전시)	미술, 박물관 관람 및 여러 전시회 관람 지역별 박물관 관람 후기 및 체험 정보 공유 지역별 축제 정보 공유 및 권유	<ul style="list-style-type: none"> ● 코로나 재확산에 따른 미술관 및 박물관 온라인 서비스에 대한 개인적 의견 제시 및 토론 <ul style="list-style-type: none"> - 비대면 관람을 통해 얻을 수 있는 경험의 부재, 실효성 여부 등 찬반 토론 진행 ● 자연 박물관, 역사 박물관, 공룡 박물관, 공룡 테마파크, 백스코, 코엑스 등 ● 벽화 마을에 대한 간략한 정보 공유 및 찬반 토론(지역 발전에 도움 vs 흉물)

			<ul style="list-style-type: none"> ● 비대면 전시와 대면 전시 중 어떤 것이 더 좋은가? ● AI가 그린 그림 전시회의 그림 소유주는 AI인가? AI 제작자인가? ● 아트테크(NFT 등)가 재산으로 가치가 있는가 없는가?
4	책/독서	작가, 독서, 소설, 시집, 종이책, 전자책(E-book) 등	<ul style="list-style-type: none"> ● 최근 읽은 소설, 시집 등 감상평에 대한 내용 정보 공유 및 추천 - 선호하는 작가 및 등장인물에 대한 정보 공유 - 선호 장르 등 발화자 간의 의견 교류 및 소통을 통한 작품 추천 ● 웹툰 및 소설의 영화화, 드라마화에 성공한 사례에 대한 의견 토론 - 원작과 영화/드라마 간의 괴리감, 접근성 등 호불호 토론 - 국내외에서 영화화, 드라마화에 성공한 사례에 대한 토론 - 원작 및 영화, 드라마 추천 및 시청 권유 ● 종이책과 전자책의 장단점에 대한 토론 - 아날로그적 감성, 환경보호 등 대립적인 주제에 대한 토론 - 전자책의 시장 점유율 및 편의성에 대한 개인적 경험과 의견 - 전자책 앱 추천 및 권유
5	스포츠/레저	경기 직관, 스포츠 종목, 올림픽, 축제, 행사, 등산, 수상 레저, 클라이밍 등	<ul style="list-style-type: none"> ● 스포츠 경기에 대한 다양한 경험 공유 및 토론 - 직관/중계 관람의 차이점 토론 - 무관중 시대에 진화한 중계방송 기술에 대한 경험 공유 - 인상 깊었던 올림픽 경기나 선수에 대한 내용 ● 선호 스포츠 종목에 대한 의견 교류 및 권유, 지역 축제 및 행사에 대한 개인적인 경험과 권유 등 의견 교류 ● 개인의 취미 및 레저 스포츠에 대한 재밌었던 경험 및 추천
6	패션/뷰티	화장품, 옷, 액세서리, 유행하는 스타일, 상황에 따른 스타일	<ul style="list-style-type: none"> ● 코로나 사태로 달라진 간소화 뷰티에 대한 개인 의견 공유 - 마스크로 인해 달라진 소비 패턴(화장품, 립스틱 등) ● 기능성/패션 마스크에 대한 의견 제시 및 경험 토론 ● 코로나 시대로 온라인 쇼핑이 주가 된 요즘 판매 사진과 실제 상품의 차이에 대한 의견 토론 ● 친환경 의류 소재에 대한 개인 의견 교류 및 토론 ● 의류 분야를 넘어 다양한 패션 브랜드 전체로 확장 (이어폰 케이스, 각종 캐릭터 상품 등)
7	음식/음료	맛집, 밀키트에 대한 생각, 군것질에 대한 생각 및 어떤 음식을	<ul style="list-style-type: none"> ● 맛집 및 음식 정보 공유 및 개인 의견 교류 - 발화자의 입맛과 취향을 고려한 식당 권유 및 메뉴 추천 - 지역별 특화 식재료 및 맛집에 대한 경험 공유 - 집에서 간단히 조리할 수 있는 요리 비법 및 음식 추천 ● 음식 배달 대행 서비스에 대한 개인 의견 교류 및 토론

		먹을지 결정	<ul style="list-style-type: none"> - 배달 음식 서비스 이용의 장/단점 - 배달 라이더, 노쇼 등 새롭게 부각 된 사회적 문제에 대한 의견 - 수수료 문제 등 가맹점/라이더/이용자 간의 상생에 대한 토론 ● 코로나로 인해 급성장한 밀키트 음식 시장에 대한 토론 - 장점: 1인 식단, 간편, 취향 선택, 장소 구매 없음, 새벽 배송 등 - 단점: 원산지 표시, 품질, 환경 오염, 보관
8	반려동물	반려동물 입양, 유기 동물 문제, 동물 등록, 동물 병원 등	<ul style="list-style-type: none"> ● 개인 혹은 지인의 반려동물에 대한 이야기 공유 - 반려동물의 장/단점 의견 교류 - 반려동물 입양/추가 입양에 대한 찬반 토론 - 반려동물 보유세에 대한 개인적 주관 찬반 토론 - 반려동물 유기에 대한 사회적 문제 토론 ● 동물 문제와 관련한 토의 진행 - 유기 동물을 돈벌이로 이용하는 유기 보호 센터 찬반 토론 - 동물 등록에 대한 장/단점 토론 - 동물 등록제에 대한 찬/반 토론 - 동물 병원, 동물 호텔 등 인프라에 대한 개인적 의견 교류
9	여행 계획	교통편, 숙박, 여행지, 여행 경비 등	<ul style="list-style-type: none"> ● 다양한 교통편을 이용한 여행에 대한 개인의 경험 이야기 ● 민박, 펜션 등 숙박지 서비스 이용에 대한 경험 토론 ● 단체 여행 시 합리적인 여행 경비 지출에 대한 노하우 공유 - 각 교통 시설 이용 및 가격, 서비스에 대한 장/단점 - 비박, 장박 등 여행 스케줄에 따른 교통편 추천 및 권유 - 이용 가격이 부담했던 경험에 대한 토론 - 예약 및 취소, 환불 과정에서 겪었던 피해 사례 대화 - 숙박업소의 시설 및 서비스에 대한 평가 의견 토론
10	여행 일반	자유 여행, 패키지 여행, 국내 여행, 해외 여행, 관광지, 휴양지 등	<ul style="list-style-type: none"> ● 패키지 여행 및 자유 여행의 장단점에 대한 대화 ● 국내/해외 여행 트렌드에 대한 대화 ● 선호하는 여행의 종류(관광/휴양)에 대한 개인 의견 토론 및 발화자들이 함께 즐길 수 있는 여행에 대한 결론 도출 ● 국내외 관광지 관광세 도입 찬반 토의 - 패키지 여행 장점: 항공, 숙소, 일정 등을 여행사에서 담당하는 편리함 - 패키지 여행 단점: 낮은 자유도 - 자유 여행 장점: 높은 자유도, 시간 제약 없음 - 자유 여행 단점: 숙소, 교통편 등 모두 스스로 준비 : 각각의 장단점을 보완할 수 있는 자유 여행과 패키지 여

			<p>행을 결합한 새로운 여행 상품 정보를 교환</p> <ul style="list-style-type: none"> - 코로나 엔데믹 이후 해외 여행 수요 증가 - 해외 자연 체험 (캠핑, 오로라 등) 형태의 여행에 대한 경험 대화 - 여전한 해외 코로나 감염 우려에 따른 안전성 있는 국내 여행 선호도 증가: 코로나가 완전히 종식된 것은 아니므로 국내 자연 체험 형태 여행의 안전성이 높다는 결론 도출 가능 - 환경 오염 책임, 실효성 의문 등
--	--	--	--

2. 전문가 자문 회의 진행

사업의 원활한 수행과 체계적인 말뭉치 구축을 위해 국어국문학, 전산언어학 및 음성 인식 분야 전문가로 자문단을 구성하고 자문 회의를 두 차례 진행하였다. 첫 번째 자문 회의는 일관성 있는 고품질의 원시 말뭉치 구축을 위한 음성 전사 지침 관련 자문을 요청하였으며, 두 번째 자문 회의는 일상 대화 말뭉치 구축 사업에서 올해 처음으로 시도 되는 비통제 환경(실환경)에서의 음성 수집 방안에 대한 전문가 자문을 요청하였다.

[표 9] 1차, 2차 자문 회의 상세

1차 자문 회의	
구분	내용
회의 일시	• 2022년 8월 26일 13:00 ~ 14:30
회의 방식	• 온라인 화상 회의(Google Meet)
자문 위원	• 임수종 박사_ETRI • 오재혁 교수_건국대학교 • 장청화_교정·교열·윤문 전문가
회의 주제	• 음성 전사 지침
회의 내용	• 억양구 규정 및 효용성 관련 논의 • 긴 쉼에 의한 전사 단위 구분 관련 규정 • 음성 분절 단위 적정 길이 • 띄어쓰기가 인공지능 학습용 데이터에 미치는 영향 • 효율적인 발음 전사 수준에 대한 논의 • 억양에 의해 의미가 달라지는 경우 물음표 사용에 대한 논의 • 준말과 축약형의 구분법
2차 자문 회의	
구분	내용
회의 일시	• 2022년 9월 16일 14:30 ~ 15:30
회의 방식	• 온라인 화상 회의(Google Meet)
자문 위원	• 임수종 박사_ETRI • 박전규 박사_ETRI • 이항섭 전무_한국음성학회(KSSS) 상임이사
회의 주제	• 비통제 환경(실환경) 음성 수집 방안
회의 내용	• 실환경 대화 화자 수집안 • 활용성 높은 고품질 데이터 구축을 위한 수집 기준에 대한 논의 • 실환경 녹음 시 발생할 수 있는 변수 및 위험을 최소화하기 위한 주의사항 • 소음이 포함된 음성 데이터가 학술 연구 및 음성 인식 기술 개발에 가지는 효용성 • 적정 소음의 유형 및 dB • 소음이 포함된 음성 데이터 활용 분야

3. 화자 구성 및 모집

발화자 모집은 설계 단계에서는 모집 인원 총 2,800명으로 2021년 통계청 지역별 인구 분포 자료를 참고하여 전국 16개 지역으로 나누어 모집하였다. 성별, 나이별, 지역별(주 성장지 기준), 주제별로 비율이 편중되지 않도록 하였다. 성별, 나이별, 지역별, 주제별 분포 비율은 작업 과정에서 정제 및 대화 내용이 적합하지 않아 제외되는 데이터를 고려하여 계획되었다.

지역은 현 거주지가 아닌 주 성장지 기준으로 서울특별시, 6대 광역시(인천, 대전, 대구, 부산, 광주, 울산), 9개 도(경기, 강원, 충남, 충북, 경남, 경북, 전남, 전북, 제주)로 할당하였다. 세종시의 경우, 2021년 출범한 세종시를 주요 성장지로 하는 대상자를 찾기 어려워 대전의 인원으로 통합하였다.

발화자의 나이대는 10세 단위로 10대, 20대, 30대, 40대, 50대, 60대 이상으로 나누었고, 10세 이하는 현실적으로 녹음이 어려워 모집 대상에서 제외하였다. 또한, 10대와 60대 이상은 녹음 수집이 어려운 관계로 인원을 통합하여 균등 할당하였다.

[표 10] 사업 초기 화자 할당표 설계 기준

구분	기준
모집단	• 2021년 통계청 지역별 인구 분포 자료 기준
고려 변수	<ul style="list-style-type: none"> • 성별: 남자/여자 • 나이대: 10대/20대/30대/40대/50대/60대 이상 • 지역(주 성장지 기준): 서울/인천/대전/대구/부산/광주/울산/경기/강원/충남/충북/경남/경북/전남/전북/제주 • 2012년 출범한 세종시를 주 성장지로 하는 대상자를 찾기 어려워 별도로 수집하지 않음
배분 방법	• 제곱근 비례 배분
표본 할당	<ul style="list-style-type: none"> • 지역별: 비례 할당 • 성별×나이별: 균등 할당

초기 설계된 성별 및 나이별 지역별 모집 목표는 아래와 같다.

[표 11] 성별 및 나이별 지역별 모집 목표(단위: 명)

	인구 (명)	비율(%)	나이						총수집 인원
			10대	20대	30대	40대	50대	60대 이상	
			10%	20%	20%	20%	20%	10%	100%
전국	51,583,722	100	270	539	539	539	539	270	2,696
서울	9,496,887	18.41	50	99	99	99	99	50	496
경기	13,581,496	26.33	71	142	142	142	142	71	710
인천	2,955,167	5.73	15	31	31	31	31	15	154
충남	2,119,661	4.11	11	22	22	22	22	11	110
대전	1,448,933	2.81	8	15	15	15	15	8	76
세종	379,340	0.74	2	4	4	4	4	2	20
충북	1,597,033	3.1	8	17	17	17	17	8	84
대구	2,376,676	4.61	12	25	25	25	25	12	124
경북	2,616,177	5.07	14	27	27	27	27	14	136
부산	3,338,167	6.47	17	35	35	35	35	17	174
경남	3,298,016	6.39	17	34	34	34	34	17	170
울산	1,116,482	2.16	6	12	12	12	12	6	60
광주	1,436,012	2.78	8	15	15	15	15	8	76
전남	1,827,674	3.54	10	19	19	19	19	10	96
전북	1,779,230	3.45	9	19	19	19	19	9	94
강원	1,539,005	2.98	8	16	16	16	16	8	80
제주	677,766	1.31	4	7	7	7	7	4	36

실환경 데이터의 수집을 위해 전문가 자문 회의를 진행하였으며, 이를 통해 확정된 통제 및 비통제(실환경) 데이터 수집의 비율과 환경은 아래와 같다.

[표 12] 통제 대화 성별 및 나이별 지역별 모집 목표(단위: 명)

	인구 (명)	비율(%)	나이						총수집 인원
			10대	20대	30대	40대	50대	60대 이상	
			10%	20%	20%	20%	20%	10%	100%
전국	51,583,722	100	236	472	472	472	472	236	2,360
서울	9,496,887	18.41	43	87	87	87	87	43	434
경기	13,581,496	26.33	62	125	125	125	125	62	624
인천	2,955,167	5.73	14	27	27	27	27	14	136
충남	2,119,661	4.11	10	19	19	19	19	10	96
대전	1,828,273	3.55	8	17	17	17	17	8	84
충북	1,597,033	3.1	8	14	14	14	14	8	72
대구	2,376,676	4.61	11	22	22	22	22	11	110
경북	2,616,177	5.07	12	24	24	24	24	12	120
부산	3,338,167	6.47	14	31	31	31	31	14	152
경남	3,298,016	6.39	15	30	30	30	30	15	150
울산	1,116,482	2.16	5	10	10	10	10	5	50
광주	1,436,012	2.78	7	13	13	13	13	7	66
전남	1,827,674	3.54	8	17	17	17	17	8	84
전북	1,779,230	3.45	9	16	16	16	16	9	82
강원	1,539,005	2.98	7	14	14	14	14	7	70
제주	677,766	1.31	3	6	6	6	6	3	30

[표 13] 비통제(실환경) 대화 권역별 데이터 수집 목표 인원 및 시간

권역	비중(%)	인원(명)	시간
수도권	50	200	50
영남	15	60	15
충청	15	60	15
호남	15	60	15
강원	2.5	10	2.5
제주	2.5	10	2.5
합계	100	400	100

[표 14] 비통제 대화 카테고리별 데이터 수집 목표 시간

카테고리	장소	비중(%)	수집 시간
무소음 환경	조용한 사무실 환경	12.5	13
	회의실 환경	12.5	13
생활 환경	생활 소음(세탁기, 선풍기, TV 등)이 포함된 가정 환경	12.5	13
	카페 환경	12.5	13
교통 환경	도로변, 버스 정류장 등(야외)	12.5	13
	지하철 대합실 및 승강장	12.5	13
자연 환경	시민 공원, 산책로 등(야외)	12.5	13
	하천 등 생태 환경(야외)	12.5	13
합계	-	100	104

데이터 수집을 진행하면서 고품질 데이터 수집 및 발화자 섭외 문제 등의 사유로 데이터 수집 계획을 변경하였으며, 담당자 회의 등을 거쳐 최종적으로 아래와 같은 기준이 추가되었다.

- 1인당 대화 수집 시간을 기존 30분에서 1시간으로 확대
- 수집되는 총 화자 인원은 2,000명 이상이 되도록 조정
- 나이별 수집 비율 완화 조치($\pm 5\%$)
- 성별 비중은 기존안의 비율을 그대로 수집함(50:50)
- 제주 및 강원에서 수집하는 통제 및 비통제 데이터는 기존 목표 인원을 최대한 섭외하여 수집하는 것을 원칙으로 하며, 이로 인하여 추가로 수집되는 데이터는 수도권 수집 개수에서 차감하여 데이터가 버려지지 않도록 비율을 조정함

[표 15] 통제 대화 발화자 모집 계획 변경안(단위: 명)

	인구 (명)	비율(%)	나이						총수집 개수
			10대	20대	30대	40대	50대	60대 이상	
			5%~15%	15%~25%	15%~25%	15%~25%	15%~25%	5%~15%	100%
전국	51,583,722	100	118~354	354~590	354~590	354~590	354~590	118~354	2,360
서울	9,496,887	18.41	19~65	59~108	59~108	59~108	59~108	19~65	397~434
경기	13,581,496	26.33	28~93	85~155	85~155	85~155	85~155	28~93	569~621
인천	2,955,167	5.73	6~20	18~33	18~33	18~33	18~33	6~18	123~135
충남	2,119,661	4.11	4~14	14~24	14~24	14~24	14~24	4~14	97
대전	1,828,273	3.55	4~12	12~21	12~21	12~21	12~21	4~12	84
충북	1,597,033	3.1	3~10	10~18	10~18	10~18	10~18	3~10	73
대구	2,376,676	4.61	5~16	16~27	16~27	16~27	16~27	5~16	109
경북	2,616,177	5.07	6~18	18~30	18~30	18~30	18~30	6~18	120
부산	3,338,167	6.47	7~22	22~38	22~38	22~38	22~38	7~22	153
경남	3,298,016	6.39	7~22	22~37	22~37	22~37	22~37	7~22	151
울산	1,116,482	2.16	2~7	7~12	7~12	7~12	7~12	2~7	51
광주	1,436,012	2.78	3~9	9~16	9~16	9~16	9~16	3~9	66
전남	1,827,674	3.54	4~12	12~21	12~21	12~21	12~21	4~12	84
전북	1,779,230	3.45	4~12	12~20	12~20	12~20	12~20	4~12	81
강원	1,539,005	2.98	7~14	14~28	14~28	14~28	14~28	7~14	70~140
제주	677,766	1.31	3~6	6~12	6~12	6~12	6~12	3~6	30~60

[표 16] 비통제 대화 발화자 모집 계획 변경안

권역	비중(%)	인원(명)	시간
수도권	45	180	45
영남	15	60	15
충청	15	60	15
호남	15	60	15
강원	5	10	5
제주	5	10	5
합계	100	380	100

화자는 네이버 카페를 활용하여 모집을 진행하였다. 녹음 아르바이트를 소개하고 각 지역 녹음 장소의 위치를 공개하였으며, 음성 녹음 아르바이트 신청, 녹음 후기 등을 작성할 수 있는 게시판을 마련하였다. 이로써 발화자들이 안심하고 안전하게 녹음 장소로 찾아올 수 있게 되었으며 다양한 발화자 섭외가 가능하게 되었다. 녹음 인원은 최소 2인 1조 이상 신청자를 기본으로 하였으며, 과제 종료가 다가올수록 인원 섭외가 어려워져 1인 신청도 허용하였다. 이때 모집 기준은 섭외가 어려운 40대 이상의 남성을 최우선으로 하였다. 또한, 섭외가 어려운 40대, 50대, 60대 등 중장년 화자를 수집하기 위해 지인 추천 이벤트도 진행하였다. 음성 데이터 수집을 완료한 대상자들에게 말뭉치 수집 관련 정보를 전달하여 부모나 가족, 친척, 직장 동료 등 화자의 주변 지인들이 데이터 수집에 참여할 수 있도록 유도하였다. 이렇게 다양한 경로로 모집된 화자를 녹음 진행 요원이 녹음 가능한 날짜를 협의해 녹음 날짜를 확정하였으며, 녹음 당일 아침 녹음 진행 요원이 다시 한번 전화를 통해 녹음 가능 여부를 확인하였다.

제목	작성자	작성일	조회
필독! 2022. 국립국어원. 일상대화 녹음 알바 공지글입니다.	스마트미디어테크	2022.06.28.	1.3만
2022_일상... 녹음 장소 : 강원 춘천시 은의동(춘천 시외버스터미널) ☎ ☎	스마트미디어테크	2022.12.15.	176
2022_일상... 녹음 장소 : 경기 수원시 팔달구(수원역) ☎ ☎ [3]	스마트미디어테크	2022.11.22.	657
2022_일상... 녹음 장소 : 경기 성남시 분당구(아람역) ☎ ☎ [7]	스마트미디어테크	2022.11.09.	806
녹음 알바 ... 합정에서 두명 녹음 잘 하고 가요 ☎	morsel11	2022.10.21.	978
녹음 알바 ... 합정 녹음 알바 후기 올려요	키오피오	2022.10.21.	603
녹음 알바 ... 서울 합정 녹음알바 후기! ☎	월소	2022.10.20.	711
2022_일상... 녹음 장소 : 인천 미추홀구(주안역) ☎ ☎ [5]	스마트미디어테크	2022.10.14.	893
2022_일상... 일상대화 녹음 알바! 이렇게 홍보해 주세요~ ☎ ☎ [29]	스마트미디어테크	2022.09.27.	1,985
2022_일상... 녹음 장소 : 광주 동구 (금남로4가역) ☎ ☎ [7]	스마트미디어테크	2022.09.06.	1,308
2022_일상... 녹음 장소 : 대전 서구 (시정역) ☎ ☎ [20]	스마트미디어테크	2022.09.06.	1,555
2022_일상... 녹음 장소 : 대구 중구 (경대병원역) ☎ ☎ [11]	스마트미디어테크	2022.08.29.	1,485
2022_일상... 녹음 장소 : 부산시 부산진구 (서면/전포역) ☎ ☎ [13]	스마트미디어테크	2022.08.19.	2,351
2022_일상... 녹음 장소 : 전라북도 익산시 (영동동) ☎ ☎	스마트미디어테크	2022.08.11.	805
2022_일상... 녹음 장소 : 서울시 마포구 (합정역) ☎ ☎ [36]	스마트미디어테크	2022.07.22.	4,650
2022_일상... 필독! 2022. 국립국어원. 일상대화 녹음 알바 공지글입니다. ☎	스마트미디어테크	2022.06.28.	1.3만

[그림 4] 말뭉치 구축 사업 참여자 모집 카페

2022년 일상대화 말뭉치 구축

지인과 함께 녹음하실 분들을 초대합니다!

(주)스마트미디어테크에서는 아래와 같이 일상 대화를 수집하고 있습니다.
수집된 자료는 국가적인 데이터 구축 사업에 활용될 예정이오니 많은 참여 부탁드립니다.

신청대상 10대(만 15세) 이상
가족, 친구, 직장 동료, 선후배, 지인 등 동반 참여 환영!
10대 미성년자인 경우 보호자(부모님) 동의가 필수입니다.

진행방법 지정 장소에 지인과 함께 참석하여 자유 대화(2인~4인)
2인 : **1시간**, 3~4인 : **2시간**(사전 교육 포함, 원하는 요일, 시간 선택 가능)
(참여자 개인의 역량에 따라 녹음 시간은 변동될 수 있음)

모집기간 2022년 8월 16일 ~ 모집 시까지

녹음장소 서울, 부산, 대구, 광주, 대전, 인천, 제주, 강원 등 순차 오픈 예정

녹음비용 2인 : 1인당 **30,000원**, 3~4인 : 1인당 **60,000원** 현장 지급

지원방법 네이버 카페 가입 후 녹음 신청 댓글 작성
카페 주소 : cafe.naver.com/smartmediatech

문의처 ㈜스마트미디어테크



※ 참여비 3~6만원을 지급하오니 주위 분들에게 많은 홍보 부탁드립니다.

본 녹음은 1인당 1회만 가능합니다.



문화체육관광부
국립국어원



SMART MEDIA TECH
1주 스마트미디어테크

[그림 5] 말뭉치 구축 사업 참여자 모집 공고

4. 작업자 선발 및 교육

4.1. 녹음 진행 요원 선발 및 교육

녹음은 전북, 서울, 부산, 대구, 광주, 대전, 경기, 인천, 강원, 제주 등 12개 지역에서 순차적으로 진행하였다. 16개의 지역 거주자 2,073명의 화자가 녹음에 참여하였으며, 다수의 화자가 참여하는 만큼 원활한 녹음 진행을 위하여 각 지역별로 녹음 진행 요원이 투입되었다. 지역별로 투입된 진행 요원은 13명으로 코로나-19의 상황을 고려하여 온라인으로 면담을 진행한 후 자격 조건과 맞는 지원자를 대상으로 오프라인 교육을 실시하였으며, 교육 과정 중 평가에 통과한 사람을 최종 선발하였다. 진행 요원은 기존 유사 작업 경험자를 우선하여 선발하였다.

[표 17] 진행 요원 선발 및 운영 방안

구분	선발 기준 및 운영 내용	
선발 기준	• 전문 녹음 장비 작동 경험이 있는 사람 (우선 선발) • 최종 교육 이수 및 평가 통과자 (필수)	
투입 인원	• 진행 요원 13명	
진행 요원 역할	• 진행 요원 1 - 화자 안내 - 코로나-19에 따른 건강 체크 - 화자 인적 사항 확인 - 화자 참석 관리 및 스케줄 관리 - 녹음 종료 후 사례비 지급	• 진행 요원 2 - 녹음 진행 개요 설명 - 저작권 이용 허락 계약 체결 및 개인정보 동의서 관리 - 녹음 장비 이상 유무 확인 - 녹음 진행

말뭉치 수집 일정 및 품질에 차질이 없도록 녹음 진행 요원 및 관리 인원을 대상으로 교육을 수행하였다. 교육은 기본 4단계로 진행하였으며, 교육 내용은 아래와 같다.

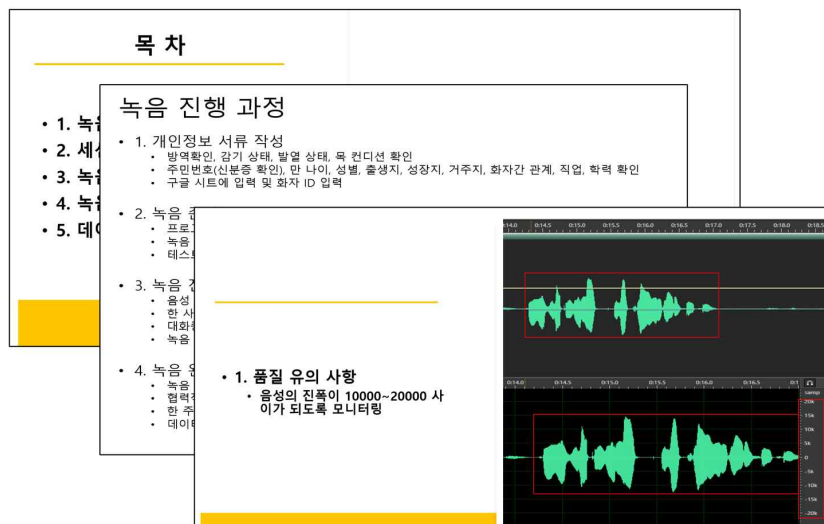
- 사업 배경 및 목적, 진행 시 유의 사항 등의 이론 교육
- 녹음 장비 작동 방법, 헤드셋 마이크 착용 방법, 녹음 진행 등의 실사 교육
- 화자 응대, 화자의 불만 제기 시 대처 방법 등의 CS 교육
- 화자 개인정보 관리, 녹음 자료 관리 등의 보안 교육

기본 교육을 마친 진행 요원은 실제 녹음으로 들어가기에 앞서 화자 응대, 녹음 장비

작동에 대한 시뮬레이션과 특이 사항 발생 시 대처 요령에 대한 모의 훈련을 실시하였다. 모의 평가에서 역할에 대한 이해도가 높은 사람을 최종적으로 선발하였으며 선발된 녹음 진행 요원들은 보안 서약서 작성 후 실제 녹음 진행에 참여하였다.

[표 18] 진행 요원 교육 내용

구분	내용
교육 일시 및 장소	<ul style="list-style-type: none"> • 2022년 08월, 서울, 익산, 부산 녹음 지역 • 2022년 09월, 광주, 대전, 대구 녹음 지역 • 2022년 10월, 인천, 수원, 성남 녹음 지역 • 2022년 12월, 강원, 제주 녹음 지역
교육자	<ul style="list-style-type: none"> • 김태권(㈜스마트미디어테크)
교육 내용	<ul style="list-style-type: none"> • 사업의 배경 및 목적 • 진행 절차 • 대화 주제 • 녹음 환경 및 녹음 장비 사용법 • 녹음 방법 • 녹음 시 주의 사항 및 녹음 진행 시 제스처 학습 • 녹음 시뮬레이션 실습 • 보안 교육 • 질의응답



[그림 6] 녹음 진행 요원 교육 자료 일부

4.2. 전사 작업자 선발 및 교육

전사 작업을 위한 인력은 2022년 9월 2일에서 2023년 1월 24일까지 총 43명의 작업자가 투입되었으며, “2021년 일상 대화 말뭉치 구축 사업” 작업 결과 우수자를 중심으로 20년 이상 경력의 교정 교열 전문가 21명, 5년 이상 도서관 원문, 선거 및 서지 목록 DB 구축 사업 참여자 등 전문 DB 구축 인력 22명으로 선발 구성하였다.

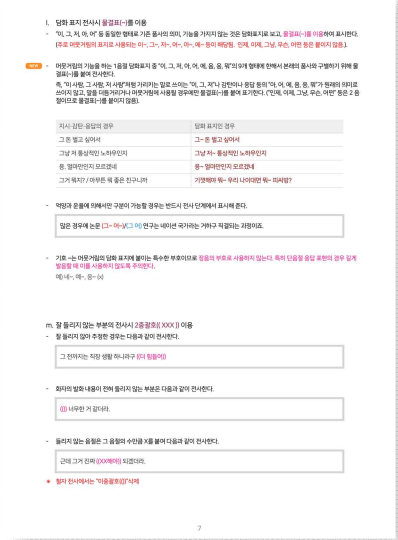
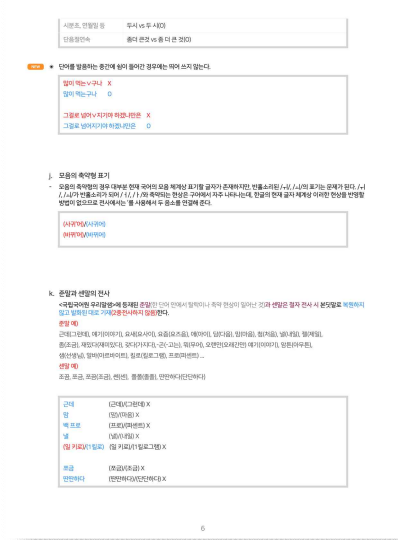
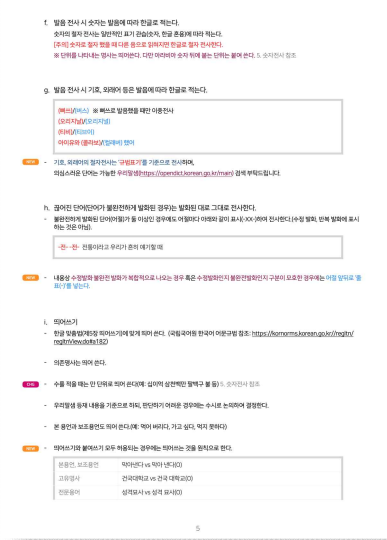
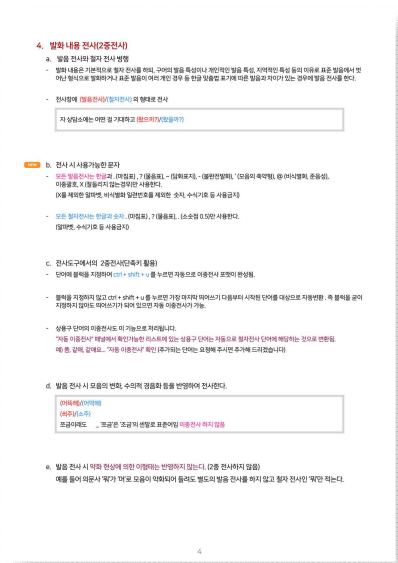
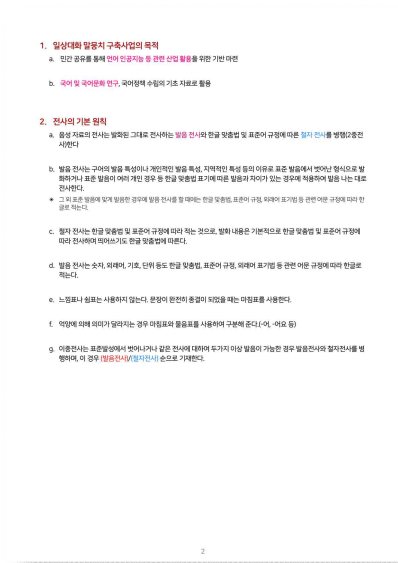
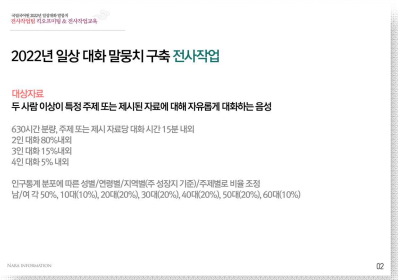
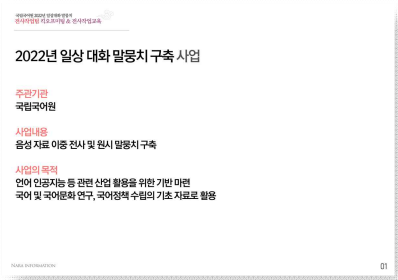
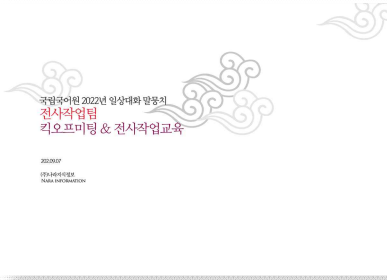
전사 인력은 전사 지침 교육, 전사 도구 활용 교육과 전사 단위(역양구) 개념 교육 후 2주간의 테스트 전사 기간을 거쳐 전사 작업에 투입되었으며, 비대면 재택근무를 통해 작업을 진행하였다.

[표 19] 전사 작업자 선발 기준 및 운영

구분	선발 기준 및 운영 내용
선발 기준	<ul style="list-style-type: none"> • 2021년 일상 대화 말뭉치 구축 사업 작업 결과 우수자 • 교정 교열 전문 작업자 • DB 구축 전문 작업자
월별 투입 인력	<ul style="list-style-type: none"> • 2022년 9월: 20명 • 2022년 10월: 23명 • 2022년 11월: 25명 • 2022년 12월: 30명 • 2023년 1월: 40명
운영	<ul style="list-style-type: none"> • 매주 목요일 전사 과정에 대한 리포트 및 변경된 전사 지침 관련 온라인 회의 진행함. • 음성 파일 중 방언 등의 특징이 있을 경우 해당 지역 출신 전사자에게 우선 배정함. • 전사자별 선호 주제를 정해 해당 주제 관련 음성 대화 우선 배정함. • 음성 발화자와 유사한 나이대의 전사자에 관련 음성 대화 우선 배정함.

[표 20] 전사 작업자 교육

구분	내용
작업자 전체 교육	<ul style="list-style-type: none"> • 일시: 2022년 9월 2일 • 장소: (주)나라지식정보 회의실 • 교육자: 박영훈(전체 교육 진행 및 공지, 나라지식정보) 박혜련(전사 지침, 유의 사항 및 맞춤법 교육, 나라지식정보) 윤기현(전사 도구 사용 교육, 바이칼AI) • 교육 참석 인원: 23명(전사 작업자 19명, 전사 작업 참여 예정자 4명) • 교육 내용: 1. 사업의 배경 및 목적, 전사 절차와 방법 2. 전사 지침 및 유의 사항 3. 전사 도구 사용 교육 4. 한글 맞춤법 주요 내용 및 오류 사례 5. 질의응답
신규 작업자 교육 1차	<ul style="list-style-type: none"> • 일시: 2022년 10월 28일 오후 5시(온라인) • 교육자: 박영훈(전체 교육 진행 및 공지, 나라지식정보) 박혜련(전사 지침, 유의 사항 및 맞춤법 교육, 나라지식정보) • 교육 참석 인원: 8명(전사 작업자 3명, 전사 작업 신규 참여자 5명) • 교육 내용: 1. 사업의 배경 및 목적, 전사 절차와 방법 2. 전사 지침 및 유의 사항 3. 전사 도구 사용 교육 4. 질의응답
신규 작업자 교육 2차	<ul style="list-style-type: none"> • 일시: 2022년 12월 28일 오후 5시(온라인) • 교육자: 박영훈(전체 교육 진행 및 공지, 나라지식정보) 박혜련(전사 지침, 유의 사항 및 맞춤법 교육, 나라지식정보) • 교육 참석 인원: 20명(11월 12월 전사 작업 신규 참여자) • 교육 내용: 1. 사업의 배경 및 목적, 전사 절차와 방법 2. 전사 지침 및 유의 사항 3. 전사 도구 사용 교육 4. 질의응답
전사 작업자 온라인 회의	<ul style="list-style-type: none"> • 일시: 2022년 8월 25일 ~ 2023년 1월 13일 사이 매월 첫째/셋째 금요일 오후 5시(온라인) • 회의 참석자: 박영훈, 박혜련, 전사 작업 및 검수자 전원 • 교육 내용: 1. 변경된 전사 지침 공유 2. 전사 관련 이슈 사항 점검 3. 전사 오류 사례 검토 4. 전사 규칙 관련 질의응답



[그림 7] 전사 교육 자료(일부)

4.3. 개인정보 보호 및 보안 교육

이 사업이 정보화 용역사업으로 편성됨에 따라, 사업의 원활한 진행을 위해 정보보안 교육과 개인정보보호 교육을 시행하였다. 먼저, 정보보안 교육의 경우 본 사업 참여 인력 전원을 대상으로 하였으며, 온라인으로 자체 교육을 진행하였다. 주요 내용은 취급 정보 보안, PC 보안, 개인정보 취급 방법, 문서 및 자료 관리와 사무실 및 장비 관리, 개인정보보호법에 따른 개인정보 처리 방법 등이었다. 문화체육관광부의 개인정보 보호 지침, 보안업무 규정 시행세칙 및 정보화 업무 규정집과 행정안전부의 개인정보보호법 주요 내용 교육 자료 등을 참고하여 교육을 진행하였다.

개인정보보호 교육은 개인정보를 직접 수집하고 처리하는 기관의 참여 인력을 대상으로 진행하였으며, 개인정보보호위원회가 운영하는 개인정보보호 포털(privacy.go.kr)의 온라인 교육인 ‘(신)개인정보보호법 이해하기’를 개별적으로 수강하고 교육 수료증을 발급받았다. 주요 교육 내용은 개인정보 보호법의 의의, 개인정보 처리 단계별 보호, 개인정보의 안전한 관리, 정보주체 권리 보장과 개인정보 유출 및 대응 등이었다.

[표 21] 개인정보 보호 및 보안 관련 교육

구분	내용
보안 교육	<ul style="list-style-type: none"> • 2022년 8월 3일 14:00~16:00 (온라인 교육) • 참여자: 사업 참여자(17명) • 내용: 사업 수행 과정에서 취득, 생산 및 유통되는 데이터 및 사용 장비와 관련하여 발생할 수 있는 위험/보안 요소에 대한 유의 사항 교육, 정보보안 처리 규정 확인 및 개인정보 취급상의 유의 사항 교육 • 참고 자료: 국가사이버안전관리규정, 문화체육관광부 보안업무규정 시행세칙, 행정안전부(KISA) 개인정보보호법_주요내용교육자료 등
개인정보보호 교육	<ul style="list-style-type: none"> • 2022년 10월 중 개별적 온라인 교육 수강(privacy.go.kr) • 참여자: 개인정보 직접 수집·처리 기관 참여 인력(4명) • 내용: 개인정보보호법 교육

5. 음성 녹음

5.1. 녹음 환경

녹음은 전국 12개 지역(전북, 서울, 부산, 대구, 광주, 대전, 경기, 인천, 강원, 제주)에서 인구 비율에 따라 최소 0.5개월에서 5개월까지 화자를 모집하여 수집하였다. 각 지역 행정구역에 광역시가 있는 경우, 음성 녹음 수집 장소를 광역시에 구축하여 해당 행정구역의 인원도 함께 모집 및 수집하였다. 녹음은 통제 녹음과 비통제 녹음으로 구분되며, 수집 장소와 수집 장비가 서로 달라 두 가지 녹음 환경을 구축하였다.

5.1.1. 통제 환경

통제 녹음은 화자가 외부와 차단된 상태로 녹음에 참여할 수 있도록 구성하였다. 화자가 편안하게 이야기할 수 있는 조용한 사무실 또는 가정집을 마련하여 상대방의 목소리가 최대한 들어가지 않도록 화자 간의 거리가 1m 이상 떨어진 공간에서 녹음을 진행하였다. 또한, 울림을 최소화하고 외부의 소음이 차단될 수 있도록 거치형 걸개를 이용하여 흡음재를 설치하였다.

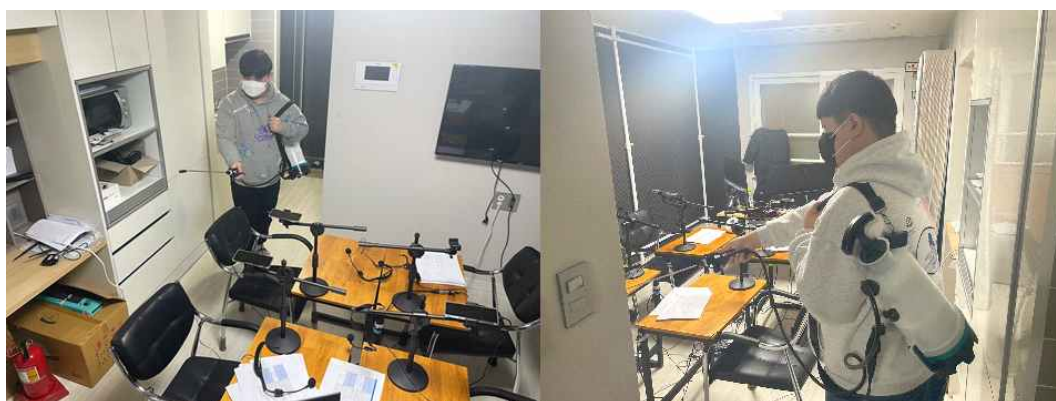


[그림 8] 통제 대화 수집을 위한 녹음 장비 및 환경

폐쇄된 공간에서 2인 이상의 대화 내용을 수집해야 하므로 코로나-19를 방지할 수 있도록 방역 수칙을 지키고 소독 및 발열 체크와 마스크 등 물품을 비치하였다. 발화자별 녹음 시간대를 조정하여 최대한 많은 인원이 부딪히지 않도록 하였고 모집 및 교육은 비대면으로 진행하는 등 코로나-19 감염에 유의하였다.

[표 22] 코로나-19 집단 감염 방지 화자 관리 방안

구분	내용
대책 분야	• 화자 모집, 음성 녹음, 참여자 교육, 회의 등 사업 전반
화자 모집	• 전화, SMS 접수
교육	• 전화, 녹음 진행 전 현장 교육
녹음 시간	• 참여자별 녹음 시간 조정
방문자	• 방문자 체온 검사, 호흡기 증상 확인
녹음실	• 녹음실 내 화자 간격 조정 • 녹음실에서는 항상 마스크 착용
방역 관련	• 소독제 및 마이크 일회용 덮개 등 방역 관련 물품 사용 • 사업장 전체를 매일 녹음 시작 전 환경 소독, 환기 실시 • 녹음 화자가 변경될 때마다 환기 실시 • 감염 관리 전담 직원 지정
인력 관리	• 방문자 및 종사자 목록 관리 • 유증상자 출근, 이용 중단 및 업무 배제



[그림 9] 녹음 시작 전 녹음 장소 방역 진행

수집 장비는 수집 기관의 과제 수행 경험을 통해 검증된 장비인 Focusrite의 Scarlett solo 18i8 오디오 인터페이스 및 Shure사의 SM35-TQC Closetalk Mic로 운용하였다. 음성 녹음은 16khz로 샘플링하였으며, 16bit로 양자화하였다.



[그림 10] 녹음 장비 및 장비 테스트

5.1.2. 비통제 환경

비통제 녹음은 실생활의 소음이 포함될 수 있는 사무실, 회의실, 카페, 가정집 등의 장소에서 진행하였으며, 포함되는 소음의 종류에 따라 크게 4가지로 구분하였으며 세부적으로 8가지 카테고리를 설정하였다. 비통제 녹음은 두 명의 남녀 화자가 각 수집 카테고리별 장소에 참석하여 수집 장비 앞에 편안하게 앉아서 주제에 맞는 대화를 나누도록 진행하였다.

[표 23] 비통제 환경 녹음 성장지별 수집 목표 시간

권역	비중(%)	인원(명)	시간
수도권	45	180	45
영남	15	60	15
충청	15	60	15
호남	15	60	15
강원	5	10	5
제주	5	10	5
합계	100	380	100

[표 24] 비통제 환경 녹음 카테고리별 수집 목표 시간

카테고리	장소	비중(%)	수집 시간
무소음 환경	조용한 사무실 환경	12.5	13
	회의실 환경	12.5	13
생활 환경	생활 소음(세탁기, 선풍기, TV 등)이 포함된 가정 환경	12.5	13
	카페 환경	12.5	13
교통 환경	도로변, 버스 정류장 등(야외)	12.5	13
	지하철 대합실 및 승강장	12.5	13
자연 환경	시민 공원, 산책로 등(야외)	12.5	13
	하천 등 생태 환경(야외)	12.5	13
합계	-	100	104

수집 장비는 발주 기관과 협의를 통해 안드로이드 기반의 스마트폰(삼성 갤럭시 S10, 삼성 갤럭시 S20)과 iOS 기반의 스마트폰(iPhone XS, iPhone 11) 등을 활용하였다. 소음이 심한 야외 환경에서는 외장 마이크(BOYA BY-MM1)를 사용(교통 환경, 자연 환경)하여 녹음을 진행하였다. 무소음 환경을 제외한 생활 환경, 교통 환경, 자연 환경의 경우 일반적으로 30dB 이상 120dB 이하의 소음이 발생하는 환경에서 녹음을 진행하였다. 약 30dB 이상의 소음이 측정되는 경우 실환경 수집의 의미가 있는 환경으로 판단하여 녹음을 진행했으며, 120dB 이상의 소음이 발생하는 경우 대화가 불가능하다고 판단하여 녹음을 진행하지 않거나 발화자가 대화를 잠시 멈추도록 하였다. 또한, 30dB~120dB 사이

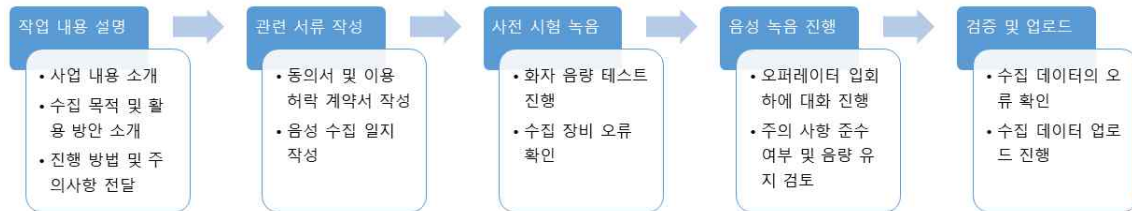
의 소음이 발생하는 환경이더라도, 오퍼레이터가 현장의 소음을 인식하였을 때 원활한 대화 및 전사가 가능하다고 판단되는 경우에만 녹음을 진행하였다. 휴대용 삼각대를 이용하여 두 명의 발화자의 입 근처에 수집 장비를 거치시키고 전사가 가능한 볼륨이 수집 되도록 발성 테스트 후 녹음을 진행하였다. 음성 녹음은 통제 환경과 마찬가지로 16khz 샘플링, 16bit 양자화하였다.



[그림 11] 외장 마이크 장착

5.2. 음성 녹음 절차

모집된 화자들이 자신들이 예약한 시간에 수집 장소에 도착하면 앞에서 설명된 코로나-19 방역 절차를 먼저 마친 후, 수집에 참여하게 된다. 이때 진행되는 절차는 아래와 같은 단계를 거치게 된다. 각 단계에 대한 상세 내용은 아래와 같다.



[그림 12] 음성 녹음 절차

5.2.1. 작업 내용 설명

화자 모집 시 간략하게 설명된 사업의 내용 및 데이터 수집 목적, 활용 방안에 대해 화자들에게 자세한 설명을 진행하고, 음성 데이터를 수집하는 진행 과정과 장비 착용 방법, 실제 수집 진행 시 화자가 주의해야 할 내용들을 충분히 설명하였다.

이때 개인정보의 수집 및 활용에 관하여 민감한 반응을 보이는 참여자들의 경우, 실제 수집된 데이터가 연구 및 학술 목적으로만 사용될 뿐 상업적으로 활용되지 않는다는 점을 충분히 설명하여, 최대한 참여자의 이탈을 막는 데에 주력하였다. 이렇게 사업 목적 및 개인정보와 수집 데이터의 활용에 대하여 화자가 동의하면 서류 작성을 진행하였다.

5.2.2. 관련 서류 작성

작업 내용 설명 단계에서 구두 동의를 한 화자들을 대상으로 저작권 동의에 관련된 서류 작성을 진행하였다. 작성된 서류들은 사업의 결과물(음성 파일, 전사 파일) 및 그 변형물에 대한 복제권, 전송권, 배포권, 2차적 저작물 작성권에 대하여 국립국어원에서 활용한다는 것을 허락하는 문서로, 수집에 참여하는 모든 화자가 작성하는 것을 원칙으로 하였다.

다만 구두 동의를 하였지만, 실제 서류 작성을 진행할 시 일부 참여자들이 서류 작성을 거부하거나, 또는 서류에 개인정보의 일부(주민등록번호 앞자리(생년월일))를 기재하는 것을 거부하는 사례가 종종 있었는데, 이 경우 앞 단계에서와 같이 설득해보고, 강경하게 거부 의사를 표시한 경우 일정 비용(교통비)을 지급하고 돌려보냈다.

[illegible]

[그림 13] 개인정보 활용 동의서(예시)²⁾

2) [붙임 2], [붙임 3], [붙임 4] 참조

화자번호	이름	나이(만)	직업	성별	출생지	성장지(~고등학생)	거주지역	학력	관계
3501	이정현	39	무직	남	부산	부산	부산	대학졸업	친구
SD2200001	김정희	48	주부	여성	부산	부산	경기	대졸	친구
SD2200002	김영숙	43	주부	여성	서울	서울	경기	고졸	친구
SD2200003	박영선	46	전문가 및 관련 종사자	남성	서울	서울	인천	대학원 이상	직장 동료
SD2200004	이영숙	46	전문가 및 관련 종사자	여성	경북	경북	인천	대학원 이상	직장 동료
SD2200005	최기훈	34	학생	남성	서울	서울	서울	대재	친구
SD2200006	고종현	34	무직/취업준비생	여성	서울	경기	경기	대졸	친구
SD2200007	최정영	34	학생	남성	서울	경기	경기	대재	친구
SD2200008	황정희	34	주부	여성	대전	대전	서울	대졸	모임_동아리 지인
SD2200009	신유진	32	주부	여성	전북	전북	서울	고졸	모임_동아리 지인
SD2200010	김지영	31	주부	여성	경북	경북	서울	고졸	모임_동아리 지인
SD2200011	박영선	30	주부	여성	전남	전남	서울	고졸	모임_동아리 지인
SD2200012	서기훈	30	주부	여성	서울	서울	서울	대졸	친구
SD2200013	김유나	29	주부	여성	서울	서울	서울	대졸	친구
SD2200014	김민복	29	학생	남성	대전	대전	경기	대재	친구
SD2200015	김재영	29	학생	남성	서울	경기	서울	대재	친구
SD2200016	이종현	29	학생	남성	서울	경기	서울	대재	친구
SD2200017	황우영	36	기타	남성	서울	경기	서울	대졸	모임_동아리 지인
SD2200018	이성훈	32	기타	남성	서울	서울	서울	대졸	모임_동아리 지인
SD2200019	윤정희	35	학생	여성	서울	서울	서울	초졸 이하	부모/자녀

[그림 16] 음성 자료 수집 일지(예시-2)

5.2.3. 사전 시험 녹음

화자들이 개인정보 활용동의서, 저작권 이용허락계약서 작성을 완료하면 진행 요원은 실제 녹음이 진행되는 공간으로 화자를 이동 시켜 자리 배치 및 수집 장비 착용을 안내해 주었다. 이후 진행 요원은 화자들이 선택한 주제로 3~5분 정도 자연스럽게 이야기를 하게 하고, 이 과정에서 화자의 목소리 크기가 충분한지, 화자의 움직임에 의해 잡음이 발생하지 않는지, 수집 장비에 문제가 없는지를 살펴보는 사전 시험 녹음을 진행하였다.

이 단계에서 녹음된 데이터가 목표한 기준을 충족하지 못할 경우, 수집 장비와 화자의 입 거리를 조정하여 충분한 음량이 유지되도록 하였다. 또한, 대화를 진행하는 과정에서 수집될 수 있는 불필요한 잡음에 대한 주의를 다시 한번 전달하여 실제 녹음 과정에서 해당 문제가 발생하지 않도록 하였다.

실제 수집 과정에서 동일한 설정으로 준비된 수집 장비라 하더라도 예기치 않은 형태의 문제로 인하여 수집 데이터에 잡음이 포함되는 경우가 있어, 이러한 사전 시험 녹음은 화자의 음량 및 발화 태도를 확인하는 것 이외에 장비를 테스트하는 목적도 있다. 특히 통제 녹음의 경우, 코로나 방역을 위해 녹음이 진행된 이후 장비들을 모두 소독하는 과정이 있었고, 헤드셋 마이크의 경우 매번 윈드 스크린⁴⁾을 교체해야 했으므로 해당 과정은 반드시 진행되어야 하는 단계이다.

4) 바람 등에 의해 발생하는 잡음을 차단하기 위한 막

5.2.4. 음성 녹음 진행

사전 시험 녹음을 통해 화자 및 장비에 문제가 없는 것이 확인되면 진행 요원은 음성 녹음을 진행한다. 주제당 12분에서 18분 사이로 대화를 진행하고, 한 화자당 최대 4개의 대화에 참여할 수 있도록 하여 한 화자의 전체 녹음 시간이 최대 60분이 넘지 않도록 하였다.

녹음을 진행하는 동안 진행 요원은 화자들이 선택한 대화 주제가 지속되는지를 살피며 화자들의 대화가 주제에서 벗어나거나 주의 사항에 위반되는 행위가 발견되면 먼저 수신 호로 화자들에게 주의를 주었다. 그럼에도 불구하고 녹음 지속이 어려울 경우 녹음을 일시 중단한 후 주의 사항을 다시 설명하고 진행하였다. 이때 화자들이 선택한 주제에 대한 대화 소재가 부족하여 대화를 계속 이어나가는 것이 어렵다고 판단될 때는 다른 주제로 변경하여 새롭게 대화를 진행하도록 하였다.

비통제 대화 녹음 시 진행 요원은 발화자들의 대화에 거의 간섭하지 않도록 교육하였다. 따라서 비통제 대화 녹음 시 발화자들은 통제 대화 때와는 다르게 자유롭게 개방적인 분위기에서 편안한 대화를 이어나갔다.



[그림 17] 통제 및 비통제 녹음 진행

야외 녹음의 경우 녹음 진행이 불가능한 상황이 종종 발생하였는데 주로 기상 상태가 악화되거나, 주변 환경이 녹음에 적합하지 않을 경우(시위, 행사, 공사 등)가 있었다. 이런 경우 다른 수집 장소로 이동하거나 녹음 일정을 재조정하는 등의 조치를 취했다. 이 밖에도 지나다니는 행인에 의해 녹음에 방해가 되어 재녹음하는 일이 없도록 진행 요원이 주변을 살피며 녹음을 진행하였다.



[그림 18] 각 카테고리별 비통제 녹음 진행 사진

5.2.5. 검증 및 업로드

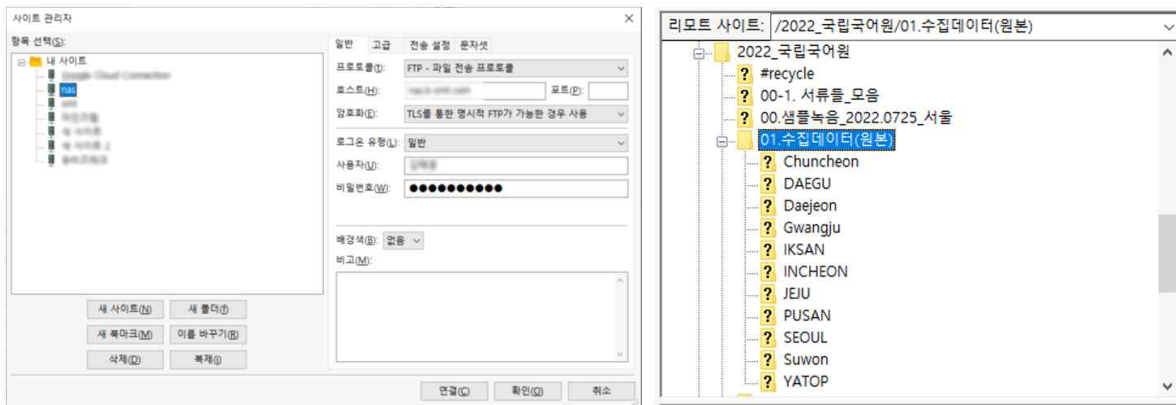
통제 대화 음성 녹음이 완료되면 각 지역의 담당자들은 실제 녹음된 파일을 청취하여 파일 자체에 문제는 없는지, 모니터링 과정에서 발견하지 못한 잡음은 없는지 등을 확인한 후 최종적으로 문제가 없으면 참여자들에게 참여 비용을 지급하고 귀가시켰다. 만약 이때 수집 데이터에서 문제(돌발적인 외부 잡음)가 발생한 경우는 화자들의 동의를 구한 후 바로 재녹음을 진행하거나 다른 날짜로 일정을 잡아 다시 녹음하였다.

비통제 대화의 경우는 음성 수집 장소에서 녹음된 파일을 청취하는 등 검증 작업을 수행할 수 없으므로 녹음이 끝난 후 사무실로 복귀한 다음 검증 작업을 진행하였다. 이때 수집 데이터의 품질에 문제가 발생한 경우 내용을 보고하고 데이터를 폐기하고 발화자들과 재녹음 일정을 협의하였다.

문제없이 수집이 완료된 원본 파일은 각 지역별 수집 지역의 진행 요원이 WAV 파일로 변환한 후 원본 파일과 WAV 파일을 지정된 경로에 등록하고, 상위 관리자에게 진행 내용 및 특이 사항을 보고하였다.



[그림 19] 수집 데이터 검증



<관리자 공유 시스템 로그인>

<지정된 경로에 음성 파일 등록>

[그림 20] 공유 시스템 로그인 및 파일 등록(예시)

6. 음성 자료 전사

6.1. 전사 규칙

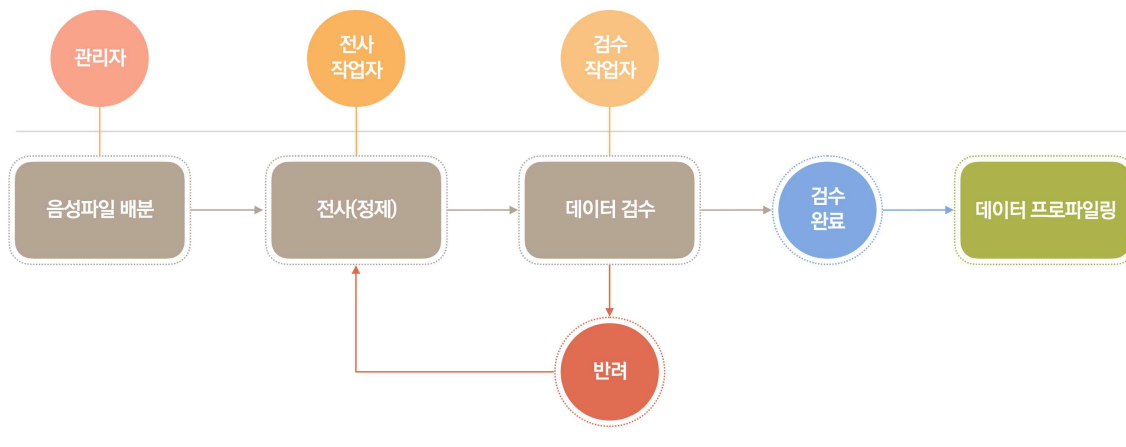
규칙은 ‘2022년 일상 대화 말뭉치 구축 전사 지침’을 적용하였다. 해당 지침은 “발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙”으로 한다. 주요한 전사 원칙 및 전사 단위를 간단히 정리하면 아래와 같다.

- 발음 전사는 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 적용하여 발음 나는 대로 적는다.
- 철자 전사는 한글 맞춤법 및 표준어 규정에 따라 적는 것으로, 발화 내용은 기본적으로 한글 맞춤법 및 표준어 규정에 따라 전사하며 띄어쓰기도 한글 맞춤법에 따른다.
- 전사 단위는 긴 휴지, 경계 억양, 경계 말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 하며, 하나의 전사 단위가 3초 이상으로 길어지는 것을 지양한다.
- 느낌표나 쉼표는 사용하지 않는다. 문장이 완전히 종결되었을 때는 마침표를 사용한다.
- 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분해 준다.
- 긴 쉼에 의해 나뉘는 경우는 통사적으로 완성이 되지 않았다 하더라도 구분하여 전사한다.
- 문장이 종결되었으나 휴지(무음)가 짧아 분할할 수 없는 경우에는 문장이 종결되는 부분에 마침표를 붙인다.

- 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. 만약 맞장구 발화가 일어나는 경우 맞장구 발화를 겹침이 발생한 주 발화 사이에 넣어 주 발화를 나눈다.
- 발음 전사 시 모음의 변화, 수의적 경음화 등을 반영하여 적는다.
- 발음 전사 시 약화 현상에 의한 이형태는 반영하지 않는다. 예를 들어 의문사 ‘뭐’가 ‘머’로 모음이 약화되어 들려도 별도의 발음 전사를 하지 않고 철자 전사인 ‘뭐’만 적는다.
- 발음 전사 시 숫자, 기호, 영문 등도 발음에 따라 한글로 적는다.
- 끊어진 단어는 발화된 대로 전사하되, 앞뒤에 ‘-’을 넣어 표시한다. 불완전하게 발화된 단어(어절)가 둘 이상인 경우에도 어절마다 다음과 같이 표시하여 전사한다.
- 구어에서는 발음의 축약 현상이 많이 나타나는데, 두 음절이 한 음절 사잇소리가 된 다거나, 두 음절이 한 음절 겹핥소리가 되는 것 등이다. 일상 대화 말뭉치에서는 발음되는 음절 수와 표기상의 음절 수를 맞추는 것이 원칙이므로 축약형의 경우 모두 표기에 반영한다.
- 모든 숫자는 이중전사를 하되, 단위를 붙일 수 없는 숫자가 단독으로 나오거나 조사가 결합되는 경우에는 이중전사를 하지 않는다.
- 이름, 이메일 주소 등 계정 정보, 주민등록번호, 카드 번호, 전화번호 등 각종 번호 및 비밀번호, 상세 주소, 출신 및 소속 등의 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다.

6.2. 전사 작업

전사 절차는 전사 도구 기획 및 개발, 전사 인력 모집 및 지침 교육, 전사 진행의 단계로 이루어졌다. 우선 음성을 듣고 전사할 수 있도록 전사 도구를 기획하고 개발하였다. 음성 재생 및 정지, 배속 설정, 음성 전사, 비식별화 등의 기능을 사용할 수 있으며, 여러 명의 작업자들이 동시에 전사 작업을 진행하는 것에 문제가 없도록 개발하였다. 개발이 완료된 전사 도구를 활용할 수 있도록 전사 인력에게 전사 도구 및 전사 지침에 대한 교육을 진행하였다. 교육이 완료된 인력들은 전사 도구를 활용하여 음성 전사를 진행하였다. 전사 도구가 탑재된 플랫폼을 이용한 전사 절차는 아래와 같다.

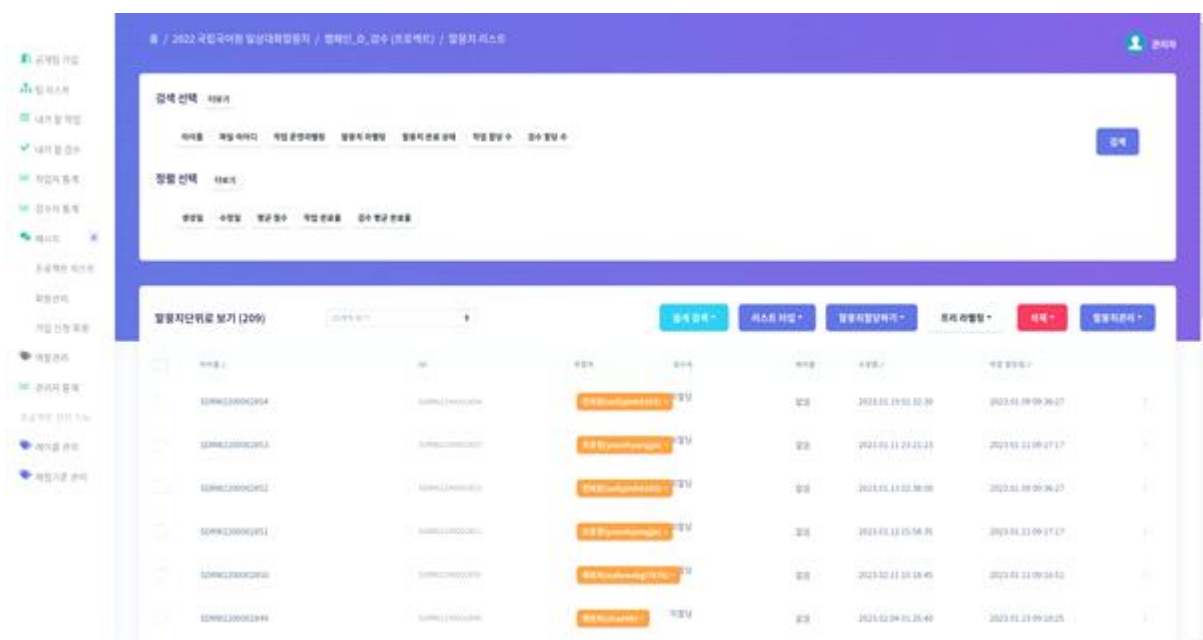


[그림 21] 전사 도구를 이용한 전사 절차

- 작업자는 전사 도구를 통해 작업 대상 음성 파일을 할당받아 작업을 시작한다. 전사 지침에 맞춰 전사(정제) 완료 후 작업 완료 상태로 만든다.
- 전사(정제)가 완료된 작업 완료 파일은 검수단계로 넘어간다. 검수자는 내용 확인 후 이상이 없을 경우 검수 완료 처리 후 데이터 프로파일링 단계로 데이터를 넘긴다. 만약 오류 사항이 발견되면 해당 파일은 작업 반려 처리로 다시 작업자에게 되돌려 보낸다.
- 반려 파일을 받은 작업자는 검수자가 작성한 반려 사유를 확인 후 재작업하여 검수 단계로 넘긴다.
- 검수자는 재작업한 파일을 검수 후 문제가 없다면 최종적으로 검수를 완료 처리하고 데이터 프로파일링 단계로 데이터를 넘긴다.



[그림 22] 전사 도구에서 전사 캠페인 보기



[그림 23] 전사 도구에서 전사 대상 대화 목록 보기

전사 인력 중 20년 이상 경력의 교정 교열 전문가에 의한 전사는 통제 대화 15분 발화 음성 기준 평균 2시간 정도(전사 1시간 30분, 자체 검토 30분)로 가장 짧은 시간이 소요되었으며 전사의 정확도 역시 가장 높았다. 언어 재활사, 언어 계열 전공자와 데이터 베이스 구축사업 유경험자의 순으로 전사의 정확도가 높았으며 15분 발화 음성 기준 평

균 2시간 30분~3시간 30분 정도의 시간이 소요되었다.

비통제 대화의 경우 녹음 장소와 환경에 따른 전사 소요 시간의 편차가 큰 편이어서 정확한 통계를 내기는 어려웠지만, 대체로 통제 대화 대비 1.5~2배 정도의 전사 시간이 소요되었으나 통제 대화 대비 정확도의 편차는 크지 않았다.



[그림 24] 전사 도구에서 전사 수정, 청취 및 결과 보기

발화의 전사는 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 이중 전사를 원칙으로 하였다. 주관기관에서 제시한 전사 지침을 준수하고 국립국어원 우리말샘, 국립국어원 한국어 어문규범 중 한국어 맞춤법, 표준어 규정 및 외래어 표기법 등을 참고하였다.

대상 지역 거주 경험이 있는 사람을 전사 인력 중심으로 배치했음에도 불구하고 방언 자료는 표준어에 비해 20~30%가량 많은 시간이 소요되었다. 또한, 발화자와 나이대가 비슷한 전사자를 배치했을 때 그렇지 않은 경우보다 전사 작업 시간과 정확도에서 20% 내외로 높은 효율을 보여 주었다.

6.3. 품질 검수

6.3.1 전사 데이터 프로파일링

전사가 완료된 파일은 데이터 프로파일링 단계를 거치고, 검수 담당자를 지정하여 전수 수작업 검수를 진행하였다. 검수 작업자가 전사 완료된 파일을 검수하기 전에 데이터 프로파일링을 진행하여 보다 효율적으로 검수를 진행할 수 있도록 하였다. 데이터 프로파일링이란, 전사 결과의 기계적 분석을 통하여 오류 가능성이 있는 후보를 탐지하는 과정으로, 이를 통해 메타데이터 및 철자 오류 발생 가능 위치와 오류 가능성이 있는 내용을 탐지할 수 있다. 데이터 프로파일링 진행 절차는 아래와 같다.



[그림 25] 데이터 프로파일링 절차

[표 25] 데이터 프로파일링 세부 공정 및 오류 예시

세부 공정		오류 예시			
형식 오류 탐지	id	original_from	form	labeler_name	
	SDRW2200000554.1.1.157	(매 돼 돼 돼) 음	매 돼 돼 돼 음	김서영	
	SDRW2200000426.1.1.173	처음에 (한 시)1시에 이제 학과 사무실	처음에 한 시1시에 이제 학과 사무실	김종철	
	SDRW22000002139.1.1.109	버스 요금이 (두 배)(2배)가 되는 거잖아요	버스 요금이 두 배2배가 되는 거잖아요	유진선	
	SDRW2200000378.1.1.143	결국에는 씨제이로부터	결국에는 CJ로부터	이근호	
	SDRW22000002118.1.1.1278	사실?	사실?	송진기	
	SDRW2200001957.1.1.1	&name& 1 아빤는 옛날에 그 은행 들어갈 때	name 1 아빤는 옛날에 그 은행 들어갈 때	조윤형	
	SDRW2200001957.1.1.1225	인제 &name& 2 이는 지금 지가 느끼고 있잖아. 왜냐	인제 name 2 이는 지금 지가 느끼고 있잖아. 왜냐	조윤형	
	SDRW2200001957.1.1.1229	&name& 2 는 보호 하에 지금 있어야 되는 상황인데	name 2 는 보호 하에 지금 있어야 되는 상황인데	조윤형	
	SDRW2200002024.1.1.159	요즘 보면 좀 옛날 드라마 ((말씀xx) 전원일기라든가	요즘 보면 좀 옛날 드라마 말씀xx 전원일기라든가	채현희	
동일 발음 전사에 나타나는 여러 철자 전사 유형 탐지	SDRW2200002024.1.1.1259	이승윤 열니이 예) 원 이었고	이승윤 열니이 예 1 이었고	채현희	
	SDRW2200001740.1.1.1231	그리고 그거 티오가 있어야지 선택을 할 수가 있는 거니까	그리고 그거 TO가 있어야지 선택을 할 수가 있는 거니까	전지영	
	SDRW2200001730.1.1.1238	어~ 어떨 땐 어스런지 않을 수 있지만=	어 어떨 땐 어스런지 않을 수 있지만=	전소연	
	그니깡	SDRW22000002745.1.1.262	그러니깡	^ 그러니깡 사람들이	박노연
	그니깡	SDRW22000002497.1.1.162	그러니깡	^ 그러니깡 상식	이은실
	그다	SDRW22000002802.1.1.261	거기다	^ 거기다 요리도	채희옥
	그다	SDRW22000001796.1.1.161	그다음	^ 그다음 금	송순화
	그다	SDRW22000002801.1.1.32	그보다	^ 그보다 요즘은	채현희
	그도	SDRW22000001796.1.1.239	것도	부친 것도 있고	송순화
	그도	SDRW22000001811.1.1.213	그래도	거지만 그래도 어느	황금희
동일 철자 전사에 나타나는 여러 발음 전사 유형 탐지	그도	SDRW22000001908.1.1.171	그래도	^ 그래도 \$	송순화
	그도	SDRW22000002083.1.1.442	그래도	^ 그래도 어	황인순
	그도	SDRW22000002338.1.1.396	그래도	^ 그래도 과감하게	서진아
	그도	SDRW22000002460.1.1.231	그래도	^ 그래도 운전은	유재현
	그도	SDRW22000002615.1.1.382	그래도	^ 그래도 강아지	최미리
	그도	SDRW22000002617.1.1.204	그래도	이제 그래도 보드보단	최미리
	그라는데	SDRW22000001490.1.1.211	그러는데	박치 그러는데 내가	채희옥
	그라는데	SDRW22000001490.1.1.123	그런데	^ 그런데 음악을	채희옥
	10억	SDRW22000002339.1.1.174	십 억	해서 십_억 모았다	전희원
	10억	SDRW22000001646.1.1.85	십억	그 십억 \$	박선희
비식별화 대상 억양구 탐지	10억	SDRW22000002065.1.1.132	십억	뭐 십억 이렇게	윤순섭
	10억	SDRW22000002156.1.1.222	십억	^ 십억 대	김혜진
	10억	SDRW22000002196.1.1.44	십억	몇 십억 단위다	남새롬
	10억	SDRW22000002339.1.1.198	십억	은행에 십억 이상	전희원
	10억으로	SDRW22000001890.1.1.254	십_억으로	대에 십_억으로 거래한	김종철
	10억으로	SDRW22000002221.1.1.246	십억으로	뭐 십억으로 지방에	전희원
	10월	SDRW22000002672.1.1.135	시 월	또 시_월 말쯤에	류시연
	10월	SDRW22000001912.1.1.92	시월	게 시월 중순일	장청화
	10월	SDRW22000002439.1.1.19	시월	그치, 시월 말에	어윤정
	10월	SDRW22000002439.1.1.20	시월	뭐 시월 중순에	어윤정
비식별화 대상 억양구 탐지	10월	SDRW22000002549.1.1.88	시월	다 시월 이전이야,	김지은
	10월	SDRW22000002803.1.1.110	십 월	^ 십_월 달에	조윤형
	1	id	form	original_from	
	2				
	3	SDRW22000000001.1.1.4	너랑 아 @이름1랑 나는 아직	너랑 아 @이름1랑 나는 아직	
	4	SDRW22000000001.1.1.11	@이름2는 어떻게 되니?	@이름2는 어떻게 되니?	
	5	SDRW22000000001.1.1.18	음 알겠어, @이름3는 뭘 전공했어?	음 알겠어, @이름3는 뭘 전공했어?	
	6	SDRW22000000001.1.1.25	@이름3는 정말 엘리트구나,	@이름3는 정말 엘리트구나,	
	7	SDRW22000000001.1.1.30	학박한친 알았어, @이름1는 어대?	학박한친 알았어, @이름1는 어대?	
	8	SDRW22000000001.1.1.50	어려운 과목이더라고. 그렇지 않니 @이름3야?	어려운 과목이더라고. 그렇지 않니 @이름3야?	
	9	SDRW22000000001.1.1.54	@이름2는 법 관련된 과목을 공부해 본 적이 있니?	@이름2는 법 관련된 과목을 공부해 본 적이 있니?	
	10	SDRW22000000001.1.1.63	그래서 @이름3가 정말 존경스럽다.	그래서 @이름3가 정말 존경스럽다.	
	11	SDRW22000000001.1.1.84	나랑 @이름2가 같은 고등학교 나온 거 알고 있니?	나랑 @이름2가 같은 고등학교 나온 거 알고 있니?	
	12	SDRW22000000001.1.1.97	@이름1 너는 복수전공을 하고 싶은 과거 있었어?	@이름1 너는 복수전공을 하고 싶은 과거 있었어?	
	13	SDRW22000000001.1.1.103	@이름2 너는 어떻게 생각해?	@이름2 너는 어떻게 생각해?	
	14	SDRW22000000001.1.1.108	혹시 @이름3는 공대에 대해서 어떻게 생각해니?	혹시 @이름3는 공대에 대해서 어떻게 생각해니?	
	15	SDRW22000000001.1.1.114	@이름3는 손 재주가 아주 좋나 보네?	@이름3는 손 재주가 아주 좋나 보네?	
	16	SDRW22000000001.1.1.147	그렇구나 @이름1는 혹시	그렇구나 @이름1는 혹시	
	17	SDRW22000000001.1.1.153	@이름2는 따로 어떤 실습을 해 봤니?	@이름2는 따로 어떤 실습을 해 봤니?	
	18	SDRW22000000001.1.1.172	-그러, 그러면 @이름2는 미래에 승무원이나 조종사가 되고 싶겠네?	-그러, 그러면 @이름2는 미래에 승무원이나 조종사가 되고 싶겠네?	
	19	SDRW22000000001.1.1.175	혹시 @이름3는 항공 쪽에 관심 있니?	혹시 @이름3는 항공 쪽에 관심 있니?	
	20	SDRW22000000001.1.1.181	혹시 @이름1 학교에서 비행기 본 적 있니?	혹시 @이름1 학교에서 비행기 본 적 있니?	
	21	SDRW22000000001.1.1.193	정말 좋은 생각이야, @이름3도 같이 참여하지 않을까?	정말 좋은 생각이야, @이름3도 같이 참여하지 않을까?	

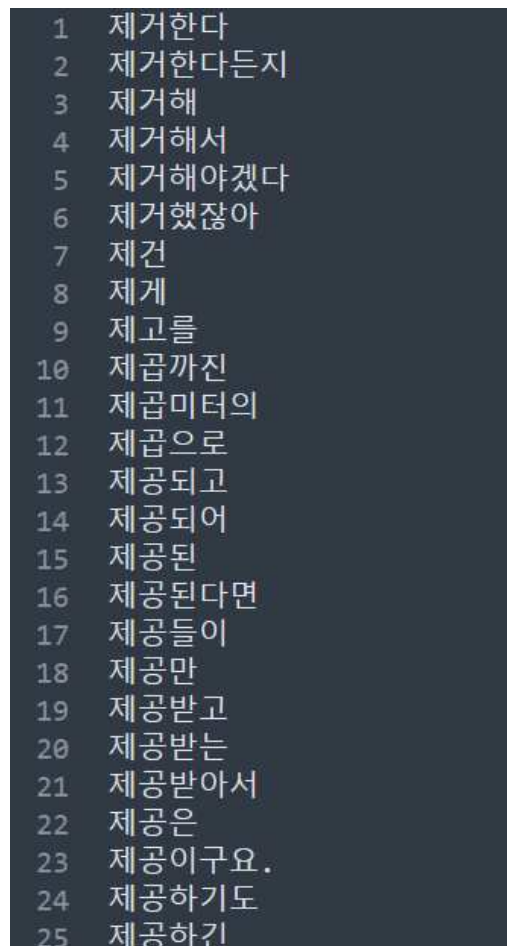
위 예와 같이 언어 분석 방법을 통해 동일 발음 전사에 대해서 서로 다른 철자 전사가 이루어진 경우, 또는 동일 철자 전사에 대해서 서로 다른 발음 전사가 나타난 예 등의 철자 및 형식 오류를 탐지하여 수정하였다.

데이터 프로파일링을 통해 기계적 검증을 마친 전사 데이터는 품질 검수 담당자들이 전수 검수를 진행하였다. 품질 검수 담당자는 전사 내용을 확인하고 지침과 상이한 경우 직접 전사 내용을 올바르게 수정하였다.

6.3.2 최종 데이터 검증

최종 데이터로 선별된 전체 데이터에 대해 최종 데이터 프로파일링을 진행하였다. 전사 데이터 전수 검수 과정에서 발견되지 못한 오류를 기계적으로 탐지하여 철자 및 형식 오류를 수정하였다.

최종 데이터 프로파일링 및 오류 수정 이후 전사 데이터의 품질 제고를 위해 전체 원시 말뭉치(JSON)의 철자 전사 내용을 추출하여 어절 단위로 분리한 후, 전체 어절에서 출현 빈도수가 3 이하인 어절들을 뽑아냈다. 이는 전체 철자 전사에서 많이 나타나지 않은 만큼 오타자 또는 철자 전사 기준에 맞지 않는 전사가 포함되어 있을 것으로 추정되는 어절들을 탐지한 것으로, 약 20만 개의 어절을 목록화하였다.



1	제거한다
2	제거한다든지
3	제거해
4	제거해서
5	제거해야겠다
6	제거했잖아
7	제건
8	제게
9	제고를
10	제공까진
11	제공미터의
12	제공으로
13	제공되고
14	제공되어
15	제공된
16	제공된다면
17	제공들이
18	제공만
19	제공받고
20	제공받는
21	제공받아서
22	제공은
23	제공이구요.
24	제공하기도
25	제공하긴

[그림 26] 전체 철자 전사 중 출현 빈도수
3 이하 어절 목록 예시

이중 올바르게 전사된 정상 어절과 오타자 등의 오류가 포함된 어절을 구분하기 위해 4인의 인력이 약 5일간 전수 검토하여 오류 어절 후보를 분류하였으며, 약 3,700여 개의 오류 의심 어절을 선별하였다. 선별 기준은 아래와 같다.

[표 26] 오류 어절 후보 선별 기준

번호	기준
1	철자 전사 지침에 맞지 않는 전사
2	외래어의 철자 전사 오류 및 외래어 규범 표기에 어긋난 전사
3	문법적으로 맞지 않는 전사 및 띄어쓰기 오류
4	오타자 추정 어절



1	흥은
2	흥을
3	흥하는
4	휘거네.
5	휘지
6	휘튼
7	흰썬
8	힐시
9	힐씨
10	휘트니스
11	휴뭉
12	휴양이가가냐
13	훅
14	훌륭한
15	흐나이는
16	흐트리스
17	흔돈이
18	흘겨들은
19	흘씨
20	흙벅
21	흙연룰은
22	희마와리라는
23	희밥
24	희안하게
25	희안한

[그림 27] 오류 어절 후보 선별 예시

오류 어절 후보 선별 시, 전사 데이터만 검토하여 선별하였으므로 해당 어절이 실제로

오류가 포함된 어절이 맞는지 확인 과정이 필요했다. 따라서 아래와 같이 선별한 오류 후보 어절에 대해 해당 어절이 포함된 파일명과 억양구 번호를 달아 목록화하였다. 이를 전사 데이터 품질 관리 담당자가 직접 해당 억양구 음성을 청취한 후 오류가 확실한 경우 전사 규칙에 맞게 수정을 진행하였다.

```

1 SDRW2200000099 ['159#월씬', '83#팬선이']
2 SDRW2200000100 ['37#일회용품을']
3 SDRW2200000101 ['263#매뉴로']
4 SDRW2200000103 ['235#이러나지']
5 SDRW2200000105 ['116#일틀', '44#쓰르', '48#주심']
6 SDRW2200000106 ['107#이렇거', '118#있었나', '232#으물', '252#없색인']
7 SDRW2200000108 ['144#매뉴들', '34#종을래나?', '88#종을']
8 SDRW2200000109 ['15#해야하는', '227#인테리어를', '300#물마드시', '325#작두공을']
9 SDRW2200000110 ['173#질리더', '186#완변', '239#싫어는데', '270#프로그램가', '57#코스토크에서']
10 SDRW2200000111 ['160#에플파이가', '25#매뉴는', '257#음질', '82#일스러워']
11 SDRW2200000112 ['166#마참가지고', '166#마참가지고']
12 SDRW2200000113 ['86#영양조']
13 SDRW2200000116 ['104#있었을', '136#콘트롤하지', '58#콘트롤', '59#콘트롤', '81#호신술를', '87#했었던거']
14 SDRW2200000117 ['186#서으']
15 SDRW2200000119 ['131#그릉']
16 SDRW2200000148 ['149#앨프', '232#제뭉', '284#웹툰으로']
17 SDRW2200000149 ['369#어쩐']
18 SDRW2200000150 ['390#운동거리지']
19 SDRW2200000156 ['32#조곤']
20 SDRW2200000157 ['169#에플이랑']

```

[그림 28] 오류 어절 후보 목록 예시

약 3,700여 개의 오류 의심 어절을 전수 검토한 결과, 약 2,600여 개(약 70%)의 어절이 실제로 오타자, 문법적으로 맞지 않는 전사 등의 오류 어절로 확인되어 전사 데이터 품질 관리 담당자가 직접 수정하였다.

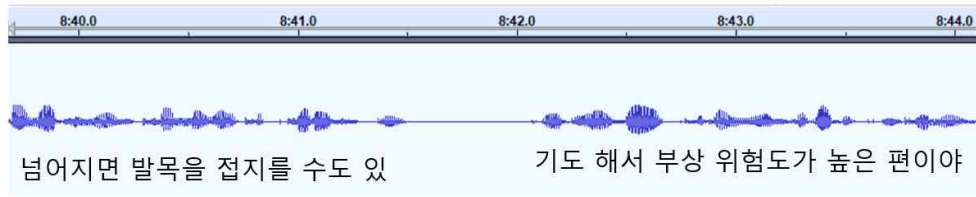
7. 음성 정제

7.1. 억양구 분할

억양구 분할은 음성을 전사 단위에 따라 분할하는 작업이다. 우선 음성 인식 엔진을 사용하여 초벌 전사를 진행하였다. 음성 인식 엔진을 활용한 초벌 전사가 끝나면 음성의 휴지 구간에 맞춰 전사 단위를 자를 수 있는 프로그램을 활용하여 자동 정제를 진행하였다. 1차로 자동 정제가 끝난 음성이 도구에 업로드되면 전사 전문 인력들이 수작업 전사를 진행하면서 1차 자동 정제가 적절하게 이루어지지 않은 문장에 대해 억양구 단위를 조정하였다.

최종 산출물 중에 아래와 같이 억양구 분할된 부분이 발생할 수 있는데, 이는 억양구 분할 오류가 아니라 발화자가 “넘어지면 발목을 접지를 수도 있” 발화 후에 일정 시간 이상의 휴지 기간을 가지고 다음 발화를 이어간 경우이다. 아래는 약 0.5초 이상의 휴지를 가지고 다음 발화를 진행한 예이다.

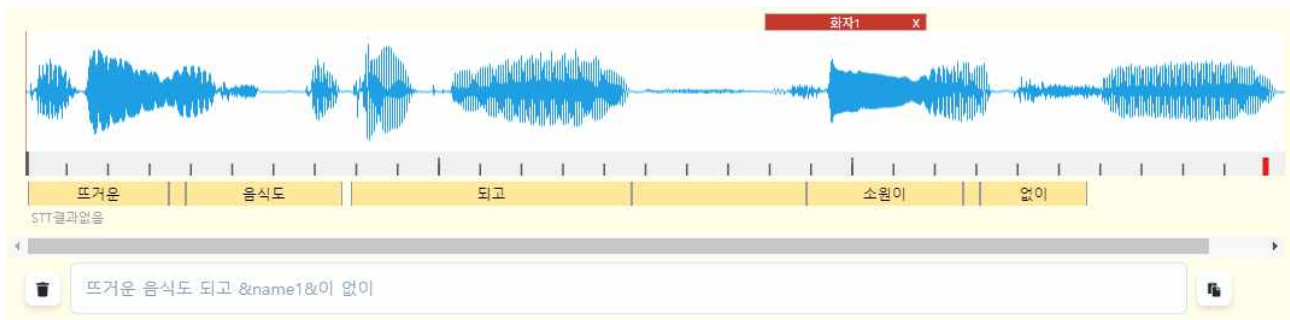
```
{
  "id": "SDRW2200000179.1.1.167",
  "form": "넘어지면 발목을 접지를 수도 있",
  "original_form": "넘어지면 발목을 접지를 수도 있",
  "speaker_id": "SD2200190",
  "start": "519.60000",
  "end": "521.59802",
  "note": ""
},
{
  "id": "SDRW2200000179.1.1.168",
  "form": "기도 해서 부상 위험도가 높은 편이야.",
  "original_form": "기도 해서 부상 위험도가 높은 편이야.",
  "speaker_id": "SD2200190",
  "start": "522.06000",
  "end": "524.19000",
  "note": ""
},
```



[그림 29] 발화 도중 휴지로 인한 억양구 분할 예시

7.2. 음성 개인정보 비식별화

말뭉치 자료 중 이름, 이메일 주소 등 계정 정보나 주민등록번호, 카드 번호 등 각종 번호 및 비밀번호와 상세 주소, 출신 소속 등 개인정보와 관련된 모든 사항들은 노출되지 않도록 전사 작업 단계에서 비식별화를 진행하였다. 단, 정치인 및 유명인의 이름은 비식별화 대상에 해당하지 않으며, 상호명 및 상품명 등은 대화 맥락상 부정적으로 언급된 경우에 한해 비식별화하였다. 개인정보에 해당하는 음성은 전사 도구의 음성 비식별화 기능을 활용하여 해당 구간 마킹 후 최종 산출물 생성 시 음성이 들리지 않도록 묵음 처리를 하는 방식으로 비식별화 처리를 진행하였다. [그림 26]의 예시에서 음성 파형 상단에 ‘화자1’로 표시된 구간은 산출물 PCM에서 묵음으로 변환되며, 전사 작업 및 검수 작업 과정에서 해당 부분에 비식별화 대상인지 여부와 적절하게 비식별화가 진행되었는지 여부를 확인하기 위해 전사 도구 내에서는 음성 파형이 표시된다.



[그림 30] 개인정보 비식별화(예시)

8. 원시 말뭉치 구축 및 메타 정보 구축

8.1. JSON 변환

전사가 완료된 말뭉치를 이용하여 JSON으로 변환하였다. JSON 포맷의 규격은 사전에 협의된 국립국어원 양식을 사용하였으며, JSON 변환 후 포맷 검증 도구를 이용하여 변환 과정에서 오류가 없는지 확인하였다. ‘일상 대화 말뭉치 구축 지침’에 따라 부여한 파일명 부여 방식은 아래와 같다.

[표 27] 대화 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	8자리 일련번호
S: 구어 말뭉치	D: 사적 대화	RW: 원시 말뭉치	22	#####

JSON 파일의 내부 구조도 “일상 대화 말뭉치 구축 지침”의 가이드를 준수하여 구성되어 있으며, 상세한 JSON 구조는 [붙임 1] “일상 대화 말뭉치 구축 지침”에 상세히 정의되어 있다. 참고로 최종 산출물 말뭉치 변환 예시 일부는 아래와 같다. 말뭉치 파일의 확장자는 json, 문자 인코딩은 유니코드(UTF-8), 줄바꿈 문자로 LF(UNIX)를 사용하였다.

[표 28] 말뭉치 변환 예시(일부)

```
{
  "id": "SDRW2200000613",
  "metadata": {
    "title": "국립국어원 구어 말뭉치 SDRW2200000613",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2022",
    "category": "구어 > 사적대화 > 일상대화",
    "annotation_level": [
      "원시"
    ],
    "sampling": "본문 전체"
  },
  "document": [
    {
      "id": "SDRW2200000613.1",
      "metadata": {
        "title": "2인 일상 대화",
```

```

"author": "개인 발화자",
"publisher": "개인 발화 녹음",
"date": "20220827",
"topic": "휴가 > 휴가 가기 좋은 여행지 공유",
"environment": "",
"speaker": [
  {
    "id": "SD2200128",
    "age": "20대",
    "occupation": "사무 종사자",
    "sex": "여성",
    "birthplace": "경기",
    "principal_residence": "경기",
    "current_residence": "경기",
    "education": "대졸"
  },
  {
    "id": "SD2200129",
    "age": "50대",
    "occupation": "주부",
    "sex": "여성",
    "birthplace": "경북",
    "principal_residence": "경북",
    "current_residence": "경기",
    "education": "고졸"
  }
],
"setting": {
  "relation": "부모/자녀",
  "device": "",
  "mic": ""
},
"utterance": [
  {
    "id": "SDRW2200000613.1.1.1",
    "form": "엄마 올해 여름휴가는 어디로 다녀왔어?",
    "original_form": "엄마 올해 여름휴가는 어디로 다녀왔어?",
    "speaker_id": "SD2200128",
    "start": "0.42000",
    "end": "4.69100",
    "note": ""
  },
  {
    "id": "SDRW2200000613.1.1.2",
    "form": "응 아빠랑 제주도 갈까 하다가",
    "original_form": "응 아빠랑 제주도 갈까 하다가",
    "speaker_id": "SD2200129",
    "start": "5.18000",
    "end": "8.57900",
    "note": ""
  }
],

```

8.2. 메타 정보 구축

수집 결과물인 음성 데이터와 텍스트 데이터의 원활한 활용을 위해 해당 데이터의 정보를 포함한 메타 정보의 구축이 필수적이다. 이러한 메타 정보의 구축은 발주기관인 국립국어원에서 제시한 양식을 활용하여 아래와 같이 진행하였다.

번호	id	metadata		document		setting	relation	시간(초)	저작권 확보여부	저작권 접촉이력
		category	title	topic1	topic2					
예시	SDRW2200000000	국어 > 사적대화 > 일상대화	2인 일상 대화	여행지(국내/해외)	해외여행 숙소, 여행 스타일	친구	0:15:06	0	0	스마트미디어테크에서 일괄 확보
1	SDRW2200000001	국어 > 사적대화 > 일상대화	3인 일상 대화	회사/학교	학교 생활 및 전공 이야기	친구	0:15:05	0	0	스마트미디어테크에서 일괄 확보
2	SDRW2200000002	국어 > 사적대화 > 일상대화	3인 일상 대화	건강/다이어트	다졌던 경험과 수술 받은 이야기, 운동하는 경험과 장단점, 약물에 관한 이야기	친구	0:15:07	0	0	스마트미디어테크에서 일괄 확보
3	SDRW2200000003	국어 > 사적대화 > 협력적대화	3인 협력적 대화	여행계획	해외여행 갈 장소 선정	친구	0:15:06	0	0	스마트미디어테크에서 일괄 확보
4	SDRW2200000004	국어 > 사적대화 > 협력적대화	3인 협력적 대화	음식/음료	저녁 식사 메뉴 선정	친구	0:15:10	0	0	스마트미디어테크에서 일괄 확보
5	SDRW2200000005	국어 > 사적대화 > 일상대화	3인 일상 대화	회사/학교	다이어트 관련 제품 이야기 및 경험 공유, 운동 관련 이야기	친구	0:15:12	0	0	스마트미디어테크에서 일괄 확보
6	SDRW2200000006	국어 > 사적대화 > 일상대화	3인 일상 대화	대중교통	가뭇던 여행지 경험 공유	친구	0:15:13	0	0	스마트미디어테크에서 일괄 확보
7	SDRW2200000007	국어 > 사적대화 > 협력적대화	3인 협력적 대화	여행계획	동아리 여행에 필요한 정보 토의	친구	0:15:12	0	0	스마트미디어테크에서 일괄 확보
8	SDRW2200000008	국어 > 사적대화 > 협력적대화	3인 협력적 대화	반려동물	저녁 식사 메뉴 선정	친구	0:15:08	0	0	스마트미디어테크에서 일괄 확보
9	SDRW2200000009	국어 > 사적대화 > 일상대화	3인 일상 대화	휴가	아이들과 같이 가기 좋은 여행지	친구	0:15:24	0	0	스마트미디어테크에서 일괄 확보
10	SDRW2200000010	국어 > 사적대화 > 일상대화	3인 일상 대화	건강/다이어트	다이어트 정보 및 보조제 정보	친구	0:15:29	0	0	스마트미디어테크에서 일괄 확보
11	SDRW2200000011	국어 > 사적대화 > 협력적대화	3인 협력적 대화	여행일반	선호하는 여행의 종류에 대한 장단점 토론	친구	0:15:06	0	0	스마트미디어테크에서 일괄 확보
12	SDRW2200000012	국어 > 사적대화 > 협력적대화	3인 협력적 대화	음식/음료	배달 음식과 일기트 음식에 대한 의견 공유	친구	0:15:19	0	0	스마트미디어테크에서 일괄 확보
13	SDRW2200000013	국어 > 사적대화 > 일상대화	3인 일상 대화	스포츠/레저/취미	운동하는 야구 팀에 대해서, 각자가 좋아하는 스포츠	부모/자녀	0:15:07	0	0	스마트미디어테크에서 일괄 확보
14	SDRW2200000014	국어 > 사적대화 > 일상대화	3인 일상 대화	먹거리	좋아하는 음식	부모/자녀	0:15:09	0	0	스마트미디어테크에서 일괄 확보
15	SDRW2200000015	국어 > 사적대화 > 협력적대화	3인 협력적 대화	음식/음료	식사 메뉴 선정 논의	부모/자녀	0:15:05	0	0	스마트미디어테크에서 일괄 확보
16	SDRW2200000016	국어 > 사적대화 > 협력적대화	3인 협력적 대화	여행계획	휴가 계획에 대한 논의	부모/자녀	0:15:23	0	0	스마트미디어테크에서 일괄 확보
17	SDRW2200000017	국어 > 사적대화 > 일상대화	3인 일상 대화	건강/다이어트	본인, 지인의 운동 경험 이야기, 다이어트 관련 경험	친구	0:15:12	0	0	스마트미디어테크에서 일괄 확보
18	SDRW2200000018	국어 > 사적대화 > 일상대화	3인 일상 대화	음악	좋아하는 가수와 음악에 관한 이야기, 직접 노래 부른 경험	친구	0:15:16	0	0	스마트미디어테크에서 일괄 확보
19	SDRW2200000019	국어 > 사적대화 > 협력적대화	3인 협력적 대화	여행계획	같이 가는 여행 계획 세우기	친구	0:15:16	0	0	스마트미디어테크에서 일괄 확보
20	SDRW2200000020	국어 > 사적대화 > 협력적대화	3인 협력적 대화	패션/뷰티	공연 무대에서 입을 의상 정하기	친구	0:15:06	0	0	스마트미디어테크에서 일괄 확보
21	SDRW2200000021	국어 > 사적대화 > 일상대화	3인 일상 대화	먹거리	좋아하는 술안주 이야기	친구	0:15:13	0	0	스마트미디어테크에서 일괄 확보
22	SDRW2200000022	국어 > 사적대화 > 일상대화	3인 일상 대화	경제/재테크	직접 하고 있는 재테크와 우리나라 경제 이야기	친구	0:15:13	0	0	스마트미디어테크에서 일괄 확보
23	SDRW2200000023	국어 > 사적대화 > 협력적대화	3인 협력적 대화	여행계획	친구들과 함께 여행 갈 장소 선정 토의	친구	0:15:09	0	0	스마트미디어테크에서 일괄 확보
24	SDRW2200000024	국어 > 사적대화 > 협력적대화	3인 협력적 대화	스포츠/레저/취미	친구에게 스포츠 추천	친구	0:15:08	0	0	스마트미디어테크에서 일괄 확보
25	SDRW2200000025	국어 > 사적대화 > 일상대화	3인 일상 대화	건강/다이어트	혈압 관리	모임 동아리 지인	0:15:12	0	0	스마트미디어테크에서 일괄 확보
26	SDRW2200000026	국어 > 사적대화 > 일상대화	3인 일상 대화	수영, 등산, 축구	수영, 등산, 축구	모임 동아리 지인	0:15:17	0	0	스마트미디어테크에서 일괄 확보
27	SDRW2200000027	국어 > 사적대화 > 협력적대화	3인 협력적 대화	반려동물	반려동물에 대한 의견 공유	모임 동아리 지인	0:15:25	0	0	스마트미디어테크에서 일괄 확보
28	SDRW2200000028	국어 > 사적대화 > 일상대화	3인 일상 대화	먹거리	대용량, 집밥	모임 동아리 지인	0:15:13	0	0	스마트미디어테크에서 일괄 확보
29	SDRW2200000029	국어 > 사적대화 > 일상대화	3인 일상 대화	휴가	휴가 때 가볼 곳 이야기	모임 동아리 지인	0:15:11	0	0	스마트미디어테크에서 일괄 확보

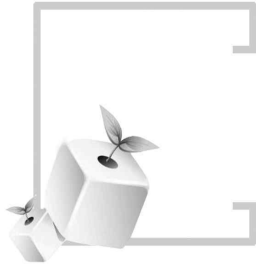
[그림 31] 메타 정보 파일 일부

2022 국립국어원 일상대화 말뭉치 발화자 정보											
대화가 다르더라도 화자가 동일하면 동일한 아이디 부여											
번호	field	id	name	age	occupation	sex	birthplace	principal residence	current residence	education	
예시	SDRW2200000001	SD22000001	홍길동	30대	학생	남성	서울	경기	경기	대재	
	SDRW2200000001	SD22000002	심정	30대	전문가 및 관련 종사자	여성	경기	경기	경기	대졸	
1	SDRW2200000001	SD22000005	최준	20대	학생	남성	서울	서울	서울	대재	
2	SDRW2200000001	SD22000006	모인	20대	무직/취업준비생	여성	서울	경기	경기	대졸	
3	SDRW2200000001	SD22000007	최현	20대	학생	남성	서울	경기	경기	대재	
4	SDRW2200000002	SD22000005	최준	20대	학생	남성	서울	서울	서울	대재	
5	SDRW2200000002	SD22000006	모인	20대	무직/취업준비생	여성	서울	경기	경기	대졸	
6	SDRW2200000002	SD22000007	최현	20대	학생	남성	서울	경기	경기	대재	
7	SDRW2200000003	SD22000005	최준	20대	학생	남성	서울	서울	서울	대재	
8	SDRW2200000003	SD22000006	모인	20대	무직/취업준비생	여성	서울	경기	경기	대졸	
9	SDRW2200000003	SD22000007	최현	20대	학생	남성	서울	경기	경기	대재	
10	SDRW2200000004	SD22000005	최준	20대	학생	남성	서울	서울	서울	대재	
11	SDRW2200000004	SD22000006	모인	20대	무직/취업준비생	여성	서울	경기	경기	대졸	
12	SDRW2200000004	SD22000007	최현	20대	학생	남성	서울	경기	경기	대재	
13	SDRW2200000005	SD22000014	김욱	20대	학생	남성	대전	대전	경기	대재	
14	SDRW2200000005	SD22000015	김현	20대	학생	남성	서울	경기	서울	대재	
15	SDRW2200000005	SD22000016	이현	20대	학생	남성	서울	경기	서울	대재	
16	SDRW2200000006	SD22000014	김욱	20대	학생	남성	대전	대전	경기	대재	
17	SDRW2200000006	SD22000015	김현	20대	학생	남성	서울	경기	서울	대재	

[그림 32] 발화자 메타 정보 일부

메타 정보 구축에는 수집에 참여한 화자의 정보(성별, 나이대, 직업, 출생지, 주 성장지, 현 거주지 등)와 수집에 참여한 화자들 간의 관계를 필수로 작성하였고, 대화 주제는 대주제(topic1)와 소주제(topic2)로 나누어서 기재하였다. 이러한 메타 정보의 구축은 데이터 수집 과정에서도 활용되었는데, 수집 후반기의 경우 결과물의 목표치를 초과하지 않도록 모니터링을 하는 도구로 사용되었다. 다수의 지역에서 수집이 진행되다 보니 성별, 나이별, 지역별, 주제별 목표 수치가 수집 지역에서 실시간으로 결과를 취합해도 진도 상황을 체크하기에는 어려움이 수반된다.

따라서 본 메타 정보 구축을 통하여, 각 지역에서 실시간으로 메타 정보가 업데이트되면서 각 수집 항목별 비율 산출이 가능하였다. 이를 토대로 나이, 지역, 제시 주제가 목표치에 도달하게 되면 해당 항목에 대한 화자의 모집이나 주제 사용을 금지하여, 불필요한 데이터의 수집을 최소화하는 데에 활용되었다. 또 메타 정보는 데이터를 수집 기관에서만 활용하지 않고 전사를 담당하는 곳에서도 활용하였다. 메타 정보는 화자의 성장지(각 지역별 방언 고려), 화자와의 관계, 대화의 주제, 대화의 내용, 성별, 나이 등 전사 작업을 진행하는 데 있어서 필수적으로 도움이 되는 정보이다. 또한, 후처리 과정에서 발견되는 오류를 함께 공유하여 수집 단계에서 발생을 보완하였고, 전사 작업 단계에서 오류를 수정하는 수단으로도 활용하였다.



제3장

사업 수행 결과



1. 주제별 수집 결과

일상 대화 말뭉치의 발화주제는 총 16개의 대주제와 세부 주제를 제시하였으며, 협력적 대화는 문화·관광의 카테고리에서 선정한 총 10개의 주제로 수집하였다.

[표 29] 주제별 대화 수집 결과

유형	대주제	세부 주제 예시	수집 개수	비율
일상 대화	휴가	여행 시 교통/숙박 선택, 자연/휴양지 소개, 여행 국가/지역 선택, 추천 여행지(국내/해외) 등	111	6.28%
	대중교통	약속 만남 시 교통 선택, 새로운 교통수단 공유 키포드, 전철, 버스, 기차, 교통 약자석 등	122	6.90%
	음악	대중 음악, 선호 가수 및 노래 추천, 선호하는 음악 장르 등	103	5.83%
	건강/다이어트	가지고 있는 질병/알레르기, 건강 보조제, 건강을 위해 하고 있는 노력, 약 부작용, 다이어트 성공/실패 경험담 등	116	6.56%
	방송/연예	인생 드라마 추천, 예능 프로그램, 선호하는 배우 등	121	6.84%
	스포츠/레저/취미	경기 직접 관람, 등산, 여름/겨울 레저, E-스포츠(게임), 독서(시집, 문학 등), 영화 관람, 만화(웹툰 등), 웹소설 등	108	6.11%
	먹거리	최근 인기있는 음식과 경험, 가장 선호하는 음식, 추천하는 맛집, 배달 음식, 밀키트 등	113	6.39%
	우정	선호하는 것, 성격, 다툼과 화해 등에 대한 친구 사이의 대화, 국경/나이를 초월한 우정 등	104	5.88%
	경제/재테크	부동산, 주식 투자, 비트코인, 고물가, 세금, 인플레이션, 금리 인상 등	114	6.45%
	회사/학교	회사 생활, 회식, 인턴 생활, 진학에 대한 정보와 결정, 입시 경쟁, 전공, 성적 등	111	6.28%
	반려동물	반려동물 입양, 유기 동물 문제, 동물 병원 고비용, 동물 등록제, 반려동물과의 추억 등	105	5.94%
	취직	취업에 대한 정보 공유, 취업에 필요한 자격증/어학 시험, 취준생, 해외 취업 등	120	6.79%
	가족/관혼상제	집안 행사, 연애/결혼, 결혼 준비, 청첩장, 축가, 축의금, 문상 예절, 조의금, 제사 준비 등	108	6.11%
	쇼핑	온라인/오프라인 쇼핑 중 선호하는 쇼핑 방법, 선물을 고르는 상황 등	106	6.00%
	생활/주거환경	이사, 주거 유형, 생활권, 장 보기, 집안일, 계절/날씨 등	104	5.88%
	기타	꿈(목표), 군대 경험 등	102	5.77%
부분 합계			1768	100.00%
협력적 대화	영화/드라마/음악(콘텐츠)	영화관, 영화 관람, OTT 플랫폼, 배우, 가수, 대중음악, 클래식 음악 등	46	9.81%
	연극/뮤지컬/콘서트(공연)	뮤지컬, 콘서트, 연극, 클래식, 오페라, 국악, 발레/무용 등	45	9.59%
	전시회/박물관 (전시)	미술, 박물관 관람 및 여러 전시회 관람 지역별 박물관 관람 후기 및 체험 정보 공유 지역별 축제 정보 공유 및 권유	42	8.96%
	책/독서	작가, 독서, 소설, 시집, 종이책, 전자책(E-book) 등	49	10.45%
	스포츠/레저	경기 직관, 스포츠 종목, 올림픽, 축제, 행사, 등산, 수상 레저, 클라이밍 등	48	10.23%
	패션/뷰티	화장품, 옷, 액세서리, 유행하는 스타일, 상황에 따른 스타일	52	11.09%
	음식/음료	맛집, 밀키트에 대한 생각, 군것질에 대한 생각 및 어떤 음식을 먹을지 결정	49	10.45%
	반려동물	반려동물 입양, 유기 동물 문제, 동물 등록, 동물병원 등	46	9.81%
	여행 계획	교통편, 숙박, 여행지, 여행 경비 등	44	9.38%
	여행 일반	자유 여행, 패키지 여행, 국내 여행, 해외 여행, 관광지, 휴양지 등	48	10.23%
부분 합계			469	100.00%

비통제 대화	휴가	여행 시 교통/숙박 선택, 자연/휴양지 소개, 여행 국가/지역 선택, 추천 여행지(국내/해외) 등	26	6.36%
	대중교통	약속 만남 시 교통 선택, 새로운 교통수단 공유 키포드, 전철, 버스, 기차, 교통 약자석 등	26	6.36%
	음악	대중 음악, 선호 가수 및 노래 추천, 선호하는 음악 장르 등	25	6.11%
	건강/다이어트	가지고 있는 질병/알레르기, 건강 보조제, 건강을 위해 하고 있는 노력, 약 부작용, 다이어트 성공/실패 경험담 등	26	6.36%
	방송/연예	인생 드라마 추천, 예능 프로그램, 선호하는 배우 등	26	6.36%
	스포츠/레저/취미	경기 직접 관람, 등산, 여름/겨울 레저, E-스포츠(게임), 독서(시집, 문학 등), 영화 관람, 만화(웹툰 등), 웹소설 등	25	6.11%
	먹거리	최근 인기있는 음식과 경험, 가장 선호하는 음식, 추천하는 맛집, 배달 음식, 밀키트 등	31	7.58%
	우정	선호하는 것, 성격, 다툼과 화해 등에 대한 친구 사이의 대화, 국경/나이를 초월한 우정 등	26	6.36%
	경제/재테크	부동산, 주식 투자, 비트코인, 고물가, 세금, 인플레이션, 금리 인상 등	23	5.62%
	회사/학교	회사 생활, 회식, 인턴 생활, 진학에 대한 정보와 결정, 입시 경쟁, 전공, 성적 등	24	5.87%
	반려동물	반려동물 입양, 유기 동물 문제, 동물 병원 고비용, 동물 등록제, 반려동물과의 추억 등	28	6.85%
	취직	취업에 대한 정보 공유, 취업에 필요한 자격증/어학 시험, 취준생, 해외 취업 등	23	5.62%
	가족/관혼상제	집안 행사, 연애/결혼, 결혼 준비, 청첩장, 축가, 축의금, 문상 예절, 조의금, 제사 준비 등	24	5.87%
	쇼핑	온라인/오프라인 쇼핑 중 선호하는 쇼핑 방법, 선물을 고르는 상황 등	27	6.60%
	생활/주거환경	이사, 주거 유형, 생활권, 장 보기, 집안일, 계절/날씨 등	26	6.36%
	기타	꿈(목표), 군대 경험 등	23	5.62%
부분 합계			409	100.00%
전체 합계			2,646	100.03%

2. 화자 모집 결과

2.1. 인구 특성별 수집 결과

일상 대화 말뭉치 구축 설계 시 한 성별, 나이별, 지역별 목표치를 충족하기 위해 다양한 방법으로 화자 모집을 진행하였다. 화자 모집은 일반적으로 구인 사이트를 활용하였으나, 코로나-19라는 특수 상황으로 인해 그 효율이 기존에 비해 저조하였다. 따라서 이번 사업에서는 구인 사이트 활용 외에 화자 모집 전용 네이버 카페를 개설하였으며 각 지역의 홍보 사이트, 맘 카페, 종교 모임, 대학 동아리 등을 통한 홍보를 진행하여 화자를 모집하였다. 특히 말뭉치 구축 사업의 정확한 안내를 위해 네이버 카페를 통해 오는 길, 녹음 신청 방법 등 다양한 정보를 제공하였으며, 녹음 후기 작성 게시판도 개설하여 본 사업에 사람들이 안심하고 참여할 수 있도록 하였다. 이러한 과정을 통해 2,073명의 발화자로부터 수집한 약 2,800여 개의 시나리오(약 700시간 분량) 데이터를 수집하였고, 최종적으로는 2,000명의 발화자로부터 수집한 2,646개의 시나리오(약 661.5시간 분량) 데이터를 전사 및 납품하였다. 수집에 참여한 화자와 납품 데이터의 결과는 아래 표와 같다.

[표 30] 통제 및 비통제 데이터 전사 및 납품 결과(단위: 명)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	19	63	39	26	29	5	21	60	36	25	32	10	365	884
	인천	9	23	14	7	6	1	8	22	11	7	7	2	117	
	경기	25	81	49	17	19	5	31	88	34	22	23	8	402	
영남권	부산	6	21	18	6	8	6	9	26	18	9	12	7	146	524
	경남	6	19	19	12	10	4	7	20	18	10	11	4	140	
	울산	3	6	3	2	2	1	0	9	5	3	2	1	37	
	대구	7	18	13	6	4	3	5	18	11	6	6	1	98	
	경북	6	14	15	6	6	2	4	15	10	11	10	4	103	
호남권	광주	3	19	9	4	3	1	3	15	5	1	3	0	66	230
	전북	2	17	8	6	4	4	5	17	7	6	11	4	91	
	전남	4	10	8	4	3	1	3	15	5	9	8	3	73	
충청권	대전	6	22	12	4	3	2	6	28	10	5	4	2	104	248
	충북	4	11	6	3	3	1	3	10	6	5	4	1	57	
	충남	7	13	11	5	3	2	5	10	9	8	9	5	87	
강원권	강원	9	8	8	2	4	4	3	9	5	6	2	3	63	63
제주권	제주	4	6	4	3	3	3	3	13	3	3	5	1	51	51
합계		120	351	236	113	110	45	116	375	193	136	149	56	2,000	

전체 2,073명을 모집하였지만, 이중 최종적으로 전사된 데이터는 총 2,000명 분량이다.

[표 31] 통제 및 비통제 데이터 전사 및 납품 결과(단위: 시나리오 개수)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	31.0	68.2	46.0	40.0	46.2	8.3	33.7	69.0	48.3	30.5	40.2	11.3	472.7	1214.3
	인천	14.0	28.7	19.0	10.0	12.0	2.0	12.0	25.0	15.8	11.0	12.0	4.0	165.5	
	경기	42.5	92.7	74.5	31.3	35.3	9.0	47.3	100.0	57.8	34.5	41.2	10.0	576.2	
영남권	부산	9.3	23.3	21.0	10.5	11.0	7.3	9.7	29.5	20.0	10.2	12.7	9.3	173.8	655.8
	경남	10.0	22.7	21.3	16.5	12.0	4.0	11.7	23.7	20.3	10.7	12.0	4.8	169.7	
	울산	5.0	5.8	3.0	3.0	4.0	0.3	0.0	11.0	5.0	6.0	3.0	1.0	47.2	
	대구	10.0	20.3	16.0	9.3	7.0	4.3	9.0	19.3	17.0	10.5	9.3	2.0	134.2	
	경북	10.0	16.5	18.0	8.5	11.0	3.0	7.0	15.7	11.3	14.2	11.5	4.3	131.0	
호남권	광주	5.0	19.5	13.3	7.0	3.8	2.0	4.0	15.5	8.0	2.0	3.7	0.0	83.8	282.2
	전북	2.7	20.7	11.0	6.5	4.7	4.7	5.0	21.0	9.3	6.7	13.5	4.3	110.0	
	전남	5.5	10.3	9.3	6.3	4.0	1.0	3.8	17.2	6.5	11.3	9.7	3.3	88.3	
충청권	대전	7.5	22.0	15.0	7.0	5.0	2.0	8.3	29.7	11.7	7.3	3.5	3.0	122.0	312.0
	충북	4.0	12.3	9.0	4.5	6.0	2.0	6.0	10.3	9.0	6.7	5.8	2.0	77.7	
	충남	10.0	15.3	13.0	8.0	4.0	1.8	9.0	10.7	11.0	9.7	15.0	4.8	112.3	
강원권	강원	16.0	10.3	16.0	3.0	8.0	6.0	6.0	11.7	8.0	10.7	3.0	4.3	103.0	103.0
제주권	제주	5.0	8.5	6.3	6.0	4.0	6.0	4.0	17.5	6.0	5.3	8.0	2.0	78.7	78.7
합계		187.5	397.2	311.8	177.5	178.0	63.8	176.5	426.7	265.2	187.2	204.0	70.7	2646.0	

전체 2,801개 시나리오를 수집하였지만, 이중 최종적으로 전사된 데이터는 총 2,646개이다.

[표 32] 통제 및 비통제 데이터 전사 및 납품 결과(단위: 시간)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대	10대	20대	30대	40대	50대	60대	지역별	권역별
수도권	서울	7.8	17.0	11.5	10.0	11.5	2.1	8.4	17.3	12.1	7.6	10.0	2.8	118.2	303.6
	인천	3.5	7.2	4.8	2.5	3.0	0.5	3.0	6.3	4.0	2.8	3.0	1.0	41.4	
	경기	10.6	23.2	18.6	7.8	8.8	2.3	11.8	25.0	14.5	8.6	10.3	2.5	144.0	
영남권	부산	2.3	5.8	5.3	2.6	2.8	1.8	2.4	7.4	5.0	2.5	3.2	2.3	43.5	164.0
	경남	2.5	5.7	5.3	4.1	3.0	1.0	2.9	5.9	5.1	2.7	3.0	1.2	42.4	
	울산	1.3	1.5	0.8	0.8	1.0	0.1	0.0	2.8	1.3	1.5	0.8	0.3	11.8	
	대구	2.5	5.1	4.0	2.3	1.8	1.1	2.3	4.8	4.3	2.6	2.3	0.5	33.5	
	경북	2.5	4.1	4.5	2.1	2.8	0.8	1.8	3.9	2.8	3.5	2.9	1.1	32.8	
호남권	광주	1.3	4.9	3.3	1.8	1.0	0.5	1.0	3.9	2.0	0.5	0.9	0.0	21.0	70.5
	전북	0.7	5.2	2.8	1.6	1.2	1.2	1.3	5.3	2.3	1.7	3.4	1.1	27.5	
	전남	1.4	2.6	2.3	1.6	1.0	0.3	1.0	4.3	1.6	2.8	2.4	0.8	22.1	
충청권	대전	1.9	5.5	3.8	1.8	1.3	0.5	2.1	7.4	2.9	1.8	0.9	0.8	30.5	78.0
	충북	1.0	3.1	2.3	1.1	1.5	0.5	1.5	2.6	2.3	1.7	1.5	0.5	19.4	
	충남	2.5	3.8	3.3	2.0	1.0	0.5	2.3	2.7	2.8	2.4	3.8	1.2	28.1	
강원권	강원	4.0	2.6	4.0	0.8	2.0	1.5	1.5	2.9	2.0	2.7	0.8	1.1	25.8	25.8
제주권	제주	1.3	2.1	1.6	1.5	1.0	1.5	1.0	4.4	1.5	1.3	2.0	0.5	19.7	19.7
합계		46.9	99.3	78.0	44.4	44.5	16.0	44.1	106.7	66.3	46.8	51.0	17.7	661.5	

전체 약 701시간이 수집되었지만, 이중 최종적으로 전사된 데이터는 약 661.5시간 분량이다.

[표 33] 통제 데이터 전사 및 납품 결과(단위: 명)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	19	35	35	26	28	5	19	40	35	25	32	10	309	741
	인천	7	14	11	7	6	1	6	9	10	7	7	2	87	
	경기	25	56	45	17	19	5	30	61	34	22	23	8	345	
영남권	부산	6	14	17	6	8	6	8	18	15	9	12	7	126	466
	경남	6	16	18	12	10	4	7	16	16	10	11	4	130	
	울산	3	5	3	2	2	1	0	8	5	3	2	1	35	
	대구	6	11	11	6	4	3	4	11	11	6	5	1	79	
	경북	6	12	13	6	6	2	4	12	10	11	10	4	96	
호남권	광주	3	7	7	4	3	1	3	8	4	1	3	0	44	184
	전북	2	9	8	6	3	4	4	9	7	6	11	4	73	
	전남	4	9	8	4	3	1	3	10	5	9	8	3	67	
충청권	대전	5	9	8	4	3	2	5	8	7	5	3	2	61	190
	충북	4	9	6	3	3	1	3	6	6	5	4	1	51	
	충남	7	9	10	5	3	2	5	6	9	8	9	5	78	
강원권	강원	5	6	8	2	4	3	3	7	5	6	1	3	53	53
제주권	제주	2	3	3	3	3	3	2	6	3	3	5	1	37	37
합계		110	224	211	113	108	44	106	235	182	136	146	56	1,671	

[표 34] 통제 데이터 전사 및 납품 결과(단위: 시나리오 개수)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	31.0	37.7	42.0	39.0	44.2	8.3	30.7	43.0	45.3	30.5	40.2	11.3	403.2	1034.8
	인천	12.0	16.7	14.0	10.0	11.0	2.0	10.0	10.0	13.8	9.0	12.0	4.0	124.5	
	경기	41.5	63.7	68.5	31.3	35.3	9.0	46.3	69.0	56.8	34.5	41.2	10.0	507.2	
영남권	부산	9.3	16.3	19.0	10.5	11.0	7.3	8.7	22.0	17.0	10.2	12.7	9.3	153.3	590.3
	경남	10.0	17.7	20.3	16.5	12.0	4.0	11.7	19.7	18.3	10.7	12.0	4.8	157.7	
	울산	5.0	5.3	3.0	3.0	4.0	0.3	0.0	10.0	5.0	6.0	3.0	1.0	45.7	
	대구	9.0	13.3	14.0	9.3	7.0	4.3	8.0	11.3	15.0	10.5	8.3	2.0	112.2	
	경북	10.0	13.0	16.0	8.5	11.0	3.0	7.0	12.7	11.3	13.2	11.5	4.3	121.5	
호남권	광주	5.0	7.0	11.3	7.0	3.8	2.0	4.0	8.0	7.0	2.0	3.7	0.0	60.8	222.7
	전북	2.7	10.7	11.0	6.5	3.7	4.7	4.0	11.0	8.3	6.7	11.5	4.3	85.0	
	전남	4.0	9.3	8.3	5.3	4.0	1.0	3.3	10.7	6.5	11.3	9.7	3.3	76.8	
충청권	대전	6.0	10.0	11.0	7.0	5.0	2.0	7.3	8.7	8.7	7.3	2.5	3.0	78.5	248.5
	충북	4.0	10.3	9.0	4.5	6.0	2.0	5.0	6.3	9.0	6.7	5.8	2.0	70.7	
	충남	9.0	9.3	12.0	8.0	4.0	1.8	8.0	6.7	11.0	9.7	15.0	4.8	99.3	
강원권	강원	8.0	8.3	15.0	3.0	8.0	5.0	6.0	7.7	7.0	10.7	2.0	4.3	85.0	85.0
제주권	제주	2.0	3.0	4.3	6.0	4.0	6.0	2.0	7.0	6.0	5.3	8.0	2.0	55.7	55.7
합계		168.5	251.7	278.8	175.5	174.0	62.8	162.0	263.7	246.2	184.2	199.0	70.7	2237.0	

[표 35] 통제 데이터 전사 및 납품 결과(단위: 시간)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대	10대	20대	30대	40대	50대	60대	지역별	권역별
수도권	서울	7.8	9.4	10.5	9.8	11.0	2.1	7.7	10.8	11.3	7.6	10.0	2.8	100.8	258.7
	인천	3.0	4.2	3.5	2.5	2.8	0.5	2.5	2.5	3.5	2.3	3.0	1.0	31.1	
	경기	10.4	15.9	17.1	7.8	8.8	2.3	11.6	17.3	14.2	8.6	10.3	2.5	126.8	
영남권	부산	2.3	4.1	4.8	2.6	2.8	1.8	2.2	5.5	4.3	2.5	3.2	2.3	38.3	147.6
	경남	2.5	4.4	5.1	4.1	3.0	1.0	2.9	4.9	4.6	2.7	3.0	1.2	39.4	
	울산	1.3	1.3	0.8	0.8	1.0	0.1	0.0	2.5	1.3	1.5	0.8	0.3	11.4	
	대구	2.3	3.3	3.5	2.3	1.8	1.1	2.0	2.8	3.8	2.6	2.1	0.5	28.0	
	경북	2.5	3.3	4.0	2.1	2.8	0.8	1.8	3.2	2.8	3.3	2.9	1.1	30.4	
호남권	광주	1.3	1.8	2.8	1.8	1.0	0.5	1.0	2.0	1.8	0.5	0.9	0.0	15.2	55.7
	전북	0.7	2.7	2.8	1.6	0.9	1.2	1.0	2.8	2.1	1.7	2.9	1.1	21.3	
	전남	1.0	2.3	2.1	1.3	1.0	0.3	0.8	2.7	1.6	2.8	2.4	0.8	19.2	
충청권	대전	1.5	2.5	2.8	1.8	1.3	0.5	1.8	2.2	2.2	1.8	0.6	0.8	19.6	62.1
	충북	1.0	2.6	2.3	1.1	1.5	0.5	1.3	1.6	2.3	1.7	1.5	0.5	17.7	
	충남	2.3	2.3	3.0	2.0	1.0	0.5	2.0	1.7	2.8	2.4	3.8	1.2	24.8	
강원권	강원	2.0	2.1	3.8	0.8	2.0	1.3	1.5	1.9	1.8	2.7	0.5	1.1	21.3	21.3
제주권	제주	0.5	0.8	1.1	1.5	1.0	1.5	0.5	1.8	1.5	1.3	2.0	0.5	13.9	13.9
합계		42.1	62.9	69.7	43.9	43.5	15.7	40.5	65.9	61.5	46.0	49.8	17.7	559.3	

[표 36] 비통제 데이터 수집 카테고리별 전사 및 납품 결과

카테고리	세부 장소	목표 시간	시나리오 수집 개수	수집 시간	목표 대비 수집 비율(%)
무소음 환경	사무실	12.5	61	15.25	122
	회의실	12.5	61	15.25	122
생활 환경	가정집	12.5	55	13.75	110
	카페	12.5	44	11	88
교통 환경	도로변 및 정류장	12.5	58	14.5	116
	대합실 및 개찰구	12.5	64	16	128
자연 환경	공원	12.5	31	7.75	62
	생태환경	12.5	35	8.75	70
계		100	409	102.25	102.25

[표 37] 비통제 데이터 발화자 성장지별 전사 및 납품 결과

권역별	권역 세부	목표 수집 수량	세부 수집 수량	수집 합계	목표 비중(%)	수집 백분율(%)	목표 대비 수집 비율(%)
수도권	서울	180	69.50	179.50	45	43.89	99.72%
	경기		41.00				
	인천		69.00				
영남권	대구	60	20.50	65.50	15	16.01	109.17%
	경북		12.00				
	부산		1.50				
	경남		22.00				
	울산		9.50				
호남권	광주	60	23.00	59.50	15	14.55	99.17%
	전북		25.00				
	전남		11.50				
충청권	대전	60	43.50	63.50	15	15.53	105.83%
	충북		7.00				
	충남		13.00				
강원권	강원	10	18.00	18.00	5	4.40	180.00%
제주권	제주	10	23.00	23.00	5	5.62	230.00%
합계		380	409.00	409.00	100	100.00	107.63%

2.2. 주제별 나이대 분포

수집에 활용된 대화 주제들은 수집에 참여하는 인원들이 자신들이 주제를 선택하는 형태로 진행되었다. 화자들이 관심 분야에 대한 대화를 할 수 있도록 최대한 다양한 선택지를 제시하였지만 사업 후반부에는 일부 주제가 목표치를 달성하여 주제 선택의 범위를 제한하였다. 각 주제별 한 화자당 대화 참여 수는 1~4개이며, 2~4인 대화가 혼용되어 있다.

각 주제들은 나이별로 다양하게 사용되었는데, 10대의 경우 음악, 회사/학교, 우정을, 20대의 경우 휴가, 스포츠/레저/취미, 기타를, 30대의 경우 생활/주거환경, 취직, 기타를, 40대의 경우 경제/재테크, 대중교통, 방송/연예를, 50대의 경우 건강/다이어트, 경제/재테크, 대중교통을, 60대 이상의 경우 건강/다이어트, 경제/재테크, 먹거리 및 가족/관혼상제 순으로 선택되었다.

10대의 경우 최근 인기 있는 K-pop과 회사/학교생활을, 20대의 경우 휴가와 취미를, 30대의 경우 주거환경과 회사와 관련된 취직, 그 밖에 반려동물에 대한 관심도 높았으며, 40~50대의 경우 경제적인 뉴스(부동산, 주식시장)에 관심이 많음을 알 수 있었고, 50대 이상의 경우 건강에 대한 관심이 증가되는 모습을 보였으며, 60대의 경우 가족에 대한 관심도가 다른 나이대보다 높다는 것을 알 수 있었다. 이는 실제 나이대에 관심이 높은 분야를 적절하게 선택한 것으로 판단된다.

[표 38] 주제별 나이대 분포(단위: 시간)

주제	10대	20대	30대	40대	50대	60대	합계	비율
휴가	2.9	12.5	8.6	3.4	5.9	0.9	34.3	6.29%
대중교통	5.8	8.9	6.1	7.3	6.8	2.1	37.0	6.80%
음악	8.7	10.1	6.5	3.0	2.9	0.8	32.0	5.88%
건강/다이어트	2.3	7.8	5.5	6.6	9.3	4.0	35.5	6.52%
방송/연예	5.8	8.9	7.9	7.0	6.2	1.0	36.8	6.75%
스포츠/레저/취미	5.5	11.3	7.1	3.5	4.0	1.9	33.3	6.11%
먹거리	5.3	10.5	8.4	4.3	5.0	2.5	36.0	6.61%
우정	9.1	9.3	7.1	2.5	3.5	1.0	32.5	5.97%
경제/재테크	2.0	5.1	8.0	8.0	8.0	3.1	34.3	6.29%
회사/학교	8.2	10.8	4.2	4.5	4.5	1.6	33.8	6.20%
반려동물	5.2	9.9	8.0	4.5	4.1	1.6	33.3	6.11%
취직	3.9	10.3	9.3	4.5	6.1	1.8	35.8	6.57%
가족/관혼상제	3.0	8.4	8.0	5.0	6.1	2.5	33.0	6.06%
쇼핑	4.7	9.2	8.9	6.3	3.7	0.5	33.3	6.11%
생활/주거환경	1.3	8.8	10.6	5.4	4.2	2.2	32.5	5.97%
기타	5.1	12.0	9.3	2.5	2.0	0.4	31.3	5.74%
	78.8	153.9	123.4	78.2	82.3	27.7	544.3	100.00%
영화/드라마/음악(콘텐츠)	1.1	5.0	1.7	2.3	1.0	0.5	11.5	9.81%
연극/뮤지컬/콘서트(공연)	1.5	6.6	1.1	1.0	0.6	0.4	11.3	9.59%
전시회/박물관(전시)	0.8	4.6	4.1	0.9	0.1	0.0	10.5	8.96%
책/독서	2.7	4.9	2.5	1.4	0.5	0.3	12.3	10.45%
스포츠/레저	0.3	6.9	1.4	1.5	1.8	0.3	12.0	10.23%
패션/뷰티	1.5	6.6	1.4	1.2	1.8	0.6	13.0	11.09%
음식/음료	1.1	4.4	2.3	1.4	2.5	0.5	12.3	10.45%

반려동물	1.0	3.3	2.7	1.2	2.0	1.4	11.5	9.81%
여행 계획	0.9	5.5	1.9	0.8	1.3	0.6	11.0	9.38%
여행 일반	1.5	4.3	1.8	1.4	1.8	1.3	12.0	10.23%
	12.3	52.0	20.9	13.0	13.2	5.9	117.3	100.00%

2.3. 대화 유형별 분포

대화 유형별 참여 인원은 초기 제시된 목표에 맞추어 화자를 섭외하고 음성 자료를 수집하였다. 다자 대화의 경우 2인 대화보다 참여비가 더 많이 지급되어서 모집 초반 상당히 많은 수의 신청을 받아 의외로 빨리 수집할 수 있었다. 그러나 수집된 데이터의 통계를 분석한 결과, 상대적으로 젊은 층과 여성 발화자들이 대부분이었다. 따라서 각 지역별, 성별, 나이별 적절한 분포치를 유지하기 위해 다자 대화는 10월부터 수량을 조절하였다.

이렇게 모집된 데이터 중 전사 및 납품을 완료한 일상 대화는 전체 분량의 약 66%인 1,768건이며, 협력적 대화가 전체의 약 17%인 469건이었다. 비통제 대화는 전체의 약 15%인 409건을 전사 및 납품하였다. 2인 대화 비율은 전체의 약 80%인 2,121건, 3인과 4인 대화 비율은 전체의 약 19%인 525건이다.

[표 39] 대화 유형 및 인원별 분포(단위: 대화 수량)

대화 유형	2인 대화	3인 대화	4인 대화	합계	비율
일상 대화	1,481	215	72	1,768	66.82%
협력적 대화	231	176	62	469	17.72%
비통제 대화	409	0	0	409	15.46%
합 계	2,121	391	134	2,646	100.00%

2.4. 주제별 성별 분포

주제별 성별 분포의 경우 모든 주제를 고르게 사용하였으나, 남성의 경우 경제/재테크, 대중교통, 취직을, 여성의 경우 방송/연예, 먹거리, 휴가가 높은 순위로 선택되었다.

[표 40] 주제별 성별 분포(단위: 대화 수량)

주제	남성	여성	합계	비율
휴가	59.7	77.3	137.0	6.29%
대중교통	79.3	68.7	148.0	6.80%
음악	69.0	59.0	128.0	5.88%
건강/다이어트	66.3	75.8	142.0	6.52%
방송/연예	66.3	80.7	147.0	6.75%
스포츠/레저/취미	73.3	59.7	133.0	6.11%
먹거리	65.8	78.2	144.0	6.61%
우정	71.4	58.6	130.0	5.97%
경제/재테크	80.8	56.2	137.0	6.29%
회사/학교	69.3	65.8	135.0	6.20%
반려동물	66.8	66.2	133.0	6.11%
취직	75.0	68.0	143.0	6.57%
가족/관혼상제	67.6	64.4	132.0	6.06%
쇼핑	69.0	64.0	133.0	6.11%
생활/주거환경	66.5	63.5	130.0	5.97%
기타	71.7	53.3	125.0	5.74%
	1117.8	1059.2	2177.0	100.00%
영화/드라마/음악(콘텐츠)	15.7	30.3	46.0	9.81%
연극/뮤지컬/콘서트(공연)	19.3	25.7	45.0	9.59%

전시회/박물관(전시)	23.5	18.5	42.0	8.96%
책/독서	29.1	19.9	49.0	10.45%
스포츠/레저	25.3	22.7	48.0	10.23%
패션/뷰티	18.3	33.8	52.0	11.09%
음식/음료	14.8	34.3	49.0	10.45%
반려동물	21.2	24.8	46.0	9.81%
여행 계획	17.8	26.2	44.0	9.38%
여행 일반	13.1	34.9	48.0	10.23%
	198.0	271.0	469.0	100.00%

2.5. 화자 관계별 분포

수집에 참여한 화자 간 관계의 경우 모집 단계에서 다양한 관계의 화자들이 섭외될 수 있도록 홍보하고 모집하였다. 그러나 실제 데이터의 수집은 자연스러운 대화가 이루어지는 형태로 진행되어야 하다 보니 참여자 간 친밀함이 필수적일 수밖에 없었다. 따라서 실제 녹음에 참여한 화자들의 관계를 살펴보면, (친구 - 부부 - 연인) 순으로 비율이 높게 나타났다. 뒤를 이어 기타(초면, 지인 소개, 친구의 가족 등), 지인, 부모/자녀, 형제/자매, 직장 동료 순으로 나타났다. 예외적으로 아르바이트 사이트를 통해 수집에 혼자 참여한 화자들이 있어 짝을 지어 녹음을 진행하였으나 대화가 원활하게 되지 않는다고 판단된 경우 수집을 중단하였다.

[표 41] 화자 간 관계별 수집 결과(단위: 개)

화자 간 관계	2인 대화	3인 대화	4인 대화	합계	비율
친구	489	223	60	772	29.18%
부부	525	4	0	529	19.99%
부모/자녀	111	24	0	135	5.10%
형제/자매	93	12	0	105	3.97%
연인	485	0	0	485	18.33%
직장 동료	47	20	8	75	2.83%
이웃사촌	0	0	0	0	0.00%
모임·동아리 지인	91	41	46	178	6.73%
대학 선후배	38	20	4	62	2.34%
고향 선후배	1	0	0	1	0.04%
교회 지인	4	4	0	8	0.30%
사제 관계	0	4	0	4	0.15%
기타 가족	31	11	0	42	1.59%
기타	206	28	16	250	9.45%
합계	2,121	391	134	2,646	100.00%

2.6. 직업별 분포

최대한 다양한 직업군을 모집하기 위해 평일과 주말 및 공휴일을 포함한 다양한 날짜에도 수집을 진행하였고, 학생(약 32%), 사무 종사자(약 24%), 무직/취업준비생(약 13%) 순으로 참여도가 높았다. 그 뒤를 이어 주부, 전문가 및 관련 종사자, 기타 순으로 나타났다.

[표 42] 직업별 수집 결과(단위: 명)

직업	모집 인원	비율
경영/관리직	34	1.70%
전문가 및 관련 종사자	103	5.15%
사무 종사자	484	24.20%
서비스 종사자	61	3.05%
판매/영업 종사자	47	2.35%
농업/임업/어업 종사자	5	0.25%
기술자 종사자(장치/기계 조작 및 조립 종사자)	11	0.55%
단순노무 종사자	1	0.05%
학생	644	32.20%
주부	248	12.40%
무직/취업준비생	263	13.15%
기타	99	4.95%
합 계	2,000	100.00%

2.7. 학력별 분포

모집 인원의 학력을 분석하면 대재와 대졸이 70% 이상을 차지하며, 그 외에 고졸, 대학원 이상, 중졸 순서로 비율을 차지하고 있다.

[표 43] 학력별 수집 결과(단위: 명)

학력	모집 인원	비율
초졸 이하	12	0.60%
중졸	103	5.15%
고졸	315	15.75%
대재	430	21.50%
대졸	996	49.80%
대학원 이상	144	7.20%
합계	2,000	100.00%

2.8. 출생지별 분포

모집 화자의 출생지별 분포의 경우 1순위는 서울(약 24%)이며, 2순위는 경기(약 14%), 3순위는 부산(약 9%)으로 나타났다. 이는 초기 수집 목표 인원을 산출하였던 지역별 인구 분포와 유사하게 서울 및 수도권, 5대 광역시, 각 광역자치단체 순으로 화자가 모집되었음을 확인할 수 있다.

[표 44] 출생지별 화자 모집 결과(단위: 명)

출생지	모집 인원	비율
서울	482	24.10%
경기	288	14.40%
인천	103	5.15%
대구	111	5.55%
경북	103	5.15%
부산	194	9.70%
경남	123	6.15%
울산	33	1.65%
광주	72	3.60%
전북	108	5.40%
전남	63	3.15%
대전	101	5.05%
충북	43	2.15%
충남	79	3.95%
강원	52	2.60%
제주	45	2.25%
합계	2,000	100.00%

2.9. 주 성장지별 분포

모집 화자의 성장지별⁵⁾ 분포의 경우 1순위는 경기(약 20%)이며, 2순위는 서울(약 18%), 3순위는 부산(약 7%)으로 나타났다. 지역별 화자 모집 비율은 주 성장지를 기준으로 설계되었고, 해당 비율을 고려하여 수집하였다.

[표 45] 주 성장지별 화자 모집 결과(단위: 명)

주 성장지	모집 인원	비율
서울	365	18.25%
경기	402	20.10%
인천	117	5.85%
대구	98	4.90%
경북	103	5.15%
부산	146	7.30%
경남	140	7.00%
울산	37	1.85%
광주	66	3.30%
전북	91	4.55%
전남	73	3.65%
대전	104	5.20%
충북	57	2.85%
충남	87	4.35%
강원	63	3.15%
제주	51	2.55%
합계	2,000	100.00%

5) 초, 중, 고등학교를 나온 지역을 의미한다. 지역이 여러 곳일 경우 가장 오래 있었던 지역을 표시하였다.

2.10. 현 거주지별 분포

모집 화자의 현 거주지별 수집 결과의 경우 1순위는 서울(약 29%), 2순위는 경기(약 16%), 3순위 부산(약 14%)이다.

[표 46] 현 거주지별 화자 모집 결과(단위: 명)

현 거주지	모집 인원	비율
서울	594	29.70%
경기	327	16.35%
인천	131	6.55%
대구	146	7.30%
경북	11	0.55%
부산	285	14.25%
경남	24	1.20%
울산	7	0.35%
광주	101	5.05%
전북	93	4.65%
전남	4	0.20%
대전	204	10.20%
충북	3	0.15%
충남	12	0.60%
강원	33	1.65%
제주	25	1.25%
합계	2,000	100.00%

3. 정책 제언

이 사업을 통해 다양한 지역의 나이와 성별을 가진 사람들이 지인들과 일상적으로 나누는 대화를 수집하고, 전사하여 말뭉치를 구축하였다. 이 말뭉치는 현재의 언어 생활을 기록하고, 자원화되어 여러 분야의 산업 발전에도 기여할 수 있을 것으로 기대된다.

본 사업을 진행하면서 느꼈던 몇 가지 생각을 정리하면 다음과 같다.

- 여성, 청년층, 대도시 중심의 연고를 가진 대화 참여자는 많지만, 남성, 노인층, 수도권 및 대도시를 벗어난 지역의 참여자는 찾기 어려웠다. 따라서 추가적인 보상이나 수집 기간 연장을 통해 참여자를 모집할 수밖에 없어 비용이 증가하게 되었다. 하지만, 일상 대화 말뭉치를 통해 개발되는 인공지능 기반 서비스는 소외 계층이나 정보 접근성이 떨어지는 계층에게 더욱 필요할 것으로 생각된다. 인구통계학적인 비율을 고려하여 균형 있게 데이터를 수집하는 것도 중요하지만, 정책적인 판단에 따라 특정 계층(사회적 약자나 수집 취약 계층)에 대한 데이터를 더 많이 수집하고 축적하는 것도 의미가 있다.
- 억양구 단위의 분할 및 전사와 관련하여 현재 전사 지침⁶⁾에는 “음성 분절 및 전사의 기본 단위는 긴 휴지, 경계 억양, 경계 말 장음화 등을 특징으로 하는 억양구가 되도록 하며, 하나의 전사 단위가 가능한 3초 이상으로 길어지지 않도록 한다.”와 같이 명시되어 있으나, 작업자들이 억양구 단위를 명확하게 이해하고 작업하기에 다소 어려움이 있다. 억양구의 필요성 및 활용 사례 등을 관련 학계의 자문 내용과 함께 구체적으로 지침에 명시하면 작업 과정에서 혼란을 최소화할 수 있을 것이다. 음성 관련 연구 및 개발에 많이 활용되는 말뭉치(Common Voice, LibriSpeech)는 문장 단위로 구축되는 경우가 많기 때문에, 일상 대화 말뭉치의 경우 억양구 단위 말뭉치 구축에 대해 논리적인 근거가 구체적으로 제시된다면 의미 있는 말뭉치로 성장해 갈 것이다.
- 전사 지침과 관련하여 전사 작업 과정 도중 내용 변경이 이루어지면, 해당 시점까지 진행된 모든 작업에 대한 재검토를 통해 수정된 지침을 반영해야 한다. 새로운 말뭉치 구축 사업을 시작하기 전에, 기존에 구축된 말뭉치를 언어 분석 관점에서

6) [붙임 1]의 3.3. 전사 단위 참조

검토하고 개선 여지가 있는 지침에 대해서는 신속히 개정하는 것이 사업의 안정성과 말뚝치의 품질을 위해서 매우 중요한 과정이라고 생각한다.

1. 파일 형식 및 개요

1.1. 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	8자리 일련번호
S: 구어 말뭉치	D: 사적 대화	RW: 원시 말뭉치	22	#####

- 예시

- SDRW2200000001.sjml 원시 말뭉치 첫 번째 파일
- ※ 참고: 음성 파일 파일명 부여 방식
- SDRW2200000001.pcm 음성 원본 첫 번째 파일
- SDRW2200000001-00001.pcm 음성 원본 첫 번째 파일의 정제본 첫 번째 파일

1.2. 음성 파일 포맷

- 기본: 샘플링 16kHz, 양자화 16bits headerless(little endian) linear PCM
- 추가: 샘플링 44.1kHz, 양자화 16bits headerless(little endian) linear PCM
- 정제본: 채널별 mono 변환

1.3. 말뭉치 파일 포맷

- UTF-8, 줄 바꿈 문자 LF(UNIX)

2. 말뭉치 형식

2.1. JSON 구조

수준 1	수준 2	수준 3	수준 4	타입	설명
id				string	말뭉치 파일 아이디
metadata				object	말뭉치 파일의 메타 정보
	title			string	말뭉치 파일 제목
	creator			string	구축자: 국립국어원
	distributor			string	배포자: 국립국어원
	year			string	구축 연도: 2022
	category			string	분류: 구어 > 사적 대화 > 일상 대화 / 협력적 대화
	annotation_level			array(string)	분석 층위: 원시
	sampling			string	샘플링 방식: 본문 전체
document				array(object)	대화 정보
	id			string	대화 아이디
	metadata			object	대화 메타 정보
		title		string	대화 제목: 2인 일상 대화
		author		string	저작권자: 개인 발화자
		publisher		string	발행자: 개인 발화 녹음
		date		string	녹음일자: YYYYMMDD
		topic		string	대화 주제: 대주제 > 세부주제
		environment		string	수집 환경 정보[비통제]: 환경 > 장소
		speaker		array(object)	화자 정보
			id	string	화자 아이디
			age	string	나이
			occupation	string	직업
			sex	string	성별
			birthplace	string	출생지
			principal_residence	string	주 성장지
			current_residence	string	현 거주지
			education	string	학력
		setting		object	환경 정보
			relation	string	화자 간 관계
			device	string	수집 단말기 종류[비통제]
			mic	string	외장 마이크 종류[비통제]
	utterance			array(object)	발화 정보
		id		string	발화 아이디
		form		string	철자 전사
		original_form		string	발음 전사
		speaker_id		string	화자 아이디
		start		num	발화 시작 시간
		end		num	발화 종료 시간
		note		string	전사자 기타 메모

- 수준에 따라 스페이스 4개로 들여쓰기를 하여 요소의 계층을 시각화한다.

```
{
  "id": "SDRW2200000613",
  "metadata": {
    "title": "국립국어원 구어 말뭉치 SDRW2200000613",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2022",
    "category": "구어 > 사적대화 > 일상대화",
    "annotation_level": [
      "원시"
    ],
    "sampling": "본문 전체"
  },
  "document": [
    {
      "id": "SDRW2200000613.1",
      "metadata": {
        "title": "2인 일상 대화",
        "author": "개인 발화자",
        "publisher": "개인 발화 녹음",
        "date": "20220827",
        "topic": "휴가 > 휴가 가기 좋은 여행지 공유",
        "environment": "",
        "speaker": [
          {
            "id": "SD2200128",
            "age": "20대",
            "occupation": "사무 종사자",
            "sex": "여성",
            "birthplace": "경기",
            "principal_residence": "경기",
            "current_residence": "경기",
            "education": "대졸"
          },
          {
            "id": "SD2200129",
            "age": "50대",
            "occupation": "주부",
            "sex": "여성",
            "birthplace": "경북",
            "principal_residence": "경북",
            "current_residence": "경기",
            "education": "고졸"
          }
        ]
      },
      "setting": {
        "relation": "부모/자녀",
        "device": "",
        "mic": ""
      }
    },
    {
      "utterance": [
        {
          "id": "SDRW2200000613.1.1.1",
          "form": "엄마 올해 여름휴가는 어디로 다녀왔어?",
          "original_form": "엄마 올해 여름휴가는 어디로 다녀왔어?",
          "speaker_id": "SD2200128",
          "start": "0.42000",
          "end": "4.69100",
          "note": ""
        },
        {
          "id": "SDRW2200000613.1.1.2",
          "form": "응 아빠랑 제주도 갈까 하다가",
          "original_form": "응 아빠랑 제주도 갈까 하다가",
          "speaker_id": "SD2200129",
          "start": "5.18000",
          "end": "8.57900",
          "note": ""
        }
      ]
    }
  ]
}
```

2.2. 각 요소별 설명

2.2.1. 말뭉치 파일

- 말뭉치 파일 아이디(id): 1.1의 파일명 부여 방식에 따른 14자리

2.2.2. 말뭉치 파일 메타 정보(metadata)

- 말뭉치 파일 제목(title): 국립국어원 구어 말뭉치 + 말뭉치 파일 아이디(예: 국립국어원 구어 말뭉치 SDRW2200000001)
- 구축자(creator): 국립국어원
- 배포자(distributor): 국립국어원
- 구축 연도(year): 2022
- 분류(category): 구어 > 사적 대화 > 일상 대화
구어 > 사적 대화 > 협력적 대화
구어 > 사적 대화 > 비통제 대화
- 분석 층위(annotation_level): 원시
- 샘플링 방식(sampling): 본문 전체

2.2.3. 대화(document)

- 대화 아이디(id): 말뭉치 파일 아이디 + . + 1(예: SDRW2200000001.1)

2.2.4. 대화 메타 정보(document > metadata)

- 대화 제목(title): 2인/3인/4인 일상 대화/협력적 대화/비통제 대화
- 저작권자(author): 개인 발화자
- 발행자(publisher): 개인 발화 녹음
- 녹음일자(date): 연월일 YYYYMMDD
- 대화 주제(topic): 대화 주제 및 대화 세부 주제

대화 주제				
일상 대화		협력적 대화		
1	휴가	1	문화	영화/드라마/음악 (콘텐츠)
2	대중교통	2		연극/뮤지컬/콘서트 (공연)
3	음악	3		전시회/박물관 (전시)
4	건강/다이어트	4		책/독서
5	방송/연예	5		스포츠/레저
6	스포츠/레저/취미	6		패션/뷰티
7	먹거리	7		음식/음료
8	우정	8		반려동물
9	경제/재테크	9	관광	여행계획
10	회사/학교	10		여행일반
11	반려동물			
12	취직			
13	가족/관혼상제			
14	쇼핑			
15	생활/주거 환경			
16	기타			

- 수집 환경 정보(environment): 비통제 대화 음성 수집 환경 및 수집 장소

수집 환경[비통제]	
1	무소음 환경
2	생활 환경
3	교통 환경
4	자연 환경
수집 장소[비통제]	
1	사무실
2	회의실
3	가정집
4	카페
5	도로변 및 정류장
6	대합실 및 개찰구
7	공원
8	생태환경

2.2.5. 화자 정보(document > metadata > speaker)

- 화자 아이디(id): 화자 고유 아이디 부여, 대화가 다르더라도 화자가 동일하면 동일한 아이디 부여
단, 화자가 교정기를 착용한 경우에는 구축 연도 다음 숫자 1을 넣어 표시(한 화자가 교정기를 뺐다 넣었다 하지 않도록 함)
(예: 교정기 미착용 화자 A: SD2200001, 교정기 착용 화자 B: SD2210002)
- 나이(age): 10대/20대/30대/40대/50대/60대 이상
- 직업(occupation): ‘한국표준직업분류’를 준용한 아래에서 선택

1) 경영/관리직	2) 전문가 및 관련 종사자
3) 사무 종사자	4) 서비스 종사자
5) 판매/영업 종사자	6) 농업/임업/어업 종사자
7) 기능원 및 관련 기능 종사자	8) 기술자 종사자(장치/기계 조작 및 조립 종사자)
9) 단순노무 종사자	10) 군인
11) 학생	12) 주부
13) 무직/취업준비생	14) 기타
- 성별(sex): 남성/여성/NA
- 출생지(birthplace): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주
- 주 성장지(principal_residence): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주
- 현 거주지(current_residence): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주
- 학력(education): 초졸 이하/중졸/고졸/대재/대졸/대학원 이상

2.2.6. 환경 정보(document > metadata> setting)

- 화자 간 관계(relation): 아래에서 선택

- | | |
|------------|--------------|
| 1) 친구 | 2) 부부 |
| 3) 부모/자녀 | 4) 형제/자매 |
| 5) 연인 | 6) 직장 동료 |
| 7) 이웃사촌 | 8) 모임·동아리 지인 |
| 9) 대학 선후배 | 10) 교회 지인 |
| 11) 고향 선후배 | 12) 사제 관계 |
| 13) 기타 가족 | 14) 기타 |

- 수집 단말기 종류(device): 비통제 대화 수집 시 사용하는 휴대전화 단말기 종류

수집 단말기 종류[비통제]			
android		iphone	
1	android_SAMSUNG Galaxy S10 SM-G973N	1	iPhone XS
2	android_SAMSUNG Galaxy S20 SM-G988N	2	iPhone 11
3	android_SAMSUNG Galaxy S22 SM-S901N	3	iPhone 12

- 외장 마이크 종류(mic): 비통제 대화 수집 시 야외 장소에서 외부 소음으로 인해 대화가 잘 들리지 않는 경우에 한해 외장 마이크를 사용

외장 마이크 종류 [비통제]
BOYA BY-MM1

2.2.7. 발화 정보(document > utterance)

- 발화 아이디(id): 대화 아이디 + . + 1 + . + 1 + . + 발화 번호(예: SDRW2200000001.1.1.4)
- 철자 전사(form): 철자 전사 결과
- 발음 전사(original_form): 발음 전사 결과
- 발화 시작 시간(start): 해당 발화의 음성 원본에서의 시작 시간을 초 단위(소수 5자리까지 필수)로 표기(예: 30.56600)
- 발화 종료 시간(end): 해당 발화의 음성 원본에서의 종료 시간을 초 단위(소수 5자리까지 필수)로 표기(예: 32.48262)
- 전사자 기타 메모(note): 녹음실 밖의 관계자의 개입으로 녹음이 중단되는 경우 등 관계자와 나눈 대화는 전사하지 않고 메모를 남김.

3. 전사 지침

3.1. 기본 원칙

- 음성 자료의 전사는 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행(이중 전사)한다.

- 발음 전사는 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 적용하여 발음 나는 대로 적는다.
*그 외 표준 발음에 맞게 발음한 경우에 발음 전사를 할 때에는 한글 맞춤법, 표준어 규정, 외래어 표기법 등 관련 어문 규정에 따라 한글로 적는다.
- 철자 전사는 한글 맞춤법 및 표준어 규정에 따라 적는 것으로, 발화 내용은 기본적으로 한글 맞춤법 및 표준어 규정에 따라 전사하며 띄어쓰기도 한글 맞춤법에 따른다.
- 발음 전사는 숫자, 외래어, 기호, 단위 등도 한글 맞춤법, 표준어 규정, 외래어 표기법 등 관련 어문 규정에 따라 한글로 적는다.
- 느낌표나 쉼표는 사용하지 않으며 문장이 완전히 종결되었을 때는 마침표를 사용한다.
- 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분한다.('응', '네', '-어', '-어요' 등)

3.2. 화자 표시

- 화자 아이디, 성별, 나이, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 'NA'로 표시한다.
- 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명하지 않을 경우에는 'NA'로 표시한다.

3.3. 전사 단위

- 음성 분절 및 전사의 기본 단위는 긴 휴지, 경계 억양, 경계 말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 하며, 하나의 전사 단위가 가능한 3초 이상으로 길어지지 않도록 한다.
※ 음성 정제본 하나가 하나의 전사 단위가 되도록 한다.
- 긴 쉼에 의해 나뉘는 경우는 통사적으로 완성이 되지 않았다 하더라도 전사 단위로 구분(음성 분절)하여 전사한다.

음성1_고등학교 때 제주도 때문에 비행기를 타봤는데
음성2_한참 뒤에 타 본거잖아.

- 음가 손실의 우려가 있는 경우에는 분절하지 않고 의미상 분절되어야 할 문장이 종결되는 부분에 종결 부호(마침표, 물음표)를 붙인다.

나는 뭐든 다 좋을 것 같아 여행을 가면. 너는 어때?

3.4. 발화 겹침

- 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다.

주 발화: 1: 딸 하나 낳아서
 맞장구 발화: 2: 네.
 주 발화: 3: 세 살 먹어 잊어버리고

3.4.1 비통제 대화의 발화 겹침

- 단일 채널로 녹음된 비통제 대화의 발화 겹침이 발생하는 경우 아래와 같이 전사한다.

주발화자: 너는 혹시 반려동물을 키워본 적이 있어?

부발화자: 예전에

1. 발화 겹침 구간의 주발화자의 음성은 명확하게 들리고, 부발화자의 음성은 전사가 불가능할 정도로 들리지 않을 때

주발화자: 너는 혹시 반려동물을 키워본 적이 있어? 발화 겹침 표시

부발화자: ((xxx)) 발화 겹침 표시

- 1-1. 주발화자의 발화 중 부발화자가 반복적으로 맞장구 발화를 하거나, 웃음, 기침 등의 준 음성 및 기타 소리들이 포함되었을 때

주발화자: 너는 혹시 반려동물을 키워본 적이 있어? 발화 겹침 표시하지 않음

2. 발화 겹침 구간의 부발화자의 음성만 명확하게 들리고, 주발화자의 음성은 전사가 불가능할 정도로 들리지 않을 때

주발화자: 너는 혹시 반려동물을 ((xxx)) 적이 있어? 발화 겹침 표시

부발화자: 예전에 발화 겹침 표시

3. 주발화자와 부발화자의 음성이 모두 전사 가능할 정도로 명확하게 들릴 때

주발화자: 너는 혹시 반려동물을 키워본 적이 있어? 발화 겹침 표시

부발화자: 예전에 발화 겹침 표시

4. 발화 겹침 구간의 주발화자와 부발화자의 발화가 모두 알아듣기 어렵고, 화자 구분이 어려워 두 발화자의 음성이 모두 전사가 불가능할 경우

주발화자: 너는 혹시 반려동물을 ((xxx)) 적이 있어? 발화 겹침 표시하지 않음

3.5. 발화 내용 전사

- 발화 내용은 기본적으로 철자 전사를 하되, 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 전사를 한다.

철자 전사: 자 상담소에는 어떤 걸 기대하고 왔을까?

발음 전사: 자 상담소에는 어떤 걸 기대하고 왔으까?

- 각 전사에 사용할 수 있는 문자는 아래와 같다.
 (X를 제외한 알파벳, 비식별화 일련번호를 제외한 숫자, 수식 기호 등 사용 금지)

	발음 전사	철자 전사
사용 가능 문자	. (마침표)	. (마침표)
	? (물음표)	? (물음표)
	~ (담화표지)	. (소수점)
	- (불완전발화)	
	' (모음의 축약형)	
	@ (비식별화, 준음성)	
	(()) (이중괄호)	
사용 불가능 문자	X (잘 들리지 않는 경우)	
	X를 제외한 알파벳	알파벳
	비식별화 일련번호를 제외한 숫자	수식 기호
	수식 기호	

- 발음 전사 시 기호, 외래어 등은 발음에 따라 한글로 적는다.
(기호, 외래어의 철자 전사는 규범 표기를 기준으로 전사하며, 우리말샘을 기준으로 한다.)

철자 전사: 오리지널
발음 전사: 오리지날

철자 전사: 티브이
발음 전사: 티비

철자 전사: 아이유와 컬래버했어
발음 전사: 아이유와 콜라보했어

- 발음 전사 시 모음의 변화, 수의적 경음화 등을 반영하여 전사한다.

철자 전사: 어떡해
발음 전사: 어뜩해

철자 전사: 소주
발음 전사: 씨주

- 발음 전사 시 약화 현상에 의한 이형태는 반영하지 않는다. 예를 들어 의문사 '뭐'가 '머'로 모음이 약화되어 들려도 별도의 발음 전사를 하지 않고 철자 전사인 '뭐'만 적는다.

3.6. 모음의 축약형 표기

- 모음의 축약형의 경우 대부분 현재 국어의 모음 체계상 표기할 글자가 존재하지만, 반홀소리된 /ㄱ/, /ㄴ/의 표기는 문제가 된다. /ㄱ/, /ㄴ/가 반홀소리가 되어 /ㅏ/, /ㅓ/와 축약되는 현상은 구어에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 전사에서는 '를' 사용해서 두 음소를 연결해 준다.

철자 전사: 사귀어
발음 전사: 사귀'어

철자 전사: 바뀌어
발음 전사: 바뀌'어

3.7. 준말과 센말의 전사

- <국립국어원 우리말샘>에 등재된 준말(한 단어 안에서 탈락이나 축약 현상이 일어난 것)과 센말은 철자 전사 시 본딴말로 복원하지 않고 발화된 대로 기재한다.

준말 예)

근데(그런데), 얘기(이야기), 요새(요사이), 요즘(요즈음), 애(아이), 담(다음), 맘(마음), 첨(처음), 널(내일), 켈(제일),
좀(조금), 재밌다(재미있다), 갖다(가지다), -곤(-고는), 뭐(무어), 오랜만(오래간만), 애틀(아무튼), 샘(선생님), 알바(아르바이트), 킬로(킬로그램), 프로(퍼센트) ...

센말 예)

조끔, 쪼끔, 쪼끔(조금), 쫄쫄(졸졸), 땀땀하다(단단하다)

3.8. 끊어진 단어(단어가 불완전하게 발화된 경우)

- 끊어진 단어는 발화된 대로 그대로 전사한다. 불완전하게 발화된 단어(어절)가 둘 이상인 경우에도 어절마다 다음과 같이 표시하여 전사한다.(수정 발화, 반복 발화에 표시하는 것은 아님)

철자 전사: 전 전 전통이라고 우리가 흔히 얘기할 때
발음 전사: -전- -전- 전통이라고 우리가 흔히 얘기할 때

- 내용상 수정 발화와 불완전 발화가 복합적으로 나오는 경우 혹은 수정 발화인지 불완전 발화인지 구분이 모호한 경우에는 어절 앞뒤로 '줄표(-)'를 넣는다.

3.9. 띄어쓰기

- 한글 맞춤법(제5장 띄어쓰기)에 맞게 띄어 쓴다.
- 의존명사는 띄어 쓴다.
- 수를 적을 때는 만 단위로 띄어 쓴다(예: 십이억 삼천백만 팔백구 불 등).
- 본용언과 보조용언도 띄어 쓴다.(예: 먹어 버리다, 가고 싶다, 먹지 못하다)
- 띄어쓰기와 붙여쓰기 모두 허용되는 경우에는 띄어 쓰는 것을 원칙으로 한다.

본용언, 보조용언	막아낸다 vs 막아 낸다(O)
고유명사	건국대학교 vs 건국 대학교(O)
전문용어	성격묘사 vs 성격 묘사(O)
시분초, 연월일 등	두시 vs 두 시(O)
단음절연속	좀더 큰 것 vs 좀 더 큰 것(O)

- 단어를 발음하는 중간에 쉼이 들어간 경우에는 띄어 쓰지 않는다.
- 우리말샘 등재 내용을 기준으로 하되, 판단하기 어려운 경우에는 수시로 논의하여 결정한다.

3.10. 담화 표지

- 머뭇거림의 기능을 하는 1음절 담화 표지 중 “이, 그, 저, 아, 어, 예, 음, 응, 뭐”의 9개 형태에 한해서 본래의 품사와 구별하기 위해 물결표(~)를 붙여 전사한다.
(인제, 이제, 그냥, 무슨, 어떤 등은 붙이지 않음.)
- 억양과 운율에 의해서만 구분이 가능할 경우는 반드시 전사 단계에서 표시해 준다.

철자 전사: 많은 경우에 논문 그 어 연구는 네이션 국가라는 거하구 직결되는 과정이죠.
발음 전사: 많은 경우에 논문 그~ 어~ 연구는 네이션 국가라는 거하구 직결되는 과정이죠.

3.11. 잘 들리지 않는 부분

- 잘 들리지 않는 부분의 전사 시 이중 괄호((xxx))를 이용한다.
(철자 전사에서는 “이중 괄호(()))” 삭제)
- 잘 들리지 않아 추정한 경우는 다음과 같이 전사한다.

철자 전사: 그 전까지는 직장 생활 하느라고 더 힘들어
발음 전사: 그 전까지는 직장 생활 하니라구 ((더 힘들어))

- 화자의 발화 내용이 전혀 들리지 않는 부분은 다음과 같이 전사한다.

철자 전사: 너무나 거 같더라.
발음 전사: (()) 너무나 거 같더라.

- 들리지 않는 음절은 그 음절의 수만큼 x를 붙여 다음과 같이 전사한다.

철자 전사: 근데 그거 진짜 xx해야 되겠더라.
발음 전사: 근데 그거 진짜 ((xx해야)) 되겠더라.

3.12. 준음성과 기타 소리들

- 웃음, 목청 가다듬는 소리, 박수, 노래 등은 다음과 같이 전사한다.

웃음: {laughing}
목청 가다듬는 소리: {clearing}
박수: {applauding}
노래: {singing}

*철자 전사에서는 삭제한다.

3.13. 숫자 전사(상세)

- 숫자의 철자 전사는 이중 전사한다.
- 발음 전사 시 숫자는 발음에 따라 한글로 적는다.
- 철자 전사 시 숫자는 일반적인 표기 관습(숫자, 한글 혼용)에 따라 적는다.

나이

철자 전사: 10살
발음 전사: 열 살

시간

철자 전사: 24시간
발음 전사: 스물네 시간

금액

철자 전사: 2 30 만 원
발음 전사: 이십삼만 원

- 숫자 철자 전사의 띄어쓰기는 “경”, “조”, “만” 단위로 띄어 쓴다.

철자 전사: 1억 2000만 원
발음 전사: 일억 이천만 원

철자 전사: 1조 2345억 6789만 1230원
발음 전사: 일조 이천삼백사십오억 육천칠백팔십구만 일천이백삼십 원

- 철자 전사 시 천 단위 분할 “,”(쉼표)는 쓰지 않는다.

3.14. 방언의 전사

- 방언(발음 전사)에 대한 표준어 대응쌍(철자 전사) 이중 전사
- 우리말샘에 등재된 방언형의 경우 발음 전사는 방언형을 소리 나는 대로 기본 형태를 살려 적고, 철자 전사는 뜻풀이의 표준 어형을 기준으로 삼는다.

철자 전사: 그런데

발음 전사: 근디

*준말의 방언형은 표준어의 본딧말로 통일

철자 전사: 먹었지

발음 전사: 묵었지

- 방언 발음 전사 시 유의 사항은 다음과 같다.

방언과 관련이 없는 표현은 표준어를 적는 방식으로 쓰되, 방언 표현은 방언의 특색이 드러나도록 표기한다. 이때 방언의 표기는 음성 그대로 소리 나는 대로 쓰지 않고 방언의 형태가 드러나는 방식으로 쓴다.

- 방언에서 흔히 나타나는 어두 된소리화의 경우, 방언의 특성으로 볼 수 있으므로 소리 나는 대로 전사하고, 표준어 대응쌍 이중 전사를 한다.

철자 전사: 저번에

발음 전사: 쨌번에

철자 전사: 다르다

발음 전사: 따르다

철자 전사: 계속

발음 전사: 께속

3.15. 비식별화를 위한 전사

- 일상 대화 자료 중 개인정보 등의 비식별화를 위해 이름, 이메일 주소 등 계정 정보, 주민등록번호, 카드 번호, 전화번호 등 각종 번호 및 비밀번호, 상세 주소, 출신 및 소속 등의 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다.
- 정치인 등 유명인의 이름은 비식별화하지 않으며, 상호명 및 상품명 등은 부정적인 경우에만 비식별화한다.
- 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다.
- 비식별화 정보는 아래와 같이 마크업한다.

이름: &name&

상호명: &company-name&

계정(아이디): &account&

주민등록번호: &social-security-num&

전화 번호: &tel-num&

카드 번호: &card-num&

기타 번호: &num&

주소: &address&

출신 및 소속: &affiliation&

기타 비식별화가 필요한 항목: &others&

3.16. 기타 지침

- 발음 전사를 위해 사용한 기호(예: -, {}, &, ())는 철자 전사에는 사용하지 않는다.

개인정보 수집·이용 동의서

(주)스마트미디어테크, (주)나라지식정보, (주)마인즈랩은 국립국어원의 “2022년 일상 대화 말뭉치 구축” 과제에 참여하여 [개인정보보호법] 제15조 및 제17조에 따라 아래의 내용으로 개인정보를 수집·이용합니다. (개인정보 수집·이용 동의에 거부할 수 있으며, 미동의시 과제참여가 불가능합니다)

개인정보 수집·이용자	개인정보 수집·이용 목적	수집·이용 개인정보 항목	보유/이용기간
(주)스마트미디어테크, (주)나라지식정보 (주)마인즈랩	◆ (주)스마트미디어테크 - 일상 대화 말뭉치 구축 과제의 음성 말뭉치 수집 ◆ (주)나라지식정보 - 과제 중 음성데이터 전사, 개인식별정보 등 제거 업무 ◆ (주)마인즈랩 - 과제관리, 최종데이터 검수 업무	발화 음성, 인구통계학적 정보 (출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)	2024년1월 31일까지
(주)스마트미디어테크	과제 참여에 대한 회계 정산 처리 및 비용 증빙	성명, 주민등록번호	2024년 1월 31일까지

귀하는 상기 개인정보 수집·이용에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 위와 같은 개인정보 수집·이용에 동의하십니까?

[☐ 동의합니다 ☐ 동의하지 않습니다.]

(만14세 미만 아동의 경우) 본인은 만14세 미만 아동의 법정대리인으로서 개인정보주체인 아동에 대한 상기 개인정보 수집·이용에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 아동의 개인정보를 수집·이용에 동의합니다.

[☐ 동의합니다 ☐ 동의하지 않습니다.]

◆ 고유식별정보의 처리에 관한 사항

(주)스마트미디어테크는 개인정보보호법에 관한 법률에 따라 회계 정산 처리 신고 목적으로 고유식별정보인 주민등록번호를 처리(수집·이용)하고자 합니다. 보유/이용기간은 2024년 1월 31일까지입니다. 이에 동의하

십니까?

(개인정보 수집 이용 동의에 거부할 수 있으며, 동의하지 않을 경우 과제 참여가 불가능합니다.)

☐ 동의합니다 ☐ 동의하지 않습니다.]

(만14세 미만 아동의 경우) 본인은 만14세 미만 아동의 법정대리인으로서 개인정보주체인 아동에 대한 상
기 고유식별정보 수집·이용에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 아동의
고유식별정보 수집, 이용에 동의합니다.

☐ 동의합니다 ☐ 동의하지 않습니다.]

2022년 월 일

신청인 성명 : _____ (자필서명)

(신청인이 만 14세 미만 아동인 경우) 법정대리인 성명 : _____ (자필서명)

(주)스마트미디어테크, (주)나라지식정보, (주)마인즈랩 귀중

[붙임 3] 개인정보 제3자 제공 동의서

개인정보 제3자 제공 동의서

(주)스마트미디어테크, (주)나라지식정보, (주)마인즈랩은 국립국어원의 “2022년 일상 대화 말뭉치 구축” 과제에 참여하여 [개인정보보호법]에 따라 아래의 내용으로 개인정보를 국립국어원에 제공합니다.(귀하는 개인정보 제3자 제공 동의에 거부할 수 있으며, 미동의시 과제 참여가 불가능합니다.)

개인정보를 제공받는 자	제공받는 목적	제공되는 개인정보 항목	보유/이용기간
국립국어원	<ul style="list-style-type: none"> ◆ 일상 대화 말뭉치 구축 과제의 음성 말뭉치 및 인구통계학적 정보의 기초정보 구분 ◆ 일상 대화 말뭉치 구축 결과물로 언어 연구 및 언어정보 처리분야 응용 기술 개발에 제공 ◆ 국립국어원 시행 타 사업 및 국립국어원 발주 타 용역 사업의 원시데이터로 활용되어 2차적저작물로 가공(외국어, 수어로 번역 가공 포함)될 수 있음 	발화 음성, 인구통계학적 정보(출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)	기본 2043년 12월 31일, 이후 5년 단위 자동 갱신

귀하는 상기 개인정보 제3자 제공에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 위와 같은 개인정보 제3자 제공에 동의하십니까?

☐ 동의합니다 ☐ 동의하지 않습니다.]

(만14세 미만 아동의 경우) 본인은 만14세 미만 아동의 법정대리인으로서 개인정보주체인 아동에 대한 상기 개인정보 제3자 제공에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 해당 아동 개인정보를 국립국어원에 제공하는 것에 동의합니다.

☐ 동의합니다 ☐ 동의하지 않습니다.]

2022년 월 일

신청인 성명 : _____ (자필서명)

(신청인이 만 14세 미만 아동인 경우) 법정대리인 성명 : _____ (자필서명)

(주)스마트미디어테크, (주)나라지식정보, (주)마인즈랩 귀중

국립국어원의 개인정보 제3자 제공(공개) 동의서

본인은 국립국어원의 “2022년 일상 대화 말뭉치 구축” 과제에 참여하여 국립국어원이 [개인정보보호법]에 따라 아래의 내용으로 개인정보를 제3자에 제공(공개)하는데 동의합니다.(귀하는 개인정보 제3자 제공 동의에 거부할 수 있으며, 미동의시 과제 참여가 불가능합니다.)

개인정보를 제공받는 자	제공(공개) 목적	제공되는 개인정보 항목	보유/이용기간
학계·연구기관· 산업체	<ul style="list-style-type: none"> ◆ 일상 대화 말뭉치 구축 결과물로 언어 연구 및 언어정보 처리분야 응용 기술 개발에 제공 ◆ 국립국어원 시행 타 사업 및 국립국어원 발주 타 용역 사업의 원시데이터로 활용되어 2차적저작물로 가공(외국어, 수어로 번역 가공 포함)될 수 있음 	발화 음성, 인구통계학적 정보(출생지/성장 지/거주지/성별/ 연령대/화자간 관계/직업/학력)	기본 2043년 12월 31일, 이후 5년 단위 자동 갱신

귀하는 상기 개인정보 제3자 제공에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 위와 같은 개인정보 제3자 제공 및 공개에 동의하십니까?

☐ 동의합니다 ☐ 동의하지 않습니다.]

(만14세 미만 아동의 경우) 본인은 만14세 미만 아동의 법정대리인으로서 개인정보주체인 아동에 대한 상기 개인정보 제3자 제공에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 해당 아동 개인정보를 제3자에 제공 및 공개하는 것에 동의합니다.

☐ 동의합니다 ☐ 동의하지 않습니다.]

2022년 월 일

신청인 성명 : _____ (자필서명)

(신청인이 만 14세 미만 아동인 경우) 법정대리인 성명 : _____ (자필서명)

국립국어원 귀중

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

저작자 및 저작권 이용 허락자 _____(이하 “권리자”이라 함)와 저작권 이용자 국립국어원(이하 “이용자”이라 함)은 아래 저작물에 관한 저작재산권 이용 허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

제1조 (계약의 목적)

본 계약은 저작재산권 이용 허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)

본 계약의 이용 허락 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

저작물: 일상 대화

저작자:

종별: ☒ 어문저작물

권리: ☒ 복제권, ☒ 공중송신권, ☒ 배포권, ☒ 2차적저작물작성권

※ 저작권 이용 허락 대상 권리의 내용

1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착, 번역(외국어, 수어, 점자, 문자 등) 등)하는 일
3. 국립국어원 시행 사업 및 국립국어원이 발주한 용역 사업의 원시자료로 활용되는 일
4. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 그 복제·변형물·2차적저작물을 학계·연구기관·산업체 등이 연구 및 기술 개발용으로 이용할 수 있도록 제공·배포하는 일
5. 대상저작물 및 그 복제·변형물·2차적저작물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물·2차적저작물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용 허락 기간)

대상저작물의 이용 허락 기간은 계약체결일부터 2043년 12월 31일까지로 하며, 계약기간 만료 시 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 계약기간 만료 시 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)

(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을 이용할 권리를 제3조의 기간 동안 비독점적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하고, 대화 녹음 등 본 계약 이행에 필요한 협조를 하여야 한다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있지 않은 경우 이용자가 요청하면 이용 허락자는 대상저작물의 저작재산권을 등록한 후 위 의무를 이행한다.

(3) 권리자는 대상저작물에 제3자의 이용 허락권, 질권 등 권리 제한 사유 또는 제3자의 권리가 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

(4) 권리자는 대상저작물의 저작재산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 한다.

제5조 (이용자의 권리 및 의무)

(1) 이용자는 대상저작물을 제3조의 이용 허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다.

(2) 이용료는 설정하지 아니한다.

(3) 이용자는 관례적으로 저작자 및 저작재산권자의 성명 등 표시를 허용하는 대상저작물을 이용하는 경우, 그 저작자 및 저작재산권자의 성명 등을 표시하여야 한다.

(4) 이용자는 대상저작물의 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 권리자에게 그 사실을 사전에 고지한 후 사소한 수정 및 편집을 할 수 있다. 특히 권리자는 이용자가 대상저작물 중 개인정보, 프라이버시, 미풍양속, 특정 상품명 등 본 계약 이행에 필요하지 않은 내용은 삭제하고 이용하는 점에 동의한다.

제6조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 대상저작물의 저작권 이용 허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
1. 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
1. 대상저작물에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.

1. 대상저작물에 적용된 이용 허락 조건에 의해서만 대상저작물 재이용을 허락할 것
1. 대상저작물을 권리자 및 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)

본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.

제8조 (계약의 해지)

(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

(2) 당사자는 상대방이 정당한 이유 없이 본 계약을 위반하는 경우에 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부 의사를 표시하였거나 위반 사항의 성격상 시정이 불가능하다는 것이 명백히 인정되는 경우에는 위와 같은 최고 없이 계약을 해지할 수 있다.

(3) 본 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

제9조 (손해배상)

당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상 책임을 면한다.

제10조 (비용의 부담)

계약 체결에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁해결)

(1) 본 계약에서 발생하는 모든 분쟁은 권리와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

(2) 제1항에 따라 해결되지 아니할 때에는 대한민국의 민사소송법 등에 따른 관할법원에서의 소송에 의해 해결토록 한다.

제12조 (비밀유지)

양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용 및 대상저작물의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

제13조 (기타부속합의)

(1) 권리와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (계약의 해석 및 보완)

본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)

본 계약의 효력은 계약 체결일로부터 발생한다.

2022년 월 일

권리자 :

성명

생년월일

주소

이용자 :

(인)

성명 국립국어원장 (인)

주소 서울특별시 강서구 금남화로 154

[붙임 6] 저작권 이용 허락 계약서 미성년자 법정대리인용 동의서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락계약서에 대한 동의서 (미성년자 법정대리인용)

본인은 미성년자의 법정대리인으로 해당 미성년자가 국립국어원의 “2022년 일상 대화 말뭉치 구축” 과제에 참여하여 별첨과 같은 “국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서”를 체결하는 점에 대해 충분히 내용을 검토하였고, 해당 계약서 체결에 동의합니다.

* 별첨 : “국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서”

2022년 월 일

미성년자 성명 :

법정대리인(보호자) : _____ (자필서명)

국립국어원 귀중

<Abstract>

2022 Korean Dialogue Corpus Construction

This project aims to build a dialogue corpus that has been ongoing since 2019, following the 1,000 hours of two-person conversations on 16 topics in 2019, 500 hours of two-person conversations on 15 topics in 2020, and 1,000 hours of multi-person conversations in 2021, and has built 630 hours of daily dialogues. The primary outcomes of this project are as follows:

Speech recording and refinement: Based on Korean demographic distribution, we collected daily and cooperative dialogues from 2,073 speakers varying by region, gender and age. In addition, unlike the existing method of collecting in a controlled indoor environment, about 20% of the uncontrolled real environment speech voices containing noise were included to collect the voices of the real environment. We categorized the collected dialogues into two types: daily dialogues and cooperative dialogues. Regarding daily dialogues, we pre-selected 16 topics that included all existing dialogue topics, and referenced newspaper articles on those topics to help the speakers talk about them. Regarding cooperative dialogues, 10 topics focused on culture and tourism were selected, along with keywords representing pros and cons arguments, newspaper articles, and detailed guides were referenced. Each dialogue consisted of two to four speakers, and the average duration of the dialogue was limited to around 15 minutes. Participants were asked to fill out a corpus license agreement form. The format of the collected and refined audio files is 16 kHz sampling, 16 bit quantization linear PCM.

Speech transcription: Speech transcription and verification was performed by people with experience in transcribing existing dialogue corpus, proofreading experts with more than 20 years of experience. For all transcriptions, primary inspectors reviewed and corrected errors which were

first filtered through semi-automatic natural language verification process. After that, the corrected transcriptions were handed over to secondary inspectors for the final review.

Construction of raw corpus and meta-information: We converted the raw corpus into JSON format according to the guidelines, using the speaker's meta-information and the transcription results. Meta-information consists of information such as the dialogue topic and its form, speaker's gender, age, place of primary residence, and relationships between them.

Keywords: dialogue corpus, raw corpus, cooperative dialogue, real environment voice, voice data transcription

Project Director: Yigyu Hwang(MindsLab)

<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 정주연 학예연구사

국립국어원 오은비 연구원

<사업 참여자>

사업 책임자 황이규 (주식회사 마인즈랩)

사업 참여자 안준환, 현영선, 이원문, 남선웅 (주식회사 마인즈랩)

박영훈, 이지현, 황주영 ((주)나라지식정보)

윤기현, 김민수, 김성진, 이재엽, 김미영 (주식회사 바이칼에이아이)

김태권, 김진호, 정현학, 채창윤 (주식회사 스마트미디어테크)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2023년 1월 31일

발행일: 2023년 1월 31일

인 쇄: 비즈카피

※ 이 책은 국립국어원의 용역비로 수행한 ‘2022년 일상 대화 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.