국립국어원 2021-01-13

발간등록번호
11-1371028-000863-01

2021년 일상 대화 말뭉치 구축

사업 책임자 | 황 이 규

국립국어원

국립국어원 2021-01-13

발간등록번호
11-1371028-000863-01

2021년 일상 대화 말뭉치 구축

사업 책임자
황이규

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2021년 일상 대화 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2021년 06월 ~ 2022년 03월

2022년 3월 23일

사업 책임자: 황 이 규 (㈜마인즈랩)

사업 수행자 ㈜마인즈랩 컨소시엄

사업 책임자 황이규

사업 참여자 박지원, 이원문, 남선웅 박영훈, 이지현, 황주영, 조정아 윤기현, 최성봉, 이현복, 김민석, 박승홍, 김진수, 김용운 김진호, 정현학, 정승현

〈사업 수행자〉 ㈜마인즈랩 컨소시엄

사업 책임자	황이규((주)마인즈랩)
사업 참여자	남선웅((주)마인즈랩)
	박지원((주)마인즈랩)
	이원문((주)마인즈랩)
	박영훈((주)나라지식정보)
	이지현((주)나라지식정보)
	황주영((주)나라지식정보)
	조정아((주)나라지식정보)
	윤기현((주)바이칼에이아이)
	최성봉((주)바이칼에이아이)
	이현복((주)바이칼에이아이)
	김민석((주)바이칼에이아이)
	박승홍((주)바이칼에이아이)
	김진수((주)바이칼에이아이)
	김용운((주)스마트미디어테크)
	김진호((주)스마트미디어테크)
	정현학((주)스마트미디어테크)
	정승현((주)스마트미디어테크)

<국문 초록>

2021년 일상 대화 말뭉치 구축

본 사업의 목표는 일상 대화 말뭉치 구축으로 1,000시간의 다자간 음성 대화를 전사하여 국어학 및 음성 처리, 자연어 처리 연구에 도움이 될 수 있는 대용량 말뭉치 구축에 있다. 이에 따른 주요 사업 내용의 성과는 다음과 같다.

음성 녹음 및 정제: 인구 통계학적 분포를 참고하여, 지역별, 성별, 연령별로 다양한 화자를 모집하였다. 전체 2,599명의 화자를 대상으로 일상 대화 및 협력적 대화 말뭉치를 수집하였다. 수집되는 대화는 일상 대화와 협력적 대화로 구분하였다. 일상 대화는 15개 주제를 사전에 선정하고, 해당 주제에 대한 신문 기사를 참고하여 대화할 수 있도록 하였다. 협력적 대화는 8개의 주제와 이에 따른 찬성과 반대를 대표하는 키워드, 신문 기사 및 공개된 영상을 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 작성하였으며, 코로나 감염을 예방하기 위해 마스크 착용 후 대화를 진행하였다. 수집 및 정제된 음성 파일의 포맷은 16kHz 표본화, 16bit 양자화 선형 PCM이다.

음성 자료 전사: 경험이 풍부한 언어학 전공자, 언어 재활사 및 전문 교정/교열 인력이 음성의 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, 연령, 주요 성장지 등과 대화 상대방과의 관계, 대화 주제와 대화 형식 등의 내용을 포함하고 있다.

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

차 례

제1장 사업 개요

제2장 사업 수행

제3장 사업 수행 결과

<표 차례>

<그림 차례>

제1장

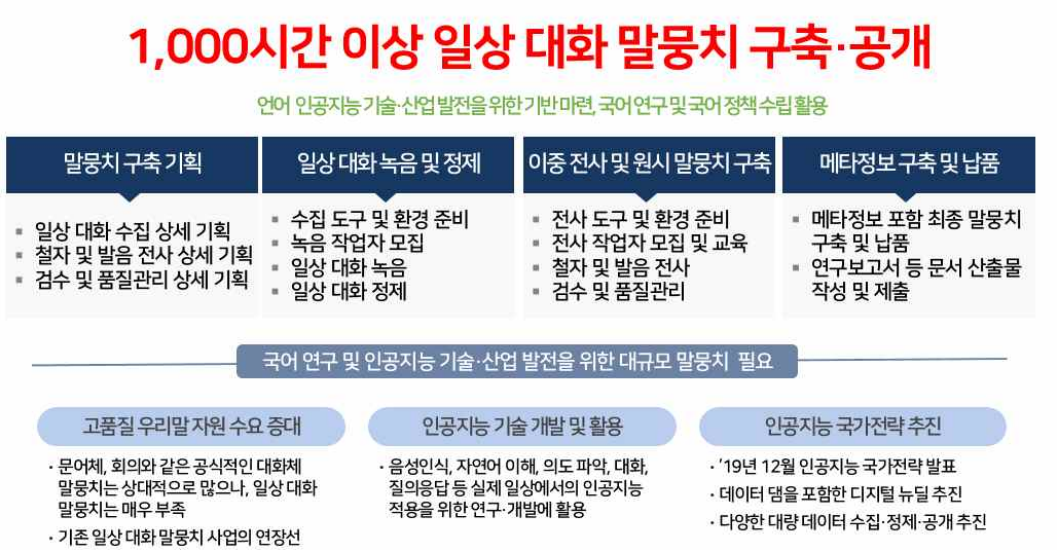
사업 개요

1. 사업 목적

인공지능 산업이 발전됨에 따라 이에 활용할 수 있는 대규모의 고품질 한국어 말뭉치에 대한 자원 수요가 증대되고 있으며, 4차 산업 혁명에 필요한 인공지능 기술을 개발 및 활용하기 위한 대규모의 고품질 우리말 자원 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 위해 중요한 데이터이다.

본 사업의 목적은 대규모 고품질의 일상 대화 말뭉치 구축을 통한 인공지능 분야 R&D와 관련 산업의 활성화에 기여하기 위해 1,000시간 이상의 일상 대화 말뭉치를 구축 및 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 구축에 이어 본 사업에서는 새로운 주제(음악, 쇼핑 등)와 2인 이상 다자 대화, 대화를 통해 결론을 도출해 가는 협력적 대화를 추가로 구축하였다. 지난 2년간 구축된 말뭉치에 이어 다양한 화자를 모집하고 풍부한 주제와 대화 내용을 수집하여 풍부한 주제와 다양한 화자, 대화 내용을 수집하고, 이를 전사 말뭉치로 구축하였다.



[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 품질관리 기준을 수립한다. 둘째, 일상 대화 말뭉치를 수집하기 위한 녹음 환경을 구축하고 대화 데이터를 수집 및 정제한다. 셋째, 녹음된 데이터를 전사 및 검수하기 위한 환경을 준비하고 작업자를 모집하여 교육 및 전사를 진행한 후 말뭉치에 대한 품질을 검수한다. 넷째, 말뭉치에 대화 주제, 화자 정보 등의 메타 정보를 부여하여 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	- 2019/2020년 구축 주제를 고려하여 다양한 주제 선정 - 협력적 대화를 위한 신규 주제 선정	- 일상 대화: 15개 주제 - 협력적 대화: 8개 주제

음성 녹음 및 정제	- 화자별 최대 녹음 시간은 1시간(4개 주제)[1] 로 제한 - 개인정보 비식별화	- 2,599명 화자 참여 - 정제 후 1,000시간 음성 수집
음성 자료 전사	- 발음 전사와 철자 전사를 병행(이중 전사) - 전사 지침에 맞게 전사하고, 전사 후 수작업 전수 검사 실시	- 최종 제출자료 1,000시간 이상

성별, 연령, 직업, 지역 등의 비율이 편중되지 않도록 초기에 선정하였으며, 사업 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 1,000시간, 대화당 12-18분 이내, 평균 15분 이내)
○ 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 날짜 등) 구축

1) 사업 초기 한 화자당 최대 녹음 시간은 30분, 세 명 이상이 대화하는 경우에는 한 화자당 최대 녹음 시간을 1시간으로 제한하였으나, 코로나 확산에 따라 화자 모집이 어려워 한 화자당 최대 1시간으로 변경함

3. 사업 수행 절차

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

[그림 3] 말뭉치 구축 수행 절차

첫 번째 준비 단계에서는 일상 대화 말뭉치를 수집하기 위한 상세 내용을 기획하였다. 수집 방법 및 장소 확보, 대화 주제, 발화자 비율, 일관된 전사를 위한 절차 및 기준 등 말뭉치를 균등한 비율과 동일한 조건에서 수집할 수 있도록 설계하였다.

두 번째로는 발화자를 모집하고, 음성 데이터를 수집 및 정제하여 전사할 수 있도록 준비하였다. 또한 음성 전사 및 개인정보 비식별화가 가능하도록 도구를 개발하고, 전사 기준에 대한 지침서를 마련하여 전사자를 대상으로 교육을 진행하였다. 또한 전사가 완료된 말뭉치는 대화 주제, 발화자 성별, 연령대, 직업 등의 메타 데이터를 생성하였다.

세 번째 단계에서는 가공이 완료된 말뭉치의 품질을 검사하였다. 전사 및 메타 데이터 정보가 부여된 데이터에 대해 전수 검사를 진행하고, 데이터 프로파일링을 통하여 메타 데이터 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 하였

다. 총 7단계[2]의 내부 품질 검증 과정을 통해 고품질의 일상 대화 말뭉치 데이터를 구축 및 공개하는데 문제가 없도록 하였다.

2) 발음 전사 형식 오류, 비식별화 표기 오류, 동일한 철자의 발음 전사 이형태 통계, 동일한 발음의 철자 전사 이형태 통계 등을 검토함

제2장

사업 수행

1. 대화 주제 및 제시 자료 선정

대화의 주제는 일상 대화와 협력적 대화로 구분되어 선정하였다. 일상 대화는 기존 구축 말뭉치 (2019년~ 2020년)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 15가지로 선정하였으며, 대화 주제와 세부 주제 예시를 주고 대화의 세부 주제를 선택할 수 있도록 하였다.

[표 2] 일상 대화 주제 및 세부 예시 주제

주제	세부 예시 주제
휴가	여행 시 교통, 숙박 선택
대중교통	약속 시간, 장소, 교통 선택
음악	대중음악 유행, 선호 가수 및 곡 추천
건강/다이어트	성인병에 대한 상식, 처방, 대응
방송/연예	드라마, 예능 프로그램 선택
스포츠/레저	직접 운동, 관람, 시청 등 참여 방법에 대한 정보 공유 및 결정
먹거리	저녁 모임의 음식 종류와 식당 선택
우정	우정의 가치, 성격, 취미
경제/재테크	집, 주식 등 투자에 대한 고려와 결정
회사/학교	취직, 진학에 대한 정보 공유 및 결정
반려동물	개, 고양이의 장단점 비교 및 입양 결정
취직	대기업/중소기업 취직의 정보 공유와 견해 교환
가족	집안 행사에 대한 검토와 결정
쇼핑	핸드폰 구매 시 기종 검토 및 결정
관혼상제	결혼, 문상, 제사, 축의금, 참석 등

기존 구축 말뭉치 (2019년~ 2020년)의 대화 주제와 2021년 구축 말뭉치를 주제별로 비교하면 다음과 같다.

[표 3] 일상 대화 주제 비교(과거 구축 사업)

2019년	2020년	2021년	세부 주제
군대			
게임			
휴일		휴가	(제주도) 여행 시 교통, 숙박 선택
자동차		대중교통	약속 시간, 장소, 교통 선택
만화			
영화	영화	음악	대중음악 유행, 선호 가수 및 곡 추천
정치			
건강/다이어트	건강/다이어트	건강/다이어트	성인병에 대한 상식, 처방, 대응
방송/연예	방송/연예	방송/연예	드라마, 예능 프로그램 선택
스포츠/레저	스포츠/레저	스포츠/레저	직접 운동, 관람, 시청 등 참여 방법에 대한 정보 공유 및 결정
먹거리	먹거리	먹거리	저녁 모임의 음식 종류와 식당 선택
자연/휴양지			
국가/지역			
문학			
연애/결혼	연애/결혼	우정	우정의 가치, 성격, 취미
경제/재테크		경제/재테크	집, 주식 등 투자에 대한 고려와 결정
	여행지(국내/해외)		
	계절/날씨		
	회사/학교	회사/학교	취직, 진학에 대한 정보 공유 및 결정
	선물		
	꿈(목표)		
	반려동물	반려동물	개, 고양이의 장단점 비교 및 입양 결정
	아르바이트	취직	대기업/중소기업 취직의 정보 공유와 견해 교환
	성격		
	가족	가족	집안 행사에 대한 검토와 결정
		쇼핑	핸드폰 구매 시 기종 검토 및 결정
		관혼상제	결혼, 문상, 제사, 축의금, 참석 등

참여자들에게는 위의 15가지 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	2021년	세부 주제	참고 자료 링크
1	휴가	(제주도) 여행 시 교통, 숙박 선택	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://newhttps://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/다이어트	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/연예	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	직접 운동, 관람, 시청 등 참여 방법에 대한 정보 공유 및 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	우정의 가치, 성격, 취미	https://www.hankookilbo.com/News/Read/A2021063023330000098 https://news.joins.com/article/24005854 http://weekly.chosun.com/client/news/viw.asp?ctcd=c09&nNewsNumb=002655100010
9	경제/재	집, 주식 등	https://www.hankyung.com/realestate/article/202107058533e

	테크	투자에 대한 고려와 결정	https://www.ajunews.com/view/20210705103146300 https://www.fetimes.co.kr/news/articleView.html?idxno=97070
10	회사/학교	취직, 진학에 대한 정보 공유 및 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 입양 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-oo%EC%98%80%EB%8B%A4/ar-AALMsIA
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	집안 행사에 대한 검토와 결정	https://imnews.imbc.com/replay/2021/nwdesk/article/6046050_34936.html https://news.joins.com/article/23882421 https://www.idailynews.co.kr/m/view.php?idx=79046
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/YTKWM46O7NAHDEKEREHWJLFJT4/ https://news.joins.com/article/24097527
15	관혼상제	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

참고한 신문 기사의 상세 내용의 예시는 아래와 같다.

[표 5] 일상 대화 주제별 참고 자료 내용(휴가)

1) <휴가-1: 머니투데이, 백식맞은 직장인 10명 중 6명 "올해 여름휴가 갈 (21.07.06)>

(1) 구인구직 매칭 플랫폼 사람인은 성인남녀 3554명에게 올해 여름휴가 계획을 조사한 결과 계획을 있다는 응답이 64.4%를 차지했다고 6일 밝혔다.

여름 휴가는 다음 달 첫 주에 가장 많이 몰릴 것으로 예상된다. 응답자의 23.6%가 여름휴가기간으로 '8월 1주(2~8일)'에 갈 것이라고 답했다. 이어 △7월 4주(26~8월1)(18.5%) △7·8월 제외 9월 이후(10.5%) △8월 2주(9~15)(8.9%) △8월 3주(16~22)(7.6%) △8월 4주(23~29)(6.1%) 등의 순이었다.

휴가 기간은 평균 3일이었으며, 코로나19(COVID-19) 영향을 국내로 간다는 이들이 96.1%였다. 휴가지로는 '바다 지역'(63.6%, 복수응답)이 가장 많았으나 사람들과 거리를 둘 수 있는 '도심 호캉스'(21.3%), '캠핑'(16.7%), '섬'(15%) 등도 많이 꼽혔다. 휴가비용으로는 평균 60만원을 예상하는 것으로 집계됐다.

여름휴가를 가지 않을 것이라고 응답한 이들 중 직장인(578명)은 '코로나19가 아직 확산세라서'(54.7%, 복수응답), '심적으로 여유가 없어서'(23.2%), '휴가 비용이 없어서'(22.8%), '휴가를 갈 필요성을 못 느껴서'(16.4%), '업무 때문에 휴가를 쓸 수 없어서'(14.5%), '휴가기간이 짧아서'(13.5%), '이직 준비 때문에 바빠서'(10.7%) 등의 이유로 여름휴가를 포기했다.

취업준비생(690명)도 '코로나19 확산세 때문에'(59.6%, 복수응답) 휴가를 가지 않는다는 이들이 가장 많았지만, '취업준비 때문에'(41.4%), '심적으로 여유가 없어서'(40.1%), '휴가 비용이 없어서'(35.9%), '휴가를 별도로 갈 필요가 없어서'(13.6%), '가족한테 눈치 보여서'(7.4%) 등의 이유로 여름휴가를 떠나지 않았다.

전체 응답자 중 10명중 4명(35.9%)은 백신접종 확산으로 인해 '휴가 계획 없다가 국내 휴가를 고려'했다고 답했다. 해외여행도 고려한다는 이들은 7.3%였다. 백신접종 계획에 대해서는 67%가 '곧 백신접종 예정'이라고 답했고, '백신접종을 이미 했다'는 이들은 18.2%였다. 14.8%는 '백신접종을 하지 않을 것'이라고 밝히기도 했다.

[표 6] 일상 대화 주제별 참고 자료 내용(반려동물)

2) <반려동물-2: 매일 경제, 팬데믹과 반려 생활-곁에 있어 줘서 고마워(21.07.06)>

우리 삶 곳곳에 속속들이 영향을 미친 팬데믹 시대. 당연히 반려 생활에도 변화가 따랐다. 동물자유연대가 반려인들을 대상으로 한 설문조사에서 분명히 확인할 수 있다. 응답자들은 한결같이 입을 모은다. 코로나 위기를 통과하는 데 반려동물이 큰 위로가 되었다고.

지난 6월 중순, 한국일보가 동물자유연대와 함께 반려인 320여 명에게 물었다. “반려동물과 함께하는 게 코로나19 극복에 도움이 되나요?” 이 질문에 응답자의 대부분인 91.64%가 “그렇다”고 동의했다. “얼마나 도움이 되나요?”라는 물음에는 10점 만점에 9점을 주었다.

‘코로나19’라는 낯설고 공포스러운 이름이 우리를 덮친 지 1년 반이 돼 간다. 그 어떤 강력한 캐치프레이즈로 불가능했을 삶의 수많은 변화들이 단시간에 큰 저항 없이 이루어졌다. 변화를 간명하게 대변하는 단어들을 떠올려 보자. 록다운, 마스크, 집합 금지, 원격 수업, 재택근무…. 사람들은 서로 거리를 두어야 했고 만남을 보류하고 혼자의 시간을 보내는 다양한 법을 익혀 나갔다.

사람끼리의 거리가 멀어지면서 반대급부로 반려동물과의 거리는 더 가까워지기도 했다. 특히 재택근무가 가져온 변화를 가장 가까이서 또 직접적으로 체감하는 존재가 바로 반려동물 아닐까. 함께 보내는 시간이 늘면서 반려동물은 안정감을 찾고, 반려인은 강제적 고립에서 오는 우울감과 무기력증을 빨리 털어 버릴 수 있었다. 위 인터뷰에서 반려인들은 다음의 이유로 ‘반려동물에게 도움을 받는다’고 했다.

“유일한 외부 활동인 반려견 산책을 하며 우울감과 무력감을 극복할 수 있었습니다. 뚱이 덕분에 작은 것에도 감사할 줄 알게 됐고, 일상에서 느끼는 소소한 행복을 전보다 많이 느끼고 있어요.” “코로나19로 수입이 줄어드는 어려움은 있지만 심바 덕분에 가족들이 한자리에 모이고 함께 이야기하며 웃게 됩니다.” “집에만 갇혀 지내다 보니 직접 대면할 수 있는 생명체는 구슬이가 유일했어요. 구슬이가 다가와 가만히 몸을 맞대거나 눈을 바라보는 것만으로 ‘내 곁엔 네가 있구나’ 하는 생각이 들어 마음의 위안이 됐습니다.” “일을 하지 못해 스스로 볼품없다 생각될 때 저만 바라보는 존재가 있음에 삶에 대한 의욕이 생겼어요.”

이는 비단 우리나라뿐만 아니라 만국 공통의 현실이기도 하다. 미국에서 실시한 조사에서도 반려인의 92%가 팬데믹 이후 반려견의 존재가 정신 건강에 도움이 됐다고 답했다. 코로나19가 닥친 첫해인 2020년 미국에서는 급작스레 늘어난 반려견 입양 트렌드를 이르는 ‘팬데믹 퍼피Pandemic puppy’라는 신조어가 등장하기도 했다. 입양이 느는 한편으로 유기나 파양은 감소했다. 영국의 유기·파양 동물 재입양 기관인 ‘도그 트러스트Dog Trust’의 북아일랜드 밸리미나 보호소는 매년 증가 일로던 보호소 입소견 수가 2020년에는 전년 대비 30%p나 감소했다고 밝혔다. 이 모두가 가리키는 팩트는 하나다. 일상이 멈추고 관계가 단절되었을 때, 사람들은 반려동물에게서 크게 위로를 받는다는 사실이다. 정서적 교감 외에도 신체적인 도움도 받는다. 산책과 식사 챙기기, 씻기기, 놀이하기 등에 더 많은 시간을 할애하면서 움직임이 늘고 규칙적인 생활이 가능해진다.

또한 자유 발화 15개의 주제 외에 8개의 협력적 대화 주제를 추가로 설정하여 제공하였다. 협력적 대화란, 문제 해결 과정에서 문제 해결과 관련된 지식을 공유하는 대화로 주제는 아래와 같다.

[표 7] 협력적 대화 주제

번호	주제
1	공공 공간의 CCTV 설치
2	가짜 뉴스에 대한 징벌적 손해배상
3	원자력 발전소의 존폐
4	지역 내 기피 시설 설치
5	안락사·존엄사 법제화
6	AI의 직업 대체
7	비대면 생활이 미치는 영향
8	청소년에게 인터넷·스마트폰이 미치는 영향

자유로운 일상 대화의 내용 외에도 대화의 주제에 대해 찬성 또는 반대 의견이 있고, 논의를 통해 결론을 도출할 수 있는 말뭉치 구축을 위해 추가로 협력적 대화 주제를 선정하였다. 대화의 원활한 진행을 위해 협력적 대화 주제에 대한 내용을 간략하게 정리하고, 주제와 관련된 키워드를 제공하였다. 협력적 대화에서는 신문 기사와 더불어 주제와 관련된 동영상 링크도 제공하여 활발하게 토론이 진행되도록 하였다.

[표 8] 협력적 대화 주제별 키워드 및 참고 자료 연결

No.	주제	내용
1	공공 공간의 CCTV 설치	- 병원, 학교, 거리 등 다양한 공간에서의 CCTV 설치 - CCTV 설치로 얻을 수 있는 효과와 부작용 - CCTV 설치가 가능한 공간에 대한 기준 - CCTV 영상을 확인할 수 있는 경우와 그 기준 - CCTV 설치 지역 거주 주민들의 동의 확보 방법
		찬성 키워드: 보안, 증거, 범죄예방, 대리수술, 의료사고 반대 키워드: 사생활 침해, 해킹, 불법 촬영, 감시 공통 키워드: 수술실, 어린이집, 설치의무화
	신문 기사 참고 링크	https://www.chosun.com/site/data/html_dir/2020/01/07/2020010701330.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news hani.co.kr/arti/society/society_general/927941.html
	유튜브 참고 링크	https://youtu.be/GHq4RfbLDKM https://youtu.be/sW7NQr7bdM8 https://youtu.be/gh-wb4_ik20
2	가짜 뉴스에 대한 징벌적 손해배상	- 악의적인 가짜뉴스, 오보에 대한 기준 - 징벌적 손해배상 도입 이후 일어날 수 있는 역효과 - 가짜뉴스 예방 및 대응책(개인/사회적 대응) - 가짜뉴스와 그 피해의 인과관계 입증 방법 - 언론의 역할과 언론에게 주어지는 자유의 범위

		찬성 키워드: 피해구제, 언론개혁, 마타도어 반대 키워드: 표현의 자유, 알권리, 언론의 자유 공통 키워드: 유튜브, SNS, 언론중재법
	신문 기사 참고 링크	https://www.chosun.com/site/data/html_dir/2020/04/13/2020041301694.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news https://www.segye.com/newsView/20201027525507
	유튜브 참고 링크	https://youtu.be/jJS0f5x0xH8 https://youtu.be/h1fK9xyEy9s
3	원자력 발전소의 존폐	- 원자력 발전소 유지와 폐지에 대한 입장 - 가정, 일반, 산업용 전기절약 방안 - 각 에너지원의 투자 대비 효율 - 주요국 에너지 정책 현황 및 추진 방향 - 원자력 발전소 설치지역 여론 - 원자력을 대체할 수 있는 대체 에너지
		찬성 키워드: 방사능, 폐기물, 체르노빌, 후쿠시마, 안전성 반대 키워드: 경제성, 탄소중립, 기후변화 공통 키워드: LNG, 재생에너지, 전기요금
	신문 기사 참고 링크	https://www.chosun.com/site/data/html_dir/2020/06/30/2020063001587.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news https://www.segye.com/newsView/20201020521364
	유튜브 참고 링크	https://youtu.be/tX1lCUaZWOQ https://youtu.be/w7IdfL8sl58 https://youtu.be/ZpDerEGxyXI
4	지역 내 기피 시설 설치	- 음식물 폐기시설, 장애인 복지시설과 같은 기피 시설 설치에 대한 반대 - 기피 시설 설치 촉진에 대한 방안 - 기피 시설을 다른 용도로 활용하는 방안 - 증가하는 지역 이기주의로 인한 사회적 문제 - 기피 시설에 대한 지자체와 시민의 의견 차이
		찬성 키워드: 복지, 친환경 에너지, 주민 친화 시설 반대 키워드: 환경오염, 집값 하락, 소음, 건강 공통 키워드: 선거공약, 지하화
	신문 기사 참고 링크	https://www.chosun.com/site/data/html_dir/2020/04/07/2020040700166.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news https://www.segye.com/newsView/20201104519954 https://www.hani.co.kr/arti/area/capital/943672.html
	유튜브 참고 링크	https://youtu.be/8n-Hc2jOdsY https://youtu.be/tccMEX0wxaE
5	안락사·존엄사 법제화	- 안락사·존엄사를 허용하는 올바른 기준 - 소극적 안락사와 적극적 안락사의 차이 - OECD 최고 수준인 우리나라의 자살률

		- 호스피스 병동이 안락사·존엄사를 대신할 수 있는지 - 본인의 의사에 반한 연명치료 중단에 대한 위험성
		찬성 키워드: 개인의 자유, 고통, 불치병, 치료비용 반대 키워드: 살인, 종교, 의학의 발달 공통 키워드: 조력 자살
	신문 기사 참고 링크	https://www.segye.com/newsView/20200225507230 https://www.hani.co.kr/arti/politics/assembly/948669.html
	유튜브 참고 링크	https://youtu.be/hbchjicelSs
6	AI의 직업 대체	- AI가 직업을 대체하는 것에 대한 의견 - 사람과 공존이 가능한 AI 활용방안 - AI로 대체가능한 직업과 불가능한 직업 - AI 발전에 따른 일자리 전환과 재배치를 위한 직업교육 - 기계가 대체할 수 없는 새로운 일자리 창출방안
		찬성 키워드: 편리함, 개발자, 기피 직종 반대 키워드: 구직자, 노령자, 단순노무직 공통 키워드: 디지털트윈[3], 자동화
	신문 기사 참고 링크	https://www.segye.com/newsView/20200601514707 https://www.hani.co.kr/arti/economy/marketing/974709.html
	유튜브 참고 링크	https://youtu.be/LPSU46gRbd0 https://youtu.be/_euD2UWholw https://youtu.be/xaIxMvLPqRc
7	비대면 생활이 미치는 영향	- 모임, 업무, 수업, 여가활동 등 다양한 비대면 생활의 효과 - 비대면 생활의 바람직한 방향 - 비대면이 효율적인 영역과 비효율적인 영역 - 사회적인 격차에 따른 비대면 생활의 차이 - 비대면 수업으로 인한 학습격차
		찬성 키워드: 수도권 집중, 공간의 자유, 교통체증, 신산업 발달 반대 키워드: 대인관계, 양극화, 아동, 교육, 사교성 공통 키워드: 코로나19
	신문 기사 참고 링크	https://www.chosun.com/site/data/html_dir/2020/05/25/2020052502666.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news https://www.segye.com/newsView/20200511517330 https://www.hani.co.kr/arti/society/society_general/935035.html
	유튜브 참고 링크	https://www.youtube.com/watch?v=jImPokx-Tas https://www.youtube.com/watch?v=Lt1fCHaGE08
8	청소년에게 인터넷·스마트폰이 미치는 영향	- 인터넷·스마트폰이 청소년에게 미치는 영향 및 보호대책 - 청소년의 적절한 인터넷·스마트폰 사용시간 - 청소년들이 인터넷·스마트폰을 제한해야 하는 경우 - 청소년 오프라인 활동 확대방안 - 청소년 인터넷·스마트폰 사용이 늘어나는 이유
		찬성 키워드: 정보습득, 기술의 발전, 소외감 반대 키워드: 중독, 시력 저하, 수면부족, 게임, 부적절한 정보의 습득 공통 키워드: 셧다운, 학습, 디지털 네이티브[4]

	신문 기사 참고 링크	https://www.chosun.com/site/data/html_dir/2020/05/13/2020051300483.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news https://www.segye.com/newsView/20201014507583
	유튜브 참고 링크	https://youtu.be/3ubU-sqX6KQ https://youtu.be/rwl7UF_H_NI https://youtu.be/BD0ZtUSj7As

3) 컴퓨터에 현실 속 사물의 쌍둥이를 만들고, 현실에서 발생할 수 있는 상황을 컴퓨터로 시뮬레이션함으로써 결과를 미리 예측하는 기술

4) 개인용 컴퓨터, 휴대전화, 인터넷, MP3와 같은 디지털 환경을 태어나면서부터 생활처럼 사용하는 세대를 말함

2. 화자 구성 및 모집

발화자 모집은 설계 단계에서는 모집 인원 총 5,500명으로 2020년 통계청 지역별 인구 분포 자료를 참고하여 전국 16개 지역으로 나누어 모집하였다. 성별, 연령별, 지역별(주 성장지 기준), 주제별로 비율이 편중되지 않도록 하였다. 성별, 연령별, 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, 9개 도(서울, 인천, 대전, 대구, 부산, 광주, 울산, 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 연령대는 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 9] 사업 초기 화자 할당표 설계 기준

구분	기준
모집단	• 2020년 통계청 지역별 인구 분포 자료 기준
고려 변수	• 성별: 남자/여자 • 연령대: 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	• 지역별: 비례 할당 • 성별×연령별: 균등 할당

초기 설계된 성별 및 연령대별 지역별 모집 목표는 아래와 같았다.

[표 10] 성별 및 연령대별 지역별 모집 목표(단위: 명)

		연령	10대		20대		30대		40대		50대		60대 이상		총수집 인원
		성비	남자	여자	남자	여자	남자	여자	남자	여자	남자	여자	남자	여자	
	인구 (천명)	비율	5%	5%	10%	10%	10%	10%	10%	10%	10%	10%	5%	5%	100%
계	51,683,025	100.0%	275	275	550	550	550	550	550	550	550	550	275	275	5,500
서울	9,575,355	18.5%	51	51	102	102	102	102	102	102	102	102	51	51	1020
부산	3,367,334	6.5%	18	18	36	36	36	36	36	36	36	36	18	18	360
대구	2,402,940	4.6%	13	13	26	26	26	26	26	26	26	26	13	13	260
인천	2,936,382	5.7%	16	16	31	31	31	31	31	31	31	31	16	16	312
광주	1,443,154	2.8%	8	8	15	15	15	15	15	15	15	15	8	8	152
대전	1,457,161	2.8%	8	8	16	16	16	16	16	16	16	16	8	8	160
울산	1,127,175	2.2%	6	6	12	12	12	12	12	12	12	12	6	6	120
세종	362,036	0.7%	2	2	4	4	4	4	4	4	4	4	2	2	40
경기	13,488,910	26.1%	71	71	143	143	143	143	143	143	143	143	71	71	1,428
강원	1,535,491	3.0%	7	7	16	16	16	16	16	16	16	16	8	8	158
충북	1,596,955	3.1%	8	8	17	17	17	17	17	17	17	17	8	8	168
충남	2,117,260	4.1%	11	11	23	23	23	23	23	23	23	23	11	11	228
전북	1,794,682	3.5%	10	10	19	19	19	19	19	19	19	19	10	10	192
전남	1,842,423	3.6%	10	10	20	20	20	20	20	20	20	20	10	10	200
경북	2,633,592	5.1%	14	14	28	28	28	28	28	28	28	28	14	14	280
경남	3,327,298	6.4%	18	18	35	35	35	35	35	35	35	35	18	18	352
제주	674,877	1.3%	4	4	7	7	7	7	7	7	7	7	3	3	70

데이터 수집이 진행되면서 전세계적으로 전염병이 확산되어 화자의 모집이 어려워져 2차례에 걸쳐 데이터 수집 계획을 변경하여 최종적으로 아래와 같은 기준이 추가되었다.

- 지역 단위는 광역자치단체에서 광역권(예를 들어, 서울/경기/인천, 부산/울산/경남 등의 광역권별로 모집 계획 변경)
- 연령별 수집 구성은 권역별로 초기 계획의 50% 이상 수집한다.
- 성별 비율은 초기 남녀 50:50에서 한쪽 성별이 모집 계획의 20% 이하가 되지 않도록 한다. 다만, 40-60대 남성은 광역권 별로 초기 모집 계획의 10% 이상을 수집한

다.[5)]

- 초기 2인 대화 80%, 3인 대화 15%, 4인 대화 5% 모집 계획에서 2인 대화 90% 이내, 3-4인 대화 10% 이내로 수집한다.

변경된 모집 목표를 감안하여, 최종적으로 모집된 인원의 숫자와 비율은 각각 [표 11]과 [표 12]와 같다. 예를 들어, 남성/10대/인천의 최종 모집 인원은 초기 목표(16명) 대비 약 0.5배(8명) 모집되었다.

[표 11] 성별 및 연령대별 지역별 모집 결과(단위: 명)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	22	81	38	24	20	6	31	133	59	70	71	31	586	1,431
	인천	8	25	17	2	1	1	10	27	16	11	6	2	126	
	경기	32	167	46	5	4	4	45	241	80	46	35	14	719	
영남권	부산	2	21	8	5	2	3	19	38	15	11	11	3	138	565
	경남	1	20	11	1	1	1	14	35	16	6	5	1	112	
	울산	0	12	3	1	1	0	11	21	5	1	3	1	59	
	대구	14	25	6	2	2	1	19	27	23	11	9	1	140	
	경북	9	25	9	2	2	1	16	26	7	6	10	3	116	
호남권	광주	6	15	12	1	1	0	6	10	11	5	9	1	77	266
	전북	5	19	6	4	3	0	2	20	11	9	14	6	99	
	전남	2	19	3	1	1	1	8	14	11	13	13	4	90	
충청권	대전	12	18	12	3	1	1	9	16	16	9	3	1	101	242
	충북	4	10	5	2	1	0	4	19	7	4	6	2	64	
	충남	4	15	4	2	1	2	7	16	12	7	5	2	77	
강원권	강원	1	4	2	1	6	1	1	16	11	9	8	5	65	65
제주권	제주	1	6	1	0	1	1	2	9	3	4	1	1	30	30
합계		123	482	183	56	48	23	204	668	303	222	209	78	2,599	2,599

5) 40-60대 남성은 직장인들이 많아, 전염병 확산에 대한 우려로 대화 참여에 매우 소극적인 성향을 보였음

[표 12] 성별 및 연령대별 지역별 기존 목표 대비 모집 비율

단위: 비율(%)		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	43.1	79.4	37.3	23.5	19.6	11.8	60.8	130.4	57.8	68.6	69.6	60.8	57.5	51.9
	인천	50.0	80.7	54.8	6.5	3.2	6.3	62.5	87.1	51.6	35.5	19.4	12.5	40.4	
	경기	45.1	116.8	32.2	3.5	2.8	5.6	63.4	168.5	55.9	32.2	24.5	19.7	50.4	
영남권	부산	11.1	58.3	22.2	13.9	5.6	16.7	105.6	105.6	41.7	30.6	30.6	16.7	38.3	41.2
	경남	5.6	57.1	31.4	2.9	2.9	5.6	77.8	100.0	45.7	17.1	14.3	5.6	31.8	
	울산	0.0	100.0	25.0	8.3	8.3	0.0	183.3	175.0	41.7	8.3	25.0	16.7	49.2	
	대구	107.7	96.2	23.1	7.7	7.7	7.7	146.2	103.9	88.5	42.3	34.6	7.7	53.9	
	경북	64.3	89.3	32.1	7.1	7.1	7.1	114.3	92.9	25.0	21.4	35.7	21.4	41.4	
호남권	광주	75.0	100.0	80.0	6.7	6.7	0.0	75.0	66.7	73.3	33.3	60.0	12.5	50.7	48.9
	전북	50.0	100.0	31.6	21.1	15.8	0.0	20.0	105.3	57.9	47.4	73.7	60.0	51.6	
	전남	20.0	95.0	15.0	5.0	5.0	10.0	80.0	70.0	55.0	65.0	65.0	40.0	45.0	
충청권	대전	120.0	90.0	60.0	15.0	5.0	10.0	90.0	80.0	80.0	45.0	15.0	10.0	50.5	40.6
	충북	50.0	58.8	29.4	11.8	5.9	0.0	50.0	111.8	41.2	23.5	35.3	25.0	38.1	
	충남	36.4	65.2	17.4	8.7	4.4	18.2	63.6	69.6	52.2	30.4	21.7	18.2	33.8	
강원권	강원	14.3	25.0	12.5	6.3	37.5	12.5	14.3	100.0	68.8	56.3	50.0	62.5	41.1	41.1
제주권	제주	25.0	85.7	14.3	0.0	14.3	33.3	50.0	128.6	42.9	57.1	14.3	33.3	42.9	42.9
합계		44.7	87.6	33.3	10.2	8.7	8.4	74.2	121.5	55.1	40.4	38.0	28.4	47.3	47.3

화자 모집은 최소 2인 1조 이상 신청자를 기본으로 하였으며, 모집이 어려운 40대 이상의 남성을 최우선으로 하였다. 1인이 개별 신청했을 경우에는 서로 모르는 사람과 대화하는 것은 자연스러운 대화가 나오지 않을 것으로 판단하고 지인과 함께 녹음할 수 있게 유도하였다. 이렇게 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.

녹음이 진행될수록 성별x연령별x지역별 할당 외에 주제 할당까지 맞는 화자를 찾는 것이 쉽지 않았다. 일부 화자는 녹음 전 확인 전화 시 녹음 일정을 일방적으로 취소하거나 연락이 두절되기도 하였으며, 연락이 되었다 하더라도 녹음 당일 녹음 장소에 오지 않기도 하였다.

[그림 4] 말뭉치 구축 사업 참여자 모집 공고

3. 작업자 선발 및 교육

3.1. 녹음 진행 요원 선발 및 교육

녹음은 서울, 인천, 수원, 대전, 대구, 울산, 부산, 광주, 익산, 강원 총 10개 지역에서 순차적으로 진행하였다. 총 17개의 지역 거주자 2,599명의 화자가 녹음에 참여하였으며, 다수의 화자가 참여하는 만큼 원활한 녹음 진행을 위하여 각 지역별로 녹음 진행 요원이 투입되었다. 지역별로 투입된 진행 요원은 11명으로 코로나-19의 상황을 고려하여 온라인으로 면담을 진행한 후 자격 조건과 맞는 인원을 오프라인 교육을 통해 최종 선발하였다. 진행 요원은 기존 유사 작업 경험자를 우선으로 선발하였다.

[표 13] 진행 요원 선발 및 운영 방안

구분	선발 기준 및 운영 내용	
선발 기준	• 전문 녹음 장비 작동 경험이 있는 자 (우선 선발) • 최종 교육 이수 및 평가 통과자 (필수)	
투입 인원	• 진행 요원 11명	
진행 요원 역할	• 진행 요원 1 - 화자 안내 - 코로나-19에 따른 건강 체크 - 화자 인적사항 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.

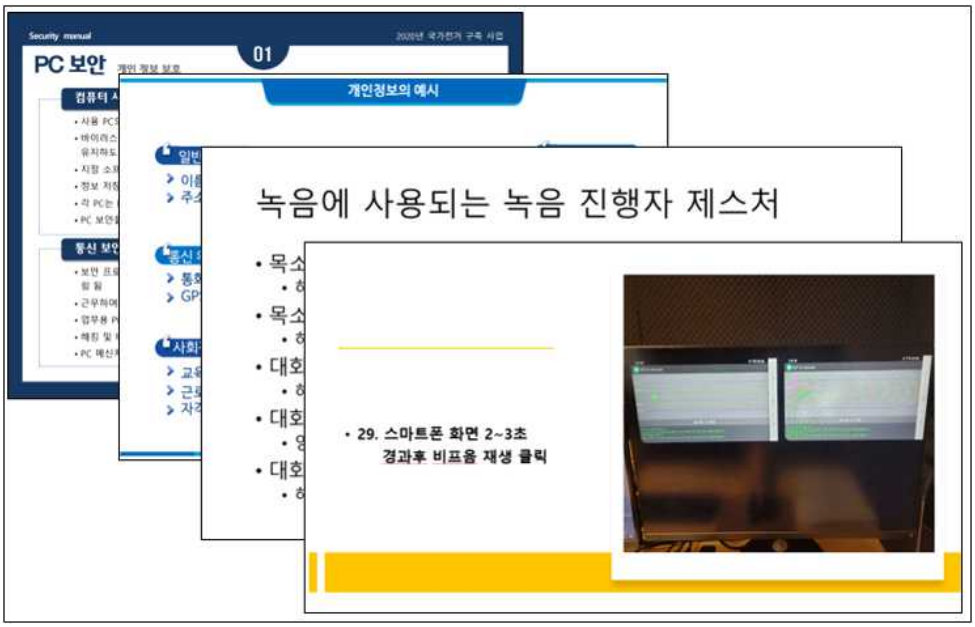
- 사업 배경 및 목적, 진행 시 유의 사항 등의 이론 교육
- 녹음 장비 작동 방법, 헤드셋 마이크 착용 방법, 녹음 진행 등의 실사 교육
- 화자 응대, 화자의 불만 제기 시 대처 방법 등의 CS 교육
- 화자 개인정보 관리, 녹음 자료 관리 등의 보안 교육

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

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

[표 14] 진행 요원 교육 내용

구분	내용
교육 일시 및 장소	• 2021년 07월, 서울 녹음 사이트 • 2021년 08월, 수원, 인천 녹음 사이트 • 2021년 09월, 부산, 익산 녹음 사이트 • 2021년 11월, 대구, 울산, 광주, 대전 녹음 사이트
교육자	• 김용운 ㈜스마트미디어테크 • 김진호 ㈜스마트미디어테크
교육 내용	• 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 5] 녹음 진행 요원 교육 자료 일부

3.2. 전사자 선발 및 교육

전사작업을 위한 인력은 7개월간 42명의 작업자가 투입되었으며, 20년 이상 경력의 교정 교열 전문가 10명, 언어 재활사 3명, 언어계열 전공자 10명(석사 이상 학위 취득자 6명 포함)와 5년 이상 도서관 원문, 전거 및 서지 목록 구축사업 유경력자(19명)로 구성하였다.

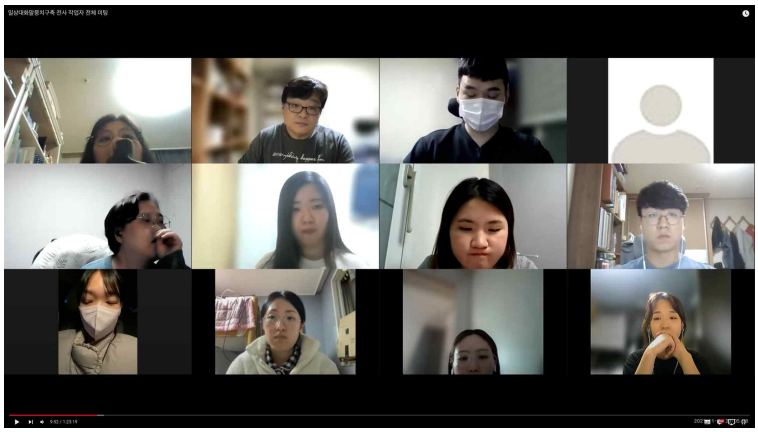
전사 인력은 전사 지침 교육, 전사 도구 활용교육과 전사 단위(억양구)개념 교육 후 2주간의 테스트 전사 기간을 거쳐 전사 작업에 투입되었으며, 비대면 재택근무를 통해 작업을 진행하였다.

[표 15] 전사자 선발 기준 및 운영

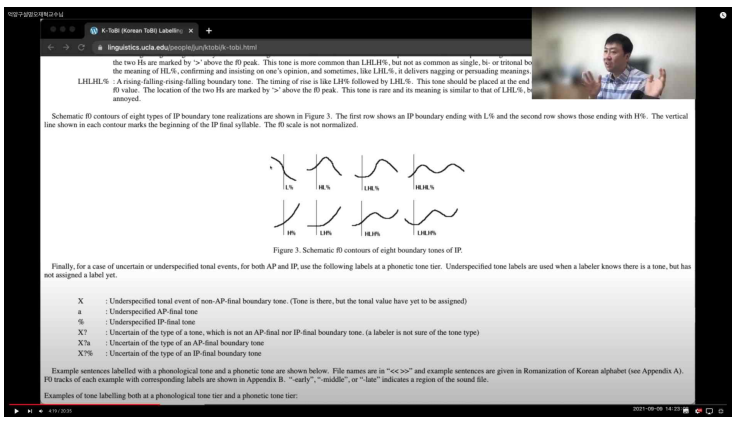
구분	선발 기준 및 운영 내용
선발 기준	• 교정 교열 전문작업자 • 언어 재활 및 언어 교정 교육과정 이수자 • 언어계열 전공자 • 유사 작업 경험자 및 관련 분야 전공자
월별 투입인력	• 2021년 9월: 12명 • 2021년 10월: 14명 • 2021년 11월: 14명 • 2021년 12월: 20명 • 2022년 1월: 35명 • 2022년 2월: 32명 • 2022년 3월: 24명
운영	• 음성 파일 중 방언 등의 특징이 있을 경우 해당 지역 출신 전사자에게 우선 배정함. • 전사자별 선호 주제를 정해 해당 주제 관련 음성 대화 우선 배정함. • 음성 발화자와 유사한 연령대의 전사자에 관련 음성 대화 우선 배정함.

[표 16] 전사자 교육

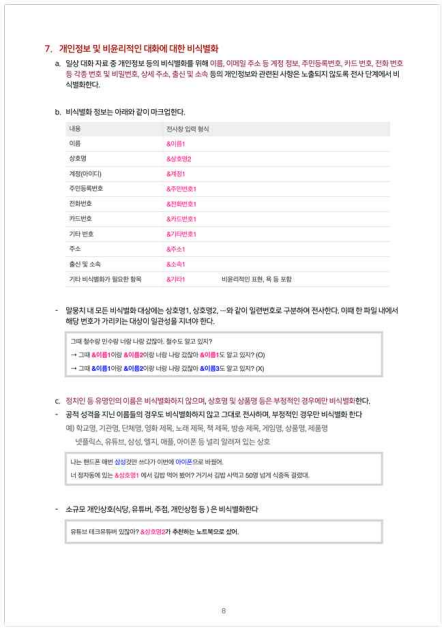
구분	내용
정기교육	• 2021년 9월 14일부터 매월 첫 주 화요일(온라인)
교육자	• 박영훈(전체 교육진행 및 공지, 나라지식정보) • 이지현(전사 지침, 유의사항 및 맞춤법 교육, 나라지식정보) • 김민석(전사 도구 사용 교육, 바이칼AI) • 오재혁(억양구 관련, 건국대학교) - 동영상 녹화 후 시청
교육 내용	• 사업의 배경 및 목적, 전사 절차와 방법 • 전사 지침 및 유의 사항 • 전사 도구 사용 교육 • 한글 맞춤법 주요 내용 • 질의응답



[그림 6] 전사자 교육(온라인)



[그림 7] 억양구 기준 관련 온라인 교육



7. 개인정보 및 비윤리적인 대화에 대한 비식별화

a. 일상 대화 자료 중 개인정보 등의 비식별화를 위해 이름, 이메일 주소 등 계정 정보, 주민등록번호, 카드 번호, 전화 번호 등 각종 번호 및 비밀번호, 상세 주소, 출신 및 소속 등의 개인정보와 관련된 시점은 노출되지 않도록 전사 단계에서 비식별화한다.

b. 비식별화 정보는 아래와 같이 마크업한다.

내용	전사할 입력 형식	
이름	&이름1	
상호명	&상호명2	
계정(아이디)	&계정1	
주민등록번호	&주민번호1	
전화번호	&전화번호1	
카드번호	&카드번호1	
기타 번호	&기타번호1	
주소	&주소1	
출신 및 소속	&소속1	
기타 비식별화가 필요한 항목	&기타1	비윤리적인 표현, 욕 등 포함

- 말뭉치 내 모든 비식별화 대상에는 상호명1, 상호명2, ···와 같이 일련번호로 구분하여 전사한다. 이때 한 파일 내에서 해당 번호가 겨리키는 대상이 일관성을 지녀야 한다.

그때 철수랑 민수랑 나랑 갔잖아. 철수도 알고 있지?
→ 그때 &이름1이랑 &이름2이랑 나랑 갔잖아 &이름1도 알고 있지? (O)
→ 그때 &이름1이랑 &이름2이랑 나랑 갔잖아 &이름3도 알고 있지? (X)

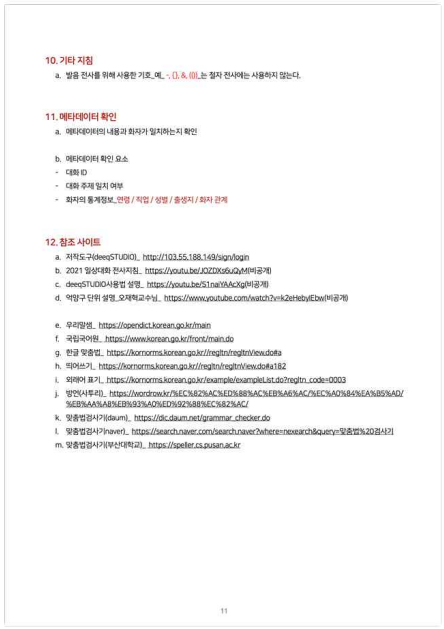
c. 정치인 등 유명인의 이름은 비식별화하지 않으며, 상호명 및 상품명 등은 부정적인 경우에만 비식별화한다.
- 공적 성격을 지닌 이름들의 경우도 비식별화하지 않고 그대로 전사하며, 부정적인 경우만 비식별화 한다
예) 학교명, 기관명, 단체명, 영화 제목, 노래 제목, 책 제목, 방송 제목, 게임명, 상품명, 제품명
'넷플릭스, 유튜브, 삼성, 엘지, 애플, 아이폰 등 널리 알려져 있는 상호

나는 핸드폰 매번 삼성갤럭시 쓰다가 이번에 아이폰으로 바꿨어.
너 정치동에 있는 &상호명1 에서 김밥 먹어 봤어? 거기서 김밥 사먹고 50명 넘게 식중독 걸렸대.

- 소규모 개인상호(식당, 유튜버, 주점, 개인상점 등) 은 비식별화한다

유튜브 테크유튜버 있잖아? &상호명2가 추천하는 노트북으로 샀어.

8



10. 기타 지침

a. 발음 전사를 위해 사용한 기호_예_ -, (), &, (()).는 철자 전사에는 사용하지 않는다.

11. 메타데이터 확인

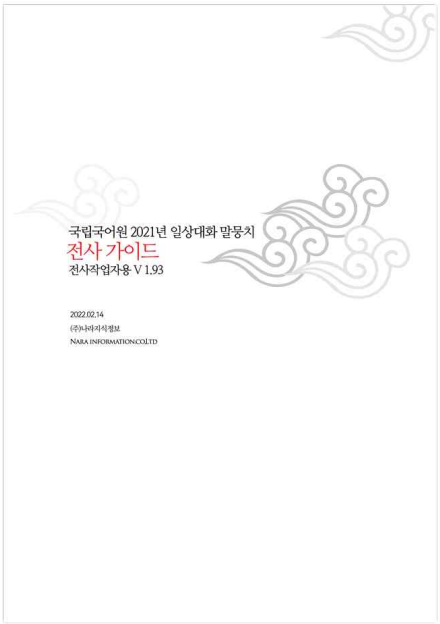
a. 메타데이터의 내용과 화자가 일치하는지 확인

b. 메타데이터 확인 요소
- 대화 ID
- 대화 주제 일치 여부
- 화자의 통계정보_연령 / 직업 / 성별 / 출생지 / 화자 관계

12. 참조 사이트

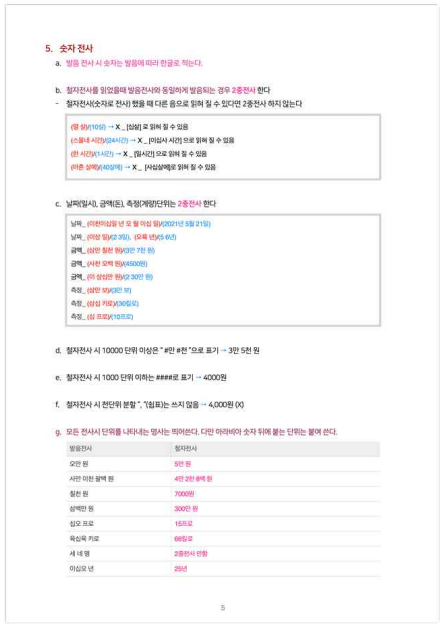
a. 저작도구(deepSTUDIO)_ http://103.55.188.149/sign/login
b. 2021 일상대화 전사지침_ https://youtu.be/JOZDXs6uQyM(비공개)
c. deepSTUDIO사용법 설명_ https://youtu.be/51naiYAAcXg(비공개)
d. 억양구 단위 설명_오재혁교수님_ https://www.youtube.com/watch?v=k2eHebytEbw(비공개)

e. 우리말샘_ https://opendict.korean.go.kr/main
f. 국립국어원_ https://www.korean.go.kr/front/main.do
g. 한글 맞춤법_ https://kornorms.korean.go.kr//regltn/regltnView.do#a
h. 띄어쓰기_ https://kornorms.korean.go.kr//regltn/regltnView.do#a182
i. 외래어 표기_ https://kornorms.korean.go.kr/example/exampleList.do?regltn_code=0003
j. 맞먹(사투리)_ https://wordrow.kr/%EC%82%AC%EB%88%AC%EB%A6%AC/%EC%A0%84%EA%B5%AD/%EB%AA%A8%EB%93%A0%ED%92%88%EC%82%AC/
k. 맞춤법검사기(daum)_ https://dic.daum.net/grammar_checker.do
l. 맞춤법검사기(naver)_ https://search.naver.com/search.naver?where=nexearch&query=맞춤법%2D검사기
m. 맞춤법검사기(부산대학교)_ https://speller.cs.pusan.ac.kr



국립국어원 2021년 일상대화 말뭉치
전사 가이드
전사작업자용 V 1.93

2022.02.14
(주)나라지식정보
NARA INFORMATION CO.LTD



5. 숫자 전사

a. 발음 전사 시 숫자는 발음에 따라 한글로 적는다.

b. 철자전사를 읽었을때 발음전사와 동일하게 발음되는 경우 2중전사 한다
- 철자전사(숫자로 전사) 했을 때 다른 음으로 읽혀 질 수 있다면 2중전사 하지 않는다

(첫 삽)(10삽) → X _ [십삽] 로 읽혀 질 수 있음
(스물네 시간)(24시간) → X _ [이십사 시간] 으로 읽혀 질 수 있음
(한 시간)(1시간) → X _ [일시간] 으로 읽혀 질 수 있음
(마흔 살째)(40살째) → X _ [사십살째]로 읽혀 질 수 있음

c. 날짜(일시), 금액(돈), 측정(계량)단위는 2중전사 한다

날짜_ (이천이십일 년 오 월 이십 일)(2021년 5월 21일)
날짜_ (이십 일)(2 3일), (오륙 년)(5 6년)
금액_ (삼만 칠천 원)(3만 7천 원)
금액_ (사천 오백 원)(4500원)
금액_ (이 십삼만 원)(2 30만 원)
측정_ (삼만 보)(3만 보)
측정_ (삼십 키로)(30킬로)
측정_ (십 프로)(10프로)

d. 철자전사 시 10000 단위 이상은 " #만 #천 "으로 표기 → 3만 5천 원

e. 철자전사 시 1000 단위 이하는 ####로 표기 → 4000원

f. 철자전사 시 천단위 분할 ", "(쉼표)는 쓰지 않음 → 4,000원 (X)

g. 모든 전사시 단위를 나타내는 명사는 띄어쓴다. 다만 아라비아 숫자 뒤에 붙는 단위는 붙여 쓴다.

발음전사	철자전사
오만 원	5만 원
사만 이천 삼백 원	4만 2만 3백 원
칠천 원	7000원
삼백만 원	300만 원
십오 프로	15프로
육십육 키로	66킬로
세 네 명	2중전사 안함
이십오 년	25년

5

[그림 8] 전사 교육 자료(일부)

3.3. 개인정보 보호 및 보안 교육

사업의 원활한 진행을 위해 개인정보 취급 및 데이터 관리와 관련하여 개인정보보호법과 개인정보의 취급에 따른 보안 교육을 진행하였다. 교육 참여자는 사업 관리, 녹음 진행요원 및 전사 도구 개발자, 전사 작업 관리자 등 사업 참여자를 대상으로 하였다. 개인정보보호위원회가 운영하는 개인정보보호 포털(privacy.go.kr)의 온라인 교육인 '(신) 개인정보보호법 이해하기(2021년도)'를 참여자들이 수강하고 교육 수료증을 발급받았다.

개인정보보호와 관련된 주요 내용은 개인정보의 범위, 개인정보의 저장/파기/양도/유출 등과 관련된 내용, 개인정보취급자의 의무/범위와 개인정보처리방침 등이 주를 이루었다. 보안 교육은 사업 참여자를 대상으로 온라인으로 자체 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 국립국어원의 정보보안업무 처리규정 등을 참고하여 교육을 진행하였다.

[표 17] 개인정보 보호 및 보안 관련 교육

구분	내용
보안 교육	• 2021년 7월 22일 16:00-17:00 (온라인 집체 교육) • 참여자: 사업 참여자(17명) • 내용: 사업 수행과정에서 취득, 생산 및 유통되는 데이터 및 사용 장비에 발생할 수 있는 위험/보안 요소에 대한 유의 사항 교육 • 참고 자료: 국가사이버안전관리규정, 문화체육관광부 보안업무규정 시행세칙, 국립국어원 정보보안업무 처리규정(안) 등
개인정보보호 교육	• 2021년 11월 ~ 12월 개별적 온라인 교육 (privacy.go.kr) • 참여자: 사업 참여자(17명) • 내용: 개인정보보호법 교육

4. 음성 녹음

4.1. 녹음 환경

녹음은 전국 10개 지역(서울, 인천, 수원, 대전, 대구, 울산, 부산, 광주, 익산, 강원)에서 인구 비율에 따라 최소 0.5개월에서 7개월까지 화자를 모집하여 수집하였다. 외부와 차단된 상태로 녹음에 참여하는 화자만이 대화할 수 있도록 구성하였고, 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 9] 대화 수집을 위한 녹음 장비 및 환경

폐쇄된 공간에서 2인 이상의 대화 내용을 수집해야 하므로 코로나-19를 방지할 수 있도록 방역 수칙을 적용하고 소독 및 발열 체크와 마스크 등 물품을 비치하였다. 또한 코

로나-19 확진자가 발생할 수 있는 상황을 우려하여 역학 조사에 동참할 수 있도록 데이터 수집에 참여하는 모든 발화자들에게 개인정보 수집을 위한 동의서를 작성하도록 하였다. 발화자별 녹음 시간대를 조정하여 최대한 많은 인원이 부딪히지 않도록 하였고 모집 및 교육은 비대면으로 진행하는 등 코로나-19 감염에 유의하였다.

[표 18] 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

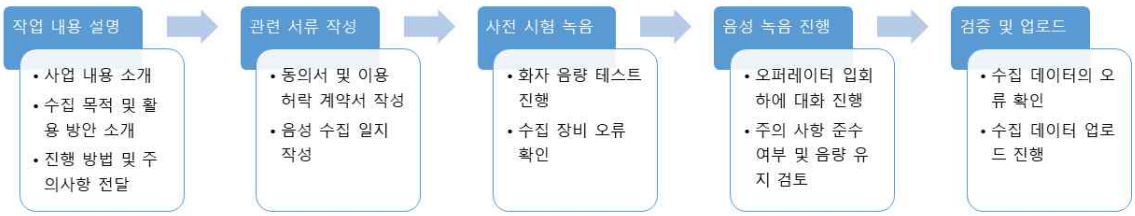
[그림 10] 녹음 시작 전 녹음 장소 방역 진행

수집 장비는 수집 기관의 과제 수행 경험을 통해 검증된 장비인 Focusrite의 Scarlett solo 18i8 오디오 인터페이스 및 Shure사의 SM35-TQC Closetalk Mic로 운용하였다. 녹음 음성은 16khz로 설정하였으며, 이외에 추가로 제안한 스마트폰 음성 수집을 위해 화자당 1대씩 스마트폰 거치대를 설치하여 음성을 수집하였다.

[그림 11] 녹음 장비 및 장비 테스트

4.2. 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

4.2.1. 작업 내용 설명

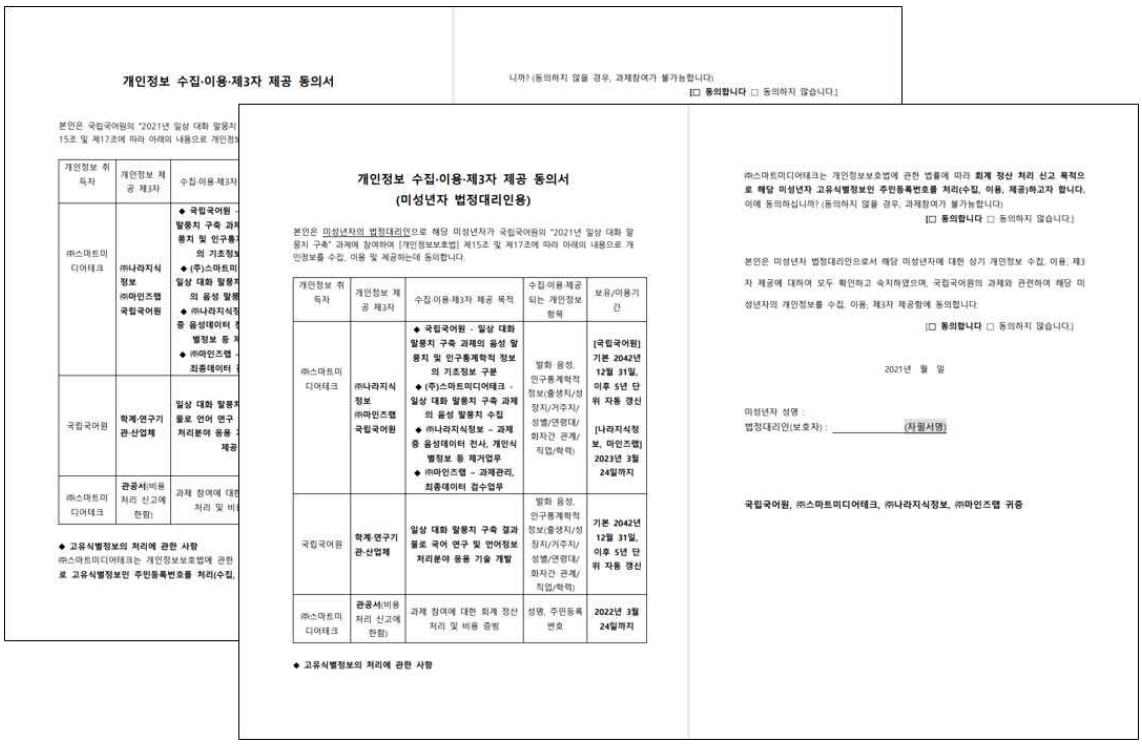
화자의 모집 시 간략하게 설명된 사업의 내용 및 데이터 수집 목적, 활용 방안에 대해 화자들에게 자세한 설명을 진행하고, 음성 데이터를 수집하는 진행 방법과 장비의 착용, 실제 수집 진행 시 화자가 주의해야 할 내용들을 충분히 설명하였다.

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 막는데 주력하였다. 이렇게 사업 목적 및 개인정보와 수집 데이터의 활용에 대하여 화자가 동의하면 다음 관련 서류 작성 단계로 넘어가게 된다.

4.2.2. 관련 서류 작성

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



개인정보 수집·이용·제3자 제공 동의서

(미성년자 법정대리인용)

본인은 미성년자의 법정대리인으로 해당 미성년자가 국립국어원의 "2021년 일상 대화 말뭉치 구축" 과제에 참여하여 [개인정보보호법] 제15조 및 제17조에 따라 아래의 내용으로 개인정보를 수집, 이용 및 제공하는데 동의합니다.

개인정보 취득자	개인정보 제공 제3자	수집·이용·제3자 제공 목적	수집·이용·제공되는 개인정보 항목	보유/이용기간
㈜스마트미디어테크	㈜나라지식정보 ㈜마인즈랩 국립국어원	◆ 국립국어원 - 일상 대화 말뭉치 구축 과제의 음성 말뭉치 및 인구통계학적 정보의 기초정보 구분 ◆ (주)스마트미디어테크 - 일상 대화 말뭉치 구축 과제의 음성 말뭉치 수집 ◆ ㈜나라지식정보 – 과제 중 음성데이터 전사, 개인식별정보 등 제거업무 ◆ ㈜마인즈랩 – 과제관리, 최종데이터 검수업무	발화 음성, 연구통계학적 정보(출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)	[국립국어원] 기본 2042년 12월 31일, 이후 5년 단위 자동 갱신 [나라지식정보, 마인즈랩] 2023년 3월 24일까지
국립국어원	학계·연구기관·산업체	일상 대화 말뭉치 구축 결과물로 국어 연구 및 언어정보처리분야 응용 기술 개발	발화 음성, 연구통계학적 정보(출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)	기본 2042년 12월 31일, 이후 5년 단위 자동 갱신
㈜스마트미디어테크	관공서(비용처리 신고에 한함)	과제 참여에 대한 회계 정산 처리 및 비용 증빙	성명, 주민등록번호	2022년 3월 24일까지

◆ 고유식별정보의 처리에 관한 사항

㈜스마트미디어테크는 개인정보보호법에 관한 법률에 따라 **회계 정산 처리 신고 목적으로 해당 미성년자 고유식별정보인 주민등록번호를 처리(수집, 이용, 제공)하고자 합니다.**

이에 동의하십니까? (동의하지 않을 경우, 과제참여가 불가능합니다)

[□ **동의합니다** □ 동의하지 않습니다.]

본인은 미성년자 법정대리인으로서 해당 미성년자에 대한 상기 개인정보 수집, 이용, 제3자 제공에 대하여 모두 확인하고 숙지하였으며, 국립국어원의 과제와 관련하여 해당 미성년자의 개인정보를 수집, 이용, 제3자 제공함에 동의합니다.

[□ **동의합니다** □ 동의하지 않습니다.]

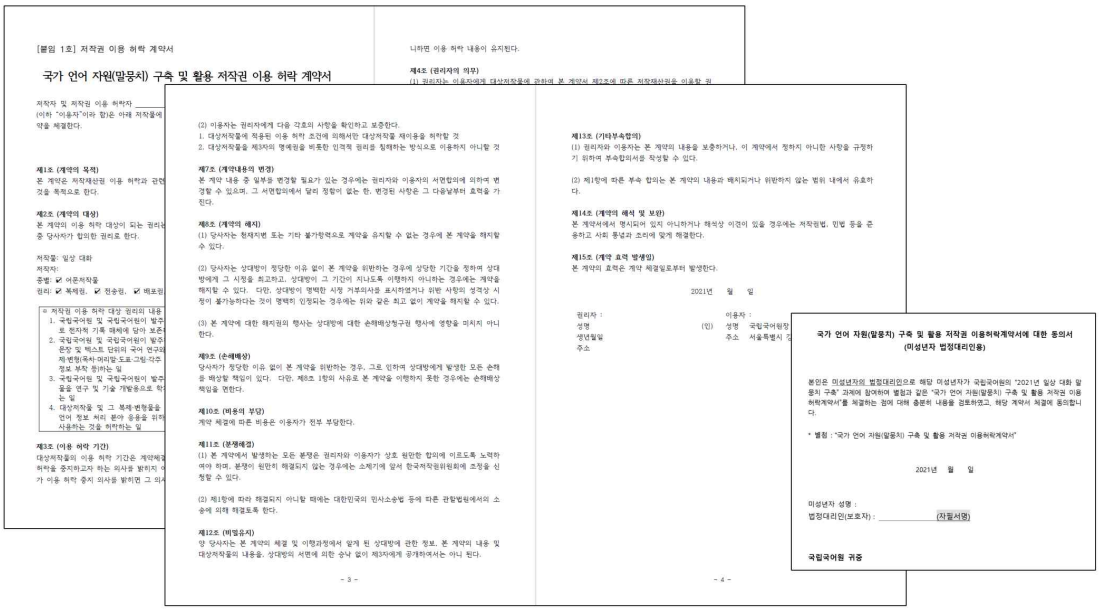
2021년 월 일

미성년자 성명 :

법정대리인(보호자) : _____ (자필서명)

국립국어원, ㈜스마트미디어테크, ㈜나라지식정보, ㈜마인즈랩 귀중

[그림 13] 개인정보 활용 동의서(예시)



[그림 14] 저작권 이용 허락 계약서(예시)

저작권 동의서 작성이 종료된 화자들을 대상으로 화자의 메타 정보(녹음 일시, 성명, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지 등)와 녹음 장소, 대화 주제의 수량, 대화 주제의 키워드를 아래와 같이 수집 일지에 작성하는 작업을 진행하였다.

녹음일시	녹음시간	이름	성별	나이	직업	신청지역	관계	출생지	성장지	현 거주지	학력
2021-08-01	10:00	A	남	31	회사원	서울	연인	서울	서울	서울	대학교_졸업
2021-08-01	10:00	B	여	29	회사원	서울	연인	서울	서울	서울	대학교_졸업

[그림 15] 음성 자료 수집 일지(예시-1)

녹음 일시 (녹음 시간)	2021-08-01(10:00)											
녹음 장소	서울											
주제 (NO.)	1_1	1_2	2_2	2_5								
키워드	휴가지, 가족여행	불만, 좋은점	언론탄압, 배상액	생명경시, 개인의 선택권 존중								

[그림 16] 음성 자료 수집 일지(예시-2)

4.2.3. 사전 시험 녹음

서류 작성이 완료된 화자들을 실제 녹음이 진행되는 공간으로 이용하여 자리 배치 및 수집 장비 착용을 진행하였다. 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 기준에 충족하지 못할 경우, 헤드셋 마이크와 화자의 입 거리를 조정하거나 오디오 인터페이스의 녹음 레벨을 조정하여 충분한 음량이 유지되도록 하였다. 잡음 이외에 대화를 진행하는 과정에서 불필요한 추임새, 발화의 겹침 등이 발생할 경우, 이에 대한 주의를 다시 한번 전달하여 실제 녹음 과정에서 해당 문제가 발생되지 않도록 하였다.

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

4.2.4. 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 총 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 하는 것이 발견되면 먼저 수신호로 화자들에게 주의를 주었다. 문제가 심해질 경우, 녹음을 일시 중단한 후 주의 사항을 다시 설명한 후 녹음을 계속 진행하였다. 이때 화자들이 선택한 주제에 대한 소재가 부족하여 대화를 계속 이어나가는 것이 어렵다고 판단될 때는 다른 주제로 변경하여 새롭게 대화를 진행하도록 하였다.

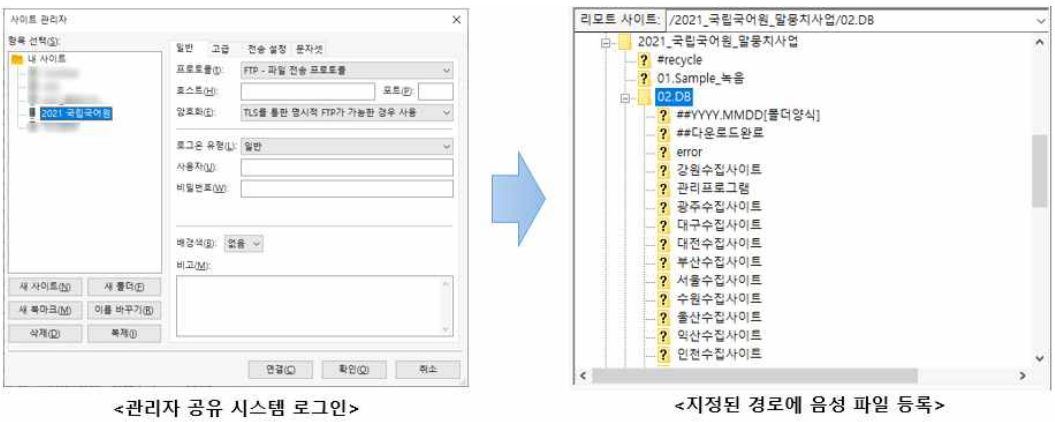
[그림 17] 음성 녹음 진행

4.2.5. 검증 및 업로드

음성 녹음이 완료되면 각 사이트의 담당자들은 실제 녹음된 파일을 다시 열어서 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 재녹음을 진행하거나 다른 날짜로 일정을 잡아 다시 수집하였다.

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

[그림 18] 수집 데이터 검증



<관리자 공유 시스템 로그인> <지정된 경로에 음성 파일 등록>

[그림 19] 공유 시스템 로그인 및 파일 등록(예시)

5. 음성 자료 전사

5.1. 전사 규칙

규칙은 전사 국립국어원 작성 ‘일상 대화 말뭉치 구축 지침’을 적용하였다. 파일명 부여 형식 및 json 구조, 대화와 관련된 화자의 메타 정보 등도 말뭉치 구축 지침에 따라 충실히 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 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 나는 대로 적는다.

○ 철자 전사는 한글 맞춤법 및 표준어 규정에 따라 적는 것으로, 한글 맞춤법 및 표준어 규정에 따라 전사하며 띄어쓰기도 한글 맞춤법에 따른다.

○ 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 하며, 하나의 전사 단위가 3초 이상으로 길어지는 것을 지양한다.

○ 긴 쉼에 의해 나뉘는 경우는 통사적으로 완성이 되지 않았다 하더라도 구분하여 전사한다.

○ 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.

5.2. 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. 플랫폼을 이용한 전사 절차는 아래와 같다.

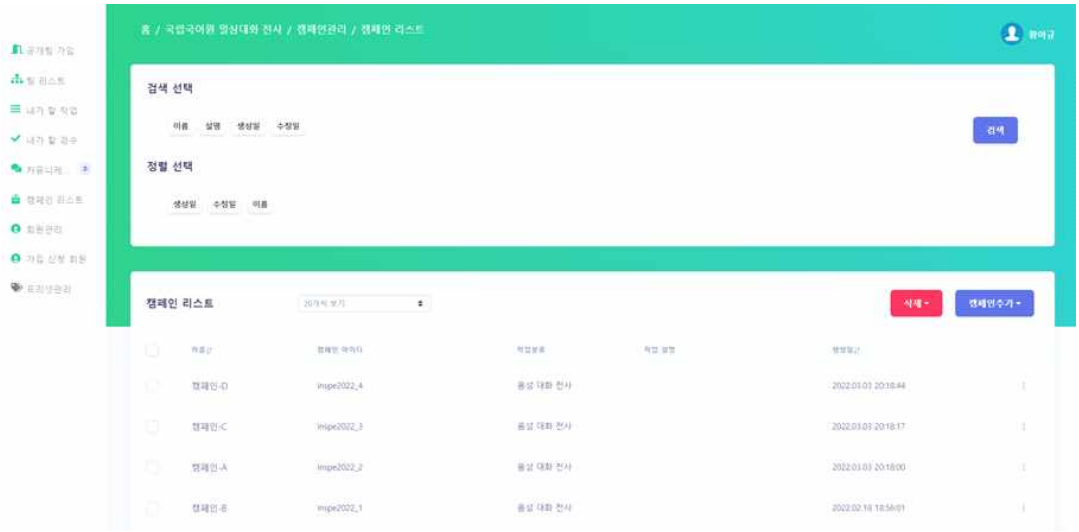
[그림 20] 전사 도구를 이용한 전사 절차

○ 작업자는 미작업 프로젝트를 할당받아 작업을 시작한다. 전사 지침에 맞춰 정제 진행 후 내용을 저장하여 작업 완료 상태로 만든다.

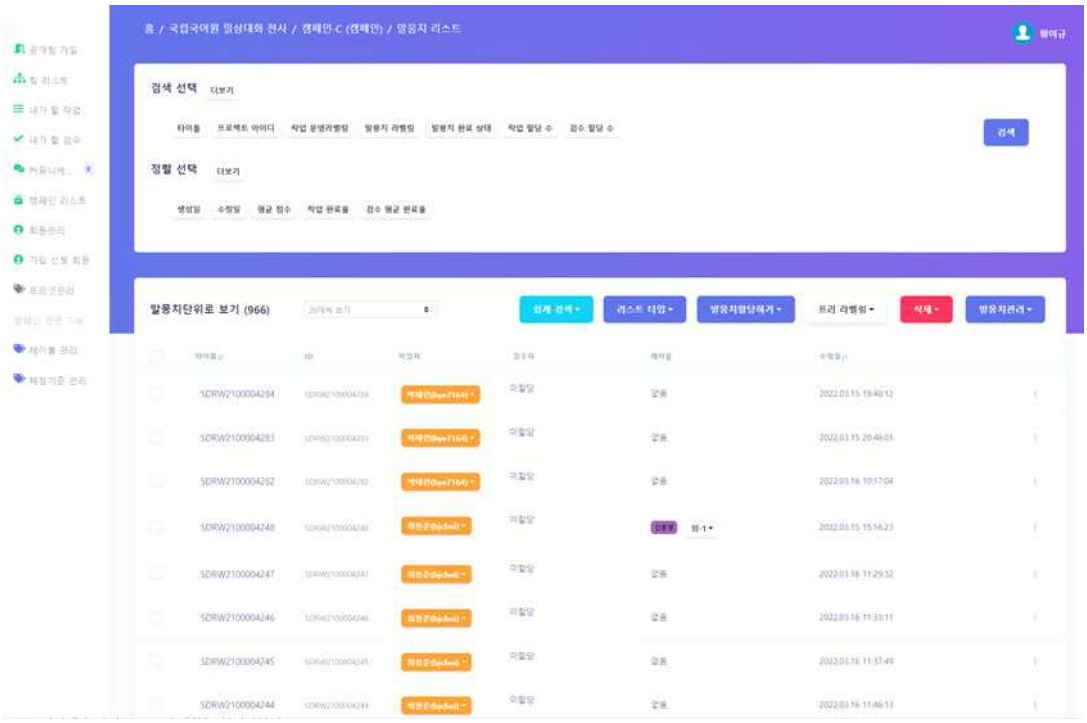
○ 정제가 완료된 작업 완료 파일은 검수자에게 넘어간다. 검수자는 내용 확인 후 이상이 없을 경우 검수 완료로 넘긴다. 만약 오류 사항이 발견되면 해당 파일은 작업 반려로 다시 작업자에게 돌아간다.

○ 반려 파일을 받은 작업자는 검수자가 작성한 반려 사유를 확인 후 재작업하여 작업 완료로 넘긴다.

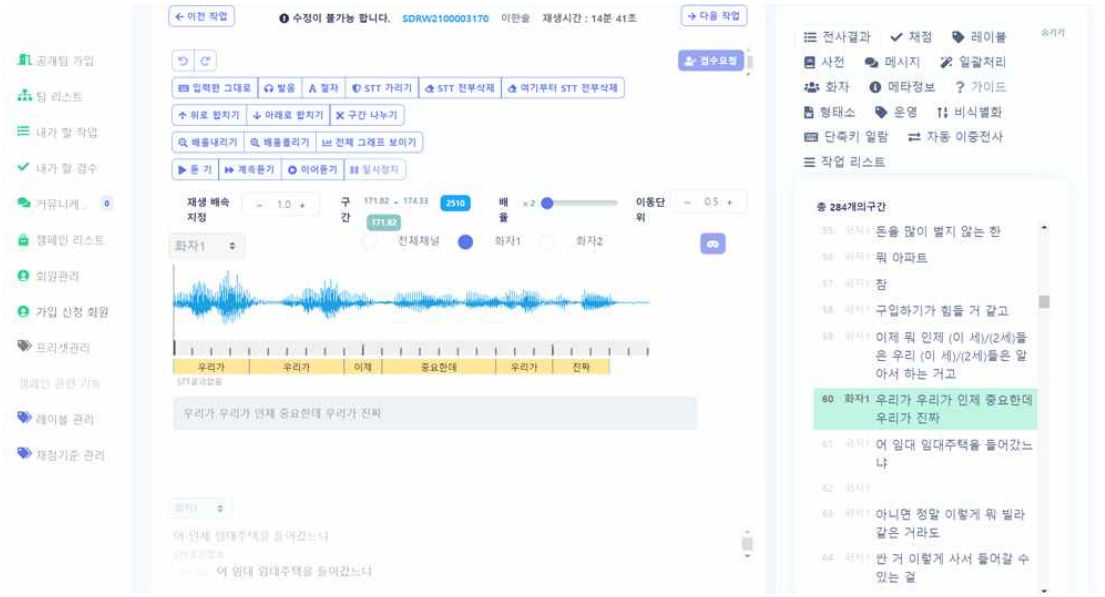
○ 검수자는 재작업한 파일을 검수 후 문제가 없다면 최종적으로 검수를 완료한다.



[그림 21] 전사 도구에서 전사 캠페인 보기



[그림 22] 전사 도구에서 전사 대상 대화 목록 보기



[그림 23] 전사 도구에서 전사 수정, 청취 및 결과 보기

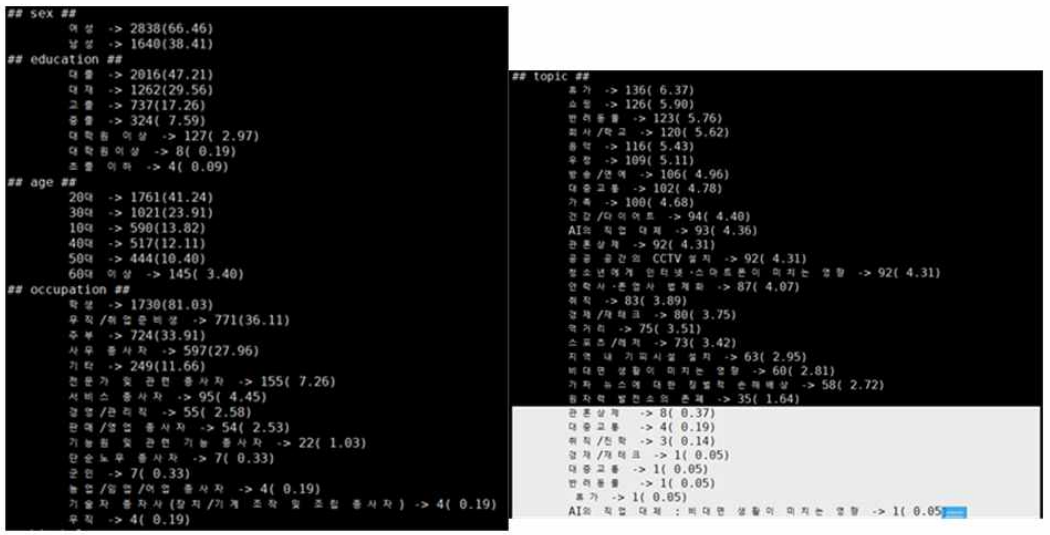
전사 인력 중 20년 이상 경력의 교정 교열 전문가에 의한 전사는 15분 발화 음성 기준 평균 2시간 정도(전사 1시간 30분, 자체 검토 30분)로 가장 짧은 시간이 소요되었으며 전사의 정확도 역시 가장 높았다. 언어 재활사, 언어계열 전공자와 데이터베이스 구축 사업 유경험자의 순으로 전사의 정확도가 높았으며 15분 발화 음성 기준 평균 2시간 30분 ~ 3시간 30분 정도의 시간이 소요되었다.

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 연령대가 비슷한 전사자를 배치했을 때 그렇지 않은 경우보다 전사 작업 시간과 정확도에서 20% 내외로 높은 효율을 보여 주었다.

5.3. 품질 검수

전사가 완료된 파일은 데이터 프로파일링 단계를 거치고, 검수 담당자를 지정하여 전수 수작업 검수를 진행하였다. 검수자가 전사 완료된 파일을 검사하기 전에 데이터 프로파일링을 통해 보다 효율적으로 검수를 진행할 수 있도록 하였다. 데이터 프로파일링이란, 전사 결과의 기계적 분석을 통하여 오류 가능성이 있는 후보를
탐지하는 과정이다. 이를 통해 메타 데이터 및 철자 오류 발생 가능 위치와 오류 가능성이 있는 내용을 탐지할 수 있다. 데이터 프로파일링으로 찾아낼 수 있는 오류 유형은 아래와 같다.



[그림 24] 메타 데이터 프로파일링(예시-1)

[그림 21]에서 보는 바와 같이 메타 데이터의 구축과정에서 메타 데이터의 띄어쓰기 오류가 있거나 메타 데이터가 기입되지 않는 경우 데이터 누적 통계 정보를 통해 오류를 확인한 후 수정하였다.

<table>
<tr><th>세부 공정</th><th>오류 내용</th></tr>
<tr><td>전사 오류 후보 탐지</td><td>똠 마음이 많이 아프다. 아무튼똠 마음이 많이 아프다. 아무튼 0
사쿠라한테 하루에 백 번씩 스쾃을를 하라고 시켰대. 그래서 사쿠라한테 하루에 백 번씩 스쿼트를를 하라고 시켰대. 그래서 40
한번 착 칠하고 한번착칠하고 7
태국 음식 중에 똠양꿍이라고 태국 음식중에 똠양꿍이라고 21
다시 뚯 많이 다시뚯많이 7
또 여성스러운 면으로 똠 더 감싸주고 또 여성스러운 면으로똠더 감싸주고 30
유명한 저~ 팟타이나 아니면 똠양꿍 유명한 저~ 팟타이나 아니면 똠양꿍 38
진짜 안 가려요. 저는 음식 가리는 게 좀 엙는 타입이에요. 진짜 안 가려요. 저는 음식 가리는 게 좀엙는 타입이에요. 54
살짝 설렜어로 컴백했는데 살짝 설렜어로 컴백했는데 10
찾아볼 수 있는 요소가 더 많은 거 같슬아요. 찾아볼 수 있는 요소가 더 많은 거 같슬아요. 49
하면서도 이제 너도 알게 되고 너랑도 이랭게 친해져서 하면서도 이제 너도 알게 되고 너랑도 이랭게 친해져서 54
얘가 이걸 뀱 긁히면얘가 이걸 -뀱-뀱히면 14</td></tr>
<tr><td>발음/철자 전사
비교 및 달리 전사한 경우 탐색</td><td>
<table>
<tr><th>발음</th><th>발음수정</th><th>타이틀</th><th>구</th><th>철자</th><th>역양구</th><th>철자수정</th></tr>
<tr><td>거_갈애</td><td>거_갈애.</td><td>SDRW2100002299</td><td>133</td><td>젓 갈아.</td><td>없는 것_갈아. $</td><td></td></tr>
<tr><td>거_갈애</td><td>거_갈애.</td><td>SDRW2100003436</td><td>71</td><td>젓 갈아.</td><td>아쉬운 것_갈아. $</td><td></td></tr>
<tr><td>젓</td><td>거</td><td>SDRW2100004023</td><td>60</td><td>거</td><td>갈 거 같았은데.</td><td>젓</td></tr>
<tr><td>젓</td><td>거</td><td>SDRW2100004124</td><td>300</td><td>거</td><td>또또한 거 본</td><td>젓</td></tr>
<tr><td>젓</td><td>거</td><td>SDRW2100000833</td><td>169</td><td>그것</td><td>^ 그것 보다</td><td>젓</td></tr>
<tr><td>구월</td><td>구 월</td><td>SDRW2100001520</td><td>110</td><td>9월</td><td>그 9월 거</td><td></td></tr>
<tr><td>구월</td><td>구 월</td><td>SDRW2100001561</td><td>179</td><td>9월</td><td>이번에 9월 일</td><td></td></tr>
<tr><td>구월</td><td>구 월</td><td>SDRW2100001849</td><td>18</td><td>9월</td><td>장사 9월 달에</td><td></td></tr>
<tr><td>구월</td><td>구 월</td><td>SDRW2100001852</td><td>107</td><td>9월</td><td>^ 9월 말쯤?</td><td></td></tr>
<tr><td>구월</td><td>구 월</td><td>SDRW2100003576</td><td>20</td><td>9월</td><td>매년 9월 달에</td><td></td></tr>
<tr><td>구월</td><td>구 월</td><td>SDRW2100003705</td><td>160</td><td>9월</td><td>^ 9월 달인가</td><td></td></tr>
<tr><td>그지_그지</td><td>그지_그지.</td><td>SDRW2100001652</td><td>480</td><td>그지 그지</td><td>^ 그지_그지 $</td><td>그렇지 그렇지.</td></tr>
<tr><td>꺼_갈애</td><td>꺼_갈애.</td><td>SDRW2100001291</td><td>102</td><td>거 갈아</td><td>생각할 거_갈아 $</td><td>거 갈아.</td></tr>
<tr><td>꺼_갈애</td><td>꺼_갈애.</td><td>SDRW2100001292</td><td>28</td><td>거 갈아</td><td>그럼 거_갈아 $</td><td>거 갈아.</td></tr>
<tr><td>꺼_갈애</td><td>꺼_갈애.</td><td>SDRW2100002298</td><td>10</td><td>거 갈아</td><td>바꿀 거_갈아 $</td><td>거 갈아.</td></tr>
<tr><td>꺼_갈애</td><td>꺼_갈애.</td><td>SDRW2100002300</td><td>16</td><td>거 갈아</td><td>없을 거_갈아 $</td><td>거 갈아.</td></tr>
<tr><td>꺼_갈애</td><td>꺼_갈애.</td><td>SDRW2100002372</td><td>51</td><td>거 갈아</td><td>할 거_갈아 $</td><td>거 갈아.</td></tr>
</table>
</td></tr>
</table>

[그림 25] 메타 데이터 프로파일링(예시-2)

위 예와 같이 언어분석 방법을 통해 철자 오류를 탐지하거나 동일한 발음 전사에 대해서 서로 다른 철자 전사가 진행된 예 등을 분석하여 오류 가능성을 확인하여 수정하였다.

데이터 프로파일링을 통해 기계식 검증을 마치고 품질 검수 담당자들이 전수 검수를 진행하였다. 품질 검수 담당자는 전사 결과를 확인하고 지침에 어긋난 경우 직접 수정하였다. 전사 및 검수 과정에서 자주 발생하는 오류의 유형을 작업자들이 공유하기 위해 전사 도구 내에 공지 게시판을 활용하였다.

6. 음성 정제

음성 정제는 음성을 전사 단위에 따라 분할하는 작업이다. 우선 주관기관의 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.

또한 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 단계에서 비식별화를 진행하였다. 개인정보에 해당하는 음성은 전사 도구를 통해 마킹 후 산출물 생성 시 묵음 처리를 하는 방식으로 비식별화 처리를 하였다.[6] [그림 26]의 예시에서 '이름'으로 표시된 부분은 산출물 PCM에서 묵음으로 변환되며, 전사 작업 및 검증 작업 과정에서 해당 부분에 비식별화 대상이 있음을 확인을 위해 음성 파형이 표시된다.

[그림 26] 개인정보 비식별화(예시)

6) 아래 그림에서 "어쩔 수 없이 제 아니지요 때문에"는 음성 인식 모듈을 통해 자동으로 변환된 것으로 오류가 있을 수 있으며 실제 수작업 전사한 부분은 "어쩔 수 없어 &name1& &name2& 때문에 여섯 시면 깨."임

7. 원시 말뭉치 구축 및 메타 정보 구축

7.1. JSON 변환

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

[표 19] 대화 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축년도	8자리 일련번호
S: 구어 말뭉치	D: 사적 대화	RW: 원시 말뭉치	21	########

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 20] 말뭉치 변환 예시(일부)

```
{       "id": "SDRW2100001383",
        "metadata": {
                "title": "국립국어원 구어 말뭉치 SDRW2100001383",
                "creator": "국립국어원",
                "distributor": "국립국어원",
                "year": "2021",
                "category": "구어 > 사적대화 > 협력적대화",
                "annotation_level": [
                        "원시"
                ],
                "sampling": "본문 전체"
        },
        "document": [
                {
                        "id": "SDRW2100001383",
                        "metadata": {
                                "title": "2인 일상 대화",
```

```
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20211117",
                "topic": "비대면 생활이 미치는 영향",
                "speaker": [
                        {
                                "id": "SD2100767",
                                "age": "20대",
                                "occupation": "사무 종사자",
                                "sex": "여성",
                                "birthplace": "대구",
                                "principal_residence": "대구",
                                "current_residence": "서울",
                                "education": "대졸"
                        },
                        {
                                "id": "SD2100768",
                                "age": "30대",
                                "occupation": "사무 종사자",
                                "sex": "여성",
                                "birthplace": "대구",
                                "principal_residence": "대구",
                                "current_residence": "서울",
                                "education": "대졸"
                        }
                ],
                "setting": {
                        "relation": "형제/자매"
                }
        },
        "utterance": [
                {
                        "id": "SDRW2100001383.1.1.1",
                        "form": "나는 코로나 때문에 비대면 생활을 하고",
                        "original_form": "나는 코로나 때문에 비대면 생활을 하고",
                        "speaker_id": "SD2100768",
                        "start": "2.35900",
                        "end": "5.97000",
                        "note": ""
                },
                {
                        "id": "SDRW2100001383.1.1.2",
                        "form": "너무 좋은 거 같아. 일단",
                        "original_form": "너무 좋은 거 같애. 일단",
                        "speaker_id": "SD2100768",
                        "start": "5.97000",
                        "end": "8.10000",
                        "note": ""
                },
                {
                        "id": "SDRW2100001383.1.1.3",
```

```
                "form": "9호선",
                "original_form": "구 호선",
                "speaker_id": "SD2100768",
                "start": "8.81000",
                "end": "9.83900",
                "note": ""
            },
...
```

7.2. 메타 정보 구축

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

2021 국립국어원 일상대화 말뭉치 메타정보

document								
id	metadata						setting	
id	title	author	publisher	data	topic1	topic2	relation	시간(s
SDRW2100000001.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210804	여행지(국내/해외)	해외여행, 숙소, 여행 스타일	친구	0:15:06
SDRW2100000001.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210805	음악	음악취향, 아이돌	친구	0:14:55
SDRW2100000002.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210805	먹거리	저녁 음식 선택	친구	0:14:37
SDRW2100000003.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210805	스포츠/레저	올림픽, 개인운동	형제/자매	0:15:02
SDRW2100000004.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210805	먹거리	저녁 음식 선택	형제/자매	0:14:46
SDRW2100000005.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210806	스포츠/레저	개인운동 및 스포츠 취향	친구	0:14:45
SDRW2100000006.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210806	휴가	휴가지 선택	친구	0:14:50
SDRW2100000007.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	음악	음악차트, 음악취향	친구	0:14:41
SDRW2100000008.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	휴가	휴가지 선택	친구	0:14:46
SDRW2100000009.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	음식 취향, 다음 외식 선정	부모/자녀	0:14:47
SDRW2100000010.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	휴가	휴가지 선택	부모/자녀	0:14:52
SDRW2100000011.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	건강/다이어트	개인운동 방식, 식단	모임_동아리 지인	0:14:53
SDRW2100000012.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	모임 저녁메뉴 선택	모임_동아리 지인	0:14:47
SDRW2100000013.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	건강/다이어트	건강식품 정보, 가족건강	모임_동아리 지인	0:14:43
SDRW2100000014.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	복날 가족식사메뉴 선택	모임_동아리 지인	0:14:54
SDRW2100000015.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	다이어트식, 저녁 메뉴 선택	연인	0:14:58
SDRW2100000016.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	휴가	휴가지 선택	연인	0:14:38
SDRW2100000017.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	외식문화	친구	0:14:44
SDRW2100000018.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	건강/다이어트	아침 운동 권유	친구	0:14:42
SDRW2100000019.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	휴가	휴가일정	모임_동아리 지인	0:14:49
SDRW2100000020.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	저녁 메뉴 선택	모임_동아리 지인	0:14:46
SDRW2100000021.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	건강/다이어트	건강검진, 식품, 건강관리	모임_동아리 지인	0:14:28
SDRW2100000022.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210809	먹거리	건강한 식사를 하자	모임_동아리 지인	0:14:27
SDRW2100000023.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	방송/연예	드라마, 예능 취향	형제/자매	0:14:33
SDRW2100000024.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	휴가	휴가지 선택	형제/자매	0:14:37
SDRW2100000025.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	가족	가족 음식 취향, 가족 건강	친구	0:14:40
SDRW2100000026.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	휴가	휴가지 선택	친구	0:14:34
SDRW2100000027.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	가족	집안 행사	모임_동아리 지인	0:15:28
SDRW2100000028.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	음악	선호 가수	모임_동아리 지인	0:14:11
SDRW2100000029.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	가족	가족사	모임_동아리 지인	0:14:40
SDRW2100000030.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20210810	휴가	휴가지 선택, 휴가 방식	모임_동아리 지인	0:14:51

[그림 27] 메타 정보 파일 일부

2021 국립국어원 일상대화 말뭉치 발화자 정보

*대화가 다르더라도 화자가 동일하면 동일한 아이디 부여

번호	fileId	id	name	age	occupation	sex	birthplace	principal_residenc	current_residenc	education
예시	SDRW2100000001	SD2100001	홍길동	30대	학생	남성	서울	경기	경기	대재
	SDRW2100000001	SD2100002	심청	30대	전문가 및 관련 종사자	여성	경기	경기	경기	대졸
1	SDRW2100000001	SD2100001	[redacted]	20대	학생	여성	서울	경기	서울	대재
2	SDRW2100000001	SD2100002	[redacted]	20대	학생	여성	전남	서울	서울	대재
3	SDRW2100000002	SD2100001	[redacted]	20대	학생	여성	서울	경기	서울	대재
4	SDRW2100000002	SD2100002	[redacted]	20대	학생	여성	전남	서울	서울	대재
5	SDRW2100000003	SD2100003	[redacted]	20대	학생	여성	서울	서울	서울	대재
6	SDRW2100000003	SD2100004	[redacted]	20대	학생	여성	서울	서울	서울	대재
7	SDRW2100000004	SD2100003	[redacted]	20대	학생	여성	서울	서울	서울	대재
8	SDRW2100000004	SD2100004	[redacted]	20대	학생	여성	서울	서울	서울	대재
9	SDRW2100000005	SD2100005	[redacted]	20대	학생	남성	서울	서울	서울	대재
10	SDRW2100000005	SD2100006	[redacted]	20대	학생	남성	서울	서울	서울	대재
11	SDRW2100000006	SD2100005	[redacted]	20대	학생	남성	서울	서울	서울	대재
12	SDRW2100000006	SD2100006	[redacted]	20대	학생	남성	서울	서울	서울	대재
13	SDRW2100000007	SD2100007	[redacted]	20대	학생	여성	서울	서울	서울	대재
14	SDRW2100000007	SD2100008	[redacted]	20대	학생	여성	서울	서울	서울	대재
15	SDRW2100000008	SD2100007	[redacted]	20대	학생	여성	서울	서울	서울	대재
16	SDRW2100000008	SD2100008	[redacted]	20대	학생	여성	서울	서울	서울	대재

[그림 28] 발화자 메타 정보 일부

메타 정보 구축에는 수집에 참여한 화자의 정보(성별, 연령대, 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 활용이 되었는데, 수집 후반의 경우 결과물의 목표치를 초과하지 않도록 모니터링을 하는 도구로 사용되었다. 해당 과제에서 목표하는 각 성별, 연령별,

지역별, 주제별 수치가 다수의 지역 사이트에서 수집이 진행되다 보니, 수집 사이트에서 아무리 실시간으로 수집 결과를 업로드해도 이를 취합하여 진도 상황을 체크하는 것에는 어려움이 수반된다.

따라서 본 메타 정보 구축을 통하여, 각 사이트에서 메타 정보가 업데이트되면서 각 수집 항목별 비율 산출이 가능하였다. 이를 토대로 연령, 지역, 제시 주제가 목표치에 도달하게 되면 해당 파트에 대한 화자의 모집이나 주제 사용을 금지하여, 불필요한 데이터의 수집을 최소화하는데 활용되었다. 또 메타 정보는 데이터를 수집 기관에서만 활용하지 않고 전사를 담당하는 곳에서도 활용하였다. 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 15개의 대주제와 세부 예시 주제를 제시하였으며, 협력적 대화는 총 8개의 주제로 수집하였다. 작년 일상 대화 말뭉치의 주제 및 수집 비율을 고려하여 신규 일상 대화 주제[7] 수집 비율을 8% 내외로 높이고, 기존 대화 주제는 4% 내외로 수집하였다.

[표 21] 주제별 대화 수집 결과

유형	대주제	세부 예시 주제	수집 쌍	비율[8]
일상 대화	휴가	여행 시 교통, 숙박 선택	181	5.51%
	대중교통	**약속 시간, 장소, 교통, 선택**	287	8.73%
	음악	**대중음악 유행, 선호 가수 및 곡 추천**	312	9.49%
	건강/다이어트	성인병에 대한 상식, 처방, 대응	169	5.14%
	방송/연예	드라마, 예능 프로그램 선택	170	5.17%
	스포츠/레저	직접 운동, 관람, 시청 등 참여 방법에 대한 정보와 결정	167	5.08%
	먹거리	저녁 모임에 대한 음식 종류와 식당 선택	178	5.42%
	우정	**친구 간 선호, 성격, 취미 토론**	338	10.29%
	경제/재테크	집, 주식 등 투자에 대한 고려와 결정	139	4.23%
	회사/학교	취직, 진학에 대한 정보와 결정	182	5.54%
	반려동물	개, 고양이의 장단점 비교 및 결정	175	5.33%
	취직	**대기업/중소기업 취직의 정보 공유와 견해 교환**	224	6.82%
	가족	집안 행사에 대한 검토와 결정	163	4.96%
	쇼핑	핸드폰 구매시 기종 검토 및 결정	333	10.13%
	관혼상제	**결혼, 문상, 제사, 축의금, 참석 등**	268	8.16%
	부분 합계		3,286	100.00%
협력적 대화	공공 공간의 CCTV 설치		102	11.90%
	가짜 뉴스에 대한 징벌적 손해배상		96	11.20%
	원자력 발전소의 존폐		90	10.50%
	지역 내 기피 시설 설치		94	10.97%
	안락사·존엄사 법제화		113	13.19%
	AI의 직업 대체		131	15.29%
	비대면 생활이 미치는 영향		112	13.07%
	청소년에게 인터넷·스마트폰이 미치는 영향		119	13.89%
	부분 합계		857	100.00%
전체 합계			**4,143**	

7) 신규 일상 대화 주제는 **'대중교통, 음악, 우정, 취직, 쇼핑, 관혼상제'**로 표 19에 굵게 표시함.
8) 수집 비율은 소수점 셋째 자리에서 반올림 한 값으로 최종 총 값은 허용 오차가 ±0.02%임.
이 비율은 아래 [표 21]에서 [표 32]까지 적용됨.

2. 화자 모집 결과

2.1. 인구 특성별 수집 결과

일상 대화 말뭉치 구축 설계 시 한 성별, 연령별, 지역별 목표치를 충족하기 위해 다양한 방법으로 화자 모집을 진행하였다. 화자 모집은 일반적으로 구인 사이트를 활용하였으나, 코로나-19라는 특수 상황으로 그 효율이 기존에 비하여 저조하였다. 따라서 이번 사업에서는 구인 사이트 활용 외에도 각 지역의 홍보 사이트, 맘 카페, 종교 모임, 대학 동아리 등을 통한 홍보를 진행하여 화자를 모집하였다. 이러한 과정을 통해 총 2,599명의 발화자를 모집하여 목표한 1,000시간 이상의 녹음 데이터를 수집하였다. 수집에 참여한 화자의 결과는 아래 표와 같다.

[표 22] 성×연령×지역별 화자 모집 결과(단위: 명)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대 이상	10대	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	22	81	38	24	20	6	31	133	59	70	71	31	586	1,431
	인천	8	25	17	2	1	1	10	27	16	11	6	2	126	
	경기	32	167	46	5	4	4	45	241	80	46	35	14	719	
영남권	부산	2	21	8	5	2	3	19	38	15	11	11	3	138	565
	경남	1	20	11	1	1	1	14	35	16	6	5	1	112	
	울산	0	12	3	1	1	0	11	21	5	1	3	1	59	
	대구	14	25	6	2	2	1	19	27	23	11	9	1	140	
	경북	9	25	9	2	2	1	16	26	7	6	10	3	116	
호남권	광주	6	15	12	1	1	0	6	10	11	5	9	1	77	266
	전북	5	19	6	4	3	0	2	20	11	9	14	6	99	
	전남	2	19	3	1	1	1	8	14	11	13	13	4	90	
충청권	대전	12	18	12	3	1	1	9	16	16	9	3	1	101	242
	충북	4	10	5	2	1	0	4	19	7	4	6	2	64	
	충남	4	15	4	2	1	2	7	16	12	7	5	2	77	
강원권	강원	1	4	2	1	6	1	1	16	11	9	8	5	65	65
제주권	제주	1	6	1	0	1	1	2	9	3	4	1	1	30	30
합계		123	482	183	56	48	23	204	668	303	222	209	78	2,599	2,599

[표 23] 성×연령×지역별 화자 모집 결과(단위: 시간)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대	10대	20대	30대	40대	50대	60대	지역별	권역별
수도권	서울	10.75	26.38	14.50	9.88	9.33	2.75	12.75	38.42	25.13	28.21	28.33	10.38	216.79	
	인천	3.75	9.75	7.50	1.00	0.50	0.50	5.00	10.88	7.75	5.25	2.38	1.00	55.25	545.33
	경기	13.25	61.17	20.50	2.25	1.88	1.25	17.96	76.63	36.63	20.96	15.83	5.00	273.29	
영남권	부산	1.00	8.25	3.88	2.25	0.75	1.50	7.88	16.29	6.88	4.71	5.25	1.25	59.88	
	경남	0.19	9.50	5.00	0.50	0.50	0.50	6.00	15.63	7.83	2.75	2.25	0.50	51.15	
	울산	0.00	5.25	1.25	0.50	0.25	0.00	5.00	9.25	2.00	0.50	1.50	0.50	26.00	241.26
	대구	4.58	9.88	3.00	1.00	0.75	0.50	6.79	11.08	10.00	4.58	3.08	0.50	55.75	
	경북	3.44	9.50	4.25	0.83	1.00	0.50	6.13	12.38	3.00	2.83	3.88	0.75	48.48	
호남권	광주	1.75	4.71	4.00	0.25	0.25	0.00	3.00	4.25	4.75	2.00	3.25	0.50	28.71	
	전북	2.00	7.67	3.00	1.75	1.50	0.00	1.00	8.92	5.13	4.13	6.38	2.50	43.96	107.92
	전남	1.00	5.25	1.25	0.25	0.50	0.50	3.50	6.38	5.00	5.63	4.75	1.25	35.25	
충청권	대전	4.06	6.27	5.50	1.50	0.25	0.13	3.42	6.19	7.50	4.50	1.50	0.50	41.31	
	충북	1.33	4.19	2.50	1.00	0.38	0.00	1.83	7.50	3.50	2.00	2.13	1.00	27.35	101.29
	충남	1.00	6.17	1.88	1.00	0.50	0.88	3.25	6.38	5.00	3.50	2.25	0.83	32.63	
강원권	강원	0.50	2.00	0.75	0.38	2.38	0.13	0.13	7.13	5.33	3.75	3.75	1.50	27.71	27.71
제주권	제주	0.50	2.25	0.50	0.00	0.38	0.25	1.00	3.50	1.00	1.88	0.50	0.50	12.25	12.25
합계		49.10	178.17	79.25	24.33	21.08	9.38	84.63	240.77	136.42	97.17	87.00	28.46	1035.76	1035.76

2.2. 주제별 연령대 분포

수집에 활용된 대화 주제들은 수집에 참여하는 인원들이 자신들이 주제를 선택하는 형태로 진행되었다. 다만, 후반부 일부 주제의 경우 목표치를 달성하여, 사용을 제한하였으나, 최대한 다양한 선택지를 제시하여, 수집에 참여하는 화자들이 관심 분야에 대한 대화를 진행하도록 하였다. 각 주제별 한 화자 당 대화 참여 수는 1~4개이며, 2~4인 대화가 혼용되어 있다. 이에 [표 22]에 표시된 인원은 총 대화 참여 2,599명이 여러 개의 대화에 참여하였기 때문에 총인원이 다르게 나타난다.

각 주제들은 연령별로 다양하게 사용되었는데, 10대의 경우 우정, 회사/학교, 음악을, 20대의 경우 음악, 우정, 쇼핑을, 30대의 경우 관혼상제, 쇼핑, 가족을, 40의 경우 건강/다이어트, 쇼핑, 경제/재테크를, 50대의 경우 관혼상제, 쇼핑, 건강/다이어트가, 60대 이상의 경우 대중교통, 건강/다이어트, 우정 순으로 선택되었다.

이는 10대의 경우 학교생활과 친구 관계를, 20대의 경우 취미와 친구 관계 및 소비생활을, 30대의 경우 애경사 및 소비생활과 가족관계에, 40대의 경우 자신의 건강 및 가족과 애경사에, 50대의 경우 애경사 및 소비생활 및 건강에, 60대의 경우 교통 및 건강, 주변 관계에 관심을 보이고 있음을 보이며, 이는 실제 나이대에 관심이 높은 분야를 적절하게 선택한 것으로 판단된다.

[표 24] 주제별 연령대 분포(단위: 시간)

주제	10대	20대	30대	40대	50대	60대	합계	비율
휴가	22.00	85.00	37.00	16.50	17.50	3.00	181.00	5.51%
대중교통	44.50	146.00	36.50	21.00	23.50	15.50	287.00	8.73%
음악	41.50	183.50	39.00	18.00	20.00	10.00	312.00	9.49%
건강/다이어트	5.00	37.00	36.67	44.83	31.50	14.00	169.00	5.14%
방송/연예	27.00	66.50	41.00	16.50	16.50	2.50	170.00	5.17%
스포츠/레저	16.00	61.25	40.42	22.33	19.25	7.75	167.00	5.08%
먹거리	20.00	78.00	34.00	24.50	15.00	6.50	178.00	5.42%
우정	61.25	178.25	45.67	18.83	21.00	13.00	338.00	10.29%
경제/재테크	4.50	26.00	40.83	30.67	28.50	8.50	139.00	4.23%
회사/학교	49.50	51.67	45.67	21.17	11.50	2.50	182.00	5.54%
반려동물	25.50	59.33	44.83	22.83	20.50	2.00	175.00	5.33%
취직	15.00	136.75	32.75	19.00	14.25	6.25	224.00	6.82%
가족	13.00	29.08	50.08	29.67	31.42	9.75	163.00	4.96%

쇼핑	33.67	161.83	53.50	40.33	33.33	10.33	333.00	10.13%
관혼상제	17.00	118.50	56.50	29.67	35.00	11.33	268.00	8.16%
공공 공간의 CCTV 설치	15.67	30.00	23.00	13.33	17.00	3.00	102.00	11.90%
가짜 뉴스에 대한 징벌적 손해배상	15.17	29.00	23.67	14.17	9.67	4.33	96.00	11.20%
원자력 발전소의 존폐	15.17	37.08	16.42	7.17	10.08	4.08	90.00	10.50%
지역 내 기피 시설 설치	17.92	22.58	25.17	16.17	9.17	3.00	94.00	10.97%
안락사/존엄사 법제화	12.00	32.33	34.17	13.00	16.50	5.00	113.00	13.19%
AI의 직업 대체	22.08	49.42	33.33	14.50	10.17	1.50	131.00	15.29%
비대면 생활이 미치는 영향	21.00	29.50	33.17	14.33	10.00	4.00	112.00	13.07%
청소년에게 인터넷/스마트폰이 미치는 영향	20.50	27.17	39.33	17.50	11.00	3.50	119.00	13.89%
합계								100.00%

[표 25] 주제별 연령대 분포(단위: 명)

주제	10대	20대	30대	40대	50대	60대 이상	합계	비율
휴가	45	172	74	33	35	6	365	5.32%
대중교통	93	301	75	42	48	31	590	8.60%
음악	95	402	79	37	41	20	674	9.83%
건강/다이어트	10	79	74	90	64	28	345	5.03%
방송/연예	54	147	83	33	33	5	355	5.18%
스포츠/레저	34	140	82	45	39	16	356	5.19%
먹거리	41	158	68	49	30	13	359	5.24%
우정	134	386	92	38	42	26	718	10.47%
경제/재테크	9	53	83	62	58	17	282	4.11%
회사/학교	101	114	92	43	23	5	378	5.51%
반려동물	54	127	91	46	41	4	363	5.29%
취직	30	286	66	38	29	13	462	6.74%
가족	28	67	101	61	65	20	342	4.99%
쇼핑	72	361	109	81	67	21	711	10.37%
관혼상제	35	249	117	61	72	23	557	8.12%
일상 대화 참여 인원	835	3,042	1,286	759	687	248	6,857	100.00%
공공 공간의 CCTV 설치	36	78	46	27	34	6	227	11.69%
가짜 뉴스에 대한 징벌적 손해배상	38	79	48	30	21	9	225	11.59%
원자력 발전소의 존폐	36	106	34	15	21	9	221	11.38%
지역 내 기피 시설 설치	46	61	54	34	22	6	223	11.48%
안락사/존엄사 법제화	24	83	69	27	35	10	248	12.77%
AI의 직업 대체	55	135	70	29	21	3	313	16.12%
비대면 생활이 미치는 영향	43	66	67	29	20	8	233	12.00%
청소년에게 인터넷/스마트폰이 미치는 영향	43	63	81	36	22	7	252	12.98%
협력적 대화 참여 인원	321	671	469	227	196	58	1,942	100.00%

2.3. 대화 유형별 분포

대화 유형별 참여 인원은 초기 제시된 목표에 맞추어 화자를 섭외하고 음성 자료를 수집하였다. 다만 다자 대화의 경우 코로나-19로 인하여 화자의 섭외가 쉽지 않아 목표한 모집 인원이 적음에도 과제 마지막까지 모집이 진행되었고, 다행히 기간 내에 모집을 완료할 수 있었다.

이렇게 모집된 대화 유형별 분포는 일상 대화가 전체의 약 80%인 3,286건이며, 협력적 대화가 전체의 약 20%인 857건이었다. 또 2인 대화 비율은 전체의 90%인 3,723건, 3인과 4인 대화 비율은 전체의 10%인 420건이다.

[표 26] 대화 유형 및 인원별 분포(단위: 대화 수량)

대화 유형	2인 대화	3인 대화	4인 대화	합계	비율
일상 대화	3,054	179	53	3,286	79.31%
협력적 대화	669	148	40	857	20.69%
합 계	3,723	327	93	4,143	100.00%

2.4. 주제별 성별 분포

주제별 성별 분포의 경우 모든 주제를 고르게 사용하였으며, 남녀 모두 일상생활과 밀접한 음악과 대중교통을 다수 선택하였다. 이 공통항목 이외에서는 남성의 경우 우정, 취직, 스포츠가 여성의 경우 쇼핑, 우정, 관혼상제가 높은 순위로 선택되었다.

[표 27] 주제별 성별 분포(단위: 대화 수량)

주제	남성	여성	합계	비율
휴가	60.00	121.00	181.00	5.51%
대중교통	92.00	195.00	287.00	8.73%
음악	135.75	176.25	312.00	9.49%
건강/다이어트	44.33	124.67	169.00	5.14%
방송/연예	40.00	130.00	170.00	5.17%
스포츠/레저	90.50	76.50	167.00	5.08%
먹거리	62.75	115.25	178.00	5.42%
우정	113.83	224.17	338.00	10.29%
경제/재테크	53.83	85.17	139.00	4.23%
회사/학교	67.83	114.17	182.00	5.54%
반려동물	51.00	124.00	175.00	5.33%
취직	105.67	118.33	224.00	6.82%
가족	46.08	116.92	163.00	4.96%
쇼핑	100.67	232.33	333.00	10.13%
관혼상제	74.83	193.17	268.00	8.16%
일상 대화 합계	1,139.07	2,146.93	3,286	100.00%
공공 공간의 CCTV 설치	36.58	65.42	102.00	11.90%
가짜 뉴스에 대한 징벌적 손해배상	32.33	63.67	96.00	11.20%
원자력 발전소의 존폐	36.17	53.83	90.00	10.50%
지역 내 기피 시설 설치	28.33	65.67	94.00	10.97%
안락사/존엄사 법제화	46.67	66.33	113.00	13.19%
AI의 직업 대체	52.42	78.58	131.00	15.29%
비대면 생활이 미치는 영향	33.00	79.00	112.00	13.07%
청소년에게 인터넷/스마트폰이 미치는 영향	40.67	78.33	119.00	13.89%
협력적 대화 합계	306.17	550.83	857	100.00%

2.5. 화자 관계별 분포

수집에 참여한 화자 간 관계의 경우 모집단계에서 다양한 관계가 섭외될 수 있도록 홍보하고 모집하였다. 그러나 실제 데이터의 수집은 자연스러운 대화가 이루어지는 형태로 진행되어야 하다 보니 참여자 간 친밀함이 필수적일 수밖에 없었다. 따라서 실제 참여 화자들의 관계를 살펴보면, (친구 - 가족관계 - 지인) 순으로 비율이 높게 나타났다. 예외적으로 아르바이트 사이트를 통해 수집에 혼자 참여한 화자들이 있어 짝을 지어 녹음을 진행하였으나 대화가 원활하게 되지 않는다고 판단된 경우 수집을 중단하였다.

[표 28] 화자 간 관계별 수집 결과(단위: 대화 수량)

화자 간 관계	2인 대화	3인 대화	4인 대화	합계	비율
친구	1,763	250	70	2,083	50.28%
부부	313	0	0	313	7.55%
부모/자녀	349	11	4	364	8.79%
형제/자매	305	14	0	319	7.70%
연인	364	0	0	364	8.79%
직장 동료	83	8	0	91	2.20%
이웃사촌	24	0	0	24	0.58%
모임·동아리 지인	330	27	8	365	8.81%
대학 선후배	94	9	7	110	2.66%
교회 지인	12	0	0	12	0.29%
선후배	19	0	0	19	0.46%
기타 가족	44	8	4	56	1.35%
기타	23	0	0	23	0.56%
합계	3,723	327	93	4,143	100.00%

2.6. 직업별 분포

최대한 다양한 직업군을 모집하기 위해 다양한 방법을 활용하였으나, 실제 데이터의 수집이 평일 낮에 이루어지는 만큼 해당 시간에 여유가 있는 학생(39.40%), 무직(17.51%), 주부(13.43%) 순으로 참여도가 높았다. 다만 이 경우 특정 직업군에 쏠림 현상이 나타나 수집 시간을 야간 및 주말로 확장하여 가능한 다양한 직업이 수집에 참여할 수 있도록 하였다.

[표 29] 직업별 수집 결과(단위: 명)

직업	모집 인원	비율
경영/관리직	33	1.27%
기능원 및 관련 기능 종사자	6	0.23%
기술자 종자사 (장치/기계 조작 및 조립 종사자)	2	0.08%
기타	121	4.66%
농업/임업/어업 종사자	9	0.35%
단순노무 종사자	5	0.19%
무직/취업준비생	455	17.51%
사무 종사자	349	13.43%
서비스 종사자	39	1.50%
전문가 및 관련 종사자	87	3.35%
주부	431	16.58%
판매/영업 종사자	36	1.39%
학생	1,024	39.40%
군인	2	0.08%
합 계	2,599	100.00%

2.7. 학력별 분포

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

[표 30] 학력별 수집 결과(단위: 명)

학력	모집 인원	비율
초졸 이하	1	0.04%
중졸	153	5.89%
고졸	421	16.20%
대재	804	30.93%
대졸	1,142	43.94%
대학원 이상	78	3.00%
합계	2,599	100.00%

2.8. 출생지별 분포

모집 화자의 출생지 분포의 경우 1순위는 서울 714명(27.47%)이며, 2순위는 경기 551명(21.20%), 3순위는 부산 177명(6.81%)로 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 31] 출생지별 화자 모집 결과(단위: 명)

출생지	모집 인원	비율
서울	714	27.47%
광주	86	3.31%
대구	149	5.73%
대전	106	4.08%
부산	177	6.81%
울산	63	2.42%
인천	115	4.42%
강원	54	2.08%
경기	551	21.20%
경북	117	4.50%
경남	107	4.12%
전북	124	4.77%
전남	83	3.19%
충북	55	2.12%
충남	72	2.77%
제주	26	1.00%
합계	2,599	100.00%

2.9. 주 성장지별 분포

모집 화자의 성장지별[9] 분포의 경우 1순위는 경기 719명(27.66%)이며, 2순위는 서울 586명(22.55%), 3순위는 대구 140명(5.39%)로 나타났다. 지역별 화자 모집 비율은 주 성장지를 기준으로 설계되었고, 해당 비율을 고려하여 수집하였다.

[표 32] 주 성장지별 화자 모집 결과(단위: 명)

주 성장지	모집 인원	비율
서울	586	22.55%
광주	77	2.96%
대구	140	5.39%
대전	101	3.89%
부산	138	5.31%
울산	59	2.27%
인천	126	4.85%
강원	65	2.50%
경기	719	27.66%
경북	116	4.46%
경남	112	4.31%
전북	99	3.81%
전남	90	3.46%
충북	64	2.46%
충남	77	2.96%
제주	30	1.15%
합계	2,599	100.00%

9) 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 890명(34.24%), 2순위는 경기871명(33.51%), 3순위 대구 142명(5.46%)이다.

[표 33] 현 거주지별 화자 모집 결과(단위: 명)

현 거주지	모집 인원	비율
서울	890	34.24%
광주	78	3.00%
대구	142	5.46%
대전	137	5.27%
부산	118	4.54%
울산	49	1.89%
인천	158	6.08%
강원	21	0.81%
경기	871	33.51%
경북	13	0.50%
경남	12	0.46%
전북	87	3.35%
전남	3	0.12%
충북	1	0.04%
충남	19	0.73%
제주	0	0.00%
합계	2,599	100.00%

3. 정책 제언

본 사업은 다양한 사람들을 모아 일상 대화를 녹음 후 전사하고 정제하여 일상 대화 말뭉치 총 1,000시간을 구축하는 사업이다. 자연스러운 일상 대화 말뭉치 구축은 현시대의 언어생활을 반영하는 언어자원으로서 다양한 국어 연구 및 인공지능 등 산업 분야에서 활용할 수 있을 것으로 기대된다.

반면, 사업을 진행하며 겪었던 문제점 및 개선 사항은 아래와 같다.

- 음성 대화 수집 과정에서 녹음 품질을 고려하고 잡음을 최소화하기 위해 밀폐된 스튜디오 환경에서 작업을 진행했는데, 이는 실생활을 오롯이 반영한 것으로 볼 수 없다. 실생활 잡음이 들어간 환경이나 야외에서 잡음이 들어간 환경에서의 일상 대화 수집도 고려해야 할 것으로 보인다.

- 다양한 연령과 지역, 생활 배경을 가진 화자의 대화이기 때문에 대화 주제, 사투리와 발성 구조의 다양성으로 인한 일부 대화에서 전사의 어려움이 발생하였다. 대화 참여자의 특성을 고려하여 다양한 배경을 가진 전사 작업자를 모집하여 해당 대화 특성에 맞는 적절한 전사자를 할당하여 작업한다면 전사 품질도 크게 향상될 것이다.

- 자연스러운 다자간 대화의 수집이 어려웠다. 코로나-19 환경에서 다자 대화에 참여하려는 참가자를 모집하기도 어려웠고, 2명 이상의 대화에서 발화 겹침 등이 많이 발생하였다. 이를 해소하기 위해 원격으로 대화할 수 있는 환경을 구축하고 이를 통해 대화를 수집할 수 있는 방안의 마련이 필요하다. 온라인 회의 환경을 적극적으로 활용하여 대화를 수집하면, 추후 비대면 환경에서의 다양한 대화를 수집할 수 있을 것이다.

- 특정 연령대와 성별에서 대화자 모집이 쉽지 않았다. 40-50대 남성의 경우, 경제활동에 참여하는 사람들이 대부분이어서 주중에 참여자를 찾기 어려웠다. 이를 해소하기 위해 주말에도 수집하는 방법을 택하기는 했지만, 적극적인 참여 유도 방법을 찾기 어려웠다. 향후 사업에서는 국가 주도의 데이터 구축 사업임을 감안하여 기업, 관련 기관 등을 대상으로 하는 홍보를 통해 집단적인 참여를 유도하는 방안 등이 필요하다.

- 대화를 구성하는 대화자 사이의 관계가 특정 관계로 많이 한정되는 경향을 보였다. 이에 따라 대화 상대의 연령, 성별과 대화 상대와의 관계 등에서 다양성의 확보가 부족하였다. 화자 간 다양성의 확보를 위한 새로운 방안이 필요할 것으로 보인다. 전혀 모르는 사이의 자연스럽고 원활한 대화(예를 들어 면접 등)를 수집하기 위해 적절한 대화 주제의 발굴이 필요하다.

[붙임 1] 상 대화 말뭉치 구축 지침

일상 대화 말뭉치 구축 지침

1. 파일 형식 및 개요

1.1. 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축년도	8자리 일련번호
S: 구어 말뭉치	D: 사적 대화	RW: 원시 말뭉치	21	########

- 예시

· SDRW2100000001.json 원시 말뭉치 첫 번째 파일

※ 참고: 음성 파일 파일명 부여 방식

· SDRW2100000001.pcm 음성 원본 첫 번째 파일

· SDRW2100000001-00001.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	구축년도: 2021
	category			string	분류: 구어 > 사적 대화 > 일상 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
		speaker		array(object)	화자 정보
			id	string	화자 아이디
			age	string	연령
			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
	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": "SDRW2100000001",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2100000001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2021",
        "category": "구어 > 사적 대화 > 일상 대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2100000001.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20210711",
                "topic": "자동차",
                "speaker": [
                    {
                        "id": "SD2100011",
                        "age": "30대",
                        "occupation": "사무 종사자",
                        "sex": "남성",
                        "birthplace": "대구",
                        "pricipal_residence": "대구",
                        "current_residence": "경북",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2100012",
                        "age": "30대",
                        "occupation": "사무 종사자",
                        "sex": "남성",
                        "birthplace": "대구",
                        "pricipal_residence": "대구",
                        "current_residence": "대구",
                        "education": "대졸"
                    }
                ],
                "setting": {
                    "relation": "동료"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2100000001.1.1.1",
                    "form": "안녕하세요.",
                    "original_form": "안녕하세요.",
                    "speaker_id": "SD2100011",
                    "start": 30.56600,
                    "end": 32.48262,
                    "note": ""
                },
                {
                    "id": "SDRW2100000001.1.1.2",
                    "form": "아~ xx님 오랜만입니다.",
                    "original_form": "아~ ((xx님)) 오랜만입니다.",
                    "speaker_id": "SD2100012",
                    "start": 33.12500,
                    "end": 34.1543323,
                    "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 SDRW2100000001)
- 구축자(creator): 국립국어원
- 배포자(distributor): 국립국어원
- 구축년도(year): 2021
- 분류(category): 구어 > 사적 대화 > 일상 대화
- 분석 층위(annotation_level): 원시
- 샘플링 방식(sampling): 본문 전체

2.2.3. 대화(document)

- 대화 아이디(id): 말뭉치 파일 아이디 + . + 1(예: SDRW2100000001.1)

2.2.4. 대화 메타 정보(document > metadata)

- 대화 제목(title): 2인 일상 대화
- 저작권자(author): 개인 발화자
- 발행자(publisher): 개인 발화 녹음
- 녹음일자(date): 연월일 YYYYMMDD
- 대화 주제(topic): 대화 주제, 제시 자료가 있을 때엔 제시 자료 파일명

2.2.5. 화자 정보(document > metadata > speaker)

- 화자 아이디(id): 화자 고유 아이디 부여(예: SD2000001), 대화가 다르더라도 화자가 동일하면 동일한 아이디 부여
- 연령(age): 10대/20대/30대/40대/50대/60대....
- 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

2.2.7. 발화 정보(document > utterance)

- 발화 아이디(id): 대화 아이디 + . + 1 + . + 1 + . + 발화 번호(예: SDRW2100000001.1.1.4)
- 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

- 발화는 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 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.

3.2. 화자 표시

- 화자 아이디, 성별, 연령, 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

- 기본 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 하며, 하나의 전사 단위가 3초 이상으로 길어지는 것을 지양한다.

  ※ 음성 정제본 하나가 하나의 전사 단위가 되도록 한다.

- 느낌표나 쉼표는 사용하지 않는다. 문장이 완전히 종결이 되었을 때는 마침표를 사용한다.

  ※ 도치된 문장의 경우 문장이 완전히 종결된 문장 성분(부사어, 목적어 등) 다음에 마침표를 넣는다 (예: 그 영화 봤어 지난달에.)

- 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분해 준다.(-어, -어요 등)
- 긴 쉼에 의해 나뉘는 경우는 통사적으로 완성이 되지 않았다 하더라도 구분하여 전사한다.

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
```

3.5. 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?
```

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

```
철자 전사: 어떡해
발음 전사: 어뜩해

철자 전사: 소주
발음 전사: 쐬주

철자 전사: 조금이라도
발음 전사: 쪼금이래도
```

- 발음 전사 시 약화 현상에 의한 이형태는 반영하지 않는다. 예를 들어 의문사 '뭐'가 '머'로 모음이 약화되어 들려도 이를 발음 전사에 반영하지 않고 '뭐'로 적는다.
- 발음 전사는 숫자나 기호, 영문 등도 발음에 따라 한글로 적는다.

```
철자 전사: 500원
```

발음 전사: 오백 원

철자 전사: 버스
발음 전사: 뻐쓰

철자 전사: 오리지널
발음 전사: 오리지날

3.6. 끊어진 단어(단어가 불완전하게 발화된 경우)

- 끊어진 단어는 발화된 대로 전사하되, 발음 전사를 할 때는 앞뒤에 '-'을 넣어 표시한다. 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

3.7. 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). 판단하기 어려운 경우에는 수시로 논의하여 결정한다(예: 오십대, 일 대 이 등).
- 본 용언과 보조용언도 띄어 쓴다.(예: 먹어 버리다, 가고 싶다, 먹지 못하다 등).

3.8. 축약형의 표기

- 구어에서는 발음의 축약 현상이 많이 나타나는데, 두 음절이 한 음절 사잇소리가 된다거나, 두 음절이 한 음절 겹홀소리가 되는 것 등이다. 일상 대화 말뭉치에서는 발음되는 음절수와 표기상의 음절수를 맞추는 것이 원칙이므로 축약형의 경우 모두 표기에 반영한다.

철자 전사: 그냥
발음 전사: 걍

철자 전사: 그러니까
발음 전사: 그니까

- 모음의 축약형의 경우 대부분 현재 국어의 모음 체계상 표기할 글자가 존재하지만, 반홀소리된 /ㅟ/, /ㅚ/의 표기는 문제가 된다. /ㅟ/, /ㅚ/가 반홀소리가 되어 /ㅓ/, /ㅏ/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

3.9. 담화 표지

- “이, 그, 저, 아, 어” 등 동일한 형태로 기존 품사의 의미, 기능을 가지지 않는 것은 담화표지로 보고, 물결표(~)를 이용하여 표시한다(주로 머뭇거림의 표지로 사용되는 이~, 그~, 저~, 어~, 아~, 에~ 등이 해당됨. 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 거하고 직결되는 과정이죠.
발음 전사: 많은 경우에 논문 그~ 어~ 연구는 네이션 국가라는 거하구 직결되는 과정이죠.

3.10. 잘 들리지 않는 부분

- 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: 너무한 거 같더라.
발음 전사: **(())** 너무한 거 같더라.

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

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

3.11. 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.**

3.12. 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. 단, 정치인 등 유명인의 이름은 비식별화하지 않으며, 상호명 및 상품명 등은 부정적인 경우에만 비식별화한다. 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다.

철자 전사: 신촌에 company-name는 진짜 맛없어.
발음 전사: 신촌에 **&company-name&**는 진짜 맛없어.

- 여러 이름이 나올 때는 일련번호를 붙여 구별할 수 있도록 한다(한 파일 내에서 지칭하는 대상이 일관성을 지녀야 함.).

철자 전사: 그때 **name1**이랑 **name2**이랑 너랑 나랑 갔잖아.
발음 전사: 그때 **&name1&**이랑 **&name2&**이랑 너랑 나랑 갔잖아.

- 비식별화 정보는 아래와 같이 마크업한다.

이름: &name&
상호명: &company-name&
계정(아이디): &account&
주민등록번호: &social-security-num&
전화번호: &tel-num&
카드 번호: &card-num&
기타 번호: &num&
주소: &address&
출신 및 소속: &affiliation&
기타 비식별화가 필요한 항목: &others&

3.13. 기타 지침

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

<Abstract>

2021 Construction of a Korean dialogue corpus

This project aims to build a large scale corpus of everyday Korean conversation by transcribing 1,000 hours of multi-speaker recordings. We expect the project to contribute towards Korean linguistics, speech processing, and language processing research. The main outcomes are as follows.

Speech recording and refinement: Based on Korean demographics, we collected daily and cooperative conversations from 2,599 speakers varying by region, gender and age. For daily conversations, the speakers were instructed to talk about one of the 15 topics selected in advance, supported by relevent newspaper articles. Similarly for cooperative discussions, one of the 8 preselected topics were presented along with for and against statements, newspaper articles, and publicly available videos. Each conversation is at maximum of around 15 minutes long and is held by 2 to 4 participants. To prevent the COVID-19 spread, all speakers were required to wear masks during conversation sessions. They were also asked to fill out a corpus license agreement form. The format of the collected and refined speech files is linear PCM with 16kHz sampling and 16bit quantization.

Voice data transcription: Experienced linguists, speech-language pathologists and qualified proofreaders verified and transcribed the conversation recordings. For all transcriptions, primary inspectors reviewed and corrected errors which were first filtered through semi-automatic natural language verification process. Then, the corrected transcriptions were handed over to secondary inspectors for the final review.

Construction of raw corpus and meta-information: We stored our transcriptions and their respective meta-information into JSON formatted files following required guidelines. Meta-information consists of information such as the conversation topic and its form, participating speakers' gender, age, hometown, and relationships between them.

Keywords: daily conversation corpus, raw corpus, cooperative dialogue, intonational phrase, voice data transcription

Project Director: Yigyu Hwang(MindsLab)

<기획·연구>
국립국어원 이승재 언어정보과장
국립국어원 이현주 학예연구사

<사업 참여자>
사업 책임자 황이규 (㈜마인즈랩)
사업 참여자 남선웅, 박지원, 이원문 (㈜마인즈랩)
박영훈, 이지현, 황주영, 조정아 (㈜나라지식정보)
윤기현, 최성봉, 이현복, 김민석, 박승홍, 김진수 (㈜바이칼AI)
김용운, 김진호, 정현학, 정승현 (㈜스마트미디어테크)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2022년 3월 23일
발행일: 2022년 3월 23일
인 쇄: 비즈카피

※ 이 책은 국립국어원의 용역비로 수행한 '2021년 일상 대화 말뭉치 구축' 사업의 결과물을 발간한 것입니다.

WWW.KOREAN.GO.KR

NATIONAL INSTITUTE OF KOREAN LANGUAGE

국립국어원