국립국어원 2020-01-02

발 간 등 록 번 호
11-1371028-000813-01

2020년 일상 대화 말뭉치 구축

사업 책임자
이 경 일

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2020년 일상 대화 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2020년 5월 ~ 2020년 12월

2020년 12월 6일

사업 책임자: 이 경 일 (㈜솔트룩스)

사업 수행자 ㈜솔트룩스

사업 책임자 이경일

사업 참여자 김예하나, 주재현,
김창환, 김진희 외

〈사업 수행자〉 ㈜솔트룩스 컨소시엄

사업 책임자	이경일((주)솔트룩스)
사업 참여자	김예하나((주)솔트룩스)
	김지영((주)솔트룩스)
	박선희((주)솔트룩스)
	이동기((주)솔트룩스)
	주재현((주)솔트룩스)
	강상규((주)소리자바)
	윤종후((주)소리자바)
	김창환((주)소리자바)
	김동희((주)소리자바)
	노희균((주)소리자바)
	정용직((주)소리자바)
	박선욱((주)소리자바)
	김진희((주)소리자바)

<국문 초록>

2020년 일상 대화 말뭉치 구축

본 사업은 일상 대화 말뭉치 구축 사업으로 화자 모집 지침과 말뭉치 구축 지침에 따라 정제본 500시간 규모의 말뭉치를 구축하여 국어 자원의 활용도와 가치를 높일 수 있도록 국어 말뭉치 확대를 위한 기초 자료를 마련하고자 하는 데 그 목적이 있다. 이에 따른 주요 과업과 사업의 성과는 다음과 같다.

음성 녹음 및 정제: 지역별, 성별, 연령별 다양한 화자를 모집하여 총 2,739명의 화자가 15분 동안 15개의 주제와 13개의 신문 기사 자료로 자연스러운 대화를 녹음하였다. 녹음한 화자 전체를 대상으로 이용 허락 계약을 체결하였고, 코로나 감염을 예방하기 위해 아크릴판으로 자리를 구분하거나 마스크를 착용한 채로 녹음하였다. 대화 주제와 무관한 대화(예: 인사말 등)는 제외하여 정제하였고, 음성 파일은 16kHz 표본화, 16bit 양자화 선형 PCM으로 저장하였다.

음성 자료 전사: 유사 사업 경험이 있는 전문 속기사를 선발하여 전사 지침 교육을 진행한 후 음성 자료 전사를 시작하였다. 전사 도구(TranscriberAG)를 사용해 전문 속기사가 발화자 표시, 전사 단위, 맞춤법, 띄어쓰기, 이중 전사 등을 전사 지침에 맞게 전사하였다. 1차 전사한 내용을 대상으로 전수 검수를 하고 오류가 있는 내용은 재작업을 통해 수정 반영하였다.

원시 말뭉치 구축 및 메타 정보 구축: 음성 파일의 대화 주제를 대분류, 소분류로 나누어 기록하고 화자 정보(성별, 연령대, 주 성장지 등), 화자 간 관계 등을 메타 정보로 저장하여 부착하였다. 전사 단위로 마크업하여 지침에 맞게 JSON 형식으로 변환하였다.

원시 말뭉치 활용: 원시 말뭉치를 음성 인지 엔진과 언어 인지 엔진에 학습 데이터 형태로 활용하여, 구축된 원시 말뭉치의 활용 사례 및 방향성을 제시한다.

주요어: 일상 대화 말뭉치, 원시 말뭉치, 음성 수집, 음성 녹음, 음성 자료 전사, 원시 말뭉치 활용

차 례

제1장 사업 개요

제2장 사업 수행

제3장 사업 수행 결과

표 차례

그림 차례

제 1 장

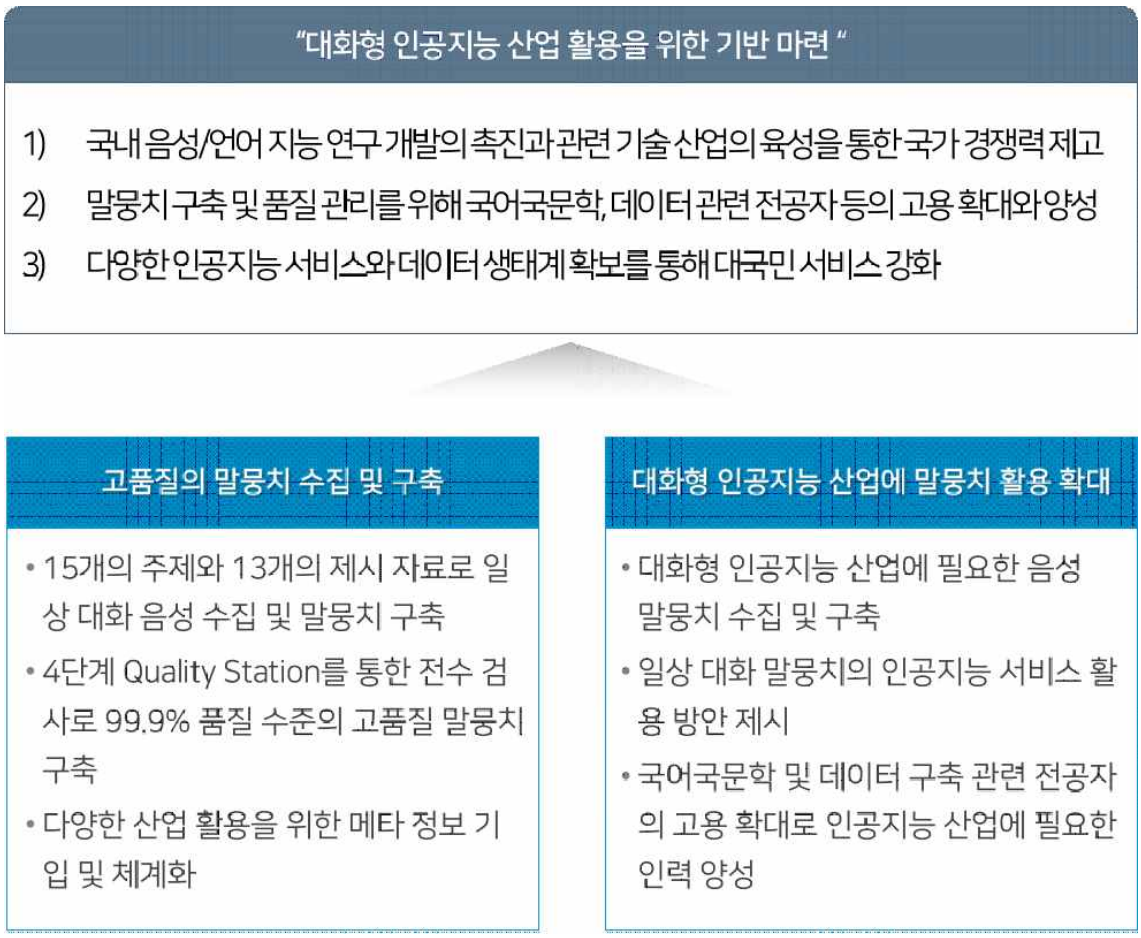
사업 개요

1. 사업 목적

대화형 인공지능 기술을 개발하고 활용하기 위해서는 많은 양의 대화형 말뭉치를 확보하는 것이 중요하며, 해외 선도국 및 글로벌 기업에서는 많은 비용을 투자하여 대화 말뭉치를 구축하고 학습하여 대화형 인공지능 서비스를 제공하고 있다.

세계 대화형 인공지능 시장에 비해 국내 시장은 한국어 대화형 말뭉치가 부족하여 기술 개발이나 서비스화가 어렵다.

본 사업은 두 사람이 특정 주제 또는 제시된 자료에 대해 자유롭게 대화하는 음성을 녹음하고 정제하는 사업으로 음성 자료의 이중 전사를 통해 메타 정보가 있는 원시 말뭉치를 구축하여 국내 대화형 인공지능 산업을 위한 핵심 자산을 확보하는 데 목적이 있다.



[그림 1] 일상 대화 말뭉치 구축 사업 기대 효과

2. 사업 수행 범위

본 사업은 대화형 인공지능 산업 발전에 필요한 일상 대화 말뭉치를 구축하는 것으로 사업의 수행 범위는 크게 세 부분으로 나눌 수 있다. 첫째는 다양한 주제로 대화할 수 있도록 대화 주제와 제시 자료를 선정하는 것이다. 둘째는 여러 방면에서 활용이 가능한 유용한 말뭉치 구축을 위해 2,000명 이상의 화자를 모집하여 다양한 연령과 지역 화자의 대화를 수집하고 화자의 신분 보장을 위해 개인 정보는 비식별화하는 것이다. 마지막으로 고품질 말뭉치 확보를 위해 발음 전사와 철자 전사를 병행하여 전사하고 전체 파일을 대상으로 품질 검증을 진행하는 것이다.

[표 1] 사업의 범위

사업의 범위	세부 내용	분량
주제 및 제시 자료 선정	• 전문가 검토를 통해 2019년도 사업 주제 외에 신규 주제 및 제시 자료 선정	• 총 15개의 주제 및 13개의 제시 자료 선정
음성 녹음 및 정제	• 화자별 최대 녹음 시간은 30분(2개 주제)으로 제한 • 개인 정보 비식별화	• 2,000명 이상의 화자 참여 • 정제 후 500시간 음성 수집
음성 자료 전사	• 발음 전사와 철자 전사를 병행 • 지침에 맞게 전사하였는지 전수 검수	• 음성 자료 500시간

특히 화자의 성별, 연령, 지역이 편중되지 않도록 지역별 최소 할당 인원과 권역별 할당 인원 목표를 세웠으며, 다양한 화자를 모집하기 위해 한 화자당 최대 녹음 시간은 30분으로 제한하였다.

사업의 주요 내용은 다음과 같다.

○ 두 사람이 특정 주제로 자유롭게 대화
○ 대화 내용 녹음 및 정제(정제 후 500시간, 대화당 15분 이하)
○ 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

3. 사업 수행 절차

한국정보화진흥원의 데이터베이스 구축 방법론(Ver.4)을 적용하여 음성 녹음, 이중 전사, 원시 말뭉치 구축에 대한 대상 자료별 과업 공정과 주요 활동 절차를 표준화하여 효율적인 말뭉치 구축 체계를 확보하였다. 일상 대화 말뭉치 구축에 적합하도록 자료의 특성을 고려하여 전체 공정을 설계하고 수행하였다.

[그림 2] 일상 대화 말뭉치 구축 전체 공정도

제 2 장

사업 수행

1. 대화 주제 및 제시 자료 선정

대화의 주제는 쉽고 편한 것으로 화자가 흥미를 가지고 대화할 수 있는 15개를 최종 선정하였다. 제시 자료는 국립국어원 신문 말뭉치 중 기사 13개를 선정하여 50쌍 내외가 하나의 기사로 대화할 수 있도록 제시하였다.

대화 주제는 화자가 직접 선택하도록 했으며, 주제에 대한 이해를 돕고자 세부 주제를 예시로 들어 주제 선택이 용이하도록 하였다.

[표 2] 대화 주제 및 세부 예시 주제

주제	세부 예시 주제
스포츠/레저	종목, 운동선수, 올림픽, 경기 관람 등
여행지(국내/해외)	장소(나라, 지역), 관광 명소, 여행 계획, 경험 등
계절/날씨	봄, 여름, 가을, 겨울, 추억 등
회사/학교	재직(재학) 중인 곳, 학창 시절, 동창, 선생님, 동아리 등
먹거리	음식, 맛집, 요리법, 요리사 등
방송/연예	연예인, 프로그램, 이슈 등
영화	영화인, 영화관, 영화제, 영화 장르 등
건강/다이어트	질병, 약, 건강 보조제, 건강 관리, 약물 부작용 등
선물	추억, 종류, 이벤트, 핸드메이드 등
꿈(목표)	꿈(과거, 금년), 장래 희망 등
연애/결혼	이상형, 데이트, 배우자, 연애관, 자녀 등
반려동물	추억, 반려동물 종류, 동물 이름, 질병 등
아르바이트	종류, 추천, 경험 등
성격	혈액형, 다혈질, 소심함 등
가족	가족 관계, 형제, 자매 등

[표 3] 자료 제시 목록

국립국어원 신문 말뭉치 아이디	기사
NWRW1900000040.15446	경향신문 “서촌 족발집 사장은 어쩌다 망치를 휘둘러 ‘살인미수범’이 됐나… 영업권보다 재산권, 법이 빚은 ‘비극’”
NWRW1900000040.15396	경향비즈 “전복 산지값 폭락에도 소비자값은 ‘찔끔’ 왜?”
NWRW1900000040.4066	경향신문 “역시 세계 최강…한국 女 쇼트트랙 3000m 계주, 넘어지고도 올림픽 신기록”
NWRW1900000040.9893	경향신문 “폐지 줍는 노인 절반, 월 10만원도 못 번다”
NWRW1900000060.15333	한겨레 “일본 지역사회의 치매 끌어안기…조금 실패해도 괜찮지 않나요”
NWRW1900000060.18354	한겨레 “‘제2의 장현수’들…예술·체육요원 절반이 ‘허위 봉사활동’”
NWRW1900000020.18519	동아일보 “김정은, 트럼프에 4번째 친서… ‘핵리스트 제출’ 제안 가능성”
NWRW1900000020.21711	동아일보 “난민 문제, 불편하지만 우리사회 문화적 체질 강화시킬 수 있어”
NWRW1900000020.24874	동아일보 “대만 징병제 67년만에 역사속으로”
NWRW1900000010.21763	조선일보 “무상복지로 경제 몰락한 브라질… '극우 포퓰리스트' 불러냈다”
NWRW1900000010.21633	조선일보 “경제 점수는 빵점… 이대로면 몰락할 것"”
NWRW1900000010.26464	조선일보 “[아무튼, 주말] 비주얼이 예술이네요… 눈에서 살살 녹는 송년 케이크”
NWRW1900000010.25960	조선일보 “反韓 정서 확산, 객관적 역사 연구 힘들게 해”

2. 화자 구성 및 모집

모집 대상은 전국 17개 시·도에 거주하는 성인 남녀로, 성별, 연령, 지역 등의 비율이 편중되지 않도록 사전에 참가자 할당표를 설계해 모집하였다. 해당 분류는 2020년 3월 기준 행정안전부 주민등록 인구통계를 기준으로 분류하였다. 지역은 현 거주지가 아닌 주 성장지 기준으로 17개 지역(서울, 인천, 대전, 대구, 부산, 광주, 울산, 경기, 강원, 충남, 충북, 경남, 경북, 전남, 전북, 세종, 제주)으로 할당하였으며, 연령은 10세 단위로 분할하여 할당하였다. 현실적으로 녹음이 어려울 것으로 예상되는 0~9세, 70세 이상은 대상자에서 제외하였다. 2012년에 출범한 세종시는 주 성장지로 하는 대상자를 찾기 어려워 할당 인원을 축소하였다.

화자 모집에 있어 최대한 인구 비율에 맞도록 진행하고자 했으나, 대상을 찾기가 어려운 경우에는 지역은 권역 전체로 할당을 맞추고 각 지역별로는 기본 할당량의 최소 50%를 맞추어 진행하였다. 성별×연령별 할당은 남성, 여성 10대와 50대 이상은 기본 할당량의 최소 50%를 맞추어 진행하였다.

[표 4] 화자 할당표 설계 기준

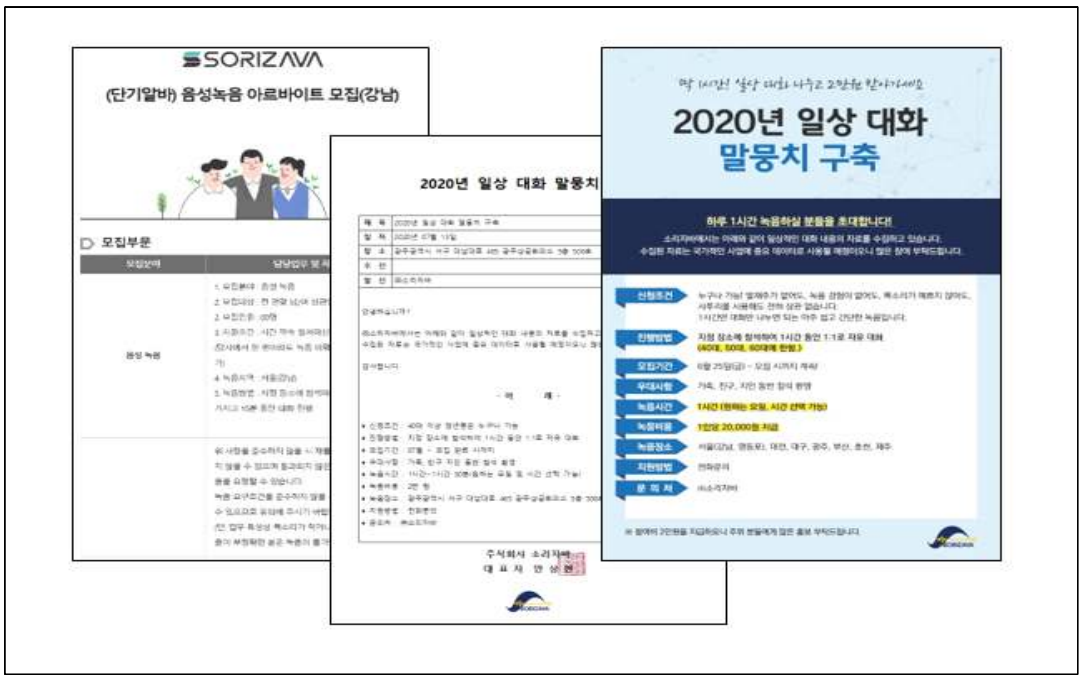
구분	기준
모집단	• 행정안전부, 2020년 3월 기준 주민등록 인구통계 기준 활용
고려 변수	• 성별: 남자/여자 • 연령대: 10대/20대/30대/40대/50대/60대 • 지역(주 성장지 기준): 서울/인천/대전/대구/부산/광주/울산/경기/강원/충남/충북/경남/경북/전남/전북/세종/제주 * 세종시의 경우, 2012년 출범한 세종시를 주 성장지로 하는 대상자를 찾기 어려워 할당 인원을 축소함.
배분 방법	• 제곱근 비례 배분
표본 할당	• 지역별: 비례 할당 • 성별×연령별: 균등 할당

[표 5] 성×연령×지역별 화자 모집 할당표

단위: 명		남성						여성						합계		
		10대	20대	30대	40대	50대	60대	10대	20대	30대	40대	50대	60대	할당	기본 할당	최소 할당
수도권	서울	25	35	35	45	40	30	25	35	35	45	40	30	1,050	420	210
	경기	33	38	41	47	54	45	33	38	41	45	52	43		510	255
	인천	6	9	11	12	12	10	6	9	11	12	12	10		120	60
충청권	대전	4	6	7	7	6	6	4	6	7	7	6	6	230	72	36
	세종	2	2	0	0	0	0	2	2	0	0	0	0		8	4
	충북	2	3	5	9	8	8	2	3	5	9	8	8		70	35
	충남	3	5	6	9	10	7	3	5	6	9	10	7		80	40
영남권	대구	8	7	7	10	10	8	8	7	7	10	10	8	520	100	50
	경북	6	9	9	13	13	10	6	9	9	13	13	10		120	60
	부산	7	10	13	15	15	15	7	10	13	15	15	15		150	75
	경남	4	9	9	11	14	8	4	9	9	11	14	8		110	55
	울산	3	4	3	4	3	3	3	4	3	4	3	3		40	20
호남권	광주	3	4	5	5	5	3	3	4	5	5	5	3	190	50	25
	전북	3	4	7	9	7	5	3	4	7	9	7	5		70	35
	전남	2	4	6	8	8	7	2	4	6	8	8	7		70	35
강원권	강원	3	5	4	7	5	6	3	5	4	7	5	6	60	60	30
제주권	제주	4	4	4	5	5	4	4	4	4	5	4	3	50	50	25
합계		118	158	172	216	215	175	118	158	172	214	212	172	2,100	2,100	1,050

대규모 화자 모집을 위해 주로 사업 수행 기관에서 자체 보유하고 있는 온라인 패널을 활용하였다. 사업 수행 기관의 홈페이지, 구인 공고 사이트에 일상 대화 말뭉치 구축 참여자 모집 공고를 게시하여 적합한 대상자를 선정하였다. 화자 모집 시 2인 1조 신청자를 최우선으로 하였으며, 1인이 개별 신청했을 경우 비슷한 연령대 및 관심사를 구분하여 조 편성하였다. 이렇게 1차 모집된 화자를 녹음 진행 요원이 전화 통화를 통해 적합한 대상자가 맞는지 다시 한번 확인하고, 녹음 가능한 날짜를 협의해 최종 대상자로 선정하였다. 기타 방법으로는 각 복지 기관에 공문을 보내 직접 모집하거나 지인 추천, 지역 커뮤니티 공지 등의 다양한 방법으로 화자를 모집하였다.

녹음이 진행될수록 성별×연령별×지역별 할당 외에 주제 할당까지 맞는 화자를 찾는 것이 쉽지 않았다. 특히, 10대, 50~60대, 일부 시 · 도 지역 거주자, 일부 주제는 모집이 어려워 해당 지역에서 직접 추천을 받는 방식으로 모집하였다. 일부 화자는 녹음 전 확인 전화 시 녹음 일정을 일방적으로 취소하거나 연락이 두절되기도 하였으며, 녹음 당일 녹음 장소에 오지 않는 등의 이유로 화자 모집에 어려움을 겪기도 했다.



[그림 4] 온라인 패널 대상 모집 공고

3. 작업자 선발 및 교육

3.1. 녹음 진행 요원 선발 및 교육

녹음에는 서울, 대전, 대구, 부산, 광주, 강원, 제주 등 총 17개 지역 거주자 2,000명 이상이 참여하였다. 다수의 화자가 참여하는 만큼 원활한 녹음 진행을 위하여 각 지역별로 녹음 진행 요원이 투입되었다. 지역별로 투입된 진행 요원은 총 11명으로, 자격 조건에 맞는 인원 중 오프라인 교육을 이수하고 사업 목적에 맞는 시뮬레이션 평가를 통과한 자를 최종 선발하여 녹음 파일의 품질 유지와 녹음 일정에 차질이 없도록 하였다. 진행 요원의 선발 기준은 유사 작업 경험자 중 전문 녹음 장비 작동 경험이 있는 자를 우선 선발하였다.

[표 6] 진행 요원 선발

구분	선발 기준 및 운영 내용	
선발 기준	• 전문 녹음 장비 작동 경험이 있는 자 • 최종 교육 이수 및 평가 통과자	
투입 인원	• 진행 요원 11명	
진행 요원 역할	• 진행 요원 1 - 화자 안내 - 화자 인적사항 확인 - 화자 참석 관리 - 녹음 종료 후 사례비 지급	• 진행 요원 2 - 녹음 진행 개요 설명 - 저작권 이용 허락 계약 체결 - 녹음 장비 작동 - 주제 이탈 및 녹음 관리

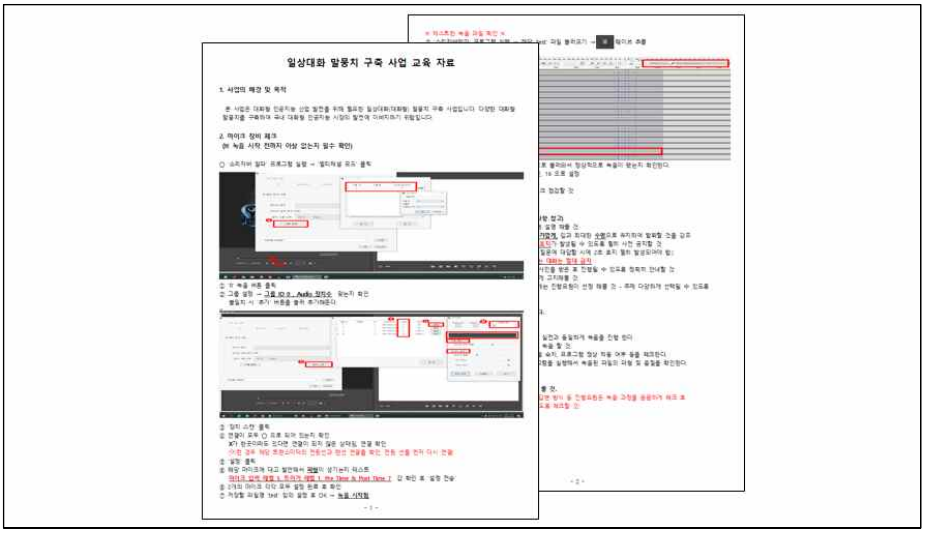
녹음 진행 요원 및 관리 인원을 대상으로 교육을 수행하였다. 교육은 기본 4단계로 진행하였으며, 교육 내용은 사업 배경 및 목적, 진행 시 유의 사항 등의 이론 교육, 녹음 장비 작동 방법, 헤드셋 마이크 착용 방법, 녹음 진행 등의 실사 교육, 화자 응대, 화자의 불만 제기 시 대처 방법, 화자의 일정 변동 시 대처 방법 등의 CS 교육, 화자 개인 정보 관리, 녹음 자료 관리 등의 보안 교육으로 나누어 진행하였다.

기본 교육을 마친 진행 요원은 실제 녹음으로 들어가기에 앞서 화자 응대, 녹음 장비 작동 등 실제와 같은 상황 시뮬레이션과 빈번하게 일어날 만한 특이 사항 발생 시 대처 요령 등을 시뮬레이션으로 실시하였다. 시뮬레이션 평가에서 역할에 대한 이해도가 높은 자를 최종적으로 선발하였고, 적정 수준 미달인 자는 부족한 점에 대한 피드백을 통해 다시 평가를 거친 후 선발하였다.

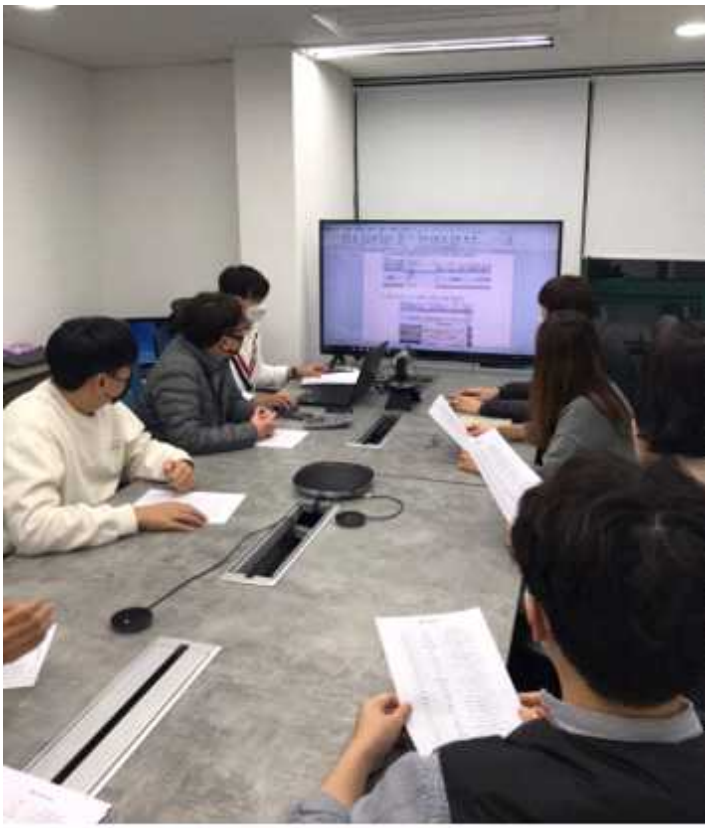
이러한 과정을 통해 최종 선발된 녹음 진행 요원들은 보안 서약서 작성 후 실제 녹음 진행에 참여하였다.

[표 7] 진행 요원 교육

구분	내용
교육 일시 및 장소	• 2020년 6월 17일 소리자바 3층 회의실
교육자	• 장윤우(㈜소리자바)
교육 내용	• 사업의 배경 및 목적 • 진행 절차 • 대화 주제 • 녹음 환경 및 녹음 장비 사용법 • 녹음 방법 • 녹음 시 주의 사항 • 시뮬레이션 실습 • 보안 교육 • 질의 응답



[그림 3] 녹음 진행 요원 교육 자료 일부



[그림 4] 진행 요원 교육

3.2. 전사자 선발 및 교육

본 사업에는 전체 2,000명 이상의 화자가 참여하여 녹음한 대화 음성 500시간(정제 기준)을 전사하는 과업이 포함되어 있으므로 다수의 전사자가 투입되었다. 전사자 선발 기준은 유사 작업 경험자, 전사 지침에 대해 정확한 이해를 하고 있는 전문 속기사를 우선 선발하였다.

원활한 전사를 위해 약 20명의 전사자를 교육하였다. 억양구 기준으로 구분하는 전사 단위와 전사 지침의 이해, 전사 완성도가 전사자 간에 차이가 생길 수 있어 전체 전사자를 대상으로 오프라인 교육을 실시하였다. 전사자 교육 내용은 전사 지침과 유의 사항, 한글 맞춤법 위주로 진행되었으며, 부수적으로 사업 배경 및 목적, 전사 절차 등을 교육하여 사업 목적에 맞게 작업에 임할 수 있도록 하였다.

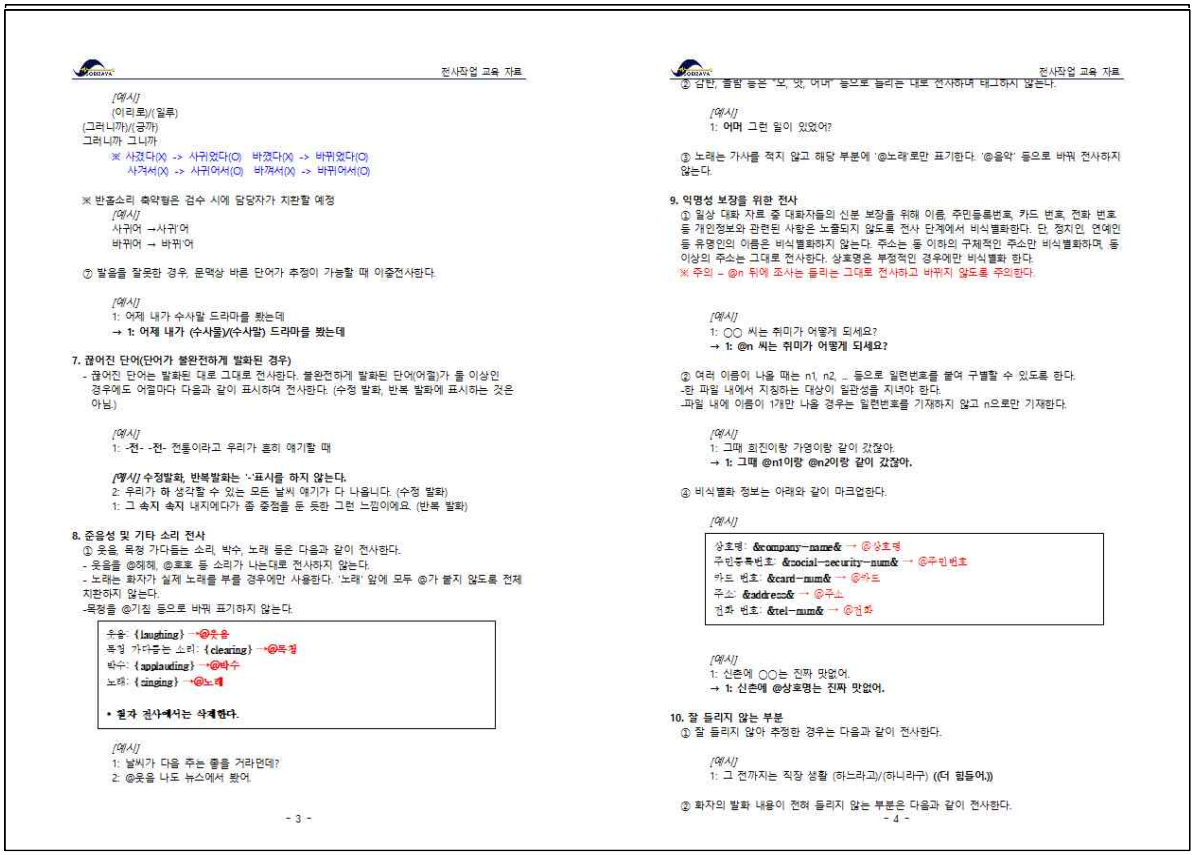
이러한 과정을 통해 최종 선발된 전사자들은 보안 서약서를 작성한 후 전사에 참여하였다. 교육을 마친 전사자는 샘플 전사 후 1차 결과물에 대해 교정을 받고, 두세 차례 수정 전사 후 충분히 전사 지침을 숙지한 상태에서 본 전사에 투입되었다. 하루에 1인당 약 15분 녹음 파일 4개를 전사하는 것을 기준으로 일평균 12명 정도가 투입되었다.

[표 8] 전사자 선발

구분	선발 기준 및 운영 내용
선발 기준	• 유사 작업 경험자 • 전사 지침에 대해 정확히 이해를 하고 있는 자 • 전사 교육 이수 및 샘플 평가 통과한 자
투입 인원	• 일평균 12명 정도 진행
운영	• 음성 파일 중 방언 등의 특징이 있을 경우 해당 지역 출신 전사자에게 우선 배정함.

[표 9] 전사자 교육

구분	내용
교육 일시 및 장소	• 2020년 6월 17일 소리자바 2층 전사실
교육자	• 김응준(㈜소리자바)
교육 내용	• 사업의 배경 및 목적, 전사 절차와 방법 • 전사 사용 도구(TranscriberAG) • 전사 지침 및 유의 사항 • 한글 맞춤법 주요 내용 • 보안 교육 • 질의 응답



전사작업 교육 자료

[예시]
(이리로)/(일루)
(그러니까)/(긍까)
그러니까 그니까

※ 사겼다(X) -> 사귀었다(O) 바꼈다(X) -> 바뀌었다(O)
사겨서(X) -> 사귀어서(O) 바껴서(X) -> 바뀌어서(O)

※ 반줌소리 축약형은 검수 시에 담당자가 치환할 예정

[예시]
사귀어 →사귀'어
바뀌어 → 바뀌'어

⑦ 발음을 잘못한 경우, 문맥상 바른 단어가 추정이 가능할 때 이중전사한다.

[예시]
1: 어제 내가 수사말 드라마를 봤는데
→ 1: 어제 내가 (수사물)/(수사말) 드라마를 봤는데

7. 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.)

[예시]
1: -전- -전- 전통이라고 우리가 흔히 얘기할 때

***[예시]* 수정발화, 반복발화는 '-'표시를 하지 않는다.**
2: 우리가 하 생각할 수 있는 모든 날씨 얘기가 다 나옵니다. (수정 발화)
1: 그 속지 속지 내지에다가 좀 중점을 둔 듯한 그런 느낌이에요. (반복 발화)

8. 준음성 및 기타 소리 전사
① 웃음, 목청 가다듬는 소리, 박수, 노래 등은 다음과 같이 전사한다.
- 웃음을 @하해, @호호 등 소리가 나는대로 전사하지 않는다.
- 노래는 화자가 실제 노래를 부를 경우에만 사용한다. '노래' 앞에 모두 @가 붙지 않도록 전체 치환하지 않는다.
-목청을 @기침 등으로 바꿔 표기하지 않는다.

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

• 원자 전사에서는 삭제한다.

[예시]
1: 날씨가 다음 주는 좋을 거라면데?
2: @웃음 나도 뉴스에서 봤어.

- 3 -

전사작업 교육 자료

② 감탄, 놀람 등은 "오, 앗, 어머" 등으로 들리는 대로 전사하며 태그하지 않는다.

[예시]
1: **어머 그런 일이 있었어?**

③ 노래는 가사를 적지 않고 해당 부분에 '@노래'로만 표기한다. '@음악' 등으로 바꿔 전사하지 않는다.

9. 익명성 보장을 위한 전사
① 일상 대화 자료 중 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. 단, 정치인, 연예인 등 유명인의 이름은 비식별화하지 않는다. 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다. 상호명은 부정적인 경우에만 비식별화 한다.
※ 주의 – @n 뒤에 조사는 들리는 그대로 전사하고 바꾸지 않도록 주의한다.

[예시]
1: ○○ 씨는 취미가 어떻게 되세요?
→ 1: @n 씨는 취미가 어떻게 되세요?

② 여러 이름이 나올 때는 n1, n2, … 등으로 일련번호를 붙여 구별할 수 있도록 한다.
-한 파일 내에서 지칭하는 대상이 일관성을 지녀야 한다.
-파일 내에 이름이 1개만 나올 경우는 일련번호를 기재하지 않고 n으로만 기재한다.

[예시]
1: 그때 희진이랑 가영이랑 같이 갔잖아.
→ 1: 그때 @n1이랑 @n2이랑 같이 갔잖아.

④ 비식별화 정보는 아래와 같이 마크업한다.

[예시]

상호명: &company-name& → @상호명
주민등록번호: &social-security-num& → @주민번호
카드 번호: &card-num& → @카드
주소: &address& → @주소
전화 번호: &tel-num& → @전화

[예시]
1: 신촌에 ○○는 진짜 맛없어.
→ 1: 신촌에 @상호명는 진짜 맛없어.

10. 잘 들리지 않는 부분
① 잘 들리지 않아 추정한 경우는 다음과 같이 전사한다.

[예시]
1: 그 전까지는 직장 생활 (하느라고)/(하니라구) **((더 힘들어.))**

② 화자의 발화 내용이 전혀 들리지 않는 부분은 다음과 같이 전사한다.

- 4 -

[그림 6] 전사 교육 자료 일부

[그림 7] 전사자 교육

3.3. 보안 교육

보안 교육은 사업에 참여한 관리자와 진행 요원, 전사자, 검수자 등 모든 사업 참여자를 대상으로 하였다. 녹음 진행 요원과 전사자는 작업 교육과 함께 실시하였고, 관리자와 검수자 등의 인력은 별도의 시간을 마련하여 교육하였다. 교육 내용은 크게 네 가지로 개인 정보 보호 교육, 자료에 대한 보안 관리 교육, 사무실·장비에 대한 보안 관리 교육, 내·외부망 접근 시 보안 관리 교육이다.

[표 10] 보안 교육 내용

구분	보안 교육 내용
개인 정보 보호	• 사업 진행 중 알게 되는 화자의 인적 사항에 대한 비밀 보장 • 화자의 대화 청취 중 알게 되는 개인 사생활 및 사적 의견에 대한 비밀 보장
자료에 대한 보안 관리	• 누출금지 대상 정보는 반드시 자료 관리 대장에 인계자·인수자가 직접 서명하여 관리 • 생산되는 모든 산출물은 보안 담당관이 지정한 PC에만 저장·관리하고 비인가자에게 제공·대여·열람 금지 • 인터넷 자료 공유 사이트 및 상용 메신저 사용으로 인한 해킹 위험 방지 • 퇴근 시 비공개 자료는 반납하고 그 외 자료는 사무실 시건 장치가 된 보관함에 보관
사무실·장비에 대한 보안 관리	• CCTV·시건 장치 등 비인가자의 출입통제 • PC는 패스워드를 설정하고 상시 점검 및 악성코드 감염 차단을 위한 저장 매체 자동 점검이 될 수 있도록 설정 • 인가된 USB 및 휴대용 저장 매체만 사용 가능
내·외부망 접근 시 보안 관리	• 용역 업체 사용 전산망은 방화벽 등을 활용, 주관 기관 업무망과 분리 구성, 업무상 필요 시 제한적 접근 허용 • 웹 시스템은 관리자에 의해 인증된 사람만 접근 가능하며 로그인 후 일정 시간 사용하지 않을 시 보안을 위해 자동 로그아웃 • 참여 인원에게 부여한 패스워드는 별도로 관리하고 수시로 내부 서버 및 네트워크 장비에 대한 접근 기록 확인

[그림 8] 보안 교육

정보 보안 교육 자료

솔트룩스 컨소시엄

2020.09.18

□ 보안 정책 및 지침 준수

◦ 참고자료

- 국가 정보보안 기본지침(국가정보원, 2019. 3.)
- 국가공공기관 용역업체 보안관리 가이드라인(국가정보원, 2014. 3.)
- 문화체육관광부 개인정보보호지침(훈령 제342호, 2018.6.11.)
- 본 사업은 위 지침 외에도 정부가 제정·공포한 관계 제 법규(지침)을 준수하여야 하며, 사업기간동안 법규가 변경될 경우 해당 법규 준수
- 용역사업 중 또는 종료 후라도 본 사업과 관련하여 국가정보원 등 관련 기관으로부터 보안에 문제가 있다고 지적될 경우 반드시 해결책을 강구하고 조치하여야 함.

※ 용역업체가 정보누출 적발 시 「국가계약법」 시행령 제76조에 근거, 해당업체를 부정당업자로 등록, 입찰 참가자격 제한 등 제재조치 가능

□ 참여인원에 대한 보안관리

◦ 용역사업 참여인원은 개인의 친필 서명이 들어간 보안서약서 및 개인정보보호법 준수를 위한 개인정보 처리위탁 계약서, 개인정보 위탁 보안서약서 제출

◦ 용역사업 수행 전 참여인원에 대해 법적 또는 주관기관 규정에 따른 비밀유지 의무 준수 및 위반 시 처벌내용, 누출금지 대상정보 및 정보누출 시 부정당업자 제재조치 등에 대한 보안교육 실시

◦ 사업수행 중 주관기관의 정기적인 보안점검에 성실히 응하여야 함.

□ 자료에 대한 보안관리

◦ '누출금지 대상정보'는 반드시 '자료관리 대장'에 인계자·인수자가 직접 서명하여 관리하고 사업완료 시 관련 자료 회수

◦ 사업수행에서 생산되는 모든 산출물은 파일서버 또는 보안담당관이 지정한 PC에만 저장·관리하고 사업담당자가 인가하지 않은 비인가자에게 제공·대여·열람 금지

◦ 주관기관의 보안정책에 P2P, 웹하드 등 인터넷 자료공유사이트 및 상용 메일·메신저 사용을 금지하고 자료전송이 필요한 경우 자체 전자우편을 이용, 첨부자료를 암호화 후 수발신(다만, 대외비 이상의 비밀은 전자우편 송수신 금지)

◦ 매일 퇴근 시 주관기관이 제공한 비공개 자료는 반납하고 그 외 자료는 사무실 시건장치가 된 보관함에 보관

□ 사무실·장비에 대한 보안관리

◦ 용역사업 수행 장소는 CCTV·시건장치 등 비인가자의 출입통제 대책 마련된 외부 사무실을 사용

◦ 상주인력에 대해서는 정보보안 S/W(바이러스 백신, 보조기억매체제어 등) 설치를 통해 우리 기관 직원과 동일한 수준의 보안 통제를 적용

◦ PC는 부팅 패스워드, 운영체제 패스워드, 화면보호기 패스워드를 설정해야 하며, 공유폴더 감시, 운영체제 업데이트, 백신 최신패치 등 상시점검 및 악성코드 감염 차단을 위한 저장매체 자동 점검이 될 수 있도록 설정

◦ 노트북 등 관련 장비를 외부에 반출·입시 악성코드 감염여부 및 자료 무단반출 여부를 확인하고 관리대장에 반드시 기록

◦ 인가받지 않은 USB 등의 휴대용 저장매체 사용을 금지하고 산출물 저장을 위해 휴대용 저장매체가 필요한 경우 보안담당관의 승인하 사용

◦ 용역사업자는 수행 장소에 대한 보안점검을 정기적으로 실시하여 사업 담당자에게 결과를 보고

□ 내·외부망 접근 시 보안관리

◦ 용역업체 사용 전산망은 방화벽 등을 활용, 주관기관 업무망과 분리구성하며, 업무상 필요 시 제한적 접근 허용

◦ 용역사업 수행 시 내부 전산망 이용이 필요한 경우

▸ 사업 참여인원에 대한 사용자계정은 하나의 그룹으로 등록하고 계정별로 정보시스템 접근권한을 차등 부여하되 기관 내부문서 접근 금지

▸ 계정별로 부여된 접속권한은 불필요 시 곧바로 권한을 해지하거나 계정을 폐기

▸ 참여인원에게 부여한 패스워드는 보안담당관이 별도로 기록 관리하고 수시로 해당 계정에 접속하여 저장된 자료와 작업이력 확인

▸ 보안담당관은 서버 및 장비 운영자로 하여금 내부서버 및 네트워크 장비에 대한 접

[그림 9] 보안 교육 자료 일부

4. 음성 녹음

4.1. 녹음 환경

녹음은 전국 7개 지역(서울, 대전, 대구, 부산, 광주, 강원, 제주)에서 동시에 진행되었으며, 외부와 차단된 상태로 대화에 참여한 두 명만이 대화할 수 있도록 구성하였고, 녹음에 임하는 두 명의 화자가 편안하게 이야기할 수 있는 사무실 환경을 마련하여 상대방의 목소리가 들어가지 않도록 화자 간 거리가 1m 이상 떨어진 공간에서 녹음을 진행하였다. 장소가 여의치 않은 경우 녹음 기간 동안 유료 회의실이나 별도 공간을 대여해서 녹음을 진행하였다.

[그림 10] 지역별 녹음실

특히, 코로나-19 집단 감염 방지를 위해 녹음 시 화자의 체온을 측정하고 호흡기 증상을 확인하였다. 손 소독제를 녹음실에 비치하였으며, 화자는 마스크를 쓰고 녹음하거나 아크릴판으로 구분된 좌석에서 녹음을 하였다. 또한, 참여자별 녹음 시간을 조정하여 화자 집단별로 마주치지 않도록 하고 모집과 교육은 최대한 비대면으로 진행할 수 있도록 하는 등 감염 방지에 각별히 유의하였다. 녹음 종료 시까지 본 사업 참여로 인한 감염은 발생하지 않았다.

[그림 11] 녹음실 환경

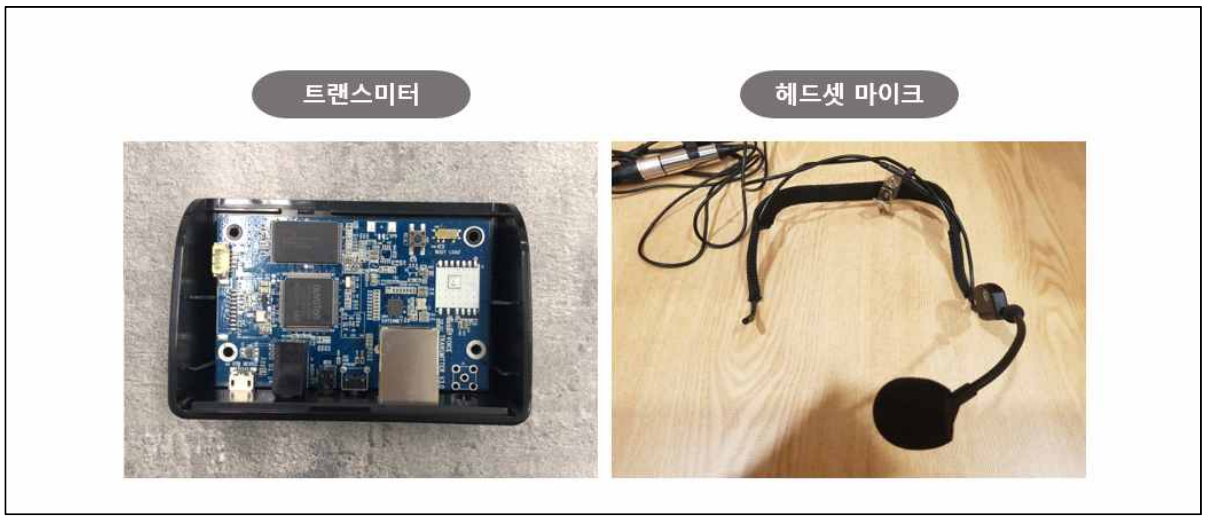
[표 11] 코로나-19 집단 감염 방지 화자 관리 방안

구분	내용
대책 분야	• 화자 모집, 음성 녹음, 참여자 교육, 회의 등 사업 전반
화자 모집	• 전화, 온라인 접수
교육	• 전화, 온라인 교육
녹음 시간	• 참여자별 녹음 시간 조정
방문자	• 방문자 체온 검사, 호흡기 증상 확인
녹음실	• 녹음실 내 화자 간격 조정
방역 관련	• 소독제 및 마이크 1회용 덮개 등 방역 관련 물품 사용 • 사업장 전체 주기적 환경 소독, 환기 실시, 감염 관리 전담 직원 지정
인력 관리	• 방문자 및 종사자 목록 관리 • 유증상자 출근, 이용 중단 및 업무 배제

녹음은 2채널로 진행되었으며, 두 명의 화자는 각자 헤드셋 마이크를 착용한 후 녹음에 참여하였다. 이때 녹음 음성의 최대 샘플값이 10,000~20,000 사이가 되도록 음량을 조절하였다.

각 지역에 마련된 녹음 장소와 녹음 장비가 세팅된 환경을 직접 확인하고 수정 보완하였으나 진행 중 발생하는 돌발적인 소음 등은 완벽하게 통제하기 힘들었다.

이렇게 구축된 녹음 환경에서 발화한 샘플을 국립국어원에 검증 받아 최종적으로 적합한 환경을 확인한 후 본 녹음을 진행하였다.



[그림 12] 녹음 장비

4.2. 음성 녹음

4.2.1. 녹음 절차

예정된 시간에 화자가 녹음 장소에 도착하면 진행 요원은 화자 모두의 개인 정보를 확인하고 녹음된 음성 자료 및 전사 결과물 등 관련 자료에 대한 저작권 이용 허락 계약서를 작성하도록 하였다. 그리고 녹음 시 유의 사항에 대해 충분히 전달한 다음 한두 차례 시험 녹음 실시 후 본 녹음을 진행하였다. 녹음은 한 화자당 최대 녹음 시간을 30분으로 제한하며, 동일 화자가 중복 참여하지 않도록 관리하였다.

녹음은 다음과 같은 순으로 진행하였다.

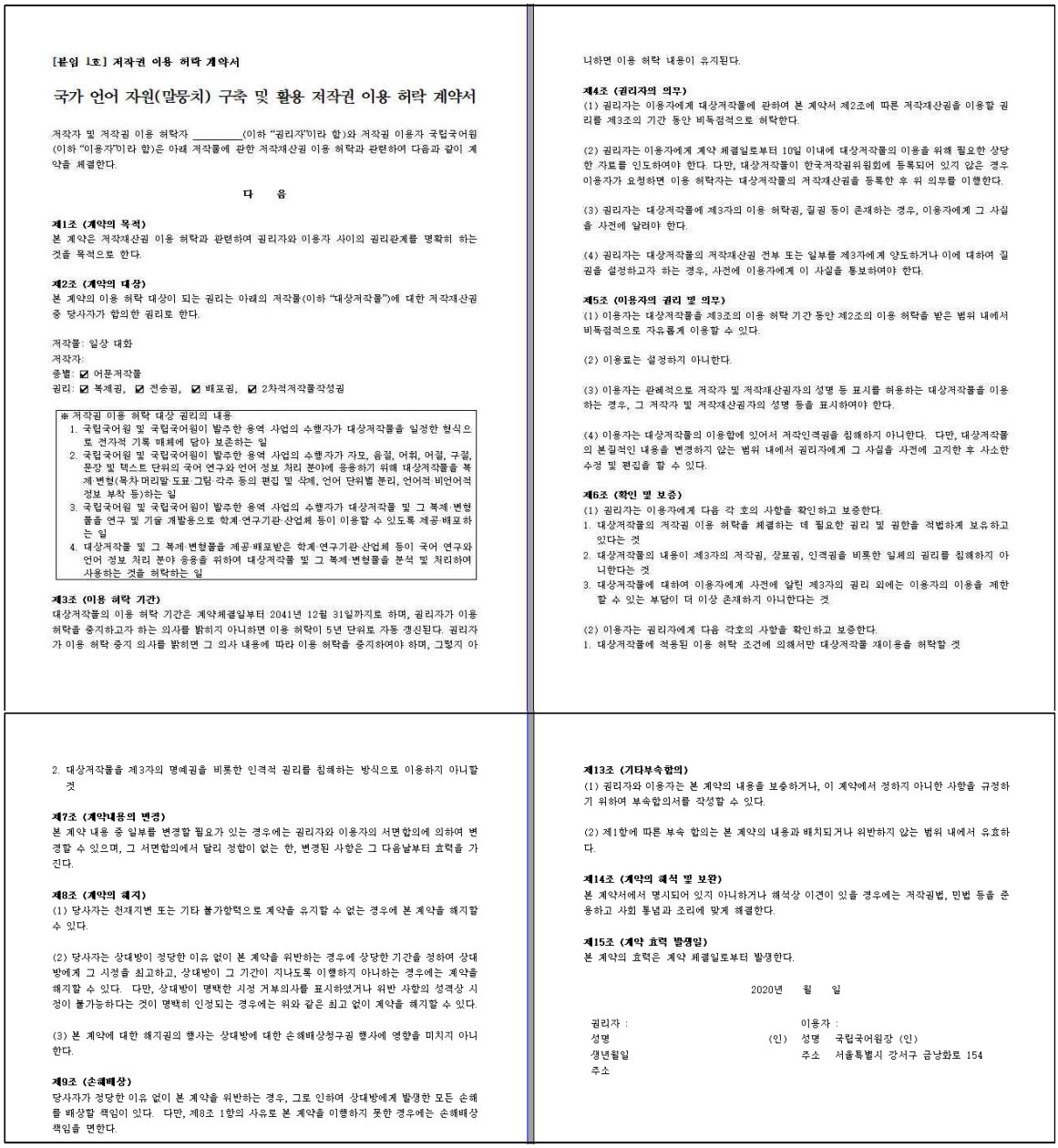
녹음 목적 설명 및 저작권 이용 허락 계약 체결	• 화자에게 녹음의 목적과 방법을 자세히 설명 • 녹음된 음성 자료 및 전사 결과물 등 관련 자료에 대한 이용 허락 계약 체결
음성 자료 수집 일지 작성	• 화자 정보와 녹음 일시, 화자 간 관계 등의 수집 일지 작성
사전 시험 녹음	• 두 화자 모두 헤드셋 마이크를 바르게 착용 • 본 녹음에 앞서 사전 시험 녹음을 실시하고 녹음이 제대로 되는지 확인 후 마이크 볼륨 및 녹음 순서, 발화 겹침, 마이크와 입과의 거리 등 조절
녹음	• 녹음 시 분명하고 큰 목소리와 명확한 발성으로 대화하도록 하고 손이나 머리카락, 옷깃 등에 마이크가 닿지 않도록 함. • 추임새 또는 맞장구 등으로 인해 화자 간 발화가 겹치지 않도록 함. • 사전 녹음 시험 때와 목소리 크기를 비슷하게 하도록 함.
관리자 공유 시스템에 음성 파일 등록	• 녹음 완료 후 지정된 경로에 음성 파일을 등록함. : 시스템 로그인 → 음성 파일 업로드 → 결과 보고 및 등록 결과 확인

[그림 13] 녹음 절차

4.2.2. 저작권 이용 허락 계약 체결 및 수집 일지 작성

녹음 전 참가자에게 본 사업의 목적과 개인 정보 보호 준수에 대해 충분히 설명하고 저작권 이용 허락 계약을 체결하였다. 이용 허락 계약서는 결과물인 음성 파일, 전사 파일, 그 변형물에 대한 복제권, 전송권, 배포권, 2차적 저작물 작성권에 대해 국립국어원에서 활용하는 것을 허락한다는 내용으로 참가자 전원 작성을 원칙으로 하였다.

참가자가 녹음실을 방문하여 녹음 전 사전 안내와 함께 같이 작성하기 때문에 참석한 대부분의 화자들이 이용 허락 계약서에 동의는 했으나, 간혹 일부 화자는 계약서 내용을 확인한 후 녹음을 거절하고 돌아간 경우도 있었다.



[붙임 1호] 저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

저작자 및 저작권 이용 허락자 _____(이하 "권리자"이라 합)와 저작권 이용자 국립국어원(이하 "이용자"이라 합)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, ☑ 전송권, ☑ 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 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 그 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
4. 대상저작물 및 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용 허락 기간)
대상저작물의 이용 허락 기간은 계약체결일부터 2041년 12월 31일까지로 하며, 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 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 인도하여야 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 등이 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.

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
2. 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
3. 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

2. 대상저작물을 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 거부의사를 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.

제13조 (기타부속합의)
(1) 권리자와 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.

2020년 월 일

권리자 :
성명 (인)
생년월일
주소

이용자 :
성명 국립국어원장 (인)
주소 서울특별시 강서구 금낭화로 154

[그림 14] 저작권 이용 허락 계약서

저작권 이용 허락 계약 체결 후 진행 요원은 녹음에 참여한 화자 모두의 정보(녹음 일시, 성명, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지 등)와 마이크 작동 수, 대화 주제, 주제의 키워드를 수집 일지에 작성하였다.

녹음 일시	녹음 시간	성명	채널			성별	연령	연락처	직업	신청 지역	화자 간의 관계	출생지	주 성장지	현 거주지	학력
2020-07-01	10:00	A	1		2	여	31	010-1234-5678	회사원	서울	제삼자	서울	서울	서울	전문대_졸업
2020-07-01	10:00	B	2	1		남	27	010-1234-5678	무직	서울	제삼자	서울	서울	서울	고등학교_졸업
2020-07-01	10:00	C		2	1	남	45	010-1234-5678	프리렌서	서울	제삼자	서울	서울	서울	대학교_졸업

[그림 15] 음성 자료 수집 일지 예시(1)

녹음 일시 (녹음 시간)	2020-07-01(10:00)											
HEAD	2	2	2									
주제(No.)	6	37	17									
키워드	상사, 동료 근무지 편의시설	가족여행 자녀 시댁	운동, 체력 건강보조제 다이어트									

[그림 16] 음성 자료 수집 일지 예시(2)

4.2.3. 녹음 진행

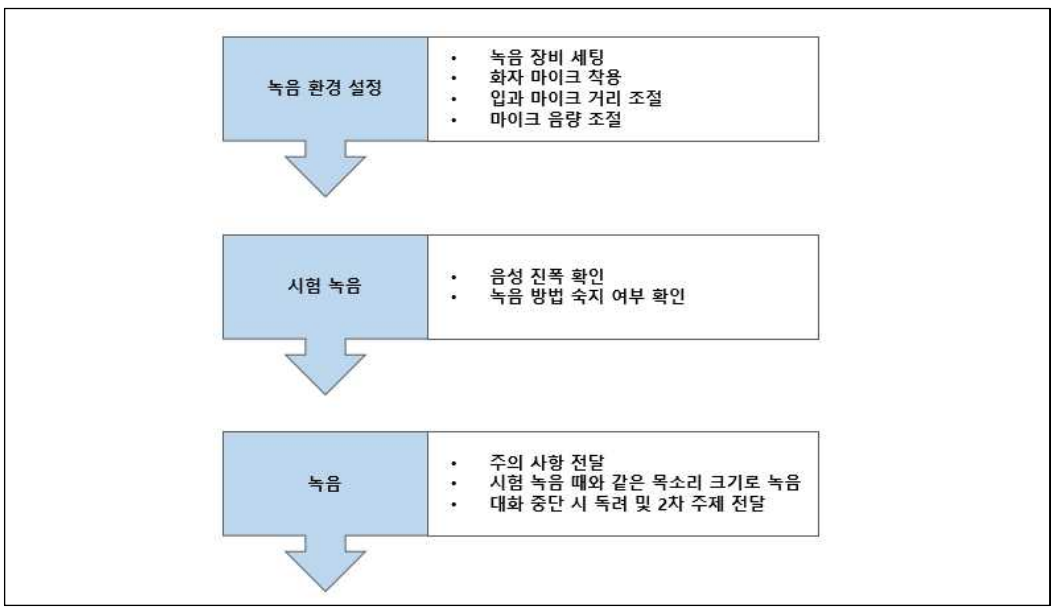
진행 요원은 녹음 장비를 세팅하고, 화자별로 헤드셋 마이크를 바르게 착용했는지 확인 후 사전 시험 녹음을 3~5분 정도 진행하였다. 반드시 실제와 같은 목소리 크기로 시험 녹음을 한 다음 녹음 상태를 확인했다. 녹음이 제대로 되는지, 음량은 적정한지, 녹음 방법을 충분히 숙지하였는지 등을 확인한 다음 본 녹음을 시작했다.

만약 녹음 상태가 올바르지 못하다면 헤드셋 마이크와 입과의 거리, 마이크 음량 등을 조절한 후 다시 시험 녹음을 진행하였고, 화자가 녹음 시 주의 사항을 숙지하지 못함으로 인해 발생한 문제에 대해서는 재차 주의 사항을 설명 후 녹음을 다시 진행했다.

본 녹음 시 분명하고 큰 목소리로 대화하도록 하고 사전 시험 녹음 때와 비슷한 목소리 크기로 대화하도록 했다. 또한 손이나 머리카락, 옷깃 등에 마이크가 닿지 않도록 했으며, 추임새 또는 맞장구 등으로 인해 화자 간 발화가 최대한 겹치지 않도록 하는 등 녹음 시 주의 사항을 다시 한번 전달하고 본 녹음을 시작했다.

대부분의 경우 녹음 시작 후 처음 몇 분 동안은 대화가 부자연스럽고 녹음 방법 등을 어려워했으나 몇 차례 재녹음이 진행될수록 본연의 일상 대화가 자연스럽게 이루어졌다.

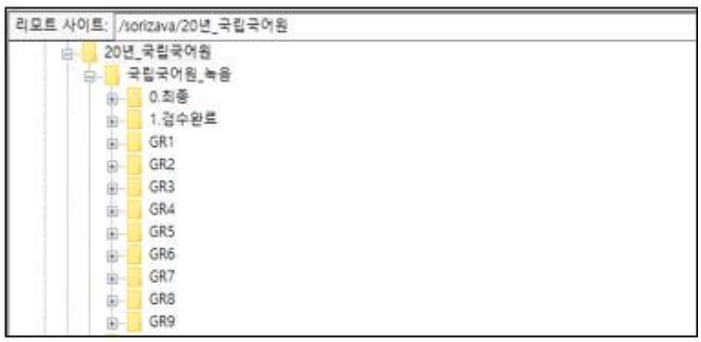
녹음 중 주제에서 벗어난 대화가 지속되거나 녹음 방법과 다르게 필요 이상으로 자연스럽게 대화함으로 인해 규칙에 벗어나는 경우 진행 요원은 녹음을 중단하고 녹음 규칙에 맞게 대화하도록 요청했다. 만약 하나의 주제에 대해 더 이상의 대화가 힘들다고 판단되는 경우 진행 요원은 화자가 선택한 두 번째 관심 주제를 제시하고 대화를 지속하도록 했다.



[그림 17] 녹음 진행

녹음이 끝나면 각 지역별 녹음 진행 요원은 녹음 원본을 WAV 파일로 변환 후, 원본과 WAV 파일을 지정된 경로에 등록하고 관리자에게 특이 사항 및 결과를 보고하였다.

■ 관리자 공유 시스템 로그인



■ 지정된 경로에 음성 파일 등록

[그림 18] 공유 시스템 로그인 및 파일 등록 예시

5. 음성 자료 전사

5.1. 전사 규칙

발화 내용은 기본적으로 한글 맞춤법에 따라 전사하는 것을 원칙으로 하며, 띄어쓰기도 한글 맞춤법을 따라 전사하였다. 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 병행하였다.

[표 12] 전사 규칙 예시

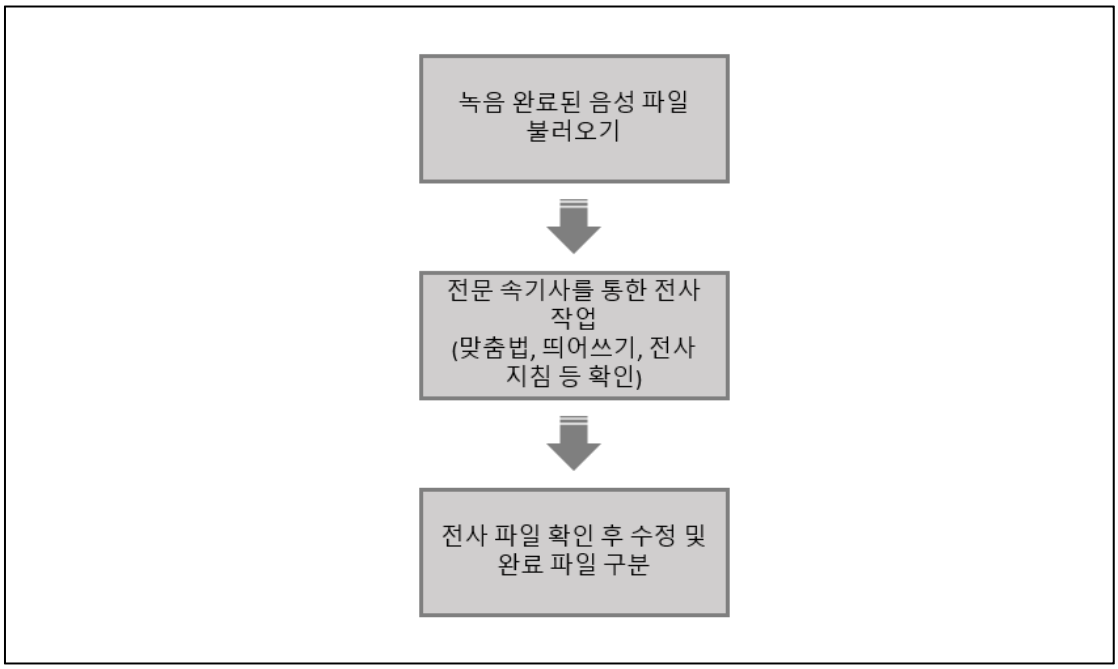
지침 항목	예시
전사 단위	기본 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 하며, 하나의 전사 단위가 3초 이상으로 길어지는 것을 지양한다. (예) 1: 내가 학교에 갔을 때/ 학생들이 막/ 길게 줄을 서 있더라고./
이중 전사 (철자 전사, 발음 전사)	발화 내용은 기본적으로 철자 전사를 하되, 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 전사를 한다. (예) • 철자 전사: 그렇게 해도 되더라고요. • 발음 전사: 그렇게 해도 되더라구요. • 철자 전사: 이렇게 적금을 3년씩 넣었더니 • 발음 전사: 이케 적금을 삼 년씩 넣었더니
끊어진 단어	끊어진 단어는 발화된 대로 그대로 전사한다. 불완전하게 발화된 단어(어절)가 둘 이상인 경우에도 어절마다 다음과 같이 표시하여 전사한다. (예) • -저- 저는 아직 미혼이기 때문에 • -아- 아이들에게 안전 교육을 하고

지침 항목	예시
축약형 표기	축약형의 경우 모두 표기에 반영한다. 모음의 축약형의 전사에서는 ’를 사용해서 두 음소를 연결해 준다. (예) • (이리로)/(일루) 가면 있더라고요. • 그 사람이랑 (사귀어서)/(사귀‘어서)
준음성 및 기타 소리	웃음, 목청 가다듬는 소리, 박수, 노래 등은 다음과 같이 전사한다. (예) • 아무래도 제주도에서 @목청 • 그래서 저는 장마를 좋아해요. @웃음 • @노래 이거 하는 거 이게 딱 나오거든요.
익명성 보장을 위한 전사	일상 대화 자료 중 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. (예) • @n 씨는 여행을 할 때… • @상호명에서 아르바이트를 하다가…
잘 들리지 않는 부분	잘 들리지 않아 추정한 경우는 다음과 같이 전사한다. (예) • ((이게)) 사실 핑계지만… • 근데 먹고 (()) 시댁에 또 살다 보니… • 또 많이 ((xx)) 거 아니야.
담화 표지	동일한 형태로 기존 품사의 의미, 기능을 가지지 않는 것은 담화 표지로 보고, 물결표(~)를 이용하여 표시한다. (예) • 어~ 키우지 않고 있는 이유는 • 그~ 준다고 해도

5.2. 전사 절차

녹음 완료된 음성 파일을 내려받아 전사 도구(TranscriberAG)를 사용해 전문 속기사가 전사 규칙에 따라 발화자를 표시하고, 전사 단위별로 맞춤법, 띄어쓰기에 따라 발화 내용을 전사하는 방식으로 진행하였다.

[그림 19] 전사 도구(TranscriberAG)



[그림 20] 전사 절차

5.3. 전사 작업

전문 속기사에 의한 전사는 15분 음성 파일 기준으로 평균 약 1.5시간가량이 소요되어, 1인당 일평균 4개 정도의 음성 파일을 전사하였다. 전사 작업은 일평균 12명 정도의 전사 인력이 투입되어 진행되었지만, 높은 연령의 화자가 참여한 경우와 강한 방언형을 사용하는 경우에 음성 전사에 상대적으로 더 많은 시간이 소요되었다.

전사는 기본적으로 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 이중 전사를 기본 원칙으로 하였고, 세부적인 전사는 주관 기관이 제시하는 전사 지침을 준수하였다. 전사자는 전사 단위 구분을 가장 어려워하였으며, 그다음으로 발음 및 철자 전사와 줄임말 이중 전사를 어려워하였다.

전사 단위는 억양구 단위로 구분하였는데, 음성을 듣고 억양구 단위인지 판단하는 과정에서 작업자의 주관이 어느 정도 개입되기 때문에 명확한 구분이 쉽지 않았다. 이런 경우 긴 휴지가 발생하는 부분에서 전사 단위를 구분하여 처리하였다.

발음 및 철자 전사와 줄임말 이중 전사는 여러 가지 경우의 수가 존재하여 혼란스러울 수 있으므로 국립국어원의 『우리말샘』을 활용하여 지침을 정하여 처리하였다.

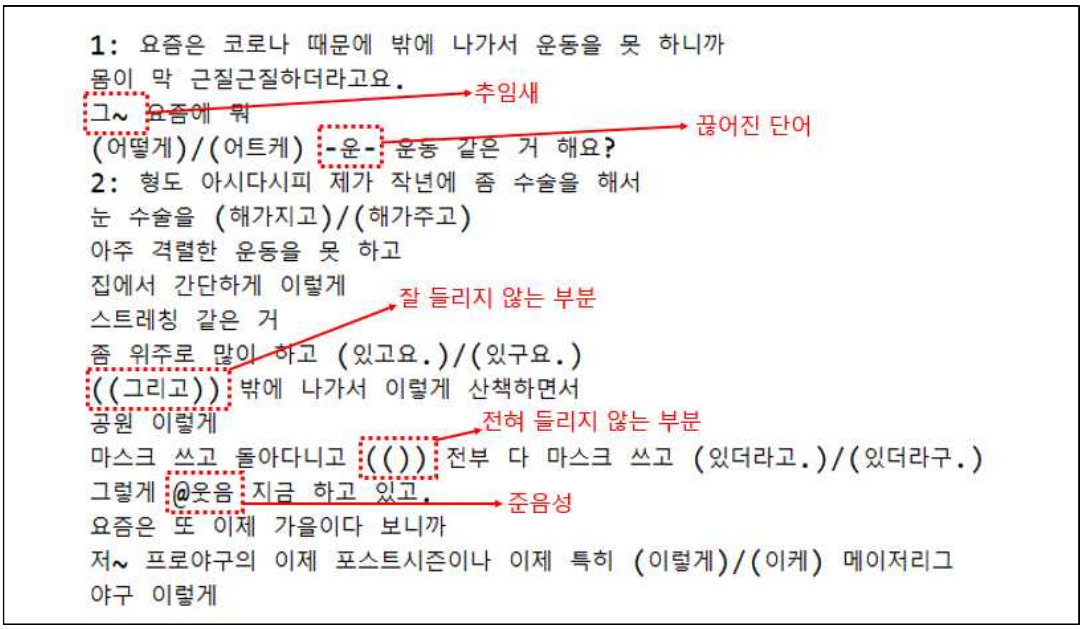
관리자는 전사자의 전사 결과물에 잘못된 부분이 없는지, 전사 규칙은 제대로 준수하였는지 등을 확인한 다음 잘못된 부분이 있다면 수정 요청하여 수정한 후 최종 완료된 전사 파일은 관리자가 취합하였다.

전사 지침에 따라 주로 작업한 내용은 다음과 같다.

[표 13] 전사 지침 및 작업 내용

발화자 표시 방법	• 첫 번째 발화자를 1로 표시하고, 1번 발화자 목소리가 2번으로 기재되지 않도록 구분하여 전사 • 화자의 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보 표시
전사 단위	• 기본 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구가 되도록 함.
문장 기호	• 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분 • 선택의문문의 경우 쉼표를 사용하지 않으므로 마지막 종결형 어미 뒤에만 물음표를 붙임. • 느낌표나 쉼표는 사용하지 않음.
발화 겹침	• 겹침 발화는 표시하지 않고 시간 순서에 따라 적음. • 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눔.

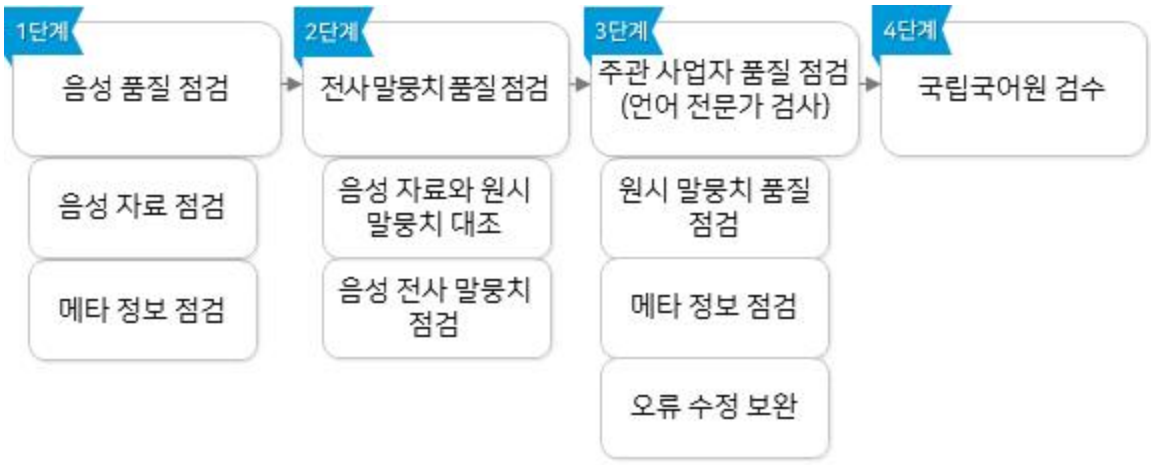
발화 내용 전사	• 발화 내용은 기본적으로 철자 전사를 하되 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 한글 맞춤법 표에 따른 발음과 차이가 있는 경우 발음 전사를 병행(예: (철자 전사)/(발음 전사)) • 모음의 변화, 된소리 등을 반영하여 적음(예: 쐬주, 쪼끔). • 약화 현상에 의한 이형태는 반영하지 않음(예: 머 → 뭐). • 숫자나 기호, 영문 등도 발음에 따라 한글로 적음. • 외래어 표기법에 반하여 발음하는 경우는 발음 전사와 병행(예: (오리지널)/(오리지날))
띄어쓰기	• 단위 띄어쓰기를 준수하고 날짜에서는 월 단위는 제외하고 띄어 씀.
축약형의 표기	• 두 음절이 한 음절의 사잇소리가 되거나, 두 음절이 한 음절 겹홀소리가 되는 등의 경우 반영 • 발음되는 음절 수와 표기상의 음절 수를 맞추어야 하므로 축약형의 경우 모두 표기에 반영함(예: (이리로)/(일루), (그러니까)/(긍까)). • 반홀소리 축약형은 ’를 표시 (예: 사귀’어, 바뀌’어)
끊어진 단어	• 단어가 불완전하게 발화된 경우, 발화된 그대로 전사하고 불완전하게 발화된 단어(어절)가 둘 이상인 경우에도 어절마다 ‘-’를 표시함(예: -전 - -전- 전통이라고). • 수정 발화, 반복 발화는 ‘-’를 표시하지 않음.
준음성 및 기타 소리	• 준음성은 소리가 나는 대로 전사하지 않고 @웃음, @목청, @박수, @노래 형태로 전사함. • 감탄, 놀람 등은 들리는 대로 전사함(예: “오”, “앗”, “어머”).
익명성 보장을 위한 전사	• 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인 정보와 관련된 사항은 비식별화함(예: @n, @상호명, @주민번호, @카드, @전화, @주소). • 상호명은 부정적인 경우에만 비식별화함. • 여러 이름이 나올 때는 일련번호를 붙여 구별하고, 파일 내에 하나의 이름만 나올 경우 n으로만 기재함(예: @n1, @n2, …).
잘 들리지 않는 부분	• 잘 들리지 않아 추정한 경우는 ‘(())’ 안에 전사(예: 내가 너보다 ((더 힘들어.))). • 발화 내용이 전혀 들리지 않는 부분은 ‘(())’ 전사(예: 내가 봐도 (()) 너무한 것 같더라.). • 들리지 않는 부분의 음절 수가 구분이 되는 경우 음절 수만큼 ‘x’를 표시 • 없는 소리를 추정하여 적지 않음.
추임새	• ‘이, 그, 저’ 등 기존 품사의 의미, 기능을 가지지 않고 주로 머뭇거림의 의미나 언어적 습관에 의해서 사용되는 것은 추임새로 보고 단어 뒤에 ‘~’를 표시함.


1: 요즘은 코로나 때문에 밖에 나가서 운동을 못 하니까
몸이 막 근질근질하더라고요.
그~ 요즘에 뭐
추임새
(어떻게)/(어트케) -운- 운동 같은 거 해요?
끊어진 단어
2: 형도 아시다시피 제가 작년에 좀 수술을 해서
눈 수술을 (해가지고)/(해가주고)
아주 격렬한 운동을 못 하고
집에서 간단하게 이렇게
스트레칭 같은 거
잘 들리지 않는 부분
좀 위주로 많이 하고 (있고요.)/(있구요.)
((그리고)) 밖에 나가서 이렇게 산책하면서
공원 이렇게
전혀 들리지 않는 부분
마스크 쓰고 돌아다니고 (()) 전부 다 마스크 쓰고 (있더라고.)/(있더라구.)
그렇게 @웃음 지금 하고 있고.
준음성
요즘은 또 이제 가을이다 보니까
저~ 프로야구의 이제 포스트시즌이나 이제 특히 (이렇게)/(이케) 메이저리그
야구 이렇게

[그림 21] 전사 결과 예시

5.4. 품질 검수

전사가 완료된 파일은 품질 검수 담당자가 품질을 점검하였다. 자체 품질 점검을 통해 전체 파일을 검수하고 오류 여부를 확인하여 수정 조치를 하였고, 재작업이 필요한 말뭉치 파일의 경우에는 전사 작업자에게 재작업을 요청하였다. 품질 점검 시 음성 파일과 전사 파일의 일치 여부 점검, 음성 파일 품질 점검, 전사 지침 사항 점검, 오류 수정 여부 점검과 전사 작업자의 평가를 진행하여 교체 여부를 결정하였다.



[그림 22] 4단계 품질 검증 단계

[표 14] 검증 세부 공정

세부 공정	작업 내용
음성 파일 점검	• 대화 주제와 무관한 대화(예: 인사말 등)가 제외되었는지 점검 • 파일이 끊기거나 소리 단절이 없는지 점검 • 마이크 지지직거리는 소리가 있거나 화자 음성이 작은지 점검
전사 말뭉치 점검	• 발음 전사와 철사 전사가 병행되었는지 점검 • 음성 자료의 전사 누락, 중복, 오탈자의 오류 점검
메타 정보 점검	• 메타 정보의 오탈자 및 항목 누락 등 항목별 오류 내용 전수 점검
오류 수정 보완	• 솔트룩스와 국립국어원의 품질 오류 점검 시 반복적으로 나온 오류에 대한 수정
점검 내역서 작성	• 자체 품질 점검 내역에 대한 점검 내역서 작성

품질 점검 담당자별로 작업자의 전사 파일을 점검하였으며, 자체 품질 점검 후 전사 작업자에게 실시간으로 오류 내역과 함께 보완을 요청하였다. 전사 작업자는 재작업 시 지침과 관련하여 궁금한 내용이 있을 때 품질 점검 담당자에게 문의할 수 있도록 관리 대장을 구성하였다. 반복적으로 나오는 오류 유형은 전사 파일 전체를 분석하여 오류가 있는 부분을 파악하고 수정하여 데이터 품질을 관리하였다.

파일명	업로드일	검수일	검수여부	검수자	수정요청일	1차 재작업 완료여부	수정완료파일 업로드일
SDRW2000000001	06-22	06-25	검수 완료	고*지/김**나	06-25	O	07-13
SDRW2000000002	06-22	06-23	검수 완료	김**나	06-24	O	07-13
SDRW2000000003	06-22	06-23	검수 완료	박*희	06-24	O	07-13
SDRW2000000004	06-22	06-23	검수 완료	강*빈	06-24	O	07-13
SDRW2000000005	06-22	06-23	검수 완료	배*영	06-24	O	07-13
SDRW2000000006							
SDRW2000000007							
SDRW2000000008							
SDRW2000000009							
SDRW2000000010							
SDRW2000000011							
SDRW2000000012							
SDRW2000000013							
SDRW2000000014							
SDRW2000000015							
SDRW2000000016							
SDRW2000000017							
SDRW2000000018							
SDRW2000000019							
SDRW2000000020							
SDRW2000000021							
SDRW2000000022							
SDRW2000000023							
SDRW2000000024							
SDRW2000000025							
SDRW2000000026							
SDRW2000000027							
SDRW2000000028							
SDRW2000000029							

파일명	검수내용
SDRW2000000005	1. 시험 주석 {79.52628} 저희 아 어멈 아버님이 {81.75181} → 끊어진 단어 표시 안함 2. 전사 단위 문제 {220.68502} 근데 중학교 {221.74043} {221.74038} (3학년짜리)/(삼 학년짜리) {222.43052} {222.43028} 사준기예요. {223.48525} → 분리될 만한 근거가 없음 {477.87345} 어~ 아~ 어~ (줌)/(쫌) 이제 준비해야 될 것도 있고 {482.16131} → 첫번째 '어~' 기준으로 휴지시간이 김 3. 이중전사 오류 {181.98734} 저 (슈퍼)/(수퍼)에서 (조금)/(쪼끔) 싸게 팔 때가 있고 이 (슈퍼)/(수퍼)에서 싸게 팔 때가 있는데 {185.76083} → '쪼끔'과 '쪼금'은 이중전사 대상 아님. {239.75970} 참 너희 엄마 (때문에)/(땜에) {241.35967} {906.39203} (그렇기)/(그러기) (때문에)/(땜에) → '땜에' 는 '때문에' 의 준말로 이중전사 대상 아님 {23.85022} 자녀가 (3명)/(세 명) (있고요.)/(있구요.) {25.50005} {26.91627} (5명이)/(다섯 명이) 있습니다. {28.68450} {28.68417} (5식구가)/(다섯 식구가) 살아가고 (있고요.)/(있구요.) {32.16256} → 하나, 둘, 셋... 은 이중전사 하지 않음 4. 오타 {364.82364} 말성을 부려도 {366.31803} → 오타, '말썽'이라고 발음함 {557.82053} 맞아 계 투투 (더치다도 가)/(더치다더 가) {559.75157}

[그림 23] 실시간 피드백 예시

6. 음성 정제

음성 정제는 음성을 전사 단위에 따라 분할하는 작업이다. 관리자 공유 시스템에서 음성 파일과 전사 파일을 내려받아 분할 작업 후 16kHz 표본화, 16bit 양자화 선형 PCM으로 저장하는 순으로 진행하였다.

이때 음성 구간 앞뒤에 200msec의 휴지가 포함되도록 저장했다. 또한 음성 구간 앞뒤에 잡음이 포함된 경우에는 잡음 외에 200msec 이상의 휴지가 포함되도록 했다.

음성 정제 시 부정적 맥락에서 사용된 상호명, 이름, 주소, 주민등록번호, 카드 번호, 전화 번호 등의 개인 정보는 익명성 보장을 위해 묵음으로 비식별 처리하였다.

[그림 24] 음성 정제 예시

7. 원시 말뭉치 구축 및 메타 정보 구축

7.1. JSON 변환

전사가 완료된 말뭉치를 대상으로 국립국어원과 협의한 기준으로 파일명을 부여하고 JSON으로 변환하여 최종적으로 원시 말뭉치를 구축하였다. 구축 완료 후 태그 오류 여부를 확인하였고, 오류가 있는 부분을 재수정하였다.

본 사업 결과물은 「공공데이터의 제공 및 이용 활성화에 관한 법률」 등에 따라 공공 데이터의 형태로 제공되어야 함을 고려하여 구축하였으며, 구축 시에는 「공공기관의 데이터베이스 표준화 지침」(행정안전부고시 제2019-20호, 2019. 3. 20.), 「공공데이터 관리지침」(행정안전부고시 제2019-71호, 2019. 9. 3.), 「공공데이터 품질관리 매뉴얼 v2.0」(2018. 1.) 등 데이터베이스 구축 관련 규정을 준수하였다.

파일명은 말뭉치 유형 구분, 매체 및 장르 구분, 분석 층위 구분, 구축 연도, 8자리 일련번호를 부여하였다.

[표 15] 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	20	########

준음성과 기타 소리, 개인 정보는 전사 편의를 위해 '@웃음', '@이름'의 형태로 전사하였으나, JSON 변환 시 지침에 맞게 마크업하였다.

[표 16] 전사 기호의 마크업 변환

분류		전사	마크업
준음성	웃음	@웃음	{laughing}
	목청 가다듬는 소리	@목청	{clearing}
	박수	@박수	{applauding}
	노래	@노래	{singing}
비식별화	이름	@n	&name&
	상호명	@상호명	&company-name&
	주민번호	@주민번호	&social-security-num&
	카드번호	@카드번호	&card-num&
	주소	@주소	&address&
	전화번호	@전화번호	&tel-num&

파일명 부여 후 기술팀에서 JSON 변환 프로그램으로 말뭉치를 변환하고, 오류 여부 및 오류의 발생 원인을 검증하였다. JSON 형식 검증을 통해 단순한 형식 오류는 자동으로 일괄 수정하였으며, 수동으로 수정해야 할 경우에는 전사팀에 전달하여 확인 후 수정하였다.

SDRW2000002219.txt	이중 전사의 괄호 표시가 맞지 않습니다.	1000만)/(천만) 관객이 넘은 영화 중에서는
SDRW2000000153.txt	발음 전사에는 숫자나 영문이 포함될 수 없습니다.	(2학기로)/(2학기로)
SDRW2000001031.txt	허용되지 않은 준음성 표기(@)가 있습니다.	저도 인제 나이를 먹다 보니까 @윳음
SDRW2000000189.txt	허용되지 않은 준음성 표기(@)가 있습니다.	@이름 님의 진로는
SDRW2000002047.txt	끊어진 단어로 발화가 끝날 수 없습니다.	라는 영화였 -였는데-
SDRW2000001232.txt	발화자 정보 표시 오류	2. 이것도 대학교
SDRW2000000202.txt	전사 내용 누락	(654.06404) (658.03902)
SDRW2000000597.txt	축약형 표기 오류	(사귀'어)/(사겨) 그러니까
SDRW2000000693.txt	잘 들리지 않는 부분의 표기 오류가 있습니다.	((이렇게)

[그림 25] 변환 오류 예시

말뭉치 파일의 확장자는 JSON이며, 문자 인코딩은 유니코드(UTF-8)이다. 형식은 수준 4개로 구분하고 수준에 따라 스페이스 4개로 들여쓰기를 하여 계층을 시각화하였다.

[표 17] 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: 2020
	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

[표 18] 말뭉치 변환 예시


```
{
    "id": "SDRW2000000002",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2000000002",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2020",
        "category": "구어 > 사적 대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2000000002.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20200602",
                "topic": " 반려동물 > 보험, 유튜버, 동물학대, 작명",
                "speaker": [
                    {
                        "id": "SD2000003",
                        "age": "30대",
                        "occupation": "전문가 및 관련 종사자",
                        "sex": "여성",
                        "birthplace": "서울",
                        "pricipal_residence": "서울",
                        "current_residence": "서울",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2000004",
                        "age": "20대",
                        "occupation": "학생",
                        "sex": "여성",
                        "birthplace": "서울",
                        "pricipal_residence": "서울",
                        "current_residence": "서울",
                        "education": "대재"
                    }
                ],
                "setting": {
                    "relation": "기타"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2000000002.1.1.1",
                    "form": "반려동물을 키우고 계신가요?",
                    "original_form": "반려동물을 키우고 계신가요?",
                    "speaker_id": "SD2000003",
                    "start": 2.78988,
                    "end": 4.93666,
                    "note": ""
                },
                {
                    "id": "SDRW2000000002.1.1.2",
                    "form": "혹시 안 키우고 계시다면은",
                    "original_form": "혹시 안 키우고 계시다면은",
                    "speaker_id": "SD2000003",
                    "start": 4.93661,
                    "end": 6.71950,
                    "note": ""
                },
```

7.2. 메타 정보 구축

녹음 날짜, 대화 아이디, 대화 주제, 화자 아이디, 화자 정보(성별, 연령대, 직업, 출생지, 주 성장지, 현 거주지 등), 화자 간 관계는 필수 항목으로 메타 정보를 구축하였고, 대화 주제는 대주제(topic 1)와 소주제(topic 2)로 나누어서 기재하였다.

2020 국립국어원 일상대화 말뭉치 메타 정보

document								
id	metadata						setting	
id	title	author	publisher	date	topic1	topic2	relation	시간(초)
SDRW2000000001.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200602	여행지(국내/해외)	해외여행, 숙소, 여	친구	0:15:06
SDRW2000000002.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200602	반려동물	보험, 유튜버, 동물	기타	0:15:08
SDRW2000000003.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200602	먹거리	좋아하는 음식, 다	기타	0:15:31
SDRW2000000004.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200602	반려동물	영상, 파충류, 길고	기타	0:15:16
SDRW2000000005.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200602	가족	관계, 에피소드, 연	기타	0:15:28
SDRW2000000006.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200602	먹거리	맛집, 디저트, 차, 디	형제/자매	0:15:42
SDRW2000000007.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	여행지(국내/해외)	제주, 부산, 카페, 남	친구	0:14:44
SDRW2000000008.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	계절/날씨	사계절, 좋아하는 계	친구	0:15:24
SDRW2000000009.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	건강/다이어트	약, 운동, 공진단	친구	0:10:15
SDRW2000000010.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	계절/날씨	사계절, 활동, 여행	기타	0:10:12
SDRW2000000011.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	여행지(국내/해외)	장소, 첫여행, 추천	기타	0:15:05
SDRW2000000012.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	가족	강아지, 자랑, 결혼,	기타	0:15:17
SDRW2000000013.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	꿈(목표)	장래희망, 아르바이	기타	0:15:18
SDRW2000000014.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	스포츠/레저	올림픽, 스포츠선수	기타	0:15:19
SDRW2000000015.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	회사/학교	선생님, 동네, 분위	친구	0:15:12
SDRW2000000016.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200603	반려동물	질병, 핸들링, 죽음,	친구	0:15:24
SDRW2000000017.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200609	반려동물	코로나, 산책, 강아	기타	0:15:06
SDRW2000000018.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200609	선물	받고 싶은 선물, 어	기타	0:15:08
SDRW2000000019.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200609	스포츠/레저	코로나, 헬스장, 운	기타	0:15:08
SDRW2000000020.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200609	영화	애니메이션, 장르,	기타	0:15:47
SDRW2000000021.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200609	계절/날씨	좋아하는 날씨, 기	기타	0:15:17
SDRW2000000022.1	2인 일상 대화	개인 발화자	개인 발화 녹음	20200609	방송/연예	예능, 나영석, 김태	기타	0:15:17

[그림 26] 메타 정보 파일 일부

2020 국립국어원 일상대화 말뭉치 발화자 정보

* 대화가 다르더라도 화자가 동일하면 동일한 아이디 부여

fileId	id	age	occupation	sex	birthplace	pricipal_residence	current_residence	education
SDRW2000000001	SD2000001	20대	학생	여성	경기	경기	경기	대재
SDRW2000000001	SD2000002	20대	전문가 및 관련 종사자	여성	경기	경기	경기	대졸
SDRW2000000002	SD2000003	30대	전문가 및 관련 종사자	여성	서울	서울	서울	대졸
SDRW2000000002	SD2000004	20대	학생	여성	서울	서울	서울	대재
SDRW2000000003	SD2000005	10대	학생	여성	서울	서울	서울	중졸
SDRW2000000003	SD2000006	20대	기타	여성	서울	경기	서울	대재
SDRW2000000004	SD2000007	10대	학생	여성	서울	서울	서울	중졸
SDRW2000000004	SD2000006	20대	기타	여성	서울	경기	서울	대재
SDRW2000000005	SD2000008	30대	무직/취업준비생	남성	서울	서울	서울	대졸
SDRW2000000005	SD2000009	40대	주부	여성	서울	서울	서울	고졸
SDRW2000000006	SD2000010	20대	학생	여성	경기	경기	경기	대재
SDRW2000000006	SD2000011	20대	사무 종사자	여성	경기	경기	경기	대졸
SDRW2000000007	SD2000012	20대	학생	여성	제주	제주	제주	대재
SDRW2000000007	SD2000013	20대	학생	여성	제주	제주	제주	대재

[그림 27] 발화자 정보 일부

제 3 장

사업 수행 결과

1. 주제별·제시 자료별 수집 결과

일상 대화 말뭉치는 주관 기관과 협의하여 15개 주제와 13개 제시 자료를 선정하였고, 특정 주제나 자료에 편중되지 않도록 수집하였다.

[표 19] 주제별 수집 결과

번호	대주제	세부 예시 주제	수집 쌍	비율
1	스포츠/레저	종목, 운동선수, 올림픽, 경기 관람 등	115	6.3%
2	여행지 (국내/해외)	장소(나라, 지역), 관광 명소, 여행 계획, 경험 등	145	8.0%
3	계절/날씨	봄, 여름, 가을, 겨울, 추억 등	118	6.5%
4	회사/학교	재직(재학) 중인 곳, 학창 시절, 동창, 선생님, 동아리 등	134	7.4%
5	먹거리	음식, 맛집, 요리법, 요리사 등	131	7.2%
6	방송/연예	드라마, 연예인, 프로그램, 이슈 등	110	6.1%
7	영화	영화인, 영화관, 영화제, 영화 장르 등	133	7.3%
8	건강/다이어트	질병, 약, 건강 보조제, 건강 관리, 약물 부작용 등	108	5.9%
9	선물	추억, 종류, 이벤트, 핸드메이드 등	105	5.8%
10	꿈(목표)	꿈(과거, 금년), 장래 희망 등	108	5.9%
11	연애/결혼	이상형, 데이트, 배우자, 연애관, 자녀 등	131	7.2%
12	반려동물	추억, 반려동물 종류, 동물 이름, 질병 등	124	6.8%
13	아르바이트	종류, 추천, 경험 등	133	7.3%
14	성격	혈액형, 다혈질, 소심함 등	106	5.8%
15	가족	가족 관계, 형제, 자매 등	117	6.4%
합계			**1,818**	**100%**

[표 20] 제시 자료별 수집 결과

번호	기사 제목	국립국어원 신문 말뭉치 아이디	수집 쌍	비율
1	경향신문 “서촌 족발집 사장은 어쩌다 망치를 휘둘러 ‘살인미수범’이 됐나…영업권보다 재산권, 법이 빚은 ‘비극’”	NWRW1900000040.15446	47	11.4%
2	경향비즈 “전복 산지값 폭락에도 소비자값은 ‘찔끔’ 왜?”	NWRW1900000040.15396	36	8.7%
3	경향신문 “역시 세계 최강…한국 女 쇼트트랙 3000m 계주, 넘어지고도 올림픽 신기록”	NWRW1900000040.4066	22	5.3%
4	경향신문 “폐지 줍는 노인 절반, 월 10만원도 못 번다”	NWRW1900000040.9893	52	12.6%
5	한겨레 “일본 지역사회의 치매 끌어안기…조금 실패해도 괜찮지 않나요”	NWRW1900000060.15333	45	10.9%
6	한겨레 “‘제2의 장현수’들…예술·체육요원 절반이 ‘허위 봉사활동’”	NWRW1900000060.18354	28	6.8%
7	동아일보 “김정은, 트럼프에 4번째 친서… ‘핵리스트 제출’ 제안 가능성”	NWRW1900000020.18519	10	2.4%
8	동아일보 “난민 문제, 불편하지만 우리사회 문화적 체질 강화시킬 수 있어”	NWRW1900000020.21711	29	7.0%
9	동아일보 “대만 징병제 67년만에 역사속으로”	NWRW1900000020.24874	22	5.3%
10	조선일보 “무상복지로 경제 몰락한 브라질… '극우 포퓰리스트' 불러냈다”	NWRW1900000010.21763	19	4.6%
11	조선일보 “경제 점수는 빵점… 이대로면 몰락할 것"”	NWRW1900000010.21633	28	6.8%
12	조선일보 “[아무튼, 주말] 비주얼이 예술이네요… 눈에서 살살 녹는 송년 케이크”	NWRW1900000010.26464	45	10.9%
13	조선일보 “反韓 정서 확산, 객관적 역사 연구 힘들게 해”	NWRW1900000010.25960	31	7.5%
합계			**414**	**100%**

2. 화자 모집 결과

2.1. 인구 특성별 수집 결과

일상 대화 말뭉치 구축 사업 특성상 대면 녹음이 불가피한 상황에서 코로나-19 대유행으로 수도권과 각 권역의 사회적 거리두기 조치가 강화되었다. 이에 주관 기관과 협의하여 지역별 사회적 거리두기 조치 상황에 따라 화자 모집 인원을 조정하였으며, 결과적으로 총 인원 2,739명으로 목표치보다 추가 모집하였다.

[표 21] 성×연령×지역별 화자 모집 결과(단위: 명)

		남성						여성						합계	
		10대	20대	30대	40대	50대	60대	10대	20대	30대	40대	50대	60대	지역별	권역별
수도권	서울	1	54	38	64	22	22	77	68	32	139	116	45	678	1,083
	경기	6	24	7	5	8	2	79	51	43	58	25	7	315	
	인천	2	9	4	2	1	0	25	15	15	10	6	1	90	
충청권	대전	10	21	15	2	16	1	6	19	19	17	29	0	155	259
	세종	1	2	0	0	0	0	1	1	0	0	0	0	5	
	충북	0	4	0	3	0	0	0	4	3	11	13	5	43	
	충남	2	5	1	0	0	0	0	14	2	19	5	8	56	
영남권	대구	2	15	9	7	0	2	5	92	17	15	31	5	200	724
	경북	1	3	3	0	0	0	1	20	7	7	14	4	60	
	부산	8	28	13	28	18	0	39	124	31	39	44	8	380	
	경남	0	2	5	2	2	2	6	24	5	5	5	2	60	
	울산	0	0	1	0	0	0	0	14	7	2	0	0	24	
호남권	광주	3	28	6	32	8	2	13	80	10	69	28	6	285	414
	전북	3	0	2	2	0	2	0	6	3	24	4	6	52	
	전남	0	5	3	4	2	3	4	13	6	14	17	6	77	
강원권	강원	5	19	2	5	2	1	7	45	8	40	10	7	151	151
제주권	제주	0	18	6	4	5	0	18	25	11	15	6	0	108	108
합계		44	237	115	160	84	37	281	615	219	484	353	110	2,739	2,739

2.2. 주제별 연령대 분포

주제별 화자의 연령대 분포를 보면 주제별로 다양한 연령이 고르게 대화하였음을 알 수 있다. 그 중 10대는 회사/학교, 방송/연예, 20대는 여행지, 회사/학교, 아르바이트를 주제로 대화를 많이 하였고, 30대는 여행과 먹거리, 40대는 스포츠/레저와 가족, 50대는 가족과 건강/다이어트, 60대는 가족, 선물을 주제로 대화를 많이 하였다.

[표 22] 주제별 연령대 분포(단위: 명)

주제	10대	20대	30대	40대	50대	60대	합계	비율
스포츠/레저	11	45	36	70	52	16	230	6.3%
여행지(국내/해외)	24	133	44	38	41	10	290	8.0%
계절/날씨	28	89	16	43	42	18	236	6.5%
회사/학교	89	119	30	19	10	1	268	7.4%
먹거리	33	85	48	49	34	13	262	7.2%
방송/연예	60	79	28	38	12	3	220	6.1%
영화	28	97	51	53	29	8	266	7.3%
건강/다이어트	14	58	21	51	56	16	216	5.9%
선물	18	52	18	59	39	24	210	5.8%
꿈(목표)	37	75	12	35	39	18	216	5.9%
연애/결혼	28	82	37	59	39	17	262	7.2%
반려동물	31	72	41	55	39	10	248	6.8%
아르바이트	18	116	42	54	28	8	266	7.3%
성격	25	87	22	41	31	6	212	5.8%
가족	11	51	27	60	58	27	234	6.4%
합계	455	1,240	473	724	549	195	3,636	100%

2.3. 제시 자료별 연령대 분포

제시 자료별 화자의 연령대 분포를 보면 이 역시 다양한 연령이 다양한 제시 자료로 고르게 대화하였음을 알 수 있다. 그 중 10대와 20대는 사회면의 기사를, 30대는 국제면의 기사를, 40대는 사회면과 국제면, 50대와 60대는 국제면의 기사로 대화를 많이 하였다.

[표 23] 제시 자료별 연령대 분포(단위: 명)

제시 자료			10대	20대	30대	40대	50대	60대	합계	비율
정치	NWRW1900000010.21633	조선일보 “경제 점수는 빵점… 이대로면 몰락할 것"”	3		5	24	21	3	56	6.8%
국제	NWRW1900000010.21763	조선일보 “무상복지로 경제 몰락한 브라질… '극우 포퓰리스트' 불러냈다”	2		1	19	13	3	38	4.6%
사회	NWRW1900000010.25960	조선일보 “反韓 정서 확산, 객관적 역사 연구 힘들게 해”	2	9	5	25	18	3	62	7.5%
사회	NWRW1900000010.26464	조선일보 “[아무튼, 주말] 비주얼이 예술이네요… 눈에서 살살 녹는 송년 케이크”	30	14	6	26	12	2	90	10.9%
국제	NWRW1900000020.18519	동아일보 “김정은, 트럼프에 4번째 친서… ‘핵리스트 제출’ 제안 가능성”	2	2	1	9	5	1	20	2.4%
문화	NWRW1900000020.21711	동아일보 “난민 문제, 불편하지만 우리사회 문화적 체질 강화시킬 수 있어”	4	8	5	30	11		58	7.0%
국제	NWRW1900000020.24874	동아일보 “대만 징병제 67년만에 역사속으로”		6	8	22	7	1	44	5.3%
경제	NWRW1900000040.15396	경향비즈 “전복 산지값 폭락에도 소비자값은 ‘찔끔’ 왜?”	2	2	3	43	18	4	72	8.7%
사회	NWRW1900000040.15446	경향신문 “서촌 족발집 사장은 어쩌다 망치를 휘둘러 ‘살인미수범’이 됐나…영업권보다 재산권, 법이 빚은 ‘비극’”	4	9	9	46	22	4	94	11.4%
스포츠	NWRW1900000040.44066	경향신문 “역시 세계 최강…한국 女 쇼트트랙 3000m 계주, 넘어지고도 올림픽 신기록”	7	7	5	16	9		44	5.3%
사회	NWRW1900000040.99893	경향신문 “폐지 줍는 노인 절반, 월 10만원도 못 번다”	27	16	7	26	24	4	104	12.6%
국제	NWRW1900000060.15333	한겨레 “일본 지역사회의 치매 끌어안기…조금 실패해도 괜찮지 않나요”	1	2	9	40	27	11	90	10.9%
정치	NWRW1900000060.18354	한겨레 “‘제2의 장현수’들…예술·체육요원 절반이 ‘허위 봉사활동’”	5	6	4	26	10	5	56	6.8%
합계			89	81	68	352	197	41	828	100%

2.4. 주제별 성별 분포

주제별 성별 분포를 보면 편차 없이 고르게 다양한 주제로 남성과 여성이 대화하였음을 알 수 있다. 그 중 남성은 스포츠/레저를 주제로 대화를 많이 하였고, 여성은 여행지를 주제로 대화를 많이 하였다.

[표 24] 주제별 성별 분포(단위: 명)

주제	남성	여성	합계	비율
스포츠/레저	110	120	230	6.3%
여행지(국내/해외)	53	237	290	8.0%
계절/날씨	65	171	236	6.5%
회사/학교	66	202	268	7.4%
먹거리	57	205	262	7.2%
방송/연예	29	191	220	6.1%
영화	94	172	266	7.3%
건강/다이어트	53	163	216	5.9%
선물	27	183	210	5.8%
꿈(목표)	55	161	216	5.9%
연애/결혼	46	216	262	7.2%
반려동물	41	207	248	6.8%
아르바이트	72	194	266	7.3%
성격	49	163	212	5.8%
가족	45	189	234	6.4%
합계	862	2,774	3,636	100%

2.5. 제시 자료별 성별 분포

제시 자료별 성별 분포를 보면 남성은 국제면을 주제로 대화를 많이 하였고, 여성은 사회면을 주제로 대화를 많이 하였다.

[표 25] 제시 자료별 성별 분포(단위: 명)

제시 자료			남성	여성	합계	비율
정치	NWRW19000 00010.21633	조선일보 “경제 점수는 빵점… 이대로면 몰락할 것””	22	34	56	6.8%
국제	NWRW19000 00010.21763	조선일보 “무상복지로 경제 몰락한 브라질… '극우 포퓰리스트' 불러냈다”	16	22	38	4.6%
사회	NWRW19000 00010.25960	조선일보 “反韓 정서 확산, 객관적 역사 연구 힘들게 해”	20	42	62	7.5%
사회	NWRW19000 00010.26464	조선일보 “[아무튼, 주말] 비주얼이 예술이네요… 눈에서 살살 녹는 송년 케이크”	18	72	90	10.9%
국제	NWRW19000 00020.18519	동아일보 “김정은, 트럼프에 4번째 친서… ‘핵리스트 제출’ 제안 가능성”	9	11	20	2.4%
문화	NWRW19000 00020.21711	동아일보 “난민 문제, 불편하지만 우리사회 문화적 체질 강화시킬 수 있어”	19	39	58	7.0%
국제	NWRW19000 00020.24874	동아일보 “대만 징병제 67년만에 역사속으로”	25	19	44	5.3%
경제	NWRW19000 00040.15396	경향비즈 “전복 산지값 폭락에도 소비자값은 ‘찔끔’ 왜?”	15	57	72	8.7%
사회	NWRW19000 00040.15446	경향신문 “서촌 족발집 사장은 어쩌다 망치를 휘둘러 ‘살인미수범’이 됐나…영업권보다 재산권, 법이 빚은 ‘비극’”	22	72	94	11.4%
스포츠	NWRW19000 00040.4066	경향신문 “역시 세계 최강…한국 女 쇼트트랙 3000m 계주, 넘어지고도 올림픽 신기록”	12	32	44	5.3%
사회	NWRW19000 00040.9893	경향신문 “폐지 줍는 노인 절반, 월 10만원도 못 번다”	18	86	104	12.6%
국제	NWRW19000 00060.15333	한겨레 “일본 지역사회의 치매 끌어안기…조금 실패해도 괜찮지 않나요”	17	73	90	10.9%
정치	NWRW19000 00060.18354	한겨레 “‘제2의 장현수’들…예술·체육요원 절반이 ‘허위 봉사활동’”	17	39	56	6.8%
합계			230	598	828	100%

2.6. 화자 관계별 수집 결과

2020년 일상 대화 말뭉치 구축 사업에서는 서로 모르는 사람과 대화한 비율이 굉장히 높아서 흥미로운 연구 자료가 될 수 있다. 처음 본 사람과 대화한 비율이 전체의 48.3%를 차지하였으며, 친구와 대화한 비율이 수집 비율의 35%를 차지하였다. 그다음으로는 부부(4.2%)가 많았으며, 직장 동료(3.0%)가 그 뒤를 이었다.

[표 26] 화자 관계별 수집 결과

화자 간 관계	수집 쌍	비율
친구	781	35.0%
부부	93	4.2%
부모/자녀	33	1.5%
형제/자매	52	2.3%
연인	55	2.5%
직장 동료	67	3.0%
이웃사촌	10	0.4%
모임·동아리 지인	32	1.4%
대학 선후배	13	0.6%
교회 지인	5	0.2%
선후배	0	0.0%
사제 관계	0	0.0%
기타 가족	12	0.5%
기타	1079	48.3%
합계	2,232	100%

2.7. 직업별 수집 결과

화자의 직업 현황을 보면 학생이 전체 인원의 31.8%를 차지하였고, 그다음으로 주부가 23.2%로 높은 비율을 차지하였다.

[표 27] 화자 관계별 수집 결과(단위: 명)

직업	모집 인원	비율
경영/관리직	39	1.4%
기능원 및 관련 기능 종사자	17	0.6%
기술자 종사자 (장치/기계 조작 및 조립 종사자)	17	0.6%
기타	191	7.0%
농업/임업/어업 종사자	2	0.1%
단순노무 종사자	20	0.7%
무직/취업준비생	415	15.2%
사무 종사자	276	10.1%
서비스 종사자	77	2.8%
전문가 및 관련 종사자	135	4.9%
주부	635	23.2%
판매/영업 종사자	45	1.6%
학생	870	31.8%
합계	2,739	100%

2.8. 학력별 수집 결과

화자의 학력 현황을 보면 대졸이 전체 인원의 52.9%를 차지하였고, 그다음으로 대학교 재학(19.9%), 고등학교 졸업(14.0%)이 뒤를 이었다.

[표 28] 학력별 수집 결과(단위: 명)

학력	모집 인원	비율
초졸 이하	15	0.5%
중졸	253	9.2%
고졸	383	14.0%
대재	546	19.9%
대졸	1,450	52.9%
대학원 이상	82	3.0%
무응답	10	0.4%
합계	2,739	100%

2.9. 출생지별 수집 결과

화자의 출생지별 수집 결과를 보면 서울 출생이 전체 인원의 26.4%를 차지하였고, 그다음으로 부산(13.0%), 광주(8.0%)가 높은 비율을 차지하였다.

[표 29] 출생지별 수집 결과(단위: 명)

출생지	모집 인원	비율
서울	722	26.4%
광주	219	8.0%
대구	182	6.6%
대전	175	6.4%
부산	356	13.0%
울산	30	1.1%
인천	54	2.0%
강원	132	4.8%
경기	221	8.1%
경북	79	2.9%
경남	86	3.1%
전북	57	2.1%
전남	169	6.2%
충북	63	2.3%
충남	82	3.0%
제주	106	3.9%
세종	1	0.0%
무응답	5	0.2%
합계	2,739	100%

2.10. 주 성장지별 수집 결과

화자의 주 성장지별 수집 결과를 보면 서울이 전체 인원의 24.8%로 가장 많고 부산이 13.9%, 광주가 10.4%를 차지하였다.

[표 30] 주 성장지별 수집 결과(단위: 명)

주 성장지	모집 인원	비율
서울	678	24.8%
광주	285	10.4%
대구	200	7.3%
대전	155	5.7%
부산	380	13.9%
울산	24	0.9%
인천	90	3.3%
강원	151	5.5%
경기	315	11.5%
경북	60	2.2%
경남	60	2.2%
전북	52	1.9%
전남	77	2.8%
충북	43	1.6%
충남	56	2.0%
제주	108	3.9%
세종	5	0.2%
합계	2,739	100%

2.11. 현 거주지별 수집 결과

화자의 현 거주지별 수집 결과를 보면 서울이 전체 인원의 28.1%로 가장 많고 부산이 14.2%, 대전이 12.2%, 광주가 11.4%를 차지하였다.

[표 31] 현 거주지별 수집 결과(단위: 명)

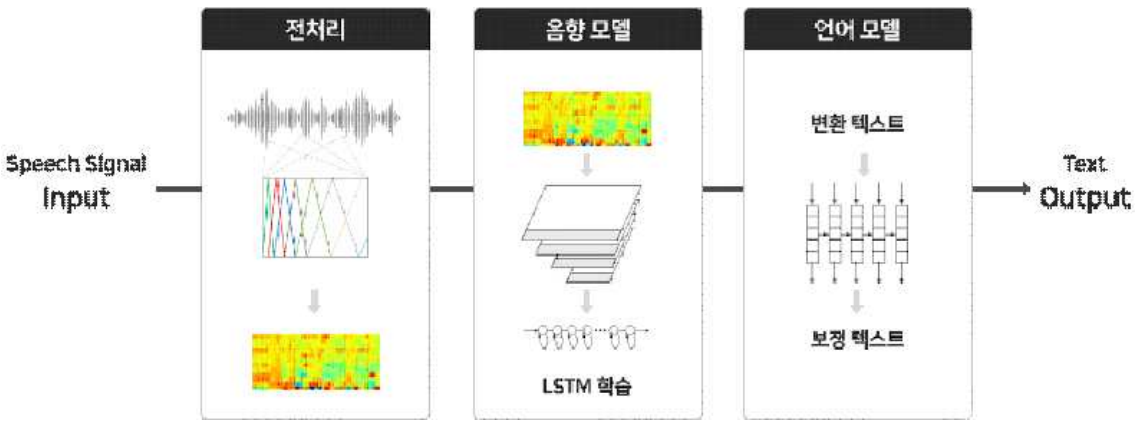
현 거주지	모집 인원	비율
서울	769	28.1%
광주	313	11.4%
대구	183	6.7%
대전	334	12.2%
부산	390	14.2%
울산	6	0.2%
인천	53	1.9%
강원	167	6.1%
경기	255	9.3%
경북	4	0.1%
경남	23	0.8%
전북	4	0.1%
전남	18	0.7%
충북	18	0.7%
충남	5	0.2%
제주	186	6.8%
세종	7	0.3%
무응답	4	0.1%
합계	2,739	100%

3. 인공지능 모델 활용

3.1. 음성 인식 모델

3.1.1. 음성 인식 모델 개요

음성 인식은 사전에 학습된 모델을 통해 음성 데이터를 텍스트 정보로 변환한다. 이 과정에서 사용되는 학습 모델은 크게 음향 모델(Acoustic Model, AM)과 언어 모델(Language Model, LM)로 구분할 수 있다. 음향 모델은 음성 데이터에서의 음향적 특성을 통계적으로 모델링하여 학습하게 되는데, 음성 인식 엔진에서 제공하는 기본 모델(baseline model)을 기반으로 실제 적용하고자 하는 음성의 특성을 추가하는 적응 학습이 가능하다. 콜센터 등 특정 분야에서 수집된 녹취 음성 데이터와 전사 데이터를 학습 데이터로 입력하여 기존 기본 모델에 적응 학습을 수행할 수 있다. 장단기 메모리(LSTM, Long Short-Term Memory) 기반으로 학습된 음향 모델은 은닉 마르코프 모델(HMM, Hidden Markov Model), 심층 신경망(DNN, Deep Neural Network) 방식에 비해 높은 음성 인식 성능을 제공하고, 해당 분야에 특화된 음성 인식 기능을 제공할 수 있다.



[그림 28] 음성 인식 학습 개요

3.1.2. 음성 인식 모델 개발 경과

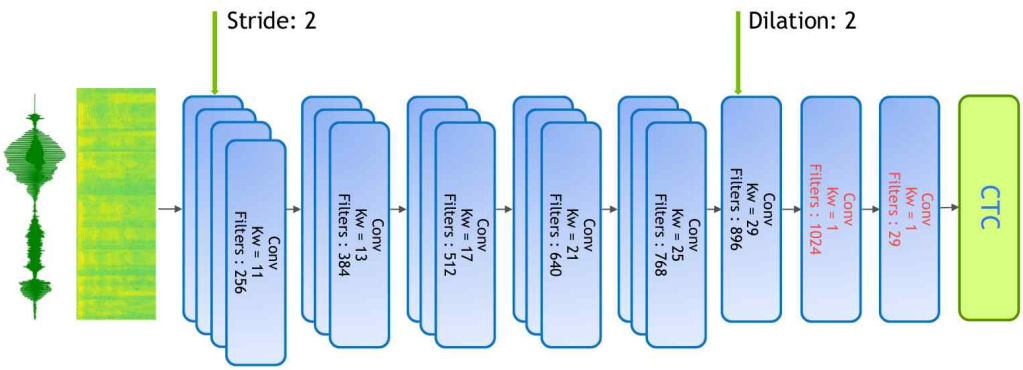
본 사업의 산출물인 500시간의 일상 대화 음성 자료와 전사 자료를 활용하여 음성 인식 엔진의 성능을 최대한 높이기 위해 다음과 같은 오픈 소스 라이브러리를 조사하였다.

[표 32] 주요 음성 인식 관련 오픈 소스

이름	언어	특징
OpenSeq2Seq	Python	기계 번역, 음성 인식, 음성 합성 중심의 인코더-디코더 모델 학습 구성 요소 제공
FAIR wav2letter++	C++	연산 효율 중심의 음성 인식 툴킷으로, wav2letter를 포함한 Facebook의 다양한 연구 논문 레시피 제공
FAIRseq	Python	다양한 레시피의 sequence-to-sequence 모델링 툴킷, wav2vec 2.0의 음성 자가 지도 학습법 제공
ESPNet	Python	음성 인식, 음성 합성 중심으로 다양한 데이터 처리, 특징 추출, 레시피 제공
Kaldi	C++, Perl	가장 널리 알려진 음성 인식 위주 툴킷 음성 인식 구현에 필요한 거의 모든 구성 요소를 제공

조사 결과, 개발에 있어서 소스 코드 복잡도가 상대적으로 낮고, 추론 속도가 빠른 종단형 모델인 OpenSeq2Seq 라이브러리 wav2letter+를 선택하여 학습 및 모델 개발에 활용하였다.

wav2letter+의 영어에 대한 음성 인식 모델은 아래 그림과 같이 19계층의 Conv1D로 구성되어 있으며 마지막 두 계층은 완전 연결(Fully-Connected) 연산과 구조적으로 동일하다. 마지막 계층의 필터 개수 29는 영어 어휘 사전의 구성인 알파벳(a-z) + 아포스트로피(') + 띄어쓰기( ) + 공백(ε)의 문자 가짓수와 같으며 공백은 연결주의 시간 분류(CTC, Connectionist Temporal Classification) 알고리즘에서 사용하는 문자로, 이 알고리즘은 음성 자료와 전사 자료의 정렬(alignment) 없이 학습 가능한 구조로 되어 있으며 적은 양의 데이터를 사용하여도 좋은 성능을 보이는 것으로 조사되었다.



[그림 29] wav2letter+ 모델

한국어 음절은 유니코드상 가짓수가 10,000가지(가~힣)가 넘어 음운 단위로 학습하고 재구축하는 방법을 사용하게 되는데, 이때는 레이블이 80개 이하로 감소하게 된다.

3.1.3. 음성 인식 모델 개발 결과

초기 선정된 알고리즘(wav2letter+)의 성능을 평가하기 위해, 공개되어 있는 학습 데이터를 활용하여 평가를 수행하였다. 그 결과 한국어 음절 단위 인식률은 80%의 성능을 보였다.

[표 33] 알고리즘 평가를 위한 학습 데이터

언어	데이터	설명
한국어	Zeroth-Korean test data	10명의 457발화, 전체 약 1.2시간

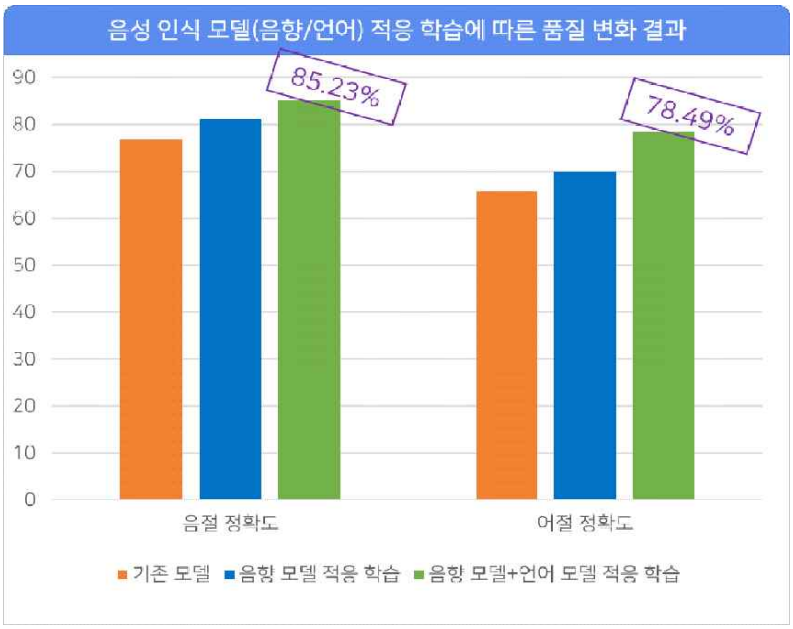
[표 34] 알고리즘 평가 결과(음절 인식률)

언어	대상 데이터	인식 단위	평균 인식률(%)
한국어	Zeroth-Korean	음절 기반	80.45%

본 사업에서 구축한 일상 대화 말뭉치를 사용해 언어 모델과 음향 모델 적응 학습을 수행하였다. 일상 대화 말뭉치 약 15시간을 대상으로 평가하였을 때, 음향 모델 적응 학습 결과 음절 인식률은 기존 모델 76.79%에서 81.15%로 약 5% 향상되었으며, 어절 인식률은 기존 모델 65.85%에서 69.89%로 약 4% 향상되었다. 여기에 언어 모델 적응 학습까지 수행한 경우에는 음절 인식률이 85.23%, 어절 인식률이 78.49%로 나타났다.

[표 35] 학습 데이터 통계

구분	내용
구축 원본 데이터	1. 전체 시간: 520시간 2. 파일 수: 2,232개 3. 파일 사이즈: 114 기가바이트 4. 오디오 스펙: 1채널, 32,000Hz 표본화, 16비트 PCM 인코딩
학습/평가 데이터	1. 비식별화 데이터 제외 2. 학습 데이터: 약 300시간, 1,200개 파일 3. 평가 데이터: 약 15시간, 60개 파일



[그림 30] 품질 변화 결과

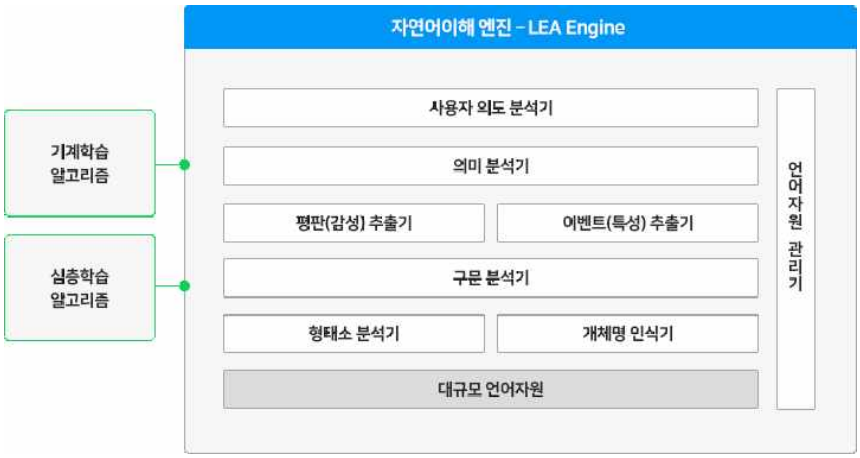
3.2. 언어 인지 모델

3.2.1. 자연어 이해 엔진 개요

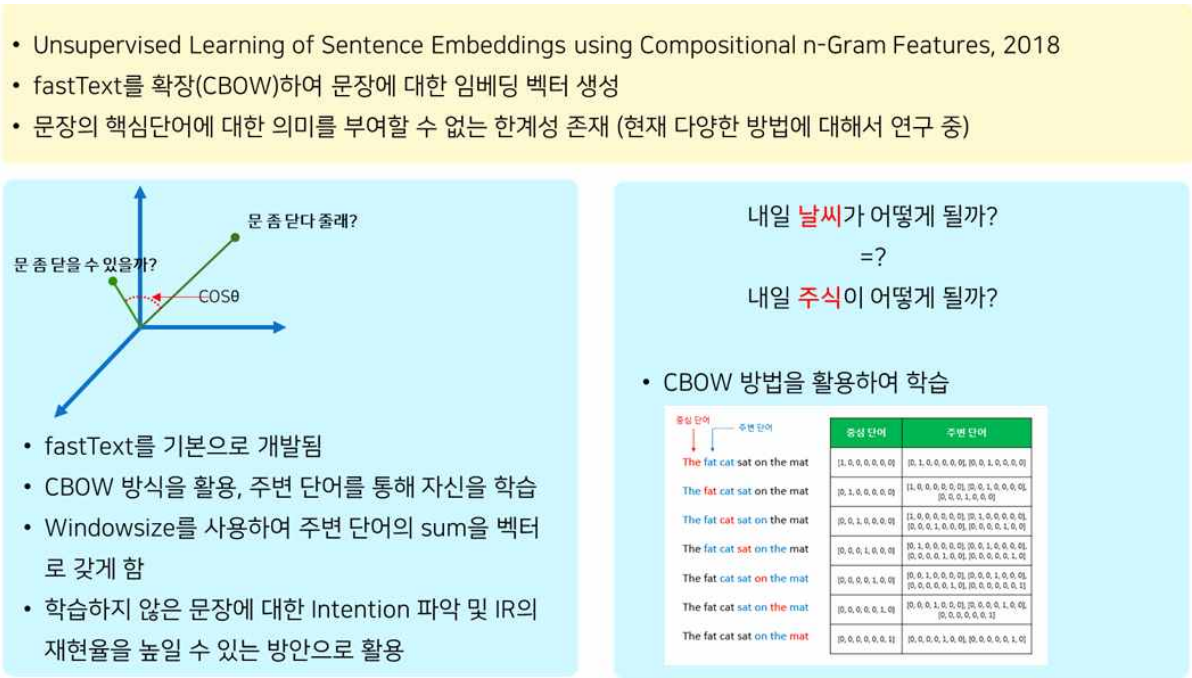
자연어 이해 엔진은 비정형 데이터 가공을 위해 형태소 분석, 개체명 인식, 구문 분석, 감성 분석 등의 텍스트 분석 기능을 처리하는 기계 학습 및 심층 학습 기반의 언어 분석 엔진이다. 그 뿐만 아니라 자연어 처리 결과를 바탕으로 문장에 숨겨진 의도를 이해하거나 질문의 유형을 파악하는 등의 한 단계 높은 수준의 분석 결과를 제공함으로써, 대화 처리를 위한 의도 이해 및 분석, 심층 질의응답을 위한 질문 의미 이해 등이 가능하다.

3.2.2. 언어 인지 모델 개발 경과

본 사업의 산출물인 일상 대화 전사 자료를 활용하여 언어 모델을 생성하기 위해 다양한 학습 알고리즘을 확인하였다. BERT 또는 GPT-II는 대화 모델 개발 및 채팅을 위한 문장 유사도를 비교하기 위한 요구 사항이 많이 생겨 학습을 수행하는 시간 및 비용이 많이 발생할 것으로 예상하였다. 따라서 sent2vec 알고리즘을 선택하였고 그에 대한 적용 및 테스트를 수행하였다.



[그림 31] 자연어 이해 엔진 구성도



[그림 32] sent2vec 개요

3.2.3. 언어 인지 모델 개발 결과

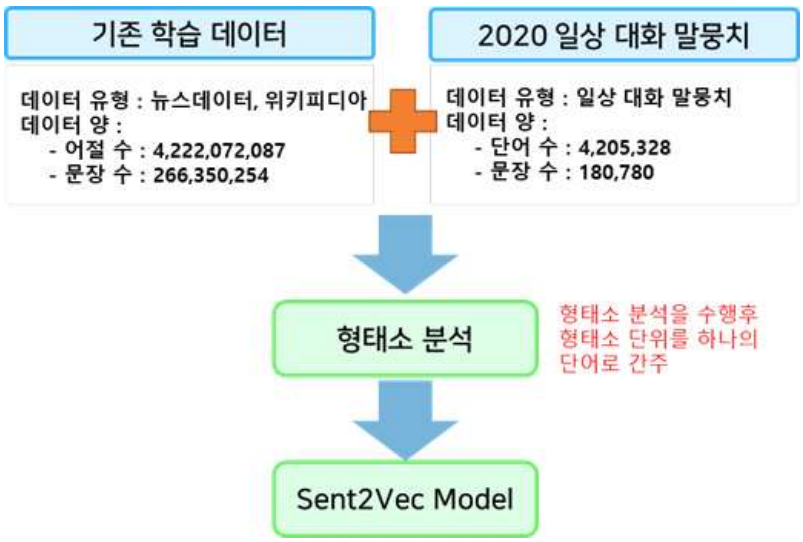
기존의 sent2vec 모델은 뉴스, 블로그, 위키피디아 자료를 사용하여 학습을 수행하였기 때문에 일상 대화에서 사용하는 구어체의 문장을 입력할 경우 기존의 문장 유사도에서 성능이 떨어지는 문제가 발생할 수 있다. 이번 언어 인지 모델 개발은 해당 일상 대화를 추가적으로 학습에 활용하였을 경우 성능에 어떠한 영향을 주었는지 확인하는 것이 가장 큰 목적이다.

일상 대화 말뭉치의 경우 발음 전사와 철자 전사가 동시에 전사되어 있어 구어 문장

과 표준어 문장으로 자동 생성하여 학습에 활용하였다.

학습 시에는 다음의 파라미터를 사용하여 학습을 수행하였다.

○ -minCount 5 -dim 300 -epoch 20 -lr 0.3 -wordNgrams 2 -loss ns –neg 10 -thread 40 -t 0.000005 -dropoutK 4 -minCountLabel 20 -bucket 2000000 –lrUpdateRate 10000



[그림 33] 학습 대상 데이터

학습 결과 전체적인 모델 성능은 약간 하락하는 문제를 보였으나, 구어체 질문에 대해서는 일부 유사도 점수가 향상되는 효과를 보였다. 향후 일상 대화와 같은 구어체 문장 데이터를 늘려 모델을 생성할 경우 구어체에도 강건한 모델이 생성될 것으로 판단하며, 이는 자연스러운 대화가 가능한 챗봇 및 인공지능 상담에 도움이 될 것으로 보인다.

[표 36] 학습 결과 예시

문장 예시	모델	유사도
토하는 아기를 흔들며 안고 있어도 괜찮나요?	기존	0.603
	신규	0.658
임신 후기인 산모에게 적더라도 출혈이 계속된다면 취해야 하는 조치는 무엇인가요?	기존	0.620
	신규	0.684
아이가 급성 기관지염에 걸렸다면 해줘야 하는 일은 무엇인가요?	기존	0.840
	신규	0.872

4. 정책 제언

본 사업은 정제 기준 500시간의 일상 대화를 녹음 후 전사, 정제하여 일상 대화 말뭉치를 구축하는 사업이다. 사업 진행 중 발생한 주요 문제점 및 개선 사항을 살펴보면 다음과 같다.

각 지역의 다양한 방언 자료를 확보하기 위해서는 청년층보다는 중장년층 위주의 데이터를 수집할 필요가 있다. 실제 지역별로 수집된 자료를 살펴보면 높은 연령층에 비해 낮은 연령층에서는 억양과 같은 운율적인 측면에서는 방언의 특성이 드러나긴 하였으나, 어휘나 표현 등에서는 방언의 특성이 크게 드러나지 않았다. 따라서 어휘나 표현의 다양성이 보장된 다양한 방언 자료를 확보하기 위해서는 청년층보다는 중장년층 화자의 비율을 높여 수집하는 것이 필요하다.

2020년 일상 대화 말뭉치 구축 사업에서는 발음 전사와 철자 전사를 병행하여 전사하였기 때문에 본 사업의 결과물은 다양한 계층에서 표준어를 어떻게 발음하는지, 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성을 연구하는 데 도움이 될 것이다.

본 사업에서는 화자의 연령층을 10대~60대로 구성하여 수집하였으나, 사실상 50대~60대 남성 화자의 모집은 어려웠다. 우리나라 사회 통념상 여성보다 남성, 남성 중 청년층보다 중장년층은 낯선 상대와의 자유로운 일상 대화를 나누는 것을 꺼리는 경향이 비교적 많기 때문인지 화자 모집이 잘 이루어지지 않았다. 덧붙여 올해 전 세계적인 코로나-19로 대면 녹음으로 진행하는 본 사업의 특성상 청년층보다 중장년층의 화자 모집에 어려움을 겪었다. 따라서 향후 진행 시 특정 화자층을 겨냥한 대화 주제 선정 등 그들이 쉽게 참여하고 자연스럽게 대화할 수 있는 환경을 조성하고, 코로나-19 확산에 대비해 적극적인 방역 대책을 마련하여 중장년층 참가자의 모집을 늘릴 필요가 있다.

향후 이러한 문제점을 보완한다면 일상 대화 말뭉치 구축 사업이 국어 및 국어 문화 연구, 4차 산업 대비 기반 기술 발전에 보다 실효성 있는 사업이 될 것으로 생각한다.

일상 대화 말뭉치 구축 지침(2020)

작성 2020. 1. 21.
수정 2020. 6. 9.

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	20	########

- 예시
  · SDRW2000000001.sjml 원시 말뭉치 첫 번째 파일

※ 참고: 음성 파일 파일명 부여 방식
  · SDRW2000000001.pcm 음성 원본 첫 번째 파일
  · SDRW2000000001-00001.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. 파일 형식 및 개요

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: 2020
	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": "SDRW2000000001",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2000000001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2019",
        "category": "구어 > 사적 대화 > 일상 대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2000000001.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20190711",
                "topic": "자동차",
                "speaker": [
                    {
                        "id": "SD1900011",
                        "age": "30대",
                        "occupation": "사무 종사자",
                        "sex": "남성",
                        "birthplace": "대구",
                        "pricipal_residence": "대구",
                        "current_residence": "경북",
                        "education": "대졸"
                    },
                    {
                        "id": "SD1900012",
                        "age": "30대",
                        "occupation": "사무 종사자",
                        "sex": "남성",
                        "birthplace": "대구",
                        "pricipal_residence": "대구",
                        "current_residence": "대구",
                        "education": "대졸"
                    }
                ],
                "setting": {
                    "relation": "동료"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2000000001.1.1.1",
                    "form": "안녕하세요.",
                    "original_form": "안녕하세요.",
                    "speaker_id": "SD1900011",
                    "start": 30.56600,
                    "end": 32.48262,
                    "note": ""
                },
                {
                    "id": "SDRW2000000001.1.1.2",
                    "form": "아~ xx님 오랜만입니다.",
                    "original_form": "아~ ((xx님)) 오랜만입니다.",
                    "speaker_id": "SD1900012",
                    "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 SDRW2000000001)
- 구축자(creator): 국립국어원
- 배포자(distributor): 국립국어원
- 구축년도(year): 2020
- 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: SDRW2000000001.1)

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/세종

- 주 성장지(principal_residence): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주/세종
- 현 거주지(current_residence): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주/세종
- 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: SDRW2000000001.1.1.4)
- 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.
- 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.

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.
- 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.

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 별도의 발음 전사를 하지 않고 철자 전사인 '뭐'만 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 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

3.7. 띄어쓰기

- 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.

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: 이리로
발음 전사: 일루

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 /ㅝ/, /ㅘ/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. 익명성 보장을 위한 전사

- 일상 대화 자료 중 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인 정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. 단, 정치인, 연예인 등 유명인의 이름은 비식별화하지 않는다. 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다.

철자 전사: 신촌에 name는 진짜 맛없어.
발음 전사: 신촌에 &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 name1이랑 name2이랑 너랑 나랑 갔잖아.
발음 전사: 그때 &name1&이랑 &name2&이랑 너랑 나랑 갔잖아.

- 비식별화 정보는 아래와 같이 마크업한다.

상호명: &company-name&
주민등록번호: &social-security-num&
카드 번호: &card-num&
주소: &address&
전화 번호: &tel-num&

3.13. 기타 지침

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

<Abstract>

2020 Daily Conversation Corpus Construction

The purpose of this corpus construction project for daily conversation corpus is to arrange preliminary data for the expansion of the Korean language corpus. Building a refined corpus of 500-hour according to the speaker recruitment guidelines and daily conversation corpus construction guidelines could increase the utilization and value of Korean resources. The results of the major tasks and projects are as follows:

Voice recording and refining: Various speakers by region, gender, and age were recruited, and a total of 2,739 speakers recorded 15 minute-long natural conversations of 15 themes and 13 newspaper articles. A contract was signed for permission of use among all recorded speakers. To prevent coronavirus infections, we divided the seats with acrylic plates, and speakers wore masks during the recording session. Conversations unrelated to conversation topics (e.g., greetings) were refined, and speech files were stored as 16 kHz sampling, 16 bits quantization linear PCM.

Voice data transcription: After selecting professional stenographers with related work experience and conducting instructional training on our transcription guidelines, we started the voice data transcription. According to these guidelines, the professional stenographers transcribed the speaker's indication, transcription unit, orthography, spacing, and double transcription using TranscriberAG. The contents of the first transcription were examined, and any errors were revised and reflected through rework.

Raw corpus construction and meta information construction: The conversation topics of voice files were divided into large categories and subcategories, and the speaker information (sex, age, place where they grew up, etc.) and the

relationship between speakers were stored and attached as meta information. The data was marked up by transcription unit and converted into JSON format according to the guidelines.

Utilization of raw corpus: Using raw corpus as a learning data form for speech recognition engine and language recognition engine, we present the application case and direction for the constructed raw corpus.

Keywords: daily conversation corpus, raw corpus, voice collection, voice recording, voice data transcription, utilization of raw corpus

Project Director: Kyung-il Lee(SALTLUX)

사업 책임자 이경일((주)솔트룩스)
사업 참여자 김예하나((주)솔트룩스)
김지영((주)솔트룩스)
박선희((주)솔트룩스)
이동기((주)솔트룩스)
주재현((주)솔트룩스)
강상규((주)소리자바)
윤종후((주)소리자바)
김창환((주)소리자바)
김동희((주)소리자바)
노희균((주)소리자바)
정용직((주)소리자바)
박선욱((주)소리자바)
김진희((주)소리자바)
담당 연구원 이승재(국립국어원 언어정보과장)
홍혜진(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2020년 12월 6일
발행일: 2020년 12월 6일
인 쇄: H&J Printing

※ 이 책은 국립국어원의 용역비로 수행한 '2020년 일상 대화 말뭉치 구축' 사업의 결과물을 발간한 것입니다.