국립국어원 2023-01-63

발간등록번호
11-1371028-000986-01

2023년 일상 대화 말뭉치 구축

사업책임자

성 기 완

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2023년 일상 대화 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 5월 2일 ~ 2023년 12월 15일

2023년 12월 15일

사업책임자: 성 기 완 (㈜솔트룩스)

사업 수행자: 주식회사 솔트룩스 컨소시엄

사업 책임자: 성기완

사업 참여자: 김 준, 김예하나, 박선희, 강수빈, 박문수,
이상준, 김응준, 노강일, 김선아, 김선희, 송혜주, 이인호, 박선욱

〈사업 수행자〉 ㈜솔트룩스 컨소시엄

사업 책임자	성기완(㈜솔트룩스)
사업 참여자	김 준(㈜솔트룩스)
	김예하나(㈜솔트룩스)
	박선희(㈜솔트룩스)
	강수빈(㈜솔트룩스)
	박문수(㈜솔트룩스)
	이상준(㈜팀벨)
	김응준(㈜팀벨)
	노강일(㈜팀벨)
	김선아(㈜팀벨)
	김선희(㈜팀벨)
	송혜주(㈜팀벨)
	이인호(㈜팀벨)
	박선욱(㈜팀벨)

<국문 요약>

2023년 일상 대화 말뭉치 구축

이 사업은 2019년부터 이어온 일상 대화 말뭉치를 구축하는 사업으로 화자 모집 계획과 말뭉치 구축 지침에 따라 정제본 500시간 규모의 말뭉치를 구축하여 활용도 높은 국어 말뭉치를 마련하고자 하는 데 그 목적이 있다. 이에 따른 주요 과업과 사업 성과는 다음과 같다.

음성 녹음 및 정제: 통계청의 인구 통계 분포를 참고하여 지역별, 성별, 나이별 다양한 화자를 모집하고 총 2,168명의 화자가 16개 주제를 기반으로 15분에서 20분간 자연스러운 대화를 녹음하였다. 참여한 화자 모두 저작권 이용 허락 계약서를 작성하였다. 코로나19 감염 예방을 위해 관리자는 마스크를 착용하고 화자는 좌석을 분리하여 녹음을 진행했다. 대화 주제(예: 인사말)와 관련 없는 부분은 정제하고 음성 파일은 16kHz 샘플링, 16비트 양자화 선형 피시엠(PCM: 펄스 코드 변조) 형식으로 저장했다.

음성 자료 전사: 관련 업무 경험이 많은 전문 속기사를 선발하고 전사 지침을 숙지할 수 있도록 교육을 진행하였다. 전문 속기사들은 전사 도구를 활용하여 지침에 따라 발음과 철자를 구분하여 전사했다. 화자 표시, 전사 단위, 맞춤법, 띄어쓰기 등 전체 말뭉치 1,973개 중 약 7.9%인 156개에서 오류가 발견되어 보완 작업을 진행하였다.

원시 말뭉치 구축 및 메타 정보 구축: 음성 파일의 대화 주제를 대범주와 하위 범주로 구분하고, 화자 정보(성별, 나이, 주 성장지 등)와 화자 간의 관계를 메타 정보 데이터에 저장하여 첨부하였다. 메타 정보 데이터는 전사 단위로 주석(마크업)되었으며 지침에 따라 제이슨(JSON) 형식으로 변환하였다.

주요어: 일상 대화 말뭉치, 원시 말뭉치, 화자 간 관계, 이중 전사, 음성 자료 전사

<Abstract>

Korean Dialogue Corpus Construction 2023

This project is to build a dialogue corpus that has been ongoing since 2019. A corpus of 500 hours of refined text will be constructed according to the speaker recruitment guidelines and corpus construction guidelines. The purpose is to provide basic data for expanding the Korean language corpus to increase the utilization and value of Korean language resources. The major tasks and business outcomes resulting from this are as follows.

Speech recording and refinement: A variety of speakers were recruited by region, gender, and age, and a total of 2,168 speakers recorded natural conversations for 15 to 20 minutes based on 16 topics. License agreements have been entered into for all recorded speakers. To prevent COVID-19 infection, the manager wore a mask and the speaker recorded in separate seats, and the speaker wore a headset microphone when recording. Parts unrelated to the conversation topic (e.g., greetings) were refined, and the audio files were saved in linear PCM format with 16 kHz sampling and 16-bit quantization.

Audio material transcription: Professional stenographers with extensive relevant work experience were selected and training was provided to ensure that they were familiar with transcription guidelines. Professional stenographers used transcription tools to transcribe pronunciation and spelling according to instructions. 156 errors were found out of 1,973 corpora, approximately 7.9% of the total, including speaker marking, transcription units, spelling, and spacing, and correction work was performed.

Construction of raw corpus and meta information: Conversation topics in audio files were divided into major categories and subcategories, and speaker information (gender, age, place of primary residence, etc.) and relationships between speakers were stored and attached to metadata. Meta information data was marked up on a transcription basis and converted to JSON format according to the instructions.

Keywords: dialogue corpus, raw corpus, voice collection, voice recording, voice data transcription, use of raw corpus

차 례

제1장 사업 개요

제2장 사업 수행

제3장 사업 수행 결과

표 차례

그림 차례

제 1 장

사업 개요

1. 사업 목적

인공지능 산업의 발전에 따라 2019년부터 일상 대화 말뭉치 구축 사업이 진행되었다. 하지만 대화형 인공지능 산업과 관련된 서비스에 활용하기 위해서는 일상 대화 말뭉치의 양이 여전히 부족하기에 대량의 고품질 말뭉치 구축이 지속적으로 필요하다.

일상 대화 말뭉치는 국내 대화형 인공지능 산업을 위한 핵심 자산으로 다양한 주제의 고품질 말뭉치를 구축하는 것은 대화형 인공지능 산업 활용을 위한 기반을 마련하는 일이 된다.

이 사업의 목적은 음성 자료의 이중 전사를 통해 메타 정보가 있는 원시 말뭉치를 구축하여 국내 언어 연구 및 인공지능 개발을 촉진하는 것이다. 또한 관련 기술 산업의 육성을 통해 국가 경쟁력을 높이고, 말뭉치 구축 및 품질 관리를 위해 국어국문학, 데이터 관련 전공자 등의 고용을 확대하여 인재를 양성하며, 다양한 인공지능 서비스와 데이터 생태계 확보를 통해 대국민 서비스를 강화하는 데 목적이 있다.

[그림 1] 인공지능 시장 전망

※ 참고: 국내 인공지능(AI) 시장은 2021년 전년 대비 24.1% 성장하여 9,435억 원의 매출 규모를 형성할 전망이며, 세계 인공지능 기술 시장이 2032년까지 1,800억 달러(약 255조 2,220억 원) 규모로 성장할 것으로 전망됐다.

2. 사업 수행 범위

이 사업은 대화형 인공지능 산업 발전에 필요한 일상 대화 말뭉치를 구축하는 것으로 사업의 수행 범위는 크게 세 가지로 나눌 수 있다. 첫째는 기존 사업 분석과 대화형 인공지능 전문가 검토를 통해 10개 이상의 다양한 대화 주제를 선정하는 것이다. 둘째는 여러 방면에서 활용이 가능한 말뭉치 구축을 위해 성별, 나이별, 지역별로 다양한 화자를 모집하는 것이다. 마지막으로 99.9% 고품질 말뭉치 확보를 위해 발음 전사와 철자 전사를 병행하여 전사하고 전체 파일을 대상으로 품질 검증을 진행하는 것이다.

구분	세부 내용	분량
주제 선정	'19년~'22년 사업 분석과 전문가 검토를 통해 '23년도 주제 선정	총 10개 이상의 주제 및 자료 선정
화자 선정 기준	성별, 나이, 직업, 지역 등의 비율 편중 없이 선정 및 구성 비율은 주관 기관과 협의	통계청 인구 통계 분포를 참고하여 성별/나이별/지역별 (주 성장지 기준)으로 화자 모집
음성 녹음 및 정제	한 화자별 최대 녹음 시간은 30분(2개 주제)으로 제한 3인 이상 대화 시 최대 녹음 시간은 1시간(4개 주제)으로 진행 가능	2,000명 이상의 화자 참여 정제 후 500시간 이상 음성 수집
음성 이중 전사	발음 전사와 철자 전사를 병행	음성 자료 500시간 이상

<표 1> 사업의 범위

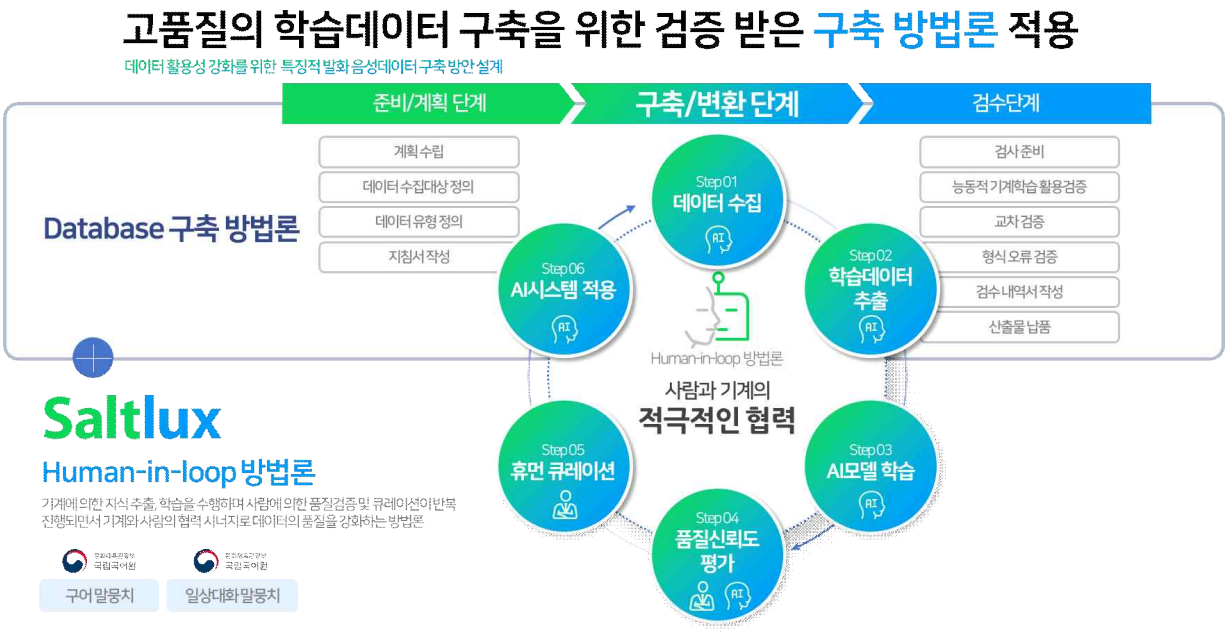
특히 화자의 성별, 나이, 지역이 편중되지 않도록 지역별 최소 할당 인원과 권역별 할당 인원 목표를 세웠으며, 다양한 화자를 모집하기 위해 화자당 최대 녹음 시간은 30분으로 제한하였다.

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

이 사업은 준비/계획 단계, 구축/변환 단계, 검수 단계로 진행되었다. 각 단계를 수립할 때는 음성 녹음, 이중 전사, 원시 말뭉치 구축에 대한 대상 자료별 과업 공정과 주요 활동 절차를 표준화하여 효율적인 말뭉치 구축 체계를 확보하였다. 또한 일상 대화 말뭉치 구축에 적합하도록 자료의 특성을 고려하여 전체 공정을 설계하고 수행하였다.



[그림 2] 데이터 구축 방법론

[그림 3] 일상 대화 말뭉치 전체 구축 공정도

제 2 장

사업 수행

1. 대화 주제 및 제시 자료 선정

대화 주제는 기존에 구축된 국립국어원 일상 대화 말뭉치의 주제를 참고하여 선정하였다. 대화 주제는 화자가 직접 선택하도록 했으며, 주제에 대한 이해를 돕고자 세부 주제를 예시로 들어 주제 선택에 도움을 주었다.

대화 진행 중 주제 이탈 방지를 위해 세부 주제마다 예시 질문들을 제시하여 화자들이 해당 질문들을 활용하여 주제에 최대한 집중할 수 있도록 준비하였다.

<표 2> 대화 주제 및 세부 예시 주제

주제	세부 예시 주제
방송/영화/연예인	텔레비전(TV) 프로그램(드라마, 예능 등), 영화, 연예인
취미	음악 이론, 음악 활동, 추천 음악, 선호 음악 장르, 미술 이론, 미술 활동, 추천 미술관/전시회, 선호 미술 장르, 헬스, 골프, 수영, 테니스, 배드민턴, 크로스핏, 마라톤, 트레일 러닝, 철인3종경기, 등산, 스키, 보드, 서핑, 스노클링, 스쿠버다이빙, 축구, 야구, 배구, 농구, 월드컵, 올림픽, 패럴림픽, 게임, 만화 등
반려 동식물	반려 동식물 관련 경험 및 조언, 추천 등
쇼핑	선호 쇼핑몰, 선호 브랜드, 쇼핑 방식, 중고 거래 등
패션/미용	패션 스타일, 색조 화장, 얼굴, 몸매, 피부 관리, 헤어 스타일링, 네일, 문신, 성형 수술, 외모 콤플렉스, 신발, 가방, 모자, 시계, 안경/선글라스, 보석, 양말, 헤어 액세서리, 방한용품, 일상복 스타일링, 외투, 상의, 하의, 속옷, 실내복, 의례복 스타일링, 정장, 드레스, 파티 복장, 하객 복장, 한복, 국외 전통 의상 등
먹거리	음식, 좋아하는 요리, 요리법, 식재료 쇼핑, 조리 도구, 조리 가전, 간편식 활용법, 맛집 추천, 맛집 후기, 주방 특선(오마카세), 고급 레스토랑, 패밀리 레스토랑, 프랜차이즈 레스토랑, 배달 음식, 길거리 음식, 카페 추천, 카페 후기, 커피, 커피 원두, 차, 다과, 제과점 등
건강/다이어트	건강, 다이어트, 식단, 건강 보조제, 질병 관련 경험과 증상, 부상 관련 경험과 증상, 우울증 등의 정신/심리 관련 어려움, 병원, 요양원, 기타 치료 시설, 양학, 한의학, 민간요법, 기타 치료 행위 등
여행/휴가	국내 여행 경험 및 계획, 해외여행 경험 및 계획, 추천 여행 지역, 추천 여행 활동, 방문 희망 지역, 자유 여행, 패키지여행, 배낭여행, 캠핑, 산, 바다, 호수, 휴양림, 추천 자연 명소 등
생활/주거 환경	가사 활동, 가사 관련 가전, 리모델링, 실내 장식, 실외 장식, 이사, 전세, 월세, 자가, 자취 여부, 행복 주택, 주택, 전원주택, 빌라, 아파트, 오피스텔, 기숙사, 원룸, 투룸 등
가족/관혼상제	가족, 결혼, 출산, 성인식, 결혼식, 장례식, 명절, 제사, 돌잔치, 환갑잔치, 기타 경조사 등
회사, 학교생활	직장 및 학교생활, 업무 내용, 업무 강도, 야근, 회식, 회의, 승진, 복지 제도, 고용 형태, 이직, 헤드헌팅, 초등학교, 중학교, 고등학교, 특수목적 고등학교, 대학교, 대학원, 공교육, 사교육, 인터넷 강의, 비대면 강의, 전공과목, 교양 과목, 선호 과목, 비선호 과목 등
취직	진로, 직업, 취직, 이직, 취준생, 해외 취업, 일반 자격증, 전문 자격증 등

인간관계	친구, 연애, 학교 동기, 직장 동료, 성격상 장점, 성격상 단점, 닮고 싶은 사람의 성격 특성, 닮고 싶지 않은 사람의 성격 특성, 성격 유형 검사(MBTI), 사주, 혈액형별 성향 등
경제/재테크	경제, 재테크, 예금, 적금, 주식, 코인, 부동산, 투자 조언, 가계, 기업, 공기업, 새싹기업(스타트업), 개인의 소비 활동, 기업의 생산 활동, 창업 등
사회 이슈	인공지능(AI) 기술발전, 누리 소통망, 환경 문제, 인구 감소, 고령화, 다문화, 인권, 복지, 지역 간 불균형, 빈부격차, 기타 사회 이슈 등
기타	군대, 추억, 꿈, 인생 목표, 인생 계획, 가치관 등

기존 구축 말뭉치(2019년~2022년)와 2023년 구축 말뭉치를 주제별로 비교하면 다음과 같다.

<표 3> 2019년~2023년 일상 대화 말뭉치 주제 비교

번호	2019년	2020년	2021년	2022년	2023년	세부 주제	비고
1	군대						기타에 포함
2	게임						취미에 포함
3	휴일		휴가	휴가	여행/ 휴가	국내 여행 경험 및 계획, 해외여행 경험 및 계획 등	
4	자동차		대중 교통	대중 교통			제외
5	만화						취미에 포함
6	영화	영화	음악	음악			취미에 포함
7	정치						제외
8	건강/ 다이어트	건강/ 다이어트	건강/ 다이어트	건강/ 다이어트	건강/ 다이어트	건강, 다이어트, 식단, 건강 보조제, 질병 관련 경험과 증상 등	
9	방송/ 연예	방송/ 연예	방송/ 연예	방송/ 연예	방송/ 영화/ 연예인	텔레비전(TV) 프로그램(드라마, 예능 등), 영화, 연예인	
10	스포츠/ 레저	스포츠/ 레저	스포츠/ 레저	스포츠/ 레저/ 취미	취미	음악 이론, 음악 활동, 미술 이론, 미술 활동, 헬스, 골프, 수영 등	
11	먹거리	먹거리	먹거리	먹거리	먹거리	음식, 요리, 좋아하는 요리, 요리법, 식재료 쇼핑, 조리 도구, 조리 가전, 간편식 활용법 등	
12	자연/ 휴양지						여행/휴가에 포함

13	국가/ 지역				사회 이슈	인공지능(AI) 기술 발전, 누리 소통망, 환경 문제, 인구 감소, 고령화 등	
14	문학						제외
15	연애/ 결혼	연애/ 결혼	우정	우정			가족/ 관혼상제에 포함
16	경제/ 재테크		경제/ 재테크	경제/ 재테크	경제/ 재테크	경제, 재테크, 예금, 적금, 주식, 코인, 부동산 등	
17		여행지 (국내/ 해외)					여행/ 휴가에 포함
18		계절/ 날씨					제외
19		회사/ 학교	회사/ 학교	회사/ 학교	회사, 학교생활	직장 및 학교생활, 업무 내용, 업무 강도, 야근, 회식, 회의, 승진 등	
20		선물					제외
21		꿈 (목표)					기타에 포함
22		반려동물	반려동물	반려동물	반려 동식물	반려 동식물 관련 경험 및 조언, 추천 등	
23		아르 바이트	취직	취직	취직	진로, 직업, 취직, 이직, 취준생, 해외 취업, 일반 자격증, 전문 자격증 등	
24		성격			인간관계	친구, 연애, 학교 동기, 직장 동료, 성격상 장점 등	
25		가족	가족				가족/ 관혼상제와 통합
26			쇼핑	쇼핑	쇼핑	선호 쇼핑몰, 선호 브랜드, 쇼핑 방식, 중고 거래 등	
27			관혼상제	가족/ 관혼상제	가족/ 관혼상제	가족, 결혼, 출산, 성인식, 결혼식, 장례식, 명절 등	
28				생활/ 주거 환경	생활/ 주거 환경	가사 활동, 가사 관련 가전, 리모델링, 인테리어 등	
29				기타	기타	군대, 추억, 꿈, 인생 목표, 인생 계획, 가치관 등	
30					패션/ 미용	패션 스타일, 얼굴, 몸매, 화장, 피부 관리, 헤어스타일링 등	

대화 주제는 자유 발화 시 도움이 될 수 있도록 기존 사업의 10~15개 주제 분류 체계를 대분류 16개, 중분류 40개로 세분화하고, 다양한 질문들을 제공하여 주제 중복을 최대한 배제하였다. 대분류의 경우 국립국어원 <신어 조사 사업>, <한국어 교육 어휘 내용 개발(4단계)>를 참고하였다. <한국어 교육 어휘 내용 개발>의 '의미 범주' 분류 체계 중 '정치'와 '종교' 등 데이터의 비윤리성이나 편향성을 일으킬 수 있는 민감한 주제는 제외하였다. 또 일상 대화에 적합한 '여행/휴가, 예술, 체육' 범주를 추가하여 총 16개로 개편하였다. 중분류의 경우 대분류 의미 범주에서 심화된 세부 주제 3~5개로 설정하여 자연스러운 일상 대화를 유도하도록 구성하였다.

<표 4> 한국어 교육 어휘 내용 개발(4단계)

대범주	3단계 소범주	4단계 수정 내용	비고
인간	사람의 종류		
	신체 부위		
	체력 상태		
	생리 현상		
	감각		
	감정		
	성격		
	태도		
	용모		
	능력		
	신체 변화		
	신체 행위		
	신체에 가하는 행위		
	인지 행위		
	소리		
		신체 내부 구성	(추가)
삶	삶의 상태		
	삶의 행위		
	일상 행위		
	친족	친족 관계	(용어 수정)
	가족 행사		
	여가 도구		
	여가 시설		
	여가 활동		
	병과 증상		
	치료 행위		
		치료 시설	(추가)
		약품류	(추가)
식생활	음식		
	채소		
	곡류		
	과일		
	음료		
	식재료		
	조리 도구		
	식생활 관련 장소		
	맛		
	식사 및 조리 행위		
의생활	옷	옷 종류	(세분화)
		옷감	(세분화)
		옷의 부분	(세분화)
	착용물	모자, 신발, 장신구	(용어 수정)
	의생활 관련 장소		
	의복 착용 상태		
	의복 착용	의복 착용 행위	(용어 수정)
	미용 행위		
주생활	건물 종류		
	주거 형태		
	주거 지역		
	가구	생활용품	(통합)
	가전제품		
	일상용품		
	주택 구성		
	주거 상태		
	주거 행위		
	집안일	가사 행위	(용어 수정)
사회생활	인간관계		
	소통 수단		
	교통수단		
	교통 이용 장소		
	매체		
	직장		
	직위		
	직업		

대화 주제와 관련된 세부 주제와 키워드, 그리고 상세 가이드 예시를 함께 제공하여 활발하게 대화가 진행되도록 하였다.

<표 5> 대화 주제별 키워드 및 상세 가이드 예시

번호	주제	세부 주제	상세 가이드 예시
1	방송/ 영화/ 연예인	텔레비전(TV) 프로그램 (드라마, 예능 등), 영화, 연예인	▪ 가장 좋아하는 텔레비전(TV) 프로그램은 무엇입니까? 왜 좋아하나요? ▪ 지금까지 본 영화 중에서 가장 좋아하는 영화는 무엇입니까? 왜 그런가요? ▪ 지금까지 본 영화 중에서 가장 기억에 남는 영화는 무엇입니까? 어떤 장면이 기억에 남나요? ▪ 가장 좋아하는 연예인이 누구입니까? 왜 좋아하게 되었나요? ▪ 웹툰이나 만화책 보는 걸 좋아하시나요? 지금까지 본 만화 중에 가장 재밌었던 만화를 소개해 주세요. ▪ 요즘 보고 있는 텔레비전 프로그램이나 드라마가 있으신가요? ▪ 스릴러, 로맨스, 판타지 등이 있는데요, 어떤 장르의 드라마를 좋아하시나요? ▪ 가장 최근에 본 영화가 무엇인가요? 왜 그 영화를 보게 되었나요? 영화가 재미있었나요? ▪ 가장 좋아하는 배우가 누구인가요? 그 배우를 좋아하는 이유는 무엇인가요?
2	취미 (스포츠/ 레저, 미술, 음악 등)	음악 이론, 음악 활동(보컬, 악기 연주 등), 추천 음악, 선호 음악 장르 등	▪ 요즘 어떤 취미를 즐기세요? ▪ 취미를 시작하게 된 계기는 무엇입니까? ▪ 취미 활동을 하면서 자신에게 어떤 변화가 있었나요? ▪ 최근에 좋아하게 된 노래의 제목과 가수, 좋아하게 된 이유를 알려 주실 수 있으신가요? ▪ 좋아하는 가수나 프로듀서, 아이돌 그룹 등 있으신가요? ▪ 클래식과 대중음악 중 주로 뭘 들으시나요? 어떤 장르의 음악을 선호하시는가요? ▪ 다룰 수 있는 악기가 있으신가요? 아니면 배우고 싶은 악기가 있으신가요? ▪ 최근에 좋아하게 된 미술 작품의 제목과 작가, 좋아하게 된 이유를 알려 주실 수 있으신가요? ▪ 최근에 미술 전시회나 사진 전시회를 가 보신 적 있으신가요? 전시회를 자주 가시는 편인가요? ▪ 좋아하는 미술 작가가 있으신가요? ▪ 어떤 장르의 미술을 선호하시는가요? ▪ 직접 그림을 그려 보신 적 있으신가요? ▪ 취미로 하는 운동이 있으신가요? 그 운동을 선택한 이유는 무엇인가요? ▪ 함께 운동하는 가족이나 친구가 있으신가요? 혹은 속해 있는 운동 모임이 있으신가요?

			▪ 지금 하는 운동의 성취 목표가 있으신가요? ▪ 훗날 혼자서 혹은 가족이나 친구들과 함께해 보고 싶으신 운동이 있으신가요? ▪ 마라톤이나 철인 3종 경기 같은 스포츠 대회에 출전해 보신 적 있으신가요? ▪ 여름에 할 수 있는 레저 활동 중에 가장 좋아하는 혹은 가장 해 보고 싶은 레저 활동이 무엇인가요? ▪ 겨울에 할 수 있는 레저 활동 중에 가장 좋아하는 혹은 가장 해 보고 싶은 레저 활동이 무엇인가요? ▪ 훗날 혼자서 혹은 가족이나 친구들과 함께해 보고 싶으신 레저 활동이 있으신가요? ▪ 스포츠 관람하는 것을 좋아하시나요? ▪ 스포츠 관람을 좋아하신다면, 집에서 텔레비전(TV)이나 휴대폰으로 보는 걸 좋아하시나요? 아니면 직접 구장에 가서 보는 걸 좋아하시나요? ▪ 꼭 한번 관람해 보고 싶은 스포츠 경기가 있으신가요? ▪ 온라인 게임이나 콘솔 게임 등을 좋아하시나요? 좋아하신다면 어떤 게임인지 알려 주실 수 있나요?
3	반려 동식물	반려 동식물 관련 경험 및 팁, 반려 동식물 추천 등	▪ 반려동물(식물)을 키우면 가장 좋은 점은 무엇입니까? ▪ 반려동물(식물)에 대한 나만의 팁은 무엇입니까? ▪ 반려 식물로 키우고 계신 식물이 있으신가요? ▪ 반려 식물을 어떤 계기로 키우게 되었나요? 또 반려 식물과 함께할 때 어떤 감정을 느끼시나요? ▪ 추천할 만한 반려 식물이 있으신가요? ▪ 반려동물로 키우고 계신 동물이 있으신가요? ▪ 반려동물을 어떤 계기로 키우게 되었나요? 또 반려동물과 함께할 때 어떤 감정을 느끼시나요? ▪ 추천할 만한 반려동물이 있으신가요? ▪ 반려동물과 함께 한 경험 중 가장 기억에 남는 경험이 무엇인가요? ▪ 반려 동식물이 아니라, 야외에서 동식물을 본 적 있으신가요? ▪ 동물원이나 식물원 등 인공적으로 조성된 곳에 가는 걸 좋아하시나요?
4	쇼핑	선호 쇼핑몰, 선호 브랜드, 쇼핑 방식	▪ 오프라인 쇼핑과 온라인 쇼핑 중 어느 것을 선호하시나요? 그 이유를 얘기해 주세요. ▪ 가장 좋아하는 쇼핑몰은 어디입니까? 온라인 쇼핑몰과 오프

		(매장 방문, 온라인 쇼핑 등), 중고 거래 등	라인 쇼핑몰로 나눠서 얘기해 주세요. ▪ 최근에 온라인 쇼핑몰에서 구매한 물품은 무엇인가요? 왜 그 쇼핑몰에서 물건을 사게 됐나요? ▪ 가장 좋아하는 브랜드는 무엇입니까? 신발, 의류, 가방 등으로 나눠서 얘기해 주세요. ▪ 쇼핑할 때 가장 먼저 고려하는 점은 무엇입니까?
5	패션/미용(화장, 액세서리 등)	패션 스타일, 메이크업, 얼굴, 몸매, 화장, 피부 관리, 헤어스타일링, 네일 등	▪ 가장 좋아하는 패션 스타일은 무엇입니까? ▪ 본인은 화장을 잘하시는 편인가요? ▪ 패션이나 미용에 대한 팁은 주로 어디서 얻나요? ▪ 패션이나 미용에 대한 팁은 무엇입니까? ▪ 가장 자주 신는 신발은 무엇인가요? ▪ 발이 편한 브랜드가 있으면 추천해 주세요. ▪ 큰 가방이랑 작은 가방 중 어떤 걸 자주 드시나요? ▪ 피어싱 뚫은 곳 있으세요? ▪ 양말까지 신경 써서 맵시 있게 꾸미시는 편이신가요? ▪ 평소에 어떤 옷을 즐겨 입으시나요? ▪ 옷을 살 때 자주 사용하는 쇼핑몰은 어디세요? ▪ 쇼핑할 때 옷 사이즈를 잘 확인할 수 있는 팁 있으세요? ▪ 격식 있는 자리에 갈 때 옷은 어떻게 입으시나요? ▪ 결혼식에 입고 싶은 드레스는 어떤 스타일이신가요? ▪ 최근에 한복 입어 보신 적 있으세요? ▪ 한복을 입은 사람들을 보면 어떤 느낌이 드나요? ▪ 외국 전통 의상을 입어 보신 적 있으세요? ▪ 문신해 보신 적 있으세요? ▪ 성형 수술하고 싶다고 생각한 적 있으세요? ▪ 선호하는 헤어스타일은 어떤 것인가요? ▪ 써 보신 화장품 중에 추천할 만한 것 있으신가요?
6	먹거리 (요리, 맛집, 커피 등)	음식, 요리, 좋아하는 요리, 요리법, 식재료 쇼핑, 조리 도구, 조리 가전, 간편식 활용법 등	▪ 가장 좋아하는 음식은 무엇입니까? ▪ 직접 요리하는 것을 좋아하시는가요? 제일 자신 있는 요리는 무엇인가요? 자신만의 요리법(레시피)이 있나요? ▪ 요즘 즐겨 찾는 음식들은 무엇입니까? ▪ 가장 자신 있는 요리가 무엇인가요? ▪ 에어프라이어로 할 수 있는 요리 중에 해 보신 게 있으신가요? ▪ 밀키트 같은 간편식을 더 맛있게 먹는 법 알고 계신 거 있으세요? ▪ 유명한 맛집에 갔는데 의외로 실망스러웠던 적 없으세요? ▪ 맛집 투어 좋아하시나요? ▪ 맛집을 찾기 위해 어떤 방법을 쓰시나요? 친구 추천, 블로그 참고, 포털 검색 등 여러 가지 방법이 있겠지요.

			▪ 자주 먹는 외식 브랜드가 있으신가요? ▪ 배달 음식 자주 시켜 드시나요? 어떤 음식을 주로 주문하세요? ▪ 가 보신 카페 중에 가장 마음에 들었던 곳은 어디였나요? ▪ 개인 카페나 프랜차이즈 카페 중에 어디를 더 선호하세요? ▪ 커피 맛을 잘 아시면 원두 종류를 추천해 주실 수 있나요? ▪ 단 거 좋아하시나요?
7	건강/ 다이어트	건강, 다이어트, 식단, 건강 보조제, 질병 관련 경험과 증상 등	▪ 건강을 유지하기 위해 무엇을 하고 계시나요? ▪ 지금까지 다이어트를 해 본 적이 있습니까? ▪ 다이어트를 성공했거나 실패했던 경험이 있나요? ▪ 알고 있는 다이어트는 방법에는 뭐가 있나요? ▪ 건강한 식단을 유지하는 가장 좋은 방법은 무엇입니까? ▪ 최근에 아팠던 적이 있으신가요? ▪ 최근에 병원에 가서 진료를 받았던 적이 있으신가요? 어디에, 어떤 증상 때문에 갔었나요? ▪ 운동하거나 레저 활동을 하는 도중 다친 적이 있으신가요? ▪ 코로나19나 기타 감염병에 걸린 적이 있으신가요? 그 당시 증상이 어땠나요? 그리고 어떻게 치료하고 격리 생활을 하셨나요? ▪ 건강 검진을 매년 또는 격년으로 받으시나요? 건강 검진 받은 경험을 이야기해 주세요. ▪ 주변에 우울증으로 어려움을 겪는 사람이 있나요? ▪ 화가 나거나 짜증이 날 때 어떻게 해결하시나요? ▪ 우울하거나 기분이 안 좋을 때 어떻게 해결하시나요? ▪ 큰 병원(종합 병원)에 입원해 보신 적 있나요? ▪ 병원이나 기타 치료 시설을 이용하시며 불편했던 적 있으신가요? ▪ 추천하고 싶은 병원이 있나요? 왜 그런가요? ▪ 몸이 불편해졌을 때 요양 병원에서 여생을 보내는 것에 대해서 어떻게 생각하세요? ▪ 아플 때 어떤 치료법을 선호하시나요? ▪ 감기에 걸렸을 때 병원에 바로 가시나요? 집에서 쉬면서 몸이 회복되기를 기다리나요? 왜 그렇게 하나요? ▪ 아픈 증상이 느껴질 때 바로 병원에 가서 치료를 받으시는 편이신가요? ▪ 성공한 다이어트 비법 있으시면 알려 주세요.
8	여행/	국내 여행 경험	▪ 지금까지 다녔던 여행지 중 가장 기억에 남는 여행지가 어디입니까?

	휴가	및 계획, 해외여행 경험 및 계획, 추천 여행 지역, 추천 여행 활동, 방문 희망 지역, 자유 여행 등	▪ 여행에서 가장 기억에 남는 추억은 무엇입니까? ▪ 여행에서 가장 힘들었던 일이 있었나요? 어떤 일이었는지 얘기해 주세요. ▪ 여행하면서 가장 재미있었던 일이 있었나요? 어떤 일이었는지 얘기해 주세요. ▪ 여행하면서 기분이 나빴던 일이 있었나요? 어떤 일이었는지 얘기해 주세요. ▪ 패키지여행 및 자유 여행의 장단점은 무엇이라고 생각하십니까? ▪ 관광지와 휴양지 중 선호하는 곳은 어디이며 그 이유는 무엇입니까? ▪ 호텔, 펜션, 민박, 게스트하우스 등 숙박지 서비스 중 선호하는 곳은 어디이며 이유는 무엇입니까? ▪ 여행 경험 중 바가지를 당했던 경험이 있습니까? ▪ 국내 여행 중 가장 좋았던 지역이 어디인가요? ▪ 국내 여행 중 추천해 줄 만한 지역과 그 지역의 명소, 축제 등이 있으신가요? ▪ 다음 국내 여행 때 어디로 가 보고 싶으신가요? ▪ 가족과 다시 한번 더 가고 싶은 국내 여행지가 있나요? 왜 그곳에 가족과 함께 가고 싶으세요? ▪ 외국인에게 국내 여행지를 소개한다면 어디를 소개하고 싶으세요? ▪ 해외여행 중 가장 좋았던 지역이 어디인가요? ▪ 해외여행 중 추천해 줄 만한 지역과 그 지역의 명소, 축제 등이 있으신가요? ▪ 해외여행으로 어디에 가고 싶으신가요? ▪ 여행하실 때 자유 여행을 선호하시나요? 아니면 패키지여행을 선호하시나요? ▪ 대중교통이나 자가용을 이용하지 않고 도보로 배낭 여행해 보신 적 있나요? ▪ 캠핑이나 배낭여행(백패킹) 등 야외에서 활동하는 여행을 해 보신 적 있나요? ▪ 산과 바다 중 어느 곳이 더 좋으신가요? ▪ 자연 휴양림이나 수목원에 가 보신 경험이 있으신가요? ▪ 추천해 주실 만한 자연 명소가 있으신가요? 그리고 왜 그곳을 좋아하시나요?
9	생활/ 주거	가사 활동 (빨래, 청소 등),	▪ 현재 살고 있는 동네의 좋은 점과 불편한 점은 무엇입니까? ▪ 당신이 그리는 이상적인 집은 어떤 모습입니까? 집의 크기,

	환경 (대중교통 포함)	가사 관련 가전(세탁기, 의류 관리기, 로봇청소기, 공기청정기 등), 리모델링, 인테리어, 외부 장식, 이사 등	위치, 가격, 인테리어 등에 대해 얘기해 주세요. ▪ 집 내부를 꾸미는 것 중에서 가장 중요하게 생각하는 것은 무엇입니까? 침실, 거실, 주방 등 공간별로 얘기해 주세요. ▪ 집을 더 편안하게 만드는 방법은 무엇이라고 생각합니까? ▪ 빨래나 청소하는 것 좋아하시나요? ▪ 의류 관리기나 건조기를 사용해 보신 적 있으세요? ▪ 추천한다면 추천하는 이유를 얘기해 주세요. 혹시 추천하지 않는다면 그 이유는 무엇인가요? ▪ 리모델링 계획이나 직접 인테리어를 할 계획이 있으신가요? ▪ 어떤 스타일의 인테리어를 좋아하세요? ▪ 내 집 마련 경험이나 계획이 있으신가요? ▪ 부모님이랑 같이 살고 계신가요? ▪ 처음 독립했을 때 기분이 어떠셨나요? ▪ 가족과 함께 살 때, 혼자 살 때로 나누어 각각의 장단점이 무엇이라고 생각하세요? ▪ 현재 살고 있는 주거 형태는 어떤가요? 주택, 빌라, 아파트인가요? ▪ 나중에 주택이나 아파트 중에 어떤 집에 살고 싶으세요? ▪ 집에 마당이나 작은 정원이 있으신가요? ▪ 기숙사에서 생활하시면 불편한 점은 없으세요? ▪ 전원주택에 대한 로망이 있으신가요? ▪ 복층 오피스텔에서 살면 불편할까요?
10	가족/ 관혼상제	가족, 결혼, 출산, 성인식, 결혼식, 장례식, 명절, 제사, 돌잔치, 환갑잔치, 기타 경조사 등	▪ 가족을 위해 할 수 있는 가장 좋은 일은 무엇이라고 생각하십니까? ▪ 결혼이나 출산에 대해 생각해 본 적이 있습니까? ▪ 결혼은 꼭 해야 한다고 생각하시나요? ▪ 가족 관계가 어떻게 되나요? ▪ 이웃들과 왕래가 있으신 편인가요? ▪ 연애 중이신가요? 어떻게 처음 만나게 됐는지, 만난 지 얼마나 됐는지 얘기해 줄 수 있을까요? ▪ 성인식 때 선물 받으셨나요? ▪ 회사 동료 결혼식에 부조금은 얼마 정도가 적당하다고 생각하세요? ▪ 돌잔치에 초대받는다면 어떤 선물을 준비할 거예요? ▪ 돌잔치에 가 본 적이 있나요? ▪ 제사를 지내나요? 일 년에 몇 번 지내나요? 몇 시에 지내나요? 가족들이 몇 명이나 모이나요? ▪ 명절이나 오랜만에 만난 조카에게 용돈을 주나요? 나이에 따라 달리 주나요? 왜 주나요?

			▪ 명절이나 생신 때 부모님께 용돈을 드리나요? ▪ 장례식에 갈 때 가장 신경 쓰는 부분은 무엇인가요? 복장, 인사 예절, 부의금 액수 등 ▪ 장례식장에서 실수한 적 있으세요?
11	회사, 학교생활	직장 및 학교생활, 업무 내용, 업무 강도, 야근, 회식, 회의, 승진, 복지 제도 등	▪ 현재 회사 생활에 대해 어떻게 생각하십니까? ▪ 학교생활에서 가장 좋아하는 수업은 무엇입니까? ▪ 학교생활에서 가장 싫어했던 수업은 무엇입니까? ▪ 원격 근무의 이점과 단점에 대해 어떻게 생각하십니까? ▪ 어떤 일을 하고 계신지 궁금해요. ▪ 업무 강도가 높은 편이신가요? ▪ 야근/회식 자주 하시나요? ▪ 회사 복지 제도가 잘 마련되어 있는 편인가요? ▪ 앞으로 대학교가 많이 없어진다는데 어떻게 생각하세요? ▪ 대학 진학이 반드시 필요하다고 생각하시나요? ▪ 외고나 과학고가 꼭 필요하다고 생각하시나요? ▪ 학창 시절에 사교육을 많이 받으셨나요? ▪ 비대면 강의 들어보신 적 있으세요? ▪ 수강 효과가 좋았던 인터넷 강의가 있으신가요? ▪ 자녀가 생기면 학원을 꼭 보내실 계획이신가요? ▪ 전공이 무엇인가요? ▪ 전공을 선택하신 계기가 있으신가요? ▪ 학창 시절에 좋아했던 과목과 싫어했던 과목은 무엇인가요? ▪ 가장 기억에 남은 수업은 어떤 수업인가요? ▪ 가장 좋은 이미지를 가진 기업은 어떤 곳인가요? ▪ 특정 기업에 대한 보이콧을 어떻게 생각하세요?
12	취직 (취업 준비, 진로, 자격증 등)	진로, 직업, 취직, 이직, 취준생, 해외 취업, 일반 자격증, 전문 자격증 등	▪ 원하는 직업은 무엇이며 왜 관심이 있습니까? ▪ 현재 자신이 하는 일에 대한 만족도를 1점에서 5점까지 중에 하나로 표현해 본다면 몇 점을 주시겠어요? ▪ 왜 그런 점수를 주셨나요? ▪ 현재 직장에 취직하기 위해 어떤 준비를 하셨나요? 준비 과정에서 가장 어려웠던 점은 뭔가요? ▪ (취업자의 경우) 본인의 직장에 취업하기 위해 준비하고 있는 사람에게 전해줄 수 있는 꿀팁이 있을까요? ▪ (미취업자의 경우) 취업을 위해 어떤 준비를 하고 있나요? 준비 과정에서 가장 어려운 점은 뭔가요? ▪ 새로운 직장을 옮긴다면 그곳에서 얻고 싶은 것은 무엇입니까? ▪ 현재 직장에서 만족하는 점과 불만족하는 점은 무엇입니까? ▪ 이직에 대해서 고민해보신 적 있으세요?

			▪ 자격증 가지고 계신 것 있으신가요?
13	인간관계 (연애, 우정, 성격 등)	친구, 연애, 학교 동기, 직장 동료, 성격상 장점, 성격상 단점, 닮고 싶은 사람의 성격 특성 등	▪ 가장 친한 친구는 누구이며 그 친구의 장점은 무엇입니까? ▪ 연애 경험이 있습니까? ▪ 인간관계에서 가장 중요한 것은 무엇이라고 생각하십니까? ▪ MBTI가 어떻게 되세요? 본인 성격과 잘 맞는 것 같으신가요? ▪ 사주나 타로점, 신년 운세 같은 걸 보신 적 있으세요? ▪ 혈액형 사랑법, 혈액형 공부법 등이 유행했는데요, 본인의 혈액형이 그 설명과 잘 맞던가요? ▪ 사주나 타로점 같은 것 보고 들은 말 중에 기억나는 거 있으세요? ▪ 이상형으로 생각하는 사람의 성격 유형이나 특징은 무엇인가요? ▪ 자주 연락하는 친한 친구가 몇 명 정도 되세요? ▪ 직장 동료 중에 친하게 지내는 사람이 있나요? 친해지게 된 계기가 있을까요? ▪ 소개팅은 몇 번 해 보셨나요?
14	경제/ 재테크 (주식, 부동산 등)	경제, 재테크, 예금, 적금, 주식, 코인, 부동산, 투자 조언, 가계(개인) 등	▪ 현재 어떤 경제 상황에 처해 있습니까? ▪ 재테크를 위해 무엇을 하고 계시나요? ▪ 재테크에 대한 자신만의 팁은 무엇입니까? ▪ 한 달에 어디에 가장 돈을 많이 쓰는 편이신가요? ▪ 이런 데 돈 쓰는 게 가장 아깝다라고 느끼는 부분이 있나요? ▪ 생각해 두신 창업 아이템 있으세요? ▪ 지금 투자하고 계신 종목은 무엇인가요? ▪ 적금은 몇 개 정도 들고 계세요? ▪ 앱테크 추천해 주실 만한 것 있나요?
15	사회 이슈 (환경, 고령화 등 사회 이슈)	인공지능(AI) 기술 발전, 누리 소통망, 환경 문제, 인구 감소, 고령화, 다문화, 인권, 복지, 지역 간 불균형 등	▪ 인공지능(AI) 기술이 우리의 관계, 업무, 건강 및 여가 시간에 미치는 영향은 무엇이 될 것이라 생각하십니까? ▪ 전기 자동차, 자율 주행 차량, 대중교통 시스템의 장단점은 무엇이라고 생각하십니까? ▪ 챗지피티 또는 바드에 대해 들어보셨나요? 알고 있는 바를 얘기해 주세요. ▪ 챗지피티 또는 바드를 사용해 보셨나요? 사용 경험을 얘기해 주세요. ▪ 누리 소통망이 우리 사회에 미치는 영향은 무엇이라고 생각합니까? ▪ 누리 소통망 중에서 어떤 걸 주로 사용하시나요? 그 이유는 뭔가요? (카톡, 인스타, 트위터 등) ▪ 누리 소통망을 하루에 몇 시간 정도 사용하시나요?

			▪ 어떤 날씨와 계절을 좋아하세요? ▪ 화창한 날씨에 야외 활동을 많이 하시는 편인가요? ▪ 엄청 더운 것과 엄청 추운 것을 비교하면 어떤 쪽이 더 견딜 만한가요? ▪ 기후 변화가 우리의 일상에 미치는 영향은 무엇이라고 생각합니까? ▪ 기억에 남는 지진, 태풍, 폭염 등 자연재해가 있으신가요? 그 때의 상황은 어땠나요? ▪ 2022년에 슈퍼 태풍인 힌남노가 와서 경북 포항과 경주 등 영남 해안지역에 집중호우가 쏟아져 곳곳이 침수되면서 이로 인한 재산 및 인명 피해가 속출했었습니다. 이러한 피해를 막기 위해 어떤 대비가 필요하다고 생각하세요? ▪ 2017년에는 포항에 5.4 규모의 지진이 나서 큰 피해가 있었고, 최근 동해안에 여러 번 지진이 있었습니다. 그 당시 지진에 대해 기억하는 바가 있나요? 우리는 이런 지진에 어떻게 대비해야 할까요? ▪ 지방에는 뮤지컬, 연극 등을 볼 시설이 부족하다고 하는데 알고 있으신가요? 이러한 지역 격차를 어떻게 해소할 수 있을까요? ▪ 지방의 교통 인프라가 부족하다고 하는데 혹시 알고 있으신가요? 이러한 지역 격차를 어떻게 해소할 수 있을까요? ▪ 고독사에 대한 뉴스를 접할 때 어떤 생각이 드시나요? ▪ 안락사에 대해서 생각해 보신 적 있으세요? ▪ 조력 자살(의료진이 자살을 돕는 행위를 하는 것)을 법적으로 허용하는 나라도 있습니다. 이에 대해 어떻게 생각하시나요? ▪ 인구가 감소하고 있다는 뉴스는 많이 접하셨을 텐데요, 이와 관련한 문제를 몸소 느끼신 적 있나요? ▪ 가장 시급하게 해결되어야 할 인권 문제가 무엇이라고 생각하세요?
16	기타 (군대, 특별한 추억, 꿈, 목표)	군대, 추억, 꿈, 인생 목표, 인생 계획, 가치관 등	▪ 어린 시절 있었던 일 중에서 가장 기억에 남은 일이 있나요? ▪ 친구나 가족과 관련하여, 오래된 특별한 추억이 있습니까? ▪ 이루고 싶은 꿈이나 목표가 있습니까? 자신의 작지만 소중한 버킷 리스트에 대해 이야기해 주세요. ▪ 인생에서 꼭 이루어야겠다고 생각하는 목표가 있으신가요? ▪ 20대, 30대, 40대, 50대, 60대 시기별 대략적인 계획이 어떻게 되시나요? ▪ 지금까지의 인생에서 가장 값진 경험은 무엇이었나요? ▪ 지금까지의 인생에서 가장 하지 말았어야 했던 경험은 무엇이었나요? ▪ 앞으로의 인생에서 꼭 유지하고 싶은 생각이나 태도가 있으신가요?

2. 전문가 자문 회의 진행

사업의 원활한 수행과 체계적인 말뭉치 구축을 위해 국어국문학, 전산언어학 및 음성 인식 분야 전문가로 자문단을 구성하고 자문 회의를 두 차례 진행하였다. 자문 회의는 일관성 있는 고품질의 원시 말뭉치 구축을 위한 음성 전사 지침 관련 전문가 자문을 요청하였다.

<표 6> 전문가 자문 회의 상세

자문 회의	
구분	**내용**
회의 일시	• 2023년 5월 19일 15:00~16:00
회의 방식	• 대면 회의
자문 위원	• 윤 승 박사_한국전자통신연구원(ETRI) • 오재혁 교수_건국대학교 • 김학수 교수_건국대학교 • 김재은 박사_솔트룩스 • 신동환 박사_팀벨
회의 주제	• 발화 단위, 억양구 단위의 인공지능 활용 효용성
회의 내용	• 데이터 활용성을 고려한 분할 기준 설정의 문제 논의 • 문장 부호 사용(‘?’, ‘~’)의 적절성 문제 논의 • 인공지능 활용을 위한 효과적인 발음 전사 방안 논의 • 활용성을 고려한 추가 메타 데이터 논의 • 화자 정보 및 화자 비율 논의

3. 화자 구성 및 모집

모집 대상은 전국 16개 시·도에 거주하는 성인 남녀로, 성별, 나이, 지역 등의 비율이 편중되지 않도록 사전에 참가자 할당표를 설계해 모집하였다. 해당 분류는 2023년 3월 기준 행정안전부 주민등록 인구 통계를 기준으로 하였다. 지역은 현 거주지가 아닌 주 성장지 기준으로 16개 지역(서울, 인천, 대전, 대구, 부산, 광주, 울산, 경기, 강원, 충남, 충북, 경남, 경북, 전남, 전북, 제주)으로 할당하였으며, 나이는 10세 단위로 나누어 할당하였다. 현실적으로 녹음이 어려울 것으로 예상되는 0~9세는 제외하였고 60세 이상의 나이대는 하나의 그룹으로 지정하였다. 2012년에 출범한 세종시는 주 성장지로 하는 대상자를 찾기 어려워 대전 지역으로 통합하였다.

화자 모집에 있어 최대한 인구 비율에 맞도록 진행하고자 했으나, 대상을 찾기가 어려운 경우에는 지역은 권역 전체로 할당을 맞추고 지역별로는 기본 할당량의 최소 50%를 맞추어 진행하였다.

<표 7> 화자 할당표 설계 기준

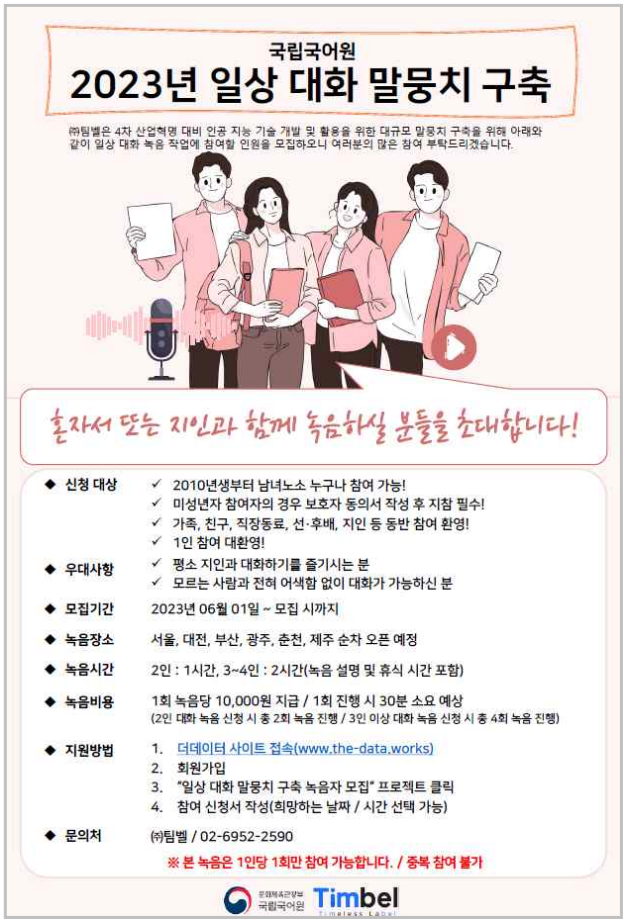
구분	기준
모집단	• 행정안전부, 2023년 3월 기준 주민등록 인구 통계 기준 활용
고려 변수	• 성별: 남자/여자 • 나이대: 10대/20대/30대/40대/50대/60대 이상 • 지역(주 성장지 기준): 서울/인천/대전/대구/부산/광주/울산/경기/강원/충남/충북/경남/경북/전남/전북/제주 * 세종시의 경우, 2012년 출범한 세종시를 주 성장지로 하는 대상자를 찾기 어려워 대전 지역에 편입함.
배분 방법	• 제곱근 비례 배분
표본 할당	• 지역별: 비례 할당 • 성별×나이별: 균등 할당

행정구역 / 연령별 성별 비율 수집 목표 (안)		연령	10대		20대		30대		40대		50대		60대 이상		총
		성별	남	여	남	여	남	여	남	여	남	여	남	여	
		비율	5%	5%	15%	15%	10%	10%	10%	10%	5%	5%	5%	5%	100%
수도권	서울특별시	20.0%	21명	21명	63명	63명	42명	42명	42명	42명	21명	21명	21명	21명	420명
	인천광역시	5.5%	6명	6명	17명	17명	12명	12명	12명	12명	6명	6명	6명	6명	116명
	경기도	25.0%	26명	26명	79명	79명	53명	53명	53명	53명	26명	26명	26명	26명	525명
영남권	부산광역시	7.0%	7명	7명	22명	22명	15명	15명	15명	15명	7명	7명	7명	7명	147명
	울산광역시	2.5%	3명	3명	8명	8명	5명	5명	5명	5명	3명	3명	3명	3명	53명
	대구광역시	4.5%	5명	5명	14명	14명	9명	9명	9명	9명	5명	5명	5명	5명	95명
	경상북도	5.0%	5명	5명	16명	16명	11명	11명	11명	11명	5명	5명	5명	5명	105명
	경상남도	6.5%	7명	7명	20명	20명	14명	14명	14명	14명	7명	7명	7명	7명	137명
호남권	광주광역시	2.5%	3명	3명	8명	8명	5명	5명	5명	5명	3명	3명	3명	3명	53명
	전라북도	3.5%	4명	4명	11명	11명	7명	7명	7명	7명	4명	4명	4명	4명	74명
	전라남도	3.5%	4명	4명	11명	11명	7명	7명	7명	7명	4명	4명	4명	4명	74명
충청권	대전광역시	2.5%	3명	3명	8명	8명	5명	5명	5명	5명	3명	3명	3명	3명	53명
	충청북도	3.0%	3명	3명	9명	9명	6명	6명	6명	6명	3명	3명	3명	3명	63명
	충청남도	4.0%	4명	4명	13명	13명	8명	8명	8명	8명	4명	4명	4명	4명	84명
강원도	강원도	3.0%	3명	3명	9명	9명	6명	6명	6명	6명	3명	3명	3명	3명	63명
제주권	제주특별자치도	2.0%	2명	2명	6명	6명	4명	4명	4명	4명	2명	2명	2명	2명	42명
	합계	100.0%	105명	105명	315명	315명	210명	210명	210명	210명	105명	105명	105명	105명	2100명

[그림 4] 성별, 나이별, 지역별 기준 수집 목표

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 나이대 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×나이별×지역별 할당에 맞는 화자를 찾기가 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 5] 화자 모집 홍보 포스터

화자 모집 진행 중 SBS 방송사에서 일상 대화 말뭉치 구축에 대한 촬영 의뢰가 들어왔다. 화자 모집의 어려움을 겪고 있는 상황에서 다양한 화자를 모집하기 위한 홍보 효과를 기대하고 국립국어원과 협의하여 촬영을 수락했다. 주제에 맞는 대화를 나누고 돈을 벌 수 있는 이색 아르바이트를 소개해 주는 내용으로 방송 촬영은 7월 28일(금) 가산디지털단지에 있는 서울 녹음실에서 진행하였고 방영은 모닝와이드 3부 8,163회 '이색 알바의 세계'라는 제목으로 8월 4일(금) 아침에 소개되었다.

[그림 6] SBS 모닝와이드 방송 화면

4. 작업자 선발 및 교육

4.1. 녹음 진행 요원 선발 및 교육

녹음에는 서울, 대전, 대구, 부산, 광주, 강원, 제주 등 총 16개 지역 거주자 2,168명이 참여하였다. 다수의 화자가 참여하는 만큼 원활한 녹음 진행을 위하여 지역별로 녹음 진행 요원이 투입되었다. 투입된 진행 요원은 총 11명으로, 대면 교육을 이수하고 시뮬레이션 평가를 통과한 자를 최종 선발하여 녹음 파일의 품질 유지와 녹음 일정에 차질이 없도록 진행하였다. 진행 요원의 선발 기준은 유사 작업 경험자 중 전문 녹음 장비 작동 경험이 있는 자를 우선 선발하였다.

<표 8> 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 인적 사항 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 진행하였다. 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 고객 만족(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 상황과 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 9> 진행 요원 교육

구분	내용
교육 일시 및 장소	• 2023년 6월 2일 팀벨 3층 교육장
교육자	• 김선아(㈜팀벨)
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 • 질의응답


[그림 7] 진행 요원 교육 사진

일상 대화 말뭉치 구축 진행요원 수집 매뉴얼

1. 진행요원 녹음 수집 절차

녹음 환경 기준 체크 → 녹음 인원 명단 체크 → 녹음 인원 안내 및 교육 → 녹음 진행 → 발화 순서 및 대화 주제 구글시트에 정리 → 녹음 파일 저장 → 파일질라 파일 업로드

1.1. 녹음 환경 기준 체크

(1) 진행 요원 및 녹음 신청자는 반드시 마스크를 착용
(2) 진행 요원 및 녹음 신청자 발열 체크, 출입 명부 작성, 입실 전 손 소독 필수
(3) 하루 3회 이상 녹음실 환기 및 소독을 진행
(4) 2-4인의 화자가 자유롭게 이야기할 수 있는 통제 환경 준비
(5) 녹음실은 외부와 차단된 상태로 대화에 참여한 화자들만 대화할 수 있도록 구성

1.2. 녹음 인원 명단 체크

(1) 담당 지역 탭으로 들어가서 참석자 인원을 확인

(2) 정리된 명단(시트) 확인 후 각 타임별 녹음 10분 전까지 미참석 시 전화하여 확인 후 특이사항 및 미참석 인원은 담당자에게 연락

1.3. 녹음 인원 안내 및 교육

(1) 녹음 안내 사항 녹음 신청자들에게 안내
1) 마이크에 소음이 들어가지 않도록 주의한다.
2) 마이크 장비를 건드리지 않는다.
3) 휴대폰은 진동으로 하고 책상 위에 두지 않는다.
4) 발언이 겹치지 않도록 발언 중간에 얘기하지 않는다.
5) 발언 시 말끝을 흐리지 말고 명확하게 마무리한다.
6) 진행요원이 녹음 끝났다고 안내하기 전까지는 말을 하지 않는다.
7) 녹음 진행 시 욕설 및 비속어는 사용하지 않는다.

(2) 녹음 시작 전 보이스레코더가 켜져 있는지 다시 한번 확인
(켜져 있으면 빨간 불이 깜박거림)
(3) 발화자의 모든 발화 시작 전에는 2초의 휴지가 발생할 수 있도록 필히 안내
(4) 녹음 시 인사말, 안부 등 주제와 벗어나는 대화는 지양하도록 안내
(5) 녹음 시작과 끝에 진행요원의 OK 사인을 받은 후 진행할 수 있도록 정확히 안내

1.4. 녹음진행

(1) 녹음 안내 사항에 따라서 녹음을 진행

(2) 음성 자료 수집 시 점검 사항

검사항목	검사내용	처리내용
발언자 문제	* 발언자의 목소리가 작음 * 발언자의 말이 빠름 * 발언자의 발음이 부정확	녹음 시 모니터링을 통한 사전 주의로 문제사항 확인 및 개선 조치
대화상 문제	* 정해진 주제에서 벗어난 발언 * 과도하게 한 사람 위주의 발언	
녹음 환경 문제	* 외부 잡음이 심함 * 마이크 착용 문제로 인한 불필요한 잡음	녹음 시간 및 주변환경을 고려하고, 여분 마이크 준비로 오류사항 조치
시스템 문제	* 녹음장비 문제로 녹음상태 불량 * 녹음 프로그램 오류로 파일 손상	녹음전후 장비 및 시스템 점검으로 오류 최소화

[그림 8] 녹음 교육 자료 일부

4.2. 전사자 선발 및 교육

이 사업에는 전체 2,168명의 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 하는 전문 속기사를 우선 선발하였다.

원활한 전사를 위해 약 14명의 전사자를 교육하였다. 억양구 기준으로 구분하는 전사 단위와 전사 지침의 이해, 전사 완성도가 전사자 간에 차이가 생길 수 있어 전체 전사자를 대상으로 대면 교육을 진행하였다. 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.

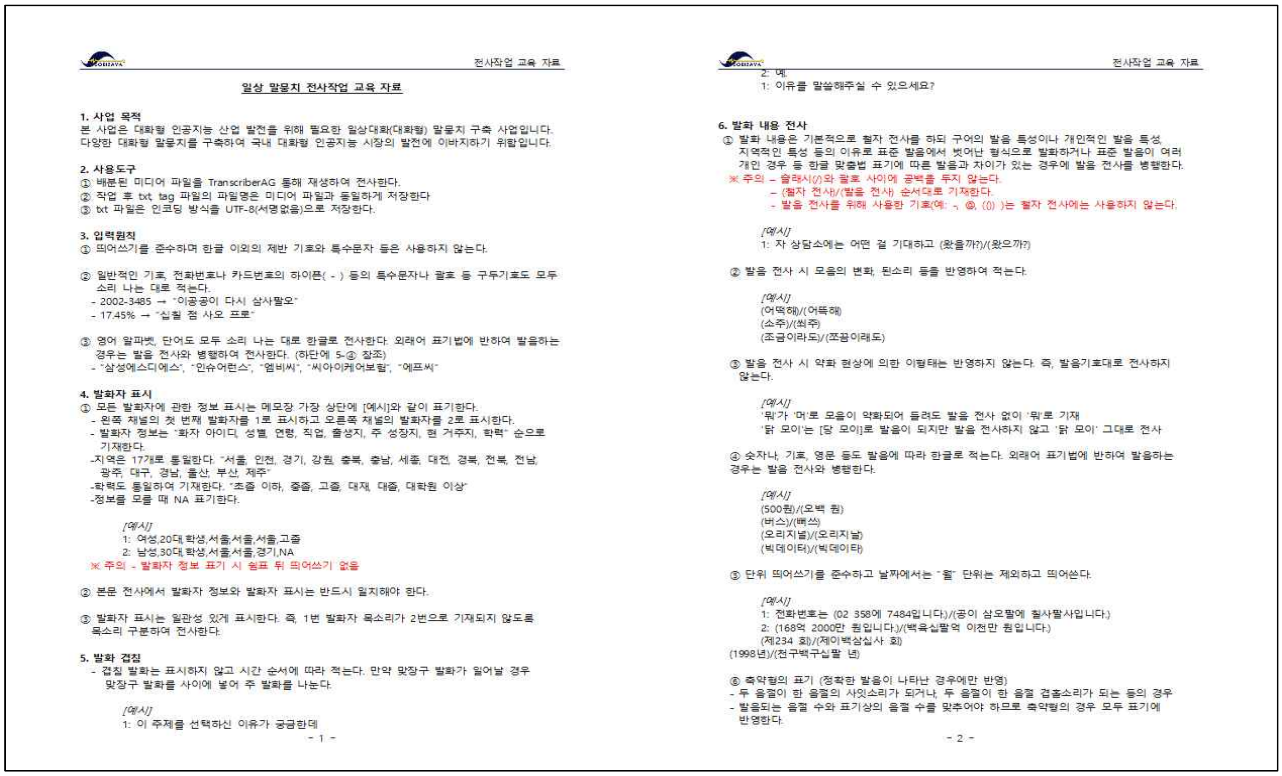
이러한 과정을 통해 최종 선발된 전사자들은 보안 서약서를 작성한 후 전사에 참여하였다. 교육을 마친 전사자는 샘플 전사 후 1차 결과물에 대해 교정받고, 두세 차례 수정 전사를 진행하며 충분히 지침을 숙지한 상태에서 본 전사에 투입되었다. 하루에 1인당 약 15분 녹음 파일 4개를 전사하는 것을 기준으로 일평균 8명 정도가 투입되었다.

<표 10> 전사자 선발

구분	선발 기준 및 운영 내용
선발 기준	• 유사 작업 경험자 • 전사 지침에 대해 정확히 이해하고 있는 자 • 전사 교육 이수 및 샘플 평가 통과한 자
투입 인원	• 일평균 8명 정도 진행
운영	• 방언이 포함된 음성 파일은 해당 지역 출신 전사자에게 우선 배정함.

<표 11> 전사자 교육

구분	내용
교육 일시 및 장소	• 2023년 6월 15일 팀벨 2층 전사실
교육자	• 송혜주 주임(㈜팀벨)
교육 내용	• 사업의 배경 및 목적, 전사 절차와 방법 • 전사 사용 도구 • 전사 지침 및 유의 사항 • 한글 맞춤법 주요 내용 • 보안 교육 • 질의응답



전사작업 교육 자료

일상 말뭉치 전사작업 교육 자료

1. 사업 목적
본 사업은 대화형 인공지능 산업 발전을 위해 필요한 일상대화(대화형) 말뭉치 구축 사업입니다. 다양한 대화형 말뭉치를 구축하여 국내 대화형 인공지능 시장의 발전에 이바지하기 위함입니다.

2. 사용도구
① 배분된 미디어 파일을 TranscriberAG 통해 재생하여 전사한다.
② 작업 후 txt tag 파일의 파일명은 미디어 파일과 동일하게 저장한다
③ txt 파일은 인코딩 방식을 UTF-8(서명없음)으로 저장한다.

3. 입력원칙
① 띄어쓰기를 준수하며 한글 이외의 제반 기호와 특수문자 등은 사용하지 않는다.

② 일반적인 기호, 전화번호나 카드번호의 하이픈(-) 등의 특수문자나 괄호 등 구두기호도 모두 소리 나는 대로 적는다.
- 2002-3485 → "이공공이 다시 삼사팔오"
- 17.45% → "십칠 점 사오 프로"

③ 영어 알파벳, 단어도 모두 소리 나는 대로 한글로 전사한다. 외래어 표기법에 반하여 발음하는 경우는 발음 전사와 병행하여 전사한다. (하단에 5-③ 참조)
- "삼성에스디에스", "인슈어런스", "엠씨씨", "씨아이케어보험", "에프씨"

4. 발화자 표시
① 모든 발화자에 관한 정보 표시는 메모장 가장 상단에 [예시]와 같이 표기한다.
- 왼쪽 채널의 첫 번째 발화자를 1로 표시하고 오른쪽 채널의 발화자를 2로 표시한다.
- 발화자 정보는 "화자 아이디, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력" 순으로 기재한다.
-지역은 17개로 통일한다. "서울, 인천, 경기, 강원, 충북, 충남, 세종, 대전, 경북, 전북, 전남, 광주, 대구, 경남, 울산, 부산, 제주"
-학력도 통일하여 기재한다. "초졸 이하, 중졸, 고졸, 대재, 대졸, 대학원 이상"
-정보를 모를 때 NA 표기한다.

[예시]
1: 여성,20대,학생,서울,서울,서울,고졸
2: 남성,30대,학생,서울,서울,경기,NA
※ 주의 - 발화자 정보 표기 시 쉼표 뒤 띄어쓰기 없음

② 본문 전사에서 발화자 정보와 발화자 표시는 반드시 일치해야 한다.

③ 발화자 표시는 일관성 있게 표시한다. 즉, 1번 발화자 목소리가 2번으로 기재되지 않도록 목소리 구분하여 전사한다.

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

[예시]
1: 이 주제를 선택하신 이유가 궁금한데
- 1 -

전사작업 교육 자료

2: 예
1: 이유를 말씀해주실 수 있으세요?

6. 발화 내용 전사
① 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 병행한다.
※ 주의 – 슬래시(/)와 괄호 사이에 공백을 두지 않는다.
– (철자 전사)/(발음 전사) 순서대로 기재한다.
- 발음 전사를 위해 사용한 기호(예: ., @, (()))는 철자 전사에는 사용하지 않는다.

[예시]
1: 자 상담소에는 어떤 걸 기대하고 (왔을까?)/(왔으까?)

② 발음 전사 시 모음의 변화, 된소리 등을 반영하여 적는다.

[예시]
(어떡해)/(어뜩해)
(소주)/(쐬주)
(조금이라도)/(쪼끔이래도)

③ 발음 전사 시 약화 현상에 의한 이형태는 반영하지 않는다. 즉, 발음기호대로 전사하지 않는다.

[예시]
'뭐'가 '머'로 모음이 약화되어 들려도 발음 전사 없이 '뭐'로 기재
'닭 모이'는 [당 모이]로 발음이 되지만 발음 전사하지 않고 '닭 모이' 그대로 전사

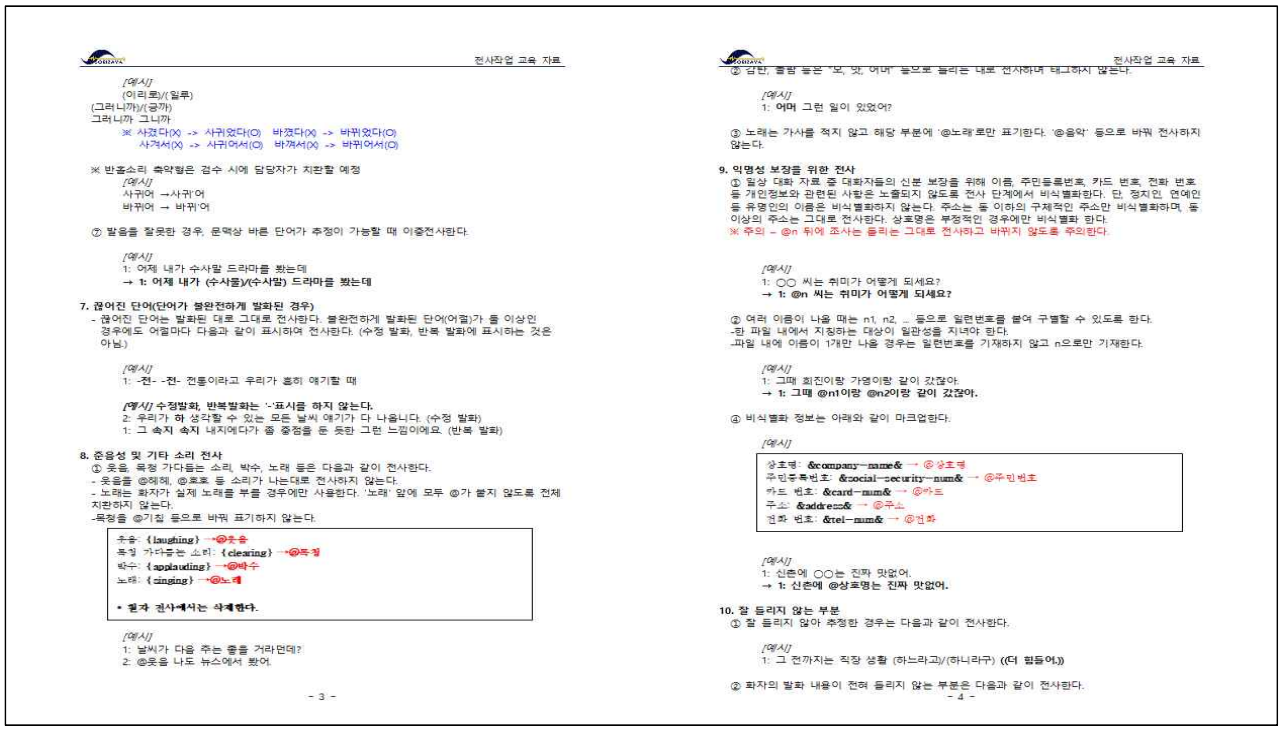
④ 숫자나, 기호, 영문 등도 발음에 따라 한글로 적는다. 외래어 표기법에 반하여 발음하는 경우는 발음 전사와 병행한다.

[예시]
(500원)/(오백 원)
(버스)/(뻐쓰)
(오리지널)/(오리지날)
(빅데이터)/(빅데이타)

⑤ 단위 띄어쓰기를 준수하고 날짜에서는 '월' 단위는 제외하고 띄어쓴다.

[예시]
1: 전화번호는 (02 358에 7484입니다.)/(공이 삼오팔에 칠사팔사입니다.)
2: (168억 2000만 원입니다.)/(백육십팔억 이천만 원입니다.)
(제234 회)/(제이백삼십사 회)
(1998년)/(천구백구십팔 년)

⑥ 축약형의 표기 (정확한 발음이 나타난 경우에만 반영)
- 두 음절이 한 음절의 사잇소리가 되거나, 두 음절이 한 음절 겹홀소리가 되는 등의 경우
- 발음되는 음절 수와 표기상의 음절 수를 맞추어야 하므로 축약형의 경우 모두 표기에 반영한다.
- 2 -



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)

※ 반홀소리 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 수사발 드라마를 봤는데
→ 1: 어제 내가 (수사물)/(수사발) 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. 준용성 및 기타 소리 전사
① 웃음, 특정 가다듬는 소리, 박수, 노래 등은 다음과 같이 전사한다.
- 웃음을 @헤헤, @호호 등 소리가 나는대로 전사하지 않는다.
- 노래는 화자가 실제 노래를 부를 경우에만 사용한다. '노래' 앞에 모두 @가 붙지 않도록 전체 치환하지 않는다.
-묵정을 @기침 등으로 바꿔 표기하지 않는다.

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

• 필자 전사에서는 삭제한다.

[예시]
1: 날씨가 다음 주는 좋을 거라면데?
2: @웃을 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 바뀌지 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.

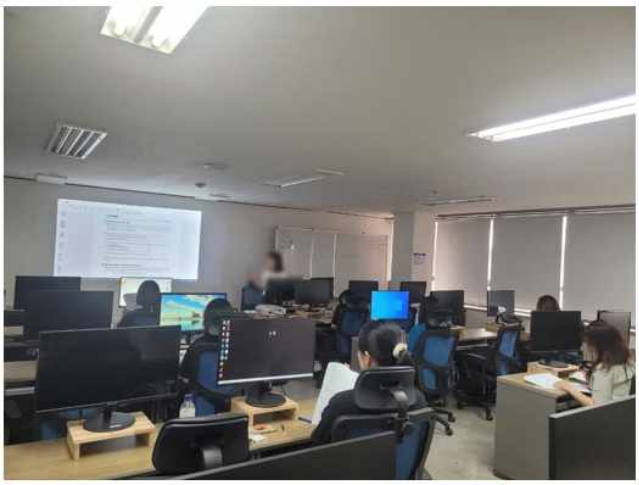
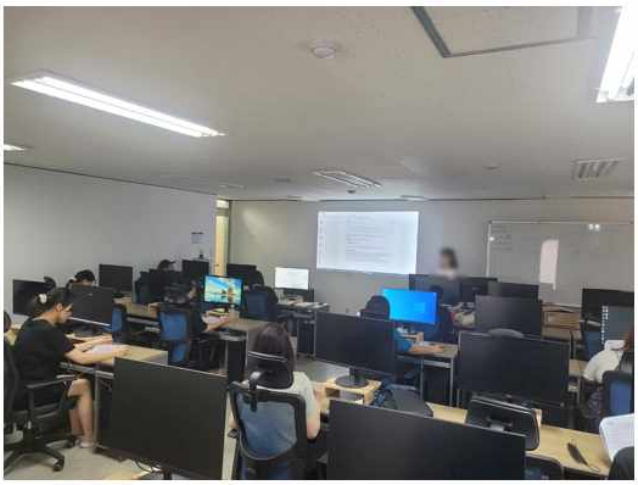
[예시]
1: 그 전까지는 직장 생활 (하느라고)/(하니라구) **((더 힘들어.))**

② 화자의 발화 내용이 전혀 들리지 않는 부분은 다음과 같이 전사한다.
- 4 -

[그림 9] 전사 교육 자료 일부

<표 12> 전사 작업자 교육 상세

구분	내용
작업자 전체 교육	• 일시: 2023년 6월 15일 • 장소: ㈜팀벨 교육장 • 교육자: 송혜주 주임 • 교육 참석 인원: 10명 • 교육 내용: 저작 도구 사용법, 전사 지침 및 유의 사항
신규 작업자 교육 1차	• 일시: 2023년 7월 10일 • 장소: ㈜팀벨 교육장 • 교육자: 송혜주 주임 • 교육 참석 인원: 2명 • 교육 내용: 전사 지침 및 유의 사항
신규 작업자 교육 2차	• 일시: 2023년 8월 7일 • 장소: ㈜팀벨 교육장 • 교육자: 송혜주 주임 • 교육 참석 인원: 2명 • 교육 내용: 전사 지침 및 유의 사항
전사 작업자 비대면 회의	• 일시: 2023년 10월 23일 • 장소: ㈜팀벨 교육장 • 교육자: 송혜주 주임 • 교육 참석 인원: 14명 • 교육 내용: 변경 지침 및 잦은 오류 유형에 대한 피드백

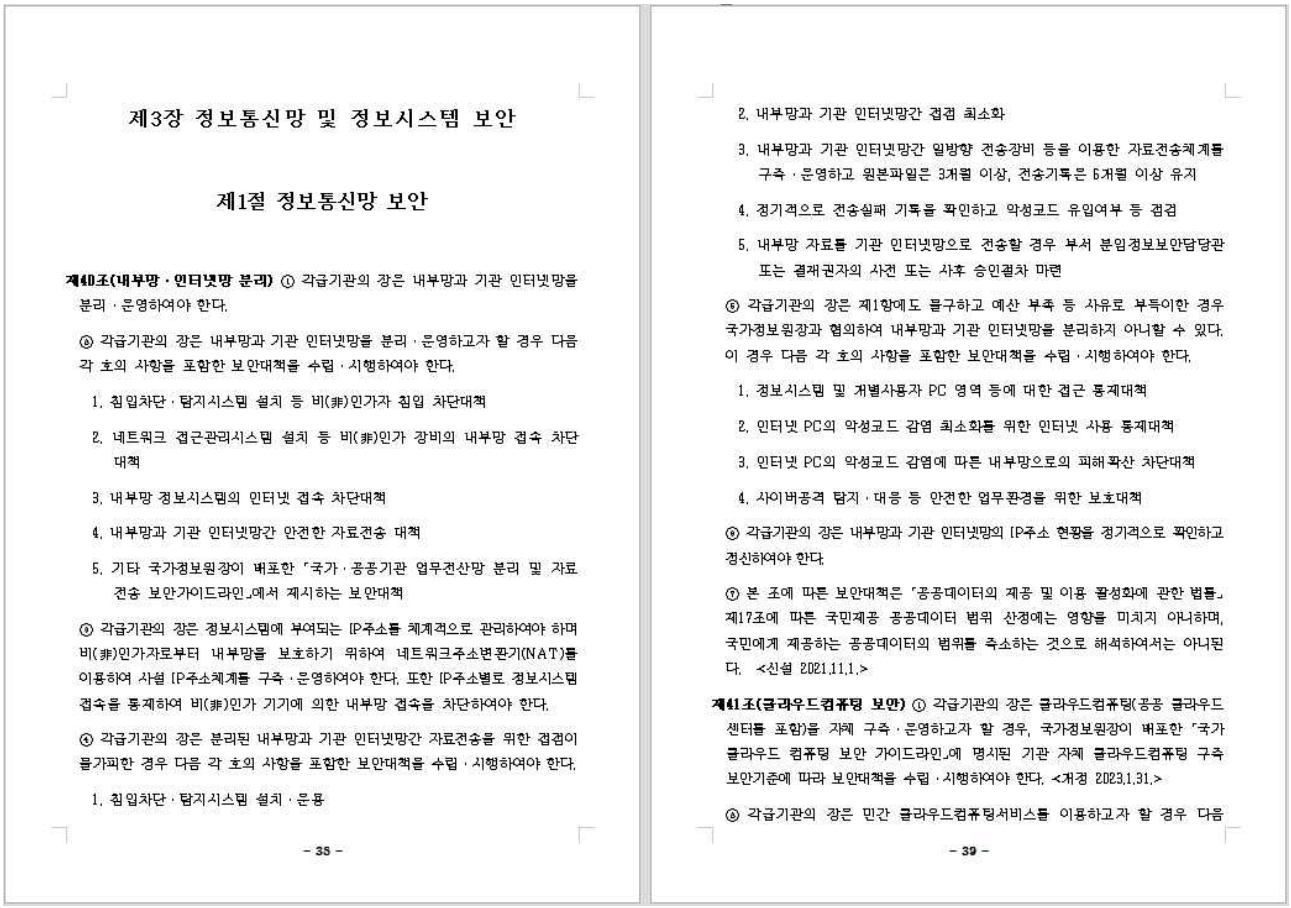
[그림 10] 전사 작업자 교육 사진

4.3. 개인정보 보호 및 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 개인정보 보호 교육, 자료에 대한 보안 관리 교육, 사무실·장비에 대한 보안 관리 교육, 내·외부망 접근 시 보안 관리 교육이다.

<표 13> 보안 교육 내용

구분	보안 교육 내용
개인정보 보호	• 사업 진행 중 알게 되는 화자의 인적 사항에 대한 비밀 보장 • 대화 청취 중 알게 되는 화자의 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 개인용 컴퓨터(PC)에만 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 잠금장치가 된 보관함에 보관
사무실·장비에 대한 보안 관리	• 폐쇄회로 텔레비전(CCTV)·잠금 장치 등 비인가자의 출입 통제 • 개인용 컴퓨터(PC)는 패스워드를 설정하고 상시 점검 및 악성코드 감염 차단을 위한 저장 매체 자동 점검이 될 수 있도록 설정 • 인가된 유에스비(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 필요시 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



제3장 정보통신망 및 정보시스템 보안

제1절 정보통신망 보안

제40조(내부망·인터넷망 분리) ① 각급기관의 장은 내부망과 기관 인터넷망을 분리·운영하여야 한다.

② 각급기관의 장은 내부망과 기관 인터넷망을 분리·운영하고자 할 경우 다음 각 호의 사항을 포함한 보안대책을 수립·시행하여야 한다.

1. 침입차단·탐지시스템 설치 등 비(非)인가자 침입 차단대책
2. 네트워크 접근관리시스템 설치 등 비(非)인가 장비의 내부망 접속 차단대책
3. 내부망 정보시스템의 인터넷 접속 차단대책
4. 내부망과 기관 인터넷망간 안전한 자료전송 대책
5. 기타 국가정보원장이 배포한 「국가·공공기관 업무전산망 분리 및 자료전송 보안가이드라인」에서 제시하는 보안대책

③ 각급기관의 장은 정보시스템에 부여되는 IP주소를 체계적으로 관리하여야 하며 비(非)인가자로부터 내부망을 보호하기 위하여 네트워크주소변환기(NAT)를 이용하여 사설 IP주소체계를 구축·운영하여야 한다. 또한 IP주소별로 정보시스템 접속을 통제하여 비(非)인가 기기에 의한 내부망 접속을 차단하여야 한다.

④ 각급기관의 장은 분리된 내부망과 기관 인터넷망간 자료전송을 위한 접점이 불가피한 경우 다음 각 호의 사항을 포함한 보안대책을 수립·시행하여야 한다.

1. 침입차단·탐지시스템 설치·운용

2. 내부망과 기관 인터넷망간 접점 최소화
3. 내부망과 기관 인터넷망간 일방향 전송장비 등을 이용한 자료전송체계를 구축·운영하고 원본파일은 3개월 이상, 전송기록은 6개월 이상 유지
4. 정기적으로 전송실패 기록을 확인하고 악성코드 유입여부 등 점검
5. 내부망 자료를 기관 인터넷망으로 전송할 경우 부서 분임정보보안담당관 또는 결재권자의 사전 또는 사후 승인절차 마련

⑤ 각급기관의 장은 제1항에도 불구하고 예산 부족 등 사유로 부득이한 경우 국가정보원장과 협의하여 내부망과 기관 인터넷망을 분리하지 아니할 수 있다. 이 경우 다음 각 호의 사항을 포함한 보안대책을 수립·시행하여야 한다.

1. 정보시스템 및 개별사용자 PC 영역 등에 대한 접근 통제대책
2. 인터넷 PC의 악성코드 감염 최소화를 위한 인터넷 사용 통제대책
3. 인터넷 PC의 악성코드 감염에 따른 내부망으로의 피해확산 차단대책
4. 사이버공격 탐지·대응 등 안전한 업무환경을 위한 보호대책

⑥ 각급기관의 장은 내부망과 기관 인터넷망의 IP주소 현황을 정기적으로 확인하고 갱신하여야 한다.

⑦ 본 조에 따른 보안대책은 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제17조에 따른 국민제공 공공데이터 범위 산정에는 영향을 미치지 아니하며, 국민에게 제공하는 공공데이터의 범위를 축소하는 것으로 해석하여서는 아니된다. <신설 2021.11.1.>

제41조(클라우드컴퓨팅 보안) ① 각급기관의 장은 클라우드컴퓨팅(공공 클라우드센터를 포함)을 자체 구축·운영하고자 할 경우, 국가정보원장이 배포한 「국가 클라우드 컴퓨팅 보안 가이드라인」에 명시된 기관 자체 클라우드컴퓨팅 구축 보안기준에 따라 보안대책을 수립·시행하여야 한다. <개정 2023.1.31.>

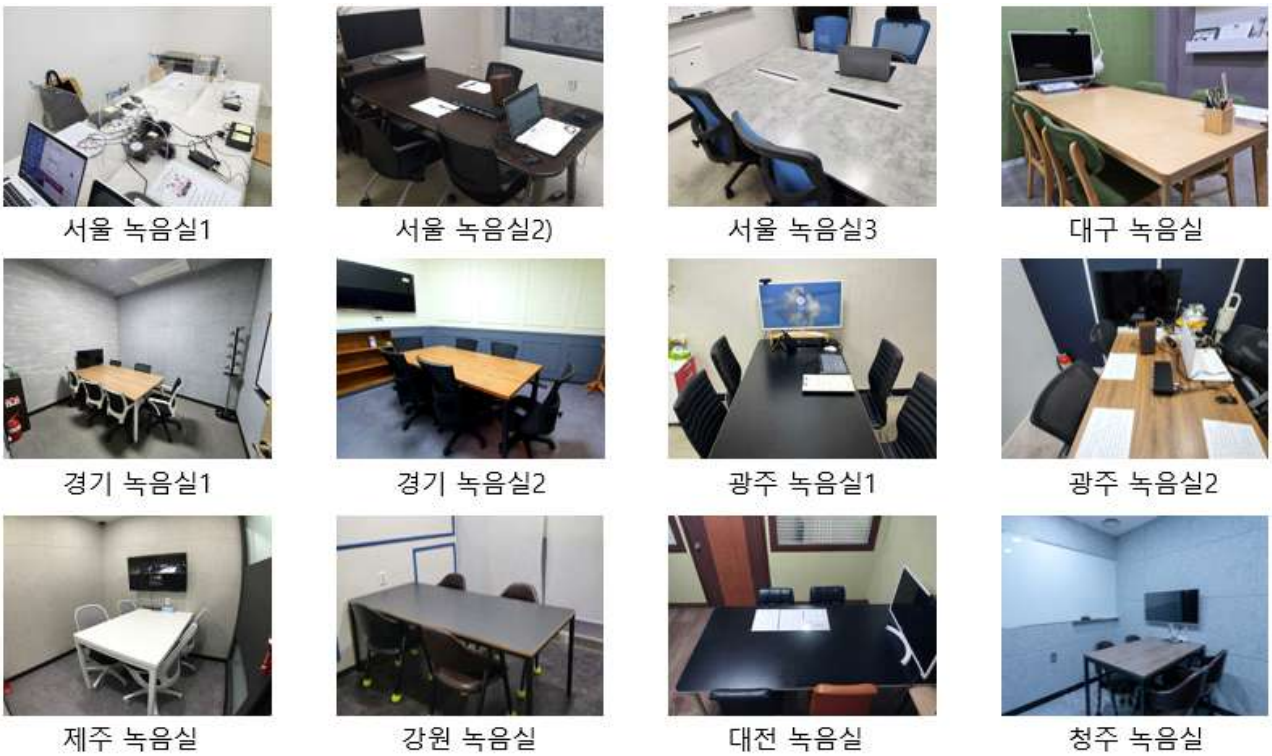
② 각급기관의 장은 민간 클라우드컴퓨팅서비스를 이용하고자 할 경우 다음

[그림 11] 보안 교육 자료 일부

5. 음성 녹음

5.1. 녹음 환경

녹음은 전국 7개 지역(서울, 대전, 대구, 부산, 광주, 강원, 제주)에서 동시에 진행되었다. 외부와 차단된 상태로 대화에 참여한 두 명 이상의 화자만이 대화할 수 있도록 구성하였다. 녹음에 임하는 화자가 편안하게 이야기할 수 있는 사무실 환경을 마련하여 상대방의 목소리가 들어가지 않도록 화자 간 거리가 1m 이상 떨어진 공간에서 녹음을 진행하였다. 장소가 여의찮으면 녹음 기간에 유료 회의실이나 별도 공간을 대여해서 녹음을 진행하였다.



[그림 12] 지역별 녹음실

특히, 코로나19 집단 감염 방지를 위해 녹음 시 화자의 체온을 측정하고 호흡기 증상을 확인하였다. 손 소독제를 녹음실에 비치하였으며, 화자는 마스크를 쓰고 녹음하거나 아크릴판으로 구분된 좌석에서 녹음하였다. 또한, 참여자별 녹음 시간을 조정하여 화자 집단별로 마주치지 않도록 하고, 모집과 교육은 최대한 비대면으로 진행하여 감염 방지에 유의하였다.

[그림 13] 녹음실 환경

<표 14> 코로나19 집단 감염 방지 화자 관리 방안

구분	내용
대책 분야	• 화자 모집, 음성 녹음, 참여자 교육, 회의 등 사업 전반
화자 모집	• 전화, 인터넷 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 일회용 덮개 등 방역 관련 물품 사용 • 사업장 전체 환경 주기적 소독, 환기 실시, 감염 관리 전담 직원 지정
인력 관리	• 방문자 및 종사자 목록 관리 • 유증상자 출근, 이용 중단 및 업무 배제

녹음은 다채널로 진행되었으며, 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 도중 발생하는 돌발적인 소음 등은 완벽하게 통제하기 힘들었다.

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

트랜스미터


헤드셋 마이크


헤드셋 마이크와 트랜스미터

[그림 14] 마이크 장비

5.2. 음성 녹음

5.2.1. 녹음 절차

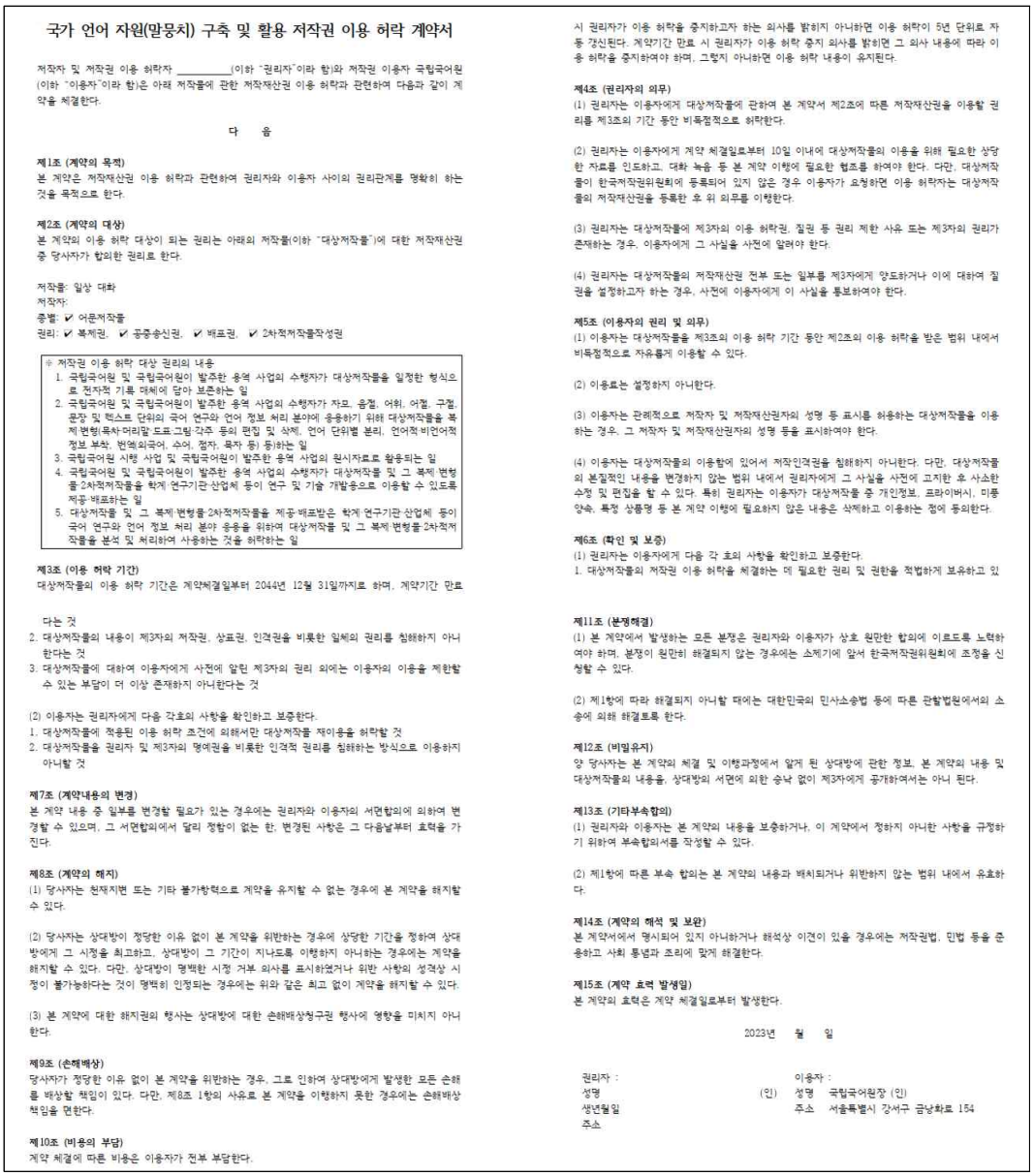
예정된 시간에 화자가 녹음 장소에 도착하면 진행 요원은 화자 모두의 개인정보를 확인하고 녹음이 진행될 음성 자료에 대한 저작권 이용 허락 계약서를 작성하도록 하였다. 그리고 녹음 시 유의 사항에 대해 충분히 전달한 다음 한두 차례 시험 녹음 실시 후 본 녹음을 진행하였다. 녹음은 2인 대화 시 화자당 최대 녹음 2회, 3인 이상 대화 시 화자당 최대 녹음 4회로 제한하며, 동일 화자가 중복참여하지 않도록 관리하였다. 녹음은 아래와 같은 절차로 진행하였다.

녹음 목적 설명 및 저작권 이용 허락 계약 체결	• 화자에게 녹음의 목적과 방법 설명 • 녹음된 음성 자료에 대한 저작권 이용 허락 계약 체결 • 개인정보 수집 이용 동의 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 (마이크 볼륨 및 녹음 순서, 마이크와 입과의 거리 등 조절)
녹음	• 녹음 시 분명하고 큰 목소리와 명확한 발성으로 대화하도록 하고 손이나 머리카락, 옷깃 등에 마이크가 닿지 않도록 함. • 사전 시험 녹음 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 15] 녹음 절차

5.2.2. 저작권 이용 허락 계약 체결 및 수집 일지 작성

녹음 전 참여자에게 이 사업의 목적과 개인정보 보호 준수에 대해 충분히 설명하고 저작권 이용 허락 계약을 체결하였다. 말뭉치 구축 및 활용 저작권 이용 허락 계약서는 결과물인 음성 파일, 전사 파일의 복제권, 전송권, 배포권, 2차적 저작물 작성권에 대해 국립국어원에서 활용하는 것을 허락한다는 내용으로 참여자 전원 작성을 원칙으로 하였다. 참여자가 녹음실을 방문하여 녹음 전 사전 안내에 따라 작성하기 때문에 대부분이 저작권 이용 허락 계약서에 동의는 했으나, 간혹 일부는 계약서 내용을 확인한 후 녹음을 거절하고 돌아간 예도 있었다.



국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

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

다 음

제1조 (계약의 목적)
본 계약은 저작재산권 이용 허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

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

저작물: 일상 대화
저작자:
종별: ✔ 어문저작물
권리: ✔ 복제권, ✔ 공중송신권, ✔ 배포권, ✔ 2차적저작물작성권

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

제3조 (이용 허락 기간)
대상저작물의 이용 허락 기간은 계약체결일부터 2044년 12월 31일까지로 하며, 계약기간 만료

다는 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 권리자 및 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)
본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.

제8조 (계약의 해지)
(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

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

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

제9조 (손해배상)
당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상 책임을 면한다.

제10조 (비용의 부담)
계약 체결에 따른 비용은 이용자가 전부 부담한다.

시 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 계약기간 만료 시 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)
(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을 이용할 권리를 제3조의 기간 동안 비독점적으로 허락한다.

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

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

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

제5조 (이용자의 권리 및 의무)
(1) 이용자는 대상저작물을 제3조의 이용 허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다.

(2) 이용료는 설정하지 아니한다.

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

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

제6조 (확인 및 보증)
(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.
1. 대상저작물의 저작권 이용 허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있

제11조 (분쟁해결)
(1) 본 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

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

제12조 (비밀유지)
양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용 및 대상저작물의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

제13조 (기타부속합의)
(1) 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.

2023년 월 일

권리자 :
성명 (인)
생년월일
주소

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

[그림 16] 저작권 이용 허락 계약서

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

녹음 날짜	선택시간	대화 제목	성명	성별	연령	연락	직업	출생지	주성장지	현거주지	학력	그룹 친밀도	화자 간 관계	1번 주제 - 세부주제		2번 주제 - 세부주제	
2023-09-01	13:00:00	2인 일상 대화	[redacted]	여성	20대	[redacted]	사무 종사자	부산	부산	서울	대졸	5	연인	H	자유 여행	C	식물 관련 경
2023-09-01	13:00:00			남성	20대		사무 종사자	서울	서울	서울	대졸						
2023-09-04	17:00:00	2인 일상 대화		여성	10대		학생	인천	경기	경기	대재	5	연인	B	게임	H	방문 희망 지역
2023-09-04	17:00:00			남성	20대		학생	경기	경기	경기	고졸						
2023-09-06	14:00:00	2인 일상 대화		여성	30대		주부	경남	경남	경기	대졸	4	친구	C	식물 관련 경	D	중고 거래
2023-09-06	14:00:00			여성	30대		서비스 종사자	부산	서울	서울	대졸						
2023-09-06	16:00:00	2인 일상 대화		여성	30대		문가 및 관련 종사	서울	서울	경기	대졸	5	친구	B	수영	C	식물 관련 경
2023-09-06	16:00:00			여성	30대		무직/취업준비생	경기	인천	인천	대졸						

[그림 17] 음성 자료 수집 일지

5.2.3. 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.

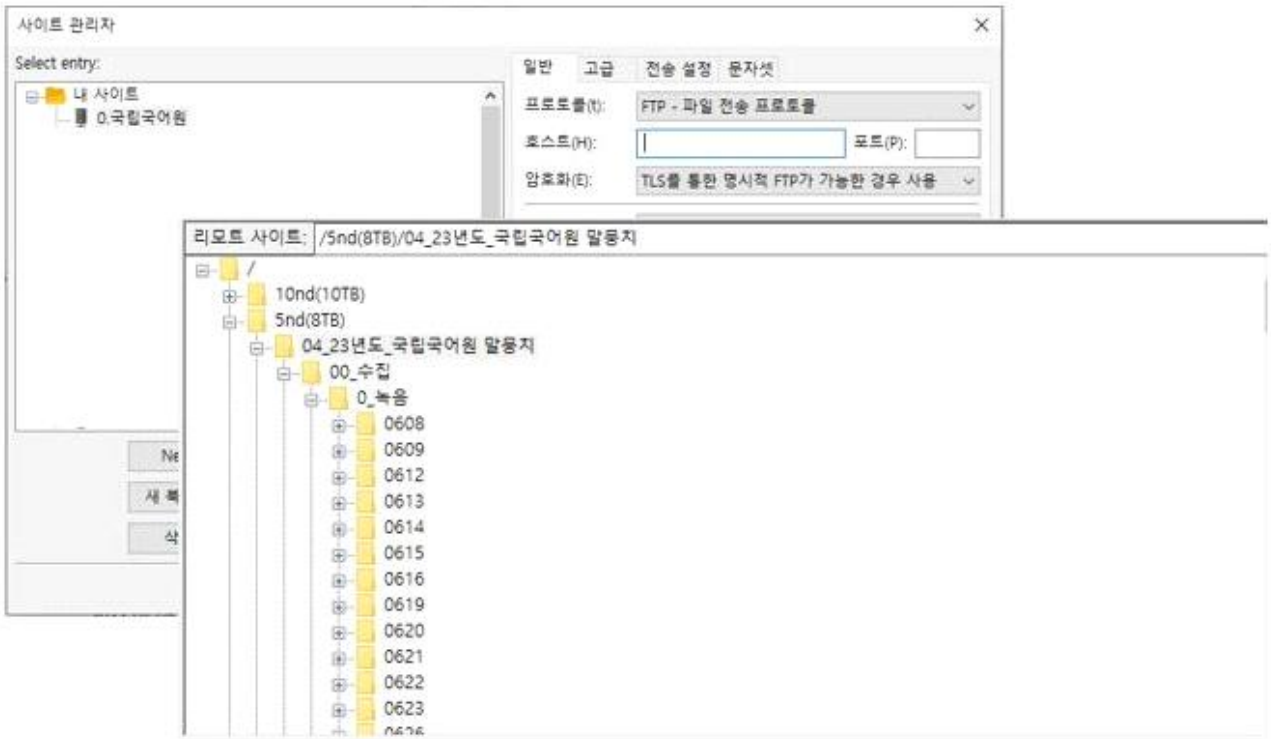
본 녹음 시 분명하고 큰 목소리로, 사전 시험 녹음 때와 비슷한 목소리 크기로 대화하도록 했다. 또한 손이나 머리카락, 옷깃 등에 마이크가 닿지 않도록 했으며, 녹음 시 주의 사항을 다시 한번 전달하고 본 녹음을 시작했다.

대부분 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 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 18] 녹음 진행 순서

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



[그림 19] 공유 시스템 로그인 및 파일 등록 예시

6. 음성 자료 전사

6.1. 전사 규칙

발화 내용은 기본적으로 한글 맞춤법에 따라 전사하는 것을 원칙으로 하였다. 발화 내용은 발음 전사와 철자 전사를 병행하여 진행하였고, 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 경우에는 전사에서 차이가 나도록 구축했다.

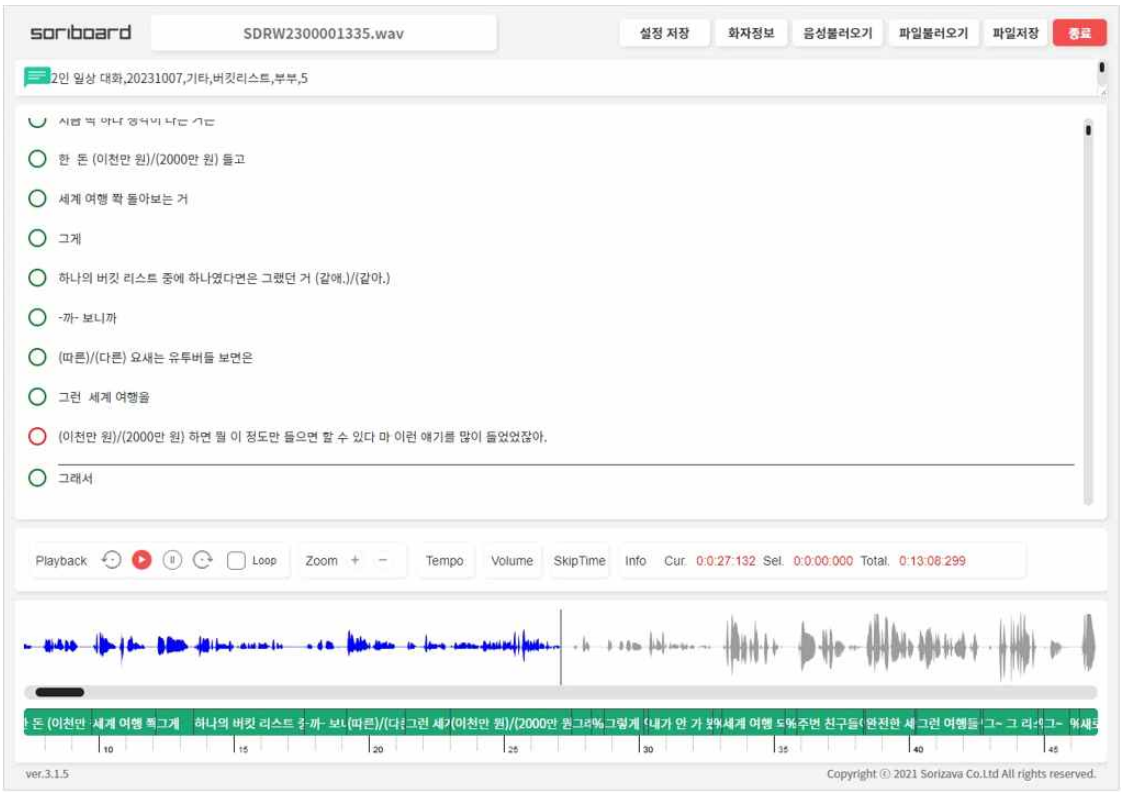
<표 15> 전사 규칙 예시

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 6초 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 발음 전사와 철자 전사를 병행하여 진행하고, 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 전사에서 차이가 나도록 적는다. (예) • 철자 전사: 그렇게 해도 되더라고요. • 발음 전사: 그렇게 해도 되더라구요. • 철자 전사: 이렇게 적금을 3년씩 넣었더니 • 발음 전사: 이케 적금을 삼 년씩 넣었더니

지침 항목	예시
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
축약형 표기	축약형은 모두 표기에 반영한다. 모음의 축약형을 전사할 때는 ’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, 전화번호 등 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. (예) • @이름1 씨는 여행을 할 때… • @상호명1에서 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

6.2. 전사 절차

전사 작업자는 녹음이 완료된 음성 파일을 내려받아 전사 도구를 사용해 전사 규칙에 따라 작업을 진행하였다. 전사 단위로는 맞춤법, 띄어쓰기에 따라 발화 내용을 전사하는 방식으로 진행하였다.



[그림 20] 전사 도구

[그림 21] 전사 절차

6.3. 전사 작업

전문 속기사에 의한 전사는 15분 음성 파일 기준으로 평균 약 1.5시간이 소요되어, 1인당 하루 평균 4개의 파일을 전사하였다. 전사 작업은 하루 평균 8명의 전사 인력이 투입되어 진행되었지만, 나이가 많은 화자가 참여한 경우와 다량의 구어체가 포함되어 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 주관기관이 제시하는 전사 지침을 준수하였다. 전사자는 전사 단위 구분을 가장 어려워하였으며, 그다음으로 이중 전사 및 담화 표지 전사를 어려워하였다.

전사 단위는 억양구 단위로 구분하였는데, 음성을 듣고 억양구 단위인지 판단하는 과정에서 속기사의 주관적 판단이 개입되기 때문에 명확한 구분이 쉽지 않았다. 이런 경우 긴 휴지가 발생하는 부분에서 전사 단위를 구분하여 처리하였다.

이중 전사는 여러 경우의 수가 존재하여 혼란스러울 수 있으므로 국립국어원의 『우리말샘』을 기준으로 지침을 정하여 처리하였다.

관리자는 전사 결과물에 잘못된 부분이 없는지, 전사 규칙은 제대로 준수하였는지 등을 확인한 다음 잘못된 부분이 있다면 수정을 요청하였다. 수정하여 최종 완료된 전사 파일은 관리자가 취합하였다.

전사 지침에 따라 주로 작업한 내용은 다음과 같다.

<표 16> 전사 지침 및 작업 내용

발화자 표시 방법	• 첫 번째 발화자를 1로 표시하고, 1번 발화자 목소리가 2번으로 기재되지 않도록 구분하여 전사 • 화자의 성별, 나이, 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 • 선택 의문문은 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 발음 전사와 철자 전사를 병행하여 진행하고, 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 전사에서 차이가 나도록 전사함(예: (발음 전사)/(철자 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, 전화번호 등 개인정보와 관련된 사항은 비식별화함(예: @이름, @상호명, @주민번호, @카드, @전화, @주소). • 상호명은 부정적인 경우에만 비식별화함. • 여러 이름이 나올 때 일련번호를 붙여 구별함(예: @이름1, @이름2, …).
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.

No.	Data ID	Class	검사 내용				검사 결과	검사 날짜 (2023.XX.XX)	검사자	오류 내용	보완완료 여부 (O, X, -)
			검사요소 1	검사요소 2	검사요소 3	검사요소 4					
1	SDRW2300000001		PASS	PASS	PASS	PASS	PASS	2023.07.03	[redacted]		-
2	SDRW2300000002		PASS	FAIL	PASS	PASS	FAIL	2023.07.03		한글 맞춤법 오류	O
3	SDRW2300000003		PASS	FAIL	PASS	PASS	FAIL	2023.07.03		한글 맞춤법 오류	O
4	SDRW2300000004		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
5	SDRW2300000005		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
6	SDRW2300000006		PASS	FAIL	PASS	PASS	FAIL	2023.07.03		한글 맞춤법 오류	O
7	SDRW2300000007		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
8	SDRW2300000008		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
9	SDRW2300000009		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
10	SDRW2300000010		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
11	SDRW2300000011		PASS	PASS	PASS	PASS	PASS	2023.07.03			-
12	SDRW2300000012		PASS	FAIL	PASS	PASS					
13	SDRW2300000013		PASS	PASS	PASS	PASS					
14	SDRW2300000014		PASS	PASS	PASS	PASS					
15	SDRW2300000015		PASS	PASS	PASS	PASS					
16	SDRW2300000016		PASS	PASS	PASS	PASS					
17	SDRW2300000017		FAIL	PASS	PASS	PASS					
18	SDRW2300000018		PASS	PASS	PASS	PASS					
19	SDRW2300000019		PASS	PASS	PASS	PASS					
20	SDRW2300000020		PASS	PASS	PASS	PASS					
21	SDRW2300000021		PASS	PASS	PASS	PASS					
22	SDRW2300000022		PASS	PASS	PASS	PASS					
23	SDRW2300000023		PASS	FAIL	PASS	PASS					
24	SDRW2300000024		PASS	PASS	PASS	PASS					
25	SDRW2300000025		PASS	PASS	PASS	PASS					
26	SDRW2300000026		PASS	PASS	PASS	PASS					
27	SDRW2300000027		PASS	PASS	PASS	PASS					
28	SDRW2300000028		PASS	PASS	PASS	PASS					

오류 종류	오류 예시	피드백 예시
1. 이중전사 누락	단계인 것 같애요.	-> (같애요.)/(같아요.)
2. 이중전사 내용 오류	거의 (백 프로)/(100퍼센트) 신뢰하고요.	-> (백 프로)/(100프로) 이중전사 안의 발화 내용이 일치해야 함.
3. 띄어쓰기 누락	여러가지 해먹을	-> 여러 가지 해 먹을
4. 불필요한 문자	기분 좋게 먹고 0 (삼십 프로)/(30%) 정도	-> 0이 불필요하게 늘림. -> %는 사용하지 않는 문자임.
5. 담화표지 물결표 오류	오~ 네~	오,네는 담화표지 하지 않습니다.
6. 화자 지정 오류	화자 표시 1번을 2번으로 잘못 체크	-> 싱크작업시 화자 설정 정확하게 해주세요.

[그림 22] 자체 품질 검사 피드백 예시

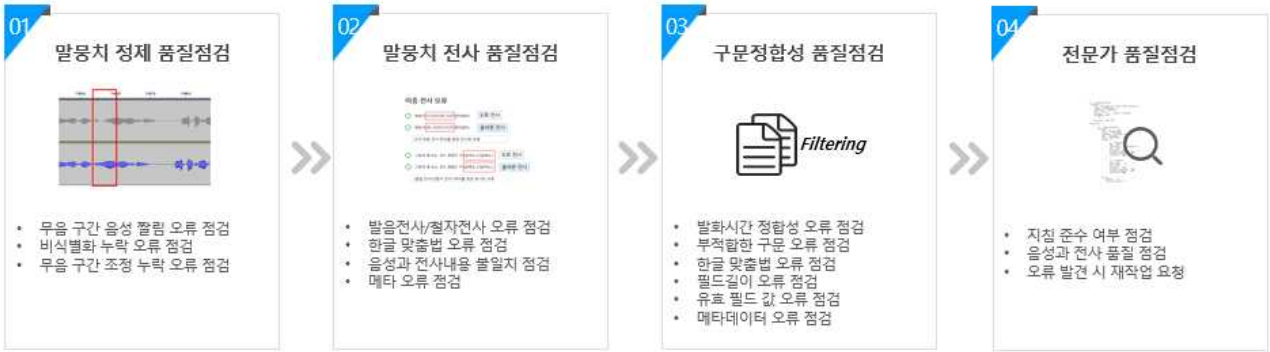
6.4. 품질 검수

전사가 완료된 파일은 검수 담당자가 품질을 점검하였다. 자체 품질 점검을 통해 전체 파일을 검수하고 오류 여부를 확인하여 수정을 진행하였다. 품질 점검 시 음성 파일 품질, 음성과 전사의 일치 여부, 전사 지침 사항 준수 여부, 오류 수정 여부 등을 점검하였고 오류가 발생한 파일의 경우에는 전사 작업자에게 재작업을 요청하였다.

<표 17> 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 • 문서 식별 표지(ID) 및 각종 필드 길이의 정합성 오류 점검 • 날짜 필드의 형식 오류 점검 • 원시 말뭉치 필수 필드 값 누락 오류 점검 • 발화 시간에서 시작 시각과 종료 시간의 정합성 오류 점검 • 철자 전사 내 담화 표지 등 부적합한 구문 오류 점검 • 메타 정보와 화자 식별 표지(ID) 불일치에 대한 데이터 오류 점검
오류 수정 보완	• 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. 최종 품질 검수 진행 시 자주 발생하는 오류 등 중점적으로 봐야 할 사항들에 대한 검수자 교육 진행하고 검수자별로 작업 분량을 배분하였다. 배분된 첫 번째 파일을 사전 분석하여 수정할 사항을 엑셀에 목록화하고 품질관리자의 검토를 통해 내용을 점검받은 후 음성 파일과 전사 파일을 동시에 검토하여 음성과 전사 품질을 확인하였다. 도출된 오류 유형들을 정리해서 작업자들에게 전달하여 데이터 품질을 관리하였다.



[그림 23] 4단계 품질 점검 단계

* 검수 이력 작성을 위한 파일 입니다.
- 주요 검사요소를 데이터에 맞게 작성

전체	1142	무음구간음성잘림	1.14
FAIL	19	비식별화누락오류	0
PASS	1123	무음구간조정누락	1.31
정확도(%)	98.34%	전체오류발생비율	1.66

No.	Data ID	Class	검사 내용			검사 결과	검사 날짜 (2023.XX.XX)	검사자	오류 내용	보완완료 여부 (O, X, -)
			검사요소 1	검사요소 2	검사요소 3					
1	SDRW2300000001		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
2	SDRW2300000002		PASS	PASS	FAIL	FAIL	2023.06.15	이정아	무음 구간 조정 누락	O
3	SDRW2300000003		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
4	SDRW2300000004		FAIL	PASS	PASS	FAIL	2023.06.15	이정아	무음 구간 추가 시 음성 잘림 오류	-
5	SDRW2300000005		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
6	SDRW2300000006		PASS	PASS	FAIL	FAIL	2023.06.15	이정아	무음 구간 조정 누락	O
7	SDRW2300000007		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
8	SDRW2300000008		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
9	SDRW2300000009		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
10	SDRW2300000010		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
11	SDRW2300000011		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
12	SDRW2300000012		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
13	SDRW2300000013		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
14	SDRW2300000014		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
15	SDRW2300000015		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
16	SDRW2300000016		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
17	SDRW2300000017		PASS	PASS	PASS	PASS	2023.06.15	이정아		-
18	SDRW2300000018		PASS	PASS	FAIL	FAIL	2023.06.15	이정아	무음 구간 조정 누락	O

* 검수 이력 작성을 위한 파일 입니다.
- 주요 검사요소를 데이터에 맞게 작성

전체	1142	이중전사 오류	1.4
FAIL	92	한글맞춤법오류	6.83
PASS	1050	음성전사불일치	1.14
정확도(%)	91.94%	메타정보오류	0

No.	Data ID	Class	검사 내용				검사 결과	검사 날짜 (2023.XX.XX)	검사자	오류 내용	보완완료 여부 (O, X, -)
			검사요소 1	검사요소 2	검사요소 3	검사요소 4					
1	SDRW2300000001		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
2	SDRW2300000002		PASS	FAIL	PASS	PASS	FAIL	2023.07.03	송혜주	한글 맞춤법 오류	O
3	SDRW2300000003		PASS	FAIL	PASS	PASS	FAIL	2023.07.03	송혜주	한글 맞춤법 오류	O
4	SDRW2300000004		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
5	SDRW2300000005		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
6	SDRW2300000006		PASS	FAIL	PASS	PASS	FAIL	2023.07.03	송혜주	한글 맞춤법 오류	O
7	SDRW2300000007		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
8	SDRW2300000008		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
9	SDRW2300000009		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
10	SDRW2300000010		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
11	SDRW2300000011		PASS	PASS	PASS	PASS	PASS	2023.07.03	송혜주		-
12	SDRW2300000012		PASS	FAIL	PASS	PASS	FAIL	2023.07.03	송혜주	한글 맞춤법 오류	O

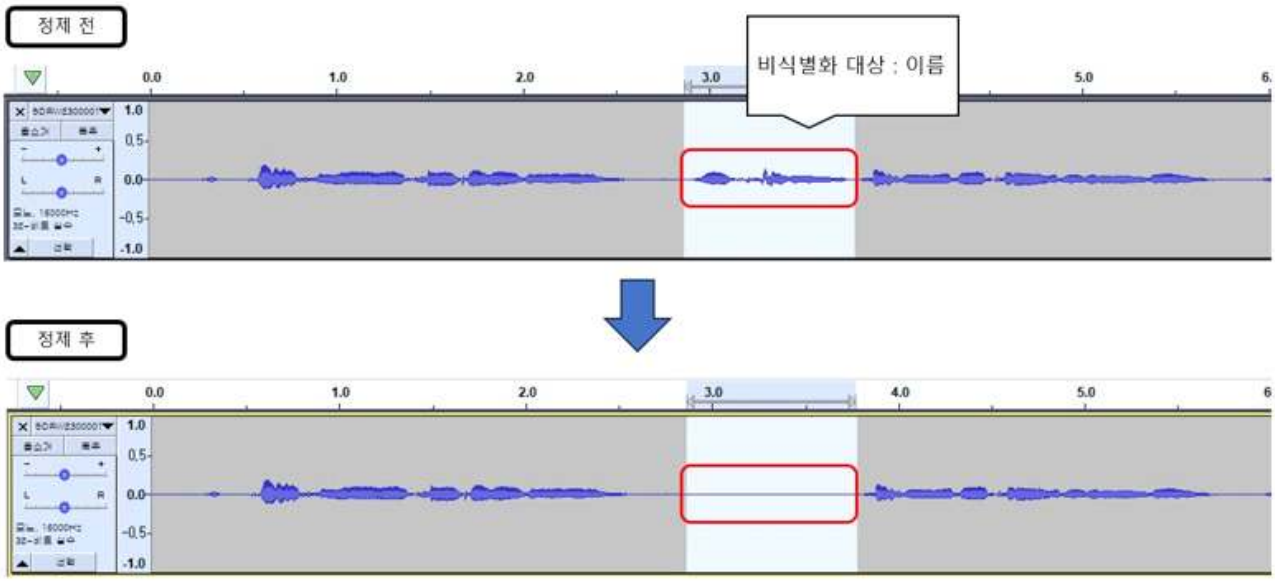
[그림 24] 품질 검사 결과 예시

7. 음성 정제

음성 정제는 음성을 전사 단위에 따라 나누는 작업이다. 관리자 공유 시스템에서 음성 파일과 전사 파일을 내려받아 분할 작업 후 16kHz 표본화, 16bit 양자화 선형 피시엠(PCM: 펄스 코드 변조)으로 저장하는 순으로 진행하였다.

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

음성 정제 시 부정적 맥락에서 사용된 상호, 이름, 주소, 주민등록번호, 카드 번호, 전화번호 등의 개인정보는 익명성 보장을 위해 묵음으로 비식별 처리하였다.



[그림 25] 음성 정제 사진

8. 원시 말뭉치 구축 및 메타 정보 구축

8.1. JSON 변환

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

이 사업 결과물은 「공공데이터의 제공 및 이용 활성화에 관한 법률」 등에 따라 공공데이터의 형태로 제공되어야 함을 고려하여 구축하였으며, 구축 시에는 「공공기관의 데이터베이스 표준화 지침」(행정안전부 고시 제2021-32호, 2021.6.7.), 「공공데이터 관리지침」(행정안전부 고시 제2021-70호, 2021.10.26.),「공공데이터 품질관리 지침 v2.0」(2018.01.) 등 데이터베이스 구축 관련 규정을 준수하였다.

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

<표 18> 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	23	########

준음성과 기타 소리, 개인정보는 전사 편의를 위해 '@웃음', '@이름'의 형태로 전사하였으나, JSON 변환 시 지침에 맞게 마크업 하였다.

<표 19> 전사 기호의 마크업 변환

분류		전사	마크업
준음성	웃음	@웃음	{laughing}
	목청 가다듬는 소리	@목청	{clearing}
	박수	@박수	{applauding}
	노래	@노래	{singing}
비식별화	이름	@이름	&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 때 전사팀에 전달하여 확인 후 수정하였다.

에러발생 파일	에러발생 위치	에러 메시지	에러 발생 값
SDRW2300000001.json	id	필드 길이가 올바르지 않습니다.	SDRW230000001
SDRW2300000001.json	metadata > distributor	유효하지 않은 필드값 입니다.	국립국어원
SDRW2300000001.json	document > 0번째 항목 > id	필드 길이가 올바르지 않습니다.	SDRW23000000001.1
SDRW2300000001.json	document > 0번째 항목 > metadata > author	유효하지 않은 필드값 입니다.	개인 발회자
SDRW2300000001.json	document > 0번째 항목 > metadata > speaker > 0번째 항목 > current_residence	유효하지 않은 필드값 입니다.	ㄱ경기
SDRW2300000001.json	document > 0번째 항목 > metadata > speaker > 1번째 항목	필수인 필드가 누락되었습니다.	age
SDRW2300000001.json	document > 0번째 항목 > utterance > 8번째 항목	필수인 필드가 누락되었습니다.	start
SDRW2300000001.json	document > 0번째 항목 > utterance > 12번째 항목 > speaker_id	필드 길이가 올바르지 않습니다.	SD230022
SDRW2300000001.json	document > 0번째 항목 > utterance > 18번째 항목	필수인 필드가 누락되었습니다.	form
SDRW2300000005.json	JSON 포맷 에러	JSON 형식이 아닙니다.	

[그림 26] 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 20> 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: 2023
	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: 대주제 > 세부주제
		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

```
{
    "id": "SDRW2300000001",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2300000001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2023",
        "category": "구어 > 사적대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2300000001.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20230608",
                "topic": "경제/재테크 > 창업",
                "speaker": [
                    {
                        "id": "SD2300009",
                        "age": "30대",
                        "occupation": "무직/취업준비생",
                        "sex": "여성",
                        "birthplace": "전남",
                        "principal_residence": "전남",
                        "current_residence": "서울",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2300011",
                        "age": "40대",
                        "occupation": "기타",
                        "sex": "여성",
                        "birthplace": "경북",
                        "principal_residence": "경북",
                        "current_residence": "서울",
                        "education": "대졸"
                    }
                ],
                "setting": {
                    "relation": "기타",
                    "contact_frequency": "0"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2300000001.1.1.1",
                    "form": "어 창업에 대해서",
                    "original_form": "어~ 창업에 대해서",
                    "speaker_id": "SD2300009",
                    "start": 0.14006,
                    "end": 2.04506,
                    "note": ""
                },
                {
                    "id": "SDRW2300000001.1.1.2",
                    "form": "좀 준비해 볼까 하는데",
                    "original_form": "좀 준비해 볼까 하는데",
                    "speaker_id": "SD2300009",
                    "start": 2.40009,
                    "end": 4.74532,
                    "note": ""
                },
```


[그림 27] 말뭉치 변환 예시

8.2. 메타 정보 구축

녹음 날짜, 대화 아이디, 대화 주제, 화자 아이디, 화자 정보(성별, 나이대, 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.

2023 일상 대화 말뭉치 구축 Meta Document										
번호	대화ID	타이틀	주제1	주제2	관계	친밀도	파일 시간	저작권 확보 여부	민감이슈여부	분류항목
1	SDRW2300000001	2인 일상 대화	경제/재테크	창업	기타	0	0:15:44	O	X	
2	SDRW2300000002	2인 일상 대화	방송/영화/연예인	영화	기타	0	0:14:29	O	X	
3	SDRW2300000003	4인 일상 대화	취직	직업	직장 동료	2	0:15:53	O	X	
4	SDRW2300000004	4인 일상 대화	인간관계	MBTI	직장 동료	2	0:15:08	O	X	
5	SDRW2300000005	4인 일상 대화	경제/재테크	개인의 소비 활동	직장 동료	2	0:14:54	O	X	
6	SDRW2300000006	4인 일상 대화	사회 이슈	소셜미디어	직장 동료	2	0:16:33	O	X	
7	SDRW2300000007	4인 일상 대화	먹거리	음식	직장 동료	2	0:13:39	O	X	
8	SDRW2300000008	4인 일상 대화	취직	직업	직장 동료	2	0:14:57	O	X	
9	SDRW2300000009	4인 일상 대화	사회 이슈	인구 감소	직장 동료	2	0:14:54	O	X	
10	SDRW2300000010	4인 일상 대화	기타	추억	직장 동료	2	0:14:15	O	X	
11	SDRW2300000011	2인 일상 대화	반려동식물	반려동식물 관련 경험 및 팁	기타	0	0:17:09	O	X	
12	SDRW2300000012	2인 일상 대화	먹거리	음식	기타	0	0:17:47	O	X	
13	SDRW2300000013	3인 일상 대화	취미	추천 음악	직장 동료	1	0:16:25	O	X	
14	SDRW2300000014	3인 일상 대화	반려동식물	반려동식물 관련 경험 및 팁	직장 동료	1	0:17:51	O	X	

[그림 28] 메타 정보 파일 일부

2023 일상 대화 말뭉치 구축 발화자 정보										
번호	파일명	화자 ID	이름	연령대	직업	성별	출생지	주 성장지	현 거주지	최종학력
1	SDRW2300000001	SD2300009	오O지	30대	무직/취업준비생	여성	전남	전남	서울	대졸
2	SDRW2300000001	SD2300011	김O은	40대	기타	여성	경북	경북	서울	대졸
3	SDRW2300000002	SD2300009	오O지	30대	무직/취업준비생	여성	전남	전남	서울	대졸
4	SDRW2300000002	SD2300011	김O은	40대	기타	여성	경북	경북	서울	대졸
5	SDRW2300000003	SD2300001	김O현	20대	사무 종사자	여성	인천	인천	인천	대졸
6	SDRW2300000003	SD2300002	김O영	30대	서비스 종사자	남성	경북	경북	서울	대졸
7	SDRW2300000003	SD2300003	최O연	30대	사무 종사자	여성	서울	서울	서울	대졸
8	SDRW2300000003	SD2300004	김O훈	30대	사무 종사자	남성	대전	대전	서울	고졸
9	SDRW2300000004	SD2300001	김O현	20대	사무 종사자	여성	인천	인천	인천	대졸
10	SDRW2300000004	SD2300002	김O영	30대	서비스 종사자	남성	경북	경북	서울	대졸
11	SDRW2300000004	SD2300003	최O연	30대	사무 종사자	여성	서울	서울	서울	대졸
12	SDRW2300000004	SD2300004	김O훈	30대	사무 종사자	남성	대전	대전	서울	고졸
13	SDRW2300000005	SD2300001	김O현	20대	사무 종사자	여성	인천	인천	인천	대졸
14	SDRW2300000005	SD2300002	김O영	30대	서비스 종사자	남성	경북	경북	서울	대졸

[그림 29] 발화자 정보 파일 일부

제 3 장

사업 수행 결과

1. 주제별·제시 자료별 수집 결과

일상 대화 말뭉치는 주관 기관과 협의하여 16개 주제와 관련된 세부 주제들을 선정하였고, 특정 주제나 자료에 편중되지 않도록 수집하였다.

<표 21> 주제별 수집 결과

번호	대주제	세부 예시 주제	수집 쌍	비율
1	방송/영화/연예인	텔레비전(TV) 프로그램(드라마, 예능 등), 영화, 연예인	121	6.13%
2	취미	음악 이론, 음악 활동(보컬, 악기 연주 등), 추천 음악 등	139	7.05%
3	반려 동식물	반려 동식물 관련 경험 및 팁, 반려 동식물 추천 등	110	5.58%
4	쇼핑	선호 쇼핑몰, 선호 브랜드, 쇼핑 방식, 중고 거래 등	117	5.93%
5	패션/미용	패션, 스타일, 메이크업, 얼굴, 몸매, 화장, 피부 관리, 헤어 스타일링 등	107	5.42%
6	먹거리	음식, 요리, 좋아하는 요리, 요리법, 식재료 쇼핑, 조리 도구 등	153	7.75%
7	건강/다이어트	건강, 다이어트, 식단, 건강 보조제, 질병 관련 경험과 증상 등	146	7.40%
8	여행/휴가	국내 여행 경험 및 계획, 해외여행 경험 및 계획, 추천 여행 지역 등	147	7.45%
9	생활/주거 환경	가사 활동(빨래, 청소 등), 가사 관련 가전(세탁기, 의류관리기) 등	108	5.47%
10	가족/관혼상제	가족, 결혼, 출산, 성인식, 결혼식, 장례식, 명절, 제사, 돌잔치 등	125	6.34%
11	회사, 학교생활	직장 및 학교생활, 업무 내용, 업무 강도, 야근, 회식, 회의, 승진 등	133	6.74%
12	취직	진로, 직업, 취직, 이직, 취준생, 해외 취업, 일반 자격증, 전문 자격증 등	120	6.08%
13	인간관계	친구, 연애, 학교 동기, 직장 동료, 성격상 장점, 성격상 단점 등	109	5.52%
14	경제/재테크	경제, 재테크, 예금, 적금, 주식, 코인, 부동산, 투자 팁 등	106	5.37%
15	사회 이슈	인공지능(AI) 기술발전, 소셜미디어, 환경 문제, 인구 감소, 고령화 등	100	5.07%
16	기타	군대, 추억, 꿈, 인생 목표, 인생 계획, 가치관 등	132	6.69%
합계			**1,973**	**100%**

2. 화자 모집 결과

2.1. 인구 특성별 수집 결과

사업 초반 화자 모집 계획은 2,100명 이상을 목표로 수립하였고, 결과적으로 2,168명의 화자를 모집하였다.

<표 22> 성×나이×지역별 화자 모집 결과(단위: 명)

구분		10대		20대		30대		40대		50대		60대		합계	
		남	여	남	여	남	여	남	여	남	여	남	여	지역별	권역별
수도권	서울	21	31	63	91	34	43	26	42	9	22	26	27	435	991
	인천	4	13	14	26	8	10	6	5	4	3	6	7	106	
	경기	29	39	74	97	27	52	11	17	4	26	30	44	450	
영남권	부산	7	11	25	41	15	23	6	15	1	7	7	7	165	625
	울산	3	4	8	13	5	5	1	5	1	3	3	3	54	
	대구	5	22	21	42	9	17	7	10	4	8	5	5	155	
	경북	5	8	17	31	9	11	3	11	1	6	5	6	113	
	경남	7	9	20	25	14	22	4	14	1	7	7	8	138	
호남권	광주	2	7	8	16	7	6	2	5	1	4	4	3	65	200
	전북	3	4	11	13	4	4	2	7	2	6	6	7	69	
	전남	2	5	9	15	4	1	1	7	1	6	9	6	66	
충청권	대전	3	8	9	27	3	5	4	6	2	3	4	3	77	233
	충북	2	3	11	17	3	6	3	6	2	3	4	8	68	
	충남	3	4	13	17	4	7	3	7	1	3	14	12	88	
강원권	강원	3	3	9	11	6	6	5	6	2	2	6	10	69	69
제주권	제주	2	2	6	8	7	4	5	4	2	3	3	4	50	50
합계		101	173	318	490	159	222	89	167	38	112	139	160	2,168	2,168

2.2. 주제별 나이대 분포

주제별 화자의 나이대 분포를 보면 주제별로 다양한 연령층이 고르게 대화하였음을 알 수 있다. 그중 10대는 회사/학교생활과 취미, 20대는 회사/학교생활과 방송/영화/연예인을 주제로 대화를 많이 하였고, 30대는 경제/재테크와 쇼핑, 40대는 경제/재테크와 건강/다이어트, 50대는 가족/관혼상제와 건강/다이어트, 60대 이상은 건강/다이어트와 먹거리를 주제로 대화를 많이 하였다.

<표 23> 주제별 나이대 분포(단위: 명)

주제	10대	20대	30대	40대	50대	60대 이상	합계	비율
가족/관혼상제	5	81	56	60	48	60	310	6.25%
건강/다이어트	15	72	61	61	38	107	354	7.13%
경제/재테크	6	83	71	70	23	12	265	5.34%
기타	51	128	55	21	17	53	325	6.55%
먹거리	51	126	43	46	36	84	386	7.78%
반려 동식물	31	103	65	36	17	9	261	5.26%
방송/영화/연예인	54	150	51	33	22	11	321	6.47%
사회 이슈	21	100	50	41	24	23	259	5.22%
생활/주거 환경	18	109	61	52	33	11	284	5.72%
쇼핑	35	130	65	33	19	9	291	5.86%
여행/휴가	30	137	60	47	32	59	365	7.35%
인간관계	54	142	42	17	21	12	288	5.80%
취미	63	128	55	21	18	40	325	6.55%
취직	36	136	44	36	9	47	308	6.21%
패션/미용	49	125	49	17	24	18	282	5.68%
회사/학교 생활	76	156	46	29	17	15	339	6.83%
합계	595	1,906	874	620	398	570	4,963	100%

2.3. 주제별 성별 분포

주제별 성별 분포를 보면 편차 없이 고르게 다양한 주제로 남성과 여성이 대화하였음을 알 수 있다. 그중 남성은 취미를 주제로 대화를 많이 하였고, 여성은 먹거리를 주제로 대화를 많이 하였다.

<표 24> 주제별 성별 분포(단위: 명)

주제	남성	여성	합계	비율
가족/관혼상제	88	222	310	6.25%
건강/다이어트	120	234	354	7.13%
경제/재테크	124	141	265	5.34%
기타	158	167	325	6.55%
먹거리	84	302	386	7.78%
반려 동식물	81	180	261	5.26%
방송/영화/연예인	101	220	321	6.47%
사회 이슈	104	155	259	5.22%
생활/주거 환경	80	204	284	5.72%
쇼핑	102	189	291	5.86%
여행/휴가	118	247	365	7.35%
인간관계	74	214	288	5.80%
취미	166	159	325	6.55%
취직	147	161	308	6.21%
패션/미용	48	234	282	5.68%
회사/학교생활	117	222	339	6.83%
합계	1,712	3,251	4,963	100%

2.4. 화자 간 관계별 수집 결과

2023년 일상 대화 말뭉치 구축 사업에서는 서로 모르는 사람과 대화한 비율이 굉장히 높아서 흥미로운 연구 자료가 될 수 있다. 처음 본 사람과 대화한 비율이 전체의 34.06%를 차지하였으며, 친구와 대화한 비율이 수집 비율의 24.63%를 차지하였다. 그다음으로는 직장 동료(11.71%)가 많았으며, 모임·동아리 지인(10.49%) 순으로 그 뒤를 이었다.

<표 25> 화자 간 관계별 수집 결과

화자 간 관계	수집 쌍	비율
고향 선후배	4	0.2%
교회 지인	1	0.05%
기타	672	34.06%
기타 가족	8	0.41%
대학 선후배	36	1.82%
모임·동아리 지인	207	10.49%
부모/자녀	84	4.26%
부부	52	2.64%
연인	105	5.32%
이웃사촌	31	1.57%
직장 동료	231	11.71%
친구	486	24.63%
형제/자매	56	2.84%
합계	1,973	100%

2.5. 직업별 수집 결과

다양한 직업의 화자 모집을 위해 평일 오전, 오후뿐 아니라 야간과 주말에도 수집을 진행하였다. 수집된 결과를 보면 전체 비율 중에서 학생이 전체의 31.32%를 차지하였고, 그다음으로 주부가 14.81%로 높은 비율을 차지하였다. 그 뒤를 이어 사무 종사자, 무직/취업준비생, 기타 순으로 나타났다.

<표 26> 직업별 수집 결과(단위: 명)

직업	모집 인원	비율
경영/관리직	37	1.71%
기능원 및 관련 기능 종사자	4	0.18%
기술자 종사자 (장치/기계조작 및 조립종사자)	23	1.06%
기타	243	11.21%
농업/임업/어업종사자	6	0.28%
단순노무 종사자	10	0.46%
무직/취업준비생	271	12.50%
사무 종사자	290	13.38%
서비스 종사자	112	5.17%
전문가 및 관련 종사자	122	5.63%
주부	321	14.81%
판매/영업 종사자	50	2.31%
학생	679	31.32%
합계	2,168	100%

2.6. 학력별 수집 결과

화자의 학력 현황을 보면 대졸 이상이 전체 인원의 45.25%를 차지하였고, 그다음으로 대학교 재학(23.57%), 고등학교 졸업(16.33%)이 뒤를 이었다.

<표 27> 학력별 수집 결과(단위: 명)

학력	모집 인원	비율
초졸 이하	81	3.74%
중졸	140	6.46%
고졸	354	16.33%
대재	511	23.57%
대졸	981	45.25%
대학원 이상	101	4.66%
합계	2,168	100%

2.7. 출생지별 수집 결과

화자의 출생지별 수집 결과를 보면 서울이 전체 인원의 25.05%를 차지하였고, 그다음으로 경기(13.88%), 대구(11.90%)가 높은 비율을 차지하였다.

<표 28> 출생지별 수집 결과(단위: 명)

출생지	모집 인원	비율
서울	543	25.05%
인천	61	2.81%
경기	301	13.88%
부산	216	9.96%
울산	18	0.83%
대구	258	11.90%
경북	116	5.35%
경남	78	3.60%
광주	93	4.29%
전북	58	2.68%
전남	72	3.32%
대전	104	4.80%
충북	65	3.00%
충남	68	3.14%
강원	78	3.60%
제주	39	1.80%
합계	2,168	100%

2.8. 주 성장지별 수집 결과

지역별 화자는 주 성장지[1]를 기준으로 선발하였고, 해당 지역의 비율을 고려하여 수집하였다. 화자의 주 성장지별 수집 결과를 보면 경기가 전체의 20.76%로 가장 많고 서울이 20.06%, 부산이 7.61%를 차지하였다.

<표 29> 주 성장지별 수집 결과(단위: 명)

주 성장지	모집 인원	비율
서울	435	20.06%
인천	106	4.89%
경기	450	20.76%
부산	165	7.61%
울산	54	2.49%
대구	155	7.15%
경북	113	5.21%
경남	138	6.37%
광주	65	3.00%
전북	69	3.18%
전남	66	3.04%
대전	77	3.55%
충북	68	3.14%
충남	88	4.06%
강원	69	3.18%
제주	50	2.31%
합계	2,168	100%

1) 주 성장지란 화자가 초, 중, 고등학교를 나온 지역을 의미한다. 주 성장지가 여러 곳일 경우 가장 오래 있었던 지역을 기준으로 한다.

2.9. 현 거주지별 수집 결과

화자의 현 거주지별 수집 결과를 보면 서울이 전체 인원의 31.32%로 가장 많고 경기가 25.46%, 대구가 13.61%, 부산이 9.23%를 차지하였다.

<표 30> 현 거주지별 수집 결과(단위: 명)

현 거주지	모집 인원	비율
서울	679	31.32%
인천	59	2.72%
경기	552	25.46%
부산	200	9.23%
울산	5	0.23%
대구	295	13.61%
경북	15	0.69%
경남	13	0.60%
광주	90	4.15%
전북	4	0.18%
전남	7	0.32%
대전	113	5.21%
충북	37	1.71%
충남	25	1.15%
강원	42	1.94%
제주	32	1.48%
합계	2,168	100%

3. 정책 제언

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

이번 사업에서 가장 어려웠던 점은 중장년층 이상의 남성화자 모집이었다. 홍보와 신청이 인터넷으로 진행되어 인터넷 사용이 취약한 60대 이상의 화자 모집은 사업 내내 수집의 어려움을 겪었다. 이를 해결하기 위해 대한노인회에 협조를 요청하여 화자 모집을 진행하였다. 사업 홍보를 위해 SBS <모닝와이드> 방송을 촬영했으나 출근 시간대인 평일 오전 8시 이후에 방영되어 큰 홍보 효과를 얻지 못했다. 남성 화자 모집의 어려움은 남녀 화자 비율 1:1 수집에 대한 어려움으로 이어졌다. 사업 후반에는 남성 화자 위주로 모집을 진행한 결과 남녀 화자 비율은 4:6의 비율로 마무리되었다.

사업 중반에는 말뭉치 구축 지침에 대한 변경으로 이슈가 있었다. 지침이 사업 중반에 변경되면 변경 전 구축된 말뭉치의 보완을 위해 많은 시간이 소요될 수밖에 없다. 이를 방지하기 위해 구축 지침 변경을 최소화하도록 진행했다. 하지만 이렇게 구축된 말뭉치 품질에는 분명 한계가 있다. 향후 고품질의 말뭉치 구축을 위해서는 작업 초반부터 전사 경험이 풍부한 인력들이 투입되어 발음 전사와 철자 전사를 정확히 이해하고 오류를 최소화할 수 있도록 진행해야 한다. 또한 분리된 마이크를 사용하여 입력되는 녹음에 대한 화자 식별 표지(ID)를 자동적으로 입력할 수 있는 시스템이 구축된다면 원시 말뭉치에서 화자별 채널 분리 오류를 미연에 방지할 수 있을 것이라 본다.

일상 대화 녹음은 단순한 아르바이트가 아니라, 인공지능 시대의 중요한 밑거름이 될 것으로 홍보한다면, 녹음 참여도를 높일 수 있을 것이다. 그리하여 풍부한 일상 대화를 수집해 인공지능 발전에 기여하는 말뭉치를 구축할 수 있을 것이다.

이와 같이 문제점을 보완한다면 향후 일상 대화 말뭉치 구축 사업은 국어 및 국어 문화 연구, 4차 산업 대비 기반 인공지능 기술 발전에 더욱 실효성 있는 사업이 될 것으로 기대한다.

[붙임 1] 2023년도 일상 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	구축 연도	8자리 일련번호
S: 구어 말뭉치	D: 사적 대화	RW: 원시 말뭉치	23	########

- 예시
  · SDRW2300000001.sjml 원시 말뭉치 첫 번째 파일
  ※ 참고: 음성 파일 파일명 부여 방식
  · SDRW2300000001.pcm 음성 원본 첫 번째 파일
  · SDRW2300000001-00001.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: 2023
	category			string	분류:구어>사적대화>일상대화
	annotation_level			array(string)	분석 층위: 원시
	sampling			string	샘플링 방식: 본문 전체
document				array(object)	대화 정보
	id			string	대화 아이디
	metadata			object	대화 메타 정보
		title		string	대화 제목: 2인 일상 대화
		author		string	저작권자: 개인 발화자
		publisher		string	발행자: 개인 발화 녹음
		date		string	녹음일자: YYYYMMDD
		topic		string	대화 주제: 대주제>세부주제
		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
			contact_frequency	string	대화 빈도
	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": "SDRW2300000728",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2300000728",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2023",
        "category": "구어 > 사적대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2300000728.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20230726",
                "topic": "패션/미용 > 스타일",
                "speaker": [
                    {
                        "id": "SD2300672",
                        "age": "20대",
                        "occupation": "무직/취업준비생",
                        "sex": "여성",
                        "birthplace": "충북",
                        "principal_residence": "충북",
                        "current_residence": "대전",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2300671",
                        "age": "20대",
                        "occupation": "기타",
                        "sex": "여성",
                        "birthplace": "충남",
                        "principal_residence": "대전",
                        "current_residence": "대전",
                        "education": "대졸"
                    }
                ],
                "setting": {
                    "relation": "기타",
                    "contact_frequency": "0"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2300000728.1.1.1",
                    "form": "혹시 평소에 선호하시는 패션 스타일이 있으실까요?",
                    "original_form": "혹시 평소에 선호하시는 패션 스타일이 있으실까요?",
                    "speaker_id": "SD2300672",
                    "start": 0.06677,
                    "end": 4.94876,
                    "note": ""
                },
                {
                    "id": "SDRW2300000728.1.1.2",
                    "form": "어 저는 이제",
                    "original_form": "어~ 저는 이제",
                    "speaker_id": "SD2300671",
                    "start": 5.03106,
                    "end": 6.39568,
                    "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 SDRW2300000001)
- 구축자(creator): 국립국어원
- 배포자(distributor): 국립국어원
- 구축년도(year): 2023
- 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: SDRW2300000001.1)

2.2.4. 대화 메타 정보(document > metadata)

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

대화 주제	
1	휴가
2	대중교통
3	음악
4	건강/다이어트
5	방송/연예
6	스포츠/레저/취미
7	먹거리
8	우정
9	경제/재테크
10	회사/학교
11	반려동물
12	취직
13	가족/관혼상제
14	쇼핑
15	생활/주거 환경
16	기타

2.2.5. 화자 정보(document > metadata > speaker)

- 화자 아이디(id): 화자 고유 아이디 부여, 대화가 다르더라도 화자가 같으면, 같은 아이디 부여 단, 화자가 교정기를 착용한 경우에는 구축 연도 다음 숫자 1을 넣어 표시(한 화자가 교정기를 뺐다 넣었다 하지 않도록 함)
  (예: 교정기 미착용 화자 A: SD2300001, 교정기 착용 화자 B: SD2310002)
- 연령(age): 10대/20대/30대/40대/50대/60대 이상
- 직업(occupation): '한국표준직업분류'를 준용한 아래에서 선택

1) 경영/관리직	2) 전문가 및 관련 종사자
3) 사무 종사자	4) 서비스 종사자
5) 판매/영업 종사자	6) 농업/임업/어업 종사자
7) 기능원 및 관련 기능 종사자	8) 기술자 종사자(장치/기계 조작 및 조립 종사자)
9) 단순노무 종사자	10) 군인
11) 학생	12) 주부
13) 무직/취업준비생	14) 기타

- 성별(sex): 남성/여성/NA
- 출생지(birthplace): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주
- 주 성장지(principal_residence): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주
- 현 거주지(current_residence): 서울/광주/대구/대전/부산/울산/인천/강원/경기/경북/경남/전북/전남/충북/충남/제주
- 학력(education): 초졸 이하/중졸/고졸/대재/대졸/대학원 이상

2.2.6. 환경 정보(document > metadata> setting)

- 화자 간 관계(relation): 아래에서 선택

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

- 대화 빈도(contact_frequency): 아래에서 선택

거의 매일	5
주 3회 이상	4
주 1~2회	3
주 1회 미만	2
월 1회 미만	1
처음(낯선 관계)	0

2.2.7. 발화 정보(document > utterance)

- 발화 아이디(id): 대화 아이디 + . + 1 + . + 1 + . + 발화 번호(예: SDRW2300000001.1.1.4)
- 철자 전사(form): 철자 전사 결과
- 발음 전사(original_form): 발음 전사 결과
- 발화 시작 시각(start): 해당 발화의 음성 원본에서의 시작 시각을 초 단위(소수 5자리까지 필수)로 표기(예: 30.56600)
- 발화 종료 시각(end): 해당 발화의 음성 원본에서의 종료 시각을 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

3.1.1 이중 전사를 하는 경우

1) 음성 자료의 전사는 철자 전사를 원칙으로 하되, 일부의 경우 '발음 전사'와 '철자 전사(맞춤법 전사)'를 병행하는 이중 전사를 한다.

- 발음 전사: 발화된 그대로를 한글로 전사하며, 숫자, 영문 등도 발화자의 발음에 따라 한글로 전사한다.
- 철자 전사: 한글 맞춤법과 표준어 규정에 따라 전사하되, 숫자, 영문의 경우는 세부 지침에 따라 이중 전사한다.

*지침 3.3 참조(숫자는 관습에 따라 숫자와 한글을 혼용하며, 외래어는 외래어 표기법의 규정에 따라 전사한다.)

2) 발음 전사의 원칙

- 발음 전사는 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우, 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우에 적용하며, 다음과 같이 발음의 차이가 드러나도록 적는다.

> 발음 전사: 자 상담소에는 어떤 걸 기대하고 왔으까?
> 철자 전사: 자 상담소에는 어떤 걸 기대하고 왔을까?

- 숫자, 외래어, 기호, 단위 등도 발음 전사에서는 한글로 적는다.
- 그 외 표준 발음에 맞게 발음한 경우에는 발음 전사를 하지 않고 한글 맞춤법, 표준어 규정, 외래어 표기법 등 관련 어문 규정에 따라 한글로 적는다.

3) 철자 전사(맞춤법 전사)의 원칙

- 철자 전사는 한글 맞춤법 및 표준어 규정에 따라 적는 것으로, 발화 내용은 기본적으로 한글 맞춤법 및 표준어 규정에 따라 전사한다.
- 띄어쓰기도 한글 맞춤법에 따르며, 띄어쓰기와 붙여쓰기가 모두 허용되는 경우 띄어 쓰는 것을 원칙으로 한다.

3.1.2. 화자 표시

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

3.1.3. 전사 단위

1) 억양구 전사

- 문어의 문장과 달리, 구어는 문장의 경계를 구분할 수 없으므로 억양구 전사를 기본 단위로 한다.
- 이때 억양구의 구획은 음성 분절 및 전사의 기본 단위는 긴 휴지, 경계 억양, 경계말 장음화 등의 음성적인 특성을 우선으로 하지만, 통사적·정보적 완결성도 고려 요인이 된다.

외할머니가 그냥 매달려서 울고불고 소화제라도 지어 달라고 환자가 눈치챈다고 그렇게 와서 돌팔이 의사라는 사람한테 진단받아 갖고 그 사람이 시키는 대로 약 갖다 먹고 <억양구 전사 사례> 외할머니가 그냥 매달려서 울고불고 소화제라도 지어 달라고 환자가 눈치챈다고 그렇게 와서 돌팔이 의사라는 사람한테 진단받아 갖고 그 사람이 시키는 대로 약 갖다 먹고

- 단, 억양구는 통사적인 특성도 어느 정도 고려되므로, 하나의 억양구 내에 완결된 문장이 둘 이상 포함된 경우, 각 문장 경계에서 억양구를 분할한다.

1) 난 치킨 좋아 치킨은 건강 음식이라 생각해. (×)
2) 난 치킨 좋아.
치킨은 건강 음식이라 생각해.(○)

1) 그렇구나 아무래도 코로나 때문에 밖에서 운동하는 일이 적어지다 보니까(×)
2) 그렇구나.
아무래도 코로나 때문에 밖에서 운동하는 일이 적어지다 보니까(○)

1) 근데 쌤 기분 맘대론 거 같다. 내가 보기엔 우리 동아리도(×)
2) 근데 쌤 기분 맘대론 거 같다.
내가 보기엔 우리 동아리도(○)

- 하나의 전사 단위가 가능한 6초 이상으로 길어지지 않도록 한다.[2)]

[주의] 쉼을 기준으로 잘못 나눈 예

오류	수정
할 말이 없더라. 근데 보건 쌤 말씀이	할 말이 없더라. 근데 보건 쌤 말씀이
근데 쌤 기분 맘대론 거 같다. 내가 보기엔 우리 동아리도 쌤 기분 따라 가지고 그냥 흘러가는 거 같아.	근데 쌤 기분 맘대론 거 같다. 내가 보기엔 우리 동아리도 쌤 기분 ~~ 거 같아.
어~ 나 name8이 것도 사 주러 가는데 이렇게 얘기하는 거야. 그래서 name8이가 그런 이미지가 아 아닌데라고	어~ 나 name8이 것도 ~~ 얘기하는 거야. 그래서 name8이가 그런 이미지가 아 아닌데라고
부회장이나 회장 그런 거 안 나가냐. 의대 가려면 그런 거 있으면 좋지 않나?	부회장이나 회장 그런 거 안 나가냐. 의대 가려면 그런 거 있으면 좋지 않나?
나이가 들수록 자식한테 짐도 안 되고 싶고 내가 스스로 살아 있는 동안은 건강한 삶을 누리고 싶어서	나이가 들수록 자식한테 짐도 안 되고 싶고 내가 스스로 살아 있는 동안은 ~~ 누리고 싶어서
벤허라는 걸 봤던 거 같아요. 그게 되게 길었어요. 한 네 시간 정도 더 됐던 거 같은데	벤허라는 걸 봤던 거 같아요. 그게 되게 길었어요. 한 네 시간 정도 더 됐던 거 같은데
그러고 나서 그다음부터 조용해 지긴 했어요.	그러고 나서 그다음부터 조용해지긴 했어요.

2) 자문위원 의견 20초도 있었으나 6초면 충분한 길이라고 사료됨.

2) 긴 쉼에 의해 나뉘는 경우는 통사적으로 완성이 되지 않았다 하더라도 전사 단위를 구분하여 전사한다.

음성1_고등학교 때 제주도 때문에 비행기를 타 봤는데 (*한참 후에 발화) 음성2_타 본 거잖아.

- 긴 쉼의 기준은 최대 1초로 한다. (Chafe 1994 기준 0.8초)

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

나는 뭐든 다 좋을 것 같아 여행을 가면. 너는 어때?

3.1.4. 문장부호의 사용

1) 느낌표나 쉼표는 사용하지 않으며 문장이 완전히 종결되었을 때는 마침표를 사용한다.

오늘의 뉴스를 시작하겠습니다!(×) 오늘의 뉴스를 시작하겠습니다.(○) 중화요리 집에서, 홀 서빙을 했었어요.(×) 중화요리 집에서 홀 서빙을 했었어요.(○)

2) 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분한다. 특히 '응', '네', '-어', '-어요' 등에서 말끝을 올리거나 내리는 것에 따라 의미가 달라질 경우, 마침표와 물음표로 구분한다.

평서문	의문문
나 이제 집에 갈래.	너도 집에 갈래?
응. 알겠어.	응? 뭐라고?
네. 잘 알겠습니다.	네? 다시 말씀해 주시겠어요?
저도 밥 먹었어요.	밥 먹었어요?

3) 도치된 문장의 경우 문장이 완전히 종결된 문장 다음에 마침표를 넣는다.

그 영화 봤어. 지난달에(×) 그 영화 봤어 지난달에.(○)

4) 억양단위의 문장이 완전히 끝나지 않았을 경우, 마침표를 붙이지 않는다. 이때 문장의 종결은 '-다, -어, -어요, -어라' 등 종결어미로 확인할 수 있다.

지난주에 그만두게 됐는데 이제.(×) 지난주에 그만두게 됐는데 이제(○) 갑자기 소리나 나더니.(×) 갑자기 소리가 나더니(○)

[주의] - 물음표를 잘못 넣은 경우(습관적으로 끝을 올리는 경우)
예) 그러니까 거기 가서 또?
우리가 또 자체적으로 만들어서도 먹고

예) 뭔가 사람들이 그 외적으로 필요한 거는 종이책도 어느 정도?
필요하다고 나는 개인적으로 그렇게 생각이 드는 거 같아.

5) [인용문 문장부호1] 인용문의 안긴 문장이 1개인 경우는 마침표, 물음표를 붙이지 않고 전체 문장에만 붙인다. 그리고 인용 조사는 안긴 문장에 붙여 쓰고, 쓴다.

예) 운동을 해야지라고(O)
운동을 해야지 라고(X)

예) 제가 늘 요즘에 친구들한테 자주 하는 말
생존을 위한 운동을 해야 <u>된다 그래요.</u> → '된다' 다음에는 마침표를 붙이지 않음.

예) 만약에 <u>운동 필요하다</u> 그러면
아 보건소에 한번 가 보시면 어떻겠냐고 이렇게 추천하기도 합니다. → '필요하다' 다음에는 마침표를 붙이지 않음.

예) 막 나중에 막 실리콘 튀어나오고
<u>그러면 재건 수술이 더 비용이 많이 든다</u> 그래서
코는 무서워서 안 -할- 안 (할려고)/(하려고) 합니다. → '든다' 다음에는 마침표를 붙이지 않음.

6) [인용문 문장부호2] 여러 문장이 안겨 있는 경우는 마지막 안긴 문장에만 문장부호를 붙이지 않는다.

예) 너 요즘 (쫌)/(좀)
유난히 버럭버럭하는 거 (같애.)/(같아.)

갱년기 오냐라고 얘기를 했대요.

→ 두 번째 줄 ‘같아’ 다음에는 마침표를 붙인다.

3.1.5. 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

3.2. 발화 내용 전사

3.2.1. 전사 일반 원칙

1) 각 전사에 사용할 수 있는 문자는 아래와 같다.

(X를 제외한 알파벳, 비식별화 일련번호를 제외한 숫자, 수식 기호 등 사용 금지)

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

2) 발음 전사 시 기호, 외래어 등은 발음에 따라 한글로 적는다.

(기호, 외래어의 철자 전사는 규범 표기를 기준으로 전사하며, 우리말샘을 기준으로 한다.)

철자 전사: 오리지널
발음 전사: 오리지날

철자 전사: 티브이
발음 전사: 티비

철자 전사: 아이유와 컬래버했어
발음 전사: 아이유와 콜라보했어

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

철자 전사: 어떡해
발음 전사: 어뜩해

철자 전사: 소주
발음 전사: 쏘주

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

3.2.2. 모음의 축약형 표기

1) 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.

사귀'어, 바뀌'어

3.2.3. 준말과 센말의 전사

1) <우리말샘>에 등재된 아래의 예와 같은 준말(한 단어 안에서 탈락이나 축약 현상이 일어난 것)과 센말은 본딧말로 복원하지 않고 발화된 대로 전사한다. (본딧말을 이중 전사하지 않음)

준말 예)(괄호 속은 본딧말)

근데(본. 그런데), 얘기(본. 이야기), 요새(본. 요사이), 요즘(본. 요즈음), 애(본. 아이), 담(본. 다음), 맘(본. 마음), 첨(본. 처음), 낼(본. 내일), 젤(본. 제일),

좀(본. 조금), 재밌다(본. 재미있다), 갖다(본. 가지다),-곤(본. -고는), 뭐(본. 무어), 오랜만(본. 오래간만), 암튼(본. 아무튼),

샘(본. 선생님), 알바(본. 아르바이트), 킬로(본. 킬로그램), 프로(본. 퍼센트) ...

센말 예)

쌤, 조끔, 쪼금, 쪼끔, 쫄쫄, 딴딴하다, ...

근데	(근데)/(그런데)	X
맘	(맘)/(마음)	X
백 프로	(프로)/(퍼센트)	X
낼	(낼)/(내일)	X
(일 키로)/(1킬로)	(일 키로)/(1킬로그램)	X
쪼금	(쪼금)/(조금)	X
딴딴하다	(딴딴하다)/(단단하다)	X

2) [주의] <우리말샘>에 등재되지 않은 경우(또는 일상대화로 등재된 경우)는 이중 전사 한다.

준말 예)

(알바비)/(아르바이트비), (왜냐면)/(왜냐하면), (그니까)/(그러니까), (이케)/(이렇게)[3], ...

센말 예)

(쫄쫄쫄)/(졸졸졸), (쫌)/(좀), ...

3) 준말과 비슷한 유형인 '줄어든 말'과 '줄여 이르는 말'은 이중 전사 하지 않는다. 이중 전사는 구어의 발음 정보를 제공하기 위한 것이므로 줄어들기 전의 형태 정보를 이중 전사 하지 않는다.

* [참고] 줄어든 말이란 '그게(그것이), 그걸(그것을)'처럼 단어의 경계를 넘어서, 조사나 어미 등이 결합하여 활용한 형태에서 탈락이나 축약이 일어나는 것을 말함.
* [참고] 줄여 이르는 말이란 '사법 시험(사시), 꾸안꾸(꾸민 듯 안 꾸민 듯하다)'처럼 두 단어 이상에서 단어마다 한 음절 이상씩 뽑아서 만든 말임.

3.2.4. 끊어진 단어(단어가 불완전하게 발화된 경우)

1) 끊어진 단어는 발화된 대로 그대로 전사한다. 불완전하게 발화된 단어(어절)가 둘 이상인 경우에도 어절마다 다음과 같이 표시하여 전사한다.

3) 비슷한 형태인 '일케'는 <우리말샘>에 줄어든 말로 등재되어 있어 이중 전사 하지 않음.

–전– –전– 전통이라고 우리가 흔히 얘기할 때

2) 다음과 같은 수정 발화나 반복 발화에 -를 붙이지 않는다.

수정 발화 학교 아니 유치원에 가게 되면(o) –학교– 아니 유치원에 가게 되면(x) 반복 발화 학교 학교 밖의 청소년(o) –학교– 학교 밖의 청소년(x)

3) 말이 꼬이는 등 의미 없이 하게 되는 발화(소리)에도 -를 붙여야 한다.(아래 예의 '츠, 트' 같은 소리)

예) **츠 트** 그~ 리더 같은 사람들 있잖아. (x)

=> -**츠-** -**트-** 그~ 리더 같은 사람들 있잖아.

4) 조사나 어미 단위에서 불완전한 발화나 자기 수정이 일어난 경우 다음과 같이 표시한다.

예1) 얼마큼 개선되고 **가까워-져-지느냐가** 이제

예2) **방문-은-만** 남았거든

예3) 그런 데 가는 거 좋아하고 하는데 다른 **분들-은-도** 그렇고

3.2.5. 띄어쓰기

1) 한글 맞춤법(제5장 띄어쓰기)에 맞게 띄어 쓴다.

2) 의존명사는 띄어 쓴다.

3) 수를 적을 때는 만 단위로 띄어 쓴다(예: 십이억 삼천백만 팔백구 불 등).

4) 본 용언과 보조 용언도 띄어 쓴다.(예: 먹어 버리다, 가고 싶다, 먹지 못하다, 가 보다, 먹어 주다, 사 가다, ...)

5) 띄어쓰기와 붙여쓰기 모두 허용되는 경우에는 띄어 쓰는 것을 원칙으로 한다.

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

6) 단어를 발음하는 중간에 쉼이 들어간 경우에는 띄어 쓰지 않는다.

7) 우리말샘 등재 내용을 기준으로 하며, 판단하기 어려운 경우에는 수시로 논의하여 결정한다.

3.2.6. 담화 표지

1) 머뭇거림의 기능을 하는 1음절 담화 표지 중 “이, 그, 저, 아, 어, 에, 음, 응, 뭐”의 9개 형태에 한해서 본래의 품사와 구별하기 위해 물결표(~)를 붙여 전사한다.

2) 즉, “이 사람, 그 사람, 저 사람”처럼 가리키는 말로 쓰이는 “이, 그, 저”나 감탄이나 응답 등의 “아, 어, 에, 음, 응, 뭐”가 원래의 의미로 쓰이지 않고, 말을 더듬거리거나 머뭇거림에 사용될 경우에만 물결표(~)를 붙여 표기한다.

담화 표지로 쓰인 예	지시·감탄·응답으로 쓰인 예
그~ 돈 벌고 싶어서	그 많은 돈을 벌고도
그냥 저~ 통상적인 노하우인지	저 사람 또 저런 소리를 하네.
응~ 얼마 만인지 모르겠네	응. 기억이 안 나. 모르겠어.
기껏해야 뭐~ 우리 나이대면 뭐~ 피시방?	그거 뭐지?/뭐든 좋아요.

3) [주의] 단, “인제, 이제, 그냥, 무슨, 어떤” 등의 2음절 이상의 담화 표지는 물결표(~)를 붙이지 않는다.

4) [주의] 물결표(~)는 머뭇거림의 담화 표지에 붙이는 특수한 부호이므로 장음의 부호로 사용하지 않는다. 특히 단음절 응답 표현의 경우 길게 발음할 때 이를 사용하지 않도록 주의한다.

예) 네~, 예~, 응~ (x)

5) [주의] 담화 표지는 이중 전사 대상이 아니다.

예) (그~)/(그) (x)

=> 그~

3.2.7. 잘 들리지 않는 부분

1) 잘 들리지 않는 부분의 전사 시 이중 괄호((xxx))를 이용한다.

2) 잘 들리지 않아 추정한 경우는 다음과 같이 전사한다.

그전까지는 직장 생활하느라 ((더 힘들어))

3) 화자의 발화 내용이 전혀 들리지 않는 부분은 다음과 같이 전사한다.

(()) 너무한 거 같더라.

4) 잘 들리지 않는 음절은 그 음절의 수만큼 x를 붙여 다음과 같이 전사한다. (음절 수 확인이 가능한 경우에만)

근데 그거 진짜 ((xx해야)) 되겠더라.

5) [주의] 정성 검수 결과, **약간 들리는데도 잘 안 들리는 부분으로 처리한 경우가 많았으므로 주의한다. (특히 '그~'로 들리는데 (())로 처리한 경우가 많았음)**

예) (()) -> ((근데)), (()) -> ((완전)), (()) -> 같아, (()) -> 개, (()) -> 거의, ...

3.2.8. 준음성과 기타 소리들

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

웃음: {laughing} 목청 가다듬는 소리: {clearing} 박수: {applauding} 노래: {singing} *철자 전사에서는 삭제한다.

2) 단, 전사창에는 다음과 같이 입력한다.

준음성 유형	전사
웃음: {laughing}	@웃음
목청 가다듬는 소리: {clearing}	@목청
박수: {applauding}	@박수
노래: {singing}	@노래

3) [주의] 위의 네 가지 준음성 외의 다른 항목은 입력하지 않는다. 간혹 숨 들이마시는 소리를 전사하는 경우가 있는데 삭제해야 한다.

예) 근데 그거 슙

=> 근데 그거

3.3. 이중 전사

1) 조사나 어미가 포함된 어절을 이중 전사하는 경우 발음 전사, 철자 전사 모두에 조사와 어미를 붙인다.

예) 엄청 액션 (씬이)/(신이) 많고

2) 마침표 등 부호를 포함한 경우 발음 전사, 철자 전사 모두에 부호를 붙인다.
예) 다큐 (좋아하구요.)/(좋아하고요.)

3.3.1. 숫자 전사(상세)

1) 숫자의 경우 이중 전사한다.

2) 발음 전사 시 숫자는 발음에 따라 한글로 적는다.

3) 철자 전사 시 숫자는 일반적인 표기 관습(숫자, 한글 혼용)에 따라 적는다.

나이_ (열 살)/(10살)
나이_ (마흔 살에)/(40살에)
시간_ (스물네 시간)/(24시간), (이십사 시간)/(24시간)
시간_ (열두 시)/(12시)
시간_ (십 몇 년)/(10 몇 년)
날짜_ (이천이십일 년 오월 이십일 일)/(2021년 5월 21일)
날짜_ (삼사일)/(3 4일), (오륙 년)/(5 6년), (일주일)/(1주일)
[띄어쓰기 주의] 이삼일, 삼사일, 오륙일, 일주일 등은 한 단어로 굳어져 사전에 등재되어 있으므로 붙여 씀
[이중 전사 주의] <우리말샘>에 한 단어로 등재되어 있지만, 숫자의 의미는 사라지지 않았으므로 이중 전사 해야 함
금액_ (삼만 칠천 원)/(3만 7000원)
금액_ (사천오백 원)/(4500원)
금액_ (이삼십만 원)/(2 30만 원)
측정_ (삼만 보)/(3만 보)
측정_ (삼십 키로)/(30킬로)
측정_ (십 프로)/(10프로)

4) 수 관형사는 이중 전사 하되 수사의 경우는 이중 전사 하지 않는다.

이 (두 개의)/(2개의) 매체는 (둘)/(2) 다 같이 공존할 수 있다고 생각해. (x)
→ 이 (두 개의)/(2개의) 매체는 둘 다 같이 공존할 수 있다고 생각해. (o)

그 (둘이랑)/(2이랑) 비교해서 어때? (x)
→ 그 둘이랑 비교해서 어때? (o)

5) [주의] 사전에 한 단어로 올라와 있는 명사, 부사로 쓰이는 경우들을 수 관형사와 구분해야 한다.

사소한 거 (하나하나를)/(1 1를) 또는 (하나하나씩)/(1 1씩) 다 일일이 신경 써야 돼? (x)
→ 사소한 거 하나하나 또는 하나하나씩 다 일일이 신경 써야 돼? (o)
※ '하나하나(씩)'은 숫자 1의 의미는 사라지고, "낱낱의 대상" 또는 "일일이"라는 새로운 의미로 쓰인 것이므로 숫자 전사의 대상이 아님

얼마 전에 유럽 (한번)/(1번) 나가 봤는데 (x)

→ 얼마 전에 유럽 한번 나가 봤는데 (o)

※ '한번'이 '기회 있는 어떤 때에'라는 의미로 쓰인 경우 '1회, 2회, ...'의 의미가 아니므로 숫자 전사 대상이 아님.

[비교] (한 번)/(1번) 실패하더라도 (o)

6) 숫자 철자 전사의 띄어쓰기는 "경", "조", "만" 단위로 띄어 쓴다.

(일억 이천만 원)/(1억 2000만 원)

(일조 이천삼백사십오억 육천칠백팔십구만 일천이백삼십 원)/(1조 2345억 6789만 1230원)

7) [주의] 철자 전사 시 천 단위 분할 ","(쉼표)는 쓰지 않는다.

예) 4000원 (O) → 4,000원 (X)

8) [주의] 외래어로 사용된 수는 외래어 표기법에 따라 이중 전사하고, 숫자를 적지 않는다.

예) 시즌 (뜨리)/(스리)

9) 숫자가 포함된 고유명의 경우 이중 전사 하지 않는다.

예) 1박 2일, 삼시 세끼

(1) 고유명(예능 프로그램)으로 사용되는 경우 - 이중 전사 X

예) 어제 일박 이 일을 봤는데

예) 어제 삼시 세끼를 봤는데

(2) 사전적 의미로 사용되는 경우

예) 우리는 (일박 이 일)/(1박 2일) 여행을 가서

=> '일박'은 <우리말샘>에 한 단어로 등재되어 있어 붙여쓰지만, 의미가 달라진 것은 아니므로 숫자 병기를 해 주어야 함

예) 삼시 세끼 챙겨 먹기는 힘들지

=> '삼시 세끼'는 굳어진 표현으로, '삼시'는 '세 번의 끼니'라는 의미로 세는 의미가 남아 있지 않고(일시, 이시, 삼시와 같이 쓰이지 않음), '세끼'는 '세 번 먹는 밥이라는 뜻으로, 하루하루의 끼니'를 의미하므로 이중 전사 하지 않는다.

예) 하루 (두 끼)/(2끼) 먹는 것도 힘들다.

예) 나는 하루에 (한 끼만)/(1끼만) 먹는다.

=> 여기서 '끼'는 밥을 먹는 횟수를 세는 단위로 쓰이므로 이중 전사 한다.

3.3.2. 외래어, 영문 및 기호의 전사

1) 외래어, 영문 및 기호의 발음 전사는 발화자의 발음에 따라 적고, 철자 전사는 한국어 어문규범 외래어 표기법에 따라 전사한다.

예) (뻐쓰)/(버스), (초콜렛)/(초콜릿)

[참고] (https://kornorms.korean.go.kr/example/exampleList.do?regltn_code=0003)

2) 고유명의 경우 발음대로 전사하고 규범에 맞는 철자 전사는 하지 않는다.
예) 유튜브, 유튭, 리플렉션, 뮬란, ...

3) 영문 약어의 철자 전사

알파벳	어문규범	알파벳	어문규범
A	에이	N	엔
B	비	O	오
C	시	P	피
D	디	Q	큐
E	이	R	아르
F	에프	S	에스
G	지	T	티
H	에이치	U	유
I	아이	V	브이
J	제이	W	더블유
K	케이	X	엑스
L	엘	Y	와이
M	엠	Z	제트

CCTV (씨씨 티비)/(시시 티브이)
SRT (에스알티)/(에스아르티)

3.3.3. 구어의 비표준 발음 이중 전사

*아래 예들은 구어 비표준 발음이므로, 이중 전사를 해야 하는 대상들임

1) 'ㅗ'가 'ㅜ'로 바뀜

▶ 조사에서: 도/두, 으로/으루

예 (오늘두)/(오늘도), (구체적으루)/(구체적으로)

▶ 어미에서: -고/-구, -고서/-구서, -고요/-구요, -더라고요/-더라구요

예 (얘기하구)/(얘기하고), (가지구서)/(가지고서), (언어구요)/(언어고요)
(뭐라구)/(뭐라고), (안다구)/(안다고), (몰라두)/(몰라도)

▶ 일부 어휘에서

예 (하두)/(하도), (별루)/(별로), (그대루)/(그대로)

2) 'ㅏ, ㅓ'가 'ㅐ, ㅔ'로 바뀜

▶ 일부 어휘에서

예 (챙피하다)/(창피하다), (멕이다)/(먹이다), (벳기다)/(벗기다),
(맨들다)/(만들다), (놀래다)/(놀라다), (같애요)/(같아요)

▶ 어미 및 어미 형태류에서

더래도(두)/더라도(두), 래서/라서

래는/라는, 대는/다는, 재는/자는

래면/라면, 대면/다면, 재면/자면

래니까/라니까, 대니까/다니까 재니까/자니까

래던데/라던데, 대던데/다던데, 재던데/자던데

래더라/라더라, 대더라/다더라, 재더라/자더라

예 (하더래도)/(하더라도), (한대는)/(한다는), (심하대니까요)/(심하다니까요),
(그랬대던데)/(그랬다던데)

3) 'ㅕ'가 'ㅡ'로 바뀜

▶ 선어말어미 -더-/-드-

더라고(요)/ 드라고(요), 던가요/든가요, 더라는/드라는

예 (가드라구요)/(가더라고요), (평범하드라는)/(평범하더라는), (좋든데)/(좋던데)

▶ 어미에서

그든요/거든요, 그든/거든

예 (당사국이그든요)/(당사국이거든요), (있그든)/(있거든)

▶ 기타

예 (이릏게)/(이렇게), (그릏게)/(그렇게), (저릏게)/(저렇게), (어뚱게)/(어떻게)

[주의] 구어에서 자주 쓰이는 조사 "(이)라든지, (이)라든가"를 각각 "(이)라던지, (이)라던가"로 발음하는 것과, 어미 "-든지"를 "-던지"로 발음하는 것은 구어의 비표준 발음이므로 이중

전사한다.

예) 엄청나게 큰 성주가 뭐~ 만들어 준 뭐~ (정원이라던가)/(정원이라든가) 이런 게 있는데

예) 밤새도록 (논다던지)/(논다든지) 친구들이랑 모여 (가지구)/(가지고)

4) [주의] 철자 전사에서 발화 내용을 과하게 수정하는 경우

- 아래와 같이 표현상의 오류인 경우는 이중 전사 하지 않음(수정하지 않음)

예) 완전 딜러 포지션 완전

=> 정말 딜러 포지션 정말 (x)

예) 공부랑은 거리가 멀게 살았었는데

=> 공부랑은 거리가 멀게 살았는데 (x)

예) 착한 학생이었어서

=> 착한 학생이어서 (x)

예) 서열 많은 대학교에 가고 그게 중요한 게 아니라

=> 서열 높은 대학교에 가고 그게 중요한 게 아니라 (x)

3.3.4. 방언의 전사

1) 방언형(발음 전사)에 대한 표준어 대응쌍(철자 전사)을 어절 단위로 이중 전사한다.

2) 우리말샘에 등재된 방언형의 경우 발음 전사는 방언형을 소리 나는 대로 기본 형태를 살려 적고, 철자 전사는 뜻풀이의 표준 어형을 기준으로 삼는다. 그 예는 다음과 같다.

지역	예시
강원	이게(다나?)/(다니?) 나도 이쪽 동네 (출신이래.)/(출신이야.) (이라)/(이렇게) 해. 모처럼 해가 난 (날에느)/(날에는) 마실이나(댕게오시우.)/(다녀오시오.) 애가 종일 (울민서)/(울면서) 쳐다보더라고. 돈이(읍어도)/(없어도) 남한테(아수운)/(아쉬운) 소리는 못 하겠다. (여서)/(여기서) 꾸물거리지 말고(얼푼)/(얼른) 가라. 마을 사람들은 뭐든 (농가)/(나누어) 먹지요.
경상	어제 어디 (갔었노?)/(갔었니?) (단디)/(단단히) 해라. 여기에 동그라미나 (꼽표)/(곱표) 치세요. 고등학교(댕길)/(다닐) 때 미역 (쭐거리)/(줄기) 반찬도 (마이)/(많이) (묵었지.)/(먹었지.) 떡을 (맹갈아)/(만들어) (묵었지.)/(먹었지.) 어제 어디 (갔었노?)/(갔었니?) 할 게 (천지삐까리다.)/(매우 많다.) 지금 (머라카노?)/(뭐라고 하니?) 내가 (하꾸마.)/(할게.) 그 직원이 (머스만데.)/(사내아인데.)

	(니맨치로)/(너처럼)
전라	혼자 다 (묵어)/(먹어) (분당께.)/(버린다니까.) (실맹키로)/(실처럼) 가는 거 그거? (그랑께)/(그러니까) 하루 종일 이엉만 (영끄고)/(엮고) 급히 약을 지었는데도 못 (나수고)/(낫고) (가부렀어.)/(가 버렸어.) 그거다 (이애기헐라면)/(이야기하려면) (미칠을)/(며칠을) 해도 안 돼. 아이들은 (훈지)/(그네) 뛰면서 놀고 있었다. 늦은 사람이 (댑대로)/(도리어) 큰소리친다.
제주	아까 (집드레)/(집으로) (가라.)/(가더라.) 너 (하구정한)/(하고 싶은) 대로(하라.)/(해라.) (모지레민)/(모자르면) (멋을)/(뭣을) 더 (가라)/(말해) (주코?)/(줄까?) 야 (무신)/(무슨) 그런 게 또 (시어.)/(있어.) 어떻든 (저디)/(저기) 다(지내치민)/(지나치면) (되우다.)/(됩니다.) 성격이 참 (요망지다.)/(야무지다.) (하르방)/(할아버지) 댁에 가는 길.
충청	동네 사람들은 (워떡헌댜?)/(어떡한대?) 가만히 (두덜)/(두질) (못햐.)/(못해.) (그려.)/(그래.) 너 (또래)/(때문에) (여기꺼지)/(여기까지) 와야(되겄어?)/(되겠어?) (오둥아를)/(오디를) 얼마나 (마이)/(많이) 먹었는지 입 안이 시커멓게 (물들었슈.)/(물들었어요.) 점심 때는(밥얼)/(밥을) 먹구 (새이)/(새참) 때는 (국시를)/(국수를) 먹는(겨.)/(거야.) 여기 (졸)/(부추) 한 단에 얼마요? (고쿠락)/(아궁이) 불이 꺼졌나 좀 봐라.

3) 일상대화는 구어가 갖는 특성이 드러나도록 표기한다. 이때 표기는 소리나는 대로 적는 것이 아니라 일상대화의 형태를 밝혀 적는다. 예를 들어, 표준형 '먹다'에 대한 경상도 일상대화는 '묵다'로, 그 활용형 '묵었지'를 소리나는 대로 쓰면 '무걷찌'이지만, 형태를 밝혀 쓰면 '묵었지'이다. 따라서 '무걷찌'로 쓰지 않고, '(묵었지.)/(먹었지.)'로 전사한다.

올바른 전사 표기	잘못된 전사 표기
점심은 아까 (묵었지.)/(먹었지.)	점심은 아까 (무걷찌.)/(먹었지.)
어제 어디(갔었노?)/(갔었니?)	어제 어디 (갇었노?)/(갔었니?) 어제 어디 (가썬노?)/(갔었니?)
오늘 날씨가 너무 (춥어서)/(추워서)	오늘 날씨가 너무 (추버서)/(추워서)
형이 (멋이간디)/(뭣이관데) 큰소리야?	형이 (머시간디)/(뭣이관데) 큰소리야?
하루에 같이 (검질멧주게.)/(김매었지.)	하루에 같이 (검질메쭈게.)/(김매었지.)
속상한 건 (위칙헌댜?)/(어떡한대?)	속상한 건 (위치컨댜?)/(어떡한대?) 속상한 건 (위치껀댜?)/(어떡한대?)

4) 일상대화에서 흔히 나타나는 어두 된소리화의 경우, 일상대화의 특성으로 볼 수 있으므로 소리 나는 대로 전사하고, 표준어 대응쌍을 이중 전사한다.

(쩌번에)/(저번에)
(따르다.)/(다르다.)
(꼐속)/(계속)

5) 지역별 일상대화 전사 시 다음을 주의한다.

(1) 경상 일상대화

① 종결어미에'-이'가 결합한 '-대이, -래이, -재이'은 소리대로 적는다.

집에 (갔대이.)/(갔다.)
전화 (해래이.)/(해라.)
다음에 (보재이.)/(보자.)

② 표준어의 '그러다'에 해당하는 '그카다, 그쿠다' 등은 소리대로 적는다.

(그카면)/(그러면) 저기 갔다 올 (끼가?)/(거가?)
(그쿠면)/(그러면) 그 일은 (끝났나?)/(끝났니?)
(그카고)/(그러고) 있지 말고 (일로)/(이리로) (온나)/(오너라)
(그쿠고)/(그러고) 잘난 척은 (자가)/(쟤가) 잘한다.

③ 받침'ㅇ'이나 'ㄴ'이 나타나지 않으면 소리대로 적는다.

(주머이)/(주머니)
(어무이)/(어머니)
(사이)/(산이) 크다
(학새이)/(학생이)

(2) 전라 일상대화

① 표준어 '-으니까'의 일상대화형은 '-응께, -응게, -응께네, -으니께' 등으로 소리 나는 대로 적는다.

(인자는)/(인제는) 약이(조응께.)/(좋으니까.)
그 공식적으로 다 (허니께.)/(하니까.)

② 둘째 음절 이하의 'ㅎ'이 나타나지 않는 말의 경우에는 소리대로 적는다.

(뭐다러)/(뭣하러) 그러냐?
잘하지도(모다고)/(못하고)
(배과점에)/(백화점에) 가(봉께)/(보니까)

으메 (답다번)/(답답한) 거 눈앞이 (깝까버다.)/(갑갑하다.)

③ 'ㄴ'이 나타나지 않으면 소리 나는 대로 적는다.

(가마이)/(가만히) (마이씩)/(많이씩) (아잉)/(아닌) 게(아이라)/(아니라)

(3) 제주 일상대화

① 자음으로 끝나는 단어와 모음으로 시작하는 단어가 만나 한 단어를 이룰 때, 제주 일상대화에서는 후행하는 단어에 선행하는 단어의 마지막 자음을 덧붙여 발음하는 특성이 있다. 이는 제주 일상대화의 특성이므로 소리 나는 대로 적는다

(맏다덜)/(맏아들) (비단놋)/(비단옷) (감못)/(감옷)

(4) 충청 일상대화

① 종결어미 '-댜'는 소리대로 적는다.

그 낮에 꿈을 (꾸니께)/(꾸니까) (그라댜)/그러더래) 우리 (야덜)/(아들) (잡어)/(잡아) (간댜.)/(간대.)

② 표준어 '어찌'의 방언형인 '워떻게, 어티기' 등은 소리대로 적는다.

(워떻게)/(어찌) 하는 줄 (아슈?)/(알아요?) 장사하려고 (어티기)/(어찌) 집을 크게 (졌는디)/(졌는데)

6) 다음은 방언으로 혼동하기 쉬운 표준어 사례이다. 이들은 이중 전사하지 않으므로 주의한다.

걸쩍지근하다	[형용사] 다소 푸짐하고 배부르다. (예) 그 사람, 정말 걸쩍지근하게 잘도 먹더군. [형용사] 말 따위가 다소 거리낌이 없고 푸지다. (예) 걸쩍지근한 사설을 늘어놓다.
인제	[명사] 바로 이때. [부사] 이제에 이르러.
식겁(食怯)	[명사] 뜻밖에 놀라 겁을 먹음.
골통	[명사] 척추동물의 두뇌를 덮어 싼 부분. [명사]'머리'를 속되게 이르는 말
꼴통	[명사] 머리가 나쁜 사람을 속되게 이르는 말.

개고생	[명사] 어려운 일이 나고 비가 닥쳐 톡톡히 겪는 고생.
삐대다	[동사] 한군데 오래 눌어붙어서 끈덕지게 굴다.
개기다	[동사] (속되게) 명령이나 지시를 따르지 않고 버티거나 반항하다.
띵하다	[형용사] 울리 듯 아프고 정신이 흐릿하다.
조지다2	[동사] (속되게) 일신상의 형편이나 일정한 일을 망치다. [동사] (속되게) 쓰거나 먹어 없애다.
조지다1	[동사] 짜임새가 느슨하지 않도록 단단히 맞추어서 박다. [동사] 일이나 말이 허술하게 되지 않도록 단단히 단속하다.
아니꼽다	[형용사] 비위가 뒤집혀 구역날 듯하다. [형용사] 하는 말이나 행동이 눈에 거슬려 불쾌하다.
대가리	[명사] 동물의 머리. [명사] 사람의 머리를 속되게 이르는 말. [명사] 주로 길쭉하게 생긴 물건의 앞이나 윗부분.
증(憎)하다	[형용사] 모양이 지나치게 크거나 괴상하여 보기에 흉하고 징그럽다.
꼽사리	[명사] 남이 노는 판에 거저 끼어드는 일.
짝퉁	[명사] 가짜나 모조품을 속되게 이르는 말.
오지다	[형용사] 마음에 흡족하게 흐뭇하다. [형용사] 허술한 데가 없이 알차다.

3.4. 비식별화를 위한 전사

1) 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.

2) 비식별화 정보는 아래와 같이 입력한다.

내용	입력 형식	비고
이름	@이름1	
계정(아이디)	@계정1	이메일 주소, 온라인 아이디 포함
주민등록번호	@주민번호1	
전화번호	@전화번호1	
카드번호	@카드번호1	
기타번호	@기타번호1	
주소	@주소1	
출신 및 소속	@소속1	
기타 비식별화가 필요한 항목	@기타1	욕설을 포함한 비윤리적인 표현 등
상호명 및 상품명	@상호명1	부정적인 평가를 포함한 경우에 한해 입력

3) 모든 비식별화 대상에는 '@이름1, @이름2, …'와 같이 일련번호로 구분하여 입력한다. 이때 한 파일 내에서 해당 번호가 가리키는 대상은 동일해야 한다.

그때 철수랑 민수랑 너랑 나랑 갔잖아. 철수도 알고 있지?	
올바른 입력	잘못된 입력
그때 @이름1이랑 @이름2이랑 너랑 나랑 갔잖아. @이름1도 알고 있지?	그때 @이름1이랑 @이름2이랑 너랑 나랑 갔잖아. @이름3도 알고 있지?

4) 정치인 등 유명인의 이름은 비식별화하지 않으며, 상호명 및 상품명 등은 부정적인 경우에만 비식별화한다. 예를 들어, 학교명, 기관명, 단체명, 영화 제목, 노래 제목, 책 제목, 방송 제목, 게임명, 상품명, 제품명, 넷플릭스, 유튜브, 삼성, 엘지, 애플, 아이폰 등 널리 알려진 상호는 비식별화 대상이 아니다.

비식별화하지 않는 경우	나는 핸드폰 매번 삼성 것만 쓰다가 이번에 아이폰으로 바꿨어.
비식별화하는 경우	너 정자동에 있는 @상호명1에서 김밥 먹어 봤어? 거기서 김밥 사 먹고 50명 넘게 식중독 걸렸대.

5) 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다.

근데 너 서대문구 연희동 살잖아. 너 지금 연희동 청구아파트 살지? → 너 지금 연희동 @주소1 살지? 교수님 댁이 연희동 136번지 맞지? → 교수님 댁이 연희동 @주소2 맞지?

6) 비윤리적인 표현은 비식별화한다. 비윤리적 표현에는 욕설, 차별/혐오 표현, 성적인 표현이 있다.

비식별화 전	비식별화 후	유형
씨발 기억이 안 나.	@기타1 기억이 안 나.	욕설
완전 미친년이네.	완전 @기타1이네.	욕설
존나 흥미가 생길 거 같은데.	@기타1 흥미가 생길 거 같은데.	욕설
틀딱들이 지하철 타면 너무 싫어.	@기타1들이 지하철 타면 너무 싫어.	차별/혐오 표현
너도 개독교냐?	너도 @기타1냐?	차별/혐오 표현
요즘 맘충이 정말 많아졌지 않아?	요즘 @기타1이 정말 많아졌지 않아?	차별/혐오 표현
여자가 죽을려고 하면 남자가 와서 뭐~ 죽기 전에 한 번 하자	여자가 죽을려고 하면 남자가 와서 뭐~ 죽기 전에 @기타1	성적인 표현

※ 비윤리적 표현의 비식별화 시 다음을 주의한다.

(1) 강조의 의미를 갖는 다음과 같은 말은 비식별화하지 않는다.

[예] 진짜 연출력 미쳤다.

→ '미치다'+'사람 명사(년, 새끼 등)'와 함께 쓰일 경우에는 비식별화하고, '미친'+ '연기력, 날씨, 가창력' 등과 결합할 경우에는 비식별화하지 않는다.

(2) 비속한 표현이지만 욕설이라고 보기 어려운 말은 비식별화하지 않는다.
[예] 존맛탱, 대존맛, 까먹다, 개좋다, 개꿀…

※ 차별/혐오 표현을 판단할 때, '네이버 국어사전'을 참고할 수 있는데, 해당 표제어 아래에 "차별 또는 비하의 의미가 포함되어 있을 수 있으므로 이용에 주의가 필요합니다. (차별 표현 바로알기 캠페인)"라는 문구가 있는 경우 비식별화 대상이 된다.
[예] 병신, 거지 등

3.5. 기타

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

2) 메타 데이터의 내용과 화자가 일치하는지 확인한다. 확인해야 할 내용은 다음과 같다.

대화I D 대화 주제 일치 여부 화자의 통계 정보 : 연령/직업/성별/출생지/화자 관계

3) 맞춤법, 일상 대화 등에 관한 정보가 필요할 경우 다음을 참고할 수 있다.

[붙임 2] 개인정보 수집 · 이용 동의서

개인정보 수집·이용 동의서

㈜팀벨, ㈜솔트룩스는 국립국어원의 “2023년 일상 대화 말뭉치 구축” 과제에 참여하여 [개인정보보호법] 제15조 및 제17조에 따라 아래의 내용으로 개인정보를 수집·이용합니다. (개인정보 수집·이용 동의에 거부할 수 있으며, 미동의시 과제참여가 불가능합니다)

개인정보 수집·이용자	개인정보 수집·이용 목적	수집·이용 개인정보 항목	보유/이용기간
㈜팀벨, ㈜솔트룩스	**◆ ㈜팀벨 - 일상 대화 말뭉치 구축 과제의 음성 말뭉치 수집, 과제 중 음성데이터 전사, 개인식별정보 등 제거 업무** **◆ ㈜솔트룩스 - 과제관리, 최종데이터 검수 업무**	**발화 음성, 인구통계학적 정보 (출생지/성장지/거주지/성별/연령대/화자 간 관계/직업/학력)**	**2023년 12월 15일까지**
㈜팀벨	**과제 참여에 대한 회계 정산 처리 및 비용 증빙**	**성명, 주민등록번호**	**2023년 12월 15일까지**

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

[☐ 동의합니다 ☐ 동의하지 않습니다.]

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

[☐ 동의합니다 ☐ 동의하지 않습니다.]

◆ 고유식별정보의 처리에 관한 사항

㈜팀벨은 개인정보보호법에 관한 법률에 따라 회계 정산 처리 신고 목적으로 고유식별정보인 주민등록번호를 처리(수집·이용)하고자 합니다. 보유/이용기간은 2023년 12월 15일까지입니다. 이에 동의하십니까?
(개인정보 수집 이용 동의에 거부할 수 있으며, 동의하지 않을 경우 과제 참여가 불가능합니다.)

[☐ 동의합니다 ☐ 동의하지 않습니다.]

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

[□ 동의합니다　　□ 동의하지 않습니다.]

2023년　월　일

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

㈜팀벨, ㈜솔트룩스

[붙임3] 개인정보 제3자 제공 동의서

개인정보 제3자 제공 동의서

㈜팀벨, ㈜솔트룩스는 국립국어원의 “2023년 일상 대화 말뭉치 구축” 과제에 참여하여 [개인정보보호법]에 따라 아래의 내용으로 개인정보를 국립국어원에 제공합니다.(귀하는 개인정보 제3자 제공 동의에 거부할 수 있으며, 미동의시 과제 참여가 불가능합니다.)

개인정보를 제공받는 자	제공받는 목적	제공되는 개인정보 항목	보유/이용기간
국립국어원	**◆ 일상 대화 말뭉치 구축 과제의 음성 말뭉치 및 인구통계학적 정보의 기초정보 구분** **◆ 일상 대화 말뭉치 구축 결과물로 언어 연구 및 언어정보 처리분야 응용 기술 개발에 제공** **◆ 국립국어원 시행 타 사업 및 국립국어원 발주 타 용역 사업의 원시데이터로 활용되어 2차적저작물로 가공(외국어, 수어로 번역 가공 포함)될 수 있음**	**발화 음성, 인구통계학적 정보(출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)**	**기본 2044년 12월 31일, 이후 5년 단위 자동 갱신**

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

[☐ 동의합니다 ☐ 동의하지 않습니다.]

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

[☐ 동의합니다 ☐ 동의하지 않습니다.]

2023년 월 일

신청인 성명 : _____(자필서명)

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

㈜팀벨, ㈜솔트룩스 귀중

[붙임4] 국립국어원의 개인정보 제3자 제공(공개) 동의서

국립국어원의 개인정보 제3자 제공(공개) 동의서

본인은 국립국어원의 “2023년 일상 대화 말뭉치 구축” 과제에 참여하여 국립국어원이 [개인정보보호법]에 따라 아래의 내용으로 개인정보를 제3자에 제공(공개)하는데 동의합니다.(귀하는 개인정보 제3자 제공 동의에 거부할 수 있으며, 미동의시 과제 참여가 불가능합니다.)

개인정보를 제공받는 자	제공(공개) 목적	제공되는 개인정보 항목	보유/이용기간
학계·연구기관·산업체	**◆ 일상 대화 말뭉치 구축 결과물로 언어 연구 및 언어정보 처리분야 응용 기술 개발에 제공** **◆ 국립국어원 시행 타 사업 및 국립국어원 발주 타 용역 사업의 원시데이터로 활용되어 2차적저작물로 가공(외국어, 수어로 번역 가공 포함)될 수 있음**	**발화 음성, 인구통계학적 정보(출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)**	**기본 2044년 12월 31일, 이후 5년 단위 자동 갱신**

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

[☐ 동의합니다 ☐ 동의하지 않습니다.]

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

[☐ 동의합니다 ☐ 동의하지 않습니다.]

2023년 월 일

신청인 성명 : _____ (자필서명)

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

국립국어원 귀중

[붙임5] 저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

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

다 음

제1조 (계약의 목적)
본 계약은 저작재산권 이용 허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)
본 계약의 이용 허락 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

저작물: 일상 대화
저작자:
종별: ☑ 어문저작물
권리: ☑ 복제권, ☑ 공중송신권, ☑ 배포권, ☑ 2차적저작물작성권

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

제3조 (이용 허락 기간)

대상저작물의 이용 허락 기간은 계약체결일부터 2044년 12월 31일까지로 하며, 계약 기간 만료 시 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 계약기간 만료 시 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)
(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을 이용할 권리를 제3조의 기간 동안 비독점적으로 허락한다.

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

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

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

제5조 (이용자의 권리 및 의무)
(1) 이용자는 대상저작물을 제3조의 이용 허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다.

(2) 이용료는 설정하지 아니한다.

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

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

제6조 (확인 및 보증)
(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.
대상저작물의 저작권 이용 허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유

하고 있다는 것
대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
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.
대상저작물에 적용된 이용 허락 조건에 의해서만 대상저작물 재이용을 허락할 것
대상저작물을 권리자 및 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)
본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.

제8조 (계약의 해지)
(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

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

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

제9조 (손해배상)
당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

제10조 (비용의 부담)
계약 체결에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁해결)

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

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

제12조 (비밀유지)
양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용 및 대상저작물의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

제13조 (기타부속합의)
(1) 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.

2023년 월 일

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

[붙임6] 저작권 이용 허락 계약서 미성년자 법정대리인용 동의서

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

본인은 미성년자의 법정대리인으로 해당 미성년자가 국립국어원의 “2023년 일상 대화 말뭉치 구축” 과제에 참여하여 별첨과 같은 “국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서”를 체결하는 점에 대해 충분히 내용을 검토하였고, 해당 계약서 체결에 동의합니다.

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

2023년 월 일

미성년자 성명 :

법정대리인(보호자) : _____ (자필서명)

국립국어원 귀중

<기획·연구>
국립국어원 강미영 언어정보과장
국립국어원 정주연 학예연구사
국립국어원 강정미 연구원

<사업 참여자>
사업책임자 성기완 ((주)솔트룩스)
사업참여자 김 준, 김예하나, 박선희,
강수빈, 박문수 ((주)솔트룩스),
이상준, 김응준, 노강일, 김선아,
김선희, 송혜주, 이인호, 박선욱 ((주)팀벨)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2023년 12월 15일
발행일: 2023년 12월 15일
인 쇄: 근아인쇄

※ 이 책은 국립국어원의 용역비로 수행한 '2023년 일상 대화 말뭉치 구축' 사업의 결과물을 발간한 것입니다.