국립국어원 2025-01-37

발간등록번호
11-1371028-100034-01

2025년 일상 대화 자료 수집 및 정제

사업책임자

이 용 주

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2025년 일상 대화 자료 수집 및 정제’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2025년 2월 17일 ~ 2025년 12월 20일

2025년 12월 20일

사업책임자: 이용주((주)나라지식정보)

사업 수행자: ㈜나라지식정보 컨소시엄

사업 책임자: 이용주

사업 참여자: 박분선, 이지현, 황주영, 박연미, 박영훈, 이재혁, 오지혜,
김희영, 박한주, 차정훈, 신은주, 김관철, 김선희

<사업 수행자> ㈜나라지식정보 컨소시엄

사업 책임자	이용주((주)나라지식정보)
사업 참여자	박분선((주)나라지식정보)
	이지현((주)나라지식정보)
	황주영((주)나라지식정보)
	박연미((주)나라지식정보)
	박영훈((주)나라지식정보)
	이재혁((주)나라지식정보)
	오지혜((주)나라지식정보)
	김희영((주)팀벨)
	박한주((주)팀벨)
	차정훈((주)팀벨)
	신은주((주)팀벨)
	김관철((주)팀벨)
	김선희((주)팀벨)

<국문 요약>

2025년 일상 대화 자료 수집 및 정제

이 사업은 2019년부터 이어온 일상 대화 말뭉치 구축 사업으로, 화자 모집 계획과 말뭉치 구축 지침에 따라 정제본 650시간 규모의 말뭉치를 구축하여 활용도 높은 국어 말뭉치를 마련하고자 하는 데 그 목적이 있다. 올해는 2인, 3인, 4인 대화와 함께 100시간 규모의 1인 발화(독백)의 수집도 포함되어 있다.

이에 따른 주요 과업과 사업 성과는 다음과 같다.

음성 녹음 및 정제: 2인, 3인, 4인 대화는 통계청의 인구 통계 분포를 참고하여 지역별, 성별, 나이별 다양한 화자를 모집하고 총 1,706명의 화자가 10개 주제를 기반으로 12분에서 18분간 자연스러운 대화를 녹음하였다.

또한 1인 발화(독백)은 총 152명이 참여하였고, 이 가운데 유튜브 기반 독백은 62명에 8개 주제로 구성되었으며, 강연·연설 등 공적 독백은 90명에 4개 주제를 바탕으로 약 30분 내외 분량으로 구성되었다.

참여한 화자 모두 저작권 이용 허락 계약서를 작성하였고, 최종 음성 파일은 16kHz 샘플링, 16비트 양자화 선형 피시엠(PCM: 펄스 코드 변조) 형식으로 저장했다.

음성 자료 전사: 관련 업무 경험이 많은 전문 전사자를 선발하고 전사 지침을 숙지할 수 있도록 교육을 진행하였다. 전문 전사자들은 전사 도구를 활용하여 지침에 따라 발음과 철자를 구분하여 전사했다.

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>

Collecting and refining dialogue corpus 2025

This project is to build a dialogue corpus that has been ongoing since 2019. The purpose of this project is to prepare a highly usable Korean language corpus by constructing a corpus of 650 hours of refined copies according to the speaker recruitment plan and corpus construction guidelines. This year, there will be 100 hours of one-person speech (monologue) along with dialogue between two, three, and four people.

The major tasks and business outcomes resulting from this are as follows.

Speech recording and refining:
The conversation of two, three, and four people recruited various speakers by region, gender, and age by referring to the demographic distribution of the National Statistical Office, and a total of 1,706 speakers recorded a natural conversation for 12 to 18 minutes based on 10 topics. In addition, a total of 152 people participated in the single speech (monologue), of which 62 people participated in the YouTube-based monologue consisted of eight themes, and public monologues such as lectures and speeches consisted of about 30 minutes based on four themes for 90 people. All the participating speakers wrote a copyright license agreement, and the final voice file was saved in the form of 16-kHz sampling and 16-bit quantized linear PCM (Pulse Code Modulation).

Speech material transcription: Professional transcribers with extensive relevant work experience were selected and training was provided to ensure that they were familiar with transcription guidelines. Professional transcribers used transcription tools to transcribe pronunciation and spelling according to instructions.

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

Key words: dialogue corpus, raw corpus, inter-speaker relationships, double tier transcription, speech transcription

차 례

표 차례

표 차례

그림 차례

제1장

사업 개요

1. 사업의 목적

2019년부터 지속적으로 구축하여 온 일상 대화 자료 수집 및 정제 사업의 2025년도 사업 목적은 언어 인공지능 등 관련 산업 활용을 위한 기반을 마련하고, 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용하기 위한 650시간 이상의 고품질의 일상 대화 말뭉치를 구축하여 민간에 공유하는 것이다.

대규모 고품질 일상 대화 말뭉치 구축을 통해, 기초 말뭉치의 양적·질적 부족에 따른 기반 기술을 개발하고, 인공지능 기술 개발 수준 지체를 해소하여 국어 자원의 활용도와 가치를 높일 수 있도록 민간에서 활용 가능한 국가 공공재로서의 말뭉치 확대 구축을 목적으로 하고 있다.

[그림 1] 일상 대화 말뭉치 구축 사업 목적

2. 사업 수행 범위

○ **말뭉치 구축 기획 및 품질관리**
- 일상 대화 녹음 관련 다양한 주제 및 배경 관련 상세 기획
- 말뭉치 구축 및 활용 관련 전문가 자문 실시
- 고품질 말뭉치 데이터 확보를 위한 단계별 품질 점검 프로세스 확보

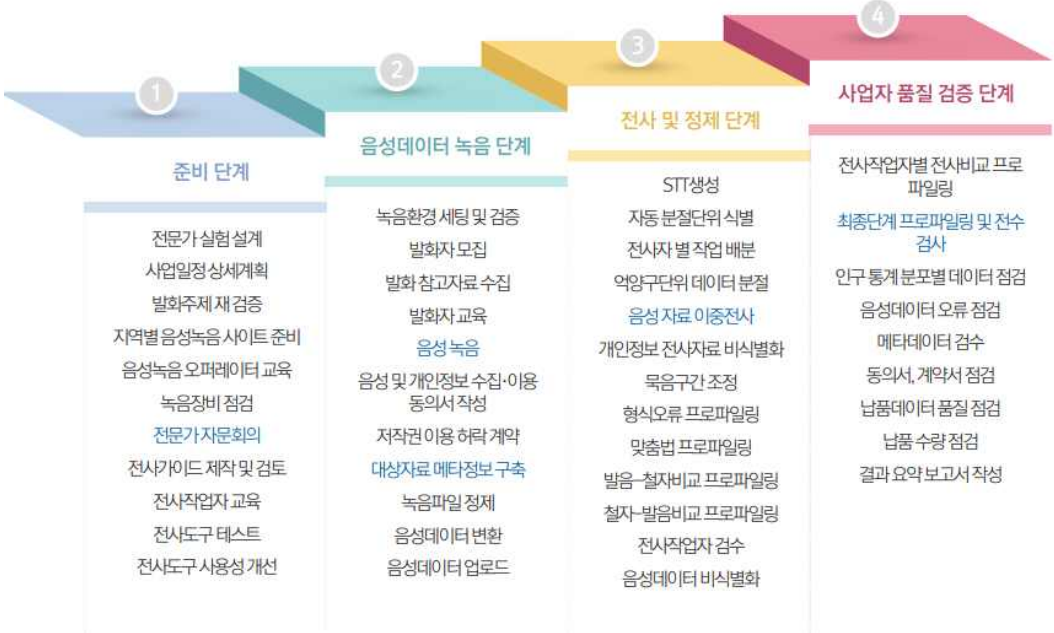
○ **일상 대화 녹음 및 음성자료 정제**
- 통계청 인구 통계 분포를 기반으로 지역별·연령별·성별 인구 통계 분포를 고려한 대화 참가자 모집
- 두 명 이상이 특정 주제에 대해 자유롭게 대화하는 음성 녹음 및 정제(정제 후 550시간 분량, 주제 또는 제시 자료당 대화 시간 12~18분)
- 1인 발화(독백)를 녹음하고 정제(정제 후 100시간 분량)
- 화자별 채널 분리, 전사 단위에 따른 음성 분할, 개인 정보 비식별(묵음) 처리
- 화자 대상 음성 및 개인 정보 수집·이용 동의, 음성 자료 대상 저작권 이용 허락 계약 체결

○ **음성 자료 이중 전사 및 원시 말뭉치 구축**
- 음성 자료의 전사는 발화된 그대로 전사하는 발음 전사
- 한글 맞춤법, 표준어 규정 등 어문 규정에 따른 철자 전사
- 전사 결과물에 대해 헤더 정보 부착 등의 표지 부착 작업을 수행하여 원시 말뭉치 형태로 가공
- 모든 개인 정보의 비식별 처리 및 혐오나 차별, 성적 표현 등 비윤리적인 내용 분리
- 음성 자료 이중 전사 및 원시 말뭉치 구축

○ **구축 대상 자료에 대한 메타 정보 구축**
- 녹음 날짜, 대화 아이디, 대화 주제/제시 자료, 화자 아이디, 화자 정보(성별, 연령대, 직업, 출생지, 주 성장지, 현 거주지 등),
- 화자 간 관계 등 구축 대상 자료에 대한 정보 구축

3. 사업 수행 절차

이번 사업은 준비 단계, 음성데이터 녹음 단계, 전사 및 정제 단계, 품질 검증의 4단계로 진행되었다. 사업 기간 내에 목표로 한 구축량 달성을 위해 각 단계별 임무를 명확히 하고, 공정별 품질검증 과정을 두어 최종 말뭉치의 품질에 문제가 없도록 하였다.



[그림 2] 사업 수행 절차

첫 번째 준비 단계는 기획 및 설계 단계로서 일상 대화 말뭉치를 수집하기 위한 상세 내용을 기획하였다. 수집 방법 및 장소 확보, 대화 주제 선정, 발화자 비율 설정, 일관된 전사를 위한 절차 및 지침서 보완 등 말뭉치를 균등한 비율로 동일한 조건에서 수집할 수 있도록 설계하였다. 아울러 지역별 녹음 사이트의 준비, 오퍼레이터의 교육을 비롯한 본격 녹음 수집 및 전사를 준비하는 단계이다.

두 번째 음성 데이터 녹음 수집 단계에서는 녹음 작업자인 오퍼레이터와 발화자를 모집하고 수집 환경을 구축하였으며, 실제적인 음성 데이터 수집 및 정제가 이루어지는 단계이다. 수집 시 모든 발화자에게 저작권 계약서 및 개인 정보 동의서를 받았고, 전사 도구에 음성 파일을 적재하기 위해 음성 파일을 알맞은 형식으로 정제 및 변환하였으며 이와 동시에 음성 데이터 및 발화자 메타데이터를 생성하여 구축하였다.

세 번째 전사 및 정제 단계에서는 전사 도구를 활용하여 음성 데이터 전사를 진행하였다. 이에

앞서 음성 전사 및 개인 정보 비식별화가 가능하도록 도구 개발 및 고도화를 진행하였으며, 전사 지침서를 활용하여 전사 작업자를 대상으로 교육을 진행하였다. 관리자가 전사 도구에 음원 파일을 업로드하면 시스템에서 자동 STT(Speech-to-Text)를 수행하여 전사본을 생성하고, 동시에 발화 구간을 기준으로 문장 단위 데이터 싱크 분할(클립 생성)을 자동으로 처리한다. 이를 통해 초기 전사 텍스트와 문장 단위 시간 정보가 포함된 1차 결과물이 마련된다. 이후 작업자에게 데이터를 배정하고 작업자는 할당된 데이터로 전사 작업을 진행한다. 원음을 직접 청취하면서 자동 전사 결과를 전사 지침에 따라 수정·보완한다. 오인식, 탈락, 중복 표기 등을 교정하고, 구어적 특성 및 발화 특수 요소를 작업 기준에 맞게 정비한다. 아울러 문장 단위 클립의 시작 및 종료 시점을 재확인하여 음성과 텍스트 간 싱크를 정밀하게 보정한다. 작업이 완료되면 검수를 요청하고, 검수자에게 할당되면 해당 데이터를 대상으로 2차 품질 점검을 수행한다. 검수 단계에서는 음원과 문장 단위 싱크의 정확성, 작업 지침 준수 여부, 전사 내용의 정확성 및 누락·오류 여부를 종합적으로 확인한다. 또한, 전사가 완료된 말뭉치는 대화 주제, 발화자 성별, 연령, 주 성장지 등의 메타데이터를 생성하고 부착하였다.

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

제2장

사업 수행

1. 대화 주제 및 제시 자료 선정

1.1. 대화 주제

대화 주제의 선정을 위하여 먼저 2019년부터 2024년까지 수집해 온 대화의 주제들을 살펴보면 다음과 같다.

<표 1> 2019년부터 2024년까지의 대화 주제

◎ 2019년

1	군대	군대 경험 등	9	방송 연예	연예인, 드라마, 방송, 프로그램, 연예계 이슈 등
2	게임	게임 종류, 게임 방법, 게임 경험 등	10	스포츠, 레저	스포츠, 운동선수, 올림픽, 운동 경기 관람, 등
3	휴일	휴일, 휴일 여행 등	11	먹거리	음식, 음료, 요리법, 요리사, 맛집 등
4	자동차	자동차, 자동차 관리, 종류, 위반 주차 등	12	자연 휴양지	산, 공원, 바다, 국내외 휴양지 등
5	만화	만화, 만화영화 만화 작가, 웹툰 등	13	국가 지역	나라, 세계 도시, 국내 도시
6	영화	영화, 영화인, 영화관, 영화제 등	14	문학	작가, 책, 오디오 북 등
7	정치	정치인, 정치적 이슈, 선거 등	15	연애, 결혼	연애, 결혼, 배우자, 애인, 결혼 생활, 데이트 등
8	건강, 다이어트	질병, 다이어트, 건강관리, 건강 검진 등	16	경제 재테크	경제 재테크, 자산 관리, 부동산, 금융, 증권, 보험, 대출 등

◎ 2020년

1	스포츠, 레저	종목, 운동선수, 올림픽, 경기 관람 등.	9	선물	추억, 종류, 이벤트, 핸드메이드 등
2	여행지(국내, 국외)	장소(나라, 지역), 관광 명소, 여행 계획, 경험 등	10	꿈(목표)	꿈(과거, 금년), 장래 희망 등
3	계절, 날씨	봄, 여름, 가을, 겨울, 추억 등	11	연애, 결혼	이상형, 데이트, 배우자, 연애관, 자녀 등
4	회사, 학교	재직(재학) 중인 곳, 학창 시절, 동창, 선생님, 동아리 등	12	반려동물	추억, 반려동물 종류, 동물 이름, 질병 등
5	먹거리	음식, 맛집, 요리법, 요리사 등	13	아르바이트	종류, 추천, 경험 등
6	방송, 연예	연예인, 프로그램, 이슈 등	14	성격	혈액형, 다혈질, 소심한 등
7	영화	영화인, 영화관, 영화제, 영화 장르 등	15	가족	가족 관계, 형제, 자매 등
8	건강, 다이어트	질병, 약, 건강 보조제, 건강 관리, 약물 부작용 등			

◎ 2021년

1	휴가	여행 시 교통, 숙박 선택	9	경제, 재테크	집, 주식 등 투자에 대한 정보와 결정
2	대중교통	약속 시간, 장소, 교통, 선택	10	회사, 학교	취직, 진학에 대한 정보와 결정
3	음악	대중음악 유행, 선호 가수 및 곡 추천	11	반려동물	개, 고양이의 장단점 비교 및 결정
4	건강, 다이어트	성인병에 대한 상식, 처방, 대응	12	취직	대기업/중소기업 취직의 정보 공유와 견해 교환
5	방송, 연예	드라마, 예능. 프로그램 선택	13	가족	집안 행사에 대한 검토와 결정
6	스포츠, 레저	직접 운동, 관람, 시청 등 참여 방법에 대한 정보와 결정	14	쇼핑	핸드폰 구매시 기종 검토 및 결정
7	먹거리	저녁 모임에 대한 음식 종류와 식당 선택	15	관혼상제	결혼, 문상, 제사, 축의금, 참석 등
8	우정	친구 간 선호, 성격, 취미 토론	16	경제 재테크	경제 재테크, 자산 관리, 부동산, 금융, 증권, 보험, 대출 등

- 협력적 대화

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

◎ 2022년

	대화주제	수집수량	수집비율		대화주제	수집수량	수집비율
01	휴가	112	6.33%	09	경제, 재테크	114	6.44%
02	대중교통	117	6.61%	10	회사, 학교	109	6.16%
03	음악	106	5.99%	11	반려동물	10	6.11%
04	건강, 다이어트	117	6.61%	12	취직	116	6.56%
05	방송, 연예	120	6.78%	13	가족, 관혼상제	106	5.99%
06	스포츠. 레저, 취미	109	6.16%	14	쇼핑	107	6.05%
07	먹거리	113	6.39%	15	생할, 주거환경	107	6.05%
08	우정	104	5.88%	16	기타	104	5.88%

- 비통제 환경 일상대화

	대화주제	수집수량	수집비율		대화주제	수집수량	수집비율
01	휴가	28	6.60%	09	경제, 재테크	24	5.66%
02	대중교통	29	6.84%	10	회사, 학교	26	6.13%
03	음악	25	5.90%	11	반려동물	28	6.60%
04	건강, 다이어트	26	6.13%	12	취직	24	5.66%
05	방송, 연예	27	6.37%	13	가족, 관혼상제	25	5.90%
06	스포츠. 레저, 취미	25	5.90%	14	쇼핑	27	6.37%
07	먹거리	31	7.31%	15	생할, 주거환경	28	6.60%
08	우정	27	6.37%	16	기타	24	5.66%

- 협력적 대화

	대화주제	수집수량	수집비율		대화주제	수집수량	수집비율
01	영화, 드라마, 음악(컨텐츠)	46	9.77%	06	패션, 뷰티	53	11.25%
02	연극, 뮤지컬, 콘서트(공연)	44	9.34%	07	음식, 음료	49	10.40%
03	전시회, 박물관(전시)	42	8.92%	08	반려동물	47	9.98%
04	책, 독서	50	10.62%	09	여행계획	44	9.34%
05	스포츠, 레저	48	10.19%	10	여행일반	48	10.19%

◎ 2023년

1	방송/영화/연예인	9	생활/주거환경
2	취미	10	가족/관혼상제
3	반려동식물	11	회사/학교생활
4	쇼핑	12	취직
5	패션/미용	13	인간관계
6	먹거리	14	경제/재테크
7	건강/다이어트	15	사회이슈
8	여행/휴가	16	기타

◎ 2024년

1	영화, 드라마, 전시회, 공연
2	먹거리, 맛집, 요리법
3	스포츠, 레저, 취미, 게임, 만화, 책, 독서
4	여행, 휴가, 휴일, 자연휴양지
5	방송 연예, 예능, 아이돌, 한류
6	건강, 다이어트, 질병
7	반려동물, 반려용품
8	우정, 성격, MBTI
9	쇼핑, 선물
10	경제, 재테크, 부동산, 금융
11	연애 결혼, 가족, 관혼상제
12	생활, 주거 환경
13	취직, 아르바이트
14	회사, 학교, 학창시절
15	스마트기기, 인공지능, 메타버스 등 IT 관련 주제

매년 다루었던 주제들의 변화 추이를 정리해 보면 다음과 같다.

<표 2> 주제의 연도별 변화 추이

주제	19	20	21	22	23	24
01 군대	■					
02 게임	■					■
03 휴일	■					■
04 자동차	■					
05 만화	■					■
06 영화	■	■				■
07 정치	■					
08 건강, 다이어트	■	■	■	■		■
09 방송연예	■	■	■	■	■	■
10 스포츠, 레저	■	■	■			■
11 먹거리	■	■	■	■	■	■
12 자연휴양지	■					■
13 국가지역	■					
14 문학	■					■
15 연애, 결혼	■	■				■
16 경제재테크	■		■	■	■	■
17 여행지(국내, 국외)		■			■	
18 계절, 날씨		■				
19 회사, 학교		■	■	■	■	■
20 선물		■				■
21 꿈(목표)		■				
22 반려동물, 반려 용품		■	■	■	■	■
23 아르바이트		■				■

주제	19	20	21	22	23	24
24 성격, MBTI		■				■
25 가족		■	■			■
26 휴가, 여행지			■	■	■	■
27 대중교통			■	■		
28 음악			■	■		
29 우정			■	■		■
30 취직			■	■	■	■
31 쇼핑			■	■	■	■
32 관혼상제			■		■	■
33 스포츠, 레저, 취미				■		■
34 가족, 관혼상제			■	■		■
35 생활, 주거환경				■	■	■
36 패션, 미용					■	
37 인간관계					■	
38 사회이슈					■	
39 **드라마, 전시회, 공연**						■
40 맛집, 요리법						■
41 예능, 아이돌, 한류						■
42 부동산, 금융						■
43 기타				■	■	

2025년 일상 대화 말뭉치 구축 사업의 대화 주제는 기수집된 주제의 대화량을 늘려 전체적인 대화의 변화가 파악될 수 있도록 하면서 주제의 다양화를 위해 지금까지의 수집 결과를 참조하여 10개의 주제를 선정하였다. 대화 참여자의 조합(성별, 연령별, 친소 관계 등)을 다양하게 구성함으로써 동일 주제 내에서도 최대한 다양한 대화의 형태가 포함될 수 있도록 하였다.

* 2025년 일상 대화 주제 선정 방향
 - 기수집된 대화 주제를 바탕으로 시간의 흐름에 따른 대화의 변화가 파악될 수 있도록 주제 선정
 - 대화 참여자의 조합을 다양하게 구성, 주제 내에서 다양한 대화의 형태가 포함될 수 있도록 유도
 - 고령화, 기후 변화 등 최신 트렌드를 반영한 키워드 추가

이상과 같은 기준에 따라 주관기관과 협의를 거쳐 기존의 주제들 중에 지속적으로 추가하여야 할 주제들과 새롭게 추가한 주제들로 다음과 같이 10개의 주제가 선정되었다.

<표 3> 2025년 선정된 주제 - 일상 대화

1	건강
2	문화예술
3	음식
4	경제
5	회사, 학교, 학창시절
6	반려동물, 반려용품
7	여행, 휴가, 휴양지, 자연휴양지
8	쇼핑
9	**새로운 기술과 우리 생활**
10	**새로운 변화와 우리 생활**

※ 신규는 굵은 글씨

2025년에는 다자 대화와 함께 1인 발화(공적 독백)를 100시간 수집하도록 되어 있다. 여기서 공적 독백이란 유튜브의 게임 채널, 당구 채널처럼 공개된 공간 또는 수단을 통해 다중과의 대화가 상호 간의 음성으로 이루어지지 않고 진행자가 혼자 말하고 시청자는 문자 등으로 일부 소통이 이루어지거나 없는 형태로 정의하였다. 공적 독백 구축 대상의 수집은 뷰티, 패션, 게임, 스포츠, 건강, 취미, 여가 등 유튜브에 게시된 다양한 분야의 콘텐츠 수집과 강연, 강의, 연설, 발표를 목표로 하였으며, 콘텐츠 당 각 18분을 넘지 않도록 하였다. 아울러 너무 특정 주제에 편중되지 않도록 유도하였다.

<표 4> 2025년 선정된 주제 - 공적 독백

1	**뷰티, 패션**
2	게임
3	스포츠, 건강
4	취미, 여가
5	푸드, 쿠킹
6	IT, 기술, 과학
7	동물, 펫
8	교육, 강의
9	**강연**
10	**강의**
11	**연설**
12	**발표**

※ 신규는 굵은 글씨

1.2. 대화 주제 제시 자료

다자 대화의 경우 선택한 주제를 중심으로 대화를 폭넓게 유도하기 위하여 다음과 같은 참고 자료들을 제시하였다.

<표 5> 2025년 주제 제시안 참고 자료 : 10개 일상 대화

번호	2025년	세부 주제	참고 대화 주제
1	건강	건강 보조제, 다이어트 식단, 질병 관련 경험이나 증상 등	• 건강을 위해 드시는 건강보조식품이 있나요? • 효과를 보신 건강 보조식품은 무엇이 있나요? • 건강검진을 받아본 경험이 있나요? • 건강검진 받은 경험을 이야기해 주세요. • 자주 이용하시는 병원이 있나요? 있다면 이유는 무엇인가요? • 병원에 입원했던 경험이 있나요? • 다이어트를 해보신 경험이 있습니까? • 알고 있는 다이어트 방법은 무엇이 있나요? • 본인이나 주변에서 질병으로 아픈 경험이 있습니까?
2	문화예술	방송/연예/문화예술/아이돌/한류 /영화/TV 프로그램 등	• 요즘 예능과 옛날 예능의 차이점은 어떤 것이 있을까요? • 옛날 예능 중 좋아하는 프로그램이 있나요? • 해외로 진출하는 아이돌에 대해 어떻게 생각하시나요? • 좋아하는 아이돌이 있나요? 있다면 누구인가요? • 좋아하는 이유는 무엇입니까? • 즐겨보는 예능 프로그램은 무엇인가요? • 기억에 남는 오디션 프로그램은 무엇인가요? • 해외에서 진행되는 예능 프로그램에 대해 어떻게 생각하십니까?
3	음식	먹거리, 맛집, 요리법 등	• 가장 좋아하는 음식은 무엇인가요? • 자신 있게 만들 수 있는 음식은 무엇인가요? • 자주 만드는 음식의 레시피는 어떻게 되나요? • 가장 맛있게 먹었던 음식을 추천해 주세요. • 맛집 투어를 좋아하시나요? • 밀키트와 같은 간편식을 알고 계신가요? • 맛집은 대체로 어떻게 찾고 계신가요? • 맛집에 갔다가 실망했던 경험이 있나요? • 오늘 저녁 메뉴로 생각하고 계신 음식이 있나요? • 추천하고 싶은 음식점이 있나요?
4	경제	재테크, 부동산, 금융, 주식 등	• 현재 거주 중인 곳의 장단점에 대해서 이야기해 주세요. • 최근 신축 아파트들의 부실 시공이 논란이 되고 있는데 어떻게 생각하시나요? • 아파트 입주 전 불량을 찾아주는 업체도 생겨났다는데

			어떻게 보시나요? • 예금이나 주식 배당 중 어떤 것이 더 좋으신가요? • 현재 경제 상황은 어떻다고 느끼십니까? • 재테크를 위해 어떤 것을 하고 계십니까? • 적금과 투자 중 어느 쪽에 비중을 두고 계십니까? • 지금 투자하고 계신 종목은 무엇인가요? • 카카오뱅크와 같은 비대면 금융사의 장점은 무엇인가요? • 시중 은행이 점점 줄어 들고 있는데 어떻게 생각하시나요?
5	회사/학교/ 학창시절	학교, 학창시절, 직장생활, 취직, 아르바이트 등	• 현재 직장 또는 학교 생활에 만족하시나요? • 학교 다니면서 좋았던 기억은 무엇인가요? • 학창시절 소풍, 수학여행 등의 경험에 대해 이야기해 주세요. • 좋아했던 과목과 싫어했던 과목에 대해 이야기해 주세요. • 사교육 관련 학원이나 과외 경험이 있으신가요? • 학창시절 좋아하던 선생님이 있었나요? • 앞으로 유망할 것으로 예상되는 직업은 무엇인가요? • 준비하고 있는 진로는 어떤 이유에서 선택하게 되었습니까? • 회사, 학교 생활을 하면서 자랑할만한 경험이 있나요? • 경험했던 아르바이트의 종류에 대해 이야기해 주세요. • 취업을 하기 위해 했던 노력에 대해 말씀해 주세요. • 취업을 준비하면서 가장 힘들었던 점은 무엇인가요? • 회사 복지제도에 추가되었으면 하는 내용이 무엇인가요? • 사회적으로 물의를 일으키는 기업들을 어떻게 생각하세요? • 재택근무에 대해서 어떻게 생각하십니까? • 현재 하고 계신 직업에 대해 간단히 소개해 주세요. • 직장생활을 하면서 가장 힘든 점은 무엇인가요?
6	반려동물/ 반려용품	반려동물 경험 및 팁, 유용한 반려용품 추천 등	• 현재 키우고 있는 반려동물(식물)이 있습니까? • 현재 키우고 있는 반려동물에 대해 소개해 주세요. • 반려동물(식물)을 키우게 된 계기가 있나요? • 반려동물 용품 중 추천하고 싶은 제품이 있나요? • 반려동물을 키우면 좋은 점은 무엇입니까? • 반려동물을 키울 때 주의해야 할 점은 무엇인가요? • 반려동물을 키우면서 기억에 남는 에피소드가 있나요? • 기억에 남는 반려동물이 있나요? • 동물병원에 갔던 경험이 있나요? • 반려동물과 외출 시 불편한 점에 대해 말씀해 주세요. • 반려동물 입양 경험에 대해 이야기해 주세요. • 반려동물이 있어 좋은 점과 불편한 점에 대해 말씀해 주세요. • 키워보고 싶은 반려동물이 있나요? 그 이유는 무엇인가요?

7	여행/휴가/ 휴일/자연 휴양지	여름 휴가, 겨울 휴가, 국내/해외 여행 경험, 산/바다 자연 휴양지 경험, 여행 계획 등	• 다녔던 여행지 중 가장 기억에 남는 여행지는 어디인가요? • 여행하면서 기억에 남는 에피소드에 대해 말씀해 주세요. • 올해 여름 휴가 계획을 세우셨나요? • 겨울에 휴가 다녀오신 경험이 있다면 말씀해 주세요. • 여행하면서 기분이 안 좋았던 경험이 있으신가요? • 여행할 때 숙소의 선택 기준은 무엇인가요? • 여행할 때 주로 이용하는 숙박 유형은 어떻게 되나요? • 주로 이용하는 숙박 유형을 선택하는 이유는 무엇인가요? • 패키지 여행과 자유여행 중 어느 쪽을 선호하시나요? • 국내 여행은 어디로 가고 싶으신가요? • 가고 싶은 해외 여행지가 있나요? • 추천하고 싶은 여행지가 있다면 이야기해 주세요. • 산과 바다 중 어느 곳을 선호하시나요? • 국내 여행 중 가장 만족했던 여행지는 어디인가요?
8	쇼핑	선호 브랜드, 선호 쇼핑몰, 쇼핑 방식, 중고 거래, 기억에 남는 선물 등	• 구매하는 물품의 브랜드를 중요하게 생각하시나요? • 온라인, 오프라인 구매의 장단점은 무엇인가요? • 마트, 백화점, 아울렛 중 자주 이용하시는 곳은 어디인가요? • 최근 구매하신 물품은 무엇인가요? • 쇼핑할 때 가장 먼저 고려하는 점은 무엇인가요? • 해외 직구를 해보신 경험이 있습니까? • 홈쇼핑 구매에 장단점은 무엇입니까? • 중고거래 경험이 있나요? • 중고거래 시 주의해야 할 사항에 대해 알려주세요.
9	새로운 기술과 우리 생활	인공지능, 블록체인, 자율주행차, 드론, 가상현실과 메타버스 등	• 스마트기기로 분류되는 전자기기는 무엇이 있을까요? • 스마트폰 이용하기 전과 후의 삶이 어떻게 바뀌었나요? • 스마트워치를 사용하고 계신가요? 사용하신다면 어떤 부분의 기능을 많이 사용하시나요? • AI 스피커 사용 시 장점은 무엇인가요? • 스마트기기들의 단점은 무엇이 있습니까? • ChatGPT에 대해서 알고 계신가요? • 인공지능의 장점은 무엇이 있을까요? • VR로 불리는 가상세계 경험을 하신 적이 있습니까? • 증강현실(AR)을 활용한 게임을 해보신 적이 있습니까? • 이케아 플레이스라는 어플을 사용해 보신 적이 있습니까? • 초고속 인터넷, 5G 기술의 발전은 어떻게 생각하십니까? • 지문인식, 안면인식 같은 생체인증은 어떻게 생각하십니까? • 자율주행의 경험이 있으신가요? 있다면 장단점에 대해서 이야기해 주세요. • 드론을 활용하여 촬영을 해본 경험이 있습니까? • 인공지능의 발전으로 우리 삶에 우려되는 사항은 뭘까요?

10	사회적 변화와 우리 생활	기후변화, 고령화 사회, 사회적 거리두기, 성평등, 청년실업 등	• 기후 변화가 우리의 일상에 어떤 영향을 미치고 있나요? • 탄소 배출을 줄이기 위한 방법은 무엇이 있을까요? • 일상에서 실천할 수 있는 환경 보호 방법은 무엇이 있나요? • 고령화 사회에서의 가장 큰 문제는 무엇일까요? • 노인 복지를 강화하려면 어떤 변화가 필요할까요? • 고령화 사회에서 일하는 방법은 어떻게 달라질까요? • 고령자들이 더 행복하게 살 수 있는 방법은 무엇일까요? • 사회적 거리두기가 경제에 미친 영향은 어떤 것이 있을까요? • 거리두기 후, 우리가 느낀 변화는 무엇인가요? • 위기 상황에서 우리가 할 수 있는 사회적 거리두기 외의 방법은 무엇이 있을까요? • 직장에서 성평등을 실현하려면 어떤 노력이 필요할까요? • 성평등이 이루어진 사회는 어떤 모습일까요? • 성별 역할 고정관념을 바꾸려면 무엇이 필요할까요? • 청년실업 문제, 어떤 방식으로 해결할 수 있을까요? • 청년들에게 어떤 직업 교육이 필요할까요? • 청년실업을 줄이기 위한 정책은 어떻게 바뀌어야 할까요? • 청년들이 직장을 찾는 데 어려운 이유는 무엇일까요? • 청년 창업을 촉진하려면 어떤 지원이 필요할까요?

[그림 3] 추가 제시 자료 - 한국의 전통 음식 162선 사진

<음식 - 한국의 전통 음식 사진 자료>

<녹음장에 비치하여 참고>

2. 전문가 자문회의 진행

사업의 원활한 수행과 체계적인 말뭉치 구축을 위해 국어국문학, 음성 언어학 및 음성정보기술(음성인식 및 합성) 분야 전문가로 자문단을 구성하고 정기 자문 회의를 두 차례 진행하였다.

특히, 두 번째 자문회의는 국립국어원과의 협의를 거쳐 지금까지 7년 동안 구축해 온 일상대화 말뭉치의 활용 및 발전 방향의 모색을 위하여 산학연 전문가로 구성된 확대 자문회의로 진행되었다. 그 결과는 다음 표와 같다.

<표 6> 1차 정례 자문회의

일시	2025년 4월 18일(금) 17시~19시
장소	나라지식정보 회의실
참석자	○ 자문위원 정민화 (서울대 언어학과 교수) 음성정보처리 (음성인식) 이호영 (서울대 언어학과 교수) 언어학(음성학) 이석재 (연세대 영어영문학과 교수) 코퍼스언어학, 음성학 도재학 (경기대 국어국문학과 조교수) 국어학 ○ 과제수행자 이용주 연구위원(나라지식정보, PM) 차원철 부장(나라지식정보, PL) ○ 수행기관 책임자 박승희 부사장(나라지식정보)
자문내용	- 24년부터 억양구에서 문장으로 분절 단위가 바뀜에 따라 다소 애매했던 분절 기준의 일부 표현을 명확하게 할 필요가 있음. - 사용자 입장의 산업계 대표 중심의 자문회의 의견 청취 기회 마련. - 자문위원 모두 공적 대화 강연 자료 수집에 참여

<표 7> 2차 산학연 확대 자문회의

안건	일상대화 말뭉치 발전 및 개선을 위한 산학연 전문가 자문
일시	2025년 12월 5일(금) 15시 ~ 17시
형식	산학연 전문가 대면 자문회의
참석자	▶ 자문위원 학계: 도재학 교수(경기대), 김지환 교수(서강대), 이석재 교수(연세대:서면 제출) 산업계: 이상준 이사(인공지능협회), 김훈 팀장(카카오), 윤재선 상무(셀바스AI), 임영현 수석(페르소나AI) 연구소: 김상훈 박사(한국전자통신연구원) ▶ 국립국어원 이현주 언어정보과장, 박미영 학예연구관, 장연지 연구원 ▶ 나라지식정보: 박승희(부사장), 이용주(PM), 박분선(PL)
자문내용	2019년 이래 7년동안 지속적으로 구축해온 일상대화 말뭉치에 관하여 현황을 소개하고 이에 대하여 각 자문위원들의 의견과의 토의 내용을 다음과 같이 정리함. 가. 데이터 전략 - 자연스러운 대화의 양적 확대 필요 - 활용 목적에 따른 품질 레벨 구분을 병행하는 방향 검토 필요 나. 대화 유형 및 수집 환경 다변화 - 전화·콜센터 등 비대면 대화 - 줌/온라인 회의 - AI·챗봇과의 상호작용 등 새로운 상호작용 환경 다. 도메인 및 주제 설계 - 프리토킹·잡담형 대화 - 실생활에서 빈도가 높은 질의·응답형(문의-응대) 대화 - 전문 용어가 많이 등장하는 특정 도메인 대화(의료, 약국, 금융 등) 라. 전사 기준·가이드라인 정비 - 발음 전사와 철자 전사 기준, 방언·개인 발음 처리 원칙의 지속적인 연구 - STT·LLM 학습 목적에 맞게 활용 가능한 형태로 정비 필요 마. 피드백·유통 체계 - 활용 기관의 오류 신고·피드백을 수용할 수 있는 다양한 채널을 마련 - 지속적으로 데이터를 보정·업데이트하는 방안 검토

3. 화자 구성 및 모집

3.1. 화자의 구성

2~4인 대화의 성별, 연령별, 지역별 화자 비율은 2024년 통계청의 인구 분포 자료를 참고하여 화자 모집 기준을 다음과 같이 수립하였고 다자 대화(정제 후 550시간)의 구축을 위하여 실제 총 녹음 시간은 660시간으로 산정하여 화자를 모집하였다.

<표 8> 2025년 일상 대화(2~4인 대화) 화자 모집 목표(단위: 시간)

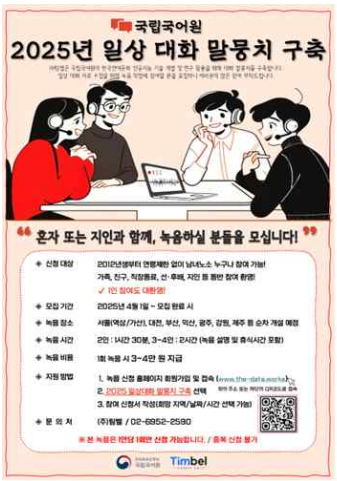
	연령비								목표 시간		
	인구통계		10대	20대	30대	40대	50대	60대~			
	인구 (단위 : 천명)	비율(%)	15%	15%	15%	20%	15%	20%	100%		
계	51,313	100.00%	82.5	82.5	82.5	110	82.5	110	660		
서울	9,394	18.29%	15.00	15.00	15.00	20.00	15.00	20.00	120.00	18.18%	
인천	3,049	5.94%	4.95	4.95	4.95	6.60	4.95	6.60	39.60	6.00%	
경기	13,861	26.99%	19.50	19.50	19.50	26.00	19.50	26.00	156.00	23.64%	
강원	1,518	2.96%	3.00	3.00	3.00	4.00	3.00	4.00	24.00	3.64%	
부산	3,265	6.36%	5.25	5.25	5.25	7.00	5.25	7.00	42.00	6.36%	
대구	2,354	4.58%	3.75	3.75	3.75	5.00	3.75	5.00	30.00	4.55%	
울산	1,103	2.15%	1.80	1.80	1.80	2.40	1.80	2.40	14.40	2.18%	
경북	2,597	5.06%	4.05	4.05	4.05	5.40	4.05	5.40	32.40	4.91%	
경남	3,249	6.33%	5.25	5.25	5.25	7.00	5.25	7.00	42.00	6.36%	
대전	1,473	2.87%	2.40	2.40	2.40	3.20	2.40	3.20	19.20	2.91%	
충북	1,630	3.17%	2.55	2.55	2.55	3.40	2.55	3.40	20.40	3.09%	
충남	2,224	4.33%	4.05	4.05	4.05	5.40	4.05	5.40	32.40	4.91%	
광주	1,456	2.83%	2.25	2.25	2.25	3.00	2.25	3.00	18.00	2.73%	
전북	1,759	3.42%	2.85	2.85	2.85	3.80	2.85	3.80	22.80	3.45%	
전남	1,757	3.42%	2.85	2.85	2.85	3.80	2.85	3.80	22.80	3.45%	
제주	675	1.31%	3.00	3.00	3.00	4.00	3.00	4.00	24.00	3.64%	

2025년도에 수집하는 1인 발화는 공적 독백(100시간)인데, 유튜브 60시간, 강연 등 40시간 분량을 목표로 하였고 성별·연령별·지역별 구성은 고려하지 않았다.

3.2. 화자의 모집

3.2.1. 녹음 참여자 모집

다자 대화 및 1인 공적 독백에 필요한 화자의 원활한 모집을 위해 먼저 구인 사이트, 온라인커뮤니티, 누리소통망(SNS), 지역 카페 등 온·오프라인 홍보활동을 강화하였다. 특히, 아르바이트 앱과 구인 사이트 또는 맘 카페 등 다양한 경로로 2인 이상 짝을 이루어 화자를 모집하였다. 모집된 화자들은 사전에 전화 통화로 작업 스케줄을 조정하고, 해당 일시에 녹음 사이트로 내방하여 녹음을 진행하였다. 녹음에 참여하는 모든 화자는 '개인 정보 활용 동의서'를 작성하여 데이터의 활용에 대한 근거를 확보하였다.



〈홍보 포스터〉

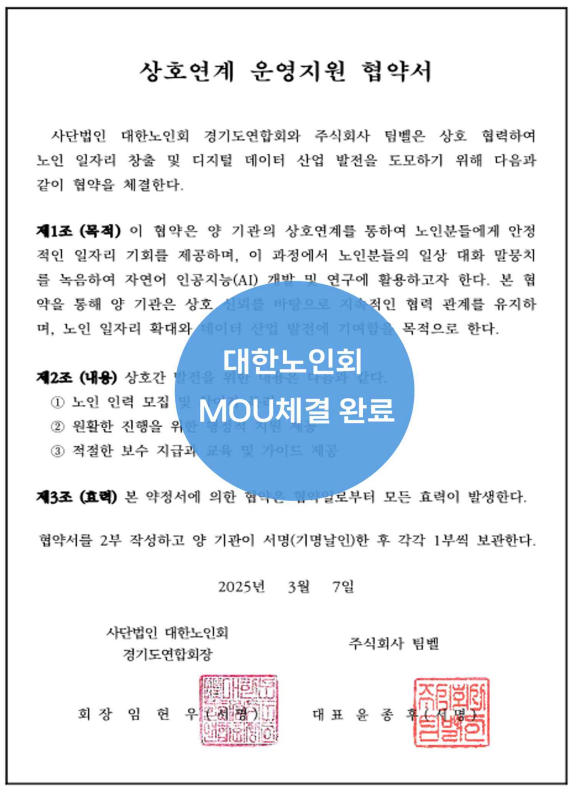
〈온라인 홍보(당근 마켓)〉

〈온라인 홍보(지역 카페)〉

〈온라인 홍보(전문 사이트)〉

[그림 4] 홍보 예시

아울러, 인터넷 활용도가 상대적으로 낮은 50대 이상 고령자 및 모집 취약 지역에서의 발화자 확보를 위하여 사단법인 대한노인회(경기도연합회)와 양해각서(MOU)를 체결하였으며, 이를 기반으로 서울연합회 및 전국 지방지회와의 협력 체계를 구축하여 안정적인 참여자 모집을 진행하였다.



상호연계 운영지원 협약서

사단법인 대한노인회 경기도연합회와 주식회사 팀벨은 상호 협력하여 노인 일자리 창출 및 디지털 데이터 산업 발전을 도모하기 위해 다음과 같이 협약을 체결한다.

제1조 (목적) 이 협약은 양 기관의 상호연계를 통하여 노인분들에게 안정적인 일자리 기회를 제공하며, 이 과정에서 노인분들의 일상 대화 말뭉치를 녹음하여 자연어 인공지능(AI) 개발 및 연구에 활용하고자 한다. 본 협약을 통해 양 기관은 상호 [illegible]를 바탕으로 지속적인 협력 관계를 유지하며, 노인 일자리 확대와 [illegible] 산업 발전에 기여할 [illegible] 목적으로 한다.

제2조 (내용) 상호간 [illegible]을 위한 내용은 다음과 같다.
① 노인 인력 모집 [illegible]
② 원활한 진행을 위[illegible]
③ 적절한 보수 지급과 [illegible] 교육 및 가이드 제공

제3조 (효력) 본 약정서에 의한 협력은 [illegible]로부터 모든 효력이 발생한다.

협약서를 2부 작성하고 양 기관이 서명(기명날인)한 후 각각 1부씩 보관한다.

2025년 3월 7일

사단법인 대한노인회
경기도연합회장 회 장 임 현 우 (서명)

주식회사 팀벨
대 표 윤 종 [illegible] (서명)

대한노인회
MOU체결 완료

[그림 5] 기관 협력을 위한 양해각서의 예

3.2.2. 공적 독백 참여자 모집

1인 공적 독백에 필요한 콘텐츠 확보는 유튜브 크리에이터와의 개별 접촉으로 콘텐츠 사용 허락 동의를 받은 후 해당 콘텐츠의 음성을 추출하는 방식으로 진행을 하였다. 아래와 같은 절차를 거쳐 223개 채널의 크리에이터에게 참여 여부를 확인하였고 61개 채널이 본 사업에 참여하였다.

① 채널 목록 정리 : 주제별 추천 채널을 검색하여 목록 정리
② 채널 선정
  - 채널별로 게시된 콘텐츠를 무작위로 확인하여 본 사업 대상 조건에 적합한지 확인하여 채널 선정
③ 콘텐츠 목록 작성 : 1인 발화 조건, 시간 조건 등을 고려하여 대상 콘텐츠 목록 정리
④ 참여 여부 확인(e-mail)
  - 본 사업과 참여 절차를 소개하고 활용할 콘텐츠 목록을 제시하며 참여 여부 확인
⑤ 저작권 이용 허락 계약 체결
  - 전자계약 서비스를 활용하여 계약 체결
⑥ 해당 콘텐츠 음성 추출

3.3. 저작권 이용 허락 계약 체결

저작권 이용 허락 계약 체결은 전자계약 서비스를 활용하여 진행하였다. 본 사업 참여 확인 및 개인 정보 동의와 관련된 양식을 국립국어원으로부터 전달받아 법적 검토 후 활용하였다. 계약 체결은 녹음 참여자의 경우 녹음이 진행된 현장에서 스마트 기기를 사용해 동의를 구하였으며, 공적 독백 참여자(유튜브 크리에이터)의 경우 전자 결재를 할 수 있도록 이메일로 링크를 전송하는 방식으로 진행하였다.

[그림 6] 전자 결재 사이트 및 동의서 산출물

4. 작업자 선발 및 교육

4.1. 녹음 진행 요원 선발 및 교육

녹음은 서울(역삼/가산), 대구, 부산, 광주, 청주, 제주, 원주 등 7개 지역에서 병렬로 진행하였다. 7개의 지역 거주자 1,706명이 화자로 녹음에 참여하였으며, 다수의 화자가 참여하는 만큼 원활한 녹음 진행을 위하여 각 지역별로 녹음 진행 요원이 투입되었다. 지역별로 투입된 진행 요원은 15명으로 면담을 진행한 후 자격 조건과 맞는 지원자를 대상으로 오프라인 교육을 실시하였으며, 교육 과정 중 평가에 통과한 사람을 최종 선발하였다. 진행 요원은 기존 유사 작업 경험자를 우선하여 선발하였다.

<표 9> 진행 요원 선발 및 운영 방안

구분	선발 기준 및 운영 내용	
선발 기준	• 전문 녹음 장비 작동 경험이 있는 사람 (우선 선발) • 최종 교육 이수 및 평가 통과자 (필수)	
투입 인원	진행 요원 15명	
진행 요원 역할	• 진행 요원 1 - 화자 안내 - 화자 인적 사항 확인 - 화자 참석 관리 및 스케줄 관리 - 사례비 지급 정보 확인	• 진행 요원 2 - 녹음 진행 개요 설명 - 저작권 이용 허락 계약 체결 및 개인 정보 동의서 관리 - 녹음 장비 이상 유무 확인 - 녹음 진행

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

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

기본 교육을 마친 진행 요원은 실제 녹음으로 들어가기에 앞서 화자 응대, 녹음 장비 작동에 대한 시뮬레이션과 특이 사항 발생 시 대처 요령에 대한 모의 훈련을 실시하였다. 모의 평가에서 역할에 대한 이해도가 높은 사람을 최종적으로 선발하였으며 선발된 녹음 진행 요원들은 보안 서약서 작성 후 실제 녹음 진행에 참여하였다.

<표 10> 진행 요원 교육 내용

구분	내용
교육 일시 및 장소	• 2025년 4월, 서울(역삼/가산), 대구, 부산 녹음 지역 • 2025년 5월, 광주, 청주 녹음 지역 • 2025년 9월, 제주, 강원(원주) 녹음 지역
교육자	신은주(㈜팀벨)
교육 내용	• 사업의 배경 및 목적 • 진행 절차 • 대화 주제 • 녹음 환경 및 녹음 장비 사용법 • 녹음 방법 • 녹음 시 주의 사항 및 녹음 진행 시 제스처 학습 • 녹음 시뮬레이션 실습 • 보안 교육 • 질의응답

[그림 7] 녹음 요원 교육 자료 예

4.2. 전사 작업자 선발 및 교육

2025년 일상 대화 말뭉치 전사 작업은 선행사업(2021~2024년 일상 대화 말뭉치 구축) 전사 데이터 구축에 참여한 작업자 중 작업효율과 정확도가 우수한 속기사 16명이 투입되었다. 전사 작업을 위한 교육은 전사지침 및 음성 분절 교육을 시작으로 격월로 3차례 온·오프라인 교육하였으며, 2주에 한 차례씩 작업자 전체 미팅을 통해 오류 사례를 공유하고, 변화된 지침에 대한 재교육을 실시하였다.

크라우드워커에 의한 작업보다는 유사사업 경험이 풍부하며, 집중 가능한 소수의 속기사를 통한 운영과 정기적인 교육을 통해 말뭉치 데이터의 품질을 높였다.

<표 11> 전사자 교육 일정 및 내용

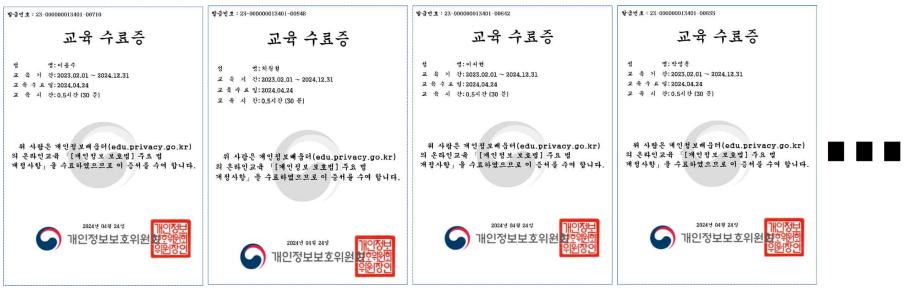
작업자 전체 교육	◆ 일시 : 2025년 5월 셋째 주부터 격월 교육 ◆ 장소 : 나라지식정보 회의실 혹은 온라인 ◆ 교육 참석인원 : 전사 작업자 및 PM, 전사 분야 사업 참여 인력 전원 <교육 내용> 1. 사업의 배경 및 목적, 전사 절차와 방법 2. 전사 지침 및 유의 사항 3. 전사 도구(음성 전사 워크벤치 시스템) 사용 교육 4. 한글 맞춤법 주요 내용 및 오류 사례 5. 질의응답
신규 작업자 교육	◆ 일시 : 3명 이상의 작업자 신규 투입 시 온·오프라인 병행 ◆ 교육 참석인원 : 전사 작업 신규참여자 및 PM, 전사 분야 사업 참여 인력 <교육 내용> 1. 사업의 배경 및 목적, 전사 절차와 방법 2. 전사 지침 및 유의 사항 3. 전사 도구(음성 전사 워크벤치 시스템) 사용 교육 4. 질의응답
전사작업자 정기 미팅	◆ 일시 : 매월 첫 번째, 세 번째 금요일 오후 5시(온라인) ◆ 회의 참석자 : 전사 작업자 및 PM, 전사 분야 사업 참여 인력 전원 <교육 내용> 1. 변경된 전사 지침 공유 2. 전사 관련 이슈사항 점검 3. 전사 오류 사례 검토 및 공유 4. 전사 규칙 관련 질의 응답

4.3. 개인 정보 보호 및 보안 교육

이 사업이 정보화 용역사업으로 편성됨에 따라, 사업의 원활한 진행을 위해 정보보안 교육과 개인 정보보호 교육을 시행하였다. 먼저, 정보보안 교육의 경우 본 사업 참여 인력 전원을 대상으로 하였으며, 온라인으로 자체 교육을 진행하였다.

개인 정보 보호 교육은 또한 참여 인력 전원을 대상으로 진행하였으며, 개인정보보호위원회가 운영하는 개인정보배움터 포털(edu.privacy.go.kr)의 온라인 교육인 '[개인정보 보호법] 주요 법 개정사항'을 개별적으로 수강하고 교육 수료증을 발급받았다.

<표 12> 개인 정보 보호 및 보안 관련 교육

구분	내용
보안 교육	• 교육일 : 2025년 4월 19일 13:00~14:00 (온라인 교육) • 참여자 : 사업 참여자(13명) • 내용 : 사업 수행을 위한 보안 정책 및 지침 확인, 정보보안, PC 보안, 개인 정보 취급 방법, 문서 및 자료 관리, 사무실 및 장비 관리 등 • 참고 자료 : 국가사이버안전관리규정, 문화체육관광부 개인정보 보호지침, 문화체육관광부 보안업무규정 시행세칙, 문화체육관광부 정보화업무 규정, 행정안전부(KISA) 개인정보 보호법_주요내용교육자료 등  
개인 정보 보호 교육	• 개별적 온라인 교육 수강(edu.privacy.go.kr) • 참여자: 개인 정보 직접 수집·처리 기관 참여 인력(4명) • 내용: [개인정보 보호법] 주요 법 개정사항 

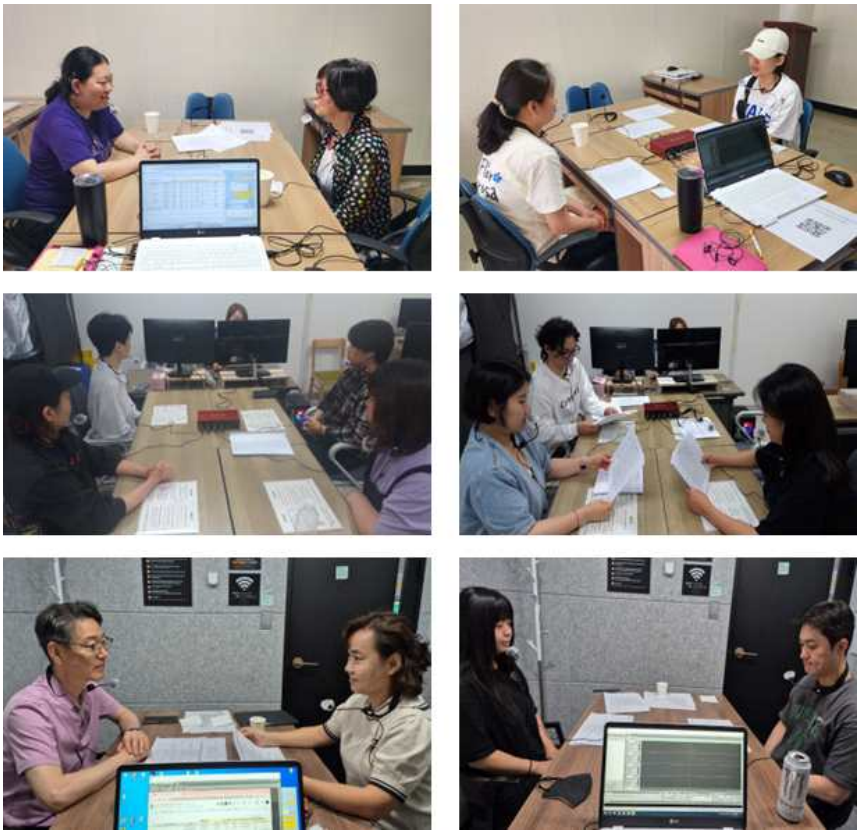
5. 음성 녹음

5.1. 녹음 환경

녹음은 전국 7개 지역 서울(역삼/가산), 대구, 부산, 광주, 청주, 제주, 원주)에 녹음 공간을 마련하고 배분된 화자 비율에 따라 최소 1개월에서 7개월까지 화자를 모집하여 수집하였다.

녹음 공간은 일상 대화 환경을 상정하여 두 명 이상의 화자가 자유롭게 대화할 수 있는 환경(예: 방음벽이 있는 회의실)을 구축하였고 최대 4인의 화자 간 자연스러운 대화를 유도하고, 대화 진행 상황을 전문 교육을 받은 오퍼레이터가 모니터링하여, 최상의 데이터를 수집할 수 있도록 녹음 환경을 구축하였다.

[그림 8] 2인 및 4인 대화의 스튜디오 및 장비 세팅 예시



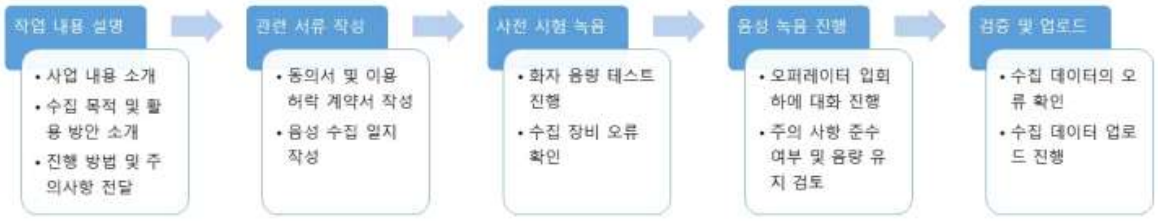
[그림 9] 각 녹음 스튜디오의 음성 녹음 모습

음성 녹음에 사용되는 장비는 최대 4인의 화자 간 대화를 녹음하기 위한 멀티 채널을 지원하는 오디오 인터페이스를 활용하였다. 또한, 다자간 대화 시 주변 노이즈를 최소화할 수 있는 클로즈톡 마이크로폰과 1인 발화를 위한 방송용 무선 외장마이크를 사용하였다. 모든 장비는 테스트 녹음 및 주관기관의 검토를 거쳐 수행하였다. 음성 데이터의 수집 도구의 상세한 내용은 다음과 같다.

[그림 10] 음성 데이터 수집 도구(하드웨어)

5.2. 음성 녹음 절차

모집된 화자들이 자신들이 예약한 시간에 수집 장소에 도착하면 다음과 같은 절차로 녹음이 진행된다. 각 단계에 대한 상세 내용은 아래와 같다.



[그림 11] 음성 데이터의 녹음 절차

5.2.1. 작업 내용 설명

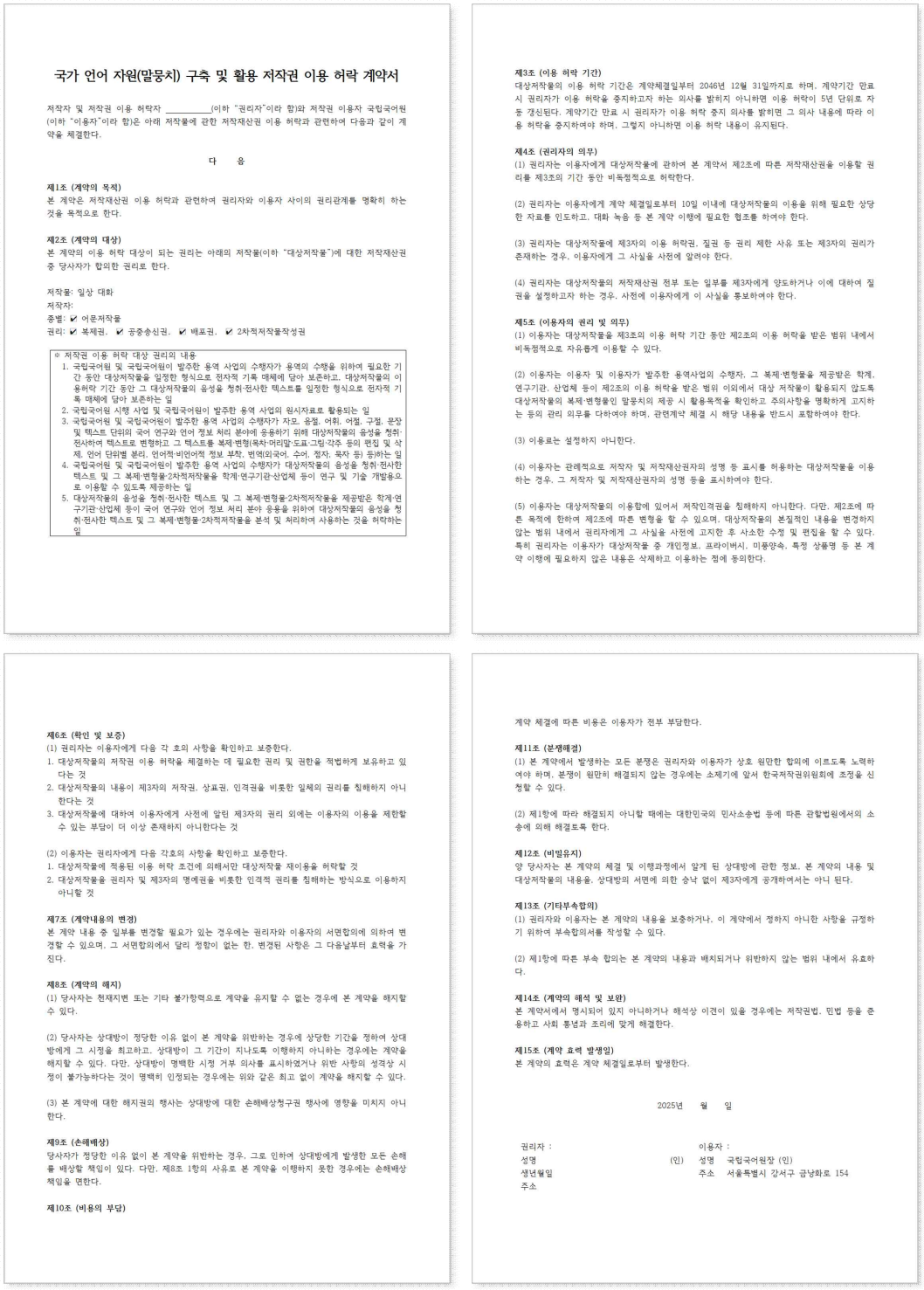
화자 모집 시 간략하게 설명된 사업의 내용 및 데이터 수집 목적, 활용 방안에 대해 화자들에게 자세한 설명을 진행하고, 음성 데이터를 수집하는 진행 과정과 장비 착용 방법, 실제 수집 진행 시 화자가 주의해야 할 내용들을 충분히 설명하였다.

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

5.2.2. 관련 서류 작성

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

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



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 용역의 수행을 위하여 필요한 기간 동안 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하고, 대상저작물의 이용허락 기간 동안 그 대상저작물의 음성을 청취·전사한 텍스트를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 시행 사업 및 국립국어원이 발주한 용역 사업의 원시자료로 활용되는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물의 음성을 청취·전사하여 텍스트로 변형하고 그 텍스트를 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착, 번역(외국어, 수어, 점자, 목자 등) 등)하는 일
4. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물·2차적저작물을 학계·연구기관·산업체 등이 연구 및 기술 개발용으로 이용할 수 있도록 제공하는 일
5. 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물·2차적저작물을 제공받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물·2차적저작물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용 허락 기간)
대상저작물의 이용 허락 기간은 계약체결일부터 2046년 12월 31일까지로 하며, 계약기간 만료 시 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 계약기간 만료 시 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다.

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

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하고, 대화 녹음 등 본 계약 이행에 필요한 협조를 하여야 한다.

(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) 이용자는 이용자 및 이용자가 발주한 용역사업의 수행자, 그 복제·변형물을 제공받은 학계, 연구기관, 산업체 등이 제2조의 이용 허락을 받은 범위 이외에서 대상 저작물이 활용되지 않도록 대상저작물의 복제·변형물인 말뭉치의 제공 시 활용목적을 확인하고 주의사항을 명확하게 고지하는 등의 관리 의무를 다하여야 하며, 관련계약 체결 시 해당 내용을 반드시 포함하여야 한다.

(3) 이용료는 설정하지 아니한다.

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

(5) 이용자는 대상저작물의 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 따른 목적에 한하여 제2조에 따른 변형을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 권리자에게 그 사실을 사전에 고지한 후 사소한 수정 및 편집을 할 수 있다. 특히 권리자는 이용자가 대상저작물 중 개인정보, 프라이버시, 미풍양속, 특정 상품명 등 본 계약 이행에 필요하지 않은 내용은 삭제하고 이용하는 점에 동의한다.

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

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.

2025년 월 일

권리자 :
성명 (인)
생년월일
주소

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

[그림 12] 저작권 이용 허락 계약서

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

화자ID	이름	연령	직업	성별	출생지	성장지	거주지	학력	관계	친밀도
SD2500020		30대	서비스 종사자	여성	대구	대구	대구	대재	기타	0
SD2500021		20대	무직/취업준비생	여성	대구	대구	대구	대졸	기타	0
SD2500020		30대	서비스 종사자	여성	대구	대구	대구	대재	기타	0
SD2500021		20대	무직/취업준비생	여성	대구	대구	대구	대졸	기타	0
SD2500004		20대	전문가 및 관련 종사자	여성	서울	서울	서울	대졸	기타	0
SD2500005		30대	무직/취업준비생	여성	서울	서울	서울	대졸	기타	0
SD2500006		20대	학생	여성	경기	경기	경기	대재	기타	0
SD2500007		20대	서비스 종사자	남성	대구	대구	경기	고졸	기타	0
SD2500004		20대	전문가 및 관련 종사자	여성	서울	서울	서울	대졸	기타	0
SD2500005		30대	무직/취업준비생	여성	서울	서울	서울	대졸	기타	0
SD2500006		20대	학생	여성	경기	경기	경기	대재	기타	0
SD2500007		20대	서비스 종사자	남성	대구	대구	경기	고졸	기타	0

[그림 13] 음성 자료 수집 일지

5.2.3. 사전 시험 녹음

화자들이 개인 정보 활용동의서, 저작권 이용허락계약서 작성을 완료하면 진행 요원은 실제 녹음이 진행되는 공간으로 화자를 이동시켜 자리 배치 및 수집 장비 착용을 안내해 주었다. 이후 진행 요원은 화자들이 선택한 주제로 1~3분 정도 자연스럽게 이야기를 하게 하고, 이 과정에서 화자의 목소리 크기가 충분한지, 화자의 움직임에 의해 잡음이 발생하지 않는지, 수집 장비에 문제가 없는지를 살펴보는 사전 시험 녹음을 진행하였다.

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

실제 수집 과정에서 동일한 설정으로 준비된 수집 장비라 하더라도 예기치 않은 형태의 문제로 인하여 수집 데이터에 잡음이 포함되는 경우가 있어, 이러한 사전 시험 녹음은 화자의 음량 및 발화 태도를 확인하는 것 이외에 장비를 테스트하는 목적도 있다.

5.2.4. 음성 녹음 진행

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

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

록 하였다.

녹음 레벨의 경우, 음성의 최대 샘플값은 10,000(16-bit Integer/PCM) ~ 20,000(16-bit Integer/PCM)을 권장하며, 32,767(16-bit Integer/PCM)을 넘지 않도록 마이크 볼륨을 조절하였다.

5.2.5. 검증 및 업로드

음성 녹음이 완료되면 각 지역의 담당자들은 실제 녹음된 파일을 청취하여 파일 자체에 문제는 없는지, 모니터링 과정에서 발견하지 못한 잡음은 없는지 등을 확인한 후 최종적으로 문제가 없으면 참여자들을 귀가시켰다. 만약 이때 수집 데이터에서 문제(돌발적인 외부 잡음)가 발생한 경우는 화자들의 동의를 구한 후 바로 재녹음을 진행하거나 다른 날짜로 일정을 잡아 다시 녹음하였다. 문제없이 수집이 완료된 원본 파일은 지역별 수집 지역의 진행 요원이 WAV 파일로 변환한 후 원본 파일과 WAV 파일을 지정된 경로에 등록하고, 상위 관리자에게 진행 내용 및 특이사항을 보고하였다.

6. 음성 자료 전사

6.1. 전사 규칙

전사는 [붙임1]의 '2025년 일상 대화 말뭉치 구축 지침' 중의 '전사 지침'을 적용하였다.

6.2. 전사 절차 및 작업

전사 절차는 전사 도구 기획 및 개발, 전사 인력 모집 및 지침 교육, 전사 진행의 단계로 이루어졌다. 우선 음성을 듣고 전사할 수 있도록 전사 도구를 개발하였다. 음성 재생 및 정지, 배속 설정, 음성 전사, 검수 등의 기능을 사용할 수 있으며, 여러 명의 작업자들이 동시에 전사 작업을 진행하는 것에 문제가 없도록 개발하였다. 먼저 관리자가 전사 도구에 음원 파일을 업로드하면 시스템에서 자동 STT(Speech-to-Text)를 수행하여 전사본을 생성하고, 동시에 발화 구간을 기준으로 문장 단위 데이터 싱크 분할(클립 생성)을 자동으로 처리한다. 이를 통해 초기 전사 텍스트와 문장 단위 시간 정보가 포함된 1차 결과물이 마련된다. 이후 작업자에게 데이터를 배정하고 작업자는 할당된 데이터로 전사 작업을 진행한다. 원음을 직접 청취하면서 자동 전사 결과를 전사 지침에 따라 수정·보완한다. 오인식, 탈락, 중복 표기 등을 교정하고, 구어적 특성 및 발화 특수 요소를 작업 기준에 맞게 정비한다. 아울러 문장 단위 클립의 시작 및 종료 시점을 재확인하여 음성과 텍스트 간 싱크를 정밀하게 보정한다. 작업이 완료되면 검수를 요청하고, 검수자에게 할당되면 해당 데이터를 대상으로 2차 품질 점검을 수행한다. 검수 단계에서는 음원과 문장 단위 싱크의 정확성, 작업 지침 준수 여부, 전사 내용의 정확성 및 누락·오류 여부를 종합적으로 확인한다. 전사 도구의 세부 기능은 아래 그림과 같다.

[그림 14] 전사 도구 세부 기능

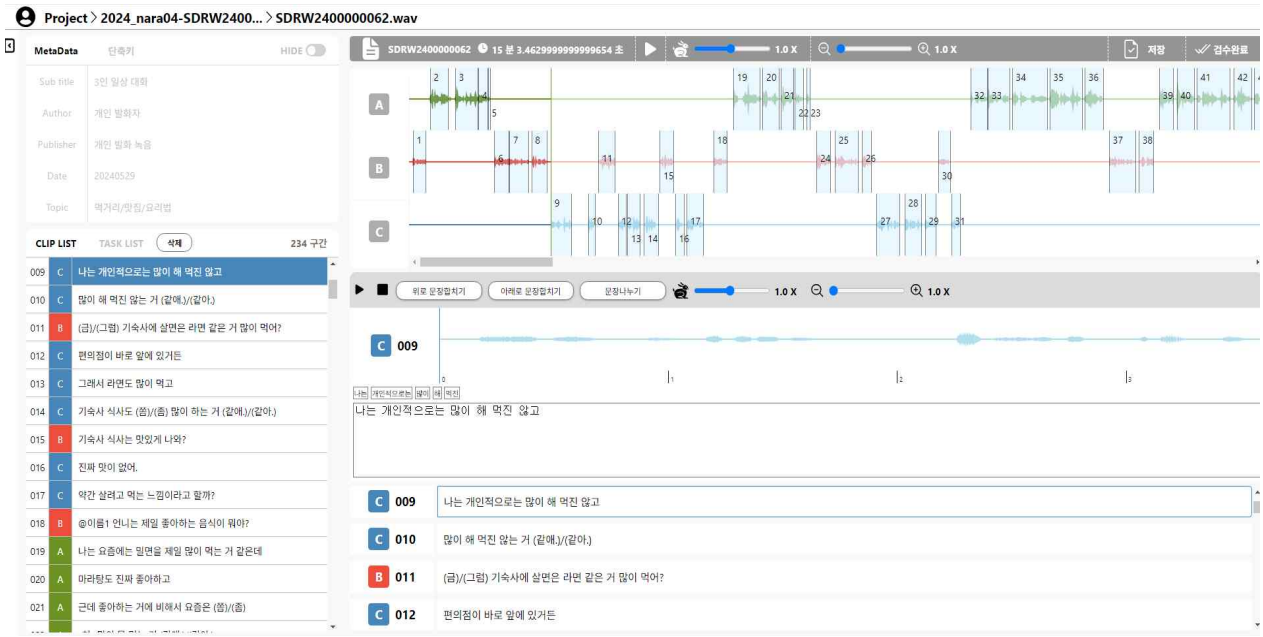
전사 작업을 위해 전사 인력을 대상으로 전사 도구 활용 방법과 전사 지침에 대한 교육을 진행하였다. 음성 전사 작업은 발화된 음성 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 이중 전사를 원칙으로 하였으며, 국어원에서 제시한 전사 지침을 준수하고 국립국어원 우리말샘, 국립국어원 한국어 어문규범 중 한국어 맞춤법, 표준어 규정 및 외래어 표기법 등을 참고하였다. 워크벤치를 활용한 관리자의 음성(메타데이터 포함) 업로드, 작업자 할당 작업이 완료되면 작업자별로 로그인 후 교육받은 지침에 따라 아래와 같은 내용으로 전사 작업을 실시하였다.

① 대화 종류 확인
- 1인 공적 독백, 2~4인 일상 대화

② 전사 작업
- 음성 분할(문장 나누기)
- 철자 확인 및 수정
- 이중 전사 처리
- 비식별 대상 표기
- 준음성 대상 표기 등

③ 메타데이터 확인 : 주제

전사 작업은 1개 음성을 처리하는데 1인 독백의 경우 40분~1시간 20분 정도의 시간이 소요되었고, 다자 대화의 경우 화자 간 간섭이 없는 경우 1시간 30분~2시간, 간섭이 심한 경우는 2시간 30분 이상이 소요되었다. 1인 공적 독백 639건, 다자 대화 2,288건, 총 2,927건의 대화에 대해 전사 작업을 진행하였다.

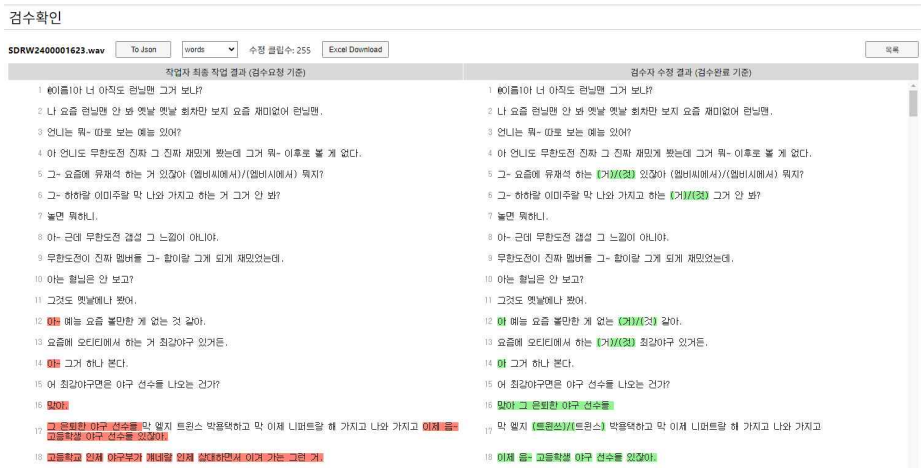


[그림 15] 전사 기능을 포함한 워크벤치(3인 대화 예)

6.3. 품질 검수

6.3.1. 전사 데이터 전수 점검

전사가 완료된 파일은 검수 담당자를 지정하여 전수 수작업 점검을 진행하였다. 전사 도구에서 전사 작업 후 검수 요청을 한 시점의 데이터와 검수 완료 시 저장된 데이터의 내용을 저장하여 검수 진행 여부를 확인할 수 있도록 하였다.



[그림 16] 워크벤치에서의 검수 여부 확인

품질 점검 시 음성 파일 품질, 음성 전사 정확성 여부, 전사 지침 준수 여부 등을 점검하여 수정을 진행하고, 반복되는 패턴 오류와 자주 발생하는 오류는 주기적인 교육을 통해 전사 작업자의 작업 품질을 높일 수 있도록 하였다.

<표 13> 품질 점검 내용

구분	점검 내용
음성 파일 점검	대화 주제와 무관한 내용이 제외되었는지 점검 녹음 상태 점검(음성 크기, 노이즈 등) 음성이 끊기거나 소리 단절이 있는지 점검
음성 전사 정확성	이중 전사 대상의 발음 전사와 철자 전사 병행 여부 확인 전사 누락, 중복, 오탈자 오류 점검 음성 분할 점검
메타 정보	메타 정보의 주제와 대화 내용 비교 점검

6.3.2. 데이터 프로파일링

검수가 완료된 파일은 데이터 프로파일링 작업을 통해 최종 검수를 진행하였다. 기계적인 검증을 통해 형식 오류 검증을 실시하였고, 발음 전사를 기준으로 한 철자 전사 검증과 철자 전사를 기준으로 한 발음 전사 검증을 통해 최대한의 인적 오류를 줄이고자 하였다.

형식 오류 점검은 입력된 데이터의 오류 유형을 분석하여 각 항목별 오류 기준을 마련하고 전사 파일에 대한 검사를 실시하여 간단히 형식 오류를 선별할 수 있었다.

<형식오류 기준>

구분	오류 유형	예	비고	조건
이중전사 형식 오류	괄호 누락 괄표 표기 오류 괄호 두 개 / 누락 단어 누락 이중전사 띄어쓰기 기타	현대 시티 아울렛이라든가 하는 (제)/(이제) 아울렛류 같은 데에서 름 (스트레쓰가)/(스트레스가)0 일 발이고 근데 아~ 선생이 그렇다고 치급하고 (체제제체))/(제외최해) 가지고 아~ 안 되겠더라. 그냥 유전으로 받는 (거라고)/(가이라고) 생각해서 나는 빈티지 (산)/() 되게 좋아해서 폼 자주 찾아보는데 그게 플라스틱으로 (랜들은) /(만든) 것이래요. 어~ 내가 배우고 싶은 (겨를) 폼 배울 수 있어서 좋고		철자전사(from)에 괄호가 포함된 것들만 찾기 ** 발음전사(original_form)에는 철 들리지 않는 음성 표기 괄괄호가 있어 철자전사에서만 확인 필요
잘 안들림	대문자 X 괄호 누락 ((또는)) 누락	그렇게 시간적으로 경제적으로 그러 ((XX)) 돈 쓰고 시간 쓰고 하면서 그런 식으로 아이들에 대해서 아이를 성 상품화 그렇게 안 좋은 비판적인 ((xx로도) 볼려고 했는데	X는 소문자 x로. 잘 안들림 표기는 앞 뒤에 쌍괄호	발음전사(original_form)에서 대문자 X가 들어간 것들 찾기 ((를 포함하고))를 포함하지 않는 것 찾기))를 포함하고 ((를 포함하지 않는 것 찾기
끊어진 단어	하이픈(-) 누락 하이픈(-) 추가 오기	저부장이 있어서 안 했~ 뭐~ 대출 이자 뭐~ 상환하는 것 땜에. 그 뒤로는 수면제를 복게 했는지 어쨌는지 모르겠는데 .	끊어진 단어는 앞 뒤에 하이픈 있어야 함	발음전사(original_form)에서 하이픈(-) 개수가 홀수인 것 찾기 하이픈(-)이 연속으로 쓰인 것 찾기
비식별화	일련번호 누락 일련번호 띄어쓰기 참욱표기 오탈자 @ 개수	저기 @창고에 @이름5 교장선생님이 얘기하신 것저럼 @이름 1 씨는 스마트 워치 사용하고 계신가요? @성로명2 며칠 전에 말이랒 갔다 왔는데 나는 @창장을4가 제일 기억에 남아. 신발 가방 모자 특히 그중에서 어떤 거 더 좋아해 @@이름2는?	일련번호는 필수 비식별 항목과 일련번호는 붙여야 함 아래 비식별항목 참고 아래 비식별항목 참고	
담화표지	대상 아님 물결표 위치 띄어쓰기 물결표 추가 오기	로~ 인자 같아지 키우는 사람들로 보면은 니가 아까 말한 ~뭐 요아정이나 아니면 그~ 포케 큔참은 거 같고. 그 ~가음 음~ 분무기로다 이렇게 뿌려 주기도 하고 하는데 그~ ~ 노래가 처음에 너무 좋아 가지고	★ '이, 그, 저, 아, 어, 예, 음, 옴, 뭐' 이외는 대상이 아님 담화표지 대상과 물결표는 붙여야 함	

<형식오류 검사 결과>

파일명	클립No	form	original_form	작업자	비고	수정여부
SDRW2400000203	SDRW2400000203.1.1.133	그리고 내가 모르는 걸 물어보	그리구 내가 모르는 걸 물어보	2024_nara03@nara.c	이중전사오류,	수정완료
SDRW2400002399	SDRW2400002399.1.1.111	그 스 그 내 안에 있는 모 그 얼	그~ 스~ 그~ 내 안에 있는 -모	2024_nara04@nara.c	끊어진단어(-)오류,	수정완료
SDRW2400001179	SDRW2400001179.1.1.28	예전에 언니랑 헬스장 같이 다	예전에 언니랑 헬쓰장 같이 다	2024_nara06@nara.c	이중전사오류,	수정완료
SDRW2400001686	SDRW2400001686.1.1.181	그렇게 에 효과적이라고 하는데	그렇게 에~ 효과적이라고 하는	2024_nara06@nara.c	담화표지,	수정완료
SDRW2400002410	SDRW2400002410.1.1.23	그 나중에 쪼끔 다시 한번 봐이	그 ~나중에 쪼끔 다시 한번 봐	2024_nara10@nara.c	담화표지,	수정완료
SDRW2400000212	SDRW2400000212.1.1.2	다시 시작하기가 (힘들드라구,	다시 시작하기가 (힘들드라구,	2024_nara12@nara.c	이중전사오류,	수정완료

[그림 17] 형식 오류 기준 및 검사 결과

효율적인 이중 전사 점검을 위해 어휘 분석 방법을 통해 동일 발음 전사에 대해서 서로 다른 철자 전사가 이루어진 경우, 또는 동일 철자 전사에 대해서 서로 다른 발음 전사가 나타난 예 등의 철자 및 형식 오류를 탐지하여 수정하였다.

<발음 기준 철자 전사 점검 결과>

파일명	클립No	발음	철자	원문	결과	비고
SDRW2400001318	SDRW2400001318.1.1.161	거에	것의	치기 공학과도 그런 (거에)/(것의) 얼종이야?		
SDRW2400001090	SDRW2400001090.1.1.56	그리구	그러고	(그리구)/(그러고) 나서 직업을 선택하라고 얘기를	철자	그리고
SDRW2400001242	SDRW2400001242.1.1.50	그리구	그러고	(그리구)/(그러고) 나서 고다음 뭐~ (알바에)/(알바의	철자	그리고
SDRW2400001376	SDRW2400001376.1.1.97	그리구	그러고	(그리구)/(그러고) 나서 나는 그 (아르바이트두)/(아	철자	그리고
SDRW2400000177	SDRW2400000177.1.1.15	그리구	그리고	(그리구)/(그리고) 요즘 만나는 남자 친구 아버님께		
SDRW2400000177	SDRW2400000177.1.1.65	그리구	그리고	(그리구)/(그리고) 코인이랑 코인은 분석이 있단		
SDRW2400000177	SDRW2400000177.1.1.128	그리구	그리고	(그리구)/(그리고) 나 내가 제일 좋아하는 이유는		
SDRW2400000177	SDRW2400000177.1.1.232	그리구	그리고	(그리구)/(그리고) 내가 이번에 아빠한테 (빡싱이)/(		
SDRW2400001376	SDRW2400001376.1.1.197	그리구	그리고	(그리구)/(그리고) 철급적으로 월급을 (그니까)/(그러		
SDRW2400001378	SDRW2400001378.1.1.234	그리구	그리고	(그리구)/(그리고) 나는 키가 큰 편이기 때문에		
SDRW2400000366	SDRW2400000366.1.1.46	그까	그니까	음~ 저는 (유투브)/(유튜브) 중에서 요즘 그~ 스케치	철자	그러니까
SDRW2400000947	SDRW2400000947.1.1.149	그까	그니까	계속 (그까)/(그니까) 대사를 남자라고 하다가	철자	그러니까
SDRW2400001226	SDRW2400001226.1.1.58	그까	그니까	저 제가 경험한 (거)/(것) 중에 이렇게 뭐 이렇게 뭐	철자	그러니까
SDRW2400000153	SDRW2400000153.1.1.25	그까	그러니까	어떤 동호회에 들었어요 (그까)/(그러니까) 독서 동		
SDRW2400000154	SDRW2400000154.1.1.94	그까	그러니까	(그까)/(그러니까) ◎운음 하여간 오프라인 (온나인)		
SDRW2400000599	SDRW2400000599.1.1.37	까	그니까	(까)/(그니까) 나도 (좀)/(좀) 열심히 해서 잘 치면 더	철자	그러니까
SDRW2400000241	SDRW2400000241.1.1.129	까	그러니까	(까)/(그러니까) 이번에 봄에 생일 때		
SDRW2400000242	SDRW2400000242.1.1.4	까	그러니까	그게 (까)/(그러니까) 먹어 버릇하면은 먹어지는데		
SDRW2400000242	SDRW2400000242.1.1.61	까	그러니까	(까)/(그러니까) 이게 음식을 하고 뭐~ 이런 (거다)/(		
SDRW2400000242	SDRW2400000242.1.1.149	까	그러니까	(까)/(그러니까) 사실은 그런 어떤 상담이나 정신과		
SDRW2400000279	SDRW2400000279.1.1.103	까	그러니까	(까)/(그러니까) -누- 누군가 사용했다는 흔적이 있		
SDRW2400000317	SDRW2400000317.1.1.181	까	그러니까	(까)/(그러니까) 지금 집에서 술을 안 먹다 보니까 요		
SDRW2400000317	SDRW2400000317.1.1.185	까	그러니까	(까)/(그러니까) 지금 취미 생활이 전 인제 여러 가지		
SDRW2400000319	SDRW2400000319.1.1.211	까	그러니까	(까)/(그러니까) 요번에 어~ 요번은 아니지, 한 몇 년		
SDRW2400000183	SDRW2400000183.1.1.76	한 번도	1번도	나는 해외 직구로 (한 번도)/(1번도) 사 본 적이 없어	이중전사 아님, '한번'	
SDRW2400000183	SDRW2400000183.1.1.214	한 번도	1번도	내가 딴 적은 (한 번도)/(1번도) 없었거든.	이중전사 아님, '한번'	
SDRW2400000184	SDRW2400000184.1.1.8	한 번도	1번도	나는 (파팅)/(과팅) (한 번도)/(1번도) 안 나가 봤는데	이중전사 아님, '한번'	

<철자 기준 발음 전사 점검 결과>

파일명	클립No	철자	발음	원문	작업자	결과	비고
SDRW2400000004	SDRW2400000004.1.1.71	100년을	백년을	그~ 사람으로 치면 (백 년을)/(100년을) 산 거라고 그	[illegible]	발음	백 년을
SDRW2400000004	SDRW2400000004.1.1.88	100년을	백 년을	(백 년을)/(100년을) 사람으로 치면 (백 살인데도)/(10	[illegible]		
SDRW2400001333	SDRW2400001333.1.1.51	100년을	백 년을	지금 재건축을 하는 아파트가 그렇게 -경- 튼튼하게 지	[illegible]		
SDRW2400001160	SDRW2400001160.1.1.58	100년이	백 년이	근데 이 엠비티아이 검사가 나온 게 (백 년이)/(100년	[illegible]		
SDRW2400000222	SDRW2400000222.1.1.198	100대	백대	지금 하나하나 다 바꾸고 있는 그 상황이라서 너무 (좋	[illegible]	발음	백 대
SDRW2400000530	SDRW2400000530.1.1.6	100대	백 대	뭐~ 등산 이야기 나오니까 어제 아래 그~ 서울에 입	[illegible]		
SDRW2400000530	SDRW2400000530.1.1.6	100대	백 대	뭐~ 등산 이야기 나오니까 어제 아래 그~ 서울에 입	[illegible]		
SDRW2400000530	SDRW2400000530.1.1.7	100대	백 대	야 (땟)/(몇) 번은 무슨 (땟)/(몇) 번 (두 번)/(2번) 갔다	[illegible]		
SDRW2400000530	SDRW2400000530.1.1.8	100대	백 대	뭐~ 일본 북알프스도 갔다 오고 (코타키나바루도)/(코	[illegible]		
SDRW2400000530	SDRW2400000530.1.1.9	100대	백 대	(그래)/(그래서) (내보고)/(나보고) (와)/(왜) (백 대)/(1	[illegible]		
SDRW2400000066	SDRW2400000066.1.1.68	100대를	백 대를	그때 아싸리 그~ 나한테 (백 대를)/(100대를) 맞고 버	[illegible]		
SDRW2400000066	SDRW2400000066.1.1.68	100대를	백 대를	그때 아싸리 그~ 나한테 (백 대를)/(100대를) 맞고 버	[illegible]		
SDRW2400000066	SDRW2400000066.1.1.69	100대를	백 대를	와 어떻게 사람이 (백 대를)/(100대를) 맞냐.	[illegible]		
SDRW2400000518	SDRW2400000518.1.1.85	100리	백 리	뭐~ (백 리)/(100리) 가다가 (한 사람)/(1사람) (이백 리	[illegible]		
SDRW2400000092	SDRW2400000092.1.1.92	100마디	백 마디	(백 마디)/(100마디) 말보다 그냥 (한 가지의)/(1가지의	[illegible]		
SDRW2400001044	SDRW2400001044.1.1.8	100마리	백 마리	근데 처음에는 구피로 시작했다가 점점 막 열대야도	[illegible]		
SDRW2400001044	SDRW2400001044.1.1.9	100마리	백 마리	근데 그렇게 (딱다)/(적다) 보니까 어항 (싸이즈도)/(사	[illegible]		
SDRW2400001044	SDRW2400001044.1.1.20	100마리를	백 마리를	근데 너는 (백 마리를)/(100마리를) 어떻게 ◎운음 하	[illegible]		
SDRW2400001617	SDRW2400001617.1.1.77	100만	백만	그래도 (백만)/(100만) 뷰 이상 넘어가고	[illegible]		
SDRW2400001164	SDRW2400001164.1.1.81	100만 엔을	백만 엔을	어떤 소녀가 교도소에 갔다가 나오게 된 이후로 이제	[illegible]		
SDRW2400001164	SDRW2400001164.1.1.83	100만 엔을	백만 엔을	또 거기서 (백만 엔을)/(100만 엔을) 구하게 되면은	[illegible]		
SDRW2400000271	SDRW2400000271.1.1.44	100만 원	백만원	막 (오십만 원)/(50만 원) (백만 원)/(100만 원) 이렇게	[illegible]	발음	백만 원
SDRW2400001266	SDRW2400001266.1.1.184	100만 원	백만원	그냥 -스- -스- 스타일러를 (살려고)/(사려고) 하면은	[illegible]	발음	백만 원
SDRW2400001104	SDRW2400001104.1.1.170	100만 원	백만 원	돌이 한 (백만 원)/(100만 원) 차이.	[illegible]		
SDRW2400000667	SDRW2400000667.1.1.213	100만 원	백만 원	어~ 엄청 막 (백만 원)/(100만 원) 하지 않는 이상	[illegible]		
SDRW2400001320	SDRW2400001320.1.1.92	100만 원	백만 원	(백만 원)/(100만 원) 이상 수술비도 들고	[illegible]		

[그림 18] 어휘 분석을 통한 검증

7. 음성 정제

7.1. 음성 정제 기준

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) 및 WAV로 저장하는 순으로 진행하였다. 이때 음성 구간 앞뒤에 200msec의 휴지가 포함되도록 저장했다. 또한, 음성 구간 앞뒤에 잡음이 포함되면 잡음 외에 200msec 이상의 휴지가 포함되도록 했다.

정제 기준은 다음 표와 같다.

<표 14> 음성 데이터 정제 기준

구분	준수 내용
기본 지침	• 대화 전체 음성 파일(원본)과 문장 단위로 분할된 파일(정제본)을 제출함. • 폴더 구조 및 파일명 부여 방식 등은 기구축 말뭉치(2019~2024년 구축 일상 대화 말뭉치 등)와 연계·통합되도록 폴더 구조 및 파일명 부여 방식 등은 주관기관과 협의하여야 함. • 녹음 및 분할 시 음성 구간이 잘리지 않도록 하여야 함. • 음성 구간 앞, 뒤에 200msec 이상의 휴지가 포함되어야 함. • 음성 구간 앞, 뒤에 잡음이 포함된 경우에는 잡음 외에 200msec 이상의 휴지가 포함되도록 함. • 16kHz 표본화, 16bit 양자화 선형 PCM 및 WAV 파일로 저장함. • 음성의 최대값이 10,000 ~ 20,000을 권장하며, 최대 32,767을 넘지 않도록 마이크 볼륨을 조절함.
파일 부여 방식	• 예시 : SDRW2500000001.json 원시 말뭉치 첫 번째 파일 ※ 참고: 음성 파일 파일명 부여 방식 · SDRW2500000001.pcm 음성 원본 첫 번째 파일 · SDRW2500000001-SD2500001.pcm 음성 원본 첫 번째 파일의 정제본 첫 번째 파일
음성 파일 포맷	• 기본: 샘플링 16kHz, 양자화 16bits headerless(little endian) linear PCM • 추가: 삭제 • 정제본: 채널별 mono 변환
말뭉치 파일 포맷	• UTF-8, 줄 바꿈 문자 LF(UNIX)

7.2. 음성 개인 정보 비식별화

말뭉치 자료 중 이름, 이메일 주소 등 계정 정보나 주민등록번호, 카드 번호 등 각종 번호 및

비밀번호와 상세 주소, 출신 소속 등 개인 정보와 관련된 모든 사항들은 노출되지 않도록 전사 작업 단계에서 비식별화를 진행하였다. 단, 정치인, 유명인, 상호명 및 상품명 등은 비식별화하지 않되 대화 맥락상 부정적으로 언급된 경우에 한해 비식별화하였다. 개인 정보에 해당하는 음성은 전사 시 표시하고 최종 산출물을 생성할 때 음성이 들리지 않도록 묵음 처리를 하는 방식으로 비식별화 하였다. [그림 21]의 예시에서 추출한 엑셀 파일의 '이름1'로 표시된 구간은 산출물 PCM에서 묵음으로 처리하였다.

<비식별화 대상 추출>

정제파일명	텍스트	구분
SDRW2400000483.1.1.2.wav	저~ 빳데리 연구하고 있는 주임 연구원 @이름1입니다 예.	비식별
SDRW2400000484.1.1.135.wav	쪼끔 @이름1처럼 계획을 해서 쪼끔 웨이팅을 하는 편이고.	비식별
SDRW2400000041.1.1.53.wav	지금 @이름1이가 얘기한 것처럼 드레스를 구매를 하면	비식별
SDRW2400000041.1.1.62.wav	아 그리고 궁금했던 게 @이름2이랑 결혼을 결심하게 된 결정적인 이유가 있	비식별

<비식별화>

○ 비식별화 전

○ 비식별화 후

[그림 19] 개인 정보 비식별화 예시

8. 원시 말뭉치 구축 및 메타 정보 구축

8.1. JSON 변환

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

<표 15> 대화 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축년도	8자리 일련번호
S: 구어 말뭉치	A: 공적 독백 D: 사적 대화	RW: 원시 말뭉치	25	########

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


```
{
    "id": "SDRW2400001372",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2400001372",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2024",
        "category": "구어 > 사적 대화 > 일상 대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2400001372.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20240719",
                "topic": "건강/다이어트/질병",
                "speaker": [
                    {
                        "id": "SD2401103",
                        "age": "30대",
                        "occupation": "사무 종사자",
                        "sex": "여성",
                        "birthplace": "대전",
                        "principal_residence": "대전",
                        "current_residence": "대전",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2401104",
                        "age": "60대 이상",
                        "occupation": "주부",
                        "sex": "여성",
                        "birthplace": "대구",
                        "principal_residence": "대전",
                        "current_residence": "대전",
                        "education": "대졸"
                    }
                ],
                "setting": {
                    "relation": "부모/자녀",
                    "contact_frequency": "5"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2400001372.1.1.1",
                    "form": "내가 작년에 살이 엄청 쪘잖아 거의 20킬로 가까이.",
                    "original_form": "내가 작년에 살이 엄청 쪘잖아 거의 이십 키로 가까이.",
                    "speaker_id": "SD2401103",
                    "start": 0.589,
                    "end": 5.059,
                    "note": ""
                },
                {
                    "id": "SDRW2400001372.1.1.2",
                    "form": "그런데 살이 찌고 나니까",
                    "original_form": "데 살이 찌고 나니까",
                    "speaker_id": "SD2401103",
                    "start": 5.349,
                    "end": 7.23056,
                    "note": ""
                },
                {
                    "id": "SDRW2400001372.1.1.3",
                    "form": "일상생활이 너무 힘들어지는 게 많이 느껴져, 숨 쉬는 것도 좀 헉헉대게 되고",
                    "original_form": "일상생활이 너무 힘들어지는 게 많이 느껴져, 숨 쉬는 것도 좀 헉헉대게 되고",
                    "speaker_id": "SD2401103",
                    "start": 7.23056,
                    "end": 13.0364,
                    "note": ""
                },
```


[그림 20] 말뭉치 변환 예시(일부)

<표 16> JSON 구조

수준 1	수준 2	수준 3	수준 4	타입	설명	Length	필수 여부	데이터
id				string	말뭉치 파일 아이디	14	Y	변동
metadata				object	말뭉치 파일의 메타 정보			
	title			string	말뭉치 파일 제목	100	Y	변동
	creator			string	구축자: 국립국어원	20	Y	고정
	distributor			string	배포자: 국립국어원	20	Y	고정
	year			string	구축년도: 2025	4	Y	고정
	category			string	분류: 구어 > 사적 대화 > 일상 대화	100	Y	고정
	annotation_level			array(string)	분석 층위: 원시	10	Y	고정
	sampling			string	샘플링 방식: 본문 전체	20	Y	고정
document				array(object)	대화 정보			
	id			string	대화 아이디	20	Y	변동
	metadata			object	대화 메타 정보			
		title		string	대화 제목: 2인 일상 대화	20	Y	변동
		author		string	저작권자: 개인 발화자	20	Y	고정
		publisher		string	발행자: 개인 발화 녹음	20	Y	고정
		date		string	녹음일자: YYYYMMDD	8	Y	변동
		topic		string	대화 주제: 대주제 > 세부주제	100	Y	변동
		speaker		array(object)	화자 정보			
			id	string	화자 아이디	9	Y	변동
			age	string	연령	10	Y	변동
			occupation	string	직업	100	Y	변동
			sex	string	성별	10	Y	변동
			birthplace	string	출생지	10	Y	변동
			principal_residence	string	주 성장지	10	Y	변동
			current_residence	string	현 거주지	10	Y	변동
			education	string	학력	20	Y	변동
		setting		object	환경 정보			
			relation	string	화자 간 관계	20	Y	변동
			contact_frequency	string	친밀도(대화 빈도)	1	Y	
	utterance			array(object)	발화 정보			
		id		string	발화 아이디	25	Y	변동
		form		string	철자 전사	1000	Y	변동
		original_form		string	발음 전사	1000	Y	변동
		speaker_id		string	화자 아이디	9	Y	변동
		start		num	발화 시작 시간	8	Y	변동
		end		num	발화 종료 시간	8	Y	변동
		note		string	전사자 기타 메모	1000	N	변동

8.2. 메타 정보 구축

메타 정보 구축은 수집에 참여한 화자의 정보(성별, 연령, 직업, 출생지, 주 성장지, 현 거주지 등)와 수집에 참여한 화자들 간의 관계를 필수로 작성하였고, 대화 주제는 대주제(topic1)와 세부 주제(topic2)로 나누어서 기재하였다.

	2025 일상 대화 말뭉치 구축 대화(파일) 메타 정보										
번호	대화ID	타이틀	주제1	주제2	관계	친밀도	녹음지	파일 시간	저작권 확보 여부	민감이슈여부	분류항목
637	SARW2500000667	1인 공적 독백	뷰티/패션	자라 역대급 여름 세일! 안사면 후회할 ZARA 필수템 추천	기타	N/A	기타	0:10:13	O	X	
638	SARW2500000668	1인 공적 독백	뷰티/패션	2만 원~30만 원 가격대별 반팔 핸리넥 추천!	기타	N/A	기타	0:08:16	O	X	
639	SARW2500000669	1인 공적 독백	강연	조용히 단단해지는 사람들의 철학	기타	N/A	기타	0:12:38	O	X	
640	SARW2500000670	1인 공적 독백	강연	무심함이 만드는 차별, 작은 관심이 만드는 변화	기타	N/A	기타	0:09:17	O	X	
641	SARW2500000671	1인 공적 독백	강연	공평한 세상은 공감에서 시작된다	기타	N/A	기타	0:08:54	O	X	
642	SARW2500000672	1인 공적 독백	푸드/쿠킹	호불호 갈리는 음식 월드컵	기타	N/A	기타	0:08:44	O	X	
643	SARW2500000673	1인 공적 독백	뷰티/패션	심은 패션 월드컵	기타	N/A	기타	0:14:45	O	X	
644	SDRW2500000003	2인 일상 대화	음식	좋아하는 음식과 요리 레시피	기타	0	대구	0:13:34	O	X	
645	SDRW2500000004	2인 일상 대화	여행/휴가/휴일/자연휴양지	가고 싶은 여행지 및 추천여행지	기타	0	대구	0:13:39	O	X	
646	SDRW2500000005	4인 일상 대화	여행/휴가/휴일/자연휴양지	여행지에 관련된 경험담	기타	0	가산	0:14:48	O	X	
647	SDRW2500000006	4인 일상 대화	회사/학교/학창시절	직업 소개 및 경험담	기타	0	가산	0:15:22	O	X	
648	SDRW2500000007	4인 일상 대화	음식	좋아하는 음식 및 맛집 투어	기타	0	가산	0:14:49	O	X	
649	SDRW2500000008	4인 일상 대화	문화예술	좋아하는 아이돌 및 문화예술 관련 얘기	기타	0	가산	0:14:13	O	X	

[그림 21] 메타 정보 파일 일부

	2025 일상 대화 말뭉치 구축 대화별 발화자 메타 정보								
번호	파일명	화자 ID	이름	연령대	직업	성별	출생지	주 성장지	현 거주지	최종학력
55	SARW2500000063	SD2502012	한O현	60대 이상	기타	남성	충남	서울	경기	대졸
56	SARW2500000064	SD2502012	한O현	60대 이상	기타	남성	충남	서울	경기	대졸
57	SARW2500000066	SD2502011	이O지	30대	서비스 종사자	여성	서울	서울	경북	대졸
58	SARW2500000067	SD2502011	이O지	30대	서비스 종사자	여성	서울	서울	경북	대졸
59	SARW2500000068	SD2502011	이O지	30대	서비스 종사자	여성	서울	서울	경북	대졸
60	SARW2500000069	SD2502011	이O지	30대	서비스 종사자	여성	서울	서울	경북	대졸
61	SARW2500000070	SD2502013	김O연	30대	전문가 및 관련 종사자	남성	충북	충북	충북	대졸
62	SARW2500000071	SD2502013	김O연	30대	전문가 및 관련 종사자	남성	충북	충북	충북	대졸
63	SARW2500000072	SD2502013	김O연	30대	전문가 및 관련 종사자	남성	충북	충북	충북	대졸
64	SARW2500000073	SD2502013	김O연	30대	전문가 및 관련 종사자	남성	충북	충북	충북	대졸
65	SARW2500000074	SD2502013	김O연	30대	전문가 및 관련 종사자	남성	충북	충북	충북	대졸
66	SARW2500000075	SD2502014	권O성	20대	사무 종사자	여성	서울	서울	서울	대졸

[그림 22] 발화자 정보 파일 일부

제3장

사업 수행 결과

1. 일상 대화 말뭉치: 2~4인 다자 대화 수집 결과

1.1. 주제별·제시 자료별 수집 결과

2~4인 대화 말뭉치는 총 2,288건을 수집하였으며, 주관기관과의 협의를 거쳐 선정한 10개 주제가 편중되지 않도록 하였다. 주제별 수집 결과는 다음과 같다.

<표 17> 다자 대화 주제별 수집 결과

순번	주제	건수	시간	비율
1	**건강**	**269**	**65.36**	**11.76%**
2	**문화예술**	**249**	**60.07**	**10.88%**
3	**음식**	**247**	**59.81**	**10.80%**
4	**경제**	**229**	**55.95**	**10.01%**
5	**회사/학교/학창시절**	**264**	**63.15**	**11.54%**
6	반려동물/반려용품	137	32.00	5.99%
7	여행/휴가/휴일/자연휴양지	244	60.03	10.66%
8	쇼핑	230	57.04	10.05%
9	새로운 기술과 우리 생활	190	44.66	8.30%
10	사회적 변화와 우리 생활	229	55.24	10.01%
합계		2,288	553.31	100.00%

1.2. 인구분포별 수집 결과

총 1,706명의 화자가 참여하였고 성별, 연령별, 지역별 분포는 다음과 같다.

<표 18> 다자 대화 성별, 연령별, 지역별 모집 결과(단위: 명)

구분	10대		20대		30대		40대		50대		60대 이상		총합계	
	남	여	남	여	남	여	남	여	남	여	남	여	남	여
합계	228		312		298		234		244		390		1,706	
	94	134	144	168	133	165	85	149	119	125	178	212	753	953
구분	남	여	남	여	남	여	남	여	남	여	남	여	지역별	권역별
서울	24	29	26	22	29	32	31	38	31	28	30	38	358	857
인천	5	8	8	7	10	12	5	6	8	7	9	15	100	
경기	31	40	37	33	33	38	17	30	26	32	34	48	399	
강원	4	6	5	5	4	6	3	6	7	2	8	9	65	65
부산	5	10	11	15	11	12	4	13	8	9	7	7	112	421
대구	0	6	9	26	6	9	8	6	4	4	6	10	94	
울산	0	3	4	7	3	5	0	2	1	2	0	4	31	
경북	0	4	8	10	5	7	6	6	1	5	8	9	69	
경남	22	9	10	13	9	13	2	8	2	5	13	9	115	
대전	0	3	4	5	1	3	1	0	2	2	7	9	37	160
충북	1	1	5	3	3	5	1	3	6	3	12	9	52	
충남	0	3	2	7	5	3	0	8	9	8	13	13	71	
광주	2	9	5	3	4	8	0	7	0	5	5	7	55	172
전북	0	0	4	5	5	7	3	3	8	6	11	11	63	
전남	0	2	6	4	1	5	1	4	6	6	9	10	54	
제주	0	1	0	3	4	0	3	9	0	1	6	4	31	31
합계	94	134	144	168	133	165	85	149	119	125	178	212	1,706	1,706

1.3. 주제별 연령 분포

주제별 연령 분포는 다음과 같다. 주제별로 다양한 연령층이 고르게 대화하였음을 알 수 있다. 그중 10대는 새로운 기술과 우리 생활, 20대는 여행, 휴가, 휴일, 자연휴양지 등을 주제로 대화를 많이 하였고, 30대와 40대는 문화예술과 쇼핑, 50대와 60대 이상은 건강과 음식 주제로 대화를 많이 하였다.

<표 19> 다자 대화 주제별 연령 분포(단위: 명)

주제	10대	20대	30대	40대	50대	60대 이상	합계	비율
건강	17	101	88	80	129	245	660	11.65%
문화예술	93	156	138	101	50	78	615	10.85%
음식	60	150	107	83	108	160	668	11.79%
경제	50	115	111	72	86	119	553	9.76%
회사/학교/학창시절	89	139	120	93	74	143	658	11.61%
반려동물/반려용품	96	38	55	34	45	49	317	5.59%
여행/휴가/휴일/자연휴양지	52	173	132	80	97	144	678	11.96%
쇼핑	89	125	143	101	88	56	602	10.62%
새로운 기술과 우리 생활	137	85	79	82	35	4	422	7.45%
사회적 변화와 우리 생활	44	56	78	62	93	161	494	8.72%
총합계	727	1,138	1,051	788	805	1,159	5,667	100%

1.4. 주제별 성별 분포

주제별 성별 분포를 보면 편차 없이 고르게 다양한 주제로 남성과 여성이 대화하였음을 알 수 있다. 그중 남성은 회사, 학교, 학창시절을 주제로, 여성은 음식을 주제로 대화가 높은 비율을 보였다.

<표 20> 다자 대화 주제별 성별 분포(단위: 명)

주제	남성		여성		합계
건강	289	43.79%	371	56.21%	660
문화예술	239	38.86%	376	61.14%	616
음식	194	29.04%	474	70.96%	668
경제	260	47.02%	293	52.98%	553
회사/학교/학창시절	329	50.00%	329	50.00%	658
반려동물/반려용품	143	45.11%	174	54.89%	317
여행/휴가/휴일/자연휴양지	214	31.56%	464	68.44%	678
쇼핑	150	24.92%	452	75.08%	602
새로운 기술과 우리 생활	263	62.32%	159	37.68%	422
사회적 변화와 우리 생활	294	59.51%	200	40.49%	494
총합계	2,375	41.91%	3,292	58.09%	5,667

1.5. 화자 간 관계별 분포

화자 간 관계를 비율 순으로 살펴보면, 기타(55.28%), 친구(14.07%), 모임·동아리 지인(13.36%), 직장 동료(5.28%) 순의 결과를 보였다.

<표 21> 다자 대화 화자 간 관계별 분포(세분류)

관계	인원(명)	비율
친구	240	14.07%
부부	36	2.11%
부모/자녀	35	2.05%
형제/자매	20	1.17%
연인	16	0.94%
직장 동료	90	5.28%
이웃사촌	66	3.87%
모임·동아리 지인	228	13.36%
대학 선후배	2	0.12%
교회 지인	10	0.59%
고향 선후배	10	0.59%
기타 가족	10	0.59%
기타	943	55.28%
합계	1,706	100.00%

<표 22> 다자 대화 화자 간 관계별 분포(대분류)

관계(대분류)	인원(명)	비율
가족 관계	91	5.33%
사회적 관계	672	39.39%
기타	943	55.28%
합계	1,706	100.00%

1.6. 화자의 직업별 분포

직업별 수집된 결과를 보면 전체 비율 중에서 무직, 취업준비생이 전체의 24.79%를 차지하였고, 학생(20.4%), 주부(20.34%), 사무 종사자(11.37%) 순으로 나타났다.

<표 23> 다자 대화 화자의 직업별 분포

직업	인원(명)	비율
경영/관리직	13	0.76%
전문가 및 관련 종사자	89	5.22%
사무 종사자	194	11.37%
서비스 종사자	93	5.45%
판매/영업 종사자	32	1.88%
농업/임업/어업 종사자	4	0.23%
기능원 및 관련 기능 종사자	1	0.06%
기술자 종사자(장치/기계 조작 및 조립 종사자)	10	0.59%
단순노무 종사자	46	2.70%
학생	348	20.40%
주부	347	20.34%
무직/취업준비생	423	24.79%
기타	106	6.21%
합계	1,706	100.00%

1.7. 학력별 분포

화자의 학력별 분포를 보면 대졸이 전체 인원의 49.3%를 차지하였고, 그다음으로 고등학교 졸업(20.46%), 중학교 졸업(10.08%)이 뒤를 이었다.

<표 24> 다자 대화 화자의 학력별 분포

학력	인원(명)	비율
초졸 이하	105	6.15%
중졸	172	10.08%
고졸	349	20.46%
대재	138	8.09%
대졸	841	49.30%
대학원 이상	101	5.92%
합계	1,706	100.00%

1.8. 출생지별 분포

화자의 출생지별 수집 결과를 보면 서울이 전체 인원의 26.2%를 차지하였고, 그다음으로 경기(17.29%), 부산(7.74%), 순으로 나타났다.

<표 25> 다자 대화 화자의 출생지별 분포

출생지	인원(명)	비율
서울	447	26.20%
인천	61	3.58%
경기	295	17.29%
강원	65	3.81%
부산	132	7.74%
대구	101	5.92%
울산	16	0.94%
경북	91	5.33%
경남	98	5.74%
대전	30	1.76%
충북	55	3.22%
충남	82	4.81%
광주	55	3.22%
전북	61	3.58%
전남	92	5.39%
제주	25	1.47%
총합계	1,706	100%

1.9. 화자의 주 성장지별 분포

지역별 화자는 주 성장지를 기준으로 선발하였고, 해당 지역의 비율을 고려하여 수집하였다. 화자의 주 성장지별 수집 결과를 보면 경기가 전체의 23.39%로 가장 많고 서울이 20.98%, 경남이 6.74%를 차지하였다.

<표 26> 다자 대화 주 성장지별 분포

주 성장지	인원(명)	비율
서울	358	20.98%
인천	100	5.86%
경기	399	23.39%
강원	65	3.81%
부산	112	6.57%
대구	94	5.51%
울산	31	1.82%
경북	69	4.04%
경남	115	6.74%
대전	37	2.17%
충북	52	3.05%
충남	71	4.16%
광주	55	3.22%
전북	63	3.69%
전남	54	3.17%
제주	31	1.82%
합계	1,706	100%

1.10. 현 거주지별 분포

화자의 현 거주지별 수집 결과를 보면 서울이 전체 인원의 40.5%로 가장 많고 경기가 31.07%, 부산이 9.03%를 차지하였다.

<표 27> 다자 대화 현 거주지별 분포

현 거주지	인원(명)	비율
서울	691	40.50%
인천	45	2.64%
경기	530	31.07%
강원	21	1.23%
부산	154	9.03%
대구	103	6.04%
울산	3	0.18%
경북	13	0.76%
경남	33	1.93%
대전	1	0.06%
충북	20	1.17%
충남	12	0.70%
광주	42	2.46%
전북	0	0.00%
전남	4	0.23%
제주	34	1.99%
합계	1,706	100%

1.11. 다자 대화 수집 목표 대비 실적

다자 대화의 수집 목표 대비 수집량은 아래 표와 같다.

<표 28> 다자 대화 화자 모집 목표 대비 실적(단위: 시간)

	연령비												합계	
	10대		20대		30대		40대		50대		60대 이상			
	목표량	수집량	목표량	수집량	목표량	수집량	목표량	수집량	목표량	수집량	목표량	수집량	목표량	수집량
	15%	13%	15%	18%	15%	18%	20%	14%	15%	14%	20%	24%	100%	100%
계	82.50	70.46	82.50	97.26	82.50	97.11	110.0	77.32	82.50	79.41	110.0	131.75	550.00	553.31
서울	15.00	16.39	15.00	14.95	15.00	19.77	20.0	22.60	15.00	18.97	20.0	23.98	100.00	116.66
인천	4.95	4.44	4.95	4.87	4.95	7.22	6.6	3.76	4.95	5.21	6.6	7.98	33.00	33.49
경기	19.50	22.10	19.50	23.13	19.50	23.32	26.0	15.82	19.50	19.20	26.0	27.42	130.00	131.00
강원	3.00	3.42	3.00	3.27	3.00	3.44	4.0	3.04	3.00	2.75	4.0	5.82	20.00	21.74
부산	5.25	4.61	5.25	8.43	5.25	7.13	7.0	5.20	5.25	5.36	7.0	4.74	35.00	35.47
대구	3.75	1.54	3.75	9.59	3.75	4.29	5.0	4.33	3.75	2.51	5.0	5.35	25.00	27.60
울산	1.80	0.87	1.80	3.14	1.80	2.36	2.4	0.74	1.80	0.86	2.4	1.15	12.00	9.11
경북	4.05	1.18	4.05	5.46	4.05	3.76	5.4	3.88	4.05	1.75	5.4	5.95	27.00	21.98
경남	5.25	8.64	5.25	7.21	5.25	6.93	7.0	3.28	5.25	2.17	7.0	7.18	35.00	35.41
대전	2.40	0.94	2.40	2.63	2.40	1.40	3.2	0.40	2.40	1.27	3.2	4.88	16.00	11.51
충북	2.55	0.71	2.55	2.30	2.55	2.63	3.4	1.56	2.55	3.07	3.4	6.85	17.00	17.12
충남	4.05	1.13	4.05	3.02	4.05	2.87	5.4	2.55	4.05	6.06	5.4	8.64	27.00	24.26
광주	2.25	3.43	2.25	2.48	2.25	4.60	3.0	2.21	2.25	1.55	3.0	4.10	15.00	18.39
전북	2.85	0.00	2.85	2.70	2.85	3.89	3.8	2.09	2.85	4.42	3.8	7.29	19.00	20.40
전남	2.85	0.75	2.85	3.15	2.85	2.07	3.8	1.73	2.85	3.88	3.8	6.50	19.00	18.09
제주	3.00	0.31	3.00	0.94	3.00	1.43	4.0	4.11	3.00	0.40	4.0	3.91	20.00	11.09

<표 29> 다자 대화 분야별 수집 목표 대비 실적

구분	목표		실적		달성률
	시간	비율	시간	비율	
2인 대화	368.5	67%	363.1	65.62%	98.53%
3인 대화	99	18%	103.3	18.67%	104.34%
4인 대화	82.5	15%	86.91	15.71%	105.35%
합계	550	100%	553.31	100%	100.60%

<표 30> 다자 대화 성별 수집 목표 대비 실적

구분	목표		실적		달성률
	시간	비율	시간	비율	
남성	275	50%	246.46	44.54%	89.62%
여성	275	50%	306.84	55.46%	111.58%
합계	550	100%	553.31	100%	100.60%

<표 31> 다자 대화 연령별 수집 목표 대비 실적

구분	목표		실적		달성률
	시간	비율	시간	비율	
10대	82.5	15%	70.46	12.74%	85.41%
20대	82.5	15%	97.26	17.58%	117.89%
30대	82.5	15%	97.10	17.55%	117.69%
40대	110	20%	77.32	13.97%	70.29%
50대	82.5	15%	79.41	14.35%	96.26%
60대 이상	110	20%	131.75	23.81%	119.77%
합계	550	100%	553.31	100%	100.60%

<표 32> 다자 대화 연령별, 성별 수집 목표 대비 실적

구분	성별	목표		실적		달성률
		시간	비율	시간	비율	
10대	남성	41.25	7.50%	26.56	4.80%	64.40%
	여성	41.25	7.50%	43.90	7.93%	106.42%
20대	남성	41.25	7.50%	47.09	8.51%	114.16%
	여성	41.25	7.50%	50.17	9.07%	121.63%
30대	남성	41.25	7.50%	44.23	7.99%	107.24%
	여성	41.25	7.50%	52.86	9.55%	128.15%
40대	남성	55	10.00%	27.58	4.98%	50.14%
	여성	55	10.00%	49.74	8.99%	90.44%
50대	남성	41.25	7.50%	40.76	7.37%	98.80%
	여성	41.25	7.50%	38.66	6.99%	93.71%
60대 이상	남성	55	10.00%	60.24	10.89%	109.52%
	여성	55	10.00%	71.51	12.92%	130.02%
합계		550	100.00%	553.31	100.00%	100.60%

<표 33> 다자 대화 지역(권역)별 수집 목표 대비 실적(단위: 시간)

구분	지역별					권역별				
	목표		실적		달성률	목표		실적		달성률
	시간	비율	시간	비율		시간	비율	시간	비율	
서울	100	18.20%	116.66	21.10%	116.70%	263	47.80%	281.15	50.80%	106.90%
인천	33	6.00%	33.49	6.10%	101.50%					
경기	130	23.60%	131	23.70%	100.80%					
강원	20	3.60%	21.74	3.90%	108.70%	20	3.60%	21.74	3.90%	108.70%
부산	35	6.40%	35.47	6.40%	101.30%	134	24.40%	129.57	23.40%	96.70%
대구	25	4.50%	27.6	5.00%	110.40%					
울산	12	2.20%	9.11	1.60%	75.90%					
경북	27	4.90%	21.98	4.00%	81.40%					
경남	35	6.40%	35.41	6.40%	101.20%					
대전	16	2.90%	11.51	2.10%	71.90%	60	10.90%	52.89	9.60%	88.20%
충북	17	3.10%	17.12	3.10%	100.70%					
충남	27	4.90%	24.26	4.40%	89.90%					
광주	15	2.70%	18.39	3.30%	122.60%	53	9.60%	56.88	10.30%	107.30%
전북	19	3.50%	20.4	3.70%	107.40%					
전남	19	3.50%	18.09	3.30%	95.20%					
제주	20	3.60%	11.09	2.00%	55.50%	20	3.60%	11.09	2.00%	55.50%
합계	550	100%	553.31	100%	100.60%	550	100%	553.31	100%	100.60%
합계	550	100%	553.31	100.00%	100.60%	550	100%	553.31	100%	100.60%

<표 34> 다자 대화 주제별 수집 목표 대비 실적(단위: 시간)

연번	구분	목표		실적		달성률
		시간	비율	시간	비율	
주제1	건강	66	12%	65.36	11.80%	99.00%
주제2	문화예술	60.5	11%	60.07	10.90%	99.30%
주제3	음식	60.5	11%	59.81	10.80%	98.90%
주제4	경제	55	10%	55.95	10.10%	101.70%
주제5	회사/학교/학창시절	60.5	11%	63.15	11.40%	104.40%
주제6	반려동물/반려용품	33	6%	32	5.80%	97.00%
주제7	여행/휴가/휴일/자연휴양지	60.5	11%	60.03	10.80%	99.20%
주제8	쇼핑	55	10%	57.04	10.30%	103.70%
주제9	새로운 기술과 우리 생활	44	8%	44.66	8.10%	101.50%
주제10	사회적 변화와 우리 생활	55	10%	55.24	10.00%	100.40%
합계		550	100%	553.31	100%	100.60%

1.12. 다자 대화 전사 결과

2~4인 대화 말뭉치 2,288건의 대화 수는 68,746건이며 말뭉치 1건 당 평균 대화 수는 30.05건이다. 발화 문장 수는 505,305문장, 말뭉치 1건 당 평균 220.85문장이며, 전사 어절 수는 총 4,069,701어절, 말뭉치 1건 당 평균 1,778.72어절이다. 또한 다자 대화 참여자 1,706명은 1명 당 평균 대화 수 40.30건, 발화 문장 수 296.19문장, 전사 어절 수 2,385.52어절 발화한 것으로 나타났다. 분야별, 주제별, 성별, 연령별 전사 결과는 다음과 같다.

<표 35> 다자 대화 분야별 전사 결과

순번	주제	건수	시간	대화 수	발화 문장 수	전사 어절 수
1	2인 대화	1,535	363.1	48,037	331,383	2,708,608
2	3인 대화	414	103.3	12,549	93,250	734,437
3	4인 대화	339	86.91	8,160	80,672	626,656
합계		2,288	553.31	68,746	505,305	4,069,701

<표 36> 다자 대화 주제별 전사 결과

순번	주제	건수	시간	대화 수	발화 문장 수	전사 어절 수
1	건강	269	65.36	7,583	60,945	481,911
2	문화예술	249	60.07	8,951	54,283	435,682
3	음식	247	59.81	9,291	56,985	437,833
4	경제	229	55.95	4,972	50,494	415,500
5	회사/학교/학창시절	264	63.15	7,394	57,459	453,143
6	반려동물/반려용품	137	32.00	5,083	29,241	238,897
7	여행/휴가/휴일/자연휴양지	244	60.03	6,563	56,623	436,626
8	쇼핑	230	57.04	8,309	52,233	429,178
9	새로운 기술과 우리 생활	190	44.66	6,280	38,502	338,624
10	사회적 변화와 우리 생활	229	55.24	4,320	48,540	402,307
합계		2,288	553.31	68,746	505,305	4,069,701

<표 37> 다자 대화 성별 전사 결과

순번	주제	시간	대화 수	발화 문장 수	전사 어절 수
1	남성	246.46	41,408	276,085	2,211,270
2	여성	306.84	27,338	229,220	1,858,431
합계		553.31	68,746	505,305	4,069,701

<표 38> 다자 대화 연령별 전사 결과

순번	연령	시간	대화 수	발화 문장 수	전사 어절 수
1	10대	70.46	19,210	62,272	499,544
2	20대	97.26	12,807	91,504	700,091
3	30대	97.10	11,429	83,580	689,964
4	40대	77.32	10,562	71,304	597,099
5	50대	79.41	6,061	72,230	612,554
6	60대 이상	131.75	8,677	124,415	970,449
합계		553.31	68,746	505,305	4,069,701

<표 39> 다자 대화 연령별, 성별 전사 결과

구분	성별	시간	대화 수	발화 문장 수	전사 어절 수
10대	남성	26.56	8,084	24,684	200,084
	여성	43.90	11,126	37,588	299,460
20대	남성	47.09	5,311	45,823	358,384
	여성	50.17	7,496	45,681	341,707
30대	남성	44.23	4,099	39,528	336,348
	여성	52.86	7,330	44,052	353,616
40대	남성	27.58	3,535	24,588	215,294
	여성	49.74	7,027	46,716	381,805
50대	남성	40.76	2,636	36,872	313,433
	여성	38.66	3,425	35,358	299,121
60대 이상	남성	60.24	3,673	57,725	434,888
	여성	71.51	5,004	66,690	535,561
합계		553.31	68,746	505,305	4,069,701

2. 일상 대화 말뭉치: 1인 발화(독백) 수집 결과

2.1. 공적 독백 수집 결과

공적 독백 대상은 앞서 기술하였듯이 유튜브에 게시된 콘텐츠와 강연 등이며, 크리에이터의 동의를 얻어 구축된 대상 채널은 61채널, 294개의 콘텐츠이다. 해당 채널 및 콘텐츠의 상세 목록은 아래와 같다.

<표 40> 공적 독백 참여 채널 및 콘텐츠 목록

채널명	콘텐츠명
사물궁이 잡학지식	인공지능(AI)도 의식을 가질 수 있을까?
사물궁이 잡학지식	유리는 왜 투명하고 잘 깨질까?
사물궁이 잡학지식	토너·스킨 안 바르고 로션만 바르면 안 될까?
사물궁이 잡학지식	한국은 왜 220V 전압을 사용하는 걸까?
사물궁이 잡학지식	정말 왼쪽으로 누워서 자면 더 좋을까?
사물궁이 잡학지식	눈으로 들어간 눈썹은 어디로 갈까?
사물궁이 잡학지식	왕조시대 때 신하들은 어떻게 타이밍을 맞춰서 합창했을까?
사물궁이 잡학지식	자동차 급발진이 정말 있다면 어떤 게 원인이 될 수 있을까?
안될과학 Unrealscience	우리는 왜 지브리에 열광하는가?! 뇌과학적으로 분석한 지브리 사진에 중독되는 진짜 이유! [안될과학 뉴런]
안될과학 Unrealscience	뇌와 컴퓨터를 인공지능으로 연결해서 뇌의 신호를 커버한다?! 이제는 사람의 뇌에도 침투하는 인공지능?! [뉴스속으로 - 뉴런]
안될과학 Unrealscience	제임스 웹 우주망원경 돌아오다! 제임스 웹이 관측한 오래된 은하의 비밀은?! [항성의 우주속으로]
안될공학 - IT 테크 신기술	구글, NVIDIA 버리고 AI 칩 시장까지... 9년 동안 설계 빛 보다 | 7세대 TPU Ironwood 발표 | 가장 빠른 슈퍼 컴퓨터보다 24배
안될공학 - IT 테크 신기술	미쳐버린 구글AI 근황… 플랫폼 강자가 앞선다 | 딥씨크 GPT4보다 좋은 소형 LLM(오픈모델) gemma3 | 고퀄 이미지 통합 LLM | 딥리서치 무료 | 진정한 개인화 답변
안될공학 - IT 테크 신기술	모르고 쓰면 개인정보 털려... 개발자 사이에 제2의 딥시크라는 AI 앱 혁명 MCP, 보안 유지하려면 어떻게?
안될공학 - IT 테크 신기술	모르면 실시간 손해? MCP, 딥시크 속도로 빠르게 확산 | 무료 AI 앱 폭발하게 된 MCP, 클로드 커서ai 동반 폭등 | 진짜 에이전트AI 시대 | 샘알트먼 급하게 지원 약속
미래채널 MyF	휴머노이드 로봇 시대에 생겨날 창업 아이템들
미래채널 MyF	1~2년 내에는 (거의)완전자율주행으로 고향 가겠는데요?
미래채널 MyF	미국 Vs 중국 | 휴머노이드 산업의 승자는 누가 될까?
미래채널 MyF	미래형 신사업 아이템 모음 1편 (비개발자도 도전 가능)
미래채널 MyF	미래형 신사업 아이템 모음 2편 (자동차, 로봇 산업도 포함)
미래채널 MyF	생성형 AI로 훨씬 똑똑해지는 Amazon Alexa+ 스마트홈
밥상차려주는남자	오늘 완벽한 쭈꾸미볶음 알려드릴테니 똑같이 만들어 보세요!!
밥상차려주는남자	얼큰한 소고기배추국 이거 한냄비면 추운겨울 걱정 없어요 👍
밥상차려주는남자	[매콤 오징어잡채] 인정할 수 밖에 없는 탱탱함!! 불지 않아요 물생기지 않아요

채널명	콘텐츠명
밥상차려주는남자	이렇게만 따라서 만들면 무조건 인정💘정말 맛있는 순두부찌개 황금레시피
밥상차려주는남자	생선조림 이렇게만 따라서 만드세요!! 어떤 생선이든 100% 성공✨
Dmonk	2025, 4월의 주목해야 할 신작 게임 5개+ : '코만 도스' #신작게임추천
Dmonk	2025, 3월의 주목해야 할 신작 게임 5개 : '물러설 곳 없는 암살자' #신작게임소개 #게임5
Dmonk	익숙하지만 새로운 맛, 그 반대인가? 메타포 : 리판타지오 리뷰
Dmonk	GAME 5: 1월의 주목해야 할 신작 게임 5개 Vol.61.2024 #game5 #게임5
정세월드	JP 일본여행	사막에서 연 날리는 장인들의 소도시 😮 시즈오카
정세월드	도쿄 MZ들의 떠오르는 여행지. 올해 가거나 가지마세요 (내년부터 붐빌 예정)
정세월드	JP 일본생활	도쿄 봄여행? 여긴 꼭 걸어봐요
정세월드	JP 일본도쿄	맨날 같은 일정은 그만~ 도쿄 11년차 추천 코스
돌봄개린이집	강아지 산소방 제발 아무 강아지나 사주지마세요. 꼭 필요한 반려동물만 구매하는 '오투맥스 산소발생기'
돌봄개린이집	강아지 초기 사회화시기 놓쳤을때 어떻게 해야할까?
돌봄개린이집	강아지 식분증 원인과 해결방법 (똥 먹는 강아지)
돌봄개린이집	강아지 오줌지옥에서 탈출하는 배변훈련의 모든것
돌봄개린이집	강아지에게 주인으로 인정받는 방법
돌봄개린이집	새끼 강아지 입양시 주의사항 (2개월 강아지)
돌봄개린이집	그동안 강아지가 말을 안들은 이유. 이제 이렇게 하면 말을 잘 듣게 됩니다.
돌봄개린이집	강아지 산책 할때 줄 당기지 않게 하는 방법
돌봄개린이집	포메라니안과 함께 살기위해서 알아야 하는것. 성격, 털빠짐, 사회성 등.
돌봄개린이집	강아지가 어릴때 반드시 해야하는 교육 3가지. 새끼 강아지 훈련
돌봄개린이집	강아지 산책훈련 할 때, 100명중 95명이 실수하는 행동 4가지. 산책훈련 꿀팁!
돌봄개린이집	강아지 교육의 핵심, 안돼를 가르치는 방법
돌봄개린이집	반려견 키우기도 템빨!! 강아지 용품 추천 4가지, 직접 사용해보고 추천합니다.
우리집베리-반려견 정보 채널	아기 강아지가 계속 문다면? 초간단 입질 교육, 새끼 강아지 무는 버릇 고치기
우리집베리-반려견 정보 채널	아기 강아지 무엇을 가르쳐야 할까? 새끼 강아지 사회성 교육, 4개월 미만 강아지 사회화 훈련
우리집베리-반려견 정보 채널	산책 줄 당기는 강아지 편하게 산책하는 법! 강아지 산책 교육, 반려견 산책 훈련
우리집베리-반려견 정보 채널	강아지에게 절대 하면 안되는 보호자의 나쁜 행동 5가지, 강아지 성격 나빠져요, 강아지 분리불안 원인
우리집베리-반려견 정보 채널	절대 하지마세요! 강아지 분리불안의 원인입니다
우리집베리-반려견 정보 채널	강아지 배변훈련 100% 성공하는 법! 2개월 강아지 배변교육
우리집베리-반려견 정보 채널	강아지 성질머리 테스트? 만지면 강아지가 싫어하는 부위
우리집베리-반려견 정보 채널	강아지가 우울할 때 하는 행동, 강아지가 우울증 걸렸을 때
숏다	모동숲에서 가장 빠르게 돈 버는 방법 2가지! 비교&꿀팁💰

채널명	콘텐츠명
숏다	호주에서 만든 모동숲?! 딩컴 해봤습니다
숏다	초보도 쉽게 따라하는 섬꾸미기✨ 비행장 입구 완벽 리모델링!
숏다	K-심즈?! 핫한 신작 게임! 인조이 사전 플레이 해봤습니다 [inZOI]
숏다	원하는 주민 데려오는 법부터 이사노가다까지 싹- 다 알려드림🔥모동숲 초보 가이드
썸니아마크라메somnia_macrame	[SUB]내가 만든 쿠키이이이🍪 아니고 캔디도어벨🍭🔔 : macrame candydoorbell:[DIY Tutorial]
썸니아마크라메somnia_macrame	[SUB]마크라메 입체 하트 키링♡❤🖤: macrame 3D heart keychain:[DIY Tutorial]
썸니아마크라메somnia_macrame	[SUB]마크라메 걱정인형 키링 만들기👌: 원숭이 주먹 매듭: Macrame worry doll keychain: Monkey's fist knot [DIY Tutorial]
썸니아마크라메somnia_macrame	[SUB]마크라메를 처음 시작하는 분들을 위한 기본 매듭 강의♥️첫번째 시간 : (종달새 머리 매듭,평매듭,평매듭응용) :Macrame Basic Knot:Square Knot
지식브런치	AI와 인간, 진짜 닮았을까?
지식브런치	CCTV가 많으면 더 안전해질까? 눈앞에 다가온 '디지털 전체주의'
지식브런치	"현실 세계는 과잉보호, 가상 세계는 과소 보호" - 디지털은 어떻게 아이들을 병들게 하나?
지식브런치	달에 못 간지 50년, 왜 다시 가려는 걸까?
지식브런치	사람들은 왜 가짜 뉴스에 쉽게 빠져들까? 진실보다 믿음이 중요한 Post-Truth 시대가 왔다
맨날 수리야	AI 편집기 UPDF로「명령」해보세요! (AI 요약, 채팅, 번역, 편집, 보안에 마인드맵까지??) PDF 파일 있다면 꼭 써보세요~!
맨날 수리야	유튜버인데 이게 없다고? 😅 필수 기능이 50개나 탑재된 역대급 프로그램...👍 (녹화, 편집, 자동자막, 용량압축, 자동썸네일, 워터마크 및 배경제거 등)
맨날 수리야	유튜버를 꿈꾼다면 이 프로그램 꼭 써보세요! (PC화면 녹화 + 영상편집 +웹캠 +자동자막 + 각종 효과 등 이거 하나로 끝~) 스트리머, 인터넷강의도 ok~
맨날 수리야	[2024년 최신판] 윈도우10 무료 설치 방법 (USB만 있으면 초딩도 따라합니다~)
맨날 수리야	[2025년 최신판] 윈도우11 무료 설치 방법, 이 영상 하나로 끝내세요 (2만원은 그냥 법니다~)
IT's okay 잇츠 오케이	노트북도 예쁠 수 있잖아
IT's okay 잇츠 오케이	(이벤트) 인공지능학과가 본 애플 인텔리전스 vs 갤럭시 AI
IT's okay 잇츠 오케이	노트북마다 붙어있는 이 스티커, 도대체 무슨 뜻임?? (부제 : 갤럭시북에서 에어드랍 쓰는 법)
IT's okay 잇츠 오케이	직장인들, AI '이렇게만' 활용해보세요 ㅣ 업무효율을 300% 높여줄 18가지 AI툴 활용법 ㅣ (챗GPT, 노션AI, AskUp)
커피하는 람쥐	같은 레시피인데 왜 맛이 다를까? 답은 '물줄기'에 있습니다
커피하는 람쥐	아이스크림 안 팔면 손해예요! 민트라벨 아이스크림 파우더 찐후기?!
커피하는 람쥐	핸드드립 2인분 내릴 때, 그냥 물 2배 때리시면 안됩니다
커피하는 람쥐	초코라떼 만들기 전에 반드시 시청하세요. 초코소스, 파우더 5종 제대로 비교해봤습니다
커피하는 람쥐	탬핑에 따라서 커피 맛이 극명하게 아예 달라집니다
오늘식탁 - Today Table	해물찜 3분컷 대박집 황금레시피 이렇게 만들면 식당처럼 맛있게 만들수있어요!
오늘식탁 - Today Table	카레 전문점보다 더 맛있게 만들수 있어요 '이것'만 준비해주세요
오늘식탁 - Today Table	집에 있는 기본재료로 만드는 단짠단짠 꽃게무침 비법
오늘식탁 - Today Table	스푼이면 인생 깍두기 만들수있어요 누구나 만들수있는 전문점 섞박지 김치
오늘식탁 - Today Table	아직도 쥬스와 김치양념을 한곳에 갈고 계신가요? 가성비 요리전용 무선 핸드블렌더가 궁금하다면-드웰러 에어스틱 공동구매

채널명	콘텐츠명
오늘식탁 - Today Table	초간단 단호박죽 꼭 이것으로 만들어 보세요! 인생 단호박죽 완성
오늘식탁 - Today Table	열무를 꼭 치데주세요 부드럽고 아삭하게 드실수잇어요초보자도 쉽게 만드는 열무물김치 열무김치
미노엔	게임에서만 가능한 스토리텔링은 없을까 (스포있음) [미노엔의 게임독학]
미노엔	닌텐도에 미친 사람의 닌텐도 스위치 2 다이렉트 후기 [미노엔의 데이원리뷰]
미노엔	즐거운 파워업을 만드는 방법 [미노엔의 게임독학]
미노엔	레이튼 교수와 이나즈마 일레븐과 요괴워치와 골판지 전기를 가지고 망해버린 회사 [미노엔의 겜관고]
축구보는기자	준비된 승리, 예고된 득점' 라인을 내려도 인천에게 뚫릴 수밖에 없는 이유 (부천 v 인천 리뷰)
축구보는기자	멀티골 손흥민' 비길 경기를 승리로 바꿔버린 손흥민의 미친 골결정력 (호펜하임 vs 토트넘)
축구보는기자	욕설 나오는 전술변화' 리그 19위 공격력의 팀한테 전반전에만 3골을 쳐먹힌 이유
축구보는기자	손흥민을 풀백 발사대로?' 손흥민도 희생, 경기력도 희생, 성적도 희생되고 있는 이유
축구보는기자	헌신적인 주장' 손흥민의 전방압박이 토트넘에게 중요한 이유
축구보는기자	고립된 손흥민' 토트넘의 공격이 답답하게 흘러가는 이유
축구보는기자	선수들이 기회를 못 살려?' 손흥민, 이강인, 황희찬이 있어도 공격이 답답한 이유
패션디자이너 힙토리	겨울 포인트 컬러코디🎨 센스가 배가되는 컬러코디 공식👩🎨	겨울에도 칙칙하지 않고 화사하게!
패션디자이너 힙토리	2024 FW 패션 트렌드 총정리	컬러, 소재, 아이템 깔끔 정리! 10가지 이것만 기억하세요!🏵️🍃🤍
패션디자이너 힙토리	여름 클래식 코디🖤 실패 없는 6가지 코디공식! 새 옷 사기 애매하죠?! 그렇다면 클릭! + 올해 트렌드 코디팁까지!📌
패션디자이너 힙토리	2024 SS 패션 트렌드 총정리	SS 2024 Fashion Trend
패션디자이너 힙토리	쉽지만 정말~세련된! 클래식 컬러코디 방법 4가지!🎨 유행 없이 언제든 꺼내어 봐주세요!
40대 여자 꾸미는 방법	이거 전부 7만원 실화?	스타일리스트 테무 쇼핑 금지령?
40대 여자 꾸미는 방법	2025 SS 패션 컬러 & 트렌드 봄 쇼핑 가기 전 필수 시청!!
40대 여자 꾸미는 방법	2025 Fur 퍼자켓 힙하게 입는 방법: 아줌마 처럼 입지 않기!
40대 여자 꾸미는 방법	겨울코트 사러 가기 전 필수 시청!! 2024-25 겨울코트 트렌드와 코트 코디
미미미누	"나는 이 정시 지원을 해봤어요!" 조기 발표 기다리면서 보는 2025대입 정시 지원 경쟁률 리뷰
미미미누	"수능에서 과학Ⅰ 선택했는데, 이거 맞아요?" 사탐런이 현실화된 2025수능 원서 접수 결과 리뷰
미미미누	"2025수능 D-4" 수능 4일 전 2025수능 수험생 유의사항 리뷰해보았습니다
미미미누	"아니 올해 수시 경쟁률 왜 이래요?" 882:1의 경쟁률과 1:1의 경쟁률이 공존하는 미쳐버린 2025대입 수시 경쟁률 리뷰
미야옹철의 냥냥펀치	고양이 털 관리, 고양이에겐 고문일지 모릅니다
미야옹철의 냥냥펀치	20분 순삭! 코숏을 키우고 있다면 꼭 알아야 할 충격적인 사실
미야옹철의 냥냥펀치	마중 나오는 유형으로 알아보는 고양이 속마음
미야옹철의 냥냥펀치	집사에게만 있고 고양이에게는 없는 감정
미야옹철의 냥냥펀치	사냥놀이 학원이나 클래스가 있으면 수강하고 싶을 정도입니다 [실전편]
시온	게임계 끝판왕들이 온다 ! 곧 출시되는 차세대 신작 게임들 총정리
시온	중세시대 공성전을 가장 실감나게 구현한 게임..
시온	진정한 게임의 해가 찾아온다. 2025년 신작 게임 총정리 !

채널명	콘텐츠명
빛오 BITOH	봄 데일리룩으로 딱ㅛ ㄱ| 꾸안꾸부터 꾸꾸꾸까지 말아온 무채색 코디ㅓㅓㅓㅓㄴ ˇ ノ♡ Spring Fashion Haul 11 ITEMSㅜ
빛오 BITOH	***HAUL***📦] 이제 봄 준비 해야지•• ? 봄 신상 택배 언박싱 💐 ▨ ✨[쓰리타임즈•보헤미안서울•시너진•더콜디스트모먼트•허그유어스킨•엘리오티]
빛오 BITOH	흔하지 않고 힙하게 일주일 코디하기GB❤🏷 디테일 끝판왕 등장.. 봄 맞이 패션룩북 🤘 ▨[봄코디•무채색코디•데일리룩•스키즘인듀싱]
빛오 BITOH	무채색 겨울코디 갓성비편.□10+ ☼만원대 아우터부터 휘뚤마뚤 이너랑 가방까지 무신사 블프 준비완료ㅛ ㄱ| [헬레네파리스•아우터•가성비룩북•이너추천•보부상백]
햄튜브	할아버지가 주인공인 가십걸
햄튜브	그게 뭔데 십덕아 😳
햄튜브	초반은 '커피토크' 중반은 '탈주' 후반은 '용과같이'인 드라마
햄튜브	이것저것 전부 리뷰합니다 들어오세요!
햄튜브	인간이랑 똑같이 생긴 외계인이 지구를 침공했다?!
미국주식으로 은퇴하기 - 미주은	[오늘의 미국주식뉴스] 하락장이 힘겨운 투자자분들을 위한 미주은의 메시지 / 트럼프의 갑작스러운 태도 전환으로 선물 지수 급등! / 테슬라, 최악의 어닝에도 급등하는 이유
미국주식으로 은퇴하기 - 미주은	[오늘의 미국주식뉴스] 스태그플레이션 발생 확률 상승! 미주은 투자 전략 재점검 / 트럼프는 파월 의장을 정말 해임할까? / 지난 주 미국 증시를 지배했던 4가지 뉴스
이기주의 스케치 KEEZOO sketch	눈 쌓인 풍경을 PEN으로만 그리는 법	눈 내린 숲과 집이 있는 풍경
이기주의 스케치 KEEZOO sketch	(수채화) 나무와 색이 가득한 가을 풍경 그리기	누구나 쉽게 하는 수채화 기법 |색을 쓰는 법
안스타	커피 '이것'만 알고 드신다면 훨씬 맛있게 즐기실 수 있습니다
안스타	입문자를 위한 1인분 핸드드립 레시피
안스타	출시만 하면 품절이라는 화제의 이마트 원두, 정말 맛있을까?
안스타	스타벅스부터 저가 프랜차이즈까지 모든 커피 가격이 오르는 이유
약사 이진수	쉽지 않은 영양제 이야기 - 신개념 미백 영양제 비라이트™에 대해 알아보자!
약사 이진수	쉽지 않은 피부 관리! 피부 전문 약사가 알려주는 증상별 맞춤 성분들! (우레아 크림, 티트리오일, 멜라토닌)
약사 이진수	약국의 주인은 오로지 약사인 이유?
약사 이진수	쉽지 않은 약사 생활 - 약국 약사 1년 후기
약사 이진수	쉽지 않은 잇몸 질환! - 약국에서 살 수 있는 잇몸약은?
삼미노_삼성라이온즈 이야기	강민호 뒤를 이을 포수는 누구일까? [삼성라이온즈의 백업포수들]
삼미노_삼성라이온즈 이야기	오늘 삼성라이온즈는 이길 생각이 없어 보였습니다. [5/7 경기 리뷰]
삼미노_삼성라이온즈 이야기	삼성라이온즈의 현실, 최대한 객관적으로 분석해보았습니다.
삼미노_삼성라이온즈 이야기	삼성라이온즈 야구.. 도대체 왜 이럴까..
삼미노_삼성라이온즈 이야기	어린 사자들의 진짜 프로야구 커리어는 이제부터 시작입니다.
삼미노_삼성라이온즈 이야기	올시즌 김성윤 무엇이 달라졌을까? [5툴플레이어, 낮아진 스윙적극성?]
바둑에듀	[바둑 입문 #22] 판 전체를 휘감는 축의 묘수를 찾아라! - '명황유월궁세'의 판축!

채널명	콘텐츠명
바둑에듀	[바둑 명국 리뷰] 영화 '승부'의 엔딩 명국, 조훈현 九단 vs 이창호 九단의 대결!
바둑에듀	[바둑 입문 #1] 단 10분 만에 바둑의 기초 배우기!
바둑에듀	[바둑 실력 향상 사활] 이 문제를 자신 있게 풀 수 있어야 진정한 수읽기 고수! (feat. 백돌을 살리는 묘수는?)
바둑에듀	[바둑 실력 향상 사활] 바둑 중급자 약 70%가 두 손을 들고 마는 문제입니다. 30초 안에 백돌을 잡는 수를 찾아보세요.
말년 호빙요	[분석] 라이스가 잉글랜드 최고 미드필더인 이유
말년 호빙요	[분석] 요즘 축구 보는 재미가 떨어진 이유
말년 호빙요	[분석] 로컬 아놀드가 리버풀을 떠나는 이유
말년 호빙요	[분석] 포르투갈은 호날두 없이 더 잘할까?
말년 호빙요	[이슈] 이제 손흥민은 대표팀에 큰 도움이 되지 못할까?
말년 호빙요	[리뷰] 예견된 실점 장면... 왜 우리나라만 뚫지 못할까...?
ssin 씬님	[ENG] 유행은 돌아온다! 2025년 뷰티 트렌드 ☑ 제가 감히 예측 해 봅니다.
ssin 씬님	[ENG] 이 클린🌱비건 시국에 존재하는 동물성 화장품을 찾아보자! 🐟🐢🐌🦍🪰
ssin 씬님	[ENG] 요즘 많이 달라진 씬님 메이크업💧오랜만에 보여드릴게요 (얼태기, 머태기 극복!)
ssin 씬님	[ENG] 요새, 쿠션 좀 괜찮은거 뭐있어?🧐2024 하반기 신상 쿠션 리뷰! 밸런스 좋은 올리브영 쿠션
김진짜 Real KIM	펩이 축구를 망쳤다는 말이 나오는 이유
김진짜 Real KIM	일본은 되고 한국은 안 되는 이유
김진짜 Real KIM	홍명보가 이정효보다 나을까?
김진짜 Real KIM	음바페가 부진하다가 갑자기 부활한 이유
김진짜 Real KIM	손흥민, 토트넘 떠나면 어디로 갈까?
새벽의 축구 전문가	"토트넘 완벽한 승리" 손흥민 결승전 간다! 드디어 찾아온 우승의 기회 (토트넘 vs 보되글림트)
새벽의 축구 전문가	"아스날이 더 수준높은 팀이었다" 아르테타 망언 인터뷰, 비판 받는 아르테타의 경기 운영능력
새벽의 축구 전문가	"여전히 훈련 못한다" 손흥민 결승전은 출전 가능할까? 4강전은 출전불가 확정
새벽의 축구 전문가	"PSG 결승전 진출" 로테이션 안했던 아스날, 무기력하게 챔스 탈락! 무관으로 끝난 아스날의 시즌
주씨 [Juicy]	[러너들 주목!] 단 10분으로 앞으로의 러닝이 달라집니다
주씨 [Juicy]	국가대표 물리치료사가 알려주는 무릎 운동ㅣ집에서 무릎 셀프 재활 🦴
주씨 [Juicy]	러닝워치 입문기 🏃▨♀🏃▨♀□ 가민포러너165 ㅣ 애플워치에서 가민워치로 💨💨
주씨 [Juicy]	러닝 Q&Aㅣ 보폭 늘리는법, 척추측만증, 마라톤
주씨 [Juicy]	러닝 좀 한다 ?!! 하시는 분들 꼭 보셔야 합니다.
주씨 [Juicy]	드릴운동, 왜 해야할까? ㅣ 선출이 알려드림
방수좋아 당구TV	세상에서 가장 빠르게 최대각을 최대한 쉽게 찾는법😄 ㅣ 뒤돌리기,옆돌리기,뱅크샷 가능 ㅣ 긴각출발
방수좋아 당구TV	앞돌리기 짧게 칠 때 가장 쉬운 설계법👍 ㅣ 기준 응용 모든게 너무 간단하다 ㅣ 45도 활용법
방수좋아 당구TV	뒤돌리기 길게치는 방법! ㅣ 영업비밀 까발립니다🚫 ㅣ 이 기준이 당구실력 올려줍니다↗
방수좋아 당구TV	이 영상은 조만간 유료전환 됩니다👆 ㅣ 앞돌리기 1타 강의 레슨 ㅣ 어렵게 접근하지 마세요✋

채널명	콘텐츠명
이상그이상	플스 6월 신작 총정리 ㅣ 총 18작품 ㅣ 데스 스트랜딩 2, 듄 어웨이크닝 등
이상그이상	닌텐도 스위치 세일💧ㅣ 할인 게임 소개 & 추천 게임 등 13가지!
이상그이상	메타크리틱 선정! 최고의 게임 퍼블리셔 TOP 10
이상그이상	플스 5월 신작 총정리 ㅣ 총 21작품 ㅣ 엘든 링 밤의 통치차, 둠 더 다크 에이지스 등
엄마의 손맛	부추전 힘들게 하지마세요✔️이 방법은 맛있는 부추전 만들기 1등 레시피입니다
엄마의 손맛	감자 힘들게 먹지마세요✔️이 방법은 맛있는 감자반찬 만들기 1등 레시피입니다
엄마의 손맛	알배추 힘들게 먹지마세요✔️ 방법은 맛있는 알배추 반찬 만들기 1등 레시피입니다
론나 Ronna	샘소나이트, 포터 말고 예쁜 남자가방 추천💼
론나 Ronna	빈티지 파타고니아를 사야하는 이유
론나 Ronna	센스 있는 양말 14켤레 추천
론나 Ronna	보풀 쩌는 니트/코트 깔끔하게 관리하는 방법
오토 플랜트	나의 방이 정말 마음에 든다 ㅣ 직접 만든 토분 ㅣ 도색 DIY ㅣ 요즘 예쁜 희귀 식물들 #10
오토 플랜트	식친에게 받은 선물들 ㅣ 식물 근황 예쁜 식물들 보고서 ㅣ 초보자를 위한 팁 과습
오토 플랜트	예쁜 꽃 보고 가세요 호야 수퍼에스키모 꽃이 폈어요 안스리움 비타리폴리움 꽃도~*^^* 요즘 예쁜 식물들 #8 ㅣ 온실 구경 신엽 파티
양빵당구	💥123만 알면💥 모든 원뱅크는 끝! 접시, 구멍, 걸어치기 모두 가능한 양샘이 만든 대박 시스템 / 큐선으로 겨냥점 빠르고 정확하게 찾기 / 보기만해도 수지업↑↑
양빵당구	큐 잡는 법 때문에 실패가 많다💥 그립의 변화로 큐질이 달라진다!!
양빵당구	💥 두께💥에 관한 모든 것! 이 영상 하나로 마스터하세요~
양빵당구	⭐1뱅크 장인들의 설정법⭐ 구멍치기, 되돌아오기 접시를 자유자제로 칠 수 있는 초간단 공략법 / 두께와 스트록은 고정, 당점만 결정
이꾸소	다이소템으로만 요즘 유행하는 다꾸 해보기 (샤오홍슈)
이꾸소	이게 요즘 유행이라며? 샤오홍슈 다꾸
이꾸소	야구덕후들은 다이소에 당장 뛰어가셈 ⚾□다이소 신상 리뷰
안녕영경 deeptwinkles	이거 너무 쉬운데..? 짧은손톱도 힙할수있다 🍭✨ 간단 자석치크네일 셀프로하기💅▨ㅣ초간단 치크네일 꿀팁ㅣ메이유어 실루엣 자석젤ㅣ셀프네일 가성비 시럽젤 추천💜
안녕영경 deeptwinkles	다이소 젤네일로 소녀st 엔젤코어 네일해보기ʚɞ·₊˚l 볼체인❌ 엔젤날개 만들기.다이소 젤네일 솔직리뷰ㅣ짧은손톱ㅣ숏톱네일 ㅣ다이소 셀프네일
안녕영경 deeptwinkles	취미에 진심인 초보다꾸러의 서일코 언박싱+다꾸 영상˚₊·꒰ა □ ໒꒱ ·₊˚ 덤의 축복이 끝이 없다..ᄃ 다꾸입문템?! 페어에 가는 이유? 다꾸는 템빨ᄅ
임뻠임	🐶강아지와 외출할 때 삶의질상승 시켜주는 꿀템 모음집!!!
임뻠임	🐶 집에서 손쉽게 만드는 강아지 간식, 화식 만들기 몰아보기.zip📁
임뻠임	🐶 간단하고 손쉽게 강아지 화식 만들기ㅣ황태국ㅣ미역국ㅣ삼계죽 🥘
임뻠임	🐶동물병원, 강아지미용샵에서 실제로 사용하는 강아지용품 추천템!!!🧻🛁✂□ 눈물자국ㅣ목욕ㅣ미용
소박sobac	구성은 더 알차고, 디자인은 간결해진 2025 올인원 다이어리 + 무료 가이드북 + 부록 스티커
소박sobac	심플하지만 알찬 2025 소박 다이어리 무료 공유, 2024년 12월 포함, 깨짐없는 굿노트 스티커 90종 😘
소박sobac	32종 템플릿 x 16색상✨ 아이패드 굿노트 필기노트는 이걸로 끝내세요😉

채널명	콘텐츠명
소박sobac	하루 일과 어떻게 기록하세요? 데일리 플래너를 작성하는 4가지 방법
윈터제뉴어리 winterjanuary	그럼에도 불구하고, 때로는 경험하지 않았으면 좋았을 것들이 내게 일어나서 다행이다 (사람과 강아지의 근황 일기 2)
윈터제뉴어리 winterjanuary	체력은 소진되었으나 마음은 충전되었습니다🔋 유기견 임시보호가 꼭 필요한 이유, 이별 파티로 바쁜 E 강아지와 그냥 I 멍집사의 힐링 일상 브이로그
윈터제뉴어리 winterjanuary	강아지 패브릭 용품 추천템 소개: 방석, 펫타올, 넥카라, 산책 가방, 산책 매트
윈터제뉴어리 winterjanuary	강아지 패딩, 우비, 기본 티셔츠, 해충방지 쿨티🐶 강아지 패션하울? 고르고 고른 사계절 옷! 가성비 기능성 디자인 모두 고려한 댕댕이 옷장 속 찐 추천템
요리왕비룡 Korean Food Cooking	이거 모르면 후회해! 매콤한 오뎅볶음 이렇게 만들면 밥반찬으로 최고입니다
요리왕비룡 Korean Food Cooking	무르지 않고 더욱 아삭한 오이소박이 정말 너무 쉬워요~!
요리왕비룡 Korean Food Cooking	제철 마늘쫑은 절여요. 멸치볶음 이렇게 만들면 최고의 반찬입니다.
요리왕비룡 Korean Food Cooking	열무무침 이것을 넣으면 맛과 영양이 최고인 반찬입니다
요리왕비룡 Korean Food Cooking	소금에 절인 채로 두면 오이가 빠르게 물러져요. 이렇게 절여보세요. 대박집 오이김치 만드는법
요리왕비룡 Korean Food Cooking	오늘 여름 오이지는 이렇게 담궈서 시원하게 보내세요. 뚝딱 만들어요~!
요리왕비룡 Korean Food Cooking	🔔감자채볶음 어렵지 않아요! 이렇게 볶으면 쫀득하고 맛있습니다.
오징어약사TV - 가장 쉬운 건강채널	뱃살 나온 사람들의 당뇨 탈출법. 오히려 더 쉽습니다!
오징어약사TV - 가장 쉬운 건강채널	당뇨에 좋은 채소는 따로 있다는 거 아셨나요? 의외로 혈당 폭발시키는 음식. 당뇨에 최악과 최고의 채소
오징어약사TV - 가장 쉬운 건강채널	췌장을 되살려 당뇨를 탈출하는 영양소, 아직 늦지 않았습니다. 한국인 당뇨 특화
오징어약사TV - 가장 쉬운 건강채널	(2025당뇨06) 매일 식전에 드시면 혈당 사라지는 당뇨 탈출 마지막 퍼즐 한 조각.
범지 Bum G	결혼식 완료💍👰🏻‍♀️🤵🏻‍♂️👰🏻‍♀️따꾸하면서 풀어보는 결혼 준비+결혼식 썰👀💬
범지 Bum G	물욕 없던 나날에도 구매한 것들?!💚 작지만 알찬 하울🐰(스티커, 노트, 라벨리필, 북파우치, 무지 잠옷)
범지 Bum G	스크랩 다이어리 꾸미기✂️최근 소비한 것들을 기록하면서 건강과 운동에 관한 토크^.^💬
생선머가리	리니지2M 중국 진출?ㄷㄷ 실화임??
생선머가리	아니 서든어택 후속작이 또 나온다고??!
생선머가리	국내 게임사가 모바일에서 PC로 갈아타고 있는 이유는 무엇일까?
생선머가리	이번 달 신작들 와꾸가 살벌한데?! 6월 모바일게임 신작 [모바일게임 추천]
생선머가리	회사를 되살리기 위해 미쳐 날뛰는 엔씨 근황 (Feat. 아이온 2)
생선머가리	텅 빈 로스트아크 모바일 프리뷰 / 혐한 베데스다, 지역락 풀었다?!
혜다다 베이킹 유치원 ᵒ,,	겉바속쫀 다음날 먹어도 맛있는 휘낭시에의 비밀! 이렇게 만들어보세요!
혜다다 베이킹 유치원 ᵒ,,	친구가 사온거 아니냐고 물어봐요! 정말 맛있는 로투스 브라우니 치즈케이크 만들기

채널명	콘텐츠명
혜다다 베이킹 유치원 ෆ,,	아무래도 어마무시한 스콘 레시피를 만들어버린 듯 하다.. 딸기 크림치즈 스콘 만들기 🍓
혜다다 베이킹 유치원 ෆ,,	아무래도 어마무시한 스콘 레시피를 만들어버린 듯 하다.. 블루베리 크림치즈 스콘 만들기 🫐
혜다다 베이킹 유치원 ෆ,,	겉바속촉 맛있는 생크림 스콘 만들기 / 초보도 성공하는 쉬운 스콘레시피 (손반죽)
혜다다 베이킹 유치원 ෆ,,	한가지 반죽으로 여섯가지맛 버터쿠키 만들기 / 크리스마스 버터쿠키 실패 없이 만들어봐요!
혜다다 베이킹 유치원 ෆ,,	바삭쫀득한 바통 휘낭시에 레시피와 예쁜 포장방법까지🎁 한 큐에 알려드립니다😭
푸우형	7년 동안 강아지 물그릇 수십개 써보고 알게된 것 (켁켁거림, 음수량부족 해결)
푸우형	지금까지 먹여본 개껌 중에 최고입니다. 양치 귀찮을 때 이렇게 해보세요
푸우형	써보고 정말 만족한 다이소 강아지 꿀템 8가지 (보이면 꼭 사세요)
푸우형	이게 애정표현이라고? 강아지가 주인을 사랑한다는 신호 5가지
푸우형	혀가 나오면 위험한 신호라고? 강아지가 힘들 때 하는 행동 7가지 (반드시 알아두세요)
푸우형	7년째 하루도 빠짐없이 해주고 있는 4가지 (강아지가 무조건 행복해집니다)
푸우형	이건 믿고 써보세요! 수백만원 쓰며 찾아낸 최고의 강아지용품 5가지 (어디서 샀냐고 다 물어봄)
푸우형	강아지에게 정말 위험한 가전 4가지⚡□
푸우형	목줄 하네스 도망가는 강아지? 착용방법 꿀팁 (이렇게 하면 강아지산책이 시작부터 달라집니다)
푸우형	강아지 사료 설마 그냥 주세요? 이렇게 주니까 엄청 잘먹어요!
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	반려견 다이소 꿀템 5가지 😎 강아지 아토피에 도움된 용품 추천👍
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	산책 못 나가는 날, 강아지🐶와 집에서 재밌게 놀기 좋은 방법 4가지
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	강아지🐶 수명을 줄이는 잘못된 습관 4가지, 지금 바로 멈추세요!✋
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	강아지🐶가 더 건강해지는 음식 7가지, 이건 꼭 기억하세요!📝
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	강아지에겐 독! 절대 주면 안 되는 음식 7가지
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	반려견 수제 간식으로 바꾼 이유🐶 아토피 강아지를 위한 수제 간식 만들기
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	애견미용 맡기기 전, 보호자가 반드시 알아야 할 4가지🐶✂□
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	강아지 용품 베스트 5👀 견주 인생 10배 편해집니다
멈추개! I 반려견 보호자의 모든 고민 멈추개 !	강아지 발 핥기… 스트레스? 질병? 진짜 원인 공개! 강아지🐶 발 냄새의 정체는??
모리츠TV	[게임추천] 2025년 하반기 출시 예정 국내 MMORPG ㅣ 1황 게임은 과연 무엇이 될까? (25.7.13)
모리츠TV	[스텔라 블레이드] 100시간 플레이 솔직 후기 ㅣ K-게임의 희망!! 후속작이 기대되는 Stellarblade Review
모리츠TV	검은사막M] 아크매지션 개선 패치 이후 클래스 리뷰 ㅣ 두 가지 전승 포함 ㅣ 직업 추천 ㅣ ARCHMAGICIAN, BlackDesertMobile ,BDM (25.1.29)
모리츠TV	검은사막M] 쿠레나이 두가지 전승 스킬 트리 및 최종 리뷰 고정 댓글 확인 ㅣ 직업 추천 ㅣ KURENAI, BlackDesertMobile,BDM (24.10.23)

채널명	콘텐츠명
모리츠TV	검은사막M] 신규,복귀 가이드 ㅣ 시즌 졸업 이후 ㅣ 전투력 42.500이 중요한 이유 ㅣ BlackDesertMobile, BDM (24.8.23)
게임잡지 GTOPIA	2탄) 그래픽 실화? 차세대 그래픽, 언리얼 엔진 5 기반 기대 신작게임 TOP 10! ㅣ BIGGEST NEW Unreal Engine 5 Games TOP 10
게임잡지 GTOPIA	어때? 그래픽 좀 쩔지? 2026년 출시될 언리얼 엔진 5 기반 기대 신작게임 TOP 10! ㅣ BIGGEST NEW Unreal Engine 5 Games TOP 10
게임잡지 GTOPIA	잠자는 스위치2여, 이제 그만 일어나요~ 25-26 출시 예정인 스위치2 기대 신작들 TOP 15! ㅣ UPCOMING NINTENDO SWITCH2 NEW GAMES
미고랑 Migowith	미고 집사의 요즘 잘쓰는 아이템 ㅣ 고양이 용품 추천 ㅣ 미고랑 Migowtih
미고랑 Migowith	슈퍼 J 예비집사의 고양이 입양 준비기 ㅣ 고양이 필수 용품 추천 ㅣ 미고랑 Migowith
미고랑 Migowith	다이소 가성비 고양이 용품 추천 ㅣ 로봇청소기, 수반, 식기, 사료 및 습식 보관용품 ㅣ 미고랑 Migowith
김원상의 당구강좌	독학당구 기초편 - 1쿠션 득점법 (응용문제 풀이)
김원상의 당구강좌	독학당구 기초편 - 1쿠션 득점법
김원상의 당구강좌	독학당구 기초편 - 직접 맞히기의 핵심
김원상의 당구강좌	독학당구 기초편 - 두께에 따른 분리각 & 연습법
알린 ALINN	솔직히 GPT-5 보다 더 놀랐다.. 당신의 능률을 미친듯이 올려줄 '레전드' 기능
알린 ALINN	제발 돈 낭비는 그만.. AI에 500만 원 쓴 사람이 알려주는 AI 선택 가이드
알린 ALINN	AI를 사용하기 전에 보면 인생이 바뀌는 영상
알린 ALINN	요즘 챗GPT보다 더 좋다고 난리 난 미친 가성비의 AI
테이스터 션 Taster Shawn	8~40만 원! 가격대별 경량 바람막이 4종 추천드립니다 (광고 X)
테이스터 션 Taster Shawn	호카,아디다스,나이키 보다 비싼 새티스파이 첫 러닝화 돈 값할까..??
테이스터 션 Taster Shawn	평생 모은' 셔츠 공개! 남자 셔츠 이걸로 종결시켜 드립니다. ㅣ 웨스턴,워크,밀리터리,레저,드레스
테이스터 션 Taster Shawn	자라 역대급 여름 세일! 🔥안사면 후회할 ZARA 필수템 만 콕 집어드립니다
테이스터 션 Taster Shawn	2만 원~30만 원 가격대별 반팔 헨리넥 추천!
피알남 피부과전문의 김홍석	뭘 발라도 건조하다면 이 방법대로 하세요 (1분도 안걸림) ㅣ 보습제 '이렇게' 발랐더니 오히려 속건조 심해지는 이유?! ㅣ 피부과전문의 추천 2가지 보습성분
피알남 피부과전문의 김홍석	피부 트러블케어 최고의 제품은? 최고의 진정케어 성분 4가지, 제품 9종 리뷰! (여드름, 민감피부 필수시청!)
피알남 피부과전문의 김홍석	비싼 화장품 말고 '이걸'로 물광피부 되세요!😭하루종일 촉촉함 유지되는 피부과의사 극찬템 '이 성분'으로 바꾸세요! (건성, 수부지 주목!) #스네일뮤신 #뮤신
피알남 피부과전문의 김홍석	선크림 '이렇게' 바르면 99%는 피부노화 쎄게 옵니다 ㅣ 전문의가 직접 발라보고 알려드리는 '선크림 잘 바르는 진짜 방법'

2.1.1. 주제별 · 제시 자료별 수집 결과

공적 독백을 위한 콘텐츠는 유튜브와 강연 등을 수집하였으며 구분별, 주제별 허락 동의를 받은 콘텐츠만을 수집하였다.

<표 41> 공적 독백 주제별 수집 결과

구분	순번	주제	목표 시간	가공 시간	비율
유튜브	1	뷰티/패션	5.0	4.8	12.54%
	2	게임	5.0	5.1	12.52%
	3	스포츠/건강	5.0	5.0	12.30%
	4	취미/여가	5.0	5.9	12.19%
	5	푸드/쿠킹	5.0	4.9	12.07%
	6	IT/기술/과학	5.0	5.1	12.53%
	7	동물/펫	5.0	5.4	13.07%
	8	교육/강의	5.0	4.8	12.78%
	합계		40	41	100%

구분	순번	주제	목표 시간	가공 시간	비율
강연 등	1	강연	15.0	15.7	24.90%
	2	강의	15.0	14.4	24.84%
	3	연설	15.0	15.1	24.95%
	4	발표	15.0	15.3	25.30%
	합계		60	60.5	100%

2.1.2. 인구분포별 수집 결과

총 152명의 공적 독백 발화자가 참여하였고 성별, 연령별, 지역별 분포는 다음과 같다.

<표 42> 공적 독백 성별, 연령별, 지역별 화자 모집 결과(단위: 명)

구분	10대		20대		30대		40대		50대		60대 이상		총합계	
	남	여	남	여	남	여	남	여	남	여	남	여	남	여
합계	0		24		65		29		8		26		152	
	0	0	15	9	34	31	22	7	7	1	25	1	103	49
구분	남	여	남	여	남	여	남	여	남	여	남	여	지역별	권역별
서울	0	0	2	4	13	12	5	1	4	0	10	1	52	89
인천	0	0	0	1	3	3	7	0	0	0	1	0	15	
경기	0	0	3	1	6	5	0	1	1	0	5	0	22	
강원	0	0	1	0	3	2	0	0	0	0	2	0	8	8
부산	0	0	0	0	1	1	1	0	1	0	0	0	4	24
대구	0	0	2	0	0	2	2	0	0	1	1	0	8	
울산	0	0	1	0	0	0	2	1	0	0	0	0	4	
경북	0	0	1	0	0	1	0	1	0	0	1	0	4	
경남	0	0	0	0	2	1	1	0	0	0	0	0	4	
대전	0	0	2	0	1	1	1	0	0	0	0	0	5	16
충북	0	0	0	0	3	0	0	1	0	0	1	0	5	
충남	0	0	1	2	1	0	0	1	0	0	1	0	6	
광주	0	0	0	0	0	2	1	0	1	0	0	0	4	15
전북	0	0	0	1	1	1	1	1	0	0	1	0	6	
전남	0	0	2	0	0	0	1	0	0	0	2	0	5	
제주	0	0	0	0	0	0	0	0	0	0	0	0	0	0
합계	0	0	15	9	34	31	22	7	7	1	25	1	152	152

2.1.3. 주제별 연령 분포

공적 독백 화자는 30대가 제일 많은 것으로 나타났다.

<표 43> 공적 독백 주제별 연령 분포(단위: 명)

주제	10대	20대	30대	40대	50대	60대 이상	합계	비율
뷰티/패션	0	2	3	2	1	0	8	5.26%
게임	0	2	5	1	0	0	8	5.26%
스포츠/건강	0	0	6	1	0	0	7	4.61%
취미/여가	0	2	4	1	0	0	7	4.61%
푸드/쿠킹	0	0	5	0	1	1	7	4.61%
IT/기술/과학	0	3	4	1	0	1	9	5.92%
동물/펫	0	0	5	3	0	0	8	5.26%
교육/강의	0	0	3	2	3	0	8	5.26%
강연	0	3	10	7	0	2	22	14.47%
강의	0	1	6	8	3	4	22	14.47%
연설	0	4	1	1	0	18	24	15.79%
발표	0	7	13	2	0	0	22	14.47%
총합계	0	24	65	29	8	26	152	100%

2.1.4. 주제별 성별 분포

본 사업에 참여한 공적 독백 화자의 성별 분포는 남성 103명, 여성 49명이다. 남성은 스포츠/건강, 교육/강의, 연설에서 비율이 높았고, 여성은 뷰티/패션, 취미/여가에서 비율이 높았다.

<표 44> 공적 독백 주제별 성별 분포(단위: 명)

주제	남성		여성		합계
뷰티/패션	3	37.50%	5	62.50%	8
게임	6	75.00%	2	25.00%	8
스포츠/건강	7	100.00%	0	0.00%	7
취미/여가	2	28.57%	5	71.43%	7
푸드/쿠킹	5	71.43%	2	28.57%	7
IT/기술/과학	7	77.78%	2	22.22%	9
동물/펫	3	37.50%	5	62.50%	8
교육/강의	7	87.50%	1	12.50%	8
강연	12	54.55%	10	45.45%	22
강의	17	77.27%	5	22.73%	22
연설	24	100.00%	0	0.00%	24
발표	10	45.45%	12	54.55%	22
총합계	103	67.76%	49	32.24%	152

2.1.5. 화자의 직업별 분포

참여한 공적 독백 화자는 ‘전문가 및 관련 종사자’가 가장 많았고 ‘기타’가 다음으로 높은 비율을 차지했다.

<표 45> 공적 독백 화자의 직업별 분포

직업	인원(명)	비율
전문가 및 관련 종사자	58	38.16%
사무 종사자	18	11.84%
서비스 종사자	7	4.61%
기술자 종사자(장치/기계 조작 및 조립 종사자)	2	1.32%
단순노무 종사자	1	0.66%
학생	8	5.26%
주부	1	0.66%
무직/취업준비생	7	4.61%
기타	50	32.89%
합계	152	100%

2.1.6. 화자의 학력별 분포

공적 독백 화자의 학력별 분포를 보면 대졸이 전체의 63.82%를 차지하였고, 그다음으로 대학원 이상(25.66%), 고등학교 졸업(5.26%), 대학교 재학(4.61%)이 뒤를 이었다.

<표 46> 공적 독백 화자의 학력별 분포

학력	인원(명)	비율
초졸 이하	1	0.66%
중졸	0	0.00%
고졸	8	5.26%
대재	7	4.61%
대졸	76	63.82%
대학원 이상	39	25.66%
합계	152	100%

2.1.7. 출생지별 분포

공적 독백 화자의 출생지별 수집 결과를 보면 서울이 전체 인원의 34.87%를 차지하였고, 그다음으로 경기가 15.79%, 인천이 5.92%를 차지하였다.

<표 47> 공적 독백 화자의 출생지별 분포

출생지	인원	비율
서울	53	34.87%
인천	9	5.92%
경기	24	15.79%
강원	8	5.26%
부산	5	3.29%
대구	7	4.61%
울산	3	1.97%
경북	6	3.95%
경남	4	2.63%
대전	4	2.63%
충북	5	3.29%
충남	7	4.61%
광주	5	3.29%
전북	7	4.61%
전남	5	3.29%
제주	0	0.00%
총합계	152	100%

2.1.8. 주 성장지별 분포

공적 독백 화자의 주 성장지별 수집 결과를 보면 서울이 전체의 34.21%로 가장 많고 경기가 14.47%, 인천이 9.87%를 차지하였다.

<표 48> 공적 독백 주 성장지별 분포

주 성장지	인원	비율
서울	52	34.21%
인천	15	9.87%
경기	22	14.47%
강원	8	5.26%
부산	4	2.63%
대구	8	5.26%
울산	4	2.63%
경북	4	2.63%
경남	4	2.63%
대전	5	3.29%
충북	5	3.29%
충남	6	3.95%
광주	4	2.63%
전북	6	3.95%
전남	5	3.29%
제주	0	0.00%
합계	152	100.00%

2.1.9. 현 거주지별 분포

공적 독백 화자의 현 거주지는 경기가 38.82%로 가장 많고 서울 36.84%, 인천이 7.89% 순으로 나타났다.

<표 49> 공적 독백 현 거주지별 분포

현 거주지	인원	비율
서울	56	36.84%
인천	12	7.89%
경기	59	38.82%
강원	3	1.97%
부산	2	1.32%
대구	2	1.32%
울산	1	0.66%
경북	2	1.32%
경남	0	0.00%
대전	4	2.63%
충북	3	1.97%
충남	4	2.63%
광주	1	0.66%
전북	3	1.97%
전남	0	0.00%
제주	0	0.00%
합계	152	100.00%

2.1.10. 공적 독백 수집 목표 대비 실적

<표 50> 공적 독백 분야별 수집 목표 대비 실적

분야	목표		실적		달성률
	시간	비율	시간	비율	
유튜브	40	40%	41.0	40.39%	102.52%
강연 등	60	60%	60.5	59.61%	100.81%
합계	100	100%	101.5	100%	101.5%

<표 51> 공적 독백 주제별 수집 목표 대비 실적

분야	구분	목표		실적		달성률
		시간	비율	시간	비율	
유튜브	뷰티/패션	5	12.50%	4.8	11.59%	95.08%
	게임	5	12.50%	5.1	12.52%	102.72%
	스포츠/건강	5	12.50%	5.0	12.30%	100.85%
	취미/여가	5	12.50%	5.9	14.43%	118.42%
	푸드/쿠킹	5	12.50%	4.9	12.07%	98.97%
	IT/기술/과학	5	12.50%	5.1	12.53%	102.77%
	동물/펫	5	12.50%	5.4	13.08%	107.24%
	교육/강의	5	12.50%	4.8	11.48%	94.12%
	소계	40	100%	41	100%	102.52%
강연 등	강연	15	25%	15.7	26.03%	104.96%
	강의	15	25%	14.4	23.72%	95.64%
	연설	15	25%	15.1	24.95%	100.62%
	발표	15	25%	15.3	25.30%	102.02%
	소계	60	100%	60.5	100%	100.81%
합계		100		101.5		101.50%

<표 52> 공적 독백 연령별, 성별 수집 실적(단위: 시간)

구분	성별	실적	
		시간	비율
20대	남성	10.7	10.63%
	여성	5.5	5.37%
30대	남성	23.9	23.46%
	여성	21.1	20.83%
40대	남성	14.9	14.69%
	여성	5.0	4.90%
50대	남성	3.9	3.87%
	여성	0.8	0.78%
60대 이상	남성	15.0	14.80%
	여성	0.7	0.67%
합계		101.5	100.00%

2.2. 공적 독백 전사 결과

1인 발화(독백) 말뭉치 639건의 발화 문장 수는 총 81,869문장, 말뭉치 1건당 평균 128문장이며, 전사 어절 수는 총 769,604어절, 말뭉치 1건당 평균 1,204어절이다. 분야별, 주제별, 성별, 연령별 전사 결과는 다음과 같다.

<표 53> 공적 독백 분야별, 주제별 전사 결과

분야	주제	말뭉치 건수	시간	발화 문장 수	전사 어절 수
유튜브	뷰티/패션	28	4.8	3,871	38,659
	게임	28	5.1	4,128	42,677
	스포츠/건강	31	5.0	4,176	42,974
	취미/여가	31	5.9	5,408	44,877
	푸드/쿠킹	37	4.9	4,545	41,363
	IT/기술/과학	34	5.1	3,826	44,239
	동물/펫	47	5.4	4,689	41,412
	교육/강의	29	4.8	4,181	38,802
	소계	265	41	34,824	335,003
강연 등	강연	87	15.7	13,386	121,201
	강의	90	14.4	12,009	110,721
	연설	118	15.1	10,707	95,114
	발표	79	15.3	10,943	107,565
	소계	374	60.5	47,045	434,601
합계		639	101.5	81,869	769,604

<표 54> 공적 독백 성별 전사 결과

순번	주제	말뭉치 건수	시간	발화 문장 수	전사 어절 수
1	남성	450	68.4	55,621	530,794
2	여성	189	33.1	26,248	238,810
합계		639	101.5	81,869	769,604

<표 55> 공적 독백 연령별 전사 결과

순번	연령	말뭉치 건수	시간	발화 문장 수	전사 어절 수
1	20대	85	16.2	12,521	122,371
2	30대	267	45.0	37,477	358,180
3	40대	133	19.9	16,343	152,943
4	50대	28	4.7	4,062	36,914
5	60대 이상	126	15.7	11,466	99,196
합계		639	101.5	81,869	769,604

<표 56> 공적 독백 연령별, 성별 전사 결과

구분	성별	말뭉치 건수	시간	발화 문장 수	전사 어절 수
20대	남성	55	10.7	8,352	79,480
	여성	30	5.5	4,169	42,891
30대	남성	153	23.9	20,570	206,840
	여성	114	21.1	16,907	151,340
40대	남성	96	14.9	12,442	118,059
	여성	37	5.0	3,901	34,884
50대	남성	23	3.9	3,463	31,490
	여성	5	0.8	599	5,424
60대 이상	남성	123	15.0	10,794	94,925
	여성	3	0.7	672	4,271
합계		639	101.5	81,869	769,604

[붙임1] 2025년 일상 대화 말뭉치 구축 지침

1. 파일 형식 및 개요

1.1. 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축년도	8자리 일련번호
S: 구어 말뭉치	A: 공적 독백 D: 사적 대화	RW: 원시 말뭉치	25	########

- 예시
  · SDRW2500000001.json 원시 말뭉치 첫 번째 파일

※ 참고: 음성 파일 파일명 부여 방식
  · SDRW2500000001.pcm 음성 원본 첫 번째 파일
  · SDRW2500000001-00001.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
- 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: 2025
	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
			contact_frequency	string	친밀도(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": "SDRW2500000001",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2500000001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2025",
        "category": "구어 > 사적 대화 > 일상 대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2500000001.1",
            "metadata": {
                "title": "3인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20250517",
                "topic": "사회적 변화와 우리 생활",
                "speaker": [
                    {
                        "id": "SD2500038",
                        "age": "50대",
                        "occupation": "주부",
                        "sex": "여성",
                        "birthplace": "전북",
                        "principal_residence": "전북",
                        "current_residence": "전북",
                        "education": "고졸"
                    },
                    {
                        "id": "SD2500039",
                        "age": "50대",
                        "occupation": "주부",
                        "sex": "여성",
                        "birthplace": "전북",
                        "principal_residence": "전북",
                        "current_residence": "전북",
                        "education": "고졸"
                    },
                    {
                        "id": "SD2500040",
                        "age": "50대",
                        "occupation": "주부",
                        "sex": "여성",
                        "birthplace": "전남",
                        "principal_residence": "전남",
                        "current_residence": "전북",
                        "education": "고졸"
                    }
                ],
                "setting": {
                    "relation": "이웃사촌",
                    "contact_frequency": "2"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2500000001.1.1.1",
                    "form": "얼마 전에 나 일하는 데 아저씨가 집 안에 어 음식물 처리 그걸 싱크대에 달았대요.",
                    "original_form": "얼마 전에 나 일하는 데 아저씨가 집 안에 어~ 음식물 처리 그걸 싱크대에 달았대요.",
                    "speaker_id": "SD2500038",
                    "start": 0.06001,
                    "end": 9.30000,
                    "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 SDRW2500000001)
- 구축자(creator): 국립국어원
- 배포자(distributor): 국립국어원
- 구축년도(year): 2025
- 분류(category): 구어 > 사적 대화 > 일상 대화, 구어 > 공적 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: SDRW2500000001.1)

2.2.4. 대화 메타 정보(document > metadata)

- 대화 제목(title): 1인 공적 독백, 2인/3인/4인 일상 대화
- 저작권자(author): 개인 발화자
- 발행자(publisher): 개인 발화 녹음
- 녹음일자(date): 연월일 YYYYMMDD
- 대화 주제(topic): 대화 주제

다자대화 대화 주제	
1	건강
2	문화예술
3	음식
4	경제
5	회사/학교/학창시절
6	반려동물/반려용품
7	여행/휴가/휴일/자연휴양지
8	쇼핑
9	새로운 기술과 우리 생활
10	사회적 변화와 우리 생활

1인 공적 독백 대화 주제	
1	뷰티/패션
2	게임
3	스포츠/건강
4	취미/여가
5	푸드/쿠킹
6	IT/기술/과학
7	동물/펫
8	교육/강의
9	강연
10	강의
11	연설
12	발표

2.2.5. 화자 정보(document > metadata > speaker)

- 화자 아이디(id): 화자 고유 아이디 부여, 대화가 다르더라도 화자가 동일하면 동일한 아이디 부여 단, 화자가 교정기를 착용한 경우에는 구축 연도 다음 숫자 1을 넣어 표시(한 화자가 교정기를 뺐다 넣었다 하지 않도록 함)
  (예: 교정기 미착용 화자 A: SD2500001, 교정기 착용 화자 B: SD2510002)
- 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

- 친밀도(contact_frequency): 0~5, N/A 중 선택

0	1	2	3	4	5	N/A
처음 (낯선관계)	월 1회 미만	주 1회 미만	주 1~2회	주 3회 이상	거의 매일	1인 독백에 한정함

2.2.7. 발화 정보(document > utterance)

- 발화 아이디(id): 대화 아이디 + . + 1 + . + 1 + . + 발화 번호(예: SDRW2500000001.1.1.4)
- 철자 전사(form): 철자 전사 결과
- 발음 전사(original_form): 발음 전사 결과
- 발화 시작 시간(start): 해당 발화의 음성 원본에서의 시작 시간을 초 단위(소수 5자리까지 필수)로 표기(예: 30.56600)
- 발화 종료 시간(end): 해당 발화의 음성 원본에서의 종료 시간을 초 단위(소수 5자리까지 필수)로 표기(예: 32.48262)
- 전사자 기타 메모(note): 녹음실 밖의 관계자의 개입으로 녹음이 중단되는 경우 등 관계자와 나눈 대화는 전사하지 않고 메모를 남김.

3. 전사 지침

3.1. 기본 원칙

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

3.2. 화자 표시

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

3.3. 음성 분절 및 전사 단위

- 음성 분절 및 전사의 기본 단위는 문장이 되도록 한다.
  ※ 음성 정제본 하나가 하나의 전사 단위가 되도록 한다.

- 문장은 휴지, 경계 억양(상승조, 하강조)을 특징으로 하는 억양구를 기준으로 분할하되, 억양구 경계의 선행절과 후행절이 각 절의 서술어가 요구하는 문장 성분을 모두 갖추었을 경우 분할한다. 단, 서술어가 생략되었을 경우에도 분할한다.

  ※ 서술어가 생략된 경우

화자1	외국인 데리고 오면 막 타자 다 아는 선수들 되게 유명한 사람 많았잖아.
화자2	어 소크라테스나 이런 애들.
화자2	근데 소크라테스 요즘 좀 안 좋더라고.

화자1	약만 보조제 빼고 약만 먹는 데도 그렇게 많은 거야.
화자2	그래서 이게 이렇게 살다가는 정말 내가 약에 중독이 되지 않을까 싶을 정도로.
화자2	그래서 나 같은 경우는 아직은 뭐 보조제같은 거 그런 거는 안 먹고

- 서술어에 뒤따르는 구가 휴지 없이 발화되어 서술어와 하나의 억양구를 이룰 경우 해당 구(후치된 표현, 접속사, 간투사 등)를 포함하여 문장을 나눈다.

- 긴 쉼에 의해 나뉘는 경우는 통사적으로 완성이 되지 않았다 하더라도 전사 단위를 구분하여 전사한다. 단, 하나의 어절에서의 긴 쉼이 있을 경우는 나누지 않는다.

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)

- 분리한 자리에 음가 손실이 우려되는 경우에는 아래와 같은 기준을 적용한다.

기준	적용
전체 발화 시간이 6초 이내	- 분할하지 않음
전체 발화 시간이 6초 초과	- 다음 억양구 경계 또는 분할 가능한 가장 가까운 휴지에서 분할 - 종결부호는 문장이 완전히 종결되는 경우에만 사용

3.4. 발화 겹침

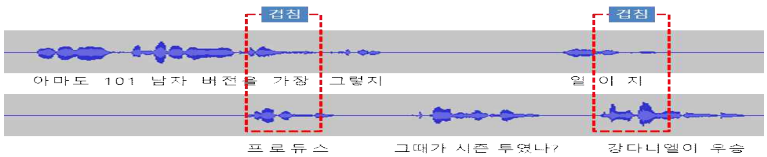
- 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다.

  ※ 맞장구 발화는 기계적으로 분리되어야 하나 기계적으로 분리되지 않은 경우에 예시와 같이 처리한다.

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

- 발화가 겹칠 경우에는 아래와 같이 처리한다.

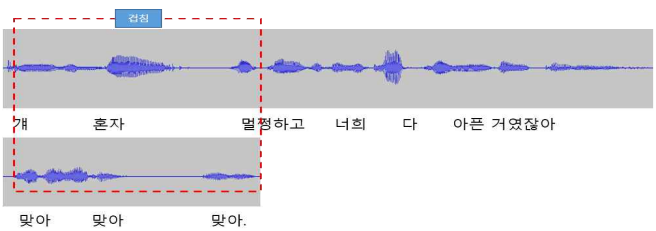
  ※ 일부 겹침




```
{
"id": "SDRW2400000882.1.1.103",
    "form": "아마도 101 남자 버전을 가장",
    "original_form": "아마도 일공일 남자 버전을 가장 ",
"speaker_id": "SD2400349",
"start": 453.79000,
"end": 457.49996,
"note": "발화겹침"
},
{
"id": "SDRW2400000882.1.1.104",
"form": "프로듀스",
"original_form": "프로듀쓰",
"speaker_id": "SD2400350",
"start": 456.98900,
"end": 457.61242,
"note": "발화겹침"
},
{
"id": "SDRW2400000882.1.1.105",
"form": "그렇지.",
"original_form": "그렇지..",
"speaker_id": "SD2400349",
"start": 457.69994,
"end": 458.24981,
"note": ""
},
```

```
{
"id": "SDRW2400000882.1.1.106",
    "form": "그때가 시즌 투였나?",
"original_form": "그때가 씨즌 투였나?",
"speaker_id": "SD2400350",
"start": 458.25238,
"end": 459.26901,
"note": ""
},
{
"id": "SDRW2400000882.1.1.107",
"form": "1이지.",
"original_form": "일이지.",
"speaker_id": "SD2400349",
"start": 459.39872,
"end": 460.10902,
"note": "발화겹침"
},
{
"id": "SDRW2400000882.1.1.108",
    "form": "강다니엘이 우승한 아 1.",
    "original_form": "강다니엘이 우승한 아~ 일. ",
"speaker_id": "SD2400350",
"start": 459.67000,
"end": 462.15965,
"note": "발화겹침"
},
```


  ※ 전체 겹침




```
{
        "id": "SDRW2400001083.1.1.76",
        "form": "개 혼자 멀쩡하고 너희 다 아픈 거였잖아.",
        "original_form": "개 혼자 멀쩡하고 너네 다 아픈 거였잖아.",
        "speaker_id": "SD2400789",
        "start": 203.42789,
        "end": 207.1,
        "note": "발화겹침"
    },
    {
        "id": "SDRW2400001083.1.1.77",
        "form": "맞아 맞아 맞아.",
        "original_form": "맞아 맞아 맞아.",
        "speaker_id": "SD2400790",
        "start": 205.279,
        "end": 206.949,
        "note": "발화겹침"
```

- 대화 상대의 발화에 대응하는 발화가 아닌 의미 없이 버릇처럼 발화되는 ‘어’, ‘응’, 등의 지속적인 발화에 대해서는 전사하지 않는다.

3.5. 발화 내용 전사

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

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

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

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

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

> 철자 전사: 오리지널
> 발음 전사: 오리지날
>
> 철자 전사: 티브이
> 발음 전사: 티비
>
> 철자 전사: 아이유와 **컬래버**했어
> 발음 전사: 아이유와 **콜라보**했어

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

철자 전사: 어떡해
발음 전사: 어뜩해

철자 전사: 소주
발음 전사: 쏘주

- 발음 전사 시 약화 현상에 의한 이형태는 반영하지 않는다. 예를 들어 의문사 '뭐'가 '머'로 모음이 약화되어 들려도 별도의 발음 전사를 하지 않고 철자 전사인 '뭐'만 적는다.
- 영화명, 드라마명 등은 대중에게 공개된 맞춤법과 띄어쓰기를 우선으로 한다.

폭싹 속았수다 (○)
폭삭 속았수다 (×), (폭싹)/(정말) (속았수다)/(수고하셨습니다) (×)

- 널리 알려진 맞춤법과 띄어쓰기보다 <우리말샘>의 표기를 우선으로 한다..

해리 포터 (○), 해리포터 (×)
모둠 초밥 (○), 모듬초밥 (×), 모듬 초밥 (×)

- 잘못된 용언의 활용형은 이중 전사 한다.

그런 까페들은 꼭 (들리는)/(들르는) 편인 거 같아요. (○)
그런 까페들은 꼭 들리는 편인 거 같아요. (×)

- '요런', '조런' 등은<우리말샘>에 등재된 단어이니 이중 전사 하지 않는다.

테무나 지그재그 에이블리 요런 데서 (○)
테무나 지그재그 에이블리 (요런)/(이런) 데서 (×)

3.6. 모음의 축약형 표기

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

철자 전사: 사귀어
발음 전사: 사귀'어

철자 전사: 바뀌어
발음 전사: 바뀌'어

철자 전사 : 할퀴어
발음 전사 : 할퀴'어

3.7. 준말과 센말의 전사

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

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

센말 예)
조끔, 쪼금, 쪼끔(조금), 쫄쫄(졸졸), 딴딴하다(단단하다)

- <국립국어원 우리말샘>에 에 등재되지 않은 경우는 이중 전사 한다.

준말 예)
(알바비)/(아르바이트비), (왜냐면)/(왜냐하면), (그니까)/(그러니까), (이케)/(이렇게)… …

센말 예)
(쫄쫄쫄)/(졸졸졸), (쫌)/(좀), …

- 준말과 비슷한 유형인 '줄어든 말'과 '줄여 이르는 말'은 이중 전사하지 않는다. 이중 전사는 구어의 발음 정보를 제공하기 위한 것이므로 줄어들기 전의 형태 정보를 이중 전사하지 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.

3.8. 끊어진 단어(단어가 불완전하게 발화된 경우)

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

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

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

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

츠 트 그~ 리더 같은 사람들 있잖아. (X)
=> -츠- -트- 그~ 리더 같은 사람들 있잖아. (O)

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

얼마큼 개선되고 가까워-져-지느냐가 이제
방문-은-만 남았거든
그런 데 가는 거 좋아하고 하는데 다른 분들-은-도 그렇고

3.9. 띄어쓰기

- 한글 맞춤법에 맞게 띄어 쓴다.
- 의존명사는 띄어 쓴다.
- 수를 적을 때는 만 단위로 띄어 쓴다(예: 십이억 삼천백만 팔백구 불 등).
- 본용언과 보조 용언도 띄어 쓴다.(예: 먹어 버리다, 가고 싶다, 먹지 못하다)

용병도 여러 번 뛰어 봤고 (○), 용병도 여러 번 뛰어봤고 (×)
한번 여쭤 본 겁니다. (○), 한번 여쭤본 겁니다. (×)
제주도에 가 보셨나요? (○), 제주도에 가보셨나요? (×)
요리를 해 주는 남편 (○), 요리를 해주는 남편 (×)
아르바이트가 너무 한번 해 보고 싶어서 (○), 아르바이트가 너무 한번 해보고 싶어서 (×)

- 띄어쓰기와 붙여쓰기 모두 허용되는 경우에는 띄어 쓰는 것을 원칙으로 한다.

본용언, 보조 용언	막아낸다 vs **막아 낸다(O)**
전문 용어	성격묘사 vs **성격 묘사(O)**
시분초, 연월일 등	두시 vs **두 시(O)**
단음절 연속	좀더 큰것 vs **좀 더 큰 것(O)**

- 단어를 발음하는 중간에 쉼이 들어간 경우에는 띄어 쓰지 않는다.

음성 : 많이 먹는∨∨∨구나
전사 : 많이 먹는구나 (O)

음성 : 그걸로 넘어∨∨∨지기야 하겠냐만은
전사 : 그걸로 넘어지기야 하겠냐만은 (O)

- 우리말샘 등재 내용을 기준으로 하되, 판단하기 어려운 경우에는 수시로 논의하여 결정한다.
- <우리말샘>에서 표제어가 띄어져 있으면 언제나 띄어 써야 한다.

고 삼 : '고등학교 삼 학년'을 줄여 이르는 말. <우리말샘>
(고 삼)/(고3) (○)
(고삼)/(고3) (×)

- <우리말샘>에서 '^' 부호는 해당 전문어 구를 띄어 쓰는 것이 원칙이나 붙여 씀도 허용한다는 의미이다. 본 사업의 전사 작업에서는 원칙에 따라 언제나 띄어 쓴다.

문화^마을 : 『지명』 충청남도 공주시 계룡면에 있는 마을. <우리말샘>
문화 마을 (○)
문화마을 (×)

- 명사 뒤의 '드리다'는 붙여 쓰지만 동사 뒤의 '드리다'는 띄어 쓴다.

말씀드리다(말씀<명> + 드리다) (○) / 말씀 드리다 (×)
알려 드리다(알려<동> + 드리다) (○) / 알려드리다 (×)

- 음식명은 모두 붙여 쓴다. 다만 지역의 특산물은 띄어 쓴다.

파인애플피자 (○), 파인애플 피자 (×)
쉑쉑버거 (○), 쉑쉑 버거 (×)
밀떡볶이 (○), 밀 떡볶이 (×)
로제마라샹궈 (○), 로제 마라샹궈 (×)
콩나물잡채 (○), 콩나물 잡채 (×)

영광 굴비 (○, 영광굴비 (×)

- 관형사나 관형어가 꾸미는 '분'은 띄어 쓰고, 명사 뒤에 오는 '분'은 붙여 쓴다. 또한 복수의 청자인 경우에는 '여러분'을, 다수의 사람들을 의미할 때는 '여러 분'을 쓴다.

반대하시는 분 계십니까? (○), 반대하시는분 계십니까? (×)
어르신분들 (○), 어르신 분들 (×)
외국분들 (○), 외국 분들 (×)
성우분들이 뮤지컬을 하세요. (○), 성우 분들이 뮤지컬을 하세요. (×)

여러분, 안녕하십니까? (○), 여러 분, 안녕하십니까? (×)
이 자리에 여러 분들이 모였네요. (○), 이 자리에 여러분들이 모였네요. (×)

- “~지” 다음에 오는 ‘못하다’는 붙여 쓴다.

> 보지 못했던 해외 영화나 (○), 보지 못 했던 해외 영화나 (×)

- 부정어(없다/못하다)와 호응하는 ‘밖에’는 앞 명사에 붙여 쓴다.

> 나밖에 없다니까 (○), 나 밖에 없다니까 (×)
> 클래식밖에 못 들어 봐서 (×), 클래식 밖에 못 들어 봐서 (×)

- 띄어 쓰는 ‘같이 하다’와 붙여 쓰는 ‘같이하다’가 서로 다르니 <우리말샘>을 참조해서 구별하여 사용한다. 헷갈리면 ‘같이’의 어순을 바꿔도 말이 되면 ‘같이 하다’로 보면 된다.

> 친구와 밤새 시험공부를 같이 했다. (○) : 친구와 같이 밤새 시험공부를 했다. (○) 친구와 밤새 시험공부를 같이했다. (×)
> 우리는 끝까지 의견을 같이할 것이다. (○) 우리는 끝까지 의견을 같이 할 것이다. (×) : 우리는 같이 끝까지 의견을 할 것이다. (×)

- “그런데”의 의미를 지니는 ‘데’를 앞말에 붙여 쓰고, “곳”이나 “것”의 의미를 지니는 ‘데’를 띄어 쓴다.

> 그 사람이 정직하기는 한데 이번 일에는 적합지 않다. (○), 그 사람이 정직하기는 한 데 이번 일에는 적합지 않다. (×)
> 매운 음식을 못 먹는 데 반해 (○), 매운 음식을 못 먹는데 반해 (×)

- “물건을 팔거나 영업을 하는 가게”를 나타내는 ‘집’은 앞말에 붙여 쓴다.

> 피자집 (○), 피자 집 (×)
> 베이글집 (○), 베이글 집 (×)
> 흑돼지집 (○), 흑돼지 집 (×)

- “능숙하게” 하면 ‘잘하다’를, “자주” 하면 ‘잘 하다’를 쓴다.

> 나는 어머니를 닮아 닭 요리는 잘한다. (○) 나는 어머니를 닮아 닭 요리는 잘 한다. (×)
> 밀 키트가 발달해서 집에서도 요리를 잘 한다. (○) 밀 키트가 발달해서 집에서도 요리를 잘한다. (×)

- ‘-화(化)’로 끝나는 명사 다음의 ‘하다’와 ‘되다’는 앞말에 붙여 쓴다.

> 객관화하다 (○), 객관화 하다 (×)
> 대중화돼 (○), 대중화 돼 (×)

- 명사 뒤의 ‘하고’가 “와/과”의 뜻이면 조사이므로 앞말에 붙여 쓴다.

> 밥 쪼끔하고 참치 한 반 캔 정도하고 굴소스 쪼끔 넣어서 이렇게 볶아 먹었다. (○)
> 밥 쪼끔 하고 참치 한 반 캔 정도 하고 굴소스 쪼끔 넣어서 이렇게 볶아 먹었다. (×)

- ‘큰’ 다음에 가족 관련어가 오면 앞말에 붙여 쓴다.

> 큰애 (○), 큰 애 (×)
> 큰아버지 (○), 큰 아버지 (×)

- 일반어 구(句) 다음에 오는 '하다'는 띄어 쓴다.

> 전사하다. (○) 이중 전사 하다. (○), 이중 전사하다. (×)
> 배송되다. (○) 무료 배송 되다. (○), 무료 배송되다. (×)
> 거래하세요? (○) 중고 거래 하세요? (○), 중고 거래하세요? (×)

- 외래어/외국어와 결합한 '하다'는 앞말에 붙여 쓴다.

> 크리미하다. (○) 크리미 하다. (×)
> 좀 액티비티한 걸 즐기시는 편이세요? (○) 좀 액티비티 한 걸 즐기시는 편이세요? (×)

3.10. 담화 표지

- 머뭇거림의 기능을 하는 1음절 담화 표지 중 "이, 그, 저, 아, 어, 예, 음, 응, 뭐"의 9개 형태에 한해서 본래의 품사와 구별하기 위해 물결표(~)를 붙여 전사한다.
- 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.
- "인제, 이제, 그냥, 무슨, 어떤" 등의 2음절 이상의 담화 표지는 물결표를 붙이지 않는다.
- 억양과 운율에 의해서만 구분이 가능할 경우는 반드시 전사 단계에서 표시해 준다.

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

- 기호 '~'는 머뭇거림의 담화 표지에 붙이는 특수한 부호이므로 장음의 부호로 사용하지 않는다. 특히 단음절 응답 표현의 경우 길게 발음할 때 이를 사용하지 않도록 주의한다.

3.11. 잘 들리지 않는 부분

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

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

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

> 철자 전사: 너무한 거 같더라.
> 발음 전사: **(())** 너무한 거 같더라.

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

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

* 철자 전사에서는 "이중괄호(())"삭제

3.12. 준음성과 기타 소리들

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

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

*** 철자 전사에서는 삭제한다.**

- 위의 네 가지 준음성 외의 다른 항목은 입력하지 않는다. 간혹 숨 들이마시는 소리를 전사하는 경우가 있는데 삭제해야 한다.

음성 : 근데 그거 습
전사 : 근데 그거

- 단독으로 발화한 준음성만 표기 진행하고, 발화에 섞인 준음성은 확실하게 구분이 가능한 경우에만 표기한다. 단, 단독으로 발화한 준음성이라도 일반 발화와 시간차가 있는 경우 표기하지 않는다.

3.13. 숫자 전사(상세)

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

나이 철자 전사: 10살 발음 전사: 열 살	철자 전사 : 40살에 발음 전사 : 마흔 살에	
시간 철자 전사: 24시간 발음 전사: 스물네 시간	**시간** 철자 전사: 24시간 발음 전사: 이십사 시간	
날짜 철자 전사 : 3 4일 발음 전사 : 삼사일	**날짜** 철자 전사 : 5 6년 발음 전사 : 오륙년	**날짜** 철자 전사 : 2021년 5월 21일 발음 전사 : 이천이십일 년 오월 이십일 일
※ 이삼일, 삼사일, 오륙일, 일주일, 일월, 이월, 삼월…십이월 등은 한 단어로 굳어져 사전에 등재되어 있으므로 붙여 씀		
금액 철자 전사: 2 30 만 원 발음 전사: 이삽십만 원	**금액** 철자 전사 : 3만 7000원 발음 전사 : 삼만 칠천 원	
측정 철자 전사 : 3만 보 발음 전사 : 삼만 보	**측정** 철자 전사 : 30킬로 발음 전사 : 삼십 키로	

- 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)

- 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)

※ 횟수의 의미가 분명한 경우에는 '한 번'으로 띄어 쓰지만, 그 외의 경우에는 '한번'으로 붙여 쓴다. 다만 '한 번도', '한 번만', '한 번씩', '한 번쯤'은 언제나 띄어 쓴다.
(한 번도)/(1번도), (한 번만)/(1번만), (한 번씩)/(1번씩), (한 번쯤)/(1번쯤) (○)

※ '한번은'은 횟수의 의미가 아니므로 이중 전사를 하지 않는다.
한번은 그런 일도 있었어요. (○)
(한 번은/1번은) 그런 일도 있었어요. (×)

- 숫자 철자 전사의 띄어쓰기는 "경", "조", "억", "만" 단위로 띄어 쓴다.

철자 전사: 1억 2000만 원
발음 전사: 일억 이천만 원

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

✶ 만(10,000)의 철자 전사는 '만'으로 표기한다.

- 철자 전사 시 천 단위 분할 ","(쉼표)는 쓰지 않는다.
- 날짜(일시), 금액(돈), 측정(계량)단위는 이중 전사한다.

구분	철자 전사	발음 전사
날짜	2021년 5월 21일	이천이십일 년 오 월 이십 일
	3 4일, 5 6년	삼사일, 오륙 년
금액	3만 7000원	삼만 칠천 원
	4500원	사천오백 원
	2 30만 원	이삼십만 원
측정	3만 보	삼만 보
	30킬로	삼십 키로
	10프로	십 프로

- 모든 전사 시 단위를 나타내는 명사는 띄어 쓴다. 다만 아라비아 숫자 뒤에 붙는 단위는 붙여 쓴다.

철자 전사	발음 전사
5만 원	오만 원
4만 2800원	사만 이천팔백 원
7000원	칠천 원
300만 원	삼백만 원
15프로	십오 프로
66킬로	육십육 키로
25년	이십오 년
1000만 관객	천만 관객

- 숫자가 포함된 고유명의 경우 이중 전사 하지 않는다.

※ 고유명(예능 프로그램)으로 사용되는 경우 - 이중 전사 X

> 철자 전사 : 어제 일박 이 일을 봤는데
> 발음 전사 : 어제 일박 이 일을 봤는데

※ 사전적 의미로 사용되는 경우 - 이중 전사 O

> 철자 전사 : 우리는 1박 2일 여행을 가서
> 발음 전사 : 우리는 일박 이 일 여행을 가서
> => '일박'은 <우리말샘>에 한 단어로 등재되어 있어 붙여 쓰지만, 의미가 달라진 것은 아니므로 숫자 병기를 해 주어야 함

- 숫자전사 띄어쓰기 오류 사례

옳은 표기	오류
철자 전사 : 3 40대 돼서 발음 전사 : 삼사십 대 돼서	철자 전사 : 3 40대 돼서 발음 전사 : 삼 사십 대 돼서
'삼사'는 한 단어(관형사) '일이', '이삼', '사오', '오륙', '육칠', '칠팔' '팔구' 또한 한 단어(관형사)	
철자 전사 : 7080노래들이 좋아. 발음 전사 : 칠공팔공 노래들이 좋아.	철자 전사 : 7080노래들이 좋아. 발음 전사 : 칠 공 팔 공 노래들이 좋아.
'칠공팔공' 한 단어	
철자 전사 : 제2의 꿈 발음 전사 : 제이의 꿈	철자 전사 : 제2의 꿈 발음 전사 : 제 이의 꿈
'제-'는 접두사이므로 뒷말과 붙여 씀	
철자 전사 : 제3자 발음 전사 : 제삼자	철자 전사 : 제3자 발음 전사 : 제 삼자
철자 전사 : 1.5배에서 발음 전사 : 일 점 오 배에서	철자 전사 : 1.5배에서 발음 전사 : 일점오 배에서

3.14. 방언의 전사

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

철자 전사: 그런데
발음 전사: 근디 *준말의 방언형은 표준어의 본딧말로 통일

철자 전사: 먹었지
발음 전사: 묵었지

※ 기본 형태를 살려 적으므로 필수적 경음화는 발음 전사에 반영하지 않는다.

<<오류>>
철자 전사 : 그래가지고 그때마다 어떻게 인제 이거 절판이에요 안 나와요. (X)
발음 전사 : 그래가꼬 그때마다 어떻게 인제 이거 절판이에요 안 나와요. (X)

<<올바른 전사>>
철자 전사 : 그래 갖고 그때마다 어떻게 인제 이거 절판이에요 안 나와요. (O)
발음 전사 : 그래 갖고 그때마다 어떻게 인제 이거 절판이에요 안 나와요. (O)

➔ 방언의 억양으로 발화하였을 것이나 실제 기본 어형을 살려 적으면 표준어와 동일한 경우임. 따라서 철저 잔서 시 본딧말로 복원하지 않음

- 방언 발음 전사 시 유의 사항은 다음과 같다.

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

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

철자 전사: 저번에
발음 전사: 쩌번에

철자 전사: 다르다
발음 전사: 따르다

철자 전사: 계속
발음 전사: 꼐속

- 소리 나는 대로 적은 “방언 전사”가 표준어 규정에서 벗어난 경우에 그에 대응하는 표준형을 함께 제시하는 것을 원칙으로 한다.

지역	철자전사	발음전사
강원	모처럼 해가 난 날에는 마실이나 다녀오시오.	모처럼 해가 난 날에느 마실이나 댕게오시우.
	돈이 없어도 남한테 아쉬운 소리는 못하겠다.	돈이 읎어도 남한테 아수운 소리는 못하겠다.
	여기서 꾸물거리지 말고 얼른 가라.	여서 꾸물거리지 말고 얼푼 가라.
경상	여기에 동그라미나 곱표 치세요.	여기에 동그라미나 꼽표 치세요.
	떡을 만들어 먹었지.	떡을 맹갈아 묵었지.
	할 게 매우 많다.	할 게 천지삐까리다.
전라	하루 종일 이엉만 엮고	하루 종일 이엉만 영끄고
	급히 약을 지었는데도 못 낫고 가 버렸어.	급히 약을 지었는데도 못 나수고 가 부렀어.
	늦은 사람이 도리어 큰소리친다.	늦은 사람이 댑대로 큰소리친다.
제주	야 무슨 그런 게 또 있어.	야 무신 그런 게 또 시어.
	어떻든 저기 다 지나치면 됩니다.	어떻든 저디 다 지내치민 되우다.
	성격이 참 야무지다.	성격이 참 요망지다.
충청	너 때문에 여기까지 와야 되겠어?	너 또래 여기꺼지 와야 되겄어?
	오디를 얼마나 많이 먹었는지 입 안이 시커멓게 물들었어요.	오둥아를 얼마나 마이 먹었는지 입 안이 시커멓게 물들었슈.
	여기 부추 한 단에 얼마요?	여기 졸 한 단에 얼마요?

- 경상방언 방언 전사 주의 사항

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

철자전사	발음전사
집에 갔다.	집에 갔대이.
전화 해라.	전화 해래이.
다음에 보자.	다음에 보재이.

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

철자전사	발음전사
그러면 저기 갔다 올 거니?	그카면 저기 갔다 올 끼가?
그러면 그 일은 끝났니?	그쿠면 그 일은 끝났나?
그러고 있지 말고 이리로 오너라.	그카고 있지 말고 일로 온나.
그러고 잘난척은 쟤 잘한다.	그쿠고 잘난척은 자가 잘한다.

※ 받침 'ㅇ'이나 'ㄴ'이 나타나지 않으면 소리대로 적는다.

철자전사	발음전사
주머니	주머이
어머니	어무이
학생이	학새이

- 전라방언 방언 전사 주의 사항

※ 표준어 '-으니까'의 방언형은 '-응께, -응게, -응께네, -으니께' 등으로 소리대로 적는다.

철자전사	발음전사
이제는 약이 좋으니까.	인자는 약이 조응께.
그 공식적으로 다 하니까.	그 공식적으로 다 허니께.

※ 둘째 음절 이하의 'ㅎ'이 나타나지 않는 말의 경우, 다음과 같이 소리대로 적는다.

철자전사	발음전사
뭣 하러 그러냐?	뭐다러 그러냐?
잘 하지도 못하고.	잘 하지도 모다고.
백화점에 가 보니까	배과점에 가 봉께
벌써 육 학년이야?	벌써 유강년이여?
으메 답답한 것.	으메 답다번거.
눈 앞이 갑갑하다.	눈 앞이 깝까버다.

※ ‘ㄴ’이 나타나지 않으면 소리대로 적는다.

철자전사	발음전사
가만히	가마이
많이씩	마이씩
아닌 게 아니라	아잉 게 아이라

- 제주방언 방언 전사 주의 사항

※ 앞에 요소가 받침이 있는 음절이고 후행하는 요소가 모음으로 시작할 때 후행 단어의 첫 음절 자리에 받침 자음을 복사하여 발음한다. 이러한 경우는 소리대로 적는다.

철자전사	발음전사
한국 음식	한국 금식
맏아들	맏다덜
비단옷	비단놋
감옷	감못

- 충청방언 방언 전사 주의 사항

※ 종결어미 ‘-댜’는 소리대로 적는다.

철자전사	발음전사
그 낮에 꿈을 꾸니까 그러더래.	그 낮에 꿈을 꾸니께 그라댜.
우리 아들 잡아 간대.	우리 야덜 잡어 간댜.

※ 표준어 ‘어떻게’의 방언형은 ‘워떻게, 어티기’ 등은 소리대로 적는다.

철자전사	발음전사
혹시 어떻게 하는 건 줄 아세요?	혹시 워떻게 하는 건 줄 아세요?
장사 하려고 어떻게 집을 크게 지었는데.	장사 하려고 어티기 집을 크게 졌는디.

3.15. 비식별화를 위한 전사

- 일상 대화 자료 중 개인 정보 등의 비식별화를 위해 이름, 이메일 주소 등 계정 정보, 주민등록번호, 카드 번호, 전화번호 등 각종 번호 및 비밀번호, 상세 주소, 출신 및 소속 등의 개인 정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다.
- 정치인 등 유명인(연예인, 스포츠 선수, 유튜버 포함)의 이름 및 닉네임, 상호명과 상품명 등 공적 성격을 지닌 이름들의 경우는 맥락상 부정적인 경우에만 비식별화한다.

  예) 정치인명, 연예인명, 스포츠 선수명, 유튜버명, 학교명, 상호명, 상품명, 기관명 등
- 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다.
- 비식별화 정보는 아래와 같이 마크업한다.

분류		태그	항목
이름		&name&	실명, 특수 애칭, 별명, 대화명, 필명, 가수 그룹명 포함
정치적 이름		&politician&	정치인의 실명, 특수 애칭, 별명, 변형된 형태(자모) 포함
출신 소속	출신 학교, 지역	&affiliation&	출신 학교, 지역 *출신 지역이 아닌 지역명은 장소로 주석
	온라인 커뮤니티		온라인 커뮤니티명
	팬클럽		팬클럽명
	기타		동아리명
	정치 조직/정당	&party&	정당명, 또는 변형된 형태이지만 맥락에서 어떤 정당인지 유추가 가능하면 포함
번호	고유 식별 번호	&social-security-num&	주민등록번호
	전화번호	&tel-num&	
	카드 번호	&card-num&	
	계좌 번호	&bank-account&	
	기타 번호	&num&	일련번호, (구매자) 식별 번호, 사업자 등록 번호, 비밀번호
온라인 계정		&online-account&	아이디, 이메일 주소
주소		&address&	상세 주소, 아파트 및 거주 건물명
상호명		&company&	기업/회사/상점 이름
장소명		&location&	나라, 도시 이름
상표명		&brand&	제품명, 브랜드명
창작물명		&art&	소설, 영화, 드라마, 만화 등의 작품명
기타		&other&	위에서 언급하지 않은 항목(비윤리적인 표현, 욕 등 포함)

- 말뭉치 내 모든 비식별화 대상에는 &company1&, &company2&와 같이 일련번호로 구분하여 전사한다. 이때 한 파일 내에서 해당 번호가 가리키는 대상이 일관성을 지녀야 한다.

> 그때 철수랑 민수랑 너랑 나랑 갔잖아. 철수도 알고 있지?
> → 그때 &name1&이랑 &name2&이랑 너랑 나랑 갔잖아 &name1&도 알고 있지? (O)
> → 그때 &name1&이랑 &name2&이랑 너랑 나랑 갔잖아 &name3&도 알고 있지? (X)

- 비윤리적인 표현은 비식별화한다. 비윤리적 표현에는 욕설, 성적인 표현, 세대, 신체, 종교 등과 관련한 차별/혐오 표현이 있다.

유형	비식별화 전	비식별화 후
욕설	씨발 기억이 안 나.	&other1& 기억이 안 나.
차별/혐오 표현	요즘 맘충이 정말 많아졌지 않아?	요즘 &other2&이 정말 많아졌지 않아?
성적인 표현	여자가 죽으려고 하면 남자가 와서 뭐~ 죽기 전에 한 번 하자.	여자가 죽으려고 하면 남자가 와서 뭐~ 죽기 전에 &other3&.

- 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.

- 비속한 표현이지만 욕설이라고 보기 어려운 말은 비식별화하지 않는다.

> [예] 존맛탱, 대존맛, 까먹다, 개좋다, 개꿀…

3.16. 영문 및 기호의 전사

- 영문 등 외래어 전사

※ 발음 전사는 발화자의 발음에 따라 한글로 적는다.

※ 철자 전사는 '한국어 어문 규범 외래어 표기법

※ 해당 외래어가 <우리말샘>에 등재되어 있는지 반드시 확인한다.

> 유니버설 (○), 유니버셜 (×)
> 미디엄 (○), 미디움 (×)

- 영문 약어의 철자 전사

알파벳	어문 규범
A	에이
B	비
C	시
D	디
E	이
F	에프
G	지
H	에이치
I	아이
J	제이
K	케이
L	엘
M	엠

알파벳	어문 규범
N	엔
O	오
P	피
Q	큐
R	알(아르)
S	에스
T	티
U	유
V	브이
W	더블유
X	엑스
Y	와이
Z	제트

※ 'R'은 '알'과 '아르' 두 가지로 쓰일 수 있으나 본 사업에서는 '알'로 통일하여 사용한다.

> CCTV_ (씨씨 티비)/(시시 티브이)
> SRT_ (에스얼티)/(에스알티)

3.17. 종결 부호 규정

- 느낌표나 쉼표는 사용하지 않는다. 문장이 완전히 종결이 되었을 때는 종결 부호(마침표나 물음표)를 사용한다.

※ 종결 어미로 끝난 경우

평서문 : -구나/-는구나, -군/-는군, -네(요), -는단다/-ㄴ단다/-단다/-란다, -다/ㄴ다/는다, -데(요), -소(오), -ㅂ습니다/습니다, -아(요)/어(요), -으마/-마, -을걸/-ㄹ걸, -을게/-ㄹ게, -을라/-ㄹ라, -을래/-ㄹ래, -지(요)
의문문 : -ㅂ니까/습니까, -담, -대, -련, -(으)ㄹ까요, -아(요)/어(요), -으니/-니, -으냐/-냐/-느냐, -으랴/-랴, -을쏘냐/-ㄹ쏘냐, -소/오, -나, -ㄴ가/는가, -지
명령문 : -아라/어라/-여라, -구려, -(으)오, -(으)십시오, -(으)라, -아/어, -게, -(으)소서, -(으)렴, -(으)려무나
청유문 : -자, -세, -(으)ㅂ시다
감탄문 : -구나/는구나, -군요/는군요, -아라/어라
기 타 : -는걸, -ㄴ걸, -은걸, -는걸요, -ㄴ걸요, -은걸요, -아/-어, -지

※ 종결 어미로 끝나지 않았지만 종결 어미가 드러나고 종결 어미 뒤에 다른 문장 성분(부사어, 목적어 등)으로 끝난 경우(후치 및 도치)

화자	대화1
1	이 가장 오랫동안 기억 남는 영화는 이제 하나가 제목은 자세히 기억이 안 나는데 내용이 아버지가 좀 가세가 좀 기울었어요 그 집안이.

화자	대화2
1	사회인 야구에서 보면 사회인 야구부 생각보다 많이 나누어져 있어 프로처럼.

화자	대화3
1	1달 후에 그 의사한테 다시 간 거예요 인제 체크하러 또.

※ 서술어가 생략된 경우

화자	대화1	
1	외국인 데리고 오면 막 타자 다 아는 선수들 되게 유명한 사람 많았잖아.	
2	분리 전	어 소크라테스나 이런 애들 근데 소크라테스 요즘 좀 안 좋더라고
	분리 후	어 소크라테스나 이런 애들.
		근데 소크라테스 요즘 좀 안 좋더라고.

화자	대화2	
1	그러니까 나도 이렇게 저기에다가 보조제까지 한다면 정말 이걸 셀 수도 없는데 약만 보조제 빼고 약만 먹는 데도 그렇게 많은 거야.	
2	분리 전	그래서 이게 이렇게 살다가는 정말 내가 약에 중독이 되지 않을까 싶을 정도로 그래서 나 같은 경우는 아직은 뭐 보조제 같은 거 그런 거는 안 먹고 그냥 오로지 그냥 정형외과 관절 약만 먹는데 난 다리하고 손가락만 안 아프면 살겠어
	분리 후	그래서 이게 이렇게 살다가는 정말 내가 약에 중독이 되지 않을까 싶을 정도로.
		그래서 나 같은 경우는 아직은 뭐 보조제 같은 거 그런 거는 안 먹고 그냥 오로지 그냥 정형외과 관절 약만 먹는데 난 다리하고 손가락만 안 아프면 살겠어.

화자	대화3	
1	내 친구가 건강했는데 전립선 암 수술 했다는 거야 아산 병 아산 병원에서.	
2	분리 전	1500만 원 주고 했다는 야 뭐 땜에 1500 그 로봇 수술을 한 거야 로봇.
	분리 후	1500만 원 주고 했다는.
		야 뭐 땜에 1500 그 로봇 수술을 한 거야 로봇.

※ 연결 어미가 종결 의미로 쓰이는 경우

① -거든

1. 청자가 모르고 있을 내용을 가르쳐 줌을 나타내는 종결 어미. 자랑이나 감탄의 느낌을 띨 때가 있다.

나는 지금 이제 직장을 슬슬 바꿔야 될 때가 왔거든.
그래서 나는 최대한 내 방에 누가 들어오는 게 싫어서 최 매일 청소는 하지 않아도 된다고 했 했었거든.

2. 앞으로 할 어떤 이야기의 전제로 베풀어 놓음을 나타내는 종결 어미.

나는 민원 분들이랑 이렇게 거의 시간을 보내다 보니까 약간 어르신들이 많이 오시거든. 그래서 약간 힘 초반에는 많이 힘들었는데 ~
그리고 근데 그때 사장님 없이 나 혼자서 일을 했거든. 혼자서 일을 해서 좀 많이 힘들었어.

② -ㄴ데(-는데, -던데, -은데)

1. 어떤 일을 감탄하는 뜻을 넣어 서술함으로써 그에 대한 청자의 반응을 기다리는 태도를 나타내는 종결 어미.

지금 그 건강 관리가 상당히 좋으신 걸로 판단이 되던데.
가리비 난 너무 맛있는데.

2. 일정한 대답을 요구하며 물어보는 뜻을 나타내는 종결 어미

그 옷은 얼만데?
누가 제일 예쁜데?

③ -다고(-는다고, -ㄴ다고)

1. 자신의 생각이나 주장을 청자에게 강조하여 일러 주는 뜻을 나타내는 종결 어미.

그렇게 시켜 먹으려면 뭐 몇만 원 드는 걸 집에서 푸짐하게 우리는 그렇게 먹고 있다고.

2. 해할 자리에 쓰여, '너의 말이나 생각이 이런 것이냐?' 하는 뜻으로 묻는 데 쓰는 종결 어미. 빈정거리거나 부정하는 뜻을 띨 때도 있다.

그게 왜 네 잘못인지 모르겠다고.

3. 해할 자리에 쓰여, 마음속에 가졌던 어떤 의문의 답이 의외로 별것이 아니었을 때에, 그 의문을 그대로 보여 주는 데 쓰는 종결 어미. 의문이나 긴장 또는 걱정이 해소되었다는 뜻이 암시된다.

난 또 누가 아프다고.
난 또 무슨 큰일이나 났다고.

④ -더니

1. 과거에 경험하여 새로이 알게 된 사실에 대해 묻는 종결 어미. 예스러운 느낌을 준다.

그 일이 참말이더니?
어머님께서는 진지는 잘 잡수시더니?

2. 주로 혼잣말에 쓰여, 과거에 직접 경험하여 알게 된 일을 회상하여 나타내는 종결 어미. 현재에 그와 대조되는 어떤 상황이 있음을 암시한다.

자라면 큰 재목이 되겠더니.
젊어서는 그렇게 고우시더니.

⑤ -라고(-으라고)

1. 자신의 생각이나 주장을 청자에게 강조하여 일러 주는 뜻을 나타내는 종결 어미.

근데 내가 실습을 나가 봤는데 어 되게 적성에는 맞는 거 같은데 좀 선배들이 많이 무섭더라고.
나 로봇 청소기가 있는데 난 그거 안 써지더라고.

2. '너의 말이나 생각이 이런 것이냐?' 하는 뜻으로 묻는 데 쓰는 종결 어미. 빈정거리거나 부정하는 뜻을 띨 때도 있다.

그것이 내가 잘못해서 그런 거라고.

3. 해할 자리에 쓰여, 마음속에 가졌던 어떤 의문의 답이 의외로 별것이 아니었을 때에 그 의문을 그대로 보여 주는 데 쓰는 종결 어미. 의문이나 긴장 또는 걱정이 해소되었다는 뜻이 암시된다.

난 또 저 꼬마가 널 때린 사람이라고.

⑥ -아야지(-어야지, -여야지)

1. 상대편의 주의를 환기하거나 동의를 구하는 뜻을 나타내는 종결 어미.

그냥 생긴 대로 살아야지.
내가 인제 어 이 뭐 자격증이나 뭔가를 준비할 때 알바를 그만 해야 되겠다. 알바 시간을 뺏긴다 그러면 알바를 접어야지.

2. 독백 투로, 화자의 의지를 나타내는 종결 어미.

- 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분해 준다.(-어, -어요 등)

※ '응', '네', '-어', '-어요' 등과 같이 말끝을 올리거나 내리는 것에 따라 의미가 달라지는 경우,

반드시 마침표와 물음표를 사용하여 구분해 준다.

평서문	의문문
응.	응?
네.	네?
밥 먹었어.	밥 먹었어?

- [인용문 문장부호1] 인용문의 안긴 문장이 1개인 경우는 마침표, 물음표를 붙이지 않고 전체 문장에만 붙인다.

> 제가 늘 요즘에 친구들한테 자주 하는 말 생존을 위한 운동을 해야 된다 그래요.
> -> '된다' 다음에는 마침표를 붙이지 않음

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

> 너 요즘 좀 유난히 버럭버럭하는 거 같아.
> 갱년기 오냐라고 얘기를 했대요.

- 맞장구 발화로 인해 끊어지거나, 휴지 구간 중 다른 화자의 끼어들기로 문장을 나눴을 경우 연결 어미로 끝나더라도 종결 용법이 아닌 경우(억양 등으로 구분)에는 종결 부호를 사용하지 않는다.

3.18. 기타 지침

- 발음 전사를 위해 사용한 기호(예: -, {}, &, (()))는 철자 전사에는 사용하지 않는다.
- STT가 숫자를 아라비아 숫자로 인식했을 때 한글로만 전사

> STT_ 1시간이나 늦었어. → 한 시간이나 늦었어.
> STT_ 40살에 애를 낳아? → 마흔 살에 애를 낳아?

[붙임2] 개인 정보 수집·이용 동의서

개인정보 수집·이용 동의서

㈜나라지식정보, ㈜팀벨은 국립국어원의 “2025년 일상 대화 자료 수집 및 정제” 과제에 참여하여 [개인정보보호법] 제15조 및 제17조에 따라 아래의 내용으로 개인정보를 수집·이용합니다. (개인정보 수집·이용 동의에 거부할 수 있으며, 미동의시 과제참여가 불가능합니다)

개인정보 **수집·이용**자	개인정보 수집·이용 목적	수집·이용 개인정보 항목	보유/이용기간
㈜나라지식정보, ㈜팀벨	◆ (주)팀벨 - 일상 대화 자료 수집 및 정제 과제의 음성 말뭉치 수집 및 전사 ◆ ㈜나라지식정보 - 과제 중 음성데이터 전사 검수, 개인식별정보 등 제거 업무, 과제관리, 최종데이터 검수 업무(하자보수 포함)	발화 음성, 인구통계학적 정보 (출생지/성장지/거주지/성별/연령대/화자간 관계/직업/학력)	2026년 12월 20일까지
㈜팀벨	과제 참여에 대한 회계 정산 처리 및 비용 증빙	성명, 연락처, 주민등록번호, 계좌정보	2025년 12월 20일까지

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 2025년 12월 20일까지입니다. 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.]

2025년　월　일

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

㈜나라지식정보, ㈜팀벨 귀중

[붙임3] 개인 정보 제3자 제공 동의서

개인정보 제3자 제공 동의서

㈜나라지식정보, ㈜팀벨은 국립국어원의 “2025년 일상 대화 자료 수집 및 정제” 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 자료 수집 및 정제 과제의 음성 말뭉치 및 인구통계학적 정보의 기초정보 구분 ◆ 일상 대화 자료 수집 및 정제 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)	2026년 12월 20일까지

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.]

2025년 월 일

신청인 성명 : _____ (자필서명)

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

㈜나라지식정보, ㈜팀벨 귀중

[붙임4] 국립국어원의 개인 정보 제3자 제공(공개) 동의서

국립국어원의 개인정보 제3자 제공(공개) 동의서

본인은 국립국어원의 “2025년 일상 대화 자료 수집 및 정제” 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 자료 수집 및 정제 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 2046년 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.]

2025년 월 일

신청인 성명 : _____ (자필서명)

(신청인이 만 14세 미만 아동인 경우) 법정대리인 관계 : _____ 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 용역의 수행을 위하여 필요한 기간 동안 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하고, 대상저작물의 이용허락 기간 동안 그 대상저작물의 음성을 청취·전사한 텍스트를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 시행 사업 및 국립국어원이 발주한 용역 사업의 원시자료로 활용되는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물의 음성을 청취·전사하여 텍스트로 변형하고 그 텍스트를 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
4. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물·2차적저작물을 학계·연구기관·산업체 등이 연구 및 기술 개발용으로 이용할 수 있도록 제공하는 일
5. 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물·2차적저작물을 제공받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물·2차적저작물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용 허락 기간)
대상저작물의 이용 허락 기간은 계약체결일부터 2046년 12월 31일까지로 하며, 계약기간 만료 시 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 계약기간 만료 시 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다.

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

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하고, 대화 녹음 등 본 계약 이행에 필요한 협조를 하여야 한다.

(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) 이용자는 이용자 및 이용자가 발주한 용역사업의 수행자, 그 복제·변형물을 제공받은 학계, 연구기관, 산업체 등이 제2조의 이용 허락을 받은 범위 이외에서 대상 저작물이 활용되지 않도록 대상저작물의 복제·변형물인 말뭉치의 제공 시 활용목적을 확인하고 주의사항을 명확하게 고지하는 등의 관리 의무를 다하여야 하며, 관련계약 체결 시 해당 내용을 반드시 포함하여야 한다.

(3) 이용료는 설정하지 아니한다.

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

(5) 이용자는 대상저작물의 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 따른 목적에 한하여 제2조에 따른 변형을 할 수 있으며, 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 계

약 이행에 필요하지 않은 내용은 삭제하고 이용하는 점에 동의한다.

제6조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 대상저작물의 저작권 이용 허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
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.

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.

2025년 월 일

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

[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 “2025년 일상 대화 자료 수집 및 정제” 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”

2025년 월 일

미성년자 성명 :

법정대리인(보호자) 관계 : _____ 성명 : _____ (자필서명)

국립국어원 귀중

<기획·연구>

국립국어원 이현주 언어정보과장

국립국어원 박미영 학예연구관

국립국어원 장연지 연구원

<사업 참여자>

사업 책임자 이용주((주)나라지식정보)

사업 참여자 박분선, 이지현, 황주영, 박연미,
박영훈, 이재혁, 오지혜 ((주)나라지식정보)
김희영, 박한주, 차정훈, 신은주,
김관철, 김선희, (주식회사 팀벨)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2025년 12월 20일
발행일: 2025년 12월 20일
인 쇄: ㈜세광디앤피

※ 이 책은 국립국어원의 용역비로 수행한 '2025년 일상 대화 자료 수집 및 정제' 사업의 결과물을 발간한 것입니다.