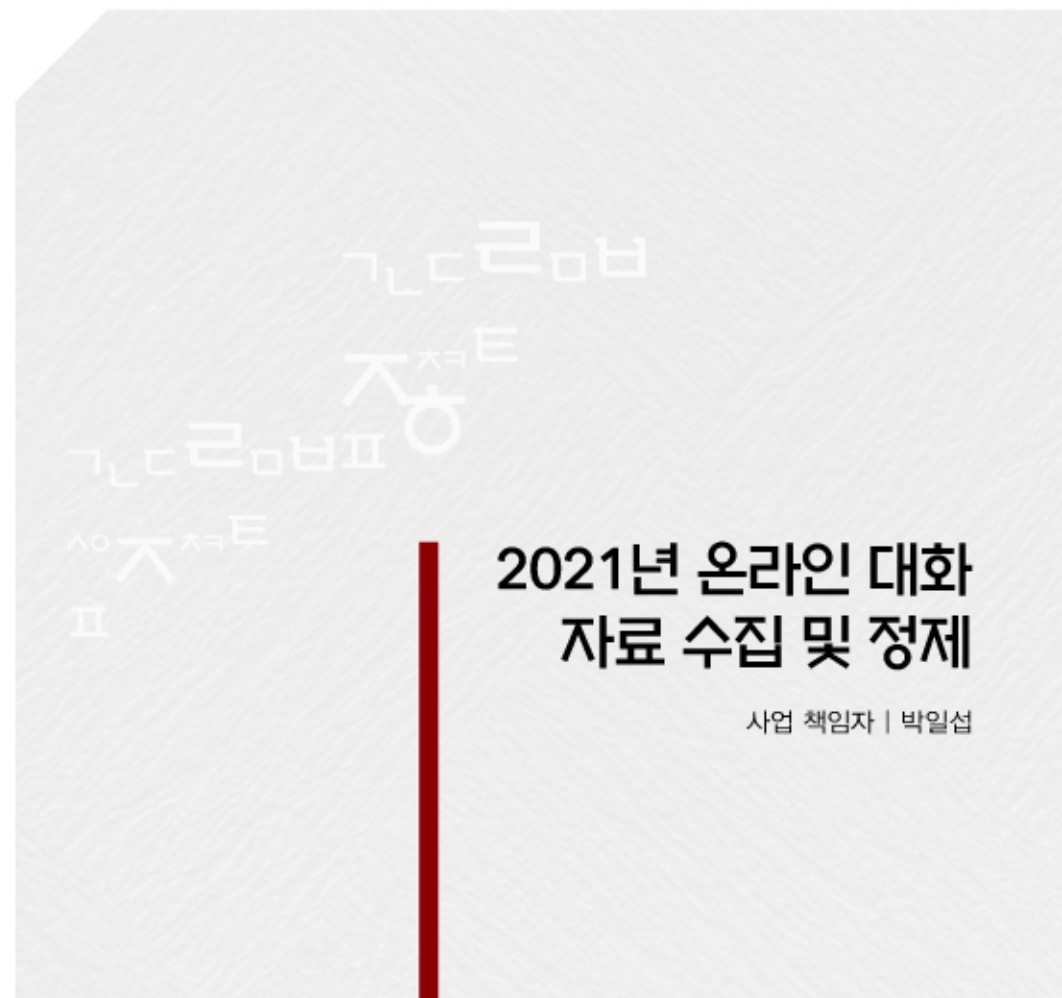




NATIONAL INSTITUTE OF KOREAN LANGUAGE



2021년 온라인 대화 자료 수집 및 정제

사업 책임자 | 박일섭

국립국어원 2021-01-09

발 간 등 록 번 호
11-1371028-000860-01

2021년 온라인 대화 자료 수집 및 정제

사업 책임자
박 일 섭

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2021년 온라인 대화 자료 수집 및 정제’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2021년 4월 ~ 2021년 11월

2021년 11월 13일

사업 책임자: 박일섭(주식회사 미디어코퍼스)

사업 수행기관	주식회사 미디어코퍼스 주식회사 다이얼로그디자인에이전시 심심이(주)
사업 책임자	박일섭
사업 참여자	신지영, 남서정, 조재원, 안 윤, 현재홍, 이수경, 양성민, 이태강, 양리아, 이나리, 이광진, 임현승, 최정희, 홍미미, 조재훈, 김민재

<사업 수행자>	주식회사 미디어코퍼스 주식회사 다이얼로그디자인에이전시 심심이(주)
----------	--

사업 책임자	박일섭(주식회사 미디어 코퍼스)
사업 참여자	신지영(주식회사 미디어 코퍼스)
	남서정(주식회사 미디어 코퍼스)
	조재원(주식회사 미디어 코퍼스)
	안 윤(주식회사 미디어 코퍼스)
	현재홍(주식회사 미디어 코퍼스)
	이수경(주식회사 미디어 코퍼스)
	양성민(주식회사 다이얼로그디자인에이전시)
	이태강(주식회사 다이얼로그디자인에이전시)
	양리아(주식회사 다이얼로그디자인에이전시)
	이나리(주식회사 다이얼로그디자인에이전시)
	이광진(주식회사 다이얼로그디자인에이전시)
	임현승(심심이(주))
	최정희(심심이(주))
	홍미미(심심이(주))
	조재훈(심심이(주))
	김민재(심심이(주))

2021년 온라인 대화 자료 수집 및 정제

이 사업의 목적은 언어와 사회 문화, 자연어 처리와 빅 데이터, 인공 지능 산업 등, 다양한 분야의 연구와 개발에 널리 쓰이도록 한국어 일상 온라인 대화를 국가 공공 언어 자원으로서 활용 가치가 높은 온라인 대화 말뭉치로 구축하는 것이다.

온라인 대화의 특성을 대표성 있게 반영하는 말뭉치가 되도록 대화 참여자 표본을 설계했다. 그리고 다양한 유형의 온라인 대화를 말뭉치에 포함하기 위하여 대화 주제와 온라인 대화의 유형을 설계했다. 대화 주제는 주제 대화와 일상 대화, 2021년의 시대적 상황을 반영하도록 선정한 시사 및 트렌드 주제 대화로 구분했다. 온라인 대화의 유형은 참여자 간 상호 작용에 따라 대화 참여자의 수, 대화 참여자 간 관계, 친밀도, 연락 빈도로 구분했다. 사용 기기와 매체 특성에 따라 카카오톡 대화와 심심이 채팅 대화, 사용 기기, 키보드 유형을 구분했다. 수집 방법에 따라 실시간 대화 수집과 기존 대화 수집으로 유형을 구분했다.

국가 공공 언어 자원으로 공개하기 위해 자료 수집 단계에서 대화 참여자 전원으로부터 개인정보 수집 및 이용 동의, 저작권 이용 허락을 받았다. 자료 가공 단계에서는 개인정보의 비식별화와 혐오 표현, 차별 표현과 같은 비윤리적인 표현을 정제했다. 이를 통해 개인 연구자, 기관, 산업체에서 법적 제한 없이 활용하도록 했다.

수집한 자료는 개인정보 비식별화와 태깅, 특수 메시지 태깅, 비윤리적 표현 정제와 태깅, 대화 분할 및 주제 태깅 이후 사전 정의한 형식과 구조에 따라 JSON 형식의 말뭉치로 가공이 이루어졌고, 정제 작업 과정과 산출물 생성 이후에 구조와 형식 및 내용의 적합성에 대한 검수를 거쳐 최종 결과물로 만들어졌다.

최종적으로 3,514명이 대화에 참여한 4,761개의 대화 파일을 수집했다. 수집한 대화 파일의 전체 규모는 대화 수 기준 151,004개, 말차례 수 기준 2,283,178개, 발화 수 기준 4,120,382개이다.

구축한 원시 말뭉치는 온라인 대화의 고유한 특성을 그대로 반영하고 있어 온라인 대화 언어의 형태와 사용 양상 특성을 연구하는 자료로 활용할 수 있다. 그리고 대화 모델 등 인공 지능 분야 연구자나 개발자가 원하는 형태로 자유롭게 변형해서 학습에 활용할 수 있도록 비교적 짧고 정제된 대화도 일정 분량을 구축했다.

주요어: 온라인 대화, 메신저 대화, 카카오톡, 대화, 원시 말뭉치, 자연어 처리

차 례

제 1 장 사업 개요

1. 사업의 목적	3
2. 사업의 범위	7

제 2 장 온라인 대화 말뭉치 구축 절차

1. 사업 수행 절차 및 진행 일정	11
1.1. 온라인 대화 말뭉치 구축 공정	11
1.2. 온라인 대화 말뭉치 구축 진행 일정	12
2. 계획 단계	13
2.1. 온라인 대화 제공자의 구성 설계	13
2.2. 온라인 대화 유형 구성 설계	15
3. 수집 단계	22
3.1. 온라인 대화 자료 수집 절차	22
3.2. 온라인 대화 수집 홍보	30
3.3. 자료 선별	34
4. 가공 단계	37
4.1. 가공 절차	37
4.2. 정제 및 태깅	38
4.3. 산출물 생성	46
5. 검수 단계	56
5.1. 정제와 태깅 작업에 대한 검수	56
5.2. 최종 산출물에 대한 검수	56

차 례

제 3 장 온라인 대화 말뭉치 구축 결과

1. 온라인 대화 말뭉치의 구성	61
1.1. 구축 규모	61
1.2. 유형별 구성	61
2. 온라인 대화 말뭉치의 참여자 구성	73
2.1. 성별 및 연령	73
2.2. 직업	74
2.3. 지역	76
2.4. 기기 및 키보드 유형	77

제 4 장 마무리 및 제언

참고문헌	85
------------	----

표 차례

〈표 1-1〉 해외 대화 데이터셋 구축 현황	4
〈표 1-2〉 국내 정부 사업 온라인 대화 구축 현황	5
〈표 1-3〉 2021년 온라인 대화 자료 수집 및 정제 사업 수행 세부 목표	6
〈표 1-4〉 2021년 온라인 대화 자료 수집 및 정제 사업 개요	7
〈표 1-5〉 2021년 온라인 대화 자료 수집 및 정제 사업 범위	8
〈표 2-1〉 국내 정부 사업 온라인 대화 제공 인원의 성·연령별 구성	15
〈표 2-2〉 2021년 온라인 대화 말뭉치 성·연령별 표본 구성 목표	15
〈표 2-3〉 2019년 국립국어원 메신저 대화 말뭉치의 유형 분류 기준	16
〈표 2-4〉 온라인 대화 말뭉치의 유형 분류 기준	16
〈표 2-5〉 온라인 대화 말뭉치 참여자 간 상호 작용 양상에 따른 유형 분류	17
〈표 2-6〉 온라인 대화 말뭉치 사용 매체에 따른 유형 분류	19
〈표 2-7〉 온라인 대화 말뭉치 사용 매체에 따른 유형 분류	19
〈표 2-8〉 온라인 대화 말뭉치 사용 매체에 따른 유형 분류	21
〈표 2-9〉 참여자 등록 사이트의 화자 정보 수집 항목	23
〈표 2-10〉 온라인 대화 말뭉치 주제 선택 및 예시 키워드 항목	25
〈표 2-11〉 6월~7월 시사/일상 트렌드 주제 목록	26
〈표 2-12〉 8월~9월 시사/일상 트렌드 주제 목록	27
〈표 2-13〉 9월~10월 시사/일상 트렌드 주제 목록	28
〈표 2-14〉 실시간 대화 지침	28
〈표 2-15〉 대화 정보 수집 항목	30
〈표 2-16〉 대화 참여자 모집 홍보 채널 운영	31
〈표 2-17〉 구축 대상에서 제외되는 실시간 대화의 선별 기준	35
〈표 2-18〉 작업 보안 유의 사항 안내 예시 및 주요 항목	38
〈표 2-19〉 온라인 대화 자동 전처리 대상 정의	40
〈표 2-20〉 2021 온라인 대화 비식별화 기본 지침	41
〈표 2-21〉 2021 온라인 대화 개인정보 비식별화 항목별 처리 지침	42
〈표 2-22〉 특수 메시지 태깅 방법	43
〈표 2-23〉 2021 온라인 대화 자료 파일명 부여 방식 및 파일명 작성 예시	47
〈표 2-24〉 2021 온라인 대화 개인정보 비식별화 항목의 산출물 표기 형식	48
〈표 2-25〉 2021 온라인 대화 특수 메시지 및 비윤리적 표현 등 산출물 표기 형식	48
〈표 2-26〉 2021 온라인 대화 말뭉치 JSON 구조	50
〈표 2-27〉 태깅 대상 항목의 txt 원문과 원시 말뭉치 JSON 형식 표기 예시	52

표 차례

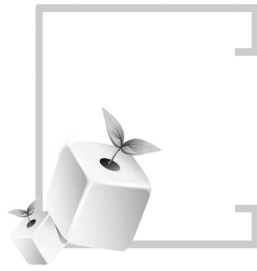
〈표 2-28〉 산출물 형식 오류 검수 및 말뭉치 가공 도구 수정 요청 예시	57
〈표 3-1〉 온라인 대화 원시 말뭉치의 구축 수량	61
〈표 3-2〉 대화 참여 인원별 온라인 대화 말뭉치 구성	62
〈표 3-3〉 수집 방법별 온라인 대화 말뭉치 구성	63
〈표 3-4〉 수집 매체별 온라인 대화 말뭉치 구성	64
〈표 3-5〉 주제 유형별 온라인 대화 말뭉치 구성	65
〈표 3-6〉 주제 대화 세부 주제별 온라인 대화 말뭉치 구성	65
〈표 3-7〉 대화 참여자 간 관계별 온라인 대화 말뭉치 구성	66
〈표 3-8〉 대화 참여자 간 친밀도별 온라인 대화 말뭉치 구성	68
〈표 3-9〉 대화 참여자 간 관계에 따른 친밀도 구성(수집 파일 기준)	69
〈표 3-10〉 대화 참여자 간 친밀도별 온라인 대화 말뭉치 구성	71
〈표 3-11〉 연락 빈도에 따른 친밀도 구성(수집 파일 기준)	72
〈표 3-12〉 온라인 대화 참여자의 성별 및 연령 구성	73
〈표 3-13〉 온라인 대화 참여자의 직업 구성	75
〈표 3-14〉 온라인 대화 참여자의 경제 활동 유무에 따른 구성	76
〈표 3-15〉 온라인 대화 참여자의 지역 구성	77
〈표 3-16〉 온라인 대화 참여자의 대화 시 주요 사용 기기 구성	78
〈표 3-17〉 온라인 대화 참여자의 키보드 유형 구성	78

그림 차례

[그림 1-1] 인스턴트 메신저 이용률 증가 추이	3
[그림 2-1] 2021 온라인 대화 말뭉치 구축 공정	11
[그림 2-2] 2021 온라인 대화 말뭉치 구축 주요 진행 일정	12
[그림 2-3] 성·연령별 인스턴트메신저 이용률	14
[그림 2-4] 성·연령별 주 이용 인스턴트 메신저 서비스	18
[그림 2-5] 키보드의 유형	19
[그림 2-6] 2019년 일반 수집 대화와 수집 봇(bot) 수집 대화 비교 예시	20
[그림 2-7] 2021 온라인 대화 자료 수집 절차	22
[그림 2-8] 참여자 등록 화자 정보 입력 과정	23
[그림 2-9] 참여자 등록 개인정보 수집·이용 동의 및 저작권 이용 허락 계약 과정	24
[그림 2-10] 개인정보 수집·이용 동의서 및 주요 내용	24
[그림 2-11] 저작권 이용 허락 계약서 및 주요 내용	25
[그림 2-12] 기타 온라인 채팅(심심이) 주제 선택 화면	26
[그림 2-13] 기타 온라인 채팅(심심이) 대화 기능 화면	29
[그림 2-14] 기타 온라인 채팅(심심이) 대화 기능의 주제 키워드 입력 화면	30
[그림 2-15] 온라인 대화 말뭉치 사업 공식 사이트의 사업 소개 내용	31
[그림 2-16] 온라인 대화 말뭉치 카카오톡 홍보 채널 및 홍보 메시지 발송 예시	32
[그림 2-17] 주관 기관 누리집 공지 및 공고	33
[그림 2-18] 대화 수집 이벤트 홍보물	33
[그림 2-19] 온라인 대화 수집 참여 인원 추이	34
[그림 2-20] 구축 대상에서 제외되는 실시간 대화의 실제 사례	35
[그림 2-21] 구축 대상에서 제외되는 기존 대화의 실제 사례	36
[그림 2-22] 2021 온라인 대화 자료 가공 절차	37
[그림 2-23] 작업자 교육 및 지침 공유를 위한 노선, 구글 공유 시트 활용 예시	39
[그림 2-24] 전처리 전 원문 및 정제 작업 파일 생성 예시	41
[그림 2-25] 비윤리적 표현 정제 기준 수립을 위한 공유 시트 활용 예시	45
[그림 2-26] 비윤리적 표현 정제 기준 수립을 위한 주관 기관 협의 진행	45
[그림 2-27] 2021 온라인 대화 비윤리적 표현 정제 관련 지침 주요 내용	45
[그림 2-28] 대화 분할 및 주제 태깅 작업 예시	46
[그림 2-29] 2021 온라인 대화 자료 산출물 생성 과정	47
[그림 2-30] 온라인 대화 말뭉치 메타 정보 JSON 형식	52
[그림 2-31] 대화 원문 txt 형식과 대화 내용 JSON 형식	52

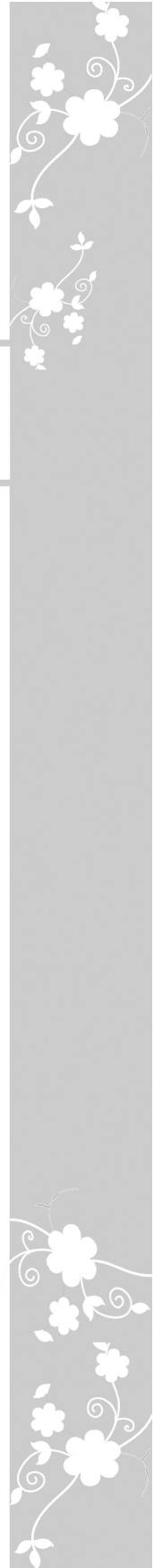
그림 차례

[그림 2-32] 구글 공유 시트를 활용한 정제와 태깅 작업 결과물에 대한 교차 검증 기록 예시 ..56	
[그림 2-33] 상용 소프트웨어를 사용한 산출물 전체 작업 오류 검수 예시58	
[그림 3-1] 키보드의 유형	78



제 1 장

사업 개요

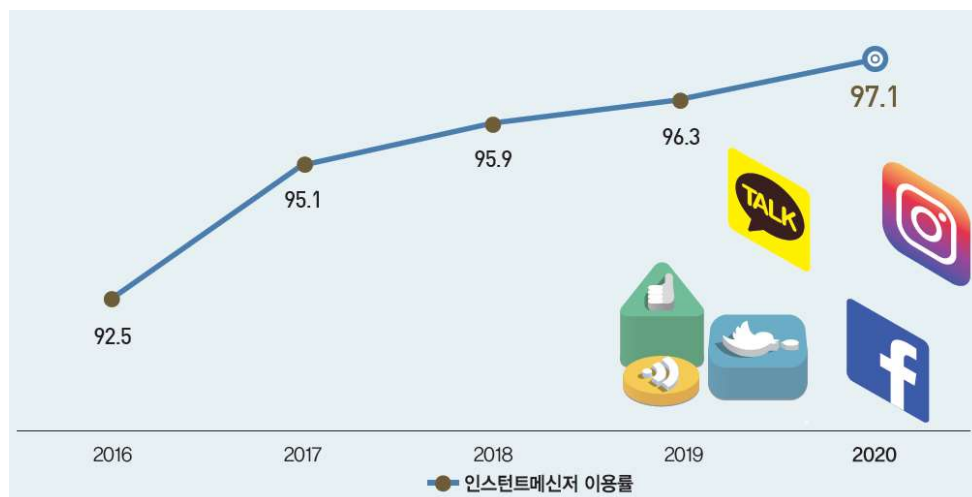


1. 사업의 목적

이 사업은 한국어 온라인 대화 자료를 수집하고 이를 정제하여 원시 말뭉치로 구축하는 사업이다.

이 사업의 목표는 다음과 같다. 먼저 사업의 최종 결과물인 온라인 대화¹⁾ 말뭉치가 언어와 사회 문화, 자연어 처리와 빅 데이터(big data), 인공지능 산업 등 학계와 산업계 등 다양한 분야의 연구와 개발에 널리 쓰이는 것이다. 이를 통해 한국어 사용자의 일상 온라인 대화가 국가 공공 언어 자원으로 활용 가치를 갖도록 하는 것이 본 사업의 궁극적인 목적이다.

온라인 대화는 우리 일상 언어 활동의 주요 방식 가운데 하나이다. 한국인터넷진흥원(2021)에 따르면 온라인 대화의 주요 매체인 인스턴트 메신저 이용률은 [그림 1-1]과 같이 2016년부터 2020년까지 꾸준한 증가 추세가 나타난다.



[그림 1-1] 인스턴트 메신저 이용률 증가 추이(한국지능정보사회진흥원, 2021:37)

특히 코로나(COVID-19)의 확산과 함께 비대면 환경에서 이루어지는 온라인 의사소통 방식은 일상 대화만이 아니라, 생활 편의를 돕는 여러 서비스에서도 활용되고 있다. 이러한 서비스는 인공지능 기술과 결합한 챗봇(chatbot)의 형태로 주문, 예약, 결제를 비롯하여 민원 처리, 상담 등 그 이용 범위가 나날이 확대되고 있다.

해외에서는 <표 1-1>과 같이 상담과 예약 모델, 자유 대화 등의 다양한 대화 데이터

1) ‘온라인 대화’는 웹과 모바일 서비스를 통한 온라인 공간에서 발생하는 모든 종류의 양방향 의사소통을 의미한다. 이러한 온라인 대화는 ‘메신저 대화’, ‘채팅 서비스를 이용한 대화’, ‘누리소통망(SNS)의 DM(Direct Message)를 통한 대화’, ‘커뮤니티 게시글의 댓글로 진행되는 대화’ 등 다양한 유형이 있지만, 본 사업에서는 다양한 유형의 온라인 대화 중에서도 ‘메신저’나 ‘채팅’과 같이 그 본래의 서비스 목적이 양방향 대화에서 출발하는 유형만을 사업의 범위로 한정한다.

를 구축하고 있다.²⁾ 그리고 이러한 데이터에는 누리소통망이나 커뮤니티 등에서 수집한 댓글과 같은 온라인 대화도 포함하는 추세이다.

데이터 명칭	분야	구축 규모	설명
MultiWOZ 2.1	예약	10,000건 가량	클라우드 소싱을 통해 수집한 호텔, 레스토랑, 기차표 예매 대화
Frames Dataset	여행 예약	20,000턴 가량 1,369건	여행 예약 모델 학습용 대화
Ubuntu Dataset	문제 해결을 위한 상담	건당 평균 8턴 대화 930,000건 가량	우분투 사용 시 기술 지원 채팅 일부 정제 가공
ConvAI2	자유 대화 (Open Domain)	훈련 데이터 9,013건 검증 데이터 968건 테스트 데이터 1,000건	일상 대화 범용 데이터 구축을 목적으로 실시되는 경진 대회를 통한 데이터 구축
DailyDialog	일상 대화	건당 평균 7.9턴 대화 13,000건 가량	잡담 형태의 대화 모델 학습에 활용
CCPE Taskmaster-1	영화에 대한 대화 주문 예약	13,700건 가량	영화 대화 500건 6개 주제의 주문 및 예약 대화 13,200건
Reddot All Comments Corpus	주제 토론	최소 4턴 이상 대화 8,000,000건 가량	1,000개의 서브레딧에서 선별된 대화 추출

〈표 1-1〉 해외 대화 데이터셋 구축 현황

온라인 대화가 우리 일상 의사소통의 주요 방식으로 자리 잡고, 챗봇과 같은 자동화된 대화 시스템 기술의 확산에 따라 온라인 대화 말뭉치의 중요도 또한 커지고 있다. 온라인 대화만의 독자적인 언어 사용 방식과 체계를 이해하고, 이에 맞는 어문 규정 등을 정립하기 위해서는 객관적으로 정량화하여 연구할 수 있는 방대한 규모의 자료가 필요하다. 또한 챗봇 형태의 대화 처리 기술을 연구하고 개발하기 위한 학습 데이터로서도 글말(文語)과는 차별화된 온라인 대화의 특성을 담고 있는 말뭉치가 필요하기 때문이다.

이러한 필요를 반영하여 한국 정부도 2019년부터 국가 공공 언어 자원으로서 연구와 개발에 활용할 수 있는 온라인 대화 데이터를 구축하고 있다. 〈표 1-2〉는 정부가 주관하여 구축한 온라인 대화 데이터의 예시이다.

2) 2021년 (주)미디어 코퍼스 자체 조사

데이터 명칭	구축 년도	주관 기관	구축 규모 ³⁾
메신저 대화 말뭉치	2019년	문화체육관광부 국립국어원	건당 평균 10턴 대화 712,291건
한국어 SNS 데이터	2020년	과학기술정보통신부 한국지능정보사회진흥원	건당 평균 4턴 대화 2,000,000건

〈표 1-2〉 국내 정부 사업 온라인 대화 구축 현황

국립국어원의 2019년도 ‘메신저 대화 말뭉치’와 한국지능정보사회진흥원(NIA)의 2020년도 ‘한국어 SNS 데이터셋’ 모두 온라인 대화 말뭉치의 필요를 반영하여 연구와 개발 등의 목적에 활용하기 위해 구축한 데이터이다. 위 사업을 통해 온라인 대화 특성 연구와 대화 처리 기술의 개발 등에 필요한 공공 언어 자원 확보가 일정 규모 이루어진 것으로 평가할 수 있다.

다만 일정 규모의 말뭉치가 확보되었다고 해서 새로운 말뭉치를 구축할 필요가 사라지는 것은 아니다. 왜냐하면 언어는 끊임없이 변하기 때문이다. 새로운 표현이 유행하기도 하고, 새로운 단어나 표현이 만들어지기도 한다. 그리고 기존 낱말과 표현의 의미나 사용 방식이 달라지기도 한다. 또한 상대방과 대화를 하고 말이 통한다는 것은 시대와 사회문화적인 상황 맥락을 이해하고 그러한 맥락을 고려하는 것까지도 포함한다. 그렇기 때문에 시간의 흐름에 따른 언어의 변화상을 반영할 뿐만 아니라, 시대상을 반영하는 사회문화적 언어 지식 사전의 역할을 하기 위해서도 말뭉치 구축은 시간적 연속성을 지녀야 한다.

앞서 언급한 사항을 고려하여 2021년 한국어 온라인 대화 말뭉치를 구축하는 것이 본 사업의 목표이며, 그 세부 내용은 다음과 같다.

첫째, 한국어 온라인 대화의 언어 자원으로서 가치 증대에 기여하는 말뭉치를 구축한다. 온라인 대화의 언어적 특성과 2021년의 시대상을 반영하도록 하여 언어 현상과 언어 현상을 둘러싼 사회문화적 현상까지 두루 연구할 수 있는 데이터로서 가치를 갖도록 한다. 또한 인공 지능 기반의 자연어처리와 한국어 대화 처리 모델의 성능 향상, 대화형 시스템 연구 개발 등 인공 지능 학습용 데이터로서 가치를 갖도록 한다.

둘째, 국가 공공 언어 자원으로서 연구와 개발 목적에 활용하려는 사람이라면 누구나 활용이 가능하도록 수집 과정에서 대화 제공자 전원으로부터 개인정보 수집에 대한 동의와 저작권 이용 허락 계약을 체결한다. 그리고 대화에 포함된 개인정보는 개인정보 보호위원회의 규정을 반영하여 비식별화한다.

이 사업의 세부 목표를 정리하면 〈표 1-3〉과 같다.

3) 국립국어원의 메신저 대화 말뭉치는 대화 한 건당 말차례가 최소 10회 이상 포함되어 있으며, 한국지능정보사회진흥원의 ‘한국어 SNS 데이터’는 대화 한 건당 말차례가 최소 4회 이상 포함되어 있다.

항목	내용
실용성, 활용도 높은 온라인 대화 말뭉치 구축	<ul style="list-style-type: none"> • 산업계 수요가 높은 주제, 도메인, 개체명 카테고리 포함 • 추가적인 데이터 전처리를 요구하는 노이즈 최소화 • 생략이 빈번하고 대화자 간의 대화 맥락이 불분명한 대화 수집 최소화
현대 한국어 온라인 대화 최적화 말뭉치 구축	<ul style="list-style-type: none"> • 2021년 기준 최신 트렌드, 대중적 관심사 반영하는 대화 포함 • 온라인 대화체와 신조어, 최신 트렌드 용어 반영된 대화 포함
법적 문제 발생 우려 없는 국가 공공재 온라인 대화 말뭉치 구축	<ul style="list-style-type: none"> • 국가 법령에 기반한 개인정보 수집 및 제공 동의, 개인정보 비식별화 실시 • 대화 자료 제공자 전원 저작권 이용 허락 동의 확보 및 상대방의 동의 없는 수집 자료 배제 • 차별, 혐오 등 민감 발언 포함 대화 별도 처리 및 관리

〈표 1-3〉 2021년 온라인 대화 자료 수집 및 정제 사업 수행 세부 목표

2. 사업의 범위

이 사업의 개요는 <표 1-4>와 같다.

항목	내용
사업명	2021년 온라인 대화 자료 수집 및 정제
사업 수행 기간	2021년 4월 13일~ 2021년 11월 13일
주관 기관	문화체육관광부 국립국어원
사업 수행자	(주)미디어 코퍼스 컨소시엄 <ul style="list-style-type: none"> • (주)미디어 코퍼스 • (주)다이얼로그 디자인 에이전시 • 심심이(주)
주요 목표	<ul style="list-style-type: none"> • 2인 대화와 다자 대화 10% 이상을 포함하는 화자 전환 8회 이상의 150,000건 온라인 대화 말뭉치 구축 • 성별, 연령, 지역 균형을 고려한 대화 제공자 모집 • 수집 자료의 개인정보 비식별화 처리 • 대화 제공자 전원으로부터 저작권 이용 허락 계약 체결

<표 1-4> 2021년 온라인 대화 자료 수집 및 정제 사업 개요

이 사업의 범위는 크게 온라인 대화 말뭉치 구축 지침 수립, 온라인 대화 자료 획득, 온라인 대화 말뭉치 구축, 온라인 대화 자료 품질 검증, 사업 운영 및 관리의 다섯 분야로 나눈다.

온라인 대화 말뭉치 구축 지침 수립은 말뭉치 구축 자료 선정 단계부터 작업 단계까지 일관성 확보를 통한 말뭉치의 품질을 보장하는 과정이다. 지침 수립 과정은 자료 수집 대상 선별 기준 수립, 대화 주제, 수집 방식 등 데이터 구성 기준 수립, 작업 방법론 및 정제와 구축 지침 수립, 품질 관리 지침 등을 포함한다.

온라인 대화 자료 획득은 사업 목표에 부합하는 대화를 수집하기 위하여 홍보와 광고 전략 수립 및 시행, 대화 참여자 모집 및 대화 수집 운영, 대화 제공자 정보 수집과 개인정보 수집 및 이용 동의 획득, 저작권 이용 허락 계약 체결 등을 포함한다.

온라인 대화 말뭉치 구축은 사업 목표에 부합하는 말뭉치를 구축하기 위하여 자료 정제, 메타 정보 구축, 개인정보 비식별화, JSON 생성 등을 포함한다.

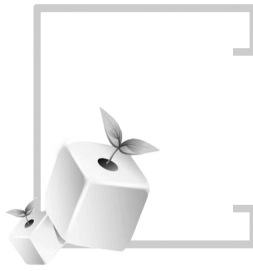
온라인 대화 자료 품질 검증은 말뭉치의 품질 확보를 위하여 품질 관리 조직과 정책을 운영하고 말뭉치 품질 기준의 검증 등을 포함한다.

사업 운영 및 관리는 단계별 산출물 납품과 사업의 전체적인 일정과 진행 관리 등을 포함한다.

이를 정리하면 <표 1-5>와 같다.

사업의 범위		세부 내용
온라인 대화 말뭉치 구축 지침 수립		<ul style="list-style-type: none"> 온라인 대화 참여 인원, 성별, 나이, 직업, 지역 등 선별 기준 수립 온라인 대화 주제, 수집 방법 등 구성 및 수집 비율 설계 온라인 대화 내 개인정보 비식별화 및 민감 발언 정제 지침, 방법론 수립 / 교육 작업 지침 수립 / 교육 품질 지침 설계 및 전 공정의 품질 관리/점검
온라인 대화 자료 획득		<ul style="list-style-type: none"> 온라인 대화 수집을 위한 홍보 및 광고 실행 온라인 대화 참여자 모집 및 대화 수집 온라인 대화 제공자로부터 개인정보 수집 및 이용 동의 획득 대화 제공자 전원으로부터 저작권 이용 허락 계약 체결
온라인 대화 말뭉치 구축		<ul style="list-style-type: none"> 온라인 대화 데이터 정제 메타 정보 구축 및 헤더 변환 개인정보 비식별화 / 특수 메시지 태깅 JSON 형식의 원시 말뭉치 구축
온라인 대화 자료 품질 검증		<ul style="list-style-type: none"> 품질 관리 조직 및 정책 운영 데이터 품질 기준 검증/시정/품질 확보
사업 운영 및 관리	단계별 산출물 납품	<ul style="list-style-type: none"> 착수 단계 - 사업수행계획서, 비밀유지계약서, 서약서 중간 단계 - 30% 분량 1차 납품, 60% 분량 2차 납품 완료 단계 - 최종 보고서, 온라인 대화 원문 및 원시 말뭉치, 메타 정보 자료, 저작권 이용 허락 계약서 원본 및 사본
	프로젝트 관리	<ul style="list-style-type: none"> 공정 관리/위험 관리/자원 관리/일정 관리/보안 관리/산출물 관리 수행 일정 계획/세부 활동 도출/중간목표 수립/자원 배분 개발 장비 확보 및 운영 보고 관리 - 정기 보고 및 비정기 보고
	프로젝트 지원	<ul style="list-style-type: none"> 품질 보증 계획 수립 및 품질 확보 교육 훈련 - 교육 훈련 방법/내용/일정/조직 운영/기술지원 하자 보수 계획 수립 및 보증 비상 대책 - 백업/복구 대책, 장애 대응 대책 수립/운영

<표 1-5> 2021년 온라인 대화 자료 수집 및 정제 사업 범위



제 2 장

온라인 대화 말뭉치 구축 절차

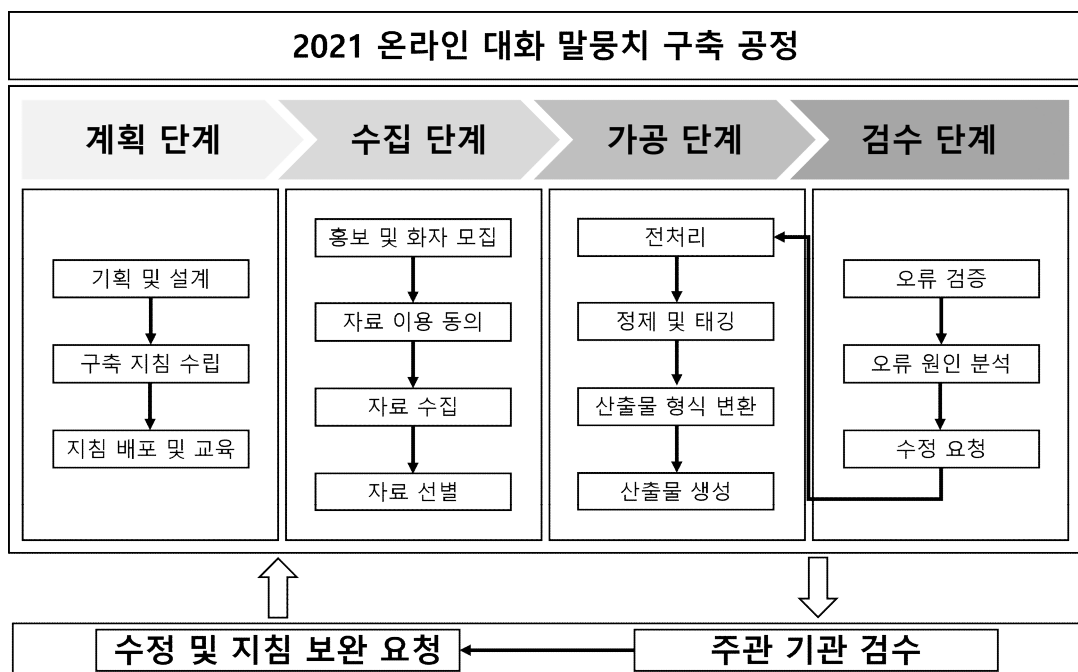


1. 사업 수행 절차 및 진행 일정

1.1. 온라인 대화 말뭉치 구축 공정

한국어 언어 자료를 말뭉치로 구축하는 과정은 원문 언어 자료의 매체 특성에 따른 절차와 방법의 차이⁴⁾를 보이지만, 대체로 말뭉치 구성 기획, 대상 자료 또는 자료 제공자 선정, 자료에 대한 이용 권리 획득, 원문 및 메타 정보 수집, 자료 정제 및 태깅, 형식 변환 및 산출물 생성, 검수의 단계를 거친다.

2021년 온라인 대화 말뭉치 구축 공정 또한 이러한 절차를 기본으로 하여 계획, 수집, 가공, 검수의 단계로 진행되었다. 말뭉치 구축 지침 수립을 비롯하여 공정 전 과정에서 주요 사항은 주관 기관과 협의하여 조정하였으며, 산출물 납품 후 주관 기관 검수 후 수정 요청에 따라 오류를 수정하고 지침을 보완하면서 공정이 진행되었다. 2021년 온라인 대화 말뭉치 구축 공정 전체는 [그림 2-1]과 같다.

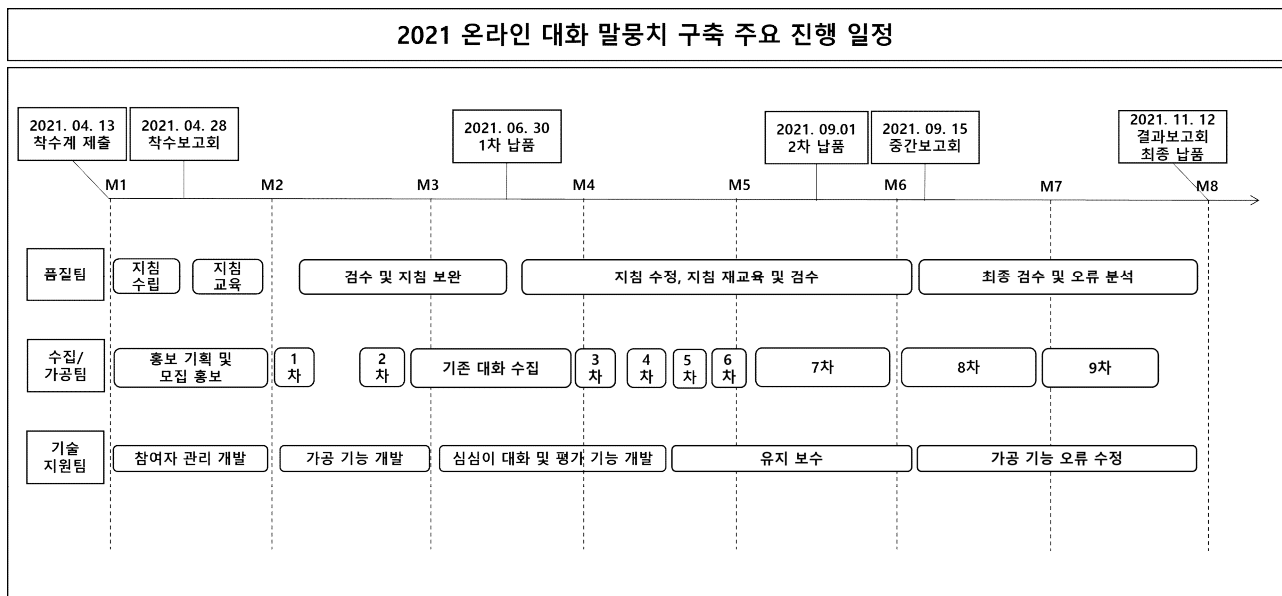


[그림 2-1] 2021 온라인 대화 말뭉치 구축 공정

4) 수집 원문이 입말인 경우는 글말 자료와 달리 음성을 녹음하고 이를 전사하는 과정이 추가로 필요하다. 그리고 음성 자료와 전사 자료 모두 정제하여 말뭉치를 구축하는 것이 일반적이다. 또한 수집 대상 자료의 소유 주체가 누구냐에 따라서도 개인을 대상으로 이용 권리를 획득해야 하는 경우와 출판사 또는 신문사 등의 기관을 통해 이용 권리를 획득해야 하는 경우와 같이 작업 절차와 방법이 달라질 수 있다.

1.2. 온라인 대화 말뭉치 구축 진행 일정

이 사업은 2021년 4월 13일 착수를 시작으로 하여 2021년 11월 12일 최종 납품까지 7개월 동안 진행되었다. 사업의 수행 범위를 기준으로 공정 업무의 효율적인 진행을 위해 품질, 수집 및 구축, 기술 지원으로 업무를 분장했다. 각 부문별 진행 일정을 조율하면서 사전 계획된 1차, 2차, 최종 납품 일정을 고려하여 사업을 진행했다. 이 사업의 주요 진행 일정은 [그림 2-2]와 같다.



[그림 2-2] 2021 온라인 대화 말뭉치 구축 주요 진행 일정

2. 계획 단계

활용도 높은 말뭉치를 구축하기 위해서는 말뭉치의 활용 목적과 수집 대상 언어 자료의 매체 특성 등을 고려하여 자료를 수집하고 가공하는 방법에 대한 기준을 설계해야 한다. 그리고 작업자가 일관된 기준에 따라 작업을 할 수 있도록 안내하는 명시적인 지침이 필요하다. 말뭉치를 계획하는 단계는 이러한 기준을 설정하고 지침을 수립하는 과정이다.

이 사업은 특수 목적의 말뭉치가 아니라, 범용성을 지니는 말뭉치를 구축하는 사업이다. 범용 말뭉치로서 온라인 대화 말뭉치는 온라인 대화 양상과 특성을 축소하여 보여줄 수 있도록 자료 제공자 표본을 구성해야 한다. 또한 다양한 온라인 대화를 균형 있게 포함할 수 있도록 대화 주제와 유형을 구성해야 한다.

그리고 온라인 대화는 컴퓨터나 휴대폰과 같은 IT 기기를 매개로 이루어진다. 글말의 형식을 지니지만, 즉각적인 상호작용이 가능하며, 발화의 특성이나 담화 구조 등은 입말의 특성이 나타나기도 한다. 또한 온라인 대화는 이모티콘을 이용한 감정이나 의도 표현이 가능하고, 사진이나 동영상, 파일, 지도, 뉴스 등 다양한 정보의 공유가 가능하다. 더 나아가 송금이나 선물 보내기 등과 같은 기능성 메시지가 발화 메시지와 혼재되어 나타난다.

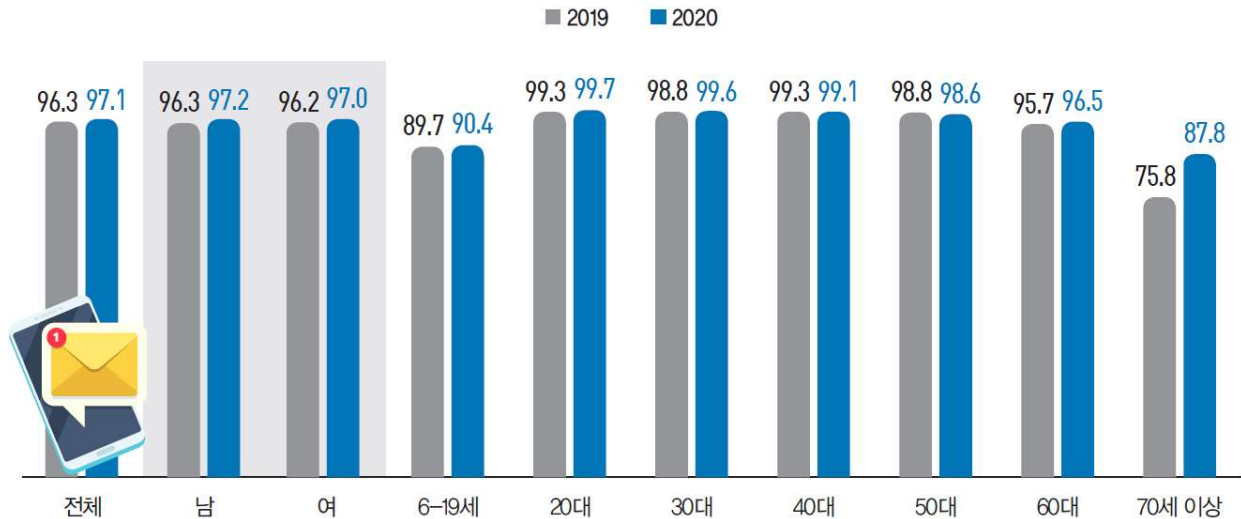
계획 단계에서 고려해야 할 또 다른 하나는 국가 공공 언어 자원으로 공개하기 위해 자료 수집과 가공에 이르는 과정의 적법성을 확보 방법과 지침을 수립하는 것이다. 특히 온라인 대화는 개인 간의 사적인 대화로부터 말뭉치를 구축하는 것이기 때문에 이 과정이 더욱 중요하다.

온라인 대화 말뭉치는 이러한 요소를 고려해서 구축해야 하며, 이를 고려한 설계의 기본 방향은 아래와 같다.⁵⁾

2.1. 온라인 대화 제공자의 구성 설계

특수 목적의 말뭉치가 아닌 범용 말뭉치로서 온라인 대화 말뭉치는 앞서 이야기했듯이 한국어 모집단의 온라인 대화 양상과 특성을 ‘대표성(representativeness)’ 있게 반영해야 한다. 한국지능정보사회진흥원(2021:132~135)에 따르면 한국인 인터넷 이용자의 97.1%는 인스턴트 메신저를 통한 온라인 대화 사용 집단이다. 그리고 [그림 2-3]과 같이 2020년 기준으로 70대를 제외한 전 연령층에서 성별에 관계없이 인스턴트 메신저를 사용하는 비중은 90% 이상으로 나타나고 있다.

5) 이 보고서 2장 계획 단계 부분에서는 위에서 언급한 내용 중 수집 대상 자료의 표본과 유형 구성에 대해서만 다룬다. 온라인 대화 언어 특성을 고려한 세부적인 가공 지침과 개인정보 비식별화 등의 지침은 4장 가공 단계에서 다룬다.



[그림 2-3] 성·연령별 인스턴트메신저 이용률(한국지능정보사회진흥원, 2021:133)

다시 말해 70세 이상을 제외한다면 한국어 온라인 대화 사용자 집단의 표본 구성은 한국의 인구 구성 비율을 따르는 것이 이상적이다.

다만 실제적인 수집 가능성도 표본 구성에 고려해야 할 사항이다. 클라우드 소싱 방식을 통해 다수의 대화 제공자를 모집하는 사업의 특성상 최적의 홍보 효과를 위해 인터넷과 누리소통망 중심의 비대면 환경에서 홍보가 진행이 된다. 그런 만큼 인터넷과 누리소통망 사용에 익숙한 성별이나 연령의 참여율이 상대적으로 높을 가능성이 크다. 또한 대화 제공으로부터 주어지는 보상에 대한 가치 평가도 성별이나 연령에 따라 차이를 보인다.⁶⁾ 그렇기 때문에 동일한 보상에 대해 가치가 높다고 여기는 집단의 참여율이 상대적으로 높을 가능성이 크다.⁷⁾

<표 2-1>의 온라인 대화 자료 수집을 대상으로 하는 2019년과 2020년 정부 사업의 실제 대화 제공자 구성에서도 알 수 있듯이, 남성에 비해 여성의 구성 비율이 높고 20대 구성 비율이 다른 연령에 비해 현저히 높다. 온라인 대화 사용자 집단의 표본 구성에서는 이러한 실제적인 수집 가능성도 고려가 필요하다.

성별과 연령 이외에 직업과 지역 등의 변인도 온라인 대화 자료 구성에서 고려가 필요한 사항이지만, 이 두 변인은 자료 분류 등을 위한 부가적인 정보로만 활용한다.

이러한 점을 고려하여 이 사업에서는 <표 2-2>를 표본 구성의 최종 목표 비율로 설정하였으며, 실제 자료 수집 과정에서 대화 제공자의 참여 현황에 따라 구성 비율을 조정하는 것으로 주관 기관과 협의했다.

6) 사업 수행 경험에 비추어 봤을 때 동일한 보상이 주어질 경우, 직장 생활을 하는 사람에 비해 직장 생활을 하지 않는 사람이 해당 보상에 대한 가치를 더욱 큰 것으로 평가하는 것으로 보인다.

7) 기존의 세종 말뭉치 구어 말뭉치 구축 사례를 보면 20대의 참여율이 상대적으로 높았다. 세종 구어 전사 말뭉치는 연령 미상의 발화자를 제외한 50%가 20대였으며(국립국어원, 2007:19), 연세 구어 말뭉치의 경우에도 연령 미상 발화자를 포함한 전체에서 20대가 차지하는 비중이 52.5%로 나타났다(서상규 외, 2013:100).

구분		남성		여성		합계	
		인원 (명)	비율 (%)	인원 (명)	비율 (%)	인원 (명)	비율 (%)
2019년 메신저 대화 말뭉치	10대	213	2.1	611	6.1	824	8.2
	20대	1,184	11.7	4,293	42.6	5,477	54.3
	30대	761	7.5	1,945	19.3	2,706	26.8
	40대	191	1.9	453	4.5	644	6.4
	50대	63	0.6	230	2.3	293	2.9
	60세 이상	47	0.5	92	0.9	139	1.4
	합계	2,459	24.4	7,624	75.6	10,083	100.0
2020년 한국어 SNS 데이터	20대 미만	-	-	-	-	120,098	2.81
	20대	-	-	-	-	31,411,169	73.54
	30대	-	-	-	-	858,289	20.09
	40대	-	-	-	-	80,797	1.89
	50대	-	-	-	-	54,557	1.28
	60대	-	-	-	-	15,608	0.37
	70세 이상	-	-	-	-	1,090	0.03
	합계	939,028	21.98	3,332,580	78.02	4,271,608	100.00

〈표 2-1〉 국내 정부 사업 온라인 대화 제공 인원의 성·연령별 구성

	10대 비율 (%)	20대 비율 (%)	30대 비율 (%)	40대 이상 비율(%)	합계 (%)
남성 비율(%)	8	12	12	8	40
여성 비율(%)	12	18	18	12	60
합계(%)	20	30	30	20	100

〈표 2-2〉 2021년 온라인 대화 말뭉치 성·연령별 표본 구성 목표

2.2. 온라인 대화 유형 구성 설계

온라인 대화 말뭉치의 설계에서 화자 구성과 함께 고려해야 할 사항은 ‘균형성(balance dness)’ 있게 다양한 유형을 포함할 수 있도록 온라인 대화의 유형을 구성하는 것이다.

국립국어원의 2019년 메신저 대화 말뭉치에서는 〈표 2-3〉과 같이 메신저 대화의 형식과 내용에 영향을 끼칠 수 있는 사항을 참여자 간 상호 작용, 사용 매체, 텍스트의 내용, 사전 통제 유무를 기준으로 메신저 대화 말뭉치의 유형을 분류하는 기준을 제시했다.

참여자 간 상호 작용의 양상은 일상적인 입말 대화도 대화의 내용과 형식에 영향을 미치는 변인이며, 온라인 대화 또한 대화 상대방과의 관계에 따라 발화의 길이나 사용하는 어휘, 이모티콘 사용의 빈도 등이 달라진다. 사용 매체 변인은 온라인 대화가 일상적인 입말 대화와 달리 PC나 스마트폰 등을 매개로 메신저상에서 이루어진다는 점

을 고려한 것이다. 다양한 종류의 메신저가 있고 대화 상대에 따라 메신저를 구분하여 사용하는 경우가 있고 이에 따라 대화의 내용이 달라지는 등 대화의 성격에 영향을 미친다.⁸⁾ 온라인 대화에 사용하는 기기의 기종에 따라 키보드의 형태가 달라지며, 키보드의 형태는 언어 형식에도 영향을 미친다. 대화의 주제는 대화 내용에 영향을 미치는 요인이다. 대화 통제 유무란 대화 수집 상황을 사전에 예고하고 주제를 부여한 후 대화 수집 상황을 인지한 상태로 이루어지는 대화인지, 그렇지 않은지에 따른 구분이다. 대화 상황에 대한 관찰 유무는 어휘 선택과 대화의 내용 등에 영향을 미친다.

기준	항목	분류
참여자 간 상호 작용	대화 참여자의 수	2인 대화 / 다자 대화
	대화 참여자 간 관계	부부 간 대화 / 학교 동기 간 대화 등
	대화 참여자 간 친밀도	친밀도가 높은 관계의 대화 / 낮은 관계의 대화 등
	대화 참여자 간 연락 빈도	거의 매일 연락하는 관계 / 처음 연락하는 관계 등
사용 매체	메신저의 종류	카카오톡 / 라인 / 페이스북 메신저 / 네이버온 등
	사용 기기 유형	PC / 스마트폰 / 태블릿
	키보드 유형	2벌식 쿼티 / 천지인 등
텍스트의 내용	주제	일상 대화 / 주제 대화
사전 통제	수집 방법	자연 대화 / 계획 대화

〈표 2-3〉 2019년 국립국어원 메신저 대화 말뭉치의 유형 분류 기준(국립국어원, 2019:10)

이 사업에서도 〈표 2-3〉에 제시된 기준과 항목을 반영한 〈표 2-4〉에 따라 온라인 대화를 수집하고, 말뭉치의 메타 정보로 기재하여 자료를 분류하거나 검색하는 기준으로 활용하도록 한다.

기준	항목	분류
참여자 간 상호 작용	대화 참여자의 수	2인 대화 / 다자 대화
	대화 참여자 간 관계	가족 / 학교, 학원 / 직장 / 지역 / 기타 / 낮은 관계
	대화 참여자 간 친밀도	0(낮은 관계) ~ 5(높은 친밀도)
	대화 참여자 간 연락 빈도	거의 매일 연락하는 관계 / 처음 연락하는 관계 등
사용 매체	메신저의 종류	카카오톡 / 온라인 채팅(심심이)
	사용 기기 유형	PC / 스마트폰 / 태블릿
	키보드 유형	2벌식 쿼티 / 천지인 / 나랏글 / 단모음 / 기타
대화 내용	주제	주제 대화 / 기타 일상 대화 / 시사, 트렌드 주제 대화
사전 통제	수집 방법	기존 대화 / 실시간 대화

〈표 2-4〉 온라인 대화 말뭉치의 유형 분류 기준

8) DMC MEDIA(2019:7)에 의하면 둘 이상의 메신저를 사용하는 사람이 82.7%이다. 이들은 대화 상대에 따라 메신저의 종류를 달리하고, 업무를 위한 메신저와 친목을 위한 메신저를 구분해서 사용한다고 한다.

2.2.1. 참여자 간 상호 작용에 따른 유형 구성

온라인 대화 말뭉치의 참여자 간 상호 작용 양상에 따른 유형 구성은 <표 2-5>와 같다. 대화 참여자의 수는 주관 기관의 요구 사항에 따라 전체 대화 중 10%는 다자 대화를 포함하여 구축하도록 했다. 대화 참여자 간 관계와 친밀도, 연락 빈도는 일정한 비율을 설계하여 구축하는 것이 현실적으로 불가능하여 메타 정보로 기재하여 자료를 분류하거나 검색하는 데 활용할 수 있도록 했다.

기준	항목	분류	
참여자 간 상호 작용	대화 참여자의 수	2인 대화	
		다자 대화(3인, 4인, 5인, 6인)	
	대화 참여자 간 관계	가족	부부
			형제 자매
			부모-자녀
			기타(조부모-손주, 그 외 친인척)
		학교/학원	동기/동창/동급생
			선후배
			교강사-제자
			기타(직원-학생 등)
		직장	동기/동료/동업자
			선후배/상사-부하
			기타
		지역	고향 및 이전 거주지 지인
			현 거주지 지인
		기타	연인
			온라인 커뮤니티
			동호회/스터디
			종교 관련
			그 외 사회적 관계
	대화 참여자 간 친밀도	0(낮선 관계) ~ 5(높은 친밀도)	
	대화 참여자 간 연락 빈도	거의 매일	
		주 3회 이상	
		주 1~2회	
		주 1회 미만	
		월 1회 미만	
		처음 연락(낮선 관계)	

<표 2-5> 온라인 대화 말뭉치 참여자 간 상호 작용 양상에 따른 유형 분류

2.2.2. 사용 매체에 따른 유형 구성

한국어 사용자가 온라인 대화에 활용할 수 있는 다양한 서비스가 있다.

구분		카카오톡	페이스북 메신저	인스타그램 다이렉트	네이버밴드 메신저	네이트온	라인
성별	남자	98.8	24.9	14.9	8.1	6.5	6.1
	여자	99.1	22.1	14.8	9.4	5.7	6.0
연령	6-19세	99.0	29.2	18.2	5.2	3.9	5.2
	20대	98.1	41.3	29.5	7.5	8.8	9.8
	30대	99.0	35.3	21.8	9.7	10.5	10.3
	40대	98.9	23.2	13.8	11.7	8.5	7.3
	50대	99.2	11.8	6.9	10.4	3.4	2.8
	60대	99.4	6.5	3.2	8.0	1.7	1.5
	70세 이상	99.8	2.0	0.8	2.3	0.7	1.2

[그림 2-4] 성·연령별 주 이용 인스턴트 메신저 서비스(한국지능정보사회진흥원, 2021:135)

[그림 2-4]에서 알 수 있듯이 20대와 30대의 경우, 다른 연령에 비해 페이스북 메신저와 인스타그램 다이렉트의 사용 비율이 높다. 그리고 앞서 언급하였듯이 업무를 위한 메신저와 개인 친목을 위한 메신저 등으로 온라인 대화 매체를 변별하여 사용하기도 한다. 따라서 매체 유형에 따른 대화 양상 관찰을 위해서는 다양한 매체를 말뭉치 구축 대상으로 삼아야 한다.

그러나 수집 절차와 가공 절차를 간소화하여 사업을 효율적으로 운영하고자 이 사업에서는 성별과 연령에 관계없이 99%에 가까운 이용률을 보이는 카카오톡을 대상으로 대화를 수집하고 말뭉치를 구축한다. 이와 함께 이 사업의 참여 기관이 보유한 상용 서비스 ‘심심이’를 카카오톡 이외의 온라인 대화 수집 창구로 활용한다.⁹⁾ ‘심심이’는 비교적 주제 맥락이 일관되게 유지되는 짧은 길이의 대화를 수집하는 데 활용해서 대화 수를 기준으로 최소 10% 이상을 포함해서 구축한다.

채팅에 주로 사용하는 기기 유형 변인과 키보드 유형 변인은 자판의 크기와 자판의 형태에 따라 달라질 수 있는 언어 형식의 차이를 관찰하기 위하여 설계한 변인이다. [그림 2-5]와 같이 온라인 대화 서비스 사용자가 주로 사용하는 네 가지 자판¹⁰⁾은 자음과 모음을 입력하는 방식에서 차이를 보이며, 이러한 차이에 따라 자모를 연속으로 입력하는 표현의 빈도나 주로 발생하는 오타 등에서 차이가 나타날 가능성이 있다. 채팅에 주로 사용하는 기기 유형과 키보드 유형은 일정한 비율을 설계하여 구축하는 것이 현실적으로 불가능하여 메타 정보로 기재하여 자료를 분류하거나 검색하는 데 활용할

9) ‘심심이’ 서비스는 인공 지능 챗봇과의 1:1 대화를 위한 서비스이다. 이 사업에서는 기존 인공 지능 챗봇과의 대화가 아니라, 대화 자료 제공자 간의 1:1 대화가 가능한 버전을 추가 개발해서 활용했다.

10) 국립국어원(2019:65)의 메신저 대화 말뭉치 참여자가 사용하는 키보드 유형 비율을 보면 2벌식(쿼티) 자판의 사용자가 65.7%, 천지인 키보드 사용자가 23.9%, 단모음 키보드 사용자가 4.5%, 나랏글 키보드 사용자가 2.3%이다. 나머지 키보드 사용자의 비율은 1% 미만으로 나타났다.

수 있도록 했다.



[그림 2-5] 키보드의 유형

기준	항목	분류
사용 매체	메신저의 종류	카카오톡
		온라인 채팅(심심이)
	채팅에 주로 사용하는 기기 유형	PC(데스크탑, 노트북)
		스마트폰
	키보드 유형	태블릿/패드
		2벌식(퀵티)
		천지인
		나랏글
		단모음
		기타

<표 2-6> 온라인 대화 말뭉치 사용 매체에 따른 유형 분류

2.2.3. 대화 내용에 따른 유형 구성

대화의 주제는 대화의 내용에 영향을 미치는 변인이다. 이 사업에서는 층위가 동일하고 다른 주제와 중복되거나 유사한 내용 전개가 되지 않도록 주관 기관과의 협의를 통해 주제 대화 항목에서는 13개의 주제를 선정했다. 그리고 2021년의 사회상을 담은 말뭉치 구축이라는 사업 목표 달성을 위해 시사과 일상 트렌드 주제를 별도로 구축했다. 시사과 일상 트렌드 주제는 참여 기관이 수집하고 있는 뉴스 검색어, 주요 실시간 검색어를 통해 2주 단위로 선별하고, 주관 기관과 협의해서 주제어를 선정했다. 기타 일상 주제는 특정 주제를 정하지 않고 자유롭게 대화한 경우에 해당하는 분류 항목이다.

개별 주제에 대한 성별이나 연령, 직업 등에 따른 관심도에 차이 등이 자연스럽게 반영될 수 있도록 주제별 수집 비율은 별도로 정하지 않고, 대화 참여자가 자유롭게 선택하여 대화를 할 수 있게 했다.

기준	항목	분류
대화 내용	주제 대화	가사 및 가족

		학교 생활
		일과 직업
		기타 사회 생활 및 활동
		연애와 결혼
		반려동물
		미용과 건강
		여행
		식음료
		쇼핑과 상품
		날씨와 계절
		콘텐츠
		공연 및 관람
	시사·트렌드 대화	시사
	기타 일상 대화	일상 트렌드
		기타 일상

〈표 2-7〉 온라인 대화 말뭉치 사용 매체에 따른 유형 분류

2.2.4. 사전 통제에 따른 유형 구성

온라인 대화는 대화 수집 상황의 인지 여부에 따라 대화의 내용과 형태가 달라질 수 있다. 이 사업에서는 대화 수집에 대한 전제 없이 이루어진 과거 대화를 추출하는 방식과 대화 수집을 전제하고 지정된 주제로 현재 시점에 이루어진 대화를 추출하는 두 가지 방식의 대화를 모두 포함하여 말뭉치를 구축했다.

일반 수집 대화 (기존 대화 제공)

```
"P1">아 강 두개살까 퐁</u>
"P1"><anon type="name" n="9"/>랑 나랑</u>
"P1">ㅋㅋㅋㅋ</u>
"P2"><anon type="name" n="1"/> 언니가 사줄게</u>
"P1">모래</u>
"P1">난 내일 결제할거임</u>
"P2">느 생일선물 이런걸로 안되게워?</u>
"P1">ㅋㅋㅋㅋㅋㅋ</u>
"P2">ㅋㅋㅋㅋㅋㅋ</u>
"P1">돼ㅏ</u>
"P1">안사줘도된다~!</u>
"P1">나만 종종 만나주면 된다~</u>
"P1">맞다 <anon type="name" n="9"/></u>
"P2"><anon type="name" n="2"/> 고민하기 싫어 니 생일선물</u>
"P2"><anon type="name" n="6"/>생일선물 고민하는걸로도 벅차</u>
"P1">독감이라 집으로 스스로 격리하려갔다</u>
"P1">ㅋㅋㅋㅋ미친</u>
"P2">ㅋㅋㅋㅋㅋㅋㅋㅋ</u>
"P1"><anon type="name" n="6"/>쓰 언제</u>
"P1">상알인데</u>
"P1">ㅋㅋㅋㅋ</u>
```

실시간 수집 (수집 봇)

```
"P2">매형은 아침부터 운동갔어</u>
"P1">그렇구나</u>
"P1">부지런하네 대단해</u>
"P2">3시간동안 운동함</u>태능선수촌인줄</u>
"P2">식스팩을 만든다고하는데 불가능각</u>
"P1">진형 때문인가</u>
"P2">나도 먹는걸줄이고있어</u>
"P1">그래도</u>
"P2">근데 움직임이 없어서 아무효과가없는것같아</u>
"P1">얼마전에 난 체력 증정했는데</u>
"P2">몸건강을생각해서 우리 부지런히 움직이자</u>
"P1">완전 저질 나왔어</u>
"P2">저질 나올것같아</u>
"P2">ㅋㅋ</u>
"P1">축정해주는 사람이 비웃었어</u>
"P1">줌</u>
"P2">난 쟁피해서축정못할수있어야</u>
"P1">속상했어</u>
"P2">우리 같이 운동하고 서로 응원하자</u>
"P2">토닥토닥 괜찮아</u>
"P2">건강은 우리맘속에 있는거야</u>
"P1">누나는 매형 갈때 같이가서 운동하면 좋겠다</u>
```

[그림 2-6] 2019년 일반 수집 대화와 수집 봇(bot) 수집 대화 비교 예시

기존 대화를 추출하는 방식은 인위적이지 않고 자연스러운 대화를 수집할 수 있다는

장점이 있는 반면, 친밀한 관계의 대화에서 두드러지는 맥락의 생략이 빈번하게 나타나고, 개인정보와 사적이고 민감한 대화 내용을 포함할 가능성이 크기 때문에 정제 작업의 부담이 증가한다.

대화 수집 상황을 인지하고 이루어진 대화를 추출하는 방식은 인위적인 대화가 될 가능성이 있는 반면, 상대적으로 대화의 내용과 형식이 정제되고 개인정보나 민감한 내용을 포함할 가능성이 적다. 그리고 지정 주제 대화 진행이 가능하다. 그렇기 때문에 인공지능 학습을 위한 말뭉치 구축이라는 관점에서는 효용성이 크다.

이 사업에서는 두 가지 방식의 대화를 모두 포함한 말뭉치를 구축하였으며 기존 대화와 실시간 수집 대화의 구축 비율은 주관 기관과의 협의를 통해 6:4로 조정했다.

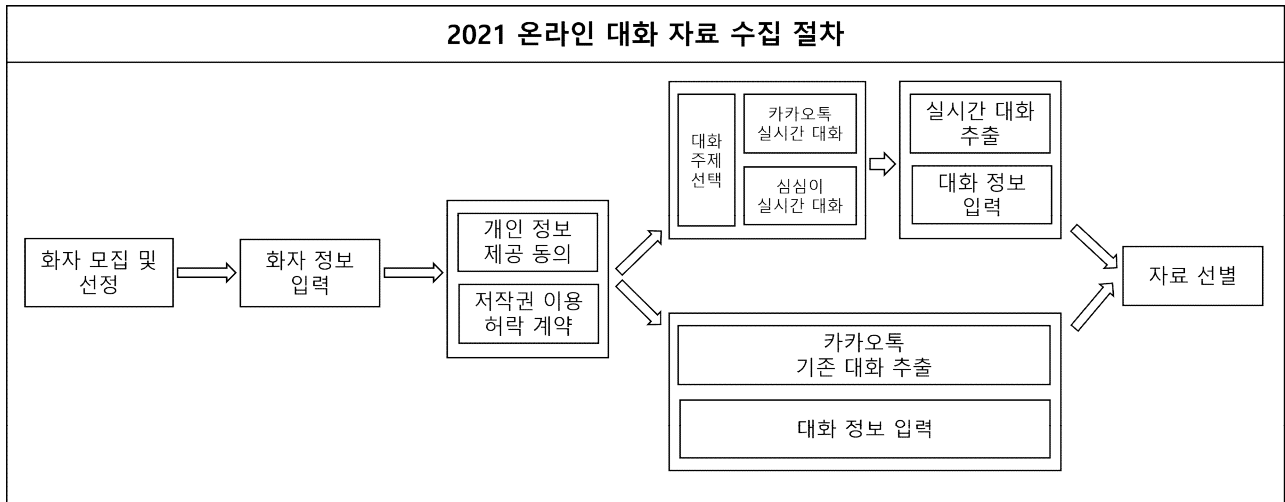
기준	항목	분류
사전 통제	수집 방식	기존 대화 수집(대화 수집 상황에 대한 인지 없음)
		실시간 대화 수집(대화 수집 상황을 인지함)

〈표 2-8〉 온라인 대화 말뭉치 사용 매체에 따른 유형 분류

3. 수집 단계

3.1. 온라인 대화 자료 수집 절차

온라인 대화 자료 수집 진행 절차는 [그림 2-7]과 같다.



[그림 2-7] 2021 온라인 대화 자료 수집 절차

먼저 이 사업의 취지와 목적, 구체적인 내용을 안내하고 홍보하여 온라인 대화를 제공할 화자를 모집하고 대상자를 선정했다. 선정된 대상자는 참여자 등록을 통해 화자 정보 입력, 개인정보 제공에 대한 동의, 말뭉치 구축 대상 자료에 대한 저작권 이용 허락 계약을 체결했다. 특히 저작권 이용 허락 계약은 대화 참여자 전원이 체결하는 것을 원칙으로 하여 대화 참여자 중 저작권 이용 허락 계약을 체결하지 않았을 경우에는 대화 참여와 말뭉치 구축 대상에서 제외했다.

저작권 이용 허락 계약 체결 이후에 대화 참여자가 대화 가능한 시간을 이용하여 카카오톡과 심심이 채팅 기능을 통한 실시간 대화 수집이 이루어졌다. 기존 대화 수집도 별도로 진행되었다. 대화 제공과 함께 대상 자료에 대한 추출과 저장, 상대방과의 관계, 친밀도, 연락 빈도 등의 대화 정보 입력을 진행했다. 이후에 자료를 분류하고 선별했다.

3.1.1. 화자 정보 수집 및 개인정보 이용 동의, 저작권 이용 허락 계약 절차

온라인 대화 말뭉치의 메타 정보 중 화자 정보 수집과 개인정보 이용 동의, 저작권 이용 허락 계약을 체결하기 위해 모바일 웹(mobile web) 형태의 참여자 등록 사이트를 개설했다.

The image shows two screenshots of a web form for participant registration. The left screenshot is titled '언어 사용 특성 연구를 위한 정보' (Information for language use characteristics research) and includes fields for '성년월일 6자리를 입력해주세요.' (Enter 6-digit birth date), '성별을 선택해주세요.' (Select gender), '타어나신 출생지를 선택해주세요.' (Select birthplace), '주로 성장하셨던 지역을 선택해주세요.' (Select region where you mostly grew up), '현재 거주하는 지역을 선택해주세요.' (Select current residence), and '직업을 선택해주세요.' (Select occupation). The right screenshot is titled '이벤트 진행 안내 및 경품 배움을 위한 정보' (Information for event progress and prize learning) and includes fields for '이메일' (Email), '회사' (Company), and '참가하시는 경품 배움의 종류를 선택해주세요.' (Select the type of prize learning you will participate in).

[그림 2-8] 참여자 등록 화자 정보 입력 과정

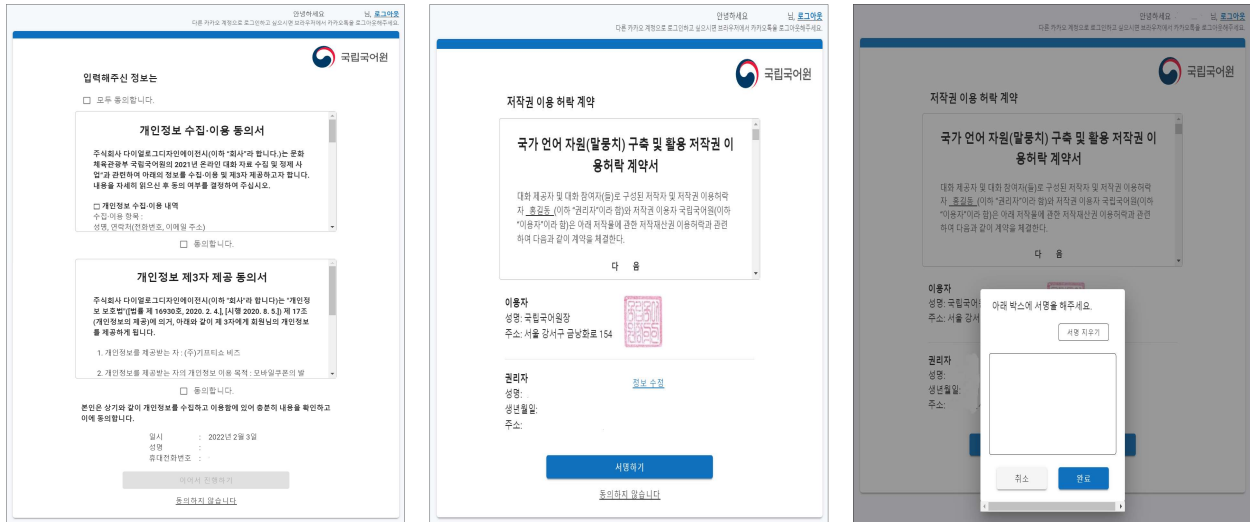
참여자 등록은 사업 기간 중 최초 1회만 진행하도록 했고, 개인의 카카오톡 ID 하나로만 등록이 가능하도록 하여 비정상적인 중복 등록을 방지했다. 참여자 등록 사이트에서 [그림 2-8]의 과정을 통해 대화 제공자가 입력한 화자 정보는 <표 2-9>와 같다.

구분	세부 항목	예시
화자 정보	대화 참여자 성별	남성/여성
	대화 참여자 연령	35세/40세
	대화 참여자 직업	경영, 관리직/전문가 및 관련 종사자
	대화 참여자 출생지	서울/경기/인천/대전
	대화 참여자 주 성장지	
	대화 참여자 현 거주지	
	대화 참여자 사용 기기	스마트폰/PC/태블릿
	키보드(자판)	2벌식(쿼티)/천지인/나랏글

<표 2-9> 참여자 등록 사이트의 화자 정보 수집 항목

화자 정보 입력 이후 개인정보 수집·이용 동의와 저작권 이용 허락 계약의 진행 과정은 [그림 2-9]와 같다.

개인정보 수집·이용 동의는 수집 항목과 수집·이용의 목적, 보유 기간, 대화에 포함된 민감 정보의 처리 방침을 고지하고 동의를 얻었다. 만약 대화 제공자 스스로 동의하지 않을 경우 대화 제공에 참여할 수 없음을 고지했다. 개인정보 수집·이용 동의 주요 내용과 그 실체는 [그림 2-10]과 같다.



[그림 2-9] 참여자 등록 개인정보 수집·이용 동의 및 저작권 이용 허락 계약 과정

개인정보 수집·이용 동의서

주식회사 다이얼로그디자인에이전시(이하 "회사"라 합니다)는 문화체육관광부 국립국어원 2021년 온라인 대화 자료 수집 및 집계 사업과 관련하여 아래의 정보를 수집·이용 및 제3자 제공하고자 합니다. 내용을 자세히 읽으신 후 동의 여부를 결정하여 주십시오.

수집 이용 항목	수집 이용 목적	보유 기간
성명, 연락처(전화번호, 이메일 주소)	자료 제공(오버일 쿠폰) 대가 지급 및 자료 제공자 관리, 저작권 계약 처리	수집된 개인정보는 원칙적으로 개인정보의 수집 및 이용목적이 달성되면 지체 없이 파기 됩니다. 다만, 관련법령의 규정에 의하여 개인정보를 보유할 필요가 있는 경우에는 해당 법령에서 정한 바에 의하여 개인정보를 보유할 수 있습니다.
온라인 대화 자료 제공자의 성별, 나이, 출생지, 거주 지역, 직업, 대화 상대방과의 관계, 대화 입력 방식(키보드 타입)	온라인 대화 내용 음성변환, 자동번역, 음성인식, 자연어 인식, 정보 검색 서비스 등 인공지능 개발 연구용 말뭉치(DB) 구축 및 검증 배포	

□ **대화 텍스트 내의 고유식별정보 및 민감정보 처리**
제공하는 대화 텍스트 내에 포함되는 고유식별정보와 민감정보(사상·신명, 노동조합·당派的 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보 등), 금융정보, 사생활을 현저히 침해할 우려가 있는 정보 등 제거되어야 할 내용은 익명 처리를 원칙으로 하며, 고유식별정보와 민감정보가 포함된 경우, 대화 제공자 스스로 제공을 거부하거나 스스로 삭제 및 삭제 의사를 표시한 후 회사에 제공할 수 있습니다. 회사는 대화 내용과 대화 제공자의 복주의로 인하여 위와 같은 정보가 포함됨으로써 발생하는 문제에 대하여 책임을 부담하지 않습니다.

□ **개인정보 수집·이용 동의의 권리 및 불이익**
귀하께서는 개인정보 수집·이용에 대한 동의를 거부하실 수 있으나, 동의를 거부하실 경우 "2021년 온라인 대화 자료 수집 및 집계"를 위한 대화 내용 수집에 참여하실 수 없습니다.

위와 같이 개인정보를 제공하는 데 동의하십니까?
☐ 동의함 ☐ 미동의함

[그림 2-10] 개인정보 수집·이용 동의서 및 주요 내용

저작권 이용 허락 계약은 제공 동의 대상이 되는 자료의 범위, 이용 허락 기간, 이용 허락 권한의 내용을 포함하여 계약을 체결하고 권리자의 이름과 생년월일, 동 단위 주소까지 기재한 후 자필로 서명을 하도록 했다. 저작권 이용 허락 계약의 주요 내용과 그 실체는 [그림 2-11]과 같다.

[illegible]

- 국립국어원의 온라인 대화 말뭉치 구축 사업 기간(2021년 4월 13일부터 2021년 11월 13일까지) 동안 제공하는 모든 온라인 대화를 대상으로 함.
- 2042년 12월 31일까지 5년 단위로 이용 허락 자동 갱신
- 저작물의 보존, 복제와 변형, 연구 및 기술 개발용으로 학계, 연구 기관, 산업체 등에 제공·배포하는 것을 포함함.
- 학계, 연구 기관, 산업체가 국어 연구와 언어 정보 처리 분야 응용을 위해 저작물 및 복제, 변형물을 분석 및 처리하여 사용하는 것을 포함함.

[그림 2-11] 저작권 이용 허락 계약서 및 주요 내용

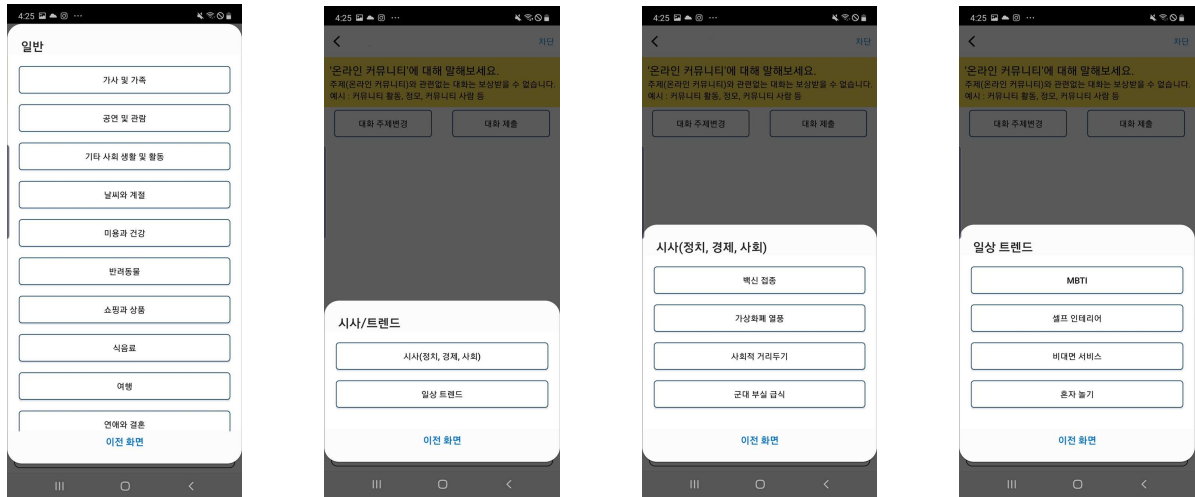
3.1.2. 온라인 대화 주제 선택 절차

실시간 대화 수집의 경우, 지정된 주제를 <표 2-10>과 같이 예시 키워드와 함께 제시한 후 대화 시작 전에 대화 참여자가 스스로 주제를 선택하게 했다.

주제	주제 키워드 예시
가사 및 가족	집안일, 육아, 가정 생활 및 활동, 가족, 룸메이트 등
학교 생활	수업, 시험, 과제, 수강 신청, 동아리, 학교 사람, 학업 스트레스 등
일과 직업	구직, 취업, 이직, 급여, 스펙, 인턴 활동, 아르바이트, 업무, 직장 동료 등
기타 사회 생활 및 활동	동호회 활동, 봉사 활동, 종교 생활, 군대 생활, 온라인 커뮤니티 활동 등
연애와 결혼	이상형, 연애편, 결혼관, 연애 근황, 미팅, 소개팅, 데이트 등
반려동물	반려동물 이름, 간식, 용품, 사료, 질병(알레르기) 등
미용과 건강	다이어트, 운동, 건강 관리, 패션, 성형, 질병 등
여행	여행지, 여행 계획, 기념품, 관광 명소, 여행 경험 등
식음료	식사 메뉴, 배달 음식, 계절 음식, 조리법, 맛집, 유명 카페 등
쇼핑과 상품	최근 구매 물건, 구매 희망 물건, 선물, 중고거래, 할인 소식 등
날씨와 계절	미세먼지, 일기예보, 월간/주간/일간 날씨, 계절과 옷차림, 기상 이변 등

콘텐츠	TV 프로그램, 넷플릭스, 영화, 음원, 게임, 도서, 연예인, 가수, 유튜버 등
공연 및 관람	전시, 관람, 스포츠 리그 및 대회, 스포츠 경기 관람, 티켓 예약 등
시사	추석 특별 방역 대책, 모더나 600만 회 분량 순차 공급 등 ¹¹⁾
일상 트렌드	비건 디저트, 달고나 세트, 10월 단풍 예상 시기, 돌봄 음식 등 ¹²⁾

〈표 2-10〉 온라인 대화 말뭉치 주제 선택 및 예시 키워드 항목



[그림 2-12] 기타 온라인 채팅(심심이) 주제 선택 화면

카카오톡 실시간 대화는 〈표 2-10〉의 주제 항목을 구글 설문지를 이용하여 대화 참여자에게 제공하여 선택 주제를 기록했고, 기타 온라인 채팅(심심이) 실시간 대화는 [그림 2-12]와 같이 주제 선택 기능을 개발하여 대화 참여자가 선택하도록 하고, 선택한 주제는 대화 내용과 함께 서버에 저장했다.

2주 단위로 선정한 시사, 일상 트렌드 주제 목록은 〈표 2-11〉, 〈표 2-12〉, 〈표 2-13〉과 같다.

항목	6월 21일~7월 4일	7월 5일~7월 18일	7월 19일~7월 30일
시사	<ul style="list-style-type: none"> 백신 접종 가상 화폐 열풍 사회적 거리 두기 군대 부실 급식 	<ul style="list-style-type: none"> 5차 재난지원금 도쿄 올림픽 광복절 대체 공휴일 기준 금리 인상 코스피 3,300 돌파 	<ul style="list-style-type: none"> 거리 두기 4단계 연장 예방 접종 사전 예약 청해 부대 전 국민 지원금 3기 신도시 사전 청약 2022년 최저 임금 실거주 2년 철회 인앱 결제 강제 금지
일상 트렌드	<ul style="list-style-type: none"> MBTI 셀프 인테리어 	<ul style="list-style-type: none"> 음성 SNS 메타버스(가상 세계) 	<ul style="list-style-type: none"> 열대야 숙면 방법 셀프 인테리어

11) 2021년 9월 6일부터 19일까지 시사 주제 중 사회 분야에 해당하는 선택 주제 항목 일부

12) 2021년 10월 18일부터 31일까지 선택 주제 항목

항목	6월 21일~7월 4일	7월 5일~7월 18일	7월 19일~7월 30일
	<ul style="list-style-type: none"> 비대면 서비스 혼자 놀기 	<ul style="list-style-type: none"> 구독 서비스 비대면 모임 싸이월드 부활 	<ul style="list-style-type: none"> 셀프 텃밭 구독 서비스 바디프로필 열풍 메타버스(가상 세계)

<표 2-11> 6월~7월 시사/일상 트렌드 주제 목록

항목	8월 2일~8월 15일	8월 16일~8월 29일	9월 6일~9월 19일
시사	사회	<ul style="list-style-type: none"> 백신 사전 예약 10부제 시행 	<ul style="list-style-type: none"> 추석 특별 방역 대책 모더나 600만 회 분량 순차 공급 ‘위드 코로나’ 단계적 전환 국세청 근로장려금 신청 접수
	경제	<ul style="list-style-type: none"> 국민 지원금 스타벅스/이케아 제외 올해 사전 청약 물량 3만 2천 호 확대 자영업 고용 한파 	<ul style="list-style-type: none"> 3기 신도시 1차 사전 청약 4,333 가구 당첨
	정치	<ul style="list-style-type: none"> 언론중재법 처리 보류 	<ul style="list-style-type: none"> 2026년 병장 월급 100만원
	세계		<ul style="list-style-type: none"> 미국 정보동맹 ‘파이프라인즈’ 한국 포함 추진 제 78회 베니스 국제 영화제 개막
	IT	<ul style="list-style-type: none"> 페북/구글/유튜브 청소년 보호 정책 발표 갤럭시 폴더블폰 언팩 행사 	<ul style="list-style-type: none"> 한국 구글 앱마켓 갑질 방지법 통과 갤럭시 폴드3/폴립3 흥행 예고 개인정보위 ‘데이터 보호’ 예산 증액
일상 트렌드	<ul style="list-style-type: none"> 개발자 교육 과정 한국 양궁 공정 경쟁 시스템 위터파크 거리 두기 홈캉스 아이디어 캠핑카 공유 서비스 라이브커머스 시대 	<ul style="list-style-type: none"> AI 학과 대세 메타버스(가상세계) 서비스 싸이월드 부활 채택근무 일자리 컬래버레이션 마케팅(곰표, 진로 두꺼비) 온라인 피트니스 서비스 	<ul style="list-style-type: none"> 지분 조각 투자 다회용 컵(리유저블컵) 할인 등 친환경 마케팅 집놀이족(홈루덴스족)용품

<표 2-12> 8월~9월 시사/일상 트렌드 주제 목록

항목		9월 20일~10월 3일	10월 4일~10월 17일	10월 18일~10월 31일
시사	사회	<ul style="list-style-type: none"> 재난지원금 신청 다자녀 지원 기준 3자녀에서 2자녀로 확대 	<ul style="list-style-type: none"> 부부 동시 육아 휴직 급여 3개월 최대 1,500만 원 지원 내년부터 퀵서비스/대리운전 기사 고용 보험 적용 10월 대체 공휴일 확대 적용 	<ul style="list-style-type: none"> 접종 완료자 백신 패스 단계적 일상 회복
	경제	<ul style="list-style-type: none"> 도시형 생활주택/30평대 주거용 오피스텔 공급 확대 배달앱 4번 주문하면 만원 환급 카카오 지주 회사 케이큐브홀딩스 사회적 기업 전환 	<ul style="list-style-type: none"> 월 10만 원 카드 캐시백 10월 1일부터 신청 소상공인 희망회복자금 10월 말 지급 	<ul style="list-style-type: none"> 수입차 4개 사 리콜 가계부채 대책 10월 발표 반값 복비 10월말 시행
	정치	<ul style="list-style-type: none"> 한국 독자 개발 SLBM 잠수함 발사 시험 성공 	<ul style="list-style-type: none"> 병사 계급체계 3단계로 단순화 플랫폼 국감 	<ul style="list-style-type: none"> 육군 간부/장병 피복류 통일
	세계		<ul style="list-style-type: none"> 미국 매체 ‘오징어 게임’ 집중 조명 	<ul style="list-style-type: none"> IMF 세계 경제 성장률 전망 소폭 하향
	IT	<ul style="list-style-type: none"> 애플 신제품 공개 민간인 4명 3일간 스페스X 우주 관광 방통위 통신 3사에 상생 일자리 창출 노력 당부 	<ul style="list-style-type: none"> LG CNS 마이데이터 사업 ‘라이프 매니징’ 서비스 개발 네이버 국내 창작자 지원금 100억원 출연 	<ul style="list-style-type: none"> 전 세계적 반도체 부족 현상 LG 유플러스/KT, 디즈니 플러스 국내 서비스 한컴 오피스 2022 출시
일상 트렌드		<ul style="list-style-type: none"> 추석 별초 대행 서비스 동네 생활권 서비스 플랫폼 한국 관광공사 ‘머드맥스’ 홍보 영상 ESG(친환경, 사회, 지배구조) 활동 	<ul style="list-style-type: none"> 파이어족 준비 폰 꾸미기 가상 인간 틱톡 댄스 챌린지 가을 글래핑 명소 	<ul style="list-style-type: none"> 비건 디저트 달고나 세트 10월 단풍 예상 시기 돌봄 음식(케어 푸드)

〈표 2-13〉 9월~10월 시사/일상 트렌드 주제 목록

3.1.3. 온라인 대화 진행

주제 선택 이후 실시간 대화를 진행했다. 실시간 대화는 카카오톡 대화와 심심이 서비스를 활용한 1:1 대화 두 가지 매체를 활용하여 수집했다. 대화에 앞서 대화 참여자에게 대화 지침을 사전 안내하여 이를 숙지한 후 대화를 진행하도록 안내했다.

대화 지침은 자연스러운 대화 진행에 관한 것과 인공 지능 학습용 데이터 구축 목적에 관한 내용을 포함하였으며, 이는 〈표 2-14〉와 같다.

-
- 대화 수집을 의식하지 않고 평소와 마찬가지로 자연스럽게 대화를 진행해 주세요.
 - 본 이벤트와 대화 수집에 관련된 내용을 직접적으로 언급하지 마세요.
-

(예: “우리 말풍선 몇 개나 되지?”, “상품권은 언제 받지?” 등)

- 대화방에 초대한 수집봇(뭉치)에게 말을 걸거나 수집봇에 대해 이야기하지 마세요.

(예: “대화방에 세 명인데, 애는 누구야?”, “너는 누구니?” 등)

- 자연스럽게 주고받는 대화가 될 수 있도록 한 사람이 계속 연속해서 말하지 마세요.

- 분량을 늘리기 위한 억지 대화는 하지 마세요.

(같은 글자 반복, 한 글자 또는 한 단어씩 끊어 말하기, 끝말잇기, 과도한 이모티콘 사용 등)

- 개인정보는 언급하지 마세요.

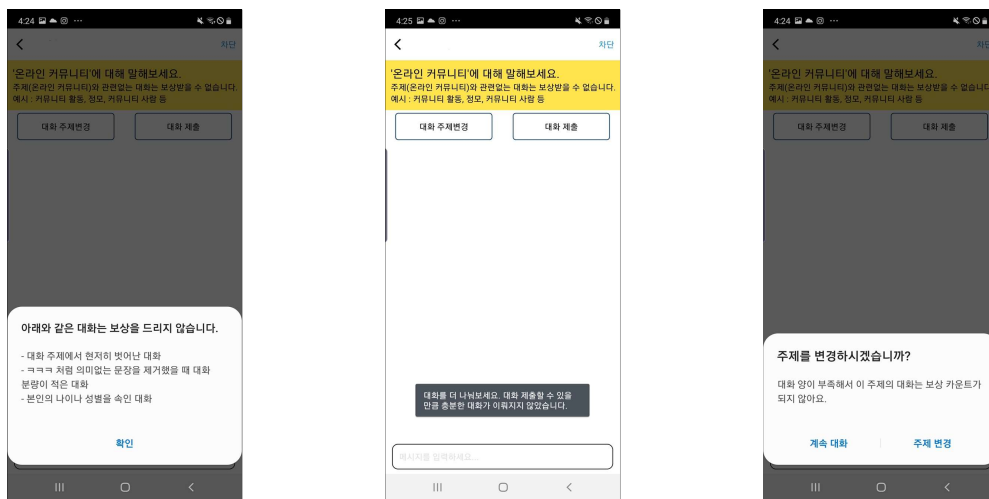
(주민등록번호, 전화번호, 주소, 소속된 직장이나 회사, 계좌번호 등)

- 자연스러운 대화를 위해 사용 가능하지만 분량 산정 시 제외되는 항목 : 이모티콘, 사진, 동영상

- 심한 욕설, 성적인 내용, 혐오나 차별, 비윤리적 내용 등 공개 시 문제가 될 수 있는 대화는 수집하지 않습니다.

〈표 2-14〉 실시간 대화 지침

심심이 대화 기능은 주제에 맞는 화자 전환(turn) 최소 12회 이상¹³⁾의 대화를 수집하는 것에 최적화하여 [그림 2-13]의 형태로 구현했다. 대화의 기본 지침과 함께 대화 시작에 앞서 다시 한번 주제와 분량 조건을 지켜야 함을 안내했고, 최소 12회 이상의 화자 전환이 이루어지지 않은 대화는 제출이 되지 않도록 했다. 그리고 대화 참여자가 원할 경우 대화 주제를 바꿀 수 있도록 하였는데, 현재 대화가 최소 분량 기준을 충족하지 않으면 알림창을 띄워서 최소 분량을 채운 후 주제를 변경하도록 유도했다.



[그림 2-13] 기타 온라인 채팅(심심이) 대화 기능 화면

3.1.4. 대화 정보 입력 절차

대화 참여자 간 대화가 종료되면 <표 2-15>의 대화 정보 항목 중에서 메신저 종류와 대화 주제를 제외한 주제 키워드, 대화 참여자 간 관계, 친밀도, 연락 빈도를 입력했다.

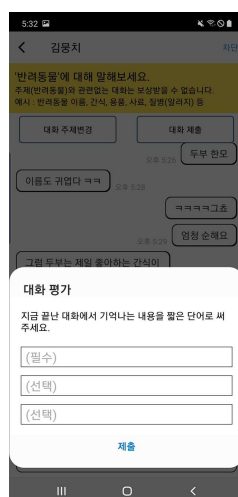
13) 주관 기관의 요구 사항은 최소 8회 이상의 화자 전환이 발생하는 것을 한 개의 대화로 보는 것이었는데, 지침에 맞지 않는 발화의 정제를 고려하여 최소 12회 이상을 수집했다.

카카오톡 실시간 대화는 화자 정보 입력과 마찬가지로 구글 설문지를 활용하여 대화 정보를 입력했고, 카카오톡 기존 대화의 경우도 대화 제출할 때 대화 정보 입력을 하도록 안내했다. 심심이 대화 기능은 화자 정보 입력과 마찬가지로 서비스 화면 안에서 대화 정보를 입력하도록 했다.

구분	세부 항목	예시
대화 정보	메신저 종류	카카오톡/심심이
	대화 주제	개인 및 관계, 주거와 생활
	주제 키워드	육퇴, 아들, 육아
	대화 참여자 간 관계	가족: 부부 / 학교: 선후배
	친밀도	0(낮선 관계)~5(친밀도가 아주 높다)
	연락 빈도	처음/주 1회 미만/(거의) 매일

〈표 2-15〉 대화 정보 수집 항목

주제 키워드는 주제 정보를 보완하여 키워드 간 연관 관계를 통한 대화의 내용이나 흐름 파악, 대화 내용 요약에 필요한 핵심 정보 역할을 할 것을 염두에 두고 수집한 대화 정보이다. 대화가 종료된 이후에 대화 참여자 스스로 대화 내용과 관련하여 기억에 남는 내용을 짧은 낱말이나 표현으로 기재하는 방식으로 수집했다. [그림 2-14]는 심심이 대화 기능을 활용한 주제 키워드 수집 화면이다.



[그림 2-14] 기타 온라인 채팅(심심이) 대화 기능의 주제 키워드 입력 화면

3.2. 온라인 대화 수집 홍보

사업 기간 동안 성별, 연령별 화자 구성 목표를 고려하여 3,000명 이상의 참여자로부터 온라인 대화 자료를 수집하기 위하여 웹과 누리소통망 등 비대면 채널 위주로 대화 참여자 모집 홍보를 진행했다.

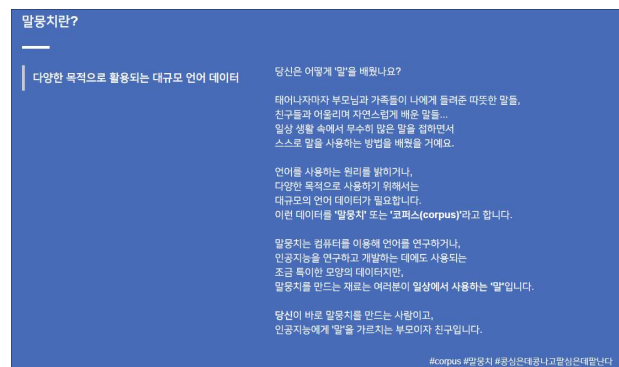
3.2.1. 온라인 대화 수집 홍보 채널 개설

항목	접속 경로	주요 내용
공식 사이트(임시)	https://www.notion.so/HOME-7d07d2e31056440493af426b53c74f3f	<ul style="list-style-type: none"> 사업 소개 참여 기관 소개 대화 제공 방법 안내
공식 사이트	https://www.mediacorpus.net/msg	<ul style="list-style-type: none"> 이벤트 소식, 대화 제공자 모집 공고 게시 자주 묻는 질문
카카오톡 채널	ID : mediaCORPUS	<ul style="list-style-type: none"> 이벤트 소식, 대화 제공자 모집 공고 게시 플러스 친구 대상 이벤트 등 알림 메시지 발송 1:1 문의 창구 운영

〈표 2-16〉 대화 참여자 모집 홍보 채널 운영

〈표 2-16〉과 같이 대화 참여자 모집을 위한 홍보 채널을 개설하고 운영했다. 사업 초기에 빠른 홍보를 위해 노션(Notion.so) 서비스를 이용한 임시 사이트와 이후 개설한 공식 사이트를 통해 사업의 목적과 취지, 참여 기관, 대화 제공 방법 및 지침 등을 공지했다.

특히 홍보 공식 사이트를 개설하면서 말뭉치에 대한 일반인의 이해도를 높이는 것과, 우리가 일상적으로 사용하는 온라인 대화가 귀중한 언어 자원으로 활용된다는 점을 알렸다. 이를 통해 향후 유사 말뭉치 사업 전반에 일반인이 말뭉치에 대한 이해를 기반으로 관심을 가지는 계기를 만들고자 하였으며, 본 사업 참여자에게는 대화 제공에 대한 책임감을 부여하고자 했다.



이제는 우리의 일상이 되어버린 메신저 대화, 우리의 온라인 메신저 언어 생활을 연구하며, 인공 지능 연구와 개발에 필요한 데이터를 위해 당신의 수다가 절실히 필요합니다.

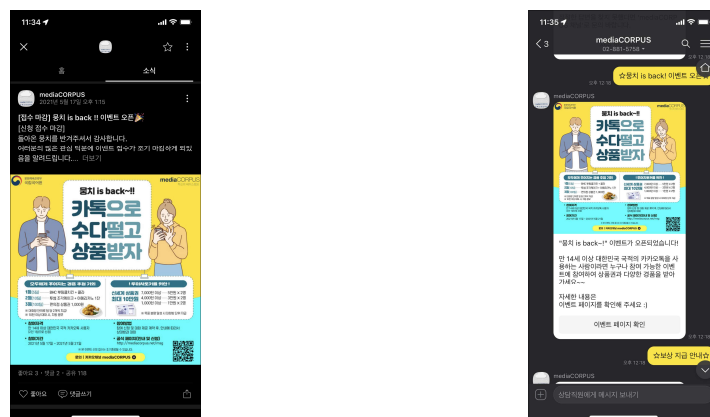
당신은 어떻게 ‘말’을 배웠나요?

태어나자마자 부모님과 가족들이 나에게 들려준 따뜻한 말들, 친구들과 어울리며 자연스럽게 배운 말들...

<p>‘2021년 온라인 대화 자료 수집 및 정제’ 사업은 4차 산업혁명과 인공 지능 시대를 대비해 일상 속 온라인 대화를 언어 연구에 사용하고, 가치 있는 데이터로 만드는 프로젝트입니다.</p> <p>국립국어원과 인공 지능 학습 데이터 전문 기업, 그리고 당신이 함께 합니다.</p>	<p>일상 생활 속에서 무수히 많은 말을 접하면서 스스로 말을 사용하는 방법을 배웠을 거예요.</p> <p>언어를 사용하는 원리를 밝히거나 다양한 목적으로 사용하기 위해서는 대규모의 언어 데이터가 필요합니다. 이런 데이터를 ‘말뭉치’ 또는 ‘코퍼스(corpus)’ 라고 합니다.</p> <p>말뭉치는 컴퓨터를 이용해 언어를 연구하거나, 인공 지능을 연구하고 개발하는 데에도 사용되는 조금 특이한 모양의 데이터지만, 말뭉치를 만드는 재료는 여러분이 일상에서 사용하는 ‘말’입니다.</p> <p>당신이 바로 말뭉치를 만드는 사람이고, 인공 지능에게 ‘말’을 가르치는 부모이자 친구입니다.</p>
---	--

[그림 2-15] 온라인 대화 말뭉치 사업 공식 사이트의 사업 소개 내용

공식 사이트의 대화 제공자 모집 공고 게시물은 카카오톡 채널 서비스를 이용한 공식 홍보 채널과 연계하여 플러스 친구 대상으로 안내 메시지를 발송했다.¹⁴⁾



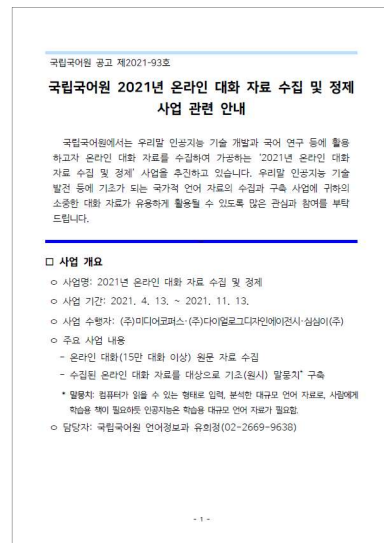
[그림 2-16] 온라인 대화 말뭉치 카카오톡 홍보 채널 및 홍보 메시지 발송 예시

3.2.2. 참여자 모집 홍보 및 모집 결과

개인 간의 지극히 사적인 대화 내용을 수집하는 사업 특성상 사업 목표의 원활한 달성

- 14) 카카오톡 채널은 해당 채널을 친구 추가한 카카오톡 사용자를 대상으로 카카오톡 메시지를 발송하거나, 챗봇을 통한 1:1 상담 기능을 제공한다. 최근에는 맞춤형 광고나 홍보 메시지 발송 등에 널리 활용되고 있다.

을 위해서는 사업에 대한 신뢰 확보가 필요하다. 이에 사업 시작 초기부터 주관 기관의 공식적인 협조와 함께 사업에 대한 홍보를 진행했다.



국립국어원에서는 우리말 인공 지능 기술 개발과 국어 연구 등에 활용하고자 온라인 대화 자료를 수집하여 가공하는 ‘2021년 온라인 대화 자료 수집 및 정제’ 사업을 추진하고 있습니다. 우리말 인공 지능 기술 발전 등에 기초가 되는 국가적 언어 자료의 수집과 구축 사업에 귀하의 소중한 대화 자료가 유용하게 활용될 수 있도록 많은 관심과 참여를 부탁드립니다.

[그림 2-17] 주관 기관 누리집 공지 및 공고

주관 기관의 공식적인 사업 안내와 자체 홍보를 동시에 진행하면서 누리꾼의 자발적인 공유로 홍보가 되도록 홍보 포스터를 [그림 2-18]과 같이 제작했다. 제작한 홍보 포스터를 네이버 카페, 다음 카페, 인터넷 커뮤니티, 누리소통망 등의 다양한 경로에 게시하고 대화 수집을 위해 마련된 이벤트임을 내세워 일반인 참여자가 부담 없이 대화 제공에 참여하도록 유도했다. 이를 통한 참여자 모집 및 홍보를 8회 이상 진행했다.



대화 수집 이벤트 1차



대화 수집 이벤트 2차



대화 수집 이벤트 3차



대화 수집 이벤트 4차



대화 수집 이벤트 5차



대화 수집 이벤트 6차

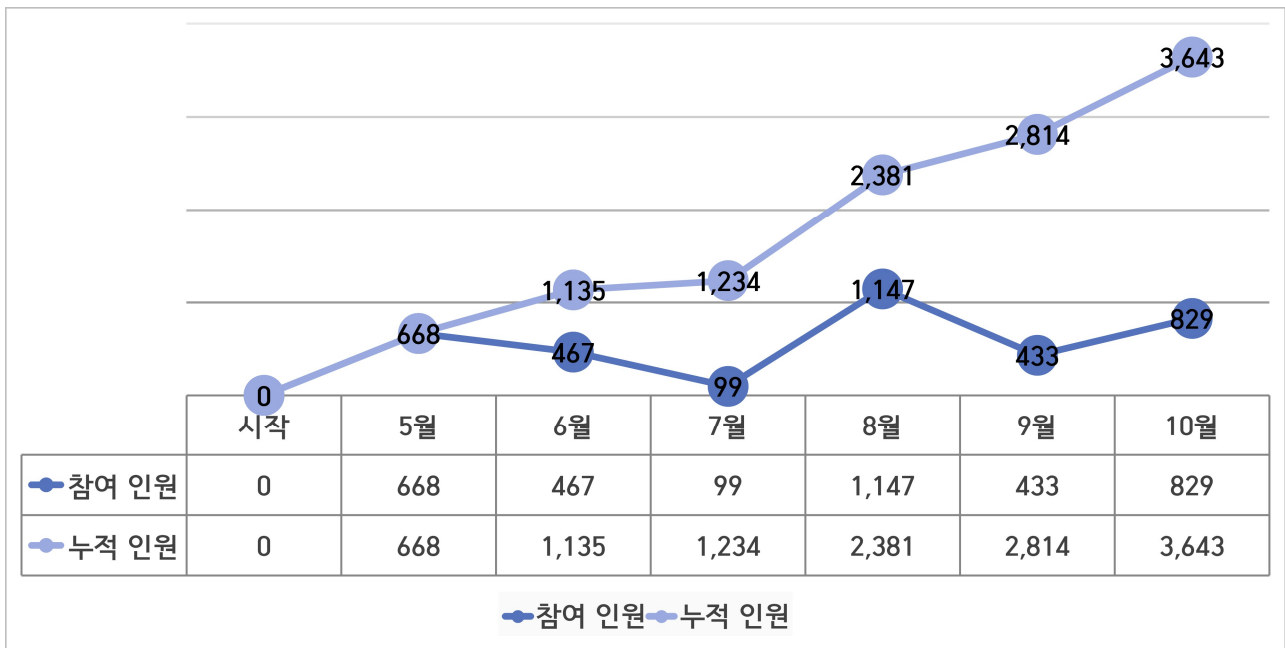


대화 수집 이벤트 7차



대화 수집 이벤트 8차

[그림 2-18] 대화 수집 이벤트 홍보물



[그림 2-19] 온라인 대화 수집 참여 인원 추이

대화 제공자 모집과 대화 수집이 시작된 5월 17일부터 대화 제공자 모집과 대화 수집이 종료된 10월 29일까지 3,643명이 대화 제공에 참여했다.¹⁵⁾ 대화 수집 이후부터 종료까지 월별 대화 제공에 참여한 인원 추이는 [그림2-19]와 같다.

3.3. 자료 선별

대화 제공자의 카카오톡 대화와 기타 온라인 채팅 대화는 본격적인 가공에 앞서 본 사업의 말뭉치 구축 목적과 대화 제공 지침에 부합하는 자료를 선별하는 과정을 거쳤다. 앞서 말했듯이, 본 사업에서는 대화 수집 상황을 전제로 하는 실시간 대화와 이미 진

15) 제공하는 대화에 대한 저작권 이용 허락 계약을 체결한 후 대화를 제공한 전체 누적 인원의 규모이다. 자료를 선별하고 정제하는 과정에서 일부 대화는 제외되기도 하기 때문에 실제 최종 산출물에 반영되는 인원과는 차이가 발생할 수 있다.

행된 대화를 제공하는 기존 대화 수집 두 가지 방식으로 대화를 수집했다.

실시간 대화 수집은 지정된 주제 맥락이 유지되는 대화를 수집할 수 있고, 개인정보나 비윤리적인 대화의 정제 부담이 크지 않다는 장점이 있는 반면, 자연스럽지 않고 인위적인 대화가 수집될 수 있다는 단점이 있다. 수집 전에 자연스러운 대화 진행을 요구하더라도 인위적인 대화가 만들어질 가능성이 있기 때문에 선별 과정을 통해 말뭉치 구축 대상에서 제외하는 과정이 필요하다.

기존 대화 수집은 자연스러운 대화를 수집한다는 장점이 있는 반면, 대화 상황 통제가 어려운 만큼 개인정보가 다량으로 포함된 대화나 비윤리적인 내용의 대화가 수집될 우려가 있다. 실시간 대화 수집과 마찬가지로 선별 과정이 필요한 이유이다.

3.3.1. 실시간 대화의 선별

대화 수집 상황을 전제하고 수집이 이루어진 실시간 대화 수집은 말뭉치 구축 목적과 유의사항을 통해 사업 취지에 맞지 않은 대화는 수집하지 않음을 사전 안내했다. 그러한 안내에도 불구하고 말뭉치 구축 대상에 포함하기 어려운 내용으로 이루어진 대화는 말뭉치 구축 대상에서 제외했다.

- 대화 수집 상황에 대해 언급하고 있는 대화
- 주제와 대화 목적이 불분명한 대화
- 지나치게 짧은 어절의 발화로만 이루어진 대화
- 비윤리적인 내용으로 이루어진 대화
- 대화 내용과 참여자가 입력한 성별, 연령, 직업, 관계 등의 정보가 일치하지 않는 대화

〈표 2-17〉 구축 대상에서 제외되는 실시간 대화의 선별 기준

uttNum	uttTime	speakerNum	utt
1	20210821 16:27	P1	여행을 떠나요
2	20210821 16:27	P2	형...가고싶다 여행
3	20210821 16:27	P2	부산..
4	20210821 16:27	P2	다음은 부산입니다
5	20210821 16:27	P1	부산은 기다린다
6	20210821 16:27	P2	그 후는 간다 해외여행..
7	20210821 16:27	P1	부산은 사람이 많다
8	20210821 16:27	P2	물론 그렇습니다
9	20210821 16:27	P1	요양 여행을 간다
10	20210821 16:27	P2	부산 공역시 인구 많습니다
11	20210821 16:27	P2	요양
12	20210821 16:27	P2	마사지 스파
13	20210821 16:27	P2	벌!
14	20210821 16:27	P1	맛있는 것
15	20210821 16:28	P2	먹을것 맛있는것 있다 많이
16	20210821 16:28	P1	제시 식당도 괜찮습니다
17	20210821 16:28	P2	물고기 괜찮습니까?
18	20210821 16:28	P1	
19	20210821 16:28	P2	먹는다 물회
20	20210821 16:28	P1	물고기 맛이 있다
21	20210821 16:28	P1	부산 앞바다에서 여행지만 여기 물고기는
22	20210821 16:28	P2	여행가서 먹습니다
23	20210821 16:28	P2	ㅋㅋㅋㅋ
24	20210821 16:29	P1	그물에 잡혀 술큰 눈으로 인간을 본다

uttNum	uttTime	speakerNum	utt
1	20210805 09:22	P1	ㅇ
2	20210805 17:59	P1	보라
3	20210805 17:59	P2	뭐
4	20210805 17:59	P1	가능
5	20210805 17:59	P1	보라니
6	20210805 17:59	P2	ㅇㅇ
7	20210805 17:59	P2	소리엄니
8	20210805 17:59	P2	힘들대
9	20210805 17:59	P2	라자
10	20210805 17:59	P2	로하자
11	20210805 18:00	P1	한종종
12	20210805 18:00	P2	뭐
13	20210805 18:00	P1	아쓰
14	20210805 18:01	P1	연결
15	20210805 18:01	P2	왜
16	20210805 18:01	P1	완료
17	20210805 18:01	P1	역시
18	20210805 18:01	P2	음
19	20210805 18:01	P1	키보드가
20	20210805 18:01	P2	좋겠네
21	20210805 18:01	P1	참이다

uttNum	uttTime	speakerNum	utt
1	20210828 15:25	P2	ㅋㅋㅋㅋ저 나중에 돌아올게요 오늘 너무 일상이 많아서 손아래요 ㅋㅋㅋㅋ
2	20210828 15:26	P1	ㅋㅋㅋㅋ오늘 뭐하든 개수 일이 제일이면 좋겠음
3	20210828 15:26	P1	안제 오빠가 저를이 ㅋㅋㅋㅋ들려주세요 저두
4	20210828 15:26	P2	넹ㅋㅋㅋ이만 줄게요
5	20210828 16:57	P2	이제 몇개 채우셨어요 ㅋㅋㅋㅋ이거 하다보니 뭔가 여행가서 사항구경하는 기분이라 재밌네요
6	20210828 17:01	P1	넹두 알람 많이 하셨네요 ㅎㅎ 하디보니 전날 똑같은 얘기하느라 지겨워요
7	20210828 17:01	P1	230개 정도요 ㅋㅋㅋㅋ
8	20210828 17:01	P2	와 진짜 많이 하셨네요 ㅋㅋㅋㅋ
9	20210828 17:01	P2	저도 이제 3월말이 됐는데 이제 170개정도 했어요 ㅋㅋㅋㅋ
10	20210828 17:03	P1	넹두 알람 많이 하셨네요 ㅎㅎ 하디보니 전날 똑같은 얘기하느라 지겨워요
11	20210828 17:05	P2	ㅋㅋㅋㅋㅋㅋ그니까요 이제 누구한테 뭘 애달 땀은지 모르겠어요
12	20210828 17:06	P1	를 위젯다가 저녁쯤에 다시 달리라고요 ㅋㅋㅋ 보니까 150개 ~300개가 가장 힘들것 같아요 그 땀면 보상 엄청해서 기분 내고 할수있을듯
13	20210828 17:06	P1	사항을 담당 느린것도 좀 답답해서 7-10명 동시에 하는 중 ㅋㅋ
14	20210828 17:07	P2	넹데 광범하실거같아요ㅋㅋㅋㅋ
15	20210828 17:07	P2	저도 한 그정도 동시에 하는 거 같네요
16	20210828 17:09	P1	ㅋㅋㅋㅋ맞습니다만 더 보보고 다시 물어보게요
17	20210828 17:09	P1	음시바라 하디바라
18	20210828 17:11	P2	지마? 저도 저녁때쯤 ㅋㅋㅋㅋㅋㅋ다시 물려요 ㅋㅋㅋㅋ
19	20210828 20:16	P1	저 다시와습다
20	20210828 20:16	P1	조바되면 주제 바카주세요 ㅎㅎ
21	20210828 20:17	P2	아직 제출안되네요 ㅋㅋㅋㅋㅋㅋ

[그림 2-20] 구축 대상에서 제외되는 실시간 대화의 실제 사례

[그림 2-20]은 수집 대화 중 구축 대상에서 제외되는 대화의 실제 사례이다.

발화 분량을 기준으로 대화 제공 사례금을 책정하였기 때문에 발화의 양을 늘리기 위

이러한 대화는 자연스러운 대화를 수집한다는 목적에 부합하지 않는 것으로 보고 구축 대상에서 제외했다.

대화 수집 시점 이전에 이미 진행된 대화를 제공하는 기존 대화 수집의 경우도 사업 취지에 맞지 않은 대화는 수집하지 않는다는 것을 사전 안내했다. 그러한 안내에도 불구하고 말뭉치 구축 대상에 포함하기 어려운 내용으로 이루어진 대화는 말뭉치 구축 대상에서 제외했다.

[그림 2-21]은 수집 대화 중 구축 대화에서 제외되는 대화의 실제 사례이다.

[illegible]

[그림 2-21] 구축 대상에서 제외되는 기존 대화의 실제 사례

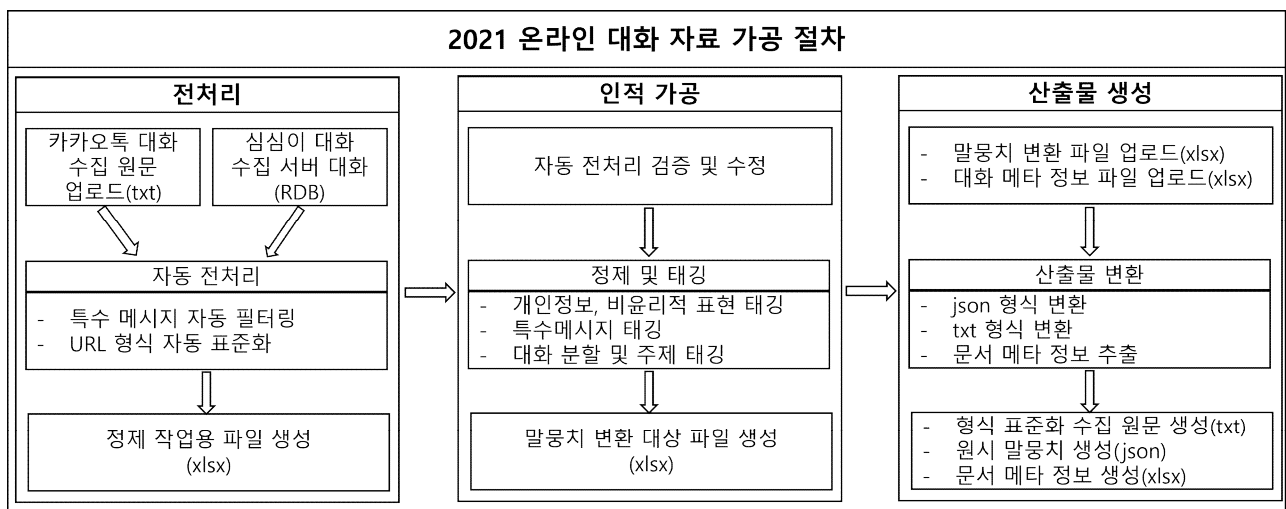
- 36 -

4. 가공 단계

수집 원문 중에서 말뭉치 구축 대상으로 선별된 대화는 정제와 태깅을 거친 후 최종 납품 형태인 원문 txt와 원시 말뭉치 JSON 형식 및 문서 메타 정보 xlsx 형식으로 가공했다.

4.1. 가공 절차

온라인 대화 자료의 가공 절차는 [그림 2-22]와 같다.



[그림 2-22] 2021 온라인 대화 자료 가공 절차

구축 대상으로 선별된 카카오톡 수집 대화는 윈도우(Window) 운영 체제 PC에서 추출해서¹⁷⁾ txt 문서(UTF-8 인코딩)로 저장한 후 말뭉치 가공 도구¹⁸⁾ 서버에 업로드했고, 화자 정보, 대화 정보는 엑셀(EXCEL) 파일로 별도 저장하여 관리했다. 심심이 채팅 기능으로 수집한 대화는 대화 수집 서버에 화자 정보, 대화 정보와 함께 관계형 데이터베이스(RDB : Relational Database) 형태로 저장하여 관리했다.

서버에 저장된 대화 원문은 1단계 자동 전처리를 통해 자동 변환이 가능한 항목을 자동으로 형식 변환한 후 작업자들에게 배포할 정제 작업용 엑셀 파일을 생성했다.

1단계 전처리 이후 생성된 정제 작업용 파일을 작업자에게 배포해서 자동으로 변환된 항목의 적절성을 검증하여 변환 내용의 확정하거나 수정했다.

17) 카카오톡에는 대화 내용 내보내기 기능이 있다. PC와 스마트폰 모두 이 기능을 지원하는데, 운영 체제(Operating System)의 종류에 따라 추출되는 형식이 다르다. 수집 대화 추출 단말기기가 다양해지면 형식을 표준화하는 작업이 진행되어야 한다. 이 사업에서는 형식 표준화의 절차를 간소화하기 위해 Window 운영 체제 기반의 PC로만 대화를 추출했다.

18) 컨소시엄이 이 사업을 위해 자체 개발했다. 사전 정의된 항목을 자동으로 필터링해서 태그를 부착하거나 사전 정의된 구조와 형식에 따라 산출물을 생성하는 도구이다.

이후 작업자의 검토를 통한 정제와 태깅이 이루어졌다. 온라인 대화를 기반으로 원시 말뭉치를 구축하여, 공공 언어 자원으로 일반에게 공개하는 이 사업의 특성을 고려하여 개인정보의 비식별화와 사용자의 실제 발화와 구분되는 시스템 메시지, 비윤리적 내용의 정제로만 정제 범위를 한정했다.¹⁹⁾

정제와 태깅 이후, 말뭉치 저작 도구에 업로드하여 서버에 저장되어 있는 화자 정보, 대화 정보 등의 메타 정보와 연결한 후, 최종 산출물을 생성했다.

4.2. 정제 및 태깅

4.2.1. 작업자 교육 및 지침 공유

작업 지침은 기본적으로 문서 파일 형태로 작업자에게 제공했다. 이와 함께 노션(Notion.so) 서비스를 이용하여 작업 가이드 웹페이지를 개설하고 사업의 목적, 유의사항, 작업 절차와 방법에 대한 교육을 진행했다.

특히 개인의 사적인 대화를 다루게 되는 작업이 갖는 특수성과 개인의 대화 내용이 외부로 유출될 경우 발생할 문제의 심각성을 강조했다. 작업 대상 대화 내용에 대해서는 어떠한 경우라도 외부로 유출되지 않아야 한다는 점을 포함하여 보안 유지에 각별히 유의할 것을 중점적으로 교육했다.

1. 보안 유의 사항

기밀 유지

- ▲ 수집 대화에는 **개인 사항**이 포함되어 있습니다.
- ▲ 업무 외 목적으로 대화 내용을 공유하지 마세요.
- ▲ 본인이 필요한 대화는 적당히 공유 공유 시스템을 이용해 주세요.
- ▼ 2021년 초에 인공지능 챗봇 '이루다' 시간에서는 발화물까지 개인정보로 대화 제공자의 대화를 기록으로 공유했던 것이 밝혀지면서 문제가 되었습니다.
- ▲ 대화 내용은 SNS나 회의로 공유할 수 없습니다.
- ▲ 대화 내용을 외부 사람과 이야기할 수 없습니다.

개인 정보 보호

- ▲ 개인 정보는 연구, 상담과 병행 목적으로만 활용합니다.
- ▲ 대화 제공자 연락처는 방화벽 내로, 상담과 병행 커리큘럼 제공 등만 답변 가능합니다.
- ▲ 대화 제공자 연락처 번호로 직접 연락하지 마세요.
- ▲ 작업 주요 링크가 외부 유출되지 않도록 해 주세요.
- ▲ 개인 정보 파일은 메신저로 주고받지 마세요.
- ▲ 개인 정보 파일에는 암호를 사용해 주세요.
- ▲ 개인 정보 파일은 디스크, 이메일로만 공유해 주세요.

- 업무 외 목적으로 대화 내용 공유 금지
- 대화 내용 SNS 공유 금지
- 대화 내용 외부 발설 금지
- 대화 제공자 개인 전화 번호로 연락 금지
- 개인정보 파일 메신저 전송 금지
- 개인정보 파일 암호 사용

〈표 2-18〉 작업 보안 유의 사항 안내 예시 및 주요 항목

작업 지침 안내 웹페이지는 작업과 관련한 변동 사항이나 작업 팁 공유 등이 실시간으로 이루어지도록 했다. 그리고 작업 지침을 해석하고 적용하는 과정에서 의문이나 이견이 있거나, 기존 지침에 적용되지 않는 새로운 사례가 나타나는 경우에는 구글의 공유

19) 이것은 온라인 대화를 수집하여 말뭉치를 구축하는 다른 정부 사업과 이 사업이 가장 차별화되는 지점이다. 실제적이고 즉각적인 활용성을 전제로 할 경우, 언어 형식의 정제와 표준화가 뒤따라야 하기 때문에 온라인 대화 언어만의 형식적인 특성이 드러나지 않는다. 반면에 이 사업을 통해 구축되는 말뭉치는 맞춤법과 띄어쓰기 등의 언어 형태가 정제되지 않은 원시 말뭉치로서 온라인 대화의 고유한 언어 형태를 그대로 반영하고 있다.

[illegible]

▶ 별첨

7. (세로 바꿈을 눌러서 확인) 국가정보기관 및 관련연 사용자 고유번호 이외에는 숫자는 반드시 7자 이상이어야 합니다.

※참고요:

- 고유 식별 정보(주민등록 번호, 주민등록번호, 운전면허증 번호 등)
- 성명(한글, 한글, 영문, 일본어 로마자)
- 생년월 주수(구 단위 미만을 제외 포함된 주수)
- 이메일, 홈페이지, URL 등
- 생일, 기념일 등 날짜 정보
- 각종 자격증 번호
- 통장 계좌, 카드 번호
- 각종 직업 코드(아이디, 사원 번호, 고객 번호 등)
- 전화 및 팩스 번호
- 의료 보험, 기록 관리 번호 및 기타 수급자 정보
- 특정 비밀번호, 쿠폰 번호, 파일명 등

※성상

7. 국가의 특성을 나타내는 정보는 생활에 따라 국가의 정보를 드러낼 수 있습니다. 뿐만 아니라 가려지

- 성별, 민족, 국적, 교육 수준, 혈액 여부, 결혼 여부, 종교, 언어, 종교, 문화
- 혈형, 신장, 체중, 머리 색깔, 눈동자 색깔, 몸무게 및 몸무게 여부, 체온, 피부
- 신체 내부 특성, 신장 높이, 기둥 높이, 건강 보호를 위해, 소득 분배, 의료 접근 자격
- 학교명, 학과, 학년, 성적, 학력 등
- 경력, 직위, 직장명, 부서명, 직급 등

[illegible][illegible]

이러한 원형 표현 체계 (국문) pdf

이러한 원형 표현 체계 (국문) pdf

제한되어 할 필요 표현

- 반드시 정제해야 할 필요 표현의 유형

제한될 수 있는 필요 표현

- 원칙적으로 정제하는 것을 권고하는 필요 표현의 유형

2. 필요 표현의 문제 발언

필요 표현은 아니지만, 문제가 될 수 있는 발언의 유형

필요 표현과 오해의 소지가 있을 만한 단어를 추려본 것이 있습니다. 표지는 판단하기 어려울 수 있어, 아래 표지를 둘러서 포함하지 어려운 부분을 골라주세요.

적당 상용 글꼴 링크

3. 부적절 발문 정제 방법

부적절 발문 기준을 적용 역발로 수정(‘발화 전체 태깅’에 ‘c’만 입력합니다. (8.6.14))

*이하 기준 방식(사용해도 됨)

부적절 발문 정제 단계는 작업물의 자동 고정 옵션을 활용하면 더 편해요.

자동 고정 옵션 활용 방법

부적절 발언

- 예시지 시각과 볼 부분에 ‘\n’를 넣어서 표시해 줍니다.
- 무엇을 할 것인지 내용이 포함된 문장 전체를 태깅해 주세요.
- 예) \nXXXX 조율이 좀 적어!\n

[illegible]

1. 물음구 위치 규칙에 맞지 않는 발화 식자

원 문맥에서나 대화 수집 이벤트와 관련된 대화

발화식자 위치가 부족한 대화 구간

2. 대화 불일치 점 추출을 위한 도구

1. 말뭉치 구조 규칙에 맞지 않는 발화 식제

원 프로젝트 대화 수집 이벤트와 관련된 대화

- 국립과학원이나 MC, dda, 심상인 등과 대한 이야기
- 수집활동, 심상인 이야기 대한 이야기
- 수집활동에 딸릴 수 있는 발화
- 수집활동으로 직접 발화한 내용(이벤트 결과 간 내 등)

말뭉치로써 가치가 부족한 대화 구간

- 역자로 분류를 놓친 대화 (출처가 아닌 내 등, 잘못된 등)

1 식자 대 발 불일치 현상 식자 대 발은 자연스러운 어휘도 많음 발화로 여러 발화를 함께 식제해야 할 수도 있음이다.

- 39 -

된 내용의 적정성을 검증할 수 있도록 자동 변환 전 발화와 변환 후 전처리 발화를 함께 제시하는 형태로 [그림 2-24]와 같이 정제 작업 파일을 생성했다.

범주	항목	txt 원문 표기 예시
감정 및 상태 표현	이모티콘	• 이모티콘
	메신저 기본 이모티콘	• (하트뽕)(하하)(우와)(심각)(힘듦)
	키보드별 기본 이모지	• 🎧📺❤️👉
시스템 메시지	선물 발송	<ul style="list-style-type: none"> • ***님이 선물과 메시지를 보냈습니다. • ***님의 “카페아메리카노 Tall”선물에 감동했어요.
	무료 통화	<ul style="list-style-type: none"> • 보이스톡 해요/페이스톡 해요 • 보이스톡 취소/페이스톡 취소 • 보이스톡 응답없음/페이스톡 응답없음 • 보이스톡 부재중/페이스톡 부재중 • 보이스톡 0:49/페이스톡 0:12
	송금	• 000 님이 돈을 보냈어요! - 받는 사람 : *** 받을 금액 : 20,000원 입금 기한 : 2021/10/21 23:33까지
	공지 등록	• 특게시판 ‘공지’: 12월 31일 연말 모임 내용 확인
	지도 공유	• 지도: 서울 송파구 송파대로 ***
	연락처 공유	• 연락처: *** 팀장님
	메시지 삭제	• 삭제된 메시지입니다.
	대화방 나감	• ***님이 나갔습니다.
	대화방 들어옴	• ***님이 들어왔습니다.
	대화방 초대	• ***님이 ***님을 초대했습니다.
콘텐츠 공유	사진 공유	• 사진, 사진 n장
	동영상 공유	• 동영상
	음악 공유	• ‘슬픈 운명 (Feat. Lexy, 황성환)-윤희중’ 음악을 공유했습니다.
	파일 공유	• 파일: 04 Beethoven_ Piano Sonata #14 In C.m4a
	음성 메시지 공유	• 음성 메시지
정보 공유	샵 검색	• 샵검색: #무간도
	블로그, 카페 등 게시글 공유	• 다음카페] [어쩌다 발견한 하루] 본인들보다 30센치 작은 여주 놀리는 남주들.jpggif
	뉴스 기사 공유	• 러 군용기 6대 KADIZ 4시간 활개..軍, F-15K 전술조치(종합2보) 【서울=뉴시스】 오종택 기자 = 전투기와 ... 긴급 출격했다...
	광고 및 이벤트 정보 공유	• [Web발신] (광고)[신한카드]신한카드-홈플러스P가 함께하는 모 바일 추가할인 혜택!!
	오픈 채팅 초대	<ul style="list-style-type: none"> • 카카오톡 오픈채팅을 시작해 보세요. • 링크를 선택하면 카카오톡이 실행됩니다.
	배송 안내	• [Web발신] [반품]안녕하세요. *** 고객님의 쿠팡맨 ***입니다. 요 청하신 반품 회수를 금일 진행할 예정입니다.
	인터넷 링크 공유	• https://corpus.korean.go.kr/

<표 2-19> 온라인 대화 자동 전처리 대상 정의

전처리 전 원문				전처리 후			
A	B	C	D	E	F	G	H
1	번호	시간	이름	원문 텍스트		파일	날짜
347	348	20200303 11:46	P1	쓰레기입니다		쓰레기입니다	
347	349	20200303 11:46	P2	난 야 애매해서 일로		난 야 애매해서 일로	
348	349	202003 11:47	P2	쓰레기입니다		쓰레기입니다	
350	349	202003 11:50	P1	이리미론		이리미론	
351	350	202003 11:50	P2	서민		서민	
352	351	202003 00:05	P1	https://www.malform.com/answers/M01716		https://www.malform.com/answers/M01716	
			P2	헌친지식		헌친지식	
			P1	http://www.office-naver.com/form/answer/		http://www.office-naver.com/form/answer/	
			P2	http://mofcncfj-2d4f7iwQdQwM2VC00N0L1T5MmZqTmNnNnNDc2MG48source?code=blue		http://mofcncfj-2d4f7iwQdQwM2VC00N0L1T5MmZqTmNnNnNDc2MG48source?code=blue	
353	352	202003 13:10	P1	아무튼 여객소		아무튼 여객소	
354	353	202003 13:18	P2	아무튼 여객소		아무튼 여객소	
355	354	202003 13:18	P1	잘못된		잘못된	
356	355	202003 13:18	P1	잘못된		잘못된	
357	356	202003 13:18	P1	잘못된		잘못된	
358	357	202003 13:18	P1	잘못된		잘못된	
359	358	202003 13:18	P1	잘못된		잘못된	
360	359	202003 13:21	P2	다들/		다들/	
361	360	202003 13:24	P1	다들/		다들/	
362	361	202003 13:24	P1	다들/		다들/	
363	362	202003 14:12	P1	https://www.instagram.com/9f9k0D3pA07w/		https://www.instagram.com/9f9k0D3pA07w/	
364	363	202003 14:15	P1	gphd=1o64sf73e7e		gphd=1o64sf73e7e	
365	364	202003 14:15	P1	오해 여객소		오해 여객소	
366	365	202003 14:15	P1	오해 여객소		오해 여객소	
367	366	202003 14:15	P1	유류료 따서 적었냐		유류료 따서 적었냐	
368	367	202003 14:15	P1	유류료 따서 적었냐		유류료 따서 적었냐	
369	368	202003 14:15	P1	유류료 따서 적었냐		유류료 따서 적었냐	
370	369	202003 14:16	P1	타다/		타다/	
371	370	202003 14:17	P2	통장잔고		통장잔고	
372	371	202003 14:17	P2	통		통	
			P2	통장잔고		통장잔고	

[그림 2-24] 전처리 전 원문 및 정제 작업 파일 생성 예시

4.2.3. 개인정보 비식별화 및 태깅

가공 대상 자료 중 대화 수집 상황이 전제되지 않은 기존 대화의 경우에는 이름과 연락처, 소속과 같이 개인의 신원이 노출될 수 있는 정보가 종종 나타난다. 실시간 대화의 경우는 개인정보 포함 가능성이 상대적으로 크지 않지만, 상대방의 이름을 부르는 경우를 비롯해서 무의식 중에 개인정보를 발화하는 경우가 있다.

최종 산출물은 화자의 성별, 연령, 직업, 출신 지역과 같은 메타 정보를 포함하고 있기 때문에 공공 언어 자원 구축이라는 사업 특성을 고려했을 때 대화에 포함된 개인정보는 철저한 비식별화가 필요하다.

이 사업의 개인정보 비식별화는 메신저 대화 특성을 고려하여 개인정보 비식별화 지침을 마련한 2019년 메신저 대화 말뭉치 구축 사업의 비식별화 지침을 따라 <표 2-20>과 같이 식별자와 속성자로 나누어 비식별화했다. 그리고 국가 개인정보보호위원회의 개인정보 관련 중차대한 변경 사항이 발생하여 사업 기본 지침과 맞지 않을 경우, 이를 고려하여 지침을 개정하기로 했다.

구분	식별자	속성자
지침	<ul style="list-style-type: none"> 개인 또는 개인과 관련된 사물에 고유하게 부여되는 값 또는 이름으로 반드시 가린다. 	<ul style="list-style-type: none"> 대화 제공자가 해당 내용을 가리지 않은 경우 해당 정보로 인해 누군가를 특정할 수 있는 상황인지 판단하여 가린다.
항목	<ul style="list-style-type: none"> 고유 식별 정보(주민 등록 번호, 운전면허증 번호 등) 성명(한글, 한문, 영문, 필명 포함) 상세 주소(구 단위 미만까지 포함된 주소) 이메일, 홈페이지 URL 등 주소 생일, 기념일 등 날짜 정보 각종 자격증 번호 통장 계좌 번호 	<ul style="list-style-type: none"> 성별, 연령, 국적, 고향, 우편 번호, 병역 여부, 결혼 여부, 종교, 취미, 동호회, 클럽 혈액형, 신장, 체중, 허리둘레, 혈압, 눈동자 색깔, 흡연 및 음주 여부, 채식 여부 세금 납부액, 신용 등급, 기부금, 건강 보험료 납부액, 소득 분위, 의료 급여자 등 학교명, 학과, 학년, 성적, 학력 등

구분	식별자	속성자
	<ul style="list-style-type: none"> • 각종 식별 코드(아이디, 사원 번호, 고객 번호 등) • 전화 및 팩스 번호 • 의료 보험, 기록 관련 번호 및 복지 수급자 번호 • 각종 비밀번호, 쿠폰 번호, 파일명 등 	<ul style="list-style-type: none"> • 경력, 직업, 직종, 직장명, 부서명, 직급

〈표 2-20〉 2021 온라인 대화 비식별화 기본 지침

기본 지침에 따른 항목별 상세 태깅 지침은 〈표 2-21〉과 같다.

범주	항목	처리	원문 표기
이름	실명	비식별화	P1 - \이름1\, P2 - \이름2\ 그 외는 등장 순서에 따라 \이름n\
	실명(변형)	비식별화	
	특수 애칭, 별명, 대화명, 필명	비식별화	
	일반 애칭 별명	비식별화하지 않음	
	공인 실명		
온라인	아이디	비식별화	\계정\
	이메일 주소	비식별화	
각종 번호 및 비밀번호	고유 식별 번호	비식별화	\신원\
	전화번호	비식별화	\전번\
	금융 번호	비식별화	\금융\
	일련번호	비식별화	\번호\
	(구매자) 식별 번호	비식별화	
	사업자 등록 번호	비식별화	
	비밀번호	비식별화	
장소	상세 주소	동 이하 비식별화	\주소\
	아파트 및 거주 건물명	비식별화	
	거주지 역명	비식별화하지 않음	
	방문 장소(비정기적)		
	상호명		
출신 및 소속	출신 및 소속 학교	비식별화	\소속\
	출신 및 소속 직장	비식별화	
	출신 및 소속 부대	비식별화	
기타	위에서 언급하지 않은 항목 ²⁰⁾	비식별화	\기타\

〈표 2-21〉 2021 온라인 대화 개인정보 비식별화 항목별 처리 지침

4.2.4. 특수 메시지 처리 및 태깅

카카오톡과 같은 온라인 채팅 서비스는 이모티콘과 같은 요소를 사용하여 감정을 더욱 효과적으로 드러낼 수 있고, 대화를 주고받으면서 각종 콘텐츠와 정보 공유, 송금과 선

20) 예를 들어 희귀한 질병이나 직업 등과 같이 명시적으로 언급하지 않았지만, 개인의 신분이 드러날 위험이 있는 내용은 기타로 분류했다.

물 보내기 등 다양한 기능을 지원하는 특수 메시지를 사용할 수 있다.

대화 이외의 특수 메시지는 앞서 제시한 <표 2-19>와 같이 패턴을 정의하여 전처리 단계에서 자동 태깅이 되도록 했다. 그러나 실제 대화에서 정의된 패턴 이외에 더욱 다양한 패턴이 나타날 가능성을 고려해서 <표 2-22>와 같이 자동으로 변환되지 않은 요소를 태깅하는 방법을 추가로 제시했다. 태깅은 특수 메시지의 시작과 끝부분에 태깅을 하는 방법과 발화 전체가 단독으로 특수 메시지에 해당할 경우, 정제 작업용 파일의 ‘발화 전체 태깅’ 칼럼에 태깅하는 두 가지 방법을 활용했다.

범주	항목	해당 메시지에 직접 태깅	정제 작업용 파일 칼럼에 태깅
시스템 메시지	선물 발송	• 메시지 시작, 끝부분에 \g\	• 전체 태깅 칼럼에 g 입력
	무료 통화	• 메시지 시작, 끝부분에 \p\	• 전체 태깅 칼럼에 p 입력
	송금	• 메시지 시작, 끝부분에 \b\	• 전체 태깅 칼럼에 b 입력
	공지 등록	• 메시지 시작, 끝부분에 \n\	• 전체 태깅 칼럼에 n 입력
	지도 공유	• 메시지 시작, 끝부분에 \a\	• 전체 태깅 칼럼에 a 입력
	연락처 공유	• 메시지 시작, 끝부분에 \t\	• 전체 태깅 칼럼에 t 입력
콘텐츠 공유	음악 공유	• 메시지 시작, 끝부분에 \m\	• 전체 태깅 칼럼에 m 입력
	파일 공유	• 메시지 시작, 끝부분에 \f\	• 전체 태깅 칼럼에 f 입력
정보 공유	샵 검색	• 메시지 시작, 끝 부분에 \i\	• 전체 태깅 칼럼에 i 입력
	블로그, 카페 등 게시글 공유		
	뉴스 기사 공유		
	광고 및 이벤트 정보 공유		
	오픈 채팅 초대		
	배송 안내		
	인터넷 링크 공유		

<표 2-22> 특수 메시지 태깅 방법

4.2.5. 비윤리적 발화 및 기타 삭제 대상 발화 처리 및 태깅

친밀한 관계에서 이루어지는 사적 대화에는 개인의 언어 습관이 자연스럽게 반영이 된다. 온라인 대화에 포함되는 비속어와 혐오, 차별 발언 등 비윤리적인 발화도 이러한 개인 언어 습관의 자연스러운 반영이다.

하지만 언어 처리를 위한 인공 지능의 연구와 개발까지도 목적에 포함하고 있는 이 사업의 특성을 고려했을 때, 인공 지능의 윤리성 문제를 가볍게 지나칠 수 없으며, 이에 따라 비윤리적인 발화를 정제할 필요가 있다.²¹⁾

21) 다만, 이 사업이 인공 지능의 윤리성을 판단하는 것에 중점을 두는 사업이 아니라는 점은 이 자료를 열람하고 활용하는 모든 사람이 고려해야 한다.

2021년 초 발생한 챗봇 서비스 ‘이루다’의 혐오와 차별 발언 논란을 시작으로 인공 지능의 윤리 문제가 한국 사회에서도 본격적으로 논의되기 시작했다. 유창하게 언어를 구사

실시간 대화는 수집 단계에서 혐오나 차별 등을 조장하는 내용으로 대화하지 말 것을 사전 안내했지만, 발화자 스스로 인지하지 못하는 문제 대화가 만들어질 가능성이 있다. 기존 대화도 혐오나 차별 등의 문제 발화는 수집하지 않는다는 점을 강조했지만, 대화 제공자가 자신의 대화를 면밀히 검증한 후 제출한 것이 아니었기 때문에 정제 단계에서 작업자가 검증하고 정제할 필요가 있었다.

작업자 간 일관된 기준에 따라 비윤리적인 발화를 정제하기 위해 비윤리적 표현을 선별하는 기준을 먼저 수립했다.

작업자가 정제 작업을 진행하면서 평소의 판단 기준보다 높은 기준을 적용하여 비윤리적인 표현과 그 표현이 포함된 맥락을 공유 시트를 통해 게시하고, 게시된 표현에 대해 모든 작업자가 6점 척도로 비윤리성 강도 판정을 진행했다.

작업자 전원의 판정을 거친 후 평균 점수를 넘어서는 표현과 점수 인력이 선정한 표현을 취합하여 주관 기관과 협의를 통해 이 사업을 위한 비윤리적 표현 정제 지침을 수립했다.

수립된 지침을 토대로 대화 내용에 포함된 비윤리적인 표현을 선별한 후, 발화 내부 비윤리적인 표현의 시작과 끝 부분에 ‘\c\’ 부호를 넣어 태깅하거나, 정제 작업용 파일의 전체 태깅 칼럼에 ‘c’를 넣는 방식으로 비윤리적인 표현을 태깅했다.

비윤리적 표현 이외에 말뭉치 구축 목적에 부합하지 않는 발화 중 대화 수집 상황 자체에 대한 언급 등 전체 맥락을 고려해서 삭제해도 문제가 없는 발화는 정제 작업용 파일의 ‘삭제 대상’ 칼럼에 ‘d’ 기호를 넣어 최종 산출물에 반영되지 않도록 했다.

하는 것을 넘어 인간과 소통하고 공감하는 인공 지능에게 윤리적인 판단 기준을 요구하는 것과, 인공 지능에게 언어를 가르치는 재료인 말뭉치 구축에 엄정한 기준을 적용하는 것도 당연하다.

그러나 인공 지능 기술 개발의 역사에 비해 인공 지능이 갖춰야 할 윤리적인 판단 기준에 대한 논의의 역사는 그리 길지 않으며, 사회적인 합의를 위한 다양한 연구가 진행되는 과정에 있다. 아울러 매체 기술이 급속히 발전하면서 언어 데이터의 생산량도 과거에 비해 폭발적으로 증가하고 있으며, 이에 따라 언어가 변화하는 속도도 과거에 비해 빠르다. 언어 표현의 윤리성을 판단하는 기준 또한 언어 변화의 속도에 맞추어 새롭게 정립이 되어야 하는 상황이다.

그렇기 때문에 최종 산출물의 비윤리적 표현 정제 결과에 부족함이 드러날 수 있다. 하지만 이 사업 체계 안에서는 일관된 기준을 갖추고자 했고, 나아가 인공 지능 언어의 윤리 기준을 만들어 나가는데 이 사업이 공헌할 수 있는 부분이 무엇인가를 찾고자 했다.

결국 이 사업 또한 완벽한 기준과 지침에 따라 표현의 비윤리성을 판정하는 사업이 아니라, 인공 지능이 갖춰야 할 윤리적인 기준을 찾아가는 과정 중에 있는 사업 가운데 하나라는 점을 고려해서 이 자료를 열람하고 활용했으면 한다.

[그림 2-25] 비윤리적 표현 정제 기준 수립을 위한 공유 시트 활용 예시

[그림 2-26] 비윤리적 표현 정제 기준 수립을 위한 주관 기관 협의 진행

- 45 -

- 욕설과 비속어는 대화의 상황 맥락에 따라 정제 여부를 판단함.
- 자모로 표현되는 욕설 중 상대방 발화에 대한 단순한 반응이 아니라, 특정 인물을 지칭하여 비난하거나 공격적인 의도가 있는 경우는 정제함.
- 일반적인 차별 표현과 노인, 아동 등을 향한 혐오 표현은 맥락에 관계없이 정제함.
- 본인 또는 타인에게 신체·정신적 장애가 있는 사람을 지칭하여 불쾌감을 유발할 수 있는 표현은 정제함.
- 사회적으로 문제가 있는 것으로 인식되는 특정 사이트에서 기원한 표현은 정제함.
- 인종, 성별, 종교, 연령, 지역, 성적 취향에 대한 명백한 차별이나 혐오 표현은 정제함.
- 명백한 성적 대상화에 해당하는 표현은 정제함.

[그림 2-27] 2021 온라인 대화 비윤리적 표현 정제 관련 지침 주요 내용

4.2.6. 대화 분할 및 주제 태깅

카카오톡으로 대화를 진행한 경우, 대화 참여자가 직접 선택한 주제로 대화를 하면서 대화가 길어지는 경우에 선택한 주제와 관계없이 다른 주제로 대화를 진행하는 경우가 있었다.²²⁾ 이런 경우에 [그림 2-28]과 같이 정제 작업용 작업 파일에 주제를 별도로 태깅하고 태깅된 주제를 단위로 별도의 대화로 분할했다.

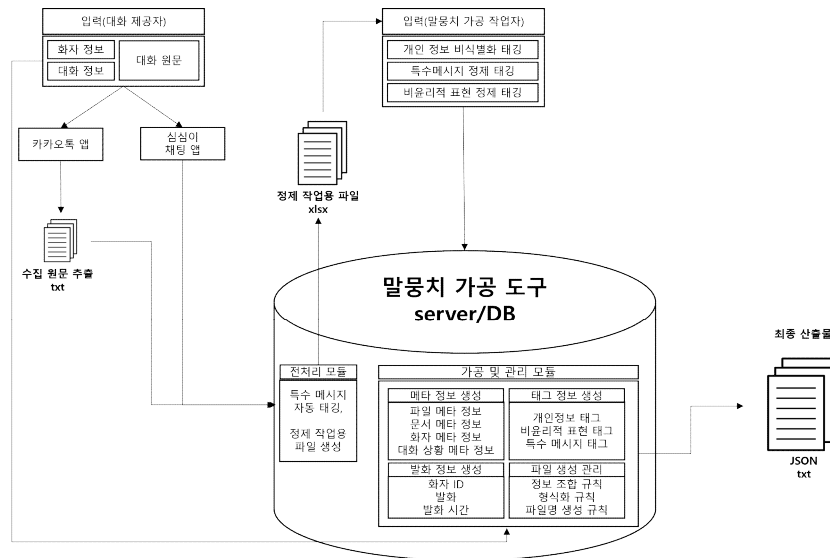
번호	시간	이름	원본 텍스트	작업자 원문 태깅	자동 변경된 발화 전체 태깅, 삭제 대상	대화 분할임(주제)
1068	20210809 16:41	P1	박자담하고 미아가하다가	박자담하고 미아가하다가		말과 직접
1069	20210809 16:41	P1	벌써 5시지롱	벌써 5시지롱		
1070	20210809 16:41	P1	놀라갈라그렸는데 실패했지롱	놀라갈라그렸는데 실패했지롱		
1071	20210809 16:41	P1	이모티콘	이모티콘		
1072	20210809 16:44	P1	캐필라리 뽏는거, 본인도 10년전에 한거래	캐필라리 뽏는거, 본인도 10년전에 한거래		
1073	20210809 18:53	P2	그거는 바늘 잘라서 쓰는거 아니면	그거는 바늘 잘라서 쓰는거 아니면		
1074	20210809 18:53	P2	대체품이 없어	대체품이 없어		
1075	20210809 19:16	P1	있대	있대		
1076	20210809 19:17	P2	헐	헐		
1077	20210809 19:17	P1	캐필라리 파는거	캐필라리 파는거		
1078	20210809 19:17	P1	있대	있대		식용료
1079	20210809 19:19	P2	이씨	이씨		
1080	20210809 19:19	P2	근데 호저	근데 호저		
1081	20210809 19:19	P2	씨받어	씨받어		
1082	20210809 19:20	P1	그거 쓰래 그냥	그거 쓰래 그냥		
1083	20210809 19:20	P2	응 담당하겠지만 머	응 담당하겠지만 머		
1084	20210809 19:20	P2	쓰다보면 적응돼	쓰다보면 적응돼		
1085	20210809 19:41	P1	이제 집앞써	이제 집앞써		
1086	20210809 19:41	P2	고생 만나까	고생 만나까		
1087	20210809 19:42	P1	이모티콘	이모티콘		
1088	20210809 19:44	P2	함뽕이 머코 시피	함뽕이 머코 시피		
1089	20210809 19:44	P2	친함뽕 사오까	친함뽕 사오까		
1090	20210809 19:44	P1	함뽕머거	함뽕머거		
1091	20210809 19:44	P1	거기있잖나	거기있잖나		
1092	20210809 19:44	P1	함뽕스도리인가	함뽕스도리인가		
1093	20210809 19:44	P2	착한 함뽕!	착한 함뽕!		
1094	20210809 19:44	P2	?	?		
1095	20210809 19:48	P1	응	응		
1096	20210809 19:49	P2	시리	시리		
1097	20210809 19:49	P2	친함뽕이 머코시포다	친함뽕이 머코시포다		
1098	20210809 19:51	P1	머경구령	머경구령		
1099	20210809 19:51	P1	머경	머경		
1100	20210809 20:04	P2	그자나	그자나		
1101	20210809 20:04	P2	사러가기	사러가기		
1102	20210809 20:04	P1	이모티콘	이모티콘		
1103	20210809 20:04	P1	머용?	머용?		
1104	20210809 20:05	P2	이모티콘	이모티콘		
1105	20210809 20:05	P2	귀자나아아아	귀자나아아아		

[그림 2-28] 대화 분할 및 주제 태깅 작업 예시

4.3. 산출물 생성

22) 심심이를 활용한 채팅에서는 대화를 진행하면서 대화 주제를 바꾸는 기능이 있었기 때문에 주제 단위로 짧은 대화를 수집하는 게 가능했다. 하지만 카카오톡 대화는 자연스러운 대화 수집을 목적으로 하면서 대화 중간에 주제가 바뀌었다는 것을 의식하면서 주제 표시를 하라는 요구를 대화 참여자에게 직접적으로 하지 않았다.

정제와 태깅이 완료된 대화문은 대화 제공자가 직접 입력한 화자 정보, 대화 정보와 함께 말뭉치 가공 도구에 업로드 이후 규칙에 따라 자동으로 생성된 문서 정보와 결합하여 사전에 정의한 형식과 구조를 따르는 산출물로 만들어졌다. 산출물이 만들어지는 과정은 [그림 2-29]와 같다.



[그림 2-29] 2021 온라인 대화 자료 산출물 생성 과정

4.3.1. 산출물의 형식과 구조

온라인 대화 자료는 말뭉치의 유형 구분, 매체 및 장르 분류, 분석 층위 구분, 구축 연도를 기본으로 하고 여기에 문서에 8자리 일련번호를 붙이는 방식으로 파일 이름을 부여했고, UTF-8 인코딩의 txt 형식 대화 원문과 JSON 형식 원시 말뭉치 두 가지 형태로 생성했다.

온라인 대화 자료의 파일명 부여 방식과 작성 예시는 <표 2-23>과 같다.

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	8자리 일련 번호
M: 온라인 대화 말뭉치	D: 2인 대화 M: 다자 대화	OR: 원문 자료 RW: 원시 말뭉치	21	00000001~ 99999999

- MDOR2100000001.txt : 2인 온라인 대화 원문 자료 첫 번째 파일 txt format
- MDRW2100000001.JSON : 2인 온라인 대화 원시 말뭉치 첫 번째 파일 JSON format
- MMRW2100000001.JSON : 다자 온라인 대화 원시 말뭉치 첫 번째 파일 JSON format




<표 2-23> 2021 온라인 대화 자료 파일명 부여 방식 및 파일명 작성 예시

개인정보 비식별화 대상의 태깅을 반영한 산출물의 표기 형식은 <표 2-24>와 같다.

범주	항목	txt 원문 표기	JSON form	JSON original_form
이름	실명	P1 - \이름1\ P2 - \이름2\ \이름n\	name1	&name1&
	실명(변형)		name2	&name2&
	특수 애칭, 별명, 대화명, 필명		nameN	&nameN&
온라인	아이디	\계정\	account	&account&
	이메일 주소			
각종 번호 및 비밀번호	고유 식별 번호	\신원\	social-security-num	&social-security-num&
	전화번호	\전번\	tel-num	&tel-num&
	금융 번호	\금융\	card-num	&card-num&
	일련번호	\번호\	num	&num&
	(구매자) 식별 번호			
	사업자 등록 번호			
	비밀번호			
장소	상세 주소	\주소\	address	&address&
	아파트 및 거주 건물명			
출신 및 소속	출신 및 소속 학교	\소속\	affiliation	&affiliation&
	출신 및 소속 직장			
	출신 및 소속 부대			
기타	위에서 언급하지 않은 항목	\기타\	others	&others&

〈표 2-24〉 2021 온라인 대화 개인정보 비식별화 항목의 산출물 표기 형식

특수 메시지와 비윤리적 표현 항목의 태깅을 반영한 산출물의 표기 형식은 〈표 2-25〉와 같다.

범주	항목	txt 원문 표기	JSON original_form	JSON form
감정 및 상태 표현	이모티콘	• 이모티콘	• {emoji}	• 공백
	메신저 기본 이모티콘	• (하트뽕)(하하)(우와)(심각) (힘듦)	• {emoji:하트뽕}{emoji:하하}{emoji:우와}{emoji:심각}{emoji:힘듦}	• 공백
	키보드별 기본 이모지	•   	• {emoji:🖱️}{emoji:💻}{emoji:❤️}{emoji:✍️}	• 공백
시스템 메시지	선물 발송	• ***님이 선물과 메시지를 보냈습니다. • ***님의 “카페아메리카노 T all”선물에 감동했어요.	• {system:gift}	• 공백
	무료 통화	• 보이스톡 해요/페이스톡 해요 • 보이스톡 취소/페이스톡 취소	• {system:call}	• 공백

범주	항목	txt 원문 표기	JSON original_form	JSON form
		<ul style="list-style-type: none"> • 보이스톡 응답없음/페이스톡 응답없음 • 보이스톡 부재중/페이스톡 부재중 • 보이스톡 0:49/페이스톡 0:12 		
	송금	<ul style="list-style-type: none"> • 000 님이 돈을 보냈어요! - 받는 사람 : *** 받을 금액 : 20,000원 입금 기한 : 2021/10/21 23:33까지 	<ul style="list-style-type: none"> • {system:money} 	<ul style="list-style-type: none"> • 공백
	공지 등록	<ul style="list-style-type: none"> • 특게시판 ‘공지’: 12월 31일 연말 모임 내용 확인 	<ul style="list-style-type: none"> • {system:notice} 	<ul style="list-style-type: none"> • 공백
	지도 공유	<ul style="list-style-type: none"> • 지도: 서울 송파구 송파대로 *** 	<ul style="list-style-type: none"> • {system:map} 	<ul style="list-style-type: none"> • 공백
	연락처 공유	<ul style="list-style-type: none"> • 연락처: *** 팀장님 	<ul style="list-style-type: none"> • {system:contact} 	<ul style="list-style-type: none"> • 공백
	메시지 삭제	<ul style="list-style-type: none"> • 삭제된 메시지입니다. 	<ul style="list-style-type: none"> • {system:delete} 	<ul style="list-style-type: none"> • 공백
	대화방 나감	<ul style="list-style-type: none"> • ***님이 나갔습니다. 	<ul style="list-style-type: none"> • 원시 말뭉치에서는 삭제함. 	
	대화방 들어옴	<ul style="list-style-type: none"> • ***님이 들어왔습니다. 		
콘텐츠 공유	대화방 초대	<ul style="list-style-type: none"> • ***님이 ***님을 초대했습니다. 		
	사진 공유	<ul style="list-style-type: none"> • 사진, 사진 n장 	<ul style="list-style-type: none"> • {share:photo} 	<ul style="list-style-type: none"> • 공백
	동영상 공유	<ul style="list-style-type: none"> • 동영상 	<ul style="list-style-type: none"> • {share:video} 	<ul style="list-style-type: none"> • 공백
	음악 공유	<ul style="list-style-type: none"> • ‘슬픈 운명 (Feat. Lexy, 황성환)-윤희중’ 음악을 공유했습니다. 	<ul style="list-style-type: none"> • {share:music} 	<ul style="list-style-type: none"> • 공백
	파일 공유	<ul style="list-style-type: none"> • 파일: 04 Beethoven_ Piano Sonata #14 In C.m4a 	<ul style="list-style-type: none"> • {share:file} 	<ul style="list-style-type: none"> • 공백
정보 공유	음성 메시지 공유	<ul style="list-style-type: none"> • 음성 메시지 	<ul style="list-style-type: none"> • {share:voice} 	<ul style="list-style-type: none"> • 공백
	샵 검색	<ul style="list-style-type: none"> • 샵검색: #무간도 	<ul style="list-style-type: none"> • {share:info} 	<ul style="list-style-type: none"> • 공백
	블로그, 카페 등 게시글 공유	<ul style="list-style-type: none"> • 다음카페] [어쩌다 발견한 하루] 본인들보다 30센치 작은 여주 놀리는 남주들.jpggif 		
	뉴스 기사 공유	<ul style="list-style-type: none"> • 러 군용기 6대 KADIZ 4시간 활개..軍, F-15K 전술조치 (종합2보) 【서울=뉴시스】 오종택 기자 = 전투기와 ... 긴급 출격했다... 		
	광고 및 이벤트 정보 공유	<ul style="list-style-type: none"> • [Web발신] (광고)[신한카드] 신한카드-홈플러스P가 함께하는 모바일 추가할인 해 		

범주	항목	txt 원문 표기	JSON original_form	JSON form
		택!!		
	오픈 채팅 초대	<ul style="list-style-type: none"> 카카오톡 오픈채팅을 시작해 보세요. 링크를 선택하면 카카오톡이 실행됩니다. 		
	배송 안내	<ul style="list-style-type: none"> [Web발신] [반품]안녕하세요. *** 고객님의 쿠팡맨 ***입니다. 요청하신 반품 회수를 금일 진행할 예정입니다. 		
	인터넷 링크 공유	<ul style="list-style-type: none"> https*** 	<ul style="list-style-type: none"> {share:url} 	<ul style="list-style-type: none"> 공백
비윤리적 표현	욕설/비속어 등	<ul style="list-style-type: none"> 원문 형태 유지 	<ul style="list-style-type: none"> {censored} 	<ul style="list-style-type: none"> 공백
기타 무의미한 내용	수집 상황에 대한 발화 등	<ul style="list-style-type: none"> 제외 	<ul style="list-style-type: none"> 제외 	<ul style="list-style-type: none"> 제외

<표 2-25> 2021 온라인 대화 특수 메시지 및 비윤리적 표현 등 산출물 표기 형식

원시 말뭉치 JSON 형식의 구조는 <표 2-26>과 같다.

1수준	2 수준	3 수준	4수준	타입	설명	예시
id				str	파일 ID	MDRW2100000001
metadata				obj	파일의 메타 정보	
	title			str	국립국어원 [말뭉치 유형 구분] [파일 ID]	국립국어원 온라인 대화 말뭉치 MDRW2100000001
	creator			str	생성자	국립국어원
	distributor			str	배포자	국립국어원
	year			str	말뭉치 구축 연도(예: 2021)	2021
	category			str	분류	온라인 대화 > 2인 대화
	annotation_level			str	분석 층위	원시
	sampling			str	샘플링 방식	실시간 대화 수집/기존 대화 수집
document				arr (obj)	문서(대화) 정보	
	id			str	문서(대화) ID	MDRW2100000001.1
	metadata			obj	문서의 메타 정보	
		title		str	문서 제목	온라인 대화
		author		str	작성자	개인 대화 참여자
		publisher		str	온라인 대화 매체	카카오톡/온라인 채팅(심심이)

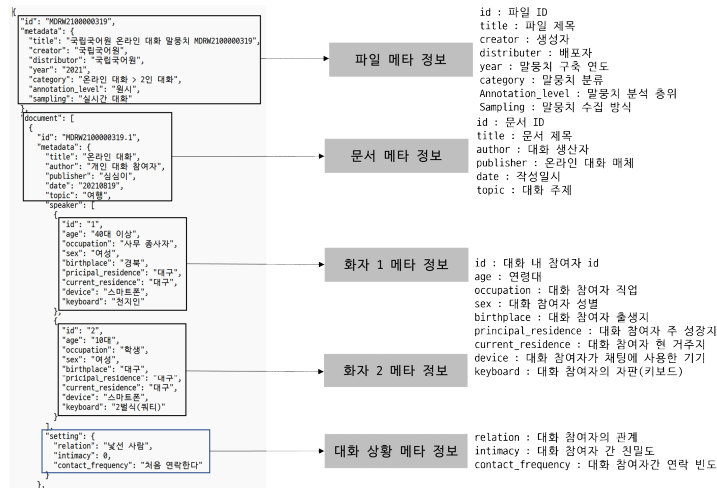
1수준	2 수준	3 수준	4수준	타입	설명	예시
		date		str	작성일시	20200101
		topic		str	대화 주제 분류	일상 생활 : 월급, 카드값 ²³⁾
		speaker		arr (obj)	대화 참여자 정보	
			id	str	대화 참여자 ID	1
			age	str	대화 참여자 연령	20대/30대
			occupation	str	대화 참여자 직업	학생/사무 종사자
			sex	str	대화 참여자 성별	남성/여성
			birthplace	str	대화 참여자 출생지	서울/경기인천/대전
			pricipal_ residence	str	대화 참여자 주 성장지	
			current_ residence	str	대화 참여자 현 거주지	
			device	str	대화 참여자 사용 기기	스마트폰/PC/태블릿
			keyboard	str	키보드(자판)	2벌식
		setting		obj	대화 정보	
			relation	str	대화 참여자 간 관계	학교/학원 : 선후배
			intimacy	num	대화 참여자 간 친밀도	0(낮선 관계)~5(높은 친밀도)
			contact_ frequency	str	연락 빈도	거의 매일 / 주 3회 이상 / ... / 처음
	utterance			arr (obj)	발화 정보	
		id		str	발화 ID	MDRW2100000001.1.1
		form		str	발화 내용(원문에서 공백 및 비식별 화 기호 등 제거)	
		original_ form		str	발화 내용(원문)	
		speaker_id		str	대화 참여자 ID	1
		time		str	발화 시간	202010101 22:00

〈표 2-26〉 2021 온라인 대화 말뭉치 JSON 구조

4.3.2. 산출물 실제

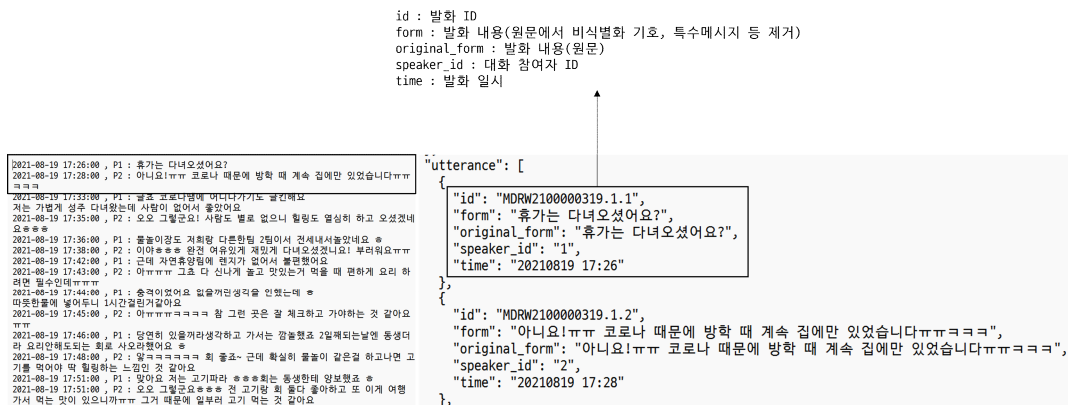
문서 정보, 화자 정보, 대화 정보는 화자가 직접 입력한 정보와 문서 정보 생성 규칙에 따라 말뭉치 가공 도구에서 자동으로 생성했다. 2021 온라인 대화 원시 말뭉치 JSON 형식의 문서 메타 정보 작성 예시는 [그림 2-30]과 같다.

23) 대화 종료 이후에 대화 참여자가 입력한 대화 키워드를 주제 뒤에 기재했다. 필수로 입력해야 하는 정보는 아니다.



[그림 2-30] 온라인 대화 말뭉치 메타 정보 JSON 형식

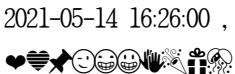
2021 온라인 대화 자료의 대화 원문 txt 형식과 원시 말뭉치 JSON 형식의 대화 내용 작성 예시는 [그림 2-31]과 같다.



[그림 2-31] 대화 원문 txt 형식과 대화 내용 JSON 형식

개인정보, 특수 메시지, 비윤리적 표현의 txt 원문과 JSON 표기 형식 예시는 <표 2-27>과 같다.

항목	txt 원문 표기 예시	JSON 표기 예시
이름	2021-05-18 16:12:00 , P2 : \이름\님 완전 대인배	<pre>{ "id": "MDRW2100000001.1.18", "form": "name1님 완전 대인배", "original_form": "&name1&님 완전 대인배", "speaker_id": "2", "time": "20210518 16:12" }</pre>

항목	txt 원문 표기 예시	JSON 표기 예시
아이디 이메일 주소	2021-05-18 20:29:00 , P2 : 아 맞네\계정\	{ "id": "MDRW2100001307.32.15", "form": "아 맞네\account", "original_form": "아 맞네&\account&", "speaker_id": "2", "time": "20210518 20:29" },
각종 번호 및 비밀번호	2021-08-05 17:02:00 , P2 : 나한 테 \전번\냐고 ㅋㅋㅋㅋ	{ "id": "MDRW2100001388.29.11", "form": "나한테 tel-num냐고 ㅋㅋㅋㅋ", "original_form": "나한테 &tel-num&냐고 ㅋㅋㅋㅋ", "speaker_id": "2", "time": "20210805 17:02" },
장소	2021-05-18 10:38:00 , P1 : 주소 지가 어제 날짜 기준으로 \주소 \동으로 바뀌었는데요	{ "id": "MDRW2100000235.1.14", "form": "주소지가 어제 날짜 기준으로 address동으로 바뀌 었는데요", "original_form": "주소지가 어제 날짜 기준으로 &address& 동으로 바뀌었는데요", "speaker_id": "1", "time": "20210518 10:38" },
출신 및 소속	2021-05-17 20:31:00 , P1 : 남편 \소속\다닐때 팀장	{ "id": "MDRW2100000161.17.6", "form": "남편\affiliation다닐때 팀장", "original_form": "남편&\affiliation&다닐때 팀장", "speaker_id": "1", "time": "20210517 20:31" },
이모티콘	2021-05-14 16:26:00 , P2 : 	{ "id": "MDRW2100000001.1.2", "form": "", "original_form": "{emoji:🌟}{emoji:💎}{emoji:❤️} {emoji:❤️}{emoji:🌟}{emoji:😊}{emoji:😐}{emoji:😐}{emoji:👋}{emoji: 👋}{emoji:👋}{emoji:👋}", "speaker_id": "2", "time": "20210514 16:26" },
선물 발송	2021-05-18 11:08:00 , P1 : 김 \이름\님이 선물과 메시지를 보 냈습니다	{ "id": "MDRW2100000088.32.4", "form": "", "original_form": "", "speaker_id": "1", "time": "20210518 11:08" },

항목	txt 원문 표기 예시	JSON 표기 예시
		<pre> "original_form": "{system:gift}", "speaker_id": "1", "time": "20210518 11:08" }, </pre>
무료 통화	2020-03-24 23:57:00 , P1 : 페이지 스톱 해요	<pre> { "id": "MDRW2100001510.1.376", "form": "", "original_form": "{system:call}", "speaker_id": "1", "time": "20200324 23:57" }, </pre>
지도 공유	2021-05-20 15:20:00 , P2 : [네이버 지도] \주소\ http***	<pre> { "id": "MDRW2100000032.13.10", "form": "", "original_form": "{system:map}", "speaker_id": "2", "time": "20210520 15:20" }, </pre>
음악 공유	2021-05-17 16:47:00 , P2 : \m\Superhuman-NCT 127' 음악을 공유했습니다.\m\	<pre> { "id": "MDRW2100000006.5.1", "form": "", "original_form": "{share:music}", "speaker_id": "2", "time": "20210517 16:47" }, </pre>
정보 공유	2021-05-18 15:21:00 , P2 : 삼겹 색 : #서울 앵무새	<pre> { "id": "MDRW2100000003.13.1", "form": "", "original_form": "{share:info}", "speaker_id": "2", "time": "20210518 15:21" }, </pre>
인터넷 링크 공유	2021-05-18 15:28:00 , P1 : https***	<pre> { "id": "MDRW2100000001.1.3", "form": "", "original_form": "{share:url}", "speaker_id": "1", "time": "20210518 15:28" }, </pre>
비윤리적 표현	2021-05-17 17:06:00 , P1 : 앵간 한 \c헬창c도 저거 없더라고	<pre> { "id": "MDRW2100000006.10.19", "form": "앵간한 도 저거 없더라고", </pre>

항목	txt 원문 표기 예시	JSON 표기 예시
		<pre> “original_form”: “앵간한 {censored}도 저거 없더라고“, “speaker_id”: “1“, “time”: “20210517 17:06“ }, </pre>

<표 2-27> 태깅 대상 항목의 txt 원문과 원시 말뭉치 JSON 형식 표기 예시

산출물 형식과 구조에 대한 검수와 수정은 말뭉치 가공 도구 개발 단계에서 사전 정의해 놓은 기능이 산출물에 바르게 적용되었는지를 확인하고, 기능이 바르게 적용되지 않았을 경우, <표 2-28>과 같이 알고리즘 개선 요청 사항을 통해 가공 도구의 알고리즘을 수정하는 방식으로 진행되었다.

예를 들어 <표 2-28>의 첫 번째 항목은 정제 작업 파일에 작업자가 비윤리적 표현과 발화를 표시한 ‘c’ 태그가 JSON의 original_form에서 사전 정의한 ‘{censored}’로 바르게 변환되지 않아 이에 대해 변환 알고리즘 수정을 요청한 예시이다.

이와 같이 산출물의 형식과 구조에 대한 검수는 생성된 산출물을 대상으로 사전 정의된 형식과 구조에 맞게 변환이 이루어졌는지를 검수하면서 말뭉치 가공 도구의 말뭉치 변환 알고리즘을 수정, 개선하는 과정을 통해 진행되었다.

기능	상세 내용	산출물 확인	개선 요청
<ul style="list-style-type: none"> 비윤리적 발화 정제 메시지 변환 	<ul style="list-style-type: none"> txt 메시지 : 욕설 txt 원문 : 욕설(수집 원문 형태 유지) JSON > original_form : {censored} JSON > form : “” 	<ul style="list-style-type: none"> MDRW2100000103.json 파일 “id” : “MDRW2100000103.6.3” “form” : “” “original_form” : “” 	<ul style="list-style-type: none"> “original_form” : “{censored}” 작업 파일 전체 미적용 상태임. 산출물 파일 전체 일괄 변환되도록 수정
<ul style="list-style-type: none"> 불필요 메시지 삭제 	<ul style="list-style-type: none"> 작업자 ‘d’ 태그 항목 txt, JSON에서 일괄 삭제 	<ul style="list-style-type: none"> 최종 산출물 txt 형식에서 해당 태그 발화 삭제 미적용 	<ul style="list-style-type: none"> 정제 작업 파일에서 ‘d’ 태그 항목은 최종 산출물 txt 파일에서도 일괄 삭제되도록 수정
<ul style="list-style-type: none"> 이모지 처리 	<ul style="list-style-type: none"> txt 메시지 : 이모티콘 txt 원문 : 이모니콘 JSON > original_form : {emoji} JSON > form : “” 	<ul style="list-style-type: none"> MDRW2100000103.json 파일 해당 파일의 txt 원문에 없는 이모티콘이 json 파일에 추가됨 “id” : “MDRW2100000103.1.12” “form” : “그러게.” “original_form” : “그러게{emoji:…}.” 	<ul style="list-style-type: none"> 오류 원인 파악하여 txt 원문에 없는 {emoji} 생성된 경우, 수정 알고리즘 적용

<표 2-28> 산출물 형식 오류 검수 및 말뭉치 가공 도구 수정 요청 예시

5.2.2. 산출물 내용에 대한 검수 및 수정

형식과 구조 오류를 제외한 작업 오류 검수와 수정은 텍스트 검색 상용 소프트웨어를 사용했다.

[그림 2-33]은 폴더 내 파일 전체를 대상으로 지정된 표현을 검색할 수 있는 상용 소프트웨어를 활용해서 작업 오류를 확인하는 예시이다. 정제와 태깅 작업 단계에서 태

킹 부호를 ‘\’ 가 아닌 ‘/’ 로 잘못 입력하는 오류가 있었는데, 최종 산출물을 대상으로 다시 한번 이런 오류가 있는 파일이 있는지를 확인하고 검출된 파일의 항목 전체를 다시 확인해서 오류를 수정했다.

이 방법을 통해 정제 작업 단계에서 정리한 비윤리적 표현 목록을 활용한 비윤리적인 표현의 추가 정제 작업도 진행했다.

The screenshot shows a file search tool interface. The top section has a search bar with 'Find: /계정/' and a 'Replace:' field. Below this is a table of search results. The table has columns: File Name, Path, Matches, File Size, Created, Modified, File Type, Encoding, and Attributes. The results list various files with names like 'MDRW210000138.json' and paths like 'C:\Users\ineopel...'. A detailed view of a file is shown below the table, displaying the content of a JSON file. The content shows a list of results, with the first two being '13216' and '13217'. The JSON structure is: 'form': '/계정/', 'original_form': '/계정/'.

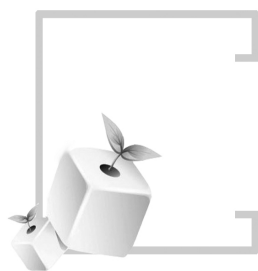
File Name	Path	Matches	File Size	Created	Modified	File Type	Encoding	Attributes
MDRW210000138.json	C:\Users\ineopel...	2	601 KB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001025.json	C:\Users\ineopel...	2	653 KB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001381.json	C:\Users\ineopel...	4	2.32 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001382.json	C:\Users\ineopel...	20	2.84 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001496.json	C:\Users\ineopel...	4	2.58 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001497.json	C:\Users\ineopel...	2	3.85 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001500.json	C:\Users\ineopel...	4	1.80 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001503.json	C:\Users\ineopel...	10	1.42 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001504.json	C:\Users\ineopel...	12	536 KB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001513.json	C:\Users\ineopel...	6	3.86 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001515.json	C:\Users\ineopel...	12	2.50 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001522.json	C:\Users\ineopel...	2	434 KB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001544.json	C:\Users\ineopel...	2	2.20 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001552.json	C:\Users\ineopel...	2	1.79 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001554.json	C:\Users\ineopel...	2	1.90 MB	2021-12-21 오후 1:12:21	2022-02-24 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001555.json	C:\Users\ineopel...	2	1.79 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001562.json	C:\Users\ineopel...	6	1.50 MB	2021-12-21 오후 1:12:21	2022-02-18 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001563.json	C:\Users\ineopel...	2	2.97 MB	2021-12-21 오후 1:12:21	2022-02-24 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001565.json	C:\Users\ineopel...	16	2.97 MB	2021-12-21 오후 1:12:21	2022-02-24 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001571.json	C:\Users\ineopel...	4	3.12 MB	2021-12-21 오후 1:12:21	2022-02-24 오후 1:12:21	JSON 파일	UTF8(NC)	A
MDRW2100001576.json	C:\Users\ineopel...	4	3.12 MB	2021-12-21 오후 1:12:21	2022-02-24 오후 1:12:21	JSON 파일	UTF8(NC)	A

Results 1-2 of 2: C:\Users\ineopel\desktop

13216 "form": "/계정/",
13217 "original_form": "/계정/",

"form": "/계정/",
"original_form": "/계정/",

[그림 2-33] 상용 소프트웨어를 사용한 산출물 전체 작업 오류 검수 예시



제 3 장

온라인 대화 말뭉치 구축 결과



1. 온라인 대화 말뭉치의 구성

1.1. 구축 규모

이 사업은 최소 8회의 말차례로 이루어진 대화 15만 개를 수집하는 것이 목표이다. 실시간 대화 수집을 통해 8회 이상의 말차례로 이루어진 대화를 수집했다. 실시간 대화의 구축 수량은 주제를 기준으로 수집된 문서의 수량으로 산정했다. 주제를 단위로 대화 상황 통제가 불가능한 기존 대화는 말차례의 수를 기준으로 구축 수량을 산정했다.²⁴⁾

이러한 분량 산정 기준에 따라 사업 기간 동안 구축한 온라인 대화 말뭉치의 규모는 <표 3-1>과 같다.

항목	수량(개)
수집 파일 수(참여자 집합 수) ²⁵⁾	47,614
대화 수	151,004
말차례 수	2,283,178
발화 수	4,120,382

<표 3-1> 온라인 대화 원시 말뭉치의 구축 수량

1.2. 유형별 구성

1.2.1. 대화 참여 인원에 따른 구성

대화 참여 인원은 상호 작용과 대화의 양상에 영향을 미치는 변인이다. 이러한 대화 양상의 변화를 고려해서 온라인 대화의 유형을 대화 참여 인원에는 따라 2인 대화와 다자 대화로 나누었다.

대화 참여 인원을 기준으로 한 온라인 대화 말뭉치의 구성은 <표 3-2>와 같다.

24) 통제된 상황에서 주제에 따른 대화 수집이 가능한 실시간 대화와 달리 기존 대화는 사전에 계획된 대화가 아니기 때문에 주관 기관이 요구한 대화 하나당 최소 말차례 기준과 주제에 맞춰 대화를 분할하는 것이 어렵다. 주관 기관과 협의를 통해 수집 방식에 따른 차이를 고려해서 기존 대화는 말차례를 기준으로 아래와 같이 분량을 산정했다.

먼저 카카오톡의 사진이나 동영상 공유, 선물 보내기 등 특수 메시지는 제외하고 유효 발화만을 대상으로 발화 분량을 산정했다. 유효 발화를 대상으로 첫 번째 화자의 최초 발화를 최초 1회로 두고 화자가 교체될 때마다 말차례 수가 증가하는 것으로 말차례의 수를 산정했다. 그리고 말차례 15개를 대화 하나로 산정했다.

25) 대화는 파일 단위로 수집했다. 수집 파일의 숫자는 대화에 참여한 참여자 쌍의 숫자와 동일하다. 중복 참여한 쌍을 포함하여 사업 기간 동안 총 47,614쌍이 대화에 참여했다.

구분		수집 파일 수		대화 수		말차레 수		발화 수	
		분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
2인 대화		47,413	99.58	123,449	81.75	1,842,154	80.68	3,435,723	83.38
다자 대화	3인 대화	103	0.22	11,768	7.79	190,070	8.32	299,568	7.27
	4인 대화	88	0.18	14,891	9.86	237,517	10.40	364,640	8.85
	5인 대화	8	>0.1	769	0.51	11,699	0.51	17,936	0.44
	6인 대화	2	>0.1	127	>0.1	1,738	>0.1	2,515	>0.1
다자 대화 소계		201	0.42	27,555	18.25	441,024	19.32	684,659	16.62
합계		47,614	100	151,004	100	2,238,178	100	4,120,382	100

〈표 3-2〉 대화 참여 인원별 온라인 대화 말뭉치 구성

수집 파일 수를 기준으로 하면 다자 대화가 차지하는 비율은 0.42%이다. 수집 파일 수를 기준으로 다자 대화의 수집 비중이 낮은 이유는 대화 참여자 전원의 대화 제공 동의와 저작권 이용 허락 계약을 체결해야 하는 조건을 충족하기가 2인 대화에 비해 상대적으로 어렵기 때문이다.

대화 수 기준으로 다자 대화가 전체 대화에서 차지하는 비율은 18.25%이다. 수집 파일 하나가 평균 2,194개의 말차레로 구성된 길이가 긴 파일로 이루어져 있어 전체 대화에서 차지하는 비중이 높아졌다.

이 사업에서는 다자 대화 참여 인원 6명 이하로 제한을 뒀다. 2019년 메신저 대화 말뭉치 사업의 경우, 다자 대화 참여 인원 6명 이하로 제한을 두지 않고 대화를 수집했다. 그러나 대화 참여 인원이 많을 경우, 대화의 맥락과 진행 양상을 파악하는 것이 어려워²⁶⁾ 언어 연구나 인공지능의 학습을 위한 말뭉치로 활용성이 떨어질 수 있다는 점을 고려했다.

1.2.2. 수집 방법에 따른 구성

이 사업 기간 동안 대화 수집 방법으로 ‘실시간 대화 수집’과 ‘기존 대화 수집’ 두 가지의 방식을 활용했다.

‘실시간 대화 수집’은 대화 제공자가 대화가 수집된다는 상황을 사전에 인식하고 통제된 상황에서 수집이 이루어지기 때문에 대화 수집 상황을 의식하는 부자연스러운 대화가 진행되거나 의식적으로 표현의 수위를 조절하는 경우도 발생할 수 있다. 반면에 이 방식은 대화 상황을 통제할 수 있기 때문에 대화의 주제 맥락이 일관되고, 개인정보나 혐오 표현의 사용 가능성이 낮고, 표현이 정제될 가능성이 상대적으로 높다. 그렇기 때문에 인공지능 학습을 위한 말뭉치 구축이라는 관점에서는 효용성이 크다.

26) 국립국어원(2019:56)에 따르면 2019년 메신저 대화 말뭉치 사업에서는 참여 인원이 최대 34명인 대화도 수집했다. 많은 인원이 참여하는 대화를 실제로 수집해서 분석해 본 결과, 대화 참여 인원이 많아질수록 대화의 맥락에 일관성이 없어 대화 내용을 파악하는 것이 어려웠다. 이런 점을 고려해서 2021년 사업 기간에는 대화 인원 6명 이하로 제한을 두었다.

이 사업에서는 이러한 효율성을 고려해서 ‘실시간 대화’를 일정 비율 수집했다. 그리고 사전에 통제되지 않고 자연스러운 온라인 대화의 언어 사용 양상을 그대로 담고 있는 ‘기존 대화’ 또한 수집했다.

‘기존 대화’는 대화 제공자의 카카오톡 대화를 txt 파일로 추출하여 구축했고, 실시간 대화 수집은 ‘카카오톡’을 통한 수집과 ‘심심이 채팅²⁷⁾’을 활용한 두 가지 방법으로 수집이 이루어졌다.

수집 방법을 기준으로 한 온라인 대화 말뭉치의 구성은 <표 3-3>과 같다.

구분		수집 파일 수		대화 수		말차례 수		발화 수	
		분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
기존 대화 수집		301	0.63	69,508	46.03	1,042,623	45.67	2,111,364	51.24
실시간 대화 수집	카카오톡	1,349	2.83	35,530	23.53	536,958	23.52	873,013	21.19
	심심이 채팅	45,964	96.53	45,966	30.44	703,597	30.82	1,136,005	27.57
실시간 대화 수집 소계		47,313	99.37	81,496	53.97	1,240,555	54.33	2,009,018	48.76
합계		47,614	100	151,004	100	2,283,178	100	4,120,382	100

<표 3-3> 수집 방법별 온라인 대화 말뭉치 구성

수집 파일 수를 기준으로 하면 기존 대화는 0.63%의 비율을 차지한다. 수집 파일 수를 기준으로 하면 기존 대화의 수집 비율은 1% 미만이지만, 기존 대화는 수집 파일 하나가 평균 3,463개의 말차례로 이루어져 있어 대화 수를 기준으로 하면 46.03% 비율을 차지하고, 발화 수를 기준으로 하면 전체 구축 수량의 절반에 가까운 51.24%를 차지한다.

카카오톡을 활용한 실시간 대화 수집은 수집 파일 수를 기준으로 2.83%의 비중을 차지한다. 카카오톡 실시간 대화의 수집 파일 하나는 평균 398개의 말차례로 이루어져 있는 반면, 심심이 채팅의 수집 파일 하나는 평균 15개의 말차례로 수집 파일 하나가 하나의 대화로 이루어져 있다. 따라서 대화 수를 기준으로 하면 카카오톡 실시간 대화는 23.53%, 심심이 채팅 대화는 30.44%의 비율을 차지한다.

심심이 채팅을 활용한 대화 수집 기간이 상대적으로 길지 않았음에도 불구하고 대화 수를 기준으로 카카오톡 실시간 대화에 비해 더 많은 양을 수집할 수 있었던 이유는 심심이 채팅을 통해 대화를 진행하고 제공하는 과정이 카카오톡 실시간 대화에 비해 복잡하지 않았기 때문이다.

1.2.3. 수집 매체에 따른 구성

27) ‘심심이’는 원래 사람과 인공지능 챗봇 ‘심심이’가 대화를 나누는 서비스다. 이번 사업 기간 동안 ‘심심이’의 기존 대화 플랫폼을 활용해 온라인 대화 수집을 위한 별도의 채팅 기능을 개발해서 실시간 대화 수집에 활용했다.

카카오톡 이외에도 라인과 페이스북 메신저, 네이트온과 같이 다양한 온라인 대화 매체가 있다. 사용자에게 따라서 메신저를 구분해서 사용하기도 하는 등, 사용 매체에 따라 대화 양상이 달라질 수 있기 때문에 다양한 매체로부터 온라인 대화를 수집할 필요가 있지만, 수집 매체가 다양해질 경우, 말뭉치를 구축하는 과정에서 형식을 표준화해야 하는 부담이 크다.

이 사업에서는 한국의 대표적인 온라인 채팅 매체인 카카오톡에 더해 온라인 공간에서 이루어지는 대화를 관찰할 수 있는 매체로서 심심이 온라인 채팅을 추가로 활용해 온라인 대화를 수집했다.

카카오톡은 감정 표현을 위해 다양한 이모티콘과 생활 편의를 돕는 다양한 기능을 제공한다. 그런데 이런 기능은 대화 참여자의 발화와 구분되는 특수한 메시지이다. 말뭉치 구축 과정에서는 대화 참여자 발화와 구분하기 쉬운 형태로 가공이 되어야 한다.

특수 메시지를 포함하고 있는 카카오톡과 달리 심심이 채팅은 이모티콘과 특수 메시지 등의 요소를 사용하지 않고 대화가 이루어지도록 개발이 되었다. 또한 심심이 채팅은 12개의 말차례를 기준으로 분할된 대화를 쉽게 제공할 수 있다. 그렇기 때문에 단일 대화의 길이가 긴 카카오톡 대화에 비해 단일 대화의 길이가 짧다.

수집 매체에 따른 말뭉치의 형태와 대화 구성 방식의 차이를 고려해 카카오톡 대화 수집과 심심이 채팅 대화 수집으로 말뭉치의 구성을 구분했다.

수집 매체를 기준으로 한 온라인 대화 말뭉치의 구성은 <표 3-4>와 같다.

구분	수집 파일 수		대화 수		말차례 수		발화 수	
	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
카카오톡 대화 수집	1,650	3.47	105,038	69.56	1,579,581	69.18	2,984,377	72.43
심심이 채팅 대화 수집	45,964	96.53	45,966	30.44	703,597	30.82	1,136,005	27.57
합계	47,614	100	151,004	100	2,283,178	100	4,120,382	100

<표 3-4> 수집 매체별 온라인 대화 말뭉치 구성

수집 파일 수를 기준으로 하면 카카오톡 대화의 비율은 3.47%, 심심이 채팅 대화의 비율은 96.53%이다. 대화 수를 기준으로 하면 카카오톡 대화는 69.56%이고, 심심이 채팅은 30.44%의 비율을 차지한다. 앞서 이야기했듯이 심심이 채팅은 파일 하나가 평균 15개 말차례의 대화 하나로 이루어져 있다. 반면에 카카오톡 대화는 수집 파일 하나가 평균 957개의 말차례와 평균 63개의 대화로 이루어져 있다.

1.2.4. 주제에 따른 구성

대화 상황 통제가 가능한 ‘실시간 대화’는 대화 참여자가 주제를 선택하고 선택한 주제로 대화하는 것이 가능하다. 반면 ‘기존 대화’는 주제를 통제하는 것이 불가능하다.

이 사업에서는 대화 참여자가 선택한 주제로 대화를 진행한 ‘주제 대화’와 사전 통제가 없는 ‘기타 일상 대화’로 구분하여 대화를 수집했다.

주제 대화와 기타 일상 대화의 구축 분량은 <표 3-5>와 같다.

구분	수집 파일 수		대화 수		말차례 수		발화 수	
	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
기타 일상 대화	301	0.63	70,303	46.56	1,053,675	46.15	2,127,857	51.64
주제 대화	47,313	99.37	80,701	53.44	1,229,503	53.85	1,992,525	48.36
합계	47,614	100	151,004	100	2,283,178	100	4,120,382	100

<표 3-5> 주제 유형별 온라인 대화 말뭉치 구성

수집 파일 수를 기준으로 하면 기타 일상 대화와 주제 대화의 구성 비율은 0.63%와 99.37%로 큰 차이를 보인다. 다만 말차례를 기준으로 할 경우에는 기타 일상 대화와 주제 대화는 46.15%와 53.85%로 구축 비율에 큰 차이를 보이지 않는다.

주제 대화의 세부 주제별 구축 분량은 <표 3-6>과 같다.

구분	수집 파일 수		대화 수		말차례 수		발화 수	
	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
가사 및 가족	2,637	5.25	6,144	7.61	93,357	7.59	147,755	7.42
학교 생활	1,701	3.38	3,469	4.3	54,859	4.46	96,181	4.83
일과 직업	2,093	4.16	4,611	5.71	71,765	5.84	121,350	6.09
기타 사회 생활 및 활동	2,052	4.08	3,764	4.66	58,356	4.75	95,048	4.77
연애와 결혼	1,661	3.3	3,261	4.04	51,505	4.19	85,833	4.31
반려동물	2,003	3.98	2,538	3.14	38,163	3.10	61,583	3.09
미용과 건강	3,651	7.26	5,832	7.23	87,858	7.15	137,839	6.92
여행	3,092	6.15	4,225	5.24	63,286	5.15	102,886	5.16
식음료	12,035	23.94	16,142	20.00	241,690	19.66	388,950	19.52
쇼핑과 상품	4,277	8.51	7,498	9.29	113,884	9.26	183,727	9.22
날씨와 계절	2,891	5.75	3,150	3.90	45,206	3.68	69,304	3.48
콘텐츠	7,989	15.89	13,901	17.23	216,994	17.65	350,067	17.57
공연 및 관람	2,304	4.58	2,572	3.19	38,827	3.16	63,672	3.20
시사	642	1.28	1,637	2.03	23,763	1.93	40,197	2.02
일상 트렌드	1,241	2.47	1,957	2.43	29,990	2.44	48,133	2.42
합계	50,269 ²⁸⁾	100	80,701	100	1,229,503	100	1,992,525	100

<표 3-6> 주제 대화 세부 주제별 온라인 대화 말뭉치 구성

식음료나 쇼핑과 상품, 가사와 가족, 미용과 건강 등과 같이 의식주와 일상생활 관련

28) 수집 파일 하나가 두 개 이상의 주제 대화 문서로 구성된 경우가 있기 때문에 전체 수집 파일 47,313개에 비해 수량이 더 많은 것으로 나타난다.

된 내용이 높은 비율로 나타났다. 2019년 메신저 대화 말뭉치에서 20% 가량의 비율로 나타난 여행 주제가 5.24%로 다소 낮은 비율을 차지하는 반면, 콘텐츠 주제의 비율이 17.23%를 차지하는 것은 코로나(COVID-19)로 인한 사람들의 관심사와 관련하여 시사하는 바가 있다고 본다.²⁹⁾

이 사업에서는 자발적으로 주제를 선택하게 했기 때문에 특정 주제의 비율이 높게 나타나는 것은 사람들의 관심사를 반영하는 것이라 볼 수 있다. 그러나 내용의 구성에서 말뭉치의 다양성을 확보하기 위해서는 특정한 주제에 대한 편중이 발생하지 않도록 면밀한 설계가 필요하다. 이 사업을 통해 구축한 자료를 통하여 온라인 대화의 특성을 고려한 주제 분류 체계를 마련하기 위한 연구가 필요하다.³⁰⁾

1.2.5. 대화 참여자 간 관계에 따른 구성

대화 참여자 간 관계는 언어의 형태와 내용에 영향을 미치는 언어 사용의 주요 변인이다. 이 사업에서는 개인 사회 활동의 근간이 되는 ‘가족, 학교, 직장, 지역, 그 외 기타’의 범주로 대화 참여자 간 관계를 구분하고 각 범주의 하위에 세부 관계를 설정하여 대화를 수집했다.

대화 참여자 간 관계에 따른 온라인 대화 말뭉치 구축 분량은 <표 3-7>과 같다.

구분		수집 파일 수		대화 수		말차례 수		발화 수	
		분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
가족	부부	2,104	4.42	8,581	5.68	127,178	5.57	211,255	5.13
	형제-자매	1,524	3.20	8,552	5.66	126,320	5.53	220,457	5.35
	부모-자녀	1,158	2.43	3,274	2.17	48,202	2.11	69,179	1.68
	기타(조부모-손주 및 친인척)	15	>0.1	2,253	1.49	33,671	1.47	66,527	1.61
	가족 복수 선택 ³¹⁾	6	>0.1	174	0.12	3,117	0.14	4,456	0.11
가족 소계		4,807	10.10	22,834	15.12	338,488	14.83	571,874	13.88
학교/	동기/동창/동급생	2,680	5.63	39,150	25.93	592,316	25.94	1,177,180	28.57

29) 다만 2019년 메신저 대화 말뭉치 주제 체계에는 ‘콘텐츠’ 범주가 독립적으로 제시되지 않고 ‘여가와 오락’ 범주 안에 포함되어 있다. 이처럼 2019년 메신저 대화 말뭉치 주제 체계와 2021년 온라인 대화의 주제 체계가 다르기 때문에 둘을 단순 비교하는 것은 무리가 있다. 실제 대화 내용의 분석을 통해 시대적 상황과 주제 선택의 관련성은 정밀한 검증이 필요하다.

30) 자체 연구와 주관 기관과의 협의를 통해 온라인 대화의 주제 분류 체계를 설계했지만, 이 사업의 주제 분류 체계는 온라인 대화의 특성을 충분히 고려했다고 보기에는 아직 미흡한 수준이다. ‘기타 일상 대화’는 주제가 분류되어 있지 않다. 이 사업에서 사전에 설계한 주제 체계가 아니라, 연구자의 관점과 역량에 따라 자유롭게 온라인 대화의 주제 분류 체계를 분석하고 연구할 수 있는 자료로 활용되기를 바란다.

31) 2인 대화가 아닌 다자 대화의 경우, 참여자 간 관계 복수 선택이 가능했다. 다자 대화 중

구분		수집 파일 수		대화 수		말차례 수		발화 수	
		분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
학원	선후배	106	0.22	1,260	0.83	19,663	0.86	344,409	0.84
	교강사-제자	0	0	0	0	0	0	0	0
	기타 (직원-학생 등)	3	>0.1	3	>0.1	39	>0.1	75	>0.1
	학교/학원 복수 선택	4	>0.1	581	0.38	9,391	0.41	16,160	0.39
학교/학원 소계		2,793	5.87	40,994	27.15	621,409	27.22	1,227,824	29.80
직장	동기/동료/동업자	667	1.4	6,824	4.52	103,867	4.55	145,585	3.53
	선후배/상사-부하	33	>0.1	886	0.59	12,835	0.56	18,993	0.46
	기타	0	0	0	0	0	0	0	0
	직장 복수 선택	3	>0.1	290	0.19	5,506	0.24	10,072	0.24
직장 소계		703	1.48	8,000	5.30	122,208	5.35	174,650	4.24
지역	고향 및 이전 거주지 지인	367	0.77	5,384	3.57	80,754	3.54	208,141	5.05
	현 거주지 지인	668	1.40	7,288	4.83	110,658	4.85	216,366	5.25
	지역 복수 선택	0	0	0	0	0	0	0	0
지역 소계		1,035	2.17	12,672	8.39	191,412	8.38	424,507	10.30
기타	연인	259	0.54	10,361	6.86	154,919	6.79	334,133	8.11
	온라인 커뮤니티	456	0.96	8,226	5.45	128,019	5.61	213,010	5.17
	동호회/스터디	139	0.29	1,786	1.18	28,883	1.27	48,734	1.18
	종교 관련	1	>0.1	51	>0.1	548	>0.1	1,119	>0.1
	그 외 사회적 관계	313	0.66	3,054	2.02	46,014	2.02	76,403	1.85
	기타 복수 선택	9	>0.1	1,213	0.80	19,073	0.84	28,017	0.68
기타 소계		1,177	2.47	24,691	16.35	377,456	16.53	701,416	17.02
낮선 사람		37,070	77.86	37,966	25.14	572,568	25.08	930,434	22.58
다른 범주 간 복수 선택 ³²⁾		29	>0.1	3,847	2.55	59,637	2.61	89,677	2.18
합계		47,614	100	80,701	100	2,238,178	100	4,120,382	100

〈표 3-7〉 대화 참여자 간 관계별 온라인 대화 말뭉치 구성

수집 파일 수를 기준으로 가장 비율이 높은 대화 참여자 관계는 77.86%를 차지하는 ‘낮선 사람’이다.³³⁾ ‘낮선 사람’의 대화는 심심이 채팅을 활용해서 수집했기 때문

동일한 범주에서 복수 선택을 한 경우를 해당 범주 내의 복수 선택 항목으로 분량을 산정했다.

32) 참여자 간 관계가 다른 범주에서 복수 선택을 한 경우를 다른 범주 간 복수 선택 항목으로 분량을 산정했다.

33) 일상적인 온라인 대화 환경에서 낮선 관계의 대화를 수집하는 것은 일반적이지는 않다.

대화 참여자 모두로부터 대화 제공에 대한 동의를 얻어야 하는 사업의 특성상 친밀도가 높은 관계의 대화의 비중이 높을 수밖에 없다. 그런데 친밀도가 높은 관계는 상호 공유하는 맥락을 기반으로 대화를 진행하기 때문에 대화 내용이 생략될 가능성과 상대적으로 언어 형태도 정제되지 않을 가능성이 크다. 반면 낮선 관계의 대화는 생략의 가능성이 낮고, 언어 형태 또한 정제되어 있을 가능성이 크다. 낮선 관계의 대화도 연구를 위한 자료로서 효용을 고려해서 일정 비율 수집이 이루어졌다.

에 수집 파일 하나가 하나의 대화로 이루어져 있다. 대화 수 기준으로 ‘낮선 사람’의 대화는 25.1%의 비율을 차지한다.

대화 수를 기준으로 할 때 학교/학원의 동기 또는 동급생, 동창 관계가 25.93%의 비율로 전체 범주 중에서 가장 높은 비율을 차지한다. 학생 관계는 수집 파일을 기준으로 낮선 관계 다음으로 높은 비율을 차지한다. 온라인 대화 참여자 중 학생 직업이 26.10%로 가장 높은 비율을 차지하고, 학교나 학원의 동기에게 대화 제공을 요구하는 부담이 상대적으로 적었기 때문이라 볼 수 있다.

낮선 관계를 제외하고 부부, 형제자매, 직장 동료, 연인 등과 같이 상하 위계로부터 자유로운 관계의 대화 제공 건수가 많은 반면, 교강사-제자, 학교 선후배, 직장 선후배 관계와 같이 상하 위계가 있는 관계의 대화는 수집 비율이 1% 미만으로 나타난다.

1.2.6. 대화 참여자 간 친밀도에 따른 구성

대화 참여자 간 친밀도도 대화 참여자 간 관계와 함께 언어의 형태와 내용에 영향을 미치는 언어 사용의 주요 변인이다. 대화 참여자 간 친밀도는 1(친밀도가 낮음)~5(친밀도가 높음) 중에서 대화 참여자 스스로 입력했다. 대화 참여자 간 관계를 낮선 사람으로 선택한 경우에는 말뭉치 가공 단계에서 자동으로 0으로 입력이 되었다.

대화 참여자 간 친밀도에 따른 온라인 대화 말뭉치 구축 분량은 <표 3-8>과 같다

구분	수집 파일 수		대화 수		말차례 수		발화 수	
	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
0(낮선 관계)	37,070	77.86	37,966	25.14	572,568	25.08	930,434	22.58
1	115	0.24	115	>0.1	1,851	>0.1	2,346	>0.1
2	6	>0.1	141	>0.1	2,327	0.1	3,975	0.1
3	1,336	2.81	5,820	3.85	91,721	4.02	166,748	4.05
4	2,680	5.63	19,014	12.59	290,872	12.74	506,472	12.29
5	6,407	13.46	87,948	58.24	1,323,839	57.98	2,510,407	60.93
합계	47,614	100	151,004	100	2,283,178	100	4,120,382	100

<표 3-8> 대화 참여자 간 친밀도별 온라인 대화 말뭉치 구성

대화 참여자 간 관계와 마찬가지로 상대방에게 대화 제출을 요구하기에 부담이 없는 친밀도 높은 관계의 대화 비율이 높다. 파일 하나가 대화 하나로 이루어진 낮선 관계 대화를 제외하면 친밀도 5인 대화는 수집 파일 수를 기준으로 13.46%, 대화 수 기준으로 58.24% 비율로 가장 높은 비율로 나타났다.

친밀도가 0인 대화는 수집 파일 수를 기준으로 가장 높은 비율인 77.86%를 차지하고 있다. 다만 친밀도 0인 대화는 심심이 채팅을 통해 수집이 이루어졌고 파일 하나가 평균 말차례 15개의 짧은 길이이다. 대화 수 기준으로는 친밀도 5 대화의 절반 이하 분

량인 25.14% 비율을 차지한다. 친밀도가 1과 2인 대화는 수집 파일 기준으로 0.3% 미만, 대화 수 기준으로는 0.1% 미만의 낮은 비율로 나타났다.

대화 참여자 간 상하 위계가 있는 관계의 대화를 수집하기가 어려운 것과 마찬가지로, 대화 참여자 간 친밀도가 낮은 대화도 수집하기가 어려운 현실적 제한이 나타난다.³⁴⁾

대화 참여자 간 관계에 따른 친밀도를 수집 파일 기준으로 나타내면 <표 3-9>와 같다³⁵⁾.

친밀도 참여자 간 관계		5		4		3		2		1		합계	
		분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
가족	부부	1,436	13.6 (68.3)	464	4.4 (22.1)	201	1.9 (9.6)	1	>0.1 (>0.1)	2	>0.1 (0.1)	2,104	20.0 (100)
	형제-자매	556	5.3 (36.5)	962	9.1 (63.1)	6	0.1 (0.4)	0	0	0	0	1,524	14.5 (100)
	부모-자녀	1,133	10.7 (97.8)	23	0.2 (2.0)	2	>0.1 (0.2)	0	0	0	0	1,158	11.0 (100)
	기타(조부모- 손주 및 친인척 등)	10	0.1 (66.7)	5	>0.1 (33.3)	0	0	0	0	0	0	15	0.1 (100)
	가족 복수 선택	3	>0.1 (50.0)	2	>0.1 (33.3)	1	>0.1 (16.7)	0	0	0	0	6	0.1 (100)
가족 소계		3,138	29.8 (65.3)	1,456	13.8 (30.3)	210	2.0 (4.4)	1	>0.1 (>0.1)	2	>0.1 (>0.1)	4,807	45.6 (100)
학교/ 학원	동기/동창/ 동급생	1,821	17.3 (67.9)	598	5.7 (22.3)	246	2.3 (9.2)	0	0	15	0.1 (0.6)	2,680	25.4 (100)
	선후배	61	0.6 (57.5)	17	0.2 (16.0)	27	0.3 (25.5)	1	>0.1 (0.9)	0	0	106	1.0 (100)
	교강사-제자	0	0	0	0	0	0	0	0	0	0	0	0
	기타(직원- 학생 등)	2	>0.1 (66.7)	0	0	1	>0.1 (33.3)	0	0	0	0	3	>0.1 (100)
	학교/학원 복수 선택	4	>0.1 (100)	0	0	0	0	0	0	0	0	4	>0.1 (100)
학교/학원 소계		1,888	17.9 (67.6)	615	5.8 (22.0)	274	2.6 (9.8)	1	>0.1 (>0.1)	15	0.1 (0.5)	2,793	26.5 (100)

34) 친밀도가 1인 대화는 수집 파일과 대화의 개수가 동일한 것으로 보아 심심이 채팅을 통해 수집된 대화이다. <표 3-9>에서도 나타나듯이 이 대화는 학교 동기, 직장 동료, 연인 간의 대화인데, 이들이 심심이 채팅을 통해 대화를 진행하고 친밀도 1이 가장 높은 것으로 혼동해서 기재한 것이 아닌가 추측한다.

35) <표 3-9>의 비율 중 괄호 항목 안에 기재된 숫자는 관계 범주 내에서 해당 친밀도의 구성 비율을 나타낸다.

친밀도 참여자 간 관계		5		4		3		2		1		합계	
		분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
직장	동기/동료/ 동업자	327	3.1 (49.0)	32	0.3 (4.8)	218	2.1 (32.7)	0	0	90	0.9 (13.5)	667	6.3 (100)
	선후배, 상사-부하	5	>0.1 (15.2)	4	>0.1 (12.1)	24	0.2 (72.7)	0	0	0	0	33	0.3 (100)
	기타	0	0	0	0	0	0	0	0	0	0	0	0
	직장 복수 선택	1	>0.1 (33.3)	2	>0.1 (66.7)	0	0	0	0	0	0	3	>0.1 (100)
직장 소계		333	3.2 (47.4)	38	0.4 (5.4)	242	2.3 (34.4)	0	0	90	0.9 (12.8)	703	6.7 (100)
지역	고향 및 이전 거주지 지인	51	0.5 (13.9)	216	2.0 (58.9)	100	0.9 (27.2)	0	0	0	0	367	3.5 (100)
	현 거주지 지인	399	3.8 (59.7)	119	1.1 (17.8)	150	1.4 (22.5)	0	0	0	0	668	6.3 (100)
	지역 복수 선택	0	0	0	0	0	0	0	0	0	0	0	0
지역 소계		450	4.3 (43.5)	335	3.2 (32.4)	250	2.4 (24.2)	0	0	0	0	1,035	9.8 (100)
기타	연인	249	2.4 (96.1)	1	>0.1 (0.4)	1	>0.1 (0.4)	0	0	8	0.1 (3.1)	259	2.5 (100)
	온라인 커뮤니티	142	1.3 (31.1)	170	1.6 (37.3)	140	1.3 (30.7)	4	>0.1 (0.9)	0	0	456	4.3 (100)
	동호회/ 스터디	23	0.2 (16.5)	24	0.2 (17.3)	92	0.9 (66.2)	0	0	0	0	139	1.3 (100)
	종교 관련	1	>0.1 (100)	0	0	0	0	0	0	0	0	1	>0.1 (100)
	그 외 사회적 관계	161	1.5 (51.4)	32	0.3 (10.2)	120	1.1 (38.3)	0	0	0	0	313	3.0 (100)
	기타 복수 선택	6	0.1 (66.7)	0	0	3	>0.1 (33.3)	0	0	0	0	9	0.1 (100)
기타 소계		582	5.5 (49.4)	227	2.2 (19.3)	356	3.4 (30.2)	4	>0.1 (0.3)	8	0.1 (0.7)	1,177	11.2 (100)
복수의 관계		16	0.2 (55.2)	9	0.1 (31.0)	4	>0.1 (13.8)	0	0	0	0	29	0.3 (100)
합계		6,407	60.8	2,680	25.4	1,336	12.7	6	0.1	115	1.1	10,544	100

<표 3-9> 대화 참여자 간 관계에 따른 친밀도 구성(수집 파일 기준)

대부분의 관계에서 친밀도 5가 가장 높은 비율을 차지한다. 그런데 형제 관계는 친밀도 5의 비율이 36.5%인 반면, 친밀도 4의 비율은 63.1%로 더 높게 나타난다. 고향 및 이전 거주지 지인 관계의 경우도 친밀도 5의 비율이 13.9%인 반면, 친밀도 4의 비율은 58.9%로 더 높게 나타난다. 직장인 간 다자 대화의 경우도 친밀도 5의 비율이 33.3%, 친밀도 4의 비율이 66.7%로 나타나지만 수집 파일의 전체 개수가 3개이기 때문에 유의미한 수치는 아니다.

1.2.7. 대화 참여자 간 연락 빈도에 따른 구성

참여자 간 친밀도가 주관적인 판단이기 때문에 이를 보완하기 위해 참여자 간 연락 빈도를 분류 항목으로 설정했다.³⁶⁾ 대화 제공자 스스로 ‘월 1회 미만, 주 1회 미만, 주 1~2회, 주 3회 이상, 거의 매일’ 항목 중에서 선택했고, 낯선 관계인 경우에 ‘처음 연락한다’ 항목을 가공 단계에서 자동 생성했다.

대화 참여자 간 연락 빈도에 따른 구성 비율은 <표 3-10>과 같다.

구분	수집 파일 수		대화 수		말차례 수		발화 수	
	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
(거의)매일 연락한다	4,379	9.20	87,234	57.77	1,316,737	57.67	2,478,533	60.15
월 1회 미만	62	0.13	222	0.15	3,370	0.15	6,357	0.15
주 1~2회	1,726	3.62	5,722	3.79	87,059	3.81	142,524	3.46
주 1회 미만	997	2.09	1,382	0.92	22,777	1.00	40,975	0.99
주 3회 이상	3,380	7.10	18,478	12.24	280,667	12.29	521,559	12.66
처음 연락한다	37,070	77.86	37,966	25.14	572,568	25.08	930,434	22.58
합계	47,614	100	151,004	100	2,283,178	100	4,120,382	100

<표 3-10> 대화 참여자 간 친밀도별 온라인 대화 말뭉치 구성

거의 매일 연락하는 관계의 참여자 쌍은 수집 파일 수 기준으로 9.2% 비율로 나타났다. 주 3회 이상 연락하는 관계의 참여자 쌍은 수집 파일 수 기준으로 7.1%의 비율로 나타났다. 주 2회 미만으로 비교적 연락 빈도가 낮은 참여자 쌍은 수집 파일 수 기준으로 5.85%로 나타났다.

대화 수를 기준으로 할 경우, 주 3회 이상과 거의 매일 연락을 하는 관계의 대화 비율은 70.01%로 나타났고, 주 1~2회 미만으로 비교적 연락 빈도가 낮은 관계의 대화 비율은 전체의 4.85%로 나타났다.

연락 빈도에 따른 친밀도를 수집 파일 수 기준으로 나타내면 <표 3-11>과 같다.³⁷⁾

36) 가족과 같이 친밀도가 높아도 연락 빈도가 낮은 관계도 있는 반면, 친밀도가 낮아도 연락 빈도가 높은 공적 관계도 있기 때문에 연락 빈도가 친밀도를 보완하는 지표로 삼기에는 부족한 점이 있다. 친밀도가 주관적이기는 하지만, 상호 간에 인식하는 친밀도가 언어의 형태나 내용에 영향을 끼치는 부분이 있기 때문에 배제할 수는 없다. 다만 친밀도의 주관성을 보완할 다른 지표를 마련할 필요는 있다.

37) <표 3-11>의 비율 중 괄호 항목 안에 기재된 숫자는 연락 빈도 범주 내에서 해당 친밀도의 구성 비율을 나타낸다.

친밀도 연락 빈도	5		4		3		2		1		합계	
	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)	분량 (개)	비율 (%)
(거의)매일 연락한다	3,098	29.4 (70.7)	949	9.0 (21.7)	310	2.9 (7.1)	2	>0.1 (>0.1)	20	0.2 (0.5)	4,379	41.5 (100)
월 1회 미만	12	0.1 (19.4)	5	>0.1 (8.1)	25	0.2 (40.3)	4	>0.1 (6.5)	16	0.2 (25.8)	62	0.6 (100)
주 1~2회	104	1.0 (6.0)	1,005	9.5 (58.2)	538	5.1 (31.2)	0	0	79	0.7 (4.6)	1,726	16.4 (100)
주 1회 미만	123	1.2 (12.3)	465	4.4 (46.6)	409	3.9 (41.0)	0	0	0	0	997	9.5 (100)
주 3회 이상	3,070	29.1 (90.8)	256	2.4 (7.6)	54	0.5 (1.6)	0	0	0	0	3,380	32.1 (100)
합계	6,407	60.8	2,680	25.4	1,336	12.7	6	0.1	115	1.1	10,544	100

<표 3-11> 연락 빈도에 따른 친밀도 구성(수집 파일 기준)

거의 매일 연락하는 참여자 집합에서 친밀도 5의 비율은 70.7%, 친밀도 4의 비율은 21.7%로 나타났다. 주 3회 이상 연락하는 참여자 집합은 친밀도 5의 비율이 90.8%로 가장 높게 나타나고, 친밀도 4의 비율은 7.6%로 나타났다.

월 1회 미만 연락하는 참여자 집합에서는 친밀도 3의 비율이 40.3%로 가장 높게 나타나고, 친밀도 5의 비율이 19.4%로, 친밀도 4의 비율인 8.1%보다 높게 나타난다. 주 1~2회 연락하는 참여자 집합과 주 1회 미만 연락하는 참여자 집합은 친밀도 4의 비율이 58.2%와 46.6%로 가장 높게 나타났다.

주 3회 이상 연락하는 관계의 경우에는 친밀도 5의 비율이 높게 나타났지만, 주 2회 미만 연락하는 관계에서는 친밀도와 연락 빈도와의 관련성이 두드러진다고 보기 어렵다.

2. 온라인 대화 말뭉치의 참여자 구성

사업 초기 계획 단계에서부터 온라인 대화 말뭉치가 온라인 대화 사용자 모집단을 대표성 있게 반영하는 것을 최대한 고려하되, 현실적인 수집 가능성까지 염두에 두고 참여자의 성별, 연령별 구성 비율을 설계했다. 실제 사업 수행 단계에서도 참여자 모집 현황을 확인하고 주관 기관과 협의를 통해 참여자의 구성을 조정했다.

2.1. 성별 및 연령

온라인 대화 말뭉치 구축 대화 참여자의 성별과 연령에 따른 분포는 <표 3-12>와 같다.

구분	남성			여성			합계	
	참여 인원 (명)	총 인원 대비 비율 (%)	범주 내 비율 (%)	참여 인원 (명)	총 인원 대비 비율 (%)	범주 내 비율 (%)	참여 인원 (명)	총 인원 대비 비율 (%)
10대	57	1.62	17.22	274	7.80	82.78	331	9.42
20대	487	13.86	27.67	1,273	36.23	72.33	1,760	50.09
30대	446	12.69	44.20	563	16.02	55.80	1,009	28.71
40대 이상	194	5.52	48.68	220	6.26	53.14	414	11.78
합계	1,184	33.69	33.69	2,330	66.31	66.31	3,514	100

<표 3-12> 온라인 대화 참여자의 성별 및 연령 구성

저작권 이용 허락 계약 등의 모든 절차를 완료하고 대화를 제공한 3,643명 중에서 자료 선별 등을 통해 말뭉치 구축 대상에 포함되지 않은 인원을 제외하고 전체 3,514명이 제공한 대화를 온라인 대화 말뭉치로 구축했다.

대화 참여자의 성별 비율은 남성이 33.69%, 여성이 66.31%로 최초 계획 단계에 목표로 한 40:60 구성 비율에 근접했지만, 여성의 참여 비율이 상대적으로 높다. 남성 참여자의 참여율을 높이기 위해 대화 참여자 중에 남성이 반드시 포함해야 한다는 조건을 제시하는 등 노력을 기울였음에도 남성 참여자의 참여율이 낮은 데에는 다양한 요인이 작용하는 것으로 보인다.

남성과 여성의 온라인 대화 매체 사용 비율이 86.5%와 94.4%로 여성의 온라인 대화 매체 사용 비율이 높고³⁸⁾, 카카오톡의 플러스 채널 등을 통해 할인이나 이벤트 정보를 접하는 비율에서도 여성 82%와 남성 56.4%로 현저한 차이를 보인다.³⁹⁾ 이 사업의 경우에도 남성에 비해 여성이 대화 참여자 모집 등의 소식을 접하고 참여하는 비중이 높았

38) DMC MEDIA(2017:4)

39) DMC MEDIA(2019:12)

던 것으로 추정할 수 있다.

대화 참여자 성별의 불균형을 해결하기 위해 남성 사용자가 많은 온라인 커뮤니티를 대상으로 온라인 대화 참가자 모집을 집중적으로 홍보하는 등 보완 대책을 수립하고 실시해서 성별 구성 목표인 40:60을 달성하고자 했다.

대화 참여자의 연령에 따른 구성 비율은 20대가 50.09%로 가장 높은 비율을 차지한다. 30대는 목표 구성 비율인 30%에 근접한 28.71%의 비율을 차지한다. 온라인 커뮤니티나 누리소통망 등을 통해 정보를 접하고 공유하고 실제로 참여하는 것에 익숙한 연령대이기 때문에 대화 참여자 모집을 위한 이벤트에도 적극적으로 반응하여 참여율이 높은 것으로 보인다.

반면 10대와 40대 이상 참여 비율이 목표 구성 비율이었던 20%에 비해 참여가 다소 저조하다. 10대와 40대 이상 연령대의 참여율이 상대적으로 낮은 요인에 대해서는 분석이 필요하다.

먼저 10대의 경우, 심심이 서비스의 10대 사용자 비중이 높아 10대의 참여율이 높을 것이라는 예상과는 달리 참여율은 9.42%로 나타난다. 대화 참여자 모집을 위해 주로 카카오톡의 플러스 채널과 온라인 커뮤니티를 통한 홍보를 진행했는데, 온라인 공간에서 10대가 정보를 얻고 주로 활동하는 커뮤니티와는 다소 차이가 있어 10대를 대상으로 하는 홍보 효과가 높지 않았던 것으로 보인다.

40대 이상 연령대의 참여율은 11.78%로 나타난다. 대화 제공 보상에 대한 가치를 상대적으로 낮은 것으로 평가하는 연령대로, 대화 제공 보상이 대화 참여를 유도하는 요인으로서 미치는 효과가 2, 30대에 비해 낮았다. 20대와 30대의 부모 세대라는 점을 고려해 대화 제공에 참여하고 있는 2, 30대가 부모님과 대화를 진행하도록 유도하는 등의 방법으로 참여율을 높이하고자 했다.

온라인 대화 말뭉치가 한국어 온라인 대화 사용 양상을 대표성 있게 반영하려면 성별과 연령별 균형을 고려한 대화 수집이 필요하며, 이를 위해서는 성별과 연령대별로 홍보 방식과 보상 체계 등을 다양하게 설계할 필요가 있다.

2.2. 직업

대화 참여자의 직업은 사회 생활의 중요한 부분을 차지하는 만큼 대화의 내용에 영향을 미치는 요소이다. 그리고 동일 직종이나 같은 직장 동료 간 대화에서는 직무, 직책 등과 관련된 용어의 사용 등, 일상적인 대화와 변별되는 요소가 나타난다.

구분	남성			여성			합계	
	참여 인원 (명)	총 인원 대비 비율 (%)	범주 내 비율 (%)	참여 인원 (명)	총 인원 대비 비율 (%)	범주 내 비율 (%)	참여 인원 (명)	총 인원 대비 비율 (%)
가정주부	4	0.11	1.89	208	5.92	98.11	212	6.03
경영/관리직	78	2.22	53.79	67	1.91	46.21	145	4.13
군인	22	0.63	100.00	0	0	0	22	0.63
기능원/관련 종사자	16	0.46	80.00	4	0.11	20.00	20	0.57
기술직 종사자	132	3.76	77.19	39	1.11	22.81	171	4.87
농림/임업/어업 종사자	6	0.17	75.00	2	0.1	25.00	8	0.23
단순 노무 종사자	30	0.85	63.83	17	0.48	36.17	47	1.34
무직/취업 준비생	76	2.16	19.49	314	8.94	80.51	390	11.10
사무 종사자	314	8.94	39.95	472	13.43	60.05	786	22.37
서비스 종사자	95	2.70	45.67	113	3.22	54.33	208	5.92
전문가/관련 종사자	91	2.59	35.14	168	4.78	64.86	259	7.37
판매/영업 종사자	37	1.05	53.62	32	0.91	46.38	69	1.96
학생	209	5.95	22.79	708	20.15	77.21	917	26.10
기타	74	2.11	28.46	186	5.29	71.54	260	7.40
합계	1,184	33.69	33.69	2,330	66.31	66.31	3,514	100

〈표 3-13〉 온라인 대화 참여자의 직업 구성

온라인 대화 참여자의 직업 구성은 〈표 3-13〉과 같다.⁴⁰⁾

온라인 대화 참여자의 직업 구성에서는 학생이 26.10%로 가장 높은 비율을 차지한다. 참여자 연령대에서 가장 높은 비중으로 나타나는 20대 대부분이 학생이기 때문에 참여율이 높게 나타나는 것으로 보인다.

사무 종사자의 참여율은 22.37%로 나타났으며, 전문가/관련 종사자의 참여율은 7.37%로 나타났다. 사무실 내에서 컴퓨터를 많이 활용하는 직종의 참여 비율이 단순 노무,

40) 직업의 세부 분류는 직업의 세부 분류는 한국표준직업분류(통계청, 2017)의 분류 항목을 일부 수정한 것이다.

판매/영업, 서비스 직종에 비해 상대적으로 참여율이 높다. 취업 준비생도 11.10%의 참여율을 보인다. 취업 준비생 다수가 참여율이 높은 2, 30대이고, 다른 직업에 비해 시간을 유동적으로 활용할 수 있기 때문에 참여율이 높게 나타나는 것으로 볼 수 있다.

직업 구성 비율을 경제 활동 인원과 비경제 활동 인원⁴¹⁾으로 구분해서 참여율을 살펴보면 <표 3-14>와 같다.

상대적으로 시간이 많지 않은 경제 활동 참여 인원의 비율과 상대적으로 시간 활용이 용이한 비경제 활동 인원의 참여율은 49.37%와 43.23%로 큰 차이를 보이지 않는다. 다만 남성 비경제 활동 인원의 참여율은 전체에서 가장 낮은 8.22%로 나타난 반면, 여성 비경제 활동 인원의 참여율은 전체에서 가장 높은 35%로 나타난 점은 주목할 만하다.

구분	남성			여성			합계	
	참여 인원 (명)	총 인원 대비 비율 (%)	범주 내 비율 (%)	참여 인원 (명)	총 인원 대비 비율 (%)	범주 내 비율 (%)	참여 인원 (명)	총 인원 대비 비율 (%)
경제 활동 인원	821	23.36	47.32	914	26.01	52.68	1,735	49.37
비경제 활동 인원	289	8.22	19.03	1,230	35.00	80.97	1,519	43.23
기타	74	2.11	28.46	186	5.29	71.54	260	7.40
합계	1,184	33.69	33.69	2,330	66.31	66.31	3,514	100

<표 3-14> 온라인 대화 참여자의 경제 활동 유무에 따른 구성

2.3. 지역

입말에서 지역은 지역 방언 구사와 관련이 되며, 언어 형태를 변별하는 주요 요인이다. 온라인 대화는 입말과 달리 글말로 표현이 되기 때문에 형태 변별이 입말에 비해 서 두드러지지 않는지만, 같은 지역 출신의 상대방과의 대화에서는 어휘나 어미 사용에 영향을 미치기도 한다.

이 사업에서는 대화 참여자의 현재 거주지뿐만 아니라, 언어 사용에 영향을 미칠 수 있는 출신지와 주 성장지까지 상세 구분했다.

한국의 지역별 인구 구성⁴²⁾과 온라인 대화 참여자의 지역별 구성을 비교하면 <표 3-15>와 같다.

41) 기타 항목을 제외하고 소속된 직장이 없고 정기적으로 급여를 받지 않는 학생, 무직/취업 준비, 가정주부를 비경제 활동 인원으로 구분하고, 그 외는 경제 활동 인원으로 구분하였다.

42) 통계청의 2020년 기준 인구 총조사를 근거로 했다.

구분		총 인구 실제		출생지		주 성장지		현 거주지	
		인원 (명)	비율 (%)	인원 (명)	비율 (%)	인원 (명)	비율 (%)	인원 (명)	비율 (%)
수도권	서울	9,586,195	18.50	1,056	30.05	952	27.09	1,054	29.99
	경기	13,511,676	26.07	602	17.13	763	21.71	919	26.15
	인천	2,945,454	5.68	166	4.72	179	5.09	199	5.66
수도권 소계		26,043,325	50.25	1,824	51.91	1,894	53.90	2,172	61.81
충청권	대전	1,488,435	2.87	118	3.36	130	3.70	125	3.56
	세종	353,933	0.68	7	0.20	11	0.31	30	0.85
	충북	1,632,088	3.15	85	2.42	87	2.48	76	2.16
	충남	2,176,636	4.20	88	2.50	105	2.99	92	2.62
충청권 소계		5,651,092	10.90	298	8.48	333	9.48	323	9.19
경상권	대구	2,410,700	4.65	189	5.38	191	5.44	154	4.38
	경북	2,644,757	5.10	169	4.81	131	3.73	102	2.90
	부산	3,349,016	6.46	271	7.71	239	6.80	226	6.43
	울산	1,135,423	2.19	75	2.13	70	1.99	52	1.48
	경남	3,333,056	6.43	213	6.06	219	6.23	172	4.89
경상권 소계		12,872,952	24.84	917	26.10	850	24.19	706	20.09
전라권	광주	1,477,573	2.85	97	2.76	102	2.90	94	2.68
	전북	1,802,766	3.48	116	3.30	111	3.16	61	1.74
	전남	1,788,807	3.45	124	3.53	104	2.96	67	1.91
전라권 소계		5,069,146	9.78	337	9.59	317	9.02	222	6.32
강원		1,521,763	2.94	109	3.10	88	2.50	72	2.05
제주		670,858	1.29	22	0.63	20	0.57	18	0.51
해외/기타		-	-	7	0.20	12	0.34	1	0.1
합계		51,829,136	100	3,514	100	3,514	100	3,514	100

〈표 3-15〉 온라인 대화 참여자의 지역 구성

대화 참여자의 출신지와 주 성장지 구성 비율은 권역별로 보면 대체로 한국의 지역별 인구 구성 비율과 큰 차이를 보이지 않는다. 그런데 현재 거주지 기준으로 보면 수도권 거주자의 구성 비율이 61.81%로, 한국의 실제 수도권 인구 구성 비율인 50.25%와 10% 이상의 차이를 보인다.

이는 온라인 대화 참여자의 연령 구성이 주로 20대와 30대로 이루어져 있고, 20대와 30대의 경우 대학교 생활과 직장 생활 등의 이유로 수도권에 거주하는 비율이 높기 때문이라 볼 수 있다. 대화 참여자의 수도권 구성 비율이 실제 인구 구성 비율보다 10% 이상 높게 나타난 반면에 충청권을 제외한 경상권과 전라권은 실제 인구 구성 비율보다 3~5% 가량 낮은 것으로 나타났다.

2.4. 기기 및 키보드 유형

온라인 대화에 사용하는 기기 유형과 키보드 유형에 따라 상대적으로 입력이 쉬운 자

모와 어려운 자모가 구분되기도 한다.



[그림 3-1] 키보드의 유형

키보드의 유형에 따라 ‘ㅇㅇ’, ‘ㅋㅋ’, ‘ㅎㅎ’, ‘ㅠㅠ’와 같은 자모를 연속으로 입력하는 형태의 실현 양상이나 사용 빈도에서 차이를 보일 가능성이 있다. 그리고 키보드의 형태에 따라 발생하는 오타의 양상도 다를 수 있다. 즉 온라인 대화에 사용하는 기기와 키보드 유형은 온라인 대화의 언어 형태를 변별하는 요인이다.

먼저 온라인 대화 참여자가 온라인 대화에 주로 사용하는 기기의 구성 비율은 <표 3-16>과 같다.

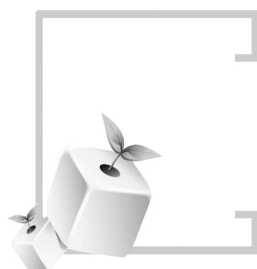
기기 유형	참여자 수	비율(%)
PC(데스크탑/노트북)	323	9.19
스마트폰	3,176	90.38
태블릿(패드)	15	0.43
합계	3,514	100

<표 3-16> 온라인 대화 참여자의 대화 시 주요 사용 기기 구성

온라인 대화 참여자의 90.38%는 스마트폰을 온라인 대화에 사용하고 있는 것으로 나타났다. 온라인 대화 참여자가 사용하는 기기의 키보드의 유형 구성은 <표 3-17>과 같다.

키보드 유형	참여자 수	비율(%)
2벌식(쿼티)	2,350	66.88
나랏글	95	2.70
단모음	86	2.45
천지인	902	25.67
기타	81	2.31
합계	3,514	100

<표 3-17> 온라인 대화 참여자의 키보드 유형 구성



제 4 장

마무리 및 제언



지금까지 2021 온라인 대화 자료 수집 및 정제 사업 수행의 과정과 결과를 살펴보았다.

1장에서는 이 사업의 수행 배경인 온라인 대화 사용과 인공지능 기반 챗봇 사용의 확산을 제시하고, 구축한 결과물을 통해 한국어 사용자의 일상 온라인 대화가 국가 공공 언어 자원으로 활용 가치를 갖도록 하는 본 사업의 목적을 제시했다. 세부적으로 실용도와 활용도가 높은 온라인 대화 말뭉치 구축, 현대 한국어 대화에 최적화된 말뭉치 구축, 법적 문제 발생 우려가 없는 국가 공공재 온라인 대화 말뭉치 구축의 세 가지 사업의 수행 목표를 제시했다.

2장에서는 온라인 대화 말뭉치를 구축하는 절차를 계획 단계, 수집 단계, 가공 단계, 검수 단계로 나누어 제시했다. 계획 단계에서 온라인 대화 제공자의 구성, 온라인 대화의 유형 구성을 설계하는 절차를 제시했다. 수집 단계에서는 자료 수집을 위한 홍보, 화자 정보 수집과 저작권 이용 허락 계약, 대화 실시, 대화 선별의 절차와 방법을 제시했다. 가공 단계에서는 전처리 단계와 인적 가공 단계 말뭉치 생성의 단계로 나누어 최종적으로 온라인 대화 원시 말뭉치가 만들어지는 과정을 제시했다. 그리고 각 단계별 작업 지침과 온라인 대화 말뭉치의 형식 구조를 제시했다. 검수 단계에서는 작업자 간 교차 검수의 과정과 최종 산출물을 검수하는 과정을 형식과 구조를 검수하고 수정하는 작업과 산출물의 내용을 검수하고 수정하는 작업으로 구분해서 제시했다.

3장에서는 온라인 대화 말뭉치 구축 결과를 제시했다. 먼저 수집 파일 기준 4,761개, 대화 수 기준 151,004개, 말차례 수 기준 2,283,178개, 발화 수 기준 4,120,382개로 구성된 말뭉치의 구축 규모를 제시했다. 유형별로는 대화 참여 인원수에 따라 2인 대화와 다자 대화로 나누어 구축한 결과를 제시했고, 수집 방법에 따라 실시간 대화 수집과 기존 대화 수집으로 나누어 구축한 결과를 제시했다. 그리고 수집 매체에 따라 카카오톡 대화 수집과 심심이 채팅 대화 수집으로 나누어 구축 결과를 제시했고, 주제별로 기타 일상 대화와 주제 대화로 나누어 구축한 결과와 주제 대화의 세부 주제별 구축 결과를 제시했다. 대화 참여자 간 관계, 친밀도, 연락 빈도에 따른 구축 결과 또한 제시했다.

다음으로 온라인 대화 참여자의 구성을 제시했다. 전체 3,514명의 대화를 말뭉치로 구축했고, 대화 참여자의 성별과 연령, 직업, 지역, 기기와 키보드 유형에 따른 참여자의 구성을 제시했다.

2019년 국립국어원이 주관한 ‘메신저 대화 자료 수집 및 말뭉치 구축’ 사업을 시작으로 온라인 대화 자료를 수집하고 말뭉치로 구축하는 사업이 국가 차원에서 지속되고 있다. 온라인 대화의 일상적 확산과 인공지능 기술 개발의 필요성 증가라는 시대적 요구에 따라 연구와 개발에 활용할 수 있는 온라인 대화 자료의 필요성도 날로 커지고 있기 때문이다.

이 사업 또한 그러한 필요를 충족하기 위한 사업으로서, 2019년 국립국어원의 메신저 대화 자료 수집, 2020년과 2021년 한국지능정보사회진흥원의 한국어 SNS 데이터 구축 사업, 대화 텍스트 데이터 구축 사업과 동일하게 ‘온라인 대화’를 구축 대상으로 삼고 있다는 점에서 관련성이 있다. 반면에 인공지능의 학습에 즉각적으로 활용하기 위한 ‘학습

데이터’를 구축하는 것이 사업의 주요 목표인 한국지능정보사회진흥원 사업과 다르게 최종 결과물 형태가 ‘원시 말뭉치’라는 점은 이 사업만의 독자적인 특성이다.

사업 계획 단계부터 이런 관련성과 독자성을 고려해 이 사업이 어떤 자리매김을 해야 할지를 고민하고 이를 반영한 결과물을 만들고자 했다. 인공 지능의 대화 능력이 기술의 수준을 평가하는 척도로 여겨짐에 따라 학습 데이터로서 말뭉치의 가치가 어느 때보다 커지고 있는 시기이다. 하지만 말뭉치는 인공 지능의 학습 데이터이기 이전에 언어 속에 반영된 사회와 문화를 관찰할 수 있는 자료이자, 언어의 형태적 특성과 사용양상과 변화까지도 관찰할 수 있는 자료로서도 가치를 지닌다.

이러한 관점에서 공공에 공개가 가능한 범위에서 수집한 온라인 대화의 형태를 그대로 보존하고 있다는 점은 이 사업만의 특성 가운데 하나이다. 인공 지능의 언어 모델 개발을 목표로 삼는 유사 사업은 인공 지능의 학습에 즉각적으로 활용하기 위해 원문의 형태를 유지하기보다는 학습에 용이한 형태로 가공이 이루어진다. 그 과정에서 온라인 대화만이 가지는 독자적인 특성이 희석되거나 사라져 버리기도 한다.

반면 이 사업의 결과물은 공공 언어 자원으로 활용하기 위해 개인정보를 비식별화하거나 비윤리 표현을 정제하는 등 필요한 부분에 한해서만 최소한의 정제 작업을 수행했기 때문에 온라인 대화 자료가 갖는 고유한 특성이 드러난다. 그렇기 때문에 어절 단위로 끊는 발화 형태, 자모만을 이용한 표현 형태, 띄어쓰기 없는 표현 형태, 표준어가 아닌 온라인 공간에서만 사용되는 어휘까지 폭넓게 관찰이 가능하다. 따라서 온라인 대화의 언어 사용 특성을 관찰하고 연구하고자 하는 사람이라면 관심을 가져야 할 자료인 동시에, 연구 과제로 삼을 만한 숙제를 던져 주는 자료인 셈이다.

물론 이 말뭉치가 온라인 대화의 특성을 밝히는 데에만 활용되지는 않는다. 이 자료를 통해 온라인 대화의 특성을 밝히는 연구 성과가 쌓이면 자연스러운 온라인 대화체를 구사하는 인공 지능 연구와 개발의 근간이 될 것이다. 그리고 인공 지능 학습을 위한 말뭉치 구축 또한 사업의 중요한 목표이다. 이를 반영하기 위해 통상적인 온라인 대화와 비교해서 상대적으로 주제 맥락이 일관되게 유지되고 표현도 정제된 짧은 길이의 대화도 일정 분량을 구축했다. 대화 모델 등 인공 지능 분야의 연구자나 개발자가 학습 목적에 맞춰 원하는 형태로 자유롭게 변형해서 활용하는 것도 고려했다.

한편, 언어는 시대와 사회의 산물이라는 점을 고려해 구축한 대화 자료에 당대의 사회문화적 상황이 반영되도록 했다. 이를 위해 일반적인 주제어 제시뿐만 아니라, 격주 단위로 해당 기간에 화제가 되고 있는 시사 주제어와 일상 트렌드 주제어를 발굴해서 이를 포함한 대화도 말뭉치에 포함했다.

이 사업이 지니는 한계도 있다. 여기에서 밝힌 사업의 한계는 이 자료를 활용하는 모두가 고민하고 해결을 위한 노력과 제안이 필요하다.

먼저 온라인 대화 자료 구축의 체계를 설계하는 과정에서 기준이 될 만한 온라인 대화에 대한 연구가 부족했다. 자체적으로 분류의 기준이나 분석의 지침 등을 마련하고 주관 기관과 긴밀히 협의했으나, 온라인 대화 자료 구축 기준의 체계성과 타당성에 대

한 검증은 부족했다. 이 자료를 통해 온라인 대화에 대한 연구가 확산된다면 현재보다 더욱 타당하고 체계적인 구축 기준과 지침 수립이 가능할 것으로 기대한다.

특정 성별이나 연령의 대화를 균형 있게 포함할 뿐만 아니라, 대화 참여자 간 다양한 관계나 친밀도를 포함하는 등 말뭉치 구축의 균형을 고려해야 한다는 관점에서도 이 사업의 한계가 있다. 앞서 말뭉치 구축 결과를 통해서도 살펴보았듯이, 대화 참여자 모두로부터 대화 제공에 대한 동의가 이루어져야 하는 사업 특성상 주로 친밀도가 높은 관계와 대등한 관계의 대화가 집중적으로 수집이 이루어졌다. 반면에 공적인 관계와 상하 위계가 있는 관계의 대화는 상대적으로 부족하다. 그리고 성별과 연령에서도 남성 참여자, 10대와 40대 이상 참여자의 비중이 높아져야 하며, 이들의 참여를 유도할 수 있는 현실적인 방안 마련이 필요하다.

앞서 언급했듯이 비윤리적 표현의 정제 기준 수립과 적용에도 한계가 있다. 인공 지능의 언어 사용 윤리가 2021년도를 기점으로 중요한 문제로 부상하고 있고, 다양한 논의와 연구가 진행 중이지만 말뭉치 언어의 비윤리성을 판단하고 정제하기 위해 합의된 기준은 아직까지 없다. 이 사업 또한 인공 지능의 언어 사용 윤리 문제에 대한 합의된 기준이 없는 상황에서 진행되었다. 사업단 내부에서 짧은 기간에 수립한 비윤리적 표현의 선별 기준에는 한계가 있을 것이다.⁴³⁾ 사업단 내부적으로 일관된 기준을 갖추고자 노력을 기울였음에도 불구하고, 비윤리 표현에 대한 판단과 지침 적용에도 개인 주관에 따른 차이가 나타날 수밖에 없다는 점을 미리 언급한다.⁴⁴⁾

마지막으로 온라인 대화 말뭉치를 열람하고 활용하는 모두에게 당부할 것이 있다. 첫째는 개인정보 보호 주체로서 연구자의 책임이다. 이 사업을 통해 구축된 말뭉치는 개인의 지극히 사적인 대화이다. 비록 개인의 신원을 알아볼 수 없는 형태로 비식별화가 이루어졌다고 해도 이 자료를 열람하고 활용하는 모두는 개인의 민감한 사생활을 취급하는 주체임을 잊지 않아야 한다. 자료를 활용하는 모두가 개인정보 보호에 대한 책임 의식을 지니고 자료를 조심스럽게 관리하기를 바란다.

43) 국립국어원에서는 말뭉치 언어의 비윤리적 표현을 선별하는 기준을 마련하기 위한 별도의 사업을 진행하고 있다. 비윤리적 표현을 선별하는 기준 수립 자체가 별도의 사업으로 운영될 만큼 시간과 인력의 투입이 필요한 작업이다.

44) 최종 결과물에 미쳐 정제가 되지 않은 표현이 나타날 가능성도 없지 않다. 2019년 국립국어원에서 구축한 메신저 대화 말뭉치의 경우에도 2020년의 ‘이루다’ 사태의 여파와 함께 비윤리적 표현 포함 등의 문제로 공개가 중단된 사례가 있다. 개인정보 노출의 문제나 혐오, 차별 표현 등을 전면 재검토하기 위한 것이지만, 국가 단위에서 최초로 구축된 대규모 온라인 대화 자료의 공개가 지연됨에 따라 이 자료를 활용한 다양한 분야의 연구도 지연된 것이다.

온라인 공간의 언어는 기존 언어 체계로 완벽히 규정하기 어렵다. 온라인 공간에서 새롭게 만들어지는 어휘도 있고, 형태만 동일하고 기존 어휘와는 다른 의미로 사용되기도 한다. 온라인 공간의 언어를 이해하는 체계와 기준을 가지기 위해서는 온라인 대화의 실제 사용 양상을 있는 그대로 보여주는 자료도 필요하다. 소수의 연구자나 말뭉치 자료를 정제하는 소수의 작업자가 내리는 판단보다는 있는 그대로를 다수의 사람들에게 공개하고 다양한 관점에서 연구와 논의하는 과정을 통해 체계와 기준을 수립해 나가는 것이 바람직하다.

다음으로 사업에 대한 제언을 단 하나의 표현으로 정리한다. ‘구슬이 서 말이라도 꿰어야 보배’이다. 이 사업의 결과물로 공개되는 자료는 결과물인 동시에 재료이다. 주어진 재료를 가공해서 저마다의 결과물을 만들어 가는 것은 이 자료를 열람하고 활용하는 모두의 몫이다.

참고문헌

- 국립국어원(2019), 메신저 대화 자료 수집 및 말뭉치 구축, 국립국어원.
- 서상규, 안의정, 봉미경, 최정도, 박종후, 백해파, 송재영, 김선희(2013). 한국어 구어 말뭉치 연구. 한국문화사.
- 통계청(2017). 한국표준직업분류. 통계청.
- 통계청(2020). 『인구총조사』 통계정보보고서. 통계청.
- 한국지능정보사회진흥원(2021), 2020 인터넷이용실태조사, 한국지능정보사회진흥원.
- DMC MEDIA(2017). 2017 모바일 메신저 앱 이용 행태. DMC리포트.
- DMC MEDIA(2019). 2019 모바일 메신저 앱 이용 행태. DMC리포트.

<Abstract>

Collection and refinement of online chat data 2021

The purpose of this project was to build a corpus from Korean online chat data as a national public language resource, so that it can be widely used for research and development in various fields such as linguistics, social culture, natural language processing, big data, and artificial intelligence industry.

We first designed a sample of speakers in order to build a representative corpus showing the characteristics of online chat conversations. We also categorized chat topics and types of dialogues to add variety in the corpus.

Topics are largely divided into ‘thematic dialogue’, ‘daily chat’, ‘current affairs and trends’. The current trending topics were selected biweekly to reflect the times in 2021.

The types of conversation are classified according to the interaction between speakers, the devices and media used, and the collection methods.

The interaction between speakers are sub-categorized by the number of speakers in each dialogue, their relationship, intimacy, and contact frequency.

We labeled the device and keyboard used for the conversation, and also the Instant Messaging apps which are mostly KakaoTalk or Simsimi.

Collection methods are labeled as ‘real-time conversation’ or ‘existing past conversation’.

To make the corpus available to the public, all of the participants agreed to collect their privacy information and gave permission to use their conversations.

We then de-identified privacy information from the data and we refined unethical expressions such as hate speech or discrimination.

Hereby, individual researchers, institutions or industries are allowed to use this corpus without legal restrictions.

After collection of raw data, topic labeling, dialogue segmentation, refinement of privacy information and unethical expressions, the data was finally converted

into JSON format with predefined structures.

In the end, we collected 4,761 conversation files from 3,514 speakers. The corpus consists of 151,004 separated dialogues, 2,283,178 speech turns, and 4,120,382 utterances.

This raw corpus reflects the unique characteristics of online chat as it is. It may be used to study linguistic features of online instant messaging.

A part of this corpus consists of relatively short and refined dialogues, so that artificial intelligence researchers or developers can freely transform and use them for dialogue models or machine learning.

Keywords: online chat, instant messaging, KakaoTalk, conversation, raw corpus, natural language processing

Project Director: Park Ilseop(mediaCORPUS Inc.)

<기획·연구>

국립국어원 이승재 언어정보과장

국립국어원 유희정 학예연구사

국립국어원 한송이 연구원

<사업 참여자>

사업 책임자 박일섭((주)미디어 코퍼스)

사업 참여자 신지영((주)미디어 코퍼스), 남서정((주)미디어 코퍼스)
조재원((주)미디어 코퍼스), 안 윤((주)미디어 코퍼스)
현재홍((주)미디어 코퍼스), 이수경((주)미디어 코퍼스)
양성민((주)다이얼로그디자인에이전시)
이태강((주)다이얼로그디자인에이전시)
양리아((주)다이얼로그디자인에이전시)
이나리((주)다이얼로그디자인에이전시)
이광진((주)다이얼로그디자인에이전시)
임현승(심심이(주)), 최정희(심심이(주))
홍미미(심심이(주)), 조재훈(심심이(주))
김민재(심심이(주))

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2021년 11월 13일

발행일: 2021년 11월 13일

인 쇄: 테라인쇄소

※ “이 책은 국립국어원의 용역비로 수행한 ‘2021년 온라인 대화 자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.”