국립국어원 2019-01-07

발 간 등 록 번 호
11-1371028-000764-01

메신저 대화 자료 수집 및 말뭉치 구축

사업 책임자
박 일 섭

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '메신저 대화 자료 수집 및 말뭉치 구축'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 7월 19일 ~ 2019년 12월 19일

2019년 12월 19일

사업 책임자: 박일섭(주식회사 미디어코퍼스)

사업 수행기관	주식회사 미디어코퍼스 주식회사 다이얼로그디자인에이전시 ㈜그립
사업 책임자	박일섭
사업 참여자	정연규, 남서정, 신지영, 이서희, 이수경, 양성민, 오가영, 이태강, 문진숙, 양한주, 송민지, 이광진, 이나리, 양리아, 이규수, 안택헌, 임명식, 최승욱, 최문성, 윤지수, 손대웅

〈사업 수행자〉	주식회사 미디어코퍼스 주식회사 다이얼로그디자인에이전시 ㈜그립

사업 책임자	박일섭(주식회사 미디어코퍼스)
사업 참여자	정연규((주)그립)
	남서정(주식회사 미디어코퍼스)
	신지영(주식회사 미디어코퍼스)
	이서희(주식회사 미디어코퍼스)
	이수경(주식회사 미디어코퍼스)
	양성민(주식회사 다이얼로그디자인에이전시)
	오가영(주식회사 다이얼로그디자인에이전시)
	이태강(주식회사 다이얼로그디자인에이전시)
	문진숙(주식회사 다이얼로그디자인에이전시)
	양한주(주식회사 다이얼로그디자인에이전시)
	송민지(주식회사 다이얼로그디자인에이전시)
	이광진(주식회사 다이얼로그디자인에이전시)
	이나리(주식회사 다이얼로그디자인에이전시)
	양리아(주식회사 다이얼로그디자인에이전시)
	이규수(주식회사 다이얼로그디자인에이전시)
	안택헌((주)그립)
	임명식((주)그립)
	최승욱((주)그립)
	최문성((주)그립)
	윤지수((주)그립)
	손대웅((주)그립)

<국문 초록>

메신저 대화 자료 수집 및 말뭉치 구축

본 사업은 자연어 처리와 빅 데이터(big data) 산업 등 인공 지능 관련 연구·개발과 관련 산업계에서 국가 공공재로 활용하기 위한 용도로 메신저 대화 자료를 수집하고 이를 원시 말뭉치로 구축하는 것을 목적으로 한다.

메신저 대화의 특성을 대표성 있게 반영하는 말뭉치를 구축하기 위하여 메신저 사용 통계를 반영하고자 했고, 현실적인 대화 수집 가능성을 고려하여 표본을 설계하였다. 그리고 다양한 유형의 메신저 대화를 말뭉치에 포함시키기 위하여 참여자 간 상호 작용 양상, 사용 매체의 특성, 내용 주제, 수집 방법 등 메신저 대화 양상에 영향을 미치는 요인을 고려한 메신저 대화 유형 분류 기준을 마련하였다. 특히 낯선 관계의 대화와 주제가 통제된 대화도 포함하기 위하여 대화를 추출하여 제공하는 방식 이외에 수집 봇(bot)을 통해 대화를 채록하는 방법도 활용하였다.

모든 대화 참여자로부터 개인 정보 제공 이용 동의를 얻은 후 자료를 수집하였고, 이름과 전화번호, 계좌 번호를 포함한 여러 유형의 개인 식별 정보는 지침에 따라 비식별화 처리를 하였다. 수집한 자료는 대화 참여자 전원을 대상으로 저작권 이용 허락 계약을 체결하여 연구자와 기관, 산업체에서 법적인 제한 없이 안정적으로 활용할 수 있도록 하였다.

개인 정보를 포함하여 사진과 동영상, 이모티콘 등의 멀티미디어 요소와 무료 통화, 송금, 선물 보내기 등의 특수 기능과 같이 메신저 대화에 포함된 발화 이외의 요소는 '걸러내기(filtering)'나 '바꾸기(replacing)' 등이 용이하도록 지침에 따라 별도의 태그를 부착하여 SJML과 JSON 두 가지 형식으로 말뭉치를 구축하였다.

최종적으로 10,083명의 화자로 구성된 7,395개의 파일을 수집하여 메타 정보가 부착된 원시 말뭉치를 구축하였다. 구축한 원시 말뭉치는 14,591,826개의 발화와 7,122,919회의 말차례 교체로 이루어진 712,291개의 대화로 구성되어 있다.

구축한 자료는 메신저 대화의 특성에 맞게 자료를 정규화하고 전처리하는 기술적인 방법론 마련과 실제 메신저 대화를 전처리하고 분석하는 기술 개발을 통해 딥 러닝(deep learning) 등의 인공 지능 학습용으로 활용할 수 있다. 그리고 메신저 언어문화를 연구하는 기초 자료로서 연구의 이론적 타당성과 체계를 정립하는 근거로 활용할 수 있다.

주요어: 모바일 메신저, 카카오톡, 대화, 원시 말뭉치, 자연어 처리, 딥 러닝(deep learning), 빅 데이터(big data)

차 례

제 1 장 사업 개요

1. 사업의 목적 ····· 3
2. 사업의 범위 ····· 4

제 2 장 메신저 대화 말뭉치 구축 절차

1. 메신저 대화 말뭉치의 설계 ····· 7
1.1. 메신저 대화 제공자 구성 ····· 7
1.2. 메신저 대화의 유형 구성 ····· 8
2. 메신저 대화 자료 수집 ····· 11
2.1. 메신저 대화 자료 수집 절차 ····· 11
2.2. 홍보 및 참여자 모집 ····· 21
2.3. 자료 선별 ····· 29
3. 정제 및 마크업 ····· 31
3.1. 비식별화 및 특수 메시지 처리 ····· 31
3.2. 추출 형식 표준화 ····· 37
3.3. 파일 형식 및 마크업 지침 ····· 39
4. 검수 ····· 50
4.1. 비식별화 항목 검수 ····· 50
4.2. 말뭉치 형식 검수 ····· 50

제 3 장 메신저 대화 말뭉치 구축 결과

1. 메신저 대화 말뭉치의 구성 ····· 55
1.1. 구축 규모 ····· 55
1.2. 유형별 구성 ····· 55
1.3. 메신저 대화 말뭉치의 참여자 구성 ····· 59
1.4. 참여자 간 관계 구성 ····· 66

제 4 장 마무리 및 제언

참고문헌 ····· 75

표 차례

<표 1> 사업의 세부 목표 ····· 4
<표 2> 국립국어원(2007a)과 서상규 외(2013)의 텍스트 유형 분류 기준 ····· 9
<표 3> 메신저 대화 말뭉치의 유형 분류 기준 ····· 10
<표 4> 이모티콘의 의미 입력 기준 표 ····· 13
<표 5> 일반 대화 수집과 수집 봇 수집의 비교 ····· 15
<표 6> 메타 정보 수집 항목 ····· 17
<표 7> 메신저 대화의 주제 분류 항목 ····· 18
<표 8> 참여자 간 관계 분류 ····· 19
<표 9> 개인 정보 수집·이용 동의 주요 항목 ····· 19
<표 10> 저작권 이용 허락 계약의 주요 내용 ····· 20
<표 11> 공식 홍보 창구 운영 ····· 21
<표 12> 대화 제공 인원 추이 ····· 29
<표 13> 메신저 대화 비식별화 기본 지침 ····· 32
<표 14> 비식별화 1차 수정 지침 ····· 33
<표 15> 비식별화 2차 수정 지침 ····· 34
<표 16> 메신저 대화 특수 메시지 항목 ····· 35
<표 17> 카카오톡의 텍스트 추출 형식 비교 ····· 38
<표 18> 원시 말뭉치의 날짜, 시간, 발화자 표준화 결과 ····· 39
<표 19> 메신저 말뭉치 파일명 부여 방식 ····· 39
<표 20> 메신저 대화 원문과 말뭉치 파일명 작성 예시 ····· 40
<표 21> 헤더 항목의 마크업 지침 ····· 42
<표 22> 원시 말뭉치 본문 마크업 지침 ····· 45
<표 23> 비식별화 항목의 원문 표기와 마크업 기호 비교 ····· 47
<표 24> 특수 메시지 항목의 텍스트 표기와 마크업 기호 비교 ····· 48
<표 25> 메신저 대화 원시 말뭉치의 분량 산정 ····· 55
<표 26> 대화 참여 인원별 구축 분량 및 비율 ····· 56
<표 27> 수집 방법별 구축 분량 및 비율 ····· 57
<표 28> 대화 주제별 구축 분량 및 비율 ····· 58
<표 29> 주제 대화 세부 주제별 구축 분량 및 비율 ····· 58
<표 30> 메신저 대화 말뭉치 참여자의 성별 및 연령 구성 ····· 60
<표 31> 메신저 대화 말뭉치 참여자의 직업 구성 ····· 62

표 차례

<표 32> 경제 활동 유무에 따른 메신저 대화 말뭉치 참여자의 구성 비율 ····· 62
<표 33> 대한민국 지역별 인구 구성과 메신저 대화 말뭉치 화자의 지역별 구성 비교 ····· 63
<표 34> 메신저 대화 말뭉치 화자의 메신저 대화 시 주요 사용 기기 ····· 64
<표 35> 메신저 대화 말뭉치 화자가 사용하는 키보드 유형 구성 ····· 65
<표 36> 메신저 대화 말뭉치의 참여자 간 관계와 구축 분량 ····· 66
<표 37> 메신저 대화 말뭉치의 참여자 간 친밀도와 구축 분량 ····· 67
<표 38> 대화 참여자 간 관계에 따른 친밀도 구성(참여자 쌍 기준) ····· 68
<표 39> 메신저 대화 말뭉치의 참여자 간 연락 빈도와 구축 분량 ····· 70

그림 차례

[그림 1] 메신저 대화 자료 수집 절차 ····· 11
[그림 2] 이모티콘과 사진의 텍스트 추출 예시 ····· 12
[그림 3] 수집 사이트 참여자 등록 과정 ····· 16
[그림 4] 저작권 계약서 정보 입력과 서명 예시 ····· 20
[그림 5] 카카오톡 채널 및 공식 블로그 ····· 22
[그림 6] 국립국어원 참여 협조 요청 공문 및 국립국어원 누리집 공지 ····· 23
[그림 7] 사업 초기 홍보 게시물 예시 ····· 24
[그림 8] 다자 대화 이벤트 홍보 게시물 예시 ····· 25
[그림 9] 1:1 대화 및 다자 대화자 모집 게시물 예시 ····· 26
[그림 10] 낯선 관계 대화 이벤트 대화자 모집 게시물 예시 ····· 27
[그림 11] 주제 특화 1:1 대화 이벤트 홍보 및 모집 게시물 예시 ····· 28
[그림 12] 수집 봇 대화 수정 전 예시 ····· 30
[그림 13] 수집 봇 대화 수정 후 예시 ····· 30
[그림 14] 카카오톡 안드로이드 운영 체제 추출 형식 예시 ····· 37
[그림 15] 카카오톡 PC 운영 체제 추출 형식 예시 ····· 37
[그림 16] 카카오톡 iOS 운영 체제 추출 형식 예시 ····· 38
[그림 17] 메신저 대화 말뭉치 SJML과 JSON 형식의 기본 구조 ····· 41
[그림 18] SJML 형식의 헤더 예시 ····· 43
[그림 19] JSON 형식의 헤더 예시 ····· 44
[그림 20] 수집 원문 예시 ····· 45
[그림 21] SJML 형식 본문 마크업 예시 ····· 45
[그림 22] JSON 형식 본문 마크업 예시 ····· 46
[그림 23] 수집 원문 비식별화 예시 ····· 47
[그림 24] SJML 형식 비식별화 마크업 예시 ····· 47
[그림 25] 수집 원문 특수 메시지 예시 ····· 49
[그림 26] SJML 형식 특수 메시지 항목 마크업 예시 ····· 50
[그림 27] 마크업 규칙 설정 오류로 인한 마크업 형식 오류 예시 ····· 51
[그림 28] 작업자 오류로 인한 마크업 형식 오류 예시 ····· 51
[그림 29] 메신저 대화 말뭉치 화자가 사용하는 키보드의 유형 ····· 65

제 1 장

사업 개요

1. 사업의 목적

본 사업의 목적은 자연어 처리와 빅 데이터(big data) 산업 분야 등의 인공 지능 관련 연구·개발과 산업 분야에 국가 공공재로 활용하기 위한 용도로 메신저 대화 자료를 수집하고 이를 원시 말뭉치로 구축하는 것이다.

디지털 기술 발전과 스마트폰 사용 확산으로 카카오톡, 페이스북 메신저, 라인 등의 메신저가 세대와 영역을 아우르는 일상 의사소통 수단으로 자리 잡았다[1]. 메신저를 이용한 채팅이 일상적이고 보편적인 의사소통 수단의 하나로 자리 잡음에 따라 산업계와 공공 부문에서도 주문, 예약, 결제를 비롯하여 민원 처리, 상담 등의 서비스 혁신에 인공 지능 기반 메신저 도입을 추진하고 있다. 더 나아가 산업계에서는 인공 지능 음성 비서와 챗봇(chatbot) 등의 인공 지능 기술을 적용한 대화 시스템이 웹(web)과 앱(app)을 대신하는 차세대 인터페이스, 차세대 플랫폼이 될 것이라 전망한다.

또한 메신저 사용이 일상적인 의사소통의 한 부분을 차지함에 따라 메신저 대화는 언어 형식에서부터 어휘, 대화의 구조, 의사소통 양상에 이르기까지 우리말 의사소통의 한 부분을 관찰할 수 있는 연구 자료라는 가치도 지니고 있다.

이러한 인공 지능 기반 대화 시스템의 연구·개발과 실질적인 활용, 그리고 메신저 의사소통에 대한 연구를 위해 메신저 언어 특성을 반영하면서 딥 러닝(deep learning) 등의 기계 학습 자료로 활용할 수 있는 고품질의 대규모 메신저 대화 말뭉치 필요성이 증대되고 있다.

그러한 필요성에도 불구하고 현재 딥 러닝 등의 기계 학습을 비롯하여 연구 자료로 실제로 활용할 수 있는 대규모의 메신저 말뭉치를 찾아보기는 힘든 실정이다. 또한 개별 연구자나 기관이 독자적으로 인공 지능 학습에 필요한 메신저 대화 자료를 수집하고 말뭉치를 구축하는 데에는 수집 규모와 구성, 개인 정보 보호 규정 준수, 저작권 이용 허락 등의 법적 문제 등 여러 제한이 따른다.

이에 본 사업에서는 메신저 대화 특성을 대표성·균형성 있게 반영하는 연구 자료이자, 인공 지능 기반 대화 시스템 기술에 최적화된 대규모의 고품질 메신저 대화 말뭉치를 구축한다. 또한 민간·공공 영역에서 제한 없이 활용할 수 있도록 개인 정보 및 저작권 이용 등의 법적 문제를 해결하여, 국가 공공재로 제한 없이 활용할 수 있는 실용적 메신저 말뭉치를 구축한다.

---

1) 한국인터넷진흥원(2018)에 따르면 국내 만 6세 이상 인터넷 이용자 중 95.9%가 메신저를 사용한다. 그리고 이들은 메신저를 통해 '대화(100%)', '사진, 동영상, 일정, 업무용 파일 공유(76%)' 기능을 주로 사용한다.

2. 사업의 범위

본 사업의 목적을 달성하기 위하여 본 사업의 범위를 크게 자료 수집, 원시 말뭉치 구축, 품질 관리, 사업 운영의 네 분야로 나눈다.

자료 수집은 대규모로 자료를 수집하기 위한 사업 내용 홍보, 대화 제공자 모집을 위한 홍보 및 참여자 모집 광고, 실제 자료 수집, 자료를 공공 데이터로 구축하기 위한 전제 조건이 되는 개인 정보 수집 이용 동의와 저작권 이용 허락 계약을 포함한다.

원시 말뭉치 구축은 수집한 자료를 분류하고 분석하여 메타 정보의 표준화, 수집 데이터 형식의 표준화, 태깅 형식의 표준화를 진행하여 사업 목적에 맞는 형태로 자료를 정제하고 가공하는 작업을 포함한다.

품질 관리는 메신저 대화의 특성과 말뭉치의 활용성을 고려하여 말뭉치를 기획하고 설계하며, 설계 내용을 기본으로 한 지침을 수립한다. 또한 작업 인력에 대한 지침 교육과 산출물 검수 과정을 포함한다.

사업 운영은 지침 수립 등을 위한 주관 기관과의 협의, 사업 내용과 산출물에 대한 보안 관리, 사업의 진행 상황 관리 등을 포함한다.

사업 범위를 기준으로 설정한 본 사업의 세부 목표는 아래와 같다.

항목	내용
자료 수집	• 대화 참여자 10,000명으로부터 50만 대화 이상 수집 • 3인 이상 다자 대화는 5만 대화 이상 수집 • 저작권 이용 허락을 포함하여 산업적 이용에 제한이 없는 수준의 개인 정보 이용 동의 계약[2] 동시 체결 • 인공 지능 학습에 필요한 어절 수 고려
원시 말뭉치 구축	• 개인 정보의 3단계 비식별화 • 메신저 대화의 다양한 특성과 기능을 반영한 마크업 • 대규모의 말뭉치 구축을 위한 자동화
품질 관리	• 인구 통계와 메신저 사용자 통계에 근거하여 성별, 연령별 대화 제공자 모집 할당 기준 수립 • 인공 지능 대화 시스템 개발과 메신저 언어문화 연구에 유용한 메타 정보 항목 추가 설계 • 메신저 대화 특성을 고려한 마크업(markup) 표준화 지침 수립

<표 1> 사업의 세부 목표

2) 법률 검토 결과에 따르면, 개인 정보의 범위가 넓기 때문에 비식별화 이후에도 개인 정보 이용 동의가 없을 경우 법적 문제 발생 우려가 있어 개인 정보 이용 동의가 필요하다.

제 2 장

메신저 대화 말뭉치 구축 절차

1. 메신저 대화 말뭉치의 설계

말뭉치가 활용도 높은 자료로서의 가치를 지니기 위해서는 활용 목적을 고려한 수집 대상의 선정과 수집 방법의 설계가 필요하며, 목적에 맞게 자료를 가공하고 관리하기 위한 지침을 마련해야 한다. 그리고 말뭉치 구축 과정에서도 사전 설계와 지침에 따라 말뭉치 구축이 진행되고 있는가를 지속적으로 점검하면서 지침을 수정하고 보완해 나가야 한다.

먼저 메신저 대화 말뭉치를 구축하기 위해서는 메신저 대화의 특성을 고려해야 한다. 메신저 대화는 문어(文語)의 형식이지만, 담화의 구조나 어휘 사용을 비롯하여 발화의 진행 양상 등이 면 대 면 대화와 유사하며[3], 비교적 구어(口語)의 특성이 두드러진다. 이를 고려하면 메신저 대화 말뭉치의 설계에 국내의 기존 구어 말뭉치 사업을 참고할 수 있다.

한편, 메신저 대화는 ICT 기기 기반의 의사소통 매체라는 특성이 있다. 문자나 이모티콘, 사진, 동영상 등의 시각화된 요소를 통해 더욱 다양한 방식으로 의미 전달이 이루어지고 띄어쓰기를 비롯한 언어 형태의 실현 양상이 기기의 유형과 특성이라는 변인과 관련이 있다. 또한 선물 보내기와 송금 등과 같이 생활 편의를 돕는 다양한 기능도 빈번하게 활용되며, 이러한 요소가 대화 내용의 일부로 나타난다.

메신저 대화 말뭉치를 구축할 때는 메신저 대화가 갖는 이러한 특성을 반영할 수 있는 설계와 지침 마련이 되어야 한다. 메신저 대화 말뭉치를 구축하기 위한 설계의 기본 방향은 아래와 같다.

1.1. 메신저 대화 제공자 구성

특수한 목적으로 만들어지는 말뭉치를 제외하고 일반적인 말뭉치가 갖춰야 할 요소에는 '대표성(representativeness)'과 '균형성(balancedness)'이 있다. 메신저 대화 말뭉치도 메신저 사용자 집단의 구성과 메신저 사용에서 드러나는 특성을 축소하여 보여주는 대표성과 다양한 유형의 메신저 대화를 포함하는 균형성을 확보해야 한다.

대표성과 균형성을 갖춘 메신저 대화 말뭉치를 구축하기 위해 먼저 말뭉치 구축 목적에 맞게 표본을 구성해야 한다. 모집단으로부터 표본을 추출하는 방법은 크게 인구 통계 정보에 근거한 확률 표본 추출과 편의를 고려한 비확률 표본 추출로 나뉘며, 이를 절충한 방법을 활용하기도 한다[4]. 메신저 대화의 경우도 화자의 성별과 연령, 지역 변

3) Naomi Baron(2008:66)에서는 메신저의 대화 양상을 담화 구조, 사용 어휘의 형태, 발화의 중단 양상 등을 통해 분석하여 면 대 면 대화와 일반적인 글쓰기와의 유사점을 비교하였다. 결론적으로 메신저의 언어는 '문자'보다는 '대화'로 느껴지는 경향이 있다고 분석하였다.

4) 강현화(2017:68)

인에 따라 달라지는 언어 사용 양상을 고려하여 인구 통계를 바탕으로 표본의 구성 비율을 설계하는 것이 바람직하다[5].

그러나 메신저 대화 자료의 실제 수집 가능성도 고려해야 한다. 구어 자료 수집의 경우를 살펴보면, 방송 매체를 제외한 구어 자료는 수집에 여러 제약이 따른다. 대면 인터뷰를 통해 자료를 수집할 경우 대화 제공자 모집에 시간과 거리 등의 물리적인 제약이 발생하고, 온라인을 통해 자료를 수집할 경우에는 인터넷과 매체 사용에 익숙하지 않은 세대나 계층의 참여가 제한된다. 또한 화자로부터 자료 제공에 대한 동의까지 얻어야 한다는 점도 구어 자료 수집의 제약 가운데 하나이다. 기존 구어 말뭉치 구축 사례에서도 나타나듯이 다른 연령에 비해 20대로부터 자료를 수집하기가 상대적으로 용이한데[6], 이는 20대가 앞서 언급한 제약의 영향을 비교적 덜 받기 때문인 것으로 추정할 수 있다. 메신저 대화 제공자의 표본 구성에서도 실제적인 수집 가능성을 고려하되, 20대에 지나치게 편중된 수집이 되지 않도록 적절한 통제 방안을 마련해야 한다.

대화 제공자의 직업과 출신 및 지역 등도 메신저 대화 자료를 구성할 때 고려해야 하는 사항이다. 다만 이러한 요소는 수집 비율을 구체적으로 설정하기보다는 포괄적으로 자료를 수집한 후 자료의 분류와 검색을 위한 정보로 활용한다.

1.2. 메신저 대화의 유형 구성

메신저 대화 말뭉치는 대화의 변인에 따른 다양한 유형의 대화를 포함할 수 있도록 해야 한다. 다양한 유형의 대화가 균형 있게 포함되어 있는가를 평가하기 위해서는 메신저 대화의 유형을 분류하기 위한 기준이 필요하다.

말뭉치의 구성을 위한 텍스트 유형 분류 기준은 국립국어원(2007a)과 서상규 외(2013)를 참고할 수 있다. 이는 <표 2>와 같다.

5) 한국인터넷진흥원(2018)의 성별과 연령에 따른 인구 통계와 연령별 인터넷과 메신저 사용률 통계를 바탕으로 성별과 연령별 표본의 구성 비율을 자체 산정한 결과는 아래와 같다. 아래의 결과에 따르면 10대부터 30대까지 청소년과 청년층에 비해 4, 50대 중장년과 60대 이상 노년층의 비율이 높다.

	10대	20대	30대	40대	50대	60대	70세 이상
남(%)	7	7	8	10	10	6	2
여(%)	7	7	8	10	10	6	2
합계(%)	14	14	16	20	20	12	4

6) 세종 구어 전사 말뭉치는 발화자 분포에서 연령 미상 발화자를 제외할 경우 20대가 50%에 가까운 비중을 차지하며(국립국어원, 2007b:19), 연세 구어 말뭉치는 20대가 연령 미상 발화자를 포함한 전체에서 52.5%의 비중을 차지한다(서상규 외, 2013:100).

구분	분류 기준	분류 내용
국립국어원(2007a)	문어/구어의 분류	문어 / 구어
	매체에 따른 분류	신문 / 잡지 / 책 / 기타 출판물 / 기타 비출판문 등
	내용에 따른 분류	총류 / 신문 / 교육 자료 / 상상적 텍스트 등
	담화 상황에 따른 분류	일상 대화 / 전화 대화 / 토론, 회의 / 연설 등

구분	1차 분류	2차 분류	텍스트 유형	분류 내용
서상규 외(2013)	독백	대면성	공적 독백	대면 강의 / 발표 / 설교 등
	대화		사적 독백	경험담 말하기 / 줄거리 말하기 등
	공적	매체	공적 대화	TV 토론 / 대면 회의 등
	사적		사적 대화	대면 일상대화 / 주제대화/ 수업 대화

<표 2> 국립국어원(2007a)과 서상규 외(2013)의 텍스트 유형 분류 기준

메신저 대화의 유형을 분류하는 기준은 메신저 대화의 양상에 영향을 미치는 다양한 요인을 고려해야 한다.

먼저 메신저 대화 참여자 간 상호 작용의 양상은 대화의 양상에 영향을 미친다. 대화 참여자의 수나 참여자 간 관계, 친밀도 등은 대화 참여자가 사용하는 언어 형식을 비롯하여 참여하는 대화 공동체에 따른 언어 사용 양상이나 의사소통 목적에 따른 대화 내용과도 밀접한 관련이 있다[7].

그리고 메신저 대화는 IT 기기와 '메신저'라는 매체를 매개로 하는 의사소통이기 때문에 매체의 종류나 특성도 유형 분류의 기준으로 고려해야 한다. 가령 친교를 위한 메신저와 업무를 위한 메신저를 구분하거나 대화 상대에 따라 메신저를 구분하여 사용하는 경우가 있기 때문에 사용하는 메신저의 종류에 따라 대화의 공공성이나 대화의 주제 또는 내용이 달라질 수 있다[8]. 또한 사용하는 기기의 키보드 유형도 언어 형식에 영향을 미칠 수 있는 요인이다.

다른 텍스트와 마찬가지로 대화의 주제도 메신저 대화의 유형을 분류할 수 있는 기준이 될 수 있다.

마지막으로 대화가 통제된 상황에서 이루어지는가, 통제가 없이 자연스러운 상황에서 이루어진 대화인가 여부도 텍스트의 유형을 분류하는 기준이 될 수 있다. 즉 대화를 수집한다는 것을 사전에 예고한 후 주제를 부여한 후 이루어지는 대화와 그렇지 않은 대화는 대화의 전개 양상과 언어 형식의 사용 양상이 다르게 나타날 수 있다. 특히 이러한 기준은 메신저 대화를 수집하는 방법[9]과 밀접한 관련이 있다[10].

---

7) 조연정·강정한(2015:108~111)에서는 대화 참여자가 참여하고 있는 대화 공동체에 따라 언어코드를 다르게 사용하고 있음을 언급하고 있다. 가령 20대 과외 교사가 공적 관계인 학부모와 주고받는 메시지는 단일 말풍선을 기준으로 발화의 길이가 대체로 길고 정리가 되어 있고 맞춤법도 대체로 주의를 기울이는 반면, 사적 관계인 친구와 주고받는 대화의 경우는 짧은 호응이나 감정 표현 위주의 짧은 길이의 메시지가 나타난다.

8) DMC MEDIA(2019:7)에 따르면 조사 대상의 82.7%가 두 개 이상의 메신저를 사용하며, 다수의 메신저를 사용하는 주된 이유로 대화 상대에 따라 주로 이용하는 메신저가 다르다는 점, 업무용과 친목용 메신저가 다르다는 점을 들고 있다.

화자 변인 이외에 메신저 대화 양상에 영향을 미치는 다양한 요인과 이에 따른 분류는 <표 3>과 같다.

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 / 낯선 관계의 대화 등
	대화 참여자 간 연락 빈도	거의 매일 연락하는 관계 / 처음 연락하는 관계 등
사용 매체	메신저의 종류	카카오톡 / 라인 / 페이스북 메신저 / 네이트온 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 3> 메신저 대화 말뭉치의 유형 분류 기준

위의 항목은 메신저 대화의 언어 형식과 내용에 영향을 미치는 변인이다. 대화 참여자 간 상호 작용뿐만 아니라 사용 매체, 텍스트 내용, 사전 통제 여부 모두 대화의 형식과 내용에 영향을 끼칠 수 있는 요인이다. 메신저 대화 말뭉치도 변별 요인에 따른 다양한 유형의 대화를 포함하는 것이 바람직하다.

다만 메신저 대화를 수집할 때 유형별로 구체적인 구성 비율을 설정하여 이에 맞춰 대화를 수집하는 것은 현실적으로 불가능하다. 유형 분류의 기준은 최대한의 규모로 대화를 수집한 후 활용 목적에 맞게 자료를 분류하거나 검색하기 위한 기준으로 활용한다.

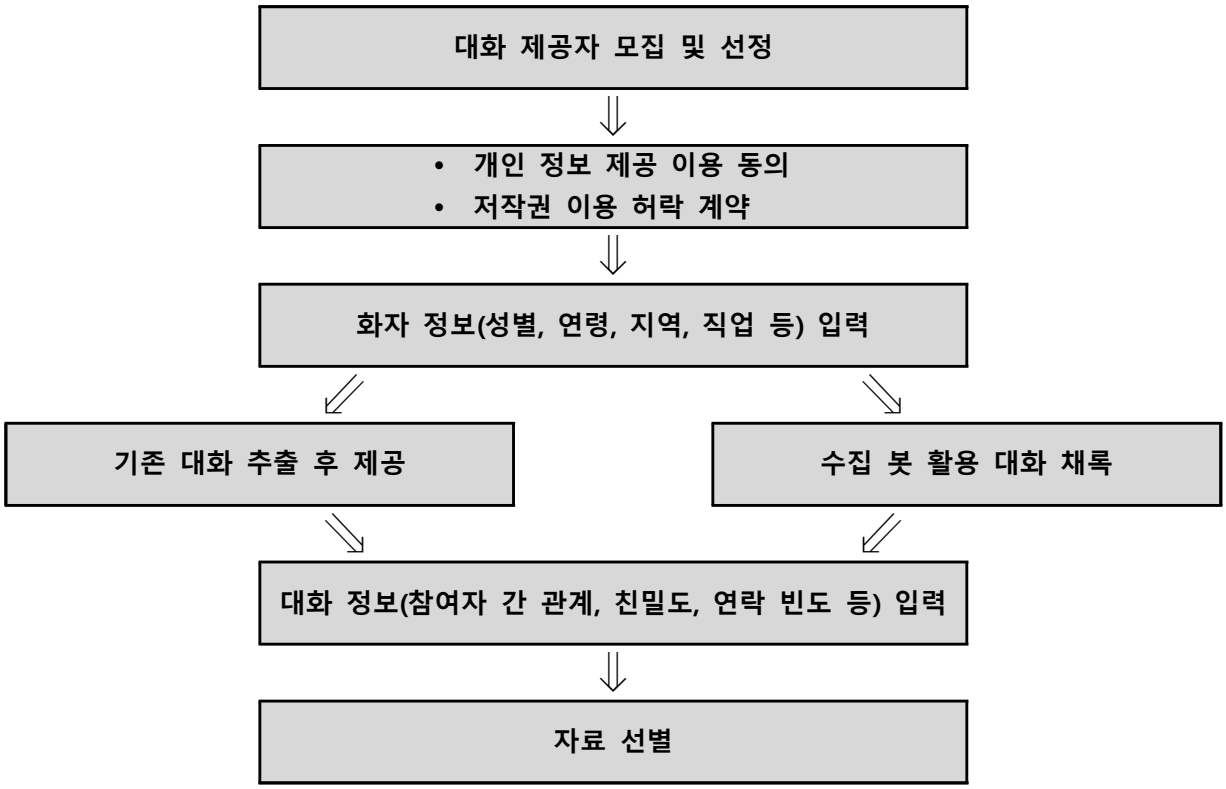
9) 대화를 수집하는 방법에 대해서는 뒤에서 다시 설명한다.

10) 본 사업을 통해서는 친밀도가 높은 관계 위주의 대화 자료가 모일 가능성이 크다. 이는 대화 참여자 모두가 대화 제공에 동의를 하고 동시에 저작권 이용 허락 계약까지 이루어져야 대화 제공이 가능하다는 본 사업의 제약과도 관련이 있다. 한 사람이 대화 제공을 할 의사가 있다고 하더라도 대화 상대방이 그럴 의사가 없다면 대화 제공이 불가능하기 때문이다. 그러나 참여자 간 상호 작용 특성에 따라 언어 형식과 대화의 내용이 달라지는 점을 고려할 때 친밀도가 높은 관계에서 이루어지는 대화에만 편중된 자료 수집은 바람직하지 않다. 조연정·강정한(2015:80~89)에 따르면 친밀한 관계의 카카오톡 대화에서는 맞춤법과 띄어쓰기가 대체로 맞지 않고 비문법적 표현과 줄임 표현, 생략 등이 빈번하게 나타나는 등 비공식적인 언어 형식의 사용이 두드러진다. 이러한 점에 비추어 보면 친밀한 관계에서 이루어지는 대화로만 자료가 구성이 될 경우 언어 형식의 정규화(normalization)나 교정 등의 전처리가 쉽지 않기 때문에 자연어 처리 분야에서 활용도가 떨어질 우려가 있다. 이를 고려하여 낯선 관계에서 이루어지는 대화까지도 포함하여 가급적 다양한 관계의 대화를 수집할 수 있는 방안을 마련해야 한다.

2. 메신저 대화 자료 수집

2.1. 메신저 대화 자료 수집 절차

메신저 대화 자료의 수집 절차는 [그림 1]과 같다.



[그림 1] 메신저 대화 자료 수집 절차

먼저 메신저 대화 자료 수집에 대한 안내와 홍보를 통해 대화 제공자를 모집하고 대상자를 선정하였다. 대화 제공자는 메타 정보의 수집과 관련한 개인 정보 이용 동의와 대화 자료의 저작권 이용 허락 계약을 체결한 후 성별, 연령, 지역, 직업 등의 화자 정보를 입력하였다. 이후 대화 제공자 스스로 기존 대화를 추출하여 제출하거나, 대화방에 수집 봇을 초대한 후 대화를 채록하는 두 가지 방법을 이용하여 대화를 수집하였다[11]. 대화를 제출하거나 채록하면서 대화 제공자 스스로 참여자 간 관계와 친밀도, 연락 빈도 등의 대화 정보를 입력하였다. 이후 작업자가 수집이 완료된 자료 중 말뭉치로 구축하기에 적절한 대화 자료를 선별하였다.

11) 본 사업에서 활용한 두 가지 자료 수집 방법은 뒤에서 다시 설명한다.

2.1.1. 메신저 대화 수집 방법

메신저 대화 수집에 활용할 수 있는 방법은 아래의 두 가지이다.

· 메신저 대화 화면 캡처 후 화면 내용 입력
· 메신저가 제공하는 텍스트 추출 기능 활용

첫 번째 방법은 메신저 대화 내에 포함된 이모티콘이나 사진 등 비언어적 요소가 무엇인가를 확인할 수 있고 필요에 따라 해당 요소의 의미나 정보를 입력할 수 있다. 그러나 대화 제공자가 대화가 진행된 개별 화면을 모두 캡처해야 하고, 타인에게 공개하기 어려운 사진은 별도의 이미지 편집을 해야 하는 등 번거로움이 따른다. 그리고 작업자는 캡처 파일의 발화 내용을 텍스트로 입력하는 비효율적인 작업을 진행해야 한다. 따라서 대규모의 자료를 수집해야 하는 본 사업의 특성상 대화 제공자의 불편함과 작업자의 비효율을 초래할 수 있는 첫 번째 방법은 적절하지 않다.

두 번째 방법은 대규모의 메신저 대화를 효율적으로 수집하기에 적합하다. 카카오톡과 라인 등 주요 메신저의 경우, 대화방의 대화 내용 전체를 텍스트 파일 하나로 변환해 주는 기능을 지원하기 때문이다.

다만 두 번째 방법은 [그림 2]와 같이 이모티콘과 사진 등이 '이모티콘', '사진'과 같은 텍스트로 변환되기 때문에 사용자가 이 요소에 대한 의미 등을 추가로 입력해 주지 않는 이상, 해당 요소의 의미를 파악할 수 없다는 제한이 있다.

2019-11-04 22:25:00 , P1 : 콜라잇오
2019-11-04 22:25:00 , P1 : 이모티콘
2019-11-04 22:25:00 , P2 : ㅋㄱㅋㄱㅋㄱㅋ콜라
2019-11-04 22:25:00 , P2 : 환타 선배다
2019-11-04 22:25:00 , P1 : 콜라도불량식품이지만ㅁㅁㅁ
2019-11-04 22:25:00 , P2 : 이모티콘
2019-11-04 22:25:00 , P1 : 맛이환타는불량맛 ㅋㅋㅋㅋㅋ
2019-11-04 22:26:00 , P2 : 콜라는 환타보다 형님같은 맛ㅋㄱㅋㄱㅋㄱㅋㅋ
2019-11-04 22:26:00 , P2 : 대선배의 맛
2019-11-04 22:26:00 , P1 : 사진

[그림 2] 이모티콘과 사진의 텍스트 추출 예시

또한 10대와 20대 연령에서 활용도가 높은 페이스북 메신저의 경우, 파일 내보내기 기능을 통한 텍스트 형식 변환을 지원하지 않는다[12]. 결국 두 번째 방법은 메신저 대화

12) 페이스북 메신저의 경우 대화 내용을 내보내는 기능은 있으나, 내보내기 기능을 통해 추출된 파일이 일반적인 문서 프로그램으로는 열 수 없는 형식이다.

에 포함된 사진이나 이모티콘 등의 의미 파악이 어렵고, 다양한 메신저로부터 대화를 수집하기 어렵다는 제한이 따른다. 그러나 본 사업에서는 대규모로 대화를 수집하고 이를 가공해야 한다는 점에서 대화 제공자의 편의와 작업자의 작업 효율이 우선이다. 이를 고려하여 두 번째 방법을 대화 자료 수집 방법으로 채택하였고, 이를 '일반 대화 수집'이라 명명하였다.

텍스트 추출 기능을 활용할 경우 대화방 전체의 대화 내용이 하나의 텍스트 파일(.txt)로 만들어지기 때문에 대화방이 만들어진 시점부터 텍스트 추출 시점까지의 대화 수집이 가능하다. 본 사업에서는 대화가 이루어진 시기에 제한을 두지 않고 과거에 이루어진 대화까지 수집 범위에 포함하였다[13].

텍스트 파일로 추출할 경우 이모티콘의 내용을 확인할 수 없다는 점을 보완하기 위하여 대화 제공자에게 사용한 이모티콘의 사용 의도나 감정을 <표 4>의 중분류 항목 중에서 선택하여 입력해 달라는 요청을 하였다[14].

범주	중분류 항목	세부 항목 예시
상태	기쁨	흥미, 행복함, 반가움, 감동, 만족, 안심, 안도, 즐거움
	슬픔	연민, 쓸쓸함, 참담함, 안타까움, 실망, 체념, 아쉬움, 허무, 그리움
	화남	격노, 분노, 억울함
	두려움	공포, 불안, 무서움
	놀람	당황, 의외
	좋아함	사랑, 좋음, 호감
	싫어함	미움, 증오, 혐오, 지루함, 짜증, 심심함, 싫증, 거북함
	바람	설렘, 기대, 부러움
	아픔	
	무표정	
행동	졸림	
	인사	인사말, 축하, 응원, 격려
	사과	
	조롱	
	부탁	

<표 4> 이모티콘의 의미 입력 기준 표

그리고 대화 제공자에게 메신저 대화에 포함된 개인 정보 중 아래의 항목은 대화 추출 후 직접 '***'으로 편집해 달라는 요청을 하였다.

- 본인이나 상대방의 이름
- 주민 등록 번호, 운전면허증 번호 등 신분과 관련된 고유 번호

13) 이 경우에 과거에 이루어진 대화까지도 수집할 수 있다는 이점이 있으나, 참여자 간 관계나 친밀도와 같이 시간의 흐름에 따라 가변적인 메타 정보 항목이 있다는 점을 고려한다면 수집하는 대화의 시간 범위를 제한할 필요가 있다.

14) 심리학적 감정 분류에 기반한 소셜 웹 자료의 감성 분석을 다룬 장문수(2012)와 카카오톡 이모티콘의 감정 분류를 시도한 김유진 외(2014)의 논의를 주로 참고하여 대화 제공자가 이모티콘의 의미를 직관적으로 입력할 수 있도록 단순화하였다.

· 계좌 번호나 카드 번호
· 본인이나 상대방의 구체적인 거주지 정보(주소, 건물명 등)
· 기타 공개하기 싫은 내용

대화 제공자가 직접 비식별화를 하였고, 이후 작업자의 점검을 통한 비식별화[15]가 추가로 이루어졌다. 대화 제공자가 스스로 비식별화할 수 있도록 함으로써 개인 정보 수집에 대한 거부감을 완화시킬 수 있었다. 그러나 대화 제공자 스스로 비식별화를 해야 하는 번거로움이 대화 제공을 어렵게 하는 한편, 기준 없이 과도한 비식별화가 이루어지기도 했다. 대화 제공 편의와 기준 없는 과도한 비식별화 등의 문제를 고려하여 대화 수집 10주차 이후부터는 대화 제공자에게 비식별화를 의무적으로 요청하지 않았다[16].

일반 대화 수집은 기존에 이루어진 대화를 제공자가 스스로 텍스트로 추출하는 것이기 때문에 자연스러운 대화 자료를 수집할 수 있다는 장점이 있다. 반면에 제공자가 직접 대화를 추출하는 과정이 다소 번거롭고, 대화 추출 방법을 모르는 사람을 위한 추가 안내가 필요하다[17]는 단점이 있다.

일반 대화 수집 이외에 본 사업에서는 수집 봇을 활용한 대화 수집을 병행하였다. 이러한 '수집 봇 수집'은 녹음기를 사용하여 구어 대화를 채록하는 방식과 유사하다. 녹음기 역할을 하는 수집 봇을 대화방에 초대한 시점부터 수집 봇이 대화방에서 나가는 시점까지의 대화가 채록되고 채록된 대화는 텍스트 파일로 변환되어 서버에 저장이 된다.

수집 봇 수집은 대화 제공자가 직접 대화를 추출하는 번거로움이 없어 대화 제공자 확보에 용이하다. 또한 대화 환경을 사전에 통제할 수 있어 본 사업 기간 동안 주관 기관의 요청에 따른 낯선 관계의 대화 수집[18]과 주제가 통제된 대화 수집 등에 활용하였다.

다만 수집 봇 수집은 일반적인 대화 채록과 마찬가지로 누군가가 대화 내용을 관찰함으로써 대화 내용이나 대화 진행이 인위적이고 부자연스러워지는 문제가 발생할 우려가 있다[19]. 그리고 대화 제공에 대한 보상이 대화의 분량과 밀접한 관련이 되어 있었기 때

15) 작업자의 비식별화 지침은 뒤에서 설명한다.
16) 개인 정보 노출에 대한 대화 제공자의 우려를 고려하여 공식 블로그를 통하여 개인의 신원이 노출될 수 있는 정보는 철저하게 비식별화가 되며, 대화 내용은 산업계 등의 연구와 개발 목적으로만 활용된다는 사항을 게시하였다.
17) 안드로이드, iOS, PC 등 기기별, 메신저별로 텍스트 추출 방법이 다르다. 본 사업에서는 공식 블로그를 통해 기기별, 메신저별 텍스트 추출 방법을 안내하였다.
18) 대화에 참여한 모두로부터 대화 제공에 대한 동의와 저작권 이용 계약이 이루어져야 한다는 제약에 따랐기 때문에 친밀도가 높은 관계의 대화 위주로 수집이 이루어질 가능성이 높았다. 다만 친밀도가 높은 관계에서 이루어지는 대화의 경우, 서로가 공유하는 대화 맥락이 발화 내에서는 생략되는 경우가 많다. 이러한 대화는 인공 지능의 학습에 유의미한 자료로서 가치가 떨어질 것을 고려하여 공유하는 대화 맥락이 많지 않아 생략 가능성도 비교적 낮은 낯선 관계의 대화도 수집하는 것으로 사전 협의하였다.
19) 수집 봇 수집 참여자로부터 '내 대화를 누군가가 보게 될 것이고 내 대화가 인공 지능 연구와 개발에 활용된다고 생각하니 평소에 비해 띄어쓰기를 비롯한 맞춤법, 속어나 비어 사용에 신경이 쓰였다'는 피드백이 있었다. 이는 표준적인 언어 사용의 구속으로부터 상대적으

문에 분량을 늘리기 위한 부자연스러운 대화 진행의 우려도 있다. 이를 방지하고 최대한 자연스러운 대화를 수집하기 위하여 대화 진행 시 유의해야 할 사항을 아래와 같이 안내하고 대화를 수집하였다.

· 대화 수집을 의식하지 않고 평소와 마찬가지로 자연스럽게 대화를 진행한다.
· 이벤트[20]와 대화 수집에 관련된 내용을 직접적으로 언급하지 않는다.
· 대화방에 있는 수집 봇에게 말을 걸거나 수집 봇에 대해 이야기하지 않는다.
· 자연스럽게 주고받는 대화가 될 수 있도록 혼자만 말하지 않는다(말풍선 5개 이상 연속으로 혼자 말하기 자제).
· 주민 등록 번호나 전화번호, 계좌 번호 등과 같은 개인 정보는 이야기하지 않는다.
· 의도적으로 말풍선 숫자를 늘리기 위한 끝말잇기와 특별한 이유 없이 한 단어 또는 한 글자씩 끊어서 말하기를 하지 않는다[21].
· 이유 없이 분량을 늘리기 위해 'ㅋㅋㅋㅋ' 또는 'ㅎㅎㅎㅎ' 등을 남발하지 않는다.
· 특별한 이유 없이 이모티콘을 남발하지 않는다.

대화 시작 전 위의 사항을 여러 차례 공지하였고, 대화 수집 후 정제 과정에서 대화방에서 이루어진 안내와 대화 제공자의 문의를 비롯한 실제 대화 진행과 관계없는 내용은 삭제하였다. 그리고 위의 항목을 과도하게 위반한 경우는 수집 대상에서 제외하였다.

본 사업에서 활용한 일반 대화 수집과 수집 봇 수집의 장단점을 비교하면 <표 5>와 같다.

	일반 대화 수집	**수집 봇 수집**
설명	메신저의 텍스트 추출 기능을 통해 대화 제공자가 직접 추출한 대화 제공	대화방에 초대된 수집 봇이 대화 채록 후 서버 저장
장점	• 과거 대화 수집 가능 • 자연스러운 대화 수집 가능	• 대화 상황 통제 가능 • 대화 제공이 용이함
단점	• 대화 제공 과정의 불편함 • 특정 상황의 대화 수집 불가능[22]	• 인위적인 대화 진행 우려 • 수집 이후 정제 부담

<표 5> 일반 대화 수집과 수집 봇 수집의 비교

로 자유로운 메신저 대화라는 관점에서 본다면 관찰자의 개입으로 인한 부자연스러움의 발생으로 볼 수 있다. 다만 인공 지능의 학습 자료라는 관점에서 본다면 표준적인 우리말을 구사함으로써 자료의 정제나 전처리의 부담을 완화시킬 수 있는 자료로서 가치를 지닌다고 볼 수 있다.

20) 메신저 대화 수집을 위한 이벤트를 여러 차례 진행하였다. 상세한 내용은 홍보 및 참여자 모집에서 설명한다.

21) 이벤트의 성공 조건과 보상 조건이 발화의 분량이었기 때문에 억지스러운 분량 늘리기를 막기 위한 항목이다.

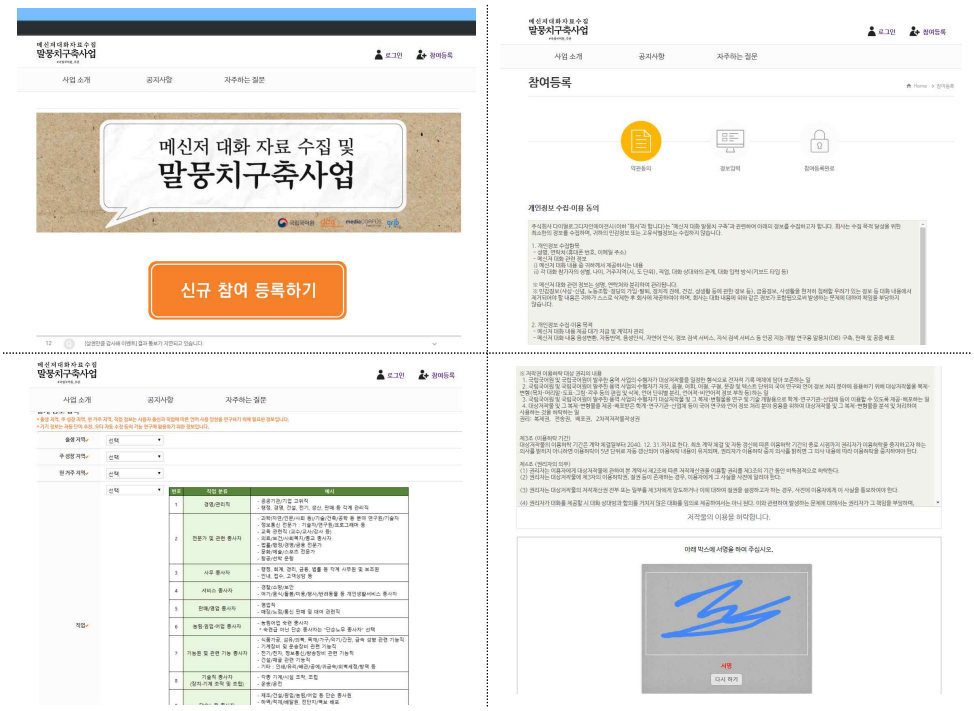
22) 친밀한 관계 이외의 다양한 관계의 대화 수집이 어렵고, 주제가 일관성 있게 유지되는 대화 수집이 불가능에 가깝다.

2.1.2. 메타 정보 수집 및 개인 정보 이용 동의, 저작권 이용 허락 계약

말뭉치의 분류와 검색 및 말뭉치의 효과적인 활용을 위하여 대화 수집 과정에서 대화 참여자의 성별과 연령, 직업을 비롯하여 대화 참여자의 관계나 대화 주제 등의 메타 정보를 수집할 필요가 있다.

이와 같은 메타 정보의 수집과 대화에 포함된 다수의 개인 정보를 법적인 제한 없이 수집하기 위해서는 개인 정보 이용에 대한 동의도 반드시 필요하며, 저작권 이용 허락 또한 자료의 안정적인 활용을 위해 필요하다[23].

본 사업에서는 메타 정보 수집과 개인 정보 이용 동의 및 저작권 이용 허락 계약이 동시에 이루어지는 수집 사이트를 구축하여 활용하였다. 수집 사이트는 PC와 모바일 접속을 고려하여 구축하였고, 구축한 사이트를 통해 대화 제공자가 메타 정보 제공, 개인 정보 이용 동의 및 저작권 이용 허락 절차를 동시에 진행할 수 있도록 하였다[24]. 수집 사이트에서 이루어지는 메타 정보 수집과 개인 정보 이용 동의, 저작권 이용 허락 계약 등의 예시는 [그림 3]과 같다.



[그림 3] 수집 사이트 참여자 등록 과정

23) 메신저 대화와 같이 사적이고 일상적인 대화가 저작권 이용 허락 대상에 포함되는지에 대해서는 현재까지 명확한 법적 기준이 없다. 본 사업에서는 메신저 대화도 대화 참여자가 상호 구성한 어문 저작물에 준하는 것으로 간주하여 대화 제공자 전원과 저작권 이용 허락 계약을 체결하였다.

24) 수집 사이트 구축 후 자체 테스트와 일부 대화 제공자 대상 테스트까지 진행하여 안정성에 문제가 없는 것으로 판단한 11월부터 공식적으로 활용하였다. 그 이전까지 메타 정보의 제공과 개인 정보 이용 동의는 구글에서 제공하는 구글 설문지를 활용하였다. 그리고 저작권 이용 허락 계약은 온라인에서 전자 서명 기능을 제공하는 전자 서명 플랫폼을 활용하였다.

메타 정보는 대화 참여자의 인구 통계 관련 정보와 대화 정보로 구분하여 수집하였고, 메타 정보를 수집하는 취지와 목적을 명확히 밝혔다. 수집한 메타 정보는 <표 6>과 같다.

구분	세부 항목	예시
참여자 정보	대화 참여자 성별	남성/여성
	대화 참여자 연령[25]	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
대화 정보	메신저 종류	카카오톡/라인/페이스북 메신저
	대화 주제[26]	개인 및 관계, 주거와 생활
	대화 참여자 간 관계	가족: 부부 / 학교: 선후배
	친밀도	0(낯선 관계)~5(친밀도가 아주 높다)
	연락 빈도	처음/주 1회 미만/(거의) 매일

<표 6> 메타 정보 수집 항목

참여자 정보 중 기본적인 인구 통계 정보 이외에 대화 참여자의 사용 기기와 키보드 유형은 해당 항목의 유형에 따라 오타 양상이 다르다는 것을 반영한 항목으로서 오타 자동 수정 등에 활용할 수 있다. 그리고 PC와 스마트폰, 키보드 유형에 따라 나타나는 띄어쓰기 양상의 차이, 문장의 길이, 입력 내용의 차이[27] 등도 비교할 수 있다.

대화 정보 중 대화의 주제는 대화의 내용과 밀접한 관련이 있는 대화의 변인으로서, 어휘의 출현과 사용 빈도 등에 영향을 미치는 요인이다. 다만, 특수한 목적을 가진 집단의 대화나, 뚜렷한 목적을 가진 대화, 또는 사전에 주제가 합의된 경우를 제외하면 일상적으로 이루어지는 자연스러운 대화의 주제를 대화 내용에 따라 파악하는 것은 쉽지가 않다. 이를 고려하여 본 사업에서는 주제 통제가 이루어지지 않은 대화 이외에도 수집 봇을 통해 대화 제공자가 스스로 주제를 선택하도록 하여 주제 통제가 이루어진 대화도 수집하였다. 대화 내용을 기준으로 하는 주제 분류는 <표 7>과 같다[28].

---

25) 성별과 연령 정보는 대화 참여자가 직접 입력하지 않고 수집 사이트에서 휴대폰 실명 인증 단계에서 주민등록번호를 입력하면 자동으로 만 연령과 성별이 입력 내용에 반영이 되게 하였다.

26) 수집 봇 수집을 통한 일부 대화만 대화 제공자가 주제를 스스로 입력하였다.

27) 예를 들어 2벌식(쿼티) 자판에서는 'ㅇㅇ'이나 'ㅎㅎ'와 같이 연속으로 자음을 입력하는 것이 어렵지 않다. 하지만 천지인 자판의 경우 연속으로 자음을 입력하는 것이 어렵다. 즉 기기나 자판의 종류에 따라 연속된 자음의 사용 빈도와 양상이 다르게 나타난다.

28) 국립국어원의 세종 말뭉치 구어 항목과 연세 구어 말뭉치 등 기존 구어 말뭉치도 내용에 따라 대화를 분류하고 있다. 그러나 이러한 분류는 '인문', '사회', '자연', '예술' 등과 같이 일상적이고 사적인 대화가 주로 이루어지는 메신저 대화의 내용에 따른 주제 분류 체계로 삼기에는 적합하지 않은 체계이다. 또한 이들 기존 구어 말뭉치의 대화 주제 분류 체계에서는 사적인 관계에서 다양한 주제가 복합적으로 나타날 경우는 '총류' 또는 '기타', '일상 대

항목	항목 예시(키워드)
개인 및 관계	이름, 전화번호, 가족, 국적, 고향, 성격, 외모, 개인의 기호(선호), 직업, 종교, 반려동물, 연애(관), 결혼(관), 이상형, 인간 관계, SNS
주거와 생활	숙소, 방, 가구, 침구, 주거비, 생활 편의 시설, 지역, 지리, 가전 제품, 자취, 집안일, 육아, 부동산, 주거 시설, 이사, 생활비, 자동차
상거래(쇼핑)	쇼핑 시설 및 장소, 식품, 의복, 가정용품, 물건 및 가격, 택배, 중고거래, 서비스, 교환 및 환불, 구매 후기
식음료	식사, 음식, 음료, 배달, 외식, 맛집, 식사 메뉴, 야식, 디저트, 요리
공공 서비스	우편, 전화, 통신, 휴대전화, 인터넷 서비스, 은행, 관공서
여가와 오락	휴일, 취미, 동아리 및 동호회 활동, 관심사, 방학, 휴가, 행사, 술, 웹서핑,
일과 직업	취업, 스펙, 직장 생활, 업무, 회식, 급여, 계약, 협상, 회의
행사 및 모임	초대, 방문, 소개팅, 약속, 가족 및 친척 행사, 공적 행사, 사적 모임(친목 모임)
미용과 건강	신체, 위생, 부상 및 질병, 치료 및 수술, 보험, 병원, 운동, 미용, 다이어트, 건강검진, 약품 및 건강 보조 식품(용품)
기후	날씨, 계절
여행	여행 장소 및 경로, 여행 계획(일정, 숙소, 교통편, 여행 경비), 여행팁, 기념품, 여행사 및 여행 상품
교통	위치, 거리, 길, 이동 수단, 이동 경로, 대중교통(지하철, 버스, 택시)
교육	학교 교육, 교과목, 진로, 학원, 진학, 입시, 시험, 자격증, 성적, 자기 계발, 외국어 학습, 스터디
시사, 사회	정치, 경제, 사회, 사건 및 사고, 법과 제도, 여론, 국제 관계, 재해 및 재난
예술, 문화 생활	문학, 음악, 미술, 공연, 전시, 스포츠 관람, 엔터테인먼트
전공/전문 지식	학문 및 학술 분야, 학회 및 세미나

<표 7> 메신저 대화의 주제 분류 항목

대화 상대방과의 관계는 어휘 사용과 문체 변화를 비롯하여 대화 양상의 주요한 변인 가운데 하나이다. 따라서 메신저를 매개로 하는 다양한 언어 사용 양상을 관찰하기 위해서는 다양한 관계의 대화 제공자로부터 자료를 수집하는 것이 바람직하다.

참여자 간 관계는 사회 활동의 영역인 가정, 학교, 직장, 지역을 큰 범주로 두고 항목별로 세부 분류하였다. 그리고 사회 활동의 범위가 다양화되고 있는 점을 고려하여 기타 범주도 세부 분류하였다. 특히 기타 범주에서는 다른 관계와 비교하여 언어 사용의 변별적 요인이 두드러질 것으로 예상되는 '연인' 범주를 별도로 설정하였다.

'친구' 관계는 그것이 지칭하는 범위가 지나치게 포괄적[29]이기 때문에 상대방과의 관계에 따른 언어적 변별성을 드러내기에 다소 부족할 수 있다는 점을 고려하여 별도의 범주로 두지 않았다.

이를 반영한 대화 참여자 간 관계 항목의 분류는 <표 8>과 같다.

화'와 같은 포괄적인 범주의 주제 분류가 이루어지고 있으며, 내용을 중심으로 세분화된 주제 분류 체계를 제시하지는 않고 있다. 이에 본 사업에서는 한국어 교육 분야에서 사적인 영역에서 공적인 영역을 포괄하여 구어 교육을 위해 설정한 국제 통용 한국어 표준 교육과정(국립국어원, 2017; 김정숙·이정희, 2018)의 주제 분류 체계를 준용하였다.

29) 지역을 기반으로 친구 관계가 형성될 수도 있고, 학교 생활을 기반으로 친구 관계가 형성될 수도 있다. 그 밖에도 사회 활동의 다양한 상황에서 친구라는 관계가 형성될 수 있다는 점에서 그 범위가 포괄적이다.

소속 범주	세부 관계 분류
가족	[가족] 부부
	[가족] 부모 - 자녀
	[가족] 형제, 자매
	[가족] 기타(조부모-손주, 그 외 친인척 등)
학교/학원	[학교/학원] 동기, 동창, 동급생
	[학교/학원] 선후배
	[학교/학원] 스승 - 제자
	[학교/학원] 기타(직원-학생 등)
직장	[직장] 동기, 동료, 동업자
	[직장] 선후배, 상사-부하
	[직장] 기타(거래처, 고객 등)
기타	연인
	종교 관련 지인
	동호회 / 스터디 등
	군대
	온라인 커뮤니티
	낯선 관계, 그 외 사회적 관계

<표 8> 대화 참여자 간 관계 분류

개인 정보의 수집과 이용 동의는 수집 사이트에 최초 참여 등록을 할 때 약관을 제공하고 해당 약관에 동의할 경우 확인 상자(check box)에 확인 표시를 하도록 하였다. 개인 정보 수집과 이용 동의 주요 내용은 <표 9>와 같다.

- 성명과 연락처를 제외한 고유 식별 정보와 민감 정보는 수집하지 않음.
- 대화 참여자가 제공하는 메신저 대화 내용을 수집함.
- 대화 참여자의 성별, 나이, 시.도 단위 거주 지역, 직업, 대화 상대방과의 관계, 대화 입력 방식을 수집함.
- 대화 내용의 음성 변환, 자동 번역, 음성 인식, 자연어 인식, 정보 검색 서비스 등 인공 지능 개발 연구용 말뭉치 DB 구축, 공중 배포 목적으로 활용함.
- 수집 및 이용 목적이 달성되면 파기하며, 보유할 필요가 있는 경우 법령 및 규정에 의거하여 보유함.

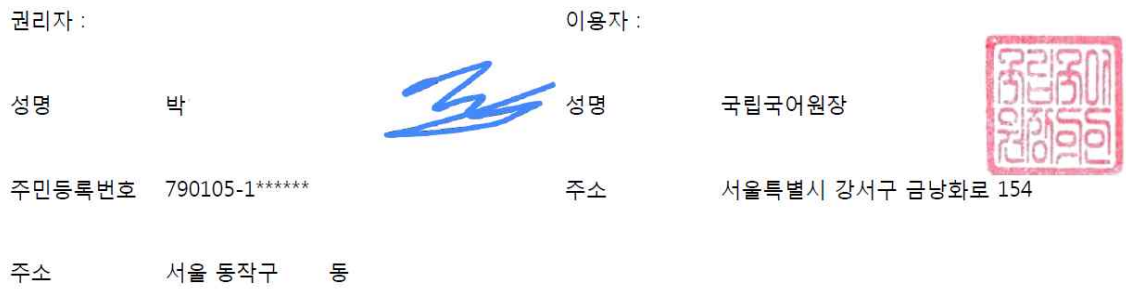
<표 9> 개인 정보 수집·이용 동의 주요 항목

메타 정보 수집과 메신저 대화에 포함된 개인 정보를 제한 없이 수집하기 위하여 위와 같은 내용에 동의를 받은 대화만을 수집하였다. 앞으로도 메신저 대화의 안정적인 수집을 위해서는 「개인정보 보호법」을 기반으로 하여 개인 정보 수집과 이용 동의에 대한 정밀한 법적 검토가 지속적으로 이루어져야 한다.

저작권 이용 허락 계약 또한 수집 사이트에 최초 참여 등록을 할 때 약관[30]을 읽고

30) 저작권 이용 허락 계약서의 약관은 주관 기관이 제공한 기본 저작권 이용 허락 계약서를 토대로 메신저 대화 수집의 특성에 맞게 변경하였다. 가령 대화 건당 계약을 하게 될 경우, 말뭉치 구축 목표인 50만 건에 해당하는 저작권 이용 허락 계약을 체결해야 한다. 50만 건의 저작권 이용 허락 계약서를 대화별로 관리하는 것에 어려움이 따를 뿐만 아니라, 대화 제공자의 입장에서도 대화를 제공할 때마다 저작권 계약을 진행해야 하는 번거로움이 따른다. 이에 본 사업에서는 대화 건당 계약이 아니라, 대화 참여자가 최초 1회 사업 기간 동안 제공한 모든 메신저 대화에 대해 저작권 이용 허락 계약을 하는 것으로 주관 기관과 협의

해당 약관에 동의할 경우 온라인상에서 서명을 하도록 하였고, 최초 계약서 작성 이후 수집 사이트에 로그인하면 언제든지 내려받기를 할 수 있도록 하였다. 서명을 할 때 기재하는 주민 등록 번호는 생년월일과 성별 표기까지, 주소는 동까지만 표기하는 것으로 주관 기관과 협의하였다[31]. 이는 [그림 4]와 같다.



[그림 4] 저작권 계약서 정보 입력과 서명 예시

그리고 저작권 이용 허락 계약의 주요 내용은 <표 10>과 같다.

• 국립국어원의 메신저 말뭉치 구축 사업 기간(2019년 7월 19일부터 2019년 12월 19일까지) 동안 제공하는 모든 메신저 대화를 대상으로 함. • 계약 체결일부터 2040년 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. • 제공 받은 학계, 연구 기관, 산업체가 국어 연구와 언어 정보 처리 분야 응용을 위해 저작물 및 복제, 변형물을 분석 및 처리하여 사용하는 것을 포함함.

<표 10> 저작권 이용 허락 계약의 주요 내용

하였다.

31) 해당 정보는 대화 제공자가 직접 기재하는 것이 아니라 휴대폰 인증 단계에서 입력한 주민 등록 번호와 주소가 계약서에 자동으로 반영된다.

2.2. 홍보 및 참여자 모집

특정 연령대와 성별에 편향되지 않는 10,000명 이상의 참여자에게서 대규모의 메신저 대화 자료를 모집하기 위하여 본 사업 기간 동안 다양한 경로를 통해 홍보를 진행하고 대화 참여자를 모집하였다.

2.2.1. 홍보 창구 운영

<표 11>과 같이 공식 홍보 창구를 운영하여 본 사업 내용, 대화 제공 방법 등 사업 관련 제반 사항을 안내하였고, 이벤트 등 대화 제공자 모집 공고 게시와 1:1 문의 등을 운영하였다[32].

항목	접속 경로	주요 내용
공식 블로그	https://msgcorpus.blog.me/	• 사업 소개 • 대화 제공 방법 안내 • 이벤트 등 대화 제공자 모집 공고 게시 • 자주 묻는 질문
카카오톡 채널	ID : 메신저말뭉치	• 이벤트 등 대화 제공자 모집 공고 게시 • 플러스 친구 대상 이벤트 등 알림 메시지 발송 • 1:1 문의 창구 운영

<표 11> 공식 홍보 창구 운영

카카오톡 채널은 본 채널을 친구로 추가한 카카오톡 사용자를 대상으로 이벤트를 비롯한 사업 관련 공지를 카카오톡 메시지 형식으로 발송할 수 있어 사업 내용에 대한 즉각적인 홍보와 참여를 유도하는 효과를 보였다. 1:1 문의 창구 또한 별도의 플랫폼을 활용하지 않고 카카오톡 채널 내에서 효율적으로 이루어졌다. 그리고 카카오톡 채널의 자동 응답 기능을 활용하여 자주 묻는 질문에 자동적으로 응답이 이루어지고 업무 시간 외에도 기본적인 사항에 대한 안내가 가능하였다.

카카오톡 채널을 홍보에 활용함으로써 적시에 참여자를 모집할 수 있었고, 사업 내용에 대한 의문점도 즉각적으로 대응할 수 있어 사업에 대한 신뢰성을 확보할 수 있었다. 이는 목표 인원 모집에 긍정적인 영향을 미친 요인 가운데 하나이다.

공식 블로그는 실질적으로 대화를 제공하기 위해 필요한 절차와 방법을 다양한 형식으로 편집하고 게시할 수 있어, 카카오톡 채널 게시글의 부족한 편집 기능을 보완하여

32) 사업 진행 초기에는 페이스북과 트위터, 인스타그램 등 SNS에도 공식 계정을 개설하여 홍보를 진행하였으나, 조회 수를 비롯한 반응도가 낮았고, 이후 효율적으로 홍보 창구를 운영하고 관리하기 위하여 공식 블로그와 카카오톡 채널 두 가지만을 홍보 창구로 활용하였다.

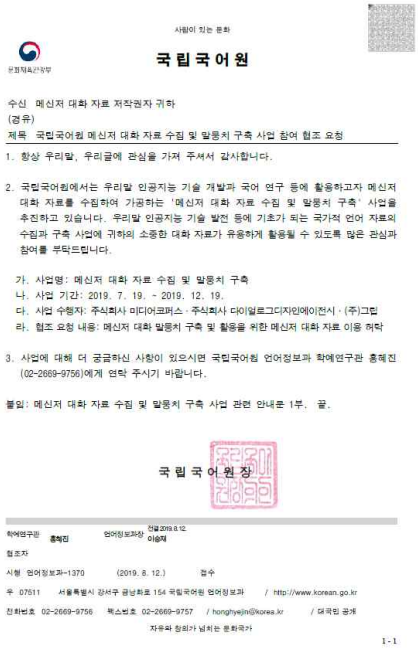
사업과 참여 방법 등에 대한 상세한 정보를 전달하는 역할을 하였다.

[그림 5] 카카오톡 채널 및 공식 블로그

2.2.2. 참여자 모집

사업 기간 동안 주관 기관인 국립국어원의 협조하에 홍보와 참여자 모집이 이루어졌다. 국립국어원의 협조는 본 사업에 대한 참여 협조 요청 공문 발행과 사업 내용의 국립국어원 누리집 게시 등으로 이루어졌다. 이를 통해 본 사업이 국가에서 추진하는 사업이라는 신뢰성을 확보할 수 있었다[33]. 이는 메신저 대화라는 지극히 사적인 영역의 대화를 수집해야 한다는 제한에도 불구하고 목표 인원을 확보할 수 있었던 요인 가운데 하나이다.

33) 사업 기간 동안 본 사업이 국립국어원 사업이 맞는지에 대한 문의가 다수였다. 그리고 본 사업 참여자가 남긴 것으로 추정되는 개인 SNS와 블로그, 카페 등의 게시글에서도 본 사업이 국립국어원 주관으로 진행되기 때문에 신뢰할 수 있다는 내용을 다수 확인할 수 있었다.



사람이 있는 문화

국 립 국 어 원

문화체육관광부

수신 메신저 대화 자료 저작권자 귀하
(경유)
제목 국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업 참여 협조 요청

1. 항상 우리말, 우리글에 관심을 가져 주셔서 감사합니다.

2. 국립국어원에서는 우리말 인공지능 기술 개발과 국어 연구 등에 활용하고자 메신저 대화 자료를 수집하여 가공하는 '메신저 대화 자료 수집 및 말뭉치 구축' 사업을 추진하고 있습니다. 우리말 인공지능 기술 발전 등에 기초가 되는 국가적 언어 자료의 수집과 구축 사업에 귀하의 소중한 대화 자료가 유용하게 활용될 수 있도록 많은 관심과 참여를 부탁드립니다.

가. 사업명: 메신저 대화 자료 수집 및 말뭉치 구축
나. 사업 기간: 2019. 7. 19. ~ 2019. 12. 19.
다. 사업 수행자: 주식회사 미디어코퍼스 · 주식회사 다이얼로그디자인에이전시 · (주)그립
라. 협조 요청 내용: 메신저 대화 말뭉치 구축 및 활용을 위한 메신저 대화 자료 이용 허락

3. 사업에 대해 더 궁금하신 사항이 있으시면 국립국어원 언어정보과 학예연구관 홍혜진 (02-2669-9756)에게 연락 주시기 바랍니다.

붙임: 메신저 대화 자료 수집 및 말뭉치 구축 사업 관련 안내문 1부. 끝.

국 립 국 어 원 장

학예연구관 홍혜진 언어정보과장 이승재 전결 2019.8.12.
협조자
시행 언어정보과-1370 (2019. 8. 12.) 접수
우 07511 서울특별시 강서구 공항대로 154 국립국어원 언어정보과 / http://www.korean.go.kr
전화번호 02-2669-9756 팩스번호 02-2669-9757 / honghyejin@korea.kr / 대국민 공개

자유와 창의가 넘치는 문화국가

1 - 1

공지 사항 상세보기

국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업 관련 안내

작성자 국립국어원 | 등록일 2019.8.12. | 조회수 21095

첨부파일 총1건 (109.74 KB) 전체 내려받기

메신저 대화 자료 수집 및 말뭉치 구축 사업 관련 안내.pdf 미리보기

<국립국어원 공고 제2019-174호>

국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업 관련 안내

국립국어원에서는 우리말 인공지능 기술 개발과 국어 연구 등에 활용하고자 메신저 대화 자료를 수집하여 가공하는 '메신저 대화 자료 수집 및 말뭉치 구축' 사업을 추진하고 있습니다. 우리말 인공지능 기술 발전등에 기초가 되는 국가적 언어 자료의 수집과 구축 사업에 귀하의 소중한 대화 자료가 유용하게 활용될 수 있도록 많은 관심과 참여를 부탁드립니다.

[그림 6] 국립국어원 참여 협조 요청 공문 및 국립국어원 누리집 공지

사업 기간 동안 진행한 홍보 및 모집 항목별 내용은 아래와 같다.

2.2.2.1. 일반 참여자 모집

사업 착수 후 1차 지침과 홍보 및 모집 계획을 수립하고 공식 블로그 등의 홍보 창구를 개설한 후 메신저 대화 자료 수집에 대한 홍보와 대화 제공자 모집을 진행하였다[34].

일반 참여자 모집 초기에는 어문학 및 자연어 처리 관련 대학교 학과, 연구소, 학회와 같이 말뭉치에 대한 이해도와 구축된 자료의 실질적 활용도가 높을 것으로 예상되는 단체를 대상으로, 기관의 홈페이지 게시와 해당 기관 구성원의 참여 독려 협조를 요청하는 방식으로 홍보가 이루어졌다. 또한 카카오톡 오픈 채팅방[35]의 자연어 처리 관련 대화방에서도 지속적인 홍보를 진행하였다. 이후 홍보의 범위를 확대하여 학과와 연구소, 학회 및 개별 연구자를 대상으로 하는 협조 요청 이외에도 페이스북과 트위터, 인스타그램 등의 SNS를 활용하여 일반인을 대상으로 하는 홍보와 대화 제공자 모집도 병행하였다.

초기 홍보를 통해 누적 인원 기준으로 157명의 대화 제공자를 모집하였다. 일부 대화 제공자의 피드백에 의하면 사업 내용과 대화 제공 방법에 대한 홍보가 본격적으로 이루어지지 않아 사업 자체에 대한 신뢰도가 높지 않았던 상황에서 지극히 사적인 영역에

34) 일반 참여자의 대화 제공은 사업 기간 동안 상시로 이루어졌다.
35) 카카오톡의 오픈 채팅방은 카카오톡 사용자 누구나 자유롭게 개설이 가능하고 익명으로 참여가 가능한 단체 대화방이다.

해당하는 메신저 대화를 제3자에게 제공하는 것에 대한 거부감이 높았다는 점, 특히 대화에 참여한 상대방에게까지 대화 제공 동의를 받는 것이 어려웠다는 점이 대화 제공을 저해한 요인 가운데 하나로 나타났다. 물론 대화 제공자 본인이 공개를 원하지 않는 개인 정보를 스스로 비식별화하여 대화를 제출할 수 있게 하였으나, 이 경우 대화 제공 당사자가 대화 파일을 스스로 편집해야 하는 번거로움이 발생하여 대화 제공을 꺼리게 하는 요인으로 작용했다.

사업 내용과 대화 제공 방법 등에 대한 홍보가 본격적으로 이루어지지 않은 사업 초기임을 감안하더라도 집중 홍보 대상이었던 말뭉치 관련 분야 연구자의 참여율이 높지 않은 점은 다소 아쉬운 부분이다.

[그림 7] 사업 초기 홍보 게시물 예시

2.2.2.2. 다자 대화 이벤트

사업 초기 대화 참여자 모집 결과와 대화 제공자의 피드백을 분석한 결과 사업에 대한 홍보가 더욱 확대되어야 하며, 대화 제공 절차를 더욱 간소하게 정비해야 할 필요성이 제기되었다.

먼저 본 사업에 대한 정보 공유 및 확산을 위해 온라인 커뮤니티 활동과 SNS 활용에 적극적인 계층을 2차 홍보의 집중 홍보 대상으로 선정하였다. 특히 2차 홍보 진행 기간이 대학생의 2학기 개강 시점임을 고려하여 대학교 재학생의 수강 신청 및 대학 생활 관련 정보 공유 커뮤니티를 대상으로 한 홍보가 집중적으로 이루어졌다. 아울러 접속자 상위 온라인 카페와 커뮤니티, 페이스북의 지역별 대표 커뮤니티[36]에서도 홍보와 참여

36) 대표적으로 페이스북의 지역별 '대신 전해 드립니다' 페이지를 중심으로 이벤트를 홍보하였다.

자 모집이 집중적으로 이루어졌다.

대화 제공자가 직접 대화 파일을 추출하고 편집과 전송을 하는 방식이 다소 번거로웠고, 이러한 점이 대화 제공을 저해하는 요인으로 작용했다는 피드백을 반영하여 대화 제공 절차를 간소화하는 수집 봇 수집 방식을 활용하였다.

다자 대화 수집을 목표로 하여 기간 내 신규 참여 인원 1,240명을 대화 제공자로 모집하였다. 대화 제공자 인원의 증가와 함께 본 사업에 대한 이해도 또한 향상되어 메신저 대화를 직접 추출하여 파일로 제출하는 참여자도 증가하였으며, 다자 대화 수집 이벤트에 대한 문의만이 아니라 본 사업에 대한 문의가 증가함에 따라 카카오톡 채널을 1:1 문의를 위한 창구로 활용하였다.

본 이벤트 기간 동안 다수의 대화 제공자를 확보하였고, 본 사업의 취지와 목적에 대한 홍보도 기대 이상의 성과를 달성하면서 남은 사업 기간 동안 목표 인원을 달성하는 발판을 마련하였다. 다만 20대 대학생 위주로 홍보가 진행이 되어 이 기간 동안 대화 제공자 중 20대 대학생 비율이 다소 높은 점과, 이벤트 참여 시 지켜야 할 대화 규칙[37]에 대한 사전 통제에 한계가 있었던 점은 다소 아쉬운 부분이다.

[그림 8] 다자 대화 이벤트 홍보 게시물 예시

37) 자연스러운 일상 대화를 수집한다는 목적을 사전 공지하여, 이벤트 자체에 대한 언급은 하지 말라는 제한 조건이 있음에도 불구하고 해당 내용을 숙지하지 않은 채 대화에 참여하여 이벤트 관련 내용으로 대화가 진행되는 경우가 있었다. 이러한 부분은 대화 정제 과정에서 대화 맥락을 심각하게 손상시키지 않을 경우 삭제하였다. 대화 규칙과 정제에 대한 내용은 자료 수집 방법과 원시 말뭉치 구축 항목에서 상세히 기술한다.

2.2.2.3. 1:1 대화 및 다자 대화자 모집

다자 대화 이벤트 진행 이후 수집 봇 운용 방식을 개선하여 1:1 대화와 3인 대화에 한정한 다자 대화자 모집을 진행하였다.

온라인 구인·구직 사이트를 중심으로 모집 공고를 게시하였고, 이전까지 카카오톡 채널을 친구 추가한 참여자를 대상으로 홍보 메시지를 발송하였다. 동일인 대화 중복을 막고 참여 인원을 확보하기 위하여 기존 대화 제공 이력이 있는 참여자 간 대화 참여는 제한하였다.

1:1 대화와 3인 다자 대화를 대상으로 한 대화 제공자 모집을 통해 기간 내 신규 참여 인원 4,725명을 모집하였다. 카카오톡 채널 친구를 대상으로 한 알림 메시지 발송이 즉각적인 홍보 효과를 보였으며, 참여자의 자발적인 홍보 게시물 공유 건수가 증가하였다.

수집 봇을 통해 대화를 수집하는 방식에 익숙해진 참여자의 증가와 1:1 및 3인 다자 대화로 한정된 대화가 이루어지면서 기존 다자 대화 이벤트에 비해 대화 규칙에 대한 사전 숙지와 이로 인한 대화 통제가 비교적 원활히 이루어졌다. 전반적으로 대화 수집을 위한 홍보가 안정적으로 정착되었으나, 구인·구직 사이트를 주요 홍보 창구로 활용함에 따라 직장인의 참여가 다소 제한되었던 점은 다소 아쉬운 부분이다.

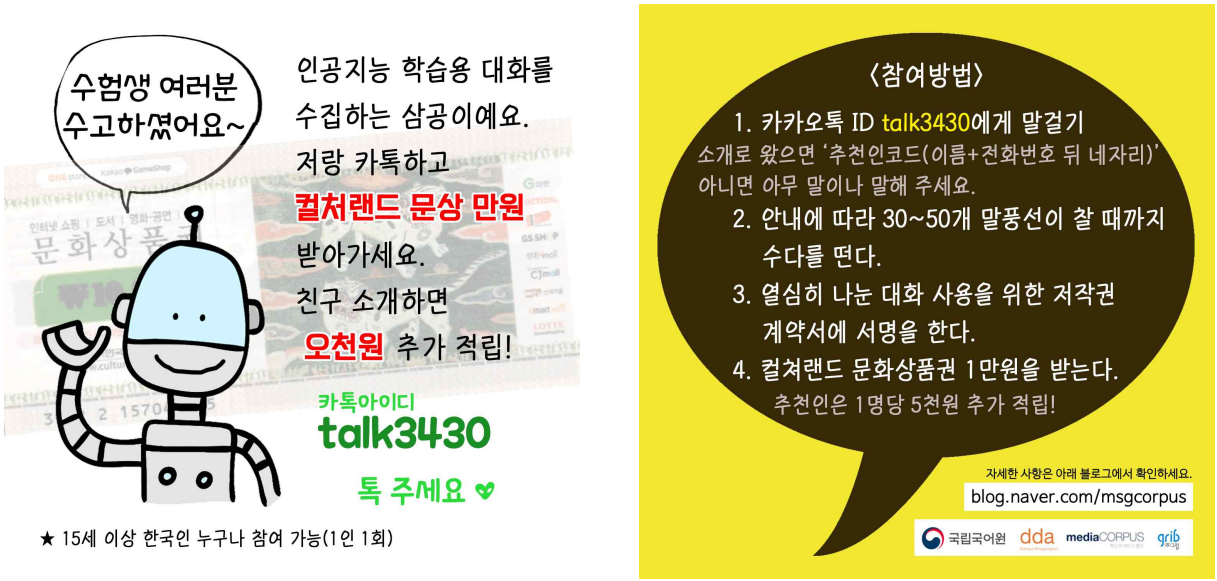
[그림 9] 1:1 대화 및 다자 대화자 모집 게시물 예시

2.2.2.4. 낯선 관계 대화 이벤트

주관 기관의 요청[38]에 따라 수집 자료에 낯선 관계의 대화를 포함시키기로 하였고, 수집 봇을 활용하여 임의로 배정된 상대방과 대화를 진행하는 이벤트를 진행하였다.

기존 대화 제공자의 참여를 제한하는 대신 추천인 제도를 운영하였고, 카카오톡 채널 친구를 대상으로 한 알림 메시지 발송 등을 활용한 홍보를 진행하였다.

낯선 관계의 대화 수집을 목적으로 한 대화 제공자 모집을 통해 기간 내 신규 참여 인원 2,627명을 모집하였다. 추천인 제도를 통해 기존 참여자가 자신의 가족이나 지인에게 자발적이고 적극적으로 이벤트를 홍보하는 효과를 얻을 수 있었다. 그리고 기존 참여자의 가족이나 지인 이외에도 온라인 카페 등의 여러 커뮤니티에서 자발적으로 해당 이벤트 내용을 게시하고 공유하는 사례가 증가하여 신규 대화 제공자의 모집이 안정적으로 진행되었다. 비교적 짧은 길이의 대화를 수집함으로써 긴 시간 동안 대화를 하지 않아도 된다는 점이 참여자 확보에 긍정적인 요인으로 작용하였으며, 40대 이상 중장년층의 참여도 또한 비교적 높게 나타났다.



[그림 10] 낯선 관계 대화 이벤트 대화자 모집 게시물 예시

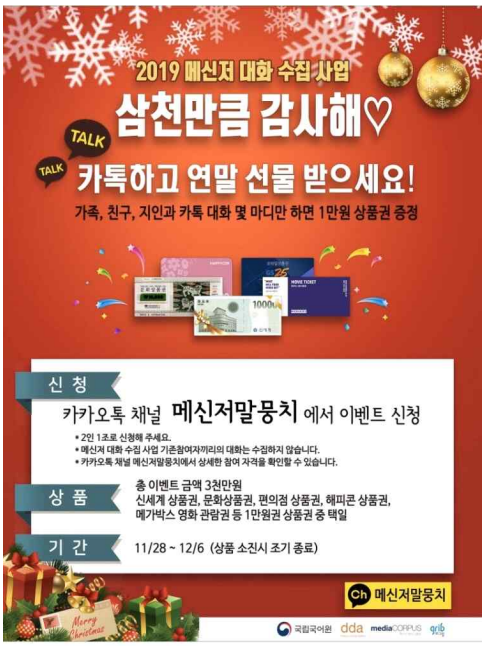
38) 대화에 참여한 모두로부터 대화 제공에 대한 동의와 저작권 이용 허락 계약이 이루어져야 한다는 제약에 따랐기 때문에 친밀도가 높은 관계의 대화 위주로 수집이 이루어질 가능성이 높았다. 다만 친밀도가 높은 관계에서 이루어지는 대화의 경우, 서로가 공유하는 대화 맥락이 발화 내에서는 생략되는 경우가 많다. 이러한 대화는 인공 지능의 학습에 유의미한 자료로서 가치가 떨어질 것을 고려하여 공유하는 대화 맥락이 많지 않아 생략 가능성도 비교적 낮은 낯선 관계의 대화도 수집하는 것으로 사전 협의하였다.

2.2.2.5. 주제 특화 1:1 대화 이벤트

구축하는 원시 말뭉치의 메타 정보에 주제가 포함되는 것을 고려하여 수집 봇을 활용한[39] 주제 특화 메신저 대화 참여자를 모집하였다.

기존 대화 제공자 간 참여와 기간 내 중복 참여를 제한하고 신규 참여자 1인 이상을 반드시 포함한 경우에만 참여가 가능하도록 하였다. 카카오톡 채널 친구를 대상으로 이벤트 알림 메시지 발송과 30대 이상과 남성 참여자를 모집하기 위한 온라인 커뮤니티 홍보를 병행하였다.

주제 특화 대화 이벤트 모집을 통해 기간 내 신규 참여 인원 1,386명을 모집하였다. 낯선 관계 대화 이벤트와 마찬가지로 기존 참여자의 자발적인 홍보와 온라인 커뮤니티 내에서 이루어진 자발적인 홍보와 이벤트 공유 등의 효과로 안정적인 대화 제공자 모집이 이루어졌다. 또한 대화 참여자가 스스로 선택한 주제 내에서 비교적 짧은 길이의 대화를 요구한 점도 다수의 참여자를 확보할 수 있었던 요인 중 하나이다.



[그림 11] 주제 특화 1:1 대화 이벤트 홍보 및 모집 게시물 예시

수집 기간 동안 메신저 대화 제공 인원의 추이는 <표 12>와 같다[40].

39) 통제된 상황에서 채록이 이루어지는 경우를 제외하면 자연스러운 일상 대화의 주제는 상황에 따라 유동적으로 변한다. 통상적인 메신저 대화도 마찬가지로 일관된 단일 주제의 분석 기준을 마련하여 기준에 따라 분석하고 분류하는 것이 쉽지 않다. 이러한 제한을 고려하여 수집 봇을 통해 대화 참여자가 선택한 주제 내에서만 대화를 하도록 주제가 통제된 대화를 수집하였다.

40) 실제로 대화 제공을 한 사람 중에서 메타 정보 입력과 저작권 이용 허락 계약까지 문제없

기간	모집 내용	신규 참여	누적 참여
8/9~8/29	일반 참여자 모집	157	157
9/2~9/12	다자 대화 이벤트	1,240	1,397
9/20~11/7	1:1 대화 및 다자 대화(3인) 수집	4,725	6,122
11/8~11/21	낯선 관계 대화 이벤트	2,627	8,749
11/15~12/5	주제 특화 대화 이벤트	1,386	10,135
합계			**10,135**

<표 12> 대화 제공 인원 추이

2.3. 자료 선별

홍보와 참여자 모집 등을 통해 수집한 메신저 대화 자료 중 말뭉치 구축 목적과 지침에 맞는 자료를 선별하였다.

2.3.1. 일반 대화의 선별

일반 대화의 경우 비속어의 사용이 포함된 대화는 자연스러운 언어 습관의 한 부분으로 간주하여 수집 대상에 포함시켰다. 그러나 지나치게 선정적인 내용이나 반사회적인 내용, 범죄 모의, 혐오나 차별 등의 내용이 다수 포함되어 있어 말뭉치로 구축하였을 경우 논란이 될 여지가 있는 메신저 대화는 수집 대상에서 제외하였다.

2.3.2. 수집 봇 수집 대화의 선별

수집 봇 수집 대화의 경우는 사전에 지나치게 선정적인 내용, 반사회적 내용, 혐오나 차별 등의 논란이 될 수 있는 대화는 수집하지 않는다고 안내하여 이러한 대화를 원천적으로 차단하였다. 사전 공지에도 불구하고 그러한 내용이 포함되어 있을 경우는 일반 대화와 마찬가지로 수집 대상에서 제외하였다.

그리고 입력한 메타 정보와 대화 내용을 검수하여 메타 정보로 입력한 성별, 연령, 직업, 상대방과의 관계 등의 정보와 실제 대화 내용이 일치하지 않는 대화[41]는 수집

이 이루어진 숫자이다. 대화를 제공한 사람 중에서 본인이 저작권 계약을 하지 않았거나 대화 참여자 상대방이 대화 제공에 동의를 하지 않아 상대방의 계약서가 누락된 경우는 참여 인원에서 제외하였다.

41) 이벤트 등을 통해 대화를 제공한 사람 중 일부는 가족이나 지인 등 타인의 휴대폰 명의를 도용하여 신청하는 경우가 있었다. 다수의 대중을 상대로 대화 제공자를 모집하는 과정에서 대화 제공자를 직접 대면할 수 없다는 점, 휴대폰 본인 인증의 절차가 단순하여 가족을 비롯한 가까운 관계의 사람으로부터 휴대폰을 잠깐 빌려 본인 인증을 한 후 PC로 메신저 접속을 하는 것이 가능하다는 점 등을 악용한 사례가 발생한 것이다. 언어 연구와 인공 지능 개발과 연구를 위한 국가 단위 사업이라는 점을 사전에 안내하여 명의 도용을 비롯한 편법을 동원한 대화 제공은 불가능하다는 점을 충분히 안내하였고, 입력한 메타 정보의 성별,

대상에서 제외하였다.

수집 봇 대화 수집의 경우, 수집 봇의 안내 발화 이외에도 [그림 12]와 같이 대화 제공자가 이벤트나 본 사업 내용에 대해 언급하거나 수집 봇에게 말을 거는 발화 등 부자연스러운 대화 진행이 나타나는 경우가 있었다.

[P6] [오후 11:27] 여러분 연구실 계시는 분 있나요
[P6] [오후 11:27] 위당관 현관문 ㅠ
[P6] [오후 11:28] 내일은 꼭 신청하고 말리라...
[P2] [오후 11:30] 이제 끝난거..?
[P2] [오후 11:30] 얼른 집에 가ㅎㅎ
[P6] [오후 11:30] 아냐 ㅋㅋㅋ이제 연구실 도착한거야 ㅋㅋㅋ운동하고
[P2] [오후 11:31] 아하ㅎㅎ 20개 채웠다..
[P2] [오후 11:31] 말풍선..
[P6] [오후 11:32] 20개 채우려고 열심히 답장했구만ㅋㅋㅋㅋㅋㅋㅋ
[P2] [오후 11:33] 아닠ㅋ그런 건 아닌데 괜히 그냥 세어 보게 되었어ㅋㅋㅋㅋ
[P6] [오후 11:38] ㅋㅋㅋㅋㅋ나도 20개 채운 듯??!
우리 막 나중에는 ㅋㅋ 연구실에서 말 안하고 카톡으로 이야기하고 그러는거 아니냐며 ㅋㅋㅋ
[P3] [오후 11:58] 오 난 아직 20개 못채우구 지금 이제서 연구실 가는중ㅠ하아 내일 태양이 안 뜨길..
[P3] [오후 11:59] 위당관 카드 출입신청해야하는거야?
--------------- 2019년 9월 4일 수요일 ---------------
[P4] [오전 12:05] 헐 지금 안열릴탠데
[P4] [오전 12:06] 위당관 출입권한 내가 신청하려니까
안됐어
[P4] [오전 12:06] 교수님이라도...혹시 연구실 계시면 안에서 열어달라고...
[P3] [오전 12:06] 아.... 밤엔 통제하는구낭

[그림 12] 수집 봇 대화 수정 전 예시

[P6] [오후 11:27] 여러분 연구실 계시는 분 있나요
[P6] [오후 11:27] 위당관 현관문 ㅠ
[P6] [오후 11:28] 내일은 꼭 신청하고 말리라...
[P2] [오후 11:30] 이제 끝난거..?
[P2] [오후 11:30] 얼른 집에 가ㅎㅎ
[P6] [오후 11:30] 아냐 ㅋㅋㅋ이제 연구실 도착한거야 ㅋㅋㅋ운동하고
[P3] [오후 11:59] 위당관 카드 출입신청해야하는거야?
--------------- 2019년 9월 4일 수요일 ---------------
[P4] [오전 12:05] 헐 지금 안열릴탠데
[P4] [오전 12:06] 위당관 출입권한 내가 신청하려니까
안됐어
[P4] [오전 12:06] 교수님이라도...혹시 연구실 계시면 안에서 열어달라고...
[P3] [오전 12:06] 아.... 밤엔 통제하는구낭

[그림 13] 수집 봇 대화 수정 후 예시

직업, 연령을 비롯하여 상대방과의 관계, 친밀도 및 기존 대화 제공 이력 등을 추적하여 본인이 아닌 것으로 판명되는 경우는 대화 수집 불가를 통보하는 등 강력한 차단 조치를 시행하였다. 수집 불가 통보가 이루어진 이후 본인임을 충분히 소명한 대화에 한해서만 수집 대상에 포함하였다.

[그림 12]와 [그림 13]은 대화 참여자 모두 하루 최소 20회 이상 발화를 해야 한다는 대화 규칙[42]과 관련하여 참여자가 대화를 나눈 예시와 이 부분을 삭제한 예시이다. 이와 같은 경우는 대화 전체를 수집 대상에서 제외하는 대신 대화 맥락을 고려하여 맥락을 해치지 않는 범위 내에서 자연스러운 대화 진행과 관계없는 부분을 삭제하였다.

3. 정제 및 마크업

메신저 대화 말뭉치의 활용성을 높이기 위해 시스템 메시지와 같이 화자의 직접 발화가 아닌 요소와 민감한 개인 정보는 '걸러내기(filtering)'나 '바꾸기(replacing)' 등이 용이한 형태로 정제가 필요하다[43]. 또한 메신저 대화는 사용 기기와 운영 체제(OS), 메신저 종류에 따라 추출 형식이 상이하기 때문에 이를 표준화하는 과정이 반드시 필요하다. 이후 마크업 지침[44]에 따라 메타 정보 부착과 태깅이 완료되면 1차적인 원시 말뭉치 구축이 이루어지는 것이다.

3.1. 비식별화 및 특수 메시지 처리

3.1.1. 개인 정보의 비식별화

수집한 메신저 대화 자료 대부분은 친밀한 관계에서 이루어지는 사적인 대화이며, 이름과 연락처, 소속 등을 비롯하여 개인의 신원이 노출될 수 있는 다양한 개인 정보가 포함되어 있다. 대화 내용에 포함된 개인 정보와 메타 정보로 수집하는 성별과 연령, 직업 등의 정보가 결합한 형태로 말뭉치로 구축될 경우 개인의 신원이 노출될 우려가 있기[45] 때문에 개인 정보에 대한 철저한 비식별화가 필요하다. 본 사업의 비식별화 기본 지침은 국무조정실 외(2016)의 지침을 토대로 본 사업의 특성을 고려하여 <표

42) 다자 대화 이벤트의 경우 대화 참여자의 숫자가 보상과도 관련이 되어 있었다. 따라서 참여 신청만 하고 발화를 하지 않는 상황을 통제하기 위하여 하루 최소 20회 이상 발화를 해야 한다는 규칙을 설정하였다.

43) 표현의 정규화(normalization)를 위한 '교정(correcting)'이나 '풀어쓰기(paraphrasing)'도 메신저 대화의 정제에 필요한 작업이다. 다만 교정이나 풀어쓰기는 현재까지 메신저 대화의 형태적 특성을 고려한 교정 또는 풀어쓰기 기준 마련이 어려우므로 본 사업에서는 제외하였다. 앞으로 본 사업을 통해 구축된 말뭉치를 토대로 하여 메신저 대화의 형태적 특성을 고려한 교정 또는 풀어쓰기 등 정규화나 전처리에 필요한 기준 등을 마련할 필요가 있다.

44) 마크업 지침을 비롯하여 비식별화 항목, 특수 메시지 항목의 태깅 지침은 주관 기관의 최초 지침을 기준으로 하여 중간 산출물 납품 후 피드백과 작업 진행 등을 고려하여 여러 차례 수정이 이루어졌다. 최종 지침은 사업 중간 보고 이후 협의를 토대로 15주차에 확정하였고, 16주차부터는 이를 반영하여 비식별화 등의 작업과 태깅 수정 등이 이루어졌다.

45) 국무조정실 외(2016 : 4)에 따르면 특정 개인을 알아 볼 수 없게 조치한 경우에도 다른 정보와의 결합으로 특정 개인을 알아볼 수 있는 경우는 개인 정보에 포함되는 것으로 간주한다.

13>과 같이 마련하였다.

구분	식별자	속성자
지침	개인 또는 개인과 관련된 사물에 고유하게 부여되는 값 또는 이름으로 반드시 가린다.	대화 제공자가 해당 내용을 가리지 않은 경우 해당 정보로 인해 누군가를 특정할 수 있는 상황인지 판단하여 가린다.
항목	• 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 • 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	• 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 • 경력, 직업, 직종, 직장명, 부서명, 직급

<표 13> 메신저 대화 비식별화 기본 지침

다만 과도한 비식별화로 인하여 대화 내용을 파악하는 것이 불가능하거나, 유효한 개체명(entity)을 추출해 내기가 어려워지는 등 자료의 활용도가 떨어질 우려가 있다. 그리고 개인 정보를 알아볼 수 있는 주체가 해당 정보를 제공받거나 처리하는 자로 한정되어 있다는 점[46]도 감안하여 과도하게 비식별화하지 않는 것을 원칙으로 하였다[47].

최초 비식별화 항목은 일괄 '***'으로 표기하였으나, 해당 요소가 어떤 항목을 비식별화한 것인지 확인이 가능하도록 비식별화 표기 방식을 1차 수정하였다. 1차 수정된 비식별화 작업 지침은 <표 14>와 같다.

46) 위의 책. 개인 정보 유출의 범위가 자료를 제공받거나 처리하는 자로 한정되어 있는 만큼 향후 주관 기관이 이 자료를 개인 및 기관 등 관련 연구자에게 공개할 경우 보안 서약 등을 통한 엄밀한 관리가 필요하다.

47) 다만 앞서 밝혔듯이 일반 대화 수집의 경우 대화 제공자 스스로가 비식별화할 것을 요구함으로써 일부 과도한 비식별화가 이루어진 경우나 비식별화의 기준이 모호한 경우도 있다.

범주	항목	처리	원문 표기
이름	실명	비식별화	₩이름₩
	실명(변형)	비식별화	
	특수 애칭, 별명, 대화명, 필명	비식별화	
	일반 애칭 별명	비식별화하지 않음	
	공인 실명	비식별화하지 않음	
온라인	아이디	비식별화	₩아이디₩
	이메일 주소	비식별화	
	인터넷 URL[48]	http, https 제외 비식별화	http*** https***
각종 번호 및 비밀번호	고유 식별 번호	비식별화	₩번호₩
	사업자 등록 번호	비식별화	
	(구매자) 식별 번호	비식별화	
	일련번호	비식별화	
	전화번호	비식별화	
	금융 번호	비식별화	
	비밀번호	비식별화	
장소	상세 주소	동 이하 비식별화	₩장소₩
	아파트 및 거주 건물명	비식별화	
	거주지 역명	비식별화하지 않음	
	방문 장소(비정기적)	비식별화하지 않음	
	상호명	비식별화하지 않음	
출신 및 소속	출신 및 소속 학교	비식별화	₩소속₩
	출신 및 소속 직장	비식별화	
	출신 및 소속 부대	비식별화	
기타	위에서 언급하지 않은 항목	비식별화	₩기타₩

<표 14> 비식별화 1차 수정 지침

이름 범주의 실명 항목은 대화 참여자 또는 제3자의 실명을 의미하며, 실명(변형) 항목은 실명의 형태적 변형과 오타를 의미한다. 특수 애칭과 별명, 대화명, 필명은 대화 참여자 또는 제3자의 특수한 별명이나 애칭, 대화명, 필명, 호칭으로 사용되는 아이디 등을 의미한다. 일반적인 애칭이나 연예인, 공인 등의 이름은 비식별화 대상이 아니다.

온라인 범주의 아이디 항목은 대화 참여자 또는 제3자가 온라인상에서 사용하는 각종 아이디를 의미하며, 이메일 주소 항목은 온라인상에서 사용하는 이메일 주소를 의미한다. 인터넷 URL은 http 또는 https로 시작하는 온라인 주소와 그 외 형식의 온라인 주소를 모두 포함한다.

각종 번호 및 비밀번호 범주의 고유 식별 번호는 주민 등록 번호나 운전면허 번호와 같이 대화 참여자 또는 제3자의 신원에 부여되는 고유한 번호를 의미한다. 일련번호 항목은 대화 참여자 또는 제3자가 소유한 자동차나 각종 기기, 컴퓨터의 네트워크 주소 등 일련번호를 의미한다. 전화번호 항목은 대화 참여자 또는 제3자의 휴대 전화 번호나 집, 직장 등 전화번호를 의미한다. 금융 번호 항목은 대화 참여자 또는 제3자의 통장 계좌 번호나 신용 카드 번호 등을 의미한다. 기타 대화 참여자 또는 제3자의 사

48) 인터넷 URL의 경우 개인 블로그나 SNS 주소인 경우, 개인이 실명으로 남긴 상품평이나 댓글 등의 주소인 경우가 있었다. 인터넷 URL 전체를 추적하여 해당 항목이 개인 정보와 관련이 없다는 것을 확인하는 것이 불가능하여 일괄 비식별화 처리하였다.

업자 등록 번호, 구매와 관련된 번호, 예매 번호, 인증 번호, 비밀번호 등도 비식별화 대상에 포함된다.

장소 범주의 상세 주소 항목은 대화 참여자 또는 제3자가 거주하는 곳의 동 이하 상세 주소를 의미하며, 아파트 및 거주 건물명 항목은 대화 참여자 또는 제3자가 거주하는 아파트 이름이나 구체적인 건물 이름을 의미한다. 거주지 역명 항목은 대화 참여자 또는 제3자가 거주하는 지역의 대중교통 역 이름을 의미하며 이는 비식별화하지 않는다. 대화 참여자 또는 제3자가 비정기적으로 방문하는 장소, 구체적인 상호명 등도 비식별화하지 않는다.

출신 및 소속 범주의 학교 항목은 대화 참여자나 대화 참여자 가족의 출신 또는 소속 학교명을 의미하며 유치원과 어린이집, 학원 등의 교육 기관도 포함한다. 직장 항목은 대화 참여자나 대화 참여자 가족의 출신 또는 소속 직장명을 의미한다. 출신 및 소속 부대도 마찬가지 기준으로 비식별화한다.

이러한 항목 이외에도 특수한 직업이나 특이 질병이나 이력, 경력 등 대화 참여자나 제3자의 신원을 파악하는 것이 가능한 특수한 경우도 비식별화한다.

이후 이름 항목은 대화에 등장한 사람을 구분 가능하도록 '이름1', '이름2', '이름3'과 같이 작업하는 것과 '번호' 항목을 세분화하여 고유 식별 번호와 전화번호, 금융 관련 번호를 구분할 수 있도록 지침을 수정하였다. 2차 수정 지침은 <표 15>와 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[49]	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	비식별화	₩기타₩

<표 15> 비식별화 2차 수정 지침

3.1.2. 특수 메시지 처리

카카오톡과 라인 등의 메신저는 대화 참여자 사이의 대화뿐만 아니라 파일이나 정보 공유, 선물 발송, 송금, 무료 통화 등의 다양한 기능을 지원한다. 이러한 기능은 대화 참여자의 직접적인 발화가 아니기 때문에 말뭉치로 구축한 이후 걸러내기 등이 용이하여야 한다. 또한 이모티콘이나 사진, 동영상 등과 같이 텍스트로 추출하였을 때 그 내용을 확인하는 것이 불가능한 경우도 마찬가지이다. 그리고 서버 자체에서 발신한 시스템 메시지 등도 대화 내용 인식이나 처리와는 무관한 항목이므로 특수 메시지 처리 항목으로 분류하였다.

특수 메시지 항목의 분류는 <표 16>과 같다.

범주	항목	텍스트 파일 표기 예시
감정 및 상태 표현	이모티콘, 스티커[50]	이모티콘 스티커
	메신저 기본 이모티콘	(하트뿅)(하하)(우와)(심각)(힘듦)
	키보드별 기본 이모지	🙌 ▨ ❤ 🍾
시스템 메시지	선물하기	***님이 선물과 메시지를 보냈습니다. ***님의 "카페아메리카노 Tall"선물에 감동했어요.
	무료 통화	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 : 2019/10/21 23:33까지
	공지 등록	톡게시판 '공지': 12월 31일 연말 모임 내용 확인
	지도 공유	지도: 경기 성남시 분당구 정자동 ***
	연락처 공유	연락처: *** 작가님
	메시지 삭제	삭제된 메시지입니다.
시스템 메시지 (서버 발신)	대화방 나감	***님이 나갔습니다.
	대화방 들어옴	***님이 들어왔습니다.
	대화방 초대	***님이 ***님을 초대했습니다.

49) URL 항목은 2차 수정 지침에서는 특수 메시지 항목으로 분류하여 처리하였다.
50) 카카오톡은 이모티콘, 라인은 스티커라는 명칭을 사용한다.

범주	항목	텍스트 파일 표기 예시
콘텐츠 공유	사진/GIF	사진, 사진 n장
	동영상	동영상
	뮤직	'슬픈 운명 (Feat. Lexy, 황성환)-윤희중' 음악을 공유했습니다.
	파일	파일: 04 Beethoven_ Piano Sonata #14 In C.m4a
	게임	[딩동! 초대장이 도착했어요~] 한 번 시작하면 멈출 수 없는 마성의 바로 그 게임, 프렌즈마블로 당신을 초대해요!
	음성 메시지	음성 메시지
정보 공유	샵 검색	샵검색: #무간도
	블로그, 카페 등 게시글 공유	다음카페] [어쩌다 발견한 하루] 본인들보다 30센치 작은 여주 놀리는 남주들.jpgif
	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 모바일 추가할인 혜택!!
	오픈 채팅 초대	카카오톡 오픈채팅을 시작해 보세요. 링크를 선택하면 카카오톡이 실행됩니다.
	배송 안내	[Web발신] [반품]안녕하세요. *** 고객님 쿠팡맨 ***입니다. 요청하신 반품 회수를 금일 진행할 예정입니다.

<표 16> 메신저 대화 특수 메시지 항목

이모티콘은 크게 세 가지 항목으로 분류가 가능하다. 먼저 이모티콘 또는 스티커는 메신저에서 제공하는 동적인 캐릭터 요소로, 텍스트로 추출할 경우 '이모티콘'으로 표기되고 어떤 내용인지 확인이 불가능하다. 두 번째는 메신저가 제공하는 기본 이모티콘으로, 첫 번째 요소와 달리 캐릭터의 움직임이 없다. 기본 이모티콘은 그 의미 정보가 고정되어 있어 텍스트로 추출하면 '(기쁨)', '(하트뿅)'과 같은 형식으로 표기가 된다. 세 번째는 스마트폰이나 태블릿에 자체적으로 내장된 키보드가 제공하는 기본 이모지이다. 이 항목은 텍스트로 추출하면 글꼴의 지원 여부에 따라 '🎧 ▨ ❤ 🍾 ' 와 같은 이미지로 표기되거나 지원하지 않는 글꼴에서는 '□□□' 등으로 표기되기도 한다.

시스템 메시지는 메신저가 지원하는 기능을 실행하였을 때 표기되는 메시지로서 선물하기와 송금 등 메시지의 발화 주체가 대화 참여자인 경우와 대화방 나감, 들어옴, 대화방 초대 등 메시지의 발화 주체가 대화 참여자가 아닌 경우로 구분할 수 있다.

그 외 다양한 유형의 콘텐츠를 공유하거나 온라인의 유용한 정보나 서비스 관련 정보를 공유하는 메시지도 메신저에서 나타나는 특수 메시지에 해당한다.

이러한 특수 메시지는 대화 내용에 포함되지만, 이모티콘이나 사진, 동영상 등과 같이 내용 인식이 불가하거나 대화 참여자가 직접 발화한 것이 아니기 때문에 대화 인식과 처리에는 불필요한 요소이며, 원칙적으로 삭제하거나 태깅을 통해 걸러내기가 용이한 형태로 가공이 이루어져야 한다.

3.2. 추출 형식 표준화

메신저 대화는 사용한 기기와 운영 체제의 종류에 따라 텍스트로 추출했을 때 형식이 다르다. 대표적으로 카카오톡의 운영 체제별 추출 형식은 [그림 14], [그림 15], [그림 16]과 같다. 그리고 이를 비교하면 <표 17>과 같다[51].

```
P2 님과 카카오톡 대화
저장한 날짜 : 2019년 7월 30일 오후 3:40

2019년 7월 30일 오전 9:56, P1 : 안녕하세요?
2019년 7월 30일 오전 9:56, P2 : 안뇽하세요
2019년 7월 30일 오전 9:56, P2 : 오늘 날이 흐리네요 ㅜㅜ
2019년 7월 30일 오전 9:56, P1 : 처음 뵙겠습니다
2019년 7월 30일 오전 9:56, P2 : 우산 안가지고 왔는데...
2019년 7월 30일 오전 9:56, P2 : 네 처음뵙겠습니다 ㅋㅋ
2019년 7월 30일 오전 9:56, P1 : 저는 가지고 왔어요
2019년 7월 30일 오전 9:56, P2 : 오 준비성이 철저하시네요
```

[그림 14] 카카오톡 안드로이드 운영 체제 추출 형식 예시

안드로이드 운영 체제에서 카카오톡 추출 형식은 'YYYY년 MM월 DD일' '발화 시간(12시간 체계), 발화자 : 발화 내용' 형식으로 표기가 된다.

```
P2 님과 카카오톡 대화
저장한 날짜 : 2019-08-14 13:25:39


--------------- 2017년 10월 29일 일요일 ---------------
[P2] [오후 10:10] 저 소가 무슨 소인지가 궁금했음 ?
[P1] [오후 10:23] ㅇㅇㅇㅇㅇ
[P1] [오후 10:23] 존나 궁금함
[P1] [오후 10:23] 작을 소냐
[P1] [오후 10:25] 아 알려줘ㅓ
```

[그림 15] 카카오톡 PC 운영 체제 추출 형식 예시

51) 사용 언어가 외국어로 설정된 형식, 맥(mac) 운영 체제에서 추출한 파일 등 기타 형식은 세 가지 기준 형태 중에서 가장 변환이 쉬운 형식으로 바꾸는 작업이 먼저 이루어진 후 표준화하였다.

카카오톡의 PC 운영 체제 추출 형식은 연월일과 요일 구분자가 '--------------- YYYY년 MM월 DD일 요일 ---------------'형식으로 상단에 표시가 되며, '[발화자] [발화 시간(12시간 체계)] 발화 내용' 형식으로 표기가 된다.

P2 님과 카카오톡 대화
저장한 날짜 : 2019. 7. 30. 오전 11:02

2019년 7월 30일 화요일
2019. 7. 30. 오전 10:22, P1 : 부먹? 찍먹?
2019. 7. 30. 오전 10:22, P1 : 뭐가 더 좋으세요 ㅎㅎㅎ
2019. 7. 30. 오전 10:22, P2 : 전 솔직히 둘다 좋아해요
2019. 7. 30. 오전 10:22, P1 : 헐 저도요!
2019. 7. 30. 오전 10:23, P2 : 탕수육은 먹는것만으로
2019. 7. 30. 오전 10:23, P1 : 그냥 같이 먹는 사람 취향대로 맞춰주는 편이에요 ㅋㅋㅋㅋㅋ
2019. 7. 30. 오전 10:23, P2 : 내 눈앞에 존재해주는것만우로
2019. 7. 30. 오전 10:23, P1 : 둘 다 맛있음 ^___^
2019. 7. 30. 오전 10:23, P2 : 저두요. 같이 먹는 사람에 따라
2019. 7. 30. 오전 10:23, P2 : 탕수육은 그냥 무조건 좋아해서요

[그림 16] 카카오톡 iOS 운영 체제 추출 형식 예시

카카오톡의 iOS 운영 체제 추출 형식은 연월일과 요일 구분자가 상단에 표시가 되는 것은 PC 형식과 동일하나, '--------------- ' 구분선이 없는 점이 다르다. 'YYYY. MM. DD. 시간(12시간), 발화자 : 발화' 형식으로 표기가 된다.

	안드로이드	PC	iOS
날짜 구분자	없음	있음(구분선 포함) YYYY년 MM월 DD일 요일	있음(구분선 없음) YYYY년 MM월 DD일 요일
발화 날짜 표기	YYYY년 MM월 DD일	없음	YYYY. MM. DD.
발화 시간 표기	12 시간 체계(오전/오후) [ ] 사용하지 않음	12시간 체계(오전/오후) [ ] 안에 시간 기재	12시간 체계(오전/오후) [ ] 사용하지 않음
발화자 표기	[ ] 사용하지 않고 대화명 기재	[ ] 안에 대화명 기재	[ ] 사용하지 않고 대화명 기재

<표 17> 카카오톡의 텍스트 추출 형식 비교

말뭉치 구축에 앞서 이러한 다양한 형식을 <표 18>과 같이 표준화하였다.

	표준화 결과
날짜 구분자	사용하지 않음
발화 날짜 표기	YYYYMMDD [ ] 사용하지 않음
발화 시간 표기	24시간 체계 [ ] 사용하지 않음
발화자 표기	대화 참여자 대화명(P1, P2 변환) 기재 [ ] 사용하지 않음

<표 18> 원시 말뭉치의 날짜, 시간, 발화자 표준화 결과

표준화 작업은 자체 말뭉치 가공 시스템[52]을 활용하였다. 이를 통해 원문 텍스트 형식의 인식과 표준화한 형태로의 변환이 자동으로 이루어졌다.

3.3. 파일 형식 및 마크업 지침

수집 자료의 선별과 대화 정제, 형식 표준화의 단계를 거친 메신저 대화 자료는 해당 대화의 메타 정보를 헤더로 부착한 후 태깅 지침에 따라 마크업이 이루어졌다.

3.3.1. 파일 형식

메신저 대화 말뭉치의 파일명은 말뭉치의 유형, 매체 및 장르 구분, 분석 층위, 구축년도, 일련번호 항목에 따라 부여하였으며, 이는 <표 19>와 같다.

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축년도	8자리 일련번호
M : 메신저 말뭉치	D : 2인 대화 M : 다자 대화	OR : 원문 자료 RW : 원시 말뭉치	19	00000001~ 99999999

<표 19> 메신저 말뭉치 파일명 부여 방식

수집 원문 자료는 텍스트 파일(.txt) 형식으로 저장[53]하였고, 원시 말뭉치는 XML 형식을 기본으로 하여 주관 기관이 요청한 SJML[54] 형식으로 작성하고, 추가로 최근의 개발 환경에서 널리 활용되고 있는 JSON(JavaScript Object Notation)[55] 형식으로도 작성

52) 본 사업 기간 동안 수집 사이트와 함께 대규모의 메신저 대화 자료의 표준화, 마크업, 태깅 등 말뭉치 구축이 자동으로 이루어지는 시스템을 구축하여 활용하였다.
53) 수집 파일의 텍스트 작업은 윈도우에서 기본으로 제공하는 메모장 또는 Notepad++ 프로그램으로 이루어졌다.
54) SJML 형식은 국립국어원(2018)에서 정의한 형식을 준용하였다.
55) JSON은 XML에 비해 유연하게 속성을 추가할 수 있고 형식과 체계가 간결하다는 장점과

하였다. 파일명 부여 방식에 따른 메신저 대화 원문과 말뭉치 파일명의 작성 예시는 <표 20>과 같다.

파일명	설명
MDOR1900000001.txt	메신저 2인 대화 원문 자료 1번째 파일
MDRW1900000001.SJML	메신저 2인 대화 원시 말뭉치 1번째 파일 XML 형식
MMRW1900000001.SJML	메신저 다자 대화 원시 말뭉치 1번째 파일 XML 형식
MDRW1900000001.JSON	메신저 2인 대화 원시 말뭉치 1번째 파일 JSON 형식
MMRW1900000001.JSON	메신저 다자 대화 원시 말뭉치 1번째 파일 JSON 형식

<표 20> 메신저 대화 원문과 말뭉치 파일명 작성 예시

수집한 파일은 ANSI와 UTF-8(BOM) 등으로 인코딩이 된 경우도 있었으나, 이는 수집 파일을 정제하는 작업 후에 일괄 UTF-8로 변환하였다.

3.3.2. 마크업 지침

메신저 대화 말뭉치의 SJML과 JSON 형식 모두 헤더(header)와 본문(text)을 기본 구조로 한다. 헤더에는 파일명, 제목 등 파일에 대한 정보와 성별, 연령, 지역, 직업 등 대화 참여자 정보 및 대화 참여자 간 관계, 친밀도, 대화 주제 등 대화 유형에 대한 메타 정보를 기재하고, 본문에는 발화 시간, 발화자, 발화 내용 등 대화 참여자가 주고받은 실제 대화 내용을 기재한다. 메신저 대화 말뭉치의 기본적인 형식과 구조는 [그림 17]과 같다[56].

메신저 대화에는 사진이나 동영상, 이모티콘 등의 멀티미디어 요소를 비롯하여 선물 보내기 등의 기능 요소와 URL 등의 각종 정보 공유 요소 등이 다수 포함되어 있다. 자연어 처리 등에 활용하기 위해서는 이러한 요소는 '걸러내기(filtering)'나 '바꾸기(replacing)'를 통해 일반적인 발화와 구분하기 용이한 형태로 가공해야 한다.

보안상의 강점 등으로 최근 개발 환경에서 XML의 대안으로 널리 활용되고 있다.

56) 개별 태그의 의미에 대해서는 뒤에서 상세히 설명한다.

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
  <header>
    <fileInfo>
      <fileId></fileId>
      <annoLevel></annoLevel>
      <class></class>
    </fileInfo>
    <sourceInfo>
      <title></title>
      <author></author>
      <publisher></publisher>
      <msg></msg>
      <topic></topic>
    </sourceInfo>
    <profileInfo>
      <personId></personId>
      <setting></setting>
    </profileInfo>
  </header>
  <text>
    <u></u>
  </text>
</SJML>
```

```
{
  "SJML" : {
    "header" : {
      "sourceInfo" : {
        "msg" :
        "author" :
        "publisher" :
        "topic" :
        "title" :
      },
      "profileInfo" : {
        "personId" : [ {
        } ],
        "setting" : {
        }
      },
      "fileInfo" : {
        "annoLevel" :
        "class" :
        "fileId" :
      }
    },
    "text" : {
      "u" : [ {
      } ]
```

[그림 17] 메신저 대화 말뭉치 SJML과 JSON 형식의 기본 구조

3.3.2.1. 헤더 항목 마크업

원시 말뭉치에 헤더(header)로 부착되는 파일 정보와 대화의 메타 정보는 <표 21>의 지침에 따라 메신저 말뭉치 가공 시스템을 통해 자동으로 태깅이 이루어졌다.

태그				설명	유형
header	fileInfo	fileId		파일명	M D/M RW 19 00000001~99999999
		annoLevel		주석 수준	원시
		class		장르	2인 대화/다자 대화(X명)
	source Info	title		제목	2인 메신저 대화 00000001~99999999/ 다자 메신저대화00000001~99999999
		author		대화 참여자	대화 참여자
		publisher		매체 유형	메신저 대화 수집
		topic		대화 주제	'메타 정보_주제' 참고
		msg		메신저 종류	카카오톡/라인/기타
	profile Info	person Id	id	대화 참여자 아이디	P1, P2.....
			sex	성별	M/F
			age	연령	(만)14~
			occupation	대화 참여자 직업	경영,관리직/전문가 및 관련 종사자/ 사무종사자/서비스종사자/ 판매,영업종사자/농업,임업,어업종사자/ 기능원및관련기능종사자/ 기술자종사자(장치/기계조작및조립)/ 단순노무종사자/군인/학생 무직,취업준비생/주부/기타
			bplace	대화 참여자 출생지	서울/경기/인천/대전/세종/충북/충남/ 대구/경북/부산/경남/울산/광주/전북/ 전남/ 강원/제주/해외, 기타
			gplace	대화 참여자 주 성장지	
			city	대화 참여자 현 거주지	
			device	기기 유형	폰/태블릿/PC
			key	사용 자판 종류	2벌식(쿼티)/천지인/ 천지인플러스/나랏글/단모음/딩굴/모아키/ 밀기글/베가(VEGA)/기타
		setting	relation	대화 참여자 간 관계	'메타 정보_관계' 참고
			level	대화 참여자 간 친밀도	1(친밀도가 낮다)~5(친밀도가 높다) 0(낯선 관계)
			freq	대화(연락) 빈도	거의 매일 주3회 이상 주1~2회 주1회 미만 월1회 미만 처음

<표 21> 헤더 항목의 마크업 지침

헤더 항목의 마크업 지침을 반영한 헤더 예시는 [그림 18], [그림 19]와 같다.

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
    <header>
        <fileInfo>
            <fileId>MDRW1900000008</fileId>
            <annoLevel>원시</annoLevel>
            <class>2인 대화</class>
        </fileInfo>
        <sourceInfo>
            <title>2인 메신저 대화00000008</title>
            <author>대화 참여자</author>
            <publisher>메신저 대화 수집</publisher>
            <msg>카카오톡</msg>
            <topic>식음료 (식사, 음식, 배달, 맛집, 요리)</topic>
        </sourceInfo>
        <profileInfo>
            <personId age="28" bplace="경기" city="경기" device="스마트폰" gplace="경기
" key="2벌식(쿼티)" occupation="가정 주부" sex="F">P1</personId>
            <personId age="27" bplace="서울" city="경기" device="스마트폰" gplace="경기
" key="천지인" occupation="기타" sex="F">P2</personId>
            <setting freq="(거의) 매일 연락한다." level="5.0" relation="학교/학원 : 동기/동
창/동급생">P1-P2</setting>
        </profileInfo>
    </header>
```

[그림 18] SJML 형식의 헤더 예시

```
{
  "SJML" : {
    "header" : {
      "sourceInfo" : {
        "msg" : "카카오톡",
        "author" : "대화 참여자",
        "publisher" : "메신저 대화 수집",
        "topic" : "식음료 (식사, 음식, 배달, 맛집, 요리)",
        "title" : "2인 메신저 대화00000008"
      },
      "profileInfo" : {
        "personId" : [ {
          "gplace" : "경기",
          "bplace" : "경기",
          "occupation" : "가정 주부",
          "city" : "경기",
          "sex" : "F",
          "device" : "스마트폰",
          "age" : 28,
          "key" : "2벌식(쿼티)",
          "content" : "P1"
        }, {
          "gplace" : "경기",
          "bplace" : "서울",
          "occupation" : "기타",
          "city" : "경기",
          "sex" : "F",
          "device" : "스마트폰",
          "age" : 27,
          "key" : "천지인",
          "content" : "P2"
        } ],
        "setting" : {
          "level" : 5,
          "freq" : "(거의) 매일 연락한다.",
          "content" : "P1-P2",
          "relation" : "학교/학원 : 동기/동창/동급생"
        }
      },
      "fileInfo" : {
        "annoLevel" : "원시",
        "class" : "2인 대화",
        "fileId" : "MDRW1900000008"
      }
    }
```


[그림 19] JSON 형식의 헤더 예시

3.3.2.2. 본문 마크업

메신저 대화의 내용을 담고 있는 본문은 표준화 형식과 <표 22>와 같은 지침에 따라 메신저 말뭉치 자동 구축 시스템을 통한 마크업이 자동으로 이루어졌다.

태그			항목	예시
text	u	n	발화 번호	n="1"
		date	발화 날짜	date="20191104"
		time	발화 시간	time="15:30"
		who	발화자	who="P2"
		|	줄 바꿈[57]	너거기절대가지마 <|/>나알지?수정할수잇는거면내가 뭐라뭐라햇겟지

<표 22> 원시 말뭉치 본문 마크업 지침

본문 마크업 지침을 반영한 예시는 [그림 21], [그림 22]와 같다.

2019-11-04 15:30:00 , P2 : 짜쟌
2019-11-04 15:30:00 , P1 : ㅋㅋㅋ
2019-11-04 15:30:00 , P1 : 오늘 나는 아침에 8시에나와서 두유만먹어가지구
2019-11-04 15:30:00 , P2 : 예스리 끝낫웅?
2019-11-04 15:30:00 , P2 : 아이구 힘들었겟다ㅠㅠ
2019-11-04 15:30:00 , P1 : 아까 쉬는시간에잠깐 삼김 2개 겨우먹음
2019-11-04 15:31:00 , P1 : ㅋㄷㅋㄷ

[그림 20] 수집 원문 예시

```
<text>
        <u date="20191104" n="1" time="15:30" who="P2">짜쟌</u>
        <u date="20191104" n="2" time="15:30" who="P1">ㅋㅋㅋ</u>
        <u date="20191104" n="3" time="15:30" who="P1">오늘 나는 아침에 8시에나와서
두유만먹어가지구</u>
        <u date="20191104" n="4" time="15:30" who="P2">예스리 끝낫웅?</u>
        <u date="20191104" n="5" time="15:30" who="P2">아이구 힘들었겟다ㅠㅠ</u>
        <u date="20191104" n="6" time="15:30" who="P1">아까 쉬는시간에잠깐 삼김 2개
겨우먹음</u>
        <u date="20191104" n="7" time="15:31" who="P1">ㅋㄷㅋㄷ</u>
```

[그림 21] SJML 형식 본문 마크업 예시

57) 하나의 발화 안에서 줄 바꿈이 있는 경우에 <|/> 태그를 사용하였다.

```
"text" : {
  "u" : [ {
    "date" : 20191104,
    "time" : "15:30",
    "n" : 1,
    "content" : "짜쟌",
    "who" : "P2"
  }, {
    "date" : 20191104,
    "time" : "15:30",
    "n" : 2,
    "content" : "ㅋㅋㅋ",
    "who" : "P1"
  }, {
    "date" : 20191104,
    "time" : "15:30",
    "n" : 3,
    "content" : "오늘 나는 아침에 8시에나와서 두유만먹어가지구",
    "who" : "P1"
  }, {
    "date" : 20191104,
    "time" : "15:30",
    "n" : 4,
    "content" : "예스리 끝낫웅?",
    "who" : "P2"
  }, {
    "date" : 20191104,
    "time" : "15:30",
    "n" : 5,
    "content" : "아이구 힘들었겟다ㅠㅠ",
    "who" : "P2"
  }, {
    "date" : 20191104,
    "time" : "15:30",
    "n" : 6,
    "content" : "아까 쉬는시간에잠깐 삼김 2개 겨우먹음",
    "who" : "P1"
  }, {
    "date" : 20191104,
    "time" : "15:31",
    "n" : 7,
    "content" : "ㅋㄷㅋㄷ",
    "who" : "P1"
```


[그림 22] JSON 형식 본문 마크업 예시

3.3.2.3. 비식별화 항목 마크업

지침에 따라 비식별화가 이루어진 요소는 메신저 말뭉치 가공 시스템을 통해 <표 23>과 같은 지침에 따라 지정된 마크업 기호로 자동 변환이 이루어졌다.

범주	항목	원문 표기	마크업
이름	실명	P1 - ₩이름1₩, P2 - ₩이름2₩ 그 외는 등장 순서에 따라 ₩이름n₩	<anon type="name" n="1"> <anon type="name" n="2"> <anon type="name" n="n">
	실명(변형)		
	특수 애칭, 별명, 대화명, 필명		
온라인	아이디	₩계정₩	<anon type="account">
	이메일 주소		
각종 번호 및 비밀번호	고유 식별 번호	₩신원₩	<anon type="social-security-num"/>
	전화번호	₩전번₩	<anon type="tel-num"/>
	금융 번호	₩금융₩	<anon type="card-num"/>
	일련번호	₩번호₩	<anon type="num"/>
	(구매자) 식별 번호		
	사업자 등록 번호		
	비밀번호		
장소	상세 주소	₩주소₩	<anon type="address"/>
	아파트 및 거주 건물명		
출신 및 소속	출신 및 소속 학교	₩소속₩	<anon type="affiliation"/>
	출신 및 소속 직장		
	출신 및 소속 부대		
기타	위에서 언급하지 않은 항목	₩기타₩	<anon type="others"/>

<표 23> 비식별화 항목의 원문 표기와 마크업 기호 비교

비식별화가 반영된 원시 말뭉치 예시는 [그림 24]와 같다.

2019-11-04 22:15:00 , P2 : 내가₩이름1₩이 울집에놀러오면데러가께ㅠㅠ
2019-11-04 22:15:00 , P1 : 방금얘기끝난거아녓어 ?
2019-11-04 22:15:00 , P2 : 아가보러ㅠㅠ5시이전에 가야되니께ㅠㅠ
2019-11-04 22:15:00 , P1 : 내가한번 가고혼자놀러가고
2019-11-04 22:15:00 , P2 : 맘편히 못놀러갈까봐ㅠㅠ
2019-11-04 22:15:00 , P2 : 아아
2019-11-04 22:15:00 , P1 : ₩이름4₩데리고는
2019-11-04 22:15:00 , P1 : ₩이름3₩랑간다구

[그림 23] 수집 원문 비식별화 예시

<u date="20191104" n="793" time="22:15" who="P2">내가<anon type="name" n="1"/>이 울집에놀러오면데러가께ㅠㅠ</u>
<u date="20191104" n="794" time="22:15" who="P1">방금얘기끝난거아녓어 ?</u>
<u date="20191104" n="795" time="22:15" who="P2">아가보러ㅠㅠ5시이전에 가야되니께ㅠㅠ</u>
<u date="20191104" n="796" time="22:15" who="P1">내가한번 가고혼자놀러가고</u>
<u date="20191104" n="797" time="22:15" who="P2">맘편히 못놀러갈까봐ㅠㅠ</u>
<u date="20191104" n="798" time="22:15" who="P2">아아</u>
<u date="20191104" n="799" time="22:15" who="P1"><anon type="name" n="4"/>데리고는</u>
<u date="20191104" n="800" time="22:15" who="P1"><anon type="name" n="3"/>랑간다구</u>

[그림 24] SJML 형식 비식별화 마크업 예시

3.3.2.4. 특수 메시지 항목 마크업

특수 메시지 항목 중 유형 지정이 가능한 항목[58]은 메신저 말뭉치 가공 시스템을 통해 <표 24>와 같은 지침에 따라 지정된 마크업 기호로 자동 변환이 이루어졌다.

유형	항목	텍스트 파일 표기	마크업
감정 및 상태 표현	이모티콘, 스티커	이모티콘 이모티콘 : 즐거움[59]	<emoji/> <emoji>#즐거움</emoji>
	메신저 기본 이모티콘	(하트뿅)	<emoji>(하트뿅)</emoji>
	키보드별 기본 이모지	🎧 ▨ ❤	<emoji>🎧 ▨ ❤ </emoji>
시스템 메시지	선물하기	***님이 선물과 메시지를 보냈습니다. ***님의 "카페아메리카노 Tall"선물에 감동했어요.	<message type="gift"/>
	무료 통화	보이스톡 해요/페이스톡 해요 보이스톡 취소/페이스톡 취소 보이스톡 응답없음/페이스톡 응답없음 보이스톡 부재중/페이스톡 부재중 보이스톡 0:49/페이스톡 0:12	<message type="call"/>
	송금	000 님이 돈을 보냈어요! - 받는 사람 :*** - 받을 금액 : 20,000원 - 입금 기한 : 2019/10/21 23:33까지	<message type="money"/>
	공지 등록	톡게시판 '공지': 12월 31일 연말 모임 내용 확인	<message type="notice"/>
	지도 공유	지도: 경기 성남시 분당구 정자동 ***	<message type="map"/>
	연락처 공유	연락처: *** 작가님	<message type="contact"/>
	메시지 삭제	삭제된 메시지입니다.	<message type="delete"/>
시스템 메시지 (서버 발신)	대화방 나감	***님이 나갔습니다.	원시 말뭉치에서는 삭제
	대화방 들어옴	***님이 들어왔습니다.	
	대화방 초대	***님이 ***님을 초대했습니다.	

58) 특수 메시지는 그 형식이 다양하여 유형 지정이 불가능한 경우도 다수이다. 본 사업 기간에는 유형이 정형화되어 유형 지정이 가능한 항목으로 한정하여 특수 메시지의 자동 변환을 실시하였다. 앞으로 정형화가 불가능한 특수 메시지 항목의 처리를 위한 작업 지침도 마련할 필요가 있다.

59) 대화 제공자가 이모티콘에 감정 등을 표기한 경우이다.

60) 정보 공유 항목 중 URL이 노출된 항목은 URL 공유 항목과 동일하게 마크업 처리한다.

유형	항목	텍스트 파일 표기	마크업
콘텐츠 공유	사진/GIF	사진, 사진 n장	<message type="photo"/>
	동영상	동영상	<message type="video"/>
	뮤직	'슬픈 운명 (Feat. Lexy, 황성환)-윤희중' 음악을 공유했습니다.	<message type="music"/>
	파일	파일: 04 Beethoven_ Piano Sonata #14 In C.m4a	<message type="file"/>
	음성 메시지	음성 메시지	<message type="voice"/>
URL 공유	URL 등 링크 공유	https://www.youtube.com/watch?v=mT-iNAUqfQc	<message type="url"/>
정보 공유[60]	샵 검색	샵검색: #무간도	<message type="info"/>
	블로그, 카페 등 게시글 공유	[다음카페] [어쩌다 발견한 하루] 본인들보다 30센치 작은 여주 놀리는 남주들.jpgif	
	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 모바일 추가할인 혜택!!	
	배송 안내	[Web발신] [반품]안녕하세요. *** 고객님 쿠팡맨 ***입니다. 요청하신 반품 회수를 금일 진행할 예정입니다.	

<표 24> 특수 메시지 항목의 텍스트 표기와 마크업 기호 비교

특수 메시지 항목의 마크업이 반영된 원시 말뭉치 예시는 [그림 26]과 같다.

[P2] [오후 9:56] 사진
[P2] [오후 9:56] 2퓨리☆
[P2] [오후 9:56] ㅠ 앱이라서
[P2] [오후 9:56] 링크쨩이 없나
[P2] [오후 9:56] 찾아봄
[P2] [오후 9:56] http***
[P2] [오후 9:56] 찹랏당☆
[P1] [오후 9:56] 오 ㄱㅅㄱㅅ
[P1] [오후 9:59] 사진
[P1] [오후 9:59] 이모티콘

[그림 25] 수집 원문 특수 메시지 예시

```
<u date="20191017" n="2479" time="21:56" who="P2"><message type="photo"/></u>
<u date="20191017" n="2480" time="21:56" who="P2">2퓨리<emoji>☆</emoji></u>
<u date="20191017" n="2481" time="21:56" who="P2">ㅠ 앱이라서</u>
<u date="20191017" n="2482" time="21:56" who="P2">링크쨩이 없나</u>
<u date="20191017" n="2483" time="21:56" who="P2">찾아봄</u>
<u date="20191017" n="2484" time="21:56" who="P2"><message type="url"/></u>
<u date="20191017" n="2485" time="21:56" who="P2">찯랏당<emoji>☆</emoji></u>
<u date="20191017" n="2486" time="21:56" who="P1">오 ㄱㅅㄱㅅ</u>
<u date="20191017" n="2487" time="21:59" who="P1"><message type="photo"/></u>
<u date="20191017" n="2495" time="21:59" who="P1"><emoji/></u>
```

[그림 26] SJML 형식 특수 메시지 항목 마크업 예시

4. 검수

1차 구축이 완료된 메신저 대화 말뭉치는 지침에 따라 정확하게 구축이 되었는지, 규격에 맞게 일관된 형식을 갖추고 있는가를 기준으로 검수가 이루어졌다.

4.1. 비식별화 항목 검수

메신저 대화 말뭉치를 법적인 제한 없이 활용하기 위해서는 철저한 비식별화가 이루어져야 한다. 비식별화의 중요성을 고려하여 비식별화 검수는 3단계로 이루어졌다.

먼저 구축한 말뭉치 파일을 대상으로 소프트웨어를 활용[61]하여 유형 식별이 가능한 전화번호, 주민 등록 번호, 계좌 번호, 여권 번호, 운전면허 번호, 이메일 주소 등에서 비식별화가 누락된 항목을 검출하였다. 그리고 누락된 항목이 검출된 파일을 1차 수정하였다.

1차 수정 파일을 포함한 전체 말뭉치 파일은 전체 작업자 간 교차 검수를 통하여 2차 수정이 이루어졌다. 교차 검수를 통해 비식별화가 누락된 항목을 다시 비식별화하였고, 과도하게 비식별화가 이루어진 경우[62]에는 비식별화 작업 이전 수집 파일을 기준으로 해당 항목을 복원하였다.

2차 수정 파일은 검수 담당 작업자가 검수하여 비식별화가 누락된 항목을 중심으로 3차 수정이 이루어졌다.

4.2. 말뭉치 형식 검수

파일명 부여 방식과 인코딩을 비롯하여 가공 시스템을 통해 이루어진 기계적 마크업

61) 에스원PS 프로그램에 포함되어 있는 민감 정보 검출 검사 기능을 활용하였다.
62) 텍스트 편집 프로그램의 '모두 바꾸기' 기능을 활용하여 이름이나 대화명을 비식별화한 경우에 해당 내용을 포함하는 다른 문자열까지 일괄 바뀌는 사례가 있었다.

의 오류를 검수하였다.

4.2.1. 파일명 및 인코딩 형식 검수

메신저 대화 말뭉치의 파일명은 가공 시스템에서 사전 설정된 파일명 부여 규칙에 따라 자동으로 부여하였다. 말뭉치 구축 초기에 나타난 일부 파일명 오류[63]를 수정하여 가공 시스템에 반영하였다.

말뭉치 파일의 인코딩이 UTF-8(BOM)과 ANSI로 저장된 경우가 일부 있었다. 이는 소프트웨어를 사용하여 전체 말뭉치 파일을 일괄 UTF-8로 변환[64]하여 수정하였다.

4.2.2. 마크업 항목 검수

구축한 메신저 대화 말뭉치 파일의 일부를 선정하여 가공 시스템에서 사전 설정된 마크업 지침에 따라 태그가 잘 부착되어 있는가를 검수하였다.

마크업 지침에 맞지 않는 태그 오류나 형식의 오류가 발생한 원인은 크게 두 가지이다. 첫 번째 유형은 [그림 27]의 예시와 같이 가공 시스템의 마크업 규칙 설정 오류로 발생한 마크업 오류이다.

```
<u date="20191017" n="1533" time="23:21" who="P2">아 너무 자연스럽게 <emoji/>으로 손이 갔어</u>
<u date="20171017" n="3" time="22:28" who="P1"><emoji/>:좋아함</u>
```

[그림 27] 마크업 규칙 설정 오류로 인한 마크업 형식 오류 예시

첫 번째 줄은 발화자가 메신저의 '이모티콘' 기능을 사용한 경우에만 '<emoji/>'로 변환되어야 하는데 발화자가 직접 발화한 단어까지도 변환하여 발생한 오류이다. 두 번째 줄은'<emoji>#즐거움</emoji>'이 정상적인 형식이나, 가공 시스템의 규칙 설정 오류로 인하여 형식에 맞지 않는 태그가 발생한 오류이다. 이러한 유형은 가공 시스템의 규칙을 수정함으로써 오류를 해결하였다.

두 번째 유형은 [그림 28]의 예시와 같이 작업자의 오류로 인하여 발생한 마크업 오류이다.

```
<u date="20190318" n="14684" time="21:51" who="P1">₩ 금융₩ 카카오뱅크 <anon type="name" n="63"/></u>
<u date="20180621" n="10446" time="09:59" who="P2">근데 난 ₩주소에서 ₩주소잘 간듯ㅇㅇ</u>
```

[그림 28] 작업자 오류로 인한 마크업 형식 오류 예시

63) 파일명에 포함되는 말뭉치 구축 년도를 '20'으로 잘못 기재한 오류가 있었다.

64) 'BOM-Remover' 프로그램을 이용하여 UTF-8(BOM) 형식을 수정하였고, 'RedUTF8' 프로그램을 이용하여 ANSI 형식을 UTF-8 형식으로 일괄 변환하였다.

첫 번째 줄은 ‘<anon type="card-num"/>’로 변환되어야 하고, 두 번째 줄은 ‘<anon type="address"/>’로 변환되어야 한다. 비식별화 항목은 ‘₩문자₩’와 같은 형식으로 기재가 되어야 가공 시스템에서 지정된 마크업 형식으로 변환이 되는데, 작업자가 [그림 28]과 같이 지정된 형식과 다르게 표기하여 이와 같은 오류가 발생한 것이다. 이러한 오류는 전체 말뭉치 파일을 대상으로 동일 유형의 오류가 있는지를 검출하여[65], 오류가 검출된 파일은 개별 수정하였다.

65) 파일 전체 대상 오류 검출은 ‘TextCrawler’ 프로그램을 이용하였다.

제 3 장

메신저 대화 말뭉치 구축 결과

1. 메신저 대화 말뭉치의 구성

1.1. 구축 규모

메신저 대화의 분량은 생성된 발화와 말차례의 수를 기준으로 산정하였다.

먼저 발화의 분량을 산정하였는데, 전체 발화에서 사진이나 동영상 공유, 이모티콘을 단독으로 사용한 경우는 발화 분량 산정에서 제외하였다. 그리고 메신저에서 제공하는 파일 공유나 송금, 선물 보내기, 무료 통화 등의 특수 메시지도 발화 분량 산정에서 제외하였고, 대화방 초대나 대화방에서 나가기, 메시지 삭제 등의 시스템 메시지도 발화 분량 산정에서 제외하였다.

이들을 제외한 후 분량 산정 대상이 되는 발화를 대상으로 하여 말차례를 산정하였다. 화자 한 명의 최초 발화를 말차례가 시작되는 것으로 두고 이후 화자가 교체될 때마다 말차례 수가 증가하는 것으로 산정하였다.

그리고 산정한 말차례 수를 기준으로 말차례 10개를 대화 하나로 산정하였다[66].

기술한 분량 산정 기준에 따른 본 사업 기간 동안 구축한 메신저 대화 원시 말뭉치의 규모는 <표 25>와 같다.

항목	분량
수집 파일(참여자 집합) 수	7,395
대화(말차례 10개 기준) 수	712,291
말차례(turn) 수	7,122,919
발화 수	14,591,826

<표 25> 메신저 대화 원시 말뭉치의 분량 산정

1.2. 유형별 구성

1.2.1. 대화 참여 인원에 따른 구성

대화에 참여한 인원에 따라 상호 작용 양상과 대화 양상이 달라진다. 이러한 점을 고려하여 메신저 대화의 유형을 2인 대화와 다자 대화로 나눌 수 있다. 대화 참여 인

66) 메신저 대화를 대화 단위로 분할하기 위해서는 명확한 대화 분할 기준과 그러한 기준에 따른 분석의 과정이 필요하다. 본 사업에서는 기준에 따른 대화 분석과 분할의 과정은 포함하지 않았고, 분량 산정을 위하여 10개의 말차례를 하나의 대화 단위로 삼았다.

원을 기준으로 한 구축 분량은 <표 26>과 같다.

대화 참여 인원	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
2인	6,997	94.6	585,424	82.2	5,854,243	82.2	12,237,008	83.9
3인	135	1.8	54,379	7.6	543,799	7.6	1,085,202	7.4
4인	50	0.7	26,239	3.7	262,394	3.7	507,187	3.5
5인 이상	214	2.9	46,248	6.5	462,483	6.5	762,419	5.2
합계	7,395	100	712,291	100	7,122,919	100	14,591,826	100

<표 26> 대화 참여 인원별 구축 분량 및 비율

다자 대화의 수집 비율은 전체 대화 수 대비 17.8%이다[67]. 제공 파일 수를 기준으로 할 경우, 다자 대화는 전체 수집 파일의 5.4% 분량을 차지한다. 자료 수집 기간 동안 2인 대화 위주의 자료 수집을 한 것이 2인 대화가 파일 수에서 높은 비중을 차지하는 첫 번째 이유이다. 다만 대화 참여 인원에 대한 제한 없이 상시 모집하였던 일반 대화를 기준으로 할 경우에도 2인 대화 파일의 비율이 78%로 나타났다. 대화 참여자 전원으로부터 대화 제공에 대한 동의와 저작권 이용 허락 계약이 체결되어야 한다는 요건이 있었기 때문에 상대적으로 3인 이상의 다자 대화가 차지하는 비중이 낮게 나타난 것으로 볼 수 있다.

본 사업에서는 다자 대화의 참여 인원에 대한 별도의 제한을 두지 않았다. 참여 인원이 34명인 대화 파일이 본 사업 기간 동안 수집한 대화의 최대 참여 인원으로 나타났고, 30명인 대화 파일도 2개를 수집하였다.

그런데 대화 참여 인원이 많을수록 대화가 산발적으로 전개되어 대화의 구조나 양상을 이해하기 어렵기 때문에 메신저 언어에 대한 연구와 컴퓨터의 대화 양상을 학습하기 위한 자료라는 관점에서 본다면 다자 대화가 전체 대화에서 차지하는 비율과 최대 대화 참여 인원은 제한을 둘 필요가 있다.

1.2.2. 수집 방법에 따른 구성

앞서 본 사업 기간 동안 대화 수집 방법으로 활용한 '일반 대화 수집'과 '수집 봇 수집' 두 가지의 특성에 대해 설명하였다.

'수집 봇 수집'은 대화 제공자가 대화가 수집된다는 상황을 사전에 인식하고 통제된 상황에서 수집이 이루어진다. 따라서 대화 수집 상황이 전제되지 않은 일반 수집 방식의 대화와는 언어 사용이나 대화 양상의 차이가 있을 것으로 예상할 수 있다.

한편 수집 봇을 대화방에 초대한 시점부터의 대화를 수집하는 수집 봇 수집과 달리 일반 대화 수집은 대화방이 최초 만들어진 시점의 대화도 수집이 된다. 이로 인해 대

67) 다자 대화는 전체 대화의 10% 이상을 수집하는 것이 본 사업의 전제 조건 중 하나였다.

화를 제공하기 위해 메타 정보를 입력한 시점과 대화 시작 시점의 격차가 나타날 수 있고, 대화 참여자 간 관계와 같이 시간의 변화에 따라 달라지는 요소는 대화 내용과 관련성이 떨어질 수 있다[68].

이처럼 메신저 대화를 수집하는 방법에 따라 자료의 성격이 달라질 수 있다는 점을 고려해야 한다.

수집 방법을 기준으로 한 구축 분량은 <표 27>과 같다.

수집 방법	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
일반 수집	742	10.0	358,311	50.3	3,583,112	50.3	7,450,361	51.1
수집봇 수집	6,653	90.0	353,980	49.7	3,539,807	49.7	7,141,465	48.9
합계	**7,395**	**100**	**712,291**	**100**	**7,122,919**	**100**	**14,591,826**	**100**

<표 27> 수집 방법별 구축 분량 및 비율

말차례를 기준으로 할 경우 일반 수집 대화와 수집 봇 수집 대화의 분량은 큰 차이가 없다. 다만 수집 파일 수의 비율로 보면 수집 봇 수집 대화는 90%, 일반 수집 대화는 10%로 큰 차이를 보인다.

일반 수집 대화의 경우는 사업 시작 이전 과거의 대화도 제공이 가능하여 개별 파일의 대화나 발화의 분량이 상대적으로 많은 반면에, 수집 봇 수집 대화는 사업 기간 내 특정 모집 기간 동안 한시적으로 자료 수집이 이루어졌기 때문에 개별 파일의 대화나 발화의 분량이 상대적으로 적다. 그러한 개별 파일의 분량 차이에도 불구하고 수집 봇 수집 방식이 대화를 제공하는 방식이 용이했던 까닭에 수집 파일 수 기준으로는 높은 비중으로 나타나고 있다.

본 사업에서는 수집 방법의 차이를 메타 정보 항목에 별도로 기재하지 않았다. 그러나 수집 방법의 차이에 따라 자료의 성격이 달라진다는 점을 감안한다면 수집 방식을 달리 할 경우, 그러한 차이를 메타 정보로 기재하는 것도 고려할 필요가 있다.

1.2.3. 주제에 따른 구성

대화 주제의 사전 통제 여부에 따라 주제의 사전 통제가 이루어진 '주제 대화'와 사전 통제가 없는 '일상 대화' 두 유형으로 구분한다. 주제 대화와 일상 대화의 구축 분량은 <표 28>과 같다.

68) 서로 잘 모르는 관계에서 대화방을 개설하여 초반에 이루어진 대화는 낯선 관계의 대화로 볼 수 있으나, 시간의 흐름에 따라 메타 정보 기재 시점에 친밀한 관계로 바뀌게 된 경우에 참여자 간 친밀도가 4로 기재된 사례가 있었다.

주제	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
일상 대화	3,413	46.2	681,118	95.6	6,811,188	95.6	13,976,966	95.8
주제 대화	3,982	53.8	31,173	4.4	311,731	4.4	614,860	4.2
합계	**7,395**	**100**	**712,291**	**100**	**7,122,919**	**100**	**14,591,826**	**100**

<표 28> 대화 주제별 구축 분량 및 비율

수집 파일 수를 기준으로 하면 일상 대화와 주제 대화의 구축 비율은 46.2%와 53.8%로 주제 대화의 파일 수가 조금 더 높은 비중을 차지하나 큰 차이를 보이지는 않는다. 그러나 말차례를 기준으로 할 경우 주제 대화가 차지하는 비율은 전체의 일부분에 불과한 4.4%이다. 이는 대화가 이어지는 동안 하나의 주제를 일관성 있게 유지하는 것이 어렵다는 점을 고려했기 때문이다. 대화의 주제를 의도적으로 하나로 통제할 경우 부자연스럽고 인위적인 대화가 될 것이라는 점을 고려하여 주제 대화의 경우는 적은 분량으로 수집이 이루어졌다.

주제 대화의 세부 주제별 구축 분량은 <표 29>와 같다.

구분	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
개인 및 관계	525	13.2	1,677	5.4	16,771	5.4	32,584	5.3
주거와 생활	370	9.3	1,467	4.7	14,668	4.7	26,384	4.3
상거래 (쇼핑)	93	2.3	227	0.7	2,268	0.7	4,602	0.7
식음료	735	18.5	9,666	31.0	96,660	31.0	193,381	31.5
공공 서비스	2	0.1	4	>0.1	39	>0.1	78	>0.1
여가와 오락	558	14.0	5,735	18.4	57,347	18.4	110,666	18.0
일과 직업	276	6.9	1,045	3.4	10,449	3.4	20,242	3.3
행사 및 모임	19	0.5	42	0.1	427	0.1	808	0.1
미용과 건강	361	9.1	2,574	8.3	25,746	8.3	51,780	8.4
날씨와 계절	137	3.4	281	0.9	2,817	0.9	5,495	0.9
여행	531	13.3	6,825	21.9	68,256	21.9	134,313	21.8
교통	28	0.7	56	0.2	563	0.2	1,190	0.2

69) 여행, 주거와 생활 두 개의 주제를 중복 선택한 경우가 있었다.

구분	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
교육 및 학교	119	3.0	394	1.3	3,938	1.3	7,876	1.3
시사, 사회	32	0.8	590	1.9	5,903	1.9	13,620	2.2
예술, 문화생활	185	4.6	550	1.8	5,500	1.8	10,980	1.8
전공/ 전문 지식	10	0.3	31	0.1	314	0.1	714	0.1
기타[69]	1	>0.1	6	>0.1	65	>0.1	147	>0.1
합계	**3,982**	**100**	**31,173**	**100**	**311,731**	**100**	**614,860**	**100**

<표 29> 주제 대화 세부 주제별 구축 분량 및 비율

대화의 세부 주제 구성 비율을 통해 알 수 있듯이 식음료나 주거와 생활을 비롯하여 여행, 여가, 미용과 건강 등과 같이 의식주나 여가 활동 등 개인의 일상생활과 관련된 내용이 높은 비율로 나타난 반면, 시사나 예술, 전공 등과 같이 일상생활과의 관련성이 떨어지는 주제가 차지하는 비율은 10% 미만으로 나타났다.

앞으로 대화 제공자에게 선택하도록 제시한 대화 주제의 체계성과 적절성은 검증이 필요한 부분이다. 본 사업을 통해 구축한 자료를 통하여 메신저 대화의 주제 분류에 더욱 적합한 체계를 마련하기 위한 연구가 필요하다.

1.3. 메신저 대화 말뭉치의 참여자 구성

메신저 대화 말뭉치의 참여자 구성은 메신저 대화 말뭉치가 메신저 사용자 모집단을 대표성 있게 반영하는가[70]를 측정하는 지표가 되기 때문에 면밀한 설계와 수집 기간 중에도 지속적인 점검이 필요하다. 본 사업에서는 성별과 연령의 구성 비율에 중점을 두고 지속적으로 확인하면서 화자를 모집하였다.

1.3.1. 성별 및 연령

메신저 대화 말뭉치 참여자의 성별과 연령에 따른 분포는 <표 30>과 같다.

70) 다만 설계 단계에서도 밝혔듯이 실제 화자를 모집할 때 대화 자료의 수집 가능성도 고려하여야 하기 때문에 메신저 사용자 모집단과 정확하게 일치하도록 표본을 설계하는 것은 어렵다.

구분	남성		여성		합계	
	참여자 수	비율(%)	참여자 수	비율(%)	참여자 수	비율(%)
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

<표 30> 메신저 대화 말뭉치 참여자의 성별 및 연령 구성

메신저 말뭉치 구축에 참여한 전체 대화 제공자는 10,083명이다. 이는 저작권 이용 허락 계약 등의 모든 절차를 완료하고 대화를 제공한 10,135명 중에서 말뭉치 구축 자료 선별 과정에서 제외한 인원을 최종적으로 산정한 결과이다.

성별 비율은 남성이 24.4%, 여성이 75.6%로 여성의 참여 비율이 상대적으로 높고, 남성의 참여 비율은 상대적으로 낮았다. 성별 제한을 두지 않고 공개적인 모집을 시행하였으나, 남성과 여성 대화 제공자의 구성 비율이 현저한 차이를 보이는 데에는 여러 요인이 작용하고 있다.

먼저 DMC MEDIA(2017:4)에 따르면 남성과 여성의 모바일 메신저 앱 사용 비율이 86.5%와 94.4%로 남성에 비해 여성의 모바일 메신저 앱 사용 비율이 높은 것으로 나타나고 있다. 그리고 DMC MEDIA(2019:5)에 따르면 외부 거래처 응대나 업무용으로 모바일 메신저를 사용하는 비중이 남성은 54.5%, 여성은 49.5%로 나타난다. 즉 여성에 비해 남성의 모바일 메신저 사용 비율이 다소 낮다. 그리고 남성의 경우 모바일 메신저를 공적인 업무를 위해 사용하는 비중이 여성에 비해 높기 때문에 공적인 관계에서 이루어지는 대화 제공에 제약이 있었던 본 사업의 특성이 남성 참여율이 낮은 한 가지 요인으로 추정할 수 있다.

다음으로 DMC MEDIA(2019:12)에 따르면 카카오톡 채널을 통해 할인이나 이벤트 정보를 얻는 것에 성별에 따른 차이가 있다. 남성은 56.4%가 카카오톡 채널을 통해 이벤트 등의 정보를 얻는 반면, 여성은 82%가 카카오톡 채널을 통해 이벤트 등의 정보를 얻는 것으로 나타나고 있다. 본 사업의 주요 홍보 창구 중 하나가 카카오톡 채널이었기 때문에 남성에 비해 여성이 메신저 대화 참여자 모집 관련 정보에 노출될 가능성이 높았고, 이로 인하여 여성의 참여율이 남성에 비해 높았던 것으로 볼 수 있다.

연령에 따른 구성 비율은 20대가 54.3%로 높은 비율로 나타났다. 사업 초반 홍보가 20대 학생에게 치중되어 있었고, 20대가 다른 세대에 비해 SNS와 온라인 활동을 통해 더욱 적극적으로 정보를 공유하는 특성이 있어 20대 위주로 대화 제공자 모집에 대한 정보가 확산된 영향으로 보인다. 다만 중간 점검 당시 70% 이상으로 높았던 20대 참여 비율은 지속적인 점검과 중장년층을 대상으로 한 집중적인 홍보와 대화 수집 등을 통하여 54.3%까지 낮아지게 되었다.

20대를 제외한 다른 연령대의 참여율이 상대적으로 낮은 요인에 대해서는 분석이 필

요하다[71]. 먼저 메신저를 적극적으로 활용할 것으로 예상되는 10대의 참여율이 8.2%에 그치고 있는데, 이는 10대 다수가 친구와의 사적인 대화 등에서는 페이스북 메신저를 더욱 적극적으로 활용하는 양상과도 관련이 있다. 본 사업에서는 대화에 활용하는 메신저의 종류를 특별히 제한하지 않았다. 그러나 페이스북 메신저의 경우는 대화를 텍스트로 추출하는 기능을 직접적으로 지원하지 않기 때문에 페이스북 메신저를 통해 친구와 주고받은 대화를 제공하기가 상대적으로 어려웠을 것으로 추정한다.

40대 이상의 경우에도 참여율이 10.7%에 그쳤다. 직장 생활로 인해 이벤트 등의 참여에 시간적인 제한이 있고, 참여를 유인하는 요인으로 대화 제공에 대한 보상이 미치는 영향력이 2, 30대에 비해 상대적으로 크지 않았다는 점이 낮은 참여율의 원인이라 추정한다.

60세 이상은 다른 연령대에 비해 인터넷의 활용도가 상대적으로 떨어지기 때문에 대화 제공자 모집 게시물에 노출되는 빈도가 높지 않았을 것이라는 점과 대화를 추출하거나, 수집 봇을 초대하는 등 대화를 제공하는 방식에도 익숙하지 않았을 것이라는 점이 낮은 참여율의 원인이라 추정한다.

메신저 사용자의 인구·통계적 특성에 근거를 두고서 우리말 메신저 사용 환경을 대표성 있게 반영하는 메신저 대화 말뭉치를 구축하기 위해서는 성별과 연령별 메신저 사용에서 나타나는 특성에 대한 더욱 면밀한 고려가 필요하며, 모집 방식과 보상 등도 성별과 연령에 따른 특성에 맞추어 다각도로 활용할 필요가 있다.

1.3.2. 직업

메타 정보 중 하나로 직업 항목을 수집하였다. 메타 정보로서 직업이 유의미하고 변별적인 가치를 지니기 위해서는 동일 직종 종사자 간에 업무에 대하여 공적으로 나누는 대화를 수집할 필요가 있다. 직무와 관련된 특수한 용어나 직책과 관련된 호칭 등이 일상적인 대화와 변별되기 때문이다.

메신저 대화 제공자 전체의 직업별 분포는 <표 31>과 같다[72].

71) 면밀한 분석을 위해서는 대화를 제공하지 않은 사람을 대상으로 조사가 먼저 이루어져야 한다. 다만 대화 제공을 하지 않은 사람을 대상으로 조사를 할 수 있는 여건이 마련되어 있지 않아 그러한 조사가 실제로 이루어지지 않았고, 대화 제공자들로부터 주변 가족이나 지인들이 참여를 꺼리는 사유에 대해 수집한 피드백에 근거하여 연령별 대화 제공의 제한 요인을 추정해 보았다.

72) 직업의 세부 분류는 직업의 세부 분류는 한국표준직업분류(통계청, 2017)의 분류 항목을 일부 수정하여 활용하였다.

구분	남성		여성		합계	
	참여자 수	비율(%)	참여자 수	비율(%)	참여자 수	비율(%)
가정주부	6	0.1	950	9.4	956	9.5
경영/관리직	124	1.2	203	2.0	327	3.2
군인	51	0.5	3	>0.1	54	0.5
기능원/관련 종사자	45	0.4	31	0.3	76	0.8
기술직 종사자	165	1.6	40	0.4	205	2.0
농림, 임업, 어업 종사자	10	0.1	7	0.1	17	0.2
단순 노무 종사자	47	0.5	31	0.3	78	0.8
무직/취업 준비생	235	2.3	941	9.3	1,176	11.7
사무 종사자	333	3.3	1,559	15.5	1,892	18.8
서비스 종사자	143	1.4	335	3.3	478	4.7
전문가/관련 종사자	353	3.5	1,200	11.9	1,553	15.4
판매/영업 종사자	90	0.9	142	1.4	232	2.3
학생	761	7.5	1,953	19.4	2,714	26.9
기타	96	1.0	229	2.3	325	3.2
합계	**2,459**	**24.4**	**7,624**	**75.6**	**10,083**	**100**

<표 31> 메신저 대화 말뭉치 참여자의 직업 구성

직업에 따른 구성 비율은 학생이 26.9%로 나타나 가장 높은 비중을 차지했다. 대학생 커뮤니티를 대상으로 한 초반 사업 홍보와 온라인 커뮤니티 활동에서 20대 학생이 차지하는 비중이 높기 때문에 직업별 인원 구성에서도 높은 비중으로 나타난 것으로 볼 수 있다.

한편, 사무 종사자와 전문과 및 관련 종사자와 같이 사무실 내에서 컴퓨터를 많이 활용하는 직종의 참여 비율이 34.2%로 나타났고, 무직(취업 준비), 가정주부와 같이 상대적으로 시간을 유동적으로 활용할 수 있는 유형의 참여자 비율도 21.2%로 나타났다. 수집 봇을 활용한 대화 이벤트 등에 참여하여 기한 내 일정 분량 이상의 대화를 생성할 수 있는 직종의 참여도가 상대적으로 높았던 것으로 보인다.

다만 직업 구성 비율을 경제 활동 인원과 비경제 활동 인원[73]으로 구분하였을 때 비경제 활동 인원의 참여 비율이 높을 것으로 예측했던 것과는 달리 <표 32>와 같이 비경제 활동 인원과 경제 활동 인원의 참여 비율은 큰 차이를 보이지 않았다.

구분	남성		여성		합계	
	참여자 수	비율(%)	참여자 수	비율(%)	참여자 수	비율(%)
경제 활동 인원	1,361	13.5%	3,551	35.2%	4,912	48.7%
비경제 활동 인원	1,002	9.9%	3,844	38.1%	4,846	48.1%
미분류(기타)	96	1.0%	229	2.3%	325	3.2%
합계	2,459	24.4	7,624	75.6	10,083	100

<표 32> 경제 활동 유무에 따른 메신저 대화 말뭉치 참여자의 구성

73) 학생, 무직/취업 준비, 가정주부와 같이 소속된 직장이 없고, 정기적으로 급여를 받지 않는 인원을 비경제 활동 인원으로 분류하고 그 외는 경제 활동 인원으로 구분하였다. 기타 항목은 미분류 항목으로 구분하였다.

이는 사업 후반부에 낯선 관계 대화 수집이나 주제 특화 대화 수집을 통해 짧은 길이의 대화도 수집함으로써 한정된 시간 내에서 일정 분량 이상의 대화를 생성해야 한다는 부담과 시간의 제한이 다소 완화되었기 때문인 것으로 보인다.

1.3.3. 지역

지역은 지역 방언 구사와 관련이 되며, 대화 상대방과의 관계에 따라 부호전환(code-switching)이나 부호혼용(code-mixing)을 발생시키는 요인으로 작용하는 등 언어 사용의 주요한 변별적 요인 가운데 하나이다. 특히 메신저 대화 자료에서는 구어 성격의 언어가 문어로 표기되었을 때 방언의 실현 양상을 지역과 관련지어 살펴볼 수 있다.

본 사업에서는 대화 제공자의 출신지뿐만 아니라 주요 성장지와 현재 거주지까지 메타 정보 항목으로 수집하였다.

대한민국 지역별 인구 구성[74]과 메신저 말뭉치 대화 제공자의 지역별 구성을 비교하면 <표 33>과 같다.

지역		총 인구 실제		출생지		주 성장지		현 거주지	
		인원	비율(%)	참여자 수	비율(%)	참여자 수	비율(%)	참여자 수	비율(%)
수도권	서울	9,673,936	18.7	2,667	26.5	2,279	22.6	2,767	27.4
	경기	13,103,188	25.4	1,670	16.6	2,275	22.6	2,772	27.5
	인천	2,936,117	5.7	505	5	565	5.6	567	5.6
수도권 소계		25,713,241	49.8	4,842	48	5,119	50.8	6,106	60.6
충청권	대전	1,511,214	2.9	313	3.1	364	3.6	321	3.2
	세종	312,374	0.6	13	0.1	14	0.1	69	0.7
	충북	1,620,935	3.1	245	2.4	227	2.3	184	1.8
	충남	2,181,416	4.2	292	2.9	269	2.7	247	2.4
충청권 소계		5,625,939	10.9	863	8.6	874	8.7	821	8.1
경북권	대구	2,444,412	4.7	748	7.4	757	7.5	724	7.2
	경북	2,672,902	5.2	509	5	455	4.5	306	3.0
경북권 소계		5,117,314	9.9	1,257	12.5	1,212	12	1,030	10.2
경남권	부산	3,395,278	6.6	861	8.5	742	7.4	627	6.2
	울산	1,150,116	2.2	246	2.4	255	2.5	185	1.8
	경남	3,350,350	6.5	632	6.3	620	6.1	430	4.3
경남권 소계		7,895,744	15.3	1,739	17.2	1,617	16	1,242	12.3
전라권	광주	1,490,092	2.9	273	2.7	283	2.8	220	2.2
	전북	1,818,157	3.5	306	3	256	2.5	166	1.6
	전남	1,790,352	3.5	388	3.8	305	3	174	1.7
전라권 소계		5,098,601	9.9	967	9.6	844	8.4	560	5.6
강원		1,520,391	2.9	321	3.2	299	3	263	2.6
제주		658,282	1.3	74	0.7	64	0.6	43	0.4
해외/기타		-	-	20	0.2	54	0.5	18	0.2
합계		51,629,512	100	10,083	100	10,083	100	10,083	100

<표 33> 대한민국 지역별 인구 구성과 메신저 대화 말뭉치 참여자의 지역별 구성 비교

74) 통계청의 2018년 12월 기준 인구 총조사를 근거로 하였다.

수도권의 실제 인구 구성 비율이 49.8%인 반면, 메신저 대화 말뭉치 화자 중 수도권 거주자의 비율은 60.6%로 10%에 가까운 차이가 난다. 그리고 수도권을 제외한 대부분 지역의 대화 제공자의 비율은 실제 인구 구성 비율에 비해 낮게 나타난다. 이는 메신저 대화 제공자의 연령 구성에서 20대와 30대가 차지하는 비중이 높다는 점과 관련이 있다. 즉 20대와 30대의 경우 유학이나 취업 등의 이유로 수도권에 거주하는 비율이 높기 때문이다.

예외적으로 대구·경북 지역은 실제 인구 구성과 대화 제공자의 비율이 큰 차이를 보이지 않는다. 특히 대구 지역 화자의 구성 비율이 7.2%로, 대구 지역의 실제 인구 구성비인 4.7%에 비해 높게 나타났다. 이는 다자 대화 이벤트 기간 중에 대구 지역 대학교 커뮤니티에 이벤트에 대한 공유와 참여가 타 지역에 비해 활성화되었기 때문이다.

1.3.4. 기기 및 키보드 유형

메신저 대화에 사용하는 기기 유형과 키보드 유형에 따라 오타 발생 유형이 다르다. 이러한 오타 유형 등을 학습시킴으로써 자동 오타 수정 등의 기능 구현에 활용할 수 있다.

이러한 오타 유형의 학습은 키보드 유형이 메신저 언어 형식의 형태 실현 양상과도 관련이 있음을 의미한다. 예를 들어 자음을 연속으로 치기가 용이한 키보드와 그렇지 않은 키보드의 경우 'ㅇㅇ', 'ㅋㅋ', 'ㅎㅎ'와 같은 자음 연속의 형태적 실현 양상이나 사용 빈도 등에서 차이를 보일 가능성이 있다.

먼저 본 사업에서는 메신저 대화를 할 때 주로 사용하는 기기를 PC, 태블릿, 스마트폰의 세 항목으로 나누어 조사하였고, 그 결과는 <표 34>와 같다.

기기 유형	참여자 수	비율(%)
스마트폰	9,202	91.3
PC	857	8.5
태블릿(패드)	24	0.2
합계	**10,083**	**100**

<표 34> 메신저 대화 말뭉치 참여자의 메신저 대화 시 주요 사용 기기

다음으로 메신저 대화 제공자가 사용하는 키보드 유형은 [그림 29]와 같이 8개 유형과 기타로 나누어 수집하였다.

[그림 29] 메신저 대화 말뭉치 참여자가 사용하는 키보드의 유형

메신저 대화 말뭉치 화자가 사용하는 키보드의 유형 구성은 <표 35>와 같다.

키보드 유형	참여자 수	비율(%)
2벌식(쿼티)	6,621	65.7
나랏글	236	2.3
단모음	451	4.5
딩굴	66	0.7
모아키	18	0.2
베가(VEGA)	80	0.8
천지인	2,412	23.9
천지인 플러스	49	0.5
기타	150	1.5
합계	**10,083**	**100**

<표 35> 메신저 대화 말뭉치 참여자가 사용하는 키보드 유형 구성

1.4. 참여자 간 관계 구성

1.4.1. 참여자 간 관계

2인 대화에서 대화 참여자 간 관계의 구성과 구축 분량은 <표 36>과 같다[75].

참여자 간 관계		수집 파일		대화		말차례		발화	
		분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
가족	부부	214	3.1	18,351	3.1	183,515	3.1	359,430	2.9
	부모-자녀	220	3.1	10,026	1.7	100,259	1.7	187,205	1.5
	형제-자매	526	7.5	58,998	10.1	589,978	10.1	1,177,417	9.6
	기타(조부모-손주 및 친인척 등)	66	0.9	8,176	1.4	81,761	1.4	173,478	1.4
가족 소계		1,026	14.7	95,551	16.3	955,513	16.3	1,897,530	15.5
학교/학원	동기/동창/동급생	1,664	23.8	250,119	42.7	2,501,194	42.7	5,373,035	43.9
	선후배	134	1.9	18,880	3.2	188,800	3.2	408,826	3.3
	스승-제자	5	0.1	389	0.1	3,895	0.1	10,448	0.1
	기타(직원-학생 등)	5	0.1	368	0.1	3,680	0.1	7,972	0.1
학교 소계		1,808	25.8	269,756	46.1	2,697,569	46.1	5,800,281	47.4
직장	동기/동료/동업자	254	3.6	36,155	6.2	361,555	6.2	750,637	6.1
	선후배, 상사-부하	64	0.9	2,651	0.5	26,506	0.5	58,375	0.5
	기타(거래처, 고객 등)	1	>0.1	46	>0.1	459	>0.1	837	>0.1
직장 소계		319	4.6	38,852	6.6	388,520	6.6	809,849	6.6
지역	고향 및 이전 거주지 지인	89	1.3	12,136	2.1	121,367	2.1	270,984	2.2
	현 거주지 지인	150	2.1	16,451	2.8	164,506	2.8	325,398	2.7
지역 소계		239	3.4	28,587	4.9	285,873	4.9	596,382	4.9
기타	동호회, 스터디	74	1.1	10,754	1.8	107,540	1.8	228,394	1.9
	연인	411	5.9	72,745	12.4	727,447	12.4	1,499,233	12.3
	온라인 커뮤니티	308	4.4	32,368	5.5	323,686	5.5	649,268	5.3
	종교 관련	73	1.0	8,288	1.4	82,876	1.4	156,645	1.3
	군대	5	>0.1	107	>0.1	1,074	>0.1	2,344	>0.1
	그 외 사회적 관계	237	3.4	22,097	3.8	220,968	3.8	468,652	3.8%
기타 소계		1,108	15.8	146,359	25.0	1,463,591	25.0	3,004,536	24.6
낯선 관계		2,497	35.7	6,318	1.1	63,177	1.1	128,430	1.0
합계		**6,997**	**100**	**585,423**	**100**	**5,854,243**	**100**	**12,237,008**	**100**

<표 36> 메신저 대화 말뭉치의 참여자 간 관계와 구축 분량

수집 파일 기준으로 전체 참여자 쌍 중 낯선 관계가 35.7%로 가장 높은 비율로 나타난다. 낯선 관계는 일반적인 메신저 대화에서는 수집이 어렵지만, 상호 전제하고 있는 맥락이 적어 내용의 생략 가능성이 적고 상대적으로 정제된 대화를 생성할 가능성이 크기 때문에 본 사업에서는 수집 봇을 통하여 낯선 관계의 대화도 수집하였다.

다만 낯선 관계의 대화는 통제된 상황에서 제한된 시간 내에서 대화 수집이 이루어

75) 참여자 간 관계 구성은 2인 대화를 기준으로 하였다. 다자 대화의 경우는 참여자 간 관계를 항목 하나로 설정하기 어려운 경우가 있어 복수 선택이 가능하도록 하였기 때문이다.

졌고 대화를 오래 지속하기 어려운 경우가 많아 말차례를 기준으로 한 구축 분량은 전체 말차례의 1.1%만을 차지한다.

발화와 말차례를 기준으로 할 경우에 가장 높은 비율을 차지하는 관계는 학교/학원의 동기 또는 동급생, 동창 관계로 각각 43.9%, 42.7%의 비율을 차지하고 있다. 학교 동기 관계는 수집 파일 기준으로도 낯선 관계 다음인 23.8%의 비율을 차지하고 있는 것으로 나타나며, 메신저 대화 말뭉치 화자의 직업 구성에서 학생의 비율이 26.9%로 가장 높은 것과도 관련이 있다.

형제, 자매를 제외하고 학교 동기나 동창 또는 직장 동기나 동료, 연인과 같이 상하 위계로부터 자유로운 관계의 대화 제공 건수가 많은 것으로 나타나고 있다. 대화를 제공하기 위해서는 상대방으로부터 동의를 얻어야 한다는 본 사업의 제한으로 인하여 상대방에게 대화 제공을 요구하는 것이 부담이 없는 관계에서 대화 제공이 많이 이루어진 결과라 볼 수 있다.

1.4.2. 대화 참여자 간 친밀도

대화 참여자 간 관계와 함께 친밀도 또한 대화 양상의 주요한 변인이 된다. 대화 참여자 간 친밀도는 대화 상대방과의 친밀도에 대해 1(친밀도가 낮음)~5(친밀도가 높음) 중에서 대화 제공자 스스로 선택하도록 하였고, 낯선 관계에서 이루어진 대화는 작업자가 직접 0으로 입력하였다.

2인 대화에서 대화 참여자 간 친밀도에 따른 구축 분량은 <표 37>과 같다[76].

친밀도	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
5	3,294	47.1	464,684	79.4	4,646,843	79.4	9,693,997	79.2
4	773	11.0	90,463	15.5	904,626	15.5	1,894,219	15.5
3	297	4.2	21,062	3.6	210,626	3.6	460,680	3.8
2	70	1.0	1,752	0.3	17,522	0.3	38,324	0.3
1	66	0.9	1,145	0.2	11,449	0.2	21,358	0.2
0	2,497	35.7	6,318	1.1	63,177	1.1	128,430	1.0
합계	**6,997**	**100**	**585,424**	**100**	**5,854,243**	**100**	**12,237,008**	**100**

<표 37> 메신저 대화 말뭉치의 참여자 간 친밀도와 구축 분량

전체 참여자 중 가장 높은 비율로 나타나는 친밀도는 5로, 수집 파일 기준으로 47.1%를 차지하고 있으며, 말차례 기준으로도 79.4%를 차지하고 있다. 상대방에게 대화 제공을 요구하는 것이 부담이 없는 관계의 대화 제공이 많이 이루어진 결과이다.

친밀도가 0인 대화는 수집 파일 기준으로 두 번째로 높은 비율인 35.7%를 차지하고 있다. 다만 친밀도 0인 대화는 참여자 간 관계 항목과 마찬가지로 통제된 상황에서 제

76) 친밀도도 참여자 간 관계와 마찬가지로 2인 대화를 기준으로 하였다.

한된 시간 내에서 수집이 이루어졌기 때문에 말차례를 기준으로 한 구축 분량은 전체 말차례의 1.1%이다.

친밀도가 1과 2로 낮은 대화가 차지하는 비중은 수집 파일 기준으로 2% 미만이며, 말차례를 기준으로 할 경우에도 0.5%에 불과하다.

대화 참여자 간 친밀도에 따른 세부 참여자 관계를 참여자 쌍 기준으로 나타내면 <표 38>과 같다.

친밀도 / 참여자 간 관계		5		4		3		2		1		합계	
		분량	비율	분량	비율	분량	비율	분량	비율	분량	비율	분량	비율
가족	부부	192	4.3 (89.7)	15	0.3 (7.0)	7	0.2 (3.3)	0	0.0 (0.0)	0	0.0 (0.0)	214	4.8 (100)
	부모-자녀	169	3.8 (76.8)	45	1.0 (20.5)	6	0.1 (2.7)	0	0.0 (0.0)	0	0.0 (0.0)	220	4.9 (100)
	형제-자매	428	9.5 (81.4)	80	1.8 (15.2)	18	0.4 (3.4)	0	0.0 (0.0)	0	0.0 (0.0)	526	11.7 (100)
	기타(조부모-손주 및 친인척 등)	45	1.0 (68.2)	12	0.3 (18.2)	7	0.2 (10.6)	1	>0.1 (1.5)	1	>0.1 (1.5)	66	1.5 (100)
가족 소계		834	18.5 (81.3)	152	3.4 (14.8)	38	0.8 (3.7)	1	>0.1 (0.1)	1	>0.1 (0.1)	1,026	22.8 (100)
학교/학원	동기/동창/동급생	1318	29.3 (79.2)	297	6.6 (17.8)	45	1.0 (2.7)	3	0.1 (0.2)	1	>0.1 (0.1)	1,664	37.0 (100)
	선후배	82	1.8 (61.2)	33	0.7 (24.6)	13	0.3 (9.7)	5	0.1 (3.7)	1	0.0 (0.7)	134	3.0 (100)
	스승-제자	1	>0.1 (20.0)	3	0.1 (60.0)	1	>0.1 (20.0)	0	0.0 (0.0)	0	0.0 (0.0)	5	0.1 (100)
	기타(직원-학생 등)	4	0.1 (80.0)	1	>0.1 (20.0)	0	0.0 (0.0)	0	0.0 (0.0)	0	0.0 (0.0)	5	0.1 (100)
학교 소계		1,405	31.2 (77.7)	334	7.4 (18.5)	59	1.3 (3.3)	8	0.2 (0.4)	2	>0.1 (0.1)	1,808	40.2 (100)
직장	동기/동료/동업자	139	3.1 (54.7)	66	1.5 (26.0)	40	0.9 (15.7)	7	0.2 (2.8)	2	>0.1 (0.8)	254	5.6 (100)
	선후배, 상사-부하	20	0.4 (31.3)	16	0.4 (25.0)	22	0.5 (34.4)	5	0.1 (7.8)	1	>0.1 (1.6)	64	1.4 (100)
	기타(거래처, 고객 등)	1	>0.1 (100)	0	0.0 (0.0)	0	0.0 (0.0)	0	0.0 (0.0)	0	0.0 (0.0)	1	>0.1 (100)
직장 소계		160	3.6 (50.2)	82	1.8 (25.7)	62	1.4 (19.4)	12	0.3 (3.8)	3	0.1 (0.9)	319	7.1 (100)
지역	고향 및 이전 거주지 지인	74	1.6 (83.1)	12	0.3 (13.5)	2	>0.1 (2.2)	1	>0.1 (1.1)	0	0.0 (0.0)	89	2.0 (100)
	현 거주지 지인	107	2.4 (71.3)	27	0.6 (18.0)	12	0.3 (8.0)	4	0.1 (2.7)	0	0.0 (0.0)	150	3.3 (100)
지역 소계		181	4.0 (75.7)	39	0.9 (16.3)	14	0.3 (5.9)	5	0.1 (2.1)	0	0.0 (0.0)	239	5.3 (100)

친밀도 / 참여자 간 관계		5		4		3		2		1		합계	
		분량	비율	분량	비율	분량	비율	분량	비율	분량	비율	분량	비율
기타	동호회, 스터디	36	0.8 (48.6)	16	0.4 (21.6)	19	0.4 (25.7)	2	>0.1 (2.7)	1	>0.1 (1.4)	74	1.6 (100)
	연인	397	8.8 (96.6)	14	0.3 (3.4)	0	0.0 (0.0)	0	0.0 (0.0)	0	0.0 (0.0)	411	9.1 (100)
	온라인 커뮤니티	144	3.2 (46.8)	63	1.4 (20.5)	66	1.5 (21.4)	18	0.4 (5.8)	17	0.4 (5.5)	308	6.8 (100)
	종교 관련	19	0.4 (26.0)	17	0.4 (23.3)	12	0.3 (16.4)	20	0.4 (27.4)	5	0.1 (6.8)	73	1.6 (100)
	군대	4	0.1 (80.0)	1	>0.1 (20.0)	0	0.0 (0.0)	0	0.0 (0.0)	0	0.0 (0.0)	5	0.1 (100)
	그 외 사회적 관계	114	2.5 (48.1)	55	1.2 (23.2)	27	0.6 (11.4)	4	0.1 (1.7)	37	0.8 (15.6)	237	5.3 (100)
기타 소계		714	15.9 (64.4)	166	3.7 (15.0)	124	2.8 (11.2)	44	1.0 (4.0)	60	1.3 (5.4)	1,108	24.6 (100)
합계		**3,294**	**73.2**	**773**	**17.2**	**297**	**6.6**	**70**	**1.6**	**66**	**1.5**	**4,500**	**100**

<표 37> 대화 참여자 간 관계에 따른 친밀도 구성(참여자 쌍 기준)

친밀도 1과 2로 나타난 참여자 쌍은 전체 참여자 쌍의 3.1%의 비율을 차지하고 있다. 친밀도 1과 2를 선택한 참여자 쌍 중 76.5%는 온라인 커뮤니티나 그 외 사회적 관계로 나타났는데, 평소 대면 접촉이 이루어지지 않은 참여자 쌍의 경우 낮은 친밀도에 해당하는 1과 2를 선택한 것으로 보인다.

1.4.3. 대화 참여자 간 연락 빈도

주관적 요소인 참여자 간 친밀도를 보충하기 위한 요소로서 참여자 간 연락 빈도를 메타 정보 항목으로 설정하였다[77]. 대화 제공자 스스로 '월 1회 미만, 주 1회 미만, 주 1~2회, 주 3회 이상, 거의 매일' 항목 중에서 선택하도록 하였고, 낯선 관계인 경우에 '처음 연락한다' 항목을 설정하여 작업자가 일괄 기재하였다.

대화 참여자 간 연락 빈도에 따른 구성 비율은 <표 39>와 같다.

77) 다만 연락 빈도가 반드시 친밀도와 양의 상관관계가 있음을 의미하는 것은 아니다. 가족인 경우, 친밀도가 높다고 하더라도 연락 빈도는 낮을 수가 있는 반면에 공적인 관계는 친밀도는 낮다고 하더라도 연락 빈도는 높을 수 있기 때문이다. 연락 빈도와 친밀도의 관련성에 대해서는 면밀한 분석이 필요하다.

연락 빈도	수집 파일		대화		말차례		발화	
	분량	비율(%)	분량	비율(%)	분량	비율(%)	분량	비율(%)
거의 매일	3,311	44.8	539,410	75.7	5,394,102	75.7	10,934,610	74.9
주 3회 이상	819	11.1	115,509	16.2	1,155,088	16.2	2,447,831	16.8
주 1~2회	343	4.6	30,721	4.3	307,214	4.3	660,667	4.5
주 1회 미만	169	2.3	13,457	1.9	134,572	1.9	283,569	1.9
월 1회 미만	184	2.5	6,219	0.9	62,197	0.9	126,343	0.9
처음(낯선 관계)	2,569	34.7	6,975	1	69,746	1	138,806	1
합계	7,395	100	712,291	100	7,122,919	100	14,591,826	100

<표 39> 메신저 대화 말뭉치의 참여자 간 연락 빈도와 구축 분량

거의 매일 연락하거나 주 3회 이상 연락을 하는 관계의 참여자 쌍이 수집 파일 전체의 55.9%를 차지하는 것으로 나타났고, 처음 연락하는 관계는 34.7%로 나타났다. 그리고 주 1~2회 미만으로 비교적 연락 빈도가 낮은 참여자 쌍은 9.4%로 나타났다.

말차례를 기준으로 할 경우, 주 3회 이상 연락을 하는 관계의 말차례 비율은 전체 말차례의 91.9%로 나타났고, 주 1~2회 미만으로 비교적 연락 빈도가 낮은 관계의 말차례 비율은 전체 말차례의 7.1%로 나타났다. 그리고 처음 연락하는 관계의 말차례 비율은 전체 말차례의 1%로 나타났다.

제 4 장

마무리 및 제언

지금까지 메신저 대화 말뭉치 구축의 과정과 구축 결과에 대해 살펴보았다.

2장에서는 메신저 대화 말뭉치를 구축하는 과정을 설계와 홍보 및 참여자 모집, 자료의 정제와 마크업, 검수로 나누어 단계별로 살펴보았다. 그리고 메신저 대화 말뭉치를 구축하기 위한 자료 수집의 방법과 작업 지침, 말뭉치의 형식 등을 살펴보면서 메신저 대화 말뭉치를 구축하기 위하여 고려해야 할 사항에 대하여 기술하였다.

3장에서는 메신저 대화 말뭉치 구축 결과를 구축 규모, 유형별 구성, 메신저 대화 말뭉치의 화자 구성, 대화 참여자 간 관계 구성의 항목으로 나누어 살펴보면서 메신저 대화 말뭉치의 구축 결과에 대하여 간략히 분석하였다.

본 사업은 국가 단위로 이루어진 최초의 대규모 메신저 대화 자료 수집과 말뭉치 구축 사업이라는 것에 첫 번째 의의가 있다. 기존에 개별 연구자가 수집하기에는 제약이 많았던 메신저 대화 자료를 대규모로 수집하였을 뿐만 아니라, 구축한 모든 자료의 대화 참가자를 대상으로 저작권 이용 허락 계약을 체결하여 계약 범위 내에서 누구나 안정적으로 자료를 이용할 수 있도록 하였다.

본 자료는 일상적인 의사소통 매체로 자리 잡은 메신저 대화를 연구함으로써 매체 환경의 변화에 따른 의사소통 방식의 변화와 메신저 환경의 언어문화를 다양한 범위와 다양한 각도로 연구할 수 있는 자료로 활용할 수 있다. 또한 일상적인 대화 속에 포함된 한국의 사회와 문화를 다각도로 관찰할 수 있는 자료라는 가치도 지닌다.

그리고 본 자료는 언어와 사회 문화를 연구하는 자료로만 그치지 않는다. 수집 봇을 활용하는 수집 방식을 통해 낯선 관계에서 이루어진 대화와 주제가 통제된 대화 등도 일부 실험적인 수집이 이루어졌으며, 이러한 통제된 방식의 수집이 자연어 처리나 딥러닝 분야 등에서도 필요로 하는 최적의 대화 자료 수집에 활용할 수 있다는 가능성도 발견하였다.

아울러 대규모의 언어 자료를 수집하기 위해서는 실제 우리말 사용자의 관심과 참여가 반드시 필요하다는 점에서 본 사업을 통해 말뭉치에 대한 홍보 효과도 있었을 것이라 예상한다.

물론 최초로 실시한 메신저 대화 자료 수집이라는 점에서 본 사업이 지니는 한계와 앞으로 개선이 필요한 부분도 분명하다.

먼저 메신저 대화 자료의 분류와 분석을 위해 참고할 수 있는 선행 연구의 부족으로 인하여 자체적으로 분류의 기준이나 분석의 지침 등을 마련했으나, 기준이나 지침의 이론적 타당성에 대한 엄밀한 검증은 다소 부족했다. 다음으로 메신저 대화 자료의 실제를 접하게 되면서 사전에 마련한 기준과 지침 등이 지속적으로 수정이 되어야 했기 때문에 지침 적용의 일관성이 다소 부족한 부분이 있다.

이러한 한계와 보완점은 앞으로 본 자료를 활용하는 연구자가 지속적으로 개선해 나가야 할 부분이다. 구축한 자료를 통해 앞으로 메신저 대화의 체계를 분류하고 분석하기 위한 기준을 정립하고 그러한 기준의 이론적 타당성에 대한 지속적인 검증이 이루어져야 한다. 또한 자연어 처리 분야에서는 이 자료를 활용하기 위한 전처리도 반드시

필요하다. 이를 위해 메신저 대화의 특성에 맞게 자료를 정규화하는 지침과 기술적인 방법론에 대한 논의가 이루어져야 하며, 실제로 메신저 대화를 전처리하고 분석하기 위한 기술 개발도 이루어져야 한다.

마지막으로 메신저 대화는 개인의 사생활과도 밀접한 관계가 있다. 최근 사생활 노출과 보호에 대한 관심이 지속적으로 높아지고 있으며, 실제로 대화 제공자 다수가 개인 정보 보호를 비롯한 사생활 노출에 대한 관심이 높다는 점도 자료를 수집하는 과정에서 확인할 수 있었다. 물론 본 자료는 기준에 따라 철저한 비식별화가 진행되어 개인 정보와 관련된 문제가 발생할 여지를 최소화하였다. 그럼에도 불구하고 본 자료를 활용하는 연구자와 관리자는 개인의 민감한 사생활을 취급하는 주체로서 책임 의식을 지니고 본 자료에 접근하고 활용할 것을 당부하는 바이다.

참고문헌

강현화(2017). 학습자 말뭉치의 구축과 활용연구. 소통.

국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부(2016), 개인정보 비식별 조치 가이드라인, -비식별 조치 기준 및 지원·관리체계 안내, 관계부처 합동 발행.

국립국어원(2007a). 21세기 세종계획 국어 기초자료 구축. 국립국어원.

국립국어원(2007b). 21세기 세종계획 국어 특수자료 구축. 국립국어원.

국립국어원(2017). 국제 통용 한국어 표준 교육과정 적용 연구. 국립국어원.

국립국어원(2018). 2018년 국어 말뭉치 연구 및 구축. 국립국어원.

김유진·김유미·김승인(2014). 모바일 인스턴트 메신저에서 감정 표현 기능에 관한 비교 연구 : 카카오톡과 프랭클리 챗을 중심으로. 디지털디자인학연구, 14권 3호. 73~82.

김정숙·이정희(2018). 국제 통용 한국어 표준 교육과정의 구성과 내용. 새국어생활, 28권 12호, 49~71.

서상규, 안의정, 봉미경, 최정도, 박종후, 백해파, 송재영, 김선혜(2013). 한국어 구어 말뭉치 연구. 한국문화사.

장문수(2012). 심리학적 감정과 소셜 웹 자료를 이용한 감성의 실증적 분류. 한국지능시스템학회 논문지, 22권 5호. 563~569.

조연정·강정한(2015). 카카오톡은 어떻게 공동체가 되었는가? 다산출판사.

통계청(2017). 한국표준직업분류. 통계청.

통계청(2018). 『인구총조사』 통계정보보고서. 통계청.

한국인터넷진흥원(2018). 2017 인터넷이용실태조사. 과학기술정보통신부 한국인터넷진흥원.

DMC MEDIA(2017). 2017 모바일 메신저 앱 이용 행태. DMC리포트.

DMC MEDIA(2019). 2019 모바일 메신저 앱 이용 행태. DMC리포트.

Baron, Naomi(2008). *Always On: Language in an Online and Mobile World. Oxford University Press*, USA.

<Abstract>

Instant messaging chat data collection and corpus construction

The purpose of this project is to collect conversation data from instant messengers (IM) and build a raw corpus usable as public goods for research and development on artificial intelligence and related industries such as natural language processing and big data.

In order to build a corpus that represents the characteristics of instant messaging, we tried to reflect IM usage statistics and designed samples in consideration of the realistic possibility of data collection. To include various types of data, we classified the conversations by factors affecting the aspects of instant messaging such as interaction between speakers, type of device, conversation subject, collection method etc. In particular, we used a method of extracting and collecting chat data with a bot especially for conversations between strangers and subject-controlled conversations.

All participants agreed to the use of their personal information, and various types of personally identifiable information, such as names, telephone numbers, or account numbers, were de-identified according to the guidelines. Every speaker also signed on a contract of copyright permission so that researchers, institutions and industries may be able to use the corpus stably without legal restrictions.

Two types of format were used to build the corpus : SJML and JSON. Inside, we tagged the de-identified personal information, multimedia elements such as photos, videos or emoticons, and non-linguistic functions like free calls, money transfers, or gifts, in order to facilitate future filtering or replacing.

Finally, we collected 7,395 files from 10,083 speakers and built a raw corpus with meta information attached. It consists of 712,291 conversations with 14,591,826 utterances and 7,122,919 turns.

This corpus can be used as machine learning data for artificial intelligence

following the development of data normalizing, preprocessing and analysis technology specialized for IM conversations. It can also be a basis for establishing theoretical validity and system when studying the language and culture of instant messengers.

Keywords : Mobile Messenger, KakaoTalk, Conversation, Raw Corpus, Natural Language Processing, Deep Learning, Big Data

Project Director: Park Il Seop(mediaCORPUS)

사업 책임자	박일섭(주식회사 미디어코퍼스)
사업 참여자	정연규((주)그립)
	박일섭(주식회사 미디어코퍼스)
	남서정(주식회사 미디어코퍼스)
	신지영(주식회사 미디어코퍼스)
	이서희(주식회사 미디어코퍼스)
	이수경(주식회사 미디어코퍼스)
	양성민(주식회사 다이얼로그디자인에이전시)
	오가영(주식회사 다이얼로그디자인에이전시)
	이태강(주식회사 다이얼로그디자인에이전시)
	문진숙(주식회사 다이얼로그디자인에이전시)
	양한주(주식회사 다이얼로그디자인에이전시)
	송민지(주식회사 다이얼로그디자인에이전시)
	이광진(주식회사 다이얼로그디자인에이전시)
	이나리(주식회사 다이얼로그디자인에이전시)
	양리아(주식회사 다이얼로그디자인에이전시)
	이규수(주식회사 다이얼로그디자인에이전시)
	안택헌((주)그립)
	임명식((주)그립)
	최승욱((주)그립)
	최문성((주)그립)
	윤지수((주)그립)
	손대웅((주)그립)
담당 연구원	이승재(국립국어원 언어정보과장)
	홍혜진(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2019년 12월 19일
발행일: 2019년 12월 19일
인 쇄: 카피랜드

※ 이 책은 국립국어원의 용역비로 수행한 '메신저 대화 자료 수집 및 말뭉치 구축' 사업의 결과물을 발간한 것입니다.