국립국어원 2022-01-34

발 간 등 록 번 호
11-1371028-000922-01

2022년 온라인 게시 자료 수집 및 정제

사업 책임자
이 영 희

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 온라인 게시 자료 수집 및 정제’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 4월 28일 ~ 2022년 11월 28일

2022년 11월 28일

사업 책임자: 이 영 희((주)버즈메트릭스)

사업 수행자 (주)버즈메트릭스

사업 책임자 이영희

사업 참여자 김수진, 신현주, 김도현, 이진상, 유지현, 권주원

<국문 요약>

2022년 온라인 게시 자료 수집 및 정제

본 사업은 4차 산업혁명을 대비하여 인공지능 기술의 개발 및 활용을 위한 대규모 말뭉치를 구축하여 국어 자원의 활용도와 가치를 제고하고, 실제 언어생활을 반영하는 온라인 게시 자료를 수집하여 산업계 및 학계의 기술 개발·연구 활용 등에 필요한 말뭉치 자원을 확보하고 제공하는 데 목적이 있다. 이에 따른 주요 사업 내용을 요약하면 다음과 같다.

첫째, 참여자를 모집하여 온라인 게시 자료에 대한 저작권 이용 허락 계약을 체결하고, 계약이 체결된 참여자의 온라인 게시 자료를 수집하였다. 현대 한국어 사용자의 언어 사용 양상을 반영한 자료를 수집하기 위해 매체별 계정 수와 계정별 수집 한도를 지정하여 게시 자료가 특정 매체와 특정 계정에 편중되지 않도록 하였으며, 분야별·주제별로 균형 있는 자료를 수집하기 위해 목표 비중을 두고 수집 과정에서 키워드에 의거하여 분야와 주제를 분류함으로써 비중에 맞추어 게시 자료가 확보될 수 있도록 하였다.

둘째, 누리소통망과 게시판에서 수집된 온라인 게시 자료 총 31만 건을 원시 말뭉치 형태로 구축하되 적합하지 않은 게시 자료를 제거하고, 비윤리적 내용의 문서는 31만 건에 포함되지 않도록 별도로 분리하였다.

셋째, 게시 자료별로 구축 대상 자료의 메타 정보를 ‘매체 분류, 게시 누리집(사이트), 글 제목, 본문, 게시 날짜, URL 주소, 조회 수, 게시자 정보, 연령, 성별 등’으로 구축하였다.

주요어: 온라인 게시 자료, 말뭉치 수집, 원시 말뭉치

<Abstract>

2022 Online posting crawling and Purification

The purpose of this project is to build a large corpus for the development and utilization of artificial intelligence technology in preparation for the 4th Industrial Revolution to enhance the utilization and value of Korean language resources. The main business contents accordingly are summarized as follows.

First, participants were recruited to sign a copyright permission contract for online posting materials, and online posting materials of the participants who signed the contract were collected. In order to collect data reflecting the language usage of modern Korean users, the number of accounts by media and collection limits by account were designated to prevent posting materials from being concentrated on specific media and accounts.

Second, 310,000 online postings collected from SNS and community site were constructed in the form of primitive corpora, but inappropriate postings were removed, and documents with unethical content were separately separated so as not to be included in 310,000.

Third, the meta information of the data to be constructed for each posting material was constructed by 'media classification, title, post, posting website, posting date, URL address, number of views, publisher information, age, gender, etc‘.

Key-words: online post corpus, web corpus, online post

<개조식 요약문>

2022년 온라인 게시 자료 수집 및 정제

1. 연구 목적

○ 4차 산업혁명 대비 기반 기술 및 인공지능 기술 개발, 활용을 위한 대규모 말뭉치 구축

○ 실제 언어생활을 반영하는 온라인 게시 자료를 수집하여 산업계 및 학계 기술 개발·연구에 필요한 말뭉치 자원 확보

2. 주요 사업 내용

가. 온라인 게시 자료 수집

○ 온라인 게시 원문 자료 매체별·분야별·주제별 수집
(누리소통망(SNS: 페이스북·인스타그램), 게시판)

○ 저작 권리자(참여자)와의 저작권 이용 허락 계약을 통한 저작권 해결

나. 온라인 게시 자료 원시 말뭉치 구축 (총 31만 건)

○ 비적합 자료 원시 말뭉치 구축 제외

○ 비윤리적 내용의 문서 분리 및 별도 말뭉치 구성

○ 개인 정보 비식별화 처리

다. 구축 대상 자료에 대한 메타 정보 구축

○ 게시 자료별 메타 정보 작성 및 목록 작성
(매체 분류, 게시 누리집(사이트), 글 제목, 본문, 게시 날짜, URL 주소, 조회수, 게시자 정보, 연령, 성별 등)

차 례

표 차례

그림 차례

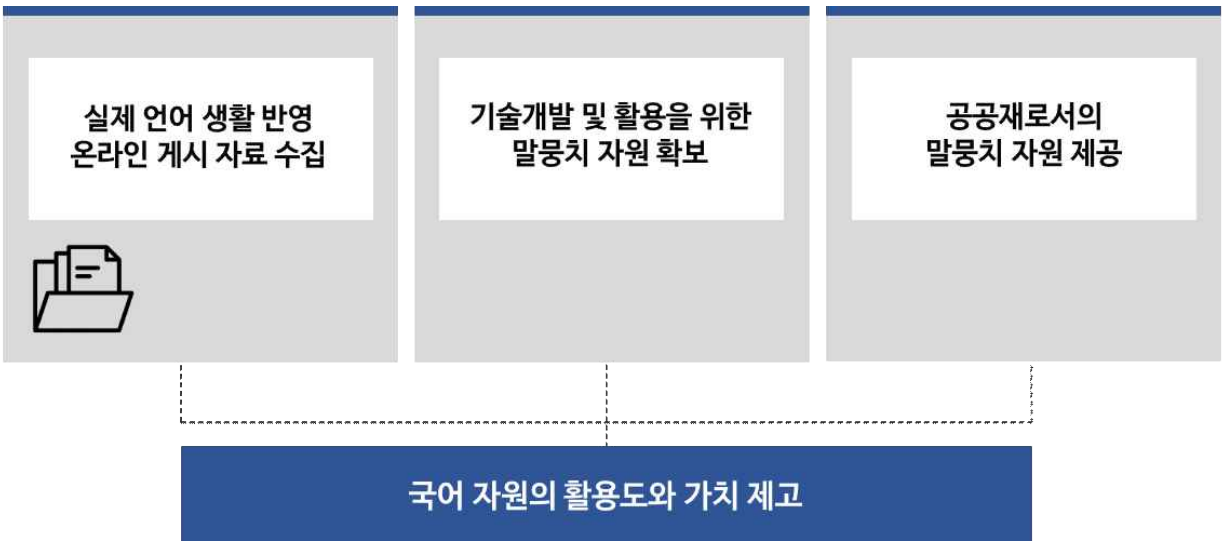
제 1 장

서 론

1. 사업 목적

본 사업은 4차 산업혁명을 대비하여 인공지능 기술의 개발 및 활용을 위한 대규모 말뭉치를 구축하여 국어 자원의 활용도와 가치를 제고하고, 실제 언어생활을 반영하는 온라인 게시 자료를 수집하여 산업계 및 학계의 기술 개발·연구 활용 등에 필요한 말뭉치 자원을 확보하고 제공하는 데 목적이 있다.

<그림 1> 사업 목적 및 필요성



2. 사업 수행 범위

본 사업은 온라인 게시 자료를 수집하여 원시 말뭉치를 구축하고, 구축한 말뭉치에 대한 메타 정보를 구축하는 것으로 구성된다. 구체적인 사업 수행 범위는 다음과 같다.

○ 온라인 게시 자료 수집
- 온라인 게시 원문 자료(누리소통망(SNS: 페이스북·인스타그램), 게시판) 매체별·분야별·주제별 균형 수집
- 저작 권리자와의 저작권 이용 허락 계약을 통한 저작권 해결

○ 온라인 게시 자료 원시 말뭉치 구축 (총 31만 건)

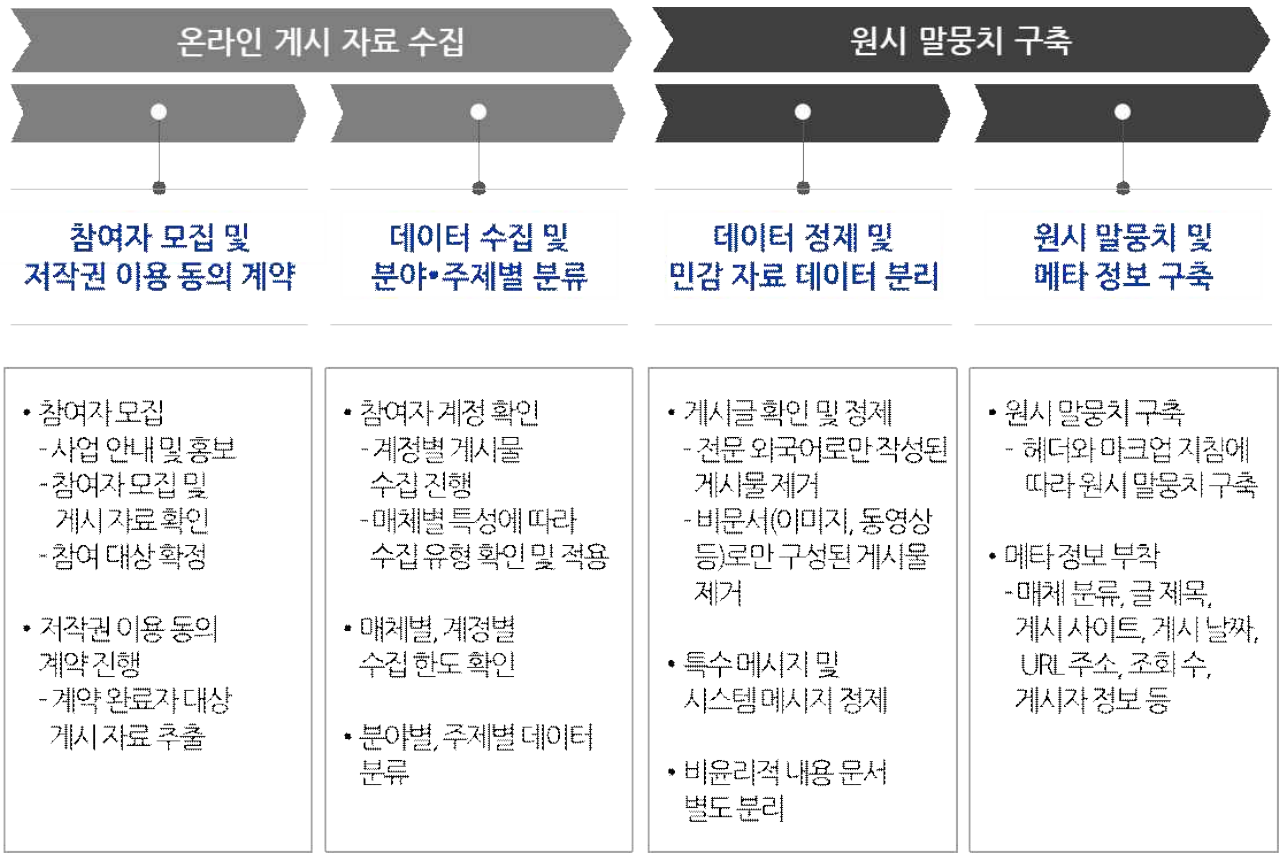
- 비적합 자료 원시 말뭉치 구축 제외
- 비윤리적 내용의 문서 분리 및 별도 말뭉치 구성
- 개인 정보 비식별화 처리

○ 구축 대상 자료에 대한 메타 정보 구축

3. 사업 수행 절차

본 사업은 온라인 게시 자료를 수집하여 원시 말뭉치를 구축하는 절차로 구성된다. 온라인 게시 자료를 보유한 참여자 모집 및 저작권 이용 허락 계약을 진행하고, 계약이 완료된 참여자의 온라인 게시 자료를 수집하여 데이터를 정제한다. 이때 비윤리적 내용이 포함된 문서는 별도로 분리하여 원시 말뭉치를 구축한다. 그리고 해당 자료의 메타 정보를 구축하는 절차로 사업을 수행하였다. 각 단계의 주요 내용은 다음과 같다.

<그림 2> 사업 수행 절차



제 2 장

온라인 게시 자료 수집

1. 참여자 모집 및 선정

본 사업은 제한된 사업 기간과 예산의 범위 안에서 저작권 이용 동의 계약이 완료된 온라인 게시 자료 보유자의 참여가 필요한 사업이다. 특히 31만 건의 온라인 게시 자료를 확보하기 위해서는 많은 참여자가 필요하다. 최소 9개 이상의 사이트에서 450개 이상의 계정을 확보하고, 분야별·주제별 균형 있는 자료를 수집하기 위해서는 다양한 분야를 내용을 담은 온라인 게시 자료가 필요하였다.

사이트별 목표 건수와 계정 수는 국립국어원의 2019년 '웹 말뭉치 구축'자료를 기준으로 하였다. 누리소통망의 경우, 인스타그램과 페이스북 간의 비중이 각각 약 90%와 10%를 차지하는 점을 고려하여 전체 30만 건 중 90%인 27만 건은 인스타그램으로, 10%인 3만 건은 페이스북으로 수집 목표를 정하였다. 게시판 역시 7개 사이트에서 수집이 이루어진 점을 고려하여, 7개 이상 사이트에서 온라인 게시 자료 수집을 목표로 하였으며, 계정 수는 450개 이상의 계정에서 수집되도록 목표를 설정하였다.

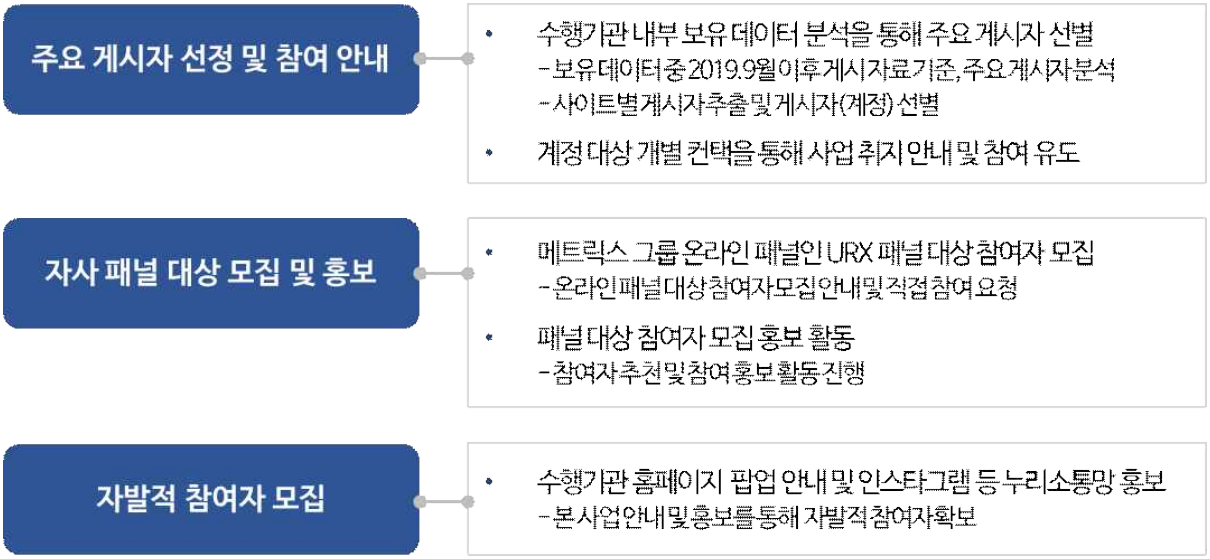
특정 사이트와 특정 계정에 게시 자료가 편중되지 않도록 1개 계정당 4천 건 이하, 1개 게시판 사이트에서 5천 건 이상이 수집되지 않도록 추가적인 기준을 마련하였다.

<표 1> 사이트 및 계정 확보 목표

사이트 구분		수집 목표	
		건수	계정 수
누리소통망	인스타그램	270,000건	400개 이상
	페이스북	30,000건	
게시판 (7개 이상)	네이버/다음 카페	4,000건	50개 이상
	기타 커뮤니티	6,000건	
합 계		310,000건	450개 이상

목표한 사이트와 계정 수를 확보하고, 분야 및 주제에 적합한 게시 자료를 확보하기 위해 여러 가지 방법을 통한 다양한 차원에서의 참여자 모집이 필수적이다. 따라서, 다음과 같은 참여자 모집 방법을 통해 수집된 게시 자료의 다양성을 확보하고자 하였다.

<그림 3> 사업 참여자 모집 방법



첫 번째 참여자 모집 방법인 '주요 게시자 선정 및 참여 유도'는 사업 수행 기관에서 내부적으로 보유한 데이터를 적극적으로 활용하였다. 내부에서 보유하고 있는 누리소통망 및 게시판 데이터를 검토하여 참여 조건에 부합하는 게시 자료를 보유한 참여자를 찾는 방법으로 진행하였다. 내부 축적된 온라인 게시 자료 중, 최근 2년 이내 작성된 게시 자료 200만 건을 분석하여, 22년 게시글 기준 200건 이상의 게시 자료를 작성한 작성자 계정을 추출하였다. 22년 게시글 기준으로 추출한 이유는 본 사업의 수집 대상이 2019년 9월 이후 게시 자료 기준이며, 현재 활발하게 온라인 활동을 하고 있는 참여자를 선별하기 위함이다. 작성자 계정 선별 과정을 거쳐 선정된 계정 소유자에게 사업의 목적과 취지에 대해 안내하고 참여를 유도하는 방식으로 진행하였다.

두 번째로는 사업 수행 기관의 온라인 패널인 'URX(메트릭스 그룹 온라인 패널) 회원을 대상으로 참여 안내와 홍보 활동'을 진행하였다. URX 패널은 2022년 5월 기준 약 130만 명이 회원으로 가입되어 있어, 사업 참여를 유도하고 사업을 홍보하는 일이 동시에 이루어지기 용이하였다. 또한, 패널 회원 본인이 직접 참여하는 것만으로는 제한된 시간 내에 충분한 모집이 이루어지기 어려우므로, 사업을 홍보하고 주변인을 추천하는 방식을 병행하여 진행하였다.

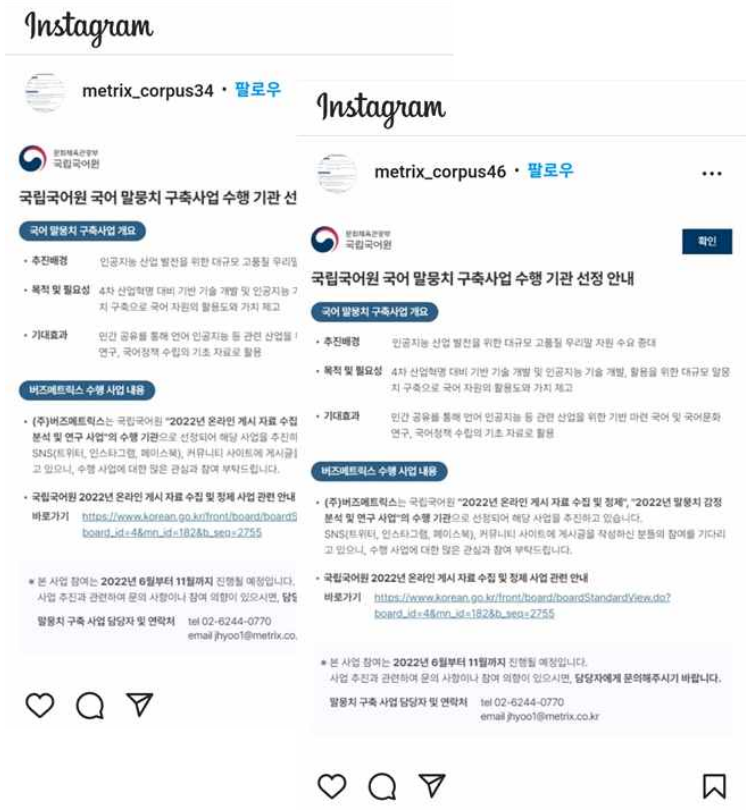
세 번째 방법은 '자발적 참여'를 통한 참여자 모집이다. 국립국어원 누리집 게시판에 사업 관련 안내 공고문을 게시하였고 사업 수행 기관 누리집을 통해 사업 안내를 진행하였

으며, 인스타그램에 사업 참여에 대한 홍보 메시지를 게시함으로써 사업의 취지에 공감한 참여자의 자발적 참여를 유도하였다.

<그림 4> 홈페이지 및 누리소통망 홍보 화면

<수행 기관 누리집 팝업 안내>

<인스타그램 게시글 홍보>



2. 저작권 이용 허락 계약 체결

본 사업 참여자는 온라인 게시 자료를 작성한 원문 자료의 저작권자로, 참여자와의 저작권 이용 허락 계약 체결이 필요하다. 특히, 본 사업은 온라인 게시 자료를 수집하여 말뭉치로 구축하는 일뿐만 아니라, 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공 및 배포하는 것을 목적으로 하므로, 저작권자로부터 저작권 이용 허락 계약이 선행되어야 한다. 향후 발생할 수 있는 법률적 분쟁을 최소화하고 민간 활용도를 제고하기 위해 저작권 이용 허락 계약 체결은 본 사업에서 필수적인 과정이다.

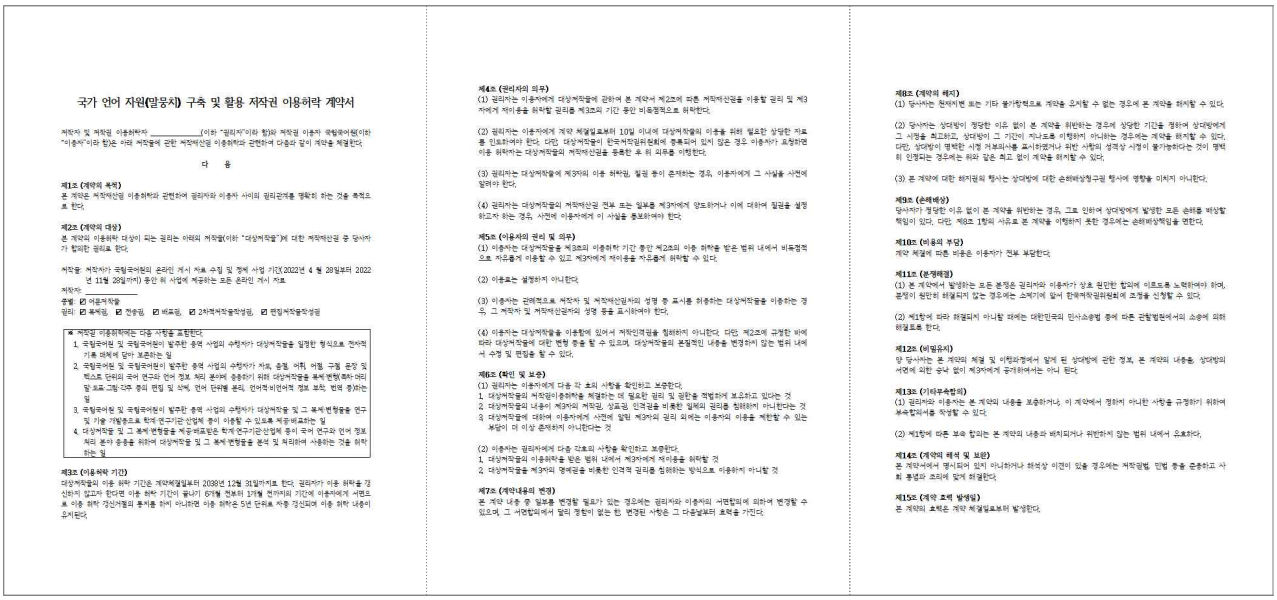
2-1. 저작권 이용 허락 계약의 내용

저작권 이용 허락 계약서 양식 및 내용은 법률 검토를 거친 후 최종적으로 확정하였다. 본 사업의 특성에 따라 대상 권리의 내용은 복제권, 전송권, 배포권, 2차적 저작물 작성권, 편집 저작물 작성권을 포함하며, 저작권 이용 허락 계약에 사용된 계약서의 세부 내용은 다음과 같다.

※ 저작권 이용 허락에는 다음 사항을 포함한다.

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

<그림 5> 저작권 이용 허락 계약서



국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

저작자 및 계약일 이용허락자 _____(이하 "권리자"이라 함)와 계약일 이용자 생활정보원(이하 "이용자"이라 함)은 아래 저작물에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

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

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

저작물: 저작자가 국립국어원의 온라인 게시 서비 수집 및 정제 사업 기간(2022년 4 월 28일부터 2022년 11월 30일까지) 동안 위 사업에 제공하는 모든 온라인 게시 자료

저작자: _____
유형: ☑ 어문저작물
범위: ☑ 복제물, ☑ 전송물, ☑ 배포물, ☑ 2차적저작물작성물, ☑ 편집저작물작성물

■ 저작물 이용허락에는 다음 사항을 포함한다.
1. 국립국어원 및 국립국어원이 발주한 후속 사업의 수행자가 대상저작물을 일정한 형식으로 전사적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 후속 사업의 수행자가 저장, 출력, 어휘, 어절, 구절 분석 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위한 대상저작물을 복제·변형(특히 어미 및 표준 그림 각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적 비언어적 정보 표적, 번역 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 후속 사업의 수행자가 대상저작물 및 그 복제·변형물을 연구 및 기술 개발목표로 학계 연구기관 산업체 종이 이용할 수 있도록 제공비교하는 일
4. 대상저작물 및 그 복제·변형물을 제공비교받은 학계·연구기관·산업체 종이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용허락 기간)
대상저작물의 이용 허락 기간은 계약체결일로부터 2038년 12월 31일까지로 한다. 권리자가 이용 허락을 갱신하지 않고자 한다면 이용 허락 기간이 끝나기 6개월 전부터 1개월 전까지의 기간에 이용자에게 서면으로 이용 허락 갱신거절의 통지를 하지 아니하면 이용 허락은 5년 단위로 자동 갱신되며 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)
(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조의 따른 저작재산권을 이용할 권리 및 제3자에게 제이용을 허락할 권리를 제3조의 기간 동안 비종행적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있어 경우 이용자가 요청하면 이용 허락자는 대상저작물의 저작재산권을 증명한 후 위 의무를 이행한다.

(3) 권리자는 대상저작물에 제3자의 이용 허락일, 철점 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

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

제5조 (이용자의 권리 및 의무)
(1) 이용자는 대상저작물을 제3조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있고 제3자에게 제이용을 자유롭게 허락할 수 있다.

(2) 이용료는 설정하지 아니한다.

(3) 이용자는 권리자로부터 저작자 및 저작재산권자의 성명 중 표시를 허용하는 대상저작물을 이용하는 경우, 그 저작자 및 저작재산권자의 성명 중을 표시하여야 한다.

(4) 이용자는 대상저작물을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 규정한 바에 따라 대상저작물에 대한 변형 중을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 수정 및 편집을 할 수 있다.

제6조 (확인 및 보증)
(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.
1. 대상저작물의 저작권리자중하여을 처결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. 대상저작물이 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
3. 대상저작물에 대하여 이용자에게 사전에 알림 제3자의 권리 피해는 이용자의 이용을 제한할 수 있는 복잡이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.
1. 대상저작물의 이용허락을 받은 범위 내에서 제3자에게 제이용을 허락할 것
2. 대상저작물을 제3자의 명예감을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

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

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

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

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

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

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

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

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

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

제13조 (기타부속합의)
(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 효력한다.

제14조 (계약의 해석 및 보완)
본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준종하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)
본 계약의 효력은 계약 체결일로부터 발생한다.

저작권 이용 허락 기간은 계약 체결일부터 2038년 12월 31일까지로 하되, 이용 허락 기간이 끝나기 6개월 전부터 1개월 전까지의 기간에 이용자가 서면으로 이용 허락 갱신거절의 통지를 하지 아니하면 이용 허락은 5년 단위로 자동 갱신되며 이용 허락 내용이 유지된다.

계약 진행 과정과 계약서 원본 관리 과정에서 개인정보 유출 등의 문제를 최소화하기 위해, 계약서 서명란에 주소나 연락처, 주민번호 뒷자리 등의 정보는 포함하지 않도록 하였다.

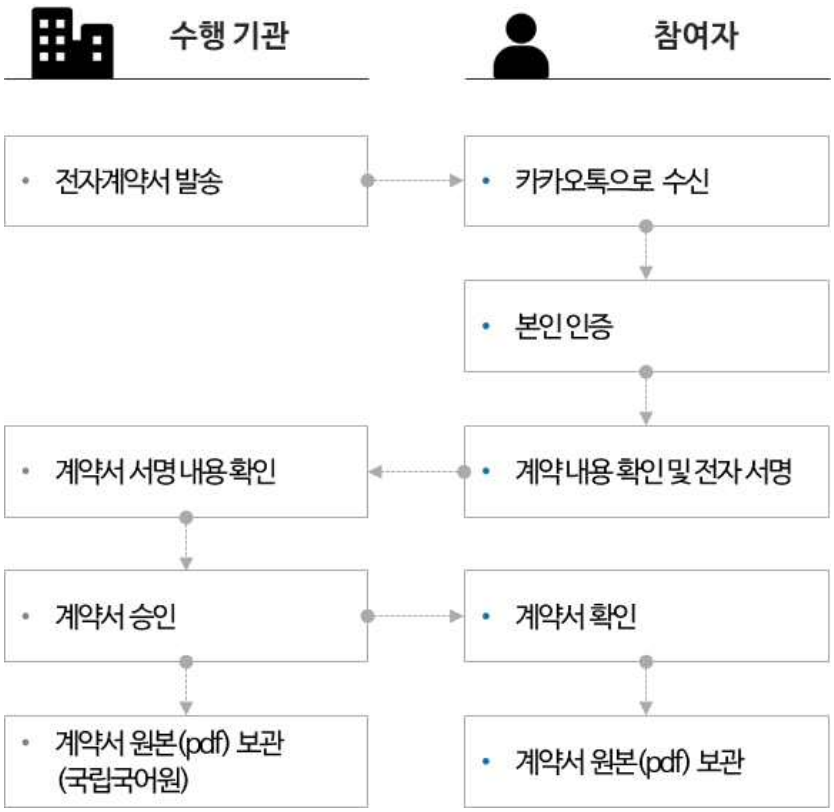
2-2. 저작권 이용 허락 계약 체결

저작권 이용 허락 계약을 체결하기 위해서는 저작권자를 만나 계약을 체결하고, 상호간 계약서 각 1부씩을 보관하는 것이 원칙이나, 대면을 통해 계약을 체결하기 위해서는 많은 시간과 비용이 소요되며, 참여자 모집과 참여자 본인의 시간과 공간적 제약이 크기 때문에 전자 계약 시스템을 이용하여 계약을 진행하였다.

전자 계약 시스템을 통해 본인 인증 절차를 거침으로써, 참여자 본인에 대한 인증은 물론 계약서에 직접 개인정보를 기재해야 하는 번거로움을 최소화할 수 있었다.

본 사업을 위해 저작권 이용 동의 계약을 완료한 참여자는 총 463명이나, <표 2>의 비적합 자료 기준에 따라 26명의 계약자는 제외하고 최종 437명의 계약자만을 선정하였다.

<그림 6> 저작권 이용 허락 전자 계약 진행 절차



3. 온라인 게시 자료 수집

온라인 게시 자료의 수집은 사업 수행 기관 자체 개발 수집기인 'Buzz Crawler'를 이용하여 수집하였다. 본 사업의 경우 참여자의 계정 또는 게시물 단위로 게시물을 수집하는 방식으로 진행하였다.

누리소통망인 인스타그램과 페이스북은 참여자 계정 기준의 게시 자료 수집, 게시판의 경우는 개별 게시 자료의 URL 주소를 추출하여 게시글 단위로 수집하는 방식을 사용하였다.

제 3 장

말뭉치 구축

1. 데이터 분류 및 정제

본 사업의 목적에 따라 비적합 자료와 비윤리적 표현이 포함된 자료는 정제 또는 별도 문서로 분리하는 과정을 진행하였다.

1-1. 비적합 자료 정제

우선 비문서, 비국문 자료는 말뭉치 대상에서 제외하였다. 비문서는 이미지, 스티커, 사진, 동영상, 파일 링크, 웹 주소 등으로만 문서가 구성된 경우이며, 해시태그로만 구성된 게시물 역시 비문서로 규정하여 제외하였다. 또한, 전문이 외국어로 구성된 게시물 역시 말뭉치 대상에서 제외하였다.

다음으로, 중복 게시 자료와 펌글 역시 제외하였다. 누리소통망의 경우 작성자는 다르지만 자료의 내용은 동일한 경우가 많아, 이 경우 동일 문서로 인식하여 1건으로 인정하고 중복 문서 건은 삭제하였다. 펌글로만 구성된 게시 자료 역시 본인이 직접 작성하지 않은 문서이므로 비적합 자료로 처리하였다. 또한, 동일한 계정에서 4천 건을 초과하여 수집된 게시 자료 역시 자료의 다양성을 위해 삭제로 처리하였다.

<표 2> 비적합 자료 기준

비적합 자료 대상	세부 기준
비대상 기간	2019년 9월 이전 작성 자료 삭제
비문서	이미지, 스티커, 사진, 동영상, 파일 링크, 웹 주소, 해시태그로만 구성된 게시 자료 삭제
비국문 자료	전문 외국어로 구성된 게시 자료 삭제
중복글, 펌글	중복 게시 자료 삭제, 펌글(기사, 타인이 작성한 게시물 등)로만 구성된 게시 자료 삭제
계정별 한도 초과 자료	계정별 4천 건 초과 게시 자료 삭제

1-2. 비윤리적 언어 표현 자료 분리

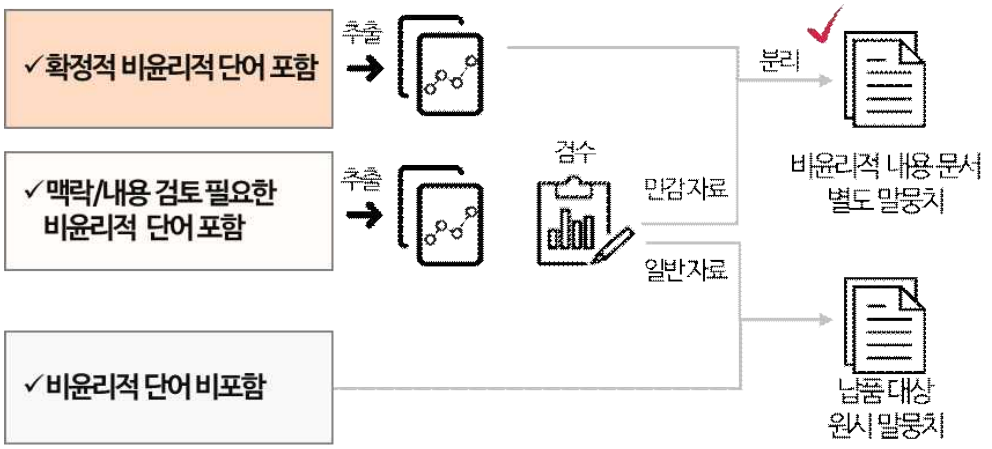
비윤리적 언어 표현이 포함된 게시 자료는 별도로 분리하여, 납품 대상 원시 말뭉치에 포함되지 않도록 하였다. 직접적인 비윤리적 언어 표현으로 지정한 단어, 맥락 및 내용 기준에서 비윤리적 표현에 해당하는 단어 단위를 기준으로 게시 자료를 선별하였다. 욕설, 비어, 속어, 차별 표현, 혐오 표현이 포함된 게시물을 선별해 별도 말뭉치 세트로 구성하였으며, 비윤리적 언어 표현은 국립국어원의 '말뭉치 언어의 사회적 인식 조사·분류'(2021) 내용을 기준으로 하였다.

최근 비윤리적 언어 표현 역시 다양한 비속어로 나타나는 경우가 많고 자음으로만 표현하는 등 유형이 다양하여 비윤리적 언어 표현을 정교하게 분리하는 과정에 어려움이 있었다. 이에 따라 표현 유형과 연관된 다양한 용어로 확장하여 비윤리적 언어 표현을 선별함으로써 원시 말뭉치에서 제외되도록 하였다.

<표 3> 말뭉치 언어의 비윤리적 표현 유형

표현 유형	내용
혐오 표현	특정 개인 및 집단과 이들이 가진 속성에 대하여 적의, 혐오의 감정을 명시적으로 드러내는 표현
성적 표현	특정 개인 및 집단을 성적으로 묘사하거나 불필요한 맥락에서 특정 신체 부위 및 성적 행위를 적나라하게 드러내는 표현
욕설 표현	격이 낮고 속된 말, 대상을 얕잡아 보고 경멸하는 태도를 드러내거나 타인에게 불쾌감을 주는 표현
차별적 표현	암묵적으로 특정 개인 및 집단을 분리하고 불평등하게 대우하는 표현
기타	위의 4가지 유형에 해당하지는 않지만, 사회적으로 용인되지 않는 표현

<그림 7> 민감 자료 데이터 분리



1-3. 비식별화 처리

참여자와 저작권 이용 허락 계약을 완료하였더라도 개인정보가 노출되지 않도록 비식별화 처리가 필요하며, 비식별화 처리 대상 기준에 따라 비식별 조치를 진행하였다. 이름, 이메일 주소 등 계정 정보, 주민등록번호, 카드 번호, 전화번호 등 각종 번호 및 비밀번호, 상세 주소, 출신 및 소속 등의 개인정보와 관련된 사항은 노출되지 않도록 비식별화 처리 유형에 포함하였다. 단, 정치인, 연예인 등 공인의 이름은 비식별화 제외하고 주소의 경우 동 이하의 구체적인 주소만 비식별화 처리하였다.

<표 4> 비식별화 처리 유형

비식별화 유형	비식별화 표지	설명
이름	&name&	개인의 실명 (정치인, 연예인 등 공인 제외)
온라인 계정(아이디)	&account&	특정 사이트의 온라인 계정
고유 식별 번호 (주민등록번호)	&social-security-num&	개인의 주민등록번호
전화번호	&tel-num&	휴대폰 번호, 사업장 번호 등
카드 번호	&card-num&	신용카드 번호 등
기타 번호	&num&	비밀번호 등 기타 비식별화 대상 번호
주소	&address&	동 이하의 상세 주소
출신 및 소속	&affiliation&	개인의 출신 및 소속
기타 비식별화가 필요한 항목	&others&	위 항목 외 기타 비식별화 대상

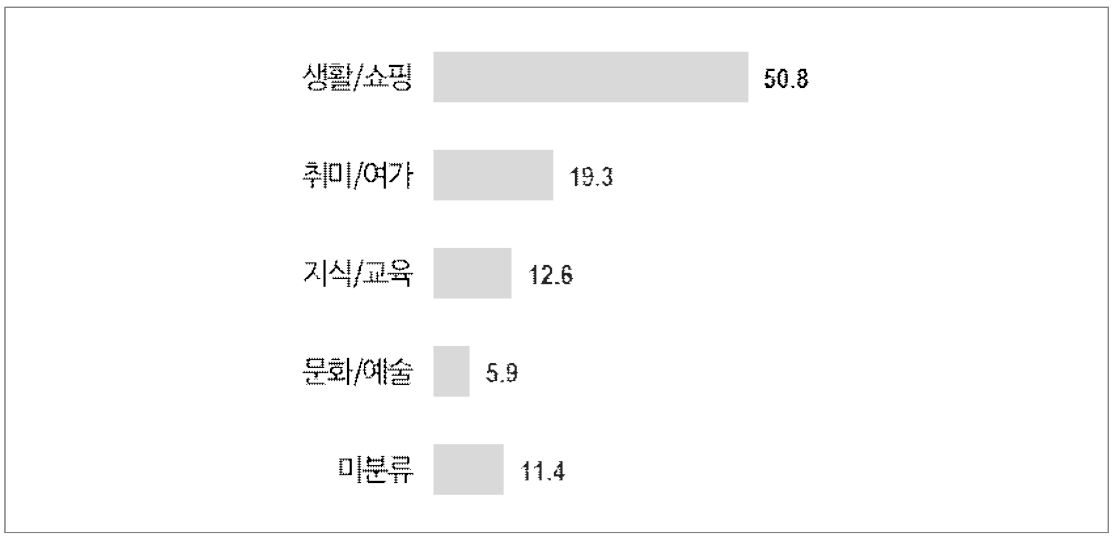
1-4. 데이터 분류(분야별, 주제별)

현대 한국어 사용자의 언어 사용 양상을 반영한 다양한 주제의 자료를 수집하기 위해 분야별·주제별 목표 비중을 설정하고 목표 비중에 맞추어 게시 자료를 수집하도록 하였다.

분야별 목표 비중의 기준을 설정하고, 실제 온라인 게시 자료의 현실적인 데이터를 반영하고자 국립국어원 '웹 말뭉치 구축'(2019년) 자료 중 인스타그램 1,000건을 사전에 임의 분류하여 분야별 비중을 분석해보았다.

분야는 포털사이트의 게시 자료 카테고리를 참고로 1차적으로 구성하였다. 그리고 임의 분류를 진행하면서 1차적인 기준을 통합 및 재편하여 문화/예술, 생활/쇼핑, 취미/여가, 지식/교육 4개 분야로 설정하였다. 임의 분류 결과, 인스타그램 특성상 일상생활, 육아, 요리 등 생활/쇼핑과 관련된 게시글이 절반을 차지하였으며, 취미/여가, 지식/교육 순으로 높은 비중을 차지하였다.

<그림 8> 분야별 임의 분류 결과

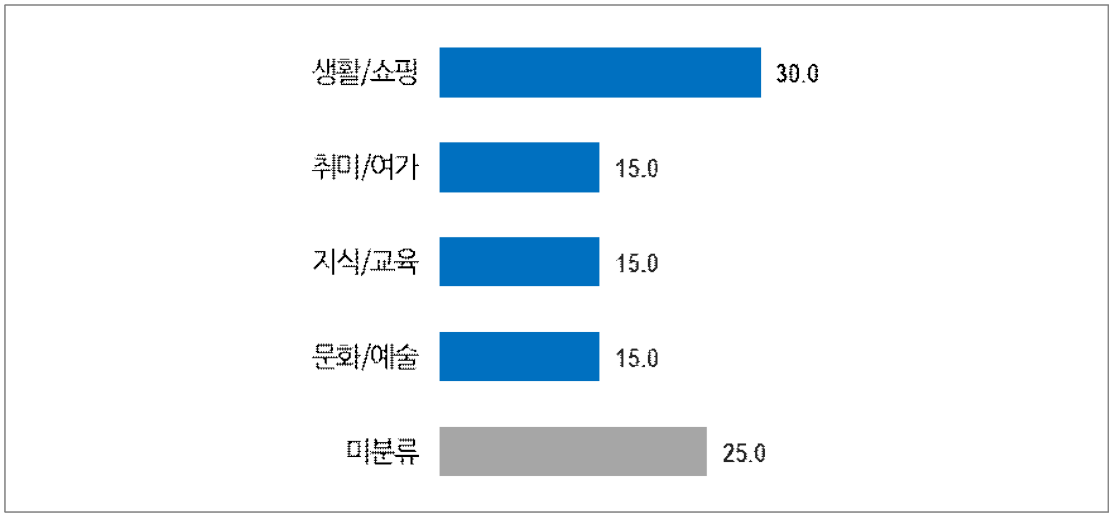


(N=1,000 / 단위 : %)

본 사업의 수집 대상 사이트가 인스타그램 외에도 페이스북과 게시판이 포함된 점을 고려해 임의 분류 결과를 반영하되, 분야별 최소 15% 이상의 게시 자료를 확보하고 균등한 자료 수집을 위해 비중이 상대적으로 낮은 취미/여가, 지식/교육, 문화/예술 3개 분야는 15% 이상, 생활/쇼핑은 30% 이상 확보를 목표로 설정하였다. 또한 글이 단순하거나 주

제 판단이 불가능하여 분야를 분류할 수 없는 경우는 25% 미만이 되도록 목표를 정하였다.

<그림 9> 분야별 수집 목표 비중



(단위 : %)

특정 주제가 없는 미분류를 제외한 4개 분야, 21개 주제에 대한 기준을 설정하여 분류하였으며, 주요 키워드를 중심으로 1차 분류 진행 후 미분류 게시 자료를 재분류함으로써 미분류 비중을 최소화하였다.

<표 5> 분야 및 주제 분류 기준

분야	주제	세부 주제	분야 및 주제 기준 예시
생활/쇼핑	일상생활	일상생활	일기, 내면 다짐, 추억 회상, 슬픈 이야기, 유머, 데일리그램, 브이로그, 출근인증, 일상스타그램, 출근스타그램
		자기계발	미라클모닝, 루틴, 오늘의 다짐, 자기계발
	가족/연애/모임	연애/결혼/가족/모임	결혼, 신혼, 가족, 기념일, 웨딩/예물, 럽스타그램, 데이트, 연애담, 결별 이야기, 파티, 모임, 행사, 동호회, 친목회
		임신출산/육아	임신, 출산, 육아, 아이, 육아맘일상, 육아맘스타그램, 육아그램, 맘스타그램, 아가그램, 아가스타그램, 아기 사진
	요리/음식	요리/음식	쿡스타그램, 밀키트, 집밥, 요리, 레시피, 홈파티, 홈카페, 홈레스토랑, 술/와인/위스키
	인테리어/집꾸미기	인테리어/집꾸미기	집꾸미기, 인테리어, 홈스타일링, 리모델링, 온라인집들이, 집소개
	패션/미용/뷰티	패션/미용/뷰티	OOTD, 코디, 헤어, 메이크업, 피부, 네일, 성형, 복부관리, 다이어트, 바디프로필

분야	주제	세부 주제	분야 및 주제 기준 예시
	반려동식물/키우기	반려동식물/키우기	개, 고양이, 파충류, 식물, 멍스타그램, 냥스타그램, 견스타그램, 독스타그램, 캣스타그램, 애견용품, 길고양이
	제품 구매/소개	제품 구매/소개	한정판, 명품, 대란템, 받은 선물, 제품 판매글, 제품 자랑, 개봉기, 할인권, 협찬 제품
문화/예술	**영화/드라마/방송**	영화/애니메이션/OTT	영화, 애니메이션, 극장, 상영작 소개, 넷플릭스, 개봉, 영화 시사회
		드라마/방송/예능	TV 방송, 드라마 시청 후기, 예능짤, 방송프로그램
	공연/전시/박람회	미술/공예/박람회	미술/그림, 조각, 미술관, 박람회, 뮤지엄, 전시회, 박물관
		뮤지컬/연극/공연	뮤지컬, 오페라, 연극, 발레/무용 공연
	도서/문학	도서/문학	문학 작품, 도서소개, 서평, 오디오북, 만화책, 웹툰
	게임/캐릭터	게임	게임, 게임 캐릭터, 모바일 게임, 일반 캐릭터 포함
	음악/음반/콘서트	음악/음반/콘서트	클래식, 재즈, 국악, 음악감상, 음반, 음질, 사운드, 연주회, 콘서트
	인물/연예인	연예인/유명인	연예인, 유명인, 유명인 인맥 자랑, 팬덤, 팬클럽
지식/교육	**교육/취업/직업**	교육/학원/강의	강의, 학원, 강연회, 수업, 과목, 외국어 학습, 코딩교육
		직업/취업/자격증	합격 수기, 공시생, 자격증 준비, 취업, 면접, 채용 정보, 이직, 퇴직, 아르바이트, 자소서
	정치/경제/사회	경제/재테크	주식, 투자, 금융, 재테크, 코인
		정치/사회	대통령, 정치인, 정당, 정책, 선거
	의료/건강	의료/건강	병원, 질병, 질환, 약품, 입원 일지, 투약 일지, 건강스타그램, 건강기능식품
	기타 지식 공유/전문 지식	공학/ IT/과학	이공계 관련 지식 정보, 전문지식, 전문가
		인문학/법학	인문/법학 관련 지식 정보, 전문지식, 전문가
취미/여행	**여행/관광**	여행/관광	여행지, 관광지, 숙소, 여행 교통편, 맛집, 풍경, 핫플, 캠핑, 차박, 여행상품
	운동/스포츠	운동/스포츠	등산, 골프, 홈트, 낚시, 라이딩, 하이킹, 헬스, 헬린이, 오하운
	종교/봉사	종교/봉사	플로킹, 봉사활동, (돕는) 챌린지, 헌혈, 예배, 교회, 사찰, 성당, 부활절
	기타 취미활동	기타 취미활동	취미 결과물 자랑, 만들기, DIY, 수채화, 뜨개질, 그림그리기, 덕질 자랑글
미분류	**특정 주제 없음 단일 주제 판단 불가**		·주제 없는 모호한 내용 ·글이 짧아 주제 판단이 불가한 내용 ·포함되는 분야 및 주제 없는 경우 ·복합적인 내용으로 단일 주제로 판단하기 어려운 내용

실제 수집 진행 결과, 인스타그램이 전체의 87%를 차지하는 자료의 특성상, 생활/쇼핑 비중이 44%로 높은 비중을 차지하였다. 미분류 비중은 재분류를 통해 최소화하여 목표 비중 대비 절반 수준으로 낮추었다.

그러나, 임의 분류 결과와 마찬가지로, 수집이 약 80% 이루어진 상태에서 중간 점검을 한 결과 '문화/예술' 비중이 5% 수준으로 매우 낮았으며, '지식/교육' 분야 역시 10% 수준으로 목표 비중 대비 부족한 상황이었다.

자발적 참여만으로는 분야별 목표 비중 달성이 불가능하여, 약 80% 수집이 이루어진 시점 이후에는 수행 기관 내부 자료를 활용하여 '문화/예술', '지식/교육' 관련 게시물 보유 계정을 선별해 추가 모집하는 과정을 통해 목표 비중을 확보하였다. 최종 수집한 분야 및 주제별 비중은 다음과 같다.

<표 6> 분야별 비중

분야	문서 수(건)	비중	※ 참고. 목표 비중
생활/쇼핑	134,606	43%	30%
문화/예술	46,562	15%	15%
지식/교육	47,291	15%	15%
취미/여행	46,516	15%	15%
특정 주제 없음/주제 판단 불가	35,025	11%	25%
합계	310,000	100%	100%

참여자 게시물을 목표 비중을 설정하지 않고 수집을 진행했다면, 사이트별로 분야의 비중을 파악하는 의미가 있을 수 있으나, 의도적으로 분야별 목표 비중을 두어 할당한 것이므로 사이트별 분야 비중은 본 사업에서는 파악하지 않았다.

분야 내 주제별 비중은 '여행/관광' 관련 비중이 전체의 10%로 가장 높은 비중을 차지하였으나, 전반적으로 다양한 주제별 게시 자료가 수집되었다. 주제별 비중은 다음과 같다.

<표 7> 분야 내 주제별 비중

분야	주제	문서 수(건)	비중 (%)
생활/쇼핑	일상생활	18,777	6.1
	가족/연애/모임	29,683	9.6
	요리/음식	24,658	8.0
	인테리어/집꾸미기	6,345	2.0
	패션/미용/뷰티	20,701	6.7
	반려동식물/키우기	14,377	4.6
	제품 구매/소개	20,065	6.5
문화/예술	영화/드라마/방송	14,783	4.8
	공연/전시/박람회	8,057	2.6
	도서/문학	11,999	3.9
	게임/캐릭터	2,876	0.9
	음악/음반/콘서트	6,388	2.1
	인물/연예인	2,459	0.8
지식/교육	교육/취업/직업	12,881	4.2
	정치/경제/사회	9,419	3.0
	의료/건강	21,351	6.9
	기타 지식 공유/전문 지식	3,640	1.2
취미/여행	여행/관광	31,747	10.2
	운동/스포츠	9,118	2.9
	종교/봉사	3,013	1.0
	기타 취미활동	2,638	0.9
미분류	특정 주제 없음/주제 판단 불가	35,025	11.3
합 계		310,000	100.0

2. 원시 말뭉치 구축 및 메타 정보 구축

수집과 분류 및 정제 작업이 완료된 온라인 게시 자료는 국립국어원의 원시 말뭉치 구축 지침에 지정된 항목과 형식을 기준으로 원시 말뭉치 자료를 구축하고 JSON 파일 형태로 출력하였다. 파일명 부여 방식과 JSON 형태의 말뭉치 형식은 다음과 같다.

<표 8> 파일명 부여 방식

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	8자리 일련번호
E: 온라인 게시 자료 말뭉치	S: 누리소통망 P: 게시판	OR: 원문 자료 RW: 원시 말뭉치	22	00000001~ 99999999

<표 9> 말뭉치 형식(JSON)

1 수준	2 수준	3 수준	설명
id			말뭉치 파일 ID
metadata			파일의 메타 정보
	title		국립국어원 온라인 게시 자료 원시 말뭉치 [파일ID]
	creator		생성자(국립국어원)
	distributor		배포자(국립국어원)
	year		말뭉치 구축 연도(2022)
	category		분류
	annotation_level		분석 층위 (원시)
	sampling		샘플링 방식(게시자 모집 후 무작위 추출)
document			문서 정보
	id		문서 ID
	metadata		문서의 메타 정보
		title	문서 제목
		author	작성자
		publisher	게시 플랫폼
		date	작성일시, 게시일시
		topic	주제
		crawl_date	크롤링 일시
		url	URL 주소
	paragraph		문단
		id	문단 ID
		form	정제된 형태
		original_form	원문 표기된 그대로의 형태(개인 정보 비식별화 후)

원시 말뭉치 구축 지침에 따라 출력한 말뭉치 형식(JSON) 납품 형태는 다음과 같다.

<그림 10> 누리소통망 말뭉치(JSON) 출력 예시


```
{
"id": "ESRW2200000001.71976",
"metadata": {
   "title": "N/A",
   "author": "__hyelyn__",
   "publisher": "인스타그램",
   "date": "20191223",
   "topic": "생활/쇼핑_가족/연애/모임",
   "crawl_date": "2022062210:06:30",
   "url": "https://www.instagram.com/p/B6Ze5i8lqcG/"
},
"paragraph": [
   {
   "id":"ESRW2200000001.71976.1",
   "form": "찌우 생애 첫 산타",
   "original_form": "찌우 생애 첫 산타😇"
   },
   {
   "id":"ESRW2200000001.71976.2",
   "form": "원생활을 늦게 시작한덕에 이제야",
   "original_form": "원생활을 늦게 시작한덕에 이제야"
   },
   {
   "id":"ESRW2200000001.71976.3",
   "form": "산타를 처음 만난 네쨜 언니",
   "original_form": "산타를 처음 만난 네쨜 언니👉👈"
   },
   {
   "id":"ESRW2200000001.71976.4",
   "form": "첫 산타는 트니트니에서 산타가",
   "original_form": "첫 산타는 트니트니에서❤ 산타가"
   },

                                    - 이하 생략 -
```

<그림 11> 게시판 말뭉치(JSON) 출력 예시


```
{
"id": "EPRW2200000001.2435",
"metadata": {
   "title": "불닭볶음면 개발의 비밀?.humor",
   "author": "스퀴니",
   "publisher": "클리앙",
   "date": "20210226",
   "topic": "생활/쇼핑_일상생활",
   "crawl_date": "2022062810:06:00",
   "url": "https://www.clien.net/service/board/park/15925539?/"
},
"paragraph": [
   {
   "id":"EPRW2200000001.2435.1",
   "form": "개발진이 회장님 암살시도 하려고 했으나...",
   "original_form": "개발진이 회장님 암살시도 하려고 했으나..."
   },
   {
   "id":"EPRW2200000001.2435.2",
   "form": "알고보니 회장님이 매운걸 좋아하신거 아니냐고..",
   "original_form": "알고보니 회장님이 매운걸 좋아하신거 아니냐고.."
   },
   {
   "id":"EPRW2200000001.2435.3",
   "form": "ㅋㅋㅋㅋㅋㅋㅋ",
   "original_form": "ㅋㅋㅋㅋㅋㅋㅋ"
   }
]
}
```

참고문헌

국립국어원(2019), 웹 말뭉치 구축, 국립국어원.

국립국어원(2021), 말뭉치 언어의 사회적 인식 조사·분류, 국립국어원.

국립국어원(2021), 2021년 온라인 대화 자료 수집 및 정제, 국립국어원.

부록

[붙임 1] 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

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

다 음

제1조 (계약의 목적)

본 계약은 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)

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

저작물: 저작자가 국립국어원의 2022년 온라인 게시 자료 수집 및 정제 사업 기간(2022년 4 월 28일부터 2022년 11월 28일까지) 동안 위 사업에 제공하는 모든 온라인 게시 자료
저작자: _____
종별: ☑ 어문저작물
권리: ☑ 복제권, ☑ 전송권, ☑ 배포권, ☑ 2차적저작물작성권, ☑ 편집저작물작성권

※ 저작권 이용허락에는 다음 사항을 포함한다.

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

제3조 (이용허락 기간)

대상저작물의 이용 허락 기간은 계약체결일부터 2038년 12월 31일까지로 한다. 권리자가 이용 허락을 갱신하지 않고자 한다면 이용 허락 기간이 끝나기 6개월 전부터 1개월 전까지의 기간에 이용자에게 서면으로 이용 허락 갱신거절의 통지를 하지 아니하면 이용 허락은 5년 단위로 자동 갱신되며 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)

(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을 이용할 권리 및 제3자에게 재이용을 허락할 권리를 제3조의 기간 동안 비독점적으로 허락한다.

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

(3) 권리자는 대상저작물에 제3자의 이용 허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

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

제5조 (이용자의 권리 및 의무)

(1) 이용자는 대상저작물을 제3조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있고 제3자에게 재이용을 자유롭게 허락할 수 있다.

(2) 이용료는 설정하지 아니한다.

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

(4) 이용자는 대상저작물을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 규정한 바에 따라 대상저작물에 대한 변형 등을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 수정 및 편집을 할 수 있다.

제6조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 대상저작물의 저작권이용허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
3. 대상저작물에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.

1. 대상저작물의 이용허락을 받은 범위 내에서 제3자에게 재이용을 허락할 것
2. 대상저작물을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)

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

제8조 (계약의 해지)

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

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

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

제9조 (손해배상)

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

제10조 (비용의 부담)

계약 체결에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁해결)

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

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

제12조 (비밀유지)

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

제13조 (기타부속합의)

(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (계약의 해석 및 보완)

본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)

본 계약의 효력은 계약 체결일로부터 발생한다.

2022년 월 일

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

〈기획・연구〉

국립국어원 강미영 언어정보과장

국립국어원 이보라미 학예연구관

국립국어원 서샛별 학예연구사

국립국어원 윤희상 연구원

〈사업 참여자〉

사업 책임자 이영희 (㈜버즈메트릭스)

사업 참여자 김수진 (㈜버즈메트릭스)

신현주 (㈜버즈메트릭스)

김도현 (㈜버즈메트릭스)

이진상 (㈜버즈메트릭스)

유지현 (㈜버즈메트릭스)

권주원 (㈜버즈메트릭스)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금낭화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2022년 11월 28일

발행일: 2022년 11월 28일

인 쇄: ㈜타라그래픽스

※ 이 책은 국립국어원의 용역비로 수행한 ‘2022년 온라인 게시 자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.