

국립국어원 2021년 온라인 대화 자료 수집 및 정제 사업 관련 안내

국립국어원에서는 우리말 인공지능 기술 개발과 국어 연구 등에 활용하고자 온라인 대화 자료를 수집하여 가공하는 ‘2021년 온라인 대화 자료 수집 및 정제’ 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.

□ 사업 개요

- 사업명: 2021년 온라인 대화 자료 수집 및 정제
 - 사업 기간: 2021. 4. 13. ~ 2021. 11. 13.
 - 사업 수행자: (주)미디어코퍼스·(주)다이얼로그디자인에이전시·심심이(주)
 - 주요 사업 내용
 - 온라인 대화(15만 대화 이상) 원문 자료 수집
 - 수집된 온라인 대화 자료를 대상으로 기초(원시) 말뭉치* 구축
- * **말뭉치**: 컴퓨터가 읽을 수 있는 형태로 입력, 분석한 대규모 언어 자료로, 사람에게 학습용 책이 필요하듯 인공지능은 학습용 대규모 언어 자료가 필요함.
- 담당자: 국립국어원 언어정보과 유희정(02-2669-9638)

□ 주요 질의·답변

1. 온라인 대화 자료를 수집하는 목적은?

- 메시지를 통해 실제로 나눈 대화 자료를 모아 컴퓨터가 읽을 수 있는 형태로 분석한 말뭉치를 국가적으로 구축하여 **우리말 인공지능 개발과 국어 연구 등에 공공 자료로 활용할 수 있도록** 하기 위해서입니다.

2. 저작권 이용 허락 범위는?

- 국립국어원과 국립국어원의 용역 사업 수행자가 귀하의 온라인 대화 자료를 말뭉치로 구축하고 배포하기 위하여 아래 일을 할 수 있도록 허락을 해 주시는 것이 필요합니다.
 - 수집 자료를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
 - 수집 자료를 형태소, 단어, 문장 등의 언어 단위별로 분리하며, 언어적·비언어적 정보를 부착하는 등 자료를 복제하여 변형하여 말뭉치를 구축하는 일
 - 구축된 말뭉치를 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
- 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용 등을 위하여 아래 일을 할 수 있도록 허락을 해 주시는 것이 필요합니다.
 - 우리말 인공지능 기술 개발과 국어 연구용으로 말뭉치를 분석 및 처리하여 사용하도록 하는 일

3. 저작권 이용 허락 기간은?

- 학계·연구기관·산업체 등이 연구 및 기술 개발에 활용하기 위해서는 **충분한 기간 동안 안정적으로 말뭉치를 이용할 수 있는 것이 중요합니다.** 예를 들어 1990년대 초반에 영국에서 구축한 BNC(British National Corpus) 말뭉치는 25년이 지난 현재까지도 안정적으로 제공되어 활용되고 있습니다. 국립국어원에서는 귀하의 소중한 온라인 대화 자료를 말뭉치로

구축하여 최소 2042년 12월 31일까지는 안정적으로 이용할 수 있도록 허락해 주시기를 바랍니다.

- 귀하께서 이용 허락 중지 의사를 밝히시면 최소 이용 허락 기간이 끝난 후 즉시 이용을 중지할 예정입니다.

4. 온라인 대화 말뭉치는 어떠한 형식으로 구축되는 것인지?

- 귀하께서 메신저를 통해 나누신 대화 자료의 원문을 수집하고, 수집된 자료에 말뭉치의 형식을 갖추기 위한 정보를 추가하여 원시 말뭉치로 구축합니다. 여기에 형태소, 어휘, 문장과 관련된 언어적 정보를 추가하여 분석 말뭉치로 구축할 수 있습니다.

〈원시 말뭉치 예시〉

```
"id":"MDRW2100000003",
"metadata":{
  "title":"국립국어원 온라인 대화 말뭉치 MDRW2100000003",
  "creator":"국립국어원",
  "distributor":"국립국어원",
  "year":"2021",
  "category":"온라인 대화 > 2인 대화",
  "annotation_level":[
    "원시"
  ],
  "sampling":"참여자 모집 후 대화 수집",
  "sampling_method":"실시간 대화 수집"
},
"document":[
  {
    "id":"MDRW2100000003.1",
    "metadata":{
      "title":"온라인 대화",
      "author":"개인 대화 참여자",
      "publisher":"카카오톡",
      "date":"20210427",
      "topic":"트렌드 검색어",
      "satisfaction":"5",
      "beneficial":"3",
      "speaker":[
        {
          "id":"1",
          "age":"20대",
          "occupation":"전문가 및 관련 종사자",
          "sex":"여성",
          "birthplace":"서울",
          "principal_residence":"해외/기타",
```

```

        "current_residence": "서울",
        "device": "스마트폰",
        "keyboard": "2벌식(쿼티)"
    },
    {
        "id": "2",
        "age": "20대",
        "occupation": "전문가 및 관련 종사자",
        "sex": "여성",
        "birthplace": "울산",
        "pricipal_residence": "전북",
        "current_residence": "경기",
        "device": "스마트폰",
        "keyboard": "2벌식(쿼티)"
    }
],
"setting": {
    "relation": "직장 동료",
    "intimacy": 4,
    "contact_frequency": "주 3회 이상"
}
},
"utterance": [
    {
        "id": "MDRW2100000003.1.1.1",
        "form": "지난주 슈퍼맨이 돌아왔다 봐췌요?",
        "original_form": "지난주 슈퍼맨이 돌아왔다 봐췌요?",
        "time": "20210412 17:11",
        "speaker_id": 2
    },
    {
        "id": "MDRW2100000003.1.1.2",
        "form": "ㅠㅠ",
        "original_form": "ㅠㅠ",
        "time": "20210412 17:11",
        "speaker_id": 2
    },
    {
        "id": "MDRW2100000003.1.1.3",
        "form": "아니여 아직 못봤어여ㅠㅠ!!! ",
        "original_form": "아니여 아직 못봤어여ㅠㅠ!!! ",
        "time": "20210412 17:11",
        "speaker_id": 1
    }
],

```

5. 개인정보가 노출될 우려는 없는지?

- 이름, 전화번호, 주소 등 개인정보는 철저히 알아볼 수 없게 처리합니다.