

국립국어원 온라인 대화 말뭉치

(버전 1.1)

- **자료명:** 국립국어원 온라인 대화 말뭉치
- **공개일**
 - (버전 1.0) 2022. 4. 1.
 - (버전 1.1) 2023. 9. 27.
 - principal_residence → principal_residence
- **자료 유형:** 텍스트
- **관련 사업:** 온라인 대화 자료 수집 및 정제(2021)
- **자료 설명**
 - **내용**
 - 두 명 이상의 대화 참여자가 온라인 공간에서 주고받은 대화 자료를 대상으로 구축한 말뭉치.
 - 대화 참여자들의 자연스러운 언어 습관이 그대로 반영되어 있으며, 개인 정보 등은 비식별화 처리함.
 - ※ **구축 방법 및 비식별화 처리에 대한 내용은 ‘국립국어원 누리집 > 자료 > 연구조사 자료’에서 ‘온라인 대화 자료 수집 및 정제’ 사업 보고서를 참고.**
- **분량**
 - 총 74,665건(대화 메시지 3,069,927개)
- **파일 형식:** JSON(UTF-8 인코딩)
- **파일 수 및 크기:** 파일 47,421개, 총 835MB
- **인용:**
 - (국문) 국립국어원(2022). 국립국어원 온라인 대화 말뭉치(버전 1.1). URL: <https://corpus.korean.go.kr>
 - (영문) National Institute of Korean Language (2022). NIKL Online text message Corpus (v.1.1). URL: <https://corpus.korean.go.kr>

· 파일 명명 규칙

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14
속성	매체	대화 참여 인원	주석 단계		구축 연도	일련번호(8자리)								
정의 값	M: 온라인 대화	D: 2인 대화 M: 다자 대화	RW: 원시 말뭉치		21: 2021년	00000001 ~ 99999999 (여덟 자리 일련번호)								
※ 예시: MDRW2100000010.json 2021년에 구축한 2인 온라인 대화 원시 말뭉치 파일 MMRW2100000010.json 2021년에 구축한 다자 온라인 대화 원시 말뭉치 파일														

· 예시

```

{
  "id": "MDRW2100000010",
  "metadata": {
    "title": "국립국어원 온라인 대화 말뭉치 MDRW2100000010",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2021",
    "category": "온라인 대화 > 2인 대화",
    "annotation_level": "원시",
    "sampling": "실시간 대화"
  },
  "document": [
    {
      "id": "MDRW2100000010.1",
      "metadata": {
        "title": "온라인 대화",
        "author": "개인 대화 참여자",
        "publisher": "카카오톡",
        "date": "20210513",
        "topic": "미용과 건강",
        "speaker": [
          {
            "id": "1",
            "age": "30대",
            "occupation": "가정 주부",
            "sex": "여성",
            "birthplace": "강원",
            "principal_residence": "충북",
            "current_residence": "경기",
            "device": "스마트폰",
            "keyboard": "나랏글"
          },
          {
            "id": "2",
            "age": "30대",
            "occupation": "기타",

```

```

    "sex": "남성",
    "birthplace": "제주",
    "principal_residence": "경기",
    "current_residence": "경기",
    "device": "스마트폰",
    "keyboard": "나랏글"
  }
},
"setting": {
  "relation": "가족>부부",
  "intimacy": 5,
  "contact_frequency": "(거의) 매일 연락한다."
}
},
"utterance": [
  {
    "id": "MDRW2100000010.1.1",
    "form": "지금 운동하러가려고하는데 반팔 반바지 입으니까 선크림을 발라야돼",
    "original_form": "지금 운동하러가려고하는데 반팔 반바지 입으니까 선크림을 발라야돼",
    "speaker_id": "1",
    "time": "20210513 16:02"
  },
  {
    "id": "MDRW2100000010.1.2",
    "form": "지금 운동갈꺼야?",
    "original_form": "지금 운동갈꺼야?",
    "speaker_id": "2",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.3",
    "form": "근데 옷만입었는데도 덥네",
    "original_form": "근데 옷만입었는데도 덥네",
    "speaker_id": "1",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.4",
    "form": "어",
    "original_form": "어",
    "speaker_id": "1",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.5",
    "form": "오늘달리기하려고",
    "original_form": "오늘달리기하려고",
    "speaker_id": "1",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.6",
    "form": "이따가 걸기라도 하려고 했는데",
    "original_form": "이따가 걸기라도 하려고 했는데",
    "speaker_id": "2",

```

```

    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.7",
    "form": "혼자 해야겠네",
    "original_form": "혼자 해야겠네",
    "speaker_id": "2",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.8",
    "form": "아 이따밤에?",
    "original_form": "아 이따밤에?",
    "speaker_id": "1",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.9",
    "form": "같이걸지 뭐",
    "original_form": "같이걸지 뭐",
    "speaker_id": "1",
    "time": "20210513 16:03"
  },
  {
    "id": "MDRW2100000010.1.10",
    "form": "저녁이나 밤에",
    "original_form": "저녁이나 밤에",
    "speaker_id": "2",
    "time": "20210513 16:03"
  }
}

```

※ “original_form”: 수집한 언어 자료의 원문을 그대로 유지한 형태(개인 정보 등은 비식별화)
 “form”: 원문에서 연속된 여러 개의 공백(스페이스, 탭 등), 특수 메시지, 비식별화 기호 등을 제거하여 전처리한 형태

※ 특수 메시지

- 이모지 {emoji}
- 선물하기 {system: gift}
- 통화 {system: call}
- 송금 {system: money}
- 공지 {system: notice}
- 지도 공유 {system: map}
- 연락처 공유 {system: contact}
- 메시지 삭제 {system: delete}
- 사진 공유 {share:photo}
- 동영상 공유 {share:video}
- 음악 공유 {share:music}
- 파일 공유 {share:file}

- 음성 메시지 공유 {share:voice}
- 정보 공유 {share:info}
- URL 공유 {share:url}
- 부적절 대화 {censored}

※ 비식별화 기호

- 이름 &name&
- 온라인 계정 &account&
- 고유 식별 번호 &social-security-num&
- 전화번호 &tel-num&
- 금융 번호 &card-num&
- 기타 번호 &num&
- 주소 &address&
- 출신, 소속 &affiliation&
- 기타 &others&

- 자료 내용 문의: 02-2669-9638