문화체육관광부
국립국어원

국립국어원 메신저 말뭉치

(버전 2.0)

· **자료명:** 국립국어원 메신저 말뭉치

· **공개일**
  · (버전 1.0) 2020. 8. 25.
  · (버전 2.0) 2022. 4. 1.
    - 이름 등 비식별화 작업 추가
    - 부적절한 내용 포함 문서 제외
      ※ 2021년 '말뭉치 언어의 사회적 인식 조사·분류' 사업에서 부적절한 내용이 포함된 것으로 평가한 문서를 제외하였음.

· **자료 유형:** 텍스트

· **관련 사업:** 메신저 대화 자료 수집 및 말뭉치 구축(2019)
말뭉치 언어의 사회적 인식 조사·분류(2021)

· **자료 설명**
  · 내용
    - 두 명 이상의 대화 참여자가 메신저로 나눈 메신저 대화 자료
    - 대화 참여자들의 자연스러운 언어 습관이 그대로 반영되어 있으며, 개인 정보 등은 비식별화 처리함.
    ※ 구축 방법 및 비식별화 처리에 대한 내용은 '국립국어원 누리집 > 자료 > 연구·조사 자료'에서 '메신저 대화 자료 수집 및 말뭉치 구축' 사업 보고서를 참고.
    문서 정제에 대한 내용은 '말뭉치 언어의 사회적 인식 조사·분류' 사업 보고서를 참고.

  · 분량
    - 총 3,836건(대화 메시지 691,535개)

  · 파일 형식: JSON(UTF-8 인코딩)

  · 파일 수 및 크기: 파일 3,836개, 총 212MB

· **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: 메신저	D: 2인 대화 M: 다자 대화	RW: 원시 말뭉치		19: 2019년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: MDRW1900000008.json 2019 년에 구축한 메신저 2 인 대화 원시 말뭉치 파일														

· **인용:**

(국문) 국립국어원(2022). 국립국어원 메신저 말뭉치(버전 2.0). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2022). NIKL Messenger Corpus (v.2.0). URL: https://kli.korean.go.kr/corpus

· **예시**


```
{
    "id": "MDRW1900000008",
    "metadata": {
        "title": "국립국어원 메신저 말뭉치 MDRW1900000008",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2019",
        "category": "메신저 대화 > 2인 대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "참여자 모집 후 대화 수집"
    },
    "document": [
        {
            "id": "MDRW1900000008.1",
            "metadata": {
                "title": "메신저 대화",
                "author": "개인 대화 참여자",
                "publisher": "카카오톡",
                "date": "20191219",
                "topic": "식음료 (식사, 음식, 배달, 맛집, 요리)",
                "speaker": [
                    {
                        "id": "1",
                        "age": "20대",
                        "occupation": "가정 주부",
                        "sex": "여성",
```

```
                    "birthplace": "경기",
                    "pricipal_residence": "경기",
                    "current_residence": "경기",
                    "device": "스마트폰",
                    "keyboard": "2벌식(쿼티)"
                },
                {
                    "id": "2",
                    "age": "20대",
                    "occupation": "기타",
                    "sex": "여성",
                    "birthplace": "서울",
                    "pricipal_residence": "경기",
                    "current_residence": "경기",
                    "device": "스마트폰",
                    "keyboard": "천지인"
                }
            ],
            "setting": {
                "relation": "학교/학원 : 동기/동창/동급생",
                "intimacy": 5,
                "contact_frequency": "거의 매일"
            }
        },
        "utterance": [
            {
                "id": "MDRW1900000008.1.1.1",
                "form": "짜쟌",
                "original_form": "짜쟌",
                "speaker_id": "2",
                "time": "20191104 15:30"
            },
            {
                "id": "MDRW1900000008.1.1.2",
                "form": "ㅋㅋㅋ",
                "original_form": "ㅋㅋㅋ",
                "speaker_id": "1",
                "time": "20191104 15:30"
            },
            {
                "id": "MDRW1900000008.1.1.3",
                "form": "오늘 나는 아침에 8시에나와서 두유만먹어가지구",
                "original_form": "오늘 나는 아침에 8시에나와서 두유만먹어가지구",
                "speaker_id": "1",
                "time": "20191104 15:30"
            },
            {
                "id": "MDRW1900000008.1.1.4",
                "form": "name1 끝낫웅?",
                "original_form": "&name1& 끝낫웅?",
                "speaker_id": "2",
                "time": "20191104 15:30"
            },
            {
                "id": "MDRW1900000008.1.1.5",
```

```
            "form": "아이구 힘들었겠다ㅠㅠ",
            "original_form": "아이구 힘들었겠다ㅠㅠ",
            "speaker_id": "2",
            "time": "20191104 15:30"
        },
        {
            "id": "MDRW1900000008.1.1.6",
            "form": "아까 쉬는시간에잠깐 삼김 2개 겨우먹음",
            "original_form": "아까 쉬는시간에잠깐 삼김 2개 겨우먹음",
            "speaker_id": "1",
            "time": "20191104 15:30"
        },
        {
            "id": "MDRW1900000008.1.1.7",
            "form": "ㅋㄷㅋㄷ",
            "original_form": "ㅋㄷㅋㄷ",
            "speaker_id": "1",
            "time": "20191104 15:31"
        },
        {
            "id": "MDRW1900000008.1.1.8",
            "form": "삼각김밥!",
            "original_form": "삼각김밥!",
            "speaker_id": "1",
            "time": "20191104 15:31"
        },
        {
            "id": "MDRW1900000008.1.1.9",
            "form": "너는 ???",
            "original_form": "너는 ???",
            "speaker_id": "1",
            "time": "20191104 15:31"
        },
```


※ “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:videoclip}

- 음악 공유 {share:music}
- 파일 공유 {share:file}
- URL 공유 {share:url}
- 정보 공유 {share:info}

※ 비식별화 기호

- 이름 &name&
- 주민 등록 번호 &social-security-num&
- 카드 번호 &card-num&
- 주소 &address&
- 전화번호 &tel-num&
- 계좌 번호 &account&
- 기타 번호 &num&
- 출신, 소속 &affiliation&
- 기타 &others&

· **자료 내용 문의:** 02-2669-9638