문화체육관광부
국립국어원

국립국어원 일상 대화 말뭉치 2023

(버전 1.0)

· **자료명:** 국립국어원 일상 대화 말뭉치 2023

· **공개일**

· (버전 1.0) 2024. 12. 26.

· **자료 유형:** 텍스트

· **관련 사업:** 2023년 일상 대화 말뭉치 구축(2023)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2023년 일상 대화 말뭉치 구축' 사업 보고서 참고

· 내용

- 16개 주제로 자유롭게 나눈 일상 대화를 전사하여 구성한 말뭉치임.
- 각 대화는 최소 두 명, 최대 네 명의 화자로 구성되어 있으며 대화의 평균 시간은 약 15분(총 2,168명 화자, 총 500시간 분량).
- 일상 대화 자료를 한글로 전사하였으며, 발음 전사(표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등에 실제 발음 나는 대로 전사)와 철자 전사(한글 맞춤법 및 표준어 규정에 따라 전사)를 병행함.
- 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구 단위로 설정함.

· 구성 및 분량

- 일상 대화 총 1,973건(16개 주제)

구분		건수
일상 대화 주제	가족/관혼상제	125
	건강/다이어트	146
	경제/재테크	106
	기타	132
	먹거리	153
	반려동식물	110
	방송/영화/연예인	121
	사회 이슈	100
	생활/주거 환경	108
	쇼핑	117
	여행/휴가	147
	인간관계	109
	취미	139
	취직	120
	패션/미용	107
	회사/학교 생활	133
합계		1,973

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 1,973개, 총 31.4MB

· **파일 명명 규칙**

자리	1	2	3 4	5 6	7 8 9 10 11 12 13 14
속성	매체	장르	주석 단계	구축 연도	일련번호(8자리)
정의 값	S: 구어	D: 사적 대화 (일상 대화)	RW: 원시 말뭉치	23: 2023년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시: SDRW2300000001.json 2023년에 구축한 일상 대화 말뭉치 파일					

· **인용:**

- **(국문)** 국립국어원(2023). 국립국어원 일상 대화 말뭉치 2023(버전 1.0). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language(2023), NIKL Korean Dialogue Corpus (transcription) 2023(v.1.0). URL: https://kli.korean.go.kr/corpus

· **예시**


```
{
    "id": "SDRW2300000001",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2300000001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2023",
        "category": "구어 > 사적대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2300000001.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20230608",
                "topic": "경제/재테크 > 창업",
                "speaker": [
                    {
                        "id": "SD2300009",
                        "age": "30대",
                        "occupation": "무직/취업준비생",
                        "sex": "여성",
```

```
                "birthplace": "전남",
                "principal_residence": "전남",
                "current_residence": "서울",
                "education": "대졸"
            },
            {
                "id": "SD2300011",
                "age": "40대",
                "occupation": "기타",
                "sex": "여성",
                "birthplace": "경북",
                "principal_residence": "경북",
                "current_residence": "서울",
                "education": "대졸"
            }
        ],
        "setting": {
            "relation": "기타",
            "contact_frequency": "0"
        }
    },
    "utterance": [
        {
            "id": "SDRW2300000001.1.1.1",
            "form": "어 창업에 대해서",
            "original_form": "어~ 창업에 대해서",
            "speaker_id": "SD2300009",
            "start": 0.14006,
            "end": 2.04506,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.2",
            "form": "좀 준비해 볼까 하는데",
            "original_form": "좀 준비해 볼까 하는데",
            "speaker_id": "SD2300009",
            "start": 2.40009,
            "end": 4.74532,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.3",
            "form": "혹시 생각해 보신 적이 있거나 조언해 주실 게 있으실까요?",
            "original_form": "혹시 생각해 보신 적이 있거나 조언해 주실 게 있으실까요?",
            "speaker_id": "SD2300009",
            "start": 5.49039,
            "end": 9.77170,
            "note": ""
```

```
},
{
    "id": "SDRW2300000001.1.1.4",
    "form": "어",
    "original_form": "어~",
    "speaker_id": "SD2300011",
    "start": 10.04674,
    "end": 10.57180,
    "note": ""
},
{
    "id": "SDRW2300000001.1.1.5",
    "form": "저는 그냥 제 전 직장이 프랜차이즈",
    "original_form": "저는 그냥 제 전 직장이 프랜차이즈",
    "speaker_id": "SD2300011",
    "start": 10.90162,
    "end": 14.31459,
    "note": ""
},
{
    "id": "SDRW2300000001.1.1.6",
    "form": "관리하는",
    "original_form": "관리하는",
    "speaker_id": "SD2300011",
    "start": 14.80956,
    "end": 15.77951,
    "note": ""
},
{
    "id": "SDRW2300000001.1.1.7",
    "form": "업종이어 가지고",
    "original_form": "업종이어 가지고",
    "speaker_id": "SD2300011",
    "start": 16.08964,
    "end": 17.27516,
    "note": ""
},
{
    "id": "SDRW2300000001.1.1.8",
    "form": "그냥 막연하게",
    "original_form": "그냥 막연하게",
    "speaker_id": "SD2300011",
    "start": 17.61518,
    "end": 18.66024,
    "note": ""
},
{
    "id": "SDRW2300000001.1.1.9",
```

```
        "form": "이렇게 슈퍼바이저를 했었는데",
        "original_form": "이렇게 슈퍼바이저를 했었는데",
        "speaker_id": "SD2300011",
        "start": 19.22509,
        "end": 21.10015,
        "note": ""
    },
    {
        "id": "SDRW2300000001.1.1.10",
        "form": "이제 점주님들 상대하고",
        "original_form": "이제 점주님들 상대하고",
        "speaker_id": "SD2300011",
        "start": 21.57948,
        "end": 23.02450,
        "note": ""
    },
    {
        "id": "SDRW2300000001.1.1.11",
        "form": "이제 오픈 나가서",
        "original_form": "이제 오픈 나가서",
        "speaker_id": "SD2300011",
        "start": 23.02450,
        "end": 24.90981,
        "note": ""
    },
    {
        "id": "SDRW2300000001.1.1.12",
        "form": "지원해 주고 이런 걸 하다 보면",
        "original_form": "지원해 주고 이런 걸 하다 보면",
        "speaker_id": "SD2300011",
        "start": 24.90981,
        "end": 26.94068,
        "note": ""
    },
    {
        "id": "SDRW2300000001.1.1.13",
        "form": "차라리 내가 돈을 모아서",
        "original_form": "차라리 내가 돈을 모아서 {laughing}",
        "speaker_id": "SD2300011",
        "start": 27.72019,
        "end": 29.48409,
        "note": ""
    },
    {
        "id": "SDRW2300000001.1.1.14",
        "form": "하나 차리는 게 낫겠다",
        "original_form": "하나 차리는 게 낫겠다",
        "speaker_id": "SD2300011",
```

```
            "start": 29.79913,
            "end": 31.35919,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.15",
            "form": "약간 이런 생각이 들어서 그냥 막연하게",
            "original_form": "약간 이런 생각이 들어서 그냥 막연하게",
            "speaker_id": "SD2300011",
            "start": 31.71931,
            "end": 34.00437,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.16",
            "form": "조금 그냥 돈을 모아서 프랜차이즈",
            "original_form": "조금 그냥 돈을 모아서 프랜차이즈",
            "speaker_id": "SD2300011",
            "start": 34.14434,
            "end": 36.39886,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.17",
            "form": "하나 차려 볼까 이런",
            "original_form": "하나 차려 볼까 이런",
            "speaker_id": "SD2300011",
            "start": 36.97402,
            "end": 38.33440,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.18",
            "form": "생각을 했었는데",
            "original_form": "생각을 했었는데",
            "speaker_id": "SD2300011",
            "start": 38.76977,
            "end": 39.90458,
            "note": ""
        },
        {
            "id": "SDRW2300000001.1.1.19",
            "form": "인제",
            "original_form": "인제",
            "speaker_id": "SD2300011",
            "start": 40.51003,
            "end": 41.04188,
            "note": ""
```

```
        },
        {
            "id": "SDRW2300000001.1.1.20",
            "form": "관리를 하다 보면은",
            "original_form": "관리를 하다 보면은",
            "speaker_id": "SD2300011",
            "start": 41.59657,
            "end": 42.84088,
            "note": ""
        },
```


※ “form”: 철자 전사
“original_form”: 발음 전사(개인 정보 등은 비식별화)
“speaker_id”: 발화자 아이디
“start”: 발화 시작 시간(초)
“end”: 발화 종료 시간(초)
“note”: 전사자 기타 메모

※ 전사 기호
- 웃음 {laughing}
- 목청 가다듬는 소리 {clearing}
- 노래 {singing}
- 박수 {applauding}
- 잘 들리지 않는 부분 ((추정 전사))
- 들리지 않는 음절 ((xx))
- 전혀 들리지 않는 부분 (())
- 담화 표지 ~
- 불완전 발화 -불완전 발화-

※ 비식별화 기호
- 이름 &name&
- 주민 등록 번호 &social-security-num&
- 카드 번호 &card-num&
- 주소 &address&
- 전화번호 &tel-num&
- 상호명 &company-name&