문화체육관광부
국립국어원

국립국어원 일상 대화 음성 말뭉치 2021

(버전 1.1)

· **자료명:** 국립국어원 일상 대화 음성 말뭉치 2021

· **공개일**
 · (버전 1.0) 2022. 12. 30.
 · (버전 1.1) 2024. 02. 29.
  - 음성-전사 불일치 오류 수정

· **자료 유형:** 음성, 텍스트

· **관련 사업:** 2021년 일상 대화 말뭉치 구축(2021)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2021년 일상 대화 말뭉치 구축' 사업 보고서 참고

 · 내용
  - 15개 주제를 대상으로 자유롭게 대화를 나눈 일상 대화와 8개 주제를 대상으로 찬성 또는 반대 의견을 내고 논의를 통해 결론을 도출하는 협력적 대화의 음성과 전사 자료
  - 각 대화는 최소 두 명, 최대 네 명의 화자로 구성되어 있으며 대화의 평균 시간은 약 15분(총 2,599명 화자, 총 1,000시간 분량).
  - 일상 대화 음성 파일은 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구 단위로 설정된 전사 단위에 따라 분할함. 음성 구간 앞뒤에 최소 200msec의 휴지가 포함되도록 함. 개인 정보는 묵음으로 비식별 처리함.
  - 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 4,143건(15개 주제 일상 대화 3,286건, 8개 주제 협력 대화 857건)

구분		건수	구분		건수
알상 대화 주제	휴가	181	협력적 대화 주제	공공 공간의 CCTV 설치	102
	대중교통	287		가짜 뉴스에 대한 징벌적 손해배상	96
	음악	312		원자력 발전소의 존폐	90
	건강/다이어트	169		지역 내 기피 시설 설치	94
	방송/연예	170		안락사·존엄사 법제화	113
	스포츠/레저	167		AI의 직업 대체	131
	먹거리	178		비대면 생활이 미치는 영향	112
	우정	338		청소년에게 인터넷·스마트폰이 미치는 영향	119
	경제/재테크	139			
	회사/학교	182			
	반려동물	175			
	취직	224			
	가족	163			
	쇼핑	333			
	관혼상제	268			
합계		**3,286**	**합계**		**857**

· 파일 형식:

- 음성: PCM(16kHz 표본화, 16bit 양자화 선형 PCM, Little Endian)
- 텍스트: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 음성 파일 1,416,216개, 텍스트 파일 4,143개, 약 100GB

· **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	21: 2021년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시 - SDRW2100000001.json 2021년에 구축한 일상 대화 말뭉치 파일 - SDRW2100000001.1.1.3.pcm 2021년에 구축한 일상 대화 말뭉치 파일 SDRW2100000001의 세 번째 발화 음성 파일(*음성 파일명: 말뭉치 파일.1.1.발화 번호.pcm)					

· **인용:**

- **(국문)** 국립국어원(2024). 국립국어원 일상 대화 음성 말뭉치 2021(버전 1.1). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language(2024), NIKL Korean Dialogue Corpus (audio) 2021(v.1.1). URL: https://kli.korean.go.kr/corpus

· **예시**


```
{
    "id": "SDRW2100000121",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2100000121",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2021",
        "category": "구어 > 사적대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2100000121.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20210816",
                "topic": "우정 > 친구의 소중함, 감사함",
                "speaker": [
                    {
                        "id": "SD2100121",
                        "age": "50대",
                        "occupation": "주부",
                        "sex": "여성",
```

```
                        "birthplace": "전북",
                        "principal_residence": "전북",
                        "current_residence": "서울",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2100122",
                        "age": "50대",
                        "occupation": "주부",
                        "sex": "여성",
                        "birthplace": "경기",
                        "principal_residence": "경기",
                        "current_residence": "서울",
                        "education": "대졸"
                    }
                ],
                "setting": {
                    "relation": "직장 동료"
                }
            },
            "utterance": [
                {
"id": "SDRW2100000121.1.1.1",
                    "form": "아 지금",
                    "original_form": "아 지금",
                    "speaker_id": "SD2100121",
                    "start": 0.89000,
                    "end": 1.99000,
                    "note": ""
                },
                {
                    "id": "SDRW2100000121.1.1.2",
                    "form": "우리 우정에 대해서 얘기하자 그랬잖아.",
                    "original_form": "우리 우정에 대해서 얘기하자 그랬잖아.",
                    "speaker_id": "SD2100121",
                    "start": 1.99000,
                    "end": 4.93900,
                    "note": ""
                },
                {
                    "id": "SDRW2100000121.1.1.3",
                    "form": "근데",
                    "original_form": "근데",
                    "speaker_id": "SD2100121",
                    "start": 4.93900,
                    "end": 6.09000,
                    "note": ""
                },
                {
                    "id": "SDRW2100000121.1.1.4",
                    "form": "너하고 나하고도 어쨌든 친구로",
                    "original_form": "너하고 나하고도 어쨌든 친구로",
                    "speaker_id": "SD2100121",
                    "start": 6.09000,
```

```
            "end": 9.12648,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.5",
            "form": "만난 지 몇 년 됐는데.",
            "original_form": "만난 지 몇 년 됐는데.",
            "speaker_id": "SD2100121",
            "start": 9.12648,
            "end": 10.99900,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.6",
            "form": "이렇게",
            "original_form": "이렇게",
            "speaker_id": "SD2100121",
            "start": 10.99900,
            "end": 11.96900,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.7",
            "form": "이렇게 좋게",
            "original_form": "이렇게 좋게",
            "speaker_id": "SD2100121",
            "start": 12.47754,
            "end": 14.22861,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.8",
            "form": "그 우정에 대해서 얘기를 나눌 수 있으니까 너무 재밌다.",
            "original_form": "그 우정에 대해서 얘기를 나눌 수 있으니까 너무 재밌다.",
            "speaker_id": "SD2100121",
            "start": 14.22861,
            "end": 18.10900,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.9",
            "form": "으?",
            "original_form": "으?",
            "speaker_id": "SD2100121",
            "start": 18.10900,
            "end": 18.74900,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.10",
            "form": "나도 그렇게 생각해.",
            "original_form": "나도 그렇게 생각해.",
            "speaker_id": "SD2100122",
            "start": 19.75900,
```

```
                "end": 21.37000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.11",
                "form": "왜냐하면 학교 친구도 아니고",
                "original_form": "왜냐하면 학교 친구도 아니고",
                "speaker_id": "SD2100122",
                "start": 21.37000,
                "end": 24.25502,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.12",
                "form": "같은 뭐 동네 친구도 아니고",
                "original_form": "같은 뭐 동네 친구도 아니고",
                "speaker_id": "SD2100122",
                "start": 24.25502,
                "end": 26.88000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.13",
                "form": "그런데 너랑 나랑 참 우연하게 만나 가지고 이렇게 좋은 인연을 맺어 가고 있는 게",
                "original_form": "그런데 너랑 나랑 참 우연하게 만나 가지고 이렇게 좋은 인연을 맺어 가고
있는 게",
                "speaker_id": "SD2100122",
                "start": 26.88000,
                "end": 32.80410,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.14",
                "form": "나는 한편으로 너무 감사하고 고맙고 그래.",
                "original_form": "나는 한편으로 너무 감사하고 고맙고 그래.",
                "speaker_id": "SD2100122",
                "start": 32.80410,
                "end": 36.77000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.15",
                "form": "나도 그런 얘기 자주 하거든.",
                "original_form": "나도 그런 얘기 자주 하거든.",
                "speaker_id": "SD2100121",
                "start": 37.30000,
                "end": 39.70000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.16",
                "form": "정말 어릴 때 친구는 어릴 때 친구 그",
                "original_form": "정말 어릴 때 친구는 어릴 때 친구 그",
                "speaker_id": "SD2100121",
```

```
            "start": 39.70000,
            "end": 43.37166,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.17",
            "form": "깊은 속정이 있다면",
            "original_form": "깊은 속정이 있다면",
            "speaker_id": "SD2100121",
            "start": 43.37166,
            "end": 45.71433,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.18",
            "form": "우리가 어느 정도 나이가 들어서 만나니까",
            "original_form": "우리가 어느 정도 나이가 들어서 만나니까",
            "speaker_id": "SD2100121",
            "start": 45.71433,
            "end": 49.07811,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.19",
            "form": "뭐 부딪칠 일도 없이 서로 이해심이",
            "original_form": "뭐 부딪칠 일도 없이 서로 이해심이",
            "speaker_id": "SD2100121",
            "start": 49.07811,
            "end": 52.66000,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.20",
            "form": "나는 많다고 생각해. 너도 나 나를 참 많이 이해해 준다고 느끼고.",
            "original_form": "나는 많다고 생각해. 너도 -나- 나를 참 많이 이해해 준다고 느끼고.",
            "speaker_id": "SD2100121",
            "start": 52.66000,
            "end": 57.33818,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.21",
            "form": "나도 예전 같았으면",
            "original_form": "나도 예전 같앴으면",
            "speaker_id": "SD2100121",
            "start": 57.33818,
            "end": 60.48000,
            "note": ""
        },
        {
            "id": "SDRW2100000121.1.1.22",
            "form": "욱 했을 수 있는 것도",
            "original_form": "욱 했을 수 있는 것도",
            "speaker_id": "SD2100121",
```

```
                "start": 60.48000,
                "end": 62.80883,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.23",
                "form": "잘 참고.",
                "original_form": "잘 참고.",
                "speaker_id": "SD2100121",
                "start": 62.80883,
                "end": 64.11567,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.24",
                "form": "아 얘는 이랬으 이래서 어쩔 수 없었을 거야 하고 뭔가 살짝 서운했던 것도",
                "original_form": "아 얘는 이랬으 이래서 어쩔 수 없었을 거야 하고 뭔가 살짝 서운했던 것도
",
                "speaker_id": "SD2100121",
                "start": 64.11567,
                "end": 70.26000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.25",
                "form": "그냥 그냥 사르르 녹아지는 그게 우리가 인제 좀 나이 들어간다는 뜻인가 봐.",
                "original_form": "그냥 그냥 사르르 녹아지는 그게 우리가 인제 좀 나이 들어간다는 뜻인가
봐.",
                "speaker_id": "SD2100121",
                "start": 70.26000,
                "end": 75.34000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.26",
                "form": "그러게 말이야.",
                "original_form": "그러게 말이야.",
                "speaker_id": "SD2100122",
                "start": 75.96000,
                "end": 77.27000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.27",
                "form": "참",
                "original_form": "참",
                "speaker_id": "SD2100122",
                "start": 77.27000,
                "end": 78.27000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.28",
                "form": "좋은 관계를 이렇게 이어간다는 것은",
```

```
                "original_form": "좋은 관계를 이렇게 이어간다는 것은",
                "speaker_id": "SD2100122",
                "start": 78.27000,
                "end": 82.36000,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.29",
                "form": "응 자기를 약간 조금 낮추고",
                "original_form": "응 자기를 약간 조금 낮추고",
                "speaker_id": "SD2100122",
                "start": 82.36000,
                "end": 85.04469,
                "note": ""
            },
            {
                "id": "SDRW2100000121.1.1.30",
                "form": "상대방을 좀 높여 주는 거.",
                "original_form": "상대방을 좀 높여 주는 거.",
                "speaker_id": "SD2100122",
                "start": 85.04469,
                "end": 87.00000,
                "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&

· **자료 내용 문의:** 02-2669-9754