문화체육관광부
국립국어원

국립국어원 일상 대화 음성 말뭉치 2020

(버전 1.4)

· **자료명:** 국립국어원 일상 대화 음성 말뭉치 2020

· **공개일**
 · (버전 1.0) 2021. 3. 30.
 · (버전 1.1) 2021. 5. 31.
  - 일부 PCM 파일명 수정
  - 주 성장지 JSON 요소명 수정(pricipal_residence → principal_residence)
 · (버전 1.2) 2021. 12. 1.
  - 음성-전사 불일치 오류 PCM 및 JSON 파일 수정
  - 일부 화자 메타 정보(연령대, 현 거주지, 학력) 수정
 · (버전 1.3) 2022. 12. 30.
  - 일부 화자 메타 정보 오탈자 수정
  - 일부 음성 누락 파일 제외
 · (버전 1.4) 2024. 2. 29.
  - 음성-전사 불일치 오류 수정
  - 일부 대화 메타 정보 수정(녹음 일자, 주제)

· **자료 유형:** 음성, 텍스트

· **관련 사업:** 2020년 일상 대화 말뭉치 구축(2020)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2020년 일상 대화 말뭉치 구축' 사업 보고서 참고

 · 내용
  - 15개 주제, 13개의 제시 자료(국립국어원 신문 말뭉치(버전 1.0)에서 선정한 신문 기사)를 대상으로 두 명의 화자가 자유롭게 대화를 나눈 일상 대화(총 2,739명 화자, 대화당 약 15분 분량, 총 500시간 분량)의 음성과 전사 자료
  - 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 2,231건(15개 주제 대화 1,817건, 13개 제시 자료 대화 414건)

구분		건수	구분		건수
주제	스포츠/레저	115	제시 자료	NWRW1900000010.21633	47
	여행지(국내/해외)	145		NWRW1900000010.21763	36
	계절/날씨	117		NWRW1900000010.25960	22
	회사/학교	134		NWRW1900000010.26464	52
	먹거리	131		NWRW1900000020.18519	45
	방송/연예	110		NWRW1900000020.21711	28
	영화	133		NWRW1900000020.24874	10
	건강/다이어트	108		NWRW1900000040.15396	29
	선물	105		NWRW1900000040.15446	22
	꿈(목표)	108		NWRW1900000040.4066	19
	연애/결혼	131		NWRW1900000040.9893	28
	반려동물	124		NWRW1900000060.15333	45
	아르바이트	133		NWRW1900000060.18354	31
	성격	106			
	가족	117			
합계		**1,817**	**합계**		**414**

· 파일 형식:

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

· 파일 수 및 크기: 음성 파일 870,162개, 텍스트 파일 2,231개, 약 54.1GB

· **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	20: 2020년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시 - SDRW2000000001.json 2020년에 구축한 일상 대화 말뭉치 파일 - SDRW2000000001.1.1.3.pcm 2020년에 구축한 일상 대화 말뭉치 파일 SDRW2000000001의 세 번째 발화 음성 파일(*음성 파일명: 말뭉치 파일.1.1.발화 번호.pcm)					

· **인용:**

- **(국문)** 국립국어원(2024). 국립국어원 일상 대화 음성 말뭉치 2020(버전 1.4). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language(2024), NIKL Korean Dialogue Corpus (audio) 2020(v.1.4). URL: https://kli.korean.go.kr/corpus

· **예시**


```
{
    "id": "SDRW2000000002",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2000000002",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2020",
        "category": "구어 > 사적 대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2000000002.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20200602",
                "topic": "반려동물 > 보험, 유튜버, 동물학대, 작명",
                "speaker": [
                    {
                        "id": "SD2000003",
                        "age": "30대",
                        "occupation": "전문가 및 관련 종사자",
```

```
                    "sex": "여성",
                    "birthplace": "서울",
                    "principal_residence": "서울",
                    "current_residence": "서울",
                    "education": "대졸"
                },
                {
                    "id": "SD2000004",
                    "age": "20대",
                    "occupation": "학생",
                    "sex": "여성",
                    "birthplace": "서울",
                    "principal_residence": "서울",
                    "current_residence": "서울",
                    "education": "대재"
                }
            ],
            "setting": {
                "relation": "기타"
            }
        },
        "utterance": [
            {
                "id": "SDRW2000000002.1.1.1",
                "form": "반려동물을 키우고 계신가요?",
                "original_form": "반려동물을 키우고 계신가요?",
                "speaker_id": "SD2000003",
                "start": 2.78903,
                "end": 4.92608,
                "note": ""
            },
            {
                "id": "SDRW2000000002.1.1.2",
                "form": "혹시 안 키우고 계시다면은",
                "original_form": "혹시 안 키우고 계시다면은",
                "speaker_id": "SD2000003",
                "start": 4.93608,
                "end": 6.70908,
                "note": ""
            },
            {
                "id": "SDRW2000000002.1.1.3",
                "form": "어떤",
                "original_form": "어떤",
                "speaker_id": "SD2000003",
                "start": 6.71906,
                "end": 7.63302,
                "note": ""
            },
            {
                "id": "SDRW2000000002.1.1.4",
                "form": "반려동물을",
                "original_form": "반려동물을",
                "speaker_id": "SD2000003",
```

```
            "start": 7.64304,
            "end": 8.86301,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.5",
            "form": "한번 키워보고 싶은 생각이 있으신지",
            "original_form": "한번 키워보고 싶은 생각이 있으신지",
            "speaker_id": "SD2000003",
            "start": 8.87302,
            "end": 12.75805,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.6",
            "form": "저는 반려동물을",
            "original_form": "저는 반려동물을",
            "speaker_id": "SD2000004",
            "start": 12.76804,
            "end": 16.48801,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.7",
            "form": "키우지 않고 있는데요.",
            "original_form": "키우지 않고 있는데요.",
            "speaker_id": "SD2000004",
            "start": 16.49803,
            "end": 18.63202,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.8",
            "form": "어~ 키우지 않고 있는 이유는",
            "original_form": "어~ 키우지 않고 있는 이유는",
            "speaker_id": "SD2000004",
            "start": 18.64208,
            "end": 21.67103,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.9",
            "form": "반려동물을 끝까지 책임질 수 있다는",
            "original_form": "반려동물을 끝까지 책임질 수 있다는",
            "speaker_id": "SD2000004",
            "start": 21.68103,
            "end": 26.22707,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.10",
            "form": "생각이 아직은 없어서",
            "original_form": "생각이 아직은 없어서",
            "speaker_id": "SD2000004",
```

```
            "start": 26.23708,
            "end": 28.19308,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.11",
            "form": "키우지 않고 있지만",
            "original_form": "키우지 않고 있지만",
            "speaker_id": "SD2000004",
            "start": 28.20305,
            "end": 30.52503,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.12",
            "form": "어~ 어느 정도 때가 된다면",
            "original_form": "어~ 어느 정도 때가 된다면",
            "speaker_id": "SD2000004",
            "start": 30.53502,
            "end": 32.30907,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.13",
            "form": "언젠가는 반려동물을 한번 꼭",
            "original_form": "언젠가는 반려동물을 한번 꼭",
            "speaker_id": "SD2000004",
            "start": 32.31904,
            "end": 34.59703,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.14",
            "form": "키워보고 싶고",
            "original_form": "키워보고 싶고",
            "speaker_id": "SD2000004",
            "start": 34.60707,
            "end": 36.33303,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.15",
            "form": "저는 그중에서도",
            "original_form": "저는 그중에서도",
            "speaker_id": "SD2000004",
            "start": 36.34306,
            "end": 38.21901,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.16",
            "form": "고양이를",
            "original_form": "고양이를",
            "speaker_id": "SD2000004",
```

```
            "start": 38.22901,
            "end": 40.02802,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.17",
            "form": "좋아해서",
            "original_form": "좋아해서",
            "speaker_id": "SD2000004",
            "start": 40.03807,
            "end": 40.99002,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.18",
            "form": "고양이를 키워보고 싶습니다.",
            "original_form": "고양이를 키워보고 싶습니다.",
            "speaker_id": "SD2000004",
            "start": 41.00004,
            "end": 43.28802,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.19",
            "form": "혹시 반려동물을",
            "original_form": "혹시 반려동물을",
            "speaker_id": "SD2000004",
            "start": 43.29804,
            "end": 45.57902,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.20",
            "form": "키우신 경험이 있으신지 아니면",
            "original_form": "키우신 경험이 있으신지 아니면",
            "speaker_id": "SD2000004",
            "start": 45.58905,
            "end": 49.09604,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.21",
            "form": "키우시고 계시다면",
            "original_form": "키우시고 계시다면",
            "speaker_id": "SD2000004",
            "start": 49.10608,
            "end": 50.39008,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.22",
            "form": "왜 키우시게 되셨는지",
            "original_form": "왜 키우시게 되셨는지",
            "speaker_id": "SD2000004",
```

```
            "start": 50.40002,
            "end": 52.35302,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.23",
            "form": "저는 반려동물을",
            "original_form": "저는 반려동물을",
            "speaker_id": "SD2000003",
            "start": 53.97708,
            "end": 57.88704,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.24",
            "form": "반려 요즘에 말하는",
            "original_form": "반려 요즘에 말하는",
            "speaker_id": "SD2000003",
            "start": 57.89703,
            "end": 59.69806,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.25",
            "form": "반려동물이라는 거를 꽤 많이 키웠었는데",
            "original_form": "반려동물이라는 거를 꽤 많이 키웠었는데",
            "speaker_id": "SD2000003",
            "start": 59.70807,
            "end": 62.80601,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.26",
            "form": "어~ 예전에 초등학교 때",
            "original_form": "어~ 예전에 초등학교 때",
            "speaker_id": "SD2000003",
            "start": 62.81602,
            "end": 65.64801,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.27",
            "form": "금붕어라든지",
            "original_form": "금붕어라던지",
            "speaker_id": "SD2000003",
            "start": 65.65802,
            "end": 67.94407,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.28",
            "form": "이런 것도 키워봤고",
            "original_form": "이런 것도 키워봤고",
            "speaker_id": "SD2000003",
```

```
            "start": 67.95403,
            "end": 69.24006,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.29",
            "form": "햄스터도 키워 본 경험이 있고요.",
            "original_form": "햄스터도 키워 본 경험이 있고요.",
            "speaker_id": "SD2000003",
            "start": 69.25003,
            "end": 71.85608,
            "note": ""
        },
        {
            "id": "SDRW2000000002.1.1.30",
            "form": "어~ 강아지도 키워 본 경험이 있어요.",
            "original_form": "어~ 강아지도 키워 본 경험이 있어요.",
            "speaker_id": "SD2000003",
            "start": 71.86607,
            "end": 74.61403,
            "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