

국립국어원 일상 대화 음성 말뭉치 2022 (버전 1.0)

- **자료명:** 국립국어원 일상 대화 음성 말뭉치 2022
- **공개일**
 - (버전 1.0) 2023. 12. 29.
- **자료 유형:** 음성, 텍스트
- **관련 사업:** 2022년 일상 대화 말뭉치 구축(2022)
- **자료 설명**
 - ※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > ‘2022년 일상 대화 말뭉치 구축’ 사업 보고서 참고
 - **내용**
 - 16개 주제를 대상으로 자유롭게 대화를 나누는 일상 대화(비통제 대화 포함)와 10개 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 2,000명 화자, 총 630시간 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 2,654건(16개 주제 일상 대화 2,184건(비통제 대화 417건), 10개 주제 협력 대화 470건)

구분		건수 (비통제 대화 건수)	구분		건수			
일상 대화 주제	휴가	137(26)	협력적 대화 주제	문화	영화/드라마/ 음악(콘텐츠)	46		
	대중교통	149(27)			연극/뮤지컬/ 콘서트(공연)	45		
	음악	128(25)				전시회/박물관 (전시)	42	
	건강/다이어트	142(26)					책/독서	49
	방송/연예	147(26)					스포츠/레저	48
	스포츠/레저/취미	133(25)				패션/뷰티	52	
	먹거리	144(31)					음식/음료	49
	우정	130(26)					반려동물	47
	경제/재테크	138(24)				관광	여행 계획	44
	회사/학교	136(25)					여행 일반	48
	반려동물	132(28)						
	취직	144(24)						
	가족/관혼상제	133(25)						
	쇼핑	133(27)						
	생활/주거환경	132(28)						
	기타	126(24)						
합계		2,184(417)	합계		470			

· 파일 형식:

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

· 파일 수 및 크기: 음성 파일 880,811개, 텍스트 파일 2,654개, 약 62.1GB

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	22: 2022년	00000001 ~ 99999999 (여덟 자리 일련번호)									
※ 예시 - SDRW2200000001.json 2022년에 구축한 일상 대화 말뭉치 파일 - SDRW22100000001.1.1.3.pcm 2022년에 구축한 일상 대화 말뭉치 파일 SDRW2200000001의 세 번째 발화 음성 파일(*음성 파일명: 말뭉치 파일.1.1.발화 번호.pcm)														

· 인용:

- (국문) 국립국어원(2023). 국립국어원 일상 대화 말뭉치 2022(버전 1.0).
URL: <https://kli.korean.go.kr/corpus>
- (영문) National Institute of Korean Language(2023), NIKL Korean Dialogue Corpus (transcription) 2022(v.1.0). URL: <https://kli.korean.go.kr/corpus>

· 예시

```
{
  "id": "SDRW2200000001",
  "metadata": {
    "title": "국립국어원 구어 말뭉치 SDRW2200000001",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2022",
    "category": "구어 > 사적대화 > 일상대화",
    "annotation_level": [
      "원시"
    ],
    "sampling": "본문 전체"
  },
  "document": [
    {
      "id": "SDRW2200000001.1",
      "metadata": {
        "title": "3인 일상 대화",
        "author": "개인 발화자",
        "publisher": "개인 발화 녹음",
        "date": "20220824",
        "topic": "회사/학교 > 학교 생활 및 전공 이야기",
        "environment": "",
        "speaker": [
          {
            "id": "SD22000005",
            "age": "20대",

```



```
        "occupation": "학생",
        "sex": "남성",
        "birthplace": "서울",
        "principal_residence": "서울",
        "current_residence": "서울",
        "education": "대재"
    },
    {
        "id": "SD2200006",
        "age": "20대",
        "occupation": "무직/취업준비생",
        "sex": "여성",
        "birthplace": "서울",
        "principal_residence": "경기",
        "current_residence": "경기",
        "education": "대졸"
    },
    {
        "id": "SD2200007",
        "age": "20대",
        "occupation": "학생",
        "sex": "남성",
        "birthplace": "서울",
        "principal_residence": "경기",
        "current_residence": "경기",
        "education": "대재"
    }
],
"setting": {
    "relation": "친구",
    "device": "",
    "mic": ""
}
},
"utterance": [
    {
        "id": "SDRW2200000001.1.1.1",
        "form": "어 여기서 학교 얘기가 나와서",
        "original_form": "어 여기서 학교 얘기가 나와서",
        "speaker_id": "SD2200007",
        "start": 0.14000,
        "end": 3.48850,
        "note": ""
    },
    {
        "id": "SDRW2200000001.1.1.2",
        "form": "아마 여기서 학력이 제일",
        "original_form": "아마 여기서 학력이 제일",
        "speaker_id": "SD2200007",
        "start": 4.11000,
        "end": 6.22450,
        "note": ""
    }
]
```

```

    "id": "SDRW2200000001.1.1.3",
    "form": "높은 사람이 말하는 게 낮 낮지 않을까?",
    "original_form": "높은 사람이 말하는 게 -낮- 낮지 않을까?",
    "speaker_id": "SD2200007",
    "start": 6.29000,
    "end": 8.43250,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.4",
    "form": "너랑 아 name1랑 나는 아직",
    "original_form": "너랑 아 &name1&랑 나는 아직",
    "speaker_id": "SD2200007",
    "start": 9.00000,
    "end": 11.92050,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.5",
    "form": "학교를 졸업을 안했으니까",
    "original_form": "학교를 졸업을 안했으니까",
    "speaker_id": "SD2200007",
    "start": 12.11000,
    "end": 13.60641,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.6",
    "form": "그래서 대학을 졸업한 먼저 배우신 분한테",
    "original_form": "그래서 대학을 졸업한 먼저 배우신 분한테",
    "speaker_id": "SD2200007",
    "start": 14.17000,
    "end": 18.10085,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.7",
    "form": "말을 넘기는 게 나올 것 같아.",
    "original_form": "말을 넘기는 게 나올 것 같애.",
    "speaker_id": "SD2200007",
    "start": 18.76000,
    "end": 20.79000,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.8",
    "form": "그래.",
    "original_form": "그래.",
    "speaker_id": "SD2200006",
    "start": 21.61000,
    "end": 22.14450,
    "note": ""
  },
  {

```

```

    "id": "SDRW2200000001.1.1.9",
    "form": "다들 전공이 어떻게 돼?",
    "original_form": "다들 전공이 어떻게 돼?",
    "speaker_id": "SD2200006",
    "start": 23.64000,
    "end": 25.13650,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.10",
    "form": "나는 경영학을 전공을 하고 있어.",
    "original_form": "나는 경영학을 전공을 하고 있어.",
    "speaker_id": "SD2200005",
    "start": 27.15000,
    "end": 29.96850,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.11",
    "form": "name2는 어떻게 되니?",
    "original_form": "&name2&는 어떻게 되니?",
    "speaker_id": "SD2200005",
    "start": 30.96000,
    "end": 32.08050,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.12",
    "form": "나도 너랑 같은 경영학을 전공하고 있어.",
    "original_form": "나도 너랑 같은 경영학을 전공하고 있어. {laughing}",
    "speaker_id": "SD2200007",
    "start": 32.75000,
    "end": 36.14450,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.13",
    "form": "맞다 우리 18 학번 동기였지.",
    "original_form": "맞다 우리 일 팔 학번 동기였지.",
    "speaker_id": "SD2200005",
    "start": 36.49000,
    "end": 38.33650,
    "note": "발화겹침"
  },
  {
    "id": "SDRW2200000001.1.1.14",
    "form": "응 너무 오랜만에 봤다.",
    "original_form": "응 너무 오랜만에 봤다. {laughing}",
    "speaker_id": "SD2200007",
    "start": 38.07000,
    "end": 40.44850,
    "note": "발화겹침"
  },
  {

```

```

    "id": "SDRW2200000001.1.1.15",
    "form": "그러게 우리 얼마만에 봤지?",
    "original_form": "그러게 우리 얼마만에 봤지?",
    "speaker_id": "SD2200005",
    "start": 40.53350,
    "end": 42.00050,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.16",
    "form": "우리 한 3개월 만에 본 거 같은데도 까먹은 거야?",
    "original_form": "우리 한 삼 개월 만에 본 거 같은데도 까먹은 거야?",
    "speaker_id": "SD2200007",
    "start": 42.59750,
    "end": 45.55250,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.17",
    "form": "미안하다. 나도 최근에 너무 바빠서.",
    "original_form": "미안하다. 나도 최근에 너무 바빠서.",
    "speaker_id": "SD2200005",
    "start": 46.61350,
    "end": 49.63250,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.18",
    "form": "응 알겠어. name3는 뭘 전공했어?",
    "original_form": "응 알겠어. &name3&는 뭘 전공했어?",
    "speaker_id": "SD2200007",
    "start": 49.87750,
    "end": 52.97650,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.19",
    "form": "나는 법학을 전공했고 복수 전공으로 사회복지학을 전공했어.",
    "original_form": "나는 법학을 전공했고 복수 전공으로 사회복지학을 전공했어.",
    "speaker_id": "SD2200006",
    "start": 53.41350,
    "end": 57.77650,
    "note": ""
  },
  {
    "id": "SDRW2200000001.1.1.20",
    "form": "복수 전공으로 사회 사회복지학?",
    "original_form": "복수 전공으로 사회 사회복지학?",
    "speaker_id": "SD2200007",
    "start": 58.46950,
    "end": 60.52850,
    "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