문화체육관광부
국립국어원

NIKL Dialogue Corpus (transcription) 2023

(version 1.0)

- **Title**: NIKL Dialogue Corpus (transcription) 2023

- **Release Date**
  - (version 1.0) 26 December 2024

- **Data Type**: text

- **Description**
  - Content
    - A corpus of natural dialogues in which speakers freely communicate about 16 topics.
    - Each conversation consists of between two and four speakers, with an average conversation time of about 15 minutes (2,168 speakers in total, 500 hours in total).
    - Daily dialogues are transcribed in Korean, and phonetic transcription (transcription as pronounced in a format that deviates from standard pronunciation or has multiple standard pronunciations) and orthographic transcription (transcription according to the Korean orthographic rules and standard language regulations) are combined.
    - The transcription unit is set as an intonational phrase unit segmented by long pauses, boundary intonations, or utterance-final lengthening.

  - Composition and size
    - 1,973 dialogues (16 topics)

Category		Count
일상 대화 주제	Family/Four ceremonial occasions (coming of age, wedding, funeral, ancestral rites)	125
	Health/Diet	146
	Economy/Financial technology	106
	Etc	132
	Food	153

	Companion plant and animal	110
	Broadcasting/Movie/Entertainer	121
	Social Issues	100
	Living/Residential environment	108
	Shopping	117
	Travel/Vacation	147
	Relationship	109
	Hobby	139
	Employment	120
	Fashion/Beauty	107
	Life at work and school	133
합계		**1,973**

- File format and Encoding: JSON(UTF-8)
- Number and size of files: 1,973 files, 31.4MB

- File Naming Rule

order	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Feature	Data source	Building method	Annotation level		Building year		Serial number (8 digits)							
Value	S: Spoken language	D: Private conversation (Daily dialogue)	RW: raw corpus		23: 2023		00000001 ~ 99999999 (8 digits serial number)							
※ Example: SDRW2300000001.json - a file of the Dialogue Corpus, built in 2023														

- Citation(s):

(Korean) 국립국어원(2024). 국립국어원 일상 대화 말뭉치(전사) 2023(버전 1.0). URL: https://kli.korean.go.kr/corpus

(English) National Institute of Korean Language (2024). NIKL Dialogue Corpus (transcription) 2023 (v.1.0). URL: https://kli.korean.go.kr/corpus

· Examples

```
{
    "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”: orthographic transcription

“original_form”: phonetic transcription (personal information is de-identified)

“speaker_id”: speaker id

“start”: start time of the utterance (in seconds)

“end”: end time of the utterance (in seconds)

“note”: transcriber’s notes

※ Transcription symbols

- laugh {laughing}
- clearing one’s throats {clearing}
- singing {singing}
- applauding {applauding}
- indistinct voice ((추정 전사))
- inaudible syllable ((xx))
- inaudible at all (())
- discourse markers ~
- incomplete utterance -불완전 발화-

※ de-identification symbols

- name: &name&
- social security number: &social-security-num&
- card number: &card-num&
- address: &address&
- telephone number: &tel-num&
- company name: &company-name&