문화체육관광부
국립국어원

국립국어원 구어 말뭉치

(버전 1.2)

· **자료명:** 국립국어원 구어 말뭉치

· **공개일**
  · (버전 1.0) 2020. 8. 25.
  · (버전 1.1) 2021. 3. 30.
    - 구어 말뭉치(버전 1.0) 포함 사적 대화(2,224건) 제외
    - 파일 내 문서 아이디 오류 수정
  · (버전 1.2) 2021. 12. 1.
    - 주 성장지 JSON 요소명 수정(pricipal_residence → principal_residence)

· **자료 유형:** 텍스트

· **관련 사업:** 2018년 국어 말뭉치 연구 및 구축(2018), 구어 자료 수집 및 원시 말뭉치 구축(2019)

· **자료 설명**
  ※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > 연구 보고서 > '2018년 국어 말뭉치 연구 및 구축', '구어 자료 수집 및 원시 말뭉치 구축' 사업 보고서 참고
  · 내용
    - 방송, 강연 등의 공적 구어 자료, 드라마 대본 등의 준구어 자료로 구성한 구어 말뭉치
    - 구어 자료는 한글로 전사하였으며, 실제 발음이 표준 발음과 다른 경우에 발음대로 표기함.

  · 분량
    - 공적 독백 2,490건
    - 공적 대화 19,104건
    - 준구어-대본 4,102건(드라마 4,102회 분량)

  · 파일 형식: JSON(UTF-8 인코딩)

  · 파일 수 및 크기: 파일 25,696개, 총 6.73GB

· **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)
정의값	S: 구어	A: 공적 독백 B: 공적 대화 E: 준구어-대본	RW: 원시 말뭉치	18: 2018년 19: 2019년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시: SBRW1900000001.json 2019년에 구축한 구어 공적 대화 원시 말뭉치 파일					

· **인용:**

(국문) 국립국어원(2021). 국립국어원 구어 말뭉치(버전 1.2). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2021). NIKL Spoken Corpus (v.1.2). URL: https://kli.korean.go.kr/corpus

· **예시**

· 구어


```
{
    "id": "SARW1800000002",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SARW1800000002",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2018",
        "category": "구어 > 공적 독백 > 뉴스",
        "annotation_level": [
            "원시"
        ],
        "sampling": "목록 선정 후 구어 자료 수집"
    },
    "document": [
        {
            "id": "SARW1800000002.1",
            "metadata": {
                "title": "EBS 정오뉴스 2018년 1월",
                "author": "박민영 외",
                "publisher": "EBS",
                "date": "20180000",
                "topic": "도서관의 변신, 메이커 스페이스에 대한 기사",
                "speaker": [
                    {
                        "id": "P1",
                        "age": "NA",
                        "occupation": "아나운서 ",
                        "sex": "여성", "birthplace": "NA",
                        "principal_residence": "NA",
                        "current_residence": "NA"
                    },
```

```
                {
                    "id": "P2",
                    "age": "NA",
                     "occupation": "리포터",
                     "sex": "여성", "birthplace": "NA",
                     "principal_residence": "NA",
                     "current_residence": "NA"
                }
            ]
        },
        "utterance": [
            {
                "id": "SARW1800000002.1.1.1",
                "form": "미국의 공공도서관들이 새로운 모습으로 변신하고 있습니다.",
                "original_form": "미국의 공공도서관들이 새로운 모습으로 변신하고 있습니다.",
                "speaker_id": "P1",
                "note": ""
            },
            {
                "id": "SARW1800000002.1.1.2",
                "form": "누구나 책을 접할 수 있는 공간에서,",
                "original_form": "누구나 책을 접할 수 있는 공간에서,",
                "speaker_id": "P1",
                "note": ""
            },
            {
                "id": "SARW1800000002.1.1.3",
                "form": "누구나 무엇이든",
                "original_form": "누구나 무엇이든",
                "speaker_id": "P1",
                "note": ""
            },
            {
                "id": "SARW1800000002.1.1.4",
                "form": "만들 수 있는 공간으로",
                "original_form": "만들 수 있는 공간으로",
                "speaker_id": "P1",
                "note": ""
            },
            {
                "id": "SARW1800000002.1.1.5",
                "form": "탈바꿈하고 있는 건데요",
                "original_form": "탈바꿈하고 있는 건데요",
                "speaker_id": "P1",
                "note": ""
            },
            {
                "id": "SARW1800000002.1.1.6",
                "form": "개인이 쉽게 접하기 힘든",
                "original_form": "개인이 쉽게 접하기 힘든",
                "speaker_id": "P1",
                "note": ""
            },
            {
                "id": "SARW1800000002.1.1.7",
                "form": "쓰리디프린터 같은 고가의 장비부터 재봉틀",
```

```
            "original_form": "쓰리디프린터 같은 고가의 장비부터 재봉틀",
            "speaker_id": "P1",
            "note": ""
        },
        {
            "id": "SARW1800000002.1.1.8",
            "form": "손 공구까지",
            "original_form": "손 공구까지",
            "speaker_id": "P1",
            "note": ""
        },
        {
            "id": "SARW1800000002.1.1.9",
            "form": "다양한 제작 도구를 갖추고",
            "original_form": "다양한 제작 도구를 갖추고",
            "speaker_id": "P1",
            "note": ""
        },
        {
            "id": "SARW1800000002.1.1.10",
            "form": "시민들을 기다리는 도서관",
            "original_form": "시민들을 기다리는 도서관",
            "speaker_id": "P1",
            "note": ""
        },
        {
            "id": "SARW1800000002.1.1.11",
            "form": "뉴스지에서 전해드립니다.",
            "original_form": "뉴스지에서 전해드립니다.",
            "speaker_id": "P1",
            "note": ""
        },
```


※ “original_form”: 수집한 구어 자료를 한글로 전사한 형태(개인 정보 등은 비식별화)
“form”: 원문에서 연속된 여러 개의 공백(스페이스, 탭 등), 전사 기호, 비식별화 기호 등을 제거하여 전처리한 형태

※ 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&

· 준구어


```
{
    "id": "SERW1900001001",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SERW1900001001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2019",
        "category": "구어 > 준구어-대본",
        "annotation_level": [
            "원시“
        ],
        "sampling": "목록 선정 후 대본 수집"
    },
     "document": [
        {
            "id": "SERW1900001001.1",
            "metadata": {
                "title": "9회말 2아웃 1회 ",
                "author": "여지나",
                "publisher": "MBC",
                "date": "20070000"
            },
            "utterance": [
                {
                    "id": "SERW1900001001.1.1.1",
                    "form": "",
                    "original_form": "",
                    "speaker_id": "",
                    "note": "시상식장"
                },
                {
                    "id": "SERW1900001001.1.1.2",
                    "form": "",
                    "original_form": "",
                    "speaker_id": "",
                    "note": "검은 화면"
                },
                {
                    "id": "SERW1900001001.1.1.3",
                    "form": "후~ 아~ 몹시 긴장이 되네요. 정말 감사합니다. 제 인생에 이런 말을 할 날이 과연 있을까 했는데.",
                    "original_form": "(E) 후~~ (떨리는 심호흡) 아~~ 몹시 긴장이 되네요. 정말 감사합니다. 제 인생에 이런 말을 할 날이 과연 있을까 했는데...",
                    "speaker_id": "난희",
                    "note": ""
                }
```


※ “original_form”: 수집한 준구어 자료의 원문을 그대로 유지한 형태
“form”: 원문에서 연속된 여러 개의 공백(스페이스, 탭 등), 기호, 대사 이외의 지문 등을 제거하여 전처리한 형태

· **자료 내용 문의:** 02-2669-9638