

## 국립국어원 문어 말뭉치 (버전 1.2)

· **자료명:** 국립국어원 문어 말뭉치

· **공개일**

- (버전 1.0) 2020. 8. 25.
- (버전 1.1) 2022. 9. 7.
- (버전 1.2) 2022. 10. 31.

· **자료 유형:** 텍스트

· **관련 사업:** 2018년 국어 말뭉치 연구 및 구축(2018)  
문어 말뭉치 원문 자료 수집(2019)

· **자료 설명**

· **내용**

- 책, 잡지, 보고서 등 저작권 문제가 해결된 저작물 10,045종의 문어 원시 말뭉치  
\*저작권 문제 미해결 자료는 추후 배포 예정

· **분량**

- 총 10,045건

· **장르별 저작물 수 및 비율**

장르	저작물	
	종 수	비율(%)
책-상상	1,393	13.87
책-정보	8,651	86.12
잡지	1	0.01
합계	10,045	100

· ‘책-정보’ 장르의 주제별 저작물 수 및 비율

장르	저작물		주제*	저작물	
	총 수	비율(%)		총 수	비율(%)
총류	415	4.80	기술과학	776	8.97
철학	938	10.84	예술	333	3.85
종교	533	6.16	언어	69	0.80
사회과학	3,399	39.29	문학	1,241	14.35
자연과학	238	2.75	역사	709	8.20

\* 주제 분류는 국립중앙도서관 도서 정보의 한국십진분류법(제6판) 백의 자리 수에 대응하는 기준으로 분류함.

- 파일 형식: JSON(UTF-8 인코딩)
- 파일 수 및 크기: 파일 10,045개, 총 4.24GB

· 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	W (책, 잡지, 보고서 등 신문 외 문어)	A (책-상상) B (책-정보)	RW (원시 말뭉치)		18 (2018년)	19 (2019년)	00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: WARW1900000001.json 2019년에 구축한 책-상상 장르의 원시 말뭉치 1번째 파일														

· 인용:

(국문) 국립국어원(2020). 국립국어원 문어 말뭉치(버전 1.2). URL: <https://kli.korean.go.kr/corpus>

(영문) National Institute of Korean Language (2020). NIKL Written Corpus(v.1.2). URL: <https://kli.korean.go.kr/corpus>

· 예시

```
{
  "id": "WARW1800000007",
  "metadata": {
    "title": "국립국어원 문어 말뭉치 WARW1800000007",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2018",
    "category": "문어 > 책-상상 > 문학",
    "annotation_level": [
      "원시"
    ],
    "sampling": "본문 전체"
  },
  "document": [
    {
      "id": "WARW1800000007.1",
      "metadata": {
        "title": "이야기꾼 구연설화",

```

```

    "author": "황인덕",
    "publisher": "박이정",
    "date": "20070000"
  },
  "paragraph": [
    {
      "id": "WARW1800000007.1.1",
      "form": "01범보다 무서운 꽃감"
    },
    {
      "id": "WARW1800000007.1.2",
      "form": "화자를 처음 만나 이야기를 들으러 왔다고 하자 서슴없이 꺼낸 첫 이야기이다. 화자로서 가장 쉽게 기억해낸 이야기인 셈이다. 설화 앞뒤에 교훈적 해석을 덧붙이고 있음은 화자의 습관화된 태도의 한 모습이기도 하다. 어려서 조모로부터 들었다고 했다."
    },
    {
      "id": "WARW1800000007.1.3",
      "form": "그랑개. 사람이 어거지루는 못 살구. 응? 어거지루 안 되능 거여. 사람이 그랑개 뭐이냐 하면 자연~간 제절루 되야지 어거지루는 못 살어, 사람이."
    },
    {
      "id": "WARW1800000007.1.4",
      "form": "그래 옛날, 그 꼭감이라능 게 말여. 사람이 먹잡야 이케? 먹지마는. 그게 참 무성(무서운) 거여."
    },
    {
      "id": "WARW1800000007.1.5",
      "form": "애기가 울어. 옛날에. 그래 할머니가 달걀(달래). 그때 호랭이가, 응? 그 집 문얍이 와 섰어 지금. 옛날이는 호랭이가 말두 하구 그랴. 그래 인제 그 할머니가 애기를 달램성 윈갓 소리를 다 하. '호랭이 왔다'구 해두 울구우, '괘이 왔다'두 울구, 윈갓 소리를 다 해두 울어."
    },
    {
      "id": "WARW1800000007.1.6",
      "form": "그랑개 꼭감을."
    }
  ]

```

· 자료 내용 문의: 02-2669-9638