

국립국어원 신문 말뭉치 2023 (버전 1.0)

- **자료명:** 국립국어원 신문 말뭉치 2023
- **공개일**
 - (버전 1.0) 2024. 2. 29.
- **자료 유형:** 텍스트
- **관련 사업:** 2023년 신문 기사 원문 자료 수집 및 정제
- **자료 설명**
 - **내용 및 분량:** 2022년 생산된 신문 기사 1,023,431건
 - 포함 신문 매체 (총 28개)

매체 종류	매체 이름
전국 종합 일간 (8개)	경향신문, 국민일보, 내일신문, 서울신문, 세계일보, 조선일보, 한겨레, 한국일보
지역 종합 일간 (10개)	강원일보, 경기일보, 경북일보, 남도일보, 대구신문, 부산일보, 전북도민일보, 중도일보, 중부일보, 충청일보
경제 일간 (7개)	머니투데이, 서울경제, 아시아경제, 아주경제, 이데일리, 이투데이, 헤럴드경제
스포츠 일간 (1개)	스포츠서울
인터넷 신문 (2개)	노컷뉴스, 뉴스핌

- 주제(TOPIC)별 기사 건수

사회	경제	생활	정치	IT/과학	미용/건강	스포츠	문화	연예
274,416	200,739	156,393	154,183	74,660	51,050	54,662	36,208	21,120

- **파일 형식:** JSON(UTF-8 인코딩)
- **파일 수 및 크기:** 28개(zip 압축파일 1개, 969MB)
- **인용:** 국립국어원(2023). 2023년도 구축 국립국어원 신문 말뭉치(버전 1.0).

URL: <https://corpus.korean.go.kr>

· 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	N: 신문	W: 전국 종합지 L: 지역 종합지 P: 전문지 I: 인터넷 기반 신문 Z: 기타	RW: 원시 말뭉치		23: 2023년		00000001 ~ 99999999 (여덟 자리 일련번호)							

※ 예시: NWRW2300000001.json 2023년도에 구축한 '신문 전국 종합지 매체의 기사 원시 말뭉치' 첫 번째 파일

· 예시

```
{
  "id": "NWRW2300000005.5",
  "metadata": {
    "title": "세계일보 2022년 기사",
    "author": "김신성",
    "publisher": "세계일보",
    "date": "20220101",
    "topic": "문화",
    "original_topic": "문화>출판, 문화>학술_문화재, IT_과학>모바일"
  },
  "paragraph": [
    {
      "id": "NWRW2300000005.5.1",
      "form": "한국어 동사 5472개 뜻풀이 사전 나왔다"
    },
    {
      "id": "NWRW2300000005.5.2",
      "form": "한국어 동사 5472개의 언어 정보를 상세히 기술한 사전이 출간됐다."
    },
    {
      "id": "NWRW2300000005.5.3",
      "form": "베가북스가 펴낸 '알뜰꼼꼼 뜻풀이 한국어 동사 사전'이다. 국립국어원장을 지낸 송철의·권재일 서울대 명예교수와 김현권 한국방송통신대 명예교수, 박만규 아주대 교수 등 언어학자 4명이 6년 넘게 편찬 작업한 끝에 내놓은 성과물이다."
    },
    {
      "id": "NWRW2300000005.5.4",
      "form": "저자들은 "문장은 동사가 꾸미는 틀로 이뤄졌다는 점에서 동사는 우리 생각의 틀"이라며 "동사가 다른 품사에 비해 특권적 지위를 지니고 있으므로 그에 대해 더 많은 지식을 얻기 위해 노력하는 것은 당연하다"고 발간 취지를 밝혔다."
    },
    {
      "id": "NWRW2300000005.5.5",
      "form": "표준어가 아니더라도 사람들이 많이 사용하는 동사는 표제어에 포함했고, 표제어별로 어원·형태·의미·속어·용례 등 다양한 정보를 담았다. 유의어, 반의어, 상위어, 하위어도 제시했다. 1358쪽 분량이다."
    }
  ]
}
```

```
"id": "NWRW2300000005.5.6",  
  "form": "베가박스 관계자는 “동사 사전이 있는 언어가 영어, 독일어, 불어, 일본어 등 10개도 되지 않는다”며 “우리말의 뼈대인 동사의 완벽한 뜻풀이와 함께 응용 사례를 기록한 한국 최초의 본격 동사 사전”이라고 설명했다."  
    }  
  ]  
},
```

※ “original_topic”: 신문 매체의 자체 분류 주제

“topic”: 매체 통합 분류 주제(정치, 경제, 사회, 생활, IT/과학, 연예, 스포츠 문화, 미용/건강)

· 자료 내용 문의: 02-2669-9636