

국립국어원 무형 대용어 복원 말뭉치 (버전 1.0)

- **자료명:** 국립국어원 무형 대용어 복원 말뭉치 2020
- **공개일**
 - (버전 1.0) 2021. 3. 30.
- **자료 유형:** 텍스트
- **관련 사업:** 주격 무형 대용어 복원 말뭉치 구축(2019), 구문 및 무형 대용어 복원 말뭉치 연구 분석(2020)
- **자료 설명**
 - ※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > 2019년 ‘주격 무형 대용어 복원 말뭉치 구축’과 2020년 ‘구문 및 무형 대용어 복원 말뭉치 연구 분석’ 사업 보고서 참고
 - **내용**
 - 문장 내 생략된 주어와 목적어를 맥락에 따라 복원한 말뭉치임.
 - 2018년 구축 국립국어원 신문 말뭉치와 구어 말뭉치에서 추출한 원시 말뭉치를 분석 대상으로 함.
 - 한국전자통신연구원 <생략어복원 태깅 가이드라인 v3.3>과 한국정보통신기술협회 표준 <의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법> 지침을 기반으로 분석하였음.
- **분석 정보**
 - 서술어(predicate): 의미적으로 필수적인 주어나 목적어가 생략된 동사 혹은 형용사
 - 보조 용언은 복원 대상 서술어에서 제외함.
 - 의사 보조 용언 구성은 복원 대상 서술어에서 제외함.
‘-ㄹ 수/리(가) 있다/없다, -ㄴ/ㄹ 것/터/뿐/따름/모양/지경/참/중/노릇/예정/길이다, -ㄹ 만/법/듯하다, -는 말이다, -ㄴ/ㄹ 듯(도) 하다, -ㄴ/ㄹ 것 같다, -ㄹ 것을(걸) 그랬다, -어서는 안 된다, -고 해서, -든지 하다’
 - 관용적으로 주어나 목적어를 생략하는 서술어는 복원 대상에서 제외함.
‘관한, 대한, 의한, 향한, 인한, 통한, 따른, 아닌, 같은’, ‘관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고, 그러면서’ 등
 - 선행어(antecedent): 생략된 표현에 의미를 부여하는 명사
 - 선행어가 문서 내에 나타나지 않으면 주어는 ‘누군가, 무언가’로 복원하고, 목적어는 ‘누구를,

무엇을로 복원함. 이때 선행어의 문장 번호(sentence_id)나 문장 내 위치(begin, end)는 모두 '-1'임.

- 선행어가 복수 어절의 명사구이면 가장 마지막 어절을 선행어로 분석하였음.
- 선행어에 부착된 조사는 분석 결과에 포함되지 않음. 선행어와 조사가 결합하여 축약형으로 쓰이면 선행어의 원형을 기준으로 기술하였음.

· 분량: 총 300만 어절

- 문어 200만 어절
- 구어 100만 어절

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 2개

- 문어 624MB
- 구어 366MB

· 인용:

(국문) 국립국어원(2021). 국립국어원 무형 대용어 복원 말뭉치 2020(버전 1.0). URL: <https://kli.korean.go.kr/corpus>

(영문) National Institute of Korean Language (2021). NIKL Zero Anaphora Corpus (v.1.0).

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

· 파일 명명 규칙

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14
속성	유형	구성 방법	주석 단계	배포 연도	일련번호(8자리)									
정의 값	N: 신문 S: 구어	X: 추출	ZA: 무형 대용어 복원	20: 2020년	00000001 ~ 99999999 (여덟 자리 일련번호)									
※ 예시: NXZA2002003180.json 2020년에 배포한 신문 자료 가공 무형 대용어 복원 분석 파일 SXZA2002003180.json 2020년에 배포한 구어 자료 가공 무형 대용어 복원 분석 파일														

· 예시

```

"sentence": [
  {
    "id": "NWRW1800000026.251.3.1",
    "form": "창원시는 기존 누비자의 단점을 보완, 내구성·안전성·편의성 등을 획기적으로 개선한 '명품 누비자'를 이달 말 보급한다고 26일 밝혔다.",
    "word": [
      {
        "id": 1,

```

```
"form": "창원시는",
"begin": 0,
"end": 4
},
{
  "id": 2,
  "form": "기존",
  "begin": 5,
  "end": 7
},
{
  "id": 3,
  "form": "누비자의",
  "begin": 8,
  "end": 12
},
{
  "id": 4,
  "form": "단점을",
  "begin": 13,
  "end": 16
},
{
  "id": 5,
  "form": "보완",
  "begin": 17,
  "end": 20
},
{
  "id": 6,
  "form": "내구성·안전성·편의성",
  "begin": 21,
  "end": 32
},
{
  "id": 7,
  "form": "등을",
  "begin": 33,
  "end": 35
},
{
  "id": 8,
  "form": "획기적으로",
  "begin": 36,
  "end": 41
},
{
  "id": 9,
  "form": "개선한",
  "begin": 42,
  "end": 45
},
{
  "id": 10,
  "form": "'명품'",
  "begin": 46,
```

```

        "end": 49
      },
      {
        "id": 11,
        "form": "누비자'를",
        "begin": 50,
        "end": 55
      },
      {
        "id": 12,
        "form": "이달",
        "begin": 56,
        "end": 58
      },
      {
        "id": 13,
        "form": "말",
        "begin": 59,
        "end": 60
      },
      {
        "id": 14,
        "form": "보급한다고",
        "begin": 61,
        "end": 66
      },
      {
        "id": 15,
        "form": "26일",
        "begin": 67,
        "end": 70
      },
      {
        "id": 16,
        "form": "밝혔다.",
        "begin": 71,
        "end": 75
      }
    ]
  },
  {
    "id": "NWRW1800000026.251.3.2",
    "form": "삼천리자전거와 공동 개발한 '명품 누비자'는 내구성 강화를 위해 프레임을 고급 알루미늄 합금 소재로 제작했고, 프레임 직경을 기존 55mm에서 57mm로 키웠다.",
    "word": [
      {
        "id": 1,
        "form": "삼천리자전거와",
        "begin": 0,
        "end": 7
      },
      {
        "id": 2,
        "form": "공동",
        "begin": 8,
        "end": 10
      }
    ]
  }

```

```
},  
{  
  "id": 3,  
  "form": "개발한",  
  "begin": 11,  
  "end": 14  
},  
{  
  "id": 4,  
  "form": "'명품",  
  "begin": 15,  
  "end": 18  
},  
{  
  "id": 5,  
  "form": "누비자'는",  
  "begin": 19,  
  "end": 24  
},  
{  
  "id": 6,  
  "form": "내구성",  
  "begin": 25,  
  "end": 28  
},  
{  
  "id": 7,  
  "form": "강화를",  
  "begin": 29,  
  "end": 32  
},  
{  
  "id": 8,  
  "form": "위해",  
  "begin": 33,  
  "end": 35  
},  
{  
  "id": 9,  
  "form": "프레임을",  
  "begin": 36,  
  "end": 40  
},  
{  
  "id": 10,  
  "form": "고급",  
  "begin": 41,  
  "end": 43  
},  
{  
  "id": 11,  
  "form": "알루미늄",  
  "begin": 44,  
  "end": 48  
},  
{
```

```
"id": 12,  
"form": "합금",  
"begin": 49,  
"end": 51  
},  
{  
"id": 13,  
"form": "소재로",  
"begin": 52,  
"end": 55  
},  
{  
"id": 14,  
"form": "제작했고",  
"begin": 56,  
"end": 61  
},  
{  
"id": 15,  
"form": "프레임",  
"begin": 62,  
"end": 65  
},  
{  
"id": 16,  
"form": "직경을",  
"begin": 66,  
"end": 69  
},  
{  
"id": 17,  
"form": "기존",  
"begin": 70,  
"end": 72  
},  
{  
"id": 18,  
"form": "55mm에서",  
"begin": 73,  
"end": 78  
},  
{  
"id": 19,  
"form": "57mm로",  
"begin": 79,  
"end": 83  
},  
{  
"id": 20,  
"form": "키웠다.",  
"begin": 84,  
"end": 88  
}  
}  
],
```

```
[...]  
  "ZA": [  
    {  
      "predicate": {  
        "form": "개발한",  
        "sentence_id": "NWRW1800000026.251.3.2",  
        "begin": 11,  
        "end": 14  
      },  
      "antecedent": [  
        {  
          "form": "창원시",  
          "type": "subject",  
          "sentence_id": "NWRW1800000026.251.3.1",  
          "begin": 0,  
          "end": 3  
        },  
        {  
          "form": "누비자",  
          "type": "object",  
          "sentence_id": "NWRW1800000026.251.3.2",  
          "begin": 19,  
          "end": 22  
        }  
      ]  
    }  
  ]  
]
```

· 자료 내용 문의: 02-2669-9638