문화체육관광부
국립국어원

# 국립국어원 개체명 분석 말뭉치 개체 연결 2022

(버전 1.1)

· **자료명:** 국립국어원 개체명 분석 말뭉치 개체 연결 2022

· **공개일**
  · (버전 1.0) 2023. 6. 30.
  · (버전 1.1) 2023. 9. 27.

· **자료 유형:** 텍스트

· **관련 사업:** 2022년 말뭉치 개체명 분석 및 개체 연결(2022)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2022년 말뭉치 개체명 분석 및 개체 연결' 사업 보고서 참고

  · 내용
    - 대상 문서 내 인식된 개체 표현(entity mention)에 대해 그 개체 표현이 문서 내에서 의미적으로 지칭하는 개체를 지식베이스에서 찾아 연결 정보를 부착함.
    - 개체 유형은 '2022년 개체명 분석 지침 ver. 2.5.'에 따라 150개 세분류 체계로 구축된 분석 말뭉치가 기준이며, PERSON(PS)과 LOCATION(LC), ORGANIZATION(OG), ARTIFACTS(AF), DATE(DT), CIVILIZATION(CV), EVENT(EV)의 세분류를 연결 대상으로 함.
    - 연결할 지식베이스는 ① '한국어 위키피디아'를 기본으로 함. ①에 없는 경우, ② '영어 위키피디아'로 연결함. ②에 없는 경우, '없음'에 해당하는 정보(NA)를 표시함.

  · 개체명 분석 표지: 150개(※ 연결 대상 분석 표지에 바탕색 표시)

| 대분류 | 세분류 |
|---|---|
| 1. PERSON(PS) | PS_NAME, PS_CHARACTER, PS_PET |
| 2. STUDY_FIELD(FD) | FD_SCIENCE, FD_SOCIAL_SCIENCE, FD_MEDICINE, FD_ART, FD_HUMANITIES, FD_OTHERS |
| 3. THEORY(TR) | TR_SCIENCE, TR_SOCIAL_SCIENCE, TR_MEDICINE, TR_ART, TR_HUMANITIES, TR_OTHERS |
| 4. ARTIFACTS(AF) | AF_BUILDING, AF_CULTURAL_ASSET, AF_ROAD, AF_TRANSPORT, AF_MUSICAL_INSTRUMENT, AF_WEAPON, AFA_DOCUMENT, AFA_PERFORMANCE, AFA_VIDEO, AFA_ART_CRAFT, AFA_MUSIC, AFW_SERVICE_PRODUCTS, AFW_OTHER_PRODUCTS |

| | |
|---|---|
| 5. ORGANIZATION(OG) | OGG_ECONOMY, OGG_EDUCATION, OGG_MILITARY, OGG_MEDIA, OGG_SPORTS, OGG_ART, OGG_MEDICINE, OGG_RELIGION, OGG_SCIENCE, OGG_LIBRARY, OGG_LAW, OGG_POLITICS, OGG_FOOD, OGG_HOTEL, OGG_OTHERS |
| 6. LOCATION(LC) | LCP_COUNTRY, LCP_PROVINCE, LCP_COUNTY, LCP_CITY, LCP_CAPITALCITY, LCG_RIVER, LCG_OCEAN, LCG_BAY, LCG_MOUNTAIN, LCG_ISLAND, LCG_CONTINENT, LC_SPACE, LC_OTHERS |
| 7. CIVILIZATION(CV) | CV_CULTURE, CV_TRIBE, CV_LANGUAGE, CV_POLICY, CV_LAW, CV_CURRENCY, CV_TAX, CV_FUNDS, CV_ART, CV_SPORTS, CV_SPORTS_POSITION, CV_SPORTS_INST, CV_PRIZE, CV_RELATION, CV_OCCUPATION, CV_POSITION, CV_FOOD, CV_DRINK, CV_FOOD_STYLE, CV_CLOTHING, CV_BUILDING_TYPE |
| 8. DATE(DT) | DT_DURATION, DT_DAY, DT_WEEK, DT_MONTH, DT_YEAR, DT_SEASON, DT_GEOAGE, DT_DYNASTY, DT_OTHERS |
| 9. TIME(TI) | TI_DURATION, TI_HOUR, TI_MINUTE, TI_SECOND, TI_OTHERS |
| 10. QUANTITY(QT) | QT_AGE, QT_SIZE, QT_LENGTH, QT_COUNT, QT_MAN_COUNT, QT_WEIGHT, QT_PERCENTAGE, QT_SPEED, QT_TEMPERATURE, QT_VOLUME, QT_ORDER, QT_PRICE, QT_PHONE, QT_SPORTS, QT_CHANNEL, QT_ALBUM, QT_ADDRESS, QT_OTHERS |
| 11. EVENT(EV) | EV_ACTIVITY, EV_WAR_REVOLUTION, EV_SPORTS, EV_FESTIVAL, EV_OTHERS |
| 12. ANIMAL(AM) | AM_INSECT, AM_BIRD, AM_FISH, AM_MAMMALIA, AM_AMPHIBIA, AM_REPTILIA, AM_TYPE, AM_PART, AM_OTHERS |
| 13. PLANT(PT) | PT_FRUIT, PT_FLOWER, PT_TREE, PT_GRASS, PT_TYPE, PT_PART, PT_OTHERS |
| 14. MATERIAL(MT) | MT_ELEMENT, MT_METAL, MT_ROCK, MT_CHEMICAL |
| 15. TERM(TM) | TM_COLOR, TM_DIRECTION, TM_CLIMATE, TM_SHAPE, TM_CELL_TISSUE_ORGAN, TMM_DISEASE, TMM_DRUG, TMI_HW, TMI_SW, TMI_SITE, TMI_EMAIL, TMI_MODEL, TMI_SERVICE, TMI_PROJECT, TMIG_GENRE, TM_SPORTS |

· 분량

총 약 500만 어절(신문 300만 어절, 일상 대화 100만 어절, 온라인 대화 100만 어절)

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 3개, 총 111MB(ZIP 파일 기준)

· **인용:**

- **(국문)** 국립국어원(2023). 국립국어원 개체명 분석 말뭉치 개체 연결 2022(버전 1.1). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language (2023). NIKL Named Entity Linking 2022 (v.1.1). 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 | 구축 연도 | 일련번호(8자리) |
| 정의값 | N: 신문<br>S: 일상 대화<br>M: 온라인 대화 | X: 추출 | EL: 개체 연결 | 22: 2022년 | 00000001 ~ 99999999<br>(여덟 자리 일련번호) |
| ※ 예시: NXEL2202211217.json 2022년에 구축한 신문 자료 가공 개체 연결 파일<br>SXEL2202211217.json 2022년에 구축한 일상 대화 자료 가공 개체 연결 파일<br>MXEL2202211217.json 2022년에 구축한 온라인 대화 자료 가공 개체 연결 파일 | | | | | |

· **예시**


```
{
    "id": "MDRW2100010933.1.10",
    "form": "캐나다 동부쪽 몬트리올 퀘백 이쪽은 프랑스어 쓰더러구요 프랑스에서 이민을 많이 왔나봐요",
    "word": [
        {
            "id": 1,
            "form": "캐나다",
            "begin": 0,
            "end": 3
        },
        {
            "id": 2,
            "form": "동부쪽",
```

```
            "begin": 4,
            "end": 7
        },
        {
            "id": 3,
            "form": "몬트리올",
            "begin": 8,
            "end": 12
        },
        {
            "id": 4,
            "form": "퀘백",
            "begin": 13,
            "end": 15
        },
        {
            "id": 5,
            "form": "이쪽은",
            "begin": 16,
            "end": 19
        },
        {
            "id": 6,
            "form": "프랑스어",
            "begin": 20,
            "end": 24
        },
        {
            "id": 7,
            "form": "쓰더러구요",
            "begin": 25,
            "end": 30
        },
        {
            "id": 8,
            "form": "프랑스에서",
            "begin": 31,
            "end": 36
        },
```

```
                {
                    "id": 9,
                    "form": "이민을",
                    "begin": 37,
                    "end": 40
                },
                {
                    "id": 10,
                    "form": "많이",
                    "begin": 41,
                    "end": 43
                },
                {
                    "id": 11,
                    "form": "왔나봐요",
                    "begin": 44,
                    "end": 48
                }
            ],
            "NE": [
                {
                    "id": 1,
                    "form": "캐나다",
                    "label": "LCP_COUNTRY",
                    "begin": 0,
                    "end": 3,
                    "kid": "05000000000251",
                    "wikiid": "1464",
                    "URL": "https://ko.wikipedia.org/wiki/%EC%BA%90%EB%82%98%EB%8B%A4"
                },
                {
                    "id": 2,
                    "form": "동부",
                    "label": "TM_DIRECTION",
                    "begin": 4,
                    "end": 6,
                    "kid": "14010000000042",
                    "wikiid": "NA",
```

```
                "URL": "NA"
            },
            {
                "id": 3,
                "form": "몬트리올",
                "label": "LCP_CITY",
                "begin": 8,
                "end": 12,
                "kid": "05030000000214",
                "wikiid": "29299",
                "URL": "https://ko.wikipedia.org/wiki/%EB%AA%AC%ED%8A%B8%EB%A6%AC%EC%98%AC"
            },
            {
                "id": 4,
                "form": "퀘백",
                "label": "LCP_CITY",
                "begin": 13,
                "end": 15,
                "kid": "05030000000521",
                "wikiid": "44525",
                "URL": "https://ko.wikipedia.org/wiki/%ED%80%98%EB%B2%A1_(%EB%8F%84%EC%8B%9C)"
            },
            {
                "id": 5,
                "form": "프랑스어",
                "label": "CV_LANGUAGE",
                "begin": 20,
                "end": 24,
                "kid": "06020000000043",
                "wikiid": "2914",
                "URL": "https://ko.wikipedia.org/wiki/%ED%94%84%EB%9E%91%EC%8A%A4%EC%96%B4"
            },
            {
                "id": 6,
                "form": "프랑스",
                "label": "LCP_COUNTRY",
```

```
                        "begin": 31,
                        "end": 34,
                        "kid": "05000000000289",
                        "wikiid": "95069",
                        "URL": "https://ko.wikipedia.org/wiki/%ED%94%84%EB%9E%91%EC%8A%A4"
                    }
                ]
            },
            {
                "id": "MDRW2100010933.1.11",
                "form": "나중에 한번 가볼만 한거같아요 고고님은 기억에 남는 여행지 있으신가요?",
                "word": [
                    {
                        "id": 1,
                        "form": "나중에",
                        "begin": 0,
                        "end": 3
                    },
                    {
                        "id": 2,
                        "form": "한번",
                        "begin": 4,
                        "end": 6
                    },
                    {
                        "id": 3,
                        "form": "가볼만",
                        "begin": 7,
                        "end": 10
                    },
                    {
                        "id": 4,
                        "form": "한거같아요",
                        "begin": 11,
                        "end": 16
                    },
                    {
```

```
                "id": 5,
                "form": "고고님은",
                "begin": 17,
                "end": 21
            },
            {
                "id": 6,
                "form": "기억에",
                "begin": 22,
                "end": 25
            },
            {
                "id": 7,
                "form": "남는",
                "begin": 26,
                "end": 28
            },
            {
                "id": 8,
                "form": "여행지",
                "begin": 29,
                "end": 32
            },
            {
                "id": 9,
                "form": "있으신가요?",
                "begin": 33,
                "end": 39
            }
        ],
        "NE": [
            {
                "id": 1,
                "form": "고고",
                "label": "PS_NAME",
                "begin": 17,
                "end": 19,
                "kid": "00000000000619",
                "wikiid": "NA",
```

```
                    "URL": "NA"
                }
            ]
        },
```


※ "kid": 국립국어원 개체 연결 표현 고유 번호
"wikiid": 위키피디아 문서 번호
"URL": 위키피디아 주소(분석 대상 개체 표현에 대응하는 위키피디아 문서가 없는 경우 NA)

· **자료 내용 문의:** 02-2669-9638