문화체육관광부
국립국어원

국립국어원 맞춤법 교정 말뭉치 2022

(버전 1.0)

· **자료명:** 국립국어원 맞춤법 교정 말뭉치 2022

· **공개일**
  · (버전 1.0) 2023. 3. 31.

· **자료 유형:** 텍스트

· **관련 사업:** 2022년 맞춤법 교정 말뭉치 연구 분석(2022)

· **자료 설명**
  · 내용
    - 온라인 대화 자료를 대상으로 한국어 처리 도구가 분석할 수 있는 수준으로 오탈자 등을 교정한 말뭉치

  · 분량
    - 약 400만 어절

  · 파일 형식: JSON(UTF-8 인코딩)

  · 파일 수 및 크기: 파일 1개, 총 517MB

  ※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2021년 맞춤법 교정 말뭉치 연구 분석' 사업 결과 보고서 참조.

· **인용**

(국문) 국립국어원(2023). 국립국어원 맞춤법 교정 말뭉치 2022(버전 1.0). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2023). NIKL Spelling Correction Corpus 2022 (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	매체	구성 방법	주석 단계		구축 연도		일련번호(8자리)							
정의 값	M: 메신저	X: 추출	EC: 맞춤법 교정		22: 2022년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: MXEC2202210100.json 2022 년에 구축한 메신저 자료 가공 맞춤법 교정 말뭉치 파일														

· 예시


```
{
    "id": "MXEC2202210100",
    "metadata": {
        "title": "국립국어원 온라인 대화 말뭉치 추출 MXEC2202210100",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2022",
        "category": [
            "온라인 대화 > 2인 대화",
            "온라인 대화 > 다자 대화"
        ],
        "annotation_level": "맞춤법 교정",
        "sampling": "부분 추출 - 임의 추출"
    },
    "document": [
        {
            "id": "MDRW2100000002.1",
            "metadata": {
                "title": "온라인 대화",
                "author": "개인 대화 참여자",
                "publisher": "카카오톡",
                "date": "20210518",
                "topic": "연애와 결혼",
                "speaker": [
                    {
                        "id": "1",
                        "age": "20대",
                        "occupation": "사무 종사자",
                        "sex": "여성",
                        "birthplace": "서울",
                        "principal_residence": "서울",
                        "current_residence": "서울",
                        "device": "스마트폰",
                        "keyboard": "2벌식(쿼티)"
                    },
                    {
                        "id": "2",
                        "age": "20대",
                        "occupation": "전문가 및 관련 종사자",
                        "sex": "여성",
```

```
                        "birthplace": "서울",
                        "pricipal_residence": "해외/기타",
                        "current_residence": "서울",
                        "device": "스마트폰",
                        "keyboard": "2벌식(쿼티)"
                    }
                ],
                "setting": {
                    "relation": "직장>선후배/상사-부하"
                }
            },
            "utterance": [
                {
                    "id": "MDRW2100000002.1.1",
                    "original_form": "하이하이",
                    "form": "하이하이",
                    "corrected_form": "하이하이.",
                    "speaker_id": "2",
                    "emoticon": [],
                    "meaningless_words": []
                },
                {
                    "id": "MDRW2100000002.1.2",
                    "original_form": "반가워욬ㅌㅋㅋ",
                    "form": "반가워욬ㅌㅋㅋ",
                    "corrected_form": "반가워요. ㅋㅌㅋㅋ",
                    "speaker_id": "1",
                    "emoticon": [],
                    "meaningless_words": []
                },
…
                {
                    "id": "MDRW2100000002.1.10",
                    "original_form": "요새 강철부대 육준서에 마음이 선덕선덕댑니다{emoji:…}{emoji:…}.^^",
                    "form": "요새 강철부대 육준서에 마음이 선덕선덕댑니다.^^",
                    "corrected_form": "요새 강철부대 육준서에 마음이 선덕선덕댑니다. ^^",
                    "speaker_id": "2",
                    "emoticon": [
                        {
                            "begin": 25,
                            "end": 27,
                            "value": "^^"
                        }
                    ],
                    "meaningless_words": []
                },
…
                {
                    "id": "MDRW2100000005.40.11",
                    "original_form": "&hate-speech&",
                    "form": "hate-speech",
                    "corrected_form": "hate-speech",
                    "speaker_id": "1",
                    "emoticon": [],
                    "meaningless_words": []
                },
```

※ “original_form”: 수집한 언어 자료의 원문을 그대로 유지한 형태(개인 정보 등은 비식별 처리됨.)
“form”: 원문(original_form)에서 ‘연속된 여러 개의 공백(스페이스, 탭 등), 특수 메시지, 비식별화 기호 등’을 제거하여 전처리한 형태
“corrected_form”: ‘form’에서 제공된 언어 표현을 바탕으로 한국어 처리 도구가 분석할 수 있는 수준으로 오탈자 등을 교정한 형태

· **자료 내용 문의:** 02-2669-9638