문화체육관광부
국립국어원

국립국어원 맞춤법 교정 말뭉치

(버전 1.0)

· **자료명:** 국립국어원 맞춤법 교정 말뭉치

· **공개일**

· (버전 1.0) 2022. 4. 1.

· **자료 유형:** 텍스트

· **관련 사업:** 2021년 맞춤법 교정 말뭉치 연구 분석(2021)

· **자료 설명**

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

· 내용

- 메신저 대화 자료 및 누리 소통망 등에서 수집한 웹 언어 자료를 대상으로 한국어 처리 도구가 분석할 수 있는 수준으로 오탈자 등을 교정한 말뭉치임.
- 세부 교정 지침은 2021년 맞춤법 교정 말뭉치 연구 분석 사업 결과 보고서를 참고할 것. (국립국어원 누리집 > 자료 > 연구·조사 자료 > '2021년 맞춤법 교정 말뭉치 연구 분석' 사업 보고서)

· 분량

- 약 250만 어절(메신저 대화 136만 어절, 웹 자료 114만 어절)

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 2개, 총 195MB

· **인용:**

(국문) 국립국어원(2022). 국립국어원 맞춤법 교정 말뭉치 2021(버전 1.0). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2022). NIKL Spelling Correction Corpus 2021 (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 E: 웹	X: 추출	EC: 맞춤법 교정	21: 2021년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시: MXEC2102112091.json 2021 년에 구축한 메신저 자료 가공 맞춤법 교정 말뭉치 파일 EXSEC102112091.json 2021 년에 구축한 웹 자료 가공 맞춤법 교정 말뭉치 파일					

· **예시**


```
{
    "id": "MXEC2102112091",
    "metadata": {
        "title": "국립국어원 메신저 말뭉치 추출 MXEC2102112091",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2021",
        "category": [
            "메신저 대화 > 2인 대화",
            "메신저 대화 > 다자 대화"
        ],
        "annotation_level": "맞춤법 교정",
        "sampling": "부분 추출 - 임의 추출"
    },
    "document": [
    {
        "id": "MDRW1900000940.1",
        "metadata": {
            "title": "메신저 대화",
            "author": "개인 대화 참여자",
            "publisher": "카카오톡",
            "date": "20191219",
            "topic": "미용과 건강 (질병과 치료, 운동, 다이어트, 미용)",
            "speaker": [
                {
                    "id": "1",
                    "age": "30대",
                    "occupation": "가정 주부",
                    "sex": "여성",
                    "birthplace": "경남",
                    "pricipal_residence": "경남",
                    "current_residence": "경남",
                    "device": "스마트폰",
                    "keyboard": "모아키"
                },
                {
                    "id": "2",
```

```
                    "age": "30대",
                    "occupation": "전문가 및 관련 종사자",
                    "sex": "남성",
                    "birthplace": "경북",
                    "pricipal_residence": "경북",
                    "current_residence": "경남",
                    "device": "스마트폰",
                    "keyboard": "천지인"
                }
            ],
            "setting": {
                "relation": "지역 : 현 거주지 지인"
            }
        },
        "utterance": [
            {
                "id": "MDRW1900000940.1.1.1",
                "original_form": "오빠야",
                "form": "오빠야",
                "corrected_form": "오빠야,",
                "speaker_id": "1"
            },
            {
                "id": "MDRW1900000940.1.1.2",
                "original_form": "요즘먹는영양제 뭐있음?",
                "form": "요즘먹는영양제 뭐있음?",
                "corrected_form": "요즘 먹는 영양제 뭐 있음?",
                "speaker_id": "1"
            },
            {
                "id": "MDRW1900000940.1.1.3",
                "original_form": "난 아는 약사가 추천해준 거 먹는데",
                "form": "난 아는 약사가 추천해준 거 먹는데",
                "corrected_form": "난 아는 약사가 추천해준 거 먹는데",
                "speaker_id": "2"
            },
            {
                "id": "MDRW1900000940.1.1.4",
                "original_form": "뭐먹는디",
                "form": "뭐먹는디",
                "corrected_form": "뭐 먹는데?",
                "speaker_id": "1"
            },
            {
                "id": "MDRW1900000940.1.1.5",
                "original_form": "너무 비싸서 다시 외산으로 갈까싶다..",
                "form": "너무 비싸서 다시 외산으로 갈까싶다..",
                "corrected_form": "너무 비싸서 다시 외산으로 갈까 싶다...",
                "speaker_id": "2"
            },
            {
                "id": "MDRW1900000940.1.1.6",
                "original_form": "ㅋㅋㅋㅋㅋㅋ뭐를얼마에먹고잇는데 ㅋㅋㅋㅋㅋ",
                "form": "ㅋㅋㅋㅋㅋㅋ뭐를얼마에먹고잇는데 ㅋㅋㅋㅋㅋ",
                "corrected_form": "ㅋㅋㅋㅋㅋㅋ 뭐를 얼마에 먹고 있는데? ㅋㅋㅋㅋㅋ",
```

```
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.7",
            "original_form": "이름은 지금은 기억 안나는데 녹십자꺼임",
            "form": "이름은 지금은 기억 안나는데 녹십자꺼임",
            "corrected_form": "이름은 지금은 기억 안 나는데 녹십자 거임.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.8",
            "original_form": "녹십자꺼 비싸?",
            "form": "녹십자꺼 비싸?",
            "corrected_form": "녹십자 거 비싸?",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.9",
            "original_form": "두달치에 6만 정도더라",
            "form": "두달치에 6만 정도더라",
            "corrected_form": "두 달 치에 6만 정도더라.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.10",
            "original_form": "그정도면 비싼건 아닌데",
            "form": "그정도면 비싼건 아닌데",
            "corrected_form": "그 정도면 비싼 건 아닌데",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.11",
            "original_form": "싸지도 않지",
            "form": "싸지도 않지",
            "corrected_form": "싸지도 않지.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.12",
            "original_form": "단백질보충제?",
            "form": "단백질보충제?",
            "corrected_form": "단백질 보충제?",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.13",
            "original_form": "아님 종합비타민?",
            "form": "아님 종합비타민?",
            "corrected_form": "아님 종합 비타민?",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.14",
            "original_form": "종합~",
            "form": "종합~",
```

```
            "corrected_form": "종합~",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.15",
            "original_form": "보충제도 사긴 해야되네",
            "form": "보충제도 사긴 해야되네",
            "corrected_form": "보충제도 사긴 해야 되네.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.16",
            "original_form": "왜 요새 체중이 줄었으",
            "form": "왜 요새 체중이 줄었으",
            "corrected_form": "왜 요새 체중이 줄었어",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.17",
            "original_form": "?",
            "form": "?",
            "corrected_form": "?",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.18",
            "original_form": "닌 저번에 가니까 이것저것 많더만",
            "form": "닌 저번에 가니까 이것저것 많더만",
            "corrected_form": "닌 저번에 가니까 이것저것 많더구먼.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.19",
            "original_form": "글치",
            "form": "글치",
            "corrected_form": "글치.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.20",
            "original_form": "난 모유수유 중이잖아",
            "form": "난 모유수유 중이잖아",
            "corrected_form": "난 모유 수유 중이잖아.",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.21",
            "original_form": "ㄷㄷㄷ",
            "form": "ㄷㄷㄷ",
            "corrected_form": "ㄷㄷㄷ",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.22",
            "original_form": "근육량이 줄어서",
```

```
            "form": "근육량이 줄어서",
            "corrected_form": "근육량이 줄어서.",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.23",
            "original_form": "칼슘이랑 철분 종합비타민 먹음",
            "form": "칼슘이랑 철분 종합비타민 먹음",
            "corrected_form": "칼슘이랑 철분, 종합 비타민 먹음.",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.24",
            "original_form": "ㅋㅋ 먹으면 좀 힘이 나나",
            "form": "ㅋㅋ 먹으면 좀 힘이 나나",
            "corrected_form": "ㅋㅋ 먹으면 좀 힘이 나나?",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.25",
            "original_form": "아니 죽지않을수있다",
            "form": "아니 죽지않을수있다",
            "corrected_form": "아니, 죽지 않을 수 있다.",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.26",
            "original_form": "힘은무슨 ㅋㅋㅋㅋ",
            "form": "힘은무슨 ㅋㅋㅋㅋ",
            "corrected_form": "힘은 무슨. ㅋㅋㅋㅋ",
            "speaker_id": "1"
        },
        {
            "id": "MDRW1900000940.1.1.27",
            "original_form": "ㅋㅋㅋ",
            "form": "ㅋㅋㅋ",
            "corrected_form": "ㅋㅋㅋ",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.28",
            "original_form": "죽을똥 말똥 할 때 머리끄댕이 잡고 당겨주는건가",
            "form": "죽을똥 말똥 할 때 머리끄댕이 잡고 당겨주는건가",
            "corrected_form": "죽을 둥 말 둥 할 때 머리끄덩이 잡고 당겨주는 건가?",
            "speaker_id": "2"
        },
        {
            "id": "MDRW1900000940.1.1.29",
            "original_form": "그런거임",
            "form": "그런거임",
            "corrected_form": "그런 거임.",
            "speaker_id": "1"
        },
```


※ “original_form”: 수집한 언어 자료의 원문을 그대로 유지한 형태(개인 정보 등은 비식별화)

“form”: 원문에서 연속된 여러 개의 공백(스페이스, 탭 등), 특수 메시지, 비식별화 기호 등을 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