우리말 말뭉치 구축과 활용

덕성여자대학교 영어영문학과 “빅데이터와 영어구문분석“ 특강

2022. 5. 10.

서 샛 별

국립국어원 학예연구사

서 샛 별

학력

2005 서강대학교 영미어문학과 문학사

2009 서울대학교 언어학과 문학 석사(통사론 전공)

"*Constraints on NP-ellipsis and DP-internal movement*"

2017 서울대학교 언어학과 언어학 박사(통사론 전공)

"*The syntax of jussives: speaker and hearer at the syntax-discourse interface*"

경력

2019 - 학예연구사

문화체육관광부 국립국어원 언어정보과

개요

- 국립국어원
- 모두의 말뭉치(한국어 빅데이터)
- 문화체육관광부 연구직 공무원

국립국어원

설립 목적: 국어의 발전, 국민의 언어 생활 향상, 체계적 정책 수립 기반 마련

- 합리적 국어 정책 추진에 필요한 체계적 조사, 연구
- 원활한 의사 소통을 위한 국어 사용 환경 개선
- 한국어 교육의 질을 높이는 기반 조성
- 국가 언어 자원 수집, 통합

조직

모두의 말뭉치
(한국어 빅데이터)

국립국어원 언어정보과

문화체육관광부
국립국어원

말뭉치 = 한국어 빅데이터

말뭉치란?

언어학 용어인 코퍼스(corpus)를 번역한 말

컴퓨터가 분석하고 처리할 수 있도록 입력된 대량의 언어 자료

책, 신문 기사, 잡지, 보고서, 일상 대화, 강연, 드라마 대본, 회의, 블로그, 메신저 대화 …… 사람의 모든 말과 글

말뭉치의 활용

언어 연구 및 어문 정책 수립

사전 편찬, 언어 교육

언어 인공지능(자연언어처리, NLP) 개발

원시 말뭉치

신문 말뭉치, 일상 대화 말뭉치, 웹 말뭉치, ...

원 자료(신문, 책, 대화 음성, 카카오톡 대화, 블로그 글, ...)

⇒ 언어 자료 정제, 메타 정보 입력, 형식 정보 입력

⇒ 원시 말뭉치

원시 말뭉치: 일상 대화 말뭉치


```
{
    "id": "SDRW2000000002",
    "metadata": {
        "title": "국립국어원 구어 말뭉치 SDRW2000000002",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2020",
        "category": "구어 > 사적 대화 > 일상대화",
        "annotation_level": [
            "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id": "SDRW2000000002.1",
            "metadata": {
                "title": "2인 일상 대화",
                "author": "개인 발화자",
                "publisher": "개인 발화 녹음",
                "date": "20200602",
                "topic": "반려동물 > 보험, 유튜버, 동물학대, 작명",
                "speaker": [
                    {
                        "id": "SD2000003",
                        "age": "30대",
                        "occupation": "전문가 및 관련 종사자",
                        "sex": "여성",
                        "birthplace": "서울",
                        "principal_residence": "서울",
                        "current_residence": "서울",
                        "education": "대졸"
                    },
                    {
                        "id": "SD2000004",
                        "age": "20대",
                        "occupation": "학생",
                        "sex": "여성",
                        "birthplace": "서울",
                        "principal_residence": "서울",
                        "current_residence": "서울",
                        "education": "대재"
                    }
                ],
                "setting": {
                    "relation": "기타"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2000000002.1.1.1",
                    "form": "반려동물을 키우고 계신가요?",
                    "original_form": "반려동물을 키우고 계신가요?",
                    "speaker_id": "SD2000003",
                    "start": 2.78903,
                    "end": 4.92608,
                    "note": ""
                },
```

분석 말뭉치

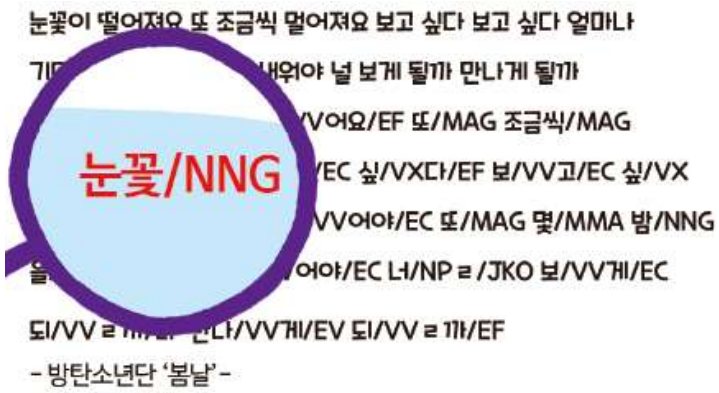
당신의 말과 글 소중한 언어 자원입니다.(12-14쪽)

구문 분석 말뭉치, 개체명 말뭉치, 감성 분석 말뭉치, ...

원시 말뭉치

⇒ 품사, 의미, 문장 구조 등 여러 가지 분석 정보 부가

⇒ 분석 말뭉치



그때는 나 어릴 때는 아무것도 몰랐네 그 다리 위를 건너가는 기분을 어디서냐고 어디냐고 여쭤보면

아버지/CV_RELATION는 항상 양화대교/AF_BUILDING, 양화대교/AF_BUILDING

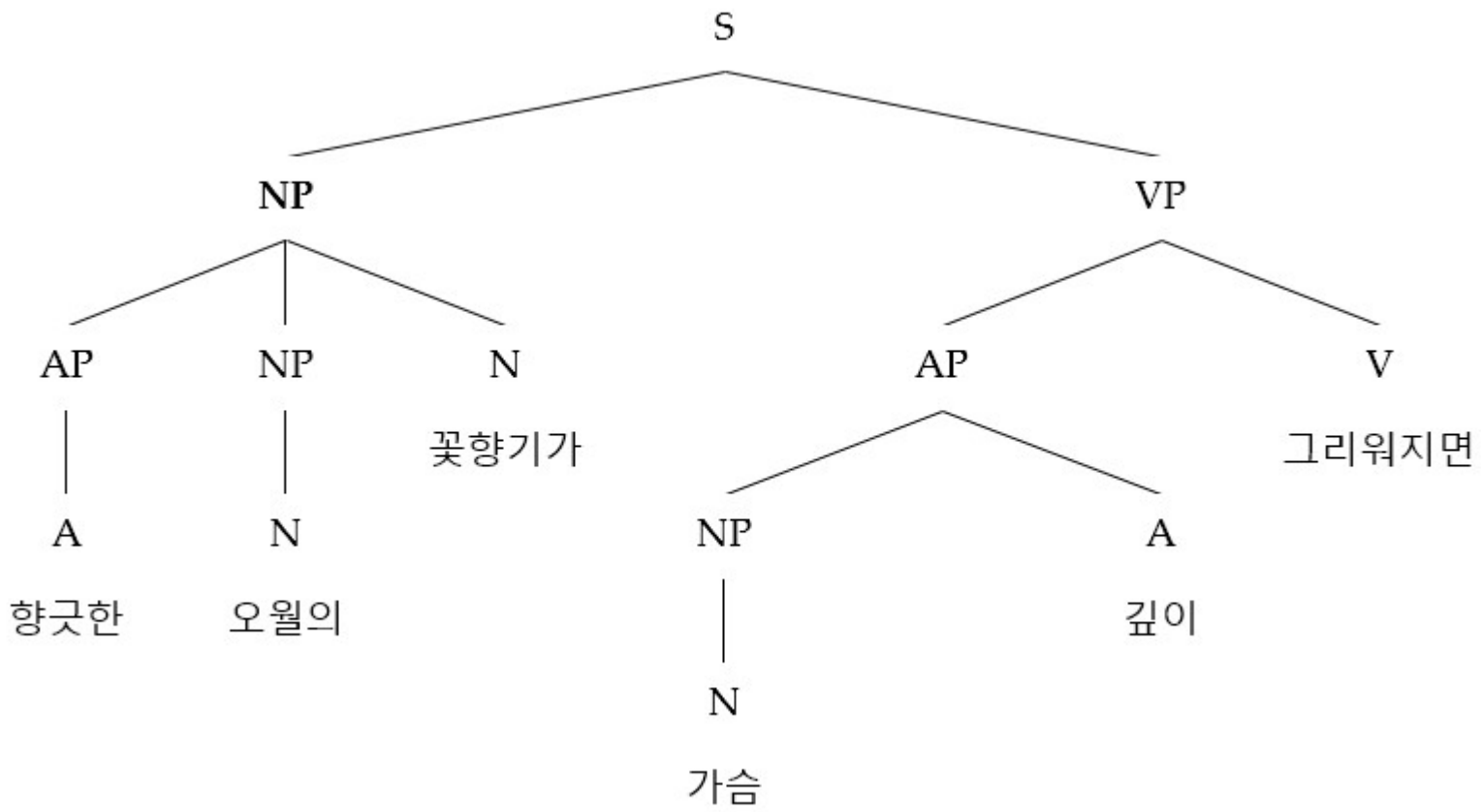
이제 나는 서 있네 그 다리 위에 그 다리에

- 자이언티, '양화대교' -

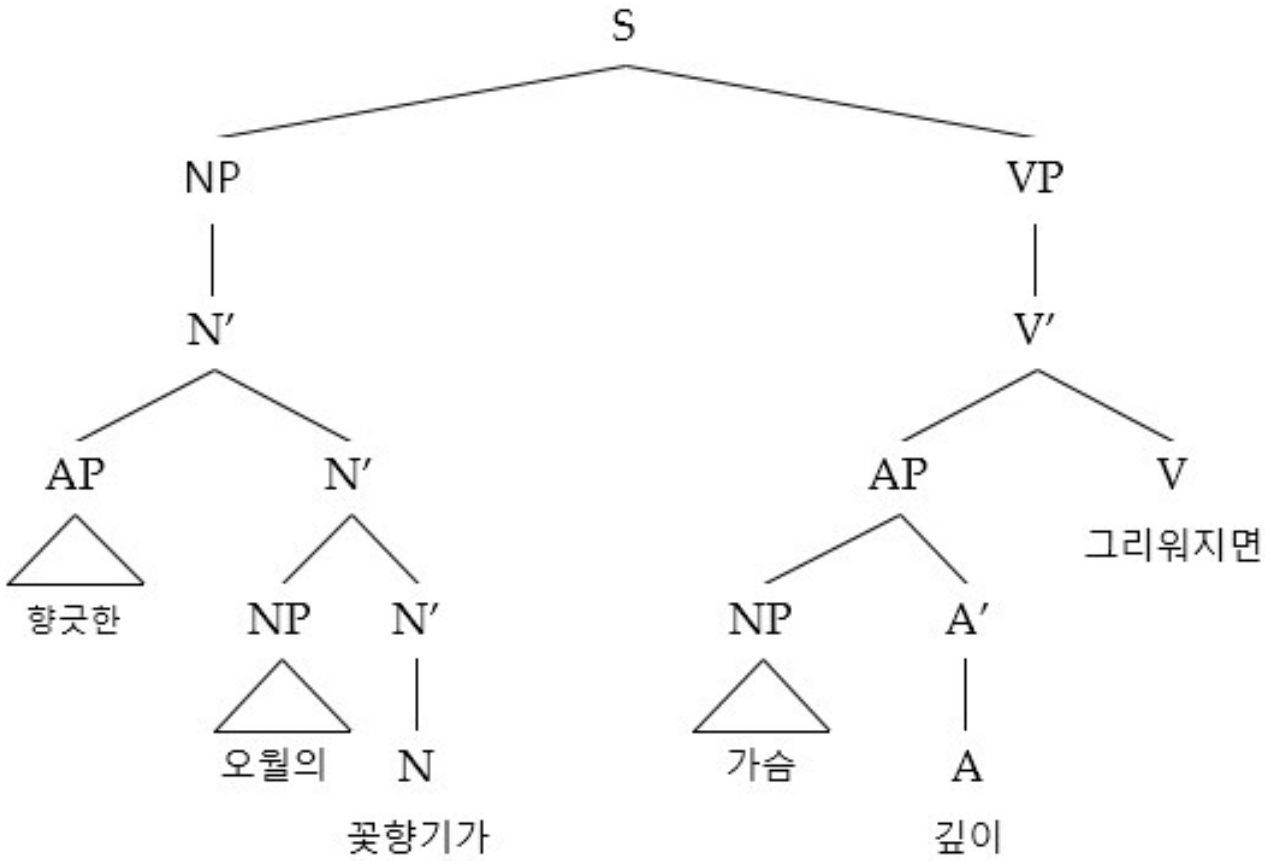
구문 분석

향긋한 오월의 꽃향기가 가슴 깊이 그리워지면
- 이문세, “광화문 연가”

구문 분석: 구 구조 규칙



구문 분석: X'-이론



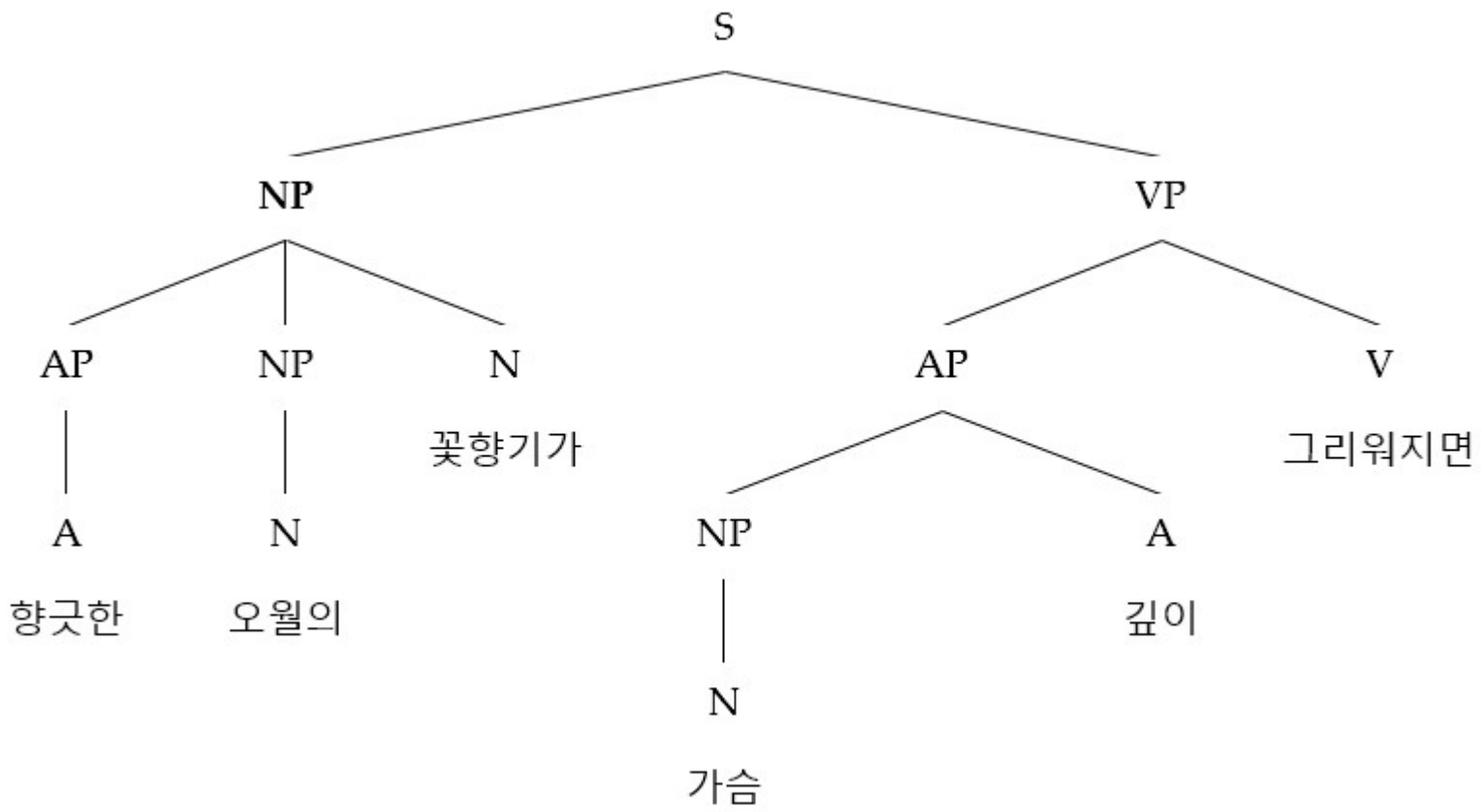
말뭉치 구축을 위한 구문 분석

단위화, 묶음(CHUNKING)이 중요

향긋한 오월의 꽃향기가 가슴 깊이 그리워지면

- [향긋한 오월의]? [향긋한 꽃향기가]?
- [꽃향기가 가슴 깊이]? [꽃향기가 그리워지면]?

말뭉치 구축을 위한 구문 분석: 구 구조 규칙



의존 구문분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법 (한국정보통신기술협회 표준)

구문 표지	기능 표지
NP(체언)	SBJ(주어)
VP(용언)	OBJ(목적어)
AP(부사구)	MOD(관형어)
VNP(긍정 지정사구)	AJT(부사어)
DP(관형사구)	CMP(보어)
IP(감탄사구)	CNJ(접속어)
X(의사구)	
L(부호-왼쪽 괄호 및 따옴표)	
R(부호-오른쪽 괄호 및 따옴표)	

의존 관계 구문 분석(그림)

의존 관계 구문 분석(표)

Form	ID	Begin	End	Head	Label
향긋한	1	0	3	3	VP_MOD
오월의	2	4	7	3	NP_MOD
꽃향기가	3	8	12	6	NP_SBJ
가슴	4	13	15	5	NP_AJT
깊이	5	16	18	6	AP
그리워지면	6	19	24	…	VP

분석 말뭉치: 구문 분석 말뭉치 예시


```
{
    "id": "NWRW1800000021.1.2.1",
    "form": "경기 성남시 판교신도시에서 이달 분양하는 중대형 아파트의 3.3m²당 분양가가 2006년보다 200만 원 정도 싼 1500만 원 후반대로 결정될 것으로 보인다.",
    "word": [
        {
            "id": 1,
            "form": "경기",
            "begin": 0,
            "end": 2
        },
        {
            "id": 2,
            "form": "성남시",
            "begin": 3,
            "end": 6
        },
        {
            "id": 3,
            "form": "판교신도시에서",
            "begin": 7,
            "end": 14
        },
        {
            "id": 4,
            "form": "이달",
            "begin": 15,
            "end": 17
        },
        {
            "id": 5,
            "form": "분양하는",
            "begin": 18,
            "end": 22
        },
        {
            "id": 6,
            "form": "중대형",
            "begin": 23,
            "end": 26
        },
        {
            "id": 7,
            "form": "아파트의",
            "begin": 27,
            "end": 31
        },
        {
            "id": 8,
            "form": "3.3m²당",
            "begin": 32,
            "end": 38
        },
        {
            "id": 20,
            "form": "보인다.",
            "begin": 85,
            "end": 89
        }
    ],
    "DP": [
        {
            "word_id": 1,
            "word_form": "경기",
            "head": 2,
            "label": "NP",
            "dependent": []
        },
        {
            "word_id": 2,
            "word_form": "성남시",
            "head": 3,
            "label": "NP",
            "dependent": [
                1
            ]
        },
        {
            "word_id": 3,
            "word_form": "판교신도시에서",
            "head": 5,
            "label": "NP_AJT",
            "dependent": [
                2
            ]
        },
        {
            "word_id": 4,
            "word_form": "이달",
            "head": 5,
            "label": "NP_AJT",
            "dependent": []
        },
        {
            "word_id": 5,
            "word_form": "분양하는",
            "head": 7,
            "label": "VP_MOD",
            "dependent": [
                3,
                4
            ]
        },
        {
            "word_id": 6,
            "word_form": "중대형",
            "head": 7,
            "label": "NP",
            "dependent": []
        },
        {
            "word_id": 7,
            "word_form": "아파트의",
            "head": 9,
            "label": "NP_MOD",
            "dependent": [
                5,
                6
            ]
```

4차 산업혁명 대비 국어 빅데이터 구축 사업

추진 배경 국어 빅데이터(말뭉치) 부족에 따른 인공지능 산업의 한국어 서비스 수준 지체
인공지능 산업 발전을 위한 대규모 고품질 우리말 자원 수요 증대

사업 기간 ’18 ~ (계속)

사업 예산 348억(’18~’22)

추진 목적 **인공지능 기술 개발 및 연구의 기반이 되는 대규모 고품질 국어 말뭉치 구축**

추진 내용

▪ **대규모 현대 한국어 원시 말뭉치 구축**

- 현대 한국어 사용 양상을 반영한 언어 자료를 수집, 정리하여 말뭉치 구축
- 활용 제약 최소화를 위해 말뭉치 구축 대상 언어 자료에 대한 저작권 처리 선행

▪ **분석, 평가용 말뭉치 구축**

- 언어 처리 기반의 응용 시스템 개발에 필요한 다양한 언어 정보 분석 말뭉치
- 한국어 인공지능 시스템 평가를 위한 평가용 말뭉치 구축

▪ **말뭉치 배포 및 활용 지원**

- 말뭉치 통합 시스템 개발
- 배포 말뭉치 정제 및 가공

모두의 말뭉치

https://corpus.korean.go.kr/

32종 24억 어절 규모의 한국어 말뭉치 공개

모두의 말뭉치

https://corpus.korean.go.kr/

신규

속성 기반 감성 분석…

(버전 1.0) 국립국어원 감성 분석 말뭉치 2020(1.0)과 동일한 문서에 속성 기반 감성 정보를 부착한 말뭉치입니다.

신청하기

신규

개체명 분석 말뭉치 2…

(버전 1.0) 문장에 나타난 개체명의 경계를 표시하고 분석 표지를 부착한 말뭉치입니다.

신청하기

신규

개체명 분석 말뭉치 …

(버전 1.0) 개체명 분석 말뭉치에 위키피디아 정보를 부착한 자료입니다.

신청하기

신규

온라인 대화 말뭉치 2…

(버전 1.0) 두 명 이상의 대화 참여자가 온라인 공간에서 주고받은 대화 자료로 구성된 말뭉치입니다.

신청하기

추론_확신성 분석 말…

(버전 1.0) 내포문에서 추출한 가설에 화자의 확신성을 추론한 정보를 부착한 말뭉치입니다.

신청하기

의미역 분석 말뭉치

(버전 1.0) 문장의 술어가 가지는 논항을 분석하고 의미 역할을 부착한 말뭉치입니다.

신청하기

수정

개체명 분석 말뭉치 2…

(버전 2.0) 문장에 나타난 개체명의 경계를 표시하고 분석 표지를 부착한 말뭉치입니다.

신청하기

비출판물 말뭉치

(버전 1.1) 개인적 글쓰기 자료(시, 일기, 편지, 감상문 등)로 구성된 말뭉치입니다.

신청하기

수정

어휘 의미 분석 말뭉치…

(버전 2.0) 다의어를 구별하여 <우리말샘>의 의미 번호를 부착한 말뭉치입니다.

신청하기

일상 대화 음성 말뭉…

(버전 1.2) 일상 대화의 음성(PCM 파일)과 전사 자료로 구성된 말뭉치입니다.

신청하기

일상 대화 말뭉치 2020

(버전 1.2) 특정 주제 또는 제시 자료로 자유롭게 대화를 나눈 일상 대화 말뭉치입니다.

신청하기

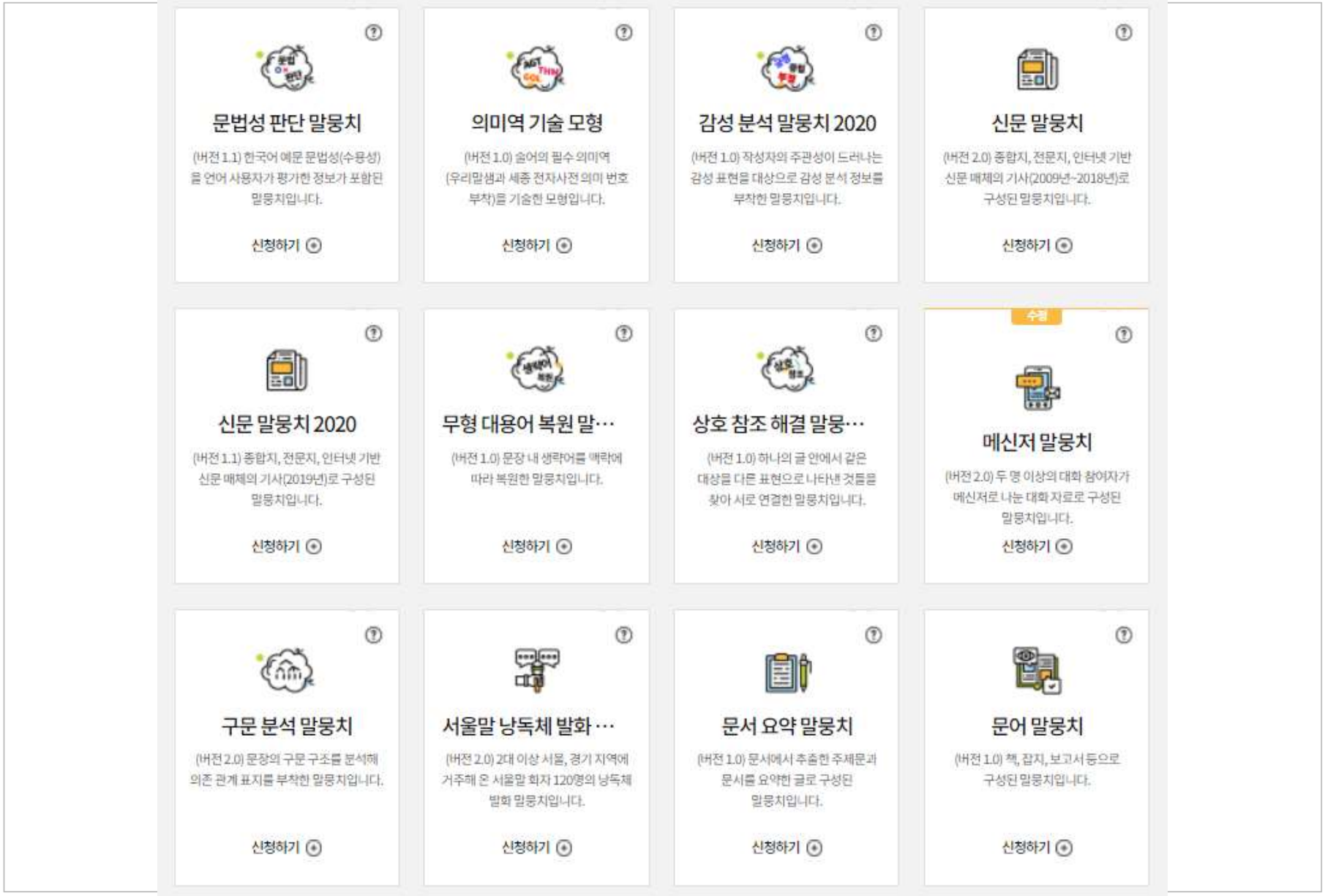
구어 말뭉치

(버전 1.2) 방송, 강연 등의 공적 구어 자료, 드라마 대본 등의 준구어 자료로 구성된 말뭉치입니다.

신청하기

모두의 말뭉치

https://corpus.korean.go.kr/



모두의 말뭉치

https://corpus.korean.go.kr/

말뭉치 활용

- 컴퓨터의 언어 학습 모델(GPT-3)
- 자동 통번역 앱(프로그램)
- 음성 인식(음성 인식 스피커, 자율 주행차)
- 챗봇 상담(관광, 보험 등)
- 문장 교정 프로그램(맞춤법 검사기 등)
- 문서 요약 프로그램
- 수어 교육 앱(프로그램)
- 점역·역점역 자동화 프로그램 등 개발

모두의 말뭉치 활용

[단독] SK텔레콤, 요약 잘하는 AI모델 공개...'자연어이해' 기술 강화

최근 깃허브에 'KoBART' 올려...3번째 AI모델 오픈
페이스북 AI모델 한국어판...'한 줄 요약' 능력 강점

SK텔레콤은 T3K 센터의 깃허브 프로젝트 소개 문구를 통해 "한국어 BART는 (페이스북 연구진이 작성한) 논문에서 사용된 Text Infilling 노이즈 함수를 사용해 40GB 이상의 한국어 텍스트에 대해 학습한 한국어 encoder-decoder 언어 모델"이라고 밝혔다. 현재 이 소스코드와 함께 공개된 요약기능 데모를 통해 한국어 뉴스 원문을 '한 줄 요약'해볼 수 있다.

코바트는 약 2억7500만개 문장 분량의 한국어 텍스트 원문을 활용해 학습했다. 이 데이터는 한국어 위키백과 문장 500만개, 청와대 국민청원에서 재작년 8월 이전 기준으로 만료된 청원 데이터 및 국립국어원이 올해 8월 25일 공개한 '모두의 말뭉치' 데이터(대화·뉴스 등)의 문장 2억7000만개 등 다양하다.

네이버 초거대 AI '하이퍼클로바', 뭐가 다를까

발행일 2021-05-25 18:31:05

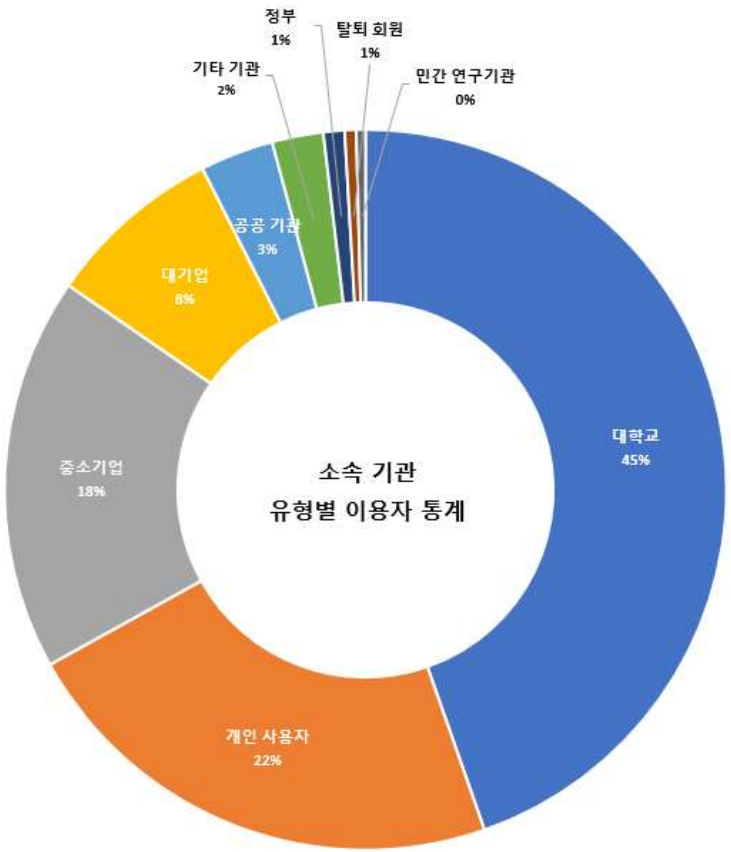
데이터는 어디서 얻었나

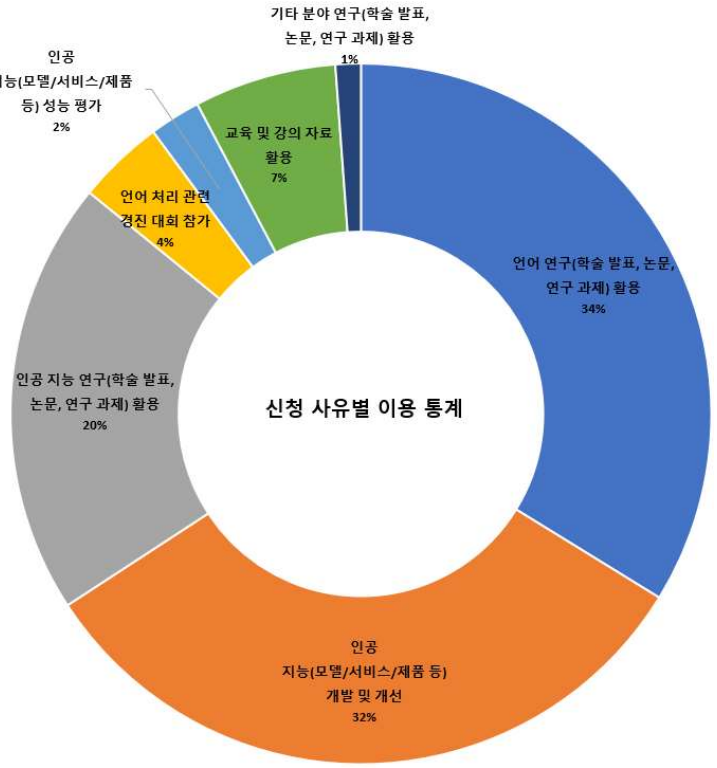
하이퍼클로바는 네이버가 가진 '자료'에 뿌리를 두고 있다. 뉴스부터 블로그·지식인·카페·웹문서가 대상이었다. 국내에 있는 전반적인 데이터를 두고 품질 순으로 데이터를 골랐다. 먼저 기반이 되는 지식은 정보 범용성·완결성 등을 고려해 객관적인 사실을 중심으로 꾸리고자 노력했다. 국립국어원의 '모두의 말뭉치'처럼 신뢰할 수 있는 출처에서 공유하는 자료들은 고품질로 분류해 데이터에 추가했다.

모두의 말뭉치 이용 현황

언어 인공지능 관련 산업계 및 학계 등에 총 12,599건 제공

(2020. 8. 25. ~ 2021. 3. 31.)





문화체육관광부 연구직 공무원

연구직 공무원 소속 기관

박물관, 미술관, 국악원, ...

국립국악원

국립국어원

국립민속박물관

국립아시아문화전당

국립중앙극장

국립중앙도서관

국립중앙박물관

국립한글박물관

국립현대미술관

대한민국역사박물관

2019년 국립국어원 학예연구직 경력경쟁채용시험 공고

2. 채용직급/인원/직무/근무예정지

직 급	직류(분야)	인 원	담당 예정 업무	근무예정지
학예연구사	국어 (국어 정보화)	**1명**	■ 국어 및 한국어의 언어 정보 처리 ■ 국어 정보 자료(말뭉치, 사전 등) 분석 및 조사·연구	국립국어원 (서울)

나. 선발 직무분야별 응시 자격요건 (최종시험(면접) 예정일 기준으로 인정)

임용예정 직급	직류 (분야)	선발 인원	응시자격 요건	
학예 연구사	국어 (국어 정보화)	1명	학위	○ 고등교육법에 의해 설치된 국내·외 대학(대학원 포함)의 관련 분야 석사학위 소지자 ※ [관련학과] 국어국문학, 언어학, 국어(언어)정보학, 국어교육학, 한국어교육학
			경력	○ 임용예정 직렬 또는 직위의 업무 내용과 같거나 유사한 분야에서 공무원 임용시험령 별표 9의 구분에 따라 임용예정 계급 상당 경력이 3년 이상인 자(연구직규정 7조2 3항) ※ [관련분야] (한)국어 정책 관련 조사 및 연구, (한)국어 정보 처리 관련 조사·분석 및 연구

2019년 국립국어원 학예연구직 경력경쟁채용시험 공고

다. 우대요건 (원서접수 마감일 기준으로 인정)

<table>
<tr><th>임용예정
직급</th><th>직류
(분야)</th><th>우대요건</th></tr>
<tr><td>학예
연구사</td><td>공통</td><td>1. 응시자격 요건을 초과한 직무분야 근무경력
2. 응시자격 요건을 초과한 직무분야 학위
3. 직무분야와 관련된 연구 논문
4. 직무분야와 관련된 정부 및 공공기관 수상 실적(상훈)
5. 국어 능력(공인 시험 점수)
- KBS한국어능력시험 2+급 이상, 국어능력인증시험 2급(169점) 이상</td></tr>
<tr><td></td><td></td><td>6. 외국어 능력(공인 시험 점수)
- 외국어시험 성적이 아래 기준 이상인 자
<table>
<tr><th colspan="2">시험의 종류</th><th>기 준 점 수</th></tr>
<tr><td rowspan="4">영어</td><td>토플(TOEFL)</td><td>CBT(220점), IBT(83점)</td></tr>
<tr><td>토익(TOEIC)</td><td>775점</td></tr>
<tr><td>텝스(TEPS)</td><td>700점</td></tr>
<tr><td>지텔프(G-TELP)</td><td>Level 2의 77점 이상</td></tr>
<tr><td rowspan="2">일본어</td><td>JPT</td><td>740점</td></tr>
<tr><td>JLPT</td><td>N2 150점</td></tr>
<tr><td>중국어</td><td>HSK</td><td>5급(210점)</td></tr>
</table></td></tr>
</table>

감사합니다!

saetbyol.s@gmail.com