





요즘 우리는 인공지능에게 날씨를 물어보고, 박물관 전시 안내를 부탁할 수도 있어요!
인공지능이 우리의 말을 어떻게 알아듣고 소통할 수 있는 걸까요?
인공지능이 우리말을 이해하고 구사할 수 있게 된 것은 바로 ‘말뭉치’ 때문입니다!
지금부터 저와 함께 인공지능 기술 속에 숨어 있는 비밀 자료 ‘말뭉치’에 대해서 알아보까요?

말뭉치란 무엇일까요?

실뭉치



말뭉치?

솜뭉치



‘말뭉치’는 언어학 용어인 코퍼스(corpus)를
우리말로 번역한 말이에요.
‘실뭉치’, ‘솜뭉치’에서 ‘뭉치’는
‘한데 뭉치거나 말거나 감은 덩이’를 말하죠.
그렇다면 ‘말뭉치’는 말[언어]을
한데 모아 놓은 덩어리가 되겠네요.

말은 왜 모을까요?

언어에 어떤 원리가 있는지, 사람들이 어떻게 언어를 쓰는지 알려면 어떻게 해야 할까요?

뇌를 들여다봐야 할까요? 사실, 뇌를 열어 봐도 알 수 없어요.

언어를 연구하기 위해서는 우리가 말하거나 글로 써 놓은 언어의 일부를 분석해서 전체 언어의 모습을 유추할 수밖에 없지요.

이렇게 언어가 어떻게 쓰이는지를 분석하기 위해 언어 자료의 일부를 표본으로 모아 놓은 것이 말뭉치랍니다.

말을 모아두기만 하면 될까요?

신문을 모으면 신문 말뭉치, 소설을 모으면 소설 말뭉치, 대화를 모으면 대화 말뭉치라고 할 수 있겠죠. 그런데 보통 쌓아둔 종이 신문이나 책꽂이에 꽂힌 소설책들을 말뭉치라고 하지는 않아요. 단순히 모으기만해서는 활용할 수 없거든요.

말뭉치는 컴퓨터로 분석하고 처리할 수 있도록 입력되어 있어야 합니다.

다시 말하면 말뭉치는
컴퓨터가 읽을 수 있도록
다양한 분야의 언어 자료를
모아 놓은 것이라 할 수 있어요.



말뭉치는 어떻게 생겼을까요?

```

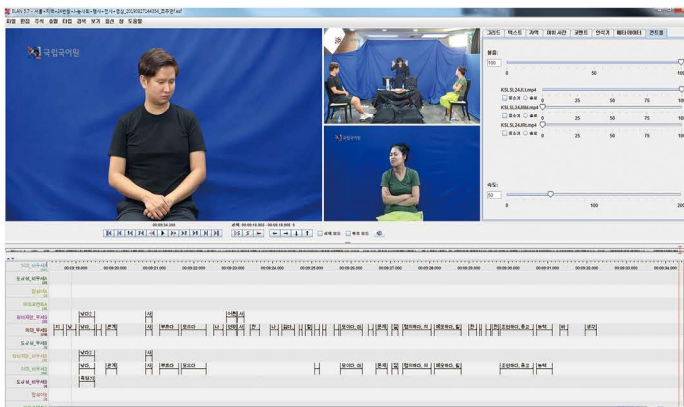
{id: "NWRW2000000002",
"metadata": {
  "title": "국립국어원 신문 말뭉치 NWRW2000000002",
  "creator": "국립국어원",
  "distributor": "국립국어원",
  "year": "2019",
  "category": "신문-한국 종합지",
  "annotation_level": "장문",
  "sampling": "본문 전체"
},
"document": {
  "id": "NWRW2000000002.1",
  "metadata": {
    "title": "한겨레 2019년 기사",
    "author": "강성원",
    "publisher": "한겨레",
    "date": "20190120",
    "topic": "문화",
    "original_topic": "문화,종교|문화,예술,문화재|문화,출판"
  },
  "paragraph": [
    {
      "id": "NWRW2000000002.1.1",
      "form": "우의 다·일 위와 나그네 한 겹 깨달을 때 희망의 시간 올 것"
    },
    {
      "id": "NWRW2000000002.1.2",
      "form": "첫 에세이 낸 최태형 신부"
    },
    {
      "id": "NWRW2000000002.1.3",
      "form": "조선이 내게 알려준 것들-(미항복) 천주교 의정부교구 최태형 신부가 최근 펴낸 책이다
    }
  ]
}

```

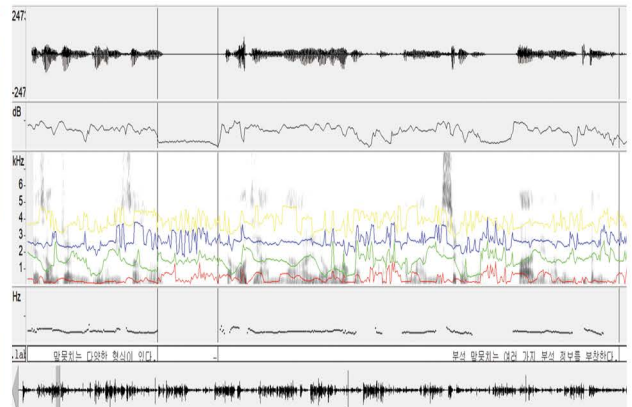
순서	장	절	문단	내용	순서	장	절	문단	내용		
97	3	4	6	1	OSCO에 가입했다고 알리고, 세계화를 외치고, 외국자본의 국미에 맞겨 금융 시스템...	128	3	4	6	5	The activities of the Korea Foundation go beyond the boundaries of ...
98	3	4	6	2	한국국제교류재단의 활동은 공공외교의 틀을 넘어 지구촌 요소에 한국과 세계...	129	3	4	6	6	they effectively create for Korea an interface with strategic points are...
99	4	1			KF 논단	130	4	1			KF Forum
100	4	2			멕시코와 한국, 40년간의 우정	131	4	2			Mexico and Korea, 40 Years of Friendship
101	4	3			Jose V. Borjon	132	4	3			Jose V. Borjon
102	4	4			국립 멕시코사절관 문화담당관	133	4	4			Cultural attache of the Embassy of Mexico
103	4	5			josevborjon@hotmail.com	134	4	5			josevborjon@hotmail.com
104	4	5	1	1	올해 1월 26일로 멕시코와 한국은 수교 40주년을 맞았다.	135	4	5	1	1	On January 26th, Mexico and Korea celebrated the 40th anniversary ...
105	4	5	1	2	이 날은 지난 40년간 경제, 문화, 과학 교류를 통해 우호 관계를 강화해 온 양국 모두...	136	4	5	1	2	This was a significant milestone for both countries because in the las...
106	4	5	2	1	한반도에 멕시코 문화가 처음 선보인 것은 멕시코의 고대문화 신교단인 잔도를 위해 ...	137	4	5	2	1	The first Mexican presence on the Korean peninsula can be traced ba...
107	4	5	2	2	1962년 11월 27일 최초의 멕시코인 선교사 루이스 루빈에 도착했고, 그후 수년간 ...	138	4	5	2	2	On November 27th, 1962 the first two Mexican missionaries arrived ...
108	4	5	3	1	한반도 문화가 처음 멕시코에 들어온 것은 1993년 5월 15일로 이 날은 멕시코 남서부 ...	139	4	5	3	1	The first presence of Korean culture in Mexico came with the arrival ...
109	4	5	3	2	1950년대 말까지 이 이민자들 중 100명 정도가 남아 남고 그들의 자손 1,000여 명 ...	140	4	5	3	2	By the end of the 1950s, it was believed that around 100 of the origi...
110	4	5	4	1	통한 문화유산을 가진 한국과 멕시코 두 나라간의 교류는 1962년 수교 이래 빠른 ...	141	4	5	4	1	With the establishment of diplomatic relations in 1962, these two co...
111	4	5	4	2	스포츠 분야에서는 60년대 말 한국의 현대화 교육이 멕시코에 음으로써 전국 곳곳에 ...	142	4	5	4	2	For example, in the area of sport, the arrival of Professor Moon Dai...
112	4	5	5	1	양국 정부가 외교관계를 수립하면서 가장 먼저 한 일은 문화협력협정을 체결한 것이었다.	143	4	5	5	1	One of the first actions both governments undertook when establish...
113	4	5	5	2	이 협정의 틀 안에서 문화교류 교류가 급격히 증가해 발전해왔다.	144	4	5	5	2	Within the framework of this agreement, cultural and educational int...
114	4	5	5	3	오늘날 양국간 다양한 전시, 영화제, 악단 교류 등을 통해 양국 문화가 볼 수 있으며, 2000...	145	4	5	5	3	Today it is common in both countries to see art exhibitions, film fest...
115	4	5	5	4	2001년 6월에는 전설의 멕시코 예술가인 알렉산더 가요의 한국에 이르기까지 ...	146	4	5	6	1	Notable was the establishment of a Korean gallery in Mexico's Natio...
116	4	5	6	1	국속분야에서는 양국 정부 양국 예술계의 도움으로 양국간의 한국인이 스페인, 라틴...	147	4	5	6	2	The exhibition of Mexican Contemporay art, inaugurated by Presid...
117	4	5	7	1	오늘 한국에는 멕시코 문화가 널리 알려져 있다.	148	4	5	7	1	In the era of education, thanks to the scholarships offered by both g...
118	4	5	7	2	한국인들은 멕시코를 통해 한 문화 유산과 축복한 문화적 정서감을 가진 나라로 인식...	149	4	5	7	2	Many Mexican students also come to Korea to study economic devel...
119	4	5	8	1	멕시코의 문화와 교육정책은 1996년 유네스코에 등록된 유산과 문화유산으로 ...	150	4	5	8	1	Mexico's rich cultural heritage with its unique cultural identity is vid...
120	4	5	8	2	이런 한국기업들의 발전에 도움을 준 요인으로서는 멕시코의 경제자유화와 북미자유무...	151	4	5	8	2	Thanks to the growing popularity of Mexican music and food, espec...
121	4	5	8	3	한자 한국은 멕시코의 5대 무역 상대국이며 투자규모로는 타겟지역의 국가들 중 2대...	152	4	5	9	1	Closer cooperative economic relations have developed since 1996 w...
122	4	5	9	1	양국의 경제관계 또한 우호적이다.	153	4	5	9	2	There has been an increasingly active presence of Korean enterprises ...
123	4	5	9	2	지난 10년간 두 양의 한국 다들들이 멕시코를 방문했으며 멕시코에서도 두 대통령...	154	4	5	9	3	An important element in the growth of Korean companies and visio...
124	4	5	9	3	지난 2001년 6월에는 전설의 멕시코 예술가인 알렉산더 가요의 한국에 이르기까지 ...	155	4	5	9	4	Currently, Korea is Mexico's 15th largest trading partner and the seco...

신문 말뭉치

병렬 말뭉치



수어 말뭉치



음성 말뭉치

말뭉치는 어디에 쓰일까요? - 사전 편찬



그러나 말뭉치를 활용하면서부터 많이 쓰이는 단어를
올림말로 결정하고 뜻을 객관적으로 풀이하며,
많이 사용하는 자연스러운 용례를 말뭉치에서 쉽게 찾고
분석해서 사전에 수록할 수 있게 되었어요.

말뭉치를 분석해서 우리 국어의 모습을 오롯이 담아낸
국어사전을 만들 수 있어요.

예전에는 사전을 편찬할 때 사전 편찬자의 직관에 의존하여
단어의 뜻을 기술하거나 용례도 사람이 일일이 수집했지요.



1980년대에
영국 버밍엄대가 콜린스 출판사와
2천만 어절 규모의 말뭉치를 구축하고
이것을 토대로 코빌드(COBUILD) 영어 사전을
편찬했어요. 우리나라의 <국립국어원 표준국어대사전>,
<연세 한국어 사전>, <고려대 한국어 대사전>도
모두 말뭉치를 활용했답니다.

말뭉치는 어디에 쓰일까요? - 언어 교육

말뭉치는 언어 교육에도 유용한 자료예요.

외국어를 가르칠 때 말뭉치를 활용하면 가장 많이 쓰는 자연스러운 표현들을 뽑아 가르칠 수 있어요.

외국어를 배우는 사람들의 언어를 수집해서 말뭉치를 만들기도 해요.

이 학습자 말뭉치를 활용하면 외국어를 배울 때 자주 틀리는 문법, 어휘, 표현 등을 분석할 수 있어서 효과적인 교수법을 찾거나 교재를 만드는 데 도움이 된답니다.

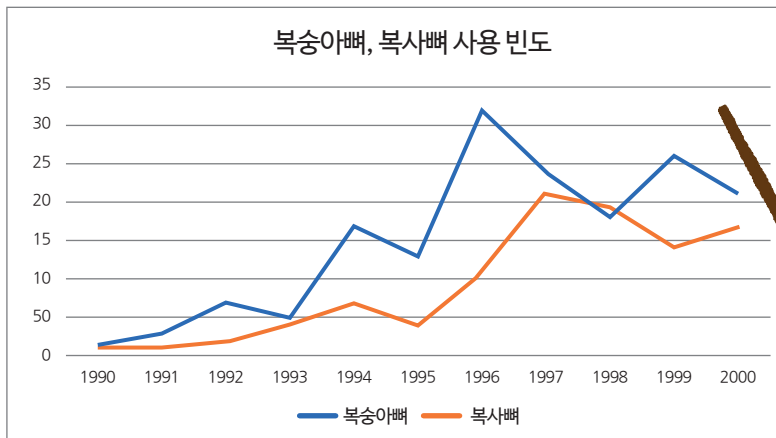


말뭉치는 어디에 쓰일까요? - 언어 연구 및 어문 정책 수립

말뭉치는 언어를 연구하고 어문 정책을 수립할 때에도 꼭 필요해요.

‘복숭아빠’, ‘복사빠’ 라는 말을 아세요?

말뭉치를 분석해서 두 단어의 사용 빈도를 측정해 보니 ‘복사빠’와 함께 ‘복숭아빠’도 많이 쓰인다는 것을 알 수 있었어요. 이러한 언어 현실을 반영하여 2011년에 두 단어 모두 표준어가 되었어요.



말뭉치를 분석해서 언제부터 그 단어가 쓰이기 시작했는지, 어느 때에 그 단어가 많이 쓰였는지, 시대에 따라 단어의 형태와 의미가 어떻게 변화되어 왔는지 등도 확인할 수 있습니다.

대학행

송리단길

혼행족

맛세권

객리단길

혼골족

소비 요정

2018년 신어
2019년 신어
2020년 신어

4차 산업 혁명 시대에 들어서면서 말뭉치는 기존의 언어 연구 분야 외에도 인공지능 기술 개발에 꼭 필요한 자원으로 관심이 모아지고 있어요.

인공 지능 스피커나 로봇 등에는 음성 인식, 언어 이해, 번역 등의 다양한 언어 처리 기술이 포함되는데, 이러한 언어 처리 기술은 최근에는 컴퓨터가 말뭉치를 학습하는 방식으로 발전하고 있어요.

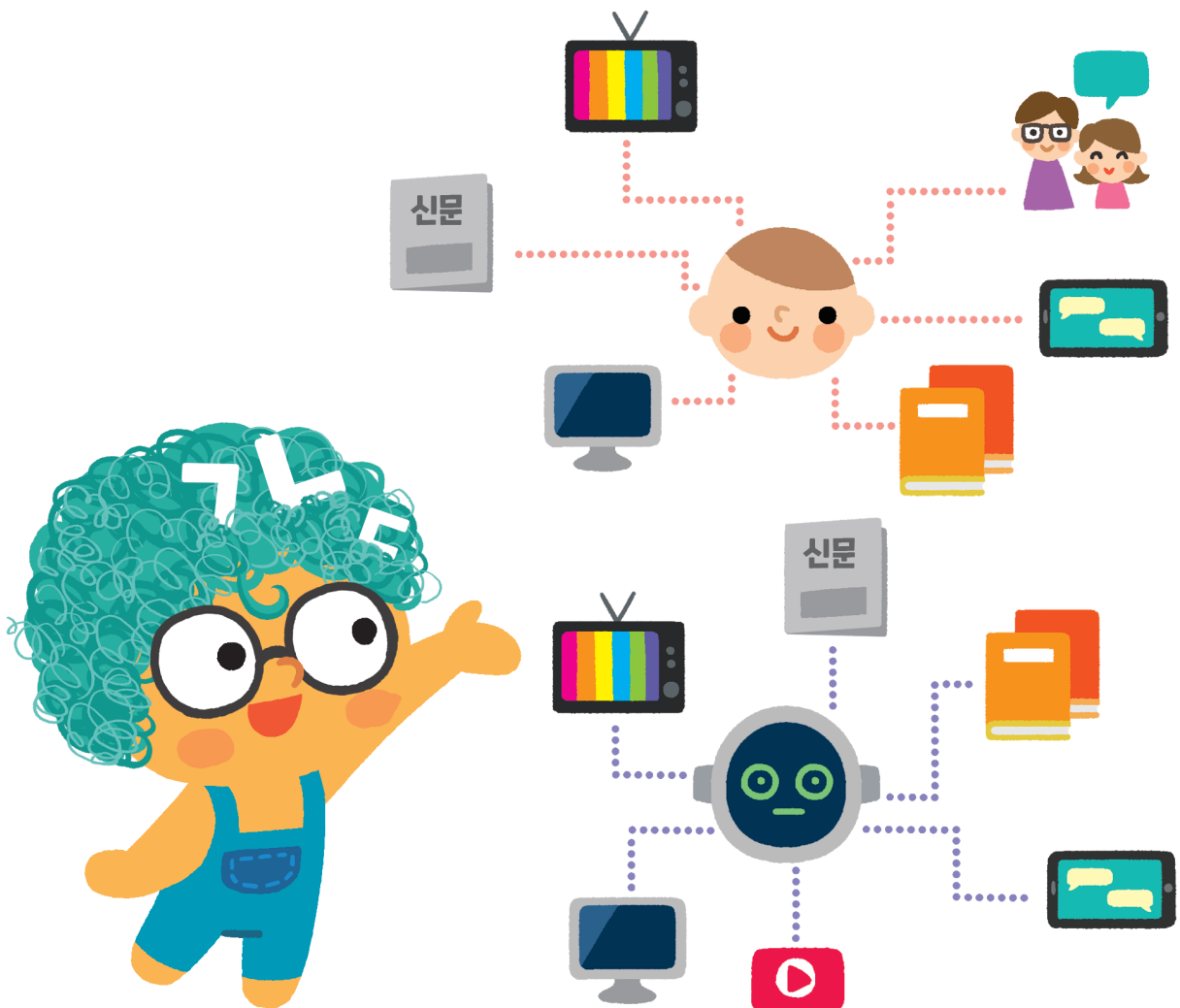


학습할 수 있는 말뭉치가 많을수록
컴퓨터가 인간의 말과 글을 제대로 이해하고
반응할 수 있기 때문에 인공지능의 발전을
위해서는 많은 양의 말뭉치가 필요합니다!

컴퓨터가 말을 학습한다고요?

세상에 태어나 한마디 말도 하지 못하던 아이들이 주변의 말과 글을 듣고 보면서
차차 말과 글을 배우고 그 말과 글 속에 담긴 의미를 이해하며 지식을 쌓아 나갑니다.

컴퓨터(인공 지능)가 아이처럼 언어를 듣고, 이해하고, 말하며, 정보를 찾아내려면 무엇이 필요할까요?
인간의 두뇌에 해당하는 알고리즘과 언어 학습 자료인 말뭉치가 필요합니다.



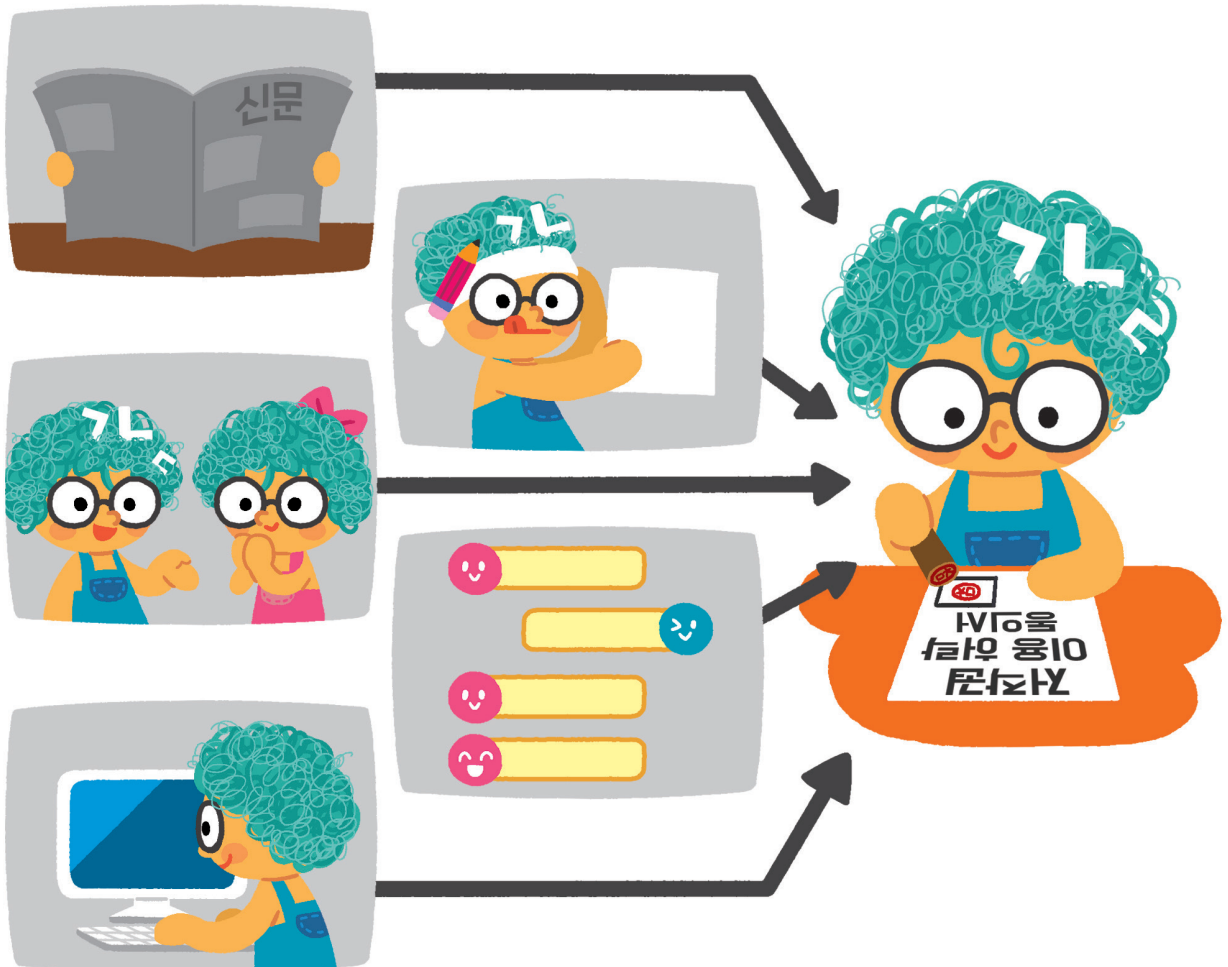
말뭉치는 어떻게 만들까요?

- 언어 자료 수집, 저작권 이용 허락

말뭉치를 만들려면 먼저 말뭉치의 재료가 될 언어 자료를 수집해야 합니다.

신문 기사, 책, 일상 대화, 메신저 대화, 블로그나 게시판의 글 등 다양한 언어 자료가 모두 말뭉치의 재료가 됩니다.

수집한 언어 자료를 말뭉치로 만들어 사용하려면 저작권자에게 이용 허락을 받아야 합니다.



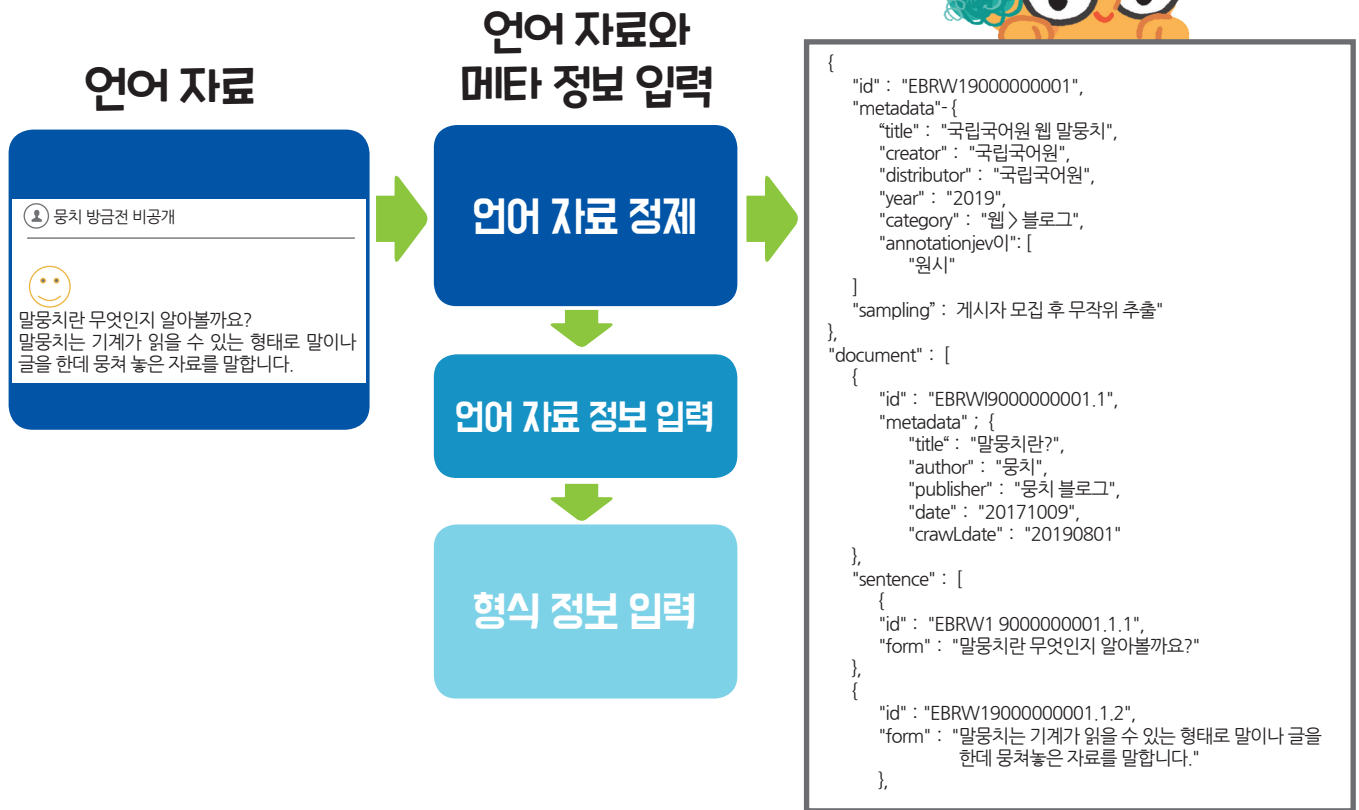
말뭉치는 어떻게 만들까요?

- 언어 자료와 메타 정보 입력

저작권 이용 허락 동의까지 마친 언어 자료는 컴퓨터가 읽을 수 있도록 입력해야 합니다.

일상 대화와 같이 글이 아닌 말로 된 언어 자료는 말을 글로 바꾸어 입력합니다.

여기에 언어 자료의 종류나 제목, 작성자, 출처 등 언어 자료의 특징이나 성격을 알려 주는 정보와 문단이나 문장 경계를 알려 주는 형식 정보 등을 컴퓨터가 읽을 수 있도록 입력하면 말뭉치의 기본이라고 할 수 있는 ‘원시 말뭉치’가 됩니다.



원시 말뭉치

말뭉치는 어떻게 만들까요?

- 분석 정보 입력

원시 말뭉치에 품사 정보, 의미 정보, 문장 구성 정보 등 여러 가지 분석 정보를 더하여 ‘분석 말뭉치’를 만듭니다.
이때 분석 정보는 컴퓨터가 읽을 수 있는 특별한 형식으로 덧붙입니다.

형태 분석 말뭉치

단어가 어떻게 구성되어 있는지, 그것의 품사 정보는 무엇인지에 대한 정보를 붙여 만듭니다.



눈꽃이 떨어졌어요 또 조금씩 떨어졌어요 보고 싶다 보고 싶다 얼마나
기다려야 내위야 널 보게 될까 만나게 될까
VV어요/EF 또/MAG 조금씩/MAG
눈꽃/NNG EC 싶/VX다/EF 보/VV고/EC 싶/VX
VV어야/EC 또/MAG 몇/MMA 밤/NNG
을/JKO 더/MAG 새우___001/VV어야/EC 너___001/NP 께 /
JKO 보___001/VV게/EC 되___016/VV 께 기/EF 만나___002/
VV게/EV 되___013/VV 께 기/EF
- 방탄소년단 ‘봄날’ -

어휘 의미 분석 말뭉치

형태는 같지만 의미가 다른 단어를 구별할 수 있도록 <우리말샘> 등 사전을 기준으로 의미 번호를 더하여 만듭니다.

눈꽃___001/NNG이/JKS 떨어지___001/VV어요/EF 또/MAG 조금씩/MAG 떨어지___001/VV어요/EF
보___001/VV고/EC 싶/VX다/EF 보___001/VV고/EC 싶/VX다/EF 얼마나/MAG
기다려___001/VV어야/EC 또/MAG 몇/MMA 밤___001/NNG
을/JKO 더/MAG 새우___001/VV어야/EC 너___001/NP 께 /
JKO 보___001/VV게/EC 되___016/VV 께 기/EF 만나___002/
VV게/EV 되___013/VV 께 기/EF
- 방탄소년단 ‘봄날’ -



개체명 분석 말뭉치

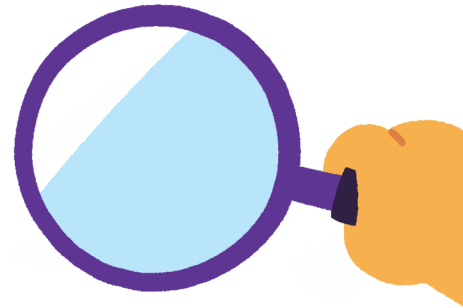
인명, 단체, 지명, 수량, 날짜 등 특정한 의미를 나타내는 단어나 어구에 대한 정보를 더하여 만듭니다.

그때는 나 어릴 때는 아무것도 몰랐네 그 다리 위를 건너가는 기분을 어디시냐고 어디냐고 여쭙보면

아버지/CV_RELATION는 항상 **양화대교/AF_BUILDING, 양화대교/AF_BUILDING**

이제 나는 서 있네 그 다리 위에 그 다리에

- 자이언티, '양화대교' -



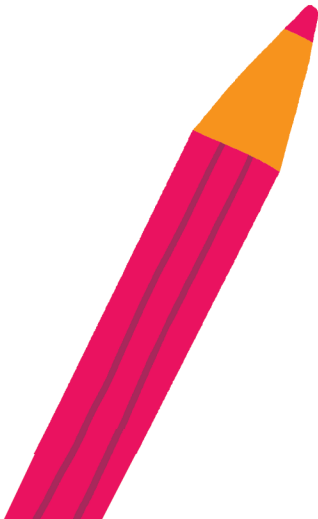
상호 참조 해결 말뭉치

하나의 글 안에서 같은 대상을 다른 표현으로 나타낸 것들을 찾아 서로 연결한 말뭉치를 말합니다.

그때는 나 어릴 때는 아무것도 몰랐네 그 다리 위를 건너가는 기분을 어디시냐고 어디냐고 여쭙보면

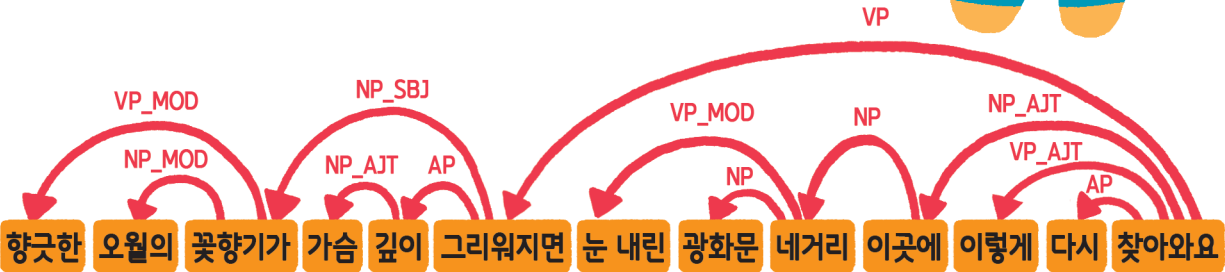
아버지는 항상 양화대교, 양화대교 이제 나는 서 있네 그 다리 위에 그 다리에

- 자이언티, '양화대교' -



구문 분석 말뭉치

문장을 분석하여 문장이 어떻게 구성되는지, 문장을 이루고 있는 단어들의 관계는 어떠한지에 대한 정보를 더해 만듭니다.



- 이문세, '광화문 연가' -

감성 분석 말뭉치

영화 관람 후기나 상품 평가 같은 글에서 내용이 긍정적인지 또는 부정적인지 등을 분석하거나 글을 쓴 사람이나 말하는 사람의 감정을 분석하여 만듭니다.
문서를 자동으로 분류하거나 어떤 문제에 대한 여러 사람의 의견을 파악할 때 사용될 수 있습니다.

대역 말뭉치

한 언어를 다른 언어로 번역하여 쌍으로 만든 말뭉치입니다.
자동 번역기를 만들기 위해서 필요합니다.

국립국어원에서는 소중한 언어 자원 '말뭉치'를 만듭니다.

원시 말뭉치 구축

국립국어원에서는 여러 가지 모습의 우리말을 담아내는 말뭉치를 만들고 있습니다.



신문 기사



책, 잡지, 기타 문서 등
신문 외의 단행본



TV, 라디오, 인터넷 방송 등의
강연, 발표, 토론, 인터뷰, 대화



메신저 대화 자료



일상 대화 녹음 자료



누리 소통망(SNS),
블로그, 웹 게시판



일기, 편지 등
일상 생활 글쓰기 자료

분석 말뭉치 구축

국립국어원에서는 우리말이 담고 있는 여러 가지 정보를 분석한 말뭉치를 만들고 있습니다.



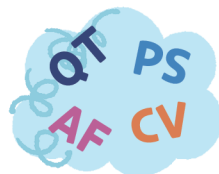
형태 분석 말뭉치



어휘 의미 분석 말뭉치



구문 분석 말뭉치



개체명 분석 말뭉치

당신의 말과 글, 소중한 언어 자원입니다.

발 행 일 2019년 12월 31일

발 행 인 국립국어원장

발 행 처 국립국어원 www.korean.go.kr

기 획 · 구 성 서혜진, 오은비, 이현주, 홍혜진

삽 화 최은영

디자인·인쇄 한국장애인상생복지회

이 책의 저작권은 국립국어원에 있습니다.



문화체육관광부
국립국어원