말뭉치 작성

말뭉치란? 언어를 연구하는 분야에서 필요로 하는 연구 재료로서 언어의 본질적인 모습을 총체적으로 드러내 보여줄 수 있는 자료의 집합 을 뜻한다.

일정 규모 이상의 크기를 갖추고 내용상으로 다양성과 균형성이 확보된 자료의 집합체로 정의하고 있으며 , 전산학적으로는 다음과 같이 정의하고 있다 .

1. 대규모 언어 데이터베이스
2. 인간의 음성 언어(문어,구어)를 대용량 컴퓨터에 저장하고 이를 필요에 따라 가공하여

언어 연구에 사용하는 것

1. 컴퓨터가 판독할 수 있는 형태로 저장된 자연어의 용례들과 이들 용례에 대한 부속 정보

말뭉치 종류

1. 원시 말뭉치 :텍스트를 컴퓨터 가독형 자료로 만들어 데이터베이스화한 말뭉치
2. 가공된 말뭉치:수집된 텍스트 데이터베이스를 형태소 분석이나 어휘, 품사 정보, 문헌 ,내용 등으로 분류할 수 있도록 인공적으로 가공한 말뭉치

2.작성 방법에 따른 종류

\*텍스트 여부에 따라

가)샘플 말뭉치:텍스트를 일정량만 수집한 것으로 텍스트의 내용이 조정된 말뭉치

나)모니터 말뭉치:변화하는 언어의 실태를 추적하기 위한 것으로 낡은 자료를 제외하고 항상 새로운 데이터베이스화한 말뭉치

\*연구 목적에 따라

가)범용 말뭉치

나)특수 목적 말뭉치

3.채취 샘플의 대상에 따라

가)공시 말뭉치 : 공시 자료를 대상으로 작성된 말뭉치

나)통시 말뭉치 : 통시 자료를 대상으로 작성된 말뭉치

4.언어 매체에 따라

가)문자언어 말뭉치 : 문자 언어에서 샘플을 채록하여 작성된 말뭉치

나)음성언어 말뭉치 : 음성 언어를 문자화하여 채록하여 작성된 말뭉치

5)기타

같은 내용을 2개 국어 이상 병렬시켜 만든 말뭉치로 병렬 말뭉치가 있다. 예를 들어 영어 소설의 원본과 한국어 번역본의 데이터로 말뭉치를 만들면 한영 소설어의 병렬 말뭉치가 된다.

원시 말뭉치 : 텍스트를 컴퓨터 가독횽 자료로 만들어 데이터베이스화한 말뭉치

가공된 말뭉치 : 수집된 텍스트 데이터베이스를 형태소 분석이나 어휘,품사 정보,문헌, 내용등으로

불류할 수 있도록 기준에 따라 가공한 말뭉치