2조 김성은, 박유빈, 임채현

우리나라 말뭉치 종류

1. 연세 한국어 말뭉치(1987): 「연세 한국어 사전」을 편찬하는 과정에서는, 1960년대로부터 1990년대 중반까지의 국내의 출판물과 일부의 구어 등을 중심으로 하여, 다양한 주제와 형태의 글을 전산으로 입력하였다. (연세대학교 언어정보연구원)

2. 고려대학교 한국어 말모둠(1995): 고려대학교 민족문화연구소에서는 1970-90년대 한국어의 1000만 어절 규모의 현대 한국어 말모둠인 ‘고려대학교 한국어 말모둠1(KOREA-Q Corpus)을 구축하였다. 이것은 분야별 균형을 고려한 ‘고른 말모둠’(balanced corpus)으로서, 국어학, 언어학 등 학문의 연구와 국어사전 편찬 등 응용의 기초가 되는 중요한 데이터베이스이다.

(고려대학교 민족문화연구소)

3. 국립국어원 말뭉치(1992): 1400년대 이후의 한국어(6800만 어절)

표준국어대사전을 만드는 데 활용(국립국어원)

-세종 말뭉치: 세종 말뭉치는 TEI(Text Encoding Initiative)의 부호화 방식에 따라 헤더(Header)와 본문(Text)으로 구성되며, 본문에는 언어구조화 방식에 따라 원시(문장, 어절), 형태(문장, 어절, 형태소/품사), 형태의미(문장, 어절, 형태소/품사/의미정보), 구문분석 말뭉치(문장, 어절, 형태소정보, 구문태그정보)로 구성된다.

  o 원시 말뭉치: 문장과 어절 정보만으로 구성된 말뭉치 형태 구조

  o 형태분석 말뭉치: 문장과 어절, 형태소, 품사의 조합으로 구성된 말뭉치 형태 구조

  o 형태의미분석 말뭉치: 문장과 어절, 형태소, 품사, 의미의 조합으로 구성된 말뭉치 형태 구조

  o 구문분석 말뭉치: 문장과 구문 태그, 형태소, 품사의 조합으로 구성된 말뭉치 형태 구조