21세기 세종계획 현대국어 기초말뭉치: 성과와 전망

김흥규[+] 강범모[++] 홍정하[+++]
고려대학교 국어국문학과[+]/언어학과[++]/민족문화연구원 전자텍스트연구소[+++]
{gardener[+]/bmkang[++]/kleist[+++]}@korea.ac.kr

21st Century Sejong Modern Korean Corpora: Results and Expectations

Hung-gyu Kim[+] Beom-mo Kang[++] Jungha Hong[+++]
Department of Korean Language and Literature[+]/Linguistics[++]
Center for Electronic Texts, Institute of Korean Culture[+++]
Korea University

요 약

현대국어 기초말뭉치는 방법론 및 표준화 연구, 그리고 소프트웨어 개발과 더불어 21세기 세종계획 국어기초자료 구축 사업의 일환으로 개발되었다. 현대국어 기초말뭉치 개발에서는 세종말뭉치 통합분 12,000만 어절을 후처리하고, 원시말뭉치 6,200만 어절, 형태분석 말뭉치 1,500만 어절, 형태의미분석 말뭉치 1,250만 어절, 구문분석 말뭉치 80만 어절을 신규 구축 완료하였으며, 이 중 일부 말뭉치에 대한 정제 작업이 2007년 말까지 완료될 예정이다. 방법론 및 표준화 연구에서는 말뭉치 구축 방법론과 분석표지 표준화, 말뭉치 활용 연구가 진행되었고, 이 밖에도 소프트웨어 개발 사업에서는 말뭉치 구축 및 활용에 필요한 도구를 개발하였다. 이 논문은 21세기 세종계획 국어기초자료 구축 사업의 연구 성과를 현대국어 기초말뭉치를 중심으로 소개하고 향후 전망을 논의하는 것이 목적이다.

1. 들어가기

현대국어 기초말뭉치는 선진 정보문화의 기본 바탕과 자원을 확충하는 국어정보화 중장기 발전계획인 21세기 세종계획 국어기초자료 구축 사업의 일환으로 지난 10년간(1998년-2007년) 개발되어 왔다. 국어기초자료 구축 사업은 현대국어 기초말뭉치 개발뿐만 아니라, 말뭉치 구축 및 활용에 필요한 방법론 및 표준화 연구, 그리고 소프트웨어 개발의 세 가지 세부 연구 사업으로 추진되어 왔다.

이 논문은 21세기 세종계획 국어기초자료 구축 사업의 연구 성과를 현대국어 기초말뭉치를 중심으로 소개하고 향후 전망을 논의하는 것이 목적이다. 이를 위해 이 논문에서는 먼저 현대국어 기초말뭉치에 대해 기술하고, 말뭉치 구축과 활용에 필요한 방법론 및 표준화 연구, 그리고 소프트웨어 개발 성과를 소개한다. 마지막으로 본 연구의 성과를 토대로 하여 향후 전망을 논의한다.

2. 현대국어 기초 말뭉치 개발

국어 연구의 과학적 토대를 구축하고 국어 특성에 적합한 정보 문화와 새로운 정보처리 이론 및 기술을 발전시키기 위해서는 그 기반으로 대규모 국어 텍스트 말뭉치가 필수적이다. 현대국어 기초말뭉치는 이런 목적에 적합하도록 말뭉치 구성 및 규모, 그리고 부착 정보가 고려되어 설계되었으며, 그 세부 말뭉치로는 세종말뭉치 통합분, 원시말뭉치, 형태분석 말뭉치, 형태의미분석 말뭉치, 구문분석 말뭉치가 있다. 여기서는 이들 말뭉치의 구성 및 규모, 부착 정보에 대해 제시한다.

2.1. 개요

종류	규모(단위: 만 어절)
세종말뭉치 통합분	12,000
원시	6,200
형태분석	1,500
형태의미분석	1,250
구문분석	80

[표 1] 현대국어 기초 말뭉치 종류 및 규모

현대국어 기초 말뭉치는 준구어를 포함하여 현대국어 문어 텍스트를 대상으로 구축되었다. 국어 연구의 과학적 토대 구축 및 관련 정보처리 이론 및 기술 발전을 고려하여 세종말뭉치 통합분을 포함하여 원시 말뭉치, 형태분석 말뭉치, 형태의미 분석 말뭉치, 구문분석 말뭉치로 구성되어 있으며, 그 규모는 [표 1]과 같다.

세종말뭉치 통합분은 표준화 지침[1]에 따라 21세기 세종계획 사업 시행 이전 국비 지원에 의해 축적된 말뭉치를 대상으로 후처리한 말뭉치이다. 이 통합분에 포함된 말뭉치는 <국어정보처리기반 구축> 사업(1994-1997) 말뭉치 약 7천만 어절과 국립국어연구원 구축 말뭉치 약 5천만 어절이다.[1)]

세종말뭉치 통합분 외의 말뭉치들은 21세기 세종계획 사업에서 신규로 구축된 말뭉치이다. 원시 말뭉치는 모든 분석 말뭉치의 분석 대상이 되는 원본 텍스트이며, 이를 기반으로 하여 언어 주석 정보가 부착된 것이 분석 말뭉치이다. 분석 말뭉치 중에서도 형태분석 말뭉치는 형태의미분석 말뭉치와 구문분석 말뭉치의 분석 대상이 되는 핵심적 말뭉치이다.

현대국어 기초 말뭉치는 [1]에서 제시한 장르 별 구성을 고려하여 구성되었다. [표 2]는 원시, 형태분석, 형태의미분석, 구문분석 말뭉치의 장르 별 분포이다.

말뭉치 / 장르	원시	형태	형태의미	구문
준구어[2)]	5%			
신문	29%	24%	28%	22%
잡지	13%	12%	15%	10%
책-정보	33%	38%	39%	40%
책-상상	17%	24%	16%	25%
기타	3%	2%	2%	3%
합계	100%	100%	100%	100%

[표 2] 현대국어 기초 말뭉치의 장르 별 분포

2.2. 원시 말뭉치

원시 말뭉치는 [1]의 원시 말뭉치 구축 기본 원칙에 따라 구축되었다.

(1) 원시 말뭉치 구축 기본 원칙
- ■ 원문유지
  문헌의 경우 원문을 그대로 유지함을 원칙으로 한다(한자, 기호 등).
- ■ 문서 정보 표기 방식
  SGML/TEI를 수용하여 문서 정보를 표기한다.
- ■ 헤더
  헤더에 기본적인 정보는 자세히 표기한다.
- ■ 본문 마크업
  본문에는 최소의 본문 마크업만을 부착한다.

(1)에서 헤더는 저자, 텍스트 제목 및 분류[3)], 말뭉치 구축 절차 및 수정 과정 등 가능한 자세한 정보를 표시한다. 이것은 각 텍스트를 확인 및 구별하는 데 유용하고 그 활용도를 높이는 중요한 단서가 되기 때문이다. 반면 본문 마크업의 경우, 대용량 말뭉치의 효율적 구축을 위하여 텍스트 본문 정보 식별에 필수적인 최소의 마크업만을 수용하였다.

2.3. 형태분석 말뭉치

형태분석 말뭉치는 하나의 어절을 대상으로 분석하되, '형태소' 차원이 아닌 '형태' 차원의 분석이다. 또한 원본 텍스트인 원시 말뭉치를 가급적 훼손하지 않으며, 국립국어원에서 출간된 '표준국어대사전'의 표제어를 기준으로 분석하였다.

[표 3]은 형태분석 말뭉치의 어절 분석의 기준이 되는 형태 분석 표지이다. 동사, 형용사 등은 학교 문법의 품사 분류 정도의 것이나, 일반 명사, 고유 명사 등은 좀 더 세분화되었고, 특히 관계언과 의존형태는 이보다 더 세분되었다.

대분류	소분류	세분류
체언	명사	일반명사NNG 고유명사NNP 의존명사NNB
	대명사	대명사NP
	수사	수사NR
용언	동사	동사VV
	형용사	형용사VA
	보조용언	보조용언VX
	지정사	긍정지정사VCP 부정지정사VCN
수식언	관형사	관형사MM
	부사	일반부사MAG
		접속부사MAJ
독립언	감탄사	감탄사IC
관계언	격조사	주격조사JKS
		보격조사JKC
		관형격조사JKG
		목적격조사JKO
		부사격조사JKB
		호격조사JKV
		인용격조사JKQ
	보조사	보조사JX
	접속조사	접속조사JC
의존형태	어미	선어말어미EP
		종결어미EF
		연결어미EC
		명사형전성어미ETN
		관형형전성어미ETM
	접두사	체언접두사XPN
	접미사	명사파생접미사XSN
		동사파생접미사XSV
		형용사파생접미사XSA
	어근	어근XR

1) 세종말뭉치 통합본에 대해서는 이 논문에서는 자세히 다루지는 않겠다. 자세한 사항은 [1]을 참조.
2) 준구어는 문어와 달리 발화를 전제로 한 텍스트로 주로 드라마, 연극, 영화 등의 상연, 상영을 위해 준비된 대본이다.
3) 텍스트 분류는 [2]의 분류를 세분화한 것이다.

대분류	소분류	세분류
기호	마침표, 물음표, 느낌표	SF
	쉼표, 가운뎃점, 콜론, 빗금	SP
	따옴표, 괄호표, 줄표	SS
	줄임표	SE
	붙임표(물결, 숨김, 빠짐)	SO
	외국어	SL
	한자	SH
	기타 기호(논리 수학기호, 화폐 기호 등)	SW
	명사추정범주	NF
	용언추정범주	NV
	숫자	SN
	분석불능범주	NA

[표 3] 형태 분석 표지

(2)는 "생산비 절감을 위해 노력해왔으나"를 형태분석하여 [표 3]의 형태 분석 표지를 부착한 예이다.

(2) 형태분석 예

생산비	생산비/NNG
절감을	절감/NNG + 을/JKO
위해	위하/VV + 아/EC
노력해왔으나	노력/NNG + 하/XSV + 아/EC + 오/VX + 았/EP + 으나/EC

2.4. 형태의미분석 말뭉치 특성

형태의미분석 말뭉치는 표준국어대사전에 동음이의어로 등재된 형태에 부가적 의미표지를 부착한 것이다. 형태분석 말뭉치의 분석 표지 중 일반명사, 의존명사, 동사, 형용사, 관형사, 일반부사를 형태의미 분석 대상으로 하며, 그 의미표지는 표준국어대사전[4]의 의미어깨번호와 동일하게 표지된다.

예를 들어 표제어 "은행"은 표준국어대사전에 다음과 같이 세 가지 의미로 구분되어 있다.

(3) 표제어 "은행"의 세 가지 의미

가. 은행1(恩倖)
임금이 총애하여 가까이 두는 신하.

나. 은행2(銀行)
예금을 받아 그 돈을 자금으로 하여 대출, 어음 거래, 증권의 인수 따위를 업무로 하는 금융 기관.

다. 은행3(銀杏)
은행나무의 열매. 식용하거나 약용한다.

형태의미분석 말뭉치는 다음과 같이 형태분석된 "은행"에 '__'와 표준국어대사전의 의미어깨번호와 일치하는 두 자리 숫자를 부착한다.

은행/NNG → 은행__02/NNG

이 밖에도 형태분석 말뭉치와 표준국어대사전의 품사 분류가 일치하지 않아 그 분석이 달라진 경우, 어깨번호 앞에 'x' 표지를 부착하고, 표준국어대사전에 등재되지 않은 어휘나 의미가 나타날 경우 그 어깨번호는 '88'로 처리한다.

2.5. 구문분석 말뭉치

구문분석 말뭉치는 형태분석된 말뭉치에 구문정보를 부착하고 구문구조를 분석한 말뭉치이다. 구문분석 말뭉치는 총 826,127어절, 77,693 문장이며, 평균 문장당 어절수 12.32 어절이다.

[그림 1]은 구문분석 말뭉치의 문장당 어절수 분포이다. 구문분석 말뭉치는 몇 어절로 구성된 간단한 문장뿐 아니라 구문구조의 복잡성을 나타내는 20어절 이상의 문장도 비교적 고르게 분포되어 있다.

[그림 1] 구문분석 말뭉치의 문장당 어절수 분포

구문분석 말뭉치는 (4)에서 제시된 기준에 따라 구축되었으며, 그 기본 원칙은 다음과 같다.[5]

(4) 구문분석 말뭉치 구축 기본 원칙

가. 자연언어처리에서 일반적으로 고려되는 일관성 유지와 효율성 제고에 초점을 두되, 일반언어학적 관점에서도 크게 벗어나지 않도록 한다.
나. 표층 구조를 중시하여 분석한다.
다. 이분지 가설을 취하며 다분지를 허용하지 않는다.
라. 공범주를 인정하지 않는다.
마. 어절을 분석의 기본 단위로 한다.
바. 보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.
사. 원칙적으로 접속과 내포를 구별하지 않으며 접속절은 모두 부사절로 분석한다.(다만, 명사구 접속만은 인정한다.)
아. 하나의 주어가 모문과 내포문 모두에 관

4) 표준국어대사전 v 1.0을 기준으로 형태의미 표지가 부착되었다.

5) 자세한 구문 구조 분석 기준에 대해서는 [3]을 참고

련되어 있을 때 모문의 주어로 우선 분석한다.

(5)는 구문분석 말뭉치에서 구문정보를 표상하는 구문분석 표지이다. 구문분석 말뭉치의 분석 표지는 크게 구문표지와 기능표지로 구분되며, 전자는 명사구, 용언구 등의 구문범주를, 후자는 주어, 목적어 등의 문법관계를 의미한다. (5-가)는 구문표지와 해당하는 범주와 사례를, (5-나)는 기능표지와 해당 범주와 사례를 나타내고 있다.

(5) 구문분석 표지

가. 구문표지

표지	범주	사례
S	문장	
Q	인용절	인용부호(“”) 안에 들어 있는 두 개 이상의 문장
NP	체언구	체언(명사, 대명사, 수사)
VP	용언구	용언(동사, 형용사, 보조용언)
VNP	긍정 지정사구	긍정 지정사 ‘이다’와 결합한 구
AP	부사구	부사
DP	관형사구	관형사
IP	감탄사구	감탄사
X	의사 구(pseudo phrase)	인용부호와 괄호를 제외한 나머지 부호나, 조사, 어미가 단독으로 어절을 이룰 때 그 구문표지 위치에 표시(예: [X_CMP])
L, R	부호	인용부호나 괄호의 구문표지에 표시로 왼쪽 부호에는 L을, 오른쪽 부호에는 R을 표시
Q, U, W, Y, Z	인용절	인용부호(“”)에 이끌려 나온 두 개 이상의 인용절을 대신하여 표기되는 부호

나. 기능 표지

표지	범주	사례
SBJ	주어	주격 체언구, 명사 전성 용언구, 명사절 (NP_SBJ, VP_SBJ, S_SBJ, VNP_SBJ)
OBJ	목적어	목적격 체언구, 명사 전성 용언구, 명사절 (NP_OBJ, VP_OBJ, S_OBJ, VNP_OBJ)
CMP	보어	보격 체언구, 명사 전성 용언구, 인용절 (NP_CMP, VP_CMP, S_CMP, VNP_CMP)
MOD	체언 수식어	관형격 체언구, 관형형 용언구, 관형절 (NP_MOD, VP_MOD, S_MOD, VNP_MOD)
AJT	용언 수식어	부사격 체언구, 문말어미+부사격조사 (NP_AJT VP_AJT, S_AJT, VNP_AJT)
CNJ	접속어	접속격 체언(NP_CNJ, VNP_CNJ)
INT	독립어	체언 (NP_INT)
PRN	삽입어구	삽입된 성분의 기능표지에 표시(예: [NP_PRN])

(6)은 문장 "그는 배밭을 향해 남쪽으로 출발했다."에 (5)의 구문분석 표지를 부착하고 구문분석한 예이다.

(6) 구문분석의 예

```
(S (NP_SBJ 그/NP + 는/JX)
   (VP (VP (NP_OBJ 배/NNG + 밭/NNG + 을/JKO)
           (VP 향하/VV + 아/EC))
       (VP (NP_AJT 남쪽/NNG + 으로/JKB)
           (VP 출발/NNG + 하/XSV
                + 았/EP + 다/EF + ./SF))))
```

2.6. 말뭉치 정제

1998년부터 2006년에 걸쳐 신규 구축된 현대국어 기초 말뭉치의 규모는 원시 말뭉치 6,200만 어절, 형태분석 말뭉치 1,500만 어절, 형태의미분석 말뭉치 1,250만 어절, 구문분석 말뭉치 80만 어절에 달한다. 그러나 오랜 기간에 걸쳐 대규모로 구축된 만큼 말뭉치의 일관성 확보가 그 무엇보다 중요하다 하겠다. 특히 그 사이 효율적인 말뭉치 구축을 위한 말뭉치 구축 지침이 여러 차례 수정되었고, 이에 따른 구축년도 별 차이를 일관되게 정제하였다.

정제 대상이 된 말뭉치는 원본 텍스트가 동일한 원시, 형태분석, 형태의미분석 말뭉치이며 그 규모는 다음과 같다.[6]

종류	규모(단위: 만 어절)
원시	500
형태분석	1,250
형태의미분석	1,250

[표 4] 정제 대상 말뭉치 규모

말뭉치 정제는 일관성 확보뿐만 아니라 형식적 오류 수정, 그리고 기존 원시말뭉치에만 부착되었던 <head>, <p> 등의 본문 마크업을 추가로 부착하였다. 이는 원본 텍스트가 동일한 원시-형태분석-형태의미분석 말뭉치 사이의 대응성 확보를 통한 다양한 말뭉치의 통합적 활용 가능성을 증대한 것이다.

3. 방법론 및 표준화 연구

21세기 세종계획 국어기초자료 구축 사업에서는 현대

6) 정제 대상 형태분석 및 형태의미분석 말뭉치 규모에 비해 원시 말뭉치 규모가 750만 어절 적은 이유는 2007년도 말뭉치 정제 사업에서 원시 말뭉치가 제외되었기 때문이다.

국어 기초 말뭉치 구축 외에도 방법론 및 표준화 연구도 수행되었다. 이 절에서는 국어기초자료 구축 사업에서 수행된 연구를 개관한다.

3.1. 말뭉치 구축 방법론 및 표지 표준화 연구

말뭉치 구축 방법론 및 표지 표준화 연구로 다음의 연구들이 수행되었다.

(7) 말뭉치 구축 방법론 및 표지 표준화를 위한 세부 연구과제
- 말뭉치 구축 방법론에 대한 연구
- 병렬 말뭉치 구축을 위한 기초 연구
- 한국어 정보처리를 위한 어절분석표지 표준화 연구
- 구문분석 방법론 및 표지의 권장 표준안 연구

먼저 "말뭉치 구축 방법론에 대한 연구"[1]에서는 크게 세 가지 주제가 연구되었다. 첫째, 말뭉치 규모의 적정성 연구에서는 언어 분석 및 정보화 기반 기술 개발에 있어 가장 효율적으로 그 결과를 제공해 줄 수 있는 대규모 말뭉치의 적정 규모 설정에 관한 연구가 수행되었다. 둘째, 말뭉치 구성의 변인 연구에서는 문헌의 장르나 주제, 시대 등이 말뭉치 구성을 결정하는 변인으로 충분한 조건인지를, 그리고 그 밖의 문체, 작가의 출신, 성별, 기타의 언어적, 언어외적 제반 변인들을 검증하였다. 셋째, 말뭉치 구성의 적정성 검증의 문제에서는 균형말뭉치라 할 때의 '균형'과 '특성'의 문제를 중심으로, 기존 대규모 말뭉치를 대상으로 하여 검증할 수 있는 방법론을 모색하였다.

"병렬 말뭉치 구축을 위한 기초 연구"[1]에서는 국어-영어, 국어-일어의 병렬 말뭉치 구축을 위한 기초연구와 병렬 말뭉치 구축 방법론 및 모형을 개발하는 것을 목적으로 하였다. 이 연구에서는 국내외 병렬 말뭉치 연구 동향을 파악하고, 병렬 말뭉치 부호화 기준 연구 및 개별언어의 특성과 언어비교 연구, 그리고 병렬 말뭉치 구축의 기본적 방법론을 연구하였다.[7)]

"한국어 정보처리를 위한 어절 분석표지 표준화 연구"[1]에서는 한국어 정보 처리를 위하여 한국어 문장이나 어구에 적용할 수 있는 어절 분석표지 집합의 표준안을 마련하는 것을 목적으로 하였다. 이 연구는 기존 어절 분석표지 집합을 수집 및 검토하였으며, 그 결과를 최대한 반영하고 한글 맞춤법이나 학교 문법과 같은 어문 규범을 최대한 준수하는 입장에서 한국어 문장이나 어절 단위에 적용할 수 있는 품사-형태 분석표지 집합의 표준안을 마련하였다.[8)]

마지막으로 "구문분석 방법론 및 표지의 권장 표준안 연구"[4]에서는 국어 문장 분석에 설정되어야 할 기본적인 통사 단위와 그것이 결합되어 형성되는 통사 구성을 범주화하고 체계화하는 목적으로 수행되었다. 이 연구는 기존 국어 문법 연구의 성과와 자연언어처리를 위해 도입된 구문분석 표지를 참고하여, 구문 분석 방법과 분석 표지 작성의 기본 원리를 수립하고, 국어 문장 분석에 필요한 분석 표지의 표준적인 집합을 마련하였다.[9)]

3.2. 말뭉치 활용 연구

현대국어 기초 말뭉치 중 원시 말뭉치와 형태분석 말뭉치를 활용한 연구는 다음과 같다.

(8) 원시 및 형태분석 말뭉치를 활용한 연구
- 형태분석 말뭉치를 이용한 국어 문법 현상의 계량적 연구
- 분석 말뭉치를 이용한 한국어 형태소 연접 관계 연구
- 국어 연구에서의 말뭉치 활용방법 연구

먼저 "형태분석 말뭉치를 이용한 국어 문법 현상의 계량적 연구"[5]에서는 형태분석 말뭉치 150만 어절에 나타나는 국어 문법 현상을 계량화하여 그 결과를 제시하였다. 여기서 제시된 결과는 문장 성분, 또는 문장 성분들이 상호 관련된 문법 현상, 조사 및 어미와 관련된 현상, 그리고 접미사와 관련된 현상 등이다.

"분석 말뭉치를 이용한 한국어 형태소 연접관계 연구"[6]에서는 350만 어절 형태분석 말뭉치를 토대로 어절 내부의 형태소 연접관계와 어절 사이의 형태소 연접관계에서 추출의 가치가 있는 문법 현상을 계량적으로 제시하였다. 체언 어절, 용언 어절, 수식언 어절, 기타 어절의 구성 양상, 그리고 형태소의 결합 양상이 이 연구를 통해 추출되었다.

마지막으로 "국어 연구에서의 말뭉치 활용방법 연구"[7]에서는 원시 말뭉치 및 형태소분석 말뭉치의 특성을 개관하고 각 말뭉치의 특성을 고려한 활용방법을 제시하였다. 이 연구에서는 기존 검색 프로그램을 활용하는 방법 및 이를 이용한 추출 자료를 제시하였다.

4. 소프트웨어 개발

방법론 및 표준화 연구 외에도 말뭉치 활용 및 구축에 필요한 소프트웨어가 개발되었다. 먼저 말뭉치 활용을 위한 소프트웨어는 다음과 같다.

(9) 말뭉치 활용을 위한 소프트웨어
- 각종 기초 도구 및 용례추출기
- 지능형 형태소 분석기
- 말뭉치 통합 응용 시스템(글잡이II)

---

7) 병렬 말뭉치는 2000년부터 특수 말뭉치로 이관되어 구축되었다.

8) 이 연구 결과를 토대로 2.3절 형태분석 말뭉치의 형태 분석 표지가 확정되었다.

9) 이 연구를 토대로 2.5절 구문분석 말뭉치의 구문분석 표지 및 구문 구조 분석 방법이 확정되었다.

· 현대국어 말뭉치 활용 시스템 개발
(한마루)

"각종 기초 도구 및 용례추출기"[1]는 원시 말뭉치에서 용례를 추출하고 유용한 통계 정보를 추출할 수 있는 도스용 소프트웨어이다. "지능형 형태소 분석기"[4]는 일반 국어 텍스트를 자동으로 형태소 분석해주는 도구이며, "말뭉치 통합 응용 시스템"[5]는 원시 말뭉치와 형태분석 말뭉치에서 용례를 검색하고 유용한 통계 정보를 추출하는 윈도우용 도구이다. 마지막으로 "현대국어 말뭉치 활용 시스템"[10]은 모든 종류의 현대국어 기초 말뭉치, 즉, 원시, 형태분석, 형태의미분석, 구문분석 말뭉치에서 유용한 정보를 검색 및 통계 추출하는 통합적 도구이다.

한편, 말뭉치 구축에 필요한 모든 소프트웨어는 구문분석 말뭉치 구축을 위해 개발되었으며, 그 종류는 다음과 같다.

(10) 구문분석 말뭉치 구축을 위한 소프트웨어
· 구문 태그 부착 말뭉치 구축도구
· 반자동 구문분석 말뭉치 구축도구
· 구문분석 말뭉치 종합 관리 도구
· 지능형 구문분석 도구

"구문 태그 부착 말뭉치 구축도구"[6]는 형태소 분석된 한국어 문장에 대한 구문 표지를 부착하여 구문분석 말뭉치를 만들고 관리하는 작업을 지원하는 도구로, "반자동 구문분석 말뭉치 구축도구"[7]는 일부의 구문분석 단계를 자동화 시킨 도구로, "구문분석 말뭉치 종합 관리 도구"[8]은 구문분석 말뭉치를 통합적으로 검색하여 수정할 수 있는 도구로, "지능형 구문분석 도구"[9]는 최대한의 구문분석 단계를 자동화시켜 수작업을 최소화시킨 도구로 개발되었다.

5. 전망

지난 10년간에 걸쳐 구축된 현대국어 기초 말뭉치는 BNC 말뭉치[11]에 비견할 만한 국어 말뭉치로 손색이 없을 뿐만 아니라, 언어학적 그리고 국어 관련 학문적 연구 및 정보 기술 수준을 양적으로나 질적으로 향상시킬 수 있는 바탕이 될 것이다. 또한 일반인들에게 현대국어 기초 말뭉치가 향후 공개된다면 국어 생활 향상에도 기여하게 될 것이다.

그러나 아무리 다양하고 가치있는 텍스트 및 언어학적 분석 정보가 담겨있을지라도, 말뭉치 규모의 방대함과 제한된 기능의 말뭉치 분석 도구로 인한 학자 및 일반인들의 접근을 어렵게 만드는 문제가 있다. 이미 말뭉치 이용에 필요한 다양한 소프트웨어가 개발되어 있기는 하지만, 사용자가 필요한 정보를 추출하기에는 그 규모가 개인적 처리 수준을 넘어서게 되었고, 다양한 목적에 부응하는 정보추출에는 한계가 있다고 할 수 있다. 따라서 현대국어 기초 말뭉치를 개별 사용 목적에 적합하도록 재처리하는 작업이 필요하며, 대규모 말뭉치에서 사용자가 요구할 수 있는 정보를 미리 추출하는 작업도 향후 진행되어야 할 것이다.

참고 문헌

[1] 김흥규 외 (1998) 21세기 세종계획 국어 기초자료 구축 분과 제1차년도 연구 보고서, 문화관광부.
[2] 김흥규, 강범모 (1996) “고려대학교 한국어 말모둠 1”, 한국어학 3, 한국어학회.
[3] 강범모, 김의수 (2004), “세종 구문분석 말뭉치를 위한 구문 분석 방법”, 코퍼스와 어휘데이터베이스(강범모 외 편), 월인.
[4] 김흥규 외 (1999) 21세기 세종계획 국어 기초자료 구축 분과 제2차년도 연구 보고서, 문화관광부.
[5] 김흥규 외 (2000) 21세기 세종계획 국어 기초자료 구축 분과 제3차년도 연구 보고서, 문화관광부.
[6] 김흥규 외 (2001) 21세기 세종계획 국어 기초자료 구축 분과 제4차년도 연구 보고서, 문화관광부, 국립국어원.
[7] 김흥규 외 (2002) 21세기 세종계획 국어 기초자료 구축 분과 제5차년도 연구 보고서, 문화관광부, 국립국어원.
[8] 김흥규 외 (2003) 21세기 세종계획 국어 기초자료 구축 분과 제6차년도 연구 보고서, 문화관광부, 국립국어원.
[9] 김흥규 외 (2005) 21세기 세종계획 국어 기초자료 구축 분과 제8차년도 연구 보고서, 국립국어원.
[10] 김흥규 외 (2006) 21세기 세종계획 국어 기초자료 구축 분과 제9차년도 연구 보고서, 국립국어원.
[11] Aston, G. & Burnard L. (1998) *The BNC handbook: Exploring the British National Corpus with SARA*, Edinburgh: Edinburgh University Press.