21세기 세종 계획 특수자료 구축 분과의 성과 (1998~2007)

서상규
연세대학교 국어국문학과
inaka@yonsei.ac.kr

The 21st Century Sejong Project Special Corpus Construction (1998~2007)

Sang-kyu Seo
Dept. of Korean Language & Literature, Yonsei University

요 약

이 발표는, <21세기 세종 계획>(문화관광부/국립국어원의 지원, 1998~2007)의 일환으로 이루어진, 특수자료 구축 분과의 지난 10년간의 성과를 소개하고자 하는 데에 목적이 있다. 특수자료 구축 분과에서는 구어, 병렬, 역사 자료, 북한 및 해외 말뭉치와 같은 특수 말뭉치의 구축을 담당하고 있다. 여기서는 특수자료 구축 소분과의 개요와 과제의 구성, 각 세부 과제별 말뭉치 구축 성과 및 각 말뭉치의 가치와 특성을 밝히고자 한다.

1. 특수자료 구축 분과의 목적과 과제 구성

오늘날의 정보화 사회에서 국민의 언어와 문화 생활을 과학적 토대 위에서 향상시키고 국어의 특성에 적합한 정보 문화와 정보 처리 이론 및 기술을 발전시키기 위해서는, 다양한 언어 자료를 축적하고 이를 정보·지식으로 가공하는 일이 매우 중요하다. <21세기 세종 계획>의 일환으로 수행된 국어 자료 구축 사업은, 바로 이러한 국어 자료의 데이터베이스를 다양하게 구축함으로써, 언어 정책의 수립과 교육의 효율화, 정보 처리의 생산성 향상, 나아가 국민들 모두가 쉽고 편리하게 언어 정보를 찾고 활용할 수 있도록 하고자 하는 목적으로 추진되었다.

특수자료 구축 분과에서는 국어정보화 중장기 사업인 <21세기 세종 계획>의 한 축으로서, 국어 기초 자료 중에서 특히 연구자들이 쉽게 구축하거나 접근하기 힘든 특수 자료를 대규모의 국가 말뭉치 수준으로 구축하는 것을 목표로 하고 있다.

본 사업은 1998~2007년의 10년 동안 3단계로 나누어 진행되었으며, 현대 국어 구어 전사 말뭉치 구축, 병렬 말뭉치 구축(한영, 한일 등), 북한 및 해외 한국어 말뭉치 구축, 역사 자료 말뭉치 구축, 전문 용어 말뭉치 등의 세부 과제로 구성되어 있다.

지난 10년간 구축된 특수자료 분과의 원시 말뭉치와 형태소 분석 말뭉치 구축 성과를 간략히 요약하면 아래의 표와 같다.[1)]

구분 \ 단계		1단계	2단계	3단계	합계
현대 국어 구어 전사 말뭉치		153만	194만	172만	519만
병렬 말뭉치	한·영	100만	307만	163만	723만
	한·일	-	65만	73만	
	한·중	-	15만	-	
	한·러				
	한·불				
북한 및 해외 한국어 말뭉치		395만	394만	294만	1,083만
역사 자료 말뭉치		245만	206만	161만	612만
전문 용어 말뭉치		-	-	200만	200만
합계		3,127만 5천			

<표-1> 21세기 세종 계획 특수자료 구축 현황

1) 이하, 1단계(1998~2000), 2단계(2001-2003), 3단계(2004~2007)

2. 과제별 목표와 성과

2.1. 현대 국어 구어 전사 말뭉치

우리의 의사소통 행위는 음성과 문자를 통해 이루어지는데, 그 가운데서도 중심이 되는 것은 당연히 구어이다. 구어는 구체적인 상황에서 실시간으로 다양하게 실현되는 언어이기 때문에, 언어의 모든 양상을 있는 그대로 생생하게 관찰할 수 있다. 따라서 구어 자료의 구축은 실제적인 언어 사용의 양상과 다양한 언어의 모습을 밝히기 위한 기초 자료로서, 문어 자료와 함께 언어 분석의 균형성을 갖추기 위한 필수적인 요소라 할 것이다.

구어는 문어와는 전혀 다른 특성을 지니기 때문에, 실제 구어 말뭉치를 구축하기 위해서는 매우 복잡한 과정을 거치게 된다. 현대 국어 구어 전사 말뭉치 구축 과제에서 구축한 구어 말뭉치의 특성은 다음과 같다.

첫째, 우리의 일상 언어생활을 대표할 수 있을 만큼 다양한 발화가 수집될 수 있도록 자료 수집 단계의 설계를 구어의 분류 체계를 치밀히 설계하여 최대한 균형적인 구어 자료가 확보될 수 있도록 노력하였다. 둘째, 발화된 음성 언어는 즉각적이고 일회적이기 때문에 구어를 데이터베이스화하기 위해서는 반드시 이를 실시간에 녹음하여 그 음성을 전사하는 과정을 거쳐야 하는데, 이 과정에서 음성 발화에 담겨진 다양한 정보들이 최대한 유지될 수 있도록 전사 체계를 확립하고 이를 철저히 지켰다. 셋째, 구어의 수집 과정에서 구체적인 발화 상황은 물론이고, 발화자의 개인 정보, 언어 자료로 활용하도록 허락하는 발화자의 자료 사용 동의서와 같은 자료들을 확보하였다.

이 과제에서 구축된 구어 전사 말뭉치는, 말뭉치를 기반으로 한 국어 연구, 사전학, 담화 분석, 실험음성학 등의 언어학적 연구와 언어 교육, 언어 병리학, 구어의 분석과 활용 기술 개발과 관련된 공학 분야 등에서 다양하게 이용될 수 있을 것이다.

구분 \ 단계	1단계	2단계	3단계	합계
원시 말뭉치	153만	144만	122만	419만[2]
형태소 분석 말뭉치	-	50만	50만	100만

<표-2> 현대 국어 구어 전사 말뭉치 구축 현황

2) <표-2>에 단계별 구축량으로 제시된 값은 만 단위 이하를 버린 값이기 때문에 합계의 수치는 실제 구축량과 최대 3만의 오차가 있을 수 있다. <표-3>의 원시 말뭉치 총 합계와 <표-2>의 제시된 값이 4천 어절 가량 차이가 있는 것도 이러한 이유 때문이다.

1998~2007년 동안에 구축된 구어 말뭉치는 총 519만 어절로, 원시 말뭉치가 419만 어절, 형태소 분석 말뭉치가 100만 어절이다.

구어 전사 원시 말뭉치는 텍스트 유형과 발화 주제, 발화 상황 등의 면에서 다양성과 균형성을 갖추도록 구성되었다. 텍스트의 상호작용성에 따라 독백과 대화로 분류된 원시 말뭉치는 다시 공공성에 따라 공적 텍스트와 사적 텍스트로 분류된다. 각각에 해당되는 텍스트의 유형과 구축량은 아래의 표와 같다.

상호작용성	공공성	텍스트 유형	어절 수
독백	공적	강연	317,383
		강연(라디오)	63,919
		강의	243,931
		강의(TV)	20,202
		개회사/폐회사	5,238
		발표	171,432
		설교	18,090
		설교(라디오)	4,093
		주례사	2,679
		축사	824
		행사 대화	15,309
		회의(발표 토론형)	18,187
	합계		**881,287**
	사적	경험담 이야기하기	227,127
		동화 들려주기	7,829
		영화 줄거리 이야기하기	21,365
	합계		256,321
독백 합계			**1,137,608**
대화	공적	구매대화	22,690
		뉴스(TV)	223,846
		방송대화(라디오)	99,309
		방송대화(TV)	289,036
		방송대화(TV)/인터뷰	59,092
		상담	252,559
		상담(라디오)	20,757
		수업대화	75,774
		인터뷰	10,239
		인터뷰(TV)	16,688
		주제대화	29,033
		진료대화	3,008
		토론	193,006
		토론(TV)	293,446
		토론식 강의	6,359
		토의	86,814
		참여식 강의	55,455
		회의	91,706
	합계		**1,828,817**
	사적	일상대화	1,003,055
		전화대화	13,947
		주제대화	220,655
	합계		**1,237,657**
대화 합계			**3,066,474**
총합			**4,204,082**

<표-3> 현대 국어 구어 전사 원시 말뭉치의 구성

원시 말뭉치가 언어학 및 제반 분야에서 실제적으로 활용되기 위해서는 원시 말뭉치에 형태소의 문법 범주에 관한 정보를 부착한 형태소 분석 말뭉치의 구축이 필수적이다. 언어의 실제 모습을 반영하고 있는 구어 자료는 발화 장면, 발화 목적, 발화 주제, 화자의 연령 및 성별, 개인적인 언어 습관, 등에 따라 문어에서는 관찰할 수 없는 다양한 변이형을 보이며, 발화의 구성 또한 복잡하게 실현된다. 이러한 구어의 특징이 반영된 대규모의 구어 전사 형태소 분석 말뭉치는 그간 문어 중심으로 이루어져 오던 국어 연구의 한계를 극복할 수 있게 할 의미 있는 자료가 될 것이다. 구어 형태소 분석 말뭉치는 독백 자료가 약 40%, 대화 자료가 약 60%이며, 그 상세한 구성은 아래의 표와 같다.

상호 작용성	공공성	텍스트 유형	어절 수	비율	합계 (비율)
독백	공적	강연	30,002	2.97%	273,780 (27.14%)
		강연(라디오)	7,575	0.75%	
		강의	161,002	15.96%	
		강의(TV)	15,087	1.50%	
		발표	40,880	4.05%	
		설교	14,049	1.39%	
		개회사/폐회사	5,185	0.51%	
	사적	경험담 이야기하기	106,487	10.56%	126,996 (12.59%)
		동화 들려주기	7,829	0.78%	
		영화줄거리 이야기하기	12,680	1.26%	
대화	공적	구매대화	22,690	2.25%	226,374 (22.44%)
		방송대화(라디오)	5,063	0.50%	
		방송대화(TV)	18,805	1.86%	
		상담	11,334	1.12%	
		상담(라디오)	20,757	2.06%	
		수업대화	23,002	2.28%	
		진료대화	2,547	0.25%	
		토론	62,828	6.23%	
		토론(TV)	45,526	4.51%	
		회의	13,822	1.37%	
	사적	식사대화	12,933	1.28%	381,519 (37.83%)
		일상대화	214,241	21.24%	
		전화대화	14,274	1.42%	
		주제대화	140,071	13.89%	
합계			**1,008,669**	**100%**	**1,008,669 (100%)**

<표-4> 현대 국어 구어 전사 형태소 분석 말뭉치의 구성

2.2. 병렬 말뭉치

국제적인 문화와 정보의 교류가 급증하면서 자동 번역 및 통역을 비롯한 자연언어처리 및 언어공학 분야의 필요성이 대두되고 있다. 이러한 현실을 반영하기 위해서는 국가 간에 언어 정보를 공유하고 자원화 하는 일이 필요하다. 또한 기계번역, 번역 검증 시스템, 언어 교육, 비교 언어학, 대조 언어학 등의 연구와 이를 뒷받침할 수 있는 소프트웨어의 개발을 위해서는 둘 이상의 언어를 대응시켜 구성한 병렬 말뭉치가 필수적이다.

당초 이 병렬 말뭉치는 다국어 병렬 말뭉치의 구축을 목표로 하였으므로, 한·영 병렬 말뭉치와 한·일 병렬 말뭉치 외에도 한·중 병렬 말뭉치, 한·러 병렬 말뭉치, 한·불 병렬 말뭉치 등의 시험 구축과 기초 연구가 수행되었다.

병렬 말뭉치 구분 \ 단계		1단계	2단계	3단계	합계
한·영	원시	100만	255만	115만	470만
	형태소 분석	-	52만	48만	100만
한·일	원시	-	60만	49만	109만
	형태소 분석	-	5만	24만	29만
한·중	원시	-	15만	-	15만
한·러					
한·불					

<표-5> 병렬 말뭉치 구축 현황

위의 표에 나타난 바와 같이, 한·영 병렬 원시 말뭉치는 총 470만 어절, 형태소 분석 말뭉치는 100만 어절이 구축되어 있고, 한·일 병렬 원시 말뭉치는 약 100만 어절, 형태소 분석 말뭉치는 29만 어절이 구축되어 있다.

한·영 병렬 원시 말뭉치는 매체에 따라 잡지, 책, 기타 출판물, 기타 비출판물로 분류되며 다시 내용 및 세부 분류 기준에 따라 아래와 같이 나뉜다.

매체 분류	내용 분류	세부 분류	어절 수	비율
잡지	잡지	총류	477,185	10.04%
책	교과서		309,263	6.50%
	상상적 텍스트	소설	566,269	11.91%
		성경	1,340,795	28.20%
	기타 비상상적 텍스트		1,508,469	31.73%
기타 출판물	안내문, 소책자, 정부 문서	전자, 사회, 경제, 인문 등	250,927	5.28%
기타 비출판물	연설문	사회, 외교	295,651	6.22%
	기타	사회, 인문, 일반	6,056	0.13%
합계			**4,754,615**	**100%**

<표-6> 한·영 병렬 원시 말뭉치의 구성

병렬 말뭉치의 활용성을 높이기 위해서는 형태론적 정보를 부착한 주석 말뭉치의 개발이 필수적이다. 한영 병렬 형태소 분석 말뭉치는 추후의 보급과 활용을 위해 원시 말뭉치 중 저작권 활용 승낙서를 받은 문서를 우선적

으로 선정하여 구성하였다.

매체별	내용별	어절 수	비율
잡지	총류	100,768	10.0%
책	상상적 텍스트	94,852	9.4%
	인문	113,025	11.2%
	사회	45,048	4.5%
	생활	99,507	9.8%
	교육자료	345,178	34.2%
기타 출판물/ 기타 비출판물	사회/외교, 일반(연설문)	211,338	20.9%
	합계	**1,009,716**	**100%**

<표-7> 한·영 병렬 형태소 분석 말뭉치의 구성[3)]

한편 한·일 병렬 말뭉치의 경우, 아래의 표에서 볼 수 있듯이 신문, 잡지, 책, 기타 출판물, 기타 비출판물, 시나리오의 다양한 장르들로 약 100만 어절 가량의 원시 말뭉치가 구축되어 있으며, 이들 중 29만 어절에 대해서는 형태소 분석 말뭉치가 구축되어 있다.

	어절 수	비율
신문	177,284	16.09%
잡지	114,289	10.37%
책-소설	370,155	33.59%
책-시	18,587	1.69%
책-기타	2,217	0.20%
책-정보	18,988	1.72%
기타 출판물	221,907	20.14%
기타 비출판물	104,294	9.47%
시나리오	74,157	6.73%
총합	**1,101,878**	**100%**

<표-8> 한·일 병렬 원시 말뭉치의 구성

2.3. 북한 및 해외 한국어 말뭉치

현재 우리가 사용하는 한국어는 공간적으로는 국내 각 지역의 방언을 포함하고, 북한 및 해외에 있는 동포들의 한국어까지를 포용하는 것이다. 그러므로 북한 및 해외 한국어에 대한 심층적 연구와 한민족의 언어적·문화적 통합의 기반을 마련하기 위한 언어 자원의 확보 또한 긴요하다. 이러한 필요성으로 인해 21세기 세종 계획 특수자료 구축 분과에서는 북한 및 해외 한국어 말뭉치를 구축하였으며, 구축된 말뭉치는 원시 말뭉치가 약 900만 어절, 형태소 분석 말뭉치가 150만 어절이다.

3) <표-8>에 나타난 한·일 병렬 원시 말뭉치 총량과 <표-5>에 제시된 총량이 2천 어절 가량 차이가 나는 이유는 각주 1)에 기술한 것과 같은 이유이다.

구분 \ 단계	1단계	2단계	3단계	합계
원시 말뭉치	395만	320만	218만	933만
형태소 분석 말뭉치	-	74만	76만	150만

<표-9> 북한 및 해외 한국어 말뭉치 구축 현황

2.4. 역사 자료 말뭉치

민족 문화 유산으로서의 가치가 있는 역사 자료는 원전의 말뭉치 구축 자체가 큰 의의를 가진다. 또한 국어의 발달 변천사나 방언론과 같은 국어 연구와 국어 교육, 그리고 어문 생활의 발전을 기하기 위해서는 역사 자료들의 정보화가 필수적이다. 고어와 방언에 대한 이해는 이해 그 자체에 그치는 것이 아니라 현대 국어의 근원을 파악할 수 있게 해 주고, 현대의 어문 생활의 길잡이가 된다는 점에서도 중요하다. 21세기 세종 계획 특수자료 구축 분과에서는 일반 국민들 모두 쉽고 편리하게 역사 자료를 검색하고 활용할 수 있도록 하기 위해 15~20세기 초기 자료들로 구성된 역사 자료 말뭉치를 구축하였다. 역사 자료 말뭉치는 원시 말뭉치가 550만 어절, 형태소 분석 말뭉치가 62만 어절이다.

구분 \ 단계	1단계	2단계	3단계	합계
원시 말뭉치	245만	154만	151만	550만
형태소 분석 말뭉치	-	52만	10만	62만

<표-10> 역사 자료 원시 말뭉치의 구성

2.5. 기타 말뭉치

그 밖에도 특수자료 구축 분과에서 구축된 말뭉치로는 전문용어 말뭉치가 있다. 전문용어 말뭉치는 KORTERM의 '전문용어의 정비' 분과의 전문용어 표준화 사업에 기반 자료를 제공하는 것을 목적으로 해당 전문학술학회에 협조를 구해 그 분야의 가장 대표적인 텍스트를 선정하여 구축하였다.

3. 특수분과 자료의 가치와 특성

3.1. 현대 국어 구어 전사 말뭉치

현대 국어 구어 전사 말뭉치의 가장 큰 가치는 실제 발화되는 소리의 정보를 담은, 생동감 있는 현실음의 입말 자료라는 점이다. 그러므로 실제 발화에서 나타나는 다양한 음성(음운) 정보를 포함하도록 구축함으로써 구어 전사 말뭉치가 실제 음성언어에 가까운 말뭉치로 구

축하기 위한 노력이 필요하다. 한편으로, 향후의 주석과 기계적 검색 단계에서의 이용의 용이성을 확보하기 위해서, 어문규정에 준한 철자법 전사를 기본으로 한 다음, 각종 음성 정보를 표시하도록 하였다.

21세기 세종 구어 말뭉치는 다음과 같은 특징을 지니고 있다.

(1) 구축 과정에서 통사론의 기본 단위인 '문장(각종 문장 기호를 포함하는)'을 기준으로 삼지 않고, 실제 발화인 입말의 단위인 '억양 단위'를 기본 단위로 하여 구축되었으며, 각 단위에서의 억양 정보를 4가지로 구분하여 기호로 표시하였다.[4)]

(2) '끊어진 어절, 불분명한 어절, 쉼' 등의 정보를 비롯하여 '웃으면서 말하는 부분, 노래 부르는 부분, 박수치면서 부르는 부분, 음성에 준하는(실제 대화가 아닌) 기타 소리들'에 대한 정보를 TAG를 이용하여 표시함으로써 실제 음성의 정보를 최대한 반영하였다.

(3) 문어에서 잘 나타나지 않는 담화표지인 '음, 이, 그, 저, 아' 등은 바로 뒤에 '~'를 표시하여 다른 형태와 구별('음~, 이~, 그~, 저~, 아~')하였고, 대화 중 상대방이 대화에 끼어든 위치, 말을 하다가 끊어지는 불완전한 발화에 대한 정보를 표시하였다.

(4) 실제 발화에서 자주 나타나는 음운 현상에 대한 정보를 포함하기 위해, /ㅟ/, /ㅢ/가 반홀소리가 되어 /ㅜ/, /ㅏ/와 축약되는 현상이나, 확실한 장음을 표시하였다.[5)]

(5) 구어 전사 말뭉치는 사적 자료와 공적 자료의 구성비가 1:1에 가까운 균형성을 갖추고 있다.[6)]

(6) 저작권 문제가 거의 대부분 해결되었다는 점은 구어 전사 말뭉치의 큰 장점으로 꼽을 수 있다. 구어 전사 말뭉치는 구축 초기부터 음원을 제공해 준 발화자들의 동의를 얻어 구축되었기 때문에 저작권이 대부분 해결되었다.[7)]

3.2. 병렬 말뭉치

병렬 말뭉치를 구축함에 있어서는 양질의 번역본을 채택하는 것이 무엇보다 중요하다. 구축 대상 텍스트를 선정할 때는 다음의 세 가지 사항을 고려하여 양질의 텍스트가 선정되도록 하였다.

첫째로, 기계번역이나 문체 연구, 한국어 교육, 대조 연구 등에 두루 활용이 가능하도록 범용성이 있는 자료를 채택하였다. 둘째, 구축된 자료가 실제 언어 연구나 교육, 기계 번역 등에 활용될 수 있도록 실용성을 고려하였으며, 셋째, 텍스트의 유형 및 번역의 방향성에 있어서 균형성을 유지하도록 매 해마다 기존의 말뭉치 지도를 토대로 부족한 장르를 보완하여 균형성 확보에 주력하였다.

그 결과 병렬 말뭉치는 신문, 잡지, 법률, 논문, 시나리오, 소설, 수필, 시, 성경, 매뉴얼 등 다양한 장르로 구성되었으며, 이를 통해 특정 장르에 편중되지 않은 대역 양상을 살펴볼 수 있다. 이러한 세 가지 고려사항 외에도 자료 선정 시 의역이 심하게 이루어진 것은 제외하여 언어 연구 및 기계 번역에 활용될 수 있는 자료들을 채택하였다. 이와 같이 양질의 다양한 대역 텍스트들로 구성되어 있다는 점이 21세기 세종 병렬 말뭉치의 가장 큰 특징이라 할 수 있다.

한편 병렬 말뭉치의 경우 장르 간 균형성 외에도 번역 방향성의 측면에서도 균형성이 고려되었다. 번역의 방향성이란 원본 텍스트의 언어와 대역 텍스트의 언어가 각기 무엇인가에 따른 것으로, 어느 언어가 원본 텍스트의 언어이냐에 따라 번역이 양상이 달라질 수 있기 때문에 병렬 말뭉치의 균형성을 확보하기 위해서는 번역의 방향성에 대한 고려가 필수적이다. 특수자료 구축 분과의 병렬 말뭉치는 이러한 번역이 방향성에 있어서도 균형성을 갖추고 있는데, 한·영 병렬 형태소 분석 말뭉치에는 한국어가 원본인 한·영 대역 말뭉치가 46%, 영어가 원본인 영·한 대역 말뭉치가 54% 포함되어 있으며, 한·일 병렬 말뭉치의 경우에도 한국어가 원본인 한·일 병렬 말뭉치가 47%, 일본어가 원본인 일·한 병렬 말뭉치가 53% 포함되어

---

4) 억양과 쉼, 통사적 정보, 글말 등을 참고하여, 내림 억양이면서 쉼이 있는 경우에는 마침표(.)를, 약간 짧은 쉼이 있고 약한 문말 오름 혹은 문말 내림 억양인 경우 쉼표(,)를, 확실한 오름 억양일 경우에는 물음표(?)를, 활기에 넘치는 기운찬 어조·감탄이 나타나는 억양일 경우에는 느낌표(!)를 사용하였다.

5) 이러한 현상은 입말에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 입말 전사에서는 ' (apostrophe)를 사용하여 두 음소를 연결해 주었으며(ex. 사귀' 어), 확실한 장음은 …로 표시하여 현실음에 최대한 가까게 표현하였다.

6) 독백 자료와 대화 자료의 비율이 4:6 비율이기는 하지만, 2명 이상의 사람이 대화를 나눈 자료의 중요성을 감안한다면 충분히 균형성을 갖추었다고 할 수 있다.

7) 다만, 방송토론이나 대담과 같은 방송 자료들은 저작권 문제가 토론자(혹은 발화자)와 방송국에 걸쳐 있어서 저작권 문제가 아직 해결되지 않은 것이 일부 포함되어 있다.

있어, 번역의 방향성에 있어서 균형성을 확보하고 있다.

3.3. 북한 및 해외 한국어 말뭉치

북한 및 해외 한국어 말뭉치는 한국어 연구자들이 사용할 수 있는 연구 자료의 폭을 대폭 확대시켜 줄 수 있다는 점에서 가장 의의가 있다. 현재까지 구축되어 있는 대부분의 말뭉치가 남한의 언어만을 반영한 것인데 비해, 북한 및 해외 한국어 말뭉치는 자료를 구하기 어려운 북한과 중국, 구소련 지역의 한국어 자료들로 구성되어 있다는 점에서 희귀 자료로서의 중요한 가치가 있다.

3.4. 역사 자료 말뭉치

역사 자료 말뭉치의 구축에 있어서 가장 중요한 점은 원전의 정보를 최대한 충실하게 반영하면서도 이를 검색에 용이하도록 구축하는 것이다. 역사 자료 말뭉치는 방점이나 한자음이 원전과 일치하도록 입력되어 있기 때문에, 중세 국어를 연구하기 위한 자료로서의 가치가 크다.

원전에는 대부분 띄어쓰기가 되어 있지 않지만 말뭉치로 가공되는 과정에서는 검색의 편의를 위해 반드시 띄어쓰기가 이루어져야 한다.[8] 이체자 역시 원전의 정보를 그대로 유지시키기 어려운 부분인데, 원본과 말뭉치의 한자 자형이 다르게 입력된 경우 해당 한자를 별도의 표로 제시하였다.

이처럼 21세기 세종 계획 특수자료 구축 분과의 역사 자료 말뭉치는 원전의 정보를 충실히 반영하는 동시에 전자 자료로서의 이용 가능성을 최대한 살렸다는 장점이 있다.

<참고문헌>

국립국어원(2005,2006). “21세기 세종 계획 국어 특수자료 구축 연구보고서”.

문화관광부(1998,1999,2000,2001,2002,2003,2004). “21세기 세종 계획 국어 기초자료구축 분과 특수자료구축 소분과 연구보고서”.

서상규(2002), “한국어 말뭉치의 구축과 과제”, 「한국어와 정보화」. 태학사.

서상규(2005), “Informatization and Use of Korean Language Data”, *The Review of Korean Studies* 8-4. 한국학중앙연구원.

서상규·구현정 공편(2002). 「한국어 구어 연구(1)-구어 전사 말뭉치와 그 활용」, 한국문화사.

서상규·김형정(2005). “구어 말뭉치 설계의 몇 가지 조건”, 「언어정보와 사전편찬」 제14·15·16합집. 연세대 언어정보연구원.

이태영(2003), “국어사 자료의 전산화와 21세기 세종계획”, 국어사학회 제15회 학술대회 발표논문집.

이한섭(2007). “한일 병렬코퍼스의 구축과 활용”, 홋카이도대학 국문학과 국제 심포지움 발표 자료집, 日本, 北海道大.

전영옥(2006). “구어 말뭉치의 구축과 활용”. 21세기 국어정보화아카데미, 연세대학교.

정태구 · 김흥규 · 김정숙(2000), “한 · 영 병렬 코퍼스의 설계, 구축 및 응용 방안 연구”, 「한국어학」 11. 한국어학회.

8) 따라서 역사 자료 말뭉치는 현대국어의 띄어쓰기에 따라 자료를 입력하되, 원전과 달라진 부분에 대해서는 #나 $의 기호를 삽입하여 원전과 다르다는 점을 표시하였다.