국어 특수 자료 구축의 성과와 전망[1)]

서상규 · 연세대학교 교수

1. 머리말

국어학을 비롯해서 국어정보학, 국어 정보 처리, 정보 검색, 언어 병리학, 대조 언어학 등의 분야에서 언어 규칙을 찾아내고 이를 활용하고자 할 때, 우리에게는 무엇이 필요할까? 또 외국인들이나 재외동포들에게 한국어를 설명하고 가르치려고 할 때는 어떤 것이 필요할까? 또한 국민들의 언어생활을 편리하고 올바르게 할 수 있도록 여러 가지 정책을 세우려고 할 때 우리에게 필요한 것은 무엇일까?

물론 어떤 물음이냐에 따라서 필요한 것들도 다를 것이다. 그러나 이 모든 질문들에 공통된 답이 되는 것이 하나 있다. 그것은 바로 국어를 잘 알아야 한다는 것이다. 국어를 잘 안다는 것은 그렇다면 무엇일까? 그것은 바로 한국인의 언어생활의 여러 현상을 있는 모습 그대로, 빠짐없이 파악하고 그 속에 숨은 모든 법칙과 원리를 밝혀내야 한다는 것을 뜻한

1) 이 글은 필자가 이전에 발표한 "한국어 특수 말뭉치의 구축 현황과 그 특징 —21세기 세종 계획의 성과를 중심으로—"(『한국사전학』제12호, 2008, 41-60쪽)를, 일반인 독자들이 읽기 쉽게 고쳐 쓰면서 내용을 새롭게 보완한 것이다.

다.

그러면 어떻게 해야 국어에 숨어 있는 원리들을 알아낼 수 있을까? 그것은 당연히 국어를 실제로 쓰는 장면과 그 자리에 있는 사람들을 자세히 관찰해야만 할 것이다. 그래야만 그때의 상황에서 어떠한 원리와 법칙에 따라서 언어 행동이 이루어졌는지를 알 수 있기 때문이다. 그러나 몇천만이 넘는 국어 화자들의 실제 언어생활의 기록을 모두 모아 놓고 관찰하기란 애초부터 불가능하다. 이런 탓에 이제까지 우리가 국어를 분석할 때는 모어 화자로서의 직관에 크게 의존하거나, 글로 쓰인 갖가지의 자료를 대상으로 삼아서 그 일을 대신해 온 것이다.

그런데 20세기 말에 이루어진 기술적 혁신에 따라서, 오늘날 우리는 잘 발달된 컴퓨터와 정보 저장 기술, 네트워크로 이어진 정보의 비단길 등의 환경을 누리고 있는데, 이러한 바탕 위에서 우리가 생각할 수 있는 가장 좋은 방법은, 누구나가 쉽게 관찰하고 실험하고 분석해 낼 수 있는 언어 자료를 잘 설계된 데이터베이스로 만들고, 이를 널리 공유할 수 있도록 하는 것이다. 국어를 본격적으로 관찰하고 분석할 만큼 충분한 국어 자료를 개개인이 혼자의 힘으로 수집하고 축적하여 데이터베이스화하는 일은 너무나도 어려운 일이기 때문이다.

데이터베이스의 대상으로서의 국어는 매우 다양한 측면을 지니고 있다. 즉 시간적으로 볼 때 아주 오래전에 쓰이던 옛말부터 오늘날의 말에 이르기까지 시대의 흐름에 따라 국어의 모습이 달라져 왔다. 뿐만 아니라 지역이나 공간에 따라서 국어의 모습이 다르게 나타나기도 하며, 글자로 표현되는 글과 소리로 표현되는 말의 모습도 종종 차이를 보이곤 한다. 이렇게 다양한 국어의 모습을 제대로 포착할 수 있을 만큼 다양하고도 충분한 국어 자료를 컴퓨터와 네트워크 상에서 자유롭게 활용하고 이를 바탕으로 국어 생활의 편리성을 향상시키는 것이야말로 바로 국어 정보화라고 할 수 있을 것이다.

'21세기 세종 계획'(이하, '세종 계획'으로 줄여 표시하기로 한다.)은 이

러한 국어 정보화의 기반을 구축하기 위해서, 1998년부터 2007년까지 10년 동안에 걸쳐서 국책 사업의 일환으로 추진되었다.

> 21세기 세종 계획 전체 사업의 목표는 우리말과 우리글을 바탕으로 하는 정보 사회 건설에 있다. 국어 기초 자료 구축 분과의 사업은 '한국어로 국어생활의 수준과 효율성을 높이며 고도 정보 사회의 요구에 부응하는 선진적 지식과 기술을 개발하고, 이를 세계화하는 일이 곧 우리나라가 국제 간의 정보화 경쟁에서 주도권을 확보할 수 있는 길이며, 또한 이것이 선진 정보 문화를 실현하는 요건'이라는 인식에서 시작된 것이다.(『21세기 세종 계획 국어 기초 자료 구축』(1998.12) 보고서, 3쪽.)

즉 '세종 계획'을 통해서 이루어진 국어 자료 구축 사업은, 국어 자료를 다양한 데이터베이스로 만들어 축적함으로써, 언어 정책의 수립과 교육의 효율화에 활용하고, 또한 정보 처리의 생산성을 향상시키며, 나아가 국민들 모두가 언어 정보를 쉽고 편리하게 찾고 활용할 수 있도록 하려는 것을 중요한 목적으로 하였다.

'세종 계획'은 크게 기초 자료 구축 분과, 특수 자료 구축 분과, 전자 사전 분과 등의 여러 분과로 구성되었다. 특수 자료 구축 분과는 기초 자료 구축 분과와 함께 국어 자료의 수집과 가공을 담당하였는데, 기초 자료 구축 분과가 주로 현대 국어 문어 자료를 대상으로 하는 데 비해, 특수 자료 구축 분과는 특수한 국어 자료를 수집하는 것을 임무로 하였다. 특수 자료란 두 가지 의미로 해석된다. 즉 현대 국어 문어 자료를 제외한 나머지 모든 국어 자료를 특수한 자료로 분류한다는 뜻도 있지만, 실제로는 국어 자료 중에서도 특히 일반인들이나 연구자들이 스스로 구하거나 대규모로 구축하기 어려운 자료를 뜻한다. 이렇게 체계적으로 수집, 구성된 특수한 언어 자료는 일반적인 국어 연구 외에도 매우 폭넓은 이용 목적과 용도가 있어서 이를 '특수 말뭉치'라고도 부르지만, 그 가리키는 바가 꼭 같지는 않다.

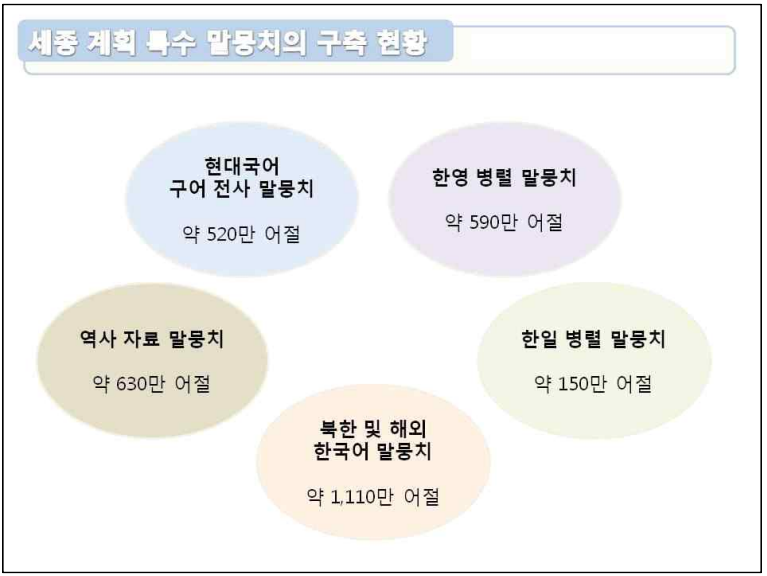
"특별한 목적이나 용도를 정하지 않고, 어휘, 문법, 담화 구조 등의 일반적인 언어 조사를 위해 텍스트들을 모은 말뭉치를 일반 말뭉치(general corpus)라 한다. (중략) 한편, 특정한 조사나 연구만을 위해 디자인된 말뭉치를 특수 말뭉치(specialized corpus)라 하는데……."(서상규·한영균(1999), 『국어정보학 입문』, 태학사, 32-33쪽.)

특수 말뭉치가 일반적으로 '한정된 목적'으로 쓰일 것을 전제로 하는데 비해, '세종 계획'의 특수 자료는 좀 다른 뜻으로 쓰이고 있다. 예컨대, 책이나 인쇄물을 통해서 비교적 쉽게 수집할 수 있는 '문어'에 비해서, '구어'의 경우에는 실제의 발화 장면을 녹음이나 녹화하여 이를 귀로 들으면서 글자로 옮기는 전사 과정을 거쳐야 하므로 구축에 인력과 경비, 시간이 많이 소요될 수밖에 없어서, 구축의 난이도가 높다는 점, 자료의 활용 범위가 문어에 비해서 매우 넓다는 점 등에서 문어 자료와 구별된다. 이런 뜻에서 특수 말뭉치로 분류되었을 뿐, 자료 자체의 가치에 따른 명칭이 아닌 것이다. 즉, '세종 계획'에서의 특수 자료(말뭉치)는 꼭 특별한 목적이나 용도가 한정된 자료를 뜻한다기보다는, '현대 국어 문어 자료'를 제외한 갖가지의 국어 자료들을 모두 포함하는 넓은 개념으로 사용된 것이다.

'세종 계획'에 의해 만들어진 특수 말뭉치는 매우 다양한데, 여기에는 '현대 국어 구어 전사 말뭉치', '병렬 말뭉치'(한영, 한일 등), '북한 및 해외 한국어 말뭉치', '역사 자료 말뭉치', '전문용어 말뭉치' 등이 포함된다.

특수 말뭉치가 이렇게 다양한 자료로 구성된 것은, 국어를 총체적으로 이해하기 위해 필요한 요소를 빠짐없이 채워 나가기 위해서이다. 즉, 〈그림 2〉에서 보인 바와 같이, 시간에 따라 변해 온 국어, 공간과 지역에 따라서 달라지는 국어, 그리고 문어와 구어의 차이라는 3가지 측면에서 국어의 전체 범위를 대표할 수 있는 말뭉치 구축을 목적으로 설계된 것이다. 〈그림 2〉는 '세종 계획'의 말뭉치들의 성격과 위상을 이해하기 쉽게

표현한 것이다.[2)]

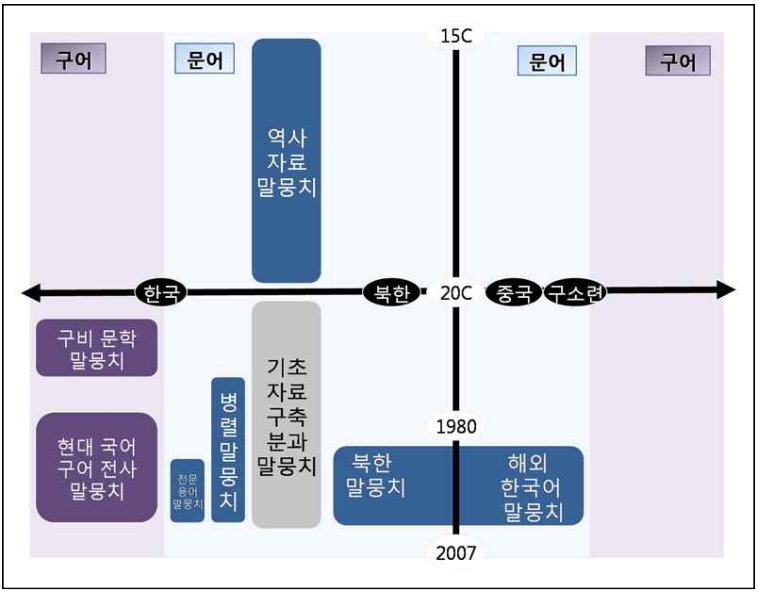


〈그림 1〉 세종 계획 특수 말뭉치의 구성과 구축 현황

〈그림 2〉의 세로축은 한국어의 시간에 따른 변이를 나타내고, 가로축은 공간적인 변이를 나타낸다. 또한 그러한 시공간의 축에 속한 각 언어 자료는 다시 문어와 구어로 갈라지게 된다. '세종 계획' 특수 말뭉치는 시간적으로는 15세기로부터 현대에 이르기까지의 한국어를 포함하고, 공간(지역)적으로는 남북한을 비롯해서 중국과 구소련 지역 등의 해외 지역의 한국어를 모두 수집의 대상으로 하는 것이다.

자연스럽게 발화된 말을 녹음해서 전사해야 하는 '구어' 자료는 〈그림 2〉에 나타난 바와 같이 과거의 발화 현장을 지금 녹음할 수 없으므로 시간적으로 20세기 말 현대의 자료에 국한될 수밖에 없다. 반면, 현재의 한

2) 〈그림 2〉에서 '기초 자료 구축 분과 말뭉치'는 '현대 국어 문어 말뭉치'를 가리키는데, 이 말뭉치를 제외한 대부분의 세종 계획 말뭉치가 특수 말뭉치에 속한다.



〈그림 2〉 '세종 계획' 특수 말뭉치의 위상

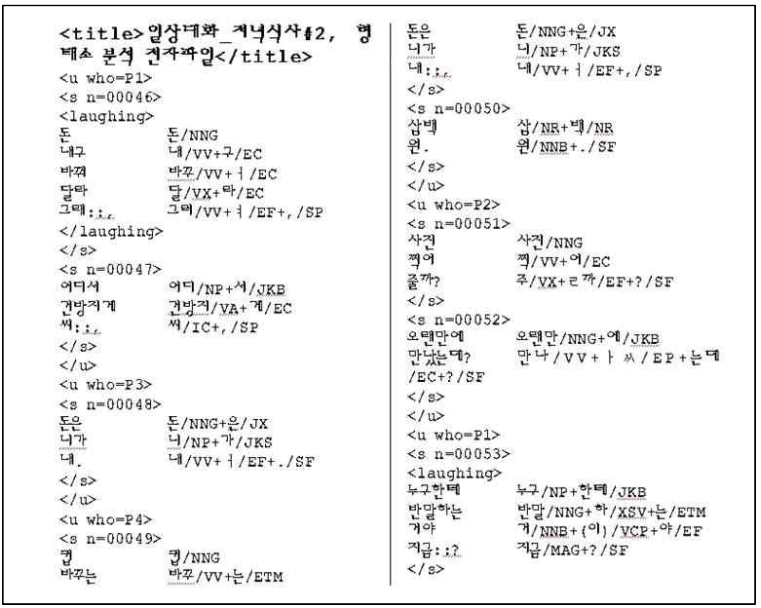
국어로 의사소통이나 사회 활동이 이루어지는 지역은 한반도뿐 아니라 중국 러시아 등 전 세계 여러 곳에 분포해 있다. 따라서 완전한 국어 정보화의 토대를 쌓기 위해서는 공간적으로 "전 세계에 흩어진 모든 한국어"를 수집하여 데이터베이스화하는 것이 바람직할 것이다. 그러나 '세종 계획'에서는 여러 가지 어려움으로 해외의 모든 지역을 포함하지는 못하고, 우선 남한의 자료를 수집하되, 향후의 사업의 확대가 이루어질 때를 대비하여 녹음과 전사, 주석 등 일련의 말뭉치 수집 구축 과정에 관한 기초 연구와 구어 말뭉치 분류 체계를 수립하는 데에도 큰 힘을 기울였다.

실제 말뭉치의 구축은 1998년~2007년의 10년 동안 3단계로 나누어 이루어졌다. 그 구축 성과를 간략히 보면 〈표 1〉과 같다.

구분 \ 단계		1단계 (1998~2000)	2단계 (2001~2003)	3단계 (2004~2007)	합계
현대 국어 구어 전사 말뭉치		153만	194만	172만	519만
병렬 말뭉치	한·영	100만	307만	163만	716만
	한·일	–	60만	71만	
	한·중	–	15만	–	
	한·러				
	한·불				
북한 및 해외 한국어 말뭉치		395만	394만	294만	1,083만
역사 자료 말뭉치		245만	206만	161만	612만
전문용어 말뭉치		–	–	200만	200만
합계					3,130만

〈표 1〉 '21세기 세종 계획'의 특수 자료 구축 성과(1998~2007) (단위: 어절)

'세종 계획' 특수 자료 분과의 자료는 원시 말뭉치와 형태소 분석 말뭉치로 구성되어 있는데, 원시 말뭉치는 원문(원래의 발화) 그대로를 컴퓨터로 입력 저장한 자료를 가리키고, 형태소 분석 말뭉치란 하나하나의 어절마다 어떤 단어와 품사 구성으로 이루어졌는지를 분석한 자료이다. 〈그림 3〉은 구어 형태소 분석 말뭉치의 실제 예이다. 각 줄의 맨 왼편이 원 어절이고 오른편이 분석된 결과인데, 여기에는 단어 또는 문법 요소들의 형태에 각각 로마자로 표시된 품사 정보가 기록되어 있다.



〈그림 3〉 구어 형태소 분석 말뭉치

2. 현대 국어 구어 말뭉치의 성과와 특성

2.1. 구어 말뭉치의 목적과 활용

인간의 의사소통 행위는 주로 말소리(음성)와 글자(문자)를 통해 이루어지는데, 특히 구어는 구체적인 상황에서 실시간으로 실현되는 언어이기 때문에 언어의 모든 양상을 있는 그대로 생생하게 관찰할 수 있는 장점이 있다. 따라서 구어 자료는 실제적인 국어의 사용 양상을 밝히기 위해 필수적이라 할 것이다.

구어 자료는 발화 장면, 발화 목적, 발화 주제, 화자의 연령 및 성별, 개인적인 언어 습관 등에 따라 문어에서는 관찰할 수 없는 다양한 변이형을 보이는 등 언어의 실제 모습을 반영하고 있는데, 실제 구어의 특징을

제대로 포착하기 위한 대규모의 구어 말뭉치는, 이제까지 문어 중심으로 이루어져 오던 국어 연구의 한계를 극복하게 해 주는 데에 중요한 몫을 할 것이다.

2.2. 구어 말뭉치의 구축 현황

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

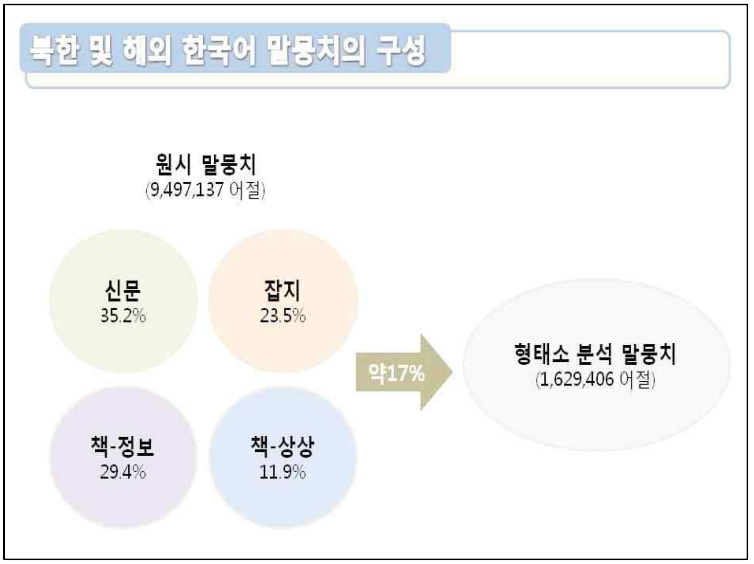
구분 / 단계	1단계 (1998~2000)	2단계 (2001~2003)	3단계 (2004~2007)	합계
원시 말뭉치	153만	144만	122만	419만[3]
형태소 분석 말뭉치	–	50만	50만	100만

〈표 2〉 현대 국어 구어 전사 말뭉치 구축 현황

구어 자료를 수집할 때는, 되도록 다양한 발화(텍스트)의 유형과 주제, 발화의 상황 등을 포함하면서도 여러 가지 자료가 골고루 포함되도록 애썼다. 텍스트의 상호작용성, 즉 말하는 사람과 듣는 사람이 함께 말하느냐 아니냐에 따라서 구어는 크게 독백과 대화로 나뉜다. 또한 말하는 장면의 공공성에 따라서 공적인 발화(텍스트)와 사적인 발화로 나눌 수 있다.

3) 〈표 2〉에 단계별 구축량으로 제시된 값은 만 단위 이하의 수를 버린 값이기 때문에 합계의 수치는 실제 구축량과 약간 오차가 날 수 있다. 〈표 2〉와 〈표 3〉의 원시 말뭉치 총 합계에 차이가 있는 것도 그 때문이다.

각각에 해당되는 텍스트의 유형과 구축량은 아래의 표와 같다.



〈그림 4〉 '세종 계획'의 현대 국어 구어 (전사) 말뭉치의 구성

상호작용성	공공성	텍스트 유형	어절 수	비율
독백	공적	TV뉴스	223,846	5.3%
		강연	381,302	9.1%
		강의	264,133	6.3%
		개·폐회사/축사/주례사	8,741	0.2%
		발표	171,432	4.1%
		설교	22,183	0.5%
		행사 대화	15,309	0.4%
		회의(발표 토론형)	18,187	0.4%
	합계		**1,105,133**	**26.3%**
	사적	경험담 이야기	227,127	5.4%
		동화 들려주기	7,829	0.2%
		영화 줄거리 이야기	21,365	0.5%
	합계		**256,321**	**6.1%**
독백 합계			**1,361,454**	**32.4%**
대화	공적	구매 대화	22,690	0.5%
		방송 대화(TV/라디오)	388,345	9.2%
		인터뷰	86,019	2.0%
		상담	273,316	6.5%
		수업 대화	75,774	1.8%
		주제 대화	29,033	0.7%
		진료 대화	3,008	0.1%
		토론	548,266	13.0%
		토의/회의	178,520	4.2%
	합계		**1,604,971**	**38.2%**
	사적	일상 대화	1,003,055	23.9%
		전화 대화	13,947	0.3%
		주제 대화	220,655	5.2%
	합계		**1,237,657**	**29.4%**
대화 합계			**2,842,628**	**67.6%**
전체 합계			**4,204,082**	**100.0%**

〈표 3〉 현대 국어 구어 전사 원시 말뭉치의 구성

'세종 계획'의 구어 말뭉치는, 말뭉치를 기반으로 한 국어 연구, 사전학, 담화 분석, 실험음성학 등의 분야에서 이루어지는 언어학적 연구와 언어 교육, 언어 병리학, 구어의 분석과 활용 기술 개발과 관련된 공학 분야 등

에서 다양하고 폭넓게 이용될 수 있도록 설계되었다. 한편, 구어 말뭉치가 여러 분야에서 실제적으로 활용되기 위해서는 원시 말뭉치 외에 각 어절의 구성단위의 문법 범주에 관한 정보를 부착한 분석 말뭉치가 더 효율적이다. 이를 위해서 구축된 구어 형태소 분석 말뭉치는 독백 자료 약 40%, 대화 자료 약 60%로 구성되어 있다.

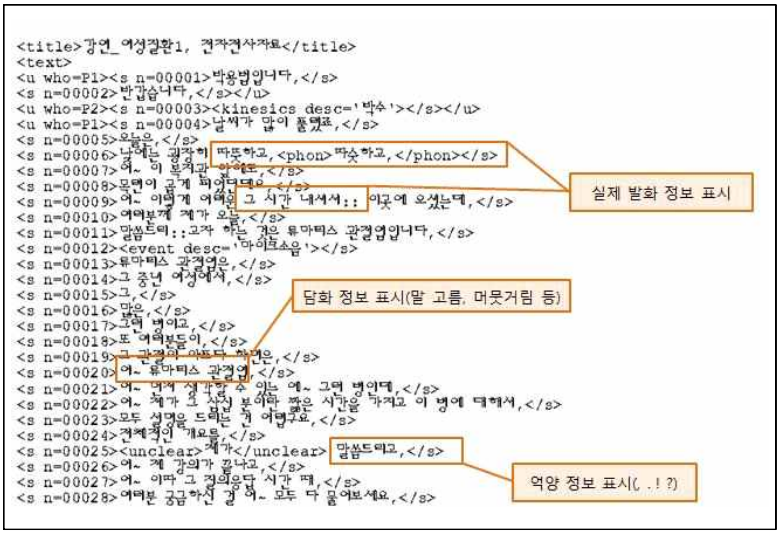
2.3. 구어 말뭉치의 특성

'세종 계획'에서 구축된 현대 국어 구어 말뭉치의 가장 큰 가치는, 실제의 자연 발화를 가능한 한 있는 그대로 전사하여 문자화한 자료라는 것이다. 실제로 발화되는 언어는 즉각적이고 일회적이기 때문에 이를 수집하여 말뭉치로 만들기 위해서는 반드시 발화를 실시간으로 녹음(또는 녹화)하여 그 음성을 전사하는 과정을 거쳐야 한다. '세종 계획'을 통해서 이 과정에서 음성 발화에 담긴 다양한 정보들이 최대한 유지될 수 있도록 전사 체계가 고안되었고 실제 전사 과정에 반영되었다.

〈그림 5〉에 보인 것처럼 실제 발화에서 나타나는 다양한 정보들이 말뭉치에 잘 드러나도록 함으로써 구어 말뭉치가 실제 음성 언어에 최대한 가까운 형태가 되도록 하는 한편, 문법 및 어휘 정보의 주석과 기계적 검색 단계에서의 이용의 용이성을 확보하기 위해서 철자법에 따라 한글로 전사하고 여기에 각종 세밀한 음성 및 발화 정보를 덧붙여 표시하도록 설계하였다.

그 몇 가지 실제적인 특성을 살펴보면 다음과 같다.

(1) 구축 과정에서 문어의 기본 단위인 '문장(각종 문장 기호를 포함하는)'을 기준으로 삼지 않고, 구어의 특징을 반영할 수 있는 '억양 단위'(억양의 바뀜을 동반하면서 한 숨에 이루어지는 발화)를 기본 단위로 하였으며, 각 단위의 경계에서 나타나는 억양 변화를 4가지로 구분하여 기호로 표시하였다.[4)]



〈그림 5〉 구어 전사 말뭉치의 특성

(2) 말할 때 흔히 나타나는 '끊어지거나 불분명한 말, 말이 한동안 끊기는 쉼' 등을 비롯하여 '웃으면서 하는 말, 박수 치면서 하는 말, 말소리 즉 실제 대화가 아닌 그 밖의 소리들'에 대한 정보를 일정한 방법을 정해서 표시하였다. 한편, 말 중간에 다른 사람이 끼어들어서 생긴 발화 중단, 다른 사람의 말이 끼어드는 발화 삽입, 둘 또는 여러 사람의 발화가 겹치는 현상 등이 일어난 발화 상의 위치와 유형을 구체적으로 표시하였다.

(3) 말을 하다가 머뭇거리거나 말을 고를 때 흔히 나타나는 '음, 이, 그, 저, 아' 등의 군소리(언어학적으로는 담화표지라고도 한다)에 '~' 기호를 덧붙여 표시함으로써('음~, 이~, 그~, 저~, 아~'), 같은 형태이면서 뜻을 가지고 쓰이는 경우와 구별되게 하였다.

(4) 실제 발화에서 자주 나타나는 음운 및 음성적 특징이 드러나도록 하기 위하여, 소리가 줄어드는 현상(축약)이나, 소리가 바뀌는 현상 등에

4) 억양과 쉼, 통사적 정보 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.

서 보이는 이형태, 표현적 장음 등의 음성 정보를 충실히 표시하였다.

또한, 구어 말뭉치에 우리의 일상적인 언어생활을 대표할 수 있을 만큼 다양한 발화가 포함될 수 있도록 하기 위해서, 자료 수집 단계부터 구어의 분류 체계를 상세히 설정하여 균형적인 구어 자료 확보에 노력하였다[5)]

아울러, 구어의 수집 과정에서는 구체적인 발화 상황에 대한 정보(녹음 장소, 녹음 시간, 담화 유형 등) 및 발화자의 개인 정보(연령, 성별, 출신 등)를 기록하였고, 언어 연구 및 분석 자료로 실제 활용할 수 있도록 발화자에게 자료 이용에 관한 동의서를 받았다. 다만 사업 초기에는 이에 대한 인식이 거의 낮았기 때문에, 사업의 수행 과정에서 문제를 발견하게 되어 문제 해결에 적지 않은 어려움을 겪을 수밖에 없었다.

3. 다국어 병렬 말뭉치의 성과와 특성

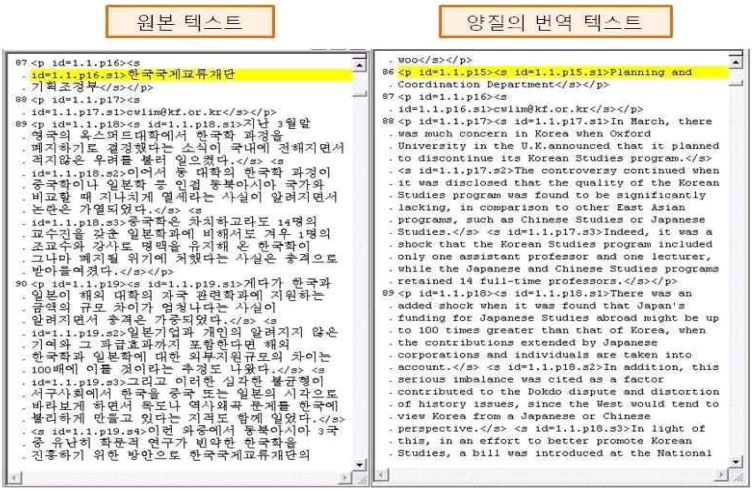
3.1. 병렬 말뭉치 구축의 목적과 활용

오늘날 문화와 정보의 국제적인 교류가 급증하면서 자동 번역과 통역, 기계 번역, 번역 검증 시스템 등의 연구를 비롯해서 언어 교육, 비교 언어학, 대조 언어학 등의 분야에서의 연구에 이르기까지 폭넓은 분야에서의 연구와 이를 뒷받침 할 수 있는 소프트웨어의 개발을 할 때 병렬 말뭉치가 적극적으로 수집, 구축되어 활용되고 있다.

병렬 말뭉치란, 〈그림 6〉에서 보는 것과 같이, 동일한 내용의 텍스트를 둘 이상의 언어로 대응시켜 구성한 자료를 뜻한다. 병렬 말뭉치는, 그림에 나타난 것과 같이, 동일한 내용의 원본과 번역 텍스트를 입력하여 전산화한 후, 이들 텍스트의 문장들을 비교하면서 서로 대응되는 문장을 찾

5) 이 글의 〈표 3〉에 제시한 분류 체계와 함께 서상규·김형정(2005:19~26)을 참조 바람.

아 그 대응 관계를 나타낸 데이터베이스를 만들어 나가는 방식으로 만들어진다.



〈그림 6〉 병렬 말뭉치의 텍스트(한국어-영어)

'세종 계획'에서는 당초 5개의 언어로 된 다국어 병렬 말뭉치의 구축을 목표로 하였기 때문에, 한국어·영어 병렬 말뭉치와 한국어·일본어 병렬 말뭉치 외에도 한국어·중국어 병렬 말뭉치, 한국어·러시아어 병렬 말뭉치, 한국어·프랑스어 병렬 말뭉치 등의 시험 구축과 기초 연구가 수행되었다.[6]

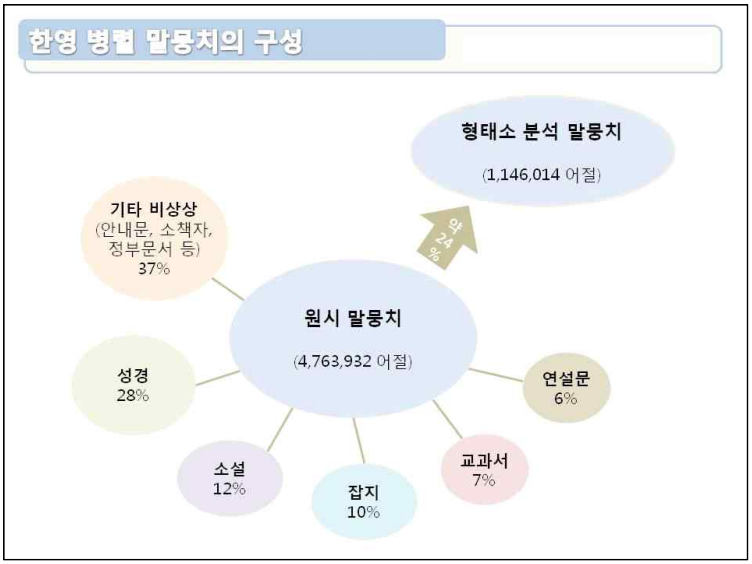
6) 물론, 원본 텍스트가 어느 언어로 되어 있느냐에 따라서, '한-영'과 '영-한'은 뜻하는 바가 달라지기는 하지만, '세종 계획'에서는 통칭하여 '한-영' 병렬 말뭉치로 부른다.

3.2. 병렬 말뭉치의 구축 현황

단계 \ 구분		1단계 (1998~2000)	2단계 (2001~2003)	3단계 (2004~2007)	합계
한·영	원시	100만	255만	115만	470만
	형태소 분석	–	52만	48만	100만
한·일	원시	–	55만	47만	102만
	형태소 분석	–	5만	24만	29만
한·중 한·러 한·불	원시	–	15만	–	15만

〈표 4〉 병렬 말뭉치 구축 현황

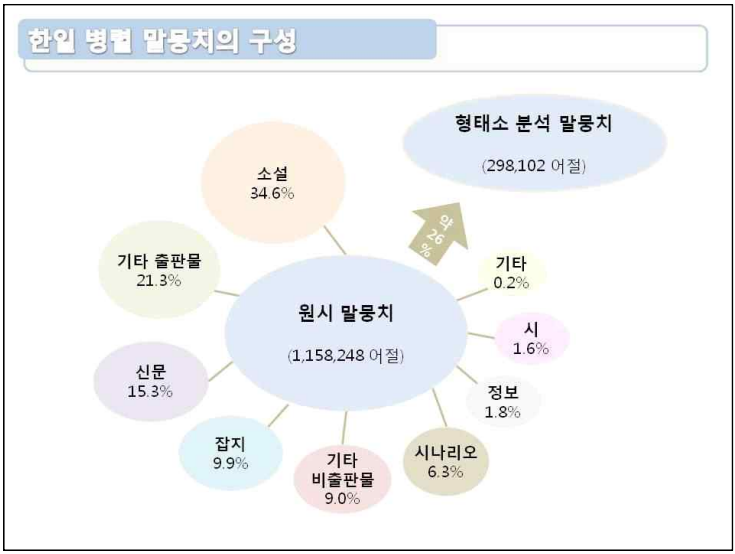
병렬 말뭉치는 텍스트의 실현 매체에 따라서 잡지, 책, 기타 출판물, 기타 비출판물로 나뉘었는데, 이에 따른 구체적인 분류와 실제로 구축된 양을 보면 다음과 같다.



〈그림 7〉 한영 병렬 말뭉치의 구성

3.3. 병렬 말뭉치의 특성

병렬 말뭉치를 구축할 때는, 원문에 대한 양질의 번역을 찾아내어 선택하는 일이 무엇보다 중요하다. 일반적으로 병렬 말뭉치 구축에서는 번역의 유형이나 품질에 대해서 크게 고려하지 않는 경우가 많지만, '세종 계획'에서는 수집 대상의 텍스트를 선정할 때에 다음의 3가지를 고려함으로써, 최대한 양질의 텍스트가 선정되도록 힘썼다.



〈그림 8〉 한일 병렬 말뭉치의 구성

첫째, 기계번역이나 문체 연구, 한국어 교육, 언어 대조 연구 등의 분야에서 두루 활용이 가능하도록 범용성이 있는 자료인가, 둘째, 실제로 언어 연구나 교육, 기계 번역 등에 활용될 수 있을 만큼 실용성이 있는 자료인가, 셋째, 이미 구축된 말뭉치의 균형성을 확보하기 위해서 부족한 장르를 보완할 수 있는 자료인가이다. 이러한 선별 과정을 거쳐서 수집된 병렬 말뭉치는, 신문, 잡지, 법률, 논문, 시나리오, 소설, 수필, 시, 성경, 매

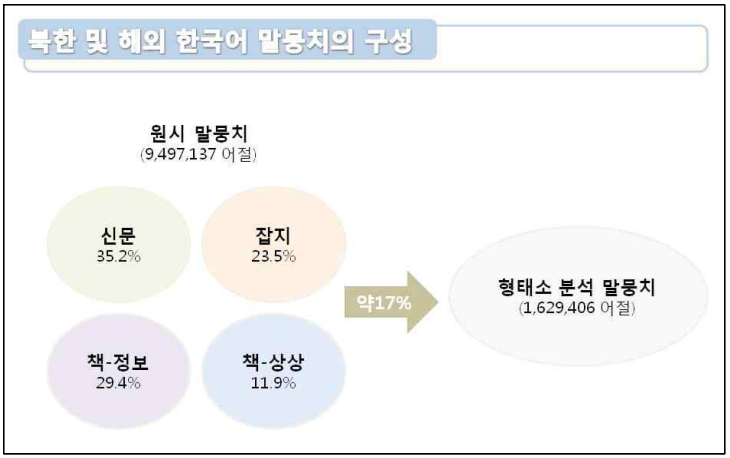
뉴얼 등 다양한 장르로 구성되었으며, 이를 통해서 특정 장르에 편중되지 않은 번역 양상을 살펴볼 수 있다.

한편, 병렬 말뭉치를 구축하면서 앞서 살펴본 장르 간의 균형성과 마찬가지로 중요하게 고려된 요소는 번역의 방향성 측면에서의 균형성이다. 번역의 방향성이란, 원본 텍스트의 언어와 대역 텍스트의 언어가 각기 무엇인가에 따라서, 즉 어느 언어가 원본 텍스트이냐를 뜻한다. 이 점을 중요하게 고려하는 까닭은 번역의 방향성에 따라서 번역 양상이 사뭇 달라질 수 있기 때문이다. 우리가 일상생활에서 흔히 경험하는 바와 같이, 원문이 외국어인 번역 문장의 문장투가 일반 문장과 달리 느껴지는 것과 같은 이치이다.

4. 그 밖의 한국어 특수 말뭉치의 성과와 특성

4.1. 북한 및 해외 한국어 말뭉치

'한국어'라고 할 때는 표준어뿐 아니라 각 지역의 방언들을 포함하며, 지역적으로 좀 더 널리 본다면 북한 및 해외 지역의 동포들이 사용하는 한국어까지 포괄할 수 있다. 그러므로 한국어의 공간적 외연을 북한 및 해외 지역으로까지 넓히고, 특히 일반적으로 개개인이 수집하거나 대규모로 전산화하기 어려워 상대적으로 접하거나 구하기 어려운 북한 및 해외 한국어 말뭉치를 구축함으로써, 우리는 한국어 연구자들이 쉽게 활용할 수 있는 연구 자료의 폭을 대폭 확대할 수 있을 것이다.



〈그림 9〉 '북한 및 해외 한국어 말뭉치'의 구성

```
1   **4NT00001          **4NT00001
34  오늘                오늘/NNG
35  위대한              위대/XR+하/XSA+ㄴ/ETM
36  수령님의            수령/NNG+님/XSN+의/JKG
37  신년사를            신년사/NNG+를/JKO
38  높이                높이/MAG
39  받들고              받들/VV+고/EC
40  새해의              새해/NNG+의/JKG
41  힘있게              힘/NNG+있/VA+게/EC
42  다그쳐나가는데서    다그치/VV+ㅕ/EC+나가/VX+는/ETM+데/NNBG+서/JKB
43  문제는              문제/NNG+는/JX
44  일군들이            일/NNG+군/XSN+들/XSN+이/JKS
45  조직지도사업을      조직/NNG+지도/NNG+사업/NNG+을/JKO
46  패기있게            패기/NNG+있/VA+게/EC
47  벌려나가는것이다    벌리/VV+ㅕ/EC+나가/VV+는/ETM+것/NNBG+이/VCP+다/EF
48  .                   ./SF
49  위대한              위대/XR+하/XSA+ㄴ/ETM
50  수령                수령/NNG
51  김일성동지께서는    김일성/NP+동지/NNG+께서/JKS+는/JX
52  다음과              다음/NNG+과/JKB
53  같이                같이/MAG
54  교시하시였다        교시/NNG+하/XSV+시/EPH+였/EPT+다/EF
55  .                   ./SF
```

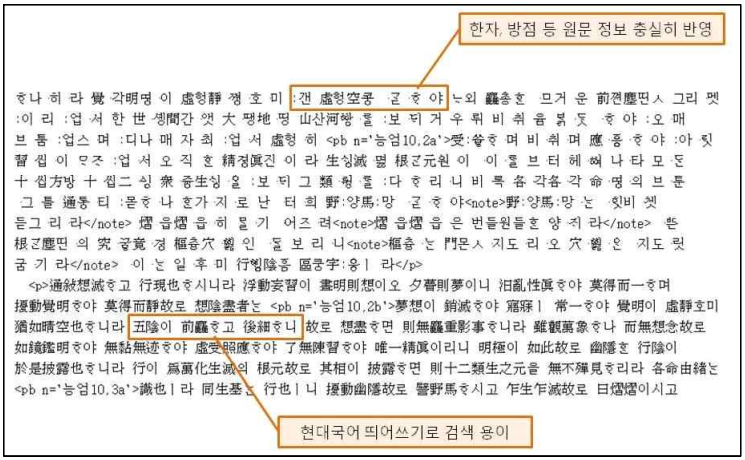
〈그림 10〉 '북한 및 해외 한국어 말뭉치' 일부

'세종 계획' 특수 자료 구축 분과에서는 북한 및 해외 지역에서 발행된 책이나 신문 등의 문어 자료를 수집, 전산화하고 가공함으로써 원시 말뭉치 약 900만 어절, 형태소 분석 말뭉치 150만 어절을 구축하였다.

4.2. 역사 자료 말뭉치

역사 자료란, 현대 국어 이전의 모든 국어 자료를 가리킨다. 당연히 역사 자료는 민족 문화 유산으로서의 가치가 크므로, 원전의 국어를 전산화하여 말뭉치로 구축하는 것 자체로도 큰 의의를 가진다. 이렇게 말뭉치로 구축함으로써, 희귀 문헌에 나타난 한국어를 전자적으로 보존할 수 있을 뿐 아니라, 일반인들이 컴퓨터를 통해서 쉽고 편리하게 옛말을 검색하고 활용할 수 있게 될 것이기 때문이다. 나아가 국어의 변천사나 방언 연구 등의 국어 연구, 국어 교육 등에도 역사 자료의 데이터베이스화는 매우 중요하다. 역사 자료에 나타난 옛 국어에 대한 이해를 통해서 현대 국어의 근원을 파악할 수 있고, 현대 어문 생활의 길잡이가 될 수 있기 때문이다. '세종 계획'에서 데이터베이스로 만든 역사 자료는 주로 15세기로부터 20세기에 걸친 자료들로 구성되어 있다.

역사 자료를 말뭉치로 구축할 때 지켜야 할 중요한 점은, 원전의 정보를 최대한 충실하게 반영하면서도 이를 검색에 용이하도록 정보화하는 것이다. 역사 자료 말뭉치에는 방점이나 한자음이 원전에 나타난 그대로 입력되어 있기 때문에 15세기 이후의 국어를 연구하기 위한 자료로서의 가치가 크다.



〈그림 11〉 역사 자료 말뭉치

한편, 역사 자료의 원전에는 대부분 띄어쓰기가 되어 있지 않지만, 말뭉치로 가공하는 과정에서는 검색과 활용의 편의를 위해 띄어쓰기를 하여 입력하였다. 역사 자료에 흔히 나타나는 이체자 역시 원전의 정보를 그대로 유지시키기 어려운 부분이기 때문에, 원본과 말뭉치의 한자 자형이 다르게 입력된 경우에는 해당 한자를 별도의 표로 저장하였다.

4.3. 전문용어 말뭉치

그 밖에도 '세종 계획' 특수자료 구축 분과에서는 '전문용어의 정비' 분과의 전문용어 표준화 사업[7]에 기반 자료를 제공하고, 전문용어의 언어학적 기초 자료를 구축할 목적으로, 각 분야의 대표적인 텍스트를 선정하여 전문용어 말뭉치로 구축하였다.

7) 카이스트의 전문용어언어공학센터(KORTERM)에서 담당하였으며, 상세한 내용은 http://korterm.kaist.ac.kr/에서 볼 수 있다.

5. 향후의 전망

이 글에서는 '21세기 세종 계획'의 특수 자료 구축 분과의 활동을 통해서 이루어진 한국어 특수 자료(말뭉치) 구축의 성과를 개괄적으로 살펴보았다. 특히 각 특수 자료들의 구축 목적, 활용 가능 분야, 구축량과 구성, 말뭉치의 특성 등을 중심으로 소개하였다.

한국어 특수 자료는 '현대 한국어 문어' 자료에 국한된 말뭉치 구축의 한계점을 극복하는 데에 중요한 초점을 두고 설계되었다. 즉, 시간과 공간적 변이, 구어와 문어의 변이를 망라하여 포함으로써, 한국어의 총체적인 언어 자원을 확보하는 데에 그 근본적인 목적을 두었다.

시간적인 면에서는 15세기로부터 현재, 공간적인 면에서는 한반도와 해외(중국, 구소련) 지역을 포함하는 문어 말뭉치가 포함되었다. 또한, 현대 한국의 자연 발화를 중심으로 한 구어 말뭉치를 대규모로 구축함으로써, 현대 국어 구어 연구뿐 아니라 문어와 구어의 비교 연구 등을 가능하게 하는 기초적인 토대를 마련하였다.

이제는 '세종 계획'을 통해서 확보된 한국어 특수 말뭉치를 토대로 하여 시간과 공간, 구어 문어의 틈새를 촘촘히 메워 나가는 노력이 필요하며, 이를 활용한 연구와 개발 또한 더욱 촉진될 것으로 기대한다.

참고문헌

국립국어원, “21세기 세종 계획 국어 특수자료 구축 연구 보고서”. 2005, 2006.

문화관광부, “21세기 세종 계획 국어 기초 자료 구축 분과 특수 자료 구축 소분과 연구보고서”. 1998, 1999, 2000, 2001, 2002, 2003, 2004.

서상규(2002), 「한국어 말뭉치의 구축과 과제」, 『한국어와 정보화』, 태학사.

서상규, 「Informatization and Use of Korean Language Data」, 『The Review of Korean Studies』 8-4, 한국학중앙연구원. 2005.

서상규, “한국어 특수 말뭉치의 구축 현황과 그 특징 —21세기 세종 계획의 성과를 중심으로—”, 「한국사전학」 제12호, 한국사전학회. 2008.

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

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

서상규·한영균(1999), 『국어정보학 입문』, 태학사.

이태영, “국어사 자료의 전산화와 21세기 세종 계획”, 「국어사학회 제15회 학술대회 발표논문집」. 2003.

이한섭, “한일 병렬코퍼스의 구축과 활용”, 「홋카이도 대학 국문학과 국제 심포지엄 발표 자료집」, 日本:北海道大. 2007.

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