한국어 전자 자료의 수집과 정리 및 활용 방안

홍윤표

단국대학교 국어국문학과 교수

1. 서론

필자는 국어 연구에서 이론과 자료가 한 수레의 두 바퀴에 비유된다고 주장한 적이 있다. 두 바퀴 중에서 어느 한 바퀴가 훨씬 더 크거나 작을 때에는 수레는 그 자리에서 맴돌고 앞으로 나아가지 못하기 때문이다. 이 비유는 이론과 자료가 다 같이 중요함을 강조한 것이지만, 그보다는 지금까지 연구 이론이 연구 자료에 비해 훨씬 더 중시되어 온 국어학계에 연구 자료도 이론 이상으로 중요한 것임을 더 강조하기 위한 것이었다.

국어 연구에서 연구 이론의 변화 과정에 대해서는 학술대회를 통해 여러 번 검토되어 왔다. 전통문법, 구조문법, 변형생성문법 등의 서구 언어학 이론이 국어 연구에 준 영향이 논의되곤 하였다. 이에 비해서 연구 자료의 수집·정리 방안이나 활용 방안에 대해서는 단 한번도 논의되거나 토의된 적이 없었다. 그만큼 연구 자료는 이론에 비해 학자들의 관심 밖에 있었다.

학문 연구가 체계적이어야 한다는 명제는 비단 이론에만 국한된 것이 아니다. 자료의 이용 및 활용도 역시 체계적이어야 한다. 왜냐하면 이론을 뒷받침할 수 있는 자료가 체계적이지 못하다면, 거기로부터 나온 이론 또한 체계

적이지 못할 것이 명약관화하기 때문이다.

그래서 비록 늦은 감은 있지만, 지금이라도 국어 연구 자료들을 체계적으로 수집·정리하여 국민들과 전문가들에게 제공함으로써 국어 인식을 제고시키고 국어 연구에 이바지하지 않으면 안 될 것이다.

이 글은 한국어 전자 자료를 수집·정리하여 이를 활용하는 방안을 모색하기 위해 쓰인 것이다. 이 글의 주제를 선명하게 하기 위하여 몇몇 용어에 대한 개념을 간략히 정의해 두도록 한다.

'한국어 전자 자료'의 '한국어 자료'는 세 가지로 해석될 수 있다. '한국어로 되어 있는 자료', '한국어를 반영하고 있는 자료', '한국어를 연구한 자료'가 그것이다. '한국어로 되어 있는 자료'가 좁은 의미의 한국어 자료임에 비하여 나머지 두 개는 이것을 포함하는 넓은 의미의 '한국어 자료'라고 할 수 있다. 이 글은 후자의 '넓은 의미의 한국어 자료'를 대상으로 한다.

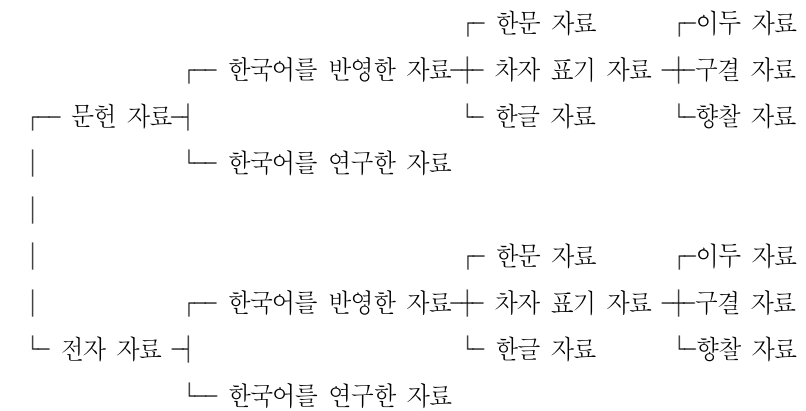
자료는 그 성격에 따라 기술의 범위가 달라진다. 자료는 형태상으로 청각 자료인 음성 자료, 그리고 시각 자료인 문자 자료와 그림 자료로 구분될 수 있다. 음성 자료는 한국어를 반영하였거나 연구한 1차적인 자료이다. 문자 자료는 음성 자료를 문자로 표기하여 놓은 것이고 그림 자료는 문자로 표기하여 놓은 문헌을 그림(이미지)으로 보여 주는 것이다. 그래서 이들은 음성 자료에 비하여 모두 2차 내지 3차 자료라고 할 수 있다. 음성 자료는 별도의 항목으로 기술될 것이기에 이 글에서는 기술 대상에서 제외하도록 한다.

자료는 전달 매체에 따라 문헌 자료, 문서 자료, 금석문 자료, 전자 자료 등으로 구분된다. 문헌 자료, 문서 자료, 금석문 등의 자료가 1차적이라고 한다면 전자 자료는 이에 비하여 2차적이라고 할 수 있다. 왜냐 하면 전자 자료들은 문헌 자료를 전산 처리할 목적으로 컴퓨터로 입력하여 놓은 자료를 말하기 때문이다. 이 글에서는 문헌 자료를 정보로서의 가치를 가지도록 가공하여 놓은 전자 자료를 기술 대상으로 한다.

문자 자료는 문자의 종류에 따라 다양하게 구분될 수 있으나, 앞에서 '한국어 자료'를 '한국어를 반영하였거나 연구한 자료'로 그 범위를 넓힌 것과 마

찬가지로, 그 표기가 한글 표기로 되어 있든, 구결이나 이두와 같은 차자 표기로 되어 있든, 한자 내지 한문으로 되어 있든 간에, 한국어를 반영하고 있다는 뚜렷한 증거만 있다면 한국어 자료라고 할 수 있다. 예컨대 향가 등이 이러한 예에 속할 것이다. 또한 문자 자료는 지나간 시기의 자료와 현대의 자료를 모두 포괄한다.

결론적으로 한국어 자료란 다음과 같은 자료들을 포함한다.



결국 전자 자료란 문헌 자료(문서 자료와 금석문 자료 포함)를 그 형태만을 달리한 것이라고 할 수 있다.

2. 전자 자료의 효용 가치

전자 자료가 문헌 자료와 그 효용 가치에서 어떠한 차이가 있을까? 전자 자료는 활용 도구로서 컴퓨터를 이용하는 것이기 때문에, 컴퓨터의 기능을 아는 것이 전자 자료의 효용 가치를 아는 첩경이 될 것이다. 전자 자료는 컴퓨터를 이용하여 활용 효과를 극대화할 수 있도록 구성된 것이어야 하는데, 이것은 곧 컴퓨터의 기능을 최대한으로 활용하는 일이다. 주지하는 바와 같이

컴퓨터는 다음과 같은 아홉 가지 기능을 가지고 있다.

① 입력(input) 기능
② 제어(control) 기능
③ 연산(arithmetic) 기능
④ 기억(memory) 기능
⑤ 출력(output) 기능
⑥ 통신(communication) 기능
⑦ 오락(entertainment) 기능
⑧ 학습(instruction) 기능
⑨ 자료 처리(data processing) 기능

전자 자료는 이 모든 기능과 연관이 있다고 할 수 있다. 이 아홉 가지 기능 중에서 전자 자료의 생산 과정에 관여하는 컴퓨터의 기능은 ①의 입력 기능이고, 유통 과정과 연관되는 기능은 ⑤의 출력 기능과 ⑥의 통신 기능이다. 그리고 나머지의 기능들은 전자 자료의 활용 단계에 관여하는 기능이다. 이러한 기능 때문에 컴퓨터는 다음과 같은 장점을 지니게 된다.

① 자료 처리의 완벽성
② 자료 처리의 신속성
③ 자료 처리의 대용량성
④ 자동적 처리
⑤ 대량의 자료를 영구적으로 기억
⑥ 많은 이용자의 동시 사용
⑦ 복합적 기능(멀티미디어 기능 : 음악, 영상, 음성 등에 대한 복합적 처리)

결국 우리가 문헌 자료를 이용하는 대신에 전자 자료를 구축하고 정리하여 이를 활용하려는 목적과 이유는 위와 같은 이득을 얻기 위한 것이라고 할 수 있다. 위와 같은 기능은 문헌 자료들이 제공해 줄 수 있는 기능과 비교하여 볼 때에 엄청난 효과를 발휘할 수 있기 때문에, 전자 자료 이용의 중요성은 시간이 흐를수록 더 증대할 것이다.

특히 21세기는 정보 사회이기 때문에 현대인들은 자신의 의사를 전달하고 전달 받는 양식을 다양화 또는 다원화하고 있다. 이러한 인간의 욕구로 말미암아 결국 문헌 자료보다는 전자 자료를 더 중요한 매개체로 인식하게 될 것이다.

3. 전자 자료 구축을 위한 사전 조사

3.1. 한국어를 반영한 전자 자료

전자 자료는 주로 문헌 자료를 가공 처리하여 입력한 것이다. 여기서 가공 처리란 자료의 수집, 분류, 재정렬, 계산, 요약, 저장 등을 포함하는 일련의 디지털화 작업을 말한다. 즉 원전 자료를 컴퓨터에서 활용할 수 있도록 하는 과정인 것이다. 이렇게 가공 처리된 전자 자료를 우리는 흔히 말뭉치(corpus)라고 한다. 전자 자료를 구축한다는 의미는 한편으로는 한국어를 반영한 자료인 말뭉치를 구축한다는 의미이다. 이 말뭉치는 다양한 기준에 따라 다양하게 분류된다. 말뭉치의 일반적인 분류와 그 종류를 보이면 다음 표와 같다.

분류기준	분 류		설 명
매체	문서 말뭉치		문서로부터 추출된 말뭉치
	음성 말뭉치		음성으로 된 말뭉치
	문자 말뭉치		문자의 글자꼴을 모은 말뭉치
부속 정보	원시 말뭉치		아무런 부속 정보를 가지고 있지 않은 말뭉치
	분석 말뭉치	문법 정보 말뭉치	단어에 부속 정보를 첨가한 말뭉치
		구문 분석 말뭉치	구문 주석을 첨가한 말뭉치
디자인 방법	균형 말뭉치		모든 장르의 문서가 균등한 비율로 포함된 말뭉치
	피라미드형 말뭉치		균형 말뭉치를 피라미드형으로 만든 말뭉치
	기회적 말뭉치		용례의 균형적 분포를 고려하지 않은 말뭉치
시대성	공시 말뭉치		어느 한 시대의 용례에 대한 말뭉치
	통시 말뭉치		각 시대의 용례에 대한 말뭉치
언어	언어의 종류	단일어 말뭉치	한 언어의 용례를 갖는 말뭉치
		이중어 말뭉치	같은 뜻을 가진 용례가 두 언어로 되어 있는 말뭉치
		다중어 말뭉치	같은 뜻을 가진 용례가 둘 이상의 언어로 되어 있는 말뭉치
	번역 여부	원문 말뭉치	외국어로 번역되어 있지 않은 원시 말뭉치
		번역 말뭉치	어느 한 언어로 번역되어 있는 말뭉치
학습	학습 말뭉치		말뭉치 분석 도구의 확률을 평가하는 말뭉치
	실험 말뭉치		말뭉치 분석 도구의 성능을 평가하는 말뭉치

이들 말뭉치들은 음성 말뭉치를 제외하고는 주로 문헌 자료를 바탕으로 한다는 공통점이 있다. 따라서 전자 자료의 구축·수집·정리는 문헌 자료에 대한 기초 조사를 선행 조건으로 한다.

문헌 자료는 그 특징에 따라 다양하게 분류되지만[1] 그중 가장 큰 분류는 역사 자료와 현대 자료로 분류하는 것이다. 현대 자료의 기점이 문제가 되기는 하지만, 대체로 자료의 양이 극도로 제한되어 분포하는 시기인 20세기 중반(특히 8.15 광복) 이전의 자료를 역사 자료로 보고, 방대한 자료를 확보할 수 있는 20세기 중반 이후 시기의 자료를 현대 자료로 보는 것이 합리적일 것으로 생각한다.[2] 따라서 지금까지 발견된 한국어 자료 중에서 가장 오래되었다고 생각하는, 414년의 '광개토대왕비'로부터 시작하여 현대까지의 한국어 자료 연표를 작성하는 것이 급선무라고 생각한다. 이 한국어 자료 연표에는 다음과 같은 사항들이 기술되어 있어야 할 것이다.

① 연도 : 간행 연도의 표시를 위한 것이다.

② 문헌명 또는 문서명

③ 문헌의 약칭 : 각 문헌이 인용될 때에 필요하다. 지금까지 남광우(1960), 남광우(1997), 유창돈(1964), 김정수(1984), 홍윤표 외(1995) 등에서 각각 제안된 약칭(또는 약호)이 있으나 아직 표준화되어 있지 않다. 이 약칭은 가능한 한 2음절 내지는 3음절로 되어 있는 것이 기억의 편의를 위하여 가장 합리적일 것이라고 생각한다. 이 약칭은 전자 자료의 파일명을 정하는 데에도 그대로 쓰일 수 있다.

④ 간지 : 대부분의 문헌이 간행 연도를 간지로 제시하고 있기 때문에 그 확인을 위하여 필요하다.

---

1) 문헌 자료의 분류에 대해서는 졸고(1997)를 참조할 것.

2) 국어사적인 면에서는 근대국어 시기가 끝나는 19세기 말까지의 자료를 역사 자료로 볼 수도 있다. 그러나 21세기에 접어든 오늘날에는 20세기 초의 자료를 현대 자료로 인식하는 사람이 적다.

⑤ 연호 : 대부분의 문헌에서 간행 연도를 연호(특히 중국의 연호)로 표시하고 있기 때문에 이 연호도 연표에 반드시 들어가야 할 사항이다.

⑥ 왕대 : 우리나라 각 왕조의 왕과 그 재위년을 표시하는 것이다.

⑦ 소장처 : 문헌 원전의 소장처뿐만 아니라 도서 번호까지도 표시하여 주는 것이 필요할 것으로 보인다. 왜냐하면 한 문헌의 이본이 많기 때문이다.

⑧ 문헌 자료의 입력 여부와 그 파일명 : 전자 자료의 수집과 정리를 위하여 절대적으로 필요한 부분이다. 이것은 텍스트 자료와 이미지 자료의 두 가지로 구분하여 조사되어야 한다.

⑨ 입력자명 : 전자 자료의 신뢰성 여부를 알 수 있도록 입력한 사람이나 입력한 기관의 이름을 밝혀 두는 것이 좋을 것이다.

이러한 내용들을 담은 '한국어 자료 연표'가 작성되어 있어야 한국어를 반영한 전자 자료의 구축·수집·정리의 계획을 체계적으로 수립할 수가 있는 것이다. 다음에 필자가 작성하고 있는 한국어 자료 연표의 일부를 예로 보이도록 한다.[3] 여기에 든 것은 훈민정음 창제 이후의 일부 예이지만, 앞에서 제시한 414년부터 연표가 되어 있어야 할 것이다.

한글 자료 연표가 작성되면 앞으로 구축해야 할 전자 자료와 정리하여야 할 전자 자료의 목록을 작성하고 이에 대한 단계적인 계획을 수립해야 할 것이다.

상당수의 문헌 자료들(특히 19세기 말까지의 역사 자료들)은 텍스트 자료로서 입력되어 있지만, 이미지 자료로 구축해 놓은 것들은 거의 없는 편이다. 설령 다른 기관(예컨대 규장각이나 국립중앙도서관 등)에서 구축해 놓은 것이 있어도 공개되어 있지 않으면 구축되어 있는 자료로 보기 힘들기 때문에, 이러한 문제까지 염두에 두고 계획이 이루어져야 할 것이다.

---

3) 필자는 이러한 연표를 이미 작성하여 놓고 수정 보완하고 있는 중이다.

연도	간지	왕	연호	문 헌 명	종류	소 장 처	입력 파일명	입력자	그림 파일명
1446	丙寅	世宗 28	正統 11	訓民正音(解例本)	漢	간송문고	KHMJU000.HWP	국	
1447	丁卯	世宗 29	正統 12	龍飛御天歌	한	가람문고	A5CA0021.HWP	세	
						규장각			
						계명대 도서관(권8,9,10)			
				釋譜詳節	한	천병식(권3)	A5CD0006.HWP (권3) A5CD0007.HWP (권6,9,11,13,19, 21,23,24)	세	
						국립중앙도서관(권6, 9, 13, 19)			
						심재완(권11)			
						개인 모씨(권20)			
						개인 모씨(권21)			
						동국대 도서관(권23,24)	seokbo20.hwp(권20)	홍	
				月印千江之曲	한	대한교과서주식회사(卷上)	A5CD0019.HWP	세	
				琴榼別給文記	吏	安東 河氏 家門			
1448	戊辰	世宗 30	正統 13	東國正韻	한	간송문고(권1, 6)			
						건국대(전질)			
1449	己巳	世宗 31	正統 14	舍利靈應記	한	고려대 도서관			
						고려대 육당문고			
						고려대 아세아문제연구소			
1450	庚午	世宗 32	景泰 1						
1451	辛未	文宗 1	景泰 2						
1452	壬申	文宗 2	景泰 3	李遇陽許與文記	吏	慶北安東周村派 眞城李氏宗家			
1453	癸酉	端宗 1	景泰 4						
1454	甲戌	端宗 2	景泰 5	鄭玉堅 朝謝牒	吏				
1455	乙亥	世祖 1	景泰 6	洪武正韻譯訓	한	고려대 화산문고(권1, 2)			
						고려대 만송문고(권9)			
						연세대 도서관(권3, 4)			
1456	丙子	世祖 2	景泰 7						
1457	丁丑	世祖 3	天順 1	雙峰寺 賜牌	吏	동국대 박물관			
				醴泉龍門寺 賜牌	吏	醴川 龍門寺			
1458	戊寅	世祖 4	天順 2						
1459	己卯	世祖 5	天順 3	月印釋譜	한	서강대 도서관(권1, 2)	A5CD0020.HWP (권15) A5CD0021.HWP A5CD0022.HWP (권4) A5CD0023.HWP A5CD0024.HWP A5CD0025.HWP	세	
						김병구(권4) (복각본)			
						동국대 도서관(권7, 8)			
						故梁柱東 舊藏(권9, 10)			
						호암미술관(권11, 12)			
						연세대 도서관(권13, 14)			
						전북 순창군 구암사(권15)			
						전남 장흥 보림사(권17)			
						강원도 홍천 수타사(권18)			
						개인 모씨(권20)			
						광흥사(권21) (복각본)			
						개인 모씨(권22) (복각본)			
						개인 모씨(권23)			
						개인 모씨(권25)			
				金潤宗原從功臣錄券	吏	규장각			
				李楨原從功臣錄券	吏	경북 안동 眞成 李氏 宗家			
				崔某原從功臣錄券	吏	규장각			

한 : 한글 문헌 吏 : 吏讀 관련 문헌 세 : 21세기 세종계획 말뭉치 국 : 국립국어연구원 홍 : 홍윤표

3.2. 한국어를 연구한 전자 자료

한국어를 연구한 연구 자료에 대한 기초 조사는 지금까지 한국어 및 한글을 연구한 연구 논저 목록을 작성하는 일일 것이다. 이 연구 논저는 편의상 세 가지로 구분할 수 있다. 단행본, 논문집에 게재된 논문, 학위 논문(석·박사)이다.

단행본은 각 연구 분야별로 조사하여 목록화할 필요가 있다. 그리하여 한국어 연구 단행본 목록을 다음과 같은 형식으로 만들어 두어야 한다.

연구 분야	저 자	책 이름	발행 연도	출판사	입력 파일명	
					이미지 파일	텍스트 파일

이 목록은 이들 단행본들이 디지털화되어 있는지를 확인하기 위한 것이다. 그런데 지금까지 국어 및 한글을 연구한 단행본 목록을 작성하는 일은 대단히 어려운 일이다. 국어사 연구 자료 목록은 그 자료가 한정되어 있어서 그 작성이 수월한 편이지만, 연구 논저 목록은 그 양이 방대하여서 작성이 수월치 않기 때문이다. 그래서 이 단행본 목록 작성은 지금까지 국어 연구에 크게 기여했다고 평가되는 연구 업적을 중심으로 하여 작성할 필요가 있다. 각 연구 분야(예컨대 음운론, 형태론, 통사론, 의미론 등등)에서 연구를 진행하는 학자들은 대부분이 자기 분야의 중요한 연구 업적들을 목록화하여 가지고 있거나 또는 그 문헌을 직접 소장하고 있기 때문에 큰 고통을 겪지 않고 목록을 작성할 수 있을 것으로 생각된다.

지금까지 간행된 국어 관련 논문집들도 기초 조사가 이루어져야 한다. 다음과 같은 목록을 만들고 확인하는 작업이 이루어져야 한다.

학술지명	학회명	세부 분야명	권·호수	입력파일 이름	
				이미지 파일	텍스트 파일
			1권 1호		
			1권 2호		

그런데 이 학술지 목록 작성도 수월치 않다. 왜냐하면 그 종류와 숫자가 만만치 않기 때문이다. 그리하여 전자 파일로 만들 자료도 제한하지 않으면 안 된다. 예컨대 전국 규모 학술지와 각 대학의 교내 학술지를 대상으로 하되, 학부생이나 대학원생들만의 논문이 실리는 학회지는 제외하는 것이 좋을 것이다.

석·박사 학위 논문 목록도 작성해 두어야 하는데, 이 목록은 국립중앙도서관 홈페이지에서 지원받아 작성할 수 있다. 그 목록 양식은 다음과 같은 것이 좋을 것이다.

분야명	필자명	논문 제목	연도	학위별	학교	입력 파일명	
						이미지 파일	텍스트 파일

4.. 전자 자료의 구축 방법

전자 자료의 구축 계획이 이루어지면 아직까지 입력되지 않은 문헌 자료를 전자 자료로 만들어야 한다. 이러한 자료 구축의 방법은 전자 자료의 종류에 따라 달라지게 된다.

4.1. 한국어를 반영한 자료의 구축

한국어를 반영한 한국어 말뭉치의 구축은 말뭉치의 종류에 따라 달라지게 되는데, 전자 자료로서 구축해야 할 말뭉치는 주로 원시 말뭉치와 분석 말뭉치, 그리고 번역 말뭉치이다. 그리고 이들 중에서 선택하여 구성된 말뭉치가 균형 말뭉치이다. 그래서 여기에서는 주로 원시 말뭉치를 중심으로 하여 기술하도록 한다. 원시 말뭉치는 가장 기본적인 말뭉치이기 때문이다. 그리고 다른 말뭉치에 대해서는 간략히 기술하도록 한다.

4.1.1. 원시 말뭉치(raw corpus)

원시 말뭉치는 기본적인 전자 자료이다. 이것은 이 자료에 대한 출전, 저자, 용량 등의 기본 정보를 나타내는 헤더(header)와 본문(text)으로 구성되어 있다.

(1) 헤더

자료의 공유를 위해서는 헤더를 붙이는 양식이 표준화되어 있어야 하는데, 현재 문화관광부에서 시행하고 있는 '국어 정보화 중장기 발전 계획'인 '21세기 세종계획'에서 마련한 헤더의 표준 양식이 있다. 다음에 '21세기 세종계획'에서 마련한 표준안에 따라 이루어진 헤더의 일례를 보이도록 한다.[4)]

```
<!DOCTYPE tei.2  SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEI.corpus "INCLUDE">
    <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
    <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
  <teiHeader>
  <fileDesc>
  <titleStmt>
  <title>청어노걸대, 전자 파일</title>
  <author>?</author>
  <sponsor>대한민국 문화관광부</sponsor>
  <respStmt><resp>국립국어연구원 말뭉치 입수, 표준화, 헤더 붙임</resp>
    <name>국립국어연구원</name>
  </respStmt>
  </titleStmt>
  <editionStmt>
  <edition><date>1996/08/12/</date>전산입력</edition>
  </editionStmt>
   <extent>10279 어절</extent>
  <publicationStmt>
  <distributor>국립국어연구원</distributor>
  <idno>a9cf0001.hwp(Kceng000.hwp)</idno>
  <availability><p>배포 불가</p></availability>
  </publicationStmt>
  <notesStmt>
  <note><p>이 텍스트는 국립국어연구원 작업 내용 유지</p></note>
  </notesStmt>
  <sourceDesc>
  <bibl><author>?</author>
    <title>청어노걸대</title>
```

4) 이 헤더에 사용된 각종 태그 부호에 대해서는『21세기 세종계획 국어 기초 자료 구축』분과의 1998년도 보고서를 참조하기 바란다.

```
    <pubPlace></pubPlace>
    <publisher>?</publisher>
    <date>철종</date>
  </bibl>
  </sourceDesc>
  </fileDesc>
  <encodingDesc>
  <projectDesc><p>21세기 세종계획 1차년도 말뭉치 구축</p>
  </projectDesc>
  <samplingDecl><p>파일 변환 정보 없음</p>
  </samplingDecl>
  <editorialDecl><p>21세기 세종계획 말뭉치 문헌 입력 지침에 따름</p>
  </editorialDecl>
  </encodingDesc>
  <profileDesc>
  <creation><date>철종</date></creation>
  <langUsage>
    <language id=KO usage=99>한국어, 고어</language>
  </langUsage>
  <textClass>
    <catRef scheme='SJ21' target='P9CF'>역사 자료: 19세기, 언해/번역 자료,
역학서류</catRef>
  </textClass>
  </profileDesc>
  <revisionDesc>
<change>
        <date>1996/08/12</date>
        <respStmt>
            <resp>입력자</resp><name>△△△</name>
        </respStmt>
        <item>1장~8장</item>
    </change>
    <change>
        <date>1996/08/20</date>
        <respStmt>
            <resp>교정자</resp><name>□□□</name>
        </respStmt>
        <item>교정</item>
    </change>
    <change>
        <date>1998/10</date>
        <respStmt>
            <resp>프로젝트 책임자</resp><name>○○○</name>
            <resp>연구원</resp><name>▽▽▽</name>
            <resp>프로그래머</resp><name>◇◇◇</name>
        </respStmt>
        <item>파일 변환, 세종 21 프로젝트 헤더 붙임, 마킹</item>
    </change>
  </revisionDesc>
  </teiHeader>
```

그런데 이 표준 양식 중에서 서지 사항을 표시하는 항목의 기술 내용이 정밀하지 않아서, 특히 이본이 많은 역사 자료는 그 서지 정보의 기술이 불완전한 편이다. 이 점만 보완한다면 21세기 세종계획에서 마련한 헤더의 표준 양식은 거의 완벽하다고 할 수 있다. 그래서 옛 문헌에 대한 서지 정보는 별도로 마련하는 것이 좋을 듯하다.

(2) 옛 문헌의 서지 정보

옛 문헌에 대한 서지 정보의 기술은 현대 문헌과는 다르다. 옛 문헌에 대한 서지 정보의 기술에 포함될 내용은 다음과 같다.

冊名	所藏處	圖書番號	板種	刊行年度
刊記	內賜記	刊行處	序文	跋文
卷·冊數	圖板 有無	冊匡	板匡	四周邊
界線有無	表紙題	內紙題	版心題	版心魚尾
行·字數	註	裝幀	紙質	印
影印本 有無	參照事項	關係文獻	參考文獻	

그러나 이러한 옛 문헌의 서지 정보를 헤더 속에 포함시키는 일은 번거로운 일이다. 따라서 이러한 서지 정보는 텍스트 자료로서 입력된 파일 속에 포함시키는 것보다는 오히려 이미지로 만들어 놓은 파일의 앞에 넣어 그 자료의 성격을 파악하도록 하는 것이 좋을 것이다.

(3) 본문 입력 양식

본문의 입력 양식은 전자 자료의 양식에서 신중하게 고려해야 할 부분이다. 컴퓨터로 이 자료들을 검색하여 활용하는 부분이기 때문이다.

본문의 입력 방식은 크게 두 가지로 구분된다. 하나는 원문의 구조와 형식까지도 그대로 입력하는 것이고, 또 하나는 원문의 표기나 방점 등은 그대

로 반영하되 형식은 가공하여 처리하는 것이다. 원문의 형식에 충실하게 입력한다면, 옛 문헌 자료의 입력 파일은 띄어쓰기가 되어 있어서는 안 된다. 왜냐 하면 옛 문헌에는 대부분이 띄어쓰기가 되어 있지 않기 때문이다. 그리고 행의 바꿈도 원문에 그대로 따라야 한다. 그러나 이러한 입력 방식은 거의 무의미하다. 왜냐하면 이것은 이미지로 처리한 자료와 다르지 않기 때문이다. 그래서 가공 처리하지 않으면 안 된다. 21세기 세종계획에 따라 입력된 옛 문헌의 본문은 다음과 같다. 이 입력 자료는 국립국어연구원에서 표준국어대사전을 편찬하기 위하여 입력해 놓은 자료를 단지 후처리 과정만을 거쳐 공개한 것이기 때문에 현대국어의 입력 양식과 차이를 보인다.

```
<text>
<body>
<pb n='법화2, 174a'>
<p>妙法蓮華經 新解品第4</p>
<pb n='법화2, 174b'>
<p>信解ᄂᆞᆫ 喩說 듣ᄌᆞ오ᄆᆞᆯ 因ᄒᆞ야 信으로 드러 法要ᄅᆞᆯ 알씨라 알ᄯᅴ 法說一周에 身子
<pb n='법화2, 175a'>ㅣ</p>
<p>喩品 처ᅀᅥᆷ메 領悟ᄒᆞ야ᄂᆞᆯ 부톄 喩品에 述成ᄒᆞ샤 記 주시고 喩說一周에</p>
<p>四大弟子ㅣ 이 品에 領悟ᄒᆞ야ᄂᆞᆯ</p>
<p>부톄 藥草品에 述成ᄒᆞ시고 授記品에 記 주시니 그러나</p>
<p>大迦葉이 爲頭 머릿 弟子ㅣ로ᄃᆡ 領悟ㅣ 身子애셔 後ᄂᆞᆫ</p>
<p>이 經은 二智ᄅᆞᆯ 노겨 어울우논디라 身子ㅣ 當ᄒᆞᆫ</p>
<p>機ㄹᄊᆡ 몬져 領悟ᄒᆞ고 諸大弟子ᄂᆞᆫ 다 안해 ᄀᆞᄎᆞ고</p>
<p>밧긔 現ᄒᆞ논디라 根이 中下ㅣ 아니며 아로미 先後ㅣ</p>
<p>업건마ᄅᆞᆫ 法化 돕소와 펴믈 爲ᄒᆞᆯ ᄊᆡ 次第로 펴 버리니</p>
<pb n='법화2, 175b'>
<p>그ᄯᅴ 慧命 須菩提와 摩訶迦旃延과 摩訶迦葉과</p>
```

이 전자 파일은 다음과 같은 특징을 가지고 있다.

① 옛 문헌 원전의 표기인 세로쓰기를 가로쓰기로 가공하여 놓았다.

② 입력자의 개인적인 기준에 따라 띄어쓰기를 하였기 때문에, 일정한 규

칙에 따라 입력되어 있지 않다.

③ 대부분, 원문이나 한자음은 입력하지 않고 언해문이나 번역문만을 입력하여 놓았다.

④ 방점을 입력하여 놓은 자료가 흔하지 않다.

⑤ 옛 문헌의 입력에는 많은 한자가 필요한데, 문서작성기에서 처리하지 못한 한자가 상당수 있다.

⑥ 거의 모든 옛 문헌 한국어 입력 자료는 원시 말뭉치(raw corpus)의 상태이다. 반면에 현대국어 자료는 그 자료에 형태 통사적 주석을 가하여 단어 구성 정보와 품사 정보를 제공하여 주는 소위 주석 말뭉치(annotated corpus)가 상당수 구축되어 있다.

⑦ 입력된 문서 자료는 문헌에 대한 정밀한 서지 정보를 제공하여 주지 않고 있다. 단지 입력자나 입력기의 이름, 입력일자, 그리고 입력문헌명과 간행 연도 등만이 기재되어 있다.

⑧ 입력된 자료의 파일명이 입력자마다 달라서 혼돈을 일으키고 있다. 예컨대『두시언해』초간본 1권을 입력한 자료라면 어느 입력자는 dusi1.hwp로, 어느 입력자는 dseh0001.hwp 등으로 입력되어 있어서, 통일이 되어 있지 않을뿐더러 그 파일 이름만 보아서는 그 자료가 어떤 성격의 자료인지를 알 수가 없다. 즉 초간본인지, 중간본인지, 그리고 이 자료가 어느 시기의 자료인지, 또는 어느 성격을 지닌 문헌인지도 알 수 없게 되어 있다.

⑨ 문헌의 장차 표시 방법이 통일 또는 표준화되어 있지 않다. 즉 문헌의 장차 표시를 〈 〉 안에다 할 것인지, 아니면 ( ) 안에 할 것인지가 통일되어 있지 않다. 21세기 세종계획에서는 〈 〉 안에 표시하고 있지만, 문헌의 약호와 장차 사이에 어떠한 기호를 넣을 것인지에 대해서는 정해 놓지 않았다. 예컨대『청어노걸대』의 장차를 표시하기 위하여서는 〈청노,1a〉, 〈청노:1a〉, 〈청노;1a〉 등을 쓸 수 있는데, 이 중에서 어느 것을 선택할 것인지도 결정되어 있지 않다. 또한 권수와 장수 표시에 한자 숫자, 로마 숫자, 아라비아 숫자를 쓸 것인지도 결정되어 있지 않다.

이들 입력 자료들에 대한 문제점 및 그 해결 방안을 제시하면 다음과 같다.

1) 띄어쓰기의 문제

띄어쓰기가 되어 있지 않은 옛 문헌도 현대의 '한글맞춤법'에 의거하여 띄어서 입력하고 있다. 그러나 현대국어와는 다른 면이 많아서 띄어쓰기의 기준을 그대로 적용하기가 수월치 않다. 옛 문헌의 띄어쓰기 기준은 21세기 세종계획에서 정한 바가 있는데, 그것을 간략히 소개하면 다음과 같다.

① 일반 원칙

㉠ 고전 자료(언해자료)의 띄어쓰기는 기본적으로 현행 맞춤법 규정을 준용한다.

㉡ 단어와 단어는 띄어 쓴다.

㉢ 어미와 조사, 파생접사는 붙여 쓴다.

㉣ 합성어는 띄어쓰는 것을 원칙으로 한다.

② 특수한 어절의 띄어쓰기

㉠ 어휘화된 '어간+접사' 결합체는 띄어쓰기를 하지 않는다(예 : 이런ᄃᆞ로).

㉡ 관형형 어미가 개재된 합성어는 이들을 어휘화한 것으로 다루어 한 단어로 처리하며 따라서 띄어쓰기를 하지 않는다.

㉢ 문법화된 명사+조사 결합체는 띄어쓰기를 하지 않는다.(예 : 그에, 손ᄃᆡ)

㉣ 관형형 어미+불완전명사 'ㅣ'로 구성된 어절은 띄어쓰기를 하지 않는다.

㉤ 어미화된 의존명사는 띄어쓰지 않는다. 예컨대 '-ㄹ씨'는 띄어쓰지 않는다.

㉥ 어원이 불명확하거나 하나로 굳어져서 분석하기 어려운 것들은 하나의 단어로 인정하여 띄어쓰기를 하지 않는다.

이 기준은 매우 일반적인 기준이기는 하지만, 이 원칙을 대원칙으로 하고 세부적인 내용까지도 고려하여, 옛 문헌의 띄어쓰기 규정을 마련할 필요가 있다.

이 규정은 현대의 한글 맞춤법처럼 규범으로서의 법적인 효력은 없어도 표준 화안으로서의 기능은 매우 클 것으로 생각한다.

2) 한자음 입력의 문제

옛 문헌을 입력할 때에 한자음을 함께 입력하여야 할 것인가 말 것인가 하는 것은 옛 문헌을 입력해 본 경험이 있는 사람에게는 한 번쯤 고민을 해 본 문제일 것이다. 지금까지 상당수의 한자음 병기 문헌들이 입력되었지만 대부분이 한자음을 무시한 채 입력되어 있는 실정이다. 그러나 원칙적으로 한자음은 입력되어야 한다. 왜냐하면 한자음 표기 부분도 한국어의 한 부분이기 때문이다.

옛 문헌에는 한자음이 한자의 아래에 병기되어 있어서 옛 문헌의 형식대로 입력한다면 '世宗御製訓民正音'은 '世솅宗종御엉製졩訓훈民민正졍音흠'과 같이 입력될 것이다. 이러한 형식으로 입력된 자료는 원문을 충실히 반영한 것이기는 하지만 전자 자료로서는 바람직한 입력 형식이라고 할 수 없다. 왜냐 하면 검색 방법에 어려움이 있기 때문이다. 즉 '訓民正音'을 검색하고자 할 때에는 검색어를 '訓民正音'으로 하지 못하고 '訓훈民민正졍音흠'으로 할 수밖에 없는데, 이렇게 되면 한자음이 표기되어 있는 문헌과 한자음이 병기되어 있지 않은 문헌을 동시에 검색할 수 없게 된다. 가장 합리적인 입력 형식은 한자로 써 놓은 단어의 오른쪽 괄호 안에 그 한자음을 입력하여 두는 방법일 것이다. 즉 '訓民正音(훈민졍흠)'과 같은 형식으로 입력되어 있으면 검색에서 문제가 발생하지 않는다. 또한 체언과 용언 어간의 조사나 어미와의 연결체를 검색하고자 할 때에는 매크로 방법을 이용하여 괄호 안의 한자음을 한꺼번에 지우고 나서 검색할 수도 있다. 그리고 지금까지 입력된 많은 전자 자료 중에서 '世솅宗종御엉製졩訓훈民민正졍音흠'의 형식으로 입력된 것은 프로그램을 이용하여 일정한 형식으로 재구성할 수 있을 것이다.

3) 방점 표시의 문제

방점 표기 문헌을 입력할 때에도 방점을 입력하지 않는 것이 지금까지 옛 한글 문헌을 입력하면서 이루어진 잘못된 관행이다. 이 방점도 반드시 함께

입력이 되어야 한다.

방점이 표기되어 있는 문헌(주로 15세기 문헌)을 입력할 때에 이 방점을 어떻게 입력하여야 할 것인가 하는 문제는 15세기의 국어 역사 자료를 입력하려고 하는 연구자들에게 가장 큰 고민 중의 하나이다.

옛 한글 문헌에서 방점은 점으로써 한글의 왼쪽에 표기되어 왔었다. 그리하여 한자와 한자음과 한자음의 방점 표기가 동시에 이루어지면 '世·솅宗종御·엉製·졩訓·훈民민正·졍音흠'과 같은 형식의 입력이 될 것이다. 이렇게 입력된 자료들은 앞에서 언급한 한자음 표기가 되어 있는 자료와 마찬가지로 검색에 어려운 점이 많이 있다. 따라서 방점에 관심이 있는 사람들만이 이 방점 자료를 이용하고 그렇지 못한 사람들에게는 이 방점 표시를 일괄적으로 지운 후에 검색 자료로 이용할 수 있도록 하는 것이 좋을 것이다.

4) 입력 파일명의 표준화 문제

입력 파일의 이름은 알파벳을 이용하는 방법과 한글을 이용하는 방법의 두 가지가 있다. 소위 도스 파일명인 알파벳식 표기의 파일명은 모두 8자리를 사용할 수 있으며, 확장자는 3자리를 이용할 수 있다. 전자 자료들은 국내 이용자들뿐만 아니라 외국인들까지도 이용할 수 있도록 하려면 전자 파일의 이름을 정하는 기준도 어느 정도 표준화되어 있어야 할 것이다.

현재까지 전자 파일명을 정하는 기준을 정한 적이 두 번 있었다. 하나는 국립국어연구원에서 정한 것이고, 또 하나는 21세기 세종계획에서 정한 것이다.

국립국어연구원의 기준은 문헌명의 각 음절자의 초성 자음을 로마자로 표기하여 정하는 방식이다. 예컨대 『두시언해』이면 '두' '시' '언' '해'의 초성 로마자 표기인 'd' 's', 'e', 'h'를 합쳐서 'dseh'를 그 파일명으로 정하는 방식이었다. 나머지 자리는 책의 권수나 순서를 표시하였다. 예컨대 '두시언해 권6'이면 'dseh0006'이 그 책을 입력한 파일 이름이 되는 것이다.

21세기 세종계획의 기준은 매우 정제되어 있다. 그 내용을 간략히 소개하면 다음과 같다.

① 파일의 첫 자리에는 자료 코드를 표시한다. 여기에는 영문자의 알파벳을 이용한다.

② 두 번째 자리에는 시기를 세기별로 표시하는데, 숫자로 표시한다. 그리하여 1은 11세기 자료, 그리고 9는 19세기 자료를 말하고 0은 연대불명 및 연대 혼합자료를 표시한다. 그리고 20세기 자료는 A를 사용한다.

③ 세 번째 자리에는 문헌의 성격에 따른 구분을 표기하되 알파벳으로 표기한다.

④ 네 번째 자리에는 텍스트 유형에 따른 구분을 표기하는데 역시 알파벳으로 표시한다.

⑤ 5~8자리에는 일련번호를 붙인다. 고서는 권과 책의 구분이 다른 경우가 적지 않다. 즉 동일한 권이 분책된 경우나 여러 권이 한 책으로 묶이는 것이다. 이를 고려하여 한 권이 하나의 파일이 되도록 한다.

이러한 원칙에 따라 붙인 전자 파일의 이름을 예로 들어 보이면 다음과 같다.

첫째 자리	둘째 자리	셋째 자리	넷째 자리	다섯째 자리	확장자
고전 자료	시기	문헌 유형	종류	일련 번호	
P	8	B	A	0001	.HWP
고전 자료	18세기 자료	원국문본	고전 시가		

21세기 세종계획에서 입력 파일에 이름을 붙이는 기준은 매우 합리적이고 체계적이라고 할 수 있다. 물론 이 기준에 의해 붙인 파일명만 보고서는 금방 그 파일이 어떠한 문헌자료를 입력한 것인지를 쉽게 알 수는 없지만, 그래도 그 분류는 체계적이라고 할 수 있다. 앞으로 파일 이름을 붙이는 기준은 21세기 세종계획에서 정한 기준에 따르는 것이 좋을 것으로 생각한다.

5) 입력 양식의 문제

전자 자료들의 본문 입력 양식은 대부분이 다음과 같다고 할 수 있다.

<煮硝,1b>
ᄀᆞᆯᆞᆷ 모ᄒᆞᆷ이라 길 우ᄒᆡ나 혹 담 밋ᄒᆡ나 나죄 볏 ᄡᅴ고 밤의 긔운이 소사 빗치 검고 맛이 ᄆᆡ온 ᄀᆞᆯᆞᆷ이 ᄀᆡ장 아ᄅᆞᆷ답고 혹 서ᄂᆞᆯ커나 혹 ᄲᅳ거나 혹 ᄃᆞᆯ거나 혹 ᄉᆡᆫ ᄀᆞᆯᆞᆷ이 지ᄎᆞ요 오직 ᄠᅳᆫ ᄀᆞᆯᆞᆷ은 나죵의 습긔 나매 됴티 아니ᄒᆞ니라

<煮硝,2a>
ᄯᅡ 빗츨 보아 ᄀᆞᆯᆞᆷ을 맛보면 흰 ᄃᆡᄂᆞᆫ 맛이 슴겁고 검은 ᄃᆡᄂᆞᆫ 맛이 두텁ᄂᆞ니 곱은 삷흐로 그 검은 거슬 엷게 긁고 깁히 말띠니 깁히 ᄒᆞ면 ᄉᆡᆼᄀᆞᆯᆞᆷ이 섯겨 맛이 엷ᄂᆞ니라 긁어 ᄇᆞᆫ 후의 사ᄅᆞᆷ도 ᄇᆞᆲ으며 볏도 ᄡᅴ야 ᄯᅩ 두어 날이 디나면 긔운과 맛이 소사 올라 검은 빗치 스스로 나ᄂᆞ니 젼대로 긁어 ᄡᅳ면 가히 진티 아니ᄒᆞ려니와 만일 비ᄅᆞᆯ 맛나면 열나믄 날이나 익ᄀᆞᆺ 볏츨 ᄡᅴ야 디낸 후의야

<煮硝,2b>
ᄯᅩ 가히 긁어 ᄡᅳ리니라

즉 한 줄에 〈 〉 부호를 하고 그 안에 문헌의 약호와 장차 또는 쪽수를 적은 후에 그 아랫 줄이나 또는 〈 〉의 바로 뒤를 이어서 원문을 입력하되 다음 장차나 쪽수가 시작되기 이전까지는 행을 바꾸지 않고 계속 이어서 입력하는 것이다. 그 글이 현대문이라면 각종의 문장 부호가 사용되어서 글이나 문장의 단위를 굳이 표시해 주지 않아도 자연히 그 단위를 인지할 수 있겠지만, 띄어쓰기만 해 준 옛 문헌의 경우에는 그 단위를 어절별로만 인식할 수 밖에 없어서, 이 자료를 가지고 검색할 때에 검색자가 원하는 대로 용례를 추출해 내기 어렵다.

용례를 검색하는 방식에는 KWIC 방식과 KWOC 방식이 있다. 전자는 검색 단어가 문장의 가운데에 배열되는 것이고 후자는 검색 단어를 포함하고 있는 예문을 문장 단위로 추출해 주는 것이다. 위와 같은 일반적인 입력 양식으로는 KWIC 방식의 용례는 추출할 수 있지만 KWOC 방식으로 용례를 추출해 낼 수 없다. 그래서 본문 입력은 언어 단위를 인식시킬 수 있는 양식으로 되어 있어야 한다. 용례를 추출해 내는 가장 일반적인 단위는 문장일 것이다. 물론 접속어와 같은 검색 단어는 문장 단위를 넘어서 단락 단위로 검색되어야겠지만, 용례의 길이가 길어져서 우리가 원하는 바와 같이 빠르고 정확한

정보를 얻는 데 오히려 방해가 될 수 있다. 그래서 정보 추출의 가장 일반적이고 또 보편적인 언어 단위는 문장이라고 생각한다.

위와 같은 이유로 인하여 본문의 입력 양식은 컴퓨터가 문장 단위를 인식할 수 있도록 해 주어야 할 것이다. 그 방법에는 여러 가지가 있을 수 있다. 하나는 매 문장이 끝나는 곳에 마침표 등의 문장부호를 표시하여서 컴퓨터가 이것을 인식할 수 있도록 해 주는 방법이며, 또 하나는 한 줄(행)이 한 문장 단위임을 컴퓨터가 인식할 수 있도록 입력하고, 한 문장의 입력이 끝났다고 생각했을 때에 엔터를 쳐서 줄을 바꾸어 주는 방식이 있다. 가장 간편한 방법은 후자일 것이라고 생각한다. 그리하여 본문은 다음과 같은 양식으로 입력하는 것이 좋을 것이다.

<煮硝,1b>
ᄀᆞᆱ 모ᄒᆞᆷ이라.
길 우ᄒᆡ나 혹 담 밋히나 나죄 ᄇᆞᆺ 쩌고.
밤의 긔운이 소사 빗치 검고 맛이 ᄆᆡ온 ᄀᆞᆱ이 ᄀᆡ장 아ᄅᆞᆷ답고.
혹 서ᄂᆞᆯ커나 혹 쓰거나 혹 ᄃᆞᆯ거나 혹 싄 ᄀᆞᆱ이 지ᄎᆞ요.
오직 ᄲᅳᆫ ᄀᆞᆱ은 나죵의 습긔 나매 됴티 아니ᄒᆞ니라.
<煮硝,2a>
ᄯᅡ 빗츨 보아 ᄀᆞᆱ을 맛보면.
흰 ᄃᆡᄂᆞᆫ 맛이 슴겁고.
검은 ᄃᆡᄂᆞᆫ 맛이 두텁ᄂᆞ니.
곱은 삷흐로 그 검은 거슬 엷게 긁고 깁히 말띠니.
깁히 ᄒᆞ면 ᄉᆡᆼᄀᆞᆱ이 섯겨 맛이 엷ᄂᆞ니라.
긁어 ᄇᆞᆫ 후의 사ᄅᆞᆷ도 ᄇᆞᆲ으며 ᄇᆞᆺ도 쩌야.
ᄯᅩ 두어 날이 디나면 긔운과 맛이 소사 올라 검은 빗치 스스로 나ᄂᆞ니.
젼대로 긁어 쓰면 가히 진티 아니ᄒᆞ려니와.
만일 비ᄅᆞᆯ 맛나면 열나믄 날이나 익ᄀᆞᆺ 볏츨 쩌야 디낸 후의야.
<煮硝,2b>
ᄯᅩ 가히 긁어 쓰리니라.

이러한 입력 양식은 옛 문헌 입력에만 한정되는 것이 아니라 모든 텍스트 자료 입력에 꼭 필요한 것이다.

4.1.2. 분석 말뭉치

분석 정보가 부여된 말뭉치는 형태 정보를 부가한 것과 구문 정보를 부가한 것이 있다. 의미 정보를 부가한 말뭉치는 현재까지 연세 한국어 말뭉치밖에 없다. 이 분석 정보가 부여된 말뭉치는 주로 각 단어의 품사 정보와 문장 부호에 대한 처리 정보를 보여 준다. 특히 각종의 통계 분석을 위한 전처리 작업의 하나인데, 정확한 통계 처리를 위해서는 반드시 필요한 절차이다.

단어에 대한 분석 정보를 제공하기 위해서는 표준화된 태그 세트(tag set)가 필요하다. 태그 세트란 말뭉치에 대한 형태적·통사적 주석을 위한 코드의 집합을 의미한다. 오늘날 몇몇 태그 세트가 제시되어 있지만, 아직은 만족할 만한 단계에 있지 않았다. 최근에 21세기 세종계획에 의해 이루어진 어절 분석 표지 표준안이 제시되었는데, 이것은 기존의 태그 세트를 일일이 검토하여 작성된 것이어서 지금까지 제시되었던 어느 태그 세트보다도 합리적이라는 평가를 받고 있다. 그러나 이 표준안을 바탕으로 한 tagger 프로그램이 아직 나와 있지 않아서 아직은 실용화 단계에 있지 않다. 이것이 실용화되어 널리 사용된다면, 분석 말뭉치의 구축에 많은 공헌을 할 것으로 생각한다.

4.1.3. 균형 말뭉치

말뭉치를 어떤 내용을 가진 용례들로 배분할 것인가는 매우 중요하다. 가장 이상적인 말뭉치는 모든 장르의 문서가 동등한 비율로 포함된 것이라고 할 수 있지만, 실제의 언어생활에서 모든 분야의 내용들을 동등한 비율로 배분하는 것은 아무런 의미가 없다. 왜냐하면 모든 장르의 가치가 동등하다고 볼 수 없기 때문이다.

21세기 세종계획에서 1000만 어절의 균형 말뭉치는 다음과 같이 구성되어 있다.

문 어	90%	신 문	20%	사설/칼럼	30%
				정치/사회/경제/외신/북한/종합	30%
				문화/매체/생활/과학	30%
				스포츠	5%
				기타	5%
		잡 지	10%		
		책, 정보	35%	총류	15%
				교육 자료	10%
				체험 기술	15%
				인문	20%
				사회	15%
				자연	10%
				예술/취미/생활	15%
		책, 상상	20%	장편	50%
				중·단편	40%
				동화	10%
		기 타	5%		
순구어	5%				
준구어	5%				

국어 자료가 어떠한 주제로 어떠한 분류 체계에 따라 어떻게 배분되든지, 다음의 말뭉치들은 한국어의 특성상 꼭 고려되어야 할 것으로 생각한다. 즉 한국어는 지역적 특성에 따라 각종의 방언으로 구분될 수 있지만, 남한의 말뭉치와 북한의 말뭉치, 그리고 재외 동포들의 말뭉치는 단순한 방언 차이만은 아니기 때문에 이들 말뭉치들은 꼭 고려되어야 할 것으로 생각한다. 뿐만 아니라, 통시 말뭉치도 고려되어야 한다. 왜냐하면 통시 말뭉치들은 현대국어에서 발견할 수 없는 많은 국어학적인 정보를 제공하여 주기 때문이다.

4.1.4. 번역 말뭉치

번역 말뭉치는 기계번역기 및 자동통역기 개발 및 외국인을 위한 한국어 학습 모형을 만들기 위해서 필요한 것이다. 특히 자동 기계번역기의 개발을 위해서는 필수적인 말뭉치이다. 외국의 중요한 정보를 모든 국민들이 쉽게 접할 수 있게 하고, 또 우리나라의 정보를 외국에 알리기 위해서는 자동 기계번

역기의 개발이 시급한 상황이다. 지금까지 우리나라에서 개발된 기계번역기는 대부분이 단어 대 단어의 1:1의 대역 말뭉치이어서 정확한 번역율이 50%를 밑도는 수준이라는 연구보고서도 있다. 따라서 번역 말뭉치는 단순한 단어와 단어의 대역 말뭉치뿐만 아니라, 문장 대 문장의 번역 말뭉치도 포함시켜야 할 것이다. 이 연구는 대조언어학적 접근방식을 필요로 한다. 그러나 대조언어학은 본격적인 연구가 시도된 바가 거의 없는, 국어학계에서는 생소한 부분이다. 앞으로 외국 어문학자들과 함께 공동으로 연구가 진행되어야 할 부문이다.

4.1.5. 문자 말뭉치

문자 말뭉치는 문자 인식기 개발에 필요한 것이다. 한글은 글자의 수가 방대하고 문자 간의 유사성이 많아, 문자 간에 변별력이 적어서 문자인식기 개발에 어려움이 많다는 지적이 있으나 글자꼴 데이터베이스를 방대하게 구축한다면 쉽게 해결될 수 있을 것으로 예상된다. 앞으로 구축될 문자 말뭉치는 인쇄체 및 필기체뿐만 아니라, 필기체 중에서도 정서체와 흘림체 등이 포함되어야 할 것이다.

4.1.6. 이미지 자료

이상의 말뭉치는 모두 전자 자료 중 텍스트와 연관된 자료이다. 이에 비하여 이미지 자료는 문헌 자료를 그림 자료로 만들어 놓은 것이다. 이 이미지 자료는 스캐너나 디지털 카메라를 이용하여 구축해 놓을 수도 있다. 그러나 고문헌 중에서 선장본의 문헌이나, 크기가 작은 고문서 등은 이러한 방법에 의하여 이미지 자료를 구축해 놓을 수가 있지만, 두루마리 자료들(예컨대 가사 자료 등)은 디지털 캠코더를 이용하여 아예 동화상으로 만들어 놓는 방법도 있다.

이 이미지 자료의 구축 의미는 다음과 같다.

(1) 종이로 된 문헌은 종이의 수명이 한정되어 있기 때문에, 문헌은 영원

히 보존되지 못한다. 한지로 된 고문헌들은 그 종이가 알칼리성이기 때문에 그 수명이 오래지만, 양지로 된 문헌은 100년을 견디기 어려우므로 이러한 문헌의 모습을 있는 그대로 영구히 보존하기 위하여 이미지 자료의 구축이 필요하다.

(2) 입력된 자료는 오타나 실수 등으로 인하여 잘못 입력될 수도 있다. 그러나 원책과 비교해 보지 않는 한, 그 잘못을 알 길이 없다. 따라서 텍스트 전자 자료를 이용하면서 의심스럽거나 이상한 부분이 있다고 생각되었을 때, 이를 확인하기 위하여 이미지 자료를 한 모니터 상에서 두 개의 창으로 비교하여 검토할 수 있다.

이 이미지 파일을 구성하는 방안에 대하여서는 후술할 것이다.

4.2. 한국어를 연구한 전자 자료

전술한 바와 같이 한국어를 연구한 자료는 편의상 단행본, 논문, 학위논문으로 구분하였다. 이들 자료들은 그것이 단행본이든 일반 논문이든 그리고 학위 논문이든 상관없이, 그리고 그것을 텍스트 자료로 구축하든 이미지 자료로 구축하든, 한국어를 반영한 말뭉치를 구축할 때의 방식과 동일하다. 따라서 여기에서는 별도의 설명을 하지 않는다.

5. 전자 자료의 수집 방안

우리나라에는 지금까지 개인이나 공공 연구기관 등에서 구축해 놓은 많은 전자 자료들이 있다. 이 자료들이 체계적으로 수집되고 정리되어 공개된다면 한국어에 대한 관심을 높일 수 있다. 또 그 연구의 질도 높게 향상될 것이다. 그래서 이 자료의 수집과 정리는 시급한 실정에 있다. 그렇지 않아도 수많은 자료들이 인터넷을 통하여 공개되고 또 서로 교환되고 있는데, 어느 자료가 믿을 수 있는 자료인가를 검증할 수 없는 형편에 있다. 이것을 이용하려는 개

인이 검증하여 이용하기도 어렵다. 따라서 공공기관 등(예컨대 국립국어연구원이나 대학의 연구소 등)에서 이를 체계적으로 수집, 정리하고 또 관리하여야 할 것이다.

5.1. 한국어를 반영한 전자 자료의 수집

전자 자료의 수집은 세 가지가 있다. 하나는 원문 텍스트를 입력한 자료이고 또 하나는 이미지 자료이고 또 하나는 이들을 이용할 수 있는 프로그램의 수집이다.

5.1.1. 원문 텍스트 입력 자료의 수집

원문 텍스트의 대규모 말뭉치를 구축해 놓은 곳은 다음과 같다.

기 관 명	연구소 및 프로젝트	명 칭	구축 기간	어 절 수
문화관광부	21세기 세종계획	세종 말뭉치	1998년-2000년	165,492,052
연세대학교	언어정보개발연구원	연세 한국어 말뭉치	1987년-1999년	말뭉치1-9 4,300만 표준말뭉치 2,900만 특수말뭉치 2,500만 품사 표지 부착 말뭉치 180만 의미 표지 부착 말뭉치 100만
고려대학교	민족문화연구소	고려대 한국어 말모둠	1995년	한국어 말모둠 1,000만 장르별 텍스트 코퍼스 40만
한국과학기술원		과기원 코퍼스	1996년	7,158만
국립국어연구원		국립국어연구원말뭉치	1992-1999년	6,765만

21세기 세종계획을 통해서 구축된 세종 말뭉치는 한국과학기술원의 과기원 코퍼스와 국립국어연구원 말뭉치를 포함한 것이다. 세종 말뭉치는 저작권법에 저촉되지 않는 것은 거의 다 공개되어 있으나, 다른 말뭉치들은 거의 공개가 되어 있지 않다. 따라서 이들 자료들을 모두 수집하여서 이용하기란 그리 쉬운 일이 아니다. 그러나 세종 말뭉치는 연구자들에게는 비공개적으로 열람을 허용하고 있어서, 가장 수집이 손쉬운 말뭉치라고 생각된다.

이 이외에도 21세기 세종계획 중 '한민족 언어 정보화' 분야의 '한국 방언 검색 시스템 개발'을 담당한 연구진에는 다음과 같은 방언 자료집들이 전자 파일로 입력된 자료가 있어서 국어 자료로서 매우 중요한 역할을 할 것으로 생각한다. 그 목록을 보이면 다음과 같다.

1) 남한 방언 자료

① 김영태(1975).『경상남도 방언연구』(Ⅰ). 진명문화사.
② 이기갑 외(1997).『전남방언사전』. 전라남도.
③ 현평효 외(1995).『제주어사전』. 제주도.
④ 한국정신문화연구원.『한국방언자료집』(1~9)
⑤ 이상규(2000).『경북 방언사전』. 태학사.
⑥ 서울대학교(1997).『한국 방언사전』
⑦『표준국어대사전』(국립국어연구원)에 등재되어 있는 방언 관련 자료
⑧『우리말큰사전』(한글학회)에 등재되어 있는 방언 관련 자료
⑨『국어대사전』(금성사)에 등재되어 있는 방언 관련 자료
⑩ 한영목(1999).『충남 방언의 연구와 자료』. 이회문화사.
⑪ 김주석·최명옥(2000).『경주 방언 자료집』
⑫ 김주석·최명옥(2000).『경주 속담 사전』
⑬ 곽충구(1999). '함북 길주·명천 지역 방언에 대한 조사 연구' 자료

2) 북한 방언 자료

① 김병제(1980).『조선 방언사전』. 과학백과사전출판사.
② 김영배(1997).『평안방언연구(자료편)』. 태학사.
③ 김태균(1986).『함북방언사전』. 경기대학교 출판국.
④ 리운규 등(1992).『조선어 방언사전』

가장 많은 한국어 전자 자료를 구축해 놓은 곳은 아마도 각종 도서관 및 공공연구소나 기관일 것으로 생각한다. 서울대학교 규장각, 한국정신문화연구원, 민족문화추진회, 국사편찬위원회에서는 2000년도에 정보통신부의 사업으로 한국학 자료의 디지털화 작업으로 엄청난 텍스트 자료와 이미지 자료를 구

축하여 놓았다. 이들의 목록을 구하고 국가 기관 간의 긴밀한 협조를 통하여 이들 자료가 한 곳에 모일 수 있도록 한다면 매우 큰 의의를 지니게 될 것이다. 뿐만 아니라 국립중앙도서관을 비롯한 각 대학의 도서관이나, 국어 관련 단체에도 교육, 연구 등을 목적으로 하여 각종 국어 교과서 및 문학 작품(시, 소설 희곡 등)들을 대단위로 입력하여 CD로 만들어 배포하고 있다. 이들은 이미 많은 양이 공개되어 있어서 쉽게 그 자료에 접근할 수 있다. 뿐만 아니라 학자들이 개인적으로도 입력해 놓은 자료들이 많아서 이들을 널리 알려 수집한다면 엄청난 자료를 수집할 수 있을 것으로 생각한다.[5)]

또한 가장 많은 전자 자료를 가지고 있는 곳은 출판사로 생각한다. 각 출판사에서는 출판된 문헌에 대한 전자 자료가 있을 것인데, 이것은 저작권과 연관되어 있어서 이들을 수집하여 두기란 그리 쉬운 일이 아닐지 모르나, 저자의 허락을 얻어서 연구용으로 사용하기만 한다면 허락해 줄 가능성이 무척 높다고 할 수 있다.

또한 각 대학의 연구소 등에서도 각종 말뭉치를 구축하여 놓은 것으로 알고 있다.

이들은 대부분 원시 자료들이다. 그리고 아직 정밀하게 검증된 자료들이라고 할 수 없다. 따라서 어느 기관에서 이 자료를 수집한다면, 수집하는 책임자가 지속적으로 작업하여서 검증받을 수 있도록 해야 할 것이다.

5.2. 한국어를 연구한 전자 자료

한국어를 연구한 연구 자료도 상당수 입력되어 있다고 할 수 있다. 우선 연구 논저 중에서 최근의 단행본들은 텍스트 입력 자료로서 출판사에서 수집

5) 예컨대 선문대학교의 번역문학연구소(소장 : 박재연)에서는 지금까지 번안 고소설을 입력하고 이것에 대한 주석을 붙여 많은 양의 문헌을 간행하였는 바, 이들은 모두 전자 파일로 되어 있어서 국어사 연구에 큰 도움을 받을 수 있다. 부분적으로는 개인에게 공개한 적이 있다.

할 수 있다. 저자와 출판사의 동의를 구한다면 꽤나 많은 파일들을 구할 수 있을 것으로 생각한다. 따라서 전자 자료를 수집하는 기관에서 심혈을 기울인다면 많은 양의 단행본 파일을 구할 수 있을 것이다.

일반 논문집들은 최근에 발행된 논문집들을 텍스트 입력 자료로 각 학회의 홈페이지에 공개한 경우가 많다. 그러나 이전의 논문집들은 입력해 놓은 것이 없기 때문에, 주로 영리를 목적으로 하는 기업에서 이를 이미지 파일로 만들어 검색하여 이용할 수 있도록 만들어 판매하고 있다. 그러나 이들은 가격이 비싸서 개인이 사용하기는 어려운 형편이다. 대부분이 도서관 등에서 구입하여 CD-net를 통하여 검색이 가능하도록 하였다.

지금까지 이러한 자료로 내 온 논문집은 국어학 관련 학회지(學會誌)만 약 30개가 된다.

5.3. 기타 전자 자료

프로그램도 엄밀히 말하면 전자 자료이다. 특히 이것이 '한국어'와 직접적인 관련이 있는 것이 많을 것이다. 이러한 프로그램은 몇 가지로 분류될 수 있다. 그것을 보이면 다음과 같다.

① 한국어 학습 프로그램　② 한글 학습 프로그램
③ 한자 학습 프로그램　④ 한국어 처리 프로그램
⑤ 한글 글꼴 자료

한국어 학습 프로그램, 한글 학습 프로그램, 한자 학습 프로그램 등은 아직까지 체계적으로 수집된 적이 없다. 전자상가에 가서 우선 눈에 띄는 대로 구입한다면 수십 종은 구할 수 있을 것이다. 한글 글꼴 자료는 한국글꼴 개발원에서 간행한 글꼴 98, 글꼴 99, 글꼴 2000에 그 목록과 함께 설명이 되어 있어서 자료의 수집에 많은 도움을 줄 것이다. 한국어 처리 프로그램은 '활용'에서 설명될 것이다.

6. 전자 자료의 정리 방안

전자 자료들이 수집되었으면 한 가지 문헌 자료에 대하여 어떠한 전자 자료들이 있는지를 조사해야 할 것이다. 그래서 한 문헌 자료에 대하여 다음과 같은 내용을 담아 하나의 CD ROM으로 만들어 보관·관리·배포하여야 한다.

한 문헌의 자료집(CD 한 장)에 포함되어야 할 내용을 들어 보면 다음과 같다.

① 한 문헌의 이미지 파일
② 원문 Text 자료
③ 각 자료에 대한 해제
④ 각 자료의 소장처 및 도서 번호
⑤ 각 자료의 영인본 목록
⑥ 각 자료에 대한 연구 논저 목록
⑦ 각 자료의 용례 사전
⑧ 각 자료에 대한 연구 논저 중 중요한 논문의 원문
⑨ 검색 프로그램

이들 하나하나에 대하여 구체적인 사항을 제시하면 다음과 같다.

6.1. 이미지 파일

① 자료를 한 면씩 디지털 카메라로 찍거나, 스캐너로 스캔을 하거나 또는 한 화면으로 처리가 불가능한 자료(예컨대 두루마리 자료)는 디지털 캠코더로 촬영하여 동화상으로 만든다.

② 흑백으로 할 것인가, 컬러로 할 것인가 하는 결정은 문헌 자료의 원본 촬영 가능성 여부에 따라 결정하도록 한다. 가능하면 컬러로 찍는 것이 당연한 것이지만, 문헌 원본을 직접 촬영할 수 없는 경우에는 영인본을 대상으로 하여 촬영할 수도 있어야 한다. 왜냐하면 컬러로 하면 눈으로 원본을 본 듯이 확인하기는 수월해도, 이것을 출력하여 선명하게 보려는 사람에게는 다시

컴퓨터를 조작해야 하는 번거로움이 있기 때문이다. 그리고 흑백으로 하면 복사를 해서 찍는 것이 더 효과적이며, 컬러로 했을 경우에는 지금까지 흑백 마이크로 필름을 이용하여 복사하여 놓은 자료들이나 일반 복사물을 이용할 수가 없다.

③ 각 이미지 파일은 JPEG(확장자 이름 jpg)나 GIF 파일로 하는 것이 좋으나, 흑백인 경우에는 PCX 파일도 무난하다. 그러나 가장 일반적인 것은 JPEG 파일로 생각된다.

④ 각 이미지 파일의 이름은 문헌의 약자(예컨대 『春香傳』은 '春香')와 출전의 각 張의 앞뒤를 밝혀 적도록 한다. 그 방법은 여러 가지가 있다. 예를 들어 첫 장의 앞뒷면을 표시하기 위해서

① '춘향 1a, 춘향 1b' 식으로 하는 방법
② '春香 1a, 春香 1b' 식으로 하는 방법
③ '춘향 1앞, 춘향 1뒤' 식으로 하는 방법
④ '春香 1앞, 春香 1뒤' 식으로 하는 방법
⑤ chun 1a, chun 1b 식으로 하는 방법

이 중에서 ①의 방법이 좋을 것으로 생각한다.

6.2. 원문 Text 자료

① 원문 텍스트 자료는 한글 word 2000과 한글 워디안, 이 두 가지의 문서작성기로 입력한 것이거나, 이 중에서 하나를 선택하는 것이 좋다. 즉 한글 word 2000과 한글 워디안 ('훈글'의 새 버전)의 두 가지이다. 왜냐 하면 한글 word 2000은 외국인을 위해서, 그리고 워디안은 국내 사람들을 위해서 필요하기 때문이다. 그리고 훈글의 이전판(예컨대, 훈글 3.0b나 훈글 96, 훈글 97)은 한글 word 2000으로는 '불러내기'가 가능하지만, 한글 word 2000으로 입력한 것은 훈글 3.0b나 훈글 96, 훈글 97 등으로는 '불러내기'가 가능하지 않으며, 또한 문자 set가 동일하지 않아 호환이 되지 않는다. 그러나 '한글과 컴퓨터사'의 '워디안'은 이 word 2000과 완전히 호환된다. 그러나 '훈민정음

오피스 2000'은 이들과 아직 호환이 되지 않아서 이 프로그램은 이용할 수가 없다.

그런데 앞으로의 또 한 가지 문제는 '윈도'를 사용하지 않는 사람들, 즉 '리눅스'를 사용하는 외국인이나 국내인들에 대해서는 어떻게 대처하여야 할지는 미지수다.

② 원문 텍스트 자료는『21세기 세종계획』의 header를 가지도록 입력한다.

③ 원문 텍스트 자료는 다음과 같은 구조를 가지는 것이 좋을 것이다.

㉠ 출전은 〈 〉 속에 쓰고 장차는 쉼표 뒤에 쓴다. 출전과 장차는 띄어쓰지 않도록 한다.(예 : 〈춘향,1a〉). 단, 〈 〉는 문자판에서 입력하는 것으로 한다.

㉡ 이 출전이 앞에 나오고 그 뒤에 본문이 나오도록 한다.

예 :

<심청,01a>
화셜 ᄃᆡ명 셩화 년간의 남군 ᄊᆞ히 일위 명ᄉᆡ 이스되 셩은 심이오 명은 현이니 본ᄃᆡ 명문거족으로 공의게 이르러는 공명의 유의치 아니 ᄒᆞ여 일ᄃᆡ명읫 되엿고
<심청,01b>
흥진비ᄅᆡ는 고금샹ᄉᆡ라 졍시 홀연 득병ᄒᆞ여 맛참ᄂᆡ 셰상을 ᄇᆞ리니 공이 크게 비도ᄒᆞ여 녜를 갓초와 안장ᄒᆞ고 녀ᄋᆞ를 품고 듀야 슬허 ᄒᆞ이

이러한 구조를 가지도록 함은 이미 나와 있는 프로그램들을 활용하기 위해서다.

㉢ 모든 텍스트 자료는 그것이 원 문헌에는 띄어쓰기가 되어 있지 않아도, 띄어서 입력하는 것을 원칙으로 한다. 단, 띄어쓰기의 원칙에 대해서는 국립국어연구원에서 정한 원칙에 따르는 것이 좋을 것이다.

6.3. 각 자료에 대한 해제

각 자료에 대한 해제는 국문 해제와 영문 해제의 두 가지로 한다. 영문

해제는 외국인을 위한 것이다. 그리고 옛 문헌에 대한 해제는 앞에서 기술한 바와 같은 '서지 정보'의 기술을 포함하는 것이 좋을 것이다.

6.4. 각 자료의 소장처 및 도서 번호

문헌을 소장하고 있는 곳과 그 특징을 다음과 같이 기술한다. 즉 이본까지도 정밀하게 기술한다.

(예)

소장처	도서 번호	비고(판본 등)
규장각	생략	활자본
서울대 고도서	생략	목판본

6.5. 각 자료의 영인본 목록

각 문헌 특히 옛 문헌의 영인본 목록은 다음과 같은 구조로 한다.

영인본 제목(판본), 간행 연도(해제자명), 출판사명.

(예)

訓民正音(解例本), 1974, 訓民正音(姜信沆 譯註), 신구문고 1, 신구문화사.
訓民正音(解例本), 1976, 譯解 訓民正音(박병채), 박영문고 150, 박영사.
訓民正音(解例本), 1988, 훈민정음 해례본, 용비어천가 훈몽자회와 합본, 대제각, 국어국문학총서 6.
訓民正音(解例本), 1995, 訓民正音新硏究(李覲洙), 보고사.

6.6. 각 자료에 대한 연구 논저 목록

각 문헌에 대한 연구 논저 목록은 다음과 같이 제시한다.

필자(간행 연도), 논저명, 출판사(잡지명).

(예)

□ 家禮諺解

金根洙(1962), 家禮諺解 解題, 國語國文學古書雜錄.

李德興(1985), 家禮諺解에 나타난 語彙形成考 -特히 漢字語를 中心으로-, 語文研究 48.

추교신(1982), 가례언해의 국어학적 연구, 인하대 교육대학원 석사학위논문.

洪允杓(1986), 家禮諺解 解題, 影印本 家禮諺解, 弘文閣.

6.7. 각 자료의 용례사전

원시 자료를 이용하여 용례사전을 만들어 둔다. 이것은 프로그램을 이용하여 자동으로 만들도록 노력해야 할 것이다. 다음에 용례사전의 한 예를 들어 보도록 한다.

<춘향上, 1a>

열여춘향슈졀가라. 슉종ᄃᆡ왕 직위 초의 셩덕이 너부시사. 셩자셩손은 계계승승ᄒᆞ사 금고옥족은 요슌시졀이요. 으관문물은 우탕의 버금이라. 좌우보필은 쥬셕지신이요 용양호위난 간셩지장이라. 조졍의 흐르난 덕화 힝곡의 폐엿시니. 사히 구든 기운이 원근의 어려 잇다. 츙신은 만조ᄒᆞ고 회자 열여 가가지라. 미지미지라. 우슌풍조ᄒᆞ니 함포고복 빅셩덜은 쳐쳐의 격랑가라. 잇ᄯᅥ 졀나도 남원부의 월ᄆᆡ라 하난 기싱이 잇스되. 삼남의 명기로셔 일직 퇴기ᄒᆞ야 셩가라 ᄒᆞ는 양반을 다리고 셰월을 보ᄂᆡ되. 연장사순의 당하야 일졈 혀륙이 업셔 일노 한이 되야 장탄슈심의 병이 되것구나. 일일은 크계 ᄭᅴ쳐 예 사람을 싱각ᄒᆞ고. 가군을 쳥입ᄒᆞ야

<춘향上, 1b>

엿자오ᄃᆡ. 공슌이 ᄒᆞ난 마리.

이 자료는 프로그램을 이용하여 다음과 같은 용례사전을 만들게 된다.

열여춘향슈졀가라.	열여춘향슈졀가라. <춘향上,1a>
슉종ᄃᆡ왕	슉종ᄃᆡ왕 직위 초의 셩덕이 너부시사. <춘향上,1a>
직위	슉종ᄃᆡ왕 직위 초의 셩덕이 너부시사. <춘향上,1a>
초의	슉종ᄃᆡ왕 직위 초의 셩덕이 너부시사. <춘향上,1a>
셩덕이	슉종ᄃᆡ왕 직위 초의 셩덕이 너부시사. <춘향上,1a>
너부시사.	슉종ᄃᆡ왕 직위 초의 셩덕이 너부시사. <춘향上,1a>
셩자셩손은	셩자셩손은 계계승승ᄒᆞ사 금고옥족은 요슌시졀이요. <춘향上,1a>
계계승승ᄒᆞ사	셩자셩손은 계계승승ᄒᆞ사 금고옥족은 요슌시졀이요. <춘향上,1a>
금고옥족은	셩자셩손은 계계승승ᄒᆞ사 금고옥족은 요슌시졀이요. <춘향上,1a>
요슌시졀이요.	셩자셩손은 계계승승ᄒᆞ사 금고옥족은 요슌시졀이요. <춘향上,1a>
으관문물은	으관문물은 우탕의 버금이라. <춘향上,1a>
우탕의	으관문물은 우탕의 버금이라. <춘향上,1a>
버금이라.	으관문물은 우탕의 버금이라. <춘향上,1a>

6.8. 각 자료에 대한 연구 논저 중 중요한 논문의 원문

각 문헌의 연구 논저 중에서 매우 중요한 업적으로 평가를 받는 논문을 이미지 파일로 만들어 CD 속에 포함시킨다. 그 선정은 전문 연구자에게 일임한다.

6.9. 검색 프로그램

원문 텍스트를 검색하면 거기에 해당하는 이미지 파일의 페이지를 찾아 보이게 하도록 한다.

7. 전자 자료의 활용 방안

자료를 많이 보유하고 있는 사람은 마치 자기가 가장 많은 사실을 알고

있는 양 생각하는 일이 있다. 그러나 자료를 많이 보유하고 있는 학자가 가장 뛰어난 학자가 아니다. 가장 뛰어난 학자는 그 자료들을 활용하는 사람이라고 할 수 있다.

자료를 수집·정리하여 제공만 해 주는 것으로서 그 연구기관이 책임을 다했다고 할 수는 없다. 그것을 활용할 수 있는 여건과 환경을 제공하여 주고 또한 활용의 결과를 다시 응용할 수 있도록 최선을 다해야 한다.

많은 전자 자료가 수집·정리되어 있어도 그것을 활용하지 않으면 아무런 가치도 없는 것이 될 것이다. 실제로 말뭉치를 비롯한 많은 전자 자료들이 공개되어 있어도 이것을 활용할 줄 아는 사람이 많지 않아서, 말뭉치의 위력이 아직은 크게 나타나지 않는 실정에 있다.

한국어 전자 자료의 이용자는 전문가인 국어학자들과 비전문가인 일반 국민들이라고 할 수 있다. 따라서 모든 사람들의 접근이 용이하도록 하여야 할 것이다. 전자 자료를 활용시킬 수 있는 가장 빠른 길은 다음의 몇 가지 방안일 것이다.

① 전자 자료를 연구자들이 쉽게 접근하여 이용할 수 있도록 이들을 CD로 만들어 배포하는 일

② 인터넷 상에서 공개하여 모든 사람들이 쉽게 검색할 수 있고, 또 쉽게 내려받아 이용할 수 있도록 하는 일.

③ 활용할 수 있는 각종의 프로그램을 만들어 배포하는 일

이러한 환경을 만들어 주면 많은 전문가들이 이 전자 자료들을 이용하여 다음과 같은 업적들을 쌓게 될 것이다.

① 각종의 국어사전 편찬이 이루어질 것이다. 특히 전문 분야별 한국어 사전이 편찬될 것이다.

② 국어 사전 편찬에 어휘 및 그 용례와 의미를 제공하여 줄 것이다.

③ 국어의 개별 어휘의 역사와 방언형을 제공하여 주어 국어 어휘사를 알 수 있게 해 준다.

④ 국어의 어원을 알 수 있게 해 준다.

⑤ 국어의 기초 어휘를 선정하게 해 준다.

⑥ 국어의 각종 빈도를 알 수 있게 해 준다.

⑦ 적절한 통계를 이용하여 국어 교육에 크게 활용된다.

⑧ 음운 변화의 유형을 알 수 있게 해 준다.

⑨ 각종 논문에서 이용하는 자료를 제공하여 줄 것이다.

그러나 이러한 자료를 추출해 낼 수 있는 능력이 없으면 이러한 작업도 가능하지 않다. 따라서 프로그램을 만들어 이를 사용하는 방법을 자세히 붙여 공개하여 활용할 수 있도록 해 주어야 한다.

다음에 지금까지 국어 자료를 처리할 수 있는 프로그램들을 소개하도록 한다.

① 어절별 색인 만드는 프로그램(halign.exe): 간단한 색인 작업을 쉽게 해 주는 것이다. 색인과 출전을 동시에 표시해 준다.

② 빈도 조사 프로그램(bindo.exe): 빈도를 조사하여 그것을 가나다순으로 정렬하여 보여 준다. 어절 빈도, 음절 빈도, 음소 빈도, 음소 연접빈도, 음소 연접빈도 용례 결과를 볼 수 있다.

③ 음절 빈도수 백분율 조사 프로그램(nsyll.exe): 각 음절 빈도의 백분율을 나타내 준다.

④ 용례사전 만드는 프로그램(kwoc.exe): 용례사전을 만드는 프로그램이다.

⑤ 정렬 프로그램(hansort.exe): 한글, 옛한글, 한자로 된 자료를 통합하여 가나다순 또는 그 역순으로 정렬해 준다.

⑥ 역순사전 만드는 프로그램(inverse.exe): 역순사전을 만드는 프로그램도 두 가지가 있다. 하나는 한 어절만 어절별 역순으로 만드는 프로그램(inverse1.exe)이고 또 하나는 모든 어절을 역순으로 만드는 프로그램(inverse2.exe)이다.

⑦ 형태소 분석 프로그램: 중세국어의 형태소를 분석하는 프로그램이다. 현재 완벽하게는 분석되지는 않지만, 약 95% 이상의 정확성을 지니고 있다.

⑧ 검색 프로그램 (morph.exe): 다음과 같은 다양한 자료를 검색할 수 있다.

ⓐ 음절의 전체 ⓑ 음절의 초성 ⓒ 음절의 중성
ⓓ 음절의 종성 ⓔ 음절의 초성+중성 ⓕ 음절의 초성+종성
ⓖ 음절의 중성+종성

⑨ 한 줄의 음절수 조사 프로그램(getum.exe): 한 줄의 음절수를 조사하여 음절수대로 목록을 만들어 준다.

⑩ 방점 자동 생성기(tone.exe): 중세국어의 방점을 처리하는 프로그램이다.

⑪ 형태소 검색기: 21세기 세종계획에서 개발한 검색기이다.

⑫ 통합 프로그램(깜짝새): 이 프로그램은 아직 완성된 것은 아니지만, 최근에 전주대 소강춘 교수팀이 개발한 통합 프로그램이다. 지금까지 개발된 어떠한 프로그램보다도 가장 쉽게 활용할 수 있도록 짜여진 것이다. 유니코드를 이용하여야 하므로 윈도 2000 환경에서만 사용이 가능하다.

이 이외에도 프로그램은 매우 다양하여서, 일일이 다 설명을 하지 못한다.

8. 맺는말

지금까지 한국어 전자 자료를 수집·정리하여 이것을 어떻게 활용할 것인가에 대해 매우 구체적으로 기술하였다. 그러나 그러한 자료를 수집·정리하여 활용할 수 있는 환경과 여건이 마련되지 않고, 단지 방법만 제시한 것이라면 이 글은 아무런 의미도 없게 될 것이다.

전자 자료를 수집 정리하여 이를 활용할 수 있도록 계획하고 실행할 수 있는 곳은 국가 기관밖에 없다. 그 중에서 가장 적당한 기관은 국립국어연구

원이라고 생각한다. 이 계획과 실행에는 많은 예산과 인원이 소요될 것으로 생각한다. 따라서 이 일은 단계적 계획을 세우고 이를 실현시킬 수 있는 구체적인 방안을 마련해 두어야 한다. 그리고 실제로 그 필요성을 인식시켜 예산을 확보하고 이를 실행하여야만 빛이 나는 것이다.

자료의 정리와 수집과 활용은 단지 전자 자료만에 국한된 것은 아니다. 다른 모든 자료를 통틀어 수집·정리하고 활용할 수 있도록 하여야 한다.

참고 문헌

南廣祐(1997).『敎學 古語辭典』. 敎學社.

劉昌惇(1964).『李朝語辭典』. 延世大學校出版部.

김정수(1984).「옛글의 준이름을 통일하기 위한 시안」.『언어학 7』. 한국언어학회.

남광우(1960).『고어사전』. 동아출판사.

홍윤표(1997).「한글 자료의 성격과 해제」.『국어사연구』. 국어사연구회 편. 태학사.