개화기 한글자료 말뭉치의 구축 방안

신 중 진*

Ⅰ. 서 론

현대는 정보화의 시대이다. 수없이 많은 정보들이 컴퓨터와 인터넷을 통해 유통된다. 더 많은, 더 빠른, 더 정확한 정보를 주고받기 위한 인간의 노력으로 컴퓨터공학은 급속히 발전하고 있다. CPU가 점점 증가하고 있고, 메인보드 용량이 기가바이트(Giga-byte) 단위를 훨씬 넘어 테라바이트(Tera-byte)까지 이르렀으며, 메모리 용량도 기가바이트 단위가 넘는 램(Ram)이 장착된 컴퓨터가 보급되고 있다. 심지어 슈퍼컴퓨터는 아주 복잡한 연산을 눈 깜짝할 사이에 해결하기도 한다. 이런 컴퓨터의 현란한 발전은 순수학문의 영역에도 영향을 주고 있다. 이제는 국어학도 컴퓨터의 무시못할 영향을 받고 있음이 주지의 사실이다.

컴퓨터가 국어학에 끼친 가장 큰 영향은 수천, 수억 어절의 데이터베이스로 구축된 국어자료의 제공에 있다. 다시 말하면 이제 우리는 컴퓨터를 통해서 원자료를 말뭉치 자료로 저장할 수 있게 되었고, 다양한 검색, 계산, 모의실험을 할 수 있게 되었다.[1] 국어학적으로 말뭉치가 갖는 가장 큰 효용은 그 동안 인상적으로 혹은 장난감 체계(Toy System)적인 예로써 국어학적

* 서울대학교 기초교육원 대우전임강사

1) 색인, 문맥색인, 핵심어 문맥색인뿐만 아니라 단어나 형태소 사이의 통합관계 및 문법적 단위의 용법을 좀더 다양하고 구체적으로 검토할 수 있고, 나아가 자소(grapheme), 음절, 어절 빈도를 조사할 수 있으며, 더나아가 인공지능 컴퓨터를 구현할 수도 있다.

진리에 도달했다면,[2] 이제는 대용량의 데이터베이스로 구축된 국어자료를 통해서 국어학적 진리에 도달할 수 있게 되었다는 것이다. 이것은 그간의 국어학적 성과들을 실제 언어자료로서 검증할 수 있고, 또 미처 인식하지 못했던 국어학적 진리를 찾을 수 있게 되었다는 말과 같다. 다시 말하면, "과연 자연 언어자료는 그 자체로서 연구대상이 될 수 있으며, 자연 언어자료 자체에 대한 연구는 언어연구의 영역에 들 수 있는가?" 라는 질문에 대하여, 연구자의 직관을 통해서 만들어지거나 다듬어진 것이 아니라, 자연 언어자료의 분석을 통해서 언어의 내적 구조를 밝히는 작업이 그 나름대로 중요성을 가지는 것이며, 경우에 따라서는 직관에 의해 창출된 자료를 바탕으로 한 언어연구보다 더 실질적이고 구체적인 결과를 얻을 수 있고, 더 나아가 기존 연구를 수정보완할 수 있으며, 미처 발견하지 못했던 새로운 진리에도 도달할 수 있다는 답변을 줄 수 있다.

이 글은 최근에 급속도로 성장한 국어 말뭉치와 그 연구 성과가 사실은 현대국어에 편중되어 있었다는 반성에서 시작되었다. 한글자료로만 보면, 현대 이전 즉 중세, 근대, 개화기의 자료도 문어텍스트로 남아 있다. 이들 시기의 자료들은 현대국어 자료와 구별되어 '역사자료'로 인식할 수 있다. 역사자료도 데이터베이스화하는 것은 국어학적 논의의 활성화를 위해 시급한 과제이다.[3]

따라서 이 글의 목적은 개화기의 주요 한글자료들의 활용 가치를 높이기 위한 방안으로 이들을 말뭉치로 구축하는 절차와 문제점 그리고 그 해결책을 제시하는 데에 있다. 즉 시기적으로 1876년부터 1910년까지 간행된 7편의 신문/잡지류, 13편의 성경류, 10편의 소설류, 3편의 사전류, 22편의 기타 교과서/기술서/윤음류 등을 합하여 모두 55편의 개화기 한글자료를 기계가독형 말뭉치로 구축하는 절차, 신뢰할 수 있는 말뭉치로 데이타베이스화하려고 할 때 생기는 문제점, 특히 역사자료를 대상으로 할 때 생기는 문제점, 그리고 그 해결방안을 제시하려 한다.[4]

---

2) 장난감 체계(Toy System)란 실험실 수준의 소규모 시스템을 말한다.

3) 그래서인지 지금은 대다수의 한글 문헌자료들이 정교함에는 문제점이 없지 않으나 어느 정도 전자자료로서 구축되어 있는 실정이다.

II. 말뭉치의 현황 및 요건

말뭉치 언어학(Corpus Linguistics)이라는 용어가 언제부터 사용되었는지는 확실하지 않다. 다만, 1950년대 미국의 실증적 구조주의 언어학자들이 실제 언어자료를 일차 연구대상으로 삼은 것을 효시로 볼 수 있다. 그러나 실질적으로 말뭉치를 통한 언어연구가 시작된 것은 영국학자들이 1959~1960년 이후 '영어 용법 조사 말뭉치(Survey of English Usage Corpus)'라는 프로젝트를 시작하면서부터이다. 이후 1963~1964년 미국 브라운(Brown) 대학의 프랜시스/쿠체라가 15개 장르에 걸친 500편의 글에서 2000어절씩 추출하여 100만 어절의 말뭉치를 구축한 것이 현대 말뭉치 시초이다(서상규 · 한영균 1999:18). 이후 대표적으로 Collins-COBUILD는 2천만 어절 말뭉치에 기반하여 사전을 내놓기 시작했고, 1990년대 초 영국의 BNC(British National Corpus)는 1억 어절 규모의 말뭉치를 완성하였으며, 현재 Bank of English는 Collins-COBUILD에서 관리하는 4억5천만 어절 규모의 초대형 영어 말뭉치를 구축하고 있다.

---

4) 개화기 한글자료의 목록을 보이면 다음과 같다(졸고 2003). **신문/잡지류** : 『경향신문』, 『(경향)보감』, 『대죠션독립협회회보』, 『대한미일신보』, 『독립신문(獨立新聞)』, 『협셩회회보/미일신문』, 『신학월보(神學月報)』 **성경류** : 『교회일과』, 『셩경젼셔』, 『셩경직히』, 『셩교빅문답』, 『셩교졀요』, 『시편촬요』, 『예수셩교젼셔』, 『쥬교요지』, 『쥬년첨례광익』, 『진교졀요』, 『치명일과』, 『텬로력뎡』, 『훈ᄋᆞ진언』 **소설류** : 『구마검(驅魔劍)』, 『귀의셩(鬼의聲)』, 『금슈회의록(禽獸會議錄)』, 『빈샹설(鬢上雪)』, 『설중미(雪中梅)』, 『송뢰금(松籟琴)』, 『은세계(銀世界)』, 『ᄌᆞ유종(自由鐘)』, 『치악산(雉岳山)』, 『혈의누(血淚)』 **사전류** : 『한불ᄌᆞ뎐』, 『國漢會語』, 『廣才物譜』 **기타(교과서/기술서/윤음류 등)** : 『경셕ᄌᆞ지문(敬惜字紙文)』, 『과화존신언해(過化存神諺解)』, 『관셩제군오륜경(關聖帝君五倫經)』, 『국문정리(國文正理)』, 『국민소학독본(國民小學讀本)』, 『남궁계젹셔』, 『명성경언해(明聖經諺解)』, 『삼성훈경(三聖訓經)』, 『서유견문(西遊見聞)』, 『소학독본(小學讀本)』, 『신정심상소학(新訂尋常小學)』, 『ᄉᆞ민필지(士民必知)』, 『어제유팔도사도기로인민등륜음(御製諭八道四都耆老人民等綸音)』, 『여사ᄉᆞ지(女士須知)』, 『역대천자문』, 『유중외대쇼민인등쳑샤(諭中外大小民人等斥邪)』, 『이무실천자문(李茂實千字文)』, 『이언언해』, 『잠샹즙요(蠶桑輯要)』, 『죠군령젹지』, 『중남포목포각국죠계쟝정(甑南浦木浦各國租界章程)』, 『진리편독삼자경(眞理便讀三字經)』.

국내에는 연세대학교의 연세 한국어 말뭉치 4,300만 어절, 과학기술원의 KAIST 말뭉치 7,158만 어절, 고려대학교 민족문화연구원의 말뭉치 1,000만 어절, 국립국어원의 말뭉치 6,765만 어절, 21세기 세종계획 말뭉치 165,492,052 어절이 대표적이고, 이들 말뭉치는 계속 증가하고 있다(홍윤표 2001a:62). 이 중 연세 한국어 말뭉치를 통해 5만 단어 규모의 중사전 『연세 한국어사전』을 출간했고, 국립국어원에서는 50만 단어의 『표준국어대사전』을 출간했다. 현재 문화관광부에서 주관하는 21세기 세종계획은 말뭉치를 통한 전자사전을 준비하고 있다.

한편 컴퓨터의 등장은 말뭉치 언어학의 성립을 가능하게 했다. 쉽게 말하면 언어학(국어학)은 컴퓨터 없이도 성립가능한 학문분야이지만, 말뭉치 언어학(국어학)은 컴퓨터가 없으면 성립불가능한 학문분야이다. 이처럼 말뭉치 언어학에서는 컴퓨터의 활용이 필수적이다.[5]

이런 말뭉치는 전산처리를 전제로 한 데이터로서 기계가독형 데이터(Machine Readable Dictionary)라고도 한다. 따라서 말뭉치는 손쉽게 수합, 분산, 저장, 복사, 전송이 가능한 것이 큰 특징이다. 이러한 특징을 갖는 말뭉치도 일정한 요건을 갖추어야 믿을 수 있는 자료가 된다. 첫째 우선 원자료의 내용과 형태가 고의로든 실수로든 달라지거나 첨삭됨이 없이 정확하여야 한다. 즉 원문 회복 가능성이 있어야 하므로 주석을 제외한 일체의 내용이 원자료와 동일해야 한다. 둘째 해당 언어의 변이가 최대한 반영되도록 말뭉치를 가공해야 한다. 즉 대표성과 균형성을 갖춘 말뭉치여야 한다.[6] 마

---

5) 컴퓨터는 말뭉치의 구축과 언어정보의 추출을 위한 말뭉치 가공과 가공된 말뭉치로부터 언어정보 추출과 추출된 언어정보의 통계적 분석과 통계 분석의 결과 및 개별 언어 단위들이 지닌 언어정보의 데이터베이스화와 데이터베이스의 운용, 유지, 확장 및 보수에 필수적인 활용가치를 가진다.

6) 말뭉치의 대표성을 획득하기 위해서는 표본이 된 말뭉치에 기반한 연구나 결과물이 해당 언어 전체에 대해서도 진실성을 가져야 한다. 대표성을 살리기 위해서는 양도 중요하지만, 표본의 모집단에 대한 정의와 표본화 방법을 고려해야 한다. 그리고 거듭되는 시행착오를 토대로 대표성을 획득할 수 있다. 그러나 실제로 대표성의 문제는 말뭉치 구축에서 해결해야 할 본질적인 목표로 아직 남아 있다. 말뭉치의 균형성은 일반 말뭉치에서 각기 다른 텍스트 영역들 간의 균형과 가중치의 문제와 직결된다. 즉 문어자료와 구어자료를 같은 비율로 구성한다고 해서 균형성

지막으로 분석결과가 해당 언어에 대한 진실성을 가지기 위해서는 해당 언어의 다양한 특성이 총체적으로 포착될 수 있을 만큼의 충분한 양을 수집해야 한다.[7]

Ⅲ. 말뭉치의 유형과 개화기 말뭉치의 위치

말뭉치는 아무런 가공이 되지 않은 순수 텍스트인 원시 말뭉치(raw corpus)과 품사를 비롯한 각종 문법 정보를 붙인 주석 말뭉치(annotated corpus)로 나눌 수 있다. 주석 말뭉치는 품사정보(POS tagged)를 단 경우가 기본이며, 이를 토대로 파싱(parsing)이 이루어질 수 있고, 통사정보, 동음이의어와 다의어 구분을 포함한 의미정보를 비롯하여 화용정보까지 가능하게 된다. 이들 정보들이 어우러지면 이해를 위한 말뭉치 활용을 넘어 표현을 위한 말뭉치 활용이 가능해진다.

그리고 특별한 목적이나 용도를 정하지 않고, 어휘, 문법, 담화 구조 등의 일반적인 언어 조사를 위해 텍스트들을 데이터베이스화한 경우를 일반 말뭉치(general corpus)라고 하며, 특정한 조사나 연구만을 위해 디자인된 경우를 특수 말뭉치(specialized corpus)라고 하는데, 학습, 실험, 방언, 지역, 학습자 말뭉치 등으로 분류될 수 있다. 한편 규모가 고정되어 있으면 정적 말뭉치(static corpus)라고 하고, 규모가 확대, 수정, 보완되면 동적 말뭉치(dynamic corpus) 혹은 모니터 말뭉치(monitor corpus)라고 한다.

한편 특정 시기의 언어를 대표하는 경우는 공시 말뭉치(synchronic corpus)라고 하며, 각기 다른 시대를 통틀어 각각의 시대에 사용된 용례들

이 완성되는 것은 아니라는 점을 유의해야 한다.

7) 실제로 말뭉치의 규모는 얼마나 많은 어절로 말뭉치를 구성하느냐의 문제라기보다는 말뭉치가 얼마나 많은 범주를 포함하고 있으며, 각 범주 안에 얼마나 많은 표본들이 포함되어 있느냐의 문제이다. 그러나 대체로 규모가 큰 말뭉치일수록 어떤 언어 사실에 대해 신뢰할 수 있을 정도로 다양하고 많은 예를 찾을 수 있다. 규모가 제한된 말뭉치로는 폭넓고 변화하는 언어 사실을 충분히 보여 줄 수 없다는 문제를 해결하기 위해 모니터 말뭉치(monitor corpus)를 구축하여 규모와 영역을 확대해 나가기도 한다.

을 모은 경우는 통시 말뭉치(diachronic corpus)라고 한다. 끝으로 현대 화자들의 언어직관과 세계에 대한 지식이 직접 닿는 언어자료를 수집한 경우는 현대 말뭉치(contemporary corpus)라고 하며, 화자의 언어직관과 세계에 대한 지식이 직접 닿지 않는 시기의 언어자료를 수집한 경우는 역사 말뭉치(historical corpus)라고 한다.

이 밖에 말뭉치의 디자인 방법, 사용 언어의 종류, 그 번역 여부, 사용 목적에 따라 유형 구분된다. 디자인의 방법에 따라 '균형 말뭉치(balanced corpus), 피라미드형 말뭉치, 기회적 말뭉치'가 있다. 모든 장르의 문서가 균등한 비율로 포함된 말뭉치를 '균형 말뭉치'라고 하고, 균형 말뭉치를 피라미드형으로 만든 말뭉치를 '피라미드형 말뭉치'라고 하며, 용례의 균형적 분포를 고려하지 않은 말뭉치를 '기회적 말뭉치'라고 한다. 사용 언어의 종류에 따라 '단일어 말뭉치, 이중어 말뭉치, 다중어 말뭉치'가 있다. 번역 여부에 따라서는 '원문 말뭉치, 번역 말뭉치'로 구분된다(홍윤표 2001a:41).

말뭉치 유형 구분을 통해 보면, 필자가 수집 데이터베이스화한 개화기 주요 한글자료 말뭉치는 원자료를 그대로 입력한 수준에서 더 이상의 주석을 달지 않았으므로 원시 말뭉치이고 특수한 조사를 위해서 구축된 것이 아니라 개화기의 일반적인 언어 사실을 포착하기 위해 구축되었으므로 일반 말뭉치이기도 하다. 그리고 현재로서는 규모가 고정되어 있으므로 정적 말뭉치이며 1876년부터 1910년까지 거의 1세대를 넘지 않는 시기의 언어사실을 구축했으므로 공시 말뭉치이다. 또 언어직관을 맹신할 수 없는 100여년 전의 문어텍스트를 대상으로 했으므로 역사 말뭉치이다. 그리고 '기회적 말뭉치, 단일어 말뭉치, 원문 말뭉치'이다.

IV. 역사자료로서 개화기 한글자료 말뭉치의 필요성

최근에 '역사 말뭉치'라는 개념이 널리 인식됨으로써 중세 한글자료부터 근대 한글자료 및 개화기 한글자료도 많이 전산화되었다. 이 글도 역사자료의 말뭉치를 염두에 두고 이루진 것이다. 어떻게 보면, 이 글은 대략 400만 어절에 달하는 방대한 개화기 한글자료를 토대로 해서 이루어지는 셈이다.

전산화된 400만 어절은 그 자체로도 활용가치가 높다. 개화기 국어 어휘에 대한 빠른 검색은 물론이고, 어휘들의 빈도, 누계 및 범주별 어휘 사용량, 그 편차, 어휘들 간의 연어관계 등을 손쉽게 확인할 수 있다. 이를 통해 언어사실의 변이도 쉽게 찾아볼 수 있음은 물론이다. 이런 점에서 역사자료의 말뭉치도 꼭 필요한 것이다. 따라서 이 글의 후반부에서는 실제 역사자료 특히 개화기 한글자료의 데이터베이스화를 시도할 때, 그 절차, 문제점 그리고 그 해결책을 제시할 것이다.

본격적인 논의에 앞서 개화기 한글자료를 역사자료로 보아야 할 몇 가지 이유를 살펴보겠다. 최근 세종계획의 일환으로 역사자료 구축이 진행되고 있다. 그런데 논의 과정에서 개화기 한글자료가 역사자료인가 현대국어자료인가 하는 다소 소모적인 논쟁이 계속되었다.

물론 역사자료인가 아닌가의 구분 더 나아가서 국어사적 시대구분은 동일한 언어학적 조건들에서 차이가 분명할 때 가능하다. 이 글은 현대국어와 개화기국어는 여러 근거와 더불어 그 차이가 분명하다는 견해를 좇는다.

구체적인 이유로 먼저 직접적인 면을 보면, 한글자료의 데이터베이스화는 국어의 정보화를 일차적인 목표로 삼는다. 한글자료는 한글로 기록된 모든 텍스트를 대상으로 하며, 한글로 기록된 텍스트의 시작은 15세기 '훈민정음'의 창제 이래로 소급된다. 세기별로 따지면 15세기 자료 및 16, 17, 18, 19, 20세기 자료를 포함하며, 국어사적 시대구분으로 보면, 후기중세국어, 근대국어, 개화기국어, 현대국어를 포함한다. 이런 거시적인 관점에서 본다면, 현대국어자료마저도 역사자료에 속하는 셈이다. 어떤 자료가 역사자료이든 아니든 한글자료라면, 국어의 정보화를 위해서 가급적 빠른 시일내에 데이터베이스화되어야 마땅하다.[8] 이쯤만으로도 개화기자료에 대해 역사자료인가, 아닌가를 논의하는 것은 적어도 국어의 정보화를 위해서는 무의미하며, 차라리 전체 한글자료를 데이터베이스화하는 데 있어서 개화기자료의 구축

8) 즉 시간의 축을 더 늘인다면, 현대국어자료는 잠재적 역사자료인 셈이다. 단지 현대국어자료는 잠재적 역사자료라는 사실과 현대국어화자들을 직접 둘러싸고 있는 자료라는 점에서, 좀 극단적인 감은 있지만, 마치 칼로 두부 자르듯이 다른 시기자료들과 뚝 떨어뜨려 놓고 우선시해 온 것이다.

이 다른 시기의 자료구축보다 우리에게 얼마만큼의 효용이 있는가를 평가해야 마땅하다. 이미 다른 시기의 자료들은 기대 이상으로 데이터베이스화되어 있다는 점은 우리들에게 개화기자료 구축의 필요성을 더 느끼게 한다.

한편 바람직한 국어의 정보화를 위해서는 어느 시기도 소홀히 다루어서는 안 되며, 혹 홀대된 시기가 있다면, 올바른 자료구축 지형을 위해서도 개화기의 자료구축에 시간과 자본이 투자되어야 할 것이다. 이런 의미에서 개화기 한글자료를 역사자료가 아닌 현대국어자료로 본다손치더라도 한글자료임이 분명하고, 다른 시기보다 관심이 소원했었다는 이유로 데이터베이스화하는 데에 주저해서는 안 된다.

변천하는 시대별 자료의 신속하고 효율적인 검색은 국어에 대한 고급 정보를 제공할 수 있다는 면에서도 개화기 자료 구축은 시급한 요구사항이다. 언어는 유기체적인 속성을 가지고 있다는 점에서 현대국어 시기마저도 엄밀한 의미에서 언어변화의 격동기, 과도기임은 앞에서도 지적하였다. 현시점에서 정확히 100 여년전 시기는 한국 역사의 특수성을 감안하더라도 언어사실의 격동기임이 틀림없다. 따라서 현대국어의 모태가 되는 개화기자료는 역사자료로서 데이터베이스화의 시급한 소재이다.

다음으로 간접적인 면을 보면, 우선 개화기자료는 현대국어 화자의 언어직관과 거리가 있다. 가령 1910년을 전후해서 양산된 신소설을 해석하는 데에 있어서 현대문학 전공자 중에 그 시대에 관심이 많은 연구자도 어려움을 느끼는 실정이다. 특히 표기법에서 뿐만 아니라 신조어, 외래어, 전문어 등은 주석의 대상이 되는바, 현대국어 어휘체계로는 개화기 텍스트 자체에 대한 이해를 제대로 해낼 수 없는 경우가 많다. 이런 의미에서도 현대국어자료와는 구별된다.

근대성, 보편성, 필연성을 추구하는 세계관이 반영되어 있는 개화기자료는 현대국어자료와 궤를 같이 한다. 이런 면에서 현대국어 전공자들이 개화기 자료를 자신들의 연구영역에 포함시킨다. 개화기자료의 내용이 갖는 근대성, 보편성, 필연성은 이전의 전근대성, 봉건성, 우연성, 민족성을 극복한 현대문학의 속성이라는 점 때문이다. 그러나 언어사실만을 두고 보면 현대국어와 음운, 형태, 통사, 의미, 특히 어휘에서 차이가 있음을 이들도 함께

공감한다.

한편 국립국어원에서 이미 국어의 시대별 변천 연구 시리즈로 개화기를 하나의 큰 역사자료 테마로 삼고 있다. 즉 1999년 발행된 『국어의 시대별 변천연구』 4에서 개화기 국어의 음운, 형태, 어휘, 문법 및 자료를 개략적으로 소개하고 있는데, 모든 필자들이 국어사자료의 관점에서 개화기 자료에 접근하였다.

끝으로 최근 홍윤표(2001a, 2001b)에서 '현대국어 자료의 기점' 문제를 언급하면서 21세기에 접어든 오늘날에는 20세기초의 국어 자료를 현대국어 자료로 인식하는 사람이 적다는 점을 지적하고, 대체로 자료의 양이 극도로 제한되어 분포하는 시기인 20세기 중반(8.15 광복) 이전의 자료를 국어사 사료로 보고, 방대한 자료를 확보할 수 있는 20세기 중반 이후 시기의 자료를 현대 자료로 보는 것이 합리적이라고 주장한다. 이것은 1945년까지의 자료를 역사자료로 삼는 견해이고 보면, 적어도 20세기 초반기까지의 자료를 역사자료로 삼는 것은 당연한 것이다.[9]

V. 개화기 말뭉치 구축 절차

말뭉치 구축의 일반론을 참조하고, 역사자료의 특수성을 감안할 때, 개화

9) 1998년부터 2000년초까지 규장각에 있는 한글 문헌이 실사되고 해제된 일이 있다. 규장각에는 『龍飛御天歌』를 비롯하여 광복이전까지의 한글문헌이 방대하게 소장되어 있다. 작업은 세기에 따라 경사자집(經史子集)별로 진행되었는데, 마무리 작업이 한창인 1999년말과 2000년초에 한 가지 큰 고민을 안게 되었다. 1900년을 전후한 소위 개화기자료를 역사자료로서 해제해야 하는가가 그것이었다. 그러나 1999년말에 개화기자료가 현대와 동세기이므로 현대국어와 언어사실에서 다소의 차이는 있더라도 제외되었지만, 2000년 원단을 기해서 개화기자료도 이제는 동세기를 넘긴 역사자료임이 분명해진 이상, 해제를 해야 마땅하다는 견해가 팽배하게 되었다. 비유컨대 방언조사에서 제보자의 선정기준 중에 제보자는 되도록이면 조사 지역에서 3대 이상을 거주해야 하는데, 이는 3대까지는 서로 방언의 간섭이 심하기 때문이다. 따라서 산술적으로는 3대 즉 90년에서 100년 이상의 기간은 언어변이에서 의미있는 구분기점이 될 수 있겠다.

기 역사 말뭉치의 구축 절차는 다음과 같다. 먼저 개화기 한글자료를 가지고 원시 말뭉치(raw corpus)를 구축하기 위해서는 첫째, 원자료를 수집해야 한다.[10] 이때 역사자료의 수집인 만큼 가급적 선본(善本)을 선택해야 한다. 서지학적으로도 선본이 가치가 있음은 물론이거니와 기계가독형 데이터베이스로 만들 때도 선본을 원자료로 삼아야 우선 입력의 편의를 도모할 수도 있고, 활용에서도 말뭉치의 신뢰를 높일 수 있다.

둘째, 수집된 자료는 텍스트의 성격에 따라 범주별로 구별해야 한다. 이것은 입력, 가공의 편의를 위해서라기보다는 구축된 자료의 가치있는 활용과 믿을 수 있는 결과를 추출하기 위해서 필요한 말뭉치 구축 절차의 하나이다.

셋째, 범주별로 원자료를 기계가독형으로 데이터베이스화한다. 이 때 가장 유의해야 할 점은 오자와 탈자 없이 정확히 입력하는 것과 띄어쓰기를 정확히 지키는 것이다. 처리된 말뭉치에 대해 신뢰할 수 있는 계량과 분석을 위해서는 오자나 탈자를 줄이는 것과 띄어쓰기를 정확히 지키는 것이 말뭉치 구축의 실제에서 대단히 중요한 작업 과정이다.

넷째, 원자료와 말뭉치자료를 일일이 비교하면서 교정을 보아야 한다. 이 때는 입력할 때에 못지 않은 시간과 인력이 요구된다. 자동분석 및 의미있는 결과를 추출하기 위해서는 이처럼 꼼꼼한 후처리 수작업이 반드시 뒤따라야 한다.

마지막으로 교정이 끝나면, 검색과 분석의 효용을 돕기 위해 헤더 마크업을 비롯하여 본문 마크업 태깅을 달아야 한다. 국내의 관행에 따라 필자는 TEI 텍스트 인코딩 체계를 따랐다. 이것은 동일 범주에 속하는 데이터베이스들에 대해 통합검색, 통합분석을 할 때도 개별 데이터들의 기본 정보를 보존할 수 있어서 반드시 필요한 절차이다.

---

10) 본고의 개화기 한글자료 말뭉치는 원시 말뭉치를 대상으로 한다. 현재 필자가 원시 말뭉치 이상으로 원자료를 가공하지 못한 이유도 있지만, 무엇보다도 원시 말뭉치가 모든 말뭉치의 기본이기 때문이다. 믿을 수 있는 원시 말뭉치의 구축은 자료 자체뿐만 아니라 연구 결과물의 신뢰를 위해서도 반드시 선결되어야 한다.

VI. 개화기 말뭉치 구축시 문제점

개화기 한글자료의 말뭉치 구축 절차를 염두에 둘 때, 생기는 문제점으로 우선 어절별 띄어쓰기의 통일성 여부를 들 수 있다.[11] 개화기국어 자료는 현대국어 자료와 달리, 띄어쓰기가 이루어져 있지 않거나 띄어쓰기를 하더라도 현대국어 어법에 따른 띄어쓰기가 적용되지 않고 편의에 따라 띄어 쓴 경우가 대부분이다. 신뢰할 수 있는 말뭉치를 구축하기 위해서는 어절별 띄어쓰기의 통일성을 기해야 한다. 다시 말하면 띄어쓰기는 말뭉치 구축의 실제에서 중요한 절차이다. 왜냐하면 띄어쓰기가 전혀 되어 있지 않으면 컴퓨터를 통한 형태분석이 불가능하게 되고, 더군다나 띄어쓰기의 통일성이 없으면 신뢰할 수 있는 검색결과나 형태분석 결과를 내놓을 수 없기 때문이다.[12]

현대국어 자료의 입력 처리에서는 발생하지 않지만, 개화기국어 자료의 데이터베이스화에서 생기는 또다른 문제점으로 옛글 처리 즉 표기법 문제를 들 수 있다. 개화기 한글자료에는 맞춤법이 반영되어 있지 않기 때문에 어두의 합용병서나 이미 음가가 사라졌지만 표기의 보수성 때문에 관례로 쓰인 'ㆍ'가 혼하게 나타난다. 그리고 동일 어휘의 표기에도 많은 혼란을 보인다. 특히 15개의 다양한 표기를 보이는 개화기 어휘도 있다.[13] 이들은 입력에도 어려움을 주지만, 활용 단계에서도 아직 자동분석 시스템이 개발되지 않았기 때문에 수작업을 통해 활용해야 한다(한영균 1994:8).

기타의 세부적인 문제점으로 원자료 자체의 오자나 탈자, 입력시에 발생

---

11) 이 점은 비단 개화기 한글자료만의 문제가 아니다. 현대국어 자료도 띄어쓰기는 크나큰 문제점이다. 현대국어 말뭉치에서의 띄어쓰기 문제에 대해 한영균(2003)에서 상세히 논하고 있다.

12) 현재 이 글의 자료가 된 개화기 한글자료들은 '혼글' 프로그램 툴(tool)을 통해서 입력되었으며, 결과물은 txt파일로 저장했다.

13) '가르치다'는 '가르치다, 가라치다, 가ᄅᆞ치다, 갈아치다, 갈오치다, 갈으치다, 갈ᄋᆞ치다, ᄀᆞ라치다, ᄀᆞ르치다, ᄀᆞᄅᆞ치다, 골으치다, 골ᄅᆞ치다, 골아치다, 골앗치다, 골ᄅᆞ치다'와 같이 무려 15개의 표기형이 나타난다.

한 오자나 탈자도 빼놓을 수 없는 것이고, 원자료 자체의 일부가 판독 불가능한 경우도 문제점으로 들 수 있다. 끝으로 자료의 저장방식, 저장 파일명의 표준화도 무시할 수 없는 중요한 문제점이다.

Ⅶ. 문제점 해결 방안

개화기국어 자료의 구축을 위해서는 무엇보다도 띄어쓰기 문제를 우선 해결해야 한다. 당연히 이것은 현대국어 자료의 구축보다 더 많은 수작업이 요구된다. 쉽게 말하면 현대국어 자료 처리에서는 필요하지 않는 가공 단계를 하나 더 거쳐야 한다는 것이다. 구체적인 개화기 한글자료의 띄어쓰기 처리절차를 소개하면, 일단 입력의 최우선 원칙은 원자료를 있는 그대로 최대한 반영하는 것이다. 따라서 원본이 띄어 써진 경우에는 띄어 써진 그대로 반영하는 것을 원칙으로 한다. 특히 우선은 현대국어 띄어쓰기에 따라 띄어 쓰는 일이 없도록 주의한다. 현대국어 어법에 맞는 띄어쓰기는 입력자료를 교정한 후, 다른 개화기 한글자료의 띄어쓰기 경향을 면밀히 파악한 후, 2차 가공시에 현대국어 어법의 띄어쓰기 지침을 바탕으로 이루어져야 한다. 데이터베이스화된 자료마다 띄어쓰기의 경향이 달라지면, 검색이나 분석의 결과를 전혀 신뢰할 수 없기 때문이다. 한편 원자료에 있는 특수 기호도 생략하지 않고 가능하면 그대로 반영한다.

개화기 한글자료는 띄어쓰기 경향상 크게 세 가지 유형으로 분류할 수가 있다.[14] 그리고 유형별로 입력 절차를 위한 방안은 다음과 같다.

(1) 이전 시기의 역사자료와 마찬가지로 띄어쓰기가 전혀 되어 있지 않은 자료
예 : 『셩경젼셔』, 『시편촬요』를 제외한 성경류, 소설 중에 『구마검』, 대부분의 언해자료, 윤음자료 등.

---

14) 신문류는 모두 띄어쓰기가 반영되어 있다. 단 『신학월보』는 구절단위로 띄어 쓰고 있다. 성경류는 띄어쓰기가 반영되어 있지 않다. 단 『셩경젼셔』, 『시편촬요』는 띄어쓰기가 반영되어 있다. 소설류는 띄어쓰기가 반영되어 있다. 단 『구마검』만 띄어쓰기가 되어 있지 않다. 한편 기타 자료에서 교과서류는 띄어쓰기를 하는 경향이 있고, 윤음이나 언해류는 띄어쓰기를 하지 않는 경향이 있다.

입력자는 원문의 의미를 최대한 살리는 선에서 현대국어 어법에 맞도록 띄어쓰기를 적용하여 입력한다. 교정자는 데이터베이스화된 자료와 원자료를 대조하면서, 원문이 최대한 반영되도록 오자와 잘못 띄어 쓴 어절을 바로 잡는다. 결과적으로 현대국어 어법에 맞는 띄어쓰기가 되도록 원시 말뭉치를 구축한다.[15]

(2) 절 단위로 띄어쓰기가 되어 있는 자료
예 : 『신학월보』

여기서 절이란 구보다는 크고 문장보다는 작은 단위로 하나 이상의 서술어를 포함하는 구 이상의 단위를 말한다.[16] 예컨대 『신학월보』에는 "쌀니흔드는대로 놉흔소리가나셔"같이 띄어쓰기되어 있다. 입력자는 원자료의 절 사이에 절 경계로 '#'를 삽입하고, 절 내부는 현대국어 어법에 맞도록 띄어쓰기를 적용했다. '#'을 삽입한 것은 결과적으로 띄어쓰기 단위와 원자료의 구절 단위를 구별하기 위해서이다. 교정자는 원문이 최대한 반영되도록 오자나 탈자와 잘못 띄어 쓴 어절을 바로 잡되, 절 경계를 반드시 확인한다. 따라서 위 예문은 "쌀니 흔드는 대로 # 놉흔 소리가 나셔"와 같은 데이터베이스화된다.

(3) 현대국어 띄어쓰기와 일치하지는 않지만 어절별로 띄어쓰기가 되어 있는 자료
예 : 『독립신문』, 『경향신문』를 비롯한 대부분의 신문류, 『혈의누』, 『치악산』을 비롯한 대부분의 소설류, 『스민필지』를 비롯한 학습서, 교과서, 기술서류

위의 (1)과 (2)의 경우는 띄어쓰기가 되어 있지 않다는 정보가 있으므로 데이터베이스 초기 입력시에 현대국어 어법에 맞는 띄어쓰기가 허락되지만, 원자료가 이미 띄어쓰기되어 있는 경우, 데이터베이스 구축을 위한 초기 입

15) 역사자료의 띄어쓰기는 홍윤표(2001a:52-3)을 참조할 수 있다.
16) 구 단위로 띄어 쓴 부분이 전혀 없는 것은 아니다.

력시에 현대국어 어법에 맞는 띄어쓰기를 임의대로 한다면, 결과적으로 원문으로 회복이 불가능하게 된다. 따라서 우선 입력자는 원문의 띄어쓰기를 그대로 반영하여 원자료를 입력한다. 교정자는 입력시에 현대국어 어법에 맞는 띄어쓰기가 적용되지 않았는지를 주의하면서, 오자와 탈자를 바로 잡는다.

이상의 세 유형의 텍스트들은 충분한 교정을 거친 이후, 믿을 수 있는 검색과 분석이 가능하도록 현대국어 어법에 맞는 띄어쓰기 지침을 마련하여 통일되게 띄어쓰기를 적용하도록 한다.

다음은 옛글 처리 즉 표기법 처리 방안이다. 앞에서 잠시 언급했다시피 개화기 한글자료에는 옛글의 흔적이 많이 남아 있다. 바로 이 점 때문에 말뭉치 구축에서 개화기 한글자료가 현대국어 자료와 달리 더 많은 어려움을 갖게 된다. 다음의 예는 그 중에 대표적인 경우를 보인 것이다.

(4)
ㄱ. 발, 쏠, 뼈, 짝총(-銃), 뼉다귀
ㄴ. 익국ᄉᆞ상(愛國思想), ᄋᆞ히, 너음ᄉᆡ
ㄷ. 쥥츄막, 쳘교(鐵橋), 인쟝(印章)
ㄹ. 념치(廉恥), 례배당(禮拜堂) 롱아원(聾啞院), 륙혈포(六穴砲)
ㅁ. 뎡거쟝(停車場), 텰도(鐵道), 텬연두(天然痘)

(4ㄱ)은 개화기 한글자료에서 보이는 합용병서의 예들이다. 이전 시기의 역사자료에서도 볼 수 있는 'ㅂ'계 합용병서와 'ㅅ'계 합용병서가 여전히 개화기 한글자료에도 나타나고 있음을 알 수 있다. (4ㄴ)은 표기에 'ㆍ'가 나타나는 예를 보인 것이다. 예 외에도 '-ᄒᆞ다' 파생어들은 거의 대부분이 'ㆍ'로 표기된다.[17] (4ㄷ)은 현대국어에서 'ㅈ, ㅊ' 다음에 오는 'ㅑ, ㅕ, ㅛ, ㅠ'는 'ㅏ, ㅓ, ㅗ, ㅜ'로 표기하기로 한 현행 한글맞춤법 때문에 차이나는 예이고, (4ㄹ) 또한 현행 한글맞춤법의 두음법칙에 반하는 예들이다. 마지막으로 (4

17) 『신학월보』에서는 '-하다'로 표기된 파생어들이 자주 보인다. 이것는 『신학월보』가 신학 잡지인바, 표기의 보수성보다는 현실발음을 더 따른 결과로 보인다. 이 경우에 결과적으로 현대국어에서의 표기와 차이가 없다.

ㅁ)은 현행 한글맞춤법의 구개음화가 반영되지 않은 표기가 그대로 드러난 예들이다. 현재로서는 이들 옛글들을 그대로 반영해서 입력하는 수밖에 없다.

한편 검색과 분석을 위한 데이터베이스화에서 이보다 더 큰 문제인 '동일 어휘에 대해 다양한 표기를 보이는 개화기의 표기 양상'에 대한 일관된 처리 방안도 필요하다. 한글맞춤법 통일안이 제안된 해는 1933년인바, 본고의 대상이 된 개화기 한글자료에는 표기에 통일성이 없다. 그래서 이전 시기의 한글자료와 마찬가지로 자연히 표기에도 혼란이 보인다. 가령 개화기에 '心'에 대해서 '마암, 마음, 마ᄆᆞᆷ, ᄆᆞᆷ'와 같이 네 가지 표기가 혼재한다. 심지어 '가르치다'는 '가르치다, 가라치다, 가ᄅᆞ치다, 갈아치다, 갈오치다, 갈으치다, 갈ᄋᆞ치다, ᄀᆞ라치다, ᄀᆞ르치다, ᄀᆞᄅᆞ치다, ᄀᆞᆯ으치다, ᄀᆞᆯᄅᆞ치다, ᄀᆞᆯ아치다, ᄀᆞᆯ앗치다, ᄀᆞᆯᄅᆞ치다'와 같이 무려 15개의 표기형이 나타남을 전술한 바 있다.[18]

말뭉치 활용을 위해서는 개화기 한글자료의 표기법에 대해 처리 방안 두 가지을 제시할 수 있다. 하나는 동일한 의미와 유사한 자소형(grapheme)을 가진 이표기들에서 대표 표기형을 설정하고 이표기들을 대표 표기형으로 바꾸어서 입력하는 것이다. 예컨대, '마암, 마음, 마ᄆᆞᆷ, ᄆᆞᆷ'의 경우 개화기 한글자료에서 가장 보편적인 표기형인 '마ᄆᆞᆷ'을 대표형으로 설정하고 입력하여 결과를 추출하는 것이다. 다른 하나는 동일한 의미와 유사한 자소형을 갖는 이표기들을 모두 드러내는 방법인데, 이표기들을 '마암/마음/마ᄆᆞᆷ/ᄆᆞᆷ'과 같이 한 어절 형식으로 입력하여 결과를 추출하는 것이다.

두 가지 이표기 처리 방법에는 일장일단이 있다. 그러나 컴퓨터를 통한 말뭉치언어학에서는 이표기형을 모두 드러내는 후자의 방법이 더 적합하다. 왜냐하면 대표형을 설정하든, 이표기형을 모두 드러내든, 거의 뼈를 깎는 절대시간 양의 수작업이 필요한데, 개화기의 이표기형을 모두 보여주는 것이 개화기의 표기 실상을 직접 반영하는 결과이기 때문에 더 친절한 처리 방법이 될 수 있다.

---

18) 이와 같은 현상은 현대국어 자료의 처리에서는 고민거리가 되지 않는 역사자료 처리만의 독특한 문제점이다.

한편 교육과 계몽 그리고 선교 등의 이유로 개화기에는 필연적으로 신문, 잡지, 소설, 성경, 교과서를 비롯한 각종 신서적들이 봇물 터지듯 발간되었다. 각종 신서적의 발간은 근대식 인쇄 시설인 활판 인쇄기의 도입을 촉진시켰다. 초기 근대적 인쇄술의 도입은 정부에 의해 추진되었는데, 1883년 일본에서 기계와 장비를 들여와 설립된 박문국(博文局)이 우리나라 최초의 근대식 인쇄소였다. 그 후 민간에서도 많은 근대식 인쇄소들이 설립되면서 인쇄업이 민간 경영 형태로 자리잡게 되었다.[19] 이러한 근대식 인쇄시설의 도입은 이제까지의 인쇄 방식에서 탈피하여 서적의 대량 생산과 더불어 대중화를 가능하게 했다. 그러나 인쇄시설의 근대화는 인쇄물의 대량화, 대중화를 낳은 이면에 인쇄물의 형식과 내용에서는 저급화를 낳은 것도 사실이다. 특히 전통적인 제책 과정을 고려할 때, 근대식 제책 과정은 상당히 엉성한 편이었다. 결과적으로 예전의 고문헌에서는 찾기 어려운 오자나 탈자가 개화기 한글자료에는 너무나 흔하게 산재한다.

이런 연유에서 생긴 오자와 탈자에 대한 데이터베이스화 방안을 알아보겠다. 우선 입력자의 실수로 생긴 오자나 탈자는 교정시에 원문대로 수정이 되어야 한다. 이보다 더 큰 문제는 위에서 언급한 이유에서 생긴 원문 자체의 오자나 탈자인데, 데이터베이스화에서는 이것에 대한 정보도 함께 담아야 한다. 예컨대 『독립신문』 1896년 5월 12일자 '논셜'에 "다ᄉᆞ리를기 ᄇᆞ라노라"에서 '다ᄉᆞ리를기'와 같은 어절이 나타난다. 여기서의 '다ᄉᆞ리를기'는 전후 문맥상 '다ᄉᆞ리기를'의 오표기이다. 말뭉치에서는 이 어절을 '다ᄉᆞ리를기[다ᄉᆞ리기를]'과 같이 처리했다. 이처럼 특수 기호와 함께 한 어절로 처리함으로써 원자료의 어절수와 동일함을 유지하고, 원의미도 담을 수 있게 된다. 한편 간혹 특수 기호 [ ] 속에 '?'를 단 경우가 있다. 이것은 오자임이 확실하지만 교정자도 확신이 서지 않아서 붙인 경우이다. 차후 검토를 위해서

19) 대표적인 민간인쇄소로는 서울의 開文堂, 廣文社, 國民敎育會, 校文舘, 大同廣智社, 大韓國文舘, 大韓每日申報社, 同文館, 萬歲報館, 文雅堂印刷所, 博文社, 博文書館, 博學書館, 普文舘, 普文社, 寶晉齋印刷所, 普印社, 誠文社, 新文舘, 右文舘, 以文社, 日新社, 昌文社, 昌新社, 塔印社, 華文舘 石版部, 皇城新聞社, 興大社, 지방의 慶南日報社印刷所(진주), 廣文社(대구), 仁川活版所 등이 우후죽순처럼 생겨났다(김봉희 1999:39~41).

[ ] 속의 '?'를 필히 입력해 둔다.

다음으로 원자료 자체의 일부가 판독 불가능한 경우의 처리 방안이다. 이를 보기 위해 아래의 예를 살펴보겠다.

(5) 쏘 경무청으로 말ᄒᆞ면 입에서 졋 가 나고 킈가 ᄒᆞᆫ자가 못 되는 이도 잇스며 턱밋히 슈염이 빅셜 ᄀᆞᆺ흔이가 간간이 셕겻스니 이사롬 ᄒᆞ나히 직무에 근간ᄒᆞ며 각 치 못ᄒᆞ면 빅셩의게 무슨 해가 밋칠지 모로거늘(광무이년 구월이십칠일 화요일 뎨일권 미일신문 뎨빅이십팔호, 밑줄은 필자의 것)

여기서 밑줄친 어절 '졋 가, 각 치'는 각각 한 음절이 빠져 있어서 판독이 불가능하다. 이런 경우에는 〈gap reason='1자 판독 불가'〉라는 태그(tag)를 달아서 입력한다. 대부분의 태거는 원자료의 데이터베이스화가 이루어진 다음에 달지만, 판독 불가 태거는 그때그때 달아서 후처리를 원활하게 한다. 위의 예를 처리한 결과를 보이면 다음과 같다.

(5') 쏘 경무청으로 말ᄒᆞ면 입에서 졋〈gap reason='1자 판독 불가'〉가 나고 킈가 ᄒᆞᆫ자가 못 되는 이도 잇스며 턱밋히 슈염이 빅셜 ᄀᆞᆺ흔이가 간간이 셕겻스니 이사롬 ᄒᆞ나히 직무에 근간ᄒᆞ며 각〈gap reason='1자 판독 불가'〉치 못ᄒᆞ면 빅셩의게 무슨 해가 밋칠지 모로거늘

이외에 자료의 저장방식도 빼놓을 수 없는 중요한 현안이다. 원칙은 통일된 저장방식을 유지해야 한다는 것이다. 많은 사람이 공유할 수 있는 말뭉치 구축을 위해서는 무시할 수 없는 문제점이다. 저장방식에는 우선 원자료로의 회복 가능성을 염두에 두고 결정되어야 한다. 이를 전제로 글자 크기부터 줄 간격, 특수기호, 글꼴 등뿐만 아니라 저장 파일명도 표준화가 필요하다. 이와 아울러 말뭉치의 공유를 위해서는 '헤더(Header)'를 다는 양식의 표준화도 반드시 필요하다. 예컨대 필자가 작업한 『협셩회회보/미일신문』의 저장방식 지침의 일부를 보이면 다음과 같다. 이것은 가급적 다른 모든 개화기 한글자료에도 동일하게 적용시켰다.

(6)

ㄱ. 글자 크기 : 10point

ㄴ. 줄 간격 : 180

ㄷ. 쪽번호매기기 : 본문에 태깅

ㄹ. 원자료의 특수기호 : 일단 그 모양대로 반드시 살릴 것.

ㅁ. 의도적으로 행이 조절된 경우 : "하ᄂᆞ님, 황뎨, 폐하" 등과 같은 어휘가 나오면, 존경의 뜻으로 줄을 바꾸는 이른바 '공격(空隔)'을 두게 되는데, 이 때는 원문에 따라 입력시에도 줄을 바꾼다.

ㅂ. 입력시 행을 띄우는 일이 없도록 한다. 단 날짜가 다른 신문과 신문 사이는 교정시 가시적 효과와 자동분석의 편의를 위해 한 행을 띄운다. 현재로서는 별도의 날짜 표시는 하지 않는다.

ㅅ. 신문 제목은 다음과 같이 통일하여 입력하기로 한다.
광무 이년 십월 륙일 목요일
뎨일권 미일신문 뎨빅삼십륙호

ㅇ. 글꼴은 명조로 하고 좌로 정렬시킨다. 교정시에 오른쪽 끝 어절의 띄어쓰기 여부를 분명히 확인하기 위해서이다.

ㅈ. 본문 태깅을 한다.

ㅊ. 헤더 마크업을 한다.

ㅋ. 저장은 txt파일, 검색의 편의를 위해 경우에 따라서는 2b파일로 한다.

끝으로 『협성회회보/미일신문』의 헤더 마크업을 보이면 다음과 같다. 헤더는 '21세기 세종계획'에서 마련한 헤더의 표준양식을 따랐다.

```
〈!DOCTYPE tei.2  SYSTEM "c:₩sgml₩dtd₩tei2.dtd" [
        〈!ENTITY % TEI.written "INCLUDE"〉
        〈!ENTITY % TEI.corpus "INCLUDE"〉
        〈!ENTITY % TEI.extensions.ent SYSTEM "K1.ent"〉
        〈!ENTITY % TEI.extensions.dtd SYSTEM "K1.dtd"〉
]〉

〈tei.2〉

  〈teiHeader〉
  〈fileDesc〉
    〈titleStmt〉
      〈title〉미일신문, 전자파일〈/title〉
      〈author〉마상-배재학당 협성회〈/author〉
```

```
        〈sponsor〉대한민국〈/sponsor〉
        〈respStmt〉〈resp〉문헌입력, 표준화, 헤더붙임, TEI markup〈/resp〉
            〈name〉한국대학교, 홍길동〈/name〉
        〈/respStmt〉
    〈/titleStmt〉
    〈extent〉20,903어절〈/extent〉
    〈publicationStmt〉
        〈distributor〉한국대학교〈/distributor〉
        〈idno〉p9bc0001.hwp〈/idno〉
        〈availability〉〈p〉배포 가능〈/p〉〈/availability〉
    〈/publicationStmt〉
    〈notesStmt〉
        〈note〉〈p〉이 말뭉치는 1898년 1월 1일부터 창간되어 1899년 4월 3
일까지 간행된 한글 신문 〈미일신문〉 권1의 전자파일이다. 이 파일에는 권1
128호부터 권1 147호까지가 입력되어 있다.
〈/p〉〈/note〉
    〈/notesStmt〉
    〈sourceDesc〉
        〈bibl〉〈author〉미상-배재학당 협성회〈/author〉
            〈title〉미일신문〈/title〉
            〈pubPlace〉배재학당〈/pubPlace〉
            〈publisher〉미상〈/publisher〉
            〈date〉1898-1899〈/date〉
        〈/bibl〉
    〈/sourceDesc〉
  〈/fileDesc〉
  〈encodingDesc〉
    〈projectDesc〉〈p〉역사자료 구축〈/p〉
    〈/projectDesc〉
    〈samplingDecl〉〈p〉직접 입력〈/p〉
    〈/samplingDecl〉
    〈editorialDecl〉〈p〉역사자료 말뭉치 문헌 입력 지침에 따름〈/p〉
    〈/editorialDecl〉
  〈/encodingDesc〉
  〈profileDesc〉
    〈creation〉〈date〉2000〈/date〉〈/creation〉
    〈langUsage〉
        〈language id=KO usage=99〉한국어, 고어〈/language〉
```

```
        〈/langUsage〉
        〈textClass〉
                〈catRef  scheme='K1'  target='P19CD'〉신문,  협성회회보  관계
〈/catRef〉
        〈/textClass〉
    〈/profileDesc〉
    〈revisionDesc〉
        〈change〉〈date〉2000/11 〈/date〉
                〈respStmt〉〈resp〉프로젝트                책임자〈/resp〉〈name〉홍길동
〈/name〉〈/respStmt〉
                〈item〉헤더 마크업 및 본문 태깅 붙임〈/item〉
        〈/change〉
    〈/revisionDesc〉
    〈/teiHeader〉
〈text〉
〈group〉
〈text〉
〈body〉
〈pb n='1'〉〈date〉광무 이년 구월 이십칠일 화요일〈/date〉
〈head〉뎨일권 뎨일신문 뎨빅이십팔호〈/head〉
```

그런데 이 표준양식은 역사자료의 서지 정보를 표시하는 항목의 기술 내용이 정밀하지 않다는 문제점이 있다(홍윤표 2001a:49). 이것은 헤더 양식 자체의 한계로 생긴 문제이다. 홍윤표(2001a:49)에서는 원자료의 서지 정보를 이미지 자료로 만들어서 헤더 앞에 붙어서 이 문제를 해결하자는 방안을 제시했다. 물론 홍윤표(2001a:49)도 한 가지 개선 방안이긴 하다. 그러나 text 자료와 이미지 자료를 혼합한 원시 말뭉치는 자료 이용자들에게 친절한 데이터라고 말할 수 없다. 이런 점을 고려해서 헤더 자체를 서지 정보가 충분히 들어가도록 개선하는 것이 나을 것이다.[20]

20) 헤더에 필요한 서지 정보는 홍윤표(2001a:49)에서 잘 소개되어 있다.

Ⅷ. 결 론

지금까지 말뭉치 구축의 일반론으로 말뭉치의 현황 및 요건, 말뭉치의 유형, 말뭉치 수집의 방법과 절차를 소개하였으며, 아울러 역사자료 말뭉치의 필요성을 갈파하고, 개화기 말뭉치 구축 절차와 그 문제점 그리고 해결 방안을 개화기 한글자료를 토대로 검토해 보았다. 본문의 주요 내용을 요약하고 앞으로의 과제를 언급하는 것으로 결론을 삼겠다.

말뭉치는 다양한 유형으로 구분되는데 결과적으로 개화기 한글자료 말뭉치는 원시 말뭉치이면서 일반 말뭉치이기도 하고, 정적 말뭉치이며 공시 말뭉치이며, 역사 말뭉치이다. 또 '기회적 말뭉치, 단일어 말뭉치, 원문 말뭉치이다.

어떤 자료가 역사자료이든 아니든 한글자료라면, 국어의 정보화를 위해서 가급적 빠른 시일내에 데이터베이스화되어야 마땅하다. 개화기자료에 대해 역사자료인가, 아닌가를 논의하는 것은 최소한 국어의 정보화를 위해서는 무의미하며, 차라리 전체 한글자료를 데이터베이스화하는 데 있어서 개화기 자료의 구축이 다른 시기의 자료구축보다 우리에게 얼마만큼의 효용이 있는가를 평가해야 마땅하다. 이미 다른 시기의 자료들은 기대이상으로 데이터베이스화되어 있다는 점은 우리들에게 개화기자료 구축의 필요성을 더 느끼게 한다.

역사자료의 특수성을 감안할 때, 개화기 역사 말뭉치의 구축 절차는 다음과 같다. 먼저 개화기 한글자료를 가지고 원시 말뭉치를 구축하기 위해서는 첫째, 원자료 수집을 해야 한다. 둘째, 수집된 자료는 텍스트의 성격에 따라 범주별로 구별해야 한다. 셋째, 범주별로 자료를 기계가독형으로 입력한다. 넷째, 원자료와 말뭉치자료를 일일이 비교하면서 교정을 보아야 한다. 마지막으로 교정이 끝나면, 검색과 분석의 효용을 돕기 위해 헤더 마크업을 비롯하여 본문 마크업 태깅을 달아야 한다.

개화기 한글자료의 말뭉치 구축 절차를 염두에 둘 때, 생기는 문제점으로 우선 띄어쓰기의 통일성을 기해야 한다는 것을 들 수 있다. 옛글 처리 즉

표기법 처리도 문제이다. 기타의 문제점으로 원자료 자체의 오자나 탈자, 입력시에 발생한 오자나 탈자, 원자료 자체의 일부가 판독 불가능한 경우도 빼놓을 수 없는 것이다. 이외에 자료의 저장방식, 저장파일명의 표준화도 문제가 된다.

띄어쓰기 해결 방안으로 개화기 한글자료의 띄어쓰기 경향에 따라 세 유형으로 띄어쓰기를 파악하되, 세 유형의 텍스트들은 충분한 교정을 거친 이후, 믿을 수 있는 검색과 분석이 가능하도록 현대국어 어법에 맞는 띄어쓰기 지침을 마련하여 통일되게 띄어쓰기를 적용하도록 한다. 옛글들은 그대로 반영해서 데이터베이스화한다. 한편 동일한 의미와 유사한 자소형을 갖는 이표기들은 모두 드러내는 방식으로 어절 처리한다. 원문의 오자는 [ ] 를 이용해서 바로잡은 어형과 함께 데이터베이스화한다. 원자료의 일부를 판독할 수 없을 때는 〈gap reason='몇 자 판독 불가'〉로 처리한다. 마지막으로 저장방식은 통일된 저장방식을 유지하는 것으로 원칙을 삼는다.

앞으로의 과제는 이렇게 해서 구축된 개화기 말뭉치를 활용하여 색인, 문맥색인, 핵심어 문맥색인뿐만 아니라 단어나 형태소 사이의 통합관계 및 문법적 단위의 용법을 좀더 다양하고 구체적으로 파악해 보고 나아가 자소, 음절, 어절 빈도를 검토하는 것이다. 더나아가 현대국어 말뭉치 처리 결과와 비교하여 그 공통점과 차이점을 밝히는 것도 중요한 과제가 될 것이다.

참고문헌

김봉희(1999), 『한국개화기 서적 문화 연구』, 이화여자대학교 출판부.

남윤진(1997), 현대국어의 조사에 대한 계량언어학적 연구, 서울대학교 박사학위논문.

서상규 · 한영균(1999), 『국어정보학 입문』, 태학사.

신중진(2003), 개화기 주요 한글자료를 찾아서, 『인하어문연구』 6, 인하대학교.

신중진(2004ㄱ), 개화기국어 신문·잡지의 인간 관련 명사 연구, 서울대 박사학위논문.

신중진(2004ㄴ), 개화기 신문·잡지에 쓰인 고빈도 동음이의어 고찰, 『한국문화』 33, 서울대 한국문화연구원.

이현희(1999), 개화기국어 자료, 『국어의 시대별 변천 연구』 4-개화기국어편, 국립국어원.

조성오(1993), 『우리역사이야기』-조선후기에서 식민지시기까지, 돌베개.

한영균(1994), 후기중세국어의 모음조화 연구, 서울대 박사학위논문.

한영균(2003), 어휘 계량적 분석과 띄어쓰기 문제, 『한국문화』 31, 한국문화연구소.

한영균(2004), 문법화와 연어 구성 변화-'있-'의 경우-, 『국어학』 44, 국어학회.

홍윤표(2001a), 한국어 전자 자료의 수집과 정리 및 활용 방안, 『새국어생활』 11-2, 국립국어원.

홍윤표(2001b), 국어사 자료 코퍼스의 구축 현황과 과제, 『한국어학회 전국학술대회 논문집』, 한국어학회.

Biber. D.&S. Conrad&R. Reppen (1998), *Corpus linguistics: Investigating language structure and use*, Cambridge Univ. Press.