일본어 코퍼스의 구축 및 활용과 관련된 문제에 대하여*

李 漢 燮**
眞島知秀***

目 次

1. 들어가며

본고는 일본어 코퍼스의 구축 및 활용 문제에 대하여 언급한 것이다. 고려대학교 일어일문학과 이한섭 교수 연구실에서는 1995년도부터 일본어 코퍼스 구축을 시작하였으며 세 단계로 나누어 구축 작업을 진행하고 있다.

제1단계(1995~2001년)에서는 원시 코퍼스 구축 작업이 이루어졌다. 이 시기에는 일본어 자료를 직접 입력하고, 공개된 자료를 수집하여 원시 코퍼스를 구축하였으며 또 이를 관리하는 방법에 대한 기초적인 연구가 진행되었다. 이 시기에 구축된 자료는 소설과 시, 수필, 신문, 잡지기사, 교과서, 고전자료 등 약 1,000만 어절 이상이다.

제2단계(2002~2006년)에서는 코퍼스 구축과 수집을 병행하면서 구축된 코퍼스를 가공하고 표준화하는 작업 단계이다. 현재는 약 1,000만 어절 분량의 코퍼스에 대하여 규격화·표준화 작업을 수행하고 있으며 범용성을 염두에 두면서 코퍼스의 가공 작업을 진행시키고

* 이 논문은 2005년도 고려대학교 특별연구비 지원에 의한 것임.
** 고려대학교 교수 일본어학
*** 고려대학교 박사과정 일본어학

있다. 또한 지금까지의 연구 성과를 바탕으로 일부 코퍼스를 인터넷에 공개하고 있다.

제3단계(2007~2011년)에서는 코퍼스의 활용 방안을 모색하는 단계이다. 이 시기에는 일본어 코퍼스의 효과적 활용을 위한 제반 연구가 이루어질 것이며 구축 자료의 형태소 주석 작업도 병행해 나갈 예정이다.

본고에서는 이한섭 교수 연구실에서 구축 중인 일본어 코퍼스를 대상으로, 코퍼스 구축과 검색시스템의 개발과정에서 발견된 제반 문제에 대하여 살펴보고 코퍼스의 활용 문제에 대해서도 몇 가지 생각을 표하고자 한다. 여기서 언급한 사항들은 금후 일본어 코퍼스를 구축하고 검색시스템의 개발하는 사람들에게 다소 참고가 될 것으로 믿는다.

2. 코퍼스 활용의 의의

코퍼스(Corpus)라는 용어는 원래 한 작가의 텍스트나 발화를 망라한 대규모 언어 자료의 총체를 의미하며 애초에는 서적 등 인쇄물을 매체로 한 문자 언어의 자료라는 의미로 쓰였다[1]. 그러나 브라운 코퍼스(Brown Corpus, 1964)와 롭 코퍼스(LOB Corpus, 1978) 등 100만 어절을 넘는 대규모 코퍼스가 구축되고 이들 코퍼스 자료를 컴퓨터를 통하여 분석하는 방법이 도입된 이후로는 코퍼스는 '컴퓨터에 의한 언어 분석용 전자화 텍스트의 집합'이라는 의미로 이해되게 되었다. 그리고 현재는 텍스트 자료뿐만 아니라 컴퓨터로 분석 가능한 음성 파일 형태도 음성 코퍼스라고 불리며 음성 언어 연구 분야에 활용되고 있다.

대규모 코퍼스의 구축은 개인 레벨의 작업으로는 시간적·능력적 제약을 받게 되므로 주로 대학이나 국가 연구 기관이 구축하는 것이 통례이다. 일본에서는 名古屋大學[2]과 京都大學[3] 등에서 각기 특색이 있는 코퍼스 구축이 이루어지고 있으며 국가 연구 기관인 國立國語研究所[4]에서도 각종 코퍼스의 연구개발이 이루어지고 있다. 한국에서는 국립국어원[5]과 한국과학기술원(KAIST)[6] 등이 코퍼스 구축 작업을 하고 있고 일본어 코퍼스는

---

1) 廣義の意味でのコーパスとは「言語分析のための文字言語、あるいは音聲言語の資料の集合体」("The body of written or spoken material upon which a linguistic analysis is based" The Oxford English Dictionary 1956)を指し、当初は書籍をはじめとする紙の媒体の資料を主に想定していた(伊藤雅光 2002 p.5)

2) 『名古屋大學學習者コーパス』 취로(就勞) 브라질인의 발화 문자화 자료 및 일본어 연수생 추적 데이터 등이 있다.

3) 『京都テキストコーパス』 每日新聞 기사에 각종 언어 정보를 부여한 코퍼스로 1995년 1월 1일부터 17일까지의 모든 기사 약 2만 문장과 1월부터 12월까지의 사설 기사 약 2만 문장에 대해 형태소 및 구문 정보가 부여되고 있다.

4) 대표적인 코퍼스로서는 『日本語話し言葉コーパス』, 『全國方言談話データベース』, 『太陽コーパス』, 『日本語學習者による日本語作文とその母語譯との對譯データベース オンライン版』 등이 있다.

5) 21세기 세종계획(1998-2007)에서 다양한 분야의 코퍼스 구축 사업이 진행되고 있다. 대표적인 것으로 국어 기초자료 구축, 국어 특수자료 구축, 전자 사전 개발, 전문용어 표준화, 한민족 언어 정보화 사업 등이 있다.

고려대일어일문학과 이한섭 교수 연구실 등에서 구축작업이 이루어지고 있다.

코퍼스 활용의 의의는 코퍼스를 통하여 언어 현상의 사용 실태를 관찰함으로써 어문 규칙과 의미, 장면에 따른 용법 등을 귀납적으로 발견할 수 있다는 점에 있다. 언어 현상의 사용 실태는 인간의 사고에 의해서도 밝힐 수 있으나 그것을 체계적으로 파악하기에는 쉬운 일이 아니다. 또한 자기 관찰에 의한 언어 현상의 규칙을 구체적·객관적으로 제시하는 것은 상당한 언어학적인 센스가 요구된다. 이와 관련하여 특정 표현이 문법적인가에 대한 판단이 어려운 경우, 코퍼스에서 검출된 내용을 관찰하면 쉽게 이를 판단할 수 있다. 특히 일본어를 모국어로 하지 않는 외국인 연구자나 학습자에게 코퍼스 자료는 하나의 커다란 판단 기준이 될 수가 있을 것이다. 일본어 코퍼스가 일본어 연구와 교육에 사용될 자료로서 주목받는 이유는 이 때문이다.

3. 일본어 코퍼스의 구축 및 활용에 관한 문제

이한섭 교수 연구실에서 구축한 제1단계 작업의 결과물은 준비 부족과 경험 미숙으로 여러 가지 문제가 내포되어 있다. 이들 문제는 구축 당시에도 어느 정도 예견한 일이었으나 제1단계 작업 완료 후 구축 자료를 정리하고 검색 시스템을 개발하는 과정에서 그 문제점이 드러났다. 이에 따라 최종작업에서는 코퍼스 전반에 걸쳐 여러 가지 수정이 이루어졌는데 본고에서는 위의 작업을 통하여 알게 된 코퍼스의 구축시의 문제와 활용방안에 대하여 초점을 맞추어 살펴보고자 한다.

3.1. 자료를 선정할 때의 문제

코퍼스를 구축하는데 우선 문제가 되는 것은 자료 선정이다. 자료를 선정할 때는 장르와 시대, 자료의 성격(문어체인가 회화체인가 등)을 고려하여 자료 상호간의 균형성을 확보하는 것이 무엇보다 중요하다. 또한 자료 선정 후 실제로 코퍼스 구축 작업에 들어갈 때는 구축 대상 자료를 입수하는 것도 문제가 되지만 저작권을 확보하는 것도 그에 못지않게 중요하다. 저작권이 확보되지 않은 자료는 코퍼스로 구축했다 하더라도 공개가 불가능하며, 따라서 전자 매체를 통하여 수많은 사용자와 자료를 공유한다고 하는 코퍼스 본래의 정신과 맞지 않게 된다.

---

6) 국어정보 베이스(http://kibs.kaist.ac.kr/)에서는 지금까지 구축해온 코퍼스 자료의 일부를 공개하고 있다. (한국어/일본어 용례검색, 전자사전 검색, 전문용어 언어공학, 한국어 정보처리 등)

3.2. 자료를 입력할 때의 문제

코퍼스를 구축한다는 것은 컴퓨터를 통하여 대상 자료를 전자화 하는 작업을 말한다. 자료의 전자화에는 스캐너를 사용하기도 하나 대부분은 키보드를 통하여 글자 하나하나를 입력해나가는 방법을 사용한다. 이 때문에 코퍼스를 구축하는 데는 컴퓨터 사용과 관련된 여러 가지 사항을 고려할 필요가 있다. 자료를 입력할 때 발생하는 문제로서는 다음과 같은 것들이 있다.

① 한자의 舊字體・新字體 차이[7]
　　樂(6A02) / 楽(697D), 國(570B) / 国(56FD), 學(5B66) / 学(5B78)
② 일반문자・영어숫자의 전각・반각 차이
　　ア(30A2) / ｱ(FF71), 1(FF11) / 1(31), A(FF21) / A(41)
③ 기호문자의 전각・반각 차이
　　？(FF1F) / ?(3F), ＜(FF1C) / <(3C), 「(300C) / ｢(FF62)
④ 동일형태 문자의 코트 값 차이
　　李(F9E1) / 李(674E), 羅(F90F) / 羅(7F85), 力(F98A) / 力(529B)

①의 字體面에서는 舊字體와 新字體의 차이에 주의해야 한다. 즉 입력하는 사람이 자기 멋대로 舊字體 또는 新字體로 입력해서는 안 된다는 것이다. 筆寫와 달리 입력 작업은 키보드를 통해 순식간에 이루어지는 만큼 漢字는 무의식적으로 다른 字體로 잘 못 입력할 가능성이 있다. 舊字體가 대량으로 쓰인 古書나 舊字體와 新字體가 섞어있는 문헌을 입력할 경우에는 이들 字體의 차이를 잘 이해하고 세심한 부분까지 신경을 쓰면서 입력 작업에 임해야 할 것이다.

②와 ③의 전각・반각의 문제는 입력 방법에 따라 여러 가지 형태로 입력되게 됨으로 가급적 입력 방식을 통일시키는 것이 향후 코퍼스 활용에 도움이 될 것이다[8]. 그렇지 않으면 관련 정보를 검색하는 데에 전각 문자와 반각 문자 양쪽을 다 검색해야 하므로 활용성이나 자료적 가치 면에서 질적으로 떨어지게 된다.

④의 경우는 특히 주의를 요하는데, 언뜻 보기에는 같은 형태의 문자로 보이나 각각의 문자에 서로 다른 코드 값이 부여된 것들이다. 이러한 예는 일본어 문자 코드(JIS, Shift-JIS, EUC)에는 존재하지 않으나 유니코드(Unicode) 문서에서는 표시 가능한 문자들이다. 엄밀하게 말하면 이 문제는 한국어 문자 코드에 존재하는 것으로서 한국어 입력 방

7) 이하 숫자와 알파벳으로 된 괄호 내 네 자리 정보는 문자 코드(Unicode) 값을 나타낸다.
8) 일본어 입력 모드로 글자를 입력할 경우 입력 변환 후보 창에 전각 문자, 반각 문자, 기타 관련 기호 등이 모두 표시되는데, 이 중에서 적절한 글자를 선택하여 입력 방식을 통일시키는 것이 중요하다

식에 따라 한글을 한자로 변환하는 과정에서 일어날 수 있는 것이라 하겠다. 예를 들면 한국인 성씨 이씨의 경우, "이"로 입력하여 한자로 변환했을 경우와 "리"로 입력하여 변환한 한자의 문자코드가 다르다는 것이다. "이"로 입력했을 경우 출력된 한자 "李"는 문자코드로는 F9E1에 해당되며, "리"의 문자코드는 674E이다. "李(F9E1)"와 "李(674E)"는 시각적으로 같은 형태의 문자이므로 인간의 인지에서는 당연히 동일한 뜻을 가진 문자로 인식되나 컴퓨터는 각각 배당된 문자 코드 값이 다르므로 결과적으로 다른 문자가 되는 것이다. 따라서 일본어와 한국어가 혼재하는 자료를 전자화 할 경우에는 이러한 점에 유의해야 할 것이다.

즉 원문에 충실하면서 문자 코드 면에서도 일관되게 입력하는 것이 매우 중요하다는 점이다. 그렇지 않으면 검색에서 같은 문자로 처리되어야 할 것이 검색 결과로부터 제외되게 된다. 결과적으로 이러한 코퍼스는 자료적 가치가 손상되어 애써서 구축한 코퍼스가 쓸모없는 것이 되고 말 것이다.

3.3. 자료를 가공할 때의 문제

인터넷상에 공개된 자료를 내려받거나 아는 사람으로부터 입수한 전자자료 대부분은 그대로 사용하기 어렵다. 그 이유는 자료를 입력한 사람이 아무런 약속 없이 데이터를 입력한 경우가 대부분이고, 또 입력 방법의 통일성이 결여되어 있기 때문이다. 이 때문에 이들 자료를 코퍼스 구축에 사용하기 위해서는 각종 事前 가공 작업이 필요하다.

코퍼스 자료를 가공할 때 고려해야 할 점으로는 앞서 언급한 한자의 字體(舊字體, 新字體) 문제와 전각·반각의 문제, 한자의 讀音의 차이에 따른 코드 값의 문제 등이 있으며 다음에 열거하는 문제들도 주의해야 할 작업들이다.

① 파일 형식의 통일 (텍스트 형식(.txt), 아래아 한글 형식(.hwp), 워드 형식(.doc) 등)
② 파일 코드 형식의 통일 (JIS、Shift-JIS、EUC、Unicode)
③ 오타 수정 (장음표시 "ー"와 대시 "—" 문자의 혼동, 한자 "二"와 가타카나 "ニ", 히라가나 "へ"와 가타카나 "ヘ" 문자의 혼동 등)
④ 강제로 改行된 부분의 편집 (문장 중에서 강제로 enter가 들어간 것)
⑤ 기타 불필요한 문자열의 삭제 (문자열 내의 불필요한 스페이스 및 탭 등)

우선 ①의 파일 형식을 통일하는 것이 중요하다. 파일 형식은 텍스트 형식(.txt)과 아래아 한글 형식(.hwp), 워드 형식(.doc) 등이 있으나 코퍼스를 구축할 때는 이들 형식을 통일시킬 필요가 있다. 여러 가지 파일 형식이 혼재된 코퍼스는 구축 자체도 어려울 뿐만 아니라 구축한 뒤에도 검색 면에서 일괄처리가 어려워지는 등 문제가 발생한다. 파일 형

식은 범용성이 높은 텍스트 파일 형식(.txt)으로 통일시키는 것이 좋을 것이다.

다음은 ②의 코드 형식의 통일 문제이다. 자료를 입력할 때 사용하는 한글 코드로는 KS 코드와 Unicode 등이 있으며 일본어 코드로는 JIS、Shift-JIS、EUC 등을 들 수 있다. 자료의 코드를 통일시키지 않았을 때는 위의 ①에서 언급한 문제와 비슷한 문제가 발생하게 된다. 코퍼스를 이용할 경우 일반적으로 복수의 파일을 대상으로 일괄적인 검색을 실시하기 때문에 코드 형식이 통일되어 있지 않으면 검색되지 않은 파일이 발생하여 정확한 검색 결과를 얻지 못 하게 된다.

③의 오타 수정은 단순 입력 오류나 한자를 변환시킬 때의 오류 등을 비롯하여 입력자 실수로 다른 문자 코드로 입력된 것들도 포함한다. 예를 들면 일본어 입력 모드에서 문자를 변환시킬 경우, 장음 표시 "ー"와 대시 "—"를 혼동하는 수가 있다. 또 한자 "二"와 가타카나 "ニ"를 혼동해서 입력하거나 히라가나 "へ"와 가타카나 "ヘ"를 같은 문자로 처리하는 것도 흔히 일어날 수 있는 문제이다. 그리고 키보드의 문자판 배열이 인접해 있는 숫자 "0"과 알파벳 "o"의 경우, 그 형태가 비슷하여 잘 못 입력하더라도 그 오류가 눈에 띄지 않을 가능성이 있다. 그 밖에도 전각 숫자의 "０"과 漢數字의 "〇", 그리고 원형 기호 "○" 등은 글꼴 설정에 따라 언뜻 보기에 구별하기가 어려운 예들이다

④의 강제로 改行된 부분의 편집도 문제이다. 이 경우는 입력자가 원본 자료의 편집 상태를 입력에 그대로 반영시키려고 원본 자료의 개행 위치나 페이지의 편집 상태를 충실하게 전자 텍스트화 하였을 경우에 나타나는 것으로, 문장 사이에 강제로 enter를 넣어 개행이 이루어진 경우이다. 이렇게 문장 중간에서 줄이 바뀔 경우 인간의 인지에서는 아무런 문제가 없으나 컴퓨터상에서는 강제로 개행을 하게 되면 다른 문장이 시작되는 것으로 인식되게 된다. 그래서 검색 대상 문자열 사이에 개행된 것이 위치하면 그 부분의 검색이 불가능하게 된다. 인터넷상에 공개된 자료를 보면 이와 같이 강제로 행이 바뀐 자료가 적지 않으므로 주의를 요한다.

⑤의 문장 사이에 나오는 불필요한 문자열의 삭제도 필요한 작업이다. 문장 사이에 빈칸(space)이나 탭(tab)이 들어가 있을 경우 분석 결과에 치명적인 영향을 미칠 수 있다. 용례를 눈으로 찾는 경우는 이들 빈칸이나 탭이 들어간 것이 있다 하더라도 아무 문제없이 찾을 수가 있으나 컴퓨터로 검색할 경우는 검색어와 검색 대상 문자열이 완벽하게 일치하지 않으면 검색되지 않기 때문이다. 이러한 점에 유의하지 않으면 잘 못 한 분석 결과를 아무 의심 없이 그대로 받아들일 수도 있는 웃지 못 할 상황이 발생하게 된다. 특히 연구 목적으로 용례 검색을 할 경우는 세심한 주의가 필요하다.

또 소위 루비(ruby)가 달린 문장도 문제가 된다. 루비는 한자 등의 문자열 바로 위에 讀音을 단 것인데 코퍼스 자료에서는 편의상 한자 단어의 바로 뒤에 루비를 표시하는 경우가 있다. 코퍼스를 이용하는데 있어서 특별히 루비 정보가 필요로 하지 않은 경우는 이를 제거하는 것이 좋다고 본다[9]. 검색 대상 문자열 사이에 루비가 위치할 경우 검색되지 않

기 때문이다.

3.4. 코퍼스를 활용할 때의 문제

코퍼스 자료를 연구에 사용하기 위해서는 각 연구 목적에 맞는 프로그램이 필요하며 이 때문에 코퍼스 데이터를 실제 사용하려면 프로그램 사용법을 습득할 필요가 있다. 코퍼스 활용과 관련된 문제로는 다음 두 가지를 들 수 있다.

① 사용할 검색 프로그램에 따른 문제
② 컴퓨터의 언어 환경(OS)에 따른 문제

우선 ①에 대하여 살펴보기로 한다. 코퍼스 자료에서 언어 정보를 추출하려면 검색 프로그램을 사용하게 되는데 이 때 사용되는 검색 프로그램으로는 워드프로세서와 전문적인 검색 프로그램으로 대별된다. 코퍼스 자료에서 특정 언어 정보 추출해내기 위해서는 "검색어 입력" → "검색" → "검색 결과 출력" 과정을 거치게 되며 이 때 파일 수가 적거나 검색어가 간단할 경우는 워드프로세서의 검색 기능만으로도 충분하다. 그러나 보다 복잡한 검색 조건으로 대량의 용례를 검색하려고 할 때는 전문적인 프로그램이 필요하다. 전문적인 검색 프로그램에서 단 한 번의 조작으로 끝낼 수 있는 작업이 워드프로세서에서는 수 십 번, 수 백 번씩 반복 작업을 해야 할 경우가 있기 때문이다. 전문적인 검색 프로그램은 단시간에 대량의 언어 데이터를 정확하게 처리할 수 있도록 개발되어 있기에 코퍼스 활용의 기본 이념에 부합된다. 소량의 코퍼스를 샘플적으로 이용할 경우를 제외하고 대량의 코퍼스를 워드프로세서를 이용하여 하나하나 수작업으로 검색해나가는 방법은 코퍼스 자료 사용법으로는 부적절하다고 하겠다.

다음은 ②의 컴퓨터의 언어 환경에 따른 문제이다. 일본어코퍼스를 이용할 경우에는 일본어를 지원하는 검색 프로그램이 필요한데 이들 검색 프로그램의 사용 환경은 대부분 일본어 OS(일본어 Windows 환경)에 맞도록 설계되어 있다. 이 때문에 일본어 OS 환경을 갖추지 않은 연구자는 문자가 깨져서 전용 프로그램의 사용이 불가능하게 된다. 일본지역

9) 青空文庫(http://www.aozora.gr.jp/)에서 공개된 코퍼스 자료에는 대부분 다음과 같이 루비가 부여되어 있다. 루비는 언어학 연구 측면에서 아주 유용한 정보가 되지만 단순 문자열 검색으로 용례를 수집할 경우에는 오히려 검색에 장애가 될 수도 있다. 예를 들어 "親譲り"라는 단어를 찾고자 할 때 코퍼스 자료에는 "親譲《おやゆず》り"라고 단어 사이에 루비가 들어가 있으므로 단순히 " 親譲り" 라는 검색어로는 검색 결과가 나오지 않은 문제를 말한다.

> 親譲《おやゆず》りの無鐵砲《むてっぽう》 で小供の時から損ばかりしている。小學校に居る時分學校の二階から飛び降りて一週間ほど腰《こし》を抜《ぬ》かした事がある。【夏目漱石 / 坊っちゃん】

외에서 일본어 코드로 작성된 코퍼스 자료를 사용하기 위해서는 일본어 OS 환경 구축이 필수불가결이다. 그러나 최근에는 컴퓨터 자체의 진보와 프로그램 기능의 향상으로 문자 코드 문제도 해소되어 가고 있으며 일본어 OS 이외의 환경에서도 일본어를 표시하고 분석이 가능한 검색 프로그램이 나오고 있다.

4. 새로운 코퍼스 검색 시스템의 개발

일본어 코퍼스 활용의 저변을 확대하기 위해서는 활용 시스템의 개발이 필요하다. 이한섭 교수 연구실에서는 제2단계 작업의 주요 과제로 '일본어 코퍼스 검색 시스템'[10] 및 '한일 병렬 코퍼스 검색 시스템'[11]을 개발하였으며 현재 이를 시험적으로 공개·운영하고 있다. 본 시스템의 동작 화면은 다음 <그림1> 과 <그림2> 와 같다

이 시스템의 특징은 우선 OS면에서 고속으로 텍스트 처리가 가능한 리눅스(Linux)[12]를 도입하였고 데이터베이스에는 마이에스큐엘(MySQL)[13]을 채택했다는 점을 들 수 있다. 그리고 이용자와의 데이터 전달은 PHP[14] 스크립트 언어를 매체로 HTML 표시를 하는 방식을 택하고 있다. 현재 등록된 파일수는 약 300개 정도이나 앞으로 자료량을 늘려 최종적으로는 저작권 문제가 없는 약 4,000개의 파일을 올릴 계획이다.

이하에서는 본 시스템의 기능별 특징에 대하여 소개하고자 한다.

① 온라인 검색으로 세계 어디에서도 검색이 가능하다

전용 검색 프로그램의 경우, 인터넷을 통해 이용이 가능한 것과 자신의 컴퓨터에 프로그램을 설치한 후 이용할 수 있는 것이 있으나 본 시스템에서는 인터넷 열람 프로그램(브라우저)을 사용한 온라인 검색 방식을 택함으로써 세계 어디에서도 검색을 가능하게 하였다. 또한 사용 문자 코드로 유니코드(UTF-8)를 채택함으로써 일본어 문자를 표시하기 위한 글꼴(Font) 설치 등 특별한 설정 없이 바로 이용할 수 있다.

---

10) 일본어 코퍼스 검색 시스템은 http://www.transkj.com/fstab/search.html
혹은 http://www.transkj.com 에서 들어갈 수 있다.

11) 한일 병렬 코퍼스 검색 시스템은 http://transkj.com/in.htm
혹은 http://www.transkj.com/ 에서 들어갈 수 있다

12) 리눅스(Linux)은 Windows와 달리 무상으로 입수 가능한 OS 시스템으로 장시간 안정된 동작이 가능하기 때문에 Website의 서버 사용에 적합하다.

13) 마이에스큐엘(MySQL)은 무상으로 입수 가능한 데이터베이스 관리 시스템이다. 기타 유상으로 제공되는 일반 상용 데이터베이스 시스템에 비해 많은 기능을 제외시켜 데이터 처리 속도 면에 중점을 둔 설계로 되어 있기 때문에 대량의 데이터를 고속으로 검색할 수 있다.

14) PHP는 Web 공간에서 많이 사용되는 프로그래밍 언어의 하나로 JavaScript와 같이 HTML 파일 안에 바로 기술하여 Website에 동적인 효과를 제공한다.

<그림1> 일본어 코퍼스 검색 시스템 [http://www.transkj.com/fstab/search.html]

<그림2>한일 병렬 코퍼스 검색 시스템 [http://www.transkj.com/in.htm]

② 대량의 코퍼스에서 용례를 검색할 수 있다

현재 인터넷상에서는 여러 가지 코퍼스가 공개되어 있으나 여러 곳에 흩어져 있는 코퍼스를 조사하여 수집하는 것 자체도 시간과 인내를 요하는 일이다. 본 시스템은 금후 양적 확대를 계획하고 있으므로 앞으로는 이를 사용하면 대량 데이터의 검색이 가능할 것으로 보인다.

③ 특별한 코퍼스 전용 검색 프로그램이 필요하지 않다

본 코퍼스 검색 프로그램은 누구나 간편하게 용례 검색을 할 수 있도록 설계되어 있다. 기본 조작 방법은 Yahoo나 Google과 같은 검색 사이트에 들어가서 검색하는 것과 유사한 방법으로 아주 간단하게 이용할 수 있다.

④ 코퍼스 활용에 필요한 별도의 작업이 필요하지 않다

자신이 코퍼스를 구축하거나 인터넷상에서 입수한 코퍼스자료를 사용하여 용례를 검색하기 위해서는 데이터를 사용하는데 편리하도록 편집·가공해야 한다는 문제가 따른다. 이들 작업은 시간과 기술을 필요로 하기 때문에 컴퓨터 사용에 자신이 없는 사람은 처리가 쉽지 않다. 본 시스템에서는 누구에게나 간편하게 사용할 수 있도록 이미 편집 가공한 데이터를 수록하였다.

⑤ 단순하면서 다양한 기능을 갖추고 있다

조작법 자체를 알기 쉽고 단순하게 하여 "찾고자 하는 문자열의 입력"이나 "검색 결과의 표시" 등 일련의 과정이 누구에게나 알기 쉽게 구성되어 있다. 또한 다중 검색이 가능하며 아직 완벽하지는 않지만 정규식 검색도 가능하다.

검색 결과는 검색어가 포함된 센텐스를 한 페이지에 10개씩 표시되며 검색된 용례의 전후 문맥을 확인할 필요가 있을 때는 앞 뒤 문맥 표시용 단추를 누르면 바로 확인이 가능하다.

⑥ 자료의 상세 정보를 얻을 수 있다

본 시스템에서는 용례 검색 이외에 연구와 교육면에도 활용할 수 있도록 수록 파일 모두에 상세정보를 표시하였다. 본 코퍼스에서 상세정보에는 다음과 같은 것들이 있다.

- ○ 자료명(漢字)
- ○ 자료명의 히라가나 표시
- ○ 저자명(漢字)
- ○ 저자 이름의 히라가나 표시
- ○ 저자의 출생년도
- ○ 저자의 사망년도
- ○ 자료의 저본
- ○ 출판사명
- ○ 초판의 출판년월일

이와 같은 자료의 상세 정보는 연구자에게 다음과 같은 편의를 제공할 수 있을 것이다. 예를 들면 외국인 연구자(학습자)가 자료(작품)명이나 저자(필자명)의 한자를 어떻게 읽어야 할지 모를 경우와 해당 자료가 어느 시기에 출판된 것인가를 알고 싶을 때 유용할 것이다. 또한 저자의 출생 및 사망 시기나 초판 출판일 데이터를 이용하면 시대별 용례 검색도 가능하게 될 것이다.

⑦ 전문가를 위한 정규식 검색이 가능하다.

문자열 검색만으로는 만족하지 못하는 전문가를 위하여 정규식 검색이 가능하도록 하였다. 정규식 검색은 상세한 조건 검색을 가능하게 함으로써 초보자에서 상급자에 이르기까지 유익하게 사용될 것이 기대된다. 단 현재의 정규식 검색기능은 아직 미완으로서 개선의 여지가 있다. 정규표현의 검색식에 사용되는 메타 문자(meta-character) [15] 는 다음과 같다.

^ (줄의 맨 앞을 의미함)
$ (줄의 맨 뒤를 의미함)
. (개행을 제외한 임의의 1글자)
| ("|" 로 문자열 패턴의 구분)
? (직전의 문자를 0회 혹은 1회)
* (직전의 문자의 0회 이상의 반복)
+ (직전의 문자의 1회 이상의 반복)
{n} (n의 수치로 반복 횟수 지정)
{n,} (n의 수치로 n회 이상의 반복을 지정)
{n,m} (n 및 m의 수치로 n회 이상 m회 이하의 반복을 지정)
( ) ("( )" 로 문자열 패턴을 그룹화)
[ ] ("[ ]" 로 지정한 문자 하나에 매치)
[-] ("[-]" 하이픈으로 문자 코드를 범위 지정)
[^] ("[^]" 안에서 부정을 나타냄)

정규식 검색을 이용한 검색 방법의 예를 간략하게 소개하면 다음과 같다.

15) 메타 문자(meta-character)란 특별한 의미를 가지는 문자이며 정규식 검색을 실행할 때 사용된다. 예를 들어 [0-9] 라는 표현은 숫자 0부터 9까지, 즉 모든 숫자를 의미하며 숫자만을 검색하고자 할 때 이용된다. 같은 식으로 [A-Z] 는 A부터 Z까지의 모든 대문자를 나타내며 [ぁ 〆] 는 모든 히라가나를 검색할 때 사용된다. 이 때 '[ ]' 는 '[ ]'로 지정된 문자 하나를 검색함을 의미하며 '-' 는 문자 코드의 범위를 지정하는 역할을 한다. 이와 같이 '[ ]' 나 '-' 등이 정규식 검색으로 특별한 의미를 갖는데 이러한 문자를 메타 문자라고 한다.

① 활용하는 품사(동사「食べる」)의 모든 활용형을 한꺼번에 검색
【검색어】 食べ(る|ます|て|た|たり|よう|ながら)
② 영문자의 연속만 검색(알파펫으로 표기된 부분만 검색)
【검색어】 [A-Za-z]+
③「暮し」「暮らし」와 같이 표기에 복수의 방법이 존재하는 것을 검색
【검색어】 暮し|暮らし

다만 현재의 정규표현으로 인해 검색 시간이 더 소요되는 문제가 있기 때문에 앞으로는 서버에 부하가 적은 최소한의 정규표현만을 적용시키는 방향으로 개선을 검토하고 있다.

5. 활용 및 파급 효과

일본어 코퍼스는 연구에게는 물론 교육과 학습현장에서 유용하게 활용될 수 있다. 이하에서는 이들이 어디에 사용되며 그 파급효과가 어떠할지를 살펴보고자 한다.

① 일본어 연구자
일본 지역 외에서 일본어를 연구하는 사람들이 겪는 가장 큰 어려움의 하나는 연구 자료를 입수하는 일이다. 일본어 자료는 高價이며 입수가 가능하다 할지라도 시간적·경제적 부담이 적지 않다. 이러한 경우 해당 자료 코퍼스가 있다면 시간적·경제적 부담 없이 자료 접근이 가능하여 연구에 큰 도움이 될 것이다.
코퍼스를 활용하면 문법과 어휘, 의미, 형태론·담화론 등 각 분야의 언어학적 정보를 수집할 수 있고, 또 한일 병렬 코퍼스를 활용하면 대조연구에도 큰 도움이 될 것이다. 특히 언어 현상의 계량적 연구를 할 때는 코퍼스의 활용이 유용하다고 본다. 각 조사 항목을 카드에 기입·분류하여 계량적 결과를 도출하던 옛날과 달리 지금은 코퍼스로 순식간에 검색하여 정확한 그 결과를 얻을 수가 있기 때문이다. 이는 코퍼스가 가장 힘을 발휘하는 부분이라 할 수 있다.

② 일본어 교육 담당자
일본어 교육 담당자에게는 코퍼스가 항목별 교육 관련 용례를 수집하여 한정적인 용법이 아닌 체계적인 시각에서 본 일본어를 교육할 수 있게 해준다. 그리고 검색된 용례는 교재 제작에도 많은 도움을 줄 수 있을 것이다.
교사 자신의 직관으로 문법적 판단이 안 서는 부분은 코퍼스가 가져다주는 대량의 용례로 판단 기준을 마련할 수 있을 것이다. 문어체는 신문 코퍼스로, 회화체는 시나리오 코퍼

스를 이용하는 등 코퍼스의 장르를 고려하면 작문지도나 회화지도에도 문체에 맞는 적절한 표현을 제시할 수 있을 것이다.

교육 현장에서의 코퍼스 도입은 실제 용례를 구체적으로 제시할 수 있어서 학습자의 호기심을 유발시키고 학습 분위기 조성에도 도움을 줄 것이다. 특히 학습초보자에게는 한국어 대역문이 병기된 한일 병렬 코퍼스를 활용하는 것이 효과적일 것이다.

③ 일본어 학습자

일본어 코퍼스는 학습자에게 실제 용례를 보면서 문법이나 어휘 등을 학습할 수 있게 한다. 한일 병렬 코퍼스의 경우 학습자가 표현하고자 하는 일본어 표현을 우선 한국어로 검색해 볼 수 있다는 점에서 활용 가치가 높다. 일본어 작문 수업에서도 병렬 코퍼스는 자국어에 대응하는 일본어 표현을 바로 찾을 수 있다는 점에서 유익하게 사용될 것이다. 물론 검색 결과 중에서는 원하는 용례 이외의 것이 포함되어 있어 바로 사용하기 어려운 점도 있으나 이를 하나하나 검증하며 어떤 용례의 용법이 자신이 찾고자 하는 용법인지를 귀납적으로 분석할 수도 있을 것이다. 일본어코퍼스의 도입은 기존의 학습 방법과 달리 학습 초기 단계부터 실제 사용되는 자연스러운 표현에 많이 접할 수 있다는 장점이 있어 학습효과 면에서 탁월한 잠재력을 가지게 할 것으로 생각된다.

6. 마치며

이상 고려대학교 일어일문학과 일본어 코퍼스의 예를 중심으로 코퍼스 구축시의 문제점과 활용 방안에 대하여 살펴보았다.

오늘날 전 세계에서는 국가적 사업으로 자국어 코퍼스 구축 작업을 하는 나라가 적지 않다. 한국에서도 21세기 세종계획 사업(1998~2007년)이 8년 전부터 진행되고 있으며 앞으로 2년 후에는 3억 어절 규모의 한국어 코퍼스가 완성될 예정이다. 일본 國立國語研究所에서도 앞으로 일본어 코퍼스 구축 작업을 본격적으로 시작할 것으로 알려졌다. 또한 원 자료의 화상 자료 구축도 눈에 띄게 늘고 있다. 일본의 주요 고전 자료는 이미 많은 수가 화상 자료 형태로 코퍼스가 구축되어 공개되고 있으며 國立國會圖書館에서 공개한 명치자료 약 59,900권이 그 대표적인 예이다[16]. 이러한 변화를 생각할 때 앞으로 일본어 연구와 교육에서의 코퍼스의 활용은 점차 확대될 것으로 예상된다.

고려대 일어일문학과 일본어코퍼스 검색 시스템은 현재 시험 공개를 하면서 각종 보완

16) 『近代デジタルライブラリー』(http://kindai.ndl.go.jp/). 2005년 11월 현재 59,900권이며, 앞으로도 계속 추가될 것임.

책을 강구하고 있다. 금후 시스템의 안정화가 확인되면 일반에게 전면 공개할 예정으로 있어 많은 사람들의 활용이 기대된다. 그러나 코퍼스가 모든 것을 해결하는 것이 아니라 코퍼스를 통해 얻어낸 정보를 어떻게 활용할 것인지는 결국 사용자에 달려 있음을 명심해야 한다. 방법론이 바뀌어도 인간의 언어 현상을 관찰하여 다룰 수 있는 것은 아직까지 인간 밖에 없기 때문이다.

【참고문헌】

강범모 편역(1997), 전자텍스트 부호화 개설 : TEI라이트, 고려대학교 민족문화연구소

김흥규・강범모(1996), 고려대학교 한국어 말모둠 설계 및 구성, 한국어학 3, 한국어학회, 國立國語研究院의 '21세기 세종계획 관계 각종 보고서

李漢燮(2000) 일본어 코퍼스 구축에 관한 기본 구상, 언어정보 3, 고려대언어정보연구소, pp.33-51 일본어판

———(2002) 코퍼스를 활용한 일본어 어휘 연구에 대하여, 日本語學研究第5緝, 韓國日本語學會, pp.131-152

眞島知秀・金暻泳(2003) 일본어 주석 코퍼스(tagged corpus)의 구축 방법에 대하여, 日本學報第57卷, 韓國日本學會, pp.93-108

문화관광부・국립국어원(2003)『21세기 세종계획 국어특수자료구축 2004년 연구보고서』, 문화관광부・국립국어연구원

————————(2004)『21세기 세종계획 국어특수자료구축 2004년 연구보고서』, 문화관광부・국립국어원

伊藤雅光(2002)『計量言語學入門』, 大修館書店

後藤齊(2003)「言語理論と言語資料 ― コーパスとコーパス以外のデータ ―」,『日本語學』第22卷4月臨時增刊号「コーパス言語學」, pp.6-15

齊藤俊雄他(1998)『英語コーパス言語學』, 研究社出版

中尾浩他(2002)『コーパス言語學の技法Ⅰテキスト處理入門』, 夏目書房

要 旨

本稿は日本語コーパスの構築およびその活用に關連する諸問題について、高麗大學校日語日文學科李漢燮敎授研究室で行われているコーパスの研究で明らかになったことを中心に述べたものである。

大規模コーパスの構築は各大學や國家の研究機關を中心に行われており、近年のコンピューター技術の著しい發達と普及により、その利用は一般の研究者や學習者の手の届く範囲になりつつある。しかし、いざコーパスを構築したり活用したりするとなると、様々な問題や課題が明らかになる。日本語コーパスの構築および活用に關する諸問題としては、資料選定の際の問題、資料入力の際の問題、コーパス加工の際の問題、コーパス活用の際の問題があげられる。

資料選定の際の問題は、バランスの取れたジャンルの選定と著作權の確保といった点にある。資料入力の際の問題は、原文に忠實かつ一貫した入力規則に則って作業を進めることができるかという点にある。その点を深く理解せずに入力作業を行ってしまうと、檢索の際に檢索されるべき文字が除外されてしまう恐れがある。コーパス加工の際の問題は、ファイル形式の統一、文字コードの統一、タイプミスの修正などが擧げられる。また正しく入力されたファイルであっても、改行コードの編集や、スペースやタブなどの不必要な文字列の削除といったファイルの加工が行われていなければ、正確な檢索結果が得られないこともある点に注意しなければならない。コーパス活用の際の問題は、コーパスを構築 · 入手しても、專用の檢索プログラムを使いこなさなければならないといった課題がある。また專用の檢索プログラムを活用する際も、OSの言語環境により、日本語檢索專用プログラムが韓國語のOSでは使えない場合が多いといった根本的な問題点も存在する。

以上のようなコーパスの構築および活用に關する諸問題を少しでも解消し、コーパス活用の底辺擴大を図るべく、李漢燮敎授研究室では誰でも何處でも簡單にコーパスを活用できるような日本語コーパス檢索システムおよび日韓並列コーパス檢索システムの開發研究を進めている。様々な機能を持ったこの檢索システムを利用しながらコーパスを活用することで、次のような波及効果が期待される。

まず日本語研究者には研究資料入手に關わる時間的 · 經濟的負担を大幅に解消する効果がある。またコーパスを利用すれば意味論 · 形態論 · 語彙論など、様々な言語學的な情報を大量に收集可能であり、日韓並列コーパスを利用すれば、それぞれの分野の日韓對照研究へも研究範囲を廣げることができる。

日本語敎育者には學習項目別の用例を收集して、限定的な用例ではなく体系的に捉えた日本語を敎育することができる。また集められた大量の用例は、敎材の製作にも大いに活用することができるだろう。學習者への實際の指導の際にも、文法的か非文法的かの判斷が難しい表現については、コーパスによる大量の用例を根據として判斷することができよう。

日本語學習者にとっては、文法的 · 語彙的な學習事項を、實際の用例を見ながら確認することができ、大量の用例から連鎖的に出てくる新しい學習項目を、自ら學ぶことができる。コーパスの利用はこれまでの學習方法とは異なり、學習初期の段階から實際に使われる自然な表現に多く触れることができるのが最大の利点であり、學習効果の面で優れた潛在力を持っていると言える。

以上のように多方面で利用可能なコーパスはこれから益々發展していくことが予想され、それに向けた準備も必要であろうと思われる。

キーワード : コーパス, 檢索システム, 文字コード, 檢索語 正規表現檢索

투 고 : 2005. 11. 30
1차 심사 : 2005. 12. 10
2차 심사 : 2005. 12. 31

[필자연락처 1] - 李漢燮 (Lee Han-seop)
住 所 : (136-701) 서울 성북구 안암동 5-1 고려대학교 문과대학 507호
電 話 : 02-3290-2144
e-mail : lhs1001@korea.ac.kr

[필자연락처 2] - 眞島知秀 (Majima Tomohide)
住 所 : (136-861) 서울 성북구 종암1동 45-171 201호
電 話 : 02-929-5593
e-mail : majima@korea.ac.kr