# 전산 자료 및 체제의 구축 활용

조 남 호
(국립국어연구원 학예연구사)

사전 편찬 작업이 구체적으로 진행되기 시작한 1992년도에 사전 편찬 방향을 검토할 때부터 컴퓨터를 사전 편찬에 이용하는 방안이 적극 모색되었다. 이미 학계에서 국어 연구에 컴퓨터를 이용하려는 노력이 활발하였기 때문에 컴퓨터가 사전 편찬에 어떻게 이용될 수 있는가에 대해서는 이견이 별로 없었다. 그러나 그 당시 여건은 불충분하다 고조차 말하기 어려운 형편이었다. 연구원 전체를 통틀어 연구용으로 컴퓨터가 불과 2대 있었을 뿐이었다. 우선적으로 컴퓨터부터 확보해야 할 형편이었다. 그 이후 3년 간 지속적으로 전산 처리를 위한 노력이 경주되면서 여건은 상당히 개선되었다. 하드웨어만 해도 개인용 컴퓨터 50여 대, 국산주전산기인 타이컴 1대를 확보하였으며, 사전 편찬 전 과정을 지원하는 프로그램도 용역으로 개발하여 전산 처리에 큰 어려움이 없게 되었다. 사전 편찬에 관련된 입력 자료 역시 적지 않은 양이 확보되어 각종 정보를 찾는 데도 크게 도움을 받고 있다.

필자는 1992년도부터 사전 편찬 작업 중에서 주로 전산 처리와 관련된 업무를 담당하여 왔다. 사전 편찬에 관련된 입력 자료를 확보하고, 전산 시설을 갖추는 것이 필자의 주요 업무였다. 여기서는 그동안 필자가 맡아 왔던 일을 입력 자료 확보와 전산 시설 확보의 둘로 나누어 경과를 소개하고자 한다.

## 1. 입력 자료 확보

국어학계에서 컴퓨터에 관심을 둔 주요한 이유로는 연구 자료를 찾고 정리하는 작업이 수작업과는 비할 수 없이 빠르다는 점을 들 수 있다. 국어 사전 편찬 과정에서 검색과 정리는 개인 연구와는 비교할 수 없이 다량이고, 또 중요하다고 할 수 있다. 연구원에서 국어 사전 편찬 방향을 논의하면서 컴퓨터의 이용을 적극적으로 모색한 것도 이 때문이다. 학계에서 사전 편찬에 관한 논의를 하면서 컴퓨터를 이용할 필요성이 제기되었다는 점도 방향 설정에 도움이 되었다.

사전 편찬 초기에 설정한 방향은 새 사전 편찬의 기초 자료를 확보하기 위하여 기존 사전의 정보를 전산화하고, 실생활에서 사용된 언어 자료를 확보하는 것이었다. 사전 편찬 작업이 진행되면서 약간의 변동이 있기는 했지만 대체로 초기의 방향이 현재까지 이어지고 있다.

전산 처리에서 먼저 시작한 일은 기존 사전을 검토하는 일이었다. 이 작업을 추진하는 방안으로 두 가지가 논의되었다. 하나는 기존 사전을 그대로 입력하는 방안이었다. 다른 하나는 전통적인 방법이라고 할 수 있는 것으로 각 사전의 표제어를 오려서 한 카드에 붙이는 방안이었다. 필자는 전자의 방안을 주장하였었다. 이 방안이 아주 부정되었던 것은 아니지만 현실적으로 추진이 어렵다는 문제점이 있었다. 우선은 컴퓨터의 확보가 필수적인데 컴퓨터는 한 대도 없는 실정이었고, 입력을 전문으로 하는 사람을 구하는 것도 어려운 일이었다. 또한 사전을 입력한 후에 이용하기 위해서는 각 사전을 연관지어 검색할 수 있는 프로그램이 있어야 하는데, 예산이 없어 이 프로그램을 개발하기 어렵다는 문제도 있었다. 그래서 카드로 작업을 하되 한 사전은 입력도 하여 검색할 수 있도록 하기로 결정되었다. 입력 대상 사전은 현행 어문 규범에 따라 편찬된 대사전으로 하기로 하였다.

대사전 한 권의 입력도 쉽게 추진될 수 있는 작업은 아니었다. 필자가 입력 방안을 검토하는 과정에서 크게 두 가지 문제가 대두되었다. 첫째는 문자의 사용이었다. 사전에는 여타 문헌과는 비교가 되지 않을 정도로 다양한 문자가 사용되고 있다. 그동안 컴퓨터상의 문자 사용에 대해서는 학계에서 지속적으로 개선해 오기는 했지만 사전을 입력하기에는 문자의 수가 턱없이 부족하였다. 각종 기호를 사용하여 편법으로 입력하는 방법이 있기는 하나 필자의 경험으로는 이 역시 많은 혼란을 감수해야 한다. 둘째는 입력이 완료된 후의 활용 방안, 구체적으로 검색 방법의 문제였다. 사전은 데이터베이스로 볼 수 있으므로 데이터베이스 관리 프로그램을 이용하는 방안을 생각할 수 있었다. 그런데 데이터베이스에서는 정보의 길이가 고정되고 제한돼야 한다는 제약이 있다. 이

제약은 특히 뜻풀이를 입력하고자 할 때 큰 부담이 된다.

이러한 문제점을 고려하여 필자가 내린 결론은 문자를 많이 지원하는 문서편집기를 이용하여 자료를 입력하고, 그 입력 파일을 검색하는 프로그램을 따로 개발한다는 것이었다. 문서편집기로는 아래한글 2.0을 이용하기로 하였다. 아래한글 2.0에서는 한자 11,000여 자를 포함하여 다른 문서편집기에 비해 월등히 많은 문자가 지원된다는 것을 알고 있었기 때문이다.

'92년 초에는 아직 아래한글 2.0이 나오지 않았기 때문에 전체 입력은 뒤로 미루었다. 그 대신에 데이터베이스 관리 프로그램을 이용하여 표제어 중심으로 사전을 입력하는 일을 우선 시작하였다. 사전 편찬과 관련된 다른 작업을 지원하기 위하여 최대한 빨리 기존 사전의 정보를 이용할 수 있도록 준비되어야 한다는 지적이 있었기 때문이다.

세부적인 입력 방법은 필자가 정리하였는데 최단 기간에 입력이 완료되면서도 최대한 많은 정보를 수록할 수 있도록 입력 계획을 세웠다. 표제어는 사전의 표기대로 입력하였으며, 부표제어도 필드를 달리하여 사전의 표기대로 입력하였다. 각종 정보는 기호를 사용하여 해당 정보가 있다는 표시만 입력하였다. 순수한 뜻풀이에 대해서는 정보를 주지 않았다. 이 일을 위해서 입력 보조원 2명을 채용하였다. '92년 3월 말에 시작하여 9월 말에 작업을 완료하였다. 입력 보조원이 입력 초기에는 컴퓨터 사용이 익숙하지 않아 속도가 상당히 더뎠기 때문에 이 정도의 시간이 소요되었다.

이 표제어 목록은 여러 가지로 활용되었다. 그 중에 대표적인 경우가 6개 사전 표제어 목록을 만드는 작업에 그대로 이용한 경우이다. 카드를 오려붙이는 작업은 표제어 목록을 입력한 사전부터 시작을 하였는데 입력된 표제어 목록을 출력하여 카드에 제목을 붙이는 데 사용하였다. 그 다음 사전부터는 먼저 붙인 사전에 없는 표제어가 나올 때면 카드를 새로 만들고, 제목 표시에 사용하기 위해 이 표제어들을 기록해 두었다가 입력을 하곤 하였다. 이 때 만들어진 입력 파일을 보관했다가 나중에 카드 작업이 완료된 후 표제어 목록 파일과 합쳐 6개 사전의 표제어 목록 파일을 만들었다. 이렇게 만들어진 6개 사전 표제어 목록을 카드와 대조하여 오류 및 추가 정보를 정리하여 새 사전 표제어 선정의 기초 자료로 활용하고 있다.

전체 입력은 '93년 3월 중순부터 시작되었다. 'ㄱ'의 일부는 '92년에 입력이 이미 되어 있었다. 작업의 문제점을 파악하기 위하여 원외의 조사 보조원을 써서 입력을 하였었다. '93년 초에 입력 보조원을 3명을 더 채용하여 5명의 입력 보조원이 있었는데 이들이 모두 이 작업에 참여하였다. 최대한 사전의 원모습을 반영하되 검색이 필요할 것으로 생각되는 정보의 입력에는 특수기호를 사용하여 변별이 될 수 있도록 하였다. 후술하겠지만 '92년에 검색 프로그램이 개발되었기 때문에 입력 형식은 이 프로그램이 요

구하는 형식을 따랐다. 필자는 입력 지침을 정리하면서 한 번 입력된 자료가 최대한 활용될 수 있도록 사전을 세밀히 검토하여 검색 가능성이 있는 정보는 모두 검색될 수 있도록 입력 원칙을 정하였다. 그런데 입력이 완료된 후에야 한 가지 아쉬운 점이 발견되었다. 예문에 대해 시작 기호는 주었지만 끝 기호를 주지 않아서 예문만을 따로 뽑는 일이 쉽지 않다는 아쉬움이다. 예문만 따로 검색할 일이 있다는 데 미처 생각이 미치지 못한 것이다.

표제어 목록 입력으로 이미 입력된 정보 중에서 표제어, 부표제어 정보는 그대로 이용하였다. 원어의 경우도 따로 작업된 것을 이용하였다. 표제어 목록에서는 원어에 대해서는 고유어, 한자어, 외래어를 구분할 수 있는 영문 기호만을 입력해 두었다. '92년도에 표제어 분과-외래어 담당에서 이 목록 중에서 외래어라고 표시된 표제어만 뽑아 원어를 입력해 둔 것이 있었다. 표제어 분과-한자어 담당에서도 한자를 입력을 했는데, 일부밖에 하지 못했다. 입력 보조원이 한자까지 입력하면 입력 시간이 훨씬 많이 걸릴 것으로 생각되어 한자를 잘 아는 원외 조사 보조원을 써서 전체 입력보다 앞서면서 한자 입력이 진행되도록 하였다. 따라서 전체 입력은 발음, 뜻풀이 등 빠진 부분을 보충하는 형식으로 진행되었다.

이 작업은 8월 말이면 끝날 것으로 예상했었으나 11월 중순까지 계속되었다. 입력 보조원이 아래한글을 사용한 입력에 익숙해지는 데 시간이 필요하고 뜻풀이 중에 한자를 비롯하여 낯선 문자를 찾아 입력하는 데 의외로 시간이 든다는 점을 고려하지 않았기 때문이다.

사전 입력 자료는 국어를 연구하는 사람들에게 매우 유용한 자료이다. 그렇지만 이 자료를 외부의 사람이 이용할 수 있도록 할 경우 여러 가지 문제가 대두될 것이라는 우려가 있어 내부에서 연구용으로만 사용하는 것으로 결정되었다. 다만 외부 기관에 공식적으로 제공된 적은 있다. 고려대 언어정보연구소에서 사전을 전체 입력할 계획을 세우면서, 이미 입력된 것을 또 입력하기보다 다른 사전을 입력하여 공동으로 이용하는 것이 좋지 않겠느냐는 문의가 있었고 우리 원에서도 자료의 공동 이용에는 별다른 문제가 없을 것으로 판단하여 자료를 공동으로 이용하기로 한 것이다. 연구소의 사전 자료는 1994년 5월 중순경에 제공받았다. 입력 형식을 달리했기 때문에 우리 원의 검색 프로그램으로 검색할 수 있도록 형식을 손질하였다.

연구원에서는 기존 사전의 문제점 중의 하나로 실제 언어 자료를 반영하지 않았다는 점을 꼽고 이를 개선하기 위해 문헌 자료를 입력하여 사전 편찬에 활용하려는 구상을 일찍부터 가져 왔다. 국내외에서 언어 자료를 기반으로 사전을 편찬하려는 논의가 있고,

그에 따라 편찬된 사전이 있었다는 점도 이러한 방향 설정에 도움이 되었다. 전문어는 몰라도 일반어에 대해서는 가급적 빠짐없이 예문을 제시하자는 방침도 일찍 정해 졌다.

이 목적을 달성하기 위해 다량의 언어 자료를 확보하려는 노력을 계속해 왔다. 그런데 다량의 언어 자료를 확보하는 일에는 검토할 사항이 여러 가지가 있다.

우선 무엇을 자료로 확보할 것인가의 문제가 있다. 지금까지 나온 모든 출판물을 입력할 수 있다면 별 문제는 없지만 그럴 수 없기 때문에 어떤 자료를 확보할 것인가가 중요한 문제가 되는 것이다. 이는 실제 언어의 모습을 토대로 현대어 사전을 만들고자 하는 경우에 특히 중요한 문제가 될 것으로 생각하는데 이에 대한 논의는 국내외에서 찾아볼 수 있다. 필자 역시 이 문제에 대해서 오랫동안 고민을 해왔다. 그런데 이 문제는 사전의 성격과 밀접한 관계를 맺는 면이 있다. 연구원에서 목표로 한 사전은 종합 국어 대사전이기 때문에 최근의 문헌만을 대상으로 자료를 확보해서는 불충분하다. 현재는 거의 쓰이지 않는 말에 대한 자료도 있어야 한다. 언어 자료는 예문 제시에 제일 많이 사용될 것으로 생각되는데 예문을 어떤 성격의 것으로 수록할 것인가에 따라 입력 자료의 대상이 바뀌어야 하기도 한다. 이에 대해서는 충분한 검토가 있어야 한다고 생각했으나 여건이 허락되지 않아 자료의 취사에 필자의 주관이 작용한 점이 없지 않다. 자료량이 많아지면서 상대적으로 이 문제는 덜 심각해지는 점도 있다.

동일한 문헌이 여러 번 출판되었을 때 어느 것을 택할 것인가도 고려 사항이다. 주로 현대국어 시기의 문헌이 문제인데, 정본이거나 그에 준하는 것이 있으면 이를 이용하면 되지만 이에 대한 확고한 의식이 없이 그때그때 표기를 고치면서 출판이 이루어지기 때문에 문제가 되지 않을 수 없다. 이 과정에서 극히 드물게 나타난 단어를 확보하지 못할 수도 있다. 활자화된 모든 판본을 구해서 비교하는 게 가장 정확하겠지만 현실적으로 불가능한 작업이다. 처음 활자화된 것이 가장 정확하다고 보고 그 문헌을 입력하는 것이 문제점을 최소화할 수 있는 방안이라고 생각되기는 한다. 그런데 최초로 활자화된 시기가 최근이라면 별로 문제될 것은 없지만 '20, '30년대까지 거슬러 올라가면 문제는 달라진다. 우선은 문헌을 구하기가 쉽지 않다는 점이 있다. 설령 구한다 하더라도 더욱 심각한 문제가 있다. 표기가 현재의 어문 규범과 다르다는 점이다. 예문에 그대로 인용한다면 문제는 없겠지만 규범 사전의 성격을 가진 새 사전에서 고어가 아닌 현대어 표제어에 어문 규범과 다른 표기를 가진 예문을 제시할 수 있는가가 문제가 아닐 수 없다. 92년도에 분과 회의를 할 때 예문의 표기 문제가 안건으로 상정된 적이 있다. 이 때 그대로 두는 것이 좋지 않겠느냐는 견해가 유력하기는 했다. 그후로도 몇 차례 논의가 있었지만 최종 입장이 정리되지는 못했다. 우선 최초 발표 작품으로 입력하는 것을 원칙으로 하였으며, 집필할 때는 예문을 그대로 인용해 두도록 하였다. 교열 과정에서 예문

만 따로 확인하여 규범에 맞게 고치는 작업을 해도 큰 무리는 없다고 판단했다. 오히려 집필자마다 달리 손질할 수 있다는 우려를 씻을 수 있는 장점이 있다.

다른 문제는 입력 형식을 통일시키는 점이다. 입력 형식이 통일되어 있지 않으면 검색이 제대로 되지 않거나 검색하고도 그 결과를 마음놓고 이용할 수 없는 경우가 생길 수 있다. 자료량이 많으면 원문헌에서 확인한다는 것도 적지 않은 부담이 따르는 일이다. 이 문제를 해결하기 위하여 입력 과정에서 발생할 가능성이 있는 문제점들을 짚어서 입력 지침을 정하여 두었다.

1992년과 1993년은 자료 구축의 방향을 모색한 시기라고 할 수 있다. 원내의 입력 보조원과 원외의 조사 보조원이 다양한 성격의 문헌을 입력하였다. 입력 보조원은 사전을 입력하지 않은 1992년 9월에서 1993년 3월 사이에 잡지, 수필, 시, 교과서, 희곡 등을 입력하였다. 최근의 것을 입력하였다. 원외의 조사 보조원은 신문과 '20년대의 잡지를 입력하였다. 1992년에는 각 연구원이 분과를 맡아 사전 편찬 작업을 진행했는데 다른 분과에서도 문헌을 입력하는 일을 하기도 했다.

1992년에는 문헌 목록을 선정하는 작업이 용역으로 진행되었다. 이 일은 총괄 분과에서 진행하였다. 이 용역에서는 장단편 소설, 수필을 대상으로 개화기부터 1990년까지 발표된 작품 목록을 조사하였으며, 전문가의 의견을 취합하여 어휘 수집에 도움이 될 전형적이고 표본적인 작품을 선정하였다.

1994년에는 자료 입력 용역 예산이 확보되어 사전 편찬에 필요한 문헌을 입력하는 일을 전국의 국문과 대학 교수 21인에게 의뢰하였다. 국어사 문헌, 고전문학 문헌, 현대 문헌을 입력 대상으로 하였다. 현대 문헌은 1992년의 용역 결과를 기반으로 하였다. 일치도가 3 이상인 문헌을 입력 의뢰하였으며, 여기에 시와 신문, 잡지를 추가하였다. 용역은 어절 대비 단가를 기준으로 2,000만 어절을 입력하는 것을 목표로 하였으나 문헌에 따라서는 작업이 까다로운 것도 있어 입력 단가를 조정하였기 때문에 실제 입력 어절은 다소 줄었다.

연구원에서는 입력 보조원의 경우와 달리 교정 인력은 많이 두지 않았다. 1992년 4월에 표제어 목록 입력 결과를 교정할 인원을 2명 채용한 이후 카드 작업이 완료되어 그 작업을 하던 인력 일부가 교정 인력으로 투입된 1994년 10월까지 그 수를 그대로 유지했다. 이에 따라 교정 인력이 충분하지 않기 때문에 교정까지 포함하여 용역을 의뢰하였다. 그동안 다양한 문헌을 입력하면서 작성한 입력 지침서를 함께 배포하여 입력이 통일될 수 있도록 하였다. 현대 문헌 용역자에게는 최초로 활자화된 문헌을 찾아서 입력해 주되 이것이 어려울 경우는 신뢰할 만하다고 판단되는 것을 선택하여 입력해 줄 것을 부탁하였다.

연구원에서 대학 교수에게 의뢰한 것은 학생들을 이용하여 입력, 교정을 진행할 수 있다고 판단했기 때문이다. 그런데 실제로 작업이 진행되면서 한자가 많이 섞였거나, 띄어쓰기가 현재와는 많이 차이나는 문헌을 학생들이 입력을 제대로 하지 못해 심한 경우 용역자가 직접 입력하기도 하였다고 들었다. 문헌을 찾는 일도 예측한 것보다 훨씬 어려웠다. 최근에 간행된 것을 선택한 경우도 적지 않았고, 끝내 찾지 못한 경우도 있었다. 연구원에서 처음에 배포한 문헌 목록만으로는 용역량에 못 미칠 경우 추가로 선정한 문헌을 입력하기도 했다. 입력 지침서를 배포했음에도 지침을 제대로 이해하지 못해 중간에 들어온 결과물을 검토하여 용역자에게 지침과 달리 입력이 된 부분을 알려준 일도 종종 있었다.

내부에서의 입력도 계속되었다. 1994년에는 입력 보조원을 2명 더 채용하여 7명으로 늘어났다. 사전 전체 입력이 끝난 후 입력 보조원들은 표제어 작업 등에서 발생하는 입력 작업이 있으면 그 일을 우선으로 했지만 일이 없는 경우에는 문헌을 입력하였다. 94년도에는 주로 소설과 교과서, 북한 문헌을 입력하였다. 북한 문헌의 경우는 외부에 입력을 의뢰하기 어려웠기 때문에 내부에서 직접 입력하였다. 연구원 내에서 사전 편찬 외의 사업을 추진하면서 문헌을 입력해야 하는 경우들이 종종 있었다. 아직 사전 편찬실에서 확보한 자료가 충분하지 않았기 때문이다. 이 때 만들어진 입력 파일도 사전 편찬실에서 그대로 이용했다.

한편으로는 외부의 입력 자료를 구하기도 하였다. 최근에 각종 자료를 컴퓨터에 입력하는 일이 빈번하다. 컴퓨터 조판이 활발해서 전산화된 자료가 적지 않으며, 컴퓨터 통신이 발전하면서 각종 자료가 입력된 상태로 통신망에 소개되기도 한다. 초기에는 이들 자료도 적극 활용할 생각이었으나 지금은 그리 높은 비중을 두지는 않는다. 우선 연구원에서 필요로 하는 자료여야 하는데, 입맛에 맞는 자료를 구하기가 쉽지 않다는 문제가 있다. 또한 내부의 입력 형식에 맞게 다시 손질해야 하는 번거로움이 있다. 그 부담도 결코 무시하기는 어렵다. 컴퓨터 조판으로 된 자료는 때로는 변환조차 되지 않아 구해 놓고도 활용하지 못하는 경우도 있었다. 1994년 초에 신문사와 출판사로부터 일정량의 입력 자료를 구한 정도에 그쳤다.

이렇게 한 결과 현재 연구원에서 확보한 자료량은 약 2,200만 어절 분량이다. 1995년에는 예산이 축소되기는 했지만 자료 입력 용역 예산이 확보되었기 때문에 자료량은 더욱 늘어날 것이다. 연구원에서는 최종적으로는 6,000만 어절을 확보할 계획이다.

자료를 많이 확보하기는 했지만 이를 효율적으로 활용하는 것도 큰 문제가 아닐 수 없다. 우선은 검색 속도의 문제가 있다. 자료량이 늘어날수록 검색에 소요되는 시간이 늘어나기 때문이다. 한 단어가 쓰인 예를 검색하는 데 최신 기종인 486DX로도 1시간이

소요될 정도이다. 자료량이 편중된다는 문제도 있다. 자주 쓰이는 동사의 경우는 수천 개도 넘는 용례가 나오는가 하면 전혀 용례를 찾을 수 없는 단어도 있다. 용례가 지나치게 많아도 집필에는 큰 부담이 된다. 또다른 문제는 검색에서 원하지 않는 예임에도 함께 검색되는 것이 많다는 점이다. 현재 연구원에서 이용하는 검색 프로그램은 단순히 문자열 검색밖에 되지 않는다. 이에 따라 예컨대 '날다'를 검색하려면 '날다,/나는' 등을 찾기 위해서 '날', '나'로 시작되는 어절을 모두 검색하는 방법을 쓰고 있다. 이렇게 되면 체언과 조사의 결합형인 '나는'도 함께 검색되어 나온다.

이러한 문제점은 개인용 컴퓨터로 해결하는 데는 한계가 있다. 개인용 컴퓨터에 자료를 보관하고 있기 때문에 한 번에 하나의 검색만 가능하여 성능을 아무리 개선해도 한계를 가질 수밖에 없다. 집필자들에게 표제어가 사용된 용례를 검색하여 내보낼 초기에는 표제어별로 검색을 하였었다. 그러다 보니 항상 시간이 부족하였고, 집필자가 늘어나면서는 도저히 작업을 쫓아갈 수 없다는 문제가 발생했다. 지금은 모든 어절을 검색을 하여 가나다순으로 정리하여 기계적으로 잘라 주는 방법을 쓰고 있다. 예를 들어 집필자에게 의뢰한 집필 표제어가 '농부'부터 '눈덩이' 사이의 표제어라면 '농부'부터 '눈덩이' 사이의 모든 어절을 용례로 준다. 이 방법이 문제를 다 해결한 것은 아니다. A4 용지로 몇백 페이지의 분량이 제공되곤 하는 일이 흔한데 그 중에는 검색 오류, 집필 의뢰가 되지 않은 단어 등 잘못된 결과를 포함하는 일이 흔하다.

중형컴퓨터를 활용하게 되면 상황은 다소 호전될 것으로 생각한다. 그렇지만 자료량이 편중된다는 문제는 여전히 남는다. 이 문제는 자료의 검색을 두 단계로 나누면서 해결할 생각이다. 입력 자료 중에서 일부를 선정하여 여기서 일차로 검색을 하고, 용례를 전혀 찾지 못했거나 보충이 필요한 경우 나머지 문헌을 검색한다는 구상이다. 일차 검색에 이용될 문헌은 형태소 분석, 엄밀하게는 어간과 어미로 나누는 분석도 할 예정이다. 이렇게 하면 동음이의어의 문제까지도 해결하고, 나아가 분석된 자료를 다른 사업에도 활용할 수 있을 것으로 생각한다.

## 2. 전산 시설 확보

자료의 확보와 아울러 필요한 것은 이 자료를 이용할 수 있도록 도와 주는 하드웨어나 소프트웨어를 구비하는 일이다.

사전 편찬 작업은 컴퓨터가 한 대도 없는 상태에서 출발하였다. 그후 지속적으로 예산을 확보하여 현재는 개인용 컴퓨터 50여대가 확보되었다. 1기가바이트의 하드 용량을 가진 컴퓨터도 5대나 확보하였다. 최종적으로는 사전 편찬에 종사하는 사람들에게 1인

1대나 그 이상으로 보급될 것으로 본다. 전산 처리를 위한 노력을 많이 한 결과 대부분의 작업이 컴퓨터를 이용하여야 하기 때문이다.

개인용 컴퓨터의 발달이 급속도로 이루어지고 있기는 하지만 아직 사전 편찬 작업을 개인용 컴퓨터에 의존하는 데는 한계가 있다. 내부에서 이 문제가 제기되면서 연구원에서는 중형컴퓨터의 도입을 추진하기로 하였으며, 1993년에 중형컴퓨터 도입 예산을 확보하였다. 정부 방침에 따라 국산주전산기인 타이컴을 구입하였다. 1993년 말에 삼성전자의 SSM7000 모델로 결정되었는데 연구원 이사 문제와 겹쳐 1994년 7월에야 설치되었다. 중형컴퓨터와 개인용 컴퓨터는 LAN으로 연결하였다.

하드웨어의 확보와 더불어 소프트웨어를 확보하려는 노력도 계속되었다.

소프트웨어의 개발은 두 차례에 걸쳐 이루어졌다. 한 번은 아래한글에서 검색을 할 수 있는 프로그램을 개발하는 일이었고, 한 번은 사전 편찬 전 과정을 지원하는 프로그램의 개발이었다.

앞에서 말했듯이 기존 사전 정보를 검색하기 위해 사전을 아래한글 2.0을 이용하여 입력하고 이를 검색할 수 있는 프로그램을 개발하기로 했다. 프로그램의 개발은 1992년에 아래한글의 개발사인 한글과컴퓨터사에 용역으로 의뢰하였다.

프로그램 개발은 한글과컴퓨터사의 개발 담당자와 필자가 협의하면서 진행했다. 필자가 먼저 어떤 종류의 프로그램이 필요한가를 정의했고, 이를 기술적으로 검토하면서 프로그램의 형태를 결정하였다. 프로그램 개발의 일차 목적은 사전 검색 프로그램의 개발이었다. 그렇지만 이와 아울러 문자열 검색 프로그램도 함께 개발하였다. 그 당시 학계에서 사용되는 문자열 검색 프로그램이 두 개 정도 있었으나 텍스트 파일만 검색이 가능하였기 때문에 아래한글로 작성된 문서를 검색할 수 없다는 문제점이 있었다. 이들 프로그램을 사용하면서 필자가 아쉬워했던 점이 최대한 프로그램에 반영되도록 했다. 이렇게 해서 개발된 프로그램이 HGREP.EXE와 HDB.EXE이다. 전자는 문자열을 검색하는 프로그램이고, 후자는 사전을 검색하는 프로그램이다.

이들 프로그램은 1994년 5월에 한 차례 수정되었다. 아래한글이 2.1로 업그레이드되면서 2.1로 작성된 문서는 검색하지 못한다는 문제가 있었기 때문이다. 개발 담당자에게 수정을 의뢰하면서 그동안 사용 중에 겪었던 불편 사항을 추가로 고쳤다.

지금 집필을 의뢰할 때 함께 나가는 용례 자료는 HGREP.EXE로 검색한 것이다. 아래는 HGREP.EXE와 HDB.EXE를 이용한 검색 결과이다.

검색내용 : "(@[ㄷ##](는¦고))" res009.hg /어절끝
☞ 어절의 마지막 두 글자에서 첫 번째는 초성이 'ㄷ'이고, 두 번째는 '는' 또는 '고' 출력형식 "표제어 : 20,출전 : 15,예문 : 40" /예문표제 : 10 /어절강조

| 표 제 | 출 전 | 예 문 |
|---|---|---|
| 도는 | 〈MAA01440.06〉 | 북한의 태{도는} 눈에 띄게 변한 |
| 데는 | 〈MAA01440.07〉 | 설치하는 {데는} 얼마 걸리지 않는다"고 |
| 도는 | 〈MAA01440.06〉 | 정가에 나{도는} 말로는 페로의 퇴진으로 |
| 다는 | 〈MAA01440.06〉 | 현상유지보{다는} 물갈이 쪽으로 쏠릴 |
| 다는 | 〈MAA01440.06〉 | 뽑았{다는} 것이 또한 언론들의 |
| 드는 | 〈MAA01440.06〉 | 뛰어{드는} 순간 온갖 추잡한 트릭이 |
| 다는 | 〈MAA01440.06〉 | 증도탈락했{다는} 것이다. |
| 돌고 | 〈MAA01440.07〉 | 나{돌고} 있어 민주당 진영을 |
| 들고 | 〈MAA01440.08〉 | 만{들고} 있다. |
| 다는 | 〈MAA01440.08〉 | 것이었{다는} 점, 그리고 무엇보다도 |
| 드는 | 〈MAA01440.08〉 | 투표에 {드는} 수고를 덜 수 있게 됐다. |
| 다는 | 〈MAA01440.07〉 | 해야 한{다는} 태도를 바탕으로 한 |
| 되는 | 〈MAA01440.08〉 | 대두{되는} 이때 비둘기들은 |

** 13 개의 항목을 찾았습니다.

출력형식 "표제 : 20,어원"
가려뽑기 (@길이([표제])==@길이( [어원])&&!(@정규식([어원],
"[#0000-#3fff]+",""))&&@길이([표제])==4) res004.hd
☞ 표제의 길이와 어원의 길이가 모두 4이면서, 어원에 영문 등이 없는 표제어

| | |
|---|---|
| 가가대소 | 呵呵大笑 |
| 가가호호 | 家家戶戶 |
| 가감부득 | 加減不得 |
| 가감승제 | 加減乘除 |
| 가감역관 | 假監役官 |
| 가공철도 | 架空鐵道 |
| 가교봉도 | 駕轎奉導 |
| 가구경행 | 街衢經行 |

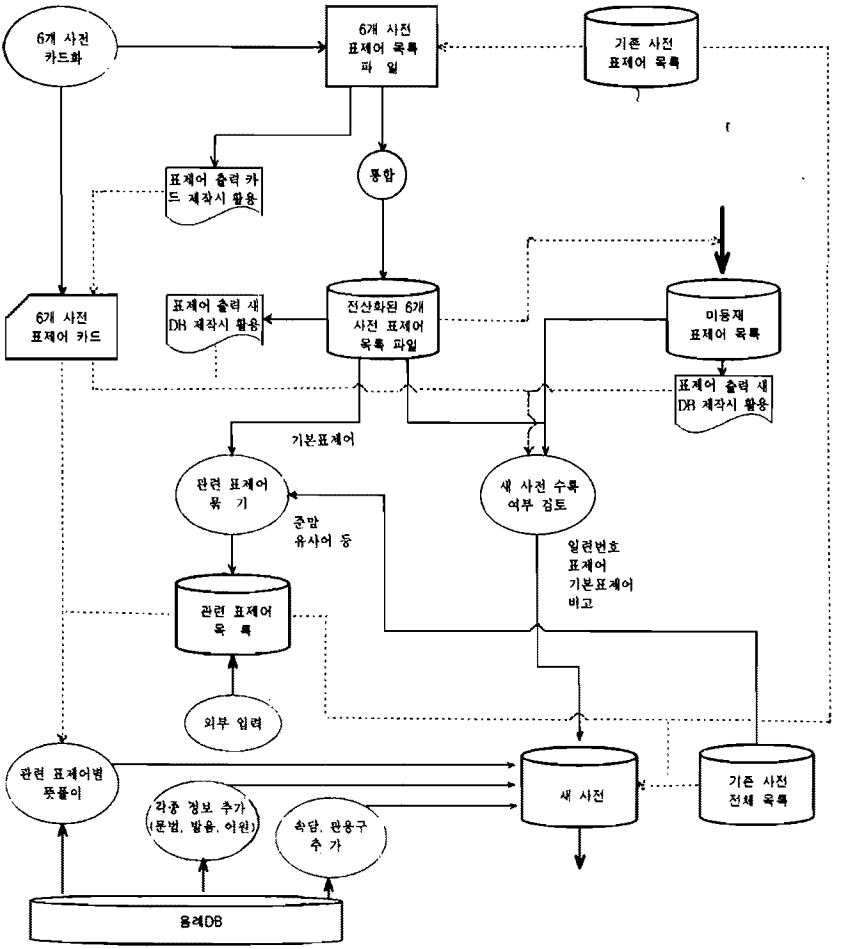
가구적간 家口摘奸
** 9 개의 항목을 찾았습니다.

사전 편찬 전과정을 지원하는 프로그램. 보통 말하는 사전 편찬 지원 시스템 개발 용역은 1993년 9월부터 1994년 12월까지 진행되었다. 이 시스템은 중형컴퓨터에서 작동되는 프로그램이다. 중형컴퓨터의 도입을 추진하면서 함께 추진한 것이다. 처음 작업을 시작하였을 때는 중형컴퓨터를 도입하면서 시스템도 함께 장만할 생각이었다.

그런데 예산이 부족하여 하드웨어와 함께 원하는 소프트웨어를 갖추기는 어렵고 예산이 확보되는 대로 순차적으로 개발을 할 수밖에 없다고 생각하고 있었다. 검색 프로그램을 먼저 개발한 이유 중의 하나도 예산 때문이었다. 그런데 1993년 6월에 중소기업 지원 자금으로 4억 7천여만 원이 시스템 개발에 배당되었다. 이에 따라 중형컴퓨터 도입과는 별도로 시스템 개발에 착수하였다. 이 일은 한국정보공학주식회사에서 용역으로 맡아 진행하였다.

중형컴퓨터에서는 개인용 컴퓨터와는 달리 KS 5601에서 정의된 문자만을 사용할 수 있다. 이 코드 체계는 그동안 뜨거운 논쟁을 불러 일으켰던 완성형 한글을 포함한 체계이다. 국어학계에서는 문자가 부족하기 때문에 이 체계를 거의 사용하지 않는다. 사전 편찬에 중형컴퓨터를 사용하기 위해서는 문자의 사용을 결코 소홀히 할 수 없다는 것이 담당자인 필자의 의견이었다. 오랫동안 검토하면서 주위의 전문가의 의견도 많이 들었었는데, 연구원 내부 코드 체계로 4바이트 코드 체계를 쓸 수 있도록 프로그램을 개발하는 것으로 결정하였다. 이에 따라 옛자모를 포함하여 240자에 이르는 자모로 만들어지는 완성자와 한자 55,000여 자를 컴퓨터상에서 사용할 수 있게 되었다. 사전 편찬에 필요한 문자의 문제는 거의 해소되었다고 생각한다.

아래에 제시한 도표는 시스템 구성에 관한 협의를 하던 초기에 전산화 계획을 도표화한 것이다.




여기서 볼 수 있듯이 사전 편찬의 전 과정을 전산화하는 것을 목표로 하였다. 문헌의 입력, 새 사전 표제어의 확보부터 용례의 검색, 뜻풀이 등의 작업이 모두 하나의 시스템 속에서 이루어지도록 하였다. 체계적으로 사전 편찬 관련 자료를 관리하면서 편찬을 진행할 수 있는 기반을 마련한 셈이다. 사전 편찬에 소요되는 시간도 상당히 단축하는 효과를 가진다.

시스템의 도입이 빨랐다면 사전 편찬의 진행에 더 많은 도움이 되었을 것이라는 아쉬움이 있는 한편, 사전 편찬 경험이 없어 작업을 확실히 정의하지 못해 시스템 개발 방향이 중간에 바뀌는 어려움도 있었다. 개발한 부분이 불필요하게 된 경우도 있다. 관련 표제어 작업이 불필요하게 된 경우도 들 수 있다. 시스템을 설계할 무렵에는 관련 표제어별로 뜻풀이를 집필자에게 의뢰한다는 계획을 세워둔 상태였다. 그런데 이 방침을 고수하지 않기로 했기 때문에 이 부분은 당장은 사용하지는 않을 부분으로 남는다.

이 시스템은 일차적으로 사전 편찬을 위해 마련된 것이다. 그렇지만 여기에는 연구에 필요한 각종 프로그램이 거의 포함되어 있다고 믿는다. 최초 설계 당시부터 연구원의 다른 업무에도 사용할 목적으로 이미 진행 중이거나 앞으로 진행할 것으로 예상되는 연구원 사업에서 어떤 프로그램이 필요할 것인가를 함께 고려했기 때문이다. 실제로는 사전 편찬에 필요한 프로그램과 다른 사업에서 사용할 프로그램이 크게 차이 나는 것은 없고 사용에 있어서 약간의 고려만 더하면 되었다. 따라서 사전 편찬 작업 외에 연구 업무를 위해서도 유용하게 이용될 것으로 기대된다.