□ 기술해설 □

국어정보베이스의 현재와 미래

한국과학기술원 최기선*
시스템공학연구소 박동인*

1. 서 론

국어정보베이스는 그 방대함과 다양성에 비추어, 국가적으로 대표적인 것이 되어야 함은 의논의 여지가 없다. 또, 앞으로 변화하는 언어와 대처해야 할 상황의 다양성에 비추어 지속적 국가적 지원이 필요함도 당연하다 하겠다. 언어의 문제, 특히 한국어에 대한 집중적 연구는 다음의 이유로 인하여 중요하다고 하겠다. 첫째, 언어의 보편화와 세계화 상황에서 생존하여야 한다. 우리 한국어 사용자는 약8천만에 이르며, 세계 제8위를 자랑한다. 그러나, 현재의 정보 네트워크의 발달을 보면, 몇 개의 중요 언어로 모든 정보가 만들어지는 현상을 볼 수 있다. 이와 같은 현상이 향후 50년이 지속되면 그 중요한 몇 개의 언어 이외의 언어는 그 가치를 잃어 버린다는 위험성을 안고 있다. 이를 방지하기 위하여, 각국은 자국 언어의 육성에 힘을 쓰고 있다. 이에 따라 국내에서도 다음과 같은 노력이 필요하다.

첫째, 한국어 연구에 필요한 모든 기반 조성을 전세계적으로 하여야 할 필요성이 있다. 둘째, 언어는 우리 삶의 공기와 같은 존재이다. 의사소통의 기본인 개념은 언어로서 표현된다. 이에 한국어의 표준적 사용과 현상의 분석에 대한 연구, 개발을 통해 컴퓨터에서의 사용이 활성화 될 필요가 있다. 셋째, 문화의 보존에 해당한다. 문화는 언어에서 비롯된다. 한국어이되 한국어가 아닌 언어화 현상을 막아야 할 것이다. 넷째, 우리 언어의 정당성 확보에 의하여, 자생적 학문 육성이 시급하다. 외국어로 된 어휘와 개념으로는 한국의 자생적 학문이 존재하지 않을 것이다.

이에 우리 한국어의 언어 현상이 국어학, 국어공학 분야에서 실증적으로 연구 개발이 이루어지기 위한 초석으로서 국어정보베이스의 존재 이유가 있다.

국어정보베이스란 국어학 및 국어공학 연구의 기반 자료 위주의 실증적 연구를 뒷받침하기 위한 방대한 양의 자료 데이타베이스이다. 국어학의 제분류로서 형태론, 통사론, 의미론, 화용론, 음운론, 음성론, 글자 형태의 기본 데이타베이스뿐만 아니라, 기본 어휘자료의 각종 통계치, 각 연구분야에 맞도록 필요 정보를 더한 가공 자료 데이타베이스, 그리고 각종 조사 연구 도구 및 관리 시스템 등을 국어정보베이스라고 총칭한다. 즉, 국어정보베이스란 모든 원시 자료 데이타베이스(혹은 코퍼스) 및 가공 자료 데이타베이스 그리고 이를 개발 및 관리하기 위한 도구, 시스템 등을 일컫는다. 가공 자료 데이타베이스는 품사가 부착된 가공 자료 데이타베이스(형태통사 태그 부착 코퍼스), 구문구조가 매겨진 가공 자료 데이타베이스(트리뱅크), 음성 레이블이 붙여진 음성 데이타베이스 등이다.

이와 같은 국어정보베이스를 개인이 개발하기에는 그 규모와 소요되는 비용이 방대하고 지속적인 개발을 해야 하기 때문에 개인이 하기에는 어려운 것이다. 또, 중복 개발하는 것도 비용 및 시간적인 면에서 그 효용성이 없다고 본다. 따라서, 국가적으로 잘 개발된 표준 국어정보베이스가 구성된다면, 국어학 및 국어공학의 모든 연구의 실증적 결과가 뚜렷해질 것임

*종신회원

은 자명한 사실이다.

이와 같은 필요성에 의하여, 1994년부터 지난 3년간 문화체육부와 과학기술처의 협력 하에 STEP 2000(Software Technology Enhancement Program)이라는 제목으로 "우리말 정보처리"에 관한 "국어기반 자료 개발"계획에 의하여 "통합국어정보베이스" 등이 이루어졌다.

제2절에서는 이 국어정보베이스의 현황 및 문제점을 살펴보고, 제3절에서는 향후의 방향 및 문제점 타개에 대하여 기술한다. 마지막으로 제4절에서 결론을 내린다.

2. 국어정보베이스의 현황과 문제점

문화체육부의 지원 하에, 1998년까지 원시 텍스트 코퍼스가 1억 어절, 품사부착 코퍼스가 5천만 어절, 과학기술처 지원 하에, 구문구조 코퍼스(혹은 트리 뱅크)가 1997년까지 2만 문장 규모 이상에 달한다. 음성 데이타베이스로서 음성 균형 코퍼스, 낭독 코퍼스, 그리고, 필기체 데이타베이스가 KS 완성형 표준에 따른 한글 글자에 대한 지역별, 연령별, 필기구별, 종이 질 별로 구축되었다.

텍스트 관련 데이타베이스의 관리 및 개발의 표준화를 이루기 위하여, 형태-통사 태그(약하여 "품사"라 칭함)에 대한 표준안을 발표하여 이에 따른 품사부착 코퍼스를 개발하고 있다. 음성 및 필기체 데이타베이스도 이와 같은 형식을 공개하였다. 이와 같은 대량의 데이타베이스 및 자체적 표준안에 따른 개발에 따라, 여러 개발 팀간의 의견 통일 및 결과물의 다른 연구 개발자 혹은 기업에서 즉시 활용이 가능하도록 하였다[1, 2].

이미 제2회 한국어정보처리 표준화 심포지움을 통해, 배포된 연구 실험용 "대한민국 국어정보베이스"시험판(CD-ROM)이 사용되고 있다. 단, 이 CD-ROM은 개발된 대용량 데이타베이스의 일부에 불과하며, 아직 그 오류가 수정되어 있지 않은 상태이다. 그러나, 이 과정에서 볼 때, 다음과 같은 자료의 문제점을 도출할 수는 있다. 첫째, 언어 데이타 및 정보베이스가 우리 언어의 전반적인 모습을 드러낼 수 있을 정도로 대표성이 있는가? 대표성을 갖기 위해서, 어느 정도의 양과 균형성을 가져야 할까? 둘째. 결과물인 데이타베이스가 그 품질을 보증하는가? 어느 정도의 오류를 허용할 수 있을까? 세째. 결과물의 활용성으로서 개발 즉시 다른 목적으로 활용이 가능한가? 이 문제점의 해결 방안에 대하여, 다음 절에서 논의하고자 한다.

3. 국어정보베이스의 미래

국어정보베이스는 앞으로 어떻게 만들어지고 어떻게 발전하여야 하는가? 이에 대한 문제점이 앞 절에서 지적되었다. 이 절에서는 이에 대한 해결 방안을 모색하고자 한다.

첫째, 개발 과정의 과학화를 들 수 있다. 방대한 양의 코퍼스 작성은 흡사 방대한 운영체계 시스템을 만드는 일보다 더 어렵다. 국어정보베이스의 개발에도 소프트웨어 공학과 지식공학의 방법론이 도입되어야 한다. 즉, 단계별 과정을 본다면, 사용자 요구분석 단계로서 코퍼스의 이용 대상의 수집, 가능한 다목적으로 쓰일 수 있도록 균형화에 대한 조사 연구를 한다. 다음은 설계 단계로서, 코퍼스의 저장 형식, 분류 체계 확립을 들 수 있다. 이제 실제 일을 할 단계이다. 보통 소프트웨어 관리에서는 프로그램 단계라고 한다. 여기서는 데이타 구축에 관한 것이므로 데이타 구축 단계라고 하여야 할 것이다. 이 단계에서는 정해진 규격에 맞도록 컴퓨터에 의한 분석과 그 분석의 사람에 의한 수정과 그 피드백에 의한 컴퓨터의 재분석 등의 작업이다. 이 때, 먼저 프로토타입 데이타베이스를 만들어 그 효용성을 검증한다. 효용성이란 최초의 사용자 요구분석에 맞는가에 대한 검증에 속한다. 이 검증을 통과하면, 대량의 데이타 구축에 들어간다. 구축에 따라 몇개의 시점에서 반복 검증을 실행하며 최초의 설계에 맞는가에 대한 검증을 다시 실행한다. 모든 과정이 끝나면, 서비스를 위한 실제 활용처에 설치하는 설치 단계 및 활용 단계를 거친다. 이 과정 전반에 걸쳐 잊어서는 안될 것은 문서화 단계이다. 이 문서화 단계는 모든 단계의 중간 및 단계마다 이루어져야 한다. 예를 들어, 사용자 요구분석서, 설계 사양서, 구축

추진계획서, 검증 계획서 및 검증 결과서, 오류 검증서, 설치 사양서, 활용 방법서, 활용 후 유지/보수계획서 등의 개발 안내 지침서, 활용 안내서 등을 이룬다.

이 과정의 복잡도가 높다는 것은 대상이 무한대의 언어라는 점이다. 무한의 언어를 제한된 양의 텍스트 수집 데이타베이스 만으로 그 언어의 구조를 추정하려 한다는 문제에서 그 난이성이 비롯 한다. 또 한가지의 어려운 점은 그 구축이 사람에 의하여 이루어지며, 그 사람이 여러 곳에 흩어져 있다는 점, 그리고 적시에 여러 곳에서 만들어진 데이타베이스를 과학적으로 완벽하게 검증하기가 매우 어렵다는 데에 있다.

둘째, 품질 유지의 과학화를 들 수 있다. 품질은 두 가지 측면을 가진다. 첫째 측면은 앞서 언급한 바와 같이 과연 "대표성"을 갖는가에 대한 문제이다. 대상 언어의 모든 것을 보여 주는 코퍼스는 과연 어떤 것이어야 하나? 대표성 유지를 위하여 어떻게 개발되어야 하나? 코퍼스의 대표성 있는 핵은 과연 무엇인가에 대한 해답을 찾는 것이 국어정보베이스에서 해결하여야 할 시급한 과제중의 하나이다. 둘째 측면은 오류의 사전 방지에 관한 방법론이다.

셋째, 사람과 기계의 조화, 즉 작업환경의 과학화를 들 수 있다. 개발 관리 도구의 방향은 완전 자동화를 지향한다. 그러나, 완전 자동화는 국어정보베이스가 완성이 되고, 그것을 기반으로 많은 연구 개발이 이루어져야 가능하다. 따라서, 현 단계에서는 자동화 도구와 사람의 가능한 대화와 학습에 의하여 가능하다. 사람과 기계가 어떻게 활용되어야 할 것인가에 대한 심층적 연구가 이루어져야 한다.

위에서 살펴본 바는 국어정보베이스의 개발을 위한 추진 방법론에 해당한다. 그 내용면에서 본다면 언어현상의 전반적 정리, 다국어와의 대조 분석, 분류 어휘 등의 해야 할 작업과 분야가 많다는 것은 주지의 사실이다. 또, 활용성의 입증을 위하여 응용 연구 및 개발, 시스템화에 의하여 검증되어야 할 것이다.

4. 결 론

국어정보베이스의 현재는 이미 존재하는 데이타의 수집과 방대한 양을 집성하는 데에 의의를 두었다. 내용면에서는 단어 중심의 정보만을 기록하였다. 따라서, 품사 위주의 정보베이스만이 존재한다. 또, 음성 데이타베이스나 필기체 데이타베이스도 그러하다.

국어정보베이스의 미래는 활용에 있다고 본다. 활용은 품질 보증과 모듈화, 규격화에 있다. 그 목적을 달성하기 위하여, 개발과정의 과학화, 품질보증의 과학화, 작업환경의 과학화에 대하여 논의하였다.

국어정보베이스는 우리의 것이다. 앞으로 적절한 절차를 거쳐 모두 공개될 것이다. 이것이 우리 국력의 한 척도로서 보일 수 있기를 바란다. 우리의 언어 데이타와 정보베이스를 바탕으로, 외국의 새로운 이론적 연구 성과가 검증되기를 원하며, 문화적인 생존과 향상의 근간을 이루는 한 개의 밀알로서의 역할을 하게 되기를 바란다.

참고문헌

[1] "제1회 한국어 정보처리 표준화 심포지움", 1996. 7. 한국어정보처리연구회.

[2] "제2회 한국어 정보처리 표준화 심포지움", 1997. 6. 한국어정보처리연구회.

최 기 선

1978 서울대학교 수학과 학사
1980 한국과학기술원 전산학과 석사
1985～1986 한국외국어대학교 전산학과 조교수
1986 한국과학기술원 전산학과 박사
1986～1987 일본 NEC C&C 정보연구소 초빙연구원
1988～1992 한국과학기술원 전산학과 조교수
1992～현재 한국과학기술원 전산학과 부교수, 인공지능연구센터 언어공학연구실장

박 동 인

1979 서강대학교 전자공학과 학사
1979~현재 시스템공학연구소 자연어정보처리연구부 부장
1994~현재 공업진흥청 산업표준 심의회 위원
1995~현재 국어정보학회 이사, 문화체육부 국어심의회(국어정보화분과) 위원

● '97 추계 정보통신 단기강좌 ●

- 일 자 : 1997년 10월 13(월)~14일(화)
- 장 소 : 한국과학기술회관
- 주 제 : 'Mobile Computing'
- 주 최 : 정보통신연구회
- 문 의 처 : 서강대학교 전자계산학과 최명환 교수
  Tel. 02-705-8495