조선어코퍼스구축에서 나서는 몇가지 문제

지 동 은

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학기술정보사업을 잘하면 큰 밑천을 들이지 않고도 나라의 과학기술을 빨리 발전시킬수 있습니다.》(《**김정일**선집》 증보판 제11권 145페지)

오늘날 정보는 재료, 에네르기와 함께 현대과학기술발전의 3대기둥으로서 인류의 생존과 발전을 결정하는 귀중한 자원으로 되고있으며 정보기술은 과학기술발전을 주도하는데서 관건적역할을 하고있다. 사회의 전반부분이 정보화되는 21세기에 한 나라의 종합적인 현대화정도는 정보산업기술의 발전수준에 따라 평가된다고 할수 있다.

정보화사회란 정보의 대량 생산, 류동, 소비를 특징으로 하는 사회를 말한다. 지식의 생산과 축적은 정보기술의 발전을 토대로 이루어지고있으며 그 대부분이 인간의 언어, 자연언어로 표현되고 전달된다.

최근에 콤퓨터에 의한 자연언어처리에서는 언어현상들을 수집, 축적하여 콤퓨터로 쉽게 처리할수 있도록 여러가지 부호형식을 갖춘 대용량의 언어자료기지를 만들고있다. 이러한 과정에 나온것이 바로 코퍼스이다.

코퍼스(corpus)라는 용어는 라틴어 《corporatia》에서 기원하였으며 어원론적으로는 《덩어리》, 《묶음》, 《모임》, 《뭉치》 등의 뜻을 나타낸다. 현재 코퍼스라는 용어는 《전자화된 대량의 자료모임》, 《언어분석용자료모임》 등의 다양한 의미를 가지고 언어학분야와 언어처리분야에서 널리 리용되고있다.

조선어코퍼스구축에서 나서는 중요한 문제의 하나는 무엇보다먼저 생코퍼스를 옳게 구축하는것이다.

생코퍼스는 전자화본문의 단순한 유한모임으로서 말그대로 아무런 가공도 하지 않은 코퍼스이며 가장 기본적인 전자자료이기도 하다. 내용적으로 보면 크게 본문코퍼스, 대역코퍼스, 오유코퍼스 등 3개 부분으로 구성되여있다. 또 매개 본문자료는 그 출처, 저자, 용량 등의 기본문서정보를 부가한 헤더(header)와 본문(text)으로 구성되였다.

우선 본문코퍼스가 있다.

본문코퍼스는 일부 언어학자들이 인위적으로 모아놓은 문장표현인것이 아니라 신문, 잡지나 문학작품 등에서 실지로 쓰인 문장표현들을 모아놓은 방대한 량의 언어자료모임이다. 코퍼스를 일명 《실례형의 자료기지》라고 하는것은 실지로 쓰인 례로 될만 한 언어표현들을 일정한 규칙과 질서에 맞추어 정돈해놓은것과 관련된다. 즉 코퍼스안의 언어표현들을 그것들이 쓰인 분야나 시기 등의 환경정보와 문맥을 보존할수 있는 본문형식으로 그안의 언어학적정보들을 쉽게 취급하도록 해놓았다는 측면에서 코퍼스를 《자료기지》라고 표현하는것이다.

이러한 《자료기지》로서의 코퍼스는 주로 각이한 시기, 신문, 잡지, 문학작품, 교재 및 사전으로 구성되였다.

또한 대역코퍼스가 있다.

대역코퍼스는 말그대로 조선어와 외국어의 번역대응을 지어놓은 자료모임이다. 대역코퍼스에서 기본은 두 언어의 표현들사이의 대응관계이다. 대역코퍼스에서는 단어급, 구절급, 문장급, 문맥급대응을 실현한다. 이렇게 하면 앞으로 문제나 분야에 따른 특수한 번역대응표현을 쉽게 찾아내고 명확한 어휘지식을 획득하여 조선어와 외국어의 기계번역을 위한 기계사전구축이나 씨소러스작성 등에 도움을 줄수 있다.

또한 오유코퍼스가 있다.

오유코퍼스는 전형오유코퍼스와 자연상태오유코퍼스로 나뉜다. 전형오유코퍼스는

자연상태오유코퍼스를 분석, 분류하는 과정을 거쳐 얻은 코퍼스를 말한다.

자연상태오유코퍼스는 그 어떠한 가공을 거치지 않은 오유코퍼스를 말하는데 주로 사유모식전환규칙에 대한 연구, 두가지 언어문화차이 등 령역에 대한 연구에 유리할뿐만아니라 특히 조선어교육과정에서 교육자가 학생들의 언어습득의 전반과정을 전면적으로 관찰할수 있고 학습자의 학습에 영향을 주는 요소를 보여줄수 있으므로 효과적인 교수방법, 교수내용, 교수목표에 도달하는데 도움을 줄수 있는 장점을 가지고있다.

조선어코퍼스구축에서 나서는 중요한 문제의 하나는 다음으로 형태정보부가코퍼스를 옳게 구축하는것이다.

문법정보가 부여된 주석코퍼스는 형태정보를 부가한것, 구문정보를 부가한것, 의미정보를 부가한것으로 나뉜다. 조선어형태단어해석기는 조선어문장속의 단어형태를 자동으로 분석하여 단어를 사전적형태부단위로 구분해주는 단어구분프로그람이다. 이 프로그람은 본문의 형태분석코퍼스작성과 문장분석의 전 처리단계 그리고 본문검색 등에 리용될수 있다.

우선 형태단어해석기의 구성에 대하여 보기로 한다.

실례로 형태단어해석기《KMAnal 1.5》는 동적실행서고들인《KAnal.dll》, 《spdll.dll》로 구성된다. 동적서고들에는 조선어형태단어해석을 위한 처리함수들이 준비되여있으며 이 함수들은 c++나 VB와 같은 개발프로그람들은 물론 오피스제품들에 제공되는 VBA에서도 쉽게 리용될수 있다.

KMANAL.EXE: 조선어형태단어해석 대면프로그람

KCHAR.DLL: 조선어문자처리 동적실행서고프로그람

KANAL.DLL: 조선어단어처리 기본부분 동적서고프로그람

SPDLL.DLL: 조선어단어처리 확장부분동적서고프로그람

자료부분은 사전과 규칙부분으로 되여있다.

형태단어해석기가 정확하게 동작하자면 사전과 규칙이 반드시 적재되여야 한다.

MainDic.DDA: 단어분석용기계사전

MainDic.EXT: 확장사전+리용자정의사전

Appromiss.RUL: 규칙사전

이밖에 리용자가 간단히 고유명사 같은것을 등록하여 리용할수 있는 간이사전을 보충적으로 등록리용할수 있다.

또한 형태정보부가코퍼스의 응용에 대하여 보기로 한다.

조선어형태정보부가코퍼스의 구축, 리용은 조선어연구, 조선어교육, 사전학, 어휘론, 언어정보처리, 문법론(특히 형태론)의 연구에서 의의가 있다.

-사전적단어, 형태소들의 사용빈도 등의 구체적인 사용정보를 얻을수 있다.

-형태소들의 가결합성, 련어지식, 문접적성구, 문법적형태갖춤의 특징 등 형태론 및 어휘론적련관지식들을 연구할수 있다.

-새로운 어휘단어의 발굴, 단어나 형태부들의 구체적인 사용법을 조사정리함으로써 사전편찬, 언어교육 등에 리용될수 있다.

-형태소가결합정보, 기계사전 등을 갱신보충함으로써 형태해석기의 성능을 높일수 있다.

-고유표현(NE)발굴, 구문분석코퍼스제작 등의 입력자료로 리용된다.

이상에서 조선어정보와 관련한 문제들에 대하여 서술하였다. 그러나 그 구축과정, 방법, 내용 등 면에서 아직 많은 부족점들을 안고있으며 또한 앞으로 보충하거나 새롭게 건설해야 할 과업들도 많다.

우리는 과학기술이 급속히 발전하는 현실적요구에 맞게 조선어정보화에서 나서는 모든 문제들을 하루빨리 해결함으로써 우리 민족의 언어와 문화를 기록, 보존, 보급하는데 적극 이바지하여야 할것이다.

(필자는 중국실습생임)