김일성종합대학학보(어문학)
주체106(2017)년 제63권 제2호

도이췰란드어-조선어입말코퍼스의 구축

김 철 준

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《나라의 과학기술을 세계적수준에 올려세우자면 발전된 과학기술을 받아들이는것과 함께 새로운 과학기술분야를 개척하고 그 성과를 인민경제에 적극 받아들여야 합니다.》 (《**김정일**선집》 증보판 제11권 138~139페지)

현시기 우리의 언어학자들앞에는 지식경제시대의 요구에 맞게 언어의 정보화를 다그치고 언어학연구를 확고한 과학기술적토대우에 올려세워야 할 과업이 나서고있다. 특히 세계 많은 나라들에서 대규모언어코퍼스를 구축하고 그를 통한 언어연구와 언어정보처리를 진행하는데 많은 관심을 돌리고있는 현실은 언어코퍼스구축의 필요성을 더욱 절실히 제기하고있다.

이 글에서는 현대도이췰란드어입말코퍼스를 우리 나라의 실정에 맞게 구축하는데서 나서는 문제에 대하여 론하려고 한다.

오늘날 언어코퍼스는 나라의 경제 및 과학발전을 위한 중요한 전략적자원으로 인정되고있으며 여러 나라들에서는 언어코퍼스구축에 많은 자금과 로력을 투자하고있다.

우리 나라에서도 지난 시기 조선어해석 및 코퍼스체계를 개발하여 형태론적, 문장론적, 의미론적준위의 방대한 코퍼스체계를 구축하였다. 이와 함께 다국어병렬코퍼스에 의한 통계식기계번역체계를 개발한것을 비롯하여 코퍼스구축과 활용에서 많은 연구성과들이 이룩되였다. 최근에는 수십만쌍의 의학부문 도이췰란드어-조선어병렬코퍼스가 구축됨으로써 빠른 기간안에 세계적수준의 도조기계번역체계가 개발도입되는 성과도 이룩되였다.

현재 다른 부문과 마찬가지로 도이췰란드어학분야에서도 여러가지 류형의 언어코퍼스들을 구축하고 그것을 언어학연구와 교수실천에 적극 활용하여야 할 필요성이 제기되고있다. 여기에서 중요한 문제의 하나가 도이췰란드어입말코퍼스의 구축이라고 말할수 있다.

도이췰란드어-조선어입말코퍼스구축에서 나서는 중요한 문제는 무엇보다먼저 입말코퍼스의 구축대상을 바로 정하는것이다.

코퍼스구축대상을 바로 정하는 문제는 구축된 코퍼스의 실용성과 관련되는 매우 중요한 문제이다. 그리고 도이췰란드에 이미 구축되여있는 입말코퍼스와의 내용상중복을 피하면서 우리 나라의 실정에 맞는 도이췰란드어입말코퍼스를 구축하기 위하여서도 중요한 문제라고 말할수 있다.

도이췰란드에서 진행되고있는 입말코퍼스구축정형을 보면 언어사용을 대표하는 입말행위속에 내재하고있는 언어적원리를 해명하는데 가장 부합되는 연구방법을 입말코퍼스에 대한 경험적분석으로 간주하고 국가적인 힘을 넣어 대규모의 도이췰란드어입말코퍼스들을 구축하고있다.

그중 대표적인 도이췰란드어입말코퍼스들로서는 도이췰란드 만하임언어학연구소에서 구축한 도이췰란드어입말코퍼스와 도이췰란드 라이프찌히종합대학에서 구축한 다국어과학입말체문장비교코퍼스를 들수 있다. 이러한 코퍼스들은 모두 인터네트상에서 접근이 가능한 언어자료들이다.

만하임연구소의 도이췰란드어입말코퍼스는 국가적인 도이췰란드어입말자료기지의 한 부분으로서 여기에는 지난 기간 도이췰란드어학연구와 교수분야에서 채취한 방대한 량의 록음자료들과 전사화된 입말코퍼스자료들이 구축되여있다. 이 코퍼스에는 19개의 기본주제에 소속된 총 8 757개의 개별담화들이 포함되여있으며 동시에 각종 록화자료, 추가자료들이 들어있고 그것들을 종합적으로 검색할수 있는 탐색기능이 구비되여있다.

라이프찌히종합대학에서 구축한 다국어과학입말체문장비교코퍼스는 도이췰란드어, 영어, 뽈스까어로 진행된 구답시험, 과학강연, 론문발표 등에서 채취한 3개 나라의 입말문장들을 호상 비교분석할수 있도록 구축된 다국어비교코퍼스이다. 이 코퍼스에서 특징적인것은 도이췰란드어에 대한 모국어사용자 및 기타 언어사용자들이 발언한 입말문장들도 자료기지화함으로써 외국어로서의 도이췰란드어교수법연구에 입말코퍼스를 활용할수 있는 가능성도 제공하고있는것이다.

도이췰란드에 구축된 입말코퍼스들을 분석종합하여보면 그것이 단순한 언어학연구만을 목적으로 한것이 아니라 도이췰란드어교육에도 리용할수 있는 보다 실천적인 언어코퍼스로 발전하고있다는것을 알수 있다.

도이췰란드어-조선어입말코퍼스구축의 대상은 도이췰란드어연구뿐아니라 우리 나라의 도이췰란드어교육을 실용화, 종합화, 현대화하는데 적극 이바지하는 방향에서 정해져야 한다.

도이췰란드어-조선어입말코퍼스의 구축에서는 우선 언어연구와 교육실천에 실지로 써먹을수 있는 자료들을 선택하여야 한다.

도이췰란드어언어연구를 위하여서는 국내외에 이미 구축되여있는 입말코퍼스들가운데서 언어학연구에 필요한 자료들을 선택하는것과 함께 도이췰란드사람들의 일상담화나 실무담화들을 해당한 법률적 및 륜리적원칙에 기초하여 언어자료로 받아들여야 한다.

도이췰란드어교육실천에 리용하기 위한 입말코퍼스를 구축하는데서는 자연언어로서의 담화와 인공적으로 만들어낸 담화들을 구분하며 내용의 난이성정도에 따라 입말자료들을 분류하여 코퍼스에 포함시키는 문제가 중요하게 제기된다.

도이췰란드어-조선어입말코퍼스의 구축에서는 또한 선택리용된 입말자료들이 단일언어코퍼스가 아니라 이중언어코퍼스 즉 도조입말병렬코퍼스로 되여야 한다.

도이췰란드어입말코퍼스가 병렬코퍼스화되자면 구축된 언어자료들에 담화분석을 위한 부가정보들뿐아니라 입말문장쌍으로 순서화되고 해당한 조선어번역문과 함께 두 언어간의 대응관계를 나타낼수 있는 각종 형태론적, 문장론적 및 의미론적정보들이 각각 입력되여야 한다.

이렇게 병렬화된 도이췰란드어입말코퍼스는 도이췰란드어통역을 비롯하여 입말로 진행되는 도이췰란드어교수실천에 이바지할수 있을뿐아니라 기계번역의 난문제로 되고있는 입말문장번역의 효과성을 높이는데 적극 리용될수 있다.

도이췰란드어-조선어입말코퍼스구축에서 나서는 중요한 문제는 다음으로 입말코퍼스의 구축을 위한 언어학적원리를 옳게 해명하고 구축도구를 우리의 실정에 맞게 개발하는

것이다.

우선 도이췰란드어입말코퍼스구축의 언어학적원리를 옳게 해명하는 문제가 중요하게 나선다.

일반적으로 입말자료들은 자료선정단계, 기술처리단계, 분석처리단계를 거쳐 코퍼스로 구축된다.

자료선정단계에서는 코퍼스에 적합한 담화주제들을 확정하고 그에 가장 알맞는다고 보아지는 담화들에 어떤것이 있겠는가 하는데 대한 가설을 세운다. 이러한 기준에 따라 주제중심의 자료선정과 재료중심의 자료선정으로 코퍼스자료탐색방향을 선정한다.

주제중심의 자료선정에서 고려하여야 할 문제는 분석에 적용할 개념과 계획, 방법에 부합되는 담화주제들을 먼저 선정하는것이다. 다시말하여 언어학적분석에 알맞는 합당한 주제를 확정하는 작업이 선행되고 그에 따라 담화자료들이 선정되는것이 바로 주제중심의 자료선정에서 견지하는 방법론이다.

반대로 재료중심의 자료선정은 이미 얻어진 많은 자료들가운데서 일정한 분석목적에 알맞는 자료들만을 고르는 방법이다. 즉 자료선정이전에 이미 여러가지 각이한 담화주제들이 먼저 주어져있고 그속에서 필요한 자료들을 선택하는 자료선정방법이다.

기술처리단계는 입말자료들을 수자식 또는 상사식자료로 만들어내는 단계이다. 일반적으로 기술처리되는 입말자료라고 할 때에는 록음록화자료들을 의미한다.

기술처리단계에서는 시간적으로 엄밀히 구분되여있는 3개의 공정을 거쳐야 한다.

첫 공정은 계획화공정으로서 여기에서는 자료선정단계에서 내린 결론에 따라 언어자료의 풍부성과 세부요소, 담화주제와의 적합성, 기술적수단, 시간소요량 등을 계획한다. 이 공정을 통하여 수집해야 할 담화들을 목록화한 계획서, 자료수집에 리용할 기재선정, 자료수집의 구체적인 일정 등이 확정된다.

둘째 공정은 조직화공정으로서 여기에서는 해당한 담화주제를 실천에 옮길 담화자들을 선정하게 된다. 이 단계에서 관심을 넣어야 할 문제는 담화자들에게 진행할 담화내용에 대하여 구체적으로 설명해주어 그들이 목적한 담화를 원만히 만들어낼수 있도록 하는것이다. 여기에서 주의해야 할것은 담화의 목적을 너무 구체적으로 설명해주는 경우 분석자가 관찰하려던 언어현상이 자연스럽게 일어나지 못하고 어떠한 요인에 의해 유도당할수 있는 약점이 있다는것이다. 이러한 조직화공정은 인공적으로 만들어내는 담화에만 해당되며 자연적인 담화에 대하여서는 적용하지 않는다.

셋째 공정은 록음공정으로서 해당한 담화주제에 따라 담화자들의 말을 록음록화하는 공정이며 여기에서는 실제적인 코퍼스원천자료들이 확보된다.

분석처리단계는 앞선 단계들에서 얻어진 록음록화자료들이 비로소 언어적분석을 위한 코퍼스로 만들어지는 단계이다. 여러개의 공정들로 이루어진 분석처리단계에서는 언어자료들이 담화의 환경과 담화자들사이의 관계, 그들의 호상작용방식에 따라 순서화되고 가공된다.

이 단계에서 가장 중요한 부분이 전사화공정이다. 전사법은 입말을 기호화하여 글말자료로 전환시키는 기술이다.

전사법에서 기호화는 단순히 입말을 서사화하는것이 아니라 사람들이 보통 쓰고있는 받아쓰기의 방법으로 서사물을 만들고 거기에 담화환경의 구체적인 정황과 목소리, 담화

진행과정 등을 시각화할수 있는 기호들을 첨부하는 적기법이다.

전사법은 음악에서 악보를 보면서 음악작품을 동시에 연주할수 있는것처럼 입말의 진행과정을 마치 눈으로 보는듯이 분석할수 있게 한다. 도이췰란드에서는 악보식적기법(Partitur-Schreibweise)과 담화분석용전사법(GAT)체계가 보편적으로 쓰이고있다.

전사법에 의하여 만들어진 도이췰란드어입말코퍼스의 형태를 다음의 실례를 통하여 보기로 한다.

례:

125a wollen nämlich mal in die geschichte zurückdenken ich genau
G
126 kann ichs belegen von den germanen von tacitus her und der
127 spricht auch von der + genau von der ehe die frau hat
E
128 bestimmte + da is es sogar so dass dass also bei den germanen
D
129 wurde ja die frau so unheimlich hoch eingeschätzt dass + der
E
130 [nann] sogar da für sie zu bezahlen hatte /↙+ [und] wenn wir
.T
(ironisch)
131b [ja] [hm]
BF BF
132 mal zurückdenken wie wars bei den römern haben die eine
133 form der ehe geführt eigentlich ja nicht ↗+
1 VF ?
134b doch ja: ↗+ [sie hatten doch auch nur ein] + ein + ein
F !T D D
135a ja du [du NATÜRLICH]
BF LG !
(überschwenglich)

(A: 한번 력사를 되새겨보겠습니다. 나는 정확히 따시뚜스가 쓴 도서인《게르만인》을 통하여 증명할수 있습니다. 이 책에서 그는 결혼에 대하여 구체적으로 언급했는데 결혼할 녀성들이 게르만인들속에서 매우 높은 가치를 가지고있어 지어 어떤 남자들은 녀인을 얻으려고 숱한 돈을 지불했다고 합니다.

B: 예, 그렇습니다.

A: 그렇다면 로마인들이 어떠한 결혼형식을 가지고있었는가를 생각해본다면 그들은 원래 구체적인 형식을 가지고있지 않았습니다.

B: 아닙니다. 그들도 일정한 형식을 가지고있었습니다.

A: 글쎄 물론 …)

우의 실례코퍼스에 리용된 각종 전사기호들을 보면 담화자들의 동시적인 발언을 꺾

쇠괄호로 표시하였으며 입말진행과정의 말차례들을 수자로 배렬하였다. 그밖에도 상대방의 발언에 대한 긍정(BF=Bestätigungsform), 각종 형태의 휴지(D, T, E)와 휴지의 길이(+), 억양의 높낮이(↙↗), 특별히 강조되는 부분에 밑줄을 그어주는것을 비롯하여 기타 심리 및 반응정보들이 부가되였다.

입말자료들의 전사화가 끝나면 매 자료들에 구별정보들을 부가해주는 작업이 진행된다. 그러한 정보들에는 자료식별번호, 작성자이름, 작성날자와 시간, 교정날자, 자료형태, 전사방법, 교제류형 등 각이한 정보들이 첨부된다.

코퍼스구축작업은 언어학적으로 과학성이 철저히 담보되여야 하며 구축설계와 준비, 실행, 완료의 모든 단계가 하나의 생산공정처럼 맞물려 진행되여야 실용적인 코퍼스를 만들어낼수 있다.

또한 도이췰란드어입말코퍼스구축도구를 개발하는 문제도 중요하게 나선다.

현재 우리 나라에서 개발리용되고있는 코퍼스구축도구들은 모두 글말자료를 기본원천으로 삼고있는것으로서 앞으로 입말코퍼스구축을 위한 전용도구를 개발하여야 할 과업이 나서고있다.

도이췰란드어에서 입말코퍼스구축에 리용되고있는 도구들로는 EXMARalDA와 FOLKER프로그람이 있다.

EXMARalDA프로그람은 담화해석용 확장표식언어체계로서 세계적으로 널리 리용되고있는 입말코퍼스구축도구이며 도이췰란드 만하임언어학연구소에서 개발한 FOLKER프로그람은 자기 연구소의 입말코퍼스구축을 위한 전용도구이다. 이러한 입말코퍼스구축도구들은 록음록화된 입말자료들로부터 전사화된 코퍼스를 자동생성하며 음성인식뿐아니라 담화의 정황과 담화자들의 심리상태 등 여러가지 구별정보들도 입력할수 있는 종합적인 도이췰란드어입말코퍼스구축도구이다.

도이췰란드어-조선어입말코퍼스구축도구를 개발하는데서 중요한것은 이미 개발된 도조병렬코퍼스구축도구에 음성인식체계와 전사화원리를 도입하는것이다. 이밖에도 글말과 다른 입말의 형태론적, 문장론적, 의미론적특성을 반영할수 있는 부가정보들을 확정하는 문제도 절실하게 나선다.

우리는 경애하는 최고령도자 **김정은**동지의 교육중시사상을 높이 받들고 도이췰란드어-조선어입말코퍼스를 비롯한 여러가지 류형의 언어코퍼스개발에 더욱 힘을 넣음으로써 언어학연구와 외국어교육을 지식경제시대의 요구에 적극 따라세워야 할것이다.