국어 정보화 사업의 미래와 전망

홍종선 · 고려대학교 교수
남경완 · 고려대학교 연구 교수

1. 머리말

오늘날 우리 사회는 수많은 정보의 흐름 속에서 움직여 간다. 온갖 종류의 정보가 만들어지고 그것이 다양한 매체를 통해 유통되며 다시금 재생산된다. 이것이 바로 현대 사회를 특징짓는 정보화 사회(information oriented society)의 모습이다. 우리가 접하는 대다수의 정보는 여전히 우리가 일상적으로 사용하는 언어 즉 말과 글로 구성되어 있지만, 이러한 단순 자료가 효율적인 정보로서 가치를 가지고 대량으로 유통되기 위해서는 필연적으로 컴퓨터나 인터넷과 같은 현대 정보 통신 매체와 결합되지 않으면 안 된다.

이 과정에서 수많은 언어 자료가 컴퓨터를 통해 체계화된 정보로서 가공되고 유통될 수 있도록 만드는 자연 언어 처리(natural language processing)에 대한 관심이 나타났고, 국어의 경우에도 국어 정보 처리의 영역이 꾸준히 연구되어 오고 있다. 그러나 이러한 연구를 각 연구자들이 개별적 차원에서 진행하기에는 무리가 따른다. 대단위의 언어 정보를 생

산하고 그것을 바탕으로 다양한 응용 프로그램을 개발하는 것에는 막대한 예산과 전문적 인력이 뒷받침되어야 하기 때문이다. 따라서 국어 정보 처리 연구를 지속적으로 수행하기 위해서는 이를 도와주는 국가 차원의 뒷받침이 반드시 필요하고, 이와 같이 국어 정보 처리를 위한 기초 자료의 구축과 프로그램 개발 등을 총괄적으로 지원하는 것이 바로 지금까지 국어 정보화 사업의 실체라고 할 수 있다.

이에 따라 국내에서도 약 20여 년 전부터 시작된 개별 학자들의 연구들이 서서히 축적되어 왔고, 더불어 '21세기 세종 계획'(이하 '세종 계획')으로[1] 대표되는 국가 정책적인 국어 정보화 사업으로 인해 현재까지 국어 정보화 부문에서 괄목할 만한 성장이 이루어져 왔다. 그러나 정보화 사회의 미래를 내다볼 때, 지금까지의 성과는 사실상 그 출발점에 불과하다고 볼 수밖에 없다. 언어 정보의 정리는 불과 몇 년 만에 끝날 수 있는 일이 아니며, 더더욱 다양화되고 있는 수많은 응용 영역에서 우리 국어가 계속 중요한 역할을 담당할 수 있기 위해서는, 그리하여 미래 사회에서도 국어의 위상과 효용성을 확보하기 위해서는 국어 정보화의 나아가야 할 길이 앞으로도 멀기만 하다.

이 글에서는 지금까지 진행되어 온 국어 정보화 사업의 내용과 성과를 간략하게 살펴보고, 이를 바탕으로 앞으로 진행될 국어 정보화 사업의 미래와 전망에 대해 논의해 보고자 한다.

1) '21세기 세종 계획'은 문화체육관광부가 국립국어원 및 관련 학계와 더불어 지난 1998년부터 10년간의 계획으로 진행한 국어 정보화 사업으로서, 2007년까지 언어 정보 문화의 기본 바탕과 자원을 확충하기 위한 '국어 정보화 중장기 발전 계획'의 일환으로 수립된 것이다. 이는 정보화 사회의 확산에 따라 보다 적극적으로 한국어의 정보화에 기여하기 위한 것으로써, 중장기적인 계획을 가지고 본격적으로 국어 정보화 사업이 정책적으로 시작된 것은 "21세기 세종 계획"이 시초라고 할 수 있다.

2. 국어 정보화 사업의 성과와 과제

국어 정보화 사업의 미래를 조망하기에 앞서 지금까지 국어 정보화의 영역이 어떤 분야에서 어느 정도의 성과가 이루어져 왔는지에 대해 개관적으로 살펴보고자 한다.[2] 이를 통해 현재까지 국어 정보화 사업이 주로 어떤 분야에 집중해 왔고, 또 어떤 분야가 아직 미진한지를 살펴볼 수 있을 것이다.

2.1. 국어 정보화 사업의 성과

지금까지 진행되어 온 국어 정보화 사업의 세부 영역은 크게 두 분야로 나누어 볼 수 있다. 첫째는 국어의 기초 자료를 구축하는 것이고, 둘째는 이를 바탕으로 다양한 정보 처리 프로그램을 만드는 전산학적 응용 과정이다.

이 가운데 기초 자료의 구축은 다시 코퍼스(corpus) 구축, 전자 사전 구축, 전문용어 정비 등으로 세분되는데, 이와 같은 기초 자료를 구축하는 것 자체는 상업적 이익을 기대하기 어렵기 때문에 국가 차원의 정책적 지원이 더욱 필수적이다.

코퍼스란 언어의 본질적인 모습을 총체적으로 보여줄 수 있는 언어 자료의 집합으로서, 어떤 종류의 글이나 말, 어떤 형식의 글이나 말이든 실제 사용된 언어 자료를 모아 놓은 것이다. 이것은 언어 연구의 각 분야에서 기초적인 자료로 이용될 수 있는데, 실제 언어의 모습을 정확히 반영하기 위해서는 필연적으로 일정 수준 이상의 대규모 자료가 확보되어야만 한다. 따라서 대규모의 국어 코퍼스를 만드는 것은 국어 정보화의 기

2) 국어 정보화 사업의 세부 영역은 매우 다양하다. 그것은 정보화 대상이 되는 언어 자료의 성격이 다양하기 때문이기도 하고, 응용 분야의 영역도 관련 전문가를 대상으로 하는 것뿐만 아니라 일반인들도 쉽게 접하고 다룰 수 있는 것 등으로 세분화되기 때문이다.

초 자료 구축 사업에서 가장 핵심적인 사업이라고 볼 수 있으며, 그에 따라 가장 먼저 시작된 부분이기도 하다.

현재까지 국내에서는 1988년부터 구축되어 온 연세 한국어 말뭉치를 비롯하여 KAIST, 고려대학교, 울산대학교 등에서 본격적으로 코퍼스를 구축해 왔으며, '세종 계획'에서도 우리 언어의 총체적인 모습을 보여줄 수 있는 대규모의 국가 코퍼스를 구축하여 국어 정보화 발전의 기반을 조성하고자 하였다.[3] 특히 세종 계획에서는 국어 기초 자료 구축 분과에서 원시 코퍼스로는 5,700만 어절, 형태 분석 코퍼스로는 1,000만 어절 규모의 코퍼스를 구축하여, 양적으로 괄목할 만한 성과를 거두었다.

전자사전은 코퍼스와 함께 국어 정보화의 다양한 영역에서 일차적인 언어 자료가 되는데, 그것은 단순히 종이 사전을 컴퓨터 화면에 옮겨 놓은 것이 아니라 한국어에 대한 정밀하고 광범위한 정보를 프로그래밍 언어로 기술함으로써 형태소 분석기나 기계 번역기 등과 같은 응용 프로그램에서 전체 모듈 사이의 통합에 기여하고 정보 데이터베이스로서의 기능을 담당한다. '세종 계획'에서는 전자사전 구축 분과에서 핵심 사전으로서의 체언 사전 및 용언 사전과 조사, 어미, 부사, 관형사, 연어, 관용표현, 복합명사구 사전 등을 포함하여 총 약 35만 어휘 규모의 전자사전을 구축하여 양적으로는 상당한 성과를 거두었다.[4]

이 외에도 전문용어의 국가 표준 정립을 위하여 '세종 계획'의 전문용어 표준화 분과에서는 전문용어의 목록을 데이터베이스로 구축하여 표준화하는 작업을 시행하기도 하였으며, 한글의 문자 코드 문제와 관련하여 문자 코드 표준화 분과에서는 문자의 기본적인 입출력 문제 및 자형이나 문자 정보의 표준화 연구, 그리고 유니코드에 불완전하게 수용됨으로써

3) 외국의 예로는 1963년부터 2년간에 걸쳐 100만 어절로 구축된 브라운 코퍼스(Brown Corpus)를 시초로, 1990년대에는 1억 어절 이상의 영국 국가 코퍼스(British National Corpus) 등이 구축된 바 있다.(서상규 · 한영균 1999:28)

4) 다만, 세종 전자사전의 현재 형식으로는 여러 응용 시스템에 곧바로 적용될 만큼 다른 모듈과의 통합성이 확보되었다고 보기 어렵고, 또한 범용 전자사전으로서의 다기능적인 역할을 담당하기에도 무리가 있는 것으로 판단된다.

정보 처리에 어려움을 겪고 있는 옛 한글의 정보 처리 연구 등을 수행하기도 하였다. 아울러 한민족 언어 정보화 분과에서는 남북한 및 국외 동포간의 언어 이질감을 해소하고 통일 이후의 국어 정보화를 위한 기초 작업으로서 남북한 언어 비교 사전을 구축하고 한민족 정서법 변환 프로그램을 개발하기도 하였다.

이상과 같은 기초 자료의 구축 성과를 바탕으로 이를 다양한 분야에서 활용할 수 있도록 전산적으로 응용한 여러 프로그램이 개발되어 왔다. 이것은 자연 언어 처리 분야나 음성 공학과 같이 전문 연구자들을 위한 것뿐만 아니라 정보 검색, 문서 요약, 텍스트 교정 등과 같이 일반인들 역시 일상생활에서 쉽게 접하고 이용할 수 있는 다양한 분야에 적용되어 왔다.

자연 언어 처리 과정에서 다른 응용 프로그램의 기반 기술이라고 할 수 있는 형태소 분석기는 현재까지 다양한 방법론이 연구되어 왔고, 실제 프로그램 역시 여러 가지가 상용화되어 있으며, '세종 계획'에서도 원시 말뭉치에 형태소 태그를 자동적으로 부착하는 '지능형 형태소 분석기'를 개발하여 배포하고 있다. 또한 음성 공학 분야 역시 음성 인식과 음성 합성의 기술을 적용한 ARS(Automated Response System)나 텍스트 음성 변환 시스템 등이 여러 민간 업체에서 개발되어 왔다.

이 외에도 문서를 작성할 때 철자법이나 띄어쓰기, 문장 구성의 오류 등을 자동적으로 검색하여 수정해 주는 문법 검사 시스템(text critique system)이나, 많은 분량의 정보 가운데 핵심적 내용만을 자동적으로 요약해 주는 자동 문서 요약 시스템(text summarization system) 등도 현재 개발되어 일반인들이 쉽게 접근할 수 있다.[5)]

이러한 응용 프로그램 가운데 자연 언어 처리의 핵심적 기술들이 총체적으로 집약되어 있는 것은 바로 기계 번역(machine-translation)이라 할 수 있는데, 여기에서는 코퍼스와 전자사전을 통한 어휘 정보의 추출과 형

5) 이런 종류의 프로그램들은 대개 '아래아한글'이나 'MS-Word'와 같은 문서 편집기 프로그램의 내부에 포함되어 있기 때문에, 일반인들이 가장 손쉽게 접할 수 있는 응용 프로그램이라고 할 수 있다.

태소/구문 분석기를 이용한 문장 구조의 분석 등이 총체적으로 결합되어 있으며, 음성 공학 영역의 연구와도 결합될 수 있다. 기계 번역 분야에서도 지금까지 다양한 방법론에 대한 이론적 연구와 제품 개발이 이루어져 왔다. 영한 번역기의 경우 이미 1980년대 중반부터 실험적인 시스템에 대한 연구가 시작되었고, 1990년대에 들어 기초적인 수준에서 상용화된 제품들 몇 종이 출시된 바 있으며, 언어적 차이가 상대적으로 크지 않은 일한, 한일 번역기의 경우에는 웹 문서의 실시간 번역 프로그램도 개발되어 왔다. 그러나 이러한 연구는 대체적으로 전산학 중심의 토대를 가지고 있었고, 그에 따라 국어에 대한 심도 있는 분석보다는 두 언어 사이의 직접 변환을 시도한 것이 대부분이었다. 이와는 달리 자연언어 처리를 위한 언어의 분석 이론을 먼저 정립함으로써 국어의 정보 처리 시스템을 구축하고자 한 연구가 비교적 최근에 진행되기도 하였다. 특히, 고창수(2002) 등에서는 정보처리를 위한 언어 분석의 이론으로서 자질연산문법(FCG; Feature Computational Grammar)을 제안한 바 있고, 이를 바탕으로 지난 2003년 고려대학교 민족문화연구원 기계번역연구실에서는 한영/영한 번역기 '트랜스마스터 2003'을 개발하기도 하였다.

이상에서 보듯이 국어 정보화 사업은 기초 자료 구축과 전산학적 응용 모두 다양한 세부 영역으로 구성되고, 그것의 확장 가능성은 앞으로도 무궁무진하다. 그러나 이러한 다양한 프로그램들은 기본적으로는 모두 국어에 대한 이론적 연구와 국어 정보화의 기초 성과를 바탕으로 만들어진 것으로서 현재까지의 성과에 만족할 수 없고 끊임없이 확대되어 나가야 한다.

2.2. 국어 정보화 사업의 반성적 회고

위에서 살펴본 바와 같이 국어 정보화에 대한 기초 연구는 지난 10여 년간 괄목할 만한 성장을 이루었다. 그것은 최근 들어 국어 정보화에 보

다 많은 학자들이 관심을 가지게 되었고, 그에 걸맞는 국가 정책적 지원이 결합한 결과로 볼 수 있다. 그러나 이와 동시에 지난 국어 정보화 사업은 몇 가지 점에서 아쉬움을 남긴 것도 사실이다. 물론 이것은 끊임없이 새로운 연구 주제를 찾아내고, 학제 간 연구가 본격화되어야 하는 국어 정보화의 근본적인 속성상 걸음마 단계에서 어쩔 수 없이 겪게 되는 과정이었다고 볼 수도 있다. 하지만 앞으로 계속되어야 할 국어 정보화 사업을 좀 더 체계적이고 효율적으로 진행하기 위해서는 지난 과정에 대한 반성적 회고가 필요하다고 할 것이다.

우선, 학문 영역 간 역할 분담과 협력이 더욱 확대 강화될 필요가 있다. 국어 정보화의 영역 가운데 국어학 영역에서는 주로 코퍼스, 전자사전 등과 같은 기초 자료 구축에 집중해 왔고, 여러 응용 프로그램 개발은 주로 전산학 영역에서 연구되어 왔다. 과거 1980년대부터 정보통신부나 과학기술부 주관의 국어 정보 처리 기술 개발 프로젝트가 다년간 진행된 바도 있으나, 이러한 연구 성과가 '세종 계획'의 연구와 효율적으로 연계되어 자연스럽게 확장되어 오지 못한 것이 사실이다. 결국 국어학과 전산학의 이러한 분업화는 각각 자신의 고유한 영역에서 연구 성과를 축적해 나가는 등의 의미 있는 성과를 나름대로 거두었으나, 앞으로는 이들 두 부문 사이의 유기적인 협동 체제, 즉 공동 연구가 좀 더 요구되며, 특히 응용 프로그램의 개발에도 국어학/언어학의 연구자들이 적극적으로 참여할 필요가 있다. 전산 프로그램 가운데 수준 높은 언어 전산화에는 고도의 언어학적 지식과 감각이 절대적인 것이기 때문이다.

국어학의 영역에서 살펴볼 때, 코퍼스 구축 등과 같은 기초 자료 구축은 상당한 정도의 수준에 이르렀으나, 이렇게 대단위로 구축된 여러 기초 자료를 광범위하게 활용하는 단계에는 아직 이르지 못하였다. 일부의 연구자들이 이들 기초 자료를 토대로 연구를 하고 있으나, 좀 더 많은 연구자들이 손쉽게 이들 자료를 적극적으로 이용할 수 있는 환경과 기반이 갖추어져야 할 것이다.

지금까지의 '세종 계획'으로 대표되는 국어 정보화 사업은 주로 기초 자료 구축에 집중되어 왔는데, 국어 정보화 영역에서 지원해야 할 연구 주제가 응용 프로그램 개발 분야로 크게 확대될 필요성이 있다. 기초 자료를 구축하는 것은 가장 근본적인 출발점임에는 틀림없지만, 방대한 언어 자료의 특성상 정제된 기초 자료를 구축하기까지는 많은 시간이 필요하다. 따라서 기초 자료로서 구축한 국어 정보들이 여러 응용 프로그램 개발과 직접적으로 연결되기 위해서는 기초 자료 구축 단계에서부터 향후 이것을 이용할 프로그램 개발 단계에서 필요로 하는 것이 무엇인지를 파악하고 있어야 한다.

이런 측면에서 데이터베이스 검색이나 문서 오류 수정, 문서 요약 등과 같은 실제적인 응용 프로그램 개발에도 관심의 폭을 넓혀 나갈 필요가 있다. 이러한 경험을 바탕으로 자연 언어 처리 기술의 총체적 집합이라 할 수 있는 기계 번역기와 같은 프로그램 개발 등으로 이어질 수 있을 것이다. 또한 문자 언어와 달리 자료 구축 단계에서부터 어려움이 많은 음성 언어에 대한 관심도 필요하다. 현재 컴퓨터를 비롯한 정보 기기의 성능은 10여 년 전과 비교해 비약적으로 발전한 상태이므로, 음성 언어 정보에 대한 기초 자료 구축도 보다 확대될 수 있을 것으로 보인다. 따라서 음성 언어 정보를 구축하는 것 역시 코퍼스 구축 단계에서부터 시작하여, 음성 인식과 음성 합성 등과 같이 일반인들의 실생활에 밀접한 연구도 심화 발전될 필요가 있다.

3. 국어 정보화 사업의 미래와 과제

인간의 언어가 살아 숨쉬며 끊임없이 변화하는 한, 언어 정보를 정리하고 가공하며 그것을 다시 응용하는 과정 역시 끝이 없는 연구 과제가 될 수밖에 없다. 또한 하루가 다르게 급성장하고 있는 현대 정보화 사회의 모습을 감안하면 국어 정보화의 영역과 수준이 어느 정도까지 발전할 것

인지 예측하는 것은 불가능할지도 모른다. 그러나 이러한 현실적인 어려움을 인정하더라도 국어 정보화 사업을 효율성 있게 추진하기 위해서는 장·단기적으로 구체적인 계획을 수립하는 것이 중요하다. 이 장에서는 국어 정보화 사업의 성과와 반성을 토대로 국어 정보화 사업이 나아가야 할 방향에 대해 기초 자료 구축, 응용 프로그램 개발, 기반 이론 연구 문제로 나누어 살펴보고자 한다.

3.1. 기초 자료 구축

코퍼스 구축으로 대표되는 기초 자료 구축 분야는 현재까지 양적으로는 괄목할 만한 성과를 이루었다고 볼 수 있다. 기본적인 원문 텍스트의 내용 그대로만 저장되어 있는 원시 코퍼스(raw corpus)나 여기에 언어학적 정보 등을 붙인 표지화 코퍼스(tagged corpus)는 여러 연구 기관에서 상당한 규모로 구축해 왔다. 이러한 양적 성장은 국어 정보화 사업의 도입기에 기본적으로 집중해야 할 과제였지만, 앞으로의 방향은 코퍼스의 양적 성장과 함께 질적 성장에도 초점이 맞추어질 필요가 있다.

이를 위해서는 우선 현재 구축된 코퍼스의 정련 작업이 이루어져야 할 것이다. 대규모의 언어 자료를 많은 연구자들이 구축하는 과정에서는 필연적으로 비일관적인 분석이나 오류가 나타날 수밖에 없는데, 이에 대한 교정 작업이 수반되어야만 양적 성장과 더불어 내적인 질적 성장을 이룰 수 있다. 더욱이 기초 자료를 이용한 여러 응용 연구가 더욱 탄탄한 기반 위에서 신뢰할 수 있는 연구 결과로 이어지기 위해서는 기초 자료의 신뢰도를 높이려는 노력이 끊임없이 계속되어야 하는데, 그것은 원시 코퍼스나 표지화 코퍼스 단계에서 이루어져야 한다.

이러한 기본 코퍼스의 정련 작업과 함께 보다 많은 언어적 정보를 포함하여 고차원적인 분석 결과를 붙인 분석 코퍼스(analyzed corpus)를 구축하기 위한 노력이 필요하다. 단어 차원의 용례 검색이나 품사 정보의

추출과 같은 기본적인 정보만으로는 앞으로 개발될 수많은 응용 프로그램의 기초 자료가 되기에 부족한 감이 없지 않다. 실제 텍스트와 담화의 언어 정보를 빠짐없이 파악하기 위해서는 단순한 형태 정보 외에 문장의 구조나 문장 간의 관계 정보를 표시해 주는 구문 분석 정보도 필요하고, 나아가 단어, 문장, 텍스트 전체의 의미적 분석 정보도 필요하다. 현재 이러한 분석 코퍼스 구축을 위한 다양한 이론적 기초 연구들이 진행되고 있고, 실제 실험 코퍼스들을 구축하려는 움직임이 있으므로 이에 대한 지원과 관심이 더욱 크게 요구된다.

코퍼스 구축에 있어서 특히 앞으로 좀 더 보완해 나가야 할 부분은 음성 언어 코퍼스이다. 현대 정보화 사회의 도래와 더불어 기존의 언어 연구 역시 고정된 텍스트 중심의 문자 언어에 대한 연구로부터 화자의 실제 발화를 대상으로 하는 음성 언어에 대한 연구가 활발히 진행되고 있으며, 이러한 연구 성과는 음성 인식과 음성 합성으로 대표되는 음성 공학의 연구와 직접적으로 연계되고 있다. 음성 공학은 생명 공학과 함께 21세기에 가장 촉망 받는 분야로 손꼽히고 있다. 이에 대한 준비가 언어 전산의 기초 부문에서 이루어져야 할 것이다.

음성 언어에 대한 자료 구축은 문자 언어의 그것과 달리 기초적인 자료 수집에서부터 많은 어려움이 있다. 말소리만을 연구 대상으로 하더라도 수많은 녹음 작업은 기본적으로 병행되어야 하고, 좀 더 정확한 자료를 만들기 위해서나 표정이나 몸짓 등의 동작 언어를 포함하기 위해서는 실제 발화 상황을 녹화하는 것이 필수적이기 때문이다. 이로 인해 현재까지 음성 언어 자료의 구축은 개별 연구자 혹은 연구팀들이 각기 소규모로 진행해 온 것이 많았으며,[6] 또한 지금까지 구축된 음성 언어 코퍼스

6) 대표적인 예로는 한국학술진흥재단 기초학문육성지원사업을 통해 진행된 "한국인의 의사소통 능력 발달 단계 연구"(장경희, 2002년), "한국어 구어 문법 기술을 위한 기초 연구:대학생 대화 말뭉치를 중심으로"(서상규, 2002년), "한국 유아의 신체적 경험에 의한 은유 습득과 발달"(이종열, 2005년), "한국어 구어 분석을 위한 문법 모형의 개발과 활용"(신지영, 2007년) 등이 있다. 이러한 다양한 기초 연구에서는 영, 유아 코퍼스에서부터 대학생의 자유 발화 코퍼스에 이르기까지 다양한 연령 대의 음성 언어

역시 기초적인 원시 코퍼스인 경우가 많고, 좀 더 상세한 정보를 담고 있는 자료도 녹음된 말소리를 전사한 소규모의 구어 전사 코퍼스 수준에 머물고 있다.[7] 그러나 실제 음성 언어 자료가 음성 공학에 직접적으로 활용되기 위해서는 녹음된 자료에 다양한 음성과 운율 정보를 포함한 음성 데이터베이스로서의 자료가 필요하므로, 이러한 음성 언어 기초 자료를 구축하려는 노력이 계속되어야 할 것이다.

또한, 코퍼스 자료가 자료 그 자체로 머물지 않고 실제 다양한 응용 영역에서 효율적으로 활용될 수 있는 방안에 대한 고민도 필요하다. 현재 주로 국어학 분야에서 코퍼스를 이용한 연구가 진행되고 있으나, 방대한 언어 자료로서의 코퍼스가 가지고 있는 활용 가능성은 무궁무진하다고 할 수 있기 때문이다. 이를 위해서는 대규모의 균형 코퍼스와 함께 다양한 종류의 특수 목적 코퍼스를 구축해 나갈 필요가 있다. 특별한 목적을 사전에 규정하지 않고 일반적인 언어 조사를 위해 텍스트를 모은 것을 일반 코퍼스(general corpus)라고 하며, 이러한 일반 코퍼스는 대개 일상생활에서 접하는 다양한 장르의 텍스트들이 균형을 이루도록 구성되는 균형 코퍼스(balanced corpus)로 구축된다. 이는 코퍼스 구축 단계에서 가장 먼저 이루어져야 하는 것으로, 당연히 국어 정보화 사업의 기초 자료 구축에서도 이 부분에 가장 큰 역량이 집중되어 왔다.

그러나 코퍼스를 이용한 국어 연구가 좀 더 많은 영역으로 확대되기 위해서는 비록 소규모일지라도 특정한 목적을 가지고 구축되는 특수 코퍼스(specialized corpus)가 다양하게 구축될 필요가 있다. '세종 계획'에서도 구축된 바 있는 역사 자료 코퍼스라든가 방언 코퍼스, 전문용어 코퍼스 등이 그것인데, 이러한 특수 코퍼스들은 특정한 분야에 직접적으로 응용될 수 있다. 특히 국어 교육이나 외국인을 위한 한국어 교육과 같은

코퍼스를 구축한 바 있다.

7) "21세기 세종 계획"에서 구축한 구어 코퍼스 역시 전체적으로는 약 500만 어절 규모이나 그 가운데 형태 분석 코퍼스는 약 100만 어절 규모로 파악되며, 이는 문어 코퍼스에 비하면 아직도 코퍼스 구축의 시작 단계라고 볼 수 있다.

영역에서는 다양한 학습자 코퍼스가 구축될 수 있다. 가령, 학습자 오류 코퍼스는 한국어 모국어 화자들이 주로 범하는 오류가 무엇인지를 파악하여 교육의 영역에서 직접 이용될 수 있으며, 한국어 교육의 영역에서도 한국어를 학습하는 외국인들이 작성한 한국어 자료를 데이터베이스로 구축하면 그것 역시 훌륭한 학습자 코퍼스가 될 수 있다. 한국어 학습자들의 코퍼스는 현재 한국어 교육 기관에서 매우 불충분하게 조금씩 구축하여 연구 등에 이용되고 있는데, 여러 교육 기관이 연합하여 다양한 코퍼스를 계획할 수 있을 것이다. 이와 같이 다양한 목적의 특수 코퍼스는 국어학의 전 영역에 걸친 연구뿐만 아니라 교육 등 여러 응용적 현장에서 기초 자료로서 효율적으로 이용될 수 있을 것이다.

이상과 같이 기초 자료를 구축하는 것은 국어 정보화 사업에서 지금까지 가장 집중해 온 분야이지만, 앞으로도 자료의 규모를 확대함과 동시에 보다 넓은 영역에서 다양하게 활용될 수 있도록 질적인 성장을 이루어 나가야 할 것이다.

3.2. 응용 프로그램 개발

국어 정보 처리의 기술력이 발전해 갈수록 이를 이용한 다양한 응용 프로그램들이 속속 개발되고 있다. 이는 형태소 분석기나 전자사전과 같이 기초 자료로서 구축된 성과에서 다양한 언어학적 정보를 손쉽게 추출하여 국어학 연구에서 활용할 수 있도록 도와주는 프로그램뿐만 아니라 정보 검색과 같이 이미 일반인들의 일상생활 속에 깊숙이 들어와 있는 프로그램에 이르기까지 매우 다양하다. 그러나 이러한 프로그램들은 일부 기초적인 종류를 제외하고는 아직 만족할 만한 수준에 이르렀다고 보기는 어려우며, 새롭게 개발되고 있는 정보 통신 기기와 결합되기 위해서는 보다 다양한 콘텐츠로 확장될 필요가 있다.

원시 코퍼스에 자동으로 품사 태그를 부착시켜 주는 형태소 분석기는

다양한 방법론에 대한 이론적 고찰과 함께 여러 연구팀에서 프로그램을 개발한 바 있으며, '세종 계획'에서도 '지능형 형태소 분석기'를 배포하고 있다. 아직은 만족할 만한 성공률을 보인다고 할 수는 없지만, 대단위의 자료를 다루는 데 있어서 자동적 처리 기술을 구현할 수 있었다는 점에서 큰 성과임에는 틀림없다.

그러나 국어의 정보 처리를 위해서는 형태소 분석을 넘어서 구문 분석과 의미 분석을 위한 프로그램 개발이 시급하다. 이것은 앞서 살펴본 기초 자료 구축 분야의 분석 코퍼스를 구축하는 것과 연계되는 것이며, 아울러 구문 분석기와 의미 분석기를 만들기 위한 구체적인 방법론에 대한 이론적 연구와도 연계된다. 또한 이런 프로그램들은 기초 자료로부터 다양한 언어 정보를 추출하고자 하는 국어학 연구자뿐만 아니라 전산학이나 기타 연계 학문의 연구자들도 쉽게 이용할 수 있도록 인터넷 웹 상에서 접근할 수 있는 사용자 인터페이스(user-interface)를 보다 효율적으로 만드는 것이 필요하다. 아울러 지금까지 '세종 계획'에서 구축된 다양한 기초 자료들을 좀 더 적극적으로 배포하고 활용도를 높이기 위한 웹 서비스의 품질을 높이는 것도 필요할 것이다.

정보 검색이나 맞춤법 검사기, 음성이나 문자 인식, 나아가 문서 요약 시스템과 같은 실용적 프로그램이 완성도가 높아져야 할 것이다. 이러한 프로그램들은 국어학적인 연구의 성과가 직접적으로 활용되는 부문일 뿐만 아니라, 일반인들의 언어생활에도 직접적으로 도움이 된다. 현재 상용화되어 있는 인터넷 검색 프로그램이나 다양한 종류의 문서 편집기에 포함되어 있는 프로그램들의 수준은 아직 만족할 만하지 못하다. 정보 검색 분야에서 단어를 넘어 문장의 검색이 가능하기 위해서나 단순한 철자 교정이 아닌 문장 단위 이상의 교정이 가능하기 위해서는 기본적으로 국어 구문 분석기가 필요하므로, 그에 대한 이론적 연구와 더불어 그것을 구현할 수 있는 장치를 만들기 위한 시도들이 계속되어야 한다. 문서 요약 시스템 역시 기존의 문서에서 문장을 발췌하여 단순 조합하여 제시하는 추

출 요약(extract) 수준을 넘어서 컴퓨터가 문서의 내용을 전체적으로 이해하고 새로운 요약문을 생성해 내는 생성 요약(abstract)이 가능한 단계로 나아가야 할 것이다. 이를 위해서는 역시 언어 정보를 총체적으로 분석할 수 있는 방법론이 필요한데, 형태소 분석과 구문 분석은 물론 의미 분석과 담화 분석 능력까지 갖춘 기술력이 필요하다. 물론 이러한 수준에 다다르는 것은 인공 지능(AI; artificial intelligence) 개발에 비견될 정도로 어려운 것이겠지만, 국어 정보화 영역에서 뒷받침해야 할 중요한 부분임에 틀림없다.

이 외에도 음성 공학이나 대화 시스템, 기계 번역 등의 분야 역시 지금까지 구축된 기초 자료를 활용하여 보다 향상된 프로그램을 만들어낼 수 있는 영역이다. 음성 공학과 기계 번역은 국어학적 측면에서의 자연 언어 처리 기술과 전산학적 측면에서의 다양한 기술이 결합되어야만 하는 분야로서, 우리들의 일상생활에 밀접히 연관될 수 있는 분야이기도 하다.

우선 음성 공학의 영역에서 인간의 말소리를 컴퓨터가 인식하거나 인간의 말소리를 컴퓨터가 합성해 내기 위해서는 기본적으로 국어의 음성에 대한 이론적 기반과 물리적 소리를 이용성 높게 디지털 자료로 구축할 수 있는 전산학적 기술력이 결합되어야 한다. 이러한 음성 공학의 성과는 현재 다양한 방식으로 이미 우리 생활에 들어와 있다. 가령 휴대폰의 문자 메시지나 웹상의 신문을 음성으로 읽어주는 장치(TTS; text-to-speech)라든가 특정 기업의 ARS 시스템, 또는 누군가의 음성으로부터 그 사람의 신원이나 내용을 식별할 수 있는 화자 인식(speaker recognition) 시스템 등은 이미 현재에도 초보적인 수준에서 활용되고 있는 시스템들이다. 미래 정보화 사회에서 음성 언어의 정보화를 활용하는 범위나 비중은 문자 언어의 그것과 대등하거나 혹은 더 커질 가능성이 높다는 점을 감안하면, 음성 공학의 미래를 대비할 국어 연구가 계속되어야 할 것이다.

기계 번역의 영역에서는 현재까지, 두 언어의 예제를 병렬적으로 구축

한 병렬 코퍼스(parallel corpus)를 이용한 예제 기반 방법론이나 병렬 코퍼스로부터 통계 정보를 추출하여 번역하는 통계 기반 방법론, 그리고 두 언어의 언어 정보와 특성에 대한 이론적 연구를 바탕으로 각각의 분석 과정과 생성 과정을 거치는 규칙 기반 방법론 등에 따라 몇몇 기계 번역 시스템이 개발되어 왔다. 문학 작품과 같은 고차원적인 텍스트가 아니라 일반인들의 생활 속에서 접하게 되는 일반 텍스트를 대상으로 하는 번역에 있어서는, 기계 번역 시스템이 상당한 도움이 될 수 있다. 그러나 현재까지 개발된 기계 번역기는 번역의 수준이 아직 실용성에 미치지 못하고 있는 실정이다. 모든 종류의 텍스트를 대상으로 완성도 높은 범용 기계 번역 시스템 개발은 앞으로 많은 연구 개발 기간이 필요하지만, 예약이나 구매 시스템, 혹은 날씨나 주식 등의 간단한 정보 제공 시스템 등으로 주어진 번역 상황을 한정한다면 다양한 체제에 따라 맞춤형 기계 번역 시스템을 개발하는 것은 비교적 접근하기가 나을 것이다. 또한 여기에서 축적된 연구 개발의 성과는 범용 시스템의 발전에 크게 기여할 수 있는 것이다.

이러한 응용 프로그램을 개발함에 있어, 이제까지는 거의가 전산학 전공자들이 전담하고 국어학/언어학 전공자들은 아주 소극적인 보조 역할을 해 왔을 뿐이다. 그러나 자연 언어를 기반으로 하는 전산화는 그것이 기초 자료를 마련하는 것이든 응용 프로그램을 개발하는 것이든 사실상 언어 분석력이 그 작품의 성취도를 높이는 데에 훨씬 더 필요한 요소라고 해도 과언이 아니다. 이는 필자들이 영-한, 한-영 기계 번역 프로그램 개발에 참여하면서 실감했던 사실이다. 어떠한 C-언어로 어떠한 전산화 엔진을 만드는가도 중요하지만, 대상 언어의 복잡하고 민감한 구조를 얼마나 정확하고 효율적으로 분석하여 체계화하는가 하는 연구 작업은 그 프로그램의 품질에 결정적인 역할을 하는 것이다. 다만 그 경우에 국어학자들이 전산의 메커니즘에 대한 충실한 인식이 전제되어야 하는데, 이제는 전산에 대해 이해가 깊은 국어학자들이 조금씩 늘어나고 있어 다행한

일이다. 앞으로 이를 감당할 수 있는 국어학자들을 많이 키우고, 언어를 전산화하여 이루어지는 각종 응용 프로그램의 개발에 국어학자들이 적극적으로 참여해야 할 것이다.

이상에서 살펴본 바와 같이 언어 정보화의 연구 성과를 활용하여 개발되고 있는 응용 프로그램들은 대단히 다양하다. 이들은 모두 지금까지 축적되어 온 국어 정보화 사업의 기초 결과를 그 토대로 하고 있으며, 앞으로는 국어 정보화 사업의 한 축을 담당하게 될 것이다. 그것은 미래에 전개될 정보, 통신 기기의 발전과 그에 따른 다양한 매체 환경의 변화에 발맞추기 위한 필수적인 요건이기도 하다.

디지털 시대를 살아가며 언어 전산화의 활용이 계속 늘어나게 되는 오늘날, 그러나 한국어의 전산화는 국내에서만 이루어지고 있지 않다. 외국의 거대 기업에서 크나큰 물량 공세로써 한국어의 전산화가 진행되고 있는 것이다. 한 예로 한국어의 기계 번역 프로그램 개발이 해외의 거대 기업에 의해 추진되고 있다. 언어 전산화의 꽃이라고 불리는 기계 번역은 그 개발이 어려운 만큼 언어 정보화에서 엄청나게 많은 부수적인 정보 요인들을 제공한다. 이것은 다만 상업적인 이윤 수준의 문제가 아니라, 자칫 한국어의 전산 정보화 체계가 두고두고 몇몇 외국 기업에 종속되는 문제를 야기할 수도 있다. 한국어를 쓰는 언중들이 한국어의 정보화를 주도적으로 이끌어 가는 것은, 민족주의나 자존심을 넘어서 그 자체로서 당위성을 갖는 현상이다. 우리가 국어 전산화에 더욱 힘을 기울여야 하는 이유가 여기에도 있는 것이다.

3.3. 기반 이론 연구

앞서 살펴본 국어 정보 처리를 이용한 응용 프로그램을 개발하기 위해서는 언어의 기초 자료를 모으는 것도 중요하지만, 그와 더불어 대규모의 언어 정보를 적절하게 분석할 수 있는 국어학적 기반 이론이 충실히 연

구되어야 한다. 언어 데이터베이스로부터 규칙을 만들어 나가는 귀납적 방법론만큼이나, 연구의 전체적인 방향과 흐름을 조율해 나갈 연역적 방법론 역시 필수적인 까닭이다.

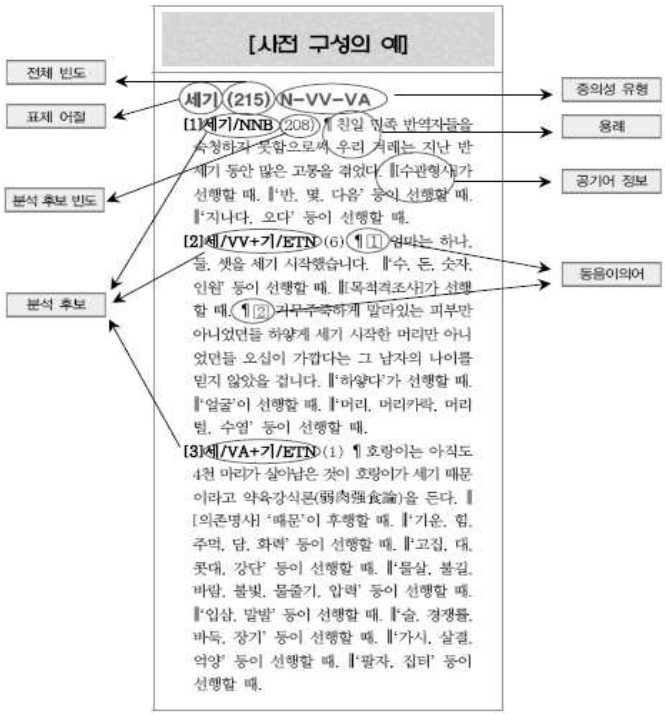
국어 정보화 과정이 기본적으로 컴퓨터를 매개로 하여 이루어지고 그로 인해 모든 언어 정보의 전산적 처리 절차가 수반되기 때문에, 전산학과 같은 이공계 중심의 연구자들이 국어 정보화 연구나 작업에 먼저 관심을 기울여 왔다. 이에 비해 전산 분야에 취약한 국어학자들은 소수의 연구자들만이 뒤늦게 그것도 국어 정보화 가운데 아주 일부분에 한하여 참여하기 시작하였다. 그러나 정보화의 대상 자체가 언어이며, 단순한 자료 구축 단계를 넘어서서 새로운 응용 영역의 질적 성장을 위해서는 언어 자체에 대한 연구가 기반이 되지 않으면 안 된다. 국어 정보화의 영역에서 보다 효율적인 연구 방법론들이 개발되기 위해서는 우리의 말과 글을 체계적으로 분석하는 국어학의 이론 연구들이 두텁고 폭넓게 축적되어야 하고, 따라서 국어학자들의 관심과 참여가 더욱 증대되어야 할 것이다.

이를 위해서는 우선 국어학 각 분야의 통합적 연구가 대폭 늘어나야 한다. 지금까지 국어학 연구는 대체적으로 음성, 음운, 형태, 구문 등 국어의 각 언어 단위에 따라 세분화되어 왔고, 그것이 의미와 화용 등 국어의 내용적 측면과 결합하여 기술되어 왔다. 그러나 앞서 서술한 바와 같이 국어의 정보화는 근본적으로 언어의 모든 측면이 총체적으로 결합되어야만 가능한 것이므로 국어 정보 처리를 위한 이론적 기반이 되는 국어학 연구 역시 모든 언어 단위에 대한 연구가 총체적으로 이루어질 필요가 있다.

아울러 현재까지 축적되어 온 국어에 대한 이론적 연구를 전산화하여 정보 처리의 시스템으로 구축하고자 했던 지금까지의 과정이 국어학의 전산화적 연구라면, 앞으로는 필요에 따라 국어 정보의 전산화를 위한 국어학 연구의 필요성도 늘어나고 있다. 하나의 규칙으로는 설명되기 어렵

고 수많은 예외와 변이형을 가지고 있는 자연 언어의 특성상 그것을 이론적으로 연구할 때에는 서로 다른 분석 체계와 이견이 존재할 수 있다.

그러나 언어학자의 관점에서는 매우 중요하고 심각한 문제가 전산적 처리 과정에서는 간단한 규칙으로 처리될 수도 있고, 반대로 언어학자의 관점에서는 그다지 중요하지 않은 문제가 여러 응용 프로그램 시스템 내에서는 반드시 선결되어야 할 부분이 될 수도 있다. 전산학 부분에서는 응용 프로그램 개발에 직접적으로 이용할 수 있는 국어 자료에 대한 요구가 절실하고, 이를 뒷받침해 줄 수 있는 국어학의 연구 성과 역시 요구되고 있다.



이런 측면에서 보면 사전 편찬과 같은 국어학의 전통적인 연구 영역도

정보 처리 시스템을 고려하여 그 체제와 편성이 달라질 수 있다. 한 예로, 위의 「한국어 중의 어절 사전」(2008)에서는 표제어 선정에서부터 내용 구성에 이르기까지 국어 정보 처리의 관점에서 기술되었다. 중의 어절이란 하나의 어절이 형태소 분석 과정에서 둘 이상의 분석 후보를 가지는 경우를 말하는데, 예를 들어 '가는'이라는 어절은 '학교에 가는 사람'에서처럼 동사 '가다'의 활용형일 수도 있고, '밭을 가는 사람'에서처럼 동사 '갈다'의 활용형일 수도 있으며, '가는 팔과 다리'에서처럼 형용사 '가늘다'의 활용형일 수도 있다. 이러한 중의 어절은 국어 모국어 화자에게는 문맥을 통해 의미를 쉽게 파악할 수 있고, 따라서 이러한 중의 어절만을 모아 놓은 자료가 반드시 필요하다고 볼 수 없다. 하지만 형태소 분석기 등을 만들기 위한 정보 처리 과정에서는 이러한 중의 어절들이 우리 국어에 어떤 양상으로 분포하고 있으며, 각각의 중의 어절이 지닌 의미적 중의성을 해소하기 위해서는 어떤 조건들이 필요한지를 파악하는 것이 매우 중요한 문제가 된다. 이에 따라 위 사전에서는 각 중의 어절의 전체 빈도와 중의성 유형 및 분석 후보 목록과 분석 후보별 빈도 정도와 같은 계량적 정보와 함께 각 분석 후보별 공기어 정보를 기술함으로써 이후 어떤 정보 처리 과정에서든 중의 어절 분석을 위한 기초 정보를 제공하고 있는 것이다.

이런 예와 같이 전산화와 정보 처리 과정에서 필요로 하는 자료와 정보가 무엇인지, 그리고 그것이 어떤 형식으로 구축되어 있어야 하는지 등을 종합적으로 염두에 둔 국어 연구가 바로 국어 정보화 사업을 위한 기반 연구라고 할 수 있다. 앞으로 이러한 연구가 크게 활발해져야 할 것이다.

이러한 연구들이 실제로 많은 연구 단위에서 실행되고 구체적인 성과로 이어지기 위해서는 국어학을 넘어서 여러 학문 영역 간 협력 시스템을 구축하는 것이 매우 중요하다. 코퍼스나 전자 사전 등 기초 자료 구축에 집중해 온 국어학 영역과 기타 응용 프로그램 개발에 노력해 온 전산

학 영역의 분업화가 이제는 서로의 연구 성과를 공유하고 서로에게 도움을 주는 상호 상승효과를 추구해야 할 것이다.

4. 마무리

지금까지 국어 정보화 사업의 성과를 개관적으로 살펴보고 국어 정보화 사업의 미래와 전망에 대해 기초 자료 구축 문제, 응용 프로그램 개발 문제, 기반 이론 연구 문제로 나누어 제시해 보았다. 국어 정보화 사업은 어느 특정한 분야에 국한되어 시행될 수 있는 것이 아니므로 여러 분야가 균형있게 발전하는 장기적인 관점에서의 접근이 필요하다. 또한 국어학이나 전산학 연구자들은 정보를 공유하며 연구와 개발을 함께 해 나가야 할 것이다. 이 문제에 비교적 소홀했고 이 과정에서 소외되어 온 국어학 연구가 사실상 국어 전산화에 결정적으로 중요함을, 국어학 연구자나 이와 관련하는 지원 기관들은 절실하게 인식해야 할 것이다.

하루가 다르게 발전해 가는 현대 정보 사회의 기술력을 감안하면 국어의 정보화라는 것은 끊임없이 업데이트되는 것이 필수적이고, 이를 위해서는 방대한 양의 국어 정보를 자료로 구축하고 활용할 수 있는 능력을 가진 연구자들을 양성해 내는 것도 빼 놓을 수 없는 장기적 과제이다. 전산 기초 자료를 잘 구축해 놓았다고 하더라도 다른 분야에서 그것을 활용하지 않는다면 그저 대규모의 데이터베이스로 존재할 뿐이다. 이는 응용프로그램의 개발에서도 마찬가지이다. 지금까지의 연구 성과를 적극적으로 활용할 수 있는 방안도 모색해야 할 것이다. 이를 위해서는 현재까지의 성과를 체계적으로 관리하고, 많은 연구자들이 손쉽게 자료나 도구에 접근할 수 있는 환경을 만들어 나가는 것 역시 중요한 일이다. 이러한 노력들이 적극적으로 이어질 때 국어 정보화의 미래가 존재할 수 있을 것이며, 미래 사회에서도 여전히 우리말과 글의 위상과 효용성이 담보될 수 있을 것이다.

참고 문헌

강범모(2003), 『언어, 컴퓨터, 코퍼스 언어학』, 고려대학교 출판부.

강승식(1999), 『한국어 정보처리와 정보검색』, 한성대학교 정보전산학부.

강승식(2002), 『한국어 형태소 분석과 정보 검색』, 홍릉과학출판사.

고창수(1999), 『한국어와 인공지능』, 태학사.

고창수(2002), 『자질연산문법이론』, 월인.

김재인(2002), "음성공학의 현재와 미래", 『음성 언어 자료와 국어 연구』, 월인.

김흥규, 「21세기 세종 계획 국어 기초 자료 구축 분과」, 중간보고서. 2005.

김흥규·강범모, 『한국어 형태소 및 어휘사용 빈도의 분석』, 고려대학교 민족문화연구원. 2000.

남경완·유혜원, "동사-형용사 중의성 해소 연구: 한국어 정보 처리를 위하여", 「어문연구」 131. 2006.

남경완·최정혜, "중의 어절의 문법 범주별 유형 연구", 「민족문화연구」 45. 2006.

남윤진(2002), "국어연구와 빈도 정보", 『한국어와 정보화』, 태학사.

배재학, "언어학적 방법론을 취하는 자동문서 요약에 대한 연구", 「공학연구논문집」(울산대) 29-2. 1998.

서상규(2002), "한국어 말뭉치 구축과 과제", 『한국어와 정보화』, 태학사.

서상규·한영균(1999), 『국어 정보학 입문』, 태학사.

서정연, 「유사 적합성 피드백 기반의 문서 요약 기법을 이용한 정보검색 요약문의 효과적인 생성」, 서강대학교 석사학위논문. 2007.

시스템공학연구소, 「영한 한영 텍스트 자동번역 기술 개발」, 정보통신부. 1996.

신지영, "음성 자료에 기초한 국어 음운론 연구의 과제와 전망", 「한국어학」 17. 2002.

심광섭(2001), “문서요약시스템 개발”, J. Basic Sci. Sungshin Vol.19, 89-103.
안홍국, 「유사 적합성 피드백 기반의 문서 요약 기법을 이용한 정보검색 요약문의 효과적인 생성」, 서강대학교 석사학위논문. 2007.
연규동·박진호·최운호(2003), 『인문학을 위한 컴퓨터』, 태학사.
유혜원(2004), 『한국어 정보 처리의 이론과 실제』, 제이앤씨.
유혜원·남경완·홍종선, “한국어의 형태론적 중의 어절 사전 구축과 표제어 선정.” 「한국어학」 31. 2006.
이경호·남경완, “한국어의 형태론적 중의 어절 사전 구축의 방법과 실제”, 「우리어문연구」 28. 2007.
이교윤·윤근수·박용욱, “문서와 각 문장의 유사도를 이용한 문서 자동 요약 시스템 구현”, 「연구논문집」(울산과학대학) 26-1. 1999.
이동혁(2007), 『한국어 관용 표현의 정보화와 전산 처리』, 역락.
이유리, 「수사구조를 이용한 텍스트 자동요약」, 한국과학기술원 석사학위논문. 1999.
최호철·이정식, “자연 언어 처리를 위한 전자 사전 구축 방안”, 「어문논집」37. 1998.
최호철·한정한·오장근, 「다국어 기계번역을 위한 중간언어 모형과 방법론 연구[2]」, 고려대학교 민족문화연구원. 2005.
한국어정보처리연구소(2001), 『C로 구현한 한글 코드 시스템 프로그래밍 가이드』, 도서출판 골드.
한정한·남경완·유혜원·이동혁(2007), 『한국어 정보 처리 입문』, 커뮤니케이션북스.
홍윤표 외(2002), 『한국어와 정보화』, 태학사.
홍재성 외, 「21세기 세종 계획 보고서」, 문화관광부. 1999.
홍재성, 「21세기 세종 계획-전자사전 개발」, 문화관광부 연구보고서. 1999.

홍종선, “한국어 기계 번역에서의 중의성 처리 연구”, 「국어학」 50. 2007.
홍종선·남경완·유혜원·이동혁·황화상(2008), 「한국어 중의어절 사전」, 태학사.
황도삼 역(2002), 『자연언어처리의 응용』, 두양사, (Hozumi Tanaka (1999), Natural Language Processing and Its Application.)
황화상(2004), 『한국어 전산 형태론』, 도서출판 월인.
황화상(2005), 『한국어와 정보』, 박이정.