사전 편찬과 국어 정보화의 과제

— 국어 발전 기본 계획을 중심으로 —

남길임 • 경북대학교 국어국문학과 교수

1. 서론

국어학 연구의 가시적인 결과물로서의 사전 편찬, 사전의 재활용, 국어 자원 구축 및 활용에 대한 연구가 최근처럼 활발하게 이루어진 적은 없는 듯하다. 정보화, 세계화를 특성으로 하는 현대 사회는 언어 정보의 효율적인 축적과 활용, 이를 위한 자연 언어 처리 기술을 요구하고 있으며, 국어 자원에 대한 연구는 국어학뿐만 아니라 전산학, 정보과학 등 언어 정보 및 정보 기술을 연구 대상으로 하는 제반 학문의 핵심적인 과제로 연구되고 있다. 대표적인 예로 국어학 분야의 말뭉치를 활용한 사전 편찬, 21세기 세종계획(1998~2007)의 말뭉치 및 전자사전 구축 사업과 공학 및 정보 기술 분야의 온톨로지 구축, 시멘틱웹 등 어휘망(word net) 구축 사업 등을 들 수 있다.

한편 최근 한국어 학습자의 양적 증대와 수요자층의 다양화로 인해 한국어 교육을 위한 언어 자원의 축적 및 활용 역시 국어 자원을 기반으로

하는 연구의 새로운 분야로 자리 잡았다. 이에 따라서 학습자 사전 및 학습자 말뭉치(Learner corpus)[1]의 구축, 한국어 교육을 위한 각종 말뭉치 구축 및 활용, 컴퓨터를 이용한 언어 학습(computer-assisted language Learning) 시스템 개발 등도 중요한 시대적 과제로 부각되고 있다.

이 글은 올해부터 시행되는 국어 발전 기본 계획(안)의 추진 과제를 중심으로 이러한 국어 자원의 축적과 활용의 현황을 분석하고 향후 지향 방향을 살펴보는 데 그 목적이 있다. 국어 자원의 축적과 활용은 (1) 사전 편찬 및 활용, (2) 말뭉치, 어휘망 등 국어 자원의 축적 및 활용 부문으로 나누어 살펴볼 수 있는데, 전자는 국어 발전 기본 계획(안) 중, '다국어 지원 한국어 학습용 웹사전 편찬(3대 중점 추진 과제)', '<표준국어대사전>의 정비 및 맞춤형 사전 편찬(10대 추진 과제)'과 관련되며, 후자는 '국어 정보망 구축과 통합 정보 시스템 운영(10대 추진 과제)'과 관련된다. 한 언어의 충실하고 체계적인 정보체로서의 사전의 정보 항목은 말뭉치 자동 주석 시스템과 사전 정보를 재활용한 어휘망의 구축에 핵심적인 요소이며, 반대로 질 좋은 주석 말뭉치와 어휘망의 정보는 사전 정보의 일관성, 체계성에 기여한다는 점에서 이 두 부문은 상호 보완적 관계에 있다.

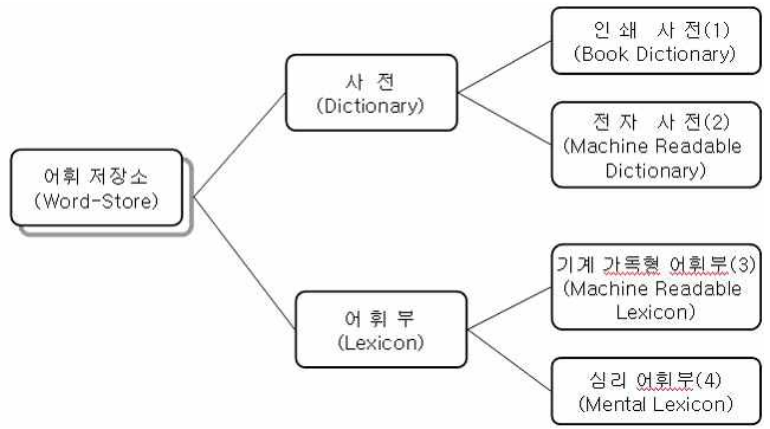
이 글은 다음과 같이 구성된다. 2장에서는 사전 편찬, 말뭉치 구축 등 국내외 언어 자원 구축 및 활용의 현황을 살펴볼 것이다. 3장에서는 2장의 논의를 기반으로 국어 발전 기본 계획(안)의 세 가지 관련 과제를 사전 편찬과 국어 자원 정보화의 관점에서 분석하고 향후 발전 방향을 점검하고자 한다.

1) '학습자 말뭉치'는 학습자에 의해 생산된 말뭉치(Corpus by Learers)로 학습자 오류 분석, 학습자의 중간언어(interlanguage) 분석, 대조 분석 등의 연구 분야에서 널리 활용되어 오고 있을 뿐만 아니라 최근에는 학습자를 위한 교재 및 사전 편찬 등 실제적인 활용 결과물들이 풍부하게 생산되고 있다. 학습자 말뭉치의 최근 연구 동향에 대해서는 Guy Aston et al(2004: 1~10)에 상세하게 소개되어 있다.

2. 사전 편찬과 국어 자원 정보화의 현황

2.1. 말뭉치, 전자 매체의 도입과 사전 편찬

전자 매체와 자연어 처리의 개념이 등장한 이후 '사전'이라는 용어는 매체와 용도, 어휘 정보 표상의 방식에 따라 매우 다의적인 의미로 쓰이게 되었다. 넓은 의미에서 어휘 저장소로서의 사전은 지금까지의 인쇄 사전뿐만 아니라 시디롬, 온라인 사전, 개인용 전자 사전(PED: Personal Electronic Dictionary)으로서의 전자 사전의 형태를 포함하며, 더 나아가서 자연 언어 처리 과정에서 사용되는 어휘 저장소로서의 기계 가독형 어휘부(Machine-Readable Lexicon)를 지칭하는 용어로 쓰이기도 한다.



<그림 1> 어휘 저장소의 유형(A typology of word-stores)

위 그림은 Handke(1995:49)의 단어 저장소(word-stores)의 유형 분류를 일부 가져온 것으로 여기서 논의할 사전 편찬은 (1), (2)의 인쇄 사전과 기계 가독형 사전으로서의 전자 사전에 해당하며 기계 가독형 어휘부로서의 사전 (3)은 2.2.에서 다루기로 한다.

새로운 매체의 등장 이후 국내 사전 편찬은 새로운 전기를 맞고 있음이 틀림없으나 본격적으로 매체와 전자 환경의 장점을 고려한 사전 편찬이 이루어지고 있다고는 말하기 어려울 듯하다. 사전 편찬에 컴퓨터가 사용되었을 때, 사전 편찬의 발달 단계를 다음의 3단계로 구분할 때, 국내 사전 편찬의 수준은 2단계에서 3단계로의 진입 단계에 위치해 있다.[2)]

1단계. 컴퓨터를 활용하여 종이 사전을 편찬하는 단계
2단계. 기존의 인쇄 사전을 전자 매체로 전환하는 단계
3단계. 전자 환경을 고려해서 새로 고안해 낸 전자 사전들

국내 최초로 말뭉치를 활용하여 표제어를 선정하고 예문을 제시한 <연세한국어사전>(1989)이나 현재 국어원 홈페이지에서 서비스되고 있는 <표준국어대사전>(이하 <표준>)의 경우를 살펴보면 국내 사전의 수준은 2단계는 완료 상태에 있으나 본격적으로 전자 환경을 고려하여 사전 편찬을 시도한 적은 없으므로 3단계에 이르지는 못했음을 알 수 있다. 물론 <표준>의 경우 종이 사전에 없는 음성 정보가 추가되어 있기는 하지만 웹 환경의 장점을 살린 정보를 구현하고 있다고 하기 어렵다. 즉, 다양한 검색 경로, 하이퍼링크나 상호 참조 기능 등이 구현되어 있지 않을 뿐만 아니라 지면의 제약에서 자유로운 전자 환경의 이점을 충분히 활용하고 있지 못하다. 전자 매체의 장점을 최대한 활용할 수 있는 정보로는 말뭉치와 연계함으로써 제공될 수 있는 풍부한 용례 정보, 용언의 경우 활용 형태 정보, 접사나 어근의 경우 해당 접사의 조어 정보 등을 들 수 있다. 물론 말뭉치 용례 정보의 경우, 정제되고 효율적으로 분류된 말뭉치를 활용할 경우 문어와 구어 등 하위 장르별 제시도 가능할 것이다.[3)] 현시점에서 전자 환경에서만 제시될 수 있는 사전 정보 항목의 유형에 대한 연구가 필요하다고 할 수 있다.

한편 전자 매체의 등장이 곧바로 인쇄 사전의 퇴조로 의미하지 않는다는 것을 밝혀 둘 필요가 있다. 하트만(Hartmann)이 언급한 바 있듯이 인

2) 위 3단계의 구분은 Cerquiglini의 견해로 Provust(2000:188)에서 재인용.
3) Hacken, P. T.(2006).

쇄 사전은 교육 사전(pedagogical dictionary)의 기능을 강화함으로써 사전 시장에서 그 영역을 분명히 할 것으로 예상된다. 최근 초등학교 교육 과정에 사전 교육이 등장함에 따라 인쇄 사전 형태의 초등 사전에 대한 수요는 꾸준한 상황이다. 또한 외국인 학습자를 위한 사전류로 <외국인을 위한 한국어 학습 사전>(2006)이나 국외의 학습 수준별 외국어 학습 사전, 교육적 기능을 강화한 Longman Dictionary of common errors(1996), Cobuid English Usage(2004) 등과 같이 교육적 기능을 특화한 실제 사전이 사용자들의 관심을 끌고 있다. 하지만 교육 사전에 있어서 오랜 전통을 가진 영어권 사전에 비해서 국내의 경우는 각 언어권별 이중 언어 사전(bilingual dictionary)이나 사전 활용 목적에 따른 학습 사전류, 즉 능동 사전(active dictionary)과 수동 사전(passive dictionary), 언어 학습 수준별 사전 등에 대한 논의는 미진한 실정이다.

지금까지 논의를 요약하면, 국내 사전 편찬의 현황은 매체를 고려한 사전 편찬과 교육적 기능을 강화한 사전 편찬의 방향으로 도약이 필요한 단계에 있다. 이는 한편으로 지금까지 주로 학문적 관점에서 평가되어온 사전이 그 평가의 영역을 넓혀 실제 사용자의 평가와 요구를 반영해야 할 필요도 있음을 의미하는 것이기도 하다.

2.2. 국어 자원 정보화의 현황

국어 자원 정보화는 기초 자료로서의 말뭉치 등 국어 자원의 구축을 핵심으로 하지만 말뭉치의 검색, 활용을 위한 프로그램 개발, 국어 자원의 응용·활용을 위한 연구 및 실제를 포함하는 광범위한 개념이다. 여기서는 (1) 말뭉치 등 국어 자원 구축·가공 부문, (2) 검색 도구 등 지원 도구 개발 부문, (3) 응용·활용 부분 등 세 가지 부문으로 구분하여 살펴보기로 한다.

우선, 국어 자원의 구축·가공 부문부터 보면, 80년대 후반 사전 편찬을 위한 자료 확보를 위해 구축되기 시작한 말뭉치 구축 사업이 21세기 세종계획(1998~2007)에 이름에 따라 기초 자료로서의 국어 자원 구축은 단순한 구축 단계를 넘어서 가공을 통한 응용·활용 단계에 올라 있다고

할 수 있다. 말뭉치, 전자 사전 등 기초 자료 구축의 결과물들이 양적으로 확충되었을 뿐만 아니라 형태 분석 말뭉치, 구문 분석 말뭉치 등 다양한 주석 말뭉치의 구축과 관련 기초 연구가 진행 중에 있으며, 비록 그 양은 적지만 역사 말뭉치, 병렬 말뭉치 등도 구축되었다. 이와 더불어 최근 몇 년 동안 학계 및 연구소 단위로 어휘 의미망 구축 사업이 활발하게 전개되고 있는 것이 두드러진 경향이다. 특히 어휘 의미망 구축 사업은 전자 사전이나 인쇄 사전의 전자 데이터베이스를 재활용하고 기존의 국내외 어휘망, 어휘 의미 부류를 적극적으로 활용하고 있다는 점에서 향후 사전 편찬 및 국어학 연구 방향에 시사하는 바가 크다.

한편, 이러한 기초 국어 자원을 활용하기 위한 검색 도구 등 지원 도구 개발 부문이나 응용·활용의 부문은 향후 장·단기적인 연구가 필요한 부분이다. 현재 배포되어 연구자들 사이에서 활용되고 있는 말뭉치 검색 프로그램 '글잡이'나 '지능형 형태소 분석기'는 연구자들이라면 누구나 불편을 호소하리만큼 '정확성'과 '신속성'의 측면에서 그리 만족스러운 상태는 아니다. 또한 검색 대상 말뭉치 역시 일부 말뭉치 유형에만 제한되어 있어서 지금까지 개발된 모든 말뭉치를 통합적으로 검색할 수 있고 연구자가 목적에 맞게 구성하여 활용할 수 있는 새로운 시스템 개발이 필요하다. 이 부분은 현재 국어원에서 계획 중인 국어 정보 통합 관리 및 검색 시스템 개발(07 완료 예정)을 통해 만족스러운 결과를 얻기를 기대한다.

또한 이미 구축되어 있는 국어 자원의 응용·활용을 위해서 말뭉치 등 기초 자료의 정제와 보완, 활용 방법론에 대한 연구도 필요한데 이는 자연 언어 처리, 언어 교육, 국어학 분야 등 각 부문별, 장·단기적 연구가 필요하다. 대표적으로 한국어 교육 분야에서의 말뭉치 및 검색 도구의 활용 방안의 연구 범위와 실례를 제시하면 <표 1>과 같다.

실제로 올해부터 추진 중인 국어 발전 기본 계획 세부 과제는 위 연구의 일정 부분을 고려하고 있으며, 이에 따른 향후 연구의 확장이 기대된다. 아래 3장에서는 사전 편찬과 국어 자원 정보화와 관련한 구체적인 관련 과제를 중심으로 향후 발전 방향에 대한 제안을 하고자 한다.

<표 1> 말뭉치 활용 방안의 예: 한국어 교육 분야

<한국어 교육을 위한 말뭉치의 활용 방안의 예>

1. 학습자를 위한 말뭉치(corpus for learners) 활용 연구
   예) 교실에서의 말뭉치 활용 연구, 말뭉치 활용 교재(corpus workbook) 개발, 학습자를 위한 온라인 말뭉치 활용 프로그램 개발
2. 학습자에 의해 생산된 말뭉치(corpus by learners) 활용 연구
   예) 학습자 오류 말뭉치를 활용한 사전 편찬 및 교재 편찬
3. 학습자와 말뭉치(corpus with learners)에 대한 연구
   예) 학습자의 말뭉치 활용 양상 연구, L1 학습자와 L2 학습자의 말뭉치 활용 양상 연구를 통한 언어 습득 비교 연구

3. 국어 발전 기본 계획 관련 과제에 대한 제안

3.1. 다국어 지원 한국어 학습용 웹사전 편찬

'다국어 지원 한국어 학습용 웹사전 편찬'은 국어 발전 기본 계획(안)의 3대 중점 추진 과제 중 하나로 사업의 예산이나 규모를 고려할 때 전체 계획 중 가장 중점적인 사업으로 판단된다. 전체 사업 기간(2007~2011) 동안 10개 언어권 대역사전 개발을 주요 목적으로 하는 이 사업은 다양한 언어권의 한국어 학습 수요에 부응하고 외국인 근로자나 국제결혼 이주 여성의 국내 정착을 도울 수 있는 시의 적절한 사업이 될 것이다. 무엇보다 개발 초기 단계부터 웹사전의 환경을 고려할 경우, 다양한 종류의 콘텐츠를 개발·확보할 수 있어서 참조물로서의 사전의 기능에 더하여 교육·학습 기능을 강화한 사전으로 기능할 수 있다. 다음에서 이 사업의 특성 및 전반적인 추진 과정에 대한 몇 가지 고려 사항을 제시하면 다음과 같다.

첫째, 이 사업은 시디롬이나 인쇄 사전을 보조적 결과물로 하되 온라인 사전 개발을 주요 목적으로 하므로, 지금까지의 사전 편찬 사업과는 매체적 관점에서 차별되는 특성을 가진다. 따라서 온라인 사전을 위한 별도의

정보 항목 구성 및 기술 지침이 필요하며, 시디롬이나 인쇄 사전으로 병행 출판하는 것을 고려하여 출판 결과물에 따른 정보의 양과 유형을 별도로 구조화하여 관리해야 할 것이다. 이를 위한 기초 연구로 한국어의 유형론적 특성을 반영한 전자 사전의 정보 항목 구성 및 기술 지침에 대한 연구와, 인쇄 사전·웹 기반 사전의 병행 출판을 염두에 둔 편집 프로그램 개발이 선행되어야 한다.

둘째, 사전의 규모 면에서 이 사전은 표제어 5만 어휘의 10개 언어권 대역사전이라는 방대한 사업이다. 표제어 5만은 학습자 사전이라는 사전의 내적 특성과 사전 편찬의 외적 환경, 즉 사업의 기간과 예산을 고려할 때 지나치게 많은 양이라 판단된다. 학습 사전의 표제어 규모를 정하는 것은 각종 기관에서 연구된 바 있는 한국어 교육 기본 어휘와 중요 어휘 목록, 한국어 교육용 말뭉치 및 균형 말뭉치의 어휘 빈도 목록 등의 분석 작업을 필요로 한다. 실제로 시판 중인 한국어 교재 총 32종 88권의 분석 결과[4] 총어휘 수는 24,293개로 나타났으며 이들 중 빈도 1인 34.5%에 해당되는 어휘 8,392를 제외하면 15,901개로 기타 생활 어휘와 문화 어휘를 합친다 하더라도 20,000 어휘를 넘지 않을 것으로 추정된다. 이 외에도 한국어 교육 기본 어휘 5,800여 개(국립국어원)의 목록이나 한국어 학습 사전의 표제어 적정 규모로 7,000 어휘(강현화: 2007), 5,000 어휘 내외(서상규: 2003)의 연구 결과 등을 참고할 만하다.

셋째, 사전의 특성 면에서 사전 편찬의 초기 단계부터 사전의 특성을 구체화할 필요가 있다는 점을 지적하고자 한다. 이 사전은 참조 기능과 교육 기능을 겸비한 사전을 지향하며, 이와 더불어 한국인 화자도 해당 언어권 언어를 참조할 수 있는 양 방향 사전(bidirectional dictionary)[5]으로서의 기능까지도 염두에 두고 있는 범용적 사전을 목적으로 하는 듯하다. 그러나 Hartmann and James(1998)에서도 밝히고 있듯이 양 방향 사전은 두 언어의 구조나 문화가 다르고 사용자의 용도가 다를 경우 실제

4) 이는 서상규(2003)의 연구 결과를 인용한 것이며 상세한 분석 자료 목록은 서상규(2003)를 참조.

5) 양 방향 사전(bidirectional dictionary): 이중 언어 사전의 한 유형으로 양쪽 언어 모두에서 대응어를 찾을 수 있는 양 방향성 사전. 일반적인 단일 방향 사전(monodirectional dictionary)과 대응되는 말(Hartmann and James: 1998).

효율적으로 쓰이기는 어렵다. 특히 두 언어의 구조나 문화가 다르고 사용자의 용도가 다를 경우, 예컨대 한국인 화자가 중국어 텍스트를 '이해'하기 위해 사전을 찾는 것과 중국인 화자가 한국어를 '표현'하기 위해 양 방향 사전을 찾는 것은 그 정보가 같을 수 없다. 용도에 따라 대역어 이상의 정보가 필요할 수도 있다. 이와 더불어 양 방향 사전이라고 했을 때 대역 정보 외에 어느 수준까지의 어휘 정보를 제시할 것인지, 표현 사전과 이해 사전으로서의 특성 중 어느 부분에 초점을 맞출 것인지에 대한 기초적 논의가 필요하다.

3.2. <표준국어대사전>의 정비 및 맞춤형 사전 편찬

'다국어 지원 한국어 학습용 웹사전 편찬' 사업과 함께 이 사업은 국어원에서 기획하고 있는 양대 사전 사업 중 하나로, <표준>의 정비 및 온라인 서비스와 맞춤형 사전 편찬 사업을 포함한다. 국내 최대 규모의 <표준>의 정비와 다양한 유형의 사전 편찬은 국어학 연구의 가시적인 결과물로서 가치를 지닐 뿐 아니라 민족 문화를 집대성한 문화의 산물이라는 점에서 문화적 의의를 가지며, 국민의 언어 생활의 요구와 수요에 부응한다는 실용적 가치를 가진다.

여기서는 <표준>의 정비 및 온라인 서비스 사업을 중심으로 살펴보기로 한다. <표준>의 정비는 정비 대상 및 방향에 따라 크게 두 가지 관점에서 논의될 수 있다. 하나는 웹서비스를 위한 사전 구조 개선 및 신어 표제어 내용 보완의 측면이고, 다른 하나는 기술 사전(descriptive dictionary)에 대비되는 개념으로서의 규범 사전(prescriptive dictionary)으로서의 표제어 정비 및 기술 내용의 보완과 관련된다. 후자와 관련해서는 민족 문화의 결정체로서의 사전이 방언, 구어 등 기타 언어 현실을 충실히 그리고 균형적으로 제시하는 기술 사전을 지향해야 한다는 점에서 수정·보완의 여지가 있다고 판단되나, 여기서는 지면의 관계상 전자에 대해서만 논의하기로 한다.

현재에도 <표준>은 국어원 홈페이지를 비롯한, 포털사이트, 개인용 전자 사전(PED) 등을 통해 보급되고 있으나, 본래 온라인 사전을 염두에 두

고 제작된 것은 아니어서 검색 경로의 다양성과 검색의 용이성, 정보의 양과 유형의 다양성, 정보의 확장성 등에 한계가 있다. 온라인 사전은 옥스퍼드 온라인 영어 사전(OED Online) 등을 참고해 볼 때 음성 정보뿐만 아니라 말뭉치와의 연동을 통해 풍부한 용례를 제공할 수 있으며, 어원·뜻풀이·활용 형태·관련 어휘로의 검색 및 구·문장 단위의 검색 등이 가능하다는 장점이 있다. 그러나 실제로 <표준>의 이용자들이 호소하는 어려움 중에는 활용 형태나 유사 어휘로 검색했을 때 해당 기본형이나 관련된 어휘가 검색되지 않는다는 점, 일상적 용례가 풍부하지 못한 점, 표제어 및 표제어 내 정보 항목의 배열 구조가 일반인들이 접근하기 용이하지 않다는 점 등이 있다.[6] 따라서 <표준>의 정비와 온라인 서비스를 위한 연구의 기초 작업으로 사용자 요구 사항을 분석할 필요가 있는 것으로 판단된다. 또한 사전 정보의 수정, 추가, 보완이 가능하도록 사전 데이터베이스 관리 구조 및 저장 구조를 개선하는 작업과 함께 정제된 균형 말뭉치와의 연동 시스템 개발 및 향후 추진될 어휘망과의 연계 작업 등도 필요할 것이다.

3.3. 국어 정보망 구축과 통합 정보 시스템 운영

이 사업은 국어 정보 통합 관리 체계 구축 및 운영, 한국어 어휘 의미망 구축, 국어 능력 향상 학습 시스템 개발, 한국어의 다양한 체험관 설립, 국어 전문 도서관 구축이라는 다소 성격이 다른 5개 하위 사업으로 구성된다. 여기서는 이 중에서 한두 가지에 대한 부분만 언급하기로 한다.

---

6) 이 외에도 화면을 통해 제시되는 정보의 가독성, 검색의 효율성 등이 언급될 수 있다. 몇 가지 예를 들면 우선 현재 국어원 홈페이지의 <표준>을 검색하면 해당 어휘가 동형어일 경우 최초의 화면이 활용 형태부터 제시되는 것을 볼 수 있다. <표준>의 1차적 사용자 대상이 외국인 화자가 아닌 것을 고려할 때 활용 정보가 처음 제시되는 것은 사용자의 검색에 아무런 도움을 주지 못한다. 불규칙 활용 형태 정보는 모국어 화자들에게는 별로 유용하지 않으며 동형어 표제어의 변별에 아무런 기여를 하지 못하기 때문이다. 이 외에 단어를 찾았을 때 어근의 형태가 먼저 제시되고 그 이하에서 정보를 찾아야 한다는 점 등도 문제이다. 우선적으로 가장 단순한 해결책은 표제항 내 정보 항목을 재배치함으로써 검색의 효율성을 높이는 방법이 있을 것이고 장기적으로는 표제항 및 사전 정보 항목 전반에 대한 보완이 지속적으로 이루어져야 할 것이다.

우선, 21세기 세종계획 등을 통해 양적인 측면에서 말뭉치, 전자 사전 등의 국어 자원이 어느 정도 확보되어 있고 말뭉치 주석에 대한 질적인 연구나 활용 방법론에 대한 논의가 활성화되고 있는 현 상황을 고려할 때 향후 국어 정보의 검색 시스템과 이를 활용한 다양한 실용적 응용·활용에 대한 논의는 시의 적절한 것으로 판단된다. 국어 정보 통합 관리 체계 구축을 통해 지금까지 각 기관이나 연구소별로 산발적으로 행해지고 있는 국어 자원 구축 사업에 대한 정보들을 종합적으로 제시하고 연구자들의 정보 이용을 증대시키며, 국어학적·실용적 활용 방안에 대한 연구를 활성화시킬 수 있을 것이다.

한편, 현시점에서는 지금까지 구축된 자원을 활용하는 실용적 접근 또는 새로운 사업도 필요하지만 동시에 이미 구축된 자료의 양적·질적 수준을 평가하고 이를 심화·확장하기 위한 논의도 필요하다. 즉, 기존 연구의 연속선상에서 이미 구축된 국어 자원의 균형적 구성을 위한 보완 작업과 다양한 차원의 주석 작업 확장, 전자 사전을 재활용한 어휘망 구축 가능성 등을 검토할 필요가 있다. 몇 가지 예를 들면, 문어 말뭉치의 경우는 형태 주석, 구문 주석, 부분적으로 이루어져 있으나 구어 말뭉치나 병렬 말뭉치는 형태 분석 말뭉치의 양이 아직 미미하며, 구문 주석 말뭉치 구축에 이르지 못하였다. 또한 학습자 오류 말뭉치의 경우도 문어에만 그 양이 집중되어 있어서 구어 학습자 말뭉치와 오류 분석을 위한 말뭉치 분석 도구에 대한 개발도 필요하다. 영미권의 경우 구어 말뭉치의 메타데이터 추출(Meta Data Extraction)이라는 전처리 작업을 통해 구어를 문어와 유사한 수준으로 정제함으로써 자동 구문 분석을 시도하고 있으며 학습자 오류 말뭉치를 활용한 교재 편찬, 언어 습득 연구에 대한 논의도 활발하다. 이와 더불어 70만 어휘의 기술을 포함하는 세종 전자 사전 역시 연구의 연속선상에서 지금까지 구축된 정보를 응용·활용하는 방안, 정보 기술 분야나 자연 언어 처리에 활용할 수 있는 방법론의 논의가 활성화되어야 할 것이다.

4. 결론

이 글에서는 국어 발전 기본 계획(안)을 중심으로 사전 편찬 및 국어 자원의 정보화의 현황과 과제에 대해 살펴보았다. 국어 발전 기본 계획(안)의 세 가지 과제는 올해로 완료되는 21세기 세종계획과 공학 분야 및 관련 업체에서 산발적으로 이루어지고 있는 온톨로지, 시멘틱웹 등 어휘망(word net) 구축, 한국어 교육을 위한 언어 자원 구축 등 국어 자원과 관련된 사업을 '국어 자원의 정보화'라는 전체적인 큰 틀 속에서 거시적으로 발전 계획을 제시한 것이다.

사전을 비롯하여 언어 자원에 대한 정보가 취약하면 현실 언어의 총체적인 기술이 불가능하고, 현실 언어에 대한 총체적인 기술 없이는 실용적인 시스템을 구현하는 데 필요한 다양한 언어 현상들을 체계적으로 처리하기가 어렵다.

21세기 세종계획(1998~2007)을 기획하고 시작한 지 10년이 지났고 그 기간 동안 정보 기술 분야나 자연 언어 처리 분야에는 새로운 연구와 방법론들이 쏟아지고 있다. '다국어 지원 한국어 학습용 웹사전 편찬', '<표준>의 정비 및 맞춤형 사전 편찬', '국어 정보망 구축과 통합 정보 시스템 운영 사업'을 통해 그동안의 사업 결과물들에 대한 접근성이 높아지고 대국민 홍보가 활성화되며, 그럼으로써 언어 자원에 대한 학문적·실용적 결과물들이 풍부해지기를 기대한다.

| 참고 문헌 |

강현화(2007), 인터넷과 어휘 의미망을 활용한 한국어 학습 사전 편찬의 현황과 과제, 한국어 어휘 의미망 구축과 사전 편찬 학술회의 발표 자료집, 국립국어원, 55~74쪽.

남길임(2005), 온라인 사전의 로그 파일(log file) 분석을 통한 사전 검색

양상 연구, “한국사전학” 5, 한국사전학회.
남길임(2007), 학습자 오류 말뭉치를 활용한 한국어 용법 사전의 편찬, 제25회 한말연구학회 전국 학술 대회 발표 자료집, 한말학회, 102～113쪽.
서상규(2003), ‘한국어 교육 기본 어휘와 학습 사전’, 제200회 개념 국제 학술 대회, 조선어연구회.
D. Stewart, S. Bernardin., & G. Aston.(2004), Introduction: Ten years of TaLC. in *Corpora and Language Learners*, Guy Aston et al, 2004, John Benjamins Publishing Company.
Guy Aston et al.(2004), *Corpora and Language Learners*, John Benjamins Publishing Company.
Hartmann, R. R. K., & James, G.(1998), *Dictionary of Lexicography*, Routledge, London and New York.
Hacken, P. T.(2006), Word Formation in an electronic Learners’ Dictionary: ELDIT, *International Journal of Lexicography 19*(3): 243～256.
Handke, J.(1995), *The Structure of the Lexicon: human versus machine*, Berlin: Mouton de Gruyter.
Provust, J.(2000), Colloquium report: Des dictionaires papier aux dictionnaires électroniques. VIIe Journée des dictionnaires (22mar 2000). *International Journal of Lexicography 13.3*: 187-93.
Yukio T.(2001), *Research on Dictionary Use in the Context of Foreign Language Learning*, Max Niemeyer Verlag Tübingen.

사전류

국립국어연구원 편(1999), “표준국어대사전”, 두산동아.
연세대학교 언어정보연구원 편(1998), “연세한국어사전”, 두산동아.

서상규・백봉자・강현화・김홍범・남길임・유현경・정희정・한송화

(2006), “외국인을 위한 한국어 학습 사전”, 신원프라임.

Turton, N. D.(1995), A*BC of Common Grammatical Errors*, Macmillan Heinemann.

Sinclair. J.(2004), *Cobuid English Usage Second Edition*, Collins Cobuild.

Turton, N. D., & Heaton J. B.(1996), *Longman Dictionary of common errors Second Edition*.

웹사이트

LinguisticData Consortium

https://project.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf