사전의 진화와 미래

배연경
국제영어대학원대학교 영어교재개발학과 교수

지난 30년간 사전이 타고 온 변화의 물살은 전에 없이 빠르고 거셌다. 1998년 유럽 사전학회의 국제 학술 대회에서 그레펜슈테트(Grefenstette, 1998:25)는 “3000년도에도 사전 만드는 사람이 있을까?(Will there be lexicographers in the year 3000?)”라는 발표문을 통해, 아무리 언어 통계 기술이 발전하여도 어휘의 의미를 분류하고 기술(記述)하는 일만은 적어도 앞으로 수백 년 동안은 사람의 몫으로 남아 있을 것이라고 전망하였는데, 그의 예측이 깨어지는 데는 채 한 세대도 걸리지 않았다. 그로부터 불과 14년 후, 2012년도에 개최된 유럽 사전학회의 원탁 좌담의 주제가 “2020년에도 사전을 만드는 사람이 존재할까?”였다는 사실만 보아도 사전의 변화가 20세기 후반과 21세기를 전후하여 얼마나 급격한 물살을 탔는지를 짐작할 수 있다. 이제 사전학 관계자들은 2020년을 기점으로 뜻풀이 같은 전통적인 인적 편찬의 영역조차 대거 기계로 넘어갈 것으로 예상하고 있다. 더 나아가 2040년 무렵이 되면 사전의 형태가 지금까지와는 완전히 달라져 있을 것이라는 전망도 나오고 있다(런델, 2012).

급속한 변화의 와중에는 대개 그렇듯이 미래와 과거가 각각 그 당위와 우위를 주장하며 경합하는 현상이 벌어진다. 사전 역시 그러한 상황에 처해

있다. 한편에는 최첨단 기술로 무장한 사전 개발 모델이 제시되는가 하면 또 다른 한편에는 전통적인 사전의 패러다임이 디지털의 외형만을 갖춘 채 존속하고 있다. 기술적 편차가 점점 더 크게 벌어지는 가운데, 사전의 형태도 더욱 다양해지고 있다. 앞으로 사전이 어떠한 길을 걷게 될 것인지는 수십 년 앞은 물론이고 향후 십 년 내외를 범위로 잡더라도 정확히 예측하기가 쉽지 않다. 그뿐만 아니라 현재 여기저기서 나타나는 사전의 변화 양상들 가운데 무엇이 앞으로 근본적인 혁신을 견인하는 큰 조류이고 무엇은 단기적인 실험이나 시도의 거품으로 꺼지게 될지도 섣불리 예단할 수 없다. 다만 지금 우리는 여러 방향으로 경합하고 혼재하는 사전의 모델과 그 발전의 전반적인 경향을 살펴 점차 분명하게 드러나고 있는 주요한 흐름을 짚어 볼 수는 있을 것이다. 이 글에서는 사전 발전의 흐름에서 두드러지게 나타나는 변화의 패턴을 사전의 제작 주체와 방식 면에서, 그리고 구조와 내용 면에서 거시적으로 조망해 보고자 한다.

1. 사전의 제작 주체와 방식의 변화

사전 편찬 주체의 경계가 스러지는 현상은 크게 '자동화'와 '대중화'라는 두 가지 핵심어로 요약이 가능할 듯하다. 전자는 사전 편찬이 전문가의 영역에서 탈피하여 대중의 역동적인 참여로 이행하고 있는 현상을, 후자는 사전 편찬 과정이 사람에서 컴퓨터로 이행하는 현상을 일컫는다. 사전 편찬 방식의 변화는 '데이터베이스-사전의 2단계 개발 모델'이 대표적인 추세가 될 것으로 전망된다.

사전의 제작에 컴퓨터가 도입된 것은 50년 전으로 거슬러 올라간다(슈타인과 어당, 1966). 컴퓨터를 제작에 도입한 최초의 사전은 1966년에 발간된 ≪랜덤하우스 영어 사전(Random House Dictionary of English Language)≫

으로 알려져 있는데, 이 시기에 컴퓨터는 교차 참조 항목을 확인하고 조판의 일관성을 유지하는 정도의 제한적인 용도로 사용되었다. 비록 제한적인 용도이기는 하였으나, 이 시점 이후로 사전의 어휘 자료는 하나의 데이터베이스로 인식되기 시작했다(런델과 킬가리프, 2011). 이후 1970년대에 들어 컴퓨터 조판을 활용한 사전 제작이 본격화하였고, 1980년대에 들어서는 편찬 단계의 후순위가 아닌 언어 자료의 수집 단계에서부터 컴퓨터가 활용되기 시작했다. 현재 컴퓨터는 언어 자료 수집과 사전 편찬 과정에 보편적으로 활용되고 있다.

컴퓨터와 인간의 협업으로 진행되는 사전 편찬의 과정은 대략 1) 사전의 기능과 목적에 따른 레마 추출, 2) 특정 레마에 대한 코퍼스 자료 인출, 3) 클러스터 분석, 4) 이에 따른 의미 및 결합 자질 분류, 5) 사전 정보의 기술(뜻풀이, 용례, 화용 정보 작성 등), 6) 기술된 정보의 제시 단계로 나눌 수 있는데, 이 가운데 1, 2, 3, 6단계는 컴퓨터에 의해, 그리고 4단계와 5단계는 인간에 의해 이루어지고 있다. 즉 사전 편찬의 많은 영역이 기계로 넘어갔지만 의미를 분류하고 기술하는 작업은 여전히 인간의 역할로 남아 있는 것이 지금까지의 추세였다고 볼 수 있다. 정리하면, 1960년대에 시작하였던 사전 편찬의 전산화는 1980년대를 기점으로 본격적으로 발전하였고, 그 이후 지금에 이르기까지 줄곧 사전 편찬의 영역을 확장하며 진행되었다고 볼 수 있다. 그런데 앞으로 진행될 사전 편찬의 전산화는 기존의 인간-기계 협업의 차원을 넘어설 것으로 보인다. 의미 영역에까지 '자동화'가 가능해지는 시점이 눈앞에 다가오고 있는 것이다.

영어를 비롯한 몇몇 유럽 언어에 국한하여 보자면, 형태소와 품사 추출은 2000년대 초반에 이미 안정적으로 자동화하였다. 이어 지덱스(GDEX)와 같은 프로그램이 도입되어 연어의 수집과 추출도 자동화하였다(런델과 킬가리프, 2011). ≪맥밀런 영어 사전(Macmillan English Dictionary for Advanced Learners, 2007)≫이 지덱스를 이용하여 편찬된 대표적인 사전

으로, 8,000개의 연어에 대한 용례를 자동화 방식으로 추출하였다고 한다. 종래에 사전 편찬에서 노동 집약적인 과정을 요구하던 용례 작성 역시 자동화하고 있다. 문장의 길이, 주변 어휘의 빈도나 고유 명사 포함 유무, 문장 길이와 복잡성, 대명사 수와 같은 필터를 적용하고 각각의 채택 기준에 가중치를 부여하면 컴퓨터가 자동으로 후보 용례들을 제시하고 사전 편찬자는 그중에서 적절한 용례를 고르기만 하면 되는 정도로 자동화가 진척되었다.

기존의 어휘에 새로운 뜻이 추가되는 어휘 변천 관련 증거 수집도 이제는 자동화될 전망이다. 2011년 전자사전 학회에 소개된 사전 편찬 로봇(dictionary-droid)이 이러한 기술적 발전을 보여 주는 예이다. 현재 신어 추출은 많은 부분 자동화된 프로세스를 이용할 수 있지만 이미 존재하는 어휘에 새로운 의미가 생성되는 경우는 기계적으로 추적하는 것이 쉽지 않았다. 그런데 특정 어휘의 의미는 그 어휘를 둘러싼 다른 어휘들과의 연관 속에서 드러나므로 종래의 결합 패턴과 동떨어진 패턴이 발견된다면 그 어휘의 의미가 달라졌음을 나타내는 신호가 될 수 있을 것이다. 사전 편찬 로봇은 바로 그 점에 착안하여 만들어진 기술이다. 이름에서 알 수 있듯이 이것은 일종의 웹 크롤러로서, 수많은 웹 사이트 링크를 따라다니면서 문서를 수집하는 과정에서 특정 레마의 결합 패턴이 기존의 패턴과 어긋난 경우를 감지한다. 이 기술은 아직 초기 단계에 머물러 있지만, 머지않은 장래에 상용화가 가능할 것으로 전망된다(매킨, 2011).

사전 편찬 단계에서 가장 '인간적인 영역'으로 남아 있는 뜻풀이마저 점차 기계로 이양될 조짐도 보인다. 사전 편찬 로봇을 활용하고 있는 워드니크(wordnik.com)와 같은 일부 인터넷 사전은 웹 코퍼스에서 정의문과 비슷한 구조를 취하는 대목(예: ** refers to~, ** means~) 등을 찾아내어 표제어의 뜻풀이를 대신하려는 시도를 하고 있다. 많은 신어들은 이 같은 과정을 통해서 상당 부분 뜻풀이에 해당하는 정보를 얻을 수 있다. 이처럼 데이터마이닝 기법을 이용하여 뜻풀이 작성을 자동화하는 것이 하나의 흐름

이라면, 또 다른 흐름으로 아예 전통적인 뜻풀이 기술 방식을 버리고 인용구 추출 방식으로 의미 기술을 대신하려는 추세도 있다. 인용구 추출 방식에 의한 뜻풀이는 특정 레마의 의미를 명시적으로 분류하여 제시하는 것이 아니라 해당 레마가 다른 어휘와 결합하는 표층 구조를 분석한 결과를 묶어서 제시하는 방식이다. 사실 특정 레마의 의미가 마치 그 레마에 내재적으로 귀속된 것인 양 의미 목록을 만들어 제시하는 전통적인 사전의 표제항 구조에 대해서는 그동안 사전학계에서도 비판이 적지 않았다(행크스, 2000). 따라서 인용구 추출 방식을 통해 귀납적으로 의미를 제시하는 방식은 기존의 사전 정보 기술상의 한계를 보완할 수도 있을 것이다. 의미 정보 기술의 자동화 추세와 관련하여 관찰할 수 있는 또 다른 변화로 웹 온톨로지와 결합한 어휘 연결망 모형을 들 수 있다. 이는 레마에 대한 뜻풀이 대신 어휘 연결망을 통해 레마의 의미 네트워크를 도식화해 보여 주는 것이다. 워드넷(WordNet)이나 프랑스어 어휘망(French Lexical Network) 프로젝트가 대표적인 예이다(폴게르, 2014).

사전 제작 주체의 대중화란 전통적인 사전 편찬 전문가와 민간이 함께 사전의 콘텐츠를 생산하는 협업적 사전 편찬을 가리킨다. 사전 편찬에 대중의 도움이 보태어진 것이 21세기에 새롭게 나타난 현상은 아니다. 19세기 중반에 시작된 ≪옥스퍼드 영어 사전(Oxford English Dictionary, 1857~1928)≫에는 당시 각계각층의 사람들이 보내 온 600만 개의 인용구가 녹아들어 있다. 그러나 웹 2.0이 몰고 온 협업화의 양상은 과거의 그것과는 확실히 다른 것이다. 사용자가 실시간으로 직접 데이터를 제공하고 서로 간에 공유할 수 있는 환경은 사전 편찬자의 고유 영역에 해당하던 작업의 경계를 허물어 놓았기 때문이다.

협업적 사전 편찬은 신어 및 속어, 전문 용어, 소수 언어와 위기 언어, 방언 및 지역어의 수집과 기술에서 잠재력이 크다. 대표적인 민간 참여형 사전 서비스의 예로는 속어 사전으로 널리 알려진 ≪어번딕셔너리(urban

dictionary.com)≫라든가 ≪윅셔너리(wiktionary.com)≫ 등을 들 수 있다. 영어가 국제 공용어가 되면서 전 세계의 지역 영어들도 협업적 사전 편찬에 가세하고 있는데, 그 한 예가 ≪중국영어 사전(cnglish.org)≫이다(친, 2015). ≪중국영어 사전≫은 영어에 들어온 중국어 및 중국의 영어 학습자들이 사용하는 영어 어휘, 중국어에 들어온 영어 외래어 등이 수집 대상이다. 협업적 사전 편찬은 다국어 사전 편찬에도 적극 도입되고 있다. 벤저민(2015)의 ≪카무시 사전(kamusi.org)≫은 자발적인 참여자들의 협업만으로 구축된 다국어 사전 사이트이다. 이 사이트는 일종의 게임 러닝 형식을 이용하여 참여자들에게 특정 단어의 대응어-정의문-용례 등을 단계적으로 기술하도록 하고, 다른 참여자들의 평가를 통해 어휘 정보의 완성도를 측정하는 방식을 채택하고 있다. 자금 조달의 어려움 등으로 이 프로젝트가 성공적으로 안착할 수 있을지는 좀 더 두고 보아야 할 듯싶다. 기성 사전 출판사들도 협업적 사전 편찬을 일부 도입하고 있다. ≪메리엄웹스터 오픈 사전(learnersdictionary.com)≫이나 ≪맥밀런 오픈 사전(macmillan dictionary.com)≫ 서비스가 그 좋은 예로, 신어나 지역 영어, 외래어와 같은 어휘 항목에 대해 일반 사용자의 참여를 적극적으로 반영하고 있다. 협업적 사전 편찬은 민간 주도로만 이루어지는 것은 물론 아니다. 대표적인 예가 국립국어원에서 서비스를 시작한 ≪우리말샘≫일 것이다.

이러한 개방적, 협업적 사전 편찬에 대한 우려의 목소리도 없지는 않다. 온라인 어휘 정보의 최대 강점이 찾으려고 하는 단어에 대한 정보를 발견할 확률이 종이 사전에 비할 수 없이 크다는 데에 있고, 이러한 장점은 협업적 사전 편찬 환경에서 극대화될 수 있을 것이다. 그럼에도 온라인 사전의 이용 만족도에 대한 대규모 조사(뮐러스피처, 2014)에서 알 수 있듯, 사용자들은 여전히 어휘 정보의 질과 신뢰성을 가장 중요한 만족 요인으로 꼽고 있는데, 개방형 사전에 대해서는 이러한 정보의 신뢰성이 높은 수준으로 유지될 수 있을지에 대한 우려가 일각에서 제기된다. 또 한편으로 사용자

참여형 환경이 특수한 사회 집단이나 이해관계가 첨예한 사안들에 정치적으로 이용될 위험이 있음을 경고하는 목소리도 나온다. 그러나 위키피디아를 비롯한 여러 개방형 지식 사이트의 전반적인 흐름으로 볼 때, 단일한 출처의 제한된 정보보다는 여러 출처의 경합하는 정보가 다수 제시될 경우 정보의 평균적인 질이 더 높아지는 집단 지성의 효과가 분명히 존재하며, 이는 개방형 어휘 사전이 효과적으로 운영될 경우 사전의 외연과 기능을 높일 수 있는 중요한 기회가 될 수 있음을 시사한다.

사전 제작 방식의 변화 흐름은 '데이터베이스-사전의 2단계 개발 모델'로 대표된다(앳킨스와 런델, 2008). 이것은 말 그대로 사전 편찬의 과정을 특정한 사전 결과물을 목표로 한 단일한 프로세스로 두는 것이 아니라 어휘 데이터베이스를 개발하는 과정을 별도로 두고 사전 제품은 이를 기반으로 그때그때 필요한 기능과 유형에 따라 편집 가공하는 이원적 프로세스를 일컫는다. 이러한 과정이 가능할 뿐만 아니라 더욱 효율적인 것으로 이해되고 있는 데는 사전의 어휘 정보에 대한 사람의 수요 못지않게 기계의 수요가 큰 데다, 사전에 기술될 정보와 그 정보의 속성을 분리할 수 있는 다층적인 메타데이터 마크업이 갈수록 정교해지고 있는 상황에 힘입은 바 크다.

특정 단어를 레마로 지정하고 그것을 표제어로 하여 세부 의미를 기술하는 위계적인 사전 정보 기술 방식은 인간 사용자의 직관적인 언어 인식을 반영하는 형식인 반면 기계에 의한 언어 정보 처리에는 매우 비효율적이다. 따라서 2단계 모델에서의 어휘 데이터베이스란 단지 특정 사전의 편찬을 위해 전자적으로 작성된 구조물이라기보다는 다른 출처의 어휘 데이터베이스 및 여타 웹 데이터와도 연결-병합-확장이 가능하고 의미 있게 활용될 수 있는 형식이어야 한다. 그러면서도 사전 개발 주체의 고유성과 저작권이 유지되어야 한다. 엑스엠엘(XML, eXtensible Markup Language)을 비롯해 링크트데이터(Linked Open Data)나 렉시컬 마크업 규격(Lexical Markup Framework) 등 사전에 기술될 정보와는 별도로 마크업이 가능한 메타언어

가 이미 사용되고 있으며 더불어 이에 대한 국제적 협약의 요구가 커지는 와중이기 때문에 단일한, 혹은 호환 가능한 데이터 포맷은 보급이 가속화할 것으로 보인다.

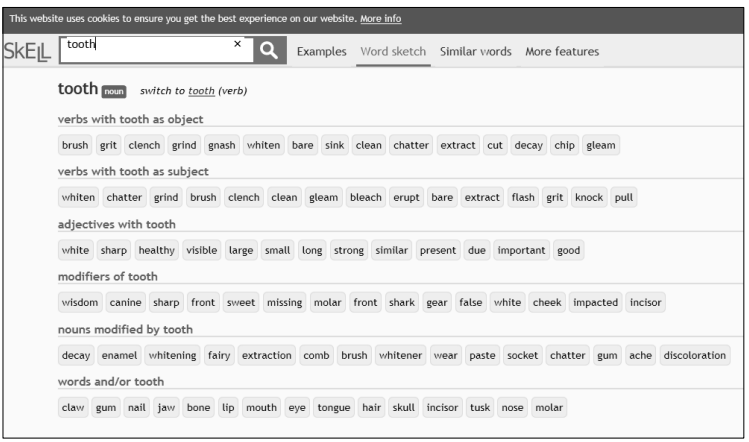
데이터베이스와 사전 기술의 2단계 개발 모델이 더욱 효율적인 사전 편찬 모델이 될 수 있는 것은 이렇게 함으로써 원자료(어휘 데이터베이스)를 각기 다른 2차, 3차 사전 제품에 중복 활용할 수 있고, 사전별로 각기 다른 구조를 취하면서도 추후에 서로 다른 사전 간의 재병합이 용이하며, 호환 가능한 데이터 포맷을 공유하는 조건이라면 타사의 사전 콘텐츠와도 연결이 가능하다는 데 있다. 2단계 모델에서의 어휘 데이터베이스는 레마를 기준으로 한 위계적 구조를 취하지 않고, 대신 '레마–의미' 쌍 하나하나가 개별적인 절대 주소를 갖는 단위(node)로 독립되어 있다. 그리고 같은 방식으로 부여된 다른 노드들과 특정 노드와의 관계를 속성으로 연결하는 구조를 취하고 있기 때문에 전체적으로 데이터의 구조가 기호 중심이 아닌 의미 중심의 네트워크를 취하게 된다. 이는 사전이 3000년 이상 속박되어 왔던 거시 구조와 미시 구조상의 경직성을 깨고 자모순 배열과 의미별 배열을 넘나들면서 재구조화할 수 있는 기반을 만들어 준다. 이러한 시도는 현재 다국어 어휘망 구축 사업에서 특히 활발히 진행 중이다. 다국어 어휘망 구축에 사활이 걸린 유럽 연합을 중심으로(예: LIDER project: liderproject-eu) 이스라엘의 커너만(Kernerman) 사는 40여 개 언어의 어휘망을, 옥스퍼드 대학 출판사에서도 100여 개 언어의 다국어 어휘망을 링크트데이터를 기반으로 하여 구축하는 중이다.

지금까지 사전 편찬 주체와 방식이라는 이 두 영역에서 사전의 변화 발전 양상을 살펴보았다. 이 두 영역은 서로 긴밀한 관련을 맺고 있는데 사전 제작의 자동화와 더불어 1차 데이터와 2차 사전 자료가 각각 독립된 구조와 내용을 가지고 개발되는 현상은 앞으로 정보 통신 및 언어 분석 기술이 발전하면서 더욱 가속화할 것으로 보인다. 더 중요한 것은 자동화가

더욱더 심화됨에 따라 1차 데이터-2차 사전 제작이라는 2단계 제작 모형조차 급속히 허물어질 수 있다는 사실이다. 부연하면, 1단계에 완성된 어휘 데이터가 사전 편찬이라는 별도의 인적 가공을 전혀 거치지 않고도 곧바로 기존의 사전과 다름없는 구성과 내용으로 변환 가능해지는 것이다. 이것이 현실화한다면, 어휘 데이터베이스가 사전을 위한 원재료인 것이 아니라 사전이 어휘 데이터베이스의 부수적 파생물이 되는 셈이다(레프, 2011).

이러한 현상은 학습자 사전 분야에서는 이미 부분적으로 나타나고 있다. 다음 [그림 1]은 렉시컬 컴퓨팅 리미티드에서 운영하는 스케치 엔진의 '스켈(SkELL, Sketch Engine for Language Learning)'이라는 영어 학습자 서비스이다(the.sketchengine.co.uk). 이 메뉴에서 단어 'tooth'를 입력하고 검색한 결과를 보자.

그림 1 스켈에서의 'tooth' 검색 결과



화면은 'tooth'와 관련하여 'tooth'의 동사 연어, 형용사 연어, 한정사 연어 및 명사 연어 등을 분류하여 보여 주고 있다. 특정 연어(예: brush)를 클릭하

면, ‘to brush teeth’가 나오는 코퍼스의 모든 예문을 선별하여 보여 준다. 홈페이지 상단의 메뉴 바에서 ‘example’을 클릭하면 ‘tooth’가 들어간 예문들이 사전의 용례처럼 정연하게 제시되어 있다. 메뉴 바에서 연관어(similar words) 항목으로 이동하면 그 밖의 신체 부위(예: body, arm, heart), 치아 관련 어휘(예: nail, surface, plate, hole), 치아와 밀접한 신체 부위 어휘(예: tongue, mouth, lip) 들이 제시된다. 스켈이 이 같은 어휘 정보를 추출하고 제시하는 전 과정은 자동으로 이뤄진다. 연관어 검색 역시 자동화된 프로세스의 결과이다. 지정한 단어와 연어 결합 구조와 성분이 최대한 유사한 단어들을 골라냄으로써 연관어를 추출할 수 있고, 반대로 연어 결합의 성분이 상반되는 것을 고른다면 반의어가 추출될 수 있다. 그림에서도 보이듯이 스켈은 매우 사용자 친화적인 정보 제시 구조를 취하고 있다. 뜻풀이 항목이 없다는 것만 제외하면 언뜻 보아 매우 잘 만들어진 학습자 사전과 구별되지 않는다. 앞으로는 점점 더 많은 코퍼스 데이터가 이와 같은 유사 사전의 형태를 취하며 대중에게 서비스될 것으로 전망된다.

2. 사전의 구조와 내용의 진화

사전의 구조와 내용 면에서의 진화 발전 양상은 크게 ‘사전의 구조적 해체와 정보 간의 융합’, 그리고 ‘사용자 맞춤형 사전’의 등장으로 설명할 수 있다. 통상적으로 사전은 특정한 분류 기준에 따라 언어 사전(辭典) 대 백과 사전(事典), 뜻풀이 사전 대 관련어 사전, 의미 사전 대 결합 사전, 단일어 사전 대 이어(또는 다언어) 사전, 일반 언어 사전 대 전문 용어 사전 등으로, 혹은 일반 사전 대 학습자용 사전 등으로 나뉠 수 있었다(하트만과 제임스, 1998). 이러한 분류법은 사전의 특정 기능 및 잠재적 사용자와 이에 따른 유형을 전제하고 이를 기준으로 사전의 내용과 구조를 갈래 지어

보려는 접근법이라고 할 수 있다. 그러나 이러한 개별적 사전 차원의 구조와 내용은 디지털 시대를 맞아 크게 흔들리고 있다. 이렇게 사전이 '헤쳐 모이는' 과정에서 생기는 역설적인 효과의 하나는, 사전이 실제 사용자 한 사람 한 사람의 구체적인 검색 패턴과 검색 욕구에 더욱 세심하게 부응할 수 있는 가능성을 열었다는 점이다.

개별 사전의 정태적 구조와 기능에 입각한 이분법의 붕괴는 전방위로 일어나고 있다. 먼저 사전 간의 연결과 통합 현상이 광범위하게 진행되고 있다. 이것은 사전의 정보 융합 현상과 내장형 표제항(embedded entry) 구조, 하이퍼링크 검색 필드의 확장 등으로 나타나고 있다. 사전들 간의 외형적 경계가 사라지고 하나의 검색 키워드 아래 통합하는 현상은 전자사전 단말기가 보급된 시점부터 활발히 이루어져 왔으며 이러한 경향은 앞으로도 지속될 것으로 보인다. 사전 콘텐츠의 통합 양상은 현재 다양하게 나타나는데, 크게 보아 1) 사전 조합(dictionary sets), 2) 포털 사전(dictionary portal), 그리고 3) 사전 정보 수집 제시형(dictionary content aggregator)으로 나뉜다(레프, 2011). 이 중 사전 조합과 포털 사전은 1990년대부터 죽 있어 왔던 방법이며, 사전 정보 수집 제시형 방식이 근래에 들어 주목받고 있다.

첫째 사전 조합 방식은 가장 오래된 사전 정보 융합 방식으로, 특정 사전 회사가 자사의 사전 제품들을 하나의 온라인 사전 사이트에 병합해 제시하는 방식을 일컫는다. 이러한 사전 조합은 온라인 사전 이전에 보급되었던 시디롬 사전에서도 흔히 볼 수 있었다. 현재는 대부분의 사전 회사들이 자사의 사전 콘텐츠를 유료 혹은 무료로 온라인상에 제공하고 있다. 조합형 사전의 사용자 인터페이스는 다양하여 페이지의 상단에 원하는 사전의 종류를 선택하는 방식과, 표제항 내부에 각기 다른 종류의 사전 정보가 병합 제시되는 방식(내장형 표제항), 이 둘을 조합한 방식 등 다양하다.

둘째 사전 포털 방식은 현재 일반 대중에게 가장 보편적으로 이용되고

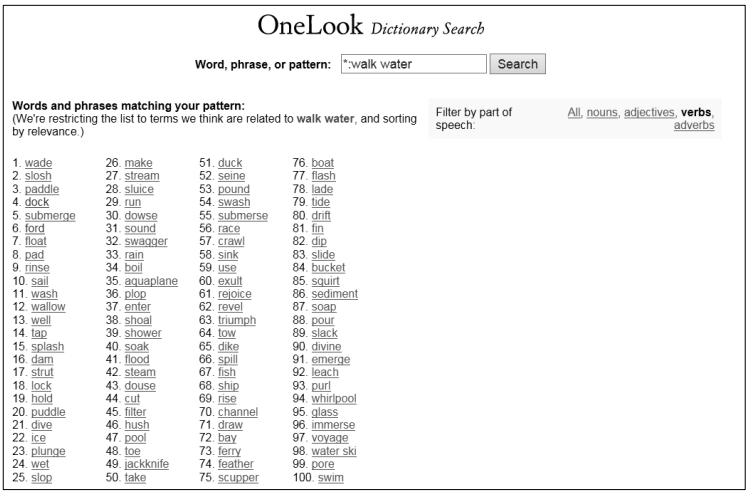
있는 사전 정보 제공 방식이라고 할 수 있을 것이다. 과거 여러 출처와 종류의 사전 수십 종이 하나의 단말기에 탑재되어 있던 휴대용 전자사전이 온라인으로 옮겨 간 것이 사전 포털 방식이라고 할 수 있다. 보통 사전 편찬과는 무관한 온라인 콘텐츠 제공 업체가 중심이 되어 서비스를 제공하고 있는데, 한국에서는 네이버나 다음의 사전 포털 서비스가 가장 많이 이용되고 있다. 사실 한국의 네이버나 다음이 제시하는 사전 포털 서비스는 상당히 독특한 유형의 서비스 제시 방식이라고 할 수 있다. 야후!나 구글과 같은 다국적 포털에서는 네이버나 다음처럼 사전 메뉴를 별도로 운영하지 않고 다른 사전 사이트의 링크 정보를 제공할 뿐이다. 국내에서 개발된, 야후!나 구글과 유사한 서비스로는 게리홈(garyshome.net)이 있는데, 이 사이트에서는 85개의 사전 링크를 제공하고 있다. 이들 포털은 여러 온라인 사전의 링크를 한데 모아 사용자가 특정 온라인 사전을 선택할 수 있도록 해 놓았을 뿐 네이버나 다음처럼 자체적으로 사전의 콘텐츠를 제휴하거나 제작하여 사용자에게 제공하지 않는다. 네이버나 다음과 같은 사전 서비스 모델은 특수한 한국적 맥락에서 개발된 사례라고 할 수 있다(배와 네시, 2014).

마지막으로 사전 정보 수집 제시형(CA)은 사전 간 정보의 융합이 가장 적극적으로 진척된 모델이다. 이 방식은 여러 종의 사전 정보를 한데 취합해서 하나의 웹 사이트에 제공하는 방식이라는 측면에서는 사전 포털 서비스와 외형적으로 비슷해 보이지만 그렇게 한데 모아진 정보를 처리하는 방식에서 사전 포털과 근본적인 차이를 보인다. 사전 정보 수집 제시형은 사전 출처의 메타데이터는 유지한 채 수집한 사전 정보를 모두 통합한다. 이것이 불러오는 검색 기능상의 효과는 사전 포털과 상당히 다르다. 사전 조합이나 포털 방식이 여전히 표제어를 중심으로 한 검색 결과를 보여 주는 반면, 수집된 사전 정보를 하나의 데이터베이스로 묶는 사전 정보 수집 제시형 방식에서는 검색 쿼리를 표제어 차원에서 더 나아가 정의항의 내용으로까지 확장하는 것이 가능하다. 사전 정보 수집 제시형 방식에서는 엔그램(n-gram)

분석을 통해 통상 사전의 좌측 핵 구조(left core structure)에 제한되어 있던 검색이 우측 핵 구조(right core structure)의 정보 내용에까지 확장되기 때문에, 사용자는 사전 거시 구조의 배열 방식에 구애됨 없이 형태 정보나 내용 정보를 자유롭게 검색할 수 있게 되었다. 아울러 우측 핵 구조의 핵심 요소인 뜻풀이문에 대한 엔그램 검색이 가능해지면서 비슷한 개념적 연관성을 지니는 표제어들을 역으로 추출할 수 있다.

사전 정보 수집 제시형 방식으로 운영되는 온라인 사전의 대표적인 예가 바로 원룩(onelook.com)이다. 원룩은 현재 1,061종의 온라인 사전의 정보를 취합하여 제공하고 있는데, 메인 홈페이지에 목표어를 입력하면 이어지는 창에서 해당 어휘가 수록된 사전들의 출처와 각각의 사전들의 표제항 내용을 볼 수 있다. 여기까지는 포털 사전과 기능이 유사하다. 그런데 원룩은 통상적인 표제어 검색 말고도 역순 검색(reverse search)이 가능하다. 위에서 언급한 것처럼 원룩에서는 제휴한 천여 종 사전 콘텐츠에 대해서 엔그램 분석이 가능하기 때문에 사용자가 특정 표제어가 아닌 개념어로 쿼리를 요청했을 때 그것과 가장 유사한 결합 구조를 가진 뜻풀이들을 걸러 내어 이에 해당하는 표제어를 제시하는 역검색을 수행한다. 예를 들어 원룩의 ‘reverse dictionary’ 메뉴에서 검색창에 ‘walk’과 ‘water’를 입력한 뒤 ‘동사(verb)’로 범위를 제한하면 다음과 같은 검색 결과가 나타난다.

그림 2 원룩에서의 'walk', 'water' 검색 결과



OneLook *Dictionary Search*

Word, phrase, or pattern: *:walk water Search

Words and phrases matching your pattern:
(We're restricting the list to terms we think are related to **walk water**, and sorting by relevance.)

Filter by part of speech: All, nouns, adjectives, **verbs**, adverbs

1. wade	26. make	51. duck	76. boat
2. slosh	27. stream	52. seine	77. flash
3. paddle	28. sluice	53. pound	78. lade
4. dock	29. run	54. swash	79. tide
5. submerge	30. dowse	55. submerse	80. drift
6. ford	31. sound	56. race	81. fin
7. float	32. swagger	57. crawl	82. dip
8. pad	33. rain	58. sink	83. slide
9. rinse	34. boil	59. use	84. bucket
10. sail	35. aquaplane	60. exult	85. squirt
11. wash	36. plop	61. rejoice	86. sediment
12. wallow	37. enter	62. revel	87. soap
13. well	38. shoal	63. triumph	88. pour
14. tap	39. shower	64. tow	89. slack
15. splash	40. soak	65. dike	90. divine
16. dam	41. flood	66. spill	91. emerge
17. strut	42. steam	67. fish	92. leach
18. lock	43. douse	68. ship	93. purl
19. hold	44. cut	69. rise	94. whirlpool
20. puddle	45. filter	70. channel	95. glass
21. dive	46. hush	71. draw	96. immerse
22. ice	47. pool	72. bay	97. voyage
23. plunge	48. toe	73. ferry	98. water ski
24. wet	49. jackknife	74. feather	99. pore
25. slop	50. lake	75. scupper	100. swim

[그림 2]에서 보이듯이 'wade(얕은 물이나 진창을 건너다)', 'slosh, paddle(물에서 첨벙거리다)', 'dock(부두에 배를 대다)', 'submerge(잠수하다)', 'ford(걸어서/차로 물을 가로지르다)' 등의 동사 표제어 목록이 하이퍼링크로 처리되어 제시되며, 해당 링크를 열면 사전별로 표제항 정보를 읽을 수 있다. 역검색은 결합 형태의 표층적인 일치 정도를 기준으로 하였으므로 검색 결과에는 목표 어휘뿐만 아니라 그것과 개념상 부분적으로만 교집합을 이루거나 반의어 관계에 있는 연관 어휘들이 망라되어 나올 수 있다.

정리하면, 원룩 등에서 채택한 사전 정보 수집 제시형 방식은 사전의 우측 핵 구조에 속해 있는 내용 정보(좌측 핵 구조 속의 형식 정보에 대비되는)를 통합적으로 검색할 수 있으면서도 사전의 출처와 종류, 해당 사전의 표제어 수, 해당 표제항이 업데이트된 날짜 등을 식별할 수 있는 메타데이터를 분리하여 제시해 주고 있기 때문에 개별 사전 편찬 주체의 저작권을 명시하면서도 정보 간의 상호 연결성이라는 수요에 부응하는 사전 정보

융합 모델이라고 할 수 있다.

사전의 구조와 내용 면에서 일어나고 있는 또 다른 흐름은 사용자 맞춤형 사전의 등장이다. 사용자 맞춤형 사전에는 먼저 사용자가 자신의 검색 욕구를 파악하여 검색의 범위와 내용을 지정하는 '사용자 지정형(user adaptable)' 방식과 사전이 특정 사용자의 검색 히스토리를 추적하여 검색 프로파일을 만들고 이에 따라 정보의 범위와 내용을 조정하여 제시하는 '사용자 적응형(user adaptive)' 방식이 있다. 이 두 가지 방식은 사전의 사용자가 누구인지, 어떤 목적으로 사전을 사용하는지에 따라 각각 장단점이 있으며, 하나의 방식만 적용되기보다는 두 가지가 적절히 결합하여 쓰일 수도 있다.

온라인 사전이 보편화하면서 사전 편찬자나 사이트 운영자는 사용자의 로그 파일을 수집하고 분석할 수 있게 되었다. 로그 파일은 시계열에 따른 사이트 방문 횟수라든가 특정 검색어의 빈도, 방문 시간 및 기간 등에 대한 정보를 제공한다. 로그 파일을 분석하여 검색된 표제어를 추출해 보면 미등재된 신어들을 확인할 수 있을 것이다. 그뿐만 아니라 자주 검색 대상에 오르는 잘못된 철자나 비표준어, 검색에 자주 오르는 굴절어형, 해당 언어에 존재하지 않는데도 검색에 오르는 단어 등을 확인할 수 있을 것이다. 이는 개발자에게 사전의 정보를 개선하고 사용자 편의에 맞게 수정하는 데 유용하게 쓰일 수 있다. 그러나 로그 파일 분석은 매우 피상적이고 표면적인 데이터에 불과한 것인 경우가 많다. 또한 검색 기록이 표제어 단위로만 남기 때문에 표제항 내부로 세션이 이동한 다음에는 사용자가 주로 어떤 정보에 눈길을 두는지, 검색은 성공했는지, 정보의 내용은 충분히 만족스러웠는지 등과 같은 유의미한 질적 정보는 수집할 수 없다(베를랭드와 비농, 2010). 로그 파일 분석의 또 다른 한계는 90퍼센트 이상의 검색이 인간 사용자가 아닌 웹 크롤러나 웹 스파이더와 같은 검색 로봇에 의한 것이라는 사실이다.

로그 파일의 이러한 한계를 극복하고 사용자 편의에 맞는 정보를 제공하

기 위해 먼저 떠오른 방법이 바로 '사용자 지정형' 사전 인터페이스이다. 사용자가 사전의 제시 정보의 범위와 내용을 스스로의 필요에 맞게 설정해 주면 그다음부터 사전 사이트는 설정한 범위에 맞게 내용을 조정하여 제시한다. 사용자 지정 기능으로는 가장 단순하게는 정보의 양을 조정할 수 있는 기능이라든가(예: show more, show less) 정보의 난이도를 조정하는 기능을 들 수 있다. 가장 복잡한 지정형 모델의 예로는 벨기에의 프랑스어 기초 사전(La Base Lexicale du Français, 현재는 Interactive Language Tool로 개칭, Verlinde와 Binon, 2010)을 들 수 있을 것이다. 이 사전은 초기 화면의 검색 지정 메뉴를 통해 사용자가 표제어에 대한 정보를 원하는지, 아니면 연어 정보를 원하는지, 불확실한 어휘 정보를 확인하려는 목적인지, 아니면 어휘 학습 목적인지, 모어 대응어 정보를 원하는지, (영어, 프랑스어, 네덜란드어 중) 어떤 언어 대응형을 알고 싶은지 등을 묻는 선택 항목을 클릭하여 자신에게 필요한 정보를 정확하게 얻을 수 있도록 하고 있다. 프랑스어 기초 사전과 흡사하게 베르겐홀츠와 욘센(2015) 등이 개발한 덴마크어 숙어 표현 사전(Danish Dictionary of Fixed Expressions: idiomordbogen.dk)에서도 사용자가 자신의 참조 목적을 지정할 수 있는 검색 옵션을 제공하고 있다.

이러한 방식은 사전 사용자 입장에서는 자신의 검색 욕구에 들어맞는 정보를 얻을 수 있다는 기능적 이점이 있고, 사전 편찬자 입장에서는 로그 파일의 피상적 데이터와는 달리 사전 사용의 동기와 욕구를 정확히 파악할 수 있다는 이점이 있다. 그러나 이 방식은 사용자 편의성 측면에서 문제가 있음이 드러났다. 사전이란 사용자가 별도의 언어 활동을 수행하다가 특정 어휘에 대한 정보 욕구를 느껴 이용하게 되는 참조 도구인데, 사용자들에게 검색 단계 전에 자세한 검색 욕구를 스스로 분석하여 그에 걸맞은 인터페이스를 지정하도록 요구하는 절차가 상당히 번거로운 것이다. 또 의도와는 달리 이 같은 인터페이스로도 사용자 욕구는 정확히 측정하기가 쉽지 않을

수 있다. 실제로 사용자는 '자신이 무슨 정보를 어떤 상황에서 원하는지'를 콕 집어 인식하지 못하는 경우가 대부분이다(트라프옌센, 2010).

이처럼 사용성에서 명백한 한계를 지닌 사용자 지정형 모델과는 달리 사용자 적응형 모델은 정보 기술만 뒷받침된다면 매우 효율적인 모델이 될 수 있다. 이것은 온라인 쇼핑몰이나 검색 엔진에서 흔히 활용되고 있는 적응형 하이퍼미디어(adapative hypermedia)와 유사하다. 즉 사용자의 사전 사용 행태를 지속적으로 분석하여 해당 사용자의 검색 패턴을 파악하고, 이에 맞춰 사전의 내용과 제시 정보의 우선순위를 조정함으로써 최적화한 기능을 구현할 수 있다는 것이다. 이 모델은, 앞서 말한 로그 파일 분석 기법 등 기술적 한계로 인해 아직까지는 본격적으로 현실화되지 못했지만, 특수 목적 사전과 같이 사용자층이 제한적인 전문 사전의 영역에서 적응형-지정형의 하이브리드 형태로 조금씩 시도되고 있는 상황이다. 대표적인 예가 콰리(2012)가 진행하고 있는 인도네시아 재무 전공자를 위한 영어-인도네시아어 재무 전문 사전(Dictionary of Finance for Indonesian Learners)이다. 일반 언어 사전과는 달리 제한된 특수 사용자층이 이용하는 특수 목적 사전은 표제어 단계에서의 검색 행태를 추적하는 것만으로도 상당히 유의미한 사용자 정보를 얻을 수 있다. 가령 검색 순위 상위에 오른 표제어들의 수준만으로도 사용자의 전문성의 정도를 추측할 수 있으며, 그에 맞춰 표제항의 난이도를 선택적으로 제시할 수 있다. 또 전문 분야 사전은 고정 사용자층의 비율이 높으며 사전의 기능도 제한적이므로 로그 파일을 통한 프로파일과 더불어 이용자 등록 시에 이용자가 사전에 대해 원하는 정보의 유형과 종류를 스스로 지정하도록 함으로써 지정형 기능도 함께 활용할 수 있는 이점을 갖고 있다. 만일 특정 사용자의 사전 사용 행태가 시간이 흐름에 따라 변화한다면, 사전은 다시금 이에 맞춰 제시 정보의 방식을 바꿀 수 있다. 사용자 지정형이든 적응형이든 이러한 기능이 최적으로 활용될 수 있으려면 사전 데이터의 마크업이 이에 맞춰 개선될 필요가 있다.

특정 사전 콘텐츠의 사용자 특성에 대한 면밀한 분석을 바탕으로 사전의 정보 내용을 세분화하고 이에 대해 사전의 사용자 및 기능과 관련한 속성을 부여함으로써 디지털 사전은 사용자의 필요와 행동 특성에 실시간으로 대응하는 유연한 구조 체계를 갖출 수 있다(보스마, 2011).

3. 맺으며

지금까지 우리는 앞으로 사전이 밟아 갈 변화와 발전의 방향을 사전 제작 주체와 제작 방식의 변화, 그리고 사전의 구조와 내용의 변화 측면에서 살펴보았다. 영어권 및 일부 유럽어의 사례를 중심으로 자동화와 제작 주체의 대중화, 어휘 데이터베이스의 발전, 사전의 거시 구조 및 미시 구조의 해체와 재구조화와 같은 사전학계의 주요 논의들이 실제로 어떤 양상으로 나타나고 있는지 알아보았다.

이 글에서 미처 다루지 못한, 아울러 논의되어야 할 몇 가지 주제들이 남아 있다. 지금까지의 변화 추세를 따를 때 사전 편찬 사업이 어떠한 방향으로 수익과 고용을 창출할 수 있는지, 그리고 저작권과 재산권을 보호하면서 아울러 전 세계적 어휘-지식 네트워크에 사전이 어떻게 기여할 수 있는지 등과 같은 수익성 및 정보 처리 상호 운영의 문제 역시 중요한 현안이다. 또 사전의 유통과 이용이 모바일 플랫폼으로 급격히 이동 중인 상황을 고려하면, 사전이 유통되고 이용되는 양상이 앞으로 어떻게 변화할 것인지와 같은 문제 역시 활발하게 논의되어야 할 것이다. 더 나아가 모바일 사전보다 모바일 번역기가 전 세계적으로 더욱 활발히 이용될 가능성이 훨씬 커지는 지금의 추세를 감안할 때, 사전의 변화는 사전 안에서가 아닌 사전 바깥에서 비롯된 변인에 의해 새로운 국면을 맞게 될 수도 있을 것이다. 사전은 존속하기보다는 진화해야 할 것이다. 전체 속의 부분이라는 오랜 구조의 제약에서

풀려나 정보와 정보가 이어지고 통합되는 연결 양상으로서의 구조를 모색하고 시도함으로써 우리는 사전을 21세기의 정보 사회에 걸맞은 도구로 쇄신할 수 있을 것이다.

참고 문헌

Atkins, S. and Rundell, M.(2008), *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.

Bae, S. and Nesi, H.(2014), Korean and English 'dictionary' questions: what does the public want to know?, *Lexicography ASIALEX 1*, 53~71.

Benjamin, M.(2015), "Crowdsourcing microdata for cost-effective and reliable lexicography" in Proceedings of Asialex 2015 Hong Kong, Hong Kong, 213~221.

Bergenholtz, H. and Johnsen, M.(2015), User research in the field of electronic dictionaries: methods, first results, proposals., *Dictionary Encyclopedia of Lexicography V*, 556~568.

Bothma, T. J. D.(2011), "Filtering and adapting data and information in an online environment in response to user needs" in Fuertes-Olivera, P. A. and Bergenholtz, H. (eds.), *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, London: Continuum, 71~102.

Grefensgtette, G.(1998), "The future of linguistics and lexicographers: will there be lexicographers in the year 3000?" in Proceedings of the 8th Euralex Congress, Liege, 25~41.

Hanks, P.(2000), Do word meanings exist?, *Computers and the Humanities 34*, 205~215.

Hartmann, R. R. K. and James, G.(1998), *Dictionary of Lexicography*, London: Routledge.

Kwary, D. A.(2012), Adaptive hypermedia and user-oriented data for online dictionaries: A case study on an English dictionary of finance for Indonesian students, *International Journal of Lexicography 25(1)*, 30~49.

Lew, R.(2011), "Online dictionaries of English" in Fuertes-Olivera, P. A. and Bergenholtz, H. (eds.), *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, London: Continuum, 230~250.

McKean, E.(2011), "Wordnik: notes from an online dictionary project"

in Electronic lexicography in the 21st century: new applications for new users. Available online at http://videolectures.net/elex2011_mckean_notes/.

Müller-Spitzer, C.(2014), *Using Online Dictionaries*, Berlin: de Gruyter.

Polguère, A.(2014), From writing dictionaries to weaving lexical networks, *International Journal of Lexicography 27(4)*, 396~418.

Qin, X.(2015), "The crowdsourced compilation of the China English Dictionary" in Proceedings of Asialex 2015 Hong Kong, Hong Kong, 74~79.

Rundell, M.(2012), "'It works in practice but will it work in theory?' The uneasy relationship between lexicography and matters theoretical' in Proceedings of the 15th EURALEX International Congress, Oslo, 47~92.

Rundell, M. and Kilgarriff, A.(2011). "Automating the creation of dictionaries: Where will it all end?" in Meunier, F., de Cock, S., Gilguin, G., and M. Paquot (eds.), *A Taste for Corpora: In Honour of Sylviane Granger*, Amsterdam: John Benjamins, 257~282.

Stein, J. and Urdang, L.(eds.)(1966), *Random House Dictionary of English Language*, New York, NY: Random House.

Trap-Jensen, L.(2010), "One, two, many: customization and user profiles in Internet dictionaries" in Proceedings of the 14th Euralex Congress, Neeuwarden, 1133~1143.

Verlinde, S. and Binon, J.(2010), "Monitoring dictionary use in the electronic age" in Proceedings of the 14th Euralex Congress, Neeuwarden, 1144~1151.