어휘정보구축을 위한 사전텍스트의 구조분석 및 변환

최병진*†
목포대학교

Byung-Jin Choi. 2002. A Structural Analysis of Dictionary Text for the Construction of Lexical Data Base. *Language and Information 6.2*, 33-55. This research aims at transforming the definition text of an English-English-Korean Dictionary (EEKD) which is encoded in EST files for the purpose of publishing into a structured format for Lexical Data Base (LDB). The construction of LDB is very time-consuming and expensive work. In order to save time and efforts in building new lexical information, the present study tries to extract useful linguistic information from an existing printed dictionary. In this paper, the process of extraction and structuring of lexical information from a printed dictionary (EEKD) as a lexical resource is described. The extracted information is represented in XML format, which can be transformed into another representation for different application requirements. **(Mokpo National University)**

Key words: 사전편찬학, 어휘정보, 전산사전, 어휘지식베이스, 어휘데이터베이스

1. 이끄는 말

현대 언어학의 다양한 언어이론에서 점차 중요시되고 있는 부분이 사전이다. 또한 전산학의 인공지능분야에서는 사전을 자연언어처리에 이용하고자 많은 학자들이 전자사전 개발에 상당한 노력을 기울여 왔다. 사전은 이처럼 언어학과 전산학의 양 영역에서 주목을 받으면서 그 중요성이 점차 부각되어 가고 있다.

국외에서는 이미 사전에 대한 연구가 언어학의 한 분야로 독립적인 위치를 차지하면서 인접학문과의 공조 속에서 그 결실을 맺고 있는 반면, 국내에서는 아직까지 사전에 대한 연구가 하나의 독립된 분야로 자리를 잡기보다는 의미론이나 통사론과의 부분적인 관련 속에서 연구가 이루어져 왔다. 90년대에 들어서야 사전편찬학에 대한 연구가 점차 주목을 받게 되었고, 2002년에 처음으로 한국사전학회가 출범함으로써 사전편찬학에 관한 연구가 독자적인 자리를 차지하기 위한 기반을 마련하고 있다.

이 연구는 사전편찬학의 한 응용 분야라고 할 수 있는 전산사전편찬학의 관점에서 단순한 텍스트 형식의 출판용 사전을 기계가독형 사전(Machine Readable Dictionary)으로 바꿀 수 있는지에 대한 생각에서 출발하였다. 이러한 연구는 이

* 534-729 전남 무안군 청계면 도림리 61 국립 목포대학교 인문과학대학 독일언어문화학과, E-mail: bjchoi@apollo.mokpo.ac.kr.
† 이 논문은 2001년 8월 언어과학회 전국학술대회에서 발표한 내용을 수정 보완한 것입니다. 이 논문에 대한 논평과 제안을 하여 주신 익명의 심사위원들께 감사를 드립니다.

미 국외에서는 이루어지고 있지만, 국내에서는 거의 이루어지지 않고 있으며[1], 사전편찬과 관련하여서는 주로 말뭉치(corpus)를 중심으로 연구가 이루어지고 있다.

사전편찬작업은 상당한시간과노력이 요구되는 작업이다. 그러나 새로운 사전을 편찬하는데, 기존의 어휘정보를 이용할 수 있다면, 많은 시간과 노력의 비용을 줄일 수 있을 것이다. 이 연구에서는 기존의 어휘정보를 재사용할 수 있는지의 가능성과 문제점을 검토함으로써 전자사전편찬을 위한 어휘데이터베이스의 구축에 기여하고자 한다.

2. 전산사전편찬학의 이해

2.1 사전

사전이란 무엇인가? 사전의 정의에 대한 이와 같은 질문에 우리는 아마도, 단어들이 알파벳 순서로 정리되어 그 의미가 설명되어 있는 책이라고 대답할 수 있을 것이다. 그러나 이러한 대답은 여러 가지 형태를 지니고 있는 사전의 다양성을 충족시키기에는 어렵다고 할 수 있다.

그 이유는 첫째로 사전에는 단어만이 표제어로 등재되는 것은 아니기 때문이다. 사전이라는 개념의 중심에는 물론 단어가 큰 비중을 차지한다. 그러나 좀 더 포괄적으로 생각해보면 사전의 어휘항목(lexical entry)으로 단어 뿐 아니라 다양한 종류의 어휘 단위(lexical unit)가 나타날 수 있다[2].

두 번째로 사전은 알파벳 순서로 정돈될 필요도 없다. 물론 대부분의 사전은 알파벳 순서로 정돈되어 있다. 그것은 사전의 본질이 사전을 처음부터 끝까지 읽어 가면서 필요한 정보를 얻는 것이 아니라, 특정한 위치의 정보를 참조하고자 하는 것이기 때문이다. 사전은 세부적으로 나누어진 정보를 포함하고 있으며, 필요한 정보는 일정한 체계 속에서 찾아볼 수 있어야 한다. 정보를 찾아볼 수 있는 가장 단순하고 객관적인 체계가 알파벳 순서라고 할 수 있다. 그러나 사전이 알파벳 순서이외에도 계층구조를 이루고 있는 개념기반의 시소로스나 동의어 사전에서와 같이 다른 형태의 정돈 체계를 가질 수도 있다[3].

세 번째로 대부분의 사전에서는 원하는 어휘항목에 대한 정의가 설명되어 있지만, 사전이라고 반드시 어휘항목에 대한 정의가 있어야 할 필요는 없다. 실제로 주어진 어휘항목에 대하여 단순한 형태의 정보만으로 충분할 때가 있다. 또한 극단적인 경우에는 어휘항목만 존재하는 것으로도 사전의 기능을 충분히 감당할 수 있다[4].

마지막으로 사전은 반드시 책의 형태로 존재할 필요가 없다. 수 백년을 거쳐 오늘에 이르기까지 사전은 주로 인쇄된 형태로 존재하였다. 그러나 최근에 와서는 사전이 마이크로 필름시트나 컴퓨터의 지원을 받아 디지털화된 형태로 바뀌고 있다. 심지어는 컴퓨터에 음성이나 기록을 통하여 입력된 단어가 음성이나 화

1. 이에 해당하는 연구로 최병진(1996), 노용균(2001)을 들 수 있다. 최병진(1996)에서는 국어사전을 기계가독형으로 만들려는 연구가 시도되었고, 노용균(2001)에서는 BBI 영어사전을 OCR로 인식하여, XML형식으로 변환하려는 시도가 있었다.
2. 예를 들어 사전의 유형에 따라 접사 내지는 형태소도 사전의 어휘항목으로 등재될 수 있다.
3. 이러한 경우에는 일반적으로 알파벳 인덱스가 첨가된다.
4. 정서법사전이나 역순사전, 또는 형태소 사전은 사전에 어휘항목만 존재하는 것으로서 사전으로서의 기능을 충분히 감당하고 있다.

면을 통해서 다른 언어로 번역되어 출력되는 사전도 있다.

이와 같은 내용을 고려하여 볼 때 "사전은 특정한 매체를 통해 특정사용자에게 필요한 정보를 쉽게 찾아볼 수 있도록 정돈된 어휘요소(주로 단어)의 목록"[5]이라는 Hausmann의 정의는 적절하다고 할 수 있다.

2.2 전산사전편찬학

전산사전편찬학의 개념을 이해하기 위해서는 어휘학과 사전편찬학의 개념을 먼저 살펴볼 필요가 있다.

언어학의 한 분야로서 어휘학(lexicology)이란 명칭은 그리이스어의 "단어에 관한, 단어와 관련된"이란 의미의 *'lexikós'*와 "이론 내지는 학문"의 의미를 지닌 *'logos'*에서 유래되었다. 한마디로 어휘학은 단어에 관한 학문으로, 어휘(lexis)[6]가 무엇인가에 대한 질문에 대답하기 위해 노력한다. 어휘를 그 연구대상으로 삼는 어휘학은 어휘정보를 서술하기 위한 언어학적 이론이나 방법론과 관련된 기술언어학의 한 분야로, 연어, 관용어, 어휘의미론[7], 어장(word fields)의 구조, 의미성분이나 의미관계와 밀접한 관련이 있다.

반면에 사전편찬학(lexicography)이란, 과거에는 사전을 편찬하는 과정이나 행위[8]를 의미하였지만, 최근에 와서는 그 과정뿐만 아니라, 응용언어학의 한 분야로 실질적인 사용을 위한 사전의 구축 및 디자인과 관련된 이론으로까지 그 개념이 확장되었다. 사전편찬학의 대상은 사전으로, 사전은 미리 주어진 것이 아니라 의도적으로 창조된 것이다. 사전은 그 사용자에게 단어나 어휘에 대한 질문에 상담자로서의 역할을 수행한다. 따라서 사전편찬학자들은 사용자들의 요구에 부합하기 위해서 어떻게 사전을 구성해야 할 지에 대해서 연구를 한다.

이러한 사전편찬학의 원칙적인 방법은 기호이론적인 연구에 기반을 두고 있다. 전통적인 사전편찬학의 기호모델은 단어의 형태(word form)와 개념으로 이해되는 의미를 구별하고 있는 Saussure의 이원적 모델이다. 이 단순한 기호모델이 어휘를 검색하는 절차적인 기준에 따라 다음과 같이 사전유형을 근본적으로 구별할 수 있는 기초를 제공한다.

> (1). 어의론 사전(semasiological dictionary): 전통적인 유형의 사전으로 검색의 기준은 단어형태이며, 단어에 대한 정보는 바로 그 의미이다.
>
> (2). 명칭론 사전(*onomasiological* dictionary): 시소러스유형의 사

---

5. Hausmann, F. J. (1985: 369)
6. 어휘가 무엇인가에 대해서는 다음과 같은 세 가지 입장이 있다.
   a. 개개인에게 있어서 정신적으로 저장된 단어로서의 어휘: 이에 대하여 우리는 mental lexicon이라는 용어를 사용한다.
   b. 자연언어의 기본으로서의 어휘: 이에 대해 우리는 어휘목록이라는 명칭을 사용한다. 사전편찬작업은 일반적으로 이러한 의미의 어휘와 연관되어 있다.
   c. 문법과의 관계속에서의 어휘: 이에 대해서는 어휘부(lexicon)라는 이름이 사용된다.
7. 최근까지도 어휘의미론은 단어의 음운적, 형태적, 통사적 특질과 별도로 다루어졌으나, 최근의 언어학 이론에서는 이러한 여러 층위의 내용들이 점차로 통합하고 있다.
8. Wiegand (1983: 38)

전으로 검색의 기준은 개념이며, 그 검색어가 가지고 있는 정보는 단어형태이다.

이와 같은 기호모델에 따라서 사전의 전체적인 구성과 관계된 거시구조(macro structure)가 결정된다. 사전이 거시구조에 있어서 어의론적 구조(semasiological structure)를 지니면, 일반적인 알파벳 순서로 정렬된 사전의 모습을 띄게 되고, 명칭론적 구조(onomasiological structure)를 지니면, 시소러스나 개념사전의 형태를 취한다.

한편 개별어휘항목의 정의부분을 어떻게 구성해야 할 지에 대한 문제는 바로 미시구조(micro structure)의 설정과 연관되어 있다.

어휘항목은 사전의 기본단위로, 표제어(lemma)와 그에 대한 설명부분으로 나눌 수 있다. 미시구조에 대한 연구는 바로 이 어휘항목의 설명부분에 대한 연구이며 설명부분에 대하여 세부항목을 설정하고 어떠한 구조를 지닐 것인가에 대한 윤곽을 정하는 것이다. Balidinger(1960)는 사전편찬학의 중심 문제를 사전에서의 어휘의 구조화로 특징짓고 있다. 어휘의 구조화는 바로 거시구조 안에서 미시구조를 정렬하는 것이다. 즉, 어휘항목들 사이의 관계(semasiological, onomasiological)의 관점에서 전체적인 거시구조를 정하고, 개별 어휘항목에 대하여 어휘정보를 어떻게 설정해야 할지의 미시구조를 정하는 것이다.

사전편찬을 위한 거시구조(사전구성)와 미시구조(어휘항목)는 언제나 긴밀하게 연결되어 있다[9].

한편, 사전편찬작업은 항상 많은 시간과 노력이 소요되는 작업이었으나, 전산학적인 처리 방법들이 사천편찬에 도입되면서 사전편찬작업은 가속화되고, 그 규모도 커지게 되었다. 특히 단어와 용례를 말뭉치에서 추출하거나, 텍스트의 분석결과를 바탕으로 어휘데이터베이스를 구성하고 다양한 형태(일반사전, 시소러스, 콩코던스)의 사전을 편찬하는데 있어서 지난 수 십년 동안 컴퓨터는 큰 도움을 주었다.

오늘날에 와서는 사전편찬작업을 위해서 컴퓨터의 이용은 필수 불가결하게 되었으며, 또한 자연언어처리를 위한 사전은 컴퓨터에 의해 처리될 수 있도록 가공되어야 하였기에 사전편찬작업과 컴퓨터는 매우 긴밀한 관계를 유지하게 되었고[10] 학문적으로도 컴퓨터를 이용한 전산처리가 불가피하게 되어 전산어휘학(computational lexicology)이나 전산사전편찬학(computational lexicography)이라는 용어까지 나오게 되었다.

사실 전산어휘학과 전산사전편찬학이라는 개념을 분명하게 설명하는 것은 그리 간단하지 않다. 부분적으로는 전산어휘학과 전산사전편찬학의 경계가 불분명하고 유동적이어서 그 경계를 명백하게 설정하기가 어렵다[11].

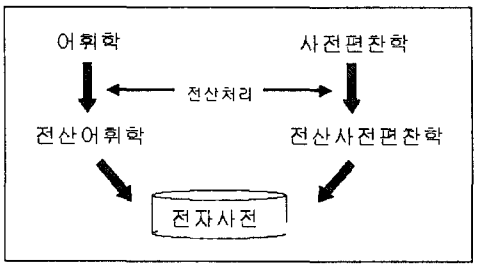
전산사전편찬학의 주된 관심사의 하나는 기계가독형태의 사전이나 말뭉치를 바탕으로 자연언어처리를 위한 대규모의 사전을 구축하는 것이다. 이러한 작

9. 소프트웨어기술의 발전으로 하나의 공통의 사전편찬을 위한 데이터베이스로부터 상이한 거시구조와 미시구조를 이끌어낼 수 있다.
10. 특히 사전편찬기업들은 사전을 편찬하는데 있어서 항상 컴퓨터에 기반한 방법을 사용하였다. 예를 들어 Birmingham 대학에서 수행되어 만들어진 the Collins Cobuild Dictionary series는 수백만 단어의 영어 텍스트와 대화를 바탕으로 거의 완벽하게 자동적으로 컴퓨터 분석에 의해 만들어진 것이다.
11. 전산사전편찬학의 대표자라고 할 수 있는 Boguraev의 경우에도 전산사전편찬학이나 전산어휘학의 개념을 혼용하여 사용하고 있다. Boguraev and Briscoe (1989)와 Boguraev and Levin (1993) 참조.

업은 어휘정보의 구조를 파악하고, 그 내용을 언어학적으로 분석함으로써 자동화 내지는 반자동화된 방법으로 어휘정보를 어휘정보의 원천으로부터 추출하는 과정도 포함하고 있다.

한편, 어휘정보의 구조를 파악하고 어휘정보를 추출하는 과정 속에서, 어휘자료(어휘지식베이스)를 구축하는데 사용될 수 있도록 현존하는 어휘자원(사전이나 말뭉치)을 구조화하거나 혹은 재가공하면서 어휘정보획득의 모델을 발전시켜 나가는 것이 전산어휘학(computational lexicology)라고 할 수 있다.

이러한 관계를 도식화하면 <그림 1>과 같이 나타낼 수 있으며, 전산어휘학과 전산사전편찬학은 궁극적으로 전자사전의 편찬을 위한 것이라 할 수 있을 것이다.



[그림 1] 전산어휘학과 전산사전편찬학

이러한 맥락에서 본다면, 전산사전편찬학이 전산어휘학을 포함하고 있다고 볼 수 있다.

2.3 전산사전편찬학의 국내외 연구동향

지금까지 전산사전편찬학 분야에서는 말뭉치와 기계가독형태의 사전으로부터 어휘자원을 추출하는 새로운 방법론에 관한 많은 연구와 프로젝트들이 수행되어 왔다.

사전편찬작업을 컴퓨터를 이용하여 자동화하려는 생각은 이미 60년대에 소그룹이지만 유능한 사전편찬학자들(불란서의 Bernard Quemada, 화란의 Felicien de Tollenaere, 독일의 Gerhard Wahrig)에 의해서 제기되었다. 독일의 Wahrig는 "사전편찬작업의 새로운 방향(Neue Wege in der Wörterbucharbeit) (Wahrig 1967)"에서 컴퓨터 지향적인 사전편찬학을 제안하고 이러한 원칙에 따라 독일어 대사전(Großes Deutsches Wörterbuch)을 편찬하였다. 그러나 Wahrig는 실제로 사전편찬작업을 자동화하기보다는 사전편찬작업의 전산처리를 위한 알고리즘만을 제시하였고, 실제 작업은 수작업으로 이루어졌다.

사전편찬작업의 초기 자동화시대에는 많은 사전들이 컴퓨터의 도움으로 편찬되고 인쇄되었지만 대부분의 경우 어휘정보가 단순히 사람에 의해 입력되고 이것들이 자동적으로 정렬되어 처리되는 정도에 불과했다. 사전편찬을 위한 기초작업에서 편찬할 때까지 컴퓨터가 읽어 처리할 수 있는 말뭉치를 중심으로 사전을 편찬하는 경우는 거의 없었다.

그러나 사전편찬작업에 컴퓨터를 이용하려는 노력이 점차로 증가하면서, 사전편찬작업을 도와줄 수 있는 콩코던스(concordance) 프로그램이나 온라인 문

서검색 시스템, 또는 사전편찬 워크벤치 등이 개발되었고, Longman Dictionary of Contemporary English (LDOCE 1978), Collins COBUILD English Language Dictionary (COBUILD 1987), Brockhaus-Wahrig (Wahrig 1981), Trésor de la Langue Française (Trésor 1971-1986)[12]와 같은 사전이 전산처리방법의 도움으로 편찬되었다.

한편, 일반 텍스트 형태의 사전을 기계가독형태의 사전으로 만든 최초의 작업은 1960년대 후반에 Systems Development Corporation(SDC)에서 Olney에 의해 이루어졌다. 그가 일반 텍스트 형태의 사전을 종이 테이프위에 키펀치를 하여 기계가독형태로 만든 사전이 바로 "Webster's Seventh New Collegiate Dictionary"이다.

그 후 Michiel(1982)은 최초로 기계가독형태의 LDOCE 사전에 대한 연구를 시작하였다. 그는 LDOCE사전에서 어휘항목의 구조를 조사하고, 표제어의 정의 부분에 대한 문법을 파악하고, 사전에 사용된 문법적 코드를 설명하면서 자연어 처리를 위한 그 적절성을 보여주었다. 이와 같이 어휘정보가 기계가독형태로 저장되면서 이 어휘정보를 자연언어처리시스템이나 기계번역시스템에 이용하려는 연구도 늘어났다. 그러나 실제로 COBUILD, LDOCE, Trésor, Wahrig 등의 언어 사전에 수록된 단어 의미의 정의를 수정이나 추가 작업 없이 그대로 전산 처리하기에는 부적합하다. 따라서 전산학, 언어학, 전산사전편찬학이나 전산언어학에서는 단어의 의미를 정형화된 형태로 나타낼 수 있는 방법을 찾으려 노력하였다.

LDOCE나 COBUILD사전의 경우와 같이 기계가독형태로 이미 형식화가 되어 있는 경우에는 동의어, 상위어, 반의어 등과 같은 의미관계를 추출하기는 비교적 용이하다. 그러나 풀이, 정의, 또는 문맥이나 예문에서 정형화된 형태의 의미관계를 추출하는 것은 쉽지 않다.

한편, 전산언어학에서는 일상텍스트로부터 이러한 어휘정보를, 특히 단어의 미에 관한 정보를 추출하려는 연구가 일찍부터 진행되었다. 이러한 연구는 구조주의 언어학의 문맥에서 요구하는 방법론과 매우 유사하다. 즉 문제가 되는 요소의 환경을 분석하여 어휘정보를 추출한다는 점에 공통점이 있다[13].

그리고 어휘풀이, 정의, 예문에서 의미관계를 발견하기 위해서는, 어휘를 인식하고 또한 풀이 텍스트의 구조를 파싱할 수 있는 프로그램이 개발되어야 하는데, Alshawi (1989), Boguraev (1991), Boguraev (1994), Copestake (1990), 또는 Neff & Boguraev (1989) 등의 연구가 이에 해당한다.

한편 국내에서는 1990년대에 들어오면서 사전편찬학에 대한 관심이 일기 시작하였으며, 사전편찬의 실제 작업과 관련하여 컴퓨터를 이용하려는 시도도 함께 이루어졌다. 국내에서는 주로 고려대학교, 국립국어연구원, 연세대학교 사전편찬실(현 언어정보개발연구원의 전신), 한국과학기술원을 중심으로 사전편찬에 대한 연구가 이루어졌다. 사전편찬작업을 수행하는 과정 속에서 개개의 사전편찬팀은 각각 개별적으로 말뭉치를 수집하고 이를 가공함으로써, 사실상 시간과 비용에 있어서 중복된 작업을 수행하였다. 이러한 상황 속에서 한국과학기술

12. B. Quemada에 의해 주도된 Trésor 사전편찬을 위해서는 상당히 많은 양의 불란서 문헌과 사전자료의 데이터가 일관성있게 모아졌다.
13. 예를 들어, 공기분석(cooccurrence analysis)을 위한 여러 가지 다른 방법론이 있고, 한 요소의 특정 의미영역을 정하기 위한 방법들, 또한 연어분석이나 Wahrig가 제안한 핵심구의 변형에 관한 방법들을 생각할 수 있다.

원의 인공지능연구센터에서는 전자사전의 표준화와 통합을 위한 목적으로 텍스트 및 사전 관리시스템(TDMS)을 개발하고 국내 전자사전의 표준화를 시도하였으나, 그 연구가 지속적으로 이루어지지 못했다. 그 뒤 문화관광부의 지원을 얻어 세종계획이라는 국책 프로젝트가 진행되었고, 이 프로젝트를 통해 세종사전[14]이라는 새로운 전자사전이 개발되고 있다.

이러한 국내의 연구상황속에서 우리가 생각해 보아야 할 것은 과연 앞으로의 국내 전자사전 개발이 어떻게 이루어져야 할 것인가를 심도있게 생각해 보아야 할 것이다. 앞으로의 전자사전개발 방향의 진로설정을 위하여 이 논문에서는 전산사전편찬학의 실제적인 측면에서 기존사전의 어휘정보를 이용하여 전자사전을 구축하는 한 방법론을 소개하고자 한다.

3. 어휘정보구축의 실제

이 장에서는 기존의 인쇄된 형태의 사전에서 어휘정보를 자동으로 변형하여 전자사전의 어휘데이터베이스로 구축하는 방법을 소개하고자 한다.

어휘정보를 어떠한 방법으로 표상하느냐에 따라 각각 다른 형태의 다양한 데이터베이스모델[15]이 이용된다. 따라서 어휘표상은 사전편찬작업 중에서 가장 중요한 이슈가운데 하나이다. 또한 어휘표상은 컴퓨터 구현을 위해서 충분히 잘 정의되어 있어야 하며, 각각 다른 종류의 어휘정보에 대해서는 다른 종류의 표상이 필요한 경우가 있다.

현재 가장 잘 알려져 있고 널리 수용되고 있는 사전편찬학적 데이터베이스 표상의 현대적 유형은 SGML(Standard Generalised Markup Language)을 이용한 텍스트의 어휘적 기술방법이다. 현재 개발되고 있는 국내의 대부분의 전자사전에서도 이와 유사한 방법을 취하고 있다. 그 이유는 사전의 표준화와 어휘정보의 재사용과 관련하여 어휘정보가 일관적이며 특정 사전에 대해서 독립적인 형태를 지녀야 하기 때문이다.

이러한 생각을 바탕으로 개발된 전자사전은 어휘지식베이스(Lexical Knowledge-Base: LKB)와 어휘데이터베이스(Lexical Data-Base: LDB)의 형태로 구분될 수 있다.

3.1 어휘지식베이스와 어휘데이터베이스

최근 들어 컴퓨터 용량과 성능의 발전은 인공지능분야에서 전문가 시스템의 발전을 가능하게 하였다. 또한 자연언어 처리기술의 발달과 함께 사전을 전문가 시스템의 한 형태로 간주하여 연구가 이루어지고 있다. 이러한 경향은 우리가 사전을 펼쳐 단어를 찾아 그 의미를 파악하는 과정이 일종의 질의응답과정과 유사하기 때문에 사전을 전문가 시스템의 형태로 만들어 볼 수 있을 것이라는 생각에서 비롯되었다. 사전을 전문가 시스템으로 간주하려는 생각은 사전을 일종의 "어휘데이터베이스(Lexical Data-Base: LDB)"로 간주하는 견해에 "어휘지식베이스(Lexical Knowledge-Base: LKB)"라는 새로운 개념을 도입하게 하였다. 전자사전에 대해 사용되는 이 두 가지 용어는 가끔 혼용되어 잘못 이해되기도 한다.

14. 세종사전은 프랑스의 M. Gross의 어휘문법(lexicon-grammar)에 기초하여 어휘정보를 구축하고 있다.
15. 어휘정보를 나타내기 위한 모델은 전통적인 관계데이터베이스형태에서부터 특수한 목적의 계층구조형태에 이르기까지 다양하다.

여기서는 이 두 용어를 구별하기 위하여 그 개념을 먼저 살펴보기로 한다.

이론적으로 어휘데이터베이스란 음운론적, 형태론적, 구문론적, 의미론적인 데이터의 구조화된 집합체로 간주된다. 어휘항목은 특별한 의미와 미리 결정된 성질의 집합을 연관시키며, 어떠한 단어에 대해서도, 그것과 관련된 어휘적 성질만이 표현된다. 즉 어휘데이터베이스는 단어를 행으로 하고 그 어휘적 성질을 열로 하는 배열 모양을 지닌다. 현존하는 대부분의 전자사전은 어휘데이터베이스의 일종으로 간주할 수 있다. 또한 대부분의 기계가독형 사전은 어휘데이터베이스의 한 형태로 온라인 상에 올려져 있다. 어휘데이터베이스에서 새로운 어휘정보를 추가하는 것은 데이터베이스에서 단어를 행으로 하고, 어휘적 성질을 열로 하는 2차원 행렬을 증가시키는 것과 같다. 이와 같은 어휘데이터베이스는 자연히 정적인 사전형태를 유지한다. 어휘데이터베이스 안에는 어휘추론시스템이 포함되어 있지 않다. 따라서 어휘데이터베이스의 틀 구조 안에서 어휘적 관계를 표현하는 것은 가능하지만, 어휘적 관계를 통해서 어떤 개념을 얻고자 한다면 이러한 틀 구조에서 벗어나야 한다.

반면 어휘지식베이스는 어휘데이터베이스의 개념에 추가로 데이터에 대한 추론을 가능하게 함으로써 보다 풍부하고 동적인 구조를 지니게 한다. 어휘지식베이스에서는 개별적인 단어의 의미가 어휘의 성질에 대해서만 정의되어 있는 것이 아니라, 다양하게 서로 연결되어 있는 개념체계 속에서 정의되어 있다. 그리고 어휘지식을 위해 사용된 개념은 '관계'를 고려하고 있다. 따라서 어휘의 의미적 관계나 일반적인 세계지식을 기호화할 수 있는 '관계'의 일반화된 기본 개념들은 모두 어휘지식베이스를 구현하는데 필요하다.

어휘지식베이스를 구성하는데 고려하여야 할 사항은 먼저, 단어의 특수한 어휘자질 이외에 무엇이 어휘지식을 구축하고 있는가를 파악하는 것이다. 즉 무엇이 언어학적으로 중요한 일반화된 지식인지를 알아내는 것이다. 그리고 이러한 일반화된 지식을 통합할 수 있는 추론 메카니즘을 어떻게 도입할 것인 지를 판단하는 것이다.

지금까지 설명한 어휘데이터베이스와 어휘지식베이스의 개념을 재고해보면, 국내에서 개발되고 있는 모든 전자사전이 대부분 어휘데이터베이스의 관점에서 개발되고 있다고 판단된다.

어휘데이터베이스의 경우에는 개별어휘항목에 대한 미시구조의 정보를 입력함으로써 어휘정보의 표상이 단순화될 수 있지만, 잉여적인 정보를 반복해서 입력해야 하는 번거로움이 있다. 한편 어휘지식베이스의 경우에는 개별 어휘항목의 정보이외에 다른 어휘항목과의 의미적인 관계를 고려해야 하는 많은 경험적 작업이 뒷받침되어야 한다. 이러한 점을 감안하면, 과연 어휘정보를 어휘데이터베이스와 어휘지식베이스의 두 가지 형태 중 어느 것을 선택하여 전자사전으로 구축하는 것이 더 효율적인지에 대한 고려가 있어야 할 것이다

3.2 전산사전편찬학의 실제

인쇄사전으로부터 필요한 어휘정보를 추출하여 전자사전으로 가공하기 위해서는 인쇄된 사전의 어휘정보를 규범화된 형태로 바꾸어야 한다. 어휘정보를 규범화된 형태로 인코딩하기 위해서는 사전항목에 대한 인식과 사전에 내재되어 있는 구조적 원칙에 대한 이해가 필수적이다. 이 장에서는 민중서림의 영영한사전

을 기계가독형태의 어휘데이터베이스로 바꾸는 과정을 설명한다[16].

3.2.1 처리과정. 민중서림의 영영한사전은 맥킨토시 운영체제에서 전자출판을 위한 형태로 편찬되어 있기 때문에 전산처리를 통하여 사전의 어휘정보를 추출하는 것이 가능하다고 판단하였다. 실제로 인쇄사전의 형태로 출판된 어휘정보는 <그림 2>와 같다.

a·bate·ment [əbéitmənt] *n.* U **1** the act of abating; the state of being abated. 감소; 감가; 감퇴; 완화. ¶ ~ *of debts* 빚의 감소 / *make an* ~ *of the price* 값을 깎다. **2** destroying the cause of; putting an end to 원인을 없앰; 끝을 냄. ¶ ~ *of a nuisance* 불법 방해 배제.

[그림 2] 인쇄사전의 어휘정보

이러한 형태의 파일을 EST(Extended Style Tag)파일로 저장하게 되면, 개개의 어휘정보와 함께 그 정보를 표현하고 있는 문자형태의 정보가 <그림 3>에서와 같이 나타나게 된다[17].

```
<DP12J0.347*F"75 Helvetica Bold">a<DP9*F"AMin">.<DP12J0.347*F"75
Helvetica Bold">bate<DP9*F"AMin">.<DP12J0.347*F"75 Helvetica Bol
d">ment <DP9.5*F"AProunA1">[2b tm2nt] <D*H110P9.5*F"I Century Ol
d Style Italic">n.<DP9*F"#중 명조"> <DP9*F"AMin">U<DP9*F"#중 명조"
> <DP9*F"75 Helvetica Bold">1<DP9*F"#중 명조"> <DP9.5*F"Bookman">
the act of abating; the state of being abated. <DP9*F"#중 명조">
 감<R><DP9*F"#중 명조">소<DP9.5*F"Bookman">;<DP9*F"#중 명조"> 감가<
DP9.5*F"Bookman">;<DP9*F"#중 명조"> 감퇴<DP9.5*F"Bookman">; <DP9*
F"#중 명조"> 완화. <DP9*F"AMin">A_ <D*H110P9.5*F"I Century Old St
```

16. 어휘데이터베이스 형태로 구축된 정보는 자연언어처리나, 사전편찬학, 그리고 전자출판에 이르기까지 다양하게 활용될 수 있다.
17. 민중서림의 영영한 사전은 맥킨토시 기반의 QuarkXpress라는 소프트웨어를 이용하여 편집되어 있다. QuarkXpress에서 작업한 파일은 PC와 호환이 되지 않기 때문에 이를 전산처리하기 위해서는 QuarkXpress에서 저장할 때에 EST형식으로 저장해야 한다. 이것은 'WORD'에서 텍스트를 RTF형식으로 저장함으로써 텍스트와 그 텍스트의 포맷과 관련된 정보를 함께 저장할 수 있는 것과 유사하다.

```
yle Italic">of debts<DP9*F"#중명조"> 빚의 감소<DP9*F"AMin">? <D*
H110P9.5*F"I Century Old Style Italic">make an<DP9*F"AMin"> _ <D
*H110P9.5*F"I Century Old Style Italic">of the price<DP9*F"#중명
조"> 값을 깎다. <DP9*F"75 Helvetica Bold">2<DP9*F"#중명조"> <DP9
.5*F "Bookman">destroying the cause of; putting an end to<DP9*F"
#중명조"> 원인을 없앰<DP9*F"Bookman">;<DP9*F"#중명조"> 끝을 냄.
<DP9*F "AMin">A_ <D*H110P9.5*F"I Century Old Style Italic">of a
nuisance <DP9*F"#중명조"> 불법 방해 배제.
```

[그림 3] "abatement"의 EST파일형식

이 연구의 목적은 바로 EST파일에 나타나 있는 단층형의 어휘정보 텍스트를 '<DP9*F"AMin">'와 같은 문자태그를 열쇠로 하여 위계형의 구조를 가진 문서로 변형하고자 하는 것이다. 다시 말하면 <그림 2>의 물리적인 형태의 텍스트를 <그림 3>에서와 같이 어휘정보의 구분을 어느 정도 가능하게 할 수 있는 문자 스타일의 태그를 바탕으로 텍스트의 구조를 분석하는 것이다.

<그림 2> 형태의 인쇄사전으로부터 정보의 손실 없이 어휘정보를 추출해내기 위해서 EST파일에 나타나 있는 문자 스타일을 위한 태그를 분석하는 '태그 분석 과정'과 이를 바탕으로 문서의 구조를 분석하여 구조화하는 '문서 구조화 과정'의 두 부분으로 나누어 작업을 수행한다.

'태그 분석 과정'은 문서의 구조와는 상관없이 다만 특정 문자나 문자열을 다른 문자나 문자열로 치환하는 전처리 과정의 일을 수행한다.

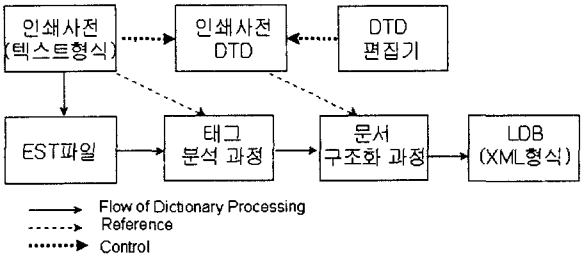
반면에 '문서 구조화 과정'에서는 <그림 3>에서와 같은 텍스트로부터 구조화된 형태의 어휘정보를 XML의 형식으로 바꾸는 일이 이루어진다. 이를 좀 더 구체적으로 살펴보면 태그 분석 과정에서는 EST파일의 내용을 검토하여 불필요한 폰트정보나 실제 사전의 문자와 다른 문자를 치환함으로써 문서 구조화 과정에서 어휘정보의 구분을 용이하게 한다. 문서 구조화 과정에서는 실제로 어휘정보추출을 위한 텍스트의 구문분석(parsing)[18]을 하는데 EST파일에서는 사전의 주 내용 이외에는 폰트정보만 표시되므로 이 정보를 최대한 이용하여 구문분석을 한다.

아래 <그림 4>는 이러한 처리과정의 절차를 도식화하여 나타내고 있다.

3.2.2 태그 분석 과정. EST파일에는 어휘정보 텍스트의 글자체를 위한 태그만이 표기되어 있다. 이 태그는 다른 글자체를 위한 태그가 나오기 전까지 유효하며 중첩되는 부분은 존재하지 않는다.

EST파일 전체에 쓰인 태그를 추출한 후, 이 태그들이 파일 내에서 어떠한 기능을 가지고 있는 지를 조사한다. 주로 사용된 태그의 목록과 그 기능은 다음과 같다[19].

18. 여기서의 구문분석이란 문장의 구조를 분석하는 통사적 분석이 아니라 텍스트의 구조를 분석하는 것을 의미한다.
19. 실제로는 155개의 태그가 분석되었으며, 각각의 태그가 위치하는 깊이(depth)도 중요한 의미를 지닌다.



[그림 4] 사전 텍스트의 분석 과정

사전 항목	폰트정보
표제어	<DP120J0.347*F"75 Helvetica Bold"> <DJ0.347*F"75 Helvetica Bold">
단어의 분절	<DJ0.347*F"AMin"> <DJ0.463*F"AMin">
발음 기호	<DP9.5*F"AProunA1">
중요서 표시	<DJ0.231> <DPJ0.231*F"#중 명조"> <DP9J0.231*F"#중 명조">
마치는 태그	<D>
이탤릭 영어 문자	<D*H110P9.5F"I century Old Style Italic">
가산명사 구분	<DP9*F"AMin"> <D*F"AMin"> <DP9J0.347*F"AMin">
겹괄호	<DP9*F"AMin"> <D*F"AMin"> <DP9J0.347*F"AMin">
어의번호	<DP9*F"75 Helvetica Bold">
참고어, 어법, 용례 기호(¶) 표시	<DP9*F"AMin"> <D*F"AMin"> <DP9J0.347*F"AMin">
관용구	<DP9*F"76 Helvetica BoldItalic">
동사의 문형	<DP9.5*F"#세나투">
어형변화(명사, 동사)	<D*K10*F"75 Helvetica Bold">

	<D*F"75 Helvetica Bold">
굵은 이탤릭 영어 문자	<DP9*F"76 Helvetica BoldItalic"> <D*F"76 Helvetica BoldItalic">
파생어 구분 문자 ·	<D%80P9*F"#중 명조">
설명에서 빈 칸(space bar)	<DP9*F"#중 명조">
줄바꿈	<R>
윗첨자	<*^*F"Bookman"> <*^P12*F"Bookman"> <*^9.5*F"Bookman"> <*^P9.5*F"AMin"> <*^P9*F"#중 명조"> <*^P12J0.347*F"Bookman"> <*^J0.347*F"Bookman">
사전 항목	폰트정보
보통 영어 문자	<DP9.5*F"Bookman"> <D*F"Bookman"> <DP9*F"Bookman">
굵은 영어 문자	<DP9.5*F"Helvetica Bold"> <D*F"75 Helvetica Bold">
보통 한글	<DP9*F"중 명조" >
영어문장의 용례에서 숫자 표시	<IP9.5>
그 밖의 특수 기호들 (문자치환)	<DP9*F"AMin"> <DP9.5*F"AMin"> <D*F"AMin"> <DP9J0.347*F"AMin">

[표 1] 문자태그의 기능

태그를 분석하는 전처리과정에서는 이미 앞에서도 언급한 바와 같이 구조에 상관없이 필요에 따라 문자의 유형에 대한 정보를 추출하고, 또한 이를 바탕으로 특수문자를 치환하여 어휘정보분석을 용이하게 한다.

다음의 <표 2>는 일정한 다른 문자로 치환될 수 있는 문자 태그의 예를 보여준다. 여기서는 모두 <DP9*F"AMin">, <DP9.5*F"AMin">, <D*F"AMin">, <DP9J0.347*F"AMin"> 태그를 중심으로 이 태그의 뒤에 나오는 문자는 태그에 의하여 정상적인 텍스트로 바뀌어 구문분석을 가능하게 한다.

예) <DP9.5*F"AMin">2<DP9.5*F"Bookman">rst cousin
==> **first cousin**

치환전	치환후	치환전	치환후
/	·(작은점)	[	〔
.	•(큰점)	]	〕
\C0\6F	?	A_	¶~
4	fl	?	/
3	ffi	_	~
2	fi	-	—
1	ff	a	ⓐ
9	fl	b	ⓑ
7	fi	c	ⓒ
<<	((	u	ⓤ
>>	))		

[표 2] 문자 치환 테이블

<표 2>에 나타난 문자 치환 이외에도 다음과 같이 특수문자를 위한 치환이 이루어진다.

<DP9*F"#중 명조">\xD5를 영문에서 작은 따옴표(')로 치환.
<DP9*F"#중 명조">\xA3\xBD(2바이트짜리 등호)의 경우에는 '='로 치환.
<DP9*F"#중 명조">\xAC\x73()을 <ARROW>로 치환.
<D%80P9*F"#중 명조">\xA1\xDC와 <D%80>\xA1\xDC(•)를 <DOT>로 치환.

3.2.3 문서 구조화 과정. 문서 구조화 과정에서는 문자정보와 그 문자정보가 사용된 위치를 중심으로 사전 텍스트의 구조를 분석한다.

텍스트의 구조를 분석하기 위하여 우선 인쇄사전의 '일러두기'와 사전에 등재된 어휘항목의 정의 텍스트를 중심으로 문서의 문서유형정의(DTD: Document Type Definition)를 다음과 같이 설정할 수 있다.

```
<?xml version="1.0" encoding="EUC-KR"?>
<!ELEMENT 어휘항목 (표제어, 발음*, 풀이+, 어원*, 파생어구*)>
<!ELEMENT 표제어 (중요도?, 철자, 구분번호?)>
    <!ELEMENT 중요도 (#PCDATA)>
    <!ELEMENT 철자 (#PCDATA | 철자-미국식 | 철자-영국식)*>
        <!ELEMENT 철자-미국식 (#PCDATA)>
        <!ELEMENT 철자-영국식 (#PCDATA)>
    <!ELEMENT 구분번호 (#PCDATA)>
<!ELEMENT 발음 (#PCDATA | 발음-미국식 | 발음-영국식)*>
    <!ELEMENT 발음-미국식 (#PCDATA)>
    <!ELEMENT 발음-영국식 (#PCDATA)>
```

```
<!ELEMENT 풀이 (품사*, 문형?, 가산성?, 어형변화*, 설명+,
        관용구*)>
    <!ELEMENT 품사 (#PCDATA | 문법설명)*>
    <!ELEMENT 문법설명 (#PCDATA)>
    <!ELEMENT 문형 (#PCDATA)>
    <!ELEMENT 가산성 (#PCDATA)>
    <!ELEMENT 어형변화 (명사복수꼴?, 동사변화?, 비교최상급?)>
        <!ELEMENT 명사복수꼴 (#PCDATA)>
        <!ELEMENT 동사변화 (#PCDATA)>
        <!ELEMENT 비교최상급 (#PCDATA)>
    <!ELEMENT 설명 (어의번호?, 특수용법?, 전치사부사?, vt목적어*,
         전문어?, 언어용법?, 영어설명+, 한글역어+, 설명-용례구*)>
        <!ELEMENT 어의번호 (#PCDATA)>
        <!ELEMENT 특수용법 (#PCDATA)>
        <!ELEMENT 전치사부사 (#PCDATA)>
        <!ELEMENT vt목적어 (#PCDATA)>
        <!ELEMENT 전문어 (#PCDATA)>
        <!ELEMENT 언어용법 (#PCDATA)>
        <!ELEMENT 영어설명 (#PCDATA)>
        <!ELEMENT 한글역어 (한글역어해설*, 한글설명+, 동의어*,
                반의어*, 참고어*, 어법*, 참고사항*)>
            <!ELEMENT 한글역어해설 (#PCDATA)>
            <!ELEMENT 한글설명 (#PCDATA)>
            <!ELEMENT 동의어 (#PCDATA)>
            <!ELEMENT 반의어 (#PCDATA)>
            <!ELEMENT 참고어 (#PCDATA)>
            <!ELEMENT 어법 (#PCDATA)>
            <!ELEMENT 참고사항 (#PCD ATA)>
        <!ELEMENT 설명-용례구 (설명-용례, 설명-동의용례*,
                설명-역어*)>
            <!ELEMENT 설명-용례 (#PCDATA)>
            <!ELEMENT 설명-동의용례 (#PCDATA)>
            <!ELEMENT 설명-역어 (#PCDATA)>
    <!ELEMENT 관용구 (관용어+, 동의관용어*, 관용어목적관계?,
            관용어설명+)>
        <!ELEMENT 관용어 (#PCDATA)>
        <!ELEMENT 동의관용어 (#PCDATA)>
```

```
        <!ELEMENT 관용어목적관계 (#PCDATA)>
        <!ELEMENT 관용어설명 (관용-어의번호?, 관용-영어설명*, 관
            용-한글역어*, 관용-대표제어참조?, 관용-용례구*)>
            <!ELEMENT 관용-어의번호 (#PCDATA)>
            <!ELEMENT 관용-영어설명 (#PCDATA)>
            <!ELEMENT 관용-한글역어 (#PCDATA)>
            <!ELEMENT 관용-대표제어참조 (#PCDATA)>
            <!ELEMENT 관용-용례구 (관용-용례, 관용-동의용례*,
                관용-용례한글설명*)>
                <!ELEMENT 관용-용례 (#PCDATA)>
                <!ELEMENT 관용-동의용례 (#PCDATA)>
                <!ELEMENT 관용-용례한글설명 (#PCDATA)>
<!ELEMENT 어원 (#PCDATA)>
<!ELEMENT 파생어구 (파생어,파생어발음?,파생어품사)>
    <!ELEMENT 파생어 (#PCDATA)>
    <!ELEMENT 파생어발음 (#PCDATA)>
    <!ELEMENT 파생어품사 (#PCDATA)>
```

[표 3] 영영한 사전을 위한 문서유형정의(DTD)

이러한 문서유형정의를 바탕으로 EST파일의 텍스트를 분석하는 문서 구조화 과정이 이루어진다. 그러나 이 과정에서 EST파일에서 추출된 태그의 기능 뿐 아니라 위치를 고려하여 태그의 기능을 보다 세분화하고, 이를 바탕으로 텍스트를 분석한 결과, 앞에서 설정한 문서유형정의에 상응하지 않는 문제점들이 발견되었다. 또한 사전의 '일러두기'에 언급되어 있지 않은 다양한 예외 규칙들이 존재한다는 사실도 텍스트를 분석하면서 확인되었다. 따라서 기존의 문서유형정의만으로는 이러한 예외 규칙들을 다 포괄할 수 없기에 문서유형정의를 다시 수정하여야 한다.

처음에는 <표 3>의 문서유형정의를 기초로 이를 확장시키는 방안을 생각했으나, 계속해서 나타나는 예외 현상들에 대하여 문서유형정의를 확장하는 것은 사실 불합리하다. 예를 들어 <표 3>의 문서유형정의에는 '품사'와 '문형'이 문자의 연속(sequence)로 나오게 되어 있고, 그 다음 반드시 나와야 하는 문자의 연속인 '설명'(필수항목)의 하위 요소(element)로서 '어의번호'가 나오도록 되어 있으나, 실제 사전에서는 다음과 같은 예들이 발견된다.

```
strip[1]
    v.  (stripped or rarely stript, stripping)  vt.  (P6,7,13)
    품사 (동사변화)  품사  (문형)  어의번호
string
    v.  (strung)  vt.  1  (P6)
    품사 (동사변화) 품사 어의번호 (문형)〔⇒ '어의번호 (문형)'이 뒤바뀜〕
```

즉, 필수로 한 번 이상 나와야 하는 '설명'이 나오지 않고 '품사'와 '동사변화

'가 먼저 나온 다음에 다시 '품사'부터 시작하는 경우이다[20]. 또한 '어의번호'와 '문형'과의 순서 및 계층 관계가 서로 뒤바뀌어 섞여있는 모습을 볼 수 있다. 이외에도 여러 가지 다른 예외 규칙들이 발견되는데, '가산성' 표시 역시 원래대로라면 트리 구조 상 '설명'과 같은 깊이에 있고, '어의번호'는 '설명'에 속한 한 단계 하위 요소이나, '가산성' 표시가 '어의번호' 뒤에 따라 나오는 경우도 있었다. 이와 같은 문제들에 일일이 대응하기 위해서 예외의 경우를 매번 선택 요소로 추가해야 한다면, 이런 식으로 추가해야 할 규칙들이 계속 늘어날 뿐만 아니라, 이미 앞서 나온 요소들이 뒤에서 의미 없이 반복된다거나, 또는 너무 복잡하게 중첩되어 꼬여버리는 경우가 생길 수도 있다. 게다가 예외 규칙이 나올 때마다 계속해서 요소(element)를 첨가하는 것은 결국 효율적으로 구조화된 문서유형정의를 설정하는 것과는 점점 더 거리가 멀어질 수밖에 없다.

이 연구를 수행하는 과정에서 영영한 사전이 체계적인 구조를 지니지 못한 채 일관성 없이 편찬되었다는 사실이 확인되었고, 이러한 사실을 통해서 이러한 연구가 과연 의미가 있을까 하는 의구심도 갖게 하지만, 그것은 궁극적으로 사전을 편찬하기 위해서는 결국 사전편찬학적인 관점에서 체계적인 구조를 갖춘 미시구조가 반드시 필요하다는 것을 반증하는 결과를 보여 준 것이다.

따라서 이 연구에서는 인쇄사전의 텍스트에 나타난 어휘정보를 완전 자동화된 방법을 통해서 분석을 하는 것이 불가능하다 할 지라도, 완전하지는 않지만 지나치게 세분화되지 않은 구조 속에서 사전의 어휘정보를 분석하기로 한다[21].

문서 구조를 분석하기 위하여 우선 태그의 종류에 따라서 상태(state)를 구분하여 상태전이테이블을 작성한다[22]. 상태전이테이블에서는 기본적으로 태그 하나에 대하여 하나의 상태가 존재한다. 분석과정에서는 전처리과정을 거친 파일을 앞에서부터 차례로 읽어들여 배열에 저장하면서 이제까지 읽어들인 문자열을 살펴보고 상태전이(state transition)가 가능한 지를 검사한다. 전이가 가능한 경우 현재 상태에 대한 태그를 출력하고 현재까지 읽은 문자열을 출력함으로써 상태의 전이가 이루어지는 것이다. 일단 읽어들인 문자열만을 가지고 동시에 한 개 이상의 다른 상태로 전이가 가능한 경우에는 중간에 상태를 하나 더 두어서 일단은 그 중간 상태로 전이시킨 후, 그 다음 문자열을 읽어들여서 알맞은 상태로의 문자열 전이가 이루어진다.

<표 4>는 문자열 전이가 일어나는 상태전이테이블을 나타낸 것이며, <표 5>는 이 상태전이테이블을 바탕으로 출력된 XML 형식의 어휘항목의 한 예로 계층구조를 이루고 있다.

20. 이러한 경우가 동사에서 자주 발견된다.
21. 일단 정의한 문서유형정의에 부합하지 않는 것은 따로 출력을 얻어 수동으로 추가 작업을 하는 방법을 선택하였다.
22. 이 논문에서는 영영한 사전의 미시구조를 DTD를 통하여 제시하였다. 그러나 민중서림의 영영한 사전은 <표 3>에서 제시한 구조를 충족시키지 못하고 있다. 따라서 영영한 사전의 사전구조를 유지하면서도, 이 논문에서 제안한 것과 일치하지는 않지만 그래도 유사한 구조의 어휘데이터베이스를 구축하기 위하여, 그 전산처리과정을 상태전이테이블로 보여준다.

현재 상태	상태 설명	다음 상태
0	처음 시작	1, 2, 22
1	단어의 중요도 표시	2
2	표제어	3, 6, 9, 10, 36
3	발음기호	2, 4, 12
4	품사	6, 7, 9, 12, 26, 34, 36
5	가산성 명사 구분	6, 7, 9, 12, 34, 36
6	영어 단어 설명	1, 2, 7, 20, 34, 36
7	한글 단어 설명	1, 2, 8, 9, 13, 22, 35, 36
8	어원	0
9	큰 설명 번호 (1, 2, 3, ...)	6, 12, 36
10	단어 번호	3, 6, 12, 36
11	어형변화(명사 복수형, 동사의 과거 과거분사, 형용사부사의 비교최상급)	32
12	특수용법, 품사설명, 관계있는 전치사부사	31
13	작은 설명 번호	6, 12, 36
15	동의어	16, 17, 33
16	반의어	15, 17, 33
17	참고어	16, 17, 33
18	어법	1, 2, 8, 9, 22, 34, 35, 36
19	참고주의사항	1, 2, 8, 9, 22, 34, 35, 36
20	용례(영어)	1, 2, 8, 9, 21, 22, 35, 36
21	용례(한글설명)	1, 2, 8, 9, 20, 22, 35, 36
22	관용구	23, 24, 26
23	관용구 영어 설명	23, 24, 26
24	관용구 설명의 구분 번호	23
25	문형	33
26	해당 사항 있음()	1, 2, 22, 35
271	파생어	272, 273
272	파생어 발음	273
273	파생어 품사	1, 2, 8, 271, 274
274	파생어 셀수있는/없는 명사구분	1, 2, 8, 271
28	전문어, 언어용법	37
29	관용구 한글 역어	1, 2, 22, 24, 35, 36
31	} 이후의 상태	6, 9,12
32	) 이후(어형변화후)의 상태	4, 6, 9, 12, 36

현재 상태	상태 설명	다음 상태
33	) 이후의 15, 16, 17에서 나온 상태	1, 2, 8, 9, 13
34	(이후의 15, 16, 17로 갈 수 있는 상태	15, 16, 17
35	파생어 loop로 들어가기 위한 상태	271
36	AMin font들의 위한 상태	5, 13, 18, 19, 271, 28, 38, 39
37	))이후의 상태	6
38	용례 loop로 들어가기 위한 상태	20
39	- 이후의 상태	3, 4

[표 4] 상태전이테이블

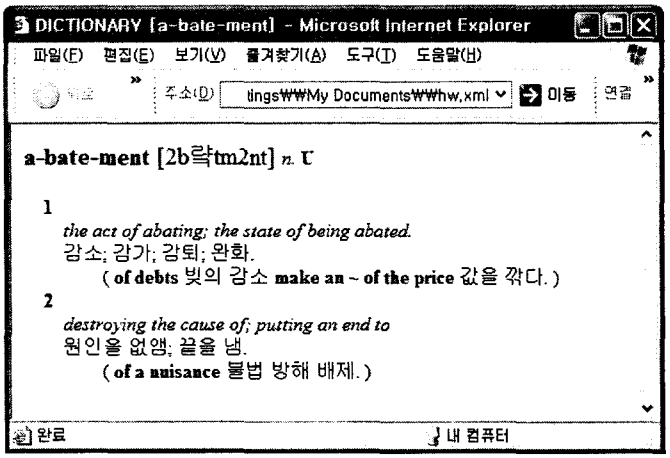
```
<어휘항목>
  <표제어>a-bate-ment </표제어>
  <발음> [əbéitmənt] </발음>
  <풀이>
    <품사>n. </품사>
    <가산성>U </가산성>
    <설명>
        <어의번호>1 </어의 번호>
        <영어설명>the act of abating; the state of being
                                abated. </영어 설명>
        <한글역어>감소; 감가; 감퇴; 완화. </한글 역어>
        <설명-용례구>
            <설명-용례>of debts </용례>
            <설명-용례역어>빚의 감소</설명-용례역어>
        </설명-용례구>
        <설명-용례구>
            <설명-용례>make an ~ of the price </설명-용례>
            <설명-용례역어>값을 깎다. </설명-용례역어>
        </설명-용례구>
    </설명>
    <설명>
        <어의번호>2 </어의 번호>
        <영어설명>destroying the cause of; putting an end
                                to </영어 설명>
        <한글역어>원인을 없앰; 끝을 냄. </한글 역어>
        <설명-용례구>
            <설명-용례>of a nuisance </설명-용례>
            <설명-용례역어>불법 방해 배제. </설명-용례역어>
        </설명-용례구>
    </설명>
  </풀이>
</어휘항목>
```

[표 5] 상태전이테이블의 결과

< 5 >의 출력 결과에서 발음 기호의 경우에 발음기호를 원래 EST파일에 있는 형태 그대로 출력하였는데, 이는 개별 발음기호를 위한 문자태그에 대응하여 각각 발음기호를 변환하는 작업이 또 필요하기 때문에 여기서는 생략하였다.

구조화 과정을 거쳐 <표 5>에서와 같이 표현된 XML형식의 데이터는 XML

문서의 출력을 제어하는 XSL(eXtensible Stylesheet Language) [23] 스타일시트를 이용하여 <그림 5>에서와 같이 웹브라우저에서 그 내용을 확인할 수 있다.



[그림 5] XSL을 이용한 어휘항목의 표현

그러나 <표 5>에서와 같은 분석 결과가 나옴에도 불구하고, 분석이 완전하게 이루어지지 못한 아쉬움이 남는다. 우선 일부 어휘항목의 경우에는 문자유형 태그의 뒤에 불필요한 문자열들이 따라 나오는 경우가 종종 발생하여 상태전이 테이블을 이용한 분석을 불완전하게 한다. 그 이유는 이미 앞에서도 언급했듯이 사전편찬을 위하여 처음부터 특정한 메타 태그에 의해서 사전의 미시구조를 설정하고 사전 텍스트가 작성된 것이 아니라, 출판을 위한 목적으로 문서편집의 형식으로 사전이 제작되었기 때문이다.

또한 어휘항목에서 문서유형정의에 따라 사전 설명이 작성되지 않은 경우 상태의 전이가 이루어지지 않고, 그 뒤의 다른 어휘항목에 대해서도 제대로 분석이 이루어지지 않는 경우가 발생하는데, 이러한 문제는 모든 상태에 대하여 어휘항목의 시작을 선택하는 상태로 갈 수 있게 만든다면, 한 어휘항목의 설명에서 나타난 에러가 다음 어휘항목에 영향을 미치는 것을 방지할 수 있을 것이다.

앞으로 보다 완전하고 지속적인 작업을 위하여서는 EST파일을 분석하여 분석이 정상적으로 이루어진 것과 불완전하게 이루어진 것을 각각 다른 파일로 저장한 후, 불완전한 분석이 이루어진 것에 대하여 다시 데이터를 분석하면서 문제를 해결해 나가는 방법을 취해야 할 것이다.

4. 결론

민중서림의 영영한사전이 체계적으로 구성된 기계가독형사전이라고 할 수는 없지만 그래도 개별 표제어에 대한 세부적인 어휘정보의 구분에 도움을 줄 수 있는 문자타입정보를 포함하고 있기에, 이 연구에서는 이러한 정보를 이용하여 일

23. <그림 5>를 위한 XSL의 자세한 내용은 부록을 참조.

반사전으로부터 어휘정보를 세분화된 형태로 추출하는 과정을 제시하면서 전산사전편찬학의 실제적인 측면을 간략히 소개하였다.

어휘정보를 데이터베이스화하려는 본 연구자의 계획은 실제로 작업을 하는 과정에서 많은 어려움을 겪을 수밖에 없었다. 그것은 우선 영영한 사전이 사전편찬학을 위한 체계적이고 계획된 과정 속에서 특정한 기술언어(markup language)를 바탕으로 편찬된 것이 아니며, 또한 어휘정보의 기술에 있어서도 사전편찬의 철저한 기본원칙 하에 이루어지지 않았고, 설령 그렇다 할지라도 개개인의 직관과는 완전히 독립적으로 이루어졌다고 판단하기는 어렵기 때문이었다.

따라서 이처럼 비체계적인 구조 아래에서 어휘정보가 기술된 사전으로부터 완전히 자동화된 방법으로 어휘정보를 구조화하여 추출하는 것은 사실상 불가능하다.

그러나 이러한 시도를 하는 이유는 어휘정보를 구축하는데 드는 많은 시간과 노력을 고려해 볼 때, 기존에 구축된 어휘정보의 재원을 재활용할 수 있고, 이를 바탕으로 전자사전의 어휘정보를 확장해 나간다면, 우리는 보다 많은 시간을 절약할 수 있고 과거에 이루어졌던 연구성과를 보다 효율적으로 이용하게 될 것이기 때문이다. 또한 이러한 과정을 통하여 사전편찬을 위한 미시구조의 설정이 얼마나 중요한지를 다시 한 번 확인하면서, 앞으로 사전편찬작업의 기준을 정하는데 도움을 줄 수 있다고 생각한다.

그러나 앞에서도 지적한 바와 같이 과연 이러한 형태의 어휘데이터베이스를 구축하는 일이 과연 얼마나 의미 있는 일인지, 그리고 앞으로 사전의 효율적인 이용을 고려할 때, 현재 진행되고 있는 전자사전 구축작업의 방향이 옳은 것인가에 대하여 다양한 측면에서 엄밀한 평가가 이루어져야 할 것이다.

참고문헌

최병진 외. 1996. 기계가독형 사전 구축을 위한 사전 항목의 논리 구조. *인지과학 제7권 2호*, 7(2).

노용균. 2001. Converting a dictionary into xml: The process and some lessons. *2001년 하계 전국 학술 발표대회, 언어과학회*.

Alshawi, H. 1989. Processing dictionary definitions with phrasal pattern hierarchies. In *Boguraev, B. and E. Briscoe(eds.)*. pages 153–170.

Boguraev, B. 1991. Special issue on computational lexicons. *International Journal of Computational Lexicography 4*.

Boguraev, B. 1994. Machine-readable dictionaries and computational linguistics research. In *A. Zampolli, N. Calzolari, and M. Palmer(eds.)*.

Boguraev, B. and Briscoe E. 1989. *Computational Lexicography for Natural Language Processing*. Longman Limited, Harlow and London.

Boguraev, B. / Levin, B. 1993. Models for lexical knowledge bases. In *Pustejovsky(ed.): Semantics and the Lexicon*. Kluwer Academic Pub.

Copestake, A. 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. In *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*. Tilburg, The Netherlands:19-29.

Hausmann, F. J. 1985. Lexikographie. In *Christoph Schwarz/Dieter Wunderlich(Hrsg.): Handbuch der Lexikologie*. Königstein- Ts: Athenäum 1985.

Neff, N. & Boguraev, B. 1989. Dictionaries, dictionary grammars and dictionary entry parsing. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.* Vancouver, British Columbia: 91-101.

Wahrig, G. 1967. *Neue Wege in der Wörterbucharbeit.* Hamburg.

Wiegand, Herbert Ernst. 1983. Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. *Germanistische Linguistik 1-4/82: 401-474.*

5. <부록>

```
<?xml version="1.0"
encoding="EUC-KR"?> <xsl:stylesheet
xmlns:xsl="http://www.w3.org/TR/WD-xsl">

<xsl:template match="/">
  <xsl:apply-templates />
</xsl:template>

<xsl:template match="어휘항목">
  <HTML>
    <HEAD>
     <TITLE>DICTIONARY [<xsl:value-of select="표제어" />]</TITLE>
    </HEAD>
    <BODY>
       <SPAN style="font-size: 15pt; font-weight: bold">
          <xsl:value-of select="표제어" />
      </SPAN>
      <SPAN style="font-size: 15pt">
          <xsl:value-of select="발음" />
      </SPAN>
      <SPAN>
          <xsl:apply-templates select="풀이" />
       </SPAN>
     </BODY>
  </HTML>
</xsl:template>

<xsl:template match="풀이">
  <SPAN style="font-style:italic">
     <xsl:value-of select="품사" />
  </SPAN>
  <SPAN>
     <xsl:value-of select="문형" />
  </SPAN>
  <SPAN style="font-weight:bold">
     <xsl:value-of select="가산성" />
  </SPAN>
  <BR/><BR/>
  <DIV style="padding-left: 1em">
     <xsl:apply-templates select="설명" />
  </DIV>
  <BR/>
```

```
  <DIV style="padding-left: 1em">
     <xsl:apply-templates select="관용어구" />
  </DIV>

</xsl:template>

<xsl:template match="설명">
  <DIV style="font-weight:bold">
     <xsl:value-of select="어의번호" />
  </DIV>
 <DIV style="padding-left: 1em">
```

접수일자: 2002년 11월 11일
게재결정: 2002년 12월 15일