한국어 사전의 논리적 구조분석과 변환
- 연세 한국어사전을 대상으로 -

최운호
서울대학교
whchoi70@freechal.com

1. 머리말

이 연구는 한국어 사전의 구조 분석을 통하여 출판을 목적으로 편찬된 사전을 어휘데이터베이스로 변환 구축하는 것을 목적으로 하며 사전의 분석과 변환 과정에서 발견된 문제점들에 대한 논의를 통해서 사전편찬과정에서 발생할 수 있는 오류와 문제점을 자동 검출하는 방안과 사전편찬과정을 자동화하는 방안에 대해 논의해 본다.

컴퓨터를 이용한 사전 편찬이 보편화되면서 사전 편찬 원고도 기계가독형 원시 파일의 형태로 존재하는 것이 일반화되었으며, 최근 편찬된 한국어 사전들은 웹을 통해 서비스 되거나 워드프로세서에 탑재되는 등 전자사전화 되어 있다. 전자화된 사전에 대한 접근성과 활용성을 높이려면 구조화된 형식으로 검색, 추출, 변환 및 재가공이 가능해야 하며, 이렇게 가공된 자료는 자연언어처리를 위한 언어자원으로 재사용될 수 있다[1].

본 연구에서는 연세대학교 언어정보개발연구원에서 편찬한 「연세 한국어사전」의 49,560 표제어(부표제어 포함 53,355)를 대상으로 삼아서, 이 자료의 계층 구조를 분석하여 구조화하였고, 관계형 데이터베이스 관리시스템을 이용하여 자료를 어휘데이터베이스로 구축하였다.

---

1) 목적에 따라서는 한국어 교육을 위한 자료로도 활용될 수 있을 것이다.

2. 선행 연구 소개 및 연구방향

사전의 구조 분석과 변환, 그리고 그러한 연구 결과에 기반한 사전의 언어자원화와 관련한 연구는 1) 사전의 구조 분석에 따른 사전 구조의 형식적 정의 및 변환에 대한 연구와 2) 구조화된 사전을 바탕으로 자연언어처리에 이용 가능하도록 어휘 의미를 분석하는 연구로 나눠볼 수 있다.

2.1. 사전의 구조 설계 및 분석에 대한 연구

사전의 구조 설계에 대한 연구는 컴퓨터를 이용한 사전편찬이 활성화되기 이전에는 사전편찬작업과 동시에 수행하기 어려운 점이 있다. 특히 대규모의 언어 사전을 편찬하는 경우 사전편찬작업 전체를 관리하는 시스템을 개발하기란 쉽지 않은 일이다. 사전의 논리 구조가 초기 원고 집필과 함께 유동적으로 변경되는 경우에는 특히나 고정된 사전편찬 통합 환경 시스템[2]을 구축하기란 어려운 일이다. 따라서 대부분의 연구는 사전을 기술하기 이전에 사전의 논리 구조에 대한 제안과 원형에 대한 연구, 또는 이미 구축된 사전의 분석을 통한 어휘정보 데이터베이스[3] 구축에 대한 연구로 집중된다.

양단희(1992)에서 제시한 '한국어 전자 사전 원형의 설계 및 구현'에는 하이퍼텍스트를 이용한 전자사전의 기능 및 요건 정의를 상세히 기술하고 있으며, 사전편찬을 위한 저작도구까지 기술되어 있다. 양단희(1992)에서 정의된 기능 및 요건은 구현상의 차이만 있을 뿐 현 시점에서도 컴퓨터를 이용한 사전 저작 도구 및 전자사전이 갖춰야 할 본질적인 내용에서는 차이가 없어 보인다[4].

---

2) 사전의 구조가 상대적으로 복잡하지 않고 많은 양의 용례 가공과 주석 편찬 위주의 내용으로 채워지는 경우 편찬작업 이전에 쉽게 구조를 확정할 수 있다. 한국 판소리문학에 대한 사전 편찬 저작 도구에 대해서는 김동건 외(2003)을 참조할 수 있다.

3) 어휘데이터베이스(LDB, Lexical Database)와 어휘지식베이스(LKB, Lexical Knowledge-Base)의 구분에 대해서는 최병진(2002) 참조.

4) 양단희(1992)에서는 사전편찬저작도구에 대한 구현을 표준 C 언어를 이용한 응용프로그

최근의 연구로는 노용균(2001)과 최병진(2002)를 들 수 있다.
노용균(2001)에서는 BBI 영어사전을 XML 문서로 변환하는 과정을 단계별로 보여 주고 있으며[5], 어떤 자료를 XML로 유지해야 하는가, 그리고 왜 사전을 XML로 변환하는가에 대한 이유를 제시하고 변환된 사전의 활용 방안에 대해서 기술하고 있다.
최병진(2002)에서는 민중서림의 영영한사전을 어휘데이터베이스로 변환하는 과정을 단계별로 보여 주고 있는데, EST(Extended Style Tag) 파일[6]로 저장된 사전 파일을 문서 구조화 과정을 거쳐 XML 형식으로 변환하는 과정을 보여 주고 있다.
외국의 연구사례를 보면, 다양한 사전에 대해서 전자사전화 하는 연구들이 수행되었고[7], 객체지향 데이터베이스 관리시스템을 이용한 어휘데이터베이스 구축 연구도 볼 수 있다[8].

2.2. 사전 정의문 분석에 대한 연구

대용량 어휘데이터베이스(large-scale lexical database)가 구축된 경우, 사전의 뜻풀이 기술에 사용된 뜻풀이 정의문(definition sentences)의 분석을 통하여 어휘 지식을 추출할 수 있으며, 자연언어처리를 위한 자료를 구축할 수도 있다.
조평옥 외(1999)에서는 사전 뜻풀이말의 분석을 통하여 한국어 명사의 의미계층구조를 구성하였으며, 이수광 외(2001)에서는 한국어

램으로 구현한 것으로 제시되어 있다. 현 시점에서 그러한 기능을 구현한다고 해도 DBMS(데이터베이스 관리시스템)를 이용하고 네트워크 기능을 추가하는 차이만 있을 뿐 기본적인 요소는 크게 달라지지 않으리라 본다.

5) 노용균(2001)에서는 digitization > detection & correction of typo-graphic errors > Document Type Definition & Grammar writing > Translation of parsed trees into XML 4단계로 작업을 정의하고 각 과정을 상세히 보여주고 있다.

6) EST 파일 포맷은 노용균(2001), Ide *et al.*(1994)에서 설명된 typo-graphic markup의 한가지로 간주할 수 있다.

7) Guthrie et al.(1996)은 다양한 프로젝트에서 수행된 대용량 사전 관련 작업들을 일목요연하게 제시하고 있다.

8) Ide *et al.*(1994)에서는 *Zyzomys* CD-ROM 사전의 50,000여 표제어를 객체지향 데이터베이스관리시스템인 IBM O2를 이용해서 어휘데이터베이스를 구축하였다.

사전의 뜻풀이말을 대상으로 한 구문분석기를 개발하였고, 허정 외(2001)에서는 사전 뜻풀이말에서 추출한 자료를 이용하여 동형어 중의성 해소 시스템을 개발했다.
Fontenelle(1997)은 2개어 사전을 분석하면서 사전 정의문의 뜻풀이 패턴을 분석하고 Melc'uk의 MTT(Meaning Text Theory)에 기반한 어휘데이터베이스 시스템을 구축하였다[9].

2.3. 자료의 소개 및 변환 작업 개요

본 연구에서 분석의 대상으로 삼은 자료는 연세대학교 언어정보개발연구원에서 편찬한 「연세한국어사전」[10]이며, 자료의 구축은 웹 서비스를 통해 제공되는 자료[11]를 사용하였다. 웹 리소스 자료는 출판본 자료와 차이가 있는데, 출판본에서 '날개' 위치에 포함되어 있는 정보인 격틀 정보, 논항 정보, 의미 항목 관련어 정보, 파생어 정보, 참고 정보, 연어 정보는 웹 리소스에 표시되어 있지 않다. 웹 리소스에 표시되지 않은 정보는 이번 연구의 자료에 포함되지 않았다. 「연세 한국어 사전」의 변환 과정은 <그림 1>과 같다.

9) Fontenelle(1997)은 영불 2개어 사전(bilingual dictionary)을 분석하였다. 사전의 뜻풀이 정의문을 뜻풀이 패턴을 이용해서 분석하는 연구에 대한 상세한 기술은 Barnbrook(2002)를 들 수 있다.

10) 연세한국어사전 초판 6쇄

11) 웹을 통해 서비스되는 「연세 한국어 사전」은 '웹 리소스'로, 출판되어 나온 「연세 한국어 사전」은 '출판본'으로 부르기로 한다.

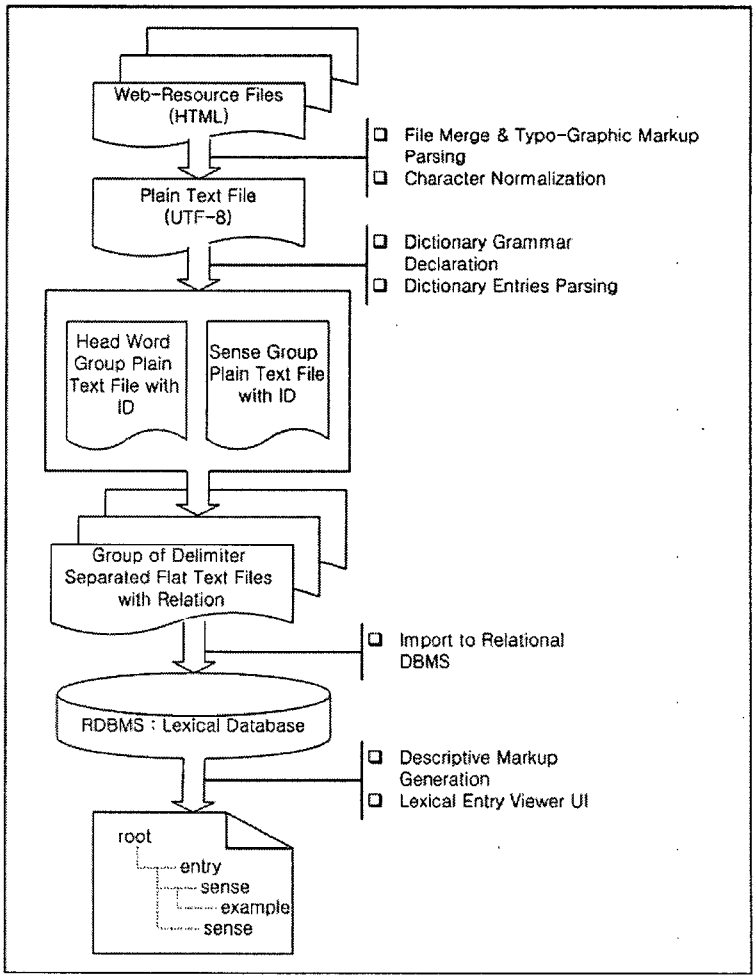

그림 1. 연세 한국어 사전 구조 분석 및 변환 작업 흐름

<그림 1>에 제시된 것처럼 변환 작업은 단계별로 진행되며, 웹 리소스 파일들을 수집한 뒤 작업의 편의를 위해 표제어의 품사, 발음, 활용형태 등과 같은 머리어 정보와 뜻풀이 구획 정보를 분리하여 각각 변환 작업을 병행해서 수행하였다. 변환 과정의 목표 구조는 개개의 사전 정보를 구분자로 분리되어 있고 관계(relation)가 설정되어 있는 파일로 만드는 것이고, 이렇게 변환된 파일은 관계형 데

이터베이스 시스템에 탑재해서 재사용 가능한 어휘 데이터베이스를 구축하는 데 사용된다.
노용균(2001)과 최병진(2002)에서 원시 파일(source file)을 구조 분석 과정을 거쳐서 직접 목적 파일(object file) 구조인 XML 형식으로 변환하는 방법을 제안하고 시도했지만, 이번 작업에서는 원시 파일을 분석하여 관계형 데이터베이스 시스템으로 관리 가능한 자료로 변환하는 것을 목표로 한다.

3. 사전의 구조 분석 및 변환

3장에서는 「연세 한국어 사전」의 구조를 설정하고 변환하는 과정에 대해서 설명한다. 사전의 구조는 계층적인 성격과 상호참조의 성격을 함께 지니고 있는 구조로 먼저 구조를 설정한 변환 과정에서 나온 자료의 소소한 오류들을 제시하고 변환 결과에 대해서 설명하도록 한다.

3.1. 「연세 한국어 사전」의 기본 구조

「연세 한국어 사전」에서 설정한 정보구획과 각각의 정보구획에서 담고 있는 내용에 대해서는 「연세 한국어 사전」의 '일러두기'와 이상섭(1992)에 상세히 설명되어 있으며, 이를 계층적으로 분석해 보면 <그림 2>와 같은 기본 구조를 설정할 수 있다.
표제어 구획에서 동형어 번호는 다른 동형어가 있는 경우에만 존재한다. 자료 변환 과정에서 동형어 번호는 1에서 21까지 할당되어 있으며, 검색 및 자료 관리의 편의를 위해 동형어 구분이 없는 경우에는 '0'을 할당했다. 한자어에서 한자를 표기해야 할 필요가 있는 경우에는 한자가 괄호 안에 병기되어 있는데, 한자 표기가 복수인 경우 현재는 구분하지 않고 하나의 정보로 관리하고 있다. 발음 및 활용에 관한 정보도 현재는 한 표제어에 대해 기술된 구획 정보로 분리하지 않고 관리되고 있다.
뜻풀이는 표제어에 대해서 다른 표제어 항목을 참조하도록 정의된

항목과 해당 표제어항에서 뜻풀이 정의문으로 기술된 항목으로 구분할 수 있다. 전자의 경우 '가 보라' 기호 '☞'를 사용하여 기술되어 있다. '가 보라' 기호만으로 뜻풀이가 기술되어 있는 경우에는 뜻풀이를 '참조' 정보로 분리하여 자료를 변환하였다. 「연세 한국어 사전」의 경우 뜻풀이의 위계 구조는 4단계로 정의되어 있다. 각각의 단계는 I, ①, ㄱ, (1) 기호를 사용한다.

부표제어 구획은 부표제어와 뜻풀이로 간략하게 정의되어 있으며, 뜻풀이를 구분해야 할 경우 ①, ② 등의 원문자만으로 구분하고 있다12).

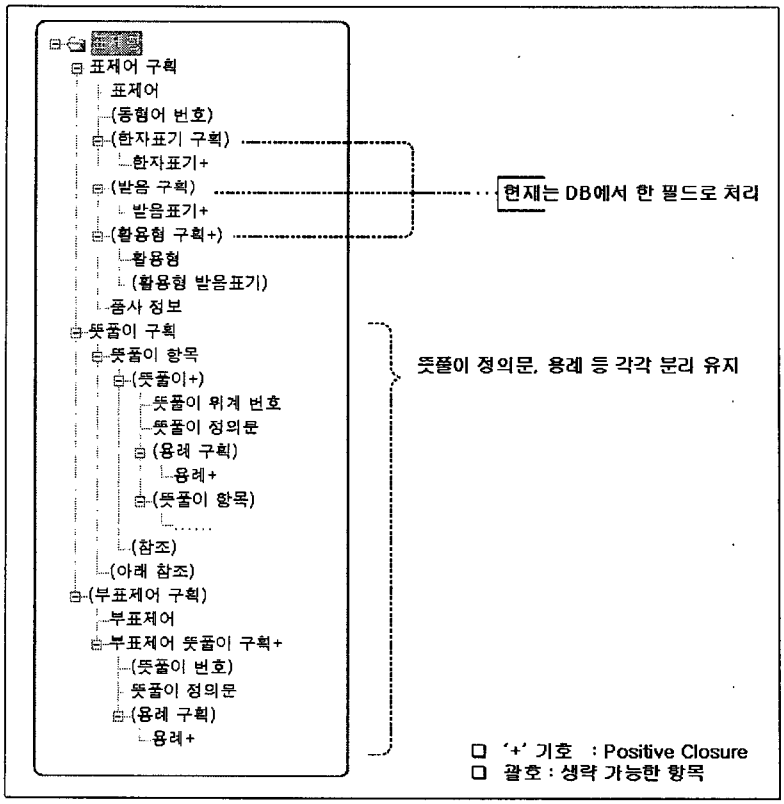

그림 2. 「연세 한국어 사전」의 기본 구조

12) 3,791개의 부표제어 중에서 뜻풀이를 구분하여 기술한 부표제어는 191개이다.

<그림 2>에서 설정된 기본 구조는 「연세 한국어 사전」의 구조를 정확히 반영하지는 않는다. 그림 구조의 특성상 뜻풀이의 계층적인 구조에 대해서 정확히 반영되어 있지는 않으며, EBNF[13]로 표기하면 <표 1>과 같다.

```
entry          ::= hwInfoGroup senseGroup
subentryGroup
hwInfoGroup    ::= entryNotation polyNo? hanjaInfo?
                   pronInfo? InflGroup? pos
hanjaInfo      ::= hanja+
pronInfo       ::= pron+
InflGroup      ::= (inflForm inflPron?)+
senseGroup     ::= refGroup | senseDef glassInfo?
senseDef       ::= simpleDef | nestedDef
simpleDef      ::= definitionString
                   | definitionString examples
examples       ::= exmpleSentence+
nestedDef      ::= defNo nestedDef
                   | defNo definitionString examples
glassInfo      ::= string
refGroup       ::= (entryNotation polyNo)+
subentryGroup ::= entryNotation senseGroup
```

표 1. 「연세 한국어 사전」의 기본 구조 : EBNF

3.2. 변환

13) Extended BNF(Backus-Naur Form). EBNF는 Context-Free Grammar를 표기하는 방법인 Backus-Naur Form에 3개의 CLosure 연산자(?, +, *)를 추가한 표기법이다. EBNF에 대한 표준은 "ISO/IEC 14977:1996"을 참조.

변환 과정은 웹 리소스로 존재하는 파일을 어휘데이터베이스로 구축할 수 있는 구조화된 파일로 변환하는 작업이다. 변환 과정은 연구자의 판단과 작업 환경에 따라서 자료의 성격과 연구자의 개발 방향에 적합한 다양한 방법이 시도되었다. 변환한 결과도 연구자(연구팀)의 작업 성격에 따라서 여러 방법이 시도되었다.

Ide *et al.*(1994)에서는 객체지향 데이터베이스에 사전을 탑재하는 장점에 대해서 논의되었다. Fontenelle(1997)에서는 관계형 데이터베이스에 어휘 의미 데이터베이스를 구축하였으며, 뜻풀이 문장의 분석을 통해서 어휘 의미 관계를 구조화하였다. 노용균(2001)에서는 프롤로그 언어를 개발 언어로 사용하여 140여개의 규칙[14]을 작성하고 프롤로그로 구현된 차트 파서를 이용, BBI 사전을 XML 문서로 생성하는 과정에 대해서 기술하고 있으며, 최병진(2002)에서는 민중서림 영영한 사전 EST 파일 태그를 분석 변환하기 위하여 유한상태변환 방식[15]을 사용하였다.

본 연구에서 사용한 웹 리소스 원시 파일은 HTML 태그가 부착된 파일이며, 다양한 HTML 태그를 사전의 구조를 밝히는 단서로 사용하여 변환 작업을 거쳤다. HTML 태그 중 의미를 구분하는 기호 부분에 사용된 이미지 태그는 사전의 의미 구분 기호로 변환하였으며, HTML 태그에 나타난 단서를 최대한 이용하여 1차 변환 과정을 수행하였고, 그 결과는 <표 2>와 같다. <표 2>에서 표제어 구획은 구분자 '##!'로 분리되어 있으며 뜻풀이 구획은 전체가 하나의 묶음으로 이루어져 있다.

---

14) 노용균(2001)에서는 140여개의 CFG 규칙을 프롤로그 언어로 구현하여 차트 파서를 작성하였다.

15) 최병진(2002)에서 유한상태변환기(FST: Finite State Transducer)를 구성하였다는 설명이 명시적으로 기술되어 있는 것은 아니지만, EST 파일의 태그를 인식하는 단계를 각각의 상태(state)로 구성하고 각 상태마다 변환 규칙을 적용하는 상태 변환 테이블을 보여주고 있다.

entry##!3##!3##!0##!가##!2##!可##![가 :]##!##!명##! ((1)) 의견이나 안건에 대해 찬성한다는 표시. @@주어진 논제로서는 가, 부의 두 방향으로밖에 생각할 수가 없는 문제점이 있다./@@ ((2)) 학교에서 학생들의 성적을 '수우미양가'의 다섯 가지로 나누어 매길 때 가장 낮은 등급. @@미술이나 음악은 수를 받을 정도로 우수했지만 산수는 언제나 가였다./@@

표 2. 「연세 한국어 사전」 '가2'의 1차 가공 결과물

<표 2>와 같이 가공된 파일에서 표제어 구획과 뜻풀이 구획은 표제어 구획에 있는 일련번호를 서로 공유하며, 각각 분리되어 개별적인 변환 과정을 거친다.

2##!2##!00##!가##!1##!##!##!##!명
3##!3##!00##!가##!2##!可##![가 :]##!##!명

표 3. 「연세 한국어 사전」 표제어 구획 텍스트 파일

<표 3>은 <표 2>에서 제시된 것과 같이 HTML 태그를 제거하고 1차 전처리 가공을 거친 자료에서 표제어 구획 정보만 분리한 자료이다. 각 필드는 '##!' 문자열을 구분자로 사용하여 구분하고 있으며 각 필드의 의미는 다음과 같다.

필드 번호	의미
F1	표제어 항목 일련번호. 표제어/부표제어 구분없이 매겨진 일련 번호
F2	표제어 고유번호. 부표제어에는 해당 부표제어의 상위 표제어 고유번호가 매겨져 있다.
F3	부표제어 번호. 한 표제어에 여러 개의 부표제어가 있을 경우 해당 표제어 내에서 각 부표제어에 매겨진 일련 번호. 기본값 '00'인 경우에는 표제어, 그렇지 않은 경우에는 부표제어.
F4	표제어
F5	동형어 번호. 동형어 번호가 없는 경우 '0'이 기본값으로 할당되어 있다.
F6	한자 표기
F7	발음 정보
F8	활용 정보
F9	품사 정보

표 4. 표제어 구획 필드 설명

<표 5>는 뜻풀이 구획을 분리해서 들여쓰기(indentation)를 적용하여 가공한 파일로, 뜻풀이 구획에 사용된 '뜻풀이 갈래 번호', 정의문(definition sentence), 용례 등을 각각 구분할 수 있도록 주석을 첨가하였다. <표 5>는 관계형 데이터베이스에 입력할 수 있는 자료 형식은 아니기 때문에, <표 6>, <표 7>과 같은 형식으로 변환하는 과정을 다시 거쳤으며, 뜻풀이 정의문과 용례를 별도로 분리하여 가공하였다[16].

```
serial@@!3@@!3@@!00
        poly@@!L2@@!1
            definition@@!의견이나 안건에 대해 찬성한다는 표시.
                ex@@!주어진 논제로서는 가, 부의 두 방향으로밖에 생각할 수가 없는 문
                        제점이 있다.
        poly@@!L2@@!2
            definition@@!학교에서 학생들의 성적을 '수우미양가'의 다섯 가지로 나누어 매
                        길 때 가장 낮은 등급.
```

16) 뜻풀이 정의문과 용례를 분리하여 별도로 가공한 이유는, 관계형 데이터베이스 내에서 뜻풀이 정의문과 용례를 별도의 테이블로 관리하려는 목적 때문이다.

ex@@!미술이나 음악은 수를 받을 정도로 우수했지만 산수는 언제나 가였다.

표 5. 「연세 한국어 사전」 '가2' 뜻풀이 구획 텍스트 파일.

3##!3##!00##!0##!1##!0##!0##!3##!의견이나 안건에 대해 찬성한다는 표시.
3##!3##!00##!0##!2##!0##!0##!4##!학교에서 학생들의 성적을 '수우미양가'의 다섯 가지로 나누어 매길 때 가장 낮은 등급.

표 6. 「연세 한국어 사전」'가2' 뜻풀이 정의문 텍스트 파일

3##!3##!00##!0##!1##!0##!0##!3##!0##!주어진 논제로서는 가, 부의 두 방향으로밖에 생각할 수가 없는 문제점이 있다.
3##!3##!00##!0##!2##!0##!0##!4##!0##!미술이나 음악은 수를 받을 정도로 우수했지만 산수는 언제나 가였다.

표 7. 「연세 한국어 사전」'가2' 용례 텍스트 파일

<표 6>, <표 7>에서 정의문과 용례를 제외한 앞부분 필드의 의미는 <표 8>과 같다.

필드 번호	의미
F1	표제어 항목 일련번호. 표제어/부표제어 구분없이 매겨진 일련 번호
F2	표제어 일련 번호
F3	부표제어 번호
F4	로마자 숫자(I, II, ...)로 매겨진 뜻풀이 번호
F5	원문자(①, ②, ...)로 매겨진 뜻풀이 번호
F6	한글 자모(ㄱ, ㄴ,...)로 매겨진 뜻풀이 번호
F7	괄호 문자((1), (2),...)로 매겨진 일련 번호
F8	뜻풀이 정의문 일련 번호. 모든 뜻풀이 정의문에 대해 매겨진 일련 번호이다.
F9	<표 7> 용례 파일에 있는 9번째 필드는 한 개의 뜻풀이 정의문에 여러 개의 용례가 있을 때, 각 용례를 구분하기 위해 매겨진 번호이다.

표 8. 표제어 구획 필드 설명

위에서 제시된 변환과정은 Perl과 AWK[17]의 정규표현과 필드 처리 기능을 이용하여 주로 처리되었으며, 파일의 문자 집합은 UTF-8 코드를 사용하였다. UTF-8을 사용한 이유는 파일을 변환하는 중간에 한자나 기호에 대한 정보가 손실되지 않도록 하기 위해서였다.

3.3. 교정 과정

「연세 한국어 사전」을 변환하여 어휘 데이터베이스를 구축하는 과정에서 몇 가지 오류가 발견되어서 발견된 오류는 수정하여 데이터베이스 구축용 텍스트 파일을 만들었다. <그림 3>과 <그림 4>는 「연세 한국어 사전」[18]에서 발췌한 내용으로, <그림 3>에는 로마자 뜻풀이 갈래 번호 'I'이 누락되어 있으며, <그림 4>에는 원문자 뜻풀이 갈래 번호 '①'이 누락되어 있음을 볼 수 있다. 웹 리소스 파일을 변환하는 과정에서 발견된 뜻풀이 구조상의 오류는 출판본 「연세 한국어 사전」과 대조하여 하나하나 검토하였다.

---

17) 모든 작업은 PC에서 이루어졌으며, Perl과 AWK는 Cygwin 패키지에 포함된 GNU Software의 perl과 gawk를 사용하였다.

18) 「연세 한국어 사전」 초판 6쇄(2002년 1월)

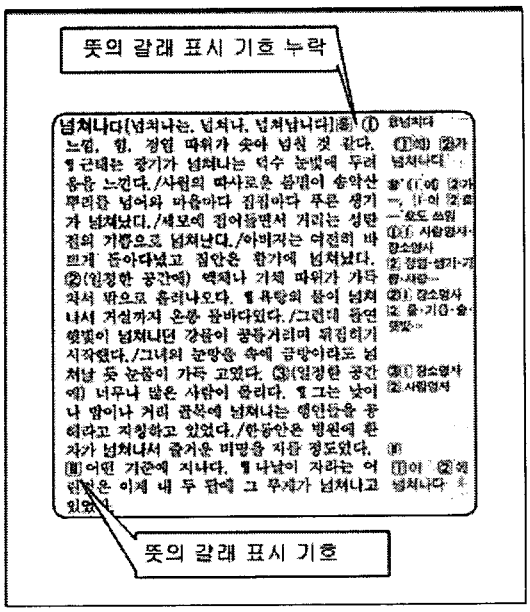

그림 3. 뜻풀이 갈래 번호 오류(로마자)

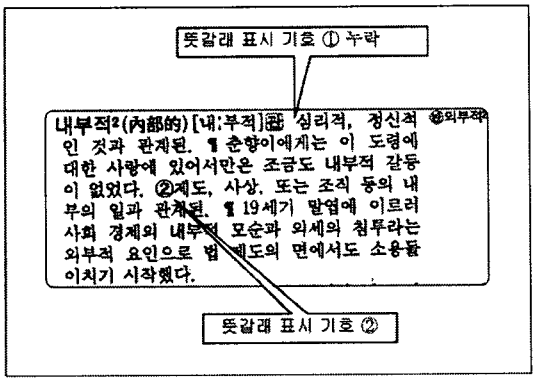

그림 4. 뜻풀이 갈래 번호 오류(원문자)

뜻풀이 구조의 오류 발견 절차는 사전을 하나의 형식적인 구조로 간주하고 사전 구조 분석기를 제작하여 <표 5>에서 보인 변환 단계의 중간 파일을 분석하는 과정에서 발견되었다[19].

표제어	웹	출판본	오류 내용
넘쳐나다	X	X	로마자 I 누락
돌아들다	X	O	로마자 I 누락
들다1	X	X	로마자 I 누락
로	X	O	로마자 I 누락
마디	X	X	로마자 I 누락
세다3	X	X	로마자 I 누락
알다	X	O	로마자 I 누락
차이다	X	X	로마자 I 누락
치우다1	X	X	로마자 I 누락
치환하다	X	O	로마자 I 누락

표 9. 로마자 뜻풀이 갈래 번호 오류

<표 9>는 로마자로 표시하는 뜻풀이 갈래 번호 'I'이 없이 'II' 이후의 갈래 번호가 등장하는 오류가 발견된 표제어 목록이다. '웹' 칸에 'X' 기호가 표시된 것은 웹 리소스에서 오류가 발견된 것이며 '출판본' 칸에 'X'가 표시된 것은 출판본 사전에서도 그러한 오류가 발견되었다는 것을 뜻한다.

동형어 번호와 관련해서, 동일한 형태의 표제어 항목에 대해서 동형어가 누락된 항목도 발견되었다. '여기(대명사)'와 '여기2(부사)'의 경우, 동형어 번호 2는 부여되어 있지만 '여기(대명사)'에 동형어 번호 1이 부여되어 있지 않다. 이 오류는 웹 리소스에서만 나타난다. 또한, '여기'와 '여기2'의 경우 출판본과 웹 리소스가 내용의 차이를 보인다.

<그림 5>와 <그림 6>을 비교해 보면, <그림 5>에서 표제어 '여기'에 동형어 번호가 빠져 있다는 것을 볼 수 있고, 부사적 용법으로 풀이된 '여기'의 II에 기술된 뜻풀이는 '여기2'에 다시 기술되어 있음을 알 수 있다. 출판본에서는 이러한 오류가 없지만 웹 리소스에 나타난 자료는 표제어 분할 과정에서 발생한 오류인 것으로 보인다. 이 부분도 어휘 데이터베이스를 구축하는 과정에서 출판본에

19) 사전 구조 분석기는 컴파일러 제작 도구인 GNU Bison과 Flex를 이용하였고 표준 C 언어로 간략하게 제작하였다. 이렇게 제작하는 도구 프로그램을 이용하여 프로그래밍 언어를 디버깅(debugging)하는 것과 같은 과정을 거쳤다.

따라 수정하여 재구성하였다.

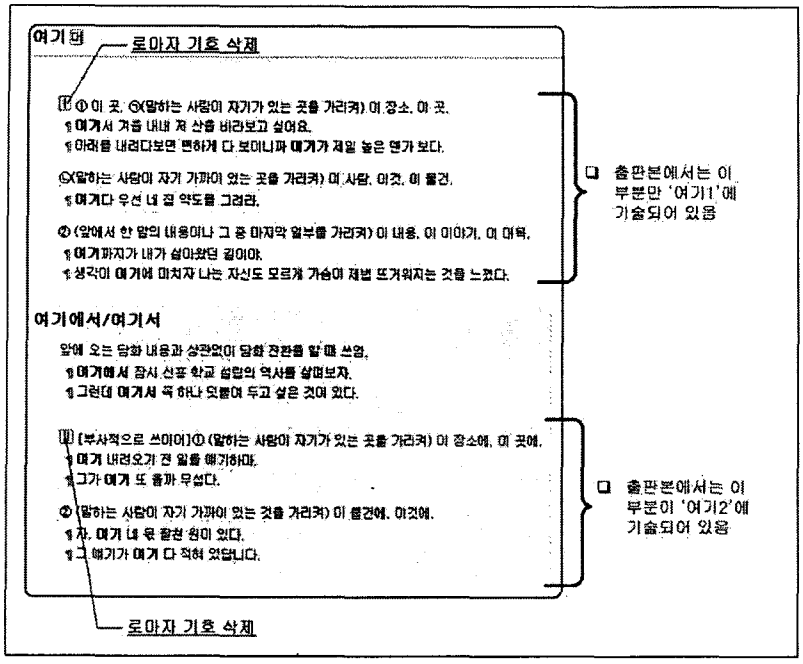

그림 5. 웹 리소스에 기술된 '여기'

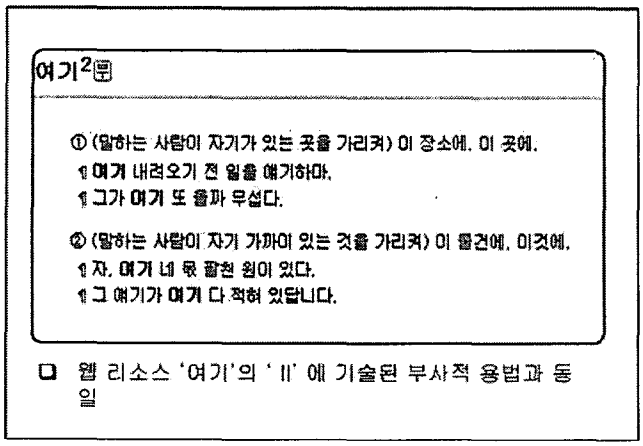

그림 6. 웹 리소스에 기술된 '여기2'

뜻풀이 갈래 번호 부여에 대해서 오류 유형이라고 봐야 할지에 대해서는 명확하지 않지만, 첫 번째 뜻풀이 갈래 번호가 부여된다는 것은 그 다음의 뜻풀이 갈래 번호가 부여된다는 것을 함의한다고 볼 수 있다.

<그림 7>을 보면 동사 표제어 '기술하다1'과 같은 경우 로마자로 구분하는 'I'의 뜻풀이는 있지만, 'II'의 뜻풀이는 없다. 뜻풀이 갈래 번호 'I'은 뜻풀이 갈래 번호 'II'가 나타나는 상황에서 구분하는 것이라고 생각할 수 있다.

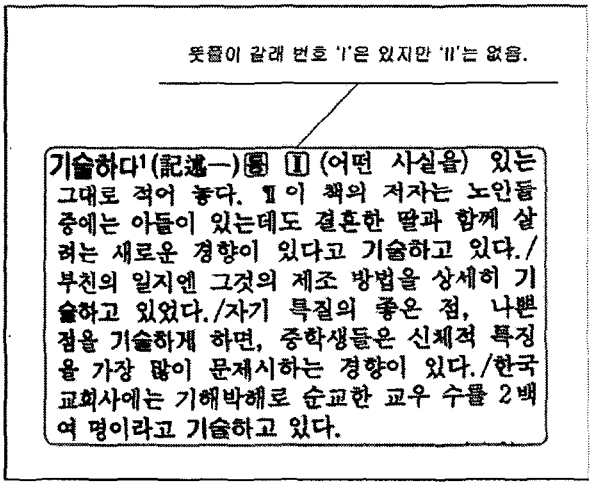

그림 7. 뜻풀이 갈래 번호

<그림 8>은 뜻풀이 갈래 번호 '①'이 있는데 뜻풀이 갈래 번호 '②' 이후의 뜻풀이가 없는 경우이다.

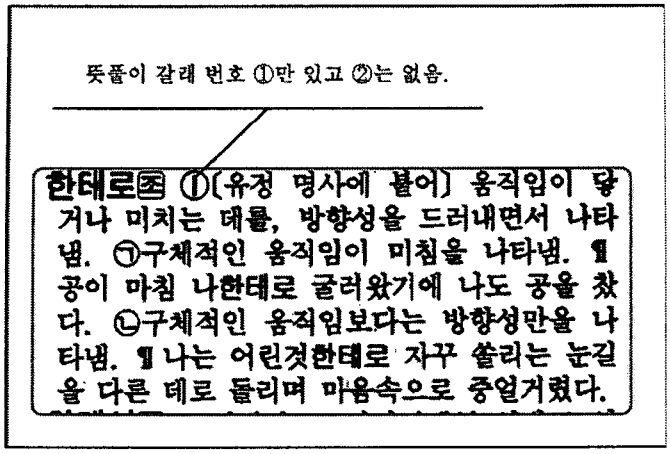

그림 8. 뜻풀이 갈래 번호

<그림 7>, <그림 8>처럼 뜻풀이 갈래 번호 함의 관계에서 벗어나는 표제어의 목록은 <표 10>과 같다.

기술하다1(I/II),	내던지다(I/II),	도사리다(I/II),
이기다1(I/II),	자6(I/II),	허덕거리다(I/II),
것(III-①/②),	고5(I-①/②),	관두다(①/②),
근거하다(①/②),	근대화되다(①/②),	내지2(①/②),
네3(①/②),	네트(①/②),	-느라고(①/②),
-는다던데(I-①/②),	대별되다(①/②),	대부하다(①/②),
더디다(①/②),	-더라고(II-①/②),	더럽다(II-①/②),
더치다(I-①/②),	덜거덕거리다(II-①/②),	덥히다(①/②),
땡기다(I-①/②),	또렷또렷(①/②),	뛰어가다(I-①/②),
-라던데2(I-①/②),	버르적거리다(①/②),	분리시키다(①/②),
분사하다3(①/②),	사이(가) 좋다(①/②),	사치하다1(①/②),
상관하다(I-①/②),	악착같다(①/②),	알1(II-①/②),
억만(I-①/②),	오다1(V-①/②),	이랑2(I-①/②),
이자3(①/②),	제조하다(①/②),	한테로(①/②),
획득되다(①/②)		

표 10 뜻풀이 갈래 번호 정형에서 벗어나는 표제어 목록

<표 10>에서 표제어 항목 옆의 괄호에 표기되어 있는 갈래 번호 쌍은, 첫 번째 갈래 번호가 뜻풀이에 나타난 번호이고 사선('/') 뒤의 갈래 번호는 뜻풀이에 앞의 번호가 있는데 뒤의 번호가 나타나지 않았다는 표시이다.

<표 9>에서 로마자 뜻풀이 갈래 번호에 오류가 있는 것과 같은 유형으로 원문자 갈래 번호에도 오류가 있는데, 해당 오류가 발견된 표제어 목록은 <표 11>과 같다. <표 11>에서 출판본 오류 항목에 'X'로 표기된 것은 출판본에서도 그러한 오류가 있다는 것이고, 'O'로 표기된 것은 웹 리소스에서만 오류가 나타났음을 뜻한다.

표제어	있음	없음	출판본 오류
내부적2	②	①	X
-대12	②	①	O
돼먹다	②	①	O
만약2	②	①	O
무쇠	②	①	X
문예	②	①	X
붓	②	①	X
비리비리하다	②	①	X
사대3	②	①	X
윤활유	②	①	X
인상2	②	①	O
자8	Ⅲ-②	①	X
자격	②	①	X
자연 경제	②	①	X
제6	②	①	O
좀5	②	①	X
중생1	②	①	X
지휘봉	②	①	X
진화하다1	②	①	X
포화1	②	①	X
피멍	②	①	O
해갈	②	①	X

표 11. 원문자 뜻풀이 갈래 번호 누락 표제어 목록

3.4. 어휘 데이터베이스 구축

구조 분석 및 변환 과정을 거친 자료는 어휘 데이터베이스로 구축하여서 관리한다. 어휘 데이터베이스는 관계형 데이터베이스 시스템을 이용해서 관리되며, 현재 5개의 테이블로 관리되고 있다. 어휘 데이터베이스에서 뜻풀이 정의문에 포함된 용법 설명[20] 등은 따로 분리되어야 할 필요가 있지만 현재는 뜻풀이 정의문과 동일한 필드로 관리되고 있다.

20) 뜻풀이 정의문에서 사용하는 ['~(의) 밥'의 꼴로 쓰이어] 같은 표현

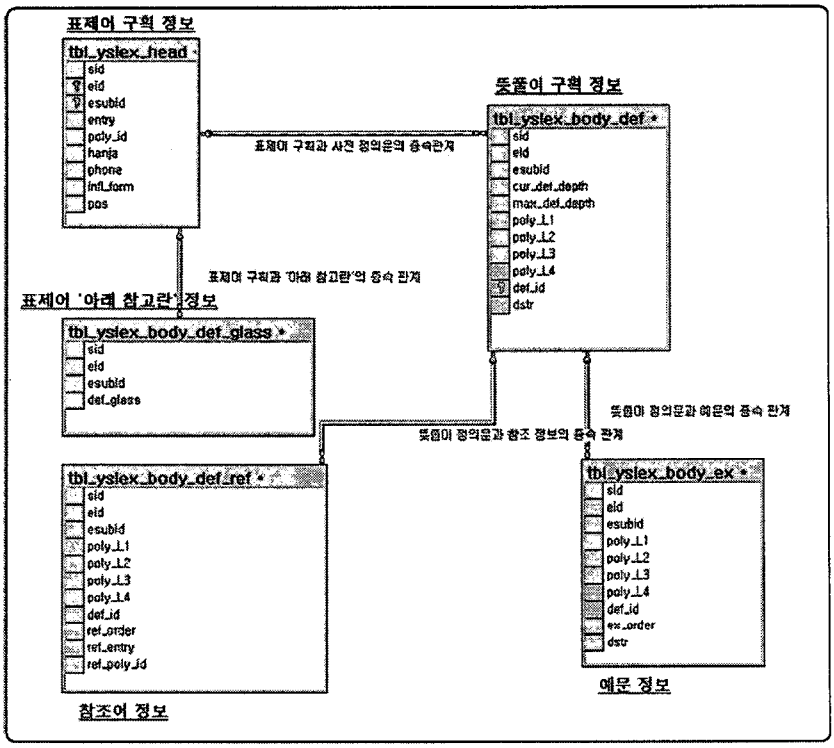

그림 9. 어휘 데이터베이스 다이어그램

관계형 데이터베이스에 구축된 어휘 데이터베이스 자료를 이용해서 참조 사전 형식의 뷰어(viewer) 프로그램을 작성하였는데, <그림 10>처럼 데이터베이스와 웹을 연결해서 어휘 리스트에 해당하는 뜻풀이와 용례를 계층적으로 보여줄 수 있도록 시험 삼아 만들어 보았다.

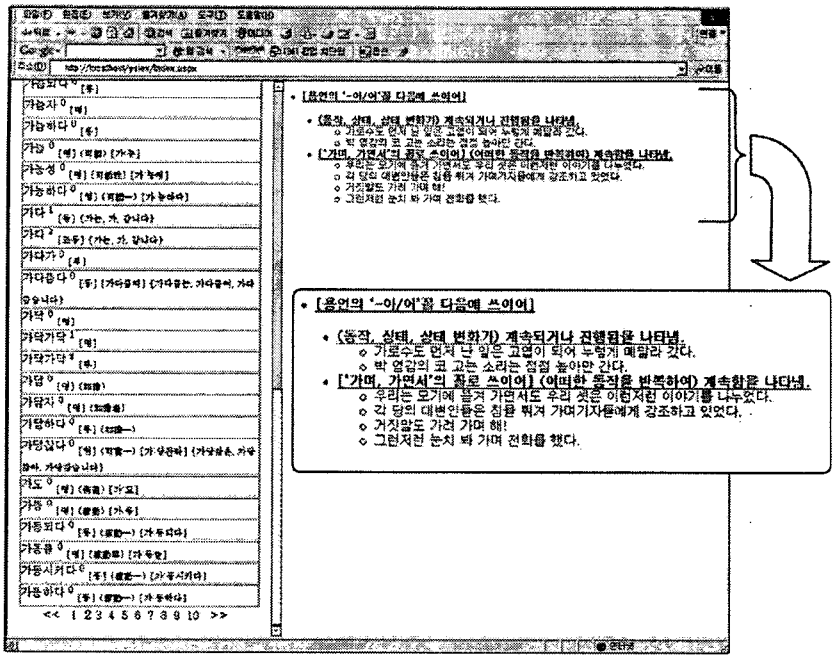

그림 10. 「연세 한국어 사전」 Viewer : '가다2'(보조동사)

어휘 데이터베이스로 구축된 자료는 기존의 데이터베이스 관리 시스템을 이용했기 때문에 다양한 통계 처리와 검색을 할 수 있으며, 대용량 자료에 대해 안정적인 성능을 보장해 주는 장점이 있다. 또 목적에 따라서 웹으로 연결해서 사용자에게 친숙한 전자 사전 기능을 구현할 수도 있으며, 자료 교환의 목적을 위해 데이터베이스의 관계 분석을 통해 XML 자료로 쉽게 변환이 가능하다.

3.5. 뜻풀이 정의문의 분포

지금까지 웹 리소스 파일의 구조 분석 및 변환 과정을 거쳐서 구축된 어휘 데이터베이스에서 뜻풀이 정의문의 분포를 추출해보면 <표 12>와 같다. <표 12>에서 '# of Def.'는 각 표제어(부표제어 제외)에 포함된 뜻풀이의 수, '빈도'는 뜻풀이가 N개인 표제어의 수를 뜻한다.

부표제어를 제외한 표제어 항목의 뜻풀이에서, 해당 표제어에 포함된 뜻풀이의 수를 추출해 본 결과 전체 표제어 중에서 78.85%의 표제어에는 하나의뜻풀이로 매겨져 있다는 것을 알 수 있다.

# of Def.	빈도	백분율	누적 백분율
1	39082	78.8579%	78.8579%
2	7187	14.5016%	93.3596%
3	1804	3.6400%	96.9996%
4	645	1.3015%	98.3010%
5	312	0.6295%	98.9306%
6	167	0.3370%	99.2676%
7	103	0.2078%	99.4754%
8	65	0.1312%	99.6065%
9	39	0.0787%	99.6852%
10	31	0.0626%	99.7478%
11	19	0.0383%	99.7861%
12	14	0.0282%	99.8144%
13	9	0.0182%	99.8325%
14	13	0.0262%	99.8588%
15	12	0.0242%	99.8830%
16	3	0.0061%	99.8890%
17	9	0.0182%	99.9072%
18	2	0.0040%	99.9112%
19	5	0.0101%	99.9213%
20	3	0.0061%	99.9274%
21	1	0.0020%	99.9294%
22	2	0.0040%	99.9334%
23	2	0.0040%	99.9374%
24	4	0.0081%	99.9455%
25	2	0.0040%	99.9496%
26	6	0.0121%	99.9617%
27	4	0.0081%	99.9697%
29	2	0.0040%	99.9738%

30	1	0.0020%	99.9758%
32	1	0.0020%	99.9778%
34	1	0.0020%	99.9798%
35	2	0.0040%	99.9839%
37	1	0.0020%	99.9859%
40	1	0.0020%	99.9879%
41	2	0.0040%	99.9919%
45	1	0.0020%	99.9939%
47	1	0.0020%	99.9960%
50	1	0.0020%	99.9980%
86	1	0.0020%	100.0000%

표 12. 뜻풀이 정의문의 분포

4. 맺음말

이번 연구에서는 「연세 한국어 사전」의 전체 표제어를 어휘 데이터베이스로 구축하는 작업을 수행하였다. 자료의 데이터베이스 구축 후에 아직 더 해결해야 할 문제점이 노출되었다. 우선 데이터베이스의 정규화(normalization) 작업이 충분히 이루어졌는지에 대한 검토가 미비했다는 것이다. 자료의 측면에서 보면 뜻풀이 문장에서 어휘의 용법을 분리해야 할 필요가 있는 것 같았다. 자료 활용 방안으로, 이번에 구축된 어휘 데이터베이스의 뜻풀이와 용례를 분석하여 어휘 의미 함수를 구축하는 연구에 활용해 볼 계획이다. 이를 위해서는 현재 사전의 뜻풀이에 사용된 어휘를 통제어(controlled vocabularies) 집합으로 간주하여 뜻풀이 체계를 분석해 볼 필요가 있을 것 같다.

「연세 한국어 사전」을 어휘 데이터베이스로 변환하는 컴파일 과정에서 사전에 포함된 몇 가지 구조적인 오류들이 발견되었지만, 이러한 오류로 인해 사전에 구조적인 문제가 있다고는 전혀 주장할 근거가 없다. 오히려, 사전 편찬과 같은 대용량 언어 자원 구축 과정에서 사소한 오류는 있기 마련이며, 이번 작업에서 발견된 오류

가 사전을 교정하시는 분들에게 도움이 되기를 바랄 뿐이다.

참고문헌

국립국어연구원 (2002), 표준국어대사전 연구분석. 국립국어연구원 2002-1-10.

김동건, 최운호 (2003), "판소리 자료 전산화 및 판소리 사전 편찬을 위한 통합 시스템의 설계 및 구현. 판소리학회 제43차 학술발표회. 판소리학회

노용균 (2001), Converting a dictionary into XML: The process and some lessons. 2001년 하계 전국 학술발표대회, 언어과학회

양단희 (1992), "한국어 전자 사전 원형의 설계 및 구현" 사전편찬학연구 제4집. 연세대학교 한국어 사전 편찬회

이경순, 김도완, 김길창, 최기선 (2001), 기계가독형사전과 코퍼스에서 추출한 의미정보를 이용한 명사열의 의미해석, 인지과학 제12권 1,2호

이수광, 옥철영 (2001), 확률적 문법규칙에 기반한 국어사전의 뜻풀이말 구문분석기. 정보과학회논문지: 소프트웨어 및 응용 제28권 제5호

조평옥, 안미정, 옥철영, 이수동 (1999), 사전 뜻풀이말에서 구축한 한국어 명사 의미계층구조. 인지과학 제10권 4호

이상섭 (1992), "전산 편찬학의 개념과 한국적 실제" 사전편찬학연구 제4집. 연세대학교 한국어 사전 편찬회

이재성, 최병진, 이운재, 최기선 (1996), 텍스트 및 전자사전 관리시스템을 위한 표준사전 표기언어의 설계. 인지과학 제7권 제4호

최병진, 이운재, 이재성, 최기선 (1996), 기계가독형 사전 구축을 위한 사전 항목의 논리 구조. 인지과학 제7권 제2호

최병진 (2002), 어휘정보구축을 위한 사전텍스트의 구조분석 및 변환. 언어와 정보 제6권 2호.

허정, 옥철영 (2001), 사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템. 정보과학회 논문지: 소프트웨어 및 응용 제28권 제9호

Alshawi, H.(1989), "Analysing the dictionary definitions", in Boguraev, B. and T. Briscoe (eds), 153-170.

Amsler, R. (1981), "A taxonomy for English nouns and verbs," Proceedings of the 19th conference on Association for Computational Linguistics. pp.133-138.

Amsler, R. A. and Tompa, F. W. (1988). An SGML-based standard for English monolingual dictionaries. Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary. Waterloo, Ontario, 61-80.

Barnbrook, G.(2002), Defining Language : A local grammar of definition sentences, John Benjamins Publishing Company.

Boguraev, B. and T, Briscoe(1989), editors, Computational Lexicography for Natural Language Processing, Longman

Boguraev, B. (1994), "Machine-Readable Dictionaries and Computational Linguistic Research," in Current Issues in Computational Linguistics: in honour of Don Walker, A. Zampoli, N. Calzolari and M. Palmer, eds. pp.119-154: Kluwer Academic Publishers.

EDR (1993), EDR Electronic Dictionary Technical Guide, Japan Electronic Dictionary Research Institute.

Fontenelle, T.(1997), Turning a Bilingual Dictionary into a Lexical-Semantic Database, Max Niemeyer Verlag, Tübingen.

Guthrie, L., Pustejovsky , J., Wilks, Y. and Slator, B. (1996), "The Role of Lexicons in Natural Language Processing," Communications of the ACM, Vol. 39, Issue 1.

Ide, N., Le Maitre J. and Véronis J.(1994), "Outline of a Model for Lexical Databases," in Current Issues in Computational Linguistics:in honour of Don Walker, A. Zampoli, N. Calzolari and M. Palmer, eds. pp.283-320: Kluwer Academic Publishers.

Raymond, D. R. and F. W. Tompa(1987), "Hypertext and the Oxford English Dictionary," Proceedings of the ACM Conference on Hypertext, pp. 143-153.