세종전자사전 활용을 위한 자연어 생성적 접근
- 어휘문법적 미시 구조의 격자 형식화를 통해

A NLG Approach to Sejong Dictionary for Computational Feasibility

저자
(Authors)

조은경
Jo Eun-kyoung

출처
(Source)

한글 , (311), 2016.3, 197-226 (30 pages)
HAN-GEUL , (311), 2016.3, 197-226 (30 pages)

발행처
(Publisher)

한글학회
The Korean Language Society

URL

http://www.dbpia.co.kr/Article/NODE06646185

APA Style

조은경 (2016). 세종전자사전 활용을 위한 자연어 생성적 접근. 한글, (311), 197-226.

세종전자사전 활용을 위한 자연어 생성적 접근

—어휘문법적 미시 구조의 격자 형식화를 통해—

조 은경

— 차례 —

〈벼리〉

세종전자사전(이하 세종사전)은 한 어휘 항목이 형태 정보, 의미 정보, 통사 정보를 가지고 여러 어휘들의 관계가 기입된 통합 어휘부이며 여러 하위 범주 사전으로 구성된 대규모 지식 베이스이다. 세종사전은 다양한 응용을 지향한 산출물임에도 불구하고 사전의 내용에 대한 연구와 자연어 처리에서의 단방향 접근에 치우쳐 있다. 무엇보다 그 산출물을 재가공하여 활용하려는 노력이 드물다.

이 논문은 자연어 생성의 방향에서 접근하여 전산 어휘부로서 세종사전의 미시 구조를 살피고, 복잡한 미시 구조의 개별 어휘 항목을 문장 생성 과정에 선택적으로 활용하기 위한 격자 형식화를 제안한다. 격자 형식화는 사전의 복잡한 기술 내용을 단순한 구조로 표현되게 하는 것이다. 이는 개별 어휘 항목이 아니라 여러 어휘 항목의 정보를 보게 하고, 어휘부 정보를 자연어로 쉽게 구성하게 한다. 특히, 용언과 체언 사전의 격자 형식화는 표현 구조를 쉽게 짜게 한다. 그리고 어미 사전의 격자 형식화는 표현 구조에서 표층 형태로 변환을 쉽게 한다.

우리는 통합 지식 베이스로서 세종사전의 활용을 위한 자연어 생성적 접근으로서 세종사전의 미시 구조를 살피고, 가능 표현 구조와 표층적인 어휘 형태의 생성을 위한 격자 형식화를 설계하였다. 또 이러한 과정을 통해 전산 어휘부로서의 개선점을 정리하였다.

주제어: 세종전자사전, 자연어 생성, 전산 어휘부, 격자 형식화, 표현 구조, 표층 형태.

1. 머리말

이 논문은 ‘21세기 세종 계획’의 중요 성과물 중 하나인 세종전자사전(이하 세종사전)을 응용 측면에서 고찰하고, 세종사전의 격자 형식화를 통한 자연어 표현 생성에의 접근 과정을 보이고자 한다.

세종사전은 현대 한국어 어휘의 체계적인 분석·기술에 근거하여 한국어 정보 처리에 보편적으로 활용될 수 있는 기반 자원을 구축하려는 사업의 성과물이다. 세종사전은 통합 지식 베이스로서 어휘 의미 정보와 어휘 통사 정보를 가진 정보화된 어휘부(Computerized Lexicon)로 자연어 이해와 생성의 전 단계에의 활용을 염두에 두었었다. 그런데 세종사전에 관한 연구는 사전의 내용과 구조에 대한 연구와 자연어 처리 혹은 이해에의 활용에 치우쳐 있다. 자연어 이해에의 활용은 사람이 표현한 자연어가 임의성이 매우 크기 때문에 지식 베이스의 규모의 한계를 드러내게 한다. 하지만, 자연어 생성에의 활용은 사전에 기술된 정보에서 출발하여 그 정보의 양과 의미 구분의 수준에 맞출 수 있기 때문에 지식 베이스로서 전자 사전의 활용성을 넓힐 수 있다. 이러한 맥락에서 세종사전의 어휘부 정보를 자연어 생성에 활용하기 위한 형식화 과정과 전산 어휘부로서의 개선점을 살피고자 한다.

자연어 생성은 또 다른 응용 연구에 활용될 수도 있다. 해외의 응용 연구 사례로 하바쉬·트라움(Habash & Traum 2003)이 있다. 이들은 기계 번역에서 어휘 개념 구조(lexical conceptual structure)로부터 자연어 문장을 생성하는 렉스젠(Lexogen)이라

는 자연어 생성 시스템을 고안하였다[1]. 이는 자연어 생성에 있어 어휘 개념과 구문 구조 지식의 역할이 핵심적이고, 이는 렉스젠이라는 이름이 시사하듯이 어휘부로부터 출발하는 시스템이다. 국내의 응용 연구로는 고 창수 외(2012)가 있다. 이 연구의 대화 시스템에는 응답문 산출 모듈이 있고, 이는 고정된 표현 템플릿(template)을 만들어 놓고 템플릿의 정보 항목인 슬롯(slot)을 채우는 방식이다. 이는 자연어 생성이 갖는 모듈적 활용성을 보여 주었다. 조 은경(2014, 2015)에서도 대화 시스템을 위한 연구에서 기계에 의해 생성된 응답 발화를 활용하지 못하여 응답 생성 모듈의 필요성을 보였다.

2장에서는 세종사전에 관한 연구과 본고와의 관련성 측면에서 세종사전의 설계 목적과 기존 연구를 살핀다. 3장에서는 자연어 생성을 위한 세종사전의 격자 형식화와 표현 생성을 살핀다. 4장에서 어휘부로서 세종전자사전의 개선점을 살핀다.

2. 세종전자사전에 관한 연구

2.1. 세종사전의 자연어 생성에의 관점

세종사전의 활용 지침서에는 세종전자사전이 한국어 정보 처리 특히 자연어 텍스트 생성에 유용하게 활용될 수 있을 것이라는 표현이 많다. 이는 전자 사전 구축 과정에 자연어 생성에의 활용도를 염두에 두었음을 말한다.

> “세종전자사전이 궁극적으로 범용 전자사전을 지향한다는 점에 비추어 볼 때, **텍스트 자동 분석과 생성**, 정보 검색 및 자동

1) http://tmt.sourceforge.net/lexgen/.

> 번역 등에 활용될 전자사전은 ....” <세종전자사전의 전산적 활용을 위한 지침서(이하 활용 지침서), 3쪽>
>
> “세종 전자사전의 조사/어미 기술 정보가 한국어 전산 처리에 효율적으로 이용될 수 있다면 한국어 자동 형태 분석, 자동 구문 분석은 물론 **한국어 텍스트의 자동 생성** 분야에 까지 많은 기여를 할 수 있으리라 기대된다.” <활용 지침서, 4쪽>

다음은 세종사전의 「사업 개요 및 사전 설명서」에 명시된 내용이다.

> “그 결과물인 세종전자사전은... **대규모 범용 한국어 전자사전**을 지향한다. 범용 전자사전이란 특정한 유형이나 영역의 기계처리 작업에 국한되지 않고, 정보 검색, **텍스트의 분석과 산출**, ..., 더 나아가서는 향후의 진보된 인공 지능 개발 환경에도 유연하게 적용될 수 있는 전자사전을 말한다.” <사업 개요 및 사전 설명서, 1쪽>

세종사전은 멜축(Mel’čuk 1999)의 어휘 함수(lexical function)의 개념에 기반해 있다(활용 지침서 51쪽). 어휘 함수 개념에 기반한 텍스트 생성에의 적용은 멜축(Mel’čuk. 1996), 요르단스카야 외(Iordanskaja et al. 1996)에서 볼 수 있고, 어휘 기능 문법의 맥락에 있는 홀리데이(Halliday 1985)의 체계 기능 문법(systemic functional grammar)에서도 담화 맥락에서 비롯되는 자연어 생성을 위한 기제를 볼 수 있다.

세종사전의 궁극적 목적은 자연어 이해와 자연어 생성이 공통적으로 사용하는 양방향(bidirectional) 자원으로의 활용이었다. 이에 우리는 세종사전의 내용과 구조에 대한 연구와 활용에 관한 연구를 살펴서 세종사전에 대한 이해를 넓히고 세종사전 활용상의 빈자리를 찾고자 한다.

2.2. 사전의 내용과 구조에 관한 연구

세종사전의 내용과 구조에 관한 연구는 매우 다양하다. 이 절에서 언급되는 대부분의 연구들은 사전 구축의 입장에서 이루어진 것들이며, 자연어 생성을 위한 직접적인 관련성을 찾기는 어려웠다. 그러나 이 연구들은 세종사전을 자료 자체로서 들여다보았을 때 이해하기 어려운 내용에 관한 이해의 폭을 넓혀준다.

임 유종(2004)은 세종사전에서 부사 사전의 정보 기술 내용과 전산적 활용을 살폈다. 목 정수(2004)는 세종사전에서 복합명사 사전에서 '어근' 목록 설정, 1음절 한자어 등에 관한 입장 차이 즉, 단일명사와 복합명사의 구분 경계가 되는 어휘 요소가 표제어 선정과 이에 따른 사전의 거시적, 미시적 구조와 관련된 복잡한 문제가 되었음을 기술하였다. 조 인식 외(2004)는 고유명사, 유행어, 시사어, 외래어/외국어, 약어, 수 표현 등을 특수어라는 범주로 묶어 미시 구조 모형과 기술 모형을 설계하였다. 임 홍빈·임 근석(2004)은 세종사전 중 체계적인 연어 사전 구축의 필요성과 미시 구조상의 특성, 연어 판별의 기준 등을 소개하였다. 송 길룡(2005)은 세종사전의 성과에 있어 세종계획의 초기 기획 단계의 기대 수준과 8년이 지난 시점에서의 전망 수준에 차이가 있음을 말하고, 2004년부터 세종사전 구축에 적용하게 된 XML 형식과 적용의 배경 및 작업의 방향을 기술하였다. 이 성헌(2004)은 술어명사 기술에 필수적인 기능동사 구문 정보를 기술하기 위한 체계를 제안하였고, 이 성헌(2005)은 세종사전에서 의미 부류 체계가 갖는 중요성을 보이기 위해 어휘 간 결합관계라는 언어 내적 속성에 근거한 형식적인 부류 체계를 설계하고 그 유용성을 보였다. 호 정은·박 만규(2005)는 관용 표현 범주들 중 속담을 구축하는 과정에서

제기된 문제 특히 변이형이 있는 속담 표현이나 어휘적 관용 표현과 경계가 모호한 속담 표현들의 표제어 선정 기준과 언어적 특성을 보이기 위한 미시 구조를 기술하였다. 그런데, 많은 연구들에서 전자사전 활용을 위한 구체적인 방향과 방법의 필요성을 말한 것은 사전 활용에 의한 응용 성과물로의 목마름을 표현한 것이라 생각된다.

또 다른 연구의 경향으로, 지식 베이스로서 세종사전의 의미 부류 체계에 관한 것이 있다. 송 도규(2005)는 세종사전의 체언과 용언의 의미 하위 부류와 동의어, 반의어, 상위어, 하위어 등의 정보를 아울(OWL: Web Ontology Language)이라는 온톨로지 형식 체계로 변환하여 활용할 수 있는 방법을 제안하였다. 이 동혁(2007)은 카이스트에서 구축한 한국어 어휘 의미망(Core Net)과 세종사전의 의미 부류 체계를 비교 분석하였으며, 의미 부류 체계에서 범주 설정과 체계화를 위해서는 그것을 어디에 활용할 것인지가 분명하게 정해져야 함을 말하였다. 전 지은·최 재웅(2008)은 세종사전의 형용사를 격틀(‘<frame>’) 항목에 기술된 형식 정보에 따라 형용사의 유형을 분류하였다. 황 순희(2009), 배 선미 외(2010), 배 선미·윤 애선(2010), 윤 애선(2010)은 세종 의미 부류와 부산대학교에서 구축한 개념 체계(KorLexNoun) 간의 통합 활용을 위해 의미 부류 체계 간 사상 방안을 논의하였다. 그 외에도 김 건희(2010)는 세종사전에서 다의어 형용사의 의미 구별의 이유가 상태성(stativity)에 있다고 보고 계량적인 분석을 한 바 있다.

2.3. 자연어 이해에 활용한 연구

세종사전을 언어 지식 자원으로서 활용하는 연구로는 주로 자연어 이해의 자원으로 활용하는 것이었다.

강 신재 외(2003)는 의미역 결정 규칙을 구축하기 위해 세종사전을 활용하였고, 김 병수 외(2007)는 부사격의 의미역 결정을 위해 세종사전을 지식 자원으로 활용하였다. 강 상욱 외(2014)는 용언의 의미 중의성 해소를 위해 세종사전과 그들이 구축한 한국어 어휘의미망을 이용한 시도를 하였다. 자연 언어 이해에의 활용에 관한 연구는 비교적 연구 성과가 적은 편이다. 이는 자연 언어의 무한한 표현력 때문에 어휘 항목 불일치로 인한 현실적인 한계 때문이다. 그래서 세종사전의 자연 언어 이해에의 활용은 수많은 정보 구조를 갖춘 규모에 비해 그 가치와 활용도 측면에서 안타까운 결과를 내기 쉽다(조 은경 2007 : 93).

3. 격자 형식화와 자연어 표현 생성

이 장에서는 세종사전 활용의 하나로서 자연어 생성을 위한 격자 형식화와 이에 의한 표현 생성의 과정을 보이고자 한다.

세종사전은 기초 사전과 상세 사전으로 나누어져 있다. 기초 사전은 용어집(glossary)에 가깝지만, 상세 사전은 표제항에 대해 형태 정보, 의미 정보, 통사 정보 등 모든 정보를 상세하게 담은 통합 어휘부로서의 전자 사전이다. 그래서 상세 사전을 고찰 대상으로 삼는다. 여러 가지의 하위 범주 사전들 중에서 특히 격틀이 기술된 용언 사전과 그 격틀의 논항 정보를 채울 체언 사전, 그리고 용언의 표층 형태를 위한 어미 사전을 살필 것이다.

3.1. 세종사전의 미시 구조와 격자 형식화

한국인이 사전을 찾아보는 이유는 상세하게 구분되는 의미를

이해하고 표현하고자 함이다. 그리하여 상세하게 구분된 의미를 살려 새로운 표현을 만든다. 기계도 이와 마찬가지이다. 세종 상세 사전과 같은 상세하게 구분된 의미를 바탕으로 결합 정보를 구성하여 적합한 구조와 표현을 생성할 수 있다. 그런데 기계가 구분된 의미에 따라 적합한 구조를 선택적으로 생성하고, 이 구조에서 적합한 표현을 선택적으로 생성하기 위해서는 현재 세종사전의 개별 어휘 항목 기준의 상세 형식과 계층적 미시 구조 형식은 적합하지 않다. 여러 어휘 항목의 형식 정보를 처리할 수 있는 형태이어야 한다. 이를 위해 이 절에서는 세종사전의 미시 구조인 구문적, 어휘적 정보를 자연어 생성을 위해 효율적으로 처리하기 위해 격자 형식화라는 방법을 기술할 것이다.

세종 상세 사전의 미시 구조상 특징은 크게 2가지로 파악되었다. 첫째는 문맥 의미 혹은 센스라 불리는 의미 구분에 따라 통사, 문형, 결합 정보가 구성되었다는 것, 둘째는 그 문맥 의미가 동형어와 다의어 범주로 구분되어 각기 다른 격틀 정보를 가지고 있다는 것이다. <표 1, 2>는 용언 상세 사전과 체언 상세 사전의 '미시 구조'를 가져와서 왼쪽에 'ㅣ'를 표시함으로써 기술 계층 구조의 가독성을 높인 것이다.

이 미시 구조를 보면, 센스 구획(sense+) 안에 의미 정보 구획(sem_grp), 통사 정보 구획(syn_grp), 문형 구성 구획(frame_grp), 결합 정보 구획(com)이 구성되어 있다. 이는 '문맥 의미(sense)'에 따라 의미 정보와 통사 정보가 구성되었음을 말한다. 의미 정보는 의미 부류, 의미 관계 등을 포괄하고 있고, 통사 정보는 명사 결합, 동사 결합, 선택 제약 등을 포괄하고 있다. 따라서 센스 구획 안의 정보 항목들로 문장의 핵심 표현 구조인 용언과 체언의 결합 구조를 구성하여 생성할 수 있다.

<표 1> 세종 용언 상세 사전의 미시 구조

superEntry	최상위 표제항 구획
orth	표기 형태
entry+	**표제항 구획**
mnt_grp	관리 정보 구획
see?	참조 표제항
morph_grp	형태 정보 구획
var*, cntr?, str?, org*, infl?, comp*, der*	변이형, 축약, 내부 구조, 원어, 굴절, 합성어, 파생어
sense*	센스 구획
sem_grp	의미 정보 구획
sem_class	의미 부류
trans+	영어 대역어
sem_rep?	의미 표시
domain?	전문 영역
lr	어휘 의미 관계 구획
syn*, ant*, hyper*, hypo*, holo*, mero*, rel*	동의어, 반의어, 상위어, 하위어, 전체어, 부분어, 관련어
frame_grp+	문형 구성 구획
frame	문형
subsense+	하위 센스 구획
sel_rst*, n_appr*, arg_rst?, ord_rst?, eg+	**선택 제약,** 적정 명사, 논항 제약, 어순 제약, 용례
com	결합 정보 구획
morph_rst?	형태론적 제약 구획
end_rst1*, end_rst2*, neg_rst*	어말어미 제약, 선어말어미 제약, 부정 제약
col_grp?	연어 구획
col*	연어 구성
morph_rel?	구성 관계 구획
ad_rel*	부사구 구성
np_rel*	명사구 구성
aux_rst*	보조용언 제약
defect*	제약된 활용형 용법 구획
infl_rst+, trans+, **sel_rst***, defect_eg+	활용 제약, 영어 대역어, **선택 제약**, 제약 용법 용례
idm_grp?	숙어 구획
idm+	숙어

<표 2> 세종 체언 상세 사전의 미시 구조

superEntry	최상위 표제항 구획
orth	표기 형태
entry+	**표제항 구획**
mnt_grp	관리 정보 구획
see?	참조 표제항
morph_grp	형태 정보 구획
var*, str?, org*, hom?, der*, comp*	변이형, 내부 구조, 원어, 동형어, 파생어 형성, 합성어 형성
idm_grp?	숙어 구획
idm*	숙어
sense+	센스 구획
sem_grp	의미 정보 구획
eg+	용례
trans+	영어 대역어
domain*	전문 영역
sem_class	의미 부류
lr?	어휘 의미 관계 구획
syn*, ant*	동의어, 반의어
syn_grp	통사 정보 구획
comb_aj*	형용사 결합 정보
comb_n*	명사 결합 정보
comb_v*	동사 결합 정보 구획
form, frame*	동사 결합 형태, 문형
max_n_str*	명사구 최대 구조
sel_rst*	선택 제약
cl*	단위 표현
prt*	조사 제약
av*	부사적 용법
s_n*	관형절 제약

세종 상세 사전의 미시 구조의 큰 특징은 동형어와 다의어가 세밀하게 구분되어 있다는 것이다. <그림 1>은 '좋다.xml'에서 동형어가 표제항 구획 번호(entry n="1")로 구분되어 있고, 다의어가 센스 구획(sense)에서 센스 번호(sense n="01")로 구분

되어 있음을 보이는 예시이다. 그리고 구분된 센스에 따라 각각의 격틀이 부여되어 있고, 격틀에서 가능한 논항의 의미 부류와 선택 제약 정보가 명세되어 있다.

<표 1, 2>에서 세종 상세 사전의 핵심 범주인 용언과 체언의 미시 구조를 보면, '문맥 의미'인 센스 구획에 따라 표현 구조 생성이 가능함을 볼 수 있다. <그림 1>에서 실제 어휘 '좋다'의 기술 내용을 보면, '문맥 의미'의 구분 범주인 동형어와 다의어가 각각의 구문 구조를 갖고, 각각이 취하는 논항 정보에 차이가 있음을 볼 수 있다. 세종사전은 하나의 어휘 항목을 중심으로 xml의 형식을 빌어 다차원의 위계화된 정보를 담고 있다. 그러나 기계로 하여금 결합 정보를 구성하고 표현을 생성하게 하기 위해서는 <표 1, 2>와 같은 계층적 개념 형식이나 <그림 1>과 같은 서술식의 상세 기술 형식은 적합하지 않다. 이에 우리는 격자 형식화라는 단순한 표현 구조로의 변환을 제안하고 이에 따른 표현 구조의 생성과 표층 어휘 형태의 생성 과정을 보이고자 한다.

격자 형식화는 세종사전의 어휘 항목을 기술한 다차원의 xml 구조를 2차원으로 변환하는 것이다. 한 개 어휘 항목의 다차원 정보를 여러 어휘 항목을 고려하여 응용 목적에 맞는 2차원의 정보 내용으로 변환[2]하는 것이다. 그리고 어느 한 어휘 항목의 정보만을 제시하는 것이 아니라 관련된 다양한 어휘 정보를 표현할 수 있다.

2) 변환을 위해서는 xml 파서 라이브러리를 활용하였고, 세종사전의 미시 구조에 관한 지침에 따라 여러 가지 스크립팅을 거쳤다. 현재 개인 연구자가 활용할 수 있는 활용 지침서에 공개된 항목 정보와 실제 항목 정보가 일치하지 않는 경우도 있었고, 유관한 항목 정보 간 유효성 검증의 과정도 있었다. 이에 대해서는 '4. 전산 어휘부로서의 개선점'에서 기술하였다.

```
<superEntry>
    <orth>좋다</orth>
    <entry n="1" pos="va">
    <entry n="2" pos="va">
    <entry n="3" pos="va">
    <entry n="4" pos="va">
        <mnt_grp>
        <morph_grp>
        <sense n="01">
            <sem_grp>
                <sem_class>상태</sem_class>
                <trans>feel so good</trans>
                <lr>
            </sem_grp>
            <frame_grp type="FA">
                <frame>X=N0-이 A</frame>
                <subsense>
                    <sel_rst arg="X" tht="THM">(마음|기분)</sel_rst>
                    <eg>내가 기분이 좋다.</eg>
                </subsense>
                <subsense>
                    <sel_rst arg="X" tht="THM">(속|몸)</sel_rst>
                    <eg>내가 몸이 안 좋다.</eg>
                    <eg>내가 속이 안 좋다.</eg>
                </subsense>
            </frame_grp>
        </sense>
        <sense n="02">
            <sem_grp>
                <sem_class>심리상태</sem_class>
                <trans>like</trans>
                <lr>
            </sem_grp>
            <frame_grp type="FA">
                <frame>Y=N1-에게는|이 X=N0-이 A</frame>
                <subsense>
                    <sel_rst arg="X" tht="THM">온</sel_rst>
                    <sel_rst arg="Y" tht="EXP">인간</sel_rst>
                    <eg>나는 영미가 좋다.</eg>
                    <eg>나는 야구가 축구보다 더 좋다.</eg>
                </subsense>
                <subsense>
                    <sel_rst arg="X" tht="THM">S것</sel_rst>
                    <sel_rst arg="Y" tht="EXP">인간</sel_rst>
                    <eg>나는 혼자 사는 것이 좋다.</eg>
                    <eg>그는 그녀를 가까이서 보살펴 주는 것이 좋았다.</eg>
                </subsense>
            </frame_grp>
        </sense>
    </entry>
</superEntry>
```

<그림 1> 세종사전의 동형어와 다의어: '좋다.xml'

2차원의 자료 구조는 1차원 x축이나 2차원 y축의 기준 축의 속성과 첨자 값(index value)을 가지고 원하는 결과를 쉽게 얻게 한다. <표 3>과 같은 격자 형식화의 예를 들어 보자.

<표 3> 체언의 형식 의미 정보를 격자로 나타냄

체언류(행) | 의미역(열)	THM	CRT	LOC	GOL	EXP
건강	신체상태	신체상태			
날씨	날씨				
마음	품격속성				
맛	속성				
머리	신체부위	능력, 두뇌			
매니저	인간				인간
맹호	구체물				짐승
머리말	서론		서언,서문		
머리맡	구체물		관계장소		
사이	추상적상태		관계장소		
쓰기	추상적대상	학문과목			

논항의 의미역 부류를 표시하는 'THM', 'CRT', 'LOC', 'GOL', 'EXP'는 각각 '대상', '기준치', '장소', '도착점', '경험주'이다(세종사전의 전산적 활용을 위한 지침서 45쪽). 각 의미역 부류 즉, 'THM, CRT, LOC' 등에는 각 단어의 의미역 부류 명세(sem_class)에 사용된 값 즉, '신체 부위, 신체 상태, 관계 장소' 등이 채워졌다. 'THM'의 의미역으로 '신체 상태'가 주어지면, '건강'을 찾아올 수 있다. 또, 'CRT'의 의미역으로 '능력, 두뇌'가 주어지면, '머리'를 찾아올 수 있다. 그리고 '머리'라는 단어가 주어진다면 '신체 부위'의 뜻일 때는 'THM'의 역할로, '능력, 두뇌'의 뜻일 때는 'CRT'의 역할로 취할 수 있다. <그림 2>는 격자 형식화된 어휘부에서 의미역 조건을 만족하는 논항을 갖춘 표현 구조 생성을 위해 항목을 가져오는 의사 코드

(pseudo code)와 실행 과정이다.

조건	실행: get * by col=val	결과
col=theme; val=신체 상태	get * by them='신체 상태'	건강
col=crt; val=능력, 두뇌	get * by (crt='능력' or crt='두뇌' or crt='능력, 두뇌')	머리
col=nominal; val=머리	get * by nominal='머리'	신체 부위, {능력, 두뇌}
조건	실행: get col.name by col.val=val	결과
val=신체 부위	get col.name by col.val='신체 부위'	THM
val=능력, 두뇌	get col.name by col.val='능력, 두뇌'	CRT

<그림 2> 격자 형식화 어휘부에서 생성 논항 가져오기

3.2. 격자 형식화에 의한 표현 구조 생성

세종사전은 의미 구분과 그 결합 정보를 상세히 기술한 자원이다. 문맥 의미 혹은 센스라 불리는 의미 구분에 따라 결합 정보가 구성되었고, 그 문맥 의미가 동형어와 다의어 범주로 구분되어 각기 다른 격틀 정보를 가지고 있다. 이러한 특성 때문에 세종사전의 구축자에 의해 이루어진 의미 구분만큼 필요 정보 항목의 설정이 가능하고, 설정된 항목의 결합 정보만큼 결합 형태의 설정이 가능하다.

이 절은 용언 사전의 동형어와 다의어로 구분된 격틀과 의미역을 자연어 생성을 위한 표현 구조 생성의 축으로 삼고 격자 형식의 설계를 제시할 것이다. 구체적으로 보일 예시로써 느낌을 표현하는 형용사를 택하였다. 우리말의 형용사는 대개 느낌 혹은 상태의 감성을 표현하는 어휘 범주이다. 느낌을 표현하는 어휘는 '좋다'와 '나쁘다'가 가장 최상위의 범주를 구성한다. 그

리고 '좋다'의 사전 기술 내용을 보면 성상 형용사로서의 '좋다'에 대한 짝으로서 '나쁘다'가 있고, 심리 형용사로서의 '좋다'에 대한 짝으로서 '싫다'가 있다. 이에 '좋다', '나쁘다', '싫다'의 3가지[3] 어휘에 대해 격자 형식화를 해보자. '좋다', '나쁘다', '싫다'의 동형어, 다의어와 각각의 격틀 및 논항 정보를 격자로 나타내면 <표 4~6>과 같다. 여기서의 논항X~Z는 이전 절에서 격자 형식화의 예시로 보였던 <표 3>의 내용과 연계되어 있다. 즉, 용언의 가능 표현 생성을 위해서는 체언 사전의 형식화가 선행되어 있어야 한다. 이는 '좋다.xml' 파일에서 의미 구분에 따른 격틀을 찾아서 격틀의 의미역 정보를 참고하여, 체언 사전들(*.xml)에서 일치되는 모든 의미역 정보를 찾는 것이 아니라 격자 형식화된 체언 항목을 곧바로 찾아오게 한다.

<표 4> 용언 '좋다'의 형식 의미 정보를 격자로 나타냄

동형어	다의어	격틀	논항X	논항Y	논항Z
entry1	sense1	X=N0-이 A	"THM">속성|상태|상황		
	sense2	X=N0-이 A	"THM">(기후|날씨|일기)		
	sense3	X=N0-이 A	"THM">속성(맛|냄새|향기)		
	sense4	X=N0-이 A	"THM">신체부위		
	sense5	X=N0-이 A	"THM">인간		
	sense6	X=N0-이 Z=N2-와 Y=N1-이 A	"THM">인간	"THM">관계(사이|관계|유대)	"COM">인간
	sense7	X=N0-이 A	"THM">사태		
entry2	sense1	X=N0-이 Y=S기1-에 A	"THM">구체물|장소	"CRT"/>	
entry3	sense1	X=N0-이 Y=N1-에|에게 A	"THM">구체물|추상적대상	"GOL">신체부위|인간|추상적대상	
entry4	sense1	X=N0-이 A	"THM">(마음|기분)		
	sense2	Y=N1-에게는|이 X=N0-이 A	"THM">은	"EXP">인간	

<표 5> 용언 '나쁘다'의 형식 의미 정보를 격자로 나타냄

동형어	다의어	격틀	논항X	논항Y
entry1	sense1	X=N0-이 A	"THM">추상적대상(품질|재질)	
	sense2	X=N0-이 A	"THM">(기후|날씨)	
	sense3	X=N0-이 A	"THM">(맛|냄새|향기)	
	sense4	X=N0-이 A	"THM">신체부위(머리|눈|어깨)	
	sense5	X=N0-이 A	"THM">인간	
	sense6	X=N0-이 Y=N1-이 A	"EXP">인간	"THM">(기분)
	sense7	X=N0-이 Y=N1-이 A	"EXP">인간	"THM">(속)
entry2	sense1	X=N0-이 Y=S기1-에|이 A	"THM">구체물|장소	"LOC">행위
entry3	sense1	X=N0-이 Y=N1-에|에게 A	"THM">구체물|추상적대상	"LOC">신체부위|인간|추상적대상

3) '문맥 의미=센스'로 본다면 3개 어휘가 아니기 때문에 '가지'라고 했다.

<표 6> 용언 '싫다'의 형식 의미 정보를 격자로 나타냄

동형어	다의어	격틀	논항X	논항Y
entry1	sense1	X=N0-에게는|이 Y=N1-이 A	"EXP">인간	"THM">온

이제 느낌을 표현하는 자연어의 표현 구조를 생성해 보자. 격자 형식으로 변환된 <표 4~6>은 <표 7>과 같은 자연어의 표현 구조 생성을 용이하게 한다. '센스'의 번호는 '동형어번호.다의어번호'이며, 각각의 '격틀'이 있다. '설명'은 xml 파일의 '<sem_class>'의 값이며, 이 값의 내용이 부실한 경우 '<sel_rst>'의 값으로 보완하였다[4]. '변항'은 격틀의 의미역에 맞는 구체적인 체언 어휘를 대입한 것이다.

<표 7> 용언과 체언에 의한 표현 구조 예시

>> 좋다

격틀	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 Z=N2-와 Y=N1-이 A
센스	1.1	1.2	1.3	1.4	1.5	1.6
설명	정적 상태	(기후) 상태	(맛,냄새)속성값	(신체부위)속성값	긍정적인성속성값	(관계,유대)상태
변항	N0=건강	N0=날씨	N0=맛	N0=머리	N0=매니저	N0=매니저, N1=사이, N2=맹호
표현	건강이 좋다	날씨가 좋다	맛이 좋다	머리가 좋다	매니저가 좋다	매니저가 맹호와 사이가 좋다

격틀	X=N0-이 A	X=N0-이 Y=S기1-에 A	X=N0-이 Y=N1-에|에게 A	X=N0-이 A	Y=N1-에게는|이 X=N0-이 A
센스	1.7	2.1	3.1	4.1	4.2
설명	(사태)속성값, 충분하다	(장소)속성값, 적합하다	(추상대상)속성값, 유용하다	(마음)상태, 기분이 좋다	심리상태, 선호되다
변항	N0=머리말	N0=머리말, S기1=읽-기	N0=쓰기, N1=머리	N0=마음	N0=매니저, N1=맹호
표현	머리말이 좋다	머리말이 읽기에 좋다	쓰기가 머리에 좋다	마음이 좋다	맹호에게는 매니저가 좋다

4) 이는 '<sem_class>'와 '<sel_rst>'의 값의 구분이 모호하고, 연관된 형식 정보임을 뜻한다. 어휘부로서의 정제가 필요한 부분이다.

>> 나쁘다

격틀	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 A
센스	1.1	1.2	1.3	1.4	1.5
설명	(품질,상태) 평가속성값	(기후) 평가속성값	(맛,냄새) 평가속성값	(신체부위) 낮은정도	(인간) 평가속성값
변항	N0=건강	N0=날씨	N0=맛	N0=머리	N0=매니저
표현	건강이 나쁘다	날씨가 나쁘다	맛이 나쁘다	머리가 나쁘다	매니저가 나쁘다

격틀	X=N0-이 Y=N1-이 A	X=N0-이 Y=N1-이 A	X=N0-이 Y=S기1-에ㅣ이 A	X=N0-이 Y=N1-에ㅣ에게 A
센스	1.6	1.7	2.1	3.1
설명	낮은정도, 기분이 나쁘다	신체상태	(행위) 평가속성값	(신체,추상) 평가속성값
변항	N0=매니저, N1=마음	N0=매니저, N1=머리	N0=머리말, S기1=읽-기	N0=맛, N1=건강
표현	매니저가 마음이 나쁘다	매니저가 머리가 나쁘다	머리말이 읽기에 나쁘다	맛이 건강에 나쁘다

>> 싫다

격틀	X=N0-에게는ㅣ이 Y=N1-이 A
센스	1.1
설명	심리상태, 좋다 4.2의 반의어.
변항	N0=매니저, N1=맹호
표현	매니저가 맹호가 싫다

이상으로 세종 상세 사전 중 용언과 체언의 미시 구조 내용인 격틀과 의미역을 격자 형식화하여 표현 구조를 생성하는 과정을 보였다. 격자 형식화는 표현 구조 생성을 위한 의미 제시 조건이 주어지면 이 조건에 맞는 표현 구조의 생성을 쉽게 한다. ‘건강’이라는 키워드(제시어)가 주어지면 이 격자 형식에서 가능한 자연어는 ‘건강이 좋다/나쁘다’이다. 마찬가지로 ‘머리말’이라는 키워드가 주어지면 이 격자 형식에서 가능한 자연어는 ‘머리말이 좋다/나쁘다/싫다’이다. 같은 맥락에서 ‘맛’, ‘건강’이 주어졌을 때는 ‘맛이 건강에 좋다/나쁘다’는 생성되지만 ‘맛이 건강에 싫다’는 생성될 수 없다. 또 ‘맹호’와 ‘매니저’가 주어졌을 때에도 ‘맹호가 좋다/나쁘다’, ‘매니저가

좋다/나쁘다', '맹호에게는 매니저가 좋다/싫다', '맹호가 매니저가 좋다/싫다' 등이 생성되지만 '맹호에게는 매니저가 나쁘다'나 '맹호가 매니저가 나쁘다'는 생성되지 않는다. 또 '매니저가 머리가 좋다/나쁘다'는 생성되지만 '매니저가 머리가 싫다'는 생성되지 못한다.

가능 표현은 최대한 생성되고, 불가능한 표현은 어느 정도 제약되어 생성된다. '최대한'이나 '어느 정도'에 대해서는 세종사전의 체언이나 용언 부류 전체를 가지고 계량적 조사를 해야겠지만, 개인 연구자가 세종사전의 '비정형화된 형식' 정보 기술을 전부 분석(parsing)하여 정당한(reasonable) 계량적 결과를 내기란 쉽지 않았다. 왜냐하면 이때의 계량적 분석은 적합하고 맞는 정보를 전제로 진행해야 하기 때문이다.

<표 9>는 용언과 체언의 표제 항 파일 수이다(사업 개요 및 사전설명서 66쪽, 80쪽 참조). 이를 바탕으로 하면, 용언과 체언의 문맥 의미를 활용하여 기계적으로 표현 가능한 2 단어짜리 표현의 수는 29,492×35,854=105,706,168개이다.

<표 9> 용언과 체언의 표제 항과 의미의 개수

	파일 수 (최상위 표제항 수)	문맥 의미 수 (센스 수)
용언 기술 항목 수	19,579	29,492
체언 기술 항목 수	25,458	35,854

3.3. 표현 구조에서 표층 형태로 변환

위의 절에서 표현 구조를 생성하는 것을 보였다면, 이 절에서는 표현 어휘 형태로 변환하는 과정을 보이고자 한다. 이는 곧 변이형, 굴절 정보를 고려한 어휘 변환에 의한 표층 실현

(surface realisation)이다.

<표 10>은 표층 실현에 필요한 어미 사전의 어휘 항목 일부인 '과거'와 '미래' 선어말 어미와 '평서'와 '의문'의 어말 어미를 격자 형식으로 변환한 것이다.

<표 10> 어미의 형식 의미 정보를 격자로 나타냄

기능 / 어휘 항목	동형어. 다의어	문법의미 :gm	하위부류 :je_subtype	격틀 :frame	문형 :st	후행 :fol	선행 :frn	변이형 :var type
았	1.1	과거	ep	V-E		겠,다,습니다,구나,더구나,사옵니다,는데		었,ㅆ,였
았	1.2	확정적예측	ep	V(시)-E		다,어,구나		었,ㅆ,였
았	1.3	현재상태	ep	Va-E		다,어,구나		었,ㅆ,였
겠	1.1	미래	ep	Vv(시)-E				
겠	1.2	추측	ep	V(시)-E, ass-E				
ㄴ다	1.1	단언	eft	Vv(시)-E	평서			
ㄴ다	2.1	자문	eft	Vv(시)-E	의문			
다	1.1	단언	eft	Va(시)-E, ass-E, gss-E	평서		으시,았,았었,겠	
다	1.2	화 계 중 화된단언	eft	V(시)-E, ass-E, gss-E	평서		으시,았,았었,겠	
아	1.1	계기	efcs	Vv-E				어,여,라
아	2.1	보조연결	efcx	V-E				
아	3.2	의문	eft	V-E	의문	요		
니	1.1	이유| 원인	efcs	V(시)-E			으시	
니	2.1	원칙단언	eft		평서			
니	3.1	친근의문	eft	V(시)-E, ass-E, gss-E	의문		으시,았,았었,겠	
냐	1.1	의문	eft	Va(시)-E, ass-E, gss-E	의문		으시,었,겠	느냐,으냐

<표 11>은 이전 절에서 생성한 '좋다'의 표현 구조에 <표 10>에서 마련한 어미 형식 구조를 연계한 표층 형태로의 생성 예시이다.

<표 11> ‘좋다’의 구조에 어미를 붙인 표층 형태 예시

격틀	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 A	X=N0-이 Z=N2-와 Y=N1-이 A
설명	정적 상태	(기후) 상태	(맛,냄새)속성값	(신체부위)속성값	긍정적인성속성값	(관계,유대)상태
가능 문법의미	과거1,미래2	과거1,미래2	과거1,미래2	과거1,미래2	과거1,미래2	과거1,미래2
가능 문형	평서1,의문2	평서1,의문2	평서1,의문2	평서1,의문2	평서1,의문2	평서1,의문2
표층 1.1	건강이 좋았다	날씨가 좋았다	맛이 좋았다	머리가 좋았다	매니저가 좋았다	매니저가 맹호와 사이가 좋았다
표층 2.1	건강이 좋겠다	날씨가 좋겠다	맛이 좋겠다	머리가 좋겠다	매니저가 좋겠다	매니저가 맹호와 사이가 좋겠다
표층 1.2	건강이 좋았냐	날씨가 좋았냐	맛이 좋았냐	머리가 좋았냐	매니저가 좋았냐	매니저가 맹호와 사이가 좋았냐
표층2. 2	건강이 좋겠냐	날씨가 좋겠냐	맛이 좋겠냐	머리가 좋겠냐	매니저가 좋겠냐	매니저가 맹호와 사이가 좋겠냐

여기서는 격자 형식화를 활용한 단순한 표층 실현 예시를 제시하는 수준에서 그쳤지만(현재 세종 사전의 활용 가능한 어미 정보에 기반하여), 용언의 격틀과 어미의 격틀을 연결하기 위한 형식화와 어미 결합에 있어 순서 관계에 대한 더 잘 다듬어진 어미 사전의 형식화가 필요하다.

4. 전산 어휘부로서의 개선점

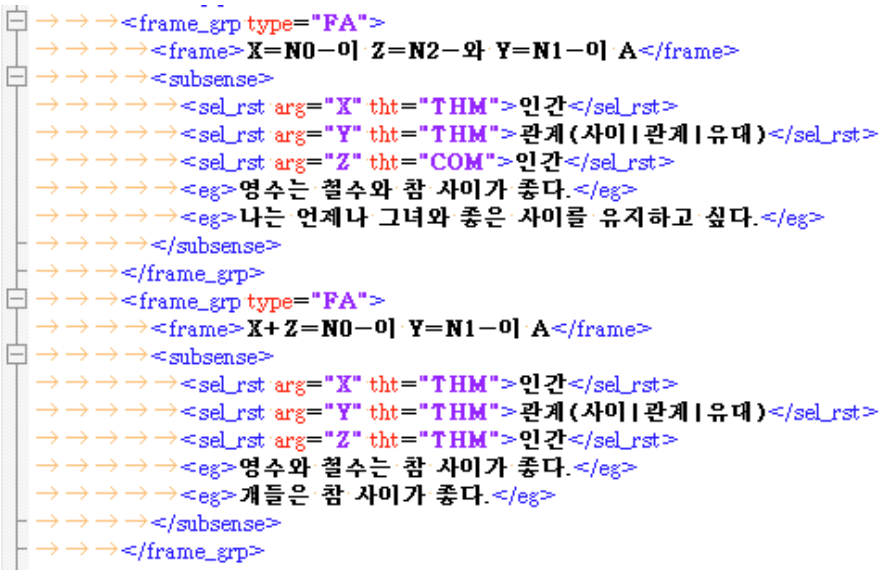
세종사전이 통합 지식 베이스로 보다 활발하게 활용되기 위해서는 정보 형식적 거시 구조나 정보 내용적 미시 구조에서 여러 가지 개선이 필요하다.

정보 형식적 거시 구조 측면에서 보자. 구조적인 형식을 갖춘 전산 어휘부로서의 활용을 위해서는 어휘 항목 간의 혹은 항목 내 정보 간의 관계를 활용 목적에 맞는 필요 정보를 추출하기 용이하도록 하는 관계형 데이터베이스 형식이 더 적합하다. 따라서 XML 형식뿐만 아니라 관계형 데이터베이스 형식의

두 가지 자료로 배포했어야 했다. 나아가 관계형 데이터베이스로부터 활용 목적에 맞는 필요 정보를 추출하여 보다 단순한 격자 형식화와 같은 자료 구조로 변환하여 쓸 수 있도록 어휘 항목 간의, 항목 내 정보 간의 관계가 전체적인 그림으로 가시화되는 개체 관계 다이어그램(ER diagram) 등을 배포했더라면 더 유용했을 것이다. 자연어 처리/생성 시스템은 하나의 단어 혹은 하나의 문장을 처리하는 데에 있어서 수많은 어휘부 혹은 이것을 여러 번 탐색하는 일이 있기에, 전산 어휘부는 XML과 같은 정보 교환용 파일 형식이 아니라 필요 항목으로 재구성되어 자주 쉽게 찾을 수 있는 메모리 탑재형 자료 구조로 변환될 필요가 있다. 따라서 어휘 항목 혹은 어휘 범주 간 관계를 필요에 따라 재구성할 수 있도록 참조할 개체 관계 다이어그램이 있어야 하는 것이다. 또, 과제 기간 후반부였을 2004년부터 적용하게 된 XML 형식의 표현에도 실제 사용을 고려한 기술이 있어야 했다. 수많은 사전 항목에 걸쳐 있는 연관 표제어 관계는 평면적인 텍스트 형태가 아니라 <a href="상대 경로 파일명">으로 기술되었어야 했다. 세종사전의 각 파일들이 하나의 상위 디렉토리 안에 여러 하위 디렉토리(=하위 사전) 구성으로 저장되어 있기 때문에 정보 기술 단계에서 충분히 가능했을 XML의 장점을 충분히 활용하지 못한 것이다.

정보 내용적 미시 구조 측면에서 보자. 세종사전은 의미(sense) 구분이 지나친 때가 종종 있으며, 논항 제약과 격틀의 형식화가 좀 더 잘 정비되었어야 했다. <그림 3>의 두 번째 격틀(frame)의 예문과 첫 번째 격틀의 예문을 보면, 격틀이 동일할 수 있다. 각기 다른 격틀에 속한 예문인 '영수와 철수는 참 사이가 좋다'와 '영수는 철수와 참 사이가 좋다'는 어순이 다를 뿐 격틀이 다르지는 않다. 또, 두 번째 격틀에 있는 형식화 표

현인 ‘X+ Y=N0’는 자연어 표현인 ‘영희와 철수는’이나 ‘걔들은’과 다르지 않다. 이러한 점들 때문에 세종사전을 활용하기 위해서는 또 다른 텍스트 처리를 해야 하는 일이 있다.



<그림 3> 지나친 의미 구분의 예시

또, ‘좋다.xml’, ‘나쁘다.xml’의 기술 내용을 들여다보면 사전 기술 과정에서 상호 참조하여 작성하였을 것도 같다. 그런데 이전 절에서 제시한 <표 5>의 (entry1, sense6)의 기술이 <표 6>에서 빠진 것은 ‘인간 관계가 나쁘다’라는 것이 불가능한 표현이 아닌데 빠져 있음을 알게 한다. 이 또한 복잡한 구조의 독립적인 지식 베이스 구조가 아니라 격자 형식화로 정리하여 병렬(parallel) 격틀 항목 유효성 검증 단계를 거치면 파악되는 문제이다. 그리고 표층 형태로 변환하는 과정에서 어미의 형식 정보가 더 잘 다듬어져야 함을 말한 바 있다. 용언의 격틀과 어미의 격틀이 ‘frame’이라는 항목으로 기술되어 있다. 이들의 연결 구성을 위한 형식 정보가 필요하다. 어미도 결합 순서에 선호도나 제약이 있을 터인데 이에 대한 조건 기술이 있어야

할 것이다. 그리고 어미 기술 중 문법 의미(gm)의 값은 너무 많은 범주가 구분되지 않고 한꺼번에 기술되어 있다. <표 11>을 보면, 문법 의미 안에 시제, 상, 서법, 양태, 문형, 명제 의미 등이 들어가 있음을 볼 수 있다. 이들은 더 구분되는 형식 의미로 나뉘어 기술되거나 이접적(disjoint)인 것만 기술되어야 한다. 따라서 전자 사전의 실제적인 활용을 위해서는 전반적인 정제 및 가공 단계가 요구된다. 이에 관한 효율적 진행을 위해서는 여러 가지 유효성 검증 단계를 두어 어휘부로서의 형식화에 맞는 정보 표현으로 구성되었는지 기계적인 점검을 하게 해야 할 것이다.

위와 같은 문제점에도 불구하고, 방대한 양으로 상세하게 형식화된 의미 기술은 목적에 따라 다양한 활용도를 만들 수 있을 것이다. 무엇보다 개인 연구자나 소그룹의 연구자가 세종사전의 내용과 같은 방대한 양의 형식 의미 정보 기술을 단기간에 축적하기란 불가능할 것이기 때문이다.

세종사전이 전산 어휘부로서 형식 정보를 담은 사전이지만, 자연어 이해에 적용하였을 때는 수많은 이표기, 오탈자, 비문, 복잡한 문서 구조를 가진 자연어의 텍스트를 처리하고 이해하기에는 너무 큰 한계가 있었다. 그리고 국내는 자연어 이해에 비해 자연어 생성 시스템에 대한 연구가 미진하고 어떤 자원을 토대로 시작해야 하는지도 미진하다. 자연어 표현 생성이라는 관점에서 접근할 때 구조적인 정보의 적합성과 필요성에 대해 더 실제적으로 접근할 수 있다. 특히 세종사전에 덜 형식화된 기술이나 의미 부류 기술의 불일치와 같은 문제점은 자연어 생성 방식으로 가능한 표현들을 만들었을 때 더 구체적인 형식화가 이루어질 것이고 이는 어휘문법의 완전성을 더할 것이다. 예를 들면, 위에서 기술한 지나친 의미 구분과 격틀 형식화의

문제점이나 문법의미(gem) 항목의 세분화 필요성 등이 그 대상이다.

또 한편, 언어학의 현상 설명 및 어휘부 기술은 대개 허용되는 표현들에 대한 상세한 기술이다. 허용되지 않는 표현들이 왜 안 되는지에 대한 형식화된 설명도 언어학적 설명 범위에서 더 잘 다루어야 한다. 이는 곧 전산 어휘부로서의 형식화에 완전성을 더할 것이다.

5. 맺음말

이 논문의 내용을 요약하면 다음과 같다.

－통합 지식 베이스를 지향한 세종사전에 관련한 기존의 연구로, 사전의 내용 연구와 자연어 이해에만 치중된 경향을 있음을 확인하였다.

－세종사전을 자연어 생성에 활용하기 위한 정보 처리 형식으로서 격자 형식화를 제안하였다.

－세종사전의 정보 내용을 격자 형식화하여 생성되는 자연어의 구조와 표현을 제시하였다.

－언어 정보 처리에의 활용을 염두에 두어 구축된 세종사전을 전산 어휘부로 더 잘 활용하기 위한 개선점을 살폈다.

기존 연구의 대부분이 세종사전의 내용에 대한 고찰이나 자연어 처리에의 활용에 그쳤다. 그런데 자연어 생성에의 활용은, 무엇보다 세종사전이 정교한 어휘 문법을 기술하고 있기 때문에, 전자 사전의 언어 정보와 지식 베이스로의 활용성을 극대화하는 결과를 기대할 수 있다. 어휘문법 자원을 자연어

생성에 활용하는 일은 사전에 기술된 정보의 양과 의미 구분의 수준에 따라 필요 정보의 집합 설정이 가능하기 때문이다. 격자 형식화는 복잡한 구조의 형식 의미를 2차원의 단순한 구조로 펼쳐 놓게 하고, 항목 정보에 쉽게 접근하여 목표 값의 활용을 용이하게 한다. 격자 형식화는 세종사전의 복잡한 구조적 지식을 단순하게 활용할 수 있게 한다. 격자 형식화는 자연어 생성에서 필요한 어휘부 정보의 선택과 결정을 편리하게 한다.

21세기 세종 계획은 1998년대에 시작하여 2007년 12월까지 지속된 대규모 국책 과제이다. 세종 계획의 여러 성과물 중 말뭉치와 전자 사전은 가장 많은 참여 인력이 동원되었고 대규모의 성과로 축적되었다. 말뭉치는 지속적으로 여러 연구에서 연구 자료로 활용되고 있지만, 세종사전은 재가공 및 활용 성과가 적다. 세종 말뭉치도 말뭉치 정제 과제가 수행된 이후 본격적이고 계량적인 연구 결과가 쏟아진 것처럼 세종사전도 지속적인 재가공과 정제의 과정이 필요하다.

〈참고 문헌〉

강 상욱·김 민호·권 혁철·전 성규·오 주현. 2014.「세종사전과 한국어 어휘의미망을 이용한 용언의 어의 중의성 해소」, 한국정보과학회 제41회 정기총회 및 추계 학술발표회. 414~416쪽.

강 신재·박 정혜. 2003.「대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축」, 정보처리학회논문지B Vol.10(2). 한국정보처리학회. 219~228쪽.

고 창수 외. 2012.『인공지능 대화 시스템 연구』. 지식과 교양.

김 건희. 2010.「다의미 형용사의 상태성에 대하여-세종계획 전자사전 형용사 기술에 대한 통계적 분석을 중심으로」, 한말연구 27. 한말연구학회. 5~39쪽.

김 병수·이 용훈·이 종혁. 2007.「비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정」, 정보과학회논문지: 소프트웨어 및 응용 제34권 제2호. 한국정보과학회. 112~122쪽.

목 정수. 2004.「복합명사의 내적 구성에 대한 연구-세종사전의 목록과 선정 기준을 둘러싸고」, 한국사전학 제3호. 한국사전학회. 105~128쪽.

배 선미·임 경업·윤 애선. 2010.「인간언어공학에의 활용을 위한 이종 개념체계 간 사상」, 인지과학 제21권 제1호. 한국인지과학회. 95~126쪽.

배 선미·윤 애선. 2010.「이종 개념체계의 상호 보완 방안 연구」, 언어와 정보 14권 1호. 한국언어정보학회. 165~196쪽.

윤 애선. 2010.「언어공학에 활용하기 위한 표준국어 대사전과 세종사전 간 용언 어의 사상」, 언어학 56. 한국언어학회. 197~235쪽.

송 길룡. 2005.「세종사전 마크업 표준화의 배경적 의미와 방향」, 한국사전학 제6호. 한국사전학회. 197~218쪽.

송 도규. 2005.「대용량 OWL 온톨로지 자동구축을 위한 세종사전 활용 방법론 연구」, 언어와 정보 Vol.9(1). 한국언어정보학회. 19~34쪽.

이 동혁. 2007.「의미 범주 체계의 구축과 사전에서의 활용」, 한국어의미학 24. 한국어의미학회. 51~82쪽.

이 성헌. 2004.「전자사전에서의 기능동사 구문 처리 문제-세종 체언 전자사전의 경우」, 한국사전학 제4호. 한국사전학회. 279~322쪽.

이 성헌. 2005.「전자사전 구축과 의미 부류—세종 명사 의미 부류 체계의 예」, 한국사전학 제5호. 한국사전학회. 103~138쪽.

임 유종. 2004.「21세기 세종사전(부사사전)의 정보 기술 내용과 전산적 활용」, 한국사전학 제3호. 한국사전학회. 153~173쪽.

임 홍빈·임 근석. 2004.「21세기 세종계획 전자사전 구축분과 연어사전의 정보 구조와 기술 내용」, 한국사전학 제4호. 한국사전학회. 99~130쪽.

전 지은·최 재웅. 2008.「한국어 형용사 유형 분류와 격틀 집합—세종사

전을 활용하여」, 한국어의미학 25. 한국어의미학회. 223~251쪽.
조 은경. 2007.「국어 공지시 해석 시스템에 관한 연구」, 연세대학교 박사학위 논문.
조은경. 2014.「인간 기계 상호작용을 위한 대화 처리 연구」, 한글 306. 한글학회. 101~130쪽.
조 은경. 2015.「대화 시스템을 위한 대화 분석과 맥락 활용」, 텍스트언어학 39. 한국텍스트언어학회. 341~370쪽.
조 인식·유 현조·신 효필. 2004.「21세기 세종 계획 전자사전의 특수어」, 한국사전학 제3호. 한국사전학회. 129~152쪽.
호 정은·박 만규. 2005.「세종 속담 전자사전에 대한 연구」, 한국사전학 제5호. 한국사전학회. 79~102쪽.
홍 재성. 2007.『세종사전의 전산적 활용을 위한 지침서』. 21세기 세종계획 전자사전 개발 분과.
홍 재성. 2007.『사업개요 및 사전설명서』. 21세기 세종계획 전자사전 개발 분과 성과물 CD.
홍 재성. 2007.『세종전자사전 최종 보고서』. 문화체육관광부.
황 순희. 2009.「상위 온톨로지를 이용한 명사의 의미 기술: KorLex Noun 1.5와『세종사전』의미 부류의 매핑을 중심으로」, 언어연구 Vol.25. 2. 한국현대언어학회. 361~387쪽.

Habash, N., Dorr, B. & Traum, D. 2003. Hybrid natural language generation from lexical conceptual structures. *Machine Translation*, 18. 2. 81~128.
Halliday, M. A. K. 1985, 2014(revised by Matthiessen, C.) (4th edition), *Halliday's Introduction to Functional Grammar*. (1985, 1st edition), Routledge.
Igor Mel'čuk, et al. 1999. Dictionnaire explicatif et combinatoire du franc ¸ais contemporain. Recherches lexico-sémantiques IV.
Mel'čuk, Igor A. 1996. "Lexical Functions and Lexical Inheritance for Emotion Lexemes in German", Lexical Functions in

Lexicography and Natural Language Processing. (Wanner, Leo, ed.) 209~278.

Iordanskaja, Lidija; Kim, Myunghee; Polguère, Alain. 1996. “Some Procedural Problems in the Implementation of Lexical Functions for Text Generation”, Lexical Functions in Lexicography and Natural Language Processing. (Wanner, Leo, ed.) 279~299.

조 은경
주소: [03722] 서울시 서대문구 연세로 50, 연세대학교 외솔관 214호
소속·직위: 연세대학교 국어국문학과 강사
누리편지: eunkyoung.jo@daum.net

<abstract>

A NLG Approach to Sejong Dictionary for Computational Feasibility

Jo Eun-kyoung

Sejong Dictionary is a knowledgebase platform which consists of various kinds of subcategory dictionaries and also a huge synthetic lexicon with morphologica, semantic and syntatic information. Although Sejong dictionary was the product of the project which was oriented to various application uses, it has been approached by a few studies of NLP and there is scarcely an effort to manipulate and use it.

This paper approaches to Sejong dictionary from NLG study and investigates the microstructure of it and then set out to format the lattice structure as a design of information makeup for developing NLG. The lattice formating is to make from the complex structure into the simple structure. This makes us have the information of various lexical entries instead of that of individaul lexical item and also makes it easy to construct natural language expression from lexicon. For one, we show that the representing structure of predicate and argument can be easily composed on lattice formatting. In addition, we show that the surface expressions also can be easily converted from the representing structures on it. Furthermore, in the process of this examination, we figure out the improvements in the macro and micro structure of this dictionary for a computational feasibility.

In all, we approached to Sejong dictionary from NLG perspective. we have examined the micro structure and tried to design the lattice formatting for computationally generating. During these process we arrange the improvements for computational lexicon.

* Key words: Sejong Dictionary, NLG: Natural Language Generation, Computational Lexicon, Lattice formatting, representing structure, surface expression.

〈논문 받은 날: 2015. 10. 12.(1차), 2015. 12. 21.(수정본)〉
〈심사한 날: 2015. 10. 28.~11. 24.(1심), 2016. 1. 19.~2. 18.(재심)〉
〈싣기로 한 날: 2016. 2. 29.〉