

표준화를 위한 일반 사전의 논리 구조

최병진*, 이재성*, 이운재*, 최기선*

*한국과학기술원 전산학과

**한국과학기술원 인공지능연구센터

The Logical Structure for Standardization of Printed Dictionary

Byung-Jin Choi*, Jae-Sung Lee*, Woon-Jae Lee* and Key-Sun Choi*

*Department of Computer Science, KAIST

**Center for Artificial Intelligence Research, KAIST

요 약

컴퓨터의 발달과 더불어 최근 자연언어 처리 분야의 일부에서는 일반 문서들(human-readable text)을 전자 문서(machine-readable text)화 하려는 노력이 이루어지고 있다. 이러한 연구 중 대표적인 것으로 사전을 전자문서화된 형태로 바꾸는 작업들 들 수 있는데, 외국에서는 이미 10 여년 전부터 이에 관한 연구가 꾸준히 진행되어 결실을 맺고 있다. 이에 반해 우리나라에는 아직 이에 견줄만한, 나아가 표준화할 만한 전자사전이 아직 개발되어 있지 않은 상황이다.

따라서 본고에서는, 일반 사전을 전자사전화 하기 위한 정형화된 논리적 구조를 검토해 보기로 한다.

1. 서론

컴퓨터로 하여금 문서 데이터를 인식하여 처리하고, 더 나아가서 다른 목적에 응용 하기 위해서는, 이 문서 데이터를 컴퓨터내의 기호 체계로 바꾸어 정형화하여야 한다. 컴퓨터가 처리할 수 있는 기호는 문서 데이터에 비해 적은 수이므로, 특별한 마크업(Markup) 체계의 개발이 필요하다.

마크업이란 문서의 형태 정보를 유지하기 위해 부가적으로 필요로 하는 정보 처리 명령을 말한다. 즉 복잡한 형태 정보를 표현

하기 위해 약속하는 일종의 기호이다. 이러한 마크업 체계의 예로는 SGML (Standardized Generalized Markup Language), HTML (HyperText Markup Language), LaTeX 등을 들 수 있다.

그런데 이러한 표시체계에서 항상 문제가 되는 것은 바로 표준화 문제이다. 표준화된 코드체계를 만들기 위한 노력은, 80년대 초 SGML 개발을 시작으로, TEI (Text Encoding Initiative)의 작성, 그리고 이를 바탕으로, 문서나 데이터를 컴퓨터가 읽을 수 있도록 인코딩 하려는 연구들이 그동안 지속적으로

추진되고 있다 [BRY88, SPE/BUR94, SPE/IDE 95].

이러한 노력의 주된 목적은 다양한 유형과 양식의 문서나 데이터를 표준적인 인코딩 방법에 따라 작성함으로써, 입력된 문서나 데이터의 정보가 응용 소프트웨어나, 하드웨어 또는 시스템에 의존하지 않고 독립적으로 사용되도록 하는데 있다.

여기서는 전자 문서화의 대표적 분야라고 할 수 있는 전자 사전 작성을 위한 노력으로서, 일반 사전의 형식을 분석, 분류하여 정형화하고, 이에 적합한 전자 사전의 논리적 구조를 설정함으로써 표준화된 형태의 전자 사전을 구현하였다. 표준화된 형태의 정보를 전자사전의 형태로 공유하게 되면, 사전 정보의 수집, 가공은 물론, 어휘 정보의 재활용 측면에서 상당한 효과를 기대할 수 있다.

2. 전산 사전 편찬학 (computational lexicography)의 추세

우선 전산 사전 편찬학에서 가장 많이 쓰이는 전자사전이라는 용어에 대해서 정의를 해보자. 여기서 이야기되는 전자사전이란 컴퓨터로 작업되는 모든 사전을 전자사전으로 정의하도록 한다¹. 이러한 폭 넓은 의미의 정의에는 크게 일반 사전의 출판을 위해 컴퓨터가 읽어서 처리할 수 있는 전산 사식

테입(type setting tape)의 형태로 만들어지는 기계 가독형 사전의 의미와, 기계 번역이나 형태소 해석, 문장 분석 등 자연언어 처리 프로그램에 사용되는 전산 처리용 사전의 의미가 모두 포함된다. 본 고에서는 전자 사전을 기계 가독형 사전과 전산 처리용 사전의 두 가지 모두가 포함된 의미에서 논하기로 한다.

약 30여 년 전부터 전산학적 처리 방법들이 사전 편찬학에 도입되어 이제는 보편적으로 사용되고 있다. 콩코던스 (concordance) 프로그램이나 on-line 문서 검색 시스템, 또는 사전 편찬자 워크벤치 (lexicographer's workbench) 등의 도움으로 사전 편찬 작업이 수월하게 되었는데, 전산학적 처리 방법의 도움으로 편찬된 사전으로는, Longman Dictionary of Contemporary English (LDOCE), Collins Birmingham University International Language Database (COBUILD), Brockhaus-Wahrig, Trésor de la Langue Française 등을 들 수 있다.

또한 이렇게 편찬된 사전을 전산 사식 테입 형태로 만든 기계 가독형 사전이 70년대에 들어 Olney 와 Ziff 에 의해 최초로 만들어졌는데, 바로 Marriam-Webster Seventh New Collegiate Dictionary 가 그것이다[OLN/RAM 72]. 이밖에 언어학적 코드와 분류 체계를 가지고 편찬된 LDOCE 사전도 기계 가독형 사전으로 만들어 졌다. 또한 The Oxford English Dictionary (O.E.D)사전의 경우에는 CD-ROM 의 형태로 사전을 저장하여 사용자로 하여금 손쉽게 사전을 이용할 수 있도록 하였다.

이처럼 텍스트의 형태가 기계 가독 형태로 입력이 되고 사용되면서, 기계 가독 형태로 저장된 어휘 정보들을 자연언어 시스템이나

¹[최기선 91]은 전자사전이란 말의 정의를 “컴퓨터의, 컴퓨터에 의한, 컴퓨터를 위한 사전” 이라고 정의하면서, 컴퓨터가 읽을 수 있고, 컴퓨터가 가공, 처리할 수 있고, 디스크나 CD-ROM 에 저장되어 컴퓨터에서 사용할 수 있는 사전이라고 설명한다.

기계번역 시스템 등에 사용하려는 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, 전산 사전 편찬학이나 전산 언어학에서는 단어의 의미를 정형화된 형태로 나타낼 수 있는 방법을 찾으려 노력하였다.

LDOCE 나 COBUILD 사전의 경우와 같이 기계 가독 형태로 이미 형식화가 되어 있는 경우에는 동의어, 상위어, 반의어 등과 같은 의미 관계를 추출하기는 비교적 용이하다. 그러나 풀이, 정의, 또는 문맥이나 예문에서 정형화된 형태의 의미 관계를 추출하는 것은 쉽지 않다.

한편, 전산 언어학에서는 일상 텍스트로부터 이러한 어휘 정보를, 특히 단어 의미에 관한 정보를 추출하려는 연구가 일찍부터 진행되었다. 이러한 연구는 구조주의 언어학의 문맥에서 요구되는 방법론과 매우 유사하다. 즉, 문제가 되는 요소의 환경을 분석하여 어휘 정보를 추출한다는 점에 공통점이 있다. 예를 들어, 공기 분석(co-occurrence analysis)을 위한 여러 가지 다른 방법론이 있고, 한 요소의 특정 의미 영역을 정하기 위한 방법들, 또한 연어 분석(collocational analysis)이나 Wahrig 가 제안한 핵심구의 변형에 관한 방법들을 생각 할 수 있다.

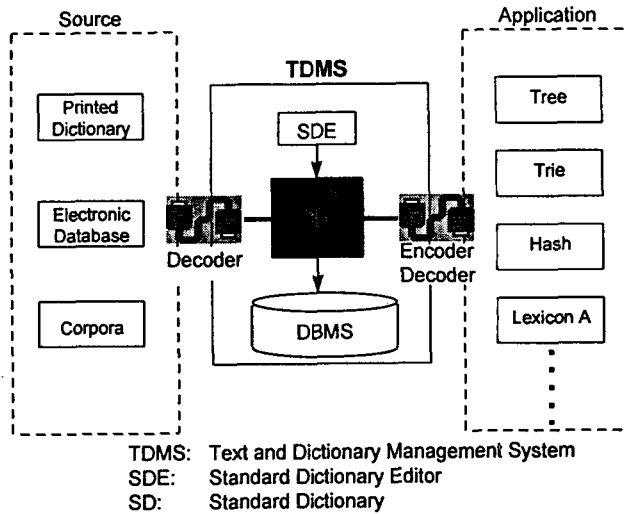
그리고 어휘 풀이, 정의, 예문에서 의미 관계를 발견하기 위해서는, 어휘를 인식하고 또한 풀이 텍스트의 구조를 파싱할 수 있는 프로그램이 만들어져야 하는데, [ALS89], [BOG91, 94], [COP90], 또는 [NEF/BOG89] 등의 연구가 이에 해당한다. 이러한 연구를 위해서는 기존 사전을 기계 가독 형태의 전

사 사전화하는 작업이 우선되어야 하고, 이것이 가공되어 어휘 연구와 자연언어 처리에 응용되어야 한다. 다시 말하면, 기계 가독형 사전이 만들어져 있다면 이를 바탕으로 전산 처리가 가능한 사전을 만드는 것이 용이하다는 것이다. 한마디로 최근 전산 사전 편찬학 연구의 특징을 요약한다면, 기존의 기계 가독형 사전을 바탕으로 자연언어 처리에 필요한 다른 형태의 사전을 가공하는 것이라 할 수 있다.

3. 표준 사전 개발과 재사용 가능성

세계적인 연구 추세와는 달리 우리 나라에는 이렇다 할 표준적 기계 가독형 사전이 아직 없는 실정이다. 따라서, 외국에서처럼 기계 가독 형태로 되어 있는 사전을 다시 파싱하여 이를 정규화 형태로 바꾸는 단계는 아직 생각도 하지 못하고 있는 실정이며, 또한 기계 가독형 사전이 어떠한 형태의 내부 구조를 지녀야 하는 지에 대한 연구도 아직 준비 단계이다. 전산 처리용 사전으로는 개개의 연구 개발자가 연구 목적과 시스템에 따라 각각 전산 처리용 사전을 개발하고 있지만, 이러한 사전 개발 작업에는 많은 시간과 비용이 든다. 또한 하나의 목적으로 개발된 사전이 있다 하더라도, 그 사전을 다른 목적에 이용하는 것도 그리 간단하지는 않다.

이러한 문제점을 극복하기 위해 여기서는 일반 국어사전의 내용을 인코딩할 수 있는 표준 사전 형태의 내부 구조를 제안하고 일반 사전을 이에 맞게 정형화 하고자 한다. 만일 우리가 표준 사전 형태로 저장된 어휘 정보를 가지고 있다면, 우리는 이것을 바탕으로 새로운 종류의 기계 가독형 사전이나



[그림 1]. 표준화된 전자 사전의 어휘 정보 구축과 재사용

전산 처리용 사전으로 가공할 수 있다. 재사용 (reuse)이 가능하도록 어휘 정보를 표시하는 일이 바로 일반 사전의 구조를 표준화하는 주된 목표의 하나이다. (그림 1 참조)

어휘 정보를 재사용한다는 의미는 두 가지 면에서 생각할 수 있다. 즉, 이미 존재하는 어휘 정보를 새로운 형태로 변환하는 것 - 예를 들어 인쇄된 사전의 어휘 정보를 기계 가독 형태로 바꾸는 것 - 과, 다른 응용 프로그램에 맞게 기존의 어휘 정보를 표준 사전에서 추출하는 일을 생각할 수 있다. [KUG/AHM/THU95] 이를 위해 우리는 ‘텍스트 및 사전 관리 시스템² (TDMS: Text and Dictionary Management System)’을 개발하고, 표준화된 사전을 위해 ‘표준 사전 마크업 언어³ (SDML: Standard Dictionary Markup Language)’라는 것을 정의하였다. 이 표준 사전

마크업 언어로 표준 사전을 위한 ‘문서유형 양식 (DTD: Document Type Definition)’을 작성하고, 이에 맞도록 ‘표준 사전 (SD: Standard Dictionary)’을 만들게 된다. ⁴ 표준화된 사전은 응용 프로그램에 관계없이 표준적인 어휘 정보를 지니고 있으므로, 사용자의 요구에 따라 언제든지 필요한 정보의 추출 및 가공이 가능하다.

여기서는 표준화로 삼을 수 있는 기계 가독형 사전의 구현을 우선적인 목표로 하고, 이를 바탕으로 하는 전산 처리용 사전의 구현은 다음으로 미루도록 한다. 이를 위한 기본 작업으로, 기존의 일반 사전을 가지고 이들의 거시(macro) 구조와 미시(micro) 구조를 비교 검토하여, 기계 가독형 사전의 내부 구조 형태를 하나의 표준안으로 제안한다.

4. 일반 사전의 구조 분석

4.1 일반 사전의 두 가지 유형

전통적인 사전 편찬학에서는 단어의 의미를

^{2,3,4} TDMS 와 SDML 및 SD 에 관련된 자세한 내용은 [이재성 96]을 참조

표현하는데 두 가지 방법론이 있다. 즉, 의미적(semasiological) 접근 방법과 개념적(onomasiological) 접근 방법이다

의미적 접근 방법에 의한 사전은 우리가 보통 사용하는 일반 사전으로, 이 사전에서는 한 언어의 어휘가 알파벳 순서로 배열되어 있다. 이러한 일반 사전은 사용자에게 어느 특정 단어나 기호의 여러 가지 의미에 관한 정보를 제공함을 목적으로 한다. 일반 사전에서는 의미가 다음과 같이 여러 가지 방법으로 설명된다.

- 정의(definitions),
- 풀이(paraphrases),
- 동의어, 반의어, 상위어, 하위어,
- 기타 관련어(cross-reference),
- 용례,
- 기타 설명.

의미 설명에서 가장 중요한 방법이 풀이이다. 풀이는 어휘를 정의하며, 대개 명제의 논리적 구조를 지닌다. 형식 의미론이 보여준 바와 같이, 명제 구조(propositional structure)는 술어와 논항을 가진 유형의 형식적인 표현으로 변형될 수 있다. 이러한 술어 논리적인 표현은 다시 말해 단어 의미의 정형화된(canonical) 표현으로 사용될 수 있는 의미 관계인 것이다.

반면, 개념 사전은 단어의 의미에 관한 정보를 표현하는 것이 아니고, 단어가 어떤 의미 영역, 또는 의미 개념에 속하는가에 관한 정보를 표현한다. 따라서 개념 사전은 어휘의 개념 순서를 위한 형식적인 체계를 제시해야 한다. 이 사전에서는 어휘를 주로 상위어(hyponymy), 동의어(synonymy), 그리고 반의어(antonymy)라는 세 개의 의미 관계에 의해 이루어진 개념 스키마에 의해 배열한다. 아마도 가장 잘 수립된 개념 스키마는 영어

에서는 Roget의 “Thesaurus of English Words and Phrases”를 들 수 있으며, 독일어에서는 R. Hallig와 W.v. Wartburg [HAL/WAR63]에 의해 만들어진 “사전학 토대로서의 개념 체계 (Begriffssystem als Grundlage für die Lexikographie)”를 들 수 있다. 이러한 개념 사전은 자동 정보검색이나 문서 시스템에 사용된다.

4.2 미시 구조와 거시 구조

사전을 편찬할 때에 표제어가 어떠한 순서에 의해 배열되는지에 대한 구조가 거시 구조이다. 즉 표제어의 배열에 있어서 자모의 종류와 이들 사이에 어떠한 차례가 정해져야 하는가가 바로 거시 구조를 어떻게 이루어야 하는지의 문제이다. 또한 표제어에 대한 정의가 어떠한 형태로 구조를 이루고 있는가가 미시 구조에 관한 연구이다. 여기서는 우선 인쇄된 사전을 거시 구조보다는 미시 구조적인 관점에서 사전 어휘 항목이 어떠한 구조를 이루고 있는지를 살펴보겠다.

일반 사전의 경우 표제어는 가나다 순이나, 알파벳순으로 배열되어 있다. 표제어가 동철자어(homograph)일 경우에는 각 표제어가 어깨번호(1, 2, 3, ...)와 함께 따로 나타난다. 영어의 경우 품사에 따라 다르게 발음되는 표제어가 있는데, 이러한 경우에도 따로 표제어로 둔다. 단, 표제어가 두 개의 품사를 지니고, 의미의 설명에 있어서 하나로 충분할 경우에는 따로 표제어로 등록하지 않는다.

또한, 하나 이상의 단어로 이루어진 구나 복합어의 경우에는 이를 따로 표제어로 설정하고 있다. 그러나 관용어의 경우에는 특정구가 따로 표제어로 나타나지 않고, 다른 표제어의 내부에 관용어로서 특수하게 표시되어 있다.

기존의 일반 사전에 개별 표제어에 나타나 있는 내용을 보면, 대체로 다음과 같은 정보들이 나타나 있다.

국어 사전: 표제어, 표제어 번호, 한자|로마자, 발음, 품사, 활용형, 어원(외래어 어원), 전문어 표시, 용법, 의미, 관련 표제어(cross-reference), 관련 표제어 관계 (준말, 늘임말, 동의어, 상위어 등), 수치, 주의, 용례, 출처, 속담, 관용어, 파생어.

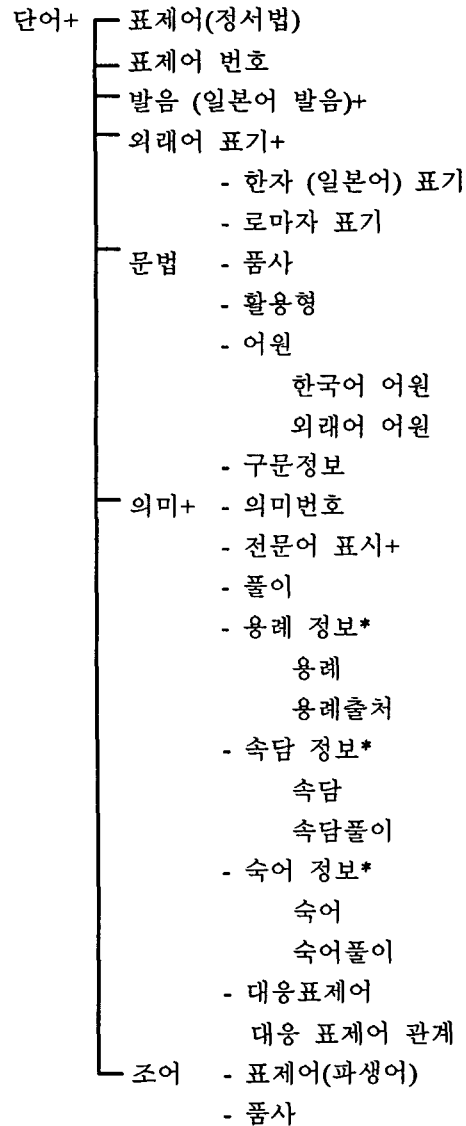
영어 사전: 표제어, 표제어 번호, 발음, 대음 표제어, 품사, 활용형, 다의어 번호, 문법적 정보, 의미, 용법, 용례, 관용어, 공기 정보.

5. 표준 사전을 위한 일반 사전의 DTD

개개의 사전 표제어들은 구조화된 형태를 지니고 있는데, 이러한 형태는 사전마다 매우 다양하게 나타난다. 다른 사전이 다른 구조를 지니는 것은 극히 자연스러운 일이지만, 같은 사전에서조차도 다양한 형태의 구조가 종종 나타난다. 이처럼 각 어휘 항목들의 구조가 사전마다 다르고, 또한 한 사전에서도 매우 다양하게 나타나기 때문에, 일반 사전을 정형화하기 위해서는 모든 사전에 응용될 수 있는 공통적 구조를 찾아내어 인코딩하는데 중점을 두어야 한다⁵⁾. 여기서 주의해야 할 문제는 우리가 중요한 정보들을 효과적으로 다시 추출하기 위해 사전의 어휘 정보들을 컴퓨터에 어떠한 방식으로 코드화 하느냐는 것이다.

사전의 DTD 작성을 위해서 우리는 앞절에서 사전 표제어의 설명에 나타나는 요소들

을 기술하였다. 사전의 DTD 작성을 위해서는 사전 표제어의 설명에 나타나는 내용을 최소한의 요소로 나누어 이것을 표준 사전의 형식에 맞게 개개의 데이터 자리에 저장한다. 사전의 어휘 정보를 세분화하면 할수록 우리는 세분화된 정보를 더 잘 사용할 수가 있다.



[그림 2]. 어휘 정보의 계층 구조

일반 사전에 포함되는 정보들을 나타내기 위해서는 다음과 같은 내용의 정보들이 주어져야 한다. 우선 최상위 요소로 표제어의 형태에 관한 정보가 있어야 한다. 여기에는 발음, 정서법, 외래어 표기에 관한 정보가 속한다. 이외에 동일한 계층에서 문법, 의미, 조어 (파생어)에 관한 정보가 주어지며, 이

들은 각각 하위 계층에 더 세분화된 정보를 지닌다. 이를 트리 형태로 나타내면 [그림 2]에서와 같이 계층 구조적인 특성을 볼 수 있다. 그리고 이러한 트리 형태를 바탕으로 표준 사전 형식을 SDML의 일부로 나타내면 [그림 3]과 같다.

<code><!DOCTYPE sd SYSTEM "sdml.dtd" [</code>	
.....	
<code><!ELEMENT entry -- (wname, body)+ ></code>	사전 어휘
<code><!ELEMENT body -- (hwn fn pr)+ , (gi si)+ , wfext* ></code>	
<code><!ELEMENT hwn - O (#PCDATA) ></code>	표제어 번호
<code><!ELEMENT fn -- (ch rom) ></code>	외래어 표기
<code><!ELEMENT ch -- (#PCDATA) ></code>	한자 표기 (일본어 포함)
<code><!ELEMENT rom -- (#PCDATA) ></code>	로마자 표기
<code><!ELEMENT pr -- (#PCDATA) ></code>	발음 (일본어 발음)
<code><!ELEMENT gi -- (pos , infl? , etym? , syn?) ></code>	문법 정보
<code><!ELEMENT pos -- (#PCDATA) ></code>	품사
<code><!ELEMENT infl - O (#PCDATA) ></code>	활용형 정보
<code><!ELEMENT etym -- (ls fs) ></code>	어원 정보
<code><!ELEMENT ls - O (#PCDATA) ></code>	어원
<code><!ELEMENT fs - O (#PCDATA) ></code>	외래어 어원
<code><!ELEMENT syn -- O (#PCDATA) ></code>	구문 정보
<code><!ELEMENT si -- (num , term+ , expl , usage* , saying* , idiom* , althw*) ></code>	의미 정보
<code><!ELEMENT num - O (#PCDATA) ></code>	의미 번호
<code><!ELEMENT term -- (#PCDATA) ></code>	전문 분야
<code><!ELEMENT expl -- (#PCDATA) ></code>	풀이
<code><!ELEMENT usage -- (us_ex , us_source?) ></code>	용례 정보
<code><!ELEMENT us_ex - O (#PCDATA) ></code>	용례
<code><!ELEMENT us_source - O (#PCDATA) ></code>	용례 출처
<code><!ELEMENT saying -- (sng , sng_expl?) ></code>	속담 정보
<code><!ELEMENT sng - O (#PCDATA) ></code>	속담
<code><!ELEMENT sng_expl - O (#PCDATA) ></code>	속담 풀이
<code><!ELEMENT idiom -- (idm , idmexpl) ></code>	숙어 정보
<code><!ELEMENT idm - O (#PCDATA) ></code>	숙어
<code><!ELEMENT idm_expl - O (#PCDATA) ></code>	숙어 풀이
<code><!ELEMENT alt_hw -- (hw , alt_hw_rel) ></code>	대응 표제어
<code><!ELEMENT alt_hw_rel - O (#PCDATA) ></code>	대응 표제어의 관계
<code><!ELEMENT wf_ext -- (hw , pos)+ ></code>	조어 확장자(파생어, 품사)
<code>] ></code>	

[그림 3] SDML로 나타낸 표준 사전 형식

6. 맺음말

대부분의 자연언어 처리 응용 프로그램들은 특정한 시스템에 맞게 정의되어 구현되어 있어서, 이식성 및 적응성이 부족하다. 전자사전의 경우에도 예외가 아니며, 다른 프로그램이나, 다른 시스템에서 다시 사용한다는 것이 쉽지가 않다. 하지만 사전이 표준화된 형태로 표현 저장되어 있다면, 필요한 프로그램에 맞게 쉽게 변환하여 사용할 수 있다.

지금까지 이러한 표준화된 사전의 논리적 내부를 설정하기 위해 기존의 사전들을 분석하였고, 이를 바탕으로 표준화된 DTD를 작성하였다. 따라서, DTD를 기준으로 하여 사전이 구성될 경우에, 언어학적으로 유용한 많은 사전(동의어 사전, 반의어 사전, 파생어 사전, 용례 사전, 속담 사전, 형태소 해석 사전, 속어 사전, 외래어 사전 등)을 자동적으로 가공할 수 있을 것이다. 또한 이를 더 확장하여, 각 사전 표제어의 자질들을 정규화된 형태로 나타낸다면, 전산언어학적인 면에서는 자연언어 처리 시스템에 응용하여 사용할 수 있다. 이를 위한 노력은 앞으로의 과제로 남겨두기로 하며, 여기서는 일반 사전의 형식등을 연구하여 표준적인 사전 형식을 제안하는 것으로 제한한다. 여기서 제안된 표준 사전 형식은 다양한 전자 사전의 요구를 충족시켜 줄 뿐만 아니라, 표준 사전 형식의 연구와 전문 용어 사전의 표준화에도 활용될 수 있다고 본다.

또한 앞으로 이러한 표준화된 사전 구조에서, 정보를 공유하고 공유된 정보를 유전시키는 비단조 논리 (non-monotonic reasoning)의 도입도 바람직하다고 본다.

참고문헌

- [이재성 96] 이재성, 최병진, 이운재, 최기선 “텍스트 및 전자사전 관리시스템의 설계”, 한글 및 한국어 정보처리, 1996.
- [최기선 91] 최기선, “전자사전 연구 개발과 정보 서비스로의 활용”, 우리말 정보화 잔치, 1991.
- [ALS89] Alshawi, H. (1989), “Processing Dictionary Definitions with Phrasal Pattern Hierarchies”, in Boguraev, B. and E. Briscoe(eds.), 153-170.
- [BOG/BRS89] Boguraev, B. and E. Briscoe(1989), Computational Lexicography for Natural Language Processing, Longman Limited, Harlow and London
- [BOG91] Boguraev, B. (1991), Special Issue on Computational Lexicons, International Journal of Computational Lexicography 4
- [BOG94] Boguraev, B. (1994) Machine-readable dictionaries and computational linguistics research, in A. Zampolli, N. Calzolari, and M. Palmer(eds.)
- [BRY88] Bryan, M.(1988), SGML: An Author's Guide to the Standard Generalized Markup Language, Addison-Wesley
- [COP90] 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

- [HAL/WAR63] Hallig, R., von Wartburg, W. (1963), Begriffssystem als Grundlage fuer die Lexicographie. Versuch eines Ordnungsschemas, Berlin.
- [IDE/VÉR95] Nancy Ide & Jean Veronis (1995). Text Encoding Initiative: Background and Context, Kluwer Academic Pub.
- [KUG/AHM/THU95] Kugler, M., Ahmad, K., Thurmair, G. (eds.) (1995), Translator's Workbench: Tools and Terminology for Translation and Text Processing, Brussels-Luxembourg.
- [NEF/BOG89] Neff, N. and B. Boguraev (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
- [OLN/RAM72] Olney, J./Ramsay, D.(1972), From machine-readable dictionaries to a lexicon tester: progress, plans, and an offer. In: Computer Studies in the Humanities and Verbal Behavior 3,4, 1972, 213 -220
- [SPE/BUR94] Sperberg-McQueen, C.M. and L. Burnard (eds.) (1994), Guidelines for Electronic Text Encoding and Interchange, TEI P3, Oxford, Chicago.
- [SPE/IDE95] Sperberg-McQueen, C.M. and Ide, Nancy M. (1995), The TEI: History, Goals, and Future, in Text Encoding Initiative: Nancy Ide & Jean Veronis (1995), Kluwer Academic Pub..
- [WAH67] Wahrig, G.(1967), Neue Wege in der Wege in der Woerterbucharbeit, Hamburg.
- [WAH84] Brockhaus-Wahrig(1984), Deutsches Woerterbuch in 6 Baenden, Wiesbaden.
- [ZAM/CAL/PAL94] Zampolli, A., Calzolari, N., and M. Palmer(1994), Linguistica Computazionale, Vol. IX.X Current Issues in Computational Linguistics: in Honor of Don Walker, Kluwer Academic Publishers, Dordrecht.