

한국어사전 인코딩체계의 확장에 관한 연구: MARTIF와 TEI를 중심으로*

Extending the MARTIF and TEI for Korean Lexical Entities

백지원(Ji-Won Baek)**, 최석두(Suk-Doo Choi)***

초 록

본 연구는 사전이 가지는 특징과 구조를 분석하고, 모든 한국어 사전항목을 손실 없이 인코딩 할 수 있는 마크업 체계를 제안하는 데 목적이 있다. 이를 위하여 국제표준인 MARTIF와 용어데이터의 상세한 인코딩형식을 제공하는 TEI를 분석하였다. TEI는 용어데이터베이스용 형식과 인쇄사전용 형식을 제공하고 있으며, 용어데이터베이스용 형식은 MARTIF와 거의 유사한 반면 인쇄사전용 형식은 확연히 다른 형식이었다. 그리고 실제 사전 및 용어 관련 자료를 분석하여 한국어사전의 종합적인 미시구조를 파악하고, 그 미시구조를 MARTIF와 인쇄사전용 TEI에 대응시켜 적용가능성을 시험하였다. 그 결과, 두 형식 모두 한국어사전의 미시구조와 정확하게 매핑되지 않았으나 확장에 따라 두 형식 모두 사용가능하다는 결론을 얻게 되었으며, 한국어사전의 특징을 살릴 수 있도록 하기 위하여 파악된 한국어사전의 미시구조 내용을 포함시켜서 새로운 두 가지 확장형식을 제안하였다.

ABSTRACT

The purpose of this study is to present a scheme to encode all possible lexical entities in dictionaries, glossaries, encyclopedias, and thesaurus, etc. First, it discussed the nature and structure of dictionaries. Second, two current major terminological data encoding schemes, MARTIF and TEI were analyzed in terms of their flexibility for extension to encompass all lexical entities. Third, an integrated microstructure of dictionaries was presented and compared with the MARTIF and TEI for print dictionaries. Then, the need and 17 suggestions for extended MARTIF and TEI formats were addressed with specific cases, which combined with the suggestions from two studies concerning MARTIF and TEI DTD modification for the markup of Korean dictionary entries.

키워드: MARTIF, TEI, 사전, 용어, 사전항목, 미시구조, 데이터교환, 인코딩, 마크업

* 이 논문은 2001년도 두뇌한국21사업에 의하여 지원되었음.

** 이화여자대학교 대학원

*** 이화여자대학교 문헌정보학과 교수(sdchoi@mm.ewha.ac.kr)

■ 논문 접수일: 2001년 5월 21일

■ 게재 확정일: 2001년 6월 10일

1 서론

1.1 연구의 목적 및 방법

용어데이터에는 인쇄용 사전데이터와 용어 데이터베이스가 있다. 이들은 편찬배경과 목적에 따라 그 종류, 내용, 체제 및 관련항목의 구성이 다양하여 매우 복잡한 구조를 가지고 있다. 따라서 용어데이터들을 가공, 재생산 혹은 교환하고자 할 때 필수적인 것이 표준기술형식이라 할 수 있다. 현재 용어데이터의 교환이나 기계처리를 위해 마련된 구조로는 국제표준으로 제정된 '기계가독형 전문용어 교환형식' (ISO/DIS 12200:1999 Machine-Readable Terminology Interchange Format—Negotiated Interchange. 이하 MARTIF라 한다)과 TEI(Text Encoding Initiative. 때로는 Text Encoding for Interchange라는 의미로 쓰이기도 한다) DTD를 들 수 있다. 또한 이들 형식에 대한 용어데이터의 표준범주를 제시하고 있는 '전문용어 처리를 위한 컴퓨터응용 - 데이터 범주' (ISO/DIS 12620:1999 Computer Applications in Terminology—Data Categories)가 있다.

용어데이터형식은 일부의 사전에만 나타나 특이한 사전항목이라 할지라도 사전정보로서의 가치가 있는 것이라면 표현할 수 있어야 한다. 특히 한국어사전이 갖고 있는 사전항목이라면 그 정보를 잃고 싶지는 않을 것이다. 따라서 국내는 물론 국제적인 용어데이터의 원활한 교환과 이용을 위해서는 정보의 손실 없이 용어데이터의 정보를 있는 그대로 살리면서 국제적인 양식에 부합되는 인코딩체계를 정의할

필요가 있다. 사전데이터의 모든 항목을 포괄적으로 처리할 수 있게 되면 그 데이터를 특정 목적용 사전으로 변환할 때 혹은 다른 매체나 형식으로 변경할 때 포함되지 않는 항목을 제외시킬 수 있기 때문이다.

이를 위하여 본 연구에서는, 1) ISO/DIS 12200:1999, ISO/DIS 12620:1999, TEI의 내용을 분석하고, 2) 한국어사전을 분석하여 사전항목의 기본요소와 개별 사전이 가지는 특수요소, 그리고 망라적인 사전데이터의 미시구조를 파악하고, 3) 이 파악된 미시구조의 인코딩을 위해 현행 사전요소 인코딩체계로 사용할 수 있는 MARTIF와 TEI의 적용가능성을 고찰하고, 4) 한국어사전의 미시구조 내용을 포함시켜서 새로운 확장형식을 제안하고자 한다.

1.2 선행연구

ISO/TC 37, TEI, LISA의 작업그룹 중 하나인 OSCAR가 수행하고 있는 SALT 프로젝트에서는 인간중심적인 용어데이터베이스와 기계번역용어로부터 고도로 구조화된 용어의 표현, 분배 및 교환을 촉진하기 위해 XLT (XML Representation of Lexicons and Terminologies)를 개발하였다(<http://www.ttt.org/salt/index.html> 참조). 또한 Brigham Young University Translation Research Group(BYUTRG)과 Kent State University Institute for Applied Linguistics (KSUIAL)의 공동연구로 CLS Framework (<http://www.ttt.org/clsframe> 참조), MSC¹⁾ (MARTIF with Standardized Constraints), 그리고 그 데이터베이스인

Reltef 등을 개발하였으며, 이들은 모두 MARTIF를 응용한 용어데이터 관련 형식이다. 한편 용어공학 분야에 있어서 ISO/TC 37과의 협력 하에 여러 산업표준을 만들고 있으며 그 대표적인 것으로는 LISA OSCAR가 책임을 맡고 있는 TMX(Translation Memory eXchange)(<http://www.lisa.unige.ch/tmx/index.html>) 참조를 들 수 있다. 이것은 데이터의 손실 없이 도구와 번역자간의 용어정보교환을 손쉽게 하기 위한 것이다.

유럽에서는 지리적 특성상 언어정보의 기계 번역 및 교환에 관한 다양한 연구가 활발히 진행되고 있다. 대표적인 용어데이터 관련 표준 작업으로는 OTELO 프로젝트에서 개발한 OLIF(Open Lexicon Interchange Format)(<http://www.otelo.lu>) 참조를 비롯하여, Interval 프로젝트에서 개발한 IIF, Inesterm 프로젝트에서 개발한 Geneter 등이 있다(Budin and Melby 2000).

국내의 용어데이터 교환 및 표준 관련 연구로는 국어사전에 나타나는 사전항목들을 TEI 확장규칙을 이용하여 표현한 연구(Kang 1998)가 있으며, MATER(ISO.6156:1987 Magnetic Tape Exchange Format for Terminology)와 TIF(ISO/DIS 12200의 전신)를 비교하고, TEI를 이용한 SGML 기반의 'TEI-K'를 개발하여 한글 전문용어 교환 데이터베이스형식으로서 제안한 연구(김성혁 1996)가 있다. SGML을 이용하여 전자사전의 논리적 구조를 제시한 연구(최병진 외 1996)와 텍스트 및 사전관리시스템을 위한 표준 사전표기언어(SDML: Standard Dictionary Markup Language)를 설계한 연구(이재성 외

1996)가 있다. 또한 XML을 이용한 문헌정보학 용어사전의 항목구조화 및 브라우저설계에 대한 이론을 제시한 연구(김진희 2001)가 있으며, 전자사전 및 용어데이터베이스 구축에 관한 다수의 연구가 있다(문화관광부 2000a; 오현주 1998; 윤성희 1997; 이상곤 1998; 이상섭, 송만석 1992; 정휘용 1998). 한편 한국 정보통신기술협회(2000a)는 정보통신단체표준의 하나로 MATER를 번역하여 『용어/사전 편찬용 자기테이프 교환형식』으로 내놓았으나 이는 특수한 경우를 제외하고 최근에는 거의 쓰이지 않는 방식으로 알려져 있다.

용어데이터 교환 및 관리 표준형식에 대한 국외의 연구는 주로 영어, 독일어, 프랑스어 등 서양의 주요언어에 초점을 두고 마련된 것이어서 한글데이터의 특성을 제대로 반영하지 못하고 있다. 또한 전술한 바와 같이 국내에서도 다양한 연구성과가 발표되고 있으나 일반적으로 통용될 수 있는 용어데이터 표준형식을 만들지 못하고 있는 실정이다. 국내의 용어데이터 형식은 원칙적으로 국제적으로 널리 받아들여지고 있는 표준을 바탕으로 미시구조, 저장구조, 기술방법 등이 연구되어야 하지만 이와 같은 관점에서의 관련 연구가 미흡하기 때문이라 사료된다.

1) MSC(MARTIF with Standardized Constraint; 이전의 Blind MARTIF의 개념과 유사)는 협상교환용 형식 MARTIF의 한계를 보완하기 위한 ISO의 표준(ISO/CD 16642:1999 Computer Applications in Terminology—Meta Model for Representing Terminology Data Collections)으로의 채택 가능성이 논의되고 있다. ISO/CD 16642에 관해서는 다음의 주소를 참조하면 된다. (<http://www.loria.fr/projets/TMF/>)

1.3 한국어 사전

1.3.1 사전의 종류

사전은, 일정한 표제어에 대하여 음운·문법·의미 등의 정보를 압축하고 체계적으로 제시하여, 그 표제어들을 일정한 순서로 배열한 하나의 텍스트를 말한다. 선정된 표제어와 관련된 데이터들이 모여 하나의 레코드를 이루고, 그러한 레코드들이 적게는 수백에서 많게는 수억 개가 모여 집합적인 의미의 용어데이터가 된다. 이와 같은 사전의 구조와 형식은 사전의 편찬목적과 종류에 따라 달라지게 된다.

사전은 보는 측면에 따라 다음과 같이 여러 갈래로 그룹화 할 수 있다. 1) 편찬목적, 기능, 규모 등에 따라 순수언어사전, 백과사전적 언어사전, 전문용어사전, 고유명사사전 등. 2) 언어단위의 최초 출현 및 통시적 변천에 대한 史的 정보를 중점으로 한 通時的 사전과 단일시기의 언어상태에 대한 정보를 제공하는 共時的 사전. 3) 표제어에 대한 선별성의 여부에 따라 일반어휘 이외에 전문용어, 신조어, 방언, 속어 등을 가능한 한 개방적으로 표제어로 실어 한 언어의 총체를 기술하고자 하는 포괄적 사전(확장형 사전)과, 방언사전, 은어사전, 속담사전, 유의어사전(시소러스), 반의어사전, 개념사전 등과 같이 한 언어의 일부 또는 특정 부분만을 표제어로 삼아 제한적 기술을 꾀하는 부분적 사전(선별형 사전). 4) 내용에 따라 일반사전과 특수사전(방언사전, 전문용어사전, 속어사전, 동사사전, 고유명사사전, 속담사전 등). 5) 단어구조에 따라 매크로구조사전(시소러스, 용어해설집, 전문용어사전, 맞춤법사전, 역순사전 등)과 마이크로구조사전(일반사전).

6) 사전체계에 따라 개념사전(의미사전, 관념사전, 시소러스 등)과 일반사전. 7) 사전의 성격에 따라 기술사전과 규범사전(새 맞춤법·표준어에 따른 바른말사전, 바른말 고운말사전, 표준 한국어 발음대사전, 표준 국어대사전 등). 8) 사전내용의 전문성 및 학술적 성격에 따라 학술사전(학문별 용어사전)과 대중사전(일반 언어사전, 꽃사전, 고사성어사전, 잡학사전 등). 9) 의미정보의 성격에 따라 원형적 개념의 사전(역사적 언어사전, 포괄적 사전), 전형적 개념의 사전(표준사전, 탁상용 사전), 연장적 개념의 사전(백과사전, 전문용어사전). 10) 이용자의 속성에 따라 외국인 학습자용 사전과 모국어 사용자용 사전. 11) 사용된 언어수에 따라 단일어사전, 二重語사전, 다국어사전. 12) 이용자의 연령에 따라 성인용과 아동용 사전 등. 13) 사전매체에 따라 출판물사전, 전자사전, 시디롬사전 및 온라인사전 등으로 나눌 수 있다(최석두 2000).

각각의 주제분야 따라 사전의 갈래를 더 세분할 수도 있는데 언어사전을 내용의 전문성에 따라 분류한 예는 다음과 같다(조재수 2000, 107-108). 즉, 일반 뜻풀이사전(큰사전, 중사전, 소사전, 학생 국어사전 등), 표준말 맞춤법사전, 국어순화사전, 방언사전, 속어·비어·은어사전, 상말·욕설사전, 옛말사전, 이두사전, 시대언어사전, 속어(관용구, 성구)사전, 속담사전, 유의어·동의어·반의어사전, 낱말사전, 개별품사사전, 상징어휘사전(한국문화상징사전, 문학상징사전), 쓰임새(용례)사전, 詩語사전, 소설어사전, 문장표현사전, 어원사전, 발음사전, 말본사전, 외래어사전, 대역사전, 한자/한자말사전, 형태소사전 등이며, 이외

에도 많은 종류의 사전이 있다.

1.3.2 사전의 미시구조

사전의 구조는 표제어 전체와 관련된 거시구조와 하나의 표제어와 관련된 미시구조로 구분할 수 있다. 사전이 복잡하게 되는 것은 음성, 음운, 형태, 통사, 어휘, 의미, 화용 등 하나의 어휘항목을 기술하는 데 필요한 모든 층위의 정보를 이 미시구조에 포함시켜야 하기 때문이다. 뿐만 아니라 컴퓨터를 사전편찬에 활용할 수 있게 되면서부터 언어항목의 기술에 필요한 여러 층위의 언어정보를 거의 제한 없이 축적해 갈 수 있게 되었으며, 어휘항목의 수에도 제한을 받지 않게 됨에 따라 사전의 미시구조는 점점 복잡하게 되었다.

이 미시구조에 포함되거나 되어야 하는 내용에 관해서는 연구자에 따라 다르다. 대표적인 제안을 보면 다음과 같다.

김현권(1989)은 미시구조의 내용으로 발음 및 문법정보의 활용, 정의의 구성, 여러 가지 뜻의 규정과 배열, 한 단어에서 여러 가지 뜻의 상호관계, 그리고 이에 대응하는 제반 계열적 의미관계(유의어, 반의어, 상위어, 하위어 등), 각 뜻에 대응하는 표제단어의 결합관계와 통사적 관계 그리고 이를 표상할 수 있는 구문의 선택과 배열, 이에 따른 예문의 선정, 부차적으로 형태론적 정보와 고정성구의 제시 등이 포함된다고 밝히고 있다.

양단희(1992)는 전자사전에 저장되는 정보를 표제어, 어원, 발음, 품사분류, 형태소표시, 불규칙 여부, 본래 뜻, 바뀐 뜻, 파생된 뜻, 변한 뜻, 동의어, 해당분야, 시기, 관련어, 보충설명, 참고낱말 등으로 나누었다. 또한 영어사전

에는 오래 전에 도입된 동사의 유형별 세분화와 이에 따른 문형, 해당 낱말과 잘 어울려 사용되는 연결어미, 연어정보, 다양한 용례와 철저한 어원정보, 풍부한 유사어, 반의어를 수록한다고 하였다.

21세기 세종계획(문화관광부 2000a)의 체언사전의 미시구조 항목은 표제어의 관리정보 구획을 따로 두고, 최상위 정보구획으로 표기 정보 및 변이형정보(약어, 본디말, 오폭기), 외래어 원어정보, 형태정보(동형어, 파생어, 합성어), 관용표현정보를 포함시켰다. 의미정보구획으로는 기본정보구획으로서 영어 대역어, 전문분야, 내포의미, 가산/불가산, 의미부류 등을 포함시켰다. 어휘관계정보구획으로는 동의어, 반의어, 상위어, 하위어, 전체어, 부분어, 관련어 등을 다루었으며 마지막 통사정보구획으로 분류사(가산명사의 경우 개체단위, 분할단위 및 집합단위, 비가산명사는 분량단위), 특정 조사와 결합하는 명사의 경우 해당 조사, 그리고 연결 어미적 용법으로서 선행문과의 결합을 다루었다.

또한 서상규(2001)는 사전 미시구조의 항목으로 표제어 맞춤법정보, 발음, 품사 및 그 하위분류, 통사구조(활용형 정보), 뜻풀이, 용례, 유의어·반의어 등 관련어, 문형과 용례, 관용적 용법과 성구/속담, 관용적 용법과 연어구조 정보, 선택제약과 문법지식, 특수한 용법, 학술용어의 용법과 그 처리 등을 들었다. 그 외에도 바람직한 사전이 더 갖추어야 할 정보로서 빈도정보(중요도 정보), 단어의 역사, 어휘구성 정보, 관련어휘정보 등을 제시하고 있다.

사전데이터 DTD를 정의하는 데 있어서 기본문제는 포함되는 항목의 범위가 최대가 되도록

표제어	공중어*
1. 문어 및 구어형태 (Information on Written and Spoken Forms)	은어*
	고어*
	북한어*
발음	부류어
이형 발음	용어 빈도
외국어의 독음	용어 수준
장음*	분류사
형태적 분리*	특수한 범주에 속하는 용어
표기	용어의 상태
이형 표기	우선어
두문자어/축약형	승인어
한자 표기**	반대어
로마자표기**	폐기어
공식, 기호, 수식	대체어
잘못된 표기	다른 표제어에 대한 참조 (Cross References to Other Entries)
2. 문법(Grammatical Information)	보라참조*
	표준어*
품사	비표준어*
활용형(불규칙)*	반의어*
3. 어의(Sense Information)	동의어*
	유사어*
정의	줄인말*
대역어	본디말*
풀이내의 다른 뜻**	거센말*
풀이내의 구분설명	여린말*
의역, 공식적 정의	큰말*
4. 어원(Etymological Information)*	작은말*
	높임말*
한국어 어원	친근체*
한자 어원	낮춤말*
외국어 어원	변한말*
5. 기타 정보(Other Information)	참조어*
	최신어
용례(Examples)	용어관계
용례	병렬적 관계
속담*	속관계
속어*	전체-부분관계
용법(Usage Information and Other Labels)	사례관계
학문 영역*	다중계층관계
방언*	기타 관계
	6. 관련항목(Related Entries)

법례: 명조체 항목은 기존의 TEI를 적용하여 변환이 가능한 부분이고, 고딕체 항목은 기존의 TEI로는 나타내지 못하는 것으로 파악되어 확장이 필요한 부분이다. 한편, *표시는 Kang(1998), **표시는 김성혁(1996)의 연구에서 확장이 제안된 부분이다.

〈그림 1〉 한국어 사전의 미시구조

록 할 것인지, 아니면 최소가 되도록 할 것인지를 선택하는 일이다. 어느 사전에나 출현하는 최소한의 항목만을 엄격히 정의하는 것도 하나의 방법이다. 그러나 개개의 사전이 가지는 특성을 살려 정보를 잃지 않으면서 경우에 따라 선택된 정보를 교환하고자 할 때는 모든 종류의 사전항목을 포괄할 수 있는 DTD를 정의할 필요가 있다.

〈그림 1〉은 분야, 종류, 크기에 관계 없이, 사전, 용어집, 시소러스, 주제명표목표 등 용어데이터를 수록한 각종 자료의 항목을 실제로 조사한 결과이다. 이는 사전에 나타나는 데이터 항목들을 모두 포괄하는 것은 물론, 사전항목의 구성에 대한 선행연구에서 제시된 항목들을 추가함으로써 사전이 잠재적으로 가질 수 있는 항목들을 망라한 한국어사전항목의 미시구조를 나타낸 것이다. 이 미시구조에 나타나는 다양한 변수들은 사전과 사전 사이에만 존재하는 것이 아니라 하나의 사전 내에서도 표제어간에 무수히 발생한다. 또한 그러한 변수들이 사전에서 언제든지, 어떤 순서로든지, 또한 몇 번이든 등장할 수 있다는 사실 역시 고려되어야 한다. 더구나 아직 조사하지 못한 사전이나 앞으로 개발될 사전의 형태는 어떠한 형태로 표현될 수 있을지 알 수 없으므로(이재성 외 1996), 사전데이터의 인코딩에 있어서 이러한 모든 변수들을 포괄할 수 있는 마크업체계를 적용하는 것이 무엇보다 중요하다.

2 용어데이터 표준형식

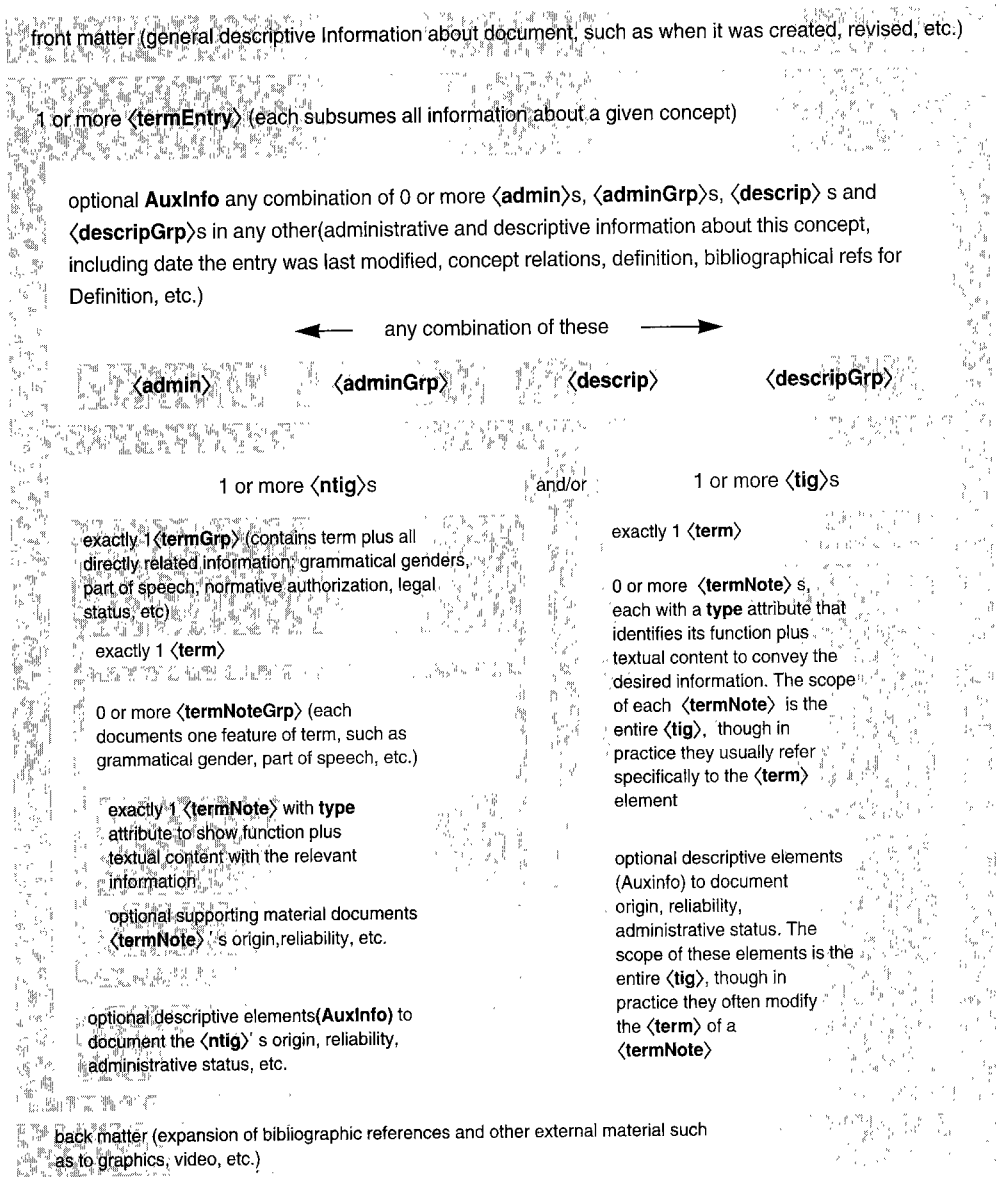
2.1 MARTIF

MARTIF는 시스템이나 하드웨어플랫폼에 독립적이고 효율적인 용어교환형식을 위한 것으로 수년에 걸친 TEI 및 LISA와의 긴밀한 협력 아래 ISO/TC 37에서 제정한 것이다. MARTIF는 ISO 8879:1986 Information Processing-Text and Office Systems—Standard Generalized Markup Language (SGML)를 기반으로 데이터를 구조화하고 적합한 마크업태그를 제시함으로써 용어교환을 위한 데이터의 형식화를 가능케 한다. MARTIF는 용어교환을 위한 큰 틀만을 제시한 것으로 실제로 용어데이터의 상세한 내용을 나타내기 위해서는 구체적인 데이터모델이 있어야 한다. 이 데이터모델은 데이터범주, 즉 데이터요소의 타입과 그들간의 관계를 구체적으로 명시하는 것이어야 하며 MARTIF는 그러한 데이터모델로서 MARTIF의 부록 A에 제시하고 있는 ISO/DIS 12620을 사용한다. ISO/DIS 12620은 MARTIF 환경에서 사용되는 용어학적 데이터항목의 마크업을 상세하게 나타내고 있으며, 태그의 속성값으로 여기에 규정된 데이터범주만을 사용하도록 규정하고 있다.

MARTIF의 용어엔트리 본문(body)은 〈그림 2〉에서와 같이 〈termEntry〉 태그로 시작해서 〈term〉 태그를 갖는 하나 혹은 여러 개의 용어를 포함한다. 하나의 전문용어 정보그룹은 하나의 용어와 관련된 데이터범주(예를 들어, 〈termNote〉, 〈descrip〉, 〈admin〉 등)로 구성된다. 같은 레벨의 이러한 요소들은 하나의 전문용어 정보그룹으로서 하나의 〈tig〉 요소에 포함된다. 만약 부가적인 하위요소가 포함되어야 한다면, 요소 〈ntig〉가 그룹요소인 〈termGrp〉,

<descripGrp>, 혹은 <adminGrp>과 함께 사용된다. <termNoteGrp> 요소는 <termGrp>내에 두 번째 정보레벨을 포함할 수 있을 때 사용된다. <termEntry>는 단일 <tig>나 <ntig>, 혹은

여러 개의 <tig>나 <ntig>를 혼합하여 구성할 수 있다. 주어진 모든 언어에서 여러 개의 <tig>나 <ntig>는 <langSet>에서 그룹화 된다(ISO/DIS 12200, 제5.2절).



〈그림 2〉 MARTIF의 기본구조
출처: <<http://www.ttt.org/cisframe/negotiated.html>>

<!ENTITY %AuxInfo	'descrip descripGrp admin adminGrp ptr ref note'
<!ELEMENT body	— (termEntry+)
<!-- termEntry = one concept entry: ideally, one concept, one entry -->	
<!ELEMENT termEntry	—((%AuxInfo:)*, (langSet tig ntig)+)
<!ATTLIST termEntry	%a.global;
	type CDATA #IMPLIED
<!-- langSet = cluster of terms in one language plus associated info -->	
<!ELEMENT langSet	—((%AuxInfo:)*, (tig ntig)+).
<!ATTLIST langSet	%a.global;
	type CDATA #IMPLIED
<!-- tig = 'terminological information group, flat' -->	
<!ELEMENT tig	—(term, (termNote)*,
	(descrip admin ptr ref note)*)
<!ATTLIST tig	id ID #IMPLIED
	lang CDATA #IMPLIED
<!-- ntig = 'terminological information group, nested' -->	
<!ELEMENT ntig	—(termGrp, (%AuxInfo:)*
<!ATTLIST ntig	id ID #IMPLIED
	lang CDATA #IMPLIED
<!ELEMENT term	—(%bText)
<!ATTLIST term	%a.global
<!ELEMENT termGrp	—(term, (termNote termNoteGrp ptr ref date note)*
<!ATTLIST termGrp	%a.global;
<!ELEMENT termNoteGrp	—(termNote, (ptr ref date note)*
<!ATTLIST termNoteGrp	%a.global;
<!ELEMENT descripGrp	—(descrip,
	ptr ref date descripNote note)*)
<!ATTLIST descripGrp	%a.global;
<!ELEMENT adminGrp	—(admin,
	(ptr ref date adminNote note)*)
<!ATTLIST adminGrp	%a.global;
<!ELEMENT descrip	—(%dText:)
<!ATTLIST descrip	%a.global;
	type CDATA #REQUIRED
<!ELEMENT admin	—(%dText:)
<!ATTLIST admin	%a.global;
	type CDATA #REQUIRED
<!ELEMENT termNote	—(%bText:)
<!ATTLIST termNote	%a.global;
	type CDATA #REQUIRED
<!ELEMENT descripNote	—(%nText:)
<!ATTLIST descripNote	%a.global;
	type CDATA #REQUIRED
<!ELEMENT adminNote	—(%nText:)
<!ATTLIST adminNote	%a.global;
	type CDATA #REQUIRED

〈그림 3〉 MARTIF 본문의 DTD

완전한 MARTIF는 MARTIF 유형의 자료 예시가 있는 프롤로그, 텍스트를 갖고 오는 헤더(<martifHeader>), 그 뒤에 순서대로 front matter(선택적), body(일련의 전문용어 엔트리), back matter(선택적)으로 구성된다. 그 중에서 MARTIF body 만의 DTD를 보면 <그림 3>과 같다.

현재 각국에서 진행중인 상당수의 용어데이터 관련 프로젝트에서는 MARTIF 구조를 기본으로 하고 ISO/DIS 12620의 데이터항목을 확장의 근거로 삼아 교환용 형식을 개발하고 있다. 선행연구에서 언급한 CLS Framework는 MARTIF를 활용한 대표적인 예로서, MARTIF의 체계를 기반으로 하되 용어교환 파트너간의 비협상교환(용어데이터의 교환시 데이터의 내용, 형식, 구조 등에 대한 도큐멘테이션을 제공할 필요가 없는 것)²⁾이 가능하도록 설계한 것이다.

2.2 인쇄사전용 TEI

TEI는 다양한 유형의 전자문헌에 대하여 기계와 프로그램에 독립적인 표준형식을 규정하여 전자자료나 전자문서를 용이하게 구축하고 교환하게 함으로써 상이한 인코딩 방식에 따른 낭비를 제거하려는 목적으로 구성된 국제적인 프로젝트이다(정광 외 1995). 이 프로젝트는 여러 단체가 공동으로 1987년에 시작하였으며, 다양한 학문분야에서 활동하는 전세계 수백 명의 학자들이 6년간의 작업결과물로 1994년 TEI 지침(*Guidelines for Electronic Text Encoding and Interchange*)을 발표하였고, 1999년에는 개정판을 간행하였다.

TEI는 텍스트유형을 산문, 시, 연설문, 드라마, 사진, 용어집 등 여섯 가지로 나누고, 400여 개의 요소를 사용하여 그 각각의 특성을 인코딩할 수 있는 표준형식을 제안하였다. 원래는 인코딩을 위해 SGML을 적용하였으나 현재는 XML도 함께 제공하고 있으며, 문자집합은 *ISO/IEC 646:1991 Information Technology -- ISO 7-Bit Coded Character Set for Information Interchange*를 참조하고 있다.

그러나 TEI가 제안한 여섯 가지 문헌의 유형별 부호화형식 중 용어데이터의 구조화와 관련해서는 인쇄사전용과 용어데이터베이스용의 두 가지 형식을 별도로 고려하고 있다. 두 형식 모두 용어데이터와 관련된 형식이지만 개발단계에서 이 두 가지를 통합하려는 노력이 수포로 돌아간 것은 용어데이터를 인쇄물과 같은 구조를 가지는 텍스트로 보느냐 아니면 데이터베이스로 보느냐에 따른 것이다(Ide and Veronis 1995). 용어데이터베이스용 TEI는 ISO/DIS 12620을 사용하므로 MARTIF와 원칙적으로 같다.

TEI 문헌은 TEI 헤더부분과 DTD에 따라 코드화된 텍스트본문으로 구성된다. 헤더는 전자텍스트의 서지기술, 부호화방법, 비서지적

2) MARTIF가 협상 교환(negotiated exchange)용 형식으로서 가지는 한계, 즉 교환 파트너들이 교환 전에 서로의 데이터를 검토하고 교환형식으로부터 그것을 받아들이기 전에 데이터에 대한 전 처리 여부를 결정해야 하는 한계를 말하며, 이를 보완하고자 하는 여러 가지 프로젝트가 진행중이다. 비협상 교환에 관한 자세한 내용은 Alder(1998), Schmitz와 Wright(1997), Melby와 Hardman(1996) 및 TTT(Translation, Theory & Technology)의 홈페이지(<<http://www.ttt.org/cslsframe/BTFdatcats.html>>)에서 찾아볼 수 있다.

기술, 개정이력을 포함하며, 이 중에서 전자문서에 관한 기본적인 정보를 제공해 주는 부분(file description) 만이 필수적인 것이다. 이 헤더부분은 텍스트와 독립적으로 서지정보를 유통하는 데 사용될 수 있다. 문헌 자체는 텍스트 본문이며 단일 또는 복합형태로 구성된다. TEI 표준을 따르는 문서(TEI conformant document)는 SGML의 모든 세부사항을 문서에 모두 명시하는 문서가 아니라, 그것과 모순되지 않으면서 그 중 일부를 선택적으로 표시하는 문서를 말한다. 단, 모든 TEI 문서는 TEI에서 규정하는 header(<teiHeader>)와 본 텍스트(<text>)로 구성되어야 한다는 제약조건이 있다. 이상과 같은 조건을 만족시키는 가장 기본적인 TEI 문서는 다음의 <그림 4>와 같이 구성된다.

인쇄사전용 TEI에서는 다양한 사전표현형식을 조사한 후, 이를 SGML의 파라미터 엔티티를 이용하여 일관되게 표현하는 방법을 제시하고 있다. 이처럼 상세하게 마련된 TEI의 항목 및 요소정의는 사전항목의 세부적인 구조표현을 가능하게 하며, 여러 가지 종류의 유용요소가 장소에 관계 없이 부동요소를 보충하는 역할을 한다. 또한 별도 태그의 추가, 사용태그명의 변경, 새로운 태그정의기능 등과 같은 확

```

<TeiHeader>
  <FileDesc>
    ...
  </FileDesc>
</TeiHeader>
<Text>
  ...
</Text>
    
```

<그림 4> TEI의 기본구조

장기법을 제공하고 있다(김성혁 1996).

기본적인 사전텍스트의 본문은 <그림 5>와 같이 여러 개의 <entry> 요소로 구성될 수 있고, 각 <entry>는 여러 개의 동형어의어 <hom> 요소로 구성될 수 있다. 사전요소는 <entry>, <hom>, 혹은 의미를 나타내는 <sense> 레벨에 나타날 수 있으며, <superEntry>는 여러 개의 <entry>를 하나로 묶는 역할을 한다(TEI 1999, 제12장).

TEI는 전자문서의 인코딩형식으로서 널리 사용되고 있다. ISO의 용어교환형식인 TIF(ISO/DIS 12200 MARTIF의 전신)는 TEI의 용어데이터베이스 DTD를 기반으로 한 것이다(김성혁 1996). 제3세대 코퍼스의 대표적인 것으로 알려져 있는 영국의 British National

```

<text>
  <body>
    <entry>...</entry>
    <entry>
      ...
      <sense>...</sense>
      <sense>...</sense>
    </entry>
    <superEntry>
      <entry>...</entry>
      <entry>...</entry>
      ...
    </superEntry>
    <entry>...</entry>
    <entry>
      ...
      <hom>...</hom>
      <hom>...</hom>
    </entry>
    ...
  </body>
</text>
    
```

<그림 5> 인쇄사전용 TEI의 본문의 구조

Corpus(BNC)와 미국전산언어학회의 ACL/DCI(Association for Computational Linguistics/Data Collection Initiative) 코퍼스도 약 1억 어절 이상을 구축하고 있으며 TEI 체계를 표준으로 삼고 있다(정광 외 1995) (TEI를 적용하고 있는 다양한 프로젝트와 상세한 내용에 대해서는 <<http://www.tei-c.org/Applications/index.html>> 참조). 또한 Oxford Text Archive는 원래 사용하던 마크업 방식을 SGML로 전환하는 과정에서 TEI Lite(TEI의 수많은 SGML 요소 및 권고사항들 중에서 일부를 추려 기술한 것)를 사용하고 있으며, Virginia 대학교와 Michigan 대학의 전자텍스트센터에서도 소장하고 있는 자료들을 TEI Lite를 이용하여 부호화하고 있다.

국내의 경우에는 연세한국어말뭉치 중 세종말뭉치에 TEI 헤더가 이용되었고, 한국과학기술원의 국어정보베이스말뭉치에 TEI 헤더를 붙여 세종말뭉치와 통합하였다. 고려대학교 한국어말모듬 역시 TEI 호환방식으로 태깅되었으며, 21세기 세종계획의 국어기초자료말뭉치에도 TEI 헤더와 마크업이 첨부되어 있는 등 국내 용어 관련 주요 프로젝트의 대부분이 TEI 체계를 적용하고 있다(문화관광부 2000b, 32-54).

2.3 적용형식

전술한 바와 같이 MARTIF는 용어교환을 위한 큰 틀만을 제시하고 실제로 용어데이터의 상세한 내용을 나타내기 위한 구체적인 데이터 모델은 MARTIF의 부록 A에 제시되어 있는 ISO/DIS 12620을 적용한다. 또한 용어데이터

베이스용 TEI도 MARTIF와 마찬가지로 ISO/DIS 12620을 적용하고 있으므로(TEI 1990, 제13장), 이 두 가지 형식은 근본적으로 거의 같은 것으로도 간주할 수 있다. 반면, 인쇄사전용 TEI는 용어데이터베이스용 TEI와는 달리 독립적으로 사전의 구조, 요소, 기술, 태그를 상세하게 정의하고 있다. 이는 또한 ISO/DIS 12620과 모든 면에서 확연히 다르다.

따라서 보다 망라적인 표준구조를 도출해내기 위해 ISO/DIS 12620와 인쇄사전용 TEI를 통합하거나 상호적용하는 것은 현실적으로 불가능하며, 무엇보다도 이 둘은 통합을 전제로 하지도 않았다. 결국 한국어사전의 미시구조를 표현하는 방법에는 상이한 두 가지의 방법을 생각할 수 있다. 즉, MARTIF를 적용하는 방법과 인쇄사전용 TEI를 적용하는 방법을 검토할 필요가 있다.

3 미시구조와 인코딩체계의 적용

3.1 MARTIF의 적용

MARTIF를 사전데이터의 인코딩을 위해 적용할 때 MARTIF가 국제표준으로서 가지는 통일된 규칙성이 큰 장점이 될 수 있다. 또한 MARTIF 자체는 단순한 구조이나 200여 개에 달하는 ISO/DIS 12620의 데이터항목을 적용하여 상세한 용어정보를 표현할 수 있도록 마련되어 있다. ISO/DIS 12620의 목적은 존재 가능한 모든 데이터베이스관리프로그램에서 다룰 수 있는 용어범주의 전집합을 제공하는 데 있다. 즉, 용어데이터의 입출력시, 현재 및

〈표 1〉 ISO/DIS 12620의 구조와 확장

		ISO/DIS 12620		확장
용어 관련 데이터	용어 유형 (term type)	표제어(main entry term)		한자(〈term type = 'hanja'〉)
		동의어(synonym)		
		유사동의어(quasi-synonym)		
		국제 학술 용어 (international scientific term)		
		완전형(full form)		
		약어형 (abbreviated form of term)	축약형(abbreviated)	잘못된 표기 (〈term type = 'wrong'〉)
			간략형 (short form of term)	
			두문자어(initialism)	
			두문자어(acronym)	
			생략형(clipped term)	풀이내의 다른 뜻 (〈termtype = 'polyseme'〉)
		이형(variant)		
		전자형(transliterated form)		
		전사형(transcribed form)		장음 (〈termNote type = 'longproun'〉)
		로마자형(romanized form)		
		기호(symbol)		
	공식(formula)			
	방정식(equation)			
	논리적 표현(logical expression)			
	자료 관리 범주 (materials management categories)	물품 표시 단위 (stock keeping unit)	단음 (〈termNote type = 'shortproun'〉)	
		부품 번호 (part number)		
구문 단위(phraseological unit)		연어(collocation)		
	구문(set phrase)			
	유의구문(synonymous)			
표준 문서(standard text)				
문법 (grammar)	품사(part of speech)	불규칙		
	문법적 성(grammatical gender)	(〈termNote type = 'irreg'〉)		
	문법적 수(grammatical number)			
	생물·무생물(animacy)	굴절형		
	명사류(noun class)	(〈termNote type = 'irrFrom'〉)		
	형용사류(adjective class)			
용법 (usage)	사용 주기(usage note)	신규어		
	지역적 사용(geographical usage)	(〈termNote type = 'newterm'〉)		
	등록(register)			

〈표 1〉 ISO/DIS 12620의 구조와 확장(계속)

		ISO/DIS 12620	확장
유 어 관 련 데 이 터	용법 (usage)	빈도(frequency)	기타 용법 (〈termNote type=domEtc〉) 용어 수준 (〈term-Note type = termlevel〉)
		일시적 한정어(temporal qualifier)	
		시간적 제약(time restriction)	
		사용 권한 제약(proprietary restriction)	
	용어 형성 (term formation)	용어 어원(term provenance)	
		어원(etymology)	
	발음(pronunciation)		
	분절법(syllabification)		
	하이프네이션 (hyphenation)		
	어형(morphology)	형태소(morphological element)	
		용어 요소(term element)	
	용어 상태 (term status)	표준어(normative authorization)	
		잠정용어(language-planning qualifier)	
관리 상태(administrative status)			
표준화 상태(process status)			
유의 정도 (degree of synonymy)			
동 등 관 계 데 이 터	등가 정도(degree of equivalence)		
	어원과 형태가 같으나 다른 뜻(false friend)		
	방향성 (directionality)		
	정확성을 나타내는 코드 (reliability code)		
	등가성에 관한 설명 (transfer comment)		
주 제 분 야	주제 영역(subject field)		
	분류 체계 (classification system)		
	분류 번호 (classification number)		
개 념 관 련 기 술	정의(definition)		풀이내의 구분 설명 (〈descrip type= subdefinition〉)
	설명(explanation)		
	문맥(context)		
	예(example)		

〈표 1〉 ISO/DIS 12620의 구조와 확장(계속)

		ISO/DIS 12620		확장
개념 관련 기술	비 문서적 요소 (nontextual illustrations)	그림 (figure)		부류어(〈descrip type= 'groupword'〉) 특수범주 (〈termNote type = 'specialdomain'〉) 풀이내의 다른 뜻 (〈term type = 'polyseme'〉)
		음향(audio)		
		비디오(video)		
		표(table)		
		기타(other binary data)		
	단위(unit)			
범위(range)				
특성(characteristic)				
개념 관계	속 관계(generic relation)			
	부분 관계 (partitive relation)			
	연속 관계 (sequential relation)	일시적 관계(temporal relation)		
		특수 관계(special relation)		
	연관 관계 (associative relation)			
개념 구조	개념 체계 (concept system)			
	개념 위치 (concept position)	상위 속/부분 관계(broader concept generic/partitive)		
		최상위 속/부분 관계(superordinate concept generic/partitive)		
		중속 속/부분 관계(subordinate concept generic/partitive)		
		등의 속/부분 관계(coordinate concept generic/partitive)		
관련 개념(related concept)				
주기				
문서 언어	시소러스명 (thesaurus name)			
	시소러스 디스크립터 (thesaurus descriptor)			
	비디스크립터 (nondescriptor)			
	키워드(keyword)			
	색인 표목(index heading)			
관 련 정 보	용어 관리 처리 기록 (terminology management transactions)			
	용어 관리 기능 (terminology management functions)	날짜 (date)	생성일(origination date)	
			입력일(input date)	
개정일(modification date)				

〈표 1〉 ISO/DIS 12620의 구조와 확장(계속)

		ISO/DIS 12620	확장
관 련 정 보	용어 관리 기능 (terminology management functions)	날짜 (date)	점검일(check date)
			승인일(approval date)
			폐기일 (withdrawal date)
			표준화된 날짜 (standardization date)
			데이터베이스에서의 출력일(exportation date)
			데이터베이스로의 입력일(importation date)
		책임 (responsibility)	생성자(originator)
			입력자(inputter)
			갱신자(updater)
			점검자(checker)
			승인자(approver)
			사용자(User)
	서브셋 식별자 (subset identifier)	고객 서브셋(customer subset)	
		최초 고객 서브셋(initial customer subset)	
		프로젝트 서브셋(project subset)	
		초기 프로젝트 서브셋(initial project subset)	
		결과물 서브셋(product subset)	
		어플리케이션 서브셋(application subset)	
환경 서브셋(environment subset)			
경영 단위 서브셋(business unit subset)			
보안 서브셋(security subset)			
권한(authorization)			
사용자 제안 (user suggestion)			
관 리 정 보	관리 용어 한정어 (administrative term qualifiers)	타용어에서 정의된 용어(entailed term)	
		정렬 키(sort key)	
		탐색 용어(search term)	
언어 코드 (language symbol)			

〈표 1〉 ISO/DIS 12620의 구조와 확장(계속)

		ISO/DIS 12620	확장	
관	외국문서(foreign text)			
	연어 순서 (collocating sequence)			
	항목유형(entry type)			
	요소 정확성 수준 코드 (element working status)			
	목표 데이터베이스 (target database)			
	항목소스(entry source)			
	개념 식별기호 (concept identifier)			
	항목 식별기호 (entry identifier)			
	리	레코드 식별기호 (record identifier)		
		파일 식별기호 (file identifier)		
보	상호참조 (cross-reference)	보라(see)		
		도보라(see also)		
		도치 용어(inverted term)		
		순열 용어(permuted term)		
		동형이의어(homograph)		
		반의어(antonym)		
소스 식별기호 (source identifier)				
이름 공간 식별기호 (namespace identifier)	URL(Uniform Resource Locator)			
	FPI(Formal Public Identifier)			
소스 작성자 (originating entity)	소스 작성자(originating person)			
	소스 작성 기관(originating institution)			
	소스 데이터베이스(originating database)			

미래의 용어데이터베이스 관리자들이 공통으로 참조하여 일관성 있는 데이터를 유지할 수 있다는 데 그 의미가 있다. 따라서 사전데이터에 대한 MARTIF의 적용을 논의하기 위해서

는 ISO/DIS 12620의 내용과 구조를 분석할 필요가 있다.

ISO/DIS 12620을 상세하게 분석하여 구조화하면 〈표 1〉과 같다. ISO/DIS 12620은 10

개의 데이터범주로 크게 구분되며, 이는 다시 용어, 용어관련정보, 설명정보, 그리고 관리정보 등 네 개의 클래스로 그룹화 될 수 있다. 용어(term)는 데이터항목의 첫 번째 클래스이며 오직 데이터항목 용어만이 데이터요소가 된다. 용어관련정보(term-related information)는 용어의 특정한 본질이나 특성을 기술한다. 설명정보(descriptive information)는 정의, 용례, 설명, 그리고 개념쌍 간의 관계 등을 기술한다. 그리고 관리정보(administrative information)는 데이터나 문서의 내용을 구체화하기 위해 사용되는 문서언어와 같은 것을 기술하여 저장과 검색에 사용할 수 있도록 한다. 관리정보에는 시소러스와 관련된 데이터항목을 포함시킬 수 있으며, 생성일, 갱신일, 생성자, 갱신자 등에 대한 데이터항목을 포함시킬 수도 있다.

표에서 가장 오른쪽 열은 ISO/DIS 12620을 적용한 MARTIF 구조로도 표현할 수 없는 한국어사전항목을 표현하기 위해 본 연구에서 제안하고 있는 확장안이다. 확장안에서는 MARTIF 및 ISO/DIS 12620에서 제시하고 있는 태그 및 속성의 양식을 그대로 따르되, 표현하고자 하는 내용이 가장 잘 드러날 수 있는 방식으로 새로운 요소 및 속성명을 추가하였다.

3.2 인쇄사전용 TEI의 적용

인쇄사전용 TEI는 사전데이터의 항목을 상세하게 나타낼 수 있는 구조로 설계되었다. TEI가 인쇄사전용과 용어데이터베이스용으로 나뉘어져 있는 것에서 알 수 있듯이, 인쇄사전용 TEI는 일반 용어데이터베이스에서는 제대로 나타내지 못하는 사전의 구조적 특성과 내

용을 최대한 반영하고자 하는 목적으로 설계되어 널리 사용되고 있으므로 사전데이터의 인코딩에 적합하다. 또한 인쇄사전용 TEI에서는 제공하지 않는 마크업요소일지라도 모든 문서 유형에 적용할 수 있도록 지정되어 있는 기본 태그 및 부가태그를 사용하면 인쇄사전용 TEI에서 제공하는 것 이상의 인코딩도 가능하게 된다.

인쇄사전용 TEI를 상세하게 분석하여 구조화하면 <표 2>와 같다. 사전데이터 관련정보를 크게 여섯 가지 범주로 나누고 그 안에서 다시 세분하고 있다. 이것은 실제 사전의 구조와 유사한 순서와 형식으로 구성되어 있으므로 사전의 다양한 항목들을 비교적 손쉽게 인코딩할 수 있다. 표의 가장 오른쪽 열은 TEI만으로는 나타내지 못하는 한국어사전항목을 표현하기 위해 본 연구에서 제안하고 있는 확장안이다. 확장안에서는 TEI에서 사용하고 있는 요소와 속성의 방식을 그대로 따랐으며, MARTIF의 확장에서와 마찬가지로, 표현하려는 내용을 가장 잘 드러낼 수 있는 방식으로 새로운 요소 및 속성명을 추가하였다.

4 한국어 사전데이터를 위한 인코딩체계 확장

MARATIF와 인쇄사전용 TEI 모두 용어데이터의 인코딩형식으로 적용할 수 있다. 그러나 두 체계 모두 망라적인 한국어사전데이터의 표현을 위해서는 어느 정도의 확장을 필요로 한다. 따라서 본 연구에서는 MARTIF의 경우에는 용어데이터의 전체적인 데이터모델인 ISO/DIS 12620(표 1 참조)과 한국어사전의

〈표 2〉 사전용 TEI의 구조와 확장

	TEI			확장
구어 및 문어 형식 (Information on Written and Spoken Forms)	표제어형식 (〈form〉)	자유로운 형태(simple)		한자(〈form type = hanja〉) 로마자 (〈form type = romanized〉) 잘못된 표기(〈form type = wrong〉)
		표제어 자체(lemma)		
		변형(variant)		
		어휘 항목으로 구성된 단어(compound)		
		표제어에서 파생된 단어(derivative)		
		일반적인 사전 형태와 다른 단어(inflected)		
	복수의 단어로 된 어휘 항목(phrase)			
	표기형식 (〈orth〉)	철자의 범위 (extent)	완전한 형태(full)	기호, 공식, 논리적 표현(〈form type = symbol〉), 〈form type = formula〉(〈form type = logical expression〉)
			접두사(pref)	
			접미사(suff)	
			부분(part)	
	발음 (〈pron〉)	발음의 범위 (extent)	완전한 형태(full)	장음(〈lenHyph〉)
			접두사(pref)	
			접미사(suff)	
			부분(part)	
하이픈 정보(〈hyph〉)				
분절 표시(〈syll〉)				
강세 유형(〈stress〉)				
기타 정보 표시(〈lbl〉)				
문법 정보 (Grammatical Information)	〈gramGrp〉	품사(〈pos〉)	동사의 불규칙(〈irreg〉) 동사의 굴절형(〈irrForm〉)	
		하위 범주화 정보(〈subc〉)		
		언어정보(〈colloc〉)		
	문법 정보(〈gram〉)			
	형태적 성별(〈gen〉)			
	굴절류(〈itype〉)			약어(abbrev)
				동사표 참조(verb table)
	문법적 수(〈number〉)			
	격(〈case〉)			
	인칭지시(〈per〉)			
시제(〈tns〉)				
법(〈mood〉)				
어의 정보 (Sense Information)	정의(def)	다국어 사전에서 정의와 대응한 대역어(〈trans〉, 〈tr〉)		
	대역어	번역 텍스트와 관련 정보(〈trans〉)		
		표제어의 역어나 사례(〈tr〉)		
		기타 정보 표시(〈lbl〉)		
어원 정보 (Etymological Information)	어원(〈etym〉)		한자 어원 (〈etym hyType = hj〉)	
	어원상 혹은 언어학적 언어명(〈lang〉)			
	어떤 형태로 된 날짜(〈date〉)			

(표 2) 사전용 TEI의 구조와 확장(계속)

	TEI		확장
어원 정보 (Etymological Information)	언급되었으나 사용하지 않은 단어나 구(<mentioned>)		외국어 어원 (<etym hyType = foreign>)
	다른 단어나 구에 대한 정의나 주석을 제공하는 단어나 구(<gloss>)		
	단어의 발음(<pron>)		한글 어원 (<etym hyType = kor>)
	사전 항목에서의 이용 정보(<usg>)		
기타 정보 표지(<lbl>)			
기타 정보 (Other Information)	용례	용례와 관련정보(<eg>)	
		인용문(<q>)	
		텍스트 외부의 특정인의 것(<quote>)	
		출처(<cit>)	
	용법 (<usg>)	지역(geo)	용어의 영역(학문, 기타, 방언)
		시간대(time)	(<domAc>, <domEtc>, <dialArea>)
		영역(dom)	분류사(<usg type = unit>)
		등록(reg)	부류어(<usg type = groupword>)
		양식(style)	용어 빈도(<usg type = frequency>)
		우선수준(plev)	용어 수준(<usg type = termlevel>)
		수용가능성(acc)	용어 상태(<usg type = normative> 등)
		외래어의 언어, 철자, 발음 기호 등(lang)	특수범주(<usg type = specialdomain>)
	다국어 사전에서의 어의적단서 (<usg>)	문법적 용례(gram)	신규어(<usg type = newterm>)
		용례로 제시된 동의어(syn)	
		용례로 제시된 상위어(hyper)	
		용례로 제시된 연어정보(colloc)	
		일반적인 보어(comp)	
		일반적인 목적어(obj)	
		일반 주어(subj)	
		일반 동사(verb)	
다른 표제어에 대한 참조	어의선택에 도움이 되는 분류할 수 없는 정보(hint)		
	텍스트 내 · 외의 특정 위치로 지시(<xr>)	관련어 : 보라(<xrsee>), 표준어(<xrstd>), 비표준어(<xrstd>), 반의어(<xrant>), 거센말(<xrstr>), 여린말(<xrsoft>), 큰말(<xrlarge>), 작은말(<xrsmall>), 높임말(<xrhon>), 친근체(<xrint>), 동의어(<xrsame>), 동음이의어(<xrsyn>), 참조어(<xrcfwd>), 줄임말(<xrshort>), 본디말(<xrlong>), 변한말(<xrchg>), 이형(<xvar>), 용어관계(<BT>, <NT>, <RT>, <IT>, 등)	
	현재 문서 내의 다른 위치 지시하는 참조 정의(<ref>)		
	현재 문서 내의 다른 위치를 지시하는 포인터정의(<ptr>)		
	확장된 포인터 기호로 다른 위치 지시(<xref>)		
	다른 위치 지시하는 포인터 정의(<xptr>)		
기타 정보에 대한 표지(<lbl>)			
주(<note>)	관용법, 문법, 내용 등에 대한 주나 주석		
관련 항목 (Related Entries)	관련항목 (<re>)	<entry>가 가지는 모든 요소를 가질 수 있음 (내포된 <xr> 제외)	
			학명(<sciName>), 속담(<prov>), 속어(<idiom>)

미시구조를 나타낸 <그림 1>의 비교를 통해, TEI의 경우에는 인쇄사전용 TEI(표 2 참조)와 역시 <그림 1>의 비교를 통해 확장안을 제안하고자 한다. 다만, 각각의 확장내용은 동일한 항목 내에서 함께 제안하며, 확장안에는 Kang (1998)과 김성혁(1996) 등의 연구에서 제안된 내용을 포함시킨다. 또한 인코딩유형과 요소의 추가부분은 결국 사용자간의 합의 또는 표준기관의 심의를 거쳐야 하는 것이므로, 본 연구에서는 일단 각 구조에 맞는 속성명 등을 만들어 확장부분이 의미하는 바가 드러날 수 있도록 하는 데 주안점을 두고자 한다. 구체적인 확장내용은 다음과 같다.

1) 발음에 대한 음의 길이를 나타낼 수 있는 요소가 필요하다. ISO/DIS 12620은 '발음' 항목을 가지고 있으나 그 안에서 구체적인 정보를 표시할 수 있는 방법이 제시되어 있지 않으므로 MARTIF에서는 상세한 내용을 나타낼 수 없다. TEI에서도 발음의 속성을 범위측면에서만 구분하여 두었기 때문에 용어의 장단음을 나타낼 수 없다. 따라서 MARTIF에서는 <termNote type= 'pronunciation'>을 기본으로 하면서 장음을 위해서는 <termNote type= 'longpronun'>, 단음을 위해서는 <termNote type= 'shortpronun'>을 사용하고, TEI에서의 음의 길이표시를 위해서는 <lenHyph>의 사용을 제안한다.

2) MARTIF와 TEI 모두 한자 및 로마자표기를 할 수 있어야 한다. 두 표준 모두 용어유형에서 표제어에 대한 각종 이형 및 표기법을 나타낼 수 있도록 하였지만 영어권에서 마련된 표준은 한자문화권의 사전데이터를 나타내는

데 있어서 한계를 지닌다. 이에 대하여 MARTIF의 형식에서는 <term type= 'hanja'>를, TEI에서는 <form type= hanja>를 사용할 것을 제안한다. 한편 로마자표기에 대하여 MARTIF는 '용어유형'에서 로마자형을 나타낼 수 있도록 하였으나 TEI에는 이러한 구분이 없다. 따라서 TEI에서는 로마자 표기를 위하여 <form type= romanized>를 사용할 것을 제안한다.

3) 잘못된 표기를 나타낼 수 있는 방법이 필요하다. 사전은 언어에 대한 규범성을 가져야 하므로 맞춤법이나 외래어표기원칙에 어긋나는 경우, 혹은 잘못 알려진 형태가 굳어져서 일반적으로 쓰이고 있는 경우에 있어서 이를 단순히 표제어형식의 변형이나 다른 형태로 다룰 것이 아니라 잘못된 표기임을 구분하여 나타낼 수 있어야 한다. 이를 위해 오폭기를 나타내는 'wrong'이라는 속성을 type 속성의 하나로 정의하고, MARTIF의 경우 <term type= 'wrong'>으로, TEI의 경우 <form>의 요소를 이용하여 <그림 6>과 같이 나타낼 것을 제안한다.

```

<form type = wrong>
  <orth>복계</orth>
  <ref>부계</ref><lbl>의 오폭기</lbl>
</form>
    
```

<그림 6> 오폭기의 예(TEI)

4) 한글용어의 열두 가지 불규칙활용형을 나타낼 수 있는 표기가 필요하다. MARTIF에서는 문법정보를 나타내는 부분에 있어서 동사의 불규칙과 굴절형을 각각 나타내기 위하여 <termNote type= 'irreg'>와 <termNote

type = 'irrForm')을 사용할 것을 제안한다. TEI에서는 <gramGrp>내에서 <irreg>과 <irrForm>을 사용할 것을 제안한다.

5) 전문용어 사전에서는 기호, 수식, 공식 및 화학원소 등이 항목 내에 포함되는 경우가 많다. 이는 사전항목으로서 매우 중요한 정보이므로 별도의 항목으로 인코딩할 수 있어야 한다. MARTIF의 경우 ISO/DIS 12620의 제 2.1.13절~제2.1.16절에 제시된 방식대로 기호, 공식, 방정식, 그리고 논리적 표현을 나타낼 수 있다.

한편, 사전용 TEI에는 수식, 공식 등을 나타낼 수 있는 방법이 구체적으로 정의되어 있지 않다. 따라서 TEI에서는 <form type =

symbol), <form type = formula>, <form type = logic>등과 같이 새로운 유형을 추가하여 나타내거나, MARTIF와 유사한 방식으로 <그림 7>과 같이 나타내는 방안을 제시한다.

6) 풀이 내의 다른 뜻을 나타낼 수 있는 장치가 필요하다. MARTIF의 형식에서는 <term type = 'polyseme' >를 제안하고, TEI에서는 <form type = polyseme>를 사용하거나 혹은 하나의 표제어 내에서 정의를 <sense> 요소의 중첩방법으로 구분할 것을 제안한다.

7) 풀이 내의 구분설명을 표현 할 수 있는 방법이 필요하다. 백과사전이나 특수사전풀이의 경우, 일반적인 용어사전에서 <def>로 나타낼 수 있는 정의보다 훨씬 길고, 풀이 내에 다시 항목으로 구분하여 상세하게 설명하고 있는 경우가 많다. 백과사전을 예로 살펴보면, 하나의 표제어에 대한 설명은 목적, 내용, 범위, 현황, 취약점, 역사, 문화, 풍습, 교통, 경제, 산업, 기후, 인구 등 많은 하위분야로 구분하여 설명된다. 표제어가 인물인 경우 인물의 성격과 배경에 따라 정의항이 이름(본명, 필명, 속명, 아명, 호, 자, 諡號 등 포함), 생몰년, 재위년(왕족의 경우), 왕조의 흥망년(역사적 인물의 경우), 본

```

<form type = wrong>
  <orth>ax2+bx+c=0(a≠0)</orth>
  <descrip type = 'definition'>미지수의 최고
                                차수가 이차인
                                방정식
  </descrip>
  <lbl>공식</lbl>
</form>
    
```

<그림 7> 공식 표기의 예(TEI)

```

<entry>
  <form>
    <orth>김동리</orth>
    <def type = occupation>시인, 소설가</def>          (!- 직업 -)
    <def type = originalname>시종</def>                (!- 본명 -)
    <def type = birthplace>경북 경주</def>             (!- 출생지 -)
    ...
    <def type = prize>예술원상, 3·1 문화상</def>      (!- 수상경력 -)
    <def type = works>《무녀도》(1947), 《역마》(1948) ... </def> (!- 주요작품 -)
  </form>
</entry>
    
```

<그림 8> 풀이내 구분설명의 예(TEI)

관, 출생지, 학력, 전공분야, 관직, 가족, 문하, 친구, 저작, 작품경향 등으로 나누어지는 경우가 흔하다. 이는 서명 및 작품이 표제어인 경우 이를 작자, 저작 연도, 내용, 그리고 평가 등으로 나누어 설명하는 경우에도 마찬가지이다.

이러한 정보를 나타내기 위하여 MARTIF의 경우 <descrip type= 'subdefinition' >과 같은 요소의 추가를 제안한다. 한편, TEI에 있어서는 정의를 나타내는 <def>에 속성을 부여하여 정의항을 구분하는 방법을 제안한다. 혹은 표제어 내에서 정의가 <sense n= '1' >, <sense n= 'a' > 등 <sense> 요소의 중첩을 이용하여 분류될 때 <sense>별로 항목 내의 구분을 나타내도록 하는 방법을 제안한다. <그림 8>은 TEI의 정의를 나타내는 <def>에 속성을 부여하는 방법에 의해 표제어 '김동리'에 대한 정의항을 항목별로 나누어 본 것이다.

8) 용어의 어원을 표시할 수 있는 방법이 필요하다. ISO/DIS 12620에서는 '용어형성'에서 용어의 어원을 다루고 있다. 한편 TEI는 '어원정보'에서 어원과 관련된 다양한 요소들을 인코딩할 수 있는 체계를 제공하고 있다. 그러나 실제 사전에서는 용어의 다양한 어원정보가 표시되어 있으므로 한국어어원과 한자어원, 그리고 외국어어원을 각각 구분하여 인코딩하기 위하여 <etym hyType=hj>, <etym hyType=foreign>, 그리고 <etym hyType=kor>를 제안한다.

9) 용례와 용법에 있어서 속담과 속어를 나타낼 수 있어야 한다. 이들을 일반용례에 포함시키는 것도 가능하나 이들을 일반용례와 구분하여 표시하기 위해 각각 <prov>와 <idiom>이라는 새로운 요소를 제안한다. 한편 MARTIF

에서는 용법에 제시되어 있는 지역적 사용요소를 이용하여 <termNote type= 'geographic-ahusage' >로 방언을 표시하고 기타 용법정의를 위해서는 속성 <domEtc>, 학문영역을 위해서는 <domEtc> 등의 추가를 제안한다. TEI에서도 이와 유사하게 <usg>의 속성으로 방언을 표시하기 위한 <dialArea>, 궁중어, 은어, 고어, 북한어 등을 표시하기 위한 <domEtc>, 그리고 학문영역을 위한 <domAca>의 추가를 제안한다.

10) 표제어에 대해 주어지는 부류어의 인코딩을 고려해볼 수 있다. 부류어는 서양의 사전이나 북한의 사전에서 등장하는 항목으로 하나의 부류로 묶일 수 있는 개념들을 함께 모아 하나의 표제어 아래에서 제공해주는 것이다. 예를 들어 '호박'이라는 표제어 아래에 '늦~, 단~, 불~, 조선~, 청동~, 꽃~, 애기~, 떡~, 씨~' 등의 부류어를 제시하고 있는 것이 보통이다. 이는 참조지시에 의해 연결되는 유사어나 동의어와는 또 다른 정보이다. 이러한 부류어에 대한 인코딩은 MARTIF의 경우 '개념관련기술'에서 <descrip type= 'groupword' > 등과 같은 방식으로 추가할 것을 제안한다. TEI의 경우에 있어서는 다른 항목에 대한 참조에서 같은 부류에 속하는 표제어로 연결시키거나 따로 부류어를 위한 적합한 요소명(예를 들어, <usg type=groupword>)을 부여하여 처리할 것을 제안한다.

11) 용어의 상태를 나타내는 인코딩방법이 필요하다. 용어에는 우선어, 승인어, 반대어, 폐기어, 대체어 등의 상태가 있을 수 있다. 북한의 사전에서 역시 옹근뜻과 반뜻이라고 하여 자립적인 뜻과 자립적인 뜻의 포괄범위 안에 존재하는 불완전한 뜻을 같은 번호 아래서 뜻

```

<ntig lang=en>
  <termGrp>
    <term>tensile strength</term>
    <termNote type= 'normativeAuthorization' >preferred term</termNote>
  </termGrp>
</ntig>
<ntig lang=en>
  <termGrp>
    <term>tensile</term>
    <termNote type= 'normativeAuthorization' >Deprecated term</termNote>
  </termGrp>
</ntig>

```

〈그림 9〉 용어상태의 예(ISO/DIS 12620)

풀이 하고 있는 경우가 있다. ISO/DIS 12620의 제2.9절 '용어상태'에서는 용어의 상태를 아홉 가지로 상세하게 구분하고 있으며 〈그림 9〉와 같은 인코딩의 예를 제시하고 있다.

한편, TEI에서 이러한 용어의 등급에 관한 정보는 기타 정보의 용법에서 〈usg type=normative〉 등과 같이 용법 내의 각 속성으로 제시할 것을 제안한다.

12) 용어의 빈도와 수준을 나타낼 수 있어야 한다. 한국어사전에는 빈도가 제시된 사전이 아직 많지 않으나 외국사전이나 북한사전에서는 용어의 빈도정보가 중요한 항목으로서 나타나며 이는 언어사전을 이용하는 외국인 학습자에게 특히 도움이 되는 항목이므로 따로 인코딩할 수 있는 체계가 마련되어야 한다. 한편 용어의 수준 또한 빈도와 마찬가지로 학습자사전에서 중요한 정보로 제공되므로 이에 대한 인코딩항목도 필요하다. MARTIF의 경우 용법에 있어서 빈도를 나타내기 위해 ISO/DIS 12620 제2.3.4절 '빈도'에서 제공하는 바와 같이 〈termNote type= 'frequency'〉와 같은 형식을 사용할 수 있다. 한편 TEI에서는 〈usg

type=frequency〉를 정의할 것을 제안한다. 한편 용어의 수준도 빈도와 같은 형식으로 나타낼 수 있다. 즉, MARTIF로는 〈termNote type= 'termlevel'〉로, TEI로는 〈usg type=termlevel〉과 같은 형식으로 용어의 수준을 표현할 것을 제안한다.

13) 가산/비가산명사의 구분에 있어서 분류사를 인코딩해야 한다. 가산명사의 경우에는 개체단위, 분할단위 및 집합단위가 지정되고, 비가산명사의 경우에는 분량단위가 지정되는 분류사가 표제어인 명사와 밀접하게 결합하여 사용된다. MARTIF에서는 ISO/DIS 12620의 '개념관련기술' 제5.6절 '단위'의 개념을 적용할 수 있다. 즉, 〈descrip type= 'unit'〉과 같이 표현할 수 있다. 반면 TEI에서는 분류사를 용법의 속성으로 추가하여 〈usg type=unit〉로 나타낼 것을 제안한다.

14) 사전의 성격에 따라 특수한 범주로 구분되어 있는 항목을 구분표시할 수 있는 방법이 마련되어야 한다. 예를 들어, 『정보통신 용어사전』(한국정보통신기술협회 2000b)의 경우, 정보소양인증제 대비용어나 표준화된 용어는 따

로 구분표시를 하고 있다. 이는 『수학대사전』(박을용 1993)에서 따로 컴퓨터 용어에 대한 표시를 하는 경우와 마찬가지로이다. 이를 위해 MARTIF의 용법에서는 <termNote type='specialdomain'>의 추가를, TEI에서는 기타 정보의 용법에서 특수영역을 나타낼 수 있도록 <usg type=specialdomain> 요소의 추가를 제안한다.

15) 다른 표제어에 대한 참조에 있어서 한국어사전은 서양언어와는 다른 독특한 요소들을 가지고 있다. 보라참조를 비롯한 표준어, 비표준어, 반의어, 동의어, 유사어, 줄임말, 본디말 등에 대한 요소와 속성은 ISO/DIS 12620의 개념관계, 개념구조, 관리정보의 상호참조 등에 걸쳐 폭 넓게 제시되고 있다. 그러나 TEI에서는 다른 표제어에 대한 참조에 있어서 참조되는 용어의 위치에 따른 구분만을 하고 있을 뿐 구체적인 구분은 하지 않고 있다. 또한 ISO/DIS 12620이나 TEI 모두 거센말, 여린말 등 앞서 제시된 한국어 고유의 관련 어형에 대한 참조표시는 제공하지 못하고 있다. 이에 다양한 참조지시속성 - 보라(<xrsee>), 표준어(<xrstd>), 비표준어(<xrxstd>), 반의어(<xrant>), 거센말(<xrstr>), 여린말(<xrsoft>), 큰말(<xrlarge>), 작은말(<xrsmall>), 높임말(<xrhon>), 친근체(<xrint>), 동의어(<xrsame>), 동음이의어(<xrsyn>), 참조어(<xrcfwd>), 줄임말(<xrshort>), 본디말(<xrlong>), 변환말(<xrchg>), 이형(<xrvar>) - 의 추가를 제안한다. MARTIF는 이러한 추가 요소들 중 ISO/DIS 12620에서 이미 제시하고 있는 요소가 아닌 것만 선택적으로 추가하고 TEI는 제안된 요소 전체를 추가하여야 한다.

16) 신규어에 대한 구분이 가능해야 한다.

현대에는 용어의 생성과 소멸이 빈번하므로 인코딩대상이 되는 사전에서 신규어라는 정보가 주어졌다면 그 정보를 따로 분리하여 나타낼 수 있어야 한다. MARTIF의 경우, '용법'에서 시간적 제약, 일시적 한정어를 나타내는 것과 유사한 방식으로 <termNote type='newterm'>을 새롭게 추가하여 사용할 것을 제안한다. TEI에서 역시 용법의 한 유형으로 <usg type=newterm>의 추가를 제안한다. 이처럼 용어의 최신성을 나타내고자 할 때에는 최신성의 기준이 되는 시간적 배경이 함께 제공되어야 하므로 소스가 되는 사전과의 참조가 명시되어야 한다. 다른 한편으로는 기존의 표제어에 대한 새로운 용어가 등장할 경우, 참조요소를 이용하여 기존의 표제어를 최신어로 연결해 줄 수 있다.

17) 용어관계를 나타내는 표제어끼리 참조를 이용하여 연결시켜주는 것이 필요하다. 이는 주로 용어데이터베이스나 시소러스에서 나타나는 형식이지만 사전항목으로 이루어진 데이터베이스도 확장된 개념의 용어데이터베이스이므로 병렬관계, 속관계, 전체-부분관계, 그리고 다중계층관계 등을 가진 항목끼리 참조해 줌으로써 정보의 활용성을 높일 수 있다. ISO/DIS 12620의 제9장에서는 TT, BT, NT, 및 RT의 개념 인코딩하기 위해 '<ref type=topTerm target=Z>...</ref>' 과 같은 형식을 제시하고 있다. 따라서 MARTIF에서는 별도의 확장 없이 용어관계를 나타내 줄 수 있다. 반면, TEI에서는 기존의 상호참조의 속성에 용어관계를 나타낼 수 있는 요소 <BT>, <NT>, <RT>, 그리고 <TT> 등의 속성을 추가로 정의할 것을 제안한다. 예를 들어, 표제어 '호

량'에 대한 용어관계 BT 및 TT를 <그림 10>과 같이 나타낼 수 있다.

```

<form>
  <orth>호랑이</orth>
  <xr><BT><ref>고양이과</ref></xr>
  <xr><BT><ref>식육목</ref></xr>
</form>

```

<그림 10> 용어관계의 예(TEI)

5 결 론

한국어 사전데이터의 활용성을 높이기 위해 일반사전, 용어집, 시소러스, 그리고 주제명표 목표 등 용어와 그 관련 정보를 나타내는 텍스트 자료를 그 항목요소와 구조의 측면에서 분석하여 망라적인 사전미시구조를 도출하였다. 미시구조의 모든 항목을 정보의 손실 없이, 그리고 이용에 효율성을 더할 수 있는 체계를 만들기 위하여 국제적으로 사용되고 있는 MARTIF와 인쇄사전용 TEI의 적용가능성을 분석하였다.

MARTIF와 용어데이터베이스용 TEI는 ISO/DIS 12620을 적용하고 있으므로 이 두 가지 형식은 근본적으로 같다고 할 수 있다. 반면, 인쇄사전용 TEI는 용어데이터베이스용 TEI와는 달리 사전에 맞게 인코딩 체계를 정의하고 있으며, ISO/DIS 12620과 구조, 내용, 특성, 요소, 태그 등 모든 면에서 확연히 다르다. 따라서 두 가지를 통합하거나 상호적용하는 것은 현실적으로 불가능하다고 보아야 하며, 상호간의 데이터변환 가능성 여부도 면밀히 검토되어야 할 문제이다.

문제가 되는 것은 MARTIF와 인쇄사전용 TEI 양쪽 모두 한국어용어데이터의 인코딩형식으로 적용할 수 있다는 사실이다. 따라서 본 연구에서는 한국어사전을 수용하기 위한 확장 부분에 대해서 독립적으로 확장의 필요성을 논함과 동시에 그 방안을 각각 제안하게 되었다. 따라서 두 가지 형식 중에서 하나를 채택해야 한다면 인쇄사전용 TEI에 비해 보다 많은 항목이 준비되어 있으며 공식적인 국제표준인 MARTIF를 채택하는 것이 바람직하리라 사료된다.

참 고 문 헌

김성혁. 1996. 우리말 용어를 위한 전문용어 교환형식. 『인지과학』, 7(4): 25-41.

김진희. 2001. 『XML을 이용한 문헌정보학용어 사전의 구조화 및 브라우저 설계』. 숙명여자대학교 대학원, 석사학위논문.

김현권. 1989. 언어사전 정의의 구성과 유형에 대하여. 『언어학』, 11: 73-93.

문화관광부. 2000a. 『21세기 세종계획: 전자사전 개발』. 제 3차년도 연구보고서.

———. 2000b. 『“한국어 말뭉치의 활용”(교재)의 개발』. 국어 정보화 연구 인력양성 사업 최종보고서.

박을용. 1993. 『수학대사전』. 서울: 한국사전연구사.

서상규. 2001. 우리 말글 연구회. 국어학 강좌: 한국어 사전의 문제. [online]. [인

- 용날짜 2001. 04. 25]. <<http://lexeme.yonsei.ac.kr/study/index.html>>.
- 양단희. 1992. 한국어 전자 사전 원형의 설계 및 구현. 『사전편찬학연구』, 4: 18-62. 서울: 탑출판사.
- 오현주. 1998. 『독·한 기계번역을 위한 전자사전 시스템에 관한 고찰』. 부산대학교 대학원, 석사학위논문.
- 윤성희. 1997. 『범용 한국어 전자사전 구축을 위한 정보 필드 구조의 설계에 관한 연구』. 『상명대산업과학연구』, 5: 157-166.
- 이상근. 1998. 영-한 기계번역을 위한 전자사전 작성 및 관리도구의 설계 및 구현. 전북대학교 대학원, 석사학위 논문.
- 이상섭, 송만석. 1992. 『전자사전 구현을 위한 어휘 데이터베이스의 설계와 국어 정보 분석 도구의 개발』. 연세논총, 28: 25-106.
- 이재성, 최병진, 이운재, 최기선. 1996. 텍스트 및 전자사전 관리시스템을 위한 표준 사전 표기언어의 설계. 『인지과학』, 7(4): 75-92.
- 정광, 이기용, 김홍규, 임해창, 강범모. 1995. 『한국어 데이터베이스의 설계 및 응용을 위한 기초 연구』. 서울: 민음사.
- 정휘웅. 1998. 『다국어지원 전자사전 개발을 위한 환경구축』. 부산대학교 대학원, 석사학위논문.
- 조재수. 2000. 辭典 編纂의 專門化에 對하여. 『第13回 言語情報 研鑽會 發表論文集: 21世紀의 辭典과 國語教育』, 105-110. 서울: 延世大學校 言語情報 開發研究院.
- 최병진, 이운재, 이재성, 최기선. 1996. 기계가독형 사전 구축을 위한 사전항목의 논리 구조. 『인지과학』, 7(2): 75-94.
- 崔錫斗. 2000. 專門用語學의 內容과 課題. 『知識處理研究』, 1(1): 61-75.
- 한국정보통신기술협회. 2000a. 『용어/사전편찬용 자기데이터 교환 형식』. 서울: 한국정보통신기술협회.
- . 2000b. 『정보통신 용어사전』. 제4판. 서울: 한국정보통신기술협회.
- Alder, Aaron D. 1998. *An Experiment in Blind Terminology Interchange: Developing and Testing Conversion Algorithms for Externally Supplied Data*. Ph.D. diss., Brigham Young University.
- “Blind MARTIF data categories, maximal set.” [online]. [cited 2000.05.09]. <<http://www.ttt.org/clsframe/BTFdatcats.html>>.
- Budin, Gerhard and Alan K. Melby. 2000. “Accessibility of multilingual terminological resources—Current problems and prospects for the future.” [online]. [cited 2001. 2.19]. <http://www.ttt.org/Salt/Athens_Paper.doc>.
- Ide, Nancy M. and Jean Veronis. 1995. “Encoding dictionaries.” *Computers and the Humanities*, 29(1): 167-179.

- ISO 6156:1987 Magnetic Tape Exchange Format for Terminology*. 1987. Geneva: ISO.
- ISO/DIS 12200:1999 Machine-readable Terminology Interchange Format (MARTIF) -- Negotiated Interchange*. 1999. Geneva: ISO.
- ISO/DIS 12620:1999 Computer Applications in Terminology -- Data Categories*. 1999. Geneva: ISO.
- Kang, Beom-Mo. 1998. "Modifying the TEI DTD: The case of Korean dictionaries." *Computers and the Humanities*, 31(1): 433-449.
- Melby, Allen and Daniel Hardman. 1996. "Importing terminology from multiple sources in three phases: inspection, adjustment and adoption." *Proceedings of TKE '96, Terminology and Knowledge Engineering*. Frankfurt: INDEKS-Verlag, 197-204.
- Schmitz, Klaus-Dirk and Sue Ellen Wright. 1997. "Terminology interchange — needs, approaches, solutions, problems." TSTT '97 Proceedings. 2nd International Conference on Terminology, Standardization and Technology Transfer. 1997년 8월 3-8일. Beijing: Encyclopedia of China Publishing House, 346-352.
- TEI. 1999. *Guidelines for Electronic Text Encoding and Interchange*. v.1-2.