한중일영 다국어 어휘 데이터베이스의 모형

차재은
(고려대학교 민족문화연구원)
nhje@chollian.net

강범모
(고려대학교 언어과학과)
bmkang@korea.ac.kr, http://ikc.korea.ac.kr/~bmkang

Abstract

This paper is a report on part of the results of a research project entitled "Research and Model Development for a Multi-Lingual Lexical Database". It is a six-year project in which we aim to construct a model of a multilingual lexical database of Korean, Chinese, Japanese, and English. Now we have finished the first two-year stage of the project. In this paper, we present the goal of the project, the construction model of items in the lexical database, and the possible (semi-)automatic methods of acquisition of lexical information. As an appendix, we present some sample items of the database as an illustration.

1. 기본 목표

본 연구[1]의 기본적인 목표는 장기적으로는 실용화될 수 있는 5만 단어 규모의 한, 중, 일, 영 4개 국어 어휘 데이터베이스 구축을 목적으로 하되, 우선 2년 간에 걸쳐 i) 그 기반이 되는 어휘정보 표상 및 획득 방법을 위한 기초 연구를 수행하고 ii) 다국어 어휘 데이터베이스 모형을 개발하며 iii) 500개 어휘를 대상으로 한, 중, 일, 영 다국어 어휘 데이터베이스를 구축하는 것이었다. 2, 3, 4절에서 다국어 어휘 데이터베이스의 설계, 어휘정보 획득 방법, 다국어

1 본고의 내용은 2000년도 한국 학술진흥재단의 지원에 의해 연구된(KRF-2000-605-Y00248) '다국어 어휘 데이터베이스의 구축 방법론 연구 및 모형 개발' 결과의 일부에 대한 요약이다. 자세한 연구 결과는 강범모, 이유선, 차재은(2002 예정)으로 출판될 것이다.

어휘 데이터베이스의 표본에 대해 소개한다.

2. 데이터베이스 설계

유로워드넷(Vossen 1998)은 다국어 DB를 설계할 때 참조할 수 있는 모델이다. 유로워드넷의 경우 각 언어의 워드넷을 먼저 구축하고, ILI(중간언어 색인)로 각 언어 워드넷에 포함된 각각의 동의어 집합들을 연결시킨다.

반면, 한, 중, 일의 경우는 자국어 워드넷이 구축되어 있지 않다. 따라서 현 단계 다국어 DB 구축은 한국어를 중심으로 영어, 중국어, 일본어 대역 정보가 평면적으로 나열된 1단계 작업을 수행하고 이를 토대로 유로워드넷의 ILI를 이용하여 다국어로 확장하는 전략이 유효하다. 이러한 방법론에 따른 다국어 DB는 다음과 같은 얼개를 가진다.

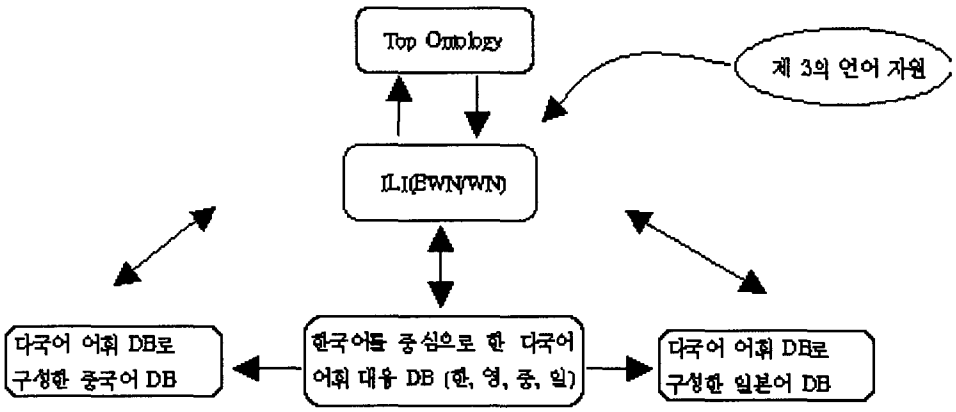

그림 1 한국어 중심의 다국어 어휘 데이터 베이스 얼개

어휘 데이터베이스의 정보 내용은 Oracle, Informix DB와 같은 상용 데이터베이스 관리 시스템에 통합되기에 앞서, 구조화된 문서 형식으로 관리하는 것이 바람직하다. 본 연구에서는 이미 구축된 전자 사전들과의 호환성을 염두에 두고 XML 문서 형식으로 고안하였다. 다국어 어휘 데이터베이스 작성은 한글 워디안을 통해 윈도우 2000의 유니코드를 기반으로 해서 구축하였는데, 이것은 향후 다양한 데이터베이스 구조로 전환될 수 있다.

내용적으로는 기존 (전자) 사전의 결과물을 바탕으로 연어, 은유 정보를 추가하고 각국어의 대역어 정보를 상세하게 부가하는 방향으로 데이터베이스의 정보를 설계하였다. 이러한 방법론에 따라 설계된 다국어 어휘 데이터 베이스는 명사의 경우 다음과 같은 계층 구조를 가진다. * 로 표시된 정보는 반복 가능한 것이다.

<entry>	<mntGrp>	<Kwrt>		
		<Ewrt>		
		<Cwrt>		
		<Jwrt>		
		<note>		
	<headGrp>	<form>		
		<org>		
		<pos>		
		<xpos>		
	<sense>*	<eg>		
		<cl>		
		<domain>		
		<sem>		
		<English>	<transE>, <E_count>, <E_xpos>	
		<Chinese>	<transC>, <C_cl>, <C_xpos>	
		<Japanese>	<transJ>, <J_cl>, <J_xpos>	
		<ILI>		
		<lr>	<syn>, <ant>, <hyper> <hypo>, <holo>, <mero>	
		<coll>	<comb_aj>*	<comb_aj_exp> <comb_aj_transE> <comb_aj_transC> <comb_aj_transJ>
			<comb_v>*	<comb_v_exp> <comb_v_transE> <comb_v_transC> <comb_v_transJ>
			<comb_n>*	<comb_n_exp> <comb_n_transE> <comb_n_transC> <comb_n_transJ>
	<idiomGrp>	<idiom>	<idiom_exp>, <idiom_transE>, <idiom_transC>, <idiom_transJ>	
	<metaphorGrp>	<meta>	<meta_exp>, <meta_transE>, meta_transC>, <meta_transJ>	

아래에 각 품사별 데이터베이스의 기본 설계 방향을 제시한다. '부사'는 명사에서 '은유' 부분이 빠지고 동사/형용사는 기본적으로 틀이 같다.

ㄱ) 명사/부사 정보 설계 (*는 반복 가능)

```
entry mntGrp
      headGrp
      Sense    *Lr
               coll*
      IdiomGrp
               Idiom*
      MetaphorGrp
               Metaphor*
```

ㄴ) 동사/형용사 정보 설계 (*는 반복 가능)

```
entry mntGrp
      headGrp
      Sense    *Lr
               caseFrame*
      IdiomGrp         Idiom*
```

이제 명사의 집필 예를 중심으로 각 필드에 대한 세부적인 사항을 살펴보도록 하자. 설명은 큰 구획을 기준으로 번호를 붙여서 달아주었으며, sense가 많을 경우 하나만 제시하기로 한다. 따라서 다음 예 '물' 에서 sense 2, 3, 4 번은 생략하였다.

①<entry>
②<mntGrp>
③<Kwrt>차준경(01/07/25)</Kwrt>
④ <Ewrt>이유선(01/07/27)</Ewrt>
⑤ <Cwrt>김세영(01/07/31)</Cwrt>
⑥ <Jwrt>소명인(01/07/31)</Jwrt>
⑦<note>차재은(02/02/04, 센스 조정, 이디엄, 은유 추가)</note>
⑧</mntGrp>

① 어휘 항목: 한 어휘 항목을 여는 태그이다. ② 관리 정보 구획(management group): 누가, 언제, 어떤 작업을 하였는지 과정을 기록하기 위한 필드이다. ③ 한국어 집필자 및 작성일 (writer and date for Korean): 한국어 정보를 누가, 언제 작성하였는지 기록한다. ④ 영어 집필자 및 작성일 (writer and date for Engli sh): 영어 정보를 누가, 언제 작성하였는지 기록

한다.⑤ 중국어 집필자 및 작성일 (writer and date for Chinese): 중국어 정보를 누가, 언제 작성하였는지 기록한다.⑥ 일본어 집필자 및 작성일 (writer and date for Japanese): 일본어 정보를 누가, 언제 작성하였는지 기록한다.⑦ 메모: 어떤 정보를 수정, 삭제하였을 경우, 기타 작업상 반드시 참조해야 할 의견이 있는 경우 해당 정보를 써 넣는다. ⑧ 관리 정보 구획을 닫는 태그이다. 관리 정보 그룹은 다국어임을 고려할 때 누가 언제 무슨 작업을 하였는지 명시적으로 기록할 필요가 있어서 필수적이다

㉠<headGrp>
㉡<form>물</form>
㉢<org> </org>
㉣<pos>n</pos>
㉤<xpos></xpos>
㉥</headGrp>

㉠ 표제정보 구획(head group): 표제 정보 구획을 여는 태그이다. 불일치 품사 정보는 다국어 의미 대응에서 발생하는 품사의 불일치를 고려하여 넣었다. ㉡ 표제어 형태(form): 하나의 어휘 항목에서 기초적이고도 필수적인 (각국어의) 표기 형태를 표시한다. 명사의 경우 단독형이 제시된다. ㉢ 표제어 어원(origin): 한, 중, 일은 공통적으로 한자를 사용한다. 한, 중, 일 한자 형태 비교 및 중의성 해소에 사용될 가능성을 고려하여 넣어준다. ㉣ 품사범주 (part-of-speech): 해당 어휘 항목의 품사를 표시한다. 명사는 n으로 표시한다. ㉤ 불일치 품사 정보: 한국어에서 동작성 명사와 그에 해당하는 -하다, -적, -되다 파생어를 기록하기 위한 필드이다.㉥ 표제 정보 구획을 닫는 태그이다.

①<sense n=1>
②<eg>아침마다 공복에 물 한 컵을 마시면 몸에 좋다</eg>
③<cl> </cl>
④ <domain> </domain>
⑤<sem>자연물</sem>

① 구체적인 명사 의미를 여는 태그이다. 해당 어휘 항목이 다의성을 가지고 있을 경우 다의 번호를 표시한다. 뜻이 하나인 경우 n= 1로 끝난다.② 전형적인 뜻을 보여주는 예문을 넣는다. 실제 데이터 구현 과정에서는 보여주지 않는다. ③ 한국어 분류사 정보를 넣는다. 개, 자루, 마리 등의 분류사가 제시된다. 한국어-외국어 번역에서 이용될 수 있다. ④ 해당 의미가 어떤 영역에 속하는지 표시한다. 주로 전문어인지 아닌지를 밝히게 되며 전문 영역의 판정 기

준은 표준국어대사전에 따른다. 번역이나 중의성 판정에 이용 가능하다. ⑤ 해당 의미가 어떤 의미 부류에 속하는지 표시한다. 의미 부류의 선정 기준은 세종 전자 사전에 따른다. 의미 기반 번역에 이용 가능하며, 상위어 추출의 참고 자료로 활용된다.

㉠<English>
㉡<transE>water</transE>
㉢<E_count>U</E_count>
㉣<E_xpos> </E_xpos>
㉤</English>

㉠ 대역어 정보 중 영어에 관한 정보를 여는 태그이다. ㉡ 해당 한국어 의미에 대한 영어 대역어 정보를 기록한다.㉢ 영어 대역어가 가산 명사인지 불가산 명사인지 표시한다. 가산이면 C, 불가산이면 U로 표시한다. ㉣ 한국어에 대한 대역어의 품사가 일치하지 않는 경우에 표시한다. 이 경우 명사가 아닌 대역어의 품사를 표시한다. ㉤ 대역어 정보 중 영어에 관한 정보를 닫는 태그이다.

①<Chinese>
②<transC>水</transC>
③<C_cl> </C_cl>
④<C_xpos> </C_xpos>
⑤</Chinese>

① 대역어 정보 중 중국어에 관한 정보를 여는 태그이다. ② 해당 한국어 의미에 대한 중국어 대역어 정보를 기록한다.③ 중국어 대역어가 가산 명사인지 불가산 명사인지 표시한다. 가산이면 C, 불가산이면 U로 표시한다. ④ 한국어에 대한 중국어 대역어의 품사가 일치하지 않는 경우에 표시한다. 이 경우 명사가 아닌 대역어의 품사를 표시한다. ⑤ 대역어 정보 중 중국어에 관한 정보를 닫는 태그이다.

㉠<Japanese>
㉡<transJ>水</transJ>
㉢<J_cl> </J_cl>
㉣<J_xpos> </J_xpos>
㉤</Japanese>

㉠ 대역어 정보 중 일본어에 관한 정보를 여는 태그이다. ㉡ 해당 한국어 의미에 대한 일본어 대역어 정보를 기록한다.㉢ 일본어 대역어가 가산 명사인지 불가산 명사인지 표시한다. 가산이면 C, 불가산이면 U로 표시한다. ㉣ 한국어에 대한 일본어 대역어의 품사가 일치하지 않는 경우에 표시한다. 이 경우 명사가 아닌 대역어의 품사를 표시한다. ㉤ 대역어 정보 중 일본어에 관한 정보를 닫는 태그이다.

①<ILI>2037290</ILI>
②<lr>
③<syn></syn>
④<ant> </ant>
⑤<hyper>액체</hyper>
⑥<hypo> </hypo>
⑦<holo> </holo>
⑧<mero></mero>
⑨</lr>

① 해당 한국어 의미에 대한 유로워드넷 ILI의 offset을 기록한다.② 한국어 사이의 어휘 의미 관계를 여는 태그이다. ③ 해당 어휘의 동의어나 유의어를 모두 기록한다.④ 해당 어휘의 반의어를 모두 기록한다.⑤ 해당 어휘의 상의어를 모두 기록한다.⑥ 해당 어휘의 하의어를 모두 기록한다.⑦ 해당 어휘의 전체어를 모두 기록한다.⑧ 해당 어휘의 부분어를 모두 기록한다. ⑨ 한국어 사이의 어휘 의미 관계를 닫는 태그이다.

㉠<coll>
㉡<comb_aj n=1>
㉢<comb_aj_exp>맑은 물;물이 맑다</comb_aj_exp>
㉣<comb_aj_transE>pure water; </comb_aj_transE>
㉤<comb_aj_transC>?水; </comb_aj_transC>
㉥<comb_aj_transJ>澄んだ水水;が澄む</comb_aj_transJ>
㉦</comb_aj>

㉠ 해당 의미에 대한 자유 결합 및 연어 정보를 여는 태그이다. 연어 정보는 다국어에서 나타날 수 있는 단어 대 구 표현의 대응 및 일대일 단어 대응의 예측 불가능성을 고려한 것으로 기계 번역 등에 이용될 가능성이 크다. (연어의 개념 및 범위에 대해서는 4.2 참조).㉡ 해당 명사 의미에 대한 형용사 연어를 여는 태그이다.㉢ 형용사 연어의 표기를 적는다. 활용하지 않

은 기본형과 관형형을 제시할 수 있다. ㉣ 형용사 연어에 대한 영어 표현을 적는다. ㉤ 형용사 연어에 대한 중국어 표현을 적는다. ㉥ 형용사 연어에 대한 일본어 표현을 적는다. ㉦ 해당 명사 의미에 대한 형용사 연어를 닫는 태그이다.

①<comb_v n=1>
②<comb_v_exp>물을 마시다</comb_v_exp>
③<comb_v_transE> drink water </comb_v_transE>
④<comb_v_transC>喝水</comb_v_transC>
⑤<comb_v_transJ>水を?む</comb_v_transJ>
⑥</comb_v>

① 해당 어휘에 대한 동사 연어를 여는 태그이다.② 동사 연어의 표기를 적는다. 주로 명사 + 동사형이 제시된다. ③ 동사 연어에 대한 영어 표현을 적는다. ④ 동사 연어에 대한 중국어 표현을 적는다. ⑤ 동사 연어에 대한 일본어 표현을 적는다. ⑥ 해당 어휘에 대한 동사 연어를 닫는 태그이다.

㉠<comb_n>
㉡<comb_n_exp> </comb_n_exp>
㉢<comb_n_transE> </comb_n_transE>
㉣<comb_n_transC> </comb_n_transC>
㉤<comb_n_transJ> </comb_n_transJ>
㉥</comb_n>
㉦</coll>
㉧</sense>

㉠ 해당 어휘에 대한 명사 연어를 여는 태그이다. ㉡ 명사 연어의 표기를 적는다. 주로 명사 + 명사형이 제시된다. ㉢ 명사 연어에 대한 영어 표현을 적는다. ㉣ 명사 연어에 대한 중국어 표현을 적는다. ㉤ 명사 연어에 대한 일본어 표현을 적는다. ㉥ 해당 어휘에 대한 명사 연어를 닫는 태그이다. ㉦ 연어 정보를 닫는 태그이다. ㉧ 의미 구획을 닫는 태그이다.

①<IdiomGrp>
②<idiom n=1>
③<idiom_exp>물(이) 오르다</idiom_exp> /*전성기에 와 있다*/
④<idiom_transE>be at the height of its prosperity</idiom_transE>

⑤<idiom_transC> </idiom_transC>
⑥<idiom_transJ> </idiom_transJ>
⑦</idiom>
⑧<IdiomGrp>

① 관용어 정보를 여는 태그이다. ② 관용어 번호를 적는다. ③ 해당 관용 표현을 적는다. ④ 관용어에 대한 영어 표현을 적는다. ⑤ 관용어에 대한 중국어 표현을 적는다. ⑥ 관용어에 대한 일본어 표현을 적는다. ⑦ 관용어 정보를 닫는 태그이다.⑧ 관용어 그룹 정보를 닫는 태그이다.

㉠<MetaphorGrp>
㉡<meta n=1>
㉢<meta_exp>물 찬 제비 같다</meta_exp> /*옷맵시가 매우 깔끔하다 */
㉣<meta_transE>be stylishness</meta_transE>
㉤<meta_transC>亭亭玉立</meta_transC>
㉥<meta_transJ> </meta_transJ>
㉦</meta>
㉧</MetaphorGrp>
㉨</entry>

㉠ 은유 그룹을 여는 태그이다. 은유 그룹은 각 언어의 은유 표현 대응을 보여주기 위한 것으로 주로 구 단위 이상의 은유가 여기에 해당한다. ㉡ 은유 번호를 여는 태그이다. ㉢ 은유 표현을 적는다. ㉣ 은유 표현에 대한 영어 대역어를 적는다. ㉤ 은유 표현에 대한 중국어 대역어를 적는다. ㉥ 은유 표현에 대한 일본어 대역어를 적는다. ㉦ 은유 표현을 닫는 태그이다. ㉧ 은유 그룹을 닫는 태그이다. ㉨ 어휘 항목을 닫는 태그이다.

이제 명사 틀의 내용을 정리해 보자. 명사 틀은 크게 표제어부, 의미 그룹, 연어 그룹, 관용어 그룹으로 구성되어 있다. 의미 그룹은 다국어 DB에서 가장 핵심적인 부분으로 영어, 중국어, 일본어 대역 정보가 들어 있으며 각 정보에는 언어간 특성에 맞게 분류사나 가산/불가산 명사 정보, 교차 품사 정보 등이 포함되어 있다. 또한, ILI가 포함되어 다국어 개념망으로 확장될 가능성을 열어 놓고 있다. ILI는 자동 추출 과정, 혹은 반자동 추출을 통하여 채워질 수 있다. 어휘 의미 관계는 기본적으로 한국어에 대한 정보를 넣어 놓고 있으나 해당 반의어나 상위어가 다국어 어휘 데이터 베이스의 항목으로 선정되어 포함될 경우 다국어 대응도 충분히 보여줄 수 있다.

3. 어휘정보 획득 방법

본 과제에서는 기존 어휘 자원을 최대한 이용하여 필요한 어휘 정보를 획득하였는데, 본 과제의 어휘 정보 획득 절차를 보이면 대략 다음과 같다.

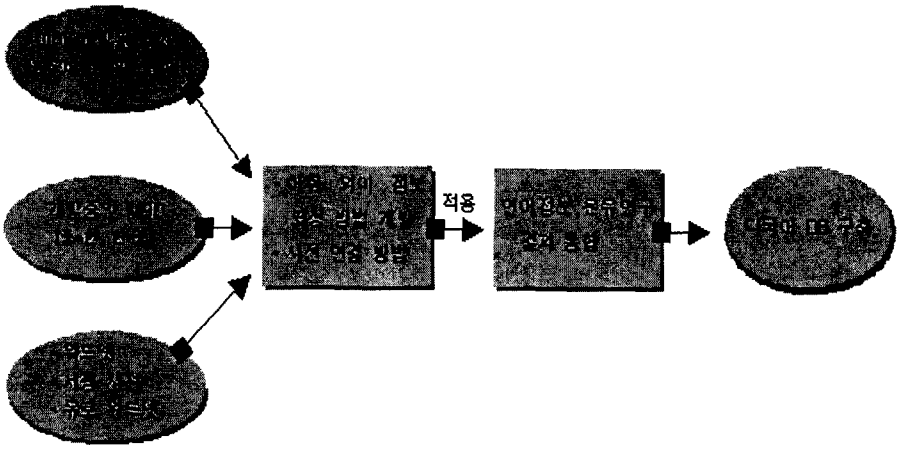

그림 2 다국어 어휘 데이터 베이스 구축 방법

즉, 이미 개발 중인 전자사전과 워드넷, 유로 워드넷 등의 어휘 자원을 통해서 자동 추출된 표제어, 품사, 대역어, 관련어휘 정보 등을 바탕으로 연어와 관용구 정보를 추가하여 다국어 어휘데이터베이스를 구축하였다. 이제 각 단계별로 어휘 정보의 획득 절차를 소개한다.

1) 세종사전

세종 전자사전(홍재성 외 1998-2001)은 각 단어들의 특성에 따라 다양한 하위사전으로 구성된다. 이 사전의 특징은 체계적인 언어학적 분석을 바탕으로 기술된 어휘정보를 컴퓨터로 활용할 수 있도록 구조화, 형식화시켰고, 기존 사전과 말뭉치 등을 활용하여 다양하고 풍부한 언어정보를 지니고 있다는 점이다. 특히 구조 자체가 SGML(Standard Generalized Markup Language)의 형식을 따르고 있어 체계적이기 때문에 데이터에 대한 프로그램을 이용한 파싱(parsing)이 용이하게 되어 있다. 다음은 체언-단일어 사전의 틀을 나타낸 것인데, ' <' 와 ' >', ' %' 나 ' @', ' =' 등으로 이루어진 기호의 정의와 표기, 용례가 형식화되어 있고, 각 정보를 이루는 내용(content)이 ' [ ]' 안에 있으므로 쉽게 정보를 추출해 낼 수 있다.

본 과제에서는 SGML 표준을 따르는 형식을 지닌 세종 전자사전으로부터 각각 명사, 동사, 형용사, 부사 정보를 추출하였다. 다음은 세종 사전으로부터 다국어 어휘 데이터베이스의 틀로

추출, 변환된 명사 정보를 나타낸 것이다(' 세종 사전의 정보구획 → 다국어 어휘 데이터베이스의 정보구획' 의 형식으로 표시되어 있다).

- 표제정보 구획/표제어 형태 <toplevel> @form → <headGrp><form>
- 표제정보 구획/품사범주 <toplevel> @pos → <headGrp><pos>
- 의미정보 구획/전문분야 <sense num> @domain → <sense><domain>
- 의미정보 구획/의미하위부류 <sense num> @sem → <sense><sem>
- 의미정보 구획/영어대역어 <sense num> @trans → <sense><transE>
- 어휘관계 구획/동의어 <sense num> @syn → <sense><syn>
- 어휘관계 구획/반의어 <sense num> @ant → <sense><ant>
- 어휘관계 구획/상위어 <sense num> @hyper → <sense><hyper>
- 어휘관계 구획/하위어 <sense num> @hypo → <sense><hypo>
- 어휘관계 구획/전체어 <sense num> @holo → <sense><holo>
- 어휘관계 구획/부분어 <sense num> @mero → <sense><mero>
- 숙어/관용 구획 <toplevel><froz> → <IdiomGrp><idiom exp>
- 통사정보 B구획 <sense num><syn_b> @comb_aj → <headGrp><sense><coll><comb_aj_exp>
  <sense num><syn_b> @comb_v → <headGrp><sense><coll> <comb_v_exp>

2) 한-외국어 사전

지금까지 구축된 전자 사전들은 완전히 전산적 처리를 염두에 두고 구축되었다기보다 인쇄 사전의 편찬을 위한 중간물로 존재하는 경우가 많고, 인터넷 서비스나 프로그램 사전의 목적으로 개발된 사전이 일부 있기는 하지만 그 수가 많지 않은 실정이다. 따라서 이와 같은 전자 사전 형태의 파일들은 가공하여 사용함에 있어 문제가 발생될 수 있다.

예를 들어 인쇄 사전의 편찬 과정에서 생성된 전자 파일을 가공이 용이한 형태의 전자 파일로 바꾸는 과정에서 구분 기호라든지 내부적으로 정의된 제어 기호들이 사라지는 등의 정보 손실이 발생된다. 이러한 정보의 손실은 가공 과정에서 자동적인 파싱을 거의 불가능하게 할 정도로 영향을 미칠 수 있다. 그리고 인터넷 서비스 등의 목적을 위한 전자 사전의 경우에는 정확한 소스 파일을 구하기 어렵기 때문에 브라우저(browser)에 표시되는 정보를 수작업으로 획득하는 방법으로 전자 파일을 얻을 수 있는데, 전자 파일로 만드는 과정에서 개행 문자(Carriage Return) 등 불필요한 문자의 삽입이 발생될 수 있고, 입력된 내용들의 구조가 체계적이지 못할 가능성이 크다.

따라서 기계 가독형 사전 형식이 갖추어져 있지 않는 사전 형태는 프로그램에서 직접 이용하는 것보다 일정 수준으로 가공하여 기계 가독형 중간 파일을 생성하는 것이 사전 처리에 있어 효율성을 증가시킬 수 있는데, 이 과정을 도식화하면 다음과 같다.

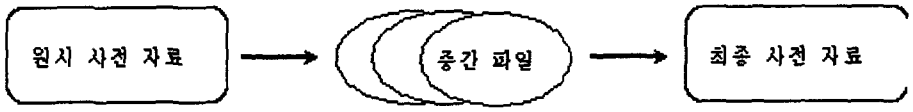

그림 3 중간 파일 생성을 이용한 한-외국어 전자 사전 구축

이러한 중간 파일의 생성을 조금 구체적으로 살펴보면 다음과 같다.

ⓐ 한-일 전자 사전을 텍스트 파일로 바꾼 형태

가깝다

[形?ㅂ變] 近(ちか)い. ①距離(きょり)または時間(じかん)のへだたりが少(すく)ない. near 「가까운 시일 : 近い時日(ちじつ)」 ②似(に)ている. resemble 「짐승에 - : 獸(けもの)に似ている」 ③血緣(けつえん)が近い. 「가까운 친척 : 近い親戚(しんせき)」 ④親(した)しい. 睦(むつ)ましい. -----(ⓐ) 「가까운 친구 : 親しい友達(ともだち)」 ↔멀다 가까-이 [副?하다他] 近(ちか)く. ※가깝디-가깝다 [形?ㅂ變] ごく近(ちか)い. 甚(はなは)だ近い ↔멀

ⓑ 일정 수준으로 형식화한 중간 파일

% 가깝다
@pos = [形]
@sem = [近(ちか)い]
@inflect = [ㅂ變]
@sense
[tab]①距離(きょり)または時間(じかん)のへだたりが少(すく)ない
[tab][tab]#trans = [near]
[tab][tab]#eg = [가까운 시일 : 近い時日(ちじつ)]
[tab]②似(に)ている
[tab][tab]#trans = [resemble]
[tab][tab]#eg = [짐승에 - : 獸(けもの)に似ている]
[tab]③血緣(けつえん)が近い
[tab][tab]#trans = []
[tab][tab]#eg = [가까운 친척 : 近い親戚(しんせき)]
[tab]④親(した)しい|睦(むつ)ましい ---------- (ⓑ)
[tab][tab]#trans = []
[tab][tab]#eg = [가까운 친구 : 親しい友達(ともだち)]
@etc
[tab]멀다 가까-이 [副?하다他] 近(ちか)く. ※가깝디-가깝다 [形?ㅂ變] ごく近(ちか)い. 甚(はなは)だ近い ↔멀다

위의 사전 텍스트는 프로그램 사전을 곧바로 텍스트 파일로 전환했을 때의 형식이다. ⓐ와 같이 순차적으로 나열해 놓은 기술 방식을 적당한 가공을 통하여 ⓑ와 같이 변환시킨다면 훨씬 체계적인 정보의 획득이 가능하다.

3) ILI

본 과제에서는 유로워드넷(Vossen 1998))의 ILI를 그대로 이용하여, 본 다국어 어휘 데이터베이스와 다른 언어들과의 대응관계를 설정하였다. 다음은 유로 워드넷의 ILI 레코드의 일부분이다.

유로 워드넷 ILI레코드의 예

```
0 @51@ ILI_RECORD
  1 PART_OF_SPEECH "n"
  1 WORDNET_OFFSET 22455
  1 GLOSS "mutual dealings or connections among persons or groups: "international relations"& 03
04 2ndOrderEntity Agentive Cause Dynamic SituationType"
  1 VARIANTS
    2 LITERAL "relations"
      3 SENSE 1
```

실제 데이터의 예에서 보는 바와 같이 각각의 ILI 레코드에는 고유의 일련번호(Offset number)가 부여되어 있으며, 또한 개별 어휘의 의미를 기술한 부분(Gloss), 그리고 실제로 어떠한 영어 어휘로 나타나는지에 대한 항목(Literal)이 포함되어 있다.

실제 본 데이터베이스 구축시에는 이러한 정보 항목들 중 Literal 부분과 Offset 부분이 데이터베이스 자료 내의 <ILI> 항목에 연결되게 된다. 원 자료 기술이 한국어 항목을 기준으로 하고 그에 영어, 중국어, 일본어 항목을 부가시키는 형식으로 되어 있으므로, 유로 워드넷의 ILI 레코드와 연결하기 위해서 <English> 항목의 하위 항목인 <transE>에 나타나는 영어 대역어 정보를 이용할 수 있다.

ILI 레코드로부터 일련번호를 얻는 작업은 자동적인 처리가 불가능하기 때문에 작업자가 해당 어휘의 센스에 대응되는 일련번호를 판단할 기준이 되는 정보를 제공해야 한다. 이러한 기준 정보는 ILI 레코드들로부터 추출할 수 있다. 그런데 다의어의 경우 여러 ILI 레코드에서 사용되기 때문에 해당 어휘를 기준으로 사용된 모든 ILI 레코드를 추출해야 한다. 따라서 이러한 작업의 편의를 위해 모든 ILI 레코드를 RDB 모델로 변환하여 저장하게 된다. 다음은 개별 어휘 항목의 영어 대역어에 해당하는 ILI 기준 데이터를 추출하는 과정을 도식화한 것이다.

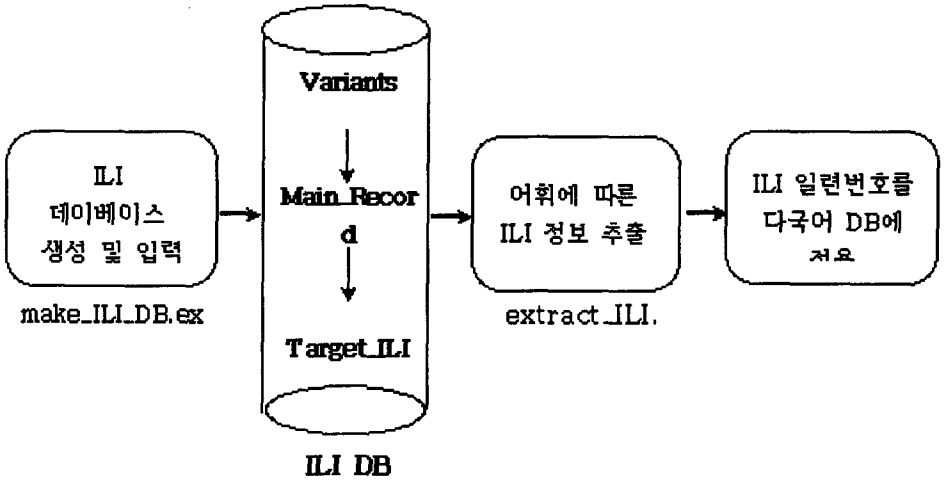

그림 4 ILI 일련번호를 다국어 데이터베이스에 적용하는 과정

본 과제에서는 이러한 방법론에 따라 다국어 어휘데이터 베이스를 구축하였는데 다음 장에서는 데이터의 표본을 제시한다.

4. 한중일영 다국어 어휘데이터 베이스 표본

각 품사별로 다국어 어휘데이터 베이스 표본을 부록으로 제시한다. 명사, 부사, 동사, 형용사의 표본으로 '땅', '다', '걸치다', '크다'를 골랐으며 센스는 하나만 제시하였다. (부록 참조)

5. 문제점 및 의의

본 과제가 가지는 방법상의 문제점이라면, 기존 사전으로부터 필요한 정보를 자동 처리하여 획득할 수 있는 방법에 한계가 있고, 획득된 정보의 질에 대한 평가 과정에 시간이 많이 걸린다는 것을 지적할 수 있다. 또, ILI로의 연결은 ' 의미' 대 ' 의미' 의 대응 작업이므로 시간이 많이 소요되고 ILI에서 해당하는 개념을 찾을 수 없는 경우도 발생한다.

그럼에도 불구하고 본 과제의 연구 결과는 다국어 정보처리를 위한 언어 정보 자원의 구축 방법의 모색이라는 측면에서 의의를 찾을 수 있으며, 앞으로 정보 검색이나 기계 번역 등 실용적 면이나 어휘 의미 연구 등 이론적 분야에서 유용하게 쓰일 수 있는, 적정한 규모의 어휘 데이터베이스를 구축하기 위한 첫 단계를 마쳤다는 데 의미가 있다.

참고문헌

강범모, 이유선, 차재은 (2002 예정) 다국어 어휘 데이터베이스의 구축 방법론 연구 및 모형 개발 1, 서울: 고려대학교 민족문화연구원.

홍재성, 강범모 외(1998-2001), 21세기 세종계획 전자사전개발 연구보고서, 문화관광부.

Vossen, P.(ed.) (1998) "Introduction to EuroWordnet," in Computers and the Humanities, Vol. 32, Nos. 2-3.

부록: 다국어 어휘 데이터베이스 표본

*아래 '땅', '다', '걸치다', '크다'의 예에서 센스는 한 개만 제시함.

1) 명사

```
<entry>
      <mntGrp>
            <Kwrt>차준경(01/07/10); 문선영(01/11/22)</Kwrt>
            <Ewrt>이유선(01/07/18)</Ewrt>
            <Cwrt>김세영(01/07/19)</Cwrt>
            <Jwrt>소명인(01/07/13)</Jwrt>
            <note> </note>
      </mntGrp>
      <headGrp>
            <form>땅</form>
            <org> </org>
            <pos>n</pos>
            <xpos> </xpos>
      </headGrp>
      <sense n=1>
            <eg>이 땅에서는 많은 유물이 나오고 있다</eg>
            <cl></cl>
             <domain> </domain>
            <sem>장소/자연</sem>
            <English>
                  <transE>ground</transE>
                  <E_count>U</E_count>
                  <E_xpos> </E_xpos>
            </English>
            <Chinese>
                  <transC>地</transC>
                  <C_cl> </C_cl>
                  <C_xpos> </C_xpos>
            </Chinese>
```

```
        <Japanese>
                <transJ>土 | 地 | 土地</transJ>
                <J_cl> </J_cl>
                <J_xpos> </J_xpos>
        </Japanese>
        <ILI>5720524_w</ILI>
        <lr>
                <syn>대지</syn>
                <ant>하늘 </ant>
                <hyper> </hyper>
                <hypo> </hypo>
                <holo> </holo>
                <mero> </mero>
        </lr>
        <coll>
                <comb_aj>
                        <comb_aj_exp> </comb_aj_exp>
                        <comb_aj_transE> </comb_aj_transE>
                        <comb_aj_transC> </comb_aj_transC>
                        <comb_aj_transJ> </comb_aj_transJ>
                </comb_aj>
                <comb_v n=1>
                        <comb_v_exp>땅을 파다</comb_v_exp>
                        <comb_v_transE>dig in the ground</comb_v_transE>
                        <comb_v_transC>?土</comb_v_transC>
                        <comb_v_transJ>土地を掘る</comb_v_transJ>
                </comb_v>
                <comb_v n=2>
                        <comb_v_exp>땅에 묻다</comb_v_exp>
                        <comb_v_transE>bury a thing in the ground</comb_v_transE>
                        <comb_v_transC>埋在地里</comb_v_transC>
                        <comb_v_transJ>地中に埋める</comb_v_transJ>
                </comb_v>
                <comb_v n=3>
                        <comb_v_exp>땅에 떨어지다</comb_v_exp>
                        <comb_v_transE> </comb_v_transE>
                        <comb_v_transC>落在地上</comb_v_transC>
                        <comb_v_transJ>地に落ちる</comb_v_transJ>
                </comb_v>
                <comb_n>
                        <comb_n_exp> </comb_n_exp>
                        <comb_n_transE> </comb_n_transE>
                        <comb_n_transC> </comb_n_transC>
                        <comb_n_transJ> </comb_n_transJ>
                </comb_n>
        </coll>
</sense>
<IdiomGrp>
        <idiom n=1>
                <idiom_exp>땅을 파먹다</idiom_exp>
```

```
                <idiom_transE>live by farming</idiom_transE>
                <idiom_transC>?庄稼</idiom_transC>
                <idiom_transJ>地面を掘って生計をたてる</idiom_transJ>
        </idiom>
        <idiom n=2>
                <idiom_exp>땅 짚고 헤엄치기</idiom_exp>
                <idiom_transE>That' s quite an easy job</idiom_transE>
                <idiom_transC>十拿九?丨完全有把握</idiom_transC>
                <idiom_transJ>朝飯前</idiom_transJ>
        </idiom>
    <IdiomGrp>
    <MetaphorGrp>
        <meta>
                <meta_exp></meta_exp>
                <meta_transE> </meta_transE>
                <meta_transC> </meta_transC>
                <meta_transJ> </meta_transJ>
         </meta>
    </MetaphorGrp>
</entry>
```

2) 부사

```
<entry>         <mntGrp>
                <Kwrt>차재은(01/11/09); 문영선(02/03/11) </Kwrt>
                <Ewrt>이유선(01/11/13)</Ewrt>
                <Cwrt>김세영(01/11/16)</Cwrt>
                <Jwrt>소명인(01/11/16)</Jwrt>
                <note> </note>
    </mntGrp>
     <headGrp>
              <form>다</form>
             <pos>adv</pos>
    </headGrp>
    <sense n=1>
             <eg>다른 사람들도 다 나와 뜻이 같았다. </eg>
             <sem>성분부사/정도부사/수량 </sem>
             <English>
                     <transE>all</transE>
                     <E_xpos> </E_xpos>
                     </English>
             <Chinese>
                     <transC>都</transC>
                     <C_xpos> </C_xpos>
             </Chinese>
             <Japanese>
                     <transJ>みな丨全部丨すべて</transJ>
                     <J_xpos> </J_xpos>
             </Japanese>
```

```
            <ILI>1722113_w</ILI>
            <lr>
                    <syn>모두|전부|모조리</syn>
                    <ant> </ant>
                    <hyper> </hyper>
                    <hypo> </hypo>
                    </lr>
            <coll>
                    <comb_aj>
                            <comb_aj_exp> </comb_aj_exp>
                            <comb_aj_transE> </comb_aj_transE>
                            <comb_aj_transC> </comb_aj_transC>
                            <comb_aj_transJ> </comb_aj_transJ>
                    </comb_aj>
                    <comb_v>
                            <comb_v_exp> </comb_v_exp>
                            <comb_v_transE> </comb_v_transE>
                            <comb_v_transC> </comb_v_transC>
                            <comb_v_transJ> </comb_v_transJ>
                    </comb_v>
                    <comb_n>
                            <comb_n_exp> </comb_n_exp>
                            <comb_n_transE> </comb_n_transE>
                            <comb_n_transC> </comb_n_transC>
                            <comb_n_transJ> </comb_n_transJ>
                    </comb_n>
            </coll>
    </sense>
    <IdiomGrp>
            <idiom>
                    <idiom_exp> </idiom_exp>
                    <idiom_transE> </idiom_transE>
                    <idiom_transC> </idiom_transC>
                    <idiom_transJ> </idiom_transJ>
            </idiom>
    </IdiomGrp>
</entry>
```

3) 동사

```
<entry>
    <mntGrp>
            <Kwrt>차재은(01/11/29); 문영선(02/03/27)</Kwrt>
            <Ewrt>이유선(01/12/05)</Ewrt>
            <Cwrt>김세영(01/12/11)</Cwrt>
            <Jwrt>소명인(01/12/06)</Jwrt>
            <note> </note>
    </mntGrp>
    <headGrp>
            <form>걸치다</form>
```

```
            <pos>pv</pos>
            <V_xpos> </V_xpos>
      </headGrp>
      <sense n=1>
            <eg>철수는 항상 잠바를 어깨에 걸치고 다닌다.</eg>
            <sem> </sem>
            <English>
                  <transE>throw on</transE>
                  <E_xpos> </E_xpos>
            </English>
            <Chinese>
                  <transC>披</transC>
                  <C_xpos> </C_xpos>
             </Chinese>
            <Japanese>
                  <transJ>掛ける</transJ>
                  <J_xpos> </J_xpos>
            </Japanese>
            <ILI> </ILI>
            <lr>
                  <syn>입다</syn>
                  <ant> </ant>
                  <hyper></hyper>
                  <tropo> </tropo>
            </lr>
            <caseFrame>
                  <frame>N0 N2-에 N1-을 V</frame>
                  <selRst>N0=인물 N1=의류(옷|망토|모자)|안경 N2=신체</selRst>
                  <thtRol>N0=AGT N1=THM N2=LOC</thtRol>
            </caseFrame>
      </sense>
      <IdiomGrp>
            <idiom n=1>
                  <Vidiom_exp>양다리를 걸치다</Vidiom_exp>/*여러 애인을 동시에 만나다*/
                  <Vidiom_transE>play a double game</Vidiom_transE>
                  <Vidiom_transC>脚??只船</Vidiom_transC>
                  <Vidiom_transJ>ふたまたを掛ける</Vidiom_transJ>
            </idiom>
      </IdiomGrp>
</entry>
```

4) 형용사

```
<entry>
      <mntGrp>
            <Kwrt>차재은(02/01/23), 문영선(02/04/08)</Kwrt>
            <Ewrt>이유선(02/01/25)</Ewrt>
            <Cwrt>김세영(02/03/22)</Cwrt>
            <Jwrt>소명인(02/03/17)</Jwrt>
```

```
        <note> </note>
        </mntGrp>
        <headGrp>
                <form>크다</form>
                <pos>pa</pos>
                <N_xpos>크기</N_xpos>
        </headGrp>
        <sense n=1>
                <eg>철수는 키가 아주 크다.</eg>
                <sem></sem>
                <English>
                        <transE>big</transE>
                        <E_xpos> </E_xpos>
                </English>
                <Chinese>
                        <transC>高</transC>
                        <C_xpos> </C_xpos>
                 </Chinese>
                <Japanese>
                        <transJ>高い</transJ>
                        <J_xpos> </J_xpos>
                </Japanese>
                <ILI>1052939_w</ILI>
                <lr>
                        <syn></syn>
                        <ant>작다</ant>
                        <hyper> </hyper>
                        <tropo></tropo>
                </lr>
                <caseFrame>
                        <frame>N0 A</frame>
                        <selRst>N0>[Ni-의 Nj] Ni=인물|사물 Nj=(키)</selRst>
                        <thtRol>N0=THM</thtRol>
                </caseFrame>
        </sense>
        <IdiomGrp>
                <idiom>
                        <idiom_exp> </idiom_exp>
                        <idiom_transE> </idiom_transE>
                        <idiom_transC> </idiom_transC>
                        <idiom_transJ> </idiom_transJ>
                </idiom>
        <IdiomGrp>
</entry>
```