碩士學位論文

국어사전의 미시정보를 이용한 동형이의어 구별 방안 연구

2010年

漢城大學校 大學院

文獻情報學科

文獻情報學專攻

梁 景 用

碩 士 學 位 論 文
指導敎授 崔錫斗

국어사전의 미시정보를 이용한 동형이의어 구별 방안 연구

A Study on Distinguishing Homographs Using Micro-information of Korean Dictionary

2010年 6月 日

漢城大學校 大學院
文獻情報學科
文獻情報學專攻
梁 景 用

碩 士 學 位 論 文
指導敎授 崔錫斗

국어사전의 미시정보를 이용한 동형이의어 구별 방안 연구

A Study on Distinguishing Homographs Using Micro-information of Korean Dictionary

위 論文을 圖書館學 碩士學位 論文으로 제출함

2010年 6月 日

漢城大學校 大學院
文獻情報學科
文獻情報學專攻
梁 景 用

梁景用의 圖書館學 碩士學位論文을 認准함

2010年 6月 日

審査委員長 _____印

審 査 委 員 _____印

審 査 委 員 _____印

목 차

표 목 차

국문 초록

언어의 중의성은 정보검색이나 자연언어처리 분야에서 정확률을 저하시키고 시스템의 효율성을 떨어뜨리는 요인으로 작용하고 있다. 언어의 중의성은 동형이의어나 다의어와 같이 단어의 형태적 동일성에서 발생되는데, 한국어는 동형이의어의 비율이 높기 때문에 효율적인 정보처리를 위해서는 동형이의어에 대한 식별이 우선적으로 해결되어야 한다.

본 논문은 지금까지 한국어의 동형이의어 구별 방안들이 가지고 있던 상호 호환성 문제를 해결하기 위하여, 국어사전의 미시정보를 이용하는 방법을 제안하고자 한다. 단어의 속성 정보를 이용하여 동형이의어의 식별 방법을 개발한다면 상호 호환성과 지속성이 보장될 수 있을 것이다.

먼저, 국어사전의 구조와 미시정보에 대한 분석을 통해, 범용적인 동형이의어의 식별자로 사용되기 위한 미시정보의 선정 기준을 다음과 같이 설정하였다. 첫째, 사전 편찬 방향이나 편찬자에 따라 내용이 달라지지 않을 것, 둘째, 대부분의 표제어에 기술될 수 있는 요소일 것, 셋째, 정보의 집필이 쉽고 명료할 것의 세 가지이다. 이 기준에 따라 '원어 정보, 발음 정보, 조어 방법, 품사 정보, 용언의 불규칙 활용 정보, 방언/비표준어의 대응 표준어' 여섯 가지 미시정보를 동형이의어의 식별자로 선정하였다.

다음으로, 여섯 가지 미시정보들이 동형이의어의 식별자로서 활용될 수 있을지 확인하기 위해 『고려대 한국어대사전』을 대상으로 각 미시정보별 식별률을 분석하였다. 식별률 분석 방법으로는, 첫째, 『고려대 한국어대사전』에 수록된 전체 표제어 가운데 여섯 가지 미시정보가 표기된 표제어를 각각 추출하고, 둘째, 미시정보별 표제어 그룹 내에서 동형이의어 관계가 발생되는 표제어를 확인한 다음, 셋째, 미시정보별 표제어 그룹의 동형이의어를 대상으로 해당 미시정보를 통해 구별되는 표제어의 비율을 통해 확인하였다. 여섯 가지 미시정보별 식별률은 방언/비표준어 97.64%, 원어 정보 95.00%, 품사 정보 12.90%, 발음 정보 6.50%, 조어 방법 5.24%, 활용 정보 1.90% 순으로 나타났다. 또한 미시정보의 종합적인 식별률 분석을 위해 여섯 가지 미시정보를 동형이의어의 식별자로 하나씩 추가함으로써 각 미시정보의 식별률 증가 추이도 함께 살펴 보았다. 처음 원어 정보의 식별률은

80.83%로 나타났으며, 이후 발음 정보 82.53%, 품사 정보 87.57%, 방언/비표준어 93.09%, 조어 방법 93.39%, 마지막 활용 정보까지 추가한 최종 식별률은 93.39%로 나타났다.

마지막으로, 『금성판 국어대사전』, 『우리말큰사전』, 『표준국어대사전』, 『고려대 한국어대사전』을 대상으로 각각의 미시정보가 사전마다 어떻게 기술되어 있는지 살펴보고, 이들 미시정보를 식별자로 활용하기 위한 기술 방법을 XML 형식으로 제시하였다. 네 개의 사전들은 위의 여섯 가지 미시정보의 기술에 있어서 대부분 일치하고 있으나, 미시정보에 대한 표기 방법에 있어서는 사전마다 약간의 차이가 있었다. 따라서 이들 미시정보를 식별자로 사용하기 위해서는 사전에서 기술된 미시정보에 대하여 정규화 과정이 필요하였다. 특히 원어, 발음, 품사와 같은 미시정보의 경우 두 개 이상이 나타날 수 있었는데, 이에 대한 처리 방법으로 XML DTD 모델과 XML 데이터를 작성하여 예시하였다.

I. 서론

1. 연구의 목적

20세기 후반 인터넷과 정보통신 기술의 발달과 더불어 생산된 지식과 정보의 양은 이전의 어느 세기와도 비교할 수 없을 만큼 빠른 속도로 증가하고 있다. 지식과 정보의 생산 속도가 빨라지고 정보의 양이 기하급수적으로 확장되면서 단순 키워드나 분류를 통한 정보 검색이 아닌, 좀더 효율적이고 지능적인 정보처리 시스템의 필요성이 절실해지고 있다. 특히 웹의 불규칙한 속성으로 인해, 사용자가 다방면에 분포된 이기종의 디지털 저장소를 동시에 검색하여 원하는 정보를 충분히 얻어내는 것이 점점 더 어려워지고 있다. 이용 가능한 자원들은 널리 분포되어 있음에도 불구하고, 정보 검색과 활용에 어려움이 따르는 직접적인 요인은 시스템 구축에 따른 기술 표준, 색인 작업뿐만 아니라 검색 시스템의 기본 언어 또한 제각각 다르게 개발됨으로써 그것들의 상호운용성을 확보하지 못한 때문이라 할 수 있다. 특히 이종의 용어 사용 및 언어의 중의성 문제는 오래 전부터 정보 검색의 커다란 장해가 되고 있다.[1)]

언어의 중의성은 정확률을 저하시키고 시스템의 효율성을 떨어뜨리는 주요한 요인으로 정보처리 분야에서 이 문제의 해결은 매우 중요한 과제이다. 언어의 중의성은 동형이의어나 다의어와 같이 단어의 형태적 동일성으로 발생되는 것으로서, 동형이의어는 두 개 이상의 단어가 형태는 같으나 의미가 완전히 다른 것을, 다의어는 하나의 단어가 두 개 이상의 의미를 가지는 것을 말한다. 특히 한국어 정보처리에 있어서 우선적으로 해결되어야 할 부분은 동형이의어에 의해 야기되는 중의성 문제이다. 한국어는 동형이의어의 비율이 매우 높기 때문에 정보의 교환과 검색 등의 정보처리를 위해서 동형이의어에 대한 식별이 우선적로 해결되어야 한다. 실제『표준국어대사전』에 대한 분석 결과를 보면, 전체 509,076개의 표제어 중에 22.4%에 해당하는 124,254개가 동형이의어로 나타나고 있다. 특히 '사장', '조사' 등과 같이 30개 이상의 동형이의어를 가지는 단어가 20개나 되며, 심지어

1) "Why is HILT important?." http://hilt.cdlr.strath.ac.uk. 2010. 6. 20.

46개의 동형이의어를 가지는 단어도 있다.[2] 따라서 자연언어에서 동형이의어를 어떻게 식별할 수 있을까의 문제 이전에, 통제어휘 사전의 차원에서 동형이의어를 어떻게 구별할 것인가의 문제가 매우 중요하다. 나아가 두 개 이상의 시소러스나 사전 정보를 자동으로 교환하거나 통합하기 위한 범용적인 식별 표지의 개발이 중요해지고 있다.

본 논문에서는 지금까지 한국어의 동형이의어 구별 방안들이 가지고 있던 상호 호환성 문제를 해결하기 위하여, 언어의 고유한 속성 정보를 이용하는 방법을 제안하고자 한다. 이전까지 사전이나 어휘목록에서 동형이의어의 구별을 위해 사용하던 방법들은 어휘 정보의 교환을 고려하여 개발되지 않았기 때문에 개별 어휘 목록의 범위를 벗어나는 순간 무의미해지거나 심각한 오류를 야기시킬 수 있다. 그러나 이종간의 어휘 정보를 교환할 수 있는 보편적 식별 표지로서 원어, 발음, 품사와 같은 언어의 보편적 속성을 이용하게 되면 어휘 목록의 통합과 상호 참조를 기계적으로 처리할 가능성이 매우 높아질 수 있다.

2. 연구의 범위와 방법

본 연구에서는 대사전 규모의 국어사전을 대상으로 동형이의어 목록을 추출하고, 실제 이들 어휘를 설명하기 위해 기술된 미시정보를 이용하여 동형이의어의 식별률을 분석함으로써, 한국어에 나타나는 동형이의어의 구별을 위한 미시정보의 활용 가능성을 확인하고자 한다.

먼저, 본 연구의 대상으로는 다음의 표 1과 같이 1988년 한글 맞춤법이 개정 고시된 이후 출간된 국어 대사전을 선정하였다. 국어사전은 일반 어휘뿐만 아니라 고유 명사와 전문어까지 폭넓게 수록하고 있어, 현대 한국어의 현황을 가장 잘 보여주고 있는 어휘 목록이라 할 수 있다. 특히, 30만 이상의 어휘를 수록하고 있는 국어 대사전은 시소러스나 전문 용어 사전과 같은 어휘 목록보다 동형이의어의 발생 비율이 높아, 한국어에 나타나는 동형이의어의 현황과 특성을 가장 정확하게 파악할 수 있는 어휘 목록이라 할 수 있다.

---

2) 이운영. 2002.『표준국어대사전 연구 분석』. 서울: 국립국어원. p.30.

표 1 국어 대사전 목록[3)]

사전 이름	편찬자	간행년도	표제어 수
금성판 국어대사전	운평어문연구소	1991	약 40만
우리말큰사전	한글학회	1992	약 45만
표준국어대사전	국립국어원	1999	약 50만
고려대 한국어대사전	고려대 민족문화연구원	2009	약 40만

Ⅱ장에서는 한국어에 나타나는 동형이의어 현황과 한국어 정보 처리시 동형이의어에 의해 발생되는 문제점에 대해 살펴보고자 한다. 한국어의 동형이의어 현황은 『표준국어대사전』과 『고려대 한국어대사전』에 대한 통계 결과를 통해서 구체적으로 확인할 것이며, 동형이의어 대한 계량적인 분석을 통해서 동형이의어의 유형과 발생적 특징에 대해서도 살펴보고자 한다. 또한 동형이의어의 처리시 발생될 수 있는 문제점에 대한 구체적인 사례를 통해서, 범용적인 어휘 식별 방안의 필요성도 확인할 수 있을 것이다.

Ⅲ장에서는 국어사전에서 동형이의어의 구별을 위해 사용하고 있는 어깨번호와 한정어의 사용 방법에 대해 살펴보고, 사전간의 어휘 정보 교환시 이러한 방법이 가지는 문제점을 짚어 보고자 한다. 또한 원어 정보를 이용하여 동형이의어를 구별하고 있는 일부 국어사전의 사례를 통해서, 어휘 식별자가 범용성을 갖기 위해 필요한 요소를 확인할 수 있을 것이다.

Ⅳ장에서는 국어사전의 미시정보를 이용한 동형이의어 식별률 분석하고, 미시정보를 동형이의어의 식별자로 사용하고자 할 때 고려해야 할 문제점에 대해서 살펴보고자 한다. 또한 국어사전에 나타나는 미시정보의 유형과 특징을 네 개의 국어 대사전을 대상으로 비교 분석하여 어휘 정보의 범용적인 식별자로 사용하기 위한 미시정보의 기술 방안을 제시하고자 한다. 1절에서는 국어사전의 구조와 미시정보의 특성에 대해서 살펴보고, 2절에서는 동형이의어의 구별을 위해 사용될 미시정보의 선정 기준과 실제 동형이의어의 식별률 분석을 위한 여섯 가지 미시정보를 제시하였다. 3절에서는 『고려대 한국어대사전』을 대상으로 동형이의어 목록을 추출하고, 2절에서 선정한 미시정보를 이용하여 동형이의어의 식별률을 분석함으로써, 해당 미시정보에 대한 동형이의어 식별 가능성을 확인하고자 한다. 또

3) 도원영, 차준경. 2009. <고려대 한국어대사전>의 종합적 고찰. 『민족문화연구』, 51: 1-54.

한『고려대 한국어대사전』의 전체 표제어를 대상으로, 선정된 미시정보들을 조합한 식별 결과를 분석함으로써 미시정보의 동형이의어 식별 능력을 계량적으로 확인할 것이다. 4절에서는 동형이의어의 식별을 위해 선정한 여섯 가지 미시정보가 국어사전에 어떻게 기술되고 있는지 검토하고, 미시정보를 동형이의어의 식별자로 사용하고자 할 때 고려해야 할 문제점과 해결 방안을 제시하고자 한다. 마지막 절에서는 여섯 가지 미시정보가 가질 수 있는 데이터 구조에 대한 분석과 XML DTD 모형을 개발하고, 실제 국어사전에 나타나는 동형이의어를 XML 데이터 구조로 작성한 예를 제시하였다.

II. 이론적 배경

1. 동형이의어의 현황

국어에서는 단어의 철자나 음운 형식이 동일한 경우를 동음어(同音語), 동형어(同形語), 동철어(同綴語) 등 편의에 따라 다양한 말로 표현하고 있다. 여기에 의미의 동일 여부를 덧붙여 동음이의어(同音異義語), 동형이의어(同形異義語), 동철이의어(同綴異義語)라고도 하는데, 이 경우 동형어와 동철어라는 말에는 이미 '이의(異義)'라는 의미를 내포하고 있다고 할 수 있다. 왜냐하면 형태가 같고 의미까지 동일하면서 다른 말이 있을 수 없기 때문이다.

일반적으로 국어에서 동음이의어는 크게 세 가지 형태로 나타나는데, 첫째, '배(梨)[배], 배(船)[배], 배(腹)[배]'와 같이 철자와 음운 형식이 동일한 경우가 있고, 둘째, '낫[낟], 낮[낟], 낯[낟]', '학문(學問)[항문], 항문(肛門)[항문]'과 같이 철자가 다르지만 음운 형식이 같은 경우가 있다. 셋째, 이와 비슷한 경우로 '눈(雪)[눈 :], 눈(眼)[눈]'과 같이 철자는 동일하지만 음운 형식이 다른 경우가 있는데, 이는 엄밀히 말하면 동음이의어가 아닌 동형어 또는 동철어라 할 수 있다.[4)]

동형어와 동철어는 음운 형식에 상관없이 철자가 동일한 경우를 일컫는 것으로 위의 첫 번째와 세 번째가 이에 해당된다. 동형어와 동철어는 같은 의미로 사용되는 용어로서 동철어보다 동형어가 좀 더 일반적이라 할 수 있으며, 동형어와 동형이의어 또한 같은 의미로 사용되는 단어이나 동형이의어라는 표기가 그 뜻을 보다 분명하게 담고 있다고 할 수 있다. 시소러스나 사전 또는 어휘 데이터베이스 등과 같은 분야에서는 단어의 음운 형식보다 표기 형태가 중요한 정보처리의 요소가 되고 있어 철자가 동일한 경우를 구별할 필요성이 매우 높다. 따라서 본고에서는 위의 세 가지 유형 가운데 첫 번째와 세 번째의 경우를 대상으로 동형이의어라는 용어를 사용하되 경우에 따라 동형어라는 용어도 같이 사용하기로 한다.

우리말에서 동형이의어의 비율은 매우 높게 나타나고 있는데, 이는 『표준국어대사전』에 수록된 동형이의어의 현황을 통해서 확인할 수 있다. 『표준국어대사전』은

4) 유현경. 2000. 사전에서의 동형어 구별을 위한 새로운 제안 : 구분자(distinguisher)의 사용에 대하여. 『사전편찬학연구』, 10: 133-157.

509,076개의 표제어를 수록한 대사전으로서 현재까지 출판된 국어사전 가운데 가장 많은 어휘를 수록하고 있다. 다음의 표 2는 『표준국어대사전』에 대한 동형이의어의 분석 결과로서, 전체 표제어의 22.4%에 해당하는 124,254개가 동형이의어로 나타나고 있다. 특히 '사장', '조사' 등과 같이 30개 이상의 동형이의어를 가지는 단어가 20개나 되고 최대 46개의 동형이의어를 가지는 단어도 있다.

표 2 『표준국어대사전』의 동형이의어 현황[5)]

최종 어깨번호	해당 표제어 개수	전체 표제어 개수	최종 어깨번호	해당 표제어 개수	전체 표제어 개수
02	24,925	49,850	25	5	125
03	7,925	23,775	26	4	104
04	3,582	14,328	27	6	162
05	1,836	9,180	28	2	56
06	1,024	6,144	29	1	29
07	667	4,669	30	2	60
08	426	3,408	31	3	93
09	286	2,574	32	3	96
10	118	1,180	33	4	132
11	147	1,617	34	1	34
12	87	1,044	35	2	70
13	77	1,001	36	0	0
14	71	994	37	1	37
15	57	855	38	1	38
16	36	576	39	0	0
17	25	425	40	0	0
18	23	414	41	0	0
19	17	323	42	0	0
20	18	360	43	0	0
21	10	210	44	1	44
22	14	308	45	1	45
23	2	46	46	1	46
24	6	144	계	41,407	124,254

그런데 위의 『표준국어대사전』을 대상으로 한 동형이의어 통계는 옛말, 북한어를 포함한 통계 결과이다. 따라서 현대 한국어 표제어만을 대상으로 할 경우 동형

5) 이운영. 2002. 전게서. p.30.

이의어 비율은 위의 통계와 다를 수 있다. 실제로 표 3에서 보는 바와 같이 옛말과 북한어를 제외한, 현대 한국어만을 표제어로 수록하고 있는『고려대 한국어대사전』의 분석 결과를 보면, 전체 386,889개의 표제어에서 99,499개의 동형이의어가 나타나고 있다. 이는 사전에 실린 전체 표제어의 25.7%에 해당되며『표준국어대사전』의 동형이의어 비율보다는 3.3% 높은 수치이다.

표 3 『고려대 한국어대사전』의 동형이의어 현황[6)]

최종 어깨번호	해당 표제어 개수	전체 표제어 개수	최종 어깨번호	해당 표제어 개수	전체 표제어 개수
02	20,507	41,014	19	14	266
03	6,289	18,867	20	12	240
04	2,806	11,224	21	11	231
05	1,453	7,265	22	8	176
06	807	4,842	23	5	115
07	503	3,521	24	1	24
08	325	2,600	25	5	125
09	221	1,989	26	5	130
10	159	1,590	27	3	81
11	112	1,232	28	0	0
12	74	888	29	3	87
13	47	611	31	1	31
14	38	532	32	3	96
15	42	630	37	2	74
16	29	464	40	1	40
17	13	221	41	1	41
18	14	252	합계	33,514	99,499

또한 강범모는『표준국어대사전』에 수록된 표제어 가운데 '방언, 옛말, 북한말, 잘못 쓴 말, 어근' 167,600개를 제외한 341,476개의 현대 표준 한국어를 대상으로 분석한 결과 표 4와 같이 동형이의어의 비율이 30.4%에 이른다고 밝히고 있다.[7)]

6) 도원영, 차준경. 2009. 전게서.
7) 강범모. 2005. 동음이의어의 사용 양상.『語學硏究』, 41(1): 1-29.

표 4 현대 표준 한국어의 동음어 비율[8)]

품사 구분	품사 구분 안함	품사 구분 함
단어형	207,517	275,455
전체 단어	339,249	339,249
동음어	103,080	96,343
동음어 비율	0.304	0.284

이와 같이 우리말에 동형이의어의 비율이 높게 나타나는 이유는 전체 어휘 가운데 2음절어 한자어의 비율이 매우 높기 때문이다. 다음에 보이는 표 5는 『고려대 한국어대사전』의 동형이의어를 대상으로 표제어의 원어 유형을 분석한 것으로, 전체 동형이의어 99,499개 가운데 한자어가 65.11%로 가장 많고 고유어가 18.57%, 고유어+한자어가 14.57%로 나타나고 있다. 한자어가 포함된 경우까지를 모두 포함할 경우에는 전체 표제어의 80% 정도가 한자어나 한자어와 결합된 어휘임을 알 수 있다.

표 5 동형이의어의 원어 유형

표제어 유형	동형이의어 수	동형이의어 비율(%)
고유어	18,476	18.57
한자어	64,784	65.11
외래어	1,671	1.68
고유어+한자어	14,499	14.57
고유어+외래어	17	0.02
한자어+외래어	48	0.05
고유어+한자어+외래어	4	0.00
합계	99,499	100.00

그리고 표 6에서와 같이 동형이의어의 음절수를 보면 2음절어가 71.03%로 매우 높은 비율을 보이고 있어, 한국어의 동형이의어 가운데 2음절어 한자어의 비중이 가장 높음을 알 수 있다.

8) 이 표의 내용은 논문에 있는 그대로를 인용한 것이다. 논문에서는 표의 제목과 품사 구분에 나타나는 '동음어'의 범위에 대해서 밝히고 있지 않으나, 전체적인 내용을 살펴보면 동형이의어를 의미하고 있다.

표 6 동형이의어의 표제어 음절 통계

표제어 음절수	동형이의어 수	동형이의어 비율(%)
1음절	4,803	4.83
2음절	70,676	71.03
3음절	7,804	7.84
4음절	15,083	15.16
5음절	572	0.57
6음절	541	0.54
8음절	20	0.02
합계	99,499	100.00

2. 동형이의어 처리의 문제점

앞에서 살펴본 바와 같이 우리말은 동형이의어의 비율이 매우 높게 나타나고 있는데, 단어의 형태적인 동일성으로 인해 발생되는 의미적 중의성은 형태소 분석이나 구문 분석과 같은 자연언어처리 분야뿐만 아니라, 정보검색이나 기계번역과 같은 응용분야에서도 문제가 되고 있다.

기계번역이나 형태소 분석, 구문 분석 등의 분야에서는 하나의 동형이의어로 인해 문장 전체를 오역하거나 잘못된 분석 결과를 초래하기도 하고, 정보의 색인과 검색 등에서는 잡음률을 증가시키는 주요한 원인이 되기도 하는데, 정보의 양이 많아질수록 이러한 문제는 더욱 커지게 된다. 특히, 정보 처리 시스템에서 사용되는 시소러스나 기계 가독형 사전(MRD)에서의 동형이의어 문제는 시스템 전체의 성능에 영향을 미칠 뿐만 아니라, 이종 시스템간의 정보 교환이나 통합 검색에 있어서도 큰 걸림돌이 되고 있다.

다음은 <21세기 세종계획>[9]의 '형태의미분석 말뭉치'에서 발생되는 동형이의어 처리의 문제점을 보인 것이다. 전체 1,250만 어절 규모로 구축된 '형태의미분석 말뭉치'는 품사뿐만 아니라, 동형이의어에 대한 의미번호까지 분석한 코퍼스로서, 어

9) 국어 정보화 기반 구축을 위해 문화관광부가 국립국어원 및 관련 학계와 더불어 지난 1998년부터 2007년까지 10년간 추진해 온 사업으로, 한국어 코퍼스 구축, 전자사전 개발, 문자코드 표준화 연구, 글꼴 개발 등의 과제가 수행되었다.

휘 빈도 통계, 사전 편찬을 위한 용례 추출, 형태소 분석기를 위한 학습 사전 등으로 다양하게 이용되고 있다. '형태의미분석 말뭉치'에서 동형이의어의 구분은 '배판__02/NNG', '사전__22/NNG'와 같이 의미번호를 부착하여 구분하고 있는데, 동형이의어를 구분하기 위한 의미번호는 『표준국어대사전』(1999)의 어깨번호를 기준으로 하고 있다는 특징이 있다.

그러나 동형이의어를 구별하기 위해 붙여진 의미번호는 코퍼스 자체로서 독립적이지 못하고, 의미번호를 인용한 사전에 의존할 수밖에 없는 문제점을 안고 있다. 만약 의미번호를 인용한 사전이 개정이나 증보되는 과정에서 어깨번호에 변동이 발생될 경우에는 코퍼스 의미번호에도 커다란 혼란이 생기게 되는 것이다.

<21세기 세종계획 형태의미분석 말뭉치>

4BS_0640198340	4·6배판에	4/SN + ·/SP + 6/SN + 배판__02/NNG + 에/JKB
4BS_0640198350	굵직한	굵직/XR + 하/XSA + ㄴ/ETM
4BS_0640198360	활자로	활자__01/NNG + 로/JKB
4BS_0640198370	된	되__01/VV + ㄴ/ETM
4BS_0640198380	이	이/MM
4BS_0640198390	무거운	무겁/VA + ㄴ/ETM
4BS_0640198400	사전을	사전__22/NNG + 을/JKO
4BS_0640198410	안고	안/VV + 고/EC
4BS_0640198420	사람들은	사람/NNG + 들/XSN + 은/JX
4BS_0640198430	잠	잠__01/NNG
4BS_0640198440	못	못/MAG
4BS_0640198450	이루어	이루__01/VV + 어/EC
4BS_0640198460	했다.	하/VX + 았/EP + 다/EF + ./SF

실제, '형태의미분석 말뭉치'의 '배판__02/NNG'는 '책의 인쇄 규격'을 의미하는 단어로 『표준국어대사전』(1999)의 어깨번호 '2'를 부착하고 있지만, 이후 국립국어원에서 발간한 『표준국어대사전』의 전자판을 보면 어깨번호가 '1'로 바뀌어 있어, 코퍼스 의미번호 체계와 일치되지 않는 문제가 있다. 뿐만 아니라 『표준국어대사전』(1999)의 '배판[1]'이 전자판 사전에서 삭제됨으로써 '배판'의 모든 동형이의어는 종이책 사전과 어깨번호가 일치하지 않게 되었다.

<『표준국어대사전』(1999)>

배판[1](背板)[배 : -] 명«동물» 곤충의 등의 판판한 부분. '뒤판'으로 순화.

배판[2](倍版)[배 : -] 명«출판» 일정한 책의 규격을 나타내는 판의 갑절 크기의 인쇄물.

배판[3](排判) 명 별러서 차림.

배판-되다(排判--)[--되-/--뒈-] 동 ⇨배판[3].

배판-하다(排判--) 동 〖…을〗 ⇨배판[3].

배판[4](褙板)[배 : -] 명 종이나 천 따위를 여러 겹으로 포개어 붙일 때, 바닥에 깔고 쓰는 널판.

<『표준국어대사전』(전자판, http://www.korean.go.kr)>

배판[1](倍版)[배 : -] 「명사」『출판』 일정한 책의 규격을 나타내는 판의 갑절 크기의 인쇄물.

배판[2](排判) 「명사」 별러서 차림.

배판-되다(排判--)[--되-/--뒈-] 「동사」 별러져서 차려지다.

배판-하다(排判--) 「동사」 〖…을〗 별러서 차리다.

배판[3](褙板)[배 : -] 「명사」 종이나 천 따위를 여러 겹으로 포개어 붙일 때, 바닥에 깔고 쓰는 널판.

이러한 경우 문제를 해결하기 위해서는 코퍼스의 의미번호를 개정된 사전의 어깨번호로 바꾸거나, 아니면 최초로 인용했던 사전만을 계속해서 사용해야 하는데, 이 두 가지 방법 모두 다음과 같은 이유로 적절한 해결책은 되기는 어렵다. 첫째, 개정되는 사전의 어깨번호 체계로 코퍼스의 의미번호를 계속 갱신하는 방법은 코퍼스의 데이터 규모가 작거나 사전의 개정이 극히 이례적이라면 혹시 가능할 수도 있지만, 개정된 사전의 어깨번호를 모두 확인하는 것도 쉽지 않은 일이며, 국립국어원의 전자판『표준국어대사전』과 같이 비정기적으로 사전의 내용이 바뀌는 경우에는 더더욱 불가능한 일이다. 둘째, 처음 인용했던 사전의 어깨번호 체계를 계속 사용하는 방법은, 더 이상 코퍼스의 확장이 없거나 기준이 되는 사전의 출판이 계속 안정적으로 유지될 때나 가능할 수 있다. 하지만 사전이 개정되지 않고 오랫동안 출판되는 경우는 극히 이례적이며, 코퍼스의 통계 결과들을 이용할 때에도 매번 동일한 사전만을 참조해야 하는 어려움이 있다. 따라서 어휘 사전의 의미 정보를 구분하는 방법에 있어서 특정 사전에 의존적이지 않는, 보편적 식별 방법의 개발이 필요하다.

3. 선행 연구

단어의 중의성 문제는 기계번역과 같은 자연언어처리 분야에서 중요하게 다뤄지는 문제로서, 1940년대 후반부터 이를 해결하기 위한 다양한 연구들이 있다. 지금까지 단어의 중의성 해소를 위해 사용했던 방법들로는 크게 수작업을 통해 개발한 규칙에 기반하거나, 코퍼스를 이용하는 방법, 또는 이 두 가지를 결합하는 방법 등이 있었다.[10] 하지만 자연언어처리 분야에서 단어의 중의성 문제는 문장에서 발생되는 다의어나 동형이의어의 의미 분별에 목적을 둔 것으로, 어휘 자체의 식별이나 어휘 정보의 교환을 위한 것은 아니다.

단어 차원에서의 동형이의어 식별 문제는 주로 색인이나 정보 검색을 위해 사용되는 주제명표목표나 시소러스와 같은 분야에서 논의되던 것으로, 오래 전부터 이들 통제어휘사전에서는 한정어를 사용하여 동형이의어를 구별하여 왔다.[11] 또한 언어 사전에서도 한정어를 사용하여 동형이의어의 중의성 문제를 해결하고 있는 사전이 나오면서, 언어 사전에서의 한정어 활용 방안에 대한 연구가 발표되고 있다. 언어 사전에서 한정어를 사용하여 동형이의어를 구별하고 있는 사전으로서 대표적인 것은 *Cambridge International Dictionary of English*[12]를 들 수 있으며, 국어사전으로는『연세 초등국어사전』[13]과『보리 국어사전』[14]이 있다.

국어사전에서의 한정어 사용 방법에 관한 논의는 김태수와 최석두에서 언급되기 시작했으며, 이후 유현경, 이희자, 박수연 등의 연구가 있다.

김태수・최석두는 이전까지 주제명표목표나 시소러스와 같은 통제어휘시스템에서 다의성이 있는 어휘를 색인할 때 한글, 한자, 영문을 섞어 한정어를 붙여 중의성을 해소하고 있으나, 여러 문자를 섞어 쓰는 것은 일관성이 없다고 지적하며, 한글 한정어의 사용을 제안하고 있다. 아울러 국어사전에서는 기존에 사용하는 어깨번호 대신, 품사, 표준말, 발음 등으로 구별할 수 있지만 이들의 내용과 순서가

---

10) Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation : Algorithms and Applications*. U.K.: Springer. pp.12-13.

11) 김태수, 최석두. 1997. 동형이의어의 구별을 위한 한글한정어 사용에 관한 연구.『情報管理學會誌』, 14(1): 107-124.

12) Cambridge University Press. 1995. *Cambridge international dictionary of English*. Cambridge; New York : Cambridge University Press.

13) 연세대학교 언어정보개발원 편. 2001.『연세 초등국어사전』. 서울: 두산동아.

14) 토박이 사전 편찬식 엮음. 2008.『보리 국어사전』. 서울: 보리.

달라 컴퓨터 처리에 적합지 않음을 지적하며, 국어사전에서도 동형이의어의 구별을 위해 한글 한정어를 사용할 수 있음을 밝히고 있다.[15]

유현경은 기존의 국어사전에서 사용하는 어깨번호는 동형어를 이루는 낱말들이 별개의 단어라는 것 이외에 더 이상의 정보를 주지 못한다고 지적하며, 이를 대신할 방법으로 구분자(distinguisher)의 사용을 제안하고 있다. 구분자는 표제어의 가장 특징적인 속성 중의 하나를 메타언어(metalanguage)의 형태로 표시한 것으로, 사용자가 구분자만으로도 자신이 원하는 정보를 쉽게 알 수 있는 장점이 있음을 밝히고 있다.[16]

이희자는 정보 검색 분야에서 동형이의어를 처리하기 위해 사용되는 한정어의 원리를 일반 언어 사전에 확대 적용시킬 수 있음을 지적하면서, 표제어를 가장 잘 나타내는 말을 붙여 동형이의어를 구별하는 길잡이말(guide words)의 사용을 제안하고 있다. 아울러 길잡이말이 사전의 이용자에게 표제어에 대한 정보를 직접적으로 제공할 수 있는 장점을 강조하면서, 길잡이말의 선정 원칙을 개략적으로 제시하고 있다.[17]

박수연은 외국인을 위한 한국어 학습사전의 측면에서 기존의 길잡이말을 사용하여 편찬된 사전들의 문제점을 지적하고, 개선된 길잡이말의 사용 원칙을 제안하고 있다. 기존의 길잡이말의 문제점으로는, 첫째, 길잡이말이 표제어보다 어려운 경우가 있고, 둘째, 길잡이말이 동형이의어인 경우도 있으며, 셋째, 표제어를 통해 길잡이말이 반응어로 연상되지 않는 경우, 넷째, 길잡이말의 통일성 부족을 지적하고 있다.[18]

그러나 이 같이 다양한 연구들이 제시하는 것처럼 동형이의어의 구별을 위해 한정어를 부여하는 방법에는 형식적인 일관성과 논리성을 유지하기 어려운 문제가 있다. 따라서 이종의 어휘 목록이나 사전의 정보를 교환하기 위해서는 목록 개발자나 사전 집필자의 자의성이 개입될 여지가 없는 단어의 고유한 속성 정보를 사용할 필요가 있는데, 이러한 측면에서 Stevenson과 Wilks의 연구가 주목할 만

15) 김태수, 최석두. 1997. 전게서.
16) 유현경. 2000. 전게서.
17) 이희자. 2001. 개별어 사전 편찬: 동음이의어의 구별을 위한 길잡이말(Guide Words) 연구.『제2차 아시아 사전학회 국제 학술대회』, 167-201.
18) 박수연. 2003. 외국인을 위한 '한국어 학습사전'에서 동음이의어의 구별에 관한 연구 : 길잡이말을 중심으로.『외국어로서의 한국어교육』, 28: 71-110.

하다.

Stevenson과 Wilks는 영어를 대상으로 단어의 의미 중의성 해결(Word Sense Disambiguation)을 위해 사전의 품사 정보를 이용하는 방법을 사용하였다. Stevenson과 Wilks에 따르면, 대략 36,000여 어휘를 수록한 롱맨 영어 사전(LDOCE)을 기반으로 품사에 따른 동형이의어 식별률을 분석한 결과 88%의 어휘가 형태적 중의성이 해결된다고 밝히고 있다. 여기에 3개 이상의 동형이의어를 가지는 어휘에서 일부 단어 만이 품사를 통해서 형태적 중의성이 해결되는 경우는 전체 동형이의어의 7%인데, 이러한 경우를 포함하면 롱맨 영어 사전에서 품사를 통한 동형이의어의 식별률은 95%로 나타난다고 밝히고 있다.[19] 롱맨 영어 사전이 학습자용 사전으로 그 규모가 크지 않기 때문에, 대사전에 적용할 경우의 식별률과 비교할 수는 없다. 그러나 단어의 형태나 의미적인 속성을 통해서 동형이의어를 구별하는 방법은 이종의 사전에서도 호환성이 유지될 수 있다.

19) Mark Stevenson and Yorick Wilks. 2001. “The Interaction of Knowledge Sources in Word Sense Disambiguation.” *Computational Linguistics*, 27(3): 321 - 349.

III. 국어사전에서의 동형이의어 처리 현황

1. 처리 현황

언어 사전은 단어, 문법 형태소, 명사 구 등을 모두 포함한 단어 집합으로서 전문어 사전이나 시소러스 등과 같은 특수 분야의 어휘 목록보다 동형이의어의 발생 비율이 높다. 따라서 대부분의 어휘 사전에서는 동형이의어를 구별하기 위해 어깨번호나 한정어를 붙이는 방법이 사용되고 있다. 이 장에서는 기존 국어사전에서 사용하는 동형이의어의 처리 방법에 대해 살펴보고, 국어사전이나 어휘 목록간의 정보처리시 기존의 구별 방법이 가지는 문제점을 살펴보고자 한다.

1.1 어깨번호

어깨번호는 개별 어휘 사전에 나타나는 동형이의어를 구별하기 위해 표제어 다음에 일련의 숫자를 붙이는 방법으로, 동형이의어의 구별 기능과 함께 표제어의 배열 순서를 정하는 방법으로 많이 사용되고 있다. 국어사전에서 어깨번호를 붙이는 방법은 크게 두 가지로 나눌 수 있다. 첫째, 표제어의 한글 표기가 같으면 모두 어깨번호를 붙이는 방법이 있고, 둘째, 한글 표기가 같더라도 원어의 한자나 로마자가 다르면 어깨번호를 붙이지 않는 방법이 있다. 『표준국어대사전』과 『고려대 한국어대사전』은 첫 번째 방법을 기준으로 어깨번호를 붙이고, 『금성판 국어대사전』은 두 번째 방법을 따르고 있다.

1.1.1 어깨번호 부여 기준

『금성판 국어대사전』에서는 고유어와 한자어, 접사와 같이 붙임줄이 있는 의존형태소를 이형으로 구분하여 어깨번호를 별도로 붙이고 있다. 한자어의 경우 한자가 동형이의어의 구분 역할을 하는 것으로 보고 한자가 같은 경우에만 어깨번호를 붙이고 있으며, 붙임줄이 있는 의존형태소의 경우 붙임줄이 앞에 있는가, 뒤에 있는가, 또는 앞뒤에 있는가에 따라 어깨번호를 달리 붙이고 있다.

가[1] 「명」<음> 음계의 여섯째 음의 이름. 곧, 라(la).
가 : [2] 「명」 ①사물의 바깥쪽 경계가 되는 부분.
가[3] 「조」 ((주로, 모음으로 끝나는 체언에 붙어)) ①그 말을 주격(主格)이 되게 하는 격조사.
가[4] 「조」<옛> 인가. '르' 이외의 자음으로 끝나는 체언에 붙어, 판정 의문문에서 쓰이는 의문 보조사.
가 : [5] '가아'(가다의 어간 '가'와 활용 어미 '아')가 준말.
-가- 「어미」(선어말) <옛> 선어말 어미 '-거-'와 '-우-'가 결합된 형태.
가 : [可] 「명」 ①옳거나 좋음. ②찬성하는 의사의 표시.
가[1] [加] 「명」 더하거나 합하는 일.
가[2] [加] 「명」<역> 고대 부여나 고구려에서 족장이나 고관을 일컫던 칭호.
가 [家] 「명」<법> 호적상 1가(家)로 등록된 친족(親族) 단체.
가 [斝] 「명」 옛날, 예식 때 사용하던 술잔.
가 : [賈] 「명」 우리 나라의 성(姓)의 하나. 본관은 소주(蘇州) 등이 현존함.
가 : - [假] 「접두」 ①'임시적인', '시험적인'의 뜻을 나타내는 말.
-가 [家] 「접미」 ①어떤 방면의 전문인. 또는, 그것을 직업으로 삼는 사람.
-가 [哥] 「접미」 ①성(姓)에 붙이어, 그 성임을 나타내는 말.
-가 [街] 「접미」 ①큰 도시의 '동(洞)'이나 '로(路)'를 작게 나눈 구획.
-가 [歌] 「접미」 노래의 이름이나 종류를 나타내는 말.
-가 [價] 「접미」 ①((어떤 명사 아래에 붙어)) '값'이라는 뜻을 나타내는 말.

『우리말큰사전』은 표제어가 자립형태소인가 의존형태소인가에 따라 어깨번호를 달리 표기하고 있다. 자립형태소의 경우에는 고유어나 한자어, 또는 외래어의 구분 없이 어깨번호를 붙이고, 의존형태소의 경우에는 『금성판 국어대사전』과 같이 붙임줄의 위치에 따라 어깨번호를 달리 표기하고 있다.

가[1] (이)<악> 서양음악의 칠음 체계에서 여섯째 음이름.
가 : [2] (이) ①어떤 바닥이 끝진 부분.
가[3] (이)<역> 부여와 고구려에서 족장이나 고관을 일컫던 말. 〔加〕
가[4] (이) =칼[2]. 〔枷〕
가[5] (이)<법> 한 호적에 올라 있는 사람의 단체. 호주와 가족으로 구성된다. 〔家〕
가[6] (이)<악> =날라리[2]. 〔笳〕
가[7] (이) 예전 제례에 쓰인 구리로 만든 술잔. 〔斝〕
가 : [8] (이) ①무던히 옳거나 좋음. 〔可〕
가 : [9] (이) 성의 하나. 〔賈〕
가 : [10] (이) 성의 하나. 〔價〕

가[11] (대) →걔.
가[12] (토) 받침 없는 임자씨 따위에 붙어, ①그 말을 임자자리로 되게 하는 토.
가[13] '가아'의 준말.
가[14] →가지고.
가 : -[1] (줄) '갈다'의 벗어난줄기.
가-[2] (앞) '덧-'의 뜻. 〔加〕
가 : -[3] (앞) ①'임시적', '정식이 아닌'의 뜻. 〔假〕
-가[1] (뒤) 그 '성'임을 나타내거나, 또는 그 성을 가진 사람을 좀 낮잡아 일컫는 뜻. 〔哥〕
-가[2] (뒤) ①'그 일에 전문적으로 종사하는 사람'의 뜻. 〔家〕
-가[3] (뒤) ①한길을 낀 큰 동리를 몇으로 나눈 한 구역. 〔街〕
-가[4] (뒤) '노래'의 뜻. 〔歌〕
-가[5] (뒤) '값'의 뜻. 〔價〕
-가[6] (끝) →-ㄴ가.

『표준국어대사전』의 경우에는 한글 표기가 같은 모든 표제어를 동형이의어로 보고 어깨번호를 부여하고 있는데, 접사와 같이 붙임줄이 있는 의존형태소도 붙임줄이 없는 자립형태소와 함께 동형이의어로 처리하고 있다.

가[1] [가 :] 「명」 ①경계에 가까운 바깥쪽 부분.
가[2] 「명」<음> 서양 음악의 칠음 체계에서, 여섯 번째 음이름.
가[3](加) '가하다[1]'의 어근.
가[4](可) [가 :] 「명」 ①옳거나 좋음.
가[5](加) 「명」<역> 부여와 고구려에서, 족장이나 고관을 이르던 말.
가[6](枷) 「명」<역> =칼[2].
가[7](家) 「명」<법> 같은 호적에 들어 있는 친족 집단.
가[8](笳) 「명」<음> 짐승의 뿔로 만든 원시적인 악기.
가[9](斝) 「명」<역> 제례 때에 쓰던 술잔.
가[10](賈) 「명」 우리나라 성(姓)의 하나. 본관은 소주(蘇州) 하나뿐이다.
가[11] 「조」 (받침 없는 체언 뒤에 붙어) ①어떤 상태나 상황에 놓인 대상, 또는 상태나 상황을 겪거나 일정한 동작을 하는 주체를 나타내는 격 조사.
-가[12](哥) 「접사」 (인명의 성(姓)을 나타내는 대다수 명사 뒤에 붙어) 그 성씨 자체 또는 그 성씨를 가진 사람 의 뜻을 더하는 접미사.
-가[13](家) 「접사」 (일부 명사 뒤에 붙어) ①그것을 전문적으로 하는 사람 또는 그것을 직업으로 하는 사람 의 뜻을 더하는 접미사.
-가[14](家) 「접사」 (고유 명사를 포함한 일부 명사 뒤에 붙어) 가문의 뜻을 더하는 접미사.

가-[15](假) 「접사」 (일부 명사 앞에 붙어) 가짜, 거짓 또는 임시적인 의 뜻을 더하는 접두사.
-가[16](街) 「접사」 (일부 명사 또는 수사 뒤에 붙어) 거리 또는 지역 의 뜻을 더하는 접미사.
-가[17](歌) 「접사」 (일부 명사 뒤에 붙어) 노래 의 뜻을 더하는 접미사.
-가[18](價) 「접사」 ①일부 명사 뒤에 붙어) 값 의 뜻을 더하는 접미사.
가[19] 「조」<옛> (모음 ㅣ 나 ㄹ 받침으로 끝나지 않는 체언류 뒤에 붙어) (의문사가 없는 의문문에 쓰여) 인가.
가[20] 「조」<옛> (모음 ㅣ 나 ·ㅣ 등으로 끝나는 체언류 뒤에 붙어) 가.
-가[21] 「어미」<옛> (ㅣ 계열 이중 모음이나 ㄹ 받침으로 끝나지 않는 어간 뒤에 붙어) (주로 자동사, 형용사 어간 뒤에 붙어) (다른 어미 앞에 붙어) (주로 1인칭 주어와 함께 쓰여) 서술어가 나타내는 동작이나 상태가 확정되거나 완료됨을 나타내는 어미.
-가[22] 「어미」<옛> (어미 -ㄴ, -ㄹ 뒤에 붙어) (의문사가 없는 의문문에 쓰여) 물음을 나타내는 종결 어미.

『고려대 한국어대사전』은 『표준국어대사전』과 같은 기준으로 어깨번호를 부여하고 있다.

가[1] [가 :] 명 ①바깥쪽 경계가 되는 가장자리 부분이나 그 부근.
가[2] 「명」<음악> 음이름의 하나. 다장조에서 계이름 '라'에 해당한다.
가[3] 〔家〕 「명」<법률> 이전에, 같은 호적에 올라 있는 친족 집단을 이르던 말.
가[4] 〔可〕 [가 :] 「명」 ①옳거나 좋음.
가[5] 〔笳〕 「명」<음악> 짐승의 뿔로 만든 고대 악기의 한 가지.
가[6] 〔加〕 「명」<역사> 부여와 고구려 때, 족장이나 고관을 이르던 말.
가[7] 〔斝〕 「명」<역사> 예전에, 제례 때 쓰는, 청동으로 만든 술잔을 이르던 말.
가[8] 〔枷〕 「명」<역사> 예전에, 죄인의 목에 씌우던 형틀의 하나를 이르던 말.
가[9] 〔賈〕 「명」 우리나라 성(姓)의 하나. 본관은 소주(蘇州) 하나뿐이다.
가[10] 「조사」 ①모음으로 끝나는 체언의 뒤에 붙어, 행위의 주체임을 나타내는 주격 조사.
-가[11] 〔-街〕 「접미」 ①행정 구역이나 거리를 세분하여 숫자로 구별한 말 뒤에 붙어, 그 행정 구역이나 거리를 작게 나눈 구역이라는 뜻을 더하는 말.
가-[12] 〔=假〕 「접두」 일부 명사 앞에 붙어, '임시의'의 뜻을 더하는 말.
-가[13] 〔-家〕 「접미」 일부 명사나 고유 명사 뒤에 붙어, 가문이나 집안의 뜻을 더하여 명사를 만드는 말.
-가[14] 〔-價〕 [까] 「접미」 ①<역사> 일부 숫자 뒤에 붙어, '그것에 결합한 원자의 수 혹은 값'의 뜻을 더하여 명사를 만드는 말.
-가[15] 〔-歌〕 「접미」 일부 명사 뒤에 붙어, '그것과 관련한 노래'의 뜻을 더하여 명사를 만드는 말.

-가[16] 〔-哥〕 「접미」 ①성(姓)의 뒤에 붙어, 그 성(姓)을 얕잡는 뜻을 더하는 말.
-가[17] 〔-家〕 「접미」 ①전문 직종을 나타내는 일부 명사 뒤에 붙어, '그것을 전문적으로 하거나 직업으로 하는 사람'의 뜻을 더하여 명사를 만드는 말.

1.1.2 어깨번호 배열 방법

표제어의 어깨번호 순서를 정하는 데에 있어서는 사전마다 큰 차이를 보이는데, 이는 사전마다 배열의 우선 순위를 다르게 정하고 있기 때문이다. 일반적으로 동형이의어를 배열하는 데에는 고유어의 여부, 품사, 발음의 길이, 표준어 여부, 한자의 부수/획수 등이 중요하게 고려되는데, 어떤 요소를 우선적으로 적용하느냐에 따라 표제어의 배열은 크게 달라지게 된다.

『금성판 국어대사전』에서는 주표제어와 부표제어로 구분하여 표제어를 배열하고 있는데, 주표제어에서 동형이의어를 배열할 때 가장 먼저 고려하는 사항은 고유어인지, 한자어인지, 외래어인지에 대한 판단이다. 이후에는 품사의 순서대로 배열하고, 위의 조건이 같은 경우 발음의 길이가 짧은 순으로 하였다. 또한 현대어를 고어보다 먼저 배열하되 현대어는 표준어를 방언보다 앞에 놓고, 한자어는 획수가 적은 것부터 싣고 있다.

① 어원 : 고유어, 한자어, 외래어 순
② 품사 : 명사, 대명사, 수사, 동사, 형용사, 관형사, 부사, 감탄사, 조사, 접두사, 접미사, 어미 순
③ 발음 : 짧은소리, 긴소리 순
④ 현대어 : 표준어, 방언 순
⑤ 고어
⑥ 한자어 : 획수 순

『금성판 국어대사전』에서 부표제어는 주표제어의 아래에 두고 있는데, 주표제어가 동형이의어일 경우 다음과 같이 주표제어의 어깨번호를 따라서 붙인다.

골 : **-골**[1] 「부」 병이 잦거나 오래 되어 늘 몸이 약한 모양.
골 : **골-하다**[1] 「동」(자) 「여불」
골 : **-골**[2] 「부」 암탉이 알겯는 소리.

골 : 골-하다[2] 「동」(자) 「여불」

『우리말큰사전』은 동형이의어를 배열하는 데 있어서 품사의 순서를 가장 먼저 고려하고 있다. 다음으로 표준말을 비표준말보다 먼저 싣고, 이후 토박이말, 한자어, 외래어 순으로 하되, 위의 사항이 같은 경우 발음의 길이, 한자의 획수, 발음란의 유무 등을 고려하여 배열하고 있다.

① 품사 : 이름씨, 대이름씨, 셈씨, 임직씨, 그림씨, 잡음씨, 매김씨, 어찌씨, 느낌씨, 토씨, 줄기, 뿌리, 앞가지, 뒷가지, 씨끝 순
② 표준말 여부 : 표준말, 비표준말(속된말, 낮은말, 방언, 변말, 곁말, 잘못 쓰는 말) 순
③ 고유어 여부 : 토박이말, 한자어, 외래어 순
④ 위의 사항이 같은 경우 : 짧은소리 , 긴소리 순 / 한자의 획수 순 / 발음란의 유무 순 / 일반어, 전문어 순
⑤ 기본 올림말, 이은말/마디 순

『표준국어대사전』은 『금성판 국어대사전』과 같이 주표제어와 부표제어로 구분하여 배열하되 일부 표제어의 경우 주표제어와 부표제어의 어깨번호가 혼합되어 있다. 먼저 주표제어의 배열에 대해 살펴보면, 현대어를 옛말보다 앞에 싣되, 어휘 형태를 문법 형태보다 먼저 배열한다. 이후 단어의 어원에 따라 고유어, 한자어, 외래어 순으로 싣고, 어원이 같은 경우 표준어, 북한어, 방언, 비표준어 순으로 싣는다. 다음으로 품사의 순서에 따라 배열하고, 품사가 같은 경우에는 일반어를 전문어 보다 먼저 둔다.

① 현대어, 옛말 순
② 어휘 형태, 문법 형태 순
③ 고유어, 한자어, 외래어 순
④ 표준어, 북한어, 방언, 비표준어 순
⑤ 품사
  - 어휘 형태 : 명사, 대명사, 수사, 동사, 형용사, 보조 용언, 관형사, 부사, 감탄사, 어근 순
  - 문법 형태 : 조사, 어미, 접사 순
⑥ 일반어, 전문어 순

『표준국어대사전』에서도 부표제어의 배열은 기본적으로 주표제어 아래에 싣되, 부표제어에 동형이의어가 있는 경우 주표제어의 순서에 따라 어깨번호를 붙인다. 다만, 주표제어와 부표제어가 동형이의어일 경우에는 먼저 실리는 순서대로 어깨번호를 붙이고 있다.

수[26](數) [수 :] 「I」「명」「1」 셀 수 있는 사물의 크기를 나타내는 값.
수-적[1](—的)[수 : 쩍] 「관」「명」 수를 기준으로 하는. 또는 그런 것.
수적[2](手迹) 〔수적만[-정-]〕 「명」 손수 쓴 글씨나 그린 그림.
수적[3](水賊) 〔수적만[-정-]〕 「명」 바다나 큰 강에서 남의 재물을 강제로 빼앗아 가는 도둑.

『고려대 한국어대사전』에서는 동형이의어의 배열에 있어 표제어의 자립성 여부를 중요하게 고려하여 어휘 형태를 문법 형태보다 앞에 배열한다. 그 다음으로 품사의 순서에 따라 배열하고, 표준어 여부를 판단하여 표준어, 비표준어, 방언 순으로 싣는다. 위의 조건이 같은 경우에는 어원을 고려하여 고유어, 한자어, 외래어 순으로 배열하고, 이후 어휘의 빈도와 조어 방법에 따라 배열하고 있다.

① 어휘 형태, 문법 형태 순
② 품사 : 명사, 대명사, 수사, 동사, 형용사, 관형사, 부사, 감탄사, 조사, 접사, 어미 순
③ 무품사어
④ 표준어, 비표준어, 방언 순
⑤ 고유어, 한자어, 외래어 순
⑥ 어휘 빈도 순
⑦ 단일어, 복합어 순

지금까지 네 개의 국어사전을 대상으로 동형이의어의 배열 순서를 살펴보았는데, 각 사전별 배열 순서를 정리하면 다음 표 7과 같다. 각 사전에서는 1단계의 순서대로 동형이의어를 배열하되, 위의 조건이 같은 경우 2단계의 요소를 고려하여 순서를 정하고 있다.

표 7 국어사전의 동형이의어 배열 순서

단계	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
1단계	1. 어원 고유어 한자어 외래어 2. 품사 명사 대명사 수사 동사 형용사 관형사 부사 감탄사 조사 접두사 접미사 어미 구	1. 품사 이름씨 대이름씨 셈씨 임직씨 그림씨 잡음씨 매김씨 어찌씨 느낌씨 토씨 줄기 뿌리 앞가지 뒷가지 씨끝 2. 표준말 여부 표준말 비표준말 3. 토박이말 여부 토박이말 한자말 들온말	1. 현대어 여부 현대어 옛말 2. 품사[어휘형태] 명사 대명사 수사 동사 형용사 보조용언 관형사 부사 감탄사 어근 3. 품사[문법형태] 조사 어미 접사 4. 전문어 여부 일반어 전문어	1. 품사[어휘형태] 명사 대명사 수사 동사 형용사 관형사 부사 감탄사 (무품사: 구) 2. 품사[문법형태] 조사 어미 접사 3. (무품사: 준말)
2단계	(고유어) 발음 짧은소리 긴소리 현대어 표준어 방언 고어 (한자어) 획수 순	(발음) 짧은소리 긴소리 (한자어) 획수 순		(표준어 여부) 표준어 비표준어 방언 (어원) 고유어 한자어 외래어 (빈도) 높은 순 (조어방법) 단일어 복합어

1.2. 한정어

일부 국어사전에서는 동형이의어에 어깨번호를 덧붙이는 방법보다, 어휘의 의미와 구별을 효율적으로 나타내기 위해 한정어를 사용하기도 한다. 한정어는 어깨번호와 달리 표제어와 밀접하게 관련된 단어나 숫자 등의 표지 정보로서, 사전의 범위와 상관없이 의미적인 구별이 지속될 수 있는 구별 방법이다. 최근 우리나라의 국어사전에서도 한정어를 사용하여 동형이의어를 구별하는 사전이 편찬되고 있는데, 『연세 초등국어사전』(2001)과 『보리 국어사전』(2008)이 대표적이다.

『연세 초등국어사전』에서는 동형이의어에 길잡이말[20]을 붙여 표제어를 구분하고 있다. 동형이의어 배열은 고유어, 한자어, 외래어 순으로 하되, 한자어의 경우 획이 적은 것에서 많은 것의 순서에 따라 싣고 있다.

가래 [침] 사람의 목구멍에 생기는 끈적끈적한 액체.
가래 [도구] 흙을 떠서 던지는 데 쓰는, 긴 삽처럼 생긴 도구.

배 [몸] ①몸에서 가슴과 다리 사이에 있는 몸의 앞부분.
배 [탈것] 물 위에 떠서 사람이나 짐을 실어 나르는 교통 수단.
배 [과일] 껍질이 누렇고 속은 희며 크고 둥글고 즙이 많고 시원한 단맛이 나는, 가을에 나는 과일
배 [세는 말] (倍) (수를 나타내는 말 뒤에 써서) 같은 수량을 여러 번 합한 만큼의 분량.

『보리 국어사전』에서도 『연세 초등국어사전』과 같이 동형이의어의 구별을 위해 올림말의 뜻을 드러낼 수 있는 '길잡이 어깨말'[21]을 표제어의 뒤에 붙이고 있다. 배열의 순서는 배우기 쉬운 말부터 어려운 말 순으로, 고유어를 한자어나 외래어보다 먼저 싣고 있다.

---

20) '길잡이말'은 『연세 초등국어사전』(연세대학교 언어정보개발연구원, 2001)의 일러두기에 "형태가 같은 말이 여럿 있을 때 그 중에서 찾고자 하는 단어가 어떤 것인지 쉽게 구분할 수 있도록 도움을 주는 말"이라고 정의하고 있는데, 이는 기존 사전에서 동형이의어의 구분을 위해 붙이는 어깨번호에 해당된다. 이에 대한 자세한 논의는 유현경(2000)과 이희자(2001)을 참조.

21) 이 말은 『보리 국어사전』(보리, 2009)의 일러두기에서 "꼴이 같은 올림말은 뜻을 드러낼 수 있는 길잡이 어깨말을 붙여 구분하였다."라고 쓰고 있다. 이는 『연세 초등국어사전』의 '길잡이말'과 같은 의미로서 이후 '동형이의어를 구별하기 위해 붙이는 말'에 대해서는 '한정어'로 통일하여 쓰도록 하겠다.

가래 [농기구] 흙을 파헤치거나 떠서 던지는 데 쓰는 농기구.
가래 [침] 사람의 목구멍에 생기는 끈끈한 액체.
가래 [풀] 연못이나 논에서 자라는 풀.

배 [몸] ①몸 가운데 부분. 사람의 배는 가슴 아래, 엉덩이 위쪽에 있다.
배 [탈것] 물 위에 떠서 사람이나 짐을 실어 나르는 탈것.
배 [과일] 배나무 열매. 크고 둥근데, 맛이 달고 즙이 많다.
배 [세는 말] (倍) 어떤 수나 양을 두번 더한 만큼.

1.3. 원어 정보

마지막으로 동형이의어를 구별하는 방법에는 원어 정보를 사용하는 경우를 들 수 있다. 『금성판 국어대사전』에서는 기본적으로 동형이의어에 어깨번호를 붙여 구별하고 있지만, 원어의 한자나 로마자가 다른 경우 어깨번호를 붙이지 않고 있다. 이는 한글의 형태적 중의성이 원어를 통해서 구별되고 있다고 보기 때문인데, 이 사전의 일러두기에는 "동음 이의어(同音異義語)는 표제어 우측 상단에 어깨 번호를 붙였으나, 음이 같더라도 한자나 로마자가 다른 어휘끼리는 굳이 그것을 보이지 않았다"라고 설명하고 있다.[22)]

나탈 : 1 [Natal] 「명」<지> 남아프리카 공화국의 동쪽에 있는 주.
나탈 : 2 [Natal] 「명」<지> 브라질 동북부, 리우그란데 두노르테 주(Rio Grande de Norte 州)의 주도.

뉴 : 턴 [Isaac Newton] 「명」 영국의 물리학자 · 천문학자 · 수학자(1642~1727).
뉴 : 턴 [newton] 「명」(의존)<물> 힘의 MKSA 단위.

사 : 기[士氣] 「명」 몸과 마음이 기운으로 넘쳐 굽힐줄 모르는 씩씩한 기세.
사 : 기[四氣] 「명」 사시(四時)의 기운.

---

22) 『금성판 국어대사전』에서는 원어가 다른 동형이의어에 어깨번호를 붙이지 않는다고 밝히고 있지만, 다음과 같이 로마자가 다른 경우에도 어깨번호를 붙이는 경우가 보이고 있다.

다 : 윈 [Darwin] 「명」<지> 오스트레일리아 북부에 있는 항구 도시. 진주 채취의 기지임.
다 : 윈[1] [Charles Robert Darwin] 「명」 영국의 생물학(1809~82).
다 : 윈[2] [Erasmus Darwin] 「명」 영국의 의사 · 철학자 · 박물학자(1731~1802).

사 : 기[1][史記] 「명」 역사적인 사실을 적어 놓은 책.
사 : 기[2][史記] 「명」『책』 중국 한(漢)나라의 사마천(司馬遷)이 적은 역사책.

위의 예들은 『금성판 국어대사전』에 수록된 동형이의어들로 한글 표기는 같으면서 원어 정보가 다른 경우이다. '나탈[Natal]'과 '사기[史記]'는 한글 표기는 물론 원어의 표기까지 동일하기 때문에 어깨번호를 붙여 이들을 구별하고 있지만, '뉴턴[Isaac Newton], 뉴턴[newton]'이나 '사기[士氣], 사기[四氣]'의 경우에는 원어 정보가 다르기 때문에 별도의 어깨번호를 붙이지 않고 있다.

이와 같이 원어 정보를 이용하여 동형이의어를 구별하고 있는 사전으로는 『동아 새국어사전』(1989)과 『DESK 국어사전』(1997) 등이 있는데, 이들 사전들은 모두 같은 출판사에서 발간된 것으로서 동형이의어의 배열 방법 또한 일치하고 있다.

2. 문제점

어깨번호는 하나의 사전 내에서 어휘를 구별하고 배열의 순서를 표시한다는 측면에서 매우 경제적이고 효율적인 방법이라고 할 수 있다. 그러나 어깨번호를 붙이는 기준과 범위가 모든 사전마다 다르기 때문에 두 개 이상의 어휘 사전을 통합하거나 정보처리를 시도할 때, 이들의 구별 기능은 무의미해지게 된다. 즉, 어깨번호는 개별 사전 내에서 동형이의어의 배열 순서를 표시하는 것 외에 해당 어휘와의 어떠한 의미적 연관성도 가지고 있지 않기 때문에 개별 사전의 범위를 넘어서는 순간 그 의미를 상실하게 된다. 설령 모든 사전들이 동일한 원칙으로 어깨번호를 붙인다고 하더라도 수록하는 어휘 수의 차이에 따라 어깨번호는 달라지게 되며, 동일한 사전에서도 개정 등의 과정을 거치면서 표제어의 가감이 있을 수 있다. 따라서 사전의 어깨번호를 통해서 어휘 정보를 교환하는 것은 사실상 불가능하다고 할 수 있다.

한정어는 어깨번호와 같이 동형이의어를 구별하는 기능과 더불어 표제어와 밀접하게 관련된 단어를 제시함으로써 표제어의 의미를 압축적으로 전달하는 의미적 기능을 함께 갖고 있다. 하지만, 한정어는 형식적 일관성과 통일성을 유지하기가 어렵기 때문에 이종 간의 사전 정보를 처리하기에는 적절한 방법이 될 수 없다. 즉, 어떤 일군의 동형이의어에 대해 한정어를 부여함에 있어 색인자의 개인적

지식의 차이나 언어에 대한 이해의 정도에 따라 부여되는 한정어가 달라질 수 있으며, 사전이 다를 경우 이러한 차이는 더욱 크게 나타날 가능성이 있는 것이다.[23]

<『연세 초등국어사전』>
가마[장치] 숯・기와・벽돌・질그릇 따위를 구워 만드는 시설.
가마[가마니] '가마니'의 수를 세는 말.

<『보리 국어사전』>
가마[도자기] 숯, 기와, 벽돌, 질그릇 들을 구워 내는 큰 아궁이.
가마[쌀] → 가마니

실제로 위의 예에서 보는 바와 같이 한정어를 통해 동형이의어를 구별하고 있는 『연세 초등국어사전』과 『보리 국어사전』을 비교해보면, '가마(그릇을 구워내는 시설)'와 '가마(가마니를 세는 말)'의 한정어가 '장치/도자기', '가마니/쌀'로 서로 다르게 붙어 있다. 이는 한정어의 목적이 단일 사전 내에서 동형이의어의 구별을 위해 설계된 것이기 때문에, 동일한 단어의 한정어라도 사전에 따라 큰 차이를 보이고 있다. 따라서 한정어 또한 사전 간의 정보 교환을 위한 식별자로서 적절한 방법이 되지 못한다고 할 수 있다.

마지막으로 원어 정보를 이용하여 동형이의어를 구별하는 방법으로, 한자어의 비율이 높은 우리말에 대한 효과적인 구별 방법이 될 수 있다. 원어 정보는 집필자의 자의적인 판단으로 달라질 수 없다는 점에서 표기의 일관성을 유지할 수 있고, 대부분의 국어사전에서 한자어나 외래어에 대해서 원어 정보를 표기하고 있기 때문에 개발의 비용 측면에서도 장점이 있다. 그런데 원어 정보만으로 동형이의어를 구별하려고 할 경우 '나탈[Natal]'이나 '사기[史記]'와 같이 원어 정보까지 동일한 형태의 동형이의어나 원어 정보가 없는 고유어에 대한 구별 방법이 문제가 된다. 이러한 경우 『금성판 국어대사전』과 같이 원어 정보를 통해 동형이의어를 구별하고 있는 국어사전에서는 어깨번호를 부여함으로써 이를 처리하고 있는데, 앞서 살펴본 바와 같이 어깨번호는 사전간의 호환성이 전혀 없다는 문제점을 가지

23) 김태수, 최석두. 1997. 전게서

고 있는 것이다. 따라서 원어 정보를 이용하여 동형이의어의 구별하기 위해서는 원어 정보로 식별되지 않는 동형이의어의 처리 방법을 찾아야만 한다.

사전간의 어휘 정보를 교환하거나 정보처리를 위한 식별자로서 원어 정보의 사용이 타당하다고 판단되는 이유는, 원어 정보가 단어의 고유한 속성으로서 사전에 따라 다르게 표기될 수 없기 때문이다. 따라서 동형이의어의 식별을 위해서는 단어의 고유한 속성을 이용하는 것이 매우 중요한데, 이러한 측면에서 국어사전의 어깨번호 배열 방법에 주목할 필요가 있다. 앞서 살펴본 바와 같이 국어사전에서는 고유어의 여부, 품사, 발음의 길이, 표준어 여부, 한자의 부수/획수 등을 고려하여 동형이의어의 배열 순서를 정하고 있는데, 이러한 정보는 단어에 내재된 언어적 속성으로서 사전에 따라 다르게 표기될 수 없는 것들이다. 원어 정보를 통해서 식별되지 않는 동형이의어들에 대해서 위와 같은 언어적 속성 정보를 추가하여 구별할 수 있다면, 사전간의 어휘 교환을 위한 호환성을 확보할 수 있다. 즉, 동형이의어의 구별을 위한 정보가 언어 보편적이고 유일하며, 모든 어휘에 공통적으로 적용될 수 있다면 범용적인 식별자로서의 역할을 할 수 있을 것이다.

IV. 사전의 미시정보를 이용한 동형이의어 구별 방안

1. 사전의 구조와 내용

사전은 표제어의 수록 범위와 선정 기준, 표제어의 표기와 배열 원칙에 해당하는 거시구조(macro-structure)와 개별 표제어에 대한 형태론적, 음운론적, 통사론적 정보, 의미의 기술, 용례 설정 등과 관련된 미시구조(micro-structure)로 구성된다.[24] 사전의 구조는 편찬 목적과 규모, 대상 언어의 특성에 따라 다양한 형태로 나타난다. 특히 언어 사전인지 백과사전인지, 일언어 사전인지 이언어 사전인지, 모어 사전인지 외국어 사전인지 등에 따라 전혀 다른 형식을 갖기도 한다. 하지만 동시대에 비슷한 목적을 가지고 편찬된 사전들은 대체로 그 구성이나 내용에 있어서 유사한 부분이 적지 않은데, 이는 오랫동안 이어져 온 사전 편찬의 과정에서 얻어진 효과적인 표현 방식을 서로 공유하는 점이 있고, 이전에 편찬된 사전의 형식에 익숙해진 사용자들을 고려하지 않을 수 없는 점이 있기 때문이다.

이 절에서는 국어사전에서 나타나는 거시적, 미시적 구조의 특성에 대해 개략적으로 살펴보고자 한다. 분석 대상으로 삼은 국어사전으로는 앞 장에서 언급한 바와 같이 1988년 한글 맞춤법이 개정 고시된 이후 출간된 대사전 규모의 국어사전으로 하였다.

1.1. 거시구조

사전의 거시구조는 표제항을 일정한 원칙에 따라 배열한 종적 구조이다. 이는 달리 말하면 미시 구조의 조합으로 짜여진 표제어 항목을 일관된 기준에 따라 종적으로 배열한 것이라 할 수 있다. 사전의 거시구조에서는 표제어의 선정 기준과 배열 원칙, 다의어와 동형어어의 처리가 주된 문제가 된다.

첫째, 표제어의 선정 기준은 통시적으로 현대어만 수록할지, 아니면 옛말까지를 모두 포함할지, 공시적으로 표준어만 수록할지 방언까지 수록할지 등에 관한 기준

24) Hartmann(2001)에서는 사전의 구조를 최대구조, 거시구조, 미시구조, 매개구조, 접근구조, 배분구조의 여섯 가지로 구분하고 있다.

을 정하는 문제이다. 다음은 『금성판 국어대사전』의 일러두기에서 표제어의 선정 범위를 밝힌 것으로, 이를 통하여 이 사전에 수록된 표제어의 대강을 짐작할 수 있다.

> 공시적(共時的)으로는 현대 표준어를 포함하여 우리나라 전 지역 모든 계층의 언어를, 통시적(通時的)으로는 15세기 이후의 중세 국어에서 현대 국어에 이르는 말들을 망라했다. 발생적으로는 고유어・한자어・외래어로 나뉠 수 있고, 속성에 따라 일상어(日常語)・속어(俗語)・비어(卑語)・신어(新語)・특수 집단어・전문어 등으로 나뉠 수 있으며, 규범성(規範性) 또는 생존성(生存性) 여부에 따라 표준어와 비표준어, 현대어와 고어(古語)로 나뉠 수 있다. 아울러, 인명・지명・단체명・책명과 같은 고유명사를 수용함으로써 백과 사전적 요소를 가미했다. 한편, 이두(吏讀)는 권말의 부록으로 수록했다[25]

『우리말큰사전』, 『표준국어대사전』, 『고려대 한국어대사전』에서도 각기 일러두기를 통해 표제어의 수록 범위를 자세히 밝히고 있는데, '옛말, 북한어, 이두'의 수록 여부를 제외하고는 네 사전이 대체로 비슷하다. 다만, 『고려대 한국어대사전』에서는 표제어를 수록함에 있어 "고려대 민족문화연구원에서 구축한 한국어 어휘 데이터베이스에서 가려 뽑은 어휘 약 39만 개를 표제어로 삼았다"고 밝힘으로써 코퍼스(corpus)를 활용하여 편찬하였음을 알 수 있다.[26]

둘째, 사전에 있어서 표제어를 배열하는 가장 대표적인 기준은 표기순에 따른 것이다. 사전의 특성에 따라 분류 사전과 같이 주제를 기준으로 배열할 수도 있고, 어기를 중심으로 표제어를 모아서 배열하기도 하지만, 대부분의 국어사전에서는 자모순으로 표제어를 배열하고 있다.[27] 현행 국어사전에서는 현대 한글을 기준으로 초성 19자, 중성 21, 종성 27을 <한글 맞춤법(1988)>에서 제시한 사전의 자모 배열 순서를 반영하여 다음과 같이 배열하고 있다.[28]

---

25) 운평어문연구소 편. 1991. 『금성판 국어대사전』. 서울: 금성출판사. p.iii.

26) 『금성판 국어대사전』과 『표준국어대사전』에서도 시, 소설, 수필 등과 같은 문학 작품과 신문, 잡지, 교과서 등의 문헌에서 뽑은 표제어와 용례를 수록하고 있으나, 대량의 문헌 자료를 컴퓨터에 저장하고 어휘를 추출하여 사전을 편찬하는 본격적인 의미의 코퍼스에 기반한 사전이라 하기 어렵다. 국내에서 코퍼스에 기반하여 국어사전이 편찬된 것은 『연세한국어사전』(연세대학교 언어정보개발연구원, 1998)이 최초이고, 대사전 규모에서는 2009년 『고려대 한국어대사전』이 편찬되었다.

27) 주제를 기준으로 표제어를 배열하고 있는 사전으로는 『우리말 갈래사전』(박용수, 1990), 『현대 한국어 학습사전: 의미로 분류한』(신현숙・김미형 등, 2000) 등이 있으며, 어기를 기준으로 한 사전으로는 『새우리말큰사전』(신기철・신용철, 1975)이 있다.

28) 김양진(2008)에서 이와 같은 국어사전의 표제어 배열은 엄밀히 말하면 자모순 배열이라기보다는

- 초성: ㄱㄲㄴㄷㄸㄹㅁㅂㅃㅅㅆㅇㅈㅉㅊㅋㅌㅍㅎ
- 중성: ㅏㅐㅑㅒㅓㅔㅕㅖㅗㅘㅙㅚㅛㅜㅝㅞㅟㅠㅡㅢㅣ
- 종성: ㄱㄲㄳㄴㄵㄶㄷㄹㄺㄻㄼㄽㄾㄿㅀㅁㅂㅄㅅㅆㅇㅈㅊㅋㅌㅍㅎ

『금성판 국어대사전』과 『표준국어대사전』에서는 명사 어근에 '-하다/-되다', '-이/-히'와 같은 파생 접미사가 붙어 생겨난 단어를 부표제어로 설정하여 다음과 같이 주표제어의 아래에 배열하기도 한다. 일반적으로 국어사전에서는 관용구나 속담도 부표제어로 설정하여 주표제어 아래에 배열하는 것이 보통이다.

<『표준국어대사전』>
용천[1] 「명사」「1」 나병, 간질 따위의 몹쓸 병.
  용천-하다[1] [I]「동사」 => 용천.
용천[2](用天) [용 : -] 「명사」하늘을 다스림.
  용천-하다[2](用天--)[용 : ---] 「동사」 => 용천.
용천[3](湧泉) [용 : -] 「명사」「1」 =용천혈.

셋째, 다의어와 동형어 처리 문제가 있다. 다의어는 하나의 표기에 여러 가지의 뜻을 가진 낱말을 이르는 것이고, 동형어는 표기 형태는 같지만 의미가 다른 말을 뜻한다. 다의어와 동형어의 구분은 어원과 의미적 관련성이 중요한 기준이 되기는 하지만 그 구분이 불분명한 경우가 매우 많다.

1.2. 미시구조

사전의 미시구조는 표제어에 대한 형태적인 표기와 의미의 기술, 용례의 설정 등과 같이 표제어와 관련된 모든 정보들을 횡적으로 나열한 구조를 말한다. 사전의 미시구조에는 원어, 어깨번호, 발음, 품사, 전문어, 뜻풀이, 용례, 유의어나 반의어, 어휘 빈도 등이 포함되는데, 사전학에서는 이들을 크게 표제어의 형식적인 측면을 다루는 형식 항목과 표제어의 의미적인 측면을 다루는 내용 항목으로 나누고 있다. 미시정보의 형식 항목에는 표제어, 어깨번호, 원어, 품사, 활용, 전문 분야 등 표제어의 형태적 표기와 형식적 분류에 관련된 정보들이 포함되고, 내용 항목

자절순 배열이라고 밝히고 있다. 자절순 배열이란 초성자를 먼저 자모순으로 배열한 뒤 각 초성자에 대하여 중성자와 종성자를 순차적으로 아울러 이룬 자절을 단위로 한 배열을 말한다.

에는 문형, 의미역, 뜻풀이, 용례, 관련어와 같은 표제어의 의미와 구문 정보들이 포함되는데, 내용과 순서를 요약하면 다음의 표 8과 같다.

표 8 국어사전 표제항의 미시구조[29)]

형식 항목(왼쪽 항목)					내용 항목(오른쪽 항목)				
표제어	원어	발음	품사	전문어	논항틀	뜻풀이	용례	관련어	참고어
형태분석 및 어깨번호 포함	어원 포함	표준발음 및 허용발음	고유명, 불규칙 활용 포함	인문과학, 사회과학, 자연과학, 생활과학	주술, 주목술, 주부술, 주목부술	기본의미 + 부가의미	출전 포함	준-본, 큰-작은, 센-거센-여린 등	유의어, 반의어, 상하위어 등

미시정보의 배열에 있어서는 형식 항목이 내용 항목에 앞서 놓이며, 형식 항목에서도 표제어가 항상 맨 앞에 자리잡는다. 표제어에는 동형어를 구별하기 위해 어깨번호를 덧붙이는 것이 일반적인데, 엄밀하게 보면 어깨번호는 표제어에 대한 어떠한 유의미한 정보도 갖고 있지 않는 것으로 단순히 동형어의 구별과 순서를 정하는 표지에 불과하다. 이와 비슷한 것으로 한정어가 있는데, 한정어는 동형어를 구별하는 기능과 더불어 표제어와 밀접하게 관련된 단어를 제시함으로써 표제어의 의미를 압축적으로 전달하는 의미적 기능을 함께 갖고 있다.

표제어와 어깨번호 다음으로는 표제어의 형태와 직접적으로 관련된 원어와 어원 및 학명이 제시되는 경우가 많다. 하지만 이러한 정보를 모두 표제어 뒤에 나열할 경우 뜻풀이와의 거리가 멀어지는 부담이 있어 내용 항목이 모두 끝난 뒤에 배열하는 경우도 있다. 『금성판 국어대사전』은 표제어 다음으로 '원어-어원-전문어-학명' 순으로 이들 정보를 모아서 보여주고 있으나, 『우리말큰사전』에서는 원어와 어원 모두 내용 항목의 참고어 뒤에 두고 있으며, 학명은 기술하지 않고 있다. 『표준국어대사전』과 『고려대 한국어대사전』에서는 원어를 표제어 뒤에 싣고 어원은 모든 정보의 맨 마지막에 두고 있으며, 학명의 경우 뜻풀이 뒤에 두고 있다.

갈-까마귀 명(Corvus monedula)『동』 까마귓과의 새. 까마귀보다 약간 작으며, 몸빛은 검은데 목 둘레와 배가 흼. **<『금성판 국어대사전』>**

갈가마귀 「이」『동』 ①까마귓과에 딸린새. …… 『한』 비거2. 여사11. 연오. ②갈가마귀, 당

29) 홍종선 외. 2009. 『국어사전학 개론』. 서울: 제이앤씨. p.118.

까마귀, 산갈가마귀를 두루 일컫는 말. **<『우리말 큰사전』>**

갈-까마귀 명『동』 까마귓과의 새. …… =비거05(鴝鵒)·여사17(鷺斯)·연오(燕烏). (Corvus dauuricus) 〔<ᄀᆞᆯ가마괴<훈몽>←ᄀᆞᆯ-+가마괴〕 **<『표준국어대사전』>**

갈까마귀 〔=갈+까마귀〕 명 「동물」 참새목 까마귓과에 속한 검은 새. …… 학명은 Corvus monedula이다. (유의)비거3(鴝鵒), 여사13(鷺斯), 연오(燕烏). ×갈가마귀. 【ᄀᆞᆯ가마괴 《훈몽上:9》】 **<『고려대 한국어대사전』>**

원어 정보 다음으로 발음, 용언의 활용, 품사가 오고 형식 항목의 마지막에는 흔히 전문어가 자리하는데, 『금성판 국어대사전』에서는 용언의 활용 정보를 품사 뒤에 두고 있으며, 『우리말큰사전』은 용언의 활용 유형만을 품사 뒤에 싣고 있다.

표제항의 형식 항목이 끝나면 내용 항목이 뜻풀이를 중심으로 배열된다. 뜻풀이에는 표제어에 대한 정의와 함께 문형 정보, 사용 영역 정보 등이 함께 제시된다. 문형 정보는 표제어가 문장 속에서 다른 언어 요소들과 결합될 때의 형식과 배치에 관한 것으로 “덩달다”라는 용언이 “‘덩달아’, ‘덩달아서’의 꼴로 쓰여”처럼 활용의 형태를 보여주거나, “가두다”가 “㉮이 ㉯을 ㉰에게”와 같은 문장 구조로 배치됨을 보이는 것이다. 또한 사용 영역이란 표제어가 쓰이는 ‘사용자의 영역’으로 계층이나 연령, 지역, 문체에 따라 달리 쓰이는 것으로, “잡수다”에 대해 “주로 아랫사람이 윗사람을 대하여 쓰는 말로”처럼 사용자의 범위에 대해 덧붙이는 설명 정보를 말한다.

뜻풀이 다음에는 해당 표제어가 실제 사용되는 단어나 구, 문장 등의 용례가 오는데, 단어 용례의 경우 ‘개-살구, 불쏘시-개’와 같이 표제어가 접두사나 접미사일 경우 제시된다. 구 용례의 경우 ‘가물이 들다, 개망신을 당하다, 일회용 반창고’와 같이 표제어와 필수적인 의미 관계를 보이는 구(句)를 수록하며, 문장 용례의 경우 문헌 자료에서 인용한 경우 출전 정보를 제시하기도 한다.

관련어는 표제어와 형태적, 통사적, 의미적으로 직접적인 관계가 있는 단어를 말한다. 형태적인 관련어에는 본말-준말, 원어-약어, 큰말-작은말, 센말-거센말-여린말, 이형태 등이 있으며, 의미적으로 관련된 어휘로는 동의어, 유의어-반의어, 상위어-하위어 등이 있다. 통사적인 관련어에는 주동-사동, 능동-피동 등을 들 수 있으며, 기타 사용 영역과 관련된 표준어-비표준어, 비어-속어 등이 있다.

참고어는 관련어와 같이 직접적인 관계는 없지만 해당 표제어를 이해하는 데

필요한 단어가 포함되는데, 번역어나 학명, 별칭, 이칭 등을 들 수 있다. 그 밖에 미시구조의 내용 항목으로는 표제어의 중요도나 빈도, 부가적인 설명이 포함된다.

2. 미시정보의 선정 기준

언어 사전은 대상이 되는 언어의 형태적 표기와 음성적 특징, 다양하게 분화되는 의미와 문법적 규칙 등에 대한 정보와 함께 동형이의어의 구별을 위해 어깨번호와 같은 표지를 덧붙이고 있다. 그러나 Ⅲ장에서 살펴본 바와 같이 표제어와 의미적 연관성이 없는 표기 정보는 개별적인 사전의 범위를 넘어서는 순간 그 기능을 상실하게 되는 문제점을 안고 있다. 사전 정보가 컴퓨터와 인터넷을 통해 전송되고 다른 지식 정보와 융합하여 활용되기 위해서는 최소한 형태적인 측면에서의 중의성이 해소되어야 한다. 그러기 위해서는 모든 동형이의어에 공통된 표기를 붙이든지, 아니면 형태적인 표기 외에 단어마다 가지는 고유한 언어적 속성을 찾아야만 한다. 사람들은 일상생활 속에서 많은 동형이의어들을 별다른 어려움 없이 구별하여 사용하고 있다. 이는 우리가 인식하지 못하는 음성적, 통사적 규칙이 언어 속에 내재되어 있음을 보여주는 것이라 할 수 있다. 사전에는 이러한 언어적 분석 정보가 일정한 원칙에 따라 기술되어 있는데, 이는 동형이의어의 구별을 위해서도 활용될 수 있다.

현재 사전에 수록되어 있는 정보 가운데 동형이의어를 식별할 수 있는 가장 확실한 방법은 사전의 뜻풀이 정보를 이용하는 것이라 할 수 있다. 그러나 뜻풀이는 모든 사전마다 표현하는 방법이 다르고, 뜻풀이 정보를 구조적으로 기술하는 것 또한 매우 어렵다. 따라서 미시정보 가운데 기술 방법이 어렵지 않고, 정보간의 중복이 비교적 적은 요소를 찾을 필요가 있다.

미시정보 가운데 표제어의 의미와 관련되는 내용 항목은 대체로 사전간의 차이가 크고 집필자의 자의성이 높아 식별 요소로 적합하지 않다. 특히 뜻풀이와 용례는 집필자에 따라 가장 많은 차이를 보이는 요소이다. 관련어 또한 동의어, 유의어, 반의어, 상위어, 하위어와 같은 경우 사전마다 그 기준이 같지 않고 집필자에 따라 다르게 판단될 수 있는 여지가 많다. 다만, 표제어와 형태적으로 관련되는

본말-준말, 원어-약어, 큰말-작은말, 센말-거센말-여린말, 이형태 등은 비교적 단어간의 관계가 명확하여 집필 기준만 같다면 일관되게 기술될 수 있는 요소이다.

이에 반해, 미시정보의 형식 항목에 해당하는 어원, 원어, 발음, 조어 방법, 품사, 용언의 활용과 같은 요소는 단어에 내재된 속성들로서 집필자가 자의적으로 개입할 수 있는 여지가 거의 없다.[30] 어원의 경우 고유어에 대해 그 말의 최초 출현형이나 변천의 과정을 보여주는 것으로, 어원의 범위를 어디까지 볼 것인가에 따라 기술의 내용이 달라질 수 있다. 어휘의 사용 영역에 관한 전문어 정보는 보통 학문의 분야에 따라 인문 과학(철학, 역사, 문학), 사회 과학(법학, 정치학, 사회학), 자연 과학(화학, 물리학, 전자공학) 등으로 세분하는데, 사전에 따라 분류 방법이 다르게 나타난다.

따라서 본 논문에서는 이상의 미시정보 가운데 원어, 발음, 조어 방법, 품사, 용언의 불규칙 활용, 방언/비표준어의 대응 표준어를 동형이의어의 식별 요소로 선정하여 동형이의어의 구별 방법과 식별률을 계량적으로 분석하고자 하는데, 이들 요소의 선정 기준은 다음과 같다.

첫째, 사전의 편찬 원칙이나 사전 집필자에 따라 내용이 달라지지 않아야 한다. 뜻풀이나 용례와 같이 편찬 원칙이나 집필자의 어휘 능력에 따라 내용이 달라지는 요소를 식별 정보로 사용하기는 어렵다.

둘째, 대부분의 표제어에 기술될 수 있는 요소여야 한다. 특정한 표제어에만 해당되는 요소를 식별 요소로 사용하게 되면 구별의 효율성이 매우 떨어지게 된다. 다만, 다른 요소로 구별되지 않거나 식별성이 뛰어난 요소의 경우 일부의 표제어에만 기술된 요소라도 식별 요소로 사용될 필요가 있다.

셋째, 해당 정보의 집필이 쉽고 명료해야 한다. 해당 요소의 집필이 매우 어렵거나 요소의 기술이 복잡하여 식별 정보의 기술에 많은 비용이 소요되어서는 안 된다.

30) 우리나라의 경우 『표준국어대사전』을 규범 사전으로 삼고 있기 때문에 위의 요소들에 대한 국어 사전의 통일성이 매우 높다고 할 수 있다.

3. 미시정보별 동형이의어 식별률 분석

3.1 미시정보별 동형이의어 식별률

지금까지 국어사전에 나타나는 미시정보의 유형과 동형이의어의 구별을 위한 미시정보의 선정 기준에 대해 살펴보았다. 그리고 국어사전에 수록되어 있는 다양한 미시정보 가운데 동형이의어의 구별에 적합할 것으로 판단되는 '원어, 발음, 조어 방법, 품사, 용언의 불규칙 활용, 방언/비표준어의 대응 표준어' 여섯 가지 미시정보를 선정하였다. 이 절에서는 위에서 선정한 미시정보들의 동형이의어 식별 능력을 구체적으로 확인하기 위해, 『고려대 한국어대사전』을 대상으로 각 미시정보별 동형이의어 식별률을 분석하고자 한다. 이 사전은 앞서 Ⅱ장에서 살펴본 바와 같이 현대 한국어만을 대상으로 386,889개의 표제어를 수록하고 있으며, 이 가운데 99,499개의 동형이의어가 출현하고 있다. 이 사전을 분석 대상으로 삼은 것은 전체 사전 자료가 데이터베이스로 구축되어 있어 각 미시정보에 대한 정확한 통계 정보와 식별률을 얻을 수 있을 뿐만 아니라, 1억 어절 규모의 현대 한국어 코퍼스를 대상으로 표제어를 추출하고 있어 지금까지 편찬된 대사전 규모의 국어사전 가운데 현대 한국어의 어휘를 가장 잘 반영하고 있기 때문이다.

미시정보별 식별률 분석 방법은, 먼저 『고려대 한국어대사전』에 수록된 전체 표제어 가운데 여섯 가지 미시정보가 표기된 표제어를 각각 추출하고, 미시정보별 표제어 그룹 내에서 동형이의어 관계가 발생되는 표제어를 확인한 다음, 미시정보별 표제어 그룹의 동형이의어를 대상으로 해당 미시정보를 통해 구별되는 표제어의 비율을 확인하고자 한다. 식별률 분석을 위해 사용된 미시정보들은 사전에 수록된 내용을 모두 사용하였으나, 미시정보의 표기 형태는 정보 처리를 위한 전처리 과정을 통해 불필요한 정보들을 삭제하거나 가공하였음을 밝힌다.

3.1.1 원어 정보

국어에서 원어 정보가 동형이의어의 구별에 유용하게 활용될 수 있을지 판단하기 위해서는 국어사전에 나타나는 원어 정보에 대한 정밀한 분석이 필요하다. 이 장에서는 『고려대 한국어대사전』을 대상으로 전체 표제어에 나타나는 원어 정보

의 비율과 유형을 살펴보고, 원어 정보의 유형별로 동형이의어 식별률을 분석하도록 하겠다.

표 9는 『고려대 한국어대사전』에서 원어 정보가 있는 표제어를 품사별로 나타낸 것으로, 표의 제목에서 '원어 정보'는 품사별로 원어 정보가 있는 표제어의 수이고, 비율은 품사별 표제어 수에 대한 원어 정보가 있는 표제어 수의 비율이다. 먼저 원어 정보의 합계를 보면 293,698개로 전체 386,889개의 표제어 가운데 75.91%를 차지하고 있다. 품사별로 살펴보면 명사 203,385개, 무품사 44,331개, 동사 34,509개, 형용사 5,820개, 부사 2,950개, 다품사 2,087개 순으로 나타나고 있다.[31] 조사와 어미는 원어 정보를 가지는 표제어가 나타나지 않는데, 이는 우리말에서 쓰이는 모든 조사와 어미가 고유어이기 때문이다.

표 9 원어 정보가 있는 표제어 통계[32]

품사	표제어 수	원어 정보 수	비율(%)
명사	253,303	203,385	80.29
대명사	380	167	43.95
수사	215	87	40.47
동사	54,921	34,509	62.83
형용사	12,271	5,820	47.43
관형사	192	32	16.67
부사	13,841	2,950	21.31
감탄사	675	38	5.63
조사	338	0	0.00
어미	1,168	0	0.00
접사	555	292	52.61
무품사	45,838	44,331	96.71
다품사	3,192	2,087	65.38
합계	386,889	293,698	75.91

품사별 원어 정보의 비율을 보면, 무품사 96.71%, 명사 80.29% 순으로 다른 품사에 비해 매우 높게 나타나고 있는데, 이는 우리말 가운데 명사나 명사 구에 해당하는 표제어가 한자어나 외래어, 또는 한자어나 외래어와 결합된 경우가 많음을 나타내주는 것이다.

31) 표 9의 품사 항목 가운데 무품사에는 '지방 자치 단체', '시민 사회', '정치 자금' 등과 같이 둘 이상의 어절이 어울려서 절이나 문장 안에서 명사의 구실을 하는 구(句)나, '갠', '게다가', '고럼'과 같이 준 말이 해당된다. 다품사는 '매일, 당분간', '근면하다, 너무하다, 마르다'과 같이 하나의 단어형이 '명사와 부사' 또는 '동사와 형용사'와 같이 두 가지 이상의 문법적 기능을 가지는 말을 이른다.

32) 도원영, 차준경. 2009. 전게서.

표 10 원어의 유형별 통계

원어의 유형	유형별 표제어 수	비율(%)
고유어	93,193	24.1
한자어	203,931	52.7
외래어	16,624	4.3
고유어+한자어	63,601	16.4
고유어+외래어	1,129	0.3
고유어+한자어+외래어	463	0.1
한자어+외래어	7,948	2.1
계	386,889	100.0

표 10은 표제어에 대한 원어 정보의 유형별 통계인데,[33] 전체 표제어 중에서 '한자어'가 52.7%로 가장 많고, 다음으로 '고유어'가 24.1%, 그리고 '고유어+한자어'가 16.4%로 그 뒤를 잇고 있다. 전체 표제어에서 한자어가 차지하는 비율은 고유어보다 두 배 이상 높게 나타나고 있으며, 한자어가 포함된 표제어를 포함할 경우에는 전체 표제어의 71.3%인 275,943개에 달한다. 이에 반해 외래어의 경우 4.3% 정도로 한국어에서 차지하는 비중이 한자어에 비해 상대적으로 매우 적게 나타나고 있다.[34] 결국 원어 정보를 통한 동형이의어의 식별 방법은 한자의 변별 능력이 매우 크게 작용될 것임을 짐작할 수 있는데, 실제 다음의 예를 통해서 한자어의 동형이의어 식별 능력을 확인할 수 있다.

조사[1] [調査] 명 어떤 일이나 사실 또는 사물의 내용 따위를 명확하게 알기 위하여 자세히 살펴보거나 밝힘.

조사[2] [祖師] 명① 「불교」 한 종(宗)이나 파(派)의 선덕(先德).

조사[3] [助詞] 명 체언의 뒤에 붙어, 그 말과 다른 말과의 문법적인 관계를 나타내거나 특별한 뜻을 더해 주는 품사.

조사[4] [弔詞/弔辭] 명 죽은 사람을 슬퍼하고 생전의 업적을 기려서 조상(弔喪)의 뜻을 나타내는 글.

33) 원어의 유형에서 2개 이상의 언어가 결합된 경우 '+' 기호로 표시하였으며, 원어의 결합 순서는 고려하지 않고 '고유어, 한자어, 외래어' 순으로 결합 유형을 제시하였다.

34) 이운영(2002)은 『표준국어대사전』 주표제어 440,594개를 대상으로 원어의 유형을 분석한 결과, 고유어 25.2%, 한자어 57.3%, 외래어 5.5%, 고유어+한자어 8.3%, 고유어+외래어 0.3%, 고유어+한자어+외래어 0.2%, 한자어+외래어 3.3%의 비율을 보인다고 밝히고 있다. 이 가운데 한자어가 포함된 표제어를 모두 포함할 경우 69%에 이르는데, 이 분석 결과는 『고려대 한국어대사전』의 원어 유형별 빈도와 매우 유사하게 나타나고 있다. 다만, '고유어+한자어'의 비율이 『고려대 한국어대사전』에 비해 매우 낮게 나타나고 있다. 이는 『표준국어대사전』에서 명사에 '-하다/-되다', '-이/-히' 등의 접사가 붙은 표제어를 부표제어로 처리한 결과 때문으로 추측된다.

조사[5] [釣師] 명 낚시를 가지고 고기잡이를 하는 사람.
조사[6] [照射] 명 ①광선이나 방사선 따위를 비추어 쬠.
조사[7] [曹司] 명 ①「역사」 조선 시대, 정삼품의 문신으로 임명한 오위장(五衛將) 두 사람을 이르던 말.
조사[8] [朝士] 명 조정에서 벼슬살이를 하는 신하.
조사[9] [措辭] 명 시가(詩歌)나 문장(文章)을 지을 때, 문구를 선택하거나 배치하는 일.
조사[10] [粗沙/粗砂] 명 굵은 모래. 대체로 지름이 1~0.5밀리미터인 모래를 이르는 말이다.
조사[11] [早死] 명 젊은 나이에 죽음.
조사[12] [照査] 명 둘 이상의 사물을 서로 대조하여 조사함.
조사[13] [造士] 명 ①학문에 통달한 사람.
조사[14] [凋謝] 명 ①초목의 잎이 시들어 떨어짐.
조사[15] [徂謝] 명 ①목숨이 끊어져 이 세상을 하직함.
조사[16] [朝事] 명 ①이른 아침에 지내는 제사.
조사[17] [朝仕] 명 예전에, 아랫자리의 벼슬아치가 윗자리의 벼슬아치를 아침마다 뵈던 일.
조사[18] [朝使] 명 조정의 사신.
조사[19] [朝謝] 명 고려 시대, 벼슬아치에 임명된 사람에게 주던 사령장(辭令狀).
조사[20] [朝辭] 명 예전에, 새로 임명된 관리가 부임하거나 외국의 사신이 떠나기에 앞서 임금께 하직 인사를 드리는 일을 이르던 말.
조사[21] [繰絲] 명 누에고치에서 실을 뽑아냄.
조사[22] [藻思] 명 글을 잘 짓는 재주.
조사[23] [助事] 명 이전에, 장로교에서 목사나 선교사를 도와 전도하는 교직(敎職)이나 그 일을 맡은 사람을 이르던 말.
조사[24] [助射] 명 활쏘기 시합에서, 둘 이상의 사람이 맞힌 화살 수가 서로 같을 때 다시 한 순을 더 쏘아서 승부를 결정함.
조사[25] [助辭] 명 한문에서, 어구나 문장의 끝에 붙어 단정이나 경탄, 의문, 반어 따위의 뜻을 나타내며 끝맺음을 도와주는 구실을 하는 말.
조사[26] [弔使] 명 남의 명령이나 부탁을 받고 조문(弔問)을 하러 가는 사자(使者).
조사[27] [詔使] 명 ①예전에, 중국 천자의 조서(詔書)를 가지고 온다는 뜻으로, 중국에서 온 사신을 이르던 말.
조사[28] [造寺] 명 절을 지음.
조사[29] [釣絲] 명 낚싯대에 낚싯바늘을 매다는, 질기고 가는 줄.

29개의 동형이의어를 가지는 '조사'의 경우 원어 정보의 한자를 통해서 모두 구별이 되고 있는데, 이처럼 동형이의어에서 한자어의 원어 정보를 통한 식별률이 높은 이유는 표의 문자인 한자의 특성과 원어 정보에서 한자가 차지하는 높은 비

중 때문이다. 그렇다면 실제 사전에서 한자어나 외래어의 유형에 따른 동형이의어의 식별률이 어느 정도인지 살펴보도록 하겠다. 다음의 표 11은 원어 정보의 유형에 따른 동형이의어의 발생 비율과 원어 정보를 이용한 동형이의어의 식별률에 대한 통계로서, 이를 통해서 원어의 유형별 동형이의어의 구별 능력을 구체적으로 확인할 수 있다.

표 11 원어 유형별 동형이의어 식별률[35]

원어의 유형	원어 유형별 표제어 수	원어 유형별 동형이의어 수	원어 유형별 동형이의어 비율(%)	원어를 이용한 동형이의어 식별률		
				식별 표제어 수	미식별 표제어 수	식별률(%)
고유어	93,193	15,128	16.23	0	15,128	0.00
한자어	203,931	62,875	30.83	59,713	3,162	94.97
외래어	16,624	1,211	7.28	757	454	62.51
고+한	63,601	13,397	21.06	13,081	316	97.64
고+외	1,129	4	0.35	0	4	0.00
고+한+외	463	4	0.86	4	0	100.00
한+외	7,948	22	0.28	14	8	63.64

표 11에서 '원어 유형별 동형이의어 비율'은 한자어나 외래어 등의 원어 유형별 표제어 그룹에서 발생된 동형이의어 비율로서, 한자어가 30.83%, 고유어+한자어가 21.06%, 고유어가 16.23%, 외래어가 7.28% 순으로 나타난다. 나머지 고유어+외래어, 고유어+한자어+외래어, 한자어+외래어의 경우에는 동형이의어 비율이 모두 1% 미만으로 매우 적게 나타나고 있다. 이는 우리말에서 한자어가 차지하는 비율이 높기도 하지만, 동시에 동형이의어 또한 많이 발생시키고 있음을 나타내는 것이다.

'원어를 이용한 동형이의어 식별률'을 보면 고유어+한자어+외래어가 100%, 고유어+한자어가 97.64%, 한자어가 94.97%, 한자어+외래어가 63.64%, 외래어가 62.51% 순으로 나타난다. 동형이의어 수가 4개뿐인 고유어+외래어, 고유어+한자어+외래어를 제외하고 보면, 앞에서 살펴본 '원어 유형별 동형이의어 비율'과 달리 한자어나 고유어+한자어 부류가 외래어의 부류보다 원어를 이용한 동형이의어의 식별률이 높게 나타나고 있다. 다시 말하자면, 한자어는 동형이의어의 발생 비율이 비교적 높으나 원어를 통해서 식별되는 비율이 매우 높고, 외래어의 경우 원어

35) 원어의 유형에서 '고+한', '한+외'와 같이 한음절로 표시된 항목들이 있는데, '고'는 고유어, '한'은 한자어, '외'는 외래어를 줄여서 표기한 것이다.

를 이용한 동형이의어의 식별률은 한자어에 비해 비교적 낮으나, 동형이의어를 발생시키는 비율이 매우 낮게 나타나는 특징을 보인다.

지금까지 『고려대 한국어대사전』을 대상으로 원어 정보에 대한 유형별 통계와 동형이의어의 식별률에 대한 통계를 살펴본 결과, 한자어나 외래어 등 원어의 유형별 동형이의어 식별률이 매우 높게 나타나고 있음을 확인할 수 있었다.

표 12 원어 정보의 동형이의어 식별률

원어 정보 표제어 수	원어 정보 동형이의어 수	동형이의어 식별 표제어 수	동형이의어 미식별 표제어 수	식별률(%)
293,698	78,950	75,006	3,944	95.00

그러면 마지막으로 원어 정보가 수록된 모든 표제어를 대상으로 동형이의어의 식별률을 살펴보도록 하겠다. 위의 표 12는 원어 정보가 있는 293,698개의 표제어를 대상으로 동형이의어의 식별률을 분석한 결과이다. 원어 정보의 동형이의어 수는 전체 표제어 386,899개를 대상으로 했을 때의 동형이의어 99,499개 보다 20,549개 적은 78,950개로 나타나고 있으며, 이 가운데 원어 정보를 통해 식별되는 표제어 수는 75,006개로 95.00%의 높은 식별률을 보이고 있다.

3.1.2 발음 정보

『고려대 한국어대사전』의 전체 표제어 가운데 발음 정보가 표시되어 있는 표제어는 153,348개로, 이는 전체 표제어 386,889개의 40% 정도에 해당된다. 다음의 표 13은 발음 정보가 수록된 표제어를 품사별로 분석한 것이다. 발음 정보가 수록된 비율을 품사별로 살펴보면, 동사가 56.48%로 가장 높고, 형용사 47.50%, 다품사 47.40%, 명사 43.34%, 수사 33.49%, 관형사 31.77% 등의 순으로 나타난다. 조사나 어미의 경우 다른 품사에 비해 상대적으로 발음 정보의 수록 비율이 낮은 편이나, 전체적으로 발음 정보는 품사의 분류와 관계없이 골고루 나타나고 있음을 알 수 있다. 다만, 무품사의 경우 발음 정보 수록 비율이 극히 낮게 나타나고 있는데, 이는 『고려대 한국어대사전』에서 구 표제어와 외래어에 발음 정보를 원칙적으로 제시하지 않기 때문이다.

표 13 발음 정보가 있는 표제어 통계[36)]

품사	전체 표제어	발음 정보	비율(%)
명사	253,303	109,783	43.34
대명사	380	81	21.32
수사	215	72	33.49
동사	54,921	31,020	56.48
형용사	12,271	5,829	47.50
관형사	192	61	31.77
부사	13,841	3,716	26.85
감탄사	675	119	17.63
조사	338	8	2.37
어미	1,168	65	5.57
접사	555	59	10.63
무품사	45,838	22	0.05
다품사	3,192	1,513	47.40
합계	386,889	152,348	39.38

표제어에 발음 정보를 표시하는 것은 발음이 표기와 다르게 소리나는 경우인데, 위의 통계를 보면 우리말의 40% 정도가 표기와 발음이 다르게 나타나고 있음을 알 수 있다. 이 가운데 두 개 이상으로 발음되는 경우는 22,224개이고, 긴소리로 발음되는 경우는 76,425개로 확인되고 있다. 그러면 전체 표제어에 40% 정도 수록되어 있는 발음 정보를 이용한 동형이의어 식별률은 어느 정도인지 다음의 표 14를 통해서 살펴보도록 하겠다.

표 14 발음 정보의 동형이의어 식별률

발음 정보 표제어 수	발음 정보 동형이의어 수	동형이의어 식별 표제어 수	동형이의어 미식별 표제어 수	식별률(%)
152,348	34,046	2,211	31,835	6.50

표 14는 발음 정보가 수록된 표제어 그룹을 대상으로 한 분석 결과이다. 발음 정보가 있는 152,348개의 표제어에서 동형이의어 수는 34,046개가 나타나고 있으며, 이 가운데 발음 정보를 통해서 구별되는 표제어는 2,211개이다. 발음 정보를 이용한 식별률은 6.50%로 매우 낮은데, 이는 우리말에서 표기가 같은 경우 93.50% 어휘들은 발음까지 동일하게 나타난다는 것을 의미하기도 한다.

36) 도원영, 차준경. 2009. 전게서.

3.1.3 조어 방법

『고려대 한국어대사전』에는 구 표제어를 제외한 모든 복합어에 대해서 그 조어 형태에 따른 형태소 분석 정보를 싣고 있다. 조어 방법에 따른 표제어의 수는 다음의 표 15와 같이, 단일어 169,700개, 복합어 172,697개, 구 표제어 44,492개로 나타나고 있다.

표 15 단어의 조어 방법에 따른 표제어 수

구분	표제어 수	비율(%)
단일어	169,700	43.86
복합어	172,697	44.64
구 표제어	44,492	11.50
총계	386,889	100.00

위의 표 15에서 복합어는 합성어와 파생어를 모두 포함한 것으로, 조어 방법에 따른 비율을 보면 단일어와 복합어가 각각 43.86%와 44.64%로 매우 유사하게 나타나고 있다. 그러면 우리말에 나타나는 동형이어어가 조어 방법에 의해 어느 정도 구별되는지 살펴보도록 하겠다.

표 16 조어 방법의 동형이의어 식별률

전체 표제어 수	전체 동형이의어 수	식별 표제어 수	미식별 표제어 수	식별률(%)
386,889	99,499	5,213	94,289	5.24

위의 표 16은 『고려대 한국어대사전』에 수록된 전체 표제어를 대상으로 '단일어, 복합어, 구 표제어'의 구분 정보를 주었을 때 동형이의어의 식별률을 나타낸 것이다. 전체 99,499개의 동형이의어 가운데 조어 방식에 따라 식별되는 표제어는 5,213개이며, 식별률은 5.24%로 낮게 나타나고 있다. 이는 우리말의 단일어와 복합어 사이에 발생되는 동형이의어가 많지 않다는 것을 의미하기도 하는데, 실제 조어 방식에 따른 동형이의어의 수를 보면, 단일어가 77,045개로 전체 동형이의어의 77.43%를 차지하고, 복합어는 22,304개로 22.42%로 나타나고 있다.

3.1.4 품사 정보

『고려대 한국어대사전』에 수록된 표제어의 품사별 통계를 살펴보면, 표 17과 같이 명사 66.47%, 동사 14.20%, 구 표제어 11.67%, 부사 3.58%, 형용사 3.17% 순으로 나타나고 있으며, 나머지 품사의 경우 모두 1% 미만의 비율을 보이고 있다. 구 표제어는 '강화 플라스틱', '정보 공개법'과 같이 둘 이상의 단어가 모여 하나의 문장 성분으로 사용되는 것으로 『고려대 한국어대사전』에 수록된 45,163개의 구 표제어는 모두 명사구에 해당된다. 따라서 전체 표제어 가운데 명사나, 명사의 성격을 가지는 표제어가 전체 표제어의 80% 가까이 차지하고 있다.[37]

표 17 품사에 따른 표제어 통계[38]

품사	표제어 수	비율(%)
명사	253,303	65.47
대명사	380	0.10
수사	215	0.06
동사	54,921	14.20
형용사	12,271	3.17
관형사	192	0.05
부사	13,841	3.58
감탄사	675	0.17
조사	338	0.09
어미	1,168	0.30
접사	555	0.14
무품사(준말)	675	0.17
무품사(구 표제어)	45,163	11.67
다품사어	3,192	0.83
합계	386,889	100.00

그러면 『고려대 한국어대사전』에서 품사를 통한 동형이의어의 구별이 어느 정도 되고 있는지 살펴보도록 하겠다. 기준은 전체 표제어를 대상으로 9품사와 어미, 접사 이외에 무품사(준말), 무품사(구 표제어), 다품사까지 포함하였다. 품사 분류에 따른 동형이의어의 식별 결과는 표 18과 같이 5,579개의 표제어가 구분되고 있

37) 품사의 항목으로 설정되어 있는 다품사어는 하나의 단어가 2개 이상의 문장 품사로 쓰이는 것으로 '갖가지, 교육적'과 같이 명사와 관형사의 성격을 동시에 가지거나, '감사하다, 마르다'와 같이 동사와 형용사의 성격을 함께 가지는 단어를 말한다.

38) 도원영, 차준경. 2009. 전게서.

는데, 이는 전체 동형이의어 99,499개의 5.61%에 해당된다. 품사 정보를 통한 동형이의어 식별률이 비교적 낮게 나타나는 것은 우리말의 대부분이 명사와 동사에 집중되어 있기 때문이다. 따라서 품사를 통한 동형이의어를 구별하고자 할 경우 품사의 하위 범주까지 식별 표지로 사용할 필요가 있다.

표 18 상위 품사의 동형이의어 식별률

전체 표제어	동형이의어 수	식별 수	미식별 수	식별률(%)
386,889	99,499	5,579	93,923	5.61

『고려대 한국어대사전』은 9품사와 접사, 어미의 품사에 덧붙여 '일반 명사, 고유 명사, 자립 명사, 의존 명사, 인칭 대명사, 지시 대명사, 자동사, 타동사, 보조 동사, 보조 형용사, 접두사, 접미사, 선어말 어미, 종결 어미, 연결 어미, 전성 어미'와 같이 16개의 품사 하위 범주를 두고 있는데, 이들 하위 범주의 품사 분류까지 이용하여 동형이의어를 구분할 경우 표 19에서 보는 바와 같이 전체 동형이의어의 12.9%인 12,837개의 표제어가 구별된다.

표 19 하위 품사의 동형이의어 식별률

전체 표제어	동형이의어 수	식별 수	미식별 수	식별률(%)
386,889	99,499	12,837	86,665	12.90

품사의 하위 범주를 고려할 경우 동형이의어의 구별 비율이 2배 이상 증가한 것은 전체 표제어의 상당 부분을 차지하는 명사와 동사가 하위 품사 범주별로 다양하게 분산됐기 때문이다.

3.1.5 용언의 불규칙 활용 정보

『고려대 한국어대사전』에는 동사 54,921개, 형용사 12,271개, 동사/형용사 547개가 수록되어 있으며, 이들 용언 가운데 44,048개의 단어가 불규칙 활용을 보이고 있다. 다음의 표 20은 용언의 불규칙 통계인데, 여 불규칙이 42,295개로 전체 불규

칙 용언의 96.02%를 차지하고 있고, 그 다음으로 ㅂ 불규칙이 1,096개, ㄹ 불규칙이 291개, ㅎ 불규칙이 122개 순으로 나타나고 있다.

표 20 용언의 불규칙 활용 통계[39]

불규칙 유형	품사	표제어 수	소계	비율(%)
ㄷ 불규칙	동사	62	62	0.14
ㅂ 불규칙	동사	21	1,096	2.49
	형용사	1,075		
ㅅ 불규칙	동사	69	70	0.16
	형용사	1		
ㅎ 불규칙	형용사	122	122	0.28
러 불규칙	동사	1	19	0.04
	형용사	18		
ㄹ 불규칙	동사	220	291	0.66
	동사/형용사	1		
	형용사	70		
여 불규칙	동사	33,776	42,295	96.02
	동사/형용사	485		
	형용사	8,034		
우 불규칙	동사	1	1	0.00
너라 불규칙	동사	92	92	0.21
합계		44,048		100.00

표 21은 용언의 전체 표제어를 대상으로 불규칙 활용 정보를 이용한 동형이의어 식별 결과를 보인 것으로, 전체 용언 67,739개 가운데 16,898개의 동형이의어가 나타나고 있으며, 이 가운데 불규칙 활용 정보를 이용하여 321개의 동형이의어가 구별되고 있다.

표 21 불규칙 활용 정보의 동형이의어 식별률[40]

용언 수	동형이의어	식별 수	미식별 수	식별률(%)
67,739	16,898	321	16,577	1.90

국어사전의 불규칙 활용 정보를 동형이의어의 식별률은 지금까지 살펴본 미시 정보 가운데 가장 낮은 수치지만, 용언만을 대상으로 동형이의어를 구별하는 기준으로 삼을 수 있다는 점에서 유용하게 활용될 수 있다.

39) 도원영, 차준경. 2009. 전게서.
40) 용언의 표제어 수가 표 17의 동사와 형용사의 표제어 수의 합보다 큰 것은, '동사/형용사' 형태의 다품사 547개가 포함되었기 때문이다.

3.1.6 방언, 비표준어

『고려대 한국어대사전』은 단어와 접사, 어미, 구 등 어휘의 문법적 범위의 제한 없이 약 3만여 개의 방언과 비표준어를 표제어로 싣고 있다. 『고려대 한국어대사전』에 수록된 방언과 비표준어는 표 22에서 보는 바와 같이, 방언 23,473개, 비표준어 6,178개로 나타나고 있다.

표 22 표준어, 방언, 비표준어의 표제어 수[41]

품사	표준어	방언	비표준어
명사	231,837	17,473	3,993
대명사	273	97	10
수사	154	51	10
동사	51,873	2,399	649
형용사	10,480	1,327	464
부사	11,905	1,275	661
관형사	140	42	10
감탄사	491	108	76
조사	210	110	18
어미	604	416	148
접사	545	8	2
준말	648	14	13
무품사	45,052	11	109
다품사	377	142	15
합계	357,238	23,473	6,178

방언과 비표준어는 대부분 고유어로서 원어, 발음, 품사와 같은 미시 정보로 식별되지 못하는 경우가 매우 많다. 표 23은 방언과 비표준어의 원어 유형별 표제어를 나타낸 것으로, 방언은 99.3%, 비표준어는 85%가 고유어로 나타나고 있다.

표 23 방언/비표준어의 원어 유형별 표제어 수

원어별 유형	방언		비표준어	
	표제어 수	비율(%)	표제어 수	비율(%)
고유어	23,312	99.31	5,257	85.09
한자어	31	0.13	429	6.94
외래어	1	0.00	97	1.57
고유어+한자어	129	0.55	377	6.10

41) 도원영, 차준경. 2009. 전게서.

고유어+외래어	0	0.00	9	0.15
한자어+외래어	0	0.00	8	0.13
고유어+한자어+외래어	0	0.00	1	0.02
합계	23,473	100.00	6,178	100.00

표를 통해서도 확인되지만 원어, 조어 방법, 발음, 품사, 용언의 활용과 같은 미시정보로 식별되지 않은 방언과 비표준어에 대한 구별 방법이 필요한데, 이런 경우 대응되는 표준어를 식별 정보로 사용할 수 있다. 방언의 경우 대응되는 표준어가 없는 경우 사용 지역 정보를 표준어 대신 주는 방법이 있다.

표 24 방언 및 비표준어의 동형이의어 식별률

구분	전체 표제어	유형별 표제어 수	동형이의어 수	식별 표제어 수	미식별 표제어 수	식별률(%)
방언	386,889	23,473	3,455	3,407	48	98.61
비표준어	386,889	6,178	325	279	46	85.85
방언+비표준어	386,889	29,651	4,238	4,138	100	97.64

표 24와 같이 방언과 비표준어의 경우 대응되는 표준어를 기준으로 동형이의어를 구별할 경우, 방언은 98.61%, 비표준어는 85.85%로 매우 높은 식별률을 보이고 있다. 방언과 비표준어를 묶어서 처리할 경우 동형이의어의 수가 약간 증가하지만 전체적인 식별률 또한 97.64%로 높게 나타남을 볼 수 있다.

3.2. 미시정보별 식별률의 비교 분석

지금까지 『고려대 한국어대사전』을 대상으로 미시정보별 동형이의어의 식별률을 살펴보았는데, 여섯 가지 미시정보에 대한 동형이의어 식별률의 결과를 종합하면 다음의 표 25와 같다. 각 미시정보의 동형이의어 식별률은 미시정보가 수록된 표제어 단위로 산출하였으나, 품사와 조어 방법은 사전에 수록된 모든 표제어에 해당 미시정보가 수록되어 있기 때문에 전체 표제어를 대상으로 하였으며, 활용 정보의 경우에는 사전에 수록된 모든 용언을 대상으로 하였다.

표 25에서 '미시정보별 표제어 수'는 전체 표제어에서 미시정보가 있는 표제어의 개수이고, '미시정보별 동형이의어 수'는 '미시정보별 표제어 수'에서 동형이의

어 관계가 나타난 표제어 개수를 말한다. ‘동형이의어 식별 표제어 수’는 ‘미시정보별 동형이의어 수’에서 미시정보를 통해 식별되는 표제어 개수를 나타낸 것이다.

표 25 미시정보의 동형이의어 식별률(미시정보별)

미시정보 유형	미시정보별 표제어 수	미시정보별 동형이의어 수	동형이의어 식별 표제어 수	동형이의어 미식별 표제어 수	식별률(%)
원어 정보	293,698	78,950	75,006	3,944	95.00
발음 정보	152,348	34,046	2,211	31,835	6.50
조어 방법	386,889	99,499	5,213	94,289	5.24
품사 정보	386,889	99,499	12,837	86,665	12.90
활용 정보	67,739	16,898	321	16,577	1.90
방언+비표준어	29,651	4,238	4,138	100	97.64

여섯 가지 미시정보의 동형이의어 식별률을 보면, 방언/비표준어 97.64%, 원어 95.00%, 품사 12.90%, 발음 6.50%, 조어 방법 5.24%, 활용 정보 1.90% 순으로 나타나고 있는데, 특히 방언/비표준어와 원어 정보를 이용한 식별률이 다른 미시정보에 비해 월등히 높게 나타나고 있다.

동형이의어 식별 표제어 수를 기준으로 보면 원어 정보 75,006개, 품사 정보 12,837개, 조어 방법 5,213개, 방언/비표준어 4,138, 발음 정보 2,211개, 활용 정보 211개 순으로, 원어 정보를 이용한 식별 표제어 수가 나머지 미시정보에 비해 월등히 높은 결과를 보여주고 있다. 방언/비표준어는 미시정보별 식별률이 매우 높지만, 동형이의어 식별 표제어 수에서는 품사나 조어 방법에 비해 적게 나타나는데, 이는 구별 대상이 되는 표제어의 개수가 적기 때문이다. 즉, 품사 정보와 조어 방법은 전체 표제어 386,889개를 대상으로 구별한 결과이고, 방언+비표준어는 방언과 비표준어에 해당되는 표제어 29,651개만을 대상으로 했기 때문이다.

다음으로 『고려대 한국어대사전』의 전체 표제어를 대상으로 동형이의어 식별률을 살펴보도록 하겠다. 표 26은 전체 386,889개의 표제어를 대상으로 각 미시정보별 동형이의어 식별률을 분석한 결과이다. 이는 『고려대 한국어대사전』이라는 어휘 군집에 대한 미시정보의 실질적인 식별률을 나타내는 것으로, 앞의 표 25와 비교해 보면 동형이의어 식별 표제어 수와 식별률에서 큰 차이를 보이고 있다. 이는 전체 표제어 가운데 미시정보가 없는 표제어가 미시정보가 있는 표제어와 동형이의어 관계에 있을 때, 미시정보의 유무의 차이로 식별되는 경우가 추가됐기 때문

이다. 즉, 원어 정보의 경우 동형이의어 식별 표제어 수가 75,006개에서 80,428개로 5,000여 개 이상 증가하였는데, 이는 원어 정보가 없는 표제어 가운데 원어 정보가 있는 표제어와 그 증가한 표제어 수만큼 동형이의어 관계를 가지고 있었기 때문이다. 다만, 조어 방법과 품사 정보의 경우 표 25와 식별률이 동일하게 나타나고 있는데, 이는 두 미시정보가 표 25에서도 전체 표제어를 대상으로 분석했기 때문이다.

표 26 미시정보의 동형이의어 식별률(전체 표제어)

구분	전체 표제어수	전체 동형이의어 수	동형이의어 식별 표제어 수	동형이의어 미식별 표제어 수	식별률(%)
원어 정보	386,889	99,499	80,428	19,071	80.83
발음 정보	386,889	99,499	14,827	84,672	14.90
조어 방법	386,889	99,499	5,210	94,289	5.24
품사 정보	386,889	99,499	12,834	86,665	12.90
활용 정보	386,889	99,499	323	99,176	0.32
방언+비표준어	386,889	99,499	10,325	89,174	10.38

전체 표제어에 대한 동형이의어 식별률을 살펴보면, 원어 정보 80.83%, 발음 정보 14.90%, 품사 정보 12.90%, 방언/비표준어 10.38%, 조어 방법 5.24%, 활용 정보 0.32% 순으로 나타나고 있다. 대체적으로 원어 정보의 경우 표 25와 큰 차이를 보이지 않지만, 발음 정보와 방언+비표준어의 경우 많은 차이를 보이고 있다. 발음 정보의 경우 식별률이 6.50%에서 14.90%로 두 배 이상 높아졌으나, 방언/비표준어의 경우에는 이와 반대로 식별률이 97.64%에서 10.38%로 대폭 낮아진 것을 볼 수 있다. 이러한 차이점은 동형이의어의 식별 표제어 수를 비교할 경우에도 볼 수 있는데, 발음 정보의 경우 2,211개에서 14,827개로 12,616개가 증가하였으며, 방언/비표준어의 경우에도 4,138개에서 10,325개로 6,187개가 증가하였다.

흥미로운 점은 미시정보별 동형이의어 식별 표제어 수에서 대상 어휘 군집의 변화에 따라 분석 결과가 큰 폭으로 변하는 미시정보와 그렇지 않은 미시정보로 나뉜다는 점이다. 원어 정보, 조어 방법, 품사 정보, 활용 정보의 경우 전체 표제어를 대상으로 한 분석 결과가 미시정보별 분석 결과와 별 차이가 없거나 완전히 일치하지만, 발음 정보와 방언/비표준어의 경우 두 개의 분석 결과에 많은 차이를 보인다.

표 25에서 매우 높은 식별률을 보인 방언/비표준어의 경우 전체 표제어를 대상으로 한 식별률이 10.38%로 낮아지기는 했지만, 전체 표제어에서 차지하는 방언/비표준어의 비율이 7.66%인 것을 감안하면 매우 높은 수치이며, 이는 전체 표제어에서의 동형이의어 식별 수가 두 배 이상 증가한 것으로도 확인할 수 있다.

발음 정보의 경우 전체 표제어를 대상으로 분석한 식별률이 14.90%로 미시정보별 식별률보다 두 배 이상 증가하였는데, 이는 미시정보별 동형이의어 수 34,046개에서 전체 표제어의 동형이의어 수 99,499개로 증가한 비율과 유사하다.

용언의 활용 정보는 전체 표제어를 대상으로 한 동형이의어 식별 표제어 수가 323개로 미시정보별 동형이의어 식별 표제어 수인 321개와 거의 변화가 없는데, 이는 용언의 표제어가 다른 품사와 형태적으로 거의 중복되지 않음을 나타낸다고 할 수 있다.

지금까지 미시정보별 식별률을 살펴본 결과, 전체 표제어에 대한 동형이의어 식별률이 높은 미시정보는 '원어 정보, 발음 정보, 품사 정보, 방언/비표준어, 조어 방법, 활용 정보' 순으로 확인되었다. 표 27은 전체 표제어에 대한 동형이의어의 식별 결과를 위의 여섯 가지 미시정보 순으로 누적하여 나타낸 것이다. 즉 미시정보 항목의 '+발음 정보'는 원어 정보와 발음 정보를 함께 적용하여 동형이의어를 구별한 결과이며, 이후의 미시정보 또한 앞의 미시정보들과 조합한 것이다. 따라서 마지막 '+활용 정보'의 통계 결과는 '원어 정보 + 발음 정보 + 품사 정보 + 방언/비표준어 + 조어 방법 + 활용 정보'까지를 모두 조합하여 동형이의어를 구별한 것이다.

표 27 미시정보의 동형이의어 누적 식별률

미시 정보	동형이의어 누적 식별 수	증가 수	증가 비율(%)	미식별 수	동형이의어 누적 식별률(%)
원어 정보	80,428	80,428	80.83	19,071	80.83
+발음 정보	82,118	1,690	1.70	17,381	82.53
+품사 정보	87,136	5,018	5.04	12,363	87.57
+방언/비표준어	92,621	5,485	5.52	6,878	93.09
+조어 방법	92,923	302	0.30	6,576	93.39
+활용 정보	92,925	3	0.00	6,574	93.39

동형이의어의 누적 식별 수를 보면, 원어 정보를 통해서 80,428개의 동형이의어가 구별되고, 이후 발음 정보 1,690개, 품사 정보 5,018개, +방언/비표준어 5,485개, +조어 방법 302개, +활용 정보 3개로 식별 표제어 수가 증가됨을 볼 수 있다. 따라서 여섯 가지 미시정보를 모두 조합한 동형이의어 누적 식별 수는 +활용 정보 항목의 92,925개이고, 누적 식별률은 93.39%이다. 그런데 동형이의어 누적 식별률을 살펴보면 원어 정보를 통해서 80.83%의 동형이의어가 식별된 이후, +발음 정보 82.53%, +품사 정보 87.57%, +방언/비표준어 93.09%로, 1.7%에서 5.5%까지 식별률이 꾸준히 증가하지만, +조어 방법과 +활용 정보에서는 단 0.3%만 증가하는 데 그치고 있다. 이는 조어 방법과 활용 정보를 통해서 식별될 수 있는 동형이의어들이 이전의 미시정보들로 이미 식별되었음을 나타내는 것이다. 다시 말하면 국어사전에서 미시정보를 이용하여 동형이의어를 식별하고자 할 때, 원어 정보, 발음 정보, 품사 정보, 방언/비표준어는 유용하게 활용될 수 있지만, 조어 방법과 활용 정보는 식별자로서의 효용성이 매우 낮다고 할 수 있다. 특히 활용 정보는 표 25의 미시정보별 식별률에서 1.90%에 그치고 있고, 동형이의어의 식별 표제어 수도 321개로 밖에 되지 않아 동형이의어의 식별자로 사용하기에 적합하지 않음을 알 수 있다.

지금까지의 분석 결과를 통해서 살펴보면 표 25에서 미시정보별 식별률이 높았던 원어 정보와 방언+비표준어가 전체 표제어에 대한 식별률도 높게 나타나고 있음을 확인할 수 있다. 따라서 동형이의어의 구별을 위해 사용할 미시정보를 선정할 때, 전체 표제어에 해당 미시정보의 수록 비율과 미시정보별 변별성을 중요한 요소로 고려해야만 한다. 또한 이상의 여섯 가지 미시정보를 조합하여 약 40만개의 어휘를 수록한『고려대 한국어대사전』의 동형이의어를 구별한 결과 전체 동형이의어의 93.39%가 식별되고 있음을 확인하였는데, 이는 국어사전의 미시정보를 조합하여 동형이의어를 식별하는 방법이 매우 유효하다고 볼 수 있으며, 어원 정보나 본말-준말, 원어-약어, 큰말-작은말, 센말-거센말-여린말, 이형태 등과 같은 관련어 정보를 식별 정보로 추가할 경우 전체 식별률을 더 높일 수 있을 것으로 기대된다.

4. 미시정보별 동형이의어 구별 방안

이 절에서는 동형이의어의 식별을 위해 선정한 여섯 가지 미시정보가 국어사전에 어떻게 기술되고 있는지 검토하고, 동형이의어의 식별자로 사용되기 위해서 고려해야할 점들을 살펴보고자 한다. 우선 여섯 가지 미시정보가 어떻게 동형이의어를 구별하고 있는지 국어사전에 수록된 구체적인 사례를 통하여 확인한다. 다음으로 제1장에서 밝힌 네 개의 국어사전을 대상으로 해당 미시정보의 표기나 기술의 특성을 비교하여 각 사전마다의 유사점과 차이점을 살펴보고, 해당 미시정보를 동형이의어의 식별자로 사용하고자 할때 고려해야 할 문제점과 해결 방안을 제시하고자 한다.

4.1 원어 정보

국어사전에서 원어 정보는 표제어의 어원과 관련된 것으로 한자어의 경우 해당 한자를, 그 외 외래어의 경우 로마자 전사를 표제어 뒤에 괄호로 표기하는 것이 일반적이다. 원어 정보는 표제어의 본래 표기를 제시함으로써 표제어의 의미를 밝히는 것이 주된 기능이지만, 한자나 로마자와 같은 문자의 표기를 통해서 표제어를 식별하는 역할도 적지 않다.[42]

다음은 『표준국어대사전』에 수록된 '가격'의 원어와 뜻풀이 정보를 보인 것이다.

ㄱ. **가격**[1](加擊) 명 손이나 주먹, 몽둥이 따위로 때리거나 침.
ㄴ. **가격**[2](家格) 명 =문벌(門閥).
ㄷ. **가격**[3](價格) 명 ①물건이 지니고 있는 가치를 돈으로 나타낸 것.

(ㄱ)과 (ㄷ)은 구어에서 사용 빈도가 높은 단어로서 원어 정보를 표기하지 않았을 경우에도 표기와 의미를 연결하여 이해하는 데 큰 어려움이 없지만, (ㄴ)과 같이 사용 빈도가 낮은 말의 경우 원어 정보가 없는 상태에서 '가격'이라는 단어를 '신분이나 지위'를 나타내는 문벌(門閥)과 연결하기가 쉽지 않다. 또한 어느 정도

42) 전통적으로 국어사전에서는 원어 정보를 표제어와 같은 수준으로 가깝게 이해하고 있으며, 『금성판 국어대사전』과 같이 원어 정보가 다른 동형이의어에 어깨번호를 붙이지 않은 사전도 있다.

한자에 익숙한 사람일 경우에는 원어 정보의 한자만으로도 직관적인 의미 구분이 가능할 수 있어, 동형이의어를 식별하는 데에 있어 원어 정보의 중요성을 알 수 있다.

원어 정보를 동형이의어의 식별자로 사용하기 위해서는 원어 정보의 표기 원칙을 정하는 것이 중요한데, 원어 정보의 위치나 표기 방법 등에서 사전마다 약간의 차이를 보이고 있다. 따라서 각각의 국어사전을 대상으로 원어 정보의 표기 방식과 원어 정보 처리시 고려해야 할 문제들에 대해서 우선적으로 살펴볼 필요가 있다.

국어사전에서 원어 정보는 표제어의 바로 뒤에 오는 것이 일반적이다. 이는 본 논문에서 분석 대상으로 삼은 『금성판 국어대사전』, 『표준국어대사전』, 『고려대 한국어대사전』에서 취하고 있는 방식이다. 하지만, 『우리말큰사전』의 경우에만 다음과 같이 특별히 원어 정보를 뜻풀이 맨 뒤에 위치시키고 있다.

가격[1] (이) 치거나 때림. 〔加擊〕
가격[2] (이) =문벌. 〔家格〕
가격[3] (이) 노래의 격식이나 품격. 〔歌格〕
가격[4] (이) ①=값. ②=금. 〔價格〕

『우리말큰사전』에서 원어 정보를 표제어와 거리를 두어 표기한 이유에 대해서 구체적으로 밝히고 있지는 않으나, 이 사전을 편찬한 한글학회의 한글 전용주의 원칙에 따른 것으로 이해할 수 있다. 이는 한글 표기는 같지만 원어 정보가 다른 동형이의어에 어깨번호를 붙이지 않는 『금성판 국어대사전』과는 대조적이라 할 수 있다. 하지만 동형이의어의 식별자로 사용하는 데 있어서 원어 정보의 위치가 중요한 문제는 아니다. 원어 정보를 동형이의어의 식별자로 이용하는 데 있어 중요하게 살펴야 할 점은 원어 정보의 표기 방법이다. 다시 말하면, 중국어나 일본어 또는 기타 외래어의 원어를 표기하는 데에 있어 언어 본래의 문자로 표기할 것인지, 아니면 한자나, 로마자 전사로 표기할 것인지, 또는 고유어가 포함된 합성어에서 고유어의 처리 방법 등을 확인하는 것이 중요하다.

4.1.1 한자어

한자어의 경우 중국식 한자어나 일본식 한자어 또는 한국식 한자로 나뉠 수 있는데, 다음 표 28에서 보이는 (ㄱ)은 중국식 한자어이고, (ㄴ)은 일본식 한자어, 그리고 (ㄷ)은 우리나라에서만 사용되는 한자어이다.[43] 여기에 (ㄹ)과 같이 우리나라에서 만들어진 이두식 한자어와 (ㅁ), (ㅂ)과 같이 본래 중국에서 다른 언어를 차음한 뒤 우리말로 들어온 단어도 있는데, 국어사전에서는 한자로 이들의 원어를 밝히는 것이 일반적이다.

표 28 한자어의 원어 표기

구분	표제어	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
ㄱ	대궐	大闕	大闕	大闕	大闕
ㄴ	견습	見習	일.見習	見習	見習
ㄷ	동생	同生	同生	同生	同生
ㄹ	사음	舍音	舍音	舍音	국.舍音
ㅁ	모란	*牧丹	<牧丹	牧丹▽	牧丹
ㅂ	보리	*菩提	菩提 <범.bodhi	菩提	@菩提

국어사전에서 한자어의 원어 표기는 네 사전 모두 동일하게 나타나고 있어, 한자어의 원어를 동형이의어의 식별자로 쓰는 데 큰 제약점은 보이지 않는다. 다만, 『우리말큰사전』의 경우 범어의 한자 표기에 있어 해당 어원을 함께 표시하고 있으며, 한자어의 음이 변한 경우에 대해서 '*, <, ▽, @' 따위의 부호를 통해 밝히거나, 일본식 한자어나 한국식 한자어에 대해 그 출처를 표시하고 있다.

표 29 한자어의 원어 표기(2개 이상)

표제어	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
기념	記念·紀念	記念. 紀念	記念/紀念	記念/紀念
감탄	感歎	感歎	感歎/感嘆	感歎/感嘆
표준	標準·表準	表準. 標準	標準	標準/表準
행적	行跡·行績·行蹟	行績. 行蹟. 行跡	行跡/行績/行蹟	行跡/行績/行蹟

43) 홍종선 외. 2009. 전게서. pp.237-238.

한자어의 원어 정보를 통해서 동형이의어를 식별하고자 할 때 문제점은 다음의 표 29와 같이 2개 이상의 한자를 원어로 가지는 표제어에 대한 처리 방법이다. 일반적으로 사전에서는 이와 같은 경우 배열 순서에 대해 일정한 원칙을 정하기 마련이지만, 각 사전마다 배열 원칙에 차이가 있거나 모호한 경우가 많기 때문에, 한자의 배열 순서를 정하여 식별자로 사용하기는 쉽지 않다.

설령 모든 사전의 배열 원칙이 일치한다 하더라도 '감탄'이나 '표준'과 같이 사전에 따라 원어를 하나만 표기하는 경우가 있어, 원어 정보를 이용하여 동형이의어를 식별하기 위해서는 n개의 정보를 비교할 수 있는 데이터 구조가 필요하다.

```
【『금성판 국어대사전』】
<표제항>
        <표제어>감탄</표제어>
        <원어>
                <원어값>感歎</원어값>
        </원어>
</표제항>

【『표준국어대사전』】
<표제항>
        <표제어>감탄</표제어>
        <원어>
                <원어값>感歎</원어값>
                <원어값>感嘆</원어값>
        </원어>
</표제항>
```

위의 데이터 구조는 표제어와 n개의 원어 정보를 XML 형식으로 표현한 것으로, 『금성판 국어대사전』과 『표준국어대사전』의 원어 정보는 1:2의 관계를 가지며, 이들 간의 비교를 통해 일치하는 값이 있을 경우 표제어가 같다고 볼 수 있다. 이와 같은 데이터 구조의 경우 원어의 개수뿐만 아니라 배열 순서에 상관없이 적용할 수 있는 장점이 있다.

4.1.2 중국어, 일본어

중국어와 일본어에서 유입된 외래어의 원어 표기를 살펴보면, 표 30처럼 한자나 로마자 전사 또는 히라가나나 가타가나 등 사전에 따라 원어를 표기하는 방식에 차이를 보이고 있다. 『금성판 국어대사전』에서는 중국어에 대해서 한자만을 사용하여 원어 정보를 표시하고 있으며, 일본어의 경우에는 한자로 쓸 수 있는 말이면 한자를 쓰되 괄호 안에 히라가나를 덧붙여 표시하고, 구미(歐美)에서 받아들인 말이면 가타카나로 표기하고 있다. 『우리말큰사전』의 경우 중국어나 일본어 모두 로마자로만 원어를 표시하고 있으며, 『표준국어대사전』과 『고려대 한국어대사전』에서는 로마자 전사 뒤에 한자를 병기하여 표기하고 있다.

표 30 중국어, 일본어의 원어 표기[44)]

구분	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
베이징	北京	×	Beijing[北京]	Beijing[北京]
마오쩌둥	毛澤東	×	Mao Zedong[毛澤東]	Mao Zedong[毛澤東]
난징조약	南京條約	nanching 條約	Nanjing[南京]條約	Nanjing[南京]條約
도쿄	東京	×	Tokyo[東京]	Tokyo[東京]
도요토미 히데요시	豊臣秀吉 (とよとみひでよし)	×	Toyotomi Hideyoshi [豊臣秀吉]	Toyotomi Hideyoshi [豊臣秀吉]
다다미	일.疊(たたみ)	일.tatami	←일.tatami[疊]	일.@tatami[疊]
빠구	일.バック	<영.back	일.bakku	일.@bakku

동형이의어의 식별자로 원어 정보를 이용하기 위해서는 해당 정보의 표기 방법이 동일해야 하지만, 『금성판 국어대사전』의 경우 다른 사전과 큰 차이를 보이고 있다. 따라서 현재의 원어 정보를 이용하여 네 개의 사전 모두 동일하게 식별자를 표시할 수는 없다. 다만, 『표준국어대사전』과 『고려대 한국어대사전』에서 영문자와 한자를 병기하여 표시함으로써 『금성판 국어대사전』과 『우리말큰사전』 가운데 한 개의 사전은 원어 표기의 통일성을 유지할 수 있다. 즉 『금성판 국어대사전』을 제외한 세 개의 사전은 로마자로 원어를 표기할 수 있으며, 『우리말큰사전』을 제

44) 『우리말큰사전』 항목의 '×'는 표제어가 사전에 수록되지 않음을 표시한 것이다.

외한 세 개의 사전은 한자로 원어로 표기할 수 있다.

그러나 표 30의 '빠꾸'와 같이 서구에서 일본으로 차용된 후 우리말로 들어 온 경우 한자로 표기할 수 없고, 중국어의 경우에도 원어가 간체자(簡體字)인 경우가 있을 수 있기 때문에, 중국어와 일본어에서 유입된 모든 외래어에 대해서 한자로 표기하는 것은 현실적으로 실현되기 어렵다. 그러면 위의 두 가지 방법 중에 로마자로 표기하는 방안이 남게 되는데, 이는 대부분의 국어사전에서 영어나 서구 외래어의 원어 표기를 로마자로 하고 있다는 점에서 일관성을 유지할 수 있는 장점이 있다. 또한 중국어와 일본어의 로마자 표기 방법은 <외래어 표기법> 제4장 제2절의 '동양의 인명, 지명 표기' 규정에서 일부를 제외한 중국과 일본의 인명과 지명을 한국식 한자음으로 읽는 것을 허용하지 않고 있는 점과도 부합한다 할 수 있다.[45)]

4.1.3 영어 및 서구 외래어

국어사전에는 영어나 서구 외래어를 다수 수록하고 있는데, 대부분 로마자로 전사하여 원어를 표기하고 있다. 『금성판 국어대사전』에서는 일러두기에서 "로마자를 사용하는 언어권은 물론이고 비사용권 국가, 러시아·그리스·아랍 등지도 로마자로 표기를 보였다"라고 밝히고 있다. 중국어와 일본어를 제외한 외래어의 원어를 로마자로 표기하는 방식은 『우리말큰사전』, 『표준국어대사전』, 『고려대 한국어대사전』에서도 마찬가지이다.

표 31 서구 외래어의 원어 표기

언어	표제어	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
그리스 어	아고라	그.agora	그.agora	그.agora	그.agora
러시아 어	아지트	러.agitpunkt	러.agitpunkt	러.agitpunkt	러.agitpunkt
라틴 어	글리세롤	glycerol	라.glycerol	glycerol	라.glycerol
아랍 어	라마단	아.Ramaḍān	아.Ramazān	아.Ramaḍān	아.Ramaḍān
히브리 어	시오니즘	Zionism	히.Zionism	Zionism	히.Zionism

45) <외래어 표기법>에서 우리 한자음으로 표기하는 것이 허락되는 경우는, 중국의 과거인 인명과 현재 쓰이지 않는 지명만 해당한다. 일본의 인명과 지명은 원칙적으로 <외래어 표기법>의 '일본어 표기법'을 따라서 표기해야 한다.

위의 표 31은 로마자를 사용하지 않는 외래어의 원어를 로마자로 전사하여 표기한 것이다. 대부분의 국어사전에서 외래어의 원어 표기를 로마자로 일관되게 표기하고 있어, 서구 외래어의 원어를 동형이의어의 식별자로 사용하는 데는 큰 제약점이 보이지 않는다.[46]

4.1.4 고유어

마지막으로 고유어와 결합한 합성어의 원어 표기 방법을 살펴보도록 하겠다. 국어사전에는 고유어에 한자어나 외래어가 결합하여 만들어진 어휘가 상당 부분 수록되어 있는데, 이들 어휘에 대한 원어 표기에서 고유어의 처리 방법이 두 가지로 나뉘고 있다. 하나는 고유어의 음절수에 상관없이 하나의 줄표(—)로 고유어를 표시하는 경우가 있고, 다른 하나는 붙임표(-)를 고유어의 음절수대로 붙여 표시하는 방법이다.

표 32 원어에서 고유어 처리 방법

표제어	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
명주실	明紬—	明紬—	明紬-	明紬-
된장	—醬	—醬	-醬	-醬
고추장	—醬	—醬	--醬	--醬
무거리고추장	—醬	—醬	-----醬	-----醬
나비넥타이	—necktie	—necktie	--necktie	--necktie
누구를 위하여 종은 울리나	—爲—鐘—	—爲—鐘—	---爲--鐘----	---爲--鐘----

표 32에서 보면 『금성판 국어대사전』과 『우리말큰사전』은 줄표를, 『표준국어대사전』과 『고려대 한국어대사전』은 붙임표를 사용하여 고유어를 표시하고 있다. 위의 두 가지 가운데 어떠한 방법을 사용하든지 원어 정보를 이용한 동형이의어의 식별률에는 차이가 없다. 하지만 '누구를 위하여 종은 울리나'와 같이 고유어가 한자어나 외래어와 복잡하게 결합된 표제어의 경우, 『금성판 국어대사전』이나 『우리말큰사전』과 같이 하나의 줄표만으로 여러 음절의 고유어를 대체하게 되면 나머

46) 『우리말큰사전』에서 '라마단'의 로마자 표기 'Ramazān'이 나머지 사전과 다르게 표기된 것은 『우리말큰사전』이 한글 학회의 <한글 맞춤법>(1980)과 조선어 학회의 <외래어 표깃법 통일안)(1941) 따라 편찬되었기 때문이다.

지 원어에 표시된 한자가 고유어의 어떤 음절과 대응되는지 쉽게 파악하기 어려운 점이 있다. 따라서 원어 정보에서 고유어의 대체 부호로 붙임표를 사용하여 고유어의 음절수만큼 표시하는 방법이 가독성의 측면에서 장점이 있다. 또한 줄표를 사용하여 고유어를 표시할 경우 '된장, 고추장, 무거리고추장'과 같이 표제어의 음절수에 따른 차이가 원어 정보를 통해서 전혀 변별되지 못하기 때문에, 고유어의 원어 표기는 붙임표를 사용하는 것이 여러 모로 유리하다 할 수 있다.

지금까지 국어사전에 나타나는 원어 정보의 유형과 동형이의어의 식별을 위한 원어 정보의 기술 방법에 대해 살펴보았다. 그러나 동형이의어의 식별 표지로서 원어 정보를 활용하기 위해서는 원어 정보를 표기하는 한자의 처리 문제가 남아 있다. 한자어는 우리말의 약 60% 정도를 차지하고 있으며, 대사전 규모의 국어사전을 편찬하기 위해서는 대략 7,000여 자의 한자를 사용해야만 한다.[47] 이는 상용한자 4,888자가 포함된 KS X 1001[48] 체계로는 수용할 수 없는 규모이며, 이들 한자를 온전히 표현하기 위해서는 유니코드(Unicode)[49] 체계를 사용해야만 한다. 하지만 유니코드 한자의 정보 처리시 호환용 한자나 이체자에 대해 각별한 주의가 필요하다.

유니코드에는 한중일 통합한자(CJK Unified Ideographs) 영역 이외에 별도로 한중일 호환용 한자(CJK Compatibility Ideographs) 301자에 대한 코드를 할당하고 있는데, 이는 유니코드 이전의 우리나라 표준 문자코드 체계인 KS X 1001에 포함되어 있던 한자와의 호환을 위한 것이다.[50] 한중일 호환용 한자는 '更(U+F901: 갱)', '車(U+F902: 거)'와 같은 이음이나, '禮(U+F9B6: 예)', '柳(U+F9C9: 유)'와 같은 두음법칙으로 표준음과 달리 발음되는 경우의 한자들로서, 하나의 문자코드 체계 내에 모양이 같은 한자가 여러 코드값을 가짐으로써 정보 처리의 문제를 발생하게 한다.[51] 이들 호환용 한자들은 사용빈도가 매우 높은 한자로서 정보 검색시

47) 『고려대 한국어대사전』의 원어 정보에 사용된 한자의 수는 6,815자이다.
48) 이전 KSC 5601-1987로서 1997년 정보기술 KS규격 번호체계 개정에 따라 바뀌었다.
49) 컴퓨터에서 세계 각국의 언어를 수용하여 하나의 체계로 표현할 수 있도록 제안된 국제적인 문자 코드로서, 현재 유니코드 버전 5.2까지 발표되었다.
50) 양경용. 2008. 유니코드 한자의 정렬 문제. 『문자코드연구센터』, 22: 6-8.
51) 이러한 코드값을 불일치 문제는 정보의 입력에서부터 발생되곤 하는데, 가장 대표적인 예가 '金(U+91D1: 쇠 금)'과 '金(U+F90A: 성 김)'이라고 할 수 있다. 국내에서 가장 대표적으로 사용되고 있는 워드 프로세서 가운데 하나인 마이크로소프트사의 'MS Word'에서 두 글자를 입력할 경우 서로 다른 코드 값으로 잘못 입력된다. 즉 한자의 자음 입력 방식으로 '금'으로 검색하면 '金(U+F90A: 성 김)'으로, '김'으로 검색하면 '金(U+91D1: 쇠 금)'자가 잘못 입력되는데, 이 한자는 우

재현율을 급격히 떨어뜨리는 요인으로 작용하고 있으며, 특히 표 33에서와 같이 '樂'자는 통합코드에 1자, 호환 영역에 3자가 포함되어 있어 이들 한자에 대해 적절한 처리가 필요하다.

표 33 한중일 호환용 한자

한중일 통합 한자	한중일 호환용 한자
更(U+66F4) 경	更(U+F901) 갱
車(U+8ECA) 차	車(U+F902) 거
金(U+91D1) 금	金(U+F90A) 김
禮(U+79AE) 례	禮(U+F9B6) 예
念(U+5FF5) 념	念(U+F9A3) 염
年(U+5E74) 년	年(U+F98E) 연
樂(U+6A02) 악	樂(U+F95C) 락 / 樂(U+F9BF) 요 / 樂(U+F914)낙

또한 유니코드에는 '峯(U+5CEF)/峰(U+5CF0)', '晉(U+6649)/晋(U+664B)'과 같은 이체자(異體字)나 속자(俗字)[52]가 많이 수록되어 있는데, 이들 한자들 또한 호환용 한자와 마찬가지로 재현율을 떨어뜨리는 요인으로 작용하며, 원어 정보를 동형이의어의 식별자로 사용시 오류를 유발할 가능성이 매우 높다. 실제 표 34와 같이 이들 한자가 국어사전에 사용된 경우를 살펴보면, 사전에 따라 정자와 이체자가 사전마다 다르게 사용되고 있음을 알 수 있다.

표 34 이체자와 속자의 사용 비교

표제어	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
천봉만학	千峰萬壑	千峰萬壑	千峯萬壑	千峯萬壑
진주 촉석루	晋州矗石樓	×	晉州矗石樓	晋州矗石樓

'천봉만학'의 원어 정보를 살펴보면 『금성판 국어대사전』과 『우리말큰사전』은 이체자인 '峰'를 쓰고 있고, 『표준국어대사전』과 『고려대 한국어대사전』은 정자인 '峯'를 사용하고 있다. '진주 촉석루' 또한 『금성판 국어대사전』에서는 속자인 '晋'를 쓰고 있는데 반해, 『표준국어대사전』과 『고려대 한국어대사전』에서는 정자인

리나라 인명의 성(姓)에 쓰이는 글자로 사용 빈도가 매우 높아 정보 검색시 잡음률을 높이는 요소로 작용하게 된다. '金'의 입력 오류는 'MS Office'뿐만아니라 마이크로소프트사 운영체제인 Windows에 내장된 한국어 입력기를 사용할 경우 공통적으로 발생되고 있다.

52) 이체자는 음과 뜻은 같으나 모양이 다른 한자를 로서, 흔히 정자(正字)에 상대되는 개념으로 사용되며, 속자는 한자에서 관습적으로 본래의 글자로부터 획을 달리하여 쓰는 한자를 이른다.

‘晉’를 쓰고 있다. 대부분 국어사전에서 원어 정보에서 한자 사용에 관한 원칙을 마련해 두고 있기는 하지만, 전체 표제어에 대해 일괄적으로 적용하기는 쉽지 않은 문제이다. 왜냐하면 ‘晝(U+756B)/昼(U+7575)’자와 같은 경우 정자인 ‘晝(U+756B)’ 보다 속자인 ‘昼(U+7575)’가 실생활에서 훨씬 많이 쓰이기 때문에 사전에서도 이를 따라 ‘昼(U+7575)’를 쓰는 경향이 높기 때문이다. 따라서 원어 정보를 동형이의어의 식별자로 사용하기 위해서는 원어 정보에 사용된 한자에 대한 정규화(regularization) 작업이 선행되어야 하는데, 이는 유니코드 한자에 대한 정보 검색용 이체자 시소러스와 같은 데이터베이스를 활용함으로써 해결될 수 있다.[53]

4.2 발음 정보

발음 정보는 사전에 수록된 표제어의 음운 변화나 소리의 길이를 밝히는 것으로, 한글 맞춤법에서 규정한 한글 자모만를 이용하여 음소 표기로 하되, ‘표준어 규정’(문교부 고시 제88-2호)의 <표준 발음법>에 따라 표준 발음을 제시하는 것이 일반적이다. 대부분의 국어사전에서는 표기 형태와 발음이 다른 경우 발음 정보를 대괄호 ‘[ ]’ 속에 표시하고 있는데, 다음의 예와 같이 표제어의 형태가 동일하지만 발음이 다른 경우가 있다.

<『표준국어대사전』>
밤[1] 명 해가 져서 어두워진 때부터 다음 날 해가 떠서 밝아지기 전까지의 동안.
밤[2] 명 [밤 :] 밤나무의 열매.
날개[1] 명 새나 곤충의 몸 양쪽에 붙어서 날아다니는 데 쓰는 기관
날개[2] 명 [-깨] 윷판의 끝에서 넷째 자리. 곧 짤밭에서 둘째 자리를 이른다.

‘밤’은 ‘해가 져서 어두운 동안’을 나타낼 때는 소리의 변화가 없지만, ‘밤나무의 열매’를 의미할 때는 길게 발음되고, ‘날개’는 ‘새나 곤충의 몸에 붙어 날아다니는 데 쓰는 기관’을 의미할 때는 표기와 발음의 차이를 보이지 않지만, ‘윷판의 끝에서 넷째 자리’를 의미할 때는 된소리로 발음됨으로써 동형이의어가 구별된다.

---

53) 김홍규, 김풍기(2002)는 유니코드(v.3.0)에 수록된 27,484자를 대상으로 이체자 관계를 조사하여 데이터베이스로 구축하였데, 국어사전에 사용된 대부분의 한자는 유니코드(v.3.0)에 포함되고 있어 원어 정보에 사용된 한자의 정규화 작업에 이와 같은 데이터베이스를 활용할 수 있다.

보통 국어사전에서 발음을 표기할 때는 표제어의 표기와 발음이 일치하지 않은 부분만 그 발음의 변화를 표시하되, 표제어의 표기와 일치하는 부분은 '-'로 표시한다. 다만 긴소리가 포함된 음절의 경우에는 음절을 생략하지 않고, 해당 음절의 뒤에 장음 기호(:)로 표시하고 있다.

표 35 발음 정보의 표기

『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
눈 : 보라	눈 : 보라	눈보라 [눈 : --]	눈보라 [눈 : --]
계 : 획서 [계획써/게훽써]	계 : 획서 :	계획서 [계 : 획써/게 : 훽써]	계획서 [계 : 획써/게 : 훽써]
둥글넓적하다 [—럽쩌카—]	둥글넓적하다 [—넙—]	둥글넓적하다 [--럽쩌카-]	둥글넓적하다 [--럽쩌카-]

발음 정보를 동형이의어의 식별자로 사용하기 위해서는 발음 정보의 표기 방법을 통일할 필요가 있다. 그런데 위의 표 35를 보면 소리의 길이를 표시하는 방법과 발음이 변하지 않는 음절의 처리에 있어 사전에 따라 차이를 보이고 있다.

첫째, 소리의 길이를 표시하는 방법을 살펴보면, 『표준국어대사전』과 『고려대 한국어대사전』은 발음란에 음운 변화와 함께 소리의 길이를 표시하고 있지만, 『금성판 국어대사전』과 『우리말큰사전』의 경우 발음란 대신 표제어에 직접 ':'를 넣어 장음 표시를 하고 있다. 『금성판 국어대사전』과 『우리말큰사전』과 같이 발음 정보를 이중으로 표기하는 방법은 비효율적일 뿐만 아니라, 가독성 측면에서도 음운의 변화와 소리의 길이를 각각 확인해야 발음해야 하는 불편함이 있다. 따라서 발음 정보를 표시하는 방법으로는 음운 변화와 소리의 길이를 발음란에 한꺼번에 표시하는 것이 여러모로 유리하다 할 수 있다.

둘째, 발음의 변화가 없는 음절의 처리 방법에서 『금성판 국어대사전』과 『우리말큰사전』은 음절의 개수와 상관없이 하나의 줄표를 사용하여 표시하고 있으며, 『표준국어대사전』과 『고려대 한국어대사전』은 붙임표를 음절의 개수만큼 사용하여 고유어를 표시하고 있다. 이는 앞의 원어 정보에서 고유어의 처리 방법과 동일한 것으로, 하나의 줄표만으로 여러 음절을 대체하게 되면 발음란의 음운 정보가 표제어의 어떤 음절과 대응되는지 쉽게 파악하기 어려우며, 붙임표를 사용하여 음절

수만큼 표시하는 방법이 가독성의 측면에서 장점이 있다.

기타 발음 정보를 동형이의어의 식별자로 사용하고자 할 때 발생될 수 있는 문제점으로는 두 개 이상의 표준 발음을 표시하는 방법에 관한 것을 들 수 있다. <표준 발음법>에서는 표준 발음이 둘 이상일 경우에 '/' 기호를 사용하여 병기하되, '/'의 왼쪽에는 원칙적인 발음을, 오른쪽에는 허용되는 발음을 제시하고 있다. 표 35의 '계획서'의 발음 정보를 살펴보면, 실제 대부분의 국어사전에서 <표준 발음법>에 따라 두 개의 발음을 표기하고 있음을 알 수 있다. 이와 같이 표준 발음이 두 개 이상인 경우에는 앞의 원어 정보의 XML 데이터 구조를 사용하거나, 아니면 원칙적인 발음 하나만을 식별 정보로 사용할 수 있다.

마지막으로 이상의 논의와 별도로 『우리말큰사전』의 표제어 '계획서'를 살펴보면, 장음과 경음 표기 모두 나머지 사전과 달리 나타나고 있음을 볼 수 있다. 즉 나머지 사전들은 모두 '계획서'의 첫음절에만 장음 표시를 한 것과 달리 3음절에도 장음 표시를 하고 있으며, 발음란에도 별도의 음운 변화를 표기하지 않아 표기와 발음이 일치하는 것으로 보고 있는데,[54] 이는 『우리말큰사전』이 '표준어 규정'(문교부 고시 제88-2호)을 따르지 않고, 한글 학회의 <한글 맞춤법>(1980)에 따라 편찬된 때문이라 할 수 있다.

4.3 조어 방식

단어는 조어 방식에 따라 단일어, 파생어, 합성어로 구분된다. 단일어는 '강아지, 가다, 행복(行福)'과 같이 하나의 자립적인 실질 형태소로 이루어진 단어이고, 합성어는 '집안, 병마개, 맛있다'와 같이 두 개 이상의 실질 형태소가 모여 새로운 뜻을 가진 단어를 말한다. 파생어는 실질 형태소에 파생 접사가 붙어서 생긴 단어로서, 명사 '신'에 접두사 '덧-'이 붙어 만들어진 '덧신'과 같은 접두 파생어와, 명사 '선생'에 접미사 '-님'이 붙은 '선생님'과 같은 접미 파생어가 있다. 따라서 국어사전에 수록된 단어를 조어 방법에 따라 '단일어, 합성어, 접두 파생어, 접미 파생어'

54) <표준 발음법> 제6항에서는 "모음의 장단을 구별하여 발음하되, 단어의 첫음절에서만 긴소리가 나타나는 것을 원칙으로 한다" 규정하고 있어 '계획서'의 3음절에 장음 표시를 한 것은 잘못된 것이라 할 수 있다. 또한 음운의 변화를 표시하는 데 있어서도 "받침 'ㄱ(ㄲ, ㅋ, ㄳ, ㄺ), ㄷ(ㅅ, ㅆ, ㅈ, ㅊ, ㅌ), ㅂ(ㅍ, ㄼ, ㄿ,ㅄ)' 뒤에 연결되는 'ㄱ, ㄷ, ㅂ, ㅅ, ㅈ'은 된소리로 발음한다"라는 제23항 경음화 규정을 지키지 않고 있다.

분류할 수 있으며, 이러한 단어의 결합 정보를 이용하여 동형이의어를 식별할 수도 있다. 실제 국어에서는 다음에 보이는 '모쟁이'나 '참새'와 같이 표기 형태가 같은 동형이의어 가운데 조어 방식이 다른 경우가 많이 나타나고 있다.

<『표준국어대사전』>
모-쟁이[1] 명 모를 낼 때에, 모춤을 별러 돌리는 사람.
모쟁이[2] 명 숭어의 새끼.
참-새[1] 명 참샛과의 새.
참새[2] 명 베틀에서, 두 개의 사침대 가운데 도투마리 쪽의 것.

다음의 표 36은 네 개의 국어사전에서 조어 방법을 표시한 것으로, 표제어에 붙임표(-)를 넣어 단어의 결합 형태를 밝히고 있다. 다만, 『고려대 한국어대사전』은 다른 사전과 달리 형태소의 결합 관계를 원어 정보에 통합하여 표시하고 있으며, 형태소의 종류에 따라 별도의 분석 표지를 붙임으로써 해당하는 단어가 합성어인지, 접두 파생어인지, 접미 파생어인지 밝히고 있다.[55]

표 36 조어 방법의 표기

『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
병-마개	병-마개	병-마개	병마개 [+甁+마개]
덧-신	덧-신	덧-신	덧신 [=덧+신]
선생-님	선생-님	선생-님	선생님 [+先生-님]

그런데 위의 예에서 보이는 것처럼 형태소의 결합 정보만으로는 그 단어가 합성어인지, 파생어인지, 또는 접두 파생어인지, 접미 파생어인지 구분하기 어렵다. 『고려대 한국어대사전』에서는 접두사나 접미사 등 형태소의 유형에 따라 별도의 부호로 표시하고 있어 어느 정도 구분이 가능하지만, '개나리꽃 [=개+나리+꽃]'과 같이 3개 이상의 형태소가 결합된 경우에는 '개나리+꽃'으로 결합된 합성어인지, '개+나리꽃'으로 결합된 접두 파생어인지 판단하기가 쉽지 않다. 따라서 현재 국어사전에 표기된 조어 정보를 통해서 단일어, 복합어, 구 표제어로 어휘를 분류하여 동형이의어의 식별에 이용할 수 있다.

55) 『고려대 한국어대사전』에서는 형태소의 유형에 따라 다섯 가지의 분석 표지를 붙이고 있는데, '+'는 자립어, '±'는 비자립어, '='는 접두사, '-'는 접미사, '_'는 어미를 의미한다.

4.4 품사 정보

품사는 공통된 성질을 가진 단어끼리 모아 놓은 갈래로서, 문법 기능과 의미 부류 등에 따라 나누는데, 국어의 경우 의미 부류에 따른 명사, 대명사, 수사, 동사, 형용사, 감탄사의 6품사와 문법 기능에 따른 관형사, 부사, 조사의 3품사를 통합하여 9품사로 나누는 것이 일반적이다. 국어사전에서는 수록된 모든 단어에 대해 품사 정보를 표기함으로써 단어의 의미 범주나 문법적 갈래를 밝히고 있어, 이를 이용하여 동형이의어를 식별할 수 있다.[56] 다음은 『표준국어대사전』에 수록된 동형이의어로서 의미나 문법적 기능의 차이로 품사가 달라지는 경우이다. '부르다'는 '사람을 부르다', '노래를 부르다'처럼 동작을 나타내는 것과, '배가 부르다'처럼 상태를 나타내는 것으로 그 의미가 확연히 다른 단어이고, '왈가닥' 또한 '성품이 덜렁대는 여자'를 나타내는 말과 '물건이 부딪치는 소리'를 나타내는 말로서 의미적으로 품사가 달라지는 경우이다. 반면에 '어쩌다'는 의미적으로는 유사하지만 문법적 기능의 차이로 품사가 달라지는 경우라 할 수 있다.

부르다[1] 동 ①말이나 행동 따위로 다른 사람의 주의를 끌거나 오라고 하다.
부르다[2] 형 ①먹은 것이 많아 속이 꽉 찬 느낌이 들다.
왈가닥[1] 명 남자처럼 덜렁거리며 수선스러운 여자.
왈가닥[2] 부 작고 단단한 물건들이 서로 거칠게 부딪치는 소리.
어쩌다[1] 동 '어찌하다'의 준말.
어쩌다[2] 부 '어쩌다가'의 준말.

대부분의 국어사전에는 학교 문법에서 정한 9품사와 접사, 어미를 표시하고 있으며, 여기에 사전의 편찬 목적에 따라 좀더 세분된 품사의 하위 범주를 덧붙이는 경우가 있다. 덧붙이는 품사의 하위 범주에는 '의존 명사, 보조 동사, 보조 형용사'가 대표적이고, '자동사'와 '타동사'를 구분해 넣기도 하며, 명사에도 '고유 명사, 자립 명사'와 같이 하위 영역을 더하는 경우가 있다.[57] 지금까지 국내에서 출판된 대사전 규모의 국어사전을 대상으로 품사의 표시 방법을 살펴보면 다음과 같다.

---

56) 강범모(2005)는 『표준국어대사전』에서 북한어, 옛말, 방언 등을 제외한 현대 한국어(표준어)의 동음이의어의 비율은 30%에 정도인데, 품사 구분을 할 경우에는 그 비율이 28.4%로 떨어지게 된다고 밝히고 있다.

57) 홍종선. 2007. 국어사전 편찬, 그 성과와 과제(5) : 풀이말 항목들의 설정, 『어문논집』, 56: 33-56.

『금성판 국어대사전』은 기본적으로 9품사에 더하여 여러 가지 하위 범주를 덧붙이고 있다. 명사에는 '자립 명사, 의존 명사'를 더하고 있고, 대명사에는 '인칭 대명사, 지시 대명사'를 추가적으로 표시하고 있다. 동사에 대해서는 '자동사, 불완전 자동사, 타동사, 불완전 타동사, 피동사, 사동사, 보조 동사'의 7가지 하위 범주를 두고 있으며, 형용사에 덧붙여서는 '보조 형용사'를 두고 있다. 접사는 '접두사, 접미사'로 구분하고 있으며, 어미의 경우 '선어말 어미'를 구분하고 있으며, 여기에 '관용구'를 추가하고 있는 것이 특징적이라 할 수 있다.

『우리말큰사전』은 한글 전용 원칙에 따라 학교 문법에서 정한 품사 명칭을 대신하여 우리말 이름을 정하여 쓰고 있다. 대체로 9품사 체제와 비슷하며 '지정사, 어간'을 추가한 것이 특징적이라 할 수 있다.『우리말큰사전』의 품사 체계를 학교 문법의 9품사를 기준으로 살펴보면, 이름씨(명사), 대이름씨(대명사), 셈씨(수사), 움직씨(동사), 그림씨(형용사), 매김씨(관형사), 어찌씨(부사), 느낌씨(감탄사), 토씨(조사)가 있으며, 여기에 추가하여 앞가지(접두사), 뒷가지(접미사), 씨끝(어미), 잡음씨(지정사), 줄기(어간)를 두고 있다. 품사의 하위 범주로는 메인이름씨(불완전명사), 제움직씨(자동사), 남움직씨(타동사), 도움움직씨(보조동사), 모자란제움직씨(불완전자동사), 모자란남움직씨(불완전타동사), 모자란그림씨(불완전형용사), 도움그림씨(보조형용사), 도움줄기(보조어간)를 두고 있다.

『표준국어대사전』은 학교 문법에 충실하여 품사의 하위 범주의 구분을 최소화하여 표시하고 있다. 9품사 이외에 추가된 품사 범주로는 '의존 명사, 보조 동사, 보조 형용사, 어미, 접사, 관용구'가 전부이다.

마지막으로『고려대 한국어대사전』을 살펴보면, 앞의 사전에서와 같이 9품사와 접사, 어미를 두고 있으며, 명사와 어미의 하위 범주를 세분화한 것이 특징이라 할 수 있다. 명사의 하위 범주로는 '일반 명사, 고유 명사, 자립 명사, 의존 명사'를 두고 있으며, 어미에는 '선어말 어미, 종결 어미, 연결 어미, 전성 어미'를 두고 있다. 이외에 '인칭 대명사, 지시 대명사, 자동사, 타동사, 보조 동사, 보조 형용사, 접두사, 접미사'의 하위 범주를 두고 있다.

지금까지 네 개의 국어사전을 대상으로 품사를 살펴 보았는데, 각 사전별 품사의 갈래를 정리하면 다음 표 37과 같다. 네 개의 사전 모두 학교문법의 9품사에 접사, 어미의 품사를 두고 있으나, 품사의 하위 범주에서는 사전마다 많은 차이를

보이고 있다. 특히 명사에서는『고려대 한국어대사전』이 4개의 하위 범주를, 동사에서는『금성판 국어대사전』이 7개의 하위 범주를, 어미에서는『고려대 한국어대사전』이 5개의 하위 범주를 두어 다른 사전들과 차이를 보이고 있다. 따라서 품사 정보를 이용하여 동형이의어를 식별하기 위해서는 현재의 사전들이 공통으로 표시하고 있는, 학교문법에서 제시한 9품사와 어미, 접사를 사용할 수 있다.

표 37 사전별 품사 분류

품사	하위 범주	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
명사	명사		○		
	일반명사				○
	고유명사				○
	자립명사	○		○	○
	의존명사	○	○	○	○
대명사	대명사		○	○	
	인칭대명사	○			○
	지시대명사	○			○
수사		○	○	○	○
동사	동사			○	○
	자동사	○	○		○
	불완전자동사	○	○		
	타동사	○	○		○
	불완전타동사	○	○		
	피동사	○			
	사동사	○			
	보조동사	○	○	○	○
	자/타동사		○		
형용사	형용사	○	○	○	○
	보조형용사	○	○	○	○
	불완전형용사		○		
지정사			○		
관형사		○	○	○	○
부사		○	○	○	○
감탄사		○	○	○	○
조사		○	○	○	○
접사	접사			○	
	접두사	○	○		○
	접미사	○	○		○
어간	어간		○		
	보조어간		○		
어미	어미		○	○	○
	어말어미	○			
	선어말어미	○			○
	종결어미				○
	연결어미				○
	전성어미				○

4.5 용언의 불규칙 활용 정보

국어에서 서술어의 역할을 하는 용언이 시제나 서법과 같은 문법적 기능을 나타내는 어미와 결합할 때 어간과 어미의 형태가 불규칙하게 활용하여 일반적인 음운 규칙으로 설명하기 어려운 경우가 있다. 보통 학교문법에서는 어간이 불규칙하게 변하는 것으로 'ㄷ불규칙, ㅂ불규칙, ㅅ불규칙, 르불규칙', 어미가 불규칙하게 바뀌는 '여불규칙, 러불규칙, 거라불규칙, 너라불규칙', 어간과 어미가 모두 바뀌는 'ㅎ불규칙'을 규정하고 있다.

용언의 활용은 동형이의어의 경우 대부분 활용 형태가 동일하지만, '누르다', '파묻다'와 같이 표제어의 표기는 같지만 활용 형태가 다른 단어도 있다. 용언의 경우 고유어의 비율이 높아 원어 정보를 통해서 동형어의 구별되는 비율이 비교적 낮게 나타나고 있는데, 이러한 동형어에 대해서는 용언의 활용 정보를 사용하여 동형이의어를 구별할 수 있다.

<『고려대 한국어대사전』>
누르다[1] [+누르_다] 르불규칙<눌러, 누르니> 동 (사람이 물체를) 표면 전체나 부분에 대하여 힘이나 무게를 가하다.
누르다[2] [+누르_다] 러불규칙<누르러, 누르니> 형 (사물이나 그 빛이) 황금이나 놋쇠와 같은 빛깔을 띤 상태에 있다.
파묻다[1] [+파+묻_다] ①땅을 파서 그 속에 묻다.
파묻다[2] [+파+묻_다] ㄷ불규칙<파물어, 파물으니> 사실을 따지면서 자세하게 묻다.

국어사전에서는 대부분 학교문법에 따라 용언의 불규칙 활용 정보를 표시하고 있지만, 일부 용언의 불규칙 현상에 대해서는 학교문법과 차이를 보이는 경우가 있다. 『금성판 국어대사전』에서는 학교문법에서 규정한 아홉 가지 불규칙 활용에다가 '우불규칙'을 더하여 표시하고 있느데, '우불규칙'은 용언의 어간이 어미와 결합하여 활용될 때 '우'가 떨어지는 것으로, 동사 '푸다'가 유일하게 이런 불규칙 활용을 보이고 있다. 『우리말큰사전』에서는 학교문법의 불규칙 유형 외에 '우벗어난끝바꿈(우불규칙)', 'ㄹ벗어난끝바꿈(ㄹ탈락)', '으벗어난끝바꿈(으탈락)'을 추가로 제시하고 있는데, 'ㄹ탈락'과 '으탈락'은 일정한 음운 규칙에 의해서 설명되기 때문에 학교문법에서는 규칙활용으로 규정하고 있다.[58] 『표준국어대사전』은 용언의 불규

칙 활용에 대해 약물로서 별도로 표시하지 않고, 어미 '-어, -의니' 등과의 결합을 통한 활용형을 제시하고 있으며, 『고려대 한국어대사전』에서는 학교문법에서 제시한 불규칙 활용 가운데 '거라불규칙'에 대해서는 밝히지 않고 있는데, 이들 네 사전에서 제시하는 용언의 불규칙 활용의 유형을 정리하면 다음의 표 38과 같다.

표 38 용언의 불규칙 활용

불규칙 유형	『금성판 국어대사전』	『우리말큰사전』	『표준국어대사전』	『고려대 한국어대사전』
ㄷ불규칙	○	○	×	○
ㅂ불규칙	○	○	×	○
ㅅ불규칙	○	○	×	○
ㅎ불규칙	○	○	×	○
러불규칙	○	○	×	○
르불규칙	○	○	×	○
여불규칙	○	○	×	○
우불규칙	○	○	×	○
거라불규칙	○	○	×	×
너라불규칙	○	○	×	○

4.6 방언, 비표준어

국어사전에 수록되는 표제어의 범위는 사전 편찬 목적에 따라 많은 편차가 있지만, 대사전 규모의 국어사전에서는 사용 빈도가 높은 방언이나 잘못된 어휘도 상당수 사전의 올림말로 수록되어 있다.[59] 일반적으로 국어사전에서는 이들 어휘에 대해 대응되는 표준어나 사용 지역을 밝히는 형태로 뜻풀이를 하고 있다. 다음은 『표준국어대사전』에 수록된 방언과 비표준어의 뜻풀이를 보인 것으로, 방언 항목의 '참봉2'를 보면 해당 방언에 대응되는 표준어를 '~의 방언'의 형태로 제시한 다음, 괄호 안에 방언의 사용 지역을 표시하고 있다. 그러나 '오지다2'와 같이 방언에 대응되는 표준어가 없는 경우에는 방언에 대해서도 뜻풀이를 하고, 괄호 안에 방언의 사용 지역을 표시한다. 비표준어의 경우 '깍정이3', '보름3'과 같이 대응되는 표준어를 '~의 잘못'의 형태로 제시하는데, 방언과 달리 모든 비표준어는 대응되

58) 『우리말큰사전』에서는 방언이나, 옛말에 대해서 'ㄹ벗어난끝바꿈', 'ㅸ벗어난끝바꿈', 'ㅿ벗어난끝바꿈', 'ㆁ벗어난끝바꿈'의 네 가지 불규칙 활용을 추가하고 있다.

59) 이운영(2002)은 『표준국어대사전』의 어휘 통계에서 방언이 20,503개, 비표준어는 9,464개가 수록되어 있다고 밝히고 있다.

는 표제어가 있어 비표준어에 대해 뜻풀이를 하지 않는다.

<『표준국어대사전』의 방언>

오지다[1] 형 【…이】=오달지다.
오지다[2] 형 『방』 마음이 흡족하다(전남).

참봉[1] 명 『방』 모든 도깨비를 이르는 말(제주).
참봉[2] 명 『방』 '장님'의 방언(평남).
참봉[3](參奉) 명 『역』 조선 시대에, 여러 관아에 둔 종구품 벼슬.

<『표준국어대사전』의 비표준어>

깍정이[1] 명 『식』 밤나무, 떡갈나무 따위의 열매를 싸고 있는 술잔 모양의 받침.
깍정이[2] 명 『역』 포도청에서, 심부름을 하며 도둑을 잡는 것을 거들던 어린아이.
깍정이[3] 명 '깍쟁이'의 잘못.

보름[1] 명 ①=보름날.
보름[2] 명 『방』 '바람'의 방언(함북).
보름[3] 명 『민』 '부럼'의 잘못.

방언과 비표준어의 경우 고유어의 비율이 높아 동형이의어가 많이 발생되고 있다. 그러나 앞에서 살펴본 다섯 가지 미시정보를 통해서는 방언이나 비표준어의 동형이의어가 구별되지 않는 경우가 많은데, 이들 어휘에 대해서 대응되는 표준어를 동형이의어의 식별자로 사용할 수 있다. 즉, '참봉2'는 '장님'을, '깍정이3'은 '깍쟁이'를 식별자로 사용하여 동형이의어를 구별하는 것으로, 표제어와 대응 표준어가 모두 동일한 동형이의어가 있을 확률이 매우 낮기 때문에 동형이의어의 식별률이 매우 높을 것으로 예상된다. 다만, 방언에서 '오지다2'와 같이 대응되는 표준어가 없는 경우가 문제가 될 수 있는데, 이러한 표제어에 대해서는 해당 방언이 사용되는 지역 정보를 대신 사용할 수 있다.

5. 미시정보의 기술 방안

지금까지 국어사전의 미시정보를 이용하여 동형이의어를 구별하고자 할 때 발생될 수 있는 문제점과 이에 대한 해결 방안을 살펴보았다. 이 절에서는 XML 기반으로 어휘 데이터를 교환할 수 있는 방안을 제시하고자 한다. 먼저, XML DTD 설계에서 여섯 가지 미시정보가 가지고 있는 데이터 구조를 분석하고, 이를 XML DTD 모델로 개발하여 제시하고자 한다. 다음으로 실제 국어사전에 나타나는 동형이의어를 앞에서 제시한 DTD 모델에 따라 XML 데이터를 작성하여 제시하고자 한다.

5.1 XML DTD 설계

국어사전의 미시정보를 동형이의어의 식별자로 사용하여 어휘정보를 교환하고자 할 때에는, 미시정보의 기술 원칙을 엄격하게 적용해야만 한다. 동일한 정보라도 기술 방법이 다를 경우 정보 처리를 위한 추가적인 비용이 소요되기도 하며, 경우에 따라 오류의 발생 가능성도 배제할 수 없기 때문이다. 따라서 XML DTD 설계 이전에 기술하고자 하는 여섯 가지 미시정보에 대한 데이터 타입과 구조에 대해 분석하고, 이를 표현하기 위한 적합한 XML 모델을 제시하고자 한다.

첫째, 표제어의 기술적 특성을 살펴보도록 하겠다. 국어사전에서 표제어는 기본적으로 한글로만 표기되지만, 접사나 어미와 같은 경우 붙임표(-)를 덧붙이거나, '타이 어(Thai語)'와 같은 명사 구에는 띄어쓰기를 표시하기 위한 빈칸이나 기호가 삽입되기도 한다. 따라서 컴퓨터가 표제어의 형태적인 표기의 동일 여부를 판단하기 위해서는 붙임표나 공백 문자가 없는 별도의 표제어 형태가 필요하며, 다음과 같이 표제어를 '표기용'과 '검색용' 두 가지로 기술할 수 있다.

둘째, 원어 정보의 경우에는 고유어와 같이 원어 정보가 없거나 '기념(記念/紀念)'이나 '삭막(索莫/索寞/索漠)'과 같이 두 개 이상의 원어가 사용되기도 한다. 따

라서 원어 정보를 기술할 때는 횟수에 대한 연산자의 사용이 필요하다. 국어사전에서는 두 개 이상의 한자 표기를 기술할 때 부수나 획수의 순서를 고려하여 배열하지만, XML 형식에서는 이러한 배열의 원칙을 고려할 필요가 없다. 그리고 원어에서 고유어 처리는 앞에서 살펴본 것처럼 붙임표(-)를 고유어의 음절수만큼 사용하여 표기하도록 한다. 또한 '시디(CD)'와 같은 외래어의 경우 두문자로 표기된 원어 정보만으로는 표제어를 식별할 수 없기 때문에 외래어의 본래 표기를 속성 정보로 줄 필요가 있다.

시디[1] [CD(<compact disk)] 광신호(光信號)로 기록된 정보를 재생하는 새로운 방식의 재생 기기.

시디[2] [CD(<cash dispenser)] 은행에 예금해 둔 돈을 은행 창구를 거치지 않고 꺼낼 수 있는 기계.

시디[3] [CD(<certificate of deposit)] 제삼자에게 양도가 가능한 무기명식(無記名式)의 정기예금.

셋째, 발음 정보의 경우에도 두 가지 이상으로 발음되는 단어가 있기 때문에, 원어 정보와 같이 횟수에 대한 제한이 필요하다. 그리고 발음이 변하지 않는 음절에 대해서는 원어 정보와 같이 붙임표(-)를 사용하여 표기하도록 한다.

넷째, 조어 방법은 '단일어, 복합어, 구' 세 가지 타입으로 유형화하여 기술하되, 리스트 연산자 '|'를 사용하여 선택을 제한할 필요가 있다.

다섯째, 품사 정보는 학교문법에서 제시한 9품사에 접사, 어미를 추가하여 기술한다. 다만, 두 개 이상의 품사 정보를 가지는 단어가 있기 때문에 '*' 부호를 사용하여 횟수를 기술하도록 한다.

여섯째, 용언의 불규칙 활용 정보는 10개의 불규칙 활용 유형을 리스트 연산자 '|'를 사용하여 선택을 제한할 필요가 있다.

마지막으로 방언과 비표준어의 경우 대응 표준어를 기술하되, 표제어가 방언인지, 비표준어인지를 속성 정보로 기술할 필요가 있다.

이상, 표제어와 미시정보에 대한 데이터 타입과 속성에 살펴보았는데, 이를 XML DTD로 표현하면 다음과 같다.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT 사전 (표제항+)>

<!ELEMENT 표제항 (표제어, 원어?, 발음?, 조어법, 품사?, 활용?, 표준어?)>

<!-- 표제항 하위 ELEMENT -->
<!ELEMENT 표제어 (표기용, 검색용)>
<!ELEMENT 표기용 (#PCDATA)>
<!ELEMENT 검색용 (#PCDATA)>

<!ELEMENT 원어 (원어값+)>
<!ELEMENT 원어값 (#PCDATA)>
<!ATTLIST 원어값 어원 CDATA #IMPLIED>

<!ELEMENT 발음 (발음값+)>
<!ELEMENT 발음값 (#PCDATA)>

<!ELEMENT 조어법 (단일어|복합어|구)>

<!ELEMENT 품사 (품사값+)>
<!ELEMENT 품사값 (명사|대명사|수사|동사|형용사|관형사|부사|감탄사|조사|접사|어미|
                  무품사)>

<!ELEMENT 활용 (ㄷ불규칙|ㅂ불규칙|ㅅ불규칙|ㅎ불규칙|러불규칙|르불규칙|여불규칙|우
                  불규칙|거라불규칙|너라불규칙)>

<!ELEMENT 표준어 (#PCDATA)>
<!ATTLIST 표준어 유형 (방언|비표준어) #REQUIRED>
```

5.2 XML 데이터 작성례

앞에서 살펴본 미시정보의 데이터 구조에 따라 XML 문서를 작성하였다. 다음의 XML 문서는 여섯 가지 미시정보가 나타날 수 있는 데이터 유형에 따라 작성한 것이다. 원어와 발음, 품사의 경우 2개 이상이 나타날 수 있는 특징을 보이고 있어 횟수 제한자를 사용하였으며, 조어법과 품사, 활용은 나타날 수 있는 유형이 고정적이기 때문에 '|' 부호를 사용하여 미시정보의 유형을 선택하도록 하였다.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE 사전 SYSTEM "사전.dtd">
<사전>
	<표제항>
		<표제어>
			<표기용>초-</표기용>
			<검색용>초</검색용>
		</표제어>
		<원어>
			<원어값>初</원어값>
		</원어>
		<조어법>단일어</조어법>
		<품사>
			<품사값>접두사</품사값>
		</품사>
	</표제항>
	<표제항>
		<표제어>
			<표기용>타이 어</표기용>
			<검색용>타이어</검색용>
		<표제어>
		<원어>
			<원어값>Thai語</원어값>
		</원어>
		<조어법>구</조어법>
	</표제항>
	<표제항>
		<표제어>
			<표기용>날개</표기용>
			<검색용>날개</검색용>
		<표제어>
		<발음>
			<발음값>-깨</발음값>
		</발음>
		<조어법>복합어</조어법>
		<품사>
			<품사값>명사</품사값>
		</품사>
	</표제항>
	<표제항>
		<표제어>
			<표기용>기념</표기용>
			<검색용>기념</검색용>
		<표제어>
		<원어>
			<원어값>記念</원어값>
```

```
                <원어값>紀念</원어값>
        </원어>
        <조어법>단일어</조어법>
        <품사>
                <품사값>명사</품사값>
        </품사>
</표제항>
<표제항>
        <표제어>
                <표기용>시디</표기용>
                <검색용>시디</검색용>
        <표제어>
        <원어>
                <원어값 어원="compact disk">CD</원어값>
        </원어>
        <조어법>단일어</조어법>
        <품사>
                <품사값>명사</품사값>
        </품사>
</표제항>
<표제항>
        <표제어>
                <표기용>계획서</표기용>
                <검색용>계획서</검색용>
        <표제어>
        <원어>
                <원어값>計劃書</원어값>
        </원어>
        <발음>
                <발음값>계:획써</발음값>
                <발음값>게:훽써</발음값>
        </발음>
        <조어법>복합어</조어법>
        <품사>
                <품사값>명사</품사값>
        </품사>
</표제항>
<표제항>
        <표제어>
                <표기용>대폭</표기용>
                <검색용>대폭</검색용>
        <표제어>
        <원어>
                <원어값>大幅</원어값>
        </원어>
        <발음>
```

```
                    <발음값>대:-</발음값>
                </발음>
                <조어법>단일어</조어법>
                <품사>
                    <품사값>명사</품사값>
                    <품사값>부사</품사값>
                </품사>
        </표제항>
        <표제항>
                <표제어>
                    <표기용>파묻다</표기용>
                    <검색용>파묻다</검색용>
                <표제어>
                <발음>
                    <발음값>--따</발음값>
                </발음>
                <조어법>복합어</조어법>
                <품사>
                    <품사값>동사</품사값>
                </품사>
                <활용>ㄷ불규칙</활용>
        </표제항>
        <표제항>
                <표제어>
                    <표기용>보름</표기용>
                    <검색용>보름</검색용>
                <표제어>
                <조어법>단일어</조어법>
                <품사>
                    <품사값>명사</품사값>
                </품사>
                <표준어  유형="방언">바람</표준어>
        </표제항>
</사전>
```

V. 결론

본 연구에서는 지금까지 한국어의 동형이의어 구별 방안들이 가지고 있던 상호 호환성 문제를 해결하기 위해, 국어사전의 미시정보를 이용하는 방법을 제안하였다. 어깨번호나 한정어를 사용하여 동형이의어를 구별하는 방법은 하나의 언어 사전 내에서 어휘를 식별하거나 배열 순서를 표시한다는 측면에서 효과적이었다. 그러나 언어 자원이 디지털화되고 이종의 데이터들이 네트워크를 통해 통합되고 있는 상황에서, 개별 언어 사전에서의 의미 구별을 위해 사용되던 어깨번호나 한정어를 이용하여 언어 정보를 효율적으로 처리하기는 매우 어려운 일이다. 따라서 본 연구에서는 언어의 중의성을 해결할 수 있는 범용적 식별 체계의 개발을 위해 사전의 미시정보를 이용할 수 있는 방법을 제시하고자 하였다.

먼저, 국어사전의 구조와 미시정보에 대한 비교 분석을 통해, 범용적인 동형이의어의 식별자로 사용되기 위한 미시정보의 선정 기준을 다음과 같이 세 가지로 설정하였다.

첫째, 사전 편찬 방향이나 편찬자에 따라 내용이 달라지지 않을 것.
둘째, 대부분의 표제어에 기술될 수 있는 요소일 것.
셋째, 정보의 집필이 쉽고 명료할 것.

다음으로 위의 기준에 따라 '원어 정보, 발음 정보, 조어 방법, 품사 정보, 용언의 불규칙 활용 정보, 방언/비표준어의 대응 표준어' 여섯 가지 미시정보를 동형이의어의 식별자로 선정하였으며, 선정된 미시정보의 실제적인 효율성은 『고려대 한국어대사전』을 대상으로 각 미시정보별 식별률을 분석함으로써 검증하였다. 식별률 분석 방법으로는, 첫째, 『고려대 한국어대사전』에 수록된 전체 표제어 가운데 여섯 가지 미시정보가 표기된 표제어를 각각 추출하고, 둘째, 미시정보별 표제어 그룹 내에서 동형이의어 관계가 발생되는 표제어를 확인한 다음, 셋째, 미시정보별 표제어 그룹의 동형이의어를 대상으로 해당 미시정보를 통해 구별되는 표제어의 비율을 통해 확인하였다. 여섯 가지 미시정보별 식별률은 방언/비표준어 97.64%, 원어 정보 95.00%, 품사 정보 12.90%, 발음 정보 6.50%, 조어 방법 5.24%,

활용 정보 1.90% 순으로 나타났다.

추가적으로 여섯 가지 미시정보의 종합적인 식별률을 확인하기 위해, 『고려대 한국어대사전』의 전체 표제어에서 나타나는 동형이의어를 대상으로 식별률 분석을 시도하였다. 『고려대 한국어대사전』은 386,889개의 표제어를 수록하고 있으며, 이 가운데 99,499개의 동형이의어가 나타나고 있다. 종합적인 식별률 분석은 여섯 가지 미시정보를 동형이의어의 식별자로 하나씩 추가함으로써 각 미시정보의 식별률 증가 추이도 함께 살펴 보았다. 처음 원어 정보의 식별률은 80.83%로 나타났으며, 이후 발음 정보 82.53%, 품사 정보 87.57%, 방언/비표준어 93.09%, 조어 방법 93.39%, 마지막 활용 정보까지 추가한 최종 식별률은 93.39%로 나타났다. 최종 식별률 93.39%는 전체 동형이의어 99,499개 가운데 6,574개를 제외한 모든 단어가 구별됨을 의미하며, 『고려대 한국어대사전』에 수록 된 전체 표제어 가운데 1.7%를 제외한 모든 모든 어휘의 형태적 중의성 문제가 해결 되고 있음을 의미한다.

다음으로 선정된 미시정보에 대해서 네 개의 국어 대사전을 대상으로 미시정보에 대한 기술 방법과 동형이의어의 식별자로 사용하기 위해서 고려해야 할 제한점에 대해 살펴보았다. 네 개의 사전들은 위의 여섯 가지 미시정보의 기술에 있어서 대부분 일치하고 있으며, 일부 표기 방법의 차이로 인해, 이들 미시정보를 식별자로 사용하기 위해서는 약간의 정규화 과정이 필요하였다. 특히 원어, 발음, 품사와 같은 미시정보의 경우 두 개 이상이 나타날 수 있었는데, 이에 대한 처리 방법으로 XML 기반의 데이터 모델을 제시하였다.

마지막으로 여섯가지 미시정보를 이용하여 어휘 목록을 교환하기 위해, 표제어와 미시정보들에 대한 데이터 구조를 분석하고, 이를 실제적으로 적용할 수 있도록 XML DTD를 개발하여 제시하였다.

결과적으로 국어사전의 미시정보를 이용한 동형이의어 식별 방법이 매우 유효하다고 볼 수 있으며, 앞으로 미시정보의 사용 원칙을 정밀하게 다듬고, 미식별어를 구별할 수 있는 식별 요소를 개발한다면, 국어사전뿐만 아니라 다양한 형태의 어휘 사전에도 적용될 수 있을 것으로 기대된다.

참고문헌

강범모. 2005. 동음이의어의 사용 양상.『語學硏究』, 41(1): 1-29.

강범모, 김흥규. 2009.『한국어 사용 빈도』. 서울: 한국문화사.

고려대 민족문화연구원 편. 2009.『고려대 한국어대사전』. 서울: 고려대 민족문화연구원.

국립국어원 편. 1999.『표준국어대사전』. 서울: 두산동아.

김경. 2002. 독-한 두말사전의 미시구조 조직.『독어학』, 5: 183-202.

김미령. 1994. 파생어와 합성어의 구분에 대하여 : '이름씨+풀이씨 줄기+뒷가지'의 짜임새를 대상으로.『國語國文學』, 31: 345-358.

김양진. 2008. 표제어 배열방식에 따른 국어사전의 거시구조 연구.『우리어문연구』, 30: 7-28.

김양진, 이현희. 2009. <고려대 한국어대사전>의 문법 및 뜻풀이 정보.『민족문화연구』, 51: 55-117.

김용권 역. 2000.『XML Bible』. 서울: 정보문화사.

김종덕. 2008. 국어사전에서의 발음 정보 처리에 대하여 :『새 연세사전』 발음 정보 제시 방안을 중심으로.『한국사전학』, 12: 61-96.

김준수, 옥철영. 2005. 정제된 의미정보와 시소러스를 이용한 동형이의어 분별 시스템.『정보처리학회논문지B』, 12(7): 829-840.

김태수, 최석두. 1997. 동형이의어의 구별을 위한 한글한정어 사용에 관한 연구.『情報管理學會誌』, 14(1): 107-124.

김흥규 외. 2007.『2007 국어 기초 자료 구축 결과 보고서』. 서울: 문화관광부.

김흥규, 김풍기. 2002.『다국어 정보 처리를 위한 유니코드(v3.0) 한자의 이체자 연구』. 서울: 정보통신부.

도원영, 김의수, 김숙정. 2007. '본말/준말'류에 대한 재고.『한국어학』, 37: 277-301.

도원영, 차준경. 2009. <고려대 한국어대사전>의 종합적 고찰.『민족문화연구』, 51: 1-54.

박미영, 정경재. 2009. <고려대 한국어대사전>의 관련어 정보.『민족문화연구』, 51: 167-213.

박수연. 2003. 외국인을 위한 '한국어 학습사전'에서 동음이의어의 구별에 관한 연구 : 길잡이말을 중심으로.『외국어로서의 한국어교육』, 28: 71-110.

서태길, 김혜령. 2009. <고려대 한국어대사전>의 문법 및 뜻풀이 정보.『민족문화

연구』, 51: 119-166.
양경용. 2008. 유니코드 한자의 정렬 문제.『문자코드연구센터』, 22: 6-8.
연세대학교 언어정보개발원 편. 2001.『연세 초등국어사전』. 서울: 두산동아.
운평어문연구소 편. 1991.『금성판 국어대사전』. 서울: 금성출판사.
유현경. 2000. 사전에서의 동형어 구별을 위한 새로운 제안 : 구분자(distinguisher)의 사용에 대하여.『사전편찬학연구』, 10: 133-157.
이승재. 2003.『한국어 정보 자료의 구축과 활용 방안 연구』. 박사학위논문, 서강대학교 대학원.
이운영. 2002.『표준국어대사전 연구 분석』. 서울: 국립국어원.
이유림 역. 2002. 주제기반 상호운용성 : High Level Thesaurus(HILT) Project에서의 현안들.『國會圖書館報』, 39(6): 93-100.
이희자. 2001. 개별어 사전 편찬: 동음이의어의 구별을 위한 길잡이말(Guide Words) 연구.『제2차 아시아 사전학회 국제 학술대회』, 167-201.
전상규. 1995. 동형이의어의 의미 식별에 관한 연구.『영남저널』, 6: 89-111.
최병진. 2002. 독한 전자사전을 위한 어휘항목의 미시구조에 대한 연구.『독어학』, 6: 1-20.
최석두, 조혜민. 2001. 다국어 시소러스의 설계.『한국정보관리학회 제8회 학술대회 논문집』, 5-10.
토박이 사전 편찬식 엮음. 2008.『보리 국어사전』. 서울: 보리
한글학회 편. 1991.『우리말큰사전』. 서울: 어문각.
홍종선. 2007. 국어사전 편찬, 그 성과와 과제(5) : 풀이말 항목들의 설정,『어문논집』, 56: 33-56.
홍종선, 최호철, 한정한, 최경봉, 김양진, 도원영, 이상혁. 2009.『국어사전학 개론』. 서울: 제이앤씨.
황경자. 1999. 사전의 미시구조 연구.『佛語佛文學硏究』, 38(2): 577-609.
Agirre, Eneko and Philip Edmonds. 2007. *Word Sense Disambiguation : Algorithms and Applications*. U.K.: Springer.
Hartmann, R. R. K. (ed.). 2001. *Teaching and Researching Lexicography*. Harlow: Pearson Education.
Stevenson, Mark and Yorick Wilks. 2001. “The Interaction of Knowledge Sources in Word Sense Disambiguation.” *Computational Linguistics*, 27(3): 321-349.

ABSTRACT

A Study on Distinguishing Homographs Using Micro-information of Korean Dictionary

Yang, Gyeong-yong
Major in Library & Information Science
Dept. of Library & Information Science
Graduate School, Hansung University

The ambiguity in language not only reduces a precision in a field of natural language processing and information retrieval but also decreases system efficiency. Especially, a method of distinguishing homographs has to be firstly developed in order to perform an efficient information processing as a rate of homographs is very high in the Korean language.

The aim of this study is to propose a method using micro-information of the Korean dictionary in order to solve interoperability matters which had the plans of distinguishing homographs in the Korean language till now. If we develop a method of distinguishing homographs using micro-information of the Korean dictionary, interoperability and sustainability of identifier will be ensured.

First of all, the criterias for selection of micro-information in order to be used as identifier of homonyms of interoperability through analysis of a micro-information and a structure of the Korean dictionary set up 3-items in the following:

1st, To be unchanged in content by direction of compilation or a compilation

committee for a dictionary.

2nd, To be a factor which enables it to be described in most main entry.

3rd, To be easy and clear to do writing and compilation of information.

In accordance with this criterias it has selected 6 kinds of micro-information as the identifiers of homographs, that is, (1)information of original language, (2)pronunciation, (3)word formation, (4)part of speech, (5)irregular conjugation of a predicate, and (6)standard language corresponded to dialect and non-standard language.

Secondly, this thesis has analyzed a identification rate by micro-information individually for the *Korea University Korean Dictionary* in order to identify whether 6 kinds of micro-information make the good use as the identifiers of homographs, or not.

A method of analyzing a identification rate in this thesis is as follows :

1st, it has individually extracted a main entry spelled in 6 kinds of micro-information among the whole entries which contained in the *Korea University Korean Dictionary*, And

2nd, After it has identified the entries that relation to homographs within an entry-group by micro-information individually is generated,

3rd, it was identified through the rate of entries, which had been distinguished through the appropriate micro-information for homographs of an entry-group by micro-information individually.

A identification rate by 6 kinds of micro-information has shown dialect and non-standard language of 97.64%, original language (95.00%), part of speech (12.90%), pronunciation (6.50%), word formation (5.24%), and conjugation (1.90%) in the order named. And, for the analysis of a comprehensive identi-

fication rate, this thesis has generally investigated a rate of increase to distinguishable rate of each micro-information by adding 6 kinds of micro-information as an identifier of homographs by ones. A identification rate of the original language information in the beginning showed 80.83%, thereafter the final identification rate which added to conjugation information showed 93.39%, as well as the pronunciation information (82.53%), part of speech (87.57%), dialect and non-standard language (93.09%) and word formation (93.39%).

Lastly, this thesis investigated what each micro-information was described in dictionary for the *Geumsung Korean Dictionary*, *Uri-Marl Dictionary*, *The Korean Standard Dictionary*, and the *Korea University Korean Dictionary* individually, and it has showed a descriptive method to make an use of these kinds of micro-information as an identifier with XML formalities. The description of the above 6 kinds of micro-information coincides mostly with 4 dictionaries but the orthography of micro-information is somewhat of difference in each dictionary. Accordingly, the micro-information which had been described in the dictionaries required a course of regularization in order to make an use of these kinds of micro-information as an identifier. Especially, the micro-information like the original language, the pronunciation, and part of speech could show two or more. this thesis has exemplified a model of XML DTD and data of XML for a processing of such a matter.

We expect that it will be a method applicable to all cases of wide-use for the information processing and distinction of homographs from a vocabulary dictionaries in diverse forms if they add a distinguishable factor capable of distinguishing un-discrimination words to arrange a principle of the micro-information use in detail, a method of distinction for homographs to use a micro-information in the Korean great dictionary in the future.