코퍼스를 이용한 정보검색용 전자사전구축에 관한 연구*

남영준
전주대학교 문헌정보학과

요 약

지능형 정보검색시스템이 효율적으로 운용되기 위해서는 여러개의 서브시스템이 필요하다. 특히, 시소러스와 색인 및 검색시스템용 전자사전은 중요한 지식베이스이다. 본 연구에서는 한글전자사전의 개발에 필요한 이론과 구축기술에 대해 조사하였다. 그 내용은 1)전자사전의 의미, 2)전자사전의 형태, 3) 전자사전개발을 위한 코퍼스 구축기술 및 방법이라는 이론적인 부분과 실제 과기원코퍼스2를 이용한 균형코퍼스를 설계하였다. 한편, 균형코퍼스를 이용한 기본적인 명사사전과 공기사전, 전문용어사전구축방법도 제시하였다.

1. 전자사전의 정의

지능형 시스템에는 인간전문가의 사고가 체계적으로 구축된 지식이 필요하다. 특히 지능형 정보검색시스템에는 전문가시스템과 하이퍼텍스트시스템, 지능형문헌정보시스템이 있다. 각 시스템마다 약간의 차이가 있지만 기본 구조속에 지식베이스는 반드시 구축되어야만 한다. 정보검색을 위해서는 반드시 색인절차가 필요하며, 색인시스템의 엔진이 용어사전이며, 이것이 지식베이스이다.

지식베이스의 구축은 전통적인 수작업 방식으로 색인전문가에 의해 주관적인 관점에서 이루어졌다. 새로운 개념의 파생과 낙후된 정보의 소멸이 과거 어느 때보다도 극심한 현대 사회에서는 보다 능동적이고 객관적인 형태의 지식베이스 구축방법이 필요하게 되었다. 한편, 전산기의 발달과 전산이론의 발달로 책자형의 사전에서 전산화된 사전의 개발이 가능하게 되었으며, 정보검색시스템에서는 이를 지식베이스로 활용하게 되었다.

일반적으로 사전은 크게 다음과 같이 구별할 수 있다.

① 사전(辭典) : 언어를 모아서 일정한 순서로 벌여 싣고 낱낱이 그 발음, 의의, 용법, 어원 등에 관하여 해설한 책.

② 사전(事典) : 여러 가지 사항을 모아 그 하나 하나에 해설을 붙인 사전

전자사전은 첫번째의 의미와 거의 유사한 구조와 쓰임새를 갖고 있다. 일반적으로 전자사전은 다음과 같은 용도로 활용되고 있다.

첫째, 지능형 워드프로세서에서 활용될 수 있다.

자판의 영한·한영자동변환과 오타자동수정, 띄어쓰기 등을 컴퓨터가 자동으로 처리한다.

둘째, 차세대 기계번역시스템에서 활용될 수 있다.

개념수준의 의미처리가 가능하다.

셋째, 지능형 정보검색분야에서 활용될 수 있다.

질문응답시스템의 질의어 처리와 키워드 자동추출, 연관어 자동탐색 등을 가능하게 한다.

넷째, 어학교육 CAI에서 활용될 수 있다.

전치사나 격조사의 용법을 교정하면, 문법상의 오류를 검출하는 지식을 제공한다.'

---

* 본 연구(과제명:국어 텍스트베이스 구축과 가공)는 1996년도 문화체육부의 지원을 받아 연구되었음.

2. 전자사전의 형태

전자사전은 책자형태의 일반 단어사전이 컴퓨터가독형으로 구성된 것과 특수한 목적을 갖고서 언어 및 용어의 여러 정보가 구축된 컴퓨터 가독형으로 크게 구분할 수 있다. 국내에서 개발된 컴퓨터가독형 일반사전은 책자형 사전에서 제공하는 품사정보와 활용정보, 간단한 용례정보가 포함된다. 그러나 특수한 목적을 갖고 개발된 전자사전은 그 목적에 따라 다양한 형태의 언어정보 사전이 된다. 그 목적에 따라 다음과 같이 구분할 수 있다.

① 대역사전
② 공기사전(共起辭典)
③ 전문용어사전

2.1 전자사전의 개발방법

2.1.1 일반단어사전을 이용한 방법

전자사전의 기본형태는 일반사전에서 갖고 있는 언어정보를 반드시 포함하고 있어야 한다. 한편, 책자형 사전의 정보수준으로 전자사전을 작성할 경우에는 상당한 연구와 시간, 경비가 소요된다. 그러므로 구미에서는 일반 책자형 단어사전을 기계가독형으로 변환하는 방법을 사용하고 있다. 이때의 장점은 다음과 같다.

① 사전개발기간이 단축된다.
② 여러 유형의 MTD를 동시에 개발할 수 있다.
③ 언어학적 연구결과를 최대한 수용할 수 있다.
④ 반복적 upgrade가 용이하다.

이때 발생할 수 있는 단점은 다음과 같다.

일반사전들은 기본적으로 사람이 읽어서 사용하려는 목적으로 제작된 것이기 때문에 어휘의 의미가 역시 그 언어 혹은 타 언어라는 자연 언어의 형태로 기술되어 있다. 자연언어로 기술된 의미를 이해하려면 매우 높은 수준의 기술이 필요하기 때문에, 이러한 표현 양식은 전자 사전으로서는 적합하지 못하다.[1] 한편, 용례로 설명된 부분에 있어 이태릭체와 같은 특수한 부호나 기호들이 어떤 의미로 사용되었는지를 구분할 수 없는 단점이 있다. 이를 위해서는 별도의 TEI(Text Encoding Initiative)와 같은 문서표준원칙이 事前에 제정되어야 하는 문제점도 있다.

2.1.2 코퍼스를 이용한 방법

기존의 책자형 단어사전은 통계적인 데이터보다는 오랜기간동안의 연구를 거쳐 국어학자들에 의해서 표제어가 선정되었다. 이러한 특성에 따라 사전의 갱신에는 많은 인력과 경비가 소요되었다. 일반 문장에 출현한 낱말의 빈도정보와 공기정보(co-occurrence information)가 객관적으로 입수될 수 있다면 이를 근거로 후보표제어를 선정할 수 있다. 빈도는 어휘의 발생빈도를 말뭉치 전체에서 조사하는 것으로서, 빈도의 많고 적음에 따라 어휘사용의 관련성을 어느 정도 알아볼 수 있다. 공기정보는 코퍼스내에 출현한 단어와 단어간의 상대적인 관련성을 파악할 수 있도록 한다. 즉, 코퍼스를 이용하여 전자사전을 구축하는 이유로는 전문적인 관점은 배제되나 객관성과 現時性을 모두 만족시킬 수 있기 때문이다. 단, 통계적인 수치가 안정적이고 신뢰성을 확보하기 위해서는 분석대상이 되는 코퍼스가 일정 규모이상이 되어야만 한다. 코퍼스를 이용하여 전자사전을 구축할 경우의 장점은 다음과 같다. [2]

---

1) 김영택 외. 자연언어처리. 교학사. p.207.
2) 남영준. 정보검색을 위한 전자사전 구축 기술. 한국어정보검색기술세미나. 한국어정보처리연구회. 1996. p.58.

① 단어의 의미분류가 용이하다
② 단어의미에 따른 용례선정이 용이하다.
③ 단어의 의미에 대한 통계치 정보를 입수할 수 있다.
④ 체계적인 품사분류가 가능하다.

한편, 코퍼스 완성후 결과의 활용방안은 다음과 같다. [3]

① 어휘의 파악 : 말뭉치로부터 직접 어휘들을 파악하여 품사에 따라서 분류함으로서, 사전의 어휘항목을 설정하고 사전정보를 구축할 수 있다.

② 구절(phase)의 파악 : 코퍼스의 문장들을 조사하여 구절들을 파악하고 그 패턴을 찾아서 일반화시킬 수 있다.

③ 관용어의 파악 ; 관용어란 그 구절을 구성하고 있는 형태소 개개의 의미의 총화와 전혀 다른 의미를 갖고 있는 구절을 의미하며, 정상적인 분석으로 얻어지는 의미와는 전혀 다른 의미를 갖게 되기 때문에 특수하게 취급하여야 할 필요가 있는 자연언어표현이다. 코퍼스를 분석할 경우 이에 대한 정보를 입수할 수 있다.

④ 연어의 파악 : 연어(collocation)는 간단한 형태의 관용어이며, 연관된 어휘성분이 자주 어울려 특정한 의미를 갖게 된다.

⑤ 용례 : 용례는 문법상 오류가 있더라도 자주 사용되는 구나 어절을 파악할 수 있다.

⑥ 문법구성 : 어휘, 구절, 관용어, 연어, 용례들의 파악이 이루어지고 나면 이를 토대로 코퍼스에 기반한 전산문법[4]을 구성할 수 있다.

3. 코퍼스의 구축

코퍼스란 언어학에서 사용되는 단어로서 국내에서는 이에 대한 해석으로 말뭉치, 말모둠등이 사용되고 있으나, 말의 의미는 '사람의 생각이나 느낌을 입으로 나타내는 소리 또는 그 행위나 내용'으로 사전적 정의를 내리고 있다. 한편, 글은 '말을 글자로서 나타낸 적발'이라고 정의하고 있다. 국내외에서 채록하고 있는 모든 코퍼스는 전산코드화한 글이기 때문에 말뭉치 혹은 말모듬이란 해석은 코퍼스의 개념을 모두 소화하지 못하고 있다. 그러므로 코퍼스는 커다란 글뭉치 혹은 글모듬, 글덩어리로 표현되어야 하고, 엄밀하게는 '말과 글덩어리'로 표현되어야 한다.

코퍼스는 자료 수집방법 혹은 사용목적에 그 유형을 다음과 같이 분류할 수 있다.

3.1 코퍼스의 유형[5]

코퍼스는 용례에 따라 부속정보나 종류나 용례의 정확성, 장르별 분포, 언어의 시대성에 따라 다음과 같이 구분할 수 있다.

① 부가정보
- 원문코퍼스
- 형태·통사 태그주석코퍼스
- 구문구조 트리구조 코퍼스

② 텍스트구성
- 균형코퍼스

---

3) 김영택 외. 자연언어처리. 교학사. p.211-212.
4) 남영준. "국어형태·통사 태그 규격. 제1회 우리말 정보처리 규격 심포지움. 한국과학기술원 인공지능연구센터. 1996. pp.37-46.
5)김덕봉. "국어 코퍼스 구축방안," 제1회 우리말 정보처리 규격 심포지움. 한국과학기술원 인공지능연구센터. 1996. p. 10.

- 피라밋코퍼스
- 과도기적 코퍼스

③ 언어의 시기
- 공시코퍼스
- 통시코퍼스

④ 언어의 수
- 단일어 코퍼스
- 병렬코퍼스

⑤ 용례의 종류에 따른 구분
- 문서코퍼스
- 음성코퍼스

3.2 코퍼스의 역사

코퍼스는 컴퓨터가 읽을 수 있는 형태로 지정된 자연어 용례들과 이들 용례에 대한 부속정보들의 묶음을 말한다.[6] 코퍼스의 개발이유는 언어학분야의 규칙기반접근방법의 한계를 극복하기 위해 자연어 처리에 도입된 통계적인 접근 방법을 이용하여 실세계의 다양하고 불규칙적인 언어현상에 관한 데이터를 입수할 수 있기 때문이다.

3.2.1 Brown 및 LOB 코퍼스

전산화된 형태의 최초 코퍼스는 미국의 브라운대학에 있는 Nelson Francis와 Henry Kucera에 의해 1961년부터 1964까지에 걸쳐 제작된 영문으로된 Brown 코퍼스이다. 국내에서는 연세대학교에서 1980년대 말부터 한국어 코퍼스를 부분적으로 구축하기 시작하였으며, 1994년부터 한국과학기술원에서 과기원코퍼스를, 1994년부터 고려대학교도 대규모 코퍼스를 구축하였으며, Brown 코퍼스가 그 기준이 되었다.

현재까지 국내외에서 개발된 대부분의 코퍼스는 균형코퍼스의 형태를 취하고 있으나 균형의 기준이 되는 것은 구축기관에 따라 조금씩의 차이를 보이고 있다. 초창기에 개발된 Brown 과 LOB[7] 균형코퍼스의 기준과 구성비율은 아래 표1과 같다.

---

6) 임해창, 이상주, 이호. "전산언어학에서의 언어데이터베이스 활용", 한국어데이터베이스의 설계 및 응용을 위한 기초 연구. 민음사. 1995. p.47.

7) LOB(Lancaster-Oslo/Bergen)코퍼스는 영국에서 개발된 초창기 코퍼스이다.

구 분		Brown	LOB
정보전달적 산문	신문 : 보도	44	44
	신문 : 사설	27	27
	신문 : 서평	17	17
	종교	17	17
	기술 및 취미	36	38
	상식자료	48	44
	美文, 전기, 수필	75	77
	정부문서 및 기타	30	30
	학술과학서적	80	80
문학적 산문	일반소설	29	29
	추리소설	24	24
	과학소설	6	6
	모험 및 서부소설	29	29
	연애소설	29	29
	유머집	9	9
계		500	500

<표 1> Brown, LOB 코퍼스의 구성비율

위 두 개의 코퍼스는 미국과 영국에서 개발된 코퍼스라는 것과 점을 제외하고는 구성적인 면에서는 거의 대동소이하다. 다만 균형코퍼스라는 관점에서 볼 때 텍스트위주의 수집방법을 사용하여 말데이터(口語資料)가 거의 수집되지 못하고, 글데이터(文語資料)위주의 코퍼스라 할 수 있다. 또한 이 코퍼스는 1960년대에 완성된 데이터로서 현대사전편찬과 언어패턴분석을 위한 코퍼스로는 부적절한 데이터라 할 수 있다. 이 코퍼스들은 균형의 관점을 정보전달적 산문과 문학적 산문으로 구분하고 있으며 주 수집대상을 전자에 중점을 두고 있다.

3.2.2 낱말빈도조사데이터

국내에서는 코퍼스에 균형의 개념을 처음 도입한 것은 1956년대 문교부에서 낱말 빈도를 위해 실시한 조사이다. 대상이 된 데이터는 초중등교과서일부와 일반정기간행물이었다. <표2>

분 류	세 부 항 목	비 율
초중등교과서	국어, 가사, 사회생활	30%
	과학, 실업류	20%
일반간행물	문학예술류	30%
	신문, 잡지, 방송원고 등	20%

<표 2> 문교부 낱말빈도조사데이터

이 조사에서의 균형의 관점은 학술적 자료와 비학술적 자료로 구분한 것이다.

3.2.3. 연세말뭉치 1

국내에서 코퍼스의 중요성에 대해 인식하고 본격적으로 데이터를 수집한 것이 연세대학교에서 최초로 구축한 연세말뭉치1이다. 이에 대한 구성비율은 <표 3>과 같다.

분 류	비 율	세 부 항 목	비 율
신 문	33%	정치	7%
		경제	6%
		사회	8%
		문화	6%
		오락/스포츠	6%
		여성/ 가정	5%
		정치/ 사회	5%
잡 지	20%	문학	5%
		취미/오락	5%
		수필	6%
소설 및 수필	18%	수필	6%
		일반소설	7%
		역사소설	5%
취미 및 교양	10%	기술/수공예/취미	4%
		생활정보	6%
수기, 전기 및 실화집	9%	수기/전기	4%
		실화집	5%
교과서	5%	초등학교 국어교과서	1%
		중등학교 국어교과서	2%
		고등학교 국어교과서	1%
		대학교 국어교과서	1%
방송스크립트	5%		5%
계	100%		100%

<표 3> 연세말뭉치 1의 구성(안)

연세말뭉치의 특징은 다른 코퍼스에 비해 학술적인 데이터가 상대적으로 작은 것이다. 또한, 방송스크립트를 일부 포함함으로서 구어데이터를 수집하였다. 이러한 구성형태는 1980년대에 영국에서 개발한 Birmingham코퍼스에서 약 75%의 글데이터와 25%의 말데이터로 구성하여 현대 영국 표준영어에 대한 데이터를 입수하려는 효과를 우리글 데이터에서도 얻으려는 방법의 일환으로 판단한다.

3.2.4 고려대학교 한국어 말모듬1

연세말뭉치와는 별도로 1995년에 1000만어절 규모의 '고려대학교 한국어 말모듬1'을 구축하였다. 이 작업의 목적은 국어사전을 위한 용례가 이 말모듬을 기반으로 구성되고 있다.[8] 이 코퍼스

8) 김홍규, 강범모. "고려대학교 한국어 말모듬 1 : 설계 및 구성", 한국어학, 제3호, 한국어학회. 1996. pp.233-258.

의 균형의 관점은 1)문/구어구분, 2)미디어별구분, 3)텍스트주제에 의한 구분 및 4)문/구어의 세분 구분등 4가지로 기준을 정했다. <표 4>

구 분	계 획	실제구현
구어/준구어	약12%(약120만)	11.7%
신문	약20%(약200만)	20.7%
잡지	약10%(약100만)	9.8%
책-정보	약35%(약350만)	33.5%
책-상상	약21%(약210만)	21.0%
기타	약 2%(약20만)	3.3%

<표 4> 고려대학교 한국어 말모둠1

한편, 고려대코퍼스는 향후 효율적인 데이터관리를 위해 데이터(문서)를 전자 문서의 표준규칙(TEI)을 일정부분 수용하고 있다. 이러한 방법은 저작권등과 같은 외부적인 상황변화에 따라 일괄적인 데이터처리에 매우 유용할 것으로 생각된다.

3.2.5 과기원코퍼스1

과기원코퍼스1의 특징은 두 개의 균형적 관점으로 코퍼스를 구축한 것이다. 하나의 관점은 예술과 비예술이라는 것이며, 또하나는 문어와 구어란 관점으로 설정하였다. 이러한 복수의 균형기준에도 불구하고 예술이라는 기준을 선정하였기 때문에 상대적으로 학술적인 자료의 구성비율이 낮아지게 되었다. <표 5>

	문 어	문어화된 구어	구 어
예 술	소설 수필 전기 시	희곡 시나리오	연극 영화
비 예 술	논설문 설명문 보고문 신문보도 일기 자서전 안내문 광고문 벽보 공고 공문 편지 광고문 표어	연설문 강의록 대담, 좌담 기록 구술기록	연설 강의 대담 좌담 일상 대화 이야기 (구술) 독백

<표 5> 과기원 코퍼스 1의 구성비율

4. 균형코퍼스의 설계

균형코퍼스는 균형의 기준을 어디에 설정하느냐에 따라 코퍼스의 성질이 크게 좌우된다. 이에 따라 Sinclair는 균형코퍼스를 구축하기 위한 코퍼스구축에 필요한 일반원칙을 다음과 같이 제안하고 있다. 1)기술이 지원하는 한 코퍼스의 규모를 최대한으로 설정할 것, 2) 코퍼스의 대표성을 획득하기 위해 여러 범주의 샘플을 선정할 것, 3)데이터 전체가 분명한 출처가 분명할 것.

한편, 균형코퍼스를 구축하기 이전에 다음과 같은 점들이 반드시 고려되어야 한다.

첫째, 코퍼스를 구축하여 입수된 통계적 정보가 신뢰도를 보일 수 있도록 대상 데이터의 양은 적정한가.

초창기 Brown과 LOB코퍼스는 100만어절 수준으로 약 500권의 자료로 코퍼스를 구축하였다. 일반적으로 언어학적으로 사용되는 사전은 10만 항목이상이 되어야 실질적인 자연어처리시스템에 유용하다고 판단되므로[9] 통계적 정보로 10만항목을 확보하기 위해서는 최소한 그 100배이상이 되어야 한다고 판단된다. 그러므로 코퍼스가 최소한의 신뢰도를 확보하기 위해서는 1000만어절이상이 되어야 할 것이다.

둘째, 코퍼스가 대표성을 띄고 있는가.

예를 들면, 용례사전과 전문용어사전을 구축할 경우 너무 일반적인 분야와 말위주의 데이터를 구축할 경우 해당 코퍼스는 대표성이 없어지게 된다. 또한 수준 이하의 잡문을 분석데이터로 설정하였을 경우 분석된 결과는 해당분야의 대표성을 갖지 못하게 된다.

셋째, 표본추출과정의 신뢰도는 있는가.

표본 추출을 데이터수집의 편의를 위해 주변에 얻기 쉬운 자료로 선정한다면 균형적인 데이터 수집이 이루어지지 않을 수 있다. 이러한 코퍼스를 이용한 분석결과는 편향적인 수치를 나타낼 수 있다.

넷째, 자료입력의 신뢰도는.

대부분의 코퍼스는 키인(key-in)방식과 OCR방식이 사용된다. 키인방식은 입력자의 성실도와 학력에 크게 좌우되므로 코퍼스의 신뢰도가 입력자에 의해 크게 좌우될 수 있다. 한편, OCR에 의한 방식은 스캐너의 성능과 OCR 프로그램의 성능에 따라 코퍼스의 입력수준이 좌우될 수 있다.

다섯째, 저작권의 저촉여부는.

현재 국내의 여러기관에서 구축한 코퍼스 데이터의 공유가 어려운 이유는 저작권에 대한 명확한 규정이 없기 때문이다. 따라서 균형코퍼스를 폭넓게 하기보다는 저작권에 크게 저촉되지 않는 데이터만을 수집하는 경향이 있다.

위에서 분석한 바와 같이 코퍼스를 구축하기 위해서는 코퍼스구축의 모든 목적을 달성하기 위한 초용량 규모(10억어절이상)의 코퍼스를 구축하는 것이외에는 코퍼스구축에 분명한 목적을 설정하고 코퍼스를 구축하여야 한다. 왜냐하면 규모면에서 제한된 규모의 코퍼스를 구축하면서 상대적으로 너무 균형적인 코퍼스를 염두에 둘 경우는 분석된 결과가 통계적 신뢰도를 상실할 우려가 있기 때문이다.

4.1 과기원코퍼스 2 (안)의 설계

본 절에서는 과기원코퍼스2를 구축하기 위해 활용한 원칙과 방법을 제시하고, 그 방법을 분석하고자 한다.

4.1.1 설계원칙

본 코퍼스는 용례사전과 전문용어사전을 구축하는 것을 목표로 한다. 이를 위해서는 수집대상으로 첫째, 표준어를 대상으로 한다. 둘째, 가능한 한 1990년대 이후 출간된 학술자료를 구축데이터로 설정한다. 코퍼스의 규모는 1500만어절을 최소선으로 한다.

---

9) 김영택 외. 자연언어처리. 교학사. p.206.

4.1.2 균형기준

본 코퍼스는 기존의 코퍼스와는 달리 글위주의 데이터를 분석자료로 한다. 또한, 말위주의 자료도 가능한 한 학술적인 데이터를 수집한다.

균형기준은 주제분야와 데이터내용수준으로 구분한다.

① 주제분야로 구분

주제분야는 한국십진분류법에서 제안한 다음 10개의 類분야로 한다.

- 총류, 철학, 종교, 사회과학, 순수과학, 기술과학, 예술, 어학, 문학, 역사

② 데이터수준

수준은 학술적 자료와 비학술적 자료를 대상으로 한다. 대표적인 학술자료는 학위논문, 교과서 등으로 설정하였으며, 비학술적 자료로는 신문, 시사정기간행물을 설정하였다. 향후 초용량코퍼스 구축을 위해 최소한의 말위주의 데이터를 수집한다. 그 종류는 방송대본과 뉴스대본으로 제한하며, 내용은 가능한 한 학술위주의 방송원고를 우선한다.

5. 전자사전의 구축

전자사전은 크게 2가지 유형의 정보를 갖고 있다. 하나는 일반 단어 및 용어사전에 수록된 내용적 정보와 언어학적인 측면에서 글과 말의 형태적 정보이다. 일본 EDR은 전자사전내에 일본어 단어사전과 영어단어사전, 개념사전, 일영대역사전, 영일대역사전, 일본어 공기사전, 일본어코퍼스, 영어공기사전, 영어코퍼스, 전문용어사전으로 구성되어 있다.[10] 어떤 목적으로 코퍼스가 구축되어도 그 부산물로서 대략 위와 같은 정보가 수록된 전자사전이 개발될 수 있다.

시스템 구축된 코퍼스에서 어떠한 정보를 추출하고, 그 정보를 어떻게 군집화(clustering)하여 어떠한 종류의 사전을 구축할 것이냐의 결정은 어떤 정보가 지능형 정보검색 지식베이스로 활용될 수 있는 가로 결정된다. 이를 위해서는 1차적으로 색인시스템용사전인 명사사전과 동사사전, 기능어 사전이 구축되어야 한다.[11] 명사사전과 동사사전은 어휘형태소사전이라 할 수 있으며 기능어사전은 문법형태소사전으로 구분할 수 있다.

5.1 색인시스템용 사전

① 명사사전 : 명사사전은 일반명사사전과 고유명사사전, 전문용어사전으로 성분에 따라 별도로 구축된다. 특히 전문용어사전은 먼저 분석대상이 되는 코퍼스가 반드시 주제별로 구분되어 있어야 한다.

② 동사사전 : 명사나 부사에 '-하다', '-되다','-스럽다'가 결합되어 용언이 되는 경우에는 일반적인 결합 규칙을 발견하기 어렵기 때문에 이는 모두 명사사전에 수록된다. 이를 제외한 동사에 대해 통사 및 의미기능에 따라 구분한 정보를 수록한다.

③ 기능어사전 : 어절에서 부수적 역할을 하는 사전으로 주로 조사정보와 어미정보로 사전이 구성된다. 명사사전에 수록된 불완전명사와 너무 일반적인 의미를 지닌 일반명사를 지칭한 불용어사전과 기능어사전은 반드시 구분되야한다.

---

10) (株)日本電子化辭書研究會. EDR電子化辭書利用マニュアル 第2.1版. 同研究會. 1994.

11) 한영균, 정보검색용전자사전. 한국통신 장기기초연구과제 1995년도 2회기술회의록. 1995.

5.2 공기사전

색인시스템용 사전은 형태소사전이라면, 공기사전은 의미분석용 사전이다. 공기정보는 단일문장내에서 한 단어와 다른 단어가 동시에 어울려 쓰이는 용례정보이다. 공기사전은 자연언어문장에서 허락된 단어의 조합형태와 개념간의 조합이 가능한 기능적인 2항의존관계에 대한 정보도 제공하고 있다. 즉, 공기사전은 표층적인 공기관계를 분석하는 것이 1차목표이지만, 의미적 정보와 문법적 정보도 입수할 수 있다. 특히, 명사와 명사간의 공기패턴은 관련어간의 관계 파악에도 중요한 정보가 될 수 있다. 공기사전의 구성은 다음과 같다.

① 단어정보 : 품사 및 관용구
② 격정보 : 기능어정보포함
③ 빈도정보 : 공기항목빈도, 표층공기정보, 共出情報
④ 용례정보

5.3 전문용어사전

전문용어는 문자인식프로그램과 자동번역시스템에서 인식 및 번역의 효율을 높여주는 중요한 지식베이스로 활용될 수 있다. 또한, 정보검색분야에서는 중요한 검색도구로 활용되고 있는 시소러스의 개발에도 활용될 수 있다. 이러한 전문용어의 중요성에도 불구하고 지금까지의 구축 및 개발된 코퍼스와 전자사전내에는 전문용어에 대한 구체적인 사용례와 구축방법이 기술된 바가 없다. 일본 EDR의 경우도 '정보처리분야'의 전문용어사전만이 개발되어 있을 뿐이다. 그 이유는 전자사전의 구축에 있어 특정주제에 대한 일정규모이상의 코퍼스가 구축되지 못했기 때문이다. 코퍼스를 이용한 전문용어사전을 구축하기 위한 대상데이터 선정방법은 다음과 같다.
1)주제분야를 선정한다. 2)코퍼스를 선정한다. 대상 코퍼스는 해당 주제분야 텍스트, 전문용어사전, 고빈도 이용전문서적, 해당분야 학위논문, 정기간행물을 선정한다.

6. 결론 및 향후과제

본 연구는 지능형 정보검색시스템에서 활용되는 지식베이스를 구축하기 위한 데이터를 입수하기 위한 것이다. 이러한 지식베이스를 얻기 위해 통계적 정보를 활용하기 위해 실제 코퍼스를 구축하였다. 본 코퍼스는 균형의 관점을 주제분야와 데이터의 수준으로 선정하였다. 분석된 결과는 전자사전의 개발에 사용되었으며, 최종적으로 분석된 정보는 명사사전과 공기사전, 전문용어사전을 개발하기 위한 데이터입수를 위한 일련의 절차와 기술이었다. 본 연구는 코퍼스를 완전하게 구축하지 못한 시점에 이루어졌으며, 그 결과도 아직 완전히 분석된 것은 아니기 때문에 계획(안)과 실제 수행결과의 차이를 분명하게 기술할 수 없었다. 한편, 완전한 전자사전을 구축하기 위해서는 다른 주제분야를 선정하여 초용량의 코퍼스를 구축하여 전분야의 색인시스템용사전과 전문용어전자사전을 구축해야 하고, 한편으로는 저작권을 해결해야 하는등의 산적한 많은 과제가 있다.

참 고 문 헌

[1] 김영택 외. 자연언어처리. 교학사.

[2] 김홍규, 강범모. "고려대학교 한국어 말모둠 1 : 설계 및 구성", 한국어학, 제3호 한국어학회 1996.

[3] 남영준. 정보검색을 위한 전자사전 구축 기술. 한국어정보검색기술세미나. 한국어정보처리학회 1996.

[4] 문화체육부, "국어 정보 처리 기반 구축을 위한 연구(2)." 1995년 학술용역 과제 보고서

[5] 임해창, 이상주, 이호. "전산언어학에서의 언어데이터베이스 활용", 한국어데이터베이스의 설계 및 응용을 위한 기초 연구. 민음사. 1995.

[6] 정광 외, 한국어 데이터베이스의 설계 및 응용을 위한 기초 연구, 서울 :민음사, 1995

[7] 한국과학기술원 "지능형 정보 검색에 관한 연구." '94 장기기초연구과제 최종보고서

[8] 한국과학기술원 인공지능연구센터, "1996년도 제 1회 우리말 정보처리 규격 심포지움."

[9] 한국정보관리학회, 한국과학기술원, "지능형 정보검색시스템." 1996년도 정보관리강좌. 한국정보관리학회. 1996.