특집

한글 코퍼스 구축에 관한 연구*

남 영 준
전주대학교 문헌정보학과

서 론

언어를 이해하기 위해 인간은 여러 가지 방법을 사용하며, 그 가운데 통계적인 방법은 인간의 언어 현상을 분석하는 중요한 알고리듬으로 사용되고 있다. 언어이해에 있어 통계적인 접근 방법은 인공지능기법에 비해 분석대상 영역의 제한이 없는 특징을 갖고 있다. 언어학 분야에서 통계적인 방법은 Corpus-based approach라 불리며, 언어 현상을 통계적으로 분석하여 그 결과를 언어해석에 다시 이용하는 방법을 의미한다. 통계적인 방법은 많은 지식을 필요로 하지 않으며, 단지 수학적인 방법에 의하여 처리를 함으로서 분석 결과에 대한 신뢰도를 적정한 수준으로 유지할 수 있다. 그러나 통계적인 방법을 사용하기 위해서는 일정규모이상의 원데이터를 분석한 통계 데이터를 가지고 있어야 하며, 이러한 통계데이터는 실제 언어를 분석함으로서 얻어진다.

언어학분야에서 많은 경비와 인력이 투입되면서도 코퍼스를 구축하는 것은 규칙기반접근방법의 한계를 극복하기 위해서이다. 즉, 자연어 처리에 도입된 통계적인 접근 방법을 이용하여 실세계의 다양하고 불규칙적인 언어현상에 관한 데이터를 코퍼스의 분석을 통하여 입수하는 것이다. 특히, 코퍼스는 코퍼스로서 가치를 지니기 위해서는 코퍼스 자체가 일정규모이상이어야 한다. 왜냐하면, 통계데이터가 그 결과의 질을 확보하기 위해서는 데이터자체가 일정규모 이상이어야하기 때문이다. 일반적으로 현대 코퍼스는 가능한 한 대규모, 대용량으로 구축하고 있는 추세이며, 이를 유지 관리하기 위해 코퍼스는 컴퓨터로 가독할 수 있는 전산화된 형태로 구축되고 있다.

한편, 본 연구에서 구축하고자 하는 코퍼스는 한국어로 이루어지는 기초국어자료로서 1980년대 이후의 자료가 그 대상이 된다. 이는 해당 자료들에 대한 저작권적인 문제와 디지털화하는데 많은 제

* 본 연구는 1996년도 전주대학교 인문과학연구소에서 지원한 연구비로 연구되었습니다.

약조건이 수반되는 것을 의미한다. 특히, 해당 자료의 디지털화는 디지털화 사업만으로도 많은 경비와 인력이 소요되기 때문에 현행 저작권법을 해결하지 못할 경우에는 코퍼스구축자체가 불가능하다. 본 연구는 이점에 착안하여 한글 코퍼스 구축에 뒤따르는 제약점과 코퍼스의 균형과 수준을 유지하는 구축방법을 제시하고자 한다.

1. 코퍼스의 정의 및 유형

Corpus는 언어현상을 분석하기 위해 인위적으로 수집한 사회 각 분야에서 활용되고 있는 언어의 집합이다. 그러므로 Corpus는 모든 언어 현상을 포함할 수 있을 정도로 크고 언어의 통계량을 대표할 수 있을 정도로 균형을 유지해야 한다. Corpus의 크기가 작은 경우에는 모든 언어 현상을 포함하지 못한다. 이러한 Corpus를 통하여 얻어진 정보(문법, 품사 정보, 의미 정보, 기타 통계 정보 등)는 실제 언어를 처리할 경우 발생하는 예외 현상에 대처할 수 없으므로 시스템이 불안정하게 된다. 안정되고 정확한 결과를 얻기 위해서는 대상 언어의 모든 현상(형태소 정보, 문법, 의미 정보, 관계 정보 등에 대한 통계 정보를)을 그대로 반영하는 Corpus를 선정하는 일이 필요하며, 이를 적절히 처리하여 원하는 정보를 추출하는 도구가 필요하다. 언어 현상이 다양하면 다양할수록, 언어사용의 변화가 심할수록 필요한 Corpus의 크기는 커진다. 코퍼스는 부속정보와 종류, 용례의 정확성, 장르별 분포, 언어의 시대성에 따라 코퍼스의 유형을 구분할 수 있다. 대표적인 것으로는 다음과 같은 것이 있다.

① 부가정보 : 원문코퍼스
② 텍스트구성 : 균형코퍼스, 피라미드코퍼스
③ 언어의 수 : 단일어 코퍼스, 병렬 코퍼스
④ 용례의 종류에 따른 구분 : 문서코퍼스, 음성코퍼스

본 연구에서 연구하는 것은 원문코퍼스로서 균형을 이룬 단일어로 이루어진 문서코퍼스이다.

2. 코퍼스의 설계

코퍼스는 특정한 시대에 사용된 언어현상을 파악하는 것을 목적으로 구축된다. 언어현상을 파악하기 위해서는 그 당시에 통용된 모든 언어행태를 수집하여 분석하는 것이 최상의 작업이 된다. 그러나 현실적으로 특정 시대의 모든 언어행태를 파악하고 이를 수집하는 것은 현실적으로 불가능한 작업이다. 차선의 방법으로는 사회 각 계층에서 통용되고 있는 언어행태를 예산과 시간이 허락하는 한 최대한 수집하는 것이다. 문제는 어느 계층의 언어를 어느 정도를, 어떤 방법으로 수집해야 하는 가이다.

코퍼스의 정형(prototype)으로 인정을 받고 있는 브라운 코퍼스(Brown Corpus)가 구축되면서부터 코퍼스는 균형을 중시하게 되었다. 즉, 코퍼스는 특성상 많은 인력과 경비, 시간이 소요되므로 가용범위내에서 특정 시기의 언어행태를 최대한 반영해야 하기 때문에 코퍼스대상의 적절한 균형이 절대 필요하다.

일반적으로 균형의 기준은 코퍼스 구축의 목적에 따라 조금씩 차이를 보이고 있으나, Sinclair는 균형코퍼스를 구축하기 위한 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. 이와 같은 기준을 고려하여 본 연구에서는 다음의 고려사항이 균형의 기준을 제시한다.

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

초창기 Brown과 LOB코퍼스는 100만 어절 수준으로 약 500권의 자료로 코퍼스를 구축하였다. 일반적으로 언어학적으로 사용되는 사전은 10만 항목이상이 되어야 실질적인 자연어처리시스템에 유용하다고 판단되므로(김영택 외. 206) 통계적 정보로 10만 항목을 확보하기 위해서는 최소한 그 100배이상이 되어야 한다고 판단된다. 그러므로 코퍼스가 최소한의 신뢰도를 확보하기 위해서는 1000만 어절 이상이 되어야 할 것이다.

둘째, 코퍼스는 대표성을 유지하는가.

예를 들면, 용례사전과 전문용어사전을 구축할

경우 너무 일반적인 분야와 말 위주(口語體)의 데이터를 구축할 경우 해당 코퍼스는 대표성이 없어지게 된다. 또한 수준 이하의 잡문을 분석데이터로 설정하였을 경우 분석된 결과는 해당분야의 대표성을 갖지 못한다.

셋째, 표본추출과정의 신뢰도를 확보한다.

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 신뢰도를 확보한다.

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, 저작권법에 저촉여부를 확인한다.

현재 국내의 여러 기관에서 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. 균형성을 확보하기 위해 국내외 대표코퍼스는 다음과 같이 코퍼스를 내용적으로 혹은 형태적으로 구분하여 수집하였다.

1) 브라운 코퍼스/LOB 코퍼스

1964년에 미국의 영어를 대상으로 브라운 코퍼스(Brown Corpus)가 구축되었으며, 이것이 균형코퍼스(balanced corpus)의 시초였다. 영국에서는 브라운 코퍼스의 균형기준을 그대로 수용한 LOB 코퍼스를 구축하였다. 양에서는 미미한 차이가 있었으나 균형기준은 차이가 없다. 이 두 코퍼스균형의 기준은 사실적(정보전달적 구분)과 비사실적 요소(문학적 구분)이다.

정보전달적(사실적) 요소	문학적(비사실적) 요소
신문 : 보도, 사설, 서평 종교서, 기술 및 취미, 상식자료 美文, 전기, 수필, 정부문서 및 기타, 학술과학서적	일반소설(일반, 추리, 과학, 모험, 연애) 유머집

2) 런던/룬드 코퍼스 (London/Lund Corpus)

1975년에 스웨덴에서 SEU 코퍼스 가운데 문자화되지 않은 데이터만을 기계가독형으로 변환하였으며, 이는 1960-70년대 영국 영어의 구어를 대상으로 구축된 코퍼스이다. 이 코퍼스의 균형기준은 인간의식을 기준으로 하였다.

균형기준	대화(의식)	독백(무의식)
내용	사적인 대화 대면대화 몰래 녹음함 미리 알리고 녹음함 전화 대화 공적인 대화	무의식적 독백 예비적 독백 말로 하기 위해 준비한 자료 글로 쓰기 위해 준비한 자료

3) 영국국립코퍼스(British National Corpus)

BNC는 코퍼스 사업가운데 최초로 국가사업으로 이루어졌다. 목표는 영국의 영어를 대상으로 2억 어절이상을 구축하는 것이다. 균형의 기준은 문어와 구어로 구분하고 있으며, 이를 세부적으로 주제와 형태별로 다시 구분하고 있다.

기준	사실적 언어 요소	사상적 기술 요소
내용	◆ 주류주제별 과학 (자연순수, 기술과학, 사회과학) 세계사, 상업과 재정, 예술, 종교 및 사상, 레저, 전기 ◆장르별 (책, 간행저작물,미간행 저작물 정기간행물, 대화기록물) ◆수준별 (전문가 초심자 일반인)	◆ 장르 대화체 소설 수필 연극대본 시 ◆ 수준 문학인 중간수준 일반인

4) 한국과학원 코퍼스(KAIST Corpus)

한국과학원 코퍼스는 국내에서는 최초로 1억어절 규모를 예상하고, 1994년도부터 코퍼스를 구축하고 있으며 1999년도에 1차 국어정보화사업이 종료되는 시점에 예상규모를 달성하는 것을 목표로 하고 있다. 이 코퍼스는 균형의 기준을 크게 2가지로 선정하였다. 첫째, 국내외 다른 코퍼스와 유사한 형태인 주제와 언어행태구분으로 구분된다. 둘째, 저작권을 기준으로 설정하여 저작권 저촉여부로 구분하였다.

균형을 위한 고려		저작권을 위한 고려	
구어체	문학저작물 드라마대본	저작권 불침해자료	법령 신문기사 학물저작물(일부:권리포기자료 초등학교 교과서(국정)
문어체	학술저작물 (교재 및 논문) 법 령 신문기사	저작권 침해자료	학술저작물(일부:권리양도유보 드라마대본 중고등학교 교과서(검인정)

5) 연세대 코퍼스

연세대 코퍼스는 2만쪽 정도의 종합 국어 대사전의 편찬을 완성하기 위한 목표로 구축되고 있다. 특징적인 것은 코퍼스 목적이 일반국어사전을 편찬하는 것이기 때문에 사회 각 분야의 언어현상을 최대한 입수할 수 있도록 설계하였다. 따라서 균형의 기준을 특정 주제에 맞추기보다는 매체특성을 기준으로 하였다.

매 체 종 류	비 율
신 문	33%
잡 지	16%
소설 및 수필	18%
취미 및 교양	10%
수기, 전기 및 실화집	9%
교과서	5%
방송스크립트	5%
계	100%

6) 고려대 코퍼스

고려대학교 코퍼스(고려대학교 한국어 말모둠)는 1995년에 민족문화연구소주관으로 1,400만 어절 규모로(KOREA-1 코퍼스) 구축되었다. 균형의 기준은 구어와 문어로 이루어졌으며, 문어는 자료의 형태(신문, 정기간행물, 책자)와 내용(사실적 묘사와 비사실적 묘사)으로 하였다.

구 분		분포도
1. 구어/준구어	(약 120만어절)	11.7%
2. 신문	(약 200만어절)	20.7%
3. 잡지	(약 100만어절)	9.8%
4. 책-정보	(약 350만어절)	33.5%
5. 책-상상	(약 210만어절)	21.0%
6. 기타	(약 20만어절)	3.3%
계 (총 10,082만어절)		100%

7) 국내외 코퍼스균형의 분석

국내에서 간행된 코퍼스는 Brown/LOB 코퍼스를 母本으로 하여 크게 구어와 문어로 구분하고 있다. 특히, Cobuild와 연세대 및 고려대 코퍼스는 사전편찬을 염두에 두고 구축하였기 때문에 현대

사회에서 통용되고 있는 언어현상을 전반적으로 입수하기 위한 기준을 설정하였다. 이에 비해 과기원코퍼스도 1996년도 이전판까지는 구어와 문어로 구분하여, 사전편찬용 균형을 이루었으나 1996년도부터 사전편찬뿐만 아니라 자동번역시스템 개발까지도 염두에 두고 코퍼스를 구축하고 있다. 특히, 1997년도부터는 저작권의 저촉여부까지도 균형의 기준으로 설정하여 코퍼스를 구축하고 있다.

3. 구축방안

1) Corpus의 수집과 선정

Corpus는 여러 경로를 통하여 얻어진다. 컴퓨터 처리를 하기 위해서는 기계 가독형(Machine readable) 형태의 문서가 필요하다. 그러나 파일 형태로 된 문서만을 수집한다면 수집된 문서의 영역이 한정되어 실제 언어행태를 반영하지 못한다. 따라서 진정한 Corpus는 신문, 문학서적, 보고서, 개인적인 메모 등을 모두 포함하여야 하며 이것은 대부분이 인쇄된 문서를 파일 형태로 변환하는 과정을 필요로 한다. 초기에는 대부분의 Corpus의 구축은 Key-in방식에 의하여 이루어 졌으며, 최근에는 OCR(문자 인식기)등의 도구를 이용하고 있다. 이렇게 선정된 문서들은 Corpus 구성 계획, 실용성, 무작위성을 고려하여 선정된다. 이중 가장 중요한 선정 기준은 무작위성이라고 볼 수 있다. 무작위성을 배제한다면 선정된 Corpus가 실제 언어 현상을 그대로 반영한다고 볼 수 없기 때문에 여기에서 얻어지는 통계 자료에 대한 신뢰성이 떨어진다. Corpus의 크기는 주어진 여건에 따라 다르게 정해진다. Corpus 수집의 목표, 자금, 기술수준에 따라 100만 어절 또는 그 이상의 문서가 Corpus로 선정되고 있다.

2) Corpus의 형태

Corpus의 형태는 쓰이는 용도에 따라 크게 세가지로 구분할 수 있다. 본 연구에서 분석하는 것은 아무런 가공을 하지 않은 문서 그 자체인 raw corpus이다. 이를 처리하여 1차 처리가 된 Corpus가 형태·통사정보가 부착된 tagged corpus 이며, 이 가공코퍼스는 이때 분석된 결과인 각 단어에 품사가 반드시 하나로 결정된 형태이다. 이를 바탕으로 구문 정보를 입수하여 구문 정보가 부착된 구문코퍼스가 있다. 일반적으로 코퍼스라 함은 raw corpus를 의미한다.

3) Corpus 구축방법

(1) 키인방법

기초자료코퍼스를 구축하기 위한 가장 원시적이고 정확하게 입력의 질을 확보할 수 있는 방안이 키인(key-in)방식이다. 이 방식은 방식명대로 기초자료로 선정된 데이터를 사람이 직접 키보드를 사용하여 입력하는 방식이다.

이 때 소요되는 경비는 자료 입력비와 입력자료 검수비로 구분된다. 이 때 소요되는 경비를 분석하면 다음과 같다.

일반적으로 시중에서 판매되는 학술문헌의 본문양은 일반적으로 약 300페이지의 규모로 구성되어 있다. 한 페이지는 200자 원고지 4-5매정도의 분량이다. 이를 원고지 4매로 환산할 경우 한 권의 학술문헌의 양은 200자 원고지 2,400매 정도의 분량을 유지한다. 이를 문자수로 환산하면 대략 480,000문자에 해당한다. 이를 전산입력한다고 가정할 경우에, 1분에 400타(문자)를 입력하는 전문입력자의 경우에 한 권의 문헌을 입력하는데 대략 1,200분이 소요되며, 시간단위로 환산하면 학술문헌을 대상으로 한 권을 입력하는데 약 20시간이 소요된다.

이를 일당으로 환산할 경우, 전문입력자가 하루 8시간 근무를 조건으로 1권 자료의 입력에 2.5일이 소요된다. 10권 입력에 대략 25일이 소요되므로 법정공휴일과 주말을 고려하면 대략 1명의 전문입력자가 10권 입력하는 소요되는 근로의 양과 한달 노동의 양은 거의 유사한 수준이라고 간주할 수 있다. 현재 국내 전문입력인의 경우, 한달 평균 인건비가 최소 100만원으로 간주된다. 이 경비는 전문 근로자가 차지하고 있는 공간비와 사무소모비를 제외한 순수 경비만으로도 한 달에 학술논문 10권만을 전산입력하는 금액이다. 그러므로 약 1권의 학술자료의 입력에 소요되는 경비는 10만원에 해당한다.

이와는 별도로 입력자료 검수비로서는 한 권에 대해 약 8시간[2]이 소요되며 비용으로는 약 1만원

이 소요된다. 그러므로 약 300페이지정도의 학술문헌 한 권을 전산입력하는데 약 11만원이 소요된다. 이 비용은 앞으로도 지속적으로 상승할 것으로 예상한다면, 기타 코퍼스 구축방법에 비해 상대적으로 많은 비용이 소요될 것이다.

장점으로 이 방안은 학술문헌에 나타난 특수기호 혹은 한자, 외국어원문을 입력자가 그대로 입력할 수 있으므로 코퍼스 대상자료선정시 자료구성 내용에 영향을 받지 않는다. 상대적으로 코퍼스 입력의 질이 가장 안정적이다.

또한, 1차 데이터 구축후 검수비용이 상대적으로 저렴하며, 검수자의 노동의 양이 상대적으로 적은 수준이다.

한편, 구축에 소요되는 전체 비용은 일반적으로 기초자료코퍼스가 언어학적 혹은 전산학적으로 활용 가치를 유지하기 위해서는 최소한 5천만 어절 이상규모의 코퍼스가 구축되어야 한다. 5천만어절의 코퍼스라 함은 2억개의 문자로 구축된 방대한 기초자료 데이터베이스를 의미한다. (5,000만×4문자)

이 크기는 국내에서 간행된 학술문헌(약 300페이지분량)가운데 약 625권분량에 해당하는 장서군이다.

이러한 규모의 코퍼스를 구축하기 위해서 소요되는 경비는 수작업으로 처리할 경우, 순수 입력경비(한권당 11만원)로 약 7천만원의 금액이 필요하며, BNC코퍼스와 같이 2억어절이상의 코퍼스를 구축한다면 약 3억원이 소요된다.

이 비용에는 별도의 2차교정에 필요한 검수비용이 포함되지 않았으며, 저작권에 관계된 일체의 경비는 제외된 비용이다.

(2) 스캐너활용방안

기초코퍼스를 구축하는데 노동 및 경비를 절감하기 위해 하드웨어를 사용하는데, 그 방안가운데 가장 일반적인 것이 스캐너를 활용하여 문자 인식을 하는 것이다.

이 방식으로 코퍼스를 구축하는 과정의 많은 부분이 컴퓨터가 활용되고 있으며, 현재 완전한 자동인식은 아직 이루어지지 않고 있다. 그 과정을 도식화하면 다음과 같다. 이 방안은 스캐너와 문자인식소프트웨어에 따라 효율이 크게 차이가 발생한다.

하드웨어(스캐너)와 소프트웨어(문자인식기)에 관계없이 스캐너를 활용하는 코퍼스 구축에는 대체적으로 다음 과정을 거친다.

자료의 선정-스캐닝-문자인식-결과 편집-구축자료검수

이 과정가운데 수작업으로 하는 부분은 ① 자료의 선정을 비롯하여 ② 스캐닝, ⑤ 결과 편집, ⑥ 구축자료검수이며 이에 대한 수작업은 정확히 구분된 데이터는 없으며 통계적으로 ①~④ 과정의 경우 하루에 약 2권정도를 처리할 수 있다.

이 때 야기되는 문제로는 현재 국내에서 개발된 문자인식소프트웨어[3]의 경우에 문자 인식율이 약 60%미만에 머무르고 있어 구축자료 검수과정에 2차적으로 많은 시간과 경비가 투입되어야 하는 점이다.

한편, 이와는 별도로 결과편집과 구축자료검수에 소요되는 시간은 12시간정도가 소요됨에 따라 약 1.5일정도의 시간이 소요된다.

1) 단면 플랫베드형 스캐너를 활용할 경우

스캐닝 방식에 소요되는 경비를 수작업 방식과의 비교를 위해 10권의 데이터를 디지털화하는데 소요되는 경비를 산출하면 다음과 같다.

스캐너로 한면인쇄에 소요되는 시간은 약 30-60초이다. 300페이지 분량의 책 한 권을 스캐닝하는데 소요되는 시간은 약 2시간 30분이 소요된다. 그러므로 책 2권을 스캐닝(흑백 300dpi 수준)하는데 소요되는 시간은 대략 5시간이 소요된다 (150면(2페이지)×60초=9,000초=150분=2시간 30분) 문자인식에 소요되는 시간은 150면의 데이터(그래픽)를 인식하는데 대략 2시간 정도가 소요된

2) 현재 전주대학교에서 구축하고 있는 기초자료코퍼스의 평균데이터.

3) 1996년도에 개발된 제품을 대상으로 실험하였음.

다. 책2권을 문자인식하는데 소요되는 시간을 환산하면, 대략 스캐닝에 소요되는 시간 5시간, 문자인식에 4시간이 소요된다. 즉, 하루 2권을 인식하는데 약 9시간으로 1일이 소요된다.

이 작업의 특징은 전문 입력가의 수고를 덜어줄 수 있을뿐더러 문자인식과정에는 인간의 노력이 최소화됨으로서 인력 낭비를 최소화할 수 있다는 점이다.

단점으로는 앞에서 설명한 바와 같이 인식소프트웨어 자체가 갖고 있는 오인식률로 인하여 후처리 작업이 수작업 방식에 비해 훨씬 많이 소요된다는 점이다. 후처리에 소요되는 시간은 1권당 4시간에서 6시간이 소요된다. 결과적으로 2권을 후처리하는데 소요되는 시간은 8-12시간으로서 1-1.5일이 소요된다.

일반적으로 소요되는 경비는 수작업처리의 2/5 정도의 비용과 노력이 투입된다.

2) 양면 스캐너를 활용할 경우

현재 국외에서 시판되고 있는 가장 일반적인 양면 스캐너는 1분에 약 200매(200dpi 수준)를 스캐닝한다. 문자 인식에 소요되는 시간은 150면의 데이터(그래픽, 150페이지)를 인식하는데 약 1시간 정도가 소요되나, 문자인식율은 플랫베드형(단면인쇄)과 거의 동일한 수준이다. 그러므로 300페이지 분량의 학술문헌을 스캐닝하고 인식하는데 소요되는 시간은 2시간 1분 30초(스캐닝 1분30초, 인식 2시간)로서 약 2시간정도가 소요된다. 하루에 8권이상을 인식할 수 있다. 책2권을 후처리하는데 소요되는 시간은 플랫베드형과 같이 8-12시간이 소요되며, 낱장단위로 스캐닝하는데 소요되는 경비는 크게 절감될 수 있다. 이를 전문입력가(키인 방식)가 하루 2권의 자료를 교정한다고 가정할 경우 한 달에 약 50권정도의 문헌을 디지털화할 수 있다. 이는 키인방식에 비해 1/5수준에 불과하며, 코퍼스 구축에 많은 시간을 절약할 수 있는 장점도 아울러 갖고 있다.

단점으로는 양면 스캐닝을 위해서는 반드시 분석대상이 되는 책등(書背: book spine)을 잘라야 하는 책의 파손이라는 결정적인 문제점을 갖고 있다. 또한, 양면 스캐너의 가격이 플랫베드형의 40배가 넘기때문에 하드웨어의 가격이 높다는 단점도 있다.

스캐너 활용의 가장 큰 제한점으로는 한자와 같이 한글과 알파벳을 제외한 외래어와 특수문자의 인식자체에 어려움이 있기 때문에 학술문헌에서 자주 출현하는 수식이나 각주를 본문대로 표현할 수 없다는 점을 들 수 있다.

소프트웨어로서 문자인식프로그램도 자연언어처리응용 분야가운데 하나로서 자주 인식에 실패하는 문자와 문장구조에 대한 통계적 데이터를 확보할 경우 현재 지적되고 있는 인식률을 획기적으로 높일 수 있는 알고리듬을 입수할 수 있다.

(2) 전자자료입수방안

전자자료입수방법은 코퍼스 구축에 가장 손쉬우면서도 가장 경제적인 방법가운데 하나이다. 전자자료는 크게 영리를 목적으로 하는 전자도서와 비영리를 목적으로 일반 통신선상에 공개된 자료이다.

① 전자도서입수방법 (상업용)

국내에서 한글로 구성된 전자도서를 입수할 수 있는 곳은 본 연구에서 조사된 바에 따르면 4개 기관을 통해 디지털자료를 구입할 수 있다. 각 기관에서 보유한 전자도서의 주제는 소설류와 만화, 수필류로 크게 구분된다.

소설의 경우는 무협소설, SF소설, 일반소설류로 구어체의 자료와 픽션적 성격이 강한 자료로 구성되어 있다. 이에 비해 수필류는 논픽션적인 자료가 많으며, 만화는 국내에서 생산된 자료가 대부분이다.

② 전자자료 입수방법 (비상업용)

웹을 비롯한 국가전산망, 상업망에 올라 있는(upload) 많은 공개데이터가 있다. 특히, 웹상에서는 매일 발간되는 신문과 방송뉴스원고를 입수할 수 있다. 이와는 별도로 국가에서 제공하는 각종 법령과 백서, 보고서 등이 공개되어 필요한 디지털 데이터를 입수할 수 있다.

한편, 첨단학술정보센터와 특정 대학교에서 제공하는 학위논문과 연구보고서 등도 온라인 상에서 입수할 수 있다. 드라마 대본과 같은 구어정보도 입수할 수 있어, 다양한 주제와 형태, 저자층을 망라할 수 있다. 주로 공개된 자료는 대부분 문어

체 자료가 많이 포함되어 있다.

4. 코퍼스구축과 저작권

국내외 저작권법은 디지털 자료로의 변환도 복제를 전제로 하는 행위로 간주하고 동일성 유지권과 같은 권리로서 이를 보호하고 있다. 이러한 보호는 디지털 데이터가 갖고 있는 여러 특성가운데 복제전송의 극적성으로 인한 원저작자의 피해를 사전에 방지하기 위해서이다. 이에 비해 코퍼스 구축은 복제의 의미보다는 형태의 변환에 해당하는 일이다. 특히, 코퍼스의 구축은 자료의 유통적 측면보다는 자료의 사적 이용에 해당한다. 왜냐하면, 코퍼스는 불특정 다수인을 대상으로한 구축되는 것이 아니며 제한된 특정 주제분야의 학자들을 대상으로 구축되기 때문이다. 이러한 차이점에도 불구하고 코퍼스의 구축은 현행법에 저촉된다는 것이 일반적인 통설이다. 단, 코퍼스의 구축행위가 다음과 같은 법적 해석의 여지가 있기 때문에 저작권적 해석이 가능할 수 있다.

① 자유 사용과 코퍼스

국제 조약상 베른 협약(제9조2항)은 자유사용(possible exceptions)에 관하여 1)복제가 저작물의 정상적인 이용(nomal exploitation)과 저촉되지 않고 2)저작권자의 정당한 이익을 부당하게 해치지 않는다는 조건이 모두 충족된다면, 특정한 경우에 동맹국은 법으로 보호받는 저작물의 복제를 인정할 수 있다고 규정하고 있다. 한편, 세계저작권협약에서도 복제권과 공연권, 방송권에 대하여도 각 권리에 합리적인 정도로 효과적인 보호를 준다면 예외를 인정하고 있다.(제4조2항)

코퍼스는 활용방안에서 알 수 있듯이 저작물의 내용 유포를 전혀 고려하지 않았기 때문에 해당 저작물의 유통력을 전혀 저하시키지 않는다. 즉, 코퍼스는 저작물의 통상적 이용인 내용의 열람을 위주로 디지털화한 것이 아니며 저작자의 정당한 이익을 부당하게 침해하지도 않고 있다.

② 인용과 코퍼스

현행법은 인용에 대해 공표된 저작물은 보도, 비평, 교육, 연구 등을 위하여 정당한 범위 안에서 공정한 관행에 합치되게 인용할 수 있다(제25조) 고 규정하여 인용의 한계로서 정당한 범위내일 것과 공정한 관행에 합치될 것을 제시하고 있으나 구체적 의미는 결국 해석과 판례에 맡겨져 있다.(저작권심의조정위원회;pp.187-188) 인용에 해당하는 기준은 법적으로 명확하게 규정되지 않고 있으나, 경쟁관계적 측면에서 이를 결정할 수 있다. 경쟁관계라 함은 동일한 분야의 이용자를 대상으로 이루어진다. 즉, 원저작물과의 경쟁관계에 있으므로 하여 원저작물을 대신하게 하거나 원저작물의 가치를 떨어뜨릴 수 있는 결과에 이르게 한다면 인용이라기 보다는 표절에 해당한다고 할 수 있다.

코퍼스가 인용에 포함될 수 있는 이유는 다음과 같은 코퍼스의 특징때문이다.

첫째, 코퍼스는 동일한 분야의 이용자를 대상으로 삼지 않기 때문에 경쟁관계가 성립되지 않는다. 원자료가운데 문자형태의 데이터를 제외하고는 어떠한 내용도 필요로 하지 않는다. 즉, 코퍼스 구축은 컴퓨터로 표현이 가능한 기호에 한하여 처리할 수 있다. 원자료의 형태(페이지나 글자크기, 삽도 여부, 기호 등)를 그대로 인용할 수가 없기 때문에 자료로서 가치가 상대적으로 작다. 둘째, 코퍼스는 근거제시를 위해 반드시 출처를 명시하기 때문에 인용의 형태로서 저작권법에 조정될 수도 있음을 의미한다. 일반적으로 코퍼스는 각각의 자료를 별도로 관리하기보다는 여러 데이터를 하나의 파일로 구축하여 분석하는 형태를 취한다. 즉, 하나의 파일로서 언어학적 데이터로 활용되기 위해 여러 자료로부터 인용의 형태를 취하고 있다.

③ 코퍼스의 보호

완성된 코퍼스는 하나의 커다란 데이터베이스이다. 이러한 데이터베이스가 개발되기 위해서는 많은 인력과 경비가 소요된다. 국내에서도 대규모의 코퍼스가 극히 일부 대학에서만 구축되는 것도 많은 예산이 필요로 하기 때문이다. 국내외적으로 데이터베이스에 대한 저작권적인 해석은 이차적 저작물로서 데이터베이스 개발자에 대한 권리를 인정하고 있다. 코퍼스도 데이터베이스이기 때문에 당연히 보호받아야 한다.

왜냐하면, 코퍼스의 보호는 특정 주제 이용자

(문헌정보학자 등)이외에 사용될 경우에 원저작자의 권리를 침해할 수 있기 때문이다. 즉, 코퍼스의 보호는 코퍼스 구축자보다는 원저작자의 권리를 보장하는 행위이다.

5. 결론

코퍼스의 사용용도는 촘스키가 언어행태 분석을 위해 제시한 이후 전산언어학과 정보검색분야까지 폭넓게 사용되고 있다. 국내외 각 기관들은 자신들의 목적을 위해 코퍼스를 각자 구축하고 있으나, 대체적으로 사전편찬을 위해서는 사회전반에 걸친 자료들을 수집하기 위한 균형을 유지하고 있다. 균형의 기준이 되는 것은 대체적으로 다음 3가지 유형으로 귀결될 수 있다.

1) 구어와 문어 2) 사실적 묘사와 비사실적 묘사 3) 주제별 구분과 비주제별 구분

이상과 같은 균형기준외에 최근에는 저작권저촉여부도 중요한 균형기준으로 제시되고 있다.

코퍼스의 구축 방법으로는 인건비의 상승과 컴퓨터의 발달로 인하여 직접 수작업으로 입력하는 키인방식보다는 컴퓨터와 스캐너, 문자인식소프트웨어(OCR)를 이용한 구축방법이 점차 확산되고 있다. 구축에 소요되는 비용은 대체적으로 일반 학술문헌의 경우 전체를 입력하는데 키인방식으로는 약 11만원이 소요되며, 스캐너 활용방식은 키인방식은 2/5정도가 소요된다.

현재 이상과 같은 비용적 문제외에 저작권의 해결은 현대와 같이 지적 소유권이 국제적인 외교문제로 비화되고 있기 때문에 코퍼스구축에 커다란 걸림돌이 되고 있다. 왜냐하면, 코퍼스의 구축은 전산화(digitized)를 의미하며 이는 원저작물의 동일성 유지권을 침범하는 행위이기 때문에, 이는 명백한 실정법 위반에 속하기 때문에 코퍼스 구축이 범법행위로 간주될 수 있기 때문이다.

이를 해결하기 위해서는 코퍼스 구축이 공정이용이나 인용과 같은 저작권 예외규정에 속한다는 사회적 인식의 확대와 이해는 향후 제정될 법과 판례에 긍정적인 영향을 미칠 수 있을 것이다.

참 고 문 헌

〔1〕 김영택 등. 자연언어처리. 교학사. 1994

〔2〕 남영준. "디지털 자료에 대한 저작권적 해석에 관한 연구", 정보관리학회지. 제14권 제1호. pp.161-181. 1997

〔3〕 남영준. "국어 형태.통사 태그 규격", 제1회 우리말 정보처리 규격 심포지움. 한국과학기술원 인공지능센터. pp.37-46. 1996

〔4〕 남영준. "한국어 정보베이스를 위한 형태·통사 태그 표준에 관한 연구", 인지과학. 제7권 제4호. pp.43-61. 1996

〔5〕 남영준. 국어 정보처리 기반 구축사업의 저작권 해결을 위한 연구. 학술용역과제보고서. 문화체육부. 1997

〔6〕 문화체육부. 국어 정보 처리 기반 구축을 위한 연구(2), 한국과학기술원. 1995년 학술용역 과제 보고서. 1996

〔7〕 임해창, 이상주, 이호. "전산언어학에서의 언어데이터베이스 활용", 한국어데이터베이스의 설계 및 응용을 위한 기초 연구. 민음사. 1995

〔8〕 저작권심의 조정위원회. 멀티미디어 시대의 저작권 대책. 동위원회. 저작권연구자료24. 1996

〔9〕 정광. 국어어휘 데이터베이스 구축에 관한 연구. 학술연구보고서. 국립국어연구원. 1996

저 자 소 개

南 泳 準

1960年 7月 25日生
1984年 2月 중앙대학교 문리과대학 도서관학과 졸업
1988年 2月 중앙대학교 대학원 도서관학과 졸업(정보학 석사)
1995年 2月 중앙대학교 대학원 문헌정보학과 졸업(정보학 박사)

1986年 3月～1992年 2月 중앙대학교 문리과대학 도서관학과 조교
1993年 3月～1997年 8月 전주대학교 이문대학 문헌정보학과 조교수
1996年 3月～1997年 8月 한국정보관리학회 이사
1996年 3月～1997年 8月 정보문화연구소 이사

주관심 분야 : 자동색인, 시스러스, 전자자작권법