디지털 자료에 대한 저작권적 해석에 관한 연구
- 코퍼스를 중심으로 -
A Study on the Copyright for the Digital Data

남영준(Young-Joon Nam)*

목 차

초 록

본 연구에서는 도서관자료가 문장분석데이터인 코퍼스로 사용(복제)될 때 해당 자료의 구축과 활용이 저작권에 저촉되는지를 분석하였다. 이를 위해 코퍼스의 정의와 사용분야, 공정이용이 저작권과의 어떠한 관련이 있는지도 별도로 분석하였다. 특히, '인간의 읽기(내용)' 보다는 '기계의 읽기(형태)' 이라는 측면에서 디지털 데이터가 사용될 때 도서관 자료에 대한 기존의 저작권적인 해석의 문제점을 고찰하였다.

ABSTRACT

The purpose of this paper is to analyze the legality of the restructuring and the using of the library data by corpus, the linguistic analysis data. In the process of the analysis, the definition of corpus is tried and its possible application areas are mentioned. It is also proved how its applications are related to the copyright.

In addition the problems in the present interpretations of the copyright for the library materials are analyzed in terms of 'to be read by machine' rather than 'to be read by mankind', especially when the data is stored in the forms of the digital data.

* 전주대학교 문헌정보학과 조교수
** 본 논문은 전주대학교 인문과학연구소 지원비에 의해 연구되었음
■ 논문 접수일 : 1997년 4월 23일

서 론

과거의 도서관들은 저작권에 대해 일종의 치외법권적인 위치에서 도서관 이용자들의 정보 요구를 훌륭히 충족시켜 주었다.

그 이유는 정보를 담고 있는 매체가 과거에는 대부분 책자형 형태로 이루어져 있기 때문에 정보가 전파되기 위해서는 해당 자료(원본 또는 복제본)가 제삼자에게 전달되어야 했다. 이러한 이유 때문에 복제의 확산 속도가 상대적으로 빠르지 않았으며, 해당 지적작업물에 대한 재산상의 피해 범위도 적었다. 특히, 책의 복제 및 사적 복사에 전통적으로 상당히 관대하였던 사회적 풍습 때문에 복제와 같은 저작권침해 행위에 대해서도 적극적인 대처가 없었다.

한편, 현대사회는 통신기반 시설의 확충 및 발전과 기억용량당 보관비용이 저렴해짐에 따라 과거의 형태와는 전혀 다르게, 정보의 전파속도와 복제 행위가 단기간 내에 이루어질 수 있는 여건이 마련되었다. 이에 따라 도서관과 같이 순수한 의미의 정보전달기관이나 연구자들은 복제행위에 대해서 과거에 무상으로 누렸던 관대함과 유보적인 재산권의 행사가 사라지게 되었다.

현재 출판된 자료들은 아직도 전통적인 책자형태를 유지하고 있으나, 전산화된 자료(digitized material)의 형태가 급증하고 있는 추세이다. 또한, 도서관에서 이미 소장하고 있는 자료들도 디지털화하여 도서관에 직접 방문하지 않고서도 가정이나 직장에서 필요한 자료의 서지 및 전문정보를 열람할 수 있도록 하는 온라인 전문서비스를 제공하기 위해 모든 도서관들은 인력과 경비를 투입하고 있다.

각 도서관들이 새로이 입수되는 자료 가운데에서 이미 디지털화한 자료는 제외하고, 나머지 모든 자료들에 대해 도서관측에서 일련의 방법(key in)을 사용하여 해당 자료를 전산소장을 원한다면, 이 행위가 새로운 저작물의 구성에 해당하는지와 아니면 '동일성 유지권'을 침해하여 저작권법에 저촉되는지의 여부, 전산화된 자료를 새롭게 가공한 데이터가 이차저작물에 해당하는지의 여부는 향후 정보화사회의 중요한 걸림돌이 되거나 아니면 발전을 위한 촉진제가 될 것이다.

이에 대한 분명한 법적 해석은 전산언어학에서 언어처리를 위해 반드시 필요한 기초 코퍼스(raw corpus)의 활용과 배포에도 동일하게 적용될 수 있다. 차이점은 도서관 내의 전산형태자료는 내용의 열람을 위한 것이라면, 언어학의 코퍼스는 문장의 구조와 형태를 필요로 하는 점이다. 본 연구에서는 이러한 유사점과 차이점을 고려하여 ASCII Code로 이루어진 디지털자료에 대한 정의와 저작권법상에서의 활용, 이에 수반되는 문제점을 분석하고자 한다. 연구의 결과는 향후 도래할 전자도서관시대의 자료활용과 전산언어학의 자료활용이라는 두 가지 측면을 만족시킬 수 있는 기초자료로 활용될 수 있을 것이다.

1. 자료 형태의 변화

미래자료의 형태가 현대인이 예측할 수 없더라도, 또한 어떠한 모습으로 변화되더라도,

미래의 도서관이 어떤 모습이 되더라도 인간은 눈으로 보고, 귀로 들으며, 손끝으로 느끼는 행위만은 변화하지 않을 것이다. 즉, 과거의 도서관에서부터 미래의 도서관에 이르기까지 도서관이 제공하는 자료의 형태와 서비스의 형태는 변할 수 있어도 도서관인들이 갖고 있는 기본적인 사명 즉, '신속하고 정확하게 이용자에게 원하는 정보를 제공한다' 는 사명도 변하지 않을 것이다.

도서관 사명은 변하지 않더라도 외부환경과 기술의 발달로 이용자 위주의 새로운 봉사가 제공되어야 함은 당연한 것이 될 것이다. 예를 들면, CD-ROM을 열람하기 위해서는 반드시 이를 열람할 수 있는 기계(컴퓨터 혹은 CD-Vision, 모니터, 스피커 등)가 필요할 것이다. 이러한 변화는 도서관으로 하여금 별도의 예산과 새로운 시설, 전문인력의 변화를 요구할 것이다. 특히, 변화가 요구되는 것은 도서관 소장자료를 이용한 복제행위에 대한 대처방안이다. 과거의 매체는 대부분 책자형태였으며, 특수자료(마이크로필름 등)는 복제행위자체가 어려웠을 뿐만 아니라 복제에 따른 소요비용이 높았기 때문에 대량적인 복사는 이루어지지 않았다. 그러나 컴퓨터를 비롯한 주변기기의 발달과 가격의 하락으로 인하여 일반인들도 누구나 복사가 가능하게 되었다. 책자형태의 자료에 대해서는 일반자료일 경우, 복사에 소요되는 비용과 원저작물을 구입하는 비용이 큰 차이를 나타내지 못하고 있기 때문에 이에 대한 복사는 원저작물의 가격에 많은 영향을 받게 된다. 오히려 저작물의 고정된 가격대와 큰 폭으로 상승하고 있는 복사비로 인하여 복사본을 구입하는 비용보다 원저작물의 구입비용이 저렴해지는 현상이 나타날 수 있다. 또한, 복사행위 자체가 복사기라는 특정한 기계를 필요로 하기 때문에 복사의 추적을 통하여 불법행위에 대한 노출과 적발이 용이하다.

그러나 디지털화된 자료일 경우에 저작물의 복제는 복사기에 비해 쉽게 이루어질 수 있으며, 또한 가격적인 면에서도 기억용량당 단가의 하락으로 인하여 원저작물가격에 비하여 훨씬 저렴하기 때문에 불법복제행위가 공공연하게 이루어지고 있다. 이러한 복제행위가 과거의 복사행위에 비해 피해규모나 피해액수가 훨씬 심각한 것은 전산복제의 경우는 컴퓨터와 같은 기기를 통해 거의 노출되지 않고서 이루어져 불법행위에 대한 추적과 적발이 현실적으로 불가능하기 때문이다.

이러한 문제점 때문에 도서관에서의 대출 특히 통신선을 이용하여 전문을 제공하는 원격정보제공서비스와 같은 디지털 데이터의 열람서비스는 위와 같은 범죄행위를 방조하는 결과를 제공할 수 있다.

2. 디지털자료의 특성

도서관에서 사용하고 있는 디지털의 의미는 일반적으로 사회에서 인식하고 있는 아날로그의 대비말은 아니다. 디지털자료는 기존의 책자형 데이터와 대비되는 말로서 컴퓨터와 같은 기계에 입력되어 모니터와 같은 출력기계를 이용하여 해당 내용을 볼 수 있는 자료를 의미한다. 디지털 자료가 기존의 책자형 자료와는 여러 가지 면에서 차이를 보이고 있

다. 그 대표적인 특성을 요약하면 다음과 같다.

① 형태적 차이 : 매체에 수록된 디지털 데이터는 육안으로 내용을 확인하기가 거의 불가능하다.

② 기술적 차이 : 디지털 데이터를 열람하기 위해서는 반드시 별도의 기기가 필요하다.

책자형태의 자료는 일반이용자의 경우 육안으로 열람을 하기 때문에 자료의 내용을 파악하기 위한 특별한 기기가 필요하지 않는 것에 비하여 디지털 데이터는 내용을 확인하기 위해서는 별도로 고가의 장비가 필요하다.

③ 복제의 속도 : 기존의 책자형태의 자료보다 복사가 용이하며, 복제에 소요되는 시간이 훨씬 짧아졌다. 특히, 복제가 이루어지고 난 후 원데이터와 복사데이터 간의 차이가 전혀 없다.

④ 전파의 신속성 : 책자형 자료가 유통되기 위해서는 물리적 형태의 전송을 담당하는 기관(예를 들면, 우편 전달 혹은 심부름)에 의해 전달이 이루어졌다. 물론 디지털 데이터의 경우도 이러한 기관을 통해 데이터를 전달할 수 있으나, 디지털 데이터의 전송하는 가장 일반적인 형태는 통신회선을 이용하는 것이다. 이에 따라 데이터의 전송속도가 기존의 전달방법보다 월등히 빨라지고, 비밀스럽게 이루어질 수 있기 때문에 디지털화된 데이터의 경우는 동시 발생적으로 전세계로 확산될 수 있다.

⑤ 공간의 무제약성 : 기존의 책자형태는 물리적 형태를 갖기 때문에 일정한 규모의 보관장소가 필요하였으나, 디지털 자료의 경우는 극히 적은 공간에 해당 데이터를 모두 보관할 수 있기 때문에 도서관과 같은 정보소장을 필요로 하는 기관들은 이러한 특징을 최대한 이용하고 있다. 미국의 경우 NSF(미국과학재단)의 도움으로 6개 도서관[1] 에서 Digital Library를 구축한 것도 시공을 극복하기 위한 투자 가운데 하나이다.

⑥ 데이터의 항구성 : 일반종이로 이루어진 책자 형태의 정보들은 햇빛과 습도, 기온 등 외부 환경 요인에 따라 다르지만 일반적인 상온에서 100년 이내에 외부적 변형이 나타난다. 이에 비해 디지털 데이터는 외부 환경에 거의 영향을 받지 않고 영구적으로 해당 데이터를 보관할 수 있다.

⑦ 복제의 용이성 : 디지털 데이터는 복제가 매우 용이하다. 기존의 책자형 자료는 복제시에 원본과 복사본에는 인쇄에 많은 차이를 보이고 있으며, 복사본을 재복사할 경우 그 질은 원본에 비해 점점 떨어진다. 또한, 복사에 따른 원본의 손상도 점차 심해진다. 이에 비해 디지털 데이터의 경우는 원본과 복사본의 차이는 전혀 존재하지 않는다. 특히, 도서관에서 제공하는 디지털 데이터는 대부분 텍스트 형태의 정보이기 때문에 일반적으로 프로텍터(복사방지프로그램)가 없어 정보이용자가 원한다면 언제든지 복제가 가능하다.

---

1) Carnegie-Mellon대학교와 Stanford대학교, UC Berkeley대학교, UC Santa Babara대학교, Illinois대학교, Michigan 대학교

3. 디지털자료의 종류

디지털 자료의 종류는 기계 가독 표현 형태에 따라 크게 두 가지로 구분할 수 있다.

3.1 그래픽 데이터

그래픽 데이터는 스캐너와 같은 기계의 도움으로 책자 형태의 데이터를 사진의 형태로 저장하는 것이다. 또한, 비디오와 같은 입력기기를 이용하여 기억공간에 직접 입력한 동·영상자료 등도 있다. 그래픽 형태의 데이터는 모든 데이터를 image형태로 저장한다. 이러한 데이터의 특성은 다음과 같다.

① 그래픽 형태로 저장된 데이터의 크기는 텍스트 데이터에 비해 기억용량을 상대적으로 많이 차지한다. 일반적으로 워드프로세서로 작성된 A4 사이즈의 내용의 경우, 텍스트 문서는 크기가 약 134kb를 차지하며, 그래픽 데이터는 약 1.51mb(비트맵이미지)를 차지한다.

② 파일 자체가 텍스트 형식에 비해 크기 때문에 하나의 파일을 다른 기억공간에 전송하는데 필요한 데이터의 전송시간이 상대적으로 길다.

③ 입력대상이 되는 자료를 원본의 형태로 입력되기 때문에 사진이나 수식과 같은 특수한 형상을 원본대로 입력이 가능하다.

④ 내용전달이 용이하다. 왜냐하면, 책자형 원본과 동일한 형태를 유지함으로써 원본에 표시된 모든 기호나 수식이 그대로 표현되기 때문이다.

⑤ 파일 형태에 따라 차이가 있다. 원자료의 형태가 컬러로 된 데이터일 때와 일반 문헌의 흑백 데이터일 때에 따라 파일의 크기에 많은 차이가 있다.

3.2 텍스트 데이터

텍스트 데이터는 ASCII 형태와 같이 Code 형태로 된 디지털 데이터를 의미한다. 텍스트 데이터는 대부분 Key-in 방법을 활용하든지, 혹은 OCR(Optical recognition character) 기기를 이용하여 그래픽 데이터를 문자코드로 변환하는 방식을 활용한다. 이 데이터는 그래픽 데이터에 비해 다음과 같은 특징을 갖고 있다.

① 문자 이외의 데이터를 표현하기가 어렵다. 예를 들면, 복잡한 수학공식이라든지 혹은 윗점, 사진을 포함한 동·영상을 텍스트 형식으로는 컴퓨터상에서 표현이 거의 불가능하다.

② 전문을 대상으로 검색시스템을 설계할 경우, 검색을 본문 전체를 대상으로 확대할 수 있기 때문에 검색알고리듬이 그래픽 데이터에 비해 상대적으로 자유롭다.

③ ASCII와 같은 표준 형태의 코드일 경우 이종간 컴퓨터 간에도 구축 데이터의 처리 및 공유가 가능하다.

3.3 디지털 데이터의 저작권법적 해석

디지털 데이터로 표현할 수 있는 저작물의 형태는 매우 다양하다. 예를 들면, 도서관 소장자료 즉, 단행본 형태의 도서와 정기간행물, 영화필름과 같은 비도서자료들이 모두 디

지털 형태로 전환이 가능하다.

국내 저작권법 상에는 디지털 자료에 대한 법리 해석이나 규정은 없으며, 다만 디지털 데이터가 수록된 매체에 상관없이 수록 내용에 따라 저작물을 구분하는 견해가 대부분이다.[2) 3)]

그 이유는 디지털로 표현할 수 있는 데이터의 형태가 너무도 다양하고, 기술의 발전에 따라 일상적이지 못한 분야까지 확대될 수 있기 때문이다. 특히, 본 연구에서 제시하고 있는 raw corpus의 경우 디지털 데이터이면서, 어문적 저작물, 2차 저작물의 성격을 모두 갖고 있다.

저작권법 제5조 1항에서 원저작물을 '번역 · 편곡 · 변형 · 각색 · 영상제작 그 밖의 방법으로 작성한 창작물(이하 "2차적 저작물"이라 한다.)은 독자적인 저작물로서 보호된다.' 이라고 규정하여, 2차적 저작물도 하나의 저작물로서 보호를 하고 있지만, 한편으로는 2차적 저작물을 구축하기에 앞서 원저작자에 대해 허락을 반드시 구할 것도 지적하고 있다.

베른 협약 제2조 5항에서 "소재의 선택과 배열에 의하여 지적 창작물이 되는 백과사전 및 선집과 같은 문학적, 예술적 저작물의 집합물은 지적 창작물로서 보호를 받으며, 그 집합물을 구성하는 각 저작물의 저작권에 하등의 영향을 미치지 않는다."라고 규정하고 있다. 이와 같이 데이터베이스를 보호하는 객체가 데이터베이스에 실려 있는 原데이터만이 아니며 집합물 자체로서도 저작성을 갖고 있음을 의미한다.[4)] 일본의 경우는 멀티미디어 제작과정에 있어서 저작물을 디지털 데이터화하는 자에 관해서, 예컨대, 디지털 데이터를 그대로 복제하는 것을 금지할 권리 등이 저작 인접권적 구성에 의해 창설해야 한다는 지적이 있다.[5)]

이상과 같이 정보가 책자 형태로 이루어져 있든지 혹은 디지털 형태로 이루어져 있든지 저작자가 구축한 지적 수고에 대한 것은 모두 저작권자의 권리로 인정하고 있으나, 디지털화 데이터에 대한 적합한 법적 해석은 국내외적으로 분명하게 제시된 것이 아직은 없는 실정이다.

3.4 디지털 데이터와 코퍼스

현대 언어학에서 코퍼스는 문헌정보학분야에서의 자료의 개념과는 전혀 무관한 데이터였다. 그러나 코퍼스의 유용성을 결정하는데 통계적 기초 데이터가 중시되면서 코퍼스는 대규모 대용량의 형태를 유지하게 되었다. 이러한 규모를 처리하기 위해서는 컴퓨터의 활용이 필연적이 되었으며, 컴퓨터의 가독을 위

---

2) 정성조. "멀티미디어 관련법 제도의 문제점과 개선방안", 한국저작권논문선집(Ⅱ), 저작권심의조정위원회. 1995. pp.47-64

3) 김진희. "초고속정보통신망과 저작권", 한국저작권논문선집(Ⅱ), 저작권심의조정위원회. 1995. pp. 83-94

4) 한승헌. 정보화시대의 저작권. 나남. 서울. 1992. pp.247-259

5) 문화체육부. 멀티미디어 시대의 저작권(1). 문화체육부, 1993.11. p.16

해서는 모든 코퍼스가 디지털 형태로 변환되었다. 이러한 코퍼스 가운데 특히 기초 자료 코퍼스는 전자도서관의 전자문헌과 매우 유사한 형태를 유지하게 되었다.

전통적으로 도서관의 장서는 그 형태의 종류에 관계 없이 내용의 열람을 위주로 이용되고 있다. 이에 비해 코퍼스는 해당 자료의 내용보다는 문장의 구조와 형태를 위주로 이용되고 있으며, 도서관 자료의 열람 형태의 이용과는 차이를 보이고 있다.

이러한 차이점에도 불구하고 도서관의 장서와 코퍼스에 대한 유사한 법적 제한이 있는 것은 두 개의 데이터가 원자료의 내용을 전산화로 변환하여 원저작자의 재산을 침해할 수 있기 때문이다.

4. 코퍼스

코퍼스란 언어학에서 사용되는 단어로서 국내에서는 이에 대한 해석으로 말뭉치, 말모둠 등이란 대역어가 사용되고 있다. 사전적 정의에서는 말의 의미를 '사람의 생각이나 느낌을 입으로 나타내는 소리 또는 그 행위나 내용'으로 해석하며, 한편, 글의 의미를 '글자로서 나타낸 적발'이라고 정의하고 있다. 국내외에서 채록하고 있는 모든 코퍼스는 전산형태로 코드화한 글이기 때문에 말뭉치 혹은 말모듬이란 제한된 개념으로는 코퍼스의 개념을 모두 소화하지 못하고 있다. 그러므로 코퍼스는 커다란 글뭉치 혹은 글모듬, 글덩어리로 표현되어야 하고, 엄밀하게는 '말과 글덩어리'로 표현하는 것이 타당하다[6].

한편, 또 다른 해석으로서, 코퍼스는 '컴퓨터가 읽을 수 있는 형태로 지정된 자연어 용례들과 이들 용례에 대한 부속정보들의 묶음'을 말한다.[7]

코퍼스의 개발 이유는 언어학분야의 규칙기반 접근 방법의 한계를 극복하기 위해 자연어 처리에 도입된 통계적인 접근 방법을 이용하여 실세계의 다양하고 불규칙적인 언어현상에 관한 데이터를 입수하기 위해서이다. 언어를 이해하기 위해 인간은 여러 가지 방법을 사용하며, 그 가운데 통계적인 방법은 인간의 언어 현상을 분석하는 중요한 알고리듬으로 사용되고 있다. 언어 이해에 있어 통계적인 접근 방법은 인공지능기법에 비해 분석 대상 영역의 제한이 없는 특징을 갖고 있다. 언어학 분야에서 통계적인 방법은 Corpus-based approach라 불리며, 언어 현상을 통계적으로 분석하여 그 결과를 언어 해석에 다시 이용하는 방법을 의미한다. 통계적인 방법은 많은 지식을 필요로 하지 않으며, 단지 수학적인 방법에 의하여 처리를 함으로써 분석 결과에 대한 신뢰도를 적정한 수준으로 유지할 수 있다. 그러나 통계적인 방법을 사용하기 위해서

---

6) 남영준, "코퍼스를 이용한 정보검색용 전자사전구축에 관한 연구", 한글 및 한국어정보처리, 제8회 한글 및 한국어정보처리학술대회, 1996. pp.440-442

7) 임해창, 이상주, 이호, "전산언어학에서의 언어데이터베이스 활용", 한국어데이터베이스의 설계 및 응용을 위한 기초연구, 민음사, 1995. p.47

는 일정 규모 이상의 원데이터를 분석한 통계 데이터를 가지고 있어야 하며, 이러한 통계데이터는 실제 언어를 분석함으로써 얻어진다. 특히, 본 연구에서 주목하고 있는 것은 현대의 코퍼스는 코퍼스로서 가치를 지니기 위해, 또한 통계데이터의 질을 확보하기 위해서는 일정 규모 이상으로 규모를 유지하여야 한다. 일정 규모 이상의 양을 유지하고 이를 처리하기 위해서는 수작업으로 처리가 불가능하기 때문에 컴퓨터를 이용하여 처리할 수 있도록 코퍼스는 전산화된 형태로 구축되어야 한다.

4.1 코퍼스의 유형

일반적으로 코퍼스라 함은 기초코퍼스에 부착된 부가정보를 기준으로 분류가 이루어진다. 가장 일반적인 구분은 기초코퍼스 〈그림 1〉와 형태.통사 태그주석 코퍼스 〈그림 2〉, 구문구조 트리구조 코퍼스 〈그림 3〉로 분류한다.

① 기초 코퍼스 : 문장을 분석하기 위한 원시데이터로서 책자형 문헌이나 음성을 ASCII형태로 변환 또는 채록한 파일이다. 외견적으로는 일반 책자형 문헌을 디지털화한 형태이다.

② 형태 · 통사 태그주석 코퍼스 : 기초코퍼스를 입력받아 형태소 단위로 분리하여 분리된 형태소들이 정당한 배열인지를 검사하여 해당 어절에 대해 일정한 기호(일반적으로 품사분류기호)를 부착한 2차 가공 형태의 코퍼스이다. 이 분석단계는 문장 분석 가운데 형태소 분석에 해당된다.

③ 구문구조 트리구조 코퍼스 : 형태 · 통사 태그주석 코퍼스를 입력받아 하나의 문장단위로 분석한 코퍼스이다. 일반적으로 수지도 형태로 제공되며, 분석 문장이 함께 출력된다. 이 코퍼스는 문장 분석 가운데 구문 분석에 해당된다.

第1章 總則
第1條 (目的)이 法은 憲法裁判所의 組織 및 운영과 그 審判節次에 관하여 필요한 사항을 정함을 目的으로 한다.
第2條 (管掌事項) 憲法裁判所는 다음 사항을 管掌한다.
1. 法院의 提請에 의한 法律의 違憲與否 審判
2. 彈劾의 審判
3. 政黨의 解散審判
4. 國家機關相互間, 國家機關과 地方自治團體間 및 地方自治團體相互間의 權限爭議에 관한 審判
5. 憲法訴願에 관한 審判
第3條 (구성) 憲法裁判所는 9人의 裁判官으로 구성한다.
第4條 (裁判官의 獨立) 裁判官은 憲法과 法律에 의하여 그 良心에 따라 獨立하여 審判한다.

〈그림 1〉 기초코퍼스의 예

기도는	기도/ncpa+는/jxc
놀라운	놀랍/paa+운/etm
모험이며	모험/ncpa+이/jp+며/ecc
아름다운	아름답/paa+ㄴ/etm
체험이다	체험/ncpa+이/jp+다/ef
.	./sf
우리의	우리/npp+의/jcm
생활과	생활/ncpa+과/jcj
환경을	환경/ncn+을/jco
변화시키는	변화/ncpa+시키/xsv+는/etm
힘을	힘/ncn+을/jco
얻을	얻/pvg+을/etm
수	수/nbn
있게	있/paa+게/ecx
된다	되/px+ㄴ다/ef
.	./sf

〈그림 2〉 형태 · 통사 태그주석 코퍼스의 예

```
: 내가 도쿄를 떠날 무렵에는 흑인남자가 아주 큰 인기를 끌고 있었다.
(SS
   (VP
      (VP
         (VP
            (NP
               (VP (NP 내/npp ) +가/jcs
                  (VP (NP 도쿄/nq ) +를/jco
                     떠나/pvg )) +ㄹ/etm 무렵/nbn ) +에는/jca
            (VP (NP 흑인남자/ncn ) +가/jcs
               (VP
                  (NP
                     (ADJP (ADVP 아주/mag ) 크/paa ) +ㄴ/etm
                     인기/ncn ) +를/jco
                  끌/pvg ))) + (AUXP 고/ecx 있/px )) + (AUXP 었/ep )) +다/ef+./sf )
```

〈그림 3〉 구문구조 트리구조 코퍼스의 예

이러한 구분 외에 코퍼스를 용례에 따라 부속정보나 종류나 용례의 정확성, 장르별 분포, 언어의 시대성에 따라 다음과 같이 구분할 수도 있다.[8]

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

② 텍스트구성
- 균형코퍼스
- 피라미드코퍼스
- 과도기적 코퍼스

③ 언어의 시기
- 공시코퍼스
- 통시코퍼스

④ 언어의 수
- 단일어코퍼스
- 병렬코퍼스

⑤ 용례의 종류에 따른 구분
- 문서코퍼스
- 음성코퍼스

4.2 코퍼스의 응용범위

코퍼스는 전통적으로 전산언어학에서 활용되었으나 시소러스와 같은 정보검색분야에서도 용어 간의 관계와 디스크립터의 설정과 같은 분야에서도 이의 구축은 반드시 필요한 것이 되었다. 일반적으로 코퍼스는 크게 다음 두 분야에서 활용되고 있다. 첫번째는 전산언어학적인 활용이며, 나머지는 문헌정보학적인 분야에서의 활용이다.

4.2.1 전산언어학 분야

전산언어학분야에서는 다음과 같은 언어학적 정보와 통계적인 정보를 입수하기 위해 코퍼스를 구축한다.

① 언어생성정보

일반적으로 언어를 이해하기 위해 언어분석이 여러 방면에서 시도되고 있으며, 이에 따른 많은 연구가 이루어지고 있다. 기계번역 등의 분야에서는 언어의 이해뿐만 아니라 언어의 생성도 중요한 역할을 한다. 언어를 분석하여 의미구조를 생성하면, 이것을 다른 언어의 표현으로 나타내는 것이 언어 생성이다. 언어생성의 경우 단순히 의미구조를 목적 언어로 변환하기만 한다면 단순하고 딱딱한 문체가 되기 쉬우며 자연스럽지 못한 문장을 생성할 수도 있다. 이러한 문제점을 다음과 같은 정보를 이용하여 해결할 수 있다.

i) 용례정보

용어정보는 문법상 오류가 있더라도 자주 사용되는 구나 어절을 파악할 수 있다. 이 정보는 문장에서의 단어의 쓰임과 문법의 쓰임 등을 검사하여 text critiquing(문장 오류분석)을 가능하게 하며, 철자 오류 검색과 철자교정, 문법 오류 검색이 포함된다.

ii) 句패턴정보 : 코퍼스의 문장들을 조사하여 구절들을 파악하고 그 패턴을 찾아서

8) 김덕봉. "국어 코퍼스 구축방안." 제1회 우리말 정보처리 규격 심포지움. 한국과학기술원 인공지능연구센터. 1996. p. 10

일반화시킬 수 있다.

iii) 문법구성정보 : 어휘와 구절, 관용어, 연어, 용례들의 파악이 이루어지고 나면 이를 토대로 코퍼스에 기반한 전산문법[9] [10]을 구성할 수 있다.

② 언어데이터베이스

통계적인 언어분석은 언어정보 즉, 사전 작성이나 문법작성, 시소러스 작성 등에 유용하다. 언어데이터베이스는 그 자체로는 의미가 없으나 언어분석시스템과 결합하여 중요한 역할을 수행한다. 사전은 형태소 해석단계에서 사용되고 형태소 해석 결과가 다시 사전을 개선하고 수정하는데 사용된다. 구문해석과 의미해석단계에서 필요한 문법, 시소러스 등도 각 시스템과 함께 개선될 수 있다. 이러한 시스템이나 언어정보의 개선은 전통적인 언어 분석이나 인공지능기법을 사용한 언어분석만으로는 이루어지기 어렵다. 코퍼스를 이용한 실제의 언어현상을 반영하고 여기서 입수되는 통계정보가 언어정보의 개선을 가능하게 한다.

4.2.2 문헌정보학 분야

문헌정보학에서도 얻고자 하는 것은 언어학적 정보이지만 그보다도 다음과 같은 통계적 정보에 기반한 용어추출 알고리듬이다.

① 어휘정보 : 말뭉치로부터 직접 어휘들을 파악하여 품사에 따라서 분류함으로써, 사전의 어휘항목을 설정하고 사전정보를 구축할 수 있다. 이는 시소러스나 혹은 전문용어사전을 구축할 때 전문가의 도움 없이 디스크립터를 생성할 수 있는 통계적 정보를 제공한다.

② 관용어정보 : 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.

③ 연어정보 : 연어(collocation)정보는 간단한 형태의 관용어이며, 연관된 어휘성분이 자주 어울려 특정한 의미를 갖게 되며, 코퍼스를 통해 해당 정보를 입수할 수 있다. 이 정보는 시소러스와 같이 용어와 용어간의 관계를 설정할 때 연관과 계층을 설정할 수 있는 중요한 정보로 활용될 수 있다.

4.2.3 코퍼스의 구축현황

앞에서 정의된 바와 같이 코퍼스는 그 중요성과 활용성으로 인해 국내외에서도 전산언어학과 문헌정보학적인 활용을 위해 대규모로 기초 자료 데이터베이스가 구축되고 있다. 그 대표적인 것은 다음과 같다.

① 연세대 코퍼스

연세대 코퍼스(연세 말뭉치)는 연세대 한

9) 남영준. "국어형태 · 통사 태그 규격", 제1회 우리말 정보처리 규격 심포지움. 한국과학기술원 인공지능연구센터. 1996. pp.37-46

10) 남영준 등. "한국어 정보베이스를 위한 형태 · 통사 태그 표준에 관한 연구", 인지과학, 제7권 4호. 1996. pp.43-61

국어사전편찬실에서 한국어에 관한 모든 정보를 체계적으로, 그리고 정밀하게 제시하는 2만쪽 정도의 종합국어대사전의 편찬을 목표로 구축하고 있다. 1993년 4월에 동아출판사의 지원으로 현대 한국어 학습사전을 우선 편찬하는데 연세대 코퍼스를 사용하고 있다. 1996년 7월 현재 기초자료는 약 4,500만 어절의 규모를 구축하였으며, 1960년대 이후의 자료를 입력 처리한 것이다. 이와는 별도로 200만 마디의 입말뭉치(口語情報)를 제작하고 있다.

② 고려대 코퍼스

고려대학교 코퍼스(고려대학교 한국어 말모둠)는 1995년에 민족문화연구소에서 1,000만 어절 규모로 구축되었다. 코퍼스의 구축은 1992년부터 이루어졌으며, 대우재단과 한샘출판사가 지원하였다. 그 후 1995년 11월에 대학생 논문자료 10만 어절이 추가되므로서 약 1,010만 어절의 규모를 유지하고 있다. 이 코퍼스는 일상 담화나 연설, 희곡과 같은 구어정보를 전사한 데이터가 포함되어 있다.

③ 한국과기원 코퍼스

한국과기원코퍼스는 1994년도부터 국가차원의 코퍼스를 구축하기 위해 매년 기초자료(raw corpus)와 가공자료(tagged corpus)를 구축 및 확장하고 있다. 1997년 현재 기초자료의 경우 7,000만 어절 수준의 자료를 구축하고 있으며, 또한 가공자료의 경우는 800만 어절을 구축하고 있다. 최종 목표는 1998년까지 기초자료의 경우 1억 어절을, 가공자료의 경우 1,800만 어절을 구축하는 것이다. 1997년까지는 대부분의 자료를 文語情報로 구축하고 있으며, 최종년도에는 일정량의 구어정보도 포함할 것으로 판단된다. 이와 함께 저작권 내에서 해당 코퍼스의 합법적 사용도 별도로 연구되고 있다.

이상과 같이 국내에서는 예산과 인력이 대규모로 (중복)투입되는 코퍼스가 구축되는 이유는 크게 다음 두 가지 이유에 기인한다. 첫째는 코퍼스는 정확한 언어정보와 통계정보를 확보하기 위해 대규모의 코퍼스가 유지되어야 한다는 학문적 특성 때문이다. 둘째는 공개나 자료 교환을 위해서는 저작권이 해결이 선결되어야 하나, 이에 대한 분명한 규정과 기준이 전혀 없기 때문이다. 왜냐하면, 코퍼스 가운데 기초자료코퍼스는 분명히 도서관에 전산소장하고 있는 자료의 형태와 조금도 차이가 없기 때문에 이의 공개는 원저작자에게 재산상의 피해를 입힐 것이며, 이는 명백한 저작권 침해에 해당하기 때문이다. 한편, 코퍼스의 사용은 해당 문헌의 내용적인 측면보다는 언어학적인 활용 방안을 위해 구축한 데이터이기 때문에, 엄밀하게 분석하면 원저작자의 재산상의 피해는 거의 없는 수준이 되고 있다. 국내법 가운데 이에 대해 법적인 해석과 기준이 가능한 것을 조사하면 다음과 같다.

4.3 저작권법상에서의 코퍼스

4.3.1 데이터베이스 보호법과 코퍼스

국내외의 저작권법에서 다른 디지털 데이터와 마찬가지로 코퍼스의 개념을 정의한 조

항이나 시행세칙은 없다. 코퍼스의 특징은 책자형태의 저작물을 컴퓨터가 읽을 수 있는 형태의 파일로 변환한 것이다. 이러한 특성에 가장 유사한 것으로는 데이터베이스에 관한 조항과 주장을 들 수 있다. 현행 법상에 데이터베이스는 인쇄 형태의 기록물과는 달리 "논문 수치 도형 기타 자료의 집합물로서 이를 정보처리장치를 이용하여 검색할 수 있도록 체계적으로 구성한 것"[11]으로 정의하고 있다. 데이터베이스에 대한 법적인 해석이 코퍼스에서 적용될 수 있는가를 판단하기 위해서는 일반텍스트 데이터베이스와 코퍼스의 유사점과 차이점을 분석할 필요가 있다.

① 유사점
- 정보처리장치를 이용하여 검색할 수 있도록 체계적으로 구성한 것이다.
- 디지털 자료이며, 2차 저작물에 포함된다.

② 차이점
- 코퍼스는 항상 코드화한 형태로 존재하나, 데이터베이스의 경우는 그래픽과 같은 다른 형태로 존재할 수도 있다.
- 코퍼스는 항상 문장에 관한 정보만을 갖고 있다.
- 코퍼스는 해당 자료에 대한 문헌 구입이 1차적으로 이루어진다.
- 데이터베이스에는 서지데이터베이스가 있어 해당자료의 서지정보를 중시하나, 코퍼스의 경우는 서지정보보다는 전문정보만을 구축한다. 특히, 전문에 출현한 사진이나 수식정보들은 코퍼스에서 제외한다.

데이터베이스에 대한 법리적인 견해는 국제적으로 데이터베이스를 어문저작물 내지 편집저작물[12]에 포함시켜 보호하고 있으며, 데이터베이스 보호를 위한 특별법이나 특별규정을 두고 있는 국가가 없을 뿐만 아니라 저작권 보호를 위한 국제협약인 베른 협약과 관련된 국제회의에서도 데이터베이스를 어문저작물로 보호하고 있다.[13] 단, 일본의 경우, 데이터베이스를 일반적 편집저작물에서는 제외하며, 정보의 선택 또는 체계적 구성에 의한 창작성이 있는 저작물로 간주하고 있다. 이러한 관점으로 코퍼스를 분석하면, 코퍼스의 경우는 문헌의 출판 시기나 글의 종류에 따라 해당하는 자료를 전산화한 것이기 때문에 소재의 선택과 배열에 창작성이 있는 편집저작물에 속한다고 볼 수 있다. 그러나 코퍼스는 데이터베이스가 갖고 있는 정보전달적인 요소가 거의 존재하지 않는 raw datum에 해당한다. 왜냐하면, 코퍼스의 구축은 대체적으로 1회적인 성격을 지니고 있기 때문이다.

기존의 코퍼스 가운데 전산화된 형태의 최초 코퍼스는 미국의 브라운대학에 있는 Nel-

11) 저작권법, 제6조 제1항

12) 어문저작물 : 소설 시 논문 강연 연술 각본 그 밖의 기타 글로 이루어진 창작물

편집저작물 : 이는 특수한 형태의 저작물로 분류되며, 소재의 선택 또는 배열에 창작성이 있는 저작물을 의미한다.

13) 저작권심의조정위원회, 저작권 상담.조정 사례집, 동위원회. 1992. pp.131

son Francis와 Henry Kucera에 의해 1961년부터 1964까지에 걸쳐 영문으로 구축된 Brown 코퍼스이다. 국내에서는 연세대학교에서 1980년대 말부터 한국어 코퍼스를 부분적으로 구축하기 시작하였으며, 1994년부터 한국과학기술원에서 과기원코퍼스를, 1994년부터 고려대학교도 대규모 코퍼스를 구축하였으며, Brown 코퍼스가 그 기준이 되었다. 이상과 같은 코퍼스들은 1회적으로 그 구축을 끝냈으며, 해당 코퍼스의 수정이나 관리, 개정은 이루어지지 않고 있다. 단, 국내 대학에서 이루어지고 있는 코퍼스 구축은 코퍼스의 양 자체가 일정 수준이 되지 않았기 때문에 일정한 양이 되도록 계속적으로 자료들을 입수하고 있다. 이는 서비스의 개선이나 새로운 정보의 입수 차원과는 전혀 상관없는 행동이다.

이에 비해 데이터베이스는 계속적인 정보 추가절차가 필요하며, 이용자에게 지속적으로 서비스를 계속해야 한다. 정보의 갱신은 데이터베이스 구축의 목적에 해당하기 때문이다. 이런 이유로 일본에서는 데이터베이스의 경우는 저작물 중 일부의 데이터만 공표되어도 데이터베이스 저작물이 공표된 것으로 보고 있다.[14]

코퍼스를 데이터베이스 보호조항의 적용에 문제가 되는 것은 데이터베이스는 해당 정보를 즉, 내용을 이용자에게 제공하는 것이지만, 코퍼스의 경우는 해당 정보의 의미와는 상관하지 않고 단지 글에 관심을 갖고 있다는 점이다. 저작권이라 함은 해당 저작물의 독창성과 아이디어의 표현, 학술적 수고를 보호하기 위한 보호벽이라면 사용한 글(한글) 자체까지 보호할 필요가 있는지에 대한 문제는 앞으로도 더 많은 연구가 필요할 것이다.

4.3.2 컴퓨터프로그램 보호법과 코퍼스

국내법에서는 소프트웨어에 대해서는 저작권법뿐만 아니라 컴퓨터프로그램 보호법을 별도로 제정하여 보호하고 있다. 이 법에서 "프로그램"이라 함은 특정한 결과를 얻기 위하여 컴퓨터 등 정보처리능력을 가진 장치(이하 "컴퓨터"라 한다.) 내에서 직접 또는 간접으로 사용되는 일련의 지시 · 명령으로 표현된 것을 말한다. (동법 제2조 제1항)

코퍼스와 언어분석용 소프트웨어는 전자의 경우는 원재료(raw materials)에 가까우며, 후자는 전자를 가공하는 응용프로그램(utility program)이라 할 수 있다. 코퍼스는 그 자체로 아무런 의미를 갖지 못한다. 만약 그 자체로서 의미를 갖는다면 이는 어문저작물의 단순한 변형에 불과하고, 저작권법에 통제를 받아야 하는 저작물일 것이다. 실질적으로 코퍼스는 반드시 어떤 목적하에 분석하는 프로그램 즉, 소프트웨어와 함께 존재하여야 그 존재의 의의가 있다. 대부분의 코퍼스는 문장분석용 도구(프로그램)와 함께 배포된다.[15]

---

14) 저작권심의조정위원회. 저작권 상담 · 조정 사례집, 동위원회. 1992. p.131

15) 한국과학기술원 코퍼스의 경우는 한국어 형태소해석기(PC용과 UNIX용)와 한국어 구문분석기, 자동태거 등과 같은 프로그램이 부착되어 있다. BNC코퍼스의 경우는 CLAWS라는 문장분석용 프로그램과 함께 배포하고 있다.

코퍼스의 활용적 측면을 위해서는 프로그램이 반드시 필요하나, 이는 원저작자와는 아무런 연관성이 없기 때문에 2차 저작물로서 코퍼스 보호측면으로는 '컴퓨터프로그램 보호법'이 적용될 수 있으나, 코퍼스를 구축하는 행위를 동법으로도 저작권 위배조항에 저촉되는지를 정의할 수는 없을 것이다.

4.3.3 멀티미디어와 코퍼스

멀티미디어에 대한 용어는 많이 사용되고 있지만, 그에 대해 명확하게 규정된 개념은 없다. 일반적으로 문자정보와 음성정보, 영상정보와 같이 여러 형태의 정보가 결합되어 복합적인 정보 또는 소프트웨어를 지칭한다. 가전회사와 같은 하드웨어 생산업자는 컴퓨터와 전화, 텔레비전, 오디오 등이 결합되어 사용자에게 다양한 정보를 생산하는 기기를 멀티미디어로 지칭한다. 여기에서 중요한 사실은 멀티미디어하드웨어는 멀티미디어용 소프트웨어를 필요로 하고, 그 소프트웨어는 0과 1의 Digital Information형태로 되어 있어야 한다는 점이다.[16] 한편, 우리히는 멀티미디어를 전통적인 텍스트데이터와 영상, 음성데이터 등이 하나로 묶여진 저작물 결합체이며, 정보기술과 통신기술의 결합체이며, 이용자와 시스템 간의 커뮤니케이션이 가능한 새로운 저작물이라고 보고 있다.[17] 즉, 다양한 형태의 정보를 동시에 수록된(multi) 매체를 멀티미디어로 정의할 수 있다.

일반적으로 우리에게 가장 알려진 멀티미디어저작물로는 CD-ROM이 있다. CD-ROM은 일반인들이 인지하고 있는 것처럼 Compact Disk Read Only Memory의 약자로서 종이나 디스크에 해당하는 단순한 신매체(New Media)일 뿐이다. CD-ROM에 하나의 어문데이터 혹은 음악데이터만이 저장이 되어 있을 경우, 이는 엄밀한 의미에서 멀티미디어저작물의 범주에 속하지 않게 된다. 그러나 시중에 유통되고 있는 대부분의 CD-ROM은 문자나 음성, 화상, 동화상으로 구성되어 전통적인 의미의 어문저작물과 음악저작물, 미술저작물, 영상저작물 등에 해당할 뿐 아니라 동시에 해당 저작물을 열람할 수 있는 프로그램도 내장하고 있다.[18] 즉, 전통적인 저작물 외에 프로그램과 같은 저작물도 동시에(multi) 갖고 있는 특징을 지니고 있다.

이러한 종합적인 저작물적 성격 때문에 멀티미디어 자료에 대해 국내 저작권 관련법[19]에서는 특별히 멀티미디어라고 명시하지 않고 단, 시디롬이라는 대표적인 미디어를 프로그램을 수록할 수 있는 매체적인 면과 저작물

16) 정성조. "멀티미디어 관련법 제도의 문제점과 개선방안", 한국저작권논문선집(Ⅱ), 저작권심의조정위원회. 1995. pp.49-50

17) Ulrich Loewenheim, "Multimedia and the European Copyright Law", IIC., Vol. 27, No..1. 1996, p.42

18) 정성조. "멀티미디어 관련법 제도의 문제점과 개선방안", 한국저작권논문선집(Ⅱ), 저작권심의조정위원회. 1995. pp.49-52

19) 컴퓨터프로그램보호법. 시행령. 제18조. 1996.6.7

로 간주한 규정을 두고 있다.

한편, 채명기는 멀티미디어제작물은 종합저작물성을 띠고 있으며 권리면에서 멀티미디어 제작물에 수록된 기초저작물을 제외하고, 멀티미디어 제작물 자체는 여러 저작자들의 기여로 이루어진 결과물로서 분리이용할 수 없는 공동저작물성을 지니고 있다고 주장하고 있다.[20]

일본의 경우, 멀티미디어 자체가 현행법으로는 처리가 불가능하다고 판단하며, 지적 소유권문제를 저작권의 현행법으로도 처리할 수 없다는 관점이 지배적이다.[21]

그러므로 코퍼스를 법적인 관점으로 해석하는 것에는 많은 무리가 있다. 왜냐하면, 멀티미디어에 대한 법적인 해석과 관점이 국내외적으로 명확하게 규정된 것이 없다는 것과 기술적으로도 코퍼스는 저장매체의 형태와 상관없이 해당 정보들을 0과 1이라는 코드만으로 기억하고 있으며, 원저작물을 그대로 저장하고 있는 점과 음성정보와 같은 비책자형 정보를 수록할 수 있다는 점을 제외하고는 멀티미디어 저작물과는 아무런 연관성이 없기 때문이다.

4.4 저작권법과 코퍼스 구축

4.4.1 저작자의 권리

저작권 보호를 위한 18세기 초의 영국 서적상 조합이 제출한 청원서를 요약하면 "많은 시간과 비용은 들여 저술을 하고 있으나,… 몇몇 업자들이 이를 똑같이 인쇄하여 원저작물이 판매되지 않아 저작의욕을 떨어지게 하고,… 재산권의 소유자에게 막대한 손해를 주고 있다…."[22] 고 주장하고 있다. 국내 저작권법에서도 복제권(제16조)과 2차적 저작물 등의 작성권(제21조)에서 저작물을 하나의 재산으로 간주하고 있다. 즉, 재산상의 피해에 대한 구제 및 보호의 의미에서 저작권법이 제정되었으며 이 정신이 지금까지 계속되고 있다. 재산상의 피해라 함은 저작물의 내용이 불법유통되어 원저작자의 수입(유무형)을 떨어뜨리는 것을 의미한다.

한편, 저작인격권 측면에서 저작권자는 해당 저작물의 내용, 형식 및 제호의 동일성을 유지할 권리를 가진다.[23] 즉, 저작권에는 동일성 유지권도 있기 때문에 저작물을 디지털화하는 것도 저작권법에 저촉될 수 있다. 이때 동일성 유지권을 원저작자에게 인정하는 것도 복제자가 해당 저작물을 복제함으로써 유무형의 이득을 추구하는데 대한 방어적 측면

---

20) 채명기, "멀티미디어시대에 있어서 저작물의 이용", 멀티미디어 시대의 저작권 대책-제2주제-, 저작권심의 조정위원회. 196. pp.32-34

21) 문화체육부. 멀티미디어 시대의 저작권(1). 문화체육부, 1993.11. p.16

22) 한승헌. 정보화시대의 저작권. 나남. 서울. 1992. pp.22-23

23) 저작권법. 제13조 1항

이라 할 수 있다.

이러한 관점은 저작물의 초기 출간 형태가 책자 형태일지라도 내용을 디지털화로 변경시킬 경우도 저작권자에게 허락을 획득해야 함을 의미한다.

영·미계의 저작권은 저작자에게 주어진 그의 저작물에 대한 배타적 지배를 내용으로 하는 경제적 권리(economic rights)로서만 파악되며, 저작자로부터 그 권리를 양도받거나 그에게 대가(royalty)를 지급하고 허락을 받지 아니한 자는 저작물의 이용이 금지된다는 의미를 갖는다.

이에 반하여 대륙법계에서의 저작자의 권리의 개념은 인간의 자유와 인격의 절대를 내세운 프랑스 혁명의 산물이라 할 수 있다. 특히, 프랑스에서는 저작자의 권리는 국왕의 특허가 아니라 저작물이 그의 지적 창작의 소산이라고 하는 자연적 사실을 바탕으로 승인되어야 한다는 인식에 생겨났다. 이 법령들은 저작자의 권리를 저작자의 손에 확고히 맡겨진 개인적인 권리로서 파악하고 있다.[24]

우리 나라 저작권법은 대륙법 특히, 프랑스식 견해가 충분히 반영되어, 저작물에 대한 절대적 권리를 저작자가 갖는 취지로 제정되었다.

4.4.2 코퍼스 구축의 법적 해석

국내외 저작권법은 디지털 자료로의 변환도 복제를 전제로 하는 행위로 간주하고 동일성 유지권과 같은 권리로서 이를 보호하고 있다. 이러한 보호는 디지털 데이터가 갖고 있는 여러 특성 가운데 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 대상으로 하여 구축되는 것이 아니며, 제한된 특정 주제분야의 학자들을 대상으로 구축되기 때문이다.

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.[25] 한편, 세계저작권협약에서도 복제권과 공연권, 방송권에 대하여도 각

---

24) 곽경직. 저작권의 제한에 관한 연구. 서울대학교 대학원 석사학위논문. 1990

25) 이상정, "디지털 시대에 있어서 저작(재산)권의 제한", 멀티미디어시대의 저작권 대책. 저작권심의조정위원회. 1996.12., pp.187-188

최근 마련된 '문학·예술 저작물의 보호에 관한 특정한 문제에 관한 초안' 제12조는 "①회원국은 국내 입법으로 저작물의 통상적인 이용과 상충되거나 저작자의 정당한 이익을 부당하게 해치지 않는 일정한 특별한 경우에 이 조약에 의해 문학·예술 저작물의 저작자에게 부여된 권리에 대하여 제한과 예외를 규정할 수 있다. ② 베른 협약의 적용에 있어 회원국은 그에 규정된 권리의 제한과 예외를 저작물의 통상 이용과 상충되거나 저작자의 정당한 이익을 부당하게 해치지 않는 일정한 특별한 경우로 한정하여야 한다."고 규정한다.

권리에 합리적인 정도로 효과적인 보호를 준다면 예외를 인정하고 있다.(제4조 2항)

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. 단, 코퍼스는 앞의 〈그림 1〉에서 제시한 바와 같이 원자료의 내용을 그대로 보여주고 있기 때문에 제3자가 코퍼스를 불법적으로 활용할 수도 있다.

이러한 것 때문에 코퍼스의 구축을 자유 사용의 측면으로 간주하는 것에 문제점이 제기될 수 있다. 즉, 현행법과 관례에 따르면 자유사용의 조항은 분명한 사례를 제시하여 문제를 해결하기보다는 앞으로 야기될 사례에 대해 사회적 통념과 법적 해석 등에 근거하여 향후 발생되는 문제점을 해결하려는 유연하고도 유보적인 법적 조항이라 할 수 있다. 코퍼스에 대한 분명한 이해와 사례를 제시하므로서 코퍼스 구축과 활용은 자유 사용으로 조정될 수 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 범위 내일 것과 공정한 관행에 합치될 것을 제시하고 있으나 구체적 의미는 결국 해석과 판례에 맡겨져 있다.[26] 즉, 인용은 공정이용에 포함시켜 저작권에 저촉되지 않는 조항으로 처리하고 있다. 일반적으로 코퍼스는 완전한 하나의 저작물을 디지털하기보다는 저작물의 일부만을 디지털화하여 언어데이터로 활용한다. 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 인용이라기보다는 표절에 해당한다고 할 수 있다.

코퍼스가 인용에 포함되며, 이를 저작권 침해로 간주하지 않을 수 있는 것은 다음과 같은 이유에 해당한다.

첫째, 코퍼스는 동일한 분야의 이용자를 대상으로 삼지 않기 때문에 경쟁관계가 성립되지 않는다. 둘째, 코퍼스는 근거 제시를 위해 반드시 출처를 명시한다. 이는 코퍼스가 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. 한편, 원자료 가운데 문자형태의 데이터를 제외하고는 어떠한 내용도 필요로 하지 않는다. 코퍼스 구축에 있어 원자

26) 곽경직. 저작권의 제한에 관한 연구. 서울대학교 대학원. 서울대학교 대학원 석사학위 논문. 1990. pp.29-45

료의 형태(페이지나 글자 크기, 삽도 여부, 기호 등)를 그대로 인용하기가 현실적으로 불가능하다.

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.

④ 국내외의 해결 방안

외국에서 코퍼스에 대한 저작권적인 해석은 크게 두 가지로 대별될 수 있다. 첫째는 기술적 해결이며, 둘째는 계약을 통한 해결이다.

i) 기술적 해결방안 : 기술적 해결방안도 크게 두 가지로 구분된다.

- 전산처리방안 : 구축된 코퍼스를 문장이나 문단단위로 절단하여 이를 소팅(sorting)한다. 소팅된 결과로는 어느 자료에서 추출된 내용인지를 파악할 수 없기 때문에, 기술적으로 저작권법을 벗어나는 방법이다. 일본의 전자사전협회(EDR)의 코퍼스 일부가 이 방법으로 구축되었다.

- 공유상태자료 활용 : 다음과 같이 공유상태에 놓인 자료를 활용한다.

첫째, 저작권법상 보호받지 못하는 저작물로서 분류되어 공유상태에 놓인 자료를 활용한다. 현행 저작권법 제7조는 법령, 고시, 공고, 판결, 또는 시사보도 등을 보호받지 못하는 저작물로 간주하고 있다. 이러한 데이터는 한국과학기술원코퍼스의 일부가 이러한 자료를 코퍼스로 구축하고 있다.

둘째, 저작물의 보호기간이 만료되어 공유상태에 놓인 자료를 활용한다. 저작권법 제36조 1항에서 명시된 바와 같이 저작물의 보호기간(사후 50년)이 지난 자료를 코퍼스로 활용하고 있다. 국립국어연구원의 코퍼스와 미국의 펜실베이니아 대학의 코퍼스도 이러한 자료를 코퍼스의 일부로 구축하고 있다.

셋째, 저작재산권이 포기되어 공유상태에 놓인 자료를 활용한 경우이다.

ii) 계약을 통한 해결방안

코퍼스 구축에 가장 합리적이고 손쉬운 방법으로서 계약을 통한 구축방안이다. 이에 해당하는 코퍼스로는 한국과기원코퍼스를 비롯하여 국내외 대부분의 코퍼스는 이러한 자료를 활용한다. 즉, 저작재산권을 포기하였다기보다는 출판사나 저자로부터 직접 코퍼스로서 구축을 허락받은 자료를 대상으로 코퍼스를 구축하는 방안이다. 고려대와 연세대 코퍼스의 일부는 국내 출판사와의 계약을 통해 코퍼스를 구축하였으나, 해당 기관에서의 이용을 허락한 것이기 때문에 실질적으로 계약에 의해 구축한 것이라고는 볼 수 없다.

결 론

어떠한 저작물이 저작권법에 의해 보호를 받을 수 있는가와 어떠한 일련의 행동이 저작권법에 저촉되는가를 판단하기 위해서는 구체적인 사례와 법적인 해석이 있기 전까지는 그 선을 규정하기가 쉽지 않다.

분명한 것은 저작권법에서 궁극적으로 보호하려는 것은 저작권자의 지적 수고이다. 이러한 정신은 지적 수고에 대한 보상을 통해 보다 지속적으로 지식의 창출을 유도하기 위해서이다.

디지털 자료 가운데 코퍼스는 저작자의 지적수고를 필요로 하지 않고 실험데이터로서 해당 자료를 필요로 하는 매우 특별한 경우이다.

코퍼스가 어문저작물을 단순히 ASCII와 같은 코드로 변환하여 제3자에게 제공하는 것이 목적이라면, 이는 동일성 유지권의 관점하에서 어문저작물의 저작자에게도 반드시 일정부분의 저작인격권이 부여되어야 한다.

왜냐하면, 디지털 형태의 자료는 기존의 책자형태의 자료보다 복제의 용이성과 전파성이 월등하기 때문에 어문저작물의 디지털화의 경우, 저작권법적인 측면에서 저작자에게 복제의 피해가 훨씬 심각하기 때문이다.

그러나 코퍼스와 같이 특정한 목적을 가지고 있으면서, 원저작물에 대해 재산상의 직접적인 피해를 입히지 않는 디지털 데이터의 경우에 다른 디지털 데이터와 동일하게 저작물 사용이 제약을 받아야 하는 것은 공공이익(도서관)과 학술활용(어문분야)에 커다란 걸림돌이 될 것이기 때문에 많은 법리적 해석과 학술적 견해가 필요하다.

그러므로 디지털 데이터에 대한 일괄적인 법적용보다는 저작물에 대한 재산상의 손해 유무로서 법의 저촉 여부를 판단하는 것이 필요하다. 왜냐하면, 앞으로의 사회는 대부분의 자료가 디지털화한 자료로 유통될 것이고, 해당 디지털 데이터가 어떠한 목적으로 사용되는지를 예측하여 저작권법도 모든 경우를 예측하여 제정될 수 없기 때문이다.

특히, 코퍼스는 자유 이용이나 인용으로 간주할 수 있는 형태 및 구축에 따른 특성을 갖고 있기 때문에 저작권이 제한될 수 있다.

끝으로, 대부분의 코퍼스를 구축하고 있는 대학과 기관에서는 자신들이 구축한 코퍼스를 공개하지 못하고 있다. 왜냐하면, 저작권은 친고죄이기 때문에 당사자가 법에 호소하기 이전에는 이의 보호에 대해서는 별다른 문제가 없을 수 있지만, 역으로 전혀 죄의식 없이 순수 연구 목적용으로 사용하고 있는 저작물에 대해 어느 날 갑자기 고발을 당할 수도 있기 때문이다.

앞으로 이에 대한 분명한 법적인 해석과 견해는 학술적 연구 특히, 모든 자료의 디지털화를 목표로 하고 있는 도서관학 분야와 언어학 분야에서는 반드시 필요한 전제조건이 될 것이다.

참고문헌

곽경직. 저작권의 제한에 관한 연구, 서울대학교. 서울대학교 석사학위논문. 1990.

김영택 등. 자연언어처리. 교학사. 1995.

남영준. "다운로딩과 저작권법의 문제", 청랑 정필모박사화갑기념논문집. 동편찬위원회. 서울. 1990.

남영준. "디지털 자료와 저작권", 신미디어에 의한 저작권 보호. 한국문예학술저작권협회. 1996.11.

문화체육부. 멀티미디어시대의 저작권(1). 문화체육부. 1995.

서울대학교 법학연구소. 미국 멀티미디어 분야 판례 번역.평석 연구. 한국컴퓨터프로그램보호회. 서울. 1995.

저작권심의조정위원회. 멀티미디어 시대의 저작권대책. 동위원회. 서울. 1996. 12.

저작권심의조정위원회. 저작권상담 . 조정 사례집. 동위원회 . 서울. 1992.

저작권심의조정위원회. 저작권에 관한 외국 판례선집(3) -일본편-. 동위원회 . 서울. 1995.

저작권심의조정위원회. 저작물의 새로운 기술적 이용에 관한 국립위원회의 최종보고서. 동위원회. 서울. 1994.

저작권심의조정위원회. 한국저작권논문선집(Ⅱ). 동위원회 . 서울. 1995.

한국문예진흥원. 문화예술정보 서비스 체계와 관련한 저작권 문제연구. 동 진흥원. 최종보고서.1996.

한승헌. 정보화시대의 저작권. 나남. 서울. 1992.