사회과학 분야 도서의 목차 텍스트에 대한 통계적 특성에 관한 연구

A Study on the Statistical Characteristics for Table of Contents Text of the Books in Social Sciences Field

이용구 (Yong-Gu Lee)*

초 록

이 연구는 최근 접근 및 활용이 높아지고 있는 목차에 대해 품사 측면과 주제 측면에서 가지는 기술통계와 비교 분석을 수행하였다. 이를 위해 대학 도서관의 수서 목록에서 사회과학분야 도서를 추출하고 해당하는 도서에 대해 종합목록으로부터 DDC 분류기호를, 인터넷 서점으로부터 목차 정보를 추출하였다. 서명과 목차를 대상으로 형태소 분석하여 명사 중심의 어휘에 대해 기술통계와 빈도 분석을 실시하였다. 그 결과 형태소 측면에서 서명과 목차는 명사가 대략 절반가량 차지하며, 서명과 비교하여 목차는 50배 정도 더 많은 명사를 가지며, 목차에 출현한 명사 중에 목차만이 고유하게 가지는 비율이 95.2%에 달하는 것으로 파악되었다. 또한 목차는 사회과학 학문분야에 따라 길이가 차이가 나는 것으로 나타났다.

ABSTRACT

Recently, the table of contents (TOC) has been becoming increasingly accessible and utilized. The study conducted descriptive statistics and comparative analysis of the table of contents in terms of parts of speech and subject in text. For this purpose, this study chose the books of the social sciences field from acquisition lists of an academic library, obtained Dewey class numbers of target books from KERIS union catalog, and extracted TOC data from online bookstore. Morphological analysis was performed on each book titles and TOCs, and descriptive statistics and frequency analysis were carried out. As a result, nouns made up roughly half of the morphemes of titles or the TOCs. TOCs had about 50 times more nouns than titles. The percentage of unique nouns that appeared only in the table of contents is estimated to be 95.2% of the TOC's total nouns. The table of contents also showed a differences in its lengths depending on the field of social science.

키워드: 목차, 통계적 특성, 목차 길이, DDC (Dewey Decimal Classification)
table of contents, descriptive statistics, length of table of contents, DDC (Dewey Decimal Classification)

* 계명대학교 문헌정보학과 부교수(yonggulee@kmu.ac.kr)

■ 논문접수일자 : 2019년 6월 24일 ■ 최초심사일자 : 2019년 6월 25일 ■ 게재확정일자 : 2019년 6월 28일
■ 정보관리학회지, 36(2), 255-273, 2019. [http://dx.doi.org/10.3743/KOSIM.2019.36.2.255]

1. 서 론

문헌이나 텍스트를 대상으로 검색하거나 자동 분류하기 위해서는 그 대상으로부터 메타데이터를 추출해야 한다. 도서관에서는 분류나 편목 작업을 통해 도서에 대해 대표적인 서명, 저자명, 주제를 나타내는 분류기호 또는 색인어를 메타데이터로 생성한다. 이미지나 멀티미디어에서는 이미지의 색상이나 질감, 소리의 속도나 높낮이 등을 내용 표현을 위해 메타데이터로 추출한다. 이러한 메타데이터는 그 대상물을 가장 잘 표현해야 하기에 무엇을 어떻게 추출하느냐가 정보 검색 시스템과 같은 응용 프로그램(application)에서 매우 중요하다.

전문(fulltext) 형태로 된 문헌이나 텍스트는 상황에 따라 많은 양의 메타데이터를 추출할 수 있다. 응용 프로그램은 목적을 달성하기 위해 이렇게 많은 양의 메타데이터를 다양하게 활용한다. 구체적으로 그 응용 프로그램이 정보 검색 시스템이면 좋은 검색 결과를 가져오도록 할 수 있으며, 기계 학습 분류기라면 더 좋은 자동 분류 결과를 가져올 수 있다. 예를 들어 학술지 논문 검색 시스템은 표제, 저자 키워드, 초록 등을 대상으로 메타데이터를 추출한다. 심지어 논문의 본문에서도 주요 키워드를 추출할 수 있다. 이러한 풍부한 메타데이터는 응용 프로그램의 다양한 활용을 가져올 수 있을 뿐만 아니라 해당 응용 프로그램의 성능을 높이는 역할도 할 수 있다.

전통적으로 도서관은 도서를 편목하고 더 나아가 탐색하기 위해 주로 표제 면(title page)에서 필요한 정보를 추출하여 사용한다. 즉 사서는 목록과 색인 과정을 통해 메타데이터를 생성하며, 이러한 정보는 도서에 대한 통합 도서관 시스템 또는 온라인 열람 목록(OPAC)의 검색 시스템의 서지레코드로 추가된다. 목록 기술 수준에서 최소 또는 표준 수준으로 작성된 도서 서지 레코드를 메타데이터 측면에서 살펴보면, 텍스트 형태로 추출 가능한 정보는 대부분 짧은 길이의 표제(서명)로 한정된다. 정보 검색이나 자동 분류와 같은 응용 프로그램에서 이러한 상황의 메타데이터는 다양한 활용이나 성능에 제약을 준다.

최근에는 예전에 비해 도서의 목차 정보를 쉽게 획득하거나 접근할 수 있다. 구체적으로 인터넷 서점들은 도서의 내용을 알리기 위해 목차 정보를 제공하고 있으며(구중억, 이응봉, 2009b, p. 311), 많은 도서관들도 자관의 OPAC 서비스에서 이용자의 정보 선택을 돕기 위해 도서의 목차 정보를 제공하고 있다(정혜미, 정재영, 2008; Byrum & Williamson, 2006). 이러한 현상은 지속적으로 많아지고 있는 것으로 보인다. 특히 도서의 목차는 도서관 장서 이용에도 영향을 미쳐, 목차 정보를 제공하는 도서가 그렇지 않은 도서보다 더 높은 대출율을 보인다는 연구 결과도 제시되었다(Morris, 2001; Chercourt & Marshall, 2013; Tosaka & Weng, 2011).

현재 국내 도서관 홈페이지에서 질의어를 사용하여 특정 서지 레코드를 검색하였을 때, 상세 보기에서 목차 정보를 종종 볼 수 있다. 대개 이는 목차 컨텐츠에 대한 단순한 링크를 제공하는 수준에 그친다. 그래도 도서관 시스템에서 목차에 대한 이러한 접근이 가능할 정도로 최근에는 도서의 목차 데이터가 보급되기 시작하였으므로, 앞으로 목차를 활용한 다양한 시도가 이루어질 필요가 있다.

이러한 배경에서 이 연구는 국내 대학 도서관이 수집하고 분류한 사회과학 분야 도서의 서명과 목차를 추출하고 이를 이용하여 텍스트로서 목차 정보의 특성을 분석하고자 하였다. 보다 구체적으로 서명, 목차, 일반 텍스트 간의 형태소 분석을 통한 단어들의 통계적 특성 및 비교 분석을 수행하였으며, DDC(Dewey Decimal Classification) 범주에 따른 목차의 차이 분석을 적용하였다. 이를 통해 목차가 가지는 정보이론적 특성을 파악하여 향후 목차 정보의 활용에 필요한 기초 정보를 제공하고자 한다.

2. 선행 연구

목차(table of contents: TOC)의 정의를 살펴보면 "도서나 각종 출판물의 내용을 나타내는 장, 절이나 계속자료의 각 기사를 차례로 적은 요약, 페이지 순서대로 정리된다"라고 하였다(문헌정보학용어사전, 2010). 이 정의 내용은 크게 두 가지 종류의 목차로 나누어짐을 알 수 있다. 하나는 일반적으로 전체 책의 내용을 지적인 측면에서 계층적으로 구분하여 나타나는 주제 기반 목차(subject-based TOC)이며, 다른 하나는 계속자료의 기사처럼 표제와 저자 중심으로 목차를 제시한 저자/표제 기반 목차(author/title-based TOC)이다. 단행본 도서의 경우 대부분 주제 기반 목차 형식을 갖지만 전집이나 논문집, 총서의 목차는 저자/표제 기반 목차의 형식을 갖는다.

목차는 문헌의 내용을 구조적으로 요약한다고 볼 수 있는데, 목차에 나타난 단어는 그 요약을 잘 드러내는, 즉 주제적 측면을 나타내는 용어로 볼 수 있다. 도서관 측면에서 이러한 목차가 가지는 주요 이점을 살펴보면 다음과 같다(Pappas & Herendeen, 2000; Chercourt & Marshall, 2013).

첫째, 목차는 이용자가 자신이 원하는 도서를 찾을 때, 그 도서가 적합한지 판단하거나 식별할 수 있게 도와준다. 문헌의 주제가 많이 함축된 짧은 표제보다, 세부적인 내용이 기술되어 있는 목차가 이용자에게 적합여부의 판정에 필요한 정보를 더 많이 제공할 수 있다.

둘째, 도서관의 검색 시스템은 목차에 출현한 용어를 추가함으로써 적합한 자료의 식별과 검색을 가능하게 하여 검색 효과성을 향상시킨다. 최근 대부분의 이용자는 키워드 탐색을 이용하므로 내용을 드러내는 풍부한 목차의 단어가 문헌에 대한 접근성을 높여 적합한 자료를 검색할 확률을 높인다.

셋째, 목차는 주제 편목의 한계를 보완한다. 주제명표목을 부여하는 규칙을 보면, 일반적으로 주제 편목(subject cataloging)은 자료의 전체 내용을 잘 요약하는 소수의 주제명 또는 주제명표목을 선정하도록 권장한다. 이렇게 소수의 주제명표목을 부여하는 규칙을 따르면, 자료가 담고 있는 다른 많은 주제어를 잃게 된다. 이러한 한계를 내용이 풍부한 목차를 추가하여 보완할 수 있다.

기타 이점으로 목록에서 부여되지 않은 추가적인 저자 및 저자 정보를 찾을 수 있거나, 자료의 전체 내용을 대표하지 않는 특화된 내용이나 주제를 포함하는 소수의 장(chapter)을 발견할 수 있다.

목차가 내용이나 주제를 나타내기 위해 어떠한 특성을 가지는지 파악하기 위해 통계적 측면

에서 목차를 살펴볼 필요가 있다. Winke(1998)는 미국 의회도서관의 목록으로부터 648권의 도서를 추출하여 이들 중 92.75%가 목차를 가지고 있으며, 평균적으로 67.75개의 단어를 포함하고, 대부분의 목차가 한 단계 또는 두 단계의 계층 구조를 가진다고 하였다. 도현호와 이용구(2014)는 사회과학분야 단행본 581권을 대상으로 서명, 목차, 책 소개 등의 텍스트를 추출하여 자질을 분석하였다. 서명과 목차에 나타난 단어의 출현빈도가 각각 3,061개(평균 5.3개)와 109,523개(평균 188.5개)로 나타났다. 다만 대상 도서가 사회과학 분야라서 목차가 상대적으로 상세하고 계층 구조의 수준이 깊어 많은 자질을 포함하고 있는 것으로 생각된다.

앞서 살펴보았듯이, 목차에 출현한 풍부한 단어는 이용자의 검색과 발견(discovery)에 중요한 원천이 된다. Van Orden(1990)은 디지털 시스템에서 목차나 초록 같이 내용이 풍부한 구성요소와 원문 정보는 향상된 탐색 방법을 가져오거나, 보다 나은 컴퓨터 인터페이스를 보이고, 원본에 나타난 지식 구조와 사용을 이해하기 좋게 한다고 하였다. 또한 검색 측면에서 보다 많은 단어를 지니는 구성요소가 짧은 표제보다 검색될 경향이 높긴 하나, 낮은 정확률을 동반한다고 주장하였다. Dillon과 Wenzel(1990)은 개선된 서지 레코드가 검색 효과성에 미치는 연구를 하였는데, 그 결과 목차가 추가된 서지 레코드가 그렇지 않은 레코드보다 10%의 재현율을 끌어올렸으나 정확률은 낮아져, 최종 검색성능은 약간 향상된 것으로 나타났다.

OPAC의 주제 탐색에서 목차와 주제명표목의 효과성을 조사한 Choi, Hsieh-Yee, Kules(2007)는 이용자가 입력한 질의어가 LCSH보다 목차에 더 성공적인 매칭을 가져왔지만, 그들의 질의는 종종 적합 자료를 찾는데 실패하여 이용자의 키워드를 목차나 주제명표목과 연결 짓는 방법이 필요하다고 강조하였다. 구중억과 이응봉 연구(2009a)에서 이용자는 검색화면의 유용성 측면에서 도서 소개, 목차, 서평, 본문 등 상세 정보를 제공하는 시스템을 그렇지 않은 시스템보다 높이 평가하였는데, 이는 이용자가 도서의 발견, 식별, 입수에서 목차나 본문 같이 풍부한 정보를 제공하는 검색 인터페이스에 더 만족한다는 것을 의미한다.

목차나 요약 정보는 이용자에게 단순히 표제나 저자 정보를 넘어 그들의 정보탐색 과정에서 자원 발견의 기회를 높이는 도구로서 역할을 한다. 도서관이 목차나 요약 정보를 그들의 서지 레코드에 추가하고 이용자가 이를 검색과정에서 브라우징 한다면, 해당 레코드는 발견의 기회와 확률이 높아져 도서의 대출을 높이는 결과를 가져올 수 있다. Morris(2001)는 온라인 목차가 도서의 이용을 증가시키는지 조사하기 위해 3,957권의 도서를 목차가 있는 그룹과 그렇지 않은 그룹으로 나누고 14개월 동안 온라인 목록의 통계를 추적하였다. 그 결과 목차가 있는 도서는 이용될 가능성이 45% 더 높아진다고 하였다. Tosaka와 Weng(2011)은 대출된 자료의 OPAC 탐색 트랜잭션을 조사하여, 4개 학문 영역에서 목차나 요약 같이 내용이 풍부한 레코드가 전반적으로 높은 대출율과 관련이 있으며, 대출과 직접적으로 관련된 검색 방식은 키워드 탐색임을 발견하였다. 특히 이들은 MARC의 505 필드가 대출율에 가장 큰 기여를 한다고 제시하였다.

앞서 연구와 반대로 Chercourt와 Marshall

(2013)은 기존의 낮은 대출을 보이는 자료를 대상으로 소급하여 서지레코드를 내용이 풍부하게 만든 후 대출 통계를 조사하였다. 그 결과 25% 이상의 자료가 최소한 한번 이상 대출되었는데 이들 중 대부분은 대출된 적이 없는 자료였다. 또한 조치된 그룹의 도서가 일반도서보다 대출에서 더 큰 증가를 보였다.

편목할 때 목차를 서지레코드에 추가하는 것이 모든 면에서 긍정적이지만은 않다. 도서관에서 이러한 작업은 표준 서지레코드 작성보다 추가적인 비용이나 인력이 투입되어야 한다. 미국의회도서관의 BEAT(Bibliographic Enrichment Advisory Team) 프로젝트는 목차를 서지레코드에 추가하는 다양한 방법과 그에 따른 비용 차이를 볼 수 있다(Byrum & Williamson, 2006). 이 프로젝트는 목차를 서지레코드에 추가하는 4가지 방법에 따른 비용을 제시하였다. 첫째 방법은 직접 수작업으로 타이핑하여 입력하는 가장 전통적인 방법이며, 둘째는 좀 더 발전된 방법으로 E-CIP(Electronic Catalog In Publication) 프로그램에 의한 방법, 셋째는 OCR 소프트웨어와 장비를 이용하는 방법, 그리고 마지막으로 출판사 제공 ONIX(ONline Information eXchange) 파일로부터 목차를 추출하는 방법이다. 수작업 입력 방법은 레코드 당 약 40달러의 비용이 소요되는데 비해 나머지는 각각 3달러, 2달러, 0.8달러로 추정하였다. 일부 제한적이기는 하지만 Dinkins와 Kirkland(2006)는 OCLC 프로그램을 통해 서지레코드 당 0.1392달러를 지불하였다. 국내의 경우는 이러한 비용에 대한 연구가 필요한 실정이다.

3. 도서 목차 수집 및 분석

3.1 데이터 수집 및 가공[1)]

이 연구는 목차 정보를 수집할 대상 도서를 실제 대학도서관의 장서에서 선정하였다. 주제별 차이분석을 위해 분석 대상 도서는 분류기호가 필요하다. 이들 도서의 분류기호로 한국학술정보원(KERIS) 종합목록의 대표 분류기호를 적용하였다. 이들 도서가 이미 대학도서관의 DDC 분류기호를 가지고 있지만, 이 기호의 객관성을 높이고자 종합목록의 대표 분류기호로 대체하였다. 또한 이들 도서들에 대한 메타데이터로 서명과 목차를 생성하였다. 목차는 인터넷 서점으로부터 Open API를 통해 획득하였다. 이후 서명과 목차에 나타난 텍스트에 대해 형태소 분석을 수행하였으며, 이러한 데이터를 기반으로 다수의 통계 분석과 차이 분석, 단어의 빈도분포 분석을 실시하였다.

우선 서명과 목차 텍스트를 얻기 위한 도서 리스트가 필요한데, 이 연구가 분석할 단행본 도서 리스트를 얻기 위해 실제 연구 중심 대학의 대학도서관 수서 목록을 사용하였다. 구체적으로 2014년과 2015년 2년 동안 한 대학도서관의 신착자료 리스트를 수집하였다. 이렇게 실제 수서 목록을 사용한 것은 도서관 장서의 측면에서 서명이나 목차를 분석하기 위함이며, 향후 도서관 시스템에 목차 관련 기술의 적용 가능성을 보다 높일 수 있도록 하기 위함이다. 실험 데이터가 실제 도서관 환경에서 비롯되면 그렇지 않은 데이터보다 도서관 상황에 더 적

1) 이 연구는 도현호(2016)의 데이터 중 일부를 이용하였으며, 이 절에서는 이해를 돕기 위해 해당 연구의 데이터 수집 및 가공 내용 중에 필요한 부분만 간략히 요약하였다.

합하게 되므로 실제 소장 도서목록을 활용하는 것이 바람직하다.

데이터 수집 대학도서관에서 2년 신착자료 리스트에 포함된 도서 총 수는 45,705종이며, 이중 국내서는 약 71%에 해당하며 사회과학 분야(DDC 300대) 도서가 가장 높은 비율을 차지했다. 인터넷 서점을 대상으로 목차 정보를 수집해야 하므로 분석 대상은 국내서로 한정하였다. 또한 주제 분야는 학문적으로 가장 잘 세분화되어 향후 다양한 활용이 가능할 것으로 예상되는 사회과학 분야를 선정하였다.

목차 정보를 수집하기 위해 인터넷 서점 '알라딘'의 Open API를 활용하였다. 신착자료 리스트에 존재하는 도서의 ISBN을 사용하여 알라딘 Open API를 통해 해당 도서를 검색하였다. 이를 통해 24,458권이 검색되었으며, 향후 폭넓은 활용을 위해 다양한 메타데이터(주로 서명, 목차, 책소개, 출판사 서평 등)가 모두 존재하는 도서만을 실험 데이터로 한정하였다. 그 결과 총 17,355건이 확인되었으며, 이 중에서 사회과학분야 도서는 4,111권으로 선정되었다.

특정 도서관에서는 수서된 도서에 대해 자관의 분류 규칙에 따라 DDC 분류기호를 부여한다. 도서관의 규모나 관종에 따라 도서 분류에 사용되는 정책이나 규칙은 상이하다. 따라서 특정 도서관의 분류기호를 사용하여 도서의 주제를 분석하면 그 도서관에는 적합할지 모르나 다른 도서관은 상황과 환경이 달라 최적합 하지 않을 수 있다. 이러한 이유로 이 연구에서는 도서의 주제를 나타내는 분류기호를 종합목록의 대표 분류기호로 적용하고자 하였다. KERIS는 학술연구정보서비스(Research Information Sharing Service, RISS)를 통해 전국 대학도서관의 장서에 대한 종합목록을 구축하고 제공한다. 이 서비스는 특정 도서의 대표 분류기호를 제시하는데, 이 연구는 이 값을 해당 도서의 대표 분류기호로 채택하였다.

대학도서관에서 입수한 신착자료 리스트 중에서 사회과학 분야(DDC 300대)이고 인터넷 서점으로부터 목차를 포함한 풍부한 메타데이터를 추출한 4,111권에 대해, KERIS 종합목록의 분류기호를 비교하면 〈표 1〉과 같다. 표를 보면 대학도서관의 신착자료 리스트에서 분류된 DDC 300대 도서(4,111권) 중 3,213권만이 KERIS 종합목록의 분류기호 300대에 해당하며, 나머지 200권은 다른 주류로 분류된 것으로 나타난다. 다만, 4,111권 중에서 697권은 결측치(missing value)에 해당하는데, 이는 신착자료 리스트에 수록된 도서의 ISBN 값에 대해 KERIS로부터 대표 분류기호를 얻지 못해 결측치가 된 것을 의미한다. KERIS가 회신한 데이터에 특정 ISBN에 해당하는 분류기호가 없었으며, 이러한 도서는 대표 분류기호가 없기에 향후 실험에서 제외하였다.

대학도서관 신착자료 리스트의 300대 주류 분류 도서 중에서 KERIS 종합목록도 300대 주류로 분류한 도서 3,213권에 대해 세부적인 강목별 빈도를 비교하는 분할표을 구해보면, 〈표 2〉와 같다. 전체에서 두 분류기호가 동일한 강목으로 된 항목(표의 대각선 값의 합)은 3,038권이며, 나머지 불일치하는 분류는 176권에 해당하였다.

특히 가장 많은 오차를 가져오는 부분은 30X대 강목으로, 대학도서관의 분류기호 30X 강목에 해당하는 50권이 KERIS의 종합목록의 분류기호에서는 300대의 다른 강목으로 분류된

〈표 1〉 대학도서관의 DDC 300대 도서별 종합목록 분류기호 분할표

KERIS	30X	31X	32X	33X	34X	35X	36X	37X	38X	39X	합계
0XX	5		3	4				1	1		14
1XX	9		4	1	2	1	1	2			20
2XX	2						1	1		1	5
3XX	684		391	809	515	159	182	303	77	93	3,213
4XX	4							2			6
5XX	2	1		4			1			1	9
6XX	10			22		7	5	6	3	4	57
7XX	6		3	2				5	3	1	20
8XX	6	1	1	3	1		2	1		1	16
9XX	12		13	3	1	3	1	3	1	17	54
결측치	128		52	141	161	45	59	73	23	15	697
합계	868	2	467	989	680	215	252	397	108	133	4,111

〈표 2〉 대학도서관의 DDC 300대 도서별 종합목록 300대 강목 분류기호 분할표

KERIS	30X	31X	32X	33X	34X	35X	36X	37X	38X	39X	합계
30X	634		10	9	7		5	2	1	2	670
31X											0
32X	7		368	10	1	5		2			393
33X	15		3	775	11		3	3	1		811
34X	4		4	7	482	3	2				502
35X	2		5	2	2	151	4				166
36X	14		1	3	9		166	1			194
37X	4			1			2	295			302
38X	1			2	3				75		81
39X	3									91	94
합계	684	0	391	809	515	159	182	303	77	93	3,213

것을 볼 수 있다. 이는 총류 성격의 강목으로, 여러 주제를 포괄하거나 특정 주제로 분류하기 어려운 분야에 해당한다. 주류 수준에서는 총류(000대)가 해당하며, 사회과학 주류에서는 사회과학 일반(30X대) 강목이 해당한다. 대개 사전류(辭典이나 事典)나 백과사전, 전집이나 총서, 잡저 등이 여기에 분류된다. 〈표 1〉과 〈표 2〉를 보면 사회과학 일반(30X대) 강목과 같은 총류 성격 분야에서 많은 오차를 보이고 있다. 이러한 분야는 수작업에 의한 도서 분류조차 일관되지 않음 또는 정확한 분류의 어려움을 보여준다.

다만 31X대(일반통계/Collections of general statistics)의 두 권 중 KERIS의 종합목록에서는 한 권은 문학 분야로 분류되었으며, 다른 한 권은 일반 통계학에 관련된 내용으로 519 세목으로 분류 되었다.

분석에 사용된 도서를 300대 강목별로 보면 33X(경제학/Economics) 분야가 가장 많은 811권이며, 그 다음으로 30X(사회과학 일반/The Social sciences) 분야 670권, 34X(법률학/Law) 분야 502권 순으로 나타났다. 31X 분야의 경우, 분석 대상에 한 권도 포함되어 않아 분석의 한계를 가진다.

3.2 형태소 분석 및 분포 특성

목차를 분석하여 그 특성을 파악하기 위해 다른 텍스트와 비교해 볼 필요가 있다. 비교를 위한 다른 텍스트는 도서의 서명과 자연언어처리를 위한 일반 말뭉치를 대상으로 하였다. 서명은 서지 레코드에서 가장 흔하게 볼 수 있는 텍스트 중 하나이다. 도서에는 대부분 서명과 목차가 존재하기에 그 차이를 살펴볼 필요가 있다. 일반 말뭉치는 우리가 사용하는 한국어의 가장 기본이 되는 특성을 가지고 있어 이 또한 목차와 비교해 볼 필요가 있다. 이 연구에서는 서명은 목차를 추출한 도서의 것으로 하였으며, 일반 말뭉치는 울산대학교 한국어처리연구실(http://nlplab.ulsan.ac.kr/)에서 공개한 학습용 말뭉치를 사용하였다. 후자의 경우 형태소 분석기를 좋은 성능을 가져오도록 만드는데 사용되는 데이터이기에 일반적인 한국어를 잘 대표한다고 할 수 있다. 이 말뭉치는 각각의 어절에 대해 이미 품사가 태깅된 상태로 제공된다.

한국어는 어근이나 어간에 접사, 어미, 조사가 붙어 이를 분리시키는 형태소 분석을 통해 어절을 최소의 의미 단위인 형태소로 나누어야 한다. 이 연구는 3,213권의 서명과 목차를 울산대학교 한국어처리연구실의 UTagger를 사용하여 형태소 분석하였으며 앞서 제시한 말뭉치를 포함하여 형태소와 그에 따른 품사 태그를 추출하였다. 이를 〈표 3〉과 같이 정리하였다.

서명, 목차, 말뭉치 텍스트로부터 추출된 형태소는 각각 42,360개, 2,126,258개, 32,213,813개이다. 이들 텍스트에서 가장 높은 비율을 차지하는 품사는 명사로 서명은 49.3%, 목차는 47.2%, 말뭉치는 25.6%를 차지했다. 일반 텍스트인 말뭉치보다 서명과 목차는 절반 정도에 해당하는 다수의 명사로 이루어져 있음을 알 수 있다. 심지어 서명은 목차 보다 명사의 비율이 약 2% 정도 조금 더 높다. 다음으로 높은 비율의 품사는 동사인데 서명은 7.0%, 목차는 6.5%, 말뭉치는 11.4%를 차지했다. 동사의 경우 명사와 반대로 일반 텍스트에서 그 비율이 가장 높게 나타났으며, 서명과 목차에서는 상대적으로 그 비율이 낮았다. 체언은 서명과 목차에서, 용언은 일반 텍스트에서 우세한 비율을 보였다. 서명과 목차는 체언과 용언이 거의 70%에 육박하는 것을 알 수 있다. 체언과 용언에 속한 품사의 형태소 비율도 서명이나 목차는 유사하여 이 두 텍스트가 비슷한 성향을 보이는 것을 알 수 있다.

관계언의 격조사는 체언과 용언의 경향과 다르게 말뭉치에서 가장 높은 비율을 보이고 그다음으로 목차, 서명 순의 경향을 보였다. 다만, '의'로 대표되는 관형격 조사는 말뭉치보다 목차와 서명에서 높은 비율을 보였는데, 관형격 조사 '의'가 주로 명사로 대표되는 체언과 다른 체언을 연결하는 역할을 하기 때문인 것으로 보인다. 이러한 조사는 제목이나 짧게 요약된 문장에 출현한 다수의 체언을 연결하기 위해 필요하며, 이러한 조사가 다수 존재하는 텍스트가 서명이나 목차임을 알 수 있다.

〈표 3〉 서명, 목차, 일반 텍스트의 형태소 분석에 따른 통계

5언	품사	태그명	태그	서명	목차	말뭉치
체언	명사	일반명사	NNG	20,877(49.3%)	1,003,962(47.2%)	8,255,186(25.6%)
		고유명사	NNP	2,180(5.1%)	68,124(3.2%)	865,166(2.7%)
		의존명사	NNB	855(2.0%)	48,671(2.3%)	1,132,910(3.5%)
	대명사	대명사	NP	476(1.1%)	43,672(2.1%)	435,703(1.4%)
	수사	수사	NR	96(0.2%)	3,977(0.2%)	118,953(0.4%)
용언	동사	동사	VV	2,956(7.0%)	138,720(6.5%)	3,657,549(11.4%)
	형용사	형용사	VA	917(2.2%)	43,032(2.0%)	1,128,051(3.5%)
	보조용언	보조용언	VX	330(0.8%)	20,891(1.0%)	664,363(2.1%)
	지정사	긍정지정사	VCP	267(0.6%)	21,316(1.0%)	545,567(1.7%)
		부정지정사	VCN	6(0.0%)	2,187(0.1%)	51,444(0.2%)
수식언	관형사	관형사	MM	288(0.7%)	16,002(0.8%)	472,325(1.5%)
	부사	일반부사	MAG	520(1.2%)	36,326(1.7%)	878,341(2.7%)
		접속부사	MAJ	70(0.2%)	2,630(0.1%)	75,090(0.2%)
독립언	감탄사	감탄사	IC	6(0.0%)	640(0.0%)	21,245(0.1%)
관계언	격조사	주격조사	JKS	476(1.1%)	27,034(1.3%)	968,007(3.0%)
		보격조사	JKC	61(0.1%)	4,244(0.2%)	75,161(0.2%)
		관형격조사	JKG	2,222(5.2%)	126,787(6.0%)	842,898(2.6%)
		목적격조사	JKO	1,115(2.6%)	62,503(2.9%)	1,388,971(4.3%)
		부사격조사	JKB	1,138(2.7%)	63,428(3.0%)	1,684,639(5.2%)
		호격조사	JKV	7(0.0%)	112(0.0%)	3,423(0.0%)
		인용격조사	JKQ	-(0.0%)	879(0.0%)	40,276(0.1%)
	보조사	보조사	JX	661(1.6%)	43,056(2.0%)	1,277,823(4.0%)
	접속조사	접속조사	JC	944(2.2%)	40,046(1.9%)	269,788(0.8%)
의존 형태	어미	선어말어미	EP	102(0.2%)	9,968(0.5%)	803,702(2.5%)
		종결어미	EF	166(0.4%)	21,771(1.0%)	1,337,009(4.2%)
		연결어미	EC	1,775(4.2%)	94,462(4.4%)	2,423,987(7.5%)
		명사형전성어미	ETN	216(0.5%)	11,596(0.5%)	205,216(0.6%)
		관형형전성어미	ETM	2,380(5.6%)	101,503(4.8%)	2,129,888(6.6%)
	접두사	체언접두사	XPN	309(0.7%)	28,961(1.4%)	35,756(0.1%)
	접미사	명사파생접미사	XSN	895(2.1%)	37,273(1.8%)	394,549(1.2%)
		동사파생접미사	XSV	41(0.1%)	2,256(0.1%)	26779(0.1%)
		형용사파생접미사	XSA	8(0.0%)	229(0.0%)	4,048(0.0%)
합계				42,360(100.0%)	2,126,258(100.0%)	32,213,813(100.0%)

어미에서는 연결 어미와 관형사형 전성 어미가 높은 비율을 보였는데, 이들은 모든 텍스트 유형에서 비교적 높은 비율을 보였다. 특히 관형사형 전성 어미는 문장에서 용언의 어간에 붙어 관형사와 같은 기능을 수행하게 하는 어미로 '-ㄴ/은/는' 등이 해당한다.

〈표 3〉은 실험 대상 텍스트의 형태소 분석에서 주로 5언 9품사 중심으로 결과를 제시하였지만, 그 외에 텍스트에 쓰인 기호 관련 통계도 추출하였다. 대표적인 기호는 마침표나 물음표, 쉼

표, 콜론, 줄표(예로 표제와 부제 사이에 오는 기호) 등과 숫자를 말한다. 전체 형태소 비율에 기호를 포함시켜 계산하였을 때 서명, 목차, 말뭉치에서 이들이 차지하는 비율은 각각 14.8%, 26.7%, 11.7%였다. 목차에서 매우 높은 비율의 기호가 사용되는 것을 알 수 있는데, 우선 그 사례로 숫자의 사용을 들 수 있다. 목차에서 숫자는 장, 절을 표현하는 동시에 쪽수를 표현하기에 대략 9.5%의 비율을 차지하였다. 나머지 기호 등도 목차의 구조나 레이아웃을 표현하기 위해 다른 텍스트에 비해 상당히 높은 비율로 사용되고 있다. 일례로 "3."이나 "6.4"와 같이 장이나 절을 표현하기 위해 사용한 마침표가 기호 중에 가장 많은 횟수를 차지하였다.

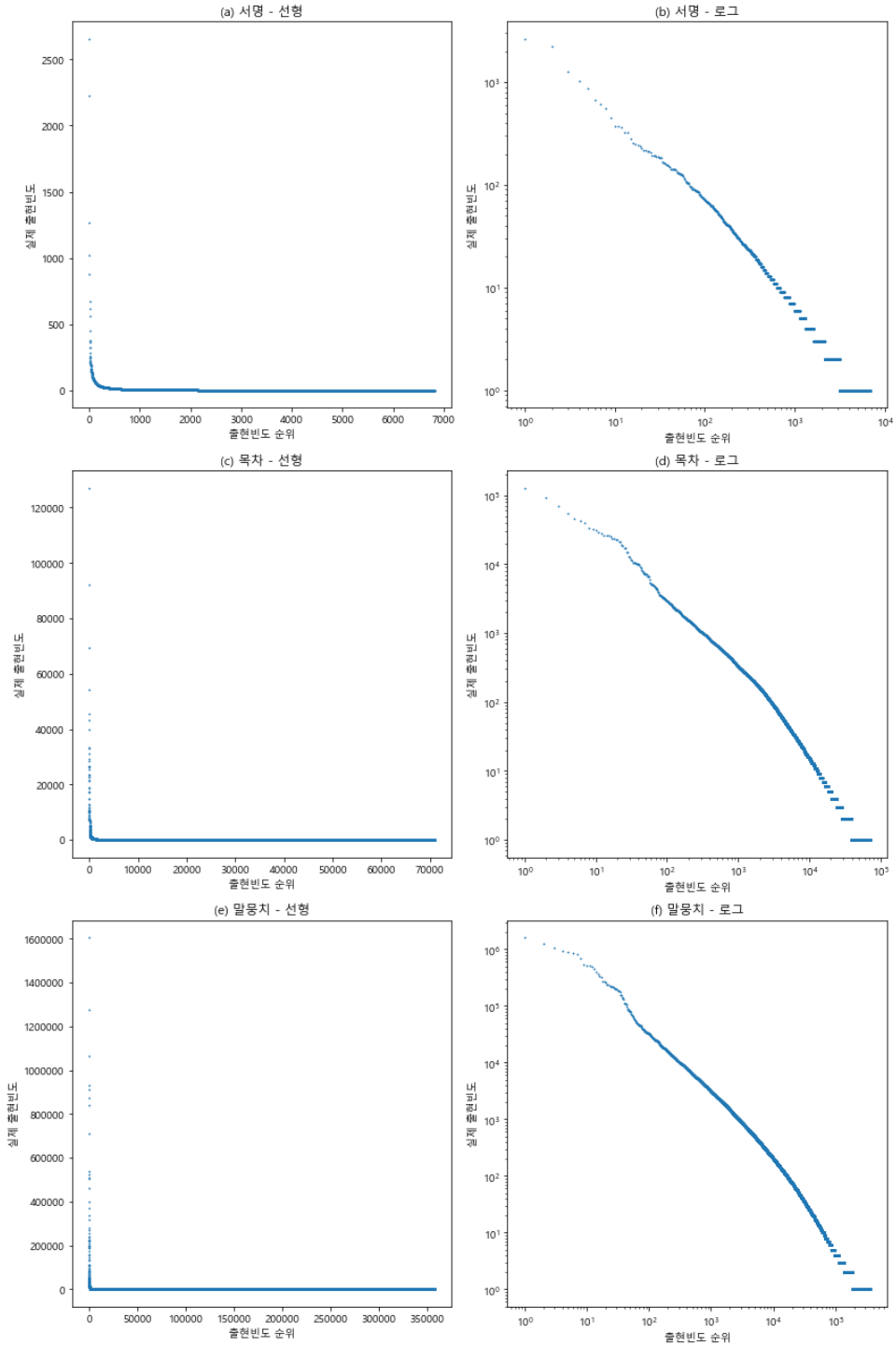
형태소 분석 결과에 따른 서명, 목차, 말뭉치의 통계적 특성을 정리하면 서명과 목차는 매우 비슷한 모습을 갖는다. 서명과 목차 모두 일반적인 텍스트보다 동사의 용언 보다는 명사의 체언이 높은 비중을 차지하며 관형격 조사가 높은 비율을 차지함을 알 수 있다.

서명과 목차가 가지는 단어의 통계적 분포가 지프(Zipf) 법칙, 더 정확히는 멱법칙(power law)을 따르는지 조사하였다. 지프 법칙은 큰 말뭉치에서 한 단어가 얼마나 출현하였는지 계산하고 그 출현빈도에 따라 순위화하였을 때 단어 순위와 출현빈도 사이에 역관계 법칙이 있다는 규칙을 밝혔으며, 이는 인간이 정보를 표현/전달이나 이해하는데 있어 최소한의 노력을 들이고자 함을 뜻한다. 세 종류의 텍스트 서명, 목차, 일반 말뭉치에 대해 형태소 분석 후에 해당 텍스트에 나타난 각각의 형태소의 출현빈도를 구하고 형태소 별로 순위화하여 출현빈도 순위와 실제 출현빈도 값을 계산하였다. 이 두 값을 쌍으로 하여 일반 선형 척도와 로그-로그 척도의 좌표 평면에 표시하면 〈그림 1〉과 같다. 이 그림에서 좌측은 서명, 목차, 일반 말뭉치의 형태소별 출현빈도 순위와 실제 출현빈도 쌍의 선형 좌표 평면이며, 우측은 해당 값을 로그 척도로 변환한 값의 좌표 평면이다. 로그 척도의 우측 세 개의 그림을 보면 약간의 차이는 있지만 모두 일정한 기울기를 보여 지프 법칙을 따르고 있음을 알 수 있다.

3.3 명사 중심의 통계 분석

정보 검색과 자동 분류와 같이 텍스트를 처리하고 이용하는 다양한 응용 프로그램에서는 키워드나 자질로 명사를 사용한다. 일례로 텍스트 범주화에서 분류 자질(features)이 필요하며, 대개 문헌에 출현한 단어들 중 불용어를 제거한 나머지 단어들을 자질로 사용한다. 이 연구에서는 도서의 서명과 목차에서 명사를 추출하고 명사 중심의 통계적 특성을 분석하였다. 앞서 진행한 형태소 분석 결과를 통해 두 텍스트로부터 명사를 추출하였다.

분석 대상 도서 3,213권의 서명과 목차에 대한 형태소 분석을 통해 어절과 명사의 기술 통계를 계산하면 〈표 4〉와 같다. 서명과 목차에 출현한 어절, 형태소 분석 결과로 추출한 명사, 그리고 이 명사에서 중복 제거한 수(고유빈도) 등이다. 표를 보면 서명에 나타난 어절, 명사, 명사의 고유빈도에 대한 평균은 각각 7.0개, 5.81개, 5.45개이다. 한 도서의 서명에서 가장 많이 출현한 어절의 횟수는 총 25개로, 이 도서는 부표제가 달린 수험서에 해당하였다. 서명에서 가장 많은 명사 빈도는 34개로 어절 빈도보다 높다.



〈그림 1〉 서명, 목차, 일반 텍스트의 형태소 빈도 분포

〈표 4〉 서명과 목차에 출현한 단어의 기술 통계

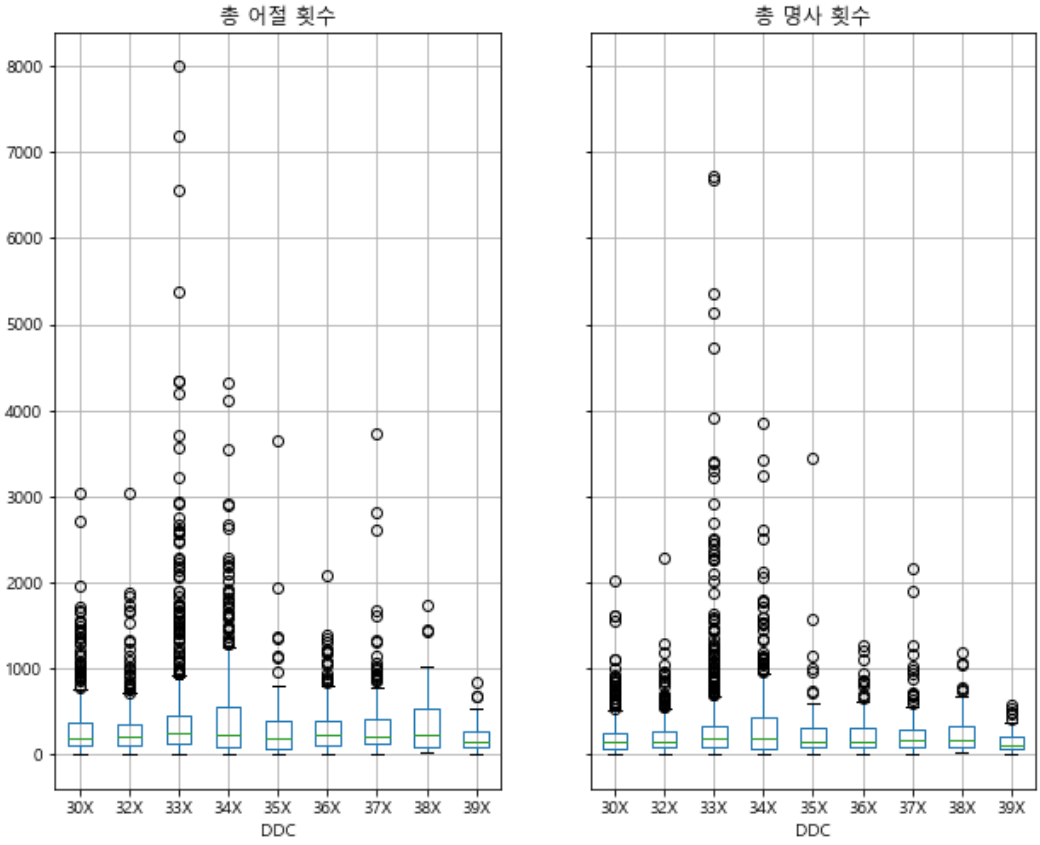
	어절 빈도		명사 빈도		고유빈도(명사)	
	서명	목차	서명	목차	서명	목차
count	3,213	3,213	3,213	3,213	3,213	3,213
mean	7.00	355.60	5.81	267.09	5.45	122.62
std	3.56	495.37	3.29	405.69	2.96	114.20
min	1	1	1	2	1	1
25%	4	104	4	84	3	49
50%	7	220.5	5	162.5	5	92
75%	9	415	7	302	7	159
max	25	7,994	34	6,718	27	1,345

이는 복합명사 때문인데, 빈칸을 기준으로 한 어절은 하나로 세어지지만, 복합명사는 형태소 분석을 거쳐 2-3개의 단일 명사로 분리되어 명사 빈도가 어절 빈도보다 더 큰 수치를 가질 수 있다. 서명이나 목차에서 어절, 명사, 명사의 고유빈도는 각각의 평균(mean)을 볼 때 대체로 차츰 작아진다.

목차의 경우 어절 빈도가 평균 355.49개이며 명사 빈도는 평균 267.01개로 서명의 7.0개, 5.81개에 비해 압도적으로 많았다. 목차의 어절 빈도와 명사 빈도는 서명의 50배 정도 되어 그만큼 더 많은 어휘를 갖는 것으로 보인다. 다만, 표준편차(std)가 매우 커 도서마다 목차의 길이가 매우 차이 남을 알 수 있다.

인터넷 서점에서 획득한 목차가 DDC 분류기호에 따라 어절 또는 명사의 통계적 특성을 보이는지 그림 상자(box plot)을 도시하였다. 〈그림 2〉는 목차에서 출현한 총 어절의 횟수(좌측)와 총 명사의 횟수(우측)를 DDC 분류기호(300대 강목)별로 나타내었다. 이 그림에서 원기호(○)는 이상치(outlier)에 해당하는 사례로 아주 큰 빈도를 의미한다. 주로 33X(경제학/Economics) 분야와 34X(법률학/Law) 분야에 다수가 분포한 것을 알 수 있다. 실제 빈도값은 1에서 7,994번까지 나타났는데, 높은 빈도값은 33X(경제학/Economics) 분야에서 출현하였으며, 이들 도서의 데이터는 내용 목차뿐만 아니라 표와 그림 목차까지 포함하며 수백 쪽의 동향과 전망 보고서 또는 분석 및 실태조사 보고서 등의 부류에 해당한다. 또한 34X(법률학/Law) 분야의 경우 각종 수험서 형태의 도서로 여러 파트(part)로 구분되거나 많은 수의 판례 등을 수록하고 있다. 이러한 결과를 볼 때 표 목차, 그림 목차나 부록 목차를 색인어나 자질로 선정해야 하는지에 대한 정밀한 분석이 필요할 것으로 보인다.

서명과 목차에 나타난 강목별 명사의 출현빈도의 평균과 표준편차를 좀 더 자세히 살펴보면, 〈표 5〉와 같다. 여기서 명사의 출현빈도는 중복을 허용하는 경우와 중복을 제거한 고유빈도로 나누어 제시하였다. 300대 강목별로 서명은 거의 비슷한 명사 빈도를 갖는다. 평균을 보면 대부분 5.15에서 6.95개의 명사로 이루어져 있으며 그 표준 편차도 2.38에서 4.21 정도로 변화하고 있다. 서명에 출현한 명사 빈도는 35X(행정학/Public administration



〈그림 2〉 DDC 300대 강목에 따른 목차의 어절과 명사 빈도 그림 상자

〈표 5〉 강목별 서명과 목차에 출현한 명사 빈도 평균 및 표준편차

강목	도서수	서명				목차			
		출현빈도		고유빈도		출현빈도		고유빈도	
		mean	std	mean	std	mean	std	mean	std
30X	670	5.15	2.81	4.83	2.58	199.33	205.98	110.68	91.97
32X	393	5.54	2.81	5.14	2.52	212.96	217.27	113.48	89.67
33X	811	5.99	3.55	5.62	3.17	348.14	621.37	146.93	146.25
34X	502	6.32	3.75	6.02	3.36	337.73	452.24	126.83	122.75
35X	166	6.95	4.21	6.46	3.57	241.66	332.38	110.11	104.14
36X	194	6.07	3.27	5.60	2.91	235.56	228.72	115.18	90.81
37X	302	5.78	2.80	5.45	2.59	222.40	238.33	107.86	93.05
38X	81	5.17	2.75	4.93	2.53	269.52	262.12	130.12	108.58
39X	94	5.31	2.38	5.04	2.23	151.07	122.69	92.05	70.30
전체	3,213	5.81	3.29	5.45	2.96	267.09	405.69	122.62	114.20

and military science)에서 가장 큰 편차를 보였다. 〈표 4〉에서 서명의 어절 빈도 평균은 7.0개이고 명사 빈도 평균은 5.81개로 이는 서명이 일반적으로 명사 중심의 짧은 단문 형태로 이루어져 있음을 뜻하며, 이러한 경향은 강목별로도 큰 차이가 없음을 알 수 있다. 목차는 이와 매우 다른 양상을 보인다. 목차에서는 39X(풍속, 민속학/Customs, etiquette, folklore) 분야가 151.07로 명사 출현 횟수의 평균이 가장 낮으며, 33X(경제학/Economics) 분야가 348.14로 가장 높다. 이러한 수치로 볼 때 목차에서는 강목 사이에 명사 출현 정도가 큰 차이가 있음을 알 수 있다. 뿐만 아니라 같은 강목 내에서도 목차에 출현한 명사 횟수는 큰 편차를 보였다. 33X와 34X의 경우 표준편차가 621.37과 452.24로 다른 강목에 비해 표준편차가 현저히 큰 것을 알 수 있다. 이는 〈그림 2〉에서 어림잡아 알 수 있듯이 목차에 출현한 단어 또는 그에 따른 길이가 도서에 따라 많은 차이가 난다는 것을 의미한다.

〈표 5〉는 목차에 나타난 명사의 출현빈도와 고유빈도의 차이를 보인다. 이는 목차에서 얼마나 많은 단어가 고유하게 출현하고 이들이 얼마나 반복 출현하는지를 간접적으로 알 수 있다. 목차에서 명사의 고유빈도 평균은 비교적 고른 것을 볼 수 있다. 최저 평균이 92.05에서 최고가 146.93 이다. 고유빈도의 표준편차도 출현빈도의 표준편차보다 최저와 최고의 폭이 크게 작아진 것을 알 수 있다. 구체적으로 33X 분야는 출현빈도의 표준편차가 621.37에서 고유빈도는 146.25로 크게 작아졌다. 34X 분야도 비슷하다. 이렇게 출현빈도와 비교해서 고유빈도의 평균과 표준편차가 현저히 작아진다는 것은 길이가 긴 목차에 특정 단어가 자주 출현한다는 의미로 그 만큼 주요 명사들의 많이 반복되고 있다는 것을 의미한다.

3.4 목차 고유(unique)의 명사 비율 분석

서명에 비해 목차의 장점으로 풍부한 어휘를 생각해 볼 수 있는데 목차가 고유하게 가지는 명사가 어느 정도 인지 파악할 필요가 있다. 이를 위해 서명과 목차 두 텍스트가 제공하는 명사의 양을 측정하고, 서명은 제공하지 않지만 목차만이 제공하는 명사의 비율을 측정하였다. 예를 들어 한 도서에서 'a'와 'b'라는 명사가 출현하였고 목차에서는 'a', 'c', 'd' 명사가 출현하였다면, 'a'는 이미 서명에서 제공되었기에 목차와 중복되지만 'c'와 'd'는 서명이 제공하지 못한 목차만이 제공하는 명사에 해당한다고 볼 수 있다. 목차의 경우 단어 출현빈도가 1 이상인 경우도 많으니까 출현 횟수도 고려하여 목차만이 제공하는 명사의 고유빈도와 출현빈도의 비율을 계산하였다. 즉 앞의 예에서 'a', 'c', 'd'의 출현빈도가 각각 2, 2, 1이라면, 서명 텍스트의 정보가 제공하는 출현빈도 비율은 전체 5개에서 2개에 해당하는 2/5이며, 목차 텍스트만이 제공하는 출현빈도 비율은 'c'와 'd'에 해당하는 3개로 3/5에 해당한다. 이러한 방식으로 목차가 고유하게 제공하는 명사의 비율(확률)을 계산하면 〈표 6〉과 같다. 목차의 명사 중에 서명과 중복되지 않은 명사의 고유빈도 비율의 전체 평균은 95.2%이며, 이를 출현 횟수에 따른 비율(전체 평균)로 표현하면 86.8%에 해당한다. 반대로 서명에 출현한 명사가 목차

〈표 6〉 목차에만 출현한 명사 비율

강목	도서수	출현빈도		고유빈도	
		mean	std	mean	std
30X	670	86.2%	12.0	95.4%	4.9
32X	393	87.1%	10.1	95.4%	4.9
33X	811	87.7%	10.4	96.0%	4.9
34X	502	88.6%	15.9	94.2%	11.1
35X	166	84.8%	13.2	93.5%	8.4
36X	194	85.4%	13.0	95.1%	6.6
37X	302	85.1%	11.4	95.4%	4.5
38X	81	84.6%	13.4	95.1%	5.2
39X	94	85.5%	12.0	94.2%	6.7
전체	3,213	86.8%	12.3	95.2%	6.6

에 출현할 확률이 13.2% 정도 된다는 의미이다. 목차에 출현한 모든 명사가 좋은 색인어가 되는 것은 아니지만, 양적으로 매우 많은 어휘를 목차가 포함하고 있다는 점이 어떤 의미를 갖는지 향후 연구할 필요가 있다.

서명에 출현한 명사 중에 목차에는 출현하지 않고 서명에만 출현한 명사 비율의 전체 평균은 2.37%이며, 서명에만 출현하는 명사의 고유빈도 비율의 전체 평균은 2.31%이다. 추가적으로 서명에 출현한 단어가 모두 목차에 출현한 사례는 733건으로 전체 3,123건 중 23.47%에 해당하며, 반대로 서명에 출현한 단어가 목차에 전혀 출현하지 않은 사례는 176건으로 5.64%에 해당하였다.

4. 논의 및 시사점

최근에 목차의 활용 가능성이 높아짐에 따라 텍스트로서 목차가 가지는 언어적 그리고 통계적 특성을 파악해 보았다. 사회과학분야 도서의 목차는 형태소 분석 결과 서명과 유사하게 절반 정도 명사로 구성되어 있다. 명사는 문헌의 핵심 개념을 나타내거나 주제를 나타내는 대상을 표현하는 형태소로 텍스트 처리와 이해 측면에서 중요하다. 목차가 다수의 명사를 중심으로 구성되므로 문헌의 내용을 대표하는 메타데이터 요소로서 활용할 충분한 가치가 있다.

서명보다 목차는 평균적으로 50배 많은 명사를 포함하고 있다. 이는 짧은 표제보다 문헌의 내용이 덜 축약되어 있는 목차가 상세하게 더 많은 정보를 제공해 줄 수 있음을 의미한다. 이러한 정보 제공은 이용자의 검색, 식별, 발견에서 중요한 역할을 담당할 수 있다. 다만, 목차에 출현한 모든 명사가 이러한 역할을 수행하는 것으로 보이진 않는다.

Dillon과 Wenzel(1990)은 목차가 포함된 서지 레코드는 검색 성능에서 재현율을 향상시키고 정확률을 떨어뜨린다고 하였다. 여기서 재현율의 향상은 서명과 중복되지 않은 명사 또는 키워드가 그 역할을 하는 것으로 보인다. 서명의 색인어만으로 질의어와 매칭이 일어나 검

색될 문헌의 양보다, 고유빈도 기준으로 20배 더 많은 목차의 명사와 매칭되어 검색될 문헌의 양이 확률적으로 더 크다. 이는 재현율의 향상으로 나타난다고 볼 수 있다. 또한 이러한 측면을 보여주는 다른 결과로, 목차에 출현한 명사 중에서 서명과 중복되지 않은 명사의 고유빈도 비율이 있다. 이 연구에서는 그 비율(전체 평균)이 95.2%에 해당한다. 반대로 전체 서명에 출현한 명사 중에 오로지 서명에만 출현한 명사의 고유빈도 비율(전체 평균)은 2.31%이다. 색인 측면에서 서명과 목차에 출현하는 명사는 양적으로 많은 차이를 보인다.

검색 성능에서 정확률도 이와 같은 측면에서 유사하게 해석할 수 있다. 목차의 모든 명사가 해당 도서나 문헌의 주제를 대표하는 색인어로서 적합하지 않다. 어떤 명사는 의례적으로 목차에 사용될 수도 있고, 어떤 명사는 해당 도서의 전체 내용을 대표하기보다 지엽적인 부분을 표현할 수 있다. 목차에 출현한 명사가 가지는 이러한 색인어로서 부정확함은 정확률을 떨어뜨리게 된다. 재현율과 정확률 측면에서 목차의 메타데이터 특성은 양날의 검과 같아 보다 효과적으로 다룰 수 있는 방법에 대한 연구가 필요하다.

5. 결 론

오래전부터 도서관은 그들의 장서가 이용자들에게 보다 효과적으로 이용될 수 있도록 하기 위해 기존의 서지 레코드를 개선하려는 노력을 해왔다. 그중 하나가 목차를 서지 레코드에 반입하고 이를 활용하는 부분이다. 최근에는 도서관 외적으로 인터넷 서점을 중심으로 도서의 목차에 대한 접근과 활용이 증가하고 있다. 목차가 도서의 내용을 풍부하게 표현하는 대체물로서의 가능성을 보이는 상황에서 전통적으로 도서를 수서하고 관리해온 도서관 관점에서 목차를 활용할 때 필요한 기초 정보를 분석할 시점이 되었다.

이 연구에서는 국내 대학도서관에서 수집한 사회과학분야 도서의 서명과 목차를 대상으로 언어학적 측면에서 통계 분석을 통해 텍스트로서 목차 정보의 특성을 분석하고자 하였다. 이들 텍스트에 대해 형태소 분석을 통해 단어들의 통계적 특성을 비교 분석하였으며, DDC 300대 강목에 따른 차이 분석을 하였다. 구체적인 방법으로 2년 동안 국내 대학도서관의 신착 자료 리스트에서 도서 목록을 수집하고, KERIS 종합목록을 통해 이들 도서에 대한 분류기호를 수집하였으며, 인터넷 서점으로부터 해당 도서의 서명과 목차 등의 메타데이터를 수집하였다. 이렇게 수집된 데이터의 텍스트에 대해 형태소 분석을 수행하였으며 필요한 요소를 추출하여 DDC 강목에 따른 통계 분석과 비교 분석을 수행하였다.

이 연구를 통해 얻은 분석 결과는 다음과 같다. 첫째, 형태소 분석을 수행한 결과 사회과학분야의 도서의 서명과 목차에 출현한 형태소 중 명사가 대략 절반 정도를 차지했으며, 동사는 상대적으로 그 비율이 낮았다. 목차는 독특하게 숫자와 기호도 상당히 높은 비율로 사용되고 있는데, 이는 목차의 구조나 레이아웃을 표현하기 위해 사용되었음을 파악하였다. 또한 서명과 목차는 일반 텍스트와 유사하게 지프 법칙, 더 정확히는 멱함수 분포를 따르고 있는 것으로 밝혀졌다.

둘째, 서명과 비교하여 목차는 어절이나 명사 빈도가 대략 50배 정도로 더 많은 어휘를 갖는 것으로 파악되었으며, 고유빈도는 20배 정도에 해당하였다. 목차의 길이는 DDC 300대 강목(사회과학분야)에 따라 편차가 매우 커서 분야 간 차이를 보였다. 구체적으로 경제학과 법률학 분야가 대표적으로 많은 어절과 명사 빈도를 보였다.

셋째, 목차를 서지 레코드에 반입시키거나 메타데이터로 추출해야 하는 근거로 목차만이 제공하는 고유한 명사의 양이나 비율을 파악하였는데, 사회과학분야 도서의 경우 목차에 출현한 전체 명사 중에 서명에는 없고 목차에만 출현한 명사의 비율(전체 평균)이 86.8%이며 고유 빈도의 비율(전체 평균)은 95.2%에 달하는 것으로 나타났다. 반대로 목차에 출현한 전체 명사 중에 서명에도 동시에 출현한 명사의 고유 빈도의 비율이 4.8%이며 출현빈도의 비율은 13.2%이다. 이러한 목차의 특성이 정보검색과 같은 응용 프로그램에 어떤 의미가 있는지는 향후 추가적인 연구가 필요하다.

이 연구는 사회과학 분야 도서의 서명과 목차에 대해 DDC 주제 분야에 따른 통계적 특성을 위주로 분석하였다. 주로 기본적인 통계량과 차이 분석과 비교 분석을 수행하였다. 사회과학 분야 이외에 다른 분야에서는 이러한 통계적 특징이 어떠한 모습을 보이는지 추가 연구가 필요하며, 보다 깊이 있는 연구로 목차 관련하여 주제적 측면이나 구조적 측면을 분석하여 응용 프로그램에서 어떻게 활용할 수 있을지 파악해 보는 것도 중요하리라 생각한다.

참 고 문 헌

구중억, 이응봉 (2009a). Open API 기반 메타 검색시스템의 사용성 평가에 관한 연구. 정보관리학회지, 26(1), 185-214. https://doi.org/10.3743/KOSIM.2009.26.1.185

구중억, 이응봉 (2009b). Open API를 이용한 서지레코드와 목록 확장에 관한 연구. 한국문헌정보학회지, 43(2), 299-328. https://doi.org/10.4275/KSLIS.2009.43.2.299

도현호 (2016). 계층적 클러스터링 기법을 이용한 도서 자동분류에 관한 실험적 연구. 석사학위논문, 계명대학교 대학원, 문헌정보학과.

도현호, 이용구 (2014). 목차, 책 소개를 이용한 단행본 문서 범주화에 관한 기초 연구. 한국정보관리학회 학술대회 논문집, 127-130.

정혜미, 정재영 (2008). 단행본 목차 정보 서비스 제공 모형에 관한 연구. 한국도서관·정보학회지, 39(1), 299-318. https://doi.org/10.16981/kliss.39.1.200803.299

한국도서관협회 문헌정보학용어사전 편찬위원회 (2010). 문헌정보학용어사전(개정판). 서울: 한국도서관협회.

Byrum Jr., J. D., & Williamson, D. W. (2006). Enriching traditional cataloging for improved access to information: Library of congress tables of contents projects. Information Technology and Libraries, 25(1), 4-11. https://doi.org/10.6017/ital.v25i1.3324

Chercourt, M., & Marshall, L. (2013). Making keywords work: Connecting patrons to resources through enhanced bibliographic records. Technical Services Quarterly, 30(3), 285-295. DOI: 10.1080/07317131.2013.785786

Choi, Y., Hsieh-Yee, I., & Kules, B. (2007). Retrieval effectiveness of table of contents and subject headings. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (pp. 103-104). ACM. DOI: 10.1145/1255175.1255195

Dillon, M., & Wenzel, P. (1990). Retrieval effectiveness of enhanced bibliographic records. Library Hi Tech, 8(3), 43-46. DOI: https://doi.org/10.1108/eb047797

Dinkins, D., & Kirkland, L. N. (2006). It's what's inside that counts: Adding contents notes to bibliographic records and its impact on circulation. College & Undergraduate Libraries, 13(1), 59-71. DOI: https://doi.org/10.1300/J106v13n01_07

Morris, R. C. (2001). Online tables of contents for books: Effect on usage. Bulletin of the Medical Library Association, 89(1), 29-36.

Pappas, E., & Herendeen, A. (2000) Enhancing bibliographic records with tables of contents derived from OCR technologies at the American Museum of Natural History Library. Cataloging & Classification Quarterly, 29:4, 61-72, DOI: 10.1300/J104v29n04_05

Tosaka, Y., & Weng, C. (2011). Reexamining content-enriched access: Its effect on usage and discovery. College & Research Libraries, 72(5), 412-427. DOI: https://doi.org/10.5860/crl-137

Van Orden, R. (1990). Content-enriched access to electronic information: Summaries of selected research. Library Hi Tech, 8(3), 27-32. DOI: https://doi.org/10.1108/eb047795

Winke, R. Conrad. (1999). An analysis of tables of contents in recent english-language books. Library Resources & Technical Services, 43(1), 14-27. DOI: http://dx.doi.org/10.5860/lrts.43n1.14

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Do, Hyun-Ho (2016). An experimental study on using hierarchical clustering method for automatic classification of books. Unpublished master's thesis, Keimyung University, Department of

Library and Information Science.

Do, Hyun-Ho, & Lee, Yong-Gu (2014). A preliminary study on text categorization of book using table of contents and book description. In Proceedings of the 21th Conference of Korean Society for Information Management, 8, 127-130.

Gu, Jung-Eok, & Lee, Eung-Bong (2009a). A study on the construction and usability test of meta search system using Open API. Journal of the Korean society for information management, 26(1), 185-214. https://doi.org/10.3743/KOSIM.2009.26.1.185

Gu, Jung-Eok, & Lee, Eung-Bong (2009b). A study on the bibliographic records and the expansion of library catalog using Open API. Journal of the Korean Society for Library and Information Science, 43(2), 299-328. https://doi.org/10.4275/KSLIS.2009.43.2.299

Jeong, Hye-Mi, & Chung, Jae-Young (2008). A study on the providing model of table of contents for monography. Journal of Korean Library and Information Science Society, 39(1), 299-318. https://doi.org/10.16981/kliss.39.1.200803.299

Korean Library Association (2010). The glossary of library and information science(revised edition). Seoul: The Korean Library Association.