말모둠의 디지털도서관 활용에 관한 연구

A Study on the Practical Use of Corpus in Digital Library

성기주, 동덕여자대학교 문헌정보학과
이은정, 동덕여자대학교 대학원 문헌정보학과

Sung Kee Joo, Lee Eun Jung
Dept. of Library and Information Science, Dongduk Univ.

말모둠은 디지털화된 텍스트의 모음으로 주로 언어 연구를 위한 자료로 이용되었다. 그러나 말모둠의 구축사례가 늘면서 검색기법이 다양해지고 내용의 범위가 확대되어 말모둠의 역할은 보다 다양해지고 있다. 본 연구는 말모둠의 개념, 구축현황, 웹사이트에서의 제공 내용 등을 조사하여 디지털도서관에서의 말모둠 활용 가능성을 알아보았다.

1. 서론

디지털도서관의 개념이 확산되면서 도서관 자료들은 웹에서도 중요한 자료로 이용되고 있다. 이 중 사전은 언어를 학습하는 초등학생에서부터 연구자에 이르기까지 유용하게 쓰이는 참고자료이며 특히 도서관 사서에게는 업무상 없어서는 안될 기본 자료이다. 말모둠은 디지털화된 일종의 용례 사전으로 대표적인 말모둠들은 제공기관이 임의로 분류하여 웹에서 서비스하고 있다. 일반적으로 말모둠에서 검색된 자료들은 전문(full text)으로 제공되는데 이것은 디지털도서관에서 제공하는 원문정보서비스와 유사한 성격이다. 현재 말모둠은 일부 언어학연구자들에 의해 소극적으로 이용되고 있다.

이에 본 연구는 말모둠의 개념 및 구축 현황, 말모둠 사이트에서의 웹서비스 내용 등을 알아봄으로써 디지털도서관에서의 말모둠 활용 가능성을 논의해보고자 한다.

2. 말모둠

2.1 말모둠의 개념 및 수록대상

코퍼스(corpus)는 라틴어에서 유래된 명칭으로 당시 coupus juris, Corpus Juris Civilis 등과 같이 집대성, 전집, 대전(大全)의 뜻으로 주로 법전에 관련된 용어로 사용되었다. 현대 사전에선 집대성, 연구를 위해 수집한 자료, 언어자료, 신소체(공학) 등의 의미로 정의하고 있다. 특히 언어자료와 관련된 명칭으로 Baronbrook은 특정 유형의 언어를 대표하기 위해서 선택되어 컴퓨터로 읽을 수 있는 형태로 되어 있는 텍스트의 모음으로 정의하였다(유석훈, 1999). 즉 컴퓨터의 등장으로 관련분야 연구자들 사이에서 디지털화된 텍스트 모음을 대표하는 의미로 코퍼스라는 용어가 사용된 것이다. 코퍼스는 가공여부에 따라 말모둠과 말뭉치(raw corpus)로 나뉜다. 여기서 말모둠은 수집된 텍스트 데이터베이스를 형태소 분석

이나 어휘별, 품사정보별, 문헌별, 장/절별, 내용별 분류가 가능하도록 가공한 것이며 말뭉치는 텍스트를 기계가독형 자료로 만들어 단순히 데이터베이스화 한 것을 의미한다.

말모둠의 수록 대상은 언어로 이루어진 모든 자료로 인문, 사회, 과학 등 전 주제 분야가 포함된다. 말모둠엔 이러한 자료들을 다시 샘플 문헌, 샘플어구로 발췌하여 싣거나 자료의 전문을 그대로 수록하는데 현재 대부분의 말모둠엔 전문을 수록하거나 수록된 자료를 전문으로 연결하여 제공하고 있다

2.2 말모둠의 전개 및 기술

말모둠은 1950년대 미국의 구조주의 언어학자들에 의해 시작되었다. 당시의 관심사는 조사된 특정 유형의 언어자료들을 어떻게 찾아보기 편하게 배열하여 구조화할 것인가에 맞춰졌는데 그 결과는 손안의 책과 같은 형태의 것이었다. 다음시기의 말모둠은 1959년 Randolph Quirk가 발표한 Survey of English Usage(SEU)와 비슷한 시기에 Brown University의 Nelson Francis와 Henry Kucera가 발표한 Brown Corpus가 대표적이다. 그러나 이 시기의 말모둠은 지금의 디지털화된 말모둠과는 차이가 있다. 1970년 후반 Lund의 Jan Svartvik이 이끈 연구팀은 Brown Corpus와 SEU의 장점을 살려 지금의 말모둠과 흡사한 Lodon-Lund Corpus(LLC)를 완성하였다. 이후 미국, 유럽, 오스트레일리아, 싱가폴, 홍콩 등지에서 대규모 말모둠 프로젝트가 진행되었는데 현재 British National Corpus(BNC), Cobuild : Bank of English, International Computer Archive of Modern and Medieval English(ICAME), International Corpus of English(ICE), Oxford Text Archive(OTA), Linguistic Data Consortium(LDC), European Language Resources Association(ELRA) 등에서 다양한 말모둠을 제공하고 있다. 국내에서는 Kaist와 고려대 언어학연구소 등에서 한국어 말모둠 구축에 관한 연구가 진행 중이다.

현재 웹에서 서비스되고 있는 말모둠들은 대부분 TEI 표준을 수용한 SGML 코드로 작성되어 운영되고 있다. 때문에 말모둠의 형식에 내용의 주제와 서지사항을 표기하는 것이 비교적 간단한데 SGML로 작성된 문서의 헤더에 주제와 서지사항을 입력하는 방법을 이용할 수 있다. 이를 검색에 활용하면 내용의 서지사항별, 주제별 검색이 가능하다. 최근 작성된 말모둠 중엔 각 내용에 제작자가 임의로 작성한 주제기호를 부여한 것도 있고 보다 효율적인 검색을 위해 통합자연어검색기법, 메타데이터기법 등을 응용한 것도 있다.

3. 말모둠 사이트

3.1 말모둠 사이트 선정 및 웹 서비스 개요

말모둠은 구축기법과 운영주체의 목적에 따라 다양한 형태로 제공될 수 있다. 본 연구에서는 영어자료를 대상으로 구축된 말모둠을 제공하는 사이트들 중 공공기관 중심 산학연 콘소시엄에서 운영하는 BNC, 사전제작업체에서 상용으로 운영하는 Cobuild : Bank of English, 대학 중심 산학연 콘소시엄에서 운영하는 LDC, 대학에서 운영하는 OTA 등의 네 곳을 선정하여 각 사이트에서 제공하는 웹서비스의 내용을 살펴보았다.

본 연구에서 조사한 사이트들에서는 말모둠 및 관련 문헌(연구논문, 자관출판물) 목록 검색 서비스, 말모둠 검색 서비스, 말모둠 이용 안내 서비스, 온라인 참고 질의 서비스 등을 공통으로 제공하고 있는데 자료를 제공받는 과정은 디지털도서관과 비슷하다. 특히 OTA 목록 검색 서비스의 경우 말모둠은 관련 문헌과 자료의 형식이 다름에도 불구하고 말모둠들과 관련 문헌을 통합 목록으로 작성하여 같은 방식으로

검색할 수 있다. 이는 말모둠 사이트들에서 제공하는 서비스와 도서관의 정보서비스가 유사한 성격이기 때문이다.

3.2 British National Corpus (BNC) - http://info.ox.ac.uk/bnc/index.html

BNC는 Oxford University Press, Lonman Group Ltd. Chambers, University of Lancaster, Science and Engineering Council(EPSRC), Joint Framework for Information Technology(JEIF)산하 DTI, British Library, British Academy 등으로 구성된 콘소시엄을 중심으로 1991년 시작되어 1994년에 완성되었다. 내용은 근대 영어를 대상으로 1억여 단어를 수록하고 있으며 문어부분과 구어부분이 각각 90%, 10%의 비율로 구성되었고 단어 단위로 University of Lancaster에서 배포한 CLAWS로 분류, 태깅되었다. 여기에서는 관련 문헌과 말모둠을 일정부분 무료로 검색할 수 있는데 말모둠은 SARA라는 전용 탐색브라우저를 이용하면 고급검색이 가능하다. 그러나 웹에서 원문을 직접 볼 순 없으며 샘플링된 구문과 서지사항을 검색결과로 볼 수 있다. BNC는 최근에 작성된 대규모의 말모둠으로 공공기관이 참여하여 작성하였기 때문에 영국에서 제작되는 다른 말모둠들의 기준이 되고 있다.

3.3 Cobuild : Bank of English - http://titania.cobuild.collins.co.uk/

Bank of English는 University of Birmingham의 School of English에서 1980년부터 사전 편찬을 목적으로 구축된 Bermingham Corpus를 바탕으로 확대 제작된 것이다. 지금의 말모둠은 Harper Collins가 참여한 1991년부터 본격적으로 제작되었으며 1994년부터 유료로 서비스되고 있다. Bank of English는 2000년 9월 기준 415백만 단어의 미국식, 영국식 영어의 용례가 수록되어 있다. 말모둠은 ozws(Australian news), ukephem(UK ephemera), ukmags(UK magazines), ukspok(UK spoken), usephem (US ephemera), bbc(BBC World Service), npr(National Public Radio), ukbooks(UK books), usbooks(US books), times(Times newspaper), today(Today newspaper) 등 11가지로 Tellnet, Java 검색을 할 수 있으며 검색 결과는 FTP로 다운받을 수 있다. 특히 여기에서는 말모둠을 단어를 이용한 게임, Wordwatch 등과 같은 기초언어학습을 위한 서비스와 같이 제공하고 있다.

3.4 Linguistic Data Consortium (LDC) - http://www.ldc.upenn.edu/

LDC는 University of Pennsylvania가 중심이 되고 북미의 대학, 기업, 정부출현 연구소 등의 지원을 받아 1992년에 설립되어 지금은 100여 곳이 넘는 회원기관이 참여하고 있다. 현재 Broun Corpus을 포함해 각 기관에서 소유한 자료를 바탕으로 제작한 총 200가지 이상의 소규모 말모둠과 그 관련 문헌을 일부는 무료로 일부는 유료로 제공하고 있다. 특히 Speech관련 자료가 풍부한데 구어 말모둠의 경우 전화 혹은 커뮤니티 활동을 통해 이용자가 직접 말모둠 제작에 참여할 수 있다. 검색 결과는 웹에서 전문으로 볼 수 있으며 결과가 방송 혹은 통화 내용인 경우 음성파일로도 받아볼 수 있다.

3.5 Oxford Text Archive (OTA) - http://ota.ahds.ac.uk/

OTA는 1976년 Lou Burnard에 의해 연구자들의 연구자료 제공과 최신정보주지를 목적으로 설립되었다. 여기에서 서비스하고 있는 자료는 라틴어, 불어 등 다양한 국적의 언어로 이루어진 연구기반자료로 다양한 언어로 작성

된 말모둠과 고대에서 중세(18C~19C)에 걸쳐 영어로 작성된 개인출판문헌을 방대하게 포함하고 있다. 현재 제공되는 모든 자료는 Arts and Humanities Data Service(AHDS)의 주관으로 유료로 제공되나 일부 무료로 이용할 수 있는 것도 있다. 제공되는 말모둠들은 일반 문헌과 함께 통합된 목록에서 저자별, 주제별, 언어별로 검색할 수 있는데 원하는 자료의 원문을 보기 위해선 웹 목록에서 검색하여 원문복사서비스를 신청해야한다. 말모둠의 경우 신청이 완료되면 CD로 우송받거나 FTP로 다운받을 수 있다.

4. 결론

말모둠은 디지털화된 언어 자료로 넓은 범위의 주제 자료를 포함하고 있으며 실제 자료의 내용을 정확하게 기술하여 작성된다. 이러한 말모둠은 기계가독형으로 바뀌는 과정에서 다양한 검색 기법이 도입되었고 수록 자료는 대부분 전문으로 제공되고있다. 그 결과 지금의 말모둠에 수록된 자료는 그 양이 방대하고 다양한 방법으로 검색이 가능하므로 말모둠을 단순한 언어 자료로 단정짓기는 어렵다.

본 연구에서 조사한 말모둠 사이트들에서는 도서관의 정보서비스와 유사한 내용을 제공하고 있다. 즉 말모둠에 관련된 각종 목록을 검색하여 필요한 부분은 전문으로 제공받을수 있으며 주제별 검색을 위한 작성방법이 연구되면서 디지털 백과사전의 역할도 가능하다. 또한 각 사이트들에서 말모둠을 CD 형태로 배포하고 개별 관리하므로 주제적 성격이 더해진다면 '손안의 도서관' 또는 'E-Book'과 같은 자료가 될 수있을 것이다. 이는 말모둠의 기본 성격이 도서관 자료와 유사하기 때문이다. 따라서 말모둠을 디지털 도서관의 자료로 흡수하여 이용한다면 이용자들의 데이터베이스 활용 범위가 확대되어, 그결과 이용자 서비스가 한단계 향상될것이다.

5. 참고문헌

김홍규, 강범모. 1995. 고려대학교 한국어 말모둠 1 (KOREA-1 CORPUS): 설계 및 구성. 한국어학, 3. 233-258.

유석훈. 1999. 언어와 컴퓨터. 서울: 고려대출판부.

Fries, U, Tottie, G and Schneider, P. 1994. Creating and Using English language corpora. Amsterdam-Atlanta; GA: Rodopi.

Sinclair, J. 1991. Corpus, Concordance, Collocation. Hong Kong: Oxford University Press.

Using the BNC : http://info.ox.ac.uk/bnc/using/index.html.

CobuildDirect User Guide : http://titania.cobuild.collins.co.uk/cdguide/svenguide.html.

About LDC : http://www.ldc.upenn.edu/About/

About the OTA : http://ota.ahds.ac.uk/ota/index.html