꼬꼬마 : 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구

(KKMA : A Tool for Utilizing Sejong Corpus based on Relational Database)

이 동 주 † 연 종 흠 †
(Dongjoo Lee) (Jongheum Yeon)
황 인 범 † 이 상 구 ††
(Inbeom Hwang) (Sang-goo Lee)

요 약 말뭉치는 언어학 분야에서 다양한 연구를 위한 기초자료로서 활용된다. 국내에서도 세종 21세기 계획 등을 통해서 몇몇 대용량 말뭉치가 구축되었으나, 다수의 사용자가 쉽게 활용할 수 있는 활용 도구에 대한 연구는 여전히 부족하다. 본 논문에서는 한국어 대용량 말뭉치 중 하나인 세종 현대 국어 말뭉치를 관계형 데이터베이스에 저장하여, 다양한 방법으로 활용할 수 있도록 지원하는 말뭉치 활용 도구에 대한 설계 및 구현 방법을 보인다. 웹 기반의 말뭉치 활용 시스템을 구축하였고, 실제로 언어학 연구자들에게 사용되고 있다.

키워드 : 말뭉치 언어학, 세종 말뭉치, 관계형 데이터베이스, 말뭉치 활용 도구

Abstract Corpus is widely used as a fundamental resource for various purposes in linguistic studies. There are several large corpora such as Sejong corpus in Korea. However, it is hard to find a tool utilizing such large corpora. In this paper, we propose a method of utilizing Sejong corpus based on the relational database. We designed the relational database scheme to store corpus and implemented a Web-based application so that many researchers can easily access and utilize the Sejong corpus.

Key words : Corpus Linguistics, Sejong Corpus, Relational Database, Corpus Utility

1. 서 론

말뭉치(Corpus)는 언어 연구를 위한 매우 중요한 자료이다. 1960년대 구축된 브라운 말뭉치(Brown Corpus)[1]는 미국 영어의 한 표본으로, 언어학, 교육학, 통계학, 사회학적인 다양한 통계 분석에 활용되었다. 이후에도 다양한 특성을 반영하는 말뭉치가 만들어졌으며, 말뭉치언어학(Corpus Linguistics)이라는 세부 분야에서 말뭉치에 대한 다양한 분석 및 연구가 수행되고 있다. 국내에서도 연세 한국어 말뭉치와 세종 말뭉치 등 대량의 말뭉치가 구축되어, 많은 연구자들에게 공개되어 다양한 목적으로 활용되고 있다. 특히 최근 발간된 세종 말뭉치는 그 규모나 질 측면에서 많은 발전을 이루었다.

그러나, 언어학 관련 연구자들의 대부분은 문서 편집기나, 기초적인 프로그래밍을 이용하여 말뭉치를 활용하고 있기에, 말뭉치 활용이 비효율적이고, 말뭉치 활용을 위한 노력이 중복 투자된다는 문제를 야기한다.

본 논문에서는, 이 같은 문제점을 극복하고자, 말뭉치를 구조화 하여 데이터베이스에 저장하고, 말뭉치에 내재된 다양한 통계 정보나 용례를 활용할 수 있도록 하는 말뭉치 활용 시스템에 대한 설계 방법과 실제 구현 사례를 보이고자 한다. 특히 관계형 데이터베이스(Relational Database)에 저장된 말뭉치는 다양한 사용자와 프로그램에서 쉽게 활용될 수 있어, 말뭉치를 저장하고, SQL을 이용하여 다양한 통계 정보를 쉽게 얻어낼 수 있음을 보여, 다양한 연구에 활용할 수 있음을 보인다. 이 같은 말뭉치 활용 도구는, 언어학자들이 말뭉치를 활용하는데 있어 불필요한 노력을 중복으로 투자하는 것을 방지하며, 데이터베이스 분야의 기술을 통해 통계 정보에 대한 접근성과 활용성을 극대화 할 수 있다는 장점을 가진다.

본 논문의 구성은 다음과 같다. 먼저 제2절에서 관련 연구에 대해서 소개한다. 제3절에서는 데이터베이스를

· 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성·지원사업(NIPA-2010-C1090-1031-0002)의 연구결과로 수행되었음
· 이 논문은 2010 한국컴퓨터종합학술대회에서 '꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용도구'라는 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 서울대학교 컴퓨터공학부
therocks@europa.snu.ac.kr
jonghm@europa.snu.ac.kr
inbeom@europa.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
sglee@snu.ac.kr

논문접수 : 2010년 8월 9일
심사완료 : 2010년 10월 6일

1) http://khnt.aksis.uib.no/icame/manuals/brown/

이용한 말뭉치 활용 시스템에 대해서 기술하고, 제4절에서는 말뭉치를 관계형 데이터베이스에 저장하기 위한 스키마 설계에 대해 설명한다. 제5절에서는 저장된 말뭉치로부터 다양한 용례 및 통계 정보를 획득하는 방법을 기술하고, 제6절에서는 구현한 '꼬꼬마 세종 말뭉치 활용 시스템'을 소개한다. 제7절에서 결론과 함께 향후 과제를 논하며 끝 맺는다.

2. 관련 연구

2.1 말뭉치 언어학

말뭉치언어학은 실제 언어나 실제 언어의 샘플을 이용하여 언어에서 나타나는 다양한 현상과 규칙들을 연구하는 응용 언어학의 한 분야이다. 말뭉치는 실제 언어 샘플의 한 형태이며, 말뭉치 구축은 어떤 문서에서 얼마만큼을 선택하느냐 하는 샘플링에서 시작하여 필요한 정보를 나타내기 위한 주석을 달기로 마무리 된다.

말뭉치 언어학의 기점은 브라운 말뭉치를 구축하고 연구한 '현대 미국 영어의 전산 분석'의 출간이다. 브라운 말뭉치는 현대 미국 영어의 한 표본으로 다양한 연구 방법에 의해 언어의 현상들을 제시했고, 이후에도 LOB말뭉치, 런던-룬트 말뭉치, 국제 영어 말뭉치(ICE, International Corpus of English) 등 다양한 말뭉치들의 모범이 되었다[1]. 국내에서도 몇몇 말뭉치가 구축되었는데, 대부분 사전 구축과 언어 용례를 분석하는데 사용되었다. 대표적으로 연세 한국어 말뭉치, 고려대학교 한국어 말모둠, 세종 말뭉치가 있다.

2.2 말뭉치 활용 도구

말뭉치 활용도구의 개발은 말뭉치 구축과 함께 해 왔으며, 다양한 그룹에서 지속적으로 개발되어 활용되고 있다. 브라운 말뭉치의 경우 Python을 기반으로 한 활용 도구가 개발되었고, 현재에도 개선되고 있다.[2] 국제 영어 말뭉치의 일부인 ICE-GB와 같이 구문 분석된 말뭉치에 대해서 다양한 검색 및 활용을 지원하는 도구가 배포되고 있기도 하다.[3]

국내에서도 90년대 초부터 전산언어학 분야에서 대용량 말뭉치에 대한 연구와 이를 활용하는 도구에 대한 연구가 진행되고 있다[2,3]. 최근에는 말뭉치의 양이 더욱 커지고, 컴퓨터의 성능이 개선됨에 따라 대용량 말뭉치에 대한 효율적인 검색을 지원하는 도구가 개발되었다[4,5].

대부분의 말뭉치 활용 도구는 말뭉치에 어느 정도 종속되는데, 본 논문에서는 국내 언어학자들이 가장 쉽게 사용할 수 있는 세종 말뭉치에 초점을 맞추고자 한다. 세종 말뭉치에 대해서는 몇 가지 활용 도구가 개발되어 배포되고 있다. 용례 검색은 국립국어원 누리집[4]을 통해서 수행할 수 있으며, 통계 정보 활용을 위한 프로그램도 몇몇 공개되어 있다. 한마루[5]는 세종 말뭉치의 현대 국어 기초 말뭉치를 대상으로 용례를 검색하고, 어절, 형태소, 품사 등에 대한 통계 정보를 조회할 수 있는 프로그램이다. 이외에도 고려대학교 임해창 교수 연구실에서 제작한 글잡이2와 같은 프로그램이 있으며, 영어권에서 사용되는 WordSmith[6]나 MonoConc[7] 같은 프로그램을 활용할 수도 있다.

그러나, 이 같은 프로그램은 말뭉치를 개인의 컴퓨터에 저장하여 활용하는 것들로 말뭉치가 개선되어 새로이 배포될 때마다 각 사용자가 이를 받아 사용해야 하며, 프로그램이 제공하는 기능만을 사용자가 활용할 수 있어 제공하지 않는 정보의 획득이 쉽지 않다. 본 논문에서는 세종 말뭉치의 현대 국어 말뭉치에 초점을 맞추어, 말뭉치를 관계형 데이터베이스에 저장하여 다수의 사용자가 쉽게 사용할 수 있도록 하고 SQL을 이용한 정보의 획득 및 기능의 확장을 용이하게 하는 프로그램에 대한 설계와 구현을 보이고자 한다.

3. 데이터베이스를 이용한 세종 말뭉치 활용 시스템

세종 말뭉치는 1998년부터 2007년까지 10년간 정부 지원 과제로 수행된 언어자원 구축 과제인 21세기 세종계획에서 구축된 것으로 현재에도 일부 자료가 꾸준히 개선되어 배포되고 있다. 세종 말뭉치는 여러 말뭉치의 집합이라 할 수 있는데 여기에는 현대 문어, 구어, 북한 및 해외, 역사 자료, 한·영 및 한·일 병렬 말뭉치가 포함된다. 세종 말뭉치에 대한 좀더 자세한 내용은 [6]에 기술되어 있다. 본 논문에서는 현대 문어와 구어를 대상으로 하고자 하며, 이후 세종 말뭉치라 함은 현대 문어 및 구어 말뭉치에 국한됨을 밝힌다.

세종 말뭉치의 모든 텍스트는 TEI(Text Encoding Initiative)의 부호화 방식을 따라서 헤더(Header)와 본문(Text)으로 구성되어 있다. 본문에 어떤 분석 내용과 이를 위한 주석을 부착하였느냐에 따라서, 원시 말뭉치와 형태, 의미, 구문 분석 말뭉치로 구분된다. 각 말뭉치는 신문, 잡지, 책 등 다양한 출처에서 선정되었으며, 다루는 주제 또한 다양하기에, 다양한 분야에서 사용되는 한국어의 통계적 특성들을 분석하는데 유용하게 사용될 수 있다. 구어 말뭉치는 화자의 연령, 직업, 성별, 거주

2) http://www.nltk.org/
3) http://www.ucl.ac.uk/english-usage/resources/icecup/
4) http://www.sejong.or.kr/
5) http://www.sejong.or.kr:8000/sjdest/jsp/pds/programdown.jsp
6) http://www.lexically.net/wordsmith/
7) http://www.athel.com/mono.html

지 등에 대한 정보를 포함하고 있어, 화자의 특성에 따른 한국어의 언어학적 특성을 연구하는데 사용될 수도 있다.

세종 말뭉치는 파일 단위로 관리된다. 각 파일의 헤더에는 말뭉치 파일의 출처, 제목, 저자, 주제 및 분야에 대한 내용과 인코딩, 변경 내역 등 관리를 위한 정보들이 태그로 기술된다. 본문은 장이나 절 혹은 단락 단위가 구분되어 있으며, 형태, 의미, 구문 분석에 따라서 다른 방법으로 주석 부착되어 있다.

파일 단위로 관리되는 말뭉치를 활용하기 위해서는 파일을 처리하고, 사용자가 필요로 하는 정보를 추출하도록 하는 프로그램을 각 기능별로 일일이 구현해야 한다. 그러나, 말뭉치는 개념적 수준에서 구조화 할 수 있고, 이를 데이터베이스로 구축한다면 좀더 유연하게 말뭉치를 활용할 수 있고, 다양한 이점을 얻을 수 있다.

우선, 말뭉치를 데이터베이스로 구조화 하여 관리하면, 그림 2에서 보이는 것처럼 말뭉치를 파일로 배포하지 않아도 많은 사용자들이 직·간접적인 다양한 경로로 최신의 말뭉치를 활용할 수 있다. 또한, 관계형 데이터베이스로 말뭉치 데이터베이스를 구축하면, 관계형 데이터베이스가 제공하는 SQL과 집계 함수를 이용해서 다양한 정보들을 쉽고 빠르게 추출할 수 있다. 이를 위해서는, 말뭉치가 관계형 데이터베이스에 저장되어야 하는데, 다음 절들에서는 정보의 손실 없이 말뭉치를 관계형 데이터베이스에 저장하고, 필요한 정보를 추출하는 방법에 대해서 기술한다.

4. 말뭉치 저장을 위한 관계형 데이터베이스 스키마의 설계

개념적 수준에서 살펴보면, 말뭉치는 분석 결과가 태그로 부착된 파일(File)들의 집합이다. 각 파일은 장, 절, 단락이 계층 구조로 구성되고, 장, 절, 단락에서 문장을 포함하고 있는 최소 단위를 부분(Part)으로 하면, 각 파

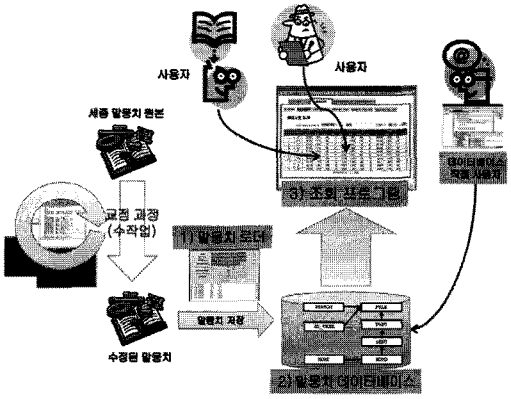

그림 2 말뭉치를 저장하기 위한 관계형 데이터베이스 구조

일은 부분들의 나열이다. 각 부분은 문장(Sentence)들의 나열이며, 문장은 어절(Word)의 나열이다. 각 어절은 형태소(Morpheme)들의 결합이다. 추가적으로, 각 부분에서의 화자를 표현하기 위한 화자(Person)와 부분에 대한 관계(Speak)를 정의할 수 있다.

이와 같은 말뭉치의 개념적 구조는 그림 1에 표현된 논리적 구조로 관계형 데이터베이스에 저장될 수 있다. 관계형 데이터베이스에서 각 테이블은 집합으로 순서가 고려되지 않기 때문에 부분, 문장, 어절 및 형태소의 순서를 고려하기 위한 구분자를 정의할 필요가 있다. 이를 위해 Part, Sentence, Word, Morpheme 테이블은 각 순서를 나타내기 위한 part_id, sentence_id, word_id, morpheme_id 등의 구분자를 가진다. 구분자를 이용하면, 다양한 수준에서의 용례, 연어 및 통계 정보를 조회할 수 있다. 이에 대해서는 다음 절에서 예를 통해 자세히 기술한다.

말뭉치 저장을 위한 각 테이블에 대해서 좀더 살펴보자. 먼저, File 테이블은 파일이 가지는 메타 정보를 포함한다. 이는 각 말뭉치 파일의 헤더에 포함된 정보들이고,

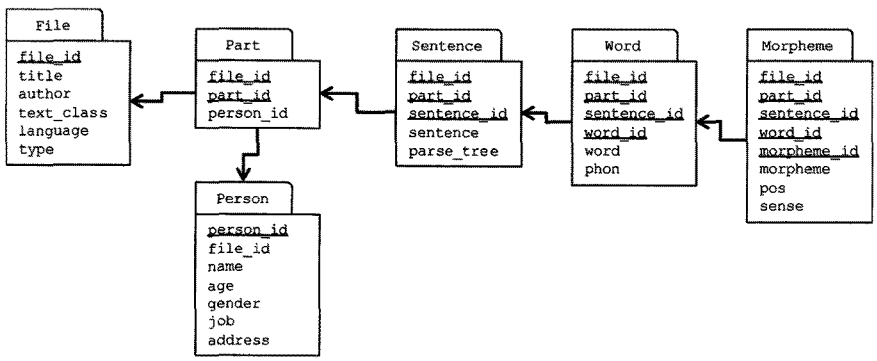

그림 1 말뭉치 데이터베이스의 구축 및 활용

그림 1에는 이중 일부만이 표시되었다. Part 테이블은 실제 파일에서의 계층 구조를 표현하기 위해 사용된 태그를 포함하고 있으나, 이는 연어나 각종 통계 정보를 표현하고 활용하는 데에는 불필요하기 때문에 본 논문에서는 생략하였다. Sentence 테이블은 문장 원본을 sentence 속성에 저장하고, 해당 문장에 대한 구문 분석 결과를 parse_tree 속성에 xml로 저장한다. 각 어절의 원본은 Word 테이블의 word 속성에 저장되며, phon 속성에는 이에 대한 발음 표기가 저장된다. 한국어는 교착어로 조사나 어미가 부착되면서 하나 이상의 단어가 하나의 어절을 구성하는데, 이 같은 과정에서 그 형태가 변하거나 줄어드는 경우가 있기에 이같이 원형을 유지하여 저장할 필요가 있다. Morpheme 테이블에는 형태소의 형태와 범주가 morpheme, pos 속성에 각기 저장된다. 형태소가 하나 이상의 의미를 가질 수 있는데, 이를 위해 sense 속성에 어깨번호로 구분된 의미를 저장한다. Part, Sentence, Word, Morpheme 테이블에는 각 요소가 어떤 상위 요소에 포함되는지를 나타내기 위해 각 상위 요소의 구분자를 저장하는데, 이를 이용해 다양한 수준에서의 연어나 용례 및 통계 정보를 쉽게 조회할 수 있다.

5. 말뭉치 활용

말뭉치가 관계형 데이터베이스에 저장되면 SQL을 이용하여 용례나 연어에 대해 조회할 수 있다. 파일단위로 관리되는 말뭉치는 원본 말뭉치, 형태, 의미, 구문 분석 말뭉치를 저장하는 파일이 따로 관리되어 이들을 유기적으로 활용하기 쉽지 않은데, 제시한 관계형 데이터베이스 스키마에는 이들이 개념적 수준에서 통합되어 저장되기에 JOIN 연산과 다양한 집계 함수를 이용해 말뭉치를 유기적으로 활용할 수 있다.

본 절에서는 용례 조회, 통계 정보 추출, 연어 조회의 예를 통해서 제시한 관계형 데이터베이스 스키마가 어떻게 사용될 수 있는지 살펴 본다.

5.1 용례 검색

용례는 어떤 어휘가 사용되는 문장의 예를 말한다. 하위 수준의 어절이나 형태소를 검색하고 해당 어절이나 형태소가 속한 문장을 얻어내는 SQL 질의문을 작성함으로써 다양한 방법으로 용례를 조회할 수 있다. 또한, 추가적인 조건을 주어 용례를 한정 지을 수 있다. 예를 들어 '고유 명사[8] 서울'에 대한 '문어'에서의 용례는 그림 3의 SQL질의를 통해서 찾을 수 있다. 제4절에서 정의된 각 테이블이 가진 속성을 이용하면 출처, 주제 등에 대한 다양한 조건을 부여할 수 있다.

8) 형태소 범주에 대한 자세한 내용은 http://sejong.or.kr에서 확인할 수 있다.

```
SELECT DISTINCT s.sentence
FROM    File f JOIN Sentence s ON
          f.file_id = s.file_id
        JOIN Morpheme m ON
          s.file_id = m.file_id AND
          s.part_id = m.file_id AND
          s.sentence_id = m.sentence_id
WHERE   m.morpheme = '서울' AND
          m.pos = 'NNP'AND
          f.type = '문어'
```

그림 3 문어에서의 용례를 찾기 위한 SQL

5.2 통계 정보 추출

말뭉치에서는 다양한 통계 정보가 조회되는데, 품사와 형태소 수준에서의 통계 정보가 많이 조회된다. 이는 SQL의 집계 함수를 이용하여 쉽게 조회될 수 있다. 예를 들어 '구어에서의 사용자 직업군별 각 품사의 사용 빈도수'는 그림 4와 같은 SQL문을 이용하여 얻을 수 있다.

```
SELECT pr.job, m.pos, count(*)
FROM    File f JOIN Part p ON
          f.file_id = p.file_id
        JOIN Person pr ON
          p.person_id = pr.person_id
        JOIN Sentence s ON
          p.file_id = s.file_id
          p.part_id = s.part_id
        JOIN Morpheme m ON
          s.file_id = m.file_id AND
          s.part_id = m.part_id AND
          s.sentence_id = m.sentence_id
WHERE   f.type = '구어'
GROUP BY pr.job, m.pos
```

그림 4 다양한 조건에 의한 통계 정보를 위한 SQL

5.3 연어 검색

연어는 빈번하게 사용되는 어휘간의 조합을 찾아내거나 사용 패턴을 분석하기 위해서 조회되기 때문에 필수적으로 빈도 정보를 함께 추출한다. 간단한 예로, '서울에 대한 굴절형'은 서울과 함께 사용된 형태소 별 빈도수를 이용하여 빈번하게 사용된 어휘를 조회함으로써 분석할 수 있다. 그림 5는 이를 위한 SQL을 보여주는데, 여기서 'm1.morpheme_id + 1 = m2.morpheme_id' 부분은 연어가 형태소 수준에서 연속됨을 나타낸다. 이에 대한 조건을 변화 시킴으로써 다양한 수준에서의 연어를 검색할 수 있다. 그림 6은 '연결어미 어'와 '보조동사 두'가 연속되어 사용되는 '-어 두~'문형에서 사용되는 용언의 빈도를 조회하는 SQL을 보여준다.

```
SELECT  m1.morpheme,  m1.pos,  m2.morpheme,
m2.pos,
        count(*) cnt
FROM    Morpheme m1
        JOIN Morpheme m2 ON
          m1.file_id = m2.file_id AND
          m1.part_id = m2.part_id AND
          m1.sentence_id = m2.sentence_id AND
          m1.word_id = m2.word_id AND
          m1.morpheme_id + 1 = m2.morpheme_id
GROUP  BY m1.morpheme,  m1.pos,  m2.morpheme,
m2.pos
ORDER BY cnt DESC
```

그림 5 서울에 대한 굴절형을 조회하는 SQL

```
SELECT m1.morpheme, m1.pos, w.word, count(*) cnt
FROM    Morpheme m1
        JOIN Morpheme m2 ON
          m1.file_id = m2.file_id AND
          m1.part_id = m2.part_id AND
          m1.sentence_id = m2.sentence_id AND
          m1.word_id = m2.word_id AND
          m1.morpheme_id + 1 = m2.morpheme_id = 2
        JOIN Morpheme m3 ON
          m2.file_id = m3.file_id AND
          m2.part_id = m3.part_id AND
          m2.sentence_id = m3.sentence_id AND
          m2.word_id + 1 = m3.word_id
        JOIN Word w ON
          w.file_id = m1.file_id AND
          w.part_id = m1.part_id AND
          w.sentence_id = m1.sentence_id AND
          w.word_id = m1.word_id
WHERE   m2.morp = '어' AND m2.pos = 'EC' AND
          m3.morp = '두' AND m3.pos = 'VX'
GROUP BY m1.morpheme, m1.pos, w.word
ORDER BY cnt DESC
```

그림 6 '-어 두~'에 대한 연어를 조회하는 SQL

6. 꼬꼬마 세종 말뭉치 활용 시스템

SQL을 작성하는 것이 프로그램을 직접 작성하는 것보다는 훨씬 단순하고 효율적이긴 하지만, 언어학자와 같이 데이터베이스에 대한 이해가 높지 않은 사용자에게는 여전히 쉬운 일이 아니다. 이를 해결하기 위해 제 4절에서 제시한 기능을 포함하여 다양한 정보를 추출하기 위한 SQL을 정형화 하여 직관적인 인터페이스에서 조회할 수 있도록 '꼬꼬마 세종 말뭉치 활용 시스템[9]'을 구현하였다.

활용 시스템은 많은 사용자가 쉽게 말뭉치를 활용할 수 있도록 웹 응용으로 구현되었다. 먼저, 별도의 처리 모듈을 구현하여 정의된 관계형 데이터베이스에 세종 말뭉치를 일괄적으로 적재하였고, 정형화된 폼에 조건을 입력하도록 하여 다양한 기능을 지원하는 조회 프로그램을 구현하였다. 조회 프로그램을 통해 세종 말뭉치의 다양한 통계 정보를 직관적인 표 형태로 조회할 수 있으며, 형태소 수준에서의 용례, 굴절형 및 연어 정보를 조회할 수 있다. 이외에도 구문 분석 결과나 연어에 대한 시각화 기능 등 다양한 기능이 포함되어 있다.

그림 7은 서울에 대한 굴절형을 꼬꼬마에서 조회한 모습을 보여준다. 말뭉치 검색 화면에서 입력 폼에 '서울'이라는 형태소를 입력하고 조회하면 해당 표현형을

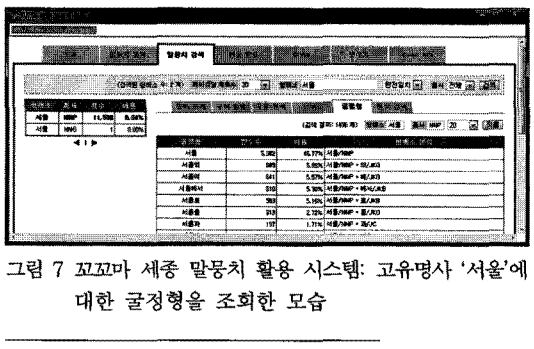

그림 7 꼬꼬마 세종 말뭉치 활용 시스템: 고유명사 '서울'에 대한 굴정형을 조회한 모습

9) http://kkma-sc.snu.ac.kr

가지는 형태소가 나열된다. 여기서 '고유명사 서울(서울/NNP)'을 선택하면 이에 대한 문어 및 구어에서의 용례와 굴절형을 조회할 수 있다. 서울은 지명을 나타내는 말이기 때문에 '에', '에서', '로'와 같은 부사격 조사와 많이 사용된다는 것을 알 수 있다.

7. 결론 및 향후 연구

본 논문에서는 말뭉치를 관계형 데이터베이스에 저장하고 이를 활용하는 시스템의 설계 및 구현에 대해서 기술하였다. 말뭉치를 관계형 데이터베이스에 저장함으로써 많은 사람들이 다양한 방법으로 말뭉치를 활용할 수 있도록 하였으며, 실제로 본 시스템은 현재 몇몇 언어학 연구자들에 의해 사용되어 실용성과 효율성을 검증 받고 있으며, 지속적인 의견 교환을 통해 기능을 개선하고 있다. 일부는 데이터베이스에 접속하여 직접 작성한 SQL을 이용해 복잡한 통계 수치들도 조회하고 있다.

몇몇 언어학자의 경우 시스템에서 제공하지 않는 정보를 조회하고자 하는 요구가 있었는데, 단순한 폼으로는 정형화 하기 어려운 것이었다. 구현된 조회 프로그램은 쉽게 SQL로 정형화할 수 있는 일부 기능만을 제공하고 있지만, 이는 더욱 확장될 수 있다. 현재, 어휘 패턴을 간단한 언어로 표현하고 이에 대한 용례를 조회할 수 있도록 하는 기능을 구현하고 있으며, 원형(Prototype)이 공개되어 있다. 추후 집계 기능을 사용할 수 있도록 패턴을 확장하여 SQL을 직접 사용하지 않아도 다양한 어휘 정보를 조회할 수 있도록 기능을 확장하고자 한다. 이 같은 패턴은 말뭉치를 활용하는 많은 연구자들에게 도움이 될 것이라 생각한다.

참 고 문 헌

[1] 서상규, 한영균, "국어정보학 입문", 태학사, 1999.

[2] Taek Chan Kang and Yoon Chul Choy, "The Design and Implementation of Tools For Dictionary Editing and Retrieval by Multi - users," *In KISS 1990*, vol.18, no.2, pp.825-828, 1990.

[3] 김경서, 김대철, 정강석, 송만석, "말뭉치를 이용한 형태소 분석 단계에서의 중의성 해결에 관한 연구", *제3 회 한글 및 한국어정보처리 학술발표논문집*, pp.36-43, 1991.

[4] Changdeok Lee, Kyeongseo Kim and Mansuk Song, "An Implementation of Concordance System in Large Corpus," *In KISS 1994*, vol.21, no.1, pp. 825-828, 1994.

[5] Jonghun Shin, Soonho Kwon and Hyuk-Chul Kwon, "Implementation of Web-based Information Retrieval System for Korean-English Parallel Corpus," *In KCC 2009*, vol.36, no.1A, pp.33-34, 2009.

[6] Beom-mo Kang, Building Corpora and Making Use of Frequency (Statistics) for Linguistic Descriptions, Journal of Korealex, no.12, pp.7-40, 2008.