‘신소설 어휘사전 편찬’ 연구 개요

연구과제명

신소설 어휘 사전 편찬 연구
Compilation of Word Dictionary of the Korean New Novels

연구진

김병선(한국학중앙연구원 교수, 연구책임자)
홍윤표(전 연세대학교 교수, 공동연구원)
송하춘(고려대학교 명예교수, 공동연구원)
김영민(연세대학교 교수, 공동연구원)
이건식(단국대학교 교수, 공동연구원)
장노현(한국학중앙연구원 전임연구원, 공동연구원)

연구 기간: 2009년 4월 ~ 2011년 11월 (3년)

연구목적

이 연구는 20세기 초반에 간행된 한국 신소설 작품의 코퍼스(linguistic corpus)를 구축하여, 연구 대상 신소설에 사용된 모든 어휘(관용구, 속담 포함)를 대상으로 어휘 사전을 제작하는 것을 목적으로 하며, 3년에 걸쳐 수행하였다.

연구내용

59편의 신소설을 연구 대상 자료로 확정, 선본을 확보하여 세그멘테이션 기법을 통해 입력하였다. 입력 자료에 대해서 기본적으로 띄어쓰기 교정 처리를 하였고, 용례사전에서 제시할 문맥의 구분 처리를 하였다. 이렇게 만들어진 입력 텍스트 파일을 어절별 용례색인(concordance)으로 가공하고, 다양한 기본형 분석 방법(수작업 포함)을 활용하여 기본형을 확정하였다. 이후 어휘사전 구성형식에 맞추어 어휘별로 문법사항, 뜻풀이, 용례 등의 어휘정보를 집필하거나 가공하였다.

연구성과

총 59편의 연구 대상 자료로부터 90만 여 개의 어휘를 추출하여, 각 어절에 대한 정규화, 기본형 확정 등의 작업을 완료하였다. 기본형 표제어 4만 여종의 어휘에 대한 뜻풀이를 완료하였다. 그 성과물은 MS Access 데이터베이스의 테이블로 수록되어 있으며, 그 목록은 다음과 같다.

1. 신소설 목록 (총 59종에 대한 서지 정보 제공)
2. 신소설 용례색인 (총 90만 어절에 이르는 어휘 용례색인)
3. 신소설 어휘사전 (기본형 어휘 4만여 종에 대한 어휘 풀이)

가. 연구 목적 및 배경

1) 연구의 배경

신소설은 고대소설과 근현대 소설의 중간 시기에 자리한 문학 장르로서, 개화기의 과도기적 시대상을 아주 잘 반영하고 있는 언어자료다. 신소설은 개화기의 여러 문장 자료 가운데서 특히 구어체 자료의 성격을 가지고 있다. 이는 관부 자료, 신문 자료, 문법서 자료, 국어사전 자료, 언해 자료, 여타 신문학 자료, 교과서 자료, 종교 자료, 학술 자료 등의 문어체 자료와 구분된다.

아울러서 신소설은 근대국어로부터 현대국어로 이행하는 과정의 자료로서, 현대의 어문 규범이 확정되기 이전의 다양한 혹은 혼란된 표기법에 의한 각종 어휘를 풍부하게 담고 있는 어휘 자료의 보고이기도 하다. 신소설은 근대적 인쇄 매체를 통해 출판 보급되었으며, 당대의 독자들에게 환영을 받은 문학 장르이기도 했다. 따라서 신소설의 어법은 우리 현대국어의 어법과 근대소설문학의 어법 및 문체 형성에 많은 영향을 미쳤다고 생각된다.

따라서 이 연구에서는 충분한 분량의 신소설 텍스트를 확보하고, 신소설 자료의 원전 확정을 통하여 품질이 높은 코퍼스를 구축하며, 코퍼스에 대한 분석·연구 및 전산 처리를 통해서 어휘 사전을 편찬함으로써, 20세기 초반의 우리말에 대한 국어학적 연구와 문학적 연구에 기여하려는 것이다.

2) 연구의 목표

- 신소설 작품의 목록을 작성한다. (제목, 판에 관한 정보를 포함한다.)
- 창작 신소설을 중심으로 하고 번안 신소설 중 당대 대표적인 신소설 작품을 대상으로 선정한다.
- 연구대상 신소설의 원전비평을 통해 텍스트를 확정한다.
- 일정한 양식으로 신소설 작품의 코퍼스(corpus)를 작성한다.
- 신소설 텍스트에 대한 정규화, 정상화(normalization) 처리를 한다.
- 신소설 어휘에 대한 기본형을 분석하고(lemmatization), 동음이의어 분석을 한다.
- 어휘 항목을 표제어로 하는 용례사전(concordance)을 제작한다.
- 사전 표제 어휘에 대하여, 관련어를 제시하고, 그 뜻풀이를 서술한다.
- 통계처리를 통해 어휘 빈도를 추출한다.

나. 연구내용 및 방법

1) 연구 내용

가) 연구 대상 자료의 확정

신소설은 수백 종이 간행되었다. 인기 있는 작품은 판을 거듭하여 발행되기도 했다. 신소설을 아직 장르 면에서 안정되지는 않은 상태였다. 창작신소설도 있고, 번안 혹은 번역소설과 개작신소설도 있고, 소위 '신작 구소설'로 불리는 작품도 있다. 이번 연구에서는 당대의 어휘사전 편찬을 목표로 하고 있기 때문에 번안소설 중에서 당대에 많이 유통된 작품에 한정하여 일부만 수록하기로 한다.

대상 자료 선별의 원칙

첫째, 1900년대부터 1910년대 말까지 발표된 단행본 소설을 대상으로 한다.

둘째, 한문소설 및 국한문소설은 제외하고 한글소설만을 대상으로 한다.

셋째, 번역 및 번안소설은 원칙적으로 제외하고 창작소설을 대상으로 한다. (단,애국부인전 등의 경우처럼 예외를 둘 수 있다.)[1)]

이 연구에서는 먼저 당대의 신소설 목록을 최대한 확보하여 데이터베이스에 수록한다. 위의 기준에 따라 학계에서 관련학자들의 주목을 받은 작품(모두 90만 어절 규모를 맞출 경우 60편으로 예상)으로 한정하여 어휘 사전을 편찬하기로 하였다. 시기적으로는 1900년대부터 1920년 이전까지 단행본으로 출판된 신소설 작품을 선정하였으며, 번안 작품일지라도 당대에 판으로 거듭한 주요 작품은 제한적으로 포함하기로 한다. 대표작품 선정에 있어서는 각 전집 수록 작품 중에서 수록 빈도가 높은 작품을 우선적으로 고려하며, 신소설 연구자와 국어학 연구자로 구성된 연구진의 회의를 통해서 대표작을 선정한다.

나) 대상 판본의 선정

- 각 작품의 초간본(초판본)을 택한다. 초간본이 없는 경우는 남아있는 판본 중에서 가장 앞서는 것으로 하였다. 이때 신문연재 후 단행본으로 출간된 작품들일지라도 신문연재본은 작업 대상에서 제외한다.
- 초간본일지라도 그 상태가 이미지 기반 입력 및 교정에 적합하지 않을 정도로 좋지 않거나 낙장이 있는 경우는 그 다음 간행본을 택한다.
- 아세아본과 계명본 등의 영인본 자료집과 국중본(국립중앙도서관 소장본)을 비교하여 보다 선명한 품질을 가진 판본을 선택한다. 대개 컬러이미지로 되어 있는 국중본은 흑백 이미지의 영인본 자료집을 스캔한 경우보다 이미지 인식율이 약간 높았기 때문에 가능하면 국중본을 활용한다.

다) 신소설 작품의 코퍼스 구축

어휘사전을 만들기 위해서 정보처리의 첫 단계로서 신소설 텍스트를 컴퓨터에 입력한다. 기술정보활동비를 활용하여 전문적으로 근세 자료의 입력을 담당할 수 있는 인력 조직에 위탁한다.

입력 자료의 품질을 높이기 위해서 오류율은 낮출 수 있는 몇 가지 방안을 적용한다. 우선적으로 상태가 좋은 선본 신소설 판본을 확보하고, 원전 비평을 실시하며, 정보처리 기술을 활용하여 텍스트의 입력 품질을 확보한다.

특히 1, 2차년도에 성과를 거두었던 세그멘테이션(segmentation) 공법을 3차년도에도 적용한다. 세그멘테이션 공법이란 원문 이미지 파일을 대상으로 글자 단위로 구획을 한 후에 유사 자형을 일괄적으로 문자로 변환하는 방법으로서, 그 정확성은 1차년도에 검증한 바 있는데, 특히 탈자의 방지에는 가장 효율적인 방법이라고 할 수 있다.

라) 신소설 코퍼스의 정상화 및 정규화 처리

---

1) 선정 작품의 서지 사항에 대한 정리는 1차년도 학술회의 자료집 참조.(장노현, 「연구 대상 자료의 간행 상황과 선본 텍스트 선정」)

신소설 텍스트는 당대의 표기를 살려서 원본대로 입력하는 것을 원칙으로 한다. 띄어쓰기의 경우는 아주 붙여 쓴 것도 있고, 띄어쓰기를 했다고 하더라도 자료에 따라 달리 나타나고 있다. 띄어쓰기는 어휘 확정에 아주 중요한 부분이므로 현대 맞춤법에 근거하여 처리한다.

1) 입력 텍스트에 판본 형태 정보를 부가한다.
입력 텍스트에는 페이지 구획, 행 구획, 들여쓰기 정보, 편집 관련 내용 표기 등을 인쇄 양식과 관련된 사항을 태깅한다. 이 자료는 어휘사전에서 문맥정보를 제시하는 데 활용된다. 아울러서 이러한 처리를 위해서 1,2,3차년도 대상 작품을 모두 하나의 파일로 묶어서 일관되게 처리한다.
- 페이지 구획 정보의 표기(작품명_쪽번호): <page id="강상루_000"/>
- 제목 정보의 표기: <t1>江上淚(강상루)</t1>
- 들여쓰기 정보: <c> ~ </c>(첫 줄부터 시작한 단락), <c1> ~ </c1>(한 줄 내려쓴 단락)
- 행 구분 표지: </br>

2) 입력 텍스트에 대하여 띄어쓰기를 적용한다.
신소설 코퍼스는 당대의 표기를 살려서 원본대로 입력하는 것을 원칙으로 한다. 신소설 텍스트는 아예 일체의 띄어쓰기가 되어 있지 않은 경우도 있고, 띄어쓰기를 했다고 하더라도 자료에 따라 달리 나타나고 있다. 띄어쓰기는 어휘 확정에 아주 중요한 부분이므로 현대 맞춤법에 근거하여 처리하기로 한다. 또한 여러 연구자가 참여하기 때문에 띄어쓰기의 일관성을 유지하기 위해서 일관성 검사를 실시하여 오류를 최소화한다.

3) 입력 텍스트에 대하여 문맥 구분 정보를 부가한다.
외부 업체에서 수행한 텍스트 입력과 교정 단계에서 교정자가 임의적으로 문맥 구분 처리를 하기로 한다. 이 자료를 모두 받아서 본 연구진에서 보다 세밀하게 문맥 단위 구분 처리를 실시한다. 문장 혹은 절 단위로 문맥을 구분하되, 하나의 문맥이 최대 55바이트(공백 포함)를 넘지 않도록 조정한다. 문맥 구분 정보는 “#” 표시로 처리한다.

4) 입력 텍스트에 대한 교정, 교열 및 정상화 처리를 한다.
정보처리 기술을 활용하여 텍스트의 입력 품질을 확보하기 위해 1차년도에 공동연구원이 개발한 띄어쓰기 일관성 검사기를 지능형으로 개조하여, 띄어쓰기 검증을 하고, 데이터베이스 상에서도 쿼리를 통해 띄어쓰기를 지속적으로 검사한다. 문장부호의 경우는 원칙을 정해서 표준화하고, 대화자 표시는 ‘[ ]’를 일괄적으로 적용하기로 한다.

2) 연구 방법

가) 표제어의 추출과 정보 처리
전산 입력된 ‘신소설2011 코퍼스’는 용례사전 전단계인 KWIC 형식의 색인으로 가공한다. 유니코드 텍스트 파일로 저장한 신소설2011 텍스트 파일에 다음과 같은 정보만 남기고, 색인용 파일로 저장한다.
* 위치 정보 (작품명과 페이지 정보)
* 문맥 구분 정보

나) 용례색인으로 가공하고 테이블로 저장

용례색인 제작 도구는 여러 가지가 있으나, 이 연구에서는 연구책임자가 개발한 “똑똑새 UniKwic 1.0”를 사용한다. 이 프로그램은 키워드, 지정한 길이의 앞뒤 문맥정보, 위치 정보 및 기타 정보를 탭(tab)으로 구분하여 텍스트로 저장하는 프로그램이다.

다) 표제어의 정리

기본적으로 위 절차를 통해서 각 어절은 하나의 표제어가 되어 문맥 정보 및 위치 정보와 더불어 하나의 레코드를 형성하게 된다.

각 어절에는 여러 가지 문장부호가 붙어 있으므로 일단 이를 정리하여 텍스트로만 구성된 어절로 변환 처리한다. 이때는 Access의 찾아 바꾸기나 정렬 기능을 적극적으로 활용하여 처리한다.

이 레코드 중에는 작품의 제목이나 주석문도 포함되어 있고, 문장부호로만 구성된 것들(특히 말없음표)도 있어서 이를 불용어로 처리하기로 한다.

라) 품사기호 적용

2010년 이전에는 품사기호를 영문 알파벳 1자~2자로 되어 있었으나, 이제는 2자로 통일하였고, 실제 코퍼스에서는 소문자로 표기한다. 그 중 첫 글자는 대분류를 반영한다. 다만 보조동사, 보조형용사는 이전과 마찬가지로 따로 구분한다.

마) 기본형 표제어 추출

위에서 정리된 활용형 표제어를 바탕으로 기본형 표제어를 추출한다. 현대국어 텍스트를 대상으로 하는 여러 종류의 형태소 분석기가 개발되어 있으나, 신소설 텍스트의 특성(옛한글이 사용되었다는 점과 오표기 및 이표기가 많다는 점) 때문에, 형태소 분석기를 통한 자동 추출이 어렵다는 판단을 하여, 2차년도부터는 별도의 방식을 강구하여 처리하였다. 그것은 기분석(활용형->기본형) 목록 사전을 적용하여 기본형 후보어들을 제시하는 방식이다.

바) 동음이의어 분석

위와 같은 방식을 적용할 경우, 기본형 추출은 물론이고, 동음이의어 분석 및 품사 분석까지 동반하고, 어휘의 품사 분석도 행하고 있으며, 일부 어휘에 대해서는 다의어 분석까지 한다. (두 개 이상의 품사로 분석되는 어휘에 대하여)

동음이의어 분석은 기본적으로『표준국어대사전』의 분류를 채택하여 첨자(superscript)를 부여하는 작업이다. 이 과정에는 한국학중앙연구원의『현대시어사전』제작 과정에서 작성한 MS Access Query를 적극 활용함으로써 그 효율성과 정확성을 높인다. 이미 구축된 활용형 어휘의 분석 사전을 활용하여 신규 어절을 분석하는 방식을 응용한다.

2010년 연구에서는, 코퍼스의 규모가 크고, 다양한 작가들의 작품으로 구성된 2차년도 신소설 코퍼스에 대해 기본형 추출과 확정 작업을 진행하였고, 그 결과를 1차년도 신소설 코퍼스에 일부 적용한 바 있다. 3차년도(2011)에는 형태소 분석과 기본형 추출의 무결성을 높이는 일에 주력한다.

사) 연어 정보의 처리

아울러서 연어 정보(2-gram) 통계 처리를 통해서 기본형 적용에 대한 검증도 실시할 계획이다. 기존 분석 목록 정보를 활용할 경우, 비교 대상의 범위를 넓히면 그만큼 애매성도 준다. 그러나 비교 대상을 3개 이상으로 늘리게 되면 효율성이 매우 떨어지게 되므로, 2-gram 활용형-기본형 목록을 제작하여, 기분석 자료의 정확성을 검증한다.

아) 자료 분석 및 처리

구축된 코퍼스(원시 말뭉치)는 기본적으로 용례 추출 절차를 거쳐서 MS Access DB로 수록된다. 수록된 자료는 MS Access의 SQL의 검증 쿼리 절차를 통해서 그 일관성과 정확성을 검증하게 된다. 이 과정을 반복하여 최고 품질의 용례 Data를 생성한다.

자) 용례색인으로 가공하고 테이블로 저장

연구책임자가 개발한 UniKwic 1.0 프로그램을 이용하여, 입력 텍스트 파일을 어절별 용례색인(concordance)으로 가공하였다. 용례색인 텍스트 파일을 MS Access의 테이블로 수입(import)하였다. 이후 모든 자료 처리는 이 테이블을 바탕으로 처리하였다.

차) 기본형 표제어의 확정

기본형 표제어 확정 작업은 이 연구사업에 있어서 가장 핵심적인 부분이고, 가장 많은 노력이 필요한 부분이다. 2010년도에 1차 활용한 바 있는 기분석 목록의 적용을 통한 기본형 분석 방법을 적극적으로 활용하였고, 수작업에 의해 일일이 재검토하고, 추가하여 기본형을 확정하였다. 기본형 표제어 확정에 있어서는 다음과 같은 원칙을 적용하였다.

- 동음이의어 분석: [표준국어대사전]의 동음이의어 분석을 바탕으로 첨자를 부여하였다.
- 다의어 분석: 일부 중요 어휘에 대해서는 위 사전에 근거하여 다의어 처리를 하였다.
- 품사 분석 및 품사기호 적용: 품사를 분석하고 그 기호를 부착하였다.
- 원어 정보 부기: 한자어는 ( ) 안에 한자를, 기타 외래어도 원어 정보를 부기하였다.
- 의미 참고 정보 부기: 동음이의어가 있는 일부 고유어 어휘에 [ ] 안에 한자로 의미 분석의 참고자료를 밝혔다.
- 관련어 정보 부기: 오류 표기, 방언, 고어 및 기타 항목에 대해 정표기, 표준어, 현대어 정보를 부기하였다.

한편 기분석 목록 적용 방식의 적용에 있어서는 활용형-기본형 목록도 적용하였지만, 2gram 목록, 즉 연어 정보를 활용하여 자동 분석의 품질을 높였고, 자동 분석의 효율성을 높였다.