Journal of the Korean Association for Science Education
Journal homepage: www.koreascience.org

교육용 과학언어 연구를 위한 범용 자료로서 과학교과서 말뭉치 K-STeC(Korean Science Textbook Corpus) 구축

윤은정[1], 김진호[2], 남길임[1], 송현주[3], 옥철영[4], 최준[1], 박윤배[1*]
[1]경북대학교, [2]강원대학교, [3]계명대학교, [4]울산대학교

Building Korean Science Textbook Corpus (K-STeC) for research of Scientific Language in Education

Eunjeong Yun[1], Jinho Kim[2], Kilim Nam[1], Hyunju Song[3], Cheolyoung Ok[4], Jun Choi[1], Yunebae Park[1*]
[1]Kyungpook National University, [2]Kangwon National University, [3]Keimyung University, [4]University of Ulsan

ARTICLE INFO

Article history:
Received 8 June 2018
Received in revised form
21 June 2018
12 July 2018
Accepted 20 July 2018
Keywords:
science textbook, corpus, science textbook corpus, scientific language, scientific terminology

ABSTRACT

In this study, the texts of science textbooks of the past 20 years were collected in order to systematically carry out researches on scientific languages and scientific terms that have not been noticed in science education. We have collected all the science textbooks from elementary school to high school in the 6th curriculum, the 7th curriculum, and the 2009 revised curriculum, and constructed a corpus comprising of 132 textbooks in total. Sequentially, a raw corpus, a morphological annotated corpus, and a semantic annotated corpus of science terms, were constructed.
The final constructed science textbook corpus was named K-STeC (Korean Science Textbook Corpus). K-STeC is a semantic annotated corpus with semantic classification and classification of scientific terms, together with meta information of bibliographic information such as curriculum, subject, grade, and publisher, location information such as chapter, section, lesson, page, and sentence, and structure information such as main, inquiry activities, reference materials, and titles. Throughout the three-year study period, a new research method was created by integrating the know-how of the three fields of linguistic informatics, computer science and science education, and a large number of experts were put in to produce labor-intensive results. This paper introduces new research methodologies and outcomes by looking at the whole research process and methods, and discusses the possibility of future development of scientific language research and how to use the results.

Ⅰ. 서론

1. 과학언어 연구 및 말뭉치 구축의 필요성

최근 발표된 TIMSS 2015의 결과에 의하면 우리나라 학생들의 과학 성취도가 TIMSS 2011에 비해 한 등급 떨어졌고 과학에 대한 흥미는 여전히 최하위권에 머무르고 있다(Martin *et al*., 2016). 우리나라 학생들이 국제 학업 성취도 평가에서 과학 성취도는 최상위권임에도 불구하고 과학에 대한 흥미가 매우 낮다는 고질적인 문제점을 개선하기 위하여 최근 흥미 위주의 수업을 강조해 왔으나, 이것이 과학에 대한 흥미는 높이지 못하고 오히려 성취도마저 떨어뜨리는 역효과를 가져온 것이 아니냐는 우려의 목소리도 있다. 우리나라 학생들은 다른 나라 학생들에 비해 과학에 대한 자신감과 흥미가 매우 낮다(Kwak, 2017). 과학에 대한 흥미가 낮은 이유로 내용이 어렵다고 하는 점을 고려하면(Kwak *et al*., 2006), 결국 학생들이 과학을 어려워하는 원인을 규명하고 이를 개선하는 것이 중요한 당면 과제로 보인다.

여러 학자들은 학생들이 과학을 어려워하는 주된 원인으로 언어적 문제를 꼽는다(Fang, 2006; Ford & Peat, 1988; Jaipal, 2001; Shaw, 2002; Yore *et al*., 2004). 과학에서 사용하는 언어는 과학적 사고방식과 지식의 체계를 반영하여 특수하게 발달해 오는 과정을 통해 (Darian, 2003) 일상의 언어와는 다른 특수한 성격을 가지고 있다 (Jaipal, 2001; Reeves, 2005; Shaw, 2002). 과학 교육에서 언어는 단순한 전달 수단이 아니며 과학언어를 배우고 이해하는 것이 곧 과학적 사고와 지식 체계의 습득으로 이어지게 된다(Maskill, 1988). 그러나 과학언어가 가지는 이러한 특수성은 학생들로 하여금 과학언어를 이질적으로 느끼게 하고 어렵게 느끼는 요인으로 작용하게 된다(Miller, 2009). 따라서 과학 교수학습 상황에서 과학언어는 매우 주의 깊게 다루어져야 하며, 학생들로 하여금 과학언어를 체계적으로 익히고 친숙해지도록 할 필요가 있다(Shaw, 2002). 한편, 과학용어의 문제 역시 학생들이 과학을 어려워하게 되는 주된 요인으로 꼽힌다(Fang, 2006; Ham *et al*., 2011; Merzyn, 1987; Nam, 2008; Wellington & Osborne, 2001). Yun & Park(2013a)을 비롯한 몇몇 조사에서는 중고등학생의 85% 전후가 과학용어를 어려워하고 있고, 95%의 과학 교사가 과학용어가 중요하다는 점은 인식하지만 어떻게 가르쳐야 할지 모르겠다며 용어 교수의 어려움을 겪고 있다고 보고한 바 있다. 어휘

* 교신저자 : 박윤배 (ypark@knu.ac.kr)
** 이 논문 또는 저서는 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2014S1A5B4038405)
http://dx.doi.org/10.14697/jkase.2018.38.4.575

라는 것이 언어를 구성하는 하나의 요소이므로 사실상 과학용어는 과학언어에 포함되는 하나의 요소이다. 그러나 언어 습득에서 어휘가 차지하는 비중이 매우 높아 어휘 교육이 독자적인 관심을 받는 것처럼 과학언어의 문제를 다룰 때 역시 과학용어는 보다 심도있게 다루어야 할 이슈라 여겨진다.

그런데 지금까지 과학언어에 대한 연구는 다소 소외되어 왔다. 과학언어 교육의 중요성과 필요성에도 불구하고 국내 과학교육 분야에서 주목받지 못하는 이유는 교육과정이나 교육내용, 탐구 등에 비해 주변적인 영역으로 인식되거나, 언어학의 영역이라는 인식 때문인 듯하다. 또 다른 이유로, 과학용어 연구의 어려움을 들 수 있는데, 과학용어 자체가 가지는 개념이 대부분 과학, 철학, 언어학, 교육학의 제 분야와 복잡하게 얽혀 있어서 그 본질을 규명하고 명확하게 정의하거나 해석, 구분하는 것이 쉽지 않기 때문이다. 이러한 복합적인 요인으로 인하여, 과학언어에 대한 연구는 체계적으로 축적되기 어려웠고 과학용어와 관련된 여러 가지 문제점들은 개선되지 못하고 있다. 일례로 과학교과서에 노출된 문장의 수준 및 과학용어의 양과 범위가 객관적인 기준에 의하기 보다는 대부분 과학교육 전문가들의 직관에 의존되어 왔고, 이는 학생들이 과학교과서의 텍스트를 읽고 이해하는 데 어려움을 겪는 문제점으로 이어졌다(Park *et al*., 2015).

과학언어와 관련된 교육 현장의 여러 가지 문제점들을 개선하고, 과학언어 교육을 효과적으로 실시하기 위해서는 과학언어에 대한 연구가 심도있게 이루어져야 하는데, 과학언어를 연구하기 위해서는 연구 대상이 되는 자료들을 수집, 정리하는 것이 선행되어야 한다. 학생들에 대한 연구를 하기 위해서는 학생들을 직접 대면하거나 연구 목적에 맞게 설계된 설문지나 검사지 등을 도구로 사용하여 학생들로부터 여러 가지 정보를 추출한 뒤, 그것들을 한 자리에 모아서 연구 목적에 맞게 다각도로 분석하고 결론을 도출하게 된다. 같은 맥락에서 과학언어를 연구하기 위해서는 과학언어 자료로부터 정보를 추출해야 하는데, 언어학 분야에서는 언어 연구를 위해 언어 자료들을 한 자리에 모아두고 다양한 정보 추출 및 분석이 용이하도록 하는 방법을 사용해오고 있다. 이 때 모은 언어 자료의 덩어리를 말뭉치(코퍼스)라고 부르며, 말뭉치를 구축하거나 혹은 말뭉치를 바탕으로 언어학적 연구를 수행하는 연구 방법을 말뭉치 언어학이라 부른다.

본 연구에서는 그 중요성과 필요성에도 불구하고 지금까지 잘 이루어지지 못했던 과학언어에 대한 연구가 체계적으로 축적될 수 있도록 말뭉치 언어학의 방법을 도입하여 지금껏 구축된 적 없는 과학언어 말뭉치를 구축하고자 하였다. 본 연구를 통해 구축되는 말뭉치는 초중등 과학 교육용 과학언어 연구라는 분명하고 구체적인 목적을 가지고 출발하지만, 그 활용 가능성은 현 시점에서 고려되는 것들 뿐만 아니라 고려되지 않는 부분들 까지도 다양하게 확장될 수 있을 것으로 기대한다. 언어학 분야에서 말뭉치는 문법, 의미, 어휘 등 다양한 언어 연구뿐만 아니라 사전 편찬, 언어 교육용 자료 개발 등 실제적 결과물 산출을 가능하게 하고, 자연언어처리나 번역, 미래 사회의 핵심 기술인 인공지능의 개발에도 결정적인 자료로 사용되고 있다(Kang, 2011). 과학언어 말뭉치 역시 과학언어와 일상적 언어와의 문법적, 어휘적 차이를 찾아내고, 말뭉치로부터 획득한 용어 목록 및 용례, 유의어 추출 등을 이용하여 과학용어 사전 편찬이 가능하다. 또한 한국어 범용 말뭉치에서 소외되었던 전문 분야의 말뭉치 확충으로 과학 교육 맥락에서의 인공지능 기술의 질적 향상에 자료로 활용될 수 있다.

말뭉치 구축에 대한 기본적인 원리나 절차는 언어학 분야에서 이미 많은 연구가 이루어져 있으므로 도입하면 될 것이나 대상 자료의 선정이나 분량, 말뭉치의 구조나 설계 등의 세부적인 사항은 개별 말뭉치의 특성 및 연구 목적에 따라 이루어져야 한다. 이에 본 연구에서는 교육용 과학언어 말뭉치를 구축함에 있어 대상 자료의 범위와 분량 설정에서부터 말뭉치 설계와 실제 구축에 이르기까지의 과정을 소개하고 구축된 말뭉치의 활용 방안과 향후 관리에 대한 논의를 간략하게 제시하고자 한다.

2. 말뭉치 언어학의 이론적 배경

말뭉치라는 개념은 언어학 분야에서는 이미 하나의 학문 분야로 자리 잡은 만큼 그 의미를 설명하는 것이 무색한 일일 것이다. 그러나 본 연구가 소개될 과학교육 분야에서는 생소할 수 있으므로 간략하게 말뭉치의 개념과 역사, 원리 등에 대해서 소개하고자 한다. 말뭉치란 모집단의 언어를 대표하는 샘플을 모아 놓은 기계가독형 텍스트로, 그 어원은 라틴어의 몸(‘corpus’), 즉 ‘말의 몸뚱이 그 자체’라는 뜻이다. 일반적으로 언어학 연구에서는 주요 말뭉치 유형을 문어 말뭉치(written corpus)와 구어 말뭉치(spoken corpus)로 구분하며, 말뭉치 자체에 품사 등 문법 표지를 부착하느냐의 여부에 따라 원시 말뭉치(raw corpus)와 주석 말뭉치(annotated corpus) 등으로 구분한다.

모집단의 언어를 대표하는 말뭉치의 정의에는 기본적으로 ‘대표성’과 ‘균형성’의 원리가 포함된다. ‘대표성’이란 수집된 자료가 연구 대상 모집단의 언어 특성을 대표할 수 있어야 한다는 것이고, ‘균형성(balance)’은 말뭉치의 실제 구성에서 개별 텍스트들이 특정 주제나 장르에 편중되지 않고 균형있게 수집되어야 한다는 뜻이다.

말뭉치의 발달 과정을 살펴보면, 1963년 미국의 브라운 말뭉치를 시작으로 컴퓨터를 활용한 말뭉치 구축이 본격화되었으며, 1990년 이후 BNC(British National Corpus), ICLE(International Corpus of Learner English), 연세한국어말뭉치, 국립국어원의 세종말뭉치와 학습자말뭉치, Trend 21 코퍼스 등과 같이 본격적인 말뭉치 구축이 시작되면서, 말뭉치는 언어 교육, 사전 편찬, 어휘 사용 빈도 조사 등 다양한 방면에서 널리 활용되어 왔다. 초기 말뭉치는 주로 언어학 분야에서 사전 편찬, 외국어교육에 많이 활용되어 왔지만, 최근 컴퓨터와 IT 기술의 발달로 말뭉치는 자동번역, 문서요약, 기계학습 등 자연언어처리와 정보과학 분야에도 주된 자원이 되고 있다(McEnery & Hardie, 2012).

말뭉치 언어학이 갖는 가장 큰 특징은 인간의 직관만으로는 결코 알 수 없는 정보를 제공한다는 것에 있다. 예를 들어 과학교과서 ‘빛과 파동’ 단원에서 가장 많이 사용되는 과학용어는 무엇인지, 과학 용어 ‘빛’과 함께 사용되는 단어들을 빈도순으로 나열하면 어떤 단어들이 상위에 올지, 과학교과서 텍스트의 평균 문장 길이는 어느 정도인지, 단문과 복문의 비율은 어떠한지 등은 직접 하나하나 세어봐야지만 확인할 수 있으며 전문가들의 지식이나 논의를 통해서는 결코 알아낼 수 없는 정보들이다. 말뭉치 언어학의 초기에는 그 가치에 대한 의심이 종종 제기되기도 했으나, 말뭉치를 통한 계량적 연구들이 다양한

의미 있는 결과들을 도출해 내면서 인정받기 시작했다. 또한 말뭉치를 통한 계량적 연구가 언어의 의미 연구로 이어지고(Mikolov *et al*., 2013), 말뭉치를 이용한 전산처리 능력 향상 및 관련 서비스의 비약적인 발전 등으로 인해 말뭉치 규모 및 사용률이 지속적으로 확대되고 있다(Gwak, 2013).

3. 과학교과서 말뭉치가 가지는 의의

과학언어를 연구함에 있어 대표성과 균형성을 갖춘 자료가 무엇인지 그 대상과 범위를 정하는 것은 자료로부터 도출될 연구 결과물들의 타당성을 결정짓는 매우 중요한 일이다. 본 연구에서 지향하는 두 가지 큰 목표는 첫째, 학생들이 과학언어를 자유롭고 능숙하게 구사하여 과학언어로 인해 겪었던 과학 학습의 여러 가지 어려움을 해소해 주는 것과, 둘째, 과학언어에 내포되어 있는 과학적 지식의 구조와 사고 체계를 학생들에게 효과적으로 전달할 수 있는 방법을 찾는 것이다. 이 두 가지 목표를 달성하기 위해서 연구해야 할 대상은 실제 과학 학습에서 사용되는 언어이면서, 정확하고 표준적이며 규범성을 갖는 자료여야 한다. 본 연구에서는 이 두 가지 조건을 모두 갖춘 최적의 자료로서 과학교과서를 지목하였다. 과학교과서는 국가 교육과정에서 정한 교육 내용을 그대로 담고 있으며, 이를 학생들에게 전달하기 적합한 언어로 구현된 자료이다. 또한 과학교과서의 저자는 집필 당시 국내 과학 혹은 과학교육 분야 최고의 전문가들로서, 과학교과서 텍스트는 이들이 가진 지식의 구조 및 사고 체계를 반영하고 여러 차례 검정을 거친 정선된 자료이다. 따라서 과학교과서 텍스트는 본 연구의 목적에 비추어 대표성은 충분히 갖춘 것으로 볼 수 있다.

다음으로 균형성의 측면에서 살펴보면 과학교과서의 모집단은 초등학교부터 고등학교까지 학년별로 구성되어 있고, 각 학년별로 다양한 출판사에서 교과서를 출판하고 있다. 또 한 권의 교과서 내에는 물리, 생명과학, 지구과학, 화학 네 분야에 해당하는 여러 단원으로 구성되어 있다. 그리고 수 년 마다 이루어지는 교육과정 개편에 따라 새로운 과학교과서가 집필된다. 본 연구에서는 교육용 과학언어 연구의 대상이 되는 말뭉치를 구축하되 폭넓은 활용성을 고려하여 특정 분야나 특정 학년, 특정 내용에 치우치지 않고 모든 분야, 모든 학년, 모든 내용을 포괄하는 자료를 구축하고자 하였다. 따라서 이들 요인들이 균형있게 반영되어 균형성이 확보된 과학교과서 말뭉치를 구축함을 목적으로 한다.

II. 연구 방법 및 결과

본 연구의 경우 과학교과서 말뭉치를 구축함에 있어 기본적으로는 언어학의 말뭉치 구축 전략을 바탕으로 하되 결과물의 과학 교육적 활용 목적에 맞도록 새로운 연구 방법들을 개발하여 적용하는 것이 주안점이다. 따라서 과학교육 연구에서 연구 방법과 결과를 분리하여 작성하는 것이 보편적이기는 하나 본 연구의 성격과 내용을 보다 잘 드러내기 위하여 연구 방법에 대한 상세한 설명과 그에 따른 결과물을 함께 제시하는 방법으로 기술하고자 한다.

1. 말뭉치 구축 대상 자료

대상 자료 선정에 있어 분야, 학년, 단원 내용을 균형있게 포함하는 것을 지향하였다. 전체 학년, 전체 단원을 포괄하는 것은 학년별로 교과서 한 권씩 선택하면 간단하게 해결이 된다. 그러나 특정 내용에 대한 텍스트를 놓고 봤을 때 저자의 개별 특성을 최대한 배제하고 해당 내용에 대한 보편적인 언어 특성을 찾아내기 위해서는 같은 내용을 여러 명의 전문가가 작성한 텍스트를 확보해야 한다. 이를 위해서는 하나의 교육과정 내에서 여러 출판사의 교과서를 포함할 수도 있고, 여러 교육과정을 포함할 수도 있다. 우리나라 과학과 교육과정의 경우 초등학교 교과서는 국정교과서 1종류이고 중학교와 고등학교 과학 교과서는 통상적으로 8~10종의 출판사에서 교과서를 출판한다. 따라서 초등학교의 경우는 여러 출판사를 포함하는 것이 불가능하므로 불가피하게 교육과정을 걸쳐서 보아야 한다. 따라서 본 연구에서는 6차, 7차, 2009 개정 교육과정의 세 교육과정을 포함하였고, 중등 수준에서는 각 학년별로 세 교육과정 동안 지속적으로 과학 교과서를 출판해 온 두 개 정도의 출판사를 포함하여 교육과정과 출판사 모두를 어느 정도 아우를 수 있도록 하였다. 2007 개정 교육과정의 경우 전체 학년이 시행된 것이 아니었기 때문에 자료의 균형성을 확보하기가 어려워 대상 자료에서 제외하였다. 최종적으로 본 연구의 대상이 된 자료는 6차, 7차, 2009 개정 교육과정의 1, 2학년 슬기로운 생활 교과서와 3~10학년의 과학, 11~12학년의 물리, 화학, 생명과학, 지구과학 교과서 각각 두 개 출판사씩 총 132권의 교과서이다 (Table 1). 말뭉치 구축 범위는 표지, 차례, 부록, 색인을 제외하고 첫 번째 대단원이 시작되는 페이지에서부터 마지막 대단원이 끝나는 페이지까지로 정하였으며, 범위 내에서 수식을 제외한 모든 텍스트를 구축 대상으로 하였다.

2. 말뭉치 구축 절차 및 방법

말뭉치 구축 절차는 크게 자료 수집, 원시 말뭉치 구축, 형태주석 말뭉치 구축, 용어주석 말뭉치 구축의 네 단계로 이루어졌으며, 연구 기간은 2014년 9월부터 2017년 8월까지 총 3년이 소요되었다. 교과서 텍스트는 최근에 출판된 교과서를 중심으로 전자 자료 형태가 있는 것은 전자 자료로 수집하였고, 그렇지 않은 경우는 종이 책 형태의 자료로 수집하였다.

가. 원시 말뭉치 구축

원시 말뭉치 구축 단계는 다시 교과서 구조 분석, 원시말뭉치 입력 도구 개발, 원시말뭉치 입력, 검토 및 수정 단계로 세분화하여 진행하였다. 소설이나 에세이, 교양서적 등과 같은 일반 서적들은 단일한 혹은 단조로운 구조로 이루어져 있는 반면, 과학교과서 텍스트는 다양한 크기의 여러 단으로 나누어져 있거나, 본문과 탐구활동이 구분되어 있는 등 매우 복잡한 구조를 가진다. 이에 과학교육 전문가 2인과 언어정보학 전문가 2인, 총 4인이 논의를 통해 말뭉치 활용도와 효율성을 고려하여 텍스트를 구조화 하였는데, 제목, 본문, 탐구활동, 참고자료, 단원마무리의 5가지 구조로 구분하였다. 각 구조에 해당하는 내용들의 예시를 Table 2에 제시하였다.

Table 1. 과학용어 DB 구축 대상 교과서 목록

6차				7차				2009 개정			
	교과서		출판사		교과서		출판사		교과서		출판사
1	초등학교	슬기로운생활1-1	A	45	초등학교	슬기로운생활1-1	D	89	초등학교	슬기로운생활1-1	F
2	초등학교	슬기로운생활1-2	A	46	초등학교	슬기로운생활1-2	D	90	초등학교	슬기로운생활1-2	F
3	초등학교	슬기로운생활2-1	A	47	초등학교	슬기로운생활2-1	D	91	초등학교	슬기로운생활2-1	F
4	초등학교	슬기로운생활2-2	A	48	초등학교	슬기로운생활2-2	D	92	초등학교	슬기로운생활2-2	F
5	초등학교	자연3-1	A	49	초등학교	과학3-1	D	93	초등학교	과학3-1	G
6	초등학교	자연4-1	A	50	초등학교	실험관찰3-1	D	94	초등학교	실험관찰3-1	G
7	초등학교	자연5-1	A	51	초등학교	과학3-2	D	95	초등학교	과학3-2	G
8	초등학교	자연6-1	A	52	초등학교	실험관찰3-2	D	96	초등학교	실험관찰3-2	G
9	초등학교	실험관찰3-1	A	53	초등학교	과학4-1	D	97	초등학교	과학4-1	G
10	초등학교	실험관찰4-1	A	54	초등학교	실험관찰4-1	D	98	초등학교	실험관찰4-1	G
11	초등학교	실험관찰5-1	A	55	초등학교	과학4-2	D	99	초등학교	과학4-2	G
12	초등학교	실험관찰6-1	A	56	초등학교	실험관찰4-2	D	100	초등학교	실험관찰4-2	G
13	초등학교	자연3-2	A	57	초등학교	과학5-1	D	101	초등학교	과학5-1	G
14	초등학교	자연4-2	A	58	초등학교	실험관찰5-1	D	102	초등학교	실험관찰5-1	G
15	초등학교	자연5-2	A	59	초등학교	과학5-2	D	103	초등학교	과학5-2	G
16	초등학교	자연6-2	A	60	초등학교	실험관찰5-2	D	104	초등학교	실험관찰5-2	G
17	초등학교	실험관찰3-2	A	61	초등학교	과학6-1	D	105	초등학교	과학6-1	G
18	초등학교	실험관찰4-2	A	62	초등학교	실험관찰6-1	D	106	초등학교	실험관찰6-1	G
19	초등학교	실험관찰5-2	A	63	초등학교	과학6-2	D	107	초등학교	과학6-2	G
20	초등학교	실험관찰6-2	A	64	초등학교	실험관찰6-2	D	108	초등학교	실험관찰6-2	G
21	중학교	과학1	A	65	중학교	과학1	C	109	중학교	과학1	C
22	중학교	과학2	A	66	중학교	과학2	C	110	중학교	과학2	C
23	중학교	과학3	A	67	중학교	과학3	C	111	중학교	과학3	C
24	중학교	과학1	B	68	중학교	과학1	E	112	중학교	과학1	H
25	중학교	과학2	B	69	중학교	과학2	E	113	중학교	과학2	H
26	중학교	과학3	B	70	중학교	과학3	E	114	중학교	과학3	H
27	고등학교	공통과학	C	71	고등학교	과학	C	115	고등학교	과학	C
28	고등학교	공통과학	B	72	고등학교	과학	B	116	고등학교	과학	B
29	고등학교	물리1	C	73	고등학교	물리 I	B	117	고등학교	물리 I	C
30	고등학교	물리2	C	74	고등학교	물리 II	B	118	고등학교	물리 II	C
31	고등학교	물리1	B	75	고등학교	물리 I	C	119	고등학교	물리 I	B
32	고등학교	물리2	B	76	고등학교	물리 II	C	120	고등학교	물리 II	B
33	고등학교	생물1	C	77	고등학교	화학 I	B	121	고등학교	화학 I	C
34	고등학교	생물2	C	78	고등학교	화학 II	B	122	고등학교	화학 II	C
35	고등학교	생물1	B	79	고등학교	화학 I	C	123	고등학교	화학 I	B
36	고등학교	생물2	B	80	고등학교	화학 II	C	124	고등학교	화학 II	B
37	고등학교	지구과학1	C	81	고등학교	생물 I	B	125	고등학교	생명과학 I	C
38	고등학교	지구과학2	C	82	고등학교	생물 II	B	126	고등학교	생명과학 II	C
39	고등학교	지구과학2	B	83	고등학교	생물 I	C	127	고등학교	생명과학 I	B
40	고등학교	지구과학2	B	84	고등학교	생물 II	C	128	고등학교	생명과학 II	B
41	고등학교	화학1	C	85	고등학교	지구과학 I	B	129	고등학교	지구과학 I	C
42	고등학교	화학2	C	86	고등학교	지구과학 II	B	130	고등학교	지구과학 II	C
43	고등학교	화학1	B	87	고등학교	지구과학 I	C	131	고등학교	지구과학 I	B
44	고등학교	화학2	B	88	고등학교	지구과학 II	C	132	고등학교	지구과학 II	B

Table 2. Category of science textbook structure

구조	해당 내용
제목	대단원, 중단원, 소단원의 단원명
본문	교과서 내용이 기술된 주 부분
탐구활동	실험, 시범실험, 탐구 확인, 탐구 활동, 해보기, 관찰, 탐구 학습, 연구 과제, 조사, 자료 해석, 탐구해보자, 현장 학습, 생각해보기, 경연대회, 역할 놀이, 조사 및 토의, 사고실험, 연습하기, 모둠별 과제, 토의해보기, 창의력 키우기, 보고 생각하기, 미니 탐구 등
참고자료	참고 자료, 읽을거리, 각주, 참고, 과학과 우리생활, 보충, 인물 이야기, 과학자, 과학&생활, 인터넷 탐방, 과학과 환경, 더 알고 싶은 과학, 과학과 생활, 역사 속의 과학, 더 알고 싶은 최첨단 과학, 잘못 알기 쉬운 과학, 아하! 그렇구나, 이곳에서 정보를!, 흥미진진 과학 이야기, 과학과 직업, 잘못 알기 쉬운 과학, 노벨상에서 배우기, 직업의 세계, 자료실, 과학과 문명, 최신 과학, 쉽게 읽는 과학 고전, 과학 역사 신문, 우리 옆집 과학자, 마음을 움직이는 과학사진, 과학이 사방팔방, 과학의 과거 현재 미래, 미래 속 과학, 미니 읽기, 지식 충전, 더 알아보기 등
단원마무리	단원 요약, 중단원 요약, 대단원 요약, 요점 정리, 학습 정리, 단원 정리, 단원 학습 마무리, 개념 정리, 과학용어 정리, 한 눈에 보기, 개념 정리하기, 학습 요점, 중단원 학습 정리, 개념도, 개념 복습, 요약 정리, 중요개념 다시 보기, 가로세로 퍼즐, 개념 관련짓기, 단원 정리하기, 되짚어 보기, 핵심 정리 등

한편, 과학교과서 말뭉치에는 교육과정, 학년, 과목, 출판사 등의 서지 정보와 페이지, 문장 순서 등의 위치 정보, 제목이나 본문, 탐구 활동 등의 구조 정보 등의 메타정보들이 포함되어야 한다. 따라서 메타정보들을 표지하고 텍스트 구조화를 구현하기 위하여 입력 소프트웨어를 개발하였다(Figure 1).

이후 개발된 도구에 132권의 과학교과서 텍스트를 입력하였고, 입력에 참여하지 않은 인력에 의해 2차례 검토 과정을 거쳐 원시말뭉치를 구축하였다. 검토 과정에서는 메타정보의 오류, 구조 구분의 오류, 오탈자 오류 등을 수정하였다. 최종적으로 입력이 완료된 말뭉치는 총 320만 어절 규모에 달하였다. 이는 특수말뭉치로서 국립국어원의 21세기 세종계획의 역사자료 말뭉치가 280만 어절, 한영병렬 말뭉치가 270만 어절, 북한 및 해외 한국어 말뭉치가 440만 어절(Jeon, 2003)인 것과 비교하면 특수 언어 연구를 위한 말뭉치로는 상당하고 충분한 분량임을 알 수 있다. 또한 국립국어원이 10여 년 간 국가적 과제로 구축한 이들 말뭉치들에서 형태주석이 된 분량은 각각 100만 어절이 채 되지 않는데 반해 본 연구에서는 320만 어절 전체에 대해 형태주석 뿐만 아니라 의미주석까지 수행했으므로 그 규모와 가치가 높다고 하겠다. XML 형태의 완성된 원시말뭉치의 예시를 Figure 2에 제시하였다.

나. 형태주석 말뭉치 구축

형태주석 말뭉치는 형태소 분석기를 이용한 자동주석과 오류를 수정하는 수동주석의 두 단계로 실시하였다. 형태소 분석기는 울산대의 Utagger(Shin & Ock, 2012)를 사용하되 본 연구에서 구축한 원시말뭉치에 최적화 하여 형태주석 결과물을 검토, 수정할 수 있는 도구를 개발하여 사용하였다. 자동으로 분석된 형태주석 결과물 검토는 석박사 이상의 형태주석 전문 인력들이 수행하였으며, 320만 어절에 대한 형태분석 결과를 하나하나 검토하여 오류를 수정하였다. 개별 작업이 완료된 뒤에는 검토 과정에서 누락된 오류 부분을 다시 한번 확인하고 형태주석 말뭉치 구축 지침 내용의 빈자리로 인해 일관성이 지켜지지 않은 분석 결과들을 확인하여 이를 수정하는 후처리 일관성 검토 단계를 거쳤다. 후처리 일관성 검토 단계는 '일관적 분석 확인 단계→일관성 적용 단계'의 두 단계로 구성되었다. 먼저 '일관적 분석 확인 단계'에서는 자동 탐색을 통해 동일 어절 형태의 형태 주석 결과가 불일치하는 목록을 확보하여 검토하는 방안과 내용어를 중심으로 기준값을 설정하여 내용어 형태 주석 결과가 불일치하는 목록을 확보하여 검토하는 방안 두 가지를 설정하여 타당성을 검토하였다. 후처리 일관성 검토까지 완료된 형태주석 말뭉치의 구조는 Figure 3과 같다.

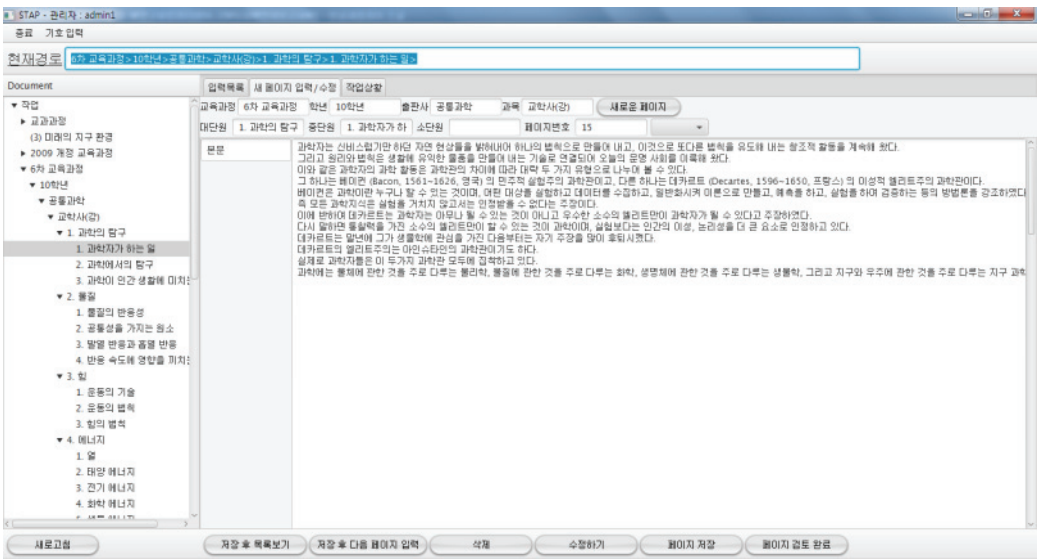

Figure 1. Software for making structure of raw corpus

```
- <books>
  - <book curriculum_version="6차 교육과정" grade="8학년" publisher="교학사(정)" subject="과학2">
    - <page num="11">
      - <struct type="타이틀">
          <sentence num="1">1 물질의 구성</sentence>
          <sentence num="2">1. 화합물과 원소</sentence>
          <sentence num="3">2. 물질을 구성하는 원자와 분자</sentence>
        </struct>
      </page>
    - <page num="11">
      - <struct type="본문">
          <sentence num="4">우리는 물질의 특성을 이용하여 혼합물로부터 순수한 물질을 분리할 수 있음을 1학년에서 배웠다.</sentence>
          <sentence num="5">이 단원에서는 화학 변화와 원소에 관하여 공부한 다음, 원자와 분자의 모형으로 화합물의 성분비와 부피비를 조사해 보며, 분자 운동으로 보일의 법칙과 샤를의 법칙
            을 설명할 수 있는지 알아본다.</sentence>
        </struct>
      </page>
    - <page num="12">
      - <struct type="참고자료">
          <sentence num="1">● 물질을 구성하는 입자에 관한 연구</sentence>
          <sentence num="2">18세기의 라부아지에의 실험실</sentence>
          <sentence num="3">왼쪽 벤반 위와 아래에 여러 가지 실험 기구가 갖추어져 있다.</sentence>
          <sentence num="4">잘 꾸며진 오늘날의 실험실</sentence>
          <sentence num="5">새로운 화합물을 만들어내기 위한 정밀한 실험 장치가 있다.</sentence>
          <sentence num="6">투과 전자 현미경</sentence>
          <sentence num="7">물질을 이루는 원자의 배열과 상을 수백만 배로 확대하여 볼 수 있다(한국 과학 기술원 제공).</sentence>
          <sentence num="8">금의 원자의 상</sentence>
          <sentence num="9">투과 전자 현미경으로 본 금 원자의 확대된 상으로, 원자가 규칙적인 배열을 하고 있다.</sentence>
        </struct>
      </page>
    - <page num="13">
      - <struct type="타이틀">
          <sentence num="1">1. 화합물과 원소</sentence>
        </struct>
      </page>
    - <page num="13">
      - <struct type="본문">
          <sentence num="2">우리는 물질의 특성을 이용하여 혼합물로부터 순수한 물질을 분리할 수 있었다.</sentence>
```


Figure 2. Sample of raw corpus

```
<?xml version="1.0"?>
- <books>
  - <book subject="물리 I " publisher="교학사" grade="11학년" curriculum_version="6차 교육과정">
    - <page num="7">
      - <struct type="타이틀">
          <sentence num="1"> I 운동과 에너지</sentence>
        - <word tid="1" sid="1">
            <orth> I </orth>
            <mor pos="SW"> I </mor>
          </word>
        - <word tid="2" sid="1">
            <orth>운동과</orth>
            <mor pos="NNG">운동__02</mor>
            <mor pos="JC">과</mor>
          </word>
        - <word tid="3" sid="1">
            <orth>에너지</orth>
            <mor pos="NNG">에너지</mor>
          </word>
          <sentence num="2">1. 물체의 운동</sentence>
        - <word tid="1" sid="2">
            <orth>1.</orth>
            <mor pos="SN">1</mor>
            <mor pos="SF">.</mor>
          </word>
        - <word tid="2" sid="2">
            <orth>물체의</orth>
            <mor pos="NNG">물체</mor>
            <mor pos="JKG">의</mor>
          </word>
        - <word tid="3" sid="2">
            <orth>운동</orth>
            <mor pos="NNG">운동__02</mor>
          </word>
```


Figure 3. Sample of morphologically annotated corpus

다. 용어주석 말뭉치 구축

끝으로 용어주석 말뭉치는 크게 일곱 단계를 거쳐서 구축하였는데, 기존 용어집을 이용한 1차 자동 주석 단계, PoS-gram(Part-of-Speech-gram)[7]을 통한 과학용어 후보 목록 추출 단계, 후보 목록 검토 및 선별 단계, 2차 자동 주석 단계, 의미 검토 단계, 정의문 선별 단계, 최종 검토 및 수정의 단계로 진행하였다.

자동 주석 단계에서는 기존의 자료를 활용하여 단순 대조를 통한 용어 선별을 하였는데, 이 때 사용한 자료는 표준국어대사전(The National Institute of the Korean Language, 2008)과 물리, 생명과학, 지구과학, 화학 분야의 각 학회에서 발간한 학술용어집의 용어 목록이다. 이들 목록은 각각 수천 개 혹은 수만 개씩의 전문용어를 수록하고는 있으나 특별한 기준이나 준거에 의해 제작되거나 관리되기 보다는 과거의 자료를 유지 혹은 일부 보완하는 소극적인 형태로 관리되고 있다. 따라서 학계에서 조차 더 이상 사용하지 않는 사어나 통일되지 않은 표기 등이 포함되어 있기도 하고 혹은 누락된 용어들이 발견되기도 하였다(Yun & Park, 2014). 본 연구에서는 과학교과서 텍스트 말뭉치에서 과학용어를 선별하여 의미 주석을 다는 것이 목적이므로 사어나 표기 불일치 용어들이 추출되는 것은 별로 문제가 되지 않으나 과학용어임에도 주석이 누락되는 것을 막는 것은 중요한 이슈라고 판단하였다. 따라서 기존 용어집을 그대로 활용하여 포괄적인 용어 추출을 실시한 뒤, 누락된 용어들을 구제하는 방안을 마련하였다. 표준국어대사전의 전문어 목록과 각 학계에서 발행한 용어집을 통합하여 물화생지 분야별로 정리한 결과 물리 분야 24,108개, 생명과학 분야 47,029개, 지구과학 분야 25,231개, 화학 분야 25,391개의 용어

7) PoS-gram(Part-of-Speech-gram) : 품사 연쇄를 의미하는데 '용어를 이루는 문법적 패턴'과 '용어를 이루지 않는 문법적 패턴'을 구분하는 방식을 취한다. POS-gram 패턴은 기 구축한 용어 사전 목록에 포함된 개별 항목의 품사 연쇄 패턴을 대상으로 한다. 본 연구에서 활용하는 기 구축 용어 사전 목록은 이전 단계에서 수집 구축한 용어 사전 목록 가운데 단어 단위를 제외한 구 단위 표제어 목록이다.

목록을 생성하였다.

누락된 과학용어들을 구제하기 위해서 가장 단순하게는 전체 말뭉치를 어절 하나하나씩 검토하며 자동 주석 단계에서 누락된 과학용어들을 찾아낼 수 있으나 너무 많은 인력과 시간, 비용이 발생하는 문제, 전문 인력들을 동원하더라도 판단의 일관성을 확보하기 어렵다는 문제 등이 있다. 따라서 본 연구에서는 이러한 문제들을 최소화하는 방안으로 PoS-gram을 도입하였다. 먼저 기존의 전문용어 목록의 형태소 분석을 통해 과학용어가 가질 수 있는 형태소 연쇄 목록을 작성하고, 작성된 목록을 기반으로 앞 단계에서 구축된 형태주석이 완료된 과학교과서 말뭉치로부터 과학용어 후보 목록을 추출하였다. 총 467개의 PoS-gram 목록을 기준으로 580,513개의 과학용어 후보가 추출되어 검토 대상이 되었다. 이 방법은 원시말뭉치에서 띄어쓰기 오류로 인해 누락된 과학용어들이나 다양한 표기로 인해 누락된 용어들을 구제할 수 있고, 기존 용어집에서 수록되어 있지 않아 자동 추출 단계에서 누락된 과학용어들을 구제할 수 있으며, 후보 추출 단계에서 작업자의 주관이 개입되지 않으므로 작업의 일관성이 확보되고 매우 짧은 시간에 자동으로 일괄 처리가 가능하다는 장점이 있다. 그러나 형태소 연쇄는 일치하나 과학용어가 아닌 결과물을 구별하는 추가 작업을 필요로 한다. 예를 들어 전문용어에서 가장 전형적으로 나타나는 형태소 연쇄인 'NNG+NNG' 검색에서 '나선 은하', '용수철 저울' 등의 과학용어와 '저울 눈금', '다음 그림' 등과 같은 과학용어가 아닌 결과물들이 동시에 추출이 된다. Table 3에 PoS-gram 으로 추출한 과학용어 후보 목록 가운데 일부를 예시로 제시하였는데, A는 과학용어로 구분한 사례이고 B는 A와 같은 형태소 연쇄를 가지지만 과학용어로 구분하지 않은 사례이다.

PoS-gram을 통한 과학용어 후보 목록에서 과학용어로 구제할 용어와 그렇지 않은 것들을 선별하는 작업은 박사과정 이상의 현장 중

Table 3. List of candidate scientific terms extracted using pos-gram

POS-gram	A	B
NNP NNG	베네딕트 용액 허블 상수	크리미아 전쟁 유럽 각지
NNP JKG NNG	샤를의 법칙 아보가드로의 분자설	폴란드의 천문학자 라부아지에의 죽음
NNG NNG NNG	분자 운동 모형 이상 기체 법칙	습도 조건 아래 주요 사업 가운데
VV ETM NNG	녹은짐	지난 며칠 기입한 후
VA ETM NNG	묽은 염산	약한 비

* NNP: 고유명사, NNG:일반명사, JKG:관형격조사, VV: 동사, ETM: 관형형전성어미, VA: 형용사

고등학교 물리, 생명과학, 지구과학, 화학 각 과목 담당 교사 12명으로 전문가 집단을 구성하여 실시하였다. 각 전공별로 3명씩 구성된 전문가 집단은 1차로 각자 선별을 실시하고, 1차 선별 결과물 가운데 3명이 만장일치로 구제하거나 삭제한 결과물은 그대로 수용하고, 3명의 의견이 엇갈린 경우는 모여서 논의 후 처리하였다. 전문가 집단에서 마지막까지 합의가 되지 않은 경우는 과학교육전문가 2인과 전문가 집단이 다시 한 번 논의를 통해 최종 결정을 내렸다. 검토 작업의 효율성과 결과물 관리의 효율성을 고려하여 검토 도구를 개발하여 사용하였다(Figure 4). 총 58만여 개의 후보 목록 가운데 23,118개의 단어가 1명 이상의 전문가로부터 과학용어로 선택이 되었고 이 가운데 2,321개가 3명의 전문가 모두로부터 과학용어로 선택이 되었다 (Table 4). 전문가 집단의 의견이 엇갈린 용어에 대한 최종 검토를 통해 과학용어로 선택된 단어는 물리가 859개, 생명과학이 932개,

Figure 4. Eample of selecting scientific terms from the candidates

Table 4. Number of selected scientific terms

분야	표준국어대사전	학술용어집	공통 용어	표준국어대사전 + 학술용어집 (중복제거)=A	3명 만장일치 과학용어로 선택한 개수=B	1명 또는 2명이 과학용어로 선택한 개수 (최종 채택 용어 개수=C)	최종 대조 사전 목록 (A+B+C)
물리	11,469	15,373	3,904	24,108	361	3,580 (859)	25,328
생명과학	26,342	46,832	26,145	47,029	1,039	9,050 (932)	49,000
지구과학	12,883	15,333	2,985	25,231	50	1,913 (608)	25,889
화학	10,342	17,342	2,293	25,391	871	6,254 (1,032)	27,294

지구과학이 608개, 화학이 1,031개였으며 이 단어들과 앞서 3명의 전문가가 만장일치로 선택한 단어들과 합하여 최종적으로 과학용어 목록에 추가하였다. 이 단계에서 선정된 과학용어들을 기존의 용어집 목록에 추가하여 최종적으로 과학용어 자동 주석을 위한 대조 사전을 완성하였고, 이를 이용하여 2차 자동 용어 주석을 실시하였다. 최종적으로 정리된 과학용어 대조 사전의 분야별 용어 수는 물리 25,328개, 생명과학 49,000개, 지구과학 25,889개, 화학 27,294개이다. Table 4에 제시한 바와 같이 표준국어대사전의 전문어와 각 학계의 학술용어집의 일치도가 매우 낮은 것을 알 수 있다. 표준국어대사전은 고등학교 수준의 전문어를 수록한 자료이고 학술용어집은 전문가가 사용하는 수준까지를 포함하는 자료임을 고려하면 학술용어집은 표준국어대사전의 용어를 모두 포함하고 있는 것이 이상적이다. 그러나 생명과학을 제외한 나머지 분야에서는 표준국어대사전에서 과학용어로 구분하고 있음에도 학술용어집에 누락된 용어가 대다수를 차지함을 알 수 있다. PoS-gram을 통해 구제한 누락 용어의 수도 분야별로 각각 일천 개 가까이므로 적은 수가 아니다. 따라서 본 연구에서 정비한 물리, 생명과학, 지구과학, 화학 과학용어 대조사전 목록은 지금까지 통일되지 못했던 표준국어대사전과 학술용어집을 통합 정비하고 누락된 과학용어들을 추가한 의미 있는 자료라 할 수 있다.

컴퓨터가 수행하는 자동 주석은 과학용어 대조 사전과 텍스트를 대조하여 형태가 일치하는 단어에 과학용어임을 표지하고 분야에 대한 정보를 태깅하는 것이다. 그러나 컴퓨터가 형태의 대조는 정확하고 빠르게 처리할 수 있으나 의미 대조는 쉬운 일이 아니다. 본 연구에서 사용한 형태소분석기 Utagger의 경우 문맥 검토를 통해 자동으로 단어의 의미를 파악하여 표준국어대사전의 의미 분류 기준인 어깨번호를 부착해준다. 그러나 Utagger가 참조하고 있는 데이터베이스에 과학 텍스트의 비율이 낮아 과학텍스트에서의 의미 분류에 대한 정확도는 매우 낮다. 따라서 하나의 표현형에 다양한 의미를 담고 있는 단어들을 대상으로 그 의미를 구분해주는 작업을 수행하였다. 예를 들어 '눈'의 경우 '빛의 자극을 받아 물체를 볼 수 있는 감각 기관'의 의미로 사용된 경우와 '대기 중의 수증기가 찬 기운을 만나 얼어서 땅 위로 떨어지는 얼음의 결정체'의 의미, 그리고 '새로 막 터져 돋아나려는 초목의 싹. 꽃눈, 잎눈 따위'의 의미로 사용되는 경우 등을 정확하게 구분하는 일은 아직은 사람이 할 수 밖에 없다. 이 작업은 물화생지 각 분야를 전공한 전문 인력들이 수행하였다. 과학용어로 자동 주석이 된 결과물에서 용어주석이 달린 단어가 포함된 문장들을 불러온 다음에 문맥을 고려하여 단어의 의미를 구분하여 일련번호를 부착하였다. 이 때 부여한 번호는 표준국어대사전의 어깨번호와 일치시키지는 않았다.

본 연구의 목적 가운데 하나는 과학용어의 교육이 포함된다. 따라서 과학용어에 대한 교과서 정의문들이 의미있게 활용될 수 있을 것으로 판단하였다. 이에 의미 검토 과정에서 정의문에 대한 표지를 동시에 부여하였다. 끝으로 과학용어 가운데는 일상적 의미와 함께 사용되고 있어서 문맥에 따라 과학용어로 사용되기도 하고 일상어로

Figure 5. Software for examination a context of the sentences

```
<sentence num="7">물체의 가속도는 작용하는 힘에 비례하고, 질량에 반비례한다.</sentence>
<word sid="7" tid="1">
<orth>물체의</orth>
<mor pos="NNG">물체</mor>
<sci area = "phy">물체</sci>
<mor pos="JKG">의</mor>
</word>
<word sid="7" tid="2">
<orth>가속도는</orth>
<mor pos="NNG">가속도</mor>
<sci area = "phy">가속도</sci>
<mor pos="JX">는</mor>
</word>
<word sid="7" tid="3">
<orth>작용하는</orth>
<mor pos="VV">작용하__01</mor>
<sci area = "phy">작용</sci>
<mor pos="ETM">는</mor>
</word>
<word sid="7" tid="4">
<orth>힘에</orth>
<mor pos="NNG">힘__01</mor>
<sci area = "phy">힘_01</sci>
<mor pos="JKB">에</mor>
</word>
```

Figure 6. Sample of semantic annotated corpus of scientific term

사용되기도 하는 경우가 있다. 이 경우 일상적 의미로 사용된 경우는 자동 주석 과정에서 과학용어로 표지되었지만 표지를 해제해야 한다. 과학용어의 의미를 구분하는 작업, 정의문을 선별하는 작업, 일상적 의미로 사용된 경우 표지를 해제하는 작업은 모두 문장을 검토하며 문맥을 파악해야 한다. 따라서 문장의 한 번 검토에서 세 작업이 동시에 이루어질 수 있도록 하나의 도구를 개발하였다(Figure 5). 이 때 각 문장별로 서로 다른 두 명의 물화생지 전공자가 작업을 하도록 배정하고 두 명의 작업이 일치하는 경우는 그대로 수용하고, 그렇지 않은 경우는 과학교육전문가의 재검토를 통해 처리하였다. 검토 과정을 거친 뒤 최종적으로 과학용어에 대한 의미 주석이 달린 용어주석 말뭉치를 완성하였다(Figure 6).

Ⅲ. 결론 및 제언

본 연구에서는 과학교육에서 그 동안 주목받지 못했던 과학언어 및 과학용어에 대한 연구를 체계적으로 수행하기 위한 목적으로 지난 20년간의 과학교과서 텍스트를 한 자리에 모아 과학교과서 말뭉치를 구축함으로써 다각도로 분석 가능한 형태의 언어 자원을 생성하였다. 총 3년여에 걸친 연구 기간 동안 언어정보학, 컴퓨터공학, 과학교육학의 세 분야 전문가들의 노하우를 융합하여 새로운 연구 방법을 창출하였고, 다수의 전문 인력들이 투입되어 노동집약적 결과물을 내었다. 본 원고에서는 전체적인 연구 절차와 방법을 조망함으로써 새로운 연구 방법론 및 결과물을 소개하고 향후 과학언어 연구의 발전 가능성 및 결과물의 활용방안에 대해 논의하고자 하였다. 처음 시도되는 방법인 만큼 각 연구 단계 마다 많은 논의와 이슈가 있었으나 과학교육학적 관점에서 본 연구가 갖는 의의를 중심으로 아래에 기술해 보았다.

첫째, 지금까지 흩어져 있었던 과학교과서 자료들을 한 자리에 모으고 전자 문서의 일관된 형태로 정리하였다. 과학교과서 말뭉치를 구축하기 위하여 6차 교육과정에서부터 2009 개정 교육과정까지의 모든 과학교과서를 수집하였다. 최근에 나온 2009 개정 교육과정의 경우는 자료를 구하기가 수월했으나, 7차 교육과정의 일부 출판사와 6차 교육과정의 자료는 수집에 어려움이 있었다. 대부분의 출판사에서는 지난 교육과정의 교과서에 대한 체계적인 보관이 이루어지지 않고 있었고, 한국교육과정평가원과 한국교육개발원의 자료실, 그리고 한국교과서연구재단에서도 모든 교과서를 다 보유하고 있지는 않았다. 이러한 기관들에서 보유하고 있지 않은 과학교과서들은 전국의 헌책방과 대학 도서관 등을 통해 구하였다. 한편 국가적 차원에서 구축된 언어자원들의 경우 매우 방대한 양의 대규모 말뭉치가 구축되어 연구용으로 공개되어 있다. 이러한 범용 말뭉치들의 경우 정제되어 있고 형태소 분석이 잘 되어 있어 단어의 추출과 빈도 분석 등의 연구가 용이하게 되어 있으나, 이러한 자원들에 교육용 텍스트의 비율은 매우 낮으며 특히 과학교과서 텍스트가 포함된 말뭉치는 전무한 상태이다. 본 연구에서는 비록 수집된 모든 교과서를 주석말뭉치로 구축하지는 못했으나, 구축된 132권의 약 320만 어절의 말뭉치는 한 분야의 특수 목적용 말뭉치로는 충분한 분량의 언어자원이라 볼 수 있다(Kang, 2011). 추후 과거의 나머지 교과서들과 앞으로 나올 과학교과서들에 대한 텍스트를 순차적으로 추가하여 과학교과서 말뭉치 K-STeC을 지속적으로 보완, 관리할 계획이다.

둘째, 발행 주체에 따라 통일되지 못하고 있던 용어집 자료들을 한 곳에 모으고 정리, 보완하였다. 현재 고등학교 수준까지의 과학용어들이 포함된 자료로는 편수자료, 표준국어대사전, 각 학계의 용어집 정도를 꼽을 수 있다. 그러나 이들 자료 사이의 불일치도가 높고, 세 자료를 합하더라도 여전히 누락된 용어들이 존재한다(Yun & Park, 2014). 본 연구 과정에서는 이들 자료들을 한 자리에 모아 정리하고 PoS-gram 방법을 사용하여 누락된 용어들을 보완하기 위한 노력이 있었다. 물론 본 연구에서 구축한 과학용어 대조 사전 역시 완벽하다고 보기는 어렵다. 그러나 지금까지 분야별로, 그리고 자료의 주체별로 통일되지 않고 상이했던 목록을 한 자리에 모아서 정리했다는 점, 어느 자료에도 포함되어 있지 않아 누락되어 있었던 과학용어들을 어느 정도 보완했다는 점에서 가치있는 자료라 여겨진다.

셋째, 과학용어의 교육적 우선 순위를 가늠할 근거 자료가 확보되었다. 지금까지 학교 교육 수준에서 과학용어를 대상으로 한 연구들은 과학용어 선정의 주관성 문제를 배제하기 어려웠다. 본 연구의 결과물이 과학용어와 과학용어가 아닌 단어의 경계를 구분짓는 문제를 해결하지는 못한다. 그러나 학생들이 알아야 할 과학용어들의 우선순위 문제의 논의를 가능하게 해주는 것은 분명하다. 물론 이 때 가장 중요한 것은 과학적 지식의 맥락에서 얼마나 중요한 개념을 담고 있는 과학용어인가의 문제일 것이며 이는 교육과정이나 전문가 집단의 논의를 통해 결정될 수 있다. 그러나 과학교과서에 사용되고 있는 과학용어들 가운데 개념적 으로 중요하게 다루어지는 것들은 극소수에 불과하며 대부분의 과학용어들은 이러한 논의에서 소외되거나 낮은 수준에서 유사한 중요성을 갖는다. 개념적 논의에서 소외된 많은 과학용어들의 교육적 우선 순위를 논의할 다른 근거가 필요하며 이 때 빈도가 중요한 단서가 된다. 지난 20년간 초중고 과학교과서에서 단 한 번 사용된 과학용어와 수천 이상의 빈도로 사용된 과학용어 가운데 학생들이 우선적으로 학습해야 할 과학용어가 어느 쪽인가에 대해서는 이견이 없을 것이다. 과거 교과서에 사용된 빈도를 바탕으로 과학용어의 교육적 우선 순위를 논의한 연구가 있었으나(Yun & Park, 2013b), 하나의 교육과정 및 물리 영역에만 국한되어 있어 활용에 제한이 있었다. 본 연구에서 구축된 과학교과서 말뭉치는 3 차례의 교육과정과 2개의 출판사를 포괄하고 있으며, 아울러 물화생지 네 영역을 모두 포함한다. 따라서 학생들이 과학 학습에 필요한 과학용어에는 어떠한 것이 있는지, 어떤 용어를 먼저 학습하는 것이 좋을지 등을 논의하기에는 충분한 자료가 될 수 있을 것이다.

넷째, 4차 산업혁명 이후의 과학교육에 대비한 중요한 자원이 될 것이다. 4차 산업혁명의 키워드 가운데 하나는 인공지능이며, 인공지능의 핵심 기술인 딥러닝 가운데 문자 기반 기법인 RNN은 말뭉치를 자원으로 한다. 따라서 어떠한 말뭉치가 들어가느냐에 따라 딥러닝의 결과와 활용 및 인공지능의 성능이 좌우된다고 볼 수 있다. 가까운 미래의 화두가 인공지능임을 인정한다면 분명 새로운 형태의 과학교육 서비스에는 본 연구의 결과물이나 본 연구에서 수행했던 연구 방법들이 유용하게 활용될 수 있을 것이다.

이 외에도 과학교과서 텍스트가 담고 있는 과학적 지식의 구조와 사고 체계를 알아볼 수 있는 자료로서의 가치, 과학 교사 교수 자료로서의 가치, 과학 학습 자료 생성 용도로의 가치, 완벽하다고는 하기 어려울 것이나 이상적인 과학 교수 언어로서 교사 교육 용도로의 가치 등 과학교과서 말뭉치가 과학교육학적으로 갖는 잠재적 가치는

매우 넓다. 아래에는 언어학에서 지금까지 있어왔던 말뭉치 활용 사례들을 기반으로 하여 과학교과서 말뭉치의 과학교육학적 활용 방안을 기술해 보았다.

첫째, 과학언어의 언어적 특징을 밝히는 데 활용할 수 있다. 언어학에서 말뭉치는 언어적 특징을 통계화하고 계량적으로 기술하며, 언어현상을 체계적이고 실증적으로 가시화하거나 언어적 직관을 뒷받침하는 증거로써 이용된다(Jeon, 2003). 과학의 언어가 일상의 언어와 달라서 학생들이 어려워한다는 추상적 논의에서 한 단계 나아가 과학언어가 구현됨에 있어 일상어와 구체적으로 어떻게 다른지, 과학적 지식의 구조와 사고체계가 어떻게 나타나고 있는지 등을 계량적으로 밝혀내는 것이 가능해진다. 말뭉치의 출발인 미국의 브라운 코퍼스가 미국 영어 사용법을 구체적으로 밝히기 위해 만들어진 것(Kang, 2011)과 같은 맥락으로 이는 가장 대표적이고 주된 활용 영역이 될 것이다.

둘째, 학습용 과학용어 사전 편찬에 활용될 수 있다. 사전 편찬을 위해서는 사전에 수록된 단어를 선정하고, 문법적 특징, 의미, 정의문, 용례 등의 정보를 수집하는 과정이 필요하다. 영국의 롱맨/랭커스터 코퍼스와 버밍엄 코퍼스처럼 사전 편찬을 목적으로 계획되고 구축된 코퍼스도 있을 만큼(Kang, 2011) 사전 편찬에 말뭉치는 유용하게 활용된다.

셋째, 과학언어의 교육에 활용될 수 있다. 언어 교육 분야의 경우 과거 문법학습, 뜻풀이식 어휘 교육에서 최근 의사소통이나 과제 해결 중심으로 목표가 바뀌면서(Jeon, 2003) 말뭉치가 언어 교육을 위한 실제적 자원으로 활용되고 있다. 과학교과서 말뭉치 역시 과학언어 교육의 전략을 수립하고 교수 자료를 확보함에 있어 실제적 자원이 된다.

넷째, 그동안 물리 영역에 제한되어 있던 과학용어 등급화의 영역 확장 및 타당성 확보에 활용될 것이다.

위에서 언급한 외에도 과학교과서 말뭉치의 활용 방안은 여러 가지가 있으며 그 가능성은 열린 집합일 것이다. 언어학에서 문법이나 어휘 연구 혹은 사전 편찬을 목적으로 말뭉치 구축을 시작하였으나, 현재 그 활용 범위와 가치가 무한히 확장되어 온 것과 같이 과학교과서 텍스트 말뭉치 역시 현재의 목적이나 활용 범위를 넘어서 앞으로 더 많은 확장 가능성을 가지고 있다고 여겨진다.

국문요약

본 연구에서는 과학교육에서 그 동안 주목받지 못했던 과학언어 및 과학용어에 대한 연구를 체계적으로 수행하기 위한 목적으로 지난 20년간의 과학교과서 텍스트를 한 자리에 모아 과학교과서 말뭉치를 구축함으로써 다각도로 분석 가능한 형태의 언어 자원을 생성하였다. 말뭉치 구축 대상 자료는 6차 교육과정, 7차 교육과정, 2009 개정 교육과정의 초등학교에서부터 고등학교까지 모든 과학교과서를 수집하고 이 가운데 두 개의 출판사에 해당하는 132권에 대한 말뭉치를 구축하였다. 원시말뭉치, 형태주석 말뭉치, 용어주석 말뭉치의 총 3단계로 구축하였다. 최종적으로 구축된 과학교과서 말뭉치를 K-STeC(Korea - Science Textbook Corpus)이라 명명하였다. K-STeC은 과학용어에 대한 의미 구분과 분야가 표지된 의미 주석 말뭉치로서 교육과정, 과목, 학년, 출판사의 서지 정보와 대단원, 중단원, 소단원의 단원 정보, 페이지, 문장번호의 위치 정보와 함께 본문, 탐구활동, 참고자료, 제목 등의 텍스트 구조 정보를 메타정보로 마크업 하였다. 총 3년여에 걸친 연구 기간 동안 언어정보학, 컴퓨터공학, 과학교육학의 세 분야 전문가들의 노하우를 융합하여 새로운 연구 방법을 창출하였고, 다수의 전문 인력들이 투입되어 노동집약적 결과물을 내었다. 본 원고에서는 전체적인 연구 절차와 방법을 조망함으로써 새로운 연구 방법론 및 결과물을 소개하고 향후 과학언어 연구의 발전 가능성 및 결과물의 활용방안에 대해 논의하였다.

주제어 : 과학교과서 말뭉치, 과학교과서, 말뭉치, 과학언어, 과학용어

References

Darian, S. G. (2003). Understanding the language of science. University of Texas Press.

Fang, Z. (2006). The language demands of science reading in middle school. International Journal of Science Education, 28(5), 491-520.

Ford, A., & Peat, F. D. (1988). The role of language in science. Foundations of Physics, 18, 1233.

Ham, J., Lee, J., & Shin, D. (2011). Middle school students' feelings of easiness and understanding of earth science terminology. Journal of Research in Curriculum Instruction, 15(4), 1045-1060.

Jaipal, K. (2001). English second language students in a grade 11 biology class: Relationships between language and learning. Proceeding of 2001 Annual Meeting of the American Educational Research Association, ED 453690.

Jeon, S. H. (2003). (21st Sejong project) Application of corpus. The National Institute of the Korean Language.

Kang, B. M. (2011). Language, computer, and corpus linguistics. Seoul: Korea University Press.

Kwak, Y. (2013). Corpus quality control for high-quality language resource construction. Doctoral Dissertation, Yonsei University.

Kwak, Y. (2017). Exploration of features of Korean eighth grade students` achievement and curriculum matching in TIMSS 2015 earth science. Journal of the Korean Association for Science Education, 37(1), 9-16.

Kwak, Y., Kim, C. J., Lee, Y. R., & Jeong D. S. (2006). Investigation on elementary and secondary students` interest in science. Journal of the Korean Earth Science Society, 27(3), 260-268.

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). TIMSS 2015 International results in science. IEA.

Maskill, R. (1988). Logical language, natural strategies and the teaching of science. International Journal of Science Education, 10(5), 485-495.

McEnery, A. & Hardie, A.(2012). Corpus linguistics: Theory, method and practice. Cambridge: Cambridge University Press.

Merzyn, G. (1987). The language of school science. International Journal of Science Education, 9(4), 483-489.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. HLT-NAACL, 746–751.

Miller, J. (2009). Teaching refugee learners with interrupted education in science: Vocabulary, literacy and pedagogy. International Journal of Science Education, 31(4), 571-592.

Nam, K. S. (2008). Middle school students' learning difficulty caused by scientific terminology and ways to solve it via writing using scientific terminology. Doctoral Dissertation, Seoul National University.

Park, Y., Gwon, S., & Yun, E. (2015). Research for improvement of science textbook(Physics) through inducing change of instruction. Korea Foundation for the Advancement of Science & Creativity, Research Report.

Reeves, C. (2005). The language of science. Routledge.

Shaw, J. (2002). Linguistically responsive science teaching. Electronic Magazine of Multicultural Education, 4(1), 24.

Shin, J. C., & Ock, C. Y. (2012). A stage transition model for Korean part-of-speech and homograph tagging. Software and Application, 39(11), 889-901.

The National Institute of the Korean Language (2008). Pyojun Korean unabridged dictionary, The National Institute of the Korean Language.

Wellington J., & Osborne, J. (2001). Language and literacy in science education. Open University Press.

Yore, L. D., Hand, B., Goldman, S. R., & Hildbrand, G. M. (2004). New directions in language and science education research. Reading

Research Quarterly, 39(3), 347-352.
Yun, E., & Park, Y. (2013a). Research on science teacher`s perception of teaching science terminology. Journal of the Korean Association for Science Education, 33(7), 1343-1353.
Yun, E., & Park, Y. (2013b). Analysis of physics terminology in science textbooks for teaching science words. Journal of the Korean Association for Science Education, 33(4), 735-750.
Yun, E., & Park, Y. (2014). Consistency among the glossary for a textbook, the Glossary of Physics Terminology and the Pyojun Korean Unabridged Dictionary on the basis of the words used in middle-school science textbooks. Sae Mulli, 64, 180-187.