초등학교 교과서의 어휘 통계 분석 연구 : 한국어 세종 코퍼스와의 비교를 중심으로

유원희† · 임희석††

요 약

본 논문에서는 초등학교 교과서 말뭉치를 구축하고, 초등교과서에서 나타나는 어휘들에 대하여 통계 분석을 실시하였다. 또한 초등 교과서가 일반생활에서 사용하는 어휘와 얼마나 유사한지를 살펴보기 위하여 스피어만 상관관계 분석을 실시하였다. 연구결과로 초등교과서의 말뭉치 구축 모습과 실제 예시를 보였고, 상관관계 분석을 통하여 초등교과서와 일반 말뭉치와의 상관관계를 수치적으로 보였다.

주제어 : 초등 교과서 말뭉치, 초등 교과서 통계, 상관관계 분석

The Study Of Lexical Statistics Analysis For Elementary School Textbook : Focusing On Comparing The SEJONG Corpus In Korean

Wonhee Yu† · Heuiseok Lim††

ABSTRACT

In this paper, we build a primary school textbook corpus and a statistical analysis was performed with respect to the vocabulary found in elementary textbooks. also We performed the Spearman's correlation coefficient in order to explore whether similar elementary textbooks in general life used vocabulary. the result of this study shows that corpus building in the form of elementary school textbooks and actual examples. then numerically shown correlation of the elementary textbooks and general corpus.

Keywords : Elementary School Corpus, Elementary School Statistics, Correlation Coefficient

† 정 회 원: 고려대학교 컴퓨터교육학과 박사수료
†† 종신회원: 고려대학교 컴퓨터교육학과(교신저자)
논문접수: 2014년 11월 24일, 심사완료: 2014년 12월 1일, 게재확정: 2015년 1월 22일
* 본 논문은 2014년 한국연구재단의 지원으로 수행되었음(NRF-2013S1A5A2A03044158)

1. 서론

학교교육은 체계상 초등교육·중등교육·고등교육으로 구분되는데, 이중 초등교육은 초등수준과 일반 기초소양을 내용으로 하여야 한다.

한국의 교육법 93조는 '초등학교는 국민생활에 필요한 기초적인 초등보통교육을 하는 것을 목적으로 한다고 규정하여 그 교육의 목적을 명시하였다.

초등교육에서 이루어지는 교육 내용은 아동이 성장·발달해가기 위한 가장 절실한 내용이 되어야 하므로 지적발달은 물론 신체적·사회적·정서적·정신적인 제반 영역에 걸쳐 광범위한 내용이 취급되어야 한다.

따라서 초등교육의 주요내용은, ① 아동의 학습에 필요한 능력을 사용하거나 또는 그 기능을 발전시켜 나아갈 수 있는 내용으로서 말하기·듣기·읽기·쓰기를 비롯하여 관찰과 가르치는 능력, 계산 능력, 문제를 분석하는 능력, 추리 능력, 물건을 만드는 능력 등을 길러 주는 내용, ② 집단생활에서 일어나는 여러 문제를 해결할 수 있는 능력을 길러 주는 내용, ③ 아동으로 하여금 인간생활의 물질적 ·자연적 환경에 관한 이해를 깊이 할 수 있는 내용, ④ 아동이 창조적 표현을 할 수 있는 내용, ⑤ 아동의 건강생활에 관한 내용 등으로 되어 있다[1].

초등 교과서는 초등 교육의 목적과 내용이 잘 반영되도록 구성되어야한다. 이는 초등 교과서의 어휘구성이 일반 성인들이 사용하는 어휘들 수준으로 구성되어 아동이 사용하는 어휘수준을 일반 성인들이 사용하는 어휘수준으로 향상시키며, 일반 성인들이 사용하는 어휘들을 잘 이해하여 아동의 성장, 발달에 기여할 수 있는 내용들로 구성되는 것이 바람직할 것이다.

그러나 실제 초등 교과서가 어떠한 어휘들을 사용하는지에 대한 연구들이나 일반인들이 사용하는 어휘와 어느 정도 차이가 있는지에 관한 연구들은 거의 이루어지지 않고 있다. 이는 초등 교과서의 어휘들을 분석할 수 있는 초등 교과서 분석 말뭉치가 구축되지 않은 이유가 제일 크다. 때문에 초등교과서 말뭉치를 구축하고 이에 대한 기초 통계 자료를 제시하는 것은 매우 의미 있는 연구라고 판단된다.

이 연구에서는 초등학교 교과서를 말뭉치 형태로 구축하고 이것을 바탕으로 기본적인 어휘 통계 분석 자료를 제시하고자 한다. 구체적인 연구 목적은 다음과 같다.

1) 초등 교과서의 말뭉치를 구축한다.

2) 초등 교과서 말뭉치에서 나타나는 어휘의 통계적 분석을 한다.

3) 초등 교과서가 얼마나 일상생활에서 사용되는 어휘로 기술되어 있는지를 분석한다.

해당 연구의 목적을 달성하기 위한 연구내용은 다음과 같다.

1) 초등 교과서를 전사한고, 전사된 교과서를 말뭉치형태로 변환한다.

2) 초등 교과서 말뭉치에서 통계 수치를 구하고, 이를 분석한다.

3) 초등 교과서 말뭉치와 일반 말뭉치와의 비교를 통해 초등 교과서 말뭉치가 실생활의 어휘 분포와 얼마나 유사한지 분석한다.

2. 관련연구

해당 연구와 관련된 논문들은 말뭉치 기반을 통한 분석 연구와 교과서 내용을 사례별 분석하거나 인터뷰를 통한 분석을 하는 두 가지 형태의 연구들로 이루어진다.

먼저 말뭉치 기반 분석연구들은 아래와 같은 연구들이 이루어졌다.

Shin, Dongkwang은 초등 영어교과서와 중등영어교과서를 통계적으로 비교하였다[2]. 초,중,고등학교의 영어교과서의 어휘 통계들을 비교하였고, 세부적인 어휘리스트 순위 비교를 하였다.

박선호는 영어 아동 영화의 코퍼스 기반 초등영어 어휘 및 연어 분석과 자료집 개발 연구를 영어권 아동 영화의 어휘들과 다른 영어 코퍼스와의 비교를 통하여 진행하였다[3].

김혜영은 신문사설에 나타나는 어휘 사용의 추이를 물결 21 코퍼스를 활용하여 해당 코퍼스내의 신문사설을 분석하는 것으로 연구 하였다[4].

교과서 내용을 사례별 분석하거나 인터뷰를 통한 분석하는 형태의 연구들은 다음과 같이 이루어 졌다.

장재원은 중학교 1학년 영어 디지털 교과서의 사용성에 대한 연구가 진행되었다[5]. 평가에 휴리스틱한 평가에 초점이 맞추어 져있다.

정혜승은 국어과 교과서 채택기준과 채택한 교과서에 대한 교사의 반응을 교사들의 인터뷰를 통하여 연구하였다[6].

정재림은 문학교과서에서 나오는 문학 이론 및 개념의 문제점과 개선 방안을 교과서에서 나오는 사례들을 비교하는 것을 중점으로 연구하였다[7].

최유현은 초등학교 실과교과서의 초등기술교육 내용 분석과 미래 지향적 내용 구성 전략에 관한 연구를 내용별로 분석하고 이것을 교과서의 변화에 따라 분석하였다[8].

남가영은 학습자 오개념 형성 요인으로서의 교과서 연구를 중학교 국어 교과서 '단어형성법'에 관련된 단원을 가르치는 교사 인터뷰를 통하여 진행하였다[9].

3. 초등학교 교과서 말뭉치 연구

말뭉치언어학(corpus linguistics)은 '실제 언어' 혹은 실제 언어의 샘플을 이용하여 언어를 공부하는 응용언어학의 한 분야이다. 말뭉치(코퍼스)란, 언어를 연구하는 각 분야에서 필요로 하는 연구 재료로서 언어의 본질적인 모습을 총체적으로 드러내 보여줄 수 있는 자료의 집합을 뜻한다.

말뭉치는 몇 가지 요건을 갖추어야 하는데, 그 요건은 아래와 같다[10].

1) 텍스트 수집이나 입력 과정에서 원래의 내용이나 형태의 누락이 있어서는 안 된다. 즉 원형을 유지하고 있다는 보장이 필요하다.

2) 언어의 다양한 변이를 담아내야 한다. 즉 언어의 특성을 잘 반영할 수 있는 구성으로 조합되어야 한다.

3) 해당 언어의 통계적 대표성을 지녀야 한다. 즉 유의미한 규모로 확보되어야 한다.

말뭉치가 지녀야 하는 두 가지 특성은 '대표성'과 '균형성'이다. 대표성은 표본이 모집단을 통계적으로 대표할 수 있어야한다는 특성이고 균형성은 저 빈도로 나타날 수 있는 단어들도 균형 있게 말뭉치에 표현되어야 한다는 것이다.

또한 말뭉치는 가공 정도에 따라 원시 말뭉치(raw corpus), 주석 말뭉치(tagged / annotated corpus), 분석 말뭉치(analyzed corpus)로 나뉠 수 있다.

본 연구에서는 초등학교 교과서는 1-2학년, 3학년, 4학년, 5학년, 6학년의 교과서의 국어과, 수학과, 과학과, 사회과 교과 총 32권을 말뭉치의 기본 여건에 따라 구축하고, 가공을 원시 말뭉치와 주석 말뭉치, 그리고 통계적 자료까지 포함된 분석 말뭉치의 수준으로 구축 하여 본다.

3.1 초등학교 교과서 전사

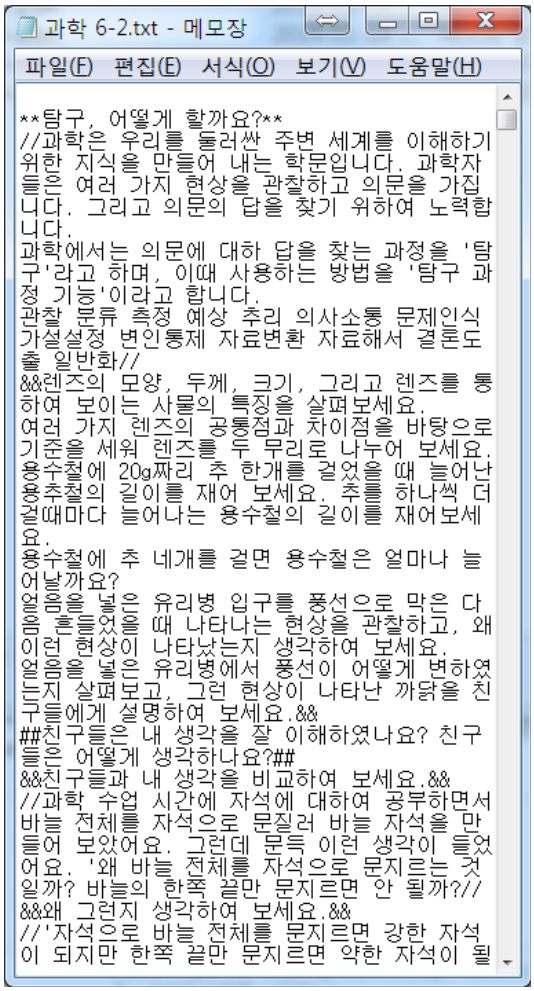
전사목표인 초등학교 교과서는 1-2학년, 3학년, 4학년, 5학년, 6학년의 교과서의 국어과, 수학과, 과학과, 사회과 교과 총 32권으로 이루어져있다. 1-2학년은 통합교과로 구성되어있는 국어, 수학, 가을, 겨울, 우리나라, 이웃의 6가지 교과서로 구성되어있고, 3학년에서 6학년의 교과는 말하기 듣기 쓰기, 읽기, 수학, 과학, 사회의 5가지 교과서로 구성되어있다.

교과서 전사는 단원 제목, 단원 목표, 단원 소목표, 문제 또는 지시문, 본문의 형태에서 나타나는 모든 텍스트를 전사하였다. 단원 제목, 단원 목표, 단원 소목표, 문제 또는 지시문, 본문을 구분하기 위하여 전사된 교과서는 <표 1> 과 같이 마크업 되어있다. 태그는 해당 교과서부분의 앞뒤를 감싸는 형태로 교과서 텍스트 앞뒤로 태그를 붙여 해당 부분을 구분 할 수 있도록 구성하였다. 실제 교과서 전사에 사용된 태그는 <표 1>과 같이 사용되었다.

<표 1> 교과서 전사 태그

교과서 부분	사용 태그	사용 예
단원 제목	**	**단원 제목**
단원목표	$$	$$단원목표$$

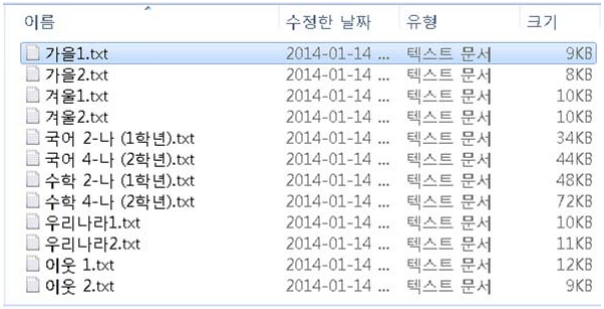
단원 소목표	&&	&&단원 소목표&&
문제(지시문)	##	##문제(지시문)##
본문	//	//text//



<그림 1> 교과서 전사 예시 - 6학년 과학교과

실제 전사된 교과서의 예시는 <그림 1>과 같다.

그림은 학년과 교과목을 나타내주는 파일명을 가지고 있고, 전사된 교과의 각 부분을 태그가 감싸고 있다.



<그림 2> 교과서 전사 텍스트 파일 예시 - 1,2학년 통합교과

전사된 교과서들은 <그림 2>와 같이 과목별로 텍스트 파일로 구성되어있다. 각 교과명의 텍스트 파일로 분리되어 구성되어있고, 학년별로 다시 분리되어있다.

3.2 어휘 분석 도구

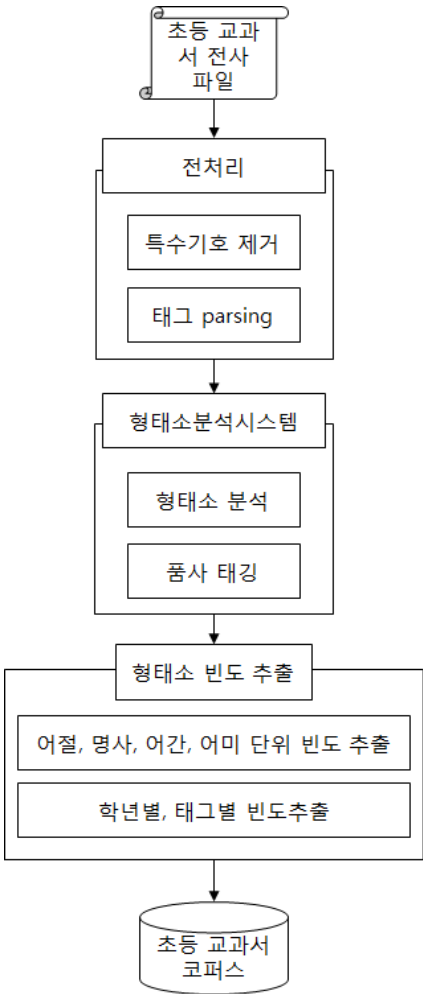
전사된 교과서를 어휘단위로 분석하기 위하여 교과서 어휘 분석 도구를 개발하였다.

어휘 분석 도구는 초등학교 교과서 전사 파일을 입력으로 받아 전처리, 형태소 분석 시스템, 분석된 형태소 빈도 추출로 구성되어있고, 동작의 흐름은 <그림 3>과 같다.

먼저 전처리 과정에서는 교과서 분석에서 필요없는 특수기호들을 제거하고, 전사된 교과서를 태그단위로 구조화한다. 특수기호 제거는 어휘 분석과 상관없는 특수기호를 전사된 교과서에서 제거하는 과정이다. 태그 parsing은 전사된 교과서를 형태소 분석한 이후에 태그별로 빈도를 추출하기 위하여 미리 텍스트를 구조화 하는 과정이다.

형태소 분석 시스템 과정에서는 전 처리된 교과서를 형태소 분석과 품사 태깅한다. 형태소 분석은 입력으로 들어온 문장단위의 텍스트를 형태소 단위로 parsing해 주는 과정이고 품사 태깅은 형태소 분석된 결과에 품사를 부착하는 과정이다.

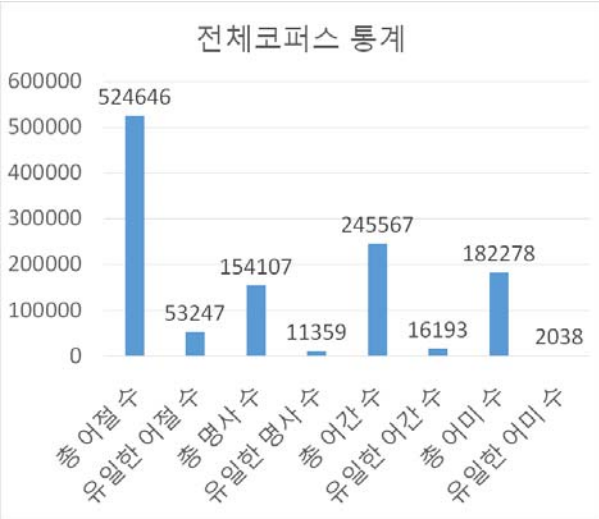
형태소 빈도 측정 과정에서는 형태소 분석된 결과를 어절, 명사, 어간, 어미 단위로 빈도를 추출하고 전처리 단계에서 구조화된 학년별, 태그별 빈도를 나누어 빈도를 추출한다.



<그림 3> 어휘 분석 도구 흐름도

<표 2> 초등학교 교과서 전체 빈도

카테고리		빈도
어절	총 어절 수	524646
	유일(unique)한 어절 수	53247
명사	총 명사 수	154107
	유일한 명사 수	11359
어간	총 어간 수	245567
	유일한 어간 수	16193
어미	총 어미 수	182278
	유일한 어미 수	2038



<그림 4> 초등학교 교과서 전체 빈도

해당 어휘 분석 도구의 분석결과는 학년별, 태그별로 분리된 어절, 명사, 어간, 어미 단위의 분석된 형태소와 빈도의 쌍으로 구성된다.

4. 어휘 분석 및 비교 분석

어휘 분석 도구에서 추출된 어휘빈도를 분석하여 초등학교 교과서의 빈도 통계 정보를 분석해 보고 일반인이 사용하는 언어로 구성된 세종코퍼스와 비교를 통하여 초등학교 교과서의 특성을 분석한다.

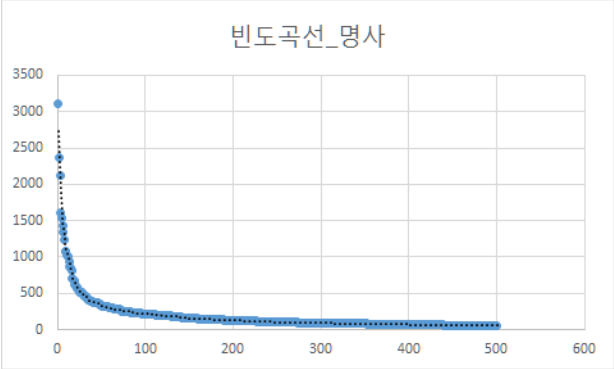
4.1 초등학교 교과서 전체 코퍼스 통계 정보

초등학교 교과서의 전체 빈도는 아래 <표2>와 같이 나타났고 이를 막대그래프로 살펴보면 <그림 4>와 같다.

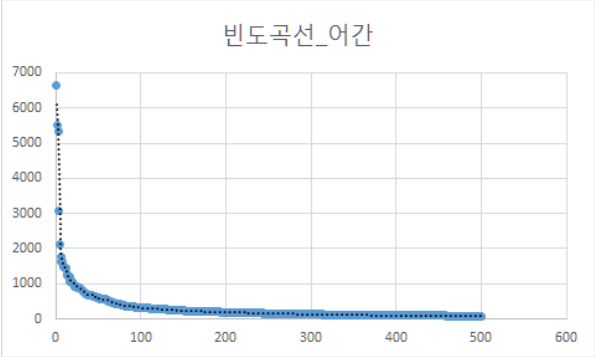
초등학교 교과서에서 나타나는 어휘들의 분포를 빈도곡선 형태로 살펴본다. 이것은 지프의 법칙(Zipf's law)에 따라 전체 초등학교 교과서의 빈도분포가 나타나는지를 확인할 수 있다. 지프의 법칙에 따르면 어떠한 자연어 말뭉치 표현에 나타나는 단어들을 그 사용 빈도가 높은 순서대로 나열하였을 때, 모든 단어의 사용 빈도는 해당 단어의 순위에 반비례한다. 따라서 가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높다. 이러한 특성으로 인간의 언어를 빈도곡선으로 나타내었을 때, 단어의 사용 빈도는 꼬리가 매우 긴 헤비 테일 분포를 보이고, 고빈도 단어의 일부가 전체 단어의 95% 이상을 차지하는 형태이다.

어휘의 빈도 측정 결과는 아래 <그림 5><그림 6> <그림 7> <그림 8>와 같다. 각각 상위 500위

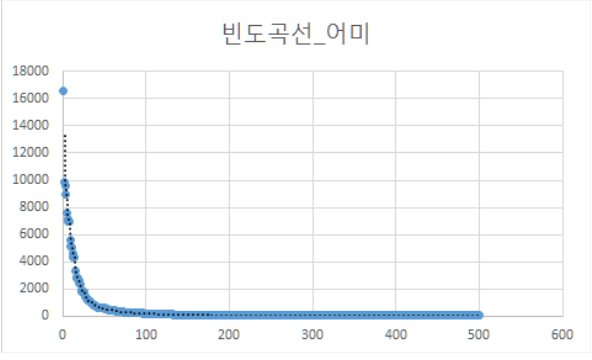
의 어휘 빈도를 측정하였으며 모두 반비례하는 추세선을 가지는 것이 관측되었다.



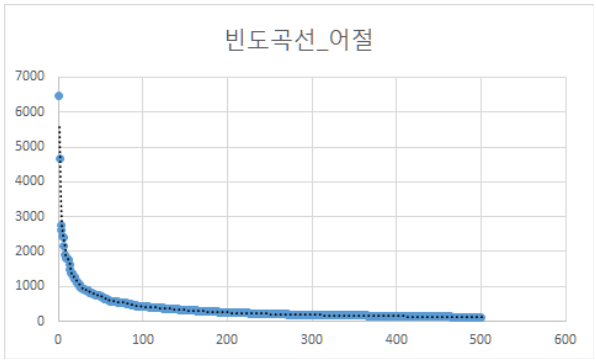
<그림 5> 초등학교 교과서 말뭉치의 명사 빈도곡선



<그림 6> 초등학교 교과서 말뭉치의 어간 빈도곡선



<그림 7> 초등학교 교과서 말뭉치의 어미 빈도곡선



<그림 8> 초등학교 교과서 말뭉치의 어절 빈도곡선

4.2 초등학교 교과서의 학년별 빈도

초등학교 교과서의 학년별 어절, 명사, 어간, 어미의 기본 빈도를 측정하였다. 초등학교 교과서의 학년별 어절 빈도는 아래 <표 3>과 같고, 명사 빈도는 <표 4>와 같다. 어간 빈도는 <표 5>와 같고, 어미 빈도는 <표 6>과 같다.

<표 3> 초등학교 교과서 학년별 어절 빈도

학년		빈도
1,2학년	총 어절 수	39,107
	유일한 어절 수	10,212
3학년	총 어절 수	81,948
	유일한 어절 수	19,751
4학년	총 어절 수	130,944
	유일한 어절 수	29,328
5학년	총 어절 수	194,785
	유일한 어절 수	42,412
6학년	총 어절 수	262,323
	유일한 어절 수	53,247

<표 4> 초등학교 교과서 학년별 명사 빈도

학년		빈도
1,2학년	총 명사 수	19,771
	유일한 명사 수	2,382
3학년	총 명사 수	24,553
	유일한 명사 수	3,009
4학년	총 명사 수	29,037
	유일한 명사 수	3,887
5학년	총 명사 수	39,243
	유일한 명사 수	5,300
6학년	총 명사 수	41,503
	유일한 명사 수	5,174

<표 5> 초등학교 교과서 학년별 어간 빈도

학년		빈도
12학년	총 어간 수	33,399
	유일한 어간 수	3,672
3학년	총 어간 수	40,195
	유일한 어간 수	4,486
4학년	총 어간 수	46,785
	유일한 어간 수	5,709
5학년	총 어간 수	60,764
	유일한 어간 수	7,632
6학년	총 어간 수	64,424
	유일한 어간 수	7,495

<표 6> 초등학교 교과서 학년별 어미 빈도

학년		빈도
12학년	총 어미 수	24,191
	유일한 어미 수	631
3학년	총 어미 수	30,143
	유일한 어미 수	760
4학년	총 어미 수	34,856
	유일한 어미 수	900
5학년	총 어미 수	45,651
	유일한 어미 수	1,084
6학년	총 어미 수	47,437
	유일한 어미 수	1,037

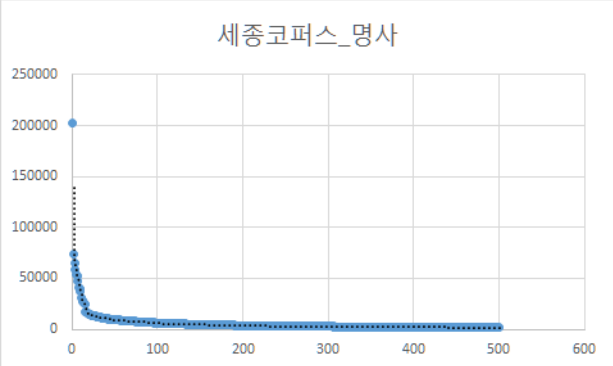
4.3 초등학교 교과서와 세종코퍼스와의 비교

초등학교 교과서 말뭉치와 세종 코퍼스 간의 상관관계 분석을 하였다. 초등학교 교과서 말뭉치와 비교는 스피어만의 순위 상관 계수(spearman's rank correlation coefficient) 방법을 사용하였다. 상관관계 분석이 필요한 X변인과 Y변인이 있을 때, 스피어만의 순위 상관 계수는 수식(1)과 같이 표현된다[11].

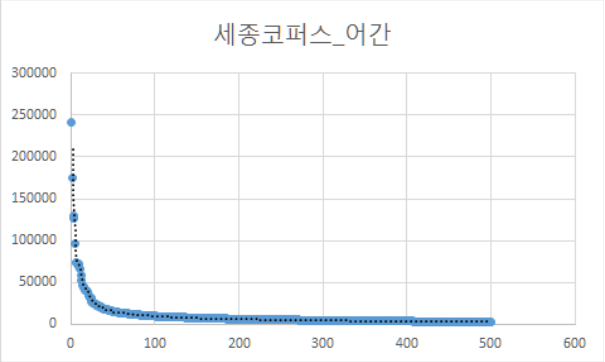
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (1)$$

di : X변인에서의 등위와 Y변인에서의 등위 간의 차이
n : X와 Y변인의 수치를 한 짝으로 한 사례수

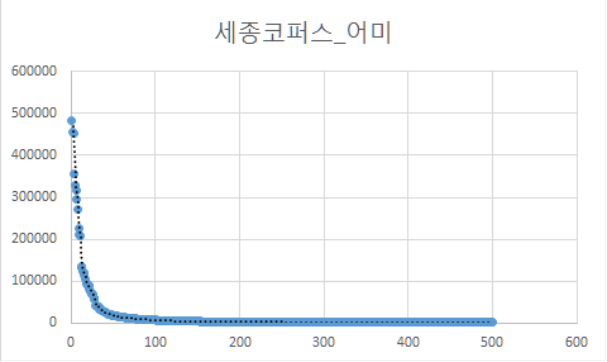
초등학교 교과서 말뭉치와 세종 코퍼스를 비교하기 전에 세종코퍼스의 기본적인 데이터 분포를 살펴보았다. 세종 코퍼스 말뭉치의 빈도곡선이 지프의 법칙을 따르는지를 확인하고, 초등 교과서 말뭉치와의 빈도곡선과 유사한 추세를 가지는지를 확인하였다. <그림 9>, <그림 10>, <그림 11>, <그림 12>는 세종코퍼스의 명사, 어간, 어미, 어절의 빈도곡선으로 지프의 법칙을 따르는 형태인 것을 확인할 수 있고, 추세선이 초등 교과서 말뭉치와 유사한 것을 확인하였다.



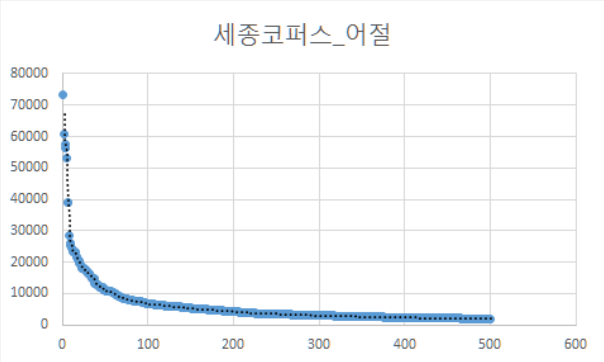
<그림 9> 세종코퍼스의 명사 빈도 곡선



<그림 10> 세종코퍼스의 어간 빈도 곡선



<그림 11> 세종코퍼스의 어미 빈도 곡선



<그림 12> 세종코퍼스의 어절 빈도 곡선

지프의 법칙대로 추세형태가 나타나는 상위 500위까지의 초등 교과서 말뭉치의 단어들을 대상으로 하여 세종코퍼스와 상관관계 분석을 실시하였다.

첫 번째 상관관계 분석은 상위 500위 순위 명사 군에서 중복되어 나타나는 명사 260개를 가지고 측정되었다.

분석결과 ρ는 0.44533이 측정되었다. 측정된 ρ의 수치로 초등 교과서 말뭉치와 세종 코퍼스간의 유사성이 있는지 추가로 t검정을 시행하였다. 검정결과 t는 7.988937로 유의수준 0.05에서 서로 관계가 있는 것으로 나타났다.

세종코퍼스와 중복되어 나타난 초등교과서 말뭉치의 명사들은 <표 7>과 같다.

<표 7> 세종코퍼스와 중복되어 나타난 고빈도 260개 초등교과서 말뭉치의 명사어휘

수,것,생각,말,사람,우리,나,때,방법,친구,무엇,일,이야기,활동,내용,글,다음,등,사용,이용,점,생활,물,가지,모습,그림,모양,개,문제,필요,나라,곳,우리나라,책,중,내,때문,마음,집,이,시간,발표,표현,자료,이름,인물,조사,위,세계,학생,부분,다양,속,학교,뜻,마을,지역,설명,해결,변화,선생님,소리,자신,안,번,사이,시,과학,거,누구,주장,뒤,경우,눈,데,사회,가족,얼마,자리,관계,엄마,결과,운동,중요,노력,반,아버지,문화,어머니,정보,소개,공부,너,앞,행동,나무,확인,동안,어디,과정,날,하나,손,조선,힘,몸,사건,아이,느낌,발전,위치,일본,준비,발생,시작,돈,지금,후,성격,어린이,꽃,열,거리,땅,명,전,역할,국민,도시,자연,길,예,이해,상황,식,이유,아저씨,그,계획,배,법,저,환경,불,이것,국가,정도,사랑,오늘,하늘,할머니,그것,자기,영향,관련,경험,바다,중심,해,밖,전쟁,정부,자,선거,말씀,원인,처음,기분,아빠,원,민족,신문,달,연구,옛날,관심,오늘날,개발,기관,노래,경제,참여,선택,뭐,세상,줄,회의,바람,삶,목소리,아래,광고,발,얼굴,결정,만,이상,하루,산,작품,판단,비,아침,전체,기술,년,머리,차례,고개,북
한,자유,정치,현상,소비자,중국,끝,대부분,발견,여성,다리,밤,미국,부족,사실,옷,건강,병,의미,장,쪽,당시,분,방향,적,제시,기준,기록,방,상태,생산,시대,가운데,경기,고려,걱정,반대,면,잠,표정,농민,문,이번,입,그때,대통령,질문,을,구성,기억,통일

세종코퍼스와 중복되어 나타나지 않은 초등교과서 말뭉치의 명사들은 <표 8>과 같다. 초등학교에서만 나타나는 명사들을 살펴보면 수학, 과학, 역사 부분에서 한 가지 혹은 그이상의 단원에 집중되어 나타나는 일련의 명사들이 많이 관찰되는 것을 확인할 수 있다. 예를 들어 관찰, 탐구, 실험, 태양, 지구, 편리와 같은 과학교과서에서 주로 쓰이는 단어들이나 도형, 선분, 부피, 크기와 같은 수학교과서에서 주로 쓰이는 단어들 그리고 백성, 촌락, 일제, 민주주의와 같이 사회교과서에서 쓰이는 단어들이다.

<표 8> 세종코퍼스와 중복되어 나타나지 않은 고빈도 240개 초등교과서 말뭉치의 명사어휘

까닭,정리,계산,비교,의견,관찰,활용,탐구,실험,동물,답,놀이,물건,규칙,문장,이동,딱지,물음,낱말,고장,물체,주제,부피,음식,공기,소수,지구,물질,모둠,길이,크기,주변,표,온도,기체,한글,특징,수단,종류,순서,도형,빛,양,할아버지,발달,이웃,식물,글쓴이,모형,상자,사진,카드,바탕,마무리,그림자,권리,준비물,분수,에너지,날씨,가을,학습,그래프,인구,무게,약속,누나,예상,근거,측정,겨울,장소,글자,면담,숫자,사회과,백성,태양,인터넷,보호,산화,전달,동생,축하,화산,원기둥,산소,성질,완성,의사소통,소,파악,소금,속력,교실,체험,주의,시각,선분,액체,촌락,화석,얼음,학급,탄소,색깔,뉴스,넓이,콩,오른쪽,사과,우주,행성,제목,우리말,여행,변,일제,부탁,피해,도움,세로,풍선,토끼,쌀,과학자,장면,가로,비커,기후,마리,독도,자동차,편지,새엄마,생각열기,단위,과자,직선,선물,새,편리,이반,생김새,값,동전,기구,자전거,실천,놀이터,연결,종이,부모님,여러분,안전,갈등,짝,색칠,직업,표시,송아지,분리,현장,지층,높이,아프리카,도깨비,이때,전개,인권,남극,화,제안,상상,직사각형,서양,총각,대한민국,지진,그릇,전통,선,키,민주주의,잎,윌버,어림,단원,버스,구멍,등장인물,의도,고양이,테이프,독서,토의,파란색,토론,특성,별자리,암석,칭찬,돌이,국제,주요,혼합물,색,물고기,합,분류,돼지,시계,혈액형,인형극,초,시설,도착,가게,자연환경,정확,개수,유리,호응,조상,눈금,열기,상품,연소,병태,막대,대왕,영감,컵,칸,방정식,기념일,지도,공원,공룡,금

두 번째 상관관계 분석은 상위 500위 순위 어간 군에서 중복되어 나타나는 어간 294개를 가지고 측정되었다.

분석결과 ρ는 0.51788이 측정되었다. 측정된 ρ의 수치로 초등 교과서 말뭉치와 세종 코퍼스간의 유사성이 있는지 추가로 t검정을 시행하였다. 검정결과 t는 10.34487로 유의수준 0.05에서 서로 관계가 있는 것으로 나타났다.

세 번째 상관관계 분석은 상위 500위 순위 어미 군에서 중복되어 나타나는 어미 499개를 가지고 측정되었다.

분석결과 ρ는 0.8813431이 측정되었다. 측정된 ρ의 수치로 초등 교과서 말뭉치와 세종 코퍼스간의 유사성이 있는지 추가로 t검정을 시행하였다. 검정결과 t는 20792.87로 유의수준 0.05에서 서로 관계가 있는 것으로 나타났다.

5. 결론

본 연구에서는 초등교육에서 실제 사용하는 교과서의 텍스트에 담겨있는 어휘들을 분석하기 위하여 초등 교과서를 전사하고 말뭉치 형태로 만들고, 초등 교과서 말뭉치에서 나타나는 어휘의 분포를 빈도별로 분석하고, 초등 교과서에서 나타나는 어휘 분포와 일반 말뭉치에서 나타나는 어휘의 분포를 비교해 보는데 그 목적이 있다. 이를 위하여 교과서를 전사하고, 교과서를 태그가 달린 형태의 말뭉치로 변환하고, 빈도를 추출하였다. 또한 어휘들의 빈도곡선을 살펴봄으로써 교과서의 어휘가 일반적인 어휘통계의 빈도 곡선과 유사한 형태를 보임을 보였으며, 세종 코퍼스와 고빈도 어휘들 간의 유사성을 비교해 보았다. 본 연구의 실험결과를 종합하여 볼 때, 초등교과서의 전체 어휘는 일반적인 말뭉치 통계에서 나타나는 경향성을 그대로 가지고 있지만, 특정 교과들에서 중점적으로 나타나는 고빈도 어휘들은 일반말뭉치에서는 고빈도로 발견되지 않는다.

본 연구는 초등교과서 전체를 말뭉치형태로 전사하고 통계자료를 추출하는 형식의 연구는 국내에서 처음 시도되어지는 연구이며, 본 연구에서 제작한 어휘, 분류 통계들은 초등 교과서를 분석하고 교육하는데 실질적인 도움이 될 것이다.

앞으로의 연구에서는 분석된 자료를 웹에 공개, 검색할 수 있는 플랫폼을 만들고, 해당 자료와 연구된 자료들을 함께 모으고 커뮤니케이션 할 수 있는 온라인 시스템 개발을 목표로 한다.

참 고 문 헌

[1] doopedia 두산백과, 초등교육의 목적 및 내용

[2] S Dongkwang, YV Chon (2010). A Corpus based analysis of curriculum based elementary And Secondary english textbooks, *Multimedia-Assisted Language Learning,* Vol.14 No.1, 2011.4, 149-175 (27 pages)

[3] 박선호, 홍성준. 영어 아동영화의 코퍼스 기반 초등영어 어휘 및 연어 분석과 자료집 개발. English Education, 18(1), 309-338. PrimaryEnglishEducation Vol.18, No.1

[4] 김혜영, 강범모. (2013). 신문 사설에 나타나는 어휘 사용의 추이-[물결 21 코퍼스] 를 활용하여. 텍스트언어학, 35(단일호), 1-22.

[5] 장재원, 김보연. (2012). 국내 디지털 교과서의 사용성 평가. 디지털디자인학연구, 12(2), 429-438.

[6] 정혜승. (2011). 국어과 교과서 채택 기준과 채택한 교과서에 대한 교사의 반응. 독서연구, (25), 347-383.

[5] 정재림. (2012). 문학교과서에 나오는 문학 이론 및 개념의 문제점과 개선 방안 (2). 한국학연구, 40, 253-274.

[8] 최유현. (2001). 초등학교 실과교과서의 초등 기술교육 내용 분석과 미래 지향적 내용 구성 전략. 한국실과교육학회지, 14(2), 21-39.

[9] 남가영. (2013). 학습자 오개념 형성 요인으로서의 교과서. 우리말글, 57, 109-137.

[10] 서상규, 한영균 (1999). 국어정보학 입문, 태학사

[11] Spearman, C. (1904). The proof and measurement of association between two things. The American journal of psychology, 15(1), 72-101.

유 원 희

2007 한신대학교
소프트웨어학과(학사)
2009 한신대학교
컴퓨터학과(석사)
2009~현재 고려대학교 컴퓨터교육과
박사과정
관심분야: 컴퓨터교육, 인공지능, 자연어처리
E-Mail: galadous@naver.com

임 희 석

1992 고려대학교
컴퓨터학과(학사)
1994 고려대학교
컴퓨터학과(석사)
1997 고려대학교 컴퓨터학과(박사)
2008~현재 고려대학교 컴퓨터교육학과 교수
관심분야: 자연어처리, 인공지능, 컴퓨터교육,
Educational data mining
E-Mail: limhseok@korea.ac.kr