https://doi.org/10.5392/JKCA.2020.20.07.274

남북한 고등학교 영어교과서 4-gram 연어 비교 분석

Comparative Analysis of 4-gram Word Clusters in South vs. North Korean High School English Textbooks

김정렬
한국교원대학교 초등교육과

Jeong-ryeol Kim(jrkim@knue.ac.kr)

요약

본 연구는 4-gram 연어분석으로 남북한 고등학교 영어교과서를 비교분석하고자 하는 것이 목적이다. N-gram 분석은 그동안 우리가 알고 있는 관습적인 관용어와는 달리 코퍼스를 구성하여 기계적인 방법으로 물리적으로 함께 공기하는 빈도가 높은 낱말군을 객관적인 방법으로 추출하여 분석하는 것이다. 본 연구의 목적은 AntConc의 N-gram 분석 도구로 4-gram 연어를 남북한 영어교과서 코퍼스에서 찾아서 비교 분석해 보는 것이다. 분석의 대상은 북한의 2013 교육개혁에 따른 북한 고등중학교 영어교과서와 남한의 2015교육과정에 따른 고등학교 영어교과서로 구성된 코퍼스에서 구어와 문어의 token과 type을 구분하여 분석 비교한다. 이를 분석대상으로 하여 코퍼스의 4-gram 연어를 문법범주와 기능범주로 나눈 준거를 통해서 분석하였다. 문법범주는 크게 명사구, 동사구, 전치사구, 부분절 그리고 기타로 나누어 범주화하고 기능범주는 지칭, 텍스트의 조직, 입장과 기타로 나누었다. 분석한 결과 4-gram 연어에 나타난 구어와 문어 모두 남한의 영어교과서가 북한의 영어교과서 보다 token과 type의 수가 상대적으로 많았다. 그리고 문법범주에는 남북한 모두 영어교과서에 동사구와 부분절 형태의 4-gram 연어가 가장 많았으며 기능범주에는 남북한 모두 영어교과서에 입장 기능과 관련된 4-gram 연어가 가장 많았다.

■ 중심어 : | 고등학교 영어 | 북한 영어교과서 | N-gram 분석 | 남북한 영어 비교 | 4-gram 키워드 | 영어교과서 비교

Abstract

N-gram analysis casts a new look at the n-word cluster in use different from the previously known idioms. It analyzes a corpus of English textbooks for frequently occurring *n* consecutive words mechanically using a concordance software, which is different from the previously known idioms. The current paper aims at extracting and comparing 4-gram words clusters between South Korean high school English textbooks and its North Korean counterpart. The classification criteria includes number of tokens and types between the two across oral and written languages in the textbooks. The criteria also use the grammatical categories and functional categories to classify and compare the 4-gram words clusters. The grammatical categories include noun phrases, verb phrases, prepositional phrases, partial clauses and others. The functional categories include deictic function, text organizers, stance and others. The findings are: South Korean high school English textbook contains more tokens and types in both oral and written languages. Verb phrase and partial clause 4-grams are grammatically most frequently encountered categories across both South and North Korean high school English textbooks. Stance is most dominant functional category in both South and North Korean English textbooks.

■ keyword : | High School English | North Korean English Textbook | N-gram Analysis | Comparison of South and North Korean English Education | 4-gram | Comparison of English Textbook |

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A5A2A01037209)
접수일자 : 2020년 05월 07일
수정일자 : 2020년 06월 18일
심사완료일 : 2020년 06월 21일
교신저자 : 김정렬, e-mail : jrkim@knue.ac.kr

I. 서 론

언어 사용자들은 코퍼스 기반 기술적 분석 중의 하나로서 N-gram의 패턴으로 나타나는 동일 totken 연결체를 언어에 대한 다빈도 노출이라는 관점에서 언어의 구조적 단위와는 상관없이 많이 접하게 된다. 따라서 이와같은 동일 token 연결체는 원어민이라면 자연스럽게 습득하게 되는 어휘의 다발(Gray and Biber, 2013; Wray, 2002; Hyland, 2012; Koprowski, 2005)로서 연어라고 한다[1-4]. 학자에 따라서 이를 연속어휘다발이라고 표현하기도 했으나[5] 본고에서는 일반적인 용어로서 N-gram을 그대로 쓰기로 한다. 영어학습자로서 N-gram의 중요성은 다빈도어가 어휘학습에서 중요하듯이 다빈도 N-gram 연어목록 또한 영어 표현에서 중요하게 학습되어져야 한다는 것이다. 실제로 Quin과 Hyland는 각각의 연구에서 독자적으로 N-gram 분석에서 외국어로서 영어 학습자 쓰기 코퍼스에 나타난 목록과 원어민의 N-gram 목록 사이에 많은 차이가 있다는 것을 발견했다[6][7]. 따라서 영어교과서에 나타난 N-gram 분석은 영어교과서 표현의 진정성을 나타낼 수 있는 잣대로 볼 수 있다.

영어교과서는 외국어로서 영어 학습을 할 때에 노출의 원천이고 영어교과서에 나타나는 N-gram의 빈도분석은 학생들의 N-gram 노출빈도와 유사할 것이라는 추정은 여러 연구에서 입증되었다[8][9]. 아울러 영어교재에 나타난 N-gram의 사용이 인위적이라서 언어의 진정성이 많이 떨어진다는 비판도 있었다[5][10]. Wood가 지적한대로 교과서 제작시에 어휘의 빈도수는 중요하게 다루어서 비교적 진정성 있는 빈도순으로 교과서에서 다루지만[11] N-gram의 경우는 교과서에서 진정성 있는 실제 언어자료의 빈도를 교과서에 반영하는 경우는 드물다는 것이다. N-gram의 종류와 관련해서 홍신철은 N-gram을 phraseology-group, lexical bundle-group, skipgram-group의 3가지 그룹으로 구분하였으나[5] 본 연구에서는 lexical bundle-group으로 정의된 어휘연속체로서 N-gram을 대상으로 N-gram 어휘연속체(연어)라고 한다. 다만 연어를 collocation으로 혼동할 수 있으므로 본 연구에서는 연어를 collocation이 아닌 어휘연속체를 연어라고 정의한다.

어휘 연속체로서 N-gram은 영어교과서에 여러 단원에서(범위: range) 여러 차례에(빈도: frequency) 걸쳐서 반복하여 나타나는 경우에 영어를 배우는 학생에게 습득의 기회를 높일 수 있다. 이러한 관점에서 남북한 영어교과서에 나타난 어휘의 비교와 함께 N-gram의 분석을 통해서 학습하는 어휘연속체의 차이가 어떤 것들이 있는지 파악해 보는 것은 남북한 영어교육의 비교를 위해서 의미있는 일이다. 따라서 본고에서는 첫째, 남북한 영어교과서 코퍼스에 나타난 N-gram 어휘연속체의 목록의 양적 비교를 한다. 둘째, 남북한 영어교과서 코퍼스에 나타난 N-gram 어휘연속체의 문법적, 기능적으로 나누어서 분포 양상을 비교분석한다.

II. 연구방법

1. 분석대상

남한의 2015교육과정에서 출간되어서 사용되는 인정교과서 가운데서 널리 채택되어서 쓰이고 있는 Y사 고등학교 영어, 영어 I, 영어 II 3권의 검정교과서와 북한의 2013 신교육강령에 따라서 집필된 고급중학교 영어교과서 1, 2, 3학년 교과서를 코퍼스로 구성해서 이들 고등학교 영어교과서 코퍼스에 나타난 N-gram을 추출해서 비교하고자 한다. 좀 더 구체적으로 분석비교대상 남북한 고등학교 영어교과서를 구분해서 학년, 교과서내의 Token과 Type수를 조사해서 남북한을 구분하고 학년별로 고등학교 영어교과서를 제시하면 [표 1]과 같다.

표 1. 남북한 고등학교 영어교과서 Token과 Type의 양적 비교

구분	학년	Token수			Type수		
		전체	구어	문어	전체	구어	문어
남한 고등학교 영어 교과서	1	36317	10895	25422	3584	1075	2059
	2	39235	11770	27465	3769	1101	2638
	3	42374	12712	29662	3997	1199	2798
북한 고등중학교 영어 교과서	1	21983	6595	15388	2361	708	1653
	2	22793	6838	15955	2513	754	1759
	3	24076	7223	16853	2843	853	1990

남한의 고등학교 영어교과서는 1학년 교과서에 나타

난 Token수가 36317개, Type수는 3584개이고, 2학년 교과서에 나타난 Token수가 39235개, Type수는 3769개이며, 3학년 교과서에 나타난 Token수가 42374개, Type수는 3997개이다. 전체 Token을 듣기, 말하기를 포함하는 구어 영역과 읽기와 쓰기를 포함하는 문어 영역으로 나누어서 살펴보면 남한의 고등학교 영어교과서는 1학년 구어 영역 Token수는 10895이고 문어 영역 Token수는 25422개이다. 2학년 구어 영역 Token수는 11770이고 문어 영역 Token수는 27465개이고 3학년 구어 영역 Token수는 12722이고 문어 영역 Token수는 29662개이다. Type수를 살펴보면 1학년 구어 영역 Type수는 1075이고 문어 영역 Type수는 2059개이다. 2학년 구어 영역 Type수는 1101이고 문어 영역 Type수는 2638개이고 3학년 구어 영역 Type수는 1199이고 문어 영역 Type수는 2798개이다. 학년이 올라갈수록 Token과 Type수는 꾸준히 증가하지만 Type-Token 비율은 1학년이 0.098, 2학년이 0.096이고 3학년이 0.094로 조금씩 낮아지는 것을 발견하였다.

북한의 고등중학교 영어교과서는 1학년 교과서에 나타난 Token수가 21983개, Type수는 2361개이고, 2학년 교과서에 나타난 Token수가 22793개, Type수는 2513개이며, 3학년 교과서에 나타난 Token수가 24076개, Type수는 2843개이다. 전체 Token을 듣기, 말하기를 포함하는 구어 영역과 읽기와 쓰기를 포함하는 문어 영역으로 나누어서 살펴보면 북한의 고등중학교 영어교과서는 1학년 구어 영역 Token수는 6595이고 문어 영역 Token수는 15388개이다. 2학년 구어 영역 Token수는 6838이고 문어 영역 Token수는 15955개이고 3학년 구어 영역 Token수는 7223이고 문어 영역 Token수는 16853개이다. Type수를 살펴보면 1학년 구어 영역 Type수는 708이고 문어 영역 Type수는 1653개이다. 2학년 구어 영역 Type수는 754이고 문어 영역 Type수는 1759개이고 3학년 구어 영역 Type수는 853이고 문어 영역 Type수는 1990개이다. 학년이 올라갈수록 Token과 Type수는 꾸준히 증가하지만 Type-Token 비율은 1학년이 0.098, 2학년이 0.096이고 3학년이 0.094로 조금씩 낮아지는 것을 발견하였다. 북한 고등중학교 영어교과서의 경우도 전반적으로 남한의 영어교과서에 비해서 Token과 Type수가 적지만 학년이 올라갈수록 Token과 Type의 수가 증가하는 것은 공통적으로 관찰할 수 있었다. 북한 고등중학교 영어교과서의 학년별 Type-Token 비율은 1학년이 0.107, 2학년이 0.11이고 3학년이 0.118로 남한의 영어교과서와 달리 학년이 올라가면서 조금씩 높아지는 것을 발견하였다. 이와 더불어 황서연, 김정렬이 지적한대로 북한의 영어교과서는 어휘의 반복율이 남한의 영어교과서에 비해서 떨어진다는 것을 알 수 있다[12].

2. 분석준거

N-gram 분석을 통한 연어 연구에서 기준 값을 영어의 경우는 보통 4-gram으로 설정한다[5][16]. 다시 말해 4개의 연속어가 일정한 횟수 이상 반복하여 나타날 때 이를 일단 1차적인 연어 추천 목록에 담는다. 일정한 횟수를 어떻게 설정하느냐는 출현 빈도의 밀집도에 따라서 다를 수 있지만 일반적으로 Biber외 4인, Chen과 Baker가 제안한 100만 단어 당 25회를 기준 횟수로 설정한다[13][14]. 이러한 기준을 우리나라 초등학교 영어교과서에 적용한 홍신철에 따르면 이 기준을 적용할 때에 교과서에 나타난 총 token수에 25를 곱해서 1,000,000 단어를 나누면 기준 횟수가 된다[5]. 이를 기준으로 보면 남한의 고등학교 영어교과서는 1학년 교과서에 나타난 Token수가 36317개, 2학년 교과서에 나타난 Token수가 39235개, 3학년 교과서에 나타난 Token수가 42374개로서 모두 합치면 Token수가 117926이다. 북한의 고등중학교 영어교과서는 1학년 교과서에 나타난 Token수가 21983개, 2학년 교과서에 나타난 Token수가 22793개, 3학년 교과서에 나타난 Token수가 24076개로서 모두 합치면 Token수가 모두 68852개이다. 따라서 일반적으로 4-gram 분석에서 통용되는 1,000,000 Token당 25회는 남한 영어교과서의 경우는 나타나는 4-gram Token수가 2.95회 정도이면 1차 4-gram 어휘연속체 목록에 들어가고 북한 영어교과서의 경우는 나타나는 4-gram의 Token수가 1.72회 정도로 나타나면 4-gram 어휘연속체 목록에 들어갈 수 있는 것으로 나타났다. 따라서 본고에서는 남한 영어교과서의 4-gram 어휘연속체가

3회이상 나타나는 경우와 북한 영어교과서의 경우는 4-gram어휘연속체가 2회이상 나타나는 경우로 1차 목록의 준거를 설정하였다.

그리고 이들 연어목록을 분류하는 기준은 Biber, Conrad와 Cortes의 분류 기준을 응용해서 다음과 같은 기준을 설정한다[15]. 문법적인 분류는 명사구, 전치사구, 동사구, 부분절, 기타로 구분하고, 기능적인 분류는 지칭어(시간, 장소, 기술, 양화), 텍스트 조직어(비교, 초점, 프레임, 토픽), 입장(인식, 의무, 가능)과 기타로 나눈다. 다시 말해 연어를 문법적인 구조와 관련된 형태적 분류와 이들 연어들의 대화나 읽기 지문에서 담당하는 언어적 역할 즉 기능적 분류로 크게 나누어서 정리하여 비교 분석한다.

Ⅲ. 연구결과

남한의 고등학교 영어교과서와 북한의 고등중학교 영어교과서를 비교해보면 우선 전체적인 구성이나 체제가 북한의 2013 교육개혁 이후에 많이 비슷해졌다는 것을 알 수 있다. 2013 교육개혁 이전에는 북한의 고등중학교 영어교과서는 문법과 읽기 위주의 교과서여서 남한의 고등학교 교과서와 듣기 말하기 영역은 직접적인 비교가 힘들었다. 그동안 제1중학교와 같은 수재학교를 제외하고 북한의 중학교 영어교과서에 나온 구어 영어를 남한의 영어교과서에 나온 내용과 직접 비교하는 것은 처음 시도하는 일이 될 것으로 판단된다. 따라서 이런 의미를 살리고 실제로 구어 영역과 문어 영역을 비교하여 4-gram연어의 양상이 어떻게 달리 나타나는지를 파악하기 위해서 구어와 문어를 분리해서 남북한 고등학교 영어교과서를 서로 비교해보고자 하였다. 연구결과의 구성은 먼저 남북한 영어교과서에 나타난 4-gram 연어의 양적비교를 구어와 문어 영역별로 살펴보고 연구방법에서 제시했던 문법적 분류와 기능적 분류로 나누어서 이들을 좀 더 자세히 비교해 본다.

1. 남북한 영어교과서에 나타난 4-gram 목록의 양적 비교

남북한 영어교과서에 나타난 4-gram 연어를 학년별로 구어와 문어를 분리해서 양적으로 비교하면 [표 2]와 같다.

표 2. 남북한 고등학교 4-gram 연어의 양적 비교

	남한 고등학교 영어 교과서			북한 고등중학교 영어 교과서		
학년	구어 빈도 %	문어 빈도 %	전체빈도 %	구어 빈도 %	문어 빈도 %	전체 빈도 %
1	142 34.05	64 30.19	206 32.8	57 30.81	76 31.80	133 31.37
2	137 32.85	73 34.43	210 33.4	61 32.97	79 33.05	140 33.02
3	138 33.10	75 35.38	213 33.8	67 36.22	84 35.15	151 35.61
전체	417 100	212 100	629 100	185 100	239 100	424 100

남한 고등학교 영어 교과서의 경우 4-gram 연어의 수가 1학년은 구어 영역에서 142개(전체 학년중 34.05%), 문어 영역에서 64개(전체 학년중 30.19%)이고 이들을 합해서 전체로는 206개였다. 북한 고등중학교 영어 교과서의 경우 1학년은 구어 영역에서 57개(전체 학년중 30.81%), 문어 영역에서 76개(전체 학년중 31.80%)이고 이들을 합해서 전체로는 133개였다. 남한 고등학교 2학년은 구어 영역에서 137개(전체 학년중 32.85%), 문어 영역에서 73개(전체 학년중 34.43%)이고 이들을 합해서 전체로는 210개였다. 북한 고등중학교 2학년은 구어 영역에서 61개(전체 학년중 32.97%), 문어 영역에서 79개(전체 학년중 33.05%)이고 이들을 합해서 전체로는 140개였다. 남한 고등학교 영어 교과서의 경우 4-gram 연어의 수가 3학년은 구어 영역에서 138개(전체 학년중 33.10%), 문어 영역에서 75개(전체 학년중 35.38%)이고 이들을 합해서 전체로는 213개였다. 북한 고등중학교 영어 교과서의 경우 3학년은 구어 영역에서 67개(전체 학년중 36.22%), 문어 영역에서 84개(전체 학년중 35.15%)이고 이들을 합해서 전체로는 151개였다. 남한 고등학교 영어교과서 1, 2, 3학년 전체로는 구어의 경우 417개이고 문어는 212개이고 전체는 629개였다. 북한 고등중학교 영어교과서 1, 2, 3학년 전체로는 구어 185개이고 문어는 239개이고 전체로는 424개였다.

전체적으로 2013 교육개혁 이후에 북한 영어교과서의 단원수와 신출 어휘수가 많이 줄고 대신에 연습 활동이 많이 들어갔기 때문에 양적인 면에서 남한의 영어

교과서에 나타난 token과 type의 수가 북한 영어교과서보다 많다는 것을 감안하고 [표 2]를 살펴볼 필요가 있다. 눈에 띄는 특징은 우선 남한의 영어교과서에는 4-gram 연어가 구어영역에 더 많이 나타나는데 북한의 영어교과서에서는 4-gram 연어가 문어영역에 상대적으로 더 많이 나타난다. 이와 같은 현상은 북한의 영어교과서에서 문어영역에는 문법적 교수요목에 의해서 특정 언어형식에 따른 표현이 많이 삽입되어 있고 구조적인 지시어가 많이 들어 있기때문으로 판단된다. 예를 들면 고등중학교 2학년 영어교과서에 나오는 4-gram 연어인 다섯 번에 걸쳐서 나오는 *It's a good idea to* (북2), *whether the following statements (are)* (북2)과 3학년 영어교과서에 여섯 번에 걸쳐서 나타나는 *read the sentences and* (북3), *and complete the rules* (북3)과 같은 4-gram 연어들이 북한 영어교과서 문어영역에서 발견되는 표현들이다. 이에 반해 남한의 고등학교 영어교과서 문어 영역에서 나오는 표현들은 읽기 내용 속에서 반복적으로 나오는 것들로 2학년 교과서에 네 번에 걸쳐서 나타나는 *(is) the only way to (practice)* (남2)와 3학년 교과서에 네 번에 걸쳐서 읽기 내용 중에 나타나는 *the most tragic thing (about poverty)* (남3)와 3학년 교과서에 세 번 나타나는 *will (always) end up as* (남3) 등이 있었다. 구어영역에서 남북한 영어교과서에 나타나는 4-gram 연어의 특이점은 없었고 *I'd like to tell you* 는 남북한 영어교과서 구어영역에서 가장 많이 등장하는 가장 긴 연어였다.

2. 남북한 영어교과서에 나타난 4-gram 연어의 문법 범주별 비교

남북한 영어교과서에 나타난 4-gram 연어를 문법범주별로 명사구, 전치사구, 동사구, 부분절과 기타로 나누어서 학년 구분없이 전체적으로 구어와 문어를 구분하고 각 구어와 문어에 대해서 token과 type의 %를 나누어서 제시하면 [표 3]과 같이 나타낼 수 있다.

표 3. 남북한 영어교과서의 문법 범주별 4-gram 연어 양적 비교

문법범주	남한 고등학교 영어교과서				북한 고등중학교 영어교과서			
	구어		문어		구어		문어	
	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)
명사구	11.5	12.4	12.2	11.3	15.7	16.2	15.3	13.9
전치사구	7.1	6.1	9.8	8.1	6.9	7.3	8.7	8.1
동사구	31.4	35.2	33.3	34.2	31.9	33.2	34.4	35.7
부분절	37.3	36.0	38.3	35.2	35.7	34.1	35.2	33.1
기타	12.7	10.3	6.4	11.2	9.8	9.2	6.4	9.2

남한 고등학교 영어교과서에 나타난 명사구는 구어 token의 경우 11.5%이고 type의 경우 12.4%이고 문어는 token의 경우 12.2%이고 type의 경우 11.3%였다. 북한 고등중학교 영어교과서에서는 명사구는 구어 token의 경우 15.7%이고 type의 경우 16.2%이고 문어는 token의 경우 15.3%이고 type의 경우 13.9%였다. 남한 고등학교 영어교과서에 나타난 전치사구는 구어 token의 경우 7.1%이고 type의 경우 6.1%이고 문어는 token의 경우 9.8%이고 type의 경우 8.1%였다. 북한 고등중학교 영어교과서에서는 전치사구는 구어 token의 경우 6.9%이고 type의 경우 7.3%이고 문어는 token의 경우 8.7%이고 type의 경우 8.1%였다. 남한 고등학교 영어교과서에 나타난 동사구는 구어 token의 경우 31.4%이고 type의 경우 35.2%이고 문어는 token의 경우 33.3%이고 type의 경우 34.2%였다. 북한 고등중학교 영어교과서에서는 동사구는 구어 token의 경우 31.9%이고 type의 경우 33.2%이고 문어는 token의 경우 34.4%이고 type의 경우 35.7%였다. 남한 고등학교 영어교과서에 나타난 부분절은 구어 token의 경우 37.3%이고 type의 경우 36.0%이고 문어는 token의 경우 38.3%이고 type의 경우 35.2%였다. 북한 고등중학교 영어교과서에서는 부분절은 구어 token의 경우 35.7%이고 type의 경우 34.1%이고 문어는 token의 경우 35.2%이고 type의 경우 33.1%였다.

남북한 영어교과서 모두 동사구와 부분절에 나타나는 4-gram 연어의 비율이 상대적으로 명사구와 전치사구에 비해서 큰 차이로 양이 많았다. 이러한 양의 차이는 4-gram 연어의 많은 양이 동사를 포함하여 이루어졌다는 것을 방증한다고 볼 수 있다. 남북한 영어교과서 4-gram 연어에 나타난 차이로 눈에 띄는 것은 구

어와 문어를 막론하고 북한 영어교과서에 나타난 명사구 연어의 4-gram 연어의 비중이 높다는 것이다. 이러한 이유는 북한 영어교과서의 경우 김정렬의 연구에서 밝힌대로 문법용어, 학술어휘와 정치나 사상적 용어의 반복에서 기인된다고 볼 수 있다[17]. 북한 영어교과서에 등장하는 명사구 4-gram 연어의 예를 들어 보면 고등중학교 영어교과서에 아홉 번 등장하는 *words in the box*, 네 번 *relative clause with when/where, the verbs in brackets*, 세 번 등장하는 *phrases in the box* 와 같은 표현들이 있었다.

3. 남북한 영어교과서에 나타난 4-gram 연어의 기능적 범주별 비교

지칭어, 텍스트 조직어, 입장, 기타로 나누어진 표현의 기능적 범주에 따라서 남북한 영어교과서에 나타난 4-gram 연어를 구분하고 이를 구어와 문어 그리고 각각에 속한 token과 type에서 차지하는 비율을 살펴보고 이를 기술하고 분석하고 해석하고자 한다. 남북한 영어교과서 4-gram 연어의 기능적 분포를 지칭어, 텍스트 조직어, 입장과 기타로 나뉘어진 기능적 범주와 남한 고등학교 영어교과서와 북한 고등중학교 영어교과서를 각각 구어와 문어로 나누고 이들을 다시 token으로 각 기능범주별로 차지하는 비율과 type으로 각 기능범주별로 차지하는 비율을 구분해서 표로 나타내면 [표 4]와 같다.

표 4. 남북한 영어교과서의 4-gram 연어의 기능적 분포 비교

기능범주	남한 고등학교 영어교과서				북한 고등중학교 영어교과서			
	구어		문어		구어		문어	
	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)
지칭어	12.3	13.7	23.7	24.3	14.2	13.9	27.6	28.6
텍스트 조직어	7.6	7.9	12.6	13.2	7.7	8.7	13.7	14.1
입장	31.8	32.9	28.7	27.6	33.2	32.9	29.1	30.2
기타	48.3	45.5	35.0	34.9	44.9	44.5	29.6	27.1

언어의 기능적 분류는 4-gram 연어가 시간, 장소, 기술, 양화 표현에 사용된 경우는 지칭어로 분류하고, 4-gram 연어가 비교, 초점, 프레임, 토픽 표현에 사용된 경우는 텍스트 조직어로 분류하고, 4-gram 연어가 인식, 의무, 가능 표현에 사용된 경우는 입장어로 분류하고, 여기에 속하지 않는 기능의 표현은 모두 기타로 분류하였다. 남한 고등학교 영어교과서에 나타난 지칭어는 구어 token의 경우 12.3%이고 type의 경우 13.7%이고 문어는 token의 경우 23.7%이고 type의 경우 24.3%였다. 북한 고등중학교 영어교과서에서는 명사구는 구어 token의 경우 14.2%이고 type의 경우 13.9%이고 문어는 token의 경우 27.6%이고 type의 경우 28.6%였다. 남한 고등학교 영어교과서에 나타난 텍스트 조직어는 구어 token의 경우 7.6%이고 type의 경우 7.9%이고 문어는 token의 경우 12.6%이고 type의 경우 13.2%였다. 북한 고등중학교 영어교과서에서는 명사구는 구어 token의 경우 7.7%이고 type의 경우 8.7%이고 문어는 token의 경우 13.7%이고 type의 경우 14.1%였다. 남한 고등학교 영어교과서에 나타난 입장어는 구어 token의 경우 31.8%이고 type의 경우 32.9%이고 문어는 token의 경우 28.7%이고 type의 경우 27.6%였다. 북한 고등중학교 영어교과서에서는 명사구는 구어 token의 경우 33.2%이고 type의 경우 32.9%이고 문어는 token의 경우 29.1%이고 type의 경우 30.2%였다.

남북한 영어교과서 모두 입장을 나타내는 표현에 나타나는 4-gram 연어의 비율이 상대적으로 지칭어와 텍스트 조직어에 비해서 큰 차이로 양이 많았다. 이러한 양의 차이는 4-gram 연어의 많은 양이 인식, 의무, 가능 표현을 포함하여 이루어졌다는 것을 보여주는 것이다. 남북한 영어교과서에 구어와 문어를 막론하고 모두 시간, 장소, 기술, 양화 표현에 사용된 지칭어가 비교, 초점, 프레임, 토픽 표현에 사용된 텍스트 조직어보다 많이 나타났다. 특히 주목할 것은 남북한 영어교과서 모두 입장을 나타내는 기능어 4-gram 연어는 문어보다 구어에서 더 많이 나타나는데 비해서 지칭어와 텍스트 조직어의 경우는 반대로 구어보다 문어 영역에 더 많이 나타난다. 이는 구어와 문어의 특성이 반영된 것으로 입장을 표현하는 *do you mind if, it is important to, It has to be, can you tell me* 등과 같은 기능어 4-gram 연어가 구어에 많이 나타났다. 문어에는 지칭어에 해당하는 *for a long time, at the beginning of, all over the world, from around*

*the world*과 같은 4-gram 연어들이 남북한 영어교과서 모두 네 번 이상 등장했다. 또한 문어에는 텍스트 조직어 기능에 해당하는 *at much younger ages, on the other hand, suppose if you are a, in the case of, the best title for, the only way to*와 같은 4-gram 연어들이 남북한 영어교과서 모두 세 번 이상씩 문어 영역에 나타났다.

Ⅳ. 결론

영어교과서 코퍼스에 대한 N-gram 분석을 통해서 연어 연구를 한 논문은 민덕기, 이승민, 심규남의 연구와 홍신철의 연구가 있다[5][16]. 두 연구 모두 초등학교 영어교과서 코퍼스를 대상으로 N-gram 분석을 통해서 발견된 연어 표현을 분류하고 원어민 코퍼스와 비교하였다. 이들 연구와 달리 본 연구는 북한의 고등중학교 영어교과서 코퍼스와 우리나라 고등학교 영어교과서 코퍼스에 대해서 N-gram을 돌려서 추출된 연어의 목록을 형태와 기능범주로 나누어서 분류하고 남북한 영어교과서에 나타난 연어의 양상을 비교하였다.

남북한 영어교과서 4-gram 연어에 나타난 차이로 눈에 띄는 것은 남한의 영어교과서에는 4-gram 연어가 구어영역에 더 많이 나타나는데 북한의 영어교과서에서는 4-gram 연어가 문어영역에 상대적으로 더 많이 나타난다. 이와 같은 현상은 북한의 영어교과서에서 문어영역에는 문법적 교수요목에 의해서 특정 언어형식에 따른 표현이 많이 삽입되어 있고 구조적인 지시어가 많이 들어 있기 때문으로 추정하였다. 또 다른 차이는 구어와 문어를 막론하고 북한 영어교과서에 나타난 명사구 연어의 4-gram 연어의 비중이 남한의 고등학교 영어교과서와 비교해보면 상대적으로 높다. 이러한 이유는 북한 영어교과서의 경우 황서연, 김정렬이 밝힌 대로 문법용어, 학술어휘와 정치나 사상적 용어의 반복에서 기인된다고 볼 수 있다[12]. 남북한 영어교과서 모두 입장을 나타내는 기능어 4-gram 연어는 문어보다 구어에서 더 많이 나타나는데 비해서 지칭어와 텍스트 조직어의 경우는 반대로 구어보다 문어 영역에 상당한 차이의 양이 더 많이 나타나는 것을 알 수 있었다.

코퍼스를 활용한 남북한 영어교과서에 나타난 N-gram 연어의 비교연구는 이제 시작 단계이고 본 연구를 통해서 밝힌 연어의 차이도 앞으로 다양한 연구를 통해서 양적으로 질적으로 분석해 볼 필요가 있다. 본 연구의 한계점인 남한에서 가장 널리 쓰이는 고등학교 영어교과서 1종으로 제한을 벗어나서 다양한 선택형 교과서 종별로 전체 교과서 코퍼스를 구성해서 분석하고 교과서 간에 N-gram 연어의 비교 분석과 더불어서 남북한 영어교과서 간에 구어와 문어를 구분해서 좀 더 세부적인 분석을 통해서 N-gram 연어의 사용 양상을 분석하고 비교해 본다면 다양한 교육적 시사점을 발견할 수 있을 것으로 생각된다.

참 고 문 헌

[1] B. Gray and D. Biber, "Lexical frames in academic prose and conversation," International Journal of Corpus Linguistics, Vol.18, pp.109-135, 2013.

[2] A. Wray, *Formulaic Language and the Lexicon,* Cambridge: Cambridge University Press, 2002.

[3] K. Hyland, "Bundles in Academic Discourse," Annual Review of Applied Linguistics, Vol.32, pp.150-69, 2012.

[4] M. Koprowski, "Investigating the Usefulness of Lexical Phrases in Contemporary Course Books," ELT Journal, Vol.59, No.4, pp.322-332, 2005.

[5] 홍신철, "초등학교 영어교과서 분석: 연속어휘다발을 중심으로," 언어과학연구, Vol.83, pp.485-506, 2017.

[6] J. J. Qin, "Use of Formulaic Bundles by Non-Native English Graduate Writers and Published Authors in Applied Linguistics," System, Vol.42, pp.220-231, 2014.

[7] K. Hyland, "Bundles in Academic Discourse," Annual Review of Applied Linguistics, Vol.32, pp.150-69, 2012.

[8] L. Chen, *An Investigation of Lexical Bundles in ESP Textbooks and Electrical Engineering*

Introductory Textbooks, In Perspectives on Formulaic Language, edited by David Wood, pp.107-128, London: Continuum, 2010.

[9] D. Wood, *Lexical Clusters in an EAP Textbook Corpus,* In Perspectives on Formulaic Language, edited by David Wood, pp.88-106, London: Continuum, 2010.

[10] J. Sinclair, *Corpus, Concordance, Collocation,* Oxford: Oxford University Press, 1991.

[11] D. Wood, *Formulaic Language and Second Language Speech Fluency,* London: Continuum, 2010.

[12] 황서연, 김정렬, "북한 초급중학교 영어교과서의 2013 개정 전후 어휘비교," 학습자중심교과교육연구, Vol.20, No.5, pp.611-634, 2020.

[13] D. Biber, S. Johansson, G. Leech, S. Susan Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English,* Harlow: Pearson, 1999.

[14] Y. H. Chen and P. Baker, "Lexical Bundles in L1 and L2 Academic Writing," Language Learning and Technology, Vol.14, No.2, pp.30-49, 2010.

[15] Biber, Douglas, Susan Conrad, and Viviana Cortes, "If You Look at...: Lexical Bundles in University Teaching and Textbooks," Applied Linguistics, Vol.25, No.3, pp.371-405, 2004.

[16] 민덕기, 이승민, 심규남, "초등학생용 영어 코퍼스 분석을 통한 뭉치말의 비교 연구," 영어교과교육, Vol.11, No.3, pp.39-64, 2012.

[17] 김정렬, "남북한 영어교과서 어휘의 차이," 한국콘텐츠학회논문지, Vol.17, No.4, pp.107-116, 2020.

저 자 소 개

김 정 렬(Jeong-ryeol Kim) 종신회원

- 1996년 ~ 현재 : 한국교원대학교 초등교육과 교수
- 2012년 ~ 현재 : 한국영어다독학회 회장
- 2013년 ~ 현재 : 초등영어교육학회 고문
- 2014년 ~ 현재 : 한국외국어교육학회 고문
- 2019년 ~ 현재 : 한국코퍼스언어학회 고문
- 2012년 ~ 현재 : Extensive Reading Foundation 이사
- 2014년 ~ 현재 : Reading in Foreign Languages 저널 편집이사

〈관심분야〉 : 컴퓨터활용 영어교육, 초등영어교육, 영어교수법