https://doi.org/10.5392/JKCA.2021.21.04.050

토픽 모델링을 활용한 한국 영어교육 학술지에 나타난 연구동향 분석

Analysis of Research Trends in Korean English Education Journals Using Topic Modeling

원용국, 김영우
서울대학교

Yongkook Won(linguistry@gmail.com), Youngwoo Kim(youngwoolearning@gmail.com)

요약

본 연구는 2000년 이후 최근 20년간 우리나라 영어교육의 연구동향을 파악해보는 것을 목적으로 한다. 이를 위해 영어교육 관련 주요 학술지 12개를 선정하여 해당 기간 동안에 게재된 논문 7,329편의 서지정보를 수집하여 분석하였다. 분석 대상이 된 영어교육 학술지의 논문 게재 현황은 2000년대부터 2010년대 전반기까지 계속 증가하였다가 2010년대 후반기에 다소 감소하였다. 그리고 2010년대 후반기에 학술지별 논문 게재 수도 비슷해졌다. 이와 같은 결과는 양적인 측면에서 영어교육 학술지의 영향력이 전반적으로 감소하면서 평준화된 것이라고 볼 수 있다. 다음으로 논문의 영문 초록을 데이터로 잠재 디리클레 할당(LDA) 토픽 모델링을 적용한 결과 34개 토픽(주제)이 추출되었다. 영어교육 분야에서 많이 연구된 토픽은 교사, 단어, 문화/미디어, 문법 등이었다. 단어, 어휘, 평가 등의 주제는 독특한 키워드를 통해 나타났고, 학습자요인 관련하여 여러 토픽들이 나타나면서 영어교육 연구의 관심 주제가 되었다. 다음으로, 상승 및 하강 토픽을 분석한 결과 상승 토픽으로 질적 연구, 어휘, 학습자요인, 평가요소 등이 있었고, 하강 토픽으로 CALL, 언어, 교수, 문법 등이 있었다. 이런 연구 주제의 변화는 영어교육 분야의 연구 관심사가 정적인 연구 주제에서 데이터 중심적이고 동적인 연구 주제로 이동하고 있음을 보여주는 것이다.

■ 중심어 : | 영어교육 학술지 | 연구동향 | 토픽 모델링 | 잠재 디리클레 할당 |

Abstract

To understand the research trends of English education in Korea for the last 20 years from 2000 to 2019, 12 major academic journals in Korea in the field of English education were selected, and bibliographic information of 7,329 articles published in these journals were collected and analyzed. The total number of articles increased from the 2000s to the first half of the 2010s, but decreased somewhat in the late 2010s and the number of publications by journal has become similar. These results show that the overall influence of English education journals has decreased and then leveled in terms of quantity. Next, 34 topics were extracted by applying latent Dirichlet allocation (LDA) topic modeling using the English abstract of the articles. Teacher, word, culture/media, and grammar appeared as topics that were highly studied. Topics such as word, vocabulary, and testing and evaluation appeared through unique keywords, and various topics related to learner factors emerged, becoming topics of interest in English education research. Then, topics were analyzed to determine which ones were rising or falling in frequency. As a result of this analysis, qualitative research, vocabulary, learner factor, and testing were found to be rising topics, while falling topics included CALL, language, teaching, and grammar. This change in research topics shows that research interests in the field of English education are shifting from static research topics to data-driven and dynamic research topics.

■ keyword : | English Education Journal | Research Trend | Topic Modeling | Latent Dirichlet Allocation |

접수일자 : 2020년 08월 13일
수정일자 : 2021년 01월 04일
심사완료일 : 2021년 01월 04일
교신저자 : 김영우, e-mail : youngwoolearning@gmail.com

I. 서 론

최근에 인공지능기술 사용이 본격화되어 기계번역, 챗봇 등과 같은 언어 관련 인공지능기술 서비스가 등장함으로써 영어교육 등 언어교육 분야에 컴퓨터 등 정보통신기술의 영향이 더욱 커지고 있다. 또한 인공지능기술의 핵심인 머신러닝 기법이 연구 방식에도 영향을 크게 주고 있다. 이에 본 연구에서는 토픽 모델링(topic modeling)이라는 머신러닝 연구 방법을 이용하여 영어교육 분야의 주요 학술지에 실린 논문들의 연구 동향을 파악해보고자 한다.

기존의 영어교육 분야의 연구동향 분석 방식은 대부분 연구자들이 수작업 방식으로 진행하였기 때문에 특정 학술지나 특정 주제를 한정하여 다루는 경향이 있었다[1]. 왜냐하면 연구자가 논문을 일일이 살펴보고 파악하는 데 시간이 많이 걸려서 주어진 시간에 살펴볼 수 있는 논문의 수가 제한될 수밖에 없었기 때문이다. 하지만, 토픽 모델링과 같은 비지도(unsupervised) 학습 방식의 머신러닝 연구 방식은 연구자가 들이는 시간을 크게 줄일 수 있기 때문에 보다 많은 논문을 분석하여 해당 분야 연구를 전체적으로 조망해볼 수 있다. 또한 토픽 모델링 활용 연구의 경우, 토픽(주제)의 명칭 부여 및 토픽 관련 키워드의 해석 부분에 해당 분야 전문가(domain expert)의 참여가 필요한 만큼 인공지능기술의 활용과 전문가의 협업이 이루어지는 좋은 사례라고 볼 수 있다.

본 연구에서는 2000년 이후 20년간 영어교육 관련 주요 학술지에 게재된 논문의 서지정보를 수집하고 결과를 분석하여 한국의 영어교육 연구동향을 제시하고자 한다.

II. 이론적 배경

1. 토픽 모델링

토픽 모델링은 대량의 텍스트를 분석하는 방법으로 문서(document)를 구성하는 단어(word)를 활용하여 잠재적인 토픽을 도출하는 확률 기반의 분석 모형이다. 기존에는 텍스트 마이닝(text mining)의 하나의 방법론으로 활용되어 왔는데, 빅데이터가 본격 활용되기 시작하고 머신러닝을 활용한 연구방법이 활성화되면서 연구동향 분석에 많이 활용되고 있다[2][3].

토픽 모델링 기법으로 많이 쓰이는 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)은 Blei, Ng과 Jordan[4]이 소개한 방법이다. LDA에서는 토픽 분포에 대해 특별한 가정을 하고 있는데, 문서는 토픽들로 구성된 집합이고, 단어는 토픽의 속성이라는 가정이다. LDA의 작동 원리를 그림으로 제시하면 다음 [그림 1]과 같다[p. 997].

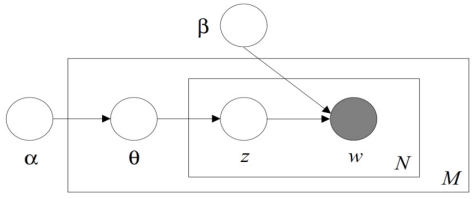

그림 1. Latent Dirichlet Allocation

LDA에는 코퍼스 수준, 문서 수준, 단어 수준처럼 세 가지 층위로 구성된다. 가장 외부인 코퍼스 수준의 파라미터인 α는 문서와 토픽 간의 밀도로 토픽 개수 설정과 관련되는 파라미터이고, β는 토픽과 단어 간의 밀도로 단어 개수 설정과 관련되는 파라미터이다. α 값이 크면(토픽 개수가 많으면) 문서를 설명할 수 있는 토픽을 많이 얻을 수 있다. 하지만 각각의 토픽을 설명하기는 어려워진다. 반대의 경우로, α 값이 작으면 토픽 수가 줄어들기 때문에 문서를 설명할 수 있는 토픽을 제대로 얻지 못하게 된다. β 값이 커지면 토픽이 더 많은 단어로 구성되고, 이 값이 작아지면 토픽 관련 단어가 줄어든다. 문서 수준(M) 변수인 θ는 개별 문서에 대한 토픽의 분포이다. 단어 수준 변수인 z는 특정 문서 내의 단어에 대한 토픽을 의미하고, w는 실제 단어를 의미한다. 변수 w는 실제로 확인할 수 있는 값(단어)이기 때문에 [그림 1]에서 원의 색상이 다르게 표시되어 있다. [그림 1]에서 N은 특정 문서에서 토픽과 단어를 반복해서 선택하는 것을 의미한다. LDA가 세 층위로 구성되어 있고, 문서 내에서 반복적으로 토픽을 선택하기 때문에 하나의 문서(본 연구의 경우 논문 초록)에 여러 키워드가 관련될 수 있게 된다.

2. 토픽 모델링 활용 연구동향 분석 연구

최근에 토픽 모델링을 활용한 연구동향 분석 연구들이 여러 분야에서 많이 이루어지고 있다. 이들 연구들은 분석 대상 논문들의 제목, 키워드, 초록 등을 데이터로 구축한 후 토픽 모델링 기법을 이용하여 해당 분야의 연구동향을 분석하고 있다.

표 1. 토픽 모델링 활용 학술지 논문 분석 연구

논문	분야	대상 논문	논문 편수	분석 영역
김성연(2020)	산업수학	해당분야 17종 (2016-2019)	6,255	제목, 초록
김세현(2018)	다문화 (검색어)	국내 발행 논문 (1993-2016)	2,334	초록
김창식외(2017)	정보시스템	해당 분야 3종 (2002-2016)	1,245	초록
박자현외(2013)	문헌정보	해당 분야 4종 (1970-2012)	3,834	초록
박종도(2019)	다문화 (검색어)	등재/등재후보 (2000년이후)	5,345	제목, 키워드, 초록
박준석외(2016)	hotel (검색어)	해당 분야 4종 (1994-2016)	706	초록
박준형외(2017)	기록관리학	해당 분야 6종 (1997-2016)	1,027	제목, 키워드, 초록
박한샘외(2019)	smart city (검색어)	Scopus DB (2010-2018)	10,955	제목, 키워드, 초록
방미현외(2020)	20대, 청년 (검색어)	국내 발행 논문 (2011-2019)	530	논문
오민정(2019)	고령 (검색어)	KISS DB (1996-2018)	1,291	초록
이효섭외(2020)	전환학습 (검색어)	등재/등재후보 (1998-2019)	82	초록
전설아외(2015)	정보학	해당분야 20종 (2009-2013)	6,545	제목, 초록
홍성연외(2017)	대학생지원 (검색어)	등재/등재후보, Scopus (2015년까지)	158 294	제목, 키워드, 초록

이 연구[5-17]들의 특징을 보면, 첫째, 문헌정보학 분야를 필두로 다양한 분야에서 토픽 모델링을 이용하여 연구가 이루어지고 있다는 점이다. 이를 보면 각 분야마다 토픽 모델링을 이용한 연구동향 연구의 유용성을 확인할 수 있다. 둘째, 분석 대상이 특정 학술지에 게재된 논문인 경우가 있고, 특정 주제어로 검색한 논문인 경우가 있다는 점이다. 전자의 경우 논문 게재 관련 학술지의 특성을 결과로 제시하고 있고, 후자의 경우 학제적 연구의 특성을 제시하는 경우도 있다. 셋째, 기존 연구동향 연구가 논문 제목과 키워드를 중심으로 한 것에 대비되어 이들 연구들은 논문 초록을 주된 데이터로 분석하고 있다는 점이다. 이는 연구자가 상대적으로 시간을 적게 들여 대량의 텍스트를 분석할 수 있는 비지도학습 방식의 토픽 모델링 기법 덕분이라고 볼 수 있다. 넷째, 연구자들은 해당 분야의 기존 연구들이 미시적이고 연구자의 주관성이 개입되는 한계가 있다고 지적하면서, 토픽 모델링을 활용한 연구가 거시적이고 객관적이라는 점을 강조하고 있다. 다섯째, 토픽 모델링 초반 연구들은 SAS Enterprise Miner, Matlab Text Analytics, Net Miner 등 여러 텍스트마이닝 툴을 사용하였고, 최근에는 R과 같은 범용 툴을 사용하는 경우가 있다는 점이다. 이는 토픽 모델링 연구가 빅데이터와 머신러닝 연구 방식이 본격 도입되기 이전부터 이루어졌다는 점을 보여준다. 여섯째, 토픽 모델링 외에 시계열 분석, 네트워크 분석과 같은 추가적인 기법을 적용한 경우가 있다는 점이다. 이는 연구의 취지로 해당 분야의 연구 주제의 변화, 그리고 토픽 간의 관련성을 함께 제시하고자 하는 경우에 많이 나타나고 있다.

3. 영어교육 연구동향 연구

영어교육 분야는 오랜 역사를 갖고 있는 만큼 오랜 기간 관련 연구가 이루어져 왔다. 영어교육 분야에서 가장 오래된 학술지인 〈영어교육(English Teaching)〉은 1965년에 창간되었고, 〈응용언어학(Korean Journal of Applied Linguistics)〉은 1983년에 창간되어 영어교육 관련 학술 논문이 게재되기 시작하였다. 이후 1990년대 말부터 2000년대 초에 영어교육 관련 학술지가 많이 창간되면서 영어교육 관련 학술 연구가 크게 늘어났다. 그러면서 영어교육 연구 동향을 살펴보는 연구도 많이 나타나고 있다.

'동향 연구'는 특정 분야 또는 특정 주제의 여러 연구 문헌들의 연구 결과를 분석하는 연구로서, 시대에 따라 혹은 특정 항목에 따라 연구 결과들을 분류하여 연구 경향을 살펴보거나 특이사항을 찾아내어 논의하면서 해당 분야의 연구 성과를 종합적으로 살펴보는 활동이다. 2000년 이후 2019년까지 20년 동안 영어교육 관련 12개 학술지에 게재된 논문을 보면 77편의 연구동향 관련 연구가 진행된 것으로 나타났다. 이 연구들의 일부는 특정 학술지의 연구 성과를 종합하여 분석하는 방식으로 학술지의 특정 시기(〈현대영어교육〉 창간 10주년, 〈영어어문교육〉 창간 20주년, 〈영어교육〉 창간

50주년 등)에 맞춰 진행된 경우이다[18-20]. 이 방식에서는 해당 학술지에 게재된 논문을 주요 영역으로 나눠 여러 연구자가 분석하였다. 반면, 동향 연구의 많은 부분은 연구자의 개별적인 관심에 따라 특정 주제의 연구동향을 조사하여 분석한 연구들이다. 이와 같은 연구동향 연구들은 연구자들이 수작업으로 대상 논문을 수집하여 분석하는 방식이라서 분석 대상 논문의 수량이 제한적이라는 특징을 보이고 있다.

이에 본 연구에서는 논문의 양적 분석 및 토픽 모델링 기법을 활용하여 거시적인 관점에서 2000년 이후 영어교육 분야의 주요 학술지에 게재된 논문들을 분석하여 다음과 같은 질문에 답을 얻고자 한다.

첫째, 2000년 이후 20년 동안 영어교육 학술지의 논문 게재 현황이 어떤가?

둘째, 동 기간 동안 영어교육의 주요 연구 주제(토픽)는 무엇인가?

셋째, 동 기간 동안 영어교육 연구 주제에서 어떤 변화가 있는가?

III. 연구 방법

1. 연구대상

본 연구의 대상은 영어교육 관련 학술지 12종에 2000년부터 2019년까지 20년간 실린 논문들이다. 학술지 목록은 다음과 같다.

표 2. 분석 대상 학술지 목록

학술지	기관
Foreign Languages Education	한국외국어교육학회
Studies in English Education	글로벌영어교육학회
멀티미디어언어교육 (Multimedia-Assisted Language Learning)	한국멀티미디어언어교육학회
영상영어교육 (STEM Journal)	영상영어교육학회
영어교과교육 (Journal of the Korea English Education Society)	한국영어교과교육학회
영어교육 (English Teaching)	한국영어교육학회
영어교육연구 (English Language Teaching)	팬코리아영어교육학회
영어문교육 (English Language & Literature Teaching)	한국영어문교육학회
응용언어학 (Korean Journal of Applied Linguistics)	한국응용언어학회
중등영어교육 (Secondary English Education)	한국중등영어교육학회
초등영어교육 (Primary English Education)	한국초등영어교육학회
현대영어교육 (Modern English Education)	현대영어교육학회

위의 학술지들은 모두 한국연구재단의 등재지로, 김영우와 김주혜[1]가 분석한 학술지 목록에 〈중등영어교육〉을 추가한 것이다.

2. 자료 수집

본 연구의 대상 학술지에 게재된 논문 관련 자료를 수집하기 위해 한국학술지인용색인(www.kci.go.kr), 기초학문자료센터(www.krm.or.kr), 그리고 각 학회의 웹사이트를 이용하였다. 먼저 각 학술지의 논문별로 서지정보(영어로 된 제목, 주제어, 초록 등)를 수집하여 엑셀 파일에 저장하였다. 수집된 논문의 서지 정보가 제대로 되었는지 파악하면서 파악되지 않은 부분은 추가 조사를 통해 수집하였다. 경우에 따라 이미지로 된 논문 파일에서 초록 텍스트를 추출하기 위해 OCR(Optical Character Recognition, 광학문자인식) 기법을 활용하였고, 추출된 텍스트를 확인하였다.

본 연구의 분석 대상이 영어교육 관련 논문으로 정해졌기 때문에, 12개 학술지에 2000년부터 2019년까지 게재된 7,622편의 논문 중에서 영어교육 분야가 아닌 한국어교육, 프랑스어교육, 일본어교육 등의 논문 280편과 초록이 파악되지 않거나 초록이 한글로 된 13편의 논문을 제외한 7,329편의 논문을 분석 대상으로 하였다.

3. 자료 분석

영어교육 관련 학술지에 게재된 논문의 연구동향을 분석하기 위해 Asmussen와 Møller[21]의 제안을 참고하여 전처리(pre-processing), 토픽 모델링, 후처리(post-processing) 과정을 거쳤다.

토픽 모델링 과정에서 논문 관련 데이터가 있는 엑셀 파일을 csv 파일로 변환한 후 필요한 필드의 데이터를

추출하였다. 본 연구에서는 논문의 영문 초록 데이터를 분석 대상으로 하였다. 분석 대상으로 논문의 제목을 제외한 이유는, 제목과 초록에서 동일 단어가 반복 등장할 경우 특정 단어의 가중치의 왜곡 현상이 있을 것으로 판단하였기 때문이었다. 키워드의 경우는 제목과 같은 이유 외에, 2000년대 초반 논문의 경우 키워드가 없는 경우도 있었기 때문이었다. 가중치 왜곡 현상의 방지를 위해 77편의 연구 동향 논문도 분석 대상에서 제외하여 토픽 모델링을 위한 최종 논문은 7,254편이 되었다.

텍스트 데이터 처리를 위해 NLTK, spaCy, gensim 등의 자연어처리 패키지를 사용하였다. 먼저 구두법 기호, 특수문자 등 일반 어휘가 아닌 요소를 제거하였다. 추출되는 단어 수를 줄이기 위해 텍스트를 모두 소문자로 변환하였다. 불용어 처리를 위해서 전술한 패키지의 불용어 목록을 통합하였다. 다만, 본 연구에서 활용되는 용어(CALL, computer 등)는 통합된 불용어 목록에서 제외하였다. 형태소 분석 과정에서는 단어의 기본형을 용어로 사용하기 위해 표제어 분석(lemmatization) 방식을 사용하였다. 추출한 단어의 품사는 기존 연구[9][11][14][17]처럼 명사로 한정하였다. 다만, 행위 동사(action verb)가 연구 주제를 파악하는데 도움이 될 수 있겠다는 판단에 따라 분석 초기 단계에 동사도 추출하여 분석에 참고하였다.

토픽 모델링 기법으로 잠재 디리클레 할당(LDA) 방법을 사용하였다. 이 과정에서는 77편의 기존 영어교육 연구동향 연구의 주제들을 참고하고, '복잡도(perplexity)' 값이 낮아지는 경향을 살펴보면서 토픽 수를 정한 후 토픽 모델링을 반복하여 실시하였다[17][22].

후처리 과정은 토픽 모델링을 적용한 결과를 확인하는 과정으로, 토픽이 제대로 분류되었는지, 그리고 각 토픽의 주요 단어(키워드)들이 적절한지 확인하였다. 각 토픽의 명칭을 적절하게 부여하기 위해 영어교육 전문가인 본 논문의 저자들이 일차적으로 판단을 하였고, 추가적으로 영어교육 전문가의 자문을 받아 명칭을 확정하였다.

다음으로 영어교육 분야의 연구동향의 변화를 파악하기 위해 전체 20년 기간 동안 상승한 토픽(hot topic)과 하강한 토픽(cold topic)을 파악하였다. 분석 과정에서 토픽 모델링의 결과와 함께 관련 자료를 추출하여 분석 결과의 타당성을 확인하고 결과 제시 및 논의자료로 활용하였다.

IV. 연구 결과 및 논의

1. 영어교육 관련 학술지의 논문 게재 현황

영어교육 관련 12개 학술지에 2000년부터 2019년까지 20년간 게재된 논문 현황을 5년 단위로 나눠 제시하면 다음 [표 3]과 같다.

표 3. 영어교육 관련 12개 학술지 논문 게재 현황

학술지	2000-2004	2005-2009	2010-2014	2015-2019	합계
Foreign Langs Edu	191	253	210	157	811
Studies in Eng Edu	78	93	108	165	444
멀티미디어언어교육	104	136	138	135	513
영상영어교육	79	102	131	194	506
영어교과교육	41	69	162	152	424
영어교육	293	272	215	156	936
영어교육연구	168	244	235	170	817
영어어문교육	110	263	285	181	839
응용언어학	106	146	201	142	595
중등영어교육		27	71	130	228
초등영어교육	89	117	190	162	558
현대영어교육	91	148	244	175	658
합계	1,350	1,870	2,190	1,919	7,329

2000년부터 2019년까지 12개 학술지에 총 7,329편의 논문이 게재되었다. 전체적으로 2000년 후반에 게재 논문 수가 크게 증가하여 2010년대 전반부까지 유지되다가 2010년 후반부에 다소 줄어들었다. 다만, 영어교육 분야 논문이 실제로 줄어들었다고 단정하기는 어렵다. 왜냐하면 한국학술지인용색인 사이트에서 '영어교육'으로 검색한 결과 2000년부터 2019년까지 10편 이상의 논문이 실린 등재후보 이상의 학술지가 76종이 되는 것으로 나타났다. 이중에서 〈학습자중심교과교육연구〉, 〈한국콘텐츠학회논문지〉, 〈영어영문학연구〉, 〈영어영문학〉, 〈교과교육학연구〉, 〈예술인문사회융합멀티미디어논문지〉 등의 학술지에 50편 이상의 영어교육 관련 논문이 게재된 것으로 나타났다. 그리고

학제적인 분야의 학술지에 영어교육 연구 동향을 조망하는 논문이 게재되기도 하였다[23]. 그러므로 본 연구의 분석 대상이 된 영어교육 관련 12종의 학술지의 영향력이 이전에 비해 줄어들었다고 짐작해볼 수 있다.

2000년대 전반기의 경우 〈영어교육〉의 게재 논문 수가 다른 학술지에 비해 월등히 많았고, 〈Foreign Languages Education〉과 〈영어교육연구〉가 중간 그룹을 형성하였고, 다른 학술지의 경우 상대적으로 적은 수의 논문이 게재되었다. 영어교육 분야에서 한국연구재단의 등재지들이 많이 생긴 2000년대 후반기에는 〈영어교육〉을 제외한 모든 학술지의 게재 논문 수가 증가하였다. 그래서 〈영어교육〉, 〈Foreign Languages Education〉, 〈영어교육연구〉, 〈영어어문교육〉이 논문이 많이 게재된 학술지 그룹이 되었다.

2010년대 전반기에는 〈영어교육〉, 〈Foreign Languages Education〉, 〈영어교육연구〉의 게재 논문 수가 줄어든 반면, 다른 학술지의 게재 논문 수는 계속 증가하였다. 2010년대 후반기에는 양상이 바뀌어, 〈Studies of English Education〉, 〈영상영어교육〉, 〈중등영어교육〉에 게재된 논문 수는 증가하였으나 그 외 다른 학술지에 게재되는 논문 수는 줄어들었다. 그로 인해, 12개 학술지의 최근 연간 평균 게재 논문 수가 비슷한 양상을 보이고 있다.

이와 같은 영어교육 관련 주요 학술지의 게재 논문 수의 감소 및 평균화 현상은 양적인 측면에서 주요 영어교육 학술지의 학문적 영향력이 약화되었고, 특정 영어교육 학술지의 영향력도 두드러지지 않게 됨을 보여준다고 할 수 있다.

2. 영어교육 분야의 주요 연구 주제

2000년부터 2019년까지 진행된 77편의 영어교육 연구동향 연구들의 주제를 보면 다음과 같다.

> 컴퓨터를 이용한 영어교육(CALL), 코퍼스(corpus), 창의력(creativity), 문화(culture), 교육과정(curriculum), 담화(discourse), 특수목적영어(ESP), 게임(game), 문법(grammar), 학습자(learner), 학습(learning), 듣기(listening), 리터러시(literacy), 교재(materials), 읽기(reading), 연구방법론(research), 말하기(speaking), 스토리텔링(storytelling), 교사(teacher), 교수(teaching), 평가(testing), 어휘(vocabulary), 쓰기(writing)

앞의 23개의 연구 주제가 영어교육의 연구 주제를 모두 포괄했다고 보기는 어렵지만, 영어교육의 주요 분야를 다루고 있다고 볼 수 있다. 이와 같은 기존 연구동향 연구의 주제 숫자를 고려하면서 복잡도 값이 낮아지는 토픽 수를 확인하였다. 확인 결과 [그림 2]처럼 토픽 수가 30개 전후로 복잡도의 값이 낮아지는 것으로 나타났다.

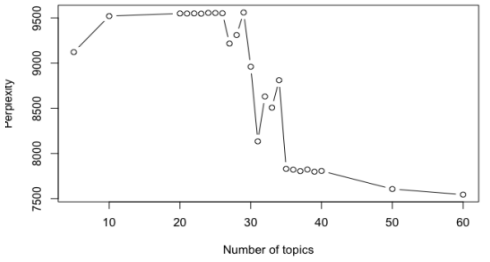

그림 2. 토픽 수에 따른 복잡도 값의 추이

복잡도 값이 낮아진 토픽 수 중 세 경우(30개, 34개, 36개)의 토픽별 키워드를 추출한 결과 토픽 수가 34개일 때 토픽별 키워드가 안정적인 것으로 나타나 토픽 수를 34개로 정한 후 토픽 모델링을 실시하였다.

토픽 모델링을 실시한 결과로 얻은 토픽별 키워드를 참고하여 토픽 명을 정하였다. 이 과정에서 토픽의 주요 키워드의 비중을 참고하였다. 일부 토픽의 경우 다양한 키워드로 인해 토픽 명을 정하지 못한 경우도 있었다[11]. 일련의 과정을 거쳐 확정한 토픽 명은 다음 [표 4]와 같다.

표 4. 주요 토픽 및 키워드와 논문 수

토픽		하강/상승	키워드 (비율)	논문 수
T23	읽기		reading (0.274), comprehension (0.157), question (0.127), skill (0.122), ability (0.063), reader (0.034)	223
T19	쓰기		writing (0.362), process (0.083), peer (0.069), essay (0.056), writer (0.045), quality (0.034)	288
T20	듣기		effect (0.155), ability (0.118), listening (0.099), control (0.077), improvement (0.062), interest (0.061)	243
T29	말하기	하강5	speaker (0.135), feature (0.076), pattern (0.06), speech (0.058), pronunciation (0.057),	198

			production (0.05)	
T17	의사 소통		interaction (0.128), communication (0.101), competence (0.097), classroom (0.057), discourse (0.048), conversation (0.047)	220
T12	단어	상승7	word (0.256), frequency (0.063), collocation (0.044), number (0.029), function (0.024), list (0.024)	333
T31	어휘	상승2	vocabulary (0.295), knowledge (0.19), term (0.035), size (0.032), effect (0.031), dictionary (0.023)	203
T25	문법	하강4	sentence (0.084), structure (0.07), verb (0.069), acquisition (0.038), construction (0.031), clause (0.024)	318
T10	문법 지도		instruction (0.232), form (0.137), grammar (0.116), input (0.064), target (0.061), output (0.042)	219
T21	평가 검정		test (0.406), item (0.114), score (0.11), taker (0.015), correlation (0.015), validity (0.015)	256
T26	수행 평가		assessment (0.11), performance (0.098), evaluation (0.072), criterion (0.048), scale (0.042), rater (0.034)	200
T01	평가요소	상승4	task (0.288), proficiency (0.111), effect (0.057), fluency (0.044), complexity (0.04), condition (0.038)	184
T04	CALL	하강1	child (0.111), computer (0.058), technology (0.05), web (0.047), tool (0.046), story (0.045)	296
T32	문화/ 미디어		culture (0.08), movie (0.078), video (0.038), world (0.026), life (0.024), film (0.02)	327
T03	교사		teacher (0.626), teaching (0.103), classroom (0.069), belief (0.031), practice (0.023), team (0.012)	424
T27	교수	하강3	teaching (0.187), method (0.15), material (0.137), model (0.135), lesson (0.076), purpose (0.047)	142
T33	학습자		learner (0.633), l2 (0.171), role (0.028), adult (0.02), autonomy (0.019), l1 (0.009)	92
T13	학습		learning (0.456), environment (0.046), application (0.046), style (0.044), project (0.03), preference (0.027)	177
T28	학습자 요인 (전략)		strategy (0.311), use (0.27), proficiency (0.097), difference (0.047), gender (0.024), questionnaire (0.023)	198
T16	학습자 요인 (동기)	상승6	factor (0.157), motivation (0.13), relationship (0.063), variable (0.055), proficiency (0.046), correlation (0.035)	231
T06	학습자 요인 (정서)		student (0.697), attitude (0.103), anxiety (0.066), questionnaire (0.044), survey (0.023), conferencing (0.002)	145
T08	학습자 요인 (기타)	상승3	student (0.288), self (0.184), feedback (0.161), achievement (0.101), efficacy (0.037), effect (0.017)	169
T24	코스 (대학)	상승5	course (0.131), university (0.123), student (0.094), college (0.089), perception (0.069), survey (0.063)	203
T22	프로그램		program (0.214), training (0.082), need (0.067), skill (0.057), development (0.049), service (0.033)	240
T34	학교 (교육 과정)		school (0.268), education (0.194), curriculum (0.106), year (0.035), policy (0.03), parent (0.028)	224
T02	교재 (활동)		activity (0.253), textbook (0.173), school (0.143), grade (0.101), purpose (0.053), grader (0.033)	229
T15	교재 (난이도)		text (0.202), problem (0.074), difficulty (0.067), stage (0.048), book (0.046), idea (0.038)	135
T05	이론 일반		approach (0.108), practice (0.076), theory (0.058), awareness (0.055), development (0.054), literature (0.051)	205
T14	언어	하강2	language (0.684), classroom (0.056), role (0.018), setting (0.016), context (0.015), target (0.014)	161
T11	오류 분석		type (0.221), error (0.123), article (0.084), category (0.035), correction (0.027), number (0.025)	174
T30	연구 (질적)	상승1	participant (0.182), interview (0.107), experience (0.091), journal (0.045), identity (0.041), case (0.035)	144
T18	연구 (기타)	상승8	class (0.365), student (0.355), participation (0.029), time (0.028), work (0.022), satisfaction (0.019)	182

영어교육 분야에서 많이 언급되는 네 기능(듣기, 말하기, 읽기, 쓰기)에 대한 연구를 보면, 읽기와 쓰기는 키워드를 통해 토픽이 명확하게 나타나는 반면, 듣기와 말하기는 상대적으로 명확성이 떨어지는 양상을 보이고 있다. 이는 해당 주제에 대한 연구 내용과 방법이 읽기와 쓰기처럼 명확하지 않는 면이 있을 수도 있고, 아니면 다른 영역과 연계된 연구가 많기 때문일 수도 있을 것이다. 그리고 '상호작용' 또는 '의사소통'이라는 토픽은 영어교육에서 강조되는 개념이지만, 통합적이고 추상적인 성격을 갖고 있다 보니 토픽으로 부각되기는 어려운 면이 있는 것으로 보인다.

'단어(word)'와 '어휘(vocabulary)'는 토픽으로 명확하게 나타나고 있으며 많은 연구가 이루어지고 있다. 이는 개별적인 단어들과 이들의 빈도와 연어 관계, 그리고 어휘 지식 등이 영어교육 분야에서 주목을 많이 받고 있기 때문이라고 해석해볼 수 있다. 그리고 IT의 발달로 코퍼스언어학과 같은 연구 분야의 연구 여건이 개선된 측면도 관련될 것이다. 그에 비해, 많은 연구가 수행된 '문법'과 '문법 지도' 토픽은 상대적으로 여러 다양한 키워드들이 해당 토픽에 영향을 주고 있는 양상이다. '평가검정'과 '수행평가'와 같은 토픽은 차별적인 키워드로 토픽이 구성된 만큼 영어교육에서 평가 영역이 고유한 분야라는 것을 잘 보여주고 있다. 시험과 같은 검정 평가와 수행 능력을 측정하는 수행평가가 별도의 토픽으로 구분되어 추출되었다는 점도 주목할 만하다.

'컴퓨터를 이용한 영어교육(CALL)'과 '문화/미디어'라는 주제는 영어교육에서 연구가 많이 이루어지고 있다. 다만, 토픽 분석에서 CALL과 관련되어 아동(child)이 키워드로 부각되는 것으로 볼 때, CALL의 적용이 학습자의 연령대에 따라 달라질 수 있다는 점을 고려해야 할 것이다. '문화/미디어'에서는 영화, 비디오와 같은 영상 자료의 활용이 많은 것으로 나타났다. 이는 영상을 많이 활용하는 최근의 사회 상황과 관련될 수 있다고 볼 수 있다.

'교사' 토픽은 관련 논문 수로 볼 때 영어교육 분야에서 가장 많은 연구가 이루어지고 있다. 이는 우리나라 영어교육이 공식적인(official) 또는 형식을 갖춘(formal) 방식으로 교사(teacher)에 의해 진행되고 있다는 점을 잘 보여준다. 교수(teaching)와 관련해서는

방법(method), 자료(material), 모형(model) 등이 관련 키워드로 나타나고 있다.

교수와 대비되어 '학습자(learner), 학습(learning)'도 중요한 토픽으로 나타나고 있고, 이에 덧붙여 여러 학습자요인이 또 다른 토픽으로 나란히 제시되고 있다. 이는 영어교육에서 학습자의 다양한 동적인 변인들이 중요한 고려사항이라는 점을 잘 보여주는 것이다. 학습자요인으로 전략(strategy)과 동기(motivation)가 상대적으로 명확히 나타나고 있다.

영어교육이 이루어지는 구체적인 방식으로 '코스(course), 프로그램(program), 교육과정(curriculum)'이 토픽으로 나타나고 있으며, '교재'에 대한 토픽도 나타나고 있다. 그리고 '학교(school)'라는 토픽도 나타나면서 우리나라 영어교육 연구가 공교육 중심으로 이루어지고 있음을 짐작해보게 하고 있다.

이론적인 측면에서 이루어지는 영어교육 연구와 언어적인 측면에서 이루어지는 영어교육 연구를 '이론일반, 언어, 오류분석'과 같은 토픽을 통해 확인해볼 수 있다. 그리고 '질적연구'와 같은 토픽이 영어교육의 연구방식으로 질적 연구가 주목받고 있다는 점을 알 수 있게 해준다.

잠재 디리클레 할당(LDA) 방식이 모든 문서가 아니라 특정 문서에 나타나는 키워드에 가중치를 두어 분석하는 방식이기 때문에 특정 단어가 여러 토픽에 나타나는 경우가 상대적으로 적은 편이다. 그럼에도 불구하고, 발달(development), 의미(meaning), 맥락(context), 능숙도(proficiency), 기능(skill), 습득(acquisition), 능력(ability) 등과 같은 단어는 3개 이상의 토픽에 등장하는 키워드들이다. 그러므로 영어교육의 세부 분야 연구 시 위와 같은 키워드를 고려해서 연구를 진행한다면 통합적이고 유기적인 방식의 연구가 될 것이다.

3. 영어교육 분야의 연구 주제의 변화

다음으로 2000년부터 20년간 시간의 흐름에 따른 연구 주제의 변화를 살펴보기 위해 시계열회귀분석에 따른 하강 토픽과 상승 토픽을 보면 다음 [그림 3]와 같다.

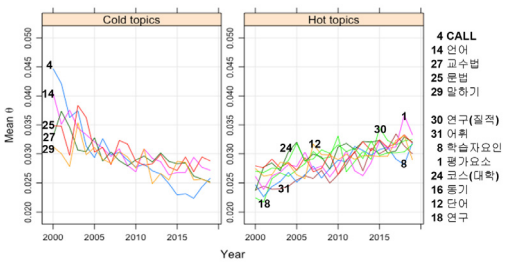

그림 3. 하강 토픽과 상승 토픽

하강 토픽으로는 CALL(T4), 언어(T14), 교수법(T27), 문법(T25) 등 5개가 나타났고, 상승 토픽으로는 질적연구(T30), 어휘(T31), 학습자요인(T1), 평가요소(T1) 등 8개가 나타났다. 하강 및 상승 토픽의 논의에서 먼저 언급할 것은, [표 4]의 논문 빈도를 통해 알 수 있듯이, 하강 토픽과 상승 토픽 대부분이 다른 토픽에 비해 많이 연구가 진행되고 있다는 점이다. 예를 들어, CALL의 경우 IT 분야가 본격적으로 발전하기 시작한 2000년대 초반에 다른 주제에 비해 월등히 많이 연구되다가 최근에 연구 수가 줄어들었지만 그래도 다른 주제만큼은 연구가 진행되고 있다고 볼 수 있다. 이런 결과는 CALL 관련 연구가 영어교육 분야에서 일반화되어 자체적인 주제의 연구에서 영어교육의 다른 주제의 연구에 통합되어 진행되고 있다고 짐작해볼 수도 있다.

토픽의 상승과 하강에서 주목할 점은 '언어', '문법'과 같은 언어 중심의 연구가 2000년대에 비해 상대적으로 줄어들고 있다는 점이다. 반면, '어휘' 등과 같은 데이터 중심의 연구와 '질적연구'와 '학습자요인'과 같은 상황(situation)과 맥락(context)에 의존적인 연구가 늘어나고 있다는 점이다. 이는 영어교육 연구의 관심사가 정적(static)인 주제보다는 동적(dynamic)인 주제로 이동하고 있음을 보여준다고 할 수 있다.

V. 결론

본 연구에서는 영어교육 연구동향을 거시적이고 양적인 측면에서 분석하였다. 분석 결과 전반적으로 영어교육 학술지에 게재된 논문이 양적으로 늘어난 것을 확인할 수 있었다. 또한 학술지별 게재 논문 수가 비슷하

게 수렴되는 것을 확인할 수 있었다. 영어교육 논문의 연구 주제에 대한 거시적 분석을 위해 토픽 모델링을 적용한 결과, 많이 연구되는 주제를 파악할 수 있었고, 상승 토픽과 하강 토픽 분석을 통해 기간에 따른 연구 논문의 수가 변화되고 있는 주제를 파악할 수 있었다. 파악 결과, 정적인 주제에서 데이터 기반의 동적인 주제로 영어교육 연구가 변하는 경향을 볼 수 있었다.

본 연구는 토픽 모델링을 이용하여 영어교육 연구동향을 파악해보았다는 면에서 연구의 의의가 있다. 그리고 향후 관련 연구의 계기가 될 수 있다는 점 또한 연구의 의의로 언급할 수 있을 것이다. 향후 연구 과제로는 본 연구에서 사용한 자료를 활용하여 네트워크 분석을 해볼 수 있을 것이다. 또한 본 연구의 대상이 된 영어교육 전문 학술지 이외의 학술지에 게재된 영어교육 논문을 파악하여 우리나라 영어교육 연구의 전반적인 연구동향을 파악해볼 수 있을 것이다.

참 고 문 헌

[1] 김영우, 김주혜, "영어교육 프로그램과 교육과정에 대한 연구 동향 분석," 영어교육연구, 제27권, 제2호, pp.57-83, 2015.

[2] X. Li and L. Lei, "A Bibliometric Analysis of Topic Modelling Studies (2000-2017)," Journal of Information Science, (September 2019). doi:10.1177/0165551519877049.

[3] 송민, *텍스트 마이닝,* 청람, 2017.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[5] 김성연, "텍스트마이닝 기법을 활용한 미국산업응용수학 학회지의 연구 현황 및 동향 분석," 한국콘텐츠학회논문지, 제20권, 제7호, pp.212-222, 2020.

[6] 김세현, "비정형자료분석을 통해 살펴본 한국의 다문화 연구," 한국인구학, 제41권, 제1호, pp.1-27, 2018.

[7] 김창식, 최수정, 곽기영, "토픽모델링과 시계열회귀분석을 활용한 정보시스템분야 연구동향 분석," 한국디지털콘텐츠학회논문지, 제18권, 제6호, pp.1143-1150, 2017.

[8] 박자현, 송민, "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석," 정보관리학회지, 제30권, 제1호, pp.7-32, 2013.

[9] 박종도, "토픽 모델링을 활용한 다문화 연구의 이슈 추적 연구," 한국문헌정보학회지, 제53권, 제3호, pp.273-289, 2019.

[10] 박준석, 김창식, 곽기영, "텍스트마이닝과 소셜네트워크분석 기법을 활용한 호텔분야 연구동향 분석," 관광레저연구, 제28권, 제9호, pp.209-226, 2016.

[11] 박준형, 오효정, "국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교: LDA와 HDP를 중심으로," 한국도서관·정보학회지, 제48권, 제4호, pp.235-258, 2017.

[12] 박한샘, 김동현, 장성주, "구조적 토픽 모델링 기반 스마트 시티 연구 동향 분석," 한국디지털콘텐츠학회논문지, 제20권, 제9호, pp.1839-1846, 2019.

[13] 방미현, 이영민, "20대 청년세대에 관한 연구 동향 분석," 한국콘텐츠학회논문지, 제20권, 제7호, pp.223-232, 2020.

[14] 오민정, "23 년간 (1996~2018) 국내학술연구의 '고령' 키워드로 살펴본 빅데이터 분석," 경영학연구, 제48권, 제2호, pp.515-532, 2019.

[15] 이효섭, 조대연, "토픽모델링과 키워드 네크워크분석을 활용한 국내 전환학습 연구 동향," 평생교육학연구, 제26권, pp.1-24, 2020.

[16] 진설아, 송민, "토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구," 정보관리학회지, 제33권, 제1호, pp.7-32, 2016.

[17] 홍성연, 최재원, "토픽 모델링 분석 기법을 활용한 대학의 학생 지원 연구 동향 분석," 학습자중심교과교육연구, 제17권, pp.21-48, 2017.

[18] 송민종, 임정완, "현대영어교육 학술지 10년: 읽기 및 쓰기 관련 연구 회고," 현대영어교육, 제11권, 제2호, pp.60-81, 2010.

[19] 남은희, "학술지 영어어문교육 창간 20주년을 통해 본 영어교육에서의 교수매체 연구 동향," 영어어문교육, 제20권, 제1호, pp.379-402, 2014.

[20] 신상근, "50년간 영어교육에 게재된 영어 평가 관련 논문 분석," 영어교육, 제70권, 제5호, pp.109-132, 2015.

[21] C. B. Asmussen and C. Møller, "Smart literature review: A practical topic modelling approach to exploratory literature," Journal of Big Data,

6, 2019. doi:https://doi.org/10.1186/s40537-019-0255-7
[22] 김지은, 백순근, “텍스트 빅데이터 분석 기법을 활용한 대학구조개혁 평가의 쟁점 분석,” 아시아교육연구, 제17권, 제3호, pp.409-436, 2016.
[23] 최원경, “초등영어교육 연구 논문의 변천: 코퍼스 기반 분석,” 한국콘텐츠학회논문지, 제19권, 제2호, pp.11-21, 2019.

저 자 소 개

원 용 국(Yongkook Won) 정회원

■ 2019년 10월 ~ 현재 : 서울대학교 교육종합연구원 연구원

〈관심분야〉 : 언어 평가, 응용언어학, 인공지능기술, 머신러닝

김 영 우(Youngwoo Kim) 정회원

■ 2018년 9월 ~ 현재 : 서울대학교 교육종합연구원 연구원

〈관심분야〉 : 인공지능기술과 언어, 정보통신기술과 언어교육, 멀티리터러시와 영어교육