대학생 영어 학습자 작문 코퍼스에 대한 주제별 분류: 계층적 군집화*

최재웅** · 송지영
(고려대학교)

Choe, Jae-Woong & Song, Ji-Young. 2013. **The Topical Classification of Essays by College Student English Learners Using Hierarchical Clustering**. *Language Information*. Volume 17. 93-115. In this study, we report on a set of experimentations for, and a successful completion of, the automatic topic classification of 3286 English essays (YELC) written by college level English learners in Korea. We adopted Hierarchical Agglomeration Clustering for our purpose. In order to find the best combination of distance measures and algorithms for hierarchical clustering, we first selected 100 essays, and then calculated precision rate on the basis of the subset of essays for each of the 15 combinations of 5 distance measures and 3 methods provided in R implementation of 'Dist' and 'hclust'. As a result, the combination of 'correlation' and 'ward' method was chosen as the optimal one for our chosen corpus, which was applied to ten sets of randomly selected 100 essays for further validation. As a final step for topic classification, the 'correlation'-'ward' combination was applied to classify the whole corpus into six topics. The precision rate was estimated to be 98.7%, a quite decent one for our purpose. We then conducted a Key word analysis on the six topic-groups, thereby showing some distributional characteristics of the words used in each group.

Key words: Learner Corpus, English, College students, YELC, Argumentative writing, Topical Classification, Hierarchical Clustering, Automatic unsupervised document classification, Statistical computing environment R, Key word analysis

---

* 본 논문은 송지영(2013: "통계적 기법을 이용한 코퍼스 파일 주제별 자동 분류", Kling Vol. 7)을 모태로 하고 있으며 연구 주제와 결론이 같다. 그러나 본 논문은 논증 방식이 보다 간결하면서도 보편성을 제고시키는 방식으로 새롭게 작성되었으며, 언어학적 논의를 크게 보충하였다. 본 논문은 초기 원고 형태로 "The 14th Korea-Japan Workshop on Linguistics and Language Processing(2013년 3월 8일, 경희대)"과 "제2회 연세 영어 코퍼스 심포지엄(2013년 4월 6일, 연세대)에서 발표된 바 있고, 특히 김종복, 이석재, 정채관 선생님의 도움과 지적을 통해 더 다듬어졌다. 본 논문의 작성 과정에 통계 방법론 논의 및 논문 완성에 도움을 주신 유석훈, 홍정하, 전희원 선생님께도 감사를 표한다. 심사위원 세 분의 구체적인 지적과 의견에도 감사드린다. 물론 본 논문의 오류나 미진한 부분은 전적으로 필자들의 책임이다.

** 교신저자, 고려대학교 언어학과, jchoe@korea.ac.kr

1. 서론

언어 연구의 새로운 방법론으로 코퍼스 활용에 대한 관심이 커지고 있고, 또 이런 기대에 부응하여 다양한 종류의 코퍼스가 구축되고 있다. 코퍼스 구축은 하나의 과정일 뿐 그 자체가 궁극적 목적이 될 수 없다. 언어에 대한 이해를 궁극적인 목표로 잡는다고 할 때, 코퍼스 구축 못지않게 중요한 것은 그것을 어떻게 활용하느냐 하는 문제로, 여기에는 아직도 개발의 여지가 많다. 이와 관련하여 구축된 코퍼스로부터 다양한 종류의 고급 정보를 어떻게 추출하느냐 하는 방법론 개발이 핵심 관심사로 떠오른다. 그 중 하나가 문서 자동 분류(automatic document classification)로 대상 코퍼스를 구성하는 문서들을 주제에 따라 분류하는 문제는 현재 주요 연구 과제로 부각되어 있다(Manning et al., 2007; Ingersoll et al., 2013). 전산적 기술을 활용한 문서 분류 연구가 문서별 주요 특질을 근간으로 하고 있다는 점에서 코퍼스별로 이미 외적으로 주어진 주요 지표(index), 또는 변수(variable)뿐만 아니라 코퍼스 구성요소로부터 도출한 여러 다양한 지표/변수의 확보는 코퍼스를 대상으로 한 연구의 핵심 과제다.

본 연구는 최근 공개된 영어 학습자 말뭉치 YELC 2011(이석재·정채관, 2012)의 일부를 대상으로 한 주제별 분류 실험에 대한 보고서다. YELC가 작지 않은 규모인 데다 보기 드물게 완전 공개된 학습자 말뭉치라는 점, 문서별 외적 정보, 즉 작성자의 성별과 영어 숙달도(proficiency)에 대한 정보도 추가되어 있다는 점 등에서 매우 가치 있는 자원이라는 점은 분명하다. 그러나 다른 한편으로 다각적인 활용을 위해서는 코퍼스를 구성하는 개별 문서별 특징을 최대한 확보할 필요성도 대두된다. 예를 들어 비교적 손쉽게 추출해 낼 수 있는 문서별 구성 어휘의 종류(type－표면형 기준) 및 어휘별 빈도수(token frequency) 등에 더해서 문서별 주제 및 그러한 주제를 구성하는 주요 특징적 어휘 등은 무엇인가를 밝혀내는 점도 주요 관심사가 된다. 본 연구에서는 바로 YELC를 주제별로 자동 분류하는 것을 일차적인 목표로 하되, 그에 못지않게 그 과정의 객관성과 효율성을 최대화시키는 방안에 대한 탐구를 또 다른 목표로 한다. 본 논문에서 시도한 방법이 필요한 수준으로 효율적이면서 객관적이라면, 본 연구의 방법론이 다른 코퍼스에도 적용되어 활용될 수 있을 것이며, 그만큼 더 보편성을 띠게 될 것이다. 더 나아가 본 연구는 YELC처럼 문서별 주제가 밝혀지지 않은 자료뿐만 아니라 이미 장르나 주제가 표기된 자원에 대해서도 활용 가능하다고 본다. 예를 들어, 코퍼스에 표시된 기분류의

타당성을 검증하거나, 추가 하위분류에도 활용이 가능할 것이다.

본 연구에서 택하는 분석 기법인 '계층적 군집화(hierarchical clustering)'는[1] 이미 전산적으로 효용성과 타당성이 입증되었다. 그리고 사용하기가 상대적으로 그리 복잡하지 않은 통계기법이고 그것이 본 연구에서 방법론으로 선택한 주요 이유이기도 하다. 그러나 계층적 군집화 내에도 수많은 하위 기법들이 제안되어 있고, 또 분석 대상과 분석 기법 사이의 적합도 등 구체적인 상황별 적용에는 추가 논의가 필요하다. 예를 들어 YELC를 분석하는 데 어떤 구체적인 기법을 적용해야 할지, 그러한 기법의 정확도는 어느 정도인지 등은 미지수다. 특히 언어학적 관점의 연구에서는 실제 이러한 기법의 활용이 거의 보이지 않고 있다. 따라서 언어학 연구의 지평을 넓힌다는 점에서도 본 논문에서와 같은 시도는 가치가 있다고 본다. 본 연구에서는 구체적인 통계 계산을 위해 무료 공개된 통계 프로그래밍 언어 소프트웨어인 R(http://www.r-project.org)을 활용한다.

본 연구의 구성은 아래와 같다. 2장에서는 우선 대상 코퍼스를 소개한다. 3장은 주제별 분류 기법인 계층적 군집화를 소개하고, 특히 구체적으로 어떤 하위 기법을 적용해야 할지에 대한 문제, 즉 선별 방법론을 논하기로 한다. 4장에서는 선택한 방법론의 정확도를 검증하는 절차를 진행한다. 5장은 주제별로 분류된 문서에 들어있는 어휘 사용양상을 핵심어 분석을 통해 일부 살펴본다. 6장은 결론이다.

2. 대상 코퍼스: YELC 2011

YELC 2011은 한국 내 한 대학에서 1학년 신입생이 작성한 영어 작문을 모아놓은 것이다(이석재·정채관, 2012). YELC는 크게 '논술' 작문(argumentative writing)과 '서술' 작문(narrative writing)으로 구성되어 있고 각 부문별로 총 3286개의 개인별 작문이 포함되어 있다. 그 중에서 본 연구는 주제의 구분이 명확한 '논술' 작문을 대상으로 한다.

우선 대상 코퍼스에 대한 큰 틀에서의 이해 차원에서 가장 기본적인 값들을 살펴보면 아래와 같다. 첫 번째는 타입(type)과 토큰(token)의 전체 값이고, 두 번째는 '중심경향성'(central tendency)의 주요 지표를 열거한 것이다.

---

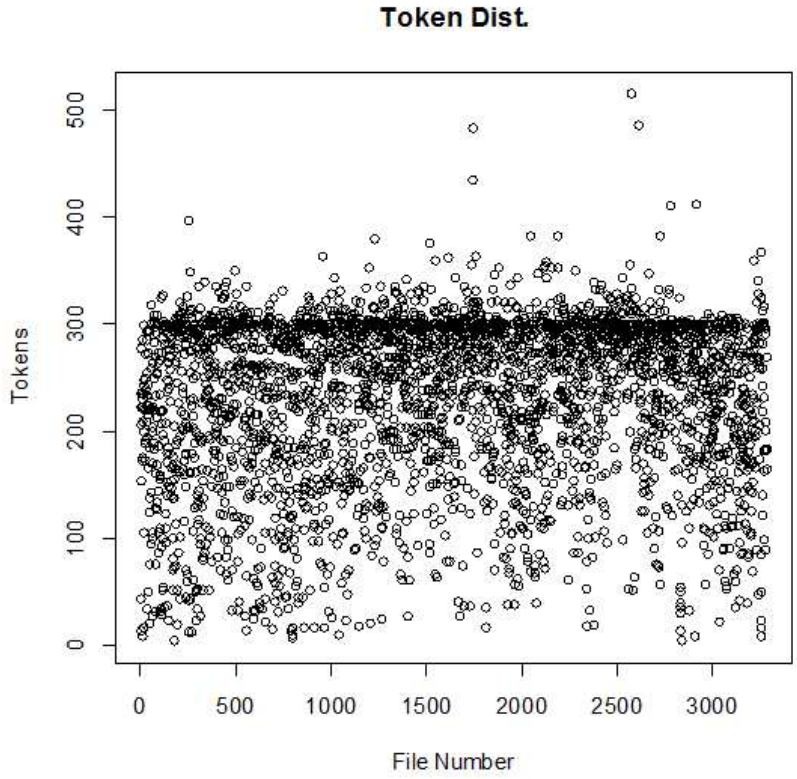
1) 본고에서 사용하는 통계용어는 주로 한국통계학회에서 온라인으로 제공하는 용어집을 따른다 (http://www.kss.or.kr/pds/sec/dic.aspx).

(1) ’논술‘ 작문의 규모 및 중심경향성 지표 (파일 3286개)

a. Tokens: 767,686; Types: 20,769

b. Min. : 4.0; 1st Qu. : 187.0; Median : 258.0; Mean : 233.6; 3rd Qu. : 296.0; Max. : 534

학생들이 작성한 개별 작문의 길이를 바탕으로 전체 평균을 구하면 233.6 어휘이고 중앙값은 258이다. 그런데 이러한 주요 지표에 더해 아래 그림은 여러 분포적 특징을 시각적으로 일목요연하게 보여준다.



<그림 1> YELC 2011 논술 작문 토큰 분포도

코퍼스에는 작문의 지시문에 대한 정보가 전혀 없고 이를 추가적으로 확인할 수도 없는 상황이지만 위 그림을 보면 지시문에 “300자”라는 말이 들어 있었을 것이라는 점을 쉽게 추정할 수 있다. 위 그림에서 작은 다이아몬드는 개별 작문의 분포로, 가로축(x축)은 3286개 개별 작문을 나타내고, 세로축(y축)은 작문별 토큰수를 나타낸다. 위 그림에서 많은 작문이 토큰수 300 근처에 짙게 몰려 있다는 점은 분명 학생들이 작문 과정에서 ‘300’이란 수치를 강하게

의식하고 있었다는 점을 드러낸다. 즉 중심경향성이나 산포도(dispersion) 수치만으로는 드러나지 않는 자료 분포의 특질을 순수 귀납적인 방법으로 도출시켜 준다.[2)]

이제 본 논문의 주 관심사인 주제의 분포라는 점에서 YELC의 특징을 살펴보자. 수작업을 통해 작문번호 1번부터 100번까지를 확인해 본 결과 개별 작문은 적어도 다음 여섯 가지의 주제중 하나를 택하여 그에 대한 찬-반의 의견을 작성한 것임을 확인할 수 있었다. 아래 표에서 병기된 숫자는 100개 중 해당 주제로 작성된 작문의 숫자이고, 이어 병기된 영어 표현은 각 해당 주제에 대한 구분자로서 이어지는 논의에서 해당 주제를 가리키는 약어로 사용하기로 한다.

<표 1> YELC 2011 논술 작문 100개 주제 분포

주제	문서 수	약식표기
학교 체벌	34	Discipline
동물대상 실험	16	Animal
공공장소에서의 흡연	16	Smoking
운전 중 휴대폰 사용	14	Cellphone
군대 의무 복무	13	Military
인터넷 실명제	7	Internet

주제 분류와 관련하여 구체적인 예 세 가지를 추가적으로 살펴보면서 자료의 특성을 조금 더 검토해 보기로 하자. 그리고 그러한 특성이 본 연구에서 택한 문서 군집화 기법에 미칠 영향에 대해서도 논의해 보기로 한다.

(2) I think smoking in all public buildings should be banned because in some reasons. First, Social must to prevent non-smoker's right. In public space, there are smoking people. But in there, so many non-smoking people are exsisting. Non-smoking people hate even the fragment of cigaret. If other people smoke in there, non-smoking people affected by cigaret's smog. That violates non-smoker's right, so smoking in public buildings must be

2) 관련 본문에서 설명하듯 (1)은 대상 코퍼스에 대한 기본적인 이해의 차원에서 제시하는 것이고, <그림 1>은 토큰 분포와 관련한 흥미로운 특징을 제시하는 것으로, 이 밖에도 더 많은 논의가 가능할 것이나 본 논문에서는 이 정도로 충분하다고 본다.

banned. ... <C2_0007>

(2)는 내용으로 보나 사용된 어휘적 특징(*smoking, smoke, cigaret, non-smoker* 등)으로 보나 Smoking과 관련된 주제임이 명확하다. 반면 (3)은 비유를 통한 주제의 전개로, 마지막 문장 전까지는 주제가 무엇인지 명확하지 않고, 사용된 어휘의 특징 역시 좀 막연한 편이다.

(3) Yes of coures. say, there are two rabbits, one in bush, the other in another bush. If you chose one of them and try to catch it Maybe you could catch at least one of them. but If you chose all of them and try to catch them you couldn't catch even one of them. because when you are in one bush to catch one of them the other will left. when people do many things at sametime, They are confused and they can't do even one of them. So They should not use cellular phones while driving. <C2_0590>

위 글은 마지막 문장에 가서야 주제에 대한 단서가 드러난다. 따라서 이런 종류의 글을 어휘의 분포만으로 주제가 무엇인지 파악하는 데는 뚜렷한 제약이 있을 법하다. 다음 예는 조금 다른 관점에서 흥미로운 예가 된다.

(4) No I disagree a this opinion because that is a important thing not only study but also physical actions. When We stydied many hours in the small room rise up Co2. so, We speak headache, and dry eyes, dry nose. And also down blood exercising speed. So, We need a physical action time. Actually many schools when P. E time they did not tougch the P. E. because school's opinion is We need go to a universty not pysical actions but study English, Koreans, Mathmatics and exam. so many schools when P. E time they tiching English, Mathmatics and many other study. But my opinion is we go to a university, we need a pysical actions. because everytime we study. So, always sit down and looking at the books and short sleep time. Actually many students speak headache, stomachache, dry eyes, and many sick. So, they usually eatting a drug. so my opinion by subject is shuld physicl punishment be allowed in all schools. <C2_0223>[필자 주: 오타에 밑줄 추가]

(4)는 글의 주제가 무엇인지 어느 정도 독자를 혼란스럽게 한다는 점에서는 (3)과 비슷하다. 그러나 (3)이 일종의 글쓰기 전략으로서의 비유적 용법으로 인한 일시적인, 의도된 혼란인 반면, (4)는 글쓴이가 '체육'과 '체벌'을 혼동한 데서 기인한, 즉 글쓴이의 착각으로 인한 혼란으로 보인다. 이 글 역시 마지막 문단 전까지는 '체육', 또는 '운동'의 필요성을 강조하고 있다. 만일 평가자의 입장에서 글 (4)가 Discipline이라는 주제에 부합하느냐는 질문을 받는다면, 관점에 따라 크게 달라질 수 있을 것이다. 그러나 만일 <표 1>에 제시된 분류 중 어디에 속하느냐고 묻는다면 일단은 '체벌'로 분류되는 것이 맞다고 본다. 이처럼 (3)이나 (4)와 같은 류의 작문이 어느 정도 규모로 존재하는지 전체적으로 확인이 불가능하고, 또 그러한 방식의 글쓰기가 문서 군집화 기법에 어떤 영향을 줄지도 확실하지 않은 상황이다.

(4)가 보여주는 또 다른 주요한 특징은 적지 않은 수의 오타다. 대상 말뭉치가 학습자 말뭉치라는 점에서 어느 정도의 오타가 들어 있을 것으로 쉽게 예상된다. 그런데 특히 그런 오타가 주제와 관련된 어휘라는 점에도 주목해 보자. 이런 오타는 사람의 눈에는 그리 큰 문제가 아닐 수 있다. 그러나 별도의 전처리나 자원을 활용하지 않는다면 문서 군집화 관점에서는 모든 오타가 전혀 다른 어휘로 인식될 수 있다. 따라서 어휘의 분포에 의존하는 문서 군집화 방식이, 어느 정도의 오타가 기대되는 학습자 말뭉치의 경우에도 별도 조치 없이 적용이 가능할지가 의문으로 남는다.

3. 분석 방법 및 절차: 비감독 기반 계층적 군집화

본 연구에서 취한 분석 방법, 절차, 관련 도구는 아래와 같다.

(5) 분석 절차: 내용(도구)
   a. 코퍼스 전처리: 형태소 주석(Stanford POS Tagger), 명사추출(Perl script)
   b. 자료변환: 문서-어휘 행렬로 변환(R)
   c. 군집화: 주제별 계층적 군집화(R)

이어지는 논의에서는 위에 제시된 단계별로 세부적인 작업 내용을 설명하기로 한다.

3.1. 코퍼스 전처리

문서 군집화 기법은 대상 코퍼스에 들어 있는 어휘의 분포와 빈도를 바탕으로 하는 것이라서 목적에 따라 대상 어휘를 정리하는 작업이 필요하다. 예를 들어 문장부호를 제거한다든지, 또는 *the*처럼 빈도는 매우 높으나 연구의 목적에 거의 도움이 되지 않고 오히려 문제를 야기할 소지가 높은 어휘들은 별도로 목록을 만들어 사전에 제거하기도 한다.[3] 본 연구에서는 작문의 주제 파악에 명사가 주요 변수가 될 것이라는 가설을 취하여 코퍼스에서 명사만을 추출하여 분석 대상으로 삼기로 하였다. 이를 위해 우선 Stanford Log-linear Part-Of-Speech Tagger(Toutanova et al., 2003)를 사용하여[4] 형태소 주석을 첨가하였고 그 중에서 명사류만을 추출한 임시 중간 코퍼스를 구축하였다.[5]

3.2. 자료변환

전산처리의 관점에서는 일반 텍스트는 구조화되지 않은 자료다. 따라서 전산처리를 위해서는 일단 텍스트를 구조화된 자료로 변환시키는 절차가 필요하다. 구조화된 자료의 대표적인 형태가 문서-어휘 행렬(Document-Term Matrix)로, 아래 표가 아주 간단한 예시가 된다.

---

3) 소위 ‘stopwords’라고 불리는 목록으로 연구자에 따라 필요한 목록을 만들어 사용한다.

4) http://nlp.stanford.edu/software/tagger.shtml 참조. 구체적으로는 BFSU_Stanford POS Tagger 1.1.2(bidirectional, left3words)를 이용하였다.

5) 주석된 자료에서 N으로 시작하는 것으로 아래와 같은 네 가지 종류가 있다. 문서별 명사 코퍼스로의 변환은 Perl script를 이용하였다.

NN : noun, common, singular or mass
NNP : noun, proper, singular
NNPS : noun, proper, plural
NNS : noun, common, plural

<표 2> 문서-어휘 행렬

Words	*people*	*animals*	*punishment*	*name*	*Internet*	*experiments*	*experiment*
c2_0003	1	0	6	0	0	0	0
c2_0004	0	0	8	0	0	0	0
c2_0006	4	6	0	0	0	5	1
c2_0008	2	0	5	1	0	0	0
c2_0009	0	0	6	0	0	0	0
c2_0010	4	10	0	0	0	1	5
c2_0023	1	0	0	10	6	0	0
c2_0024	5	10	0	0	0	8	1

위 표에서 열(column)은 어휘의 목록이고, 줄(row)은 문서 목록으로 열과 줄이 교차하는 개별 셀(cell)에 나오는 숫자는 각각 해당 어휘가 해당 문서에 몇 회 나오는가를 표시해 주고 있다. YELC 2011 내 문서가 3286개고, 사용된 명사의 종류가 8692이므로, 그 전체는 8692*3286개로 된, 약 3천만 개에 육박하는 셀로 구성된 행렬식으로 변환되는 셈이다.[6] 그리고 개별 셀에는 해당 어휘가 해당 문서에 몇 회 사용되었는가가 수치로 표시된다. 그런데 그러한 행렬의 주요한 특징으로 상당 비중의 셀 값이 0이 될 것이고 이러한 현상은 통계에서 '희박자료(sparse data)'의 문제로 불린다. 이 점은 추후 더 논의하기로 한다. 이처럼 행렬식으로 변환된 자료는 앞서 언급한바 '구조화된 자료'로 여러 통계적 일반화를 도출해 낼 수 있는 바탕이 되고, 군집화도 그중 대표적인 일반화 기법 중 하나다.

3.3. 군집화

문서 분류는 크게는 데이터 마이닝, 또는 텍스트 마이닝(text mining)의 한 가지 하위 분야로, 자동화된 문서 분류 기법은 보통 감독 기반 방식(supervised method)과 비감독 기반 방식(unsupervised method) 두 가지로 나뉜다(Manning

6) 이러한 변환을 자동으로 해 주는 도구가 여럿 있다. 본 연구에서는 R 패키지 'tm'의 'Corpus', 'DocumentTermMatrix' 함수 등을 주로 이용하였다 (http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf).

et al., 2007; Ingersoll et al., 2013).[7] 본 연구에서는 그 중에서 비교적 적용이 간단한 비감독 기반 방식을, 그리고 그중에서도 정확성이 상대적으로 우수하다고 알려져 있는 계층적 응집 군집화(hierarchical agglomeration clustering) 방식을 택하기로 한다.

군집화를 위해서는 크게 두 가지 절차를 거친다(Gries, 2010). 하나는 문서 또는 어휘간 거리를 계산하는 단계고, 두 번째는 그러한 거리를 기준으로 계층적 군집화를 만드는 단계다. 그런데 각 단계별 여러 하위 기법들이 개발되어 있으므로 그 중 어느 것을 택할지를 결정해야 한다. 이를 위해 본 연구에서는 R의 패키지 'amap'에 구현된 여러 거리 계산 기법들 중 적합한 것과, 그러한 결과에 적용되는 군집화 방법(알고리즘) 중 하나를 선택하되 그 중 최상의 결과를 도출하는 쌍을 찾는 작업을 우선 시도하였다.

거리 계산과 알고리즘 선택을 위한 비교 분석 대상으로 파일 번호 기준 상위 100개를 이용한다. 이는 4절에서 시도할 정확도 검증에서의 규모와 같은 규모다. 상위 100개 파일은 이미 직관에 따라 파일별 주제 분류를 해 둔 상태이므로 그 분류를 일단 '황금표준'(gold standard)으로 잡았다. R 패키지 'amap'에서 함수 'Dist'에 포함된 거리 기법 중에서는 "eucl(idean)", "manh(attan)", "pear(son)", "corr(elation)", "kend(all)"을, 그리고 R의 함수 'hclust'에서 제공되는 군집화 방법 중에서는 "ward", "sing(le)", "comp(lete)"를 선택하여 검토하였다. 따라서 그 두 가지 조합 15개를 모두 상기 100개의 파일 분류에 적용한 뒤에 그 결과를 위의 '황금표준'과 비교하여 보았다. 그 각 결합별 정확도는 아래와 같다.

<표 3> YELC 기준 거리 계산-알고리즘 결합별 정확도

번호	'Dist' 거리 계산	'hclust' 알고리즘	문서 수	불 일 치 개수	정확도
1	eucl	ward	100	19	81%
2	eucl	sing	100	62	38%
3	eucl	comp	100	30	70%

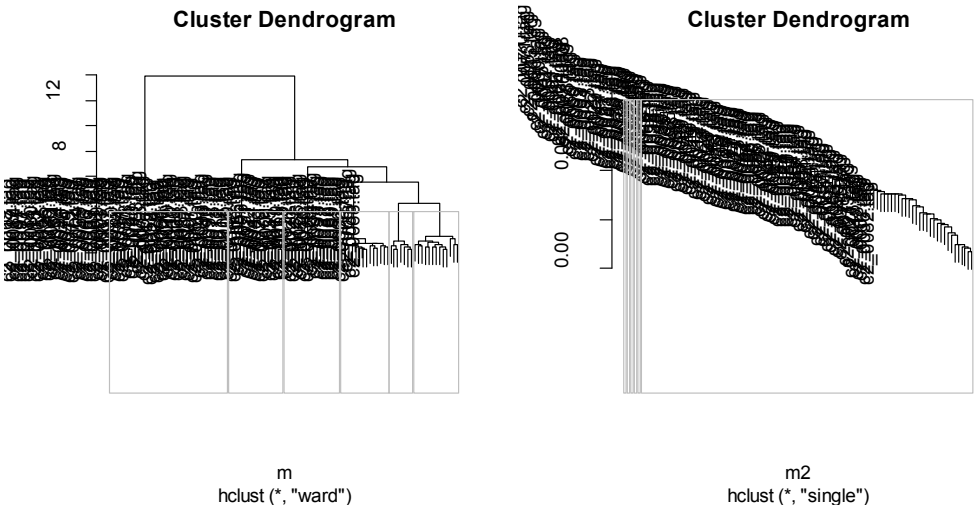
7) 비교적 간결하게 텍스트 마이닝에 관한 소개를 담고 있는 다음 웹사이트도 좋은 참고가 된다.
https://en.wikipedia.org/wiki/Document_classification,
http://cogsys.imm.dtu.dk/thor/projects/multimedia/textmining/node11.html

4	manh	ward	100	14	86%
5	manh	sing	100	64	36%
6	manh	comp	100	31	69%
7	pear	ward	100	0	100%
8	pear	sing	100	19	81%
9	pear	comp	100	10	90%
10	corr	ward	100	0	100%
11	corr	sing	100	19	81%
12	corr	comp	100	0	100%
13	kend	ward	100	7	93%
14	kend	sing	100	63	37%
15	kend	comp	100	51	49%

위 표를 보면 거리 계산과 알고리즘을 어떤 조합으로 선택하느냐에 따라 정확도가 크게 달라지는 것을 알 수 있다. 물론 위의 결과가 모든 대상에 대해 각 결합간 보편적인 우열관계를 보여준다고는 볼 수 없을 것이나, 적어도 YELC 2011이라는 특정 코퍼스를 대상으로 한 특정 목적, 즉 주제 분류라는 관점에서는 결합별로 차이를 보이고 있다.[8] 정확도의 차이는 <그림 2>에서와 같은 두 개의 수형도(dendrogram)를 통해서도 분명해진다.

<그림 2>는 R의 plot 기능을 이용해 그린 두 수형도로 세로로 그어진 회색선이 6개 군집을 서로 구분해 주고 있다. 두 그림을 비교해 보면, 왼쪽은 <표 1>에 제시된 주제별 파일 분포를 잘 반영하고 있는 반면, 오른쪽은 파일 대부분이 하나의 주제에 몰려 있는 것을 볼 수 있다. 왼쪽은 'correlation'-'ward' 결합으로 100% 정확률을 나타내는 경우고, 오른쪽은 'kendall'-'single' 결합으로 37%의 정확률을 보이는 군집화 결과다.

8) 위 표에서 "불일치 개수"는 최소한도로 계산한 것으로, 계산법에 따라 더 커질 수도 있다. 군집화 결과에 나오는 '주제'는 임의의 숫자로 되어 있으므로 이러한 결과와 '황금표준' 사이의 일치 여부는 별도로 맞추어 보아야 하고, 그런 과정에서 숫자와 주제 사이의 상관관계를 어떻게 설정하느냐에 따라 일치 여부 및 그에 따른 정확도는 달라질 수 있다. 예를 들어 <표 3> 14번의 경우 대부분 문서가 '1'번 그룹에 속하는 것으로 나왔다(이어지는 <그림 2> 참조). 이런 경우에는 기존의 '황금표준'에서 가장 많은 비중(34개)을 차지하는 'Discipline'과 일치하는 것으로 보았다. 반대로 '1'번 그룹이 'Internet'과 상응하는 것으로 했다면 전체 정확도는 10%에도 미치지 못한다.



<그림 2> 거리계산-알고리즘 결합 결과 비교

위 표를 살펴보면 거리 계산법에서는 ‘correlation’이, 군집화 알고리즘으로는 ‘ward’가 비교적 좋은 결과를 산출해 주는 경향이 있는 것으로 보인다. 따라서 본 연구에서는 그 둘의 결합(<표 3>의 10번)을 군집화 방법으로 선택하여 활용하기로 한다.

4. 검증 및 적용

3절에서의 논의를 통해 YELC 2011을 대상으로 한 주제별 분류에서 ‘correlation’-‘ward’ 기법의 결합으로 100% 정확한 결과가 산출된다는 점을 보인 바 있다. 그러나 이러한 결과가 우연에 의한 결과일 가능성도 배제할 수 없다. 다시 말해 최초 100개의 파일이 우연히도 좋은 결과를 산출하는 집합일 가능성을 배제할 수 없다는 말이다. 본 절에서는 이러한 문제에 대처하기 위해 임의의 파일 집합을 반복적으로 선택하여 각각 정확성을 검증해 보는 방식을 통해 위에서 제시된 결과의 타당성을 검증해 보고, 이어서 전체 자료에 동일 방법론을 적용해 보기로 한다.

4.1. 임의의 파일 집합 선택을 통한 반복 검증

선택의 우연성에 따른 결과값의 왜곡 가능성을 넘어서기 위해 흔히 취하는

방식은 임의 샘플링을 반복적으로 실시해 보는 것이다. 즉 대상 코퍼스에서 일정 규모로, 임의로 파일 집합을 선택한 다음 그것에다 위 방식으로 군집화를 시도해 보는 작업을 몇 차례 반복해 보는 것이다. 그러한 작업을 몇 회까지 반복해 보아야 하는가에 대한 정답은 없다. 물론 많을수록 정확도 수치가 그만큼 더 안정적이겠지만, 작업량이 이에 비례해서 커지게 마련이다. 본 연구에서는 그러한 작업을 10회 반복하였다. 그리고 매번 임의로 선택하는 집합별 파일의 개수를 3절에서의 규모와 같게 100개씩으로 하였다.[9)]

<표 4> 임의선택 반복을 통한 정확도 검증

	Animal	Military	Discipline	Cell phone	Internet	Smoking	불일치	정확도	불일치 예
RS01	17	11	22	15	19	16	0	100%	
RS02	16	14	27	12	14	17	0	100%	
RS03	19	11	30	10	16	14	0	100%	
RS04	13	14	28	13	13	18	1	99%	C2_0417 Ani->Mil
RS05	13	15	26	24	9	12	1	99%	C2_2739 Cel->Mil
RS06	13	11	26	16	17	17	0	100%	
RS07	14	14	34	6	9	23	0	100%	
RS08	9	14	30	13	15	18	1	99%	C2_3123 Smo->Dis
RS09	17	18	25	12	14	14	0	100%	
RS10	8	11	28	23	12	18	0	100%	

위 표에서 각 줄(row)은 임의의 파일 100개씩을 선택한 경우들(RS01~RS10)을 나타낸다. 상위 각 열(column) 제목은 매 실험에서 100개의 파일이 각각 어떤 주제로 분류가 됐는지를 나타내며, 마지막 세 열은 각각 불일치 개수, 실험별 정확도, 그리고 불일치된 사례와 해당 파일을 표시한다.

위 결과를 보면 평균적으로 99.7%에 이르는 정확률을 보인다. 그리고 10회에 걸친 검증에서 별 등락폭이 없이 고르게 나타난다. 그렇다면 계층적 군집화가 YELC 2011의 '논술' 작문의 주제별 분류에 적용될 타당성이 충분히 입증되었고 그 정확률은 100%에 육박한다고 결론 내릴 수 있다.

9) 임의의 파일 100개 선택에는 R의 함수 'sample'을 이용하였다.

끝으로 위에서 잘못 분류된 3가지 예의 내용을 살펴보았는데, 겉으로 뚜렷이 드러나는 특징을 찾기는 어려우나 오분류의 원인에 대한 추정은 가능하다. 다음 예를 보자.

(5) I think this question is not valid. Do you think killing a person is right? You think smoking cannot kill person directly, but I don't think killing is just crime. Being harm to some person is not crime? I wish you think about this question more. I can't respond about it. People who think right way, the answer is one. <C2_3123.txt>

위 내용은 *smoking*이란 어휘가 한 번 등장하므로 주어진 여섯 개의 주제 중 하나로 분류해야 한다면 Smoking이 될 것이다. 그러나 그 점을 제외하면 사실상 주제가 무엇인지 확실히 파악하기 어려울 정도로 글이 매우 막연하게 전개되고 있다. (5)는 *harm*, *crime*, *person* 등의 어휘가 있기 때문에 아마도 Discipline과 관련된 글로 잘못 분류됐을 것으로 짐작된다.

나머지 두 개 중, 주제가 Cellphone인 <C2_2739.txt>의 경우엔 *people* 9회, *thing(s)* 5회, *danger* 2회 등으로 조금 많이 쓰이고 있다는 점이 약간 특기할 만하나, *traffic*이나 *accidents* 등도 2회씩 쓰이고 있다. <C2_0417.txt>에서는 *animals*와 *experiments*가 각각 1회씩 쓰인 반면 *technology* 2회, *country/countries* 2회, *economics* 1회 등이 쓰였다는 점이 아마도 주제를 Military로 오분석하게 한 요인이 아닐까 추정해 볼 수 있다.

4.2. 전체 자료 분석

앞 절의 논의를 통해 본 연구에서는 'correlation'-'ward' 결합을 이용한 계층적 군집화가 적합하다는 점을 밝혔다. 이제 그러한 방식으로 전체 3286개의 분석을 시도해 본다. 그런데 100개라는 소규모 자원에 적용하던 방식을 그보다 훨씬 큰 규모의 코퍼스에 적용할 때 추가적으로 고려해야 할 점들이 생겨난다. 가장 큰 문제는 앞에서도 언급한 희박자료(sparse data) 문제로, 두 가지 측면에서 이 문제를 검토해야 한다.

희박자료와 관련하여 가장 큰 문제는 3천만여 개 셀의 상당수 값이 0이라는 점이다. 그것을 그대로 둔 채로 거리 계산과 군집화를 시도할 경우 전산처리에 시간도 어마어마하게 걸릴 뿐만 아니라 중간과정에서 여러 가지 문제가 야기되

어 원하는 결과를 얻기가 어렵다. 이러한 문제의 해결방향은 '차원감소' (dimension reduction) 기법을 활용하는 것으로, 기본적으로 문서 분류에 별로 기여하지 않을 뿐만 아니라 오히려 문제를 야기할 소지가 있는 어휘를 대폭 배제하는 방법이다. 본 연구에서는 이를 위해 R패키지 'tm'에 들어있는 'removeSparseData'라는 함수를 활용하였다. 그리하여 행렬의 열에 본래 제시된 8천여 개의 어휘를 영향력이 크다고 판단되는 100개 남짓으로 줄여서 작업을 진행하였다.

희박자료와 연관된 또 한 가지 문제는 개별 파일에 내용이 거의 없는 경우다.

(3) I don't think so. <C2_0175.txt>

작문 <C2_0175.txt>는 위에서 보듯 내용이 거의 없다. 더구나 위 자료에서 명사만 추출할 경우에는 내용이 빈 상태로 바뀐다. 이처럼 내용이 없는 파일이 존재할 경우 전산처리에 커다란 문제가 생겨날 수 있다. 대체로 내용 분량이 일정 기준에 못 미치는 경우는 일단 전산처리에서 완전히 배제하는 것도 처리 결과의 안정성을 높인다는 차원에서 일반적으로 취하는 한 가지 방법이기는 하나, 본 연구에서는 입력 파일을 최대한 살린다는 차원에서, 파일에 어휘가 전혀 들어있지 않은 경우만 배제하고 진행하였다. 그렇게 해서 배제된 파일이 두 개로, 전체 분류 대상 파일은 3284개가 된다.[10)]

아래 표는 YELC 2011 논술 작문 전체에 대한 분류 결과다.

10) 여기에 더해 행렬식 내의 값을 빈도에서 비율로 바꾸는 가중치 방식도 추가로 채택하였다. 이는 절대 빈도수를 상대적인 비율값으로 바꾸어 주는 것으로, 일반적으로 결과 값의 정확도를 높여주는 기법으로 알려져 있다.

<표 5> 군집화를 통한 YELC 2011 논술 작문 주제 분포

주제	파일 수	비율
Animal	440	13.4%
Cellphone	462	14.1%
Discipline	927	28.2%
Internet	473	14.4%
Military	482	14.7%
Smoking	500	15.2%
합계	3284	100.0%

위 결과의 정확도는 얼마나 될까? 물론 수작업을 통해 모든 자료를 검토해서 위 결과의 정확도를 확인해 볼 수도 있겠으나, 전체 수작업 분류는 본래부터 본 연구에서는 논외로 했었다. 정확도와 관련해서 한 가지 추정해 볼 수 있는 것은 위 결과의 정확도가 4.1절에서 반복 검증을 통해 도출한 정확도와 크게 다르지 않을 것이라는 점이다. 다만 한 가지 불투명한 부분으로, 4.1절에서의 검증은 100개를 단위로 한 검증이었던 반면, 위 분석은 3천 개 이상의 자료를 대상으로 한 것이라는 점에서 정확도에 영향을 미칠 수 있는 한 가지 주요 변수가 생겼다. 즉 대상 규모가 커지는 데 따른 오차 가능성이 얼마나 될까라는 의문이다.

이러한 의문에 대해 확인해 볼 수 있는 자원을 우리는 이미 확보하고 있다. 4.1절에서 반복 실시한 평가에서 이미 수작업을 통해 상당수 작문의 주제에 대한 '황금표준'을 확보해 둔 상태로 그것의 규모는 이미 1000개 수준, 그 중에서 중복 선택된 경우를 감안하면 정확히는 897개가 된다.[11] 그 897개의 주제가 전체를 대상으로 적용한 군집화 결과와 어느 정도 일치하는지를 확인해 본다면, <표 5> 결과의 정확성을 가늠해 볼 수 있을 것이다. 이렇게 검토해 본 결과 오분석된 경우들은 아래 표에 제시된 12개로, 897개의 정확도는 98.7%가 된다.

11) 4.1절에서 활용한 R의 임의선택 함수에 따라 동일 파일이 두 차례 이상 선택될 수 있다. 실제로 89개의 파일이 두 차례 선택되었고, 또 7개 파일이 세 차례 선택된 것으로 나온다. 거기에 1회만 선택된 801개의 파일을 더하면 모두 897개가 된다.

<표 6> 군집화 분석 결과 중 오분석 사례

파일	황금표준	군집 오분류	성별	숙달도	토큰 수
c2_0291.txt	Military	Discipline	F	A1+	54
c2_0795.txt	Military	Discipline	M	A1	10
c2_0797.txt	Military	Discipline	M	A1	10
c2_0911.txt	Animal	Discipline	F	A2	18
c2_0916.txt	Animal	Military	M	A2	234
c2_1378.txt	Animal	Discipline	F	B1+	343
c2_2145.txt	Smoking	Discipline	M	B1	126
c2_2445.txt	Military	Discipline	M	B1	103
c2_2845.txt	Cellphone	Discipline	F	B1	264
c2_3123.txt	Smoking	Discipline	M	A1+	66
c2_3260.txt	Cellphone	Discipline	M	A1	24
c2_1128.txt	Internet	Discipline	M	B1+	346

앞의 <표 5>에서 Discipline을 주제로 한 작문 수가 다른 주제에 비해 두 배가 넘는다는 것을 알 수 있었다. <표 6>에 나온 오분석 사례를 보면 Discipline 쪽으로 약간의 쏠림현상이 드러난다. 그런데 그보다 더 흥미로운 점은 '토큰 수'가 오분석과 상관성이 뚜렷하게 보인다는 점이다. 앞의 (1)에 제시된 토큰 분포 수치와 비교해 보면, 제1사분위 경곗값 187보다 적은 수치가 <표 6>에서 8개로 전체 12개의 2/3나 된다. 즉 토큰수가 상대적으로 적은 파일들에서 오분석의 가능성이 부쩍 증가한다는 점이다. 저빈도나 희박자료가 통계적 분석에 문제를 제기한다는 관점에서 이러한 높은 상관성은 어느 정도 예측된다. 따라서 정확도를 높이기 위해서는 저빈도 사례 처리에 대한 여러 대책이 필요할 것이다.

<표 6>의 오분석 사례는 <표 4>에 나온 오분석 사례와 비교해서도 특기할 만한 점이 있다. 동일 자료임에도 100개씩 쪼개서 군집분석 했을 때와 전체 3284개를 한꺼번에 군집분석을 할 때의 결과가 정확도나 오분석 사례 면에서 서로 다르다. 정확도는 99.7%에서 98.7%로 미미한 수준이지만 1% 하락했다는 점을 확인할 수 있다. 또한 오분석 사례에서도 <c2_3123.txt>의 경우만 일치할 뿐 나머지는 서로 다르다. 이러한 차이는 군집분석이 선택된 파일들을 대상으로 상대적인 어휘 분포에 기초한 것이라는 점 때문이다. 직관에 바탕을 둔

주제 구분에서는 작문별로 하나씩 읽으며 주제가 무엇인지 확인하는 과정을 거치는 반면 군집화 기법은 대상으로 삼은 파일 전체를 기준으로 그 안에서 어휘가 상대적으로 어떤 분포를 보이는가를 구분해 주는 것이다. 따라서 상대 비교 대상을 무엇으로 잡느냐가 분류결과에 영향을 미치게 되는 것이다.

결론적으로 100개를 대상으로 10회 반복 평가했을 때의 평균 정확도가 99.7%였고, 전체를 대상으로 잡았을 때도 임의의 자료 897개를 확인해 본 결과 정확도가 98.7%로 거의 대등한 수준을 유지한다.

5. 군집별 핵심어 분석

1절에서 언급하였듯이 작문에 대한 주제별 분류는 추가적인 통계 작업을 위한 주요 특질로 활용될 수 있다.[12] YELC 2011과 관련하여 한 가지 흥미로운 연구 과제로, 주제별로 학생들이 어떤 어휘를 주로 선택하는가라는 질문을 들 수 있다. 이러한 질문에는 '핵심어 분석'이 한 가지 답이 된다. 이는 각 주제별로 특징적으로 사용되는 어휘의 종류나 분포를 분석해 보는 것으로서 이미 잘 알려져 있는 기법이다(Key word Analysis: Scott & Tribble, 2009; Jeon & Choe, 2009; 전지은, 2010).

핵심어 분석을 위해선 일단 대상 코퍼스와 참조 코퍼스가 정의되어야 한다. 대상 코퍼스는 물론 <표 5>에 제시된 주제별 파일들이 될 것이다. 참조 코퍼스도 흔히 취하는 방식에 따라 전체 코퍼스 중에서 대상 코퍼스를 제외한 것으로 한다(Culpeper, 2009). 코퍼스의 대상에 대한 가공 여부, 또는 가공 정도도 결정해야 한다. 원시 코퍼스로 할지, 형태소 분석 코퍼스로 할지, 아니면 그 이상의 선별과정을 거쳐야 할지의 문제이다. 본 연구에서는 앞 절에서 주제별 분류에 활용한 명사 추출 임시 코퍼스를 활용하기로 한다. 이러한 판단은 앞에서 언급한 대로 명사가 주제를 가장 잘 드러내는 특징적 어휘일 것이라는 가정과 또 명사만으로도 주제 분류가 성공적으로 이루어졌다는 점에 바탕을 둔 것이다.

위의 방식에 따라 주제별로 대상 코퍼스와 참조 코퍼스를 구분한 뒤에 핵심어를 추출하였다. 추출 도구로는 AntConc가 활용되었다(Anthony, 2012). 아래 표는 추출 결과이다.[13]

---

12) 대표적으로 로지스틱 분석(Logistic Analysis: Bresnan & et al., 2007)이나 로그리니어 분석(Log-linear Analysis: Jun, 2010; Choe, 2012) 등의 주요 변인으로 활용될 수 있다.

<표 7> 주제별 핵심어 목록 (핵심도 200이상, 상위순서대로)

Discipline	*punishment, students, teachers, teacher, children, school, schools, student, education, punishments, class, parents, way, teaching, violence, child, behavior, ways*
Internet	*internet, name, people, information, names, cyber, users, privacy, nickname, opinion, words, rumors, opinions, crimes, comments, id, online, world, site, rumor, crime, netizens, sites, web, suicide, reply*
Military	*service, men, korea, army, war, country, north, military, nation, south, soldiers, women, years, duty, force, soldier, man, korean, job, weapons, power*
Smoking	*smoking, buildings, smokers, non, smoke, building, smoker, health, places, smell, area, fire, cancer, cigarette, air, place, room, people, lung, areas, cigarettes, nonsmokers, rooms*

위 표를 보면 주제별로 특징적인 어휘들의 목록이 일목요연하게 제시되어 있다. 예를 들어 '동물 실험'의 경우 *animals, experiments, animal, experiment* 등 주제어 자체가 주요 특징어로 드러나고 있고, 거기에 더해 그런 주제와 직결되는 '의학, 인간생명, 과학' 등과 연관된 어휘들이 목록의 상위를 차지하고 있다. 주제별로 사용된 어휘의 분포나 빈도를 보면 주제와 직결되는 소수의 어휘가 반복 사용되는 현상이 뚜렷하다. 예를 들어 Animal 주제의 글 440편(<표 5>)에 사용된 총 명사 25032개 중에서 *animals, experiments, animal, experiment*가 사용된 빈도 합계가 7615로 명사 세 개 중 하나꼴로 주제어가 반복 사용되고 있는 셈이다. 이러한 특징은 다른 주제의 작문에서도 마찬가지이다.[14]

<표 7>에서 또 한 가지 눈에 띄는 점은 *people*이 Internet과 Smoking 두 곳에 핵심어로 나온다는 사실이다. 사실상 *people*은 전체 파일에서 빈도수가 가장 많이 나오는 어휘 타입이다. 상위 빈도 10위까지의 어휘 유형 목록은 아래와 같다.

13) AntConc에서의 Keyword추출시 Log-likelihood를 이용하였다.

14) 물론 이것은 평균값이기 때문에 작문에 따라서 그런 주제어가 적게 사용된 경우도 있고 그런 경우가 분류상 어려움을 야기할 것이다. 그리고 심지어는 해당 주제어가 들어있다고 해서 군집분류에서 동일 주제로 당연히 분류되는 것도 아니다(앞의 예 (5) 참조). 즉 군집분류는 대상 어휘 전체의 분포를 동시에 고려하여 계산된 결과이다.

<표 8> 토큰 빈도상 상위 10위까지의 어휘타입 목록

	토큰빈도	어휘 유형	관련 주제
1	9797	*people*	Internet, Smoking
2	7279	*punishment*	Discipline
3	4786	*smoking*	Smoking
4	4504	*students*	Discipline
5	3531	*animals*	Animal
6	3515	*internet*	Internet
7	3089	*name*	Internet
8	2739	*service*	Military
9	2693	*children*	Discipline
10	2571	*teachers*	Discipline

<표 8>에서 Discipline 관련 핵심어가 네 개씩이나 나오는 것은 그 주제의 작문이 다른 작문에 비해 두 배 정도로 많다는 데 있을 것이다. 가장 많이 쓰인 *people*의 경우 35% 정도는 Internet 주제의 글에, 그리고 25% 정도는 Smoking을 주제로 한 글에 사용되고 있다. 이러한 상대적인 분포의 비중이 아마도 *people*이 두 주제 모두에 핵심어로 등장하는 까닭으로 보인다. 여섯 가지 주제가 모두 일상적인 생활과 관련된 주제라서 의미적인 측면에서는 *people*이 굳이 특정 주제와 연관이 있다고 판단되지 않으나 실제 어휘의 사용면에서 특정 주제와 연관될 수 있다는 점은 흥미롭다. 특히 <표 7>에서 '사람'을 가리키는 표현들이 적지 않게 각 주제별 핵심어로 선정되어 있는 바, 이는 주제에 따른 어휘 선택의 차이를 보여준다.

6. 결론

본 연구에서는 YELC 2011 '논술' 작문 3286개의 주제별 분류를 위해 비감독 기반 방식인 계층적 응집 군집화(hierarchical agglomeration clustering)를 이용한 실험 결과를 제시하였다. 그 결과 98.7% 수준의 정확도로, 전체 파일을 여섯 가지 주제로 분류할 수 있음이 확인되었다. 또한 그러한 분류를 바탕으로 각 주제에 따라 특징적으로 사용되는 핵심어 분석을 통해 해당 주제와의

연관성을 일부 살펴보았다.

언어학 내에서도 코퍼스 및 관련 통계에 대한 관심이 급증하고 있는 시점에서, 코퍼스에 대한 통계적 분석방법이 여러 가지로 시도되어야 함은 당연할 것이다. 본 연구는 그러한 연구의 일환으로 계층적 군집화를 통한 주제별 문서 분류가 성공적으로 이루어질 수 있다는 점을 보였다. 작문별 주제가 제공되지 않는 코퍼스의 경우엔 본 연구에서 제시한 방법론을 활용하여 분류도 하고 그 결과에 대한 정확도도 가늠해 볼 수 있을 것이다. 본 연구에서처럼 순수 귀납적인 방식으로 코퍼스 전체의 특성을 파악하는 방식은 앞으로도 다양하게 시도되어야 할 것이다. 거기에 더해 군집화처럼 안정적이면서도 비교적 적용이 쉬운 기법들이 목적에 따라서 앞으로도 적극 활용될 여지가 있다.

물론 본 연구 결과의 정확도는 대상 코퍼스 자체의 특성과도 밀접하게 연결되어 있을 것이다. 본래 작문이 주제별로 작성되었고, 또 학습자 말뭉치인 만큼 어휘 사용이 다채롭지 못하여 그만큼 성공적인 분류가 가능했다고 볼 수도 있기 때문이다. 그렇지 않은 코퍼스는 이처럼 높은 성공률을 보이기 어려울 것이라는 점을 예측하기가 어렵지 않고, 실제로 보다 고도화되고 일관성을 갖춘 문서분류 기법들이 꾸준히 개발되고 있다. 이러한 새로운 기법들이 언어학적 논의에 적극 이용된다면 보다 효율적이면서도 체계적인 코퍼스 관리 및 활용에 큰 도움이 될 것이다.

참고문헌

이석재·정채관. 2012. “연세영어학습자코퍼스(YELC): 한국인 영어 사용자 연구를 위한 새로운 언어자원.” 「제1회 연세 영어코퍼스 심포지엄 프로시딩」. 2012년 3월. 연세대학교. 26-36.

전지은. 2010. “핵심어 분석을 통한 성별 어휘 사용 양상 연구.” 고려대학교 대학원 박사학위 논문.

Anthony, L. 2012. AntConc (Version 3.3.5w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. “Predicting the Dative Alternation.” In *Cognitive Foundations of Interpretation*, ed. by G. Boume, I. Kraemer, and J. Zwarts. Amsterdam: Royal Netherlands Academy of Science. 69--94.

Choe, J. 2012. “The role of gender and other variables in the distribution of English tag questions: A log-linear analysis.” *Language, Communication, and Culture* Vol. 1. 9-25.

Culpeper, J. 2009. “Keyness: Words, part-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet.” *International Journal of Corpus Linguistics* 14(1). 29-59.

Gries, Stefan Th. 2010. *Statistics for Linguistics with R: A Practical Introduction*, Berlin: Mouton Textbook.

Ingersoll, Grant S., Thomas S. Morton and Andrew L. Farris. 2013. *Taming Text: How to Find, Organize, and Manipulate It*, Manning Publications.

Jeon, J., and Choe, J. 2009. “A Key word Analysis of English Intensifying Adverbs in Male and Female Speech in ICE-GB.” *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*. City University of Hong Kong. 210-219.

Jun, J. S. 2010. “Log-linear regression in Corpus Linguistics.” *ESSLLI 2010*. Course Material. University of Copenhagen, Denmark.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Scott, M., & Tribble, C. 2006. *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam; Philadelphia: John Benjamins.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.” *Proceedings of HLT-NAACL* 2003. 252-259.

▪ 최재웅 (Choe, Jae-Woong)
▪ 소속: 고려대학교 언어학과
▪ 전자우편: jchoe@korea.ac.kr

▪ 송지영 (Song, Ji-Young)
▪ 소속: 고려대학교 언어학과
▪ 전자우편: dramania@naver.com

▪ 논문투고일: 2013. 07. 12.
▪ 논문심사수정완료일: 2013. 09. 12.
▪ 논문게재확정일: 2013. 09. 20.