JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 2, pp. 237-246, Feb. 2023

Check for updates

문화 요소의 유사도와 네트워크 분석을 통한 문화적 특징의 정량적 분석

한 경 수*
*성결대학교 컴퓨터공학과 교수

Quantitative Analysis of Cultural Characteristics through Similarity and Network Analysis of Cultural Elements

Kyoung-Soo Han*
*Professor, Department of Computer Engineering, Sungkyul University, Anyang 14097, Korea

[요 약]

사회적 동물인 인간의 삶에 관한 이야기를 표현하고 기록하는 말과 글에는 자연스럽게 문화의 내용이 포함된다. 본 연구는 대용량 말뭉치에 나타난 문화적인 특징을 정량적으로 분석하였다. 문화를 표현하는 동사를 11개 범주에 따라 58개를 정의하고 이 동사와 한 문장에서 함께 등장한 명사를 문화 요소로 추출하였다. 평균 오버랩과 순위 편향 오버랩(rank-biased overlap)을 사용하여 문화 요소 리스트 간의 유사도를 측정하였고, 문화 요소 네트워크를 생성하여 문화 요소의 중심성을 검토하였다. 한국어 말뭉치에서 분석한 결과, 감정, 생사, 생활 범주는 타 범주와 유사한 문화 요소를 가지고 있었고, 식, 의, 경제 범주는 타 범주와는 구분되는 특유의 문화 요소를 가지고 있었다. 가족과 관련된 어휘들의 중심성이 높아 가족 중심의 문화적인 특징을 확인할 수 있었다.

[Abstract]

The content of culture is naturally included in words and writings that express and record stories about human life as a social animal. This study quantitatively analyzed the cultural characteristics of the large corpus. 58 verbs expressing culture were defined according to 11 categories, and the nouns that appeared together in one sentence with these verbs were extracted as cultural elements. The similarity between cultural element lists was measured using average overlap and rank-biased overlap, and the centrality of cultural elements was reviewed by creating a cultural element network. As a result of analyzing the Korean corpus, emotion, life&death, and living categories had similar cultural elements to other categories, and eating, clothing, and economy categories had unique cultural elements that were differentiated from other categories. The centrality of vocabulary related to family was high, and it was possible to confirm family-oriented cultural characteristics.

색인어 : 문화 동사, 문화 요소, 문화 요소 클라우드, 문화 요소 네트워크, 문화 마이닝

Keyword : Cultural verb, Cultural element, Cultural element cloud, Cultural element network, Culture mining

http://dx.doi.org/10.9728/dcs.2023.24.2.237

Received 22 December 2022; **Revised** 06 January 2023
Accepted 26 January 2023

***Corresponding Author; Kyoung-Soo Han**

Tel: +82-31-467-8189
E-mail: kshan@sungkyul.ac.kr

Ⅰ. 서 론

문화란 사회 구성원이 가지는 의식주, 언어, 풍습, 종교, 학문, 예술, 제도 등의 행동 및 생활 양식의 과정과 그 과정의 산물이다[1][2]. 사회적 동물인 인간의 삶과 그 주변의 이야기를 표현하고 기록하여 전달하는 말과 글에는 자연스럽게 문화의 내용이 포함될 것이다. 그렇다면 어떤 사회를 반영한 말과 글로부터 그 사회의 문화를 구성하는 요소들을 파악해 낼 수 있지 않을까? 그것을 어느 정도 자동화해 볼 수 있지 않을까? 이 질문들이 본 연구의 출발점이다.

다음과 같은 문장들을 살펴보자.

- (문장 1) 죽이 싫은 사람은 밥 먹어도 돼.
- (문장 2) 여럿이 먹을 때는 반찬이나 찌개 등의 그릇에 공용 젓가락과 조그만 국자를 곁들여 놓는다.
- (문장 3) 한 대접 막걸리를 단번에 마시고 나니 뱃속이 출렁였다.
- (문장 4) 배추김치를 담글 때 젓갈은 최근에 와서 멸치액젓을 구입하여 사용한다.
- (문장 5) 된장찌개는 채소와 된장을 넣고 금방 끓여낸다.

위 문장들로부터 다음과 같은 식문화의 일부를 파악할 수 있겠다.

- (문장 1), (문장 2): '젓가락', '국자' 등의 도구를 사용하여 '죽', '밥', '반찬', '찌개' 등을 먹는다. 여럿이 먹기도 한다.
- (문장 3): '대접'을 사용하여 '막걸리'를 마신다.
- (문장 4): '젓갈', '멸치액젓' 등의 재료를 사용하여 '배추김치'를 담근다.
- (문장 5): '채소', '된장' 등의 재료를 넣어 '된장찌개'를 끓인다.
- 위 모든 단어들은 식문화와 관련된다.

위와 같이 각 사회와 관련된 대량의 말뭉치(corpus)가 있다면 그 사회의 문화를 구성하는 요소들을 추출해 볼 수 있다. 이를 바탕으로 그 사회의 문화적인 특징을 분석하고, 여러 사회에 대한 문화 요소의 공통점과 차이점 등을 연구할 수 있다.

본 연구는 말뭉치에서의 단어 쓰임을 바탕으로 여러 문화 범주에서 문화를 표현하는 데 사용되는 문화 요소들이 어떤 특징을 갖는지를 살펴보기 위해 문화 요소 유사도와 문화 요소 네트워크에서의 중심성 등 정량적인 분석을 시도한다.

Ⅱ. 관련 연구

여러 방면에서 말뭉치 기반의 문화 관련 연구가 진행되고 있다. [3]은 신문 말뭉치에서 '다문화'와 한 문장에서 함께 등장하는 단어들을 연도별, 품사별로 분석하여 다문화 논의의 언어적 양상을 연구하였다. 또한 '다문화'-'이민자'와 '다문화'-'이주민'의 공통 공기어(co-occurrence)를 비교하여 '다문화'가 '이주민'과 공유하는 속성이 크므로 '다문화 가정'을 '이주민 가정'으로 명명할 것을 제안하였다. [4]는 뉴스 기사 말뭉치에서 '다문화'와 한 문서에서 공기한 키워드들을 추출하여 연대별 키워드 관계도를 분석하였고, '다문화'와 선정된 몇몇 키워드 사이의 상관계수를 계산하였다. [5]는 2000년대 초와 2010년대 초 신문 말뭉치에서 '여성', '문화', '남성'과 한 문단에서 공기한 단어들을 t 점수에 따라 랭킹하여 관련어들을 추출하고, 이를 바탕으로 여성 및 남성과 문화의 관련성을 분석하였다. 분석 결과, 여성이 남성보다 훨씬 많이 문화와 관련되었으며, 특히 사회적 문제에서 여성이 더 많이 관련되었다. [6][7]은 여성 잡지 기사 중 주거 관련 기사만으로 구성된 특수목적 말뭉치에 대해 단어의 출현 확률을 기반으로 각 단어의 KL 다이버전스(Kullback-Leibler Divergence)를 계산하여 상위 단어를 분석하였다. 이를 통해 시대에 따라 주거 관련 명사나 형용사의 쓰임이 어떻게 달라졌는지를 연구하였다. [8]은 21세기 세종계획 말뭉치와 KAIST 말뭉치에서 '입다', '벗다'의 목적어로 사용된 명사들의 쓰임을 바탕으로 '옷'의 개념이 어떻게 확장되어 사용되고 있는지를 유형화하였다. [9]는 뉴스 기사 말뭉치를 사용하여 의식주 관련 관용 표현의 사용 양상을 연구하였다. 의식주 범주별 관용 표현에 자주 등장하는 명사가 사용된 문장을 분석 대상으로 삼았다. 의 범주는 '호주머니', '탈'을, 식 범주는 '밥', '떡'을, 주 범주는 '벽', '문턱'을 사용하였다.

한편, 한국어, 중국어, 일본어 말뭉치를 기반으로 한중일 문화 요소를 추출하는 시스템을 개발하고[10][11], 이를 활용하여 어휘 기반 문화적 특징을 분석하는 연구들이 진행되었다[12][13]. [10][11]에서는 대용량 말뭉치를 색인해 두고, 질의 형태소와 한 문장에서 함께 등장한 의미 형태소들을 공기 빈도와 t 점수를 기반으로 랭킹하여 제공함으로써 말뭉치에 등장한 문화적인 특징을 분석할 수 있는 기본 자료를 제공한다. [12]는 이 시스템을 사용하여 화장실을 의미하는 한국어와 일본어 각 단어에 대해 각 말뭉치에서 공기하는 명사와 동사를 바탕으로 한일 화장실 문화의 공통점과 차이점을 분석하였다. [13]은 '배우다', '공부'에 해당하는 한국어, 중국어 각 단어와 공기하는 단어를 사용하여 프레임넷에 설정된 studying 프레임의 핵심 요소와 비핵심 요소를 비교 분석함으로써 한중 문화적인 특징을 연구하였다.

지금까지 살펴본 바와 같이 기존 연구들은 '다문화', '여성', '남성', '입다', '벗다', '화장실', '배우다', '공부' 등 특정 어휘의 쓰임을 바탕으로 한정적인 문화의 한 단면만을 분석하였다. 주거 문화[6][7]나 의식주 문화[8][9] 등으로 다소 확대된 연구도 있기는 하였으나 보편적인 문화의 특징 분석으로는 부족함이 있었다.

본 연구는 전반적으로 문화를 구성하는 구성 요소들을 추출하여 이들의 특징을 살펴보고자 하였다. '문화'나 '다문화' 등과 같이 문화 자체를 의미하는 특정 명사로는 문화가 표현된 문장들을 찾는 데는 한계가 있다. [8]에서 옷의 개념을 파악하기 위해 '입다', '벗다'와 같은 동사를 기준으로 용례를 분

석했던 것처럼, 본 연구에서는 문화를 표현할 때 사용될 만한 동사들을 먼저 선별하여 유형화하고 이들과 함께 사용된 명사들을 문화 요소로 간주하여 분석하는 방법을 취함으로써 문화 전반의 특징을 분석하였다. 특정 어휘 몇 개의 연관어들을 하나씩 살펴보는 기존 연구와는 달리, 본 연구는 58개 동사의 문화 요소 리스트에 대해 각각의 유사도를 비교한다. 결국 1,600회 이상의 유사도 비교가 필요하여 본 연구에서는 문화 요소의 유사도를 정량적으로 계산하는 방법을 사용하였다. 또한 각 문화 요소의 중요도를 파악하기 위해 문화 요소 네트워크를 구성하여 문화 요소의 중심성을 측정하였다.

Ⅲ. 연구 방법

본 연구는 말뭉치에 나타난 문화적 특징을 파악하고자 한다. 이를 위해 문화를 표현하는 동사에 집중하는데, 이런 류의 동사를 문화 동사(cultural verb)라고 칭하기로 한다. 문화 동사를 중심으로 주변 문맥에서 문화적인 요소들이 등장하게 되는데 이를 문화 요소(cultural element)라고 한다. 문화 요소 추출 시스템[10]을 통해 문화 동사의 문맥에서 문화 요소를 추출하고 문화 요소 리스트의 유사도와 문화 요소 네트워크 분석을 통해 문화적인 특징을 분석한다.

3-1 문화 동사

말뭉치에 나타난 문화 요소를 파악하기 위하여 문화 동사를 [1]의 정의를 바탕으로 표 1과 같이 선정하였다. 이 문화 동사는 21세기 세종계획 현대 문어 형태분석 말뭉치[14]에서 고빈도로 등장하는 동사 중에서 선별한 결과이다. 감정, 경제, 교육, 교통, 생사, 생활, 식, 여가, 예술, 의, 주 등 11개의 문화 범주로 구분하였고 범주별 3~9개의 문화 동사를 선정하여 총 58개의 문화 동사가 정의되었다.

3-2 문화 요소와 문맥 형태소

전산 시스템을 통해 자동으로 문화 요소를 추출하려면 문화 요소를 명확히 정의할 수 있어야 하지만, 아직 문화 요소를 명시적으로 정의하기는 어려우므로 문화 동사가 사용된 문장에서 문화 동사와 함께 등장한 문맥 형태소들을 문화 요소로 간주하여 추출한다. 본 연구에서는 문화 동사와 하나의 문장에서 함께 등장한 일반명사를 문화 요소로 추출한다. 예를 들어, 서론에서 언급한 (문장 1), (문장 2)로부터 문화 동사 '먹다'의 문화 요소로 '죽', '사람', '밥', '반찬', '찌개', '그릇', '젓가락', '국자' 등이 추출된다. 이 문화 요소들이 중복해서 등장할 경우 공기 빈도 등을 기준으로 문화 요소들을 랭킹할 수 있다[10].

표 1. 문화 동사
Table 1. Cultural Verbs
* In order to clearly convey the linguistic meaning, Hangeul is written together.

Category	Cultural Verbs
Emotion (감정)	Smile(웃다), Cry(울다), Surprise(놀라다), Like(좋아하다), Dislike(싫어하다)
Economy (경제)	Buy(사다), Sell(팔다), Earn(벌다), Borrow(빌리다), Be-sold(팔리다), Repay(갚다)
Education (교육)	Learn(배우다), Teach(가르치다), Train/Raise(키우다), Develop(기르다), Master(익히다)
Traffic (교통)	Ride(타다), Drive(몰다), Pick-up(태우다)
Birth&Death (생사)	Live(살다), Die(죽다), Bear(낳다), Be-born(태어나다), Kill(죽이다), Survive(살아가다), Save(살리다)
Living (생활)	Create(만들다), Meet(만나다), Build(짓다), Believe(믿다), Get-along-with(지내다), Argue(싸우다), Break-up(헤어지다)
Eating (식)	Eat(먹다), Drink(마시다), Set(차리다), Feed(먹이다), Boil(끓이다), Brew/Make(빚다), Crush(다지다), Make(담그다), Chew(씹다)
Leisure (여가)	Play(놀다), Enjoy(즐기다), Rest(쉬다)
Art (예술)	Sing(부르다), Draw/Paint(그리다), Dance(추다)
Clothing (의)	Wear(입다), Take-off(벗다), Put-on(신다), Wear-around(두르다)
Shelter (주)	Sleep(자다), Wake-up(깨다), Stay(머무르다), Stay(계시다), Stop-over(머물다), Put-up(묵다)

3-3 문화 요소의 유사도 비교

본 연구에서는 문화적인 특징을 파악하기 위하여 같은 문화 범주 내에서 문화 동사의 문화 요소들을 비교하고, 문화 범주 간에 문화 요소들은 어떻게 다른지를 비교 분석하였다. 이를 위해서는 서로 이질적인 두 문화 요소 랭킹을 비교할 수 있어야 한다. 두 랭킹은 서로 다른 원소들로 구성될 수 있으며 길이 또한 다를 수 있다. 또한 랭킹은 모든 원소들의 중요성을 동일시하는 집합이 아니며 상위에 랭킹된 원소가 더 중요한 의미를 갖는 리스트이다. 이런 특징을 갖는 문화 요소 랭킹의 유사성을 비교하기 위하여 본 연구에서는 평균 오버랩(average overlap; AO)[15]-[17]과 순위 편향 오버랩(rank-biased overlap; RBO)[17]을 사용한다. 이 척도들은 정보검색시스템의 검색 결과 랭킹을 비교하는 데 활용되었던 것들이다. 검색시스템 A의 검색 결과에 존재하는 항목이 검색시스템 B의 검색 결과에는 존재하지 않을 수 있으므로, 검색 결과 랭킹은 순위가 다를 뿐만 아니라 구성 원소 자체가 다를 수 있다. 또한 검색 결과 랭킹의 길이 또한 다를 수 있다. 검색 결과 전체를 비교할 때, 관측된 상위 k개의 비교 결과를 바탕으로 전체 결과에 대한 유사도를 추정하게 된다. 문화 요소 랭킹도 이와 같은 특징을 가지므로 문화 요소 랭킹의 유사도를 비교하는데 AO와 RBO를 사용한다.

1) 평균 오버랩

다음과 같은 원소들로 구성된 두 개의 리스트 A, B가 있다고 하자.

- A = [a, b, c, d, e, f, ...]
- B = [z, a, c, y, d, x, ...]

오버랩은 리스트의 각 위치에서 계산되는데, 위치 d에서 오버랩 O_d는 다음과 같이 첫 번째 위치부터 그 위치까지의 원소 중 일치하는 원소의 개수로 정의된다.

$$O_d = \left| A_{1:d} \cap B_{1:d} \right| \quad (1)$$

$A_{1:d}$와 $B_{1:d}$는 각 리스트의 첫 번째 원소부터 d번째 원소까지로 구성된 부분집합을 의미한다. 위 예의 경우 첫 번째 원소가 일치하지 않으므로 첫 번째 위치에서의 오버랩 $O_1 = 0$이고, 두 번째 위치까지는 공통 원소 a가 하나 포함되므로 $O_2 = 1$이다. 같은 방식으로 $O_3 = 2$, $O_4 = 2$, $O_5 = 3$이다.

오버랩을 바탕으로 일치율(O_d/d)을 계산할 수 있는데 이 일치율을 사용하여 평균 오버랩을 계산한다. 위치 k에서의 평균 오버랩 $AO(k)$는 위치 1부터 위치 k까지의 일치율들을 누적한 후 평균한 값이다[15]-[17].

$$AO(k) = \frac{1}{k} \sum_{d=1}^{k} \frac{O_d}{d} \quad (2)$$

위 예의 경우 $AO(1) = 0$, $AO(2) = 0.25$, $AO(3) = 0.389$, $AO(4) = 0.417$, $AO(5) = 0.453$ 등으로 계산된다. 평균 오버랩은 교차 척도(intersection metric)[15], 평균 정확도(average accuracy)[16]로도 불린다.

평균 오버랩을 사용하여 두 문화 요소 랭킹의 유사성을 측정할 수 있다. 그러나 비교하는 랭킹 결과가 전체 리스트가 아닌 상위 일부이므로 전체 리스트에 대한 유사성을 추정할 수 있는 방법이 필요한데 순위 편향 오버랩이 이런 목적으로 사용된다.

2) 순위 편향 오버랩

순위 편향 오버랩 RBO는 위치 k까지의 일부 리스트에 대한 유사도를 바탕으로 관측하지 못한 전체 리스트의 유사도를 외삽법(extrapolation)을 통해 추정하는 방법으로서, 다음 수식과 같이 계산된다[17].

$$RBO_{EXT}(p, k) = \frac{O_k}{k} p^k + \frac{1-p}{p} \sum_{d=1}^{k} \frac{O_d}{d} p^d \quad (3)$$

각 위치에서 발생하는 오버랩은 가중치가 부여되는데, 위치 d에서 발생한 오버랩에는 $(1-p)p^{d-1}$의 가중치가 부여된다. 모든 가중치의 합이 1이 되도록 부여된 가중치이다. p는 $0 < p < 1$의 값을 가지는데, p는 순위가 내려갈수록 가중치가 얼마나 가파르게 낮아질 것인지를 결정하는 파라미터이다. p가 작을수록 가중치가 급격히 줄어들어 상위 오버랩만을 고려하여 추정하게 되고, p가 1에 근접하여 클수록 가중치 감소 폭은 거의 없어 하위 원소들까지 고려하게 된다.

3-4 문화 요소 네트워크 분석

소셜 네트워크나 키워드 네트워크에서 노드의 중심성(centrality)을 측정하여, 한 노드가 네트워크에서 갖는 중요도나 영향력을 파악할 수 있다[18]-[20]. 본 연구에서는 각 문화 요소가 문화 표현에 얼마나 중요한지를 알아보기 위해, 문화 요소 네트워크를 구성하고 각 문화 요소의 중심성을 측정하였다.

문화 요소를 노드(node)로 하고 동일한 문화 동사의 문화 요소로 등장했는지의 여부에 따라 간선(edge)을 연결하는 문화 요소의 네트워크를 구성할 수 있다. 이 네트워크에서 연결 중심성(degree centrality), 근접 중심성(closeness centrality), 아이겐벡터 중심성(eigenvector centrality), 매개 중심성(betweenness centrality) 등을 사용하여 각 문화 요소의 영향력을 측정하였다. 중심성이 높을수록 문화를 표현하는데 중요한 역할을 수행하는 문화 요소로 해석할 수 있다.

1) 연결 중심성

노드 N_i의 연결 중심성 C_D는 다음과 같이 각 노드의 연결 개수를 바탕으로 측정된다[19].

$$C_D(N_i) = \frac{\sum_{j=1}^{g} x_{ij}}{g-1}, \ i \neq j \quad (4)$$

여기서 g는 노드의 총 개수를 나타내고, x_{ij}는 노드 N_i와 N_j가 연결된 경우 1 아니면 0의 값을 갖는다. 즉, 노드 N_i의 연결 중심성은 N_i에 연결된 노드의 개수를 계산한다. 네트워크 크기로 정규화하기 위해 $g-1$로 나눈다. 따라서 연결 중심성은 0에서 1 사이의 값을 갖는다. 노드의 활동성을 측정하는 중심성으로서, 다른 문화 요소와 직접적으로 많이 연결될수록 연결 중심성이 높다.

2) 근접 중심성

노드 N_i의 근접 중심성 C_C는 다음과 같이 계산된다[19].

$$C_C(N_i) = \frac{g-1}{\sum_{j=1}^{g} d(N_i, N_j)}, \ i \neq j \quad (5)$$

여기서 $d(N_i, N_j)$는 노드 N_i와 노드 N_j 사이의 최단경로거리이다. 직접적인 연결이 아니더라도 다른 노드와의 거리가 가까운 노드의 근접 중심성이 높게 측정된다. 다른 문화 요소에 의존하지 않고 독립적으로 문화 요소들에 신속하게 도달할 수 있는 문화 요소의 근접 중심성이 높다.

3) 아이겐벡터 중심성

아이겐벡터 중심성은 노드의 개수뿐만 아니라 연결된 노드의 중요도까지 고려한다. 중심성이 높은 노드와 연결될수록 더 높은 중심성을 갖도록 계산된다. 노드 N_i 의 아이겐벡터 중심성 C_E는 다음과 같이 계산된다[19].

$$C_E(N_i) = \lambda \sum_{j=1}^{g} x_{ij} C_E(N_j), \ i \neq j \quad (6)$$

여기서 λ는 아이겐값(eigenvalue)을 의미한다. 노드 자체의 연결 정도가 낮더라도 연결된 노드의 연결 정도가 높다면 아이겐벡터의 중심성은 높아질 수 있다. 아이겐벡터 중심성이 높은 문화 요소는 문화 표현에서의 인기도나 영향력이 높은 것으로 해석할 수 있다.

4) 매개 중심성

매개 중심성은 다른 노드들 사이에서 노드들을 서로 연결해주는 역할을 강조한 중심성이다. 노드 N_i 의 매개 중심성 C_B는 다음과 같이 계산된다[19].

$$C_E(N_i) = \frac{2}{(g-1)(g-2)} \sum_{j<k} \frac{g_{jk}(N_i)}{g_{jk}}, \ i \neq j \neq k \quad (7)$$

여기서 g_{jk}는 노드 N_j와 N_k 사이의 최단 경로의 개수를, $g_{jk}(N_i)$는 노드 N_j와 N_k 사이의 최단 경로에 노드 N_i가 포함된 경로의 개수를 의미한다. 네트워크 크기와 관계없이 중심성 값이 0에서 1 사이의 값을 갖도록 $\frac{2}{(g-1)(g-2)}$ 값을 곱한다. 매개 중심성은 직접 연결되지 않은 노드 간의 관계를 통제 및 중개하는 정도를 측정하므로, 문화 요소들에 대한 통제력이 높은 문화 요소의 매개 중심성이 높게 측정된다.

Ⅳ. 실험 결과 및 분석

실험에는 21세기 세종계획 현대 문어 형태분석 말뭉치[14]를 사용하였다. 표 2에 보인 바와 같이, 이 말뭉치의 크기는 약 1천만 개 어절이며, 책, 신문, 잡지, 기타 비출판물, 전자출판물로 구성되어 있다.

표 2. 21세기 세종계획 현대 문어 형태분석 말뭉치의 구성
Table 2. Composition of the Modern Written Morphological Analysis Corpus of the 21st Century Sejong Project

Media	Eojeols	Ratio
Book	7,253,857	72.1%
Newspaper	1,625,713	16.1%
Magazine	1,012,017	10.1%
Non-published Work	103,996	1.0%
E-publication	71,139	0.7%
Total	10,066,722	100.0%

이 중 책이 72.1%로 가장 많은 비중을 차지한다. 문화 동사와 한 문장에서 함께 사용된 일반명사를 문화 요소로 추출하되, 공기 빈도를 기준으로 랭킹하였다. 노이즈를 최소화하기 위해 공기 빈도가 3 이상인 것만 추출하였다. '말', '사람', '때', '일', '속', '정도' 등은 모든 문화 동사의 문맥으로 추출되어, 이들을 포함하여 너무나 일반적인 명사들은 불용어(stopword)로 미리 정의하여 이들은 문화 요소로 추출되지 않도록 하였다. 평균 오버랩과 순위 편향 오버랩은 Python RBO 모듈[21]을, 문화 요소 네트워크의 중심성은 NetworkX 모듈[22]를 사용하여 계산하였다. RBO_{EXT} 계산 시 파라미터 p는 0.98로, k는 100으로 설정하였다.

4-1 문화 범주 내 문화 요소 유사도 비교

표 3은 문화 범주별 한 문화 범주에 속한 여러 문화 동사들의 문화 요소 리스트 간의 유사도를 평균한 결과이다. 예를 들어, 감정 문화 범주에 속한 5개의 문화 동사에 대한 문화 요소 리스트 간의 유사도를 평균 오버랩(AO)으로 각각 계산한 후 이들을 평균한 결과가 0.4103이었다. 실험 결과 평균 오버랩 기준으로 문화 요소가 가장 유사한 문화 범주는 교육, 생사, 감정 순이었으며 문화 요소가 가장 상이한 문화 범주는 식, 주, 예술 순이었다. 순위 편향 오버랩(RBO)으로 추정한 유사도도 같은 경향성을 보였다.

표 4는 가장 유사한 범주인 교육 범주에서 가장 유사했던 '배우다'와 '가르치다'의 문화 요소와 가장 상이했던 '키우다'와 '익히다'의 문화 요소를 상위 20위까지만 나열한 것이다. '배우다'와 '가르치다'는 '학교', '교육', '사회', '영어', '학생', '선생', '필요', '대학', '공부' 등의 많은 문화 요소가 일치하여 평균 오버랩이 0.7048이었다. '배우다'와 '가르치다'는 의미적으로 상반되지만, 주체만 바뀔 뿐 그 의미 표현에 동반되는 요소들은 유사함을 확인할 수 있었다. 반면 '키우다'와 '익히다'의 문화 요소는 상대적으로 일치하는 것이 적어 평균 오버랩이 0.2330이었다. 표 4의 교육 범주 4개 문화 동사의 문화 요소에는 '학교', '교육', '사회', '필요' 등의 문화 요소가 공통적으로 존재하였다.

표 3. 문화 범주 내 문화 요소 리스트 사이의 유사도
Table 3. Similarity between Cultural Element Lists within a Cultural Category

Category	AO	RBO
Emotion	0.4103	0.3855
Economy	0.3388	0.3357
Education	0.4323	0.4002
Traffic	0.3234	0.3209
Birth&Death	0.4132	0.3845
Living	0.2987	0.2687
Eating	0.1990	0.1845
Leisure	0.2760	0.2435
Art	0.2586	0.2307
Clothing	0.2916	0.2753
Shelter	0.2316	0.2193

표 4. 교육 범주의 문화 요소
Table 4. Cultural Elements in the Education Category
* Since it is the result of automatic extraction from the Korean corpus, Hangul was used to clearly convey the linguistic meaning.

Rank	Learn	Teach	Train/Raise	Master
1	학교	교육	능력	기술
2	교육	학생	교육	필요
3	사회	선생	집	방법
4	영어	학교	사회	생활
5	학생	교사	자식	법
6	선생	방법	아들	사회
7	기술	대학	꿈	글
8	방법	영어	기업	기본
9	컴퓨터	공부	남편	음식
10	필요	집	힘	교육
11	책	길	아기	사용
12	대학	사회	여성	컴퓨터
13	지식	문제	필요	공부
14	공부	연구	애	얼굴
15	글	글	여자	감각
16	나라	내용	사랑	책
17	언어	법	어머니	학교
18	과정	문학	개	기능
19	법	과정	엄마	대학
20	역사	필요	학교	효과

표 5는 범주 내 문화 요소가 가장 상이했던 식(eating) 문화 범주의 예이다. 식 문화 범주에서 가장 유사한 문화 요소는 '먹다'와 '씹다', '먹다'와 '마시다'이었다. 또한 식 문화 내에서 가장 상이한 문화 요소는 '빚다'와 '씹다'이었다.

'먹다'는 먹는 대상인 '밥', '음식', '고기', '빵', '떡' 등과 끼니와 연관되어 '아침', '점심', '저녁'이 문화 요소로 추출되었다. '마시다'는 '술', '물', '커피', '차' 등의 대상과 '잔', '병', '친구', '이야기' 등이 추출되었다. 삼시세끼에 음식을 먹는 상황과 일상에서 다소 일탈하여 편한 분위기에서 마시는 상황이 그려지는 문화 요소의 구성이다.

식 범주 4개 문화 동사에서 공통으로 존재하는 문화 요소는 '술', '물' 뿐이었다. 식 범주의 문화 동사별 문화 요소는 매우 다양하였지만, 식문화에서 '술'과 '물'은 빠질 수 없음을 확인할 수 있었다.

실험에는 표 1의 모든 문화 동사에 대한 모든 문화 요소가 사용되었으나 지면 관계상 표 4, 5에는 일부 동사에 대한 일부 문화 요소만 제시하였다.

표 5. 식 범주의 문화 요소
Table 5. Cultural Elements in the Eating Category
* Since it is the result of automatic extraction from the Korean corpus, Hangul was used to clearly convey the linguistic meaning.

Rank	Eat	Drink	Brew/Make	Chew
1	밥	술	술	입
2	음식	물	물의	껌
3	집	커피	차질	맛
4	저녁	잔	갈등	밥
5	마음	차	흙	음식
6	점심	소주	논란	빵
7	술	맥주	문제	여자
8	물	집	사태	얼굴
9	맛	병	마찰	입술
10	약	친구	현상	오징어
11	돈	이야기	사회	다리
12	아침	한잔	결과	배
13	나이	우유	반죽	소리
14	배	담배	송편	모습
15	고기	밤	물	술
16	어머니	여자	가루	목
17	입	맛	누룩	물
18	빵	입	말썽	집
19	떡	혼자	정부	잔
20	식사	얘기	소동	조각

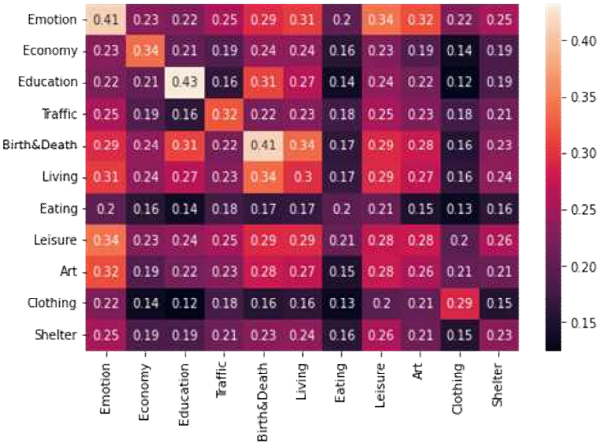

그림 1. 문화 범주 간 문화 요소 유사성
Fig. 1. Similarity of Cultural Elements across Cultural Categories

한편, '키우다'나 '익히다', '빚다'는 단어에 의미적인 중의성이 존재하여 '개', '음식', '물의' 등이 문화 요소로 추출되었는데 이는 향후 연구에서 개선이 필요한 사항이다.

4-2 문화 범주 간 문화 요소 유사도 비교

문화 범주 간에 문화 요소가 얼마나 유사한지를 살펴보기 위해 각 문화 동사의 문화 요소 리스트 간의 유사도를 계산한 후 문화 범주간의 유사도를 계산하였다. 예를 들어, 의(clothing) 문화 범주와 식 문화 범주 사이의 유사도는 의 범주에 해당하는 4개 문화 동사의 문화 요소 리스트와 식 범주에 해당하는 9개 문화 동사의 문화 요소 리스트 각각의 유사도를 계산한 후 이들을 평균하여 계산하였다.

그림 1은 평균 오버랩을 기반으로 각 문화 범주 사이의 문화 요소 유사도를 히트 맵으로 표현한 것이다. 히트 맵에서 숫자가 크고 음영이 옅은 것이 유사도가 높은 것을 의미한다. 문화 범주 간 유사도를 종합적으로 파악하기 위해 표 6는 자신을 제외한 타 문화 범주와의 유사도를 평균한 결과이다. 즉, 표 6에서 감정 범주는 감정을 제외한 나머지 10개 범주와의 유사도를 계산한 후 평균한 값이 평균 오버랩 기준으로 0.2646이었다. 표 7은 문화 범주별 가장 유사한 범주와 가장 상이한 범주를 정리한 것이다.

그림 1, 표 6, 표 7에서 확인할 수 있듯이 감정 범주는 전체적으로 다른 문화 범주와의 유사도가 높다. 여가, 생사, 생활 범주도 다른 문화 범주와 유사도가 높은 편이다. 반면, 식, 의 범주는 다른 범주와 유사도가 매우 낮다. 감정 범주와 여가 범주 사이의 유사도가 가장 높았으며, 교육 범주와 의(clothing) 범주 사이의 유사도가 가장 낮았다.

그림 2는 이 범주들의 문화 요소들로 워드 클라우드를 생성한 결과이다. 감정, 여가, 교육, 의 등 4가지 범주 이외에도 비교를 위하여 생사, 생활, 식, 주 범주를 포함하였다.

표 6. 타 문화 범주와의 문화 요소 리스트의 유사도 평균
Table 6. Average Similarity of Cultural Element Lists with Other Cultural Categories

Category	AO	RBO
Emotion	0.2646	0.2370
Economy	0.2001	0.1791
Education	0.2078	0.1804
Traffic	0.2113	0.1884
Birth&Death	0.2528	0.2250
Living	0.2512	0.2246
Eating	0.1666	0.1452
Leisure	0.2587	0.2305
Art	0.2353	0.2081
Clothing	0.1677	0.1491
Shelter	0.2089	0.1888

표 7. 문화 범주별 최대 최소 유사 범주
Table 7. Most and Least Similar Category by Cultural Category

Category	Max	Min
Emotion	Leisure	Eating
Economy	Birth&Death	Clothing
Education	Birth&Death	Clothing
Traffic	Emotion	Education
Birth&Death	Living	Clothing
Living	Birth&Death	Clothing
Eating	Leisure	Clothing
Leisure	Emotion	Clothing
Art	Emotion	Eating
Clothing	Emotion	Education
Shelter	Leisure	Clothing

워드 클라우드는 범주별 각 문화 요소의 공기 빈도를 합산한 결과를 바탕으로 생성하였다.

감정 범주의 문화 요소로는 '소리', '얼굴', '어머니', '여자', '집' 등이 자주 등장하였고, 생사 범주에서는 '집', '삶', '인간', '세상', '사회' 등이 자주 등장하였다. 생활 범주는 '집', '인간', '사회', '여자', '이야기' 등이, 여가 범주에서는 '집', '한숨', '새', '놀이', '하루' 등이 자주 등장하였다. 교육 범주에서는 '교육', '학교', '학생', '사회', '선생' 등이, 식 범주에서는 '술', '밥', '물', '음식', '집' 등이, 의 범주에서는 '옷', '여자', '집', '남자', '바지' 등이, 주 범주에서는 '잠', '집', '밤', '방', '어머니' 등이 자주 등장하였다.

* Hangeul was used because the picture was automatically generated from the raw Korean data.

그림 2. 문화 요소 클라우드
Fig. 2. Cultural Elements Cloud

4-3 문화 요소 네트워크 분석

표 8은 문화 요소들로 네트워크를 구성하여 문화 요소의 연결 중심성, 근접 중심성, 아이겐벡터 중심성, 매개 중심성을 측정한 결과 상위 20개의 문화 요소들이다. 문화 요소 네트워크는 총 58개 문화 동사의 문화 요소로 등장한 단어들을 노드로 생성하고, 동일한 문화 동사의 문화 요소로 등장한 단어 쌍에 대해 모두 간선을 연결하였다. 이처럼 노드 간의 직접 연결이 많아, 연결 중심성과 근접 중심성의 순위가 동일하게 나타났다. '집', '마음', '여자', '길', '어머니' 등이 다른 문화 요소와 빈번히 연결되었음을 뜻하며, 이는 이들이 여러 문화 범주나 문화 동사에서 문화 요소로 사용된다는 의미이다. 아이겐벡터 중심성으로 측정된 문화 요소의 영향력이나 매개 중심성으로 측정된 문화 요소의 통제력 또한 비슷한 경향을 보인다.

표 8. 문화 요소의 중심성
Table 8. Centrality of Cultural Elements
* Since it is the result of automatic extraction from the Korean corpus, Hangul was used to clearly convey the linguistic meaning.

Rank	Degree Centrality	Closeness Centrality	Eigenvector Centrality	Betweenness Centrality
1	집	집	집	집
2	마음	마음	마음	어머니
3	여자	여자	여자	마음
4	길	길	길	길
5	어머니	어머니	아버지	아버지
6	소리	소리	어머니	소리
7	아버지	아버지	소리	여자
8	모습	모습	모습	물
9	남	남	남	모습
10	자리	자리	이야기	자리
11	물	물	돈	문제
12	선생	선생	선생	남
13	돈	돈	남자	선생
14	이야기	이야기	자리	얼굴
15	남자	남자	문제	남자
16	문제	문제	사이	돈
17	얼굴	얼굴	얼굴	술
18	사이	사이	사회	이야기
19	사회	사회	물	사회
20	인간	인간	인간	사이

중심성 측정 결과를 종합해보면 4가지 중심성에서 모두 '여자', '남자', '어머니', '아버지', '인간' 등 문화의 중심에 있는 인간을 지칭하는 단어들의 중심성이 큰 것으로 분석되었다. 특히, '집', '어머니', '아버지' 등이 문화를 표현하는 데 매우 중요한 영향력을 가지고 있음을 확인할 수 있었다. 이는 가족 중심의 문화적인 특징이 반영된 결과로 해석된다. 또한 기존 연구 [5]의 결과와 마찬가지로 '여자'가 '남자'보다 문화 표현에서 더 중요한 역할을 차지하고 있음을 확인할 수 있었다. 4가지 중심성 모두 비슷한 경향을 보이는데, 이는 같은 동사의 문화 요소로 출현한 경우에만 간선을 연결하여 문화 요소 명사들 사이의 관계 생성이 제한되었기 때문으로 추정된다.

V. 결 론

본 연구는 문화를 11개의 범주로 구분하여 각 범주의 문화를 표현하는 문화 동사를 선정하였다. 대용량 말뭉치로부터 문화 동사와 한 문장에서 함께 등장한 일반명사들을 문화 요

소로 추출하였고, 공기 빈도를 기반으로 랭킹하였다. 각 문화 요소 랭킹의 유사도를 평균 오버랩과 순위 편향 오버랩 척도를 사용하여 측정함으로써 문화적 특징을 정량적으로 분석하였다. 분석 결과, 문화 동사들이 비슷한 문화 요소를 공유하는 경향을 띠는 범주는 교육, 생사, 감정 등이었으며, 식, 주, 예술 등의 범주는 같은 범주이더라도 문화 동사들이 서로 상이한 문화 요소를 동반하는 경향을 보였다. 범주 사이의 유사도를 검토했을 때는 감정, 생사, 생활 등의 범주가 타 범주와 유사한 문화 요소를 가지고 있었고, 식, 의, 경제 범주는 타 범주와는 구분되는 범주 특유의 문화 요소를 가지고 있었다. 문화 요소 네트워크를 통해 분석한 결과 가족과 관련된 어휘들의 중심성이 높아, 가족 중심의 문화적인 특징을 정량적인 분석을 통해서 확인할 수 있었다.

문화 동사를 중심으로 공기한 명사들을 문화 요소로 추출하는 과정에서 '키우다', '익히다', '빚다' 등 어휘 자체의 의미적 중의성으로 인해 추출된 문화 요소가 적절치 못한 경우들이 발생하였다. 향후 이런 노이즈를 최소화하기 위한 개선이 필요하다. 한편, 본 연구의 실험은 형태분석까지 완료된 정제된 말뭉치를 통해 수행하였으나, 향후 대량의 원시 텍스트와 자동 형태소 분석기를 사용하여 실험 대상을 확장할 필요도 있다. 또한 여러 시대에 걸친 말뭉치를 바탕으로 시대적인 문화 특징의 변화를 비교하는 것도 향후 연구로서 의미가 있겠다.

참고문헌

[1] National Institute of Korean Language. Standard Korean Dictionary − Culture [Internet]. Available: https://stdict.korean.go.kr/search/searchView.do?word_no=424006&searchKeywordTo=3.

[2] C.H. Paek, "Cultura," *Philosophy and Reality*, Vol., No. 26, pp. 294-303, September 1995.

[3] Y.S. Choi, "A Study of the Co-occurring Words with Muliticulture -focusing on the Korean Newspaper Corpus," *Journal of Multi-Cultural Contents Studies*, Vol., No. 24, pp. 275-301, April 2017. https://doi.org/10.15400/mccs.2017.04.24.275

[4] C. Lee, "Multicultural Aspects Analysis Using News Articles Corpus," *The Journal of Humanities and Social Science*, Vol. 12, No. 1, pp. 1405-1418, February 2021.

[5] B.M. Kang, "The Relation between Women/Men and Culture and its Changes Reflected in the Newspaper Corpus," *Eoneohag: Journal of the Linguistic Society of Korea*, Vol., No. 94, pp. 31-55, December 2022. https://doi.org/10.17290/jlsk.2022..94.31

[6] Y. Oh, "Lexical Changes in the Korean Residential Culture Corpus: Focused on Semantic Classification of Nouns," *Language and Information*, Vol. 24, No. 3, pp. 27-45, November 2020. https://doi.org/10.29403/LI.24.3.2

[7] Y. Oh, "Contemporary Korean Residential Sensibility Represented in the Corpus," *The Korean Cultural Studies*, Vol. 39, pp. 173-209, January 2020. https://doi.org/10.17792/kcs.2020.39..173

[8] H.Y. Jeon, "Aspects of Conceptualization Manifested in the [clothing] Metaphors by Koreans-with a focus on the verbs 'ipda(put on)/budda(take off)'," *The Korean Cultural Studies*, Vol. 19, pp. 129-162, January 2010. https://doi.org/10.17792/kcs.2010.19..129

[9] H.J. Song, "Motivation of Korean Food, Clothing, and Shelter Idiomatic Expressions," *Korean Semantics*, Vol. 58, pp. 185-209, December 2017.

[10] J.S. Lee and K.S. Han, "A Trial of Constructing Multi-lingual Cultural Element Mining System(CEMS)," *Journal of Japanese Language and Literature*, Vol. 99, No. 1, pp. 289-304, November 2016. http://doi.org/10.17003/jllak.2016.99.1.289

[11] J.S. Lee, K.S. Han, and W.G. Roh, "A Trial to Construct the Cultural-Image-Frame-Network Based on Big-data Framework," *The Japanese Language Association of Korea*, Vol., No. 65, pp. 131-142, September 2020. http://doi.org/10.14817/jlak.2020.65.131

[12] H.Y. Kim, "A Possibility of Korean-Japanese Vocabulary and Culture Education Utilizing Text Mining: Vocabulary and Culture Related to toire and hwajangsil," *Korean Journal of Japanese Language and Literature*, Vol. 1, No. 92, pp. 139-156, March 2022. https://doi.org/10.18704/kjjll.2022.03.92.139

[13] Y.J. Yun, "A Comparative Study on the Cultural Elements of the 'Studying' Frame between Korean and Chinese using Text Mining," *The Journal of Chinese Cultural Studies*, Vol., No. 55, pp. 191-215, February 2022. https://doi.org/10.18212/cccs.2022..55.009

[14] National Institute of Korean Language, *21st Century Sejong Project Final Result*, Revised Edition, 2011.

[15] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing Top k Lists," *SIAM Journal on Discrete Mathematics*, Vol. 17, No. 1, pp. 134-160, 2003. https://doi.org/10.1137/S0895480102412856

[16] S. Wu and F. Crestani, "Methods for Ranking Information Retrieval Systems without Relevance Judgments," in *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC '03)*, New York: NY, pp. 811-816, 2003. https://doi.org/10.1145/952532.952693

[17] W. Webber, A. Moffat, and J. Zobel, "A Similarity Measure for Indefinite Rankings," *ACM Transactions on Information Systems*, Vol. 28, No. 4, Article 20, 38 pages, November 2010. https://doi.org/10.1145/1852102.1852106

[18] P. Bonacich, “Power and Centrality: A Family of Measures,” *American journal of Sociology*, Vol. 92, No. 5, pp. 1170-1182, March 1987. https://doi.org/10.1086/228631

[19] K.Y. Kwahk, *Social Network Analysis,* 2nd ed. South Korea, Seoul: Cheonglam Pub., pp. 182-229, 2017.

[20] S. Jang, D. Han, and C. Oh, “Identifying the Ethical Issues of Virtual Human: A Semantic Network Analysis of Media Reports,” *Journal of Digital Contents Society*, Vol. 23, No. 11, pp. 2307-2316, November 2022. https://doi.org/10.9728/dcs.2022.23.11.2307

[21] Python Software Foundation. Rank-biased Overlap (RBO) [Internet]. Available: https://pypi.org/project/rbo/.

[22] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring Network Structure, Dynamics, and Function using NetworkX,” in *Proceeding of the 7th Python in Science Conference (SciPy2008)*, Pasadena: CA, pp. 11-15, 2008.

한경수(Kyoung-Soo Han)

1998년 : 고려대학교 컴퓨터학과 (학사)
2000년 : 고려대학교 대학원 (이학석사)
2006년 : 고려대학교 대학원 (이학박사-전산학)

2006년～2009년: SK텔레콤
2009년～현 재: 성결대학교 컴퓨터공학과 교수
※관심분야 : 텍스트 마이닝(Text Mining), 질의응답시스템(Question Answering System), 정보검색(Information Retrieval) 등