# Vocabulary Analysis of the Archives of Design Research: Utilizing Corpus and Text Mining Techniques

Eunyoung Kim[1], Byunghak Ahn[2*]

[1]Department of Visual & Media Design, Lecturer, Sookmyung Women's University, Seoul, Korea
[2]Department of Visual Communication Design, Professor, Hongik University, Seoul, Korea

## Abstract

**Background** There have been noticeably fewer studies of design terminology in comparison to the studies of the practical design field of South Korea. The study of design terminology, especially the study that looks into the aspect of the use of vocabulary, corresponds to the foundation of design research. This study examines whether the aspect of the use of vocabulary in the design field of South Korea can be technically analyzed using a text mining technique and whether it will be a meaningful method for the study of design terminology.

**Methods** Having constructed the corpus from 2,214 papers written in Korean published in the Archives of Design Research*(Korean Society of Design Science)*, I intended to grasp the aspect of the use of vocabulary with a text mining technique. In the analyzing process, Python libraries and analytical techniques of frequency and distribution(similarities) were used which are widely utilized in natural language processing.

**Results** We looked into a possibility to find a meaningful vocabulary list and similarity among the vocabulary from the design corpus through text mining. We confirmed that the sufficiently interesting findings from the study are related to the design terminology through a series of results.

**Conclusions** This study is the first case to construct a corpus in the design field and to analyze the aspect of the use of vocabulary with a text mining technique. I expect that this study contributes to the vitalization of the study of design terminology with various viewpoints in the future.

**Keywords** Design Terminology, Text Mining, Natural Language Processing, Design Corpus, Design Research

*Corresponding author: Byunghak Ahn (ahn.hisd@gmail.com)

*Citation:* Kim, E., & Ahn, B. (2020). Vocabulary Analysis of the Archives of Design Research: Utilizing Corpus and Text Mining Techniques. *Archives of Design Research, 33*(1), 205-217.

http://dx.doi.org/10.15187/adr.2020.02.33.1.205

**Received :** Oct. 16. 2019 **; Reviewed :** Dec. 17. 2019 ; **Accepted :** Dec. 17. 2019
**pISSN** 1226-8046 **eISSN** 2288-2987

## 1. 서론

디자인은 매해 그해의 트렌드가 발표될 정도로 기술과 환경 변화에 민감한 분야이다. 그만큼 분야 내에 존재하는 용어의 변화 속도도 빠를 수밖에 없다. 그러나 한국에서 디자인 용어에 대한 연구는 디자인의 실천적 분야 연구에 비해 현저히 적은 실정이다. 디자인 분야는 최신 기술과 분야에 대한 적극적 수용을 통해 영역 확장을 거듭하고 있으며, 경영학과 공학 등의 연관 분야에서는 제 분야에서 디자인의 중요성을 인식하고 주의 깊은 관찰을 통해 '디자인적 사고(디자인씽킹)'와 같은 개념을 제안하며, 새로운 용어들을 생성하고 있다. 디자인의 융합적 속성상 이러한 현상은 당연하고 자연스럽다. 그러나 이와 더불어 디자인의 고유 지식과 개념에 대한 불확실성이 증가하고 있는 현실은 우려스럽다. 달리 말하면, 디자인 지식과 연관 지식을 구분하고 설명하는 디자인 분야의 주체적 관점과 기준이 더욱 모호해 지고 있다는 의미이다.

언어의 최소 단위로서 단어는 시간에 따라 변화하고 생성되고 사멸되면서 자신이 속한 세상을 표상한다. 디자인 분야의 어휘 즉 디자인 용어 역시 디자인이라는 세상을 표상한다. 디자인이 타 분야들이 이루는 교집합의 합으로서가 아닌 독자적인 분야로서 존재하기 위해서는 디자인 분야가 스스로를 규명하고 체계를 세우는 연구가 필수적이며, 디자인 용어에 대한 연구, 특히 용어의 사용 양상을 살피는 연구가 그 토대에 해당한다.

한국에서 그간 디자인 용어 또는 디자인 어휘를 분석하고자 한 연구는 학술지 논문 6건 정도가 있다. 이중 5건은 '디자인'이라는 단일 어휘를 깊이 있게 다뤘으며 1건은 디자인과 관련된 감성어휘에 초점을 맞추고 있다.

Table 1 Domestic studies on design terms or design vocabulary

| 연구자 | 논문 제목 (발행일) | 발행처 |
|---|---|---|
| 신수길, 연현정, 이진승 | 3대 일간지를 통한 디자인용어 변천에 관한 연구 (2012.4) | 한국디지털디자인협의회 |
| 임영빈, 나건 | 디자인에서의 고급감에 관한 의미 및 감성어휘 연구 (2015.6) | 한국디자인문화학회 |
| 김효진, 나건 | 다 매체 시대에 있어서 디자인 용어 변천에 관한 연구 (2018.2) | 디지털정책학회 |
| 김효진, 나건 | 디자인 관련 키워드 분석을 통한 디자인 용어 개념 변화에 대한 연구 (2018.3) | 한국디자인문화학회 |
| 김효진, 나건 | 디자인 용어 활용 확장에 따른 본질적 의미 변화에 관한 고찰 (2018.6) | 한국디자인문화학회 |

위의 연구들은 일간지 혹은 디자인 관련 전문 매체를 주도적으로 혹은 부분적으로 분석에 활용하고 있으나 한국 디자인 전반의 어휘 사용 양상을 분석하고자 하는 이 연구와는 연구 대상과 분석 기법이 사뭇 다르다. 현재 디자인 분야에서 텍스트마이닝 기법을 사용하는 연구는 대개 디자인 인식 조사, 사용자 경험 측정 등 디자인 리서치 측면으로 진행되고 있으며 건축, 무용, 컴퓨터공학 등의 분야와 같이 학문 차원에서 텍스트마이닝 기법을 연구 동향이나 지식 구조, 용어 분석에 활용한 예는 찾아보기 어렵다.

## 2. 분석 대상과 분석 방법

연구자는 이 연구에서 텍스트마이닝 기법을 이용하여 한국 디자인 분야의 어휘 사용 양상을 데이터 기반으로 분석할 수 있을지, 그것이 디자인 용어 연구에 유의미한 방식일지를 살피고자 한다. 제목과 본문에서 '용어' 대신 '어휘' 혹은 '단어'를 사용한 이유는 이 연구의 분석 범위를 '디자인 용어'가 아닌 '디자인 분야에서 쓰는 말'로 설정했기 때문이다. 디자인 분야에서 쓰는 모든 말이 디자인 용어일 수는 없으며, 기술적 분석을 토대로 디자인 용어의 정의와 기준을 제시하기 위해서는 지속적인 연구가 필요하다. 이 연구의 연구 질문은 아래와 같다.

첫째, 텍스트마이닝을 통해 한국 디자인 분야의 주요 어휘를 추출하고 그 출현 양상을 파악할 수 있는가?

둘째, 텍스트마이닝을 통해 위의 주요 어휘와 다른 어휘들의 의미 관계를 파악할 수 있는가?

'주요 어휘'의 정의는 연구자마다 다를 수 있다. 연구자는 중요한 어휘는 다른 어휘에 비해 자주 사용될 것이라는 보편적 믿음을 전제로, 정보검색에서 주로 활용되는 단어 빈도와 분포 분석 기법을 활용했다.

일반적으로 전문용어는 특정한 단일 개념을 지칭하기 위해 인위적으로 만들어지며, 일상과 거리가 있는 전문적 개념을 지칭하기 때문에 보통 낯설고 어렵다. 김한샘(2015:133)은 “전문용어는 이러한 특성 때문에 어휘의 생명 주기로 보았을 때 갓 생겨난 신어일 가능성이 높고, 새로운 지식과 개념을 어휘와 함께 국외에서 들여올 가능성이 높기 때문에 외래어의 비율이 높다.”고 설명하며 “새로 생긴 전문용어의 정착은 반드시 시간의 흐름을 동반”하기 때문에 “결과에 초점을 맞춘 공시적 연구와 함께 일정 시간 경과 후의 변화를 살펴보는 통시적 연구가 필요하다.”고 했다.

연구자는 앞서 제시한 두 가지 연구 질문에 답하면서도 위의 맥락에서 공시적, 통시적 비교가 가능한 말뭉치를 구축하고자 했다.

이 연구에서는 선택한 대상 데이터는 〈디자인학연구〉에 게재된 논문이다. 한국 디자인 어휘의 분석 대상으로서 한국디자인학회에서 발행하는 단일 학술지를 선택한 이유는 다음과 같다.

첫째, 어휘 분석을 위해서는 말뭉치 곧 대량의 텍스트 데이터가 필요하며 개인 연구자가 디자인이라는 특정 분야에 국한된 양질의 텍스트를 대규모로 구할 수 있는 방법은 현재 논문 외에는 존재하지 않는다. 둘째, 어휘 분석을 위해서는 앞서 수집한 대량의 텍스트를 분석 가능한 상태로 정제하는 전처리 작업이 필요하며 이는 문서의 구조에 일정한 패턴이 있고 띄어쓰기가 올바로 되어 있을 때 효율이 높다. 논문은 학회에서 지정한 서식에 맞춰 작성해야 하므로 단일 학술지의 게재 논문을 대상으로 하면 전처리 시간을 단축할 수 있다. 셋째, 〈디자인학연구〉는 위의 조건을 만족하면서도 한국에서 가장 오래되고 인지도 높은 통합 디자인 학술지로서 디자인 분야에서 충분한 대표성을 가지고 있다.

연구자는 2만 쪽 이상의 텍스트 데이터로부터 의미 있는 패턴을 찾아내기 위해 텍스트마이닝 기법을 이용하였고 첫 번째 연구질문에 답하기 위해 빈도 분석을, 두 번째 연구질문에 답하기 위해서는 분포 분석을 실시했다. 분석을 수행하는 도구로는 프로그래밍 언어인 파이썬과 관련 자연어처리 라이브러리를 사용했다.

## 3. 말뭉치 구축과 전처리

연구자는 〈디자인학연구〉에 1989년부터 2018년까지 게재된 한글(한문 혼용 포함) 문서 pdf 2,214건을 사용하여 말뭉치를 구축했다. 1980년 창간호에는 단 두 편이 실렸고 그중 하나는 창간사인데다가 이후 두 번째 호까지 꽤 긴 공백기가 있었기에 제외했다. 역시 평균 2쪽짜리 발표문으로 구성된 학술대회 논문집과 영어 논문, 각 호의 발간사, 추천사, 논문 투고 안내문 등은 분석에서 제외했고, 대담이나 논평류는 포함했다. 이로써 정리한 분석 대상 문서 수와 단어 수는 표 2와 같다.

Table 2 Number of articles (volumes) and words analyzed per year

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1989(1) | 9 | 42,519 | 1997(2) | 45 | 182,895 | 2005(4) | 111 | 474,538 | 2013(4) | 67 | 265,865 |
| 1990(1) | 9 | 22,013 | 1998(3) | 73 | 286,091 | 2006(6) | 127 | 554,014 | 2014(4) | 32 | 121,746 |
| 1991(1) | 5 | 18,909 | 1999(4) | 85 | 349,333 | 2007(6) | 130 | 511,160 | 2015(4) | 36 | 131,773 |
| 1992(1) | 6 | 17,753 | 2000(4) | 93 | 375,015 | 2008(6) | 110 | 415,310 | 2016(4) | 37 | 131,901 |
| 1993(1) | 6 | 21,191 | 2001(4) | 91 | 372,417 | 2009(6) | 123 | 437,578 | 2017(4) | 27 | 111,049 |
| 1994(1) | 3 | 12,932 | 2002(4) | 131 | 561,433 | 2010(7) | 140 | 539,668 | 2018(4) | 37 | 165,200 |
| 1995(2) | 27 | 116,451 | 2003(4) | 157 | 670,474 | 2011(6) | 134 | 529,679 | | | |
| 1996(4) | 82 | 378,310 | 2004(4) | 165 | 698,897 | 2012(5) | 116 | 432,328 | | | |

텍스트마이닝은 크게 데이터 전처리 단계와 데이터 분석 단계로 나뉜다. 얼마나 잘 처리된 데이터를 사용하는지에 따라 데이터 분석 결과의 품질이 달라지므로 전처리 과정은 상당히 중요하다.

연구자는 앞서 수집한 pdf 파일을 텍스트 분석이 가능한 상태로 만들기 위해 먼저 광학문자인식(OCR) 프로그램을 이용하여 htm 파일로 변환했다. 다음으로 문서 파일에 포함된 학술정보 포털업체의 저작권 안내, 이용정보 등 텍스트 분석에 필요하지 않은 텍스트, 그림과 표, 단락 표시자 외에 불필요한 htm 태그를 삭제하고 줄바

꿈 과정에서 발생한 띄어쓰기 오류 등을 전반적으로 교정하는 클리닝 작업을 수행하여 xml 문서를 생성하였다. 이 과정에서 파악한 2,214개 문서에 사용된 어휘 수는 총 8,948,442개이다.
마지막 처리 과정으로서 텍스트를 형태소 분석하여 품사를 태깅하고 문장, 문서 단위로 구조화한 json 파일을 생성했다. 형태소 분석에는 오픈소스 형태소 분석기인 Mecab을 사용했다. 이때 Mecab 사전에 미등록된 디자인 고유명사의 임의 분절을 피하기 위해 한국에서 출간된 디자인 사전과 타이포그래피 사전의 수록어 2,650여 개를 목록화하여 사용자 정의 사전으로 활용했다.

## 4. 데이터 분석

### 4. 1. 단어 빈도 기반 분석

앞서 전처리한 json 파일을 활용하여 일반명사(NNP), 고유명사(NNP), 어근(XR) 세 개 품사를 가진 2음절 이상의 단어로 한정하여 1989~2018년 〈디자인학연구〉에 나타난 고빈도 어휘를 분석했다. 전문용어는 보통 2음절 이상의 명사 형태를 띠기 때문이며, 어근을 분석에 포함한 이유는 어근의 범위에 명사가 포함되기 때문이다.
하나의 문서에서 출현 빈도가 높다는 것은 해당 단어가 중요함을 의미하지만, 하나의 문서가 아닌 모든 문서에서 출현 빈도가 높다는 것은 해당 단어가 보편적으로 자주 사용되는 단어이거나 분석 대상 텍스트의 유형적 특성에서 비롯된 단어일 가능성을 보여준다. 출현 빈도는 높지만 실질적 의미 분석에는 필요하지 않은 이런 단어들을 불용어(stopword)라고 한다. 예를 들어, 디자인 연구 논문에서 '디자인'(119,236)과 '연구'(60,944)라는 단어가 가장 많이 출현하는 것은 당연하다. 이런 단어를 불용어로 처리하지 않으면 연도별 혹은 문서별 텍스트를 비교할 때 모든 텍스트의 고빈도 어휘가 비슷하게 도출되어 유의미한 발견을 하기 어렵다. 텍스트에 출현한 단어들의 통계적 특성으로 색인어를 선정하는 방법을 최초로 제시한 Luhn은 단어의 빈도 분포에서 중간빈도의 단어들이 문헌 내용의 식별력이 가장 크며 이를 색인어로 선정하도록 제시했다. 그러나 중간빈도의 최고 한계치와 최저 한계치를 산출하는 구체적인 공식까지 함께 제시되어 있지는 않다. (정영미, 2012:76-77)
연구자는 빈도 분석 결과에서 가장 빈도가 높은 단어 100개를 추출한 뒤 4.2.의 방법으로 주변 단어 분포를 함께 살피면서 어느 지점까지 불용어로 처리해야 할지 단위별로 검토했다. 이와 별개로 논문에서 소제목과 캡션에 일관되게 등장하는 어휘 5개와 학술지 이름인 '디자인학연구'를 불용어 목록에 추가하여 최종적으로 표 3과 같이 48개를 확정했다.

Table 3 48 stopwords applied to 〈ADR〉 analysis (in order of frequency)

| | |
|---|---|
| 고빈도 어휘 (42) | (디자인(119236), 연구(60944), 사용(57605), 제품(33837), 공간(33432), 분석(32290), 요소(29791), 이미지(29185), 표현(28733), 정보(28051), 방법(26350), 형태(25954), 환경(23093), 개발(23010), 문화(23000), 결과(22409), 과정(22400), 그림(22331), 이러(21512), 다양(20833), 경우(20774), 기능(19946), 의미(19799), 광고(19549), 구성(19084), 특성(18387), 시각(18072), 평가(17991), 사회(17936), 가능(17931), 변화(17747), 필요(17067), 개념(16780), 조사(16764), 기술(16398), 활용(15660), 교육(15563), 대상(15243), 구조(15012), 관계(14977), 중요(14929), 산업(14394) |
| 논문어휘(6) | 참고문헌, 참고, 문헌, 배경, 논문, 디자인학연구 |

불용어로 처리하기는 했지만, 표 3의 고빈도 어휘는 꽤 흥미롭다. 해당 목록은 논문에서 주로 쓰이는 '분석, 요소, 방법, 결과, 과정' 등 단어 외에도 다양한 성격의 단어들을 포함하고 있다. '제품, 공간, 정보, 환경, 광고, 교육, 문화, 사회' 그리고 '이미지, 표현, 형태, 구성, 시각' 등의 단어들은 〈디자인학연구〉에 게재된 논문들의 주된 관심 분야와 영역을 드러낸다. 현재의 말뭉치를 추후 균형적으로 확장한다면, 한국 디자인 일반의 경향 역시 엿볼 수 있을 것으로 판단한다.

연도별 주요 어휘를 분석하는 데 활용한 TFIDF는, 정보검색 이론에서 한 단어가 특정 문서에 출현한 빈도와 전체 문서에서 출현한 빈도를 비교하여 그 문서에서 얼마나 중요하게 사용되었는지 평가하는 데에 널리 사용된다. 이 가중치를 이용하면 전체 출현 빈도는 적지만 특정 문서에서 비중 있게 사용된 어휘를 버리지 않을 수 있다.

TFIDF는 단어 출현 빈도(TF)와 역문헌 빈도(IDF)의 곱으로 계산하며, 한 문서에 출현하는 빈도가 높으면서 전체 문서에 출현하는 빈도가 낮을수록 높은 점수를 얻는다. 이 가중치를 적용하여 추출한 연도별 주요 어휘 상위 10개 단어와 앞서와 동일하게 불용어를 제거한 뒤 단어 출현 빈도(TF)로 추출한 10개 단어를 표 4로 정리했다. 한문 사용이 빈번했던 90년대 초반 논문의 경우 불용어 처리가 되지 않으므로 최종 결과에서 한문을 한글로 바꾼 뒤 불용어를 걸렀다. 지면 한계상 목록을 다 실을 수 없어 표 4에서는 잘 드러나지 않지만, 현재 연도별 TFIDF 결과에는 사람 이름이 많이 포함되어 있다. 이는 논문의 특성상 타 연구의 인용 시 저자 이름을 기입해야 하고, 참고문헌 목록 역시 포함해야 하기 때문으로 보인다. (이번 탐색 과정에서는, 참고문헌 제목에 해당 논문에서 다루는 주제와 연관된 키워드가 포함되는 경우가 많고 본문에 내주로 사용한 저자 이름까지 제거하기는 거의 불가능하므로 전처리 과정에서 참고문헌 부분을 제거하지 않은 상태이다.)

Table 4 Top 10 words by year in 〈ADR〉 (where * represents Chinese characters)

| | | |
|---|---|---|
| 1989 | TF | 형태*, 생산*, 설계*, 생활*, 염색*, 항목, 자인, 편조*, 체계*, 문제* |
| | TFIDF | 편조*, 단지*, 왕정*, 부업*, 금제*, 실태*, 기존*, 주옥*, 중요*, 정도* |
| 1990 | TF | 관광, 민예품, 지역, 상업, 상품, 가치*, 쇼핑, 생산, 산업, 척도* |
| | TFIDF | 문항*, 간격*, 구간*, 판단*, 계열*, 상한*, 평정*, 물건*, 수상*, 민예품 |
| 1991 | TF | 전시, 오피스, 사무, 조직, 전문, 인간, 계획, 기업, 전시회, 부분 |
| | TFIDF | 아가미, 속성*, 요구*, 조작*, 직적*, 콩깍지, 집합*, 잠수함, 변환*, 사무원 |
| 1992 | TF | 도시, 산업, 학문, 현금, 포스트, 자동, 한국*, 오픈, 지급기, 스페이스 |
| | TFIDF | 지급기, 실내*, 소공원, 점무늬, 금자동, 혁신*, 주거*, 한옥*, 본식*, 내부* |
| 1993 | TF | 가공, 중심*, 중량, 온도, 습도, 상품, 소비자, 중합, 표시, 비례* |
| | TFIDF | 그래프트, 독성*, 견섬유, 모노머, 타닌, 산신*, 공견, 중합체, 중합, 단량체 |
| 1994 | TF | 객체, 시스템, 그래픽, 스케치, 지향, 하이퍼, 발상, 처리, 컴퓨터, 텍스트 |
| | TFIDF | 뷰트, 염화금산, 염호, 발색제, 환원제, 듀어, 견사, 염욕, 프로테아제, 염산 |
| 1995 | TF | 영상, 언어, 문제, 인간, 컴퓨터, 전통, 세계, 사람, 한국, 시대 |
| | TFIDF | 푸치, 빼로, 숙련가, 혜강, 토큰, 라지에타, 칠성, 최한기, 논구, 관념시 |
| 1996 | TF | 색채, 인간, 산업, 문제, 생산, 조형, 단계, 예술, 작품, 기업 |
| | TFIDF | 산드, 포락면, 이중직, 관두, 시넥틱스, 연설회, 편색, 유권자, 모노그램, 까시나 |
| 1997 | TF | 기업, 인간, 조형, 문제, 예술, 포장, 단계, 생산, 소비자, 디자이너 |
| | TFIDF | 건어물, 퍼즈, 원시주의, 인랜드, 도봉구, 우미, 리버티, 고드윈, 우버, 문촌 |
| 1998 | TF | 산업, 효과, 기업, 컴퓨터, 인간, 상품, 내용, 디자이너, 조형, 작업 |
| | TFIDF | 개더, 사외보, 워스트, 창작동화, 강아지똥, 크워스, 비개이, 캐랙터, 레토르트, 앙케트 |
| 1999 | TF | 대학, 산업, 인간, 컴퓨터, 효과, 시스템, 관련, 내용, 속성, 작업 |
| | TFIDF | 납석, 염포, 공동육아, 신발장, 임꺽정, 낙농, 별신굿, 각문, 청옥, 안마루 |
| 2000 | TF | 내용, 산업, 인간, 디자이너, 시스템, 효과, 문제, 작업, 적용, 모델 |
| | TFIDF | 아바칸, 아바카, 굴삭기, 커뮤터, 바카노, 장농, 골프장, 코벨, 탈구, 팔메트 |
| 2001 | TF | 인간, 이용, 제작, 효과, 인터넷, 소비자, 단계, 기업, 작업, 사이트 |
| | TFIDF | 와편, 죽재, 크리스토, 산드, 동명대학, 조직선, 장방전, 극공, 눈박이, 편죽 |
| 2002 | TF | 기업, 소비자, 산업, 내용, 한국, 이용, 감성, 인간, 효과, 시간 |
| | TFIDF | 조직선, 순천시, 윗글, 샤프펜슬, 소학교, 브로셔, 반달리즘, 의장도, 통경, 등가성 |
| 2003 | TF | 적용, 사이트, 이용, 효과, 인간, 내용, 브랜드, 단계, 경험, 색채 |
| | TFIDF | 배사, 배열표, 스넬러, 버너, 삼륜차, 태극무늬, 노블티, 홍보인쇄물, 짜개, 혁필화 |
| 2004 | TF | 산업, 인간, 한국, 감성, 중심, 이용, 관련, 효과, 경험, 내용 |
| | TFIDF | 옥죽, 프레즌스, 디에이, 홍삼, 효제, 쓰레드, 은물, 앞다리, 무신도, 유량계 |
| 2005 | TF | 브랜드, 지역, 중심, 인간, 감성, 산업, 학습, 소비자, 방식, 한국 |
| | TFIDF | 인제군, 눈축제, 소격, 이노센스, 은파, 본당, 담양군, 오시이, 페티시, 마모루 |
| 2006 | TF | 브랜드, 소비자, 요인, 한국, 실험, 유형, 제작, 제시, 가치, 기업 |
| | TFIDF | 이브랜드, 하울, 약령시, 유산소, 갑천, 서울우유, 쌍폭, 아이원, 각분, 회복실 |
| 2007 | TF | 브랜드, 한국, 색채, 실험, 게임, 인터페이스, 시간, 이용, 영향, 감성 |
| | TFIDF | 기억색, 의안, 올드보이, 연판문, 트레일, 약령시, 연창, 임곡, 홍채, 펜션 |

| | | |
|---|---|---|
| 2008 | TF | 브랜드, 중심, 감성, 한국, 인간, 사람, 분류, 게임, 디지털, 제작 |
| | TFIDF | 웨이팅, 주밍, 시맨틱, 시큐리티, 스타니슬라프스키, 어윈, 하우어, 브랜치, 인벤, 부여군 |
| 2009 | TF | 기업, 브랜드, 경영, 적용, 문제, 인식, 유형, 중심, 시스템, 사례 |
| | TFIDF | 시즈오카, 정군, 무하, 결착, 퍼셉트론, 지봉, 영비, 풍류도, 득정, 반사회 |
| 2010 | TF | 기업, 도시, 브랜드, 적용, 중심, 실험, 영향, 단계, 전략, 사례 |
| | TFIDF | 포충, 엘이디, 분전반, 서비스스케이프, 에니어그램, 하이틴, 우키요에, 빅게임, 앙굴렘, 토커 |
| 2011 | TF | 감성, 기업, 사례, 중심, 한국, 브랜드, 색채, 경험, 단계, 도시 |
| | TFIDF | 재클린, 모예, 온열, 아우토반, 저탄소, 정촌, 어버니즘, 서플리, 로서리, 안주인 |
| 2012 | TF | 경험, 브랜드, 가치, 서비스, 한국, 기업, 영향, 중심, 작용, 사례 |
| | TFIDF | 카시트, 화장비누, 한흥, 서울색, 레터마크, 예악, 인라인스케이트, 이노션, 오판, 디자인유산 |
| 2013 | TF | 서비스, 브랜드, 경험, 제공, 산업, 적용, 관련, 기업, 사례, 소비자 |
| | TFIDF | 파티나, 복대, 막걸리, 족저, 둘레길, 창간사, 권태민, 운반자, 오마주, 사이클링 |
| 2014 | TF | 시간, 시장, 소재, 역량, 실험, 디자이너, 경험, 단계, 자전거, 영향 |
| | TFIDF | 르펜, 이리스, 태양새, 군론, 문복, 화고, 데님, 피로회, 봉황*, 집필진 |
| 2015 | TF | 브랜드, 색채, 경험, 지역, 가치, 도시, 실험, 주거, 패션, 아이콘 |
| | TFIDF | 마라도, 콜라보레이션, 아동화, 홍비, 왜색, 미토스, 장상, 성형외과, 정왕동, 홍시 |
| 2016 | TF | 국가, 문제, 영향, 참여, 사례, 공공, 서비스, 요인, 프로세스, 정책 |
| | TFIDF | 민원서류, 재능기부, 민원인, 기부자, 파장동, 모빌리티, 밥솥, 자아상, 위임장, 사리면 |
| 2017 | TF | 지도, 범죄, 효과, 거리, 활동, 지역, 장소, 게임, 명화, 보행 |
| | TFIDF | 게이미피케이션, 안전교육, 인플레, 접지층, 성혈, 태식, 느와르, 취식, 추격자, 보유량 |
| 2018 | TF | 시장, 전통, 영향, 개선, 적용, 중심, 문제, 색채, 행위, 사례 |
| | TFIDF | 아이소타이프, 필머, 서비스스케이프, 달개, 심정지, 편성표, 아이스하키, 아카이빙, 슈타이너, 이송 |

표 4에서 TF 목록으로는 연도별로 당시 디자인 연구자들이 공통적으로 빈번하게 사용한 어휘를 알 수 있고, TFIDF 목록으로는 같은 시기에 개별 논문들에 비중 있게 사용된 단어들을 살필 수 있다. 예를 들어 1994년에는 컴퓨터와 디지털 관련 단어의 출현 빈도가 전반적으로 높아졌지만, 동시에 공예 관련 단어를 비중 있게 사용하는 논문 역시 존재함을 알 수 있다.
다만, 현재 TF, TFIDF 목록은 연도별 문헌 길이의 차이를 고려하지 않은 단순 빈도를 활용한 값으로 이후 활용 목적에 따라 문헌 길이에 대해 정규화를 거친다면 일부 결과가 달라질 수 있다.

다음으로 연구자는 연도별 고빈도 20개 어휘의 빈도 추이를 그래프로써 살펴보고자 했다. 이를 통해 각 단어의 생명주기와 더불어 서로 생명주기가 겹치는 단어 간에 상관관계를 파악할 수 있지 않을까 생각했기 때문이다. 이 단계에서는 전체 말뭉치 중 1989~1994년의 논문은 어휘 규모가 다른 해에 비해 지나치게 적어 결과를 신뢰할 수 없기에 부득이 제외했다. 나머지 데이터에서도 2002~2012년의 말뭉치 규모가 이후 기간의 규모에 비해 두 배 이상 크기 때문에 연도별 단어 빈도를 해당 연도의 총 단어 수로 정규화하는 과정을 거쳤다.

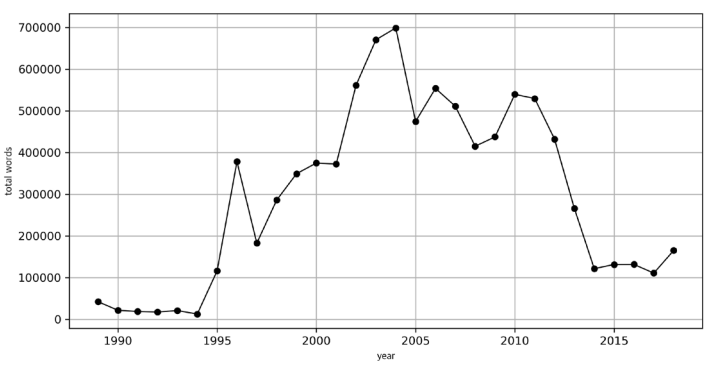


**Figure 1** 〈ADR〉 annual difference in corpus size

연구자는 단어별로 생성한 그래프를 아래와 같이 크게 7개 유형으로 분류하였다.
첫째, 전 기간에 걸쳐 일정 빈도 이상으로 꾸준히 나타나는 유형으로 '감정, 관련, 내용, 다음, 단계, 모델, 목적, 문제, 방식, 방향, 부분, 분야, 사람, 생활, 시간, 시스템, 영역, 예술, 유형, 이용, 이해, 인식, 자연, 작용, 작품, 적용, 전략, 전체, 제공, 제시, 중심, 차이, 체계, 측면, 한국, 행위, 효과'가 있으며 둘째, 꾸준히 나타나되 연도별 편차가 상당히 큰 유형으로 '경영, 로봇, 문양, 미술, 상호, 색채, 속성, 전통, 지역, 포장'이 있다.
셋째, 특정 시기에만 출현이 집중된 유형으로 '공방, 소재, 언어, 윤리, 자전거, 토큰'이 있으며 넷째, 그 특정 시기가 최근에 집중되었거나 갑작스레 최근 2~3년 내에 빈도가 급증한 유형으로 '거리, 게임, 계획, 국가, 남성, 명화, 범죄, 보행, 시장, 신체, 아이콘, 역량, 장소, 재질, 정책, 주거, 지도, 프로세스, 해학, 활동'이 있다.

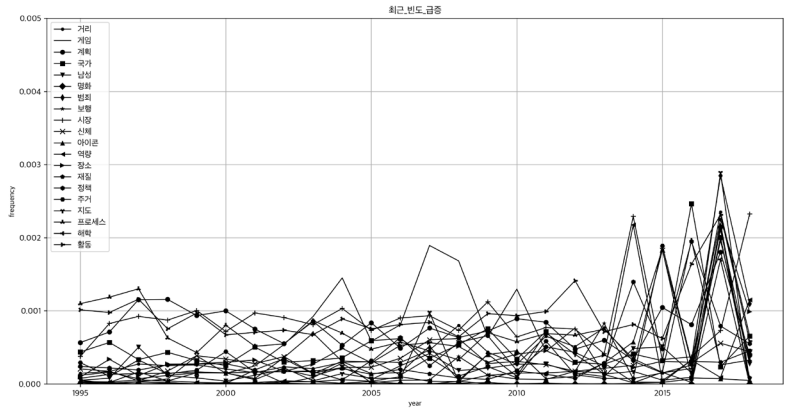


**Figure 2** 20 words with a sharp increase in frequency in the last two to three years

다섯째, 연도별 변동 폭은 다소 있으나 전반적으로 감소 추세를 보이는 유형으로 '기본, 기호, 대학, 디자이너, 발전, 상품, 생산, 세계, 소비자, 시대, 양식, 인간, 작업, 재료, 조형, 컴퓨터'가 있으며 여섯째, 전반적으로 증가 추세를 보이는 유형으로 '개선, 사례, 사업, 상황, 실험, 영향, 요인, 장애, 정부, 패션'이 있다.
일곱째, 출현 빈도가 한동안 일정한 수준을 유지하거나 증가하다가 점차 감소 추세를 보이는 유형으로 '가치, 감성, 경험, 공공, 기업, 도시, 디지털, 분류, 브랜드, 사이트, 서비스, 애니메이션, 영상, 인터넷, 인터페이스, 제작, 참여, 캐릭터, 커뮤니케이션, 학습'이 있다.

### 4. 2. 단어 간 유사도 분석

디자인 분야에서는 오랜 시간에 걸쳐 하나의 개념이 하나의 용어로 다듬어지기보다 시대 흐름에 맞춰 새로운 개념과 용어가 빠르게 나타났다가 사라지는 경향이 있다. 비슷한 개념이 서로 다른 용어로 제시되어 혼선을 일으키는 경우가 잦으며, 같은 개념이라도 누군가는 외국 용어를 발음 그대로 혹은 로마자 표기법에 따라 적고 누군가는 서로 다른 말로 번역하여 표기하곤 한다. 연구자는 이러한 어휘 사용 양상을 파악하기 위해서는 단어들의 의미 관계 특히 유의 관계를 볼 수 있어야 한다고 판단했다. 유의 관계에서 동의어, 반의어 등 세부 관계를 파악하기 위해서는 상당히 전문적인 연구가 필요하지만, 단어나 문장 간 유사도 분석에 대해서는 기계학습 분야에서 이미 많은 연구가 진행되었고 검증된 라이브러리도 존재한다.
단어 간 유사도는 단어 임베딩을 통해 단어의 의미를 벡터화한 뒤 다차원 공간에 표현하는 방식으로 계산한다. 단어 임베딩을 통해 벡터화된 단어들은 연산도 가능하다. 단어 임베딩은 단어의 분산 표현이라고도 하며 타깃 단어의 주변 단어와의 관계를 담은 밀집벡터를 표현한 것이다. 최근에는 이 작업에 2013년 구글에서 발표한 추론 기반 기법인 Word2Vec 알고리즘을 주로 사용한다. Word2Vec는 비지도학습의 일종으로 CBOW와 skip-gram 두 가지 학습 방법 중 하나를 선택하여 사용할 수 있다. 연구자는 앞뒤 단어를 통해 타깃 단어를 예측하는 skip-gram 방식을 학습에 사용했다. 학습 데이터는 앞서 만들어 둔 json 파일을 이용하였고, 분석 도구로는 파이썬 genism 라이브러리의 Word2Vec 클래스를 사용했다. 빈도 분석할 때와 마찬가지로 단어 유형은 일반명

사, 고유명사, 어근 세 개 품사를 가진 2음절 이상 단어로 한정했다. 학습 시 벡터공간의 차원 크기는 100개, 단어의 맥락을 뜻하는 윈도(window) 크기는 5개, 학습 횟수는 100회로 설정하였고 전체 출현 빈도가 50 미만인 어휘는 분석 효율을 위해 제외했다. 학습 토큰(어휘) 수는 5,545,190개이며, 학습에 걸린 시간은 00:15:55이다.

학습을 마친 Word2Vec 모델에 등록된 어휘는 6,915개이다. 코사인 유사도를 이용하여 주요 어휘에 대한 유의단어 상위 13개를 정리하면 표 5와 같다. 코사인 유사도가 클수록 벡터공간에서 두 어휘 간의 관계가 가까움을 의미한다.

Table 5 Similar words by word for 10 high-frequency words

| | 기준 단어에 대한 유사 어휘 상위 13개와 코사인 유사도 |
|---|---|
| 산업 | (공업, 0.6743673086166382), (제조업, 0.6697900295257568), (수출, 0.6144658923149109), (경제, 0.614357054233551), (육성, 0.6121785044670105), (진흥, 0.610385537147522), (진흥, 0.5961518287658691), (선진국, 0.5948867797851562), (고부, 0.5941715836524963), (경쟁력, 0.5767738819122314), (설립, 0.5753041505813599), (기술, 0.5682474374771118), (정책, 0.5640063285827637) |
| 브랜드 | (브랜, 0.7984333634376526), (자산, 0.7178537249565125), (아이덴티티, 0.695055365562439), (랜딩, 0.6639623045921326), (소비자, 0.6507193446159363), (기업, 0.6408684849739075), (인지도, 0.6380709409713745), (쿼티, 0.6208349466323853), (로고, 0.616268515586853), (네임, 0.614116370677948), (전략, 0.6131088733673096), (스토어, 0.6114202737808228), (브렌드, 0.6089605689048767) |
| 인간 | (육체, 0.6809476017951965), (본성, 0.6658341884613037), (본능, 0.6454765200614929), (유기체, 0.6339219808578491), (생명체, 0.6038764715194702), (외계, 0.5923397541046143), (사람, 0.5885531902313232), (인류, 0.5718858242034912), (사물, 0.5687993764877319), (우주, 0.5657267570495605), (자연, 0.5656843781471252), (인공, 0.5651911497116089), (원초, 0.564018726348877) |
| 기업 | (경영, 0.7493628859519958), (자사, 0.7467575073242188), (중소기업, 0.7079539895057678), (다국적, 0.6566455960273743), (전략, 0.6523627042770386), (브랜, 0.6441393494606018), (브랜드, 0.6408684253692627), (경영자, 0.6252350807189941), (마케팅, 0.6229760646820068), (회사, 0.6194640398025513), (소비자, 0.6182168126106262), (자산, 0.6090850830078125), (사내, 0.6016970872879028) |
| 효과 | (극대, 0.6937471628189087), (시너지, 0.6384539604187012), (효율, 0.6151406764984131), (기대, 0.6105225682258606), (활용, 0.5780884027481079), (증대, 0.5568416714668274), (방안, 0.5466914176940918), (유발, 0.5414680242538452), (촉진, 0.5352616310119629), (제고, 0.5326005220413208), (파급, 0.5325725078582764), (방법, 0.52884441614151), (기법, 0.5257129669189453) |
| 중심 | (위주, 0.6467703580856323), (중점, 0.6394760608673096), (초점, 0.6068586111068726), (한편, 0.5736681222915649), (기반, 0.5680197477340698), (주도, 0.5416107177734375), (주요, 0.5403997898101807), (핵심, 0.524506688117981), (관점, 0.5228261947631836), (바탕, 0.5132979154586792), (근간, 0.5132592916488647), (한양, 0.5057476758956909), (대표, 0.5008050203323364) |
| 이용 | (이용한, 0.6817575097084045), (활용, 0.6792874932289124), (사용, 0.6448403596878052), (기법, 0.5348072648048401), (적용, 0.5211700201034546), (각종, 0.5196045637130737), (고안, 0.5122220516204834), (응용, 0.498163640499115), (용이, 0.4961937367916107), (기타, 0.490857005119323 73), (방법, 0.4893469214439392), (제작, 0.48470646142959595), (일반, 0.484178364276886) |
| 한국 | (일본, 0.6945984363555908), (우리나라, 0.6605163812637329), (나라, 0.6568058133125305), (학회, 0.649865984916687), (한국인, 0.6125739812850952), (중국, 0.6016074419021606), (흥원, 0.5893231630325317), (협회, 0.5773638486862183), (학회지, 0.5756537914276123), (전통, 0.5557723641395569), (산업자원부, 0.5455669164657593), (통권, 0.544651985168457), (연합회, 0.5296597480773926) |
| 관련 | (포함, 0.6700097322463989), (각종, 0.6346242427825928), (연관, 0.6218062043190002), (선행, 0.6162725687026978), (기타, 0.6077620983123779), (국내외, 0.5928460955619812), (문헌, 0.587039589881897), (분야, 0.5850914716720581), (주요, 0.5842400789260864), (미비, 0.5733816623687744), (밀접, 0.5732154846191406), (서적, 0.5631219148635864), (이외, 0.5626921057701111) |
| 적용 | (활용, 0.6974302530288696), (응용, 0.6842111945152283), (제안, 0.67865389585495), (접목, 0.660029947757721), (구현, 0.6400392055511475), (사례, 0.6314212679862976), (제시, 0.6218027472496033), (개발, 0.5976897478103638), (도입, 0.5787987112998962), (가이드라인, 0.5693635344505 31), (적합, 0.5358203053474426), (바탕, 0.533772349357605), (모색, 0.5328781604766846) |

표 5의 결과는 주어진 말뭉치로 Word2Vec 학습을 통해 추론한 값이며 데이터 전처리 상태에 따라 학습이 잘못될 가능성이 존재한다. 그러나 연구자가 결과값을 살핀 바로는 즉각적으로 맥락을 유추하기 어려운 단어가 일부 존재하였으나 전반적으로는 결과를 수긍할 만했다. 경험적으로 봤을 때 오류로 느껴지거나 단어 간 유사 관계를 직관적으로 알기 어려웠던 예는 아래와 같다.

- 산업: (진흥, 0.5961518287658691), (고부, 0.5941715836524963)
- 브랜드: (브랜, 0.7984333634376526), (랜딩, 0.6639623045921326), (퀴티, 0.6208349466323853)
- 한국: (흥원, 0.5893231630325317), (통권, 0.544651985168457)
- 중심: (한양, 0.5057476758956909)

말뭉치에서 위의 어휘들이 사용된 문장들을 검토해 본 결과 위의 문제는 다음과 같은 이유에서 비롯되었다.

1) OCR에서 글자 인식 오류
   예. 한국 디자인 '진흥' 체제의 현황 및 문제점 (진흥)
2) OCR에서 띄어쓰기 인식 오류 또는 미등록어의 형태소 오분석
   예1. '고부+가+가치' 데님 패션+의 (고부가가치)
   예2. 기업+은 감성 '브+랜딩'+을 통하+여 (브랜딩)
   예3. '브랜드 에+퀴티'+의 이론+적 고찰 (브랜드 에퀴티)
3) 형태소 오분석으로 일어난 학습 오류
   예1. '한국'+디자인학회 '통권' 00호
   예2. ('중심'이 포함된 논문 제목) '한양'+대학교

위의 항목 중 1과 2는 OCR이나 형태소 분석 등 데이터 전처리 단계에서 발생한 오류이며 3은 전처리 오류가 학습의 오류로 이어진 경우이다. OCR 과정에서 발생하는 오류는 해당 작업이 필요 없도록 원 텍스트 파일을 말뭉치 구축에 활용하거나 말뭉치의 규모를 크게 키워서 해당 오류의 발생률을 평균으로 수렴시키는 방식이 이상적이다. 형태소 분석 사전에 미등록어, 단어가 형태소 분석 사전에 등록되어 있지 않아서 발생하는 오분석은 당장 근본적 해결이 어렵다. 분석 텍스트에 출현하는 모든 단어를 일일이 찾아내어 사전에 추가하는 것은 불가능하기 때문이다. 이 문제를 개선하기 위해서는 오분석이 많이 발생하는 어휘군에 대해서 사용자 정의 사전을 지속해서 보강해 나가야 한다.

마지막으로 연구자는 〈디자인학연구〉에 사용된 전체 어휘들이 어떤 의미 단위로 분포되어 있는지 t-SNE 그래프를 생성하고, 밀집별 어휘 목록을 밀도 기반 군집화 알고리즘인 DBSCAN을 이용하여 추출하였다. t-SNE는 고차원 벡터들의 상관관계를 유지하면서 저차원으로 변환하는 알고리즘으로, 다른 차원 축소 알고리즘에 비해 군집 간 변별력을 잃지 않기에 데이터 시각화를 위해 많이 사용된다.
t-SNE와 DBSCAN는 모두 scikit-learn의 클래스로 수행하였으며, DBSCAN 추출 시에는 군집별 최소 어휘 수를 5로 설정하고, 데이터 포인트로부터 이웃 데이터까지 거리(epsilon)의 최댓값을 1.75로 설정하였다. 추출된 군집 수는 총 93개이다. 이중 최소 어휘 수를 만족하지 못해 잡음(-1)으로 분류된 경우, 주로 고빈도 어휘가 모이는 군집으로 군집의 성격이 명확히 드러나지 않는 경우(0)를 제외하고 소속 어휘 개수 기준의 상위 10개 군집을 정리하면 표 6과 같다.

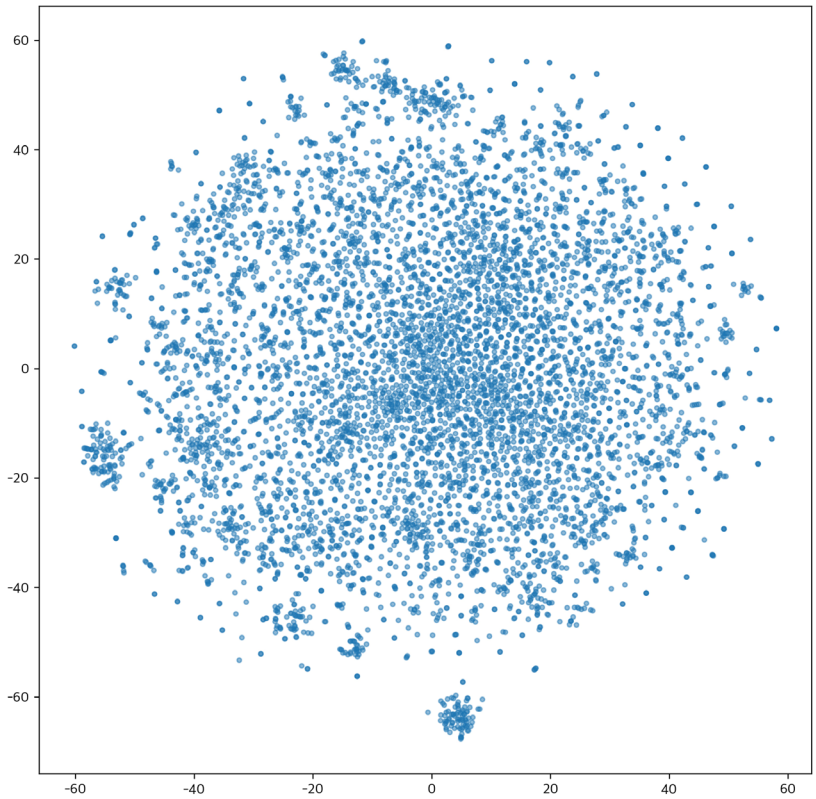


**Figure 3** Distribution of words used in 〈ADR〉 (t-SNE, minimum frequency above 50)

Table 6 Top 10 clusters by cluster size

| | |
|---|---|
| 군집1 (170) | 재료, 소재, 장식, 재질, 가공, 금속, 유리, 나무, 투명, 플라스틱, 부품, 타일, 마감, 패널, 목재, 조립, 온도, 장신구, 공예품, 도자기, 마감재, 화학, 성형, 귀금속, 첨가, 도자, 대나무, 알루미늄, 궁궐, 도장, 건조, 식기, 가죽, 핸드, 사찰, 광택, 스틸, 그릇, 주조, 금형, 대문, 견고, 보석, 석재, 내구, 민예품, 수지, 고무, 소성, 주석, 바닥재, 대리석, 기물, 카펫, 창덕궁, 원료, 가옥, 벽돌, 용접, 중량, 경량, 합금, 습도, 콘크리트, 아크릴, 비닐, 폴리, 유약, 레스, 접시, 천연, 도료, 동합금, 단단, 시편, 우드, 벤딩, 수분, 기와, 요사, 합판, 원재료, 착색, 접착제, 공법, 함유, 페인트, 구리, 항아리, 루시, 메탈, 사출, 반투명, 모래, 장식품, 밥그릇, 금속재, 소나무, 코팅, 유리창, 담장, 주전자, 가람, 화병, 합성수지, 패브릭, 청동, 화강석, 트튼, 파이프, 판넬, 가마, 견고성, 글라스, 원목, 부식, 껍질, 한지, 점토, 전선, 주재료, 용해, 황토, 액체, 탄성, 매듭, 판재, 게이트, 적층, 토기, 재로, 도금, 다이아몬드, 강철, 도기, 황동, 석회, 찻잔, 실리콘, 진주, 약품, 장미, 장판, 장식물, 납석, 대방, 중합, 기름, 깃털, 진공, 몰딩, 하회탈, 찰흙, 전사지, 철제, 투명도, 몰드, 스테인, 구슬, 죽재, 라미네이트, 세공, 골판지, 무늬목, 칠기, 스텐, 석공예, 결착, 시유, 하우징 |
| 군집2 (121) | 교육, 대학, 학습, 대학교, 학교, 학생, 전공, 학과, 대학원, 학위, 교수, 석사, 여자, 수업, 과목, 교과, 학년, 실기, 학부, 홍익, 박사, 남자, 교사, 강의, 실습, 초등, 입시, 개설, 교과목, 교과서, 학기, 졸업, 연세대, 시험, 강사, 학원, 교실, 고등학교, 단원, 실업, 입학, 학점, 수강, 취업, 한국과학기술원, 서울대, 한양, 러닝, 학년도, 교육자, 전북, 교재, 인학, 정규, 재학, 선발, 학업, 연수, 강좌, 건국대, 교원, 숙명, 이수, 성균관, 청구, 경희대, 공과, 청주, 교수자, 모집, 커리, 학사, 연세, 교수진, 상명, 학급, 이화, 평생, 전임, 계명, 커리큘럼, 고등교육, 이화여대, 성취도, 전문대, 국민대, 수능, 자격증, 논총, 신입, 고등, 단국, 여대, 부교수, 임용, 고학년, 도학, 진학, 재직, 단과, 조교수, 교수법, 취도, 중학교, 성신, 렌디, 소집단, 익대, 한양대, 전문학교, 내신, 소묘, 시립, 동국, 동명대학, 공대, 조경학, 학과목, 이재국, 박영원, 사립 |
| 군집3 (97) | 색채, 색상, 컬러, 계열, 배색, 칼라, 명도, 검정, 색조, 채도, 녹색, 모드, 청색, 무채색, 흰색, 팔레트, 회색, 외장, 빨강, 조색, 초록, 선명, 원색, 백색, 파랑, 노랑, 계통, 측색, 적색, 경색, 보라색, 색명, 황색, 화이트, 색체, 난색, 블랙, 고명, 파스텔, 중명, 색감, 저채, 보색, 노란색, 빨간색, 파란색, 그레이, 단색, 고채, 색도, 갈색, 유채색, 색이름, 뉘양스, 보라, 오방색, 색은, 한색, 배합, 초록색, 앤드, 색상환, 주황, 삼원색, 오렌지, 홍색, 주황색, 색표, 흑색, 표색계, 그라데이션, 자색, 다크, 자연색, 베이지, 통색, 푸른색, 색지, 뉴트럴, 바탕색, 간색, 색계, 순색, 퍼플, 검은색, 강색, 분홍색, 중간색, 표색, 중채, 남색, 유행색, 금색, 붉은색, 하늘색, 브라이트, 기미 |

| | |
|---|---|
| 군집4 (79) | 形態, 硏究, 生産, 機能, 製品, 設計, 造形, 價値, 過程, 環境, 論文, 文化, 韓國, 方法, 生活, 時代, 空間, 傳統, 社會, 朝鮮, 構造, 槪念, 尺度, 色彩, 人間, 意味, 中心, 對象, 問題, 中國, 結果, 關係, 心理, 建築, 體系, 使用, 屬性, 構成, 評價, 技術, 染色, 視覺, 要素, 開發, 現代, 藝術, 要求, 可能, 科學, 編造, 變化, 特性, 狀況, 表現, 創造, 製作, 比例, 必要, 利用, 多樣, 裝飾, 學會, 判斷, 計劃, 測定, 素材, 道具, 存在, 材料, 經濟, 民俗, 側面, 影響, 歷史, 背景, 全體, 處理, 高麗, 問項 |
| 군집5 (78) | 공간, 배치, 설치, 면적, 동선, 거실, 수납, 부엌, 계단, 구역, 테이블, 화장실, 스페이스, 통로, 주방, 오픈, 공중, 욕실, 공용, 입구, 휠체어, 정원, 통과, 복도, 식당, 체어, 출입구, 좌석, 침실, 코너, 로비, 안방, 구획, 수납장, 샤워, 엘리베이터, 출입, 비상, 현관, 부스, 식탁, 평형대, 실외, 세면대, 발코니, 변기, 지주, 간호, 사적, 수납공간, 화재, 호수, 베란다, 매입, 중정, 승강기, 평면도, 출입문, 다용도실, 목욕, 욕조, 횡단, 개폐, 경사로, 에스컬레이터, 작업대, 현수막, 출구, 주출, 리프트, 접이식, 병실, 테라스, 부대, 싱크대, 접수대, 동장, 샤워기 |
| 군집6 (74) | 감성, 반응, 감각, 자극, 감정, 심리, 유도, 흥미, 정서, 유발, 청각, 동기, 즐거움, 촉각, 재미, 몰입, 유머, 유희, 생리, 진동, 호기심, 촉감, 각성, 호소, 미각, 감수, 오감, 공감각, 기분, 만족감, 감동, 기쁨, 쾌락, 의외, 후각, 충동, 웃음, 흥분, 애착, 지성, 입감, 쾌감, 유쾌, 구복, 촉발, 유인, 생리학, 유회, 이입, 볼거리, 이완, 차이역, 희극, 수용기, 유머러스, 슬픔, 절정, 촉구, 예민, 불쾌, 분노, 감정이입, 선천, 감수성, 흥미, 놀라움, 손끝, 유대감, 지루함, 위트, 감흥, 입도, 식욕, 비렌 |
| 군집7 (62) | 브랜드, 기업, 소비자, 전략, 경영, 마케팅, 고객, 구매, 아이덴티티, 성과, 비즈니스, 중소기업, 자산, 자사, 인지도, 컨설팅, 수익, 무형, 매니지먼트, 포지셔닝, 랜딩, 경영자, 스토어, 전술, 이윤, 리뉴, 충성, 브랜, 트랜스, 다국적, 타깃, 프로모션, 마켓, 리스크, 믹스, 정경원, 매니아, 저탄소, 필립스, 사내, 포지션, 프랜차이즈, 플레이스, 재무, 네임, 쿼티, 티티, 인벤, 공략, 로열티, 틈새시장, 비자, 충동구매, 덴티, 신선도, 샴푸, 브랜드디자인, 마아, 경영진, 브렌드, 아커, 에이전시 |
| 군집8 (57) | 패션, 의상, 의류, 착용, 복식, 의복, 신발, 컬렉션, 디테일, 드레스, 마스크, 실루엣, 데님, 모자, 가방, 체형, 소매, 오뜨, 주름, 갑옷, 코디, 티셔츠, 상의, 코트, 스카프, 바지, 셔츠, 재킷, 시즌, 액세서리, 개더, 스커트, 목걸이, 레이스, 하의, 비즈, 매치, 줄무늬, 주머니, 착용감, 우산, 필머, 옷감, 원피스, 치마, 자켓, 웨이스트, 네이트, 발끝, 슈트, 착장, 여성복, 장갑, 비니, 허리둘레, 헬멧, 발가락 |
| 군집9 (54) | 문자, 글자, 한글, 서체, 가독성, 글꼴, 활자, 폰트, 타이포그라피, 자간, 한자, 활자체, 글씨, 부호, 창제, 행간, 네모, 알파벳, 굵기, 줄기, 훈민정음, 고딕, 받침, 글줄, 네모꼴, 발음, 글자꼴, 판독성, 중성, 모음, 타자기, 인쇄술, 윗글, 낱글자, 대문자, 성체, 네모틀, 자판, 명조체, 세벌식, 조판, 文字, 글쓰기, 기준선, 세리프, 탈네모틀, 소문자, 가독, 손글씨, 활자꼴, 상형, 고딕체, 음절, 닿자 |
| 군집10 (51) | 장애, 노인, 편리, 편의, 안전, 복지, 배려, 고령자, 편안, 병원, 쾌적, 접근성, 의료, 범죄, 환자, 장애인, 보육, 여유, 예방, 약자, 시력, 위생, 이동성, 안락, 쾌적성, 청결, 감염, 안전성, 회관, 보건, 세심, 장애자, 임산부, 편리성, 진료, 색각, 재활, 장애아, 출산, 공평, 요양, 실버산업, 노약자, 임신, 보호자, 안락감, 보건복지부, 애인, 안심, 소아, 비장 |

오분석된 어휘들과 연관성이 적어보이는 어휘가 일부 섞여 있으나 군집의 성격은 확연히 드러나는 편이다. 이번 분석에서는 어휘별 유사 어휘가 아닌 분포별 의미 단위를 살피는 것이 목적이므로 거리값을 성기게 주었으나 목적에 따라 속성값을 조절하여 군집을 더 정밀하게 추출할 수도 있다.

군집 분석 결과, 〈디자인학연구〉에 사용된 어휘 목록에서는 도시 및 지역 관련 어휘가 가장 큰 비중을 차지하고 있으며 재료, 교육, 색상, 공간, 감성, 브랜드, 패션, 타이포그래피, 노약자 및 장애 관련 어휘 역시 비중이 높았다. 이 결과는 빈도와는 관계가 없으므로 해당 군집의 어휘가 많이 출현했다거나 해당 주제의 논문이 많음을 의미하지 않는다.

표 6에서는 흥미로운 부분이 있는데 타이포그래피와 관련된 군집9에 '타이포그래피'가 아닌 '타이포그라피'가 포함되어 있다는 점이다. 이는 외래어 표기에 맞으며 절대 빈도가 높은 쪽은 '타이포그래피'이지만, 타이포그래피 관련 어휘와 연관도가 높은 표기는 '타이포그라피'임을 뜻한다. 이 분석에서 '타이포그래피'는 '타이포그래피, 타이포, 타이, 키네틱, 그래피, 캘리그래피, 구체시, 그라, 무빙'과 하나의 군집을 이루고 있다.

## 5. 결론과 제언

연구자는 이 연구에서 텍스트마이닝 기법을 이용하여 한국 디자인 분야의 어휘 사용 양상을 데이터 기반으로 분석할 수 있을지 그것이 디자인 용어 연구에 유의미한 방식일지 살피고자 했다. 이를 위해 해당 기법을 활용한 주요 어휘의 추출과 그 출현 양상의 파악, 주요 어휘에 대한 유사어 파악 등 두 가지 연구 질문을 설정하였고, 〈디자인학연구〉에 게재된 한글 논문으로 말뭉치를 구축한 뒤 자연어처리에서 널리 사용되는 파이썬 라이브러리를 활용하여 빈도 분석과 분포(유사도) 분석을 수행했다.

빈도 분석에서는, 불용어 추출에는 단순 빈도를 활용하고 연도별 주요 어휘 추출에는 단순 빈도와 TF-IDF를 활

용했고, 주요 어휘의 연도별 출현 빈도를 그래프로 표현할 때는 말뭉치의 출현 빈도를 문서 길이로 정규화한 TF 값을 사용했다. 불용어를 제거하였기에 단순 빈도였음에도 연도별 고빈도 어휘는 사뭇 다르게 나타났으며 특정 연도에 비중 있게 출현했지만 절대적인 출현 빈도가 낮아서 자칫 버려질 수 있었던 어휘들은 TF-IDF를 통해 파악할 수 있었다.

고빈도 어휘의 연도별 빈도 그래프는 단어별 출현 추이를 유형화하고 해당 그룹의 어휘들을 모아서 살피는 데 유용했다. 최근 빈도가 급증한 어휘들은 한국디자인학회 연구자들의 최근 관심사를 반영하는 것으로, 추후 말뭉치의 규모가 커진다면 디자인 분야의 최신 이슈를 파악하는 데에도 활용할 수 있을 것으로 기대한다. 다만 현재의 분석 수준에서 유형별 혹은 유형 간 어휘 목록 사이에 명확한 상관관계나 차별점까지 발견하기는 어려웠다.

빈도 분석과 같은 통계적 기법을 사용할 때는 어휘의 표기 통일이 잘 되어 있지 않은 경우, 하나의 어휘가 서로 다른 표기로 적혀 빈도가 흩어질 가능성을 배제할 수 없다. 대표 표기는 형태소 분석 사전에 등록되어 있지만, 다른 표기들이 등록되어 있지 않아 오분석되고 결과적으로 빈도 측정이 되지 않았을 가능성 역시 존재한다. 추론 기반의 분포 분석은 이런 문제를 파악하고 보완하는 데 도움을 준다. Word2Vec의 skip-gram 학습의 경우 앞뒤 맥락을 분석하여 중심에 나타날 타깃 단어를 예측하는 방식으로, 동일한 맥락에서 나타날 수 있는 유사어 목록을 제공한다. 이를 통해 특정 단어가 어떤 맥락에서 사용되었는지 역으로 유추할 수 있으며 문서에서 그 단어의 중요도까지도 판단할 수 있다. t-SNE 그래프와 밀도 기반의 군집 분석을 통해 도출한 어휘 군집들은 〈디자인학연구〉에 어떤 맥락들이 존재하고 이를 표현하기 위해 어떤 어휘들이 사용되었는지 분석하는 데 유용했다.

이 연구는 소규모이지만 한국 디자인 분야의 텍스트를 말뭉치로 구축하고 텍스트마이닝 기법으로 디자인 어휘의 사용 양상을 분석하고자 한 첫 사례이다. 다만 이 연구는 〈디자인학연구〉라는 단일 학술지를 대상으로 넓은 범위에서 탐색적으로 분석을 수행했기에 어휘 분석 결과를 직접 활용하기에는 한계를 가지고 있다. 이 연구를 바탕으로 후속 연구로서 문서 군집을 병행하여 디자인학 연구 동향을 분석하거나 말뭉치 규모를 확대하여 디자인 분야의 학회지들 간에 어휘 사용 양상을 비교하는 연구가 가능할 것으로 판단한다.

**References**


1. Goki, S. (2019). *밑바닥부터 시작하는 딥러닝2 [Deep Learning from scratch 2]*. Seoul:Hanbit Media
2. Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2015). *자연어 텍스트 처리를 통한 검색 시스템 구축 [Taming Text: How to Find, Organize, and Manipulate it]*. Seoul: Acorn Publishing Co.
3. Jeong, Y. (2012). *정보검색연구 [Research in information retrieval]*. Seoul:Yonsei University Press
4. Kim, H. J. (2017). LOVIT x DATA SCIENCE. Retrieved October, 2019. from https://lovit.github.io
5. Kim, H. S. (2015). 전문용어의 일반어화에 대한 소고 [A Study on the Generalization of Technical Terms]. *Hanminjok Emunhakhoe, 71*, 129-154
6. Müller, A. C., & Guido, S. (2019). *파이썬 라이브러리를 활용한 머신러닝 [Introduction to Machine Learning with Python]*. Seoul:Hanbit Media

# 말뭉치와 텍스트마이닝 기법을 활용한 〈디자인학연구〉의 사용 어휘 분석

김은영[1], 안병학[2*]

[1]숙명여자대학교 시각영상디자인과, 강사, 서울, 대한민국

[2]홍익대학교 시각디자인과, 교수, 서울, 대한민국

**초록**

**연구배경** 한국에서 디자인 용어에 대한 연구는 디자인의 실천적 분야 연구에 비해 현저히 적은 실정이다. 디자인 용어에 대한 연구 특히 어휘의 사용 양상을 살피는 연구는 디자인학의 토대에 해당한다. 이 연구에서는 텍스트마이닝 기법을 이용하여 한국 디자인 분야의 어휘 사용 양상을 데이터 기반으로 분석할 수 있을지 그것이 디자인 용어 연구에 유의미한 방식일지 살피고자 했다.

**연구방법** 한국디자인학회 학술지인 〈디자인학연구〉에 게재된 한글 논문 2,214편으로 말뭉치를 구축한 뒤 텍스트마이닝 기법으로 어휘 양상을 파악하고자 했다. 분석 과정에는 자연어처리 분야에서 널리 활용되는 파이썬 라이브러리와 빈도 분석, 분포(유사도) 분석 기법을 사용했다.

**연구결과** 연구자는 텍스트마이닝을 통해 디자인 말뭉치로부터 유의미한 어휘 목록과 어휘 간 관계를 찾아낼 수 있을지 탐색하고자 하였으며, 일련의 결과를 통해 디자인 용어 연구에서 텍스트마이닝 기법으로 충분히 흥미로운 발견들을 이끌어 낼 수 있음을 확인했다.

**결론** 이 연구는 비록 소규모이지만 디자인 분야의 텍스트를 말뭉치로 구축하고 텍스트마이닝 기법으로 어휘의 사용 양상을 분석하고자 한 첫 사례이다. 이 연구가 이후 다양한 관점을 가진 디자인 용어 연구의 활성화에 일조하길 기대한다.

**주제어** 디자인 용어, 텍스트마이닝, 자연어처리, 디자인 말뭉치, 디자인 연구

*교신저자 : 안병학 (ahn.hisd@gmail.com)