NRF R&D BRIEF

노벨과학상 수상자와 한국의 리더연구자 연구제목 키워드 연관 분석

2021-11호 (2021.7.5.) / 정책혁신팀 이성민, 양정모

1. 분석대상

□ (분석자료)노벨과학상 역대 수상자의(1901년~2020년) 연구제목을 대상으로 물리학, 화학, 생리의학 한글 코퍼스(말뭉치)를 구축함

<표 1> 부문별 수상자 수

구 분	물리학상	화학상	생리의학상	합 계
노벨과학상	216	186	222	624

□ (분석방법) 코퍼스 언어학(Corpus Linguistics)의 키워드 분석함

○ **키워드**란 서로 다른 두 개의 코퍼스에서 만든 **단어 목록**을 사용하여 컴퓨터 알고리즘으로 **텍스트의 특징**이 될 만한 단어를 선별한 것을 말함[1)]

○ **기계학습**기법인 **교차분석(Cross Validation)**으로 코퍼스 간 비교

○ Rayson and Garside(2000) **로그-라이클리후드** 알고리즘에 의한 키워드로 **유의수준 0.01 이상의 임계값(critical value)**을 키워드 산출 기준으로 설정함

(1) $E_i = N_i \frac{\sum_i O_i}{\sum_i N_i}$ (2) $LL = 2*((a*\log(\frac{a}{E_1})) + (b*\log(\frac{b}{E_2})))$

(1) E_i는 기대빈도를 나타내고 코퍼스의 크기는 N_i(N1=c, N2=d로 표시) 그리고 비교하는 단어 'a'와 'b'의 빈도 합계(a+b)는 관찰빈도는 O_i로 표시됨

(2) 기대빈도 계산은 아래와 같음
- a = 단어1의 빈도, b = 단어2의 빈도, c = 코퍼스1의 모든 단어, d = 코퍼스2의 모든 단어
- 첫 번째 기대빈도: E1 = c*(a+b) / (c+d)
- 두 번째 기대빈도: E2 = d*(a+b) / (c+d)

□ (분석도구)

○ **코퍼스언어학** 분석을 위하여 WordSmith Tools 7.0을 사용하였으며, 부록의 **클라우드** 빈도 분석 시각화를 위해 NetMiner 4.0을 활용함

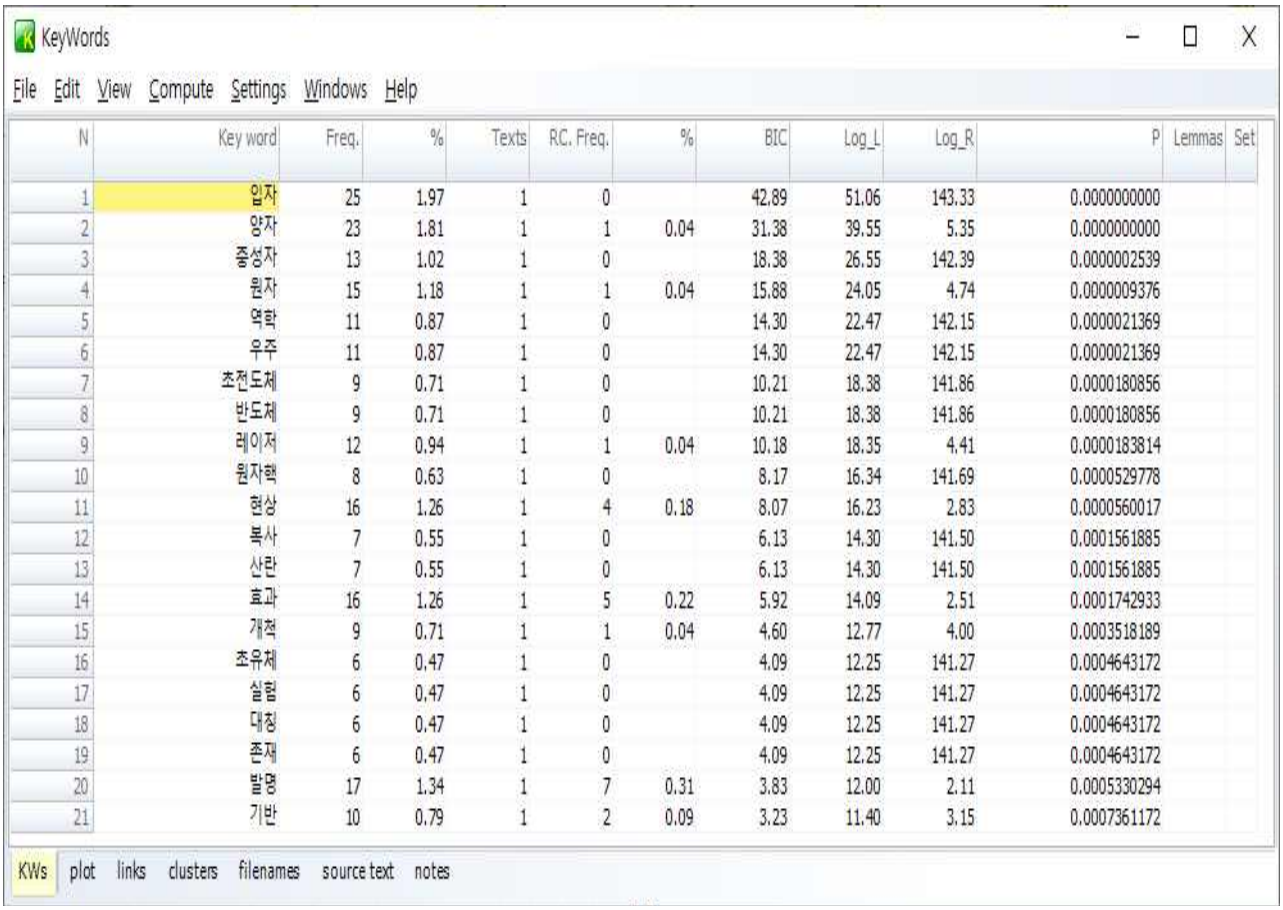
1) Mike Scott(2020)에 따르면 키워드란 연구코퍼스에 사용된 단어의 빈도가 참조코퍼스 사용단어의 빈도보다 통계적 유의성이 있을 정도로 더 많은 빈도로 사용되면 긍정키워드(positive keyword)로, 더 적게 사용할 경우 부정키워드(negative keyword)로 나타난다고 함

2. 분석결과

□ 물리학 분야 키워드

○ 물리학의 연구제목을 대상코퍼스로 두고, 화학과 생리의학을 참조코퍼스로 하여 산출한 물리학 분야 키워드 추출함[2)]

<그림 1> 물리학 분야 키워드 목록


KeyWords

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	%	BIC	Log_L	Log_R	P	Lemmas	Set
1	입자	25	1.97	1	0		42.89	51.06	143.33	0.0000000000		
2	양자	23	1.81	1	1	0.04	31.38	39.55	5.35	0.0000000000		
3	중성자	13	1.02	1	0		18.38	26.55	142.39	0.0000002539		
4	원자	15	1.18	1	1	0.04	15.88	24.05	4.74	0.0000009376		
5	역학	11	0.87	1	0		14.30	22.47	142.15	0.0000021369		
6	우주	11	0.87	1	0		14.30	22.47	142.15	0.0000021369		
7	초전도체	9	0.71	1	0		10.21	18.38	141.86	0.0000180856		
8	반도체	9	0.71	1	0		10.21	18.38	141.86	0.0000180856		
9	레이저	12	0.94	1	1	0.04	10.18	18.35	4.41	0.0000183814		
10	원자핵	8	0.63	1	0		8.17	16.34	141.69	0.0000529778		
11	현상	16	1.26	1	4	0.18	8.07	16.23	2.83	0.0000560017		
12	복사	7	0.55	1	0		6.13	14.30	141.50	0.0001561885		
13	산란	7	0.55	1	0		6.13	14.30	141.50	0.0001561885		
14	효과	16	1.26	1	5	0.22	5.92	14.09	2.51	0.0001742933		
15	개척	9	0.71	1	1	0.04	4.60	12.77	4.00	0.0003518189		
16	초유체	6	0.47	1	0		4.09	12.25	141.27	0.0004643172		
17	실험	6	0.47	1	0		4.09	12.25	141.27	0.0004643172		
18	대칭	6	0.47	1	0		4.09	12.25	141.27	0.0004643172		
19	존재	6	0.47	1	0		4.09	12.25	141.27	0.0004643172		
20	발명	17	1.34	1	7	0.31	3.83	12.00	2.11	0.0005330294		
21	기반	10	0.79	1	2	0.09	3.23	11.40	3.15	0.0007361172		

KWs plot links clusters filenames source text notes

- 분석결과 총 21개의 키워드가 타 분야(화학, 생리의학)와 비교하여 물리학 분야에서 통계적 유의미성을 가지며 높은 임계값을 보임

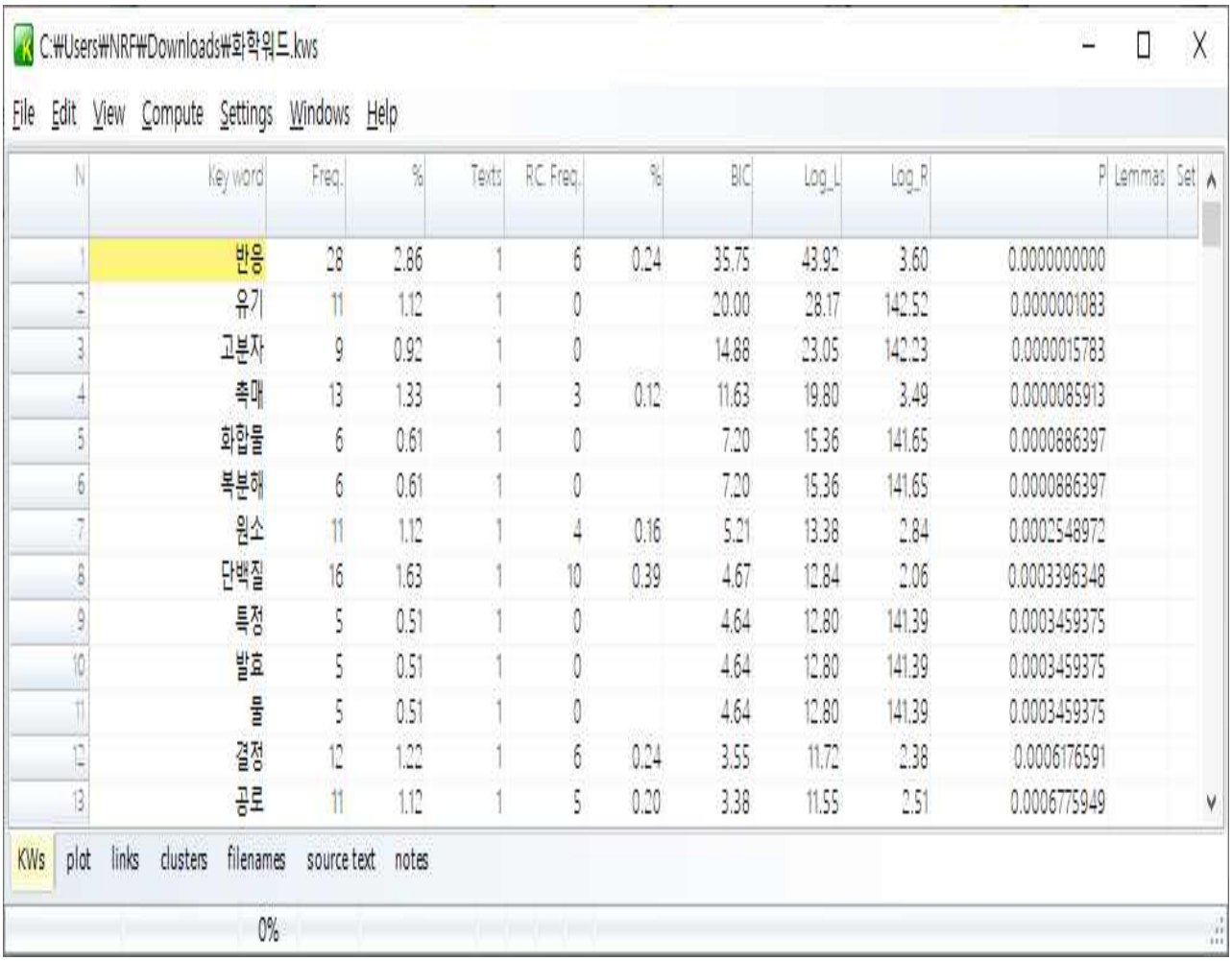
- 임계값이 높은 10개 연구 관련 내용어는 입자(1위), 양자(2위), 중성자(3위), 원자(4위), 역학(5위), 우주(6위), 초전도체(7위), 반도체(8위), 레이저(9위), 원자핵(10위) 등이었음

2) 그림에서 'Key word'는 물리학의 키워드이고 'Freq'는 키워드 사용빈도, '%'는 사용 비율을 의미함. 'RC.Freq.'는 화학과 생리의학으로 구성한 코퍼스에서 해당 키워드의 사용 빈도를 의미함.

□ 화학 분야 키워드

○ 화학의 연구제목을 대상코퍼스로 두고, 물리학과 생리의학을 참조코퍼스로 하여 산출한 화학 분야 키워드 추출함

<그림 2> 화학 분야 키워드 목록



N	Key word	Freq.	%	Texts	RC Freq.	%	BIC	Log_L	Log_R	P	Lemmas	Set
1	반응	28	2.86	1	6	0.24	35.75	43.92	3.60	0.0000000000		
2	유기	11	1.12	1	0		20.00	28.17	142.52	0.0000001083		
3	고분자	9	0.92	1	0		14.88	23.05	142.23	0.0000015783		
4	촉매	13	1.33	1	3	0.12	11.63	19.80	3.49	0.0000085913		
5	화합물	6	0.61	1	0		7.20	15.36	141.65	0.0000886397		
6	복분해	6	0.61	1	0		7.20	15.36	141.65	0.0000886397		
7	원소	11	1.12	1	4	0.16	5.21	13.38	2.84	0.0002548972		
8	단백질	16	1.63	1	10	0.39	4.67	12.84	2.06	0.0003396348		
9	특정	5	0.51	1	0		4.64	12.80	141.39	0.0003459375		
10	발효	5	0.51	1	0		4.64	12.80	141.39	0.0003459375		
11	물	5	0.51	1	0		4.64	12.80	141.39	0.0003459375		
12	결정	12	1.22	1	6	0.24	3.55	11.72	2.38	0.0006176591		
13	공로	11	1.12	1	5	0.20	3.38	11.55	2.51	0.0006775949		

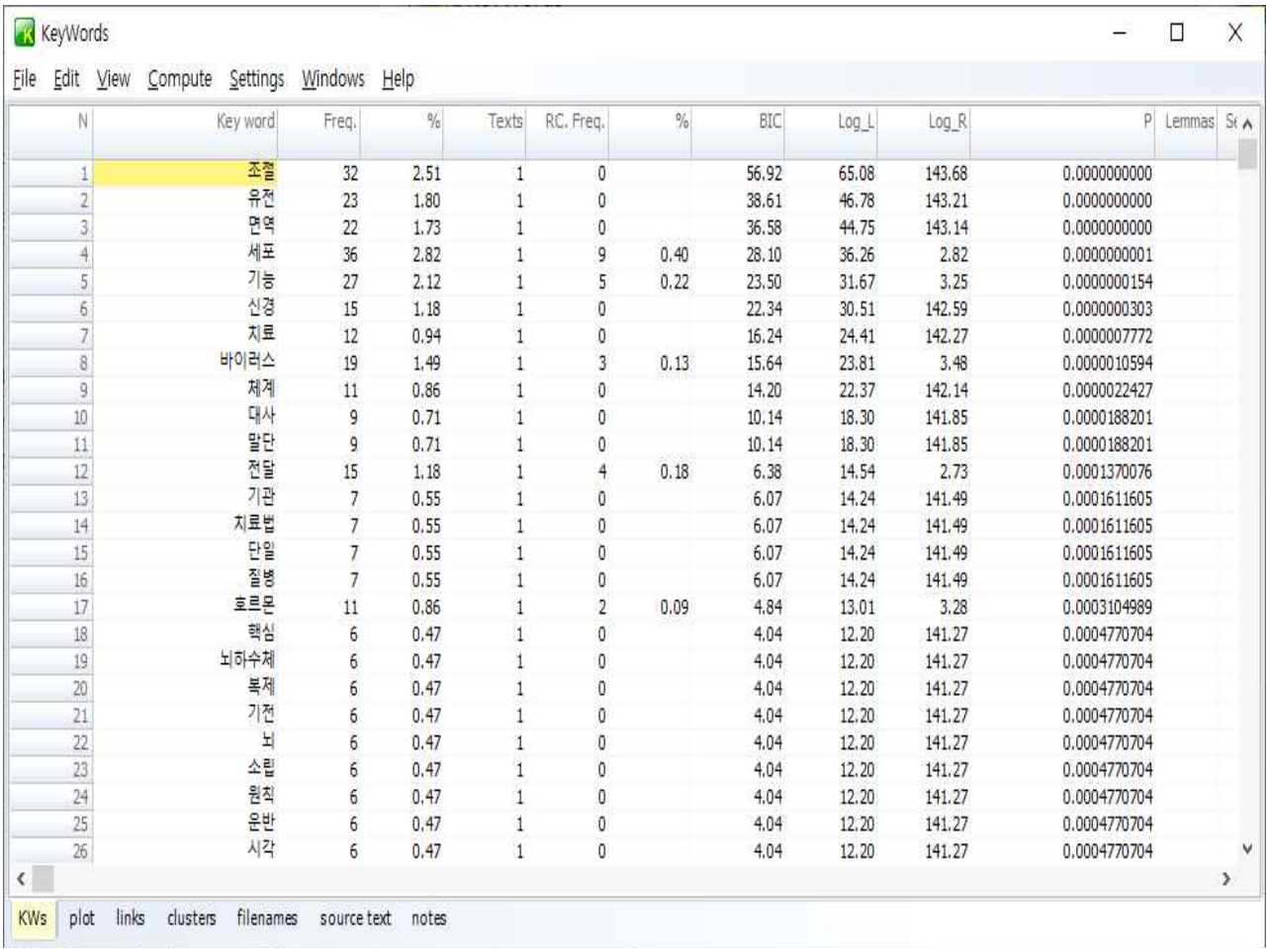
- 분석결과 총 13개의 키워드가 타 분야(물리학, 생리의학)와 비교하여 화학 분야에서 통계적 유의미성을 가지며 높은 임계값을 보임

- 임계값이 높은 10개 연구 관련 내용어는 반응(1위), 유기(2위), 고분자(3위), 촉매(4위), 화합물(5위), 복분해(6위), 원소(7위), 단백질(8위), 특정(9위), 발효(10위) 순이었음

□ 생리의학 분야 키워드

○ 생리의학의 연구제목을 대상코퍼스로 두고, 물리학과 화학을 참조코퍼스로 하여 산출한 생리의학 분야 키워드 추출함

<그림 3> 생리의학 분야 키워드 목록


KeyWords

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	%	BIC	Log_L	Log_R	P	Lemmas	Sc
1	조절	32	2.51	1	0		56.92	65.08	143.68	0.0000000000		
2	유전	23	1.80	1	0		38.61	46.78	143.21	0.0000000000		
3	면역	22	1.73	1	0		36.58	44.75	143.14	0.0000000000		
4	세포	36	2.82	1	9	0.40	28.10	36.26	2.82	0.0000000001		
5	기능	27	2.12	1	5	0.22	23.50	31.67	3.25	0.0000000154		
6	신경	15	1.18	1	0		22.34	30.51	142.59	0.0000000303		
7	치료	12	0.94	1	0		16.24	24.41	142.27	0.0000007772		
8	바이러스	19	1.49	1	3	0.13	15.64	23.81	3.48	0.0000010594		
9	체계	11	0.86	1	0		14.20	22.37	142.14	0.0000022427		
10	대사	9	0.71	1	0		10.14	18.30	141.85	0.0000188201		
11	말단	9	0.71	1	0		10.14	18.30	141.85	0.0000188201		
12	전달	15	1.18	1	4	0.18	6.38	14.54	2.73	0.0001370076		
13	기관	7	0.55	1	0		6.07	14.24	141.49	0.0001611605		
14	치료법	7	0.55	1	0		6.07	14.24	141.49	0.0001611605		
15	단일	7	0.55	1	0		6.07	14.24	141.49	0.0001611605		
16	질병	7	0.55	1	0		6.07	14.24	141.49	0.0001611605		
17	호르몬	11	0.86	1	2	0.09	4.84	13.01	3.28	0.0003104989		
18	핵심	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
19	뇌하수체	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
20	복제	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
21	기전	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
22	뇌	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
23	소립	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
24	원칙	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
25	운반	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		
26	시각	6	0.47	1	0		4.04	12.20	141.27	0.0004770704		

KWs plot links clusters filenames source text notes

- 분석결과 총 26개의 키워드가 타 분야(물리학, 화학)와 비교하여 생리의학 분야에서 통계적 유의미성을 가지며 높은 임계값을 보임

- 임계값이 높은 10개 연구 관련 내용어는 조절(1위), 유전(2위), 면역(3위), 세포(4위), 기능(5위), 신경(6위), 치료(7위), 바이러스(8위), 체계(9위), 대사(10위) 순이었음

3. 결론 및 시사점

○ 코퍼스언어학의 유용한 도구 중 하나인 키워드분석을 통해서 노벨과학상 분야별(물리학, 화학, 생리의학)로 사용된 특징적인 단어를 빠르게 파악할 수 있었음

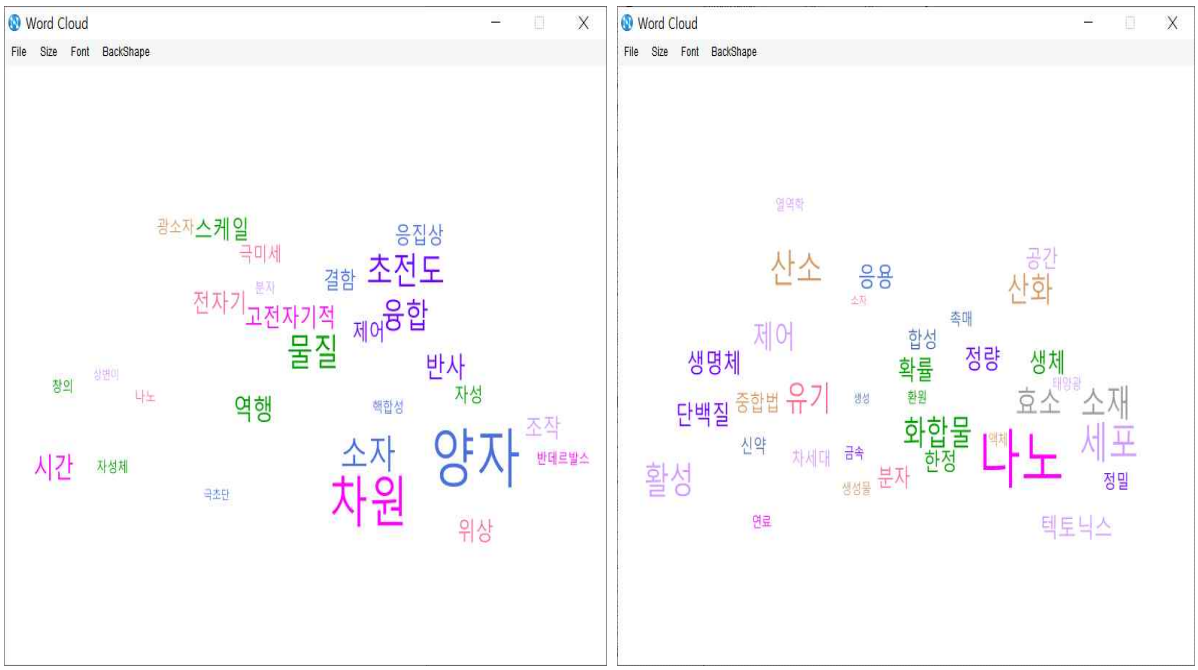
– 키워드 수는 물리학 21개, 화학 13개, 생리의학 26개 순으로 생리의학에서 가장 많은 키워드가 나타남. 생리의학 코퍼스에는 타 분야에서 0회 혹은 매우 적은 빈도로 사용된 단어를 많이 포함하고 있음을 알 수 있었음

○ 노벨과학상의 물리학과 화학 분야 키워드를 한국연구재단의 리더연구사업*(2012~2021, 10년)의 물리 및 화학 분야 선정 과제의 키워드와 비교함

* 미래의 독자적 과학기술과 신기술 개발을 위해 세계적 수준에 도달한 연구자의 심화연구 집중 지원을 통해 글로벌 연구리더로 육성하는 사업

※ 생리의학의 경우 다양한 분야를 포괄하고 있어 비교대상에서 제외

<그림 4> 리더연구 선정 과제 중 물리학(왼쪽), 화학(오른쪽)의 키워드



– 리더연구 물리학의 빈출 키워드는 **양자(1위)**, 차원(2위), 소자(3위), 물질(4위), **초전도**(5위), 융합(6위), 반사(7위), 시간(8위), 역행(9위), 고전자기적(10위) 등으로 노벨물리학상 키워드와 2개가 일치

– 리더연구 화학의 빈출 키워드는 나노(1위), 세포(2위), 산소(3위), 소재(4위), 활성(5위), 산화(6위), **유기(7위)**, 화합물(8위), 효소(9위), **단백질(10위)** 등으로 노벨화학상 키워드와 2개가 일치

- 노벨과학상과 비교하여, 재단 리더연구사업은 사업기간과 선정 과제 수가 상대적으로 적어, 키워드 비교에서 큰 유사점과 차이점은 없었으나, 양자, 초전도, 유기, 단백질 등 유사 키워드가 상위에 노출되었음

- 향후, 노벨과학상의 수상자 연구주제를 살펴보고, 재단의 지원사업을 비교해 나간다면, 한 걸음 더 노벨상 수상에 가까워지는 증거를 제공할 수 있을 것으로 생각됨

참 고 자 료

1. **노벨상 공식 홈페이지** www.nobelprize.org

2. 2018, 2019 **노벨과학상 종합분석 보고서 - 수상 현황과 트렌드를 중심으로 -**

3. Rayson & Garside, 2000, Comparing Corpora Using Frequency Profiling. in the Proceedings of the workshop on Comparing corpora. pp.1-6.

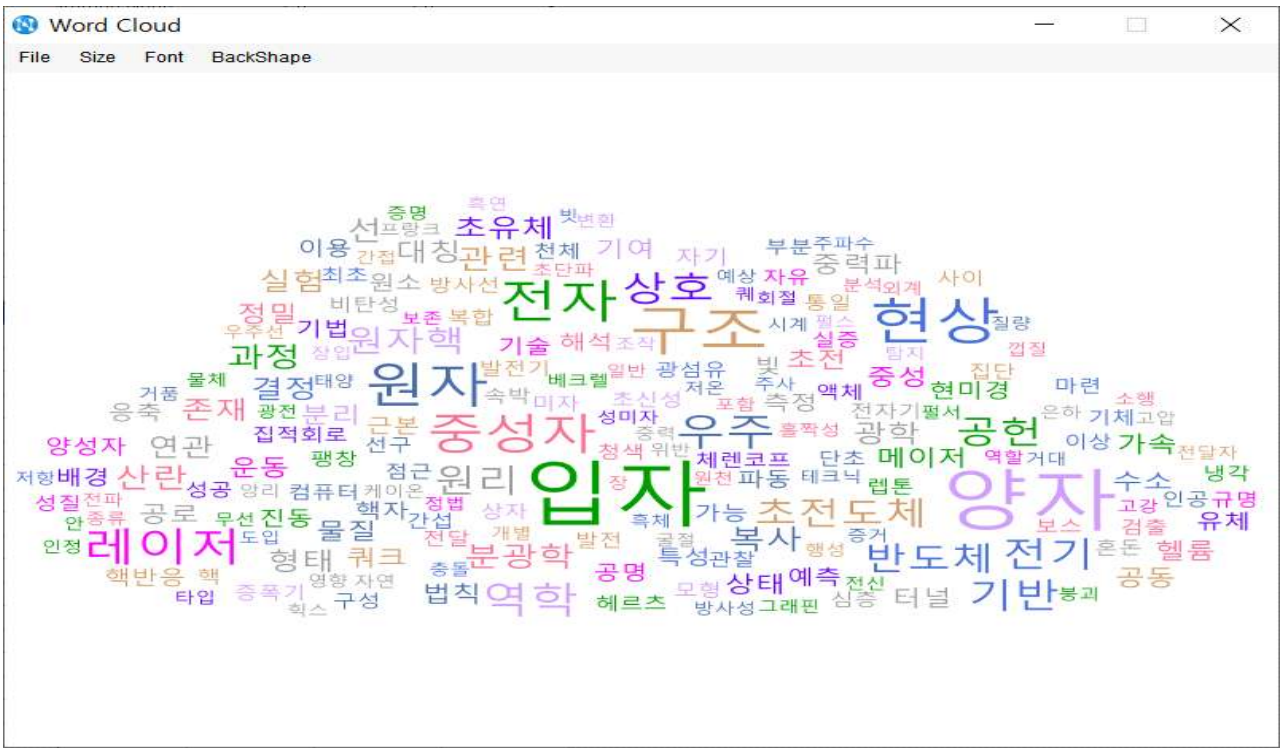
4. Scott, 2020, WordSmith Tools version 8, Stroud: Lexical Analysis Software.

부록 **노벨과학상 연구제목 클라우드 분석(단순 빈도)**

□ 물리학 분야 216명의 키워드

○ 빈출 키워드는 입자(25회), 양자(23회), 구조(17회), 원자(15회), 전자(14회), 중성자(13회), 레이저(12회), 역학(11회), 우주(11회) 전기(10회) 순이었음

<그림 1> 물리학 분야 키워드 클라우드



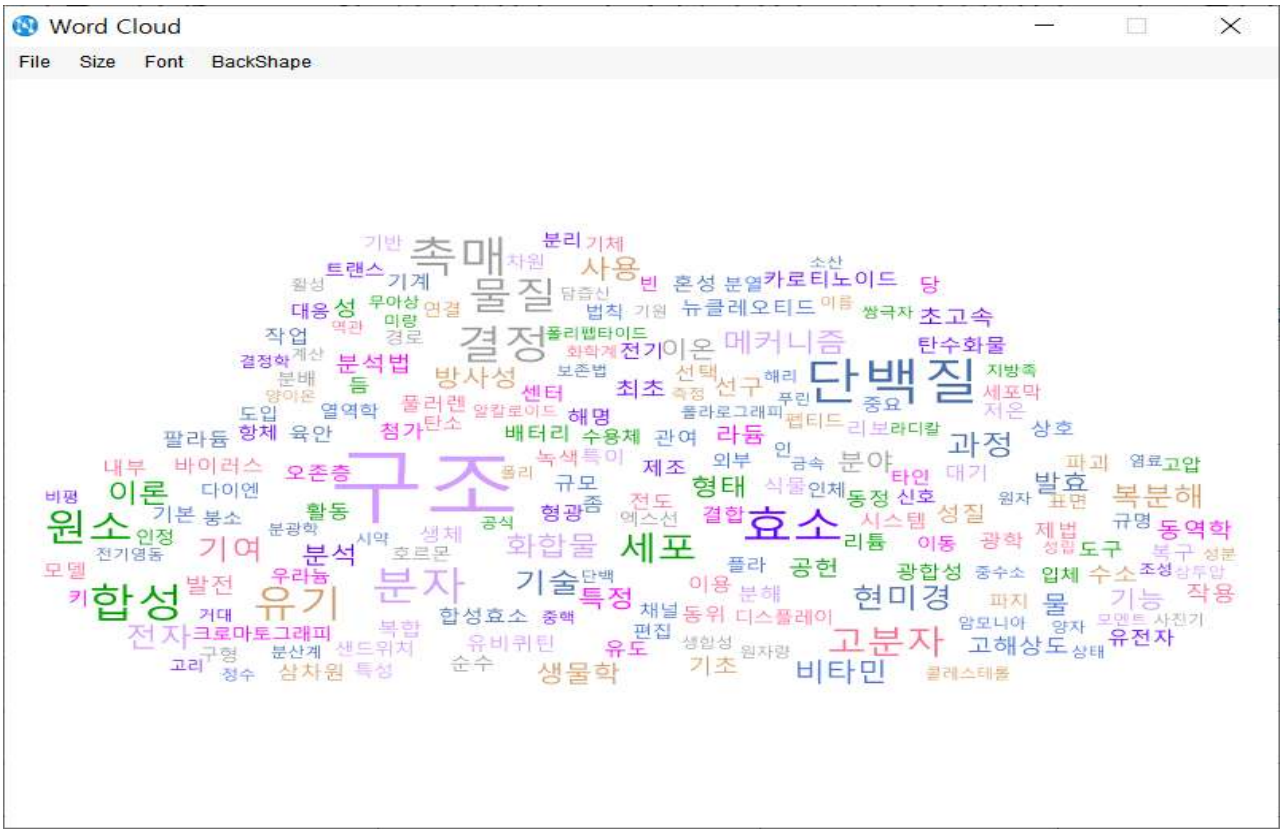
<표 1> 물리학 분야 빈출 키워드

명사	Part of Speech(POS)	Frequency	Word length
입자	Common Noun	25	2
양자	Common Noun	23	2
구조	Common Noun	17	2
원자	Common Noun	15	2
전자	Common Noun	14	2
중성자	Common Noun	13	3
레이저	Common Noun	12	3
역학	Common Noun	11	2
우주	Common Noun	11	2
전기	Common Noun	10	2

□ 화학상 분야 186명의 키워드

○ 빈출 키워드는 구조(29회), 단백질(16회), 촉매(13회), 효소(13회), 결정(12회), 분자(12회), 합성(12회), 물질(11회), 유기(11회), 원소(11회) 순이었음

<그림 2> 화학 분야 키워드 클라우드



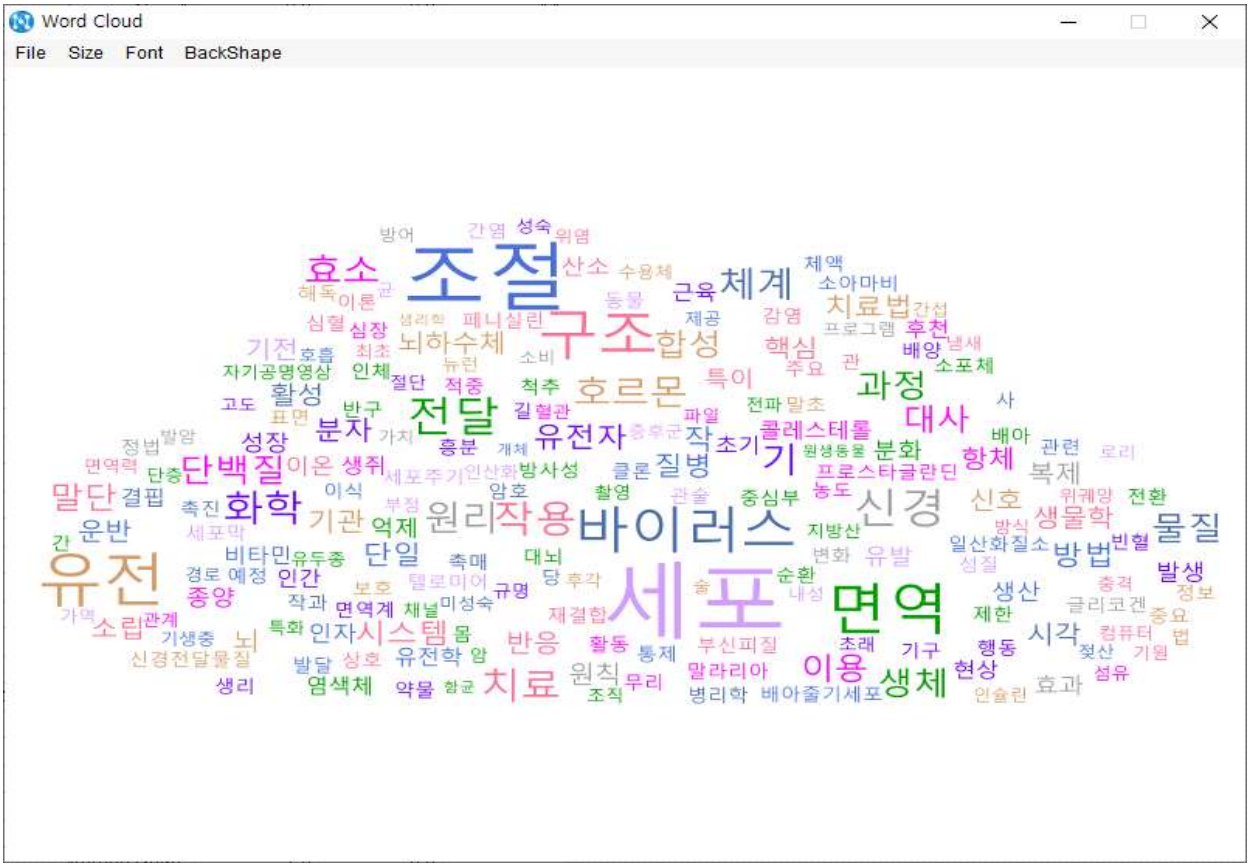
<표 2> 화학 분야 빈출 키워드

명사	Part of Speech(POS)	Frequency	Word length
구조	Common Noun	29	2
단백질	Common Noun	16	3
촉매	Common Noun	13	2
효소	Common Noun	13	2
결정	Common Noun	12	2
분자	Common Noun	12	2
합성	Common Noun	12	2
물질	Common Noun	11	2
유기	Common Noun	11	2
원소	Proper Noun	11	2

□ 생리의학 분야 222명의 키워드

○ 빈출 키워드는 세포(36회), 조절(32회), 유전(23회), 면역(22회), 구조(21회), 바이러스(19회), 신경(15회), 전달(15회), 작용(13회), 치료(12회) 순이었음

<그림 3> 생리의학 분야 키워드 클라우드



<표 3> 생리의학 분야 빈출 키워드

명사	Part of Speech(POS)	Frequency	Word length
세포	Common Noun	36	2
조절	Common Noun	32	2
유전	Common Noun	23	2
면역	Common Noun	22	2
구조	Common Noun	21	2
바이러스	Common Noun	19	4
신경	Common Noun	15	2
전달	Common Noun	15	2
작용	Common Noun	13	2
치료	Common Noun	12	2

□ 노벨과학상 전 분야 624명의 키워드

○ 지금까지 역대 수상자 전체의 연구제목에서 추출한 상위 10개 키워드는 세포(45회), 반응(34회), 조절(32회), 작용(31회), 단백질(26회), 물질(26회), 입자(25회), 양자(24회), 효소(24회), 유전(23회) 순이었음

<그림 4> 노벨과학상 전 분야 키워드 클라우드

<표 4> 노벨과학상 전 분야 빈출 키워드

명사	Part of Speech(POS)	Frequency	Word length
세포	Common Noun	45	2
반응	Common Noun	34	2
조절	Common Noun	32	2
작용	Common Noun	31	2
단백질	Common Noun	26	3
물질	Common Noun	26	2
입자	Common Noun	25	2
양자	Common Noun	24	2
효소	Common Noun	24	2
유전	Common Noun	23	2