한국어 의미학 32, 1-28 (2010. 08. 30.)

공기 명사에 기초한 의미/개념 연관성의 네트워크 구성

강 범 모
(고려대학교)

<Abstract>

Kang, Beom-mo, 2010. Constructing Networks of Related Concepts Based on Co-occurring Nouns. *Korean Semantics 32.* A method of constructing a network of related words(nouns) is proposed. We start with discussing the significance of relatedness of words as a kind of lexical relation. Related nouns are nouns "somehow (felt) related" and proposed to be found by acquiring nouns occurring in the same paragraph to a significant degree. The degree of co-occurrence is measured by t-score, which is a widely used method for calculating the degree of collocation. The networks of related words(nouns) are themselves the networks of things and concepts denoted by the words. Therefore, by this method we can understand features of culture in the society, a language community. The data is based on morpho-semantic analysis component of Sejong Korean Corpora. Network representation is provided by means of Pajek, a popular network analysis tool.

핵심어: 어휘관계(lexical relation), 관련어(related word), 연관성(relatedness), 공기(co-occurrence), 연어(collocation), t-점수(t-score), 코퍼스(corpus), 세종말뭉치(Sejong Korean Corpora), Pajek, 개념(concept), 문화(culture), 네트워크(network)

1. 어휘관계와 관련어

소쉬르(1990[1916])는 언어 체계의 구조, 즉 언어 요소들의 관계를 중요시하였다. 그 관계들은 크게 보아 결합관계(syntagmatic relation)와 치환관계(paradigmatic relation)로 나눌 수 있다. 예를 들어 '어떤 학생이 떠났다'에서 '어떤'과 '학생'은 결합적 관계에 있으며, '학생'은 그것 대신 올 수 있는 '교수'와 치환관계에 있다. 말하자면, 관계들이 언어 구조를 결정한다.

언어 요소들 중에서도 단어들 사이의 관계인 어휘관계(lexical relation)는 구조주의 의미론에서 중심적인 역할을 했다(Lyons 1977, Cruse 1966, 2000). 하의관계('동물'-'고양이'), 동의관계/유의관계('친구'-'벗'), 반의관계('성공'-'실패'), 부분관계('얼굴'-'코') 등 단어들 사이의 여러 가지 어휘관계가 단어의 의미를 기술하는 중요한 기제이다. 이러한 구조주의적 관점은 근래의 전산적 응용에서도 수용되어 WordNet 같은 어휘데이터베이스 구축의 기반이 되었다(Miller 1991, Felbaum 1998). 그렇다면 어휘관계들은 결합관계인가 치환관계인가? 우선 전통적인 어휘관계들이 대개 명사-명사, 동사-동사, 부사-부사 등 동일한 품사의 단어들 사이의 관계로 인식된다는 것을 고려한다면, 같은 품사의 단어들이 동일한 통사적 맥락에 나타날 수 있다는 의미에서 어휘관계들은 치환관계이다. 그러나 다른 관점에서 볼 수도 있다. 어휘관계 속에 있는 단어들은 같은 언어 맥락 속에서 사용될 가능성이 많다. 어떤 단어를 사용한 후 동일한 단어를 한 문장 속에서 반복하지 않기 위하여 유의어를 사용할 수도 있을 뿐만 아니라(이 경우 유의어가 한 문장에 나타난다), '성공'과 '실패' 같은 반의어까지도 다음과 같이 한 문장 안에서 자연스럽게 사용되는 경우가 많다.

(1) ㄱ. 성공을 하기 위해서는 먼저 실패를 경험해야 한다.
ㄴ. 성공과 실패는 동전의 양면이다.

말하자면, 어휘관계는 치환적 관계이면서 동시에 결합적 관계이다. 본 연구의 관심사인 관련어(related word)는 '병원'과 '의사' 같은 것들이다. 그것들

은 앞서 언급한, 하의관계, 부분관계, 동의/유의관계, 반의관계 등의 어휘관계들로 묶이지 않지만, 어떤 이유에서든지 연상되고 관련되는 단어들이다. 한 단어의 관련어 속에는 그 단어의 하의어, 부분어, 동의어 등도 섞여 있을 수 있지만 그것들에 한정되지 않는다. 한 마디로 말한다면, 관련어는 (적어도 잠재적으로) 우리가 "상대적으로 많이 관련/연관되어 있다고 느끼는 단어"이다. 따라서 관련어는 다른 어휘관계와 달리 정도성이 개입한다.

관련어 관계는 일반적인 어휘관계들과 또 하나의 다른 점이 있다. 하의관계, 부분관계, 동의/유의관계, 반의관계 등은 모두 단어들의 개념으로부터 한정적으로 결정되고, 대개 그 결정에 대해서 사람들 사이의 이견이 없다('성공'이 '실패'의 반의어라는 것에 반대할 사람은 별로 없을 것이다). 그러나 예를 들어, '병원'의 관련어는 '의사'뿐만 아니라 '환자', '간호사', '병', '수술', '약' ... 등 한정되지 않는다. 그리고 관련어들의 목록이나 관련성 정도는 사람마다 다를 수 있다. 이러한 관련어들은, 심리적으로 보자면, 연상작용에 의해 연결되는 단어들로서, 일찍이 심리학자들이 단어 연상 테스트의 방법을 이용하여 측정해 왔다. 영어 단어에 대한 최초의 대규모 조사는 1910년 Kent와 Rosanoff에 의해 수행되었는데, 100개 단어들에 대하여 1000명의 피험자들을 실험하였다(Miller 1991). 예를 들어, 'chair'에 대하여 'table' > 'seat' > 'sit' > 'furniture' > 'wood' 등의 순으로 연상의 결과가 나왔다. 이 경우 우리는 대체로 'chair'에 대한 관련어 순위가 위의 순서라고 할 수 있다.

관련어는 연상작용과 관계되고 그것은 언어 사용자의 경험과 연관된다. 어떤 사물과 다른 사물이 물리적, 인지적으로 같은 맥락에 존재함으로써 그 사물들을 언급(지시)하는 단어들이 서로 관련어의 관계를 맺는 것이다. 그것은 단어들의 연관성이면서 동시에 그 단어들이 지시하는 사물들의 연관성이다. 이러한 연관성이 피험자의 연상작용에 의해 추출될 수 있지만, 이 논문에서 제시할 좀 더 객관적이고 광범위한 방법이 있다. 즉 산출된 언어의 동일한 문맥에 자주 같이 나타나는 단어들이 관련어라고 볼 수 있다. 이것은 일종의 연어(collocation) 즉 공기관계이지만 그 맥락의 범위는, 일반적인 연어 연구에서처럼 핵심어의 앞뒤 몇 단어 혹은 한 문장에 국한되지 않는다. 문장보다 더 큰 단위가 필요한데, 적어도 문단 단위에서의 공기성이 필요하다.[1] 기본

적으로 문장은 주술 관계를 중심으로 제한된 수의 단어만이 참여하고(물론 귀환성 때문에 이론적으로는 그 길이가 무한할 수 있으나 실제 코퍼스에는 평균 10여 개의 단어가 하나의 문장을 구성한다: 홍정하 외 2008), 따라서 문장에서는 제한된 범위의 일부 관련어만 추출할 수 있다. 그러나 하나의 화제와 내용적 응집성을 가지고 전개되는 문단에는 보다 많은 수의 관련된 단어들이 나타난다. 따라서 문장보다 문단이 보다 넓은 범위의 관련어 추출에 유용하다.

이 논문에서 우리는 코퍼스를 기반으로 텍스트의 문단 내에서 의미 있게 공기하는 단어들을 추출하여 관련어 관계를 확립하고 나아가 그것을 기반으로 관련어의 네트워크를 구성하는 방법을 제시하려고 한다. 이것은 단어들이 지시하는 사물, 나아가 개념들의 네트워크를 구축하는 방법인 셈이다.[2] 네트워크를 시각화하기 위해서는 네트워크 시각화 프로그램인 Pajek을 이용한다 (Batagelj, Vladimir and Andrej Mrvar 2010).[3]

다음 절로 넘어가기 전에 한 가지 언급할 것이 있다. 본 논문에서는 '문단'을 사전에 풀이된 바와 같이 "글에서 하나로 묶을 수 있는 짤막한 단위"라는 내용적 정의를 따르기보다는 줄바꿈의 단위라는 기계적인 방식을 따른다(컴퓨터 워드프로세서로 문서를 만들 때 "Enter" 키로 구분되는 단위이다). 추상적인 정의와 기계적인 구분이 대개 일치할 것이다.[4]

2. 문단 내의 공기관계: t-점수

연어, 즉 공기관계는 두 단어가 일정한 문맥에 "의미 있게" 많이 공기하는 것을 말한다. "의미 있게" 공기한다는 것은 예상보다 많이 공기한다는 것이

1) 조은영(2010)이 기존의 여러 연어 연구를 개관한 바와 같이, 한국어 연구에서 연어는 좁은 범위 혹은 문법 관계에 있는 언어요소들 사이의 공기성과 의미전이성에 초점을 맞추었다. 다음 절에서 이 점을 부연할 것임.

2) 이기황, 이재윤(2008)이 사전의 올림말과 풀이말을 기초로 한 네트워크 구성을 연구한 바가 있으나, 이것은 본 연구에서 관심을 가지고 있는 공기어의 네트워크와는 다른 것이다.

3) Pajek 프로그램 홈페이지: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

4) 세 심사자들의 적절한 지적과 질문 덕분에 전체적으로 논문의 주장과 의의가 좀 더 분명하게 드러나도록 논문을 보완할 수 있었다.

다(잠시 후에 상술함). 여기서 문맥은 어떤 중심어(keyword)를 중심으로 좌우 n 단어일 수도 있고, 문장 내에서 중심어와 문법관계를 이루는 단어일 수도 있다. 전자의 예로서, Sinclair(1991)는 중심어의 앞뒤 네 단어들 연어 맥락으로 간주하였다. 후자의 예는, 우리말 연구에서 보통 연어로 간주하는 형용사+명사('새빨간 거짓말'), 명사+동사('나이를 먹다' 같은 것인데, 21세기세종계획의 전자사전 구축에서도 그러한 방식을 채택하였다(홍재성 외 1998~2007). 김진해(2000), 임근석(2009) 등 한국어의 연어 연구는 대략 후자의 관점을 취하였는데, 예외적으로 홍종선 외(2001)는 후자의 관점에서뿐만 아니라 전자의 관점에서도 한국어의 연어 조사를 하였다.

두 단어가 하나의 문맥에 공기하는 것은 단순히 그 두 단어의 공기 빈도가 높은 것을 의미하지는 않는다. 어떤 형용사 A1과 명사 N1가 있을 때, 단순히 전체 코퍼스에서 N1의 빈도가 높기 때문에 A1+N1 공기빈도가 우연히 높을 수 있다. '사람, 말, 때, 일' 등이 그러한 명사들이다. 의미 있게 많이 공기하는 것은 전체 빈도 때문에 우연히 많이 공기하는 것이 아니라 예상보다 더 많이 공기하는 것이다. 즉, 실제 공기 빈도를 O라고 하고 예상 공기 빈도를 E라고 할 때 이 둘 사이의 차이를 고려해야 하는 것이다.

연어 연구에서 실제 공기 빈도 O와 예상 공기 빈도 E를 비교하는 수식으로 자주 사용되는 것이 t-점수(t-score)와 상호정보(mutual information: MI)이다. 아래의 t-점수는 간략화한 식으로서 Church et al. (1991), Stubbs(1995), Barbrook(1996), 강범모(2003), 박병선(2005) 등에서 사용한 수식이다.

(2) $t = \frac{(O-E)}{\sqrt{O}}$

(3) $I = \log\frac{O}{E}$

일반적으로 작은 빈도의 단어들과 관련된 예상빈도는 아주 작기 때문에 상호정보는 그것들(작은 빈도의 단어들)에 대하여 연어 지표를 과장하는 경향이 있어서 많이 사용되지 않는다(분모의 E가 아주 작다). 결국 t-점수가 연어성을 측정하는 데 더 많이 사용되며, 따라서 여기서도 t-점수를 연어성/공

기성 측정에 사용하도록 하자. 어떤 단어 w에 대하여 말하자면, 위 수식에서 사용되는 O는 중심어 k가 나타나는 문단들에 출현하는 w의 빈도이다. E는 1) w가 전체 코퍼스에서 나타나는 빈도, 2) 전체 코퍼스의 크기(어절 수), 3) k가 나타나는 문단들의 크기(어절 수)를 고려하여 계산할 수 있다. 즉, 전체 코퍼스가 10,000어절이고 거기에 단어 w가 100번 나타난다고 하자. 그리고 중심어 k가 나타나는 문단들의 전체 크기가 1000어절 이라고 하자. 그러면 예상 빈도 E = 100 (1000 / 10000) = 10 이다. 실제로 k가 나타나는 문단에서 w가 20번 사용되었다면 t-점수는 다음과 같이 계산된다. (단, 특정 품사의 단어에 집중한다면 어절 크기 대신 그 품사 단어들의 빈도를 기준으로 계산할 수도 있다.)

(4) $t = \frac{(20-10)}{\sqrt{20}} = 2.236$

원래 t-점수의 분포로 위 점수는 99% 확률로 의미가 있는 연어성을 보여준다고 할 수 있지만, 실제 연어 계산에서는 점수 자체보다는 여러 단어들 사이의 연어성의 순서가 중요하다(Manning and Schütze 1999).

이제 코퍼스를 이용하여 실제로 특정 단어의 연어성, 즉 t-점수를 계산해 보자. 사용할 코퍼스는 세종말뭉치(김흥규 외 1998-2007)의 형태의미분석 코퍼스를 단어 단위로 재분석한 1500만 어절 규모의 코퍼스이다.[5] 전체 크기를 어절을 기준으로 측정할 수도 있지만 여기서는 명사에 집중하기 위하여 명사 빈도를 기준으로 한다. 코퍼스에 나오는 전체 명사 빈도 6,927,965이고, 단어 '병원'을 중심어로 할 때 그것이 나타나는 문단들의 명사 빈도는 68,521이다(전체 크기의 약 1/100). 문제의 문단들에 나타나는 단어들 중 '환자'는 그 문단들에서 637회 출현하는데, '환자'는 전체 코퍼스에서 4,049회 출현한다. 따라서 '환자'에 대하여 t-점수는 다음과 같다.

(5) O = 637

5) 원래 21세기세종계획에서 구축한 세종말뭉치는 형태소 단위로 분석되어 있기 때문에(예: '공부+하+는') 그것을 단어 단위로 재분석한 것이다(예: '공부하+는')

$$E = 4049 \times (68{,}521 / 6{,}927{,}965) = 40.047$$

$$t = \frac{(O-E)}{\sqrt{O}} = \frac{(637-40.047)}{\sqrt{637}} = 23.65$$

실제로 ‘환자’는 ‘병원’과 실제 공기빈도가 가장 많으면서 t-점수도 가장 높다. 그러나 실제 공기빈도가 높다고 반드시 t-점수가 높은 것은 아니다. 예를 들어(표 1 참조), ‘진료’는 전체 코퍼스에 562회 그리고 ‘병원’이 나오는 문단에서 164회 사용되어, t-점수가 12.37이다. 반면에 ‘집’은 전체 코퍼스에서 22,237회 나타나고 ‘병원’이 나오는 문단에서 434회 사용되어, t-점수는 10.28이다. 즉, ‘병원’이 나오는 문단에서 ‘집’이 ‘진료’보다 3배가량 더 많이 사용되지만 t-점수는 ‘진료’가 ‘집’보다 더 높다. 그것은 ‘집’이 전체 코퍼스에서 ‘진료’보다 40배 이상 많이 나오고 따라서 ‘병원’이 나오는 문단에서도 (우연히) 많이 나타났기 때문이다. 실제로 ‘병원’에 대하여 우리가 느끼는 관련어는 ‘집’보다는 ‘진료’이다.

위와 같은 방법으로 우리는 ‘병원’의 관련어들을 다음과 같은 순서로 계산해 낼 수 있다(상위 50개). 아래에서 보는 바와 같이 ‘환자, 의사, 의료, 치료, 수술, 단체, 진료, 약, 정신, 비영리, ...’ 순으로 t-점수가 높고, 따라서 그 순서로 ‘병원’과 연관성이 높다고 할 수 있다. 10위 바깥이긴 하지만 ‘어머니, 아내, 엄마, 아버지, 아들’ 등의 친족어가 ‘병원’의 공기어로 나타나는 것은 병원이 등장하는 상황 중 가족의 치료나 입원이 많이 있기 때문으로 보이며, ‘집’도 마찬가지 이유로 많이 나타나는 것으로 보인다(가족과 집은 매우 관련성이 높다). 일견 이러한 것들은 다른 것들보다 연상 작용과는 다소 거리가 있어 보인다. 하지만 의식적인 연상이 아니라 연관성에 대한 무의식이 있을 수 있다. 혹은 개인적인 경험에서 병원과 가족 및 집은 밀접한 (의식적인) 연관이 있을 수도 있다. 또한 다른 관점에서 보자면, 경험이 개인적인 것이듯이 코퍼스에 있는 텍스트도 개별적인 것이므로 연상에서나 코퍼스에서나 특수한 예가 나타날 가능성은 존재한다. 마지막으로 이러한 예상치 못한 공기어가 새로운 관련어와 관련성의 발견 절차의 시작이 될 수 있다.[6]

6) 한 심사자가 언급한 대로, ‘병원’은 공기어 추출에서 특수한 지위에 있다. 즉, 병원이 Fillmore의 틀 의미론(Frame Semantics: Fillmore and Atkins 1992)에서 말하는 틀(frame), 즉 하나의

순위	관련어 (명사)	'병원' 문단 내 빈도(O)	코퍼스 빈도	예상 빈도(E)	t-점수
1	환자	637	4049	40.047	23.652
2	의사	488	3201	31.659	20.658
3	의료	394	3092	30.581	18.309
4	치료	287	2006	19.840	15.770
5	수술	241	1384	13.688	14.642
6	단체	311	7023	69.461	13.696
7	진료	164	562	5.558	12.372
8	약	182	1753	17.338	12.206
9	정신	245	6120	60.530	11.785
10	비영리	131	738	7.299	10.808
11	병	157	2360	23.342	10.667
12	어머니	302	12388	122.523	10.328
13	집	434	22227	219.836	10.280
14	종합	136	2009	19.870	9.958
15	노인	148	3398	33.608	9.403
16	남편	198	6703	66.296	9.360
17	검사	116	1696	16.774	9.213
18	입원	84	229	2.265	8.918
19	아내	164	5348	52.894	8.676
20	보험	110	2248	22.234	8.368
21	경찰	139	4243	41.965	8.230
22	응급	71	233	2.304	8.153
23	암	80	738	7.299	8.128
24	간호사	72	315	3.116	8.118
25	엄마	173	6814	67.394	8.029
26	약국	72	411	4.065	8.006
27	진단	76	787	7.784	7.825
28	건강	97	2059	20.365	7.781
29	응급실	62	116	1.147	7.728
30	아이	307	17472	172.807	7.659
31	아버지	230	11754	116.253	7.500
32	서비스	111	3301	32.649	7.437
33	기관	142	5432	53.725	7.408

의미 영역을 담당하는 역할을 하기 때문에 공기어가 보다 일관적으로 추출될 수 있었다. 그러나 틀을 가리키지 않은 단어에 대해서 반드시 공기어의 일관성이 감소하는지는 여러 경우를 비교해 보아야 확인될 것이다.

34	아들	137	5199	51.421	7.312
35	돈	215	10912	107.925	7.302
36	차	122	4215	41.688	7.271
37	원장	62	713	7.052	6.978
38	치료비	50	110	1.088	6.917
39	병실	52	232	2.295	6.893
40	사고	95	2832	28.010	6.873
41	날	229	12645	125.065	6.868
42	가족	138	5803	57.395	6.862
43	증세	57	564	5.578	6.811
44	아기	98	3171	31.363	6.731
45	몸	209	11294	111.703	6.730
46	혈액	55	515	5.094	6.729
47	시설	97	3116	30.819	6.720
48	딸	100	3397	33.598	6.640
49	임신	55	758	7.497	6.405
50	진찰	42	120	1.187	6.298

표 1 '병원'의 관련어(t-점수 순, 상위 50개)

3. 관련어와 관련 사물/개념

앞에서 우리는 세종말뭉치 코퍼스를 대상으로 특정 단어에 대하여 동일 문단에 공기하는 단어들의 빈도를 t-점수로 계산함으로써 관련어들과 그 정도를 추출할 수 있음을 보였다. 한편, 관련어들은 연관되는 단어들인 동시에 연관되는 사물들을 표상하며, 나아가 연관되는 개념들을 표상한다. 일찍이 오그든과 리처즈가 의미(표의작용)의 삼각형을 제안한 이래(Ogden and Richards 1923), 단어(언어), 개념(생각), 사물(세계) 사이의 삼각관계는 심리학적 의미론의 기반이 되었다(Lyons 1977 등). 즉, 단어와 사물은 직접적인 연관성이 없고 그 사이에 있는 개념을 통하여 간접적으로 연관된다는 이론이다. 반면에 논리학에 기초한 형식의미론에서는(Montague 1974, Portner and Partee 2002 등) 개념의 층위가 없이 단어와 사물이 직접 연결된다고 주장한다. 단어의 의미에 관하여 심리적인 관점을 채택하든 논리적인 관점을 채택하든 단어가 (직접적으로 혹은 간접적으로) 세상의 사물과 연관된다는 것은

이의가 없다. 따라서 단어들의 공기관계로 나타나는 단어들의 연관성은 바로 그 단어들이 가리키는 사물들의 연관성을 표상한다고 볼 수 있다. 나아가 의미에 대한 심리적인 관점을 취한다면 단어들의 연관성은 개념들의 연관성을 표상한다고 할 수도 있다.[7)]

예를 들어, 앞에서 단어 '병원'이 단어 '환자, 의사, 의료, 치료, 수술, 단체, 진료, 약' 등과 이 순서로 밀접한 관련이 있다고 하였다. 이것은 코퍼스에서 나타난 단어들의 공기관계(연어)로부터 추출한 결과이다. 그렇다면 이것은 세상에 있는 사물인 병원이 세상의 사물인 환자, 의사, 의료, 치료, 수술, 단체, 진료, 약 등과 이 순서로 연관된다는 것을 말한다. 나아가 "병원"이라는 개념이 "환자, 의사, 의료, 치료" 등의 개념과 연관되어 있다는 것을 보여준다.

이러한 단어, 사물, 개념의 연관성은 언어를 사용하는 공동체에 의존한다. 언어 공동체 내에서 사람들의 생활양식이 언어에 반영되고, 언어를 통하여 언어 공동체의 생활양식(문화), 나아가 정신세계의 단면을 알 수 있다. 이러한 연관성은 한 단어의 관련어들이 아니라 여러 많은 단어들의 관련어들을 조사함으로써 확장될 수 있고, 이러한 과정을 통하여 언어 공동체의 문화와 정신세계에 대한 이해도 확장될 수 있다.

물론 엄격히 말하자면 본 논문에서 실제로 텍스트를 통하여 발견한 것은 단어의 연관성이고 관련어의 네트워크이다. 한 단어에 여러 뜻이 있을 수 있고 두 개의 단어가 하나의 뜻과 연결될 수도 있다. 그러나 앞에서 논의한 전통적인 의미삼각형을 고려하고, 나아가 단어의 사용까지도 단어의 의미를 형성하는 데 기여한다는 코퍼스언어학의 관점을 수용한다면(Firth 1957, Teubert and Krishnamurthy 2007) 단어의 사용이 다르면 의미가 다르다고 보는 것, 따라서 단어의 연관성을 곧 의미/개념의 연관성으로 보는 것이 의의가 있을 것이다. 그럼에도 불구하고 독자는 이 논문에서 제시하는 모든 네트워크를 단어의 네트워크로만 받아들일 여지는 있다.

---

7) 형식의미론의 관점에서, 프레게(Frege)의 뜻(sense)를 개념에 대응시킬 수도 있다. 한편, 최근에는 형식의미론의 관점을 취하면서도 사물뿐만 아니라 개념 및 그 표상을 중요시하기도 한다(Cann, et al. 2009). 언어에는 대용, 생략 등 맥락이 중요하게 작용하는 형상들이 큰 부분을 차지하고 있기 때문에 순전히 언어와 세계만으로 자연언어의 의미론이 충분하게 성립할 수 없기 때문이다.

4. 관련 단어/사물/개념의 네트워크

한 단어에 대한 공기어들은 단순한 그래프를 구성할 수 있다. 즉 중심어(‘병원’)와 공기어들(‘환자, 의상’ 등)의 결점들(vertex, node), 그리고 중심어와 공기어들 사이의 연결선(arc, edge, link)의 그래프이다. 앞의 방식으로 계산된 t-점수가 반영된 연결선은 중심어로부터 공기어로 가는 방향성이 있으므로(그래프 이론의 용어를 따르자면 edge가 아니라 arc이다) 이 그래프는 방향적 그래프(directed graph)이다. 다른 단어에 대해서도(예: ‘학교’) 그것의 공기어들을(‘교육, 학생, 교사’ 등) 구할 수 있고 그것들의 그래프를 구할 수 있다. 이러한 여러 그래프들을 동시에 표상할 때 그것은 좀 더 복잡한 하나의 그래프, 즉 네트워크(network)를 구성한다.

좀 더 정확하게 정의하자면 그래프는 결점들, 그리고 두 결점의 쌍들을 이어주는 연결선들의 집합이며, 네트워크는 “하나의 그래프, 그리고 그 그래프의 결점들과 연결선들에 대한 부가적 정보로 구성된다”(Nooy, Mrvar, and Batagelj 2005). 결점의 부가정보는 기본적으로 그 결점의 표지이며 관련어들의 네트워크인 경우 그것은 단어 자체이다.[8] 연결선에 대한 부가정보는 그 방향성과 연결의 강도이다. 앞서 공기어의 t-점수를 구했는데 관련어들의 네트워크에서 그 점수가 중심어와 공기어(관련어) 사이의 연결의 강도이다.

네트워크는 그래프이므로 시각화할 수 있다. 몇 개의 결점들만 있다면 사람이 대략적으로 그 그래프를 그릴 수 있겠지만, 수백 개 혹은 그 이상의 결점들이 있고 그것들이 복잡하게 연결되어 있는 네트워크를 사람이 그릴 수는 없다. 그렇게 때문에 네트워크의 시각화를 도와주는 컴퓨터 프로그램이 필요하다. 본 연구에서는 네트워크 시각화에서 가장 많이 사용되는 Pajek(파이엑) 프로그램을 사용한다. Pajek은 슬로베니아의 수학자들이 개발한 프로그램으로 이 연구에서 사용한 것은 Pajek 1.26 버전이다(Batagelj and Mrvar 2010).

앞에서 제시한 방식대로 ‘병원’, ‘학교’, ‘회사’, ‘은행’의 관련어들의 t-점수를 계산하고 각 단어의 상위 10개 관련어들의 네트워크를 Pajek으로 시각화하면 그림 1와 같다. 관련어 네트워크는 관련 개념의 네트워크를 표상하므로

8) 결점의 다른 부가적인 정보들도 있지만 우리의 논의에서는 중요하지 않다.

앞으로의 논의에서 관련어와 관련개념 그리고 관련 사물을 혼동하여 사용하기로 한다.9)

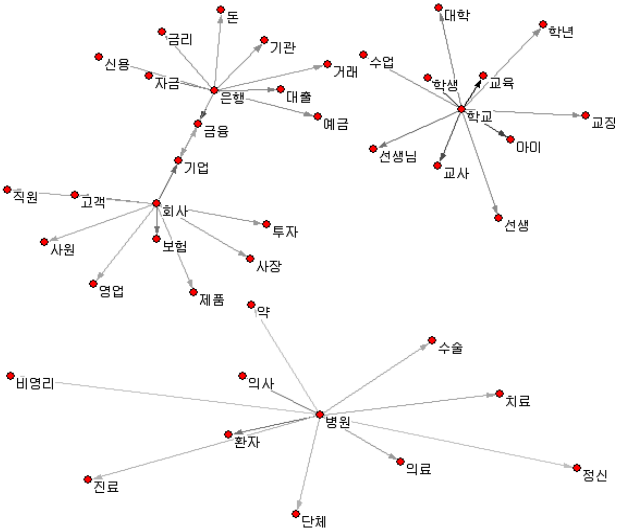

그림 1 병원, 학교, 은행, 회사의 관련어/관련개념 네트워크 (상위 10개씩)

이 네트워크에서 병원과 학교의 네트워크는 독립적인 네트워크를 구성하지만 은행과 회사의 네트워크는 금융과 기업의 결점으로 연결되어 있다. 연결선의 길이는 그것의 강도를 반영하는데, 그 강도가 셀수록 연결선이 짧다(즉, 결점들이 더 가깝다). 예를 들어, 학교의 네트워크를 보면, 교육, 학생, 교사, 아이가 선생, 선생님, 대학 등보다 더 가깝게 연관된다. 은행이 금리, 대출, 금융과 밀접하게 연관되고 회사가 기업, 고객, 투자와 밀접하게 연관되는 것도 우리 상식에 맞는다. 금융과 기업은 은행과 회사에 모두 연관되지만 금융은 은행에, 기업은 회사와 더 가깝게 연관된다.

위의 간략한 네트워크를 보면 병원, 학교, 은행, 회사가 어느 정도 독립적인 것으로 보이지만 더 많은 관련어들을 보면 그것들은 상호 연결되고 네트

9) 예를 들어, a와 b가 연관된다고 기술할 때, 그것은 단어 'a'와 'b', 사물 a와 b, 개념 "a"와 "b"가 연관됨을 말한다.

워크는 더 복잡해진다. 그림 2은 각 단어들의 관련어들 중 상위 50개씩이 구성하는 네트워크이다.

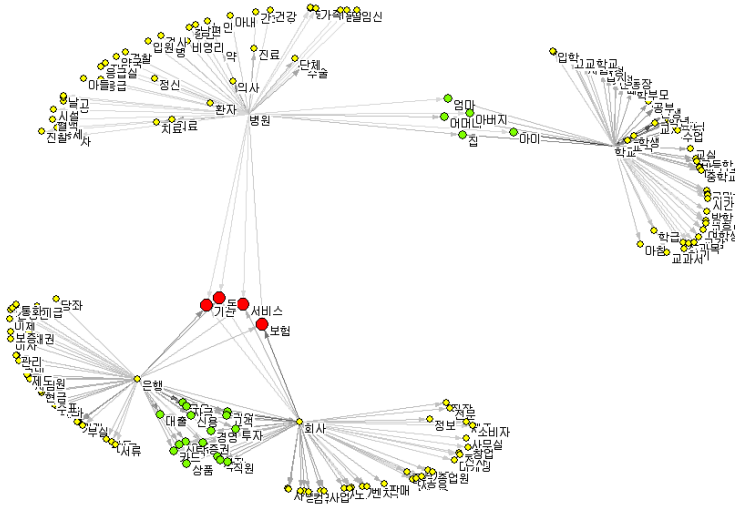

그림 2 병원, 학교, 은행, 회사의 관련어/관련개념 네트워크 (상위 50개씩)

이 네트워크를 보면 네 개의 하위 네트워크가 약하게나마 연결되어 있는 것을 볼 수 있다. 네 개의 중심어에 공통된 관련어는 없지만 병원, 은행, 회사에 공통된 것들로 돈, 기관, 서비스, 보험이 있다. 학교는 넷 중에서는 가장 독립적인데, 엄마, 어머니, 아버지, 집, 아이 등의 일상적인 개념을 통하여 병원과 연결될 뿐 회사 및 은행과는 공통된 것이 없다. 은행과 회사는 가장 밀접하게 연관되는데, 금융, 자금, 기업, 고객, 신용, 대출, 경영, 투자, 업무, 신탁, 카드, 증권, 상품, 시장, 주식, 직원이 공통으로 연관된다.[10]

병원, 학교, 은행, 회사의 관련어/관련개념 네트워크는 최상위 것들에 대해서 우리가 어느 정도 예상할 수 있지만 100개, 200개 등 확장될수록 우리의 예상은 불확실해진다. 코퍼스에 기초한 관련어 추출 방법은 (t-점수의 유의성

10) 공간적인 제약으로 모든 결점의 표지가 네트워크 표상에 보이지 않지만 Pajek은 그래프의 일부를 확장해서 볼 수 있는 방법을 제공함으로써 모든 결점을 확인할 수 있게 해 준다.

이 허용하는 한에서) 1000개 이상의 관련어를 찾아줄 수도 있다.

한편 어떤 단어들은 최상위 관련어에 대해서도 우리의 예상이 불확실하다. 예를 들어, 고유명사 '미국', '중국', '일본'의 최상위 관련어들은 무엇이며 그것들의 네트워크는 어떤 모양일까? 이 질문에 대해서는 코퍼스에 기초하여 그것들을 계산해 볼 수밖에 없다. 그 결과가 우리의 상식에 맞는지는 계산 이후에 확인할 수 있다(즉, 우리는 관련어를 예상할 수는 없지만, t-점수에 의해 나온 관련어들이 적절한 것인가를 확인할 수는 있다). 그림 3은 '미국', '중국', '일본'과 공기하는 일반명사의 네트워크이다(상위 50위).

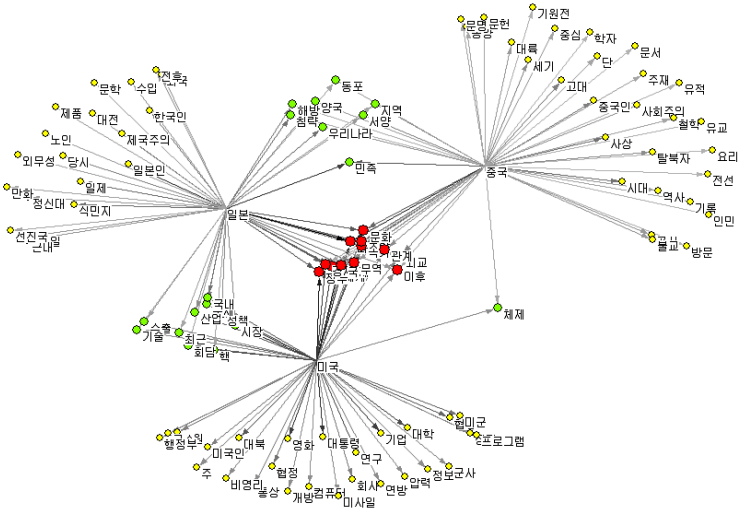

그림 3 미국, 중국, 일본의 관련어/관련개념 네트워크 (상위 50개씩)

위 그래프를 살펴보기 전에 확인할 것이 있다. 즉, 미국, 중국, 일본의 관련어/관련개념은 한국어 텍스트에 나타난 것이고 우리나라 사람들의 관심과 생각을 반영한 것이라는 점이다. 우리의 관점에서 미국은 무엇 무엇과 연관되고 중국은 무엇 무엇과 연관된다는 말이다.

미국, 중국, 일본 모두는 나라, 정보, 경제, 전쟁, 세계, 국제, 무역, 국가, 조약, 문화, 관계, 외교와 연관되어 있다.[11] 미국과 일본은 (4국 공통의 관련어/

관련개념을 제외하고) 공통적으로 구출, 회담, 핵, 시장, 수출 등과 연관되고, 일본과 중국은 동포, 해방, 침략, 민족 등과 연관된다. 미국과 중국은 공통적으로 체재와만 연관된다. 그러면 각국에 (우리나라의 관점에서) 고유한 관련어/관련개념은 무엇인가? 미국은 미군, 대학, 영화, 컴퓨터 등과 배타적으로 연관되고, 일본은 식민지, 제국주의, 만화 등과 연관되며, 중국은 대륙, 사회주의, 탈북자, 유교 등과 연관된다. 미국과 영화, 일본과 만화, 중국과 탈북자의 연관성은 현대 한국 사회와 문화의 실상을 잘 반영한다. 그 밖에도 상식에 부합하는 여러 연관성들이 발견된다. 반면에 어떤 것들은 의외의 연관관계를 보인다.[12] 예를 들어 '일본'-'일본인', '중국'-'중국인', '미국'-'미국인'이 당연한 연관관계인 반면, '한국인'은 '일본'과만 연관된다. 일본에 대한 언급과 고려의 과정에서 한국인이 더 많이 언급되고 고려된다는 것을 보여준다. 예를 들어 한국인과 일본인, 한국 문화와 일본 문화의 비교가 더 많이 이루어지고 있을 수 있다. 이것은 새로운 발견의 단서를 제시한다.

한편, 텍스트에서의 공기어를 기반으로 찾아내는 단어와 개념의 연관성들이 상식에 부합하기 때문에 코퍼스언어학이 필요 없다고 하는 사람이 있다면 그것은 잘못이다. 이러한 연관성들은 그것들의 발견 이후에는 적절한 것으로 인식될 수 있지만 직관으로 발견되기 힘들기 때문에 우리의 방법이 정당성을 갖는다.

특히 추상적인 개념들과의 관련개념은 직관적으로 발견하기가 쉽지 않다. 그러나 코퍼스에 기초한 관련어 추출 방법을 이용하면 쉽게 그 관련어/관련개념들을 발견할 수 있다. '자유, 정의, 진리, 평등'의 관련어들을 추출하여 네트워크로 구성하여 보자(그림 4, 상위 50개).

11) 그래프에 나타나는 '이후'는 개념적으로는 일견 관련어의 범위에서 벗어나지만, '미국, 중국, 일본'이 나타나는 문단에서 자주 사용되기 때문에 관련어로 나타난다(예: '세계대전 이후...'). 이것은 오히려 연상으로의 관련어 발견 절차를 넘어서는 좋은 결과일 수도 있다.

12) 이 점, 그리고 '한국인'은 예는 한 심사자자가 지적한 것이다.

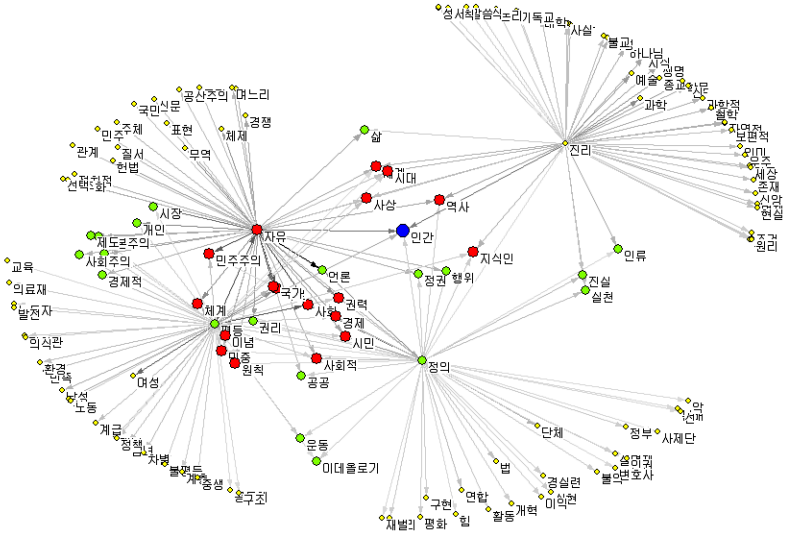

그림 4 자유, 정의, 진리, 평등의 관련어/관련개념 네트워크 (상위 50개씩)

자유, 정의, 진리, 평등 유일한 공통 관련개념은 인간이다. 그 네 가지는 모두 인간의 보편적 가치이며 인간의 존엄성과 관련되기 때문이다. 자유, 정의, 평등은 민주주의, 국가, 체계, 사회, 시민, 이념, 민중 등을 매개로 서로 연관되며, 진리는 이 셋과는 다소 떨어져 있다(그래프 상의 거리가 그것을 반영한다). 물론 진리도 지식인을 매개로 정의와 평등, 그리고 역사를 매개로 자유와 정의와 연관되기는 한다. 네 가지 개념의 독자적인 관련개념들은 보면, 진리는 과학, 예술, 종교 등과 연관되고, 자유는 민주, 무역, 경쟁 등과 연관되고, 평등은 여성, 노동, 계급 등과 연관되며, 정의는 법, 평화, 개혁 등과 연관된다. 위 네트워크 그래프는 개념들의 연관성 자체만이 아니라 그것들 사이의 연관성의 정도까지 표시해 준다(넷 중에서 자유, 정의, 평등이 서로 더 많이 연관된다). 정의와 평등이 개념적으로 매우 밀접하지만 관련어로 보면 이 둘의 연관성이 자유와의 연관성보다 크지는 않다.

여기서 한 가지 호기심을 자극하는 질문은 비슷한 뜻의 단어들이 공통의 연관어를 많이 가지고 있는가 하는 것이다. 한 가지 예로서 '언어'와 '말'의 관련어 네트워크를 살펴보자(그림 5, 상위 100개씩).[13]

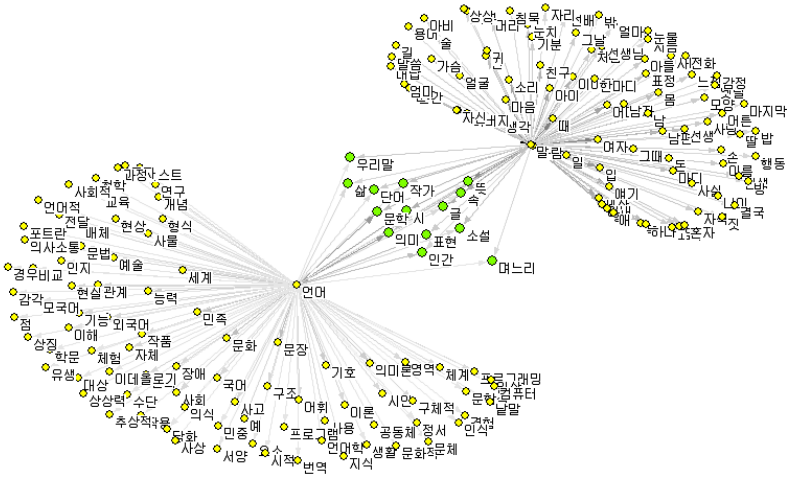

그림 5 언어와 말의 관련어/관련개념 네트워크 (상위 100개씩)

이 네트워크에 나타나는 특징은 ‘언어’와 ‘말’이 유의어임에도 불구하고 공통의 관련어가 많지 않다는 것이다. 우리말, 삶, 단어, 작가, 문학, 의미, 인간, 뜻 등의 14개 공통의 관련개념 이외에 나머지 172개(각각 86개) 개념들은 언어 혹은 말 하나의 개념과만 연관된다. 이것은 비슷한 뜻을 가진 단어들의 사용 환경(레지스터 포함)이 매우 다를 수 있다는 것을 의미하며, 이러한 사실을 언어 교육이나 사전편찬을 할 때 특히 염두에 두어야 할 것이다.

그렇다면 반의어의 경우는 어떠한가? 하나의 예로서 ‘남자, 여자, 남성, 여성’의 관련어 네트워크를 구성해 보자. 먼저 ‘남자’와 ‘여자’의 관련어/관련개념 네트워크는 그림 6과 같다.

13) ‘말’은 중의적이지만(“언어”, “동물”) 본 연구를 위해 사용한 형태의미분석 코퍼스에 동음이의어가 구분되므로 언어의 뜻으로의 ‘말’만 고려되었다.

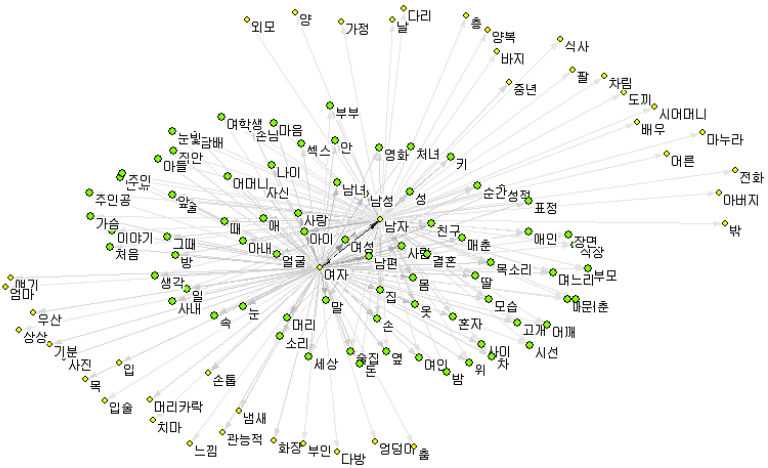

그림 6 남자와 여자의 관련어/관련개념 네트워크 (상위 100개씩)

사랑, 아내, 남편, 여성, 남성, 등 대부분의 관련개념들은 남자와 여자에 공통으로 연관된다(79개). 남자 혹은 여자에만 연관되는 관련개념들은 각각 21개밖에 없다(남자 - 중년, 양복, 마누라 등, 여자 - 머리카락, 관능적, 춤 등). 이것은 반의어가 가장 가까운 단어라는 일반적인 사실을 다시 한 번 일깨워준다(Lyons 1977, Cruse 1996, 2000, Croft and Cruse 2004). 반의어는 개념적으로 반대의 뜻이어야 할 뿐만 아니라 그 단어 자체가 동일한 레지스터(사용역)에서 사용되는 쌍인 것이다.[14] 예를 들어, '남자'의 반의어는 '여자'이지 '여성'이 아니다. '여성'은 '남성'의 반의어이다. 남성과 여성까지 포함한 그림 7의 네트워크가 그러한 사실을 더욱 잘 보여준다.[15]

14) 반의어의 공기에 대하여 이기황(1988)이 논의한 바 있다.

15) 한 심사자가 지적한 대로, 위 네트워크는 여자를 성적 대상으로 보는 시각을 많이 반영한다. 그림 7에서 나타나는 바와 같이 '여성' 관련어와 대조된다.

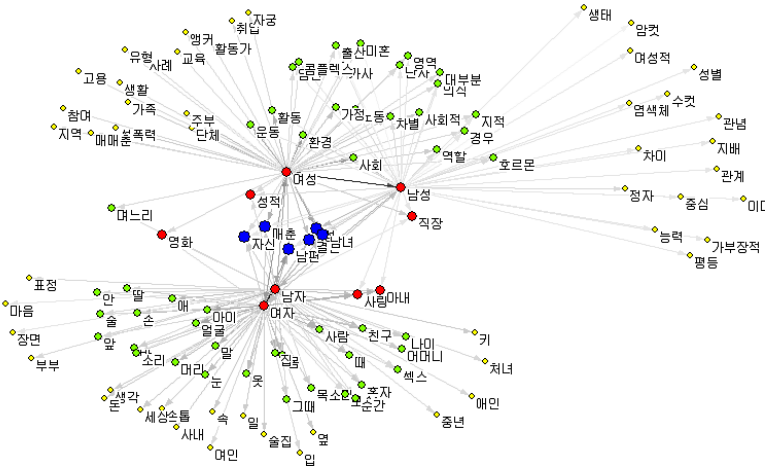

그림 6 남자, 여자, 남성, 여성의 관련어/관련개념 네트워크 (상위 50개씩)

남자, 여자, 남성, 여성 모두에 연관된 것은 자신, 남편, 결혼, 성, 남녀, 매춘뿐이다. 영화, 직장, 아내, 사랑 등, 세 개의 개념과 연관된 것이 몇 개 있기는 하지만 대부분 남자-여자의 공통 관련개념, 남성-여성의 공통 관련개념이 40개 이상이다. 이것은 남자와 남성, 여자와 여성의 공통으로 연관된 개념이 많지 않은 것과 대비된다. 역시 뜻이 비슷한 것보다 (반의어에 기초한) 반대의 뜻을 가진 개념들이 공통의 연관성을 보일 수 있다는 사실을 드러낸다. 아울러 이 네트워크는 남자-여자의 연관성이 남성-여성의 연관성보다 더 강하다는 것도 보여준다.

지금까지 두 개 내지 네 개의 단어에 대한 관련어에 기초한 네트워크를 제시하고 논의하였다. 더 많은 단어들의 관련어들의 네트워크는 매우 복잡해지므로 시각화가 힘들 수 있다. Pajek은 기본적인 네트워크 속성을 비시각적으로 보여주는 장치들을 가지고 있는데, 그것들을 여기서는 논의하지 않는다. 또한 여러 중심 단어들의 목록을 정하고 그 단어들 사이의 관련성을 본다면, 많은 단어들 사이의 양방향 관계를 포함하는 네트워크를 구성할 수도 있다. 이것 역시 여기서는 논의하지 않는다. 모두가 차후 연구에서 고려할 것들이다.

5. 공기관계의 정교화

앞의 관련어/관련개념 네트워크는 중심어 k가 나타나는 문단에 공기하는 단어 w의 문단 내 빈도를 바탕으로 t-점수를 계산함으로써 결정되었다. 여기서 중심어 k가 나타나는 문단이라는 것은 k가 한 번이라도 나타나는 문단이라는 뜻이다. 즉, 중심어 k가 한 번이라도 나타나는 문단에 발생하는 단어 w의 문단 내 빈도가 n이라면 n 자체를 계산에 반영한 것이다.

그런데 한 문단에는 중심어 k가 두 번 이상 발생할 수도 있다. 실제로 단어 '은행'이 한 문단 내에서 발생하는 빈도는 다음과 같다.[16]

빈도	1회	2회	3회	4회	5회	6회	7회	8회	9회	합계
문단 수	2087	359	93	25	11	6	3	2	2	2588
비율 (%)	80.6	13.9	3.6	1.0	0.4	0.2	0.1	0.1	0.1	100

표 2 '은행'의 한 문단 내 발생빈도

한 문단에서 중심어가 2회 이상 발생하는 것은 문단 수로 20%에 미치지 못한다. 그것도 2회 발생한 경우가 대부분이고 그것보다 많이 발생하는 경우는 별로 없다. 그러니까 우리가 t-점수를 계산했던 방식이 크게 문제될 것은 없는 것으로 보인다. 그렇다 하더라도 중심어가 1회 발생한 경우와 2회 이상 발생한 경우 그 문단에 발생하는 공기어의 무게가 다르다고 주장할 수도 있다. 이러한 점을 고려하여 생각해 볼 수 있는 방식은 중심어 k의 문단 내 빈도 n을 고려하여, 공기어 w의 문단 내 빈도가 m이라면 t-점수 계산에서 공기어 빈도를 m n으로 생각하는 것이다. 이 경우 공기어가 나타나는 환경, 즉 문단의 어절 빈도도 그만큼 늘려주어야 할 것이다. 즉, 중심어 k가 n번 나타나는 문단의 어절 크기가 p이라면 t-점수를 계산을 위한 중심어 발생 문단의 어절 크기를 p n으로 계산하는 것이다. 이렇게 하면 문단 내 중심어의 빈도를 공기어 빈도와 문단 크기에 모두 반영하는 셈이다.

실제로 이러한 방식으로 '은행'의 공기어 네트워크를 구성하여 원래의 방

16) 물론 기반 코퍼스의 동음이의어 구분이 되어 있으므로 식물 '은행'은 고려되지 않았다.

식, 즉 문단 내 중심어가 두 번 이상 출현하더라도 한 번 출현한 것과 마찬가지로 계산하여 구성한 네트워크와 비교하면 그림 8과 같다. 이 네트워크에서 '은행'은 원래의 방식(중심어 빈도 상관 없음)에서 관련어의 중심이고 '은행5'는 중심어 빈도를 고려한 방식에서 관련어의 중심이다.

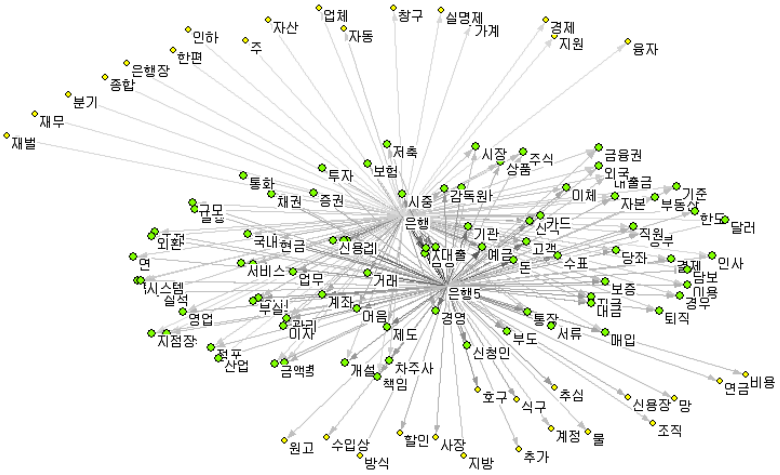

그림 8 두 가지 방식으로 구축된 '은행' 관련어 네트워크

이 네트워크를 보면 두 개의 방식이 공통적으로 83개의 관련어와 관련되어 있고 각각 17개의 관련어들이 서로 다르다. 중심어 빈도를 고려하지 않은 방식에서만 유효한 관련어들과 중심어 빈도를 고려한 방식에서만 유효한 관련어들을, 상식에 비추어 적절한 것과 그렇지 않은 것으로 나누어 제시하면 다음과 같다.[17]

17) 이것은 연구자 개인의 상식이므로 사람마다 판단이 다소 다를 수 있다.

적절성 판단	문단 내 중심어 빈도 고려하지 않음	문단 내 중심어 빈도 고려함
관련어로 적절	재벌, 재무, 은행장, 분기, 인하, 자산, 업체, 창구, 실명제, 가계, 경제, 지원, 융자, 자동 (14개)	수입상, 할인, 사장, 계정, 추심, 신용장, 조직, 망, 연금, 비용 (10개)
관련어로 부적절	종합, 한편, 주 (3개)	원고, 방식, 지방, 호구, 추가, 식구, 물 (7개)

표 3 두 가지 방식으로 구축된 '은행' 네트워크 비교

'은행'의 예를 보면 중심어를 빈도를 고려하지 않는 것이 중심어 빈도를 고려하는 것보다 상대적으로 더 적절한 결과를 산출하는 것으로 보인다. 2절에서 예로 들었던 '병원'을 같은 방식으로 조사해 보아도 역시 동일한 결과를 보인다(친족어는 관련어로 판단함, 네트워크 그래프 생략).

적절성 판단	문단 내 중심어 빈도 고려하지 않음	문단 내 중심어 빈도 고려함
관련어로 적절	질병, 의약, 선생님, 보건소, 요양, 이상, 약사, 질환, 정신과, 민간, 할머니, 부인 (12개)	박사, 의료원, 뇌, 보건, 치과, 감기, 형, 딸아이 (8개)
관련어로 부적절	학교, 때, 전화 (3개)	급, 크레파스, 그때, 후, 미끄럼틀, 박쥐, 곳 (7개)

표 4 두 가지 방식으로 구축된 '병원' 네트워크 비교

이러한 일이 발생하는 이유는 다음과 같은 것으로 추정된다. 즉, 일반적으로 문단의 길이가 길수록 중심어가 더 많이 발생할 가능성이 있다. 어떤 문단에 관련성이 적은 단어가 발생하면(이것도 긴 문단일수록 더 가능성이 많다) 그것의 중요성이 중심어의 다중 빈도로 인하여 더 높아진다고 볼 때, 중심어 빈도를 고려하는 것이 관련성이 적은 단어들을 관련성이 높은 것으로 만들 가능성이 있는 것 같다. 물론 '은행'뿐 아니라 모든 예들을 확인해야 확실하게 말할 수 있지만, 적어도 중심어 빈도를 고려하지 않는 단순한 방식이 빈도를 고려하는 방식보다 못할 것이 없다는 사실을 보여준다고 할 수 있다.[18]

관련어 추출을 위한 방식으로 이 연구에서 채택한 t-점수는 여러 연구에서

채택하는 연어 계산 방식이다(Church et al. 1991, Stubbs 1995, Barbrook 1996, 강범모 2003). t-점수 방식이 방향성을 가진 네트워크를 구성하는 데 사용될 수 있는 반면에 정보과학에서 흔히 사용하는, 확률을 이용한 상호정보(mutual information)는 방향성이 없이 공기관계를 포착한다. 어떤 단어 x와 y가 있을 때 상호정보 I는 다음과 같이 계산된다.

(6) $$I = \log \frac{p(x,y)}{p(x) \times p(y)}$$

여기서 p(x)와 p(y)는 단어 x와 y의 발생확률로서 x와 y가 우연히 같이 나타날 확률은 p(x)와 p(y)를 곱한 값이다. 실제로 x와 y가 같이 나타난 확률이 p(x,y)인데, p(x,y)가 p(x) p(y) 보다 크다면 (즉, I가 + 값이라면) 그것은 x와 y가 의미 있게 공기한다는 것을 보여준다(즉, 공기하는 것이 우연이 아니다). 반면에 p(x,y)와 p(x) p(y)가 동일하다면 (즉, I = 0이면) x와 y는 우연히 같이 나타난 것이고 의미 있는 공기관계가 없다고 할 수 있다. 그런데 이 방법의 문제는 O/E를 고려할 때와 마찬가지로 작은 빈도의 단어들의 확률이 아주 작은 만큼 p(x) p(y)가 0에 가깝게 되고 상대적으로 I의 값을 많이 증가시킨다는 것이다. 즉, 빈도가 작은 단어들에 상대적으로 더 많이 공기성을 부여하게 된다. O/E의 경우와 마찬가지로 하한 빈도를 정해 놓고 관련어를 찾아낼 수도 있으나 지나치게 임의적이다.

또한 상호정보를 그대로 적용하기 힘든 이유는 위 수식은 두 단어 x와 y가 연속하여 나타날 때에만 적용된다는 것이다.[19] 우리의 연구는 한 문단 내에 나타나는 두 단어의 공기관계의 크기를 구하는 것인데, 이 경우에 위 수식은 맞지 않는다. 만일 p(x)와 p(y)를 모든 단어의 쌍(bigram)의 수를 고려한 확률

18) 그런데 이러한 논리는 문단이라는 단위에 국한될 수도 있다, 좀 더 큰 단위, 예를 들어 신문의 기사 하나 혹은 책의 절 하나같은, 문단보다 훨씬 큰 단위를 고려할 때에도 같은 결과가 나올지는 불확실하다. 오히려 중심어가 영향을 미치는 범위가 너무 넓어 관련이 없는 단어들까지도 관련어로 만들 수도 있다. 이 경우 다른 방도를 고려해야 할 것이다.

19) 원래 MI는 반드시 인접한 두 요소에만 적용되는 것은 아니다. 그러나 단순히 p(x)와 p(y)를 곱한 것은 x와 y가 우연히 인접하여, 이 순서로 나타날 경우의 확률이다. n 단어 거리 내에 (0~n 단어 거리에) 두 요소가 우연히 나타날 확률은 조합(combination)을 고려해야 한다.

로 계산한다면 모든 단어에 대하여 그 확률은 거의 0에 가까워져서 그 차이가 없게 되고 분자, 즉 공기빈도(공기확률)만이 상호정보에 영향을 미친다. 즉 두 단어 x와 y의 공기성 순서는 x와 y의 공기빈도 순서와 일치하게 된다. 이것은 절대적 공기빈도에 상관없이 기대보다 공기빈도가 높은 경우 공기성이 높은 것으로 파악하는 기본적 방법에 위배된다. t-점수를 공기성의 확률로 변환하여 제시한 Manning and Schütze(1999)의 다음 식에 대해서도 마찬가지로 말할 수 있다. 아래 식에서 N의 코퍼스의 크기이고, f(x), f(y)는 x, y의 빈도, f(x,y)는 x와 y의 공기빈도이다.[20]

$$(7)\ t = \frac{\frac{f(x,y)}{N} - \frac{f(x)}{N} \times \frac{f(y)}{N}}{\sqrt{\frac{\frac{f(x,y)}{N}}{N}}} = \frac{p(x,y) - p(x) \times p(y)}{\sqrt{\frac{p(x,y)}{N}}}$$

이상에서 t-점수를 정교화하기 위한 방법 그리고 t-점수가 아닌 상호정보를 사용하는 방식을 고려해 보았다. 결과적으로 상호정보는 작은 빈도의 단어의 공기성을 과장하는 문제가 있음이 드러났고 아울러 공기어의 방향성을 무시한다는 점에서도 문제가 있다. t-점수를 계산할 때 문단 내 중심어의 빈도를 고려하지 않는 것이 고려하는 것보다 단순하면서도 더 적절한 결과를 내는 것으로 나타났다.

6. 코퍼스 기반의 문화 연구를 위하여

관련어들을 기반으로 구성한 네트워크는 연관된 단어들의 네트워크인 동시에 그 단어들이 지시하는 사물이나 그 단어들의 개념들의 연관성 네트워크이다. 그리고 그 네트워크는 코퍼스, 즉 산출된 언어 자료로부터 추출된다. 말하자면 언어자료를 기반으로 생활양식(문화)과 정신의 네트워크를 구성하

20) 이와 같은 식은 임근석(2009)과 같이 언어 요소들이 연속하여 나타날 때의 연어성 계산에만 사용될 수 있다.

는 셈이다. 이것은 어떻게 보면 당연하다. 정신과 생활양식이 언어 사용을 결정하며 따라서 언어는 정신과 생활양식의 반영이다. 즉, 언어의 네트워크는 정신과 문화의 네트워크이다. 본 연구의 관점은 언어, 정신, 문화의 영역을 네트워크로 구성함으로써 언어와 정신과 문화를 네트워크 과학의 영역으로 들여왔다는 의의가 있다(참고: 바라바시 2002, 김용학 2007, Nooy et al. 2005 등). 이것은 언어를 복잡계(complex system)로 파악하려는 Bybee, Croft 등의 언어학적 관점과도 일치한다(Bybee and Hopper 2001, Bybee 2006, Beckner, et al. 2009 등). 복잡계(complex system)는 수많은 요소들이 복잡한 관계로써 연관되고, 독립적으로 외부에 적응하는 체계를 말한다(윤영수, 채승병 2005). 언어가 복잡계인 이유는 첫째, 언어 속의 수많은 언어 요소들이 복잡한 상호관계에 의해 연관되어 있고; 둘째, 언어 사용자들이 언어를 사용하면서 서로 복잡한 상호작용을 하기 때문이다. 여기서 제시한 관련어와 관련개념의 네트워크는 첫 번째 의미로서의 언어 복잡계의 일부이다. 참고로 네트워크는 복잡계 과학의 가장 중요한 연구 분야들 중 하나이다.

본 연구에서는 세종말뭉치를 기반으로 하여, 언어 자료를 기반으로 언어, 정신, 문화의 네트워크를 구성하는 방법을 제시하였다. 이 코퍼스가 1990년대와 2000년대 초의, 균형을 고려한 여러 장르의 언어자료로 이루어졌음을 상기할 때 세종말뭉치에 기초한 네트워크는 이 시기의 한국인, 한국 사회, 한국 문화의 네트워크이다. 물론 사회와 문화는 변한다. 그렇다면 다른 시기의 언어 자료는 다른 언어의 네트워크, 정신과 문화의 네트워크를 보여줄 것이며 나아가 언어와 정신과 문화의 변화를 보여줄 것이다. 이러한 목적을 위해서 신문 텍스트를 연구 자료로 활용하는 연구를 진행 중에 있다(김홍규, 강범모 외 2010).[21]

21) 한 심사자가 언급한 대로, 본 연구에서 이용한 자료는 동음이의어가 구분된 것으로서 공기어 추출의 과정 및 단어-개념 대응성 주장의 바탕이 되었다. 그러나 현재 동음이의어 구분을 자동적으로 하기 어려운 상황을 고려하면, 임의의 텍스트에 대하여 이 방법이 어려움이 있을 수 있다. 그러나 강범모(2005)가 밝힌 대로 사전에 동음어가 많이 등재되어 있다고 하더라도 실제 사용의 약 98%는 한 형식을 한 뜻으로만 사용함을 고려할 때, 그 어려움이 그렇게 크지 않을 수도 있다. 물론 '배', '상', '신부' 같이 두 가지 이상의 뜻으로 많이 사용되는 것들은 여전히 문제일 것이다.

참고문헌

강범모 (2003), 언어, 컴퓨터, 코퍼스언어학, 서울: 고려대학교 출판부.
강범모 (2005), “동음이의어의 사용 양상,” 어학연구 41-1, 1 - 29.
강범모, 김흥규 (2009), 한국어 사용 빈도, 서울: 한국문화사.
김용학 (2007), 사회 연결망 분석, 개정판. 서울: 박영사.
김진해 (2000), 국어 연어 연구, 경희대학교 국어학 박사학위 논문.
김흥규 외 (1998~2007), 21세기세종계획 기초자료구축 연구보고서, 문화관광부.
김흥규, 강범모 외 (2010) 물결 21: 신문 텍스트 기반의 장기간 언어·사회·문화 연구, 제1차 보고서, 서울: 고려대학교 민족문화연구원.
바라바시, A. L. (2002), 링크, 서울: 동아시아.
박병선 (2005), 한국어 계량적 연구 방법론, 서울: 역락.
소쉬르, 페르디낭 드 (1990 [1916]), 일반언어학 강의, 샤를 바이, 알레르 세슈에 역음, 최승언 옮김, 서울: 민음사.
윤영수, 채승병 (2005), 복잡계 개론: 세상을 움직이는 숨겨진 질서 찾기, 서울: 삼성경제연구소.
이기황 (1998), “말뭉치에 나타난 반대말의 쓰임새,” 서상규 편, 언어정보의 탐구 1, 서울: 월인.
이기황, 이재윤 (2008), “한국어 사전 어휘의 네트워크 분석,” 제3회 복잡계 컨퍼런스: 복잡계와 인문학/사회학의 만남.
임근석 (2009), “통계적 방법을 이용한 연어 후보 추출,” 한국어학 45, 305 - 333.
조은영 (2010), “어휘적 연어의 형성과 유추,” 제54차 한국어학회 전국학술대회 논문집, 67 - 87.
홍재성 외 (1998~2007), 21세기세종계획 전자사전구축 연구보고서, 문화관광부.
홍정하, 김주영, 강범모 (2008), “세종 구문분석 말뭉치의 구축과 통사적 범주 및 기능의 분포,” 민족문화연구 49집, 285 - 331.
홍종선, 강범모, 최호철 (2001), 한국어 연어 관계 연구, 서울: 월인.
Barnbrook, G. (1996), Language and Computers: A Practical Introduction to the Computer Analysis of Language, Edinburgh: Edindurgh University Press.
Batagelj, Vladimir and Andrej Mrvar (2010), Pajek: Program for Large Network Analysis, version 1.26. Homepage: http://vlado.fmf.uni-lj.si/pub/networks/pajek/.
Beckner, C., R. Blythe, J. Bybee, M. Christensen, W. Croft, N. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, T. Schoenemann (2009), “Language is a Complex Adaptive System: Position Paper,” in Language Learning 59: Suppl 1, 1 - 26.
Bybee, Jean (2006), “From Usage to Grammar: The Mind’s Response to

Repetition," in Language 82-4.
Bybee, Joan and Paul Hopper (2001), "Introduction to Frequency and the Emergence of Linguistic Structure," in J. Bybee and P. Hopper (eds), Frequency and the Emergence of Linguistic Structure, Amsterdam: John Benjamins Publishing Co.
Cann, Ronnie, Ruth Kempson, and Eleni Gregoromichelaki (2009), Semantics: An Introduction to Meaning in Language, Cambridge: Cambridge University Press.
Church, K., W. Gale, P. Hanks, and D. Hindle (1991), "Using Statistics in Lexical Analysis", in U. Zernik (ed.), Lexcial Acquisition: Exploiting on-line resources to build a lexicon. Hilldale: Lawrence Erlbaum, 115 - 164.
Croft, William and D. Alan Cruse (2004), Cognitive Linguistics, Cambridge: Cambridge University Press.
Cruse, D.A. (1986), Lexical Semantics, Cambridge University Press, Cambridge.
Cruse, D.A. (2000), Meaning in Language, Oxford: Oxford University Press.
Fellbaum, Christiane (ed.)(1998), WordNet: An Electronic Lexical Database, Cambridge, Mass: The MIT Press.
Firth, J.R. (1957) Papers in Linguistics, London: Oxford University Press.
Fillmore, Charles J. and Beryl T. Atkins (1992), "Towards a Frame-Based Lexicon: The Semantics of RISK and its Neighbors", in Lehrer and Kittay (eds.), 75 - 102.
Lyons, John (1977), Semantics 1, 2, Cambridge: Cambridge University Press.
Manning, C. and H. Schütze (1999), Foundations of Statistical Natural Language Processing, Cambridge, Mass: The MIT Press.
Miller, George (1991), The Science of Words, Scientific American.
Montague, Richard (1974), "The Proper Treatment of Quantification in Ordinary English", in R. Thomason (ed.), Formal Philosophy, New Haven, Yale University Press.
Nooy, Wouter de, Andrej Mrvar, and Vladimir Batagelj (2005), Exploratory Social Network Analysis with Pajek, Cambridge: Cambridge University Press.
Ogden, C.K. and I.A. Richards (1923), The Meaning of Meaning, London: Routledge and Kegan Paul.
Portner, Paul and Barbara H. Partee (eds.), (2002), Formal Semantics: The Essential Readings. Oxford: Blackwell Publishing.
Stubbs, Michael (1995), "Collocations and Semantic Prosodies: On the cause of the trouble with quantitative studies," in Foundations of Language 2-1, 23 - 55.
Teubert, Wolfgan and Ramesh Krishnamurthy (2007), "General Introduction," in W.

Teubert and R. Krishnamurthy (eds.) Corpus Linguistics: Critical Concepts in Linguistics, London: Routledge, 1 – 37.

서울시 성북구 고려대학교 안암동
고려대학교 문과대학 언어학과
136-701
전화번호: 3290-2173
전자우편: bmkang@korea.ac.kr

원고 접수일: 2010년 7월 15일
원고 수정일: 2010년 8월 9일
게재 결정일: 2010년 8월 23일