응용통계연구 제 18권 3호, 2005년, pp. 583-595

언어의 공기관계 분석을 위한 임의화검증의 응용

양경숙[1] 김희영[2]

요 약

언어의 공기관계를 파악하는 데는 여러 가지 연관성 통계량들이 이용된다. 그러나 일부 통계량을 제외한 나머지 통계량들은 분포가 알려져 있지 않아 정작 통계량 값을 구하고도 명확한 설명을 하지 못하는 경우가 있다. 따라서 언어의 공기관계 분석을 위해서 정규근사나 t통계량을 이용하여 가설검증을 하는 경우가 많다. 그러나 공기관계에 있는 어휘빈도가 전체 빈도에서 차지하는 백분율이 매우 작기 때문에 정규근사에는 무리가 있어 보인다. 따라서 본 논문은 여러 논문에서 자주 언급되는 연관성 통계량의 특성을 임의화검증(randomization test)을 통해 고찰함으로써 계량언어학의 연어분석에서 데이터의 특성을 고려하여 보다 정확하게 언어의 공기관계를 이해할 수 있도록 도모하고자 한다.

주요용어: 공기관계, 연관성, 카이제곱 통계량, 상호정보, 연어(連語).

1. 서론

언어연구는 관심을 갖는 언어 표현의 적형(well-formed)여부에 주로 관심을 가져왔다고 할 수 있다. 이런 연구는 객관적인 자료를 기반으로 하기보다는 주로 모국어 화자의 직관에 의존하는 경향이 있다. 예를 들어 a) 차가운 공기/차가운 라디오, b) 찬 바람/차가운 바람, c) 찬 현실/차가운 현실에서 '차가운 공기', '차가운 바람', '차가운 현실'은 자연스런 문구이다. 그러나 '차가운 라디오'는 어색한 표현임을 쉽게 알 수 있다. 여기서 '차갑다'라는 형용사가 수식할 수 있는 대상의 전체적인 양상을 직관적으로 정리하기는 어려우며 '찬'과 '차가운' 중 어느 것이 더 많이 사용되는지 또 피수식어 공기관계 어휘들의 분포상의 차이 등을 직관적으로 전개하기는 쉽지 않다.

이처럼 한 문장 안에서 어떤 단어(중심어)와 함께 출현하는 단어(연결어) 사이의 여러 언어학적 관계를 언어학 분야에서는 공기관계(co-occurrence)라고 부르며 이와 같은 언어연구를 위해 어휘 말미에 형태소를 붙여 축적한 대규모 텍스트 데이터를 말뭉치(코퍼스: corpus)라고 한다. 통상적으로 사용되는 대규모 말뭉치는 1000만 어절 이상을 축적하고 있고 모집단 대표성을 위해 1억 어절의 데이터를 다루는 경우도 있다. 따라서 계량언어학 연구에서는 대규모의 자료를 다루기 위해 전산처리와 통계이용이 필수적이다.

1) (136-701) 서울특별시 성북구 안암동 5가, BK21 한국학 교육·연구단 박사후 연구원
E-mail: myksyang@naver.com
2) (136-701) 서울특별시 성북구 안암동 5가, 고려대학교 통계연구소 연구조교수
E-mail: starkim@korea.ac.kr

외국의 연구에서는 공기관계의 연구에 있어서 말뭉치 자료를 이용하여 계량적으로 접근하려는 노력이 비교적 활발함을 알 수 있다. 특별히 유의미한 공기관계를 추출하기 위해서 적절한 통계량의 연구가 필수적인데 바로 이 분야에 통계학의 역할이 필요하다.

용어 연관성을 측정하기 위해 공기 빈도를 이용하는 경우 연관계수(association measure), 또는 유사계수(similarity measure)라고 불리는 통계량을 이용하여 평가한다. 쿨진스키 계수, 오키아이 계수, 페이거 맥고언의 계수, 율의 계수, 상호정보가 그 예이다 (Scott Songlin Piao, 2002; Michael P. Oakes,1998). 그런데 언어 특성들 사이의 공기관계를 측정하는데 사용되는 이들 통계량들은 일반적으로 분포가 알려져 있지 않고 행태적 특징 또한 구체적으로 정립되어 있지 않다. 물론 연관성을 측정하기 위해서는 분석단위나 공기범위 등을 사전에 결정해야 한다. 여기서 분석단위라 함은 빈도 측정 대상을 가리키며 공기범위란 두 분석단위가 공기했다고 판정할 범위를 말한다. 연어분석과 같은 경우 주로 인접 단어를 범위로 하며 다른 경우에는 3 단어 이내, 5 단어 이내와 같이 일정한 크기의 문맥창(context window)을 범위로 삼거나 동일 문장, 동일 문헌 등을 이용하게 된다 (홍종선, 강범모, 최호철, 2001).

그런데 언어의 특성상 문맥창의 범위를 변화시켜도 공기하는 빈도는 전체 관측빈도에 비해 상대적으로 매우 작게 관측된다. 따라서 데이터의 특성상 정규분포 가정아래 연관성을 검증하는 것은 잘못된 결과를 초래할 수 있다.

본 연구의 목적은 국내 통계학의 응용분야로서 비교적 생소한 계량언어학에서의 공기관계와 두 단어의 공기관계 파악을 위한 몇 가지 통계량들을 고찰하고 이들 통계량 중 비교적 많이 사용되는 통계량의 분포적 행태를 파악하고자 한다. 따라서 2절에서는 공기관계의 연관성 검정에 주로 사용되는 쿨진스키 계수, 오키아이 계수, 페이거 맥고언의 계수, 율의계수, 상호정보, 로그 유사도 계수 등을 고찰해보고 3절에서는 임의화검증을 통해 2절에서 언급한 통계량 중 그 쓰임새가 많은 율의 계수와 상호정보 통계량의 유의확률을 다양한 조합에 대하여 비교한 결과를 살펴볼 것이다.

2. 공기관계와 연관성 통계량

일반적으로 한 문장을 단위로 할 때 언어의 공기관계 파악을 위해 사용되는 데이터 형태는 아래의 표 2.1과 같다. 이때 공기어의 위치는 연구자의 관심사에 따라 조금씩 달라질 수 있다.

표 2.1: 공기관계 파악을 위한 데이터 형태

		공기어(Y)		
		O	X	합계
중심어 (X)	O	a	b	a+b
	X	c	d	c+d
합계		a+c	b+d	N

여기서 각 칸의 빈도는 한 문장을 단위로 다음을 나타낸다.(N=a+b+c+d)

a : 중심어와 공기어(연결어) 두 단어가 함께 단어 쌍으로 나타나는 빈도

b : 중심어는 포함되고 공기어가 나타나지 않는 빈도

c : 중심어는 포함되지 않고 공기어만 나타난 빈도

d : 중심어, 공기어 모두 포함되지 않는 빈도

통상적으로 언어의 공기관계 파악에서 중심어와 공기어가 함께 나타나는 빈도 a는 나머지 칸의 빈도에 비해 상대적으로 매우 작게 출현한다. 그리고 이들 중심어와 공기어간의 연관성을 파악하는데 사용되는 통계량들은 쿨진스키 계수(Kulczinsky coefficient: KUC), 오키아이 계수(Ochiai coefficient: OCH), 페이거 맥고언의 계수(Fager and McGowan coefficient : MAG), 율의 계수(Yule coefficient), 상호정보(mutual information : MI), 로그 유사도 계수(log-likelihood coefficient : LL) 등이 있다. 이들 통계량의 특징은 다음과 같다 (Rodham, E. Tulloss, 1997).

1) 쿨진스키 계수(Kulczinsky coefficient : KUC)

$$\text{KUC} = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right), \qquad 0 \le \text{KUC} \le 1 \tag{2.1}$$

이 계수는 a,b,c가 고정된 상태에서 d의 빈도를 아무리 증가시켜도 동일한 계수값을 갖게 된다. 표 2.2와 같이 중심어의 빈도는 3이고 공기어의 빈도가 28일 경우, 두 케이스가 비슷한 패턴을 보이지만 좌측 테이블에 대한 KUC는 0.35이고 가운데 테이블에 대한 KUC는 0.55로 계산된다. 그런데 표 2.2의 우측 테이블과 같이 중심어와 공기어의 출현빈도가 동일한 경우에도 KUC 통계량은 0.55로 계산되어 두 항목간의 연관성을 살펴보기 위한 통계량으로 문제점을 내포하고 있다.

표 2.2: 쿨진키스 계수 (예)

		(Y)		
		O	X	합계
(X)	O	2	1	3
	X	26	d	26+d
합계		28	1+d	N

		(Y)		
		O	X	합계
(X)	O	3	0	3
	X	25	d	25+d
합계		28	0+d	N

		(Y)		
		O	X	합계
(X)	O	55	45	100
	X	45	d	45+d
합계		100	45+d	N

2) 오키아이 계수(Ochiai coefficient : OCH)

$$\text{OCH} = \frac{a}{\sqrt{(a+b)(b+c)}}, \qquad 0 \le \text{OCH} \le 1 \tag{2.2}$$

오키아이 계수는 <a, b, c>의 구성이 <1, 0, n-1>과 같이 구성될 경우 $\frac{1}{\sqrt{n}}(n \ge 1)$로 계산된다. 구체적으로 <1,0,1>의 테이블에 대해서는 OCH=0.71이고 <1,0,69>의 경우 OCH=0.12로 빈도가 단 한번 나타남에도 바람직하지 않게 큰 수치를 보인다.

표 2.3: 오키아이 계수 (예)

		공기어		
		O	X	합계
중심어	O	1	0	1
	X	1	d	1+d
합계		2	d	N

		공기어		
		O	X	합계
중심어	O	1	0	1
	X	69	d	69+d
합계		70	d	N

Rodham(1997)에 의하면 <a,b,c>의 빈도가 <n, 0, 8n> <n, 2n, 2n>, <24, 3, 168>, <24, 24, 84>, <24, 30, 72> <24, 40, 57>의 경우, 모두 OCH=0.33으로 계산된다고 보고하고 있다. 또한 b와 c의 빈도 차가 비교적 비슷할 경우에는 쿨진스키 계수값과 비슷하게 계산되지만 차이가 크게 날 경우에는 오키아이 계수는 쿨진스키 계수값과 차이가 크게 벌어진다.

3) 페이거 맥고언 계수(Fager and McGowan coefficient : FAG)

페이거 맥고언의 상관계수도 d가 커지는 것과 무관되게 고정된 <a,b,c> 테이블에 대해서는 동일한 계수값으로 계산되는 문제점을 내포하고 있다. <a,b,c>의 빈도가 작을수록 FAG계수값은 음수를 나타내며 클수록 양수로 계산되나 1을 넘지는 않는다.

$$\text{FAG} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{\sqrt{a+b}}, \qquad -\infty \le \text{FAG} \le 1 \tag{2.3}$$

4) 율 계수(Yule coefficient)

이 통계량은 표본크기나 특정 칸 빈도가 작을 경우 이에 의한 영향을 받지 않는 것으로 알려져 있다. 두 항목이 서로 독립이면 율의 계수값은 0으로 계산되며 양의 방향으로 상관되어 있으면 +1, 음의 방향으로 상관되어 있으면 -1의 값으로 계산된다.

$$\text{YUL} = \frac{ad - bc}{ad + bc}, \qquad -1 \le \text{YUL} \le 1 \tag{2.4}$$

5) 상호정보(Mutual information : MI)

상호정보는 특성상 두 단어가 공기하는 빈도의 중요성 뿐 아니라, 한 문장에서 두 단어가 동시에 나타나지 않는 빈도인 d 값에도 동등한 중요성을 부여해서 저빈도 공기어의 경우 상호정보 값이 크게 확대되어 나타나는 경향을 보인다. 따라서 상호정보를 적용할 때 해당 연결어가 전체 자료에서 저빈도이고 대부분 중심어와 공기하는 단어라면 즉, a값이 작고 c값도 작은 경우, 결과가 왜곡되어 나탄다는 점을 주의해야 한다 (박병선, 2003).

$$\text{MI} = \log_2 \frac{aN}{(a+b)(a+c)}, \qquad -\infty \le \text{MI} \le \infty \tag{2.5}$$

이 외에도 로그 가능도 계수(log-likelihood coefficient : LL), 파이제곱 계수(Phi-square coefficient), 단순일치 계수(simple matching coefficient: SMC) 등이 연구되었으나 잘 사용

되고 있지 않다. 또한 쿨진스키상관계수, 오키아이 계수는 위에서 언급한바와 같이 공기관계의 유의미성을 검증하는 통계량으로는 문제를 안고 있어 실제로는 많이 사용되지 않는다. 따라서 본 연구에서는 언어의 연관성 분석에서 많이 사용되는 율계수, 상호정보 통계량 등을 중심으로 임의화검증을 하여 그 특성을 살펴보기로 한다.

3. 임의화검증(randomization test)

2절에서 설명한 여러 개의 연관성 통계량에 대해서 다음의 가설을 고려해보자.

귀무가설 : 중심어와 공기어는 서로 독립이다.

대립가설 : 중심어와 공기어는 서로 독립이 아니다.

이 가설을 검증하는데 많이 사용되는 통계량은 피셔의 정확검증과 카이제곱 검증이다.

행간 동질성 검증에서 사용되는 피셔의 정확검증(Exact test)은 행합과 열합이 고정된 상태에서의 조건부 확률에 대한 유의확률을 계산함으로써 적용된다. 즉 정해진 행합과 열합을 갖는 모든 임의의 분할표에 대해서 아래와 같은 조건부 확률을 계산하여 관측된 2원 분할표의 p값을 산출하는 방법이다 (허명회, 1997).

$$p(a,b,c,d|a+b,a+c,b+c,b+d) = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{N}{a+b}} \tag{3.1}$$

그런데 열합과 행합의 빈도가 커질수록 정확검증을 위해 작성되는 이원분할표의 개수는 늘어난다. 따라서 2절에서 언급한 여러 연관성 통계량 중 널리 이용되는 율의 계수, 상호정보와 카이제곱 통계량을 순열 모의실험을 통해 검증하고자 한다.

임의화검증에 의해 각 통계량에 대한 유의확률은 다음과 같이 계산한다.

$$\#(|YUL_1| \geq |YUL_0|)/\text{총반복회수}$$
$$\#(|MI_1| \geq |MI_0|)/\text{총반복회수}$$
$$\#(|\chi_1| \geq |\chi_0|)/\text{총반복회수}$$

여기서 아래첨자 0으로 표시된 통계량은 기준이 되는 분할표로부터 계산되며 아래첨자 1로 표시된 통계량은 임의순열 모의실험으로 새롭게 구성된 분할표로부터 계산됨을 나타낸다. 또한 χ^2 은 다음 통계량을 나타낸다.

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}, \quad \text{여기서 } n_{ij} \text{ :관찰값, } e_{ij} \text{ :기대값}$$

이들 통계량과 함께 자유도 1의 카이제곱 분포로부터 계산된 유의확률을 함께 비교할 것이다.

칸 빈도를 여러 형태로 적용한 모의실험결과가 필요한데 여기서 임의화검증은 Rodham(1997)이 실험한 형태의 데이터에 우선적으로 적용하였다. 표 3.1은 a, b, c는 1로 고정시키고 d를 점차 증가시키면서 유의확률을 계산한 결과이다. 매번의 분할표에 대해서는 1000번씩 임의화검증을 하였다. 표 3.1에서 p(1)은 자유도 1의 카이제곱 분포로부터 계산

된 유의확률을 나타낸다. chi_p , $yule_p$, MI_p는 모두 각 통계량을 이용하여 순열 모의실험으로부터 계산된 유의확률을 나타낸다. 마지막 3개의 열들은 각 유의확률에 대한 표준오차를 나타낸다.

표 3.1의 결과로부터 카이제곱분포표로부터 계산된 유의확률과 임의화검증을 통해 계산된 유의확률간에 차이가 매우 크다는 것을 알 수 있다.

표 3.1: <1,1,1>의 테이블에 적용한 임의화검증

num	a	b	c	d	p(1)	chi_p	$yule_p$	MI_p	chi_{se}	$yule_{se}$	MI_{se}
1	1	1	1	0	0.8865	1.0000	1.0000	0.6670	0.0000	0.0000	0.0149
2	1	1	1	1	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
3	1	1	1	2	0.7094	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
4	1	1	1	3	0.5403	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
5	1	1	1	4	0.4274	1.0000	0.5250	1.0000	0.0000	0.0158	0.0000
6	1	1	1	5	0.3458	1.0000	0.4540	0.4540	0.0000	0.0157	0.0157
7	1	1	1	6	0.2840	0.4260	0.4260	0.4260	0.0156	0.0156	0.0156
8	1	1	1	7	0.2357	0.3670	0.3670	0.3670	0.0152	0.0152	0.0152
9	1	1	1	8	0.1971	0.3590	0.3590	0.3590	0.0152	0.0152	0.0152
10	1	1	1	9	0.1659	0.2990	0.2990	0.2990	0.0145	0.0145	0.0145
11	1	1	1	10	0.1402	0.3180	0.3180	0.3180	0.0147	0.0147	0.0147
12	1	1	1	11	0.1190	0.2570	0.2570	0.2570	0.0138	0.0138	0.0138
13	1	1	1	12	0.1013	0.2500	0.2500	0.2500	0.0137	0.0137	0.0137
14	1	1	1	13	0.0865	0.2370	0.2370	0.2370	0.0134	0.0134	0.0134
15	1	1	1	14	0.0740	0.2240	0.2240	0.2240	0.0132	0.0132	0.0132
16	1	1	1	15	0.0634	0.2390	0.2390	0.2390	0.0135	0.0135	0.0135
17	1	1	1	16	0.0545	0.1850	0.1850	0.1850	0.0123	0.0123	0.0123
18	1	1	1	17	0.0469	0.2060	0.2060	0.2060	0.0128	0.0128	0.0128
19	1	1	1	18	0.0404	0.1970	0.1970	0.1970	0.0126	0.0126	0.0126
20	1	1	1	19	0.0348	0.1680	0.1680	0.1680	0.0118	0.0118	0.0118
21	1	1	1	20	0.0300	0.1770	0.1770	0.1770	0.0121	0.0121	0.0121
22	1	1	1	21	0.0260	0.1400	0.1400	0.1400	0.0110	0.0110	0.0110
23	1	1	1	22	0.0225	0.1410	0.1410	0.1410	0.0110	0.0110	0.0110
24	1	1	1	23	0.0194	0.1680	0.1680	0.1680	0.0118	0.0118	0.0118
25	1	1	1	24	0.0168	0.1410	0.1410	0.1410	0.0110	0.0110	0.0110
26	1	1	1	25	0.0146	0.1440	0.1440	0.1440	0.0111	0.0111	0.0111
27	1	1	1	26	0.0127	0.1290	0.1290	0.1290	0.0106	0.0106	0.0106
28	1	1	1	27	0.0110	0.1350	0.1350	0.1350	0.0108	0.0108	0.0108

그림 3.1은 칸빈도 <a,b,c>를 동일하게 고정시키고 d를 1씩 증가시키면서 유의확률이 어떻게 달라지는지 순열조합 모의실험한 결과를 플롯으로 나타낸 것이다.

표 3.1:<1,1,1>의 테이블에 적용한 임의화검증 (계속)

num	a	b	c	d	p(1)	chi_p	$yule_p$	MI_p	chi_{se}	$yule_{se}$	MI_{se}
29	1	1	1	28	0.0095	0.1430	0.1430	0.1430	0.0111	0.0111	0.0111
30	1	1	1	29	0.0083	0.1390	0.1390	0.1390	0.0109	0.0109	0.0109
31	1	1	1	30	0.0072	0.1340	0.1340	0.1340	0.0108	0.0108	0.0108
32	1	1	1	31	0.0063	0.1190	0.1190	0.1190	0.0102	0.0102	0.0102
33	1	1	1	32	0.0055	0.1240	0.1240	0.1240	0.0104	0.0104	0.0104
34	1	1	1	33	0.0047	0.1030	0.1030	0.1030	0.0096	0.0096	0.0096
35	1	1	1	34	0.0041	0.1200	0.1200	0.1200	0.0103	0.0103	0.0103
36	1	1	1	35	0.0036	0.1130	0.1130	0.1130	0.0100	0.0100	0.0100
37	1	1	1	36	0.0031	0.0770	0.0770	0.0770	0.0084	0.0084	0.0084
38	1	1	1	37	0.0027	0.0880	0.0880	0.0880	0.0090	0.0090	0.0090
39	1	1	1	38	0.0024	0.1020	0.1020	0.1020	0.0096	0.0096	0.0096
40	1	1	1	39	0.0021	0.0940	0.0940	0.0940	0.0092	0.0092	0.0092
41	1	1	1	40	0.0018	0.0800	0.0800	0.0800	0.0086	0.0086	0.0086
42	1	1	1	41	0.0016	0.0930	0.0930	0.0930	0.0092	0.0092	0.0092
43	1	1	1	42	0.0014	0.0720	0.0720	0.0720	0.0082	0.0082	0.0082
44	1	1	1	43	0.0012	0.0950	0.0950	0.0950	0.0093	0.0093	0.0093
45	1	1	1	44	0.0011	0.0850	0.0850	0.0850	0.0088	0.0088	0.0088
46	1	1	1	45	0.0009	0.0650	0.0650	0.0650	0.0078	0.0078	0.0078
47	1	1	1	46	0.0008	0.0890	0.0890	0.0890	0.0090	0.0090	0.0090
48	1	1	1	47	0.0007	0.0840	0.0840	0.0840	0.0088	0.0088	0.0088
49	1	1	1	48	0.0006	0.0800	0.0800	0.0800	0.0086	0.0086	0.0086
50	1	1	1	49	0.0005	0.0670	0.0670	0.0670	0.0079	0.0079	0.0079
51	1	1	1	50	0.0005	0.0700	0.0700	0.0700	0.0081	0.0081	0.0081
52	1	1	1	51	0.0004	0.0890	0.0890	0.0890	0.0090	0.0090	0.0090
53	1	1	1	52	0.0004	0.0680	0.0680	0.0680	0.0080	0.0080	0.0080
54	1	1	1	53	0.0003	0.0710	0.0710	0.0710	0.0081	0.0081	0.0081
55	1	1	1	54	0.0003	0.0780	0.0780	0.0780	0.0085	0.0085	0.0085
56	1	1	1	55	0.0002	0.0730	0.0730	0.0730	0.0082	0.0082	0.0082
57	1	1	1	56	0.0002	0.0760	0.0760	0.0760	0.0084	0.0084	0.0084
58	1	1	1	57	0.0002	0.0760	0.0760	0.0760	0.0084	0.0084	0.0084
59	1	1	1	58	0.0002	0.0700	0.0700	0.0700	0.0081	0.0081	0.0081
60	1	1	1	59	0.0001	0.0570	0.0570	0.0570	0.0073	0.0073	0.0073
61	1	1	1	60	0.0001	0.0600	0.0600	0.0600	0.0075	0.0075	0.0075
62	1	1	1	61	0.0001	0.0620	0.0620	0.0620	0.0076	0.0076	0.0076
63	1	1	1	62	0.0001	0.0560	0.0560	0.0560	0.0073	0.0073	0.0073
64	1	1	1	63	0.0001	0.0660	0.0660	0.0660	0.0079	0.0079	0.0079
65	1	1	1	64	0.0001	0.0540	0.0540	0.0540	0.0071	0.0071	0.0071

표 3.1:<1,1,1>의 테이블에 적용한 임의화검증 (계속)

num	a	b	c	d	p(1)	chi_p	$yule_p$	MI_p	chi_{se}	$yule_{se}$	MI_{se}
66	1	1	1	65	0.0001	0.0550	0.0550	0.0550	0.0072	0.0072	0.0072
67	1	1	1	66	0.0001	0.0680	0.0680	0.0680	0.0080	0.0080	0.0080
68	1	1	1	67	0.0000	0.0710	0.0710	0.0710	0.0081	0.0081	0.0081
69	1	1	1	68	0.0000	0.0600	0.0600	0.0600	0.0075	0.0075	0.0075
70	1	1	1	69	0.0000	0.0540	0.0540	0.0540	0.0071	0.0071	0.0071
71	1	1	1	70	0.0000	0.0700	0.0700	0.0700	0.0081	0.0081	0.0081
72	1	1	1	71	0.0000	0.0490	0.0490	0.0490	0.0068	0.0068	0.0068
73	1	1	1	72	0.0000	0.0490	0.0490	0.0490	0.0068	0.0068	0.0068
74	1	1	1	73	0.0000	0.0480	0.0480	0.0480	0.0068	0.0068	0.0068
75	1	1	1	74	0.0000	0.0540	0.0540	0.0540	0.0071	0.0071	0.0071
76	1	1	1	75	0.0000	0.0480	0.0480	0.0480	0.0068	0.0068	0.0068
77	1	1	1	76	0.0000	0.0460	0.0460	0.0460	0.0066	0.0066	0.0066
78	1	1	1	77	0.0000	0.0520	0.0520	0.0520	0.0070	0.0070	0.0070
79	1	1	1	78	0.0000	0.0420	0.0420	0.0420	0.0063	0.0063	0.0063
80	1	1	1	79	0.0000	0.0390	0.0390	0.0390	0.0061	0.0061	0.0061
81	1	1	1	80	0.0000	0.0470	0.0470	0.0470	0.0067	0.0067	0.0067
82	1	1	1	81	0.0000	0.0420	0.0420	0.0420	0.0063	0.0063	0.0063
83	1	1	1	82	0.0000	0.0550	0.0550	0.0550	0.0072	0.0072	0.0072
84	1	1	1	83	0.0000	0.0520	0.0520	0.0520	0.0070	0.0070	0.0070
85	1	1	1	84	0.0000	0.0510	0.0510	0.0510	0.0070	0.0070	0.0070
86	1	1	1	85	0.0000	0.0330	0.0330	0.0330	0.0056	0.0056	0.0056
87	1	1	1	86	0.0000	0.0370	0.0370	0.0370	0.0060	0.0060	0.0060
88	1	1	1	87	0.0000	0.0410	0.0410	0.0410	0.0063	0.0063	0.0063
89	1	1	1	88	0.0000	0.0520	0.0520	0.0520	0.0070	0.0070	0.0070
90	1	1	1	89	0.0000	0.0400	0.0400	0.0400	0.0062	0.0062	0.0062
91	1	1	1	90	0.0000	0.0500	0.0500	0.0500	0.0069	0.0069	0.0069
92	1	1	1	91	0.0000	0.0510	0.0510	0.0510	0.0070	0.0070	0.0070
93	1	1	1	92	0.0000	0.0380	0.0380	0.0380	0.0060	0.0060	0.0060
94	1	1	1	93	0.0000	0.0460	0.0460	0.0460	0.0066	0.0066	0.0066
95	1	1	1	94	0.0000	0.0370	0.0370	0.0370	0.0060	0.0060	0.0060
96	1	1	1	95	0.0000	0.0250	0.0250	0.0250	0.0049	0.0049	0.0049
97	1	1	1	96	0.0000	0.0360	0.0360	0.0360	0.0059	0.0059	0.0059
98	1	1	1	97	0.0000	0.0370	0.0370	0.0370	0.0060	0.0060	0.0060
99	1	1	1	98	0.0000	0.0380	0.0380	0.0380	0.0060	0.0060	0.0060
100	1	1	1	99	0.0000	0.0420	0.0420	0.0420	0.0063	0.0063	0.0063

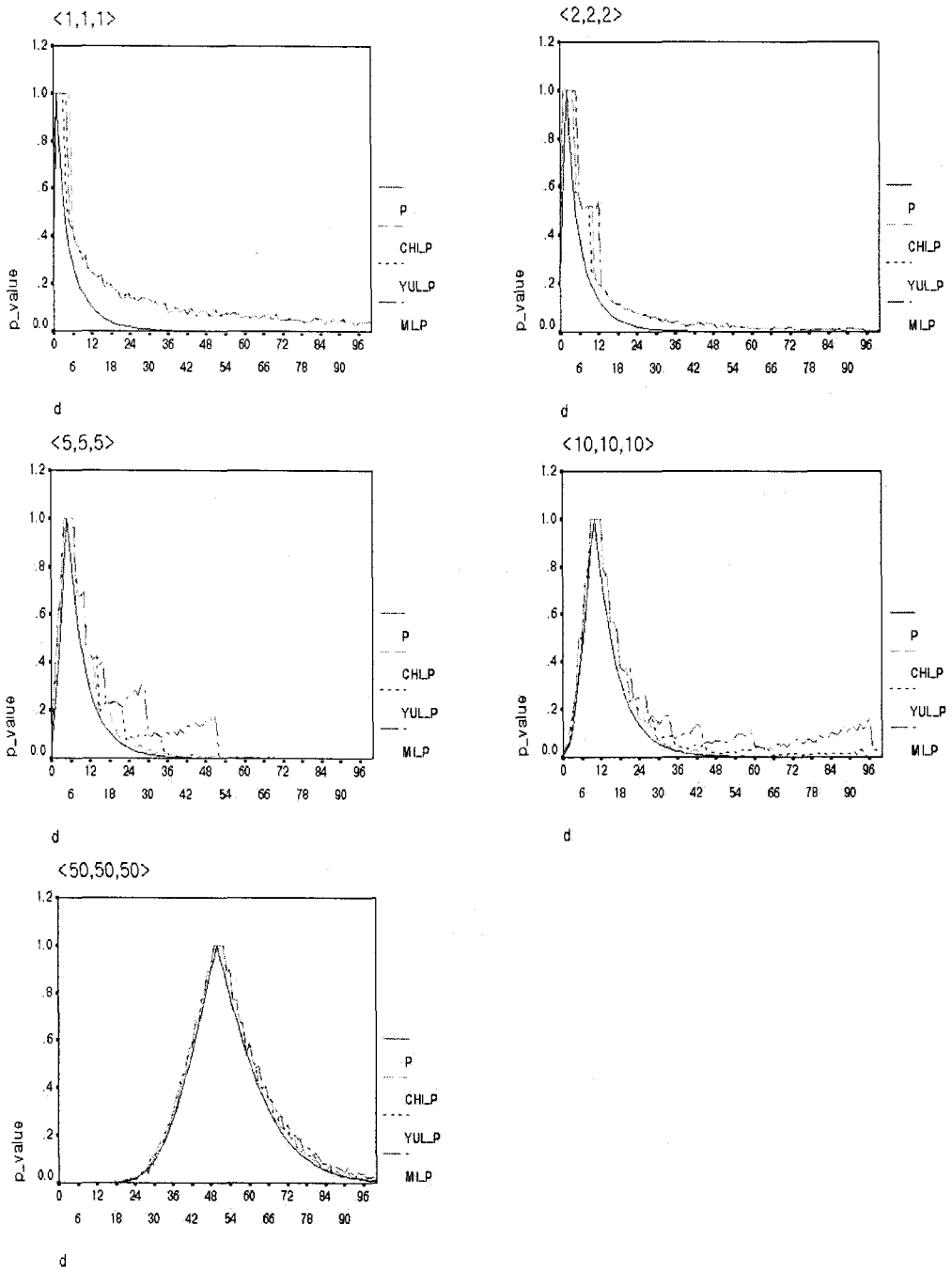

그림 3.1: <a,b,c>를 동일하게 고정시켰을 때의 유의확률의 변화

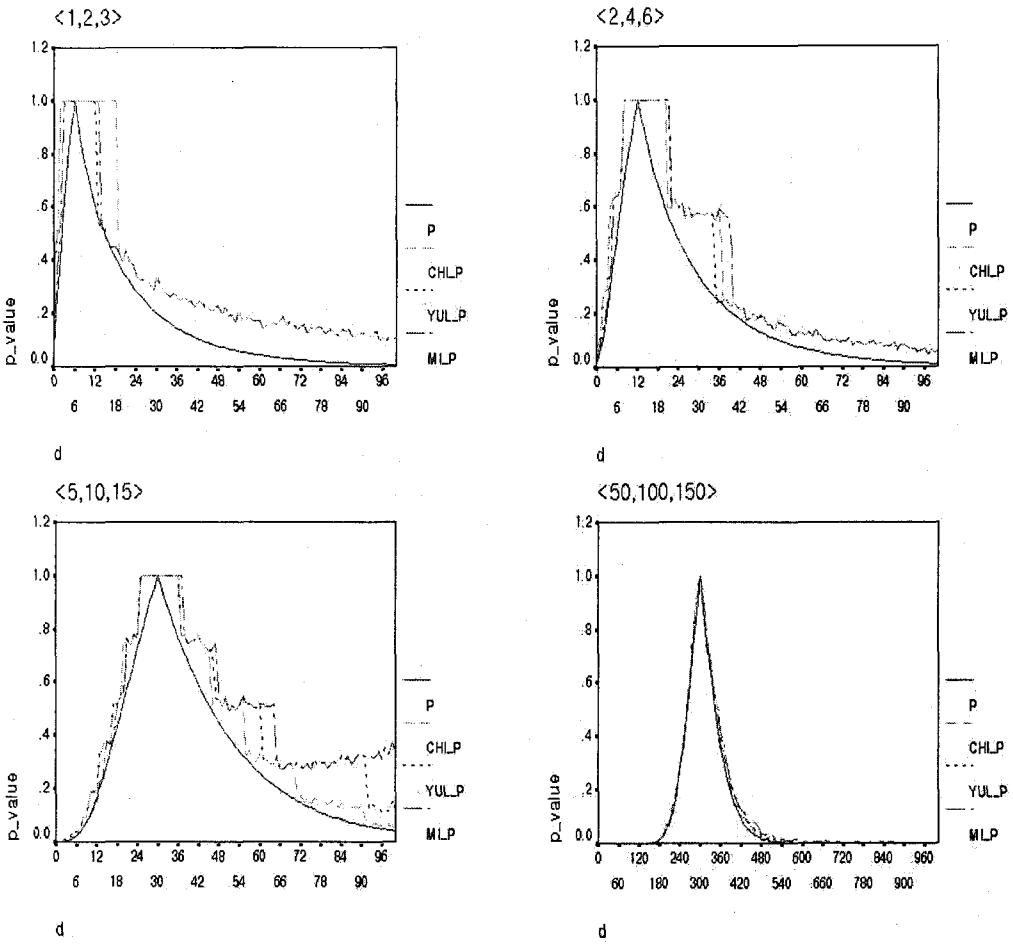

그림 3.2: <a,b,c>를 다르게 구성했을 때의 유의확률의 변화

그림 3.1은 d를 100까지 증가시키면서 살펴본 결과 a, b, c가 50정도 되어야 정규분포와 비슷한 패턴을 보여주며 비교하는 4가지 통계량에 대한 유의확률이 거의 비슷하게 됨을 알 수 있다. a, b, c가 상대적으로 작을 경우에는 d가 증가할수록 MI 통계량을 이용한 유의확률이 다른 통계량에 비해 상대적으로 다소 크게 계산됨을 살펴볼 수 있다.

그림 3.2는 a, b, c를 서로 다르게 고정시키고 d값을 증가시켰을 때 유의확률을 도식화한 결과이다. 그림 3.1에 비해 치우침 정도는 조금 덜 하지만 a, b, c값이 상대적으로 작을 경우 MI나 율의 계수를 이용한 유의확률이 더 높게 계산됨을 알 수 있다. 물론 전체 표본 크기 증가도 함께 고려하여 살펴보아야 할 것이다.

표 3.2는 실제로 한국어 공기어에 대해서 세종태거로부터 1000어절 기준으로 어휘빈도를 조사한 후 작성된 테이블이다. 이들 분할표에서도 알 수 있듯이 실제로 한국어의 공기

관계를 살펴볼 때 흔히 접하는 어휘들의 분포는 d 빈도가 상대적으로 매우 크고 a의 빈도는 상대적으로 매우 작은 경우가 빈번하다. 그런데 많은 텍스트 데이터를 다루는 연구자들은 보통 정규분포 가정 하에서 가설검증을 많이 하고 있다. 하지만 그림 3.1과 그림 3.2에서 보듯이 a의 빈도가 상대적으로 매우 작을 경우 정규분포 가정을 두는 것이 어렵다는 것은 자명한 일이다. 결국 정규분포 가정하에서 공기관계를 검증하는 것은 잘못된 결과를 초래하게 된다. 이 경우 유용한 것이 임의화검증을 통한 가설검증이다.

표 3.2: "수 있다"와 "그 얘기"에 대한 분할표

		있다		
		O	X	합계
수	O	2	11	13
	X	2	985	987
합계		4	996	1000

		얘기		
		O	X	합계
그	O	2	10	12
	X	7	981	988
합계		9	991	1000

표 3.2에서 '수 있다' 에 대해서 임의화검증을 통해 계산된 유의확률은 4가지 통계량 모두 0.000으로 계산되었고 '그 얘기'에 대한 유의확률은 0.007로 계산되어 두 어휘가 공기관계가 있음을 부정하기 어렵다는 결론에 도달하게 된다.

4. 맺음말

본 논문에서는 언어의 공기관계에 대한 유의성 검증통계량들을 소개하고 특징들을 비교하였다. 이들 통계량은 분포가 정확히 알려져 있지 않음에도 불구하고 보통 정규분포나 t분포로 근사시켜 유의성을 판단한다. 그러나 총빈도에 비해 특정 칸 빈도가 매우 작을 경우 이들 통계량의 정규근사에 무리가 있음을 짐작할 수 있다. 또한 2원 분할표에서 특정 칸의 기대빈도가 5 이하일 경우 카이제곱분포로 근사시키는 것도 무리가 있다. 따라서 본 논문에서는 계량언어학의 연관성 분석에서 많이 사용하는 율의 계수, 상호정보 통계량의 유의확률을 몇 가지 다양한 조합에 대하여 모의실험을 통해 살펴보고 그 결과를 제시하였다.

참고문헌

박병선 (2003). 국어 공기관계의 계량언어학적 연구, 고려대학교 대학원 박사학위논문.

허명회 (1997). 2원 분할표의 소표본 검증법, <응용통계연구>, **10**, 339-352.

홍종선, 강범모, 최호철 (2001). <한국어 연어관계 연구>, 서울: 월인.

Rodham, E. Tulloss (1997). Assessment of similarity indices for undesirable properties and a new tripartite similarity index bases on cost functions, *Mycology in Sustainable Development : Expanding Concepts, Vanishing Borders* (Palm, M.E and I. H. Chapela eds.), Parkway Publishers, Boone, North Carolina, 122-143.

Scott Songlin Piao (2002). Word alignment in English-Chinese parallel corpora, *Literary and Linguistic Computing*, **17**, 207-230.
Michael P. Oakes (1998). *Statistics for Corpus Linguistics*, Edinburgh University Press.

[2005년 1월 접수, 2005년 7월 채택]

Applying Randomization Tests to Collocation Analyses in Large Corpora

Kyung-Sook Yang[1] HeeYoung Kim[2]

ABSTRACT

Contingency tables are used to compare counts of n-grams to determine if the n-gram is a true collocation, meaning that the words that make up the n-gram are highly associated in the text.

Some statistical methods for identifying collocation are used. They are Kulczinsky coefficient, Ochiai coefficient, Frager and McGowan coefficient, Yule coefficient, mutual information, and chi-square, and so on.

But the main problem is that these measures are based on the assumption of a normal or approximately normal distribution of the variables being sampled. While this assumption is valid in most instances, it is not valid when comparing the rates of occurrence of rare events, and texts are composed mostly of rare events.

In this paper we have simply reviewed some statistics about testing association of two words. Some randomization tests to evaluate the significance level in analyzing collocation in large corpora are proposed. A related graph can be used to compare different test statistics that can be used to analyze the same contingency table.

Keywords: Co-occurrence, Collocation, Association, Chi-square statistic, Mutual information

1) Post Doctoral Researcher, Brain Korea 21 The Education and Research Group for Korean Studies, Korea University. Anam-Dong 5-Ga, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: myksyang@naver.com

2) Research Assistant Professor, Institute of Statistics, Korea University. Anam-Dong 5-Ga, Sungbuk-Gu Seoul 136-701, Korea.
E-mail: starkim@korea.ac.kr