코퍼스 빈도 정보 활용을 위한 적정 통계 모형 연구:

코퍼스 규모에 따른 타입/토큰의 함수관계 중심으로

양경숙, 박병선*
고려대학교

Kyungsook Yang and Byungsun Park. 2003. The Statistical Relationship between Linguistic Items and Corpus Size. *Language and Information 7.2*, 103–115. In recent years, many organizations have been constructing their own large corpora to achieve corpus representativeness. However, there is no reliable guideline as to how large corpus resources should be compiled, especially for Korean corpora. In this study, we have contrived a new statistical model, ARIMA (Autoregressive Integrated Moving Average), for predicting the relationship between linguistic items (the number of types) and corpus size (the number of tokens), overcoming the major flaws of several previous researches on this issue. Finally, we shall illustrate that the ARIMA model presented is valid, accurate and very reliable. We are confident that this study can contribute to solving some inherent problems of corpus linguistics, such as corpus predictability, corpus representativeness and linguistic comprehensiveness. (**Korea University**)

Key words: 타입수와 토큰수 간의 관계(relationship between the number of types and the number of tokens), 1억 어절(100 million word corpus), 코퍼스 규모(corpus size), ARIMA 모형(ARIMA model), 모의실험(simulation), 빈도(frequency)

1. 서론

지금까지 코퍼스를 구축하는 데 있어서, 자료의 다양성을 고려한 자료 균형성 문제와 더불어 코퍼스 구축 규모의 문제는 매우 중요한 고려사항이다. 이런 문제는 일찍이 영어 코퍼스를 중심으로 많은 연구가 진행된 바가 있지만 한국어를 대상으로 한 엄밀한 연구는 많이 이루어지지 않았다.

* 136-701 서울시 성북구 안암동 5가 고려대학교 한국학관 고려대학교 BK21 한국학 교육연구단 정보화팀, Email: ksyang27@korea.ac.kr, bpark@ikc.korea.ac.kr.

이 논문의 일차적 목적은 엄밀한 통계적 방법을 적용하여 코퍼스 규모(크기)에 따른 타입과 토큰간의 관계를 밝히고자 하는 것이다. 특히 이 연구에서는 현재까지의 국내 연구에서 다루지 않았던 규모인, 현대 한국어 코퍼스 1억여 어절을 대상으로 코퍼스 크기 증가에 따른 타입과 토큰의 상관성을 통계적 모델을 이용하여 규명하였다. 그리고 이를 위해 통계적 방법을 엄밀히 적용하여 좀더 명시적인 결과를 도출하였다.

계량적 언어 연구 방법에서 이용하는 언어 자료의 중요성은 여기서 더 상세히 언급할 필요는 없다. 그런데 이렇게 중요한 언어자료가 실제 언어 사용 양상을 얼마나 잘 보여줄 수 있는가에 대한 논의가 모든 계량적 연구에 앞서 선행되어야 함이 필수이다. 그러나 순수 언어학적 이론이나 전산적 방법으로는 언어자료의 대표성 확보에 대한 근거를 분명히 제시할 수 없다.

계량언어학에서 대규모 자료를 다루기 위해서는 전산적 자동처리와 그 결과를 효과적으로 해석하여 활용하기 위해 통계활용이 필수적이다. 그런데 언어처리를 위한 통계 방법론이 한국어 연구에는 그리 활발하게 이루어지지 않고 있다. 그리고 기본적으로 한국어 코퍼스 자체의 특성을 통계적으로 충분히 분석하여 그 결과를 활용할 필요성이 있다. 이를 위해서 본 논문은 한국어 코퍼스의 규모에 대한 특성을 통계적으로 밝혀 보고자 하였다.

따라서 이 논문에서 보이는 계량적 방법론을 활용하여 각 장르별 통계적 모형과 또 단어의 빈도와의 상관성도 통계적 모형으로 설명한다면, 각 언어자료별로 연구에 적정한 코퍼스의 규모를 엄밀히 예측하여 보다 객관적이고 명시적인 연구 방법을 세울 수 있다고 본다.

예를 들어 자연언어처리의 관점에서 보면, 이 연구는 어떤 자연언어처리 도구에서 내부적으로 활용하는 어휘 데이타베이스의 규모를 확인하여 그 도구가 효율적으로 다룰 수 있는 언어자료의 양을 제시할 수 있다. 기존 자연언어처리 연구 분야의 도구들의 효용성을 보여주는 논문들에서 다루었던 검증 자료의 양이 대부분 수십만 어절 규모에 지나지 않는다. 그리고 그 결과들을 실제 대규모 언어자료를 처리하기 위해 얼마나 잘 사용할 수 있는지 실험적으로 검증된 예가 드물고 실험하기도 여러 힘든 면이 있다. 따라서 이 논문의 연구 결과를 활용하면 자연언어처리 도구 평가에서 정량적 평가에 효과적으로 활용할 수 있다고 본다.

2. 연구 방법

2.1 연구 대상 자료

이번 연구에 사용한 현대 국어 코퍼스는 '21세기 세종계획-국어 기초자료 구축' 분과에서 1998년부터 2001년까지 구축한 현대국어 자료 1억여 어절이다. 이들 자료들은 비교적 언어 자료 분포의 균형성을 고려하여 구축한 것으로, 본문 오류 수정과 기본적인 TEI mark-up등의 표준화를 거친 것이다.

이 연구에서 말하는 토큰이란 통상적으로 텍스트에서 띄어쓰기를 기준으로 하는 어절로서, 어절의 갯수는 코퍼스의 규모를 나타내는 것이다. 그리고 타입이란 토큰을

유형별로 나눈 것으로 타입의 수는 결국 토큰의 종류 다양성을 나타내는 것이다. 그런데 이 연구에서는 완전히 기계적 처리의 관점에서 언어적으로 같은 표현이더라도 기호나 띄어쓰기 형식이 다른 경우는 모두 다른 타입으로 간주하였다. 예를 들어 '대한민국', '대한 민국', '대한민국!'등은 같은 표현이지만 모두 다른 타입으로 보았다.

이렇게 기본형을 고려하지 않은 것은 기계적 처리 용이성을 우선 고려한 것이지만, 언어학적으로도 대규모의 코퍼스에서 토큰과 타입의 상관성을 밝히는 연구에서 크게 문제가 되지 않는다고 본다. 이런 점에 대한 내용을 보여주는 연구는 Sanchez and Cantos(1997)과 D. Yang, et al(2002)이 있다. 한국어의 문법형태소는 100개 미만이고 실제 빈번하게 사용되는 것은 그 수가 더 한정되어 있다. 앞서 언급한 논문들에서도 굳이 단어의 기본형을 고려하지 않는다 할지라도 코퍼스 규모인 토큰과 타입의 상관성을 밝히는 데에는 큰 문제가 없음을 여러 연구사례를 통해 밝히고 있다. 따라서 이 논문에서 1억여 어절의 코퍼스를 활용하였고, 각 통계식들을 100회 모의실험을 통해 얻어진 토큰과 타입간의 관계식은 매우 유용하게 활용할 수 있는 연구 결과라고 생각한다.

이 자료들의 토큰과 타입의 수를 세는 데 있어서는 앞에서 언급한 바와 같이 띄어쓰기(스페이스)를 단위로 해서 여러 굴절형과 기호 포함 어절 등을 모두 다른 어절로 간주하였다. 이는 기계처리에는 기본 인식 단위가 스페이스를 기준으로 한, 어절 단위임을 고려한 것이다. 언어학적으로 좀 더 엄밀한 자료처리를 요구할 수 있는데, 이를 위해서는 한국어 형태소 정보가 부착된 말뭉치를 이용할 수 있다. 언어학적으로 전처리한 자료에 대한 연구는 이번 연구 결과를 이용하여 추후에 진행할 예정이다.

이번 연구에 사용한 코퍼스는 총 114,387,008 어절 규모이고 [표 1]은 코퍼스 구성을 표로 보인 것이다.

장르	어절	백분율
구어	5,404,036	4.72
기타비출판	633,176	0.55
기타출판	228,789	0.20
신문	18,536,093	16.20
잡지	13,371,937	11.69
전자출판	1,031,809	0.90
책-교육	4,424,814	3.86
책-사회	5,998,403	5.24
책-상상	34,308,182	29.99
책-생활	1,383,343	1.20
책-예술	4,176,877	3.65
책-인문	13,245,445	11.57
책-자연	2,521,551	2.20
책-체험	8,002,216	6.99
책-총류	1,120,337	0.97
총합	114,387,008	100

[표 1] 코퍼스 구성 비율

[표 1]에서 보듯이 비교적 다양한 장르의 글들이 코퍼스를 구성하고 있다. 이 논문에서 보인 코퍼스의 타입과 토큰의 상관성 연구 방법을 각 장르별로 확대해서 그 특징을 밝힌다면, 코피스 구축시 각 장르별 구축량도 효과적으로 추정해 낼 수 있다.

2.2 자료 처리

'21세기 세종계획-국어기초자료 구축'분과에서 구축한 자료들은 모두 '한글'파일이다. 이들의 기계적 처리를 위해 총 6999개의 파일을 텍스트 파일로 변환하였다.

전체 코퍼스의 규모는 총 114,387,008어절로 이번 연구에서는 매100만 어절별로 토큰을 증가시키면서 누적된 타입 빈도를 계산하였다. 타입수의 변화 양상이 자료의 장르 특성에 영향을 받을 수 있기 때문에 분석에 사용된 코퍼스 파일 6999개를 난수를 이용하여 완전 랜덤화[1] 하였다. 그러나 한번만 파일의 순서를 랜덤화 하는 것으로써 장르의 성격이 토큰 타입간의 관계에 미칠 수 있는 영향력을 배재시키기 어렵다고 판단하여 완전랜덤화 하는 작업을 100회 실시하였다.[2] 결국 매번 완전 랜덤화된 데이터 파일로부터 토큰을 100만개씩 증가시키면서 누적된 타입수를 계산하였고 이와 같은 랜덤화 과정을 100회 반복한 것이다. 물론 이와 같이 완전 랜덤화한 데이터로부터 타입수를 계산하는 과정을 100번 반복한 것은 장르특성으로부터 영향을 가능한 받지 않도록 만들면서 토큰과 타입간의 관계를 규명하기 위함이다.

결국 최종적으로 타입과 토큰간의 관계 규명을 위해 사용한 데이터는 100회 반복하여 계산된 타입수의 평균으로 큐빅모형(cubic model), 파워모형(power model) 그리고 ARIMA모형(autoregressive integrated moving average model)[3]에 적합시켜 그 결과를 비교하였다.[4] 이들 3가지 모형에 대한 평가는 물론 잔차분석과 모형에 대한 설명력 등을 이용하여 이루어졌다.

3. 연구 결과

100만개의 토큰 단위별로 누적 관측된 평균타입개수와 토큰간의 관계를 살펴보기 위해 일차적으로 선도표를 작성하였다. [그림 1]은 1억 어절에 대하여 100만어절 단위별로 누적 관측한 타입수(세로축)과 토큰빈도(가로축(단위:백만어절))간의 관계를 나타내는 선도표(line plot)이다. 거의 일직선과 유사한 관계를 나타냄을 살펴볼 수 있다.

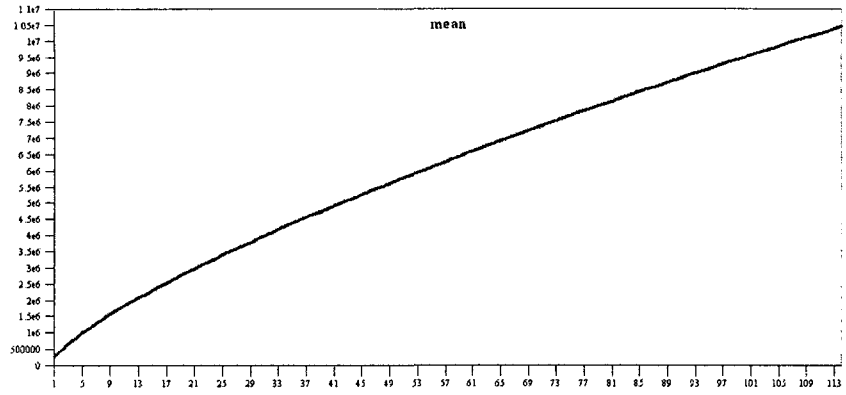
직선적인 결합강도를 나타내는 타입과 토큰간의 피어슨 상관계수(Pearson's correlation)는 0.995로 1%의 유의수준(significance level)에서 통계적으로 유의미한 결

[1] 여기서 완전 랜덤화는 6999개 파일의 나열을 비복원추출에 의해 무작위로 나열한 것을 의미한다. 이와 같이 무작위 추출로 파일을 나열한 이유는 토큰별 타입빈도가 장르나 기타 언어 특징에 의해 영향받는 것을 최대한 배제시키기 위함이다.

[2] 실제로도 1회 모의실험한 데이터에 대한 모형과 100회 모의실험하여 평균타입수에 적용한 모형은 다르다.

[3] ARIMA모형은 시차를 두고 관측된 데이터의 앞뒤 관계가 자기회귀와 이동평균모형으로 표현되는 것을 나타낸다.

[4] 여기서 큐빅모형과 파워모형은 데이터 적합을 위한 기본 가정을 충족시키지 않고 있으므로 본 논문에서는 해당모형에 의해 추정된 결과는 기술하지 않았다.



[그림 1] 토큰수와 평균타입개수의 선도표

과를 보여 매우 강한 양의 상관성을 보였다.

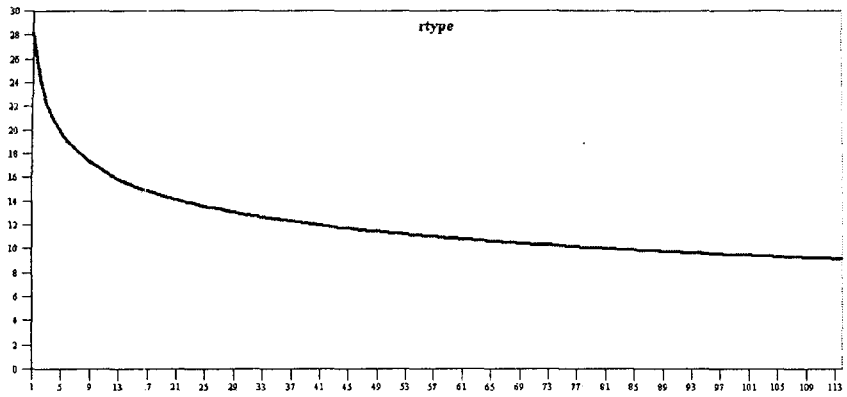
앞의 [그림 1]에서도 보이듯이 단순히 코퍼스 규모를 나타내는 토큰수와 그 증가에 따른 타입수의 증가의 상관성은 강한 비례관계이다. 기존의 일부 연구에서는 이 증가 추세가 어느 규모에서 급격히 감소하는지를 파악하여 코퍼스 구축에 있어서의 규모 적정성을 구하고자 했다. 그러나 본 논문의 연구를 진행하면서도 확인하였지만, [그림 1]과 같이 1억여 어절 규모의 코퍼스를 이용하여서도 토큰수에 대한 타입수의 증가 추세가 어느 규모 시점에서 급격한 변화를 보이는지는 찾아내기 힘들었다. 장석배(1998)에서는 약 4천만 어절 규모의 코퍼스를 사용하여 그 관계를 밝혀 보고자 한, 의미있는 연구였으나 토큰과 타입의 상관성만을 비교적 단순히 보였을 뿐이고 토큰과 타입의 상관성을 명시적으로 밝히는 통계 결과를 도출해 내지 못했었다.

따라서 이 논문에서 시도하는 방법은 단순하게 토큰수 증가에 대한 타입수 증가 비율만을 고려하는 방법으로는 토큰과 타입의 상관성을 분명히 밝힐 수 없었음을 고려하여, 토큰과 타입의 상관성을 엄밀한 통계적 모형 적합도를 구하고 함수화하여 그 관계를 밝히고자 한다.

[그림 2]는 세로축에 토큰 대비 평균타입개수 백분율을 표시하고 가로축은 [그림1]과 마찬가지로 100만 어절의 토큰수를 표시한 선도표이다. [그림1]과는 달리 [그림 2]에서는 토큰의 증가에 따른 평균타입 백분율은 약 천만어절 이전에선 급격히 감소하다 그 이후로는 지수적으로 완만히 감소하고 있다고 할 수 있겠다. 그리고 이들 두 변수간의 피어슨 상관계수는 -0.833으로 계산되어 음의 상관을 나타내었다.

[그림 1]의 선도표로부터 토큰의 변화량에 따라 타입의 개수가 얼마나 증가하는지를 예측하기 위하여 일차적으로 큐빅모형과 파워모형을 고려하였다.

큐빅모형과 파워모형에 대한 데이터 적합 결과, 설명력은 약 99%로 매우 높게 나왔지만 [그림 3], [그림 4]와 같이 오차에 대한 가정을 충족시키지 못하였다. 일반적으



[그림 2] 토큰대비 평균타입개수의 백분율에 대한 선도표

로 선형회귀모형에 적합시킬 경우 두는 오차항에 대한 독립성[5]이나 선형성[6] 가정을 잘 충족시킬 경우 잔차플롯은 [그림 3]이나 [그림 4]와 같이 특정한 패턴을 보이지 않는다.

더욱이 오차의 독립성을 살펴볼 수 있는 더빈-왓슨 통계량값[7]이 큐빅모형에서는 0.074로, 파워모형에서도 0.338로 계산되어 1차 자기상관관계[8]가 있음을 나타내었다. 이는 비록 결정계수 값이 높게 계산되었다 하더라도 파워모형이나 큐빅모형이 적합한 모형이 아님을 나타낸다. 아울러 토큰이 100만 어절씩 증가할수록 이에 따라 증가하는 타입수의 변화가 그 이전 시점의 타입수와 상관관계가 있음을 나타내는 것이다.

통계학에서 다루는 시계열 자료분석은 관측시점간의 데이터가 서로 상관되어 있는 경우에 적용할 수 있는 분석 방법의 하나이다. [그림 5a]와 [그림 5b]는 원자료인 평균타입에 대해 토큰수의 변화에 따른 자기상관함수(ACF)[9]와 편자기상관함수(PACF)[10]를 플롯으로 나타낸 것이다. [그림 5a]와 같은 패턴을 보이는 경우, 시계열이 시간이 흐름에 따라 변하는 확률적 추세를 보이는 비정상적인 시계열 자료임을 나타낸다.

시계열 자료분석에서 고려하는 백색잡음(white noise)[11]에 대한 가정 충족을 위해 종종 차분을 이용하는데 가령 1차 차분이란 현재시점에서 이전 시점의 데이터간의 차이를 말한다. 본 연구에서 사용한 100번 모의실험한 평균타입수의 경우 3차 차분이 유용하였다.

---

5 오차의 독립성은 서로 다른 시점의 관측값들간에 서로 독립임을 가정하는 것이다.

6 모수에 대해 선형관계임을 나타냄.

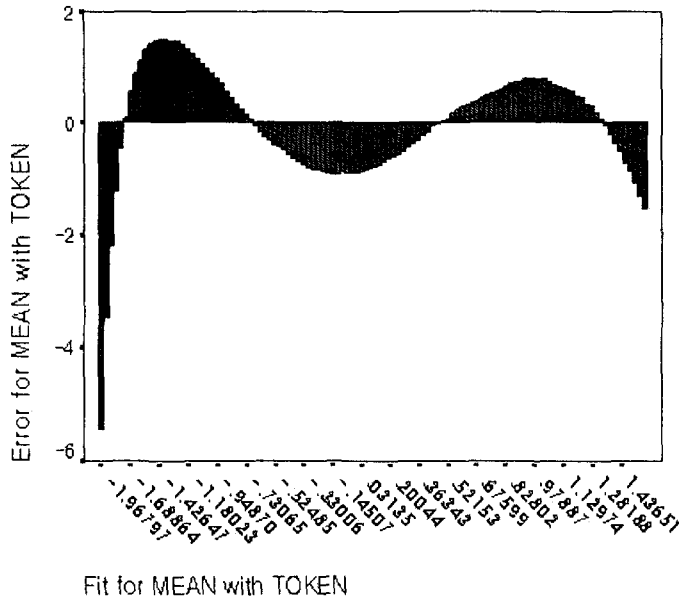
7 현재 시점과 한 시점전의 데이터간의 관계를 살펴볼 수 있는 통계량으로 데이터간에 독립성이 유지되면 더빈왓슨 통계량값은 2근방의 값으로 계산된다.

8 현재시점과 한 시점전의 데이터간의 상관관계를 가르킴.

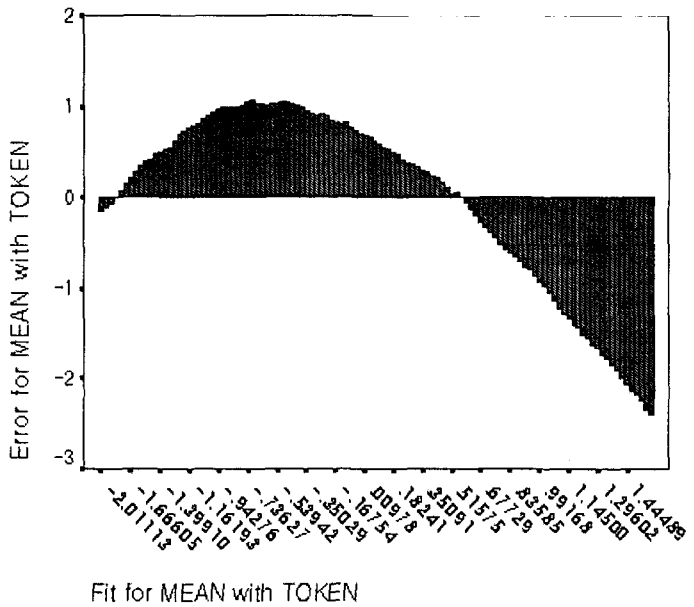
9 동일 변수에서 연속적인 관측값들 사이의 상호연관 관계를 나타내는 척도

10 나머지 시점에 의한 영향력을 제거한 서로 다른 두 시점간의 순수한 상관관계

11 서로 독립이고 평균 0, 분산이 σ^2인 정규분포를 따르는 확률변수를 나타냄.

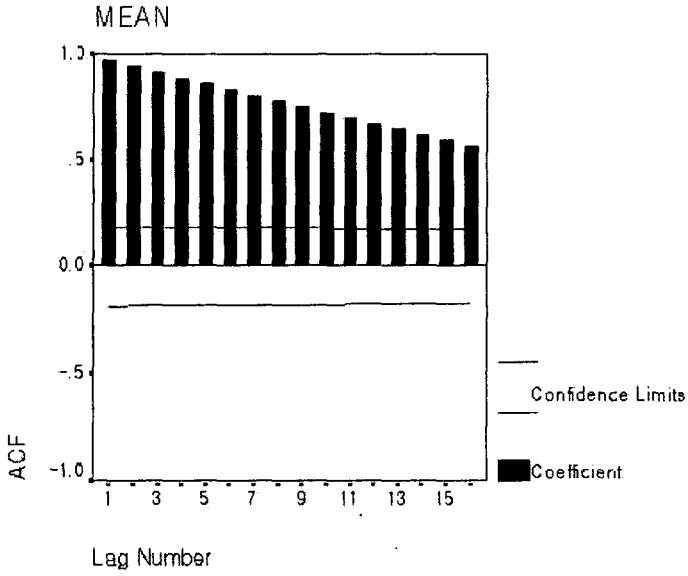


[그림 3] 큐빅모형에 적합시킨 후의 표준화 잔차 플롯(bar chart)

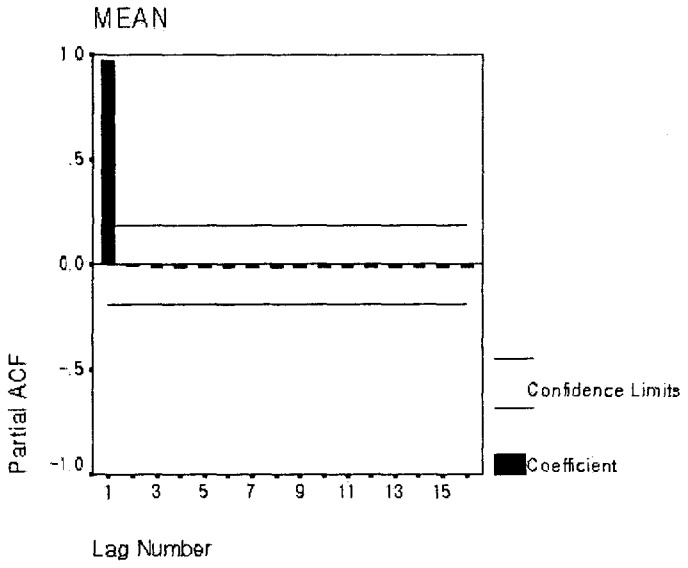


[그림 4] 파워모형에 적합시킨 후의 표준화 잔차 플롯(bar chart)

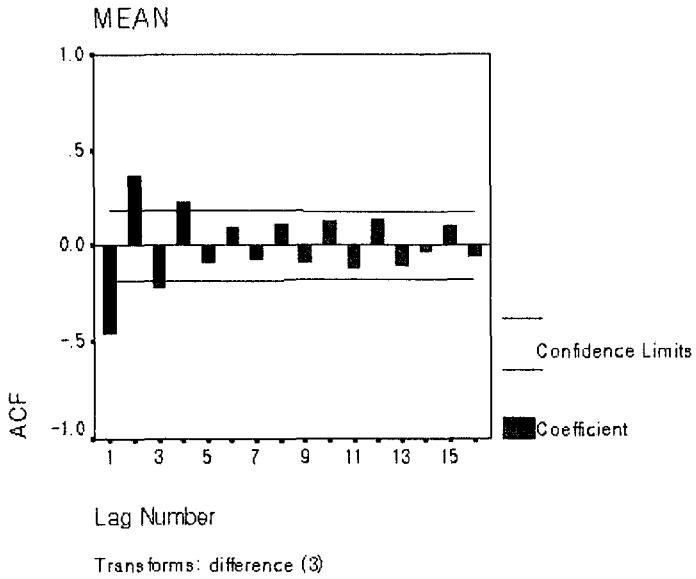
타입을 3차 차분한 후 자기상관함수와 편자기상관함수를 살펴본 결과 [그림 6a], [그림 6b]와 같이 특정 시점을 제외하고 모두 95% 신뢰구간 안에 들어오거나 감소하는 경향을 보일 경우 오차항에 대한 가정이 만족되었음을 나타낸다.



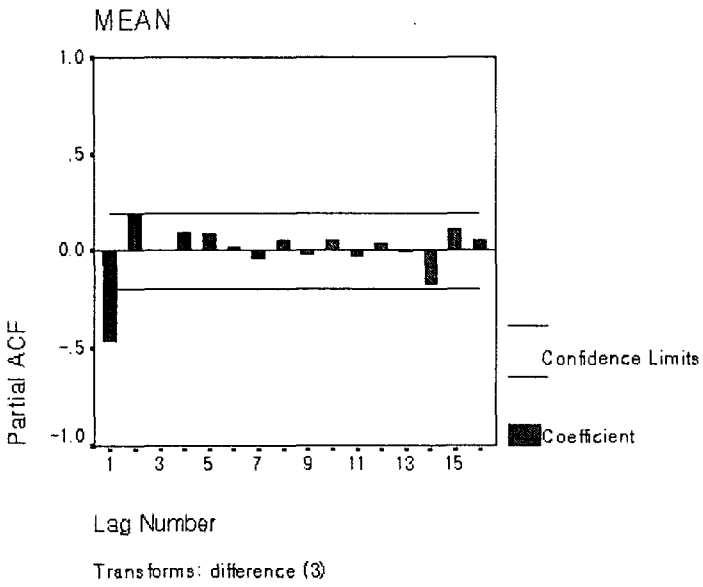
[그림 5a] 원자료 타입에 대한 ACF



[그림 5b] 원자료 타입에 대한 Partial ACF



[그림 6a] 3차 차분된 타입의 ACF



[그림 6b] 3차 차분된 타입의 Partial ACF

시계열이 AR 1의 특성을 보일 경우 자기상관계수는 [그림 6a]와 같은 패턴을 보이고 편자기상관계수는 시점1에서만 특정값을 나타내며 그 이후 시점에서는 0의 값으

로 계산된다. [그림 6b]에서도 시점1에서만 -0.5정도의 값을 나타내며 나머지 시점에서는 모두 신뢰한계(confidence limits)안에 들어오므로 0이라고 생각할 수 있다. 따라서 [그림 6a]와 [그림 6b]로부터 AR1 임을 확인하여 타입의 개수를 추정하기 위한 모형으로 ARIMA(1,3,0)모형에 적합시켜 [표 2]의 결과를 구하였다. [표 2]에서 AIC는 아카이케 정보량 기준값이고 SBC는 Schwartz의 베이즈통계량값을 나타낸다. AIC와 SBC는 시계열모형을 평가할 때 주로 사용되는 통계량으로 이들 통계량들은 오차에 대한 분산 추정값에 로그변환을 취한 후 설정된 모형의 모수개수를 더하는 형태로 백색잡음의 분산이 작을수록 데이터를 잘 적합시킨다고 생각할 수 있는 것이다. 따라서 여러 모형을 비교할 때 이들 통계량 값이 작게 계산된 것을 최적모형으로 선택하게 된다.

Number of residuals	111			
Standard error	1953.5667			
Log likelihood	-998.31631			
AIC	1998.6326			
SBC	2001.3422			
Analysis of Variance:				
	DF	Adj. Sum of Squares	Residual Variance	
Residuals	110	421496414.2	3816422.7	
Variables in the Model:				
	B	SEB	T-RATIO	APPROX. PROB.
AR1	-.59981001	.07611627	-7.8801812	.0000000

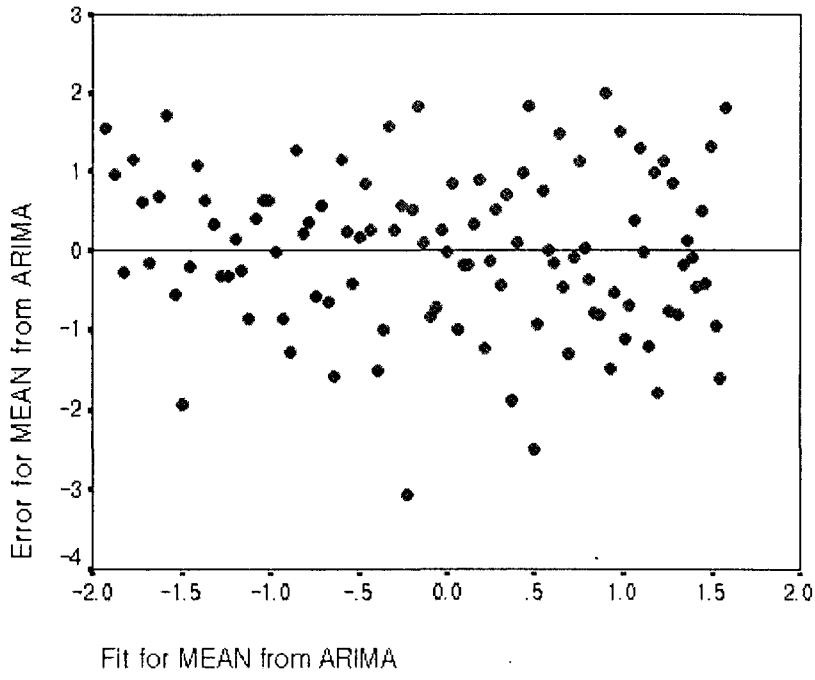
[표 2] ARIMA(1,3,0) 모형적합 결과

위 결과로부터 100만 어절 단위로 관측된 평균타입수(Y)에 대해 적합된 모형식은 다음과 같이 수립된다.

$$Y_t = Y_{t-1} - 0.599Y_{t-1} - Y_{t-2} + 0.599Y_{t-2} + Y_{t-3} - 0.599Y_{t-3} \quad (1)$$
$$+0.599Y_{t-4} + a_t$$

여기서 아래첨자 t는 관측단위별 시점을 나타내는 것으로 토큰 100만 어절 단위로 표시하였다. 즉 t=10 이면 1000만어절의 토큰일 경우의 평균 타입수를 나타낸다. 즉 t시점의 평균 타입개수는 한 시점 이전의 평균타입수에 의해서는 양의 영향력을 받고 두 시점 이전의 평균타입수에 의해서는 음의 영향력이 크다고 할 수 있다. 그리고 5시점 이전의 평균타입수에 의한 영향력은 무시해도 좋다고 생각할 수 있는 것이다.

ARIMA(1,3,0) 모형에 적합시킨 이후의 잔차플롯을 보면 일부 몇 개의 잔차만이 ±3을 벗어나 있고 나머지 잔차들은 모두 그 범위 안에 들어옴을 확인할 수 있다. 이 모형으로부터 계산된 잔차플롯에서 ±3을 벗어나는 2개 잔차를 제외하면 [그림 7]과 같이 백색잡음에 대한 가정을 잘 충족시키고 있는 잔차플롯임을 확인할 수 있다.



[그림 7] ARIMA(1,3,0) 모형 적합후의 잔차플롯

[표 3]은 ARIMA(1,3,0)에 의해 추정된 타입수와 실제 관측된 평균타입빈도간의 차이를 나타내는 잔차를 앞에서 비교한 3개 모형에 대해 계산한 결과를 나타낸 테이블이다.[12]

전반적으로 살펴볼 때 잔차크기는

파워모형 > 큐빅모형 > ARIMA(1,3,0)

의 관계를 나타낸다. 따라서 100번 모의실험을 통해 장르의 영향력을 최대한 배제시킨 후 구한 토큰과 타입수와의 함수적 관계는 ARIMA(1,3,0)가 가장 잘 설명됨을 확인할 수 있다.

4. 맺음말

본 논문에서는 이제까지 연구가 없었던 약 1억 어절이라는 대용량 데이터에 대하여 코퍼스의 크기 증가와 타입간의 상관성을 함수적 관계로 규명하였다; 더욱이 장르의 영향을 최대한 배제시키기 위해 100회의 모의실험을 통해 함수적 관계를 규명하였다.

대규모 코퍼스 데이터를 통해 형태소를 분석할 경우 대용량 사전을 준비하게 되

[12] 여기서 잔차는 100회 반복 실험된 데이터에 3가지 모형을 적용시킨 후 계산한 잔차이다. 세 가지 모형을 비교함은 현재 타입과 토큰수의 관계를 규명하는데 파워모형이 이용되나 통계적 관점에서 볼 때 오차항에 대한 기본 가정을 충족시키면서 보다 잘 데이터를 설명하는 모형을 살펴보기 위함이다.

토큰 (단위:백만)	잔차1 (큐빅모형)	잔차2 (파워모형)	잔차3 (ARIMA(1,3,0))
60	-24606	14876	1243
61	-23230	13204	-1423
62	-21332	11901	162
63	-20167	9721	-356
64	-18282	8128	726
65	-15469	7338	1105
66	-13450	5641	-1571
67	-11071	4201	147
68	-8415	2944	635
69	-5935	1431	-477
70	-2285	1016	1101
71	-924	-1748	-2627
72	913	-4084	716
73	2566	-6641	1024
74	7029	-6413	2625
75	8833	-8859	-3648
76	10857	-11085	-375
77	12052	-14131	703
78	13271	-17130	249
79	13750	-20835	-226
80	16093	-22629	2171
81	17814	-24986	-897
82	19088	-27718	-1290
83	19743	-30984	-42
84	21494	-33057	1638
85	23189	-35076	-98
86	25269	-36588	-224
87	26416	-38897	-1028
88	26597	-42022	-797
89	28576	-43186	2771
90	28399	-46332	-2273
91	28348	-49162	-63
92	28815	-51273	1787
93	28080	-54370	-1459
94	26549	-58034	-600
95	24513	-61960	561
96	23567	-64540	1795
97	22587	-66884	-141
98	20858	-69693	-1364
99	20316	-71019	1532
100	17199	-74606	-2573

[표 3] 잔차 (일부)

는데 일상생활에서 사람들이 사용하는 어휘는 사전에 실린 어휘의 10-15%라고 한다. 따라서 본 연구에서 밝힌 토큰수 별 예상되는 타입수 보다 실제로는 더 많은 사전이 대용량 데이터를 다룰 경우에는 필요할 것으로 판단된다.

그리고 세 모형을 통계적으로 비교하는 과정에서 오차의 가정을 벗어나는 관측값들을 살펴보면서 통계적 처리를 위해서는 최소한 1000만 어절 이상의 규모가 필요할 것으로 판단되었다. 따라서 지금까지 '21세기 세종계획'등에서 구축한 균형 코퍼스(balanced corpus)의 규모가 1000만 어절 규모인데 이 연구에 따르면, 그 규모는 통계적 처리에 비교적 적정 구축량임을 확인할 수 있었다.

이 연구에서 밝혀진 통계모형을 코퍼스 구축에 있어서 장르별 특징과 시대별 특징을 대상으로 확대 발전시킨다면, 각 연구별 적정 코퍼스 구축량을 예측하고 가장 적절한 통계처리 연구를 할 수 있는 방법론의 근거를 제시한 의의가 있다고 하겠다. 그리고 저빈도 타입의 증가와 코퍼스 크기 증가에 따른 변화 특징 등의 연구에 응용하여, 저빈도 단어들의 적절한 통계 처리 방법을 개발해 낼 수 있다. 따라서 이 논문은 앞에서 언급한 여러 의미있는 연구를 진행하기 위한 기본적이면서 가장 중요한 검증과 방법론을 보였고 앞으로 계속 연구를 진행할 것이다.

<참고문헌>

강범모. 2003. 언어, 컴퓨터, 코퍼스 언어학-컴퓨터를 이용한 국어 분석의 기초와 이론. 고려대학교 출판부.

장석배. 1999. 말뭉치 규모와 어절 유형 증가간의 상관성에 대한 연구. 언어 정보의 탐구 1, pp. 159-210.

Curch, K.W, and R.L. Mercer. 1994. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. In *Using Large Corpora.* Ed. Susan Armstrong. The MIT Press, pp. 1-24.

Yang, Dan-hee, Ik-hwan Lee, Pascual Cantos. 2002. On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition. *Computers and the Humanities* 36. pp. 171-190.

Sanchez, A. and P. Cantos. 1997. Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-million-word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics*, 2.2, pp. 259-280.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics.* John Benjamins.

Wei, William W.S. 1990. *Time Series Analysis.* Addison Wesley.

Zernik, U. 1991. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates, pp. 1-26.

접수일자: 2003년 11월 10일
게재결정: 2003년 12월 6일