코퍼스 규모에 따른 타입과 토큰의 상관성 연구

고려대학교 통계학과, 국어국문학과, 컴퓨터학과
양경숙 · 박병선 · 임준호

The Statistical Relationship between Types and Tokens

Kyung Sook Yang, Byung Sun Park, Jun Ho Lim
Department of Statistics, Korean Lit. & Lan., Computer Science Korea University, Seoul, Korea

요 약

이 논문의 목적은 코퍼스 크기에 따른 타입과 토큰간의 관계를 엄밀한 통계적 방법으로 그 특징을 밝히고자 하는 것이다. 지금까지 코퍼스를 구축하는 데 있어서, 자료의 다양성을 고려한 자료 균형성 문제와 더불어 코퍼스 구축 규모의 문제는 매우 중요한 고려사항이었다. 이런 문제는 일찍이 영어 코퍼스를 중심으로 많은 연구가 진행된 바가 있지만 한국어를 대상으로 한 엄밀한 연구는 많이 이루어지지 않았다. 이 연구에서는 현재까지 구축한 현대 한국어 말뭉치 1억여 어절을 대상으로 말뭉치 크기 증가에 따른 타입과 토큰간의 통계적 관계를 3가지 모형에 대해 비교하였으며 최종적으로 ARIMA모형을 이용하여 그 함수적 관계를 밝혀보았다. 연구 결과에 따르면 한국어 자료는 약 1천만 어절의 토큰을 기준으로 타입의 변화가 다소 둔화되는 결과를 보인다. 연구에 의해 도출된 함수식을 이용하면 소규모의 자료를 이용하더라도, 대규모 자료에서의 타입수를 계산해 낼 수 있으므로, 더욱 다양하고 정확한 통계처리의 근거를 제시할 수 있게 된다.

서 론

이 논문의 목적은 코퍼스 규모(크기)에 따른 타입과 토큰간의 관계를 엄밀한 통계적 방법으로 그 특징을 밝히고자 하는 것이다. 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. 특히 이 연구에서는 현재까지의 연구에서 다루지 않았던 규모인 현대 한국어 말뭉치 1억여 어절을 대상으로 말뭉치 크기 증가에 따른 타입과 토큰의 상관성을 통계적 모델(ARIMA)을 이용하여 그 함수적 관계를 밝힌 의의가 있다. 그리고 통계적 방법을 엄밀히 적용하여 좀더 명시적인 결과를 도출한 의의도 크다고 생각한다.

E-mail : ksyang27@korea.ac.kr
E-mail : bpark@ikc.korea.ac.kr
E-mail : jhlim@nlp.korea.ac.kr

연구 결과에 따르면 한국어 자료는 약 1천만 어절을 기준으로 타입의 변화량이 다소 둔화되는 결과를 보인다. 연구에 의해 도출된 함수식을 이용하면 소규모의 자료를 이용하더라도, 대규모 자료에서의 타입수를 계산해 낼 수 있으므로, 더욱 다양하고 정확한 통계처리의 근거를 제시할 수 있게 된다. 또 이와 더불어 이 논문에서 시도한 방법을 코퍼스를 구성하는 각 장르별, 크기별로 그 함수적 특징을 밝히는 데 적용한다면, 각 언어자료별로 연구에 적정한 코퍼스의 규모를 엄밀히 예측하여 보다 객관적이고 명시적인 연구 방법을 세울 수 있다고 본다.

연구방법

1. 연구 대상 자료

이번 연구에 사용한 현대 국어 말뭉치는 '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-up 등의 표준화를 거

친 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.

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.

2. 자료 처리

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

전체 코퍼스의 규모는 총 114,387,008어절로 이번 연구에서는 매 100만 어절별로 토큰을 증가시키면서 누적 타입을 계산하였다. 타입수의 변화가 언어 자료의 장르적 특성에 영향을 받을 수 있기 때문에 분석에 사용된 코퍼스 파일 6999개를 난수를 이용하여 완전 랜덤화 하였다.

타입과 토큰간의 관계 규명을 위해 사용한 모형은 큐빅모형(cubic model), 파워모형(power model) 그리고 ARIMA모형(autoregressive integrated moving average model)이다. 이들 3가지 모형에 대한 평가는 물론 잔차분석과 모형에 대한 설명력 등을 이용하여 이루어졌다.

연구결과

타입과 토큰간의 관계를 살펴보기 위해 일차적으로 선도표를 작성하였다. Fig. 1은 1억 어절에 대하여 100만어절 단위별로 누적 관측한 타입과 토큰간의 관계를 나타내는 선도표(line plot)이다. 거의 일직선과 유사한 관계를 보이고 있다.

일직선적인 연관정도를 나타내는 타입과 토큰간의 피어슨 상관계수(Pearson's correlation)는 0.995로 1%의 유의수준(significance level)에서 통계적으로 유의미한 결과를 보여 매우 강한 양의 상관성을 지니고 있음을 나 알 수 있다.

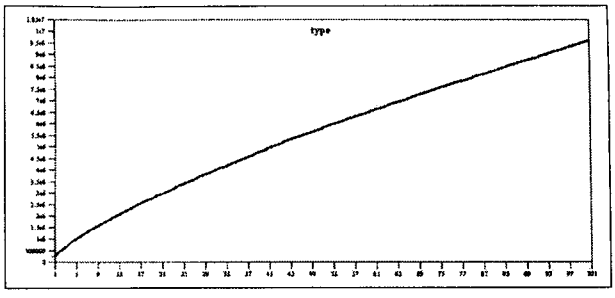

Fig. 1. 타입과 토큰의 선도표.

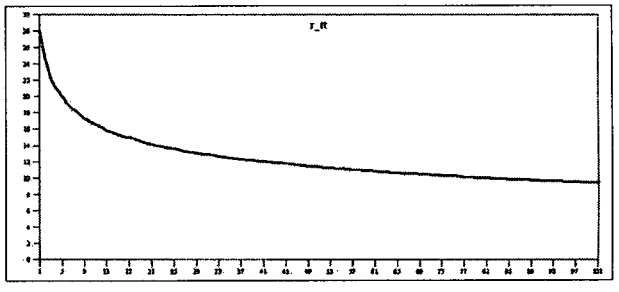

Fig. 2. 토큰대비 타입의 백분율의 선도표.

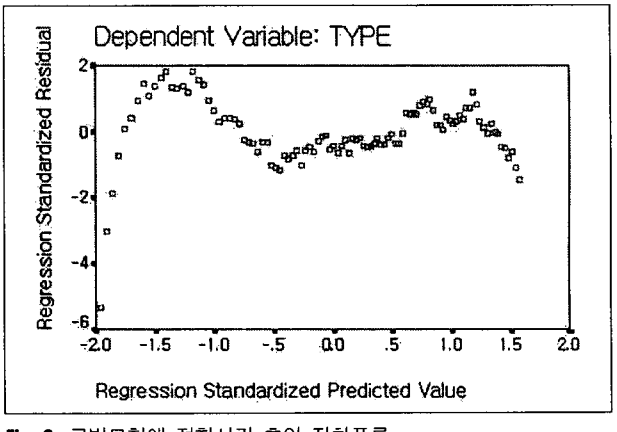

Fig. 3. 큐빅모형에 적합시킨 후의 잔차플롯.

Fig. 2는 토큰에 대한 타입의 백분율을 나타낸 선도표이다. 토큰의 증가에 따른 타입의 백분율이 약 천만어절 이전에선 급격히 감소하다 그 이후로는 지수적으로 완만히 감소하고 있음을 알 수 있다. 이들 두 변수간의 상관계수는 -0.837로 계산되어 음의 상관을 나타내었다.

Fig. 1의 선도표로부터 토큰의 변화량에 따른 타입의 개수를 예측하기 위하여 일차적으로 큐빅모형과 파워모형을 고려하였다.

큐빅모형과 파워모형에 대한 데이터 적합 결과, 설명력은

약 99%로 매우 높게 나왔지만 Fig. 3, Fig. 4와 같이 오차에 대한 가정을 충족시키지 못하고 있음을 확인할 수 있다. 더욱이 오차의 독립성을 살펴볼 수 있는 더빈-왓슨 통계량 값이 큐빅모형에서는 0.17로 계산되었고 파워모형에서도 0.338로 계산되어 1차 자기상관관계가 있음을 나타내었다.

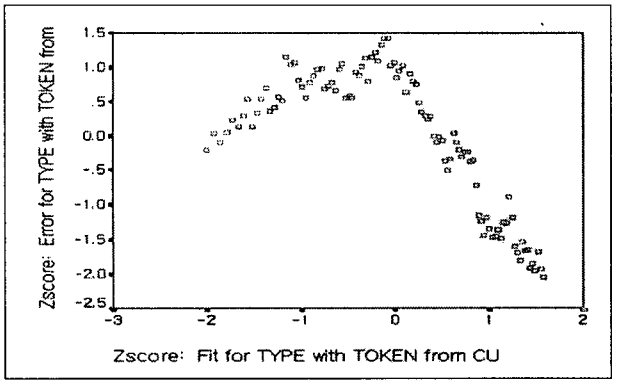

Fig. 4. 파워모형에 적합시킨 후의 잔차플롯.

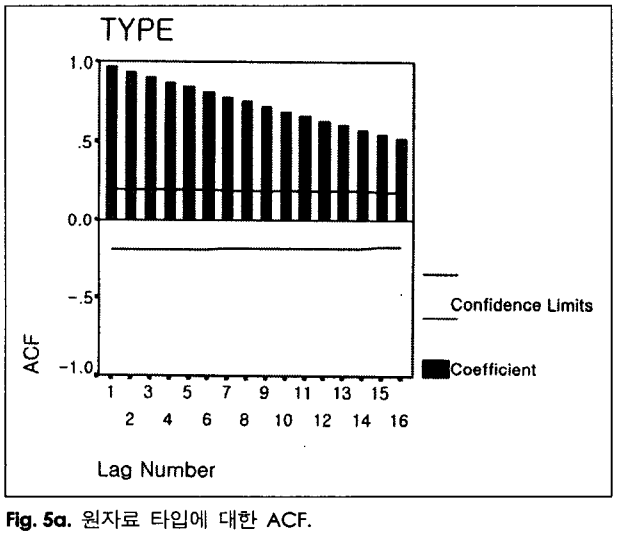

Fig. 5a. 원자료 타입에 대한 ACF.

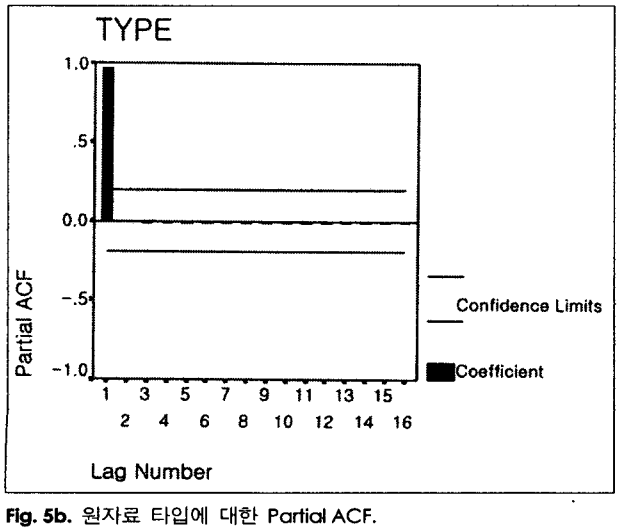

Fig. 5b. 원자료 타입에 대한 Partial ACF.

이와 같은 더빈-왓슨 통계량값은 비록 결정계수 값이 0.99 정도로 높게 계산되었다 하더라도 파워모형이나 큐빅모형이 적합한 모형이 아님을 나타낸다. 아울러 토큰이 100만 어절씩 증가할수록 이에 따라 증가하는 타입수의 변화가 그 이전 시점의 타입수와 상관이 있음을 나타내는 것이다.

통계학에서 다루는 시계열 자료분석은 관측시점간의 데이터가 서로 상관되어 있는 경우에 적용할 수 있는 분석방법의 하나이다. Fig. 5a와 Fig. 5b는 원자료 타입이 토큰수의 변화에 따른 자기상관함수(ACF)와 편자기상관함수(PACF)를 나타내며 비정상적인 시계열 자료임을 보여준다.

시계열 자료분석에서 고려하는 백색잡음(white noise)에 대한 가정 충족을 위해 차분을 이용하는데 본 연구에서 사용한 데이터의 경우 2차 차분이 유용하였다.

타입을 2차 차분한 후 자기상관함수와 편자기상관함수를 살펴본 결과 Fig. 6a, 6b와 같이 95% 신뢰구간 안에 들어오므로 오차항에 대한 가정이 만족되었음을 알 수 있다.

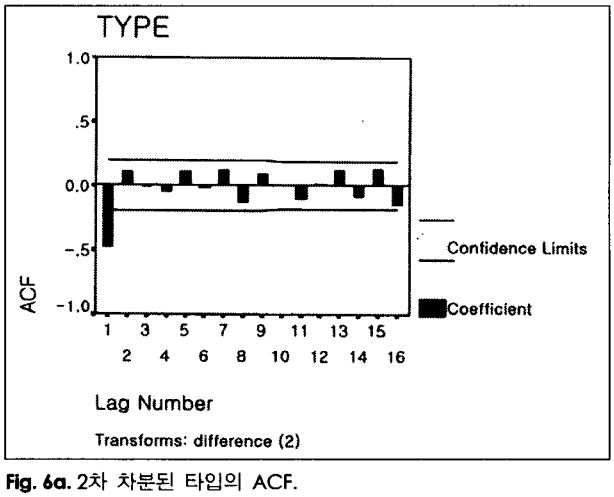

Fig. 6a. 2차 차분된 타입의 ACF.

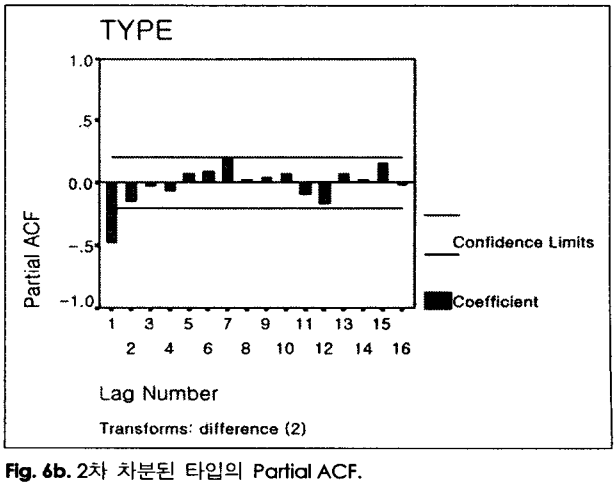

Fig. 6b. 2차 차분된 타입의 Partial ACF.

Table 1. ARIMA(0,2,1) 모형적합 결과

Number of residuals	99			
Standard error	11212.2			
Log likelihood	−1062.8			
AIC	2129.6			
SBC	2134.8			
Analysis of Variance :				
	DF	Adj. Sum of squares	residual variance	
Residuals	97	12244596411.2	125713816.4	
Variables in the Model :				
	B	SEB	T-RATIO	APPROX. PROB.
MA1	.5788	.085	6.7648	.00000000
CONSTANT	−1203.86	481.366	−2.5009	.01406283

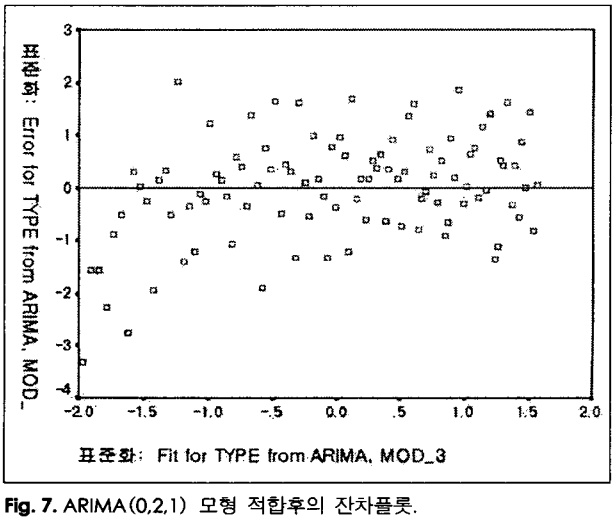

Fig. 7. ARIMA(0,2,1) 모형 적합후의 잔차플롯.

Fig. 6a와 Fig. 6b의 플롯이 MA 2(moving average 2)임을 나타내므로 타입의 개수를 추정하기 위한 모형으로 ARIMA(0,2,1)모형에 적합시켰다.

그 결과는 Table 1과 같다.

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

$$Y_t = -1203.86 + 2Y_{t-1} - Y_{t-2} + a_t - 0.5788a_{t-1}$$

여기서 아래첨자 t는 관측단위별 시점을 나타내는 것으로 토큰 100만 어절 단위로 표시하였다. 즉 t=10 이면 1000만어절의 토큰일 경우의 타입수를 나타낸다.

ARIMA(0,2,1) 모형에 적합시킨 이후의 잔차플롯을 살펴보면 Fig. 7과 같이 Fig. 3이나 Fig. 4의 잔차플롯과 달리 백색잡음에 대한 가정을 잘 충족시키고 있음을 알 수 있다.

Table 2. 추정된 타입수(일부)

토큰수	관측된	ARIMA(0,2,1)		
		추정된 타입수	95% 신뢰하한	95% 신뢰상한
92	8940502	8938383	8957545	8919222
93	9024858	9008331	9027493	8989170
94	9093485	9095641	9114802	9076479
95	9167653	9163037	9182199	9143876
96	9231272	9237492	9256653	9218330
97	9307132	9298970	9318132	9279809
98	9375761	9375910	9395072	9356749
99	9459172	9443941	9474226	9413656
100	9522556	9511225	9552296	9470153
101	9590357	9577761	9629848	9525673

Table 2. continue

관측된 타입수	큐빅모형 적합값	파워모형 적합값
8940502	8941384	9018752
9024858	9017199	9092749
9093485	9093304	9166552
9167653	9169719	9240164
9231272	9246467	9313587
9307132	9323569	9386824
9375761	9401047	9459876
9459172	9478922	9532747
9522556	9557216	9605438
9590357	9635950	9677952

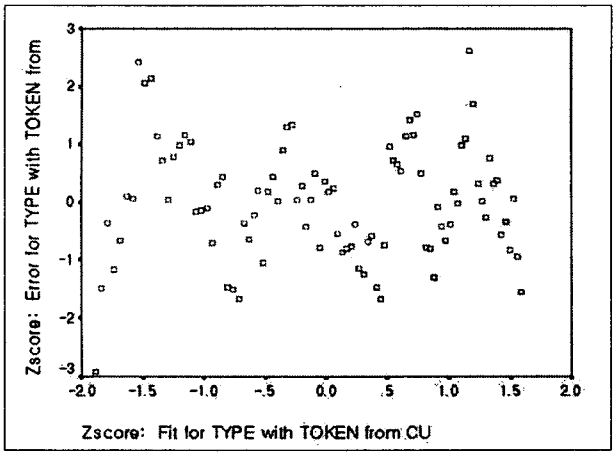

Fig. 8. 큐빅 모형에 적합시킨 이후의 잔차플롯(900만 이하 토큰 제외).

Table 2는 ARIMA(0,2,1)에 의해 추정된 타입수와 그에 대한 95% 신뢰구간의 일부결과를 나타낸 테이블이다.

타입의 변화율도 마찬가지로 ARIMA(0,2,1)모형에 다음과 같이 적합된다.

$$Rtype_t = 0.035 + 2Rtype_{t-1} - Rtype_{t-2} + a_t - 0.5122a_{t-1}$$

Correlations

		TYPE	TOKEN	R_TT
TYPE	Pearson Correlation	1	.997**	-.964**
	Sig. (2-tailed)	.	.000	.000
	N	92	92	92
TOKEN	Pearson Correlation	.997**	1	-.944**
	Sig. (2-tailed)	.000	.	.000
	N	92	92	92
R_TT	Pearson Correlation	-.964**	-.944**	1
	Sig. (2-tailed)	.000	.000	.
	N	92	92	92

**. Correlation is significant at the 0.01 level (2-tailed).

한편 Fig. 8은 1000만 어절 이상의 토큰만을 고려하여 큐빅모형에 적합시킨 이후 작성한 잔차플롯이다. 완전 독립적인 패턴은 아직 불완전하지만 Fig. 3보다 잔차플롯이 훨씬 더 많이 잔차 가정을 충족시키고 있음을 알 수 있다.

한편 1000만어절 이상의 규모만 고려한 TYPE, TOKEN, R_TT(=Rtype) 변수간의 상관계수는 아래의 도표에서 살펴보듯이 특히 토큰대비 타입의 백분율간의 관계가 −0.964로 매우 강한 음의 상관을 나타내었다. 따라서 본 데이터 분석결과로부터 대용량의 코퍼스 처리에서 약 1000만 어절 이상의 규모가 되어야 토큰 수 변화에 따른 타입개수의 변화가 안정적으로 변화하게 된다고 유추할 수 있겠다. 물론 이 결과는 장르를 구분하지 않은 데이터에 대한 결과이다.

맺 음 말

이 논문에서 보인 바와 같이 코퍼스의 크기 증가와 타입의 증가와의 상관성은 함수적 관계로 나타낼 수 있었다. 지금까지 '21세기 세종계획' 등에서 구축한 균형 코퍼스(balanced corpus)의 규모가 1000만 어절 규모인데 이 연구에 따르면, 그 규모는 통계적 처리에 비교적 적정 구축량임을 증명할 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 있다. 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.

REFERENCES

장석배(1999) : "말뭉치 규모와 어절 유형 증가간의 상관성에 대한 연구", 언어 정보의 탐구1, *pp159-210*

Yang DH, Lee IH, Cantos P(2002) : *"On the corpus size needed for compiling a comprehensive computational lexicon by automatic lexical acquisition", Computers and the Humanities 36, pp171-190*

McEnery T, Wilson A(1996) : *Corpus Linguistics*