한국어 코퍼스에서 나타나는 빈도효과 비교

정재범*,임희석**,남기춘*
* 고려대학교 심리학과
** 천안대학교 정보통신학부

A Comparison of Frequency Effects among Korean Corpus

Jaebum Jung*, Huisoek Lim**,Kichun Nam*
* Department of Psychology, Korea University
** Department of Information Communications, Chonan Universtity
E-mail : bluevet@korea.ac.kr

Abstract

This research studied the correlation of word frequency effect in Korean corpus. Experiment 1 showed that word frequency of each other corpus was significant correlated. Experiment 2 showed significant correlation between word frequency of each corpus and lexical decision time of participants. These results support that 4 corpus in this research should have stability to word frequency effect of participants

I. 서론

단어의 시각재인에 있어서 더 빠르고 쉽게 재인하는 데에 영향을 미치는 요소는 여러 가지가 있지만, 가장 연구가 많이 되고 그 효과가 예측가능 한 변인 중의 하나는 단어의 발생빈도문제이다. 단어 재인의 효율성을 결정하는데 사용되는 몇 가지 방법 중에서, 많은 연구자들은 고빈도 단어가 저빈도 단어보다 더 쉽게 처리된다는 것을 발견하였다[1]. 그 이후로 다양한 연구가 진행되어, 순간노출기법[2],단어읽기과제[3],어휘판단과제[4]등에서 단어빈도 효과를 관찰 하였다. 특히 언어심리연구의 비교적 초기에 단어빈도의 효과를 관찰하였기 때문에 수많은 단어재인 모형들은 모두 단어빈도 효과를 적절히 설명해야만 하였다. 예를 들면 탐색모형(Search model)에서는 단어는 빈도순서대로 bin이라는 곳에 저장되고, 어휘집에 접근시에는 고빈도순서로 탐색되기 때문에 단어빈도 효과가 나타난다고 설명하였다[3]. 다른 설명으로는 로고젠 모형(logogen model)이 있다[5]. 이 모형에서는 로고젠이라는 곳에 단어의 빈도정보가 저장되어있어 빈도가 높을 경우 단어 활성화가 저빈도 단어보다 빠르게 일어난다고 주장한다.

기존의 연구에서는 시각단어재인에서 빈도효과가 결정적인 영향을 미치는 것을 보고하고 있다. 그러나 빈도효과를 알기 위해서는 첫 번째로, 단어의 빈도와 둘째, 단어에 빈도에 사람들이 반응한 행동측정치가 필요하다. 여기서 단어의 빈도를 알기 위해서는 다시 첫째, 실험 참가자가 단어의 친숙도를 평정하는 방법과 둘째로, 기존에 미리 제작된 대규모 코퍼스(corpus)를 참조 하는 방법이 있다. 본 연구에서는 특히 국내에서 제작된 corpus를 이용하여 단어의 빈도를 추출하여 단어빈도 효과를 측정하는데 사용하였다. 특히 사전, 소설, 잡지, 신문 기사 등으로 이루어진 text corpus의 경우 인간의 언어생활을 그대로 반영한다고 할 수 있다. 그러나 국내의 여러기관에서 제작된 코퍼스는, 빈도를 세는 기준이나 형태소의 정의가 일부 통일 되어있지 않고, 원자료가 특정분야에 편향되어있는 등의 문제점이 아직 해결되어있지 않고, 서로 다른 특성을 보일 수 밖에 없다. 코퍼스에서 추출된 빈도는 각종 언어와 관련된 각종 실험에서, 반응시간에 가장 영향을 많이 끼칠 수 있는 변인이기 때문에, 대부분의 언어를 주제로 하는 실험재료는 빈도를 통제함으로써 결과의 안정성을 추구하고 있다. 다시 말하면, 각 코퍼스에서 추출된 빈도는 심리학에서 실험재료 구성과,

결과 분석에 전반적으로 영향을 미치기 때문에, 정확도와 신뢰도가 높아야 하고, 코퍼스에서 추출한 빈도를 실제 사람에게 적용하였을 때, 단어빈도효과를 잘 보여주어야 한다. 그러나 코퍼스의 특징 차이 때문에 코퍼스에서 추출한 빈도가 서로 차이가 나서 결과적으로 단어빈도 효과를 보여주지 못할 경우, 기존 실험심리적 언어연구 결과의 해석이나 실험재료 구성에 지대한 영향을 끼칠 것이다. 따라서 본 연구는 다음 두가지 질문을 해결하기 위해 실시 되었다. 첫째로 국내의 코퍼스에서 추출한 빈도는 서로 높은 상관을 보일 것인가와 둘째로 이러한 국내의 코퍼스에서 추출한 단어의 빈도는 실제 사람의 반응시간 자료와 얼마나 일치할 것인 가이다. 첫 번째 질문을 해결하기 위해 실험연구에서 많이 쓰이는 K대학,Y대학,E연구소,S기획의 코퍼스를 중심으로 단어의 빈도를 추출하여, 코퍼스간의 단순상관을 비교하고, 두번째로는 실험참가자의 반응시간 데이터와 국내 코퍼스에서 추출된 빈도와의 상관을 조사함으로서 간접적으로 단어빈도 효과를 추론하고자 한다. 만약 참가자의 반응시간과 추출한 빈도가 유의미한 역상관이 있다면, 국내의 코퍼스중 어떤 것이 참가자의 단어빈도효과를 더 잘 설명하는지 알아보고자한다.

II. 실험 1

실험 1에서는 기 선정된 단어에 대해 각 코퍼스가 인간의 언어생활을 충분히 반영하도록 제작되었다면, 일정한 단어에 대해 가 코퍼스에서 추출된 빈도가 서로 높은 상관 관계를 가질 것으로 예상된다. 그러나 각 코퍼스는 제작 방법과 기준, 그리고 코퍼스를 구성한 원자료의 차이가 존재함으로 관련성이 떨어질 가능성이 여전히 존재한다.

실험재료 및 설계 : 실험 재료로 Y대학의 코퍼스를 기준으로 200개의 단어를 선정하였다. Y대학의 코퍼스를 기준으로 한 이유는 현재 여러 언어관련 실험에서 가장 많이 사용되고, 가장 많은 어절에서 빈도를 추출했기 때문이다. 빈도의 효과를 확실히 반영하기 위해 200개의 단어중에 양 극단의(고/저빈도) 단어를 가중적으로 설계하고 표1에 제시하였다. 동일한 단어에 대해서 각각 E연구소, S기획,K대의 코퍼스에서 빈도를 추출하고 각각의 빈도를 1,000,000어절단위로 표준화하여 표 2에 제시하였다.

<표1> 실험재료의 구성

Y대코퍼스빈도	개수	평균	표준편차
7~15	50	9.9	2.7
15~390	125	204.0	108.6
1012~3221	24	2114.8	705
11284	1	11284	

<표2> 각 코퍼스의 특징

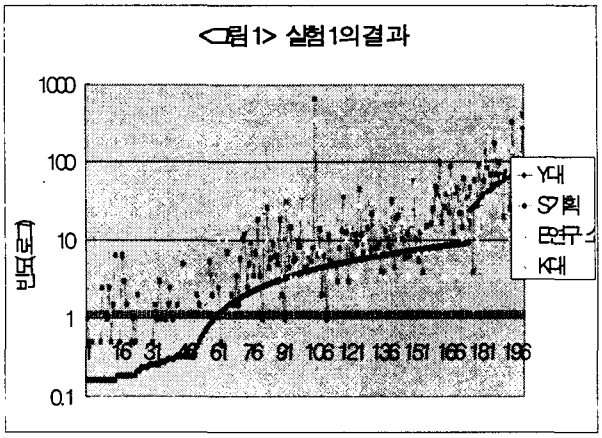
	Y대	S기획	E연구소	K대
총 어절수	42,464,052	2,039,315	288,291	151,328
미추출단어*	0	31	90	105
표준화상수**	0.02	0.49	3.46	6.6
평균	10.3	27.5	36.7	53.2
표준편차	24.4	66.8	58.1	89.3

* 코퍼스에서 출현하지 않은 단어
** 빈도를 1,000,000어절로 표준화할 때 필요한 상수

빈도추출에 기반이 되는 총 어절수는 Y대,S기획,E연구소,K대 순으로 많았다.

결과 및 논의 : 각 코퍼스의 빈도를 로그로 바꾸어 그림 1로 제시하고, 상관값과 유의도를 표 3에 제시하였다

<그림1> 실험1의결과



<표3> 실험 1의 결과

	Y대	S기획	E연구소	K대
Y대	1.0	0.57*	0.65.*	0.58*
S기획	0.57*	1.0	0.81*	0.52*
E연구소	0.65*	0.81*	1.0	0.54*
K대	0.58*	0.52*	0.54*	1.0

* $p < 0.01$로 유의미

실험 1의 결과는 비교 조건으로 선정한 모든 코퍼스의 빈도가 유의미한 상관관계가 있음을 보여준다. 이것은 절대적인 코퍼스 크기와, 세부적인 특징들이 다름에도 불구하고 각 코퍼스는 비슷한 특징을 보여 줌을 의미한다. 특히 E연구소와 S기획의 코퍼스는 가장 높은 상관관계를 보여주었고, S기획과 K대의 코퍼스는 가장 낮은 상관관계를 보여주었다. 특히 E연구소와 S기획의 코퍼스의 경우 사람이 직접 품사를 태깅(tagging)입력함으로서 코퍼스내에서 단어에 대한 빈도 계산이 다른 코퍼스 보다 더 충실했기 때문으로 분석된다. K대 코퍼스는 빈도를 계산하기 위한 총 어절수가 작고 미추출단어가 다른 코퍼스 보다 많이 발생했기 때문에 다른 코퍼스와의 상관이 미미하게 줄어들었다.

III. 실험 2

실험 2는 실험1의 코퍼스 빈도가 실제 사람에게도 적용 되는지를 알아보기 위하여 실시하였다. 실험 1에 사용된 코퍼스가 사람의 언어생활을 잘 반영한다고 가정하면, 실제 사람에게서 측정 가능한 몇몇 행동 지표가 코퍼스의 빈도와 높은 역상관을 보일 것으로 예상된다. 실험2 에서는 행동지표로 단어에 대한 어휘판단(Lexical decision) 시간을(Response time) 측정하였다.

참가자 : 고려대학교에 재학중인 정상시력을 가진 34명이 실험에 참가 하였다.

실험 재료 및 설계 : 실험 1에서 사용된 200개의 단어를 목표단어로 설정하고 어휘판단을 위해 비단어 200개를 포함한 총 400개의 어휘판단 세트를 만들었다.

실험 절차 : 실험2에서는 목표단어에 대한 어휘판단과제(lexical decision task)를 실시하였다. 참가자가 컴퓨터 앞에 앉으면, 실험 진행 방법에 대해 설명을 하고 연습 시행을 실시한 후 보충 설명을 하였다. 참가자는 각 시행에서 화면 중앙에 제시되는 목표 단어를 본 후, 컴퓨터 키보드의 '단어'키와 '비단어'키를 가능한 한 빠르고 정확히 누르도록 지시 받았다. 참가자는 양손 검지로 단어 키와 비단어 키를 누르게 하고, 참가자에 따라 단어 키를 비단어 키와 무선적으로 서로 바꾸었다. 실험이 진행되면서 자극이 제시된 후 참가자가 반응키를 누를 때까지의 시간을 측정하였다.. 참가자에게 검사 자극에 대한 판단 시간으로는 2초가 주어졌으며, 검사 자극에 대한 판단이 이루어지면 다음 시행으로 넘어갔다. 2초안에 반응을 하지 못했을 경우 결측자료(missing data)로 처리하고, 다음시행으로 넘어갔다. 참가자의 피로도를 고려하여 200개의 자극을 제시한후 5분의 휴식을 주고 나머지 200개의 자극을 제시하였다. 한 시행이 끝나고 다음시행이 시작될 때까지의 시행간 간격은 1.5초였고, 시행간 간격동안에는 '*' 모양을 한가운데에 제시하였다. 자극은 무선적인 순서로 17인치 컬러모니터에 제시되었다. 제어프로그램은 Superlab 2.0을 사용하였다.

실험 결과 및 논의 : 참가자의 어휘판단 시간을 정리하고, 결측 자료(missing data)를 처리하였다. 결측 자료는 두 종류로써 첫째는 단어를 비단어로 판단하거나 비단어를 단어로 판단한 것의 어휘판단 시간과 둘째, 전체의 어휘판단 시간의 평균과 표준편차를 구한 뒤 3 표준편차 이상이나 이하의 값들을 결측 자료로 처리하였다. 그 후 극단치의 영향을 배제하기 위하여 중앙치를 사용하였다. 결측 자료는 전체의 5.8 % 였다 참가자의 어휘판단 시간과 실험 1에서 추출한 코퍼스 빈도간의 상관관계를 표 4에 제시하였다.

<표4> 실험 1의 결과

	Y대	S기획	E연구소	K대
반응시간	-0.31*	-0.26*	-0.25.*	-0.13*

* $p < 0.01$로 유의미

각 코퍼스에서 추출한 빈도는 모두 참가자의 반응시간에 대해 유의미한 역상관을 보여주었다. 즉 코퍼스에서 추출한 빈도가 높아질수록 반응시간은 빨라지는 경향을 보여주었다. K대의 코퍼스가 다른 코퍼스에 비해 상관이 낮은 것은 실험 1에서와 마찬가지로 미추출된 단어가 많고 코퍼스를 구성하는 전체 어절 수가 작기 때문으로 분석된다.

결론적으로 각 코퍼스에서 추출된 빈도는 인간의 언어생활을 잘 반영하고, 단어빈도 효과를 예언할 수 있는 지표가 될 수 있음을 시사한다.

IV. 논의

기존연구에서 확고히 밝혀진 단어 빈도 효과를 측정하기 위해서는 먼저 단어의 빈도를 조사해야 한다. 단어의 빈도를 조사하는 것은 크게 단어의 친숙도를 평정한는 방법과 기존에 구성된 대규모 코퍼스에서 빈도를 찾는 방법이 있다. 만약 대규모 코퍼스에서 추출된 빈도가 유의미한 차이를 보인다면 단어 빈도 효과가

코퍼스 마다 틀리게 측정될 가능성이 있다. 이러한 가능성을 알아보기 위해 실험 1에서는 200개의 단어의 빈도를 추출하여 그 빈도간의 상관을 조사하였다. 결과는 국내 코퍼스에서 추출된 빈도는 코퍼스의 특성과 구성방법이 다름에도 불구하고 유의미하게 일치하는 경향을 보였다. 실험 2에서는 비교적 안정적인 국내 코퍼스 빈도가 실제 단어 반응시간과는 어떤 관련성이 있는가를 조사하였다. 결과는 역시 유의미했고, 역상관이 나타나면서 단어빈도효과를 보여주었다. 실험 1과 2를 종합하면 실험에 쓰인 코퍼스는 대부분 서로 안정적이였으며 반응시간자료를 효과적으로 설명할 수 있었다. 실험 2의 결과를 약간 다른 입장에서 해석해 보면 단어재인과정에는 빈도이외의 다른 변인들이 더 많은 영향을 끼칠 수 있다는 것을 시사한다.

이후의 실험으로는 실험 참가자들이 9점척도로 평정한 단어 친숙도를 가지고 반응시간과의 상관을 조사하고 그 결과와 실험2의 결과 중 어느 것이 단어 빈도효과를 더 많이 설명할 수 있는 지를 조사하는 것이 필요하다.

참고문헌

[1] Savin, H. B.(1963)."Word-frequency effect and errors in the perception of speech". *Journal of the Acoustical Society of America*, 35, 200-206.

[2] Jacoby, L. L., & Dallas, M. (1981). "On the relationship between autobiographical memory and perceptual learning", *Journal of Experimental Psychology : General*, 110,306-340

[3] Forster, K., & Chambers, S. M. (1973). "Lexical access and naming time". *Journal of Verbal Learning and Verbal Behavior*, 12, 627-635

[4] Rubenstein, H., Garfield, L., & Millikan, J. (1970). "Homographic entries in the internal lexicon". *Journal of Verbal Learning and Verbal Behavior*, 9, 487-494.

[5] Morton, J. (1982). Disintegrating the lexicon. In J. Mehler,E. T. C. Walker, & M. Garrett (Eds.), *Perspectives in mental representation* (pp. 89-109). Hillsdale, NJ: Erlbaum.