한국심리학회지 : 인지 및 생물 http://dx.doi.org/10.22172/cogbio.2017.29.4.004
The Korean Journal of Cognitive and Biological Psychology
2017, Vol. 29, No. 4, 395-410

한국어 심성어휘집 연구를 위한 어휘판단 데이터베이스*

이 광 오[1)†] 구 민 모[2)] 남 기 춘[2)] 박 기 남[3)]
박 태 진[4)] 배 성 봉[1)] 이 창 환[5)] 이 혜 원[6)] 조 증 열[7)]

[1)]영남대학교 심리학과 [2)]고려대학교 심리학과 [3)]고려대학교 정보창의교육연구소
[4)]전남대학교 심리학과 [5)]서강대학교 심리학과
[6)]이화여자대학교 심리학과 [7)]경남대학교 심리학과

한국어 심성어휘집의 구조와 검색 과정을 연구하기 위해 한국어 단어 30,930개와 비단어 30,930개에 대한 어휘판단시간을 수집하였다. 참가자는 52명이었고 각 참가자는 61,860개의 자극에 전부 반응하였다. 자극은 단일 단어 방식으로 제시되었으며, 참가자는 화면 중앙에 제시되는 자극이 단어인지 비단어인지 판단하였다. 수집된 데이터베이스의 신뢰성을 검증하기 위해 빈도, 연습, 품사, 길이 등의 영향을 분석하고 가상실험(virtual experiment)을 실시하였다. 두 개의 가상실험 결과는 실제실험 결과와 상당히 일치하였다. 결과를 바탕으로 한국어 심성어휘집 연구에서 메가스터디(megastudy)의 이용에 대해 논의하였다.

주제어 : 메가스터디, 단어재인, 어휘판단, 한글읽기

* 이 논문은 2012년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012S1A5A2A03034555).
† 교신저자 : 이광오, 영남대 심리학과, (38541) 경상북도 경산시 대학로 280
E-mail: yiko@yu.ac.kr

단어(word)는 언어를 구성하는 기본 요소 중 하나이다. 단어는 음운, 형태, 통사, 의미에 대한 정보를 모두 가지고 있다. 단어보다 작은 단위는 이런 정보 중 적어도 어떤 하나를 결여하고 있다. 단어는 문장을 구성하는 기본 단위로서 언어 사용자의 마음속에 중요한 층위를 구성한다.

현대 산업사회를 살고 있는, 정규 고등학교 교육을 받은 일반 사람들이 알고 있는 단어의 수는 5만개 정도로 추정된다(Miller, 1991). 개인이 이해하고 산출할 수 있는 단어의 집합을 심리학자들은 심성어휘집(mental lexicon) 또는 심성사전(mental dictionary)이라고 부른다(Aitchison, 2003). 심성어휘집의 각 단어는 고유의 발음, 표기, 통사, 의미 등의 정보를 포함하고 있으므로 심성어휘집은 하나의 방대한 데이터베이스라고 할 수 있다.

최근에 심성어휘집에 대한 연구는 두 가지 새로운 연구 방법에 의해 강한 영향을 받았다. 뇌영상화(neuroimaging) 방법은 활동 중 뇌를 실시간으로 촬영하여 분석한다. 기능적자기공명영상(fMRI), 양전자단층촬영(PET), 사건관련전위(ERP), 뇌자도(MEG) 등이 대표적이다. 다른 하나는 코퍼스(corpus)의 수집과 분석이다. 코퍼스 기반 연구는 컴퓨터 등장 이후 보편화된 것으로 최근의 빅데이터(big data) 연구 및 데이터사이언스(data science)와 같은 맥락 속에 있다. 심성어휘집 연구에서 이런 방향을 대표하는 것이 메가스터디(megastudy)이다.

메가스터디는 기존의 요인설계(factorial design) 연구에 대한 보완으로 등장하였다. 전형적 요인설계 연구는 200-300개의 자극과 20-30명의 참가자들을 대상으로 극히 통제된 실험을 실시한다. 실험자가 조작하는 변인은 2-3개에 불과하다. 단어재인에 영향을 줄 수 있는 무수히 많은 다른 변인을 모두 통제해야 하지만 사실상 불가능하다. 실험자가 예측하지 못하는 자극 변인 또는 상황이 실험 결과에 유의한 영향을 줄 수 있다. 메가스터디는 요인설계의 단점을 극복하기 위해 적어도 수천 개의 자극 또는 수백 명의 참가자를 사용한다. 메가스터디는 특정 변인이 단어재인에 미치는 영향을 다양한 자극 변인과 참가자 변인을 고려하여 분석한다. 메가스터디는 수많은 요인들이 상호작용하는 단어 재인의 전체 패턴을 파악하는 데 유리하다.

메가스터디가 표적으로 하는 행동 데이터는 어휘판단시간(lexical decision time), 명명시간(naming time), 안구고정시간(eye fixation time) 등을 포함한다. 행동 데이터를 분석하기 위한 자극 변인 중 많은 것은 코퍼스를 분석하여 얻어진다. 빈도, 길이, 품사, 발음규칙성, 다의성, 의미투명성 등의 객관적 어휘 특성이 그것이다. 아울러 친숙도, 습득연령, 구체성, 심상성 등의 주관적 어휘 특성도 분석에 사용된다. 주관적 어휘 특성은 언어 사용자의 단어에 대한 평가이며 심리적 차원을 나타낸다. 또한 참가자의 어휘력, 음운인식력, 형태소인식력, 독해력 등 개인차도 행동 데이터를 분석하기 위해 사용된다.

최초의 메가스터디는 Seidenberg와 Waters (1989)에 의한 것으로 알려져 있다. 2,897개의 단음절 단어가 사용되었으며 30명이 실험에 참가하였다. 1만개가 넘는 단어가 사용된 메가스터디는 2007년 발표된 “The English Lexicon Project(ELP)”가 처음이다(Balota, Yap, Cortese,

Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007). ELP는 4만개의 영어 단어와 4만개의 영어 비단어(nonword)에 대해 객관적 어휘 특성을 수집하고 또한 1천 명에 가까운 참가자로부터 어휘판단시간과 명명시간을 수집하였다(Table 2 참조).

이후, 프랑스어(Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, Augustinova, & Pallier, 2010), 네덜란드어(Keuleers, Diependaele, & Brysbaert, 2010), 영국 영어(Keuleers, Lacey, Rastle, & Brysbaert, 2011)에 대한 메가스터디가 뒤를 이었다. 일본어에서는 "NTT[1] 어휘 특성 연구 프로젝트"가 유명하다(Amano, Kasahara, & Kondo, 2007; Amano & Kondo, 1999, 2003). 중국어에 대해서는 Sze, Rickard Liow, 그리고 Yap(2014)와 Tse, Yap, Chan, Sze, Shaoul, 그리고 Lin(2017)의 연구가 있고, 말레이시아어에 대해서는 Yap, Rickard Liow, Jalil, 그리고 Faizal (2010)의 연구가 있다. 동아시아 언어에 대한 메가스터디는 아직 수적으로 열세이며, 유럽어를 대상으로 한 연구들에 비해 대상 단어수가 상대적으로 적다.

한국어에 대한 심리학적 연구에서 메가스터디는 아직 시도된 적이 없다. 본 연구는 한국어 심성어휘집에 대한 메가스터디의 일환이다. 본 연구는 한국어 단어 3만개에 대한 어휘판단시간을 측정하고 이를 데이터베이스로 만들어 연구자들과 공유하는 것이 목적이다.

방 법

본 연구는 한국어 단어에 대한 어휘판단시간 데이터를 수집한다. 어휘판단시간은 주어진 자극의 어휘성(lexicality)을 판단하는 데 걸리는 시간을 가리킨다. 어휘판단시간은 심성어휘집의 구조와 검색 특성을 반영하는 것으로 간주되며, 심성어휘집 연구에서 가장 많이 사용되는 지표이다. 어휘판단실험에서 실험참가자는 자극으로 주어진 문자열이 의미가 있는 단어인지 그렇지 않은지를 정확하고 신속하게 판단해야 한다. 단일 단어 어휘 판단(single-word lexical decision)은 가장 흔하고 단순한 방법으로 각 시행마다 한 개씩 자극 문자열이 제시된다. 이에 비해 점화 어휘 판단(primed lexical decision)은 점화어(primes)를 먼저 제시하고 이어서 표적 자극(targets)을 제시한다.

현재까지 대부분의 메가스터디는 단일 단어에 대한 어휘판단을 조사하였다. 메가스터디에 사용되는 단어수는 보통 2만에서 4만 사이이다(Table 2 참조). 실험 절차는 두 가지가 있다. 표집 조사 방식은 전체 조사 대상 단어 중 일부를 표집하여 참가자에게 제공한다. 1,000개에서 2,000개 사이의 단어가 각 참가자에게 주어진다. 이런 방식으로 하면 각 참가자의 부담은 적어지지만 많은 수의 참가자가 필요하게 된다. 만약에 전체 조사 단어가 4만개이고 참가자 1인당 2천개의 단어를 할당하면, 조사 단어 전체에 대해 1회 반응을 얻기 위해 필요한 인원은 20명이다. 만약에 한 단어 당 30회의 반응시간을 얻고자 하면 필요한 전체 인원은 600명이 된다.

표집 조사 방식은 장점이 있지만 다양한 개인 특성을 통제하기 어렵다는 문제점이 있다. 이에 대해 전집 조사 방식은 각 참가자

1) Nippon Telegraph and Telephone

가 대상 단어 전체에 대해 반응하도록 한다. Keuleers 등(2010)의 DLP(Dutch Lexicon Project)에서 처음 시도되었다. 참가자는 30명이었으며, 이들은 19,000개의 자극 단어(+19,000개의 비단어) 전부에 대해 반응하였다. 한 사람이 합계 38,000개의 자극 항목에 반응하는 데는 많은 시간이 소요된다. 시작 시기와 종료 시기에 반응 방식도 다를 수 있다. 시작에서 종료에 이르는 기간 동안 연습효과(practice effect)와 피로효과(fatigue effect)가 나타날 수 있다. 연습효과는 전반부보다 후반부에서 반응시간이 짧고 오반응이 적은 것을 가리킨다. 반면에 반복적인 단순 작업이 피로효과를 가져올 수 있다. 후반부로 갈수록 피로가 가중되어 오히려 반응시간이 길고 오반응이 많이 나타날 수 있다. 표집 조사 방식과 전집 조사 방식은 서로 다른 장단점을 가지고 있지만 최근의 연구들은 전집 조사 방식을 선호한다(Tse et al., 2017). 본 연구는 전집 조사 방식으로 한국어 단어에 대한 어휘판단시간을 수집한다.

참가자 실험 참가자들은 국내 소재 4개 종합대학에서 선발하였다. 수도권의 K대, E대, 영남권의 Y대, 호남권의 J대 등 4개 대학의 재학생 58명이 실험에 참가하였다. 참가자들은 모두 32시간 이상 소요되는 실험에 참여하였다. 실험은 1학기(6개월)를 넘지 않도록 하였다. 학기 중 자신에게 편리한 시간을 예약하여 실험실에 오도록 하였으며, 하루에 최대 4시간 한도로 실험에 참여하였다. 실험은 모두 124개의 블록으로 구성되었으며 모든 블록을 하나도 빠트리지 않고 반응하도록 하였다. 참가자 일인당 실험에 소요된 시간은 짧게는 3주, 길게는 6개월이었다. 실험 완료 후 사례를 지급하였다. 마지막 블록까지 모두 참가하지 않은 참가자들은 제외하였다. 최종적으로 52명의 자료를 확보하였다. 이들의 평균 연령은 21.9세(범위: 18세-25세)였고, 성별은 여자 29명 남자 23명이었다.

자극재료 Kang과 Kim(2009)이 세종계획21에서 얻은 약 1500만(15,268,017)개의 토큰 단어 빈도를 이용하여 자극을 선정하였다. 빈도(개별빈도 혹은 다의어의 경우에는 통합빈도)가 14 이하인 단어는 제외하였다. 빈도가 14이하인 단어는 100만 단어 단위로 환산하는 경우 빈도가 0에 근접하며, 친숙도가 지나치게 낮아 참가자들이 모르는 경우를 우려하였기 때

Table 1. Number of Stimulus Words in Different Length and Part of Speech

	Word length (in syllables)					Overall
	1	2	3	4	5	
Noun	528	12,125	7,303	1,554	186	21,696
Adverb	54	452	460	294	4	1,264
Verb	NA	263	877	4,195	891	6,226
Adjective	NA	41	255	1,230	218	1,744
Overall	582	12,881	8,895	7,273	1,299	

문이다. 고유명사, 조사, 감탄사, 은어, 속어 등 특수 범주의 단어도 제외하였다. 단어 길이는 1-5 글자(=음절)의 범위로 제한하였다. 품사는 명사, 동사, 부사, 형용사 등으로 국한하였으며, 동사와 형용사를 제시할 때에는 어미변화 없는 원형을 사용하였다. 세종코퍼스에 포함된 단어는 매우 많기 때문에 위의 요건을 모두 만족시키는 단어라도 전부 포함시킬 수는 없었으며 실험 실시의 편의를 위해 30,930개로 제한하였다. 단, 최종 선정된 단어들이 어종, 빈도, 품사, 길이 등에서 세종코퍼스의 구성과 일치되도록 노력하였다. Table 1에 단어 길이와 품사에 따른 자극단어 수를 제시하였다. 단어와 동일한 수의 비단어는 단어를 변형하거나 재배열하여 만들었다. 단어를 변형하는 방식은 단어 철자 중 자모 하나만 다른 것으로 바꾸는 것이었다. 이 방식은 한 글자 비단어를 만들 때만 사용되었다. 재배열 방식은 두 글자 이상의 비단어를 만들 때 적용하였다. 우선 글자 길이에 따라 단어들을 구분한 후, 동일한 위치의 글자들을 무선적으로 재배열하였다. 예를 들면, 어두 글자는 어두 글자 간에 재배열되었으며 어말 글자는 어말 글자 간에 재배열되었다. 재배열한 결과 단어가 된 경우에는 그런 것들만 모아서 재배열 절차를 다시 적용하였다.

장비 실험에는 IBM PC/AT 호환기종인 펜티엄급 개인용 컴퓨터, 해상도가 1024x768화소인 17인치 모니터(LG Flatron 795FT), VGA 그래픽 어댑터 등이 사용되었다. 자극의 제시, 반응의 측정, 실험의 통제에는 Forster와 Forster(2003)가 개발한 실험 생성 소프트웨어 DMDX를 이용하였다. 참가자의 반응은, 반응 버튼 상자를 통해, PC에 장착된 병렬입출력보드(MC PCI-DIO 24)에 입력되었다. 사용된 장비는 4개 대학에서 동일하였다.

절차 단어 또는 비단어는 PC 모니터 화면의 정중앙에 제시되었다. 글자 크기는 30포인트였고 고딕체 폰트가 사용되었다. 각 시행은 화면 중앙에 응시점(십자모양 '+')이 나타나는 것으로 시작되었다. 응시점이 나타나면 참가자는 화면에 집중하도록 지시하였다. 응시점은 200ms 이후에 사라지고 이어서 200ms 공백 후에 표적 자극이 나타났다. 참가자는 표적 자극이 의미가 있는 단어인지 아닌지를 판단하여, 단어이면 버튼박스의 오른쪽 버튼을 누르고, 비단어이면 왼쪽 버튼을 누르도록 하였다. 판단은 정확하게 버튼 누르기는 가능한 한 빨리 할 것을 지시하였다. 참가자가 버튼을 누르면 그것으로 시행이 종료되고 400ms 후 다음 시행이 시작되었다. 실험은 블록 단위로 실시되었으며 한 개의 블록은 512회의 시행으로 구성되었다. 블록은 도중에 정지할 수 없었으며, 중간에 두 번의 짧은 휴식이 제공되었다. 한 블록에 평균 30분 정도가 소요되었다. 참가자는 블록과 블록 사이에 충분한 휴식을 취하도록 요구되었다. 적절한 휴식이 없이 연달아 2블록 이상 실시하는 것은 금지되었다. 하루에 최대 4시간 이상을 참가하지 못하게 하였다(예: 오전 2시간, 오후 2시간).

결 과

본 연구에서 수집된 어휘판단시간은

3,213,716개였다. 평균 오반응률은 9.3%였으며, 단어에서 10.7%, 비단어에서 7.8%로 나타났다. 평균 오반응률이 25%를 넘는 참가자는 제외되었다. 제외된 참가자수는 6명이었다. 반응시간이 300ms보다 짧거나 1500ms보다 긴 시행은 오류로 간주하여 제외되었다. 참가자들의 평균 반응시간은 643ms였다(Table 2 참조). Table 2에 본 연구와 ELP, FLP, DLP의 일반적 특성을 제시하였다. 본 연구(KLP)는 자극 구성과 반응 결과에서 다른 메가스터디들과 큰 차이를 나타내지 않는 것으로 보인다.

비단어에 비해 단어는 어휘판단시간이 더 짧았으나(667ms 대 620ms) 반대로 오반응률은 더 높은 경향을 보였다(7.8% 대 10.7%). 이것은 속도-정확 교환(speed-accuracy tradeoff)의 개입을 의심하게 한다. 속도-정확 교환이 있었는지 확인하기 위하여 자극별로 평균 반응시간과 오반응률을 계산하고 상관을 구하였다. 반응시간과 오반응률의 상관은 단어에서 +.73이었고 비단어에서 +.66이었으며 둘 다 통계적으로 유의하였다($p<.0001$). 상관계수가 양수인 것은 반응시간이 길수록 오류율이 높다는 것이며, 이는 속도-정확 교환의 개입 가능성을 배제할 수 있음을 의미한다.

참가자 특성에 따른 수행의 차이는 유의하지 않았다. 남자에 비해 여자가 더 빠른 반응을 보였으나(628ms 대 601ms), 통계적으로 유의한 차이는 아니었다, $F(1, 50)=1.846$, n.s. 또한 학교 간 차이도 유의하지 않았다.

메가스터디에서 얻은 데이터베이스를 분석

Table 2. Descriptive Statistics of the Korean Lexicon Project, English Lexicon Project, French Lexicon Project, and Dutch Lexicon Project. The ELP, FLP, and DLP Statistics Were Adapted from Keuleers et al. (2010)

	KLP	ELP	FLP	DLP
Number of words	30,930	40,481	38,840	14,089
Length in letters	7.1 (2-14)	8.0 (1-21)	8.5 (1-21)	6.3 (2-12)
Length in syllables	2.9 (1-5)	2.5 (1-8)	2.5 (1-8)	1.8 (1-2)
Frequency per million	23.9 (0.13-13,213)	25.2 (0.02-41,857)	21.1 (0-25,988)	59.7 (0.02-39,883)
Percent error for words	10.7% (0-100%)	16% (0-100%)	9% (0-98%)	16% (0-100%)
RT for words	620 ms (360-1,440 ms)	784 ms (415-1,755 ms)	740 ms (515-1,464 ms)	654 ms (312-1,382 ms)
Percent error for nonwords	7.8% (0-100%)	12% (0-100%)	7% (0-98%)	6% (0-98%)
RT for nonwords	667 ms (446-1,366 ms)	856 ms (508-1,135 ms)	807 ms (589-1,814 ms)	674 ms (519-1,604 ms)

Note. in brackets: the range of the variables

하는 방법은 연구 가설이나 목적에 따라 달라질 수 있다. 본 논문은 특정 연구 가설의 검증을 위한 것이 아니고 데이터베이스의 존재를 공개하는 것이 목적이므로 데이터베이스의 일반적 특성을 보여주는 초보적 분석에만 집중하였다. 빈도, 연습, 길이, 품사 등의 영향을 분석하였다.

빈도효과 단어 재인이 빈도에 의해 영향을 받는다는 사실은 오래전부터 알려졌다(Rayner, Pollatsek, Clifton, & Ashby, 2012). 빈도가 높을수록 단어는 더 빨리, 더 정확하게 재인될 수 있다. 단어 빈도 효과는 심성어휘집 연구에게 가장 친숙하고 확실한 현상이다. Figure 1은 어휘판단시간에 미치는 빈도 효과를 보여준다. 여기에 사용된 빈도는 1,500만 어절을 분석하여 얻은 '세종코퍼스 빈도'이다(강범모, 김흥규, 2009). Figure 1은 단어의 어휘판단시간이 로그빈도에 비례하여 감소하는 전형적 패턴을 잘 보여준다(Keuleers et al., 2010). 회귀선의 기울기는 -38.03이었으며, 빈도는 전체 변량의 14.3%를 설명하는 것으로 나타났다. 빈도를 독립변인으로 하는 선형모형은 유의한 것으로 나타났다, $F(1, 27397)=4557$, $p < .0001$.

연습효과 본 연구는 모두 124개의 블록으로 구성되었으며 최초 블록에서 시작하여 최종 블록을 수행하는 데 최소 3주, 최대 6개월이 소요되었다. 나중의 블록으로 갈수록 연습효과(또는 피로효과)가 나타날 수 있으며, 이에 대한 고려는 데이터베이스의 분석에서 중요한 사항이 된다. Keuleers 등(2010)은 DLP에서 처음으로 블록 효과에 대한 문제를 제기하였다. 그들은 블록 진행에 따른 수행의 변화를 분석하여 상당히 선명한 연습효과를 찾아내었다. 그러나 연습효과는 반응시간에서만 뚜렷했고 정확률에서는 약하였다. 오히려 단어 자극에 대한 정확률에서는 피로효과의 경향이 나타났다.

본 연구의 결과는 Keuleers 등(2010)과 다른 패턴을 나타냈다. Figure 2와 Figure 3에서 볼 수 있는 것처럼 반응시간과 오반응률에서 본 연구의 참가자들은 뚜렷한 연습효과 양상을 나타내지 않았다. DLP에서는 초기는 물론 후

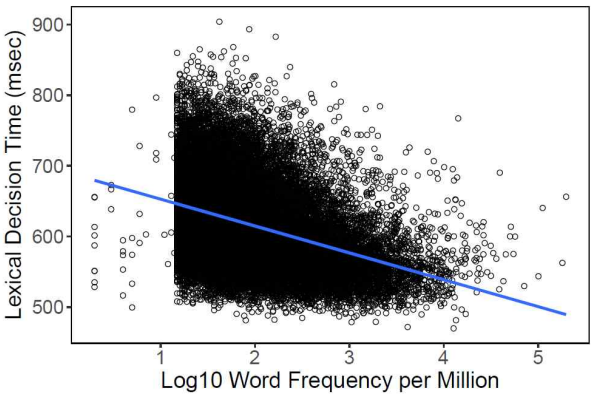

Figure. 1. Effects of Frequency on RTs (msec)

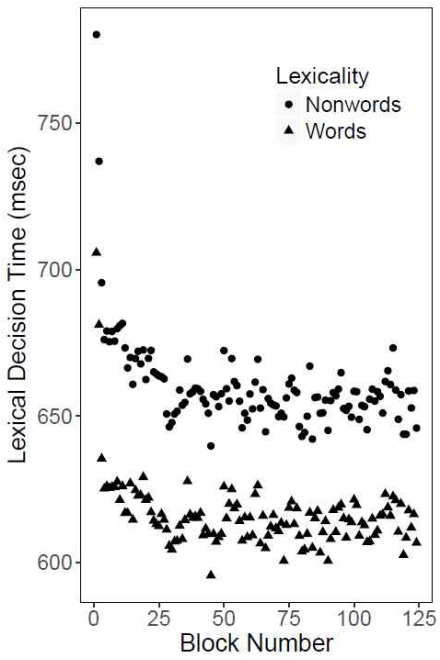

Figure. 2. Effects of Practice on RTs

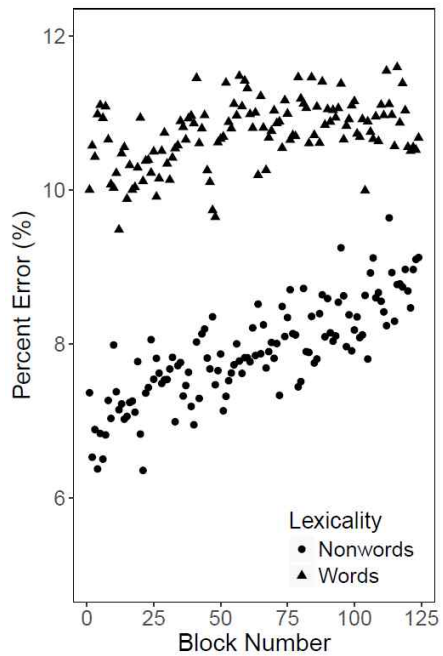

Figure. 3. Effects of Practice on Percent Errors

기 블록에서도 연습효과가 크게 나타났으나, 본 연구에서는 초기 20블록까지만 반응시간이 짧아지는 연습효과가 나타났다. 그 후에는 블록 진행에 따른 반응시간의 감소는 나타나지 않았다. 이것은 본 연구의 참가자들이 어휘판단 실험에 상대적으로 빨리 적응하였음을 시사한다. 반면에 오반응률에서는 연습효과가 거의 나타나지 않았다. 이것은 DLP의 결과와 유사하다. 특이한 것은 비단어에 대한 오반응률에서 피로효과가 나타난 것이다. 초기 블록에서 후기 블록으로 이동하면서 비단어에 대한 오반응률은 계속 증가하였다. 반면에 단어에 대한 오반응률에서는 피로효과가 상대적으로 작았다.

블록 효과에서 언어 간 차이에 대한 설명은 여러 가지 가능성이 있다. 가장 먼저 실험 환경과 참가자 특성의 차이를 들 수 있다. 더 쾌적한 환경과 여유 있는 스케줄은 과제에 대한 집중도를 높이고 유의한 연습효과를 가져올 수 있다. 또한 참가자들의 성실성과 집중력도 피로효과를 방지하는 데 중요하다. 그러나 본 연구의 참가자들이 DLP 참가자보다 실험 환경과 집중력에서 부족함이 있었다고 판단할 근거는 없다. 오히려 초기부터 후기까지 계속하여 연습효과가 나타난 DLP의 결과는 설명이 필요하다. 일반적으로 초기의 연습효과와 후기의 피로효과가 자연스런 현상이기 때문이다.

길이효과 단어들은 자모수, 음절수 등에서 서로 다르다. 단어 길이는 단어 재인에 어떤 영향을 주는가? 일반적으로 길이가 긴 단어의 재인은 길이가 짧은 단어에 비해 느린데 이것

을 억제적 단어 길이 효과라 한다. 영어를 비롯한 많은 연구들은 억제적 단어 길이 효과를 지지한다. 단어 길이가 길수록 재인이 우수한 현상을 가리키는 촉진적 단어 길이 효과는 보고되지 않았다. 그러나 최근의 메가스터디는 촉진적 효과와 억제적 효과를 모두 보여주고 있다. New, Ferrand, Pallier, 그리고 Brysbaert (2006)는 ELP 자료(Balota et al., 2007)를 분석하여 영어에서 단어 길이 효과가 U자형 곡선을 그린다고 보고하였다. 5개 이하의 문자로 구성된 단어들에서는 촉진적 길이 효과가 나타나고 8자 이상에서는 억제적 효과가 나타나고, 5-8 사이의 문자로 이루어진 단어에서는 촉진 효과도 억제효과도 나타나지 않았다.

본 연구의 결과도 U자 모양의 단어 길이 효과를 보여준다. Table 3은 글자(=음절)를 단위로 사용하여 단어길이와 반응시간의 관계를 보인 것이다. 1글자에서 3글자까지는 촉진효과가 나타나고 3글자 이후는 억제효과가 나타났다, $R(4, 27197)=76.97$, $p<.0001$. 이런 패턴의 길이효과는 선행연구에서 이미 시사되었다. '한 글자 열등 효과'라고 불리는 현상이 그것인데, 짧은 단어가 긴 단어보다 수행이 나쁠 수 있음을 가리킨다(Bae, Park, Lee, & Yi, 2016; Park, 1993). 길이효과는 사용되는 단위에 따라 달라질 수 있다. 한국어의 경우, 획, 자모, 글자 등에 따라 길이효과가 다르게 나타날 수 있다(Nam, Seo, Choi, Lee, Kim, & Lee, 1997). 한국어 단어의 읽기에서 길이효과는 흥미있는 연구 주제이며 본 데이터베이스는 단어길이효과를 해명하는 데 유용할 것으로 기대된다.

품사에 따른 차이 본 연구에는 네 종류의 서로 다른 품사(명사, 동사, 형용사, 부사)가 사용되었다. 선행연구들이 단일 품사, 특히 명사에 집중한 것과 대비된다. Table 4에 품사에 따른 반응시간을 제시하였다. 부사와 명사에

Table 3. Effects of Word Length (in Syllable) on Lexical Decision Times

	Word Length (in Kulja)				
	1	2	3	4	5
RT	622	618	605	619	623
SD	60	62	56	59	56

Note. RT: response times; SD: standard deviation

Table 4. Lexical Decision Times (msec) According to Part of Speech (POS)

	POS			
	adverb	noun	verb	adjective
RT	606	611	624	625
SD	55	60	59	62

Note. RT: response times (msec); SD: standard deviations

대한 반응시간은 짧고 동사와 형용사에 대한 반응시간은 길었다. 이런 차이는 유의한 것으로 나타났다, $F(3, 27198)=94.41$, $p<.0001$. Scheffe 검사를 사용하여 평균 간 다중 비교를 실시한 결과, 부사와 명사는 동사와 형용사보다 반응시간이 짧았다. 부사와 명사 간에는 유의한 차이가 없었고, 동사와 형용사 간에도 유의한 차이가 없었다($p<.01$). 품사에 따라 반응시간의 차이가 나타난 것은 흥미로운 결과이다. 이런 결과는 품사에 따라 심성 어휘 표상과 처리에 차이가 있다는 견해(예를 들어, Vigliocco, Vinson, Druks, Barber, & Cappa, 2011)를 지지한다.

가상실험 가상실험(virtual experiment)은 메가스터디에서 수집된 행동 데이터의 신뢰성을 검증하는 방법 중 하나이다. 가상실험은 실제의 실험 결과와 메가스터디의 데이터 분석 결과를 비교한다. 이를 위해 실제실험에 사용된 자극 단어를 확보하는 것이 필요하며, 데이터베이스에서 해당 단어의 행동 데이터(어휘판단시간)를 인출한 후 통계적 분석을 실시한다.

가상실험을 실시하기 위해 최근에 학술지에 발표된 연구들을 검토하였다. 선택을 위한 필요조건은 실험에 사용된 자극이 모두 공개되어 있어야 하는 것이다. 또한 단일 단어 실험의 경우에만 가능하다. 실제실험이 점화 과제인 경우에는 가상실험은 불가능하다. 아래에서는 두 개의 가상실험의 결과를 보고한다.

먼저 Lee, Choi, Kim, 그리고 Kim(2011)에서 사용된 자극단어에 대한 반응시간과 오반응률을 본 데이터베이스에서 추출하여 가상실험을 실시하였다. 원래 연구는 단어습득연령이 단어재인에 미치는 영향을 조사한 3개의 실험으로 구성되었다. 그 중 어휘판단시간을 측정한 실험 3에 대해 가상실험이 가능하였다. 실험은 단어습득연령과 단어빈도를 조작한 2x2 요인설계였으며 조건별 단어수는 25개였다.

Table 5는 실제실험 결과로 원논문에서 인용한 것이며 Table 6은 가상실험 결과이다. 가상실험에 사용된 단어는 실제실험의 단어 100개 중 94개였다(6개는 본 데이터베이스에 포함되지 않은 단어였다). 가상실험의 결과는 실제실험 결과와 유사한 패턴을 나타냈다. 전체적으로 가상실험에서 반응시간이 더 길었으나(566ms 대 580ms) 오반응률은 유사하였다

Table 5. The Original Results of Lee et al. (2011): Mean Lexical Decision Times (msec) and Percentage Errors (%) as a Function of Frequency and AoA (Age of Acquisition)

Word frequency	AoA				RT Difference
	Early acquisition		Late acquisition		
	RT	PE	RT	PE	
High	521 (84)	0.25 (0.98)	569 (93)	1.38 (1.93)	48
Low	562 (94)	2.13 (2.69)	612 (89)	4.00 (4.06)	50

Note. RT: response times; PE: percentage errors.

Table 6. The Results of a Virtual Experiment for Lee et al. (2011)

Word frequency	AoA				RT Difference
	Early acquisition		Late acquisition		
	RT	PE	RT	PE	
High	542 (69)	0.7 (2.0)	589 (81)	1.8 (2.9)	47
Low	583 (70)	2.4 (3.0)	607 (82)	2.8 (4.2)	34

Note. RT: response times; PE: percentage errors.

(1.94% 대 1.92%). 다만 가상실험에서는 후기 습득 저빈도 단어 조건에서 원래 실험보다 오반응율이 낮게 나타난 것이 달랐다.

실제실험의 결과 분석에서는 빈도 효과가 유의하고, $F_1(1, 31) = 9.86$, $p < .01$; $F_2(1, 96) = 31.01$, $p < .0001$, 습득연령의 주효과도 유의하였다, $F_1(1,31) = 113.08$, $p < .0001$; $F_2(1, 96) = 44.05$, $p < .0001$. 상호작용은 유의하지 않았다. 가상실험 결과 분석에서는, 빈도 효과가 유의하였고, $F_1(1, 51) = 34.54$, $p < .0001$; $F_2(1, 93) = 15.30$, $p < .001$, 습득연령 효과도 유의하였다 $F_1(1, 51) = 56.14$, $p < .0001$; $F_2(1, 93) = 22.00$, $p < .0001$. 빈도와 습득연령 간의 상호작용은 F_1에서만 유의하였다, $F_1(1, 51) = 5.93$, $p = 0.0184$.

전체적으로 가상실험은 실제실험과 상당히 유사한 패턴을 보였다. 한 가지 차이가 있다면, 실제실험에서는 빈도와 습득연령의 상호작용이 유의하지 않았지만, 가상실험에서는 F_1 분석에서 상호작용이 유의하게 나타났다는 것이다. 이런 차이가 난 이유를 본 논문에서 설명하기는 어렵다(본 가상실험에서는 자극단어가 6개 적었다는 것도 이유가 될 수 있다.). 현재로서는 습득연령을 다룬 한국어 실험이 한 개밖에 없기 때문에 문제의 정확한 이해를 위해서는 추후 관찰과 검증을 기다릴 필요가 있다고 생각된다.

또 다른 가상실험은 Bae, Yi, 그리고 Park (2012)의 실험 1을 대상으로 하였다. 실제실험은 한자어의 빈도와 의미투명성을 조작하고 단일 단어 어휘판단을 과제로 사용하였다. Table 7에 실제실험 결과를 제시하고 Table 8에 가상실험 결과를 제시하였다. 가상실험에는 실제실험 자극 80개 중 74개가 사용되었다. 반응시간은 실제실험보다 가상실험에서 더 짧게 나타났고(626ms 대 563ms) 오반응률도 가상실험에서 더 낮았다(7.9% 대 4.8%). 두드러진 차이점은 저빈도 단어 조건의 수행이 가상실험에서 훨씬 우월하게 나타난 것이다.

실제실험의 결과 보고에서는, 단어빈도의 주효과가 유의하고, $F_1(1, 46) = 273.11$, $p < .0001$; $F_2(1, 79) = 87.43$, $p < .0001$, 의미투명성의 주효과는 F_1에서만 유의하였다, $F_1(1, 46) = 6.32$, $p < .05$. 상호작용은 F_1과 F_2에서 유의하였다, $F_1(1, 46) = 35.67$, $p < .0001$; $F_2(1, 79) = 5.72$, $p < .05$. 상호작용이 유의했던 이유는

Table 7. The Original Results of Bae et al. (2012): Mean Lexical Decision Times (msec) and Error Rates (%) as a Function of Frequency and Semantic Transparency

	Semantically opaque		Semantically transparent	
	RT	PE	RT	PE
Low frequency	713 (101)	23.5	671 (91)	6.6
High frequency	551 (72)	0.6	568 (74)	0.9
Total	632 (119)	12.1	620 (97)	3.7

Note. RT: response times; PE: percentage errors.

Table 8. The Results of the Virtual Experiment for Bae et al. (2012)

	Semantically opaque		Semantically transparent	
	RT	PE	RT	PE
Low frequency	605 (60)	11.1	585 (62)	4.7
High frequency	532 (60)	2.0	530 (63)	1.2
Total	568 (70)	6.5	558 (68)	3.0

Note. RT: response times; PE: percentage errors.

의미투명성의 효과가 저빈도 단어에서는 크고 고빈도 단어에서는 작았고, 또한 효과의 방향이 반대였기 때문이다.

가상실험에서는 단어빈도의 주효과가 유의하고, $F_1(1, 51) = 248.80$, $p < .0001$; $F_2(1, 73) = 69.59$, $p < .0001$, 의미투명성의 주효과는 (실제실험에서처럼) F_1에서만 유의하였다, $F_1(1, 51) = 5.35$, $p < .05$. 상호작용은 F_1 분석에서만 유의하였다, $F_1(1, 51) = 4.66$, $p < .05$.

의미투명성 효과에 대해서도 가상실험은 실제실험과 유사한 결과를 산출하였다. 두 실험에서 의미투명성 효과는 저빈도 단어에서 크게 나타났다. 두 실험의 차이는 고빈도 단어에서 나타났다. 실제실험에서는 의미투명성 효과는 아주 작았을 뿐만 아니라 방향도 저빈도 단어 조건과 반대로 나왔다. 반면에 가상실험에서는 의미투명성 효과가 거의 사라졌다. 이런 차이가 나타난 이유는 현재로서는 정확히 알 수 없다. 가상실험은 실제실험보다 자극수가 6개 적었는데 그 때문일 수도 있다. 또한 참가자 특성, 비단어 유형, 자극 목록, 실험 맥락 등에서도 두 실험은 서로 달랐다. 가상실험과 실제실험의 결과가 다르게 나온 경우 어떻게 할 것이냐의 문제는 연구자들 사이에 아직 합의된 바가 없다. 앞으로 많은 검증과 논의가 요청되는 부분이다.

논 의

본 연구는 한국어 단어 30,930개와 비단어 30,930개, 합계 61,860개의 언어 자극에 대해

참가자 합계 52명으로부터 어휘판단시간을 수집하여 데이터베이스를 만들었다. 품사, 빈도, 길이, 의미성 등에서 다양한 단어들이 포함되었다. 최종적으로 데이터베이스에 수집된 어휘판단반응은 320만개를 초과하였다.

본 연구에서 수집된 데이터의 특성을 보이기 위해 빈도, 연습, 길이, 품사 등이 미치는 영향을 분석하였다. 어휘판단시간은 로그빈도에 반비례하는 전형적 패턴을 나타냈다. 초기 블록에서는 연습효과가 나타났으며 후기 블록으로 가면서 피로효과가 나타났다. 단어길이 효과는 U자 모양을 나타냈으며, 동사와 형용사는 명사에 비해 반응시간이 긴 것으로 나타났다. 아울러 두 개의 가상실험을 실시하였으며, 가상실험 결과는 실제실험 결과와 상당히 유사한 것으로 나타났다. 이런 결과들은 본 연구에서 수집한 메가데이터가 상당히 신뢰롭다는 것을 가리킨다.

앞으로의 과제로서 본 연구에서 수집된 어휘특성 데이터베이스의 신뢰성을 검증하고 활용 가능성을 보여주기 위해 다양한 분석과 가상실험(virtual experiment)을 실시할 필요가 있다. 아울러 본 데이터베이스의 유용성을 보여주기 위한 새로운 분석 방법들도 시도되어야 한다. 그것은 한국어 단어의 주관적, 객관적 특성들을 이용하여 본 데이터베이스에 포함된 어휘판단시간을 분석하는 작업이며, 예를 들어 선형혼합효과모형(linear mixed effect model)을 이용한 분석이 기대된다.

본 연구의 제한점은 다음과 같다. 우선 참가자들이 대학생에 국한되어 있다. 연령에 따라 심성어휘집의 크기와 검색 방법이 달라질 수 있다. 중고등 학생과 노인들을 포함시킨다면 한국어 화자에서 심성어휘집의 형성과 변화를 이해하는 데 크게 기여할 수 있을 것이다. 또한 본 연구에서 사용한 어휘의 제한점도 있다. 연구의 편의상 30,930개의 단어에 국한하였으나 더 중요한 어휘와 다양한 어휘가 포함되지 않았을 가능성이 있다. 물론 다른 언어와 비교하여 적지않은 수의 단어를 조사하였지만 영어 연구인 ELP의 4만 단어에 비해서는 적었다. 또한 본 연구는 어휘판단시간을 대상으로 하였지만, 명명시간, 안구고정시간 등도 심성어휘집의 메가스터디에서 중요한 측정치들이다. 또한 단일 단어에 대한 반응 이외에 점화 실험 데이터도 포함할 필요가 있다. 다른 언어에서는 이미 그런 시도가 이루어지고 있다(Adelman et al., 2014; Yap et al., 2016).

[한국어 심성어휘집 어휘판단 데이터베이스는 KLP(the Korean Lexicon Project) 홈페이지에서 내려 받아 사용할 수 있다. http://klexicon.org]

References

Adelman, J. S., Johnson, R. L., McCormick, S. E., McKague, M., Kinoshita, S., Bowers, J., Perry, J. R., Lupker, S. J., Forster, K. I., Cortese, M. J., Scaltritti, M., Aschenbrenner, A. J., Coane, J. H., White, L., Yap, M. J., Davis, C., Kim, J., & Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods, 46,* 1052-1067.

Aitchison, J. (2003). *Words in the Mind: An Introduction to the Mental Lexicon.* Malden, MA:

Blackwell.

Amano, S., Kasahara, K., & Kondo, T. (2007). Reliability of familiarity rating of ordinary Japanese words for different years and places. *Behavior Research Methods 39,* 1008-1011.

Amano, S., & Kondo, T. (1999). *Nihongo-no Goi-Tokusei (Lexical properties of Japanese) Vol. 1.* Tokyo: Sanseido.

Amano, S., & Kondo, T. (2003). *Nihongo-no Goi-Tokusei (Lexical properties of Japanese) Vol. 1-6* (CD-ROM version). Tokyo: Sanseido.

Bae, S., Park, K., Lee, H., & Yi, K. (2016). The mono-syllabic word inferiority effect within Korean word recognition. *The Journal of Linguistic Science, 77,* 109-125.

Bae, S., & Yi, K., & Park, H. (2012). Semantic transparency effects in the recognition and learning of Sino-Korean words. *The Korean Journal of Educational Psychology, 26*, 607-620.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283-316.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445-459. http://elexicon.wustl.edu/

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms?. *The Quarterly Journal of Experimental Psychology, 64,* 545-559.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). New York: Academic Press.

Ferrand, L., New, B., Brysbaert, M, Keuleers, E., Bonin, P., Méot, A., Augustinova, M. & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42,* 488-496.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments,* & *Computers, 35*, 116-124.

Huey, E. (1908/1968). *The Psychology and Pedagogy of Reading* (Reprint). MIT Press 1968 (originally published 1908).

Juphard, A., Carbonnel, S., Ans, B., & Valdois, S. (2006). Length effect in naming and lexical decision: the multitrace memory model's account. *Current Psychology Letters: Behaviour, Brain* & *Cognition, 19*(2).

Kang, B., & Kim, H. (2009). *Frequency of Korean Usage.* Seoul: Hankookmunhwasa.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1,* 1-15.

Keuleers, E., Lacey, P., Rastle K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*, 287-304.

Lee, H., Choi, J., Kim, Y., & Kim, S. (2011). The age of acquisition effects in Korean word recognition. *The Korean Journal of Cognitive and Biological Psychology, 23*, 465-485.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39*, 92-98.

Miller, G. A. (1991). *The Science of Words*. New York: W. H. Freeman & Co.

Nam, K., Seo, K., Choi, K., Lee, K., Kim, T., & Lee, M. (1997). Word length effect on Hangul word recognition. *Korean Journal of Experimental and Cognitive Psychology, 9*, 1-18.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: new evidence from the English Lexicon Project. *Psychonomic Bulletin and Review, 13*, 45-52.

Park, K. (1993). Mental code involved in Hangul word recognition. *Korean Journal of Experimental and Cognitive Psychology, 5,* 40-55.

Rayner, K., Pollatsek, A., Clifton, C., & Ashby, J. (2012). *The Psychology of Reading,* 2nd Edition. New York: Psychology Press.

Seidenberg, M. S., & Waters, G. S. (1989). Word recognition and naming: A mega study. *Bulletin of Psychonomic Society, 27,* 489.

Sze, W. P., Rickard Liow, S. J. R., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods, 46*, 263-273.

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods, 49*, 1503-1519.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews, 35,* 407-426.

Yap, M. J., Hutchison, K. A., & Tan, L. C. (2016). Individual differences in semantic priming performance: Insights from the Semantic Priming Project. In M. N. Jones (Ed.), *Big data in cognitive science: From methods to insights.* New York: Psychology Press.

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods, 42,* 992-1003.

1 차원고접수 : 2017. 10. 05
수정원고접수 : 2017. 10. 16
최종게재결정 : 2017. 10. 16

한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2017, Vol. 29, No. 4, 395-410

The Korean Lexicon Project: A Lexical Decision Study on 30,930 Korean Words and Nonwords

Kwangoh Yi[1] *Min-Mo Koo*[2] *Kichun Nam*[2] *Kinam Park*[3]
Taejin Park[4] *Sungbong Bae*[1] *Chang H. Lee*[5] *Hye-Won Lee*[6] *Jeung-Ryeul Cho*[7]

[1]Department of Psychology, Yeungnam University
[2]Department of Psychology, Korea University
[3]Creative Information & Computer Institute, Korea University
[4]Department of Psychology, Chonnam University
[5]Department of Psychology, Sogang University
[6]Department of Psychology, Ewha Womans University
[7]Department of Psychology, Kyungnam University

Fifty-two Korean students were recruited for the first Korean megastudy to construct a database of lexical decision times on 30,930 Korean words and nonwords. The stimuli were 1-5 syllable in length and included nouns, verbs, adjectives, and adverbs. Participants were asked to decide if a stimulus on the screen was a word or a nonword. Each participant responded to a total of 30,930 words and 30,930 nonwords. The analysis of the lexical decision data obtained showed significant effects of frequency and length. Response times were also affected by POS, supporting the proposal to take into account of POS in the study of the mental lexicon. Practice effect was strong only in early blocks, and some signs of fatigue were apparent at later blocks. Two virtual experiments replicated the results of actual experiments, showing that the database was reliable. Based on the results, the use of megastudy in the understanding of the Korean mental lexicon was proposed.

Key words : megastudy, word recognition, lexical decision, Hangul reading