한국어 어휘뭉치의 표본 선정 기준

정 찬 섭 (연세대 심리학과)

Selection Criteria of Texts for the Korean Corpus

Chan Sup Chung
Dept. of Psychology Yonsei University

요 약

신뢰롭고 타당한 우리말 어휘뭉치의 표본 선정 기준을 정하기 위하여 언어 관계 전문가와 일반인을 대상으로 독서물의 중요성 및 독서량을 묻는 현장 조사를 실시하였다. 어휘 뭉치 표본으로서 17개의 독서물 유목 및 구어 대본이 선정되었으며 각 독서물 유목별 표집 비율이 산출되었다.

I. 머리말

한 말의 낱말 빈도는 그 말의 어휘 구조, 그 말을 사용하는 사람들의 말 사용 습관, 의식 구조, 사회적 관습 등과 같은 매우 복합적인 정보를 반영한다. 따라서 한 말의 어휘 구조를 분석한다든가, 사전을 편찬한다든가, 언어 습관 및 언어 교육 문제를 다룬다든가, 한 개인 또는 사회 전반의 의식이나 인지 구조를 분석하려 할 때 낱말 빈도는 결정적인 효용 가치를 발휘할 수 있다.

예를 들어, 사전을 편찬하기 위해 사전에 올려질 말의 자료를 구하는 조사 단계에서 단순히 말의 갈래를 수집하여 분류하는 것뿐 아니라 낱말 빈도에 대한 정보를 함께 조사함으로써 편찬될 사전에 여러가지 바람직한 속성들을 부여할 수 있다. 낱말 빈도는 말의 쓰임새 측면에서 낱말의 상대적 중요성을 가늠하는 준거가 되며, 한 낱말이 여러 갈래의 뜻을 지니고 있을 때 그것들 간의 중요성을 순서지을 수 있는 기초 정보를 제공한다. 따라서 사전에 올려지는 낱말의 상대적 중요성이나 뜻풀이의 상대적 중요성을 구분해줄 수 있는 준거로서 낱말 빈도를 사용하게 되면 사전 이용의 효율성을 크게 향상시킬 수 있다.

낱말 빈도는 언어교육의 기본 방향을 설정하는 데에도 절대적인 가치를 발휘할 수 있다. 외국어나 국어교육시 기본 어휘를 설정하고 말 가르침의 위계적인 절차를 확립하는데 낱말 빈도가 결정적인 도움을 줄 수 있다(Arts and Meijs, 1984).

최근 컴퓨터를 이용하여 글자와 글의 자동인식을 구현하려는 노력이 본격화되고 있는데 그 과정에서 신뢰로운 낱말 빈도 정보의 필요성에 대한 인식이 점점 고조되고 있다. 컴퓨터를 이용한 문서 편집이나 관리에서는 철자의 오류를 자동 점검하고 수정하는 프로그램들이 실용화되고 있는 단계에 있는데 이 프로그램들은 낱말빈도 조사에서 얻어진 철자간

이행확률 또는 n-gram 통계에 기초하여 잘못 누락되었거나 대체된 글자를 찾아내어 자동수정을 한다(예: Riseman and Hanson, 1974; Srihari, Hull, and Choudhari, 1983; Sheil, 1978). 날말 빈도는 이밖에도 광학기 철자판독(OCR), 어휘집을 이용한 글의 자동 인식 등에 필수적인 정보로서 사용된다.

날말 빈도는 지금까지 열거한 실용적인 목적 이외에도 언어학, 심리학, 사회학, 문학 등 다양한 학문에 걸쳐 정보 처리나 구조 분석의 실마리를 제공하는 주요 변수로서 취급된다. 날말빈도 정보를 담고 있는 어휘 뭉치(corpus)가 언어의 여러가지 양상을 이해하는데 결정적인 단서들을 제공하기 때문에 언어학에는 뭉치 언어학(corpus linguistics)라는 학문 영역이 있으며, 컴퓨터로 어휘뭉치를 분석하여 어휘구조 및 구문 구조를 연구하고 사전을 만드는 전산 언어학 (computational linguistics)이라는 학문영역이 있다. 심리학에서는 날말 재인 속도(Solomon and Postman, 1952; Morton, 1970; 김민식과 정찬섭, 1989)나 기억망 속에서의 정보인출(Conrad, 1972; Anderson,1983) 연구 등에서 볼 수 있는 것 처럼 날말 빈도에 관한 지식이 인지구조 및 정보처리 과정을 이해하는데 중요한 설명 변수로서 다루어진다. 날말 빈도는 이상섭(1984)의 한용운의 시에 나타난 날말 빈도 분석에서 볼 수 있듯이 한 개인, 또는 사회의 지배적인 사상이나 의식을 분석하는 데에도 이용될 수 있다.

지금까지 살펴본 바와 같이 날말 빈도 정보가 여러 분야에 걸쳐 중요하며 그 활용가치가 다양하기 때문에 일찌기 1920년대부터 미국, 영국, 프랑스, 독일, 일본 등에서 날말 빈도 조사가 실시되어 다양한 형태의 어휘 뭉치가 만들어졌으며 그 결과가 광범위한 영역에 걸쳐 활용되어왔다. 우리나라에서도 문교부의 주관으로 1951년 날말 빈도 조사가 실시되어 1956년 어휘뭉치가 출간된 바 있다.

II. 영미어의 어휘뭉치 및 우리말 어휘 뭉치

Thorndike-Lodge Word Frequency Count(1944): 최초의 날말 빈도 조사가 1920년대부터 1930년대에 걸쳐 미국에서 Edward Thorndike에 의해 수행되었는데, 이것이 일련의 수정작업을 거친 뒤 1944년 우리에게 잘 알려진 'Thorndike-Lodge Word Frequency count'라는 어휘뭉치의 이름으로 출간되어 나왔다. 조사에서 수집되어 출간된 날말의 마디수(여기서 마디라 함은 표집된 날말들의 빈도 계수 단위를 말한다)는 잡지, 신문, 책 등에서 뽑은 2천만 개였다.

Brown Corpus(1967): Thorndike-Lodge 날말 빈도 계수가 나온 후 좀 더 체계적인 방법을 사용하여 대표성이 있는 어휘 뭉치를 얻기 위한 날말빈도 조사 작업이 미국 Brown 대학의 Kucera와 Francis(1967)에 의해서 수행되었다. Brown Corpus라고 알려진 이들의 어휘뭉치는 1961년에 미국에서 출간된 15개의 유목의 일반 독서물 500편에서 수집된 100만이 조금 넘는 날말 마디들의 빈도 조사 결과를 담고 있다. Brown Corpus는 조사 대상인 날말 마디수가 100만개에 불과하지만 조사표본의 대표성이 우수하여 현재까지도 다양한 분야에서 널리 이용되고 있다.

Lancaster Oslo/Bergen(LOB) Corpus(1978): 미국어 어휘뭉치인 Brown Corpus와 성격이 같은 영어 어휘뭉치를 만들기 위하여 1970년에서 1978년에 이르기까지 영국의 Lancaster와 Norway의 Oslo, Bergen 지역에 걸쳐 LOB Corpus 라고 불리우는 어휘뭉치가 완성되었다. LOB Corpus는 1961년도에 영국에서 출판된 책을 조사대상 표본으로 하여 Brown Corpus에서와 같이 15개 유목의 일반 독서물 500편에서 100만 개의 말 마디를 표집하여 조사한 결과를 담고있다.

Birmingham Collection (1980-): 영국의 Birmingham 대학에서는 사전 편찬을 비롯하여 여러가지 목적에 두루 쓰일 수 있는 어휘뭉치를 만들기 위하여 영문과 내에 COBUILD라고 불리우는 연구과제의 담당부서를 두고 1980년부터 날말빈도를 조사해오고 있다. 이렇게 형성된 대규모의

어휘뭉치인 Birmingham Collection은 필기나 인쇄로 된 글과 구어(입말)를 옮겨적은 것의 집합체로서 계속적으로 규모를 확장하여 1984년에 이미 1천 2백만 마디가 되었으며, 1987년 현재 2천만 마디의 어휘 뭉치로 자라났다(Sinclair, 1987).

우리말 어휘뭉치(문교부, 1955; 1956): 우리나라 최초이며 최대였던 우리말의 낱말빈도 조사는 당시 문교부 편수국장이었던 외솔 최현배 선생의 주도하에 이루어졌다. 조사 작업은 1951년 4월에 착수되었으며 6.25 동란의 와중에서 어렵게 진행되어 조사가 시작된지 4년만에야 완성되었다. 문교부의 보고서는 독일의 F.W. Koeding, 미국의 E. L. Thorndike, 일본 국립 국어 연구소의 낱말 빈도 조사방법을 참조하여 조사를 수행했다는 것을 밝히고 있어 당시 연구진이 합리적이고 체계적인 조사 방법의 적용을 위하여 각별히 노력하였다는 흔적을 남기고 있다. 조사의 대상이 된 표본은 초, 중, 고교의 교과서, 소설, 시, 신문, 국회속기록, 방송대본, 소책자, 잡지 및 기술 서적 등 총 93개의 독서물에서 표집된 2,218,727 개의 낱말 마디였다. 이 표본을 낱말별로 분류하여 계수한 결과 표본 내에 포함된 낱말의 갈래가 56,485 개임이 밝혀졌다.

III. 어휘뭉치 표본 선정의 문제점

낱말 빈도 조사와 어휘뭉치 작성은 일견 단순한 작업처럼 보이나 신뢰롭고 타당한 어휘뭉치를 얻기 위해서는 조사를 계획하고 낱말을 분류하여 빈도를 계수하는 과정에서 여러 변수를 세심하게 고려하여야 한다.

어휘뭉치의 타당성은 그것의 대표성에 의해 가늠될 수 있다. 어휘뭉치의 대표성이라 함은 어휘뭉치가 일반 사람들이 일상적으로 쓰는 말과 글을 얼마나 근사하게 반영하고 있는가를 말한다. 한편, 어휘뭉치의 신뢰성은 뜻에 따라 낱말이 제대로 분류되었는가, 또 낱말 빈도가 얼마나 정확히 계수되었는가를 나타낸다. 어휘뭉치의 대표성과 신뢰성이 모두 중요하지만 신뢰성은 자료를 분류하고 빈도를 계수하는 단계에서 문제가 되는데 비해 대표성은 자료를 수집하기 위한 표본의 선정 단계에서부터 문제가 되기 때문에 신뢰성보다 대표성이 어휘뭉치의 가치를 좌우하는 더 큰 변수가 된다.

IV. 우리말 어휘뭉치의 표본 선정 기준

본 연구에서는 대표성이 보장되는 우리말 뭉치의 표본을 얻기 위하여 언어 및 독서 관계전문가와 일반인에 대한 조사를 토대로 표본선정 기준을 마련하기로 하였다.

1. 언어 및 독서관련 전문가 면접 연구

대표성이 있는 우리말 뭉치의 표본선정 기준을 마련하기 위하여 언어 및 독서 관계 전문가들을 대상으로 면접 연구를 수행하였다. 연구 목적은 교과서의 표집비율 조정 문제, 번역물의 표집 여부에 관한 문제를 해결하고 일반인들을 대상으로 한 독서실태 조사 질문지에 포함될 독서물 유목을 결정하기 위한 것이었다.

1) 면접 조사대상

조사 대상이 되는 전문인들은 언어 및 독서관계 전문가들로서 한글학자 4명, 대학교수 12명, 문학가 2명, 출판 관계인 3명, 언론인 1명, 서적 판매업자 3명 등 모두 25명이었다.

2) 질문지

면접에 사용된 질문지는 어휘 뭉치 표본의 작성을 위한 독서물 유목의 적절성 및 번역물의 표집, 교과서의 표집에 대한 전문가의 의견을 묻는 부분적으로 구조화된 문항과 자유 문항으로 구성되어 있었다. 독서물 유목은 외국(특히 Brown과 LOB Corpus를 주로 참조) 및 문교부(1955)의 독서물 분류 유목을 참조하여 작성하였다. 최종적으로 질문지에 포함된 독서물 유목은 19 종류 였다.

면접 조사에서는 응답자들에게 낱말 빈도 조사의 취지와 표본 선정의 중요성을 설명한 다음 이들 유목들을 하나씩 제시하면서 그 유목을 조사대상 표본에 포함시키는 것이 얼마나 적절한가를 물었다.

유목의 적절성은 상, 중, 하 및 제외의 네 답지중 하나를 택하여 평가하도록 하였다.

3) 결과

독서물 유목의 적절성 평가에서 얻은 조사 결과가 <표 1>에 제시되어 있다. 이 결과를 토대로 유목의 상대적 중요성을 나타내는 지수를 나타내기 위하여 평정치 '상'은 4, '중'은 2.5, '하'는 1, 그리고 '제외'는 -4를 응답자 수에 곱하여 그것들을 모두 더하였다. 따라서 응답자 모두가 '상'에 평정을 하면 중요도 지수가 100, 모두 '제외'에 평정을 하면 -100의 값이 나오도록 하였다. 이와같은 과정에 의해 산출된 유목의 중요도 지수가 <표 1>의 마지막 난에 제시되어 있다.

< 표 1 > 우리말 날말 빈도 조사 표본에 사용될 유목의 적절성 평정치

독서물 분류유목 /	중요도
신문: 정치면 ———	80.5
신문: 경제면 ———	68.5
신문: 사회면 ———	91
신문: 문화면 ———	88
신문: 오락 및 스포츠면	77.5
종교서적 ———	46.5
기술, 수공예, 취미 -	50.5
위인전, 전기, 수기 —	73
수필 ———	94
잡지 (정부 간행물,사보, 각종 보고서 및 요람)	23.5
잡지 (여성, 가정) -	82
잡지 (정치, 사회) -	77.5
잡지 (문학) ———	86.5
잡지 (취미, 오락) -	71.5
학술서적 및 논문 —	40.5
단행본 소설 ———	94
역사 소설 ———	64.5
실화집 ———	65
생활정보 서적 ——	60.5

우리말 날말빈도 조사를 위하여 초,중,고,대학의 교과서를 포함시키는 것이 좋은지, 만일 좋다면 어느 수준부터 표집시켜야 할 것인지를 묻는 질문에서 찬성이 21명(84%), 반대가 3명(12%), 무응답자가 1명(4%)으로 나타나 다수의 응답자들이 교과서의 포함을 찬성하는 것으로 밝혀졌다.

우리말 날말빈도의 표집 대상으로서 번역물(또는 번안물)을 포함시켜야 하는지에 대한 물음에서는 번역물을 포함시켜야 한다가 13명(52%), 번안물을 포함시켜야 한다가 7명(28%), 조사 표본으로부터 번역물을 제외하는 것이 좋다가 5명(5%)으로 나타났다. 이같은 결과는 번역물의 포함 여부에 대해 응답자들의 의견이 대체로 양분되어 있다는 것을 보여준다.

2. 독서물 유목별 독서 실태 조사

우리말 어휘빈치의 표본 선정 기준을 전문가의 견해에만 의존해서 결정하기에는 날말 빈도 조사 관계 참고 자료의 빈약함 및 어휘빈치 관련 연구 사례의 절대 부족 등 여러가지 미흡한 점이 많다. 따라서 본 연구에서는 직접 일반인을 대상으로 독서 실태 조사를 수행함으로써 어떤 종류의 독서물이 일반인들에게 널리 그리고 많이 읽히는가를 알아보고 그 결과에 근거하여 어휘빈치의 표본선정 기준을 마련하고자 하였다.

1) 조사 대상자

연구의 성격과 목적상 조사 대상자로서 일반인의 대표 표본(representative sample)이 필요하여 전국 60개 지역에서 3단층화 무선 표집 방법을 사용하여 만 20세 이상의 사람 1,206명을 선정하였다.

2) 질문지

사람들이 어떤 독서물을 얼마나 읽고 있는가를 알기 위해서 본 연구에서는 전문가의 면접 조사에 사용된 19개의 유목중 중요도 지수가 50을 넘는 16개 유목만을 골라 독서량을 추정하기로 하였다. 응답자들로부터 답을 쉽게 얻어내기 위해서 본 연구에서는 질문지로 제시되는 16개 독서물 유목을

위계적으로 상위 유목과 하위 유목의 두 묶음으로 나누어 질문 수를 줄이고, 직접 척도화 방법 보다는 쌍대 비교법(pair comparison method)이라는 간접 척도화 방법(Thurstone, 1927)을 써서 독서량을 측정하였다(예를 들어, "소설과 신문중 어느 것을 더 많이 읽습니까"?). 또한 독서량의 상대 평가 방법인 쌍대 비교법의 보완 자료를 구하기 위하여 16개 독서물의 상위 유목인 신문, 잡지, 소설과 수필, 취미와 교양, 수기와 전기의 5유목에 관해 한달간의 평균 독서 시간을 물어 독서량의 절대 평가를 시도하였다.

3) 조사 방법

조사에 동원된 면접원은 보수를 받기로 하고 모집된 연세대학교 남녀 상급학년 재학생 20명이었다. 이들은 조사에 임하기 전 두 명의 조사 책임자로부터 질문지의 내용검토 및 면접 요령 실습 등 이틀간에 걸친 훈련을 받았다. 면접은 미리 선정된 가구로 찾아가 만 20세 이상의 사람을 대상으로 개별적으로 실시되었다.

4) 결과

Thurstone 방법(1927)에 의해 독서물 유목별, 응답자 배경 변인별로 독서량의 척도값이 계산되었다. 척도값은 질문지의 위계적 구성내용에 따라 상위척도 1개와 하위척도 5개가 따로 계산되었다. 척도값은 쌍대 비교판단을 요하는 하나의 독서물 집단에서 가장 독서량이 많은 유목의 척도값이 3이 되도록 원래의 계산에서 얻어진 척도 값에 일정한 값을 가감하여 정하였다. 상위 유목에 대한 전체 응답자 및 응답자 배경변인별 척도값만을 계산한 것이 <표 2>에 제시되어 있다.

<표 2> 독서물의 상위 유목별 척도값

신문	잡지	소설/수필	취미/교양	수기/전기/실화집
3.00	1.81	1.62	1.59	0.89

각 상위 척도에 따른 하위 표본의 유목별 독서량을 순서를 나열해보면 신문은 정치면, 사회면, 오락/스포츠면, 경제면, 문화면 순으로; 잡지는 취미/오락,정치/사회, 여성/가정, 문학잡지 순으로; 소설과 수필류는 일반소설, 역사 소설, 수필의 순으로; 취미/교양은 생활정보서적, 기술/수공예/취미관계서적 순으로; 수기/전기/실화집에서는 실화집, 수기/전기의 순으로 많이 읽는 것으로 나타났다.

독서 시간을 근거로 하여 독서량을 살펴보면 신문, 잡지, 소설/수필, 취미/교양, 수기/전기/실화집의 5개 유목에 대해 수집된 전체 응답자의 한달 평균 독서시간은 55.9시간인 것으로 나타나 하루 평균 약 1.9시간의 독서를 하고 있는 것으로 나타났다. 또한 그 유목을 읽는다고 답한 응답자의 수는 신문, 잡지, 소설/수필, 취미/교양, 수기/전기/실화집의 순으로 많았고, 그외의 독서물로는 경전 등 종교서적과 학술서적/전문서적/교과서의 순으로 많았다.

3. 날말 빈도 조사의 표집비율 결정

전문가 및 일반인에 대한 면접 조사 자료가 수집, 분석됨에 따라 이 두 조사의 결과를 결합하여 날말빈도 조사 표본의 표집 비율을 결정하는 일이 남게 되었다. 문제를 체계적으로 풀어나가기 위하여 우선 일반인 면접 조사에 근거한 새로운 척도를 구성하고 그 다음 이 척도와 전문가 면접 조사 결과를 합하여 날말빈도의 조사 결과를 사정하였다.

쌍대 비교법으로 얻은 척도는 비율을 적용할 수 없다. 따라서 독서물을 신문, 잡지, 소설/수필, 취미/교양, 수기/전기의 5유목으로 크게 나누어 각 유목에 대한 척도 점수를 낸 뒤, 독서시간으로 추정한 독서량을 참조하여 가장 많이 읽는 독서물이 가장 적게 읽는 독서물의 6 배수가 되도록 척도 점수를 조정하였다. 이렇게 하여 얻어진 척도 점수의 총합을 100으로 하여 백분율로서 5개 유목의 독서물 유형별 표집 비율을 정하였다. 각 5개 유목의 하위 유목의 표집 비율은 각 유목별 척도 평균을 중심점으로 삼아 척도 점수를 다시 조정하여 그 유목의 평균 표집 비율에 가감하는 것으로서 정하였다.

일반인이 많이 읽는 독서물을 중심으로 날말 빈도 조사의 표본을 구성한다면 위의 결과를 독서물 표집 비율로 삼아야 할 것이다. 그러나 독서율 못지 않게 독서물의 영향력을 표본 선정의 중요한 지표로 삼아야 한다. 본 연구는 이러한 점을 반영하기 위하여 두 조사 결과를 절충한 표집 비율을 정하기로 하였다.

교과서와 번역물의 표집문제는 전문가들의 의견과 일반인들의 응답, 그리고 본 연구 연구진들의 토의를 거쳐 결정되었는데, 교과서를 전 표본의 5%(초 1%, 중 2%, 고 1%, 대 1%)가 되도록 포함시키고 희곡 및 시나리오를 역시 5%가 되도록 포함시키기로 하였다(< 표 3 > 참조).

< 표 3 > 전문가 및 일반 조사 결과를 종합한 최종 백분 표집 비율

	최종 표집비율
1. 신문	33%
정치	7
경제	6
사회	8
문화	6
오락/스포츠	6
2. 잡지	20%
여성/가정	5
정치/사회	5
문학	5
취미/오락	5
3. 소설 및 수필	18%
수필	6
일반소설	7
역사소설	5
4. 취미 및 교양	10%
기술/수공예/취미	4
생활정보	6
5. 수기.전기 및 실화집	9%
수기/전기	4
실화집	5
6. 교과서 (국어만 포함)	5%
초	1
중	2
고	1
대	1
7. 방송 스크립트	5%
계	100%

VI. 맺는 말

눈앞에 닥쳐 온 정보화 사회에 적절하게 대비하기 위해서는 신뢰할 수 있는 우리말 대사전의 편찬, 한글 공학, 문서 및 언어 정보 처리의 자동화 등 서둘러 해결해야 될 문제들이 산적해 있으며 그를 위해 쓰여질 자료 밀천인 어휘 뭉치의 개발이 절실한 형편에 있다. 본 연구에서는 이러한 취지에서 어휘뭉치의 다양한 효용가치, 외국의 어휘뭉치 실태, 우리말 어휘뭉치의 실태 및 문제점을 개관해 보았으며 신뢰롭고 타당한 우리말 어휘 뭉치를 작성하는 기초 작업으로서 두편의 현장조사 결과를 토대로 어휘뭉치의 대표성을 보장받기 위한 날말 빈도조사 표본의 표집 비율을 마련하였다.

우리말 날말 빈도 조사 표본의 표집 비율을 전문가와 일반인을 대상으로 한 현장 조사의 결과에 근거하여 정한 것은 외국에서와 같이 독서물 유형별로 독서율을 파악하고 영향력이 있는 독서물을 선정할 수 있도록 해 주는 자료들이 없기 때문이다.

표집비율을 정하는 것은 어휘뭉치 작성의 아주 적은 일부분의 작업에 지나지 않는다. 이제 이 표집 비율을 근거로 독서물 표본을 가려 뽑은다음 막대한 시간과 노력 및 경비를 요하는 대규모의 날말 분류와 빈도계수 작업을 하여야만 한다. 어휘체(lexis) 수준으로 날말을 정리하여 어휘뭉치를 작성하려면 컴퓨터를 이용한 우리글과 말의 효율적인 분류 및 분석방법이 모색되어져야 하며 언어학 관련 전문가의 힘든 분류 및 확인 작업이 뒤따라야 된다. 아직 현존하는 컴퓨터가 우리말과 글을 효율적으로 처리할 수 있도록 개발되어 있지 않다는 점과 우리 말은 토씨, 어미 변형들 날말 분류작업을 어렵게 만드는 요인들을 지니고 있다는 점을 고려하면 이러한 일련의 어휘뭉치 작성 작업이 생각보다 쉽지 않으리라는 것이 명백해진다. 어휘뭉치 작성은 어렵지만 우리가 반드시 해내야만 되는 필수적인 작업이다. 앞으로 많은 사람들이 우리말 어휘뭉치작성에 관심을 가지어 가까운 장래에 우리말 전산 사전의 개발등 우리말 관련 연구 및 정보 산업의 발전이 실현되어야 할 것이다.

참고 문헌

김민식, 정찬섭 (1989). 한글의 자모 구성 형태에 따른 자모 및 글자인식. 한국인지과학회지, 1.

문교부 (1955). 우리말에 쓰인 글자의 잦기 조사-문자 빈도 조사. (우리말 말수 사용의 잦기 조사 두째 엮음).

문교부 (1956). 우리말 말수사용의 잦기 조사-어휘사용빈도 조사.

이상섭 (1984). 님의 침묵의 어휘와 그 활용 구조. 서울: 탑구당.

Aarts, J., & Meijs, W. (1984). Corpus Linguistics: Recent development in the use of computer corpora in English language research. Amsterdam: Rodopi.

Anderson, J. R. (1983). The Architecture of Cognition. Cambridge: Harvard University Press.

Conrad, C. (1972). Cognitive economy in semantic memory. Journal of Experimental Psychology, 92, 149-154.

Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence: Brown University Press.

Leech, G. N.,& Leonard, R. (1974). A computer *corpus of* British English. Hamburger Phonetische Beitrage, 13, 41-57.

Morton, J. (1970). A functional model of memory. In D.A. Norman(Ed.), Models of human memory. New York: Academic Press.

Riseman, E. M. & Hanson, A. R. (1974). A contextual postprocessing system for error correction using binary n-grams. IEEE Transactions on Computers, 480-493.

Sheil, B. A. (1978). Median split trees: A fast lookup technique for frequently occurring keys. Comm. ACM, 21, 11, 947-958.

Sinclair, J. M. (1987). Sense and structure in lexis in linguistics in a Systemic perspective, J. Benson, M. Cummings, and W. Greaves(eds.). Glendon College, York University, Toronto.

Solomon, R. L., & Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. Journal of Experimental Psychology, 43, 195-201.

Srihari, S, N, Hull, J. J., & Choudhari, R. (1983). Integrating diverse knowledge sources in text recognition. ACM Trans. Off. Inform. Sys., 1, 68-87.

Thurstone, L. L. (1927). A law of comparative judgment. Psychol. Rev., 34, 273-286.

Thorndike, E. I., & Lodge, I. (1944). The teacher's word book of 30,000 words. New York: Teacher's College Press, Colombia University.