

# 계산주의적 모델을 이용한 한국어 어휘습득 모델 설계 및 구현<sup>1)</sup>

유원희\*, 박기남\*\*, 류기곤\*, 임희석\*  
\*한신대학교 컴퓨터정보소프트웨어학부  
\*\*고려대학교 컴퓨터교육과  
galadous@hs.ac.kr

## Design and Implementation of Korean Lexical Acquisition Model using Computational Model

Wonhee Yu\*, Kinam Park\*\*, Kigon Lyu\*, Heuseok Lim\*  
Information and Software, Hanshin University\*  
Department of computer Education, Korea University\*\*

### 요 약

본 논문은 인간의 언어정보처리과정 중 초기 어휘획득(lexical acquisition)과정을 한국어에 적용시켜 Full-List 모형과 Decomposition 모형의 하이브리드한 형태의 계산주의적(computational) 어휘정보처리 모델을 구현하고 실험하였다. 실험결과 학습을 통한 언어적 입력의 인간의 어휘획득 과정을 모사(simulate) 할 수 있었고, 특정 문법범주 습득 순서에 대한 이론적 근간을 제시할 수 있었다. 또한 본 연구의 모델에서 자동으로 생성된 Full-List 사전과 Decomposition 사전을 통해 인간의 대뇌 심성표상(mental representation) 형태를 유추할 수 있는 증거를 보였다.

### 1. 서론

인간의 초기 어휘획득(early lexical acquisition) 과정을 살펴보면 외부자극인 사물(object)이나 명칭(name)의 명명(naming)을 통해 어휘의 개념과 의미를 형성하고, 기본적인 언어학적 형태를 어휘를 통해 완성해 나간다. 하지만 최근 까지도 초기 어휘획득 과정에 대한 여러 이론이 대립하고 있음은 물론이고, 인간의 대뇌 심성표상(mental representation)에 대한 의견도 분분하다.

초기 어휘획득 과정의 대표적인 두 가지 이론 중에 첫 번째는 인간은 단어를 자신의 주변 환경과 상호 연결시켜 인지하게 되는데, 예를 들어 명사는 동사에 비해 개념적으로 분명하기 때문에 동사보다 쉽게 인지할 수 있다. 즉, 인지적으로 덜 복잡한 현상은 어느 언어에서나 원칙적으로 일찍 습득 된다. 이는 선형적으로 정해진 보편적인 인지적 제약성(constraints)에 의해 어휘를 습득해 나간다는 이론이다[1]. 둘째는 언어적 입력에 의해 즉, 개별 언어

인 모국어의 특성에 근거하여 어휘를 습득한다는 이론이다[2].

또, 심성 어휘집에 형태소, 단어, 혹은 어절이 어떻게 표상되어 있을 것인가에 대한 논의가 다양하게 전개되어 왔다. 현재 형태소 분석에서 거론되고 있는 심성어휘집의 모형은 크게 세 가지로 요약될 수 있다. 첫 번째는 Full-List 모형으로 하나의 어절은 그 자체로 심성어휘집에 등록되어 있다는 이론이다[3]. 이 이론에 따르면 하나의 기본형에서 굴절되거나 파생이 된 어휘는 기본형과는 별개로 모두 심성어휘집에 등록되어 있다. 두 번째는 Decomposition 모형으로서 하나의 기본형에서 파생되거나 굴절된 어휘는 어근과 나머지 부분으로 나뉘어져서 각각 따로 저장되어 있다는 이론이다[4]. 세 번째는 Hybrid 모형으로 어휘의 품사, 빈도, 형태소 활용 규칙 등에 따라서 Full-List 혹은 Decomposition 으로 저장된다는 이론이다[5].

이에 본 논문은 한국어 초기 어휘획득과정에서 어휘획득이 언어적 입력에 의한 것이고, 심성어휘집의 형태가 Full-List 모형과 Decomposition 모형의 하

1) 본 연구는 한국학술진흥재단의 연구지원(2004-074 -HM 0004)에 의하여 수행되었음.

이브리드한 형태로 표상되었다는 이론을 바탕으로 계산주의적 모델을 설계하고 구현하고자 한다. 이를 통해 본 논문에서 한국어 어휘획득 과정의 계산주의적 모델을 제안하고, 특정 문법범주 습득 순서에 대한 이론적 근간을 제시하고자 한다. 또 본 연구의 모델에서 자동으로 생성된 Full-List 사전과 Decomposition 사전을 통해 인간의 대뇌 심성어휘집(mental lexicon) 표상 형태를 알아보하고자 한다.

## 2. 계산주의적 어휘정보처리 모델

본 논문은 Full-List 모형과 Decomposition 모형의 하이브리드한 형태의 계산주의적(computational) 어휘정보처리 모델을 제안한다.

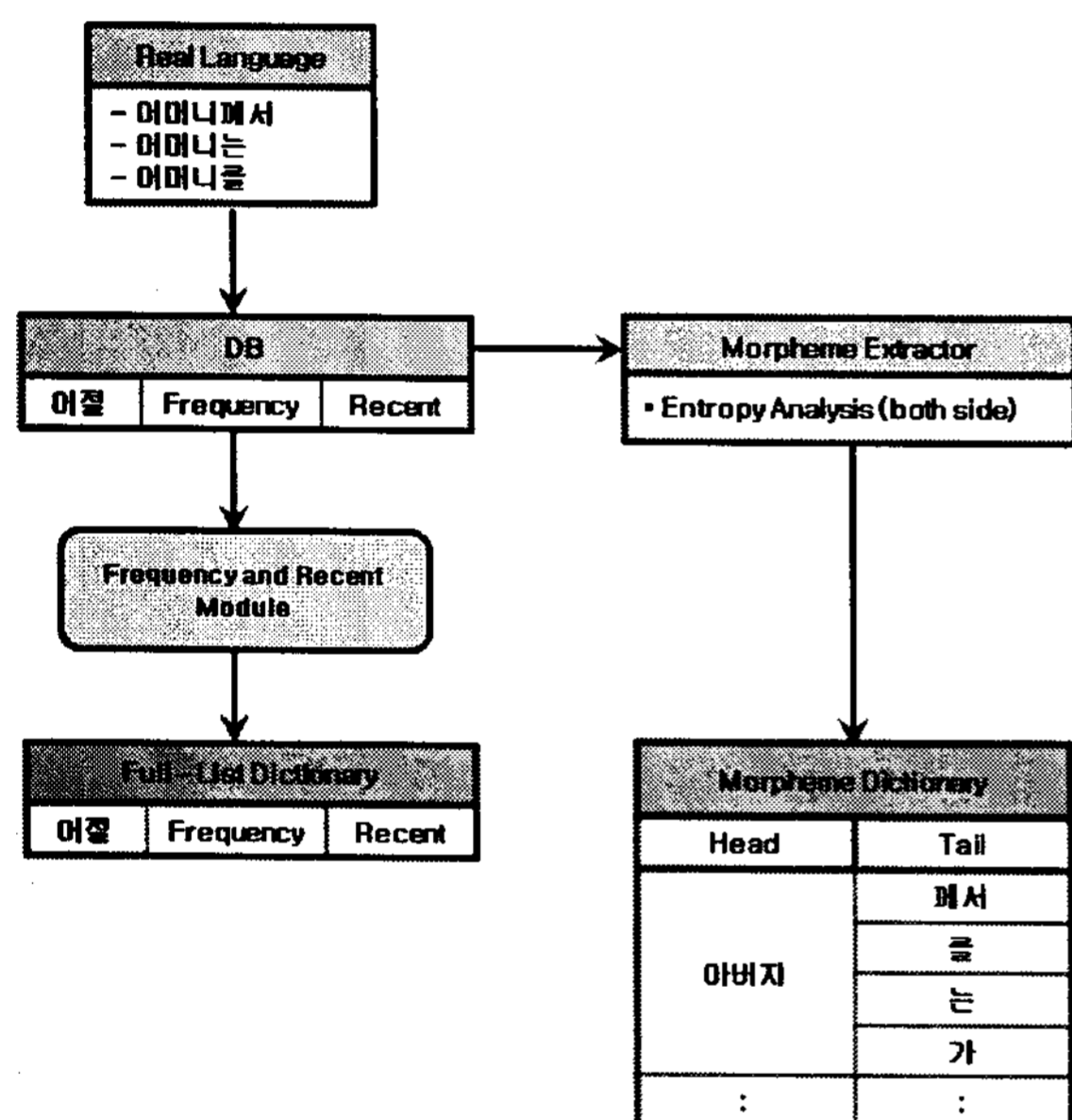


그림 1. 계산주의적 어휘정보처리 모델

본 논문에서 제안한 모델은 그림 1과 같이 Real-Language를 입력받아 일련의 과정을 거치게 되면 Full-List 사전과 Decomposition 사전이 자동으로 생성되는 구조로 설계하였다.

### 2.1 학습 데이터

입력값인 Real-Language는 21세기세종계획의 문어 말뭉치(corpus)로 1천만 어절과 구어 말뭉치 85만 어절을 사용하였다.

### 2.2 Full-List 사전 학습

본 논문의 Full-List 사전은 입력값인 Real-Language를 차례로 한 어절씩 입력하여 해당

사전을 학습시켰다. 학습 시 어절은 자체 빈도(frequency)와 해당어절의 최근성(recent)을 고려하여 특정 임계치(threshold) 이상이면 학습되도록 하였다. 예를 들어 “어머니가”라는 어절이 입력값으로 주어졌을 때 “어머니가”의 자체 빈도를 통해 특정 임계치(threshold) 이상의 빈도이면, Full-List Model에 등록 한다. 최근성은 최근 노출된 어절을 특정 window-size 만큼 목록으로 가지고 있다가 window-size안의 목록 중 특정 임계치 이상의 값이 목록에 출현했다면 Full-List Model에 등록한다.

### 2.3 Decomposition 사전 학습

Decomposition 모형은 엔트로피(entropy)를 이용하여, 한 어절을 어근과 어미 부분으로 분리 하였다. 한 어절의 음절에 대한 엔트로피는 식(1)과 같이 구할 수 있다. 여기서  $P_i$ 는 음절  $i$ 의 확률 값이다.

$$-\sum_{i=0}^n (P_i \log P_i)$$

[식 1] Entropy

## 3. 실험 결과

그림 2는 Full-List 사전의 입력 어절 수에 따른 학습 데이터양을 보여주고 있다.

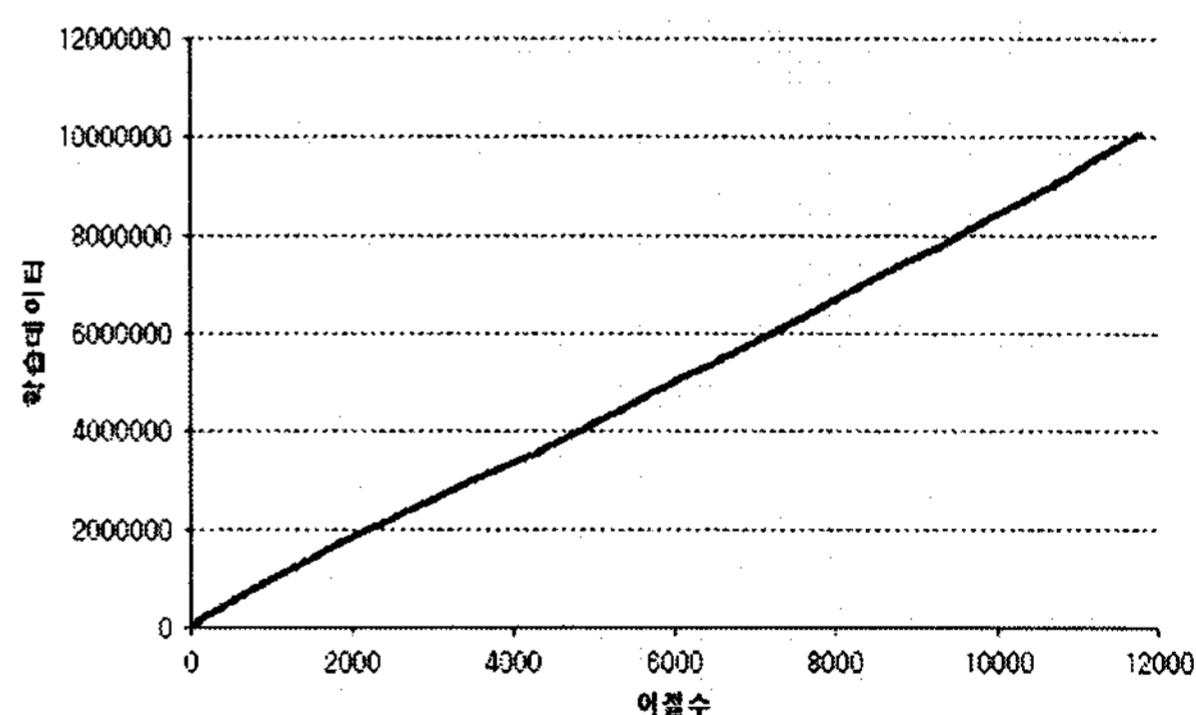


그림 2. 입력어절 수에 Full-List 사전의 학습량

학습되어진 어절은 전체어절의 50%의 비중을 가지고 있으며, 전체 어절 1백만 개 중 약 1만여 개에 해당하였다. 표 1은 학습된 Full-List 사전 결과물 중 일부이다. 1천말 어절 중 어떠한 순서대로 학습이 되었는지 볼 수 있으며 문어말뭉치를 사용한 결과물중 일부이기 때문에 그, 수, 이 등이 앞쪽에 학습된 것을 확인할 수 있다.

표 1. Full-List 사전

	단어	빈도	최근성	등록당시 노출 어절 수
1	수해는	100	2	5397
2	있었다	100	3	13219
3	있었다	100	3	18340
4	수해가	100	2	18678
5	그	100	3	21394
6	있는	100	1	24865
7	한	100	2	28509
8	너의	13	10	32748
:	:	:	:	:
10	나의	100	4	33503

Decomposition 사전은 전체 입력어절에 대해 앞, 뒤 양방향 엔트로피를 이용하여 어절을 분리한 후 학습되도록 하였다. 그림 2는 Full-List 사전의 입력어절 수에 따른 학습 데이터양을 보여주고 있다.

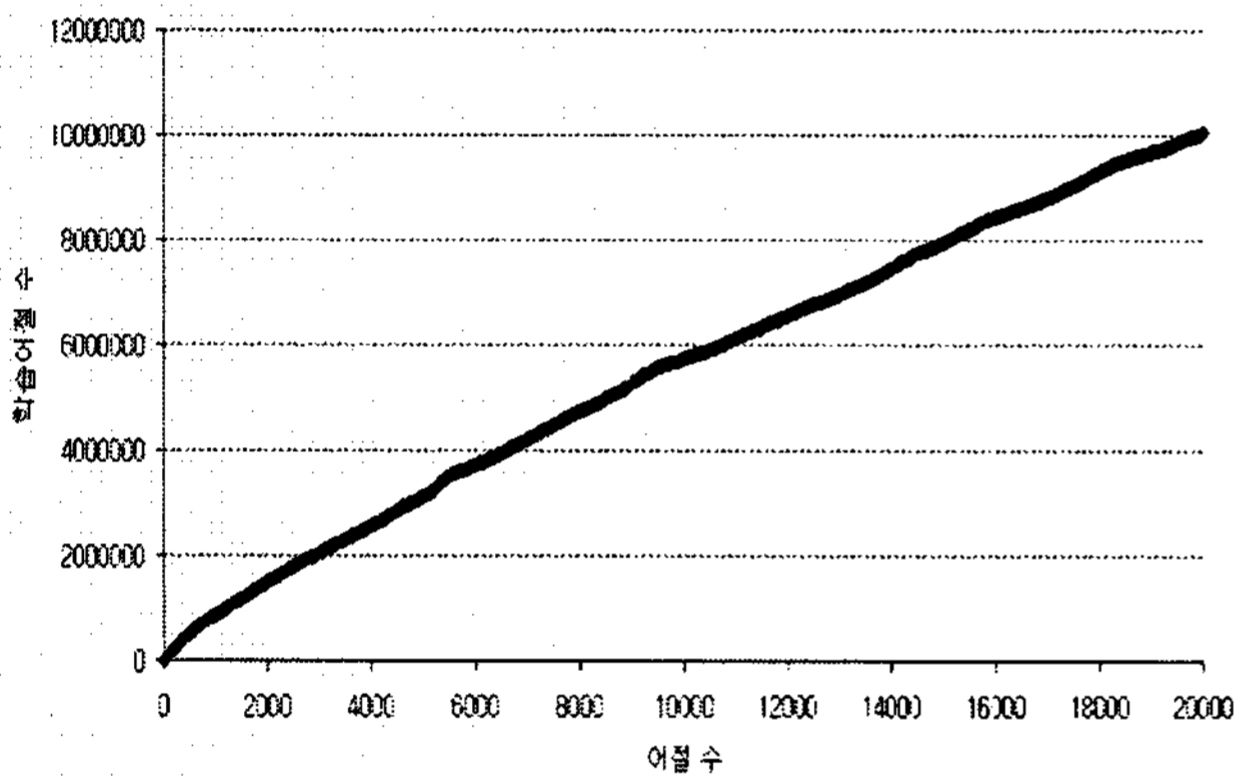


그림 3. 입력어절 수에 Decomposition 사전의 학습량

[표 2]는 decomposition model의 결과물중 일부이다. Head 부분이 어근부분을 나타내는 것이고 Tail 부분이 나머지 부분을 나타낸다. 대략 1천만 어절 중 13000여개의 어근부분이 학습 되었다.

표 2. Decomposition 사전

Index	Head	Tail
1	독도獨島	에는
2	초과일	이었다
3	양로원	에서
4	우주진화	사상과
5	인문사회	계열의
6	칠현금	에서
7	움짤	거리는
8	만화영화	처럼
:	:	:
10	토문강	에서

4. 결론

본 논문은 인간의 언어정보처리과정 중 초기 어휘 획득 과정을 한국어에 적용시켜 Full-List 모형과 Decomposition 모형의 하이브리드한 형태의 계산주 의적 어휘정보처리 모델을 구현하고 실험하였다. 실험결과 학습을 통한 언어적 입력의 인간의 어휘획득 과정을 모사 할 수 있었고, 문법범주 습득에 대해서 동사류에 비해 명사류 어휘습득량이 상대적으로 많다는 것을 본 논문에서 제안한 모델을 통해 알 수 있었다. 또 논문에서 제안한 모델이 한 한국어 어휘 습득에 있어 심성어휘집의 형태가 Full-List 모형과 Decomposition 모형 형태로 이루어졌음을 유추할 수 있는 결과물이라 할 수 있을 것이다.

참고문헌

- [1] Gentner, D., "Why nouns are learned before verbs: linguistic relativity versus natural partitioning", *Language Development*, Vol.2, pp.301-334, 1982.
- [2] Gopnik, A., Choi, S., "Early acquisition: rate, content, and the vocabulary sput", 22, pp.497-529, 1995.
- [3] Butterworth, B., "Lexical representation of derivational relation", In M. Aronoff & M.L. Kean(Eds.), *Juncture*, 37-55. Saratoga, CA: Anma Libri., 1983.
- [4] Anshen, F., Aronoff, M. "Producing morphologically complex words", *Linguistics*, 26, 1988.
- [5] Cole, P., Segui, J., Taft, M., "Words and Morphemes as Units for Lexical Access", *Journal of Memory and Language*, 28, 1997.