한국어의 어휘계층과 음운론적 복잡성*

박선우** · 홍성훈*** · 변군혁
(한신대학교) (한국외국어대학교)

Park, Sunwoo, Sung-hoon Hong and Koonhyuk Byun. 2013. Lexical strata in Korean and Phonological Complexity. *Studies in Phonetics, Phonology and Morphology* 19.2. 255-274. This paper explores the structure of the Korean lexicon based on the notion of "Phonological Complexity (PC)." The Korean lexicon is composed of three lexical strata: native and Sino-Korean words, and loanwords. For this study, we obtain 500 most frequent nouns for each class from the word list compiled by Kang and Kim (2004). We then calculate the PC values of the selected nouns using the bigram model proposed by Goldsmith (2002). A comparison of the PC values reveals that the average PC value is highest for loanwords, and lowest for Sino-Korean. The average PC value of native nouns is placed in the middle, and in fact, the PC values of native vocabulary are distributed widely from low to high values. The distribution of native nouns is in contrast with the distributions of loanwords and Sino-Korean, which are biased toward high and low PC values, respectively. Considering the relation between PC and markedness, we assert that Sino-Korean vocabulary is clustered around the unmarked portion of the lexicon, while loanwords are predominantly placed in the marked portion of the lexicon. Native vocabulary, on the other hand, is distributed in both marked and unmarked parts of the lexicon. **(Hanshin University and Hankuk University of Foreign Studies)**

Keywords: lexical strata, native Korean, Sino-Korean, loanword, Information Theory, Phonological Complexity, bigram model

1. 머리말

한국어의 어휘는 종류에 따라 고유어, 한자어, 차용어로 구분된다. 이러한 어휘계층의 구분은 주로 어원적, 사회적 개념으로 논의되어 왔으나, 언어학적으로 무관한 개념이라고 보기는 어렵다. 예를 들어 영어로부터 차용된 단어들은 구개음화와 두음법칙의 제약이 적용되지 않으며 한자어에서는 끽(喫), 쌍(雙), 씨(氏)와 같은 예외적인 한자 외에는 경음이 관찰되지 않는다. 채서영(1999)에서는 한국어의 어휘 계층을 가장 핵심적인 위치를 차지하고 있는 고유어(S1), 핵심부와 주변부 사이에 있는 한자어(S2), 핵심부로부터 가장 멀리 떨어져 주변부에 있는 차용어(S3)로 구분하고 핵심부에서 멀어질수록 적용되는 음운규칙이나 제약은 제한된다고 논의하였다. 예를 들어 /e/(에)와 /æ/(애)의 합류는 모든 어휘 계층에 적용되지만 두음제

* 이 연구는 2012학년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어졌으며, 2012년 10월 27일 서울대학교에서 열린 한국언어학회 가을학술대회에서 발표한 내용을 보완하고 수정한 것입니다. 아울러 본 연구의 문제점과 후속 연구의 방향을 제안해 주신 심사위원진과 일본어의 어휘계층 구조에 대하여 조언해 주신 한국외대 손범기 선생님께 감사드립니다.
** 제1저자, *** 교신저자

약(老人 *[loin], 利子 *[lidʑa])은 고유어와 한자어에만 적용되며 /o/와 /u/의 상승(-hago > -hagu, seda > sida)은 고유어에서만 일어난다.

본 연구에서는 한국어의 어휘계층을 '음운론적 복잡성'의 관점에서 살펴보고자 한다. Goldsmith (2002), Hume (2006), Hong (2006) 등에서는 정보이론(Information Theory, Shannon 1949)을 바탕으로 음운론적 복잡성과 유표성을 측정할 수 있는 분석방법을 제안하였다. 이러한 분석방법에 의하면 단어를 구성하는 음소의 출현빈도를 바탕으로 음소의 확률과 정보량을 구하고, 정보량을 통하여 단어의 '음운론적 복잡도'(Phonological Complexity, PC)를 측정할 수 있다. 음운론적 복잡도는 단어의 유표성을 판정하는 지표로 활용될 수 있다. 유표적 혹은 무표적으로 이분법적 기준을 적용하는 기존의 유표성 이론과 달리 음운론적 복잡도를 활용하면 개별적인 음소와 단어의 복잡도를 가시적인 수치로 측정할 수 있다.

구체적인 분석을 위하여 본 연구에서는 Goldsmith (2002)에서 제안한 'bigram 모델'에 따라 출현빈도가 높은 일반 명사들을 고유어, 한자어, 차용어로 구분하여 음운론적 복잡도를 측정하였다. 측정된 어휘계층별 복잡도는 다시 단어의 출현빈도, 음절개수, 단어를 구성하는 bigram의 유형빈도 등 다양한 지표들과 비교하여 분석하였다. 세 가지 계층을 구분하여 음운론적 복잡도를 측정한 결과 고유어와 한자어가 음운론적 제약상 하나의 범주의 묶을 수 있다는 강용순(1998)이나 한국어 어휘계층의 핵심부를 구성하는 고유어가 가장 무표적이라는 채서영(1999)의 논의와 달리 '차용어>고유어>한자어'의 유표성 위계를 확인할 수 있었다.

어휘계층별 복잡도의 분석을 통하여 본 연구에서 구체적으로 논의할 세 가지 문제는 다음과 같다. 첫째, 고유어, 한자어, 차용어의 세 가지 어휘계층은 어떠한 유표성 위계를 갖고 있으며, 이들 사이의 차이는 통계적으로 유의미한 수준인가? 둘째, 단어의 출현빈도나 음절의 개수, 단어를 구성하는 음소의 유형빈도는 복잡도와 상관 관계를 갖는가? 셋째, 어휘계층별로 복잡도가 높은 단어와 낮은 단어, 달리 말하자면 유표적인 단어와 무표적인 단어의 특성은 무엇인가? 그리고 복잡도를 높이는 유표적 특성들은 무엇인가? 이러한 문제들을 통하여 한국어의 어휘계층에 대한 음운론적 특성은 물론 정보이론에서 제안된 음운론적 복잡도의 효용성에 대해서도 함께 논의하겠다.

이후 전개될 내용은 다음과 같이 구성되어 있다. 2장에서는 강용순(1998)과 채서영(1999)를 중심으로 한국어의 어휘계층에 대한 기존의 논의를 정리하고 한국어와 유사한 어휘계층을 갖고 있는 일본어에 대한 논의도 함께 살펴보겠다. 3장에서는 bigram 모델(Goldsmith 2002)을 이용하여 음운론적 복잡도를 구하는 구체적인 과정과 방법을 소개하겠다. 정보이론에 대한 개요와 bigram 모델과 unigram 모델의 차이점에 대해서도 설명하겠다. 4장에서는 고빈도 일반명사를 중심으로 고유어, 한자어, 차용어의 음운론적 복잡도를 분석하겠다. 단어의 음운론적 복잡도를 단어의 출현빈도, 단어를 구성하는 bigram의 유형빈도, 음절의 개수 등과 비교하면서 복잡도가 높은 단어와 낮은 단어의 특성도 설명하겠다. 마지막 5장에서는

본론의 논의를 요약하면서 유표성의 지표로서 음운론적 복잡도의 효용성과 분석방법의 보완 방향을 전망하겠다.

2. 한국어의 어휘계층론

한국어의 음운론적 연구 가운데 고유어, 한자어, 차용어에 대한 어원적 특성이나 음운현상을 다룬 연구는 상당히 많다. 그러나 거시적인 차원에서 고유어, 한자어, 차용어의 특성을 통합하여 살펴보거나, 한국어의 어휘적 계층을 전반적으로 논의한 연구는 드물다. 이러한 연구 가운데 강용순(1998)과 채서영(1999)는 모두 Itô and Mester (1995)에 따라 어휘부를 핵심부와 주변부로 구분하는 논의를 참고하였으나 서로 다른 방식으로 수용하고 있다.

(1) 일본어 충실성 제약 위계 (Itô and Mester 1995: 187)

a. Yamato	b. Sino-Japanese	c. Foreign	d. Alien
SYLLSTRUC	SYLLSTRUC	SYLLSTRUC	SYLLSTRUC
|	|	|	|
NOVOIGEM	NOVOIGEM	NOVOIGEM	**FAITH**
|	|	|	|
NO [P]	NO [P]	**FAITH**	NOVOIGEM
|	|	|	|
POSTNASVOI	**FAITH**	NO [P]	NO [P]
|	|	|	|
FAITH	POSTNASVOI	POSTNASVOI	POSTNASVOI

Itô and Mester (1995)에 의하면 위와 같이 일본어의 음절이나 분절음에 대한 유표성 제약들(SYLLSTRUC, NOVOIGEM 등)은 그 위계가 고정되어 있으나 고유어(Yamato), 한자어(Sino-Japanese), 차용어(Foreign) 등 어휘의 종류에 따라 충실성 제약(FATIH)의 위계가 달라진다고 보았다. 어휘의 핵심부인 고유어(Yamato)로부터 주변부로 확대될수록 위계상 충실성 제약의 위치는 상승한다.

강용순(1998)에서는 한국어는 고정된 유표성 제약의 위계를 가정할 수 없으므로 Itô and Mester (1995)의 핵심부-주변부 가설을 받아들일 수 없다고 보았다. 강용순(1998)에 의하면 한국어는 일본어와 달리 어휘의 종류에 따라 충실성 제약의 위치가 변동되지 않고 충실성 제약들 사이의 위계 자체가 바뀐다.

(2) 한국어 고유어와 차용어의 위계 (강용순 1998: 59)

a. 고유어 '삶'

/salm/	SYLLSTRUC	DEP-IO	MAX-IO
salm	*!		
☞ sam			*
salmɨ		*!	

b. 차용어 ‘토스트’

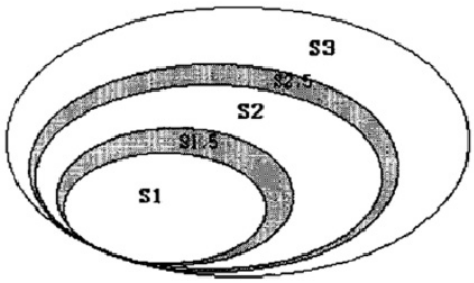
/toust/	SYLLSTRUC	MAX-IO	DEP-IO
toust	*!		
tous		*!	
☞ tousiti			**

위의 도표를 살펴보면 한국어에서 허용되지 않는 자음군을 해소하는 방식이 고유어와 차용어에서 전혀 다르다는 점을 확인할 수 있다. 충실성 제약들인 MAX-IO와 DEP-IO의 위계 차이에 따라 고유어에서는 모음의 삽입보다 자음의 탈락이 선호(DEP-IO≫MAX-IO)되지만 차용어에서는 자음의 탈락보다 모음의 삽입이 선호(MAX-IO≫DEP-IO)된다.

강용순(1998)에서는 핵심부-주변부 가설은 물론 고유어와 한자어의 계층적 구분도 인정하지 않았다. ‘/nl/’이나 ‘/h/+저해음’의 연쇄를 금지하는 제약이나 음절경계 전후로 /-k.n-/, /-p.m-/ 등을 허용하지 않는 ‘음절접촉 법칙’(syllable contact law) 등 지배적인 제약들이 고유어와 한자어를 구별하지 않고 적용되기 때문이다. 다만 한자어에는 ‘갓, 굿, 빗, 젖, 닻, 부엌, 밑, 솥, 팥, 앞, 옆’과 같이 고유어에는 가능한 음절말 장애음이 없으며, ‘그, 느, 드’처럼 [i]로 끝나는 음절도 없다는 점을 지적하였다.

강용순(1998)과 달리 채서영(1999)에서는 Itô and Mester (1995) 핵심부-주변부 이론을 받아들여 일본어와 비슷한 방식으로 한국어의 어휘계층을 설명하였다. 다음과 같이 핵심부(S1)에 고유어가 위치하고 차용어는 주변부 가장자리(S3)에 있으며 고유어와 차용어 사이(S2)에 한자어가 위치하는 어휘계층 구조를 제안하였다. 고유어와 한자어 사이(S1.5), 한자어와 차용어 사이(S2.5)에는 각각 고유어화를 겪고 있는 한자어(호두, 제사)나 두음법칙이 적용되지 않는 한자어(랭면 冷麵) 등이 있다고 설명하였다.

(3) ‘고유어, 한자어, 차용어’의 층위구조 (채서영 1999: 232)



채서영(1999)에서는 Itô and Mester (1995)와 거의 동일한 계층 구조를 가정하고 있으나 위계상 충실성 제약의 위치가 아니라 통시적인 음운변화인 ‘/o/>/u/’를 기준으로 어휘의 계층을 구분하였다. 예를 들어 ‘하루’나 ‘고추’처럼 ‘/o/>/u/’가 완료된 단어는 핵심부에 위

치한 고유어나 이미 고유어화된 한자어이다. ‘삼촌’([samtɕʰon]~[samtɕʰun])처럼 변이가 관찰되는 단어는 S2에서 S1으로 이행되는 단계, 즉 S1.5에 위치하는 한자어로 볼 수 있으며 ‘황도, 라디오’ 등 ‘/o/>/u/’의 변화가 전혀 관찰되지 않는 단어들은 주변부인 S2(한자어)와 S3(차용어)에 위치한다.

(4) /o/>/u/ 모음 상승변화로 본 한국어 어휘의 계층 (채서영 1999: 230)

S1. 핵심부: 고유어 중심의 구어	/o/>/u/
고유어: 하루, 골무, 가죽, 너무, 얼굴, 모두, -루, -두	<변화>
고유어화 된 한자어: 자두, 고추, 후추	<변화>
S1.5 핵심부로 편입되고 있는 영역: 변이를 보임	
고유어화 되고 있는 한자어: 삼촌, 호도, 장고	<변이>
한자어 사용계층/상황의 고유어: -고	<변이>
S2.5 주변부 I : 한자어 중심의 문어	
한자어: 황도, 가족, 결속, 과오	<무변화>
현학적, 고답적 고유어: 자고로, 볼모	<무변화>
한자어로 인식되는 외래어: 남포	<무변화>
S3. 주변부 Ⅱ: 외래어	
일본에서 유래한 외래어: 무대포, 벤또, 시보리	<무변화>
일본을 경유한 서구 외래어: 크락숀, 쓰봉	<무변화>
서구 외래어: 라디오, 데모, 초콜렛	<무변화>

결론적으로 핵심부-주변부 가설의 수용은 어떠한 기준으로 어휘부의 계층구조를 구분하느냐에 달려 있다. 강용순(1998)과 같이 공시적 제약만을 고려한다면, 고유어와 한자어 사이의 계층을 구분하기가 쉽지 않다. 기저형에서부터 음절말 장애음이나 모음 /ɨ/의 사용이 제한되는 한자어를 고유어와 동일한 제약들이나 위계로 설명하기는 어렵기 때문이다. 반면 채서영(1999)와 같이 단어의 통시적인 변화를 고려한다면 고유어와 한자어 혹은 고유어화된 한자어와 그렇지 않은 한자어를 차이를 구분할 수 있다.

본 연구에서는 한국어의 어휘계층을 다룬 기존 연구의 이론적 문제점을 지적하기보다는 다른 방식으로 이러한 한계를 극복해 보고자 한다. 통시적인 음운변화나 공시적인 제약 등 일부분의 양상을 고려하기보다는 ‘음운론적 복잡도’라는 지표를 통하여 고유어, 한자어, 차용어의 특성과 차이를 논의하겠다.

3. 정보이론과 음운론적 복잡도

3.1 정보이론

유표성 이론에서는 음소의 빈도나 분포, 음운규칙의 적용 여부, 자질의 구조 등 유표성을 판단하는 기준이 충돌하는 경우 분절음이나 유표성을 객관적으로 판단하기 어렵다. 또한 유표성의 원리에 어긋나는 예외적 사례들도 관찰되지만 대상을 이분법적으로, 상대적으로 평가하는 기존의 유표성 이론으로는 이러한 예외들을 설명하기 어렵다. Hume (2006)에서는 유표성은 모순적 요소를 포함하고 있다고 비판하면서 유표성 이론을 가시적 지표로 측정할 수 있는 '정보 이론'(Information Theory)으로 대체해야 한다고 주장하였다.

간단히 말하자면 '정보이론'이란 언어에 담긴 정보를 수학적으로 측정하고 이러한 지표를 활용하여 언어 현상을 설명하는 이론이다. 예를 들어 문자메시지에서 흔하게 관찰되는 '축약어'에서는 자음보다는 모음이 생략되는 현상이 언어보편적으로 관찰된다.

(5) 문자메시지 모음 문자의 생략 (이주희 · 박선우 2012: 152)
a. 영어: Texting, The Great Debate → Txtng, The Gr8 Db8
b. 한국어: 감사→ㄱㅅ, 죄송→ㅈㅅ, 응응→ㅇㅇ, 노노→ㄴㄴ

정보이론에 의하면 생략되는 분절음은 정보량의 차이에 의해 결정된다. 대부분의 언어에서는 모음의 종류가 자음보다 훨씬 적으므로 생략된 모음이 무엇인지 자음보다 쉽게 예측할 수 있다. 그러나 종류가 다양한 자음이 생략될 경우에는 본래의 메시지를 제대로 복원하기 어렵다. 아래의 텍스트를 보면 이러한 차이를 확인할 수 있다.

(6) 문자메시지 모음과 자음의 생략 비교 (Crystal 2008: 27)
a. ths sntnc hsnt gt ny vwls. (자음만 표기)
→ This sentence hasn't got any vowels.
b. i eee a o a ooa (모음만 표기)
→ This sentence hasn't got any consonants.

메시지의 수신자가 모음을 쉽게 예측할 수 있는 까닭은 모음의 정보량이 자음보다 적기 때문이다. 쉽게 예측할 수 없는 자음은 정보량이 많다고 볼 수 있다. 예측 가능성은 정보량과 반비례의 관계를 갖는다. 예측 가능성을 분절음이 나타날 확률이라고 본다면 분절음의 정보량은 확률에 반비례한다. 확률이 높아서 쉽게 예측할 수 있는 모음은 정보량이 적은 반면, 확률이 낮아서 예측하기 어려운 자음은 정보량이 많다.[1]

정보이론에서는 정보량과 반비례하는 확률을 기준으로 정보량을

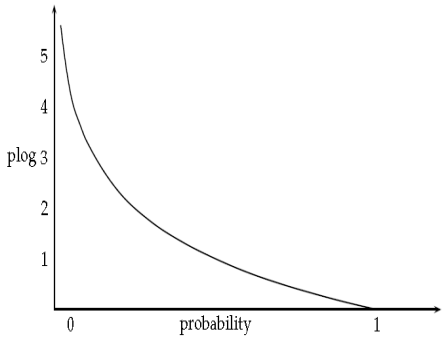
[1] '유표'(有標)와 '무표'(無標)라는 용어의 사전적 의미를 고려한다면 유표적 분절음은 무표적인 분절음에 비하여 더 많은 표지를 갖는 분절음이다. 따라서 정보량이 많은 분절음은 유표적인 분절음과 관련된다고도 볼 수 있다.

표시한다. 확률의 특성상 어떠한 분절음이 나타날 가능성은 언제나 1보다 작은 소수점 이하의 값이므로 두 분절음이 가진 확률의 차이를 직관적으로 깨닫기 어렵다. 따라서 아주 작은 확률의 차이를 극대화하기 위하여 *log*의 지수로 변환하여 사용한다. 음소의 정보량 *Plog*는 다음과 같은 공식으로 구할 수 있다.

(7) 음소의 정보량(*Plog*)을 구하는 공식
a. P(*x*) = 음소 *x*의 출현빈도 / 모든 음소의 출현빈도의 합
b. $Plog(x) = -\log_2 P(x) = \log_2(1/P(x))$

만약 어떠한 말뭉치에서 해당 분절음이 출현할 확률이 1/2이라면 *Plog*는 $-\log_2(1/2)$이므로 1이 된다. 확률이 1/4, 1/8로 낮아지면 *Plog*는 각각 2와 3으로 증가한다. 확률이 낮을수록 정보량은 오르고, 확률이 높을수록 정보량은 낮아지므로 확률과 정보량은 아래의 도표와 같이 반비례의 관계를 갖는다.

(8) 확률과 정보량의 반비례 관계 (Goldsmith 2011: 2)



A probability	its plog
0.5	1
0.25	2
0.125	3
$\frac{1}{16}$	4
$\frac{1}{32}$	5
$\frac{1}{1024}$	10
…	…
$\frac{1}{1,000,000}$	almost 20

예를 들어 '21세기 세종계획 연구교육용 현대국어 균형말뭉치'(국립국어원 2009)에서 모음 /a/는 17,445,870번 나타난다. 이 말뭉치에서 모든 음소의 출현빈도를 합친 값은 217,566,779번이므로 모음 /a/가 나타날 확률은 약 0.08019이고, 이 확률을 (7)의 공식에 대입하여 구할 수 있는 /a/의 정보량(*Plog*)은 약 3.640이다.

(9) 모음 /a/의 *Plog*
a. P(*a*) = 17,445,870 / 217,566,779 = 0.08019
b. $Plog(a) = -\log_2(0.08019) = 3.640$

Goldsmith (2011), Hume and Bromberg (2005), Hong (2006, 2008) 등 정보이론을 적용하는 음운론적 연구에서는 *Plog*의 값을 이용하여 음운현상과 음운체계의 변화를 분석한다. 예를 들어 음운체계 안에 포함된 각 분절음들의 정보량을 계산하면 정보량이 적어서 삽입이 잘 되는 무표적 모음이나 자음을 예측할 수 있다. 또한 동일한 모

음체계를 가지고 있음에도 두 언어의 무표적 삽입모음이 다른 원인도 모음 정보량의 차이로 설명할 수 있다(Hume and Bromberg 2005, Hume 2006).

3.2 음운론적 복잡도

음운론의 기본 단위인 분절음의 확률을 구할 수 있다면 정보량을 구할 수 있고, 정보량을 이용하여 단어의 음운론적 복잡도를 구할 수 있다. 앞서 살펴보았듯이 분절음의 확률은 출현빈도를 측정하여 구하는데 어떠한 언어든 음소의 배열에는 제약이 있으므로 단순한 출현빈도를 계산하기보다는 앞과 뒤에 오는 분절음을 고려하여 빈도를 측정해야 한다(Goldsmith 2002). 예를 들어 '아이'(/ai/)라는 단어의 정보량을 측정하려면 단순히 /a/와 /i/의 출현빈도를 측정하기보다는 단어가 /a/로 시작되는 빈도, /a/에 /i/가 후행하는 빈도, 단어가 /i/로 마무리되는 빈도를 측정해야 한다. 따라서 분절음이 나타나는 단순한 확률이 아니라 후행 분절음 앞에서 출현하는 조건부 확률을 구해야 한다.

조건부 확률을 이용하면 단어를 구성하는 모든 분절음들의 정보량에 대한 평균을 구할 수 있다. 이러한 지표를 '평균 정보량'(average *Plog*)이나 '음운론적 복잡도'(Phonological Complexity, PC) 혹은 '단어의 엔트로피'(entropy)라고 부른다. '음운론적 복잡도'는 다음과 같은 공식으로 구한다.

(10) 음운론적 복잡도 (Goldsmith 2011: 3)

$$PC = 1/n \times \sum_i \{ -\log_2 P(w_i) \} = 1/n \times \sum_i \log_2 Plog(w_i)$$

위의 공식에 따라 단어를 구성하는 각 분절음의 정보량을 합한 다음 단어를 구성하는 분절음의 개수로 나누어 평균 정보량을 구할 수 있는데 이 값이 바로 음운론적 복잡도이다. 분절음의 연쇄를 고려하려 분절음을 두 개씩 묶은 bigram 단위로 조건부 확률을 구한 이후에 *Plog*로 환산하여 평균값을 구한다.

(11) *a*의 뒤에 *b*가 나타날 조건부 확률

$P(b \mid a) = P(ab) / P(a) =$ *ab*의 출현빈도 / *a*의 출현빈도

예를 들어 '아이'(#ai#)의 음운론적 복잡도는 다음과 같은 조건부 확률의 평균으로 구할 수 있다.

(12) 음운론적 복잡도의 계산 과정 (#ai#)

a. 어형의 철자를 분절음 표기로 변환
   아이 → #ai#

b. 분절음 표기를 bigram 단위로 구분
   #ai# → # + #a + ai + i#

c. bigram의 *Plog* 합계
   $Plog(\#) + Plog(\#a) + Plog(ai) + Plog(i\#)$

$= -\{\log_2P(\#) + \log_2P(\#a) + \log_2P(ai) + \log_2P(i\#)\}$
$= -\{\log_2P(\#) + \log_2P(a \mid \#) + \log_2P(i \mid a) + \log_2P(\# \mid i)\}$
$= -\{\log_2C(\#)/n + \log_2C(\#a)/C(\#) + \log_2C(ai)/C(a) + \log_2C(i\#)/C(i)\}$

d. *Plog*의 평균

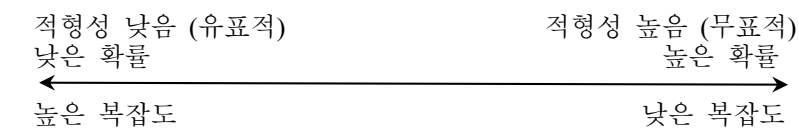
$-1/4 \times \{\log_2C(\#)/n + \log_2C(\#a)/C(\#) + \log_2C(ai)/C(a) + \log_2C(i\#)/C(i)\}$

(※ n: 분절음 전체의 출현빈도, C: 해당 분절음의 출현빈도)

위와 같이 전후의 분절음을 2개씩 묶어서 정보량을 구하는 'bigram 모델'과 달리 'unigram 모델'에서는 전후의 분절음을 고려하지 않고 단일 분절음의 확률만으로 정보량을 측정한다. 다시 말하자면 unigram 모델은 다른 분절음과의 관계를 고려하지 않고 단일 분절음의 정보량만으로 계산하고, bigram 모델은 해당 분절음과 이웃하는 분절음을 함께 묶어서 정보량을 계산한다. 실제로 unigram 모델과 bigram 모델을 통하여 '아이'의 구한 값을 비교해 보면 다소 복잡하더라도 bigram 모델을 적용하는 것이 바람직하다. 21세기 세종계획 말뭉치(국립국어원 2009)를 바탕으로 '아이'라는 단어의 음운론적 복잡도를 구해보면 unigram 모델로는 3.357이고, bigram 모델로는 5.029인데, '아이'가 음절초성(onset)도 없고 모음의 충돌이 일어나는 유표적인 음절구조(V.V)를 가지고 있다는 점을 고려한다면 unigram 모델로 구한 3.357는 이 단어의 음운론적 유표성이 반영된 복잡도로 보기 어렵다. unigram 모델의 경우 다른 모음에 비하여 상대적으로 높은 /a/와 /i/의 출현빈도만으로 평균값을 구하므로 단어의 유표성을 포착하기 어렵다. 반면 bigram 모델로 구한 음운론적 복잡도(5.029)는 '하늘'(2.348), '다리'(2.968)와 같이 모음충돌이 일어나지 않는 단어들에 비하여 상당히 높다.

위와 같이 '음운론적 복잡도'를 이용하면 단어의 유표성을 측정할 수 있다. 음운론적 복잡도는 단어를 구성하는 분절음의 빈도와 순서에 의해 결정된다. 익숙하고 흔한 유형일수록 복잡도가 낮고, 낯설고 드문 유형일수록 복잡도가 높다. 특별하고 유표적인 유형들은 복잡도가 높고, 일반적이고 무표적인 유형들은 복잡도가 낮다. 따라서 음운론적 복잡도와 유표성의 관계는 아래와 같은 도식으로 이해할 수 있다.

(13) 음운론적 복잡도와 유표성 (Hong 2006: 212)



유표적인가 무표적인가를 이분법적으로 판정하는 기존의 유표성 이론과 달리 음운론적 복잡도는 단어나 분절음의 연쇄가 음운론적으로 얼마나 자연스러운지 구체적으로 보여준다. 따라서 복잡도를 활용하면 적형과 비적형, 유표적 형태와 무표적 형태의 이분법적 구분보다 정밀한 평가가 가능하고 개별적인 단어들 사이의 유표성

도 비교할 수 있다. 다음의 표는 음운론적 복잡도를 이용하여 영어의 단어들을 분석한 결과이다.

(14) 영어 단어의 음운론적 복잡도 순위 (Goldsmith 2011: 4)

rank	orthography	phonemes	$avg.\ plog_2$
1	the	ðə	1.93
2	hand	hǽnd	2.15
12,640	plumbing	plʊ́mɪŋ	3.71
12,642	Friday	fraýdɪ̯	3.71
25,281	tolls	tólz	4.01
25,282	recorder	r ɪ̯ k ó r d ɚ̯	4.01
37,922	overburdened	óvɚ̯bɚ́dənd	4.32
37,923	Australians	ɔ̆streýlyənz	4.32
50,563	retire	rɪ̆taýr	4.75
50,564	poorer	púrɚ̯	4.75
63,200	eh	ɛ́	9.07
63,201	Oahu	óáhʊ̆	9.21

위의 표를 살펴보면 각 단어의 복잡도를 확인할 수 있다. 영어에서 가장 무표적인 단어는 the(1위, 1.93)와 hand(2위, 2.15)이다. 반면 가장 유표적인 단어는 retire(50,563위, 4.75), Oahu(63,201위, 9.21)와 같이 순위가 낮은 단어들이다. 이와 같이 단어의 음운론적 복잡도를 통해 유표적인 단어와 무표적인 단어를 구분할 수 있고, 유표성의 차이도 확인할 수 있다.

물론 이러한 음운론적 복잡도의 장점은 기존 유표성 이론에서 가정된 유표성과 음운론적 복잡도가 동일한 성격의 지표라는 것을 전제할 경우에만 주장할 수 있다. Hume(2006)에 의하면 기존의 유표성 이론에서는 음운의 분포와 빈도, 음운현상의 대상과 결과, 언어습득, 크레올(creole)의 형성과정 등을 고려하여 유표성을 결정한다. 이러한 기준들 가운데 가장 객관적인 것은 구체적인 수치로 측정할 수 있는 분절음의 유형빈도와 출현빈도인데 음운론적 복잡도는 분절음의 출현빈도를 측정하고 확률적으로 분석한 결과이다. 따라서 음운론적 복잡도는 기존의 유표성 이론에서 가정하고 있는 유표성의 개념과 무관하지 않다.

문제는 음운론적 복잡도의 계산에는 음운규칙이나 언어습득 과정 등 유표성을 판정하는 다른 기준들이 고려되지 않는다는 점이다. 이러한 결정은 음운론적 유표성을 객관적으로 평가하기 위한 불가피한 선택이라 할 수 있다. 만약 유표성을 판정할 때, '무표적인 분절음들이 약화, 탈락, 동화, 도치의 대상이 되거나 삽입, 중화 등의 결과로 출현한다'는 기준(Hume 2006: 2)을 고려한다면 두 가지 문제가 발생할 수 있다. 첫째는 규칙의 적용에 일관성이 없는 경우

가 있다. 예를 들어 약화나 탈락에서는 무표적으로 보이는 분절음이 동화나 삽입에서는 유표적일 수 있다. 또한 빈도가 높아서 무표적으로 보이는 분절음이 음운규칙에서는 유표적으로 행동하는 경우도 있다. 둘째 동일한 규칙이라도 언어에 따라서 다른 방식으로 적용될 수 있다. 예를 들어 '스리랑카-포르투갈 크레올'(Sri Lankan Portuguese Creole)에서는 언어보편적으로 유표적이라고 가정해 온 순음과 연구개음이 무표적인 설정음에 동화된다(Hume 2003: 301).[2]

결론적으로 유표성과 음운론적 복잡도는 공통점과 차이점을 갖고 있다. 음운론적 복잡도란 지표는 분절음의 빈도와 같이 객관적인 기준만을 적용하므로 기존의 유표성 이론에 나타나는 일관성의 문제를 피할 수 있으며 언어개별적 유표성도 분석할 수 있다. 다만 분절음의 빈도와 확률만을 고려하므로 음절의 구조나 단어의 길이, 음운규칙과 관련된 유표성을 포착하지 못하는 경우가 발생할 수 있다.

4. 어휘계층별 복잡도 분석

본 장에서는 한국어의 고유어, 한자어, 차용어를 대상으로 복잡도의 차이를 비교하고 분석해 보겠다. 일단 한국어에서 한자어와 차용어는 대부분 명사로 사용되므로 분석대상은 일반명사로 제한하였다. 강범모 · 김흥규(2004)에서 일반명사로 분류된 단어들 가운데 가운데 출현빈도가 가장 높은 고유어, 한자어, 차용어를 각각 500개씩 총 1,500개의 단어를 선정하였다. '선생님'이나 '달러화'와 같은 혼종어(hybrid)는 제외하였다.

(15) 분석 단어 목록 예시 (각 30개, 빈도순)
a. 고유어: 사람, 때, 말, 일, 속, 집, 앞, 소리, 생각, 아이, 나라, 눈, 뒤, 곳, 위, 마음, 어머니, 동안, 안, 날, 길, 다음, 얼굴, 아버지, 이번, 손, 사이, 돈, 이야기, 몸
b. 한자어: 문제, 사회, 경우, 자신, 시간, 사실, 정부, 정도, 인간, 여자, 전, 세계, 점, 문화, 관계, 운동, 교육, 학교, 여성, 경제, 시대, 이상, 지역, 지금, 생활, 의미, 국가, 학생, 역사, 자리
c. 차용어: 컴퓨터, 프로그램, 버스, 아파트, 팀, 이미지, 텔레비전, 그룹, 아나운서, 커피, 이데올로기, 뉴스, 에너지, 게임, 올림픽, 비디오, 서비스, 가스, 리얼리즘, 모델, 스포츠, 택시, 라디오, 소프트웨어, 퍼센트, 캠페인, 프로(프로페셔널), 시스템, 아이디어, 드라마

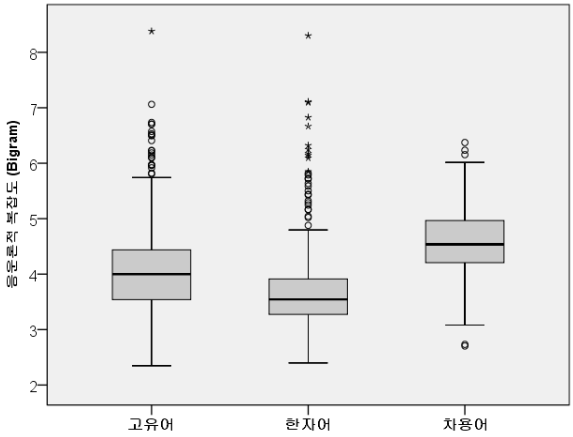
빈도가 가장 높은 단어를 기준으로 고유어, 한자어, 차용어를 각각 500개씩 bigram 모델에 따라 분석하고 평균을 구하였다. 복잡도

[2] 박선우(2011)에서는 유표성 이론에서 제안된 여러 기준으로 한국어 순음, 설정음, 연구개음의 조음위치별 유표성을 평가할 경우, 기준에 따라 모순된 결과가 나타날 수 있다는 점을 보여주고 있다.

계산의 근거가 되는 음소와 bigram의 확률은 국립국어원에서 '21세기 세종계획'을 통해 1999년에서 2004년 사이에 구축한 형태소 분석 말뭉치에서 추출하였다.[3] 실제의 음성과는 차이가 있는 문어코퍼스를 활용하였으나, 문제점을 줄이기 위하여 한글로 표기된 단어를 분절음의 표기로 변환하여 분석하였다. 음가가 없는 음절초성의 'ㅇ'은 (12b)와 같이 삭제하고 한글에서 하나의 단위로 표기되는 이중모음 '야, 여, 요, 유, 의, 위' 등은 'ja, jə, jo, ju, ɨj, wi'로 구분하여 분석하였다.

분석 결과 음운론적 복잡도는 한자어가 가장 낮고 차용어가 가장 높았으며 고유어는 한자어보다 높고, 차용어보다는 낮았다. 아래의 도표에서 확인할 수 있듯이 한자어(0.715)나 차용어(0.585)에 비하여 고유어의 표준편차(0.798)는 큰 편이었다.

(16) a. 고유어, 한자어, 차용어의 음운론적 복잡도의 상자도표



b. 고유어, 한자어, 차용어의 평균과 표준편차

	평균(mean)	개수	표준편차	최솟값	최댓값
고유어	4.085	500	.798	2.348	8.381
한자어	3.699	500	.715	2.398	8.301
차용어	4.560	500	.585	2.705	6.373

복잡도 평균에 대하여 일원배치분산분석(Oneway ANOVA)을 시행한 결과 고유어, 한자어, 차용어의 차이는 모두 유의미한 것으로

[3] 익명의 심사위원으로부터 실제의 음성과 거리가 있는 문어 코퍼스, 즉 단어의 음성이 아니라 형태적 표기를 바탕으로 음운론적 복잡도를 구하는 것은 문제가 있다는 지적을 받았다. 실제의 발화를 고려한다면 음성 코퍼스를 이용하는 것이 최선의 방법이지만, 음성을 바탕으로 분절음의 빈도와 확률을 계산할 수 있는 한국어의 음성코퍼스가 없는 상황에서는 문어 코퍼스를 이용하여 음운론적 복잡도를 구하는 것이 차선의 연구방법이라 생각된다. 확률을 추출하는 방법에 대한 자세한 설명은 Hong (2006, 2008)을 참조하기 바란다.

검증되었다. 세 집단 사이의 유의확률은 모두 0.1%이하(P<.000)였다.

(17) 음운론적 복잡도에 대한 일원배치분석 결과 (Dunnett T3)

	평균차	유의확률	95% 신뢰구간	
			하한값	상한값
고유어:한자어	.38646951	.000	.2718775	.5010615
한자어:차용어	.86065624	.000	.7618520	.9594605
차용어:고유어	.47418674	.000	.3682943	.5800792

음운론적 복잡도의 분석에서 주목할 점은 고유어보다 한자어의 복잡도가 낮다는 것이다. 이러한 결과는 채서영(1999)의 핵심부-주변부 가설과 차이가 있다. 복잡도와 유표성의 관계에 의하면 한국어의 어휘계층에서 핵심부에 위치하는 고유어가 한자어보다 음운론적으로 유표적인 셈이다. 그러나 고유어와 한자어를 구성하는 분절음이나 음절구조를 살펴보면 고유어가 한자어보다 유표적이다. 고유어와 달리 한자어에서는 설정 장애음인 /t/, /s/, /tɕ/와 유기음 /k^h/, /t^h/, /p^h/, /$tɕ^h$/ 경음 /k'/, /s'/가 음절말에 올 수 없다. 물론 이들 분절음 가운데 /t/를 제외한 나머지는 고유어의 음절말에도 올 수 없다. 그러나 고유어의 제약은 출력형에 국한되는 반면 한자어에서는 입력형부터 이러한 음절말 분절음들이 제한된다. 강용순(1998: 61)에서 지적한 바와 같이 한자어에서는 '그, 느, 드, ...'와 같이 모음 [ɨ]으로 끝나는 음절도 제한된다. 고유어에 비하여 이중모음의 분포도 상당히 제한적이며 이중모음 /wə/는 전혀 나타나지 않는다. 본 연구에서 정밀한 논의를 하기는 어렵겠으나 한자어의 음운체계에는 한자음이 한국어에 도입되던 시기 중국 한자음이 가진 제약과 한국어 음운체계의 제약이 동시 적용되었다고 볼 수 있다.[4]

이러한 제약들을 고려한다면 한자어에 포함되는 분절음과 음절의 종류는 고유어보다 단순할 수밖에 없다. 이러한 결론은 본 연구에서 선정한 1,500개의 단어를 분석한 결과를 통해서도 확인할 수 있다. 아래의 표는 본 연구에서 선정한 단어를 대상으로 고유어, 한자어, 차용어에서 관찰되는 bigram의 유형과 출현빈도, 음절수 등을 분석한 결과이다.

(18) bigram의 유형빈도, 출현빈도의 합계, 평균 출현빈도

	유형빈도	출현빈도 합계	평균 출현빈도
고유어	360	2,805	7.791
한자어	296	3,139	10.605
차용어	298	3,613	12.124

[4] 한국어와 유사한 어휘계층을 갖고 있는 일본어에 대한 연구에서도 한자어(Sino-Japanese)가 고유어(Yamato)보다 무표적이라는 논의가 있다. Kawahara et al. (2003)에서는 일본 한자어의 세 가지 특성을 근거로 한자어가 고유어보다 무표적이라고 주장하였다. 세 가지 특성은 (1) 단어의 길이 제한(size restriction), (2) 어휘적 악센트의 상실(the non-preservation of lexical accents), (3) 제2음절에 대한 제약(segmental restriction in the second syllable)이다.

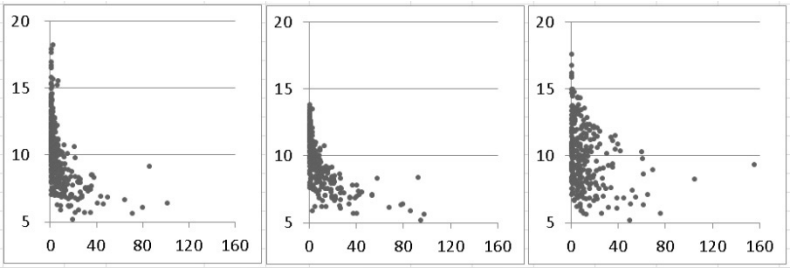
‘21세기 세종계획 현대국어 균형말뭉치’(국립국어원 2009)에서 발견된 bigram의 유형은 모두 794가지였다. 이들 가운데 고유어 500개에서 관찰된 bigram은 360가지였으나, 한자어 500개, 차용어 500개에 나타난 bigram은 각각 296가지, 298가지로서 300개에 미치지 못하였다. 이러한 결과에 의하면 고유어는 한자어나 차용어보다 다양한 분절음과 bigram으로 구성된 반면 한자어나 차용어에서는 고유어보다 적은 종류의 bigram이 반복적으로 사용된 셈이다. 각각의 bigram이 출현하는 평균빈도를 살펴보면 고유어(7.791번)가 한자어(10.605번)나 차용어(12.124번)보다 적은 편이다. 결론적으로 한자어는 고유어와 달리 비슷한 유형의 bigram이 반복적으로 나타나는 무표적인 구조를 갖고 있다.

(18)을 통하여 지금까지 한국어의 어휘부에서 핵심적 부분이라고 가정되어 온 고유어의 음운론적 복잡도가 한자어보다 높은 까닭을 이해할 수 있다. 그러나 (18)은 한자어와 차용어 사이 복잡도의 차이를 보여주지 못한다. 차용어에서 관찰되는 bigram의 유형빈도는 한자어와 유사하며 각각의 bigram이 출현하는 평균빈도도 오히려 한자어보다 높다. 따라서 한자어와 마찬가지로 차용어에서도 적은 수의 bigram이 반복적으로 나타난다고 볼 수 있다. 그럼에도 불구하고 차용어의 복잡도가 한자어는 물론 고유어보다도 높은 이유는 무엇일까?

차용어에 나타나는 bigram의 유형빈도는 한자어와 비슷하므로 복잡도가 높게 나타나는 까닭은 차용어에서 관찰되는 bigram의 정보량에서 찾을 수밖에 없다. 즉 차용어에는 고유어만큼 다양한 유형의 bigram이 나타나지는 않으나 각각의 bigram이 갖는 정보량이 고유어나 한자어보다 높다고 예상할 수 있다. 다음의 도표는 고유어, 한자어, 차용어에서 관찰되는 bigram의 정보량을 출현빈도를 기준으로 표시한 것이다.

(19) bigram의 출현빈도(X축)와 *Plog*(Y축)

a. 고유어 b. 한자어 c. 차용어

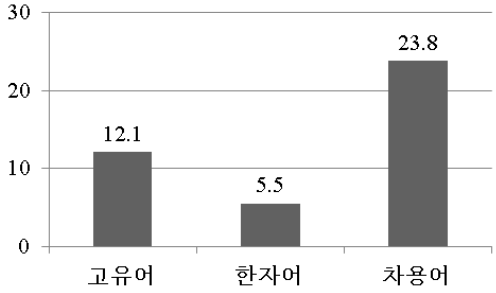


대체로 세 가지 도표 모두 *Plog*의 정의에 따라 bigram의 빈도와 *Plog*가 반비례 곡선에 가까운 모양을 띄고 있으나 다소 차이가 있다. 고유어의 bigram은 *Plog* 값을 기준으로 5부터 20 사이에 들이 넓게 분포하고 있다. 한자어는 대체로 정보량이 낮은 편인 Y축 15 이하의 구간에 분포하고 있다. 반면 차용어는 한자어와 비교하여

정보량이 10보다 높은 구간에도 상당히 많다. 따라서 한자어에 비하여 차용어는 *Plog*가 상당히 높은 bigram으로 구성되었다고 볼 수 있다.

어종별로 *Plog*가 높은 bigram의 비율을 측정해 보면 차용어에서 나타나는 bigram의 특성을 확인할 수 있다. 고유어, 한자어, 차용어 사용된 bigram 가운데 중복되는 유형을 제외하면 1,500개의 단어에 사용된 bigram의 유형빈도는 477개이고, 이들의 *Plog* 평균은 10.582 이었다. 아래의 도표는 고유어, 한자어, 차용어에 나타나는 bigram 가운데 평균보다 높은 출현빈도의 비율을 표시한 것이다.

(20) *Plog*가 평균(10.582)보다 높은 bigram 유형빈도의 비율 (%)



고유어의 2,805개 가운데 10.582보다 높은 bigram의 비율은 12.1%였다. 한자어에서는 평균 이상의 비율이 고유어의 절반 이하인 5.5%인 반면 차용어에서는 고유어의 두 배 가까이 되는 23.8%였다. 따라서 차용어에 나타나는 bigram의 유형빈도는 한자어와 비슷하지만 상대적으로 *Plog*가 높은 bigram의 비율이 높다. 이로 말미암아 차용어의 복잡도는 한자어는 물론 고유어보다도 높다.

한편 고유어, 한자어, 차용어에 사용된 bigram을 음절의 개수에 따라 분석해 보면 음절의 구조와 단어의 유표성은 기존의 이론과는 다소 다르다는 점을 확인할 수 있다. 다음의 표는 고유어, 한자어, 차용어를 구성하는 음절이 몇 개의 bigram으로 구성되었는지를 분석한 결과이다.

(21) bigram의 유형빈도와 출현빈도

	bigram 합계	음절수 합계	음절당 bigram 개수
고유어	2,805	968	2.898
한자어	3,139	997	3.148
차용어	3,613	1,455	2.483

분석 결과에 의하면 하나의 음절을 구성하는 평균 bigram의 개수는 음운론적 복잡도와 반대의 순위를 갖는다. 복잡도가 높은 차용어가 가장 낮고, 복잡도가 낮은 한자어가 가장 높다. birgram의 개수는 음절을 구성하는 분절음의 개수에 1을 더한 값이다. 예를 들어 V는 2개의 bigram으로 CV, GV, VC는 3개의 bigram으로, CVC, CGV, GVC,

VCC는 4개의 birgram으로 구성된다. 따라서 음절을 구성하는 bigram의 개수만을 고려한다면 한자어를 구성하는 음절이 가장 복잡하고, 오히려 차용어를 구성하는 음절이 가장 단순한 편이다. 다만 이러한 방식의 분석은 무표적인 CV 음절과 유표적인 VC 음절을 같은 개수의 bigram으로 취급한다는 단점을 갖는다. 그러나 복잡도가 낮은 단어들을 살펴보면 CV와 같은 무표적 구조가 음운론적 복잡도에는 큰 영향을 주는 것 같지 않다.

(22) 복잡도가 낮은 10개의 단어
a. 고유어: 하늘, 그늘, 글, 며느리, 바늘, 한글, 마련, 들, 달, 실 (평균 2.631)
b. 한자어: 한, 면, 전, 선, 잔, 산, 신, 시간, 정신, 반 (평균 2.621)
c. 차용어: 신(scene), 정글, 골(goal), 드라마, 바나나, 봉고, 핸들, 비전, 프로, 머신 (평균 3.097)

각각의 어휘계층에서 가장 복잡도가 낮은 단어들을 10개씩 꼽아보면 고유어, 한자어, 차용어의 특성이 조금씩 다르다. 고유어의 경우 /l/(ㄹ)로 끝나는 단어들이 많다. '며느리'와 '마련'을 제외하면 10개 가운데 8개 단어가 /l/로 끝난다. 반면 한자어는 10개 단어가 모두 /n/(ㄴ)으로 끝나는 단어들이며 '시간, 정신'을 제외한 8개 단어가 모두 1음절이다. 차용어에서는 '드라마, 바나나, 프로'와 같이 무표적인 CV 음절구조를 가진 단어들을 발견할 수 있으나 음절말음을 가진 단어들도 많다. 반면 위의 30개 단어들 가운데 음절의 초성 없이 모음으로 시작되는 단어는 전혀 없다. 따라서 음절의 초성이 있느냐 없느냐는 복잡도에 영향을 미치는 것으로 보인다. 단어들 가운데 복잡도가 가장 높은 단어들을 보면 음절 초성이 없는 단어들이 많이 포함되어 있다.

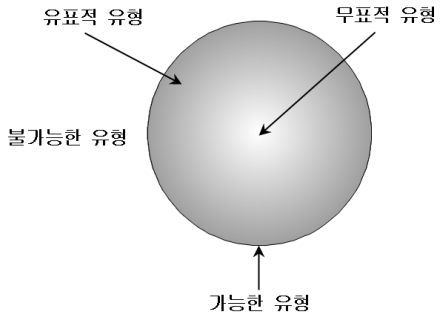
(23) 복잡도가 높은 10개의 단어
a. 고유어: 쇠, 앞, 왼쪽, 애, 열쇠, 부엌, 끝, 잎, 뜻밖, 밖 (평균: 6.763)
b. 한자어: 이외, 의회, 해외, 범죄, 외국, 사회, 최초, 대회, 기회, 최고 (평균: 6.698)
c. 차용어: 에이즈, 룸, 에어, 램, 오디오, 오브, 라이프, 라디오, 레코드, 렌즈 (평균: 5.992)

위의 단어들은 어휘계층별 500가지 단어들 가운데 복잡도가 가장 높은 것들이다. 복잡도가 높다는 공통점을 가진 단어들이지만 고유어, 한자어, 차용어의 음운론적 성격은 다르다. 고유어의 경우 '앞, 부엌, 끝, 잎, 뜻밖, 밖'에는 음절말에 ㅌ, ㅍ, ㄲ, ㄸ과 같은 유기음과 경음이 분포하고 있으며 '쇠, 왼쪽, 열쇠'에는 이중모음 '외'가 포함되어 있다. 복잡도가 높은 한자어들은 고유어와 달리 음절종성이 거의 없으며 모두 이중모음 '외'를 갖고 있다. 복잡도가 높은 차용어는 대부분 고유어와 한자어에 적용되는 두음제약을 적용

되지 않은 단어들(룸, 램, 라이프, 라디오, 레코드, 렌즈)과 모음충돌이 일어나는 단어들(에이즈, 에어, 오디오)로 구분된다. 또한 앞서 지적하였듯이 복잡도가 높은 단어들 가운데는 모음으로 시작되는 단어들이 상당히 많다.[5]

한국어의 일반 명사들을 고유어, 한자어, 차용어로 나누어 음운론적 복잡도를 기준으로 검토해 본 결과 단어의 유표성에는 특정한 제약이나 음절구조만으로는 설명하기 어려운 다양한 음운론적 현상들이 관련되었음을 확인할 수 있다. 유기음이나 경음과 같이 유표적인 음절 종성과 특이한 이중모음, 어두 자음의 유무와 모음충돌, 두음제약의 적용과 같이 다양한 음운론적 특성들이 반영된다. 복잡도를 통하여 확인할 수 있는 거시적인 특징은 고유어에는 복잡도가 낮은 단어와 높은 단어들이 섞여 있으나, 한자어는 복잡도가 낮은 단어를 중심으로 구성되어 있고, 차용어는 복잡도가 높은 단어를 중심으로 구성되어 있다는 점이다. 유표성을 기준으로 본다면 한국어의 어휘계층은 무표적인 한자어와 유표적인 차용어, 그리고 무표적인 단어와 유표적인 단어들이 혼재되어 있는 고유어로 구성되어 있은 셈이다.

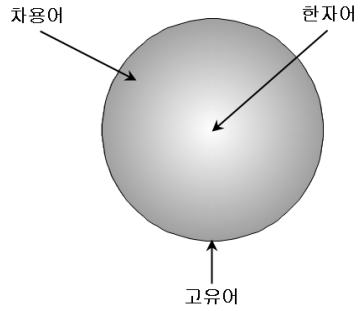
(24) 유표성과 언어학적 유형 (Hume 2011: 82)



위의 그림은 Hume (2011)에서 유표성을 기준으로 언어학적 유형을 표시한 도식이다. 원의 안과 밖은 언어에서 실제로 나타나는 가능한 유형과 관찰되지 않는 불가능한 유형을 의미한다. 원 안의 중심은 무표적 유형, 원 안의 주변은 유표적 유형의 영역이다. 유표성의 도식을 이용하여 한국어의 어휘계층 구조를 표현한다면 다음과 같은 도식이 될 것이다.

[5] 박선우(2013:213-221)에 의하면 '음운론적 복잡도'는 단어의 '출현빈도'나 '유형빈도'와는 상관이 없는 반면, 청각적 자극에 대한 적합도의 평가 결과와는 어느 정도의 상관관계를 보여준다.

(25) 음운론적 유표성과 한국어의 어휘계층



이러한 어휘 계층의 분포는 (3)에서 살펴본 채서영(1999: 232)의 동심원 구조와 다르다. 음운론적 복잡도가 낮은 한자어는 원 안의 중심부에, 복잡도가 높은 차용어는 주로 외곽부의 주변부에 분포한다. 채서영(1999)에서 고유어는 어휘부의 중심에 위치하였으나 (25)에서는 중심부와 주변부에 골고루 분포한다. 고유어의 경우 복잡도가 높은 단어들과 낮은 단어들이 혼재되어 있으므로 무표적인 한자어와 유표적인 차용어를 모두 포괄하는 원 전체의 영역을 차지하고 있다.

5. 결론

지금까지 Goldsmith (2002, 2011)에서 제안된 '음운론적 복잡도'란 지표를 활용하여 고유어, 한자어, 차용어로 구성된 한국어 어휘계층을 거시적으로 분석해 보았다. 음운론적 복잡도를 통하여 한국어의 어휘계층에 대하여 두 가지 새로운 사실을 확인할 수 있었다. 첫째, 음소의 확률을 기반으로 측정한 음운론적 복잡도의 분석 결과 고유어가 한자어보다 오히려 유표적이었다. 둘째 한자어와 차용어는 각각 음운론적 복잡도가 낮은 어휘와 높은 어휘를 중심으로 구성되었으나 복잡도의 편차가 큰 고유어는 복잡도가 낮은 어휘로부터 가장 높은 어휘까지 음운론적으로 유표적인 단어와 무표적인 단어가 혼재되어 있었다. 이러한 결과는 구어나 문어의 사용, 어휘 차용의 시기와 방법 등 언어외적 원인이나 두음법칙, 구개음화 등의 음운규칙이나 제약만으로는 파악하기 어려운 한국어 어휘계층의 거시적 특성이라 할 수 있다.

본 연구에서는 일부의 자료만을 다루었으나 앞으로 정보이론을 기반으로 하는 다양한 분석방법과 지표를 도입하고 분석대상 자료를 확대한다면 한국어 어휘계층의 거시적 특징에 대하다 보다 상세한 결과를 얻을 수 있을 것이다. 예를 들어 형태분석 말뭉치 활용하여 분석 대상을 모든 일반명사로 확대한다면 어휘의 출현빈도나 단어를 구성하는 음절 개수와 음운론적 복잡도의 상관관계에 대한 통계적 분석이 가능하다. 또한 단어의 무표성과 관련된 지표인 '상호 정보량'(Mutual Information, Goldsmith 2002, Hong 2008)을 '음

운론적 복잡도'와 함께 적용한다면 보다 정밀한 분석이 가능할 것으로 기대된다.

참고문헌

강범모 · 김흥규. 2004. *한국어 형태소 및 어휘사용 빈도의 분석 2*. 고려대학교 민족문화연구원.

강용순. 1998. 한국어 어휘부 구조. *음성 · 음운 · 형태론 연구* 4, 55-67.

국립국어원. 2009. *21세기 세종계획 연구교육용 현대국어 균형말뭉치*. CD-ROM.

박선우. 2011. 한국어 조음위치의 유표성에 대한 검토. *한국어학* 53, 249-279.

_____. 2013. 한국어의 음운현상과 음운론적 복잡도: 모음충돌과 이중모음화 현상의 비교를 중심으로. *한국어학* 59, 203-225.

이주희 · 박선우. 2012. 한국어 문자메시지의 표기와 특성: 20대 대학생을 중심으로. *음성 · 음운 · 형태론 연구* 18.1, 131-161.

채서영. 1999. 음운변화에 나타난 한국어 어휘의 층위구조. *음성 · 음운 · 형태론 연구* 7, 217-236.

CRYSTAL, DAVID. 2008. *Txtng, The Gr8 Db8*. Oxford University Press

GOLDSMITH, JOHN. 2002. Probabilistic models of grammar: Phonology as information minimization. *Phonological Studies* 5, 21-46.

_____. 2011. Information Theory for linguists: A tutorial introduction. Paper presented at the workshop on Information Theory in linguistics at the LSA Summer Institute.

HONG, SUNG-HOON. 2006. What do phonologically good and bad words look like in Korean? *Inquiries into Korean Linguistics* Ⅱ, 167-178.

_____. 2008. Hiatus resolution in Korean: From the perspective of Information Theory. *Language and Linguistics* 41, 417-436.

HUME, ELIZABETH. 2004. Deconstructing markedness: A predictability-based approach. Proceedings of the Berkeley Linguistic Society 2004.

_____. 2006. Language specific and universal markedness: An Information-theoretic approach. Paper presented at the Colloquium on Information Theory and Phonology, LSA Wintermeeting 2006.

_____. 2011. Markedness. In M. Van Oostendorp, C. Ewen, E. Hume and K. Rice (eds.). *Companion to Phonology*, 79-106. Blackwell.

HUME, ELIZABETH and ILANA BROMBERG. 2005. Predicting epenthesis: An Information-theoretic account. Paper presented at the 7th Annual Meeting of the French Network of Phonology.

ITÔ, JUNKO and ARMIN MESTER. 1995. The core-periphery structure of the lexicon and constraints on reranking. In Jill N. Beckman, Laura Walsh-Dicky and Suzanne Urbanczyk (eds.). *University of Massachusetts Occasional Papers in Linguistics* 18, *Papers in Optimality Theory*, 181-

209. Amherst, MA: GLSA.
KAWAHARA, SHIGETO, KOHEI NISHIMURA and HAJIME ONO. 2003. Unveiling the unmarkedness of sino-Japanese. *Japanese/Korean Linguistics* 12, 140-151. Stanford: CSLI.
SHANNON, CLAUDE. 1949. *A Mathematical Theory of Communication*. Urbana: University of Illinois Press.

박선우
447-791 경기도 오산시 양산동 411번지
한신대학교 正祖교양대학
e-mail: sunwoopark@naver.com

홍성훈
130-791 서울시 동대문구 이문동 270번지
한국외국어대학교 영어학과
e-mail: hongshoon@hufs.ac.kr

변군혁
130-791 서울시 동대문구 이문동 270번지
한국외국어대학교 언어연구소
e-mail: khbyun70@hanmail.net

received: March 28, 2013
revised: August 7, 2013
accepted: August 12, 2013