Linguistic Research 33(2), 349-374
DOI: 10.17250/khisli.33.2.201606.007

정보이론 기반 한국어 준말의 분석*

홍성훈** · **박선우*****
(한국외국어대학교 · 계명대학교)

Hong, Sung-hoon and Sunwoo Park. 2016. Korean reduced words from the perspective of information theory. *Linguistic Research* 33(2), 349-374. In casual speech of Korean, words are often reduced to shorter forms by reducing one or more syllables from the original words. This word reduction comes in three types: simple deletion of one or more syllables from the source word, reduction motivated by hiatus avoidance, and reduction created by deletion of subsyllabic segment(s) with subsequent resyllabification. The third type, which we call "subsyllabic reduction (SSR)," is particularly interesting because it is not clear what motivates this specific deletion of segment strings and the subsequent phonological restructuring. This paper examines the issues surrounding SSR from the perspective of information theory (Shannon 1948) and proposes that SSR occurs to improve string well-formedness and collocational strength between segments, which are measured respectively by information-theoretic notions of 'positive logarithm' (*plog*; Goldsmith 2002, 2011) and 'mutual information'. To elaborate this proposal, we compiled a list of reduced words created by SSR and examined how *plog* and mutual information vary with SSR. We obtained a partially positive result: Among the five types of SSR, the values of *plog* and mutual information change as we expected in three types, which account for 270 out of 309 cases of SSR. We suggested a tentative explanation for the remaining cases based on reverse derivation and word-edge effect. (Aitchison 2011). **(Hankuk University of Foreign Studies · Keimyung University)**

Keywords reduced word, phonological deletion, information theory, phonological complexity, mutual information

1. 머리말

'준말'은 음운론적 분석이 쉽지 않은 대상이다. 다양한 유형의 단어들이 준말이라는 용어로 불리고 있으며 이들을 단일한 원리, 규칙, 제약 등으로 일관성 있게

* 이 연구는 2015학년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것입니다. 본 논문의 내용에 대하여 조언해 주신 익명의 심사위원들께 깊이 감사드립니다.
** 주저자, *** 교신저자

설명하기 어렵다. 음운론적 기준을 바탕으로 한국어의 준말은 다음의 세 가지 유형으로 분류할 수 있다. 본 연구에서는 아래 세 가지 유형 가운데 (1c) '음절축약형 준말'을 대상으로 정보이론(Information Theory, Shannon 1948) 기반의 음운론적 분석을 진행하고자 한다.

(1) 한국어 준말의 유형 (송철의 1993, 이지양 1998, 이지양 2003, 정희창 2004)
   a. 음절탈락형: 국산(국산품), 어사(암행어사), 노조(노동조합)
   b. 모음충돌 회피형: 베다(베이다), 겉띄다(겉뜨이다), 새(사이), 에구(어이구), 눼(누에), 껴(끼어)
   c. 음절축약형: 물골(물고랑), 좀(조금), 건들다(건드리다), 골짝(골짜기), 진디(진드기)

(1a)는 단어를 구성하는 1개 이상의 음절이 떨어져나간 유형이다. 이러한 유형에서는 일반적인 음운현상과 달리 탈락이 일어나는 음운론적 동기나 규칙성을 찾기가 어렵다. (1b)는 모음충돌이 일어나는 경우인데 충돌이 일어나는 두 모음 가운데 일부가 탈락, 융합, 활음화를 겪으면서 모음충돌을 회피하는 유형들이다. (1a)와 달리 모음충돌을 회피하기 위하여 음절의 개수가 줄어든다는 음운론적 동기가 뚜렷하다. (1c)는 모음충돌과 상관없이 음절핵을 포함한 음절의 일부분이 탈락되고 재음절화가 일어나는 유형들이다.

(1c) 음절축약형 준말들은 (1b)와 비교하여 음운론적 동기가 뚜렷하지 않으나 음절구조상 몇 가지 종류로 유형화할 수 있으므로 준말의 음운론적 구조를 어느 정도는 예측할 수 있다. 예를 들어 (1a)와 달리 (1c)에서는 본딧말의 첫 음절이 준말에서도 그대로 유지된다. 따라서 (1c) 음절축약형 준말은 음절 탈락의 음운론적 동기를 찾기 어려운 (1a)와 모음충돌을 피하려는 음운론적 동기가 뚜렷한 (1b) 사이에 있는 유형이라 할 수 있다. 준말의 음절구조를 어느 정도는 예측할 수 있으나 기존의 음운론적 규칙이나 제약만으로는 설명이 쉽지 않다.

준말과 관련된 음운론적 연구들(송철의 1993, 이지양 1998, 김성규 1999, 이지양 2003, 정희창 2003, 정희창 2004, 김선철 2011)에서는 대부분 일정한 기준과 조건에 따라 준말을 정의하고, 준말의 형성과정에서 나타나는 음운현상의 경향을 분류하여 정리하고 있다. 준말은 필수적인 음운현상의 적용 결과가 아니며, 일반적으로는 본딧말과 공존한다. 따라서 본딧말과 준말을 기저형과 표면형, 입력형과 출력형으로 가정하여 분석한 기존 논의들은 다음과 같은 두 가지 한계를 가질 수밖에 없었다.

첫째, 준말의 목록 전체를 다루지 않았기 때문에 준말의 형성과 관련된 음운론

적 경향을 분류하여 기술할 수 있으나, 이들 가운데 어떠한 유형들이 가장 높은 비율을 차지하는지, 일반적인 유형과 특이한 유형들 사이 그 비율의 차이가 얼마나 되는지 설명할 수 없었다. 예를 들어, 송철의(1993:14)에서는 '가을 → 갈:', '다음 → 담:'과 같은 유형들은 둘째 음절의 모음이 /ɨ/이거나, 음절말 자음이 /l/, /m/과 같은 공명음인 경우, 둘째 음절의 모음이 탈락되면서 준말이 형성된다고 설명하였다. 그러나 준말의 일부 사례들을 분석한 결과이므로 '어제저녁 → 엊저녁'과 같이 다른 유형들에 비하여 이러한 사례들이 얼마나 자주 관찰되는지 알 수 없었다.

둘째, 모음의 탈락이나 활음화와 같은 수의적 음운현상을 중심으로 논의를 진행하였으므로 준말의 형성에 관련된 본딧말의 음운론적 조건을 구체적으로 논의할 수 없었으며, 준말과 본딧말 사이의 음운론적 차이가 무엇인지 논의하기 어려웠다. 예를 들어 '/꽂/→ [꼳]', '/같이/ → [가치]'과 같은 음운현상들을 보면 기저형의 유표적 구조, '음절말 파찰음'(/tɕʰ/)과 '치경폐쇄음+전설고모음'의 연쇄(/tʰ/+/i/)를 해소하기 위하여 음절말 평폐쇄음화와 치경폐쇄음의 구개음화가 일어나고, 그 결과 무표적인 구조가 형성되었다고 설명할 수 있다. 그러나 준말의 형성과정에서 적용되는 음운현상들은 수의적이며, 본딧말에 뚜렷한 유표적 구조가 나타나지 않는다. 따라서 준말의 형성에 관여하는 요인이 무엇인지, 음운론적으로 본딧말과 준말 사이의 결정적인 차이가 무엇인지 알 수 없다.

본 연구에서는 음절축약형 준말을 유형화하고 정보이론을 바탕으로 음절축약형 준말의 음운론적 특성과 형성 원인을 설명하고자 한다. 이후의 내용은 다음과 같이 구성되어 있다. 2장에서는 준말의 수집 방법을 소개하고 음절축약형 준말을 유형별로 분류하였다. 3장에서는 음절축약형 준말의 분석을 위하여 정보이론의 분석 지표들인 '정보량($plog$), 음운론적 복잡도, 상호정보량'을 소개하였다. 4장에서는 정보이론을 바탕으로 음절축약형 준말을 유형별로 분석하고, 그 결과를 통계적으로 검증하였다. 마지막 5장에서는 연구의 결과를 요약하여 정리하였다.

2. 자료의 수집과 준말 유형의 분류

음절축약형 준말의 분석을 위하여 본 연구에서는 '표준국어대사전'(국립국어원 1999)과 김희진(2003)으로부터 모음충돌 회피형 및 음절축약형 준말 1,453개를 수집하였다. 1,453개의 준말 가운데 음절축약형 준말은 717개이었다. 음절축약형 준말 717개 가운데 음절의 축약이 일어나는 부분이 동일한 어형들을 묶어서 계산한 결과 음절축약형 준말의 유형빈도는 332개이었다. 이들을 음절구조에 따라 다음과 같이 5가지 유형의 본딧말과 8가지 유형의 준말로 분류하였다.[1]

(2) 음절구조별 유형: 본딧말과 음절축약형 준말

본딧말	준말	유형빈도	사 례
CV.CVC	**CVC**	**156**	물고랑 /mul.**go.laŋ**/ → 물골 /mul.**gol**/
	CVCC	1	기스락 /gi.**si.lag**/ → 기슭 /gi.**silg**/
CV.CV	**CVC**	**130**	골짜기 /gol.**j'a.gi**/ → 골짝 /gol.**j'ag**/
	CV	**23**	진드기 /jin.**di.gi**/ → 진디 /jin.**di**/
CVC.(C)V	CV	10	꼬챙이 /g'o.**cæŋ.i**/ → 꼬치 /g'o.**ci**/
	CVCC	4	안하다 /**an.ha**.da/ → 않다 /**anh**.da/
CVC.(C)VC	CVC	7	나흗날 /na.**hid.nal**/ → 나흘 /na.**hil**/
CV.CVCC	CVC	1	매흙질 /**mæ.hilg**.jil/ → 맥질 /**mæg**.jil/

본 연구에서는 이들 유형 가운데 가장 많은 수를 차지하는 'CV.CVC → CVC'(156개), 'CV.CV → CVC'(130개), 'CV.CV → CV'(23개) 유형의 준말 309개를 중심으로 분석을 진행하였다. 이러한 유형들은 용언 어미(-다는 → -단), 보조사(-까짓 → -깟), 대명사(그거 → 거), 의존명사(때문 → 땜) 등에서 자주 관찰되었다.

음절구조를 기준으로 분류하면 'CV.CVC'와 'CV.CV'의 구조를 갖는 본딧말은 (3)과 (4)의 다섯 가지 유형으로 분류된다. (3)과 (4)에서 O와 C는 각각 음절의 초성(onset)과 종성(coda)을, 아래 첨자 1, 2는 음절의 위치를 의미한다. 예를 들어 O_1, V_1은 각각 첫 음절의 초성과 모음을 O_2, V_2, C_2는 각각 둘째 음절의 초성, 모음, 종성을 의미한다.[2] 예를 들어 (3)의 '$O_1V_1.O_2V_2C_2$(본딧말) → $O_1V_1O_2$(준말)'은 본딧말에 포함된 첫째 음절(O_1V_1)과 둘째 음절($O_2V_2C_2$) 가운데 첫째 음절의 초성(O_1)과 모음(V_1), 둘째 음절의 초성(O_2)이 결합되어 준말이 형성되었음을 의미한다. 여기서 준말의 표기 '$O_1V_1O_2$'가 준말의 음절구조가 아니라 본딧말의 음절구조를 반영하고 있다는 점을 주의해야 한다. 마지막 분절음(O_2)은 준말에서는 음절의 종성이지만 본딧말에서는 둘째 음절의 초성이므로 O로 표기하였다.

---

[1] 음절축약형 준말 717개와 축약유형별 332개의 목록은 'http://maincc.hufs.ac.kr/~hongsh/ssr_dataFile(final).xls'에 공개되어 있다.

[2] 음절축약형 준말은 '거머잡다 → 검잡다'의 경우처럼 본딧말을 기준으로 형태소 내부의 두 음절 사이에서 형성되는 경우가 많다. '어떠+하다 → 어떻다', '나+는 → 난'과 같이 형태소 경계에서 축약이 일어난다면 두 번째 형태소는 의존형태소인 경우가 많다. 3음절 단어에서는 '시치미→시침'처럼 보통 둘째 음절과 셋째 음절 사이에서 축약이 일어난다. 그러나 CV.CVC와 CV.CV의 음절구조를 가진 단어가 모두 준말을 형성하는 것은 아니다. '도미 → 돔'은 가능하지만 '가자미 → *가잠'은 불가능하다.

(3) 음절축약형 준말의 유형 Ⅰ (CV.CVC → CVC)

본딧말	준말		유형빈도	사 례
$O_1V_1.O_2V_2C_2$	(a)	$O_1V_1\underline{O_2}$	40	물고랑 /mul.**go.laŋ**/ → 물골 /mul.**gol**/ 수풀 /**su.pul**/ → 숲 /**sup**/
	(b)	$O_1V_1\underline{C_2}$	24	조금 /**jo.gim**/ → 좀 /**jom**/ 무엇을 /mu.**ə.siɬ**/ → 무얼 /mu.**əl**/
	(c)	$O_1\underline{V_2C_2}$	16	잘그랑 /jal.**gi.laŋ**/ → 잘강 /jal.**gaŋ**/ 그만두다 /**gi.man**.du.da/ → 간두다 /**gan**.du.da/
	(x)	$O_1V1\underline{O_2}$ $O_1V1\underline{C_2}$ (O_2=C_2)	23	제기랄 /je.**gi.lal**/ → 제길 /je.**gil**/ 나는 /**nanin**/ → 난 /**nan**/
	(y)	$O_1V_1\underline{C_2}$ $O_1\underline{V_2C_2}$ (V_1=V_2)	53	덜커덩 /dəl.**kə.dəŋ**/ → 덜컹 /dəl.**kəŋ**/ 삐꺼덕 /p'i.**kə.dəg**/ → 삐걱 /p'i.**kəg**/

(3)에서는 $O_1V_1.O_2V_2C_2$의 음절구조를 가진 본딧말을 준말의 음절구조에 따라 (a), (b), (c), (x), (y) 총 5가지 유형으로 분류하였다. (x)는 본딧말의 O_2와 C_2가, (y)는 본딧말의 V_1과 V_2가 동일하므로 준말도 각각 두 가지로 해석될 수 있는 유형들이다. 본 연구에서는 준말의 음절구조를 고려하여 둘째 음절의 초성과 종성이 같은 (x)는 준말 '$O_1V_1O_2$'의 유형인 (a)로, 첫째 음절의 모음과 두 번째 음절의 모음이 같은 (y)는 준말 '$O_1V_1C_2$'의 유형인 (b)로 분류하였다. 따라서 이후에는 (a)는 63개(40+23), (b)는 77개(24+53)로 확대하여 분석을 진행하겠다.

(4) 음절축약형 준말의 유형 Ⅱ (CV.CV → CVC, CV.CV → CV)

본딧말	준말		유형빈도	사 례
$O_1V_1.O_2V_2$	(d)	$O_1V_1O_2$	**130**	골짜기 /gol.j**'a.gi**/ → 골짝 /gol.**j'ag**/ 버물리다 /bə.**mu.li**.da/ → 버물다 /bə.**mul**.da/
	(e)	O_1V_2	**23**	진드기 /jin.dɨ.gi/ → 진디 /jin.di/ 그저께 /gɨ.jə.g'e/ → 그제 /gɨ.je/

본 연구에서는 (3)과 (4)에서 분류한 본딧말과 음절축약형 준말의 유형 309개

를 대상으로 정보이론의 계량적 분석방법론을 적용하여 두 가지 문제를 설명하고자 한다. 첫째, 음절축약형 준말의 형성 과정에서 대부분 본딧말의 CV.CVC, CV.CV 부분이 축약되는 까닭은 무엇인가? 둘째 본딧말의 CV.CVC나 CV.CV는 왜 $O_1V_1O_2$, $O_1V_1C_2$, $O_1V_2C_2$, O_1V_2와 같이 일정한 유형의 준말로 축약되는가? 두 가지 문제를 해결하기 위하여 다음의 가설을 중심으로 논의를 진행하겠다.

(5) 음절축약형 준말의 형성 원리
   a. 음절축약형 준말의 형성은 분절음 연쇄의 음운론적 적형성과 인접 분절음들 사이의 연결 빈도와 관련되어 있다.
   b. 분절음 연쇄의 음운론적 적형성 (본딧말＜준말)
   c. 인접 분절음들 사이의 연결 빈도 (본딧말＜준말)

(5)는 본딧말과 준말을 비교했을 때 준말은 출현빈도가 높은 분절음, 서로 자주 연결되는 분절음의 연쇄로 구성된 익숙한 유형으로 형성된다는 가정을 담고 있다. 여기서 '분절음 연쇄의 음운론적 적형성'과 '인접 분절음들 사이의 연결 빈도'는 각각 '음운론적 복잡도'와 '인접 분절음의 상호정보량'으로서 단어의 '출현빈도'나 '유형빈도'와는 상관이 없다.[3] (5)의 가설이 옳다면 준말을 구성하는 분절음들 사이의 연결 강도가 강하고 음운론적 적형성도 본딧말보다 높을 것이다. 분절음의 자질이나 음절구조만으로는 단어의 음운론적 적형성이나 분절음 사이의 연결 강도를 측정하기 어려우며, 분절음의 빈도와 확률을 바탕으로 계량적인 분석을 해야 한다. 다음 장에서는 이러한 계량적 분석의 기준이 되는 정보이론의 지표들을 살펴보겠다.

3. 정보이론

다섯 가지로 유형으로 분류되는 음절축약형 준말들이 형성되는 조건과 과정을 일관된 원리나 규칙으로 설명하기는 어렵다. 문제는 주로 CV.CVC와 CV.CV의 음절구조를 가진 본딧말들만 음절축약형 준말로 바뀌는 까닭이 무엇이며, 이론적으로 가능한 여러 가지 축약형 가운데 어떠한 유형이, 어떠한 원인에 의하여 실제의 준말로 선택되는가이다. 본 연구에서는 이러한 문제를 해결하기 위하여 음절축약형 준말이 분절음 연쇄의 음운론적 적형성 및 인접 분절음들 사이의 연결 빈도와 관련되어 있다고 가정하고 '정보이론'(Information Theory, Shannon 1948)의 몇 가

[3] '음운론적 복잡도'와 '상호정보량'은 3.2.와 3.3.에서 자세하게 소개될 것이다.

지 지표를 통하여 검증하였다. 규칙이나 제약 중심의 음운이론과 달리 정보이론에서는 적형성과 분절음의 연결 관계를 양적으로 측정할 수 있는 지표가 마련되어 있다.

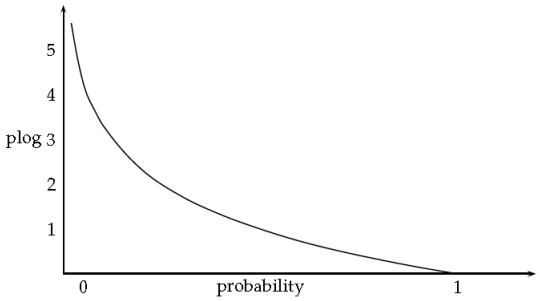
3.1 정보량

간단한 수학적 공식을 활용하면 의사소통 체계 안에서 어떠한 신호나 메시지를 전달하기 위하여 얼마나 많은 양의 정보가 필요한지 수학적 지표로 표시할 수 있다(Hume and Malihot 2013: 4). 정보이론에서는 어떠한 성분이 갖고 있는 정보량을 *plog*(positive logarithm: Goldsmith 2001, 2002, 2006 등)로 표기한다. *plog*는 해당 성분이 나타날 확률(*Prob*(x))을 밑수 2를 갖는 로그 함수로 계산한 결과이다.

(6) a. *plog* 공식

$$plog(x) = -\log_2 Prob(x)$$

b. 확률과 *plog*의 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

이분법과 2진수는 정보를 표시하는 가장 효율적인 방법이므로 정보이론에서는 밑수 2를 갖는 로그 함수를 이용하여 정보량을 측정하고 표시한다. (6a)의 공식에 의하면 만약 어떠한 분절음의 빈도가 높아서 해당 분절음이 나타날 확률이 상승한다면 *plog*는 낮아진다. 따라서 분절음이 나타날 확률과 *plog*는 반비례의 관계를 갖는다(6b). 예를 들어 분절음이 나타날 확률이 1/8이라면 *plog*는 3이지만, 분절음의 빈도가 높아서 출현 확률이 1/4, 1/2로 상승하면 *plog*는 각각 2와 1이 된다.

분절음의 확률은 출현빈도를 측정하여 계산되는데, 정보량을 구하는 방법으로는 유니그램(unigram) 모델과 바이그램(bigram) 모델을 적용할 수 있다. 유니그램

모델에서는 해당 분절음의 출현빈도를 측정하고 전체 분절음의 출현빈도를 합한 값으로 나누어 확률을 계산하며 인접 분절음은 고려하지 않는다. 예를 들어 '아이'(#ai#)와 '어머니'(#əməni#) 두 단어가 있다면 유니그램 모델에서는 아이의 /i/와 어머니의 /i/를 동일한 분절음으로 처리하므로 출현빈도는 2회이다. 반면 바이그램 모델에서는 분절음을 두 개씩 묶어서 처리하므로 '아이'와 '어머니'의 /i/는 앞 뒤의 분절음과 함께 묶어서 빈도를 측정한다. '아이'의 /i/는 /ai/와 /i#/에 포함되고, '어머니'의 /i/는 /ni/와 /i#/에 포함되므로 바이그램 /ai/, /ni/, /i#/의 출현빈도는 각각 1회, 1회, 2회이다. 인접 분절음과의 조합을 고려하므로 바이그램 모델에서는 유니그램 모델보다 분절음의 유형빈도와 출현빈도가 늘어난다.

어떠한 언어든 음소는 일정한 제약에 따라 배열되므로 앞뒤에 인접하는 분절음을 함께 고려하여 빈도를 측정하는 바이그램 모델을 적용하는 것이 바람직하다(Goldsmith 2002). 주의할 점은 바이그램의 확률과 정보량을 계산할 때 아래와 같은 조건부 확률로 구해야 한다는 점이다. '아이'(#ai#)라는 단어의 정보량을 측정하려면 /#/, /a/, /i/ 각각의 빈도를 측정하기보다는 인접한 분절음을 두 개씩 묶어서 바이그램 단위로 빈도를 측정하고 (6)의 공식에 따라 선행 분절음의 빈도와 바이그램의 빈도를 기준으로 측정되는 조건부 확률을 계산한다. 따라서 바이그램의 /#a/, /ai/, /i#/의 확률은 /#/에 /a/가 후행하는 빈도, /a/에 /i/가 후행하는 빈도, /i/ 뒤에 /#/이 인접하는 빈도를 측정해야 하고 각각 /#/, /a/, /i/의 빈도로 나누어 구한다.

(7) a의 뒤에 b가 나타날 조건부 확률

$$p(b|a) = \frac{p(ab)}{p(a)} = \frac{ab\text{가 나타날 확률}}{a\text{가 나타날 확률}} = \frac{ab\text{의 출현빈도}}{a\text{의 출현빈도}}$$

3.2 음운론적 복잡도

'음운론적 복잡도'(Phonological Complexity, PC)는 분절음의 정보량을 이용하여 단어나 어절의 음운론적 적형성이나 유표성을 측정하는 지표이다(Goldsmith 2002, 2006, 2011). 음운론적 복잡도는 $plog$에 의해 결정된다는 점에서는 정보량과 유사하지만, 평균을 구할 때 단어나 음절을 구성하는 분절음이나 바이그램의 개수로 나눈다는 점에서는 $plog$와 다르다. 단어의 '음운론적 복잡도'는 다음과 같은 공식으로 구한다. 간단히 말하자면 음운론적 복잡도는 '단어를 구성하는 모든 분절음의 $plog$를 합한 후에 분절음의 개수로 나누어 평균을 구한 결과'이다.[4]

---

4 '음운론적 복잡도'와 동일한 개념의 용어로서 '평균 정보량'(average $plog$)이나 '단어 엔트로피' 등도 사용된다.

(8) 음운론적 복잡도 (PC)

$$PC(w = x_1 x_2 ... x_n) = \frac{1}{n} \sum_{i=1}^{n=length} (-\log_2 P(x_i))$$

음운론적 복잡도는 단어의 음운론적 적형성을 반영한다. 해당 언어의 음소배열 유형을 잘 따르는 단어라면 음운론적 복잡도가 낮다. 반면 외국어의 음소배열 구조가 반영된 차용어는 고유어나 한자어보다 음운론적 복잡도가 높다(박선우·홍성훈·변군혁 2013). 앞서 살펴본 바이그램의 조건부 확률과 정보량을 구할 수 있다면 일련의 바이그램으로 구성된 단어의 음운론적 복잡도도 계산할 수 있다. 예를 들어 '아이'(#ai#)의 음운론적 복잡도는 다음과 같은 과정을 거쳐 조건부 확률의 평균으로 구한다.

(9) 음운론적 복잡도의 계산 과정 '아이'(#ai#)

a. 어형의 철자를 분절음 표기로 변환

아이 → #ai#

b. 분절음 표기를 바이그램 단위로 구분[5]

#ai# → # + #a + ai + i#

c. 바이그램의 정보량 합계

$plog$(#) + $plog$(#a) + $plog$(ai) + $plog$(i#)

= −{log_2P(#) + log_2P(#a) + log_2P(ai) + log_2P(i#)}

= −{log_2P(#) + log_2P(a | #) + log_2P(i | a) + log_2P(# | i)}

= −{log_2C(#)/n + log_2C(#a)/C(#) + log_2C(ai)/C(a) + log_2C(i#)/C(i)}

d. $plog$의 평균: 음운론적 복잡도

−1/4×{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)

분절음을 2개씩 묶어서 정보량을 구하는 바이그램 모델과 달리 유니그램 모델에서는 앞뒤의 분절음을 고려하지 않고 단일 분절음의 확률만으로 정보량을 측정

---

[5] 단어의 경계, 즉 어두와 어말의 경계는 음소의 배열에 영향을 미치는 요소이다. 따라서 바이그램 모델에서는 단어의 경계와 분절음을 함께 고려하고 선행 성분 *a*와 후행 성분 *b*로 구성된 바이그램 단위로 묶어서 조건부 확률에 따라 정보량을 계산한다. 단순히 *b*가 출현할 확률이 아니라 *a*뒤에 *b*가 출현할 확률을 바탕으로 바이그램 *ab*에 대한 정보량(-log_2C(*ab*)/C(*a*))을 계산한다. 다만 단어의 첫째 성분인 #은 선행 성분이 없으므로 #이 출현하는 별다른 조건이 없다. 따라서 조건부 확률이 아니라 #이 출현하는 확률만으로 정보량(-log_2P(#))을 구한다.

한다. 유니그램 모델과 바이그램 모델을 통하여 구한 '아이'(/ai/)의 음운론적 복잡도를 비교해 보면 앞서 지적하였듯이 음운현상의 적용 양상을 예측하기 위해서는 유니그램 모델보다는 바이그램 모델을 적용하는 것이 바람직하다. 21세기 세종계획 말뭉치(국립국어원 2009)를 바탕으로 '아이'라는 단어의 음운론적 복잡도를 구하면 유니그램 모델에서는 3.357, 바이그램 모델에서는 5.029이다. '아이'에는 음절 초성이 없으며, 모음의 충돌이 일어나는 유표적인 음절구조(V.V)가 있다는 점을 고려한다면 유니그램 모델로 구한 3.357은 '아이'의 음운론적 유표성이 반영된 복잡도로 보기 어렵다. 유니그램 모델에서는 다른 모음에 비하여 상대적으로 빈도가 높은 /a/와 /i/의 출현빈도만으로 정보량의 평균을 구하므로 단어의 유표성을 포착하기 어렵다. 반면 바이그램 모델로 구한 음운론적 복잡도(5.029)는 '하늘'(2.348), '다리'(2.968)와 같이 모음충돌이 일어나지 않는 단어들에 비하여 상당히 높다.

음운론적 복잡도는 단어를 구성하는 분절음의 빈도와 순서에 의해 결정된다. 익숙하고 흔한 유형일수록 낮으며, 낯설고 드문 유형일수록 높다. (10)은 바이그램 모델을 적용하여 영어 단어의 음운론적 복잡도를 계산한 결과 가운데 일부이다. 분석 결과에 의하면 영어에서 가장 무표적인 단어는 복잡도가 가장 낮은 *the* (1위, 1.93)와 *hand* (2위, 2.15)이다. 반면 *eh* (63,200위, 9.07), *Oahu* (63,201위, 9.21)는 복잡도가 매우 높은 유표적 단어들이다. 음운론적 복잡도와 유표성의 관계는 (11)과 같은 도식으로 이해할 수 있다.

(10) 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	plʌ́mĭŋ	3.71
12,642	Friday	fraýdĭ	3.71
25,281	tolls	tólz	4.01
25,282	recorder	rĭkórdɚ	4.01
37,922	overburdened	óvɚbɚ́dənd	4.32
37,923	Australians	ŏstreýlyənz	4.32
50,563	retire	rĭtaýr	4.75
50,564	poorer	púrɚ	4.75
63,200	eh	é	9.07
63,201	Oahu	óáhŭ	9.21

(11) 음운론적 복잡도(PC)와 유표성 (Hong 2006: 212)

적형성 낮음 (유표적)	적형성 높음 (무표적)
낮은 확률	높은 확률
←	→
높은 복잡도	낮은 복잡도

3.3 상호정보량

상호정보량(Mutual information, *MI*)은 두 가지 성분의 상호의존성을 나타내는 지표이다. 두 요소 사이에 상호의존성을 알려주는 척도로서, 상호정보량이 클수록 두 요소의 상호의존도가 높고, 작을수록 상호의존도가 낮다. 따라서 두 요소가 함께 나타나는 빈도가 높을수록 상호정보량이 높고, 반대로 함께 나타나는 빈도가 낮다면 두 요소 사이의 상호정보량은 낮아진다. 상호정보량은 실제로 두 요소가 함께 나타나는 '관측빈도'(observed frequency)와 함께 나타날 것이라고 예상되는 '예측빈도'(predicted frequency)의 비율로 계산된다(Goldsmith 2006, 2011). 예를 들어, 두 요소 x와 y 사이의 상호정보량 MI(x, y)는 x, y가 실제로 나타나는 확률 P(xy)와 함께 나올 것이라 예측되는 확률 P(x)⊥P(y)의 비율을 말하는데, 구체적으로는 관측빈도 P(xy)를 예측빈도 P(x)⊥P(y)로 나눈 후 로그로 변환하여 얻는다(홍성훈 2014).

(12) 상호정보량 (MI)

$$MI(x,y) = \log_2 \frac{prob(xy)}{prob(x) \times prob(y)}$$

$$= \log_2 prob(xy) - \log_2 prob(x) - \log_2 prob(y)$$

인접한 성분들 사이의 상호정보량이 높으면 두 성분이 함께 나타날 가능성이 높다는 것을 의미한다(Hume and Bromberg 2005). 예를 들어 한국어에서는 '-을, -를'과 같은 조사의 빈도가 높아서 모음 /i/(으)와 자음 /l/(ㄹ)이 함께 나타나는 빈도도 상당히 높다. 이와 같이 /il/의 빈도가 /i/와 /l/ 각각의 확률을 통하여 계산한 예측빈도보다 높으면 /i/와 /l/의 상호정보량은 양의 값을 갖는다.

앞서 살펴본 단어의 음운론적 복잡도와 단어를 구성하는 분절음들의 상호정보량은 대체로 반비례 관계를 갖는다. 음운론적 복잡도가 낮은 단어들, 즉 화자들에게 익숙한 음소배열 구조를 갖는 단어들의 경우, 단어를 구성하는 분절음들의 상

호정보량은 대체로 양의 값을 갖는다. 반면 음운론적 복잡도가 높은 차용어나 음소배열 구조가 특이한 음성상징어의 경우, 분절음들의 상호정보량은 음의 값을 갖는다. 두 분절음의 상호정보량이 음의 값을 갖는다는 것은 분절음들이 서로 인접하는 경우가 흔치 않다는 것을 의미한다(Goldsmith 2002:38). 다음 장에서는 지금까지 소개한 정보이론의 세 가지 지표인 정보량($plog$), 음운론적 복잡도, 상호정보량을 바탕으로 음절축약형 준말에 대한 분석을 진행하겠다.

4. 음절축약형 준말의 분석

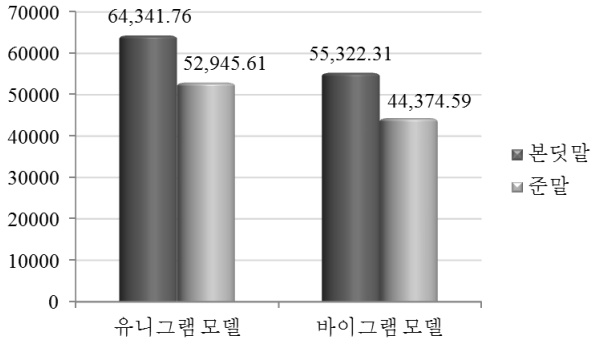
4.1 분절음의 출현빈도와 정보량

준말을 구성하는 분절음의 빈도와 정보량은 1999년부터 2004년 사이에 구축된 21세기 세종계획 형태분석 말뭉치를 바탕으로 두 가지 모델을 적용하여 측정하였다. 세종계획 말뭉치는 769개의 파일로 구성되어 있는데, 이들 가운데 북한어 자료와 역사자료 말뭉치는 제외하였다. 말뭉치를 구성하는 모든 어절의 출현빈도는 12,628,487개이었으며 모든 어절의 유형빈도는 1,936,705개이었다. 말뭉치로부터 어절을 추출하여 정렬하고 음소표기 부호로 변화하여 각 분절음들 정보량을 측정하였다. 정보량의 계산에는 1개 단위의 분절음을 기준으로 측정하는 유니그램(unigram) 모델과 분절음을 두 개씩 묶어서 측정하는 바이그램(bigram) 모델을 함께 적용하였다.

본딧말과 준말의 정보량을 측정하여 합계를 구한 결과, 바이그램 모델의 정보량이 유니그램 모델의 정보량보다 낮았다. Goldsmith (2002, 2006)에 의하면 정보이론에서는 확률이 높을수록, 정보량이 낮을수록 이론의 예측력이 높다고 판단하므로, 정보량이 낮아야 한다는 기준을 적용한다면 유니그램 모델보다는 바이그램 모델이 준말의 분석과 예측에 효과적이라고 볼 수 있다. 따라서 이후에 본 연구에서 언급되는 단어의 정보량은 모두 바이그램 모델로 계산되었다.

(13) 1,493개 본딧말-준말의 $plog$ 합계 (단위: bit)

$plog$ 합계	유니그램 모델	바이그램 모델
본딧말	64,341.76	55,322.31
준말	52,945.61	44,374.59

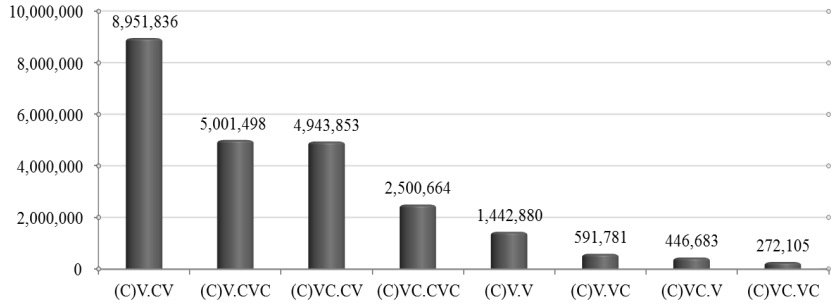


바이그램 모델에 의한 본격적인 분석에 앞서 음절축약형 준말을 형성하는 본딧말의 조건에 대하여 살펴보자. 세종코퍼스에 포함된 12,628,487개의 어절을 음절구조의 유형에 따라 분석한 결과 음절축약형 준말을 형성하는 본딧말은 대부분 (C)V.CV와 (C)V.CVC의 음절구조를 갖고 있었다. 일단 음절구조의 유형에 따라 출현빈도를 측정한 결과 아래의 도표에서 확인할 수 있듯이 CV.CV와 CV.CVC의 빈도가 가장 높았다.[6] 이러한 경향은 2장에서 살펴본 음절축약형 준말을 형성하는 본딧말의 유형빈도와 유사하였다. 음절축약형 준말을 형성하는 본딧말 332개 가운데 음절구조상 CV.CVC가 157개, CV.CV가 153개로서 93.37%를 차지하였다. 반면 CVC.CV의 빈도도 상당히 높은 편이었으나 이러한 음절구조를 가진 본딧말은 14개로서 4.22%에 불과하였다.

6 축약 유형별 분석에서 첫째 음절의 음절 초성은 준말의 형성에 결정적인 영향을 미치지 않으므로, 첫 음절의 음절초성이 있는 경우와 없는 경우를 구분하지 않고 합쳐서 빈도를 구하였다. 첫 음절의 초성까지 구분하여 분석한 2음절 유형들의 빈도는 'http://maincc.hufs.ac.kr/~hongsh/ssr_dataFile(final).xls'에서 확인할 수 있다.

(14) 음절구조 유형별 출현빈도 (내림차순)

음절구조 유형	출현빈도	비율 (%)
(C)V.CV	8,951,836	37.07
(C)V.CVC	5,001,498	20.71
(C)VC.CV	4,943,853	20.47
(C)VC.CVC	2,500,664	10.35
(C)V.V	1,442,880	5.97
(C)V.VC	59,1781	2.45
(C)VC.V	44,6683	1.85
(C)VC.VC	27,2105	1.13

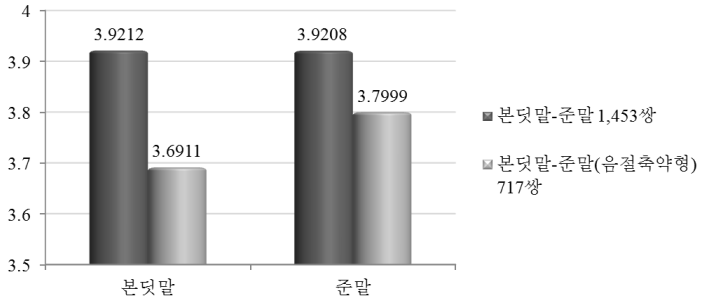


4.2 음운론적 복잡도

바이그램의 조건부 확률, 선행 분절음을 기준으로 후행 분절음이 나타날 확률을 바탕으로 음절축약형 준말과 본딧말의 음운론적 복잡도를 분석한 결과, 음운론적 복잡도는 준말의 형성과 직접 관련된 요인으로 볼 수 없었다. 본딧말과 준말 총 1,453쌍의 음운론적 복잡도를 측정하여 평균을 비교해 본 결과 차이가 크지 않았으며 통계적으로도 유의미한 수준이 아니었다. 음절축약형 준말과 본딧말 717쌍을 따로 분석한 결과에서도 본딧말의 음운론적 복잡도는 오히려 준말보다 낮았다(p<.05, t=-7.91).

(15) 본딧말과 준말의 음운론적 복잡도

음운론적 복잡도 (평균)	본딧말-준말 1,453쌍	본딧말-준말(음절축약형) 717쌍
본딧말	3.9212	3.6911
준말	3.9208	3.7999



이러한 결과는 준말의 유표적인 음절구조를 통해서도 예측할 수 있다. (2)에서 살펴본 '골짜기 → 골짝'과 같이 본딧말의 무표적인 2음절 구조(CV.CV)가 준말의 유표적인 1음절 구조(CVC)로 바뀌는 경우가 상당히 많은데, 이러한 음절구조의 변화는 오히려 준말의 음운론적 복잡도를 높일 수밖에 없다. 달리 말하자면 준말의 음소배열 순서나 음절 구조는 본딧말보다 유표적이었으며, 음운론적 복잡도는 준말의 형성을 결정하는 요인으로 볼 수 없었다.[7]

4.3 인접 분절음들 사이의 상호정보량

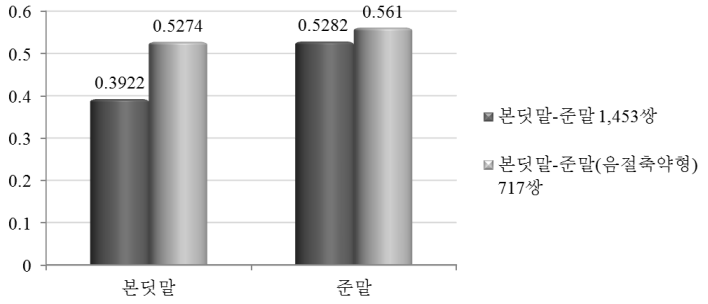
준말의 형성에 영향을 미치는 요인을 찾기 위하여 이번에는 본딧말과 준말을 구성하고 있는 인접 분절음 사이의 '상호정보량'을 비교 분석하였다. 음운론적 복잡도와 달리 단어 안에서 서로 인접하는 모든 분절음들 사이의 상호정보량을 분석한 결과 본딧말과 준말 사이에는 유의미한 차이가 있다는 사실을 확인하였다. 본딧말과 준말 1,453쌍의 상호정보량을 측정한 결과 평균의 차이는 통계적으로 유의미한 수준이었다($p<.005$, $t=-8.367$). 음절축약형 준말 717쌍만 따로 비교한 결과, 모든 표본을 비교한 결과보다는 준말과 본딧말의 차이가 크지 않았으나 부분적으

7 논문의 심사과정에서 익명의 심사위원은 단어나 어절을 구성하는 음절 개수의 변화, 즉 준말을 형성하면서 음절이 줄어드는 현상이 단어의 음운론적 유표성에 영향을 줄 수 있다고 제안하였다. 그러나 정보이론에서 제안된 '음운론적 복잡도'는 바이그램 정보량의 평균을 기준으로 측정되므로 음절 개수의 영향을 반영하지 못한다. 따라서 음절의 개수나 단어의 길이와 관련된 음운론적 유표성을 반영하지 못하는 한계를 갖고 있다.

로 유의미한 차이를 확인할 수 있었다(p<.1, t=-1.792, 단측검정). 분석 결과에 의하면 준말에 배열된 인접 분절음들의 상호정보량은 본딧말의 상호정보량보다 높았다. 따라서 분절음들이 연결되어 함께 출현하는 빈도인 '연어 관계'(collocational strength)는 본딧말보다 준말이 높았다.

(16) 본딧말과 준말의 상호정보량

상호정보량 (평균)	본딧말-준말 1,453쌍	본딧말-준말(음절축약형) 717쌍
본딧말	0.3922	0.5274
준말	0.5282	0.5610



한편, 준말의 형성과정에서 축약이 일어나는 두 음절의 분절음들에 대해서만 상호정보량을 분석한 결과 본딧말과 준말의 차이는 유의미한 수준이 아니었다. 본딧말에서 축약이 일어나는 CV.CVC나 CV.CV의 평균 상호정보량은 0.845이었으며, 축약된 준말의 CVC나 CV의 평균 상호정보량은 0.807이었다. 단어 전체의 상호정보량과 반대로 본딧말의 상호정보량이 준말보다 높았으나 통계적으로는 유의미하지 않았다. 요컨대 준말이 형성되면서 단어를 구성하는 모든 인접 분절음들 사이의 상호정보량은 증가하지만, 축약이 일어나는 음절의 분절음에서는 상호정보량의 유의미한 증가를 확인할 수 없었다.

(17) 본딧말과 준말의 상호정보량 비교
a. 단어를 구성하는 모든 분절음: 본딧말＜준말
a. 축약 부분의 음절을 구성하는 분절음: 본딧말 ≒ 준말

준말의 형성 과정에서 전체적으로 분절음들 사이의 상호정보량이 증가하는 원인은 무엇일까? 상호정보량의 증가에 영향을 주는 세부적 원인을 파악하기 위하

여, 본 연구에서는 준말의 형성 과정에서 분절음의 탈락이 일어나는 경계에 인접한 두 분절음의 상호정보량을 측정하였다. 그 결과 (3)-(4)에서 분류한 다섯 가지 유형 가운데 세 가지 유형에서 상호정보량의 증가를 확인할 수 있었으며 그 차이는 통계적으로 유의미한 수준으로 검증되었다.

(18) 분절음 탈락이 일어나는 경계의 상호정보량 (본딧말 → 준말)

음절축약형 준말 유형		상호정보량 비교
a.	$\underline{O_1V_1}.O_2V_2C_2 \rightarrow \underline{O_1V_1}O_2$	$\underline{V_1}.O_2$ (0.98) < $\underline{V_1}O_2$ (1.28) ** (t=-1.93, df=62)
b.	$\underline{O_1V_1}.O2V2C_2 \rightarrow \underline{O_1V_1}C_2$	$\underline{V_1}.O_2$ (0.42) < $\underline{V_1}C_2$ (1.45) ** (t=-4.87, df=76)
c.	$\underline{O_1}V1.O2V_2C_2 \rightarrow \underline{O_1}V_2C_2$	$\underline{O_1}V_1$ (1.99) > $\underline{O_1}V_2$ (1.16) ** (t=2.77, df=15)
d.	$\underline{O_1V_1}.O_2V2 \rightarrow \underline{O_1V_1}O_2$	$\underline{V_1}.O_2$ (0.90) < $\underline{V_1}O_2$ (1.39) ** (t=-3.59, df=129)
e.	$\underline{O_1}V_1.O_2V_2 \rightarrow O_1\underline{V_2}$	$\underline{O_1}V_1$ (0.89) < $\underline{O_1}V_2$ (1.32) ** (p=.29, t=.55, df=22)

(*p < .05, **p < .005; 단측 검정)

(18)은 준말의 형성과정에서 분절음의 탈락 이전과 이후의 상호정보량 변화를 통계적으로 검증한 결과이다. (18b), (18c)는 분절음의 탈락으로 인하여 인접 분절음과 음절구조에 변동이 일어난다. 예를 들어 (18b)의 유형인 본딧말 '조금'(jo.gim)과 준말 '좀'(jom)을 비교해 보면 준말의 형성과정에서 둘째 음절의 초성 /g/와 중성 /ɨ/가 탈락된다. 탈락의 결과로 첫째 음절의 중성 /o/는 본딧말에서는 둘째 음절의 초성 /g/와 연결되었으나 준말에서는 둘째 음절의 종성 /m/과 연결되므로 본딧말 /-ɨ.g-/와 준말의 /-im/의 상호정보량을 비교하였다. (18b)는 탈락이 일어나는 분절음의 경계를 기준으로 '조금-좀'과 동일한 유형의 음절축약형 준말들을 모두 비교하여 분석한 결과이다. (18a)와 (18d)와 같은 유형(고랑-골, 골짜기-골짝)에서는 분절음의 탈락 이후에도 본딧말과 준말 사이에 연결되는 인접 분절음의 변화가 없으나 준말의 형성과정에서 'V.C'의 음절 경계가 사라지면서 'VC'의 상호정보량이 증가한다.

분절음 탈락으로 인접 분절음이나 음절구조가 변경된 경우, 변화가 일어난 두 분절음의 상호정보량을 분석한 결과 (18a), (18b), (18d)에서는 본딧말보다 준말에서 상호정보량이 높았다. 따라서 준말이 형성됨으로써 변경된 음절구조나 인접 분

절음으로 인하여 '연어 관계'(collocational strength)가 강화되었다고 볼 수 있다.

(19) 분절음 탈락에 의한 변화와 상호정보량의 증가
a. 인접 분절음, 음절구조 변화: (18b)
조금 /j**o.g**im/ → 좀 /j**om**/, 잘그랑 /jal.**gi**.laŋ/ → 잘강 /ja**l.g**aŋ/
b. 음절구조 변화: (18a, 18d)
고랑 /g**o.l**aŋ/ → 골 /g**ol**/, 골짜기 /gol.j'**a.g**i/ → -짝 /gol.j'**ag**/

그러나 (18c)에서는 준말의 상호정보량이 오히려 줄어들었고, (18e)에서는 상호정보량의 증감이 통계적으로 확인되지 않았다. 따라서 (18c)와 (18e)는 상호정보량의 변화로도 설명이 되지 않는 준말이다. 이러한 유형의 준말에 대해서는 4.5에서 탈락 분절음의 정보량을 중심으로 다시 논의하겠다.

4.4 탈락 분절음의 정보량

대부분의 준말에서 이론적으로 가능한 여러 가지 축약형들 가운데 왜 하나의 축약형만 준말로 선택되는 것일까? 이 문제에 대한 답은 본딧말과 준말 사이의 정보량에서 찾을 수 있다. 음절의 수를 줄여서 준말을 형성하더라도 본딧말에 담긴 정보를 온전히 전달하려 한다면 본딧말과 준말 사이 정보량의 차이를 최소화하는 것이 바람직하다. 준말과 본딧말 사이 정보량의 차이를 최소화하려면 본딧말에서 탈락되는 분절음들의 정보량 수치가 낮아야 한다. 따라서 정보량이 가장 적은 분절음들이 탈락을 겪을 가능성이 높다.

(20) 음절축약형 준말의 선택

$O_1V_1.O_2V_2C_2$ → $O_1V_1O_2$ 고랑 → 골, *공, *강
$O_1V_1.O_2V_2C_2$ → $O_1V_1C_2$ 조금 → 좀, *족, *즘
$O_1V_1.O_2V_2C_2$ → $O_1V_2C_2$ (웽)그렁 → -겅, *-글, *-긍
$O_1V_1.O_2V_2$ → $O_1V_1O_2$ 도미 → 돔, *디
$O_1V_1.O_2V_2$ → O_1V_2 (진)드기 → -디, *-득

음운현상에 의한 정보량의 차이를 최소화하는 경향은 무표 모음의 삽입과 탈락에서도 확인할 수 있다. 어떠한 언어든 출현빈도가 가장 높은 분절음이 삽입되거나 탈락되는 경향이 나타나는데 출현빈도가 가장 높은 분절음은 정보량이 가장 적으므로 삽입되거나 탈락되어도 단어의 정보량에 큰 영향을 주지 않는다(Hume

et al. 2011). 이러한 가설에 의하면 탈락되는 분절음들의 *plog*가 가장 낮을 것이라고 예측할 수 있다.

(21) 음절구조 유형과 탈락 분절음

본딧말	준말	사례	탈락 분절음
$O_1V_1.O_2$(**V_2C_2**)	$O_1V_1O_2$	고랑 /go.laŋ/ → 골 /gol/	aŋ (**V_2C_2**)
$O_1V_1.$(**O_2V_2**)C_2	$O_1V_1C_2$	조금 /jo.gɨm/ → 좀 /jom/	gɨ (**O_2V_2**)
O_1(**$V_1.O_2$**)V_2C_2	$O_1V_2C_2$	(웽)그렁 (weŋ)/gɨ.ləŋ/ → -경 /gəŋ/	ɨ.l (**$V_1.O_2$**)
$O_1V_1.O_2$(**V_2**)	$O_1V_1O_2$	도미 /domi/ → 돔 /dom/	i (**V_2**)
O_1($V_1.O_2$)V_2	O_1V_2	(진)드기 (jin)/di.gi/ → -디 /di/	i.g (**$V_1.O_2$**)

정보량이 가장 적은 분절음들이 탈락된다는 가설을 검증하기 위하여 본딧말의 음절구조를 기준으로 탈락이 가능한 위치에 있는 두 분절음의 *plog*를 측정하고 그 차이를 통계적으로 검증해 보았다. 그 결과 상호정보량에 의한 분석과 마찬가지로 (18a), (18b), (18d) 유형에서는 정보량이 적은 분절음들이 탈락된다는 가설에 부합되는 결과를 얻었으나 (18c)와 (18e) 유형에서는 가설과 다른 결과가 나왔다.

(22) 탈락 가능한 분절음의 정보량: (18a)~(18c) 유형

음절축약형 준말 유형		탈락 분절음			통계적 분석 결과
		V_2C_2	O_2V_2	V_1O_2	
a.	$O_1V_1.O_2$**$\underline{V_2C_2}$** → $O_1V_1O_2$	**6.73**	7.37	7.78	$V_2C_2 < O_2V_2$** $V_2C_2 < V_1O_2$**
b.	$O_1V_1.$**$\underline{O_2V_2}$**C_2 → $O_1V_1C_2$	**7.76**	**7.89**	8.51	$V_2C_2 < V_1O_2$** $O_2V_2 < V_1O_2$**
c.	O_1**$\underline{V_1}.\underline{O_2}$**$V_2C_2$ → $O_1V_2C_2$	7.20	7.79	7.20	??

(*p < .05, **p < .005; 단측 검정)

본딧말의 음절구조가 CV.CVC인 (18a-c)를 분석한 결과 (18a)에서는 탈락되는 V_2C_2의 정보량이 탈락이 가능한 다른 분절음의 정보량보다 적었으며, t-검정(t-test)을 시행한 결과 이러한 차이는 통계적으로도 유의미한 것으로 검증되었다. (18b)에서도 탈락되는 분절음 O_2V_2의 정보량이 가장 적었다. 통계적 분석 결과 O_2V_2의

정보량(7.89)은 V_1O_2보다는 적었으나 또 다른 탈락 후보인 V_2C_2와는 유의미한 차이를 보이지 않았다. 반면 (18c)에서는 탈락되는 분절음 V_1O_2의 정보량과 다른 탈락 후보들의 정보량 사이 통계적으로 유의미한 수준의 차이를 확인할 수 없었다.

(23) 탈락 가능한 분절음의 정보량: (18d)~(18e) 유형

음절축약형 준말 유형		탈락 분절음		통계적 분석 결과
		V_2	V_1O_2	
d.	$O_1V_1.O_2\underline{V_2} \rightarrow O_1V_1O_2$	**4.30**	8.10	$V_2 < V_1O_2$**
e.	$O_1\underline{V_1.O_2}V_2 \rightarrow O_1V_2$	**4.40**	7.27	$V_2 < V_1O_2$**

(*$p < .05$, **$p < .005$; 단측 검정)

본딧말의 음절구조가 CV.CV인 (18d-e)를 분석한 결과 (18d)에서는 탈락되는 V_2의 정보량이 탈락되지 않는 V_1O_2보다 적었으며, 통계적으로 유의미한 수준의 차이가 확인되었다. 반면 (18e)에서는 탈락되는 분절음 V_1O_2의 정보량이 오히려 V_2보다 많은 편이었다.[8]

탈락되는 분절음을 기준으로 분절음의 정보량과 인접 분절음과의 상보정보량을 측정한 결과 (18a), (18b), (18d)에서는 정보량이 가장 낮거나 비교적 낮은 위치의 분절음들이 탈락되면서 인접 분절음들 사이의 상호정보량은 증가한다는 결론을 얻을 수 있었다. 반면 (18c)와 (18e) 유형에서는 정보량이 낮은 분절음이 탈락된다고 볼 수 없고, 준말에서 인접 분절음 사이의 상호정보량이 증가하는 경향도 찾을 수 없었다. 요약하자면 둘째 음절의 모음이 탈락되는 (18a), (18b), (18d)에서는 탈락 분절음의 정보량을 최소화하면서 분절음들 사이의 상호정보량을 늘리는 방향으로 준말이 형성되지만 첫째 음절의 모음이 탈락되는 (18c), (18e)에서는 이러한 경향을 찾을 수 없었다. 다음 절에서는 지금까지의 논의와는 다른 전제를 바탕으로 (18c)의 유형들에 대하여 논의하겠다.

4.5 모음 삽입과 본딧말

준말의 형성 과정에서 첫째 음절의 모음이 탈락되는 (18c)와 (18e)는 앞서 살

[8] V_2에 대해서는 바이그램의 $plog$를 구할 수 없으므로 유니그램 모델로 $plog$를 구하여 V_1O_2의 바이그램 정보량과 비교하였다. V_1과 O_2의 유니그램 $plog$를 각각 구하여 합한 뒤에 평균을 구하면 V_1O_2의 정보량은 (d)에서는 4.40, (e)에서는 4.20이었다. 이 값은 유니그램 V_2의 정보량과 통계적으로 유의미한 차이를 보이지 않았다(23d: p=0.84, 23e: p=0.17).

펴본 정보이론의 지표만으로는 설명할 수 없었다. 이들 가운데 (18c)의 사례들은 음운론적으로 동질적인 성격을 가지고 있다.

(24) $O_1\underline{\mathbf{V_1.O_2}}V_2C_2 \rightarrow O_1V_2C_2$ 본딧말-준말 목록 (18c 유형)

a. 그랑-강: 갈그랑거리다-갈강거리다, 왱그랑댕그랑-왱강댕강, 잘그랑거리다-잘강거리다
   그런-건: 그런데-건데
   그렁-겅: 걸그렁거리다-걸겅거리다, 글그렁거리다-글겅거리다, 웽그렁뎅그렁-웽겅뎅겅
   그만-간: 그만두다-간두다
b. 드덕-덕: 푸드덕거리다-푸덕거리다
   드럭-덕: 무드럭지다-무덕지다
   들입(드립)-딥: 들입다-딥다
   뜨락-딱: 까뜨락거리다-까딱거리다
c. 르랑-랑: 가르랑-가랑
   르렁-렁: 거르렁거리다-거렁거리다, 그르렁거리다-그렁거리다
d. 지락-작: 깨지락거리다-깨작거리다
   지럭-적: 께지럭거리다-께적거리다
   지작-작: 만지작거리다-만작거리다
   찌했-쨌: 어찌했든-어쨌든
   치작-착: 배치작거리다-배착거리다
   치적-척: 비치적거리다-비척거리다

위의 목록을 살펴보면 (18c) 유형의 본딧말에는 '연구개음+/ɨ/'(24a), '치경음+/ɨ/'(24b-c), '경구개음+/i/'(24d)로 구성되는 바이그램이 포함되어 있다. 알다시피 영어 차용어가 한국어의 음절구조에 수용되는 경우 무표모음이 삽입되는데, 그 양상은 (24)의 본딧말과 비슷하다. 일반적으로 한국어의 모음체계에서 가장 무표적인 /ɨ/가 삽입되지만 경구개 자음 뒤에서는 모음 /i/가 삽입된다. 모음삽입의 결과로 (24)와 마찬가지로 '비경구개음+/ɨ/', '경구개음+/i/'의 바이그램이 형성된다.

(25) 영어 차용어의 모음 삽입: 선행자음과 삽입모음 (박선우 1998:82-83)

a. 비경구개음+/ɨ/

s：	logos [lougɑs]	로고스 [rogosɨ]
	circus [sɚ:kəs]	서커스 [səkʰəsɨ]

z：	jazz [dʒæz]	재즈 [tɕɛʥi]
	lens [lenz]	렌즈 [ɾenʥi]
f：	staff [stæf/stɑ:f]	스태프 [sɨtʰepʰɨ]
v：	stove [stouv]	스토브 [sɨtʰobɨ]
	drive [draiv]	드라이브 [dɨɾaibɨ]
	curve [kə:v]	커브 [kʰəbɨ]
ʦ/ʣ：	sports [spɔ:ʦ]	스포츠 [sɨpʰotɕʰɨ]

a. 경개구음+/i/

ʧ/ʃ：	scratch [skræʧ]	스크래치 [sɨkʰɨɾɛtɕʰi]
	speech [spi:ʧ]	스피치 [sɨpʰtɕʰi]
	rush [rʌʃ]	러시 [ɾəʃi]
ʤ/ʒ：	beige [beiʒ]	베이지 [peiʤi]
	lounge [launʤ]	라운지 [ɾaunʤi]

본 연구에서는 본딧말이 음절의 축약을 겪으면서 준말로 형성된다고 가정하였으나, (24)의 본딧말과 모음삽입의 결과인 (25) 차용어들을 비교해 보면 거꾸로 모음의 삽입을 통하여 지금까지 준말로 가정해 왔던 1음절 형태(CVC)가 2음절(CV.CV)의 본딧말로 확장되었을 가능성이 있다.[9] 본딧말로부터 형성된 준말이 아니라고 가정한다면 본 연구에서 논의했던 음절축약형 준말의 유형 309개 가운데 (18c) 유형이 차지하는 비율이 7.44%(23개)에 불과하다는 점도 설명할 수 있다.

차용어와의 비교 외에도 두 가지 측면에서 이러한 가설의 가능성을 검토할 수 있다. 첫째 (24)에 제시된 (18c) 유형의 본딧말과 준말들이 대부분 음성상징어라는 점이다. (24d)의 '어찌했든-어쨌든'을 제외하면 (18c) 16가지 유형에 포함되는 단어들은 모두 음성상징어이다. 음성상징어는 운율적 구조와 제약에 민감하며 완전 중첩(빵 → 빵빵)과 부분 중첩(빵 → 빠방)을 겪는 경우가 많다. 따라서 본딧말로부터 준말이 형성된 것이 아니라 준말이 운율적 원인에 의해 본딧말의 형태로 확장되었다고 볼 수 있다.

모음삽입의 가설에서 검토해야 할 두 번째 사항은 4.3.에서 살펴보았던 상호정보량의 분석 결과가 (18c)에서는 반대로 나타난다는 점이다. 아래의 표는 (18)에서 제시했던 분절음 탈락 전후 상호정보량의 변화를 간략하게 정리한 것이다.

---

9 지금까지 가정해 왔던 '준말 → 본딧말'의 과정과 반대로 '준말 → 본딧말'의 과정을 가정한다면 '본딧말'(줄지 않은 본디 음절의 말)과 '준말'(본딧말의 일부분이 줄어든 말)이라는 용어는 부적절하다. 하지만 본 연구에서는 논의의 편의상 (18c) 유형의 1음절 형태를 준말로, 2음절 형태를 본딧말로 지칭하겠다.

(26) 분절음 탈락이 일어나는 경계의 상호정보량 (본딧말 → 준말)

음절축약형 준말 유형		상호정보량의 변화
a.	$\underline{\mathbf{O_1V_1}}.O_2V2C_2 \rightarrow \underline{\mathbf{O_1V_1}}O_2$	본딧말＜준말
b.	$\underline{\mathbf{O_1V_1}}.O_2V_2C_2 \rightarrow \underline{\mathbf{O_1V_1}}C_2$	본딧말＜준말
c.	$\underline{\mathbf{O_1}}V_1.O_2V_2C_2 \rightarrow \underline{\mathbf{O_1}}V_2C_2$	**본딧말＞준말**
c′.	$\underline{\mathbf{O_1V_1}}C_1 \rightarrow \underline{\mathbf{O_1}}V_i.O_i\underline{\mathbf{V_1}}C_1$	**준말＜본딧말**
d.	$\underline{\mathbf{O_1V_1}}.O_2V_2 \rightarrow \underline{\mathbf{O_1V_1}}O_2$	본딧말＜준말
e.	$\underline{\mathbf{O_1}}V_1.O_2V_2 \rightarrow O_1\underline{\mathbf{V_2}}$	본딧말＜준말

상호정보량이 증가하는 다른 준말들과 달리 유독 (26c)만이 분절음의 탈락 결과 상호정보량이 감소하는 변화를 겪는다. 그러나 다른 유형들과 반대로 (18c)는 ‘준말 → 본딧말’의 과정을 가정한다면 (26c′)과 같이 상호정보량이 증가하는 일반적인 경향을 벗어나지 않는다. 요약하자면 축약이 일어나는 음절구조의 유형빈도, 삽입모음과 유사한 분절음의 연결 유형, 단어의 성격, 상호정보량의 변화 등을 고려한다면 (18c)는 본딧말로부터 형성된 준말이 아니라 준말로부터 확장된 본딧말일 가능성이 높다.

5. 결론 및 과제

정보이론을 바탕으로 음절축약형 준말의 여섯 가지 유형을 분석한 결과, 다음과 같은 몇 가지 결론을 얻을 수 있었다. 첫째, 준말과 본딧말의 음운론적 복잡도 사이에는 통계적으로 유의미한 차이가 나타나지 않았다. 준말의 음운론적 적형성은 본딧말보다 높지 않았으며 음절축약형 준말의 형성은 분절음의 음운론적 적형성이나 유표성으로 예측할 수 없었다.

둘째, 음절축약형 준말을 구성하는 인접 분절음들의 상호정보량을 측정한 결과 준말의 형성 과정에서 관찰되는 상호정보량의 증가는 통계적으로 유의미한 수준이었다. 셋째, 분절음이 탈락되는 경계에 인접한 두 분절음의 상호정보량을 분석한 결과 음절축약형 준말의 다섯 가지 유형 가운데 세 가지 유형에서 음절 경계의 상호정보량이 증가하였다. 따라서 음절축약형 준말의 형성은 음절의 수를 줄이면서 상호정보량이 높은 두 분절음을 인접시켜 연결하는 과정으로 이해할 수 있다. (18c)의 유형만은 준말이 형성되면서 상호정보량이 감소하였다. 그러나 (18c)의 유형이 본딧말로부터 형성된 준말이 아니라 거꾸로 준말로부터 확장된 본딧말

의 형태라고 가정한다면 (18c) 유형 역시 상호정보량이 감소된다고 볼 수 있다.

넷째, 음절구조상 정보량이 가장 낮은 분절음들이 탈락되어 본딧말과 준말 사이의 정보량의 차이를 최소화한다는 가설을 적용하면 이론적으로 가능한 다양한 축약형들 가운데 실제의 준말을 예측할 수 있다. 분석 결과 다섯 가지 준말의 유형 가운데 세 가지 유형은 이러한 가설을 따르는 것으로 검증되었다. 셋째와 넷째 논의를 유형별로 정리한 결과는 다음과 같다.

(27) 음절축약형 준말의 유형과 형성 원인 (상호정보량, 탈락분절음의 정보량)

음절축약형 준말 유형		경계 분절음의 상호정보량 증가	탈락 분절음 = 최소정보량
a.	$O_1V_1.O_2\underline{\mathbf{V_2C_2}} \rightarrow O_1V_1O_2$	○	○
b.	$O_1V_1.\underline{\mathbf{O_2V_2}}C_2 \rightarrow O_1V_1C_2$	○	△
c.	$O_1\underline{\mathbf{V_1.O_2}}V_2C_2 \rightarrow O_1V_2C_2$	×	?
c′.	$\underline{\mathbf{O_1V_1}}C_1 \rightarrow \underline{\mathbf{O_1}}V_i.O_i\underline{\mathbf{V_1}}C_1$	○	?
d.	$O_1V_1.O_2\underline{\mathbf{V_2}} \rightarrow O_1V_1O_2$	○	○
e.	$O_1\underline{\mathbf{V_1.O_2}}V_2 \rightarrow O_1V_2$	?	×

※ ○: 가설을 뒷받침 (통계적으로 유의미한 결과를 보임)
△: 가설을 부분적으로 뒷받침
?: 통계적 분석 결과 불분명 (통계적으로 유의미한 결과가 나오지 않음)

마지막으로 (18e)는 탈락분절음의 정보량이나 4.3-4에서 살펴본 *plog*, 상호정보량, *CS-plog*만으로는 설명하기 어려운 유형이다.[10] 이 유형은 본딧말과 준말이 모두 무표적인 음절구조(CV)를 가지고 있다. '진드기-진디'와 같은 예를 보면 정보량과 상관없이 음절의 개수와 단어의 길이와 관련된 유형으로 보인다. '그저께-그제'에서는 인지적으로 단어나 형태의 첫 분절음과 마지막 분절음을 유지하는 '가장자리 효과'(edge effect, Aitchison 2012)도 나타나는 것 같다. 현재까지는 본 연구에서 제안한 정보이론의 지표로서는 (18e)의 유형을 설명하기 어려우며, 앞으로 분절음뿐만 아니라 음절의 구조와 개수, 단어의 가장자리 효과 등도 함께 고려해야 설명이 가능할 것으로 보인다.

---

[10] (18e) 유형에 포함된 본딧말과 준말은 23개이며 그 목록은 다음과 같다.
구가-가, 그거-거, 그러-거, 꺼기-끼, 끄러-꺼, 다래-대, 드기-디, 드라-다, 드러-더, 드리-디, 들이-디, 떠해-때, 라기-리, 마무-무, 머니-미, 머무-무, 무나-마, 사리-시, 저께-제, 적에-제, 츠리-치, 커니-키, 크러-커

참고문헌

국립국어원. 1999.『표준국어대사전』. 두산동아.

김선철. 2011. 통신언어 준말의 형성에 대한 음운론·형태론적 고찰.『언어학』61: 115-129.

김성규. 1999. 빠른 발화에서 음절수 줄이기.『애산학보』23: 109-137.

김희진. 2003.『현대 국어의 준말 목록』. 국립국어연구원.

박선우. 1998.『현대국어 영어차용어의 음운론적 연구: 최적성 이론을 배경으로』. 고려대학교 석사학위논문.

박선우, 홍성훈 & 변군혁. 2013. 한국어의 어휘계층과 음운론적 복잡성.『음성·음운·형태론 연구』19(2): 255-274.

송철의. 1993. 준말에 대한 형태·음운론적 고찰.『동양학』23: 25-49.

이지양. 1998.『국어의 융합 현상』. 태학사.

이지양. 2003. 국어 준말의 성격.『성심어문논집』25: 285-316.

정희창. 2004.『국어 준말의 연구』. 성균관대학교 박사학위논문.

홍성훈. 2014.『음운론의 계량적 방법론』. 한국문화사.

Aitchison, Jean. 2012. *Words in the mind: An introduction to the mental lexicon.* 4th Edition. John Wiley & Sons.

Goldsmith, John. 2001. On information theory, entropy and phonology in the 20th century. *Folia Linguistica* XXXIV(1-2): 85-100.

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

Goldsmith, John. 2006. Probability for linguists. *Mathmatiques et Sciences Humaines* 180(4): 73-98.

Goldsmith, John. 2011. Information theory for linguists: A tutorial introduction. Paper presented at *the LSA Summer Institute: Workshop on information theory in linguistics*. University of Colorado at Boulder.

Hong, Sung-Hoon. 2006. Quantitative analysis of English hypocoristics: Wellformedness and phonological complexity. *Studies in Phonetics, Phonology and Morphology* 12(1): 211-229.

Hume, Elizabeth and Frédéric Malihot. 2013. The role of entropy and surprisal in phonologization and language change. In Alan Yu (ed.), *Origins of sound patterns: Approaches to phonologization*. Oxford University Press.

Hume, Elizabeth and Ilana Bromberg. 2005. Predicting epenthesis: An information-theoretic account. Paper presented at t*he 7th Annual Meeting of the French Network of Phonology*. Aix-en-Provence, France.

Hume, Elizabeth, Kathleen Currie Hall, Andrew Wedel, Adam Ussishkin, Martine Adda-Decker, and Cdric Gendrot. 2011. Anti-markedness patterns in French Epenthesis: An information-theoretic approach. In Chundra Cathcart, I-Hsuan Chen, Greg Finley,

Shinae Kang, Clare S. Sandy, and Elise Stickles (eds.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics.* Berkeley, CA.

Shannon, Claude. 1948. A mathematical theory of communication. Reprinted with corrections from *The Bell System Technical Journal* 27: 379-433/623-656.

홍성훈
(02450) 서울특별시 동대문구 이문로 107
한국외국어대학교 영어학과
E-mail: hongshoon@hufs.ac.kr

박선우
(42601) 대구광역시 달서구 달구벌대로 1095
계명대학교 국어교육과
E-mail: sunwoopark@kmu.ac.kr

접수일자: 2015. 07. 10.
수정일자: 2015. 09. 10.
게재일자: 2015. 09. 10.