

한국어 모음조화의 약화 현상에 대한 정보이론 기반 분석*

박선우
(계명대학교)

Park, Sunwoo. 2016. The diachronic declination of Korean vowel harmony from the perspective of Information Theory. *Studies in Phonetics, Phonology and Morphology* 22.2. 453-475. This study investigates the declination of vowel harmony in Korean from the 15th century to the 18th century. Between the 15th and the 18th century, restructuring of the vowel system resulting from vowel shift brought about diachronic declination of palatal vowel harmony between front and back vowels. Since the 15th century, vowel harmony in Korean has been in steady decline. Noting this, the present paper examines the frequencies of vowel harmony from the 15th to the 18th century. To do so, it analyzes a historical Korean corpus from the perspective of Information Theory, measuring the quantitative analysis index of vowel harmony by the information-theoretic notions of ‘positive logarithm’ and ‘mutual information.’ Mutual information (MI) between the vowels in autosegmental tiers leads to two findings about the declination of vowel harmony in Korean. Firstly, there were two sharp declines of vowel harmony in the historical Korean corpus. There was first a declination between the 15th and 16th centuries, and the second declination occurred between the 17th and 18th centuries. Secondly, the prohibition of non-harmonic vowels is more effective than the preference for harmonic vowels in Korean vowel harmony. (Keimyung University)

Keywords: Korean vowel harmony, vowel system, palatal harmony, Information Theory, mutual information

1. 머리말

어떠한 음운론적 단위가 출현하거나 나타난 단위들이 서로 연결될 확률을 구할 수 있다면 이를 바탕으로 여러 가지 음운현상이 적용될 가능성과 유형들을 예측할 수 있다. 분절음의 확률을 바탕으로 음운현상을 분석하고 예측하는 연구에서는 주로 ‘정보이론(Information Theory, Shannon and Weaver

* 본 연구는 2014년도 계명대학교 비사(신진)연구기금으로 이루어졌습니다. 본 논문의 내용에 대하여 조언해 주신 익명의 심사위원들께 감사드립니다.

1949)’에서 제안된 지표들을 적용하고 있다. 정보이론에서 제안된 지표를 활용하면 의사소통 체계 안에서 신호나 메시지를 전달하기 위하여 얼마나 많은 양의 정보가 필요한지 수학적 지표로 표시할 수 있다(Hume and Mailhot 2013). 정보이론의 지표들은 전통적인 생성음운론에서 제안된 이분법적 음운자질과 달리 수학적 확률로 표시되므로 명확한 범주로 설명하기 어려운 음운현상이나 다양한 음운론적 변이의 양적 분석에 유리하다(홍성훈 2014).

본 연구에서는 정보이론의 지표를 통하여 한국어에서 관찰되는 모음조화 현상의 통시적 변화를 계량적으로 분석하고자 한다. 한국어의 모음조화는 모든 어형과 단어에 적용되지 않는다. 한국어의 모음조화는 훈민정음이 창제된 15세기 중엽 이후의 한글 자료에서부터 정확하게 확인할 수 있는데, 이미 15세기의 자료부터 ‘체언+조사, 어간+어미, 어근+접사’의 결합에서 모음조화에 어긋한 예들이 관찰된다(한영균 1996). 한국어의 모음조화는 다수의 예외가 존재하는 현상임에도 불구하고, 모음조화에 대한 연구는 모음조화의 비율보다는 모음조화와 모음체계의 관계, 모음조화 붕괴의 원인 등을 중심으로 논의되어 왔다.

음운론적 단위와 음운현상을 ‘이산적 범주(discrete category)’로 구분하는 구조음운론과 생성음운론의 분석 방법으로는 모음조화의 통시적 추세와 변화를 분석하기 어렵다. 중세한국어 시기의 모음조화는 어느 정도로 적용되었는지, 근대한국어 시기에 모음조화가 얼마나 약화되었는지 등에 대해서는 분석하기가 쉽지 않다. 본 연구에서는 Goldsmith (2002)에서 제안된 분절음 사이의 ‘상호정보량(mutual information)’을 통하여 15세기 중세한국어부터 18세기 근대한국어 시기까지 인접하는 ‘단모음(monothong)’들의 상호정보량을 측정하였다. 상호정보량의 통시적 분석을 통하여 15세기부터 18세기까지 중세한국어와 근대한국어 사이 모음조화의 양상이 어떻게 변화하였는지 계량적인 수치로 확인할 수 있다.

본 연구의 내용은 다음과 같이 구성되어 있다. 2장에서는 한국어 단모음체계의 변화와 모음조화의 관계에 대하여 개관하고, 한국어의 모음조화에 대한 선행연구들 가운데 양적 분석방법을 적용한 것들을 살펴보았다. 3장에서는 모음조화의 분석과 관련된 정보이론의 지표로서 ‘정보량(정보 엔트로피)’과 ‘상호정보량’의 개념을 소개하였다. 4장에서는 상호정보량 측정을 위한 말뭉치의 가공과 분석 방법을 설명하였다. 5장에서는 상호정보량을 바탕으로 15세기부터 18세기까지의 중세한국어, 근대한국어 자료의 모음조화를 분석하고, 분석 결과에 대하여 논의하였다. 6장에서는 주요 논의 내용을 정리하고, 정보이론 기반 연구의 의의를 전망하였다.

2. 한국어의 모음체계와 모음조화

2.1. 한국어 모음체계의 변화

모음체계에 대한 통시적 연구에 의하면 조화의 근간이 되는 15세기 훈민정음(訓民正音) 체계의 ‘양성모음(陽性母音)’과 ‘음성모음(陰性母音)’은 각각 고대한국어 모음체계의 후설모음과 전설모음에 대응된다. 고대한국어와 같이 모음조화의 대립 체계가 ‘구개적 조화(palatal harmony)’를 반영하는 경우 ‘합치(合致)’라는 용어로 불렸다(이기문 1972: 134). 구개적 모음조화 체계와 합치되는 고대한국어의 모음체계는 모음의 연쇄적 변화, /ɔ/의 비음운화, 이중모음 /에/와 /애/의 단모음화를 겪으면서 현대한국어에서는 전후설의 대립과 동떨어진 체계가 되었다.

(1) 모음체계의 변화 (이기문 1972, 1998)

- a. 고대한국어 (10세기 이전) b. 전기 중세한국어 (13세기)



- c. 후기 중세한국어 (15세기) d. 근대한국어 (18세기)



고대한국어의 모음체계 (1a)는 실선으로 표시된 음성모음과 점선으로 표시된 양성모음이 각각 전설모음과 후설모음의 분포를 제약하는 모음조화의 대립 체계와 일치되었다. (1a)와 같은 모음체계에서는 음성모음과 전설모음 (/우/ü, /으/ɔ̃, /어/ä)과 동일하며, 양성모음은 후설모음(/오/u, /으/ɔ, /아/a)과 동일하였다. 모음조화의 기준이 되는 [back] 자질을 기준으로 [+back]은 양성모음, [-back]은 음성모음과 일치된다고 볼 수 있다.

이후 모음의 연쇄적인 변화에 의해 중세 및 근대한국어의 모음체계 (1b), (1c), (1d)는 구개적 모음조화의 체계와 동떨어진 체계로 변화하였다. 전기 중세한국어 시기에는 (1a)와 같이 전설저모음 /어/가 원래 자리에서 /이/ 아래의 중모음 위치로 이동하고, 연쇄적으로 /아/가 /어/의 자리로 이동하면

서 ‘전설모음=음성모음, 후설모음=양성모음’의 공식은 깨졌다. 양성모음과 음성모음의 대립을 더 이상 [back] 자질로는 설명할 수 없는 체계로 변화되었다. 후기 중세한국어 시기에는 /어/가 /으/의 자리로 이동하고 /으/가 그 위에 있는 /우/의 위치로, /우/는 후설모음 /오/의 위치로 /오/는 /으/의 위치로 이동되는 모음추이의 결과 (1c)와 같이 음성모음과 양성모음이 전설과 후설의 위치에 뒤섞이게 되었다. 근대한국어에서는 음성모음 /으/의 짝이었던 /으/가 소실되고, 이중모음이었던 /애/([aj])와 /에/([ej])가 각각 /ε/와 /e/로 단모음화되면서 (1d)와 같이 /어/-/아/와 /오/-/우/만 남은 사선적 대립의 체계가 되었다. 따라서 전기 중세한국어부터 근대한국어 시기에 이르기까지 양성모음과 음성모음의 대립은 고대한국어 시기 [back]의 대립으로부터 점점 멀어져서 전후설의 대립과 동떨어진 체계로 변화되었다.

모음추이에 의하여 전설모음과 후설모음, 음성모음과 양성모음의 양 대립 체계가 어긋날 경우 모음조화의 재조정(재조정)이 일어날 가능성도 있으나, 한국어에서는 모음조화의 재조정 대신 체계 사이의 ‘불합치’를 유지하면서 모음조화가 고대한국어 시기의 구개적 조화와 동일한 방식으로 유지되었다(최태영 1990: 73). 따라서 한국어의 모음조화는 음성학적 자질인 [back] 만으로는 설명할 수 없고, 양성(양성)과 음성(음성)이라는 추상적인 자질로 설명할 수 밖에 없는 현상으로 변화되었다. 이러한 변화에 따라 조화 모음들 사이의 음성학적 특성([±back])이 사라지면서 고대한국어 시기에 엄격했던 것으로 추정되는 모음조화 현상은 중세 및 근대한국어 시기를 거치면서 차츰 약화하였다. 따라서 (1)과 같은 모음체계의 변화를 고려한다면 모음체계와 모음조화의 체계 사이의 관계는 다음과 같이 정리할 수 있다.

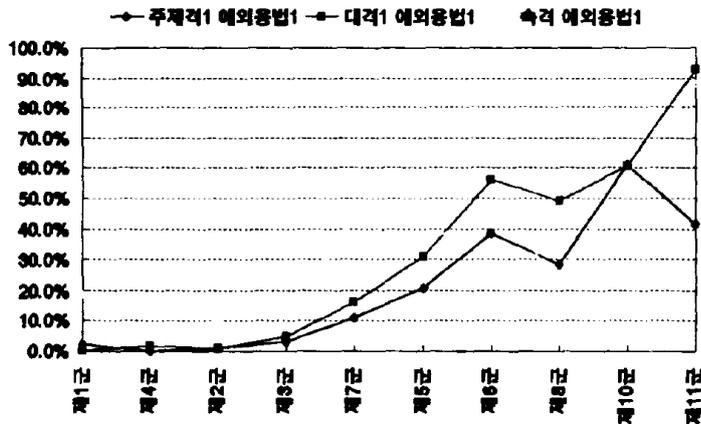
(2) 한국어 ‘모음체계’와 ‘모음조화’의 관계

- a. 고대한국어 시기에는 모음체계와 구개적 모음조화(전설:후설, 음성:양성)의 체계가 합치되었다.
- b. 중세한국어, 근대한국어 시기에 ‘모음추이, /으/의 비음운화, /에/와 /애/의 단모음화’와 같은 통시적 변화를 겪으면서 모음체계와 모음조화의 체계는 분리되었다.
- c. 모음체계와 모음조화 체계가 합치되지 않더라도 기존의 모음조화는 적용될 수 있었다.
- d. 모음체계가 고대한국어 시기의 구개적 모음조화 체계로부터 멀어질수록 모음조화 현상도 약화되었다.
- e. 고대한국어 시기에 엄밀하게 적용되었던 모음조화 현상은 모음체계의 통시적 변화를 겪으면서 후대로 내려올수록 점차 약화되었다.

2.2. 모음조화의 양적 분석

한국어의 모음조화를 통시적으로 논의한 대부분의 연구들(김완진 1978, 조성문 2001등)에서는 대부분 고대한국어로부터 현대한국어에 이르기까지, 모음체계의 변화로 인한 모음조화의 점진적 붕괴를 지적하고 있다. 그러나 모음조화가 어느 정도의 수준으로 약화되었는지, 계량적으로 분석한 연구는 흔하지 않다. 한영균(1996)에서는 ‘체언+격조사’에서 모음조화가 적용되지 않는 예외 형태들을 1유형(양성+음성), 2유형(음성+양성)으로 구분하고, 15세기와 16세기 사이에 모음조화를 따르지 않는 예외의 비율이 어떻게 변화되는지 분석하였다. 분석 문헌의 성격이나 체언의 특성(고유어, 한자어)에 따라 차이가 있었으나, 후대로 올수록 모음조화의 예외가 늘어나며, 1유형과 2유형의 예외가 정비례한다는 점을 밝혔다.

(3) ‘고유어 체언+격조사’ 모음조화의 예외 비율 (한영균 1996: 395)



위의 도표는 후기 중세한국어 시기 1447~1590년 사이의 문헌에서 관찰되는 모음조화의 예외(양성+음성)를 분석한 결과이다. 모음조화와 무관한 한자어를 제외하고 양성모음을 가진 고유어 체언이 음성모음의 조사 ‘-은, -을, -의’과 결합한 예외들을 조사한 결과 제1군 자료(1447)로부터 제11군 자료(1590)까지 후대로 내려올수록 ‘체언(양성)+조사(음성)’로 구성된 모음조화의 예외들이 대폭 증가한다는 사실을 보여 준다.

현대한국어의 모음체계는 고대한국어나 중세한국어와 전혀 다르지만 고

유어와 음성상징어에서는 여전히 모음조화가 적용되고 있다. Hong (2010)에서는 ‘관측빈도(observed frequency)’와 ‘기대빈도(expected frequency)’를 바탕으로 단일형태소로 구성된 고유어에 나타나는 현대한국어의 모음조화를 분석하였다. Hong (2010)에서는 고유어 단어에서 모음의 층렬을 분리하여 각 모음의 출현 확률을 바탕으로 두 모음이 함께 나타날 빈도를 예측하고, 기대빈도를 두 모음이 함께 출현하는 관측빈도와 비교하였다. 관측빈도는 분석 대상 자료에서 실제로 관찰된 빈도를 의미하며, 기대빈도는 각 모음의 출현 확률을 바탕으로 두 모음이 함께 나타날 빈도를 계산한 결과이다. 두 모음 V_1 과 V_2 가 함께 나타날 기대빈도 $E(V_1, V_2)$ 는 $P(V_1) \times P(V_2) \times N$ 의 공식으로 계산된다. 여기서 N 은 두 모음이 연결되는 모든 ‘경우의 수’를 의미한다. 예를 들어 ‘가위바위보 놀이’를 두 번 시행한다면 각각의 시행에서 가위가 나올 확률은 1/3이고 모든 경우의 수는 9이므로 연속으로 두 번 모두 ‘가위’가 나올 기대빈도는 $1(=1/3 \times 1/3 \times 9)$ 이 된다.

(4) 첫째 모음과 둘째 모음의 ‘관측빈도/기대빈도’ (Hong 2010: 288)

$V_1 \setminus V_2$	[+hi, +ATR] (i ɨ u)	[-hi, +ATR] (e ə)	[-ATR] (æ a o)	Total
[+high, +ATR] (i ɨ u)	O: 8093 E: 7484.79 O/E: 1.08	3290 2314.78 1.42	3446 5029.43 0.69	14829
[-high, +ATR] (e ə)	4980 4923.24 1.01	2652 1522.58 1.74	2122 3308.18 0.64	9754
[-ATR] (æ a o)	11204 11868.97 0.94	1566 3670.64 0.43	10745 7975.39 1.35	23515
Total	24277	7508	16313	48098

($N=48098, \chi^2=4428.359, p < .001$)

(4)의 표는 [-ATR]로 분류된 [æ](애), [a](아), [o](오)와 [+ATR]로 분류된 [i](이), [ɨ](으), [u](우), [e](에), [ə](어)의 공기 관계를 분석한 결과이다. 첫 모음과 둘째 모음의 [ATR] 자질이 어긋날 경우(음영 표시) 관측빈도(첫째 줄)가 기대빈도(둘째 줄)보다 훨씬 낮아서 ‘관측빈도/기대빈도’(O/E, 셋째 줄)가 0.69, 0.64, 0.43에 불과하였다. 하지만 두 모음의 [ATR] 자질이 일치할 경우에는 관측빈도가 기대빈도보다 높아서 O/E의 값이 1보다 컸다. 따라서 모음조화를 따르는 모음의 쌍들은 O/E의 값이 대부분 1보다 크고, 모음조화를 따르지 않는 모음의 쌍들은 O/E의 값이 1보다 작은 것으로 분석되었다. Hong (2010)의 계량적 분석에 의하면 현대한국어에서는 [ATR] 자

질에 의한 모음조화 현상이 있으며, 조화가 일어나는 유형 혹은 조화가 일어나지 않는 유형들 사이에도 세부적인 차이가 나타난다는 점을 확인할 수 있다.

본 연구에서는 한영균(1996)과 마찬가지로 모음조화의 통시적 변화를 고려하되, 정보이론의 상호정보량을 바탕으로 모음조화의 약화를 통시적으로 분석해 보고자 한다. 한영균(1996)에서는 15세기 ‘체언+조사’에서 관찰되는 모음조화를 미시적으로 분석하였으나, 본 연구에서는 15세기부터 18세기 사이의 역사자료 말뭉치를 토대로 거시적 분석을 진행하겠다. 다음 장에서는 상호정보량의 개념을 소개하고, 말뭉치를 가공하여 상호정보량으로 모음조화의 강도를 측정하는 과정을 설명하겠다.

3. 상호정보량과 모음조화

모음조화 현상은 모음의 분포가 선행 모음에 의하여 제약되는 현상이므로 어느 정도의 수준으로 모음조화가 일어나는지 파악하기 위해서는 모음의 출현 빈도, 선행 모음과 후행 모음 사이의 ‘공기 빈도(co-occurrence frequency)’를 측정하고 그 결과를 확률적으로 분석해야 한다. Goldsmith (2002)에서 제안된 확률적 모델이나 Hume and Bromberg (2005), Hume and Mailhot (2013) 등에서 적용된 음운이론은 흔히 ‘정보이론’이란 이름으로 불리고 있다. 정보이론에서는 단어 중심의 빈도분석과 달리 말뭉치에 나타나는 음소의 빈도를 기반으로 분절음이나 ‘바이그램(bigram)’의 확률을 계산하고 확률을 바탕으로 계산된 ‘정보량’을 이용한다. 정보량을 이용하면 삽입이나 탈락이 선호되는 분절음을 예상할 수 있고, 분절음과 단어의 음운론적 복잡성을 측정하여 어떠한 음운현상이 선호되는지도 예측할 수 있다. 여기서는 모음조화의 분석을 위하여 정보이론의 기본 개념인 분절음의 정보량과 두 분절음의 공기 빈도를 반영하는 상호정보량에 대해서만 살펴보겠다.

‘정보 엔트로피’(information entropy)로도 불리는 ‘정보량’은 정보의 불확실성을 나타내는 척도이다¹. 정보량의 개념은 우리에게 친숙한 ‘경우의 수’와 비슷하다. 예를 들어 가위바위보 놀이를 한다면 세 가지 경우의 수를 예측해야 한다. 그런데 항상 바위만 내는 상대가 있다면 경우의 수는 하나로 줄어든다. 정보량의 비교한다면 경우의 수가 셋이고, 바위를 낼 확률이 1/3인 전자의 경우가 경우의 수가 하나이고 바위를 낼 확률이 1/1인 후자

¹ ‘정보의 양’이 아니라, ‘신호를 받는 수신자가 예측해야 하는 정보의 양’ 정도로 이해하는 것이 적절하다.

의 경우보다 정보량이 높다. 상대가 무엇을 낼 것인지 예측해야 하는 정보의 양이 많기 때문이다. 따라서 바위를 낼 확률이 높을수록 정보량은 낮아지며 확률과 정보량은 반비례의 관계를 갖는다. 마찬가지로 어떠한 분절음이 나타날 확률은 정보량과 반비례한다. 일반적으로 유형빈도가 낮아서 출현 확률이 높은 모음은 정보량이 낮은 반면, 유형빈도가 높아서 출현 확률이 낮고 예측이 어려운 자음은 정보량이 높다.

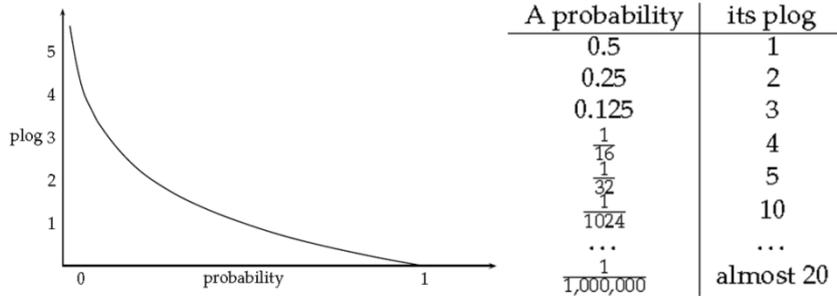
정보이론에서는 확률을 기준으로 정보량을 표시한다. 확률의 특성상 어떠한 분절음이 나타날 가능성은 언제나 1보다 작은 소수점 이하의 값을 가지므로 확률을 *plog*로 환산한다. *plog*는 다음과 같은 공식으로 구한다.

(5) *plog* (정보량)

$$plog(x) = -\log_2(x) = \log_2\left(\frac{1}{x}\right) \quad (x\text{는 확률})$$

양분법과 2진수는 정보를 표시하는 가장 효율적인 방법이므로 정보이론에서는 밑수 2를 갖는 로그 함수로 정보량을 측정한다. 어떠한 말뭉치 안에서 해당 분절음이 출현할 확률이 1/2이라면 *plog*(x)는 $\log_2 2$ 이므로 1이 된다. 확률이 1/4, 1/8로 낮아지면 *plog*(x)는 각각 $2(\log_2 4)$ 와 $3(\log_2 8)$ 으로 증가한다.

(6) 확률과 *plog*의 반비례 관계 (Goldsmith 2011: 2)



Goldsmith (2011), Hume and Bromberg (2005), Hong (2006) 등 정보이론을 적용하는 음운론적 연구에서는 *plog*의 값을 활용하여 음운현상과 음운체계의 변화를 분석하였다. 예를 들어 음운체계 안에 포함된 각 분절음들의 정보량을 계산하면 정보량이 적은 무표적 모음이나 자음을 예측할 수 있다 (Hume and Bromberg 2005, Hume 2006).

한국어의 모음조화는 선행 모음에 따라 후행 모음 분포가 제한되는 현

상이므로 조건부 확률로 분석할 수 있다. ‘상호정보량(mutual information, MI)’은 선행 모음에 대한 조건부 확률을 기반으로 다음과 같은 공식으로 계산한다. 간단히 말하자면 인접하는 두 분절음 a 와 b 의 상호정보량은 a 의 정보량과 b 의 정보량을 합한 값에서 ab 의 정보량을 빼서 구한다.

(7) 상호정보량: Unigram 모델과 Bigram 모델 사이의 \log_2 함수값 차이

$$MI(a; b) = \log \frac{\text{pr}(ab)}{\text{pr}(a)\text{pr}(b)} = \log \text{pr}(ab) - \log \text{pr}(a) - \log \text{pr}(b) \\ = -p\log(ab) + p\log(a) + p\log(b)$$

상호정보량은 a 와 b 가 동시가 나타날 확률과 관련된 지표이다. 만약 a 와 b 가 아무런 관계를 갖지 않는다면 a 와 b 가 동시에 나타날 확률은 a 가 나타날 확률과 b 가 나타날 확률에 의해 결정된다. 예를 들어 가위바위보 놀이를 두 번 반복해서 처음에는 가위가 나오고 다음에는 바위가 나올 확률 $1/9$ 은 가위가 나올 확률 $1/3$ 과 바위가 나올 확률 $1/3$ 을 곱한 값과 같다. 이러한 경우 (7)의 공식에서 분자 $\text{pr}(ab)$ 와 분모 $\text{pr}(a)\times\text{pr}(b)$ 의 값이 동일하므로 상호정보량은 0이 된다. 만약 처음에 가위를 낼 경우, 그 다음에는 바위를 내기로 약속되어 있다면 가위와 바위가 순서대로 나올 확률은 처음에 가위를 낼 확률인 $1/3$ 이 된다. 따라서 상호정보량은 ‘ $\log_2 3 (= -\log_2 3 + \log_2 3 + \log_2 3)$ ’이므로 0보다 높은 양의 값을 갖는다. 반면 가위 다음에는 되도록 바위를 내지 않는다는 제약이 있어서 가위와 바위가 순서대로 나올 확률은 $1/9$ 이하로 내려가면 상호정보량은 0보다 낮은 음의 값을 갖는다. 상호정보량을 모음조화의 분석에 적용한다면 두 모음의 관계에 따라 다음과 같이 해석할 수 있다.

(8) 확률과 상호정보량: 모음 V_1 과 V_2 의 관계

a. $\text{pr}(V_1V_2) = \text{pr}(V_1)\times\text{pr}(V_2)$

V_1 과 V_2 사이의 상호정보량은 0, 두 분절음의 관계는 독립적
(V_1 은 V_2 의 출현에 영향을 미치지 못함, 모음조화와 무관, 중립모음)

b. $\text{pr}(V_1V_2) > \text{pr}(V_1)\times\text{pr}(V_2)$

V_1 과 V_2 사이의 상호정보량은 0보다 높다.
(V_1+V_2 선호, 모음조화 반영, 양성+양성, 음성+음성 선호)

c. $\text{pr}(V_1V_2) < \text{pr}(V_1)\times\text{pr}(V_2)$

V_1 과 V_2 사이의 상호정보량은 0보다 낮다.
(V_1+V_2 회피, 모음조화 반영, 양성+음성, 음성+양성 회피)

모음조화가 일어나면 선행 음절의 모음에 의해 후행 음절의 모음이 결정되므로 ‘양성+양성, 음성+음성’의 경우 $pr(V_1V_2)$ 이 상승하고, ‘양성+음성, 음성+양성’의 경우 $pr(V_1V_2)$ 이 하강한다. 따라서 고대한국어나 중세한국어 시기와 같이 현대한국어보다 모음조화가 엄격하게 적용된다면 조화가 일어나는 모음 사이의 상호정보량은 양수가 되고 조화를 겪지 않는 모음 사이의 상호정보량은 음수가 될 것이다. 모음조화가 엄격할수록 $pr(V_1V_2)$ 와 상호정보량이 보다 큰 쪽으로 상승하거나 하강하므로 ‘조화 유형’(양성+양성, 음성+음성)과 ‘회피 유형’(양성+음성, 음성+양성) 사이 상호정보량의 차이도 커질 것이라고 예상할 수 있다. 반면 중립모음 /i/는 선행음 모음을 제한하지 않으므로 중립모음이 포함되면 V_1 과 V_2 의 상호정보량은 0에 수렴될 것이다.

4. 말뭉치의 가공과 상호정보량 측정

본 연구에서는 15세기부터 18세기까지 모음조화의 통시적 변화를 분석하기 위하여 21세기 세종계획을 통해 구축된 역사자료 말뭉치를 바탕으로 7가지 단모음(양성: /ㅏ/, /ㅓ/, /ㅗ/, 음성: /ㅑ/, /ㅕ/, /ㅛ/, 중립: /ㅣ/)의 빈도와 확률, 정보량을 측정하였다². 단모음의 정보량을 바탕으로 자립분절적 층렬에서 인접하는 모음의 상호정보량을 측정하여 모음조화의 강도를 측정하였다. 말뭉치는 다음과 같은 과정에 따라 가공하였다.

- (9) 원시 말뭉치 가공과 상호정보량 분석 과정
- a. 15세기, 16세기, 17세기 18세기 말뭉치 파일 분리
 - b. 어절 단위 가나다순 정렬
 - c. 한자 표기 어절 삭제
 - d. 단모음의 출현빈도 측정
 - e. 두 모음의 연결 출현빈도 측정
 - f. 인접한 두 모음의 상호정보량 분석

SynKDP에서 제공하는 21세기 세종계획 역사말뭉치는 15세부터 19세기

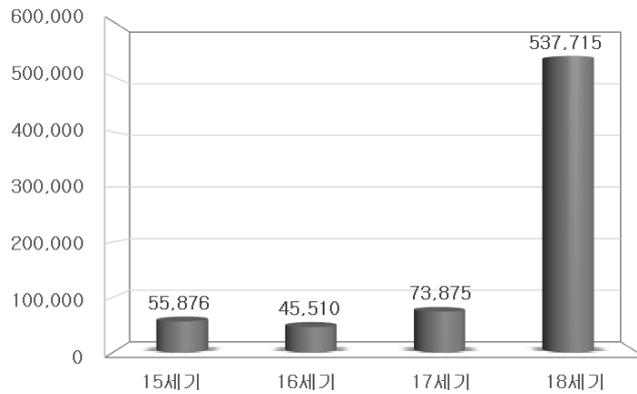
² 본 연구에서 분석한 ‘21세기 세종계획의 역사자료 말뭉치’는 홍윤표(2012)를 통하여 배포된 ‘SynKDP(감짜새) 1.5.1’에 포함된 원시말뭉치 자료이다. 이 자료에는 15세기부터 19세기까지의 한글문헌 자료들이 연대순으로 입력되어 있다.

의 자료가 하나의 파일로 통합되어 있어서 문헌자료의 간행 연도를 기준으로 15세기(~1500), 16세기(1501~1600), 17세기(1601~1700), 18세기(1701~1800) 네 가지 원시말뭉치 파일로 분리하였다. 18세기 말엽에는 /으/의 비음운화와 이중모음 /애/와 /에/의 단모음화가 완료되므로, 19세기 이후 자료는 분석에서 제외하였다. 본 연구에서는 양성모음 /아/, /오/, /우/와 음성모음 /어/, /우/, /으/의 조화를 다루므로, 모음체계 상으로 양성과 음성의 균형이 깨진 19세기 이후의 다루는 자료는 이전의 자료와 동일한 방식으로 분석하기 어려웠다.

세기별로 분리하여 어절별로 정리한 원시말뭉치의 분량에는 차이가 있었다. 15세기 파일에는 55,876개의 어절과 150,879개의 단모음, 16세기 파일에는 45,510개의 어절과 107,404개의 단모음, 17세기 파일은 73,875개의 어절과 176,524개의 단모음, 18세기 파일은 537,715개의 어절과 1,371,004개의 단모음이 포함되었다.

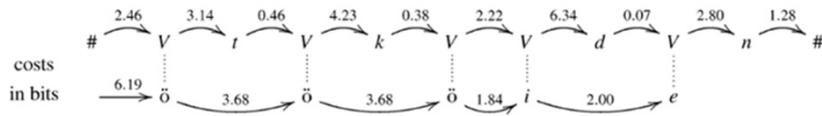
(10) 세기별 분석 어절과 단모음의 개수

a. 어절 개수 (합계: 712,976개)



모음조화는 개재 자음과 상관없이 일어나는 현상이므로 어절로부터 모음을 추출하여 독립된 층렬을 기준으로 인접 모음의 연결빈도를 측정하였다. (12)는 Goldsmith and Riggle (2012)에서 제안된 핀란드어의 모음조화 분석 모델로서 모음의 층렬을 분리하여 ‘/ø/-/ø/-/i/-/e/’에서 연결되는 두 모음의 정보량을 측정한 결과이다.

(12) 모음의 자립분절적 층렬 모델 (Goldsmith and Riggle 2012: 877)

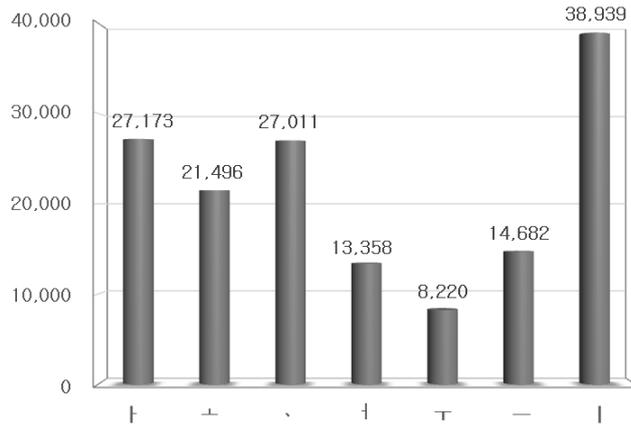


핀란드어의 모음조화에서는 전설모음인 /ä/, /ø/, /y/와 후설모음 /a/, /o/, /u/가 대립되며, 중립모음 /e/와 /i/는 전설모음 및 후설모음과 모두 어울릴 수 있다. (12)와 같이 모음층렬을 분리하여 ‘/ø/-/ø/, /ø/-/i/, /i/-/e/’의 정보량을 측정한 결과 전설모음끼리 연결된 ‘/ø/-/ø/’의 상호정보량(2.54)은 전설모음과 중립모음이 연결된 ‘/ø/-/i/’의 상호정보량(0.32)이나 중립모음끼리 연결된 ‘/i/-/e/’의 상호정보량(0.62)보다 훨씬 높았다(Goldsmith and Riggle 2012: 874).

인접 모음의 연결 빈도는 모음의 층렬을 분리하여 선행 음절과 후행 음절의 모음이 모두 단모음인 경우를 기준으로 측정하되, 어절 전체를 기준으로 측정하기 않고 인접한 두 음소를 기준으로 측정하였다. 예를 들어 ‘가곡흐면’이라는 어절을 분석하는 경우, 모음 층렬을 /아/-/으/-/으/-/여/로 분리하고 이 가운데 단모음의 연결빈도를 측정하였다. ‘가곡흐면’의 모든 모음을 고려한다면 단모음 /아/와 /으/는 양성모음이고, /여/는 음성모음이므로 모음조화의 예외로 처리되겠지만 본 연구에서는 단모음만을 고려하므로 모음조화를 따르는 모음의 연결 /아/-/으/와 /으/-/으/가 각각 1번씩 출현하는 것으로 계산되었다. 단모음과 단모음 사이의 연결빈도를 측정한 결과를 바탕으로 ‘조화 유형’(양성+양성, 음성+음성), ‘회피 유형’(양성+음성, 음성+양성)과 ‘중립모음이 포함된 유형’(양성+중립, 중립+양성, 음성+중립, 중립+음성, 중립+중립)으로 구분하여 연결빈도를 통합하였다. (13b)와 (13c)에서 음영으로 표시된 영역은 모음조화가 적용된 ‘양성+양성, 음성+음성’ 모음의 연결빈도이다.

(13) 단모음의 빈도와 모음의 연결빈도 (15세기)

a. 단모음의 빈도



b. 단모음의 연결 빈도 (V_1+V_2)

		양성			음성			중립
		ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	
양성	ㅏ	2,320	2,521	4,114	383	102	265	3,381
	ㅑ	3,121	1,433	2,844	276	132	242	3,745
	ㅓ	2,139	2,765	3,450	573	86	145	6,128
음성	ㅕ	1,154	658	867	914	901	2,719	3,074
	ㅗ	230	411	621	571	522	1,348	1,792
	ㅛ	535	1,021	798	1,305	494	1,504	2,994
중립	ㅣ	5,521	3,854	3,156	1,904	1,040	1,719	5,246

c. 양성, 음성, 중립 모음의 연결빈도 (V_1+V_2)

$V_1 \backslash V_2$	양성	음성	중립
양성	24,707	2,204	13,254
음성	6,295	10,278	7,860
중립	12,531	4,463	5,246

최종적으로 단모음의 출현빈도와 연결빈도를 (7)의 공식에 대입하여 인

접하는 두 단모음의 상호정보량을 산출하였다. (14)에 제시된 두 모음의 상호정보량을 보면 이론적 예측과 달리 실제 상호정보량은 ‘조화 유형’(양성+양성, 음성+음성)에서도 0보다 낮은 음수인 경우가 많았다.

(14) 인접 모음의 상호정보량(MI)과 보정 값(MI+0.905) (15세기)

V ₁	V ₂	상호정보량 (V ₁ V ₂)	상호정보량+0.905
양성	양성	-0.620	0.285
음성	음성	0.012	0.917
양성	음성	-3.271	-2.366
음성	양성	-1.757	-0.852
중립	중립	-1.164	-0.259
중립	양성	-0.867	0.038
양성	중립	-0.786	0.119
중립	음성	-1.231	-0.326
음성	중립	-0.478	0.427

조화유형의 상호정보량이 음수로 나타나는 원인은 (15)의 사례와 같이 모음조화와 무관한 형태소나 단음절 어절로 인하여 연결빈도의 확률이 낮아진 결과로 보인다.

(15) 모음조화의 예외적 요소 (15세기)

- a. 조사: -ㄱ장, -ㄷ려, ...
- b. 어미: -술/술/줄-, -고, -다, -ㄷ록, -잇든, ...
- c. 파생접미사: -곤ㅎ-, -ㅎ-, -곤ㅎ-, -롭-, -ㄷ뵈/ㄷ뵈-, ...
- d. 단음절 어절: 가, 간, 갈, 값, 감, ...

조화 유형의 상호정보량이 음의 값을 갖는 문제점을 해결하기 위하여 모음조화와 무관한 중립 유형의 평균이 0이 되도록 각 유형의 상호정보량에 보정 상수 ‘0.905’를 더하였다. (14)의 오른 편에는 보정된 상호정보량을 제시하였는데, 보정 상수를 더한 결과 조화 유형은 양의 값을, 회피 유형은 음의 값을, 중립 유형은 0에 가까운 값을 갖게 되었다. 다만 세기별로 구한 보정 상수는 차이가 있었다. 중립유형의 평균 상호정보량이 0이 되도록 보정 상수를 설정한 결과, 16세기 보정 상수는 0.528, 17세기는 0.886, 18

세기는 4.271이었다.

(8)에서 설명하였듯이 상호정보량은 모음조화의 세 가지 양상을 반영한다. 첫째 ‘양성+양성, 음성+음성’과 같이 모음조화가 일어나는 유형에서는 상호정보량은 0보다 큰 양의 값을 가지며 조화가 엄격할수록 상호정보량도 상승한다. 둘째 ‘양성+음성, 음성+양성’과 같이 모음조화에서 회피되는 유형의 상호정보량은 0보다 작은 음의 값을 가지며 조화가 엄격하여 회피 현상이 강할수록 상호정보량은 하강한다. 셋째, 모음조화와 무관한 경우, 예를 들어 중립모음과 양성, 음성, 중립모음이 연결되는 중립 유형에서 상호정보량은 0에 가까운 값을 갖는다. 모음조화가 강화될수록 조화 유형의 상호정보량은 상승하고, 회피 유형의 상호정보량이 하강하므로 모음조화의 강도는 다음과 같은 지수로 공식화할 수 있다.

(16) 모음조화 지수 (Vowel Harmony Index)

- a. 모음조화 지수 = 상호정보량(조화 유형) - 상호정보량(회피 유형)
- b. 모음조화 현상이 강화되면 모음조화 지수는 상승한다.
- c. 모음조화 현상이 약화되면 모음조화 지수는 하강한다.

모음조화 지수의 공식에 따라 조화 유형의 상호정보량과 회피 유형의 상호정보량을 보정한 이후에 두 보정값의 차이를 계산하여 세기별 모음조화 지수를 구하였다. 모음조화 지수가 가장 높을 것으로 예상되는 15세기의 경우 조화 유형의 평균 상호정보량은 0.601이고, 회피 유형의 평균 상호정보량은 -1.609로서 모음조화 지수는 2.210으로 계산되었다.

(17) 조화 유형별 인접 모음의 상호정보량(MI)과 보정 값(MI+0.95) (15세기)

유형	상호정보량 (V_1V_2)	상호정보량+0.905
조화 유형	-0.304	0.601
회피 유형	-2.514	-1.609
중립 유형	-0.905	0.000
모음조화 지수	2.210	

5. 분석 결과와 논의

단모음의 정보량과 인접하는 두 단모음의 정보량을 통하여 15세기부터 18세기까지 모음조화의 유형별로 상호정보량을 측정할 결과는 다음과 같았다.

(18) 15~18세기 인접 단모음의 상호정보량 (양성, 음성, 중립 모음)

V ₁	V ₂	15세기	16세기	17세기	18세기
양성	양성	-0.620	-0.367	-0.833	-4.180
음성	음성	0.012	0.140	-0.202	-4.072
양성	음성	-3.271	-1.807	-1.876	-4.694
음성	양성	-1.757	-1.084	-1.272	-4.402
중립	중립	-1.164	-0.813	-1.172	-4.405
중립	양성	-0.867	-0.516	-0.710	-4.215
양성	중립	-0.786	-0.404	-0.925	-4.103
중립	음성	-1.231	-0.830	-1.210	-4.554
음성	중립	-0.478	-0.074	-0.414	-4.075

음영으로 표시된 ‘조화 유형’(양성+양성, 음성+음성)의 상호정보량이 회피나 중립 유형의 상호정보량보다는 전반적으로 높은 편이었지만, 앞서 지적하였듯이 모음조화와 무관한 형태소나 단음절 어절의 영향으로 조화 유형의 상호정보량도 0보다 낮은 음수가 되는 경우가 많았다. 특히 18세기 자료의 분석 결과에서는 모든 유형의 상호정보량이 -4보다 낮은 편이었다. 이러한 문제를 해결하기 위하여 중립유형의 평균 값이 0이 되도록 보정 상수를 더한 결과 다음과 같은 결과를 얻었다.

(19) 15~18세기 인접 단모음의 상호정보량 (+보정 상수)

V ₁	V ₂	15세기 (+0.905)	16세기 (+0.528)	17세기 (+0.886)	18세기 (+4.271)
양성	양성	0.285	0.161	0.053	0.091
음성	음성	0.917	0.668	0.684	0.199
양성	음성	-2.366	-1.279	-0.990	-0.423
음성	양성	-0.852	-0.556	-0.386	-0.131
중립	중립	-0.259	-0.285	-0.286	-0.134
중립	양성	0.038	0.012	0.176	0.056
양성	중립	0.119	0.124	-0.039	0.168
중립	음성	-0.326	-0.302	-0.324	-0.283
음성	중립	0.427	0.454	0.472	0.196

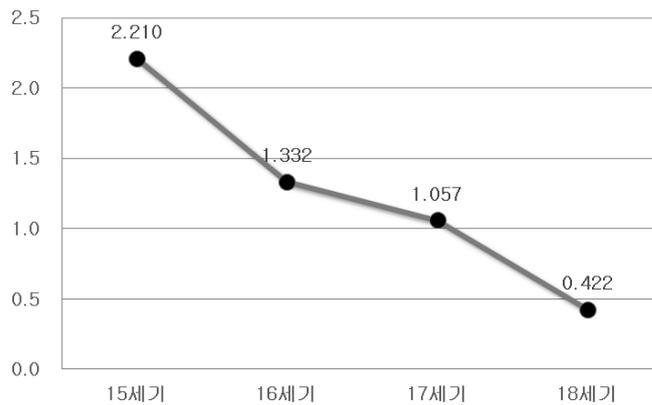
상수를 더하여 보정한 상호정보량을 구한 결과 ‘조화 유형(양성+양성, 음성+음성)’과 ‘회피 유형(양성+음성, 음성+양성)’의 차이가 뚜렷하였다. 조화 유형의 상호정보량은 모두 양의 값으로, 회피 유형의 상호정보량은 모두 음의 값으로 측정되었다. 전반적으로 회피 유형의 절대값이 높은 편이었는데, 조화 유형에서는 ‘양성+양성’보다는 ‘음성+음성’이 선호되고, 회피 유형에서는 ‘양성+음성’이 회피되는 것으로 분석되었다. 상호정보량의 측정 결과에 의하면 선행행 모음의 자질을 일치시키는 모음조화는 선행 음절이 음성모음인 경우에 잘 적용되고, 선행 모음과 성격이 다른 후행 모음을 회피하는 모음조화는 선행 음절이 양성모음인 경우에 잘 적용된다고 볼 수 있다.

조화 유형과 회피 유형 사이의 차이는 15세기가 가장 큰 편이었으며 18세기까지 점진적으로 줄어들었다. 조화 유형과 회피 유형의 평균을 기준으로 두 유형의 차이를 통하여 모음조화 지수를 구한 결과는 다음과 같았다.

(20) a. 15~18세기 인접 단모음의 상호정보량 (조화, 회피 유형)

유형	15세기	16세기	17세기	18세기
조화	0.601	0.415	0.369	0.145
회피	-1.609	-0.918	-0.688	-0.277
모음조화 지수 (=조화-회피)	2.210	1.332	1.057	0.422

b. 15~18세기 인접 단모음의 모음조화 지수

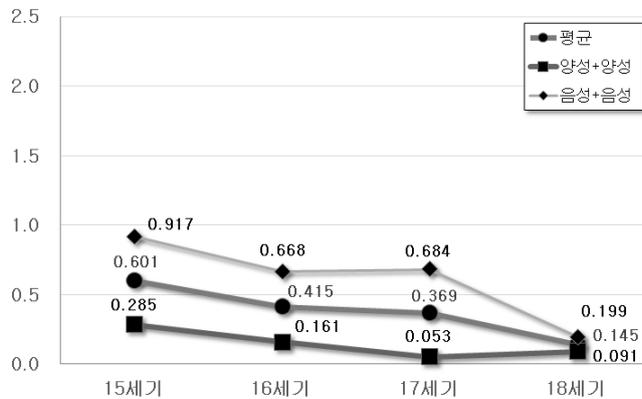


15세기부터 18세기까지의 상호정보량을 바탕으로 계산된 모음조화 지수의 변화를 고려한다면 모음조화의 강도는 이론적 가정대로 15세기로부터 후대로 내려올수록 지속적으로 약화되었다. 15세기와 16세기 사이의 하락폭(0.878)과 17세기와 18세기 사이의 하락폭(0.635)이 16세기와 17세기 사이의 하락폭(0.275)보다 큰 편이었다. 상호정보량의 분석에 의하면 모음조화는 중세한국어와 근대한국어의 전환기인 16~17세기보다는 후기 중세한국어 시기인 15~16세기와 근대한국어 시기인 17~18세기에 더 큰 변화를 겪었다.

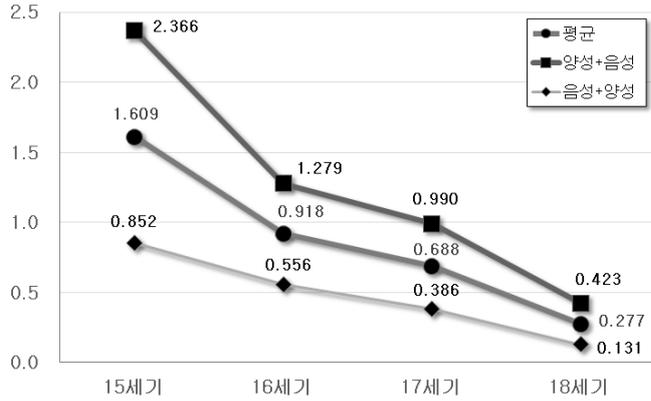
모음조화 지수와 관련된 조화 유형과 회피 유형의 상호정보량을 분리하여 살펴보니, 전반적으로 조화보다는 회피의 효과가 강한 것으로 분석되었다. 특히 모음조화 지수가 가장 높은 15세기의 분석 결과에서는 회피의 상호정보량(-1.609)이 조화의 상호정보량(0.601)보다 훨씬 강한 편이었다. 반면 16~18세기의 분석 결과, 후대로 내려올수록 모음조화 지수가 낮아지면서 조화와 회피 양 유형의 상호정보량 차이도 점진적으로 줄어들었다.

(21) 15~18세기 양성/음성모음 상호정보량의 변화

a. ‘조화 유형’의 상호정보량 (양성+양성, 음성+음성)

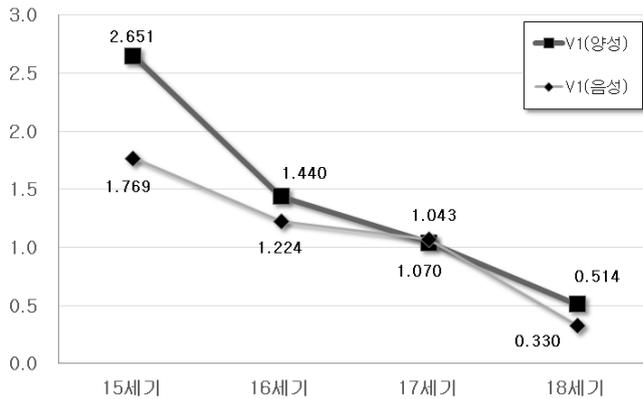


b. ‘회피 유형’의 상호정보량 (절댓값: 양성+음성, 음성+양성)



마지막으로 선행 모음(V_1)이 양성모음인 경우와 음성모음인 경우를 구분하여 모음조화 지수를 분석하였다. 전반적으로 선행모음이 양성모음인 경우가 선행모음이 음성모음인 경우보다 모음조화 지수가 높은 편이었으며, 그 차이는 15세기에 가장 컸다. 이러한 차이는 16~18세기에는 대폭 줄어들었다. 따라서 본래 후설모음이었던 양성모음이 선행하는 경우 모음조화의 효과가 뚜렷하게 나타났지만, 모음체계의 변화로 모음조화가 붕괴되면서 양성모음과 음성모음의 영향력에도 큰 차이가 없어진 것으로 보인다.

(22) 15~18세기 모음조화 지수의 변화



6. 맺음말

지금까지 한국어 모음조화 현상의 통시적 변화를 정보이론의 지표인 정보량과 상호정보량을 통하여 분석해 보았다. 후기 중세한국어와 근대한국어 시기에 간행된 한글자료의 역사자료를 말뭉치를 가공하여 어절 단위로 구분하여 정렬하고, 어절을 구성하는 단모음의 층렬을 분리하여 인접하는 두 가지 단모음 사이의 상호정보량을 측정하였다. 상호정보량의 분석 결과, 기존 연구의 이론적 논의에 부합되는 결론을 얻을 수 있었다. 15세부터 18세기까지 모음조화는 지속적으로 약화되었으며, 중세한국어와 근대한국어의 교체기인 16~17세기보다는 15~16세기와 17~18세기 사이에 심하게 약화되었다.

15세기에는 선행 모음과 후행 모음의 자질을 일치시키려는 조화의 효과보다는 선행 모음과 다른 자질을 가진 후행 모음을 회피하는 효과가 더 강하게 나타났다. 그러나 16~18세기 사이에 모음조화가 점진적으로 약화되면서 조화 효과와 회피 효과의 차이도 점차 줄어들었다. 후행 음절의 조화를 요구하는 영향력은 양성모음보다 음성모음이 강한 반면, 성격이 다른 후행 음절의 모음을 회피하는 영향력은 양성모음이 음성모음보다는 강한 것으로 분석되었다.

본 연구는 단모음만의 연결빈도를 고려하여 모음조화의 강도를 비교하였으므로 이중모음이 포함된 어절이나 어절 전체의 모음조화 양상은 분석하지 못하였다. 그러나 약 70만 어절에 포함된 180만여 개의 단모음을 대상으로 상호정보량을 분석한 결과, 모음조화의 약화를 가시적으로 확인할 수 있었다. 보다 정밀한 분석을 진행하기 위해서는 모음조화와 무관한 단음절 어절을 제외하고 이중모음이 포함된 음절도 함께 분석해야 한다.

정보이론의 상호정보량은 두 성분의 선호 관계 혹은 회피 관계를 분석하기에 유용하다. 음운론적 분석에 상호정보량을 도입하면 모음조화와 같이 앞뒤의 음소와 관련된 음운현상을 양적으로 분석할 수 있다. 또한 단어나 형태소의 연어 관계나 문장 성분의 통사적 호응관계도 분석이 가능하다. 따라서 상호정보량을 이용하면 음소나 단어 사이의 연결망을 구축하고 성분들 사이의 연결 관계를 분석하는 연구도 가능할 것으로 기대된다.

REFERENCES

- 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, July 16, 2011, 1-6.
- HONG, SUNG-HOON. 2006. Quantitative analysis of English hypocoristics: Wellformedness and phonological complexity. *Studies in Phonetics, Phonology and Morphology* 12.1, 211-229. The Phonology-Morphology Circle of Korea.
- _____. 2010. Gradient vowel cooccurrence restrictions in monomorphemic native Korean roots. *Studies in Phonetics, Phonology and Morphology* 16, 279-295. The Phonology-Morphology Circle of Korea.
- HUME, ELIZABETH. 2006. Language specific and universal markedness: An information-theoretic approach. Paper presented at annual meeting of the Linguistic Society of America, Colloquium on Information Theory and Phonology, January 2006, 1-10. Albuquerque, NM.
- HUME, ELIZABETH and ILANA BROMBERG. 2005. Predicting Epenthesis An Information-Theoretic Account. Proceedings of 7th Annual Meeting of the French Network of Phonology.
- HUME, ELIZABETH and FRÉDÉRIC MAILHOT. 2013. The role of entropy and surprisal in phonologization and language change. In A. Yu (ed.). *Origins of Sound Patterns: Approaches to Phonologization*. Oxford: Oxford University Press.
- SHANNON, CLAUDE E. and WARREN WEAVER 1949. *The Mathematical Theory of Communication*. Urbana and Chicago: University of Illinois Press.
- 김완진. 1978. 모음 체계와 모음조화에 대한 반성. *어학연구* 14.2, 127-139. 서울대학교 언어교육원.
- 이기문. 1972. *국어음운사연구*. 서울: 탑출판사.
- _____. 1998. *신정판 국어사개설*. 과주: 대학사.
- 조성문. 2001. 최적성이론에 의한 모음조화의 변화 분석. *음성·음운 형태론연구* 7.1, 191-213. 한국음운론학회.
- 최태영. 1990. 모음조화. *국어연구 어디까지 왔나*, 68-76, 동아출판사.
- 한영균. 1996. 모음조화 예외 출현 비율에 대한 통시적 해석. *관악어문연구* 21, 377-405. 서울대학교 국어국문학과.

홍성훈. 2014. *음운론의 계량적 방법론*. 한국문화사.
홍윤표. 2012. *국어정보학*. 태학사.

Sunwoo Park
Department of Korean Language Education
Keimyung University
1095 Dalgubeol-daero, Dalseo-gu, Daegu
South Korea 42601
e-mail: sunwoopark@kmu.ac.kr

received: November 21, 2016
revised: December 12, 2016
accepted: December 16, 2016