한국어 명사의 음소배열제약에 대한 기계학습*

박나영
(서울대학교)

Park, Nayoung. 2014. Machine learning of phonotactic constraints in Korean nouns. *Studies in Phonetics, Phonology and Morphology* 20.3. 297-322. For the purpose of investigating not only gradient but also categorical phonotactic constraints in Korean, this study explores the machine learning of phonotactic constraints in Korean. In so doing, it focuses on phonotactic differences between native and Sino-Korean nouns. Employing 5,543 native Korean and 29,869 Sino-Korean words as the input training data, I ran a learning simulation, using a Maximum Entropy phonotactic model (Hayes and Wilson 2008). Based on the statistical distribution of the input data, markedness constraints were created with their own weights, the size of which reflects their gradient strength. The simulation results mostly confirm previous descriptions of phonotactics in Korean (for instance, no word-initial tense consonants in Sino-Korean). In addition, some previously unreported patterns were found. **(Seoul National University)**

Keywords: Maximum Entropy phonotactic model, categorical phonotactics, gradient phonotactics, native-Korean words, Sino-Korean words

1. 서론

‘음소배열제약(phonotactic)’이란, 음소 단위의 결합 또는 회피에 대한 문법인식을 말하며, 일반적으로 가능한 음소 결합과 불가능한 음소 결합을 나누어 이분법적으로 분석하였다. Chomsky and Halle (1965: 101)는 영어 화자들이 아래 (1)과 같이 실재하는 ‘brick’뿐만 아니라, 비단어(nonce word) ‘blick’까지 적형적인(well-formed) 단어로 인식한다고 보았다. 반면, 비단어 ‘lbick’은 비적형적인(ill-formed) 단어로 인식된다고 지적하였다. 이같은 ‘적형-비적형’의 이분법적 문법성으로 한 언어 내에서 불가능한 연쇄, 즉, ‘범주적인 음소배열제약(categorical phonotactics)’이 포착되었다.

(1) 영어 어두 자음군에 대한 문법 인식
ㄱ. brick: 실재 단어(existing word)
ㄴ. blick: 적형 단어(well-formed word)
ㄷ. lbick: 비적형 단어(ill-formed word)

* 본 논문은 제5회 한국음운론학회 국제학술대회(2014. 7. 3-7.5)에서 발표한 내용을 수정하고 보완한 것입니다. 지속적으로 지도해 주신 전종호 선생님과 본 연구의 부족한 점을 검토해 주신 심사위원분들께 감사드립니다. 또한 본 연구를 시작할 수 있도록 입력 자료의 표본을 제공해 준 장하연에게 고마움을 표합니다. 물론 본 논문에서 드러나는 오류는 모두 필자의 책임임을 밝힙니다.

한편, 이보다 세분화된 문법 인식도 탐색되고 있다. 특히, 연쇄에 대한 수용도(acceptability)가 연쇄의 출현 빈도를 반영한다는 보고가 있다(Coleman and Pirrehumbert 1997, Hay et al. 2003 등). Hay et al. (2003: 5-6)은 고빈도 /nt/에 대한 수용도에 비해, 중빈도인 /mk/, /nf/에 대한 수용도가 낮다는 것을 보여 비음-장애음 연쇄에 대한 수용도와 빈도의 상관성을 밝혔다.

(2) 영어 비음-장애음 연쇄의 빈도와 수용도
ㄱ. 고빈도: /nt/
ㄴ. 중빈도: /ms/</mk/</nf/
ㄷ. 수용도: /ms/</mk/</nf/</nt/

이러한 관점에서 빈도가 낮은 음소 연쇄 중 통계적으로 유의미하게 회피되는 연쇄는 문법으로 포착될 수 있다. 이를 이른바 '비범주적 음소배열제약(gradient phonotactics)'으로 일컫는다.

본 연구는 한국어 명사를 대상으로 범주적 음소배열제약과 비범주적 음소배열제약을 종합적으로 다루고자 한다. 이제까지 한국어 음소 연쇄에 대해서는 주로 범주적인 음소배열제약이 포착되었다(허웅 1985, 신지영·차재은 2003). '활음과 모음의 결합(/ji, ɨi, wu, wo, wi/)이 출현하지 않는다'와 같이 한국어에서 불가능한 음소 연쇄에 대한 기술이 이에 해당한다. 한편, 한국어의 비범주적 음소배열제약에 대한 단서도 찾을 수 있다. 일례로, 신지영·차재은(2003)은 한국어 어휘 형태소에서 [ai]와 같은 모음 연쇄가 허용되기는 하지만, 선호되지 않는다고 지적한다.

다수의 양적 연구(김경일 1985, 진남택 1992, 유재원 1997, 이상억 2001, 신지영 2005, 2008, 2010, Lee 2007, Hong 2010, 김미란 외 2014)는 음소 연쇄가 고르게 분포하지 않는다는 것을 보이고, 비범주적 음소배열제약의 실재를 시사한다. 그러나 탐색 범위가 두 음소 연쇄에 국한되었고 양적 정보가 문법에 직접 반영될 수 있는 기제도 없다는 한계가 있었다.

음소배열제약을 전반적으로 찾고, 양적 정보를 문법에 포함하기 위해서는 보다 엄밀한 모델이 필요하다. 이를 위해 본 연구는 최대 엔트로피 음소배열제약 모델(Maximum Entropy Model of Phonotactics; Hayes and Wilson 2008)을 채택한다. 이 모델은 표면형(surface)으로부터 음소 연쇄의 출현 확률을 계산하고, 이 확률을 도출할 수 있는 제약과 그 가중치를 출력하는 기계학습방법이다. 이 방법은 영어를 포함한 다수의 언어에 대해 적용되어, 음소배열제약 및 강도를 예측하였다. 그리고 그 실재가 해당 언어의 화자를 대상으로 하는 실험을 통해 증명된 바 있다 (Hayes and Wilson 2008, Danald et al. 2011, Kager and Pater 2012, Colavin 2013, Hayes and White 2013).

이에 더하여 한국어의 음소배열제약은 고유어와 한자어를 구분하여 탐색될 필요가 있다. 다수의 연구(송기중 1992, 권인한 1997,

신지영·차재은 2003, 신지영 2009, 안소진 2009)는 한자어가 고유어에 비해 초성과 종성 위치에 오는 자음이 한정된다고 지적하였다. 또한 적어도 어원적으로는 구성 음절이 별개의 형태소인 바 개별 음절 단위로 음소배열제약을 탐색하였으며, 단어내 분절음 연쇄가 고유어보다 자유롭게 나타날 것을 시사하였다. 이와 함께 몇몇 연구(채서영 1999, 이주희 2005, 박선우 외 2013)는 '어휘적 계층(lexical strata; Itô and Mester 1999)'이란 개념을 도입하여 고유어와 한자어의 음운론적 지식을 구분하여 설정한 바 있다.

그럼에도 음소 결합에 대한 조사(진남택 1992, 유재원 1997, 이상억 2001, Lee 2007, 김미란 외 2014)는 대부분 고유어와 한자어를 구분하지 않는다. 앞서 Cho (2012)는 한국어에 대해 최대 엔트로피 음소배열제약 모델을 도입하여 기존에 간과된 음소배열제약을 새롭게 밝힌 바 있다. 그러나 어휘의 어종에 따라 음소배열제약을 나누어 다루지는 않았다.

이러한 점을 고려해 본 연구는 고유어와 한자어가 각각의 어휘 항목, 즉, 별도의 어휘부(lexicon)를 구성한다고 보고 각 어휘부 내에서 세분화된 음소배열제약(phonotacitcs)을 찾고자 한다. 이를 위해, 고유어와 한자어를 각각 분류하여 개별적인 기계학습을 진행한다. 다른 품사에는 한자어가 제한적으로 분포하기 때문에 두 어종에 대한 충분한 자료를 보장받을 수 있는 명사만을 대상으로 하였다.

우선, 2절에서는 최대 엔트로피 음소배열제약 모델을 소개하고, 3절에서는 실제 학습 자료 및 조건에 대해서 다룬다. 4절에서는 범주적 음소배열제약과 비범주적 음소배열제약에 대한 학습 결과를 살펴보고, 출력된 제약이 예측하는 바를 5절에서 다룬다. 마지막 6절에서는 결과의 특징을 요약하고, 본 모델의 의의를 밝힌다.

2. 최대 엔트로피 음소배열제약 모델

최대 엔트로피 음소배열제약 모델(Hayes and Wilson 2008)은 표면형에 출현하는 음소 결합 확률을 계산하여, 음소배열제약을 자동적으로 학습하는 방법이다. 학습 과정에서 연구자가 제약을 미리 입력하지 않기 때문에, 학습 문법은 입력 자료의 분포를 반영하는 귀납적인(inductive) 특성을 보인다.

이 모델은 음소 연쇄의 낮은 출현 확률(probability)이 낮은 적형성(well-formedness)으로 해석된다고 가정하고, 회피되는 음소 연쇄와 회피 정도를 제약 및 가중치로 포착한다. 학습자(learner)는 해당 언어에서 주어진 자질 목록(feature set)과 제약의 길이(the number of feature matrices)를 기준으로, 가능한 모든 제약을 탐색 범위로 생성한다. 이 중 정확도(accuracy)와 일반성(generality)의 척도에 따라, 음소배열제약이 선정된다.

제약(Ci)의 정확도(accuracy)는 예측되는 위배 빈도보다 실제 위배

되는 빈도가 훨씬 적은 제약을 선택하는 지표다. 한 제약이 실제 위배되는 빈도, 즉, 관찰 빈도(O[Ci])가 적다는 의미는 해당 제약이 입력 자료에서 빈도가 아주 낮은 연쇄를 포착하였다는 것이다. 또한 제약의 예측되는 위배 빈도, 즉, 기대 빈도(E[Ci])가 높다는 의미는 논리적으로 가능한 모든 표상 집합[1]에서 해당 제약을 위배하는 연쇄가 다수 출현한다는 것이다. 정확도는 (3)에서 보인 바와 같이 관찰 빈도를 기대빈도로 나누어, 실제 위배되는 빈도가 유의미한 정도를 가리킨다.[2] 관찰 빈도가 0이라도 기대 빈도가 큰 제약이 더 유효한 것으로 판단된다(Mikheev 1997).

(3) 정확도(accuracy): 제약 위배의 관찰 빈도/기대 빈도 비율

$$\text{제약의 정확도} = \frac{\text{관찰 빈도(O[Ci]): 실제 관찰된 제약의 위배 빈도}}{\text{기대 빈도(E[Ci]): 예측된 제약의 위배 빈도}}$$

이와 함께, 각 정확도 수준별로 더 일반적인(general) 제약을 학습한다. 예를 들어, 같은 정확도 수준에서 *[+전방성, +설정성][-후설성,+성절성]보다 더 포괄적인 자질로 구성된 *[+설정성][-후설성]이 선택된다. 그리고 결합되는 자질 매트릭스의 수가 적은 제약이 학습된다. 그리고 각 제약의 가중치는 연쇄의 출현 확률을 최대화하고 불가능한 연쇄의 발생 확률을 최소화하여 출력할 수 있도록 할당된다.

(4) 음소배열제약의 학습 알고리즘(Hayes and Wilson 2008: (10))

ㄱ. 입력형
  . 자질 집합에 의해 분류된 분절음들의 집합
  . 가능한 모든 표상의 집합
  . 정확도 수준(accuracy)의 집합
  . 최대 결합할 수 있는 자질 매트릭스의 수

ㄴ. 1단계: 가능한 모든 제약 집합을 탐색 범위로 생성하고 정확도가 일정 수준 이하인 제약 가운데 일반적인 제약 선정

ㄷ. 2단계: 선택된 제약에 대한 가중치 부여[3]

ㄹ. 일정 수준의 정확도에 이를 때까지 반복적으로 학습

---

[1] 실제 학습에서는 가능한 모든 표상의 집합을 구성하기 어렵다. 이 때문에 가능한 모든 표상을 임의적으로 추출하여 가상의 어휘부를 구성한다. 기대 빈도를 계산하는 구체적인 기제는 Hayes and Wilson (2008: (8))과 Eisner (2001, 2002)를 참고할 수 있다.

[2] 직관적으로 제약의 위배빈도는 음소 연쇄의 출현빈도에 대응한다. 이에 따라 정확도는 예측보다 제한되는 음소 결합을 제약으로 선택하는 기준으로 이해될 수 있다.

[3] Hayes and Wilson (2008)은 최적의 가중치를 찾기 위해 관찰 빈도와 기대 빈도의 차이(O[Ci]-E[Ci])에 바탕을 두고 반복적인 hill-climing search 방법을 채택한다. 구체적인 기계학습방법은 Hayes and Wilson (2008: 385-389)에서 소개한다.

특정 음소 연쇄의 확률은 위배되는 제약의 가중치를 더한 것으로부터 계산된다. 이 더한 값, 즉, 비적형성 점수(score)가 음의 지수로 변환되어 최대 엔트로피 값(maxent value)으로 출력된다. 이 값은 출현 확률에 대응되며, 상대적인 적형성을 가리킨다. 예를 들어 (5)에서 '자음-모음(CV)'은 위배하는 제약이 없기 때문에 비적형성 점수는 0이며, 최대 엔트로피 값은 1이다. 즉, CV는 항상 출현하며, 적형성이 매우 높다는 것을 의미한다. 반면, '자음-자음- 모음(CCV)'은 *[+단어 경계][+자음성][+자음성]의 제약을 위배하기 때문에 비적형성 점수가 3이며 이에 따른 최대 엔트로피 값은 0.05에 가깝다. 또한 '자음-자음-모음-모음(CCVV)' 인 경우, *[+단어 경계][+자음성][+자음성]과 *[+성절성][+성절성]을 모두 위배하기 때문에 비적형성 점수는 5이며, 이에 따른 최대 엔트로피 값은 0.006에 가까워 출현확률 및 적형성이 매우 낮을 것으로 보인다. 이처럼 가중치가 큰 제약일수록 제약을 위배하는 연쇄에 대한 비적형성 점수를 높여 해당 연쇄의 출현을 강하게 제한한다.

(5) 학습 문법이 특정 연쇄의 확률을 할당하는 예

ㄱ. 제약 및 제약의 가중치

① *[+단어 경계][+자음성][+자음성] (*#CC) 가중치: 3

② *[+성절성][+성절성] (*VV) 가중치: 2

ㄴ. 제약의 위배 및 최대 엔트로피값 계산

	*#CC	*VV	비적형성 점수 (score)	최대 엔트로피 값
	3	2		
CV	0	0	(0x0)+(0x0)=0	exp(-0) = 1
CCV	1	0	(3x1)+(2x0)=3	exp(-3)≒0.05
CCVV	1	1	(3x1)+(2x1)=5	exp(-5)≒0.006

3. 학습 자료 및 학습 조건

최대 음소배열제약 모델은 소프트웨어 'UCLA 음소배열제약 학습자(UCLA phonotactic learner; Hayes and Wilson 2008)' 로 구현되었다.[4] 표면형에 해당하는 어휘 목록과 자연 부류인 자질 목록이 이 학습자에 입력되면, 가중치가 부여된 제약이 출력된다. 이 절에서는 본 학습에서 입력한 자료와 학습 조건을 다룬다.

3.1 어휘 목록

본 학습에서 입력한 어휘 목록은 강범모·김흥규(2009)의 '한국어 사용빈도: 1500만어절 세종 형태의미분석말뭉치 기반'에서 선정하였다. '일반명사(NNG)' 중 빈도가 5이상인 단일어 및 복합어를 대상으로

[4] http://www.linguistics.ucla.edu/people/hayes/Phonotactics

삼았다.[5] 본 연구에서는 차용어 및 혼종 단어를 제외하고 고유어와 한자어를 선택하였다.

한자어인 경우, 현재 한자 독음과 대응하는 단어만을 선정하였다(송기중 1992, 박선우 2006). 어원은 한자어에서 비롯되었으나, 소리 등이 바뀌어 현재 한자에 대응되지 않는 단어는 고유어로 취급하였다. 예를 들어 '우엉'과 같은 경우, 표준국어대사전에서 '우웡<牛蒡'과 같이 한자 기원을 밝히고 있다. 그러나 '웡'은 현재 한자 독음에 직접 대응되지 않는 한편, '우'도 '牛'의 의미에 부합하지 않기 때문에 본 연구는 고유어로 분류한다. 이에 따라 고유어 5,543 단어, 한자어 29,869 단어를 선정하였다.

선정된 단어의 입력형태는 표준국어대사전[6]의 발음정보를 참고하였다.[7] 자음 앞 종성에 대해 종성 중화 규칙을 적용하여 실제 음성형태를 입력형으로 삼았다. 다만, 어말의 종성은 모음으로 시작하는 조사와 결합할 때 실현될 수 있는 바, 음운론적으로 가정되는 기저형을 입력형으로 채택하였다.

3.2 자질 목록

한국어 자모음 음소에 대하여, 각각 자음성 및 모음성 자질을 구분하여 명세하였다. 전방성(anterior) 자질은 설정음(coronal)에 대해서만 명세하고 그 외의 자음에 대해 미명세한 점을 제외하면 완전명세 방식을 따랐다. 활음에 대해서는 모음성 자질도 명세하였다. 그리고 신지영·차재은(2003)에 따라, 이중모음 '의'를 상향이중모음 활음 /ɯi/로 보아 단모음 /ɨ/와 이중모음 /ɯi/의 분포 차이를 반영하고자 하였다. 8모음 체계를 기준으로 하여, /e/와 /ɛ/의 구분을 유지하였고 '외'와 '위'를 이중모음 /we/, /wi/로 입력하였다.

---

[5] 본고는 단일어와 복합어를 구분하지 않고 일정 빈도 이상의 단어를 입력형으로 가정하고, 이로부터 음소배열제약을 학습하였다. '복합어'도 하나의 단어처럼 저장될 수 있다는 입장(예. Bybee 2001)을 참고한 것이다. Hayes and White (2013)는 영어 음소배열제약의 학습 대상으로 단일어만을 삼았으나, 실험 과정에서 복합어의 음소배열제약이 단일어 판단에 영향을 주는 점도 부분적으로 관찰하였다. 물론 Martin (2011)이 지적한 바와 같이 형태소 내부의 음소배열제약이 복합어에서는 다소 느슨하게 드러나기 때문에, 추후에 단일어만을 대상으로 학습한 이후 결과 비교가 이루어져야 할 것이다.

[6] 국립국어원 www.korean.go.kr

[7] UCLA 음소배열제약 학습자는 한글을 인식하지 못하기 때문에, 선정된 단어를 예일 로마자로 변환하고 음운규칙을 적용하였다. 이에 대한 R 스크립트와 Python 스크립트를 제공해 준 박수지와 이상아에게 감사한다.

(6) 자음의 자질 목록 (공통자질: [-성절성, +자음성])

	공명	지속	비음	전방	기식	긴장	조찰	양순	설정	연구개
p	-	-	-	0	-	-	-	+	-	-
p^h	-	-	-	0	+	-	-	+	-	-
p’	-	-	-	0	-	+	-	+	-	-
t	-	-	-	+	-	-	-	-	+	-
t^h	-	-	-	+	+	-	-	-	+	-
t’	-	-	-	+	-	+	-	-	+	-
c	-	-	-	-	-	-	+	-	+	-
c^h	-	-	-	-	+	-	+	-	+	-
c’	-	-	-	-	-	+	+	-	+	-
k	-	-	-	0	-	-	-	-	-	+
k^h	-	-	-	0	+	-	-	-	-	+
k’	-	-	-	0	-	+	-	-	-	+
s	-	+	-	+	-	-	+	-	+	-
s’	-	+	-	+	-	+	+	-	+	-
h	-	+	-	0	+	-	-	-	-	-
m	+	-	+	0	-	-	-	+	-	-
n	+	-	+	+	-	-	-	-	+	-
ŋ	+	-	+	0	-	-	-	-	-	+
l	+	-	-	+	-	-	-	-	+	-

(7) 모음의 자질 목록

	성절	자음	공명	고설	저설	후설	원순
j	-	-	+	+	-	-	-
w	-	-	+	+	-	+	+
ɯ	-	-	+	+	-	+	-
i	+	-	+	+	-	-	-
e	+	-	+	-	-	-	-
ɛ	+	-	+	-	+	-	-
ɨ	+	-	+	+	-	+	-
ʌ	+	-	+	-	-	+	-
a	+	-	+	-	+	+	-
o	+	-	+	-	-	+	+
u	+	-	+	+	-	+	+

UCLA 음소배열제약 학습자는 위 자질 외에 자동적으로 단어 경계 자질([+/-word_boundary])을 포함하여 제약을 학습한다.

3.3 학습 조건

*[+단어 경계][+자음성][+자음성]과 같이 단어 경계에서 제한되는 음소 연쇄까지 포착하기 위해서, 조합할 수 있는 자질 매트릭스의 수(the number of feature matrix)는 3으로 지정하였다. 정확도(O/E)의 한계치(threshold)는 0.3으로 정하여 위배 가능하지만, 제한적으로 분포하는 음소 결합을 포착하고자 했다.[8] 추가적으로, 기존 연구에서 언급한 빈칸(gap)을 모두 포착하기 위해, 시그마(sigma)는 1.2로 설정하였다.[9] 최대 학습할 수 있는 제약의 수는 100개이며, 고유어 어휘부와 한자어 어휘부 각각에 대해 5회 반복하여 학습한다.

4. 결과

이 절은 고유어 어휘부와 한자어 어휘부 각각에 대한 문법을 제시한다. 제약과 그 가중치를 살펴보고, 각 어휘부에 분포하는 음소 결합 관계를 파악한다. 가중치는 제약이 음소 결합을 금지하는 강도를 의미하며, 개별 어휘부의 문법 내에서 상대적인 적형성을 예측한다.[10]

앞서 2절에서 언급한 바와 같이 제약 위배의 기대 빈도(E[Ci])가 임의적으로 계산되기 때문에 매 학습마다 제약 및 가중치가 다소 다르게 출력된다. 본고는 5회 학습 중 3회 이상 출력된 제약과 그 가중치 평균을 학습 결과로 삼는다.

본고는 고유어(N)와 한자어(S) 어휘부에만 해당하는 제약을 중심으로 살펴 본다. 우선 범주적 음소배열제약(categorical phonotactics)을 제시하고, 비범주적 음소배열제약(gradient phonotactics)을 논의하겠다.

4.1 범주적 음소배열제약

우선, 고유어와 한자어에서 출현하지 않는 음소 연쇄를 보고한다. 해당 연쇄 모두 출현빈도가 0이지만, 제약의 기대 빈도에 따라 예측되는 출현 확률이 상이하다. 이에 따라 각 제약은 다른 가중치를 가지며, 연속적인 문법 인식을 예측한다.

---

[8] Hayes and Wilson (2008)과 Cho (2012)의 정확도 수준을 따른 것이다.

[9] 시그마(sigma)의 초기값은 1이며, 값이 커질수록 입력 자료의 패턴에 적합하여 예외없는 제약이 비중있게 학습된다. 본고는 임의로 1.2를 설정하여 기존에 언급된 제약을 모두 포착하고자 했다.

[10] 고유어 문법과 한자어 문법은 독립적으로 학습되었으므로, 가중치는 해당 문법 상에서만 상대적인 의미를 보인다. 따라서, 고유어의 가중치와 한자어의 가중치를 상호 비교할 수 없다.

4.1.1 고유어

고유어에서는 네 제약만이 예외를 허용하지 않는 범주적 제약의 자격이 있다. 가장 가중치가 높은 제약(N1)은 어말 위치에서 [e]와 [ɛ] 뒤에 장애음이 오지 못하는 것이다. 고유어 문법에서 N1을 위배하는 '객(客), 앱(App)' 과 같은 단어는 적형성이 매우 낮을 것으로 예측된다.

그 다음으로 높은 가중치가 할당된 제약은 어말 [ɨ]를 허용하지 않는 것이며, [cʰɨ]를 제한하는 제약은 이보다 낮은 가중치를 보인다. N4는 [e]에 후행하는 원순 모음 [u] 또는 w 계 이중 모음을 제한하며, 상대적으로 낮은 가중치로 출력된다. N4만을 위배하는 [eu] 연쇄는 N1-N3을 위배하는 연쇄보다 적형성이 다소 높을 것으로 예측된다.

표 1. 고유어의 범주적 음소배열제약(4개)

	제약	가중치	예	cf. 비고유어명사
N1	*[-고설,-후설][-공명]#	4.60	*ɛk#	객 [k**ɛk**]
N2	*[-고설,-후설,-원순]#	3.93	*ɨ#	버그 [pʌk**ɨ**]
N3	*[-전방,+기식][+고설,+후설,-원순]	3.21	*cʰɨ	측량 [**cʰɨ**ŋnjaŋ]
N4	*[-고설,-저설,-후설][+고설,+원순]	2.79	*eu	에우다 [**eu**ta]

N1-N4의 제약은 기존연구에서 관찰된 범주적 음소배열제약을 대부분 포함한다. N2는 [ɨ]로 끝나는 고유어와 한자어가 없다는 기술(강용순 1998)과 일치한다. N3은 Cho (2012)에서도 학습된 바 있으며, 경구개음 [c, c', cʰ] 다음에 [ɨ]가 회피된다는 기술(진남택 1992)과 부분적으로 일치한다.

한편, 본 연구가 자료를 명사로 제한하였기 때문에, 기존연구에서 분명하게 언급되지 않은 N1과 N4를 관찰할 수 있었다. 신지영·차재은(2003)은 한국어에서 출현 가능한 연쇄를 제시하면서, 용언 어간 및 의성어에서 [ɛ] 뒤에 장애음이 출현한다는 점(예: '맵-', '빽빽')과 동사 어간인 경우 [eu-]가 출현한다는 점(예: '에우-')을 언급하였다.

4.1.2 한자어

한자어인 경우, 고유어에 비해 다수의 범주적 음소배열제약이 학습되었다. 우선 가중치 상위 5위까지의 제약을 살펴보면, 어두 및 어말 음소 제약이 상대적으로 높은 가중치를 보인다. 어말 제약인 S1-S4가 [ɨ], 격음, 자음군 그리고 경음의 발생을 강하게 제한하며, S5가 어두 [e]의 출현을 저지한다.

표 2. 한자어의 범주적 음소배열제약 (상위 5개)

	제약	가중치	예	cf. 비한자어
S1	*[-고설,-후설,-원순]#	6.18	*i#	버스 [bus**i**]
S2	*[+기식]#	6.14	*c^{h}#	꽃 [k'o**c^{h}**]
S3	*[-성절성][-성절성]#	6.02	*ps#	값 [kaps]
S4	*[+긴장]#	5.93	*k'#	밖 [pa**k'**]
S5	*#[-고설,-저설,-후설]	5.75	*#e	에누리 [**e**nuli]

한자어에 출현하지 않는 음소 및 음절이 단어 내에서 세부적인 제약으로 학습되며, 해당 제약에 높은 가중치가 할당된다. S6-S8은 한자어에서 [jɛ], 어중 설정음 종성, 그리고 모음 앞 [i]로 끝나는 음절을 금지한다.

표 3. 특정 음소 및 음절에 관한 제약

	제약	가중치	예	cf. 비한자어
S6	*[-원순,-성절][+저설,-후설]	5.34	*jɛ	얘기 [**jɛ**ki]
S7	*[-공명,-양순,-연구개][+자음]	5.12	*tk'	젓갈 [jʌtk'al]
S8	*[+고설,+후설,-원순,+성절][-자음]	4.48	*ii	쓰이다 [s'**ii**ta]

다수의 연구는 한자 음절의 종성 위치에 격음과 경음이 허용되지 않고, 독음이 '으, 에, 얘'인 한자가 없다고 지적하였다(권인한 1997, 신지영·차재은 2003, 신지영 2009, 안소진 2009).

표 2와 표 3의 제약은 기존연구가 관찰한 한자어의 음소 분포를 대부분 포착한다. 나아가 각 제약의 가중치는 발생하는 어휘 예가 전혀 없는 음소 연쇄에 대해서도 적형성의 차이를 예측할 수 있다. 예를 들어, S1(가중치: 6.18)을 위배하는 어말 [i]가 S8(가중치: 4.48)을 위배하는 [i]+[모음]보다 적형성이 낮다고 볼 수 있다.

또한, 표 4의 제약은 [공명음]+[e, ʌ]가 어두와 자음 뒤에 오는 것을 막는다. 이 제약은 한자어의 구성 음절단위로 독음이 '너, 머'인 한자가 없다는 관찰과 이를 일반화한 제약 '*/ㄴ, ㄹ, ㅁ/(공명음) + /ㅓ, ㅟ, w/'(신지영 2009)보다 구체적으로 출현양상을 포착한다.

표 4. [공명음] + [e, ʌ] 제약

	제약	가중치	예	cf. 비한자어
S9	*#[+자음,+공명][-고설,-저설,-원순]	4.44	*#nʌ,	너구리 [**nʌ**kuli]
S10	*[-성절][+자음,+공명][-고설,-저설,-원순]	3.94	*lmʌ	할머니 [ha**lmʌ**ni]

한편, [k^{h}]에 대한 제한이 세부적인 제약으로 학습된다. 이제까지는 초성이 'ㅋ'인 한자 음절은 '쾌'밖에 없다고 언급되었다(권인한 1997,

강용순 1998, 신지영·차재은 2003, 신지영 2009). 그러나 단어의 발음을 고려하면, 격음화 규칙으로 인해 모음 뒤에서는 [k^h]이 출현할 수 있다(예: 국화 [ku.k^hwa]). 본 시뮬레이션 결과 [k^hi, k^he]의 출현을 막는 S11이 학습되는 한편, 설정음(coronal)이 아닌 자음 뒤에 [k^h]를 저지하는 S12가 학습되었다.

표 5. [k^h]에 관한 제약

	제약	가중치	예	cf. 비한자어
S11	*[-지속,+기식,-조찰,-양순] [-저설,-후설,+성절]	4.89	*k^hi	키 [**k^h**i]
S12	*[-설정][+기식,+연구개]	4.27	*mk^h	암캐 [a**mk^h**ɛ]

설정음과 모음의 결합도 표 6에서 보듯이 매우 제한된다. 설정음과 j, ɯ계 이중모음의 결합이 회피되는 것은 한국어 음운론 일반에서 관찰되었으며(허웅 1985, 진남택 1992, 신지영·차재은 2003), 본 한자어 문법에서도 높은 가중치의 제약 S13으로 출력된다. 또한, 치경음(alveolar)과 전설 모음 [i, e]의 결합도 회피된다. 어두 제약 S14는 공명 치경음과의 결합을 추가적으로 제한하는 한편, S15는 장애 치경음이 포함된 연쇄만을 금지한다. 이와 같은 제약은 한자어의 [ti]와 같은 연쇄가 근대 국어의 구개음화를 예외없이 거쳤다는 것(박선우 2006)을 포착한다고 볼 수 있다.

표 6. [설정음]+[모음] 제약

	제약	가중치	예	cf. 비한자어
S13	*[-공명,+설정][-원순,-성절]	4.99	*tj	디뎌 [ti**tj**ʌ]
S14	*#[-지속,+전방][-저설,-후설,+성절]	4.15	*#ne	네모 [**ne**mo]
S15	*[-공명,-지속,+전방][-저설,-후설]	3.72	*t'e	껍데기 [k'ʌpt'eki]

한편, 한자어 어두에서 양순음 뒤에 w계 이중모음이 제한되는 것은 한국어 일반의 제약으로 지적되었으나(허웅 1985, 진남택 1992, 신지영·차재은 2003), 한자어에 대해서는 어두 제약 S16으로만 학습된 것이 특징적이다.

표 7. 어두 위치: [양순음]+[활음] 제약

	제약	가중치	예	cf. 비한자어
S16	*#[+양순][+후설,-성절]	4.18	*#pw	봐 [**pw**a]

어말 위치에서도 다수의 연쇄가 제한된다. S17은 [e, ɛ] 뒤에 연구개음 외 자음을 강하게 제한하는 한편, S18은 원순 모음 [o, u] 뒤에

치경음 및 양순음 출현을 막는다. ‘앱(App), 뱀, 굽, 톱’과 같은 비한자어는 S17과 S18을 위배하기 때문에 ‘한자어’로서의 적형성이 낮을 것으로 보인다. 또한, S17의 가중치가 S18보다 높기 때문에 S17을 위배하는 ‘앱, 뱀’과 같은 단어의 적형성이 S18을 위배하는 단어의 적형성보다 낮을 것으로 예측된다. 이 외에도 [wʌ]로 끝나는 한자어와 [t, t’, tʰ]+[i, e, ʌ]로 끝나는 한자어를 허용하지 않는 제약이 출력되었다.

표 8. 어말 위치: 두 음소 결합 제약

	제약	가중치	예	cf. 비한자어
S17	*[-고설, -후설][-연구개]#	5.53	*ɛm#	뱀 [p**ɛm**]
S18	*[+원순][-공명,-조찰,-연구개]#	4.76	*up#	굽 [k**up**]
S19	*[+후설,-성절][-저설,+후설]#	4.16	*wʌ#	타워 [tʰ**wʌ**]
S20	*[-공명,-지속,+전방][-저설,-원순]#	4.05	*tʰʌ#	집터 [cipt**ʰʌ**]

어말 제약 S17-18은 [ɛ, o, u] 뒤에 치경음 및 순음이 결합되지 못한다는 관찰과 부분적으로 일치한다(강용순 1998, 신지영·차재은 2003, 신지영 2009). 다만, 본고에서는 입력 자료에 ‘품’과 결합된 단어가 다수 있는 바, 기존 기술과 달리 [um, om]의 연쇄는 제한되지 않았다. 한편, S20은 한자 ‘터(攄)’가 실재함에도 [tʰʌ]로 끝나는 한자어가 없다는 점을 새롭게 찾은 것이다.

이제까지 범주적 음소배열제약에 대한 학습 결과를 살펴보았다. 고유어인 경우, 어말에 위치한 [ɛp], [ɛk] 등이 가장 강하게 제한될 것으로 예측된 반면, [eu, ew]는 다소 가능한 연쇄로 판단될 수 있다. 한자어인 경우, 어말 [ɨ], 격음, 경음이 매우 제한되는 반면, 어말 [tʰʌ]는 상대적으로 약하게 제한될 것이 예측된다. 이러한 제약을 바탕으로, 화자들이 출현 빈도가 0인 연쇄를 포함한 새로운 단어에 대해서도 해당 어휘부 단어로서의 적형성을 판단할 수 있을 것으로 예측된다.

4.2 비범주적 음소배열제약

이 절은 출현 빈도가 유의미하게 낮은 음소 결합 관계를 다룬다. 기존연구에서 예외가 언급된 음소배열제약을 확인하는 한편, 새로운 제약을 탐색하고 그 강도를 예측한다. 이를 통해 음소 연쇄의 출현 빈도로만은 예측하기 어려운 제약 간의 상대적인 적형성을 논의한다.

4.2.1 고유어

고유어 어휘부에 대해서는 다수의 비범주적 음소배열제약이 포착

된다. 우선, 소수의 예외만을 허용하는 제약이 높은 가중치로 학습되었다. 어말 경음을 제한하는 N5가 가장 높은 가중치를 할당받았다. 다음 순위로 가중치가 높은 N6과 N8은 각각 어두 [e, ɛ]와 어두 [ɨ]를 허용하지 않는다. N5를 위배하는 '밖'은 N6, N8을 위배하는 '에누리, 으뜸'보다 '고유어'로서의 적형성이 낮을 것으로 예측할 수 있다. 이에 더하여 파생어 '기쁨, 아픔'을 제외하고는 [양순 장애음]+[ɨ]가 출현하지 못한다는 점이 제약 N9로 학습되었다.

한편, 제약 N7은 비원순 후설모음 다음에 중모음 및 저모음이 회피되는 것을 새롭게 포착하였다. 모음 연쇄에 대한 제약 중 가장 가중치가 높은 제약이며, '어안 [ʌan]'과 '가오리 [kaoli]'만을 예외로 허용한다.

표 9. 고유어의 비범주적 음소배열제약 (상위 5개)

	제약	가중치	예	예외
N5	*[+긴장]#	4.44	*k'#	밖 [pa**k'**]
N6	*#[-고설,-후설]	3.54	*#e	에누리 [**e**nuli]
N7	*[+후설,-원순][-고설]	3.38	*ʌa	가오리 [k**ao**li]
N8	*#[+고설,+후설,-원순]	3.10	*#ɨ	으뜸 [**ɨ**t'im]
N9	*[-공명,+양순][+고설,+후설,-원순]	3.04	*p'ɨ	기쁨 [ki**p'ɨ**m]

한성우(2006)가 보고한 바에 따르면, 고유어 어말에서 가장 빈도가 낮은 음소는 경음 [k']이며, 어두에서도 모음 [e, ɛ, ɨ]의 빈도가 상대적으로 낮다. 이와 같이, 유의미하게 낮은 출현 빈도가 계산되어 각각 N5, N6, N8이 학습된 것이다. N9는 한국어 음운론에서 [양순음]+[ɨ]를 회피한다는 기술(허웅 1985, 진남택 1992, 신지영·차재은 2003)과 부분적으로 일치한다.

경음 외에도 어말에서 [t, c] 및 [k^h]가 회피되는 제약이 학습되었다. 제약 N10과 N11은 어말 경음 제약 N5에 비해 상대적으로 가중치가 낮아, 경음보다는 [t, c, k^h]가 어말에서 제한되지 않을 것이 예측된다. 그러나 어말 [t]는 *[t, c]# (N10)과 함께 범주적 음소배열제약 *[t]#을 위배하여, 비적형성 점수 4.98을 부여받는다. 이 값은 어말 경음에 대한 비적형성 점수 4.44보다 다소 높은 것이며, 어말 [t]의 출현이 강하게 억제되는 것을 나타낸다.

표 10. 어말 자음 제약

	제약	가중치	예	예외
N10	*[-공명,-지속,-기식,+설정]#	3.00	*t#, *c#	낮 [na**c**]
N11	*[+기식,-양순,-설정]#	2.72	*k^h#	부엌 [puʌ**k^h**]
cf.	*[-공명,-지속,+전방,-기식]#	1.98	*t#	-

한편, [경음, 격음]+[모음] 제약이 어말 음소 제약보다 낮은 가중치로 학습되었다. 우선, [k^hu, hu]의 출현이 자유롭지 못하고(N12), [p^he, p^hʌ, k^he, k^hʌ]가 선호되지 않는다(N13). 이 두 제약은 Cho (2012)에서도 보고한 제약이다. Cho (2012)는 N12에 대해 [k^hu]가 [p^hu]로 혼동되는 음성학적 동기를 지적하였다. 반면, 격음 [p^h, k^h] 뒤에 [e, ʌ]가 오지 못하는 이유는 음성학적으로 해석될 수 없다고 보았다.

추가적으로, N14와 N15는 연구개음을 제외한 경음과 격음이 [e]와 결합하는 것을 억제한다. 이 중 N15는 N13과 함께 [p^he]를 금지한다. 두 제약의 가중치를 더하면 [p^he]에 대한 비적형성 점수는 5.7로, N13만을 위배하는 [k^he], N15만을 위배하는 [t^he, he]보다 고유어로서의 적형성이 매우 낮을 것으로 예측된다. [e]가 포함된 다른 단어를 고려하면, [p^he]는 어두 제약 N6(*#[e, ɛ], 가중치: 3.54)을 위배하는 단어(예: 에누리 [enuli])보다 적형성이 낮을 것이 예측되는 한편, [k^he, t^he, he]보다는 적형성이 높을 것으로 예측된다.

그 외에 N16(*[c'ɨ])은 범주적 제약인 N3(*[c^hɨ], 가중치: 3.21)과 음성적 동기가 같지만, 가중치가 다소 낮으며 예외를 허용한다.

표 11. [경음, 격음]+[모음] 제약

	제약	가중치	예	예외
N12	*[+기식,-양순,-설정] [+고설,+원순,+성절]	2.93	*k^hu, *hu	넝쿨 [nʌŋ**k^hu**l]
N13	*[-지속,+기식,-설정] [-고설,-저설,-원순]	2.87	*p^hʌ, *k^he	올케 [ol**k^he**]
N14	*[+긴장,-연구개] [-고설,-저설,-후설]	2.87	*s'e, *t'e	곱셈 [kop**s'e**m]
N15	*[+기식,-연구개] [-고설,-저설,-후설]	2.83	*t^h e	테 [**t^he**]
N16	*[-전방,+긴장] [+고설,+후설,-원순]	2.60	*c'ɨ	그쯤 [ki**c'ɨ**m]

위의 제약 중 일부는 한자어와 분명히 구별되는 고유어의 음소 결합 관계를 포착한다. N12에 의해 회피되는 [k^hu, hu]는 한자어(예: 낙후 [nak^{h}u], 후보 [hupo])에서 자유롭게 출현할 수 있다. 또한, 한자어 다수가 포함하는 [s'e, c'e, c^he]는 N14 및 N15에 의해 고유어에서 선호되지 않는다. 아울러 [c'ɨ]는 '즙, 증'과 결합하는 한자어에서 관찰되지만, 고유어에서는 N16에 의해 출현이 제한된다.

어말에서만 금지되는 [자음]+[모음] 연쇄도 포착되었다. 어말에서는 [p, p^h, k, k^h, h]와 [i, ʌ, a]가 결합되지 못하는 한편(N17), 경음 뒤에 후설 고모음 및 중모음도 제한된다(N18). 고유어 어말에서 제한되는 [ka, p^ha, s'u, t'o]등은 한자어 어말에서 자유롭게 출현할 수 있다.

표 12. 어말 위치: 두 음소 결합 제약

	제약	가중치	예	예외
N17	*[-공명,-긴장,-설정][+후설,-원순]#	2.68	*ka#	아가 [a**ka**]
N18	*[+긴장][-저설,+후설]#	2.62	*k'u#	애꾸 [ɛ**k'u**]

자음과 활음의 결합 제약(허웅 1985, 진남택 1992, 신지영·차재은 2003)은 표 13과 같이 확인된다. N19는 [양순음]+[w, ɯ]를 금지하고, [mwe]를 포함한 4단어가 이 제약을 위배한다. 또한, [설정 장애음]+[ja, jʌ, jo, ju, wa, wʌ]은 N20에 의해 제한되며, '똬리 [t'wali]'만이 예외로 출현한다. 이에 더하여, N21은 설정 장애음 및 [h]와 [j, ɯ]의 결합을 제한한다. [설정 장애음]+[j, ɯ]는 N20과 N21을 모두 위배하여 4.87의 높은 비적형성 점수를 받아, [hj]보다 더 강하게 제한된다는 것을 보인다.

표 13. [자음]+[활음] 제약

	제약	가중치	예	예외
N19	*[+양순][+후설,-성절]	2.91	*mw	뫼 [**mw**e]
N20	*[-공명,+설정][+공명,-성절][+후설]	2.71	*t'wa	똬리 [**t'w**ali]
N21	*[-공명,-양순,-연구개][-원순,-성절]	2.16	*tj, *hj	혀 [**hj**ʌ]

장애음에 이어 격음 [p^h, t^h, k^h]이 오지 못하는 제약도 다소 낮은 가중치로 학습되었다. 이 제약만을 위배하는 [p, t, k]+[p^h, t^h, k^h]는 '쪽파'와 같은 예외를 허용하며 다소 약한 회피를 보인다.[11] 그러나 제약을 위배하지 않는 [p, t, k]+[c^h] 보다 적형성이 낮을 것으로 예측할 수 있으며, 이는 Cho (2012: 351, (7))의 예측과도 일치한다.

표 14. [장애음]+[격음] 제약

	제약	가중치	예	예외
N22	*[-공명][-긴장,-조찰]	2.15	*pt^h, *kp^h	쪽파 [c'o**kp^h**a]

한편, N7 *[+후설,-원순][-고설](가중치: 3.38) 외에도 세부적인 모음 연쇄 제약이 학습되었다. 우선, 모음과 후설모음이 어말 위치에서 이웃하지 않는 제약 N23이 포착되었다.

[11] N22는 장애음에 이어 조찰음이 아닌 평음 및 격음이 나타나는 것을 제한한다. 이 제한되는 연쇄 중 본고는 [p, t, k][p^h, t^h, k^h]에 중점을 두어 논의한다. 그 외 연쇄는 한국어 음운론에서 필수적인 *[-공명][-기식,-긴장][-단어 경계], *[+기식][+자음], *[+긴장][+자음], *[+지속][+자음]에 의해 이미 포착되었다.

표 15. 어말 위치: [모음]+[모음] 제약

	제약	가중치	예	예외
N23	*[+성절][+후설]#	2.95	*io#, *ua#	부아 [p**ua**]

이에 더하여, 후설 고모음과 원순 모음의 결합이 회피된다(N24). N25는 고모음 및 중모음 뒤에 평순 모음과 j, ɯ계 이중모음을 금지한다. N26은 두 원순 모음의 연쇄 또는 원순 모음과 w 계 이중 모음의 연쇄를 허용하지 않는다.

표 16. [모음]+[모음] 제약

	제약	가중치	예	예외
N24	*[+고설,+후설][+원순,+성절]	2.80	*ʉo, *uu,	어두움 [ʌt**uu**m]
N25	*[-저설,+성절][-원순]	2.48	*oi, *eʌ	모이 [m**oi**]
N26	*[+원순][+원순]	2.30	*oo, *ou	도우미 [t**ou**mi]

위 제약의 가중치를 더하여, 모음 연쇄 간의 강도를 예측할 수 있다. 예를 들어, N7 및 표 16의 제약만을 살펴 보았을 때 가장 비적형성 점수가 높은 [ʉo]는 고유어에서 매우 제한되리라 본다. [ʌa], [uo]도 각각 두 개의 제약을 위배하여, 적형성이 상대적으로 낮을 것으로 보인다.

표 17. 모음 연쇄에 대한 적형성 평가 예

제약 / 연쇄	*[+후설,-원순] [-고설]	*[+고설,+후설] [+원순,+성절]	*[-저설,+성절] [-원순]	*[+원순] [+원순]	비적형성 점수
	3.38	2.8	2.48	2.3	
ʉo	1	1			6.18
ʌa	1		1		5.86
uo		1		1	5.1

학습된 제약을 종합하여, 고유어의 모음충돌 회피 경향을 구체적으로 분석할 수 있다. N4, N7 및 표 16의 제약만을 고려하면, 두 모음이 이웃하는 64연쇄(8모음×8모음) 중 15연쇄만이 고유어에서 자유롭게 결합할 수 있을 것으로 예측한다. 이 중 N23이 어말 위치에서 [모음]+[후설모음]이 출현하는 것을 금지하므로, [ɛi, ɛe, ɛɛ, ai]만이 자유롭게 나타날 것으로 예측된다. 그리고 앞서 살핀 어두 모음 제약 N6(*#[e, ɛ])까지 고려하면, [ai]의 출현만이 억제되지 않을 것으로 보인다.

고유어의 두 모음 연쇄의 결합 관계를 살펴보면 표 18과 같다. 음영 부분은 N4, N7 및 표 16의 제약이 금지하는 모음 연쇄를 나타낸다. 그 외 모음 연쇄 중 어두 제약 N6과 어말 제약 N23가 제한

하는 연쇄를 각각 '/'와 '\'로 표시한다. 이를 바탕으로, 고유어에서 모음 연쇄가 회피된다는 이전의 언급(유재원 1997, 하세경 2000)을 구체적으로 확인할 수 있다.[12]

표 18. 두 모음 연쇄의 결합 관계

V1 \ V2	i	e	ɛ	ɨ	ʌ	a	u	o
i							\	\
e								x
ɛ	/	/	/	x	x	x	x	x
ɨ								
ʌ							\	
a				\			\	
u								
o								

/: N6 위배
\: N23위배
x: N6과 N23 위배

이에 더하여 세 [-자음성]의 결합이 회피된다. N27은 [비원순모음]+[후설모음]+[비고모음], [ju, jʌ, jo, ja]+[비고모음], 그리고 [비원순모음]+[we, wɛ, wʌ, wa] 등의 결합을 다소 낮은 강도로 제한한다. 한편, N28은 세 모음 연쇄 또는 두 모음 연쇄와 활음의 결합이 회피되는 것을 포착하였다.

표 19. [-자음성]+[-자음성]+[-자음성] 제약

	제약	가중치	예	예외
N27	*[-원순][+후설][-고설]	2.36	*ʌwa	너와 [n**ʌwa**]
N28	*[+성절][+성절][-자음]	2.32	*VVV	사내아이 [san**gai**]

4.2.2 한자어

한자어 비범주적 음소배열제약은 예외가 매우 적게 출현하며, 범주적인 음소배열제약만큼 가중치가 높다. '곶' 외에 설정음으로 끝나는 단어가 없는 한편(S21), '쌍, 끽'을 포함한 경우 외에는 단어가 경음으로 시작하지 않는다는 것이 포착된다(S22). 또한 세 모음 연

[12] 표 18에 어두 제약인 *#[ö](N8)와 어말 제약인 *[ö]#(N2)를 표시하지 않았다. N24와 N25에 의해 [ö]로 시작하는 모음 연쇄가 모두 제한되는 한편, N23에 의해 [ö]로 끝나는 연쇄가 제한된다.

쇄를 저지하는 제약 S23이 여섯 단어만을 예외로 허용하고, S24가 [s, s', c, c', c^h]+[e]를 제외한 [자음]+[e]의 결합을 막는다. 그리고 S25에 의해 어두에서 모음과 후설모음의 결합이 회피된다.

표 20. 한자어의 비범주적 음소배열제약(상위 5개)

	제약	가중치	예	예외
S21	*[-공명,-양순,-연구개]#	6.16	*s#, *c#	곶(串) [ko**c**]
S22	*#[+긴장]	5.63	*#k', *#s'	쌍 [**s'**aŋ]
S23	*[+성절][+성절][-자음]	5.63	*VVV	고아원 [k**oaw**ʌn]
S24	*[-조찰][-고설,-저설,-후설]	5.60	*ke, *k'e	게시 [**ke**si]
S25	*#[+성절][+후설,+성절]	4.95	*#ao, *#ʌʌ	오악 [**oa**k]

기존연구(권인한 1997, 강용순 1998, 신지영 2009, 안소진 2009)가 지적한 한자어의 특징이 높은 가중치가 할당된 제약으로 포착되었다. 한자어의 구성 음절에서 설정 장애음 종성을 허용하지 않는다는 기술이 어말 제약인 S21로 학습되었고, 경음 초성을 제한한다는 관찰이 어두 제약인 S22로 확인되었다.

이에 더하여 본 모델은 출현여부만 기술된 음소 결합관계를 보다 정확하게 파악할 수 있다. 일례로, 기존연구(강용순 1998, 신지영 2009)는 '게'를 '세, 제, 체'와 함께 출현하는 한자로만 보고하였다. 그러나 본 연구에서는 '게'가 포함된 단어가 '세, 제, 체'와 달리 S24를 위배하며 한자어에서 제한된다는 점을 포착할 수 있다.

한편, S23과 S25는 본 모델에서 새롭게 예측한 제약으로 고유어에 대한 결과와 유사하다. [-자음성]의 세 연쇄를 제한하는 S23은 고유어에서도 학습된 바 있다(N28). [모음]+[후설모음]에 대해 고유어인 경우에는 어말 제약 N23으로 학습된 반면, 한자어인 경우에는 어두 제약 S25로 출력되었다.

상위 5개 제약 외에 8개의 비범주적 제약이 학습되었다. 우선, [p^h, t^h, k^h]와 [e, ʌ]가 결합하지 못하는 바가 학습되었다. S26을 위배하는 연쇄 중 [p^he, t^he, k^he]는 상위 제약 S24를 추가적으로 위배하기 때문에 [p^hʌ, t^hʌ, k^hʌ]보다 적형성이 더 낮을 것으로 예측된다. 한편, 한자어 제약 S26과 고유어 제약 N13이 모두 [p^hʌ, k^hʌ]를 제한하지만, [t^hʌ]는 한자어에서만 회피된다.

표 21. [격음]+[모음] 제약

	제약	가중치	예	예외
S26	*[-지속,+기식,-조찰] [-고설,-저설,-원순]	4.37	*t^hʌ, *p^he	터득 [**t^hʌ**tik]

양순음 다음에 [ɨ]가 오지 못한다는 제약 S27도 확인되었다. 다만, 한국어 전반에 대한 기술(허웅 1985, 진남택 1992, 신지영·차재은 2003)과 달리, 장애음에 대해서만 학습되었다.

표 22. [양순음]+[모음] 제약

	제약	가중치	예	예외
S27	*[-공명,+양순] [+고설,+후설,-원순,+성절]	3.81	*p'ɨ	잡음 [ca**pɨ**m]

또한 S28은 양순 장애음과 설정 장애음에 이어 격음 및 평음이 오지 못하도록 금지한다. 이 제약만을 위배하는 [p, t]+[격음]은 비적형성 점수가 1.08로 매우 낮아, 상대적으로 허용될 것으로 예측된다.[13]

표 23. [자음]+[자음] 제약

	제약	가중치	예	예외
S28	*[-공명,-연구개][-긴장]	1.08	*pc^h, *pt^h	법치 [pʌ**pc^h**i]

S25(*#[+성절][+후설,+성절]) 외에도 모음 연쇄에 대한 세부적인 제약이 학습되었다. 우선, 단어 경계에 나타나는 제약을 살펴보면, 어말 위치에서 [i, ɨ, e, ʌ]+[i, ɨ]가 제한된다.

표 24. 어말 위치: [모음]+[모음] 제약

	제약	가중치	예	예외
S29	*[-저설,-원순,+성절] [+고설,-원순]#	4.46	*ei#,*ʌi#	이이제이 [iicei]

또한, S30은 고유어와 마찬가지로 [ɨ, u]+[o, u, w]를 금지한다. S31에 의해 [모음]+[e, ɛ]가 회피되며, S32, S33에 의해 [ɨ, ʌ]에 이어 [i]를 제외한 모음이 오지 못한다.

표 25. [모음]+[모음] 제약

	제약	가중치	예	예외
S30	*[+고설,+후설][+원순,+성절]	4.67	*uu, *io	우울 [**uu**l]
S31	*[+성절][-고설,-후설]	4.65	*ʌɛ, *uɛ	거액 [k**ʌɛ**k]

[13]각주 11에서 언급한 것과 같은 이유로, 다른 음소배열제약(예: [-son][-asp, -tense])으로 포착되는 연쇄를 제외하고 [p, t][격음]에 중점을 두어 제약 S28의 의미를 파악하였다.

S32	*[-저설,+후설,-원순][-고설]	3.95	*ʌa, *ʌʌ	어업 [**ʌʌ**p]
S33	*[-저설,+후설,-원순][+후설]	2.79	*ʌi, *ʌu	어음 [**ʌi**m]

표 25의 제약을 종합하여, 두 모음 연쇄의 적형성을 평가할 수 있다. 세 제약을 위배하는 [ɨo]는 비적형성 점수가 가장 높아, 매우 낮은 적형성을 보일 것으로 예측된다. 두 개의 제약을 위배하는 [ʌɛ], [ɨu]도 다른 모음 연쇄보다 회피될 것으로 보인다.

표 26. 모음 연쇄에 대한 적형성 평가 예

제약 / 연쇄	*[+고설,+후설] [+원순,+성절]	*[+성절] [-고설,-후설]	*[-저설,+후설,-원순] [-고설]	*[-저설,+후설,-원순] [+후설]	비적형성 점수
	4.67	4.65	3.95	2.79	
ɨo	1		1	1	11.41
ʌɛ		1	1		8.6
ɨu	1			1	7.46

학습된 모음 연쇄 제약은 한자어에서 회피되는 두 모음 연쇄를 예측한다. 우선, S8 및 표 25의 제약을 모두 위배하는 연쇄는 두 모음이 이웃하는 64연쇄(8모음X8모음) 중 29연쇄이다. 표 27의 음영 부분에서 알 수 있듯이, 첫번째 모음이 [ɨ, ʌ]인 모음 연쇄와 두번째 모음이 [e, ɛ]인 모음 연쇄가 대부분 회피될 것으로 보인다. 그 외 어두 제약 S5와 S25를 위배하는 연쇄와 어말 제약 S1, S29를 위배하는 연쇄를 제외하면, 한자어에서 [ɛi, ai, oi, ui]만이 자유롭게 나타날 것으로 예측된다.

표 27. 두 모음 연쇄의 결합 관계

V1 \ V2	i	e	ɛ	ɨ	ʌ	a	u	o
i	\			x	/	/	/	/
e	x			x	/	/	/	/
ɛ				x	/	/	/	/
ɨ								
ʌ	\							
a				x	/	/	/	/
u				x	/	/		
o				x	/	/	/	/

/ = S5 또는 S25 위배

\ = S1 또는 S29 위배

x = S25와 S1, S29 위배

앞서, 신지영·차재은(2003)은 한국어 어휘 형태소 전반적으로 모음 연쇄가 저지된다고 지적하였다. 그리고 유재원(1997), 하세경(2000)은 고유어의 모음 연쇄가 매우 제한되지만, 한자어의 모음 연쇄는 다소 허용되는 것으로 언급하였다. 본 학습 결과는 이러한 언급을 구체적으로 확인할 수 있다. 우선, 어두 및 어말 제약에 의해 한자어의 모음 연쇄 또한 상당수 제한될 수 있다는 것을 보였다. 그러나 어두 및 어말 제약을 제외하면, 한자어의 모음 결합이 고유어인 경우보다 자유롭게 출현할 것으로 예측되었다.

4.2절에서는 고유어와 한자어에 대한 비범주적 음소배열제약을 제시하였다. 우선, 고유어와 한자어 모두 어두와 어말 자음의 제약이 높은 가중치로 학습되었다. 고유어인 경우는 어말 경음 제약과 어두 모음 [e, ɛ] 제약이 가장 높은 가중치가 할당되었다. 한자어인 경우 어말 설정 장애음 제약과 어두 경음 제약이 가장 상위에 위치하였다. 그리고 두 어휘부에서 모두 다소 낮은 가중치로 [자음]+[e, ʌ]에 대한 제약이 포착되었다. 특히, 한자어인 경우, [e]가 조찰음과 결합되기 어려운 한편, 고유어인 경우 연구개음이 아닌 경음 및 격음과 결합되기 어려운 점이 특징적으로 학습되었다. 이에 더하여 모음 연쇄 제약을 세부적으로 포착하여, 각 어휘부에서 허용될 수 있는 모음 연쇄를 파악하고, 그 강도를 예측하였다.

이와 같은 제약을 바탕으로 화자들은 해당 어휘부 내에서 단어간의 적형성 차이를 연속적으로 인식하고, 나아가 새로운 단어가 각 어휘부 문법에 부합하는지 판단할 수 있을 것이다. 특히, 높은 강도의 제약을 위배하는 단어는 어휘부에 출현하더라도 해당 어종의 단어로서 적형성이 매우 낮을 것이 예측된다.

5. 논의

이상에서 범주적 음소배열제약과 비범주적 음소배열제약을 제시하였다. 각 제약은 가중치에 따라 그 강도를 나타내는 바, 범주적 음소배열제약과 비범주적제약을 모두 포함하여 연속적인 문법 인식을 포착할 수 있다. 이러한 모델의 특성을 바탕으로 다음 두 가지를 예측할 수 있다.

첫 번째는 비범주적 제약이 범주적 제약보다 높은 가중치를 가지는 것이 가능하다는 것이다. 한자어의 어두 제약을 예로 들어 살펴본다. 어두 경음을 회피하는 제약 S22는 ‘쌍수, 끽연’과 같이 예외를 허용하지만 높은 가중치 5.63이 할당되었다. 반면, 어두 [nʌ]에 대한 제약 S9는 예외가 출현하지 않지만, 이보다 낮은 가중치 4.44가 할당되었다. 이 가중치 값에 따라, 화자들은 경음으로 시작하는 형태를 [nʌ, mʌ]로 시작하는 형태보다 한자어로서의 적형성이 낮다고 판단하리라 기대한다.

두 번째는 표 28-29과 같이, 실제 단어 사이에서도 세밀한 적형성의 차이가 보일 수 있다는 것이다. 예를 들어, 고유어 어휘부에

서 ‘애꾸’는 N6(*#[ɛ])과 N18(*[k’u]#)을 모두 위배하면서 높은 비적형성 점수를 보인다. 반면, ‘집터’와 같은 단어는 장애음 다음에 [p^h, t^h, k^h]를 제한하는 제약만을 위배하며, 상대적으로 낮은 비적형성 점수를 가진다. 이에 따라 고유어 문법은 ‘애꾸’를 강하게 제한할 것이 예측된 반면, ‘집터’와 같은 단어는 어느 정도 허용될 수 있다는 것을 의미한다.

표 28. 고유어의 비적형성 점수 예

단어	비적형성 점수	위배하는 제약(가중치)
애꾸 [ɛk’u]	6.16	N6 *#[-고설,-후설] (3.54) N18 *[+긴장][-저설,+후설]# (2.62)
부엌 [puʌk^h]	5.20	N11 *[+기식,-양순,-설정]#(2.72) N25 *[-저설,+성절][-원순] (2.48)
밖 [pak’]	4.44	N5 *[+긴장]#(4.44)
도우미 [toumi]	2.30	N26 *[+원순][+원순] (2.3)
집터 [cipt^hʌ]	2.15	N22 *[-공명][-긴장,-조찰] (2.15)

한자어인 경우, 어두 경음 제약 S22를 위배하는 ‘끽연’은 높은 비적형성 점수를 보인다. 반면, ‘처우’과 같은 단어는 [ʌ, i] 다음에 후설모음의 결합을 막는 제약 S33을 위배하며 그 비적형성 점수가 낮다. 이에 더해, ‘급파’와 같은 단어는 [p, t] 다음에 격음을 저지하는 제약 S28에 의해 아주 낮은 비적형성 점수만을 가진다. 따라서 ‘끽연’은 한자어로서의 적형성이 매우 낮은 반면, ‘처우, 급파’는 이보다 한자어의 문법에 부합하여, 적형성이 다소 높을 것으로 예측된다.

표 29. 한자어의 비적형성 점수 예

단어	비적형성 점수	위배하는 제약(가중치)
끽연 [kikjʌn]	5.63	S22 *#[+긴장] (5.63)
게시 [kesi]	5.60	S24 *[-조찰][-고설,-저설,-후설] (5.60)
이앙 [iaŋ]	4.95	S25 *#[+성절][+후설,+성절] (4.95)
처우 [c^hʌu]	2.79	S33 *[-저설,+후설,-원순][+후설] (2.79)
급파 [kipp^ha]	1.08	S28 *[-공명,-연구개][-긴장] (1.08)

이처럼 본 연구는 해당 연쇄에 대한 적형성을 수치화하여 예측할 수 있다. 이러한 예측은 실제 화자들이 보이는 적형성 판단 자료(예: 1-7점 적형성 판단 실험)와 직접 대응하여 검증할 수 있다는 장점이 있다.

물론 다수의 문어 및 구어 코퍼스 내에서 한국어 음소 분포 및 이웃한 음소의 전이 빈도를 보고한 바 있다(김경일 1985, 진남택

1992, 유재원 1997, 이상억 2001, 한성우 2006, 신지영 2005, 2008, 2010). 일부 연구(Lee 2007, Hong 2010, 김미란 외 2014)는 엄밀한 통계적 기법을 도입하여, 보다 유의미한 음소 결합 관계를 탐색하기도 하였다.[14] 그리고 한자 단음절에 대해 음소 전이 빈도를 다루고 기존에 포착되지 않은 제약(예: */ㅟ, ㄲ, ㄳ, ㅙ, ㅞ, ㅖ/ + 종성)을 일반화하여 제시한 연구(신지영 2009)도 있다. 그러나 각 양적 지표의 의미하는 바는 연구자마다 상이하였으며 무엇보다 양적인 정보가 문법에 반영될 수 있는 기제가 없었다. 이에 따라 어느 수준의 적은 빈도가 제약으로 인식되는지 규정하기 어렵고 제약 간의 상대적인 강도를 나타내는 지표도 일관적으로 논의되지 못했다는 한계가 있다.

본 연구는 통계적으로 뒷받침된 모델을 바탕으로 음소배열제약과 그 강도를 구체적으로 제시하여, 적형성을 예측하였다. 이를 바탕으로 각 어휘부의 제약과 가중치가 화자의 적형성 인식에 그대로 반영되어 실재하는지를 확인해 볼 수 있다.

6. 결론

본 연구에서는 범주적 음소배열제약과 비범주적 음소배열제약을 연속적인 모델로 단일하게 포착하였다. 기존에서 관찰된 음소배열제약을 단어 경계 정보를 포함하여 세부적인 제약으로 학습하였고, 기존연구에서 보고된 바 없는 새로운 음소배열제약도 출력하였다. 이에 더하여 부여된 가중치에 따라 제약의 강도를 예측할 수 있었다.

무엇보다 학습된 제약은 연구자가 미리 입력한 것이 아니라 컴퓨터 프로그램을 통해 가상 어휘부에 접근하여 귀납적으로 찾은 것이다. 이로써 다른 이론적 음소배열제약 모델에 대한 기준 모델로서 기능할 수 있으며, 보다 발전된 음소배열제약 모델을 구현하는 시작점이 될 수 있을 것이다.

참고문헌

강범모·김흥규. 2009. *한국어 사용 빈도*. 서울: 한국문화사.
강용순. 1998. 한국어 어휘부 구조. *음성·음운·형태론연구* 4, 55-67. 한국음운론학회.
권인한. 1997. 현대국어 한자어의 음운론적 고찰. *국어학* 29, 243-260. 국어학회.
김경일. 1985. *한국어 음절구조에 관한 통계분석*. 서울대학교 언어

[14] 이용은·임선희(2014)는 몽골어를 대상으로 음소배열제약에 대한 3가지 통계적 기법을 비교하여 제시하였다.

학과 석사학위논문.
김미란·최재웅·홍정하. 2014. 한국어 초성-중성 결합의 분포적 특성 및 모음의 군집분석 연구. *음성·음운·형태론연구* 20.1, 23-49. 한국음운론학회.
박선우. 2006. 현대국어 한자어의 재구조화에 대한 검토. 최남희, 정경일, 김무림, 권인한(편). *國語史와 漢字音*, 265-291. 서울: 박이정.
박선우·홍성훈·변군혁. 2013. 한국어의 어휘계층과 음운론적 복잡성. *음성·음운·형태론연구* 19.2, 225-274. 한국음운론학회.
송기중. 1992. 현대국어 한자어의 구조. *한국어문* 1, 1-85. 한국정신문화연구원.
신지영. 2005. 한국어 음소의 전이 빈도: 3세~8세 아동의 자유 발화 자료를 바탕으로. *한국어학* 28, 81-109. 한국어학회.
_____. 2008. 성인 자유 발화 자료 분석을 바탕으로 한 한국어의 음소 전이 빈도. *언어청각장애연구* 13.3, 477-502. 한국언어청각임상학회.
_____. 2009. 한국 한자음의 빈도 관련 정보 및 음절 구조 제약. *말소리와 음성과학* 1.2, 129-140. 한국음성학회.
_____. 2010. 한국어 사전 표제어 발음의 음소 및 음절 빈도. *언어청각장애연구* 15.1, 94-106. 한국언어청각임상학회.
신지영·차재은. 2003. *우리말 소리의 체계*. 서울: 한국문화사.
안소진. 2009. 한자어 구성 음절의 특징에 대하여. *형태론* 11.1, 43-59. 형태론학회.
유재원. 1997. 한국어 음소 결합 제약에 대한 계량언어학적 연구. *한글* 238, 67-118. 한글학회.
이상억. 2001. *계량언어학*. 서울: 박이정.
이용은·임선희. 2014. 비범주적 음소배열제약 계량화 방법 비교 연구. *언어학 연구* 31, 227-249. 한국중원언어학회.
이주희. 2005. 최적성 이론과 음운론적 어휘부 연구. *돈암어문학* 18, 383-413. 돈암어문학회.
진남택. 1992. *한국어 음소의 기능부담량과 음소연쇄에 관한 계량언어학적 연구*. 서울대학교 언어학과 석사학위논문.
채서영. 1999. 음운변화에 나타난 한국어 어휘의 층위구조. *음성·음운·형태론연구* 7, 217-236. 한국음운론학회.
하세경. 2000. *국어 모음충돌 회피현상에 관한 연구*. 서울대학교 언어학과 석사학위논문.
한성우. 2006. 국어 단어의 음소 분포. *어문학* 91, 163-191. 한국어문학회.
허웅. 1985. *국어 음운학*. 서울: 샘문화사.
BYBEE, JOAN. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.
CHO, HYESUN. 2012. Statistical learning of Korean phonotactics. *Studies in*

Phonetics, Phonology and Morphology 18.2, 339-370. The Phonology-Morphology Circle of Korea.

CHOMSKY, NOAM and MORRIS HALLE. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1, 97-138.

COLAVIN, REBECCA IRENE VICTORIA. 2013. *Phonotactic Probability in Amharic*: *a Psycholinguistic and Computational Investigation.* PhD Dissertation. University of California.

COLEMAN, JOHN and JANET PIERREHUMBERT. 1997. Stochastic phonological grammars and acceptability. In John Coleman (ed.). *Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, 49-56. East Stroudsburg, PA: Association for Computational Linguistics.

DALAND, ROBERT, BRUCE HAYES, JAMES WHITE, MARC GARELLEK, ANDREA DAVIS and INGRID NORRMANN. 2011. Explaining sonority projection effects. *Phonology* 28.2, 197-234. Cambridge: Cambridge University Press.

EISNER, JASON. 2001. Expectational semirings: Flexible EM for finite-state transducers. In Gertjan van Noord (ed.). *Proceedings of the ESSLLI Workshop on Finite-State Methods in NLP (FSMNLP).* Extended abstract (5 pages).

_____. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 1-8. East Stroudsburg, PA: Association for Computational Linguistics.

HAY, JENNIFER, JANET PIERREHUMBERT and MARY BECKMAN. 2003. Speech perception, well-formedness, and the statistics of the lexicon. In John Local, Richard Ogden and Rosalind Temple (eds.). *Papers in Laboratory Phonology VI,* 58-74. Cambridge: Cambridge University Press.

HAYES, BRUCE and JAMES WHITE. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44.1, 45-75.

HAYES, BRUCE and COLIN WILSON. 2008. A Maximum Entropy Model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3, 379-440.

HONG, SUNG-HOON. 2010. Gradient vowel cooccurrence restrictions in monomorphemic native Korean roots. *Studies in Phonetics, Phonology and Morphology* 16.2, 279-295. The Phonology-Morphonology Circle of Korea.

ITÔ, JUNKO and ARMIN MESTER. 1999. The phonological lexicon. In Natsuko Tsujimura (ed.). *The Handbook of Japanese Linguistics*, 62-100. Blackwell Publishers.

KAGER, RENE and JOE PATER. 2012. Phonotactics as phonology: knowledge of a complex restriction in Dutch. *Phonology* 29.1, 81-111.

LEE, YONGEUN. 2007. Effects on inter-phoneme probabilities on the acceptability judgment of Korean CVC nonwords. *Speech Sciences* 14.4, 41-52. The Korean Society of Speech Sciences.
MARTIN, ANDREW. 2011. Grammars leak: Modeling how phonotactic generalizatiosn interact within the grammar. *Language* 87.4, 751-770.
MIKHEEV, ANDREI. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics* 23, 405-423.

박나영
151-745 서울특별시 관악구 관악로 1
서울대학교 언어학과
e-mail: arimnet@naver.com

received: October 5, 2014
revised: November 21, 2014
accepted: November 27, 2014