언어학박사 학위논문

한국어 음소배열제약의 통계적 학습과 적형성 판단

2020년 2월

서울대학교 대학원
언어학과 언어학전공
박 나 영

한국어 음소배열제약의 통계적 학습과 적형성 판단

지도교수 전 종 호

이 논문을 언어학박사 학위논문으로 제출함
2019년 11월

서울대학교 대학원
언어학과 언어학전공
박 나 영

박나영의 박사 학위논문을 인준함
2020년 2월

위 원 장 이 호 영 (인)

부위원장 전 종 호 (인)

위 원 조 혜 선 (인)

위 원 강 희 조 (인)

위 원 정 선 우 (인)

초록

본 논문은 한국어 음소배열제약에 대한 연구로서, 기계 학습(machine learning) 방법을 한국어 어휘부에 적용하여 가능한 모든 제약을 자동으로 학습하였으며, 학습된 제약의 심리적 실재 가능성을 확인하고 있다. '음소배열제약'이란 음소 단위의 결합 회피 혹은 선호에 대한 모국어 화자의 직관적 판단을 나타내고, 이 판단은 전통적으로 '적형/비적형'이라는 범주적인 관점에서 다루어졌다(예: ***bl***ick [적형] vs. ***lb***ick [비적형], Chomsky & Halle 1965). 그러나 영어 음소배열제약에 대한 연구를 포함한 다수의 최근 연구에서 통계적 방법론을 적용하여 비범주적 문법을 탐색하고, 그 문법의 심리적 실재를 증명하는 데 성공하였다(Coleman & Pierrehumbert 1997, Hayes & Wilson 2008, Albright 2009 등). 이러한 배경에서 본 연구는 비범주성을 중심으로 한국어 음소배열제약의 실체를 체계적으로 탐색하고자 하였으며, 이를 위해 기계 학습 방법과 적형성 판단 조사를 시행하였다.

먼저, 기계 학습을 시행하여 한국어 어휘부에서 유효한 음소배열제약 목록을 제시하였다. 학습 자료는 단일 형태소인 명사 단어로 구성하였고, 어휘 부류의 차이를 고려하여 고유어 어휘 목록과 한자어의 어휘 목록을 구분하여 학습하였다. 학습 모델은 음운론 문법 습득 모델 중 가장 일반적이며 효과적인 '최대 엔트로피 음소배열제약 모델(Hayes & Wilson 2008)'을 이용하였다. 학습된 고유어 및 한자어 문법은 과거 한국어 음소배열제약에 대한 연구에서 산발적으로 제시한 제약 및 통계적 경향을 대부분 포함하고 있다. 그 밖에 과거 연구에서 언급된 적이 없는 제약 및 통계적 경향도 새롭게 학습되었다. 결과적으로 특정 현상 및 주제에 집중하면서 통계적으로 분명하지 않은 방식으로 진행된 대부분의 과거 연구와 비교해서, 본 연구는 총체적이고 통계적으로 정당화될 수 있는 문법을 제시하고 있다.

그리고, 어휘부에서 학습된 제약들이 한국어 화자들의 인식에 실재하는지를 확인하기 위하여 '후두자질 발생 및 공기 제한'에 초점을 맞추어 적형성 판단 실험을 시행하였다. 실험 결과, 후두자질 발생 및 공기에 대한 인식이 상당 부분

실재하며, 고유어 및 한자어 어휘부 문법 각각이 독립적으로 적형성 판단에 영향을 미친다는 것을 확인할 수 있었다.

본 연구는 한국어 어휘부에서 학습될 수 있는 음소배열제약 목록을 총괄적으로 제공하고 학습된 제약의 심리적 실재 여부를 모국어 화자를 대상으로 직접 조사하였다. 본 연구에서 제시하고 있는 문법은 최적의 한국어 음소배열제약 모델 개발의 기준 모델로 이해될 수 있다.

주요어: 음소배열제약, 비범주적 적형성, 후두자질 공기 제약, 적형성 판단 조사
최대 엔트로피 음소배열제약 학습 모델

학번: 2013-30853

목 차

초록 i

1. 서론 1

2. 이론적 배경 5

2.1. 기술 통계량 5

2.1.1. 관찰 빈도/기대 빈도 비율 6

2.1.2. 전이 확률 10

2.1.3. 요약 12

2.2. 분절음 N-gram 모델 13

2.2.1. 바이그램 모델 13

2.2.2. 음소배열제약 확률 계산기 14

2.2.3. 음절두음-운모 모델 16

2.3. 자질 기반 N-gram 모델 17

2.4. 일반화 이웃 모델 19

2.5. 입력형-출력형 대응 조화 문법 22

2.5.1. 구성 및 상대적 수용도 23

2.5.2. 학습: 점진적 학습자 알고리즘 24

2.6. 최대 엔트로피 음소배열제약 모델 29

2.6.1. 발생 확률 계산 26

2.6.2. 음소배열제약의 학습 과정 27

2.7. 문법 모델 예측: 영어 화자의 적형성 판단과 비교 26

3. 기존연구: 한국어의 음소배열제약 33

3.1. 범주적 음소배열제약(전체 한국어 및 고유어 어휘부) 33

3.2. 비범주적 음소배열제약(전체 한국어 및 고유어 어휘부) **38**
3.2.1. 인접 연쇄 38
3.2.2. 비인접 연쇄 42
3.3. 한자어 어휘부 특정 제약 **48**
3.4. 논의 **50**
3.4.1. 기존 음소배열제약 탐색 방법의 한계 50
3.4.2. 한국어 화자의 비범주적 적형성 51
3.4.3. 최대 엔트로피 음소배열제약 모델의 도입 필요성 52
4. 학습 **55**
4.1. 학습 방법 **55**
4.1.1. 학습 어휘부 55
4.1.2. 자질 목록 57
4.1.3. 학습 조건 60
4.2. 결과 **61**
4.2.1. 고유어 및 한자어 문법 공통 음소배열제약 61
4.2.2. 고유어 문법 특정 음소배열제약 70
4.2.3. 한자어 문법 특정 음소배열제약 76
4.3. 논의 **79**
5. 모국어 화자의 적형성 판단 조사 **81**
5.1. 목적과 대상 **81**
5.2. 연구 방법 **82**
5.2.1. 조사 단어 생성 82
5.2.2. 조사 과정 83
5.2.3. 피실험자 84
5.3. 결과 **85**
5.3.1. 실험 관찰과 문법 예측 비교 86

5.3.2. 격음/경음 발생 유형별 학습 점수와 응답 점수 89

5.4. 논의 94

6. 결론 97

참고문헌 101

[부록1] 학습 제약: 고유어 111

[부록2] 학습 제약: 한자어 116

[부록3] 비단어 목록 121

[부록4] 지시 문항 123

[부록5] 비단어에 대한 비적형성 점수와 응답 점수 127

ABSTRACT 139

1. 서론

음소배열제약이란, 화자의 모국어 음소 연쇄에 대한 허용 및 회피에 대한 문법 인식을 말한다. 예를 들어, 한국어 화자는 /j/ 음소와 /i/ 음소의 연속이 가능하지 않다고 판단하는데, 대부분의 음운 이론에서는 이런 문법적 판단에 대해 */ji/와 같은 형태의 제약을 가정하여 포착하고 있다.

전통적인 관점에서 음소배열제약은 '문법'과 '비문법' 혹은 '적형'과 '비적형'이라는 이분법적 관점으로 논의되었다(이하 적형/비적형으로 통일). 예 (1)은 영어 화자가 임의의 어두 자음군이 실현된 단어에 대해 적형성 판단을 보인 것이다.

(1) 영어의 어두 자음군(Chomsky & Halle 1965:101)

a. 실재 단어: **br**ick	적형
b. 비단어: **bl**ick	적형
c. 비단어: **bn**ick	비적형

예 (1)에서 보면 영어 화자는 실제 단어인 'brick'뿐만 아니라 비단어 'blick'도 적형적인 단어로 판단하지만, 'bnick'은 비적형적인 단어로 판단한다는 것을 알 수 있다. 전통적인 음소배열제약 연구에서는 화자의 이러한 판단을 토대로 영어에서 /br/과 /bl/ 연쇄가 적형적인 음절두음인 반면, /bn/ 연쇄는 비적형적인 음절두음으로 분석하여 설명하였다.

그러나 최근 다수의 연구는 비단어의 적형성을 판단하는 실험을 통해, 음소 연쇄의 적형성이 기본적으로 비범주적임을 밝히고 있다. 특히, 이러한 비범주적 문법성은 영어 화자를 대상으로 비단어의 단어성 판단 조사에 대한 다수의 실험 결과에 바탕을 두고 있다(Greenberg & Jenkins 1964, Ohala & Ohala 1986, Scholes 1966, Coleman & Pierrehumbert 1997, Frisch et al. 2000, Bailey & Hahn 2001, Albright & Hayes 2003, Hay et al. 2003, Hayes & Wilson 2008, Albright 2009, Daland et al. 2011). 이런 실험 연구들의 결과에 따르면, 영어 화자는 실제 영어 단어에 포함된 연쇄(attested sequences, 이하 어휘부 발생 연쇄)를 어떤 영어 단어에도 포함되지 않는 연쇄(unattested sequences, 이하 어휘부 비발생 연쇄)보다 '영어답다'고 판단하는 수용도의 차이를 보인다. 그러나 수용도의 차이는 여기서

그치지 않고, 발생 연쇄 및 비발생 연쇄들 각각의 사이에서도 적형성의 차이가 있음이 보고되었다. 예 (2)를 보면 /dl/ 연쇄와 /bz/ 연쇄는 모두 영어에서 발생 빈도가 0이나, /dl/ 연쇄가 포함된 'dlap' 형태가 /bz/ 연쇄가 포함된 'bzack' 형태보다 더 높은 문법성 수용도를 보인다. 한편, /st/ 연쇄와 /bl/ 연쇄는 모두 발생 연쇄이지만, 영어 화자는 'stin' 형태를 'blafe' 형태보다 더 문법적으로 수용한다 (Albright 2009: 11).

(2) 임의의 음소 연쇄에 대한 비범주적 수용도의 예(Albright 2009)

(A > B: A의 수용도가 B의 수용도보다 높음)

a. 발생하지 않는 연쇄: **dl**ap > **bz**ack

b. 발생하는 연쇄: **sti**n > **bl**afe

(2a)의 비발생 연쇄 간 수용도 차이는 전통적인 생성 음운론에서는 보편적 문법 원리인 공명도 연속 원리(Sonority Sequencing Principle)를 이용하여 설명되었다. 그러나, (2b)의 발생 연쇄들 간 수용도 차이는 생성 음운론에서 대체로 문법적 설명의 대상에서 제외되었으며, 어휘부 내 단어와의 유사성(유추) 또는 연쇄의 빈도 등 비문법적 요소를 이용한 설명이 제시되어 왔다.

이러한 배경에서 이 연구는 한국어 음소배열제약의 비범주적 성질을 포착하고 그 실체를 파악하는 것을 목표로 삼아 한국어 화자의 음소배열제약 문법을 전반적으로 탐색하고자 한다. 음운론 문법의 습득 시뮬레이션에 쓰이는 기계 학습 모델 중 가장 일반적이면서 효과적으로 알려진 '최대 엔트로피 음소배열제약 학습 모델(Maximum Entropy Phonotactic Learning Model, Hayes & Wilson 2008)'을 이용하여 한국어 어휘부에 존재하는 음소 연쇄의 회피 경향 및 강도를 파악하고 이를 문법적 제약 형태로 포착할 것이다.

최대 엔트로피 음소배열제약 학습 모델을 이용한 기계 학습은 선험적인 제약을 전제하지 않으며 어휘부를 바탕으로 제약이 귀납적으로 습득된다. 이 연구의 학습은 고유어와 한자어의 음소배열제약이 상이하다는 기존연구의 주장을 고려하여 고유어 어휘부와 한자어 어휘부를 나누어 별개로 진행하였다. 이 연구에서 학습된 문법들은 기존연구에서 전통적인 기술 방법 및 분류를 통해 밝힌 제약들

을 대부분 포함하였다. 나아가 고유어 어휘부와 한자어 어휘부를 대상으로, 기존 연구에서 제약으로 보고되지 않은 일반화들도 포착하였다.

그리고 기계 학습된 문법의 제약들이 한국어 화자의 심리적 문법에 실재하는지 여부와 실재하는 경우 반영 정도를 확인하기 위해, '후두자질 공기 제약(laryngeal co-occurrence restriction)'에 초점을 맞추어 비단어의 적형성 판단 실험을 실시하였다. 실험 결과는 기계 학습된 문법이 한국어 화자의 심리적 문법을 상당 부분 반영함을 시사하고 있다. 한국어 화자의 적형성 판단에 대한 이 연구의 분석에 따르면, 학습된 고유어 문법과 한자어 문법이 한국어 화자의 적형성 판단에 모두 독립적인 기여를 하고 있다.

이 논문의 구성은 다음과 같다. 2장에서 비범주적 음소배열제약을 찾기 위한 기술 통계량 및 통계적 문법 모델을 소개한다. 그리고 다수의 모델 중 '최대 엔트로피 음소배열제약 학습 모델'이 회피 연쇄를 자연 부류로 체계적으로 일반화하며, 제약 강도를 통계적으로 뒷받침된 방식으로 계산할 수 있다는 점을 논의한다. 3장에서는 기존연구에서 관찰된 한국어의 음소배열제약 및 연쇄의 회피 경향을 종합한다. 특히 한국어 화자의 비범주적 적형성 인식의 실재가 기존연구들에서 시사되어 왔지만, 비범주적 인식을 포괄하는 체계적 음소배열제약 탐색이 아직 수행되지 못했다는 점을 확인한다. 4장에서는 이 연구에서 최대 엔트로피 음소배열제약 학습 모델을 적용한 방법과 그 결과를 논의한다. 먼저, 고유어 문법과 한자어 문법이 공통적으로 학습한 제약을 제시하고, 이들이 기존연구에서 발견한 범주적 음소배열제약 및 비범주적 회피 경향들을 대부분을 포함한다는 것을 확인한다. 그리고 고유어 문법에 특정된 제약과 한자어 문법에 특정된 제약을 서술한다. 이들 문법은 기존연구에서 보고된 음소배열제약을 포함할 뿐만 아니라, 새로운 회피 경향성을 포착할 수 있다. 5장에서는 '후두자질 발생 및 공기 제약'을 중심으로 모국어 화자의 적형성 판단 조사를 다룬다. 실험 관찰과 문법 예측을 비교하여, 한국어 화자의 비범주적 음소배열제약의 실재를 밝힌다. 마지막으로 6장에서 연구 결과를 요약하고, 남은 연구 과제를 논의한다.

2. 이론적 배경

2장에서는 음소배열제약에 대한 기존연구가 음소배열제약의 비범주성을 어떻게 탐색하여 왔는지 검토한다. 다수의 연구에서 음소배열제약의 비범주성과 어휘부 통계량(lexical statistics)의 관계를 밝히려고 노력해 왔다. 특히 어휘부 내에 특정한 음소 연쇄를 회피하는 제약이 존재한다고 판단을 내릴 수 있는 양적 기준의 정립이 시도되었으며, 회피되는 연쇄의 유형을 분절음 또는 자질 단위를 이용하여 일반화하려는 시도가 꾸준히 행해져 왔다.

이러한 시도들을 크게 두 가지로 나눌 수 있다. 하나는 특정한 음소 연쇄의 분포상 제약의 존재를 드러내는 기술 통계량 연구이다. 2.1절에서 여러 통계량 가운데 널리 쓰이는 '관찰 빈도/기대 빈도 비율(O/E ratio)'과 '전이 확률(transitional probabilities)'을 소개하고, 이들 값의 통계적 적절성을 검토한다. 다른 하나는 특정 언어의 어휘부를 구성하는 음소 연쇄의 분포를 학습하고 이를 이용하여 임의의 연쇄에 대한 수용도를 예측하는 통계적 학습 모델이다. 2.2절부터 2.6절에서는 지금까지 제안된 여러 모델을 개괄하여 각 모델이 음소 연쇄의 분포를 계산하는 방식과 특정한 음소 연쇄의 회피 경향을 제약으로 일반화하는 기제를 검토한다. 2.7절에서는 임의의 연쇄에 대한 각 모델이 예측하는 수용도와 실제 화자의 적형성 판단을 비교한 논의를 토대로, '최대 엔트로피 음소배열제약 모델(Maximum Entropy Phonotactic Model, Hayes & Wilson 2008)'이 한 언어의 음소배열제약을 포괄적으로 탐색할 수 있는 최적의 기준 모델(baseline model)임을 밝힌다.

2.1. 기술 통계량

비단어를 이용한 실험에서 확인된 바에 따르면, 특정한 음소의 연쇄가 어휘부에서 적게 관찰될수록 해당 연쇄로 구성된 비단어의 수용도가 낮아지는 경향이 있다. 예를 들어 영어에는 [비음][저해음]인 /nt/, /ms/, /mk/, /nf/ 연쇄가 발생하는데, 음소 연쇄의 빈도가 낮을수록 음소 연쇄에 대한 수용도가 낮다.

(3) 영어 [비음][저해음] 연쇄의 어휘부 빈도와 수용도(Hay et al. 2003)

a. 고빈도: /nt/

b. 중빈도: /ms/</mk/</nf/

c. 수용도: /ms/</mk/</nf/</nt/

그런데 음소 연쇄의 절대 빈도는 제약의 정당성 및 세기를 적절히 나타낸다고 보기 어렵다. 절대 빈도는 전체 자료의 크기에 영향받는 가변적 통계량이기 때문이다. 또한, 연쇄를 구성하는 각 음소의 빈도가 낮다면 연쇄의 절대 빈도도 낮을 수밖에 없다. 이러한 문제를 피하기 위해 절대 빈도를 대신하여 관찰 빈도/기대 빈도 비율과 전이 확률이라는 두 가지 통계량이 제안되었다. 아래에서 각각 살펴보기로 한다.

2.1.1. 관찰 빈도/기대 빈도 비율

Pierrehumbert (1993)은 특정한 음소 연쇄 발생의 회피 여부 및 그 세기를 기대 빈도를 관찰 빈도로 나눈 값으로 포착하였다. 기대 빈도란, 특정 언어에서 아무런 제약 없이 음소 *x*와 *y*가 자유롭게 연쇄를 이룬다고 가정할 때에 해당 언어에서 *xy* 음소 연쇄가 발생할 수 있는 빈도를 뜻한다. 특정한 음소 연쇄 *xy*의 기대 빈도는 연쇄를 구성하는 *x*와 *y*의 발생 확률을 곱한 값에 해당 언어에서 두 개의 음소로 이루어질 수 있는 연쇄의 총 빈도를 곱하여 구할 수 있다.

(4) 임의의 음소 연쇄 *xy*의 관찰 빈도와 기대 빈도의 비율 계산[1]

$$= (xy\text{의 관찰 빈도})/(\text{연쇄의 총합} \times \frac{x\cdot\text{의 빈도}}{\text{연쇄의 총합}} \times \frac{\cdot y\text{의 빈도}}{\text{연쇄의 총합}})$$

특정 음소 연쇄의 회피와 선호는 관찰 빈도/기대 빈도 비율(이하 O/E 비율)을 1과 비교하여 결정한다. 해당 연쇄의 O/E 비율이 1보다 작으면 어휘부에서 과소 표상(under-represented)된 것으로 보고 회피 연쇄로 판단한다. 그리고 해당 연쇄

[1] $x\cdot = x$로 시작하는 두 음소 연쇄

$\cdot y = y$로 끝나는 두 음소 연쇄

의 O/E 비율이 1보다 크면 과표상(over-represented)된 것으로 보고 선호 연쇄로 판단한다. 이 때 O/E 비율의 유의미성은 카이제곱 검정으로 확보한다.

음소배열제약의 존재를 판단할 때에 O/E 비율이 어떻게 사용되는지, 영어 단음절어의 음절두음과 음절말음의 분포를 예로 들어 살펴보겠다. 논의의 편의를 위해, 영어 단음절어에서 음절두음과 음절말음 위치에 발생하는 자음 부류를 조음위치자질별로 정의하고 해당 음소 연쇄의 절대 빈도를 다음 (5)와 같이 가정하자.

(5) 영어 단음절어를 구성하는 조음위치자질 연쇄의 절대 빈도

음절말음 / 음절두음	양순음	설정음	설배음	음절두음 합계
양순음	26	256	42	324
설정음	204	428	160	792
설배음	22	110	10	142
음절말음 합계	252	794	212	1258

(5)의 값들을 바탕으로, 자음 부류의 조합으로 이루어진 연쇄의 기대 빈도를 구할 수 있다. 예를 들어, 영어 단음절어의 음절두음과 음절말음이 모두 양순음으로 구성되는 [양순음][양순음] 연쇄인 경우의 기대 빈도는 $1258 \times \frac{324}{1258} \times \frac{252}{1258}$=64.9이다. [양순음][양순음] 연쇄의 실제 관찰 빈도는 26이므로, 이를 기대 빈도 64.9로 나누면 [양순음][양순음] 연쇄의 O/E 비율은 0.4가 된다. 아래 (6)은 각 조음위치자질별 자음 부류로 구성된 연쇄의 기대 빈도를 계산한 값이고, (7)은 (5)의 관찰 빈도와 (6)의 기대 빈도를 이용하여 계산한 각 연쇄별 O/E 비율이다.

(6) 영어 단음절어를 구성하는 조음위치자질 연쇄의 기대 빈도

음절말음 / 음절두음	양순음	설정음	설배음	음절두음 합계
양순음	64.9	204.5	54.6	324
설정음	158.7	499.9	133.5	792
설배음	28.4	89.6	23.9	142
음절말음 합계	252	794	212	1258

(7) 영어 단음절어를 구성하는 조음위치자질 연쇄의 O/E 비율

음절말음 / 음절두음	양순음	설정음	설배음
양순음	0.4	1.25	0.77
설정음	1.29	0.86	1.20
설배음	0.77	1.23	0.42

(7)에서 [양순음][양순음]과 [설배음][설배음]의 연쇄는 O/E 비율이 1보다 작기 때문에 과소 표상으로 판단된다. [설정음][설정음]의 연쇄는 (5)에서 확인되는 절대 빈도가 상당히 높은 편이지만 O/E 비율이 1보다 작기 때문에 과소 표상으로 판단된다. 이와 같은 방식으로, O/E 비율을 기준으로 영어에 동일 조음위치자질 연쇄를 회피하는 제약이 있다는 결론을 내릴 수 있다. 다수의 연구는 이 수치가 개별 음소의 절대 빈도에 치우치지 않고, 실제 두 음소 연쇄에 적용되는 제약을 파악할 수 있다고 설명하였다.

그러나 Wilson & Obdeyn (2009)에 따르면, O/E 비율은 각 분절음이 발생하는 위치에 제한이 있으면 해당 분절음이 포함된 연쇄의 과소 표상 또는 과표상의 정도를 정확하게 반영하지 못한다. Wilson & Obdeyn (2009)은 다음 (8)과 같이 분절음의 발생 위치 분포가 제한되는 가상의 언어 사례를 통해 O/E 비율의 문제점을 지적한다.

(8) 가상의 음소 연쇄를 이루는 자음의 위치별 발생 확률 가정

(Wilson & Obdeyn 2009: 104a)

자음2 / 자음1	P(1/3)	T(1/2)	K(1/6)
P(1/3)	1/2	1	1
T(1/2)	1	1/2	1
K(1/6)	1	1	1/2

(8)에서 기술된 가상의 언어에서 두 자음의 분포를 보면 첫 번째 위치에 P, T, K가 올 확률은 각각 1/3, 1/2, 1/6이고, 두 번째 위치에도 마찬가지로 P, T, K가 올 확률은 1/3, 1/2, 1/6이다. 그리고 동일 자음이 공기한 PP, TT, KK 연쇄의 발생 확률은 모두 1/2이고, 그 외 연쇄의 발생 확률은 1이다. 이와 같은 발생 확률을 가정함으로써 이 언어에 동일한 자음이 공기하는 PP, TT, KK에서만 공기 제약을

가정한다. 이 언어에서 P, T, K로 구성된 연쇄 4,999개를 관찰했다고 가정하자. (8)에서 기술한 발생 확률에 따라 각 연쇄의 절대 빈도를 계산하면 (9)와 같다.

(9) 가상의 언어에서 P, T, K로 이루어진 연쇄 4,999개의 분포 자료

(Wilson & Obdeyn 2009: 104d)

자음1 \ 자음2		P_2	T_2	K_2	총합
		1724	2327	948	4999
P_1	1724	**345**	1034	345	
T_1	2327	1034	**776**	517	
K_1	948	345	517	**86**	

(9)의 분포 자료를 이용하여, P, T, K로 이루어진 각 연쇄의 기대 빈도를 계산하면, (10)에서 굵은 선으로 표시된 부분과 같다. 예를 들어, 동일한 자음 P가 공기하는 P_1P_2 연쇄의 기대 빈도는 $4999 \times \frac{1724}{4999} \times \frac{1724}{4999} = 594.6$이다.

(10) 가상의 언어에서 P, T, K로 이루어진 각 연쇄별 기대 빈도

자음1 \ 자음2		P_2	T_2	K_2	총합
		1724	2327	948	4999
P_1	1724	594.6	802.5	326.9	
T_1	2327	802.5	1083.2	441.3	
K_1	948	326.9	441.3	179.8	

(10)에서 계산한 기대 빈도를 (9)에서 가정한 관찰 빈도로 나누면, (11)과 같이 각 연쇄별 O/E 비율을 구할 수 있다. (11)에 제시한 각 연쇄별 O/E 비율을 보면, P_1P_2, T_1T_2, K_1K_2 각각의 발생 확률은 0.58, 0.72, 0.48로 계산되어 (8)에서 입력한 발생 확률 0.5와 일치하지 않는다.

(11) 가상의 언어에서 P, T, K로 이루어진 각 연쇄별 O/E 비율

(Wilson & Obdeyn 2009: 104e)

자음1 \ 자음2	P_2	T_2	K_2
P_1	0.58	1.29	1.06
T_1	1.29	0.72	1.17
K_1	1.06	1.17	0.48

기대 빈도는 구성 음소들의 발생이 서로 독립적이라는 가정에 기반을 두고 있으므로, 음소들 사이의 발생 제한이 존재하는 자료에서 기대 빈도를 토대로 O/E

비율을 계산하고 이를 기준으로 음소배열제약 여부를 기술하는 것은 정당화되기 어렵다.

지금까지 O/E 비율의 계산 방식과 그 문제점을 확인하였다. O/E 비율을 이용하여 특정한 음소 연쇄가 회피 또는 선호되는 양상을 파악하려는 시도가 다수 있었다. O/E 비율을 이용한 음소배열제약 포착은 특히 개별 언어에서 드물게 나타나는 연쇄를 효과적으로 다루며, 비인접 자음 연쇄에서 동일 자질 간 회피 원리(OCP: Obligatory Contour Principle) 등을 밝히는 성과가 있었다(Pierrehumbert 1993, Coetzee & Pater 2008 등). 그러나 O/E 비율은 분절음별로 위치에 따른 발생 제한이 있을 때 그 위치 효과를 배제하지 못하기 때문에 음소배열제약의 존재 여부 및 강도를 판단할 수 있는 기준으로서 적절하지 않음을 알 수 있었다.

2.1.2. 전이 확률

'전이 확률(transitional probabilities)'은 '상대 빈도(relative frequency)'라고도 하는데, 한 음소가 다른 분절음 앞 또는 뒤에 올 확률을 뜻한다. 임의의 연쇄 *xy*에 대하여 전이확률은 *y*가 *x*에 후행할 순방향 전이 확률과 *x*가 *y*에 선행할 역방향 전이 확률을 계산할 수 있다. 일반적으로 전이 확률이라고 하면 전자의 순방향 전이 확률을 가리키며, *xy*의 연쇄 빈도를 *x*로 시작하는 두 음소 연쇄 빈도(*x*·)로 나누어 계산한다. 연쇄의 전이 확률 값이 클수록 해당 언어에서 두 음소의 결합이 선호된다고 해석하고, 연쇄의 전이 확률 값이 작을수록 음소에 제약성이 있다고 해석한다.

(12) 임의의 음소 연쇄 *xy* 순방향 전이 확률 계산

(Saffran, Newport & Aslin 1996: 610)

$$\frac{\text{xy의 빈도}}{\text{x·의 빈도}}$$

Albright (2009)에 따르면, 한 언어에서 관찰되는 두 음소 간 전이 확률과 해당 언어 화자가 지니는 비범주적 음소배열제약 인식 사이에 긴밀한 상관관계가 있다는 연구가 다수 존재한다. (13)에서 볼 수 있듯이, 화자의 반응 시간, 인식률

등 화자의 언어 처리 과정과 관련된 각종 수치가 전이 확률과 대응하는 양상이 보고된다.

(13) 전이 확률과 비범주적 인식의 상관관계(Albright 2009: 13 재인용)

a. 음소 인지 과제(Pitt & McQueen 1998)

b. 따라 읽기 과제(Vitevitch et al. 1997, Vitevitch & Luce 1998, 2005)

c. 동일성/차이성 판단 과제(Vitevitch & Luce 1999)

d. 9개월 아동을 대상으로 한 응시 시간 관찰(Jusczyk et al. 1994)

전이 확률의 계산을 예를 들어 살펴보겠다. 우선 가상의 언어에서 관찰되는 [자음][모음] 연쇄의 절대 빈도를 (14)와 같이 가정하자.

(14) [자음][모음] 연쇄의 절대 빈도

모음 / 자음	ɨ	u	…	[자음] 계
p	3	1479	…	5682
t	310	255	…	4504
…	…	…	…	…
[모음] 계	3076	10681	…	73690

이 언어에서 자음 [p]와 모음 [ɨ]로 이루어진 [pɨ] 연쇄의 순방향 전이 확률을 구하면, [pɨ]의 빈도를 [p]로 시작하는 연쇄의 빈도로 나눈 0.0005이다. 이 값은 모음 [ɨ]가 자음 일반에 후행할 확률 0.0419보다 현저히 낮아 [pɨ] 연쇄 발생이 회피된다고 판단할 수 있다(15a). 계산 방향을 바꾸어 [pɨ] 연쇄의 역방향 전이 확률을 구하면, [pɨ]의 빈도를 [ɨ]로 끝나는 연쇄 빈도로 나눈 0.0010이다. 이 값은 [p]가 모음 일반에 선행할 확률 0.0781보다 낮아, [pɨ] 연쇄의 회피 경향을 판단할 수 있다(15b).

(15) [자음][모음] 연쇄의 전이 확률

a. 순방향 전이 확률(모음의 상대 빈도)

모음 / 자음	ɨ	u	…	계
p	0.0005	0.2603	…	1
t	0.0688	0.0566	…	1
…	…	…	…	1
[모음] 비율	0.0419	0.1458	…	1

b. 역방향 전이 확률(자음의 상대 빈도)

모음 / 자음	ɨ	u	…	[자음] 비율
p	0.0010	0.1393	…	0.0781
t	0.1016	0.0240	…	0.0619
…	…	…	…	…
계	1	1	…	1

이와 같이 연쇄를 이루는 두 분절음 중 어느 한 분절음의 빈도를 기준으로 해당 연쇄의 회피 여부를 파악할 수 있다. 한 가지 지적할 점은 전이 확률을 계산할 때 두 분절음 중 무엇을 기준으로 삼느냐에 따라 제약의 방향성이 정해진다는 것이다. 일반적으로 연구자들은 순방향과 역방향 중 하나를 임의로 선택한다. 그러나 음운 규칙이 적용되는 방향이 연쇄마다 다르다는 것을 고려하면, 전이 확률을 계산할 때 방향을 임의의 하나로 고정하면 다양한 음소배열제약을 충분히 포착하지 못할 가능성이 있다. 또한, 전이 확률의 통계적 유의미성을 검정하는 방식도 충분히 논의되지는 않았다.

2.1.3. 요약

화자의 비범주적 음소배열제약 인식과 임의의 음소 연쇄가 어휘부에 분포하는 양상 간의 관계를 밝히기 위하여, 지금까지 O/E 비율과 전이 확률이라는 두 가지 기술 통계량이 척도로서 제안되었다. 이들 척도를 이용하여 임의의 두 음소로 이루어진 연쇄가 특정한 언어에서 발생할 가능성을 계량화하고 해당 연쇄의 적형성을 비범주적으로 설명할 수 있었다. 그러나 이들 기술 통계량은 실제 어휘에 나타나는 연쇄 빈도만을 토대로 계산되기 때문에, 어휘부 발생 연쇄의 비범주적 적형성만을 설명할 수 있으며 어휘부 비발생 연쇄의 비범주적 적형성을 포착하는 것은 불가능하다. 또한, 전통적으로 논의되어 온 음운론적 구조 및 표상에 이들 기술 통계량이 직접적으로 반영될 기제가 부재하였다. 이러한 문제를 해결하는 대안으로 통계적 학습 모델이 제안되어 왔다. 다음 절부터 지금까지 제안된 주요 음소배열제약 학습 모델을 개괄하고, 각 학습 모델이 어떻게 어휘부의 연쇄 확률을 계산하여 화자의 적형성 판단을 예측하는지 살펴본다.

2.2. 분절음 N-gram 모델

N-gram 모델은 단어 또는 문장의 발생 확률을 부분/요소(part)의 발생 확률을 조합하여 계산한다. 부분/요소를 조합하여 계산하는 방향에 따라, 한 요소에 선행 또는 후행하는 다른 요소를 예측할 수 있다. 이 모델은 품사 태깅, 음성 인식, 표기 교정 등 다수의 자연언어처리 연구에서 채택된다(Jurafsky & Martin 2000, 6장 참고).

음소배열제약 모델로서 채택된 N-gram 모델은 단어의 발생 확률이 그 단어의 적형성을 나타내는 것으로 가정한다. 단어의 발생 확률은 구성 요소의 발생 확률과 요소가 조합되는 방식에 따라 결정된다. 구성 요소가 무엇인지에 따라, N-gram 모델을 분류할 수 있다. 이하에서 각 N-gram 모델이 음소배열제약 모델로서 작동하는 구체적인 방식을 서술한다.

2.2.1. 바이그램 모델

가장 일반적으로 쓰이는 N-gram 모델은 두 분절음을 구성 요소로 삼아 전이 확률을 계산하는 바이그램 모델이다. 학습 자료(training data)인 어휘부를 대상으로 두 음소로 구성된 연쇄의 유형 빈도를 선행 음소로 시작하는 연쇄의 유형 빈도로 나누어 전이 확률을 구한다. 그리고 단어를 구성하는 모든 연쇄의 전이 확률을 곱하여 발생 단어의 확률을 추정한다. 아래 (16)에서 영어 '[stɪn]'과 '[bleɪf]'를 예로 살펴보자.

(16) 바이그램 모델: [stɪn]과 [bleɪf] (Albright 2009: 13, 표 1)

전이 확률	stɪn	전이 확률	bleɪf
P(#→s)	0.118	P(#→b)	0.057
P(s→t)	0.205	P(b→l)	0.106
P(t→ɪ)	0.192	P(l→eɪ)	0.042
P(ɪ→n)	0.108	P(eɪ→f)	0.007
P(n→#)	0.105	P(f→#)	0.067
단어의 확률(log)	−4.12	단어의 확률(log)	−6.93
Albright & Hayes (2003)			
수용도	5.28	수용도	4.21

[#stɪn#]의 발생 확률은 전이 확률 P(#→s), P(s→t), P(t→ɪ), P(ɪ→n), P(n→#) 모두를 곱한 결과이고, [#bleɪf#]는 P(#→b), P(b→l), P(l→eɪ), P(eɪ→f), P(f→#) 모두

를 곱한 결과이다. 계산 결과, 위 (16)에 제시된 바와 같이 [stɪn]의 확률이 [bleɪf]의 확률보다 높고, 이는 [stɪn]의 수용도가 [bleɪf]의 수용도보다 높은 영어 화자의 직관(Albright 2009)을 설명할 수 있다.

2.2.2. 음소배열제약 확률 계산기

음소배열제약 확률 계산기(Phonotactic Probability Calculator: Vitevitch & Luce 2004)는 2.2.1절의 바이그램 모델과 마찬가지로 단어를 구성하는 각 부분의 확률을 구하고, 이를 조합하여 단어의 적형성을 나타낸다. 그러나 바이그램 모델과 달리, 음소배열제약 확률 계산기는 단어를 구성하는 특정 분절음 또는 연쇄가 단어 내에서 몇 번째로 위치하는지 명세한 다음, 명세된 위치별로 특정 분절음에 해당하는 유니그램의 확률과 연쇄에 해당하는 바이그램 확률을 구한다. 유니그램은 (17a)에서 볼 수 있듯이 지정 위치 *a*에 특정 분절음 *i*를 갖는 단어의 사용 빈도를 모두 더한 다음, 지정 위치 *a*에 분절음 일반이 발생하는 단어의 사용 빈도를 나눈 값이다. 바이그램은 (17b)에서 볼 수 있듯이 지정 위치 *a*와 *b*에 분절음 일반이 발생하는 단어의 사용 빈도로 지정 위치 *ab*에 분절음 *i*와 *j*로 구성된 연쇄를 갖는 단어의 사용 빈도의 합을 나눈다. 이 때, 각 사용 빈도의 합은 $\log_{10}$을 취하여 변환한다.

(17) 유니그램과 바이그램의 계산식(Colavin 2013: 16–17)

a. 유니그램 식

- $KF\mathrm{i}_{\mathrm{a}}$: 빈도 사전에서 *a*번째 위치에 분절음 *i*를 갖는 단어들의 집합
- N_a: *a*번째 위치에 분절음이 있는 단어들의 사용 빈도
- *frequency* : 분절음 *i*가 *a*번째 위치한 단어들의 사용 빈도

$$F(i_a) = \frac{\sum_{x \in KF_{i_a}} log_{10}\,(frequency\ x)}{log_{10}(N_a)}$$

b. 바이그램 식

- $KFi_a j_b$: 빈도 사전에서 a 위치에 분절음 i를 가지고, b 위치에 분절음 j가 있는 단어들의 집합
- Nab: ab 위치에 분절음이 있는 단어들의 사용 빈도
- *frequency*: 사용 빈도

$$F(i_a j_b) = \frac{\sum_{x \in KF_{i_a j_b}} log_{10}\,(frequency\ x)}{log_{10}(N_{ab})}$$

영어 [stɪn]을 예로 들어보자. [stɪn]에 대한 유니그램을 구하기 위해서, s가 단어 첫 번째에 위치할 확률, t가 단어 두 번째에 위치할 확률, ɪ가 단어 세 번째에 위치할 확률, n이 단어 네 번째에 위치할 확률을 구해야 한다. 예를 들어, [s]가 첫 번째에 위치할 확률은 s가 첫 번째에 위치하는 단어의 사용 빈도를 모두 더한 다음, 그 합을 단어 첫 번째에 분절음 일반이 발생하는 단어의 사용 빈도의 합으로 나누어 구한다.

(18) [s]가 단어의 첫 번째로 올 확률 [F(s_1)] =

$$\frac{log_{10}\ (\text{s가 첫 번째로 위치하는 단어들의 사용 빈도 합})}{log_{10}\ (\text{분절음이 첫 번째로 위치하는 단어들의 사용 빈도 합})}$$

이와 같이, 단어의 구성 음소 또는 음소 연쇄의 확률값을 더하여 단어의 유니그램과 바이그램값을 구할 수 있다. 음소배열제약 확률 계산기 프로그램을 이용하여, [stɪn]과 [bleɪf]에 대하여 구성 음소가 단어 내 해당 위치에 있을 확률(유니그램)과 특정 연쇄가 단어 내 해당 위치에 있을 확률(바이그램)을 구하고, 이에 대한 합을 구하였다. (19)와 (20)을 보면, 유니그램값과 바이그램값 모두 [stɪn]이 [bleɪf]보다 크고, 이는 [stɪn]이 [bleɪf]보다 수용도가 높다는 보고(Albright & Hayes 2003)와 대응된다. [2]

[2] 2018년 11월 13일 접속 https://calculator.ku.edu/phonotactic/English/words

(19) 유니그램: [stɪn]과 [bleɪf]

위치	위치 확률	stɪn	위치 확률	bleɪf
1	$F(s_1)$	0.1024	$F(b_1)$	0.0512
2	$F(t_2)$	0.0274	$F(l_2)$	0.0447
3	$F(ɪ_3)$	0.0350	$F(eɪ_3)$	0.0283
4	$F(n_4)$	0.0467	$F(f_4)$	0.0159
	$F(x)$의 합	0.2115	$F(x)$의 합	0.1401

(20) 바이그램: [stɪn]과 [bleɪf]

위치	위치 확률	stɪn	위치 확률	bleɪf
12	$F(s_1t_2)$	0.0177	$F(b_1l_2)$	0.0050
23	$F(t_2ɪ_3)$	0.0018	$F(l_2eɪ_3)$	0.0072
34	$F(ɪ_3n_4)$	0.0034	$F(eɪ_3f_4)$	0.0012
	$F(x)$의 합	0.0229	$F(x)$의 합	0.0134

이 모델은 영어, 아동 영어, 아랍어, 스페인어를 대상으로 음소배열제약을 탐색할 수 있는 웹 프로그램(https://calculator.ku.edu/phonotactic)으로 구현되었으며, 심리학 연구 등에서 널리 쓰이고 있다. 그러나 각 분절음 및 연쇄의 단어 위치를 표상하는 근거가 분명하지 않아, 이 모델이 화자의 적형성을 잘 설명한다고 보기는 어렵다(Hayes 2012).

2.2.3. 음절두음-운모 모델

음절두음-운모 모델(onset-rhyme model, Coleman & Pierrehumbert 1997)은 단어를 구성하는 분절음의 발생 확률을 음절두음(onset)과 운모(rhyme)의 위치를 구분하여 계산하고, 이들을 곱한 결과를 해당 단어의 발생 확률로 간주한다. 발생 확률을 순차적으로 곱한다는 점에서 전통적인 바이그램과 유사하다. 그러나 음절두음-운모 모델은 계산 단위가 두 분절음 연쇄가 아니라 음절두음과 운모다. 아래 예 (21)을 살펴보자. 영어 [stɪn]의 확률은 영어 어휘부에서 [st]가 음절두음일 확률과 [ɪn]가 운모일 확률을 구하고 이 두 확률을 곱하여 계산된다. [bleɪf]에 대해서도 마찬가지로 계산할 수 있다.

(21) 음절두음-운모(onset-rhyme) 모델 계산

음절 구조	stɪn의 확률	bleɪf의 확률
음절두음	P(onset[st])	P(onset[bl])
운모	P(rhyme[ɪn])	P(rhyme[eɪf])
단어의 확률	P(onset[st]) × P(rhyme[ɪn])	P(onset[bl]) × P(rhyme[eɪf])

이 모델은 앞선 두 모델에 비해, 음운론적 구조를 반영하여 화자들의 인식을 포착할 수 있다는 장점이 있다. 그러나 음절두음(onset) 경계와 운모(rhyme)의 경계 간, 음절 경계 간 음소배열제약에 대해서는 다루지 못한다(Clements & Keyser 1983: 20−21).

2.3. 자질 기반 N-gram 모델

Albright (2009)는 발생 연쇄뿐만 아니라 빈도 0인 연쇄에 대한 수용도까지 포착하기 위해, N-gram 모델을 활용한 자질 기반 바이그램 모델을 제시한다. 이 모델에서 연쇄의 문법성은 구성 음소가 포함된 자연 부류의 연쇄가 발생할 최대 가능성(likelihood)으로 구한다. 두 음소 a, b로 구성된 연쇄의 발생 확률은 각 음소가 포함된 자연 부류 A, B가 전체 어휘부에서 나타날 확률, 분절음 a가 [자연 부류 A]에 포함될 확률, 분절음 b가 [자연 부류 B]에 포함될 확률을 모두 곱하여 구한다.

(22) 연쇄 ab가 [자연 부류 A][자연 부류 B]로 나타낼 확률
(Albright 2009: (4))

$$\frac{\text{자연 부류 AB의 발생 빈도}}{\text{전체 두 분절음 연쇄 빈도}(biphone)} \times P(a|A) \times P(b|B)$$

자연 부류 AB의 발생 빈도를 최대화하기 위해서는 자연 부류 A와 B에 대한 자질 명세가 포괄적이어야 한다. 그러나 다른 한편으로 P(a|A) 값과 P(b|B) 값을 최대화하기 위해서는 자연 부류 A와 B가 구체적일 필요가 있다.

예를 들어, 연쇄 [st]가 [자음][자음] 연쇄일 확률은 자음이라는 성질을 [+consonantal]로 나타내어 다음 (23)과 같이 구할 수 있다.

(23) 연쇄 [st]가 [자음][자음] 연쇄일 확률

$$\frac{\text{자연 부류 } [+consonantal][+consonantal] \text{의 발생 빈도}}{\text{전체 두 분절음 연쇄 빈도}} \times \frac{1}{20} \times \frac{1}{20}$$

[+consonantal][+consonantal] 연쇄의 발생 빈도는 높지만, [+consonantal] 부류에서 s와 t가 선택될 확률은 각각 1/20으로 매우 낮다.

한편, 연쇄 [st]에 대해서 한 분절음만을 나타낼 수 있도록 구체적으로 자질을 명세하면 다음 (24)와 같다.

(24) 연쇄 [st]가 [s][t] 연쇄일 확률

$$\frac{[-voice, +continuant, +coronal][-voice, -continuant, +anterior] \text{의 발생 빈도}}{\text{전체 두 분절음 연쇄 빈도}} \times \frac{1}{1} \times \frac{1}{1}$$

이와 같이 계산하면, [s][t] 연쇄([−voice,+continuant,+coronal][−voice,−continuant, +anterior])의 발생 빈도는 [+consonantal][+consonantal]의 발생 빈도보다 낮지만, 자연 부류에서 [s]와 [t]를 선택할 확률은 각각 1이 된다. 이와 같이, 이 모델은 연쇄 [st]에 대한 가능한 자연 부류의 연쇄들을 생성할 수 있고, 이 자연 부류의 연쇄들 중에서 발생 확률이 가장 큰 것이 선택된다.

이 모델은 어휘부에서 해당 분절음의 연쇄 빈도가 0이더라도, 그 분절음이 포함된 자연 부류의 연쇄 빈도를 계산하여 연쇄 간 수용도 차이를 계산할 수 있다. 예를 들어, 영어 어두 [bn] 연쇄와 영어 어두 [bd] 연쇄는 모두 빈도가 0이지만, 영어 어두 [bn] 연쇄는 b[+sonornat] 자연 부류 연쇄라는 점에서 어휘부에 존재하는 [bl], [sn] 연쇄 등과 다르지 않다. 이런 자연 부류 연쇄의 빈도를 발생 빈도 계산에 포함함으로써, 영어 어두에서 [bn] 연쇄가 [bd] 연쇄보다 수용도가 높다는 것을 예측할 수 있다.

2.2절과 2.3절에 걸쳐, 분절음 기반 N-gram 모델과 자질 기반 N-gram 모델을 살펴보았다. 이들 모델은 단어의 수용도를 단어의 구성 요소 확률을 조합하여 예측하며, 각 구성 요소(예: 분절음 연쇄, 음절두음, 운모, 자연 부류 연쇄) 및 부여된 확률은 문법에 대응될 수 있다. 이와 같이 N-gram에 기반한 문법은 전통적인 문법과 달리, 어휘부의 분포를 반영하고 비범주적인 문법 직관을 설명할 수 있다. 다만 이 과정에서 문법이 음운론적으로 자연스러운지 여부는 고려되지

않는다. 다음 2.4절에서는 문법을 사용하지 않고, 어휘부로부터 직접 단어의 수용도를 계산하는 모델을 다룬다.

2.4. 일반화 이웃 모델

'일반화 이웃 모델(Generalised Neighbourhood Model, Bailey & Hahn 2001)'은 임의의 비단어에 대한 수용도를 예측하기 위해 임의의 비단어가 기존단어와 얼마나 유사한지를 계산한다. 이 모델은 유추 모델의 일종으로, N-gram 모델과 달리 연쇄의 발생 확률을 조합하지 않고, 문법의 개입없이 어휘부에서 단어의 수용도를 직접 예측한다.

임의의 비단어와 기존 어휘부에 있는 단어들과의 유사성은 둘 사이의 비교를 통해 임의의 비단어를 각각의 기존단어로 바꿀 때 적용되는 수정(교체, 삽입, 탈락), 이른바 '분절음열 편집 거리(string-edit distance)'에 근거한다. 임의의 비단어를 *i*라고 하고 기존단어를 *j*라고 할 때, 분절음열 편집 거리는 비단어 *i*와 기존단어 *j*의 분절음의 개수 차이와 분절음 간 공유 자연 부류 비율을 뜻한다.

예를 들어, 비단어 'scride [skraɪd]'와 기존단어 'shine [ʃaɪn]'의 유사도를 측정해 보자(Albright & Hayes 2003). 이 두 단어의 거리를 가장 최소로 하는 정렬을 찾아 분절음을 대응시키고, 각 분절음별로 편집 거리를 (25)에 제시된 식에 따라 구할 수 있다.

(25) 두 분절음 간 거리(Frisch 1996, Frisch et al. 1997)

a. 공유 자연 부류 비율 = $\frac{\text{공유 자연 부류의 수}}{\text{공유 자연 부류의 수+비공유 자연 부류의 수}}$

b. 거리 $d = 1-$(공유 자연 부류 비율)

해당 두 단어에서 각각 a와 ɪ는 동일하여 편집 거리가 0인 한편, 어두에 위치한 음소 [ʃ]와 음소 [s]의 거리, 어말에 위치한 음소 [n]과 음소 [d]의 거리는 1에서 두 분절음의 공유 자연 부류 비율을 뺀 것이다. 만일 기존단어와 분절음의 수가 다를 때에는 상이한 각 분절음의 삭제(또는 삽입)에 대해 0.6부터 1까지의 거리값을 부여한다. 아래 (26)을 살펴보면, [skraɪd]의 [k]와 [r]은 기존단어 [ʃaɪn]에 대응하는 분절음이 없어 0.7의 값을 부여하였다(Shademan 2007).

(26) scride와 shine의 분절음별 편집 거리 계산

scride	s	k	r	a	ɪ	d	
편집 거리	0.155	0.7	0.7	0	0	0.667	2.222
shine	ʃ	-	-	a	ɪ	n	

두 단어의 전체 편집 거리는 분절음 간의 거리값을 모두 더하여 구한다. 위 (26)에서 볼 수 있듯이, 비단어 'scride [skraɪd]'와 기존단어 'shine [ʃaɪn]'간의 편집 거리는 구성 분절음 간의 거리값을 모두 더한 2.222가 된다.

이러한 두 단어 간 전체 편집 거리값을 이용하여, 두 단어의 유사도를 구할 수 있다. 두 단어의 유사도는 상수 e(≒2.72)를 밑으로 하고, 두 단어의 편집 거리값을 음의 지수로 취하여 계산된다. 이 때, 음의 지수값에 임의의 매개 변수 D를 곱하는데, D 값이 클수록 단어 간 편집 거리 d를 민감하게 반영한다. 이러한 과정을 다음 (27)과 같은 식으로 정리할 수 있다.

(27) 단어 i와 단어 j 간 유사도 계산

$$유사도_{ij} = \exp(-D*d_{ij})$$

예를 들어, (28)과 같이 'scride'와 'shine'의 분절음 간 편집 거리를 대입하여 유사도를 구할 수 있다.

(28) 'scride'와 'shine' 간 유사도 계산

$$유사도_{scride-shine} = \exp(-D*d_{ij}) = \exp(-D*(2.222))$$

비단어 i를 기존단어 j와 쌍을 이루어 구한 유사도를 구하여 모두 더하면, (29)에서 제시한 바와 같이 비단어 i의 유사도 점수(similarity score)를 구할 수 있다.

(29) 비단어 i의 유사도 계산

$$유사도\ 점수_i = \sum_j 유사도(i,j)$$

유사도 점수 계산시, 개별 어휘의 사용 빈도와 사용 빈도 효과를 반영할 수 있다. 유사도 점수 계산 과정에서 개별 어휘 j의 사용 빈도(f_j)는 2가 더하여진 후 로그값으로 변환되며, 이 변환값이 유사도와 곱해진다. 유사도 점수가 단어의

수용도에 비단조적(non-monotonic)으로 나타난다는 가정 아래, 사용 빈도 효과 정도를 조정할 수 있는 이차 함수의 가중치 항(quadratic weighting scheme)인 변수 A, B, C도 포함된다. A, B, C 값은 개별 연구마다 다를 수 있다.[3]

(29') 기존단어 j의 사용 빈도를 반영한 비단어 i의 유사도 점수 계산

유사도 점수$_i$ = $\sum_j (A \log(f_j + 2)^2 + B \log(f_j + 2) + C) \cdot \exp(-D \cdot d_{ij})$

앞서 예를 든 'scride'의 유사도 점수는 (28)과 같이 구한 유사도에 기존단어 'shine, stride, drive'등의 사용 빈도를 변환한 값 $\log(f_{shine}+2)$, $\log(f_{stride}+2)$, $\log(f_{drive}+2)$을 곱하여 구할 수 있다. 사용 빈도 효과 정도는 변수 '−0.47, 2.02, −0.289'와 같은 변수에 의해 조정된다. 아래 (30)에 계산식을 제시한다.

(30) 유사도 점수$_{scride}$[4]

$=((-0.47)*\log(f_{shine}+2)^2+2.02*\log(f_{shine}+2))+(-0.289))\cdot \exp(-D \cdot d_{\text{scride,shine}})$
$+((-0.47)*\log(f_{stride}+2)^2+2.02*\log(f_{stride}+2))+(-0.289)) \cdot \exp(-D \cdot d_{\text{scride,stride}})$
$+ ((-0.47)*\log(f_{drive}+2)^2+2.02*\log(f_{drive}+2))+(-0.289)) \cdot \exp(-D \cdot d_{\text{scride,drive}}) \cdots$
$+ ((-0.47)*\log(f_n+2)^2+2.02*\log(f_n+2))+(-0.289)) \cdot \exp(-D \cdot d_{\text{scride,n}})$

▪ n=어휘부 내 단어

Bailey & Hahn (2001)은 어휘 유사성이 연쇄의 전이 확률과 독립적으로 기능할 수 있다는 것을 보이고, 나아가 유추 모델의 수정을 통해 화자의 비범주적 인식을 효과적으로 포착할 가능성을 제기하였다. 실제로 Bailey & Hahn (2005)에서는 탈락, 삽입, 교체 등의 편집이 대상 분절음 및 환경에 따라 달라질 가능성을 제기하고, 유사성과 비범주적 인식 간의 관계를 보다 세밀하게 포착하고자 하였다.

---

[3] Shademan (2007)의 경우, 사용 빈도 효과가 단조적으로 나타난다고 가정하고 A값을 부여하지 않고 각각 B와 C에 1을 대입하였다. 한편, Daland et al. (2011)은 어휘부와 테스트 항목의 차이를 반영하여 다양한 변수(parameter)를 대입하였다 (Daland et al. 2011: 214).

[4] 변수 A, B, C 값은 Daland et al. (2011: 241, 표 3)의 oral task 값을 대입하였다.

그러나 일반화 이웃 모델에서 제안한 어휘 유사성은 음운론적 정보를 반영하지 않는다. 이 때문에, 화자의 적형성 판단을 체계적으로 예측하기에는 근본적인 한계가 있다(Albright & Hayes 2003, Shademan 2007). 앞서 전이 확률과 비범주적 인식 간의 상관관계를 언급하였듯이, 화자들은 단어의 수용도 판단의 상당 부분을 연쇄의 적형성의 조합에 의존하여 인식하기 때문이다. 2.7절에서 다른 모델과의 비교를 통해 일반화 이웃 모델이 어휘부에서 나타나지 않는 연쇄에 대한 예측을 거의 할 수 없다는 것을 보일 것이다.

2.2절부터 2.4절까지 N-gram 모델 및 일반화 이웃 모델을 살펴보았다. 이러한 학습 모델들은 어휘부에서 관찰 가능한 연쇄의 발생 확률을 조합하여 비단어의 적형성을 예측하거나, 기존단어와 비단어의 '유사도'를 직접 계산하여 비단어의 수용도를 예측할 수 있었다.

그러나 이들 모델은 분절음 하나 또는 앞뒤의 분절음 연속에 대해서만 제약을 포착하기 때문에 음운론적 적형성을 충분히 예측하지는 못한다. 실제 언어에서는 '모음 조화'와 같이 비인접 연쇄에 대한 제약도 실재한다. 즉, 특정 모음이 분절음 층위에서는 인접 분절음과 제약을 이루는 동시에, 모음자질 층위에서 다른 모음과 제약을 이룰 수 있는 것이다(Hayes & Wilson 2008: 381). 이에 따라, 다층적인 음운론적 제약을 충분하게 포착할 수 있는 모델이 필요할 것이다.

2.5. 입력형-출력형 대응 조화 문법

기존 입력형-출력형 대응 형식의 제약 기반 모델을 채택하여, 음운론적 교체 현상과 마찬가지로 음소배열제약을 포착하고자 하는 시도도 있었다. Coetzee & Pater (2008)은 Muna어의 조음위치자질 공기 제약(homorganic consonants co-occurence restriction)을 대상으로 삼고, 이 제약에 대한 비범주적 인식을 밝혔다. 그리고 이를 형식화하기 위해, 제약 강도를 수치로 부여할 수 있는 조화 문법 모델(Legendre et al. 1990, 2006, Smolensky & Legendre 2006)을 적용하였다.

다음 2.5.1절에서 Coetzee & Pater (2008)이 도입한 조화 문법 모델을 소개하고 연쇄의 수용도를 예측하는 방식을 살펴본다. 2.5.2절에서는 학습 모델 '점진적 학습 알고리즘(Gradual Learning Algorithm)'이 어휘부의 분포를 어떻게 제약의 가중치에 할당하는지를 서술한다.

2.5.1. 구성 및 상대적 수용도

조화 문법(Coetzee & Pater 2008)은 입력형에 대해 후보형을 생성하고 제약 간의 상호 작용으로 최적형을 선택한다. 각 제약은 가중치를 할당받으며, 각 후보형의 적형성은 후보형이 위배하는 제약의 가중치와 위배 횟수를 곱한 값의 합에 대응된다. 이 값을 조화값(harmony score)이라 부르며, 계산식을 형식화하면 (31)과 같다.

(31) 조화값 계산식(Coetzee & Pater 2008: 308, (6))

$H(R)=W_1C_1(R)+W_2C_2(R)+ W_3C_3(R)+ ... W_nC_n(R)$

- R = 표상(representation)
- $\{C_1(R), C_2(R), C_3(R), ..., C_n(R)\}$ = 제약의 집합
- $\{W_1, W_2, W_3, ..., W_n\}$ = 가중치의 집합

최적형으로는 계산된 조화값이 가장 큰 후보형이 선택된다. 입력형 α와 γ에 대한 최적형이 선택되는 과정을 살펴보면 (32)와 같다.

(32) 조화 문법에서 최적형의 선택

a. 입력형 α

가중치	3	2	1	조화값
/입력형 α/	제약 1	제약 2	제약 3	
[후보형 α]	−2			−6
☞ [후보형 β]		−2	−1	−5

b. 입력형 γ

가중치	3	2	1	조화값
/입력형 γ/	제약 1	제약 2	제약 3	
[후보형 γ]		−2		−4
☞ [후보형 δ]	−1			−3

(32a)에서 입력형과 같은 형태인 후보형 α는 가중치가 3인 제약을 2회 위배하기 때문에 조화값 −6을 가지며, 후보형 β는 가중치가 2인 제약을 2회, 1인 제약을 1회 위배하기 때문에 조화값 −5를 가진다. 계산 결과, 후보형 β의 조화값이 후보형 α의 조화값보다 크기 때문에 β가 최적형으로 선택된다. 마찬가지로, (32b)

에서 입력형 γ에 대하여 후보형 γ의 조화값은 −4, 후보형 δ의 조화값은 −3을 가지며 후보형 δ가 최적형으로 선택된다.

그런데, 입력형이 다른 후보형들의 조화값은 직접 비교될 수 없다. 위의 (33a)에서 최적형 β는 (33b)에서 최적형이 아닌 후보형 δ보다도 조화값이 낮다. 그러나 실제 화자는 최적형이 아닌 후보형 δ를 최적형 β보다 적형적으로 인식하지 않는다.

Coetzee & Pater (2008)은 입력형이 다른 표면 형태 간의 적형성을 비교하기 위해서, 상대적 수용도(relativized acceptability)를 제안한다. 상대적 수용도란, 동일한 입력형에 대한 후보형 간의 적형성을 구하는 것이다. 해당 후보형의 조화값에서 그 외 후보형 중 가장 큰 조화값을 뺀다. 예를 들어, (32a) 후보형 α의 상대적 수용도는 α의 조화값 –6에서 α의 가장 조화로운 경쟁 후보형 β의 조화값 −5를 뺀 –1이다. 마찬가지 방식으로 (32b) 후보형 γ의 수용도는 –1로 구할 수 있다. 이를 통해, α와 γ의 상대적 수용도는 같다는 것을 알 수 있고, 실제 화자는 두 후보형이 동등한 정도의 적형성을 가지고 있다고 판단할 것으로 예측된다.

(33) 후보형 α와 후보형 γ의 상대적 수용도

a. 수용도(α) = 조화값(α) – 조화값(β) = (–6) – (–5) = –1

b. 수용도(γ) = 조화값(γ) – 조화값(δ) = (–4) – (–3) = –1

2.5.2. 학습: 점진적 학습자 알고리즘

Coetzee & Pater (2008)은 어휘부의 통계량을 문법에 반영하기 위하여, 점진적 학습자 알고리즘(Gradual Learning Algorithm)을 채택한다. 이 학습 알고리즘은 입력형에 대한 적절한 출력형을 찾아가는 실시간 오류 기반(online error-driven) 기제이다. 만약 현재 문법이 실제 관찰 형태와 다른 오류형을 출력한다면, 각 제약의 가중치(W_i)에 조정값을 더해 갱신한다. 조정값은 오류형이 해당 제약을 위배하는 횟수(vE)에서 실제 관찰 형태가 해당 제약을 위배하는 횟수(vC)를 뺀 값이다. 이 값에 0 초과 1이하의 n을 곱하여 값의 변화 속도를 조절한다.

(34) 제약 i의 가중치 갱신

$$W_{i'} = W_i + \{n \times (vE - vC)\} \qquad 0 < n \leq 1$$

예 (35)를 살펴보자. Coetzee & Pater (2008: 319)에 따라, 학습 초기 단계에서 유표성 제약(*AA)의 가중치를 100, 충실성 제약(IDENT)의 가중치를 50으로 둔다.

(35) 조화 문법의 제약 가중치 학습

가중치	100 →	←50	조화값
입력형: /AA/	*AA	IDENT	
✓ AA	−1		−100
☞ AA’		−1	−50

이 문법은 실제 발화형 AA가 아닌, 오류형인 AA’를 출력하기 때문에 가중치 조정이 필요하다. 제약 *AA는 오류형 AA’에서 위배되지 않고(vE = 0) 실제 발화형 AA에서만 1회 위배되기 때문에(vC =1), 가중치 100에 조정값 n×(0–1)을 더해준다. 한편 IDENT는 오류형 AA’에서 위배되고(vE = 1) 실제 발화형 AA에서 위배되지 않기 때문에(vC =0), IDENT 가중치 50에 조정값 n×(1–0)을 더해준다. 이와 같이, 실제 발화형이 최적형으로 출력될 때까지 실제 발화형이 위배하는 제약은 가중치가 낮아지고 오류형이 위배하는 제약은 가중치가 높아진다. 이 가중치 갱신 과정에서 형태의 관찰 빈도가 반영된다.

지금까지 Coetzee & Pater (2008)이 조화 문법을 채택하여, 음소배열제약의 수용도를 설명하는 방식과 문법 습득 과정을 살펴보았다. Coetzee & Pater (2008)은 앞서 살펴본 N-gram 모델과 일반화된 이웃 모델과 달리, 음운론적 교체 현상과 마찬가지로 입력형과 출력형의 대응에 대한 음소배열제약 모델을 제시하였다. 제약 집합이 보편 문법에 의해 전제되며 어휘부 빈도는 제약의 가중치에만 반영된다. 이러한 접근은 기존의 음운론적 교체 현상에 대한 분석과 음소배열제약에 대한 설명을 통합할 수 있다는 장점이 있다. 그러나 제약의 범위가 음운론적으로 자연스러운 것에만 한정되며, 어휘부 자체에서 제약이 습득될 가능성은 고려되지 않는다.

다음 2.6절에서는 Coetzee & Pater (2008)과 마찬가지로 가중치가 부여된 제약 모델인 최대 엔트로피 음소배열제약 학습 모델을 설명한다.

2.6. 최대 엔트로피 음소배열제약 모델

'최대 엔트로피 음소배열제약 모델'(Hayes & Wilson 2008)은 음운론적 위배 형태의 비적형성 정도를 나타내는 제약에 수치를 부여한다는 점에서 조화 문법의 변이형(variant)이라고 볼 수 있다. 최대 엔트로피 음소배열제약 모델은 가중치가 부여된 제약의 상호 작용으로 연쇄의 적형성을 계산한다. 그러나 Coetzee & Pater (2008)이 '입력형-출력형'의 대응을 보편적으로 전제된 제약 목록으로 평가하는 것과 달리, Hayes & Wilson (2008)은 출력형(표면형)을 대상으로 제약을 귀납적으로 학습한다. Hayes & Wilson (2008)의 모델은 어휘부 학습을 통해 다음 (36)과 같은 자질이 명세된 유표성 제약과 이에 따른 가중치를 부여한다.

(36) 학습 결과 예(Hayes & Wilson 2008: 표 16)

제약	가중치
*[−sonorant]C	5.63
*[+dorsal][+coronal]	4.48

이하에서 최대 엔트로피 음소배열제약 모델이 적형성을 계산하는 방식을 살펴본 뒤, 제약 탐색 과정을 설명한다. 그리고 귀납적으로 학습된 제약과 실제 화자의 인식을 비교하는 논의들을 다루며, 모델의 특징을 파악한다.

2.6.1. 발생 확률 계산

이 모델에서는 표면형의 발생 확률이 그 형태에 대한 적형성에 대응된다고 가정한다. 연쇄 적형성은 조화 문법에서와 마찬가지로 연쇄가 위배하는 제약의 가중치를 더하여 구한다. 이 점수를 이른바 '비적형성 점수(score)'라고 부른다. 아래 (37)에 보인 바와 같이 e를 밑으로 하고 비적형성 점수를 음의 지수값으로 취함으로써 최대 엔트로피 값(maxent value)을 구할 수 있고, 전체 연쇄의 최대 엔트로피 값의 합으로 나누면 연쇄의 발생 확률을 구할 수 있다.

(37) 최대 엔트로피 값과 발생 확률의 계산

a. 최대 엔트로피 값(e^H) = $e^{-(\text{비적형성 점수})}$

b. 해당 연쇄의 발생 확률 = $\frac{\text{해당 연쇄의 최대 엔트로피 값}}{\text{최대 엔트로피 값의 합}}$

예 (38)에서 보면, CV는 아무런 제약을 위배하지 않는 반면, CCV는 제약 *#CC를 위배하고, CCVV는 제약 *#CC와 제약 *VV를 위배한다.

(38) 예: 비적형성 점수와 발생 확률

a. *#[+consonantal][+consonantal]([*#CC]) 가중치: 3

b. *[+syllabic][+syllabic]([*VV]) 가중치: 2

	*#CC	*VV	비적형성 점수 (score)	최대 엔트로피 값 (e^H)	발생 확률 (P)
	3	2			
CV	0	0	(0×0)+(0×0)=0	exp(−0) = 1	0.84
CCV	1	0	(3×1)+(2×0)=3	exp(−3) ≒ 0.05	0.11
CCVV	1	1	(3×1)+(2×1)=5	exp(−5) ≒ 0.006	0.04

그 결과, 세 후보형 CV, CCV, CCVV는 각각 0, 3, 5의 비적형성 점수를 가지고, 이를 바탕으로 최대 엔트로피 값을 구하면 각각 1, 0.05, 0.006에 해당하는 최대 엔트로피 값을 갖게 된다. 최대 엔트로피 값의 합 1.056으로 각 연쇄의 최대 엔트로피 값을 나누면 각각 발생 확률이 0.84, 0.11, 0.04로 예측된다.

2.6.2. 음소배열제약의 학습 과정

최대 엔트로피 음소배열제약 모델에서 문법 모델을 구성하는 제약과 가중치를 학습하는 과정을 제시하면 다음과 같다. 먼저, 학습자의 보편 문법에는 자질 목록과 자질 매트릭스로 구성된 유표성 제약 형식(*[자질][자질])이 주어진다. 주어진 자질 목록과 제약 형식을 바탕으로, 가능한 모든 표상 집합이 생성된다.

이 표상 집합을 대상으로 계산된 정확도와 일반성을 기준으로 각 제약이 선택된다. 정확도란, 제약 위배의 관찰 빈도/기대 빈도 비율($O_{[Ci]}/E_{[Ci]}$)을 뜻한다. 기대 빈도보다 위배하는 형태들의 빈도가 일정 수준 이하로 낮으면 해당 제약은 어휘부에서 효과적인 것으로 선택된다. 관찰 빈도가 동일한 제약들 중에서는 가능한 표상 집합 중 기대 빈도가 큰 제약이 더 유효한 것으로 판단된다(Mikheev 1997). 그리고 같은 정확도 수준에서는 더 일반적인 제약이 선택된다. 결합되는 자질 매트릭스 수가 짧고, 제약을 이루는 자질의 수가 적으며 더 많은 자연 부류를 포함하는 제약이 선택된다. 예를 들어, *[+high][+high][+back]와 *[+high][+high]의 정확도 수준이 동등하다면, 결합 자질 매트릭스가 짧은 *[+high][+high]가 제약으로 선택된다. 또한, 제약을 이루는 자질의 수가 많은

*[+anterior,+coronal][−back,+syllabic]와 제약을 이루는 자질의 수가 적은 *[+coronal][−back]의 정확도 수준이 동등하다면, 더 많은 자연 부류를 포함할 수 있는 *[+coronal][−back]이 제약으로 선택된다.

이와 같이 선정된 제약에는 가중치가 할당된다. 학습자는 어떠한 연쇄가 금지되는지에 대한 부정적 증거에 접근할 수 없기 때문에, 관찰된 형태의 발생 확률을 최대화함으로써 관찰되지 않는 형태의 발생 확률을 최소화하는 가중치를 찾아야 한다. Hayes & Wilson (2008)은 최적의 가중치 찾기 방법으로 $O_{[Ci]}-E_{[Ci]}$가 0이 될 때까지 가중치를 찾는 Hill-Climbing search 방법을 채택한다. 아래 (39)에서 음소배열제약의 학습 알고리즘을 요약한다.

(39) 음소배열제약 학습 알고리즘 (Hayes & Wilson 2008: 394, (10))

a. 학습 입력형: 분절음에 명세된 자질 목록과 어휘부

b. 학습 조건

- 정확도 수준(accuracy)의 집합: {0.001 .... 1}
- 최대 결합할 수 있는 자질 매트릭스의 수

c. 학습 단계

- 1단계: 가능한 모든 제약 집합을 탐색 범위로 생성하고 정확도가 일정 수준 이하인 제약 가운데 일반적인 제약을 선정
- 2단계: 선택된 제약에 대한 가중치를 부여
- 일정 정확도에 이를 때까지, 1–2 단계를 반복적으로 학습

최대 엔트로피 음소배열제약 모델의 장점은 다음 세 가지로 볼 수 있다.

첫째, 제약 선정이 통계적으로 뒷받침된다는 것이다. 앞서 다룬 기술적 O/E 비율 또는 N-gram과 달리, 한 언어의 자질 목록에 따라 예측되는 모든 제약의 집합을 기준으로 제약 위배의 기대 빈도와 관찰 빈도를 계산하여 제약을 선정한다. 그 결과, 특정 위치 및 방향에 영향을 받지 않고 빈도가 낮은 연쇄를 효과적으로 포착할 수 있다.

둘째, 문법을 이루는 제약은 자연 부류를 포착할 수 있는 자질(feature)로 구성된다. 다수의 연구에서 분절음 자체의 빈도를 기술하여 음운론적 일반화가 어려웠으며, 음운론적 자질을 활용하더라도 연구자가 임의적으로 정한 것이었다. 이

에 비해, 최대 엔트로피 음소배열제약 모델은 통계적으로 유의미한 자연 부류를 연구자의 개입없이 일반화할 수 있다.

셋째, 다양한 음운론적 이론을 반영하고 평가할 수 있다. 인접한 바이그램뿐만 아니라 세 분절음, 네 분절음 연쇄, 비인접 제약(예: 모음 조화, 음절두음 공기 제약)을 다룰 수 있다.

이러한 세 가지 장점을 바탕으로, 특정 언어의 어휘부를 대상으로 세부적인 비범주적 인식을 예측할 수 있다. 다만, 기존연구(Albright 2009, Daland et al. 2011, Colavin 2013)는 최대 엔트로피 음소배열제약이 선호 연쇄에 대한 비범주적 인식은 포함하지 않으며, 어휘부의 양적 정보를 민감하게 반영하기 때문에, 음운론적으로 자연스럽지 않은 제약까지 학습될 수 있다는 한계를 지적한다. 다음 절에서 각 모델이 예측하는 적형성과 실제 인식 실험 결과를 개괄함으로써, 각 모델의 설명력을 구체적으로 비교해 보고자 한다.

2.7. 문법 모델 예측: 영어 화자의 적형성 판단과 비교

최근 연구(Hayes & Wilson 2008, Albright 2009, Daland et al. 2011)는 영어 어두 자음군에 대한 수용도를 음소배열제약 모델로 설명하고자 하였다. 그 중 Daland et al. (2011)은 영어 어두 자음군의 공명도 척도에 대한 실험을 진행하였는데, 실험 결과가 다수의 음소배열제약 모델의 예측값과 유의미한 상관관계를 보였다. 그런데 아래 (40)에서 볼 수 있듯이, 연쇄의 발생 여부에 따라 수용도와 모델의 예측력 사이의 상관관계가 달라진다. [5]

[5] (40)에 쓰인 모델명 약어의 의미는 다음과 같다.
- Bigram: (분절음) 바이그램 모델
- C & P (1997): 음절두음-운모 모델(Coleman & Pierrehumbert 1997)
- V (2004)[bi]: 음소배열제약 확률 계산기: 바이그램(Vitevitch & Luce 2004)
- Albright (2009): 자질 기반 바이그램 모델
- B & H (2001): 일반화 이웃 모델(Bailey & Hahn 2001)
- H & W (2008) [100]: 최대 엔트로피 음소배열제약 모델(Hayes & Wilson 2008), 제약 100개 학습

(40) Daland et al. (2011): 응답과 모델의 상관관계

모델 \ 연쇄 유형		발생 연쇄 (attested)	예외 연쇄 (marginal)	발생 빈도 0 (unattested)	전체
분절음 N-gram	Bigram	0.19	0.16	0.22	0.78
	C & P (1997)	0.35	0.31	−0.01	0.55
	V (2004) [bi]	0.30	0.06	0.27	0.56
Albright (2009)		0.21	0.03	0.55	0.51
B & H (2001)		0.32	0.23	−0.22	0.31
H & W (2008) [100]		0	0.02	0.76	0.83

분절음 N-gram 모델은 발생 연쇄에 대한 수용도를 잘 예측한다. 학습 단어에 음절 위치를 명세하거나(Coleman & Pierrehumbert 1997), 단어 위치를 명세하는 경우(Vitevitch & Luce 2004) 모델의 예측력이 더욱 높아진다. 그러나 분절음 N-gram 모델은 발생 빈도가 모두 0인 비발생 연쇄의 수용도 차이에 대해서는 잘 설명하지 못한다.

이에 비해, 자질 단위 바이그램 모델(Albright 2009)은 음운론적 단위인 '자질(feature)'을 매개로 연쇄 분포를 일반화한다. 그 결과, Albright (2009) 모델은 발생 연쇄뿐만 아니라 비발생 연쇄들 사이의 수용도 차이를 의미있게 예측할 수 있어서, 균형된 설명력을 가진 모델로 볼 수 있다. 다만, 발생 연쇄에 대해서는 분절음 N-gram보다 예측력이 떨어지며, 비발생 연쇄에 대해서는 Hayes & Wilson (2008) 모델보다 예측력이 떨어진다.

또한, 유추 모델의 일종인 일반화 이웃 모델(Bailey & Hahn 2001)도 발생 연쇄에 대한 수용도를 분절음 N-gram만큼 예측할 수 있는 한편, 발생하지 않는 연쇄에 대해서는 수용도 차이를 예측하지 못한다. 이 모델은 단어 간 유사성에 기반을 두기 때문에 체계적인 음운론적 구조가 고려되지 않는다. 일부 연구(Bailey & Hahn 2001, Shademan 2007)는 일반화 이웃 모델이 N-gram 모델과 상호 보완적 관계라고 보았다.

위의 모델들과 달리, 최대 엔트로피 음소배열제약 모델(Hayes & Wilson 2008)은 발생 빈도가 0인 연쇄 간의 수용도 차이를 잘 설명할 수 있다. 이는 자질을 단위로 삼아 회피 연쇄를 제약으로 일반화하였기 때문이다. 이 모델은 다양한 음운론적 구조를 반영할 수 있다는 점에서 영어 어두 자음군뿐만 아니라, 다른 언어에 대해서도 적용되었다(암하라어: Colavin 2013, 케추아어: Gallagher 2013, 아이마라어: Gallagher et al. 2019). 그 과정에서 개별 언어 화자의 비범주적 적형

성과 모델의 예측이 유의미한 상관관계를 보였다. 이러한 연구 성과를 바탕으로 이 연구에서는 최대 엔트로피 음소배열제약 모델이 개별 언어를 종합적으로 파악하고 화자의 비범주적 적형성을 확인하는데 적합하다고 판단한다.

3. 기존연구: 한국어의 음소배열제약

이 장에서는 기존연구에서 논의된 한국어 음소배열제약을 개괄한다. 먼저, 한국어 음운론 개론서를 비롯한 다수의 연구에서 논의된 범주적 제약들을 살펴본다. 둘째, 복수의 연구가 전체 한국어 및 고유어 어휘부에서 발생이 제한된다고 기술한 연쇄, 즉 비범주적 음소배열제약을 살펴본다. 이들은 공식적인 제약으로 형식화하여 제시되지는 않았지만, 한국어 화자들이 연쇄 발생 제한을 유의미하게 판단할 가능성을 제시한다. 셋째, 고유어와 구별되어 연구된 한자어 음소배열제약을 제시한다. 이러한 기술은 어휘부에 따라 음소배열제약의 종류 및 효과가 상이할 수 있다는 것을 시사한다. 이와 같은 검토를 바탕으로, 3.4절에서는 기존연구의 시도들이 한국어 화자의 비범주적 적형성을 충분하게 예측하지 못한다는 점을 지적하고 통계적 음소배열제약 학습 모델이 적용될 필요성을 제기한다.

3.1. 범주적 음소배열제약(전체 한국어 및 고유어 어휘부)

다수의 연구는 주로 한국어 전체 한국어 및 고유어에서 발생 빈도가 0인 연쇄를 공식적인 제약으로 간주하였다. 다만, 일부 예외를 인정하는 제약도 범주적 제약과 함께 형식화되어 제시되었다. 기존연구에서 논의한 음소배열제약들을 다음 (41)과 같이 분류할 수 있다.

(41) 기존연구에서 논의된 한국어 음소배열제약

a. 음절구조 제약

b. [활음][모음] 제약

c. 필수적 음운 규칙 관련 제약

d. 최근 차용어에서만 예외가 존재하는 제약

e. 고유어와 한자어에서 예외가 존재하는 제약

첫째, 한국어의 음절구조와 관련된 제약들을 (42)에 요약하였다.

(42) 음절구조 제약 ($=음절 경계)

번호	제약	예
a	음절두음 위치(어두, 자음 뒤)에 자음군이 금지된다.	*$[tɾ]
b	음절두음 위치(어두, 자음 뒤)에 [ŋ]가 오지 못한다.	*$[ŋɑ]

(42)의 제약은 어두 및 자음 뒤, 즉, 음절두음 위치에 자음군 또는 [ŋ] 발생 금지를 요구한다. 이 제약에 따라, 한국어 화자는 자음군(예: 영어 strike) 또는 [ŋ](예: 베트남어 người)로 시작하는 단어를 들을 때, [ɨ]를 삽입하여 듣는다(예: [sɨtɨr], [ɨŋ]).

둘째, [활음][모음]과 관련된 음소배열제약을 (43)에 요약하였다.

(43) [활음][모음] 관련 제약

A. 범주적 [활음][모음] 제약

번호	제약	예
a	[ji], [w][u, o]가 금지된다.	*[wu]
b	[j, w][ɨ]가 금지된다.	*[jɨ]
c	[ɨj] 외 하향 이중모음은 금지된다.	*[ɨw]
d	[ɯ]에 후행하여, [i] 외 모음이 오지 못한다.	*[ɯa]

B. [자음]//[활음][모음] 제약(A//B: AB 혹은 BA)

번호	제약	예
a	이중모음 [ɨj/ɯi, je, jɛ] 앞에 음절두음이 오지 못한다.	*$[자음][ɯi, je, jɛ]
b	이중모음 [ɨj/ɯi, je, jɛ] 뒤에 음절말음이 오지 못한다.	*[ɯi, je, jɛ][자음]$

활음과 모음이 성절성([±syllabic])을 제외한 자질이 동일한 경우, 활음과 모음 연쇄가 허용되지 않으며(43Aa), 모음 [ɨ] 앞에 활음이 금지된다(43Ab). (43Ac–d)는 이중모음 '의'와 관련된 제약으로, '의'의 지위에 따라 달리 정의될 수 있다. '의'를 하향 이중모음 [ɨj] (정연찬 1991)로 본다면, 한국어에서 [ɨj] 외 하향 이중모음을 금지하는 제약을 정할 수 있다(43Ac). 한편, '의'를 상향 이중모음 [ɯi]으로 보는 관점(신지영 1999, 김무식 2001, 김영선 2007)을 따른다면, 활음 [ɯ]의 실재를 인정하고 [ɯ]에 후행하여 [i] 외 모음이 오지 못하는 제약을 설정할 수 있다(43Ad).

이 외에도 [활음][모음]이 [자음]과 결합하는 경우의 제약이 서술되었다. (43B)에서 보듯이, [자음]은 [ɨj/ɯi, je, jɛ]에 선·후행하여 오지 못한다. 다만, 연구자마다 [음절두음][활음]의 실현 정도를 다르게 판단한다. 신지영·차재은(2003)은 기저형에서 음절두음과 이중모음 [ɨj/ɯi, je, jɛ]의 연쇄가 아예 불가능하다고 보는 한편,

허웅(1985), 이진호(2014)는 표면형에서 이중모음이 불안정하게 실현된다고 서술한다.[6]

셋째, 필수적 음운 규칙과 관련된 음소배열제약을 (44)에 요약하였다.

(44) 필수적 음운 규칙과 관계된 제약[7]

번호	제약	음운 규칙	규칙 적용의 예	
a	*[저해음][평음]	저해음 뒤 경음화	입-고	/ip-ko/ → [ip.k'o]
b	*[저해음][h]	격음화	축하	/c^{h}uk-ha/ → [c^{h}u.k^{h}a]
c	*[격음, 경음]$	음절말 저해음 중화	잎	/ip^h/ → [ip]
			쫓-고	/c'oc^h-ko/ → [c'ot.k'o]
d	*[자음][자음]$	음절말 자음군 단순화	값	/kap^h/ → [kap]
			밟고	/palp-ko/ → [pal.k'o]
e	*[저해음][공명음]	비음화	먹-는	/mʌk-nɨn/ → [mʌŋ.nɨn]
			독립	/toklip/ → [toŋ.nip]
f	*[m, ŋ][l]	비음화	음료	/ɨmlyo/ → [ɨm.nyo]
g	*[nl]	비음화	생산-량	/sɛŋsan-lyaŋ/ → [sɛŋ.san.nyaŋ]
		유음화	관리	/kwanli/ → [kwal.li]
h	*[ln]	유음화	달님	/tal-nim/ → [tal.lim]
i	*[구개음][j]	구개음 뒤 j 탈락	쳐	/c^{h}jʌ/ → [c^{h}ʌ]

[6] 표준국어대사전에서는 이중모음별 발음을 다음과 같이 제시한다.

a. [음절두음]+[je]: [j] 실현/[j] 탈락 (예: 계[계:/게])

b. [음절두음]+[jɛ]:

- 발음 정보 제시 안함 (예: 걔, 재)
- [j] 실현 (예: 뒷얘기[뒨:냬기])

c. [음절두음]+[ɨj]:

- [ɨ] 탈락 (예: 무늬[무니], 희망[히망])
- [ɨ] 실현 (예: 협의[혀븨/혀비], 산의[사늬/사니])

[7] (44)의 제약 및 규칙들 중 일부는 적용이 필수적이지 않다고 관찰한 연구들이 있다. 예를 들어, (44d)의 '자음군 단순화'와 관련해서 용언 어간말에 위치한 자음군 [lp, lk](Kim-Renaud 1974, Cho 1999, Cho & Kim 2009)은 모두 실현될 수 있음이 관찰된다. (44f-h)의 비음화 규칙과 관련하여서도, [비음][n]뿐만 아니라 [비음][l]로 실현됨이 보고된다(Jun 2000, 서윤정 2016). 그러나 어휘부를 대상으로 음소배열 제약을 탐색할 때 수의적인 음운 규칙의 실현 정도를 세밀하게 반영하는 것에는 한계가 있어, 기존 음소배열제약 논의에서도 이러한 점은 반영되지 않았다. 이 연구도 이에 따라 사전 발음형을 기준으로 논의한다.

특정 연쇄는 기저형 층위에 발생하더라도 음운 규칙이 필수적으로 적용됨에 따라 표면형에서 발생하지 않는다. 다수의 연구는 이러한 연쇄를 음운론적 교체의 동기로 관련지으면서, 회피 제약으로 서술한다. 예를 들어 음운 규칙 '저해음 뒤 경음화(44a)'의 동기는 [저해음]에 후행하여 [평음] 발생을 회피하는 제약으로 설명된다. 또한, '유음화 및 비음화(44g)'의 동기는 [nl]의 발생을 저지하는 제약으로 서술된다.

넷째, 고유어와 한자어에서는 발생이 금지되지만, 최근 차용어에서만 발생이 허용되는 분절음(연쇄)에 대한 제약들을 (45)에 요약하였다.

(45) 최근 차용어에서만 예외를 인정하는 제약

번호	제약	예
a	[l]로 시작하는 단어가 없다.	*#[l]
b	어두에 [ni, nj]가 금지된다.	*#[ni, nj]
c	[ɨ]로 끝나는 단어가 없다.	*[ɨ]#

이 제약은 모두 단어 경계에서 정의된다. 어두 위치에 [l] 음소와 [ni, nj] 연쇄가 금지된다(45a–b). 그리고 어말 위치에 [ɨ] 음소가 오지 못한다(45c).

다섯째, 고유어와 한자어에서 예외가 존재하는 제약을 (46)에 요약하였다. 다수의 연구가 해당 연쇄가 일부 형태·음운론적 조건에서만 발생하다는 것을 밝히고 해당 연쇄의 제한적 발생을 제약으로 정의하였다. 또한, 일부 연구는 특정 연쇄 빈도가 극히 낮다는 것을 계량적 조사를 통해 객관적으로 제시하고자 하였다.

(46) 고유어와 한자어에서 예외가 존재하는 제약

번호	제약	예외
a	[유음]에 후행하여 [설정 평음] 발생이 회피된다.	털실, 몰지각
b	[양순음]이 [ɨ] 앞에 오지 않는다.	예쁘-[jep'ɨ-]
c	[w]는 [양순음]에 후행하지 않는다.	봐[pwa]
d	[j]는 [설정 저해음]에 후행하지 않는다.	디뎌[titjʌ]

먼저 (46a)에서 제시하였다시피, 유음에 후행하여 [설정 평음] 발생이 회피된다. 고유어에서는 유음에 후행하여 설정 경음 및 격음이 올 수 있지만(예: 갈치, 글씨), 설정 평음의 발생은 저지된다(고광모 1996). 또한, 한자어에서는 유음 뒤 설정 평음이 오면, 설정 평음이 주로 경음으로 발생한다고(예: '결단[kʌlt'an]', '발

상[pals'aŋ]') 보고된다(권인한 1997, 신지영·차재은 2003). 이 제약은 주로 복합어에서 예외가 허용되는데, 그 예외로 고유어 '칼질, 털실, 돌다리'(고광모 1996)와 한자어 '별도리, 고별식, 몰지각'(이호영 1996: 163) 등이 관찰되었다. 이 제약의 동기는 형태소 내부 [t, s, c]에 선행하는 [ɾ]이 [l]로 미파화되고, 이에 따라 [t, s, c] 경음화되는 통시적 변화로 주장되기도 하였다(엄태수 1988, 고광모 1996). 다른 한편으로는 S. Kim (2016)은 고유어 합성어 현상('사잇소리 현상')에서도 유음에 후행하는 [t, s, c]가 [p, k]보다 더 빈번하게 경음화되는 것이 관찰하며(예: 돌솥[ls] vs. 발굽[lt]), 현대 한국어 합성어 경음화에서도 해당 제약의 유효성을 제기한다.

다음으로, [양순음][ɨ]의 연쇄 발생이 제한된다(46b). 어두에서는 발생 빈도가 0이고, '예쁘-[jep'ɨ-], 슬프-[sɨlp^{h}ɨ-]'와 같이 비어두 위치에서만 예외가 발생한다(신지영·차재은 2003). 해당 연쇄의 낮은 상대 빈도가 계량적 연구(진남택 1992, 유재원 1997, 김미란 외 2014)에서도 보고되었다. 이 제약은 양순음에 후행하는 [ɨ]가 [u]로도 실현되는 현상, 즉, 조음상의 경제성을 위한 원순성 자질동화(석주연 1996, 김경아 1996, 2001, 유필재 2001, 신우봉 2010)와 관련지어 해석된다. 해당 동화 현상은 통시적으로도 발생한 것으로 보이는데, 허웅(1985)에 따르면 일부 [양순음][ɨ] 연쇄는 17–18세기 원순모음화(예: 믈[水]>물[水])를 겪었다.

다음으로, [m, p, p', p^h][w] 연쇄는 고유어 어휘 형태소에서는 발생하지 않고 한자어(예: 입원[ipwʌn]), 용언 활용(예: 봐[pwa])등에서만 제한적으로 관찰된다((46c), 신지영·차재은 2003, 강옥미 2011). 일부 연구(유재원 1997, 김미란 외 2014)는 [양순음][w]의 낮은 상대 빈도를 제시한다. 강옥미(2011)은 제약의 동기를 [양순음] 자질과 [w]의 [원순음] 자질이 동일/유사하기 때문으로 보고 OCP 제약(동일 자질 회피 제약)의 작용으로 분석한다. 이 제약은 한자어, 용언에서 발생하는 [양순음][w] 연쇄에서 [w]가 수의적으로 탈락되는 현상으로도 정당화된다(Kang 1998, 구희산·한혜승 1999, 강옥미 2011).

또한, [설정 저해음] 뒤에 [j]가 잘 오지 못한다(46d: 허웅 1985, 신지영·차재은 2003).[8] 용언 활용형(예: 디디+어 → 디뎌)에서 예외가 허용되나, 계량적 연구(유

[8] 발생 제한의 정도는 선행 자음의 종류에 따라 다르다. 구개 파찰음 [c, c', c^h]에 후행하여 [j]가 전혀 실현되지 않고, [t, t', t^h, s, s']과 [j] 또한 어휘 형태소 내에서는

재원 1997, 김미란 외 2014)는 해당 연쇄의 상대 빈도가 매우 낮은 것을 밝힌다. 이는 [양순음][w]와 마찬가지로, [설정 저해음]과 [j]의 자질이 동일/유사하기 때문으로 분석된다(강옥미 2011). 통시적인 변화의 결과로도 이해되는데, 17세기 초 구개음화에 따라 그 당시 [t, t^h][j]는 [c, c^h][j]로 바뀌었으며 1800년 전후 [s, c, c^h]에 후행하는 [j]가 탈락되는 것도 관찰되었다.

지금까지 한국어에서 작용하는 것으로 기존연구에서 보고한 제약들을 검토하였다. 대부분의 제약은 발생 빈도가 0인 분절음에 대해 정의되었으며, 음운론적 동기 위주로 논의되었다. 그러나 일부 발생 빈도가 0은 아니지만 발생 빈도가 낮은 연쇄들이 제약으로 포착되기도 하였으며, 이를 위해 계량적 기준이 부분적으로 도입되기도 하였다. 이에 비해, 예외를 다수 허용하거나 음운론적 동기가 분명하지 않은 연쇄의 회피, 즉, 비범주적 음소배열제약은 주된 연구 대상 및 형식화 대상이 아니었다. 그러나 일부 연구는 계량적 방법론을 적용하여, 범주적 또는 음운론적으로 정의되지 않는 회피 연쇄를 탐색하였다. 기존연구에서 논의된 비범주적 음소배열제약을 아래 3.2절에서 논의한다.

3.2. 비범주적 음소배열제약(전체 한국어 및 고유어 어휘부)

기존 계량적 연구는 전체 한국어 어휘부 또는 고유어 어휘부를 대상으로 전통적인 제약 외에도 발생 빈도가 유의미하게 낮은 연쇄를 관찰한다. 이 중 일부는 범주적 음소배열제약과 마찬가지로 음운론적으로 자연스러우며, 음운론적 교체 현상의 직·간접적인 환경으로 기능하기도 한다. 아래에서 인접 연쇄 유형과 비인접 연쇄 유형을 나누어 살펴본다.

3.2.1. 인접 연쇄

인접 연쇄에 대해서 정의된 제약들을 (i) 기존연구에서 음운론적 동기가 제시되고 논의된 것, (ii) 기존연구에서 음운론적 동기가 제시되지 않은 것, (iii) 기존연구에서 논의된 [격음]과 [경음]의 제한적 분포, 세 부류로 나누어 순서대로 살

연속해서 발생하지 않지만, 용언 활용 등에서 예외적으로 발생한다(예: 디뎌[titjʌ], 모셔[mosjʌ], 허웅 1985).

펴본다. 첫 번째로, 기존연구에서 회피 동기가 제시된 경향들을 (47)에서 요약한다.

(47) 비범주적 제약: 음운론적 동기를 갖는 인접 연쇄 제약

번호	연쇄 회피	제약 형식
a	치경 저해음이 [i] 앞에 오지 않는다	*[t, t’, t^h][i]
b	[구개음]이 [ɨ] 앞에 오지 않는다.	*[c, c’, c^h][ɨ]
c	[원순모음]에 후행하여, [양순음]이 회피된다.	*[o, u][m, p]
d	모음 연쇄가 회피된다.	*[모음][모음]

(47a)부터 살펴보면, 전체 한국어 및 고유어에서 [t, t’, t^h][i] 연쇄의 회피가 보고된다. 제약을 범주적으로만 다룬 연구들에서는 *[t, t’, t^h][i] 제약을 형태·음소 교체 현상인 구개음화(예, 맏-이 [마지])에 근거해서 형태소 경계에서만 유효한 것으로 가정되었다. 그러나 진남택(1992)는 전체 한국어 어휘부를 대상으로 [i]에 대한 [자음] 일반([자음][i])의 비율이 16.8%인데 비하여, [t, t’, t^h][i] 연쇄 비율은 1.2%로 낮다는 것을 보고하며, Chong (2017)도 [t, t’, t^h][j, i]의 O/E 비율이 고유어, 한자어, 차용어 어휘부에서 모두 1보다 낮아, [t, t’, t^h][j, i]의 회피 경향이 존재한다고 주장한다.

[c, c’, c^h][ɨ] 연쇄의 회피(47b)도 관찰된다. 진남택(1992)는 [ɨ]에 대한 자음일반([자음][ɨ])의 비율이 13.4%인데 비하여 [c, c’, c^h][ɨ] 연쇄 비율은 1.8%로 낮아 [c, c’, c^h][ɨ] 연쇄의 회피 경향이 존재한다고 기술한다. 김미란 외(2014)에서도 전체 한국어 어휘부를 대상으로 [ɨ]에 대한 [c, c’, c^h]의 상대 빈도를 0.05% 이하로 계산하고 [c, c’, c^h][ɨ] 연쇄의 회피를 확인한다.

*[t, t’, t^h][i] 제약과 *[c, c’, c^h][ɨ] 제약은 동화(assimilation)와 관련하여 분석된다. 유형론적으로 설정 저해음 [t, t’, t^h]는 후행하는 [j, i]의 전설성을 닮아 구개 파찰음 [c, c’, c^h]이 되는 경향이 있다(Gordon 2016). 또한, [ɨ]는 구개 파찰음 [c, c’, c^h]의 구개성을 닮아 전설 모음 [i]가 되는 것이 자연스러우며, 일부 방언에서 전설모음화(서남 방언 예: 맺-+-으X → 메징께, 동남 방언 예: 쫓-+-으X → 쪼치무)가 일어나기도 한다(홍은영 2012). 통시적으로도 18세기 구개음화(예: 디-[ti-] > 지-[ci-])와 19세기–20세기 초 전설모음화(예: 어즈럽-[ʌcɨlʌp-] > 어지럽-[ʌcilʌp-])가 발생하였다.

[원순모음]에 후행하여 [양순음]이 잘 나타나지 않는 경향(47c)도 관찰된다. 유재원(1997)은 전체 한국어를 대상으로 [원순모음][자음][9] 비율(25.05%)보다 [원순모음][양순음] 연쇄의 발생 비율(15.43%)이 낮아 [원순모음][양순음] 제약을 설정한다. 한자어 음절에서도 '품'과 같은 예외가 발생되지만, [원순모음][양순음] 제약을 정의한다(신지영 2009). 반면, 고유어 어휘부에서는 관련 제약이 기술되지는 않았다. [원순모음][양순음] 제약 설정은 [원순음] 자질과 [양순음] 자질이 유사하여 회피되는 동기로 뒷받침한다. 실제로 해당 연쇄는 많은 언어에서 회피된다. 광동어의 운모에서는 [ow], [øy] 외 [원순모음][양순음] 연쇄가 관찰되지 않으며(Yu 2017), Tashlhiyt Berber어에서도 원순모음 뒤에 오는 양순성이 탈락된다(**aqʷ**lil > **uq**lil 'rabbit' , Gordon 2016, Odden 1994: 317 재인용).

또한, [모음][모음] 연쇄의 회피 경향(47d)에 대해서, 유재원(1997)은 '표준한국어 발음 대사전(1993, 한국방송공사)'의 표제어 자료를 바탕으로, 두 분절음 연쇄 397,064개 중에서 [모음][모음] 연쇄 빈도가 0.87%에 불과하다고 보고하고 있다. 고유어 ·소에서는 모음 연쇄의 회피 경향이 더욱 강하다(하세경 2000, 신지영·차재은 2003). 이와 관련해서, 필수적 또는 수의적인 모음충돌회피 현상이 생산적으로 발생한다(J. Kim 2000, 하세경 2000). 아래 (48)과 같이, 모음으로 끝나는 어간과 모음으로 시작하는 어미가 결합할 때 모음 탈락 및 활음화 등의 현상이 생산적이며, 어휘 형태소 내부에서는 축약 현상이 산발적으로 발생한다.

(48) 모음충돌회피 현상의 예

a. 용언 활용

- 모음 탈락 　 뜨-어 　 /t'ɨ-ʌ/ → [t'ʌ] 　 *[t'ɨʌ]
- 활음화 　 주-어 　 /cu-ʌ/ → [cwʌ] ~ [cuʌ]

b. 형태소 내 줄임말

- 축약 　 아이 　 /ai/ → [ɛ] ~ [ai]
- 탈락 　 싸움 　 /s'aum/ → [s'am] ~ [s'aum]

언어 보편적으로도, 모음 연쇄가 잘 나타나지 않는 경향이 있으며 모음충돌회피 현상이 자연스럽게 발생한다. 이러한 경향은 최적성 이론에서 *VV 제약(두

[9] 음절말음과 음절두음을 모두 포함한다.

모음의 연쇄를 허가하지 않는다, Rosenthall 1994, Casali 1996) 또는 ONSET(모든 음절은 음절두음을 가져야 한다, Prince & Smolensky 1993) 제약으로 포착된다.

모음 연쇄를 구성하는 모음의 종류에 따른 상대적인 빈도 차이는 유재원(1997)에서 제한적으로 탐색되었다. 분절음 단위로 보면, 고모음 [i, u]가 다른 모음 앞에서 나타날 수 있는 반면, [ɨ]는 다른 모음 앞에서 거의 나타나지 못한다. 또한, [e, ɛ] 앞뒤에서는 다른 모음이 잘 나타나지 못한다(유재원 1997: 104, 표 5-18). 자질 단위로 보면, 동일한 자질을 공유하거나 유사한 자질을 가진 모음 연쇄가 회피된다. 동일한 [평순/원순성], [전/후설성] 연쇄에 대한 회피를 밝히는 한편, [원순모음]과 [후설모음]의 연쇄의 발생 제한을 보고한다. 그러나 모음의 높이 자질에 따른 회피와 선호 경향은 뚜렷하지 않다고 보았다.

두 번째로, 다수의 연구에서 음운론적 동기를 언급하지 않고 제시한 회피 경향을 (49)에 요약하였다.

(49) 비범주적 제약: 음운론적 동기가 불분명한 인접 연쇄 제약

번호	연쇄 회피	제약 형식
a	공명음은 [w] 앞에 잘 오지 못한다.	*[공명음][w]
b	치경음은 [w] 앞에 잘 오지 못한다.	*[n, l, t, t’, t^h][w]
c	[p^h]는 [e, ʌ] 앞에 잘 오지 못한다.	*[p^h][e, ʌ]
d	[k^h]는 [ʌ] 앞에 잘 오지 못한다.	*[k^hʌ]
e	[e]에 후행하여 음절말음이 잘 오지 못한다.	*[e][음절말음]
f	[ɛ]에 후행하여 공명말음이 잘 오지 못한다.	*[ɛ][공명말음]

[공명음] 또는 [치경음]이 [w] 앞에 잘 발생하지 않는다(49a–b: 허웅 1985, 유재원 1997). 또한, [p^h, k^h]와 [e, ʌ]의 연속이 회피되고(49c–d: 신지영·차재은 2003, Cho 2012), [e]를 포함하는 폐음절과 [ɛ][공명 자음]을 포함하는 폐음절이 회피된다(49e–f).

세 번째로, 다수의 연구가 관찰한 [격음]과 [경음](이하, ‘후두자음’)의 제한적인 분포를 (50)에 요약한다(고유어: 김경일 1985, 전체 한국어: 유재원 1997, Cho 2012).

(50) 비범주적 제약: [격음]과 [경음]의 제한적 분포에 대한 인접 연쇄 제약

번호	연쇄 회피	제약 형식
a	[저해음]에 후행하여, [격음]이 회피된다.	*[저해음][격음]
b	[비음]에 후행하여, [경음]이 회피된다.	*[비음][경음]
c	[모음]에 후행하여, [격음]과 [경음]이 회피된다.	*[모음][격음, 경음]

[저해음]에 후행하여 [격음] 발생이 저지되는 한편(50a), [비음][경음] 연쇄와 [모음][격음, 경음] 연쇄의 회피 경향(50b–c)이 보고되었다.

이제까지 인접 연쇄에 대한 비범주적 음소배열제약을 살펴보았다. 범주적이며 음운론적으로 정의된 제약에 비해, 이 절에서 다룬 제약들은 예외를 더 허용하며 그 일부는 음운론적 동기가 제시된 바도 없고 불분명해 보인다. 그러나 복수의 연구에서 같은 경향이 반복적으로 관찰된다는 점을 고려할 때, 이들이 한국어 화자의 인식에 실재할 가능성도 연구의 대상이라고 할 수 있다.

3.2.2. 비인접 연쇄

일부 연구는 자음 자질 층위 또는 모음자질 층위를 별도로 설정하고, 비인접 연쇄에 대한 O/E 비율 및 상대 빈도를 계산하였다. 아래 (51)에 요약한 바와 같이, 비인접 연쇄에 대한 회피 경향은 (i) 모음자질 층위, (ii) 자음의 조음위치자질 층위, 그리고 (iii) 음절두음의 후두자질 층위 각각에 대해 탐색되었다.

(51) 비인접 연쇄 회피

번호	연쇄 회피	제약 형식
모음 층위: ATR(또는 양/음성)		
a	ATR(또는 양/음성)의 부조화형이 회피된다.	*[αATR][−αATR]
		*[양성]//[음성]
자음 층위: 조음위치자질 ([음절두음][음절말음], [음절두음][음절두음])		
b	동일한 조음위치자질 연쇄가 회피된다.	*[αplace][αplace]
음절두음 층위: 후두자질 공기		
c	[격음] 자질 공기가 회피된다.	*[+aspirate][+aspirate]
d	[격음]과 [경음] 자질 공기가 회피된다.	*[+tense]//[+aspirate]

이하에서 (51)에서 제시한 비인접 연쇄에 대한 회피 경향을 순서대로 살펴본다. 첫째, (51a)에 제시한 바와 같이 모음 층위에서의 회피 경향이 보고되었다.

김경일(1985)와 Hong (2010)은 모두 고유어를 대상으로, 모음 공기 관계를 조사하였다. 김경일(1985)는 모음 조화의 자질을 [양성]과 [음성]으로 설정하고, [i]를 조화/비조화에 참여하지 않는 중립 모음으로 보았다. 아래 (52)는 이러한 세 부류의 모음을 대상으로 첫 번째에 위치한 모음에 대한 두 번째 모음의 상대 빈도를 구한 것이다.

(52) 비인접 [모음][모음] 비율(김경일 1985: 34, 표 15)

σ_{n+1} / σ_n		양성			음성				i	계
		a	ɛ	o	ʌ	e	u	ɨ		
양성	a	0.502			0.270				0.228	1
	ɛ	0.481			0.249				0.271	1
	o	0.506			0.259				0.235	1
음성	ʌ	0.261			0.519				0.220	1
	e	0.315			0.400				0.286	1
	u	0.199			0.585				0.216	1
	ɨ	0.272			0.524				0.202	1
i		0.309			0.525				0.167	1

그 결과, 조화형의 평균 발생 비율([양성][양성]: 0.494, [음성][음성]: 0.507)은 부조화형의 평균 발생 비율([양성][음성]: 0.259, [음성][양성]: 0.262)보다 높다는 관찰을 하였다.

한편, Hong (2010)은 모음 조화 관련 자질로 '양성/음성'을 대신하여, [ATR](Advanced Tongue Root; Y-S. Kim 1984, J-S. Lee 1992, Y-S. Lee 1993)을 채택하고, [+ATR] 자질에 대하여 [+/−high] 자질을 추가적으로 명세하였다. 그리고, 세 개의 모음을 포함하는 단어에서 위치별로 두 개 모음 사이의 O/E 비율을 계산하였다. 보고한 O/E 비율 중 비조화형 [αATR][−αATR] O/E 비율만을 (53)에서 제시하였는데, 비조화형 O/E 비율이 대부분 1보다 낮아 비조화형의 과소표상을 나타내는 것을 알 수 있다.

(53) 비인접 모음 연쇄: 비조화형 O/E 비율[10]

위치 \ 연쇄	[+high,+ATR][−ATR]	[−high,+ATR][−ATR]	[−ATR][−high,+ATR]
V_1V_2	0.69	0.64	0.43
V_2V_3	1.07	0.63	0.42
V_1V_3	0.77	0.75	0.61

다만, 세 모음 연쇄 중 두 번째 모음(V_2)에 [+high,+ATR], 세 번째 모음(V_3)에 [−ATR]이 위치하는 경우에는 O/E 비율이 1에 가깝다. 이는 모음 비조화 유형이 모든 위치에서 일괄적으로 회피되지 않음을 보인다.

이와 같이 고유어 어휘부 내에서 모음 층위에 국한하여 모음 조화 자질 비조화형의 상대적인 회피 경향이 보고되었다. 그러나, 조화형과 비조화형의 상대 빈도 차이 또는 O/E 비율 차이가 통계적으로 유의미한지에 대해서는 분명히 밝히지 않았다. 더구나 현대 한국어에서 모음 조화를 포착하는 자질은 [ɛ, a, o]와 나머지 모음들을 기술적으로 표시하기 위해 도입된 경향이 있는 한편(Jun 2018), 용언 활용에서 제한적으로 발생하며 지속적으로 규칙의 생산성이 줄어들고 있다(Kang 2012, Jang 2016). 이러한 배경에서, 인접 제약과의 상호작용 가운데에서도 모음 비조화형에 대한 제약이 어휘부에서 유의미하며, 이를 화자들이 인식할 수 있을지에 대한 검토도 필요하다.

둘째, 자음의 조음위치자질(place feature)의 공기 관계에 대해 김경일(1985)와 Ito (2007)은 모두 고유어를 대상으로 동일 조음위치자질 회피 제약의 효과 유무 및 정도를 탐색하였다. Ito (2007)은 단음절어에서 음절두음과 음절말음이 동일한 조음위치자질을 갖지 않는다고 보고한다. (54)에 제시한 바와 같이 O/E 비율을 채택하여, [양순음][양순음], [설배음][설배음]뿐만 아니라, [설정 저해음][설정 저해음]이 과소 표상된다는 것을 밝히며, 이를 동일 자질 회피 제약(OCP)의 효과로 분석하였다.

[10] Hong(2010: 288–290)에서 제시된 O/E 비율 중 모음 조화 자질의 비조화형에 해당하는 부분만 발췌하였다.

(54) 조음위치자질 공기 관계: [음절두음][음절말음] O/E 비율(Ito 2007)

	[양순음][양순음]	[설정 저해음][설정 저해음]	[설배음][설배음]
O/E	0.65	0.71	0.48

Ito (2007)은 다수의 언어에서 동기관적 자음 회피 제약이 유효함을 근거로, 이 제약이 음운론적으로 자연스럽고 보편적이라는 점을 지적한다(아랍어: Frisch et al. 2004, 자바어: Mester 1986, 러시아어: Padgett 1995, 일본어: Kawahara et al. 2006, Muna어: Coetzee & Pater 2008).

이에 앞서, 김경일(1985)는 상대 빈도를 채택하여, 이음절어의 음절두음들 사이에서도 동일 조음위치자질이 회피되는 것을 발견하였다. [자음][양순음] 연쇄의 발생 빈도를 1로 보면, [양순음][양순음] 연쇄의 발생 비율은 0.136로 낮으며, [자음][설배음]의 비율을 1로 볼 때, [설배음][설배음] 비율은 0.117로 낮음을 보이고 있다.

(55) 조음위치자질 공기 관계: [음절두음][음절두음] 상대 빈도(김경일 1985)

음절두음$_2$ / 음절두음$_1$	양순음	설정음	설배음	…
양순음	**0.136**	0.234	0.228	…
설정음	0.413	**0.375**	0.448	…
설배음	0.250	0.207	**0.117**	…
…	…	…	…	…
계	1.000	1.000	1.000	1.000

한편, 동일 조음위치자질 회피 제약이 언어 처리에 영향을 미칠 수 있다는 점도 논의되었다. Kang (2015)는 한국어 화자는 음절두음과 음절말음의 조음위치자질이 동일한 경우 해당 음절을 잘 기억하지 못함을 밝혔다.

이처럼 동일 조음위치자질 회피 제약의 실재가 시사되었지만, 회피 경향에 대한 기술은 단음절어 또는 이음절어에 국한되었으며, 통계적 유의미성은 충분하게 뒷받침되지 못하였다. 동일 조음위치자질 회피 제약이 음절수를 제한하지 않은 어휘부에 대해서도 제약이 유의미하게 포착될 수 있는지에 대한 검토가 필요하며, 적형성 인식 조사를 통해 그 실재가 뒷받침되어야 한다.

셋째, ‘후두자질 공기 제한(laryngeal co-occurrence restriction)’이 보고되었다. ‘후두자질 공기 제한’이란, 후두자질이 다른 후두자질의 발생을 저지하거나 선호

하는 것을 뜻한다. 한국어 어휘부를 대상으로 어두에 위치한 [경음]과 [격음]의 공기 관계를 탐색한 연구들이 있다(김경일 1985, Ito 2014, Kang & Oh 2016). (56)–(57)을 보면, 이 기존연구들은 [경음][경음] 선호를 공통적으로 포착한다.

(56) 후두자질 공기 관계: [음절두음]$_1$[음절두음]$_2$ 상대 빈도(김경일 1985)

음절두음$_2$ / 음절두음$_1$	평음	경음	격음	…	Ø
평음	0.546	0.459	0.586	…	0.557
경음	0.066	**0.116**	0.090	…	0.076
격음	0.098	0.116	0.090	…	0.108
…	…	…	…	…	0.130
계	1.000	1.000	1.000	1.000	1.000

(57) 후두자질 공기 관계: [음절두음]$_1$[음절두음]$_2$ O/E 비율

	[격음][격음]	[격음][경음]	[경음][격음]	[경음][경음]
고유 단일어 Ito 2014	**0.44**	**0.80**	0.94	**2.30**
전체 한국어 Kang & Oh 2016	1.07	**0.33**	0.95	**2.35**

그러나 그 밖의 공기 관계는 대상 어휘부 및 측정 방식에 따라 달리 예측된다. [격음][경음] 회피는 고유 단일어와 전체 한국어에서 O/E 비율로 포착되고, [격음][격음] 회피는 고유 단일어에서 O/E 비율에서만 확인되었다(Ito 2014).

후두자질 공기 제약은 다수의 언어에서 정적인(static) 음소배열제약의 형태로 관찰되었다(MacEachern 1999, Gallagher 2010). 그러나 한국어에 대해서는 후두자질 공기 관계가 동적인 음운교체 과정인 '합성어 경음화'와 '어두 경음화'에도 관찰될 수 있다(Ito 2014, Kang & Oh 2016, S. Kim 2016, H. Kim 2017). (58)에서 예를 든 바와 같이 '합성어 경음화'란 두 어근이 결합하여 합성어를 이룰 때, 두 번째 어근을 시작하는 평음 저해음이 경음이 되는 합성어 경계 표지 현상이며, '어두 경음화'는 고유어 어두 평음이 의미 변화를 수반하지 않으면서 경음으로 실현되는 수의적인 음운 현상이다.

(58) 후두자질이 관여된 형태·음운론적인 교체

a. 합성어 경음화(사잇소리 현상)

- 말+솜씨 /mal+soms'i/ → [mal**s'**oms'i] ~ [malsoms'i]
- 물+갈퀴 /mul+kalk^{h}y/ → [mul**k'**alk^{h}y] ~ [mulkalk^{h}y]

b. 어두 경음화

- 곱빼기 [kopp'εki] ~ [**k'**opp'εki]

기존연구(Ito 2014, S. Kim 2016)가 한국어 화자를 대상으로 두 음운 과정 발생 비율을 조사한 결과, '합성어 경음화'는 어근에 [경음] 또는 [격음]이 존재하면(선행/후행 어근: Ito 2014, 후행 어근: S. Kim 2016), [경음] 발생이 저지되었다. 반면, 어두 경음화(Kang & Oh 2016, H. Kim 2017)는 후행하는 음절두음이 경음인 경우 더 활발하게 발생한다. Kang & Oh (2019)는 이러한 경향을 아래 (59)와 같이 요약하였다.

(59) 경음화 발생 비율에 미치는 후행 자음 효과(Kang & Oh 2019: 7, 표 1)[11]

C_2	합성어(합성어 경음화)			단일어(어두 경음화)	
	Ito (2014)		S. Kim (2016)	Kang & Oh (2016)	H. Kim (2016)
	실제 단어	비단어	실제 단어	실제 단어	실제 단어(차용어)
경음	▼	▼	▼	▲	▲
격음	▼	▼	▼	-	-

Kang & Oh (2019)는 어두 경음화에서 '[경음][경음]'이 선호되는 것은 (56−57)에서 제시한 정적인 음소배열제약에 부합하나 합성어 경음화에서는 [경음][경음] 회피가 드러나 정적인 음소배열제약과 상충된다는 점을 지적하였다. 이러한 차이를 적형성 판단 조사를 통해 재확인하였고, 단일어와 합성어의 차이로 분석하였다.

이와 같이, 복수의 연구가 한국어 화자를 대상으로 형태·음운론적 교체에서 드러나는 후두자질 공기 관계를 직접 조사하였다. 그러나 정적인 후두자질 공기 관계에 대해서는 화자의 인식이 직접 탐색되지 않았다. 이에 따라, 비인접 연쇄

[11] '▼'는 발생 저지, '▲'는 발생 촉진을 뜻한다. '-'은 발생에 영향을 미치지 않음을 뜻한다.

의 회피를 엄밀하게 포착하고, 정적인 후두자질 공기 제약의 적형성을 검토할 필요가 있다.

지금까지 비인접 연쇄에 관한 회피 경향성을 살펴보았다. 이러한 경향성은 음운론적으로 자연스러운 것으로 논의되었으며, 형태·음운론적 교체 현상과 관련되기도 한다. 이러한 경향을 밝히기 위해서, 기존연구는 개별 비인접 연쇄만을 대상으로 해당 연쇄의 회피 정도를 계산하였다. 이 경우 다른 인접/비인접 제약의 관여 가능성까지는 고려되지 않는다.

이제까지 살펴보았듯이, 다수 연구가 한국어 전체 한국어 또는 고유어를 중심으로 인접/비인접 연쇄에 대한 회피 경향을 조사하였다. 보고된 회피 경향은 다수의 연구가 어휘부에서 직접적으로 관찰할 수 있을 만큼 뚜렷하다는 것으로 볼 수 있으며, 이를 화자가 인식할 가능성도 시사한다. 그러나 이와 같이 보고된 회피 경향은 범주적 음소배열제약이 포착하는 일반화와는 달리 예외를 다수 허용하는데, 회피 경향 자체도 통계적으로 뒷받침되지 않아 한국어 화자의 제약으로 확정하기 어려운 측면이 있다.

3.3. 한자어 어휘부 특정 제약

다수의 기존연구(권인한 1997, 강용순 1998, 신지영·차재은 2003, 신지영 2009, 안소진 2009)는 한자어 특정적인 음소 분포 및 회피 경향을 관찰하였다. 그러나 관찰 범위가 주로 음절 단위에 국한되었으며, 절대 빈도 보고(신지영 2009, 마야 아타예바 2016) 외에는 충분한 양적 탐색이 이루어지지는 않았다. 기존연구에서 서술한 제약과 일반화를 (i) 음절 내에서 기술된 것, (ii) 음절 경계에서 기술된 것으로 나누어 논의한다. 덧붙여 고유어 제약과 한자어 제약의 차이를 바탕으로, 한자어 어휘부를 별도로 상정한 어휘부 계층 이론을 소개한다.

첫째, 한자어 음절은 고유어 음절보다 음소 분포가 제한된다. 아래 (60)에 요약하였듯이 한자어 음절에 대해 총 다섯 개의 제약이 보고된다.

(60) 한자어의 음절 내 음소 분포 및 회피 연쇄

번호	연쇄 회피	제약 형식
a	음절두음 위치에 [경음]의 발생이 제한적이다.	*$[경음]
b	음절말음 위치에 [설정 저해음]의 발생이 제한적이다.	*[설정 저해음]$
c	음절 내 [양순음, 공명음][e, ʌ] 발생이 제한적이다.	*[m, p, p’, p^h][e, ʌ]
		*[m, n, l][e, ʌ]
d	음절 내 [치경음][ʌ]의 발생이 제한적이다.	*[n, l, t, t’, t^h][ʌ]
e	비인접 연쇄 제약 [음절두음]과 [음절말음]이 동시에 양순음을 갖지 못한다.	*[양순음]$_o$[양순음]$_c$

가장 두드러지는 것은 음절두음의 [경음], 음절말음의 [설정 저해음]이 회피되는 점이다(60a–b). 그리고 [양순음] 또는 [공명음]과 [e, ʌ]의 연쇄 및 [치경음][ʌ]의 연쇄가 발생하지 않는다(60c–d). 비인접 연쇄에 대해서는 (60e)에서 제시된 바와 같이, [음절두음]과 [음절말음]이 양순음을 갖지 못한다(신지영 2009, 예외: 폄, 범, 법, 품).

둘째, 음절 경계를 넘는 음소배열제약의 실재가 대부분 인정되지 않았다. 이는 한자어 형태소가 음절 단위라는 가정에 근거한다. 일부 연구(하세경 2000, 신지영·차재은 2003)는 고유어에서 모음 연쇄 회피가 분명하지만, 한자어와 차용어에서는 모음 연쇄의 발생이 빈번한 것으로 보았다(예: 대응[tɛ̱iŋ]). 그러나 실제로 음절 경계를 포함하는 연쇄의 분포가 직접 조사된 연구는 없는 것으로 보인다.

이러한 관찰들을 근거로, 일부 연구(채서영 1999, 이주희 2005, 박선우 외 2013, 남성현·김선회 2018)는 고유어와 구별되는 한자어의 어휘부를 상정할 수 있다고 보았다. 채서영(1999)와 이주희(2005)는 한자어와 고유어의 통시적/공시적 음운 규칙의 차이를 어휘부 계층을 설정하는 근거로 제시하였다. 채서영(1999)는 통시적 /o/ > /u/ 변화가 고유어에서는 나타나지만(예: 하루), 한자어에서는 나타나지 않는다고 하였다(예: 황도). 그리고 이주희(2005)는 고유어 어휘부에서는 모음 축약(사나이~사내)이 발생하지만, 한자어 어휘부에서는 모음 축약이 발생하지 않는다는 차이를 지적하였다. 기술 통계량을 활용하여 어휘부별 연쇄 분포의 특징을 포착하고, 어휘 계층 구조를 상정한 연구도 있다. 박선우 외(2013)은 ‘음운론적 복잡도(Goldsmith 2002, 2011)’를 기준으로, 어휘 계층별 연쇄 분포

의 특징을 거시적으로 살펴보았으며, 남성현·김선회(2018)은 [자음][모음] 연쇄에 집중하여 어휘 계층별 [자음][모음] 연쇄 분포의 차이를 보고하였다.

이와 같이 기존연구에서는 한자어 특유의 음소 연쇄 분포가 탐색되었으며, 고유어와 구별되는 한자어 어휘부가 가정되기도 하였다. 그러나 한자어 음소 연쇄에 대한 총체적이고 체계적인 음소배열제약의 탐색 및 기술은 이루어지지 못한 것으로 보인다.

지금까지 한국어 어휘부에 나타나는 음소 분포 및 음소배열제약에 대한 기존 연구를 살펴보았다. 이를 통해, 한국어 음소 연쇄에 나타나는 특징들을 종합할 수 있었으나, 기존연구에서 제시하는 범주적/비범주적 음소배열제약을 한국어 화자의 정신적 어휘부에 실재하는 음소배열제약으로 확정하기에는 어려운 측면이 있다. 이에 대해서는 3.4절에서 논의한다.

3.4. 논의

3.4.1. 기존 음소배열제약 탐색 방법의 한계

다수의 기존연구가 한국어 음소배열제약 및 연쇄 빈도를 보고하였으나, 한국어 화자의 비범주적 적형성을 파악하기에는 미비하였다. 기존연구의 한계점은 크게 두 가지로 보인다.

첫째, 특정 연쇄의 관찰 빈도가 기준 빈도(기대 빈도)보다 낮다는 것만이 서술될 뿐 빈도의 회피/선호에 대해 통계적 검증이 충분히 진행되지 않았다. 때문에, 연쇄 빈도를 바탕으로 기술된 경향을 그대로 유의미한 제약으로 보기 어렵다. 앞서 다룬 연쇄들은 대부분 상대 빈도 또는 O/E비율을 통해 과소 표상성(under-represented)이 계량화되었다. 그러나 2.2절에서 보았듯이, 두 기술 통계량이 가정하는 기준 빈도 및 기대 빈도는 음소 위치 및 계산 방향 등에 따른 문제가 있어 각 결과값이 음소 간 제약성을 적절히 나타낸다고 보기 어렵다.

둘째, 일반화의 단위가 연구자마다 다르다. 대부분의 연구자들은 분절음 단위 빈도를 조사하였고, 포함된 자연 부류를 기준으로 삼은 연구도 있다. 유재원(1997)이 자질을 단위로 일반화를 시도하였으나 자질 선택이 임의적이다. 이에 따라, 선호/회피가 파악된 연쇄도 있었으나, 별다른 선호와 회피를 보이지 않는

조합도 다수 기술되었다. 양적인 정보에 따라, 적절한 자연 부류를 포착할 수 있는 기제가 필요하다.

3.4.2. 한국어 화자의 비범주적 적형성

한국어 화자의 비범주적 적형성은 일부 연구에서 부분적으로 시사되었다. 다수 심리언어학적 연구(예: 권유안 2006, 구민모 외 2012)는 음절 빈도 및 음절의 이웃 단어의 수에 따라, 화자의 음운 정보 처리가 다름을 보였다. 그러나 음운론적으로 유의미한 연쇄 및 연쇄의 위치에 대해서는 충분히 고려되지 않아, 언어 처리(language processing) 이상의 체계적인 적형성 인식까지 직접적으로 보여준다고 보기 어렵다.

한편, Lee & Goldrick (2008)은 연쇄 분포와 이에 대한 한국어 화자의 인식을 조사하고, '음절하위 구성소'의 심리적 실재를 주장하였다. 비단어에 대한 '단기기억 과제(short-term memory task)'를 진행하고, 이 과제 응답과 한국어 [음절두음][모음]과 [모음][음절말음]의 빈도가 관계가 있다는 것을 보인다. 나아가 과제 응답이 [음절두음][모음]의 빈도를 [모음][음절말음]의 빈도보다 민감하게 반응한다는 것을 제시하여, '음절하위 구성소'의 심리적 실재를 뒷받침하고자 하였다. 그러나 연쇄 빈도에 따른 비범주적 인식은 '[음절두음][모음]'의 표상을 상정하는 것만으로는 충분히 포착되기 어렵다.

이러한 시도와 달리, 보편적 문법 제약에 기반한 비범주적 인식이 보고되기도 하였다. Berent et al. (2008)은 (61)과 같이 한국어에서 발생 빈도가 0인 어두 자음군에 대한 비범주적 인식을 밝혔다.

(61) 공명도에 따른 어두 자음군 유형

a. **<u>bl</u>**if (공명도의 높은 상승), **<u>bn</u>**if (공명도의 낮은 상승)

b. **<u>bd</u>**if (공명도 동일), **<u>lb</u>**if(공명도의 하강)

Berent et al. (2008)은 공명도 수준에 따라 네 가지 유형의 어두 자음군을 구성하고, 한국어 화자를 대상으로 지각 조사를 진행하였다. 그 결과, 공명도 상승 정도가 크지 않은 어두 자음군을 오지각할 확률이 공명도 상승 정도가 큰 어두 자음군을 오지각할 확률보다 크다는 것을 밝혔다(lbif < bdif < bnif < blif). 이를

통해, 한국어 화자가 공명도 투사 원리에 따라 발생 빈도가 0인 연쇄들에 대해서 비범주적으로 적형성을 인식할 수 있다는 것을 보였다.

이상의 연구를 바탕으로, 보다 체계적이고 구체적인 음소배열제약을 산출할 수 있는 모델에 근거하여 한국어 화자의 비범주적 적형성을 종합적으로 파악할 필요가 있다.

3.4.3. 최대 엔트로피 음소배열제약 모델의 도입 필요성

기존연구에 대한 검토를 통해, 한국어 음소배열제약 탐색을 위한 최적의 모델 두 가지 조건을 파악할 수 있다. 첫 번째 조건은 통계적 방법에 근거한 계산 및 일반화를 할 수 있어야 한다는 것이고, 두 번째 조건은 발생 빈도가 0인 연쇄를 포함하여 한국어 화자의 비범주적 적형성을 구체적으로 예측할 수 있어야 한다는 것이다. 이 연구는 이러한 조건에 모두 부합하는 모델로서, 최대 엔트로피 음소배열제약 모델을 채택한다. 이 모델은 어휘부 내 회피 경향을 통계적 제약으로 포착할 수 있어, 기존에 밝혀지지 않거나 인상적(impressionistically)으로 제시된 제약의 실재에 다가갈 수 있다.

앞서, Cho (2012)는 최대 엔트로피 음소배열제약 학습 모델을 이용하여 한국어 음소배열제약을 탐색한 바 있다. Cho (2012)의 학습 자료는 '한국어 학습용 어휘 선정 결과 보고서'(조남호 2003)에서 추출한 5,702개의 단어로 구성되었다. 해당 단어들의 품사를 살펴보면, 명사, 동사, 형용사 등이 포함되었는데 명사가 3,404개로 가장 비중이 크다. 어종별로 살펴보면, 고유어와 한자어가 비슷한 비중으로 어휘부의 대부분을 차지한다(고유어 2,399개, 한자어 2,474개, 고유어와 한자어 혼종어 829개). 투사 자질 층위는 따로 설정되지 않았고 제약의 길이는 최대 두 자질 매트릭스의 결합으로 제한하였다.

위와 같이 학습한 결과 얻어진 Cho (2012)의 문법은 기존연구에서 보고된 제약들을 모두 포함하였으며, 기존에 포착하지 못한 저빈도 연쇄도 제약으로 학습하였다. 즉, 음소배열제약의 기계 학습을 통해 한국어 어휘부에서 발생하는 연쇄와 발생하지 않은 연쇄에 대한 비범주적 인식을 예측할 수 있었다. 또한, Cho (2012)에서는 최대 엔트로피 음소배열제약 모델 학습에서 발생할 수 있는 문제를 언급하고 이러한 문제를 해소할 수 있는 방안을 제시하였다.

다만, Cho (2012)의 음소배열제약 학습은 투사 자질 층위를 따로 설정하지 않았기 때문에 비인접 제약을 포함하지 않았으며, 고유어와 한자어 문법의 개별적인 특징이 고려되지 않았다. 그리고 예측된 비범주적 인식에 대한 적형성 판단 조사가 진행된 것은 아니다.

이러한 배경에서, 이 연구는 고유어와 한자어 어휘부를 구분하고 비인접 연쇄를 탐색할 수 있는 최대 엔트로피 음소배열제약 모델을 적용한다. 고유어와 한자어 어휘부에서 보이는 회피 경향을 효과적으로 포착하며, 이에 대한 한국어 화자의 비범주적 인식을 밝히고자 한다.

4. 학습

이 장은 최대 엔트로피 음소배열제약 학습 모델을 이용하여, 한국어 음소배열제약 및 해당 제약들이 포착하는 일반화를 조사한다. 4.1절에서는 학습 어휘부 선정 과정과 모델이 구현된 UCLA 음소배열제약 학습 프로그램의 시행에 대해 소개한다. 4.2절에서는 한국어 고유어 및 한자어 어휘부를 바탕으로 학습된 문법을 세 부분으로 나누어 제시한다. 먼저, 고유어 문법과 한자어 문법에서 공통적으로 포함된 제약을 논의한다. 다음으로, 고유어 특정 문법, 한자어 특정 문법 순서로 살펴본다. 4.3절에서는 학습 결과가 실제 한국어 화자의 인식에 부합하는지를 검토하며, 학습 제약의 심리적 실재 여부를 논의한다.

4.1. 학습 방법

학습 시뮬레이션은 최대 엔트로피 음소배열제약 학습 모델을 구현한 'UCLA 음소배열제약 학습 프로그램(UCLA Phonotactic Learner)'을 사용하였다. 앞서 2.6.2절에서 설명하였듯이, 최대 엔트로피 음소배열제약 모델은 어휘부의 연쇄 분포를 모델에 주어진 자질 목록과 제약 형식을 이용하여 제약 및 그 가중치를 습득한다. 아래 4.1.1절에서 4.1.3절까지 UCLA 음소배열제약 프로그램에 입력할 어휘부와 자질 목록, 그리고 학습 조건을 기술한다.

4.1.1. 학습 어휘부

학습 자료는 단일 형태소인 명사 어휘로만 구성하였다. 그 이유는 한국어 화자가 '비단어(nonce word)'의 적형성을 판단할 때, 비단어를 단일 형태소인 명사 어휘로 인식하고 관련 음소배열제약을 사용할 가능성이 높다고 보았기 때문이다. 영어 음소배열제약을 다룬 Hayes & White (2013)에서도, 단일 형태소를 학습 대상으로 삼은 바 있다. 이를 통해, 영어 화자가 비단어를 단일 형태소로 인식할 가능성을 고려하는 동시에 복합어의 특수한 음소배열제약을 배제할 수 있었다. 이에 더하여, 이 연구가 품사를 명사로 제한한 이유는 명사가 한국어 전체 어휘의 절대 다수를 차지하며, 차용어/신조어 등 새롭게 수용된 단어의 품사 대부분이 명사이기 때문이다. 이 연구는 한국어 화자의 실제 어휘부에 가까운 학습 자료를 구성하고자, 「한국어 사용 빈도」(강범모·김흥규 2009)의 일반 명사(NNG)

목록에서 사전에 등재되고 사용 빈도(token frequency)가 5이상인 단일어를 선별하였다.[12]

기계 학습은 전체 학습 자료를 고유어와 한자어로 구분하여 두 개의 목록을 대상으로 진행하였다. 고유어는 기존연구(김경일 1985, 한성우 2006, Ito 2007, Hong 2010)에서 한국어 화자의 어휘부를 대표한다고 가정되었으며, 앞서 3장에서 보았듯이 고유어 관련 음소배열제약이 한국어 화자의 문법으로 여겨졌다. 한편, 한자어는 한국어 명사 어휘부에서 높은 비율을 차지하고 있기 때문에 한자어 관련 음소배열제약이 한국어 화자의 문법에 실재할 가능성이 있다. 고유 단일어 명사 목록에는 1,749개 단어, 한자 단일어 명사 목록에는 5,590개 단어가 포함되었다. 최대 엔트로피 음소배열제약 모델 학습을 위해서는 최소 3,000개 이상의 단어가 필요하기 때문에, 고유어 어휘 목록은 두 배로 복사하여 사용하였다.

이상의 학습 대상 단어들은 대부분 표준국어대사전에서 제공하는 발음형으로 학습 프로그램의 입력 자료로 투입되었다.[13]

사전 발음형은 10모음 체계를 따르면서 [e]-[ɛ]와 [y]-[ø]를 각각 구분하고 있다. 해당 음소들의 합류 가능성이 기존연구에서 제기되고는 있으나, 합류의 완성 여부는 이견이 있는 바, 해당 음소들에 대해서는 사전발음형을 그대로 사용하는 보수적인 접근을 취했다.

효과적인 음소배열제약 학습을 위해 사전의 발음형을 그대로 사용하지 않은 경우들도 있다. 모든 단어에 대해서 한 개의 발음형만 입력자료로 사용하였는데, 사전에서 복수의 발음을 제시하고 있는 단어의 경우에도 좀 더 전형적이고 보수

---

[12] 최대 엔트로피 음소배열제약 학습 모델(Hayes & Wilson 2008)은 연쇄의 유형 빈도에 기반을 두고 제약을 학습한다. 이에 따라, 이 연구의 시뮬레이션에서도 단어의 사용 빈도는 반영하지 않았다. 관련 기존연구(Bailey & Hahn 2001, Albright 2009 등)는 학습 모델이 사용 빈도를 반영하는 경우 모델의 예측력이 오히려 떨어지는 경우를 제시하며, 적형성 인식 판단에 단어의 사용 빈도의 영향이 거의 없다는 것을 보인 바 있다.

[13] 변이음 교체 규칙(예: [s] 구개음화)과 일부 수의적 음운 규칙(예: 조음위치 자질 동화)은 사전 발음에 반영되어 있지 않다. 그 밖에 사전 발음에 반영된 음운규칙은 「표준어 규정」 제 2부 표준 발음법(문화체육관광부 고시 제2017-13호[2017. 3. 28.])에서 확인할 수 있다.

적인 발음으로 간주될 수 있는 한 개의 발음형만 입력하였다. 예를 들어, ‘ㄹ’을 제외한 음절두음이 [je]에 선행하는 경우 [j]의 탈락이 수의적으로 관찰되며, 사전 발음에도 [j] 탈락형이 병기되어 있으나(예:계[kje~ke], 은혜[ɨnhje~ɨnhe]), 본 연구에서는 [j]를 포함한 발음만을 입력자료로 사용하였다.

그 밖에, 현대 한국어 화자의 실제 발화를 반영하면서 기계 학습을 용이하게 할 목적으로 사전의 발음형을 그대로 사용하지 않은 경우들을 아래 (62)에 요약하였다.

(62) 사전 발음형을 수정한 경우

a. 장단을 구분하지 않음

b. ‘의’를 상향 이중모음 [ɯi]로 전사

c. [음절말음][ɯi] 연쇄에서 [ɯ] 삭제

학습 어휘부의 발음형에는 어두 위치에서만 제한적으로 표시되는 장단 구분을 반영하지 않았다(62a). 또한, ‘의’를 하향 이중모음 [ɨj]로 전사하여 입력하면, (아래에서 설명하는 바와 같이) 이중모음 단순화가 적용되는 [음절두음][의] 연쇄와 [음절두음][으] 연쇄가 모두 [음절두음][모음]의 연속을 포함하는 형태가 되어, ‘의’가 (‘으’와 같은) 단순모음과 상이한 발생 분포를 보이는 것을 포착하지 못하는 바, ‘의’를 상향 이중모음 [ɯi]로 전사하여 입력하였다(62b). 이와 관련하여, 사전에서는 [음절두음][의] 연쇄에 대해 [ɯ]를 삭제한 [음절두음][i]를 발음형으로 제시하고 있다(예: 희망[히망], 무늬[무니]). 철자상 [음절말음][의] 연쇄에 대해서는 [ɯ]가 실현된 형태도 가능한 발음으로 삭제된 형태와 함께 병기하여 제공하고 있으나(예: 문의[무늬/무니]), 해당 단어에서 연음화가 늘 적용된다는 가정하에서 [ɯ]가 삭제된 형태만을 입력 자료로 사용하였다(62c).[14]

4.1.2. 자질 목록

자질 목록은 한국어 음소를 대상으로 아래 (63)과 같이 구성되었다. 대부분의 자질 명세는 기존연구(Hayes & Wilson 2008, Cho 2012)를 따랐다. 자음, 활음, 모

[14] 다만, [ŋ][모음] 연쇄는 전통적으로 연음화가 적용되지 않는다고 가정되므로, [ŋ][의] 연쇄는 [ŋɯi]로 전사하였다.

음 모두에 성절성[syllabic], 자음성[consonantal], 접근성[approximant], 그리고 공명성[sonorant] 자질값이 할당되었다. 자음의 주요 조음위치자질은 +자질 하나만을 명세하며, 음절두음과 음절말음 정보를 제약에 반영하기 위하여, [+/−rhyme] 자질(Hayes & White 2013)을 사용한다. 한편, 모음은 전/후설, 고/저, 원순/평순 자질, 그리고 ATR 자질을 부여한다. ATR 자질은 모음 조화 관련 제약을 포착하기 위해 명세하였다(강옥미 2011, Hong 2010).[15] 활음은 자음과 마찬가지로 [+/−rhyme] 자질을 명세하는 동시에 ATR 자질을 제외한 모음자질을 명세한다.

(63) 자질 목록[16]

a. 자음

	syl	cons	appr	son	cont	str	nas	ant	asp	tns	lab	cor	dor	rhy
p	−	+	−	−	−	0	−	0	−	−	+	0	0	−
p_{coda}	−	+	−	−	−	0	−	0	−	−	+	0	0	+
p^h	−	+	−	−	−	0	−	0	+	−	+	0	0	−
p^h_{coda}	−	+	−	−	−	0	−	0	+	−	+	0	0	+
p’	−	+	−	−	−	0	−	0	−	+	+	0	0	−
p'_{coda}	−	+	−	−	−	0	−	0	−	+	+	0	0	+
t	−	+	−	−	−	−	−	+	−	−	0	+	0	−
t_{coda}	−	+	−	−	−	−	−	+	−	−	0	+	0	+
t^h	−	+	−	−	−	−	−	+	+	−	0	+	0	−
t^h_{coda}	−	+	−	−	−	−	−	+	+	−	0	+	0	+
t’	−	+	−	−	−	−	−	+	−	+	0	+	0	−
t'_{coda}	−	+	−	−	−	−	−	+	−	+	0	+	0	+
c	−	+	−	−	−	+	−	−	−	−	0	+	0	−
c_{coda}	−	+	−	−	−	+	−	−	−	−	0	+	0	+
c^h	−	+	−	−	−	+	−	−	+	−	0	+	0	−
c^h_{coda}	−	+	−	−	−	+	−	−	+	−	0	+	0	+
c’	−	+	−	−	−	+	−	−	−	+	0	+	0	−
c'_{coda}	−	+	−	−	−	+	−	−	−	+	0	+	0	+

[15] 이는 다수의 연구가 모음 조화 자질로 [ATR] 또는 [RTR]을 채택함도 고려한 것이다(Jun 2018: (31)).

[16] 자질을 나타내는 약어의 의미는 다음과 같다.

syl=syllabic, cons=consonantal, appr=approximant, son=sonorant, cont=continuant, strid=strident, nas=nasal, ant=anterior, asp=aspirate, tns=tense, cor=coronal, lab=labial, dor=dorsal, rhy=rhyme, bk=back, rnd=round

	syl	cons	appr	son	cont	str	nas	ant	asp	tns	lab	cor	dor	rhy
k	−	+	−	−	−	0	−	0	−	−	0	0	+	−
k_{coda}	−	+	−	−	−	0	−	0	−	−	0	0	+	+
k^h	−	+	−	−	−	0	−	0	+	−	0	0	+	−
k^h_{coda}	−	+	−	−	−	0	−	0	+	−	0	0	+	+
k’	−	+	−	−	−	0	−	0	−	+	0	0	+	−
k'_{coda}	−	+	−	−	−	0	−	0	−	+	0	0	+	+
s	−	+	−	−	+	+	−	+	−	−	0	+	0	−
s_{coda}	−	+	−	−	+	+	−	+	−	−	0	+	0	+
s’	−	+	−	−	+	+	−	+	−	+	0	+	0	−
s'_{coda}	−	+	−	−	+	+	−	+	−	+	0	+	0	+
h	−	+	−	−	+	0	−	0	+	−	0	0	0	−
h_{coda}	−	+	−	−	+	0	−	0	+	−	0	0	0	+
m	−	+	−	+	−	0	+	0	0	0	+	0	0	−
m_{coda}	−	+	−	+	−	0	+	0	0	0	+	0	0	+
n	−	+	−	+	−	0	+	+	0	0	0	+	0	−
n_{coda}	−	+	−	+	−	0	+	+	0	0	0	+	0	+
ŋ	−	+	−	+	−	0	+	0	0	0	0	0	+	−
$ŋ_{coda}$	−	+	−	+	−	0	+	0	0	0	0	0	+	+
l	−	+	+	+	+	0	−	+	0	0	0	+	0	−
l_{coda}	−	+	+	+	+	0	−	+	0	0	0	+	0	+

b. 활음 및 모음

	syl	cons	appr	son	high	low	bk	rnd	ATR	rhy
j	−	−	+	+	+	−	−	−	0	-
j_{coda}	−	−	+	+	+	−	−	−	0	+
w	−	−	+	+	+	−	+	+	0	-
w_{coda}	−	−	+	+	+	−	+	+	0	+
ɯ	−	−	+	+	+	−	+	−	0	-
$ɯ_{coda}$	−	−	+	+	+	−	+	−	0	+
i	+	−	+	+	+	−	−	−	+	0
e	+	−	+	+	−	−	−	−	+	0
ɛ	+	−	+	+	−	+	−	−	−	0
ɨ	+	−	+	+	+	−	+	−	+	0
ʌ	+	−	+	+	−	−	+	−	+	0
a	+	−	+	+	−	+	+	−	−	0
o	+	−	+	+	−	−	+	+	−	0

	syl	cons	appr	son	high	low	bk	rnd	ATR	rhy
u	+	−	+	+	+	−	+	+	+	0
y	+	−	+	+	+	−	−	+	+	0
ø	+	−	+	+	−	−	−	+	−	0

4.1.3. 학습 조건

인접 제약의 경우, 최대 결합 자질 매트릭스의 수를 둘로 제한한다. 비인접 제약의 경우에는 아래 (64)와 같이 투사 층위(projection tier)를 설정하여 학습한다.

(64) 비인접 제약 학습을 위한 투사 층위 설정

번호	투사 층위	투사 자질
a	[+syllabic]	high, ATR
b	[+consonantal]	labial, coronal, dorsal
c	[−sonorant,−rhyme]	tense, aspirate

첫째, 모음 조화 제약을 포착하기 위해서, 모음자질만을 기준으로 학습할 수 있도록 하는 모음 층위를 설정하였다. 투사 자질은 Hong (2010)에 따라, [+/−high]와 [+/−ATR]로 정하였다. 둘째, 동일 조음위치자질 회피 제약을 학습할 수 있도록 자음 층위를 더한다. Ito (2007)의 기술을 고려하여, 투사 자질을 주요 조음위치자질인 [labial], [coronal], [dorsal]로 할당하였다. 셋째, 음절두음 사이의 후두자질 공기 관계를 파악하기 위해서 저해음인 음절두음 층위를 설정하고, 투사 자질을 [+/−aspirate]와 [+/−tense]로 정하였다.

모음 조화와 후두자질 공기 제약은 기존연구에서 최대 세 개의 분절음에 걸쳐 기술되었으므로(Hong 2010, Kang & Oh 2016), 비인접 제약 학습 시 자질 매트릭스가 최대 세 개까지 조합될 수 있도록 하였다. 정확도(O/E)는 0.3 수준으로 맞추고, 최대 학습 제약의 수는 제한하지 않는다. 또한, 학습 시뮬레이션에서 '상보적 자연 부류(complement natural classes)'를 허용한다. '상보적 자연 부류'란 명세된 자질이 가리키는 자연 부류 외 분절음을 가리키며, 기호 '^'로 나타낸다. 예를 들어, [^−aspirate,−tense,+labial]는 '[p]를 제외한 분절음'을 의미한다. 이 자연 부류를 사용함으로써, 보다 일반적이고 해석이 용이한 제약을 학습할 수 있다(Hayes & Wilson 2008: 391).

기계 학습 시행은 어휘부마다 5회씩 진행되었다. Hayes & Wilson (2008) 모델에서 제약이 통계적으로 선택되기 때문에 각 학습 문법은 다소 제약 구성이 달라 보이지만, 궁극적으로 예측하는 적형성은 유사하다. 이 연구는 각 학습 문법 중 Hayes & Wilson (2008)의 방식을 따라 5장에서 보고할 한국어 화자의 적형성 판단 조사 결과와 가장 낮은 상관관계를 보이는 문법을 보고한다.

4.2. 결과

이 절에서는 고유어 어휘부를 대상으로 학습된 고유어 문법과 한자어 어휘부를 대상으로 학습된 한자어 문법을 논의한다. 이 연구에서 학습된 고유어 문법은 118개의 제약으로 이루어졌으며, 한자어 문법은 109개의 제약으로 구성되었다. 고유어 전체 제약 목록은 [부록1], 한자어 전체 제약 목록은 [부록2]에 제공한다.

학습 결과는 다음과 같은 순서로 살펴보겠다. 4.2.1절에서는 고유어 및 한자어 어휘부 문법에 공통적으로 포함된 제약을 기술한다. 4.2.2절에서는 고유어 어휘부 문법에만 포함되고, 한자어 어휘부 문법에는 포함되지 않은 제약을 제시한다. 4.2.3절에서 고유어 어휘부에서는 학습되지 않고, 한자어 어휘부에서 특정적으로 학습된 제약을 제시한다. 해당 제약들을 제시하면서 학습된 제약이 기존연구에서 범주적 제약 및 비범주적 경향으로 기술한 일반화를 포착하는지 확인하고, 새로운 제약 및 경향의 학습 여부에 대해서도 논의한다.

4.2.1. 고유어 및 한자어 문법 공통 음소배열제약

고유어 문법과 한자어 문법에서 공통적으로 학습된 음소배열제약 대부분은 기존연구에서 논의한 제약 및 회피 경향을 포착한 것이다. 이하에서는 이 연구에서 학습된 제약을 (i) 기존연구에서 정의한 범주적 제약에 대응하는 것, (ii) 기존연구에서 관찰한 비범주적 제약에 대응하는 것, 그리고 (iii) 기존연구에서 포착되지 않은 새로운 제약, 세 부류로 나누어 살펴보기로 한다.

첫째, 기존연구에서 범주적으로 정의한 음소배열제약 및 그 효과(3.1절 참조)에 대응하는 학습 결과를 살펴보자. 기존연구에서 범주적으로 정의한 음소배열제약 및 그 효과는 고유어 어휘부와 한자어 어휘부에서 대체로 높은 가중치를

할당받은 개별 제약으로 학습되거나 해당 효과를 갖는 복수의 제약들로 학습되었다. 다음 (65)에서 해당 제약의 예를 볼 수 있다.

(65) 학습 결과: 기존연구에서 범주적 음소배열제약으로 정의한 예

번호	제약	고유어	한자어	의미	
음절구조 제약					
a	*[-rhy][+cons]	5.89	7.07	*$[자음][자음]	
b	*[+son,+dor,-rhy]	4.74	4.96	*$[ŋ]	
[활음][모음] 제약					
c	*[+rnd,-syl][^-high,-rnd]	1.86	4.22	*[wu, wo]	
d	*[+bk,-syl][^-rnd,+syl]	1.12	1.48		
필수적 음운 규칙 관련 제약					
e	*[-son,+rhy][^-son,-rhy]	2.94	3.46	*[저해음]$[공명음]	비음화
f	*[-nas][+nas,+ant]	2.43	2.83	*[저해음,유음]$[n]	
최근 차용어에서만 예외가 존재하는 제약					
g	*#[+cons,+appr]	3.98	4.31	어두 [l] 금지	
h	*[+hi,+bk,-rnd]#	2.30	2.87	어말 [ɨ] 금지	
고유어와 한자어에서 예외가 존재하는 제약					
i	*[+lab][+high,+bk,-rnd]	2.65	2.13	*[m, p, p’, p^h][ɨ]	
j	*[+lab][+bk,-syl]	2.56	3.39	*[m, p, p’, p^h][w]	
k	*[-son,+cor][-rnd,-syl]	2.13	2.78	*[t, t’, t^h, s, s’, c, c’, c^h][j]	

이러한 제약들을 위배하는 연쇄는 상당히 높은 비적형성 점수를 받는다. 예를 들어, 음절두음의 어두 자음군은 제약 (65a) ‘*[−rhy][+cons]’을 위배하여 고유어 문법에서는 5.89, 한자어 문법에서는 7.07의 비적형성 점수를 받는다. ‘[k$n]’ 연쇄를 포함하는 형태는 제약 (65e) ‘*[−son,+rhy][^−son,−rhy]’와 제약 (65f) ‘*[−nas][+nas,+ant]’을 위배하여, 고유어 문법에서 5.37, 한자어 문법에서 6.29의 비적형성 점수를 받는다. 이러한 비적형성 점수는 한국어 화자가 해당 연쇄를 심각하게 회피할 것이며, 발생 빈도가 0인 연쇄 간에도 적형성 인식이 다를 수 있음을 나타낸다.

둘째, 기존연구에서 논의된 비범주적 제약 및 회피 경향(3.2절 참조)과 관련된 제약을 논의한다. 학습 제약은 음운론적 동기를 갖는 제약, 음운론적 동기가 불

분명한 제약, 그리고 [격음]과 [경음]의 제한적 분포에 대한 인접 연쇄 제약 순서로 살펴본다(이하 (66–74) 참조).

먼저, 기존연구에서 포착된 비범주적 제약 중 음운론적 동기를 갖는 제약을 논의한다. 이들 제약이 포착한 효과는 기존연구에서 공시적 또는 통시적 변화 과정에서 음운론적으로 자연스러운 것으로 분석되었으며 언어 유형론적으로도 보편적인 것으로 보고되었다((47) 참조). 이 연구에서는 인접 [자음][모음] 연쇄와 [모음][모음] 연쇄에 대하여 음운론적 동기를 갖는 제약이 학습되었다.

(66)에서 요약한 바와 같이, 고유어 문법과 한자어 문법은 [자음][모음] 연쇄 중 [설정 저해음][i] 연쇄와 [구개음][ɨ] 연쇄의 발생 제한을 공통적으로 요구한다.

(66) 음운론적 동기를 갖는 제약: [자음][모음] 연쇄

번호	어휘부	제약	가중치	의미	예외
a	고유어	*[−strid,−asp][+high,−bk]	2.87	*[t, t’][j, i]	마디
b		*[−ant][+high,+bk,−rnd]	2.93	*[c, c’, c^h][ɨ]	짜증
c	한자어	*[−str][+high,−bk]	3.43	*[t, t’, t^h][j, i]	-
d		*[+str,+asp][+high,+bk,−rnd]	2.13	*[c^hɨ]	층

다만, 고유어 문법은 한자어 문법과 달리 [t, t’, t^h][i] 중 [t^hi] 발생을 제한하지 않고, 한자어 문법은 고유어 문법과 달리 [c, c’, c^h][ɨ] 중 [c, c’][ɨ] 발생을 제한하지 않는다. (66)에 학습된 제약들은 다수의 예외를 허용하기 때문에 일부 연구(진남택 1992, 김미란 외 2014, Chong 2017)에서만 보고되었다. 관련 연구에서는 이 제약들은 자음과 모음의 동화와 관련하여 분석되었으며 음운론적으로도 자연스러운 것으로 간주된다. 해당 제약은 예외를 허용함에도 이 연구에서 제약으로 포착되었으며, 기존연구에서 분석한 바 음운론적인 동기도 뒷받침되기 때문에 해당 제약이 한국어 화자의 인식에 실재할 가능성이 있다.

다음으로, 고유어와 한자어 문법은 [모음][모음] 연쇄에 대해서도 발생 제한을 요구한다. 한국어에서 발생 가능한 [모음][모음] 연쇄(100개) 중에서 고유어 문법은 세 개([a][i, y, ɨ])의 연쇄를 제외한 97개의 모음 연쇄 발생이 제한되고, 한자어 문법은 80개 연쇄 발생이 제한된다. 이 중 두 어휘부 문법이 [모음][모음] 연쇄에 관하여 공통적으로 포착한 제약 및 일반화를 (67)에 정리하였다.

(67) 음운론적 동기를 갖는 제약: [모음][모음] 연쇄

번호	어휘부	제약	가중치	의미	예외
a	고유어	*[−high][−high,−bk]	1.01	*[저모음, 중모음][e, ø, ɛ]	-
	한자어		2.10		사액
b	고유어	*[+high,+bk,−rnd][−high]	2.48	*[ɨ][비고모음]	-
c		*[−low,+bk,−rnd][−high,−low]	0.87	*[ɨ, ʌ][e, ø, ʌ, o]	-
d		*[+high,+bk,−rnd][+bk]	2.57	*[ɨ][w, ɨ, u, ʌ, o, a]	-
e	한자어	*[+bk,−rnd,+ATR][−rnd,+syl]	2.30	*[ɨ, ʌ][i, ɨ, e, ɛ, ʌ, a]	-
f		*[−low,+bk,−rnd][−high]	1.81	*[ɨ, ʌ][비고모음]	-

고유어와 한자어 어휘부에서 모두 [저모음, 중모음][e, ø, ɛ] 연쇄가 회피된다 (67a). 또한, 고유어 제약 (67b–d)와 한자어 제약 (67e–f)는 [ɨ, ʌ][모음] 연쇄의 발생 제한을 요구한다. 이 제약은 [ɨ]가 다른 [모음] 앞에서 잘 나타나지 않으며, [e, ɛ]가 다른 [모음]에 후행하지 않는다는 일반화(유재원 1997)를 포착한다. 앞서 3.2절에서 밝힌 바와 같이 [모음][모음] 연쇄 회피 제약은 언어 유형론적으로 보편적이라고 볼 수 있으나, 관련 분절음의 종류에 따라 모음 연쇄의 회피 정도가 다른 것은 구체적으로 논의되지 않은 것으로 알고 있다.

다음 순서로, 기존연구에서 보고한 비범주적 제약 중 음운론적 동기가 불분명한 제약 및 그 제약이 포착하는 일반화를 논의한다((49) 참조). 이들 제약 및 일반화는 기존연구에서 반복적으로 관찰되었으나 제약 설정의 동기가 구체적으로 설명되지 않았다. 아래 나열하였다시피, 이 연구에서 관련 제약들은 [자음][활음] 연쇄, [자음][모음] 연쇄, [모음][자음] 연쇄에 대해 학습된다.

[자음][활음] 연쇄 유형부터 보면, (68)에서 요약한 바와 같이 [공명 자음] 또는 [설정 저해음]이 [w] 앞에 잘 오지 못한다.

(68) 음운론적 동기가 불분명한 제약: [자음][활음] 연쇄

번호	어휘부	제약	가중치	의미	예외
a	고유어	*[+son,−syl][+bk,−syl]	2.41	*[m, n, l, ŋ][w]	-
b		*[−tns,+cor][+son,−syl]	2.33	*[t, t^h, s, c, c^h][w, j]	돼지
c		*[−ant][−syl]	1.86	*[c, c’, c^h][w, j]	-
d	한자어	*[+son,−rhy][+bk,−syl]	2.08	*[m, n, l][w]	단원
e		*[−son,+ant][−syl]	2.44	*[t, t’, t^h, s, s’][w, j]	쇄도
f		*[−son,−cont,+cor][−syl]	2.06	*[t, t’, t^h, c, c’, c^h][w, j]	계좌
g		*[+asp,+cor][−syl]	1.48	*[t^h, c^h][w, j]	촬영
h		*[−str][−syl]	0.69	*[t, t’, t^h][w, j]	-

(68)의 제약을 포함한 고유어 및 한자어 문법은 [공명 자음][w] 연쇄와 [설정 저해음][w] 연쇄에 대해서 (69)와 같이 비적형성 점수를 부과한다.

(69) 비적형성 점수: [공명 자음][w] 연쇄와 [설정 저해음][w] 연쇄

a. [공명 자음][w]

자음	고유어	한자어	자음	고유어	한자어
m	4.96	5.48	n	2.41	2.08
ŋ	2.41	0[17]	l	2.41	2.08

b. [설정 저해음][w]

자음	고유어	한자어	자음	고유어	한자어	자음	고유어	한자어
t	2.33	5.18	s	2.33	2.44	c	4.19	2.06
t’	0[18]	5.18	s’	0[19]	2.44	c’	1.86	2.06
t^h	2.33	6.66				c^h	4.19	3.54

(69)에서 보면, 고유어 문법과 한자어 문법은 [m, n, l][w] 연쇄 발생을 공통적으로 제한하는 반면, [ŋw] 연쇄의 발생은 고유어 문법에서만 저지된다. 또한, [설정 저해음][w] 연쇄 중 [t’w, s’w]는 고유어 문법에서만 허용된다. 이 제약들의 동기를 유표적인 [활음]의 분포 제한으로도 볼 수 있으나, [공명음]과 [설정음]이 [w]와 제한을 보이는 것에 대한 이유는 구체적으로 언급되지 않았다.

[자음][모음] 연쇄 중에서는 [p^h][e, ʌ]와 [k^hʌ]가 고유어 및 한자어 어휘부에서 모두 발생이 제한된다. 다만, 한자어 어휘부 문법은 [양순음][e] 연쇄와 [k^h][ATR 모음] 연쇄의 회피까지 요구하여 고유어 어휘부 문법보다 더 많은 연쇄 회피를 포착한다. 관련 제약을 아래 (70)에 제시한다.

(70) 음운론적 동기가 불분명한 제약: [자음][모음] 연쇄

번호	어휘부	제약	가중치	의미	예외
a	고유어	*[+asp,+lab][−high,+ATR]	2.45	*[p^h][e, ʌ]	펄
b		*[+asp,+dor][−high,+bk,+ATR]	2.26	*[k^hʌ]	-
c	한자어	*[+asp,+lab][−low,+bk,−rnd]	3.30	*[p^h][ɨ, ʌ]	입헌
d		*[+lab][−high,−low,−bk]	3.01	*[p, p’, p^h][e, ø]	입회
e		*[+asp,+dor][+ATR]	2.74	*[k^h][i, y, ɨ, u, e, ʌ]	국헌

[17] 한자어 발생 예: 공원[koŋwʌn], 풍월[p^huŋwʌl]

[18] 고유어 발생 예: 똬리[t’wali]

[19] 고유어 발생 예: 쐐기[s’wɛki]

[p^h][e, ʌ] 연쇄와 [k^hʌ] 연쇄의 발생 제한은 고유어와 한자어를 구분하지 않고 학습한 Cho (2012)의 문법에도 제약으로 포함된 바 있으며, 일부 연구(허웅 1985, 신지영·차재은 2003)에서 이른바 '우연한 빈칸'으로도 기술되었다. 해당 연쇄는 최근 차용어에서(예: 페인트, 퍼펙트, 커브) 자주 발생하는점을 고려하면, 제약 (70)이 현재 한국어 화자의 적형성 판단에 실제로 영향을 줄 수 있을지에 대해서는 추가적인 검토가 필요하다.

또한 [모음][자음] 연쇄 유형 제약을 보면, 고유어 문법과 한자어 문법 모두 폐음절에서 [e, ɛ] 발생 제한을 요구한다. 관련 제약을 (71)에 요약하고, 각 어휘부 문법이 [e, ɛ][음절말음] 연쇄에 부과한 비적형성 점수를 (72)에 제시한다.

(71) 음운론적 동기가 불분명한 제약: [모음][자음] 연쇄

번호	어휘부	제약	가중치	의미	예외
a	고유어	*[−high,−bk][−nas,+rhy]	2.68	*[e, ø, ɛ][p, t, k, l]$	맵시
b		*[−high,−low,−bk][+lab,+rhy]	2.34	*[e, ø][m, p]$	셈
c		*[−high,−bk][−appr,+cor,+rhy]	1.31	*[e, ø, ɛ][t, n]$	맨드라미
d		*[−low,−bk][+dor,+rhy]	2.22	*[i, y, e][k, ŋ]$	싱아
e	한자어	*[−high,−bk,+ATR][+rhy]	3.32	*[e][음절말음]	-
f		*[−high,−bk][+lab,+rhy]	4.24	*[e, ø, ɛ][m, p]$	-
g		*[−high,−bk][+cor,+rhy]	3.96	*[e, ø, ɛ][t, n, l]$	-
h		*[−low,−bk][+son,+dor]	2.04	*[i, y, e, ø][ŋ]	횡
i		*[−high,−low,−bk][+son,+rhy]	0.40	*[e, ø][m, n, ŋ]$	횡포

(72) 비적형성 점수: [e, ɛ][음절말음] 연쇄

음절말음	e		ɛ	
	고유어	한자어	고유어	한자어
p	5.02	7.56	2.68	4.24
t	6.07	13.28	6.07	9.95
k	4.89	3.32	2.68	0[20]
m	2.34	7.96	0[21]	4.24
n	3.40	9.41	3.40	5.69
ŋ	2.22	5.75	0[22]	0[23]
l	2.68	7.68	2.68	3.96

[20] 한자어 발생 예: 액(厄)

[21] 고유어 발생 예: 뱀

[22] 고유어 발생 예: 앵두

[23] 한자어 발생 예: 갱(坑)

이에 따라, [e][음절말음], [ɛ][p, t, n, l]$ 연쇄 발생은 고유어와 한자어 어휘부에서 모두 저지된다. 다만, [ɛm]$의 발생은 한자어 어휘에서만 제한되고 [ɛk]$의 발생은 고유어에서만 회피된다.

[e, ɛ][음절말음] 연쇄 발생 제한에 대한 동기는 기존연구에서 구체적으로 제공되지 않았다. 다만 15세기 한국어에서 [e, ɛ]가 하향 이중모음으로 실현되며, 이 하향 이중모음에 후행하여 음절말음이 오지 못하는 것이 관찰되었다(김남미 2004). 이러한 통시적 흔적이 현대 한국어 어휘부에 반영되어 제약으로 학습된 것으로 볼 수 있다. 그러나 한국어 화자들이 공시적으로는 동기가 불분명한 제약을 통계적 기준에 의해 습득하고 인식할 수 있을지에 대해서는 더 논의가 필요하다.

기존연구에서 논의된 비범주적 제약 중 [격음]과 [경음]의 제한적 분포((50) 참조)도 고유어 문법과 한자어 문법의 일부로 학습되었는지를 확인하고 그 양상을 정리한다. [격음]부터 살펴보면, [격음]은 [저해음] 뒤에 잘 오지 못한다. 아래 (73)에서 제시하였듯이, 고유어와 한자어에서 [p, t, k][k^h], [pp^h]가 공통적으로 제한되고 고유어 문법만이 [pt^h] 연쇄의 발생을 추가적으로 저지한다.

(73) 후두자음 분포에 관한 인접 제약: [저해음][격음] 연쇄

번호	어휘부	제약	가중치	의미	예외
a	고유어	*[−son][−tns,+dor]	1.91	*[p, t, k][k^h]	-
b		*[−son,+cor][−tns]	1.91	*[t][p^h, t^h, c^h, k^h, h]	-
c		*[−son,+lab][+lab]	1.97	*[p][p’, p^h]	-
d		*[−son,+lab][+ant,−tns]	1.81	*[p][t^h]	-
e	한자어	*[−syl][+asp,+dor]	2.41	*[자음][k^h]	-
f		*[−son,+cor,+rhy]	3.35	*[t]$	곶
g		*[−son,+lab][−tns,+lab]	1.65	*[pp^h]	집필

이 제약들은 [저해음][격음]의 회피를 기술한 기존연구(고유어: 김경일 1985, 전체 한국어: 유재원 1997)와도 일치하며, Cho (2012)의 학습 문법에도 포함된 바 있다.

한편, [경음]은 고유어와 한자어 어휘부에서 [공명음] 뒤에 잘 오지 못한다. 특히, 기존연구에서 관찰한 바와 같이 [모음, 비음]에 후행하여 [경음] 발생이 저지되는 편이다. 이를 (74)에 정리하여 제시한다.

(74) 후두자음 분포에 관한 인접 제약: [공명음][경음] 연쇄

A. 고유어

번호	제약	가중치	의미	예외
a	*[+ATR][+tns,+lab]	2.19	*[i, y, ɨ, u, e, ʌ][p’]	삐삐
b	*[+bk,−rnd][+tns,+lab]	2.14	*[ɨ, ʌ, a][p’]	아빠
c	*[+nas][+tns,+lab]	2.37	*[n, m, ŋ][p’]	-
d	*[^+cons,+rhy][+tns,+cor]	2.18	*[모음][t’, s’, c’]	버찌
e	*[+nas,+ant][+tns,+cor]	2.51	*[n][t’, s’, c’]	-
f	*[+son,+lab][−cont,+tns,+cor]	2.32	*[m][t’, c’]	-
g	*[+cont][+tns,+dor]	1.91	*[lk’]	-
h	*[+high,+bk,−rnd][+tns,+dor]	2.12	*[ɨk’]	-

B. 한자어

번호	제약	가중치	의미	예외
a	*[^−nas,+rhy][+tns]	3.94	*[n, m, ŋ, 모음][경음]	만끽
b	*[^−cont,+rhy][+tns,+lab]	1.84	*[모음,유음]$[p’]	-
c	*[^−nas,+rhy][+tns,+cor]	1.64	*[모음,비음]$[t’, s’, c’]	-
d	*[^−cont,+rhy][+tns,+dor]	2.48	*[모음,유음]$[k’]	태권

다만, 한자어 문법은 [모음, 비음][경음] 일반을 제한하지만(74B), 고유어 문법은 [모음, 비음] 뒤에 경음 [k’]의 발생을 비교적 허용한다. [유음][경음]에 대해 살펴보면, 고유어 문법과 한자어 문법은 [lk’] 연쇄 발생을 공통적으로 저지하고(74Ag, 74Bd), [lp’] 연쇄 발생은 한자어 문법에서만 제한된다(74Bb). 이러한 회피 경향은 한국어 구문에서 관형사형 어미 ‘-ㄹ[−l]’에 후행하여 [p’, k’]가 생산적으로 경음화되는 현상(예: 할 것[lk’]이다)과는 배치된다.

이상 (73)과 (74)에서 정리하였듯이, 기존연구에서 논의된 격음과 경음의 제한적인 분포가 고유어 및 한자어 어휘부 문법의 일부로 학습되었다. 다만, 학습 어휘부에 따라 학습된 격음과 경음 포함 연쇄의 회피 제약 종류가 다른데, 한국어 화자의 문법이 어떤 어휘부의 학습 결과를 어느 정도로 반영하는지 앞으로도 계속 논의할 필요가 있다.

셋째, 기존연구에서 구체적으로 언급되지 않고 이 연구에서 새롭게 학습한 제약을 살펴보자. 해당 제약들은 (75)에서 요약한 바와 같이 과거 단독 연구자에 의해 언급되었거나 산발적으로만 보고된 회피 경향을 포착한다.

(75) 새롭게 학습한 제약

A. 고유어

번호	제약	가중치	의미	예외
a	*#[−high,−bk]	2.92	*#[e, ɛ, ø]	애꾸
b	*[−rnd][−bk,−ATR]	2.99	*[jɛ]	-
c	*[+bk,−syl][−high,−low,−bk]	1.74	*[w, ɯ][e]	꿰미
d	*[+high][+high]	1.63	*[i, ɨ, u][w, j, ɯ]	-
e	*[−son,+lab][+lab]	1.97	*[p][p’, p^h]	-
f	*[+lab][+tns,+lab]	0.98	*[m, p][p’]	-

B. 한자어

번호	제약	가중치	의미	예외
a	*#[−high,−low,−bk]	2.75	*#[e, ø]	외경
b	*[−low,−rnd][−bk,−ATR]	4.07	*[jɛ]	-
c	*[+bk,−syl][−high,−bk]	2.27	*[w][e, ɛ]	궤도
d	*[+high,+bk,−rnd][−cons,−syl]	1.92	*[ɨ][w, j, ɯ]	
e	*[−son,+lab][−tns,+lab]	1.65	*[pp^h]	집필

먼저, 어두 위치에서 고유어와 한자어에서 모두 공통적으로 [e, ø]가 오지 못한다(75Aa, 75Ba). 관련 제약은 예외를 허용하기 때문에 기존연구에서 거의 관찰되지 않았고, Cho (2012)에서만 제약으로 학습되었다. 다음 [jɛ]와 [we] 연쇄의 발생 회피가 고유어와 한자어에서 포착된다(75Ab–c, 75Bb–c). 이들은 [활음][모음]의 낮은 절대 빈도로만 언급된 바 있다(고유어: 한성우 2006, 전체 한국어: 신지영 2011). 또한, *[ɨ][활음] 제약이 고유어 문법과 한자어 문법에 모두 포함되는데(75Ad, 75Bd), 이는 유재원(1997)에서만 언급된 바 있다. 한편, 고유어와 한자어 어휘부에서 [pp^h]가 모두 제한되며(75Ae, 75Be), 고유어 문법은 *[m, p][p’] 제약(75Af)까지 포함한다. 인접한 음절말음과 음절두음이 모두 양순음일 때 회피된다는 점은 수의적 조음위치자질 동화 현상(예: 꽃보다[꼳뽀다~꼽뽀다])과도 배치되며, 제약 설정의 동기를 설명하기는 어렵다. 다만, 신지영(2011)은 전체 한국어 어휘부에서 음절말음과 음절두음 위치에서 모두 양순음의 빈도가 낮은 것을 관찰하였는데 이러한 양순음의 분포 제한이 제약으로 포착되었을 가능성도 있다.

지금까지 이 연구의 고유어 문법과 한자어 문법에서 공통적으로 포함한 제약을 정리하였다. 기존연구에서 범주적 제약 및 비범주적 경향성으로 포착한 일반

화가 고유어 및 한자어 문법의 일부로 학습되었다는 것을 확인하였으며, 두 어휘부 문법이 공통적으로 포착하는 새로운 회피 경향도 보고하였다.

4.2.2. 고유어 문법 특정 음소배열제약

이 절에서는 이 연구의 학습에서 한자어 문법으로는 포착되지 않고, 고유어 문법으로만 포착된 제약을 살펴본다. 해당 제약들을 (i) 기존연구에서 관찰된 제약에 대응하는 것, (ii) 기존연구에서 구체적으로 관찰되지 않은 제약 및 회피 경향, 두 부류로 나누어 정리한다.

첫째, 이 연구에서 고유어 문법으로만 학습된 제약은 기존연구에서 비범주적으로 정의된 제약을 포함하였다. 아래 (76–79)에서 인접 [모음] 연쇄 제약, 비인접 후두자질 층위 제약, 비인접 모음자질 층위 제약 순서로 제시한다.

먼저, 고유어 문법에서는 (76)에 요약하였듯이 다수의 인접 [모음] 연쇄 제약이 학습되었다.

(76) [모음][모음] 제약

번호	제약	가중치	의미	예외
a	*[−low,+syl][+syl]	2.36	*[i, y, ɨ, u, e, ø, ʌ, o][모음]	거울
b	*[−high][−high]	2.86	*[e, ø, ʌ, o, ɛ, a][e, ø, ʌ, o, ɛ, a]	가오리
c	*[−bk,+syl][−bk,+syl]	1.79	*[i, y, e, ø, ɛ][i, y, e, ø, ɛ]	-
d	*[+high][+high]	1.63	*[i, y, ɨ, u][i, y, ɨ, u]	-
e	*[+high,+bk][+rnd,+syl]	1.49	*[ɨ, u, w][y, ø, u, o]	추위

앞서 (67)에서 언급하였듯이 한자어 문법도 [모음][모음] 회피 제약을 포함하지만, 고유어 문법에서 한자어 문법보다 다수의 [모음][모음] 회피 제약이 학습된다. 제약 (76a)는 [ɛ, a]를 제외한 [모음] 일반이 다른 [모음] 앞에 오지 못할 것을 요구한다. 또한, 제약 (76b–d)는 [비고모음] 연쇄, [전설모음] 연쇄, 그리고 [고모음] 연쇄 발생을 전반적으로 제한한다.

앞서 3.2.1절에서 다루었듯이, 유재원(1997)은 동일 [+/−back] 자질 또는 [+/−round] 자질로 구성된 모음 연쇄의 회피를 보고한 바가 있으나 그 통계적 유의미성까지는 논의되지 않았다. 이 연구에서 학습된 고유어 문법은 동일 모음 자질 연쇄 회피, 그 중에서도 전설모음 연쇄 회피를 포착한다는 것은 유재원(1997)과 같지만, 모음의 높이 자질이 동일한 연쇄가 회피된다는 것은 유재원

(1997)과 다르다. 한 가지 언급할 점은 동일 모음자질 연쇄 회피 경향은 언어 유형론적으로 일반적이지 않은 제약이며, 과거 연구에서 어떠한 모음자질이 동일할 경우 회피되는지에 대해서는 구체적으로 논의되지 않았다. 한편, 제약 (76e)는 [w][원순모음] 연쇄 발생 회피와 더불어 [ɨ, u][원순모음] 연쇄 발생 회피를 포착하는데 해당 제약은 Cho (2012)에서도 학습된 바 있다.

다음으로 후두자질 층위 제약을 (77)에 요약해 보면, 고유어 문법으로 후두자질 공기 제약을 포함하여 네 제약이 학습되었다.

(77) 후두자질 층위 제약

번호	제약	가중치	의미	예외
a	*[^−asp,−tns][+asp]	1.95	*[격음, 경음][격음]	까치, 해파리
b	*[+asp][+tns]	1.79	*[격음][경음]	토끼, 팔찌
c	*[^−asp,−tns][+asp][]	1.36	*[격음, 경음][격음][]	-
d	*[][][^−asp,−tns]	1.53	*[][][격음, 경음]	그저께, 버들치

고유어 문법으로 *[경음][경음]을 제외한 후두자질 공기 제약(77a–b)이 학습되었다. 또한 고유어 문법에서는 저해음인 세 음절두음 연쇄에서 [격음, 경음][격음]이 첫 번째와 두 번째에 위치하는 것이 회피되고(77c), [격음, 경음]의 단독 발생이 세 번째 위치에서 제한되는 것(77d)이 포착되었다.

후두자질 층위 제약 (77)의 효과를 파악하기 위해, 음절두음 층위에서 두 자음 연쇄(C_1-C_2)와 세 자음 연쇄(C_1-C_2-C_3)에 대한 비적형성 점수를 구하여 보면 (78)과 같다.

(78) 음절두음 층위 후두자질 공기 위치별 비적형성 점수: 고유어

위치 / 공기 유형	음절두음 두 개	음절두음 세 개		
	C_1-C_2	C_1-C_2	C_2-C_3	C_1-C_3
[격음][격음]	1.95	3.31	3.48	1.53
[경음][격음]	1.95	3.31	3.48	1.53
[격음][경음]	1.79	1.79	3.32	1.53
[경음][경음]	0	0	1.53	1.53

인접한 음절에서 발생하는 [경음][경음]을 제외한 후두자질 공기 형태들은 제약 (77a) *[^-asp,−tns][+asp](가중치 1.95)와 제약 (77b) *[+asp][+tns](가중치 1.79)를 위배한다. 삼음절어에 나타나는 후두자질 공기 유형은 발생 및 공기 위치별로 제약 (77a–b)에 더하여 추가적인 제약을 위배하기도 한다. 삼음절어 C_1-

C_2에 위치하는 후두자음 연쇄는 제약 (77a)에 더하여 제약 (77c) *[^−asp, −tns][+asp][](가중치 1.36)까지 위배하고, 삼음절어 C_2-C_3에 위치하는 후두자음 연쇄는 제약 (77a)와 함께 제약 (77d) *[][][^−asp,−tns](가중치 1.53)가 위배된다. 한편, [격음][경음] 연쇄가 삼음절어 C_2-C_3에 위치할 때에는 제약 (77a)와 함께 제약 (77b) *[+asp][+tns] (가중치 1.79), 제약 (77d) *[][][^−asp, −tns] (가중치 1.53)가 위배된다. 한편, 삼음절어에서 후두자음이 C_1-C_3에 위치하는 경우 모두 세 번째 후두자음 단수 발생 제약(77d)만을 위배하고, 비인접 후두자질 공기 관계를 예측하지 않는 것을 알 수 있다.

또한, 모음자질 층위에서는 모음 조화와 관련된 제약이 학습되었다. 제약(79a)는 모음자질 층위의 세 모음에 대하여 [−high,+ATR]과 [−ATR]이 각각 두 번째와 세 번째 위치에 발생하지 않을 것을 요구한다. 또한, (79b) 제약은 모음자질 층위의 세 모음에 대하여 [+ATR], [−ATR], [−high,+ATR]이 차례로 발생하는 것을 제한한다.

(79) 모음자질 층위 제약

번호	제약	가중치	의미	예외
a	*[][−high,+ATR][−ATR]	3.62	*[][e, ʌ][ø, ɛ, o, a]	-
b	*[+ATR][−ATR][−high,+ATR]	2.44	*[i, y, ɨ, u, e, ʌ][ø, o, ɛ, a][e, ʌ]	치다꺼리

기계 학습된 고유어 문법은 앞서 3.2.2절에서 다룬 Hong (2010)의 예측보다 모음 조화 관련 제약의 적용 범위가 좁다. 이는 기계 학습 모델이 모음자질 층위에 대해서만 연쇄 분포를 조사한 것이 아니라 다른 음소 연쇄를 포괄적으로 탐색하기 때문일 것이다. 이와 같은 탐색으로, 한국어 화자의 문법에 모음 조화 자질 제약이 실재하지만 그 회피 정도가 약한 것을 효과적으로 포착할 수 있다고 본다.[24]

이상 이 연구의 고유어 특정 문법이 [모음][모음] 인접 제약, 후두자질 층위 비인접 제약, 그리고 모음자질 층위 비인접 제약을 포함한다는 것을 확인하였다. 많은 기존연구에서 고유어 어휘부는 한국어 화자의 어휘부를 대표하는 것으로

[24] 실제로 고유어 이음절어에서는 모음 부조화형 '점잔[ʌ-a], 고섶[o-ʌ]' 등이 허용되며, 최근 차용어에서도 모음 조화를 위배하는 연쇄가 발생하는 예('불도저, 레퍼토리') 등이 관찰된다.

보았으며 이에 따라 고유어 문법은 한국어 화자의 음소배열제약을 포함한다고 가정되었다. 특히 이들 제약들은 3.2절에서 논의하였듯이 기존연구에서 음운론적 동기가 있는 것으로 분석되어 한국어 화자의 적형성 판단에 영향을 줄 수 있다.

둘째, 기존연구에서 구체적으로 관찰되지 않은 회피 경향이 고유어 문법의 일부로 학습된 것을 제시한다. 음소 연쇄의 유형별로 정리하여, 단어 경계 자질 명세 제약, [음절두음]//[모음] 제약, [음절말음][음절두음] 제약, 그리고 [활음] 제약 순서로 나열한다(이하 (80–84)).

단어 경계 자질이 명세된 제약부터 살펴본다. 제약 (80a–c)는 단어 경계 자질 [+word_boundary](#)를 포함하여 어두와 어말 위치 제약을 나타낸다. 반면, 제약 (80d)는 단어 경계 자질 [−word_boundary]([]: 분절음)을 포함하여 어중 위치 제약을 뜻한다.

(80) 단어 경계 자질 명세 제약

번호	제약	가중치	의미	예외
a	*[−low,+bk,−rnd]#	2.34	*[ɨ, ʌ]#	건너
b	*#[−bk,+rnd]	2.57	*#[y, ø]	위
c	*#[+high,+bk]	2.11	*#[w, ɨ, u]	우리
d	*[][+cont,+asp]	2.89	어중 [h] 금지	나흘

제약 (80a)는 기존연구에서 보고된 어말 [ɨ] 발생 제한뿐만 아니라 어말 [ʌ] 발생 제한까지 포착하였고, 제약 (80b–c)는 어두 위치에 [y, ø, w, ɨ, u] 발생을 제한하였다. 한편, 제약 (80d)는 어중에서 [h]가 잘 오지 못하는 것을 포착한다. 공명음에 후행하여 [h]가 수의적으로 탈락되는 현상이 분석된 바 있으나(신지영·차재은 2003, 박선 2015), 정적인 제약의 형태로는 보고되지 않은 것으로 보인다.

다음으로, [음절두음][모음] 또는 [모음][음절두음](이하, [음절두음]//[모음]) 연쇄에 대한 제약을 (81)에 요약하였다.

(81) [음절두음]//[모음] 제약

번호	제약	가중치	의미	예외
a	*[+high,+bk,−rnd][−son,+lab,−rhy]	2.33	*[ɨ][p, p’, p^h]	-
b	*[+high,+bk,−rnd][+asp]	1.92	*[ɨ][격음]	끄트머리
c	*[−bk][+asp,+dor]	2.22	*[전설모음][k^h]	-
d	*[−low,−rnd][+asp,+dor]	1.98	*[i, ɨ, e, ʌ][k^h]	서캐
e	*[+high][+asp,+dor]	1.66	*[고모음][k^h]	수크령
f	*[+high,+bk,+rnd][−str,+asp]	2.13	*[ut^h]	-
g	*[−high,+ATR][−str,+asp]	1.94	*[e, ʌ][t^h]	허탕
h	*[−high,−bk][−str,+asp]	1.07	*[e, ø, ɛ][t^h]	-
i	*[+cont,+tns][−high,+rnd]	2.11	*[s’][ø, o]	쏘가리
j	*[+str,+tns][−high,−low,−bk]	2.00	*[s’, c’][e, ø]	족제비

제약 (81a)는 [ɨ][양순음] 연쇄 발생 저지를 포착한다. 이 제약은 (65i) *[+lab][+high,+bk,-rnd] 제약과 구성 자질 매트릭스의 순서만 바꾼 것처럼 보이나, 이 제약에 대해서는 기존연구에서 별다른 논의가 되지 않았다. 이 제약이 *[양순음][ɨ] 제약과 같은 방식으로 이해되고 실재할 수 있을지 향후 연구가 진행되어야 할 것이다. 또한, 제약 (81b–h)는 [모음][격음] 연쇄 중 [ɨ][격음] 연쇄, [모음][k^h] 연쇄, [일부 모음][t^h]의 발생 제한을 요구하고 제약 (81i–j)는 [경음][모음] 연쇄 중 극히 일부인 [s’o, s’e, c’e] 연쇄 발생을 제한한다. 제약 (81b–j)는 격음과 경음의 유표성 회피 외에는 자연스러운 제약 설정 동기를 찾기 어렵다. 이들의 음소배열제약의 심리적 실재에 대한 추후 검토가 필요하다.

다음으로 (82)에서 [음절말음][음절두음]과 관련된 제약을 제시한다.

(82) [음절말음][음절두음] 제약

번호	제약	가중치	의미	예외
a	*[+lab][−tns,+dor]	2.11	*[m, p][k, k^h]	임금
b	*[−son,+cor][+lab]	1.84	*[t][p, p’, p^h]	-
c	*[−appr,+cor][−str,+asp]	1.88	*[t, n][t^h]	-
d	*[−appr,+cor][+asp,+dor]	1.80	*[t, n][k^h]	-
e	*[−appr,+cor][+asp,+lab]	1.61	*[t, n][p^h]	-
f	*[+lab,+rhy][^-son,−rhy]	1.24	*[m]$[m, n]	엄마

제약 (82a–b)는 [양순음][설배음], [t][양순음] 발생을 저지한다. 이들은 수의적인 조음위치자질 동화의 양상과 유사하여 보인다. 한국어에서 주로 [양순음][설

배음]에서 [양순음]이 피동화주로, [설배음]이 동화주로 기능하고, [t][양순음]에서 [t]가 피동화주로 [양순음]이 동화주로 기능한다. 그러나 이들 제약은 앞서 4.2.1 절에서 논의한 *[양순음][양순음] 제약((75) 참조)과는 배치된다. 한국어 화자의 인식에 조음위치자질에 관한 어떠한 음소배열제약이 실재하는지 더 살펴볼 필요가 있다. 제약 (82c–e)는 [n]의 뒤에 [p^h, t^h, k^h]는 [n] 발생을 저지하고,[25] 제약 (82f)는 [m][m, n] 발생 회피를 포착한다. 이들 제약 (82c–f)에 대해서는 제약 설정 동기가 구체적으로 논의되지 않은 것으로 보인다.

또한, 고유어 문법은 [활음]이 포함된 연쇄에 관한 제약도 다수 포함한다. 아래 (83)에서 제시하였듯이, [k'j], [ju], [ɯi] 발생이 제한되며 [모음][활음]에 관한 제약도 학습되었다.

(83) [활음] 제약

번호	제약	가중치	의미	예외
a	*[+tns,+dor][−rnd,−syl]	2.14	*[k'j]	-
b	*[+high][+high]	1.63	*[j, w, ɯ][i, ɨ, u]	-
c	*[+bk,−rnd,−syl]	1.04	*[ɯ]	고의
d	*[−low,+bk,−rnd][−bk,−syl]	2.33	*[ɨ, ʌ][j]	-
e	*[−ATR][+rnd,−syl]	1.34	*[ø, o, ɛ, a][w]	-
f	*[−high,−bk][+bk,−syl]	0.46	*[e, ø, ɛ][w, ɯ]	-

앞서 다룬 [모음][모음] 연쇄 제약 (67), (76)에 더해, 활음이 명세된 제약 (83d–f)는 [모음][활음] 연쇄에 (84)와 같은 비적형성 점수를 부과할 수 있다.

(84) [모음][활음] 비적형성 점수

[모음][활음]	비적형성 점수	[모음][활음]	비적형성 점수
[i, u, y][j, w]	1.63	[e][w]	4.03
[ɛ][j]	2.21	[e][j]	5.78
[ʌ][j]	2.33	[ø][j]	7.68
[o, a][w]	3.53	[ø][w]	9.46
[ɛ][w]	3.99		

[25] 고유어 문법은 [nc^h] 연쇄 회피를 포착하지 않는데, 이는 소신애(2010)에서 논의한 파찰음에 선행한 [n] 삽입 현상의 흔적으로도 볼 수 있다. 통시적 변화의 흔적이 실제 화자의 정적인 음소배열제약 판단 인식에 미치는 영향이 존재하는지에 대한 논의는 다른 기회로 미룬다.

(84)를 보면, [모음][활음] 중 [oj, aj, ʌw]만이 허용된다는 것을 알 수 있다. 그런데, 한국어 용언 활용에서 어간이 [i, e, ɛ]으로 끝날 때 [j]가 생산적으로 삽입되는 것이 관찰되는데(예: 기-어[kiʌ~k**ij**ʌ]), 이는 이 연구가 포착한 *[ij] 제약과는 상충된다.

지금까지 최대 엔트로피 음소배열제약 모델로 학습된 한국어 고유어 문법 특정 제약을 개괄하였다. 이들 제약은 기존연구에서 관찰되고 음운론적 동기에도 부합하는 제약들을 대부분 포함하였다. 추가적으로, 자질 명세가 복잡하게 명세되며 제약 설정의 동기가 구체적으로 설명되지 않은 제약도 학습되었다. 이 과정에서 제약의 일부가 화자의 인식에 실재할 가능성을 부분적으로 논의하였다.

4.2.3. 한자어 문법 특정 음소배열제약

여기에서는 한자어에서 특정적으로 학습된 제약을 살펴본다. 다수의 연구에서 한자어 음절 내에서 고유어보다 제한되는 음소 및 연쇄가 보고되었고, 음절 경계에 걸치는 음소배열제약이 유효하지 않다고 가정되었다. 그러나 앞서 3.3절에서 밝혔듯이, 한자어 어휘부가 한국어 어휘부에서 차지하는 양적 비중이 높은 만큼 한자어 음소배열제약이 체계적으로 탐색될 필요가 있다. 이 절에서는 이 연구의 학습에서 고유어 문법으로는 포착되지 않고, 한자어 문법으로만 포착된 제약을 살펴본다. 학습 제약들은 (i) 기존연구에서 전체 한국어 및 고유어를 대상으로 논의된 음소배열제약과 대응하는 것, (ii) 기존연구에서 한자어 특정적으로 논의된 제약과 대응하는 것, (iii) 이 연구에서 새롭게 학습된 제약, 세 부류로 나누어 제시한다.

첫째, 기존연구가 전체 한국어 및 고유어 어휘부에 대해 관찰한 제약이 이 연구에서 한자어 문법의 일부로 학습된 경우를 보자. 아래 (85)에서 요약하였듯이, *[원순모음][양순음]$ 제약과 음절말음과 음절두음에 상관없이 [양순음] 세 개 또는 [설배음][양순음][양순음]의 연쇄가 회피되는 제약이 학습되었다.

(85) 기존연구가 전체 한국어 어휘부에서 관찰한 제약

번호	제약	가중치	의미	예외
a	*[^−rnd,+syl][+lab,+rhy]	4.87	*[원순 모음][m, p]$	서품
b	*[^+cor][+lab][+lab]	2.02	*[양순/설배음][양순음][양순음]	감미, 흡입

제약 (85a)는 한자어 음절(신지영 2009)뿐만 아니라 전체 한국어(유재원 1997)에 대해서 보고되었으며, [원순성]과 [양순음]이 유사한 자질이라는 점이 회피 동기로 논의되었다. 제약 (85b)는 자음 층위에서 [양순음][양순음] 연쇄가 회피되는 것을 일부 포착한 것이다. 이는 Ito (2007)이 고유어 음절을 대상으로, [음절두음][음절말음] 위치에 음운론적 동일 조음위치 자질의 동시 발생이 제한된다고 제시한 원리(OCP)와 유사해 보인다. 그러나 제약 (85b)는 세 자음에 걸쳐서 적용된다는 점, 그리고 [양순음]에만 동일 자질 회피가 국한된다는 점은 음운론적으로 자연스러운 동일 조음위치자질 회피로만 분석되기 어려워 보인다. 오히려, 한자어 특정적으로 음절 내 [음절두음][음절말음] 위치에 양순음 동시 발생이 회피된다는 기술(신지영 2009)과 부분적으로 일치한다고 볼 수도 있다.

둘째, 다수의 기존연구(권인한 1997, 강용순 1998, 신지영·차재은 2003, 신지영 2009, 안소진 2009)에서 고유어와 변별하여 직접 관찰한 특징이 이 연구의 문법에 어떻게 반영되어 있는지 살펴본다. 개별 음소 분포에 대한 제약, [음절두음][모음] 제약 순서로 제시한다. 먼저 개별 음소 분포에 대한 제약을 정리하여 (86)으로 요약하였다.

(86) 기존연구가 한자어에서 관찰한 개별 음소 분포 제약

번호	제약	가중치	의미	예외
a	*#[+asp,+dor]	3.97	*#[k^h]	쾌자
b	*#[−cont,+tns,+cor]	2.65	*#[t’, c’]	-
c	*[−son,+cor,+rhy]	3.35	*[t, s, c]$	곶

이 연구에서 학습한 한자어 문법이 예측하는 개별 음소 분포제약은 기존연구에서 기술한 한자어 음절 특징과 일치된다. 제약 (86a)는 어두 위치의 [k^h] 발생을 저지하고, 제약 (86b)는 어두 위치의 [t’, c’]의 발생을 제한한다. 또한 제약 (86c)는 예외가 발생하지 않는 [설정 저해음]인 음절말음 발생을 저지한다.

이 연구에서 학습한 한자어 문법은 기존연구가 밝힌 경음 회피 경향을 후두자질 층위 제약으로도 포착한다. (87)에서 요약하였듯이, 한자어 문법은 [경음]이 다른 [저해음]에 앞서거나(87a) 어두 위치에 발생하는 것(87b)을 제한한다.

(87) 후두자질 층위: 경음 회피 제약

번호	제약	가중치	의미	예외
a	*[+tns][]	3.55	*[경음][저해음]	끽연
b	*#[+tns]	1.77	*#[경음]	쌍, 약간

다음으로 다수의 연구가 한자어의 구성 음절 특징으로 지적한 [음절두음][모음] 제약을 살펴본다. (88)에서 볼 수 있듯이, [치경 저해음][e, ʌ] 연쇄 제한(88a)과 [공명 자음][e, ʌ] 연쇄 제한(88b–c)이 포착된다.

(88) 기존연구가 한자어에서 관찰한 제약: [음절두음][모음]

번호	제약	가중치	의미	예외
a	*[−cont,+ant][−high,+ATR]	3.41	*[t, t’, t^h, c, c’, c^h][e, ʌ]	덕
b	*[+cons,+son,−rhy][−low,+bk,−rnd]	2.62	*[m, n, l][ɨ, ʌ]	금언
c	*[+cons,+son][−high,−low,−bk]	2.12	*[m, n, ŋ, l][e, ø]	뇌

셋째, 이 연구의 한자어 문법에서 새롭게 학습된 제약을 (89)에 정리하여 제시한다.

(89) 새롭게 학습된 제약

번호	제약	가중치	의미	예외
a	*[−son,+dor][−high,−low,−bk]	1.98	*[k, k’, k^h][e, ø]	게시
b	*[+high,+bk,−rnd][−nas,+cor]	3.23	*[ɨ][l, 설정 저해음]	슬하
c	*[−high,+rnd][−nas,+cor,+rhy]	1.79	*[ø, o][l, t]$	돌기
d	*[−low,+bk,−rnd][+asp,+lab]	1.73	*[ɨ, ʌ][p^h]	저포
e	*[^+bk,+syl][+asp,+dor]	1.97	*[전설모음,자음][k^h]	백합
f	*[+cont][+asp,+lab]	1.67	*[lp^h]	살포
g	*[+high,+bk,−rnd][−appr,+cont]	2.21	*[ɨ][s, s’, h]	-

이들 제약은 제약을 구성하는 자질 매트릭스의 자질 명세가 복잡하고, 제약 동기도 분명하지 않다. 예를 들어, 제약 (89c)는 [ol] 연쇄 발생을 제한하지만, [l] 앞에 [o]만이 회피될 동기는 찾기 어렵다. 이에 따라, 한국어 화자들이 ‘골, 솔’ 등의 단어에 대한 비적형적이라고 인식할 수 있을지 의문을 남긴다.

추가적으로 언급할 것은 한자어 어휘부에서 삼음절어의 발생을 강하게 회피하는 제약이 학습되었다. 제약 (90a)는 모음자질 층위에서 세 모음의 발생, 즉, 삼

음절어의 발생을 강하게 막는 한편, 제약 (90b)는 후두자질 층위에서 저해음인 음절두음 세개를 금지한다. 이는 학습 자료인 한자어 어휘부가 예외 한 개(신기루) 외에 모두 이음절어로 구성된 것이 반영된 것이다. 이제까지 실제 화자의 정적인 제약에 대한 적형성 판단에서 음절수 자체의 영향이 보고된 바 없다는 점을 고려하면, 제약 (90)은 화자의 적형성 판단 인식과 다소 거리가 있을 것으로 보인다.

(90) 삼음절어 제약

번호	제약	가중치	의미	예외
a	*[][][](모음자질 층위)	8.94	단어 내 세 개의 모음 금지	신기루
b	*[][][](후두자질 층위)	1.62	저해음인 음절두음 세 개 금지	-

지금까지 한자어 문법 특정 음소배열제약을 다루었다. 학습된 한자어 문법은 기존연구에서 전체 어휘부를 대상으로 관찰한 제약 일부를 포함하며((85) 참조), 기존연구에서 한자어에 특정하여 기술한 제약도 학습하였다는 것을 확인하였다((86–88) 참조). 또한, 이 연구에서 새롭게 학습된 제약((89–90) 참조)을 제시하고 제약의 실재 가능성을 논의하였다.

한자어 문법은 전통적으로 한국어 화자의 문법을 대표하지 않는다고 여겨졌으며, 한자어 문법에 포함된 제약 중 (85)를 제외하면 기존연구에서 제약의 설정 동기를 분명히 언급한 경우도 없는 것으로 보인다. 그러나 복수의 연구에서 직관적으로 고유어와 변별되는 한자어의 특징을 기술하고 있으며, 한국어 명사에서 한자어가 차지하는 양적 비중을 고려할 때 이 연구에서 학습된 한자어 음소배열제약이 한국어 화자의 적형성 판단에 영향을 미칠 가능성이 있다.

4.3. 논의

지금까지 이 연구에서 학습한 고유어 및 한자어 어휘부 문법이 포함하는 음소배열제약을 제시하고 논의하였다. 기계 학습을 통해, 기존연구에서 관찰된 전통적 제약뿐만 아니라, 통계적으로 정의될 수 있는 다수의 제약을 연구자의 개입 없이 일괄적으로 포착할 수 있었다. 또한, 고유어와 한자어 목록을 구분하여 학습함으로써 고유어 어휘부 문법과 한자어 어휘부 문법의 공통점과 차이점을 체계적으로 살펴볼 수 있었다.

그러나 4.2절에서 다룬 제약 모두가 화자의 인식에 실재한다고 보기는 어렵다. 앞서 2.6절에서 밝힌 바와 같이, 최대 엔트로피 음소배열제약 모델 학습은 어휘부의 연쇄 분포를 귀납적으로 계산하여 문법을 구성하며, 학습된 문법은 이른바 '우연한 제약(accidentally-true constraint)'을 포함할 수 있다(Hayes & Wilson 2008). 일부 연구(Hayes & White 2013, Prickett 2015)는 우연한 제약도 화자의 인식에 실재할 수 있지만 그 정도가 약하고 자연스러운 제약을 강하게 인식하는 일종의 자연성 편향이 있다고 보고한 바 있다. 이 연구에서 학습한 문법에서도 일부 제약(예: (81f) *[+high,+bk,+rnd][−str,+asp]; *[ut^{h}])은 음운론적으로 자연스럽지 않고 직관적으로 이해되지 않아, 해당 제약이 한국어 화자의 인식에 분명하게 실재한다고 판단하기 어렵다.

한편, 4.2절에서 정리하여 제시한 바와 같이 학습 자료인 고유어 어휘부와 한자어 어휘부별로 문법이 구성된다. 고유어 어휘부 문법과 한자어 어휘부 문법 모두가 한국어 화자의 적형성 판단 인식에 기여할 수 있을지에 대한 의문도 남는다.

이에 따라, 한국어 화자의 적형성 판단 조사가 필수적으로 요구된다. 학습 문법이 예측한 음소배열제약 인식은 한국어 화자의 적형성 판단 조사를 설계하는 기준이 될 수 있으며, 어휘부별 문법의 기여도를 테스트할 수 있는 근거를 제공한다. 5장에서는 이러한 학습 결과를 바탕으로 한국어 화자 대상의 적형성 판단 조사를 수행한다.

5. 모국어 화자의 적형성 판단 조사

5.1. 목적과 대상

4장에서 최대 엔트로피 음소배열제약 모델을 사용하여 한국어 어휘부를 바탕으로 습득된 음소배열제약 문법을 제시하였고, 그 문법이 범주적 및 비범주적 적형성을 예측한다는 것을 보였다. 이 장에서는 어휘부에서 학습된 비범주적 음소배열제약의 심리적 실재를 확인하는 한편, 어휘부의 종류에 따라 다르게 습득된 문법들이 한국어 모국어 화자의 적형성 판단에 어떻게 반영되는지를 알아보기 위해 한국어 화자를 대상으로 적형성 판단 실험을 실시하였다.

습득된 모든 제약을 대상으로 하는 실험은 현실적으로 가능하지 않은 바, 기존연구의 검증 및 실험 결과에 대한 해석이 비교적 용이하다고 판단된 '후두자질 발생 및 공기 제한'을 중심 주제로 선택해서 실험을 준비하였다. 이 제약은 범언어적으로 관찰되는 후두자질상 유표적인 '방출음(ejective)', '유기음(aspirate)', '내파음(implosive)' 등의 발생 및 공기 제한을 가리킨다(MacEachern 1999, Gallagher 2010 등).

3장에서 밝힌 바와 같이, 한국어에 대해서도 [경음]과 [격음]의 규칙적 공기 제약이 보고되었다. 기존연구에서 한국어 화자를 대상으로 '합성어 경음화'와 '어두 경음화'의 [경음] 발생 및 인식 조사가 실시되었고, 그 결과 한국어 화자의 '합성어 경음화' 발생 비율의 결정에 있어서 [경음]//[격음, 경음]의 회피가 기능하고(Ito 2014, S. Kim 2016), '어두 경음화' 발생 비율에는 [경음][경음]의 선호가 기능한다는 점이 밝혀졌다(H. Kim 2017, Kang & Oh 2016). 이 가운데 [경음][경음]의 회피(이화)와 선호(동화)가 상충되는 측면도 지적되었다(Kang & Oh 2019). 이에 반해, 어휘부 내에서도 정적인 제약형태로 보고된 바 있으나(김경일 1985, Ito 2014, Kang & Oh 2016), 이에 대한 화자의 심리적 실재 여부가 직접적으로 탐색된 바는 없는 것으로 보인다. 4장에서 제시한 바와 같이, 이 연구의 학습 결과에서도 고유어와 한자어에 대해서 후두자질 발생 및 공기 제한이 예측되었기 때문에 '후두자질 발생 및 공기 제한'에 대한 한국어 화자의 심리적 판단을 조사할 가치가 있다고 보았다. 그 결과를 4장에서 제시한 문법들의 예측과 비교함으로써 기계 습득된 문법의 현실성 및 설명력에 대한 검증을 하고자 한다.

5.2. 연구 방법

한국어 비단어를 만들어 실험 자극으로 제시하고, 한국어 모국어 화자들을 대상으로 실험 단어의 적형성을 리커트(likert) 척도에 따라 판단하도록 하였다. 실험에 사용한 비단어는 아래 (91)에 제시한 바와 같이 이음절어와 삼음절어로 구성되어 있으며, 철자 및 음성녹음 두 가지 형태로 피실험자들에게 제시되었다. 모든 피실험자는 한 가지 유형에만 참가하였다.

(91) 조사 구성

유형	실험 자극 음절수	실험 자극 형태
1	2	철자
2		음성
3	3	철자
4		음성

실험 단어의 생성과 실험의 진행 과정을 이하에서 구체적으로 기술한다.

5.2.1. 조사 단어 생성

실험 자극은 포함된 후두자음의 종류(격음, 경음), 개수(0, 1, 2) 및 단어 내 위치(1, 2, 3음절)를 기준으로 다양하게 생성되었다. 각 음절은 음운론적 복잡성을 최소화하기 위해 개음절(CV)로만 구성되었고, 음절두음은 저해음(obstruent) 중에서 조음위치자질 회피 제약(OCP)의 영향을 받지 않도록 조음위치가 다른 것들로 이루어졌으며, 모음은 [i, ɨ, u, ʌ, o, a] 중에서 선택되었다.

위의 조건을 만족하는 모든 가능한 단어들로 후보군을 일차 작성하였고, 그 중에서 피실험자들의 응답결과에 대한 해석이 용이하도록 후두자질 공기 제약 이외의 제약들을 최소로 위배하는 단어들을 위주로 선별하였다. 구체적으로는 고유어 문법(고유 단일어 명사 어휘부를 입력자료로 삼아 습득된 문법)이 부여한 비적형성 점수가 낮은 것 위주로 이음절어 139개(C_1VC_2V, 예: 차파, 꼬빠), 삼음절어 114개($C_1VC_2VC_3V$, 예: 파코두, 또빠기)를 선정하였다. 이에 더하여, 기타 제약까지 위배하는 단어를 임의적으로 선별하였다. 이에 해당하는 것은 이음절어 18개(예: 카티, 띠커), 삼음절어 38개(예: 크프더, 프디커)다(전체 실험 단

어 목록은 [부록3]에 제시). 덧붙여, 사전에 등재되어 있고 후두자음을 포함하는 12개의 실제 한국어 단어들의 발음 형태를 필러(filler)로 포함시켰다.

(92) 필러: 사전 등재어 12개

번호	조사 제시형	사전 등재형
a	뽑깨	뽑개
b	떡삐	떡비
c	삽까	삽가
d	직뿌	직부
e	다끼	닦이
f	가삐	가삐
g	흐코	흑호
h	처푸	첩후
i	구콰	국화
j	이팍	입학
k	섭코	섭코
l	죽피	죽피

5.2.2. 조사 과정

피실험자의 중심 과제는 제시된 실험 단어의 적형성, 즉 한국어 단어다움을 판단하여 해당하는 점수로 응답하는 것이다. 단어의 발음만을 기준으로 실제 한국어 단어로 들릴 가능성이 높으면 7점, 그럴 가능성이 낮으면 1점에 가깝게 응답하도록 요청했다.

철자 형태의 실험 자극을 사용한 실험의 진행 과정은 다음과 같다. 전체 소요 시간은 약 30–40분 정도 걸렸다.

1단계: 실제 한국어 단어의 발음형 20개[26]를 임의적인 순서로 제시하여, 연구 참여자가 발음에 집중할 수 있도록 하였다.

2단계: 연습 단계로써, 기존연구에서 언급된 비적형 연쇄(예, [pɨ, ci, ti])를 포함하는 비단어 12개를 제시하여 피실험자가 1–7점 사이의 점수 매기기에 익숙하게 하였다.

[26] 실제 단어는 고유 단일어 명사 중 고빈도 단어 200개로 구성되었다. 발음형은 한글로 제시되는데, 예를 들어 '할아버지'를 '하라버지'로 제시하는 식이다.

3단계: 본 조사 단계로써, 제시된 실험 단어에 대해 1–7점 사이의 점수를 사용하여 적형성을 표시하게 하였다.

음성 녹음 형태의 실험 자극을 사용한 실험의 진행 과정은 다음과 같다.

1단계: 연습 단계로써, (위 2단계에서 쓰인) 연습 단어들을 듣고 점수로 응답을 하게 하였으며, 그리고 들은 발음을 한글로 직접 쓰도록 하였다.

2단계: 본 조사 단계로써, 들은 실험 단어에 대해 1–7점 사이의 점수를 이용하여 적형성을 표시하게 하였다.

전체 소요 시간은 약 40–60분 정도 걸렸다. 조사에 쓰인 음성 녹음은 서울말 화자인 저자의 발화를 녹음한 것으로 68dB로 정규화하였다.

5.2.3. 피실험자

서울대 학생 커뮤니티 사이트(스누라이프) 및 페이스북에 조사 링크를 게시하여 만 19세 이상 한국어 모국어 화자를 모집하였다. 조사 항목에 모두 응답한 참여자는 총 112명이다. 이 중 모든 비단어에 대해서 똑같은 점수로 응답한 두 명을 제외한 110명을 분석대상자로 삼았다.[27]

(93) 피실험자의 수

유형	실험 자극 음절수	실험 자극 형태	피실험자	분석 대상자
1	2	철자	43명	42명
2		음성	17명	16명
3	3	철자	32명	32명
4		음성	20명	20명

[27] 두 명의 피실험자는 단어의 적형성을 판단하라는 과제에 대해 잘 이해하지 못한 것으로 보인다. 철자 형태의 이음절어에 참여한 한 명의 피실험자는 비단어 157개뿐만 아니라, 사전 등재어 12개에 대해서도 모두 1점을 부여하였다. 한편, 음성 형태의 이음절어에 응답한 한 피실험자는 비단어에 대한 유효 응답 72개 중 한 개를 제외하고 모두 1점을 부여하였다. 사전 등재어 12개에 대한 응답도 살펴보면, 사용 빈도가 높은 '국화'와 '입학'에 대해서만 7점을 부여하였다. 이 피실험자는 단어의 적형성이 아닌 실재 여부만을 판단하였을 가능성이 있다.

분석된 전체 응답의 실험유형별 분포는 다음 (94)와 같다.

(94) 실험 유형별 응답 분포

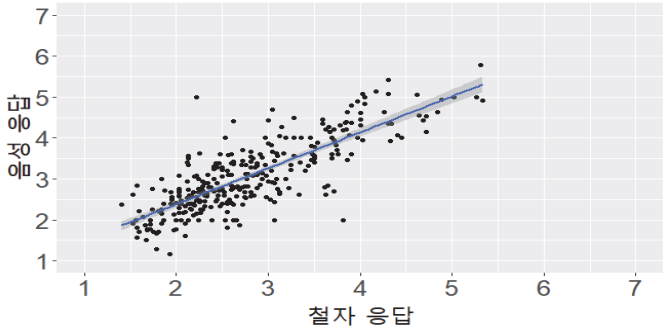
유형	자극 음절수	자극 형태	응답수	분석 응답
1	2	철자	157 (개) × 42 (명) = 6,594	6,594
2		음성	157 (개) × 16 (명) = 2,512	1,691
3	3	철자	152 (개) × 32 (명) = 4,864	4,864
4		음성	152 (개) × 20 (명) = 3,040	1,959

자극 형태가 음성인 경우, 피실험자는 비단어를 이 연구에서 의도한 것과 달리 인지하기도 하였다. 이에 따라, 피실험자가 비단어 발음형을 이 연구에서 의도한 것에 부합하여 기입한 경우만을 분석에 반영하였다. 예를 들어, 피실험자가 '파찌'에 대해서 '팥지, 팟지, 파치' 등으로 발음형을 기입한 경우, 이에 대한 적형성 판단 점수는 분석에서 제외된다.

5.3. 결과

본 장에서는 실험 자극에 대해 피실험자들이 부여한 적형성 판단 점수의 분석 결과를 제시한다. 우선 자극으로 철자와 음성녹음을 사용한 실험의 결과는 평균 점수도 유사하고(철자 2.79, 음성 3.15), 아래 (95)번 그래프에서 볼 수 있듯이 상관관계도 상당히 높게 나타났다($r(307) = 0.807$, $p<0.001$).[28]

(95) 철자 응답(x축)과 음성 응답(y축)과의 상관관계



[28] 철자와 음성녹음 두 가지 실험 자극을 사용한 기존연구의 결과와 비교해 볼 때, 본 실험의 상관관계는 높은 편인 것으로 보인다. 예를 들어 Bailey & Hahn (2001: 580)은 0.60의 상관관계를 보고하고 있다.

따라서, 이후 제시하는 결과분석은 필요한 경우가 아니면 자극의 형태를 구분하지 않기로 한다.

5.3.1. 실험 관찰과 문법 예측 비교

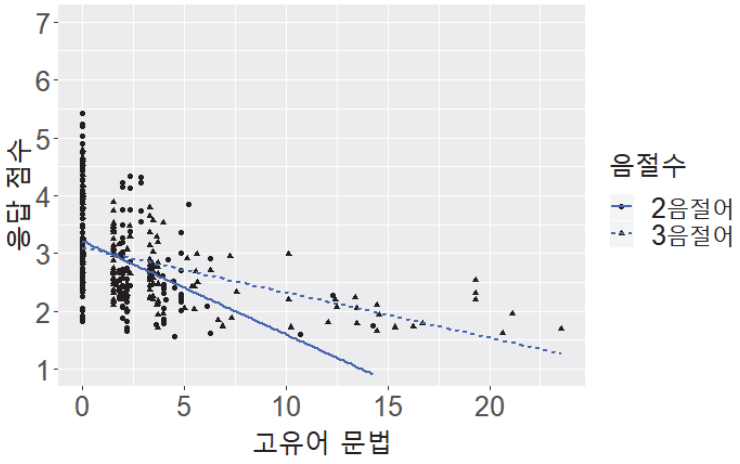
4장에서 습득된 문법들이 실험 결과로 관찰된 모국어 화자의 적형성 판단과 일치하는 예측을 하는지 알아보기 위해, 실험 자극별로 피실험자 적형성 판단의 평균 점수와 문법들이 부여한 비적형성 점수를 비교하기로 한다.

우선, 전체 자극에 대한 적형성 판단 점수와 고유어 및 한자어 문법의 비적형성 점수의 상관관계를 계산하면, 두 문법 모두 유의미한 음의 상관관계를 보인다(고유어 문법: $r(307)= -0.460$, 한자어 문법: $r(307)= -0.500$). 이 결과는 어휘부를 바탕으로 습득된 문법들의 예측이 실제 한국어 화자들의 적형성 판단과 어느 정도 일치하고, 한자어 문법이 고유어 문법보다 더 일치도가 높은 것을 보여준다. 아래 (96–98)에서는 자극의 음절수에 따른 문법 예측-실험 관찰 사이의 상관관계를 제시하고 있는데, 음절수에 관계없이 한자어 문법이 고유어 문법보다 더 높은 상관관계를 보이고 있음을 알 수 있다.

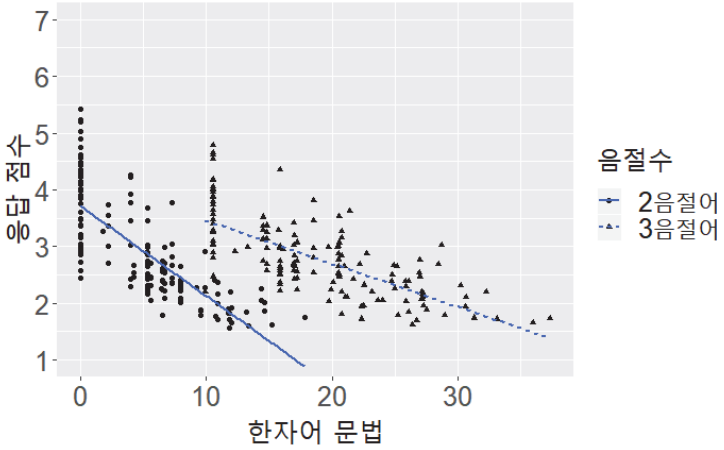
(96) 적형성 판단 점수와 어휘부 문법별 비적형성 점수의 상관 계수

자극 음절수 / 어휘부 문법	전체	이음절어	삼음절어
고유어	−0.460	−0.412	−0.574
한자어	−0.500	−0.779	−0.723

(97) 고유어 문법 비적형성 점수(x축)와 적형성 판단 점수(y축)의 상관관계



(98) 한자어 문법 비적형성 점수(x축)와 적형성 판단 점수(y축)의 상관관계



문법 예측과 실험 관찰 사이의 일치 여부 및 일치 정도에 대해서 좀 더 엄밀한 통계적 분석을 실시하기 위해, 혼합 효과 선형 회귀 분석 모델을 채택하였다. 통계 분석은 R (R Development Core Team 2019)로 진행하였으며, lmerTest 패키지(Kuznetsova et al. 2017)의 lmer 함수를 이용하였다.

피실험자의 '적형성 응답 점수'를 종속 변인으로 삼았다. 독립 변인은 고정 요인과 임의 요인 두 가지를 모두 포함하였는데, 고정 요인은 문법이 부여한 비적형성 점수와, 실험 자극 음절수, 그리고 실험 자극 유형이고, 임의 요인은 피실험자와 비단어다. 고정 요인 중 각 어휘부의 학습 비적형성 점수는 연속적인 수

치인 한편, 실험 자극 유형과 실험 자극 음절수는 범주적인 요인이다. 이 두 범주적 요인에 sum coding을 할당한다.

(99) 독립 변인

고정 요인	임의 요인
고유어 문법 비적형성 점수	피실험자
한자어 문법 비적형성 점수	비단어
실험 자극 음절수(이음절 vs. 삼음절)	
실험 자극 형태(음성 vs. 철자)	

문법 비적형성 점수와 실험 자극 음절수의 상호작용항을 포함하여 우도비율 검정(likelihood ratio test)을 진행한 결과, 최소 요인 적합 모델은 다음 (100)과 같다. 모델 (100)에서 고유어 문법 비적형성 점수와 실험 자극 음절수의 상호 작용은 모델 적합에 기여하지 못하여 제외된다($x^2(1) = 0.049, p = 0.83$).

(100) 혼합효과 선형 회귀 분석 결과(고정 요인)

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.00	0.12	218.22	32.19	<0.0001***
고유어 문법	−0.04	0.01	311.50	−4.52	<0.0001***
한자어 문법	−0.10	0.01	307.42	−16.43	<0.0001***
음절수(삼음절)	0.13	0.12	236.97	1.08	0.28
자극 형태(철자)	−0.20	0.10	107.35	−1.91	0.06
한자어 문법:음절수(삼음절)	0.05	0.01	305.64	8.03	<0.0001***

(100)의 결과는 각 어휘부 학습 문법이 유의미하다는 것을 보여 준다. 첫째, 고유어와 한자어 문법의 비적형성 점수가 높을수록 모두 적형성 판단 점수는 낮다. 둘째, 계수의 절대값을 기준으로 볼 때, 한자어 문법의 비적형성 점수가 고유어 문법의 비적형성 점수보다 적형성 판단 점수에 영향이 더 크다는 것을 알 수 있다. 그럼에도 우도비율 검정 결과, 고유어 문법은 최종 통계 모델에서 제외될 수 없었다($x^2(1) = 20.0, p < 0.001$). 이는 한자어 문법뿐만 아니라 고유어 문법도 실험 결과를 설명하는 것에 독립적인 기여를 하고 있다는 것을 의미한다.

실험 자극의 음절수는 적형성 판단 점수에 유의미한 영향을 끼치지 않는다. 이는 한자어 문법이 삼음절어 회피 제약(*[][][], 가중치: 8.94)이 예측한 것과는 일치하지 않으며, 한자어 문법의 비적형성 점수가 적형성 판단 점수에 미치는 영향이 실험 자극의 음절수에 따라 다른 것으로 분석되었다.

이상에서, 한국어 화자의 심리적 적형성 판단을 나타내는 실험 결과와 어휘부를 바탕으로 습득된 문법의 예측을 비교하였다. 그 결과, 어휘부에서 학습된 제약이 한국어 화자의 인식과 유의미한 상관관계를 보인다는 점을 확인하였으며, 개별 어휘부 문법의 부분적인 기여도를 파악할 수 있었다.

5.3.2. 격음/경음 발생 유형별 학습 점수와 응답 점수

이 절에서는 격음/경음 발생 유형에 대한 한국어 화자의 인식을 살펴보고, 학습 모델의 예측과 비교한다. 5.3.2.1절에서는 실험의 주요 조건인 후두자음 발생 유형에 따라, 고유어 문법과 한자어 문법이 비단어에 부과한 비적형성 점수를 살펴본다. 5.3.2.2절에서는 실제 피실험자가 비단어에 응답한 적형성 판단 점수를 제시하고, 이를 5.3.2.1절의 학습 점수와 비교한다.

5.3.2.1. 학습 점수

앞서 4장에서 제시한 학습 문법을 적용하여 비단어(이음절어: C_1VC_2V, 삼음절어: $C_1VC_2VC_3V$)에 부여한 비적형성 점수를 구한다. 비단어는 각 어휘부의 인접 제약(예: (74Ad) *[^+cons,+rhy][+tns,+cor]) 또는 비인접 후두자질 공기 제약(예: (77a) *[^−asp,−tns][+asp])을 함께 위배할 수 있으며, 위배한 제약의 가중치를 더하여 비단어의 비적형성 점수를 구한다. 비단어의 유형을 후두자음의 개수 및 종류, 단어 내 발생 위치에 따라 나누고, 각 유형이 나타내는 비단어 적형성 판단 점수를 제시한다.

첫째, 고유어 문법과 한자어 문법 모두 후두자음 발생이 두 개인 경우가 한 개인 경우보다 비적형성 점수가 높다. 후두자음 개수가 많을수록 화자의 적형성 판단 점수가 낮을 것으로 예측된다.

(101) 후두자음 개수에 따른 비적형성 점수

	전체		이음절어		삼음절어	
	고유어	한자어	고유어	한자어	고유어	한자어
0	0.99	4.45	0.68	0.40	1.56	11.79
1	1.65	8.48	1.06	3.65	2.59	16.34
2	4.05	14.46	2.96	7.20	4.98	20.71

둘째, (102)와 같이 후두자음 종류에 따른 비적형성 점수가 어휘부에 따라 달리 예측되었다.

(102) 후두자음 종류에 따른 비적형성 점수

후두자음 종류		비적형성 점수	
격음	경음	고유어	한자어
Yes	No	2.61	6.92
No	Yes	2.24	15.18

- Yes = 후두자음 발생 한 개, 두 개
- No = 후두자음 비발생

고유어 문법이 격음을 포함한 자극(격음 Y, 경음 N)에 부과한 비적형성 점수의 평균(2.61)은 경음을 포함한 자극(격음 N, 경음 Y)의 비적형성 점수의 평균(2.24)보다 다소 높다. 반면, 한자어 문법이 경음을 포함한 자극(격음 N, 경음 Y)에 부과한 비적형성 점수의 평균(15.18)이 격음을 포함한 자극(격음 Y, 경음 N)에 부과한 비적형성 점수의 평균(6.92)보다 두 배 이상 높다. 이에 따라, 한자어 문법에서 경음 발생 회피를 강하게 예측하는 것을 확인할 수 있다.

셋째, 각 어휘부 문법은 단어 내 후두자음의 공기 위치에 따라 비단어의 적형성 판단 점수가 다를 것을 예측한다. 고유어 문법에 관해 (103)에서 살펴보면, 고유어 문법은 삼음절어인 비단어($C_1VC_2VC_3V$)에서 후두자질 공기가 C_2-C_3 위치에서 발생할 때 비적형성 점수를 가장 높게 부과한다.

(103) 후두자질 공기 위치별 비적형성 점수: 고유어[29]

위치 / 공기 유형	이음절어	삼음절어		
	C_1-C_2	C_1-C_2	C_2-C_3	C_1-C_3
[격음][격음]	3.05	5.11	5.92	4.2
[경음][격음]	3.44	5.52	7.6	4.01
[격음][경음]	3.52	4.55	6.52	3.95
[경음][경음]	1.63	3.19	4.92	4.29

위와 같은 연쇄가 비적형성 점수를 높게 받는 이유는 삼음절어 내 후두자음이 C_2-C_3에 위치하는 경우(예: 그뻐투, 보쿠뚜), 후두자질 공기 제약((77) 참조)뿐만 아니라 *[모음][경음]((74A) 참조) 제약 또는 *[모음][격음]((81a–h) 참조) 제약이 복수로 위배될 수 있기 때문이다.

한편, 한자어 문법은 비단어 내 [경음] 발생 위치에 따라, 비단어 간 비적형성 점수의 차이를 예측한다. (104)에서 살펴보면, 한자어 문법은 후두자질 공기 유형 중 삼음절 비단어에서 경음이 C_2에 위치할 때 비적형성 점수가 높은 편이다.

(104) 후두자질 공기 위치별 비적형성 점수: 한자어

위치 / 공기 유형	이음절어	삼음절어		
	C_1-C_2	C_1-C_2	C_2-C_3	C_1-C_3
[격음][격음]	2.37	14.71	14.51	15.28
[경음][격음]	7.33	17.71	22.47	20.3
[격음][경음]	7.71	24.48	18.46	19.97
[경음][경음]	12.48	27.88	28.1	24.64

앞서 (87)에서 제시한 한자어 후두자질 층위 제약 *[+tns][](가중치 3.55), *#[+tns] (가중치 1.77)은 경음이 세 개의 음절두음 중 C_1에 위치할 경우 비적형성 점수가 높을 것을 예측하였다. 그러나 이 연구에 쓰인 삼음절어인 비단어에 대해 비적형성 점수를 실제로 구하면, 후두자질 층위 제약뿐만 아니라 인접 연쇄 제약이 관여하여 [경음]이 세 개의 음절두음 중 C_2에 위치할 경우의 비적형성 점수가 가장 높다. [경음][격음], [경음][경음] 유형이 C_2-C_3 위치할 때, 비적형

[29] 밑줄은 후두자질 공기 유형별로 가장 높은 비적형성 점수를 받을 때의 점수를 표시한 것이다.

성 점수가 높은 이유는 후두자질 층위 제약 *[+tns][](가중치 3.55)과 함께 (74B)에서 제시한 *[모음][경음] 제약(예: *[^-nas,+rhy][+tns], 가중치: 3.94)이 위배되기 때문이다. 한편, [격음][경음] 유형은 C_1-C_2 위치에서 비적형성 점수가 가장 높다. 그 이유는 이 유형의 비단어에서도 [경음]이 C_2에 위치하여, 비단어의 [모음][경음] 인접 연쇄가 *[모음][경음] 제약((74B) 참조)을 위배하며, 비단어의 C_2-C_3에 위치한 [경음][평음] 비인접 연쇄가 제약 (87a) *[+tns][](가중치 3.55)를 위배하기 때문이다.

지금까지 고유어 및 한자어 문법이 비단어에 대해 부과한 비적형성 점수를 후두자음의 발생 개수, 종류, 그리고 후두자질 공기 위치에 따라 살펴보았다. 고유어 문법과 한자어 문법은 모두 비단어의 후두자음 발생 개수가 많을수록 비적형성 점수를 높게 예측한다는 점에서 공통적이다. 그러나 고유어 문법은 후두자음의 종류에 따른 비적형성 점수의 차이를 크게 예측하지 않지만, 한자어 문법은 경음이 포함된 비단어의 비적형성 점수를 격음이 포함된 비단어의 비적형성 점수보다 높게 예측한다. 또한, 고유어 문법은 삼음절어에서 후두자질층위 제약이 포착한 후두자질 공기 위치에 따른 비적형성 점수의 차이를 예측한다. 반면, 한자어 문법은 삼음절어에서 경음의 발생 위치에 따라 비적형성 점수의 차이를 예측한다. 이러한 학습을 바탕으로 각 어휘부 문법에서 포착된 후두자음 발생 및 공기에 대한 제약이 화자의 직관에 실재하는지를 알아보도록 한다.

5.3.2.2. 적형성 판단 점수

실제 피실험자의 적형성 판단 점수를 확인하고, 각 문법에서 예측된 학습 점수와 비교한다. 첫째, 피실험자의 적형성 판단 점수는 후두자음의 개수가 0, 1, 2의 순서로 낮아지는 것으로 나타났다(없음: 4.18, 한 개: 3.22, 두 개: 2.45). 이는 고유어와 한자어 문법이 예측하는 바와 일치한다.

(105) 후두자음 개수에 따른 적형성 판단 점수

전체 자극 단어	응답 점수	이음절어	응답 점수	삼음절어	응답 점수
0	4.18	0	4.27	0	4.00
1	3.22	1	3.25	1	3.17
2	2.45	2	2.37	2	2.52

둘째, 후두자음 중에서 격음만으로 구성된 실험 자극과 경음만으로 구성된 실험 자극을 대상으로, 적형성 판단 점수를 구한다(격음: 3.32, 경음: 2.52). 그 결과, 경음을 포함한 자극에 대한 피실험자의 적형성 판단 점수는 격음을 포함한 자극보다 낮다. 이러한 결과는 고유어 문법의 예측과는 일치하지 않고 한자어 문법에서 예측된 바와 일치한다.

(106) 후두자음 종류에 따른 적형성 판단 점수

후두자음 종류		적형성 판단 점수
격음	경음	
Yes	No	3.32
No	Yes	2.52

- Yes = 후두자음 발생 한 개, 두 개
- No = 후두자음 비발생

셋째, 후두자질 공기 위치에 따라 적형성 판단 점수의 차이가 다소 있는 것으로 보인다. 다음 (107)에서 요약하였듯이, 삼음절어인 비단어에서 [격음][격음], [격음]//[경음] 유형이 C_2-C_3에 위치하는 경우에 적형성 판단 점수가 가장 낮은 편이며, [경음][경음] 유형은 각각 C_2-C_3, C_1-C_3 에 위치하는 경우의 적형성 판단 점수(C_2-C_3: 2.18, C_1-C_3: 2.15)가 C_1-C_2보다 약간 낮다.

(107) 후두자질 공기 위치에 따른 적형성 판단 점수[30]

위치 / 공기 유형	이음절어	삼음절어		
	C_1-C_2	C_1-C_2	C_2-C_3	C_1-C_3
[격음][격음]	2.98	3.06	<u>2.69</u>	2.95
[경음][격음]	2.22	2.41	<u>2.30</u>	2.65
[격음][경음]	2.21	2.47	<u>2.38</u>	2.51
[경음][경음]	1.90	2.29	2.18	<u>2.15</u>

이러한 경향은 후두자질 공기가 C_2-C_3 위치한 삼음절어 비단어에 대해 고유어 문법이 높은 비적형성 점수를 부과한다고 예측하는 것과 어느 정도 일치하는 것

[30] 밑줄은 각 후두자질 공기 유형별로 적형성 판단 점수가 가장 낮은 것을 의미한다.

으로 보인다. 특히, [격음][격음] 유형이 포함된 비단어에서 격음의 공기 위치에 따라 보이는 적형성 판단 점수 차이는 고유어 문법에서만 예측된다.

격음/경음 발생 유형별 학습 점수와 응답 점수를 비교해 본 결과, 어휘부를 중심으로 학습된 문법이 '후두자질 발생 및 공기 제한'에 대한 화자의 인식을 어느 정도 예측할 수 있음을 확인하였다. 특히, 경음 발생 저지에 대한 인식은 한자어 문법을 따르는 것으로 보이고 후두자질 공기 위치에 대한 적형성 인식은 고유어 문법을 따르는 것으로 보인다.

5.4. 논의

이 연구는 적형성 판단 실험을 통해, 한국어 화자의 '후두자질 발생 및 공기 제한'의 심리적 실재를 직접 탐색하였다. 실험 결과, 어휘부를 중심으로 학습된 문법이 '후두자질 발생 및 공기 제한'에 대한 화자의 인식을 어느정도 예측할 수 있다는 것을 파악할 수 있었으며, 고유어 문법과 한자어 문법이 독립적으로 적형성 판단에 역할하고 있다는 것을 알 수 있었다. 이러한 실험 결과를 기존연구에서 밝힌 경향과 비교하고, 실험 결과에서 드러난 어휘부 기여도에 대한 해석 가능성을 논의한다.

첫째, 기존연구(Ito 2014, S.Kim 2016, Kang & Oh 2016, 2019, H. Kim 2017)가 동적인 형태·음운론적 교체 현상(합성어 경음화, 어두 경음화)에 대해 논의한 바와 학습 결과를 비교한다. 기존연구에서는 합성어 경음화 발생 비율에 *[경음][경음], *[경음][격음], *[격음][경음]의 제약 효과가 공통적으로 포착된 반면, 어두 경음화 발생 비율에 [경음][경음] 연쇄에 대한 선호가 관찰되었다(3.2.2절 (59) 참조). 본 연구의 적형성 판단 조사 결과, 한국어 화자가 후두자질 공기를 회피하는 경향과 경음을 회피하는 경향을 확인할 수 있었으며, 이는 각각 고유어 문법에서 포착된 후두자질 공기 제약과 한자어 문법에 포함된 [경음] 회피 인접/비인접 제약의 효과로 어느 정도 해석될 수 있었다. 이러한 점을 고려할 때, 본 실험 결과는 후두자질 공기 제약이 합성어 경음화 발생 비율에 영향을 미치는 양상과 어느 정도 일치한다고 볼 수 있다. 다만, 이 연구의 결과는 정적인 적형성 판단 과제에서만 비롯되었기 때문에 동적인 합성어 경음화와의 연관성을 직접적으로 판단하기 어렵다.

둘째, 적형성 판단 조사 응답에서 한자어 문법의 영향이 고유어 문법의 영향보다 더 큰 이유는 두 가지 측면에서 논의할 수 있다. 한 가지 가능한 해석은 본 실험의 비단어 구성에서 한자어 학습 제약의 관여도가 고유어 학습 제약의 관여도보다 컸다는 것이다. 이 연구는 후두자질 공기 제약을 포함하는 고유어 문법을 기준으로, 후두자질(격음, 경음) 관련 제약만을 최소로 위배하는 비단어를 주로 선별하였다. 이에 비해, 한자어 문법은 단어 선별에서 특별히 고려되지 않았다. 이로 인해, 적형성 판단 결과에서 한자어 제약의 영향이 크게 드러난 것으로 볼 수 있다. 또다른 가능한 해석은 한자어 어휘부가 한국어 명사에서 차지하는 비중이 크기 때문에 화자들이 기본적으로 비단어를 한자어에 가깝게 보았을 수 있다는 것이다. 이로 인해, 적형성 판단 조사시 한국어 화자들이 한자어 문법을 활용하였을 가능성이 있다. 이에 대해서는 후두자질 발생 및 공기 제약을 제외한 다른 유형의 제약에 대해서도 한자어 어휘부 영향이 클 수 있을지에 대한 추가적인 조사가 요구된다.

6. 결론

이 연구는 한국어 명사의 음소배열제약을 체계적이고 포괄적으로 조사하고, 제약의 심리적 실재를 탐색하였다. 먼저 2장에서는 기존연구가 비범주적 음소배열제약 인식과 어휘 통계량의 밀접한 관계를 밝히며 어휘부 내 연쇄 분포를 제약으로 정의하는 방식을 살펴보았다. 그리고 다수의 모델 중 '최대 엔트로피 음소배열제약 학습 모델'이 자연 부류에 기반을 두고, 통계적으로 정당화된 제약을 학습할 수 있다는 점을 논의하였다. 3장에서는 과거의 기존연구에서 제시된 한국어 음소배열제약을 개괄하였다. 우선, 전체 어휘부 및 고유어 어휘부에서 발생빈도가 0인 연쇄가 한국어 음운론의 공식적인 제약으로 형식화되었음을 보았다. 다음으로 전체 어휘부 및 고유어 어휘부에서 공시적인 제약으로 형식화되지는 않았지만 보고된 비범주적 음소배열제약을 정리하였다. 그리고 한자어 어휘부에서만 관찰된 음소배열제약을 언급하였다. 이러한 검토를 바탕으로 기존 음소배열제약 탐색 방법이 연쇄 분포를 제약으로 정의하는 계산 방식이 통계적으로 정당화되기 어려운 측면이 있고, 제약을 구성하는 일반화 단위도 연구자마다 임의적이라는 한계가 있음을 지적하였다.

이러한 배경에서 이 연구는 '최대 엔트로피 음소배열제약' 모델을 이용하여 한국어 음소배열제약 문법을 학습하고, 학습 문법의 심리적 실재를 모국어 화자의 적형성 판단 조사를 통해 확인하였다.

4장에서 한국어 음소배열제약 학습 과정과 그 결과를 논의하였다. 기계 학습은 단일 형태소인 명사 어휘를 학습 자료로 하여 고유어와 한자어 어휘 목록을 구분하여 시행하였다. 또한, 과거에 논의된 비인접 음소배열제약까지 포괄적으로 학습하기 위하여, 조음위치자질 층위, 후두자질 층위, 모음자질 층위를 설정하였다. 학습 결과, 기계 학습된 문법은 기존연구에서 관찰된 음소배열제약을 대부분 포함하였으며 새로운 회피 경향도 포착하였다. 학습 제약은 어휘부 문법별 공통점과 차이점을 체계적으로 파악하기 위하여 고유어 및 한자어 문법 공통 제약과 고유어 문법 특정 제약, 그리고 한자어 문법 특정 제약으로 분류하여 제시하였다. 고유어 및 한자어 문법 공통 제약은 기존연구에서 범주적으로 정의한 제약(예: (65a) *[-rhy][+cons]; *$[자음][자음]), 기존연구에서 비범주적으로 정의한 제약(예: (66c) *[−str][+high,−bk]; *[t, t', t^h][i]), 그리고 새롭게 학습된 제약을 포함

하였다(예: (75Ae) *[−son,+lab][+lab]; *[pp^h]). 고유어 문법 특정 제약은 기존연구에서 비범주적으로 정의한 제약 중 음운론적 동기를 갖는다고 논의된 제약들(예: (77a) 후두자질 층위 제약 *[^−asp,−tns][+asp]; ***t'ac^h**i)을 포함하였다. 또한, 고유어 문법에서만 포함된 새로운 학습 제약을 연쇄 유형별로 제시하였다(예: (82a) [음절말음][음절두음] 제약 *[+lab][−tns,+dor]; *[m, p]$[k, k^h]). 한자어 문법 특정 제약은 기존연구에서 전체 한국어 및 고유어에서 포착한 제약(예: (85a) *[^−rnd,+syl][+lab,+rhy]; *[u, o][m, p]$), 기존연구가 한자어에 특정하여 관찰한 제약(예: (88a) *[−cont,+ant][−high,+ATR]; *[t, t',t^h][e, ʌ]) 등의 순서로 살펴보았고, 새롭게 학습된 제약(예: (89f) *[+cont][+asp,+lab]; *[lp^h])을 소개하였다.

5장에서는 학습 제약의 심리적 실재 여부를 파악하기 위하여, 한국어 화자의 적형성 판단 조사를 진행하였다. 이 실험은 '후두자질 발생 및 공기 제한'에 초점을 맞추었으며, 후두자음의 발생 개수, 종류, 그리고 발생 위치를 조건으로 하여 실험 자극을 구성하였다(이음절어 139개(C_1VC_2V, 예: 차파, 꼬빠), 삼음절어 114개($C_1VC_2VC_3V$, 예: 파코두, 또빠기)). 한국어 화자 112명은 실험 자극인 비단어의 적형성을 1–7점의 점수로 응답하였다. 응답 점수를 분석한 결과, 적형성 판단에 미치는 고유어 문법과 한자어 문법의 독립적인 영향을 확인할 수 있었다. 격음/경음 발생 유형별로 응답 점수를 살펴보면, 비단어의 후두자음 개수가 많을수록 적형성 판단 점수가 낮아지고 경음만으로 구성된 실험 자극이 격음만으로 구성된 실험 자극보다 적형성 판단 점수가 낮다는 것을 확인할 수 있었다. 후두자질 공기 위치에 따라 적형성 판단 점수의 차이도 다소 있는 것으로 보인다. 특히, 삼음절어인 비단어($C_1VC_2VC_3V$)에서 [격음][격음], [격음]//[경음] 공기 유형이 다른 위치보다 C_2-C_3에 위치하는 경우 적형성 판단 점수가 낮은 편이다. 이러한 응답 점수를 학습 문법이 비단어에 대해 예측한 비적형성 점수와 비교하여 학습 제약의 효과를 파악하였다. 후두자음 개수가 많아짐에 따라 적형성 판단 점수가 낮아지는 것은 고유어 문법과 한자어 문법 모두에서 분석될 수 있고 경음 발생에 대한 강한 회피는 한자어 문법으로 분석될 수 있는 것으로 보였으며, 고유어 문법은 후두자질 공기 위치에 따른 적형성 판단 점수의 차이에 어느 정도 영향을 미칠 수 있다는 것을 논의하였다.

이와 같이 이 연구는 통계적 학습 방법을 적용하여 기존연구에서 간과된 비범주적 음소배열제약을 일관적인 기준으로 학습할 수 있었으며 한국어의 음소배열제약을 총체적으로 파악할 수 있었다. 또한, '후두자질 발생 및 공기 제약'을 중심으로 적형성 판단 조사를 진행하여 어휘부에서 학습된 문법의 실재를 확인하였으며 고유어 문법과 한자어 문법의 독립적 역할을 확인할 수 있었다.

한국어 화자의 음소배열제약의 실체를 밝히기 위해서 남은 과제는 다음과 같다. 첫 번째 과제는 4장에서 탐색된 다수의 제약의 실재를 파악하는 것이다. 이 모델은 통계적 기제만을 고려한 일종의 기준 모델(baseline model)로서, 실제 화자의 인식을 모두 예측한다기 보다는 이를 찾아가는데 유용한 출발점으로 기능한다. 한국어 화자들이 비범주적 음소배열제약을 인식하는지를 확인하기 위해서는 적형성 판단 조사를 비롯한 행동 실험이 요구된다.

두 번째 과제는 학습 어휘부를 확장하는 것이다. 본 연구에서 다룬 명사 및 단일어 외 다른 품사, 복합어 등을 포함할 수 있다면 한국어에 실재하는 다양한 음소배열제약을 탐색할 수 있을 것이다.

세 번째 과제는 고유어 문법과 한자어 문법의 상호 작용을 밝히는 것이다. 적형성 판단 조사 결과 고유어 문법과 한자어 문법의 독립적인 역할이 드러났지만, 이들 문법 간의 구체적인 관계에 대해서는 계속 탐구가 필요하다. 특히, 고유어 문법과 한자어 문법이 공통적으로 포함한 제약도 다수 존재하는데 공통적인 제약과 각 어휘부의 특정적인 제약이 어떠한 조건에서 어떻게 활용될 수 있는지에 대하여 다양한 측면에서 조사되어야 한다. 이와 관련하여 비단어에 대해 '고유어' 또는 '한자어'라는 어휘 부류 정보를 분명하게 제시할 때, 어휘부별 문법이 비단어의 고유어성 또는 한자어성 판단에 어떠한 영향을 줄 수 있을지 의문이 남는다.

본 연구 결과에 더하여 이와 같은 추후 과제가 수행된다면, 궁극적으로 한국어 화자의 음소배열제약에 대한 최적의 통계적 문법 모델 개발이 가능할 것으로 기대한다.

참고문헌

강범모·김흥규(2009),「한국어 사용 빈도: 1500만어절 세종 형태의미분석말뭉치 기반」, 서울: 한국문화사.

강옥미(2011),「한국어 음운론」, 서울: 태학사.

고광모(1996), "'ㄹ'과 관련된 두 음운 변화",「언어학」18, 31–50.

강용순(1998), "한국어 어휘부 구조",「음성·음운·형태론 연구」4, 55–67.

구민모·백연지·한종혜·남기춘(2012), "한국어 단음절 단어의 시각 재인에서 음절 빈도효과",「언어과학연구」63, 1–20.

구희산·한혜승(1999), "양순음 후행 양순전이음 /w/의 음향음성학적 연구",「음성과학」6(1), 53–62.

권유안(2006), "한국어 시각단어재인에서 나타나는 이웃효과",「말소리」60, 29–45.

권인한(1997), "현대국어 한자어의 음운론적 고찰",「국어학」29, 243–260.

김경아(1996), "위치동화에 대한 재검토",「국어학」27, 131–155.

김경아(2001), "원순모음화와 원순성 동화",「인문논총」8, 51–63.

김경일(1985),「한국어 음절구조에 관한 통계분석」, 서울대학교 언어학과 석사학위논문.

김남미(2004),「15세기 국어의 중모음 연구」, 서강대학교 국어국문학과 박사학위논문.

김무식(2001), "음형대분석을 이용한 이중모음 'ㅢ'의 특징 연구",「어문학」27, 1–15.

김미란·최재웅·홍정하(2014), "한국어 초성-중성 결합의 분포적 특성 및 모음의 군집분석 연구",「음성·음운·형태론연구」20(1), 23–49.

김영선(2007), "j계 하향이중모음 '의'의 단모음화 연구",「동남어문논집」23, 5–27.

남성현·김선회(2018), "한국어 자음-모음 연쇄의 어휘 계층 간 비교",「언어」43(3), 485–506.

마야 아타예바(2016),「고유어와 한자어의 음절구조 비교 연구-2음절어를 중심으로-」, 서울대학교 국어국문학과 석사학위논문.

박선(2015),「한국어 /h/ 탈락의 음운 변이-확률적 최적성 이론 분석-」, 서울대학교 언어학과 석사학위논문.

박선우·홍성훈·변군혁(2013), “한국어의 어휘 계층과 음운론적 복잡성”,「음성·음운·형태론연구」19(2), 255–274.

서윤정(2016),「순행적 유음화의 실현 양상 연구」, 고려대학교 국어국문학과 석사학위논문.

석주연(1996), “중세국어 원순성 동화 현상에 대한 일고찰”,「관악어문연구」21, 217–228.

소신애(2010), “파찰음 앞 /ㄴ/ 삽입 현상에 관하여”,「국어국문학」154, 5–32.

신우봉(2010),「순행적 유음화의 실현 양상 연구」, 고려대학교 국어국문학과 석사학위논문.

신지영(1999), “이중모음 /ㅢ/의 통시적 연구”,「민족문화연구」32, 473–497.

신지영(2009), “한국 한자음의 빈도 관련 정보 및 음절 구조 제약”,「말소리와 음성과학」1(2), 129–140.

신지영(2011),「한국어의 말소리」, 서울: 지식과 교양.

신지영·차재은(2000), “공명자음 뒤에 위치한 /h/”,「21세기 국어학의 과제」, 도서출판 월인, 821–835.

신지영·차재은(2003),「우리말 소리의 체계」, 서울: 한국문화사.

안소진(2009), “한자어 구성 음절의 특징에 대하여”,「형태론」11(1), 43–59.

엄태수(1988), “국어학 : 국어 표면음성제약의 상위원리”,「서강어문」6, 5.

유재원(1997), “한국어 음소 결합 제약에 대한 계량언어학적 연구”,「한글」238, 67–118.

유필재(2001),「서울지역어의 음운론적 연구」, 서울대학교 국어국문학과 박사학위논문.

이주희(2005), “최적성 이론과 음운론적 어휘부 연구”,「돈암어문학」18, 383–413.

이진호(2014),「국어 음운론 강의」, 서울: 삼경문화사.

이호영(1996),「국어 음성학」, 서울: 태학사.

정연찬(1991), “현대국어 이중모음 체계를 다시 생각하여 본다”, 「石靜 李承旭 先生 회갑기념논총」, 379–402.

조남호(2003), 「한국어 학습용 어휘 선정 결과 보고서」, 서울: 국립국어연구원.

진남택(1992), 「한국어 음소의 기능부담량과 음소 연쇄에 관한 계량언어학적 연구」, 서울대학교 언어학과 석사학위논문.

채서영(1999), “음운변화에 나타난 한국어 어휘의 층위구조”, 「음성·음운·형태론 연구」 7, 217–236.

하세경(2000), 「국어 모음충돌 회피현상에 관한 연구: 최적성 이론을 중심으로」, 서울대학교 언어학과 석사학위논문.

한성우(2006), “국어 단어의 음소 분포”, 「어문학」 91, 163–191.

허웅(1985), 「국어 음운학」, 서울: 샘문화사.

홍은영(2012), 「한국어 전설고모음화 현상 연구」, 서울대학교 국어국문학과 석사학위논문.

Albright, Adam (2009), Feature-based generalisation as a source of gradient acceptability, *Phonology* 26(1), 9–41.

Albright, Adam & Bruce Hayes (2003), Rules vs. analogy in English past tenses: A computational/experimental study, *Cognition* 90, 119–161.

Bailey, Todd M. & Ulrike Hahn (2001), Determinants of wordlikeness: Phonotactics or lexical neighborhoods, *Journal of Memory and Language* 44, 568–591.

Bailey, Todd M. & Ulrike Hahn (2005), Phoneme similarity and confusability, *Journal of Memory and Language* 52(3), 339–362.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno & Paul Smolensky (2008), Language universals in human brains, *Proceedings of the National Academy of Sciences* 105, 5321–5325.

Casali, Roderic (1996), *Resolving Hiatus*, PhD Dissertation, UCLA, ROA-215.

Cho, Hyesun (2012), Statistical learning of Korean phonotactics, *Studies in Phonetics, Phonology and Morphology* 18(2), 339–370.

Cho, Taehong (1999), Intra-dialectal variation in Korean consonant cluster simplification: A stochastic approach, *Chicago Linguistics Society* 35, 43–57.

Cho, Taehong & Sahyang Kim (2009), Statistical patterns in consonant cluster simplification in Seoul Korean: Within-dialect interspeaker and intraspeaker variation, *Phonetics and Speech Sciences* 1(1), 33–40.

Chomsky, Noam & Morris Halle (1965), Some controversial questions in phonological theory, *Journal of Linguistics* 1, 97–138.

Chong, Adam (2017), *On the Relation between Phonotactic Learning and Alternation Learning*, PhD Dissertation, UCLA.

Clements, George & Samuel Jay Keyser (1983), CV phonology. a generative theory of the syllable, *Linguistic Inquiry Monographs Cambridge*, *Mass* 9, 1–191.

Coetzee, Andries & Joe Pater (2008), Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic, *Natural Language and Linguistic Theory* 26, 289–337.

Colavin, Rebecca Irene Victoria (2013), *Phonotactic Probability in Amharic: a Psycholinguistic and Computational Investigation*, PhD Dissertation, University of California, San Diego.

Coleman, John & Janet Pirrehumbert (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 Dvais & Ingrid Norrmann (2011), Explaining sonority projection effects, *Phonology* 28, 197–234.

Frisch, Stefan (1996), *Similarity and Frequency in Phonotactics*, PhD Dissertation, Northwestern University, ROA-198.

Frisch, Stefan, Michael Broe & Janet Pierrehumbert (1997), Similarity and phonotactics in Arabic, ROA-223.

Frisch, Stefan, Nathan Large & David Pisoni (2000), Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords, *Journal of Memory and Language* 42, 481–496.

Frisch, Stefan, Janet Pierrehumbert & Michael Broe (2004), Similarity avoidance and the OCP, *Natural Language and Linguistic Theory* 22(1), 179–228.

Gallagher, Gillian (2010), Perceptual distinctiness and long-distance laryngeal restriction, *Phonology* 27, 435–480.

Gallagher, Gillian (2013), Speaker awareness of non-local laryngeal phonotactics in Cochabamba Quechua, *Natural Language and Linguistic Theory* 31, 1067–1099.

Gallagher, Gillian, Maria Gouskova & Gladys Camacho Rios (2019), Phonotactic restrictions and morphology in Aymara, *Glossa* 4(1), 1–39. [DOI:http://doi.org/10.5334/gjgl.826]

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

Goldsmith, John (2011), Information Theory for linguists: A tutorial introduction, Paper presented at the workshop on Information Theory in linguistics at the LSA Summer Institute.

Gordon, Matthew K. (2016), *Phonological Typology*. Oxford University Press.

Greenberg, Joseph & James Jenkins (1964), Studies in the psychological correlates of the sound system of American English, *Word* 20, 157–177.

Hay, Jennifer, Janet Pierrehumbert & 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 (2012), The role of computational modeling in the study of sound structure. Paper presented at the conference on Laboratory Phonology, Stuttgart.

Hayes, Bruce & James White (2013), Phonological Naturalness and Phonotactic Learning, *Linguistic Inquiry* 44(1), 45–75.

Hayes, Bruce & Colin Wilson (2008), A maximum entropy model of phonotactics and phonotactic learning, *Linguistic Inquiry* 39. 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.

Ito, Chiyuki (2007), Morpheme structure and co-occurrence restrictions in Korean monosyllabic stems, *Studies in Phonetics, Phonology and Morphology* 13(3), 373–394.

Ito, Chiyuki (2014), Compound tensification and laryngeal co-occurrence restrictions in Yanbian Korean, *Phonology* 31, 349–398.

Jang, Hayeun (2016), /o/-stems as faster late-starters in decay of Korean vowel harmony, Paper presented at Japanese-Korean Linguistics Conference 2016.

Jun, Jongho (2000), Preliquid nasalization, *Korean Journal of Linguistics* 25(2), 191−208.

Jun, Jongho (2018), Morpho-phonological processes in Korean, in Mark Aronoff (ed.), *Oxford Research Encyclopedia of Linguistics*, New York: Oxford University Press. [DOI: 10.1093/acrefore/9780199384655.013.241]

Jurafsky, Daniel & James Martin (2000*), Speech and Language processing : an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Upper Saddle River, NJ: Prentice Hall.

Jusczyk, Peter, Paul Luce & Jan Charles-Luce (1994), Infants' sensitivity to phonotactic patterns in the native language, *Journal of Memory and Language* 33, 630–645.

Kang, Hijo (2012), *Diachrony in Synchrony: Korean Vowel Harmony in Verbal Conjugation*, PhD Dissertation, Stony Brook University.

Kang, Hijo (2015), Interaction of perception and memory in segmental OCP, *Altai Hakpo* 25, 145–165.

Kang, Hijo & Mira Oh (2016), Dynamic and static aspects of laryngeal co-occurrence restrictions in Korean, *Studies in Phonetics, Phonology and Morphonology* 22(1), 3−34.

Kang, Hijo & Mira Oh (2019), The Asymmetric tense consonant effects in compound and word-initial tensifications in Korean*, Studies in Phonetics, Phonology and Morphonology* 25(1), 3−30.

Kang, Hyeon-Seok (1998), The deletion of w in Seoul Korean and its implications, *Korean Journal of Linguistics* 23(3), 367–397.

Kawahara, Shigeto, Hajime Ono & Kiyoshi Sudo (2006), Consonant co-occurrence restrictions in Yamato Japanese, *Japanese/Korean Linguistics* 14, 27–38, Stanford: CSLI Publications.

Kim, Hyoju (2017), *Phonological Trends in English Loanword Word-initial Tensification in Korean*, MA Thesis, Seoul National University.

Kim, Jong-Kyoo (2000), *Quantity-sensitivity and Feature-sensitivity of Vowels: a Constraint-based Approach to Korean Vowel Phonology*, PhD Dissertation, Indiana University.

Kim, Seoyoung (2016), *Phonological Trends in Seoul Korean Compound Tensification*, MA Thesis, Seoul National University.

Kim, Young-Seok (1984), *Aspects of Korean Morphology*, PhD Dissertation. University of Texas, Austin.

Kim-Renaud, Young-Key (1974), *Korean Consonantal Phonology*, PhD Dissertation, University of Hawaii.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojesen Christensen. (2017), lmerTest Package: Tests in Linear Mixed Effects Models, *Journal of Statistical Software* 82(13), 1–26.

Lee, Jin-Seong (1992), *Phonology and Sound Ssymbolism of Korean Ideophones*, PhD Dissertation, Indiana University.

Lee, Yong-Sung (1993), *Topics in the Vowel Phonology of Korean*, PhD Dissertation, Indiana University.

Lee, Yongeun & Matthew Goldrick (2008), The emergence of sub–syllabic representations, *Journal of Memory and Language* 59, 155–168.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990), Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations, University of Colorado, Boulder.

Legendre, Géraldine, Antonella Sorace & Paul Smolensky (2006), The optimality theory–harmonic grammar connection, in Paul Smolensky & Géraldine Legendre (eds.), *The Harmonic Mind: from Neural Computation to Optimality Theoretic Grammar*, 903–966, Cambridge, MA: MIT Press.

Smolensky, Paul & Géraldine Legendre (2006), *The Harmonic Mind: from Neural Computation to Optimality Theoretic Grammar* (*Linguistic and Philosophical Implications*), Cambridge, MA: MIT Press.

MacEachern, Margaret (1999), *Laryngeal Cooccurrence Restrictions*, New York: Garland.

Mester, Armin (1986), *Studies in Tier Structure*, PhD Dissertation, University of Massachusetts, Amherst.

Mikheev, Andrei (1997), Automatic rule induction for unknown word guessing, *Computational Linguistics* 23, 405–423.

Odden, David (1994), Adjacency parameters in phonology, *Language* 70(2), 289–330.

Ohala, John & Manjari Ohala (1986), Testing hypotheses regarding the psychological reality of morpheme structure constraints, in John Ohala & Jeri Jaeger (eds.), *Experimental Phonology*, 239–252. San Diego, CA: Academic Press.

Padgett, Jaye (1995), *Stricture in Feature Geometry*, Stanford, CA: CSLI Publications.

Pierrehumbert, Janet (1993), Dissimilarity in the Arabic verbal roots, *Proceedings of NELS* 23, 367–381.

Pitt, Mark & James McQueen (1998), Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language* 39, 347–370.

Prickett, Brandon (2015), *Complexity and Naturalness in First Language and Second Language Phonotactic Learning*, MA Thesis, University of North Carolina at Chapel Hill.

R Core Team (2019), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R–project.org.

Rosenthall, Sam (1994), *The Phonology of Vowels and Glides*, PhD Dissertation, University of Massachusetts, Amherst.

Saffran, Jenny , Richard Aslin & Elissa Newport (1996), Statistical learning by 8-month-old infants, *Science* 274 (5294), 1926–1928.

Scholes, Robert (1966), *Phonotactic Grammaticality*, The Hague: Mouton.

Shademan, Shabnam (2007), *Grammar and Analogy in Phonotactic Well–formedness Judgments*, PhD Dissertation. UCLA.

Smolensky, Paul & Alan Prince (1993), Optimality Theory: Constraint interaction in generative grammar, *Optimality Theory in phonology* 3.

Vitevitch, Michael & Paul Luce (1998), When words compete: levels of processing in perception of spoken words, *Psychological Science* 9, 325–329.

Vitevitch, Michael & Paul Luce (1999), Probabilistic phonotactics and neighborhood activation in spoken word recognition, *Journal of Memory and Language* 40, 374–408.

Vitevitch, Michael & Paul Luce (2004), A Web-based interface to calculate phonotactic probability for words and nonwords in English, *Behavior Research Methods, Instruments, and Computers* 36, 481–487.

Vitevitch, Michael & Paul Luce (2005), Increases in phonotactic probability facilitate spoken nonword repetition, *Journal of Memory and Language* 52, 193–204.

Vitevitch, Michael, Paul Luce, Jan Charles-Luce David Kemmerer (1997), Phonotactics and syllable stress: implications for the processing of spoken nonsense words, *Language and Speech* 40, 47–62.

Wilson, Colin & Marieke Obdeyn (2009), Simplifying subsidiary theory: Statistical evidence from Arabic, Muna, Shona, and Wargamay, Ms., Johns Hopkins.

Yu, Alan (2017), Phonotactic constraints in Chinese dialects, in Rint Sybesma (ed.), *Encyclopedia of Chinese Language and Linguistics*, 415–422.

[부록1] 학습 제약: 고유어

번호	제약	가중치	의미
1	*[−rhy]#	6.39	어말 음절두음 금지
2	*[−rhy][+cons]	5.89	어두 자음군 금지
3	*#[+rhy]	5.13	어두 음절말음 금지
4	*[+son,+dor,−rhy]	4.74	음절두음 위치 ŋ 금지
5	*[+asp,+rhy]	4.73	음절말음 위치 격음 금지
6	*[+str,+rhy]	4.59	음절말음 위치 조찰음 금지
7	*[−cons,+rhy]	4.39	음절말음 위치 활음 금지
8	*[+tns,+rhy]	4.25	음절말음 위치 경음 금지
9	*[+cor,+rhy][^+cons,−rhy]	4.04	[설정음]$[모음, 음절말음] 금지
10	*#[+cons,+appr]	3.83	어두 유음 금지
11	*[−syl][+rhy]	3.80	음절말음 위치 자음군 금지
12	모음자질 층위 *[][−high,+ATR][−ATR]	3.62	*[][e, ʌ][ø, ɛ, o, a]
13	*[+lab][+high,+bk,−rnd]	3.23	*[m, p, p’, p^h][ɨ]
			예외: 그믐
14	*[+lab,+rhy][^−appr,−rhy]	3.06	*[m, p]$[비음/저해음인 음절두음 외 분절음]
15	*[−rnd][−bk,−ATR]	2.99	*[j, ɯ, 평순모음][ø, ɛ]
16	*[−ant][+high,+bk,−rnd]	2.93	*[c, c’, c^h][ɨ]
			예외: 짜증
17	*#[−high,−bk]	2.92	어두 [e, ɛ] 금지
			예외: 애꾸, 앵두
18	*[][+cont,+asp]	2.89	어중 h 금지
			예외: 나흘, 사흘
19	*[−nas][−asp,−tns,+cor]	2.87	*[l][t, s, c]
20	*[+lab][−back,+round]	2.86	*[m, p, p’, p^h][y, ø]
			예외: 뫼
21	*[−high][−high]	2.86	비고모음 연쇄 회피
			예외: 가오리, 배알
22	*[−appr,+rh][+appr,−syl]	2.73	*[비음, 저해음]$[유음, 활음]
23	*[−son,+rhy][^−son,−rhy]	2.70	[저해음]$ [저해음 음절두음 외 분절음] 금지
24	*[−high,−bk][−nas,+rhy]	2.68	*[e, ø, ɛ][저해음, 유음]$
			예외: 맵시, 멜빵

번호	제약	가중치	의미
25	*#[−bk,+rnd]	2.57	*#[y, ø]
			예외: 외, 위
26	*[^−rhy][−bk,+rnd]	2.56	*[음절두음 외 분절음][y, ø]
27	*[+lab][+bk,−syl]	2.56	*[m, p, p’, p^h][w]
28	*[+nas,+ant][+tns,+cor]	2.51	*[n][t’, s’, c’]
29	*[+high,+bk,−rnd][−high]	2.48	*[ɨ][비고모음]
30	*[+asp,+lab][−high,+ATR]	2.45	*[p^h][e, ʌ]
			예외: 펄
31	모음자질 층위 *[+ATR][−ATR][−high,+ATR]	2.44	*[i, y, ɨ, u, e, ʌ][ø, o, ɛ, a][e, ʌ]
			예외: 치다꺼리
32	*[+son,−syl][+bk,−syl]	2.41	*[공명 자음][w]
33	*[+nas][+tns,+lab]	2.37	*[n, m, ŋ][p’]
34	*[−bk][+high,−bk,+ATR]	2.36	*[전설모음][i, y]
35	*[−low,+syl][+syl]	2.36	*[고모음, 중모음][모음]
			예외: 거울, 모이
36	*[+son,−syl][−bk,+rnd]	2.35	*[공명음][y, ø]
			예외: 뉘
37	*[−high,−low,−bk][+lab,+rhy]	2.34	*[e, ø][m, p]$
			예외: 셈
38	*[−low,+bk,−rnd]#	2.34	*[ɨ, ʌ]#
			예외: 건너, 겨, 뼈
39	*[−low,+bk,−rnd][−bk,−syl]	2.33	*[ɨ, ʌ][j]
40	*[+high,+bk,−rnd][−son,+lab,−rhy]	2.33	*[ɨ]$[p, p’ p^h]
41	*[−tns,+cor][+son,−syl]	2.33	*[t, t^h, s, c, c^h][비음, 유음, 활음]
			예외: 돼지
42	*[+son,+lab][−cont,+tns,+cor]	2.32	*[m][t’, c’]
43	*[+high,+bk,−rnd]#	2.30	*[ɨ]#
44	*[−son][−asp,−tns]	2.27	*[저해음][평음]
45	*[+asp,+dor][−high,+bk,+ATR]	2.26	*[k^hʌ]
46	*[−bk,+rnd][^−rhy]	2.26	*[y, ø][음절두음 외 분절음]
			예외: 회오리
47	*[−high,−bk][+tns,+lab]	2.23	*[e, ø, ɛ][p’]
48	*[+high,+bk,−rnd][−str,−asp,−tns,−rhy]	2.23	*[ɨ]$[t]
49	*[−low,−bk][+dor,+rhy]	2.22	*[i, y, e][k, ŋ]$
			예외: 싱아, 익살

번호	제약	가중치	의미
50	*[−bk][+asp,+dor]	2.22	*[전설모음][k^h]
51	*[−high,−bk][+high,−rnd]	2.21	*[e, ø, ɛ][i, ɨ, ɯ] 예외: 내음, 대야
52	*[+ATR][+tns,+lab]	2.19	*[i, y, ɨ, u, e, ʌ][p’] 예외: 삐삐, 지빠귀
53	*[−ATR][+back,+round]	2.19	*[ø, o, ɛ, a][u, o] 예외: 가운데, 배웅
54	*[^+cons,+rhy][+tns,+cor]	2.18	*[음절두음, 모음][t’, s’, c’] 예외: 메뚜기, 버찌
55	*[+bk,−rnd][+tns,+lab]	2.14	*[ɨ, ʌ, a][p’] 예외: 아빠
56	*[+rhy][+high,+bk,−rnd]	2.14	*[음절말음][ɨ, ɯ]
57	*[−nas][+nas,+ant]	2.14	*[저해음, 유음][n]
58	*[+tns,+dor][−rnd,−syl]	2.14	*[k’j]
59	*[−son,+cor][−rnd,−syl]	2.13	*[설정 저해음][j]
60	*[+high,+bk,+rnd][−str,+asp]	2.13	*[ut^h]
61	*[+high,+bk,−rnd][+tns,+dor]	2.12	*[ɨk’]
62	*#[+high,+bk]	2.11	*#[w, ɨ, u] 예외: 우리, 으름
63	*[+cont,+tns][−high,+rnd]	2.11	*[s’][ø, o] 예외: 쏘가리
64	*[+lab][−tns,+dor]	2.11	*[양순음][k, k^h] 예외: 임금
65	*[−high,−bk,−rnd][−appr,+ant]	2.09	*[e, ɛ][n, t, t’, t^h, s, s’] 예외: 개나리, 헤살
66	*[−back,+rnd][+nas,+ant]	2.07	*[y, ø][n]
67	*[+high,−bk,+rnd][+lab]	2.05	*[y][양순음] 예외: 휘파람
68	*[−nas,+rhy][^+cons,−rhy]	2.05	*[저해음, 유음]$[모음, 음절말음]
69	*[+str,+tns][−high,−low,−bk]	2.00	*[s’, c’][e, ø] 예외: 족제비
70	*[−low,−rnd][+asp,+dor]	1.98	*[i, ɨ, e, ʌ][k^h] 예외: 서캐
71	*[−son,+lab][+lab]	1.97	*[양순 저해음][양순음]
72	후두자질 층위 *[^−asp,−tns][+asp]	1.95	*[격음, 경음][격음] 예외: 까치, 해파리

번호	제약	가중치	의미
73	*[−ant][−bk,+rnd]	1.95	*[c, c’, c^h][y, ø]
			예외: 쥐, 자취
74	*[−bk,+rnd][−str,−asp,−tns]	1.94	*[y, ø][t]
75	*[+tns,+lab][−high,−low,−bk]	1.94	*[p’][e, ø]
76	*[−high,+ATR][−str,+asp]	1.94	*[e, ʌ][t^h]
			예외: 허탕
77	*[−high,−low,−bk][+appr,−rhy]	1.93	*[e, ø]$[l, w]
			예외: 세로
78	*[+tns][+high,−bk,+rnd]	1.93	*[경음][y]
			예외: 꼭뒤, 여뀌
79	*[+high,+bk,−rnd][+asp]	1.92	*[ɨ][격음]
			예외: 끄트머리
80	*[+cont][+tns,+dor]	1.91	*[lk’]
81	*[−son,+cor][−tns]	1.91	*[설정 저해음][평음, 격음]
82	*[−son][−tns,+dor]	1.90	*[저해음][k, k^h]
83	*[−high,−bk,+rnd]	1.90	*[ø]
			예외: 되, 쇠
84	*[−appr,+cor][−str,+asp]	1.88	*[t, n][t^h]
85	*[+cont][+high,−bk,+rnd]	1.88	*[l, s, s’, h][y]
			예외: 쉬, 쉬리
86	*[−ant][−syl]	1.86	*[c, c’, c^h][w, j, 자음]
87	*[+rnd,−syl][^−high,−rnd]	1.86	*[w][ʌ, a가 아닌 분절음]
88	*[−str,−asp][+high,−bk]	1.86	*[t, t’][j, i, y]
			예외: 띠, 마디
89	*[−son,+cor][+lab]	1.84	*[설정 저해음][양순음]
90	*[−son,+lab][+ant,−tns]	1.81	*[양순 저해음][t^h]
91	*[−appr,+cor][+asp,+dor]	1.80	*[n, t][k^h]
92	*[−bk,+syl][−bk,+syl]	1.79	*[전설모음][전설모음]
93	후두자질 층위 *[+asp][+tns]	1.79	*[격음][경음]
			예외: 토끼, 팔찌
94	*[−cons,−syl][+high,+bk,−rnd]	1.79	*[활음][ɨ]
95	*[+high][+str,+tns]	1.77	*[고모음][s’, c’]
96	*[−high,+bk][−str,+tns]	1.77	*[ʌ, o, a][t’]
97	*[+bk,−syl][−high,−low,−bk]	1.74	*[w, ɯ][e, ø]
			예외: 꿰미

번호	제약	가중치	의미
98	*[+high][+asp,+dor]	1.66	*[고모음][k^h]
			예외: 수크령
99	*[−high,−low,−bk][+high]	1.64	*[e, ø][고모음]
100	*[+high][+high]	1.63	*[고모음, 활음][고모음, 활음]
			예외: 기와, 누이
101	*[−appr,+cor][+asp,+lab]	1.61	*[t, n][p^h]
102	*[+rnd][+str,+tns]	1.56	*[y, ø, u, o][s’, c’]
103	후두자질 층위 *[][][^−asp,−tns]	1.53	*[][][격음, 경음]
			예외: 그저께, 버들치
104	*[+high,+bk][+rnd,+syl]	1.49	*[ɨ, u, w][y, ø, u, o]
			예외: 추위
105	후두자질 층위 *[^−asp,−tns][+asp][]	1.36	*[격음, 경음][격음][]
106	*[−ATR][+rnd,−syl]	1.34	*[ø, o, ɛ, a][w]
107	*[−high,−bk][−appr,+cor,+rhy]	1.31	*[e, ø, ɛ][t, n]$
			예외: 맨드라미
108	*[+lab,+rhy][^−son,−rhy]	1.24	*[양순음]$[비음, 유음, 활음, 모음]
			예외: 엄마, 심마니
109	*[+bk,−syl][^−rnd,+syl]	1.12	*[w, ɯ][평순 모음 외 분절음]
110	*[−high,−bk][−str,+asp]	1.07	*[e, ø, ɛ][t^h]
111	*[+bk,−rnd,−syl]	1.04	*[ɯ]
			예외: 거의, 고의
112	*[−high][−high,−bk]	1.01	*[저모음, 중모음][e, ø, ɛ]
113	*[+lab][+tns,+lab]	0.98	*[m, p][p’]
114	*[+rhy][−bk,+rnd]	0.95	*[음절말음][y, ø]
115	*[−low,+bk,−rnd][−high,−low]	0.87	*[ɨ, ʌ][e, ø, ʌ, o]
116	*[+strid,+tense][−back,+round]	0.47	*[s’, c’][y, ø]
117	*[−high,−bk][+bk,−syl]	0.46	*[e, ø, ɛ][w, ɯ]
118	모음자질 층위 *##	0.26	한 개의 모음은 실재해야함

[부록2] 학습 제약: 한자어

번호	제약	가중치	의미
1	모음자질 층위 *[][][]	8.94	단어 내 세 개의 모음 금지
			예외: 신기루
2	*[−rhy]#	7.55	어말 위치 음절두음 금지
3	*[−rhy][+cons]	7.07	음절두음 자음군 금지
4	*#[+rhy]	6.72	어두 위치 음절말음 금지
5	*[+asp,+rhy]	5.55	격음 음절말음 금지
6	*[−syl][+rhy]	5.53	음절말음 자음군 금지
7	*[+son,+dor,−rhy]	4.96	음절두음 [ŋ]
8	*[^−rnd,+syl][+lab,+rhy]	4.87	[평순모음][m, p] 허용
			예외: 서품
9	*[−bk,−syl][+high,−rnd]	4.80	*[j][i, ɨ]
10	*#[+cons,+appr]	4.36	어두 [l] 금지
11	*[−high,−bk][+lab,+rhy]	4.24	*[e, ø, ɛ][m, p]
12	*[+rnd,−syl][^−high,−rnd]	4.22	*[w][e, ɛ, ʌ 외 분절음]
13	*[−low,−rnd][−bk,−ATR]	4.07	*[j, i, ɨ, e, ʌ][ø, ɛ]
14	*#[+asp,+dor]	3.97	*#[k^h]
			예외: 쾌자
15	*[−high,−bk][+cor,+rhy]	3.96	*[e, ø, ɛ][설정음]$
16	*[^−nas,+rhy][+tns]	3.94	*[비음][경음]
			예외: 산보[sanp'o], 만끽
17	*[−son][−asp,−tns]	3.92	*[저해음][평음]
18	*[−appr,+cor,+rhy][^−appr,−rhy]	3.91	*[n, t]$[유음]
19	*[+high,+bk,−rnd]#	3.87	어말 [ɨ] 금지
20	*[−cons,+rhy]	3.74	활음인 음절말음 금지
21	후두자질 층위 *[+tns][]	3.55	*[경음][저해음]
			예외: 끽연
22	*[−son][+cont,−tns]	3.51	*[저해음][h]
23	*[−nas][−asp,−tns,+cor]	3.43	*[저해음, 유음][t, s, c]
24	*[+lab,+rhy][^−appr,−rhy]	3.42	*[양순음]$[유음]
25	*[−son,+rhy][^−son,−rhy]	3.42	*[저해음][공명음]
26	*[−asp,+cor][−bk,+rnd]	3.42	*[t, t', s, s', c, c'][y, ø]
			예외: 죄
27	*[−cont,+ant][−high,+ATR]	3.41	*[t, t', t^h, c, c', c^h][e, ʌ]

번호	제약	가중치	의미
			예외: 덕, 선언
28	*[+lab][+bk,−syl]	3.39	*[양순음][w]
			예외: 삽화, 입원
29	*[−son,+cor,+rhy]	3.35	*[설정 저해음인 음절말음]
			예외: 곶
30	*[+high,−bk,+rnd][+rhy]	3.34	*[y][음절말음]
31	*[−nas][+bk,−rnd,−syl]	3.33	*[유음, 저해음][ɯ]
32	*[−high,−bk,+ATR][+rhy]	3.32	*[e][음절말음]
33	*[+asp,+lab][−low,+bk,−rnd]	3.30	*[p^h][ɨ, ʌ]
			예외: 입헌
34	*[+lab][+high,+bk,−rnd]	3.24	*[양순음][ɨ]
35	*[+high,+bk,−rnd][−nas,+cor]	3.23	*[ɨ][l, 설정 저해음]
			예외: 슬하, 금슬
36	*[−cons,−syl][−syl]	3.08	*[활음][자음]
37	*[+lab][−bk,+rnd]	3.02	*[양순음][y, ø]
			예외: 범위, 섭외
38	*[+lab][−high,−low,−bk]	3.01	*[양순음][e, ø]
			예외: 섭외, 입회
39	*[−appr][+cons,+appr]	3.00	*[비음, 저해음][l]
40	*[−son,+cor][−rnd,−syl]	2.78	*[설정 저해음][j]
41	*#[−high,−low,−bk]	2.75	*#[e, ø]
			예외: 외설, 외경
42	*[+asp,+dor][+ATR]	2.74	*[k^h][i, y, ɨ, u, e, ʌ]
			예외: 척후, 국헌
43	*[−nas][+nas,+ant]	2.73	*[저해음, 유음][n]
44	*#[−cont,+tns,+cor]	2.65	*#[t’, c’]
45	*[^−high,+bk,+rnd][−son,+cor,+rhy]	2.64	*[ø 외 분절음][t]
46	*[−con,−syl]#	2.63	어말 위치에 활음 금지
47	*[+cons,+son,−rhy][−low,+bk,−rnd]	2.62	*[공명 음절두음][ɨ, ʌ]
			예외: 금어, 신음
48	*[+high,+bk,−rnd][+bk]	2.57	*[ɨ][w, ɨ, u, ʌ, o, a]
49	*[^−cont,+rhy][+tns,+dor]	2.48	*[유음, 모음][k’]
			예외: 태권[태퀀]
50	*[−son,+ant][−syl]	2.44	*[t, t’, t^h, s, s’][활음]
			예외: 쇄도, 인쇄

번호	제약	가중치	의미
51	*[−syl][+asp,+dor]	2.41	*[활음, 자음][k^h]
52	*[−nas,+rhy][^+cons,−rhy]	2.39	*[저해음, 유음]$[활음, 모음]
53	*[+bk,−rnd,+ATR][−rnd,+syl]	2.30	*[ɨ, ʌ][i, ɨ, e, ɛ, ʌ, a]
54	*[+bk,−syl][−high,−bk]	2.27	*[w][e, ø, ɛ]
			예외: 궤도, 발췌
55	*[−str][−low,−bk]	2.24	*[t, t’, t^h][j, i, e, ø]
			예외: 퇴비, 대퇴
56	*[+high,+bk,−rnd][−appr,+cont]	2.21	*[ɨ][s, s’, h]
57	*[+appr,+rhy][^+cons,−rhy]	2.16	*[유음]$[활음, 모음]
58	*[+cont,+asp][−bk,+ATR]	2.15	*[h][i, y, e]
			예외: 희석[히석], 무희[무히]
59	*[+str,+asp][+high,+bk,−rnd]	2.13	*[c^hɨ]
			예외: 예측, 층
60	*[+cons,+son][−high,−low,−bk]	2.12	*[m, n, ŋ, l][e, ø]
			예외: 뇌, 수뢰
61	*[−high][−high,−bk]	2.10	*[비고모음][e, ø, ɛ]
			예외: 사액
62	*[+son,−rhy][+bk,−syl]	2.08	*[m, n, l][w]
			예외: 단원, 설원
63	*[−son,−cont,+cor][−syl]	2.06	*[t, t’, t^h, c, c’, c^h][자음, 활음]
			예외: 계좌, 발췌
64	*[−low,−bk][+son,+dor]	2.04	*[i, y, e, ø][ŋ]
			예외: 결빙, 횡
65	조음위치자질 층위 *[^+cor][+lab][+lab]	2.02	*[양순/설배음][양순음][양순음]
			예외: 감미, 흡입
66	*[−son,+dor][−high,−low,−bk]	1.98	*[k, k’, k^h][e, ø]
			예외: 게시, 괴리
67	*[^+bk,+syl][+asp,+dor]	1.97	*[후설모음이 아닌 분절음][k^h]
			예외: 백합, 식혜
68	*[+syl][−high,−low]	1.94	*[모음][e, ø, ʌ, o]
			예외: 기억, 수온
69	*[+high,+bk,−rnd][−cons,−syl]	1.92	*[ɨ][활음]
70	*[+ant][+high,−bk,+rnd]	1.91	*[n, t, t’, t^h, s, s’][y]
			예외: 단위, 권위
71	*[^−cont,+rhy][+tns,+lab]	1.84	*[유음, 모음][p’]

번호	제약	가중치	의미
72	*[−low,+back,−round][−high]	1.81	*[ɨ, ʌ][비고모음]
			예외: 서옥
73	*[−high,+ATR][+high,+bk,+ATR]	1.80	*[e, ʌ][i, u]
74	*[−high,+rnd][−nas,+cor,+rhy]	1.79	*[ø, o][l, t]$
			예외: 돌기, 골
75	*#[−high,−bk,−rnd]	1.78	*#[e, ɛ]
			예외: 애도, 액
76	*[+appr,−syl][+high,−bk,+rnd]	1.78	*[ly]
77	후두자질 층위 *#[+tns]	1.77	*#[경음]
			예외: 쌍, 약간
78	*[−low,+bk,−rnd][+asp,+lab]	1.73	*[ɨ, ʌ][p^h]
			예외: 저포, 서품
79	*[−high,−bk][+nas,+ant]	1.73	*[e, ø, ɛ][n]
			예외: 개념, 체납
80	*[−high,+bk][+bk,−rnd,+syl]	1.70	*[ø, o, a][ɨ, ʌ, a]
			예외: 파악, 보안
81	*[+cont][+asp,+lab]	1.67	*[lp^h]
			예외: 달필, 살포
82	*[−bk,+rnd][+high,+ATR]	1.67	*[y, ø][i, y, ɨ, u]
			예외: 취입
83	*[−son,+lab][−tns,+lab]	1.65	*[pp^h]
			예외: 납폐, 집필
84	*[^−nas,+rhy][+tns,+cor]	1.64	*[유음, 모음]$[t', s', c']
85	후두자질 층위 *[][][]	1.62	저해음인 음절두음 세 개 금지
86	*[−cons,−syl][−bk,+rnd]	1.61	*[활음][y, ø]
87	*[−str][+high,−bk]	1.53	*[t, t', t^h][j, i]
88	*[+asp,+dor][−bk,−rnd,+syl]	1.50	*[k^h][i, e, ɛ]
			예외: 독해, 식해
89	*[−bk,+rnd][+high,+rnd]	1.49	*[y, ø][w, u]
90	*[−high,−bk][+low]	1.49	*[e, ø, ɛ][ɛ, a]
			예외: 개안, 세안
91	*[+bk,−syl][^−rnd,+syl]	1.48	*[w][평순모음이 아닌 분절음]
92	*[+asp,+cor][−syl]	1.48	*[t^h, c^h][j, w, 자음]
			예외: 발췌, 촬영

번호	제약	가중치	의미
93	*[−str][−low,−bk,−rnd]	1.43	*[t, t’, t^h][j, i, e]
94	*[−low,+bk,+rnd][+bk,+ATR]	1.40	*[w, u, o][ɨ, u, ʌ]
95	*[−appr,+cont,+rhy]	1.39	마찰음인 음절말음 금지
96	*[−bk,+syl][−high,−bk]	1.33	*[전설모음][e, ø, ɛ]
97	*[+cont,+asp][−high,−bk,+ATR]	1.33	*[he]
98	*[−bk,+rnd][+cor,+rhy]	1.28	*[y, ø][설정음]$
99	*[−low][−bk,+rnd]	1.25	*[고모음, 중설모음][y, ø]
			예외: 호위, 비위
100	*[−nas,+cor][+tns,+lab]	1.25	*[t, l][p’]
101	*[+bk,−syl][+high,+bk]	1.22	*[w][ɨ, u]
102	*[+tns,+rhy]	1.04	경음인 음절말음 금지
103	*[^−rhy][−high,−bk,+ATR]	1.01	*[음절두음이 아닌 분절음][e]
104	*[+high,+bk,−rnd][+asp,+lab]	0.96	*[ɨp^h]
105	*[+ant,−asp][−bk,+rnd]	0.94	*[t, t’, s, s’][y, ø]
106	*[−low][+tns]	0.89	*[고모음, 중설모음][경음]
107	*[−str][−syl]	0.69	*[t, t’, t^h][w, j, 자음]
108	*[−high,−low,−bk][+son,+rhy]	0.40	*[e, ø][공명 자음]$
			예외: 전횡, 횡포
109	모음자질 층위 *##	0.02	한 개의 모음은 실재해야함

[부록3] 비단어 목록

A. 이음절어

조건	응답 단어 [총 181개, 본 조사: 169개]
연습 [12개]	즌밤, 산즘, 디날, 감디, 브가, 브나, 나늘, 살투, 잘무, 눈고, 미글, 바신
실제 단어 [12개]	흐코(흑호), 처푸(첩후), 구콰(국화), 이팍(입학), 섭코(섭코), 죽피(죽피), 뽑깨(뽑개), 떡삐(떡비), 삽까(삽가), 직뿌(직부), 다끼(닦이), 가삐(가삐)
격음2 [20개]	차히, 차파, 초키, 허치, 하타, 하코, 허푸, 코차, 카하, 코파, 코타, 파차, 파히, 파코, 파타, 토코, 타푸, 타히, 티커, 카티
	티커, 카티
경음2 [16개]	짜까, 쪼빠, 까찌, 꼬빠, 까싸, 꾸따, 빠찌, 빠까, 빠싸, 뿌따, 싸까, 쑤뽀, 따까, 또빠
	뽀떠, 뻐쩌
격음1, 경음1 [19개]	치꾸, 초빠, 하찌, 호꾸, 하싸, 후뚜 호삐, 카찌, 코빠, 카싸, 쿠뚜, 파찌, 파까, 파싸, 푸뚜, 투까, 토빠
	커뻐, 처빠
경음1, 격음1 [19개]	짜하, 짜코, 짜파, 까차, 까하, 까파, 까타, 빠치, 빠하, 빠키, 빠투, 싸하, 싸파, 싸카, 따하, 또코, 따파, 띠커, 뚜쿠
	띠커, 뚜쿠
격음1 [38개]	처구, 초바, 허고, 허바, 허소, 하두, 허저, 코자, 코수, 코두, 코부, 피주, 푸구, 푸수, 푸두, 투가, 토바, 자히, 주키, 자푸, 가초, 가허, 가푸, 기토, 바차, 바호, 바코, 바타, 사히, 사코, 사푸, 다후, 다코, 다파
	커디, 푸거, 디커, 버처
경음1 [32개]	쩌가, 짜바, 까주, 까부, 까수, 까두, 빠자, 빠구, 빠사, 빠두, 씨기, 싸비, 뚜가, 따부, 주꾸, 조빠, 고빠, 가싸, 구뚜, 기찌, 버쩌, 비꾸, 부또, 바싸, 수꺼, 소빠, 도까, 도삐
	쏘거, 뽀다, 고떠, 기뚜
평음 [13개]	지거, 주바, 기바, 구다, 바주, 버가, 바사, 부다, 수바, 다가, 다바
	디버, 거버

B. 삼음절어

조건	응답 단어 [총 176개, 본 조사: 164개]
연습 [12개]	즌바기, 산즈마, 디나리, 가마디, 브가라, 다브나, 소나빈, 돌마누, 골보주, 바손기, 자노짐, 두러밀
실제 단어 [12개]	흐코(흑호), 처푸(첩후), 구콰(국화), 이팍(입학), 섭코(섭코), 죽피(죽피), 뽑깨(뽑개), 떡삐(떡비), 삽까(삽가), 직뿌(직부), 다끼(닭이), 가삐(가삐)
격음2 [24개]	코파두, 코타부, 파코두, 파타기, 타코부, 타파기, 가파투, 가타피, 보코투, 바타키, 도코푸, 다파키, 코바투, 코다피, 파가투, 파다키, 타가피, 파카투, 파다키, 타가피, 타바키
	크프더, 토크부, 드프커, 보크티, 프디커, 트거포
경음2 [24개]	꼬빠두, 꾸따부, 빠까두, 뿌따기, 따까부, 또빠기, 고뿌뚜, 구또삐, 바꾸뚜, 부따끼, 다꼬삐, 도빠끼, 까부뚜, 까도삐, 빠구뚜, 빠도꾸, 따고삐, 따바끼
	끄쁘더, 끄뽀두, 거뽀떠, 도쁘끼, 쁘고떠, 뿌거띠
격음1, 경음1 [24개]	코빠두, 쿠따부, 파까두, 푸따기, 타까부, 토빠기, 코부뚜, 코도삐, 파구뚜, 파다끼, 타고삐, 타바끼, 가푸뚜, 가토삐, 바타끼, 보쿠뚜, 다코삐, 다파끼
	크쁘더, 토쁘가, 브커떠, 기파따, 커보떠, 포기띠
경음1, 격음1 [24개]	까파두, 까타부, 빠타기, 뽀코두, 따코부, 따파기, 고빠투, 구따피, 바까투, 부따키, 다꼬피, 도빠키, 까바투, 까다피, 빠가투, 빠도키, 따가피, 따바키
	꺼트부, 쁘커더, 드쁘커, 그뻐투, 쁘디커, 뿌구티
격음1 [24개]	코바두, 코다부, 푸거두, 파다기, 투거부, 타바기, 가파두, 가타부, 보코두, 바타기, 다코부, 도파구, 가바투, 가다피, 바가투, 보다키, 다가피, 다바키
	크브더, 코비디, 브커더, 다파거, 브디커, 기부티
경음1 [24개]	까바두, 까다부, 빠가두, 빠다구, 따바구, 따가부, 고빠두, 구따부, 바까두, 부따구, 다까부, 도빠기, 가부뚜, 가도삐, 바구뚜, 보다끼, 다고삐, 다바끼,
	쁘도거, 까두버, 그쁘더, 버까다, 브고떠, 부다꺼
평음 [8개]	가바두, 가다부, 부거두, 보다기, 두거부, 다바기
	그브더, 다그부

[부록4] 지시 문항

A. 철자 형태 자극 제시 (유형 1, 3)

(1) 실제 한국어 단어의 발음형 제시

본 조사 전에, 실제 한국어 단어 20개를 발음형으로 제시합니다. 한 번씩 소리내어 읽어 보세요. 나는 '단어'(을/를) 말해

(2) 점수 응답에 대한 설명

이 연구는 특정 소리 연쇄가 현대 한국어의 단어에서 실제로 쓰일 수 있을지 파악하는 조사입니다. 단어의 발음만을 기준으로, 실제 한국어 단어로 들릴 가능성을 높으면 7점, 들릴 가능성이 없으면 1점에 가깝게 응답해 주세요. 1점: 전혀 한국어 단어 같지 않다. 4점: 한국어 단어일 수 있지만, 좀 이상하다. 7점: 전형적인 한국어 단어 같다.

(3) 연습

본 조사에 앞서, 12개 단어에 대해서 연습을 해 보겠습니다.

다음 "단어"를 한 번 읽어보세요.

"단어"

발음만을 기준으로 할 때, 이 단어는 얼마나 한국어 단어 같습니까?

(1점-전혀 한국어 단어 같지 않음, 7점-전형적인 한국어 단어 같음)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

(4) 본 조사

이제 본 조사를 시작합니다.

다음 "단어"를 한 번 읽어보세요.

"단어"

발음만을 기준으로 할 때, 이 단어는 얼마나 한국어 단어 같습니까?

(1점-전혀 한국어 단어 같지 않음, 7점-전형적인 한국어 단어 같음)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

B. 음성 형태 자극 제시 (유형 2, 4)

(1) 점수 응답에 대한 설명

이 연구는 특정 소리 연쇄가 현대 한국어의 단어에서 실제로 쓰일 수 있을지 파악하는 조사입니다. 단어의 발음만을 기준으로, 실제 한국어 단어로 들릴 가능성을 높으면 7점, 들릴 가능성이 없으면 1점에 가깝게 응답해 주세요.

1점: 전혀 한국어 단어 같지 않다.
4점: 한국어 단어일 수 있지만, 좀 이상하다.
7점: 전형적인 한국어 단어 같다.

(2) 연습

본 조사에 앞서, 12개 단어에 대해서 연습을 해 보겠습니다.
▶ 를 눌러 “단어”의 발음을 들어 보세요.
▶

발음만을 기준으로 할 때, 이 단어는 얼마나 한국어 단어 같습니까?
(1점-전혀 한국어 단어 같지 않음, 7점-전형적인 한국어 단어 같음)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

단어의 발음을 한글로 적어 주세요.
[ ]

(3) 본 조사

이제 본 조사를 시작합니다.

▶ 를 눌러 “단어”의 발음을 들어 보세요.

▶

발음만을 기준으로 할 때, 이 단어는 얼마나 한국어 단어 같습니까?

(1점-전혀 한국어 단어 같지 않음, 7점-전형적인 한국어 단어 같음)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

단어의 발음을 한글로 적어 주세요. [ ]

[부록5] 비단어에 대한 비적형성 점수와 응답 점수

A. 이음절어

조건	비단어	응답 점수		비적형성 점수		조건	비단어	응답 점수		비적형성 점수	
		음성	철자	고유어	한자어			음성	철자	고유어	한자어
격음2	차파	3.60	3.48	1.95	0	경음2	까싸	2.22	1.60	2.18	10.90
	차히	3.20	3.40	4.84	2.15		까찌	2.60	2.31	2.18	10.90
	초키	3.20	2.55	1.95	4.24		꼬빠	2.40	1.86	0	11.99
	카티	2.25	2.29	1.95	9.17		꾸따	1.29	1.79	2.18	11.79
	카하	3.07	2.02	4.84	3.97		따까	3.40	2.12	0	14.39
	코차	4.42	2.62	1.95	3.97		또빠	2.50	1.95	0	14.63
	코타	3.07	2.19	1.95	3.97		빠까	2.75	1.74	0	11.74
	코파	4.83	4.05	1.95	3.97		빠싸	2.00	2.00	2.18	10.90
	타푸	2.46	2.43	1.95	0		빠찌	2.67	2.02	2.18	10.90
	타히	3.09	2.60	4.84	2.15		뻐쩌	1.57	1.57	4.52	11.79
	토코	3.30	2.40	1.95	0		뽀떠	1.92	1.52	6.28	15.20
	티커	2.75	2.02	12.41	9.91		뿌따	2.75	1.74	2.18	11.79
	파차	3.46	3.86	1.95	0		싸까	1.17	1.93	0	11.74
	파코	3.00	2.93	1.95	0		쑤뽀	1.50	1.67	2.19	11.99
	파타	3.67	3.05	1.95	0		짜까	2.00	2.05	0	14.39
	파히	3.50	2.83	4.84	2.15		쪼빠	2.00	1.86	0	14.63
	하코	2.67	2.74	1.95	0						
	하타	3.54	2.81	1.95	0						
	허치	4.62	4.00	1.95	0						
	허푸	3.62	3.17	1.95	1.73						

조건	비단어	응답 점수		비적형성 점수		조건	비단어	응답 점수		비적형성 점수	
		음성	철자	고유어	한자어			음성	철자	고유어	한자어
격-경	처빠	2.40	2.05	6.13	6.67	경-격	까차	3.08	2.98	1.95	5.32
	초빠	2.00	2.24	1.79	6.67		까타	2.93	2.12	1.95	5.32
	치꾸	3.00	2.24	1.79	7.30		까파	2.36	2.19	1.95	5.32
	카싸	2.85	1.57	3.97	9.55		까하	2.60	2.17	4.84	5.32
	카찌	1.70	1.81	3.97	9.55		따파	2.47	2.21	1.95	7.97
	커뻐	1.80	1.57	10.72	13.38		따하	2.13	1.98	4.84	7.97
	코빠	2.64	2.33	1.79	10.64		또코	2.56	1.95	1.95	7.97
	쿠뚜	1.90	1.83	3.97	13.18		뚜쿠	1.67	1.79	3.61	10.71
	토빠	2.71	2.29	1.79	6.67		띠커	2.08	1.64	14.27	17.88
	투까	2.00	2.38	1.79	7.30		빠치	3.00	2.24	1.95	5.32
	파까	2.25	2.26	1.79	6.42		빠키	2.22	1.81	1.95	9.56
	파싸	2.00	2.05	3.97	5.58		빠투	2.33	2.31	1.95	5.32
	파찌	2.60	2.45	3.97	5.58		빠하	2.47	2.05	4.84	5.32
	푸뚜	2.25	1.74	3.97	6.47		싸카	2.17	2.26	1.95	5.32
	하싸	2.75	2.24	3.97	5.58		싸파	2.54	2.12	1.95	5.32
	하찌	2.33	2.48	3.97	5.58		싸하	2.14	2.21	4.84	5.32
	호꾸	2.80	2.64	1.79	7.30		짜코	2.17	2.07	1.95	7.97
	호뻐	2.55	2.74	1.79	6.67		짜파	2.83	2.60	1.95	7.97
	후뚜	2.64	2.31	3.97	6.47		짜하	2.73	2.02	4.84	7.97

조건	비단어	응답 점수		비적형성 점수		조건	비단어	응답 점수		비적형성 점수	
		음성	철자	고유어	한자어			음성	철자	고유어	한자어
격-평	처구	4.00	4.45	0	0	평-격	가초	5.08	4.31	0	0
	초바	4.00	3.98	0	0		가푸	3.57	3.10	0	0
	커디	2.25	2.17	4.12	11.91		가허	4.20	3.71	5.23	0
	코두	5.00	4.05	0	3.97		기토	4.64	3.86	0	0
	코부	2.80	3.62	0	3.97		다키	2.60	2.52	0	4.24
	코수	4.15	3.67	0	3.97		다파	3.50	3.50	0	0
	코자	4.07	3.88	0	3.97		다후	3.93	4.33	2.89	0
	토바	4.69	3.05	0	0		디커	2.55	2.21	12.32	9.91
	투가	4.31	3.79	0	0		바차	4.44	3.60	0	0
	푸거	3.33	2.71	2.34	0		바코	2.63	3.33	0	0
	푸구	2.69	3.71	0	0		바타	4.00	3.19	0	0
	푸두	4.00	3.45	0	0		바호	4.36	4.31	2.89	0
	푸수	4.00	3.36	0	0		버처	4.27	3.14	2.34	0
	피주	5.00	5.26	0	0		사코	2.75	3.10	0	0
	하두	4.55	4.64	0	0		사푸	2.78	2.88	0	0
	허고	5.00	5.02	0	0		사히	3.78	3.50	2.89	2.15
	허바	4.19	3.81	0	0		자푸	3.29	2.86	0	0
	허소	4.92	5.33	0	0		자히	3.60	3.76	2.89	2.15
	허저	4.79	3.90	2.34	0		주키	2.69	2.40	1.66	4.24

조건	비단어	응답 점수		비적형성 점수		조건	비단어	응답 점수		비적형성 점수	
		음성	철자	고유어	한자어			음성	철자	고유어	한자어
경-평	까두	2.93	3.00	0	5.32	평-경	가싸	2.82	2.67	2.18	5.58
	까부	2.91	3.05	0	5.32		고떠	2.50	3.02	6.28	9.88
	까수	3.38	2.43	0	5.32		고빠	2.75	2.69	0	6.67
	까주	2.56	2.98	0	5.32		구뚜	2.50	2.57	2.18	6.47
	따부	2.67	2.24	0	7.97		기뚜	2.60	2.14	2.18	6.47
	뚜가	2.57	2.33	0	7.97		기찌	3.29	2.50	3.95	6.47
	빠구	2.57	2.81	0	5.32		도까	2.00	3.07	0	7.30
	빠두	2.43	2.55	0	5.32		도삐	2.43	3.07	0	6.67
	빠사	2.20	2.55	0	5.32		바싸	2.80	2.36	2.18	5.58
	빠자	3.07	2.50	0	5.32		버쩌	2.89	2.31	4.52	6.47
	뽀다	3.60	3.69	0	5.32		부또	2.82	2.74	2.18	6.47
	싸비	3.07	2.83	0	5.32		비꾸	3.00	2.76	0	7.30
	쏘거	2.38	2.55	4.45	5.32		소빠	2.67	2.52	0	6.67
	씨기	2.83	3.64	0	5.32		수꺼	3.00	2.40	2.34	7.30
	짜바	1.82	2.55	0	7.97		조빠	3.60	2.48	0	6.67
	쩌가	2.45	2.31	0	7.97		주꾸	2.00	3.81	0	7.30

조건	비단어	응답 점수		비적형성 점수	
		음성	철자	고유어	한자어
평-평	거버	3.83	3.74	2.34	0
	구다	4.93	4.88	0	0
	기바	4.31	4.00	0	0
	다가	5.79	5.31	0	0
	다바	5.13	4.17	0	0
	디버	2.86	2.90	4.19	5.20
	바사	4.64	4.26	0	0
	바주	5.42	4.31	0	0
	버가	4.00	3.69	0	0
	부다	5.07	4.62	0	0
	수바	5.08	4.02	0	0
	주바	4.38	3.83	0	0
	지거	4.08	4.40	2.34	0

B. 삼음절어

주요 조건	후두자음 위치	비단어	응답 점수		비적형성 점수	
			음성	철자	고유어	한자어
격음2	C_1C_2	코타부	3.53	2.94	3.31	14.53
		코파두	3.40	3.34	3.31	14.53
		크프더	1.87	1.66	15.36	33.14
		타코부	3.13	2.78	3.31	10.56
		타파기	3.61	3.91	3.31	10.56
		토크부	2.67	3.13	5.64	13.29
		파코두	2.87	2.91	3.31	10.56
		파타기	3.88	3.53	3.31	10.56
	C_2C_3	가타피	3.80	3.47	3.47	10.56
		가파투	3.31	2.88	3.47	10.56
		다파키	2.53	2.75	3.47	14.79
		도코푸	2.38	2.50	3.47	10.56
		드프커	2.25	1.84	21.12	22.53
		바타키	2.75	2.50	3.47	14.79
		보코투	2.60	2.84	3.47	10.56
		보크티	1.88	2.69	5.39	21.73
	C_1C_3	코다피	3.56	3.50	1.53	14.53
		코바투	2.80	2.72	1.53	14.53
		타가피	3.41	3.50	1.53	10.56
		타바키	3.29	2.81	1.53	14.79
		트거포	3.60	2.59	5.14	12.29
		파가투	2.00	2.63	1.53	10.56
		파다키	3.00	3.19	1.53	14.79
		프디커	2.44	2.25	19.31	30.23

주요 조건	후두자음 위치	비단어	응답 점수		비적형성 점수	
			음성	철자	고유어	한자어
경음2	C_1C_2	꼬빠두	3.33	2.13	0	26.10
		꾸따부	2.71	1.88	2.18	25.90
		끄뽀두	2.62	1.53	6.66	26.10
		끄쁘더	2.36	1.41	14.46	35.98
		따까부	3.25	2.41	0	28.49
		또빠기	2.89	3.06	0	28.74
		빠까두	2.67	2.25	0	25.85
		뿌따기	5.00	2.22	2.18	25.90
	C_2C_3	거뽀떠	1.75	1.97	14.58	30.65
		고뿌뚜	2.40	2.03	3.71	27.24
		구또삐	2.33	2.09	3.71	27.24
		다꼬삐	2.11	2.19	1.53	27.19
		도빠끼	2.00	2.13	1.53	27.19
		도쁘끼	1.70	1.75	6.88	31.32
		바꾸뚜	3.08	2.22	3.71	26.99
		부따끼	4.00	2.53	3.71	26.99
	C_1C_3	까도삐	2.73	2.13	1.53	22.54
		까부뚜	2.14	1.84	3.71	22.35
		따고삐	2.58	2.66	1.53	25.19
		따바끼	2.77	2.63	1.53	24.94
		빠구뚜	1.77	1.69	3.71	22.35
		빠도꾸	2.40	2.13	1.53	23.18
		뿌거띠	3.00	1.84	7.33	27.54
		쁘고떠	2.00	1.72	13.47	28.99

주요 조건	후두자음 위치	비단어	응답 점수		비적형성 점수	
			음성	철자	고유어	한자어
격-경	C_1C_2	코빠두	3.33	2.13	1.79	24.75
		쿠따부	1.78	2.00	3.97	27.29
		크쁘더	1.78	1.72	16.25	37.37
		타까부	2.80	2.38	1.79	20.53
		토빠기	3.83	3.03	1.79	20.77
		토쁘가	2.40	1.94	5.02	24.01
		파까두	2.71	2.47	1.79	20.53
		푸따기	4.50	3.28	3.97	20.58
	C_2C_3	가토삐	2.38	2.19	3.32	17.22
		가푸뚜	2.82	2.28	5.50	17.02
		기파따	3.70	2.72	7.26	16.14
		다코삐	2.50	2.44	3.32	17.22
		다파끼	2.38	2.63	3.32	16.97
		바타끼	2.85	2.59	3.32	16.97
		보쿠뚜	1.60	2.09	5.50	19.76
		브커떠	1.70	1.59	20.66	26.41
	C_1C_3	커보떠	2.17	2.03	12.50	27.14
		코도삐	2.42	2.00	1.53	21.19
		코부뚜	3.38	2.47	3.71	20.99
		타고삐	2.42	2.59	1.53	17.22
		타바끼	3.23	2.50	1.53	16.97
		파구뚜	3.62	2.25	3.71	17.02
		파다끼	3.10	2.56	1.53	16.97
		포기띠	2.54	2.75	5.56	22.22

주요 조건	후두자음 위치	비단어	응답 점수		비적형성 점수	
			음성	철자	고유어	한자어
경-격	C_1C_2	까타부	3.50	2.13	3.31	15.88
		까파두	3.28	2.44	3.31	15.88
		꺼트부	2.47	2.25	7.57	15.88
		따코부	2.89	2.44	3.31	18.52
		따파기	2.82	2.78	3.31	18.52
		빠타기	2.00	2.53	3.31	15.88
		뽀코두	2.60	2.03	3.31	15.88
		쁘커더	1.88	1.72	16.71	25.27
	C_2C_3	고빠투	2.27	2.16	3.47	20.77
		구따피	2.00	2.66	5.65	20.58
		그뻐투	2.00	1.72	12.07	20.77
		다꼬피	2.38	2.72	3.47	20.53
		도빠키	2.67	2.19	3.47	25.01
		드쁘커	1.75	1.69	23.53	26.75
		바까투	2.67	2.81	3.47	20.53
		부따키	3.00	2.41	5.65	24.81
	C_1C_3	까다피	3.13	2.94	1.53	15.88
		까바투	3.57	2.13	1.53	15.88
		따가피	4.06	3.13	1.53	18.52
		따바키	3.43	2.63	1.53	22.76
		빠가투	2.94	2.28	1.53	15.88
		빠도키	2.31	2.38	1.53	20.12
		뿌구티	2.20	2.09	3.66	21.08
		쁘디커	2.67	1.94	19.31	32.26

주요 조건	후두자음 위치	비단어	응답 점수		비적형성 점수	
			음성	철자	고유어	한자어
격음1	C_1	코다부	3.67	2.81	0	14.53
		코바두	3.88	3.00	0	14.53
		코비디	2.64	3.13	1.86	19.73
		크브더	2.57	2.03	10.13	27.15
		타바기	4.35	3.59	0	10.56
		투거부	2.79	3.22	0	10.56
		파다기	4.43	4.69	0	10.56
		푸거두	3.33	2.97	0	10.56
	C_2	가타부	3.94	3.72	0	10.56
		가파두	4.40	3.84	0	10.56
		다코부	3.62	3.09	0	10.56
		다파거	3.33	3.25	2.34	10.56
		도파구	4.64	4.84	0	10.56
		바타기	3.94	4.03	0	10.56
		보코두	4.00	2.97	0	10.56
		브커더	2.40	2.16	13.40	19.94
	C_3	가다피	4.28	3.66	1.53	10.56
		가바투	3.31	3.44	1.53	10.56
		기부티	2.63	3.63	3.66	15.75
		다가피	3.71	3.75	1.53	10.56
		다바키	3.57	3.28	1.53	14.79
		바가투	3.20	3.03	1.53	10.56
		보다키	3.23	3.25	1.53	14.79
		브디커	3.09	2.34	19.31	26.94

주요 조건	후두자음 위치	비단어	응답 점수		비적형성 점수	
			음성	철자	고유어	한자어
경음1	C_1	까다부	3.44	2.72	0	15.88
		까두버	2.74	2.59	2.34	15.88
		까바두	3.10	2.88	0	15.88
		따가부	3.33	2.75	0	18.52
		따바구	4.20	3.63	0	18.52
		빠가두	2.80	2.38	0	15.88
		빠다구	4.35	4.34	0	15.88
		쁘도거	2.00	1.56	10.23	22.35
	C_2	고빠두	3.58	2.56	0	20.77
		구따부	2.89	2.34	2.18	20.58
		그쁘더	2.30	2.03	14.46	30.66
		다까부	3.43	2.81	0	20.53
		도빠기	4.43	3.00	0	20.77
		바까두	3.35	2.63	0	20.53
		버까다	3.50	3.69	0	21.41
		부따구	3.60	2.91	2.18	20.58
	C_3	가도삐	3.38	2.59	1.53	17.22
		가부뚜	3.44	3.06	3.71	17.02
		다고삐	3.00	3.09	1.53	17.22
		다바끼	3.38	2.88	1.53	16.97
		바구뚜	3.11	2.75	3.71	17.02
		보다끼	4.00	3.28	1.53	16.97
		부다꺼	2.73	2.69	6.30	16.97
		브고떠	1.91	2.09	13.47	23.67

주요 조건	비단어	응답 점수		비적형성 점수	
		음성	철자	고유어	한자어
평음-평음-평음	가다부	4.38	3.91	0	10.56
	가바두	4.45	4.00	0	10.56
	그브더	3.15	2.91	10.13	20.44
	다그부	3.47	3.47	2.33	10.56
	다바기	4.53	4.72	0	10.56
	두거부	4.20	3.84	0	10.56
	보다기	4.14	4.72	0	10.56
	부거두	4.90	3.97	0	10.56

Abstract

Stochastic learning and well-formedness judgement in Korean Phonotactics

Park, Nayoung
Department of Linguistics
The Graduate School
Seoul National University

The present study not only provides a machine-learning-based investigation of Korean phonotactic grammar, but also tests its psychological reality through judgment experiments on native speakers of Korean. Phonotactics refers to language-specific restrictions on segments and segment sequences. Speakers' phonotactic well-formedness judegments have been mostly described as categorical: e.g., ***bl***ick [well-formed] vs. ***lb***ick [ill-formed]. However, a growing number of studies on phonotactics argue that phonotactic grammaticality is in fact gradient (Coleman & Pierrehumbert 1997, Hayes & Wilson 2008, Albright 2009 and others).

Given these backgrounds, this study provides a systematic investigation of Korean phonotactics, focusing on aspects of gradience. This study is composed of two main parts, learning Korean phonotactic grammars based on Korean lexicons, and testing the learned grammars by conducting judgment experiments on Korean speakers.

In the first part, using a Maximum Entropy Phonotactic Model (Hayes & Wilson 2008), we ran a learning simulation. Native Korean and Sino-Korean lexicons were separately adopted as the training data. Based on the statistical patterns of each lexicon, phonotactic constraints were created with their own weights, the magnitude of which reflects their gradient strength. The resulting native and Sino-Korean grammars confirmed most, if not all, of categorical and gradient phonotactic patterns reported in the previous studies on Korean phonotactics. Furthermore, some previously unreported patterns were found. Thus, this study explores the overall Korean phonotactic constraints that are justified with statistical support, improving on

the previous studies that focused on the specific phonotactic patterns and did not provide clear statistical justification.

The latter part of this study concerns the psychological validity of the phonotactics that were learned. Specifically, the well-formedness judgment test about laryngeal co-occurrence restrictions was conducted on native speakers of Korean. The test results suggest that Korean speakers are aware of most of the laryngeal co-occurrence restrictions which are parts in the learned grammars. It is also shown that native and Sino-Korean grammars make independent contributions to explaining speakers' judgments found in the present experiment.

In sum, this research shows that Korean phonotactic grammars can be learned from the lexicons of Korean, and at least some important parts of the learned grammars are psychologically real. The learned grammars can function as a baseline model for Korean speakers' knowledge of phonotactics.

Keywords : phonotactics, gradient well-formedness, laryngeal co-occurrence restriction, well-formedness judgement test, Maximum Entropy Phonotactic learning model

Student Number : 2013-30853