통계적 학습 방법을 이용한 한국어 음소배열제약 탐색

서울대학교 언어학과 박사과정

박나영 (arimnet@naver.com)

(사)한국언어학회 창립 60주년 기념 여름학술대회

2016. 6. 17(금)

본 연구의 목적

- 음소배열제약 음소 단위의 결합 또는 회피에 대한 문법적 인식
  - ✓새로운 단어에 대해서도 적형성(well-formedness)을 판단 가능

    예: 영어 blif

- 한국어 음소배열제약에 대한 탐색 ⇨ 비범주적 인식
  - ✓대상: 한국어 명사 37,160 단어(차용어 제외)
  - ✓방법: 최대 엔트로피 음소배열제약 모델 도입

- 전개

1. 기존연구: 한국어의 음소 결합 관계
2. 본 연구의 방법 및 대상 소개
3. 최대 엔트로피 음소배열제약 모델 및 학습 시뮬레이션
4. 결과: 제약 및 제약의 가중치
5. 논의: 통계적 음소배열제약 모델 vs. 실제 한국어 화자의 인식

기존연구: 한국어의 음소 결합 관계

- 출현하지 않는 연쇄 + 음운론적 분석
  - 예1: 집+만 [cimman] *[pn] ⇨ 음운과정의 동기
  - 예2: 밭+도 [patt'o] 종성 *[t^h] ⇨ 음절구조 제약
  - 예3: *[jɨ, wu, wo] ⇨ 음소 결합 제약

- 허웅(1985): 두 음소의 결합 **여부** 보고
  - 한국어 전체를 대상으로 저자의 직관에 기반을 둔 관찰
  - 일부 연쇄에 대한, 음운론적/통시적 해석
  - [모음]-[모음] 결합 여부는 다루지 않음

- 다음과 같은 관찰은 '**문법**'에 포함되기 어렵다.
  - 우연한 빈칸: *[mju] cf. 뮤직(music), 퓨즈(fuse)
    - 음성학적 동기 부재, 외래어 영향으로 **잘 쓰이게 될** 연쇄 (허웅 1985: 231)
  - 낮은 빈도로 출현하는 연쇄에 대한 산발적 보고
    - 예: [l]은 w-계 이중모음, [je, jɛ]과 결합이 **잘 나타나지** 않는다.

음소 연쇄에 대한 계량적 접근(1): 연구 대상

- 대량의 어휘 수 대상, 하위 어휘 부류를 고려하지 않음
  - 자료의 특징
    사전: 용언의 표제어 '-다'를 포함　　발화: 문법형태소의 '반복적인' 출현이 두드러짐

cf. 고유어의 결합 관계가 한국어 화자들의 문법 지식 반영(김경일 1985, 한성우 2006, Ito 2007)

- [자음]-[자음], [모음]-[모음]에 대한 연구 부족

	사전 (표제어)	텍스트, 발화자료
[자음]-[자음]	유재원(1997), Cho (2012)	신지영(2008)
[자음]-[활음]	유재원(1997), 이상억(2001), Lee (2011), Cho (2012), 김미란 외(2014)	진남택(1992), 박동근,이석재 (2005), 신지영(2008)
[자음]-[모음]	유재원(1997), 이상억(2001), Lee & Goldrick(2008), 이용은(2009), Lee (2011), Cho (2012),김미란 외(2014)	
[모음]-[자음]	1) [자음]=종성: Lee & Goldrick(2008), 이용은(2009) 2) 음절 종성 명세 X: 유재원(1997),이상억(2001), Cho (2012)	박동근,이석재 (2005), 신지영(2008)
[활음]-[모음],	유재원(1997), Lee (2011), Cho (2012)	신지영(2008)
[모음]-[모음], [모음]-[활음]	유재원(1997), Cho (2012)	

음소 연쇄에 대한 계량적 접근(2): 연구 방법

- 빈도: 두 음소의 결합 정도를 파악하기 부족하다.
  - 유재원(1997): 선행 음소에 대한 자질과 후행 음소에 대한 자질 상관표
  - 예: 모음([평순, 원순]) + 자음([조음 위치]) → [원순모음]+[양순음] 회피
  - 두 자질 부류의 결합 내에서 상대적인 우세/회피 정도를 확인

    ➔ 통계적으로 유의미성을 판단하기 어렵다.

	계	잇소리	입천장소리	입술소리	목청소리
평순모음	74.95	75.08	71.2	84.6	67.2
원순모음	25.05	24.92	28.8	15.4	32.8
계(%)	100	100	100	100	100

- 통계적 기법
  - 상관계수 (Lee & Goldrick 2008, 이용은 2009, Lee 2011),
    음운론적 복잡도 (박선우 외 2013), 군집분석 (김미란 외 2014)

  - 일부 결합 관계에 대해서만 적용
  - 계산된 결합 정도가 화자의 문법 인식에 투영될 수 있는 문법 기제 부재
  - 화자의 문법 인식에 대한 증명 부족

    cf. Lee & Goldrick 2008, 이용은 2009, Lee 2011 → 단음절어, 2음절어 대상

Cho (2012) – 최대 음소배열제약 모델(H&W 2008) 채택

- 모델의 특징
  - 어휘 목록 내에서 통계적으로 유의미한 제약을 자동적으로 학습
  - ‘우연한 빈칸’이 제약으로 포착되고, ‘결합 정도’가 제약에 반영됨
  - 각 연쇄에 대한 화자의 인식을 예측할 수 있다.

- 기존에 포착되지 않은 제약 및 정도성 학습 ➔ 일부 제약 제시
  - 예: *[+asp, +lab][-high,-low,-round] (*$p^h\Lambda$) 가중치: 2.99

- 가능한 연쇄와 불가능한 연쇄 사이의 차이를 포착하지 못하는 학습 문제 발생
  - 자질 목록 조정을 통한, 해소에 중점을 둔다.

(a) tt, pt, pp, … (b) tt^h,pt^h,pp^h… 격음: **[-tense]? [+tense]?**

(c) wɨ,wu,wo,jɨ,ji (d) uo, oo

(e) wɨ (f) pɨ,p^hɨ,p’ɨ,mɨ

본 연구의 방법

- 최대 엔트로피 음소배열제약 도입
  - 한국어 음소배열제약에 대한 귀납적 탐색
  - 한국어 화자의 인식을 점검할 수 있는 기준 모델 제시

- Cho (2012)와 다른 점
  - 자료: 명사 어휘 목록에 한정 + 37,160 단어
    - 용언 표제어 '-다' 형태 제외
    - 대량의 단어 ➔ Cho (2012)에서 다룬 학습 문제가 발생하지 않음
  - 각 제약이 예측하는 비범주적 문법성을 세부적으로 살펴보고자 함

어휘 목록

- 명사: 37,160 단어
  - 빈도 5 이상인 일반명사(강범모·김흥규 2009); 단일어와 복합어 포함
  - 표준국어대사전(http://www.korean.go.kr)에 등재된 단어

	단일어	복합어	단일어+복합어
고유어	1,872	4,029	5,901
한자어	16,205	12,897	29,102
기타	18	2,139	2,157
합계	**18,095**	**19,065**	**37,160**

- 입력형
  - 표준국어대사전 발음형 - 한 가지 발음형만을 선택
  - 어말 자음은 모음으로 시작하는 조사와 결합할 때, 실현될 수 있는 바 , 음운론적으로 가정되는 기저형을 입력형으로 채택 예: 꽃 [k’oc^{h}]

- 연구대상을 '명사'로 한정한 이유
  - 한국어 품사의 대다수를 차지
  - 용언 표제어 '-다' 제외

    - 용언의 어떠한 형태를 배울 것인지를 설정하기 어려움

      예: '먹-' vs. '먹-다' vs. '먹-어'
  - 한국어 화자들이 새로운 단어를 대부분 '명사'로 인식

    - 외국어를 차용할 때, '명사' 또는 '명사+하다'의 형태로 인지

      예: "고(Go) 해야지"
- 하위 어휘 부류(예: 고유어, 한자어)를 구분하지 않음
  - 한국어 대상 대다수의 계량 연구를 따름

최대 엔트로피 음소배열제약 모델(Hayes & Wilson 2008)

- **가정: 음소 연쇄의 출현 비율 = 음소 연쇄에 대한 적형성**

- 입력: **단어** 목록 + **자질** 목록 → 출력: (유표성) **제약** + **가중치**
  - 음소배열제약에 대한 귀납적 모델
  - 음소의 결합 관계를 종합적, 체계적으로 살펴봄
  - 양적 정보를 문법(제약+가중치)으로 투영할 수 있는 기제 제시

- 제약의 형태(***[자질]-[자질], 가중치**)
  - 유표성 제약만이 학습된다.
  - 가중치: 각 연쇄의 결합 정도를 나타냄, 값이 클수록 회피되는 제약
  - 각 연쇄의 비적형성 점수(harmony)
    - 각 연쇄는 제약을 위배할 때,
      위배되는 제약의 가중치 합을 비적형성 점수로 받는다.

각 제약이 특정 연쇄의 확률을 예측하는 방식

- 제약 및 제약의 가중치
  - ① *[+단어 경계][+자음성][+자음성] (*#CC) 가중치: 3
  - ② *[+성절성][+성절성] (*VV) 가중치: 2

- 각 연쇄에 대한 비적형성 점수 및 확률 계산

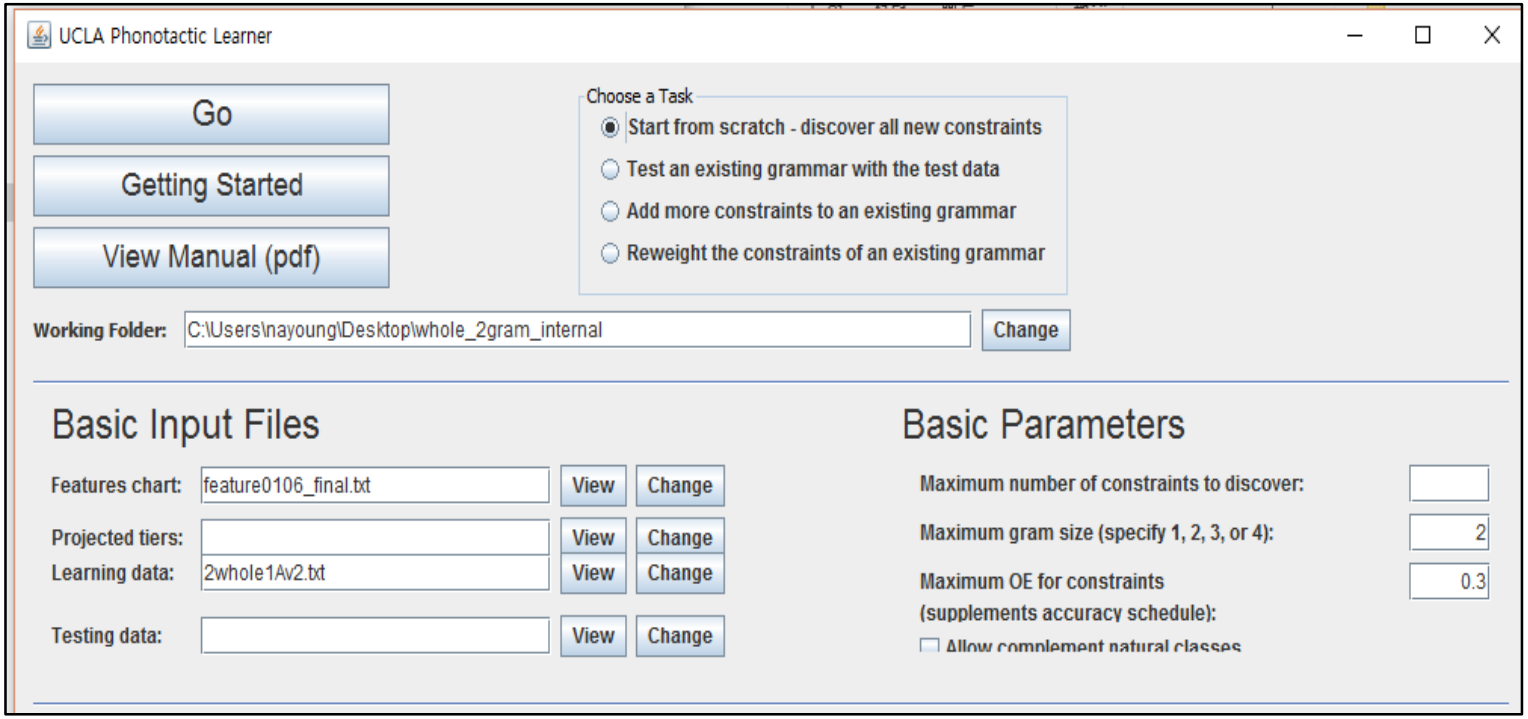
$$\text{최대 엔트로피 값} = e^{\text{비적형성 점수}}$$

$$\text{해당 연쇄의 출현 확률} = = \frac{\text{해당 연쇄의 최대 엔트로피}}{\text{최대 엔트로피 값의 합}}$$

	*#CC	*VV	비적형성 점수 (Harmony)	최대 엔트로피 값 (e^H)	출현확률 (P)
	3	2			
CV	0	0	(0x0)+(0x0)=0	exp(-0) = 1	0.84
CCV	1	0	(3x1)+(2x0)=3	exp(-3)≒0.05	0.11
CCVV	1	1	(3x1)+(2x1)=5	exp(-5)≒0.006	0.04

학습 시뮬레이션

- 소프트웨어: UCLA phonotactic learner
- 학습 조건
  - 입력형: 어휘 목록 + 자질 목록 ➔ 출력형: 제약 + 가중치
  - 어휘 목록: 37,160 X 2 (단어), 제약의 길이: 2, 정확도: 0.3
  - 학습할 제약의 수 - 제한을 두지 않는다.



자질 목록(10모음 체계)

- 단일 자질 명세 가능(예: [Dorsal], [Labial], [Coronal]) ➔ +만 명세하고, 나머지 자질을 0으로 명세

	anterior	aspirate	tense	labial	coronal	dorsal	glottis
p	0	-	-	+	0	0	0
ph	0	+	-	+	0	0	0
pp	0	-	+	+	0	0	0
t	+	-	-	0	+	0	0
th	+	+	-	0	+	0	0
tt	+	-	+	0	+	0	0
c	-	-	-	0	+	0	0
ch	-	+	-	0	+	0	0
cc	-	-	+	0	+	0	0
k	0	-	-	0	0	+	0
kh	0	+	-	0	0	+	0
kk	0	-	+	0	0	+	0
s	+	-	-	0	+	0	0
ss	+	-	+	0	+	0	0
h	0	+	-	0	0	0	+
m	0	0	0	+	0	0	0
n	+	0	0	0	+	0	0
ng	0	0	0	0	0	+	0
l	+	0	0	0	+	0	0

어휘 목록

k a
k a k e
k a kk a i
k a n a n
k a n x m
k a t a k
k a l a c i
k a l a k
k a l a k
k a l a k cc i
k a l a ng i

결과

- 182개 제약 학습(187개 제약 중 가중치 0인 제약 제외)
- 제약의 구성
  - 1개 학습된 제약
    - [자음]-[자음, 활음]: *[-sonorant,+anterior][+sonorant,-syllabic](*cw, *cj): 2.05
    - [자음]: *[+aspirate,+dorsal](*k^h): 2.321
    - [모음]: *[+high,-back,+round](*y): 2.175
  - 가중치 평균 높음: [활음]-[모음] 결합, 단어 경계 제약
  - 같은 유형이더라도 개별 제약 간의 가중치 차이가 크다. ⇨ 개별 제약에 대한 탐색 필요

제약 유형	제약의 수	가중치평균
[활음]-[모음]	5	4.37
#[자음]	5	3.79
[자음]#	9	3.07
[자음]-[자음]	35	3
[모음]#	3	3.02
#[모음]	4	2.65
[자음]-[활음]	11	2.16
[자음]-[모음]	35	1.9
[모음]-[자음]	34	1.7
[자음]-[모음, 활음]	6	1.61
[모음]-[모음, 활음]	4	1.56
[모음]-[모음]	11	1.51
[모음]-[활음]	6	1.45
[모음, 활음]-[모음]	11	1.07

[자음]-[자음] (1): [저해음]-[격음]의 비적형성 점수

자음 1/ 자음2	ㅍ	ㅌ	ㅊ	ㅋ
ㅂ	2.95	3.56	1.79	6.81
ㄷ	3.32	3.48	3.48	6.39
ㄱ	0	1.64	0	4.54

- [k^h]의 출현 및 [t][자음]에 대한 회피
- [kp^h, kc^h] 허용
- [pt^h]: 두 제약 위배 예외: 밥투정, 업태 (12개)

제약	의미	가중치
*[-sonorant,+labial][-tense,+coronal]	[**p**][**t^h**, c^h]	1.79
*[+labial][+anterior,+aspirate]	[m, **p**][**t^h**]	1.77

❖ Cho (2012)와 일치

- [저해음]-[격음] 제약 학습: 형태소 경계에만 출현하는 연쇄로 보고
- 비적형성 차이 보고 **[pt^h]:** 6.96 vs. [kc^h]: 1.73

[자음]-[자음] (2): [비음]-[경음]의 제한

- 비적형성 점수 - [m][경음], [nt’] 제한

자음 1/ 자음2	ㅃ	ㄸ	ㅆ	ㅉ	ㄲ
ㅁ	1.57	2.00	2.18	1.91	1.76
ㄴ	0	1.81	0	0	0
ㅇ	0	0	0	0	0

- 학습 제약

제약	의미	가중치	예외
*[+sonorant,+labial][+continuant,+tense]	**[m][s’]**	2.18	솜씨, 숨소리
*[+sonorant,+labial][-anterior,+tense]	**[m][c’]**	1.91	짐짝, 염증
*[+labial][-continuant,+anterior,+tense]	**[m, p][t’]**	2.00	범띠, 디딤돌
*[+sonorant,+labial][+tense,+dorsal]	**[m][k’]**	1.76	심부름꾼
*[+nasal,+anterior][-continuant,+anterior,+tense]	**[n][t’]**	1.81	안뜰, 판돈

- 기존연구 기술과 부분적으로 부합
  - [비음]+[경음] 낮은 결합빈도 – 한국어 전체(유재원 1997: 표 2-8)
  - [n, m]+[경음] 출현 금지 – 고유 2음절어(김경일 1985)

- 검증이 필요한 부분: [ŋ]+[경음]은 허용되는가?

- cf. [l]-[경음], [l]-[t를 제외한 평음]은 제약을 위배하지 않는다.
  - [lt]만을 위배하는 제약이 학습됨
  - 기존 연구에 부분적으로만 부합
    - 한국어 형태소 내부, [l]-[t, s, c]이 허용되지 않는다. (고광모 1996)
    - 한국어: [l]에 후행하는 경음: [설정음] > [비설정음] (유재원 1997)
    - 한자어: [l]에 후행하는 [t, s, c]가 경음화 됨 (권인한 1997, 신지영, 차재은 2003)

[자음]-[활음]: [저해음]-[활음]의 비적형성 점수

	w	j
ㅂ	2.445	0
ㅃ	3.946	0
ㅍ	1.994	0
ㄷ	2.048	8.585
ㄸ	2.048	8.585
ㅌ	4.424	7.888
ㅅ	2.048	6.127
ㅆ	2.048	6.127
ㅈ	0	6.71
ㅉ	1.895	8.605
ㅊ	2.376	6.013
ㄱ	0	0
ㄲ	0	0
ㅋ	2.321	2.321

- 자음의 조음위치별
  - [양순음]+[w] 회피
  - [설정음]+[w, j] 회피
    cf. [cw] 허용
  - ❖ 기존연구와 일치

+ [k^h][w, j] 회피

- cf. [공명음]+[w, j] 허용
  - [mw]만이 회피
  - 기존연구와 불일치
    - 허웅(1985): *[lw]
    - 유재원(1997): *[비음,유음][w]

- 제약 * [격음, 경음] + [활음]
  - [경음]과 [격음]을 회피하는 경향, 그러나 일부 연쇄가 허용됨
  - [경음]: [k’] [w, j], [p’j] 허용, [격음]: [p^hj] 허용
    - 기존연구(유재원 1997): 허용되는 자음은 다른 자음에 비해 높은 비율로 보고
    - 경음의 상대빈도: /s’/(0.10%)와 /t’/(0.40%), /c’/(0.52%), /p’/(1.29%)
      cf. /k’/(4.79%)
    - 격음의 상대빈도: /t^h/(0.47%), /k^h/(1.62%), /c^h/(1.37%)
      cf. /p^h/(5.11%)

- 기존연구: 경음과 격음을 겹자음으로 분석
  - [경음, 격음] + [활음]의 결합은 세 음소의 결합으로 회피 (유재원 1989)
  - 고유어 단음절 어간에 복합 분절음 두 가지 이상 포함되지 않는 경향 (Ito 2007)

- 겹자음 표상을 가정하는 분석(유재원 1989, Ito 2007)은
  각 분절음에 대한 세부적인 비율 차이까지는 ‘문법적’으로 포착하지 않는다.

- [자음]+[모음]
  - 대부분의 [자음]+[모음]의 결합 가능 → 결합 정도성의 차이
  - *[자음]+[ɨ, e, y, ø] 다수 학습 -'[자음]+[e]' 에 대한 회피가 두드러짐

자음+에	비적형성점수	한자어 내 출현여부	자음+에	비적형성점수	한자어 내 출현여부
페	8.41	X	메	3.24	X
케	7.04	X	네	2.94	X
뻬	6.56	X	게	2.44	
헤	4.99	X	쩨	0.97	
ㅇ+에	4.91	X	쎄	0.97	
데	4.18	X	세	0	
테	4.18	X	제	0	
떼	3.43	X	체	0	
께	3.41		레	0	
베	3.24	X			

  - [p^he, k^he, p'e] – 명사에서 거의 출현하지 않음
  - 대부분, 한자어에서 출현하지 않는 연쇄(신지영, 차재은 2003, 신지영 2009)
    비교 1> [se, ce, c^he] 허용 → 한자어 출현 예> 세금, 제국,
    비교 2> [le]: 한자어 출현X, 명사 허용 예> 병치레, 벌레, 모레…

- 제약 '*[자음]+[에]'의 실재에 대한 검증이 필요하다.
  - '투게더, 데이트 ……' 다수의 차용어에서 [자음]+[에] 출현
  - 음성학적 동기가 뚜렷하지 않다.

- cf. 음성학적 동기 有: *[t, t', t^h]+[i]　　비적형성 점수: 2.46
  - 형태소 경계에서 발생하는 구개음화 영향
  - 계량적 연구(진남택, 1992, 유재원 1997)에서도 드러남

- [자음]+[모음] 간 비적형성 점수 비교
  - 마디([ti] 2.46) ≒ 휴게([hjuke] 2.44)
  - 페([p^he] 8.41) 〉 디([ti] 2.46)

    ✓ 차용어　페달[p^hetal] vs. 피디[p^hiti]

기타 [자음]+[모음] 연쇄

연쇄	비적형성 점수	예외
ŋ + [애]	2.46	장애[caŋɛ]
째	1.76	탑재[tapc'ɛ]
패	1.70	실패[silpʰɛ]
빼	1.69	올빼미[olp'ɛmi]
깨	1.60	깨달음[k'ɛtalɨm]
연쇄	**비적형성 점수**	**예외**
퍼	4.62	입헌[ipʰʌn]
더	1.72	덕[tʌk]
터	1.72	터[tʰʌ]
너	1.57	은어[ɨnʌ]
연쇄	**비적형성 점수**	**예외**
ŋ + [우, 오]	1.94	방울[paŋul]
후	1.64	노후[nohu]

- [모음]+[자음]
  - 음절 종성에 허가되지 않거나 출현 빈도가 낮은 경음, 격음, [t] 출현이 제한
  - [경음]은 모든 모음에 후행하여 회피 예외: 가짜, 토끼, 서쪽, 이끼
  - [ɨ, ʌ]+[설정 저해음]을 제한하는 제약

제약	의미	가중치	예외
*[-low,+back,-round][+aspirate,+coronal]	*[ɨ, ʌ][t^h, c^h]	2.46	허탈, 서체
*[+high,+back,-round][-aspirate,+coronal]	*[ɨ, ʌ][t, s, c]	2.38	어르신, 그즈음

  - [모음]+[공명음]

제약	의미	가중치	예외
*[-high,-back,-round][+sonorant,+anterior]	*[e, ɛ][n, l]	2.2	밸, 배래, 체내
*[-high,-low,-back][+sonorant,+anterior]	*[e, ø][n, l]]	1.14	세로, 옛날
*[-low,-back][+sonorant,+dorsal]	*[i, e, y, ø][ŋ]	2.03	핑계, 결빙
*[-high,-back][+sonorant,+labial]	*[e, ɛ, ø][m]	1.53	보탬, 제물

  - 음절 경계 명세되지 않음 → **밸 vs. 배래**

- [모음]+[모음]
  - [+syllabic][+syllabic](가중치: 3.19) → 모음 연쇄에 대한 전반적인 회피
  - 비적형성 점수 – 추가적으로 제약을 위배하는 모음 연쇄에 음영 표시
  - 특정 음소의 결합 회피

    [으]+[모음], [모음]+[에, 애] ([위, 외]+[모음], [모음]+[위, 외])

    + [어]-[모음]: [어어], [어아], [어오]

    + 기타 [모음]: [에우], [우우], [우오]

V1 \ V2	이	에	애	으	어	아	우	오	위	외
이	3.19	6.05	5.70	3.19	3.19	3.19	3.19	3.19	5.36	5.24
에	3.19	8.65	4.15	3.19	3.19	3.19	4.61	3.19	6.78	4.01
애	3.19	9.27	4.67	3.19	3.19	3.19	3.19	3.19	5.36	3.19
으	5.67	16.17	14.60	5.20	9.62	9.62	7.32	13.45	9.97	17.60
어	3.19	12.00	6.34	3.19	5.59	5.59	3.19	5.48	5.36	7.39
아	3.19	9.27	4.67	3.19	3.19	3.19	3.19	3.19	5.36	3.19
우	3.19	6.24	6.24	3.19	3.19	3.19	3.78	5.59	5.95	8.18
오	3.19	6.60	4.01	3.19	3.19	3.19	3.19	3.19	5.36	4.01
위	5.36	6.18	6.18	6.54	5.56	5.36	5.82	5.82	8.92	8.80
외	3.19	8.06	5.46	3.39	3.39	3.19	3.64	3.64	6.74	6.85

기존 기술과 비교

- 유재원(1997)-동일 자질 회피 기술 → 분명하게 드러나지 않음
  - 혀위치 자질: [전설]+[전설], [후설]+[후설]을 회피
  - 원순자질: [평순]+[평순], [원순]+[원순]을 회피
  - 혀높이자질: 특별한 선호도는 없다. / [중모음]-[중모음]에 대한 약한 회피

- 여러 제약으로 세분화되어 모음 연쇄 간의 인식 정도를 예측한다.
  - *[-low,+back,-round][-high](*[ɨ, ʌ][e, ʌ, o, ɛ, a], 가중치: 2.3)

  - *[+high, +back][+round, +syllabic] (*[wu, wo, uo], 가중치: 0.59)
    →본 연구와 Cho (2012)에서 공통적으로 보고한 제약
    →유재원(1997): [원순]+[후설], [후설]+[원순] 회피 ➔ 부분적인 일치

단어 경계 제약 – [모음] 제약

- 어말 제약
  - 차용어 외에 [ɨ]로 끝나는 단어는 없다. (강용순 1998, 신지영, 차재은 2003)
  - [ʌ]로 끝나는 단어도 잘 출현하지 않는다.

제약	가중치	가중치	예외
*[+high,+back,-round]#	5.19	[ɨ]#	없음
*[-low,+back,-round]#	1.78	[ɨ, ʌ]#	구어, 놀이터

- 어두 제약: [i, y] 외 모음으로 시작하는 단어는 회피된다.
  - 한성우 2006, Cho 2012와 부분적 일치: *#[e, ɛ, u]

제약	가중치	의미	예외
*#[-high,-low,-back,-round]	4.20	#[e]	없음
*#[-high,-back]	2.32	#[e, ø, ɛ]	애견, 외부
*#[-low,+back]	2.18	#[w, ɨ, u, ʌ, o]	어장, 오산
*#[+low]	1.92	#[ɛ, ɑ]	악기, 아이

결과 요약

- 기존연구와 부분적 일치
  - *[저해음]-[격음] 허용: [kp^h, kc^h]
  - *[비음]-[경음] 허용: [n][p', s', c', k'], [ŋ][경음]
  - *[경음, 격음]-[활음] 허용: [k'][w, j][p^h][j]

- 저빈도 연쇄를 세분화된 제약으로 포착
  - *[자음]-[모음]: *[자음]+[e, ɛ, ʌ], *[ŋ][u, ɛ]

- 새로운 음소배열제약 학습
  - [모음]-[모음] 연쇄
  - [모음]-[자음] 연쇄
    예: 밸 [pɛl] vs. 배래 [pɛlɛ]: 같은 문법성 정도를 예측
  - 단어 경계 모음 제약: *#[e, ø, ɛ, ɨ, u, ʌ, o, a], *[ɨ, ʌ]#

단어에 대한 비적형성 점수 예측

명사	비적형성 점수	위배되는 제약
뜻밖 [t'ɨtp'ak']	13.9	1) *[k']# 2) *#[t'] 3) *[ɨt] 4) *[tp'] 5) *[ak']
나이테 [naitʰe]	9.06	1) *[ai] 2) *[itʰ] 3) [tʰe]
잡티 [captʰi]	6.02	1) *[ptʰ] 2) *[tʰi]

- 비적형성 점수가 높을수록 한국어 단어로서의 문법성이 상당히 낮음을 예측

남은 문제 (1)

- 본 연구: 귀납적 모델, 기준 모델 제안
  - ☞ 실제 한국어 화자들의 인식은 어떠한 지 검토할 필요성이 있다.

- 이제까지 보고된 화자들의 인식 = 통계적인 지식 + **편향성**
  - 자연스러운 제약이 부자연스러운 제약보다 강도가 강하게 인식이 된다. (Hayes and White 2013)

- Pizzo (2015)
  - 구체적인 제약 vs. 일반적인 제약
  - 누적적 위배: 제약을 위배하는 만큼, 비적형성을 인식할 수 있는가?

남은 문제 (2)

- 본 연구: 명사

  ☞ '고유 단일어'가 한국어 화자들의 인식을 가장 잘 반영하는가?

- 가정-김경일(1985), 한성우(2006)
  - '순수한' 한국어 음운론을 보존한 어종

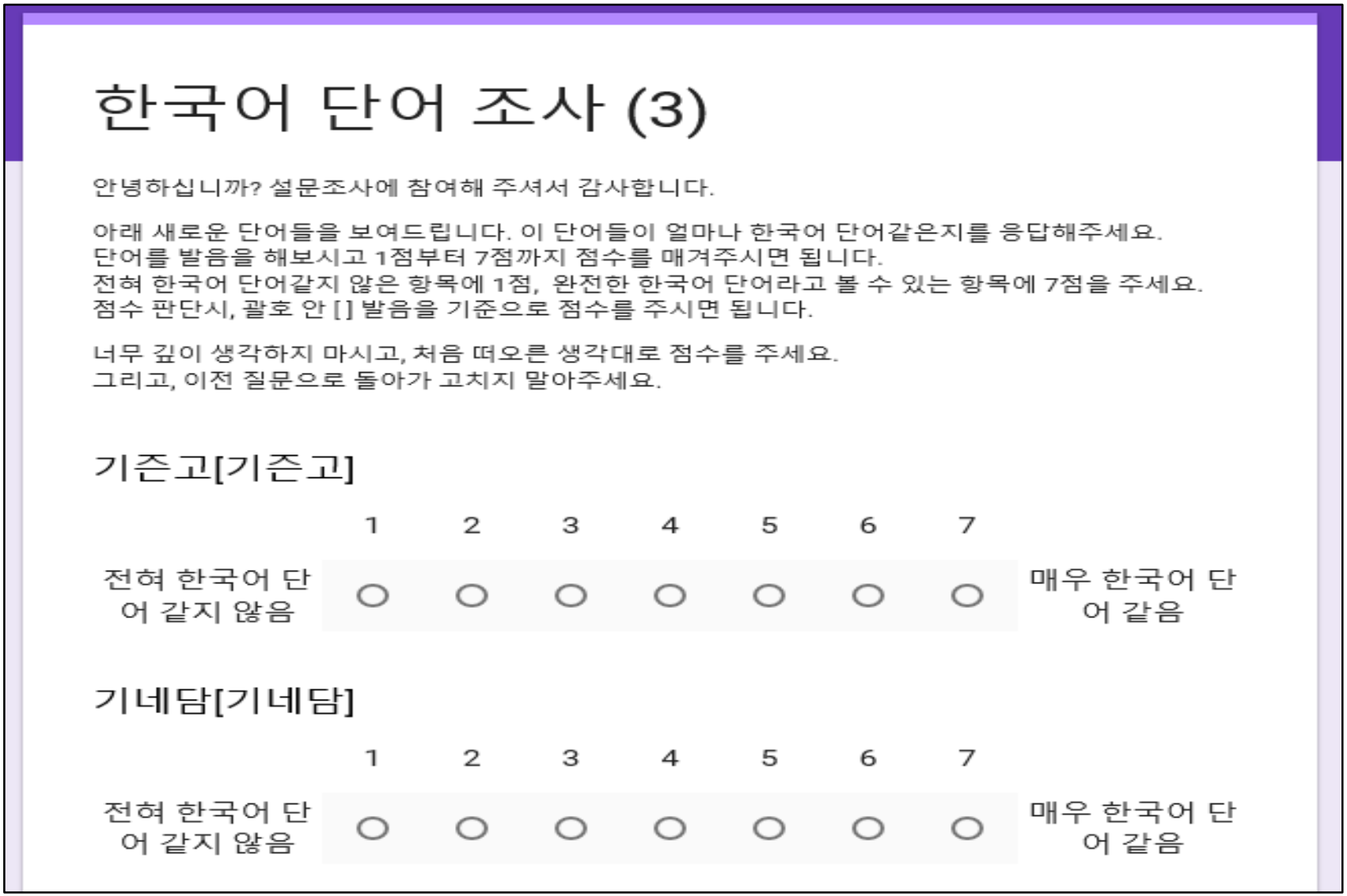
- 고유 단일어를 대상으로 학습
  - 기존 연구에서 보고된 경향성 등이 보다 뚜렷하게 학습된다.

    예: *[-sonorant][-tense] 가중치: 2.78

- 고유 단일어의 한정된 어휘 수로 인해, 학습된 제약의 실재를 증명해야 한다.
  - 예 *[+labial][+dorsal] 가중치: 2.73

cf. 한자어 및 복합어에서는 '감각, 밤거리' 등과 같이 자유롭게 나타날 수 있다.

비단어 조사 예시 – 철자, 기존단어의 영향 등을 고려–
제약을 위배하는 비단어 vs. 제약을 위배하지 않는 비단어



한국어 단어 조사 (3)

안녕하십니까? 설문조사에 참여해 주셔서 감사합니다.

아래 새로운 단어들을 보여드립니다. 이 단어들이 얼마나 한국어 단어같은지를 응답해주세요.
단어를 발음을 해보시고 1점부터 7점까지 점수를 매겨주시면 됩니다.
전혀 한국어 단어같지 않은 항목에 1점, 완전한 한국어 단어라고 볼 수 있는 항목에 7점을 주세요.
점수 판단시, 괄호 안 [ ] 발음을 기준으로 점수를 주시면 됩니다.

너무 깊이 생각하지 마시고, 처음 떠오른 생각대로 점수를 주세요.
그리고, 이전 질문으로 돌아가 고치지 말아주세요.

기즌고[기즌고]

	1	2	3	4	5	6	7	
전혀 한국어 단어 같지 않음	○	○	○	○	○	○	○	매우 한국어 단어 같음

기네담[기네담]

	1	2	3	4	5	6	7	
전혀 한국어 단어 같지 않음	○	○	○	○	○	○	○	매우 한국어 단어 같음

참고문헌

강범모·김흥규(2009), 「한국어 사용 빈도」, 서울: 한국문화사.

강용순(1998), “한국어 어휘부 구조”, 「음성. 음운. 형태론연구」 4, 55-67. 한국음운론학회.

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

김미란·최재웅·홍정하(2014), “한국어 초성-중성 결합의 분포적 특성 및 모음의 군집분석 연구”, 「음성. 음운. 형태론연구」 20.1, 23-49. 한국음운론학회.

박나영(2014), “한국어 명사의 음소배열제약에 대한 기계학습.” 「음성·음운·형태론연구」 20.3, 297-322. 한국음운론학회.

박동근·이석재(2005), 대학생 구어의 음운 실현 연구, 서상규·구현정(편), 「한국어 국어 연구(2)-대학생 말뭉치를 중심으로-」, 서울: 한국문화사.

박선우·홍성훈·변군혁(2013), “한국어의 어휘계층과 음운론적 복잡성”, 「음성. 음운. 형태론연구」19.2, 225-274. 한국음운론학회.

신지영(2008), “성인 자유 발화 자료 분석을 바탕으로 한 한국어의 음소 전이 빈도”, 「언어청각장애연구」 13.3, 477-502. 한국언어청각임상학회.

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

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

이상억(2001), 「계량언어학」, 서울: 박이정.

이용은(2009), “음소연속체 빈도가 비단어의 단어성 판단에 미치는 효과에 관한 연구”, 「영어영문학 연구」, 51.2, 215-234. 한국중앙영어영문학회.

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

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

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

허웅(1985), 「국어 음운학」, 서울: 샘문화사.

Berent, I., Steriade, D., Lennertz, T & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104, 591-630.

Cho, Hyesun (2012), Statistical learning of Korean phonotactics, *Studies in Phonetics, Phonology and Morphology* 18.2, 339-370. The Phonology-Morphology Circle of Korea.

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

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 & 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* 39.3, 379-440.

Hong, Sung-Hoon (2010), Gradient vowel cooccurrence restrictions in monomorphemic native Korean roots. *Studies in Phonetics, Phonology and Morphology* 16.2, 279-295. The Phonology-Morphonology Circle of Korea.

Lee, Yongeun (2007), Effects on inter-phoneme probabilities on the acceptability judgment of Korean CVC nonwords. *Speech Sciences* 14.4, 41-52. The Korean Society of Speech Sciences.

Lee, Yongeun (2011), The Gradient Nature of Syllabic Affiliations of Korean Glides, *Korean Journal of Linguistics* 36.3, 735-764. The linguistic society of Korea.

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

감사합니다.

음소배열제약의 정의

- 음소 단위의 결합 또는 회피에 대한 문법적 인식
  - ✓새로운 단어에 대해서도 적형성(well-formedness)을 판단 가능
  - ✓영어의 어두 자음군 인식(Chomsky and Halle 1965:101)

brick	실재 단어	적형
blick	비단어	적형
lbick	비단어	비적형

- '적형/비적형'의 이분법적 인식으로만 분석될 수 있는가?

비범주적 문법 인식

- 출현하지 않는 연쇄 간 비범주적 인식(Berent et al. 2007)
  - ✓영어 자음군: [적형] blif > bnif > bdif > lbif [비적형]
  - ✓보편적 음운론적 원리(공명도 원리)로 해석

- 연쇄의 출현 빈도 차이가 문법 인식에 반영된다.
  - ✓영어 [비음]-[저해음] (Hay et al. 2003: 5-6)

    고빈도 /nt/ 중빈도 /nf/ > /mk/ > /ms/

    수용도: /nt/ > /nf/ > /mk/ > /ms/

- **음소 결합 빈도가 문법 인식에 반영될 수 있다.**
  - ✓ 음소 결합 정도를 측정하는 척도
  - ✓ 음소 결합 정도를 반영하는 모델

기존연구의 한계

- 기존연구: 범주적 음소배열제약
  - 우연한 빈칸 ⇨ 화자들의 인식 여부?
  - 예외에 대한 인식이 산발적으로 언급

- 기존연구: 계량적 접근
  - 연구자마다 연구 대상이 다르다.
  - 결합 정도를 측정하는 척도가 연구자마다 상이하다.
  - 음소 연쇄의 출현 빈도가 문법으로 반영될 기제가 없다.
  - 결합 정도에 대한 화자들의 인식이 충분히 검증되지 않았다.

- 기존연구: Cho (2012)
  - 일부 음소배열제약 보고
  - 학습 문제 발생

← 적은 어휘 목록(명사: 3,404, 동사: 1,345, 형용사 및 기타: 376)

[첨부 1] 제약 학습 과정

- 1 단계: 제약 선택
  - 탐색 범위: 자질 목록을 바탕으로 모든 표상 집합을 구성
    - 정확도(O/E): $\frac{\text{제약 위배의 관찰 빈도}}{\text{제약 위배의 기대 빈도}}$
    - ⇨ 기대보다 위배되지 않을 제약 선택
    - ⇨ 기대보다 적게 나타나는 연쇄가 제약으로 포착
    - 일반성: 결합되는 자질의 수가 적고, 더 많은 분절음을 지시하는 제약 선택
      - *[+고설성][+고설성][+후설성] vs. *[+고설성][+고설성]
      - *[+전방성, +설정성][-후설성,+성절성] vs. *[+설정성][-후설성]

- 2 단계: 가중치 학습
  - 가정: 학습자는 대상 언어로부터 관찰된 형태에 접근할 수 있다.
  - 관찰된 형태의 확률을 최대화하고, 관찰되지 않는 형태의 확률을 최소화하는 가중치 학습

- 일정 수준의 정확도에 이를 때까지 반복적으로 학습

	*#CC	*VV	비적형성 점수 (Harmony)	최대 엔트로피 값 (e^{H})	출현확률 (P)
	3	2			
CV	0	0	(0x0)+(0x0)=0	exp(-0) = 1	0.84
CCV	1	0	(3x1)+(2x0)=3	exp(-3)≒0.05	0.11
CCVV	1	1	(3x1)+(2x1)=5	exp(-5)≒0.006	0.04

[첨부 2] 비교음소배열제약

- Hayes (2014)
  "(절대) 음소배열제약"과 "비교음소배열제약"이
  별도의 문법 체계일 수 있음을 시사

- 뜻밖 : 어두, 어말에 경음이 위치 제한
  - ✓ 문법성 매우 낮음 → "한국어 같지 않음"
  - ✓ 한자어에서 강하게 회피되는 연쇄 → "(한자어)에 비해 고유어같음"

[첨부 3] “한국어 단어 같음”에 대한 인식 점검

- 학습 자료에 포함되지 않은 차용어
  - 비적형성 점수가 높을수록, ‘한국어답지’ 않음을 시사

차용어	비적형성 점수	위배되는 제약
템포 [$t^h emp^h o$]	7.28	1) [$t^h e$] 2)[mp^h] 3)[em]
모토 [$mot^h o$]	0	

- *[pt^h] 명사: 밥투정, 업태
  cf. 차용어 허용: 캡틴

- *[ie] 명사: 지에밥, 귀엣말
  cf. 차용어 허용: 비엔날레, 피에로