고유어와 한자어의 비교 음소배열제약

서울대학교 언어학과 박사과정

박나영 (arimnet@naver.com)

한국어학회 제 68차 전국학술대회

2015년 2월 27일(금) 서강대학교

고유어 vs. 한자어

- 어원의 차이 ➔ 경쟁 관계 ➔ 하위 어휘 목록 형성
  - ✓음운, 형태, 의미의 차이
  - ✓공시적인 문법지식으로 실재

- 고유어와 한자어의 음운론적 차이

  - 한자어: 구성음절의 음소분포 차이(송기중 1992, 권인한 1997, 신지영 2009, 안소진 2009)
    - 초성: 경음 및 'ㅋ'의 제한
    - 중성: 고유어에 비해 이중모음(특히, 'ㅖ, ㅘ')이 많이 쓰임
    - 종성: 'ㄷ, ㅅ, ㅈ'과 경음, 격음, 자음군 제한

  - 한자어: 일부 음운 규칙이 적용되지 않는다.
    - /o/ > /u/(채서영 1999) 고유어: 하루 한자어: 삼촌
    - 모음축약(이주희 2005) 고유어: 아이~애 한자어: X

cf. 고유어와 마찬가지로 한자어에서도 음운규칙(비음화, 유음화, 경음화)이 적용(강용순 1998)

본 연구의 목적

- 기존연구
  - 한자어 대상
  - 구성 음절 구조 또는 개별 음운 규칙 분석

- 본 연구
  고유어와 한자어를 구별할 수 있는 음운론적 제약 탐색

1. '비교 음소배열제약' 개념 도입
2. 제약 탐색 및 선정: 통계 & 기계학습 방법
3. 결과: 제약 및 제약의 강도
4. 검토: 개별 단어가 각 어휘 부류에 속할 확률 점검

비교 음소배열제약(comparative phonotactics)

- 음소배열제약: 음소의 결합을 선호하거나 규제하는 문법 지식

- 비교 음소배열제약(Hayes, in press)
  - 어휘 부류들 사이의 선호도를 결정하는 제약
    Y([A] 선호): 음운론적 형태인 Y는 어휘 부류 [A]를 선호한다.

    - Y를 포함한 단어가 '경쟁 어휘 부류 [B]'에 속하는 것을 저지
    - 가중치(weight): 제약의 강도

비교 음소배열제약 적용 (1)

- ‘꿩 [k’wʌ□]’에 대한 비교 음소배열제약
  - 어두 경음 [k’] **한자어 회피** **고유어 선호** ➔ 한자어 후보 위배
  - 이중 모음 [wʌ] **한자어 선호** **고유어 회피** ➔ 고유어 후보 위배

단어	후보	#[k’] (고유어 선호)	[wʌ] (한자어 선호)
		가중치: 10	가중치: 7
꿩 [k’wʌ□]	고유어		1
	한자어	1	

비교 음소배열제약 적용 (2)

- 조화값(Harmony) : 각 후보가 위배하는 제약의 가중치를 더한 벌점
  - 해당 단어가 해당 어휘 부류에 부적합한 정도를 나타낸다.

단어	후보	#[k’] (고유어 선호)	[wʌ] (한자어 선호)	조화값
		가중치: 10	가중치: 7	
꿩 [k’wʌ□]	고유어		1	7
	한자어	1		10

- 조화값(부적합도) 비교
  - 고유어인 ‘꿩’ 조화값 **7** vs. 한자어인 ‘꿩’ 조화값 **10**
  - **고유어에 가까움**

개별 단어가 특정 어휘 부류에 속할 확률

- 전체 조화값(부적합도)에 대한
  특정 어휘 부류 조화값(부적합도)의 비율로 계산

  - e(≒2.718)를 밑으로 삼고 조화값을 음의 지수로 취한다. ➔ e^{-H}

    확률(어휘 부류 A) =

$$\frac{\exp(-\text{후보 } A\text{의 조화값})}{\exp(-\text{후보 } A\text{의 조화값}) + \exp(-\text{후보 } B\text{의 조화값})}$$

  - '꿩'이 고유어일 확률= $\frac{\exp(-7)}{\exp(-7)+\exp(-10)}$ = 0.952

고유어와 한자어의 비교 음소배열제약

단어	후보	#[k’] (고유어 선호)	[wʌ] (한자어 선호)	조화값	예측 확률
		가중치: 10	가중치: 7		
꿩 [k’wʌ□]	고유어		1	7	0.952
	한자어	1		10	0.048

1. 어휘 부류를 선호하는 제약
   - 1단계: 각 어휘 목록에서 회피되는 제약 탐색
   - 2단계: 두 어휘 부류를 구별하는 제약 선정

2. 가중치 학습
   - 각 제약이 개별 단어의 실제(입력) 어휘 부류를 최대한 예측할 수 있도록 가중치를 배운다

3. 개별 단어의 어휘 부류를 예측
   - 새로운 단어의 어휘 부류를 예측할 수 있다.

어휘 목록

- 어휘 목록=고유어(5,542 단어) + 한자어(29,869 단어)
  - 빈도 5 이상인 일반명사(강범모·김흥규 2009); 단일어와 복합어 포함
  - 표준국어대사전(http://www.korean.go.kr) 참조
    - 고유어와 한자어의 판별 기준: 한자음과 1:1 대응 여부
    - 표준국어대사전 발음형 기준

✓ 추후 형태론적 정보

불규칙적인 발음형

수의적인 발음형 등이 추가적으로 고려되어야 한다.

- 어말 자음은 모음으로 시작하는 조사와 결합할 때, 실현될 수 있는 바 , 음운론적으로 가정되는 기저형을 입력형으로 채택하였다.

제약 탐색

- 고유어와 한자어 각각에서 회피되는 연쇄를 자동적으로 학습 (박나영 2014)
  - 기계학습 방법: 최대 엔트로피 음소배열제약(Hayes and Wilson 2008)
    - 표면형으로부터 음소 연쇄의 출현 확률을 계산
    - 유의미하게 빈도가 낮은 음소 연쇄를 제약으로 포착
    - 자질 결합을 회피하는 유표성 제약
      예: *#[+긴장음], *[+성절][+성절]
  - 소프트웨어: UCLA 음소배열학습자(UCLA phonotactic learner)
    - 입력형: 어휘 목록 + 자질 목록 ➔ 출력형: 제약 + 가중치
    - 각각 100개씩 학습 ➔ 5회 반복 학습 중 3회 이상 학습된 제약
- 결과: 156 제약
  한자어 회피 제약 91개 + 고유어 회피 제약 82개 – 공통제약 17개

제약 선정

- 각 어휘 목록에서 회피되는 제약이 모두 두 어휘 부류에서 변별된다고 보기 어렵다.
- 두 어휘 부류를 모두 위배하지 않는 제약을 제외
  - 대상 제약: 116개 (한자어 학습 제약 67개 + 고유어 학습 제약 53개 – 공통 제약 4개)

➢ 로지스틱 회귀분석(logistic regression)
- 범주 분류에 영향을 주는 요인을 검정
- 두 어휘 부류를 판별할 수 있는 제약을 찾을 수 있다.
  - 계수(coefficient): 제약이 선호하는 부류와 영향력
  - 유의성(significance): 제약이 우연한 것인지

전체모델: bayesglm(어휘 부류 ~ 제약 1 + 제약 2 + 제약 3 + ... + 제약 116)
반응 변수 설명 변수

1단계: (다른 입력형, 새로운 단어)에 대해
예측력이 높은 제약 선택 모델 선택법: **라소(Lasso)**

2단계: 유의 수준 0.05 이하인 제약 **우연한 제약 제외**
유의 수준 0.05 이상, 계수 6이상 제약 **영향력 큰 제약 포함**

최종 결과: 65개 제약(한자어 학습 제약 42개 + 고유어 학습 제약 23개)

제약의 가중치 학습

- 선호하는 어휘 부류를 제약에 명세
  한자어에서 회피(학습) ➔ 고유어 선호 제약
  고유어에서 회피(학습) ➔ 한자어 선호 제약
- ✓ 제약을 위배하지 않는 단어가 없도록,
  한자어 기본 선호 제약, 고유어 기본 선호 제약을 더하였다.

- MaxEnt Tool(소프트웨어) 이용: 각 제약에 가중치 할당

단어	후보	입력 부류	#[k’] (고유어 선호)	[wʌ] (한자어 선호)
			가중치	가중치
꿩[k’wʌ□]	고유어	1		1
	한자어	0	1	

- ‘꿩’과 같은 단어가 고유어로서 제약을 위배하는 **빈도 → 가중치 반영**
  한자어로서 제약을 위배하는 **빈도**

결과

- 고유어 선호 제약 (42개) > 한자어 선호 제약 (23개)
  - 대체로 고유어 선호 제약이 두 어휘 부류 판별에 기여

- 본 모델은 기본적으로 한자어를 선호한다.
  - 한자어 기본 선호 제약: 1.936
  - 고유어 기본 선호 제약: 0

  - 입력단어의 구성 비율: 한자어 어휘 목록 > 고유어 어휘 목록
    ← 기본 선호 정도는 입력단어의 구성 비율에 따라 달라진다.

- 한국어 화자가 새로운 단어의 어휘 부류를 응답
  - 새로운 단어가 아무런 음소배열제약을 위배하지 않을 때
    - 응답 확률: 한자어 부류 > 고유어 부류
  - 새로운 단어가 음소배열제약으로 포착될 때,
    -한자어에서 회피되는 연쇄 多 고유어에서 회피되는 연쇄 少

고유어 선호 제약 (1)

- 기존연구: 초성, 중성, 종성 ➔ 어두, 어중, 어말 제약
  - 한자어에서 회피되는 연쇄를 단어 내 위치를 포함하여 포착

어중, 어말: 고유어 선호 제약 (해석)	가중치	고유어 예
가. [t][자음]	7.240	깃털 [kitt^{h}ʌl]
나. [t^{h}, s, c]#	5.943	옷 [os]
다. [t^{h}, c^{h}, p^{h}, k^{h}] #	5.764	앞 [ap^{h}]
라. [□m]#	4.877	뱀 [p□m]
마. [up, op]#	4.714	굽 [kup]
바. [자음][자음]#	4.654	흙 [h□lk]

어두: 고유어 선호 제약 (해석)	가중치	고유어 예
사. #[n, m][ʌ]	5.858	너머 [nʌmʌ]
아. #[k', p', s', t']	5.068	또래 [t'ol□]

고유어 선호 제약 (2)

- [자음] + [e, ʌ]
  - ‘세, 제, 체’를 제외한 [자음]+[e] 연쇄 선호; [격음]+[ʌ]; ‘터, 테’에 대한 선호

고유어 선호 제약 (해석)	가중치	고유어 예
가. [자음][n, m, l][e, ʌ]	5.467	들머리 [t□lmʌlij]
나. [p, p’, t, t’, t^h, k, k’][e]	4.540	떼 [t’e]
다. [t^h][e]	3.503	테두리 [t^hetuli]
라. [k^h, p^h, t^h][ʌ]	2.824	터[t^hʌ]

- [자음] , [모음] 연쇄 제약 – 음소 결합을 포착
  - ‘ㄹ’ 에 이어지는 설정 경음화 ➔음소배열제약<마>

고유어 선호 제약 (해석)	가중치	고유어 예
마. [l, p, t, k][t, s, c]	5.697	길잡이 [kilcapi]
바. [m, □][k^h]	4.677	넝쿨 [nʌ□k^hul]
사.[i, e, ʌ][i]#	3.249	어이 [ʌi]
자. [ɨ, ʌ][후설모음]	3.093	거울 [ʌul]

고유어 선호 제약 (3)

- [자음]+[모음]+[자음] 제약

고유어 선호 제약 (해석)	가중치	고유어 예
가. [t, t', t^h][i, ɨ, e, ʌ][m, n, l]	4.812	무덤 [mutʌm]
나. [p, p', p^h][ɨ, ʌ][k, □]	3.998	범벅 [pʌmpʌk]
다. [t^h, k^h][u][n, □]	3.851	귀퉁이 [kwit^hu□i]
라. [t, t', t^h][ɨ][l]	3.410	들 [t□l]

- 세부적인 환경을 명세하여 한자어에 잘 나타나지 않는 음절을 포착
  - 털, 덤, 덥, 틈, 듭 [가]
  - 벅, 벙 [나]
  - 퉁 [다]

들, 뜰, 틀 [가, 라]

한자어 선호 제약 (1)

- 어말 제약: [ɛk]# (가장 높은 가중치 제약)

한자어 선호 제약 (해석)	가중치	한자어 예
가. [ɛk]#	4.255	여객 [jʌk□k]
나. [p, p^h, k, k^h][ʌ, a]#	1.771	검거 [kʌmkʌ]
다. [p', t', k'][u, ʌ, o]#	1.155	각도 [kakt'o]

- 자음과 모음의 결합 제약
  cf. 기존연구: 한자어 고빈도 음절 정보만으로 다 포착하기 어려움

한자어 선호 제약 (해석)	가중치	한자어 예
마. [c^hɨ]	3.395	측정 [c^h□kc'ʌ□]
바. [c'ɨ]	2.520	녹즙 [nokc'□p]
사. [c^h][e]	2.281	체육 [c^hejuk]
아. [s', c'][e]	2.181	법제 [pʌpc'e]
자. [k^h, h][u]	2.169	후각[hukak]

한자어 선호 제약 (2)

- [자음]과 [활음]의 결합 제약

한자어 선호 제약 (해석)	가중치	한자어 예
가. [□] [w, ɯ, j]	4.095	강연 [ka□jʌn]
나. [p, k] [ju]	3.195	규정 [kjucʌ□]
다. [hj]	3.114	휴가 [hjuka]
라. [t, s, c] [wʌ, wa]	2.960	좌석 [cwasʌk]
마. [p, m, k,h, □] [ɯ]	2.505	심의 [simɯi]

- 이중모음 제약: j계 이중 모음인 경우(사, 아, 자), 선·후행 모음에 상관없이 한자어에서 더 선호될 수 있다.

한자어 선호 제약 (해석)	가중치	한자어 예
바. [wʌ]	3.243	월급 [wʌlk□p]
사. [ɛ, a][j, ɯ]	2.742	가열 [kajʌl]
아. [u][j, ɯ, w]	2.200	수요 [sujo]
자. [i, u, e, ʌ, o][j, ɯ, i, a]	1.153	제약 [ejak]

한자어 선호 제약 (3)

- 그 밖의 제약

한자어 선호 제약 (해석)	가중치	한자어 예
가. [a][ɛ, ʌ, o, a], [ʌɛ]	2.591	가액
나. #[ɨ]	2.443	응답
다. [p, t, k][p^h, t^h, k^h]	0.999	목표

결과 정리

- 고유어 선호 제약(한자어 회피 연쇄)
  - 기존연구 확인: 한자어 초성, 종성 제한 ➔ 어두, 어말 제약으로 포착
  - 구성 음절에 대한 양적 분석과 대응
    - 기존연구: 한자어에서 회피되는 결합(양적연구)
    - 본 연구: 고유어 선호 제약
      -'세, 제, 체'를 제외한 [자음]+[e, ʌ]연쇄 선호,
      - [모음] 연쇄 선호
    - 세부적인 환경을 명세하여, 한자어에 잘 나타나지 않은 음절을 포착

- 한자어 선호 제약(고유어 회피 연쇄)
  - 가장 변별적인 제약: [ɛk]#
  - 이중 모음에 대한 제약
  - [자음]과 [모음]의 결합: [츠, 즈, 세, 제, 체 후]

장점

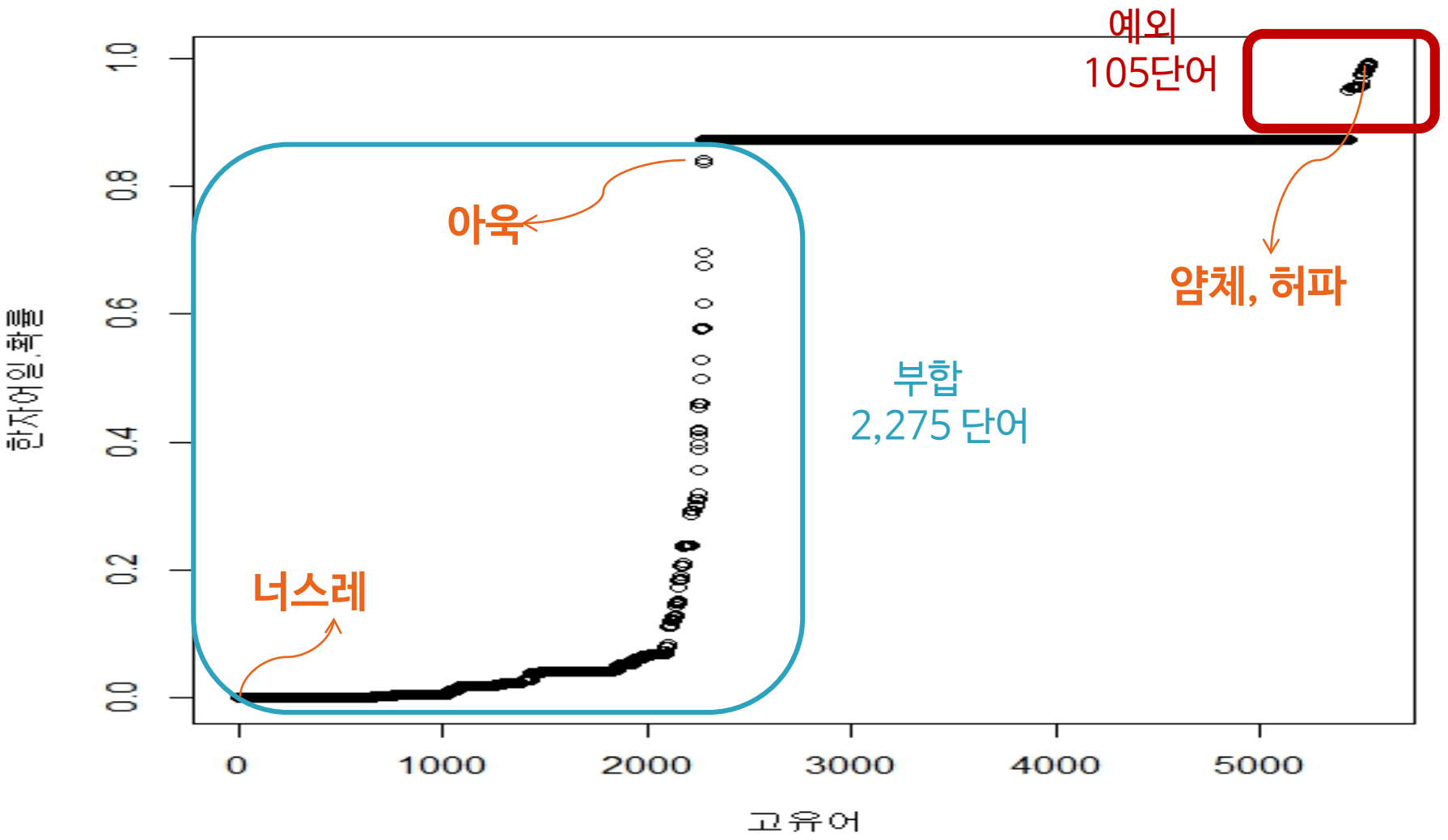
1. 낮은 빈도로 출현하는 연쇄를 제약으로 포착　　예: 게 [憩, 揭]

2. 한 단어에 두 어휘 부류의 속성이 혼재하는 경우에도
   그 단어의 어휘부류를 예측할 수 있다.

- '휴게'　　입력형: 한자어　　출력형: 고유어에 다소 가까움

단어	후보	[ke] (고유어 선호)	[hj] (한자어 선호)	조화값
		4.54	3.114	
휴게[hjuke]	고유어		1	3.114
	한자어	1		4.54

개별 단어에 대한 예측 확률 -고유어-



개별 단어에 대한 예측 확률 -고유어-

- 한자어 기본 선호 제약 1.936
  - 음소배열제약을 위배하지 않는 단어 (고유어: 3155 단어, 한자어: 22045 단어)
    한자어일 확률: 0.874 기준

- 한자어일 확률: 고유어
  - 한자어일 확률이 기준(0.874)보다 낮은 고유어
    - P < 0.2(2168 단어)

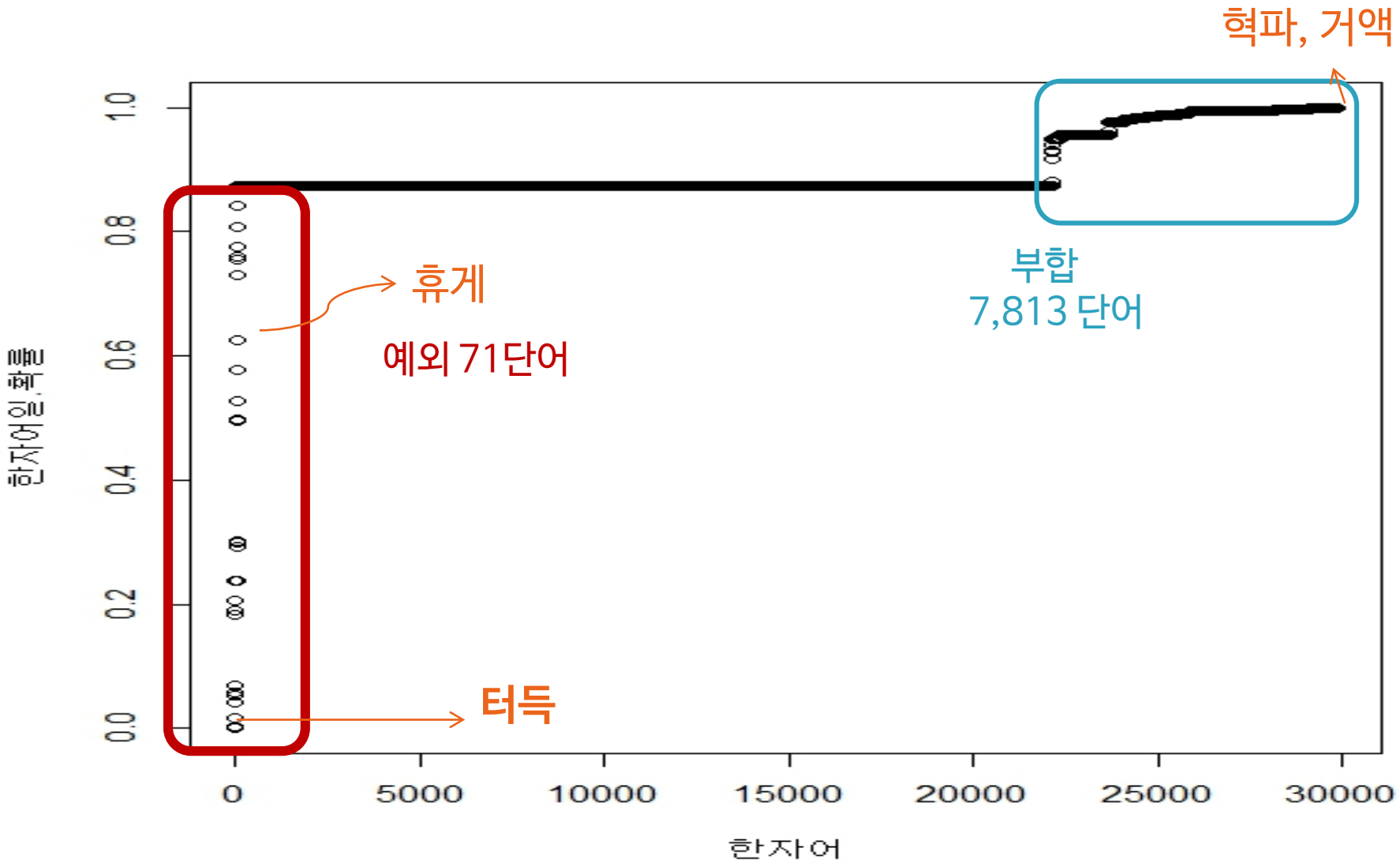
      너스레(0), 고삐(0.023), 뱀(0.05), 또아리(0.121), 키다리(0.173)
    - P ≥ 0.2(107 단어)

      겨울(0.239), 오빠(0.3), 이엉(0.578), 애꾸(0.576), 아욱(0.618)

  - 한자어일 확률이 기준(0.874)보다 높은 고유어(105 단어)

    흉내(0.994), 응석(0.988), 얌체(0.985), 허파(0.976), 비아냥(0.956)

개별 단어에 대한 예측 확률 -한자어-



- 한자어일 확률: 한자어
  - 한자어일 확률이 기준(0.874)보다 높은 한자어 (7741단어)
    - 혁파(1), 좌석(0.993), 억제(0.984), 소아(0.956), 여왕(0.881)
  - 한자어일 확률이 기준(0.874)보다 낮은 고유어 (71단어)
    - 터득(0.03), 쌍(0.042), 틈입(0.053), 별게(0.069), 잡음(0.207), 합헌(0.292), 우울(0.301), 이앙(0.578), 휴게(0.625), 파업(0.776)

- 실제 화자를 대상으로 그 인식을 점검할 필요가 있다.

- 보다 많은 단어들을 포착하기 위해

  고빈도 음절, 형태론적 제약 등을 포함할 필요 있음

남은 문제

- 한자어 에서 회피되는 제약(고유어 선호 제약)
  - 공명음에 이어지는 경음 제약

예: [-자음] [-지속,+긴장] [-후설,+성절] 가중치: 5.702
[모음][p', t', k', c'][i, e, ɛ] 해당고유어: 새끼

  - 하나의 운율 단위로 정의되지 않는 [자음]+[모음]+[모음], [모음]+[모음]+[자음] 제약

예: [+공명,-양순] [+고설,+후설,+성절] [-원순,+성절] 가중치: 5.030
[n][ɨ, u][i, ɨ, e, ʌ, ɛ, a] 해당 고유어 '누이'

  - ✓ 실재하는 제약인가?

- 개별 단어의 사용빈도(token frequency) 고려
- 한자어는 고유어와 다른 형태론적 구조와 입력형을 가정해야 하는가?
- 단일어 vs. 복합어: 음소배열제약? 형태론적 제약?

참고문헌

강범모·김흥규. 2009. 「한국어 사용 빈도」 한국문화사.

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

권인한. 1997. “현대국어 한자어의 음운론적 고찰.” 「국어학」(국어학회) 29, 243-260.

김창섭. 2001. “한자어 형성과 고유어 문법의 제약.” 「국어학」(국어학회) 37, 177-195.

김한샘. 2005. 「2005년 신어」 국립국어원.

박나영. 2014. “한국어 명사의 음소배열제약에 대한 기계학습.” 「음성·음운·형태론연구」 (한국음운론학회), 20(3), 297-322.

박선우·홍성훈·변군혁. 2013. “한국어의 어휘계층과 음운론적 복잡성.” 「음성·음운·형태론연구」 (한국음운론학회) 19(2), 225-274.

송기중. 1992. “현대국어 한자어의 구조.” 「한국어문」(한국정신문화연구원) 1, 1-85.

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

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

안소진. 2011. 「심리어휘부에 기반한 한자어 연구」 서울대학교 국어국문학과 박사학위논문.

안소진. 2014. “한자어 형태론의 제 문제와 어휘부.” 「한국어학」(한국어학회), 62, 373-394.

이운영. 2002. 「표준국어대사전 연구분석」 국립국어연구원.

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

참고문헌

Capodieci, Frank, Bruce Hayes and Colin Wilson. 2008. UCLA Phonotactic learner. http://www.linguistics.ucla.edu/people/hayes/Phonotactics/

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

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

Friedman, Jerome, Trevor Hastie, Noah Simon, Rob Tibshirani. 2014. glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.9-8. http://cran.r-project.org/web/packages/glmnet/.

Gelman, Andrew, Yu-Sung Su, Masanao Yajima, Jennifer Hill, Maria Grazia Pittau, Jouni Kerman, Tian Zheng, Vicent Dorie. arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.7-07. http://cran.r-oject.org/web/packages/arm/

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy Model. Proceedings of the Workshop on Variation within Optimality Theory, Stockholm University, 2003.

Hayes, Bruce. in press. Comparative phonotactics. Proceedings of the 50th meeting of the Chicago Linguistic Society. April 10, 2014. http://www.linguistics.ucla.edu/people/hayes/Papers/HayesComparativePhonotacticsCLS50_2014.pdf

Hayes, Bruce and Colin Wilson. 2008. A Maximum Entropy Model of phonotactics and phonotactic learning.Linguistic Inquiry 39(3), 379-440.

Hayes, Bruce, Colin Wilson, Ben George. 2009. MaxEnt Grammar Tool. http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool.

Hayes, Bruce and Kie Zuraw. 2013. Class 10 More on Ratings vs. Probability, Lecture handouts. Linguistic 251 Variation in Phonology. UCLA. http://www.linguistics.ucla.edu/people/zuraw/251_2013/AllLecturesLing251HayesZuraw.zip

Itô, Junko and Armin Mester. 1995. The core-periphery structure of the lexicon and constraints on reranking. In Jill N. Beckman, Laura Walsh-icky and Suzanne Urbanczyk (eds.). University of Massachusetts Occasional Papers in Linguistics 18, Papers in Optimality Theory, 181-209. Amherst, MA: GLSA.

R Core Team. 2014. R: A language and environment for statistical computing. Version 3.1.1. R Foundation for StatisticalComputing,Vienna,Austria. www.R-project.org.

Zheng, Vicent Dorie. arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.7-07. http://cran.r-project.org/web/packages/arm/

감사합니다.