Exceptions

Class 8: Phonotactic probability

Adam Albright

CreteLing 2024

creteling2024.phonology.party

Taking stock

- So far: examined predictability of specific features/properties, and distribution of exceptions
  - Morphological regularity
  - Phonological feature values
- Although final segments show expected frequency distribution (exceptional = higher token frequency), properties elsewhere show a tendency in the opposite direction
- Tentative suggestion: phonotactically improbably items are avoided
- A more powerful test: low global phonotactic probability

Bigram probability and acceptability

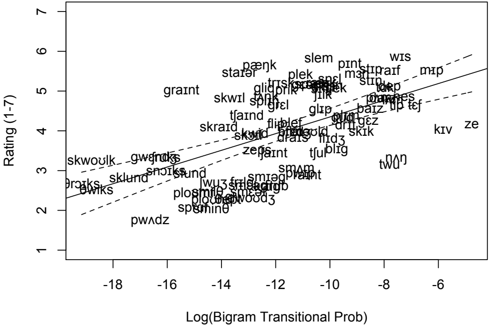
- We're interested in the distribution of words that speakers treat as *exceptional*
  - Exist, but disallowed/penalized by the grammar
- Such words should be phonotactically *unacceptable*
- It's hard to ask speakers about the acceptability of existing words, but we can estimate it using existing models
- First step: a holistic measure of phonotactic probability
  - Transitional bigram probability in the English lexicon

Modeling phonotactic acceptability

- Transitional bigram probability $P([abc \ldots x_n]_{\mathrm{Wd}}) =$
  $P(a \mid [_{\mathrm{Wd}})P(b \mid a)P(c \mid b) \ldots P(x_n \mid x_{n-1})P(]_{\mathrm{Wd}} \mid x_n)$
- Calculated over segments, or featurally defined natural classes (Albright, 2009)

Transitional bigram probability models acceptability

- Phonotactic acceptability judgments (Albright & Hayes, 2003)



- Transitional bigram probability from CELEX
- Not a perfect estimate of acceptability, but one of the best available for attested combinations
- For similar results, see Hayes & Wilson (2008), Albright (2009)

Testing this for two languages

- English
- Korean
  - Nouns
  - Verbs

English: restricting to monosyllables

- Transitional probability decreases rapidly with the length of the string
  - Hard to compare predictions for words with different numbers of syllables
- A restricted test: monosyllables
- Various choice points
  - Bigram probabilities calculated over monosyllables, or all words
  - Sensitive to syllabic role or not
  - Segments or features (natural classes)
- Report here on segmental bigrams, calculated on syllabified monosyllables

Approximating the English lexicon

- Frequency data: Open American National Corpus (OANC), second release[1]
- Combined inflected forms of lemmas, single entry with sum of counts
  - 23,451 distinct lemmas

[1] http://www.anc.org/data/anc-second-release

- Spoken portion: 3.8M tokens, mostly from SWITCHBOARD Godfrey & Holliman (1993)
- Automated tagging and lemmatization
- 41,463 distinct wordforms

Phonetic transcriptions

- American English transcriptions from CMU pronouncing dictionary
  - First ('primary') CMU pronunciation, converted to IPA
  - First pass: no POS or homophone differentiation
  - 19,367 transcribed entries, of which 4,657 are monosyllables
- Automated syllabification
  - Goal: distinguish onset vs. rhyme/coda consonants
  - Coda consonants given diacritic
- I will largely ignore stress here, except as reflected indirectly through vowel reduction

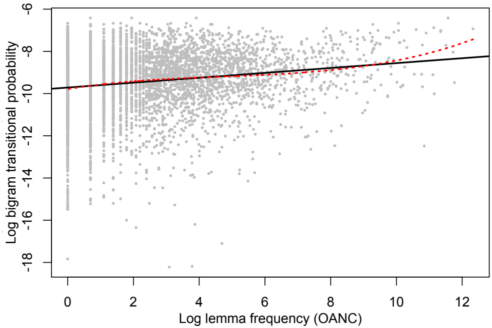
Probable and improbable monosyllables (C = coda)

you	ju	−6.417	*frisked*	fɹɪskt	−15.205
for	fɔɹ	−6.669	*swooshed*	swuʃt	−15.224
see	si	−6.720	*valve*	vælv	−15.278
hoe	hoʊ	−6.790	*garbed*	gɑɹbd	−15.290
rue	ɹu	−6.809	*briefs*	bɹifs	−15.320
core	kɔɹ	−6.850	*oomph*	umf	−15.364
do	du	−6.881	*dweeb*	dwib	−15.428
why	waɪ	−6.907	*tongs*	tɑŋz	−15.453
be	bi	−6.934	*glimpse*	glɪmps	−16.002
coo	ku	−6.946	*sixth*	sɪksθ	−16.195
co.	koʊ	−6.953	*midst*	mɪdst	−16.352
we	wi	−6.967	*length*	lɛŋkθ	−17.103
too	tu	−7.010	*depths*	dɛpθs	−17.834
ray	ɹeɪ	−7.012	*strength*	stɹɛŋkθ	−18.187

The general strategy

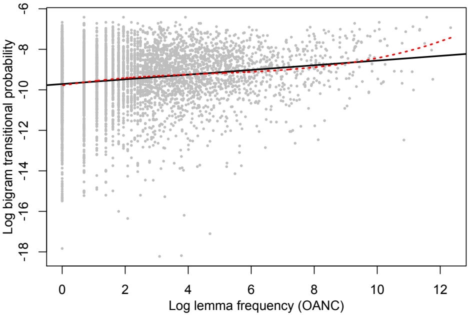
- Now we have an estimate of how probable (phonotactically ordinary) vs. improbable (phonotactically ‘exceptional’) each word is
- Next step: examine frequency distribution
  - Do improbable ($\sim$exceptional) words tend to have higher token frequency?
  - Conversely, do more high frequency words tend to be more probable ($\sim$regular)?

Phonotactic probability and token frequency



- Small but highly significant effect: phonotactically less probable words tend to have *lower* token frequency
- Holds even when differences in segment count are taken in account

Phonotactic probability and token frequency



- Model: Bigram trans. prob. $\sim$ segment count $+$ log lemma freq

	Est	Std Err	t val	P($>$ $\|$t$\|$)
Intercept	-5.691	0.071	-80.69	$<$2e-16
segment count	-1.129	0.0187	-60.53	$<$2e-16
log lemma freq	0.047	0.007	6.66	3.05e-11

Constructing a comparable test for Korean

- As with English, desirable to restrict comparisons to words of comparable length
- Ideally, test with a lexicon of comparable size, with comparable frequency distribution
- Significant phonotactic differences between nouns and verbs/adjectives in Korean, and also frequency differences
    - Potential confound: if there are many more nouns than verbs, but verbs tend to have higher token frequency, high frequency words could look phonotactically ‘unusual’ just because they are verbs
    - Approach: model nouns and verbs separately

A Korean lexicon

- Started with the 90,257 lemmas in the Sejong corpus
  - Removed symbols, letter names, suffixes, entries in Hanja, etc.
- Nouns
  - Small number of monosyllables compared to English OANC corpus (only 587), so took 15,386 mono- and disyllables
- Verbs
  - Small number of verbs compared to English OANC corpus, so took all 3,750 verbs
- Within each set, calculated bigram transitional probability

Phonetic transcriptions

- Converted Sejong entries to phonetic transcription
  - Transliterated and applied regular phonological processes
- Phonotactics of morphemes, or surface forms?
  - In principle, also curious whether morphemes ending in clusters are ‘exceptional’ and require high frequency
  - Retained coda clusters in phonetic transcription

Nouns: probable and improbable monosyllables

이	i	−4.929	룸	rum	−12.720
시	ʃi	−5.024	삯	sagd	−13.077
지	ji	−5.197	렛	red	−13.353
연	yən	−5.423	램	rɛm	−13.357
부	bu	−5.542	삶	salm	−13.803
전	jən	−5.546	몫	mogd	−13.840
도	do	−5.562	흙	hɨlg	−13.865
구	gu	−5.575	뺨	Byam	−13.908
고	go	−5.577	꾐	Gwem	−13.997
영	yəŋ	−5.602	샷	syad	−14.173
사	sa	−5.641	랩	rɛb	−14.206
기	gi	−5.646	슛	syud	−14.432
수	su	−5.689	앎	alm	−14.458
정	jəŋ	−5.724	넋	nəgd	−16.132

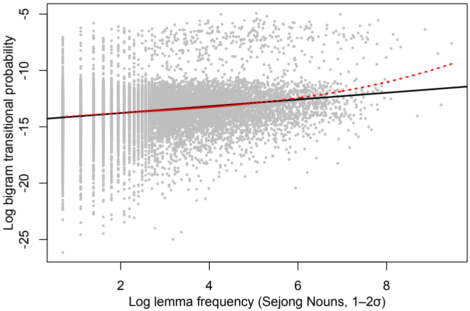
Nouns: probable and improbable disyllables

사이	sai	−10.132	뒤곁	dwiGyəd	−22.830
연시	yənʃi	−10.259	흙밭	hɨlgBad	−22.966
이지	iji	−10.283	흙물	hɨlŋmul	−23.005
도시	doʃi	−10.322	칼슘	kalsyum	−23.076
구이	gui	−10.331	링겔	riŋgel	−23.219
부시	buʃi	−10.335	벨벳	belbed	−23.257
고시	goʃi	−10.337	캡프	kɛbpɨ	−23.462
이리	iri	−10.373	뒷켠	dwidkyən	−23.507
지시	jiʃi	−10.375	튜브	tyubɨ	−23.783
전시	jənʃi	−10.381	룸펜	rumpen	−24.001
연지	yənji	−10.385	귀띔	gwiDyim	−24.346
영시	yəŋʃi	−10.431	헬멧	helmed	−24.474
바이	bai	−10.446	뜀틀	Dwimtɨl	−24.581
사시	saʃi	−10.454	캡슐	kɛbsyul	−24.978

Same strategy as for English

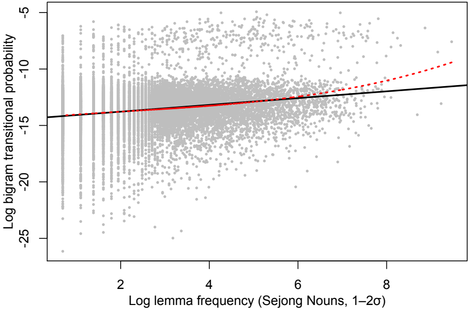
- Examine relation between n-gram probability and token frequency
  - Do improbable ($\sim$exceptional) words tend to have higher token frequency?
  - Conversely, do more high frequency words tend to be more probable ($\sim$regular)?

Phonotactic probability and token frequency



- As for English, phonotactically less probable words tend to have *lower* token frequency
- Holds even when differences in segment count are taken in account

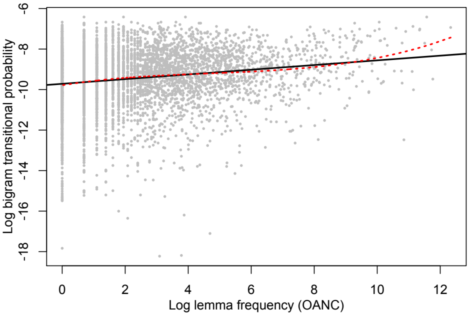
Phonotactic probability and token frequency



- Model: Bigram trans. prob. $\sim$ segment count $+$ log lemma freq

	Est	Std Err	t val	$P(>\lvert t\rvert)$
Intercept	-9.275	0.0788	-117.7	<2e-16
segment count	-0.940	0.0136	-69.1	<2e-16
log lemma freq	0.228	0.009	24.8	<2e-16

Phonotactic probability and token frequency



(Comparison with English)

Phonetic transcriptions for verbs

- A perennial problem for calculating well-formedness in highly inflected languages: what form to use?
- Interest here is really on the 'stem', but not pronounceable in isolation
  - Stem-final simplifications and irregular allomorphy
- Abstraction: stem+"V"
  - That is, a faithful surface form of the stem, as it would occur before a vowel
  - Ignores allomorphy due to irregularity, hiatus resolution, glide formation etc.

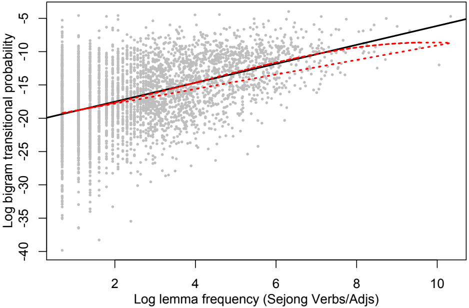
Verbs: probable and improbable monosyllables

지	ji-	−3.981	꺽	GəG-	−12.887
이	i-	−4.435	곪	goḻ.m-	−12.909
기	gi-	−4.479	삶	saḻ.m-	−12.919
시	ʃi-	−4.578	깍	GaG-	−12.925
하	ha-	−4.705	뱉	bɛt-	−12.997
치	ci-	−4.731	젊	jəḻ.m-	−13.033
가	ga-	−5.024	짧	Jaḻ.b-	−13.050
닿	da-	−5.070	섞	səG-	−13.101
나	na-	−5.129	얇	yaḻ.b-	−13.503
마	ma-	−5.171	솎	soG-	−13.560
드	dɨ-	−5.280	좇	joc-	−13.766
비	bi-	−5.375	읊	ɨḻ.p-	−13.902
자	ja-	−5.667	훑	huḻ.t-	−13.944
피	pi-	−5.771	쫓	Joc-	−14.092

Verbs: probable and improbable disyllables

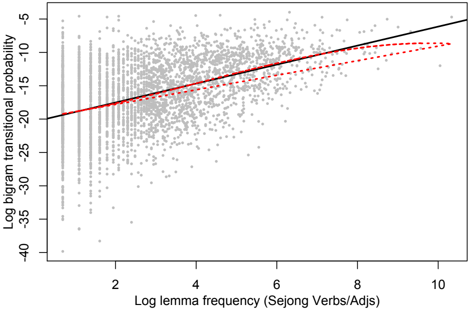
들이	dɨri-	−6.321	벗삼	bə(t).sam-	−20.418
어리	əri-	−6.999	헛짚	hə(t).jip-	−20.459
거리	gəri-	−7.034	점찍	jəm.Jig-	−20.519
그리	gɨri-	−7.059	맴돌	mɛm.dol-	−20.562
부리	buri-	−7.540	객적	gɛg.jəg-	−20.670
여리	yəri-	−7.556	뒤쫓	dwi.Joc-	−20.796
우리	uri-	−7.574	끝맺	Gɨt.mɛj-	−20.855
꺼리	Gəri-	−7.574	흉보	hyuŋ.bo-	−21.080
가리	gari-	−7.778	짱박	Jaŋ.bag-	−21.126
서리	səri-	−7.787	설삶	səl.sal.m-	−21.279
달이	dari-	−7.824	손쉽	son.swib-	−21.635
쓰리	Sɨri-	−7.923	샘솟	sɛm.sos-	−22.389
벌이	bəri-	−7.947	있잖	it.jyan-	−22.828
오리	ori-	−7.952	폭넓	poŋ.nəl.b-	−23.671

Phonotactic probability and token frequency



- A consistent result: phonotactically less probable words tend to have *lower* token frequency
- Similar trends for both verbs and adjectives

Phonotactic probability and token frequency



- Model: Bigram trans. prob. $\sim$ segment count $+$ log lemma freq

	Est	Std Err	t val	$P(>\|t\|)$
Intercept	-3.302	0.193	-17.10	$<$2e-16
segment count	-1.963	0.020	-95.85	$<$2e-16
log lemma freq	0.278	0.026	10.73	$<$2e-16

Summary of whole-word probability

- Contrary to predictions, low frequency words are not phonotactically more probable, at least as measured holistically by transitional bigram probability
- This runs contrary to the expectation that low frequency words should be more ‘regular’
- In fact, low frequency words tend to be phonotactically more unusual/improbable
    - Similar effect seen for English, and Korean (nouns, verbs/adjs)
- Cannot be reduced to independent effect of high frequency words having fewer segments
- May still be consistent with other types of ‘reduction’ among high frequency words

From acceptability to grammaticality

- Result in this section focuses on bigram probability as a proxy for *phonotactic acceptability*
- Indirectly linked to grammatical exceptionality
  - Unacceptable $\Rightarrow$ improbable $\Rightarrow$ grammatically dispreferred
- If grammar regulates the distribution of specific features in specific contexts, then perhaps we need a more pinpointed test
  - Exceptions to the distribution of specific features
  - Exceptions to specific grammatical constraints

References

Albright, Adam. 2009. Feature-based generalization as a source of gradient acceptability. *Phonology* 26.9–41.

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90.119–161.

Godfrey, John, and Edward Holliman. 1993. *Switchboard-1 release 2 ldc97s62. web download*. Philadelphia: Linguistic Data Consortium.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.