Learning alternations using sublexical phonology

Blake Allen (University of British Columbia) • Michael Becker (Stony Brook University) • LSA 2015, PDX

Finding affixes without supervision

(1) Minimum edit distance alignment finds changes:

[bɹʌʃ ∼ bɹʌʃɪz] 'brush(es)' →

b	ɹ	ʌ	ʃ	∅	∅
b	ɹ	ʌ	ʃ	ɪ	z

(2) Challenge I: finding the position of affixes

Ulwa: left-oriented *ka* sometimes appears on the right

base	possessed	
bas	bas.*ka*	'hair'
sa.paa	sa.paa.*ka*	'forehead'
suu.lu	suu.*ka*.lu	'dog'
kuh.bil	kuh.*ka*.bil	'knife'

Is the [ɪz] of [bɹʌʃ-ɪz] right-oriented?

(3) Challenge II: finding affix material

English: what is the analysis of [ɹoʊz ∼ ɹoʊzɪz] 'rose'?

ɹ	oʊ	z	∅	∅
ɹ	oʊ	z	ɪ	z

vs.

ɹ	oʊ	∅	∅	z
ɹ	oʊ	z	ɪ	z

The alignment algorithm greedily matches edge segments.

Even worse in Russian [lʲef ∼ lʲvof] 'lion':

lʲ	e	f	∅	∅
lʲ	∅	v	o	f

vs.

lʲ	∅	e	f
lʲ	v	o	f

(4) Solution: generate many hypotheses about the affix, consume the smaller ones.

k	ɪ	s	∅	∅
k	ɪ	s	ɪ	z

→ **'add [ɪz] at right edge'**; 'add [ɪz] after 3rd segment'

b	ɹ	ʌ	ʃ	∅	∅
b	ɹ	ʌ	ʃ	ɪ	z

→ **'add [ɪz] at right edge'**; 'add [ɪz] after 4th segment'

ɹ	oʊ	∅	∅	z
ɹ	oʊ	z	ɪ	z

→ 'add [zɪ] before last segment'; 'add [zɪ] after 2nd segment'

(5) Hypothesis reduction (simplified):

Hypothesis H_L is consumed by hypothesis H_W iff
• H_W and H_L are *idempotent* on H_L's bases
• and H_W covers *more lexical items* than H_L.

Learning the distributions of affixes

(6) The distribution of an affix is learned by fitting a MaxEnt phonotactic grammar to the words that take it.

Grammar for the words that take plural [–s]:

	obs. p	*[+voice]# $w = 2.5$	*[+strid]# $w = 4.6$	$\mathscr{H}$	exp. p
peak	1			0	.32
bead	0	−1		−2.5	.03
quiche	0		−1	−4.6	.003
chief	.8			0	.32
shelf	.1			0	.32
⋮					

(7) The "add [–s]" **sublexicon** =
- The list of words that take plural [–s]
- The operation "add [–s] at right edge"
- The phonotactic grammar that rejects final stridents and final voiced segments

Similarly for the "add [–z]" and "add [–ɪz]" sublexicons.

(8) Three sublexicons → three derivatives for [wʌg]:
- [wʌgs] violates *[+voice]#
- [wʌgɪz] violates *[−strident]#
- [wʌgz] is well-formed, gets most of the probability.

Learner summary

training paradigms + feature matrix
↓ *alignment*
hypotheses
↓ *hypothesis reduction*
+ constraint set
↓
sublexicons
+ testing = wugs
↓
nonce derivatives

The English past tense

(9) Data from Albright & Hayes (2003): Training 4253 real verbs, testing 58 nonce verbs.

(10) Finding supra-segmental changes from the data:

[wɪn ∼ wʌn] → **'change last nucleus to [ʌ]'**; 'change penult segment to [ʌ]'

[dɹɪŋk ∼ dɹʌŋk] → **'change last nucleus to [ʌ]'**; 'change antepenult seg to [ʌ]'

(11) Finding supra-segmental changes without support:

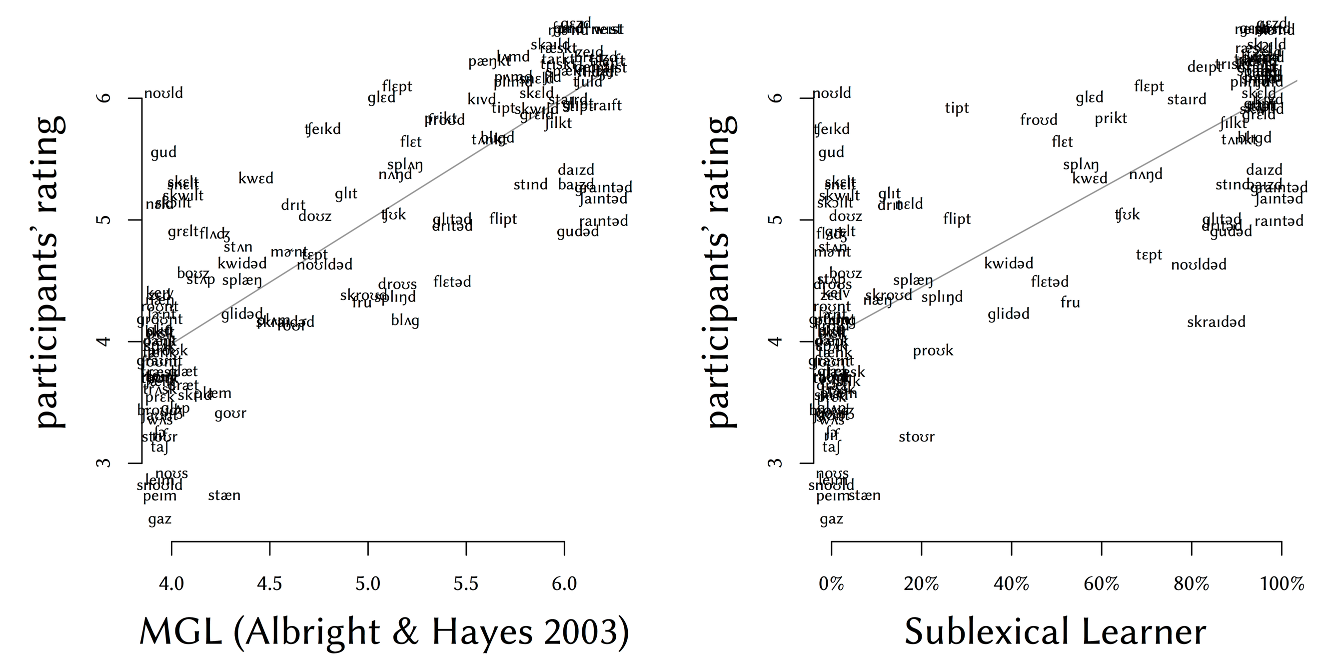
[wiv ∼ woʊv] → 'change last nucleus to [oʊ]'; 'change penult segment to [oʊ]'

We added a bias for supra-segmental hypotheses.

(12) Sublexicons found with threshold of ≥12

Sublexicon	Coverage	
Add [d] at right edge	2104	ʤasəl, spɑɹ, vɪʒuəlaɪz, ...
Add [əd] at right edge	1146	hænd, koʊɹt, sʌfəkeɪt, ...
Add [t] at right edge	791	ɪŋkɹis, tɹɛspæs, mʌf, ...
Make last nucleus [oʊ]	30	wiv, stɹaɪv, fɹiz, teɪɹ, ...
No change	29	lɛt, ʌpsɛt, bit, ɹisɛt,...
Make last nucleus [ʌ]	20	spɹɪŋ, dɪg, dɹɪŋk, wɪn, ...
Make last nucleus [ɛ]	18	fɔl, fid, hoʊld, ...
Make last nucleus [æ]	15	ɹʌn, sɪt, ɹɪŋ, bɪgɪn, ...
Add [t] at right edge, make last nucleus [ɛ]	12	slip, kɹip, nil, fil, dil, ...

(13) Both learners do really well: Albright & Hayes (2003) $\rho = .82$, our $\rho = .79$ (difference not significant)



Check it out! sublexical.phonologist.org