Complex segments or clusters? A learnability approach

Overview. Languages are thought to differ in the status of sequences such as [ts]: they are single segments in some (e.g. Russian) but clusters in others (e.g. English), suggesting that the distinction is learned. Phonological reasoning about complex segments is usually phonotactic: if the sequence has the same distribution as a single segment, treating it as a segment allows for a simpler characterization of the syllable structure (Riehl 2008, Lin 2011, a.o.). Thus, Boumaa Fijian has [mb, nd, ndʒ, ŋg, tʃ, nr], but no other clusters (see (1)). If these sequences are analyzed as single segments, then Fijian can be analyzed as a strict (C)V language (Dixon 1988):

(1) Fijian (Austronesian): complex segments or clusters? (Syllable breaks assumed)

ta.ŋ͡ga	'bag, pocket'	n͡ra.no	'lake'	n͡dʒi.na	'truly'
ŋ͡ga.ta	'snake'	n͡do.βu	'sugar cane'	m͡bo.to	'frog'

Phonotactic reasoning implies that learners use phonotactics to decide whether a given consonant sequence is a segment or cluster. We demonstrate instead that complex segments can be learned from distributional information, before any phonotactic learning takes place. We show that our proposal makes more restrictive predictions regarding extant limitations on the size and composition of complex segments than do current alternatives (Inkelas & Shih 2013, Inkelas & Shih 2018, a.o.).

Procedure. Our implemented computational learner constructs complex segment representations by tracking distributions. Initially, the learner has only simplex consonants, featurally defined. The learner tracks the frequencies of consonant bigrams (C_1C_2) and of all the individual consonants C_1, C_2... C_n. These frequencies are used to calculate an *inseparability measure* for each bigram: $\text{Prob}(C_1C_2)/\text{Prob}(C_1) * \text{Prob}(C_1C_2)/\text{Prob}(C_2)$. This measure exceeds 1 when the likelihood of either C being in the sequence is greater than the likelihood of C being on its own or in other clusters. Any sequence whose inseparability measure exceeds 1 is reanalyzed as a complex segment and is given a composite feature representation. The learning data are analyzed again with the C_1C_2 sequence rewritten as a new C_3; the process is *iterated* until no sequences exceed the inseparability threshold.

Iteration allows the learner to identify complex segments that have more than two parts (e.g. [ndʒ] in Fijian) in a computationally efficient and conceptually simple way: only bigrams have to be examined on any given iteration. In addition, sometimes complex segments have Cs that appear in more than one sequence (e.g. [n] in Fijian [nd], [ndʒ], [nr]), which means that multiple iterations can be necessary for all C-containing sequences to pass the inseparability threshold.

Results. We have tested the learner on a number of language corpora, with representative results summarized in the table. Two illustrative case studies are discussed in more detail below.

Language	Corpus (wds)	Complex segs. found	Other clusters?	Iterations
English	92,969	tʃ, dʒ	yes	1
Fijian	17,642	mb, nd, ŋg, ndʒ, tʃ, nr	no	2
Latin	22,192	–	yes	1
Mbay	4,046	mb, nd, ŋg, nɟ	yes	1
Ngbaka	5,420	gb, kp, mb, nd, ŋg, ŋmgb	no (marginal)	2
Russian	101,530	ts, tɕ	yes	1
Sundanese	13,405	mb, nd, ŋg, mp, nt, ŋk	yes	2
Wargamay	5,910	mb, nd, ŋg	yes	1

Case 1: Fijian. We tested the learner on a corpus of Fijian (http://crudaban.org, transcribed according to Dixon 1988 and cleaned). On the first iteration, the learner identified affricates and prenasalized stops, which had high inseparability measures (ŋg=30.91, mb=25.23, nd=22.55, tʃ=16.29, dʒ=8.97). This is unsurprising, as the phones [ʃ, ʒ, b, d, g] do not occur outside of these sequences. The prenasalized trill [nr] came close to the threshold (insep=0.91) but did not pass it until the second iteration. This is because both of its parts are frequent in other CC sequences ([n]) and/or separately ([n], [r]). On the

second iteration, [ndʒ] and [nr] were identified: since they were the only sequences left, they necessarily had high inseparability values. The learner thus converged on Dixon's (1988) inventory after two iterations.

Case 2: Mbay (Nilo-Saharan). We tested the learner on a digitized dictionary of Mbay, which is described as having prenasalized voiced stops [mb, nd, nɟ, ŋg] (Keegan 1996, 1997). Unlike Fijian, Mbay has sequences that are not analyzed as complex segments (rk, mk, etc.), so the learner's problem is harder: it must identify some sequences as single segments, while others remain clusters. Moreover, in Mbay, subparts of complex segments [g, b, ɟ, d, m, n, ŋ] occur as singletons. The learner arrives at the right analysis in one iteration: there is a clear cut-off point between the high inseparability measures of prenasalized stops (ranging from 2.32 to 12.46) vs. clusters (rk=.06, mk=.04, etc.). The measures of the clusters rise slightly on the next iteration (rk=.34) but are nowhere near the threshold of 1. Thus the learner does not create any more complex segments.

A phonotactic alternative. Suppose the learner has a restricted set of hypotheses about possible complex segments: they could be affricates, prenasalized stops, or labiovelars. The learner tries phonotactic learning on increasingly more complex representations of the data: clusters only, then clusters plus affricates, etc. The more complex representations are kept if the phonotactic grammar achieves a better fit to the learning data than the grammar with a simpler segmental inventory.

We tried the phonotactic procedure by applying the UCLA Phonotactic Learner (Hayes & Wilson 2008) to versions of the English corpus transcribed with increasing numbers of complex segments (none, then adding [tʃ] and [dʒ], then [ts] and [dz], then prenasalized affricates and stops). The result: the more complex the inventory, the higher the log probability of the grammar. The phonotactic approach suggests that English has prenasalized affricates and stops. This result generalizes: phonotactic reasoning applied to Russian leads to the phonologically dubious conclusion that it has prenasalized affricates and stops.

The problem with phonotactic reasoning is that *replacing consonant clusters with single-segment representations always improves the phonotactic grammar*. Since individual segments can be banned from certain positions (e.g. word-initial [ŋ] in English), representing clusters as segments gives the learner a shorthand for explaining why certain sequences are positionally restricted without having to induce a full set of constraints to describe them. On the other hand, our distributional approach clearly separates affricates from clusters in both Russian and English. It does not attempt more complex analyses because they are not justified by the learning data.

Typological implications. Complex segments appear to be typologically limited in *complexity* and *composition*. In terms of complexity, complex segments often have two subparts, sometimes three, rarely four. In terms of composition, homorganicity is common – [mb, nd, ts] – but not obligatory ([kp, t^w]; Zsiga and Tlale 1998 document a labiocoronal fricative [ɸʃ] in Tswana). Segments like [ɬb, ʃʒ, χl] are thought to be unattested. Recent theories such as Q theory (Shih & Inkelas 2013 et seq.) explain the limitation on complexity through stipulation (every segment has three parts), but offer no explanation for the limitation on composition. Our proposal derives both typological tendencies. Restrictions on segment complexity emerge from restrictions on cluster complexity: segments containing more than three subparts are rare because consonant clusters containing more than three subparts tend to be rare (at least morpheme-internally). Similarly, restrictions on the composition of segments emerge from restrictions on the composition of clusters: segments like [mb, nd, ts] are common because homorganic nasal-stop and stop-fricative clusters are common, and frequent in languages that allow them; segments like [ɬb, ʃʒ, χl] are likely unattested because these clusters are dispreferred.

Selected References. Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379-440. Lin, Y. H. (2011). Affricates. In van Oostendorp et al., eds. *The Blackwell companion to phonology*. Riehl, A. (2008). *The phonology and phonetics of nasal-obstruent sequences* (Doctoral dissertation, Cornell University). Shih, S. S., & Inkelas, S. (2018). Autosegmental Aims in Surface-Optimizing Phonology. *Linguistic Inquiry*, *50*(1), 137-196.