Selecting for Size: Survival of the Frequent

Introduction. The precise nature of affixal subcategorization for phonological size has long been controversial: is it a glimpse into markedness, or a parochial lexical statement? The Emergence of the Unmarked (TETU, McCarthy and Prince 1994) attributes size restrictions to markedness, active elsewhere in the language and/or in broader typology. For example, in English, the minimal word is bimoraic: CVC [bɪt], CV: [biː] but not CV *[bɪ]. This is coextensive with the moraic trochee, the foot type in the language's stress system. It is also the size template for truncation ("Steph"), and the subcategorization frame of the suffix *-en* (*redden, stiffen* vs. **stupiden*; Siegel 1974, a.o.). Thus, even though English does not limit words to bimoraic monosyllables, an affix can use that template to cap bases.

The alternative I pursue is Sublexical Phonotactics: size subcategorization is an emergent phonotactic generalization over a list of stems (*sublexicon*, Becker & Allen 2015 et seq.). For English, the sublexicon of *-en* includes {*red, stiff, sick, loose*} but not {*blue, odd, hot, stupid, beautiful*}; the learner examines the sublexicon to discover (among others) the generalization that only monosyllables can combine with *-en*. The strength of the phonotactic generalization is proportional to the quantitative support for it in the sublexicon, and the discrepancies between the sublexicon and the relevant portion of the lexicon. I apply this to two case studies from Russian: a small sublexicon suffix with a size restriction, and a large sublexicon suffix without one.

Case Study: ***-ast***. The Russian adjectival suffix *-ast* means "possessing a big X", and usually attaches to nouns referring to body parts (e.g., [mórd-a] 'face (derog)', [mord-ást-ɨj] 'big-faced'. Shvedova (1980) describes the suffix as quite productive, but does not mention phonological restrictions. I identify a size restriction: the affix combines with maximally disyllabic stems. From the point of view of markedness, this is surprising. The complex lexical stress system of Russian (Melvold 1989, a.o.) does not supply unambiguous evidence for any kind of metrical foot, with both iambic and trochaic analyses on the market (Alderete 1999 vs. Revithiadou 1999) and debates about the location of default stress. Russian also does not restrict stems or words in size (other than the trivial "at least 1σ"). I suggest that the disyllabic maximum is a generalization over the sublexicon that learners encounter in frequent use.

Method. To study the affix's distribution, I examined Sharoff's (2005) frequency dictionary, which contains 32,000 lemmas with at least 1 occurrence per million in the Russian National Corpus (RNC). There are just 17 lemmas derived with the *-ast* suffix in this list: the affix is not type-frequent. Despite this, it is productive by other measures, such as the incidence of hapax legomena in the Aranea Russicum III Maximum corpus (19.8 bill. tokens). The Aranea corpus was searched for ≈14,000 forms constructed by script from nominal stems in Sharoff's frequent lemma list. These forms respected the phonological generalizations (incl. variable ones) but not semantic ones; this allowed to identify creative extensions of the affix's use. The search returned 213 *-ast* adjectives (Table 1): 82% monosyllables, 17% disyllables, mirroring the smaller subset in Sharoff.

	lexicon (13,104)		*sublexicon (17)*		*productive extension (213)*	
	All noun stems in Sharoff (2005)		*-ast* adjectives in Sharoff		Constructed form hits in Aranea	
σ	zúb 'tooth'	2,211	zub-ást-ɨj 'toothy'	14	sis^{j}-ást-ɨj 'big-boobed'	174
$\sigma\sigma$	golov-á 'head'	4,859	golov-ást-ɨj 'big-headed'	3	bitseps-ást-ɨj 'big bicepsed'	37
$\sigma\sigma\sigma$	boj-ev-ík 'action film/hero'	3,381	—	0	bojevik-ást-ɨj 'full of action'	2
$\sigma\sigma\sigma\sigma$	pere-nós-its-a 'bridge of nose'	1,823	—	0	— (*perenositsastɨj)	0
5+	fizionómij-a 'face, mug'	830	—	0	— (*fizionomiastɨj)	0

Table 1: The *-ast* "has a big X" suffix: overall lexical stem shapes vs. sublexical statistics, and productive extension of disyllabic size restriction in a corpus

Analysis. Crucially, *-ast* is semantically restrictive: it refers to external attributes only (e.g., there is no *[serts-ást-ɨj] 'big-hearted'), and while it can refer to textures/patterns (e.g. [tsvet-ást-ɨj] 'flowery'), most of the frequent uses denote body parts. The frequent stems are overwhelmingly native, monomorphemic, and short. Learners notice the generalization that certain phonological traits are absent from the *-ast* sublexicon. This can be demonstrated in a Monte Carlo simulation: if 17 noun stems were drawn at random from Sharoff (2005), the likelihood of picking only σ and $\sigma\sigma$ is low. With 10,000,000 random draws of 17 stems from the list of 13,104, only 298 draws included σ and $\sigma\sigma$ stems but not longer ones. Learners should conclude that the generalization is a strong one. Under the law of frequency matching (Hayes et al. 2009), speakers should then extend the trends in their productive use: mostly body parts, mostly monosyllables, occasionally disyllables. The two uses of *-ast* on trisyllabic stems (*bojevik-ast-ɨj* 'full of action', *amerik-ast-ɨj* 'American (derog)') are predicted if learners do not encode the disyllabic maximum in their rule for the affix: the absence of longer stems is a (mostly reliable) statistical fact, not a markedness effect.

Case study: ***-ost^j***. Compare this with another Russian suffix, *-ost^j/-est^j* [glásn-ost^j] 'transparency', which forms nouns from adjectives. This suffix is type-frequent, with 589 *-ost^j* noun lemmas in Sharoff, and it is not phonologically selective, combining with stems of varying lengths. Correspondingly, most of the adjectival stems in Sharoff's lemma list occur with the *-ost^j* suffix in the Aranea corpus, as shown in Table 2. (The corpus was searched for forms constructed from adjectives in Sharoff, as in the *-ast* study). Since this suffix is not at all semantically selective, its productivity is expected. Random Monte Carlo draws of 589 stems in 10,000,000 simulations produced 80,659 cases with the same exact syllable count types as in the *-ost^j* sublexicon. The syllable distribution in the sublexicon is likely random.

	lexicon (6,428)		*sublexicon (589)*		*productive extension (4,679)*	
	All adj stems in Sharoff (2005)		-ost^j nouns in Sharoff		Constructed hits in Aranea	
σ	jún-ɨj 'young'	680	jún-ost^j 'youth'	137	bós-ost^j 'barefooted. . .'	584
$\sigma\sigma$	opásn-ɨj 'dangerous'	2,420	opásn-ost^j 'danger'	170	bezlúnn-ost^j 'moonless. . .'	1,698
$\sigma\sigma\sigma$	blagodárn-ɨj 'grateful'	1,843	blagodárn-ost^j 'gratitude'	164	vodoródn-ost^j 'hydrogen. . .'	1,406
$\sigma\sigma\sigma\sigma$	izvorótliv-ɨj 'squirmy'	969	izvorótliv-ost^j 'squirminess'	82	zakl^{j}utʃítel^{j}n-ost^j 'conclusive. . .'	701
5+	(too long for examples!)	515		35		289

Table 2: The -ost^j nominalizing suffix: overall lexical stem shapes vs. sublexical statistics,

Discussion. The claim is that affixes do not specify size subcategorization in their formal rules. Instead, the generalizations, if any are present, emerge from interrogating the sublexicons for shared features, and identifying those properties that are mismatched between the relevant portion of the lexicon and the sublexicon. This decouples subcategorization from markedness considerations, which seems warranted. Kager (1996), assuming TETU, claims that no language has affixes that select for maximally trisyllabic stems, and yet in Russian, there is a suffix *-onok* 'baby X' (Gouskova & Bobaljik 2022), which combines with mono, di-, and trisyllabic stems ([barakud^j-ónok], 'baby barracuda'), but not four-syllable ones (*[orangutan^j-ónok] 'baby orangutan'). The proposal also explains why some generalizations are fragile, despite being inviolable (e.g., all 17 *-ast* adjectives in the frequent slice of the Russian lexicon are $\leq \sigma\sigma$): they are learned from quantitative evidence rather than hard-coded in the grammar, and small datasets are less reliable. Finally, the case is relevant to alternative theories of statistical learning, such as Yang's (2016) Tolerance Principle, which predicts exceptionless extension of *-ast* to monosyllables rather than frequency matching for disyllables.