Phonological Selection in Small Sublexicons

Maria Gouskova
New York University

1 Introduction

It is well-known that affixes can select for phonological properties of stems, such as syllable count or stress location (Siegel 1974 et seq.). I am interested in two questions. First, how do learners figure out these restrictions when the evidence comes from very few words? Second, why do the restrictions tend to be, in Pierrehumbert's (2001) terminology, *coarse-grained*? They usually concern syllable count, stress location, and C/V structure, and are rarely feature/natural class-based. I argue that phonological selection results from statistical generalization over *sublexicons*: lists of stems that combine with the affixes and that learners encounter frequently. Learners compare the composition of these sublexicons with a morphosyntactically comparable portion of the rest of the lexicon, zeroing in on generalizations that are unlikely to be due to chance. In a statistical learning theory, small sublexicons can only support generalizations for which every stem in the sublexicon is informative. Every stem has a syllable count and C/V structure, but generalizations about segmental composition require a bigger search space and are correspondingly data-greedy.

I examine the Russian adjective-forming suffix *-ast*, which imposes a disyllabic maximum on its stems (see (1)). The suffix is semantically restrictive in a way that parallels its phonological restrictions. As I show in two corpus studies, the suffix is nonetheless productive. In its productive uses, we see a frequency-matching pattern at the coarse level of syllable count. I also look at the pattern of generalization over stem-final consonants: Russian speakers extend the suffix to stems that end in consonants from natural classes unseen in the learning data, but this turns out to be unsurprising given the likelihood of encountering various types of stem-final consonants in Russian nominal stems.

(1) Russian adjectival *-ast* in brief: max 2 syllables

a.	σ	✓	pup	'belly button'	pup-ást-ij	'big-belly-buttoned'
b.	$\sigma\sigma$	✓	b^{j}ítseps	'biceps'	b^{j}itseps-ást-ij	'big-bicepsed'
c.	$\sigma\sigma\sigma$+	NO	trap^{j}étsij-a	'trapezius (muscle)'	*trap^{j}etsij-ást-ij	

For comparison, I demonstrate that size restrictions do not hold for two phonologically similar suffixes, *-ist* and *-izm*. For these suffixes, we see frequency matching of a different sort: the frequencies of stems of various sizes match what one would get by drawing from the lexicon at random.

(2) Russian *-ist, -izm* in brief: no max

a.	$\sigma\sigma$	komún-a	ko.mun^{j}-íst	komun^{j}-ízm	'commune'
b.	6σ	internatsionál	in.ter.na.tsi.o.na.l^{j}-íst	internatsional^{j}-ízm	'international'

The sublexicon theory is contrasted with several alternatives. Subcategorization Frames (Lieber 1980;

[*] I would like to thank the audiences at AMP 2023, Harvard, NYU, and UCLA for feedback on this and related work, and Christine Gu for comments on the manuscript. Supplementary materials are at `https://github.com/gouskova/smallsublex2025`.

Proceedings of *AMP 2023/2024*

Paster 2006) encode selection[1] as phonological contexts in the affix's morphological rule: e.g., Kalin and Rolle (2023) formulate a dual-condition frame for a Nancowry infix that encodes its post-vocalic position as $V\_\_\_\_$, and its selection for disyllabic bases as $\_\_\_\_\ Ft_{\sigma\sigma}$. I discuss several problems with this approach, both general (e.g., dealing with complex conditions and failing to be predictive) and specific to the Russian case (the selectional restriction does not target a constituent). I also relate my findings to Scheer's (2016) proposal for modular separation of coarse generalizations (stress, syllable structure, C/V) from granular feature-based selection, and to tolerance theory of Yang (2016).

2 A Russian affix that selects for mono- and disyllables

2.1 *Background on morphosyntax and phonology of* -ast. Shvedova (1980,§637) describes the Russian adjectival suffix *-ast* as quite productive in colloquial Russian. Adjectives with *-ast* mean something like "possessing a large or prominent X", where X is often a body part. Most examples in frequent usage (see (5)) are derived from body part nouns, *glaz* 'eye' ∼ *glaz-ast-ij* $\sqrt{\overline{\text{eye}}}$-ADJ-MASC.NOM.SG 'big-eyed',[2] though there are a handful of exceptions such as *ot͡ʃʲk-ást-ij* 'bespectacled'.

The phonological properties of *-ast* require some background. First, Russian stress is lexical and can fall anywhere in the word. There is no secondary stress except in some compounds (see Gouskova 2010). As a result, words can contain long strings of unstressed syllables (lapses), as shown in (3).

(3) Free stress location, long lapses allowed

a.	*σ́σσσσσ*	ví.ka.rab.ka.je.t͡sa	'will crawl out'
b.	*σσσ́σσσσ*	prʲi.buk.sʲí.ro.va.no.vo	'one that has been towed GEN'
c.	*σσσσσ́σσ*	prʲi.vʲi.lʲe.gʲi.ró.va.nij	'privileged NOM.SG'
d.	*σσσσσσσ́σσ*	tʲe.lʲe.do.ku.mʲen.ta.lʲís.tʲi.ka	'TV documentary making'

Stress interacts in complex ways with morphology (see Melvold 1989; Alderete 1999): roots can be stressed, unstressed, or follow a final stress pattern. Affixes fall into the same classes, with a further division into recessive or dominant. With recessive affixes, stress falls on the leftmost underlyingly stressed vowel, else on the first syllable. With dominant affixes, such as *-ast*, any stress markings on the stem are overriden, and a consistent pattern is imposed: dominant accented suffixes bear stress even if the roots are accented (see (4)):

(4) Dominant accented [-ást]

		stressed	unstressed	final
		/pálʲet͡s/	/volos/	kadik*/
a.	/-Ø/ NOM.SG	pálʲet͡s	vólos	kadík
b.	/-a$_{\text{REC}}$/ GEN.SG	pálʲt͡s-a	vólos-a	kadik-á
c.	/-ámʲi$_{\text{REC}}$/ INST.PL	pálʲt͡s-amʲi	volos-ámʲi	kadik-ámʲi
d.	/-ást$_{\text{DOM}}$-ij/ big X-MASC.SG	palʲt͡s-ást-ij	volos-ást-ij	kadik-ást-ij
		'finger'	'hair'	'Adam's apple'

The analysis of Russian stress is a long-standing controversy (Halle 1973, Revithiadou 1999, many others), and there is little consensus on questions such as whether there is a default, or what type of metrical foot (if any) the system deploys. The evidence on the latter is delicate. One argument is from vowel reduction: pretonic vowels follow a different pattern from other unstressed vowels, suggesting iambicity (Crosswhite 1999). Other arguments involve narrow morpho-phonological corners such as hypocoristics, which have penultimate stress ([iván] ∼[vánʲa] 'Ivan') consistent with trochaicity. None of this helps with *-ast*. Under an iambic analysis, the stem syllables preceding *-ast* are either foot-internal or split by a foot boundary (as

[1] Another theory of selection is the Emergence of the Unmarked (see McCarthy and Prince 1994, Mascaró 1996, vs. Paster 2006 and Embick 2010). This theory is usually discussed in the context of suppletion, because markedness constraints are supposed to select between two (or more) listed allomorphs. This does not straightforwardly extend to cases where the choice is between attaching the affix or not attaching it. There are filter-based approaches (e.g. Orgun and Sprouse 1999), which encounter the problem of how to handle contradictory conditions imposed by different affixes. Paster and Embick discuss other problems with this approach to selection.

[2] All examples are my own, transcribed phonemically in IPA. I omit word-final devoicing and vowel reduction.

in *(zu.b-ás)t-ij,* ‘toothy’, or *ko(r^{j}e.n-ás)t-ij* ‘stocky’). Under a trochaic analysis, stem syllables are unfooted. Under neither analysis do the stem syllables form a prosodic constituent.

This presents a problem for subcategorization frames: the original argument for feet in such selection cases was that they were constituents, so rules could refer to mono- and di-syllables without directly counting them (McCarthy and Prince 1986). But the mono- and disyllables that *-ast* attaches to are not constituents. I argue below that words derived with [-ast] obey a restriction on stress lapses above a certain length. This is a restriction on the sublexicon of *-ast*, not on Russian in general.

2.2 *The* -ast *suffix in the Russian National Corpus.* The semantics of the *-ast* suffix makes it difficult to investigate experimentally, since the world of body parts is a small one and does not readily admit new members. Instead, I looked at two corpora of Russian that represent different types of usage. First is a subset of the Russian National Corpus (RNC), specifically the 32,000 lemmata that occur at least once per million (Sharoff 2005, henceforth the Sharoff list). The RNC currently has 2 billion tokens, and it is curated by the Russian Academy of Sciences. The second corpus I considered is Aranea Russicum III Maximum, compiled by web-crawling and considerably larger at 19.8 billion tokens (henceforth Aranea). My working assumption is that most Russian speakers have encountered the Sharoff list lemmata by the time they reach adulthood, so the list makes a plausible model of their learning data. By contrast, Aranea is a good model of productive morphological extension, as it contains quite a few hapax legomena (one-off coinages; see Baayen and Lieber 1991). The Aranea corpus is also practical as it is uncensored and allows unlimited searches.

Sharoff’s list has only 17 adjectives derived with *-ast*. For scale, compare *-ost^j* ‘-ness’, as in *glasnost^j* ‘opennness’ (589 noun lemmata) or *-sk*, as in *sov^{j}etsk^{j}ij* ‘Soviet’ (856 adjective lemmata). All 17 *-ast* adjectives are in (5), with pooled lemma/token frequency. The stems are all consonant-final and respect a disyllabic size limit; most are monosyllabic (σ: 13, vs $\sigma\sigma$: 4).[3] Stress is inconsistent in the disyllabic bases; they represent four accentual types (annotated with subscripts). The semantic generalizations are similar: 13 of the stems refer to body parts, 4 are meronymic but are not body parts (‘root’, ‘flower’, ‘spectacles’, ‘crack’). Another salient generalization is that the list includes only external attributes (no heart, liver, etc.; see (13) below).

(5) Adjectives formed with *-ast* with a frequency of ≥1 per million in RNC (per Sharoff 2005)

				Occ. p/m	Corresp. noun (+pl)	
1.	$\sigma\sigma$	ko.r^{j}e.n-á.st-ij	‘stocky’	6.86	kór^j(e)n^j, kórn^j-i$_{\text{UNACC1}}$	‘root’
2.	σ	ʦv^{j}e.t-á.st-ij	‘colorful/flowery’	5.94	ʦv^{j}ét	‘color, flower’
3.	σ	mor.d-á.st-ij	‘big-faced’	4.53	mórd-a	‘face, mug’
4.	σ	sku.l-á.st-ij	‘w. big cheekbones’	4.47	skul-á	‘cheekbone’
5.	σ	oʧ^j.k-á.st-ij	‘bespectacled’	3.37	oʧ^{j}k^j-í	‘spectacles’
6.	σ	lo.b-á.st-ij	‘w. a big forehead’	3.31	l(ó)b, lb-í	‘forehead’
7.	σ	gor.l-á.st-ij	‘loud’	2.69	górl-o	‘throat’
8.	σ	zu.b-á.st-ij	‘toothy’	2.69	zub	‘tooth’
9.	σ	gla.z-á.st-ij	‘big-eyed’	2.26	glaz	‘eye’
10.	σ	u.ʂ-á.st-ij	‘w. prominent ears’	2.20	úx-o, úʂ-i	‘ear(s)’
11.	σ	gru.d-á.st-ij	‘chesty’	2.08	grud^j	‘chest’
12.	σ	gu.b-á.st-ij	‘big-lipped’	1.65	gub-á	‘lip’
13.	σ	ʃʃ^{j}e.l^j-á.st-ij	‘full of cracks’	1.41	ʃʃ^{j}el^j	‘crack, slot’
14.	$\sigma\sigma$	go.l^{j}e.n-á.st-ij	‘big-calved’	1.35	gól^{j}en^j$_{\text{STRESSED}}$	‘leg calf’
15.	$\sigma\sigma$	v.i.xr-á.st-ij	‘tufty, tousled’	1.29	v^{j}ix(ó)r, v^{j}ixr-í$_{\text{FINAL}}$	‘tuft’
16.	$\sigma\sigma$	go.lo.v-á.st-ij	‘big-headed, smart’	1.10	golov-á$_{\text{UNACC2}}$	‘head’
17.	σ	kli.k-á.st-ij	‘with big fangs’	1.04	klik	‘fang’

The lack of longer stems is not due to their overall paucity. Russian has plenty of polysyllabic body part nouns: *jagod^{j}íʦ-a* ‘buttock’, *p^{j}er^{j}enós^{j}iʦ-a* ‘bridge of the nose’, *podboródok* ‘chin’, *f^{j}iz^{j}ionóm^{j}ij-a* ‘face (derog)’. These occur in the Sharoff list, but do not form *-ast* adjectives (even in the larger Aranea corpus,

[3] Syllable count depends on how one handles vowel deletion (“yer rule”, Gouskova 2012 & refs therein). It applies inconsistently in the set of 17: *kor^j(e)n^j, l(o)b-* keep vowels, while *oʧ^j(o)k-* and *v^{j}ix(o)r-* lose them. (For oʧ^{j}k-í, the vowel appears only in the diminutive, [oʧ^{j}óʧ^j-k-i]). I count stems as disyllabic if they have two syllables anywhere in the inflectional paradigm, but the 2σ maximum holds even if UR vowels are counted.

described next). Moreover, many frequent monosyllabic body part nouns lack *-ast* adjectives: *l^{j}it͡s-ó* ‘face’, *ruk-á* ‘arm/hand’, and *nog-á* ‘leg’. This absence likely tips the learner off to the lexical selectivity of the suffix and prompts the creation of the sublexicon (Becker and Gouskova, 2016).

2.3 *The* -ast *suffix in the Aranea corpus.* To investigate the productivity of *-ast*, I extracted the 13,113 noun stems from Sharoff’s list and constructed *-ast* adjectives from them by script. For inconsistent alternations, more than one adjective was generated (e.g., with and without vowel deletion, with and without mutation of the stem-final consonant). The resulting 16,260 *-ast* adjectives were submitted as a lemma search in Aranea. This search netted 213 adjective lemmata, consistent with the low type frequency of the suffix. Quite a few of these were hapax legomena. Once again, for scale, a comparable search for 6,428 deadjectival nominalizations with *-ost^j* netted 4,679 lemmata. Thus, 73% of adjectives in Sharoff’s list have corresponding *-ost^j* nominalizations, but only 1.3% of nouns have corresponding *-ast* adjectives.

The prosodic profiles of three types of Russian nominal stems are summarized in Table 1 and graphically in Fig. 1. The leftmost column/plot shows the lengths of nouns in the Sharoff list, our reference lexicon. Of these, disyllabic stems are most frequent, with monosyllables being only the third most common type. The middle column summarizes the properties of the *-ast* stems in Sharoff, which is a model of the learners’ sublexicon. On the right are the productive uses in Aranea. There is a striking degree of frequency matching between the Sharoff list sublexicon and Aranea: monosyllables are ≈82%, disyllables ≈17.5%. There are two trisyllabic *-ast* adjectives in Aranea, discussed below.

	reference lexicon			*sublexicon*		*productive extension*		
	Sharoff list, all nouns			Sharoff *-ast* adj.		Aranea corpus		
0σ	vs̩-á ‘louse (colloq.)’	9		0	–		0	
σ	zúb ‘tooth’	2,211	17%	14	82.4%	s^{j}is^{j}k-ást-ij ‘boob’	174	81.7%
$\sigma\sigma$	golov-á ‘head’	4,859	37%	3	17.6%	b^{j}itseps-ást-ij ‘biceps’	37	17.4%
3σ	boj-ev^j-ík ‘action film/hero’	3,381	26%	0	—	bojev^{j}ik-ást-ij ‘action’	2	0.9%
4σ	p^{j}er^{j}e-nós^j-its-a ‘bridge o’nose’	1,823	14%	0	—	(*p^{j}er^{j}enos^{j}itsastij)	0	
5+	f^{j}iz^{j}ionóm^{j}ij-a ‘face, mug’	830	6%	0	—	(*f^{j}iz^{j}ionomi^{j}astij)	0	
		13,113	100%	17	100%		213	100%

Table 1: The Russian *-ast* suffix: overall lexical stem shapes vs. sublexical statistics, and frequency matching in the productive extension in corpus use

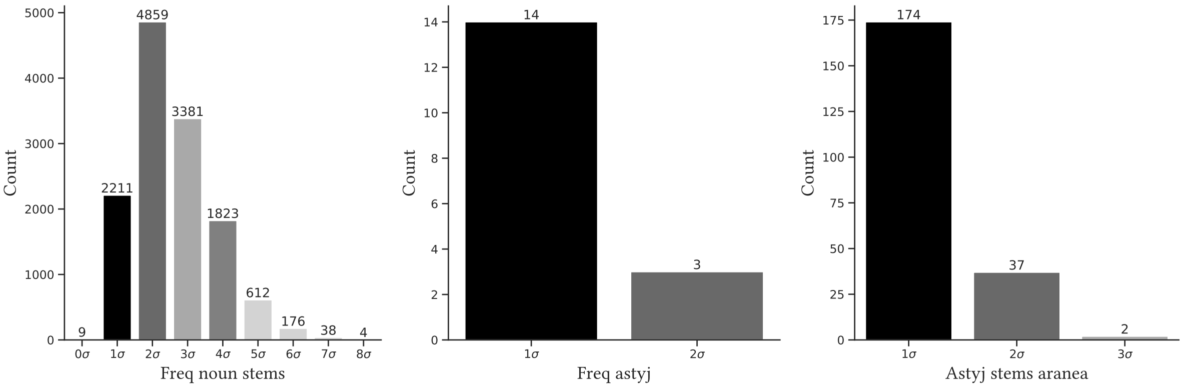

Figure 1: Stem sizes of nouns in the lexicon, in the sublexicon, and in the productive extension

Below are some examples that occurred in Aranea but not in the Sharoff list. Strikingly, there are a couple of trisyllabic stems in the productive extension of the *-ast* suffix: *bojev^{j}ik-ast-ij* ‘action-packed’ and *am^{j}erik-ast-ij* ‘American-looking’ (analogy to *s^{j}erp-ast-o-molotk-ast-ij* ‘hammer-and-sickled’, a compound of two *-ast* adjectives that appears in Mayakovsky’s poem about the Soviet passport).

Examples (6i–j) are significant because they are unexpected if phonological selection is encoded in a subcategorization frame. If the *-ast* suffix comes with a rule limiting syllable count to 2, these examples violate the rule. By contrast, if the generalization about syllable count is extracted statistically from a small sublexicon, it is unsurprising that some speakers are willing to go beyond two syllables; they may have treated the size generalization as an accident of the small sample.

(6) A sample of the 213 *-ast* adjectives in Araneum Russicum III Maximum

a.	bo.ro.d-á.st-ij	'big-bearded'	cf.	borod-á	'beard'
b.	mu.ʐi.k-á.st-ij	'mannish, peasant-like'		muʒík	'man, dude'
c.	ko.ʂel^j.k-á.st-ij	'moneyed'		koʂel^j-ók, . . .l^j-k-á	'money purse'
d.	or.d^{j}e.n-á.st-ij	'with lots of medals'		órd^{j}en	'medal'
e.	ʧ^{j}e.ʂu.j-á.st-ij	'with big scales (fish)'		ʧ^{j}eʂuj-á	'scale'
f.	b^{j}el^j.m-á.st-ij	'with cataracts'		b^{j}el^{j}m-ó	'cataract'
g.	xo.bo.t-á.st-ij	'big-trunked (elephant)'		xóbot	'trunk'
h.	ro.t-á.st-ij	'big-mouthed'		rót, rt-á	'mouth'
i.	bo.je.v^{j}i.k-á..st-ij	'action-packed'		boj-ev^j-ik	'action film/hero'
j.	a.m^{j}e.r^{j}i.k-á.st-ij	'American-looking'		am^{j}er^{j}ik-a	'America'

I next turn to stem-final consonants, which sometimes figure in phonological selection. Thus, English *-en* selects for obstruent-final stems (*short-en* vs. **tall-en*, Siegel 1974). The stem-final consonants found in Russian *-ast* stems are shown in Table 2:

a. Stem-final consonants in Sharoff list				b. Stem-final consonants in Aranea			
p **b** p^j b^j	**t** **d** t^j d^j ts	ʧ^j	**k** g k^j g^j	**p** **b** p^j b^j	**t** **d** **t^j** d^j **ts**	**ʧ^j**	**k** **g** k^j g^j
f **v** f^j v^j	s **z** s^j z^j	ʃʃ^j **ʂ** ʐ	x x^j	f **v** f^j v^j	**s** **z** **s^j** z^j	**ʃʃ^j** **ʂ** **ʐ**	**x** x^j
m m^j	**n** n^j			**m** m^j	**n** **n^j**		
	l **l^j** **r** r^j	j			**l** **l^j** **r** **r^j**	**j**	

Table 2: Stem-final consonants in Russian (plain font) vs. stem-final consonants in *-ast* adjectives (underlined): the Sharoff list vs. the Aranea corpus.

The smallest natural class that contains all the underlined Cs in 2a is [−syllabic]. Various smaller natural classes are missing from stem-final position in Table 2a: no velar fricatives, no labial nasals, no [−back] (palatalized) nasals or obstruents, etc. As shown in Table 2b, though, speakers do extend the suffix to those classes: in Aranea, we find 25 occurrences of *b^{j}el^{j}m-ast-ij* 'with cataracts'. This tells us that Russian speakers generalize coarsely: just because some natural classes are not represented in the sublexicon does not mean they are disallowed; the sample of [−syllabic] is sparse but treated as representative. The reason for this is explained in the next section.

2.4 *Monte Carlo Simulations.* To investigate the statistical support for size and segmental generalizations supplied by the *-ast* sublexicon, I ran Monte Carlo simulations. These could be seen as a model for learning the restrictions. Learners know from positive evidence that the *-ast* suffix occurs on only a small subset of stems they encounter on a regular basis; the one clear generalization about these stems is that they are morphosyntactically nouns. But learners do not a priori know what is limiting the productivity of this suffix: is it semantics[4] or phonology, and if it is phonology, which aspect is relevant? The search space of phonological generalizations is vast (see Hayes and Wilson 2008, Jarosz 2019, a.o.). By comparing the contents of the sublexicon to the contents of a morphosyntactically comparable reference lexicon, we can assess which generalizations can be ascribed to chance.

To explore the size restriction, a random set of 17 nominal stems was drawn 100,000 times from the ≈13,000 noun stems in Sharoff (2005), and their sizes were tracked. The likelihood of drawing 17 maximally disyllabic nominal stems is small but non-null; this happens 2-4 times on average. By contrast, in the simulation in Fig. 2, maximally trisyllabic stems were drawn 2,116 times, or 2.1%. The vast majority of random draws netted stems that were maximally 5 syllables:

[4] While semantics is not my main concern here, there is evidence that semantic generalization works similarly to phonological size generalization. In the sublexicon of 17 *-ast* adjectives, there are several non-body part uses ("full of cracks", "bespectacled", "flowery"). Body parts predominate in the big corpus, too, but there are some non-body part uses (e.g., "with lots of medals"). I do not discuss semantics further for lack of space.

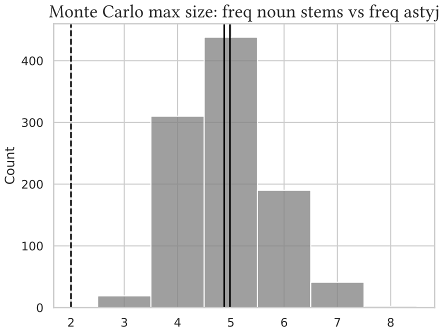

Figure 2: Max syllable counts in Monte Carlo simulations (100,000 draws of 17 stems from nouns in Sharoff's list). Solid lines=95% conf. intervals, dashed line = max size in *-ast* sublexicon.

Thus, there is a strong asymmetry between the sample of 17 *-ast* stems that learners see vs. what one might expect in a random sample of 17 noun stems of comparable frequency.

A different picture emerges for stem-final consonants. As explained above, the smallest natural class containing all the stem-final consonants in the *-ast* sublexicon is [−syllabic]. In a Monte Carlo simulation with 100,000 draws of 17 noun stems from the Sharoff list, stem-final consonants formed that same natural class 74% of the time; there were also draws including only stems ending in non-nasal segments, true (non-glide) consonants, velarized consonants, coronals, and so on (see Table 3). We can also examine which of the natural classes are missing in stem-final position both in the sublexicon and in randomly drawn samples of Russian stems; I do not discuss this in detail here but return to the question in section 4.1 on coarse generalizations.

Natural class	Draws	Natural class	Draws
[−syllabic]	74,269	[+coronal]	50
[−nasal]	14,157	[−continuant]	9
(all segments)	9302	[−sonorant]	4
[+consonantal]	1980	[+voice]	3
[+back]	225	[−voice]	1

Stem-final consonants in sublexicon			
p **b** pʲ bʲ	**t** **d** tʲ dʲ ts	ʧʲ	**k** g kʲ gʲ
f **v** fʲ vʲ	s **z** sʲ zʲ	ʃʃʲ **ʂ** ʐ	x xʲ
m mʲ	**n** nʲ		
	l **lʲ** **r** rʲ	j	

Table 3: Left: natural classes formed by the stem-final consonants in Monte Carlo simulations (100,000 draws of 17 noun stems from Sharoff's list). Right: Russian consonants that can occur in stem-final position (plain font) vs. in the *-ast* sublexicon (underlined).

The Monte Carlo exercise proves that the disyllabic size cap in the *-ast* sublexicon is strongly statistically supported when those stems are compared to the reference lexicon of noun stems, while the sample of stem-final segments is one that could arise simply by chance, since 74% of the stems in the lexicon end in consonants. Presumably, learners know both generalizations, since they routinely track all sorts of statistical information about the lexicon. This information is extracted as part of speakers' knowledge about the *-ast* suffix, and it is used when they decide how to extend it productively (see Gouskova et al. 2015; Becker and Gouskova 2016).

2.5 *Analysis.* I now turn to the formal statement of the rule governing the distribution of *-ast*. My rule for this suffix encodes the morphosyntactic context in which it occurs—namely, the sister is a categorized noun. The rule also includes a sublexicon list to indicate that it is lexically selective (I show in (13) that semantically and phonologically similar nouns of comparable frequency do not combine with *-ast*, so learners must track which nouns do).

(7) [a, INALIENABLE] $\leftrightarrow$ -ást$_{Dominant}$ / X]$_{\mathrm{N}}$ ____ Where X = { golov-, korʲenʲ-, zub-, glaz-,. . . }

I exclude any mention of phonological or semantic generalizations about the sublexicon; these are extracted separately via statistical learning. An easy way to encode the generalization about size is as a grid-based anti-lapse constraint (Selkirk 1984:52): assign a violation mark for every sequence of three unstressed syllables in a row (cf. Gordon 2002:502 on *EXTENDEDLAPSE). The traditional *LAPSE constraint ("no x x sequences") is weakly satisfied in the sublexicon (4 out of 13 stems violate it).

(8) *LAPSE3: Assign a violation for every x followed by x x.

The *LAPSE3 constraint has scope over the *-ast* sublexicon, where stem syllables are necessarily unstressed (recall 4), and can therefore track the size of stems as a phonotactic generalization. Following Hayes and Wilson (2008), I assume that the metrical grid level of representation is available to learners, and that learners can induce grid-based constraints longer than the trigram window with relative computational ease, compared to segmental/natural class-based trigram constraints.

(9) Sublexical phonotactic generalization: no x x x.

		Metrical grid	*LAPSE3	*LAPSE
a.	*σ-σ́ . . .*	$\underline{\text{x}}$ - X . . .	✓	✓
b.	*σσ-σ́ . . .*	$\underline{\text{x x}}$ - X . . .	✓	*
c.	*σσσ-σ́ . . .*	$\underline{\text{x x x}}$ - X . . .	*	**
d.	*σσσσ-σ́ . . .*	$\underline{\text{x x x x}}$ - X . . .	**	***

While *LAPSE3 holds of stems that combine with *-ast*, it does not apply generally in Russian, which is demonstrated in the next section.

3 No Size Limit for *-ist* and *-izm*

I chose *-ist* and *-izm* because they are similar to *-ast* in stress properties.[5] The main reason to look at them is to demonstrate that the statistical distributions of stem properties for other suffixes are different in a non-trivial way. As shown in (10), both suffixes allow long lapses in their stems—as is generally allowed in Russian (recall (3)). These two suffixes are also quite productive, just as their English counterparts. A good number of *-ist* and *-izm* stems are free-standing nouns, just as for *-ast*, but some are bound stems or adjectives:

(10) Long initial lapses with stress on the suffix: *-ist, -izm*

a.	-ist	*σσσ-σ́*	rʲe.t͡si.dʲi.vʲ-íst	'recidivist'	rʲe.t͡si.dʲív	'recidivist act'
		σσσσ-σ́	av.to.mo.bʲi.lʲ-íst	'motorist'	avtomobílʲ	'automobile'
		σσσσσσ-σ́	in.ter.na.t͡si.o.na.lʲ-íst	'cosmopolitan'	in.ter.na.t͡si.o.nál	'Internationale'
b.	-izm	*σσ-σ́*	bolʲ.ʂ̪e.vʲ-ízm	'Bolshevism'	bolʲ.ʂ̪e.vʲ-ík	'Bolshevik'
		σσσ-σ́	al.ko.go.lʲ-ízm	'alcoholism'	al.ko.gólʲ	'alcohol'
		σσσσ-σ́	an.tʲi.sʲe.mʲi.tʲ-ízm	'anti-Semitism'	an.tʲi.sʲe.mʲít	'anti-Semite'
		σσσσ-σ́	to.ta.lʲi.ta.r-ízm	'totalitarianism'	to.ta.lʲi.tár.nij	'totalitarian'
		σσσσσ-σ́	pro.fʲe.sʲi.o.na.lʲ-ízm	'professionalism'	pro.fʲe.sʲi.o.nál	'professional'

I do not present an Aranea corpus study due to space limitations, but the differences between *-ast* and these two suffixes are evident from examining Sharoff's list alone. There are 82 *-izm* noun lemmata and 125 *-ist* ones in Sharoff. A comparison between the reference lexicon of all noun (or adjective) stems in Sharoff vs. the *-izm* and *-ist* stems shows straightforward frequency matching for stem size; the relative prevalence of stems of various lengths is proportionally very similar (see Figure 3).

I ran a Monte Carlo simulation with 100,000 random draws of 82 and 125 noun stems from the Sharoff list. Given the similarities in phonological size between the stems in the *-izm/-ist* sets and the reference lexicon of noun stems, it is unsurprising that random draws net the same types of stems close to (or exactly) half

[5] Intuitively, having a predictable effect on stress ought to make a suffix easier to extend to novel stems, but the relationship between stress dominance and productivity is not a straightforward one in Russian: thus, *-ostʲ,* a recessive unaccented suffix, is inconsistent in its effects on stem stress (see Zaliznjak 1985) but is quite productive, with 589 lemmata in the Sharoff list and 4,679 lemmata derived from frequent adjectives in Aranea. There are, however, dozens of unproductive stress-dominant suffixes in Russian (see Shvedova 1980).

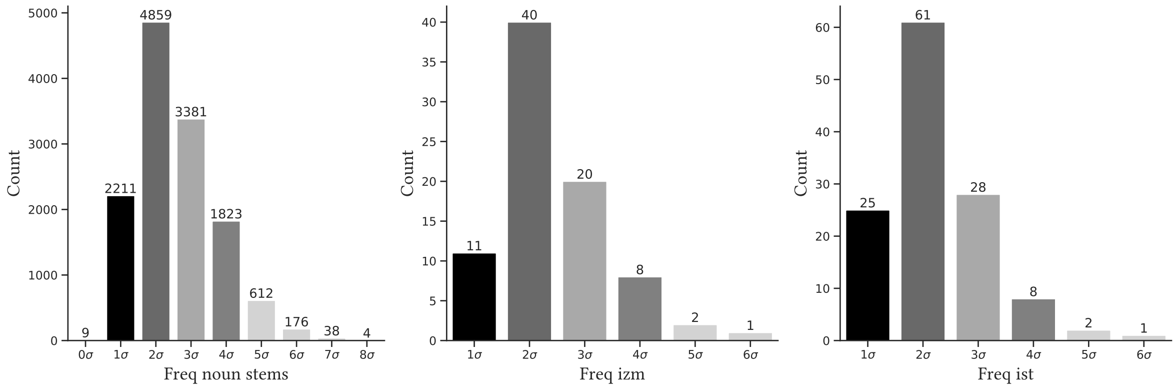

Figure 3: Noun stems in Sharoff's list vs. *-izm* and *-ist* stems in Sharoff

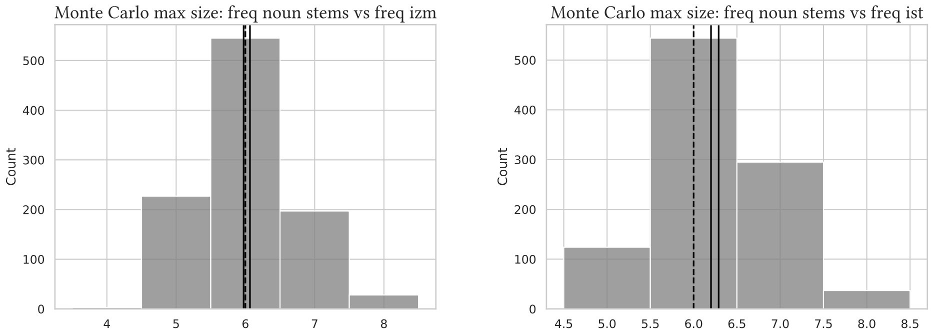

Figure 4: Monte Carlo simulation plots: drawing 82 and 125 stems respectively from 13K noun stems. Solid lines show 95% confidence intervals for maximum syllable count across draws; dashed lines show maximum syllable count in the sublexicons.

the time (see Table 4). As for segmental generalizations, these suffixes are even more permissive than *-ast*: they attach to both C- and V-final stems (e.g., [al^{j}.tru.ízm] 'altruism'). The smallest class that includes all the segments in stem-final position for each suffix is all segments; this is also true of 57% draws for *-ist* and 43% draws for *-izm*.

-izm: draws of 82	100,000
same inventory:	47,898 (47%)
same 6-syll size cap:	51,708 (51%)

-ist: draws of 125	100,000
same inventory:	50,268 (50%)
same 6-syll size cap:	54,889 (54%)

Table 4: Monte Carlo simulations for *-izm* and *-ist* sublexicons

A statistical learner ought to conclude that there are no useful generalizations to track as far as the sizes of stems that combine with each suffix, since the sets could have arisen by chance. The rules for such suffixes are simple: affixation is constrained only by the requirement that the resulting word be semantically interpretable in an appropriate way.

(11) n, PERSON ↔ *-íst*$_{Dominant}$

(12) n, PROPERTY ↔ *-ízm*$_{Dominant}$

This brief examination of additional suffixes supports the conclusion that the two-syllable size restriction of *-ast* can be learned by a statistical comparison between its sublexicon and a morphosyntactically similar reference lexicon. The comparison detected an asymmetry in the case of *-ast*, but not in the case of *-ist/-izm.* It would be disturbing indeed if every affix supplied evidence of random size restrictions on its stems, or if the lexical statistics allowed a sample of stems under some minimal phonological size to be drawn by accident. The study of *-ist/-izm* acts as a sanity check: these suffixes are indeed unselective.

4 Discussion

4.1 *Coarse Generalizations.* The literature on affixal selectional restrictions is replete with examples of selection for stress location, syllable count, and C/V at stem edges, but selection for subsegmental natural classes (e.g., labials or obstruents) is less common (though certainly not unheard of; see Paster 2006; Nevins 2011; Gouskova et al. 2015). From the point of view of a statistical learner, this is unsurprising: every stem encountered with the affix is informative as to its syllable count. We have seen that Russian speakers treat the two-syllable maximum seen in common *-ast* adjectives as a productive restriction; they moreover match the relative frequency of mono- and disyllables when they attach *-ast* onto novel stems. Even though the suffix's sublexicon is small, the restriction is noticeable given the usual lengths of Russian noun stems.

Unlike syllable count, subsegmental generalizations involve a larger space of analytical possibilities, both for learners and linguists. Recall that Russian speakers generalize *-ast* to stems ending in segments they do not necessarily encounter in the learning data (such as palatalized obstruents; see Table 2). They see a subset of [-syllabic], and they assume that other [-syllabic] stem-final consonants are allowed, too. How can they be so confident?

Scheer (2016) answers this question with a modularity stipulation: morphological rules (such as those governing suppletion or selectional restrictions) can access only C/V features and natural classes relevant to sonority, but they cannot see the melodic tier. In this theory, it would be impossible for a learner to formalize a selectional restriction that prevents a suffix from attaching to, say, stem-final palatalized obstruents or palatalized labials. There is a better motivated answer, though, which relies on quantiative reasoning. The more fine-grained the generalization, the more statistical support the learner needs to conclude that it is not an accidental gap. My study has identified a method for identifying accidental gaps: sublexical comparison. The likelihood of not encountering certain classes of segments stem-finally can be estimated by looking at how often they occur outside the sublexicon. Fine-grained generalizations on the melodic tier involve smaller natural classes, by definition.

Let's work through one specific example, palatalized obstruents, [-back, -son]—which Russian speakers apparently treat as an accidental gap for *-ast*. There are 12 segments in this class: [p^{j}, b^{j}, f^{j}, t^{j}, d^{j}, s^{j}, z^{j}, ʃʃ^{j}, ʧ^{j}, k^{j}, ɡ^{j}, x^{j}]. By contrast, Russian has 21 [+son] segments, 25 [-son] obstruents, and 36 [-syllabic] consonants. But what matters for selection is not the number of segments in each class but how often they occur stem-finally in the relevant portion of the Russian lexicon. The likelihood of drawing 17 random nominal stems and have none of them end in palatalized obstruents is about 26%; basically chance. Since the lack of palatalized obstruents in the learning data could be due to chance, learners do not assume there is a prohibition against *-ast* attaching to such stems. Forms such as [s^{j}is^{j}-ast-ij] 'big-boobed' and [koɡt^{j}-ast-ij] 'big-clawed' testify to this.

Thus, the reason Russian speakers generalize beyond the segments they see in stem-final position is not that they lack the grammatical means for encoding generalizations at the segmental level. Rather, it is because the lexicon supplies quantitative evidence for which gaps to ignore. Occamite reasoning would suggest that, given such evidence, bolder claims about the visibility of the melodic tier to the morphosyntactic module are not justified.

4.2 *What to Count? Frequency Matching Vs. Tolerance Principle* My corpus study of *-ast* found frequency matching: the frequencies of mono- and disyllables matched between the Sharoff list sublexicon and the productive extension in the Aranea corpus, but not the nominal stems. We also saw that, even though only 14 out of 17 stems in the Sharoff list refer to body parts, speakers extend the suffix to broader meanings: 'medal', 'money bag', 'shoe', etc.

Frequency matching of the kind seen with *-ast* is a challenge for tolerance theory (Yang 2016). Tolerance envisions the grammar in terms of good/bad judgments. Rules can be productive only if they pass the tolerance threshold: a rule with N eligible undergoers cannot be productive if it has more than N/ln(N) exceptions. Thus, a productive rule with 100 eligible undergoers can tolerate ≈22 exceptions, but not 25. Yang dismisses frequency matching as a task effect: "… experimental psychologists have long known that categorical tasks are likely to elicit categorical responses, and gradient tasks such as rating are likely to elicit, alas, gradient results" (Yang 2016:39).

But task effects cannot explain my results. The corpus of web texts represents spontaneous productive usage, not gradient responses on a Likert scale. Thus, frequency matching is a fact in need of an explanation. Explanations of frequency matching generally attribute it to the grammar itself being probabilistic (Boersma and Hayes 2001 et seq.): if the grammar produces probabilities rather than good/bad judgments, then speakers can mirror the probabilities in their own productions. This is the assumption in my theory, as well: sublexicons are characterized by miniature probabilistic grammars, with constraints that have sublexicon-specific weights (see Gouskova et al. 2015; Becker and Gouskova 2016). The *LAPSE constraints in (8) penalize two- and three-syllable stress lapses. Two-syllable lapses are sparsely attested in the learning data, and three-syllable lapses are unattested. Segmental lacunae in the sublexicon appear to be accidental gaps. I assume that speakers use this knowledge to evaluate potential novel stems when generalizing *-ast*.

The problem with tolerance theory is that it envisions the search for rules as a successive splitting of the hypothesis space, but it is not clear why the learner would contemplate syllable-size generalizations after semantic ones are found, or where segmental features fit in. What gets counted varies depending on the case. In Yang's discussion of the English past tense, the morphosyntax/semantics determine the scope. The rule is "form past tense by adding /-d/", so *ring/rang* and *hit/hit* are exceptions. But for English nominalizations, the scope is defined as containing the phonological substring. The rule for *-ment* is regular when the result is semantically transparent (e.g., *state/state-ment* and *compart-ment*), while exceptions are either non-affixed (*cement*) or semantically non-transparent (*department*). Thus, the *-ment* affix is deemed productive because it has 40 transparent examples in a child-directed corpus, while the exceptions are 5 *ment*-final words are of the *cement/department* ilk (the threshold for N=45 is 12 exceptions). Thus, Yang identifies the semantic transparency generalization for *-ment*, but he also notes that recent examples of *-ment* nouns are longer than a syllable (just 6 monosyllables like *pave-ment*). The implication is that monosyllabicity is the exception rather than the rule (2016:116), and should not be productive (since there are 45 stems with *-ment* but 39 of them do not follow the monosyllable generalization).

While it is usually possible to find some way of counting undergoers/exceptions that would allow the tolerance calculation to succeed, it is unclear why one should consider phonological generalizations if semantic ones are found, or vice versa. Looking at Russian *-ast* along the lines of Yang's examination of *-ment*, the rate of monosyllables (13) vs. disyllables (4) is consistent with disyllables being the exception under the tolerance equation (the exception threshold for N=17 is 6). Similar calculations over semantics should identify body parts as being the productive category (14 vs. 3 non-body part uses). Under either calculation, either phonology should be ignored or semantics. This theory does not predict productive extensions to disyllabic, non-meronymic stems such as *koʂelʲkastij* 'with money bags'. It also has no explanation for frequency matching.

At a deeper level, tolerance fails to provide an explicit characterization of how selection works. A tolerance learner flits between generalizations that sometimes end up defining the eligible pool of undergoers and other times end up defining the rule itself. By contrast, in sublexical theory, the learner is capable of considering generalizations at all levels—semantic, stress-based, syllable-based, segmental. Whenever the generalizations are non-accidental, they should be extended in proportion to the statistical support for them, which is what we saw in Russian. Selection isn't just semantic or phonological; in reality, it often involves multiple unrelated generalizations that apply simultaneously.

4.3 *Subcategorization Frames.* Some of my criticisms of modular separation and tolerance theories can also be applied to subcategorization frames. Frames often reify a single level of generalization that lends itself to a simple formal treatment but neglect other, equally valid correlates of an affix's productivity. For example, the English comparative/superlative *-er/-est* suffixes impose a disyllabic maximum on stems (e.g., **beautiful-er*), with a preference for initial stress that has been equated with the moraic trochee (McCarthy and Prince 1986). But the *-er* suffix also comes with segmental generalizations, which kick in only in disyllables (e.g., *blacker* vs. **lilac-er*). Such restrictions are often noted but rarely formalized in a satisfactory way; in the case of disyllables ending in obstruents, the condition is negative (no obstruent codas in stem-final unstressed syllables), which does not fit into the subcategorization formalism neatly (see Gouskova and Ahn 2016). The *-er* suffix also fails to attach to many phonologically conforming stems, which are apparently lexical exceptions (cf. **iller* vs. *fuller*). Lexical exceptions abound in the case of Russian *-ast*: many frequent body part nouns have no corresponding *-ast* adjectives either in Sharoff or in Aranea (and which I would judge as ungrammatical). Presumably, learners have to list lexically the stems that do and do not occur with *-ast*:

(13) Frequent body part nouns that do not occur with *-ast* in Sharoff's list or Aranea

	Gloss	Occ p/m in Sharoff	*-ast* forms (don't exist!)	
l^{j}iʦ-ó	'face'	915.59	l^{j}itsastij	
gólos	'voice'	632.06	golosastij	cf. golos^{j}istij
pl^{j}eʧ^j-ó	'shoulder'	385.14	pl^{j}eʧ^{j}astij	cf. pl^{j}eʧ^{j}istij
t^{j}él-o	'body'	319.64	t^{j}elastij	
s^{j}érʦ-e	'heart'	291.79	s^{j}ertsastij	
krov^j	'blood'	258.31	krov^{j}astij	
kol^{j}én-o	'knee'	162.45	kol^{j}enastij	
lókot^j	'elbow'	63.35	lokt^{j}astij, lokot^{j}astij, etc.	
b^{j}elók	'white of the eye'	37.01	b^{j}elkastij	
r^{j}esn^{j}íʦ-a	'eyelash'	27.30	r^{j}esn^{j}itsastij	

Also problematic is the generalization that mono- and disyllabic sequences of unstressed syllables do not form a constituent (unlike the trochees of English), which again suggests a constraint against *-ast* attaching to something, rather than a positive condition on what it attaches to.

The biggest challenge is that frames fail to be predictive. English verbalizing *-en* selects for obstruent-final monosyllables (*black-en* vs. **blue-en*) but fails to attach to numerous compliant adjectives (**voiden*, **snuggen*). This damps the appeal of phonological subcategorization frames: if they capture a predictive generalization, as linguistic rules are supposed to do, why do stems respecting the generalization fail to follow the rule? Finally, this theory does not have anything to say about frequency matching, or the occasional extensions beyond the subcategorization frame limit; these all figure in the Russian case study.

References

Alderete, John. 1999. Morphologically-Governed Accent in Optimality Theory. Doctoral Dissertation, University of Massachusetts, Amherst, Amherst, MA.

Baayen, Harald, and Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 801–843.

Becker, Michael, and Maria Gouskova. 2016. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry* 47:391–425.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32:45–86.

Crosswhite, Katherine. 1999. Vowel reduction in Optimality Theory. Doctoral Dissertation, UCLA, Los Angeles, CA.

Embick, David. 2010. *Localism versus globalism in morphology and phonology*. Cambridge, MA: MIT Press.

Gordon, Matthew. 2002. A factorial typology of quantity-insensitive stress. *Natural Language & Linguistic Theory* 20:491–552.

Gouskova, Maria. 2010. The phonology of boundaries and secondary stress in Russian compounds. *The Linguistic Review* 17:387–448.

Gouskova, Maria. 2012. Unexceptional segments. *Natural Language and Linguistic Theory* 30:79–133.

Gouskova, Maria, and Suzy Ahn. 2016. Sublexical phonotactics and English comparatives. Ms.

Gouskova, Maria, Sofya Kasyanenko, and Luiza Newlin-Łukowicz. 2015. Selectional restrictions as phonotactics over sublexicons. *Lingua* 167:41–81. URL `http://ling.auf.net/lingbuzz/002673`.

Halle, Morris. 1973. The accentuation of Russian words. *Language* 49:312–348.

Hayes, Bruce, and Colin Wilson. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39:379–440.

Jarosz, Gaja. 2019. Computational modeling of phonological learning. *Annual Review of Linguistics* 5:67–90. URL `https://doi.org/10.1146/annurev-linguistics-011718-011832`.

Kalin, Laura, and Nicholas Rolle. 2023. Deconstructing subcategorization: Conditions on insertion versus conditions on position. *Linguistic Inquiry* .

Lieber, Rochelle. 1980. On the organization of the lexicon. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Mascaró, Joan. 1996. External allomorphy as emergence of the unmarked. In *Current Trends in Phonology: Models and Methods*, ed. Jacques Durand and Bernard Laks, 473–483. Salford, Manchester: European Studies Research Institute, University of Salford.

McCarthy, John J., and Alan Prince. 1986. Prosodic Morphology. Technical Report #32, Rutgers University Center for Cognitive Science, version of 1996.

McCarthy, John J., and Alan Prince. 1994. The emergence of the unmarked: Optimality in prosodic morphology. In *Proceedings of the North East Linguistic Society 24*, ed. Merce Gonzalez, 333–379. Amherst, MA: GLSA Publications.

Melvold, Janis. 1989. Structure and stress in the phonology of Russian. Doctoral Dissertation, MIT, Cambridge, MA.

Nevins, Andrew. 2011. Phonologically-conditioned allomorph selection. In *Blackwell companion to phonology*, ed. Colin Ewen, Elizabeth Hume, Marc van Oostendorp, and Keren Rice, 2357–2382. Wiley-Blackwell.

Orgun, C. Orhan, and Ronald Sprouse. 1999. From MParse to control: Deriving ungrammaticality. *Phonology* 16:191–220.

Paster, Mary. 2006. Phonological Conditions on Affixation. Doctoral Dissertation, UC Berkeley, Berkeley, CA.

Pierrehumbert, Janet. 2001. Why phonological constraints are so coarse-grained. *Language and cognitive processes* 16:691–698.

Revithiadou, Anthi. 1999. *Headmost Accent Wins: Head Dominance and Ideal Prosodic Form in Lexical Accent Systems*. The Hague: Holland Academic Graphics.

Scheer, Tobias. 2016. Melody-free syntax and phonologically conditioned allomorphy. *Morphology* 26:341–378.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Sharoff, Serge. 2005. Methods and tools for development of the Russian Reference Corpus. In *Corpus Linguistics Around the World*, ed. Andrew Wilson, Dawn Archer, and Paul Rayson, Language and Computers, 167–180. Amsterdam and New York: Rodopi. `http://www.comp.leeds.ac.uk/ssharoff/frqlist/frqlist-en.html`.

Shvedova, Natalia Y. 1980. *Russkaja grammatika [Russian grammar]*. Moscow: Academy of Sciences USSR.

Siegel, Dorothy. 1974. Topics in English Morphology. Doctoral Dissertation, MIT, Cambridge, MA.

Yang, Charles. 2016. *The price of linguistic productivity*. Cambridge: MIT Press.

Zaliznjak, Andrej A. 1985. *Ot praslavjanskoj akcentuacii k russkoj. [From Proto-Slavic to Russian accentuation.]*. Moscow: Nauka.