Sonority Sequencing in Polish: the Combined Roles of Prior Bias & Experience

Gaja Jarosz and Amanda Rysling
University of Massachusetts Amherst

1 Introduction

It is well known that speakers' behavior in phonotactic and phonological behavioral tasks, such as acceptability judgments (Coleman & Pierrehumbert 1997; Vitevitch *et al*. 1997; Albright 2009; Hayes & Wilson 2008) or wug tests (Hayes & Londe 2006; Ernestus & Baayen 2003; Becker, Nevins & Levine 2012), is sensitive to the statistical patterns in the ambient language. At the same time, there is evidence that phonological learning may be constrained by biases: not all robust statistical patterns are learned equally well (Hayes & White 2013; Hayes *et al*. 2009; Becker, Ketrez & Nevins 2011; Becker, Nevins & Levine 2012; Wilson 2006). The existing evidence is therefore consistent with a combined influence of statistical learning and bias, but the exact nature of phonological learning biases and how they may affect the generalizations learners extract from the language input remains unclear.

The Sonority Sequencing Principle (Clements 1990; Selkirk 1984) figures prominently in recent debates concerning the sources of phonotactic and phonological knowledge. According to the Sonority Sequencing Principle (SSP), languages prefer syllables whose segments rise in sonority toward the syllable nucleus and fall in sonority away from the nucleus into the coda. The SSP is an ideal domain for examining the interactions of learning biases and statistical learning: the typological generalizations concerning relative well-formedness are well documented, and there are computational models available to examine the predictions of (unbiased) statistical learning (Hayes & Wilson 2008; Albright 2009). If learning is strongly constrained by a bias encoding these preferences, which could either be innate or derived from phonetic experience (Hayes 1999), then the predictions for learning are relatively clear: learners should exhibit preferences consistent with the SSP regardless of their language experience. If, on the other hand, learners' phonological knowledge is entirely derivable from the statistical patterns in the language input, then differences in language experience should drive differences in acquired preferences. Intermediate positions on the role of SSP are also possible, where sonority sequencing preferences are shaped both by prior bias and experience to some extent, though explicit implementations of how exactly these factors may interact and what predictions are made for particular languages are lacking.

Existing behavioral investigations of SSP sensitivity focus on poverty of the stimulus cases, examining sonority sequencing preferences in languages that have a restricted inventory of initial consonant clusters. The findings demonstrate cross-linguistic sensitivity to the SSP. Specifically, native speakers of English, Mandarin, and Korean exhibit gradient sonority sequencing preferences among illicit initial clusters, and these preferences mirror the SSP (Daland *et al.* 2011; Berent *et al*. 2007; Berent *et al.* 2008; Ren, Gao & Morgan 2010). These studies demonstrate that speakers have preferences among sequences they never experienced in their language input. For example, English speakers prefer words beginning with a sequence like #bn to words beginning with a sequence like #bd, even though neither sequence occurs in any words of English. Some argue that these results can only be explained by an innate bias (Berent *et al*. 2007; Berent *et al*. 2008; Ren, Gao & Morgan 2010). In support of this conclusion, Berent *et al.* (2007) show that basic lexical statistics and analogical models do not predict SSP preferences based on English statistics.

However, recent modeling studies question this conclusion, showing that some computational models can in fact detect SSP preferences from the lexical statistics in these languages given the right

* We are grateful to Adam Albright, Michael Becker, Iris Berent, Robert Daland, Brian Dillon, Gillian Gallagher, Maria Gouskova, John Kingston, Joe Pater, Colin Wilson, audiences at NELS 2016 and AMP 2016, and the UMass Amherst Sound Workshop and Experimental Labs Meetings for valuable feedback on portions of this work.

Proceedings of *AMP 2016*

representational abilities (Daland *et al*. 2011; Hayes 2011). To succeed, the models must have the ability to generalize on the basis of phonological features and natural classes, and they must be capable of expressing phonological contexts, such as syllable position. Daland *et al*. found that several models with these capabilities are sufficient for deriving sonority sequencing preferences based on English lexical statistics. They found the UCLA phonotactic learner (Hayes & Wilson 2008) to be particularly successful in capturing acceptability scores on unattested clusters. By generalizing broadly on the basis of existing syllables, these models can learn to favor syllables that are phonologically most similar to the predominant syllable shapes in the language input. These models are able to succeed in deriving sonority sequencing preferences because English is overwhelmingly characterized by syllables that begin low in sonority and rise into the nucleus, and the models are capable of representing this fact. English words often begin with large rises like #bl, rarely with plateaus like #st, and never with falls like #lb. The models utilize phonological similarity to favor clusters with similar profiles (Daland *et al*. 2011).

For languages like Mandarin or Korean, which arguably lack onset clusters altogether, Hayes (2011) has argued that equipping such models with constraints regulating sonority feature combinations may be sufficient to favor initial sonority rises. The set of sonority regulating constraints proposed by Hayes includes 32 bigram constraints on combinations of sonority levels defined by the features ±sonorant, ±approximant, ±syllabic, and ±consonantal, such as *[+son][-son] and *[-son][+son]. Crucially, the constraint set makes it possible to penalize either falls or rises in principle, depending on the input, and so does not inherently encode universal sonority sequencing preferences. Hayes shows that weighting the constraints of this minimal UG on the basis of toy Mandarin / Korean data can yield sonority projection. The model learns to favor words with rising sonority by generalizing from its substantial experience with initial #CV transitions, which share many features with sonority rising #CC transitions. These findings weaken the argument that only an innate bias can explain sonority projection: phonologically sophisticated statistical learning models can also derive these preferences from the input without building in the SSP a priori. Indeed, these predictions may be expected to hold for all languages with highly restricted cluster inventories since in these languages, words overwhelmingly begin with sonority rises of some kind.

Jarosz (to appear) tackles this conundrum from a different perspective: she examines the evidence for the SSP bias in a language, Polish, that utilizes the entire SSP scale. In Polish, there is no poverty of the stimulus: Polish speakers experience the entire scale. Jarosz shows that Polish nonetheless presents a particularly good test case for differentiating these competing hypotheses because the lexical statistics of Polish actually *contradict* the SSP, at least in part. This is due primarily to the fact that sonority plateaus make up nearly half of all initial clusters (Table 1), which means plateaus are actually the most preferred initial clusters from a statistical perspective. Jarosz presents evidence from phonological development that children acquiring Polish do in fact exhibit sensitivity to the SSP: in spontaneous productions, children produce onset clusters with larger sonority rises more accurately than those with smaller rises. These preferences run counter to the predictions based on input statistics. Jarosz shows that the same phonotactic models that succeeded in projecting SSP for English, Mandarin, and Korean do not predict that higher rises should be preferred to lower rises and plateaus for Polish. Thus, this recent finding contributes further support for a cross-linguistic SSP preference; this time, however, the observed preference is found early in development and cannot be derived from the input statistics.

Sonority Rise	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
Frequency	0.1%	0.2%	0.1%	45.3%	6.4%	28.0%	19.9%

Table 1 - Frequencies of Sonority Rises in Polish Initial Clusters from Jarosz (to appear)

One interpretation of the existing results is that there is a universal SSP bias that guides phonological learning, encoding an inviolable SSP scale and making it impossible for learners to diffuse or reverse these preferences, regardless of experience. However, since the only evidence for the SSP bias in Polish so far comes from young children's spontaneous productions, other interpretations are also possible. It could be that children's phonotactic knowledge differs from that of adults, who have had more persistent exposure to the lexical statistics of Polish. Indeed, if learners are equipped with a universal SSP bias, it is expected, under various common assumptions about how biases are formalized, that the bias would have greatest effects early in learning and its effects would diminish as more and more language data are encountered.

To begin to unravel some of these competing interpretations, the present paper examines the phonotactic knowledge of adult Polish speakers. We report the results of an online acceptability judgment experiment focusing on initial clusters in Polish. We examine both attested and unattested clusters to determine whether speakers' preferences are generalized by sonority to novel clusters and to determine whether any observed sonority sequencing preferences are modulated by direct experience with existing clusters. We also present the results of computational simulations evaluating the ability of phonotactic models to predict participants' ratings on the basis of the lexical statistics of Polish. Our main findings are that 1) SSP is predictive of adults' ratings, 2) sonority projection arises in both attested and unattested clusters, 3) while phonotactic models have significant predictive value, they do not subsume the SSP preferences observed in the participants' ratings, and 4) participants' sonority sequencing preferences are not entirely compatible with the SSP, suggesting a combined effect of prior bias and experience.

2 Phonotactic Judgment Experiment

This section describes the methods and results of the online phonotactic judgment experiment with adult Polish native speakers investigating sensitivity to the SSP in attested and unattested clusters.

2.1 *Methods*

2.1.1 *Materials* Test items were constructed by systematically concatenating biconsonantal initial clusters (heads) with tri-syllabic vowel-initial strings (tails) to form nonce words of Polish. 53 heads across five levels (-2 to 2) of sonority rise were chosen. 28 of these are attested initial clusters in Polish (Table 2), and 25 with comparable sonority profiles are unattested (Table 3). Within each combination of attestedness and sonority rise, heads were selected to utilize as broad a range of phonological properties and frequency ranges as possible. This was done both to minimize any accidental phonotactic asymmetries across conditions and to introduce sufficient variability within conditions to get a reliable estimate of how any potential effects of sonority transcend the variability observed within conditions.

	-3/-2	-1	0	+1	+2/+3
High token			pʂ (psz)	sm (sm)	mw (mł)
High type			ʂp (szp)	mr (mr)	pw (pł)
High token			gʥ (gdzi)	dɲ (dni)	ʑl (źl)
Low type			mn (mn)	xm (chm)	ʧw (czł)
Low token			ʂf (szw)	gn (gn)	xr (chr)
High type			gv (gw)	sn (sn)	gl (gl)
Low token	**rʐ (rż)**	mʐ (mż)	ʐv (żw)	ʐm (żm)	ʐr (żr)
Low type	**wʐ (łż)**	lɲ (lni)	ʧk (czk)	lj (lj)	ʐw (żł)

Table 2 - Attested Heads in IPA (orthography in parentheses) by Sonority Rise

	-3/-2	-1	0	+1	+2/+3
Fricatives	**jf (jf)**	ɲv (ńw)	xɕ (chsi)	ʑm (źm)	ʑw (źł)
	lʑ (lzi)	mʑ (mzi)	ɕx (ści)	fn (fn)	ʐj (żj)
Stops	**jʥ (jdz)**	np (np)	kʧ (kcz)	ʤm (dżm)	ʧl (czl)
	lʧ (lcz)	mʥ (mdzi)	bg (bg)	ʥɲ (dzni)	ʥj (dzj)
Sonorants	**wm (łm)**	wr (łr)	nm (nm)	rw (rł)	nw (nł)

Table 3 - Unattested Heads in IPA (orthography in parentheses) by Sonority Rise

We sought to utilize a broad range of segmental material and place, manner, voicing combinations in each condition. In light of these considerations, we decided to collapse clusters with a profile of -3 and 3 with clusters with profile of -2 and 2, respectively, so that clusters at the end-points of the scale, which crucially rely on a limited inventory of liquids and glides, could also have sufficient internal variability. For unattested clusters, we systematically balanced the conditions to ensure each sonority rise condition was

represented by two clusters involving fricatives, two involving stops or affricates, and one involving only sonorants. This ensures there are no major manner asymmetries across conditions, other than those specifically encoded in the SSP. To make sure relative frequency of attested clusters did not create accidental confounds with sonority, we systematically varied frequency within sonority rise conditions as well. Where possible, we selected clusters that varied in having both high and low estimates for both type and token frequency[1]. This variety was not possible for sonority falls because all initial falls have low type and token frequency in Polish. The statistical analyses in the next section address the resulting asymmetry.

For the purposes of stimuli generation, heads were sorted into ten groups, each of which contained either five or six clusters (Table 4). Head group membership was chosen so as to distribute the following properties as evenly as possible across all groups: unattestedness, attestedness, and within this, high versus low type and token frequency, sonority sequencing profiles, places, and manners of articulation.

1	2	3	4	5	6	7	8	9	10
jf	jʤ	wr	lɲ	lʑ	wm	mʑ	mʤ	ɲv	np
xɕ	ktʃ	ʤj	nm	fn	bg	nw	wzˌ	tʃw	rw
ʤm	ʑm	mzˌ	tʃl	mn	ʤɲ	gʤ	pʂ	zˌm	zj
ʂp	ʂf	sn	rzˌ	xr	mw	zˌv	xm	ɕx	tʃk
gn	dɲ	ʐl	pw	sm	lj	ltʃ	ʐw	gv	gl
zˌr	zˌw								mr

Table 4 – Head Groups

Thirty tri-syllabic VCVC(C)V(C) tails were created for pairing with heads to make complete test words. These tails were controlled so as to avoid near phonological neighbors and major phonotactic violations. Tails were selected so that they would not create near phonological neighbors with any of the heads. 10 tails ended in unambiguously nominal morphology, 10 ended in unambiguously adjectival morphology, and 10 ended in unambiguously verbal morphology (Table 5). The first vowel of every tail was always one of /a/ (4 nominal, 4 verbal, 3 adjectival), /o/ (3 nominal, 3 verbal, 4 adjectival), or /u/ (4 nominal, 3 verbal, 3 adjectival), because /a/, /o/ and /u/ can follow any consonant of Polish (unlike /i/, /ɛ/, and /ɨ/, which are restricted). These tails were also sorted into ten groups, each of which contained one nominal, one adjectival, and one verbal tail. To the extent possible, tail-initial vowels and tail-internal consonants were distributed as evenly as possible across all groups.

Group	1	2	3	4	5	6	7	8	9	10
Nominal	axatsjɛ	aʂoɕtɕik	awutʃɛk	arupnik	usatsjɛ	umatsɛk	uzˌɛvik	ɔwavik	ɔdɔʂɛk	ɔtsunɛk
Adjectival	usuwɛ	ubazˌwɛ	udʑɛzˌwɛ	ɔjuwɛ	ɔruzˌwɛ	ɔɕapwɛ	azˌavɛ	atʃɨpwɛ	alazˌwɛ	amɔzˌwɛ
Verbal	ɔɕɐ̨wa	ɔɕitɕɛtɕ	ɔzˌɔvuj	uzʲɐ̨wa	azʲijtɕɛ	aʂɐ̨wa	aɲitɕɛtɕ	aɕɔvuj	uzʲijmɨ	umitɕɛtɕ

Table 5 - Tails (and Groups) of Experimental Stimuli in IPA

Head groups and tail groups were systematically combined to create ten counterbalanced presentation lists, each of which included each head combined with three distinct tails (one nominal, one adjectival, one verbal), and each tail combined with however many heads were in the group it was paired with (five or six). The first presentation list combined head groups numbered *n* with tail groups numbered *n.* In each subsequent presentation list, the tail group number was shifted up by one. This means that the second presentation list paired head group *n* with tail group *n+1*, the third list paired head group *n* with tail group *n+2*, and so on. For example, the first head group comprised [jf-, xɛ-, ʤm-, ʂp-, gn-, zˌr-], and the first tail group comprised [-axatsjɛ, -usuwɛ, -ɔɕɐ̨wa]. All the participants who saw the first presentation list, in which head group 1 is paired with tail group 1, saw the words [jfaxatsjɛ, xɛaxatsjɛ, ʤmaxatsjɛ, ʂpaxatsjɛ, gnaxatsjɛ, zˌraxatsjɛ, jfusuwɛ, xɛusuwɛ, ʤmusuwɛ, ʂpusuwɛ, gnusuwɛ, zˌrusuwɛ, jfɔɕɐ̨wa, xɛɔɕɐ̨wa, ʤmɔɕɐ̨wa, ʂpɔɕɐ̨wa, gnɔɕɐ̨wa, zˌrɔɕɐ̨wa]. All the participants in the second presentation list, in which head group 1 is

[1] Frequencies were estimated from the largest available corpus of child-directed speech in Polish (Haman et al. 2011). For both type and token frequency, 'high' was defined as occurring in the top 50 percentile, and 'low' was defined as occurring in the bottom 50 percentile. Where possible, clusters with frequencies closer to the extremes were favored.

paired with tail group 2 ([-aʂoɛtɕik, -ubazʐwɛ, -ɔɛitɛɛtɛ]) saw the words [jfaʂoɛtɕik, xɛaʂoɛtɕik, ʤmaʂoɛtɕik, ʂpaʂoɛtɕik, gnaʂoɛtɕik, ʐraʂoɛtɕik, jfubazʐwɛ, xɛubazʐwɛ, ʤmubazʐwɛ, ʂpubazʐwɛ, gnubazʐwɛ, ʐrubazʐwɛ, jfɔɛitɛɛtɛ, xɛɔɛitɛɛtɛ, ʤmɔɛitɛɛtɛ, ʂpɔɛitɛɛtɛ, gnɔɛitɛɛtɛ, ʐrɔɛitɛɛtɛ]. In this way, every participant saw every test cluster three times and every tail five or six times.

This stimulus creation and presentation scheme allows for the use of mixed effects regression models that include full random effects structure for participant and tail. Every participant saw every head as a word of each one of the three parts of speech represented within the tails and saw every tail at all of the levels of sonority rise (-2 to 2) and attestedness (attested and unattested) represented within the test heads. Because every tail appeared at all levels of the head factors for each participant, but not with all heads, tails are appropriately treated as a random effect. This method of stimulus creation was preferred over, for example, randomly drawing tails from a much larger list to be paired and presented to each participant, because such random generation could result in accidental gaps in factor combinations or under-/over-presentation of some tails. Our counterbalancing design ensures equitable presentation rates across conditions, thereby maximizing the reliability of the random effects estimates that account for the accidental effect of any one tail or participant. This in turn increases the likelihood that our estimates for the effect of sonority on ratings, after accounting for random effects, are as reliable as possible.

The resulting 159 test words (53 heads x 3 tails) in each presentation list were mixed with 240 list-invariant fillers, which varied in word length (1 to 4 syllables) and onset length (0 to 3 consonants). These were chosen so as to ground the ends of a well-formedness scale, with some fillers clearly very good (e.g. ‘czownik’ and ‘skatościami’), some clearly very bad (e.g. ‘fmkłyrzłe’ and ‘śmłarstw’), with a few hypothesized to be in-between (e.g. ‘siskr’ and ‘mlicowatych’). By including fillers instantiating a broad spectrum of wellformedness, we minimize the possibility that ratings on the test stimuli could be subject to floor or ceiling effects at the endpoints of the scale. All of these words were written in standard Polish orthography, which is relatively phonetically transparent. In cases where there was more than one way to write the intended sequence of sounds, the most phonetically transparent way of doing so was selected. Orthographic transparency was judged by one trained linguist native Polish speaker who is not one of the authors and two linguistically naive native Polish speakers. Only orthographic representations that all three of these judges considered unambiguous were used.

2.1.2 *Procedure* The experiment was run online via Ibex Farm (Drummond 2013). Each participant who clicked on the study link was randomly assigned to one of the ten presentation lists. All of the stimuli words in a list were presented orthographically one-by-one in randomized order. Participants were instructed to pronounce each word out loud and rate its naturalness as a word of Polish on a scale from 1 (low) to 7 (high). Before beginning the main test phase of the experiment, they received training for the rating task that consisted of showing example "very unnatural" (e.g. ‘kjamś’), "very natural" (e.g. ‘szczog’), and rather "middle" (e.g. ‘chpoty’) nonsense words. All participants saw the same set of training items. The study took participants an average of 45 minutes to complete. At one-third and two-thirds of the way through the study (after approximately 133 and 266 items, respectively), participants were encouraged to take a short break, and offered the option of watching humorous internet videos. After finishing the study, participants were given the option of entering an email address into a web form, and receiving payment of 28 PLN ($7.25) via PayPal. 75 participants took advantage of this option, and received payment.

2.1.3 *Participants* The authors' contacts in Poland distributed the link to the study to their acquaintances, family members, and, in one case, students at Adam Mickiewicz University in Poznań. 115 participants completed the study. 40 identified as male and 75 identified as female. The mean age of participants was 34 years, and the median was 25 years. A participant's data were excluded from analysis for any one of the following reasons: (i) if the participant clearly did not use the response scale in the appropriate direction (as assessed by ratings of expected very good and bad fillers), (ii) if more than twenty-five percent of a participant's responses were rendered in less than 1000 milliseconds (indicating the participant was not performing the task as instructed), (iii) if the participant reported learning another language before Polish or being exposed to another language in childhood, (iv) if the participant reported having a speaking or hearing disorder, (v) if the participant reported spending more than two years outside of Poland in a country the majority language of which she spoke with a high degree of proficiency. This

left data from 81 native Polish-speaking adults without speaking or hearing disorders or substantial time spent abroad, who responded sensibly to fillers in a reasonable time frame.

2.2 *Results* Mean ratings (standard error) by sonority rise are given in Table 6. The average ratings for each head plotted by sonority distance and attestedness (the points), as well as average ratings by sonority and attestedness overall (the curves), are depicted visually in Figure 1. The R package Ordinal was used to perform statistical analysis (Christensen 2015). An ordinal mixed effects model was fitted with rating as the dependent variable, fixed effects of SSP (-2 to 2), attestedness (unattested=-1 vs. attested=1), and their interaction, and random slopes and intercepts by participant and tail (maximal random effects structure). This revealed a significant main effect of SSP (β=0.28, z=9.38), indicating participants gave higher ratings to higher sonority rises, a significant main effect of attestedness (β=0.90, z=17.48), indicating higher ratings for attested than unattested clusters, and no significant interaction between SSP and attestedness (β=-0.005, z=-0.30), indicating sensitivity to SSP did not differ significantly between attested and unattested clusters. The overall patterns remain the same when predictors for high type and high token frequency are added to the model with de-correlated random effects: SSP (β=0.22, z=7.11), attestedness (β=0.91, z=18.26), high type (β=0.53, z=14.40), high token (β=0.11, z=3.86), and the interaction between SSP and attestedness (β=-0.07, z=-5.02) were all significant. This indicates that the effects of SSP and attestedness are robust to the frequency imbalance in our design; their negative interaction indicates that SSP is more predictive for ratings of unattested than attested clusters, as expected given that acceptability increases for unattested clusters from -3/-2 to 1, but only increases from -3/-2 to 0 for attested clusters.

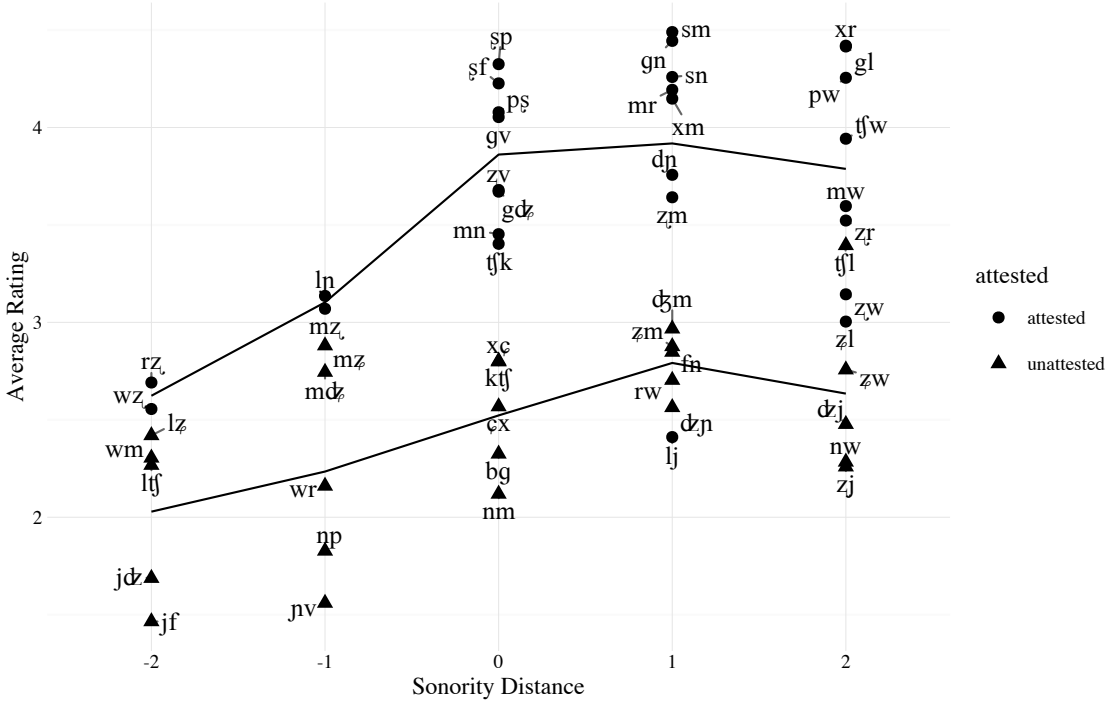

Figure 1 - Average Ratings for Each Head (points) or Overall (line) by Sonority and Attestedness

Sonority Rise	-3/-2	-1	0	1	2/3
Attested clusters	2.62 (0.13)	3.10 (0.13)	3.86 (0.17)	3.91 (0.17)	3.79 (0.17)
Unattested clusters	2.03 (0.11)	2.23 (0.12)	2.52 (0.12)	2.79 (0.14)	2.63 (0.14)
All clusters	2.20 (0.15)	2.48 (0.17)	3.35 (0.20)	3.48 (0.20)	3.34 (0.20)

Table 6 – Mean Ratings (standard error) by Sonority Rise

2.3 *Discussion* These results demonstrate that Polish speakers are sensitive to differences among levels of sonority rises across both attested and unattested clusters. Polish speakers seem to generally

disprefer clusters they have never encountered before in initial position, irrespective of their sonority profile. This could stem from orthographic unfamiliarity, difficulty coordinating novel sequences of articulation, and/or an implicit awareness of other (gradient) phonotactic generalizations of Polish, orthogonal to any effects of SSP. Our results cannot definitively differentiate among these possibilities, but we return to this question in the next section, where we consider the predictions of phonotactic models for well-formedness in Polish.

While SSP is strongly predictive of participants' ratings overall, the ratings also noticeably diverge from the preferences expected on the basis of SSP alone. Contrary to the SSP, Polish speakers do not prefer all rises (rise = 1 and rise = 2/3) to plateaus (rise = 0), and they do not prefer higher rises (rise = 2/3) to lower rises (rise = 1). When a model with the same structure as above was fitted to only the 0 to 2/3 range of sonority profiles, it found no significant effect of SSP (β=0.01, z=0.20), indicating there is no overall significant effect of SSP in the 0-2/3 range. We also fitted a model with attestedness, forward difference coding for the levels of SSP, and de-correlated random effects to all the data to perform pairwise comparisons between adjacent levels on the sonority scale. It revealed significant differences for all adjacent levels of the SSP (-2 vs -1: β=0.48, z=6.11, -1 vs 0: β=0.74, z=9.79, 0 vs 1: β=0.23, z=3.58, 1 vs 2: β=-0.22, z=-3.05), attestedness (β=0.75, z=16.97) and significant interactions between the -1 vs. 0 and 0 vs. 1 differences of SSP and attestedness (-2 vs -1 by attest: β=0.16, z=2.22, -1 vs 0 by attest: β=0.21, z= 2.94, 0 vs 1 by attest: β=-0.14, z=-2.60, 1 vs 2 by attest: β=0.02, z=0.30)[2]. A model with forward difference coding for the levels of SSP fitted to just the attested clusters found significant differences between -2 vs -1 (β=0.63, z=5.16), -1 vs 0 (β=0.93, z=8.25), and again a negative difference in 1 vs 2 (β=-0.20, z=-2.43), but no significant difference in 0 vs 1 (β=0.09, z=1.27). These analyses confirm the observations above that ratings rise from -3/-2 to 0/1 (with preferences for 1 vs 0 carried by unattesteds), but then fall from 1 to 2.

To summarize, the results confirm that SSP and attestedness are both significant predictors of participants' ratings, and the pattern of results across the sonority distance scale is largely consistent for both types of clusters. Polish provides a rare opportunity to disentangle sonority sequencing from attestedness entirely, and our results are largely consistent with independent effects of these factors. Although SSP is predictive overall, participants' ratings in the 0-2/3 sonority distance range raise questions about the influence of SSP on learning. These questions are especially pertinent given the frequency distribution for initial clusters in Polish discussed earlier, which contradicts the SSP exactly in this range. The next section turns to these questions, investigating the extent to which the behavioral results can be explained by the lexicon, orthography, and generalization from lexical statistics.

3 Phonotactic Modeling

This section describes the methods and results of computational modeling of the phonotactic knowledge of adult Polish native speakers. After introducing the models and their training, we present a general evaluation that considers each model's ability to predict participants' ratings. Our main concern however, is whether the statistical properties of the Polish lexicon can conceivably account for the above findings, namely, that adult Polish speakers exhibit gradient sensitivity to the SSP for both attested and unattested clusters, especially in the transition from sonority falls to plateaus. We therefore also present a sonority-based evaluation that focuses specifically on the models' abilities to capture the participants' patterns of preference across the sonority distance scale for both attested and unattested clusters.

3.1 *Methods* We report the predictions of an array of computational models that have been investigated in previous work, including the kinds of models that have been argued to be sufficient to derive sonority projection in languages like English, Mandarin, and Korean (Daland *et al*. 2011, Hayes 2011). All models were trained on a phonetically-transcribed lexicon (Jarosz to appear) derived from the largest available corpus of spontaneous, child-directed speech in Polish (Haman et al. 2011).

We consider several models to investigate the possible effects of orthography, basic lexical statistics and lexical analogy. While such models have not been successful in deriving sonority projection effects in previous studies (Berent *et al.* 2007, Daland *et al.* 2011), we include them here to explore the possible

[2] The Bonferroni-corrected z value for an alpha of 0.05 across the 4 ordinal mixed effects models reported here is 2.24.

influence of these various factors on our behavioral findings and to explore their limitations in deriving sonority-based preferences in the present case. To investigate the possible influence of orthographic regularities on participants' ratings, we consider two models of orthotactics: a classical bigram (Grapheme Bigram) and a classic trigram (Grapheme Trigram) model. These models are computed over graphemes, utilizing simple Laplace smoothing with a low pseudocount of 0.001 to allow nonce forms with unattested clusters to receive non-uniform, non-zero probability (Jurafsky & Martin 2008). This allows the models to differentiate among words with unseen sequences based on the other sequences that make up these words. For evaluating the potential influence of basic segmental co-occurrence regularities, we consider bigram (Phoneme Bigram) and trigram (Phoneme Trigram) models over phonemes, estimated with Laplace smoothing as for the grapheme models. Finally, as in Daland *et al.*, we also consider a model that assigns wellformedness scores based on lexical analogy (GNM): the Generalized Neighborhood Model (Bailey & Hahn 2001)[3]. This model favors nonce forms with many similar existing lexical items.

The second group of models includes the kinds of models that have previously been argued to be sufficient for deriving sonority projection. We present several variants of the UCLA Phonotactic Learner (Hayes & Wilson 2008), allowing the model to induce either 100 (HW2008 100) or 200 (HW2008 200) constraints on the basis of the input. Daland *et al.* (2011) found the 100 constraint model most successful at deriving sonority projection for English. We also consider the set of 32 sonority-regulating constraints proposed by Hayes (2011), which succeeded on toy Mandarin / Korean data (H2011 UG). In this case, the UCLA Phonotactic Learner is used only to weight pre-existing constraints on the basis of lexical statistics.

Following Daland *et al*. (2011), we evaluated two versions of some models, one trained on the basic transcriptions and another trained on syllabified transcriptions. As in Daland *et al*., syllabification was assigned automatically using the maximum onset principle (Selkirk 1982): that is, word-medial clusters were syllabified with the longest onset observed word-initially. Again following Daland *et al.*, for the models that rely on phonological features (HW2008, GNM), syllabification was encoded using a [±rhyme] feature, while for models relying on atomic segments (phoneme n-grams), onset and coda consonants simply utilized distinct alphabet symbols. We did not consider a syllabified variant for the Hayes (2011) constraint set since these constraints do not refer to syllable position.

	Unsyllabified				Syllabified			
	Overall	Attest	Unattested	SSP β (t)	Overall	Attest	Unattested	SSP β (t)
Grapheme Bigram	**0.65**	**0.52**	0.20	**0.24** (10.52)				
Grapheme Trigram	**0.84**	**0.84**	-0.03	**0.20** (8.78)				
Phoneme Bigram	**0.63**	**0.37**	0.15	**0.25** (10.65)	**0.79**	**0.47**	0.15	**0.13** (5.67)
Phoneme Trigram	**0.78**	**0.69**	-0.21	**0.16** (7.34)	**0.81**	**0.70**	-0.03	**0.15** (7.22)
GNM	**0.42**	**0.50**	0.10	**0.30** (13.31)	**0.42**	**0.51**	0.11	**0.30** (13.31)
HW2008 100	**0.64**	0.06	**0.45**	**0.23** (10.09)	**0.60**	**0.37**	**0.40**	**0.19** (8.19)
HW2008 200	**0.63**	0.06	**0.54**	**0.22** (9.71)	**0.70**	**0.31**	**0.49**	**0.15** (6.53)
H2011 UG	0.14	0.01	0.25	**0.23** (10.31)				
SSP Only	**0.48**	**0.43**	**0.54**					

Table 7 - Correlations Between Model Predictions and Average Ratings, Effect of SSP on Residuals

3.2 *General Evaluation* Table 7 presents a quantitative evaluation of the models, showing the correlations between the models' predictions for each head and the average ratings for each head. To compute the model predictions for each head, each model's score was transformed using linear regression to predict ratings. Then, the best-fit predictions of each model were averaged by head. Correlations that significantly differ from zero at a two-tailed alpha of 0.05 are bolded. Following recent work, correlations are reported for model predictions and ratings across all 53 heads ('Overall'), and separately for just the 28 attested ('Attested) and just the 25 unattested ('Unattested') heads. For comparison to the computational

[3] We report results with the parameters ('lin') that Daland *et al*. found most successful in their simulations. We ran the model using software kindly provided to us by Robert Daland.

models, the table also shows the correlations ('SSP Only') between average ratings for each head and each head's sonority distance (-2, -1, 0, 1, or 2), quantifying the association between ratings and SSP.

The correlations overall and the correlations just for attested clusters are impressively high for the grapheme n-gram, phoneme n-gram, and analogical models, indicating the models have significant predictive value. However, none of these models generalize appropriately to unattested clusters. The correlations between these models' predictions and ratings for unattested clusters are near zero or negative, indicating that they cannot capture participants' preferences among these novel stimuli. Previous studies have demonstrated similar results for English (Albright 2009; Daland et al. 2011). The n-gram models are not sensitive to featural similarity of unseen clusters to existing clusters, while the analogical model is not sensitive to the positional and co-occurrence restrictions that regulate initial clusters in the Polish lexicon. These limitations prohibit these models from generalizing appropriately to novel clusters.

The HW2008 models fare well overall and perform the best on unattested clusters. Indeed, the HW2008 models trained on syllabified input achieve significant positive correlations for both attested and unattested clusters, suggesting these models are successfully capturing important aspects of phonotactic wellformedness. The fact that syllabification is necessary for this level of performance coincides with Daland *et al*.'s conclusions. These models are picking up on under-represented feature combinations that reflect systematic regularities participants are apparently sensitive to. However, we show in the next section that these positive correlations do not reflect systematic sonority sequencing preferences. The fact that the correlations between participant ratings and the SSP scale itself are higher than those predicted by these models is another indication that these models do not derive robust SSP preferences.

Finally, when the model is restricted to considering sonority regulating constraints (H2011 UG), the overall predictions are poor. Recall that it was crucial for modeling sonority projection for languages like Mandarin and Korean that these sonority regulating constraints not be restricted to applying only to onset clusters. The present results demonstrate, however, that such unrestricted generalization about sonority sequencing does not produce an overall preference for rising sonority when these constraints are weighted based on Polish lexical statistics. This is not especially surprising given the statistical distribution of Polish initial onsets discussed earlier and the further observation that an abundance of other phoneme sequences in Polish words support sonority *falls*, such as nucleus-rhyme, coda-coda, and coda-onset transitions.

To summarize, the results reviewed in this section indicate that nearly all models, with the probable exception of H2011 UG, have some predictive value for participants' ratings. This indicates these models likely capture aspects of phonotactic wellformedness that drive participants' ratings. Most models, however, fail to generalize appropriately to unattested clusters, with the exception of the HW2008 models trained on syllabified data. In the following section we show that nonetheless even these models fail to capture the systematic sonority-based preferences that characterize participants' ratings.

3.3 *Sonority-Based Evaluation* Here we consider each model's ability to capture the effects of sonority distance for both attested and unattested clusters. The failure of the models to capture effects of SSP is shown quantitatively in the 'SSP β (t)' columns of Table 7. Each value indicates the coefficient and t statistic for an SSP predictor in a mixed effects linear regression model with full random effects structure fit to the residuals of the phonotactic models' best-fit predicted values. These coefficients represent the contribution of SSP after accounting for the phonotactic model's effects on ratings. For all models, SSP remains highly significant, indicating the phonotactic models fail to fully account for the influence of SSP on ratings. The relationship between model predictions and sonority is also depicted graphically in Figure 2. Solid lines show the participants' average ratings and standard error, while the dashed lines depict each model's best-fit predictions for attested (circle) and unattested (triangle) clusters at each sonority distance level on average. These figures reveal the models' predictions by sonority distance. They also show how, if at all, the various factors captured by these different models might be affecting participants' ratings across levels of sonority distance and attestedness. The graphs for the unsyllabified Phoneme Trigram and GNM models are nearly identical to their syllabified counterparts and are not shown. The graphs for the HW2008 100 models are not shown as they are nearly identical to the HW2008 200 unsyllabified variant.

Confirming the claims made above based on correlations, these plots reveal that the n-gram and analogical models fail to account for any systematic preferences among the unattested clusters. Although some of these models show an up and down pattern (unsyllabified Phoneme and Grapheme bigrams) for unattested clusters, most predict essentially flat preferences across the scale. Crucially, however, none of

these patterns resembles the rising curves across the sonority scale observed in the participants' ratings for unattested clusters. For attested clusters, only the models with access to syllabification resemble the curves for the ratings at all, but since the same models fail to yield any systematic correspondence to the ratings for unattested clusters, we conclude that none of the n-gram or analogical models can explain the systematic sonority-based preferences evident in the behavioral results.

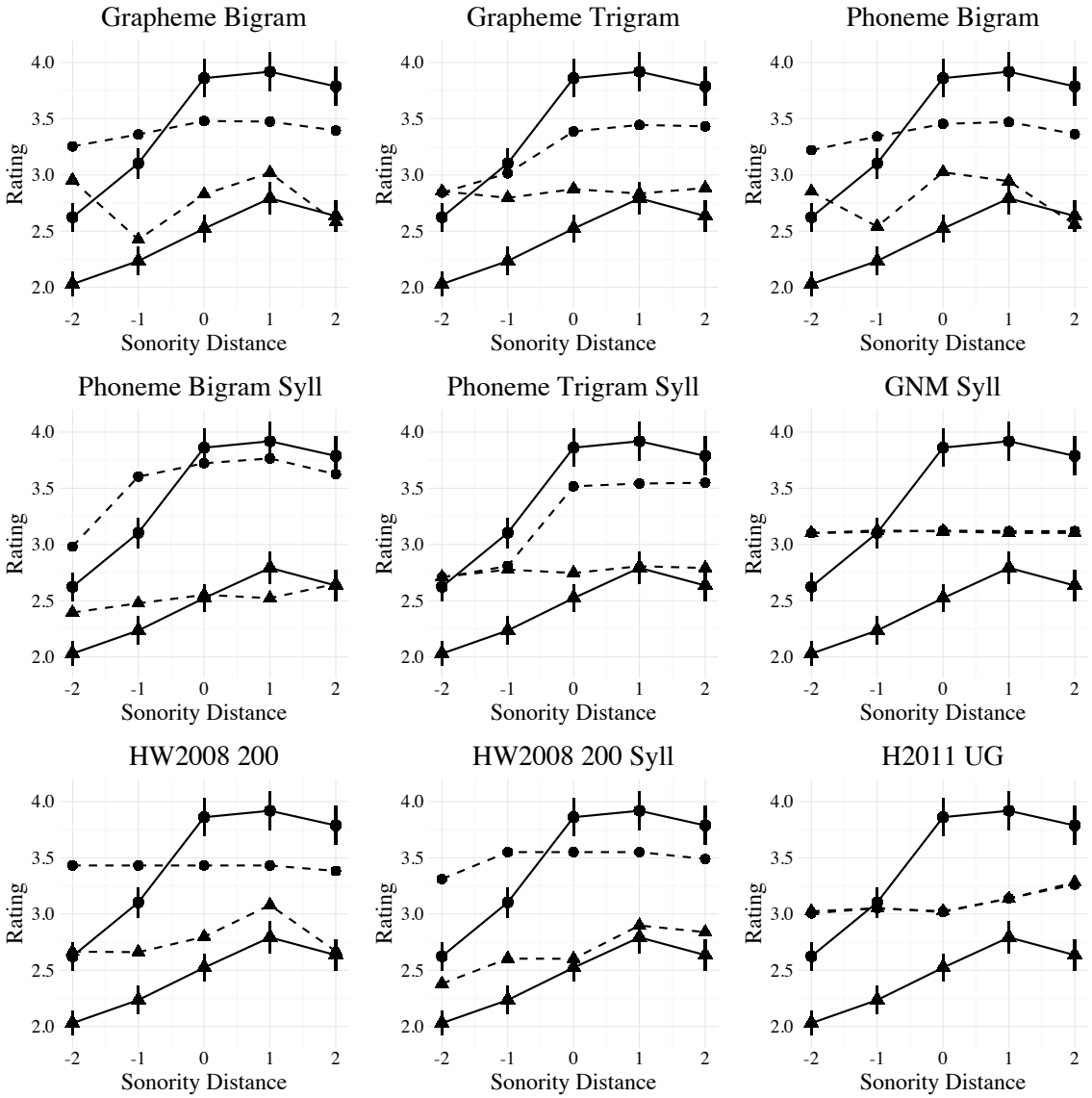

Figure 2 - Average Ratings by Sonority and Attestedness v. Model Predictions

Turning to the remaining models, each makes qualitatively somewhat different predictions. The HW2008 200 model trained on unsyllabified input incorrectly predicts essentially flat preferences across the scale for attested clusters and a preference for sonority rises of 1 for the unattested clusters. The flat preferences reflect the fact that the model fails to induce constraints penalizing attested clusters. Neither of these predictions captures the rising preferences seen in the ratings. As might be expected based on the high correlations reported in the previous section, the HW2008 200 model trained on syllabified input appears the closest to capturing the trends in the ratings. However, it too fails to explain the robust preferences participants exhibited for plateaus over sonority falls. The plot does not convey its dramatic failure, like that of HW2008 100, to generate systematic predictions for attested clusters. Its predictions are flat because the model's constraints penalize just two of the 28 attested clusters (wz̨, z̨r), making no distinctions

whatsoever among the remaining attested clusters, unlike our participants. Finally, the H2011 sonority-regulating constraints predict essentially flat preferences across the scale, and because these constraints only evaluate sonority features, this model additionally fails to distinguish attested from unattested clusters.

Based on these observations, we conclude that although some of these models capture aspects of participants' wellformedness ratings, none of the models fully explain the participants' sonority related preferences. The divergence between the correlations of the previous section and the sonority scale plots shown here – for example, the relatively high overall correlations of the GNM model and the flat predicted preferences for the GNM model across the sonority scale – emphasize the fact that strong correlations alone do not demonstrate sonority projection. Strong correlations can reflect aspects of wellformedness that are entirely orthogonal to sonority sequencing. As the results show, the HW2008 models consistently picked out constraints that penalized unattested clusters more severely than attested clusters. Indeed, the HW2008 models generally induced constraints that penalized only one or two of the attested clusters. This indicates the unattested clusters are indeed characterized by under-represented feature combinations that are largely orthogonal to the SSP and reflect participants' robustly lower scores for unattested clusters. These findings suggest that sensitivity to gradient phonotactic constraints that are largely independent of sonority is likely responsible, at least in part, for the preference for attested clusters.

4 General Discussion

Our main findings are that adults' phonotactic judgments exhibit gradient sensitivity to the SSP for both attested and unattested clusters. The observed pattern of results is not predicted by any of the computational models we tested, which cover a broad spectrum and include the models that have been argued to be sufficient for English, Mandarin, and Korean. Consequently, we conclude that Polish speakers' gradient acceptability judgments are affected by sonority sequencing in a way that is not derivable from unconstrained generalization from the statistics of the lexicon, at least with the models currently available.

At the same time, participants' sonority sequencing preferences are not entirely predictable from the SSP: the average ratings for both the attested and unattested clusters fail to increase across the 0-2 range overall and indeed decline for large rises. Considering our results together with those of Jarosz (to appear), who found that Polish-acquiring children strongly favor rises over plateaus, suggests that prior bias and experience jointly influence learning of high-level phonological generalizations over time. The hypothesis most consistent with the current results posits that prior phonetic or grammatical bias shapes early preferences, but sufficient exposure over time can distort these preferences.

Although we have argued the models do not succeed in deriving the sonority sequencing preferences of Polish speakers, the results indicate the models are successful in capturing some aspects of acceptability. If other phonotactic or orthotactic factors are influencing results, this raises the possibility that these effects could be affecting different levels of the sonority sequencing scale disproportionately. Could this explain the flattening and reversing of preferences we observe in the 0-2 range? Although we cannot rule out this possibility entirely, we note that none of the models provide consistent evidence that would predict the observed drop in acceptability for both attested and unattested large rises. Nonetheless, the possibility remains that participants responded to some orthotactic or phonotactic patterns that were not evenly distributed across conditions. Given this and the novelty of this aspect of our results, this pattern should be replicated before drawing strong conclusions about whether Polish speakers indeed reverse the SSP scale in this range. Several alternative possibilities should also be considered. While much prior work has implicitly assumed the effects of any inherent SSP bias should be linear, it is not obvious a priori why this should be the case. It is also possible that an inherent SSP bias could become warped – but not obliterated – by experience. These are important questions for future modeling and experimental work to clarify.

A necessary part of the answer to these questions will require the formulation and testing of explicit models of phonotactic learning that encode the interactions between inherent SSP bias (or other constraints on learning) and language experience. We have argued that unconstrained generalization from the lexicon, as instantiated by the models tested here and elsewhere, is not sufficient for deriving the sonority sequencing preferences of Polish speakers. Providing models like the UCLA Phonotactic Learner access to syllable structure is not enough: nothing prevents these models from ignoring syllabification if no systematic sonority-based regularities in the lexical statistics rely on it. Likewise, nothing prevents such models from inducing constraints with contradictory effects to the SSP, given appropriate evidence. On the other hand, narrowing the models' focus to syllable structure- or sonority-based generalizations limits their

ability to detect other aspects of wellformedness native speakers are reliably sensitive to. Thus, the task of accounting for the basic behavioral findings on sonority sequencing and phonotactics to date is more complex than it may first appear. Assuming future experimental studies confirm sonority projection effects exist in Polish and other languages where unconstrained generalization from the lexicon appears to make the wrong predictions, successful computational models will need to incorporate sonority sequencing biases in a way that is simultaneously flexible and sensitive to linguistic experience yet powerful and persistent enough not to be obliterated entirely by contradictory input.

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26(01). 9–41.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44(4). 568–591.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87. 84–125.

Becker, Michael, Andrew Nevins & Jonathan Levine. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language* 88(2). 231–268.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno & Paul Smolensky. 2008. Language universals in human brains. *Proceedings of the National Academy of Sciences* 105(14). 5321–5325. doi:10.1073/pnas.0801469105.

Berent, Iris, Donca Steriade, Tracy Lennertz & Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104(3). 591–630.

Christensen, Rune Haubo Bojesen. 2015. *Package "Ordinal": Regression Models for Ordinal Data*. Available at: http://www.cran.r-project.org/package=ordinal/.

Clements, George. 1990. The role of the sonority cycle in core syllabification. In John Kingston & Mary Beckmann (eds.), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech.*, 283–333. Cambridge: Cambridge University Press.

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, ed. by John Coleman, 49–56. East Stroudsburg, PA: Association for Computational Linguistics.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology* 28(02). 197–234.

Drummond, Alex. 2013. *Ibex Farm*. http://spellout.net/ibexfarm.

Ernestus, M. & R. H. Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*. 5–38.

Haman, Ewa, Bartłomiej Etenkowski, Magdalena Łuniewska, Joanna Szwabe, Ewa Dąbrowska, Marta Szreder & Marek Łaziński. 2011. *Polish CDS Corpus*. Available from http://childes.psy.cmu.edu.

Hayes, Bruce. 1999. Phonetically driven phonology: The role of Optimality Theory and inductive grounding. In Michael Darnell, Edith Moravcsik, Frederick Newmeyer, Michael Noonan & Kathleen M. Wheatley (eds.), *Formalism and Functionalism in Linguistics, Volume 1: General Papers*, 243–285. Amsterdam: John Benjamins.

Hayes, Bruce. 2011. Interpreting sonority-projection experiments: the role of phonotactic modeling. *Proceedings of the 17th international congress of phonetic sciences*, 835–838.

Hayes, Bruce & Zsuzsa Cziraky Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23(01). 59–104. doi:10.1017/S0952675706000765.

Hayes, Bruce & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44. 45–75.

Hayes, Bruce & Colin Wilson. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39(3). 379–440. doi:10.1162/ling.2008.39.3.379.

Hayes, B., K. Zuraw, P. Siptár & Z. Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4). 822–863.

Jarosz, Gaja. to appear. Defying the stimulus: acquisition of complex onsets in Polish. To appear in *Phonology: Special Issue on Computational Phonology*.

Jurafsky, Daniel & James H. Martin. 2008. *Speech & Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Second edition. Prentice Hall.

Ren, Jie, Liqun Gao & James L. Morgan. 2010. Mandarin Speaker's Knowledge of the Sonority Sequencing Principle. *20th Colloquium on Generative Grammar, University of Pompeu Fabra, Barcelona*.

Selkirk, Elisabeth. 1982. The Syllable. *The structure of phonological representations*. (Part 2). 337–384.

Selkirk, Elisabeth. 1984. On the major class features and syllable theory. In Mark Aronoff & Richard T. Oehrle (eds.), *Language sound structure*, vol. 107, 107–136. Cambridge MA: MIT Press.

Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce & David Kemmerer. 1997. Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. *Language and Speech* 40(1). 47–62.

Wilson, Colin. 2006. Learning Phonology With Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science* 30(5). 945–982.