Phonotactic Probability and Sonority Sequencing in Polish Initial Clusters

Gaja Jarosz, Yale University

The Sonority Sequencing Principle (SSP) states that syllables with a sonority rise in the transition from the onset to the nucleus are preferred cross-linguistically. Experimental evidence indicates that English speakers exhibit gradient sensitivity to the SSP even for onset clusters that are not attested in English (Davidson 2006, 2007; Berent et al. 2007; Daland et al. 2011). Berent et al (2007) argue that there is no direct lexical evidence for these preferences and suggest that the principle may therefore be innate. However, Daland et al. (2011) show that computational models that are able to form generalizations on the basis of phonological features can detect SSP preferences on the basis of English lexical statistics. This paper explores this controversy using computational and developmental approaches in a language (Polish) with very different sonority sequencing patterns from English. There are two main contributions: 1) using computational modeling, we show that the basic lexical statistics of Polish contradict the SSP, favoring onset clusters with sonority plateaus, and 2) we show that children nonetheless exhibit sensitivity to the SSP, favoring onset clusters with a sonority rise.

The data for all analyses comes from the Weist-Jarosz Polish Corpus available via CHILDES (Weist and Witkowska-Stadnik 1986; Weist et al. 1984; Jarosz 2010). The corpus includes transcriptions of child-directed speech, which were used to estimate the phonotactic probabilities of initial clusters with various sonority profiles in the language input. Following Daland et al., we use Clement's (1988) coarse-grained sonority scale: Obstruent (0) < Nasal (1) < Liquid (2) < Glide (3) < Vowel (4). Table 1 shows the relative frequencies of clusters C_1C_2 varying in the degree of sonority rise from C_1 to C_2. Regardless of whether input statistics are computed based on the lexicon (type frequency) as in Daland et al. or based on word tokens (token frequency), the input provides strong evidence in favor of sonority plateaus (sonority rise of 0), contrary to the SSP. The situation is not improved by considering the relative frequencies of individual sonority profiles (Table 2) or by splitting fricatives and plosives into separate sonority classes (not shown) because roughly half of all onset clusters in the input are composed of two obstruents, and fricative-plosive clusters are the most frequent among these. Thus, if speakers of Polish are using these statistics to construct preferences about the relative well-formedness of onset clusters, a preference in favor of sonority plateaus is expected.

We further examine these predictions by performing several computational analyses with the developmental data. Numerous previous studies have used developmental data to demonstrate children's sensitivity to phonological principles like the SSP (Gnanadesikan 2004; Pater 1997; Pater and Barlow 2003) and other principles of syllable well-formedness (Fikkert 1994; Demuth 1995; Fee and Ingram 1982; Vihman 1992; Levelt et al 2000; Lleó and Prinz 1996). Although such sensitivity has been demonstrated in phonological acquisition across many languages and domains, the debate regarding the source of such knowledge (whether it can be learned, and, if so, from what information) continues. Part of the challenge in disentangling competing hypotheses regarding the origin of principles such as the SSP is due to the fact that cross-linguistic generalizations are often mirrored by language-internal statistics. For example, English complex codas are less frequent than simple codas, and therefore predictions based on phonotactic probability and innate principles align. Polish therefore provides an important test case because several straightforward ways of calculating phonotactic probability for onset clusters (as described above) make predictions for acquisition that are contradictory to the SSP.

In the first developmental analysis we test whether Polish children are sensitive to the SSP. We fit logistic regression models to analyze the children's production accuracy on initial clusters varying in sonority profile while statistically controlling for other factors that may influence accuracy. We considered control predictors for phonological context (stress, identity of the adjacent vowel), subject, age, and word frequency. The predictor of interest, SSP, assigns a numerical value to each cluster (n=1334) based on the degree of sonority rise, as in Table 1. After inclusion of all five control predictors, SSP was significantly predictive of children's production accuracy ($\beta = 0.367$, $z = 6.9$; $p < 0.0001$). The direction of the effect was positive, indicating that children were significantly more accurate on clusters with a higher sonority rise. Therefore, the developmental data from this corpus of child Polish is consistent with the SSP.

In the second set of analyses, we considered several measures of phonotactic probability as predictors of accuracy in the regression models. Log token frequencies of sonority rises (Table 1) and sonority profiles (Table 2) were significantly predictive of production accuracy. However, crucially, the predictions were in the wrong direction: *higher* log frequency of rises ($\beta = -0.89$, $z = -6.2$; $p < 0.0001$) and of profiles ($\beta = -0.38$, $z = -2.9$; $p < 0.005$) was associated with significantly *lower* accuracy. Under the usual assumption that higher probability or frequency predicts greater well-formedness, this confirms the observation made above that these input statistics make predictions for acquisition that conflict with the SSP. Results for type frequency were similar although type frequencies were in general less predictive of accuracy.

In sum, this talk identifies a critical test case in Polish where basic input statics make predictions that are contrary to the SSP. We also demonstrate that children acquiring Polish show sensitivity to the SSP and behave inconsistently with these input statistics. The paper also presents results exploring alternative formulations of phonotactic probability. We find that some phonotactic probability formulations (those that make sonority sequencing generalizations on the basis of simple onsets and those that rely on segmental phonotactics) are more successful at predicting the developmental patterns, and we discuss the properties of these alternative models. However, none of these models fully capture the SSP effect. We discuss the implications of these findings for models of phonotactics and for phonological theory more generally.

Rise	Token Frequency (n = 5215)	Type Frequency (n = 1456)	Examples
-2	0.04%	0.07%	[lvɨ] ‘lions’
0	51.3%	47.9%	[ptak] ‘bird’, [mɲɛ] ‘me’
1	3.8%	7.3%	[mlɛkɔ] ‘milk’, [dmuxa] ‘blows’
2	23.5%	22.5%	[klɔʦɛk] ‘block’, [mjastɔ] ‘town’
3	21.3%	22.3%	[pjɛs] ‘dog’

Table 1 - Relative Frequency of Sonority Rise Profiles

Sonority Profile	LF	OO	NN	ON	NL	OL	NG	OG
Token Frequency	0.04%	50.90%	0.50%	3.60%	0.20%	20.40%	3.00%	21.30%
Type Frequency	0.10%	47.70%	0.20%	6.70%	0.60%	19.60%	2.90%	22.30%

Table 2 - Relative Frequency of Sonority Profiles (O=obstruent, N=nasal, L=liquid, G=glide)

Select References

Berent, I., Steriade, D., Lennertz, T. and Vaknin, V. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104(3). Elsevier. 591-630.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A. and Norrmann, I. 2011. Explaining sonority projection effects. *Phonology* 28(02). Cambridge University Press. 197-234.

Davidson, Lisa. 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics* 34(1). 104 - 137.