Language Acquisition

ISSN: 1048-9223 (Print) 1532-7817 (Online) Journal homepage: www.tandfonline.com/journals/hlac20

Input frequency and the acquisition of syllable structure in Polish

Gaja Jarosz, Shira Calamaro & Jason Zentz

To cite this article: Gaja Jarosz, Shira Calamaro & Jason Zentz (2017) Input frequency and the acquisition of syllable structure in Polish, Language Acquisition, 24:4, 361-399, DOI: 10.1080/10489223.2016.1179743

To link to this article: https://doi.org/10.1080/10489223.2016.1179743

LANGUAGE ACQUISITION
2017, VOL. 24, NO. 4, 361–399
http://dx.doi.org/10.1080/10489223.2016.1179743

Routledge
Taylor & Francis Group

Input frequency and the acquisition of syllable structure in Polish

Gaja Jarosz[a], Shira Calamaro[b], and Jason Zentz[b]

[a]University of Massachusetts Amherst; [b]Yale University

ABSTRACT

This article examines phonological development and its relationship to input statistics. Using novel data from a longitudinal corpus of spontaneous child speech in Polish, we evaluate and compare the predictions of a variety of input-based phonotactic models for syllable structure acquisition. We find that many commonly examined input statistics can make dramatically different predictions, as do different assumptions about the representational units over which statistics are calculated. We find that development is sensitive to multiple abstract units of phonological representation, supporting a crucial role for feature-based generalization. We also identify departures between the predictions of the best phonotactic models and children's production patterns that indicate that input sensitivity alone cannot fully explain the developmental patterns. We discuss the role of universal markedness and phonetic difficulty and argue that a full explanation requires reference to these biases.

ARTICLE HISTORY
Received 19 December 2014
Accepted 15 April 2016

1. Introduction

Recent computational and experimental work has led to a deeper understanding of language learning by investigating the complex ways in which learning outcomes depend on various statistical properties of the language input. In phonology, one growing area of research involves statistical and computational modeling of the primary language data to test the abilities of input-based models to explain adults' knowledge of their native language. A number of studies compare the results of well-formedness judgment tasks or wug test experiments (Berko 1958) probing adults' knowledge of phonotactics or phonological alternations to the predictions of explicit statistical or computational models derived from the language input (Albright 2009; Albright & Hayes 2003; Becker, Ketrez & Nevins 2011; Daland et al. 2011; Ernestus & Baayen 2003; Hayes & Londe 2006; Hayes & White 2013; Hayes & Wilson 2008; Hayes et al. 2009; Kager & Pater 2012). Evaluating the systematic predictions of input-based models is a crucial step in answering a fundamental question of cognitive science: What are the relative roles of experience and innate biases in language learning? These explicit evaluations characterize the extent to which language learning outcomes are predictable from the input, and they identify specific ways in which adults' knowledge of their language diverges from the predictions derived from modeling the language input. Where divergences occur, these findings support concrete elaborations of the input-based models and highlight areas where biases may be needed to constrain learning outcomes.

The relationship between the input and learning outcomes is also investigated in a substantial body of experimental work examining representational and computational constraints on adults' and infants' learning of artificial languages (Carpenter 2010; Finley & Badecker 2009; Moreton 2008; Newport & Aslin 2004; Pycha et al. 2003; Saffran, Aslin & Newport 1996; Seidl & Buckley 2005; Wilson 2006; see Culbertson, Smolensky & Wilson 2013; Moreton & Pater 2012a, 2012b) for recent literature reviews). Collectively, both strands of research have shown that human learning of phonological patterns is exquisitely sensitive to the statistical properties of the language input. At the same time, the findings also

CONTACT Gaja Jarosz jarosz@linguist.umass.edu Department of Linguistics, University of Massachusetts, 650 North Pleasant Street, Amherst, MA 01003-1100.

demonstrate that not all phonological patterns are learned equally well, indicating that there are biases interacting with and affecting the inferences learners make on the basis of the language input.

The debate regarding the relative roles of experience and biases or universals has also played a central role in research on first language acquisition of phonology. The ways in which phonological development abides by linguistic universals and grammatical principles have been well documented (Barlow 2007; Demuth 1995; Fikkert 1994; Gerken 1996; Gnanadesikan 2004; Jakobson 1968; Jesney & Tessier 2011; Pater 1997; Pater & Barlow 2003; Salidis & Johnson 1997). However, there is also substantial evidence for the important role of input frequency and phonotactic probability in the acquisition of phonological patterns (Coady & Aslin 2004; Edwards, Beckman & Munson 2004; Ingram 1988; Saffran, Aslin & Newport 1996; Zamuner 2009; Zamuner, Gerken & Hammond 2004), with a growing body of work investigating a role for both factors (Boersma & Levelt 2000; Edwards & Beckman 2008; Jarosz 2010, 2011; Levelt, Schiller & Levelt 2000; Roark & Demuth 2000; Stites, Demuth & Kirk 2004; Tessier 2009; Zamuner, Gerken & Hammond 2005). Like the modeling studies reviewed, this work suggests that both probabilistic phonotactics and biases play important roles in learning, but the exact nature and interaction of these factors remain poorly understood. Various definitions of frequency have been explored in the literature, and it is not clear to what extent predictions for acquisition depend on the exact characterization of frequency. There is little work examining the systematic predictions of computational or statistical models relying principally on input statistics for the developmental trajectory (see however Edwards & Beckman 2008; Zamuner, Gerken & Hammond 2004). Such work is critical for determining the relative contributions of input statistics and biases and for understanding the precise computations and representations underlying these statistics in phonological acquisition.

In this article we seek to better understand the nature of frequency effects in natural language acquisition by exploring the systematic predictions of explicit, input-based models for natural language development. In this way, we integrate aspects of the research areas discussed above. Like the computational and statistical modeling work discussed above, we use statistical models to make and test predictions for learning based on properties of the language input. However, rather than modeling learning *outcomes*, our focus is on modeling *development* to probe the input statistics that affect the acquisition process in a naturalistic setting. That is, we examine developing phonologies in children rather than modeling adults' end-state phonological knowledge. A principal goal of the present work is to determine the extent to which input statistics are predictive of phonological development. A related goal is to identify the consequences of adopting different hypotheses about frequency and the representations over which it operates. A secondary goal is to examine the divergences between the predictions of input models and how these divergences relate to biases of various kinds. We investigate these questions by analyzing the acquisition of syllable structure by four typically developing Polish children and by building statistical models testing the abilities of a range of frequency models to predict production accuracy. We also present qualitative analyses that highlight concrete areas where the frequency models succeed and where they fail in their predictions.

In the remainder of this section, we review previous work on the roles of input statistics and biases in phonological acquisition and discuss the goals of the current study. In the following sections, we discuss the data, including the spontaneous production data and the methods for calculating production accuracy and for extracting the frequency measures from child-directed speech (section 2), before turning to the analyses. Section 3 presents in-depth acquisition analyses that establish novel empirical findings on the development of syllable structure in Polish, while section 4 evaluates and compares the predictive capacities of seven frequency-based models. In the final section, we discuss implications for theories of phonological learning.

1.1. *Background: Statistics and universals in phonological learning*

Experimental work indicates that input statistics play a significant role in phonological learning and processing. Infants can extract phonotactic probabilities from brief exposure to artificial languages and use these statistics to segment continuous speech (Saffran, Aslin & Newport 1996). Before the

age of 1, infants can use phonotactic probabilities in their native language to segment continuous speech (Mattys & Jusczyk 2001). Phonotactic probability influences production accuracy of nonwords by children between 2 and 3½ years of age (Coady & Aslin 2004; Zamuner, Gerken & Hammond 2004). Adults' knowledge and use of their native language is also highly sensitive to phonotactic probability. Adults employ sophisticated knowledge of the statistical properties of their native language in producing phonological alternants for nonwords (Becker, Ketrez & Nevins 2011; Ernestus & Baayen 2003; Hayes & Londe 2006; Hayes et al. 2009). Adult spoken production is also sensitive to phonotactic probability: more probable sound sequences are produced more accurately under experimental conditions designed to induce speech errors even when complexity of these sequences is controlled (Goldrick & Larson 2008).

There is also evidence from cross-linguistic comparisons of phonological development that supports a causal role for frequency, indicating that sensitivity to input frequency constrains and guides acquisition. Some studies reveal different ages or orders of acquisition for the same structures across languages. For example, Ingram (1988) found that word-initial /v/ is mastered later in English than it is in Swedish, Estonian, and Bulgarian, where its frequency in the lexicon is significantly higher than in English. Ingram proposed that the language-particular frequency of /v/ predicts its ease of acquisition, with higher frequency corresponding to earlier acquisition. A similar example is found with respect to acquisition of /l/ in French and English: /l/ is mastered earlier by French-acquiring children than by English-acquiring children, and its frequency in running speech is higher in French than in English (Edwards & Beckman 2008; Vihman 1993).

At the same time, the debate about the relative roles of input frequency and innate or universal biases in the learning of phonological structure continues. There is no consensus about the exact characterization of frequency that is needed to capture phonological development, and there is no agreement about the nature or scope of the learning biases that are required. Contrasting with the studies reviewed above is a large body of work on phonological development that places a primary emphasis on innate biases rooted in universal markedness. It has long been observed that typologically rare and marked structures tend to be acquired later by children (Jakobson 1968; Stampe 1969). In influential work, Jakobson (1968) proposed that language acquisition mirrors language typology, with typologically less marked or more frequent structures acquired universally earlier than typologically marked or infrequent structures. Some of Jakobson's empirical predictions have been strongly supported by subsequent research. For example, children tend to acquire stops before fricatives and affricates across languages (Kent 1992; Kim & Stoel-Gammon 2010; Smit et al. 1990; Edwards, Beckman & Munson 2015).

In the domain of syllable structure, research across diverse languages has identified a set of systematic implicational laws between basic syllable types. CV syllables occur in all languages, while complex onsets, codas, complex codas, and null onsets occur only in some languages (Blevins 1995). This gives rise to implicational markedness laws stating that some syllable types (e.g., CVC) asymmetrically imply other syllable types (e.g., CV). Findings on the acquisition of syllable structure have mirrored these typological patterns. For example, examining the development of syllable structure in 12 children acquiring Dutch, Fikkert (1994) found that the first onsets to appear were singleton onsets, followed by optional onsets, and finally complex onsets. This developmental trajectory parallels typological universals, with less-marked structures acquired earlier. Fikkert also found that development of codas in Dutch obeys universal markedness: null codas were acquired first, then singletons, and finally complex codas. Early stages of syllable structure development in English follow a similar trajectory, with CV syllables appearing before closed syllables, open syllables with long vowels, and syllables with onset clusters (Demuth 1995; Fee & Ingram 1982; Vihman 1992). Similarly, Spanish and German development also begins with the CV core syllable, with closed syllables and clusters developing later (Lleó & Prinz 1996). At a finer-grained level of analysis, studies have found that universal sonority preferences (Clements 1990; Selkirk 1984) play an important role in syllable structure development. For example, multiple studies have found that cluster reduction patterns in child productions of onsets follow universal preferences, retaining either the least-sonorous consonants (Gnanadesikan 2004; Pater 1997; Pater & Barlow 2003; Ohala 1999) or those that form structural heads (Goad & Rose 2004). Other studies have found that development of

clusters, codas, and syllabification is sensitive to universal sonority principles (Stites, Demuth & Kirk 2004; Demuth & McCullough 2009; Łukaszewicz 2006; Ohala 1999).

Acquisition of basic syllable structure is consistent with universal principles, but the attested acquisition paths appear to be constrained by additional factors. Levelt, Schiller & Levelt (2000) found that while Dutch children's acquisition paths followed universal markedness principles, the observed paths among 12 children were further constrained in a way that coincided with the relative frequency of syllable types in spoken Dutch. For example, in their study, all 12 children acquired CVC syllables before V syllables, even though neither of these structures is more marked than the other cross-linguistically (neither asymmetrically implies the other). They proposed that a combination of universal principles and frequency guides development, with universal markedness determining possible developmental paths and frequency mediating among these possible paths. Levelt, Schiller & Levelt showed that this hypothesis was consistent with the detailed acquisition paths observed in Dutch, and interestingly, the only variation they observed among children in Dutch was in the relative order of acquisition of complex onsets and complex codas, neither of which is more marked than the other and which have roughly equal relative frequencies in Dutch. Work in other languages has shown that the relative order of acquisition of complex onsets and complex codas varies by language, with existing findings supporting earlier acquisition of complex codas in English and German (Kirk & Demuth 2005; Lleó & Prinz 1996) and earlier acquisition of complex onsets in French and Polish (Demuth & Kehoe 2006; Jarosz 2010). Furthermore, Jarosz (2010) showed that this cross-linguistic variation in acquisition order covaries with the relative frequencies of complex onsets and complex codas in spoken speech in these languages. These cross-linguistic findings support a causal role for input frequency.

While the evidence reviewed above suggests that both universals and frequency may play a causal role in phonological acquisition, the exact nature of the interaction between these factors remains unclear. There is no consensus regarding the precise roles of these factors, with recent work supporting divergent conclusions. Hua & Dodd (2000) assert that cross-linguistic frequency is not predictive of acquisition, and several recent studies argue that frequency is a primary factor in the acquisition of syllable structure (Edwards, Beckman & Munson 2004; Zamuner, Gerken & Hammond 2004, 2005). For example, Zamuner, Gerken & Hammond (2005) found that acquisition of singleton codas in English is better explained by input frequency than by cross-linguistic frequency. On the other hand, Ohala (1999) found support for universal sonority principles in the development of singleton codas in English. Others advocate a central role for universals without explicitly evaluating frequency (Demuth 1995; Fikkert 1994; Gerken 1996; Gnanadesikan 2004; Pater & Barlow 2003). Still others find that frequency is only partially compatible with development (Kirk & Demuth 2005; Stites, Demuth & Kirk 2004; Stoel-Gammon 1998). For example, Kirk & Demuth (2005) found that the frequency of initial as opposed to final clusters in English was predictive of production accuracy on these classes overall but that frequency did not correlate well with accuracy of particular consonant clusters. Similarly, Stoel-Gammon (1998) found that frequency predicts the overall shapes of children's early words but that certain patterns, such as word-initial /b/, were overrepresented in the children's early words as compared to the input. A final group of studies advocates explanations crucially incorporating both syllable structure universals and frequency (Boersma & Levelt 2000; Edwards & Beckman 2008; Jarosz 2010, 2011; Levelt, Schiller & Levelt 2000; Roark & Demuth 2000; Tessier 2009). For example, Edwards and Beckman (2008) show that the inherent phonetic difficulty of certain sound classes, such as affricates, can be modulated by language-particular frequency. However, accounts vary substantially in how these factors and their interaction are implemented.

1.2. *Goals of the current study*

Thus, there are many findings supporting the role of universals, others that support input frequency, and still others that suggest crucial roles for both factors. Distinguishing among these hypotheses is difficult for a number of reasons. One challenge is that there is a strong correlation between frequency and universals: the same structures that are rare cross-linguistically tend to be rare within individual languages (Zamuner, Gerken & Hammond 2005). For this reason, many developmental findings

consistent with an account in terms of universals may also be consistent with an input-based account. Thus, it is not clear to what extent markedness or other biases are *necessary* to explain particular findings even when those findings are compatible with universals. Since there is abundant evidence that learners are sensitive to frequency, and frequency is highly correlated with markedness, it is important to determine how much input statistics can explain about development without reference to other biases. A second challenge is that investigations of input frequency in phonological learning have relied on a wide range of input statistics (see discussion in section 0), and it is not clear how much different assumptions about frequency affect the predictions of input-based models. A final challenge in evaluating these hypotheses is a lack of studies that evaluate the systematic predictions of models derived from the language input for a wide range of structures. Of particular relevance is the fact that the vast majority of studies on phonotactic probability have relied on segment-level statistics to study segment-level production or perception. However, a number of studies have found that statistics over coarser-grained representations play a role in development at higher levels of organization such as syllable structure (Vihman 1993; Levelt, Schiller & Levelt 2000; Demuth & McCullough 2009; Roark & Demuth 2000). As a result, it is difficult to estimate the utility of any particular frequency model for phonological learning more generally, and it is unclear how input statistics calculated at lower levels may interact with development at higher levels and vice versa.

In this study, we address these challenges by building explicit statistical models of the input for a range of frequency measures and comparing their predictive capacities on a wide range of syllable structures in Polish. The syllable plays a central role in phonological theory, and there is an extensive literature on the universal principles underlying the organization of syllables (for reviews, see, e.g., Blevins 1995; Zec 2007). This makes syllable structure a critical test case for investigating input-based models of learning and determining how much of development can be predicted based on language-internal statistics. Polish syllable structure provides a particularly interesting arena for investigating these questions because Polish permits syllables of significant complexity (Rubach & Booij 1990). The large range of marked syllable types provides the opportunity to test predictions across structures varying widely in their complexity. Since many marked structures are relatively frequent in the input (e.g., complex onsets, a wide range of coronal fricatives and affricates, and sonority plateaus), Polish is an ideal testing ground for hypotheses that seek to explain language development on the basis of input statistics alone.

Our results indicate that different formulations of frequency can make dramatically different predictions, as can different hypotheses about the representational units over which statistics are calculated. We also find that some frequency models perform surprisingly well in predicting major effects in production accuracy but that the most successful frequency measure depends on the representational level examined. Overall, the most successful measures of frequency reference multiple abstract units of phonological representation, indicating that phonological development is sensitive to both frequency and phonological structure.

Our focus is on comparing and pushing input-based models to their limits, but we also identify and discuss departures between the models' predictions and children's production patterns. These discrepancies indicate that input statistics, even ones that reference abstract structure, do not fully account for development, suggesting a crucial role for biases. The discrepancies have many possible sources, such as articulatory or perceptual difficulty, universal grammar, domain-general learning biases, and domain-specific inductive biases. Our approach is to allow these discrepancies to emerge from the analysis to highlight concrete areas where some sort of biases are needed. Although this portion of our results is largely descriptive and cannot definitively differentiate between possible sources of these biases, we consider the potential role of a handful of universal markedness laws and articulatory constraints that have been widely discussed in the literature and are of particular relevance to syllable structure development.

2. The data and processing

This section motivates the models of frequency we investigate (section 2.1) and describes the Polish corpus of spontaneous child productions that we use for all analyses (section 2.2). We then present our methodology for automatically aligning the child productions with target pronunciations in order to calculate the children's production accuracy for various structures and representational levels (section 2.3). Finally, section 2.4 presents the methods for calculating the frequency measures from a sample of child-directed speech spoken to these children.

2.1. *Frequency models*

Since our primary goal is to determine the extent to which input-based models are capable of predicting acquisition, we focus our attention on seven measures of frequency that provide relatively broad coverage of the kinds of input statistics that have been implicated most strongly in studies of syllable structure development in previous work. Another goal is to determine the extent to which predictions vary as a function of input statistic, so we also aim at diversity in coverage: we identify two dimensions along which most statistics used in previous work vary—type vs. token and level of granularity—and we systematically vary these options. By considering this targeted and relatively broad range of statistics, we aim to give input-based models a strong chance at succeeding overall while simultaneously investigating the consequences of these choices. Our conclusions about the limits of input-based models nonetheless depend on these choices and hold to the extent that these input statistics can jointly represent the significant statistical patterns that are predictive of acquisition.

Previous studies have found both type and token frequencies to be predictive of phonological development. *Type frequency* refers to the incidence of a given structure in the lexicon, while *token frequencies* measure the raw level of exposure a language learner has to the structure in the linguistic input. For example, the type frequency of the onset [st] is the number of distinct lexical items that begins with [st], while the token frequency of [st] is the number of word tokens that begin with [st]. The choice of type versus token frequency has serious consequences: only the latter is overwhelmingly sensitive to the phonological characteristics of common words. Numerous studies have found type frequency to be predictive of production accuracy in children (Edwards & Beckman 2008; Edwards, Beckman & Munson 2004; Ingram 1988; Munson 2001; Richtsmeier, Gerken, & Ohala 2009; Archer & Curtin 2011), and several of these studies argue explicitly that type frequency is more predictive of phonological generalization than token frequency (Archer & Curtin 2011; Edwards & Beckman 2008). For example, Edwards, Beckman & Munson (2004) posit that storing words in the lexicon enables learners to form abstract generalizations about words' internal structure that can be applied to form novel combinations. In support of this conclusion, they also find that production accuracy in nonwords improves with a larger lexicon. Consistent with this, a number of studies examining adult speakers' phonological and morphological generalization have found type frequencies to be superior to token frequencies (Albright & Hayes 2003; Becker, Ketrez & Nevins 2011; Bybee 1995; Albright 2009; Richtsmeier 2011; Hayes & Wilson 2008). On the other hand, other studies of phonological development, primarily examining spontaneous production, have supported a central role for token frequency (Kirk & Demuth 2005; Roark & Demuth 2000; Stites, Demuth & Kirk 2004; Zamuner, Gerken & Hammond 2004, 2005; Levelt, Schiller & Levelt 2000). Finally, a number of authors recognize the potential for a combined role and interaction of type and token frequency (Bybee 1995; Pierrehumbert 2003; Richtsmeier, Gerken, & Ohala 2011). For example, Pierrehumbert (2003) posits that both phonetic token variability and type variability are critical to forming phonological abstractions at various levels. Manipulating frequency of exposure experimentally, Richtsmeier, Gerken & Ohala (2011) show that both are necessary for children to form generalizations about nonword acceptability. However, no studies we are aware of systematically compare predictions of type and token frequency for spontaneous production accuracy in children. We consider both type and token frequency in hopes of shedding some light on how these input statistics may function in this context.

As discussed earlier, previous work also varies widely in the representational level over which input statistics are calculated. While the vast majority of studies examine segment-level statistics, studies examining development of syllable structure have also identified a role for statistics computed over abstract phonological classes, especially sonority (Kirk & Demuth 2005; Stites, Demuth & Kirk 2004; Demuth & McCullough 2009) and CV (Boersma & Levelt 2000; Jarosz 2010; Kirk & Demuth 2005; Levelt, Schiller & Levelt 2000; Roark & Demuth 2000). The CV and sonority classes are particularly relevant to syllable structure. The CV level is useful for characterizing basic syllable shapes, and it is at this level that implicational markedness laws about basic syllable shapes are stated. There is also a large literature on the importance of sonority for the organization of syllables and subsyllabic constituents (see Zec 2007 for a review). It is over the sonority level that standard principles about the well-formedness of various syllabic constituents, including nuclei, onsets, and rhymes, have been stated (Clements 1990; Selkirk 1984). These include cross-linguistic preferences for low-sonority onsets, high-sonority rhymes, and sequences with sonority rises in onset and falls in codas. Since CV and sonority-level representations play such an important role in characterizing typological preferences about syllable shapes and subsyllabic sequencing, it is important to examine how development of syllable structure may be sensitive to these abstract representations. In addition to statistics calculated at a segmental level, we therefore also consider statistics computed over CV and sonority levels of representation to determine what roles the frequency of abstract classes and granularity play in the development of higher levels of phonological organization. We utilize a standard sonority hierarchy that differentiates six levels: stops, fricatives, nasals, liquids, glides, and vowels. We are particularly interested in examining whether syllable structure development can be understood solely in terms of low-level segmental regularities or whether sensitivity to abstract classes is required. This question is also motivated by recent findings that sensitivity to phonological classes is crucial to modeling both adults' (Albright 2009; Finley & Badecker 2009; Daland et al. 2011) and children's (Cristià & Seidl 2008) phonological generalization.

In sum, this work explores a range of input statistics, varying along two dimensions. We consider input frequencies calculated from child-directed speech at three representational levels: CV, sonority, and segment. We consider both type and token statistics calculated at each of these three levels, yielding a total of six frequency measures. In order to evaluate the extent to which these levels of phonological representation are important for modeling acquisition, we also consider and compare to these measures a structure-blind input statistic: word frequency. We refer to this statistic as structure-blind since it does not make reference to any phonological structure or units below the word level. As discussed earlier, we model the acquisition of a wide range of syllable structures, allowing for the various statistics to be evaluated across many levels of granularity and complexity. To provide as comprehensive a picture of the development of syllable structure as possible, we analyze the children's spontaneous productions across the same three levels of granularity. It is possible that the best frequency measures vary by unit of analysis. For example, it could be that the development of fricatives and stops overall is best predicted by sonority-level statistics but that frequency of particular coda consonants is the best predictor of the acquisition of these specific codas. Ultimately, a successful input-based theory must encode a single hypothesis about the formulation of frequency that makes accurate predictions when tested across a wide range of phonological structures. We explore these issues by evaluating each of the seven measures of frequency at each level of analysis.

2.2. *Data and participants*

The data come from a novel extension of the Weist corpus of child Polish (Weist & Witkowska-Stadnik 1986; Weist et al. 1984). The existing corpus includes orthographic transcripts and audio recordings of the spontaneous speech of four typically developing children interacting with their caregivers and is publicly available via CHILDES (MacWhinney 2000). Our analyses are based on newly constructed phonetic transcriptions of this corpus, as described below, which will also be made publicly available as part of the CHILDES database.

We produced a phonetic transcription of the children's productions using broad phonetic transcription with the help of the open-source Phon software (Rose et al. 2006), building on the

preliminary phonetic transcripts of the Weist corpus produced by Jarosz (2010). The orthographic transcripts in the Weist corpus were used as the basis for creating phonetic transcriptions of the children's target pronunciations, and the audio recordings were used to phonetically transcribe the children's actual productions and align them with the target transcriptions word by word. Target pronunciations were citation form transcriptions except that they took into account voicing assimilation that spans word boundaries in Polish (Gussman 1992): Word-final sequences of voiceless obstruents, which in isolation are always voiceless due to final devoicing, are voiced when immediately followed by voiced obstruents or sonorants in connected speech. The transcription of all child productions was first performed independently by two transcribers trained in phonetic transcription, at least one of whom was a native speaker of Polish. Then two Polish speakers trained in phonetic transcription worked together to create a consensus transcription of all productions, relying on a third phonetically trained native speaker of Polish to adjudicate in cases when agreement could not be reached. The resulting corpus includes phonetic transcriptions of the children's spontaneous productions in all the available audio files for the Weist corpus, providing word-by-word alignment of target pronunciations and actual pronunciations in all utterances.

The analyses reported here use all available phonetic transcripts for Bartosz (6 transcripts), Marta (3 transcripts), and Kubuś (7 transcripts). The full phonetic corpus includes 19 transcripts for Wawrzon, spanning ages 2;02 through 3;02. We include only the earliest eight sessions (ages 2;02 through 2;06) in our analyses since the data for the other children does not include sessions beyond the age of 2;06, and we do not want our data to overrepresent any single child.

We extracted from all the children's transcripts the phonetic transcriptions of their productions and intended targets for each word token with the following exceptions. We did not include utterances involving onomatopoeia, wordplay, and child-specific forms, as we could not be certain of the intended targets in these cases. We excluded word tokens that were incomplete, wholly or partially unintelligible, continuations of adult prompts, or repetitions or memorized passages. These cases were systematically labeled as such during the phonetic transcription of the corpus. Tokens of the third person singular copula *jest* [jɛst] were excluded because the pronunciation of the copula is highly variable in casual adult speech, often being reduced to [jɛ], [jɛs], or even [js], making it difficult to judge accuracy on this highly frequent function word. Finally, we excluded the phonological words formed with the consonantal proclitics *z* and *w* because we could not be certain that children would treat the word onsets created by these proclitics (e.g., *z wilkiem* [zvilkjɛm] 'with the wolf') comparably to other word onsets.

Together, these restrictions resulted in a corpus of roughly 9,000 words (as shown in Table 1) that was submitted for subsequent processing. The age ranges and sizes for the subcorpora corresponding to each of the children are shown in Table 1. Relative to previous studies examining phonological and prosodic development in transcribed, spontaneous child speech, this corpus is moderately sized. Some prior studies focused on particular aspects of phonological development such as clusters (Kirk & Demuth 2005; Lleó & Prinz 1996), fricatives (McAllister Byun 2011), and codas (Stites, Demuth & Kirk 2004) have examined corpora smaller than 2,000 words. Corpora of 2,000 to 5,000 word tokens have been used to study development of basic syllable structure (Rose 2000; Salidis & Johnson 1997) and to analyze sonority effects in cluster acquisition (Demuth & McCullough 2009). Finally, larger corpora (greater than 20,000 utterances) have been used to develop more comprehensive analyses of children's phonological systems and to study variation across children (Compton & Streeter 1977; Fikkert 1994; Levelt, Schiller & Levelt 2000; Pater 1997).

2.3. *Segmental alignment and accuracy coding*

To compute the children's accuracy, their actual productions were automatically aligned segment by segment with the intended targets. This was accomplished using a slightly adapted variant of ALINE (Kondrak 2000), a dynamic programming algorithm that incorporates knowledge of phonological features in order to maximize similarity between segments. We made minor adjustments to the set of phonological features and their weights in order to allow the algorithm to encode and compare the similarity of all the

Table 1. Corpus Characteristics.

	Sessions	*Ages*	*Words*
Bartosz	6	1;07–1;11	2192
Kubuś	7	2;01–2;06	2772
Marta	3	1;07–1;08	1218
Wawrzon	8	2;02–2;06	2913
Corpus	24	1;07–2;06	9095

segments in Polish. The resulting alignments were checked by all authors to make sure that the algorithm dealt appropriately with a wide range of structures and production errors. Examples are given in (1). Alignment makes it possible to encode, detect, and localize a variety of mismatches between the produced forms and the target, including deletions (1a–e), insertions (1f), and substitutions involving similar sounds (1b–h).

(1) Examples of aligned targets and productions from Kubuś (2;01):

a. 'first'

n	a	j	p	j	ɛ	r	f
n	a	–	p	j	ɛ	–	f

b. 'little bird'

p	t	a	ʂ	ɛ	k	
–	t	a	ɕ	ɛ	k	

c. 'little snowman'

b	a	w	v	a	n	ɛ	k
b	ɔ	w	v	a	n	ɛ	–

d. 'they (masc.) fell over'

p	ʂ	ɛ	v	r	u	t͡ɕ	i	l	i	ɕ	ɛ	w̃	
p	–	ɛ	v	l	u	t	i	l	i	ɕ	ɛ	–	

e. 'adventure (acc.)'

p	ʂ	ɨ	g	ɔ	d	ɛ	w̃
–	s	ɨ	g	ɔ	d	ɛ	–

f. 'little animals'

z	–	v	j	ɛ	ʐ̩	ɔ	n	t	k	a				
ʑ	ɨ	v	j	ɛ	ʑ	a	n	t	k	a				

g. 'watch'

z	ɛ	g	a	r	ɛ	k
d	ɨ	g	a	l	ɛ	k

h. 'I will show'

p	ɔ	k	a	ʐ̩	ɛ	w̃	
p	ɔ	k	a	ʑ	ɛ	m	

Once the targets and productions were aligned, the word margins and their alignments—shaded in (1)—were extracted. We focus our analyses of syllable structure development on initial and final word margins because this avoids making potentially problematic assumptions about the syllabification of medial clusters, which is controversial and highly variable even in adult Polish (Rubach & Booij 1990). We define *initial word margins* as the sequence of zero or more consonants preceding the first vowel in the target word and *final word margins* as the string of zero or more consonants following the last vowel of the target word. We include null word onsets and codas for two reasons. First, one of our goals is to compare development and input statistics across a range of complexities. Second, implicational markedness laws about basic syllable shapes reference null onsets and codas, and, as discussed in the introduction, previous work has shown that development is sensitive to these principles. To determine accuracy on each word margin, we consulted the alignments to find the phonological string corresponding to each target word margin. Specifically, the initial word margin

Table 2. Sample Accuracy Coding for Example Initial Margins.

Segment			*Sonority*			*CV*		
Target	*Produced*	*Accuracy*	*Target*	*Produced*	*Accuracy*	*Target*	*Produced*	*Accuracy*
#ps	#–s	0	#PF	#–F	0	#CC	#–C	0
#z	#d	0	#F	#P	0	#C	#C	1
#ʒ	#ʒ	0	#F	#F	1	#C	#C	1
#v	#v	1	#F	#F	1	#C	#C	1

for the child productions was defined as the string up to but not including the vowel aligned to the first vowel of the target word. Similarly, the final word margin in the children's productions was defined as the string of phonemes starting after the segment aligned to the target's last vowel. For example, in (1a) the initial word margin aligns target [#n] with the correct actual production [#n], where '#' is used to highlight word boundaries for readability. The final word margin for the same form aligns target [rf#] with reduced [f#]. In (1f) the initial target [#zvj] is aligned with actual [#ʒɨvj], while the null final margin [a#] (for readability, we retain the vowel for null initial and final margins) is aligned with the identical margin [a#] of the actual production. In this way, each word token contributes information about the production of one initial margin and one final margin.

The extracted word margins were then evaluated for accuracy at the segmental, sonority, and CV levels. At the *CV level*, the only phonological contrast was between consonants (C) and vowels (V). At the *sonority level*, consonants were subdivided into plosives (including affricates) (P), fricatives (F), nasals (N), liquids (L), and glides (G), alongside the class of vowels (V). The *segmental level* of representation maintained all the degrees of phonological contrast found in the IPA transcriptions for consonants, but all vowels were labeled simply as V (encoding the presence or absence of a vowel rather than their identity). At each of these levels of representation, a word margin was coded as correct (1) if the produced word margin was identical to the target word margin and incorrect (0) otherwise.[1] As shown in Table 2, accuracy at the segmental level entails accuracy at the sonority level, but it is possible for a word margin to be accurate at the sonority level and not the segmental level. Likewise, accuracy at the sonority level entails accuracy at the CV level, but it is possible for a word margin to be accurate at the CV level and not the sonority level. These three levels of accuracy coding allow us to differentiate between certain strategies children may use to avoid phonological structures: accuracy at the segmental level is sensitive to any mismatch between the target and child's production, mismatches at the sonority level ignore phoneme substitutions within the same sonority class, and finally, CV-level mismatches primarily signal insertions and deletions.

2.4. *Calculating frequency from child-directed speech*

Corpus analysis of language acquisition makes it possible to study the relationship between children's language development and the properties of the language input those children received. In addition to transcribing the child productions, the Weist corpus provides orthographic transcriptions of the adult speech directed at the children. This provides a sizeable sample of child-directed speech from which properties of the input can be estimated. We used the transcripts of child-directed speech in the corpus to estimate the various frequency measures we evaluate in the second part of this article. We extracted the child-directed utterances spoken by the primary caregivers to create the corpus of child-directed speech, which results in a corpus of 9,362 utterances, 34,122 word tokens, and 5,030 word types.[2]

[1]The only exception is that word-final productions of orthographic *ę* were coded as correct at all levels of representation when produced as either [ɛ] or [ɛw̃] to reflect the optionality of the nasal offglide in this context.

[2]The statistical properties of child-directed speech corpora of comparable size in a variety of languages have been extensively studied in the domain of word segmentation (Batchelder 2002; Blanchard, Heinz & Golinkoff 2010; Brent & Cartwright 1996; Goldwater, Griffiths & Johnson 2009; Jarosz & Johnson 2013; Johnson 2008; Venkataraman 2001) and to a lesser extent in work on phonological development (Zamuner, Gerken & Hammond 2005). Somewhat larger child-directed speech corpora (80–120k word tokens) were examined in a handful of previous studies on phonological development (Demuth & McCullough 2009; Kirk & Demuth 2005; Levelt & van de Vijver 2004; Roark & Demuth 2000).

Table 3. Example Estimates of the Six Frequency Measures.

		Segment			*Sonority*			*CV*	
		Token	*Type*		*Token*	*Type*		*Token*	*Type*
#ml	*f*(#ml)	10	6	*f*(#NLV)	13	9	*f*(#CC)	5209	1456
nt#	*f*(nt#)	39	13	*f*(NP#)	133	50	*f*(CC#)	1506	162

This sample of orthographically transcribed child-directed speech was then transcribed phonetically using automatic methods based on standard pronunciation. Similar automatic methods have been used to construct many other child-directed speech corpora in a variety of languages (Batchelder 2002; Blanchard, Heinz & Golinkoff 2010; Brent & Cartwright 1996; Gambell & Yang 2006; Goldwater, Griffiths & Johnson 2009; Hockema 2006; Jarosz & Johnson 2013; Johnson 2008; Roark & Demuth 2000; Venkataraman 2001; Zamuner, Gerken & Hammond 2004, 2005). We were careful to use the same transcription conventions as those we used for transcribing the child productions. In addition to using the same alphabet and level of phonetic encoding for both corpora, we encoded voicing assimilation processes that apply across words in Polish as described earlier. Since the phonetic transcriptions of the children's targets and actual productions encode contextual variation of this sort, and since this variation occurs in connected speech, we applied regressive voicing assimilation to word-final obstruent clusters followed by voiced obstruents or sonorants in the same utterance.

We used the resulting phonetic transcripts to estimate the frequencies of both initial and final word margins using each of the six measures of frequency discussed earlier. Both type and token frequencies were calculated for all word margins at each of the three levels of representation. This yields six different measures of frequency for each word margin: CV type, CV token, sonority type, sonority token, segment type, and segment token. Example frequencies for word-initial [#ml] and word-final [nt#] are shown in Table 3. As the frequencies in Table 3 illustrate, the representational level can dramatically affect the relative frequencies of two structures. At the CV level, there are nearly nine times as many words beginning with biconsonantal clusters [#CC] as there are words ending in biconsonantal clusters [CC#]. However, there are four times *fewer* tokens of initial [#ml] than there are of final [nt#]. If acquisition is driven by segment- or sonority-level frequencies, earlier acquisition of [nt#] is expected, while the reverse is expected if CV-level frequencies guide development. The level of representation is important for both frequency measures and calculation of accuracy. Based on CV-level frequencies, one might expect earlier overall development of initial clusters; however, if there are many rare initial clusters, low segment-level frequencies for clusters such as [#ml] could bring down average performance on initial clusters. Our analysis investigates all of these possibilities by considering each of these six frequency measures' predictions for development at each of the three levels. In order to investigate the possible influence of lexical frequency on development, we also use the sample of child-directed speech to estimate word frequency.

3. Acquisition analyses

Our approach to analyzing the production data follows a number of recent studies utilizing regression modeling strategies to analyze spontaneous corpus data (Bane, Graff & Sonderegger 2014; Jaeger 2010; Jarosz & Johnson 2013; Roland, Elman & Ferreira 2006). We employ logistic regression models, created using the glm function in R (R Development Core Team 2008), to analyze the children's production accuracy. We present two sets of acquisition analyses. The first establishes the relative accuracies across all margin types while controlling for a number of potentially confounding variables by building regression models with accuracy as the dependent variable and the margin types and control variables as independent variables. We build three such models, one for each level of representation. We then systematically simplify these models to establish which accuracy differences are meaningful and to describe the overall developmental pattern in more general terms. In addition to establishing the empirical results, these analyses serve as a reference for the subsequent evaluations of the frequency models.

3.1. *Relative production accuracy of word margins*

3.1.1. *Method*

The principal goal of this section is to determine the relative accuracy with which children produce various margin types on the basis of the spontaneous production data described earlier. An important advantage of the present methodology relying on spontaneous speech data is ecological validity and an ability to examine the relationship between language acquisition and the input. However, since these are spontaneous data, there are a number of factors in addition to the factor of interest (syllable margin types) that may affect production accuracy and which could not be systematically controlled during data elicitation. Instead, we rely on a set of control variables and regression modeling to control statistically for as many potential confounds as possible. By incorporating as many control variables into the models as feasible, we can be relatively confident that any significant differences detected in the production accuracies of two different margin types are in fact due to the factor of interest.

All models treat PRODUCTION ACCURACY as the dependent variable and MARGIN TYPE as a predictor. In order to control for potential confounds, we also considered six independent variables for inclusion in the models. We include AGE (a continuous variable measured in months) as a possible control variable in order to account for developmental progression over time. To account for individual differences in overall production accuracy, we consider SUBJECT as a four-level factor.[3] We also control for the possibility that mere experience with a particular word form affects the accuracy with which its word margins are produced by including (log) WORD FREQUENCY as a candidate predictor.[4] Finally, we consider three independent variables to control for potential effects of phonological and prosodic context of the target words within which the word margins are found. We include WORD LENGTH (counted in number of syllables), VOWEL (a six-level factor indicating the identity of the vowel adjacent to the margin), and STRESS (a binary variable indicating that the syllable the margin belongs to carries primary stress).[5] Effects of prosodic position and prominence have been repeatedly observed in child production studies (see, e.g., Fikkert 1994; Demuth 1995), and Edwards & Beckman (2008) found effects specific to particular vocalic contexts. To summarize, the six control variables and the associated number of parameters with each are: AGE (1), SUBJECT (3), WORD FREQUENCY (1), WORD LENGTH (1), VOWEL (5), and STRESS (1), for a total of 12 parameters.

Successful analysis of unbalanced corpus data requires careful attention to model evaluation and modeling assumptions (Jaeger 2010). Our primary factor of interest is the syllable margin type. Including this factor in the model contributes one parameter for each distinct margin type, and the number of margin types varies by representational level. At the CV level there are only six types: (#V, #C, #CC, V#, C#, CC#),[6] but at the sonority and segment levels, there are dozens of possible margins. To prevent over-fitting and to allow for the coefficients corresponding to each margin type to be reliably estimated, we collapsed levels for any margin type that had fewer than five observations in each cell (accurate and inaccurate). This means only those margin types that had greater than five accurate productions and greater than five inaccurate productions were retained as individual levels. We created a dummy level OTHER where the remaining, infrequently attempted margins were collapsed and retained for analysis. We did this separately for each level of representation. At the CV level, all margin types were adequately represented and retained as individual categories. At the sonority level, 25 margin types were

[3]Although subject-level effects are often included in mixed effects regression models as random rather than fixed effects (see, e.g., Jaeger 2008), the subject factor in these data has only four levels and hence does not provide sufficient information to estimate group-level variation (Gelman & Hill 2006:247). We therefore consider it for inclusion in the model as a fixed effect.

[4]Because log(0) is undefined and because a handful of children's targets do not occur in the child-directed speech corpus, we use log(frequency + 1) for all of the frequency predictors.

[5]Primary stress was assigned automatically to penultimate syllables according to the regular stress pattern of Polish lexical stress (Rubach & Booij 1985).

[6]Although Polish allows longer sequences of consonants, these patterns are rare and are seldom attempted by the children. Due to lack of reliable data, we excluded triconsonantal and longer margins from all analyses. There were only 170 word tokens with three or more initial consonants, and only 8 word tokens with three or more final consonants. This resulted in a reduction of less than 2% of word margins in the corpus.

retained as individual categories, and 14 infrequent margin types were collapsed into the category OTHER. These 14 infrequent margin types together accounted for only 0.33% of the children's production data. At the segment level, there were sufficient data to retain 77 margin types as individual categories. 110 infrequent margin types were collapsed into the segment-level OTHER category, which constituted only 3.2% of the data.

The children's overall accuracy on each of the margin types at each of the three levels of representation is summarized in Table 4. The table lists the total number of tokens of each type (*n*), and the overall accuracy on that type (%). Sonority-level margins are grouped by CV level of representation, and segment-level margins are grouped first by CV level, and then by sonority level within CV-level groups. As is evident from the table, the individual levels retained for analysis are diverse both in terms of the range of phonological structures represented and in terms of the range of average accuracy. For example, Polish is well known for allowing marked clusters, and the data retained for analysis includes a wide range of onset clusters, including relatively unmarked obstruent-sonorant combinations as well as marked sonority plateaus involving sequences of obstruents (#PP, #PF, #FP, #FF).

We fit logistic regression models with these control variables and MARGIN TYPE as predictors, one for each level of representation. The level of representation determined which set of margin types and which accuracy measure was used. For example, for the CV-level model, the set of six CV-level margin types (#V, #C, #CC, V#, C#, CC#) was used, and CV-level accuracy coding was used. This way, the CV-level

Table 4. Overall Accuracy on Margin Types (MTs) in Production Data.

MT	*n*	%	*MT*	*n*	%	*MT*	*n*	%	*MT*	*n*	%
CV Level											
#V	681	93.4	#C	6959	93.5	#CC	1354	56.9			
V#	5395	92.2	C#	3510	82.1	CC#	182	35.2			
Sonority Level			Segment Level								
#V	681	93.4	#V	681	93.4	m#	548	83.0	#pj	38	73.7
V#	5395	92.2	V#	5395	92.2	n#	166	81.3	#pw	29	69.0
						ɲ#	25	36.0	#bj	17	52.9
#P	3526	93.8	#b	366	93.2	t#	45	82.2	#gw	13	38.5
#N	1573	89.2	#t	1390	91.2	k#	468	74.8	#dr	55	63.6
#G	556	82.2	#d	359	89.1	p#	22	59.1	#tr	52	55.8
#L	275	78.5	#p	722	88.5	t͡ɕ#	171	56.7	#kl	24	54.2
#F	1029	77.3	#k	295	76.3	g#	60	48.3	#kr	36	33.3
N#	739	83.2	#t͡ɕ	73	71.2	d#	23	47.8	#pr	47	21.3
P#	886	74.8	#g	41	68.3	t͡ʂ#	30	23.3	#gr	37	13.5
L#	85	71.8	#d͡z	92	50.0	t͡s#	34	20.6	#zd	14	64.3
G#	1364	71.0	#t͡s	118	44.9	l#	25	80.0	#zb	13	46.2
F#	436	65.4	#t͡ʂ	67	23.9	r#	60	38.3	#sp	27	33.3
#NG	33	84.8	#n	251	87.3				#st	32	31.3
#FN	34	76.5	#m	656	86.7	w̃#	773	78.3	#sk	20	25.0
#FG	189	72.0	#ɲ	666	78.2	j#	244	73.4	#xt͡s	125	6.4
#PG	111	60.4	#j	529	82.6	w#	347	50.7	#sx	19	42.1
#PL	279	59.5	#w	27	66.7	ɕ#	85	78.8	#kɕ	15	40.0
#FF	52	51.9	#l	117	76.1	f#	15	66.7	#pʂ	106	7.5
#PF	193	50.3	#r	158	49.4	z#	12	41.7	#kt	27	37.0
#FL	58	39.7				ʑ#	23	34.8	#gd͡z	77	31.2
#PP	118	38.1	#v	276	89.1	s#	53	26.4	ɲt͡ɕ#	15	33.3
#FP	267	31.8	#x	64	89.1	ʂ#	165	24.2	nt#	16	31.3
other	60	38.3	#ɕ	228	74.6	ʐ#	55	18.2	ɕt͡ɕ#	46	10.9
			#z	231	49.8	#vw	52	82.7	other	573	38.0
NP#	62	50.0	#s	121	43.8	#vj	67	71.6			
PF#	21	28.6	#ʂ	24	20.8	#zw	18	66.7			
FP#	59	22.0	#ʐ	59	15.3	#sw	16	62.5			

Table 5. Regression Parameters and Upper Limits.

	Parameters				
	Control	*Margin Type*	*Total*	*Upper Limit*	*Minority Outcome*
CV	12	6	18	112	2244
Sonority	12	26	38	146	2927
Segment	12	78	90	199	3975

models have the desired interpretation: low accuracy on a CV-level type (e.g., #CC) means that this type was produced incorrectly by the children as a distinct CV-level type (e.g., #C); whereas high accuracy means children produced this type with the correct CV-level structure (although they may have made some substitutions within the class of consonants). Likewise, the sonority-level model is indicative of children's ability to correctly produce margins with the target sonority levels, and the segment-level model considers segment-level accuracy across all segment-level margin types.

To summarize, we begin by considering three logistic regression models with the full control structure discussed earlier. We have collapsed some of these levels in order to prevent over-fitting and allow for reliable estimates of the coefficients. As a further check for potential over-fitting, we compare the number of parameters in each model to the total number of observations with the minority outcome (in all cases this is the number of inaccurate productions). The rule of thumb for logistic regression models is that the number of parameters should be smaller than the number of observations with the minority outcome divided by 10 or 20 (Harrell 2001:61). The numbers for the models considered here and the corresponding upper limits are summarized in Table 5. For all models, the number of parameters is well below the safe upper limits recommended by Harrell.

3.1.2. *Results*

We first consider the CV-level regression. After inclusion of all six control predictors, the factor of interest, MARGIN TYPE, is highly significant as a whole, as shown by a nested model comparison using a likelihood ratio test, $\chi^2(5) = 1338.0$, $p < .0001$. This means that different CV margin types exhibit different accuracy patterns, which we explore further below. As for the control predictors, WORD FREQUENCY and SUBJECT are not significant ($p > .1$) given the other predictors in the model. In a nested model comparison evaluated based on likelihood ratio tests, neither SUBJECT, $\chi^2(3) = 4.76$, $p = 0.19$, nor WORD FREQUENCY contributed to model fit, $\chi^2(1) = 0.25$, $p = .62$. All other predictors were found to be significant in the CV-level regression, and these are summarized in Table 6. Not surprisingly, AGE is positively associated with accuracy: older children are more accurate ($\beta = 0.05$, $z = 7.6$, $p < .0001$). STRESS is also positively associated with accuracy: margins in stressed syllables are produced more accurately ($\beta = 0.29$, $z = 3.8$, $p < .001$). On the other hand, WORD LENGTH is negatively associated with accuracy: margins in longer words are produced less accurately ($\beta = -0.14$, $z = -5.1$, $p < .0001$). Finally, the identity of the margin-adjacent vowel (the reference category for this factor is the vowel [a]) affects accuracy as well, $\chi^2(5) = 15.7$, $p < .01$.

Returning to the factor of interest, MARGIN TYPE, the regression model indicates that the predictor as a whole is meaningful, and it also provides information about the relative accuracies of each margin type after adjusting for the control variables. This information is reflected by the coefficients associated with each margin type shown in Table 6, and the standard errors provide confidence estimates for these coefficients.[7] To provide a more readable and interpretable representation of these results, the coefficients and standard error ranges for MARGIN TYPE from Table 6 have been translated into predicted accuracy along with 95% confidence intervals using the logistic function $(1 + e^{-x})^{-1}$. This is shown in the upper left portion of Figure 1. These predicted accuracies reflect the variation the model attributes to the margin types themselves after adjusting for the control predictors. As this graph shows, the model captures significant variability in the production accuracy of the CV-level margin types, ranging from less than 20% accuracy on CC# to close to 90% accuracy on #C. The confidence intervals give a sense of the reliability of these

[7]The regression was fit with no intercept so that each margin type could receive its own coefficient and error range.

Table 6. Regression Model for Individual CV Margin Types.

Predictor	*β*	*Standard Error*	*z*	*p*
MARGIN TYPE				
#CC	− 0.64	0.18	− 3.6	<0.001
#C	1.78	0.17	10.4	<0.0001
#V	1.77	0.23	7.6	<0.0001
V#	1.69	0.17	10.0	<0.0001
C#	0.71	0.17	4.1	<0.0001
CC#	− 1.49	0.23	− 6.5	<0.0001
WORD LENGTH	− 0.14	0.03	− 5.1	<0.0001
STRESS	0.29	0.08	3.8	<0.001
AGE	0.05	0.01	7.6	<0.0001
VOWEL				
ɛ	− 0.10	0.06	− 1.5	<0.13
i	− 0.09	0.10	− 0.9	>0.35
ɨ	− 0.27	0.11	− 2.5	<0.013
u	− 0.24	0.09	− 2.7	<0.01
ɔ	0.05	0.07	0.7	>0.50

relative accuracy comparisons. The analyses in the next section will provide another way to evaluate relative accuracy among the margin types.

We perform an analogous analysis at the sonority level.[8] At the sonority level, the factor of interest, MARGIN TYPE, is highly significant as a whole after inclusion of all six control predictors, $\chi^2(25) = 1780.0$, $p < .0001$. Different sonority margin types exhibit distinct accuracy patterns after adjusting for control predictors. The lower left portion of Figure 1 shows the predicted accuracy by MARGIN TYPE at the sonority level. As with the CV level, there is a wide range of performance on margin types. SUBJECT is not significant, $\chi^2(3) = 1.67$, $p > .64$, given the other predictors in the model, but all the remaining control predictors were found to be significant. WORD FREQUENCY is positively associated with accuracy: there is higher accuracy on more frequent words ($\beta = 0.04$, $z = 2.9$, $p < .01$). Once again, AGE is positively associated with accuracy: older children are more accurate ($\beta = 0.04$, $z = 7.7$, $p < .0001$. STRESS is once again positively associated with accuracy: margins in stressed syllables are produced more accurately ($\beta = 0.24$, $z = 3.3$, $p < .001$). As before, WORD LENGTH is negatively associated with accuracy: margins in longer words are produced less accurately ($\beta = -0.09$, $z = -2.6$, $p < .01$). Finally, the identity of the margin-adjacent vowel affects accuracy, $\chi^2(5) = 30.5$, $p < .0001$.

At the segment level, results are similar. The primary factor of interest, MARGIN TYPE, is highly significant as a whole after inclusion of all six control predictors, $\chi^2(77) = 3662.4$, $p < .0001$. The right portion of Figure 1 shows the predicted accuracy by MARGIN TYPE at the segment level. Although the sign on the coefficients is consistent with the CV- and sonority-level models, neither WORD FREQUENCY ($\beta = 0.02$, $z = 1.2$, $p > .21$) nor STRESS ($\beta = 0.1$, $z = 1.5$, $p > .12$) reach significance for the segment-level regression. Once again, AGE is positively associated with accuracy ($\beta = 0.03$, $z = 2.5$, $p < .014$), and WORD LENGTH is negatively associated ($\beta = -0.19$, $z = -4.0$, $p < .0001$. Finally, VOWEL, $\chi^2(5) = 54.8$, $p < .0001$, and SUBJECT, $\chi^2(3) = 8.2$, $p < 0.05$, are significant as well.

Before turning to the discussion, we report several further steps that were taken to perform model assessment and validation. For each model, we computed Somers' D_{xy} rank correlation, which provides an interpretable measure of model fit ranging between –1 and 1 that can be used to assess and compare models. The D_{xy} values for the CV-level, sonority-level, and segment-level models were 0.440, 0.494, and 0.596 respectively. Unsurprisingly, the model with the most parameters has the best ability to predict outcomes. The D_{xy} statistics for the models that have the freedom to encode a separate parameter for each margin type will serve as a helpful reference point for the models presented in subsequent sections. To investigate potential over-fitting, we also performed bootstrap validation of all three models with 200 bootstrap samples each using the validate function in the rsm

[8]The full model results for the sonority and segment levels are shown in Table A1 and Table A2 in Appendix A.

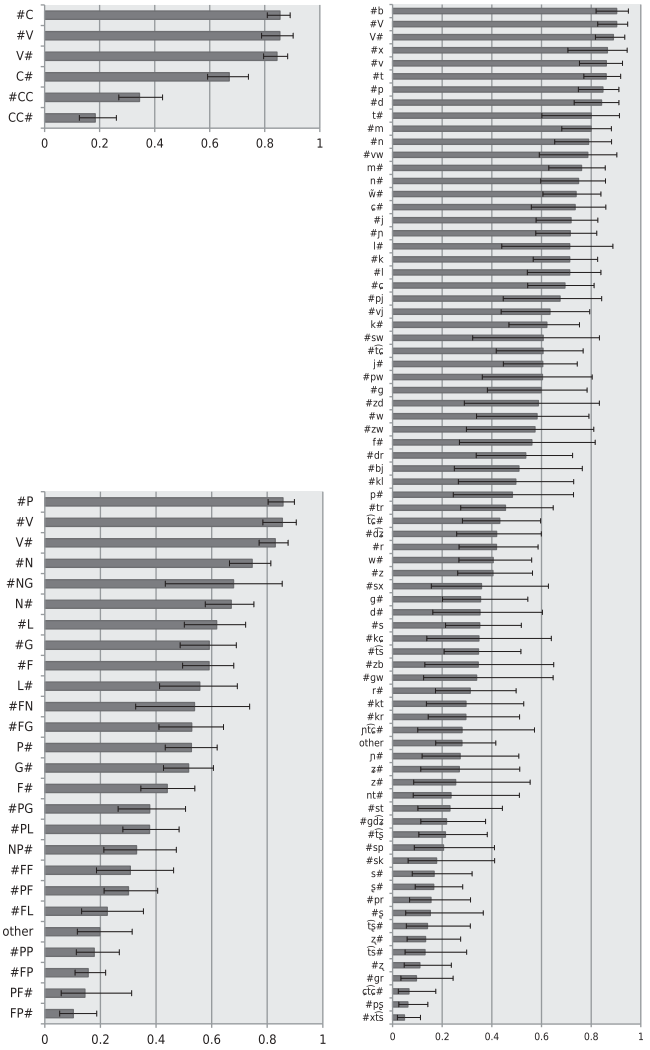

Figure 1. Predicted accuracies for margin types.

package in R (Harrell 2014). Over-fitting was minimal in all three models with optimism for D_{xy} estimated at 0.0036, 0.0053, 0.0095, for the CV, sonority, and segment levels, yielding corrected D_{xy} indices of 0.437, 0.489, and 0.587 respectively.

3.1.3. *Discussion*

Regression modeling makes it possible to adjust observed accuracies for various margin types based on potentially confounding effects of a number of variables. While it is not possible to statistically control for every possible influence on production accuracy in spontaneous speech due to data size, data sparsity, and limitations on what can be feasibly measured, the six control variables included here cover a range of possible influences (from prosodic and phonological context, to quantity of exposure, to within-subject variability) that are typically targeted for control in experimental studies of children's production, reducing the possibility that conclusions about accuracy of different margin types will be distorted by confounding variables.

Regression modeling also makes it possible to estimate a level of confidence for each of the overall production accuracies, automatically taking into account the degree of variability and the quantity of data available for each of the margin types. This lends perspective in the interpretation of these differences, making it possible to discern when differences are likely to be meaningful and when they may be spurious. The results in Figure 1 reveal that there are accuracy differences of both kinds. On the one hand, the range in production accuracy across margin types is substantial, and some margin types have tight confidence limits that are clearly separated from others. For example, the CV-level plot suggests that accuracy of #C, #V, and V# is meaningfully higher than that of C#, which is higher than #CC, which in turn is higher than CC#. Likewise, many of the margin types with lowest accuracy in the sonority- and segment-level plots are likely meaningfully lower than those at the top of the plots. At the same time, there is clearly a lot of overlap in the confidence intervals, especially in the segment- and sonority-level regressions. This indicates that it is not possible to draw firm conclusions on the basis of these data about relative accuracy of many of the overlapping margin types. Although inferences about significance can be made by inspection of these confidence intervals, doing so constitutes (implicit) multiple comparisons, and it is difficult to extract information about the overall production accuracy patterns from these large regressions. So, in the next section we present an alternative approach to interpreting the accuracy differences by systematically simplifying the models to capture these broader trends. Rather than relying on the confidence intervals directly to infer significance, we rely on hypothesis testing using model comparison to identify ways in which the models can (and cannot) be simplified without sacrificing explanatory power. In addition to providing quantitative results for individual margin types, the models presented in this section also serve as a benchmark for the simplified models of the next section.

Before turning to the overall trends, several qualitative observations can be made about the results in Figure 1. As noted earlier, there is substantial variability in the production accuracy of various margin types. Comparing accuracy of particular structures across levels of representation also reveals that there is substantial variability *within* margin type classes. For example, at the CV level the #C margin type is produced as a #C margin nearly all the time (the plot does not incorporate the effect of age, which would push all accuracies even higher). However, scanning the sonority-level plot for #C structures reveals that there is variability within the class of #C, with plosive-initial #Cs being produced more accurately than other sonority classes. More dramatically, plosive-initial #C margins are overall very accurate—children produce singleton initial plosives (including affricates) as plosives with high reliability—however, accuracy varies substantially within the class of plosives. While initial #b has the highest accuracy of all segment-level margin types, initial #t͡ʂ is one of the least accurate types of all. It is also possible to find seemingly contradictory accuracy patterns by examining patterns at different representational levels. As noted earlier, accuracy on #P is higher than accuracy on #F; however, it is not the case that children are uniformly more accurate on initial singleton plosives than initial singleton fricatives when performance within these classes is considered. Indeed, children are more accurate on #x and #v than they are on a number of initial plosives, such as #t͡ʂ or #t͡s. These observations are not specific to stops/affricates and fricatives; many similar patterns can be observed within and across the other classes. The fact that there are segmental (and sonority) effects of this sort is not particularly surprising, but it underscores the importance of considering multiple levels of representation. Without this broader picture, the low accuracy on a range of initial plosives could mask the fact that children can actually reliably produce certain plosives.

This variability also highlights how complex predicting developmental patterns on the basis of frequency (or any other factor) can be since the patterns can look quite different depending on the level of analysis. A frequency measure calculated from sonority-level representations cannot hope to make distinctions that would predict variability among different #P onsets. Frequency that is sensitive to segmental differences within the class of plosives is necessary to make such predictions. On the other hand, it is not clear whether segment-level frequency is capable of modeling developmental patterns at higher levels of representation, such as the higher accuracy of #C than #CC. All of these patterns reflect important aspects of children's phonological development, and the predictions of input-based models must be evaluated on their abilities to make general predictions for these and other aspects of development. This variability is the main

motivation for the systematic comparisons we present in the second half of the article evaluating each measure of frequency's ability to predict production at each level of representation.

3.2. *Overall trends in production accuracy*

3.2.1. *Method*

Our goal in this section is to provide a more general description of the accuracy patterns. We take the models developed in the previous section as a starting point and systematically simplify them as warranted by hypothesis tests using model comparison (Crawley 2013:324–329). The hypothesis tests investigate whether distinctions between margin types can be collapsed without sacrificing explanatory power. The model reduction process tests whether the added complexity in the model (due to making a distinction between margin types) is warranted by improvement in model fit. The reduction process utilizes hypothesis tests on the margin type coefficients (Harrell 2001:183–192). We consider two ways to simplify the model. The first is to test whether a given margin type warrants its own coefficient in the model: this is equivalent to testing the null hypothesis that coefficient $\beta_i = 0$ for margin type *i*. The other way is determining whether the data support a distinction between margin type *i* and *j*: this is equivalent to testing the null hypothesis that the coefficients for the two margin types are equal, e.g., that $\beta_i = \beta_j$. To test the possibility of collapsing and removing margin type predictors, the margin type factor for each level was recoded using dummy binary indicator variables with #V as the reference category.

To provide the most general description, we begin by determining the meaningful distinctions at the CV level. We then move on to the sonority level and determine which sonority-level margin types warrant their own parameters beyond those needed for the CV level. Finally, we proceed to the segment level, including only those predictors that significantly improve fit beyond the CV- and sonority-level predictors. As in the previous section, each model relies on accuracy coded at the corresponding representational level. At each level we attempt to collapse categories, starting with the most accurate categories, and continue collapsing categories as long as likelihood ratio tests are not significant and AIC (a measure that balances fit and parsimony) does not get worse. At the CV level we consider collapsing all categories; at the sonority and segment levels, we only attempt to collapse categories within CV-level classes. For example, at these levels, we never consider collapsing singleton codas with complex onsets even if their accuracies are very close. The result is a substantially reduced model that describes the general accuracy patterns rather than modeling each margin type individually.

3.2.2. *Results*

At the CV level, the three most accurate margin types, #C, V#, and #V, exhibit overlapping accuracy confidence limit ranges, as shown in Figure 1. Model comparison confirms that the model that collapses these three types is not significantly worse than the model that differentiates among all six CV-level types, $\chi^2(2) = 1.4$, $p > .5$. We therefore cannot reject the null hypothesis that the coefficients corresponding to these three margin types are the same, and we accept the reduced model that fits a single coefficient for this collapsed group. Indeed, since #V is the reference category, all three types get a coefficient of 0 (e.g., are removed from the model). Any further attempts to eliminate or collapse margin types at the CV level result in a highly significant decrease in model fit, however. For example, #CC cannot be collapsed with CC#, $\chi^2(1) = 25$, $p < .0001$. Likewise, none of the remaining margin types can be removed as predictors from the model. We conclude that the data provide sufficient evidence to warrant the remaining distinctions at the CV level. The overall fit of the reduced model ($D_{xy} = 0.438$) is nearly as good as the maximal CV-level model ($D_{xy} = 0.440$).

The resulting reduced model at the CV level is shown in Table 7.[9] The reference category is the collapsed #C/#V/V# class, and the coefficients for the other predictors should be interpreted relative

[9]For conciseness, we omit the control predictors from the tables in this section. Their effects are qualitatively similar as in the previous section.

Table 7. Simplified CV Model.

Predictor	*β*	*Standard Error*	*z*	*p*
(Intercept)	1.73	0.17	10.5	<0.0001
CV TYPE				
C#	− 1.03	0.06	− 17.5	<0.0001
#CC	− 2.39	0.07	− 33.9	<0.0001
CC#	− 3.22	0.16	− 20.0	<0.0001

to it. Since the remaining CV types have successively larger, negative coefficients that could not be collapsed, the model supports an accuracy hierarchy of {#C, V#, #V} > C# > #CC > CC#. This lines up intuitively with the confidence intervals plotted in Figure 1 and provides confirmation that those distinctions are statistically significant.

The sonority-level simplification process begins by considering the three CV predictors from the simplified CV model together with the sonority-level margin types, systematically removing and collapsing distinctions at the sonority level that are superfluous. Including the CV-level predictors makes it possible to reduce the model substantially since only those sonority-level types that behave differently from their CV-level class as a whole must remain in the model. Indeed, given the CV-level predictors, it is possible to remove 11 of the sonority margin types from the model completely, $\chi^2(11) = 9.3$, $p > 0.59$. Among the remaining #C types, #L, #G, and #F can be collapsed, but #N cannot be grouped with any of these. Among the remaining #CC types, #PG, #PL, #FF, #PF, and #FL can be collapsed, and #FP and #PP can be collapsed, but no further collapsing is possible within the #CC class. No collapsing is possible within the C# class, and only one type remains in the CC# class. Once these particular sonority-level onset clusters and codas are accounted for in the model, the distinction between CV-level #CC and C# is eliminated, and these are also collapsed. The full details of the simplification process are shown in Table B1 in Appendix B. The final reduced model for the sonority level is shown in Table 8. Rather than using 26 individual sonority-level margin types as in the previous section, this model relies on 2 CV-level predictors and 7 sonority-level predictors to describe the general patterns at the sonority level and above. The overall fit of the reduced model ($D_{xy} = 0.491$) is nearly as good as the maximal sonority-level model ($D_{xy} = 0.494$).

To interpret the model coefficients, recall that #C/#V/V# is the reference category at the CV level and that 11 sonority margin types (#P, G#, P#, L#, #NG, #FG, #FN, PF#, FP#, "other," V#) have been excluded, as their coefficients did not differ significantly from 0 given the other predictors. We interpret the results for each CV margin type class in turn. For #C, #P has been removed from the model, indicating that its accuracy can be predicted from the coefficient for the CV-level class. However, #N requires a separate parameter, and has a negative coefficient, indicating that its accuracy is lower than that of #P. The remaining #C types (#L, #G, and #F) are collapsed but require a separate and lower coefficient still, indicating that their accuracy is significantly lower. This interpretation of the model supports an accuracy hierarchy of #P > #N > {#L, #G, #F} for the

Table 8. Simplified Sonority Model.

Predictor	*β*	*SE*	*z*	*p*
(Intercept)	1.66	0.18	9.0	<0.0001
CV TYPE				
C# or #CC	− 1.56	0.06	− 24.4	<0.0001
CC#	− 3.70	0.23	− 16.5	<0.0001
SONORITY TYPE				
#N	− 0.58	0.09	− 6.3	<0.0001
#L or #G or #F	− 1.29	0.07	− 17.8	<0.0001
#PG or #PL or #FF or #PF or #FL	− 0.81	0.16	− 5.3	<0.0001
#FP or #PP	− 1.74	0.12	− 14.6	<0.0001
N#	0.62	0.11	5.7	<0.0001
F#	− 0.34	0.11	− 2.9	<0.005
NP#	1.36	0.34	4.0	<0.0001

singleton onsets. Similar reasoning yields the accuracy hierarchy of {#NG, #FN, #FG} > {#FL, #FF, #PF, #PG, #PL} > {#FP, #PP} for the complex onsets, N# > {G#, L#, P#} > F# for the singleton codas, and NP# > {FP#, PF#} for the complex codas.

For the segmental level, an analogous process was used: starting with the predictors from the previous reduced sonority-level model and full set of segment-level margin types, segment-level margin types were first removed and then collapsed within CV classes. 27 margin types were removed (#bj, #b, #dr, #gw, #kɕ, #kl, #kt, #ɲ, #pj, #pw, #st, #sw, #sx, #tr, #v, #vw, #vj, #x, #zb, #zd, #zw, V#, ɕ#, l#, ɲ#, ɲt͡ɕ#, nt#, t#, w̃#). All segment-level margin types within the same CV-level class that could be collapsed without adversely affecting model fit were collapsed. Once the segment-level predictors were included in the model and collapsed, they made some of the sonority-level predictors redundant, so these were removed from the model. Appendix B summarizes the whole reduction process. The final reduced model (D_{xy} = 0.594) has nearly as good model fit as the maximal segment-level model (D_{xy} = 0.596), and AIC is substantially improved. Rather than modeling segment-level patterns using 78 individual segment-level parameters as in the previous section, the reduced model has 2 CV-level predictors, 4 sonority-level predictors, and 13 segment-level predictors.

The resulting model (shown in Table 9) can be thought of as a refinement of the sonority-level model that additionally accounts for segmental variation within CV- and sonority-level classes. Recall that the sonority model supported an accuracy hierarchy of #P > #N > {#L, #G, #F} for the singleton onsets. The segmental model identifies six "strata" of accuracy for singleton onsets: the five explicitly modeled in the regression plus the stratum corresponding to those singleton onsets that are excluded (equivalent to having coefficients set to 0). These strata reveal interesting patterns both within and across sonority types.

Table 9. Simplified Segmental Model.

Predictor	*β*	*SE*	*z*	*p*
(Intercept)	2.22	0.30	7.4	<0.0001
CV TYPE				
#CC or C#	−1.09	0.09	−12.4	<0.0001
CC#	−2.90	0.23	−12.5	<0.0001
SONORITY TYPE				
#N	−1.19	0.11	−10.4	<0.0001
#PG or #PL or #FF or #PF or #FL	−1.02	0.12	−8.4	<0.0001
#FP or #PP	−1.64	0.17	−9.4	<0.0001
N#	−1.99	0.43	−4.6	<0.0001
SEGMENT TYPE				
#m or #n	0.47	0.15	3.1	<0.005
#t or #p or #d	−0.32	0.08	−3.9	<0.0005
#j or #l or #k or #ɕ or #w or #g or #t͡ɕ	−1.25	0.08	−15.0	<0.0001
#d͡ʑ or #r or #s or #z or #t͡s	−2.53	0.09	−28.2	<0.0001
#ʂ or #ʐ or #t͡ʂ	−3.76	0.21	−17.7	<0.0001
m# or n#	2.12	0.43	4.9	<0.0001
j# or k# or f# or p#	−0.56	0.11	−5.0	<0.0001
w# or d# or g# or t͡ɕ# or r#	−1.41	0.11	−12.6	<0.0001
s# or ʂ# or ʐ# or t͡s# or t͡ʂ# or ʑ# or z#	−2.63	0.14	−18.3	<0.0001
#gd͡ʑ or #sp or #sk or #kr	−0.75	0.22	−3.4	<0.001
#pr or #gr or #xt͡s or #pʂ	−2.30	0.21	−10.7	<0.0001
ɕt͡ɕ#	−1.85	0.53	−3.5	<0.005
other	−1.23	0.12	−10.6	<0.0001

Within sonority classes, the strata indicate that there is an accuracy hierarchy across different #P types, with #b > {#t, #d, #p} > {#k, #g, #t͡ɕ} > {#d͡ʑ, #t͡s} > #t͡ʂ. The two glides #j and #w pattern together, but #m and #n have higher accuracy than #ɲ, and accuracy is higher on #l than on #r. Finally, there are a wide range of accuracy profiles for the #F class, with {#x, #v} > #ɕ > {#s, #z} > {#ʂ, #ʐ}. Across sonority classes, it is clear that not all segment-level types fall neatly into the accuracy hierarchy established at the sonority level. In particular, while the #F class is overall one of the least accurate sonority level types, there are some fricatives (#x, #v) that pattern together with the most accurate #P onset (#b). Similarly, while the #P class is overall most accurate, some members of the #P class (#t͡ʂ) pattern together with the least accurate group of all. This analysis therefore presents a more nuanced picture of production accuracy that reveals considerable variation within classes, especially the plosives and fricatives.

Similar patterns hold for singleton codas. For P#, segment-level types span four accuracy strata: t# > {k#, p#} > {g#, d#, t͡ɕ#} > {t͡s#, t͡ʂ#}. Once again the nasals are split, with m# and n# more accurate than ɲ#. For glides, the hierarchy is w̃# > j# > w#, and for liquids it is l# > r#. For F#, there are three strata: ɕ# > f# > {ʐ#, z#, s#, ʂ#, ʑ#}. All the classes show substantial internal variability. Although there is low accuracy on P# and F# as a whole, some members of these classes pattern with the most accurate groups. Thus, as with singleton onsets, singleton codas reveal substantial variability across the sonority accuracy hierarchy established by the sonority-level model (N# > {G#, L#, P#} > F#) when behavior of individual segments is examined.

For complex onsets, only a few refinements are needed beyond those already made at the sonority level: {#NG, #FN, #FG} > {#FL, #FF, #PF, #PG, #PL} > {#FP, #PP}. While #PP clusters are already penalized at the sonority level, #gd͡ʑ is particularly bad and receives an additional penalty at the segment level. Similarly, while the #FP onsets overall have low accuracy and receive a negative coefficient at the sonority-level, accuracies of #st, #zb, and #zd are predictable from this, while #sk and #sp require a negative coefficient at the segmental level, and #xt͡s an even lower one, supporting the hierarchy {#zd, #zb, #st} > {#sp, #sk} > #xt͡s. The #PF class also receives a separate sonority-level parameter, but one of its members, #pʂ, is singled out at the segment level as having particularly low accuracy. Finally, some #PL clusters (#tr, #dr, #kl) are more accurate than others (#kr), which are more accurate than others (#pr, #gr). There are only three coda clusters distinguished at the segmental level, and these yield the hierarchy {ɲt͡ɕ#, nt#} > ɕt͡ɕ#, which is consistent with the sonority-level result NP# > {FP#, PF#}.

3.2.3. *Discussion*

These results establish concrete acquisition effects that provide a broader description of the production patterns in this corpus. This description makes an empirical contribution to existing studies of phonological development, in particular providing novel findings for development of phonology in Polish. In addition to serving as a basis for the theoretical investigations in the next section, these results provide a reference point future work can use to examine other hypotheses about phonological development, cross-linguistic variation, and the role of the input. The results presented here also reinforce the observations made in the previous section that acquisition patterns can appear quite different at different levels of representation. Substantial differences between classes at higher levels of representations may mask substantial variability within these classes at lower levels of representation. This highlights the multifaceted nature of the phonological system being acquired by the child, the challenges of evaluating theories of development, and the need to consider numerous levels of representation when testing developmental theories.

With these results in hand, we consider the potential role of the markedness and articulatory difficulty constraints discussed earlier. Universal markedness predicts specific developmental effects for basic syllable shapes (Blevins 1995; Levelt, Schiller & Levelt 2000; Jarosz 2010). Singleton onsets and null codas are predicted to be acquired earliest. In onset position, null and complex onsets are predicted to be disfavored. In coda position, singleton codas are predicted to be acquired after null codas, and complex codas after singleton codas. Consistent with previous findings on basic syllable structure development across multiple languages (see Jarosz 2010 for a review), our results are

consistent with all of these predictions, with the exception of null onsets, which pattern with simple onsets and null codas. Since production accuracy is very high for several of these structures, this could be a ceiling effect. Not predicted by markedness is the higher accuracy of initial clusters compared to final clusters. As discussed earlier, previous work (Jarosz 2010) has shown that complex onsets are more frequent in Polish than complex codas. Thus, the CV-level results are consistent with markedness but also suggest a role for frequency.

The patterns at the sonority and segmental levels are more complex, and there is reason to suspect that some sonority-level patterns are affected by lower-level phonetic pressures. As noted earlier, fricatives as a class tend to be acquired late by children across languages, and there are articulatory reasons for this: fricatives require precise control of a narrow constriction needed to produce frication, while stops and nasals require only a ballistic closure gesture (Kent 1992). Our results are consistent with this: #F and F# are in the lowest accuracy strata for simple onsets and codas. Similar considerations are involved in the articulation of affricates, which also tend to be acquired late (Edwards & Beckman 2008, Edwards, Beckman & Munson 2015). Our segment-level results suggest this holds for Polish as well. For example, the hierarchy established for the #P class places all stops at the top and all affricates at the bottom: #b > {#t, #d, #p} > {#k, #g, #t͡ɕ} > {#d͡ʑ, #t͡s} > #t͡ʂ. Liquids are also disadvantaged in phonological acquisition cross-linguistically (Edwards, Beckman & Munson 2015), and articulatory difficulty may be responsible. Work by Proctor (2009) suggests that liquids require complex articulatory coordination of both dorsal and coronal gestures: this fine motor coordination is likely to be challenging to children. Our results suggest that liquids may indeed be acquired late in Polish, with particularly low accuracy on both initial and final [r]. Within the class of fricatives and affricates, predictions are less clear. Edwards & Beckman (2008) find that highly perceptible sibilants like [s] are acquired earlier than nonstrident fricatives like [θ]. However, their work also indicates that there is cross-linguistic variation with regard to which coronal fricatives are favored by children. They find that children learning English and Japanese favor alveolar over postalveolar fricatives; however, Hua & Dodd (2000) find the opposite pattern in Putonghua (a dialect of Mandarin Chinese). In our results, there is a general preference for alveolo-palatal fricatives, over alveolar, over retroflex sounds, as most clearly evidenced by the accuracy hierarchy established for simple onset fricatives: {#x, #v} > #ɕ > {#s, #z} > {#ʂ, #ʐ}. Language-internal frequency could be responsible for this variability; however, it is also interesting to note that our results most closely parallel those documented for Putonghua, which, like Polish, but unlike other languages studied in previous work, has a three-way place contrast in the sibilants contrasting alveolar, alveolo-palatal, and retroflex place. Both languages also show earlier acquisition of noncoronal (labio-dental and velar) fricatives, unlike English. Since frequency analyses of Putonghua are not available, it is not possible to draw firm conclusions about the causal role of frequency, but the commonalities between the two languages in terms of the system of contrasts warrants further attention.

These lower-level phonetic considerations provide possible explanations for some apparent inconsistencies at the sonority level. As discussed earlier, universal markedness predicts a preference for low-sonority onsets, high-sonority codas, and clusters that rise sharply in sonority toward the nucleus and fall in sonority away from the nucleus. The preference for low-sonority onsets has been found in several previous studies (Gnanadesikan 2004; Pater 1997; Pater & Barlow 2003; Ohala 1999), and the hierarchy #P > #N > {#L, #G, #F} established in our results is largely consistent with this. The fact that accuracy on fricatives is lower than expected could be attributed to articulatory difficulty of fricatives in general. The generalizations for simple codas, N# > {G#, L#, P#} > F#, are less clear. While the most accurate class is the sonorant N# and the least accurate is the obstruent F#, the observed scale does not neatly line up with sonority. Previous studies disagree on the role of universals in coda acquisition. Stites, Demuth & Kirk (2004) and Ohala (1999) found some support for universals, while Zamuner, Gerken & Hammond (2005) did not. This suggests that an explanation relying solely on sonority and cross-linguistic tendencies is unlikely to provide a complete explanation. The observed hierarchies for complex onsets {#NG, #FN, #FG} > {#FL, #FF, #PF, #PG, #PL} > {#FP, #PP} and complex codas NP# > {FP#, PF#} are partially consistent with sonority-based predictions, but the fit is not perfect. For example, #PL and #PG are among the least marked initial clusters (together with #FG), but they are not the earliest sonority

sequences to be acquired. Class-level variability and interactions with segment-level accuracy may be partially responsible. As shown in Figure 1, children are most accurate on #vw, #pj, #vj, #sw, #pw onsets, all of which have the optimal obstruent-glide sequence. Accuracy on the #PG class overall is diminished by accuracy on particular plosive-glide combinations like #gw. Likewise, late acquisition of liquids (and trills in particular) is likely at least partially responsible for lowering the overall accuracy on the #PL class.

Altogether, this qualitative examination suggests a complex interaction of effects rooted in universal syllable structure markedness, language-particular frequency, and phonetic difficulty. We have also tried to highlight in our discussion the ways in which developmental effects at different levels of representation may interact. Our study cannot definitively tease apart all these possible sources of influence on production accuracy, but one goal of the following section is to explore quantitatively the extent to which these various developmental effects can and cannot be explained by input statistics at various representational levels.

4. Frequency analyses

The regression analyses in the preceding section modeled variability in margin-type accuracy using a separate predictor variable for each phonological category (e.g., margin type). In this section we utilize the same general methodology but instead examine the extent to which the *frequency* of margin types can predict variability in margin-type accuracy. That is, we systematically replace margin-type predictor variables with their corresponding frequencies, resulting in substantially more restrictive models that predict accuracy on the basis of frequency alone. This approach provides a way to evaluate and compare models relying solely on frequency to one another and to the models of the previous section.

We rely on this methodology to address three questions. In section 4.1, we compare the predictive capacities of the seven frequency measures to one another and evaluate their relative abilities to capture the observed variability in margin type accuracy. In section 4.2, we investigate how the frequency measures interact with one another and whether production is sensitive to a combination of input statistics. Finally, in section 4.3 we ask how the predictions of the frequency-only models compare with observed development, identifying systematic divergences that signal that other factors may be at work.

4.1. *Evaluating individual models of frequency*

4.1.1. *Method*

Here we consider the extent to which each frequency measure predicts production accuracy. If frequency predicts acquisition, then higher relative frequency should correspond to higher relative accuracy. Rather than examining raw frequency counts and assuming a linear relationship between frequency and accuracy, we use regression modeling to make predictions for accuracy on the basis of frequency and allow each frequency measure to be fit to the data as well as possible. In addition to maximizing the utility of each frequency measure, regression modeling makes it possible to allow other variables to contribute to the predictions and to use principled statistical tests to compare and evaluate different models. A final advantage of this approach is that each of the frequency measures can be used to make predictions for each of the representational levels, and the quality of these predictions can be directly compared.

We evaluate and compare the predictions of each of the seven measures of frequency individually by fitting a separate logistic regression model for each measure at each level of representation. As before, we include the control predictors, AGE, LENGTH, VOWEL, STRESS and SUBJECT, in the models to account for variability in production accuracy that is not directly attributable to syllable structure. We also report the predictions of a baseline model relying only on the control predictors for each of the levels of representation to highlight the independent contribution of each frequency measure.

We subsequently evaluate each model in two ways. The first method involves calculating Somer's D_{xy}, yielding an absolute and comparable statistic indicating each model's ability to account for all variability in the data. It compares all frequency models across all margins in the corpus and all levels of representation. In performing this evaluation, we also investigate the reliability of the frequency estimates in our analyses and the extent to which sample size may be affecting our conclusions. We collected 200 bootstrap samples for each frequency measure and fit models using these reestimated frequency measures to construct 95% confidence intervals around the D_{xy} statistic for each measure. These intervals provide an indication of the extent to which D_{xy} is sensitive to the idiosyncrasies of the particular corpus we use to estimate statistics.

The second method calculates the D_{xy} statistic relative to the classifications made by the three margin type models in section 3. Since frequency can only differentiate among margin types and not other aspects of the data, the margin type models of section 3 provide an upper bound on how much variability frequency can possibly capture at each level. This more effectively isolates the models' abilities to capture variability attributable to the margin types themselves. In effect, it reflects each frequency measure's ability to reproduce the variability attributable to the 6-, 26-, 78-level MARGIN TYPE factors at the CV, sonority, and segment levels respectively.

4.1.2. *Results*

Several noteworthy patterns emerge from the results of the first evaluation, shown in Figure 2. The confidence intervals are small, indicating that differences between the frequency measures are quite robust overall. Since they are all well above the performance of the baseline, all the frequency measures capture important information about relative accuracy of margin types, even the word-frequency measure. However, there is also substantial variability among the measures. One robust, dramatic effect is that the token frequencies provide a better fit to the data than the corresponding type frequencies for all levels of representation. Segment token frequency is a better predictor of accuracy than segment type frequency for each of the levels of production accuracy. Sonority token is better than sonority type, and CV token is better than CV type. These effects dramatically overshadow the small amount of variability represented by the confidence intervals. Another notable effect is that word frequency is not a good predictor of accuracy compared to the frequency measures referring to margin types, with the exception of CV type frequency, which performs quite poorly at the sonority and segment levels. This indicates that abstract representations that enable the encoding of statistics sensitive to abstract phonological units like onsets and codas are important for modeling the production patterns.

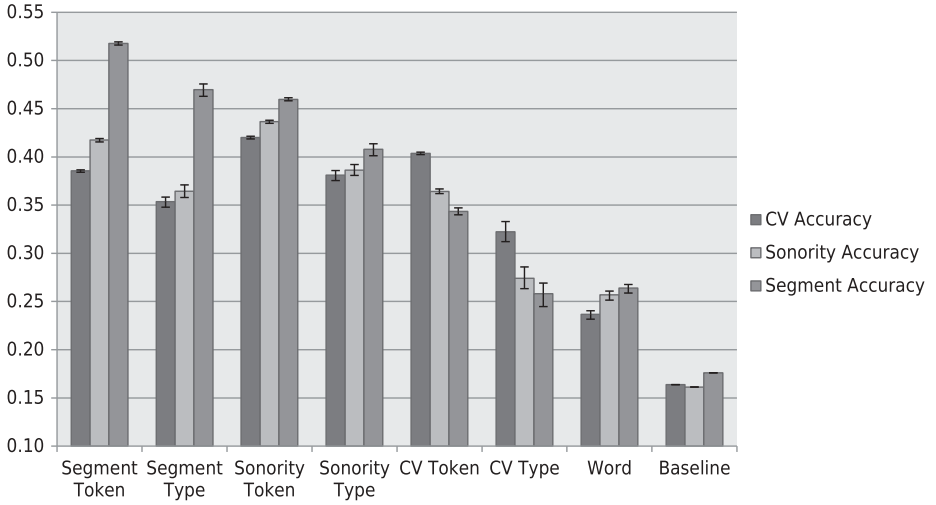

Figure 2. D_{xy} of each frequency measure with 95% bootstrap confidence intervals.

Another important generalization is that the best measure of frequency depends on the level at which accuracy is evaluated. When accuracy is evaluated at the segment level, segment token frequency provides the best fit. When accuracy is evaluated at the sonority level, sonority token frequency provides the best fit. Finally, when accuracy is evaluated at the CV level, sonority token frequency (closely followed by CV token frequency) is the best predictor. It is not especially surprising that segment-level frequency is best at predicting segment-level accuracy among margin types. After all, segment-level frequency measures have the capacity to differentiate among segment-level types, while sonority-level and CV-level frequency cannot. For example, segment-level frequency can reflect the fact that #p is more frequent (and more accurate) than #g, while sonority-level and CV-level frequencies cannot differentiate these types since they fall within the same #P and #C classes respectively. Thus, segment-level frequency has more power to differentiate among different types than sonority-level and CV-level frequency. However, this capacity is apparently not advantageous when it comes to modeling accuracy at these more abstract levels. Instead, sonority- and CV-level accuracy is better predicted by more abstract, and less expressive, frequency measures calculated at the sonority and CV levels. Thus, whether children produce, e.g., #P singleton onsets reliably as #P singleton onsets is better predicted by frequency of the #P class as a whole than by the frequencies of individual plosive onsets. This suggests that there is an important role for more abstract representations and their overall frequency in predicting aspects of children's phonological development.

The D_{xy} measures fall consistently below 0.52, suggesting that there is significant variability in accuracy that frequency cannot predict relative to the maximum possible of 1. However, this is arguably not a fair evaluation, since the models with one parameter for each margin type reported in section 3.1 achieved D_{xy} statistics in an overall similar range. They were 0.440, 0.494, and 0.596, for the CV, sonority, and segment levels respectively. The results of the second evaluation, shown in Figure 3, relativize the frequency models' success to these upper limits for each level of representation. Although the patterns are qualitatively somewhat different, they generally corroborate those of the first evaluation. In particular, type frequencies perform worse than corresponding token frequencies in all cases. Word frequency is not as predictive as any of the structure-sensitive frequency measures, but it does capture some variability as evidenced by its superiority over the baseline. Finally, just as in the previous evaluation, the best frequency measure depends on the level of representation, although the variability is less pronounced. Segment token frequency is best at predicting the variability across segment-level types, while CV token frequency is best at predicting variability across CV-level and sonority-level types.

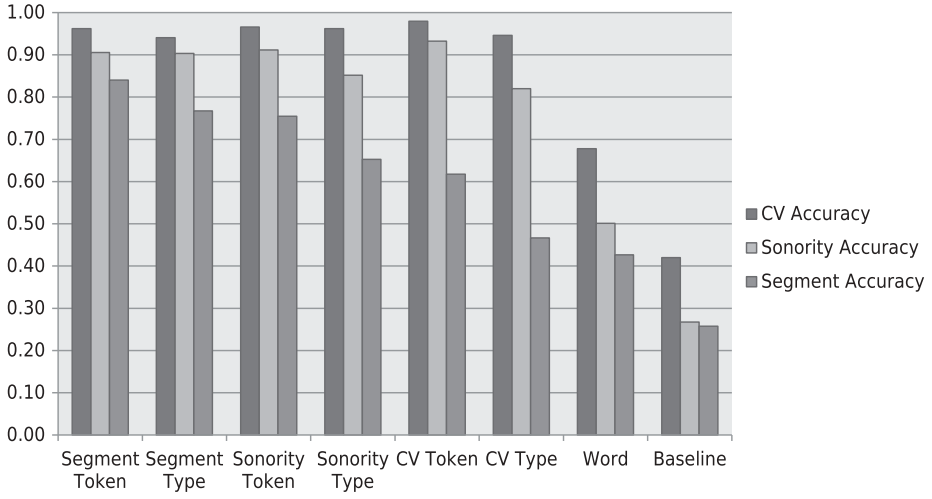

Figure 3. D_{xy} of frequency models relative to margin type models.

Beyond these similarities, the main striking observation here is the high D_{xy} for many of the frequency models. Except for word frequency, all the frequency models achieve D_{xy}'s of 0.94 or above relative to the CV-level margin type model. This means that children's relative accuracy on the six CV-level types is largely predictable from a variety of frequency measures. More impressively, several frequency measures do quite well at predicting the relative accuracy on the 26 sonority-level types: all the token frequency measures and the segment type frequency measure achieve D_{xy}'s of 0.9 or above. The frequency models show a decreased ability to capture the variability in accuracy on the 78 segment-level types, with sonority-level and CV-level token frequencies achieving D_{xy}'s of 0.755 and 0.617 respectively. However, segment-level token frequency does considerably better, reaching a D_{xy} of 0.840.

4.1.3. *Discussion*

These evaluations have shown that different input statistics vary substantially in their capacities to predict syllable structure development in Polish. They also indicate that development of higher-level phonological structure is sensitive to statistics calculated over abstract phonological classes. This is consistent with and builds on recent work showing that the capacity to reference abstract phonological classes is critical for modeling adult phonological generalization (Albright 2009; Daland et al. 2011; Hayes & Wilson 2008; Finley & Badecker 2009).

The finding that token frequencies reliably outperform type frequencies at all levels of representation is more unexpected. As discussed earlier, previous studies have found both type (Edwards & Beckman 2008; Edwards, Beckman & Munson 2004; Ingram 1988; Munson 2001; Richtsmeier, Gerken, & Ohala 2009; Archer & Curtin 2011) and token (Kirk & Demuth 2005; Roark & Demuth 2000; Stites, Demuth & Kirk 2004; Zamuner, Gerken & Hammond 2004, 2005; Levelt, Schiller & Levelt 2000) frequencies to be predictive of phonological development. Interestingly, the former studies have largely investigated production accuracy in nonword repetition tasks, while the latter group of studies has largely focused on accuracy in spontaneous production. These distinct foci may provide a hint about the divergent conclusions, since different mechanisms may be involved in each case. One difference is the role of generalization. Nonword repetition tasks targeting novel items require listeners to rely on generalization, while accurate production of familiar words could in principle occur without generalization. Although the studies with adults that found type frequency to be more predictive than token frequency also emphasized generalization (Albright 2009; Hayes & Wilson 2008; Albright & Hayes 2003; Richtsmeier 2011), it seems unlikely that this dimension can fully explain our results. Our results strongly suggest that generalization is at work in spontaneous production accuracy—otherwise, we would expect production accuracy to be driven primarily by word frequency. However, word frequency is not nearly as predictive of accuracy as the measures referring to abstract phonological entities. Children's production seems to be more sensitive to the frequency with which these abstract phonological patterns occur in the input than to frequency of particular words.

Another difference concerns the role of articulatory practice as opposed to perceptual learning (see also Richtsmeier et al. 2009), and we suspect that this factor may be the more relevant distinction between the studies. Nonword repetition tasks provide little opportunity for articulatory practice: The frequency manipulations in these studies are likely to affect perceptual learning. Likewise, the adult studies supporting type frequency involve wordlikeness ratings and wug tests with little role for articulatory practice. In contrast, accuracy on spontaneous production is more likely to be affected by articulatory practice as children repeatedly attempt the same phonetic configurations. Token frequency provides a better indication of the total amount of articulatory practice children have with a particular configuration. To confirm this, we computed correlations between token and type frequencies in the input to the frequencies with which children attempted various margin types in the child speech portion of the corpus. For all levels of representation, correlations between frequency in the children's production targets and input frequency was higher for token (range: 98.5–99.8) than type (range: 94.9–96.7) frequency.

Although token frequency is more predictive, this does not preclude a role for type frequency. Neither does articulatory practice preclude a role for perceptual learning. Richtsmeier, Gerken &

Ohala (2011) argue that both type and token variability are important in perceptual learning: "For each level of analysis, variability helps to signal invariance. At the phonetic level, talker variability signals an invariant word shape. At the phonological level, varying words signal an invariant phonotactic sequence." Frequency and variability are not identical, however. It may be that variability at various levels is required to form robust perceptual representations but that sufficient practice is also essential for mastering various articulatory configurations.

4.2. *Multiple frequency measures?*

So far our models have relied on individual measures of frequency, but it is possible production accuracy is affected by multiple types of input statistics simultaneously. To explore this possibility, we performed likelihood ratio tests comparing the models with individual frequency measures and control predictors to superset models relying on multiple measures of frequency.

To corroborate the findings of the previous sections that the structure-sensitive frequency measures capture variability in production accuracy that is not predicted based on word frequency alone, we compare a model with word frequency and controls as the only predictors to a model that includes word frequency as well as the three token frequency measures and controls. When the dependent variable is segment-level accuracy, the superset model is significantly more predictive, $\chi^2(3) = 2111.8$, $p < .0001$. Likewise, the model with word frequency and all three type frequencies included as predictors is superior to the model with only word frequency, $\chi^2(3) = 1877.7$, $p < .0001$. In both cases, the results are similar for accuracy at other levels of representation. This indicates that information from some structure-sensitive frequency measure is useful for predicting production accuracy.

We can also examine the combined token frequency and combined type frequency models to determine which structure-sensitive frequency measures are contributing independent information. In the combined type frequency model, all frequency measures are significant: word frequency ($\beta = 0.12$, $z = 10.4$, $p < .0001$), CV type frequency ($\beta = 0.10$, $z = 2.0$, $p < .05$), sonority type frequency ($\beta = 0.07$, $z = 2.3$, $p < .05$), and segment type frequency ($\beta = 0.49$, $z = 25.9$, $p < .0001$). In the combined token model, segment token frequency ($\beta = 0.48$, $z = 24.0$, $p < .0001$) and CV token frequency ($\beta = 0.15$, $z = 3.4$, $p < .001$) are significant, but word frequency ($\beta = 0.02$, $z = 1.4$, $p > .17$) and sonority token frequency ($\beta = 0.05$, $z = 1.5$, $p > .13$) are not. In both models, therefore, multiple structure-sensitive frequency measures are significant, indicating that statistical information from multiple representational levels provides at least partially complementary and useful predictive information about production accuracy.

This evidence from model comparisons indicates that combining information from multiple frequency measures is helpful. These significant effects are also reflected in modest gains in D_{xy} as compared to the best individual measures of frequency. The D_{xy} of the combined token frequency model is 0.521, which is slightly higher than the D_{xy} of the segment token frequency model presented in section 4.1 (0.518). Similarly, the D_{xy} of the combined type frequency model is 0.490, which is somewhat higher than the performance of the segment type model from section 4.1 (0.470). Similar effects can be observed when considering the second evaluation method. For example, while the D_{xy} statistics for the token segment frequency model were 0.962, 0.906, 0.840 at the CV, sonority, and segment levels respectively, the corresponding D_{xy} statistics for the combined token frequency model are 0.964, 0.930, and 0.867.

These statistics also corroborate the earlier finding that token frequencies are more predictive: the D_{xy} of the combined token frequency model (0.521) is substantially higher than that of the combined type frequency model (0.490). Interestingly, however, a superset model with all seven frequency measures included is superior to the combined token frequency model, $\chi^2(3) = 97.1$, $p < .0001$), $D_{xy} = 0.526$. While the contribution of the type frequency measures to the D_{xy} is not substantial, this comparison does indicate that type frequencies, while less predictive overall, have an independent effect on production accuracy beyond that captured by the more predictive token frequency measures. This is consistent with recent work discussed earlier arguing that both type and token frequency play essential roles in acquisition.

Finally, we also performed model comparisons to check whether each of the children in the corpus exhibits the same overall pattern of results individually. For each child, models relying on multiple structure-sensitive frequency measures are superior to models relying only on word-frequency and controls. For all children, frequency measures from multiple representational levels are significant in the combined frequency models. For each child, the combined token frequency model is more predictive than the combined type frequency model, and the model with all seven frequency measures is superior to the combined token frequency model. Thus, the data for each child individually supports the same general conclusions as the data in aggregate.

4.3. *Mismatches between frequency and accuracy*

Overall, therefore, many aspects of the production patterns are largely predictable from frequency measures. There is significant variability in the predictiveness of different frequency measures, but the structure-sensitive token frequencies, especially segment token frequency and the combined token frequency model, capture substantial variability in production accuracy across the 6 CV-level margin types, the 26 sonority-level margin types, and, to a lesser degree, the 78 segment-level margin types. We conclude that these input statistics, calculated over abstract representations at multiple levels, are highly predictive of production accuracy.

However, this does not mean that these input statistics alone can fully account for phonological development. On the contrary, there is a great deal of variability in production accuracy across margin types that these models fail to capture. This is most easily demonstrated by a likelihood ratio test between the combined token frequency model and a superset model that also includes the margin type predictors from the reduced segmental model from Table 9 in section 3.2. The difference between these two models is highly significant, $\chi^2(19) = 1568.7$, $p < .0001$. Not only are these 19 margin type predictors significant en masse after controlling for all three token frequency measures, but all of them remain highly significant ($p < .01$) individually. This indicates that the combined effects of these statistics do not fully capture the general production accuracy patterns. The importance of the margin type predictors is also reflected in the substantial increase in D_{xy}, from 0.521 for the combined frequency model to 0.595 when the target margin predictors are added. We therefore conclude that none of the frequency measures or combinations thereof can fully account for all of the acquisition effects.

To explore the predictions of frequency qualitatively, we compare the predictions of the combined token frequency model to those of the models fit with one parameter for each margin type. As discussed earlier, the combined token frequency model has access to frequency calculated at each of the levels of representation and provides the best fit to the data, thereby representing the most successful frequency predictions. The models are used to make predictions for each observation in the data, and then these predictions are averaged within margin types so that matches and systematic mismatches on margin types can be inspected. These results are plotted in Figure 4 for CV margin types, Figure 5 for sonority margin types, and Figure 6 for segment margin types. The plots are sorted by accuracy as predicted by the margin type models. In addition to these qualitative inspections, we discuss results of model comparisons that test whether the frequency models succeed in capturing the markedness and phonetic constraints discussed earlier.

Consistent with the analysis in section 4.1, the frequency model does a good job at matching the predictions of the CV margin type model. The main divergence in the predictions is for the #V type. Null onsets are quite rare in Polish, yet the children produce them reliably with high accuracy. The frequency model actually predicts slightly lower accuracy for #V than C#, which is contrary to development. A less dramatic mismatch occurs for #CC, whose high frequency fails to predict its moderate accuracy. Returning to the CV-level markedness constraints discussed in section 3, we can also ask whether the combined token frequency model succeeds in fully capturing the observed production effects. Recall that universal markedness theory predicts that codas and clusters should

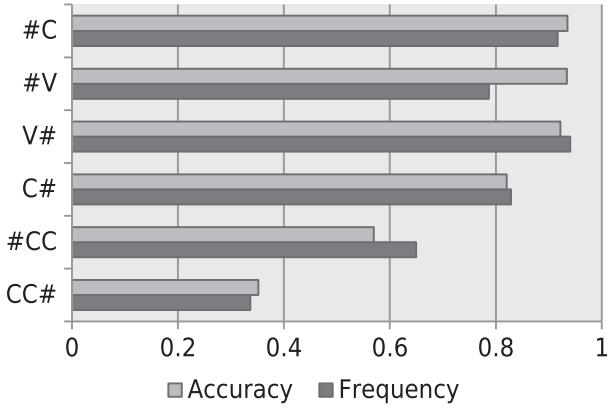

Figure 4. Frequency predictions vs. accuracy for CV margin types.

be disfavored and that the acquisition results are consistent with this. Although frequency correctly predicts that codas and clusters should be disadvantaged, model comparison reveals that production on both is significantly worse than predicted by accuracy. In a superset model including all token frequencies, word frequency, and controls, both codas ($\beta = -0.8$, $z = -10.16$, $p < .0001$) and clusters ($\beta = -1.57$, $z = -10.1$, $p < .0001$) receive significant, negative coefficients. This indicates that frequency fails to predict the extent to which these structures are disfavored by children.

The frequency predictions for the sonority-level margin types show more dramatic mismatches. #NG onsets are rare in Polish, but children's accuracy on them is very high. Frequency also fails to predict the relatively high accuracy on #FN onsets, L# codas, and NP# codas. In contrast, frequency predicts overly high accuracy on F#, #FL, #PP, #FP, and FP#. Codas and clusters composed of obstruents are very frequent in Polish, but children's accuracy on these structures is poor, especially on the clusters. Many of these discrepancies can be related to the markedness and phonetic pressures discussed earlier. The clusters for which frequency underpredicts accuracy have preferred sonority profiles, while those for which frequency overpredicts accuracy have sonority plateaus or involve phonetically disfavored segment classes F and L. On the other hand, the higher than expected accuracy on L# is consistent with the preference for higher sonority in coda. Model comparisons incorporating all these factors confirm that the frequency models do not fully capture these phonetic and markedness pressures. Accuracy on fricatives ($\beta = -0.55$, $z = -8.79$, $p < .0001$), liquids ($\beta = -0.71$, $z = -6.94$, $p < .0001$), and obstruent sequences ($\beta = -1.48$, $z = -12.75$, $p < .0001$) is significantly lower than frequency predicts. Consistent with sonority universals, simple obstruent onsets are more accurate than predicted ($\beta = 0.14$, $z = 2.05$, $p < 0.05$), and simple obstruent codas are less accurate than predicted ($\beta = -0.8$, $z = -10.97$, $p < .0001$). Each of these effects highlights a role for markedness or phonetic difficulty that cannot be explained by these frequency measures.

At the segmental level there are quite a few dramatic mismatches. For singleton onsets, frequency substantially underpredicts the accuracy of #x but substantially overpredicts the accuracy of #z, #r, #s, #t͡s, #ʂ, #ʐ and #t͡ʂ. Initial [x] is not very frequent compared to other singleton onsets, but children are very accurate on this onset. In contrast, coronal affricates, fricatives, and trills are relatively frequent, but children do not produce them reliably. Similar patterns hold for singleton codas. Frequency substantially underpredicts the accuracy of t#, l#, f#, p#, and d#, but it substantially overpredicts accuracy on w# and a number of coronal fricatives and affricates: s#, ʂ#, t͡s#, #ʐ and #t͡ʂ. There are numerous mismatches on clusters as well. For example, children are much more accurate on #vj, #pw, #pj, #zw, #zd, and #bj than predicted by frequency, but they are much less accurate on ɕt͡ɕ#, #xt͡s, #pʂ, #pr, and #st than frequency predicts. Obstruent–glide onsets are relatively rare, but children produce them accurately. Conversely, obstruent sequences are relatively frequent, especially those composed of coronal obstruents, yet children struggle

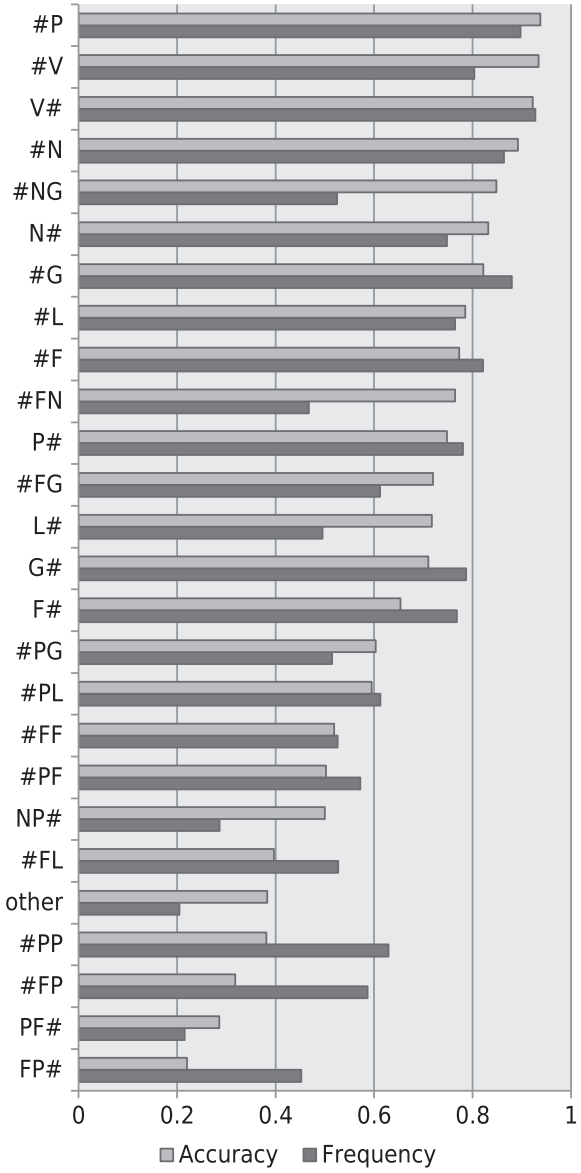

Figure 5. Frequency predictions vs. accuracy for sonority margin types.

with these clusters. Once again, many of these discrepancies can be related to the phonetic pressures discussed earlier. A superset model incorporating all token frequency measures, word frequency, controls, the sonority- and CV-level markedness predictors, and additional phonetic predictors indicates that children are less accurate than expected on several articulatorily disadvantaged sound classes: affricates ($\beta = -0.27$, $z = -2.1$, $p < .05$), stridents ($\beta = -1.17$, $z = -10.3$, $p < .0001$), trills ($\beta = -1.5$, $z = -8.26$, $p < .0001$), and retroflexes ($\beta = -1.4$, $z = -11.12$, $p < .0001$). These are all significant in a combined model, indicating that segments that fall into multiple classes (e.g., retroflex affricates) are especially disadvantaged. In general, this strongly suggests that there are phonetic pressures at work, since development cannot be explained by these frequency measures alone.

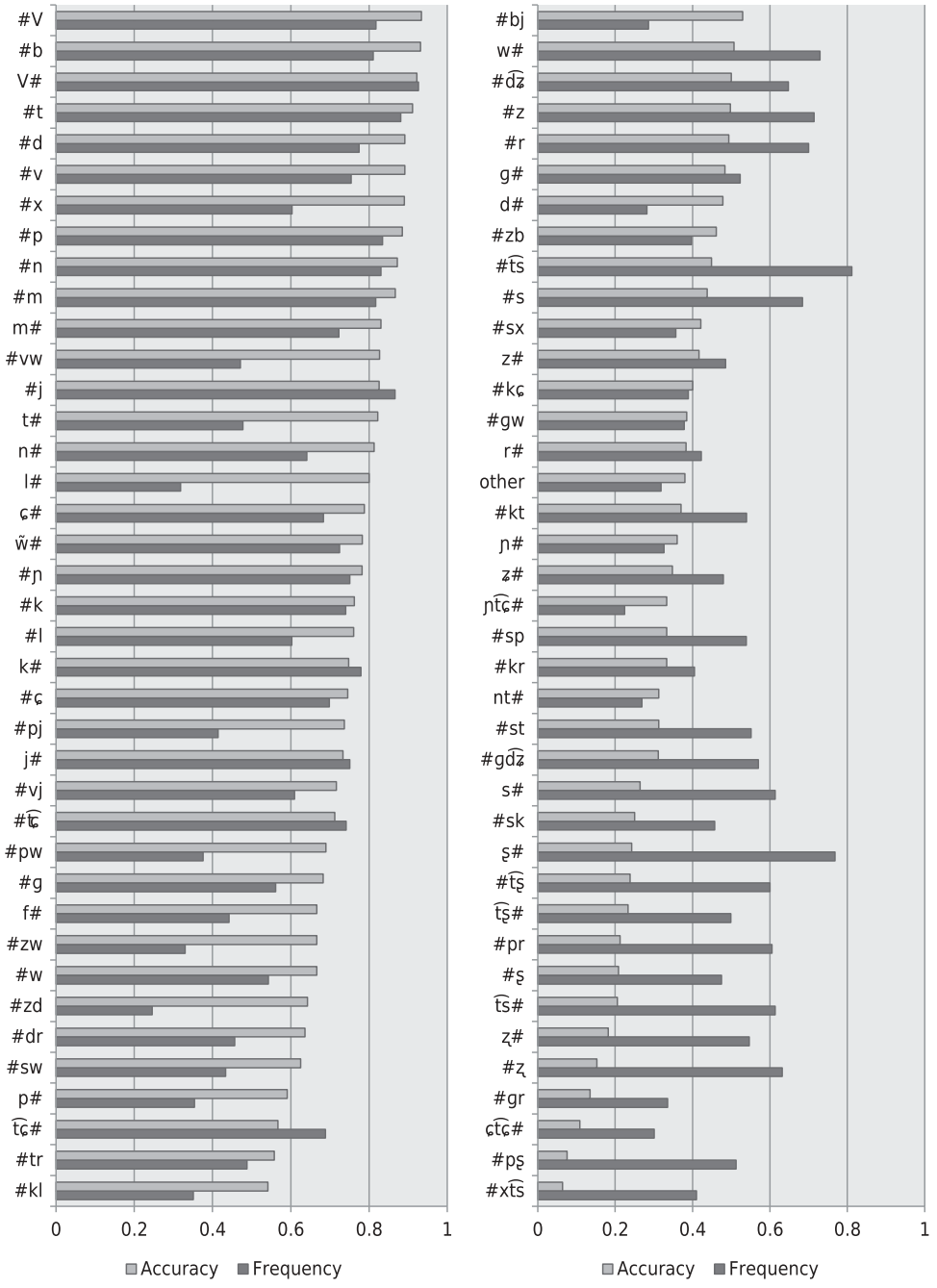

Figure 6. Frequency predictions vs. accuracy for segment margin types.

5. General discussion and implications

Returning to the research questions outlined in section 1.2, the results of the analyses in section 4 support several conclusions. All measures of frequency are highly predictive of production accuracy, but different measures of input frequency make dramatically different predictions, and their quality varies widely. Across all analyses, the frequency measures calculated over representations that make

reference to abstract phonological categories like onset and coda were better predictors of production accuracy than word frequency. Across all analyses, token frequency measures were better predictors of production accuracy than corresponding type frequency measures. These consistent, robust results in our analyses are somewhat surprising given that prior work has found type frequency to be more predictive in several contexts. We have suggested that these differences may relate to the role that articulatory practice plays in spontaneous production as opposed to nonword repetition tasks, but this is a hypothesis that requires further attention. Our analyses cannot reliably dissociate between perceptual exposure and articulatory practice, as our best estimates of both of these figures are highly correlated.

Despite these consistent generalizations, there is no single frequency measure that best accounts for production patterns across all levels of representation. Broadly speaking, more abstract, higher-level production patterns are better predicted by frequency measures calculated over more abstract, higher-level representations, while lower-level production patterns are best predicted by lower-level representations. The models that are most predictive of production patterns overall rely simultaneously on frequency measures calculated at multiple levels of representation. In addition, as mentioned earlier, frequency measures that access abstract phonological configurations are consistently more predictive than word frequency. These results in combination point to a multifaceted role of frequency in phonological development, one that references multiple, abstract levels of phonological representation. A number of recent studies support similar conclusions, indicating that representations such as natural classes (Albright 2009; Hayes and Wilson 2008), tiers (Hayes & Wilson 2008), and syllables (Daland et al. 2011; Kager & Pater 2012) are critical for modeling phonological learning. Our results and those of the related studies strongly suggest that access to abstract phonological representations is necessary for successful learning and successful modeling of acquisition.

Several frequency models—especially the combined token frequency, the segment token frequency, and the combined type models—go a long way toward capturing the variability in accuracy among margin types at the CV, sonority, and segment levels. This indicates that input statistics calculated over appropriate representational units can account for a wide range of developmental patterns at various levels of analysis for structures ranging widely in their complexity. Overall, this lends considerable support to the hypothesis that appropriately defined input statistics have the potential to explain a great deal about phonological development. At the same time, it is important to remember that the evaluations of input-based models considered here examine only certain aspects of the developing phonological system. Indeed, some of the previous studies that have strongly argued for alternative developmental theories relying primarily on innate biases or universal constraints have examined other aspects of developmental phonology, such as error patterns or productive processes and alternations, which are beyond the scope of the current investigations. Furthermore, the predictions of input-based models must be evaluated cross-linguistically to determine if the kinds of models found to be successful in this analysis for Polish are highly predictive in other languages with very different input statistics. While the present results support a substantial role for appropriate input statistics, a more comprehensive examination of input-based models is needed before more general conclusions about the primary role of input statistics in phonological development can be reached.

In addition, the results in the preceding sections have shown that there are many systematic discrepancies between children's production patterns and the predictions made on the basis of frequency. This indicates that none of these measures of frequency (or their combinations) can fully account for the variability in accuracy across margin types. This naturally raises the question of whether there could be other formulations of frequency or phonotactic probability not considered here that would improve the performance of input-based models. The answer is almost certainly yes. We considered frequency measures calculated at two abstract representational levels—the sonority and CV levels—each of which can be defined in terms of abstract phonological features, such as [±SYLLABIC], [±CONSONANTAL], [±SONORANT], and [±APPROXIMANT]. However, there are countless other

representational levels, and configurations of natural classes more generally, that play important roles in the sound systems of natural languages that were not considered here. For example, our input-based models were endowed with the capacity to penalize the class of initial consonant clusters if these structures had low frequency (by virtue of access to the CV level). However, none of the models had the capacity to represent structures in terms of other features such as [CORONAL], [±ANTERIOR], or [±DISTRIBUTED]. Although the segment-level frequency measures enabled the models to penalize segments belonging to these classes on an individual basis, it is possible that some kind of input statistic calculated at these more abstract levels would have revealed these classes to be relatively underrepresented and hence predicted their lower accuracy. Therefore, it is possible that a richer model of input frequency with access to a much wider array of phonological features would predict accuracy more successfully on the basis of the input. This is an empirical question that requires further work. More generally, the extent to which models of phonotactics endowed with richer phonological representations could be predictive of phonological development is an important question for follow-up work.

Nonetheless, it seems unlikely that access to additional phonological features could explain all the production patterns. The models we considered failed to account for the relative accuracy on many classes that they did have access to. For example, none of the models could explain the relatively high accuracy on null initial onsets or the relatively low accuracy on complex onsets even though these structures could be represented at multiple levels. None could explain the relatively low accuracy on clusters composed of obstruents even though the sonority level allows the frequency of these cluster classes to be encoded directly. Accuracy was predicted to be too high on numerous margins composed of coronal obstruents at the segmental level as well; however, allowing the model to abstract frequency over a feature such as [CORONAL] would be unlikely to help, since this would only serve to highlight how frequent these structures are as a class. It is important for future work to explore other theories of frequency and phonotactic probability that may improve on these results, but there seem to be genuine discrepancies between production accuracy and frequency in Polish that go beyond the specific assumptions made about input statistics here. Children produce certain structures unreliably even though they are abundantly represented in the input.

This suggests that some additional mechanism, besides sensitivity to input representations, is involved in phonological development. We have considered two candidates in our discussion: universal syllable structure markedness and articulatory difficulty. We found that the best frequency measures could not capture children's poor accuracy on clusters, codas, obstruent sequences, obstruent codas, stridents, affricates, trills, and retroflexes, nor could they fully capture children's high accuracy on low-sonority onsets. These results strongly suggest that language development is sensitive to many competing pressures: the language input influences the degree of exposure and practice children get with different configurations, but this influence is modulated by the inherent structural complexity and phonetic difficulty of these structures.

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26(1). 9–41.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.

Archer, Stephanie L. & Suzanne Curtin. 2011. Perceiving onset clusters in infancy. *Infant Behavior and Development* 34(4). 534–540.

Bane, Max, Peter Graff & Morgan Sonderegger. 2014. Longitudinal phonetic variation in a closed system. *Chicago Linguistic Society* 46(1). 43–58.

Barlow, Jessica A. 2007. Grandfather effects: A longitudinal case study of the phonological acquisition of intervocalic consonants in English. *Language Acquisition* 14(2). 121–164.

Batchelder, Eleanor Olds. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition* 83(2). 167–206.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87(1). 84–125.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14. 150–177.
Blanchard, Daniel, Jeffrey Heinz & Roberta Golinkoff. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language* 37(3). 487–511.
Blevins, Juliette. 1995. The syllable in phonological theory. In John A. Goldsmith (ed.), *The handbook of phonological theory*, 206–244. Cambridge, MA: Blackwell.
Boersma, Paul & Claartje Levelt. 2000. Gradual Constraint-Ranking Learning Algorithm predicts acquisition order. In Eve V. Clark (ed.), *Proceedings of 30th Child Language Research Forum*, 229–237. Stanford, CA: CSLI.
Brent, Michael R. & Timothy A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61(1–2). 93–125.
Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455.
Carpenter, Angela C. 2010. A naturalness bias in learning stress. *Phonology* 27(3). 345–392.
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.
Coady, Jeffry A. & Richard N. Aslin. 2004. Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology* 89(3). 183–213.
Compton, Arthur J. & Mary Streeter. 1977. Child phonology: Data collection and preliminary analyses. In Eve T. Clark & Pamela Tiedt (eds.), *Papers and reports on child language development 7*, 99–109. Stanford, CA: Stanford University.
Crawley, Michael J. 2013. *The R book*, 2nd edn. Hoboken, NJ: John Wiley & Sons.
Cristià, Alejandrina & Amanda Seidl. 2008. Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development* 4(3). 203–227.
Culbertson, Jennifer, Paul Smolensky & Colin Wilson. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science* 5(3). 392–424.
Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology* 28(2). 197–234.
Demuth, Katherine. 1995. Markedness and the development of prosodic structure. *North East Linguistic Society (NELS)* 25. 13–26.
Demuth, Katherine & Margaret Kehoe. 2006. The acquisiton of word-final clusters in French. *Catalan Journal of Linguistics* 5. 59–81.
Demuth, Katherine & Elizabeth McCullough. 2009. The longitudinal development of clusters in French. *Journal of Child Language* 36(2). 425–448.
Edwards, Jan & Mary E. Beckman. 2008. Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Language Learning & Development* 4(2). 122–156.
Edwards, Jan, Mary E. Beckman & Benjamin Munson. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language and Hearing Research* 47(2). 421–436.
Edwards, Jan, Mary E. Beckman & Benjamin Munson. 2015. Frequency effects in phonological acquisition. *Journal of Child Language* 42(2). 306–311.
Ernestus, Mirjam & R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1). 5–38.
Fee, Jane & David Ingram. 1982. Reduplication as a strategy of phonological development. *Journal of Child Language* 9 (1). 41–54.
Fikkert, Paula. 1994. *On the acquisition of prosodic structure*. Leiden, The Netherlands: University of Leiden dissertation.
Finley, Sara & William Badecker. 2009. Artificial language learning and feature-based generalization. *Journal of Memory and Language* 61(3). 423–437.
Gambell, Timothy & Charles Yang. 2006. Word segmentation: Quick but not dirty. Ms. Yale University.
Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
Gerken, LouAnn. 1996. Prosodic structure in young children's language production. *Language* 72(4). 683–712.
Gnanadesikan, Amelia. 2004. Markedness and faithfulness constraints in child phonology. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 73–109. Cambridge: Cambridge University Press.
Goad, Heather & Yvan Rose. 2004. Input elaboration, head faithfulness and evidence for representation in the acquisition of left-edge clusters in West Germanic. In René Kater, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 109–157. Cambridge: Cambridge University Press.
Goldrick, Matthew & Meredith Larson. 2008. Phonotactic probability influences speech production. *Cognition* 107(3). 1155–1164.
Goldwater, Sharon, Thomas L. Griffiths & Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1). 21–54.
Gussmann, Edmund. 1992. Resyllabification and delinking: The case of Polish voicing. *Linguistic Inquiry* 23(1). 29–56.

Harrell, Frank E. 2001. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.
Harrell, Frank E. 2014. RMS: Regression Modeling Strategies: R package version 4.1-3. http://biostat.mc.vanderbilt.edu/wiki/Main/RmS.
Hayes, Bruce & Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23(1). 59–104.
Hayes, Bruce & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44(1). 45–75.
Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.
Hayes, Bruce, Kie Zuraw, Péter Siptár & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4). 822–863.
Hockema, Stephen A. 2006. Finding words in speech: An investigation of American English. *Language Learning and Development* 2(2). 119–146.
Hua, Zhu & Barbara Dodd. 2000. The phonological acquisition of Putonghua (Modern Standard Chinese). *Journal of Child Language* 27(1). 3–42.
Ingram, David. 1988. The acquisition of word-initial [v]. *Language and Speech* 31(1). 77–85.
Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.
Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62.
Jakobson, Roman. 1968. *Child language, aphasia and phonological universals*. The Hague, The Netherlands: Mouton.
Jarosz, Gaja. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language* 37(3). 565–606.
Jarosz, Gaja. 2011. The roles of phonotactics and frequency in the learning of alternations. In Nick Danis, Kate Mesh & Hyunsuk Sung (eds.), *Proceedings of the 35th annual Boston University Conference on Language Development [BUCLD 35]*, 321–333. Somerville, MA: Cascadilla Press.
Jarosz, Gaja & J. Alex Johnson. 2013. The richness of distributional cues to word boundaries in speech to young children. *Language Learning and Development* 9(2). 175–210.
Jesney, Karen & Anne-Michelle Tessier. 2011. Biases in Harmonic Grammar: The road to restrictive learning. *Natural Language and Linguistic Theory* 29(1). 251–290.
Johnson, Mark. 2008. Unsupervised word segmentation for Sesotho using Adaptor Grammars. In *Proceedings of the 10th Meeting of ACL SIGMORPHON*, 20–27. Columbus, OH: Association of Computational Linguistics.
Kager, René & Joe Pater. 2012. Phonotactics as phonology: Knowledge of a complex restriction in Dutch. *Phonology* 29 (1). 81–111.
Kent, Ray D. 1992. The biology of phonological development. In Charles A. Ferguson, Lise Menn & Carol Stoel-Gammon (eds.), *Phonological development: Models, research, implications*, 65–90. Timonium, MD: York Press.
Kim, Minjung & Carol Stoel-Gammon. 2010. Phonological development of word-initial Korean obstruents in young Korean children. *Journal of Child Language* 38(2). 316–340.
Kirk, Cecilia & Katherine Demuth. 2005. Asymmetries in the acquisition of word-initial and word-final consonant clusters. *Journal of Child Language* 32(4). 709–734.
Kondrak, Grzegorz. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 288–295. Stroudsburg, PA: Association for Computational Linguistics.
Levelt, Claartje, Niels Schiller & Willem Levelt. 2000. The acquisition of syllable types. *Language Acquisition* 8(3). 237–264.
Levelt, Claartje & Ruben van de Vijver. 2004. Syllable types in cross-linguistic and developmental grammars. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 204–218. Cambridge: Cambridge University Press.
Lleó, Conxita & Michael Prinz. 1996. Consonant clusters in child phonology and the directionality of syllable structure assignment. *Journal of Child Language* 23(1). 31–56.
Łukaszewicz, Beata. 2006. Extrasyllabicity, transparency and prosodic constituency in the acquisition of Polish. *Lingua* 116(1). 1–30.
MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum.
Mattys, Sven L. & Peter W. Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78 (2). 91–121.
McAllister Byun, Tara. 2011. A gestural account of a child-specific neutralisation in strong position. *Phonology* 28(3). 371–412.
Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25(1). 83–127.
Moreton, Elliott & Joe Pater. 2012a. Structure and substance in artificial-phonology learning, Part I: Structure. *Language and Linguistics Compass* 6(11). 686–701.

Moreton, Elliott & Joe Pater. 2012b. Structure and substance in artificial-phonology learning, Part II: Substance. *Language and Linguistics Compass* 6(11). 702–718.
Munson, Benjamin. 2001. Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research* 44(4). 778–792.
Newport, Elissa L. & Richard N. Aslin. 2004. Learning at a distance I: Statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48(2). 127–162.
Ohala, Diane K. 1999. The influence of sonority on children's cluster reductions. *Journal of Communication Disorders* 32(6). 397–422.
Pater, Joe. 1997. Minimal violation and phonological development. *Language Acquisition* 6(3). 201–253.
Pater, Joe & Jessica Barlow. 2003. Constraint conflict in cluster reduction. *Journal of Child Language* 30(3). 487–526.
Pierrehumbert, Janet B. (2003). Probabilisitic phonology: Discrimination and robustness. In Rens Bod, Jennifer Hay & Stephanie Jannedy (eds.), *Probabilistic linguistics* (pp. 177–228). Cambridge, MA: MIT Press.
Proctor, Michael. 2009. *Gestural characterization of a phonological class: The liquids*. New Haven, CT: Yale University dissertation.
Pycha, Anne, Pawel Nowak, Eurie Shin & Ryan Shosted. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. In Gina Garding & Mimu Tsujimura (eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL)*,101–114. Somerville, MA: Cascadilla Press.
R Development Core Team. 2008. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
Richtsmeier, Peter T. 2011. Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology* 2(1). 157–183.
Richtsmeier, Peter T., Louann Gerken, Lisa Goffman & Tiffany P. Hogan. 2009. Statistical frequency in perception affects children's lexical production. *Cognition* 111. 372–377.
Richtsmeier, Peter T., Louann Gerken & Diane K. Ohala. 2009. Induction of phonotactics from word-types and word-tokens. In Jane Chandlee, Michelle Franchini, Sandy Lord & Gudrun-Marion Rheiner (eds.), *Proceedings of the 33rd Boston University Conference on Language Development [BUCLD 33]*, 432–443. Somerville, MA: Cascadilla Press.
Richtsmeier, Peter T., Louann Gerken & Diane K. Ohala. 2011. Contributions of phonetic token variability and word-type frequency to children's phonological representations. *Journal of Child Language* 38(5). 951–978.
Roark, Brain & Katherine Demuth. 2000. Prosodic constraints and the learner's environment: A corpus study. In S. Catherine Howell, Sarah A. Fish & Thea Keith-Lucas (eds.), *Proceedings of the 24th annual Boston University Conference on Language Development [BUCLD 24]*, 597–608. Somerville, MA: Cascadilla Press.
Roland, Douglas, Jeffrey L. Elman & Victor S. Ferreira. 2006. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition* 98(3). 245–272.
Rose, Yvan. 2000. *Headedness and prosodic licensing in the L1 acquisition of phonology*. Montreal: McGill University dissertation.
Rose, Yvan, Brian MacWhinney, Rodrigue Byrne, Gregory Hedlund, Keith Maddocks, Philip O'Brien & Todd Wareham. 2006. Introducing Phon: A software solution for the study of phonological acquisition. In David Bamman, Tatiana Magnitskaia & Colleen Zaller (eds.), *Proceedings of the 30th annual Boston University Conference on Language Development [BUCLD 30]*, 489–500. Somerville, MA: Cascadilla Press.
Rubach, Jerzy & Geert E. Booij. 1985. A grid theory of stress in Polish. *Lingua* 66(4). 281–320.
Rubach, Jerzy & Geert Booij. 1990. Syllable structure assignment in Polish. *Phonology* 7(1). 121–158.
Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274 (5294). 1926–1928.
Salidis, Joanna & Jacqueline S. Johnson. 1997. The production of minimal words: A longitudinal case study of phonological development. *Language Acquisition* 6(1). 1–36.
Seidl, Amanda & Eugene Buckley. 2005. On the learning of arbitrary phonological rules. *Language Learning and Development* 1(3). 289–316.
Selkirk, Elisabeth. 1984. On the major class features and syllable theory. In Mark Aronoff & Richard T. Oehrle (eds.), *Language sound structure: Studies in phonology*, 107–136. Cambridge, MA: MIT Press.
Smit, Ann Bosma, Linda Hand, J. Joseph Freilinger, John E. Bernthal & Ann Bird. 1990. The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders* 55(4). 779–798.
Stampe, David. 1969. The acquisition of phonetic representations. *Chicago Linguistic Society (CLS)* 5. 443–454.
Stites, Jessica, Katherine Demuth & Cecilia Kirk. 2004. Markedness vs. frequency effects in coda acquisition. In Alejna Brugos, Linnea Micciulla & Christine E. Smith (eds.), *Proceedings of the 28th annual Boston University Conference on Language Development [BUCLD 28]*, 565–576. Somerville, MA: Cascadilla Press.
Stoel-Gammon, Carol. 1998. Sounds and words in early language acquisition: The relationship between lexical and phonological development. In Rhea Paul (ed.), *Exploring the speech–language connection*, 25–52. Baltimore: Brookes.
Tessier, Anne-Michelle. 2009. Frequency of violation and constraint-based phonological learning. *Lingua* 119(1). 6–38.
Venkataraman, Anand. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics* 27(3). 352–372.

Vihman, Marilyn M. 1992. Early syllables and the construction of phonology. In Charles A. Ferguson, Lise Menn & Carol Stoel-Gammon (eds.), *Phonological development: Models, research, implications*, 393–422. Timonium, MD: York Press.
Vihman, Marilyn M. 1993. Variable paths to early word production. *Journal of Phonetics* 21. 61–82.
Weist, Richard M. & Katarzyna Witkowska-Stadnik. 1986. Basic relations in child language and the word order myth. *International Journal of Psychology* 21 (1–4). 363–381.
Weist, Richard M., Hanna Wysocka, Katarzyna Witkowska-Stadnik, Ewa Buczowska & Emilia Konieczna. 1984. The defective tense hypothesis: On the emergence of tense and aspect in child Polish. *Journal of Child Language* 11(2). 347–374.
Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30(5). 945–982.
Zamuner, Tania S. 2009. Phonotactic probabilities at the onset of language development: Speech production and word position. *Journal of Speech, Language & Hearing Research* 52(1). 49–60.
Zamuner, Tania S., LouAnn Gerken & Michael Hammond. 2004. Phonotactic probabilities in young children's speech production. *Journal of Child Language* 31(3). 515–536.
Zamuner, Tania S., LouAnn Gerken, & Michael Hammond. 2005. The acquisition of phonology based on input: A closer look at the relation of cross-linguistic and child language data. *Lingua* 115(10). 1403–1426.
Zec, Draga. 2007. The syllable. In Paul de Lacy (ed.), *The Cambridge handbook of phonology*, 161–194. Cambridge: Cambridge University Press.

Appendix A

Table A1. Regression Model for Individual Sonority Margin Types.

Predictor	*β*	*SE*	*z*	*p*	*Predictor*	*β*	*SE*	*z*	*p*
MARGIN TYPE									
#P	1.79	0.20	9.1	<0.0001	#L	0.48	0.24	2.0	<0.047
#PP	− 1.53	0.27	− 5.7	<0.0001	#G	0.37	0.22	1.7	<0.087
#PF	− 0.84	0.24	− 3.6	<0.0005	#V	1.77	0.24	7.3	<0.0001
#PL	− 0.50	0.22	− 2.3	<0.024	P#	0.11	0.19	0.6	>0.57
#PG	− 0.50	0.27	− 1.9	<0.062	FP#	− 2.17	0.36	− 6.1	<0.0001
#F	0.37	0.20	1.9	<0.062	NP#	− 0.71	0.31	− 2.3	<0.021
#FP	− 1.69	0.21	− 7.9	<0.0001	F#	− 0.24	0.20	− 1.2	>0.23
#FF	− 0.81	0.34	− 2.4	<0.017	PF#	− 1.78	0.51	− 3.5	<0.0005
#FN	0.15	0.45	0.3	>0.73	N#	0.71	0.20	3.5	<0.0005
#FL	− 1.24	0.33	− 3.8	<0.0005	L#	0.23	0.30	0.8	>0.43
#FG	0.11	0.24	0.5	>0.64	G#	0.07	0.19	0.4	>0.71
#N	1.08	0.20	5.4	<0.0001	V#	1.58	0.19	8.4	<0.0001
#NG	0.75	0.52	1.4	<0.15	other	− 1.40	0.32	− 4.4	<0.0001
AGE	0.04	0.01	7.7	<0.0001	WORD FREQUENCY	0.04	0.01	2.9	<0.005
STRESS	0.24	0.07	3.3	<0.001	WORD LENGTH	− 0.09	0.03	− 2.6	<0.01
VOWEL									
ɛ	− 0.17	0.06	− 2.7	<0.01					
i	− 0.02	0.09	− 0.2	>0.82					
ɨ	− 0.13	0.11	− 1.2	>0.22					
ɔ	− 0.35	0.07	− 5.2	<0.0001					
u	− 0.23	0.09	− 2.6	<0.01					

Table A2. Regression Model for Individual Segment Margin Types.

Predictor	*β*	*SE*	*z*	*p*	*Predictor*	*β*	*SE*	*z*	*p*
MARGIN TYPE									
#b	2.23	0.37	6.1	<0.0001	#vj	0.55	0.41	1.3	<0.18
#bj	0.04	0.58	0.1	>0.95	#vw	1.30	0.48	2.7	<0.01
#ɕ	0.82	0.33	2.5	<0.013	#w	0.33	0.51	0.6	>0.51
#d	1.67	0.34	4.9	<0.0001	#x	1.86	0.50	3.7	<0.0005
#dr	0.15	0.42	0.3	>0.72	#xt͡s	−2.99	0.47	−6.3	<0.0001
#d͡ʑ	−0.33	0.37	−0.9	>0.38	#z	−0.39	0.33	−1.2	>0.23
#g	0.40	0.45	0.9	>0.37	#zb	−0.64	0.64	−1.0	>0.31
#gd͡ʑ	−1.28	0.39	−3.3	<0.005	#zd	0.35	0.64	0.6	>0.58
#gr	−2.24	0.57	−4.0	<0.0001	#zw	0.30	0.59	0.5	>0.61
#gw	−0.67	0.65	−1.0	>0.30	#ʐ	−2.09	0.47	−4.4	<0.0001
#j	0.94	0.32	2.9	<0.005	#V	2.22	0.34	6.6	<0.0001
#k	0.91	0.33	2.8	<0.01	ɕ#	1.02	0.40	2.6	<0.011
#kɕ	−0.63	0.61	−1.0	>0.30	ɕt͡ɕ#	−2.65	0.56	−4.7	<0.0001
#kl	−0.01	0.51	0.0	>0.97	d#	−0.61	0.53	−1.2	>0.24
#kr	−0.87	0.47	−1.9	<0.063	f#	0.25	0.63	0.4	>0.69
#kt	−0.87	0.50	−1.7	<0.084	g#	−0.60	0.40	−1.5	<0.14
#l	0.91	0.38	2.4	<0.016	j#	0.43	0.33	1.3	<0.20
#m	1.38	0.32	4.4	<0.0001	k#	0.50	0.32	1.6	<0.12
#n	1.32	0.35	3.8	<0.0005	l#	0.91	0.59	1.6	<0.13
#ɲ	0.92	0.31	3.0	<0.005	m#	1.16	0.32	3.6	<0.0005
#p	1.71	0.32	5.4	<0.0001	n#	1.10	0.36	3.1	<0.005
#pj	0.73	0.48	1.5	<0.14	nt#	−1.18	0.62	−1.9	<0.059
#pr	−1.70	0.47	−3.6	<0.0005	ɲ#	−0.98	0.52	−1.9	<0.059
#pʂ	−2.72	0.47	−5.8	<0.0001	ɲt͡ɕ#	−0.95	0.63	−1.5	<0.14
#pw	0.42	0.51	0.8	>0.40	p#	−0.07	0.54	−0.1	>0.89
#r	−0.33	0.35	−1.0	>0.34	r#	−0.79	0.40	−2.0	<0.048
#s	−0.62	0.35	−1.7	<0.081	s#	−1.60	0.43	−3.7	<0.0005
#sk	−1.53	0.60	−2.6	<0.011	ʂ#	−1.62	0.35	−4.6	<0.0001
#sp	−1.35	0.51	−2.7	<0.01	t#	1.39	0.50	2.8	<0.01
#st	−1.20	0.50	−2.4	<0.016	t͡ɕ#	−0.27	0.34	−0.8	>0.41
#sw	0.43	0.60	0.7	>0.46	t͡s#	−1.90	0.53	−3.6	<0.0005
#sx	−0.58	0.56	−1.0	>0.30	t͡ʂ#	−1.81	0.52	−3.5	<0.001
#ʂ	−1.72	0.59	−2.9	<0.005	w#	−0.38	0.32	−1.2	>0.23
#t	1.82	0.31	5.9	<0.0001	w̃#	1.04	0.31	3.3	<0.001
#tr	−0.18	0.40	−0.5	>0.64	z#	−1.08	0.66	−1.6	<0.11
#t͡ɕ	0.43	0.39	1.1	>0.26	ʐ#	−1.88	0.46	−4.1	<0.0001
#t͡s	−0.64	0.36	−1.8	<0.078	ʑ#	−1.00	0.54	−1.9	<0.062
#t͡ʂ	−1.31	0.42	−3.1	<0.005	V#	2.09	0.30	6.9	<0.0001
#v	1.82	0.36	5.0	<0.0001	other	−0.95	0.31	−3.1	<0.005
WORD LENGTH	−0.11	0.03	−4.0	<0.0001	AGE	0.03	0.01	2.5	<0.014
VOWEL					SUBJECT				
ɛ	−0.41	0.07	−5.9	<0.0001	Kubuś	0.01	0.11	0.1	>0.90
i	−0.16	0.09	−1.8	<0.079	Marta	−0.17	0.07	−2.4	<0.015
ɨ	−0.44	0.11	−4.0	<0.0001	Wawrzon	0.11	0.12	0.9	>0.38
ɔ	−0.37	0.07	−5.4	<0.0001					
u	−0.33	0.09	−3.7	<0.0005					

Appendix B

Table B1. Summary of Model Simplification.

Hypothesis/Model	*df*	χ^2	*p*	D_{xy}	*AIC*
CV Level					
Maximal CV Model				0.440	11 910
$H_0 : \beta(\#V) = \beta(V\#) = \beta(\#C)$	2	1.4	>0.5	0.438	11 908
Sonority Level					
Maximal Sonority Model + Reduced CV Predictors				0.494	13 791
Remove parameter for each of SON TYPE = {#FG, #FN, #P, G#, P#, PF#, FP#, other, L#, #NG, V#}	11	9.3	>0.59	0.492	13 783
$H_0 : \beta(\#L) = \beta(\#G) = \beta(\#F)$	2	0.4	>0.82	0.491	13 779
$H_0 : \beta(\#PG) = \beta(\#PL) = \beta(\#FF) = \beta(\#PF) = \beta(\#FL)$	4	8.3	>0.08	0.491	13 779
$H_0 : \beta(\#FP) = \beta(\#PP)$	1	0.55	>0.45	0.491	13 778
$H_0 : \beta(\#CC) = \beta(C\#)$	1	0.07	>0.78	0.491	13 776
Segment Level					
Maximal Segmental Model + Reduced CV & Reduced SON Predictors				0.598	14 834
Remove parameter for each of SEG TYPE = {#bj, #b, #dr, #gw, #kɕ, #kl, #kt, #ɲ, #pj, #pw, #st, #sw, #sx, #tr, #v, #vw, #vj, #x, #zb, #zd, #zw, V#, ɕ#, l#, ɲ#, ɲt͡ɕ#, nt#, t#, w̃#}	27	31.9	>0.23	0.596	14 815
$H_0 : \beta(\#m) = \beta(\#n)$	1	0.07	>0.78	0.596	14 813
$H_0 : \beta(\#p) = \beta(\#t) = \beta(\#d)$	2	0.64	>0.72	0.596	14 809
$H_0 : \beta(\#j) = \beta(\#l) = \beta(\#k) = \beta(\#ɕ) = \beta(\#w) = \beta(\#g) = \beta(\#t͡ɕ)$	6	8.4	>0.21	0.595	14 806
$H_0 : \beta(\#r) = \beta(\#s) = \beta(\#z) = \beta(\#d͡ʑ) = \beta(\#t͡s)$	4	3	>0.55	0.594	14 801
$H_0 : \beta(\#ʂ) = \beta(\#ʐ) = \beta(\#t͡ʂ)$	2	2.6	>0.27	0.594	14 799
$H_0 : \beta(m\#) = \beta(n\#)$	1	0.06	>0.80	0.594	14 797
$H_0 : \beta(k\#) = \beta(j\#) = \beta(f\#) = \beta(p\#)$	1	2	>0.57	0.594	14 793
$H_0 : \beta(w\#) = \beta(d\#) = \beta(g\#) = \beta(t͡ɕ\#) = \beta(r\#)$	4	3.5	>0.47	0.594	14 789
$H_0 : \beta(s\#) = \beta(ʂ\#) = \beta(ʐ\#) = \beta(t͡s\#) = \beta(t͡ʂ\#) = \beta(ʑ\#) = \beta(z\#)$	6	3.2	>0.77	0.594	14 780
$H_0 : \beta(\#kr) = \beta(\#sp) = \beta(\#sk) = \beta(\#gd͡ʑ)$	3	0.21	>0.97	0.594	14 774
$H_0 : \beta(\#gr) = \beta(\#pr) = \beta(\#pʂ) = \beta(\#xt͡s)$	3	4.5	>0.21	0.594	14 773
Remove SON TYPE = #L or #G or #F, SON TYPE = F#, SON TYPE= NP#	3	3	>0.39	0.594	14 770