Prosodic Constraints and the Learner's Environment: a Corpus Study

Brian Roark & Katherine Demuth
Brown University

1 Introduction

Jakobson (1968) was one of the first to propose that children's early utterances would exhibit 'unmarked' linguistic structures. For example, he predicted that children would first use segments such as /p/, /n/, and /a/ that are widely attested in the world's languages. Given the variability found in children's early utterances, even within a specific language, this hypothesis has been somewhat difficult to test, and has therefore remained somewhat controversial.

Recent developments in phonological theory, where 'unmarked' structures such as binary feet (or 'minimal words') play an important role, have renewed interest in the status of 'unmarked' structures in natural language (e.g. McCarthy and Prince, 1994). Researchers of language acquisition have been quick to recognize the importance of such developments, especially with respect to higher level prosodic units such as syllables and prosodic words (e.g. Fikkert, 1994; Demuth, 1995; Gnanadesikan, 1995; Pater, 1997; Ota, 1999). For example, it has been noted that many children learning English and Dutch initially avoid the use of coda consonants, producing early words with unmarked 'core' CV syllables, where CVC word targets are often realized as CVCV, with an epenthetic final vowel:

(1) /bʊk/ → [bʊkə] Matthei (1989)

Markedness constraints seem to operate at the level of prosodic words as well, with children learning English and Dutch tending to produce words that conform to a trochaic foot (either ˈCVCV, or CVC once codas can be produced). This entails the deletion of syllables that cannot be mapped into a binary foot, with pretonic syllables being especially vulnerable to omission:

(2) /bəˈnænə/ → [ˈnænə] Matthei (1989)

1.1 Emergence of the Unmarked

With the development of Optimality Theory (Prince and Smolensky, 1993), researchers have realized that children's early prosodic structures can best be understood in terms of a series of hierarchically interacting linguistic 'constraints'

(cf. Demuth, 1995; Gnanadesikan, 1995; Pater, 1997; Ota, 1999). For example, the lack of early codas indicates the relative high ranking of a constraint *Coda (No-Coda). CVC target forms are then realized as either CVCV (with an epenthetic vowel) or CV (with a deleted segment) depending on the relative ranking of faithfulness constraints on the realization of segments. Likewise, the early omission of pretonic unfooted syllables indicates the higher ranking of markedness constraints such as Exhaustivity (parse all syllables into feet) over faithfulness constraints that map segments of the lexical (input) form into the surface (output) realization of the word.

In sum, research over the past 5 years indicates that Markedness (or structural) constraints tend to dominate Faithfulness constraints in children's early grammars, i.e.:

(3) Markedness >> Faithfulness

1.2 Crosslinguistic Differences in the Acquisition of 'Marked' structures

Children learning languages like English and Dutch therefore appear to start with 'unmarked' structures in their early grammars, and gradually move to more 'marked' types of structures given positive evidence from the input. Or at least this is what the literature has led us to believe. However, evidence from Spanish indicates that crosslinguistic differences in early prosodic word shape occur much earlier than initially thought (Gennari and Demuth, 1997; Demuth, in press; Lleó and Demuth, 1999). In particular, the timing of the appearance of marked structures differs depending on the language being learned, with unfooted syllables appearing much earlier in Spanish, and coda consonants appearing much earlier in Germanic languages like English, Dutch, and German, as illustrated in Figure 1.

These differences in the course of acquisition can be captured in terms of the reranking of constraints, where *Coda is demoted earlier in English, and Exhaustivity is demoted earlier in Spanish. Thus, at a stage of acquisition where some markedness constraints are being demoted, the relative ranking of constraints can

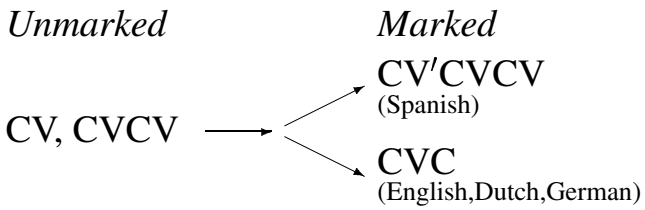

Figure 1: Differences in the timecourse of word shape acquisition

be characterized as follows:

(4) English: Exhaustivity >> *Coda
(5) Spanish: *Coda >> Exhaustivity

A constraint-based perspective on these issues is useful in terms of describing the phenomena found. However, it does not provide an explanatory understanding of why these crosslinguistic differences should occur. This is a general problem found in theories of parameter-setting as well. That is, what triggers the resetting of a parameter, or the reranking of a constraint? Given that both English and Spanish permit coda consonants, and that the 'cue' for coda consonants is therefore 'available' for both languages from the beginning, why do codas tend to appear earlier in English? We suggest that frequency effects, or the relative 'strength' of cues to constraint reranking, play an important role in determining the time course of acquisition. That is, young language learners are sensitive to statistical properties of the input, and this influences the course of language development.

1.3 Sensitivity to the Statistical Properties of the Input

Although language acquisition cannot be completely accounted for by the statistical frequency of grammatical phenomena (e.g. grammatical morphemes), there is ample evidence that infants and children are sensitive to the frequency with which certain linguistic phenomena occur, and tailor their early grammars accordingly. For example, infants are extremely sensitive to high-frequency phonological and syllabic effects of the ambient language (Jusczyk, 1997; Morgan, 1996; Saffran, Aslin, and Newport, 1996). But this sensitivity is not restricted to phonology alone: Demuth (1989; 1990) shows that children are sensitive to the frequency of constructions such as passives as well, and research in the area of psycholinguistic processing confirms that adults take longer to process low frequency constructions even though these are grammatical (e.g. MacDonald, 1994). Furthermore, work by Cutler and Norris (1988) and colleagues indicates that sensitivities to stress and to word shape are maintained by adults in the context of second language learning. Thus, it appears that theories of both language learning and language use must allow for frequency effects. This requires further research into what the relevant statistical properties of the input language actually are.

To address this issue given the problem children's early word productions, it is necessary to determine the relative frequency with which coda consonants and unfooted syllables appear in languages like English and Spanish. Given the acquisition observations mentioned above, we predict that coda consonants will appear much more frequently in English, and that unfooted syllables will appear much more frequently in Spanish. If these predictions are upheld, then our hypothesis that frequency effects play a role in the time course of constraint demotion will be confirmed. In order to carry out this investigation we examined a large set of corpora containing adult child-directed speech, as encoded in the CHILDES database

(MacWhinney, 1996).

2 Method

Very generally, we used the following method: first, we extracted parent utterances from corpora in the CHILDES database; from these utterances, we created a lexicon, and assigned a prosodic structure to each word in the lexicon; finally, using each word's prosodic structure, we returned to the extracted utterances and counted the frequency of occurrence of the prosodic structures of interest. In this case, we were interested in the syllable structures (e.g. CV,CVC), in particular the occurrence of coda consonants, and in the primary stress location for each word token.

This general method, as was noted in Swingley (1999), makes a couple of simplifying assumptions. First, by treating the prosodic structure as strictly lexical, it removes any variability that may be present in the actual speech through contextual effects (e.g. prosodic processes such as resyllabification). Further, it assumes that the words are always pronounced identically, and that the relevant cues (in this case coda consonants and lexical stress) are always perceptible in the continuous speech stream. Working from the CHILDES corpora, where parent utterances are almost never transcribed phonetically, these assumptions are necessary to get the project off of the ground. We did, however, write a routine to simulate cliticization, which is a step towards accounting for prosodic processes, and this is described below.

To obtain our sample parent utterances, we followed slightly different strategies for English and Spanish, as a result of the amount of data available in each language. In English, we extracted, from each corpus in the CHILDES database, all utterances by MOT and FAT (standard identification tags for the mother of the target child and the father of the target child) when there was a single target child identified as CHI. In this way, we were able to quickly gather a corpus that we are confident consists in large part of child-directed utterances, from the people that produce a significant amount of the child's language environment. Further, we were able to identify the unique target child's age. This yielded over 450,000 utterances. The amount of Spanish data is far less, so we were able to review each corpus for the name of the father, mother, and target child, in the case that they were not identified with the standard tags, and all such cases were included in our sample. This yielded approximately 18,000 parent utterances in Spanish.

There were two primary questions to be addressed: how different are the two languages with respect to the distributions of coda consonants and primary stress; and how variable are these distributions within the two languages. With this within/between language variation question in mind, we divided the Spanish utterances into nine equally sized samples of 2,000 utterances, and then randomly selected nine English samples of the same size. Note that the samples were chopped blindly from the corpus, so that each sample contained utterances from more than one session. The 18,000 utterances in English and Spanish contained

approximately 80,000 words each.

From these samples, we created a lexicon in each language, and assigned a syllable structure and primary stress location to each word in the lexicon. For English, we used an on-line English dictionary (Parks, 1999), which provided the appropriate information for the bulk of the words. The remainder were assigned by hand. Spanish has an orthography which is much more systematic than that of English, and we wrote a deterministic syllable parser to assign the structure, using conventional rules and exceptions (Kattán-Ibarra and Pountain, 1997). Those words that failed to receive a structure from the parser were reviewed, and a structure was assigned by hand.

In addition to this, we wrote a routine to attach clitics within utterances to their neighbor 'host' words, to form larger prosodic units. Clitics are words that do not, in general, carry primary stress, and they prosodify as an additional unstressed syllable in words that occur either to their right or left. For example, *the circle* prosodifies as a single prosodic word, with a weak initial syllable (the), followed by strong and weak syllables (cir-cle). This routine to attach clitics is an attempt to account for some of the prosodic processes that occur in speech. While it does not exhaustively account for these processes, it does give an indication of the impact of such a process on the distributions that we are examining. Results will be provided both with and without this cliticization process.

3 Results and discussion

The first thing to look at in trying to get an idea of the differences in the distributions between the two languages is the actual distribution of words with particular shapes, i.e. number of syllables and stress placement. Figure 2 shows the frequency of occurrence in each language of different word shapes, after the cliticization process was applied. Even with cliticization (which in English applied to the determiners *the* and *a*) monosyllabic and disyllabic words with initial stress account for about 90 percent of the tokens in English. In contrast, the Spanish distribution is much more balanced, with disyllabic iambic stressed tokens and multisyllabic tokens with penultimate stress accounting for over thirty percent of the tokens. Interestingly, nearly thirty percent of the tokens in Spanish are monosyllabic, which may run counter to expectations. This is due to the high frequency of a small number of generally closed class words. The following ten words account for 88 percent of the monosyllabic tokens in the Spanish sample: *con, en, es, no, por, qúe, sí, ver, y, ya*. In contrast, the most frequent ten words in English (*and, do, to, that, is, it, a, what, the, you*) account for only 30 percent of the monosyllabic tokens. Overall, the monosyllabic words in Spanish consist of a small number of generally closed class items, whereas in English, the set of monosyllabic words is much larger, and contains many open class in addition to closed class words.

Table 1 gives the mean percentages and standard deviations for all of the measures displayed in the subsequent graphs. Figures 3 and 4 show the mean percent-

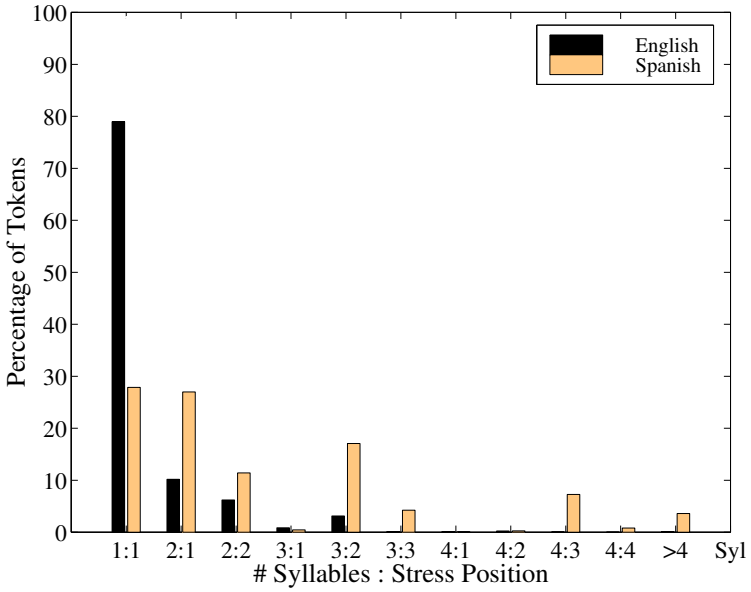

Figure 2: Distribution of cliticized word shapes in Spanish and English

age, across the nine samples, of weak initial syllables in both languages, before and after cliticization respectively. The means are dramatically different, even before cliticization, but more interestingly, the standard deviation across samples (shown in the graph as an error bar) is remarkably small, indicating that these frequencies are very stable in each language. Cliticization increased the gap between the two languages, but also further reduced the standard deviation.

The same situation holds with respect to the frequency of coda consonants in the two languages. Figure 5 shows the mean percentage, across the nine random samples, of coda consonants in the two languages. Once again, the differences across the two languages is very large, and the standard deviation within the lan-

Language	Cliticized?	Weak Initial Syllable		Coda Consonants	
		Mean	Standard Deviation	Mean	Standard Deviation
English	No	3.8	1.33	59.3	2.46
	Yes	10.0	1.25		
Spanish	No	28.3	4.03	25.2	1.04
	Yes	44.6	2.42		

Table 1: Mean percentages and std. deviations from English and Spanish samples

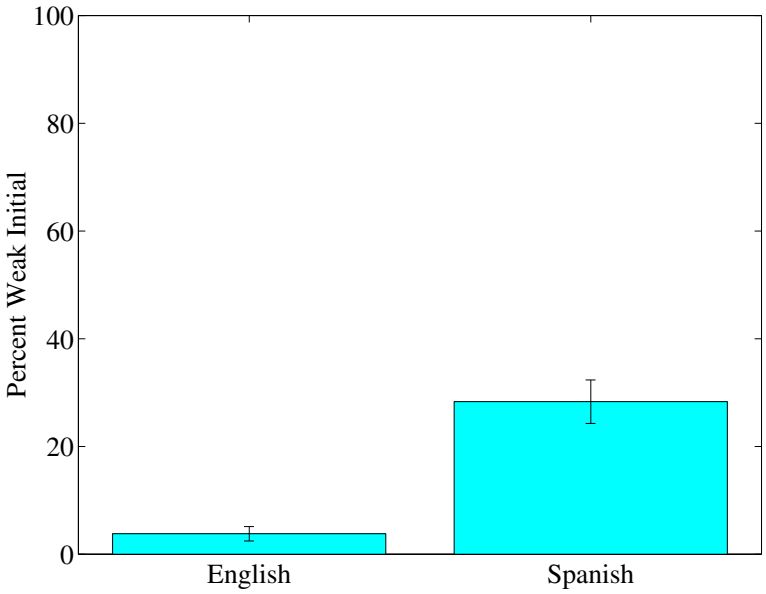

Figure 3: Mean percentage of weak initial syllables, before cliticization

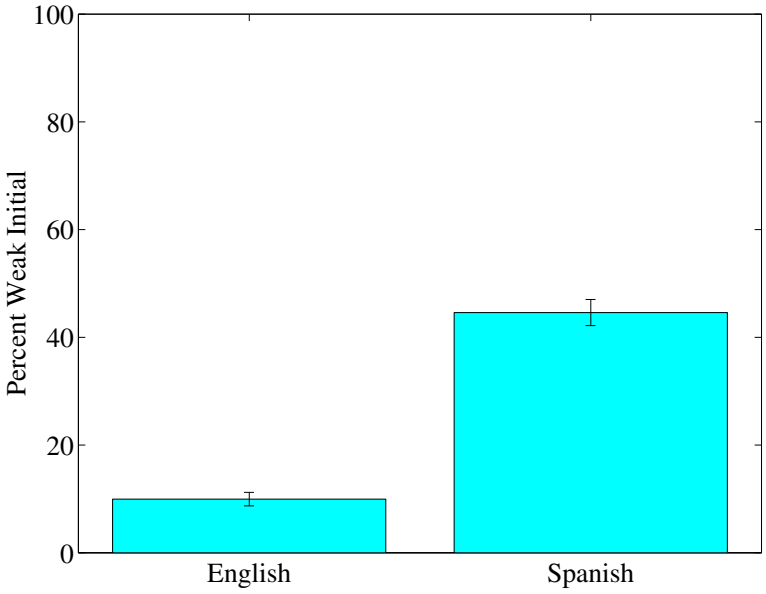

Figure 4: Mean percentage of weak initial syllables, after cliticization

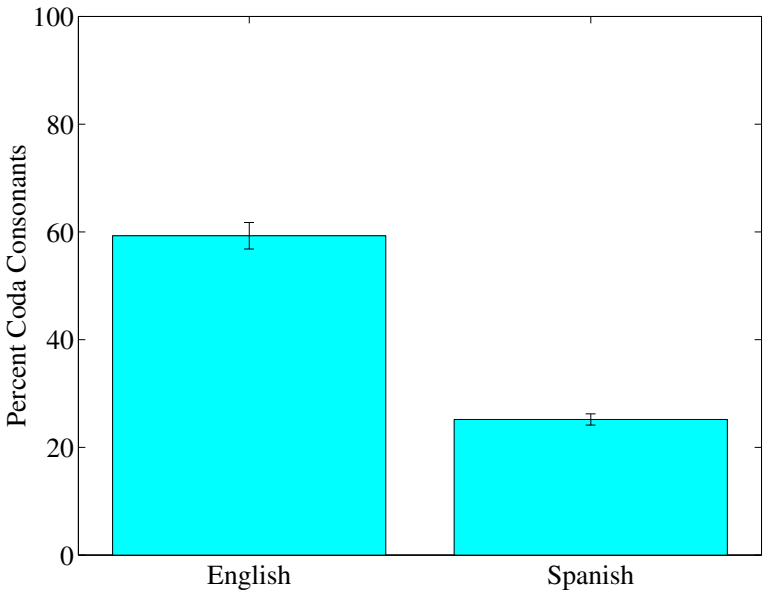

Figure 5: Mean percentage of coda consonants

guages is astonishingly small.

One question that might be asked at this point is whether or not any difference might be expected across samples with different target child ages. It could be that the random samples are balanced enough so that we are missing differences in the environments of younger children versus that of older children. Such a difference would be particularly pertinent since it has been demonstrated that the timecourse of acquisition in the two languages diverges at a very early age. To address this question, we divided our 18,000 utterance samples by the age of the target child, and performed the same measures. Figures 6 and 7 show the mean percentage of coda consonants and weak initial syllables, respectively, by the age of the target child. As these graphs dramatically illustrate, there is no difference with respect to these measures of the parents' speech as the age of the child varies.

What is particularly suggestive about these graphs is that the raw frequency level corresponds in a direct way to the observed age of acquisition in the two languages. The highest frequency that we found was for coda consonants in English, at just about sixty percent of syllables. Some researchers have claimed that the production of coda consonants in English speaking children can occur as early as the babbling stage, or with the child's very first words (Salidis and Johnson, 1997). Next most frequent are weak initial syllables in Spanish, which have been shown to occur at before 1;6 years (Lleó, 1997). Coda consonants in Spanish, which occur in our corpus 25 percent of the time, are generally acquired around

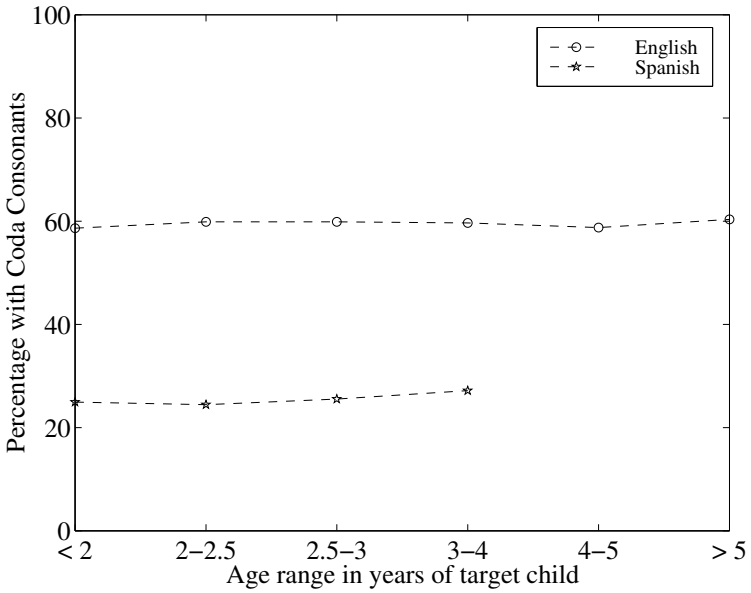

Figure 6: Mean percentage of coda consonants, by age of target child

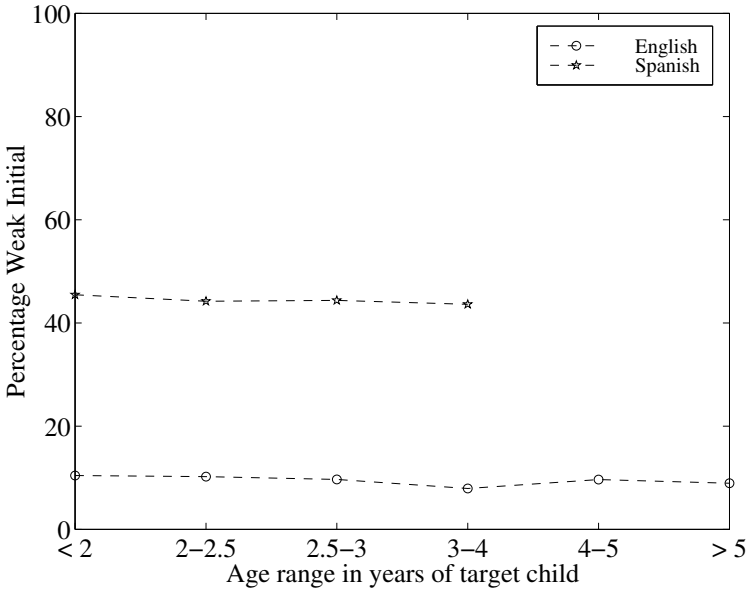

Figure 7: Mean percentage of weak initial syllables, by age of target child

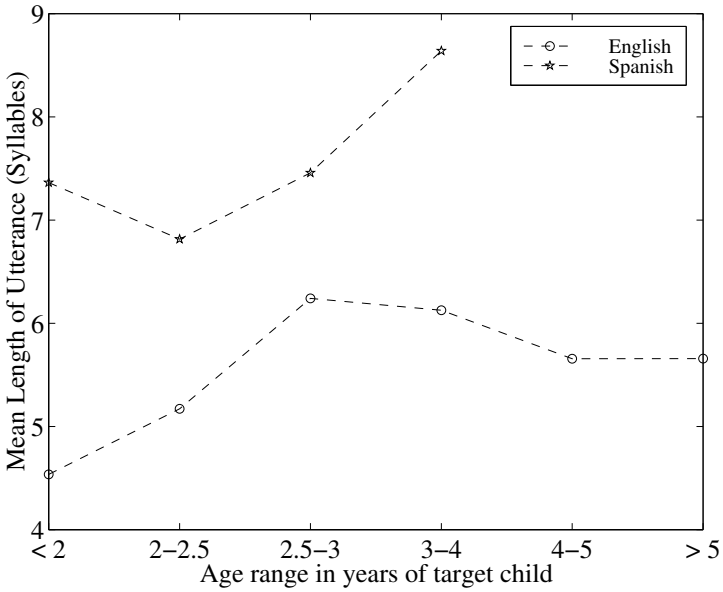

Figure 8: Mean length of utterance (syllables), by age of target child

1;10 (Gennari and Demuth, 1997; Lleó, 1997); and weak initial syllables in English can appear as late as two and a half years (Smith, 1973; Gerken, 1994)! This suggests that the raw frequency of the features in the input may impact directly the timecourse of acquisition.

Fearing that any measure of the parent utterances would have this kind of invariability over the age of the target child, we also measured something that we would expect to differ by the age of the child: the mean length of utterance. For simplicity, we measured this in syllables, and Figure 8 shows the results. While the trends in this graph are difficult to interpret, due to the sparsity of data outside of the 2-3 year age range, there is certainly some difference in length between the ages, which contrasts markedly with the flat trends shown previously. What this establishes is that the uniformity observed with respect to coda consonants and primary stress is not the result of some kind of general uniformity of the samples.

4 Conclusions

In conclusion, we have shown that certain prosodic properties of language are very statistically stable, with frequencies that mirror certain differences in the timecourse of acquisition across languages. This argues for a statistical component to theories of language learning, wherein the salience of a particular cue, here measured in terms of raw frequency, can influence, for example, constraint

re-ranking in Optimality Theory.

That said, there are some large unresolved questions. How might these kinds of statistics influence learning? Is it a question of reaching some threshold before a constraint is demoted (or a parameter set)? If so, is it a cumulative effect, or does the threshold change over time? Furthermore, what besides raw frequency may indicate salience? Might there not be some more complex interactions between different features of a language that make certain items, of perhaps a fairly low frequency, salient in some way to the learner, and thus acquired earlier than other, more frequent, features? We hope that this paper will stimulate further research along these lines.

References

Cutler, Anne, and Dennis G. Norris. 1988. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14:113–121.

Demuth, Katherine. 1989. Maturation and the acquisition of sesotho passive. *Language*, 65:56–80.

Demuth, Katherine. 1990. Subject, topic and the sesotho passive. *Journal of Child Language*, 17:67–84.

Demuth, Katherine. 1995. Markedness and the development of prosodic structure. In *Proceedings of the North East Linguistic Society 25*, pages 13–25. GLSA, University of Massachusetts.

Demuth, Katherine. in press. Prosodic constraints on morphological development. In J. Weissenborn and B. Höhle, editors, *Approaches to Bootstrapping: Phonological, Syntactic, and Neurophysiological Aspects of Early Language Acquisition*. John Benjamins, Amsterdam.

Fikkert, Paula. 1994. *On the acquisition of prosodic structure*. Ph.D. thesis, University of Leiden.

Gennari, Silvia, and Katherine Demuth. 1997. Syllable omission in spanish. In *The Proceedings of the 21st Annual Boston University Conference on Language Development*, pages 182–193.

Gerken, LouAnne. 1994. A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of Child Language*, 21:565–584.

Gnanadesikan, Amalia. 1995. Markedness and faithfulness constraints in child phonology. Ms., University of Massachusetts, Amherst (ROA).

Jakobson, Roman. 1968. *Child language, aphasia, and phonological universals*. Mouton, The Hague & Paris. Originally published in 1941.

Jusczyk, Peter W. 1997. *The discovery of spoken language*. MIT Press, Cambridge, MA.

Kattán-Ibarra, Juan, and Christopher J. Pountain. 1997. *Modern Spanish Grammar: a Practical Guide*. Routledge, New York, NY.

Lleó, Conxita, and Katherine Demuth. 1999. Prosodic constraints on the emergence of grammatical morphemes. In *The Proceedings of the 23nd Annual Boston University Conference on Language Development*.

Lleó, Conxita. 1997. Filler syllables, proto-articles and early prosodic constraints in spanish and german. In *Proceedings of GALA*, pages 251–256.

MacDonald, Maryellen C. 1994. Probabilistic constraints and syntactic ambigu-

ity resolution. *Language and Cognitive Processes*, 9(2):157–201.
MacWhinney, Brian. 1996. The CHILDES system. *American Journal of Speech Language Pathology*, 5(1):5–14.
Matthei, Edward. 1989. Crossing boundaries: More evidence for phonological constraints on early multi-word utterances. *Journal of Child Language*, 16:41–54.
McCarthy, John, and Alan Prince. 1994. The emergence of the unmarked: Optimality in prosodic morphology. In *Proceedings of the North East Linguistic Society 24*, pages 333–379. GLSA, University of Massachusetts.
Morgan, James L. 1996. A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language*, 33:666–689.
Ota, Mitsuhiko. 1999. *Phonological Theory and the Acquisition of Prosodic Structure: Evidence from Child Japanese*. Ph.D. thesis, Georgetown University.
Parks, Robert. 1999. The wordsmyth educational dictionary-thesaurus. http://www.wordsmyth.net/.
Pater, Joe. 1997. Minimal violation and phonological development. *Language Acquisition*, 6(3):201–253.
Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Rutgers University, New Brunswick, and University of Colorado, Boulder.
Saffran, Jen R., Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.
Salidis, Joanna, and Jacqueline S. Johnson. 1997. The production of minimal words: A longitudinal case study of phonological development. *Language Acquisition*, 6:1–36.
Smith, Neil V. 1973. *The Acquisition of Phonology*. Cambridge University Press, Cambridge, UK.
Swingley, Daniel. 1999. Conditional probability and word discovery: A corpus analysis of speech to infants. In *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society*, pages 724–729.