Integration of multiple probabilistic cues in syntax acquisition

Padraic Monaghan and Morten H. Christiansen

1. Introduction

Before the child can understand the relationship between words and referents in the world, the child must know the roles of words within the utterance. But in order to learn the roles of words, the child must know their referents. How does the child begin to solve this circular problem? One solution is that the child can learn to cluster words according to their similarities into groups, thus information about the roles and structures of the language are present within the language itself. In this chapter we focus on corpus analyses of child directed speech that have indicated various sources of information for helping the child to derive syntactic knowledge from the speech signal. We discuss studies of distributional, or contextual, information about the role of words, and studies about the effectiveness of grouping words according to phonological and prosodic properties. We discuss the information that is valuable for forming a sense of the syntactic role of words, and also consider evidence for the availability of these sources of information to the child learning the structure of their first language. Finally, we provide some future challenges for corpus analyses of language acquisition, in particular taking into account developmental trends and the potential for novel computational approaches to assimilate these constructivist processes.

2. The chicken and egg problem of syntax acquisition

To understand or produce spoken language a child must learn how sounds can be combined to form words and how words may be strung together to construct meaningful sentences. By one year of age, infants have already learned a great deal about the sound structure of their native language (for reviews see Jusczyk 1997, 1999; Kuhl 1999; Pallier, Christophe and Mehler 1997; Werker and Tees 1999). In contrast, acquiring knowledge about the grammatical structure of sentences takes several years (for reviews, see O'Grady 1997; Tomasello 1992, 2000b). When acquiring grammatical knowledge, children face a difficult "bootstrapping" problem. Discovering the syntactic

constraints governing the child's native language requires being able to assign individual words to grammatical classes, such as nouns and verbs. Grammatical classes, on the other hand, are only useful for acquisition insofar as they support syntactic constraints. This interdependence of syntactic constraints and grammatical categories presents the child with a seemingly insurmountable bootstrapping problem, apparently requiring simultaneously searching through *combinations* of syntactic constraints and grammatical categories. Yet, children typically acquire grammatical knowledge with accuracy and without apparent effort.

Such a perspective on language acquisition suggests a modular view of grammar, where phonology, syntax, and semantics are seen as separate and functionally-isolated representations. In this chapter, however, we indicate ways in which phonology and syntax are likely to interact in language development which aid us to conceive how the bootstrapping problem may be solved by the child – indeed we argue that the apparent poverty of the stimulus in the child's language environment holds only if one conceives that each representational type is immune to the (helpful) influence of other representational levels. Equally, we support views of the interaction of syntax and semantics that may also assist in this process (Blackburn and Bos 2005; Croft 2001; Lakoff 1987; Pinker 1984; Tomasello 2003) in particular when these are proposed to be generated from exposure to correlations between the linguistic environment and events and objects in the world (e.g., Yu and Smith 2006).

Students learning an academic subject such as physics face a similar "bootstrapping" problem: understanding momentum or force presupposes some understanding of the physical laws in which they figure, yet these laws presuppose the concepts they interrelate. But the bootstrapping problem solved by young children seems vastly more challenging, both because the constraints governing natural language are so intricate, and because young children do not have the intellectual capacity or explicit instruction available to the academic student. Determining how children so readily solve this bootstrapping problem is crucial for understanding language acquisition and, more generally, the relation between biological and environmental factors in development.

3. Solutions to the chicken and egg problem – innate categories don't help

There are three sources of information that children could potentially bring to bear on solving the bootstrapping problem: innate knowledge in the form of linguistic universals, intra-linguistic and extra-linguistic information. The intra-linguistic information is present within the physical speech signal itself, including patterns of phonological and prosodic information within the word and distributional patterns or semantic features that have a morphological realization such as gender or number, that indicate the relation of various parts of language to each other. Extra-linguistic information concerns the observed relationships between language and objects, actions, and relations in the world. In the remainder of this article we use "semantic information" in this restricted sense: to refer to information that is not present within the language signal itself.

Although some kind of innate knowledge may play a role in language acquisition, it cannot solve the bootstrapping problem. Even with built-in abstract knowledge about grammatical categories and syntactic rules (e.g., Pinker 1984), the bootstrapping problem remains formidable: innate knowledge can only help address the bootstrapping problem by building in universal aspects of language, and relationships between words and grammatical categories clearly differ between languages (e.g., the sound /su/ is a noun in French (*sou*) but a verb in English (*sue*)). Crucially, children still have to map the right sound strings onto the right grammatical categories while determining the specific syntactic relations between these categories in their native language. Moreover, there now exists strong experimental evidence that children do not initially use abstract linguistic categories, but instead employ novel words as concrete items, thereby challenging the usefulness of hypothesized innate grammatical categories (Tomasello 2000b). Thus, independently of whether or not innate linguistic knowledge is hypothesized to play an important role in language acquisition, it seems clear that other sources of information nevertheless are necessary to solve the bootstrapping problem.

Extra-linguistic information is likely to contribute substantially to language acquisition. Correlations between environmental observations relating prior semantic categories (e.g., objects and actions) and grammatical categories (e.g., nouns and verbs) may furnish a "semantic bootstrapping" solution (Pinker 1984). However, given that children acquire linguistic distinctions with little or no semantic basis (e.g., gender in French, Karmiloff-Smith 1979), semantics cannot be the only source of information involved in solving the bootstrapping problem. Another extra-linguistic factor is cultural learning, whereby children may imitate the pairing of linguistic forms and their conventional communicative functions (Tomasello, Kruger and Ratner 1993). For example, by observing the idiom *John spilled the beans* used in the appropriate context, the child by reproducing it can discover that it means that John has revealed some sort of secret, and not that he is a messy eater. However, to break down the linguistic forms into relevant units, it appears that cultural learning must be coupled with intra-linguistic learning.

Though not the only source of information involved in language acquisition, we suggest that intra-linguistic information is fundamental to bootstrapping the child into syntax. However, although intra-linguistic input appears to be rich in potential cues to linguistic structure, there is an important caveat: the individual cues are only partially reliable, and none considered alone provides an infallible bootstrap into language. Thus, a learner could use the tendency for English nouns to be longer than verbs to determine that *elephant* is a noun, but the same strategy would fail for *investigate*. Similarly, although speakers tend to pause at linguistically meaningful places in a sentence (e.g., following a phrase or a clause, Cooper and Paccia-Cooper 1980), pauses also occur elsewhere. And although it is a good distributional bet that a determiner

(e.g., *the*) will be followed by a noun, there are other possibilities (e.g., adjectives, such as *big*). To acquire language successfully, it seems that the child needs to integrate a great diversity of multiple probabilistic cues to language structure in an effective way. Fortunately, as we shall see next, there is a growing bulk of evidence showing that multiple probabilistic cues are available in intra-linguistic input, that language learners are sensitive to them, and that learning is facilitated through multiple-cue integration.

In the remainder of this chapter, we provide a review of approaches to multiple cue integration in syntax acquisition. The majority of work has concentrated on intra-linguistic cues, and so the majority of our chapter deals with these topics. We catalogue studies that have explored the richness of potential cues in the child's language environment, we indicate ways in which these sources of information may plausibly give rise to the development of syntactic categories, and we review studies that suggest these cues are actually used by the child in constraining their grammar. Though there is a multitude of potential sources of information available to the child, we limit the search for cues to linguistically-relevant material – infants, for instance, structure acoustic speech input into phoneme categories at an early stage (Kuhl, Williams, Lacerda, Stevens and Lindblom 1992) – though the generic learning mechanisms we describe could in principal be applied by the child to determine the relevance of *any* language property with respect to language structure. The search-space to grammatical structure becomes constrained by the overlap of multiple cues available in speech. We also indicate that intra-linguistic cues are just the tip of the iceberg in terms of the available environmental information to assist in bootstrapping syntax, and we discuss the potential for co-relation of intra- and extra-linguistic cues to assist in bootstrapping the syntax of the child's first language. We conclude with some future directions for studies of multiple cues in language acquisition.

4. Intra-linguistic cues in the utterance: from statistics to structure

The purpose of corpus analysis studies in language acquisition is to assess a representative sample of the child's linguistic environment and measure the information present within this environment that may assist the child in learning their language. There are two approaches for such corpus analyses. First, the aim of the analysis may be to reveal the potential information present within the child's environment. Such approaches typically employ descriptive statistics of the corpus' characteristics. Second, the aim may be to determine how such information can be used to generate or support syntactic categories within the child's language. This latter approach attempts to determine how general purpose mechanisms may give rise to a structuring of the language environment with which the child is presented. Many of the published studies have limited the assessment of grammatical category distinctions to function and content words and nouns and verbs. The former is important for the child to discover, as this enables the child to restrict their attention to link only the content words to

referents in the world. Such a distinction has been shown to be a productive constraint in Cartwright and Brent's (1997) and Dominey, Hoen, and Inui's (2006) models. The noun/verb distinction has been a focus because these categories provide the largest token categories of words, and so provides a lower-bound for potential grammatical category information present in the stimulus.

4.1 Measuring potential information in the corpus

Maratsos and Chalkley (1980) proposed that grammatical categories of words could be predicted with some accuracy from certain "frames" within which those category words occurred. For instance, they suggested that only verbs could occur between a noun phrase and the *-ed* inflection. Similarly, Fries (1952) identified 19 frames within which only words of certain grammatical categories could occur, e.g., *(The)* ____ *is/was/are good.*

Using corpus analyses of small corpora of child directed speech, Mintz (2003) provided an empirical test of the idea that static frames may serve as indicators of grammatical categories. The corpora were selected as reflecting the speech directed to very young children. Furthermore, the study of speech directed toward particular children enabled an estimate of the type of information presented to an individual child. The alternative – to select a corpus of speech directed to several children – may obscure somewhat the precise usages to which a child becomes familiar in early stages of language development (e.g., Tomasello 2000b). The flip side of analyses of speech directed to multiple children, of course, is that it provides an averaged view of children's language environment, and is less prone to local fluctuations in the words that children may be exposed to on a particular recording session. Certain vagaries may be present within small samples of speech directed toward particular children. For instance, in one of the corpora studied in Mintz's (2003) study (anne01a-anne23b, Theakston, Lieven, Pine and Rowland 2001) *penguin* occurs with a frequency of 440 per million. In 5.5 million words of child directed speech from the CHILDES corpus for English (MacWhinney 2000), *penguin* has a frequency of 92 per million (see Rowland, Fletcher and Freudenthal, this volume, for further discussion of this issue).

Within each corpus, the most frequent frames were selected, and the words that occurred within these frames were classified according to their grammatical category. Mintz then measured the extent to which words of the same grammatical category occurred inside the same frames. He found that there was a high degree of accuracy for these groupings within frames – approximately 90% for each child corpus, meaning that words of the same category grouped together within the frame. The weakness of the analysis, however, was that words of the same grammatical category did not always occur in the same frame, reflected by the measure of completeness which was below 10%, meaning that the frames specified subsets of grammatical categories, and also problematic was the fact that the connection between the words occurring in the frames *you__to* and *we___to* (both verb contexts) was not determined by the analysis.

Monaghan and Christiansen (2004) explored the extent to which high accuracy but low completeness may facilitate learning the grammatical categories of the language. We contrasted the frames analysis with a grouping of words based only on the preceding word – so words that occurred after *you* were grouped together. In this latter analysis, we found overall lower accuracy – as verbs, auxiliaries, and adverbs were all likely to occur after *you* – but a much larger proportion of the words of the corpus belonging to a grouping: 69.9% compared to 14.3% in the frames analysis. A connectionist model trained to learn the grammatical categories of words based either on bigrams or frames resulted in much better learning from bigrams, suggesting that more information was present in the bigram analysis which was more conducive to learning the categories.

Pursuing this view that bigram information may be a useful indicator on its own about grammatical category, Monaghan, Chater and Christiansen (2005) assessed the extent to which combined bigrams could give information about the grammatical category of a word. This contrasts with the analyses of Mintz (2003) and Monaghan and Christiansen (2004) where contexts were considered individually for their categorization of words. In the Monaghan *et al.* (2005) analyses, the extent to which overlaps between information from bigrams could be discovered was studied. In these analyses, the strength of association between a set of high-frequency context words and the target word was assessed using a log-likelihood statistic. So, if the target word occurs after the context word more than expected by chance (e.g., the occurrence of *apple* following *the*) then this association is given a positive score, if the target word co-occurs with the context word at chance level then the value is close to zero, and if the co-occurrence is less than chance then the association is negatively scored (e.g., the co-occurrence of *you apple*).

The information contained in the associations for the 1000 most frequent words in English child directed speech was then combined using discriminant analysis. The results indicated that accuracy of classification based on these associations was up to 85.8% for distinguishing nouns from verbs, an improvement over the use of bigram cues singly, as in Monaghan and Christiansen (2004). Such a result indicates that, even within a particular type of information, integrating cues, enabling the connections between *the*___ and *a*___, for instance, increases accuracy of categorization. Similar results pertain for other languages.

Corpora of child directed speech in Dutch, French, and Japanese also indicate that combined bigram information cues provide highly accurate reflections of grammatical category distinctions between function and content words, and also between nouns and verbs (Monaghan, Christiansen and Chater 2007). In this study, the 25 most frequent words were taken from each language corpus and associations with these frequent words either preceding or succeeding each of the other words in the corpus were assessed (the 25 words contained both function and content words). Analyses taking the frequency of words into account resulted in correct classification for function and content words of 91.0% for Dutch, 78.8% for English, 85.4% for French, and 95.2% for Japanese. For nouns and verbs, classifications were also extremely accurate: 93.0% for Dutch, 93.0% for English, 84.1% for French, and 84.4% for Japanese.

However, one limitation of these studies is that the analyses are supervised – the discriminant analyses are provided with the grammatical categories and have to use the given information in the bigram cues to match the categories as closely as possible. Such an approach is extremely useful for indicating the potential information available within the child's language environment, but it does not provide an indication of how the grammatical categories may be generated by the child learning their language. An alternative approach is to combine information using unsupervised analyses, where the model's solution emerges from the structure of the information itself.

4.2 Deriving syntactic structure from the corpus

Redington, Chater and Finch (1998), in a landmark study, showed that not only was there category information present within distributional information in child directed speech, but that a cluster analysis based on this information resulted in clusters that respected the grammatical categories of the words extremely well. So, nouns tended to occur in very similar contexts to one another, as did verbs, and adjectives, and so on. The distributional information utilized in their model was based on the co-occurrence of each word with each of the 150 most frequent words in the corpus of child directed speech in one of four positions: two words before the target word, one word before the target word, one word after the target, or two words after the target. The counts were added throughout the corpus, resulting in a 600-dimensional vector for each word. Similarities between the contexts of words could then be computed by measuring the distances between the context vector for each word. The clustering procedure grouped together words that were closest in this 600-dimensional vector space.

The accuracy and completeness of this approach were spectacularly effective, resulting in high accuracy and completeness for words within the syntactic categories. Also, the clustering resulted in a structuring of words into grammatical categories with nested degrees of specificity. So, for example, adverbs were grouped at one level of the clustering tree as they were all used in similar distributional contexts, but this grouping was comprised of subordinate clusters of adverbs distinguished by more specific usage patterns. The presence of graded similarity within the clusters corresponds to constructionist grammar, whereby grammatical categories are based on patterns of usage, of varying degrees of specificity (Croft 2001, 2003). However, the plausibility of this analysis of the distributional information as a reflection of cognitive processing within the developing child is a matter of debate. One problem is that the amount of information to be tallied for each word in the child's environment is vast: maintaining counts of 600 co-occurrences for every word presumably exceeds the child's working memory. In addition, the clustering analysis performed on the assessment required distances between all words to be considered simultaneously, and there are also difficulties in deciding how many clusters there should be in the final clustering. There are two adaptations to the account that address each of these objections, without altering the principles of the approach.

First, to address the cognitive overload objection, a developmental approach can be taken for deriving the contextual co-occurrence counts, with just a few context words employed for the co-occurrence counts at early stages of language development, and counts only recorded for a small number of words within the lexicon. Then, as the child's language knowledge develops, additional context words could be iteratively introduced to refine the contextual information for words in the child's vocabulary. Furthermore, retaining all the co-occurrences between words is not necessary to determine the words' context, and may indeed be an impediment to learning distinctions between categories. Storing counts only for context words that provide variation in their co-occurrence pattern is one such way to limit storage requirements and maximize category information. For example, interjections may co-occur with all open class words (nouns, verbs, adjectives, adverbs) and as such they do not provide discrimination between these grammatical categories, and blur the distinctions that the child is required to make. In contrast, articles tend to co-occur frequently with nouns and seldom with verbs, and so such context words provide a great deal of information about grammatical categories.

To address the second objection, alternative approaches to the clustering algorithm can be used that exploit the information contained in the similarity space for a set of words Pothos and Chater's (2002, 2005) simplicity model of clustering, for instance, determines both the number of clusters *and* which items group together within those clusters by computing the optimal solution in information-theoretic terms for the similarity space. Such a clustering approach respects intuitions and experimental results in terms of the characteristics of generating categories of perceptual stimuli. The simplicity approach exploits the fact that if two similar items can be grouped together then it requires less information to store and access these two items than if they are stored separately, due to redundancies between the properties of the items (in the case of corpus analyses, similarities between the contexts in which the two words occur).

An alternative unsupervised approach to learning syntactic categories was developed by Cartwright and Brent (1997). Their model searched child directed speech for minimal pairs, where two phrases differed by only one word. When two phrases like this occurred, the differing words were grouped together, and the phrase omitting the differing words was stored as a pattern (see also Dominey *et al.* 2006). So, if the corpus comes across *the dog sat down* and *the cat sat down*, then *dog* and *cat* would be grouped together, and the frame *the* ___ *sat down* would be extracted. Similarly, if *the* ___ *sat up* also occurred with *cat* or *dog* in the gap, the frame would become *the* ___ *sat* ____, and *down* and *up* would be clustered together. The model learned to categorize words from the child directed speech corpus. In addition, the model learned the categories more accurately when semantic information about concrete nouns was also included in the simulation (see Solan, Horn, Ruppin and Edelman 2005 for a recent update on this type of approach as well as the section below on combining intra- and extra-linguistic information).

5. Intra-linguistic cues in the word: Phonology to structure

Table 1. Phonological and prosodic cues found to distinguish grammatical categories in English

Cue	Description	Grammatical category distinctions	References
Phoneme length	How long is the word in terms of phonemes?	Function < Content Noun > Verb	(Kelly 1992; Morgan, Shi and Allopenna 1996)
Syllable length	How long is the word in terms of syllables?	Function < Content Noun > Verb	(Cassidy and Kelly 1991; Kelly 1992; Morgan *et al.* 1996)
Presence of stress	Does the word receive lexical stress?	Function < Content	(Gleitman and Wanner 1982)
Position of stress	Which syllable receives lexical stress?	Noun earlier than Verb	(Kelly and Bock 1988)
Onset complexity	How many consonants in the word's onset?	Function < Content	(Shi, Morgan and Allopenna 1998)
Word complexity	How many consonants per syllable?	Function < Content	(Morgan *et al.* 1996)
Reduced syllables	What proportion of syllables contain schwa or syllabic consonant?	Function > Content	(Cutler 1993)
Reduced first syllable	Does the first syllable contain a schwa?	Function > Content	(Cutler 1993; Cutler and Carter 1987)
-ed inflection	Does the word end in /əd/ or /Id/?	Adjective > other categories	(Marchand 1969)
Coronals	What proportion of consonants are coronals?	Function > Content	(Morgan *et al.* 1996)
Initial /ð/	Does the word begin with /ð/?	Function > Content	(Campbell and Besner 1981)
Final voicing	Does the word finish with a voiced consonant?	Noun > Verb	(Kelly 1992)

Cue	Description	Grammatical category distinctions	References
Nasals	What proportion of consonants are nasals?	Noun > Verb	(Kelly 1992)
Stressed vowel position	Is the stressed vowel more likely to be a front vowel?	Noun < Verb	(Sereno and Jongman 1990)
Vowel position	Are the vowels more likely to be front vowels throughout the word?	Noun < Verb	(Monaghan *et al.* 2005)
Vowel height	Are the vowels more likely to be high throughout the word?	Noun < Verb	(Monaghan *et al.* 2005)
Plosives	Are plosives more likely to occur in the word?	Function < Content	(Monaghan *et al.* 2007)
Fricatives	Are fricatives more likely to occur in the word?	Function > Content	(Monaghan *et al.* 2007)
Dentals	Are dental consonants more likely to occur in the word?	Function > Content	(Monaghan *et al.* 2007)
Velars	Are velar consonants more likely to occur in the word?	Function < Content Noun < Verb	(Monaghan *et al.* 2007)
Bilabials in onset	Are bilabials more likely to occur in the word onset?	Function < Content Noun > Verb	(Monaghan *et al.* 2007)
Approximants in onset	Are approximants more likely to occur in the word onset?	Noun < Verb	(Monaghan *et al.* 2007)

Besides the potential utterance-level information indicated by distributional information in language, a similar cornucopia of cues in the speech sounds of individual words has been found to relate to different grammatical categories. Many different phonological and prosodic cues have been reported in the literature for reflecting grammatical category distinctions in English, and 22 of them are reported in Table 1, containing

16 cues reported in Monaghan *et al.* (2005), extended by 6 additional cues found to be significantly different in a study by Monaghan *et al.* (2007). Establishing that these phonological and prosodic cues are significantly distinct for grammatical categories has generally been accomplished by assessment of large corpora or subsets of the lexicon of English. However, each cue considered alone does not provide very reliable information about grammatical categories. Monaghan *et al.* (2005), for instance, demonstrated that, using the length in syllables cue, classifying all words of length two syllables or greater as nouns, and words of length one syllable or less as verbs, resulted in 54.5% correct classification of nouns and verbs from the 5000 most frequent words in English child directed speech. Though highly significant ($p<.001$), such a classification is unlikely to be useful as a sole basis for determining grammatical categories. The key question, then, is whether combined cues provide useful information for grammatical categorization, and if so, how this combination may be achieved.

5.1 Individual cues in categorization

Assessment of the use of phonological cues in language learning has also tended to be in isolation, so their combined usage is unclear. For example, Cassidy and Kelly (2001) tested children's classification of nonsense words as either nouns or verbs, based only on the number of syllables that each had. In English, as noted in Table 1, nouns tend to be longer than verbs, and children were much more likely to assign trisyllabic nonsense words to roles as nouns, whereas monosyllabic words were more likely to be assigned verb roles in describing pictorial scenes. Whereas this experiment was designed to test the single cue of syllable length, interestingly there were a number of other cues that also contributed to the noun/verb distinction for these materials. The trisyllabic words differed significantly from the monosyllabic words in terms of length in phonemes, stress position (trisyllabic words had less word-initial stress), and proportion of nasal consonants (trisyllabic words had more nasals). Monosyllabic words were also marginally significantly more likely to contain complex syllables. Each of these cues may have been integrated in affecting the child's performance.

Acoustic-level cues have also been proposed as beneficial for segregating grammatical categories. Blanc, Dodane and Dominey (2003) investigated whether the function/content word distinction could be reflected by variations in F0 peaks. They found that the F0 signal alone was sufficient for classifying over 73.1% of a small corpus of function and content words in French and 64.5% in English, with performance greater than chance levels (52% and 54% respectively). Content words were found to have higher peaks and greater variety in pitch in both languages.

5.2 Combined cues for categorization

Shi *et al.* (1998) provided an investigation of multiple cues combining to aid classification. They assessed acoustic, phonological and prosodic cues, together with distributional cues about utterance position (sentence initial, medial, or final) and frequency of occurrence in two small corpora of child directed speech in two languages – Mandarin and Turkish. In Mandarin, seven of ten cues assessed were significantly differently distributed between function and content words for both corpora. These included frequency, syllable coda (more likely to be present for content words), reduplication (repetition of a syllable), syllable nucleus (diphthongs more likely in content words), marked tone (where tonal sequence does not alternate between high and low, which is more likely in content words), and longer and louder syllables in content words. Combining these analyses into a self-organising computational model indicated that content and function words could be correctly classified to a level of 71.5%-69.9% for each mother-child corpus.

For Turkish, 8 cues were assessed, including the distributional cues of frequency and utterance position, three phonological cues, and three acoustic cues. For both corpora, 7 cues were significantly differently distributed for content and function words. Function words had higher frequency and were more likely to occur utterance medially or finally than content words. Function words tended to have fewer syllables in the morpheme, and content words were more likely to have syllable codas. Function words were more likely to exhibit vowel harmony (first vowel influences later vowels in the word in terms of vowel position, height and roundedness), and syllables were longer and louder in content words. A self-organising model based on these cues correctly classified 69.1%–63.7% of the words occurring in the corpus.

Durieux and Gillis (2001) also attempted to assess combined phonological cues for categorising words in English and Dutch in a model of instance-based learning. They employed four cues from Kelly (1996) suggested to be useful for classifying nouns and verbs: position of stress, word complexity, nasals, and vowel height. They found that, individually, cues correctly classified between 62.42% and 67.38% of words from the English CELEX database into a broad set of categories, and combined cues performed slightly better, with 67.68% correctly classified. In the same study, Durieux and Gillis (2001) also tested the extent to which these cues generalized to Dutch. They found that the individual cues classified between 57.9% and 66.8% of Dutch words from CELEX, with stress position being the poorest, and word complexity being the best. Combined cues produced much better accuracy than any cue alone, classifying correctly 75.2% of the words. The same set of cues was also found to be useful for forming further individuations among the content word categories in English and Dutch, correctly classifying 66.6% of English nouns, verbs, adjectives, and adverbs, and 71.0% of these categories in Dutch.

In an assessment of Dutch, English, French, and Japanese, Monaghan *et al.* (2007) used discriminant analyses to find the extent to which distributions of phonological properties, principally in terms of manner and place features of consonants within the words, distinguished function from content words, and nouns from verbs. They assessed child directed speech in each of the four languages and used combined cues to ascertain the extent to which these cues could potentially operate in concert for categorization. For token-based analyses, where the frequency of lexical items was taken into account in the classifications, they found extremely high accuracy for classification. For function and content words, correct classification reached 82.9% for Dutch, 68.7% for English, 85.2% for French, and 93.4% for Japanese. For nouns and verbs, the results were also extremely high, with 89.6% of Dutch words correctly classified, 67.5% for English, 82.0% for French, and 82.2% for Japanese. Combined cues, then, for these four languages, resulted in very impressive classifications, all well above chance levels ($p<.0001$). Similarly, Onnis and Christiansen (2005) found that the beginning and ending phonemes of words in the same four languages provide sufficient information for reliably distinguishing between nouns and verbs. The discrimination of words based on the intra-linguistic cues we have reported results in some words being misclassified according to the benchmark categories. However, achieving accuracy approaching 90% is a persuasive demonstration that there is a high degree of useful information available within these cue types, that can be employed to constrain grammatical categories in the child's developing language and be utilized to begin the process of bootstrapping categories. Yet, the benchmark categories ignore the graded nature of grammatical categories (e.g., Croft 2003; Labov 1973), and a more graded system of grammatical categories may well result in greater accuracy of classification based on these cues.

6. Combining intra-linguistic cues

The intra-linguistic cues we have discussed in the previous two sections – both at the utterance and the word level – have been revealed by corpus analysis studies to provide a large amount of information about a word's grammatical category. Combined distributional cues provide greater reliability and accuracy of information about categories, and the same applies for the combined analyses of phonological and prosodic cues. Furthermore, these patterns of benefit have been shown to be language-general, extending across Germanic, Romance, and Japonic language families in Durieux and Gillis (2001) and Monaghan *et al.*'s (2007) studies, and Sino-Tibetan and Turkic language families in Shi *et al.*'s (1998) study.

Considered separately, each set of cues can provide impressive classification, demonstrating that the language environment is extremely rich in supporting the development of grammatical categories in the child learning her first language. But what benefit would accrue from connecting utterance-level and word-level intra-linguistic information for classification? Shi *et al.* (1998) have already pointed to the potential accumulative benefits of integrating distributional, phonological, and acoustic cues to

reflect the function/content word distinction. However, our multiple cue integration analyses have revealed that not only does accuracy increase based on converging cues, but that cues from different modalities interact in surprising ways for categorization.

Monaghan *et al.* (2005) used discriminant analyses of combined bigram distributional cues and phonological cues for categorising function and content words and nouns and verbs in English. They found that combined cues provided more accurate classification than just distributional or phonological cues alone. However, the relative benefit of these types of cues varied for different frequency groupings. For higher-frequency words, distributional cues were more reliable than phonological cues, whereas for lower-frequency words, the effectiveness of phonological cues overtook the contribution of distributional cues. For these low-frequency words, there were few occurrences of them in the child's environment, and so the certainty with which the language learner can form associations between words is limited. Fortunately, under these circumstances basing grammatical category judgement solely on the phonological cues is, in terms of the information present in the language, an effective approach (see also Durieux and Gillis 2001).

Christiansen and Monaghan (2006) focused on the classification of nouns and verbs and the contribution of distributional and phonological information for classifying words of each category. They found that distributional cues tended to be more reliable for classifying verbs than nouns, whereas verbs were more accurately classified by phonological cues. Verbs tend to occur in a wider variety of contexts (McDonald and Shillcock 2001), and so associations formed for verbs with the set of high-frequency context words is less likely to be reliable, due to being less constrained. Nouns, in contrast, tend to occur in more prescribed contexts. Yet, once again, serendipitously for the language learner, phonological information is more effective for classifying verbs than nouns and so can compensate for the lower accuracy in distributional information.

In a set of crosslinguistic analyses, Monaghan *et al.* (2007) found that the pattern of effects reported in Christiansen and Monaghan (2006) for the interaction of different modalities of cue with the noun/verb distinction were also found for Dutch, French, and Japanese. In addition, a similar interaction of reliability of cue types was found for categorising function and content words. Classification of function words was less accurate using distributional cues – despite their higher mean frequency, they tended not to co-occur reliably with other high-frequency words as they tended to occur immediately before or after content words, and could occur in a wide variety of contexts. Yet again, phonological information was more reliable for these words. Morgan *et al.* (1996) discuss the pressures on function words, due to their high-frequency and predictability, to be produced with minimal effort in speech, hence vowel-reductions and consonant-voicing are common in this category of words. A consequence of this is that such function words may be easily identified as such given minimal, but coherent phonological information in the speech signal (see also Shi, Werker and Morgan 1999).

7. Converging evidence for the use of multiple cues

The corpus analyses we have thus far discussed have indicated the potential benefits from drawing together multiple cues in order to determine the extent to which grammatical categories may be learned from the environment during language acquisition. However, experimental approaches provide converging evidence for the *use* of these cues in acquiring language structure. There are two paradigms in the literature for such experimental investigation of cues for language acquisition. The first we discuss are artificial language learning (ALL) studies of segmentation, the second are word categorization studies.

7.1 Learning to segment artificial language with multiple cues

ALL tasks involve training human participants on artificial miniature languages with particular structural constraints, and then testing their knowledge of the language. Importantly, the ability to acquire linguistic structure can be studied independently of semantic influences. Because ALL permits researchers to investigate the language learning abilities of infants and children in a highly controlled environment, the paradigm is becoming increasingly popular as a method for studying language acquisition (for reviews, see e.g., Gómez and Gerken 2000; Saffran 2003). We suggest that ALL can be applied to the investigation of issues pertaining to the usefulness of cues in language acquisition in much the same way as computational modelling is currently being used. One advantage of ALL over computational modelling is that it is possible to show that hypothesized cues actually *affect* human learning and processing and are not only potentially useful for language acquisition.

The role of distributional cues in ALL has been extensively investigated, both in terms of the extent to which this information can be exploited to learn the structure in terms of word boundaries within speech, and also the relationships between words within sentences. Saffran, Aslin, and Newport (1996) in an influential study on language acquisition in infants devised a nonsense language composed of words of three syllables each. Words were concatenated into continuous speech, such that transitions for syllables between words were of low probability and transitions for syllables within words were of high probability. Infants were found to be sensitive to such distributional information in determining their preference for sequences of syllables that respected the word structure over those that crossed word boundaries.

This approach to cues available to language learners for learning the structure of words within the language has been extended to test the benefit of including cues from other modalities for revealing language structure. In English, final syllables of words tend to be slightly longer than initial or medial word syllables in articulation. Adding this cue into an ALL segmentation task resulted in better performance for adult participants than when the first syllable was lengthened (Saffran *et al.* 1996). In English, also, words tend to be stressed on the first syllable and infants can use this to segment

artificial languages even in the absence of distributional information (Curtin, Mintz and Christiansen 2005). Moreover, Johnson and Jusczyk (2001) found that the distributional information was more accurately discovered when the first syllable was stressed, and that performance appeared to be worse than chance when the final syllable was stressed. These results indicate that multiple cues are beneficial for aiding speech segmentation, but only when the cues are consistent with the patterns found in the hearer's language environment. In the case of stress, the misplacement appeared to over-ride the use of distributional information in language processing.

ALL studies have also been used to reveal the extent to which more complex distributional information can be used for learning language structure. Peña, Bonatti, Nespor and Mehler (2002) devised a language composed of trisyllabic words, where the transitional probabilities between all adjacent syllables in the speech were identical, but where syllables within the same word (*Ai* and *Bi*), but separated by an intervening syllable (*X*), had a dependency of 1. Hence, the language was of the form *AiX-BiAjXBj*... where *Ai* and *Bi* always co-occurred in the speech. Peña *et al.* found that participants could use these non-adjacencies to discover word structure, demonstrating a preference for words over sequences composed of part-words from the speech. However, Onnis, Monaghan, Richmond and Chater (2005) demonstrated that such structure could only be discovered when it was supported by similarities in terms of the phonemes from which the non-adjacent syllables were composed. In Peña *et al.*'s (2002) studies, the non-adjacent syllables were stop consonants, whereas the intervening syllables were always continuants. If this phonology was removed then learning was not observed, but learning did take place if the non-adjacent dependent syllables were both continuants and the intervening syllable was a stop. Such a view of multiple cues conspiring to support learning is consistent with the pattern of effects found in Newport and Aslin (2004) for supporting learning non-adjacencies – learning was only found to occur when there was phonological similarity between distributionally dependent syllables.

The innovation in Peña *et al.*'s (2002) study, however, was that not only could discovery of the structure of the words in the language occur, but also that the ability of participants to generalize the structure of the language could be measured. The language was identical to before, but participants were this time tested on their preference for rule-words compared to part-words. Rule-words were composed of *Ai* and *Bi* pairs with another *Aj* or *Bj* syllable intervening between them, so they respected the non-adjacent structure but had not been encountered during training. Participants were found to prefer the rule-words over the part-words only when an additional cue of a short gap was placed before the first syllable in each word during training on the speech. Such a cue was sufficient to enable realization of the generalizable structure of the language to participants.

7.2 Learning to categorize artificial language with multiple cues

Multiple, converging cues have also been used for investigating learning categories of words in ALL studies. Valian and Coulson (1988) investigated the extent to which two categories of words could be learned from simple sentences containing reliable bigram information about category membership. Sentences were of the form *aAibBj* or *bBiaAj* where *a* and *b* were invariant marker words and *Ai* and *Bj* were category words, drawn from a set of 8 words each. Valian and Coulson (1988) found that adult participants could learn to accept valid sentences that did not occur during training on the language and correctly reject sentences where there was a mismatch between the marker words and the category words (e.g., *aBibAj*). Monaghan *et al.* (2005) extended this experimental paradigm to incorporate phonological information within the *Ai* and *Bi* category words. Words in each category had distinct phonological properties according to those found to be useful for distinguishing different grammatical categories in English. Monaghan *et al.* (2005) found that when the category words were supported by phonological information then learning of the categories was much more accurate than when phonological information was not present.

Learning the categorical distinction between different genders has been of particular interest in language learning studies. This is because extra-linguistic information provides few cues about how many genders a language may have and about gender membership for particular lexical items. However, Corbett (1991) noted that there are various potential sources of information to grammatical gender. Certain semantic groupings may have the same gender, for instance in German superordinate terms are neuter (Köpcke 1982). In addition, there is correspondence between phonological and morphological cues and gender. For example, in French the endings *-age* and *-ble* tend to indicate masculine nouns. However, there remains the difficulty of how these gender categories are developed, and the question of how and when the correspondence between particular semantic groups and gender may be exploited by the learner.

In learning to categorize words, Braine, Brody, Brooks, Sudhalter, Ross, Catalano and Fisch (1990) found that categories could be learned in their study only when there was a consistent morphological marker that applied only to that category. However, Brooks, Braine, Catalano, Brody and Sudhalter (1993), in a related study, trained participants on word categories some of which had a consistent morphological marker. They found that the category of words that were marked with the consistent morphology was judged accurately, but also performance was good on words that were of the same category but did not have the morphological marker. Generalization to phrases that had not occurred in the training but were consistent with the structure was also seen to occur, but only under the conditions of consistent marking of the categories (see also Frigo and McDonald 1998).

8. How are multiple cues integrated?

In sum, the ALL experiments indicate that multiple, multimodal cues to language structure assist language learners in developing a sense of their language, both to be able to respond to what is an accurate form in the language and also in order to generalize from this structure. Taken together with the corpus analysis studies, there are two possibilities for the benefit of multiple cues for learning. One possibility is that, as Braine (1987) contends, phonological cues that are consistent with language structure are necessary in order for that structure to be learned, in other words without the phonological correspondence the structure is unlearnable. Hence, the distributional information is emphasized by its reflection in differences in the phonology of words belonging to different categories (whether this information is in the word root or in morphosyntactic marking). Under such a view, the learner benefits from the alignment of structures in these different modalities. Monaghan *et al.* (2007) discuss a possible mechanism for this alignment, based on the Redington *et al.* (1998) corpus analyses. Redington *et al.* (1998) suggested that words are clustered according to their similarity in terms of the contexts in which they occur – words that have the same distribution in the language that the child is exposed to will be grouped together. A similar clustering could take place based on the phonological properties of words. Indeed, Farmer, Christiansen and Monaghan (2006) illustrate the fact that nouns cluster more closely to other nouns than they do to verbs in terms of their phonological properties, and similarly verbs cluster with other verbs, and furthermore the closeness of the word to the centre of its cluster in terms of phonology has an influence on accuracy and speed of its processing in sentence contexts. The two clusters – one based on distributional information, and the other based on phonological information – can then be combined to provide a more refined clustering that more accurately reflects the objective grammatical categories in the language.

The alternative is that the importance of multiple cues is due to their redundancy. Learning a language where there is only just enough information to be able to transmit the structure would be extremely difficult, and inattention to a crucial feature of the language to reveal its structure would be catastrophic. It may be that multiple cues provide a safety-net for learning. This latter perspective is consistent with the use of multiple cues in modalities other than language. In determining depth perception, for instance, viewers are sensitive to a multitude of cues (Cutting and Vishton 1995), however each cue alone is unreliable and consistent with a range of depths under different viewing conditions (for a review see Jacobs 2002). In each viewing situation, the viewer must determine which cues are valid for the current environment. For instance, distant objects tend to appear bluer, but in some scenes blue-coloured objects may be closer than yellow objects, say. The literature on multiple-cue use in depth perception has principally attended to the issue of how the viewer decides on the reliability of particular cues in the environment. There are two views for how this may proceed. Either reliability is determined according to the ambiguity of the cue, or cues that

correlate well with other cues are processed as more reliable. Once the reliability of cues has been determined, then viewers tend to place more weight in their decision about depth from cues with greater reliability. Given the quirks and variations of the visual environment, selecting from a large set of cues is doubtless a useful process in maintaining accuracy of judgement. The same reasoning applies to the noisy language environment through which the child navigates, and Christiansen and Dale (2001) provided an indication for how this may apply in a simulation of multiple-cue integration in language learning. They found that when the model was presented with four syntactically relevant cues and four distractor cues, the model learned to take advantage of the informative cues while ignoring the irrelevant cues.

Our view is that the advantage of multiple cues is some combination of the benefits of both the redundancy of overlapping information and the safety-net of being able to draw on several cues for the categorization decision. The corpus analyses, for instance, indicate that for some words (high frequency, content words, nouns) distributional information is extremely useful, whereas for other word types (lower frequency, function words, verbs) phonological information is more reliable for determining grammatical categories. Hence, reliability of different cue types varies, and placing more weight on the more reliable distributional cues for these words is likely to be a useful procedure. Yet, phonological information is still present for high-frequency words, content words, and nouns, and so this weaker but potentially useful information is still contributing toward the child's ability to learn these grammatical categories.

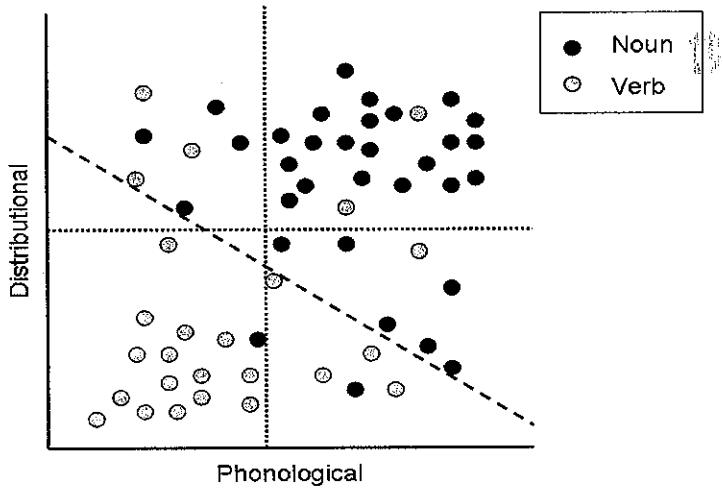

Figure 1. Classifications of nouns and verbs based on distributional cues alone (horizontal dotted line), phonological cues alone (vertical dotted line), and combined cues (oblique dashed line)

Figure 1 illustrates an idealized view of the benefit of combined cues. Words can be conceived as points in space defined by their distributional features and phonological/prosodic properties, and in the Figure we represent idealized nouns and verbs. Phonological properties are illustrated by the x-axis, where words that are close together share phonological cues. Similarly, distributional cues are illustrated by the y-axis. Using just distributional information to classify nouns and verbs results in a distinction around the horizontal dotted line – points above this line are classified as nouns according to their distributional properties, and points below are classified as verbs. A classification based only on phonological cues is shown by the vertical dotted line. In this case, points to the right of the line have phonological cues in line with the majority of nouns, and points to the left are more verb-like. In each of these single-modality classifications, several words are misclassified. However, when both cue types are combined, a line can be drawn in two-dimensional space, illustrated by the oblique dashed line, resulting in greater accuracy of classification. In this case, nouns are above right of the line, verbs below left of the line, and there are many fewer misclassifications than using a single modality of cue alone.

9. Extra-linguistic cues and language learning

In this chapter, so far, we have focused our discussion on the integration of phonological, prosodic (such as stress presence and position), and distributional cues as potential contributors to grammatical categorization in the language learner. Yet, there is certainly a vast array of other potential cues available in the child's environment.

One such source of information is contextual information from the child's environment. Language is not spoken in a vacuum, as it is in certain laboratory experimental conditions, and the surroundings – objects and actions – in the child's environment and their parallels in the spoken language may contribute enormously to language structure. Such information is likely to be extremely valuable in constraining language learning, though empirical investigations of this correlation of environment and language have not been explored until only recently. Yu (2006) investigated corpora of parents narrating a picture book with children aged between 1 and 3 years. The study correlated the words spoken with the objects present on that page of the book. The corpus analysis revealed that there was a great deal of ambiguity for the referents of words spoken at a particular page in the picture book, with a mean of more than 8 words and 3 objects for each learning situation (situations were individuated by speech pauses).

Given this input, a computational model was trained to learn grammatical categories of words, based on a system similar to that of Cartwright and Brent (1997), where words were extracted from common sentence patterns. This "syntactic" system learned in collaboration with a "semantic" system, which learned the associations between particular words and objects in the environment. Words that occurred in similar sentence contexts *and* had strong associations with objects in the visual world provided a high

degree of confidence that those words belonged to the same category (though "semantic" categories were given to the model in this case). This assisted both nouns by accentuating nodes representing nouns, as well as other grammatical categories which were boosted in activation when no objects were present in the environment. A computational model that had both syntactic and semantic learning performed better than a semantic learning model alone in learning the associations between words and objects.

In related work, Yu and Smith (2006) provided a laboratory-based paradigm for assessing learning of the co-occurrences of words and objects in the environment. Undergraduate participants were exposed to a set of words spoken in the presence of a set of pictures of uncommon objects. The number of words spoken and the number of different objects occurring within each picture was altered between conditions, varying between 2 and 4. The participants' task was to learn what the names of objects were in the language, and this was only possible if the associations between particular objects and names were learned across learning situations. In all cases, participants learned the names of objects better than chance, indicating that the correlations with environmental objects were available to learners of the language.

In computational modelling of learning mappings between words and meanings, Monaghan and Christiansen (2006) explored the potential benefit of contextual information that limits the possible assignations of words to objects in the environment. We found that, when contextual information was present, mappings between phonological and meaning forms for words was accomplished effectively. Furthermore, when phonological similarity within a syntactic category was available then words of the same category were more easily grouped together.

These studies point toward future research in analysing multiple cues that assist in language learning. Yu and Smith's (2006) research indicates that visual contextual information can be effectively combined with syntactic information in human language learning, and Monaghan and Christiansen (2006) show that there are potential benefits from integrating distributional and phonological information. There is clearly potency in studies that combine all three information sources (see e.g., Tomasello 2003 for a similar view of the future of multiple cue integration). In the case of each of these new sources of cues for grammatical category, an additional dimension can be added to Figure 1 to enable greater accuracy of classification. The more dimensions there are in designing a plane through the space, the greater the accuracy of distinguishing different grammatical categories.

10. Future directions for multiple cue research

The past decade has seen a growing bulk of corpus-based analyses that provide support for the multiple-cue integration perspective on language acquisition. Many different types of cues have been found to be potentially helpful for learning about syntactic structure, often attested to across a variety of different languages. However, several

challenges remain for multiple-cue integration research, including finding ways of quantifying new cues and incorporating cues to phrasal structure, integrating and utilising cues across different levels of linguistic analyses (such as speech segmentation and lexical category discovery, and relationships between discourse and syntax, e.g., Allen, Skarabela and Hughes this volume), and the development of more comprehensive models to explain when and how children use various cues across development. In this final section, we discuss some of these outstanding challenges for future research in multiple-cue integration.

10.1 Quantifying new cues

We noted earlier that recent work had begun to look at how extra-linguistic cues might be integrated with intra-linguistic information (e.g., Monaghan and Christiansen 2006; Yu 2006). Although this work has yielded promising results, integration of extra-linguistic cues is complicated by the difficulties involved in describing and quantifying this type of information. For example, extra-linguistic cues not only include the visual environment in which the language learner is situated but also the social context within which the interactions take place as well as the internal mental states of the learner. Each of these information sources are very hard to capture adequately in a way that facilitates the kind of computational analyses that now have become typical of corpus-based research. Nonetheless, progress is under way in the context of the CHILDES database, which now includes digital video that can be linked to corpus transcripts (see Behrens this volume; MacWhinney this volume for further discussion). However, more technical innovations are likely to be needed before most extra-linguistic cues can be incorporated into multiple-cue integration research at the same level and amount of detail as is currently the case for phonological and distributional information.

One type of cue that may be more readily amenable to corpus-based analyses is information about sentential prosody. Whereas most multiple-cue integration research so far has focused on cues relevant for lexical category discovery, intonational sentence prosody provide potential cues for phrasal structure (for reviews, see Gerken (1996); Gleitman and Wanner (1982); Jusczyk and Kemler-Nelson (1996); Morgan (1996), though see Fernald and McRoberts (1996) for cautionary remarks). Infants seem highly sensitive to language-specific prosodic patterns (Gerken, Jusczyk and Mandel 1994; Kemler-Nelson, Hirsh-Pasek, Jusczyk and Wright Cassidy 1989) – a sensitivity that may start *in utero* (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini and Amiel-Tison 1988). Prosodic information also improves sentence comprehension in two-year-olds (Shady and Gerken 1999). Results from artificial language learning experiments with adults furthermore show that prosodic marking of syntactic phrase boundaries facilitates learning (Morgan, Meier and Newport 1987; Valian and Levitt 1996). Yet, few corpus-based efforts have tried to quantify just how useful such prosodic information may be. One notable exception is the acoustic analyses by Fisher and Tokura (1996), suggesting that differences in pause length, vowel duration, and pitch provide probabilistic cues to phrase boundaries in both English and Japanese child directed speech. Although this study was on a relatively small scale due to the labor intensiveness of acoustic analyses, progress in automated acoustic analyses may provide a way in which large-scale future quantitative studies can be carried out across different languages.

10.2 Cues for different levels of language learning

Another important challenge for future work on multiple-cue integration is when and how cues might be utilized to facilitate learning at different levels of linguistic analyses. For example, as discussed above, phonology provides useful cues for distinguishing between words from different lexical categories, such as nouns and verbs. However, phonological information is also crucial for discovering words in fluent speech. Christiansen, Hockema and Onnis (2006) conducted a two-stage analysis of a large corpus of child-directed speech to determine whether information about phoneme distributions could be used first to segment speech and then as a cue to lexical categories. They found that the distribution of biphones essentially is bimodal with the phonemes either being inside a word or straddling a word boundary (see also Hockema 2006). Using these biphone distributions more than 70% of the corpus could be correctly segmented. When initial and final phonemes then were used to distinguish nouns and verbs from other words from the segmented corpus, 62% of the words were correctly classified. This indicates that the same type of information may function as a probabilistic cue at different levels of linguistic analyses. Developing a comprehensive model that allows for such cue use and integration across various levels of linguistic representation is an important nontrivial challenge for future multiple-cue integration research.

Given that different languages employ different constellations of cues to signal different syntactic distinctions, an important question for further research is exactly how children (or rather, their learning mechanisms) determine which cues are relevant for which aspect of syntax and which are just noise. This problem is even further compounded by the fact that the same cue may work in different directions across different languages. A case in point is that nouns tend to contain more vowels and fewer consonants than verbs in English, whereas nouns and verbs in French show the opposite pattern (Monaghan *et al.* 2007). So how can the child learn which cues are relevant and in which direction?

One method may be that the child exploits the correlations between cues in the environment, as discussed above. This view is further underscored by mathematical analyses couched in terms of the Vapnik-Chervonenkis (VC) dimension (Abu-Mostafa 1993), showing that the integration of multiple sources of correlated information is likely to reduce the number of hypotheses a learning system has to entertain. The VC dimension establishes an upper bound for the number of examples needed by a learning process that starts with a set of hypotheses about the task solution. Cue information may lead to a reduction in the VC dimension by weeding out bad hypotheses and thereby lowering the number of examples needed to learn the solution. In other words,

the integration of multiple cues may reduce learning time by reducing the number of steps necessary to find an appropriate implementation of the target function as well as decrease the number of candidate functions for the target function being learned, thus potentially ensuring better generalization.

10.3 Computational and developmental approaches to multiple cues

More generally, the development of computational multiple-cue integration models is still in its infancy. By now there are many studies indicating the usefulness of a variety of different cues for language acquisition, and although theoretical models abound (e.g., Gleitman and Wanner (1982); and contributions in Morgan and Demuth (1996); Weissenborn and Höhle (2001)), only a few psychologically plausible computational models for multiple-cue integration currently exists (e.g., Cartwright and Brent 1997; Christiansen and Dale 2001; Reali, Christiansen and Monaghan 2003). The existing models, however, tend to model the end-state of learning rather the developmental process itself. This ignores the different time-scales at which different cues may become important for acquisition. For example, the ability to use visually-based contextual information to constrain the interpretation of a syntactically ambiguous sentence does not appear until about eight years of age, considerably later than sensitivity to constraints on the possible kinds of constructions that may follow specific verbs (Snedeker and Trueswell 2004). To fully understand multiple-cue integration and how it develops, models must be devised that capture the developmental trajectory of cue use across different stages of language acquisition. We anticipate that the availability of so-called "dense" corpora (e.g., Behrens 2006; Maslen, Theakston, Lieven and Tomasello 2004) will facilitate the development of this kind of more constructivist-oriented computational models of language acquisition.

A further issue that remains underdeveloped is attention to the *development* of language. The studies of cue availability in the child's environment and the computational treatments of this information consider all information simultaneously. This is an over-simplification: children's productions indicate that the whole language is not acquired in one step, but that multiple stages are observed, where learning progresses based on what has already been learned. Attempts to explain and exploit these learning stages in computational models have met with considerable success for simulating early processing constraints that facilitate later learning of complex syntactic structures (Elman 1993), phrasal productions and errors in young children (Freudenthal, Pine and Gobet 2006), and the development of the lexicon (Steyvers and Tenenbaum 2005). Such approaches could equally be applied to the searches for cues we have presented in this chapter: the reliability of phonological, prosodic, or distributional cues could be based on the most frequent, or earliest-learned words, and constructed incrementally, and such a constructionist approach would enhance the cognitive plausibility of the availability and process of use of such cues by the developing child.

11. Conclusion

The analyses of multiple cues for syntax acquisition reviewed in this chapter indicate the inherent richness of the language environment for the child. The child's first tentative cognitive steps are supported by a wealth of information to help in bootstrapping the structure of their first language. This "wealth of the stimulus" argument indicates that assumptions that the child's linguistic environment is inadequate for constraining the language should be reformulated. Principles of parsimony in scientific research require that the vast array of interacting cues that correlate with syntactic distinctions should not be under-estimated in terms of their contribution to constructing syntactic categories. This chapter has indicated that the cues we know about – that we have begun to measure in the child's environment – provide coherent and reliable information when considered in concert. This chapter also catalogues some potential cues that we know about but haven't yet taken into consideration in large-scale studies of the child's linguistic environment. In the words of Donald Rumsfeld (US Secretary of Defense under President George W. Bush), we've covered the known knowns, the known unknowns, but that still leaves the unknown unknowns – the cues we don't know we don't know – within the empiricist fold.

Corpora in-Language Acquisition Research

History, methods, perspectives

Edited by

Heike Behrens
University of Basel

John Benjamins Publishing Company
Amsterdam / Philadelphia

Library of Congress Cataloging-in-Publication Data

Corpora in language acquisition research : history, methods, perspectives / edited by Heike Behrens.
p. cm. (Trends in Language Acquisition Research, ISSN 1569-0644 ; v. 6)
Includes bibliographical references and index.
1. Language acquisition--Research--Data processing. I. Behrens, Heike.
P118.C6738 2008
401'.93--dc22 2008002769
ISBN 978 90 272 3476 6 (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA