LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

CONCEPTUAL REVIEW ARTICLE

Combining Language Corpora With Experimental and Computational Approaches for Language Acquisition Research

Padraic Monaghan[a,b] and Caroline F. Rowland[c]
[a]Lancaster University, [b]Max Planck Institute for Psycholinguistics, and [c]University of Liverpool

Historically, first language acquisition research was a painstaking process of observation, requiring the laborious hand coding of children's linguistic productions, followed by the generation of abstract theoretical proposals for how the developmental process unfolds. Recently, the ability to collect large-scale corpora of children's language exposure has revolutionized the field. New techniques enable more precise measurements of children's actual language input, and these corpora constrain computational and cognitive theories of language development, which can then generate predictions about learning behavior. We describe several instances where corpus, computational, and experimental work have been productively combined to uncover the first language acquisition process and the richness of multimodal properties of the environment, highlighting how these methods can be extended to address related issues in second language research. Finally, we outline some of the difficulties that can be encountered when applying multimethod approaches and show how these difficulties can be obviated.

Keywords first language acquisition; second language acquisition; computational modeling; corpus analysis; multiple cues

---

This work was supported by the International Centre for Language and Communicative Development at Lancaster University and at the University of Liverpool, funded by the Economic and Social Research Council (UK) [ES/L008955/1].
The copyright line for this article was changed on 28 February 2017 after original online publication.

Correspondence concerning this article should be addressed to Padraic Monaghan, Department of Psychology, Lancaster University, Lancaster LA1 4YF, UK. E-mail: p.monaghan@lancaster.ac.uk

DOI: 10.1111/lang.12221

Introduction

Multiple methods in language acquisition research are now well established, although they have not been introduced without difficulty. In this article, we describe the challenges of combining corpus, experimental, and computational approaches to research in first language acquisition. We discuss the benefits of multimethod approaches and show how these allow us to address fundamental questions in first language acquisition, with relevance to related issues in second language learning. Through three examples of successful combination of multiple methods, we illustrate these benefits and suggest how some of the difficulties of their application may be circumvented for second language acquisition research.

Historically, first language acquisition research has been dominated by attempts to describe formal mechanisms that can explain why children acquire the same language structures in the same order, despite great variation in the language environment (Chomsky, 1955/1975; Pinker, 1984). Consequently, much effort in language acquisition research has focused on determining the Universal Grammar that described the deep structure (or logical form) of children's linguistic constructions and how this is activated by exposure to a particular language (Chomsky, 1981). Similar arguments have been applied in second language acquisition research in terms of whether we need to posit an innately specified grammar to explain acquisition or whether there is sufficient positive and negative evidence for the learner to be able to acquire the language without innate structure (Ellis, 2013; Flynn, Martohardjono, & O'Neil, 1998; Hawkins, 2001; White, 1996).

This theoretical approach has been largely unconcerned with combining multiple approaches to investigate language acquisition, such as corpus data, experimental methods, or computational approaches. Instead, it focuses on creating descriptions of algebraic mechanisms that can explain particular isolated patterns of data (e.g., Crain & Nakayama, 1987). However, curiously, this research has for many years run in parallel with other productive streams of research that have investigated nonsyntactic aspects of the language process, such as the use of speech segmentation to isolate words from continuous speech, or of morphological segmentation to identify lexical structure (Chomsky, 2005). Rather than focusing on formal descriptions of the developing language, these investigations used multiple methods to determine how children might segment words and discover morphological structure (see, e.g., Monaghan & Christiansen, 2010; Yang, 2002). These approaches have been very productive in uncovering the richness of the environment and defining the computations in the learner that can apply to discover language structure (MacWhinney, 2005; Pullum & Scholz, 2002).

So, why has research in the acquisition of syntax been slower to take up these alternative methods and alternative perspectives? A major problem has been that corpus analyses of data, and computational models that take these corpora as input, have frequently been dismissed as irrelevant to the study of syntax. It has been argued that such approaches are unable to provide us with insight into the logical form of language, only surface structures (Chomsky, 1980), though see Sakas and Fodor (2001) for a data-driven approach to parameter setting. Furthermore, the critical data to test the development of key syntactic constructions are, by their nature, not present in corpora (Crain & Nakayama, 1987). This is because these constructions are only of interest in the first place because they are largely unattested in the learner's language experience. For example, the lack of example utterances demonstrating structure dependence in children's input (long-distance questions such as *is the dog that is running black*?) is taken as evidence that children must be innately constrained only to consider structure dependent grammars (Crain & Nakayama, 1987).

Arguments that deny the relevance of corpus data, computational models and behavioral studies have presented a substantial challenge to interdisciplinary research in language learning. This is because these arguments are, in theory, impervious to a change of perspective on the basis of these approaches. Despite this, interdisciplinary research has, in fact, made substantial headway in first language acquisition research. Below we summarize how this was achieved, which we hope provides a roadmap for constructive application of these methods to debates in second language acquisition (Cook & Singelton, 2014).

We propose three principal arguments against the irrelevance of corpus and computational methods in informing language acquisition. First, taking into account actual corpora of language motivates an understanding of language in its natural habitat, rather than in elicitation studies in a laboratory. This immediately leads to the realization that any sort of rule-based or categorical description of data requires, at the very least, some fuzzy boundaries. Thus, the constructivist approach to language emerged to describe the very subtle and complex interactions between lexical items and syntactic structures (Ellis, 2013). From a different tradition, but largely consistent with this constructivist approach, usage-based approaches to first language acquisition highlight the multifarious ways in which language is acquired and the close connection between children's actual exposure and their productions (Lieven & Brandt, 2011; Lieven, Salomo, & Tomasello, 2009; MacWhinney, 2005; Tomasello, 2003; Wonnacott, Boyd, Thomson, & Goldberg, 2012). Such usage-based approaches are now also beginning to gain currency in second language acquisition research (Ellis, 2017).

Second, even if the key data are not directly observable in the language learner's input, they may be observable indirectly through their overlap with other structures that are present in the learner's input (Pullum & Scholz, 2002). The idea that precise transformations or constructions must be within the learner's experience for them to be learned ignores the possibility that there may be multiple partial constraints within the child's experience that together are sufficient for learning. For instance, Reali and Christiansen (2005) tackled, head on, one of the key phenomena of the generative grammar approach: the fact that children do not make errors in auxiliary fronting, even when such constructions seldom, or never, occur in their input (Crain & Nakayama, 1987). Children never make the error "Is the man who hungry is ordering dinner?" but are able to reliably produce, or select as acceptable, "Is the man who is hungry ordering dinner?" Reali and Christiansen's (2005) model demonstrated that, even if there is no direct information about the movement of the correct auxiliary in long-distance questions in the input, learners' judgments could be guided by statistical information about co-occurrences of words in phrases. Ambridge, Rowland, and Pine (2008) also found that the pattern of correct use and error in 6- and 7-year-olds' long-distance questions could be explained by this type of sensitivity to surface co-occurrence patterns. Relatedly, MacWhinney (2005) demonstrated, with reference to child-directed corpus analyses, the abundance of indirect positive and negative evidence in child-directed input, which can constrain which constructions are permissible in a language, and point to no poverty of the stimulus if children are assumed to be able to generalize from their input. Similarly, in second language acquisition research, determining the sources of indirect, as well as direct, evidence in the language learning environment is of primary importance in determining the learning mechanisms that apply to language exposure (Cook, 2013; Gass, 2013; McEnery & Xiao, 2011).

The third argument against assuming that corpora, experimental, and computational work are irrelevant to studying language acquisition is that, without actually implementing processing mechanisms, such as innate constraints on grammar, it is never entirely clear if such mechanisms are sufficient or necessary to account for the data. By combining computational models with more explicit descriptions of the richness of the linguistic environment as the learner acquires language, we can test whether certain domain-general or domain-specific mechanisms are required. For instance, we can use computational models that apply domain-general statistical learning mechanisms to language input to discover how much structure can be developed via statistical learning. When the data are not effectively replicated by such models, this means that such domain-general approaches may not be sufficient. There is thus a clear place for computational models to test for sufficiency and necessity of assumptions in both first and second language learning research.

The opportunities that recent advances in data availability (e.g., Child Language Data Exchange System [CHILDES]; MacWhinney, 2000; McEnery & Xiao, 2011), corpus analysis techniques (e.g., McEnery & Hardie, 2012), and understanding of the range, and constraints on, human statistical processing (e.g., Frost, Armstrong, Siegelman, & Christiansen, 2015; Frost & Monaghan, 2016) mean that language acquisition research is undergoing something of a renaissance. Corpus research has enabled us to recover the richness of the stimulus and to more effectively ascertain the available information in the environment of the language learner. Alongside this, computational methods have enabled us to construct models that are able to respond to this language input and to test possible theories for how the learner interfaces with the environment. Then, these theories can be tested by determining how accurately they simulate behavioral data and, more importantly, how accurately they predict the interrelations among different constructions in a language in terms of when they are acquired. Also, predictions about how different languages or different experiences of the same language (such as reduced language input through an impoverished environment or perceptual impairment, or influence of first on second language representations, or effects of different cognitive developmental stages of first and second language learners) might affect this acquisition profile can be generated and tested.

In the next section we provide three case studies that indicate how multiple methods can be combined to increase our understanding of the process and phenomena of language acquisition. We use the outcomes of these case studies in first language acquisition to highlight how they illustrate opportunities for second language acquisition research. We then conclude by summarizing our view of the future promise of multimethodological approaches for both first and second language acquisition.

Three Case Studies of Multidisciplinary Approaches to Language Acquisition

There are numerous opportunities and challenges of working across disciplinary and methodological boundaries when using combined multiple methods. We describe three examples across three aspects of language acquisition: learning grammatical categories, learning morphological structure, and learning syntactic structures in terms of dative and double object constructions. In each case, the opportunities that are now available to researchers in each area demonstrate how detailed empirical studies have afforded us insight into important and surprising features of linguistic environments, how advances in computational modeling have increased our understanding of the complexity that can result from simple statistical functions when applied to real-world data, and how the dynamic interaction of the learner with the environment is also revealed through these current techniques. However, there are still substantial challenges faced by researchers using these methods, not least to resolve apparent disagreements over how psychological and computational data can inform linguistic theory. Reviewing these challenges can prevent similar pitfalls from occurring as multiple methods are developed for second language acquisition research.

Case 1: Grammatical Category Acquisition

A dominant position in linguistics regarding the acquisition of grammatical categories, such as Noun and Verb, was the assumption that the input was not sufficiently rich to result in their construction (e.g., Chomsky, 1955; Pinker, 1984). However, these nativist perspectives ran, for several years, alongside empiricist approaches that worked to uncover the potential richness of the stimuli (e.g., Fries, 1952). These empiricist approaches have led to recent comprehensive analyses of linguistic input that demonstrate the extent to which grammatical categories can emerge from the application of general statistical mechanisms. The nativist view arose initially as an important reaction to the radical behaviorist approach to language learning (e.g., Bloomfield, 1933; Skinner, 1957), where internal processing of language structure was considered irrelevant. However, as a consequence, the nativist view then denied the possibility that data-driven, structuralist approaches to language acquisition could inform the mainstream generativist approach to language learning (for a review, see Redington, Chater, & Finch, 1998).

Fries (1952) noted that classes of words systematically varied in terms of their syntagmatic relations and that, by contrasting usages of these classes, grammatical categories could be described. Thus, "the sum of all its environments" (Harris, 1954) could be used to determine the word's (syntactic) role. For instance, only words occurring within the frame *the__is/was/are/were good* can be nouns and only those occurring within *you__to* are verbs (Fries, 1952). Maratsos and Chalkley (1980) noted that these syntagmatic relations used to define categories of words may be useful for the process of acquisition of the categories in the first place. Consistent with the approach of Fries (1952), they proposed a series of computationally tractable local contexts in which words only from certain grammatical categories occurred. Furthermore, these local contexts were identified as occurring in child-directed speech and were sufficiently simple that they could feasibly be used to constrain learning of the categories.

Kiss (1973) provided an early attempt to describe clusters of words according to the context in which they occur in child-directed speech corpora. His model operated over 15,000 words of transcribed child-directed speech, and words were classified into clusters according to their co-occurrence with a set of 31 high-frequency words. If different words co-occurred with a similar set of other words, then they were determined to be similar in usage. The resulting clusters approximated grammatical category distinctions, such as *put* being clustered with some degree of accuracy with other verbs such as *see, is, are,* and *do*. The potential of grammatical category information being derived from even small corpora of speech was thus illustrated.

Once larger corpora became available for analysis, Redington et al. (1998) demonstrated the true power of the language environment for constructing grammatical categories. They took 2.5 million words of speech from the CHILDES database (later MacWhinney, 2000) and performed a cluster analysis of the most frequent 1,000 words according to whether they occurred one or two words before or after the 150 highest frequency words used as context words. The results were spectacular, with words clustered to a high degree of accuracy with words of the same category. Hence, the development of searchable and sufficiently extensive corpora of child-directed speech permitted the investigation of how effective such distributional cues might be for grammatical categorization.

There are two criticisms of the approach taken by Redington et al. (1998), however. One issue is of tractability: the clustering required 1,000 words × 150 high frequency words × 4 co-occurrence positions to be recorded, which presumably exceeds the working memory limitations of a child acquiring a language (Freudenthal, Pine, Jones, & Gobet, 2016). The second issue is that the clustering does not perfectly respect the grammatical roles of words in language: The clusters were not always populated by a single grammatical category, and some grammatical categories spanned several clusters.

To address the first of these, Mintz (2003) proposed a small set of constrained contextual co-occurrences in which words could occur as defining their category, thus providing a corpus-based implementation of Maratsos and Chalkley's (1980) proposals of local context defining the syntax role of a word. In analyses of small, but dense, corpora of individual child-directed speech, he showed that highly frequent co-occurring words could predict, with a high degree of accuracy, the category of the intervening word (e.g., *the__is* defines nouns). St. Clair, Monaghan, and Christiansen (2010) demonstrated that flexible frames, where the mechanism just considers preceding words (e.g., words following *the__*) and additively the succeeding words (e.g., words preceding *__is*), resolved the problem of overspecification, whereby words of the same category tended to occur in different contexts. Thus, highly computable information, consistent with children's cognitive capacity constraints, could result in effective grammatical categorization. Another solution to tractability was implemented by Li, Farkas, and MacWhinney (2004) in their DevLex model. This model generated a semantic representation for words that was based on co-occurrences, but that expanded according to the learner's growing vocabulary. So, the model started by storing co-occurrences among a small set of known words and gradually supplemented this as more words become known to the learner. A self-organizing map with the co-occurrence information as input reflected different grammatical categories topologically, such that words from the same category tended to occur close together in the map. With the exception of nouns, which were highly accurate throughout training, the categorization tended to become more accurate as the vocabulary grew.

However, these tractable methods are also subject to the second criticism of the Redington et al. (1998) approach, such that the clusters are not entirely coherent with regard to category. In one sense, such corpus analyses demonstrate that precise category boundaries are not available from the input. This is partly because utterances are noisy, being replete with false starts and other speech production errors. Furthermore, the categories themselves are noisy: ambicategoriality is profuse in natural language and there is also a richness to the internal structure within categories, such as subcategories of transitive and intransitive verb. In English, for instance, many nouns can be verbed or can be adapted to be adjectivey (Conwell & Morgan, 2012). These properties of language result in reduced accuracy within a category defined in terms of co-occurrences. In addition, a lack of coherence within categories can result from words of the same category not co-occurring in the same way with other words, resulting in reduced completeness of words in a defined category. For instance, subtle constraints on subclasses of words within a category, such as "strong" but not "powerful" co-occurring frequently with "tea," even though these words are both adjectives, they do not occur in the same contexts (Halliday, 1966), as reflected in constructionist grammars.

Yet hypotheses about grammatical categories and lexical membership of those categories can be based on sources of information in the child's environment that take into account other information available in the environment. For instance, Moeser and Bregmann (1972) showed that conjunctions of semantic categories with distributionally defined grammatical categories in an artificial language promoted learning the language structure. Similarly, there is cross-situational information (where an object or an action is usually present when the word is used), and pragmatic and social cues toward the referent being discussed (e.g. eye gaze or pointing), which occurs alongside grammatical distinctions within the language and can be used to identify the meaning of a word and its grammatical category membership (Monaghan, Mattock, Davies, & Smith, 2015). Yu and Ballard (2007) showed that a computational model based on small-scale child-directed speech corpora could use the co-occurrence of words with possible referents in the child's environment, as well as co-occurrence information within speech to constrain word categories.

Furthermore, there are other sources of information within the utterance itself that can constrain the acquisition of categories. This includes phonological and prosodic information. These sources are not considered in standard linguistic analyses, but can be critical in ascertaining the information present in children's environment available for language acquisition. Such a view requires a change in perspective from the linguistic convention of the autonomy of syntax (e.g., Jackendoff, 2002), whereby other aspects of language and communication (such as phonology, or discourse-level phenomena) are assumed to be modular and not involved in syntactic construction, a view that still dictates the design of descriptive models of speech production (e.g., Ferreira, 2010).

We know, for example, that phonological and prosodic information does distinguish words belonging to different grammatical categories. Function words tend to be shorter, and contain more voiced consonants and centralized vowels, than content words (Cutler, 1993). Furthermore, these phonological distinctions are perceptible to infants as early as 3 days of age (Shi, Werker, & Morgan, 1999). Within content words, further distinctions are available, such as the fact that, in English, nouns containing more phonemes and syllables than verbs on average and are more likely to have first-syllable stress than verbs (Kelly, 1992). The usefulness of such cues for categorization, however, can again only be appraised by empirical investigations of the learner's actual language exposure. In a corpus analysis of five million words of speech spoken in the presence of children taken from the CHILDES database (MacWhinney, 2000), Monaghan, Chater, and Christiansen (2005) distinguished the grammatical categories of words from a small set of phonological and prosodic distinctions. Furthermore, these sound cues were found to be most reliable when the cues from distributional, co-occurrence information were weaker at constraining the grammatical categories. Monaghan, Christiansen, and Chater (2007) found that the interactive effects of phonological and distributional information sources were also observable in child-directed Japanese, Dutch, and French speech and were thus generalizable from English. Hence, these multimodal analyses of corpora enabled the interplay of information sources in the learner's environment to be discovered.

In summary, the challenges of alternative approaches to language acquisition research—alternatives to traditional generativist and structuralist perspectives—have previously been limited by our understanding of the statistical mechanisms that are available to process language input, and by our limited understanding of the rich, multimodal input that children receive. Combining computational and corpus-based approaches has been key to improving the validity of early structuralist accounts that aimed to show how domain-general mechanisms could apply to language, but did not have sufficient data to effectively reflect the language learner's experience. The development of ever larger second language acquisition corpora (Granger, Gilquin, & Meunier, 2015; McEnery & Xiao, 2011)—when complemented with a description of multiple information sources: distributional as well as prosodic and environmental features—can similarly inform knowledge about the process of second language acquisition. The results from this approach applied to first language acquisition suggest that innate grammatical categories are not required to describe behavior. Parallel arguments in second language learning can address claims that innate structure precedes language experience (Flynn et al., 1998; Hawkins, 2001) and give a clearer indication of the mechanisms of second language learning.

Case 2: Morphological Development: The Optional Infinitive Phenomenon

Behavioral studies show that some patterns in first language acquisition appear to be systematic across children, and relatively stable, in that they are sustained for some time. One such pattern in children's productions is the omission of agreement and tense markers in morphological acquisition. These markers are relatively late acquired, thus, children's first multiword utterances have a telegraphic feel (Brown & Fraser, 1963). Children say, for example, "Daddy eat" instead of "Daddy is eating" and "he want more" instead of "he wants more." However, when they occur, they are produced correctly, with relatively few errors.

Such observations have been explained by theoretical accounts that take as their starting point an internalized morphological grammar that becomes gradually more expressed with age, but is underspecified at an early age (e.g., Brown, 1973; Legate & Yang, 2007). An alternative account describes general cognitive constraints such as limited working memory, which results in shorter utterance lengths, thus reducing the constructions of polymorphemic words (Bloom, 1990). However, these theories have been somewhat Anglo-centric, as Wexler (1998) noted that children's early productions in other languages indicate that it is the infinitive form that seems to be used in place of the finite form, such as in the Dutch, "papa eten" instead of "papa eet" (direct translations of the English example above). Hence, these errors are referred to as optional infinitive (OI) errors.

So, what accounts for use of the infinitive in place of the finite verb form? Freudenthal, Pine, and Gobet (2006) constructed a *model of syntax acquisition in children* (MOSAIC) that was based on general principles of memory processing. A key feature of the model is that it responds incrementally to input to develop an internal representation of the language. It stores sequences of increasing length with exposure and produces utterances based on its current knowledge state, which allows the researchers to test its knowledge at different points of development. Critically, the model's incorporation of input into its internal representation of sequences is constrained by memory limitations, whereby lexical items from the end of an utterance are more likely to be stored than those at the beginning, in line with apparent observations of salience at different points in child-directed speech (Shady & Gerken, 1999).

The MOSAIC model was applied to child-directed speech corpora to determine whether these general cognitive constraints on sequence processing and memory representation were sufficient to account for the pattern of optional infinitive errors in children. Freudenthal et al. (2006) assessed the explanatory adequacy of these computational mechanisms for corpora of English and Dutch child-directed speech. An important requirement of the corpora was that they were longitudinal, such that a child's changing representation could be unfolded over time and their productions over development could be related to the exposure they receive. They also had to be intensive, such that a representative input that the child receives can be ascertained from the data. Testing generalization over languages also entails that the mechanisms are generalizable across questions and languages, and not just fitted to produce a mapping between a particular input and output in a particular language.

The corpora used came from CHILDES (MacWhinney, 2000) and comprised 1-hour recordings of the same children every 2 weeks for 2 years for the Dutch corpora (Bol, 1996) and approximately every 10 days for 1 year for the English corpora (Theakston, Lieven, Pine, & Rowland, 2001). Children were aged between 1;5 and 2;0 years when recordings began. The model was trained by inputting the corpora chronologically and was stopped and tested at various points during training to simulate its productions at different stages of development (as measured by mean length of utterance). A substantial benefit of the model is that the effect of infinitive forms in the corpus can be distinguished (in English) from the surface form similarity of first-person forms (e.g., the model producing “go” derived from input “to go,” and from “I go” can be discerned).

The model was effective in simulating the relation between occurrence of OI errors and utterance length in both languages, showing a close correspondence between the children’s OI productions and those that the model predicted. Furthermore, the model’s mechanisms were shown to interact with differences in word order from the different language corpora. Dutch is constrained to have nonfinite verb forms largely in sentence final position, whereas they occur to a greater degree sentence internally in English. This makes the nonfinite verb forms more salient in Dutch and hence represented more robustly in the model, resulting in a greater incidence of OI errors in Dutch than English.

Freudenthal, Pine, Aguado-Orea, and Gobet (2007) further showed that a slightly adapted MOSAIC model could be applied across four languages: Dutch, English, German and Spanish. The Spanish simulation was particularly interesting, because Spanish children produce very few OI errors, despite superficial similarities to Dutch and German in the number of finite and nonfinite verb forms that are present in the input. Using the same parameterization of the MOSAIC model across languages, the researchers modeled the different degrees of OI productions in the child learners of the different languages. The difference between languages came from an interaction between the distributional statistics of the language and MOSAIC’s utterance final bias. Although Spanish children hear similar numbers of nonfinite verb forms as Dutch and German children, only 26% of these occur in utterance final position, which means that they are far less likely to be learned by the model. In other words, the simple, general computational mechanisms within MOSAIC react differently with the corpora to which they are exposed and thus provide a better fit to crosslinguistic data than qualitative models designed to describe the data from a generativist, rather than a data-driven, perspective (Freudenthal, Pine, & Gobet, 2010).

The application of cognitive constraints implemented in domain-general computational modeling to language learning has permitted greater specification of the features of acquisition that cannot be explained only with domain-general mechanisms and that may require language-specific mechanisms for their explanation. As in the case of the modeling approach taken by Freudenthal and colleagues, a whole range of morphological properties of children’s productions can be explained by only very general constraints interacting with the rich complexity of the language environment. This case study also exemplifies how computational models provide extra value over corpus analyses alone, because the representations of the input can be tracked from the way in which they are internally stored by the system, through to how they are realized in productions by the system. Then, linking these computational data to children's actual behavior enables deeper insight into the child's knowledge about their language that are observed in articulation.

Similar models could be applied to longitudinal corpora in second language acquisition, given that these corpora are now being developed with sufficient detail (Granger et al., 2015). Applying computational models of acquisition to second language corpora also enables testing of some of the fundamental issues in second language acquisition research, such as the fact that the cognitive capacity of second language learners varies from those of first language learners (Andringa, 2014; DeKeyser, 2013; DeKeyser, Alfi-Shabtay, & Ravid, 2010; Johnson & Newport, 1989). MOSAIC could be adapted, for instance, in terms of its memory span, to simulate changes in working memory, or speech production capacity, in younger and older learners (Cook, 2010; Pienemann, 1998). Furthermore, the influence of learning a first language on the structures acquired in a second language (DeAnda, Poulin-Dubois, Zesiger, & Friend, 2016) can also be explicitly tested in such models, and the extent to which first and second languages are similar or distinct can then be characterized explicitly in an implemented model (Li, 2013). For instance, the extent to which morphological feature discovery can transfer from one language to another, using similar principles to MOSAIC in a bilingual version, can raise specific predictions about exactly where, in the representation of structure, morphology is processed.

Case 3: The Acquisition of Sentence Structure

In the case studies above, we have focused on corpus data and computational models. Our third and final case study concerns the debate over how children acquire sentence structure and demonstrates how combining methodological approaches can help explain apparently contradictory experimental behavioral data. The debate centers on the nature of children's early knowledge of the syntactic structures of their language, for example, their knowledge of how to form active transitives (e.g., *the boy pushed the girl*) or prepositional and double object datives (e.g., *the boy gave the girl an orange/an orange to the girl*).

On the one hand, early abstraction theorists argue that children form sentences using abstract categories from the beginning; mapping words onto semantic (e.g., agent, patient) or syntactic categories (subject, object), and then combining these categories to form sentences, aided by innate mapping rules (see Fisher, 2001; Pinker, 1984) and/or the triggering of parameterized principles (Gibson & Wexler, 1994). On the other hand, item-based theorists suggest that children start with knowledge only of how to sequence lexical items (words) and build their language from the bottom up (see MacWhinney, 2014, for a historical perspective); initially forming sentences using inventories of item-based constructions (e.g., using a [pusher]-push-[pushee] construction to form sentences like *I pushed the girl* or *he pushed me*; Akhtar, 1999). These are later, slowly built, via generalization and analogy, into more abstract categories (Lieven, 2014; Tomasello, 2003).

The behavioral experimental data used to test the predictions of these theories yields apparently contradictory results. Studies of children's comprehension seem to support the early abstraction view, demonstrating that children are capable of parsing abstract transitive sentences correctly from 2 years of age (Naigles, 1990), if not earlier (Yuan, Fisher, & Snedeker, 2012), and dative sentences from 3 years of age (Rowland & Noble, 2011). Children can do this even when such sentences contain novel verbs, which rules out the possibility of them using a verb-based formula (e.g., [pusher]-push-[pushee]) to guide interpretation. For example, Gertner, Fisher, and Eisengart (2006) reported that 21-month-old children were above chance at using word order to identify the correct referent of the sentence *the duck is gorping the bunny* in the presence of a foil referent in which a bunny was acting on a duck. This suggests that children have abstract knowledge of English word order that generalizes to novel verbs from at least 21 months of age.

However, data from elicited production paints a very different picture that seems to support the item-based view. In production, 2-year-olds seem unable to use a novel verb in a transitive sentence unless they have already heard it modeled in that structure. For example, Olguin and Tomasello (1993) showed that children who heard novel verbs with only one argument (e.g., *Cookie Monster's gorping*) were unable to subsequently produce transitives with those verbs (e.g., *Cooking Monster's gorping Mickey Mouse*). Similarly, Akhtar (1999) found that 2-year-olds who heard novel verbs in weird word orders (e.g., *Elmo's the car gopping*) were significantly less likely to correct them to standard English (*Elmo's gopping the car*) than 3- and 4-year-olds. These studies contradict the findings from comprehension and suggest instead that 2-year-olds are unable to access the abstract syntactic knowledge necessary to produce correct transitive sentences with novel verbs until much later in life.

Fortunately, computational modeling provides a solution that explains both sets of data, in the form of Chang, Dell, and Bock's (2006) and Chang, Janciauskas, and Fitz's (2012) Dual-path model. This is a connectionist model comprising two pathways: a sequencing system that learns how to sequence words correctly in syntactic structures and a meaning system that learns to link words with meanings (concepts and roles) and contains the event semantics that represents, for example, number of arguments, tense and aspect. The dual-route nature of the model and the fact that the sequence system only connects directly with the roles, not the concepts or words, in the meaning system means that it can do what traditional simple recurrent neural networks cannot do: it can generalize in sentence production (Chang, 2002). For example, when the sequencing system learns how to sequence the sentence *The dog carries the flower*, it is learning how to sequence the roles associated with the words (i.e., equivalent to agent-action-object, though note that in the model the roles are characterized differently). Thus when it later is asked to produce *The cat carries the flower*, the fact that the cat is linked to the same role as the dog means that the model can immediately transfer what it has learned about how to sequence this role to the new sentence (see Chang et al., 2006, for a more detailed description).

Unlike in MOSAIC, the input to the model is a simplified, toy input of 8,000 different sentence-message pairs. However, the toy input was designed to approximate the range of simple syntactic structures in children's real input: intransitives, active transitives, passives, and datives, as well as simulating different tenses, aspect and the correct use of determiners. In learning, the model uses back-propagation of error to learn to sequence roles based on this input, calculating the difference between the predicted and the actual next word and gradually converging on adultlike representations of syntactic structure.

Testing the model during learning allowed Chang and colleagues to explain the apparent contradiction between results from production and comprehension. To do this, the model was given both preferential looking tasks (given novel-verb transitives sentences and was then checked to see if it matched these sentences to the correct causative meaning) and elicited production tasks (given a causative message and required to output a matching sentence). Crucially, the model was given both these tasks at the same time points, every 2,000 epochs of the learning cycle.

Doing so revealed a potential explanation of the apparently contradictory results from the behavioral studies. Like children, the model exhibited different levels of performance on the production and preferential looking tasks despite having the same underlying level of grammatical knowledge at each developmental stage. The model's ability to produce transitive sentences with novel verbs developed very gradually; by the 12,000 epoch it was still producing correct productions only 35% of the time. In contrast, performance on the preferential looking task developed much more quickly; the model was more than 50% correct, on average, at the 12,000 epoch. The difference in performance across the two tasks came from the nature of the tasks themselves. The production task required the model to make a sequence of correct decisions, making a choice about each word of the produced utterance, meaning that there were multiple opportunities for error early in development, when the model still had only partial form–meaning mappings. The preferential looking task, however, was less reliant on a series of decisions, so partial form mappings allowed the model to choose the correct match more often than not, which is all that is required for correct performance.

The contribution of the Dual-path model to this debate has been significant, not only in resolving an apparently contradictory evidence base, but in emphasizing how important it is to get converging evidence from multiple methods when assessing children's performance in language acquisition studies. In this case, if we had studied only elicited production or preferential looking data, we would have received a distorted picture of children's knowledge of syntactic structure at different ages. By combining multiple methodologies, and by building computational models that simulate both the complexity of the environment and of the learning mechanisms, we get a much more accurate, detailed picture of children's syntactic development.

Again, investigations of second language learning using computational models such as the Dual-path model can provide insight into co-influence of first and second languages. The extent to which such models co-opt previously acquired language structure, or construct representations anew are issues that can be directly addressed with such computational models (e.g., Li & Zhao, 2013). They can then be related closely to behavioral data to decide between apparently competing behavioral results and also to hone theoretical proposals for when and how co-influence of language might affect performance.

Future Directions

These three case studies demonstrate the importance of combining computational modeling to extract the structure available in natural language corpora to inform behavioral observations of the processes involved in language acquisition. Of primary importance has been the collection and accessibility of large corpora of child-directed speech, collected intensively—such that individual differences between children can be observed within the environment and related to particular development of language structures—but also collected longitudinally—such that an understanding of how the richness of the stimulus unfolds over developmental time can also be plotted. These observations have enabled the field of first language acquisition to change radically its perspective on the learnability of language from input and has facilitated the emergence of a new, data-driven approach to investigating language acquisition in all its diversity and complexity.

Similarly, we predict that the expansion of data and descriptions of the environment for second language acquisition will facilitate parallel debates on learning in second language and allow more explicit tests of the extent to which performance can be predicted from input (see, e.g., Ellis, 2017). Describing the environment, and the learner's place in that environment, will be important also for addressing questions about differences between younger and older second language learners acquiring language (Cook, 2013; DeKeyser, 2013; Johnson & Newport, 1989) and the interaction between first and second language processing (DeAnda et al., 2016). Two further questions in second language learning are also well served by combining corpus, computational, and experimental methods (e.g., Li, 2013): the extent to which learning at different ages is affected both by extralinguistic and linguistic differences in input (Gass, 2013; Long, 1996); and by differences in prior exposure or capacity (such as working memory, Cook, 2010). For both first and second language research, we argue that the starting point for language acquisition research should now be investigation of the potential structure present in the environment, rather than assuming structure within the individual.

However, there are future innovative techniques that will further facilitate the development of theoretical views of both first and second language acquisition. New technology is making it easier to collect, code, and analyze naturalistic data and to perform experiments with language learners in the community. We already have a rich corpora of child-directed speech on CHILDES (MacWhinney, 2000) and growing corpora of second language learner's experience (Granger et al., 2015; McEnery & Xiao, 2011). However, more information always provides a better indication of the actual language environment. Automated language analysis systems such as LENA (e.g., Zimmerman et al., 2009) provide rough, but accurate-enough, global measures of the number of utterances that learners are exposed to on a daily basis. Transcription aids such as Blitzscribe automatically identify and segment speech in audio data, making hand-transcription up to six times faster (Roy & Roy, 2009). Further developments that enable automatic encoding of the actual words, and not just summative statistics about quantity, will provide a sea change in our ability to determine the precise input that learners receive, the variation in that input, and the importance of variation in language development. Though technological advances in this area are understandably slow given the scale of the task, there are recent advances in speech recognition technology that bring this ever closer to the researcher's toolbox (Hinton et al., 2012).

Furthermore, corpora are beginning to be collected that embed language in its broader environmental context—so including multimodal information about gesture, objects in the environment, and even the viewing direction of children and adults during communicative exchanges. This enables the full richness of the language learning environment to be uncovered (Smith, Yu, Yoshida, & Fausey, 2015). Accompanying these are formalisms by which such information can be hand coded within multimodal corpora (e.g., ELAN; Lausberg & Sloetjes, 2009). One notable instance of the benefit of this approach in first language acquisition is a study by Yurovsky, Smith, and Yu (2013), who found that identifying the referent of a word is substantially easier than previously assumed when the child's view is taken into account. Instead of the multiple alternative possibilities that were assumed to be present for each uttered word, head-mounted cameras on both children and adults demonstrated that, whereas alternatives were present for adults speaking to children, the child's view was reduced such that referential ambiguity was almost entirely avoided. Thus, the interaction of attention, environment, and language conspire to reduce uncertainty and promote useful information for the child in language acquisition. How these multiple cues play out in second language learning could be a key contributor to understanding the challenges and opportunities that a multimodal environment provides to learners. However, critical to permitting advances in the field is open-source and widely available corpora and tools (e.g., Talkbank, MacWhinney, 2007), as without publically available technologies and corpora, progress in first language acquisition would not have been possible.

Developments in computational modelling have proceeded in tandem to accommodate these multimodal sources of information. For instance, there is potential to extend models of sentence production (e.g., Dell & Chang, 2014), such that linguistic input interacts with information from a visual scene to constrain learning of objects and actions, and even thematic roles such as agent and patient. Smith, Monaghan, and Huettig (2014) have developed a model where information about the visual scene, phonology, and meaning all interact in simulating processes of language comprehension. Models of word learning are beginning to include information about visual attention (Samuelson, Jenkins, & Spencer, 2015), and even also the learner's actions which in turn affect their environment (Morse, Benitez, Belpaeme, Cangelosi, & Smith, 2015). Yet, developments to accommodate realistic representations of the language learner's experience are still at an early stage of progress.

The development of more automated methods of collecting behavioral data is another future direction for the field. There are technological advances that permit finer-grained investigations of children's responses to comprehension questions, such as touchscreen tablets, where data can be collected without the overhead of hand coding of responses post hoc. Similarly, eye-tracking equipment is now portable, unintrusive, and vitally useful for determining eye gaze for learners of all ages, enabling implicit processes as well as explicit decisions to be recorded. Yet, experimental work on learners' language comprehension and elicited production, and the predictors of these language skills, needs to keep up with the fast pace of corpus development and sophistication of the computational models. The dependencies between different language learning tasks, such as the role of speed of processing in early vocabulary development (Marchman & Fernald, 2008), and the interactive effects of learning to segment speech and acquire grammatical categories (e.g., Frost & Monaghan, 2016) require intensive, longitudinal assessments to fully understand the role of experience on all aspects of language learning.

To conclude, we have shown that recent technological advances, coupled with the collaborative accumulation of open-source and increasingly detailed corpora of child-directed speech, have enabled the field of language acquisition to address questions of the nature and process of language acquisition from an empirical perspective. We suggest that parallel developments in second language acquisition research will benefit from the lessons learned by combining methods for first language acquisition. Fundamentally important to this enterprise is interdisciplinarity, which means that behavioral studies of language development align with advances in our understanding of processing through implementation in computational models. Applying computational modeling to growing specification of the language learner's environment enables a description of the processes by which language structure—vocabulary, morphology, and syntax—can be learned from the input. For second language learning, how first language structure constrains second language learning can also be addressed by applying computational models with prior experience to acquisition of an additional language (Cuppini, Magosso, & Ursino, 2013; Li & Zhao, 2013; MacWhinney, 1987). Attention to the learner's environment as a whole must be taken into consideration, and not only the content of linguistic utterances; this is critically important to understanding the task facing the language learner. The variety and variation of language experience and language exposure is now, for the first time through these multimethodological approaches, being revealed.

Final revised version accepted 17 September 2016

References

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language*, *26*, 339–356. doi:10.1017/S030500099900375X

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science*, *32*, 222–255. doi:10.1080/03640210701703766

Andringa, S. (2014). The use of native speaker norms in critical period hypothesis research. *Studies in Second Language Acquisition*, *36*, 565–596. doi:10.1017/S0272263113000600

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry*, *21*, 491–504.

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

Bol, G. W. (1996). Optional subjects in Dutch child language. In C. Koster & F. Wijnen (Eds.), *Proceedings of the Groningen Assembly on Language Acquisition* (pp. 125–135). Groningen, Netherlands: Centre for Language and Cognition Groningen.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Brown, R., & Fraser, C. (1963). The acquisition of syntax. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behavior and learning: Problems and processes* (pp. 158–97). New York: McGraw-Hill.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609–651. doi:10.1207/s15516709cog2605_3

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272. doi:10.1037/0033-295X.113.2.234

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass*, *6*, 259–278. doi:10.1002/lnc3.337

Chomsky, N. (1955/1975). *Logical structure of linguistic theory*. New York: Plenum.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Chomsky, N. (1981). *Lectures on government and binding. Studies in generative grammar*, *vol. 9*. Dordrecht, Netherlands: Foris.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*, 1–22. doi:10.1162/0024389052993655

Conwell, E. & Morgan, J. L. (2012). Is it a noun or is it a verb? Resolving the ambicategoricality problem. *Language Learning and Development*, *8*, 87–112. doi:10.1080/15475441.2011.580236

Cook, V. J. (2010). The relationship between first and second language acquisition revisited. In E. Macaro (Ed.), *The continuum companion to second language acquisition* (pp. 137–157). London: Continuum.

Cook, V. (2013). *Second language learning and language teaching*. London: Routledge.

Cook, V., & Singleton, D. (2014). *Key topics in second language acquisition*. Bristol, UK: Multilingual matters.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, *63*, 522–543. doi:10.2307/415004

Cuppini, C., Magosso, E., & Ursino, M. (2013). Learning the lexical aspects of a second language at different proficiencies: A neural computational study. *Bilingualism: Language and Cognition*, *16*, 266–287. doi:10.1017/S1366728911000617

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research*, *22*, 109–131.

DeAnda, S., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2016). Lexical processing and organization in bilingual first language acquisition: Guiding future research. *Psychological Bulletin*, *142*, 655–667. doi:10.1037/bul0000042

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*, 52–67. doi:10.1111/j.1467-9922.2012.00737.x

DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*, 413–438. doi:10.1017/S0142716410000056

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1634), 20120394. doi:10.1098/rstb.2012.0394

Ellis, N. C. (2013). Second language acquisition. In G. Trousdale & T. Hoffmann (Eds.), *Oxford handbook of construction grammar* (pp. 365–378). Oxford, UK: Oxford University Press.

Ellis, N. C. (2017). Cognition, corpora, and computing: Triangulating research in usage- based language learning. *Language Learning*, *67*(S1), 41–66.

Ferreira, V. S. (2010). Language production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 834–844. doi:10.1002/wcs.70

Fisher, C. (2001). Partial sentence structure as an early constraint on language acquisition. In B. Landau, J. Sabini, J. Jonides, & E. L. Newport (Eds.), *Perception, cognition, and language: Essays in honor of Henry and Lila Gleitman* (pp. 275–290). Cambridge, MA: MIT Press.

Flynn, S., Martohardjono, G., & O'Neil, W. (1998). *The generative study of second language acquisition*. Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science*, *31*, 311–341. doi:10.1080/15326900701221454

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in Dutch and English using MOSAIC. *Cognitive Science*, *30*, 277–310. doi:10.1207/s15516709cog0000_47

Freudenthal, D., Pine, J., & Gobet, F. (2010). Explaining quantitative variation in the rate of optional infinitive errors across languages: A comparison of MOSAIC and the variational learning model. *Journal of Child Language*, *37*, 643–669. doi:10.1017/S0305000909990523

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2016). Developmentally plausible learning of word categories from distributional statistics. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 674–679). Austin, TX: Cognitive Science Society.

Fries, C. C. (1952). *The structure of English*. London: Longmans.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*, 117–125. doi:10.1016/j.tics.2014.12.010

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74. doi:10.1016/j.cognition.2015.11.010

Gass, S. M. (2013). *Input, interaction, and the second language learner*. London: Routledge.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*, 684–691. doi:10.1111/j.1467-9280.2006.01767.x

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*, 407–454.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learning corpus research*. Cambridge, UK: Cambridge University Press.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of J R Firth* (pp. 148–162). London: Longman.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 140–162. doi:10.1080/00437956.1954.11659520

Hawkins, R. (2001). The theoretical significance of Universal Grammar in second language acquisition. *Second Language Research*, *17*, 345–367.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*, 82–97.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99. doi:10.1016/0010-0285(89)90003-0

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349–364. doi:10.1037/0033-295X.99.2.349

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, *7*, l–41. doi:10.1016/S0079-7421(08)60064-X

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers*, *41*, 841–849. doi:10.3758/BRM.41.3.841

Legate, J. A., & Yang, C. (2007). Morphosyntactic learning and the development of tense. *Language Acquisition*, *14*, 315–344. doi:10.1080/10489220701471081

Li, P. (2013). Computational modeling of bilingualism: How can models tell us more about the bilingual mind? *Bilingualism: Language and Cognition*, *16*, 241–245. doi:10.1017/S1366728913000059

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*, 1345–1362. doi:10.1016/j.neunet.2004.07.004

Li, P., & Zhao, X. (2013). Self-organizing map models of language acquisition. *Frontiers in Psychology*, *4*, 828. doi:10.3389/fpsyg.2013.00828

Lieven, E. (2014). First language development: A usage-based perspective on past and current research. *Journal of Child Language*, *41*, 48–63.

Lieven, E., & Brandt, S. (2011). The constructivist approach. *Infancia y Aprendizaje*, *34*, 281–296. doi:10.1174/021037011797238586

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*, 481–508. doi:10.1515/COGL.2009.022

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.

MacWhinney, B. (1987). Applying the competition model to bilingualism. *Applied Psycholinguistics*, *8*, 315–327. doi:10.1017/S0142716400000357

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

MacWhinney, B. (2005). Item-based constructions and the logical problem. In *Proceedings of the workshop on psychocomputational models of human language acquisition* (pp. 53–68). Ann Arbor, MI: Association for Computational Linguistics.

MacWhinney, B. (2007). The TalkBank project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora: Synchronic databases*, *Vol.1* (pp. 163–180). Houndmills, UK: Palgrave-Macmillan.

MacWhinney, B. (2014). Item-based patterns in early syntactic development. In T. Herbst, H.-J. Schmid, & S. Faulhaber (Eds.), *Constructions, collocations, patterns* (pp. 33–70). Berlin, Germany: DeGruyter Mouton.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (Vol. *2*; pp. 127–214). New York: Gardner Press.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16. doi:10.1111/j.1467-7687.2008.00671.x

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. *2*, pp. 364–380). London: Routledge.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117. doi:10.1016/S0010-0277(03)00140-9

Moeser, S. D., & Bregman, A. S. (1972). The role of reference in the acquisition of a miniature artificial language. *Journal of Verbal Learning and Verbal Behaviour*, *11*, 759–769. doi:10.1016/S0022-5371(72)80010-0

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential contribution of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*, 143–182. doi:10.1016/j.cognition.2004.09.001

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*, 545–564. doi:10.1017/S0305000909990511

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*, 259–305. doi:10.1016/j.cogpsych.2006.12.001

Monaghan, P., Mattock, K., Davies, R., & Smith, A. C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive Science*, *39*, 1099–1112. doi:10.1111/cogs.12186

Morse, A. F., Benitez, V. L., Belpaeme, T., Cangelosi, A., & Smith, L. B. (2015). Posture affects how robots and infants map words to objects. *PloS one*, *10*(3), e0116012. doi:10.1371/journal.pone.0116012

Naigles, L. R. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*, 357–374. doi:10.1017/S0305000900013817

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, *8*, 245–272. doi:10.1016/S0885-2014(93)80001-A

Pienemann, M. (1998). Developmental dynamics in l1 and l2 acquisition: Processability theory and generative entrenchment. *Bilingualism: Language and Cognition*, *1*, 1–20. doi:10.1017/S1366728998000017

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pullum, G. K., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review*, *19*, 9–50. doi:10.1515/tlir.19.1-2.9

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028. doi:10.1207/s15516709cog0000_28

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic structures. *Cognitive Science*, *22*, 425–469. doi:10.1207/s15516709cog2204_2

Rowland, C. F., & Noble, C. H. (2011). The role of syntactic structure in children's sentence comprehension: Evidence from the dative. *Language Learning and Development*, *7*, 55–75. doi:10.1080/15475441003769411

Roy, B. C., & Roy, D. (2009). *Fast transcription of unstructured audio recordings*. *Proceedings from INTERSPEECH*, Brighton, UK. Retrieved from http://www.isca-speech.org/archive/interspeech_2009/index.html

Sakas, W., & Fodor, J. D. (2001). The structural triggers learner. In S. Bertolo (Ed.), *Language acquisition and learnability* (pp. 172–233). Cambridge, UK: Cambridge University Press.

Samuelson, L. K., Jenkins, G. W., & Spencer, J. P. (2015). Grounding cognitive level processes in behavior: The view from dynamic systems theory. *Topics in Cognitive Science*, *7*, 191–205. doi:10.1111/tops.12129

Shady, M., & Gerken, L. (1999). Grammatical and caregiver cue in early sentence comprehension. *Journal of Child Language*, *26*, 163–176. doi:10.1017/S0305000998003730

Shi, R., Werker, J., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*, B11–B21. doi:10.1016/S0010-0277(99)00047-5

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology*, *75*, 28–54. doi:10.1016/j.cogpsych.2014.07.002

Smith, L., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*, 407–419. doi:10.1080/15248372.2014.933430

St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*, 341–360. doi:10.1016/j.cognition.2010.05.012

Tomasello, M. (2003). *Constructing a language: A usage-based theory of child language acquisition*. Cambridge, MA: Harvard University Press.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152. doi:10.1017/S0305000900004608

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua*, *106*, 23–79. doi:10.1016/S0024-3841(98)00029-1

White, L. (1996). Universal Grammar and second language acquisition: Current trends and new directions. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 85–120). San Diego, CA: Academic Press.

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*, 458–478. doi:10.1016/j.jml.2011.11.004

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford, UK: Oxford University Press.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–2165. doi:10.1016/j.neucom.2006.01.034

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development*, *83*, 1382–1399. doi:10.1111/j.1467-8624.2012.01783.x

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*, 959–966. doi:10.1111/desc.12036

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., et al. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, *124*(1), 342–349. doi:10.1542/peds.2008-2267