Exploiting corpora for language acquisition research

Katherine Demuth

1. Introduction

Language corpora have long provided a rich source of information about children's language development. Many of these first appeared in the form of diary studies (Darwin 1877; Deville 1891), and this continues to be a rich source of information still exploited today (e.g., Bowerman 1974). However, the increasing affordability of audio/video recording equipment, computers and memory, plus the creation of a central public storage venue for child language corpora (CHILDES, MacWhinney 2000), has led to a recent surge in language acquisition corpora (see MacWhinney this volume). The further development of tools useful for exploiting these computerized corpora (e.g., CLAN (MacWhinney 2000), PHON (Rose, MacWhinney, Byrne, Hedlund, Maddocks and O'Brien 2005)) has enhanced the usability of these corpora for addressing research questions at multiple levels of linguistic structure (e.g., phonology, morphology, the lexicon), and in children as well as adults. This growth in the use of large datasets follows a larger trend that is now common in fields such as computational linguistics, speech research, sociolinguistics, and historical linguistics.

Although technological developments have facilitated the ability to collect and analyze these large corpora, the primary motivation for corpus construction (which is still tedious and labour intensive to transcribe) has been to provide the data needed to address certain theoretical issues. In particular, corpora have been useful for examining the course of language acquisition over time, as well as characteristics of the input language learners typically hear. The amount of data collected, how it is collected, and how it is prepared and transcribed, all influence the utility of a particular corpus. This chapter reviews some of the issues that are important to the creation and use of corpora, and their potential for assessing children's knowledge of language.

2. Corpus creation

Ideally, any corpus should be collected with specific theoretical issues in mind. This will guide decisions about the corpus design. This involves decisions regarding the number of children to be included in the study, the setting for recording (home, lab, school), the interlocutors (parent, siblings, experimenter), the activities ('natural', prompted with a specific set of toys/tasks), the amount of data recorded (how long recordings should be), the number of sessions/ages recorded per child (i.e., longitudinal or not, how frequently sampled), the placement and type of microphones used (critical for conducting acoustic analysis), and the use of video. Similar decisions arise at the level of transcription and coding (orthographic, phonetic, situational information, etc.).

3. Corpus size

The quantity of data available in particular corpus is an issue of critical importance. As Rowland, Fletcher and Hughes (this volume) discuss, estimating both errors and productivity present different problems depending on corpus size. Various statistical procedures can be used to estimate the probability of both. However, to some extent, corpus construction can also be designed to address some of these issues. For example, the examination of certain relatively high-frequency phonological phenomena (e.g., segmental acquisition, the acquisition of coda consonants in Germanic languages) can more easily be addressed with fewer hours of data than can the acquisition of lower-frequency syntactic phenomena (e.g., the acquisition of passive constructions in English). Since many researchers are interested in aspects of syntax acquisition, this has led to the collection of dense corpora (several hours per week) for more effectively examining morphological and syntactic development (e.g., the Manchester Corpus – Theakston, Lieven, Pine and Rowland 2001). However, the context of recording (location, activating, interlocutors, time of day) may also be critical in terms of encouraging more utterances on the part of the child.

4. Longitudinal case studies

Much of the field of language acquisition has been conducted using cross-sectional experiments, where several children are tested at a given age to determine if they have mastered a certain grammatical structure. Thus, much of the field of acquisition provides us with a snap shot of children's grammatical competence at a particular point in time. This type of information is extremely valuable for providing norms of typical development that can be used by theoreticians and clinicians alike. However, it less clearly addresses one of the primary goals of the field, which is to understand how a given child's knowledge of language develops over time. Given enough data, longitudinal case studies can provide exactly the type of detailed, fine-grained information need to examine how children's grammars move from one stage of generalization to the next, providing a much-needed window into the language learning process. Such studies can also expose individual differences in the learning process (cf. Lieven this volume), providing critical information about the types of generalizations different language learners make. This in turn can inform our theories about how language is learned.

5. Early production data (ages 1–2)

The field of infant speech perception has pioneered several different methods for examining children's sensitivities to various types of phonetic, phonological, morphological, lexical and distributional information before the age of two. However, it is not yet clear what the relationship is between perception and production. Recent research on early comprehension, and children's ability to process lexical and morpho-syntactic information, begins to provide a better understanding of what children 'know' about language, and how they can begin to put this to use in both language processing (e.g., Lew-Williams and Fernald 2007). However, it is extremely difficult to conduct elicited production studies with children much below the age of 2 (though see Kehoe and Stoel Gammon (2001) for success at 1;6 years). For those children who begin to produce their first words by 11 months, the second year of life provides an extremely rich arena for exploring aspects of both phonological and morphological development. Longitudinal spontaneous production corpora during this time provide rich source of information regarding language development during this period (Demuth, Culbertson and Alter 2006; Demuth and Tremblay 2007; Fikkert 1994; Levelt, Schiller and Levelt 2000).

6. Nature of the input and learnability issues

Much of the research on language acquisition has been conducted in a context that is oblivious to what language learners actually hear. This has often proved problematic for language learning theories, which assume that the target grammar for the child is the full adult model. However, recent research suggests that the model to be learned is actually quite close to that of everyday speech directed toward the child. If so, this means that we need a much more complete model/description of child-directed speech at all levels of structure. Only then can we more effectively begin to understand the nature of the learning problem. Information about the frequency of occurrence and distribution of different phonological, lexical, morphological and syntactic phenomena is therefore needed to inform the design of our experiments and the interpretation of the behavioural results. For example, Ravid, Dressler, Nir-Sagev, Korecky-Kröll, Soumann, Rehfeldt, Laaha, Bertl, Basbøll, and Gillis (this volume) show that, across languages, plurals account for a small percentage of the total nouns children hear, and that the frequency distribution of morphological marking of plurals is the same as that found in early child speech. This is consistent with other findings in the field. For example Demuth (1989) suggests that the early acquisition of passives in Sesotho (as compared to English) is due to the much higher use of passives in Sesotho everyday speech. Once again, corpora provide a means for evaluating these issues, and help to explain the behavioural results found.

Information about the nature of the input learners hear is also important for designing models of how language learning might proceed. Monaghan and Christensen (this volume) explore what types of distributional information and phonological properties might be useful for clustering together certain natural classes of words. Other models take a more probabilistic, Bayesian approach to morphological segmentation, exploring the contributions of learning across types versus tokens (e.g., Goldwater 2006). Corpora of child-directed speech therefore play an important role in helping to explore not only the nature of the input, but also how learners can use this input in constructing their earlier grammars.

7. Discourse context and the structure of language

Information about the input also provides the context needed for exploring the acquisition of discourse-dependent aspects of language. For example, Allen, Skarabela and Hughes (this volume) use both video and audio information to examine the role of discourse context in licensing null objects. Thus, although much acquisition research often focuses on words or sentences, learners must be aware of the larger discourse context to be able to use and interpret both overt and null pronouns/objects in an appropriate fashion. This is critical for our understanding of how children learn the argument structure of verbs.

Recent corpus research on the argument structure of Sesotho verbs discovered that null objects are permitted in that language as well, even though this was not mentioned in any grammars (Demuth, Machobane, Moloi and Odato 2005). Since linguists often elicit grammaticality judgments at the level of the sentence, such discourse related issues are often overlooked. Thus, corpora may be especially useful for exploring discourse-related aspects of the syntax of lesser-studied languages, again providing the background needed for a full assessment of language learning issues.

8. Interactions between corpus and experimental studies

Corpora can also provide a wealth of pilot and subsequent data for designing and interpreting experimental results. For example, corpus analysis revealed that certain double object applicative constructions never occurred in 98 hours of adult and child speech in the Demuth Sesotho Corpus (Demuth 1992). Experiments were then needed to determine when Sesotho-speaking children learned that the animate object must be immediately ordered after the verb, rather than the benefactive argument, as in other Bantu languages (Demuth *et al.* 2006). Since there is no MacArthur CDI for Sesotho, the corpus analysis was extremely useful for identifying the high-frequency verbs which Sesotho-speaking 2-year-olds should be know. Further analysis showed that the worst experimental performance occurred on the highest-frequency verbs. This suggested that children expected these verbs to occur in their high-frequency syntactic frame (i.e., with one of the objects realized as a preverbal clitic rather than a lexical object). This suggests that certain high-frequency verbs may 'prime' high-frequency frames (Bock and Loebell 1990).

In another corpus study, Song and Demuth (2005) found that some children exhibit phonotactic complexity effects on the production of 3rd person singular morphemes. This provided the impetus for further cross-sectional experimental study, where an interaction was found between phonotactic complexity and position within the utterance. This in turn is prompting a return to the corpus to examine possible positional effects. Thus, information from experiments and corpora can often exist in a symbiotic relationship, each providing a piece of the evidence needed for understanding the factors that influence how language is acquired.

9. Areas ripe for further corpus research

Many early corpora contain data from children who are productively using language, often from the age of 2 onwards (e.g., Brown 1973). The focus of such studies has typically been morphological and syntactic development, where the data were 'orthographically' transcribed. As a result, most of the language acquisition studies that have used corpora have explored (morpho) syntactic issues. Less corpus research has focused on earlier aspects of phonological and morpho-phonological development. However, this is beginning to change with the increasing availability of longitudinal, phonetically transcribed corpora and the tools needed to exploit them (see Demuth (in press) for review). Many of these corpora are also linked to acoustic files, providing the means for conducting acoustic analysis of children's early speech productions (Song and Demuth submitted). In addition, many of these corpora contain data on child-directed speech, providing much-needed information about the early input children hear. Importantly, many of these new corpora come from a variety of languages, providing a critically-needed cross-linguistic perspective on the input children hear, and how this influences the realization of their early speech productions (see Demuth (2006) for review). Ultimately, this type of investigation should lead to developing a model of early language production, which may help account for some of the variability in children's early speech.

10. Limitations of corpus research

As discussed above, longitudinal language acquisition corpora provide a rich source of information for examining phonological, morphological, lexical and syntactic development over time. As with any method, however, there are limitations on what it can tell us about the development of linguistic representations. For example, many of the corpora gathered to date contain information on only a few children. Given that there is also a large amount of individual variation, data from more children are needed in order to provide a robust picture of how language develops over time, even for English, and for the adult input as well. In addition, there is a need for denser corpora than that usually collected, with several hours of speech collected at certain points in time. Even with optimally dense data, with several children, it is difficult to know what the frequency of certain lexical items is for a given child. Furthermore, the contexts in which these appear may be highly variable, making it difficult to control for possible context effects (e.g. position within the sentence, prosodic factors). Even with ideal corpora, it may be necessary to complement these studies with experiments, where novel words and/or carefully controlled contexts can be used.

Finally, corpus studies may overestimate or underestimate children's grammatical knowledge of a certain form. It has long been observed that children's perceptual abilities are often in advance of production abilities, and this is typically the case with comprehension as well. However, full comprehension and/or knowledge of a particular morphological or syntactic construction may take years to reach adult-like competence. For example, Demuth *et al.* (2006) found that, although 4-year-olds were above chance in placing the animate object immediately after the verb in double object applicative constructions, 8-year-olds still performed significantly worse than adults. Only by 12 years did Sesotho-speaking children show adult-like word-order performance in experiments. Since these constructions are relatively rare in everyday speech, such findings would have been almost impossible to find in corpus analysis.

11. Converging evidence from corpus and experimental studies

As discussed above, the use of corpora for addressing questions of how language is learned has certain limitations. However, experiments are also limited in what they can tell us, and experimental artefacts abound – especially when experiments are designed with little understanding of what children actually hear, and the frequency/priming biases they may have. Thus, the field can greatly benefit from a research paradigm that draws on converging evidence from multiple sources of information, including both corpora studies and experimental results. Several laboratories are now beginning to take this approach, with students trained in both corpus analysis and experimental techniques. With the growing availability of new corpora, and the tools needed to exploit them, the field of language acquisition is now better prepared to probe the processes of language acquisition more effectively than every before.