J. Child Lang. **41** (S1) (2014), 18–25. © Cambridge University Press 2014
doi:10.1017/S0305000914000191

Data and beyond*

PAUL FLETCHER
University College Cork

ABSTRACT

Perusing early issues of this Journal from nearly forty years ago, we find many topics on the contents lists that could appear in a current issue. But the resources available to contributors for addressing questions of interest have changed dramatically, not least in the public availability of data. Here we look briefly at two assets which have accrued to the field in the intervening decades, and at the contributions they make. CHILDES provides a level playing field on which debates about the import of language sample data, especially longitudinal data, can be played out. The MBCDI and its offspring encourage the collection of data from large samples of individuals and so promote the exploration of factors that may explain individual differences. Both have been major contributions to the field.

The first volume of the Journal that I was concerned with – volume two – appeared in 1975. A glance at the contents list of the two issues published in that year reveals articles dealing with topics in grammatical and lexical development which remain of contemporary interest. However, the circumstances for researchers then were very different. Looking back on that time I can vouch for the fact that the data on which those articles depended was hard-won, especially if it derived from language samples. Forty years ago, aside from the samples an individual researcher could collect using a relatively cumbersome reel-to-reel tape-recorder, which then had to be typed up and analyzed by hand, there were no generally available extensive language samples to work on. This is an obvious limitation for individual researchers, but it is also a serious constraint on the field: if there is no readily accessible databank there is no common source of material for scholars to challenge analyses, attempt replications, and generally argue over. So while it may seem an obvious choice, especially for those not faced with the early practical difficulties of studying child language, my first vote for

[*] Address for correspondence: Paul Fletcher, Department of Speech and Hearing Sciences, University College Cork, Brookfield Health Sciences Centre, College Road, Cork, Ireland. e-mail: p.fletcher@ucc.ie

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

an innovation of considerable influence in the field is the Child Language Data Exchange System (CHILDES; MacWhinney, 2000, 2013, 2014). Two reasons: one, it democratizes the field – the data and the tools for analyzing it are available on an equal basis to graduate students and senior researchers alike. And second, its availability can amplify our understanding of development – there are gains that go beyond the provision of data, as I hope is apparent from a brief (and inevitably selective) overview of studies of *wh*-interrogatives.

1. CHILDES AND THE DEVELOPMENT OF INTERROGATIVES

Among the modalities of the simple sentence that Brown (1973) identified as characteristic of his Stage III of development were interrogatives. An early article that focused on the development of one type of these, *wh*-questions, was Labov and Labov (1978). This account of their daughter Jessie's learning of these forms over an 18-month period (3;3–4;9) includes a graph showing percentage inversion rates over this period – the term 'inversion' reflecting the theoretical position that interrogatives were derived from declaratives via the movement of the relevant auxiliary verb from a post-subject position. The graph showed a gentle upward trajectory over this period, from a 10% to about 80% inversion rate. (Which is what we would expect for just about any measure of grammatical development in this age-range). A second graph showed the breakdown of percentage inversion rates across different *wh*-words. A different picture emerges from this graph, depending on the *wh*-word. For *how, where*, and *what* questions, an 80% inversion rate applies by about 3;9. But *why*-questions seem to remain stubbornly resistant to pre-subject auxiliary placement until about 4;6, when they suddenly start to rise. Because the data is private to the authors of the study, we can only speculate on the reasons for this difference. Fletcher (1985) suggests that the interpretation of forms such as *what's* and *where's*, with contracted auxiliary/copula, as inverted elements, may have inflated the inversion values for these *wh*-words. But until the existence of a suitable body of accessible data it was not possible to take the issue further. Fast forward to volume 27 of the Journal, in which Rowland and Pine (2000) take advantage of the availability of 55 hours of data in CHILDES from Adam, one of Brown's subjects, to look at the issue of inversion in his development of *wh*-questions. The contributions of this study, I would argue, are three-fold: descriptive, theoretical, and practical.

First, access to a database of this size, and the potential that CLAN affords for automatic analysis, forces a realistic perspective on the sheer complexity of the learning task facing the child, simply by describing the number of combinations of *wh*-word + auxiliary occurring in the data. The possible combinations of just four of the *wh*-words with twenty non-negative

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

uncontracted copula/auxiliary forms (*do does did am is are was were have has had* plus nine modals) total eighty (and the figure increases if we include contracted forms). If we include negatives the number doubles. Rowland and Pine (2000, p. 172) tabulate forty-nine different combinations of *wh*-word + auxiliary in Adam's questions. This was at a point in his development when he was using auxiliaries in these structures in both pre-subject and post-subject position – producing examples like *how do you open that?* as well as *what he can ride in?* A wide range of auxiliary forms of *be, have*, and *do* are found in the forty-nine combinations – all of those listed above, aside from *were* – along with six modals (*can could may shall should will*). Negative forms also appear: *doesn't, don't, didn't, can't, couldn't*. But in only two of the forty-nine *wh*-word + auxiliary combinations – *how* + *can* and *why* + *is* – do we find the auxiliary appearing in both pre-subject and post-subject positions. In all other cases combinations occur in one position but not in another, e.g.:

Auxiliary pre-subject		*Auxiliary post-subject*
What are	14	0
Why + can't	0	10

Piecemeal development in relation to auxiliaries is also apparent in the elicitation studies of Ambridge, Rowland, Theakston, and Tomasello (2006), Theakston and Rowland (2009), and Rowland and Theakston (2009). The descriptive detail afforded by the analysis of large bodies of data through CHILDES, and the validation of hypotheses that arise from these studies via experimental approaches, lead to a different perspective on how the child is developing knowledge of syntax. The asymmetric character of *wh* + auxiliary forms suggests that an account which recognizes low-scope lexically specific knowledge of *wh*-word + auxiliary sequences may provide a truer perspective on this phase of syntactic development than the invocation of a movement rule. It may be that an accumulation of lexically specific sequences leads eventually to full recognition of a relationship between auxiliaries in pre- and post-subject position, but this could well be a late abstraction from accumulated examples or even a post-literacy realization.

The constructivist account, I would argue, can also contribute to the practical problem of how we think about intervention for children with language impairment. Fey and Loeb (2002; see also Fey & Finestack, 2009) report on a study which was intended to improve the control of auxiliary verbs by children with specific language impairment (SLI). The question was whether children with SLI who were modelled auxiliary verbs (*is* and *will* being primary targets) in initial (inverted) position in *yes/no* questions would be likely to develop these and other auxiliaries. Sixteen children with

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

SLI were involved in the research, and were subject to enrichment in the form of interrogative recasts of declarative sentences:

CHILD: Mummy flying
ADULT: Is Mummy flying

The children with SLI, who were on average aged 3;2, were not using auxiliaries at the time of recruitment into the study. The experimental group attended twenty sessions of enrichment over eight weeks, and during each session heard thirty recasts. There was a control group who were involved in a play condition with no systematic recasting in the same number of sessions over the same time period. There was no evidence at the end of the study that the children in the enrichment group had improved their control of auxiliaries *is* or *will*, or indeed any other auxiliary form. Fey and Finestack (2009, p. 525) put this down to the recasts (the adult response to the child in the form of a question, as in the example above) being too complex for the child to make use of. This may be the case, and there may also be other reasons for the lack of success of the intervention. But it is also possible that the underlying assumption of the study, that there is an intrinsic relationship between auxiliaries in pre-subject and post-subject position, is not valid, at least for children with language impairment at a certain stage of development. Perhaps an intervention strategy should be guided by a different take on typical development, and start from a constructivist position – one that assumes the child gradually accumulates lexically specific sequences rather than one that takes the relationship between pre-subject and post-subject auxiliaries for granted.

A constructivist view of *wh*-question development is not uncontested. Westergaard (2009), using data from Norwegian and the same data from English that Rowland and Pine (2000) analyze, argues that the acquisition of *wh*-questions must be the result of a rule-based process. While I would be more inclined to a constructivist view, my reason for seeing CHILDES as one of the most important developments in the field is not because the available data automatically favours one position over another, but because it provides an open data source which can act, among other things, as a level and public playing-field for the evaluation of competing explanatory accounts.

2. THE MACARTHUR-BATES COMMUNICATIVE DEVELOPMENT INVENTORY – MBCDI

The other milestone in child language studies which I would highlight is also a data source – the MBCDI (Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007), together with its many adaptations into languages other than English (http://www.sci.sdsu.edu/cdi/). Relying on parental response

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

to a checklist of vocabulary items and grammatical features was in the 1990s an entirely novel way of tracing language development between ages 0;8 and 2;6. By comparison with the work involved in generating and transcribing language samples, parent reports are an efficient way of collecting extensive data on large numbers of children relatively quickly, either via paper forms or, more recently, on-line (Kristoffersen *et al*., 2013). The CDIs have proved to be reliable and valid instruments for monitoring early child language, especially vocabulary. Inter-child variation in rate of language development had been apparent even in the data from Adam, Eve, and Sarah, the three children in Brown (1973, p. 57), and was confirmed in subsequent studies based on larger samples, e.g., Wells (1985), which involved 128 preschool children. But the increase in numbers of children in this project had to be set against a much more limited amount of data from each child. This is a trade-off that the CDI is not subject to. With parent report, much larger samples can be accumulated, samples that are appropriately representative of a population. This enables us to see the true scope of individual variation, and facilitates the exploration of factors relevant to inter-child differences.

Stokes and Klee (2009), looked at vocabulary recorded via a UK CDI (Klee & Harrison, 2001) from over 200 British children between 2;0 and 2;6 (mean age 26·84 months). The average number of words in the sample was 329, but the range of vocabulary items went from 15 to 666. This degree of variation in large samples fits with information from the CDI on children in the USA, and also with adaptations of this procedure in more than forty languages around the world. Similar vocabulary ranges for two-year-olds can be found for languages as typologically distinct as Irish, French, and Cantonese (O'Toole & Fletcher, 2010; Stokes, Kern & Dos Santos, 2012: Tardif, Fletcher, Zhang & Liang, 2008). Stokes and Klee (2009) explored the relevance of demographic, background, and processing variables to vocabulary scores, but were only able to identify one predictor which accounted for a sizeable proportion (36%) of the variance in vocabulary attainment. One of the processing tasks that the children underwent was a non-word repetition (NWR) procedure. It was the children's performance on the NWR (containing novel words from one to three syllables in length) which was the strongest predictor of vocabulary level. The child's ability to repeat novel words has been interpreted as indicative of the efficiency of the phonological loop, a component of short-term memory that allows the child to hold on to auditory material briefly before transferring it to longer-term memory. So this finding links vocabulary size in two-year-olds to a component of the child's learning mechanism. A deficiency in the phonological loop has been identified in children with specific language impairment (e.g., Gathercole, 2006).

Stokes (2010) has further exploited the data from this large cohort of two-year-olds by examining the effects of the neighbourhood density

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

(ND) and word frequency (WF) of the items in the parent report on expressive vocabulary size. She found that these factors together accounted for a remarkable 61% of the variance (ND 47% and WF 14%) for the sample overall. But children with low vocabulary scores (below −1 SD of the distribution) scored significantly higher on neighbourhood density and lower on word frequency than children with high vocabulary scores. A very similar picture is seen in a comparable study on French two-year-olds (Stokes *et al*., 2012). ND and WF together predict 62% of the variance in vocabulary size in a sample of over 200 French-speaking children. And the children at the low end of the French vocabulary continuum also prefer words with higher ND and lower WF values. It seems that the processes the low vocabulary children in either language are employing, in extracting and storing words from the ambient language, are restricting their development of expressive vocabulary.

These studies demonstrate the potential of CDI data. It can be mined in a way that sheds light on the learning process. In the studies by Stokes and colleagues it is similarities in learning style cross-linguistically that beg for an explanation. We can also cite examples of differences that emerge from CDIs across languages, which need to be accounted for and can lead to promising proposals. Bleses *et al*. (2008) compare the Danish CDI results to those from a number of other languages. A comparison of vocabulary comprehension shows that, for Danish children,from age 1;0 on their score is the lowest of all the languages examined. This is attributed to the characteristics of Danish phonology, which involves vowel reduction and obstruent vocalization processes, leading to potential word segmentation problems for the learner. It is striking that children learning a closely related language, Swedish, where syllable boundaries are more clearly identifiable, show a faster rate of comprehension vocabulary development from age 1;0 on than their Danish counterparts.

Tardif, Fletcher, Liang, and Kaciroti (2009) also exploit differences in CDIs between closely related languages to illuminate factors relevant to acquisition, this time in relation to grammar. Separate CDIs were normed in Hong Kong for Cantonese, and in Beijing for Mandarin. Although mutually unintelligible, these languages are closely related. Nevertheless, significant differences were found between the two populations in the speed at which they acquired vocabulary. For children in Beijing at two years of age, the median vocabulary size was 657, while for children in Hong Kong at the same age, it was 322. Some of this discrepancy seems to be due to the single child status of children in Beijing, and also to their monolingual language environment. This is, however, not the full story, which may well, as with the Danish/Swedish comparison, involve phonological factors (Tardif *et al*., 2009, pp. 1139–1140). The interest here is in a specific grammatical category common to the two languages, that of classifiers.

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

If we simply look at number of classifiers used by age, we find the overall difference in vocabulary score reflected in this value: two-year-old Mandarin speakers – and indeed children at each month up to 2;6 – are reported to use significantly more classifiers than their Cantonese-speaking counterparts. But if we equate children by vocabulary size, this difference reverses, and the Cantonese-speaking children have the advantage. This can be shown for the grammatical category of classifiers overall, and for individual classifier types. So from a vocabulary size of 200 words on, Cantonese-speaking children have the advantage over their Beijing counterparts. What we see revealed here, it can be argued, is the attentiveness young children pay to subtle features of the input. In Cantonese, classifiers are obligatory in more grammatical contexts than in Mandarin, and a wider range of semantically specific classifiers (as opposed to the general classifier, which exists in both languages) tends to be used in the speech of Cantonese adults. We may surmise that it is these input factors that encourage Cantonese-speaking children to tune into the precise functions and extensions of classifiers more readily than their Mandarin-speaking counterparts.

CHILDES and the CDI then both contribute in significant ways to the availability of data in the field, within and across languages. CHILDES is an open data source, primarily for language samples. CLEX (Jørgensen, Dale, Bleses & Fenson, 2010) continues to extend the range of languages on which CDI data is available, with ten now represented. CLEX is a web-site available to all, that allows researchers to query all aspects of the CDI dataset or sets they are interested in, construct tables and graphs from the data, and make comparisons across languages. The benefits CHILDES and the CDIs bring to the field of child language in terms of data provision are substantial, and represent a dramatic change to the situation at the time of this Journal's foundation. But as I hope I have sketched out above, the contribution they make to our profiling of the course of language development is considerable, and their impact on the understanding of the process of language learning is not negligible.

REFERENCES

Ambridge, B., Rowland, C., Theakston, A. & Tomasello, M. (2006). Comparing different accounts of inversion errors in children's non-subject wh-questions: 'What can experimental data tell us?' *Journal of Child Language* **33**, 519–57.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: a CDI-based comparison. *Journal of Child Language* **35**, 619–50.

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

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S. & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual* (2nd ed.). Baltimore: Brookes Publishing.

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press

Fey, M. & Finestack, L. (2009). Research and development in child language intervention: a five-phase model. In R. G. Schwartz (ed.), *Handbook of child language disorders*, 513–31. New York/Hove: Psychology Press.
Fey, M. & Loeb, D. F. (2002). An evaluation of the facilitative effects of yes-no questions on the acquisition of auxiliary verbs. *Journal of Speech, Language, and Hearing Research* **45**, 160–74.
Fletcher, P. (1985). *A child's learning of English*. Oxford: Basil Blackwell.
Gathercole, S. E. (2006). Nonword repetition and word learning: the nature of the relationship. *Applied Psycholinguistics* **27**, 513–43.
Jørgensen, R. N., Dale, P., Bleses, D. & Fenson, L. (2010). CLEX: a cross-linguistic lexical norms database. *Journal of Child Language* **37**, 419–28.
Klee, T. & Harrison, C. (2001). CDI Words and Sentences: validity and preliminary norms for British English. Paper presented at Child Language Seminar, University of Hertfordshire, July.
Kristoffersen, K., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E. & Henriksen, L. (2013). The use of the Internet in collecting CDI data – an example from Norway. *Journal of Child Language* **40**, 567–85.
Labov, W. & Labov, T. (1978). Learning the syntax of questions. In R. Campbell & P. Smith (eds), *Recent advances in the psychology of language: language development and mother–child interaction*, 1–44. London/New York: Plenum Press.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
MacWhinney, B. (2013). *The CHILDES Project: Tools for Analyzing Talk – Electronic Edition. Part 1: the CHAT transcription format*. Online: <http://childes.psy.cmu.edu/manuals/chat.pdf> (last accessed 17 January 2014).
MacWhinney, B. (2014). *The CHILDES Project: Tools for Analyzing Talk – Electronic Edition. Part 2: the CLAN programs*. Online: <http://childes.psy.cmu.edu/manuals/clan.pdf> (last accessed 17 January 2014).
O'Toole, C. & Fletcher, P. (2010). Validity of a parent report instrument for Irish-speaking toddlers. *First Language* **30**, 199–217.
Rowland, C. F. & Pine, J. (2000). Subject–auxiliary inversion errors and wh-question acquisition: 'What do children know?' *Journal of Child Language* **27**, 157–81.
Rowland, C. F. & Theakston, A. L. (2009). The acquisition of auxiliary syntax: a longitudinal elicitation study. Part 2: the modals and auxiliary DO. *Journal of Speech, Language, and Hearing Research* **52**, 1471–92.
Stokes, S. (2010). Neighbourhood density and word frequency in toddlers. *Journal of Speech, Language, and Hearing Research* **53**, 670–83.
Stokes, S., Kern, S. & Dos Santos, C. (2012). Extended statistical learning as an account for slow vocabulary growth. *Journal of Child Language* **39**, 105–29.
Stokes, S. & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry* **50**, 498–505.
Tardif, T., Fletcher, P., Liang, W.-L. & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. *Journal of Child Language* **36**, 1115–44.
Tardif, T., Fletcher, P., Zhang, Z. X. & Liang, W.-L. (2008). *Chinese Communicative Development Inventories: user's guide and manual*. Beijing: Peking University Medical Press.
Theakston, A. L. & Rowland, C. F. (2009). The acquisition of auxiliary syntax: a longitudinal elicitation study. Part 1: auxiliary BE. *Journal of Speech, Language, and Hearing Research* **52**, 1449–70.
Wells, C. G. (1985). *Language development in the pre-school years*. Cambridge: Cambridge University Press.
Westergaard, M. (2009). Usage-based vs. rule-based learning: the acquisition of word order in wh-questions in English and Norwegian. *Journal of Child Language* **36**, 1023–51.

https://doi.org/10.1017/S0305000914000191 Published online by Cambridge University Press