Interlanguage corpora and second language acquisition research

Florence Myles

Second Language Research 21,4 (2005); pp. 373–391

Review article

Interlanguage corpora and second language acquisition research

Florence Myles *University of Newcastle*

This article presents a selective review of the work carried out recently in second language acquisition (SLA) research which makes use of oral learner corpora and computer technologies. In the first part, the reasons why the field of SLA needs corpora for addressing current theoretical issues are briefly reviewed. In the second part, recent literature on corpora and SLA is presented, as well as corpora currently available. The final part of the article demonstrates the way in which computerized methodologies can be used, by presenting a case study of a project whose aim was to construct a database of French Learner Oral Corpora, and by illustrating how the CHILDES tools have assisted in addressing a specific research agenda.

Aston, G., Bernardini, S. and **Stewart, D.** 2004: *Corpora and language learners*. Amsterdam: John Benjamins. £87. ISBN 1588115747.

Granger, S., Hung, J. and **Petch-Tyson, S.** 2002: *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins. £51. ISBN 1588112942.

Hunston, S. 2002: *Corpora in applied linguistics*. Cambridge: Cambridge University Press. £17. ISBN 052180583X.

Pennington, M. and **Stevens, V.,** editors, 1992: *Computers in applied linguistics*. Clevedon: Multilingual Matters. ISBN 185359119X.

I Introduction

This article presents a selective review of the work carried out recently in second language acquisition (SLA) research which makes use of oral learner corpora and computer technologies. The impetus for

Address for correspondence: Florence Myles, School of Modern Languages, Old Library Building, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK; email: Florence.Myles@ncl.ac.uk

 10.1191/0267658305sr252oa

this article partly follows from Rutherford and Thomas' (2001: 211) article advocating the use of the CHILDES tools for SLA research, which concludes:

> Whereas the potential value of the CHILDES mechanisms for research in post-pubescent language acquisition has always been with us, it has (inexplicably) not been demonstrated until now. It is clear that the rich CHILDES resources refined over considerable time in CHAT and CLAN can eminently serve a wide array of research pursuits in adult language acquisition. We therefore have every reason to believe that the impact CHILDES has had upon the shape of first language acquisition research will in time be replicated for second language acquisition research.

One aim of the article is to assess if things have developed since then, and we report on recent efforts to use CHILDES tools in SLA research. The scope of this article is broader, however. In the first part, I briefly review the reasons why the field of SLA needs to make use of corpora for addressing current theoretical issues. In the second part, I present a brief review of recent literature on corpora and SLA (both written and oral), as well as of the corpora currently available. I concentrate on four texts that are representative of the field. One of these was published in 1992 and will serve as a benchmark to assess the progress the field has made since then. The final part of the article concentrates on demonstrating the way in which computerized methodologies can be used, by presenting a case study of a project whose aim was to construct a database of French Learner Oral Corpora, and by illustrating how the CHILDES tools have assisted in addressing a specific research question.

II Need for corpora in SLA

SLA research crucially depends on good datasets to work from. Researchers aim to build models of the underlying mental representations and developmental processes which shape and constrain second language (L2) productions. The language produced by learners, whether spontaneously or through various elicitation procedures, remains a central source of evidence for these mental processes, and the success of SLA research therefore relies on having access to good quality data. If we are to make generalizations about learner development, we need to be able to capture the various stages that learners go through. For this we either need longitudinal data of a number of learners over a lengthy period of time (as the acquisition of a L2 requires a

considerable amount of time), or we need very large cross-sectional datasets, so that the number of learners in each well-defined stage is big enough for us to be confident that the results of the analysis are generalizable (or to capture what is variable in language development across learners for that matter).

In a field that is traditionally not very well funded, the cost of collecting such corpora is often prohibitive, and datasets tend to remain small and limited in scope (e.g., containing very focused data aiming to answer a specific research question). Moreover, these datasets are usually not accessible to the research community as a whole, for a number of reasons. However, it is now extremely easy to store and manage data digitally on any PC, and just as easy to make it available to other researchers on the web. The field of first language (L1) acquisition research has been doing this for some considerable time, and is consequently much more sophisticated than we are in its use of computerized methodologies.

For the purposes of fundamental SLA research, oral data is an important window into learners' underlying mental grammars, and may be relatively freer of metalinguistic interference than written data, which is complicated by additional layers of learnt knowledge and monitoring processes. This is not to say that written data is not highly suitable for some research agendas, for example in the fields of second language writing or literacy development, as well as to assist in the design of various pedagogical tools such as L2 dictionaries or L2 reference grammars, but it is less suitable for other kinds of research. For this reason, we only briefly review written corpora here, concentrating on the collection and analysis of oral corpora.

L2 written corpora, however, are much more readily available than oral ones. Part of the reason, of course, is that they are much easier and cheaper to collect, and therefore do not require as large an investment as oral corpora. Such corpora often consist of learner essays, which are usually part of the assessment the learners have to produce as part of their course, already in electronic form. The pioneering work done by Sylvianne Granger and her team in constructing the International Corpus of Learner English (ICLE; Granger, 1998), or the extensive Longman Learners Corpus, are good examples of this. Oral corpora, on the other hand, remain rare.

Why is it imperative at the present time for SLA research to make use of new technologies, rather than to carry on handling data manually? The reasons are numerous. In its exploratory phase, the SLA research community has learnt a lot from detailed case studies or studies involving small numbers of learners. This has served its hypothesis-building endeavours very well, in the same way as in L1 acquisition, a somewhat older field of research, detailed descriptive case studies of individual children formed the starting point for hypothesis generating. Time has now come, though, to test some of the current hypotheses on larger and better constructed datasets, as has happened in L1 acquisition. Not only do we need large datasets in order to be able to generalize our findings, but some of the structures which are crucial for informing current debates are rarely found in learner data. They therefore either must be elicited specifically, or large datasets are needed in order to maximize their chance of being present. It is therefore now crucial that, as a field, we generate and make available to the research community a good range of well-constructed datasets.

Another reason for advocating the creation of electronic databases of oral corpora is that they become amenable to computerized analysis. As I demonstrate later on in this article, software tools allow us to manipulate the data in a number of ways, from the automatic morphosyntactic tagging of large datasets, to the carrying out of complex searches on batches of files. Having digitized sound files also allows us to use dedicated speech analysis software in order to address, e.g., research questions in language processing or phonology. An added, if incidental, benefit of sharing corpora across the research community, is that we will have to agree on procedures for so doing, from transcription conventions, to ethical codes of practice, for example, and thus we will become more rigorous in our practices. At present, for example, there are nearly as many transcription systems as there are researchers in SLA, which makes it difficult to share data, not to mention the fact that some coding systems might not meet expected standards.

III Corpora and SLA

In this section I present a brief review of four key texts dealing specifically with corpora in Applied Linguistics / SLA. The aim is to present

a critical evaluation of how corpora have been used to enhance (or not, as the case may be) research in SLA.

I start with a book published in 1992, *Computers in Applied Linguistics*, an edited collection of papers presented at an international conference on this topic in 1987 (Pennington and Stevens, 1992). This enables me to assess any significant differences some 10–15 years later. A whole section of this book is dedicated to 'Research on applications of computers in second language acquisition', containing 4 chapters (out of the total of 13), the other two sections being devoted to 'Frameworks for computer-assisted language learning in the 1990s' (3 chapters) and 'Analysis tools for a new generation of language applications' (5 chapters, one on the computational analysis of language acquisition data). The volume overall is primarily about the use of computers in the classroom and on how to evaluate its efficiency. Even in the SLA section, only one chapter (by Catherine Doughty) reports an SLA study making use of computers both in the study design (in order to control the instructional context) and in order to assist with data analysis in an interesting if somewhat limited way. The chapter on 'Computational analysis of language acquisition data' (by Manfred Pienemann) presents the first software designed specifically for the analysis of L2 learner data, outlining promising developments in this area. This software, however, remained little used by the research community, and is no longer available. Overall, the focus of this book is quite explicitly on CALL, which of course was developing at a very fast rate at the time, and the application of computers to SLA research remained very limited.

Hunston's *Corpora in Applied Linguistics* was published in 2002, and explicitly focuses on English corpus linguistics in general, and on its application to teaching. In the whole book, only 6 pages are devoted to learner corpora and, even then, exclusively to advanced learner written English (the ICLE corpus). Although the ICLE corpus (Granger, 1998) is undoubtedly the best established learner corpus available to date, it remains limited in a number of ways (cross-sectional, written data, advanced learners only), and the view that emerges from this chapter is somewhat narrow in terms of the possibilities learner corpora can offer in this area. Given the almost exclusive focus of the book on the relevance of corpora to language teaching, no mention is made

anywhere of the use of corpora in language acquisition research more generally (e.g., the huge impact of the CHILDES database in L1 acquisition research), and the immense possibilities offered for the analysis of many different aspects of language acquisition are not alluded to. This volume, if thin on SLA applications, nonetheless contains an excellent description of the different kinds of tools that have been used in corpus research, and is well worth reading by novices to this field.

Granger, Hung and Petch-Tyson's volume (2002), after a critical evaluation of the field, is split into 2 sections: the first section contains three studies illustrating the use of corpus-based approaches to the analysis of learner language. The second section, comprising 5 chapters, is about the pedagogical relevance of learner corpus work, and need not concern us here.

Granger's excellent introductory chapter 'A bird's eye view of learner corpus research' gives a very good overview of learner corpus research to date, as well as very sound advice on how to design and exploit corpora. She stresses two main approaches to the linguistic analysis of L2 corpora: Contrastive Interlanguage Analysis (CIA), which compares learner productions to either native corpora or corpora of other learners (e.g., different levels, different L1s), or both, and computer-aided error analysis. The latter is undoubtedly of interest to teachers, but probably less so to SLA researchers who have moved away from looking at learner language in terms of its deviations, and have focused instead on analysing the linguistic system in its own terms. Various software tools available to researchers are then presented, from text retrieval to annotation (Part of Speech tagging and error tagging). Somewhat surprisingly, the CHILDES suite of software is not even given a mention here, in spite of its success in supporting a similar research agenda in L1 acquisition, and the fact SLA researchers have appealed for its use (Rutherford and Thomas, 2001; Ellis, 2002), or used it themselves, e.g., in the very volume Granger is introducing here (Housen, 2002). Also, one of the main research agendas in SLA, namely the documentation and explanation of learner development over time, for which longitudinal oral corpora are crucial, is not given a mention either.

The first two chapters presented in the volume are typical in many ways of current L2 research using corpora. Altenberg investigates the

use of *make* in advanced learners of English (French and Swedish L1), using the ICLE corpus, comparing their written production to a corpus of native US English (Altenberg, 2002). He suggests that transfer is taking place, and this is explained in terms of prototypicality.

The second study (Aijmer, 2002) compares the use of key modal words in native English and in advanced L2 learners (L1 Swedish; some L1 German and French). It uses the ICLE corpus too, and makes interesting suggestions about possible pedagogical implications.

Both these chapters are good examples of the kind of studies that corpus linguistics have made much easier to carry out: they rely on large written corpora and focus on the use of discrete items in different cross-sectional populations (typically one or more L1s compared to native use). Concordancers and the availability of written corpora have made this kind of investigation very easy to carry out.

Housen's chapter in this volume is more ambitious (Housen, 2002). This is the only study that is about the acquisition process itself, and that relies on oral data produced by learners at different stages of development, combining a longitudinal dataset of 6 learners followed over 3 years to a cross-sectional dataset of 40 learners. The corpus used is the *Corpus of Young Learner Interlanguage* (CYLIL), consisting of English L2 data elicited from European School pupils from different L1 backgrounds (Dutch, French, Greek and Italian; 500,000 words). A control group of 8 native English-speaking children performing the same tasks was also used. The whole corpus has been transcribed, segmented, coded and annotated using the CHILDES range of software tools, enabling a very sophisticated analysis of the development of the verb system in these learners to be carried out. This study exemplifies very well the tremendous possibilities offered by the use of well developed software and good quality datasets. This is, however, atypical of the studies presented in this volume.

In the Aston *et al*. volume, *Corpora and Language Learners* (2004), a whole section is devoted to learner corpora (or corpora by learners, as they call them), totalling 6 out of 15 altogether, one of which is on the L1 acquisition of English, and which therefore will not be included in this discussion. This leaves 5 chapters. The range of L1s (Japanese, Swedish, Polish, German and Chinese) has grown when compared to other volumes. The target language, however, remains English in all

cases. All the corpora used are written, except for one study which complements a large written corpus with a small spoken one. Two of the studies also include an L1 corpus for comparison purposes, and all studies except one use a native corpus of the target language, English. The British National Corpus (BNC) sampler is used for this purpose in two cases, ELT books in one case, and government reports in another case. All the learners investigated are advanced (although it is difficult to determine from the information given in the study by Tono), and a wide range of areas are investigated, as follows:

- subcategorization frames of verbs (Tono);
- L1 syntactic transfer, as evidenced by POS (Part of Speech) sequences (Borin and Prütz);
- demonstratives (Lenko-Szymanska);
- support verb constructions (Nesselhauf);
- lexis (Flowerdew).

These studies employ of a range of software in order to assist the analysis. Two studies made use of the concordancer WordSmith (Scott, 1999), in one case complemented by APPRAISAL. Two studies made use of POS tagging or syntactic parsing, complemented in one case by pattern matching, ChaSen and the Comlex Lexicon. So, a wider range of L1s, of target structures, and of computerized tools are used in the studies presented in this volume, when compared with the earlier volumes reviewed here. However, the focus remains overwhelmingly on written advanced English L2 productions, with the limitations this implies.

To conclude this section, although the reasons why L2 corpora seem to be almost exclusively written and produced by advanced learners (ease of collection, and more readily comparable to equivalent native corpora than corpora from beginners), it still remains puzzling why so little use has been made in SLA research of oral corpora, or of tools such as CHILDES. Out of all the work reviewed in this section, only Housen capitalizes on it. Most of the studies using corpora make little use of software other than concordancing, and remain for the most part rather unambitious in their use of new technology. They also often remain rather descriptive, documenting differences between learner and native language rather than attempting to explain them, and the developmental dimension is almost totally lacking. Corpus-based L2

studies are also often not sufficiently informed by SLA theory, and tend to assume that finding out differences in use between learners and native speakers will have direct pedagogical implications, which is of course not necessarily the case. Such research is useful nonetheless, as we need to have good descriptions of learner language in order to inform our understanding of what shapes its development, but it is now time that corpus linguists and SLA specialists work more closely together in order to advance both their agendas.

IV The FLLOC database: a case study

The last section in this article briefly presents a current research initiative, which has led to the construction of a database of oral French L2 corpora, available from the internet and comprising transcripts, sound files, and morphosyntactically tagged transcripts. This database will be used to demonstrate the possibilities offered by such technology. After a brief presentation of the software and database, I illustrate how the various tools available in CHILDES can assist in investigating specific research questions, making the results more generalizable and reliable than would be the case with a small sample of learners. For more details of the database and its methodology, see, e.g., Myles and Mitchell (2005).

1 The CHILDES system

In the early 1990s, a research team at the University of Southampton collected large amounts of data from classroom learners of French. The 'Progression Project' (Mitchell and Dickson, 1997) followed 60 children (12 to 14 year-olds) longitudinally for their first 27 months of learning French in the classroom, and includes some 650 transcripts, primarily of children engaged in 1:1 oral tasks with a researcher. In the context of a later project (Myles, 2002), a cross-sectional study of classroom learners at the next stages of development, the team was acutely aware that the extremely rich data from the Progression Project was not exploited to its full potential, as it was not readily available to other researchers for further study, and only a relatively small subset of the database had been analysed manually (Mitchell and Martin, 1997; Myles *et al*., 1998; Myles *et al*., 1999; Myles, 2003; 2005). The research team therefore decided to investigate the possibility of using

computerized methodologies in helping them manage, store, share and analyse the data. Two attempts were made in the 1990s to develop analysis software specifically to deal with L2 data: COALA developed by Pienemann (1992) and COMOLA by Jagtman and Bongaerts (1994). However, it seems these programs have been discontinued, and no POS tagging software is currently available that has been specifically designed for SLA data handling and analysis. By contrast, in the L1 acquisition field, software for the storage, management, sharing and analysis of L1 data (the CHILDES system), has been developed in an ongoing way since the early 1980s, and is now very widely used in L1 research. After investigating the various options, the team came to the conclusion that the CHILDES system was the most suitable option, with the proviso that some SLA-specific adaptations must be practicable. The reasons were many:

- It is a well developed and well supported system, used as standard within the L1 research community and constantly updated and refined.
- It seemed relatively flexible and capable of being adapted to specific needs.
- The policy of open access adopted by CHILDES makes accessibility and data sharing very straightforward.

CHILDES offers a suite of software options including word-based programs that can be used to carry out concordancing, frequency counts, etc. Most importantly, CHILDES also makes available POS taggers for a range of languages (currently 10 languages,[1] with 4 more in preparation). The POS taggers are relatively easy to modify according to specific criteria and research needs. In its latest version, the CHILDES system is also now XML compatible.

The CHILDES set of tools was originally conceived for L1 acquisition research, but has also been used for research into language disorders and by some L2 researchers (Paradis *et al*., 1998; Housen, 2002; Malvern and Richards, 2002). CHILDES tools have been used in well over 1300 published studies ranging from L1 acquisition to computational linguistics, language disorders, narrative structures, literacy development, phonological analyses and sociolinguistics

[1]Cantonese, Danish, Dutch, English, French, German, Hungarian, Italian, Japanese and Spanish.

(MacWhinney, 2000a; 2000b). All CHILDES tools are available free of charge on the internet (http://childes.psy.cmu.edu). CHILDES consists of three integrated components:

- A large and diversified database (TalkBank) consisting primarily of child speech recordings and transcriptions, but also including some language disorder data and bilingual data. It is a condition of using CHILDES tools that the data becomes part of the TalkBank database. There are increasing numbers of second language acquisition datasets available in TalkBank; for example:
  - The Ionin corpus (Russian immigrant children learning English; 22 participants aged 2;4 to 12;5).
  - The Reading corpus (34 oral GCSE examinations taken by 16-year-old English learners of French within the UK education context). The speech recordings and transcripts from these corpora can be downloaded from TalkBank for further study. An online browsing facility should soon be available to locate data for further in-depth analysis.
  - The LIDES corpus (The Language Interaction Data Exchange System) is the database arising from the Language Interaction in Plurilingual and Plurilectal Speakers Project (LIPPS). The researchers involved have used their own list of coding conventions based on the CHILDES conventions. For example, they specifically tag each word/morpheme to indicate its language (http://talkbank.org/data/LIDES/). The LIPPS research team have produced a coding manual which is available to other researchers.
  - Part of the FLLOC database (Progression project and Linguistic Development project, mentioned above) has now also been added to TalkBank, and the rest will follow soon.
- CHAT (Codes for the Human Analysis of Transcripts) are the transcription procedures that have been developed to be compatible with the analysis programmes.
- CLAN (Computerized Language Analysis) consists of about 40 core computer commands for carrying out searches and counts, along with a range of 'switches' that can be used to customize each command. This is a powerful and flexible software package that can carry out rapid and detailed analyses and is designed to recognize the tagging conventions of CHAT.

2 *Transcription and analysis*

a Tiers: The CHAT transcription system is organized in tiers: the main speaker tier, starting with *, contains the language actually produced. In addition to this main tier, a limitless number of 'dependent tiers' can be added on separate lines, which always start with %, and contain any coding of the data. Examples of such tiers are the *%err* tier (on which errors are coded), the *%mor* tier (containing morphosyntax tagging) a *%com* tier (commentary) and a *%pho tier* (phonological coding).

The transcription conventions are specified in the CHILDES manual available on line (http://childes.psy.cmu.edu) and must be adhered to for the CLAN programs to run successfully on the data. It is, however, possible to add various adaptations, e.g., for SLA specific purposes. For details of how this has been achieved in the context of the case study database, see Marsden *et al*. (2003) and Rule *et al*. (2003).

b %mor tier: A particularly useful tool in CLAN (MOR) generates morphosyntactic tagging of the main line automatically. Versions of MOR have been produced for a range of languages (10 at present); the MOR parser for French was developed by Christophe Parisse in 2001. Below is an excerpt from a transcript with an added MOR tier from the Linguistic Development Corpus (see database content below). The extract comes from the file of a year 10 learner (a 15-year-old; 39-months of classroom French) carrying out an elicitation task focusing on the use of negation. *29N is the speaker line for the learner and *ELD the speaker line for the researcher.

```
*29N:   mais il n' aime pas le musique.
%mor:   conj|mais pro:subj|il&MASC&_3S adv:neg|ne v|aimer-PRES&_3SV
        adv:neg|pas det|le&MASC&SING n|musique&_FEM .
*29N:   um il ne joue pas le basket # + /.
%mor:   co|um pro:subj|il&MASC&_3S adv:neg|ne v|jouer-PRES&_3SV
        adv:neg|pas det|le&MASC&SING n|basket&_MASC .
*29N:   +, et il aime le cola .
%mor:   conj|et pro:subj|il&MASC&_3S v|aimer-PRES&_3SV
        det|le&MASC&SING n|cola&_MASC .
*ELD:   mmm .
*29N:   il ne mange pas le glace .
%mor:   pro:subj|il&MASC&_3S adv:neg|ne v|manger-PRES&_3SV
        adv:neg|pas det/le&MASC&SING n|glace&_FEM .
```

The %mor line can be generated very quickly on large batches of files, and other CLAN commands then enable searches to be carried out directly for morphosyntactic strings on this output.

3 CLAN commands

The range of CLAN tools aiming to assist analysis includes standard concordancing facilities as well as various searches on the different tiers. Depending how the data has been coded, lexical, morphosyntactic, discourse and phonological analyses, amongst others, can be carried out. CLAN programmes such as FREQ, KWAL and COMBO facilitate analyses of the frequency and linguistic context of interlanguage features. FREQ creates a file that provides the lexical range and frequency in a given file or group of files. KWAL operates like a concordancer and will search the data for specified words and outputs these keywords in context. FREQPOS does a frequency analysis by sentence position and MLU calculates the mean length of utterance. COMBO searches for specific words, word sequences or combinations of lexical items on the main tier, as well as directly on any of the dependent tiers, for example, for morphosyntactic or 'error' codes. In addition, the results of one analysis can be 'piped' through another analysis, allowing multiple analyses.

4 Using CLAN to address specific research questions

This section illustrates the use of CHILDES to answer specific research questions, by demonstrating how searches can be carried out directly on the morphosyntactic output (the *%mor* line). For example, you can search at the touch of a button for all instances of, say, masculine determiners followed by a feminine noun. You can do that for a specific learner over time (in the case of longitudinal corpora), or for batches of learners who share a given characteristic (e.g., school year; or performing a specific task). For details of how to carry out such searches, see e.g., Marsden *et al*. (2003); Rule *et al*. (2003); Myles and Mitchell (2005).

A recent study (Rule and Marsden, in press) investigated the development of negatives in relation to finite and nonfinite verbs in the emerging grammars of early learners of French. Differences in negative placement in French and English are a result of the verb raising past the negative in French but not in English (strong vs. weak Infl): the French negative particle *pas* occurs after a finite lexical verb in French (1), but in English *not* cannot occur after a finite lexical verb (2) and the dummy auxiliary *do* needs to be inserted (3). The verb does not raise over negation if it is nonfinite (4) (for more details of verb raising in French, see Hawkins, 2001).

1) Jean ne regarde **pas** la television.
2) Jean (ne) watches **not** the television.
3) John does **not** watch television.
4) Jean regarde la télévision pour ne **pas** s'endormir
   John watches television to not sleep, i.e., in order not to fall asleep

Rule and Marsden therefore traced all occurrences of the following in the data:

- *pas* followed immediately by infinitive;
- *pas* followed immediately by verb in present;
- verb in present followed immediately by *pas*;
- infinitive followed immediately by *pas*.

This could be done extremely easily, across the whole corpus, using a COMBO directly on the %mor tier. For details, see Rule (2004); Rule and Marsden (in press). After this brief survey of the possibilities offered by the CHILDES system to assist L2 analysis, let us now turn to a brief description of the database which supported these analyses.

5 The database (www.flloc.soton.ac.uk)[2]

The French Language Learner Oral Corpora (FLLOC) website contains an electronic database freely available to the research community, in the form of digital sound files and transcripts formatted using CHILDES conventions. The database also comprises a search facility, which enables researchers to select the sound files and transcripts they wish to access, according to criteria such as the level of the learners, the elicitation task used, the sex of the participants, etc.

The corpora included in the database have all been digitized and edited to ensure anonymity, and the transcripts have been (re)formatted according to the CHILDES system. Details of the CHILDES tools are given, and additional transcription conventions are specified in the context of each project. Additionally, most of the transcripts have been tagged using the French MOR program, and the resulting files, i.e., transcripts that have been tagged morphosyntactically, have been made available also.

[2]This project was supported by grants from the Economic and Social Research Council (RES000220070) and the Arts and Humanities Research Board (RE-AN9657/APN15456). I wish to thank the research team involved in these projects: Rosamond Mitchell, Sarah Rule, Emma Marsden, Vladimir Mircevski.

The corpora included in the database come from diverse sources, but have all been donated for shared use by SLA researchers in the United Kingdom and in mainland Europe. There are 5 in total to date, representing instructed learners of L2 French from complete beginners to final year university undergraduates, undertaking a range of interactive and narrative tasks. The database currently includes 1375 transcripts and accompanying sound files, as well as tagged files in some cases, as follows:

- 60 learners in years 7, 8 and 9 in the UK context: beginners at outset of data collection; age 12–14; longitudinal over 27 months; range of 1:1 narrative and interactive tasks, e.g., story retelling, information gap, structured conversation, etc., 650 transcripts and sound files, approximately 10–15 minutes each;
- 20 learners in each of years 9, 10 and 11 in the UK: post-beginners; age 14–16; cross-sectional; four 1:1 tasks each, some repeated from above project; 240 transcripts/sound files, around 10–15 minutes each;
- 34 learners: post-beginners; age 16; UK GCSE oral examination; 26 native controls; 60 transcripts;
- 12 university undergraduate learners in the UK: intermediate to advanced; longitudinal; narrative and interactive tasks; 300 transcripts and sound files; approximately 5–10 minutes each;
- 125 Dutch learners of French: intermediate; narrative task; 18-year-olds; 125 transcripts.

Each corpus is accompanied by a project description, which includes details of the learners and the tasks used, any additional transcription conventions used, plus an overview of the files contained in the database, and how each one is organized. Files (sound files, transcripts and tagged transcripts) can be directly downloaded for use by bona fide researchers who sign up to an explicit users' code.

V Conclusions

What can we conclude about the progress made in the last decade or so in the use of corpora in SLA? Some progress has undoubtedly been made, and there are now some L2 corpora available. My concern is that

the kind of studies that are being undertaken are too closely dependent on what corpora are at hand, and what software tools are available. For reasons explored in Section I, the kind of corpora that are readily available are not necessarily those most suited to the investigation of SLA acquisition processes; they are nearly always written, cross-sectional and overwhelmingly from advanced learners of English. Although these corpora clearly have their place in the range of SLA studies the field needs to undertake – especially at the level of lexis, discourse and pragmatics at an advanced level – and are of interest to university teachers of such learners, the field needs to become much more ambitious in its use of new technologies, and in the kind of corpora it collects in order to address its current research agenda. For this purpose, we need good quality longitudinal oral corpora, in a range of different L1/L2 combinations, for the reasons outlined in Section II. The possibilities offered by the computerized analysis of corpora are considerable, as I hope to have demonstrated. SLA researchers, however, need to make sure that not only the corpora they collect but also the computerized tools they use are adapted to their research agendas, rather than the other way round, i.e., adapting their research questions to the corpora or the tools readily available. Some sophisticated tools can be used, and it is high time that the pioneering work of L1 acquisitionists in this area is emulated by L2 researchers.

References

Aijmer, K. 2002: Modality in advanced Swedish learners' written interlanguage. In Granger, S., Hung, J. and Petch-Tyson, S., editors, *Computer learner corpora, second language acquisition and foreign language learning.* Amsterdam: John Benjamins, 55–76.

Altenberg, B. 2002: Using bilingual corpus evidence in learner corpus research. In Granger, S., Hung, J. and Petch-Tyson, S., editors, *Computer learner corpora, second language acquisition and foreign language learning.* Amsterdam: John Benjamins, 37–54.

Borin, L. and **Priitz, K.** 2004: New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In Aston, G., Bernardini, S. and Stewart, D., editors, *Corpora and language learners.* Amsterdam: John Benjamins, 67–87.

Doughty, C. 1992: Computer applications in second language research: design, description, and discovery. In Pennington, M. and Stevens, V., editors, *Computers in applied linguistics.* Clevedon: Multilingual Matters, 127–54.

Ellis, N.C. 2002: Reflections on frequency effects in language processing. *Studies in Second Language Acquisition* 24, 297–339.

Flowerdew, L. 2004: The problem-solution pattern in apprentice vs. professional technical writing: an application of appraisal theory. In Aston, G., Bernardini, S. and Stewart, D., editors, *Corpora and language learners.* Amsterdam: John Benjamins, 125–35.

Granger, S. 1998: *Learner English on computer.* London/New York: Addison Wesley Longman.

—— 2002: A bird's eye view of learner corpus research. In Granger, S., Hung, J. and Petch-Tyson, S., editors, *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins, 3–33.

Hawkins, R. 2001: *Second language syntax: a generative introduction.* Oxford: Blackwell.

Housen, A. 2002: A corpus-based study of the L2 acquisition of the English verb system. In Granger, S., Hung, J. and Petch-Tyson, S., editors, *Computer learner corpora, second language acquisition and foreign language learning.* Amsterdam: John Benjamins, 77–116.

Jagtman, M. and **Bongaerts, T.** 1994: Report -COMALA: a computer system for the analysis of interlanguage data. *Second Language Research* 10, 49–83.

Lenko-Szymanska, A. 2004: Demonstratives as anaphora markings in advanced learners' English. In Aston, G., Bernardini, S. and Stewart, D., editors, *Corpora and language learners.* Amsterdam: John Benjamins, 89–107.

MacWhinney, B. 2000a: *The CHILDES project: tools for analyzing talk,* volume 1. Transcription format and programs. 3rd edition. Mahwah, NJ: Lawrence Erlbaum.

—— 2000b: *The CHILDES project: tools for analyzing talk,* volume 2. The database. 3rd edition. Mahwah, NJ: Lawrence Erlbaum.

Malvern, D. and **Richards, B.** 2002: Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19, 85–104.

Marsden, E., Myles, F., Rule, S. and **Mitchell, R.** 2003: Using CHILDES tools for researching second language acquisition. In Sarangi, S. and van Leeuwen, T., editors, *Applied linguistics and communities of practice,* volume 18. London: British Association for Applied Linguistics/ Continuum, 98–113.

Mitchell, R. and **Dickson, P.** 1997: *Progression in foreign language learning.* Report of a project funded by the Economic and Social Research Council, 1993–96. Centre for Language in Education: Occasional Paper no. 45: University of Southampton.

Mitchell, R. and **Martin, C.** 1997: Rote learning, creativity and 'understanding' in classroom foreign language teaching. *Language Teaching Research 1,* 1–27.

Myles, F. 2002: *Linguistic development in classroom learners of French: a crosssectional study.* No. End of ESRC award report R000223421. Southampton: University of Southampton.

—— 2003: The early development of L2 narratives: a longitudinal study. *Marges Linguistiques* 5, 40–55.

—— 2005: The emergence of morpho-syntactic structure in French L2. In Dewaele, J.-M., editor, *Focus on French as a foreign language: multidisciplinary approaches.* Clevedon: Multilingual Matters.

Myles, F., Hooper, J. and **Mitchell, R.** 1998: Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning* 48, 323–63.

Myles, F. and **Mitchell, R.** 2005: Using information technology to support empirical SLA research. *Journal of Applied Linguistics* 1, 69–95.

Myles, F., Mitchell, R. and **Hooper, J.** 1999: Interrogative chunks in French L2: a basis for creative construction? *Studies in Second Language Acquisition* 21, 49–80.

Nesselhauf, N. 2004: How learner corpus analysis can contribute to language teaching: a study of support verb constructions. In Aston, G., Bernardini, S. and Stewart, D., editors, *Corpora and language learners.* Amsterdam: John Benjamins, 109–24.

Paradis, J., Corre, M.L. and **Genesee, F.** 1998: The emergence of tense and agreement in child L2 French. *Second Language Research* 14, 227–56.

Pienemann, M. 1992a: Computational analysis of language acquisition data. In Pennington, M. and Stevens, V., editors, *Computers in applied linguistics.* Clevedon: Multilingual Matters, 201–43.

—— 1992b: COALA - a computational system for interlanguage analysis. *Second Language Research* 8, 59–92.

Rule, S. 2004: French interlanguage corpora: recent developments. *Journal of French Language Studies* 14, 343–56.

Rule, S. and **Marsden, E.** in press: The acquisition of negatives in classroom learners of French. *Second Language Research.*

Rule, S., Marsden, E., Myles, F. and **Mitchell, R.** 2003: Constructing a database of French interlanguage oral corpora. In Archer, D., Rayson, R., Wilson, E. and McEnery, T., editors, *Proceedings of the corpus linguistics 2003 conference,* volume 16. University of Lancaster: UCREL Technical Papers, 669–77.

Rutherford, W. and **Thomas, M.** 2001: The *child language data exchange system* in research on second language acquisition. *Second Language Research* 17, 195–212.

Scott, M. 1999: *WordSmith tools.* Oxford: Oxford University Press.

Tono, Y. 2004: Multiple comparisons of IL, Ll and TL corpora: the case of the L2 acquisition of verb subcategorisation patterns by Japanese learners of English. In Aston, G., Bernardini, S. and Stewart, D., editors, *Corpora and language learners.* Amsterdam: John Benjamins, 45–66.

Appendix 1 List of corpora mentioned

- Cambridge Learner Corpus: http://uk.cambridge.org/elt/corpus/clc.htm
- French Learner Language Oral Corpus: www.flloc.soton.ac.uk
- ICLE Corpus: http://www.i6doc.com
- Longman Learners' corpus: www.longman.com/dictionaries/corpus/lclearn.html