Using information technology to support empirical SLA research

Florence Myles and Rosamond Mitchell

Abstract

This article argues the case for more widespread adoption of information technology in second language acquisition research. The extent to which electronic datasets and analysis tools such as concordancers and parsers are being adopted across applied linguistics research more widely is briefly reviewed. Reasons for relative lack of adoption of such tools within SLA to date are discussed, and a case is made for use of the CHILDES suite of tools to support SLA research. The background to the CHILDES project in L1 acquisition research is explained, and a detailed account is provided of how CHILDES has been adapted to support a corpus-based programme of research into French L2. The article concludes with a brief assessment of the strengths and weaknesses of CHILDES, and an outline of issues to be addressed if the full benefits of information technology are to be realised in SLA.

Keywords: second language acquisition; corpus linguistics; information technology; research methodology; French

Affiliations

Florence Myles: University of Newcastle, UK.
Rosamond Mitchell: University of Southampton, UK.
Corresponding author Florence Myles, School of Modern Languages, Old Library Building, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK.
email: Florence.Myles@ncl.ac.uk
1 Introduction

The enormous potential contribution of computer-aided methods to language research has been recognised within many areas of applied linguistics for a couple of decades, despite some expressions of skepticism (see e.g. Cook, 1998). The rise of corpus linguistics has led to major developments in applied domains such as lexicography and pedagogic grammar, which can be illustrated by influential database projects such as COBUILD/Bank of English (Sinclair, 1987; 1991; www.cobuild.collins.co.uk) and CANCODE (Carter, 1997; 1998, uk.cambridge.org/elt/corpus/cancode.htm) and publications such as the Longman Grammar of Spoken and Written English (Biber et al., 1999). All these initiatives have promoted computer-aided analysis of a variety of extremely large electronic corpora of contemporary English, written and/or spoken, so as to develop empirically grounded descriptions of English grammar, discourse and vocabulary and exploit them for learners’ purposes. There have also been considerable developments in the use of corpus techniques in discourse analysis, both for the study of written genres on a very large scale (e.g. Stubbs, 1996) and also spoken genres (e.g. Biber, 1988; Carter & McCarthy, 2004). This work has relied largely on the application of concordancers such as Wordsmith Tools (www.lexically.net/wordsmith), though major advances have also been made in the development of software supporting other types of analysis, notably part-of-speech taggers such as the University of Lancaster’s CLAWS (www.comp.lancs.ac.uk/computing/research/ucrel/claws/). An overview of these developments and their overall impact in applied linguistics is offered by Hunston (2002).

By comparison, the domain of Second Language Acquisition (SLA) research has been until recently relatively unaffected by new technologies. As we argue in more detail below, the preoccupation of SLA researchers with theory and conceptualisation has meant that the community has largely been content to work with relatively small and/or non-naturalistic datasets. Our overall judgment is that the field is at a point where the development of expertise in handling larger naturalistic datasets using new technologies is appropriate and necessary, if further progress is to be made in testing competing theoretical positions and developing stronger empirical grounding for SLA theory.

This article therefore reviews some of the recent progress which has been made, in bringing new technologies to bear on SLA problems, and in undertaking corpus-based work in SLA. The article does not review exhaustively all ongoing work in this field, but concentrates primarily on the work of a research team at the University of Southampton, led by the present authors. First of all we will briefly outline the rationale for using new technologies in SLA research, before examining in more detail the possibilities they offer in terms of data storage, management and analysis. In a second section, the software
tools used by the research team (the CHILDES system devised and housed at Carnegie Mellon University, and already well known in first language acquisition research: http://childes.psy.cmu.edu) will be described. In particular, how these tools can be adapted and used to facilitate SLA data analysis will be illustrated in some detail. The last main section will introduce a new web-based database of French learner oral corpora (FLLOC. www.flloc.soton.ac.uk), containing a range of corpora of learners of French as L2 at different developmental levels, in the shape of sound files, transcripts, and files which have been tagged morphosyntactically. The article concludes with some evaluative remarks on the progress made to date in the adaptation of CHILDES to SLA purposes, and some assessment of future priorities.

2 Rationale

It seems self-evident that one of the most precious resources in SLA research, alongside a clear conceptual framework, is a good quality dataset to work on. The language produced by learners, whether spontaneously or by means of various elicitation procedures, remains our main window into trying to understand how second languages are acquired. The hope is, through careful analysis of learners’ L2 productions, to capture both general trends in learner development, and what is variable in this development. From these descriptive accounts of language in use, it should be possible to build more satisfactory models of the underlying knowledge structures and developmental processes which make up the learner’s interlanguage system. In our view this work crucially relies on good quality data, from large sets of learners, whose profile is clearly documented.

For the purposes of fundamental SLA research, should datasets comprise spoken or written data? To us it seems that oral data must be a better window into the learner’s underlying interlanguage system than written data, which may be complicated by extra layers of ‘learned’ knowledge and monitoring processes. Research on L2 writing is itself worthwhile, but is a distinctive field with its own research agenda, more akin to L1 literacy research. For this reason, while there are some very large electronic corpora of L2 writing already in existence, including the pioneering International Corpus of Learner English (Granger et al., 2002), and the 10-million word Longman Learners Corpus, consisting of written work produced by L2 learners of English from a wide range of L1s and levels (see www.longman.com/dictionaries/corpus/lclearn.html), we do not review this work in detail here, but will concentrate instead on issues to do with the creation and analysis of oral learner corpora.

Collecting oral data is a labour intensive endeavour, involving a range of specialised skills. Access to learners needs to be negotiated, tasks need to be carefully constructed in order to elicit the kind of data which will provide the best
window into competence, and then need to be administered, recorded and transcribed, before analysis can take place. Because of the complexity of this process, researchers have traditionally only managed to collect small datasets; even very ambitious projects, such as the 1980s European Science Foundation (ESF) project (Perdue ed., 1993), which involved learners from 11 source and target languages, reports in detail on developmental trends among quite small numbers of learners in each setting (for some language pairs, only 3 or 4 learners). The dataset which resulted has been the source of significant proposals concerning the course of learner development (e.g. proposals regarding the existence of a common ‘Basic Variety’ as a general developmental stage evident across different pairs of source and target languages: Perdue & Klein, 1997). However, since the published analyses are often based on these limited numbers, it is difficult to evaluate the generalisability of the ESF theoretical claims. In order to address more definitely such issues as e.g. the existence (or not) of any ‘Basic Variety’, larger datasets are needed than are currently referred to in the literature.

This felt need is relatively new; SLA is an area of research which only really took off as an independent field in the 1970s. Initially, its research agenda was largely concentrated on hypothesis generation, which can be achieved effectively by careful analysis of small samples of data. In research on first language acquisition, a slightly older and more mature field, early work also concentrated on very small samples, typically diary studies of single children. However in the case of L1 acquisition research, these have now largely been replaced by analyses of much larger and more focused datasets. Similarly, the research agenda in SLA is changing. Many detailed hypotheses have now been generated, under the umbrella of varying theoretical approaches. The field is busily engaged in testing these (often competing) hypotheses, as well as refining them and devising new ones. Often, though, conclusions are still reached on the basis of language samples produced by very small numbers of learners, leading to what may be premature generalisation. The need for basing SLA analyses on larger samples has been argued by an increasing number of researchers in the field (Rutherford and Thomas, 2001).

The first need therefore, is that much larger corpora of oral interlanguage data should be collected for use in SLA research. New technologies cannot play a lead role in this area; computers cannot (yet) design data elicitation tasks, administer them, or transcribe oral data with sufficient accuracy for research purposes. All this still needs to be done by specialists. There are three ways in which new technologies can help at this stage, however: data storage, data management and above all, accessibility. These contributions of information technology to data collection are briefly outlined in the next section, before a more extended discussion of the main potential contribution of IT to SLA research, its role in supporting data analysis.
3 Data storage, management and accessibility

3.1 Data storage

Oral data can now be easily collected digitally in locations convenient to research participants, using lightweight portable equipment such as digital video cameras, recording minidisks, or voice recorders using different types of flash memory. The resulting digital sound files, usually of much better quality than analogue data, can then be stored on a computer. When sound files have been transcribed, the resulting protocols can also be stored electronically. Provided that legal data-protection requirements have been observed, and that the usual precautions necessary to safeguard any type of electronic data are taken, all the data can be archived in a safe and durable manner, yet all types of file are easily accessible and replicable.

3.2 Data management

Computers have made data management much easier, and many software programs can be used to facilitate the management of large datasets. Programs such as the University of Michigan’s Soundscriber (www.lsa.umich.edu/eli/micase/soundscriber.html) support transcription, and programs such as Adobe Audition (formerly known as Cool Edit Pro: see www.adobe.com/special/products/audition/syntrillium.html) make it easy to handle and manipulate sound files, making it much easier to locate a particular excerpt, to edit files for anonymisation purposes, or to time pauses precisely, for example.

3.3 Data accessibility

The main advantage of using new technologies in order to handle large datasets of oral L2 data is accessibility to the wider research community, via a web interface or other system for file sharing. This applies both to transcribed protocols of L2 speech, to digital audio files, and increasingly to digital video files too. The advantages of data sharing and opening up access to large L2 corpora to legitimate users are evident:

- SLA researchers have the option of using existing L2 corpora to pursue a range of focused research agendas, and are not inevitably lumbered with expensive and time consuming processes of interlanguage data collection. Decisions as to the necessity for fresh data collection can depend on the nature of the research questions being asked.
Well designed datasets have a long ‘shelf life’, and can undergo a range of different types of analysis by different research teams, which can be shared with others. This means that such datasets will be much more fully exploited than if used by the original creators alone.

The claims made by a given research team regarding the interpretation of phenomena within the L2 corpus can easily be scrutinised and cross-checked by other researchers with full access to the original dataset.

Different research groups, with different expertise, can contribute different types of material to a pool of secondary resources such as analysis files (for example, the same oral L2 corpus could be tagged for its lexical or syntactic features by one group of researchers, while another group could work on prosodic or phonological development, or sociolinguistic and discourse features).

It is possible that different theoretical interpretations and explanations of interlanguage phenomena can be compared and evaluated more directly, if tested on the same large datasets.

Creation of large shared datasets could be a stimulus to the field to work with more standardised (and better theorised) transcription procedures, and in particular encourage researchers to address more systematically the problems involved in representing oral interlanguage with all its non-standard features, in written language.

We should note here that these advantages are much more likely to accrue if the software being used for analysis procedures is made available as freeware and access to learner databases themselves is made available without charge to bona fide researchers and to students. In some areas of applied linguistics, corpus-based work has been undertaken as a commercial or semi-commercial venture (e.g. the various corpora supported by ELT publishers), and in such cases access to software and/or to corpora has been restricted to some degree, or may be available only through subscription. (The Cambridge Learner Corpus referred to earlier, for example, is only available for use to authors and staff of Cambridge University Press, and the International Corpus of Learner English is relatively expensive to buy.) We believe that where data has been collected and/or software developed using public funds, access both to data and to software should be unrestricted. The CHILDES project is a long-standing example of good practice in terms of open access, whose access procedures we recommend strongly to the applied linguistics/SLA community, as the creation and use of L2 learner corpora develops further.
4 Data analysis: what kinds of tools are available?

It is in the field of data analysis that the major potential benefits of computer based technologies for SLA research lie at present. Hunston (2002) divides the general types of analysis which the overall development of corpus linguistics has made possible into ‘word-based’ and ‘category-based’ methods.

Word-based methods involve the use of concordancers such as *Wordsmith Tools*, which identify words (or rather, letter strings separated by blanks!) within a corpus, and perform a variety of operations upon them. The best known of these methods involves concordancing, i.e. the identification and collation of all examples of a given word or phrase within the corpus, together with their immediate context (typically, a line of text). This basic procedure allows for a variety of further analyses, e.g. identification and statistical analysis of lexical collocates of the target word or phrase, or of the grammatical relations it enters into. (See Hunston, 2002: Chapters 3 and 4.) Other word-based methods include the use of software to calculate the mean length of utterances (MLU) in a given corpus, to work out the frequencies of lexical types and tokens in a corpus, or to identify ‘keywords’ within a text (Hunston, 2002: 67–8).

Category-based methods involve some form of annotation of the corpus, so that the words it contains are allocated to different categories proposed by researchers and program writers, which can then also be used as the basis for corpus searches and statistical procedures (Hunston, 2002: 79–94). Corpora may be annotated manually, semi-automatically or automatically; Hunston discusses the trade-offs in terms of corpus size on the one hand, and the accuracy and refinement of annotation on the other, depending on which method is used (2002: 82–3). One of the most common category-based methods is the use of so called ‘part-of-speech’ (POS) taggers, such as CLAWS or the *Bank of English* tagger, which typically achieve over 90 per cent accuracy in assigning the words in a corpus to different grammatical categories, through a mix of application of rules and probabilistic judgments. (For example, the latest version of CLAWS has now been used to POS tag over 100 million words of the *British National Corpus*: UCREL, 2004.) Computer programs have also been written in order to parse previously annotated corpus texts, i.e. to identify and label sentences, clauses and phrases of different types (see e.g. Leech & Eyes, 1997, quoted in Hunston, 2002: 84). However, the output of parser software still typically requires editing and correction by researchers.

Apart from tagging to support various kinds of grammatical analysis, category-based methods have been developed to annotate corpus texts for a wide variety of other features, such as semantic content and discourse functions (see e.g. Garside, Leech & McEnery, 1997 for examples). Annotation may be comprehensive, or selective, as in e.g. the annotation of learner errors in a
corpus of anonymous examination scripts collated from the Cambridge ESOL English examination suite (the Cambridge Learner Corpus: http://uk.cambridge.org/elt/corpus.clc.htm).

5 Data analysis options for second language acquisition

If it is accepted that new technologies have great potential for enhancing the storage, management, sharing and analysis of datasets in SLA research, in the ways outlined above, what progress has been made to date in using them?

Despite slowly increasing availability of electronic datasets (e.g. the L2 oral datasets becoming available via CHILDES, discussed later), the SLA research community still seems relatively slow to make use of software for the analysis of large learner corpora. (See Granger, 2002 for an overview of existing work.) There are a number of reasons for this, of which the main one is arguably the (non) availability of (or lack of familiarity with) a full range of suitable software, as well as the fact that applied linguistics researchers are seldom trained in the use of computerised methodologies. For some applications, programs such as concordancers can be very effective and useful. However, as we have seen, this type of program only supports searches for words/phrases/morphemes, rather than more abstract categories such as parts of speech, or phonological patterns. (They can of course retrieve POS if the corpus has been tagged, but they do not include tools for tagging the corpus automatically.) For this reason, they have been used primarily by acquisitionists working on vocabulary or discourse (e.g. Aijmer, 2002; Hasselgren, 2002). Much of the SLA research agenda to date has been concerned with the acquisition of morphosyntax, for which concordancers have limited use.

Some researchers have undertaken manual tagging of L2 corpora, e.g. the work of Granger and her colleagues who have annotated written corpus of learner English and of learner French for error analysis purposes (Dagneaux, Denness & Granger, 1998; Granger, Vandeventer & Hamel, 2001). People have also experimented with using POS taggers such as CLAWS with L2 data, as a support for morphosyntactic work. However, most well established tagger programs used in corpus linguistics have been designed to handle standard languages as used by native speakers, whether spoken or written, and error rates go up greatly when such programs are used with learner corpora with any significant proportion of non-standard interlanguage forms, code switching, or other phenomena distinctive to L2 productions. (See Granger’s comments on this problem, 2002: 18.)

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 & Bongaerts (1994). However, it seems these programs have been
discontinued, and no POS tagging software is currently available which 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.

When the Southampton team decided in 2001 to investigate what new technologies could contribute to SLA research for all these purposes, in the context of a research project which aimed to collect a large amount of oral French L2 data (Myles, 2002), we came to the conclusion that the most promising existing option was the CHILDES system, 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 which can be used to carry out concordancing, frequency counts etc. Most importantly for us given our research agenda, CHILDES also makes available POS taggers for a range of languages (currently 8 languages, with 6 more in preparation). It has also proved relatively easy to modify the POS taggers according to our own criteria and research needs. In its latest version, the CHILDES system is also now XML compatible.

The next section presents this system in more detail and outlines the ways in which we have used it to facilitate the management and analysis of French L2 oral corpora.

6 The CHILDES system

6.1 Introduction

The Child Language Data Exchange System (CHILDES) was first conceived in the early 1980s by MacWhinney and Snow, in order to share first language acquisition data. The resulting project has grown and developed ever since, forming the backbone of much L1 research of the last 20 years or so. (See MacWhinney, 2000a and 2000b, for a complete overview.) The reasons for developing the system were very similar to those outlined in the first part of this article in the context of SLA. But whereas it is now used by virtually all L1 researchers, and training in its use forms an integral part of postgraduate programs in this field, very limited use has been made of it up to now in SLA (with a few exceptions, e.g. Paradis et al., 1998; Housen, 2002; Malvern and
Richards, 2002), in spite of pleas by leaders in the field (e.g. Rutherford and Thomas, 2001; Ellis 2002):

… It is vital that the enormous effort dedicated to the development of the tools and methods of the CHILDES project (MacWhinney, 2000a; 2000b), which has allowed such strides forward in child language acquisition research, be built on by researchers of SLA. (Ellis, 2002: 319)

The CHILDES project is based at Carnegie Mellon University under MacWhinney’s direction, and is easily accessible on the web. The system comprises three components: the database, the transcription system, and the tools for analysis, as well as various resources including the manuals explaining the system, an online bibliography containing some 24,000 L1 acquisition references, etc.

**Talkbank**, the database of corpora, is freely downloadable, and contains an impressive variety of language learning data, primarily L1 English, although the learning of some 26 languages is represented. There are also some corpora from language impaired children and adults, bilingual children, and a few second language corpora. All the corpora include computerised transcripts which are in CHAT format (see below), and many also include digitised soundfiles and some video data. The number and variety of corpora available is increasing steadily; it is a condition of use of the CHILDES tools that any project making use of them should donate their data to TalkBank.

**CHAT** (Codes for the Human Analysis of Transcripts) is the transcription and coding system, which is formatted to be compatible with the analysis tools. The conventions and notation system are described in detail in the manual (http://childes.psy.cmu.edu/manuals/CHAT.pdf).

**CLAN** (Computerised Language Analysis) is the suite of analysis programs which enable automatic tagging and searching of the transcripts. There are currently some 37 such programs (see Table 1) enabling various computations and powerful searches to be carried out on batches of files (http://childes.psy.cmu.edu/manuals/CLAN.pdf).

The following sections will briefly present some of the CHAT conventions before focusing in greater detail on the possibilities for analysis the system offers.

### 6.2 The CHAT transcription system

We will not review the transcription conventions in detail here; they are similar to other orthographic transcription conventions meant to deal with oral data, and are now standard within the field of L1 acquisition. They must however be adhered to strictly in order for the CLAN tools to work on a given dataset.
<table>
<thead>
<tr>
<th>Command</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAINS</td>
<td>Tracks sequences of interactional codes across speakers.</td>
</tr>
<tr>
<td>CHECK</td>
<td>Verifies the accuracy of CHAT conventions in files.</td>
</tr>
<tr>
<td>CHIP</td>
<td>Examines parent-child repetition and expansion.</td>
</tr>
<tr>
<td>CHSTRING</td>
<td>Changes words and characters in CHAT files.</td>
</tr>
<tr>
<td>COLUMNS</td>
<td>Reformat transcripts into columnar form.</td>
</tr>
<tr>
<td>COMBO</td>
<td>Searches for complex string patterns.</td>
</tr>
<tr>
<td>COOCUR</td>
<td>Examines patterns of co-occurrence between words.</td>
</tr>
<tr>
<td>DATES</td>
<td>Uses the date and birthdate of the child to compute age.</td>
</tr>
<tr>
<td>DIST</td>
<td>Examines patterns of separation between speech act codes.</td>
</tr>
<tr>
<td>DSS</td>
<td>Computes the Developmental Sentence Score.</td>
</tr>
<tr>
<td>FLO</td>
<td>Reformats the file in simplified form.</td>
</tr>
<tr>
<td>FREQ</td>
<td>Computes the frequencies of the words in a file or files.</td>
</tr>
<tr>
<td>FREQMERG</td>
<td>Combines the outputs of various runs of FREQ.</td>
</tr>
<tr>
<td>FREQPOS</td>
<td>Tracks the frequencies in various utterance positions.</td>
</tr>
<tr>
<td>GEM</td>
<td>Finds areas of text that were marked with gem markers.</td>
</tr>
<tr>
<td>GEMFREQ</td>
<td>Computes frequencies for words inside gem markers.</td>
</tr>
<tr>
<td>GEMLIST</td>
<td>Lists the pattern of gem markers in a file or files.</td>
</tr>
<tr>
<td>KEYMAP</td>
<td>Lists the frequencies of codes that follow a target code.</td>
</tr>
<tr>
<td>KWAL</td>
<td>Searches for word patterns and prints the line.</td>
</tr>
<tr>
<td>MAKEDATA</td>
<td>Converts data formats for CHAT files across platforms.</td>
</tr>
<tr>
<td>MAKEMOD</td>
<td>Adds a %mod line for the target SAMPA phonology</td>
</tr>
<tr>
<td>MAXWD</td>
<td>Finds the longest words in a file.</td>
</tr>
<tr>
<td>MLT</td>
<td>Computes the mean length of turn.</td>
</tr>
<tr>
<td>MLU</td>
<td>Computes the mean length of utterance.</td>
</tr>
<tr>
<td>MODREP</td>
<td>Matches the child’s phonology to the parental model.</td>
</tr>
<tr>
<td>MOR</td>
<td>Inserts a new tier with part-of-speech codes.</td>
</tr>
<tr>
<td>PHONFREQ</td>
<td>Computes the frequency of phonemes in various positions.</td>
</tr>
<tr>
<td>POST</td>
<td>Probabilistic disambiguator for the %mor line</td>
</tr>
<tr>
<td>POSTLIST</td>
<td>Displays the patterns learned by POSTTRAIN</td>
</tr>
<tr>
<td>POSTTRAIN</td>
<td>Trains the probabilistic network used by POST</td>
</tr>
<tr>
<td>RELY</td>
<td>Measures reliability across two transcriptions.</td>
</tr>
<tr>
<td>SALTIN</td>
<td>Converts SALT files to CHAT format.</td>
</tr>
<tr>
<td>STATFREQ</td>
<td>Formats the output of FREQ for statistical analysis.</td>
</tr>
<tr>
<td>TEXTIN</td>
<td>Converts straight text to CHAT format.</td>
</tr>
<tr>
<td>TIMEDUR</td>
<td>Uses the numbers in sonic bullets to compute overlaps.</td>
</tr>
<tr>
<td>VOCD</td>
<td>Computes the VOCD lexical diversity measure.</td>
</tr>
<tr>
<td>WDLEN</td>
<td>Computes the length of utterances in words.</td>
</tr>
</tbody>
</table>

**Table 1:** CLAN commands (source: CHILDES on-line manual, http://childes.psy.cmu.edu/manuals/CLAN.pdf pp 37–8)
For details see MacWhinney, 1996; 2000a; Marsden et al., 2003; Rule et al., 2003. Here we will only outline the aspects of CHAT which are necessary to understand fully the use of the CLAN analysis programs.

All files must start with *headers*, which contain crucial information for identifying the file and the participants, and are necessary for requesting CLAN to carry out analyses on specific files or parts of files.

The most important feature of CHAT, from the point of view of analysis, is the organisation of the transcripts in a number of *tiers*, called the *main tier* and a range of *dependent tiers*. The main tier is the line on which is transcribed what speakers have actually said; it always starts with an asterisk followed by a three letter code representing the speaker, e.g. *21N*. Dependent tiers (lines) are where tags or comments are inserted (either manually or automatically, depending on the tier), and always start with a % sign. For example, the %mor tier will contain the morphosyntactic tagging of the transcript, the %err tier contains a description of errors. A number of standard dependent tiers are described in the CHILDES manual, but researchers can choose whether to use all/any of these, and can easily add their own, should their research agenda require it. A selection of the standard tiers is described briefly below:

*Action Tier* %act: This tier describes the actions of the speaker or the listener.

*Comment Tier* %com: A general purpose comment tier.

*Error coding Tier* %err: This tier codes errors.

*Gloss Tier* %gls: This tier can be used to provide a “translation” of the learner’s utterance into the researcher’s working language, if different.

*Morphosyntax Tier* %mor: This tier codes morphemic segments by type and part of speech.

*Paralinguistic Tier* %par: This tier codes paralinguistic behaviours such as coughing and crying.

*Phonology Tier* %pho: This tier is used to describe phonological phenomena.

*Situational Tier* %sit: This tier describes situational information relevant only to the utterance.

*Speech Act Tier* %spa: This tier is for speech act coding.

*Syntax Tier* %syn: This tier is used to code syntactic structure with trees.

The advantage of organising the data in this way is that researchers can include whatever tiers are necessary for their data analysis, given their own research agenda. Some are generated automatically, others by hand coding; either way, the new levels of coding can then be added to the database for use by other
researchers. The data does not become increasingly cluttered, however, as transcripts can be viewed and handled without the various dependent tiers. Furthermore, the powerful CLAN tools can then be applied to any of the tiers. We will illustrate in some detail how these tiers can be used to facilitate SLA analysis after having described some of the CLAN tools available.

6.3 Overview of CLAN

A list of the programs currently available in CLAN is presented in Table 1. The list is lengthy, and most researchers will only use a few of these commands. We will only explain here a few of these programs, selected for their potential usefulness for SLA research, and illustrating different types of analysis (e.g. lexical, morphological, syntactic). In each case, we will illustrate the working of the programs with small scale analyses conducted on our own database of oral L2 French (FALLOC; www.floc.soton.ac.uk). This has been collected from classroom learners with English as L1, aged from 12 years upwards, and is presented more fully in the following section.

6.3.1 Word-based measures: MLU

The mean length of utterance is used extensively in L1 acquisition research as a measure of a child’s developmental level in the early stages of acquisition. It is primarily useful for short utterances (MLUs of 5 and under), and has been little used in SLA research because this typically concentrates on learners who are much more advanced than this. However we have found it useful as a rough measure in studying development among early L2 learners. For example, the following MLUs were found for the same instructed L2 French learner (coded *01) undertaking a conversational task with an adult researcher at two year’s interval (at age 12 and age 14):

Example 1

From file <01A7iiPRD.cha>
MLU for Speaker: *01A:
Number of: utterances = 98, morphemes = 232
Ratio of morphemes over utterances = 2.367
Standard deviation = 1.847

Example 2

From file <01L9viJVH.cha>
MLU for Speaker: *01L:
Number of: utterances = 33, morphemes = 197
Ratio of morphemes over utterances = 5.970
Standard deviation = 3.504
We can see that the MLU for learner *01 has increased from 2.36 to 5.97. Scaled up for use with group data, calculations of mean MLU can give a rough general measure of the level of complexity of learners’ interlanguage at different stages.

6.3.2 Word-based measures: Lexical analysis

A number of CLAN commands can assist in the lexical analysis of a corpus. As well as various concordancing searches (KWAL, COMBO) which will be illustrated later, programs such as FREQ and FREQPOS compute the frequency of the words produced in a file or batch of files, according to sentence position in the case of FREQPOS.

In Table 2 we present the output of FREQ applied to the same transcript of an early conversation task for learner *01, arranged in descending order (various switches allow for different presentation formats, e.g. in alphabetical order, ascending or descending order of frequency, etc).

| 20 non  | 2 la  | 1 heures |
| 15 oui  | 2 madame  | 1 hill |
| 14 euh  | 2 mercredi  | 1 informatique |
| 11 um  | 2 pardon  | 1 je |
| 9 ehm  | 2 quatre  | 1 je@g |
| 7 mmm  | 2 quelle  | 1 je@n |
| 6 le  | 2 salle  | 1 kitchen@d |
| 6 oh  | 2 technique  | 1 la@g |
| 5 un  | 2 un@n  | 1 livres@g |
| 4 cantine  | 2 une  | 1 mardi |
| 4 de  | 2 verts  | 1 matin |
| 4 j’  | 2 à  | 1 ne@g |
| 4 xx  | 1 a  | 1 ok |
| 3 aime  | 1 ai  | 1 p@l |
| 3 deux  | 1 bibliothèque  | 1 pas@g |
| 3 football  | 1 bruns  | 1 piscine |
| 3 français  | 1 c’  | 1 pizza |
| 3 récréation  | 1 cahiers  | 1 post |
| 3 sept  | 1 cour@g  | 1 pourpres |
| 3 trois  | 1 crème  | 1 préfere |
| 3 vingt+cinq  | 1 cuisine@g  | 1 préfère |
| 2 ah  | 1 d’  | 1 religieux |
| 2 anglais  | 1 de@g  | 1 rouges |
| 2 bâtiment  | 1 dix  | 1 récréation@g |
| 2 chips  | 1 douze  | 1 s@ss |
| 2 cinq  | 1 déjeuner  | 1 sais@g |
| 2 classe  | 1 détente  | 1 six |
| 2 des@g  | 1 e@l  | 1 sport |
| 2 eh  | 1 en@g  | 1 spring |
| 2 et  | 1 est  | 1 tarte |
| 2 fille  | 1 est@g  | 1 tennis |
| 2 frites  | 1 février  | 1 tom |
| 2 grand  | 1 garçon  | 1 vingett+xx |
| 2 il  | 1 glaces@g  | 1 vous@g |
| 2 information  | 1 haut@g  | 1 éducation |

Table 2: FREQ output for Learner *01, conversation task, age 12
(From file <01A7iiPRD.cha>)
At the bottom of the output, it can be seen that the program also gives the total number of word types, word tokens, and type/token ratio. Such frequency data can be used in a variety of ways, e.g. as a progress measure, or as the basis for further analyses of categories of words known/used, of contexts of use, etc. Again, these analyses can be scaled up to include several files for the same learner, or batches of files from entire groups of learners, to produce group information.

6.3.3 Category-based measures: morphosyntactic analysis

From the point of view of our own research agenda, which was primarily morphosyntactic in nature, probably the most significant factor in choosing CHILDES for assisting our analyses was the fact that it incorporates an automatic morphosyntactic tagger, MOR, with versions for a number of languages including some of those most commonly studied in SLA (English, French and Spanish). Below is part of a file from Learner *19 from the Southampton dataset, undertaking a picture based narrative task in L2 French. This has been transcribed using CHAT conventions, and is shown both before it has been tagged with a %mor tier (Example 3), and after processing using the MOR and POST programs, plus final manual disambiguation (Example 4).

Example 3: Original excerpt

"19L: euh après+midi ma mère et garçon petit # est magasin et voi(ture) [/] voiture.
*JVH: mmm.
*19L: euh # grand+mère # et fille et garçon # euh # dit au revoir.
*JVH: c’est bien ok.
*19L: # um # ma mère # oh # le magasin.
*JVH: oui ok.
*19L: euh grand+père pêche [/] # [%eng: no] [+eng] peint +/.
*JVH: peint oui.
*19L: +, les bouées [*] # euh # une garçon et fille # [%eng: oh # carry][+eng] +/.
%err: boues = bouées
%com: sounds like this throughout

Notes: *19 is the learner number; JVH is the researcher code; # indicates a pause; [/] indicates a retaking; [%eng: xx][+eng] indicates a word in English; [*] indicates an error on the main tier; () indicates an unfinished word; x/+.x, indicate an utterance carried over from one speech turn to the next.

Example 4: After automatic tagging

"19L: euh après+midi ma mère et garçon petit # est magasin et voi(ture) [/] voiture.
%mor: co|euh n|après+midi det:poss|ma&SING n|mère&_FEM conj|et n|garçon&_MASC
adj|petit&MASC v:exist|être&PRES&3SV n|magasin&_MASC conj| et n|voiture&_FEM.
*JVH: mmm.
*19L: euh # grand+mère # et fille et garçon # euh # dit au revoir.
%mor: co|euh n|grand+mère&_FEM conj|et n|fille&_FEM conj|
The process of tagging the data involves three steps. First the MOR command is run on a file or batch of files within the CLAN environment. The output produced contains all the possible tags for individual words. For example, the French word *pas* has two different meanings/word classes (negative particle ‘not’, and common noun ‘step’). The MOR program therefore initially tags it both as a negator and as a noun. Similarly, *les* (‘the’; ‘them’) is tagged as both a determiner or a pronoun, *aime* (‘like/s’) as either a first person singular present or subjunctive, a third person singular present or subjunctive, or an imperative. The next step involves running the POST program on the MOR output. By taking probabilistic features of the immediate linguistic context into account, this program automatically disambiguates around 90 per cent of the data. When POST has not been able to decide according to context, final disambiguation takes place manually, using a command in the CLAN window which takes you directly to all remaining ambiguities and offers a selection of options.

Once morphosyntactic tagging has been completed, complex searches can then be carried out directly on this morphosyntactic output, as illustrated in the next section.

### 6.4 Sample searches with CLAN

Complex searches can be carried out on any of the tiers created for a given CHAT transcript, provided the tiers to be analysed are correctly specified. KWAL and COMBO are the main search tools in CLAN. We will illustrate COMBO in this section. A number of symbols can be used to define a search string (see CHILDES online manual, http://childes.psy.cmu.edu/manuals/CLAN.pdf, p. 55):
### 6.4.1 COMBO on main tier

A simple search for the word *français* as a search string on the learner tier, carried out for Learner *01* for the same file as that used in Example 1 above, is shown as Example 5:

**Example 5**

```plaintext
combo +t*01A +s"français" 01A7iiPRD.cha
```

**Combo:** 
- **CLAN search command**
- **+ t*01A:** specifies the tier (in this case we only want to do a word-based search on the main tier for Learner *01*, excluding the researcher’s utterances)
- **+s:** start of the search string
- **"word":** word searched for
- **01A7iiPRD.cha:** name of file the search is carried on

The output of this command is as follows:

**Example 6**

```plaintext
From file <01A7iiPRD.cha>

*** File “01A7iiPRD.cha”: line 23.
  *01A: (1)français .

*** File “01A7iiPRD.cha”: line 87.
  *01A: classe de (1)français .

*** File “01A7iiPRD.cha”: line 120.
  *01A: classe de (1)français .

Strings matched 3 times
```

The output indicates the file which has been searched (when the analysis has been carried out on batches of files, this is very useful), then lists all utterances in which the string has been found, indicating at the end how many times the string has been found.
6.4.2 COMBO on %mor tier

Category-based searches can also be carried out with COMBO, on any of the dependent tiers. The tier(s) to be searched are specified in the command line. This allows searches for tags or strings of tags, enabling the systematic retrieval of syntactic patterns or morphological phenomena. For example, all instances of verbs in the past tense can be retrieved, or all relative pronouns etc. More complex strings can also be searched for. For example, all instances of a negator directly followed by a verb in the infinitive, or all adjectives following a noun, or all masculine determiners followed by a feminine noun etc. The possibilities are as diverse as the research questions.

The following COMBO command is written to search directly on the morphosyntactic output, i.e. the %mor tier, for instances of incorrect assignment of gender concord, in this case instances where a masculine determiner is followed by a feminine noun. For illustrative purposes the command has again been written for a single file (Learner *48 undertaking the same picture narrative task):

Example 7

```
combo +%mor +s"det*MASC*^n*FEM" 48L11EMA.mor.pst
*: indicates any metacharacter (some symbols might follow the MASC code on the mor output)
MASC: the code for masculine
^: followed by
n: the code for noun
FEM: the code for feminine
48L11EMA.mor.pst: the name of the file. ‘.mor.pst’ is the file extension after it has been ‘mored’ and ‘posted’.
```

The output is as follows:

Example 8

From file <48L11EMA.mor.pst>
----------------------------------------
*** File "48L11EMA.mor.pst": line 14.
*48L: euh près de Loch Ness et il y a une grand+mère les trois enfants et euh le mère de les enfants euh.
%mor: co|euh adv:place|près prep|de n:prop|Loch n:prop|Ness conj|et pro:subj|il&MASC&_3S proy|v:poss|avoir&PRES&3SV det|une&FEM&SING n|grand+mère&_FEM det|les&PL adj|trois&_PL n|enfant-_PL conj|co|euh (1)det|le&MASC&SING (1)n|mère&_FEM prep|de det|les&PL n|enfant-_PL co|euh
----------------------------------------
*** File "48L11EMA.mor.pst": line 20.
*48L: ils sont à la Loch et le grand+mère euh peint .
This analysis shows that Learner *48 produced 5 instances of masculine determiners followed by feminine nouns. Complementary searches for other combinations of masculine and feminine attribution on nouns, determiners and adjectives make it very easy to find patterns of gender attribution across large amounts of data.

As another example, the following COMBO will find all instances of ne directly followed by a verb in the infinitive followed by pas, for all the files ending in .pst, i.e. all the output files within the current working directory that have already been analysed using MOR and POST:

Example 9

COMBO +t%MOR +s"^ne\^v"^\*inf\*\*^pas\*".pst

These examples illustrate the kinds of potentially powerful analyses which the CHILDES tools allow. As will be evident from the long list of CLAN programs given in Table 1 above, our presentation has been far from exhaustive, but rather aimed to outline the main possibilities offered for SLA lexical and morphosyntactic analysis when using these tools.
7 Adapting CHILDES to SLA-specific needs

Our decision to use CHILDES software to support our SLA analysis was not theory-neutral, and has influenced the way in which we transcribed the data in several respects. Furthermore, we had to make a number of minor adaptations to the CHILDES conventions and programs in order to deal with our French L2 learner oral data, which we outline below.

7.1 Transcription of SLA data within CHAT

Transcribing oral data always involves a certain amount of interpretation on the part of the transcriber, especially when the decision has been made to transcribe orthographically rather than phonetically. When dealing with L2 learner data, especially from beginners, it is often tricky to interpret what a learner has intended to say, not only because their pronunciation might be deviant, but also because we may be predisposed to impose a target language form onto their interlanguage form. When a learner of French L2 produces a verb with a zero ending, for example, it may be hard to tell whether they intended producing a first person singular, an imperative, a second person singular or third person singular or plural verb form in the present or in the subjunctive, all phonetically identical in target French. We may not know whether their interlanguage grammar distinguishes between all these forms. The only way to be sure is to trace in learner production when these forms start appearing in complementary distribution with other forms marking other tenses or person or number, e.g. the [õ] ending for the first person present plural, or the [e] ending for the infinitive, past participle or second person present plural. Computerised searches subsequently make it easier to retrieve these instances and analyse the complementary distribution of all verb forms. The easy availability of digital soundfiles, and the facility available within CHILDES for linking these to related transcripts, also facilitates re-listening and re-checking, should doubts arise about the accuracy of any transcribed segment.

A basic principle of using computerised tools for data analysis is that the software has to be able to recognise all the data entered. When using a MOR grammar to tag the data, the program has to recognise all the words on the main tier, otherwise it will not be able to parse them. It does this by a fairly belt and braces procedure – each language-specific version of MOR includes a lexicon for that language, in standard orthography, and matches words encountered in the transcript with the relevant lexicon entries, in order to produce the first-order tags.

When dealing with learner data which is full of non-standard forms, neologisms and code switching between languages, this matching process can be problematic. These problems can be handled in a number of ways. At a lexical level, if the intended word is clearly retrievable from the context, one option is
for the word to be entered by the transcriber on the main tier as the target word, followed by the symbol [*] indicating an error. The %err tier can then indicate what the learner actually said. This allows the morphosyntactic output to remain faithful to the learner’s sentence structure. For example, if a learner produces [lǝ garkõ] instead of [lǝ garsõ] (‘the boy’), we want the parser to recognise that the learner has produced a version of the noun garçon. This will allow the program to assign this production the structure of a noun phrase, while retaining on the error tier the information that the word was mispronounced, as follows:

Example 10

*021:  le garçon [*] jouer.
%err:  garkõ = garcon

At a morphological or syntactic level, errors do not need to be entered separately on an error line. As we have seen in the COMBOs illustrated in Examples 7–9 above, it is very easy to retrieve specified errors by interrogating the morphosyntactic output. Mismatches of gender on nouns and determiners, or of person on subject and verb, can readily be searched for. A problem remains, however, when learners produce a novel form with deviant morphology. For example, our learners often generalised the regular infinitive verb ending [e] to irregular verbs, e.g. prener (target prendre, ‘to take’). In this case we instructed the tagger to assign the following analysis:

Example 11

prené  [neo:v:inf] ”prendre” (neo = neologism; v = verb; inf = infinitive)

We can therefore retrieve such forms as examples of untensed verbs, and also identify for analysis all neologisms constructed by the learners.

Other transcription issues we commonly had to deal with included cases where morphology was obscured by poor pronunciation and/or indeterminacy within the interlanguage system. For example, very often, French L2 learners produced a default determiner form which sounded like a mixture of le (definite masculine) and la (definite feminine), or un (indefinite masculine) and une (indefinite feminine). Transcribing such forms as either masculine or feminine would have over-interpreted the data, and would also not have allowed us to track when they start making the distinction between masculine and feminine determiners. We therefore transcribed these forms as ‘le@n’ and ‘un@n’ and instructed the MOR grammar to assign them a new analysis, e.g.:

Example 12

*32L:  le@n fille
%mor:  neo|le n|fille&_FEM
Another characteristic of our data was code switching, in our case the very common use of English in early learners’ productions, either in the form of single words within French sentences, or entire sentences. The French MOR grammar will obviously not recognise these forms, and therefore cannot assign them an analysis. When an utterance was entirely in English, we instructed the parser to ignore it. When, however, learners used a single English word in an otherwise French utterance, we wanted the grammar to recognise the sentence structure of that utterance. For example, if a learner produced *j’aime manger les frites et les ‘sausages’* (‘I like eating the chips and the *sausages*’), we did not want the parser to analyse the second noun phrase as consisting of a determiner and nothing else, which would misinterpret this learner’s grammar, suggesting it allows determiners not to be followed by a noun, which is clearly not the case. In this case, we transcribed any English word as ‘word@d’ in the case of nouns, ‘word@v’ in the case of verbs etc, and instructed the MOR grammar to tag them as English nouns, verbs or adjectives (‘word@a’), as exemplified here:

Example 13

n:eng|sausages.

There were many other instances in our data which we wanted to code in specific ways for later analysis, and CHILDES allows for additions to the transcription conventions to meet such needs (for further details, see www.flloc.soton.ac.uk). The next step involved ensuring the MOR grammar would tag our data in the way we needed.

7.2 Adaptations to the MOR program

The examples we have given above of ways in which we adapted the transcription conventions for SLA purposes show the flexibility of the CHILDES tools. Once decisions have been made about added coding, care must be taken to add them into the lexicon and the MOR grammar. This is straightforward, however, and MOR analyses can easily be changed. For example, the CHILDES MOR grammar for French analysed the synthetic form *au* (‘to the’) as a preposition; because of our interest in the emergence of the determiner system, and a tendency in our learners to initially use an analysed neologism *à le*, we changed the MOR analysis of *au* from *prep|au* to *prep:det|au&-_MASC&-_SING*.

Some words will also need to be added to the MOR lexicon, but this is easily done, as one of the CLAN programs can pick out all the words that have not been recognised (this is a good way of correcting errors in transcriptions too!). Again, the FLLOC website provides more details of further adaptations to the MOR grammar.
This section on the CHILDES tools has aimed to show how they can assist L2 data analysis. The following section briefly presents the way in which adopting computerised procedures for data storage, management and analysis, has enabled us to construct a database of French learner oral corpora (FLLOC), freely available to the research community.

8 The FLLOC database (www.flloc.soton.ac.uk)

The FLLOC website (Figure 1) contains an electronic database freely available to the research community, in the form of linked 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 download, 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 digitised 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, some of the transcripts have been tagged using the French MOR program, and the resulting analysis files, i.e. transcripts which have been tagged morphosyntactically, have been made available also.

Figure 1: The FLLOC website
The corpora included in the database come from diverse sources, but have all been donated for shared use by SLA researchers in the UK and in Europe. There are five 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 analysis files in some cases, as follows:

- 60 learners in years 7, 8 and 9 in the UK context (ages 12–14; longitudinal; range of narrative and interactive tasks – 650 transcripts)
- 20 learners in each of years 9, 10 and 11 in the UK (ages 14–16; cross-sectional; four tasks each, some repeated from above project – 240 transcripts)
- 34 learners (age 16; UK GCSE oral examination; 26 native controls – 60 transcripts)
- 12 university undergraduate learners in the UK (longitudinal; narrative and interactive tasks – 300 transcripts)
- 125 Dutch learners of French (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 organised. Files (sound files, transcripts and analysis files) can be directly downloaded for use by *bona fide* researchers who sign up to an explicit users’ code.

9 Conclusion

This article has aimed to show the necessity for SLA researchers to engage with new technologies in order to share the extremely valuable resource that datasets represent, and bring more powerful analysis procedures to bear on SLA problems. The time-consuming task of collecting high quality longitudinal corpora becomes more worthwhile if these can be shared across the SLA community. The advantages of using computerised methods for assisting in data management and analysis have been exemplified through the detailed examination of how one research team has used and adapted the CHILDES software in order to assist their French SLA research agenda.

As users who are still relatively new to the use of computer-aided techniques in general, and to CHILDES tools in particular, we are still learning, but are also very aware of some of the limitations and dangers involved in a methodological shift of this kind. Some are general issues already noted by commentators on
the use of computer aided techniques in other fields of applied linguistics, some are CHILDES-specific. A summary list follows:

- Transcription systems and software programs encompass theoretical assumptions about the nature of language, which to date have been relatively conservative (Hunston, 2002: 93); they facilitate and promote certain kinds of analyses better than others. Thus for example, CHILDES is well adapted to conducting sentence level morphosyntactic analysis of L2 data, but is less well adapted to deal with e.g. discourse level analyses, yet these may be of most interest to many researchers working with advanced learners. Also, some types of analysis can be automated more easily than others. Retrieving all the occurrences of a word is an easy process and so most programs have functions to do that. Extracting data from syntactic trees (parsed data), on the other hand, is more complex and so very few programs are equipped to perform such actions. Similarly, POS-tagging can now be performed automatically with good results – thanks to all sorts of statistical calculations – and with the help of a good dictionary. By contrast, semantic annotation is more difficult without human intervention and so few programs have this type of annotation.

- Well established projects such as CHILDES have many advantages for the novice, such as continuous updating, well known procedures and a large and supportive user community. However there are also some problems in working within a well established project of this type. First, it is perhaps more difficult to think outside well established analysis procedures and promote innovative methodological thinking. Second, an evolutionary project may have some problems in keeping up with advances in computer technology and program design; while CHILDES is now XML compatible for instance, it may have looked very different, if XML had been available when CHILDES itself was first developed. Similarly, CHILDES tools are not menu driven as with most modern software, but involve the writing of command lines, a process which is not error free and can lead to some user frustration!

- The continual central updating of the CHILDES programs by the team at Carnegie Mellon can sometimes create conflicts for local teams who have introduced their own adaptations to individual programs. Additionally, the MOR and POST programs for French are not error-free and can sometimes be inconsistent. Consequently, outputs need to be carefully checked.

- While powerful therefore, we have to acknowledge that the CHILDES system is not especially user friendly or easy to learn, and systematic training is necessary to become a proficient user who can exploit its facilities to the full. In addition to database creation and researcher training, the
ongoing maintenance and updating of computer based corpora and associated analysis software also raise new issues regarding the levels of technical skills and support required for longer term programmes of SLA research using these methods.

- As well as technical issues, a new set of ethical issues arise when making learner data widely available to an unknown community of international researchers over the internet. Careful attention has to be given to the issue of participant consent and to the anonymisation of data made available in this way.

Despite these recognised problems and unfinished issues, we are convinced that the advantages to the SLA research field of engaging more fully with new technologies has the potential to significantly transform our work, as has already happened in other applied linguistics fields. We would encourage others to engage with the range of emerging L2 projects and methods, including CHILDES but not exclusively so, so that we can collectively develop the necessary expertise to use these tools in innovative ways and serve our collective longer term goals of building better explanations of language acquisition, grounded more securely in learner data.

Notes

1 We are thankful to the other members of the project team, without whom this work would not have been possible: Sarah Rule, Emma Marsden, Vladimir Mircevski. We are also very grateful to the various SLA researchers who have donated corpora of learner oral French to the database.

2 Grammars currently exist for Cantonese, Dutch, English, French, German, Italian, Japanese, Spanish, with Arabic, Danish, Hebrew, Hungarian, Mandarin and Swahili in progress.

3 XML (Extensible Mark-up Language): a flexible text format increasingly used in the exchange of a wide variety of data on the web and elsewhere.

4 This project was supported by grants from the Economic and Social Research Council (RES000220070) and the Arts and Humanities Research Board (RE-AN9657/APN15456).

5 We are thankful to an anonymous reviewer for making this point.

References


