This paper describes several methodological decisions made during a study of linguistic development of French in British classroom learners, highlighting the significance of choosing suitable tools for collecting, transcribing, and analyzing oral interlanguage data and noting the usefulness for interlanguage research of the CHILDES (Child Language Data Exchange System) procedures, which consists of three integrated components: the Talkbank database, CHAT (Codes for Human Analysis of Transcripts), and CLAN (Computerized Language Analysis). This paper is based on the Linguistic Development in Classroom Learners of French research project, which documents linguistic progression among classroom learners of French in grades 9-11, analyzes the development of morphosyntactic structures in spoken learner French, and evaluates the creative construction process and its interaction with formulaic languages among instructed learners. The paper notes general problems with the transcription and coding of French interlanguage but suggests that experiences to date with using CHILDES is encouraging. Three appendixes include elicitation tasks, CHAT symbols, and examples of preliminary transcription. (Contains 14 references.) (SM)
Oral French Interlanguage Corpora: Tools for Data Management and Analysis

Emma Marsden, Florence Myles, Sarah Rule & Rosamond Mitchell

Centre for Language in Education: Occasional Paper No. 58

Research & Graduate School of Education
University of Southampton
Southampton SO17 1BJ

September 2002
Oral French Interlanguage Corpora: Tools for Data Management and Analysis

Emma Marsden, Florence Myles, Sarah Rule & Rosamond Mitchell
University of Southampton

0. Introduction

This paper discusses several methodological decisions taken during a study of linguistic development of French in classroom learners. In particular, the significance of choosing suitable tools for the collection, transcription and analysis of oral interlanguage data is highlighted, and the usefulness for interlanguage research of the CHILDES procedures developed originally for the study of first language acquisition is evaluated.

1. The Project: Linguistic Development in Classroom Learners of French

The research project "Linguistic Development in Classroom Learners of French" is directed by Myles, and funded by the Economic and Social Research Council (Award No. R000223421). Further details can be found on the project website at http://www.lang.soton.ac.uk/lingdev2002/. The project has the following overall aims:

- to document linguistic progression among classroom learners of French in Years 9, 10 and 11, extending an existing corpus of oral French interlanguage data for Years 7, 8 and 9 (arising from the 1993-6 project "Progression in Foreign Language Learning": Mitchell & Dickson 1997).1
- to analyse the development of a number of morphosyntactic structures in spoken learner French, including sentence structure, verbal morphology, gender, interrogation, negation, embedding, pronominal reference etc.

1 For further details of this project see the ESRC information retrieval system at http://www.regard.ac.uk.
to analyse the creative construction process, from the Initial State and beyond, and its interaction with formulaic language among instructed learners.

The sample, balanced for gender and academic ability (as measured by the school), consists of three groups of twenty learners in Years 9, 10, 11 in an English secondary school. Each learner was given four oral tests (see Appendix 1), which were administered on a one-to-one basis with native or near-native speakers of French. In order to compare performance across year groups, the tasks were the same for all learners. Three of the tasks had previously been developed and used in the Progression Project, facilitating comparability across the studies. The current project is collecting a database of approximately 50 hours of spoken French and, together with data from the previous project, the Southampton data set will constitute a corpus of some 250 hours.

Efficient means of carrying out detailed linguistic analyses on such data, given the nature of the research questions and the size of the sample, are crucial. The “Progression in Foreign Language Learning” project, mentioned above, produced a large dataset from beginner learners of French, comprising analogue audiorecordings archived on C90 cassettes, plus a full set of transcriptions. The resulting publications, however, (see for example, Myles, Mitchell, & Hooper 1999) drew on relatively small subsets of learners from the corpus, partly due to the fact that the techniques used for data collection and storage did not facilitate rapid analyses of the complete dataset.

This illustrates an issue that has been increasingly discussed in second language acquisition studies, where theoretical claims have proliferated while the scale of empirical research to test these claims has often remained quite small. There have been calls (e.g. Ellis 1999) for a change of scale when documenting linguistic development amongst learners, and testing rival explanations for observed developmental phenomena. We need to make use of methodological developments that enable sophisticated linguistic analyses to be carried out with larger datasets, producing data which can be subjected to more rigorous statistical testing.
This paper presents a selection of electronic tools that have the potential to fulfil these aims. In particular, we argue for the potential of the CHILDES (Child Language Data Exchange System) tools (MacWhinney 2000a) for the study of second language acquisition data.

2. Storage, Transcription and Analysis

Recording the data

All tasks were recorded digitally using Sony Memory Stick IC recorders and stored as 8-bit .wav files (this is necessary in order to use Soundscriber software, described later, and it is also becoming the standard format adopted by those using CHILDES tools). Nowadays it may be commonly accepted that all data must be digital, but the advantages of digital data are perhaps worth spelling out, as they have important consequences for maximising the potential of linguistic data. Digital recording machines themselves are less intrusive (lapel mikes are not necessary), there is no ‘noise’ from the machine itself, the quality and durability of the sound is much better, negotiating your way through files is infinitely more efficient than working with traditional audiocassettes, and noting timings of pauses is easily done. Digital soundfiles can be ‘linked’ to the transcript (using tools provided by CHILDES), enabling simultaneous access to the written and spoken forms. Furthermore, digital data can be more easily shared across the internet2.

Transcription

Soundscriber (freeware at http://www.lsa.umich.edu/eli/micase/soundscriber.html) facilitates transcription of digital sound files. The keyboard is used to play, pause, auto-rewind, fastforward or ‘walk’ through the soundfile (e.g. every 5 second segment is repeated x times). Without such software, which replaces traditional transcribing machines, transcription of digital data can be extremely time-consuming.

2 The CHILDES research group offer free digitisation of data that will be offered to TALKBANK.
Attempts were made in the 1990s to develop software dedicated to the analysis of L2 oral data: COALA (Pienemann 1992) and COMOLA (Jagtman & Bongaerts 1994). However, both are now inactive, and so, rather than developing our own transcribing, coding and analysis procedures using XML (a mark-up code becoming increasingly popular for tagging and sharing a wide range of data), we investigated whether another ‘off-the-shelf’ package could meet our requirements - CHILDES (The Child Language Data Exchange System).

3. CHILDES

This set of tools was originally conceived for first language acquisition data, but it has also been used, in a limited way, by second language researchers. Together with studies ranging from computational linguistics, language disorders, narrative structures, literacy development, phonological analyses and adult sociolinguistics, CHILDES tools have been used in more than 1300 published studies (for a useful introduction to CHILDES see MacWhinney 1999).

Besides the features of specific interest to language researchers discussed in the following sections, CHILDES has several obvious and important advantages. First, the tools are constantly being up-dated by a well-funded team of programmers. Developments are regularly reported via an active community of users (see the main CHILDES website: http://childes.psy.cmu.edu/ ). The system actively supports data-sharing and all the tools discussed in this article can be downloaded free of charge from the internet.

CHILDES consists of three integrated components:

- The large and diversified database (Talkbank) consists primarily of child speech recordings and transcriptions, but also includes some language disorder data and bilingual data. It is a condition of using CHILDES tools that our data will become part of the Talkbank database, and will thus be made easily available in anonymised form for an international research audience.
+ CHAT (Codes for the Human Analysis of Transcripts) are the transcription procedures, a system for notation and coding which has been developed to be compatible with the analysis programmes. This ‘tagging system’ is now being developed to be XML compatible and a CHAT to XML converter has been written (MacWhinney 27 November 2001, personal communication).

+ CLAN (Computerized Language Analysis) is a set of computer programs for carrying out advanced searches of your data. This is a powerful and flexible software package that can carry out rapid and detailed analyses and is designed to recognise the tagging conventions of CHAT. Some CLAN commands can be used with transcriptions that are not in strict CHAT format.

3.1 CHAT transcribing and coding procedures

Every file has a set of ‘headers’ so that the computer can recognise each file (see Appendix 2). Anything that the researchers feel could potentially influence the findings (e.g. participants, elicitation task, date, researcher and transcriber) can be recorded here. Warnings are included in the file headers are so that other researchers wishing to use the data know what decisions have been made (for example, overlapping and precise phonological codes were not applied to the data in our study). The CHAT manual (MacWhinney 2000a) contains codes (see Appendix 2 for a very small selection) that have been developed by various contributors addressing a wide variety of linguistic research agendas (including, for example, codes for Conversation Analysis and the analysis of written data). However, the system also allows new codes to be developed to address project-specific questions.

3.1.a Transcribing words on to the ‘main tier’

The data is transcribed on to a main line as a set of standard language word forms. Each utterance is transcribed on to a separate line and starts with * followed by the speaker code; this line shows what was actually said, by contrast with lines starting with a % sign which contain linguistic tags.
In addition to the main line or tier, there can be multiple 'dependent tiers' that provide ancillary information. These tiers are preceded by a % sign to indicate they are strings of tags. Researchers can decide how many dependent tiers are appropriate for their own purposes. For our research questions we are using a %err tier (error), a %mor tier (morphology) and %com tier (for any additional comments), though researchers using our data in the future are free to add other coding tiers depending on their interests.

### 3.1.c %err tier

The title of the 'error' tier suggests that it is perhaps a remnant of the Error Analysis perspective still popular when the CHILDES tools were first conceived at the beginning of the 1980s. However it offers one way of enabling researchers to code the intended functions of interlanguage. By marking interlanguage features that are of interest to the researcher on the main line with [*], specific features of the interlanguage can be coded on the %err tier as appropriate to the research questions. For example, in our corpus, the emerging grammars of our instructed French learners include many uninflected verb forms which, in isolation, often lack any indication of person, number and/or tense. Thus, je jouer* is used where the context indicates that one or other of the standard forms je joue, il joue, j'ai joué, je vais jouer, tu joues? might have been expected. By tagging the interlanguage form with a suitable code indicating the 'underdeveloped' functional category (for example tense or agreement), we can begin to trace the emergence of such features systematically. This means that it is possible to retrieve both the 'target forms' and the corresponding interlanguage automatically, without having to search the data manually for contextual clues. The %err line already has a fully developed system of codes in the CHAT manual and can, for example, be used if phonological errors are of particular interest.

### 3.1.d %mor line

The %mor line can be used to study the development of morphology and syntax; it encodes syntactic categories and morphological inflections, indicating person, number and gender features. It is now possible to generate a morphological description of the main line semi-automatically by using another CLAN tool, the MOR programme. Versions of this
programme have been produced for a range of languages (ten at present\(^3\)); the parser for French has recently been developed by Parisse, see MacWhinney 2000b. For the programme to parse data from a particular corpus correctly, some time must be spent adding to the lexicon in the programme to ensure it recognises all the words in the corpus. The parsing done initially by the MOR programme produces a redundant description, tagging words on the main line with a variety of possible morphosyntactic analyses. The product of MOR must then be ‘disambiguated’, which is mainly done by POST. This programme checks for permissible morphosyntactic combinations and eliminates discordant/unwanted tags. For example, an initial analysis using MOR might tag the item ‘le’ both as an object pronoun and also as a determiner. A second analysis using POST usually works out from the linguistic context which category was intended and eliminates the redundant tags. The researcher then has to do the final disambiguating semi-manually, for around 5% of the data, by deciding which parsing options need to be rejected and which accepted, for example, whether ‘aiment’ should be parsed as a 3rd person plural indicative or subjunctive. During this disambiguation, researchers can write their own morphosyntactic codes if none of those offered are suitable.

3.2 Analysis using CLAN

Before analyses using CLAN programmes such as MOR and POST are possible, another CLAN programme called CHECK can ensure your file meets minimum requirements to be recognised by CLAN (for example by indicating where the human transcriber has not followed procedures, such as starting each main line with *).

CLAN can carry out lexical, morphosyntactic, discourse and phonological analyses, amongst others, depending on how the data has been coded. As we are interested in aspects of linguistic development (verb morphology, phrase structure, development of negatives and interrogatives, use of formulaic language etc) we can, for example, extract all negative particles according to their context (before and/or after tensed and/or untensed verbs) or all subject clitics in tensed or untensed clauses. By using further CLAN programmes such as FREQ, KWAL and COMBO we can look at the frequency and linguistic context of interlanguage features, by searching for specific words, combination of words and strings of particular morphological codes or ‘error’ codes. POSFREQ does a frequency analysis by

\(^3\) Cantonese, Danish, Dutch, English, French, German, Hungarian, Italian, Japanese and Spanish.
sentence position and MLU calculates the mean length of utterance. In addition, the results of one analysis can be 'piped' through another analysis, allowing multiple analyses. A very useful feature of CLAN is that it can take out all codes, leaving a ‘friendly’ transcript, useful for eyeballing and presentations.

3.3 Flexibility and Project-Specific Problems

The CHAT and CLAN tools are reasonably flexible, so as to accommodate project-specific issues. We illustrate this with a couple of theoretically important areas from our study:

- English learners of L2 French often use a phonologically indistinguishable default form of both the definite and indefinite article, something that lies between ‘le’ and ‘la’ and something that lies between ‘un’ and ‘une’. Similar forms are frequently used for ‘a’/‘est’ and ‘je’/‘j’ai’. CHAT suggests @n can be used to code such forms as morphological neologisms, which can be added to the lexicon and interpreted by the computer as the researcher decides.

- French has differences between its phonetic and orthographic systems (for example, regular present tense er verbs have 5 orthographic but just 3 phonetic inflections). The transcription of verb endings can therefore be problematic, especially where learners frequently use what may be default null or infinitive forms. For example, it is hard to know how to transcribe verb endings that end with the sound /e/ (written aller and allé), when there is no auxiliary to tell us which would be a more accurate written representation of the spoken form. Similarly if a learner appears to be using default null ending forms regardless of subject, for example, le garçon et la fille il* joue* (for written ils jouent), how do we transcribe il and joue? We could choose to transcribe entirely phonetically, using a %pho tier but this would be diversionary from our research objectives. We have therefore opted for mainly orthographic transcription, wherever assumptions can be made consistently, but we are making use of some phonetic symbols for certain neological forms such as fair/e/, prend/e/.
These issues illustrate the fact that two of the goals of any corpus-building process can be contradictory: the first one is to keep the main line as clutter-free and user-friendly as possible. The second one is to be as true to the actual sounds made by the learners. This area is obviously even more contentious in French given the complex relationships between the grapheme and phoneme. In addition we have found that as these learners are from a classroom context where the written word is given high priority, written forms are probably interacting with their oral performance in complex ways.

4. Conclusion

The issues discussed in this paper illustrate general problems with the transcription and coding of French interlanguage. We are in contact with other researchers who have used and are using CHILDES for similar purposes (Malvern & Richards 2002, Housen in press, Paradis, Le Corre, & Genesee 1998), and note that they have reported similar issues. However, our experience to date with using CHILDES is encouraging, and we support Rutherford & Thomas (2001) and Ellis (2002) in that these are powerful tools, capable of both top-down and bottom-up analyses, which will enable SLA researchers to test hypotheses on large datasets and to remain flexible in terms of the frames of reference used (whether this is the target language or some other hypothesis of interlanguage development).
Appendix 1 - Elicitation Tasks

• **Picture Story:** In this task, learners have to tell a story on the basis of a series of pictures. The purpose of this task is to elicit a narrative that will enable us to study sentence structure, verbal morphology, pronominal reference, gender and embedding (see Appendix 3 for a short sample of transcript from this task).

• **Interrogative elicitation task:** This task is an information gap activity in which the subjects have to find out from the researcher missing information regarding the appearance, location and actions of people on a picture.

• **One-to-one interview with photos:** a directed conversation in which the subject has to ask questions related to a set of photos and also respond to questions. The main purpose of this task is to elicit all the structures investigated, with a particular focus on past tense and future verbal morphology.

• **Negative elicitation task:** The subject has to describe a famous person by saying what they do and don't do, and the researcher has to guess who they are from a selection.
Appendix 2 - Small Selection of CHAT symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>unintelligible speech, not a word</td>
</tr>
<tr>
<td>xx</td>
<td>unintelligible speech, treated as a word</td>
</tr>
<tr>
<td>[?]</td>
<td>best guess</td>
</tr>
<tr>
<td>[*]</td>
<td>error on main line</td>
</tr>
<tr>
<td>[//]</td>
<td>repeated material</td>
</tr>
<tr>
<td>0</td>
<td>non completion of a word</td>
</tr>
<tr>
<td>0word</td>
<td>word omitted</td>
</tr>
<tr>
<td>+</td>
<td>compound word</td>
</tr>
<tr>
<td>=</td>
<td>'target' on error tier</td>
</tr>
</tbody>
</table>
Appendix 3 - Example of Preliminary Transcription

From a Year 11 pupil, after about 370 hours of learning French (5 years of lessons), picture-story narration task.

@Begin
@Participants: 45P Subject, SAR Investigator
@ID: fre.devp.45P11S.45P
@Coder: EM
@Group of 8PH: 11Fra
@Stim: Loch Ness Narration
@Transcriber: EM
@Warning: These data are not useful for the analysis of overlaps because overlapping was not necessarily transcribed accurately.

*45P: un [*] famille est en vacances uh au bord <de le> [*] lac.
%mor: *Odetlla nlfamille v:exist&tre&PRES&3SV p|en vacances co|uh prep|au bord de * det|le&MASC&SING * n|lac.
*45P: c'est le Lac Ness.
%mor: pro|ce v:exist&tre&PRES&3SV det|le&MASC&SING n:prop|Lac Ness.
*45P: regarde le [*] grand+mère et <le [*] trois> /// gar() les trois uh enfants deux garçons et une fille la femme le [*] mère des enfants.
%mor: v|regarde-2S det|le&MASC&SING * n|grand-mère conj|et det|le* num|trois [sample only]
%err: le = la $MOR $AGA le = les $MOR $AGA le = la $MOR $AGA
@end
References


Recent Occasional Papers (each £2.50 with post and packing):

40. An Introduction to Differentiation in Modern Language Teaching. Alison Miller, February 1996


44. Instructional Method and Approaches to Reading Development in Beginning Readers. Vincent Connelly, Rhona Johnston & G Brian Thompson, January 1997

45. Progression in Foreign Language Learning. Rosamond Mitchell & Peter Dickson, January 1997


52. Marrying Form and Function: a place for grammar and total target language in the secondary Modern Foreign Languages classroom. Ivy Hogg, September 1998


55. "Before William Shakespeare": the educational implications of teaching oral narratives at junior secondary school level in Sierra Leone. Mohamed Combo Kamanda, April 1999

56. Some Have Credit Cards and Others Have Giro Cheques: a study of New Labour's 'individuals' and 'people' as lifelong learners in late modernity. Alison Piper, May 2000


Also available:
CLE Working Papers 1 107 pp 1990
CLE Working Papers 2 Reading 150 pp 1992
CLE Working Papers 3 134 pp 1994
CLE Working Papers 4 140 pp 1996

(each £6.00 with postage and packing)
CLE Briefing Document No. 3: "Advanced Language Training for English Teachers". Christopher Brumfit.
CLE Briefing Document No. 4: "Adult Language Classes available in Southampton Area". Christopher Brumfit.
CLE Briefing Document No. 5: "Recent Government Reports on Language in the National Curriculum". Christopher Brumfit.
CLE Briefing Document No. 6: "Television, Children and Violence". Andrew Hart.
CLE Briefing Document No. 7: "EMT, ESOL, ESL, EFL, etc: Acronyms and Stereotypes in English Teaching". Christopher Brumfit.
CLE Briefing Document No. 8: "British Studies: Current Discussion". Lynne Pollard.
CLE Briefing Document No. 9: "Literacy as Social Practice: Contemporary International Debate". Mohamed Kamanda.
CLE Briefing Document No. 10: "The Battlefield of Language: The interplay of power and ideology in language policy". Elena Ioannidou.
CLE Briefing Document No. 11: "Interaction in Pre-school and School Settings Around the Time of School Entry". Jane Payler.

(each £2.00 with postage and packing)

Copies of all publications are obtainable from: Research Office, Research & Graduate School of Education, University of Southampton, Southampton SO17 1BJ.

Email: <jm10@soton.ac.uk>

Centre for Language in Education website:
http://www.education.soton.ac.uk/research_and_centres/centres_and_divisions/
NOTICE

REPRODUCTION BASIS

This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").