# Constructing a database of French interlanguage oral corpora

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

## 1 Introduction

There is an increasingly recognised need for datasets of oral production in the second language acquisition research community. This requirement arises in response to two complementary agendas: to advance the second language acquisition research agenda and also for educational reasons. At the University of Southampton, a series of projects have concentrated on the collection of learner oral production data in French. In the mid 1990s the 'Progression in Foreign Language Learning' project collected data from Key Stage 3 classroom learners (Mitchell and Dickson 1997 R000234754). In 2001 –2, the Linguistic Development in Classroom Learners of French' project collected similar data from Key Stage 4 learners (directed by Florence Myles 2001 R000223421). In 2002-3 further projects are running which are archiving these datasets in a standard format, and also collecting and archiving other existing datasets for more advanced learners

This paper outlines the aims of these projects and the rationale behind the development of a database of French interlanguage oral corpora. The second part will report on the methods used for the collection, storage, analysis and sharing of such data, and decisions taken about methodology, considering both project specific needs and the requirements of the wider research community. There is a need for tools that can handle large data sets of oral production data, as working with such data presents several challenges: representing speech in written form, devising coding that is helpful both for specific hypothesis testing and carrying out efficient and reliable analyses.

## 2 The arguments for developing French interlanguage corpora

There are both linguistic and educational reasons for investing in the development of electronic interlanguage corpora.

The field of second language acquisition research is still relatively young and the early work was predominantly exploratory and descriptive, using case studies and small-scale projects. As the field has developed the research questions are increasingly more focussed and theoretically informed. However the scale of empirical research has generally remained small. If larger, more easily accessible datasets were available, it should be possible to test a range of existing theoretical claims and hypotheses more fully.

Complementing the theoretical research agenda are the needs of modern foreign language education. As the field of education renews its interest in specifying the linguistic content of foreign language programmes, for example in the UK Key Stage 3 MFLs Strategy being piloted in 2002-3, it is vital that the process be informed by detailed documentation of the course of interlanguage development. Solid empirically based accounts of learner development provide an extremely valuable perspective on curriculum proposals. The availability of large scale tagged interlanguage corpora will allow much more effective and systematic cross-checking of curriculum proposals against what is known about learner development.

## 3 The research projects

The Progression in Foreign Language research project took place in two English secondary schools over a period of three years (1993-6). It was a longitudinal study that tracked a cohort of 60 pupils from their first year of learning in year 7 to their third year of learning in year 9. The learners undertook a range of oral production tasks on six occasions spread over the period. Some tasks were carried out in learner pairs and some with a researcher. Some tasks were repeated between rounds in order to track progress more closely. The dataset from the project comprises some 200 hours of analogue audiorecordings and the 650 transcripts are in plain text files appropriate for use with concordancing tools such as Wordsmith. The resulting publications, however (see for example, Myles,

Mitchell and Hooper 1999) have so far drawn on relatively small subsets of data from the corpus, partly due to the fact that the tools available did not facilitate rapid analyses of the complete dataset.

The Linguistic Development Project (2001) was conceived to build on the knowledge base that was developed in the Progression Project. Its principle aim was to collect a new dataset of French oral learner language in order to be able to document and analyse progression in the years following on from those covered by the Progression Project, with a particular focus on linguistic development. Some of the same test instruments were used to compare performance on the same tasks at different educational stages. The data from the two projects combined will enable us to build a comprehensive picture of how children develop in French during their secondary education up to GCSE.

The Linguistic Development in Learners of French Project had the following more specific aims

- To analyse the development of a number of morphosyntactic structures in spoken learner French, including phrase structure, verb morphology, gender, interrogation, negation and pronominal reference.
- To analyse the creative construction process, from Initial State and beyond, and its interaction with formulaic language among instructed learners.

The sample consisted of three groups of 20 learners in each of Years 9, 10 and 11 in an English secondary school. Each learner undertook four oral tasks with an adult interlocutor (see Appendix one for description of tasks used in both projects). The project collected approximately 50 hours of spoken French and the data for the two projects combined constitutes a corpus of some 250 hours.

The fact that in the Progression project only small subsets of learner data were used to produce theoretical research papers illustrates an issue that has been increasingly discussed in second language acquisition research: theoretical claims have proliferated while the scale of empirical research to test these have often remained quite small. Efficient means of carrying out detailed linguistic analyses on such data, given the nature of the research questions and size of the sample, are crucial. Another important issue was being able to share the data with the wider research community, given the expensive nature of corpus gathering. So, in the early stages of the Linguistic Development project decisions had to be made with regards to how the data was to be recorded, transcribed and stored in order to meet these objectives.

Our investigations discovered that two attempts to develop software dedicated to the analysis of second language (L2) oral data are now inactive: COALA (Pienemann 1992) and COMOLA (Jagtman & Bonagerts 1994). The research team also investigated the possible use of Extensible Markup Language (XML) to tag and share our data. However, time considerations in developing the necessary skills and analysis software led us to use an 'off-the-shelf' language-specific package which we felt could meet our requirements: CHILDES (The Child Language Data Exchange System). In any case, CHILDES has now been converted to an XML compatible format. Moreover the recent addition of a French morphosyntactic parser to the CHILDES system meant that it was applicable to our research needs.

## 4 A brief introduction to CHILDES

The CHILDES set of tools was originally conceived for first language acquisition, but has also been used for research into language disorders and by some second language researchers (Malvern and Richards 2002, Housen 2003, Paradis, Le Corre and Genesee 1998). CHILDES tools have been used in well over 1300 published ranging studies ranging from L1 acquisition to computational linguistics, language disorders, narrative structures, literacy development, phonological analyses and sociolinguistics (Macwhinney 2002).

Besides the features of specific interest to language researchers discussed in the following sections, CHILDES has several important advantages. First, the tools are constantly up-dated by a well-funded team of programmers, there is an active community of users, the system promotes data-sharing and all the tools 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. There are increasingly more Second Language Acquisition (SLA) datasets available in Talkbank.

- CHAT (Codes for the Human Analysis of Transcripts) are the transcription procedures, which 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 customise each command. 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.

## 5 Recording the data

Although digital soundfiles are not required in order to use the CHILDES tools, the TALKBANK database has now been entirely digitised[1]. This clearly facilitates complete data-sharing. The advantages of digital data also have important consequences for realising the potential of linguistic data. Digital recording machines themselves are quieter and less intrusive, the quality and durability of the sound is much better, negotiating through files is significantly more efficient than working with audiocassettes and the timing of pauses can be more easily done. Digital soundfiles can be 'linked' to the transcript, enabling simultaneous access to the written and spoken forms, compensating, to some extent, for the inevitable shortcomings of written representation of speech, and giving other researchers the opportunity to code the spoken data. Waveforms of the digital file can also be displayed and linked precisely with the transcription.

## 6 CHAT transcribing and coding procedures

### 6.1 Headers

Every file has a set of 'headers' so that the software can process relevant details from each file. Anything that the researchers feel could potentially influence the findings (e.g. elicitation task, date, transcriber etc) can be recorded in these headings. A 'readme' file must accompany each dataset giving a brief description of the project and sample, and information to other researchers regarding the transcription and coding decisions taken (for example, in the current study, precise pauses and phonological codes were not documented).

### 6.2 Main line

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, in contrast with lines starting with a % sign which contain linguistic tags or other relevant information. The CHAT manual (MacWhinney 2000a) contains codes that have been developed by various contributors addressing a wide variety of linguistic research agendas, including, for example, codes for Conversation Analysis, the analysis of written data, sign language, and for phonetic, prosodic, morphological and syntactic features of speech. CHAT also allows new codes to be used to address project-specific questions.

[1] The CHILDES research group offer free digitisation of data that will be offered to TALKBANK. For this project the digital recordings were stored as 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.

A programme called CHECK can ensure your file meets minimum requirements to be recognised by the analysis software CLAN (for example by indicating where the human transcriber has not followed basic procedures, such as starting each main line with *)[2].

### 6.3 Tiers for coding

In addition to the main line, there can be multiple 'dependent tiers' that provide ancillary information. These tiers are preceded by a *%* sign to indicate they are strings of tags and not the primary data. Researchers can decide how many dependent tiers are appropriate for their own purposes. Our data has a *%err* tier (error) and a *%mor* tier (morphosyntax), though researchers using our data in the future are free to add other coding tiers depending on their interests[3], e.g. a % pho tier with phonological coding.

#### 6.3i %err tier

The *%err* line (error tier) enables researchers to indicate errors in the interlanguage. It has a well-developed system of codes to be entered manually onto a tier beneath the main line. The *% err* tier is used by some SLA researchers but had several disadvantages for us e.g. it would have been a particularly arduous task given the interlanguage of beginners, resulting in cluttered coding. More importantly, project-specific error tiers can restrict the questions other researchers can ask of the data. In fact, most grammatical errors are retrievable systematically by searching in the morphosyntactic parsing produced by one of the CLAN programmes (MOR). For this to happen, the written forms on the main line have to be recognised by this programme as belonging to the French lexicon so that forms are correctly parsed. This meant that we did retain one feature of the error tier. Some words that were pronounced incorrectly but the context clearly indicated what was intended by the learner, were transcribed as target-like so that they were recognised by the automatic parser. For example, *bouée* (rubber ring) (given to the learners as a written prompt) was frequently mispronounced, but *'bouée'* was written on the main line and what was actually said was recorded on the error tier. This illustrates that even with a highly systematized transcription procedure, project-specific aims still influence methodological decisions.

#### 6.3ii %mor tier

The *%mor* tier encodes syntactic categories and morphological inflections, indicating tense, aspect, person, number and gender features. It is possible to generate a morphosyntactic description of the main line semi-automatically by using two CLAN tools, MOR and POST. Versions of MOR have been produced for a range of languages (ten at present[4]); the parser for French was developed by Christophe Parisse in 2001[5]. 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 (though one of the CLAN programmes helpfully extracts all 'unrecognised' forms).

Example of transcript with an added MOR tier from the Linguistic Development Corpus: A year 10 learner carrying out the negative elicitation task.

| | |
|---|---|
| *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 . |

[2] *Some* CLAN commands can be used with transcriptions that are not in strict CHAT format by simply typing +y next to the normal command so the program would consider each line as one tier. +y1 would consider utterances as delimited by full stops, question marks, and exclamation marks.
[3] One CLAN command can take out all codes and tiers, leaving a 'friendly' transcript, useful for eyeballing and presentations.

[4] Cantonese, Danish, Dutch, English, French, German, Hungarian, Italian, Japanese and Spanish.
[5] Christophe Parisse's advice during the project is gratefully acknowledged, as is Brian MacWhinney's.

%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 .

### 6.4 Other CLAN commands

CLAN can carry out lexical, morphosyntactic, discourse and phonological analyses, amongst others, depending on how the data has been coded. CLAN programmes such as FREQ, KWAL and COMBO facilitate analyses of the frequency and linguistic context of interlanguage features. POSFREQ does a frequency analysis by 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.

Analyses can be carried out on specific tiers by customising the commands. For example, COMBO searches can search for specific words, word sequences or combinations of lexical items and morphosyntactical and/or 'error' codes. The COMBO command given here was one used in Rule & Marsden 2002 for the analysis of the expression of negation; 'ne' followed by a verb in the present then 'pas'.

COMBO combo +t%mor 10N 10N9ELD.mor.pst +s"*ne^v*pres*^*pas"

*A definition of what each part of the command means:*

combo = specifies the command

+t = specifies the tier

%mor = we want the command to be carried out on the %mor tier

10N9ELD = name of file

.mor.pst = extension given to file after it has undergone MOR and POST programmes

+s = the switch to say look for the string

" = a metacharacter to mark the start of the string you want to look for

* = any character (on the MOR tier there can be some other symbols / characters between 'det' and the following word's code)

ne = you are looking for the word 'ne'

^ = followed by

v*pres = you are looking for a verb with the code 'pres' which is given to any verbs inflected in the present

^ = followed by

pas

" = marks the end of the string you are looking for

All commands can be used to search individual files or whole sets of files stored in the same directory.

## 7 Project-specific issues and flexibility of the tools

Certain aspects of early emerging interlanguage grammars led us to make use of the flexibility of the CHAT and CLAN tools. Two such issues are illustrated below:

### 7.1 Representing interlanguage forms

English learners of French often use 'approximate' forms of definite and indefinite determiners, lying between '*le*' and '*la*' and between '*un'* and '*une'*. Similarly for '*a*' / '*est'* and '*je'* / '*j'ai'*. CHAT suggests *@n* can be used to code such morphological 'neologisms', which can be added to the lexicon and analysed by the parser as the researcher decides in order to track their use. For example we instructed MOR to code such items as follows (neologisms underlined):

*21L: le@n famille
%mor: neo:def:det|le@n n|famille&_FEM
*21L: un@n grand+mère
%mor: neo:indef:det|un@n n|grand+mère&_FEM

Given our interest in morphosyntactic development this enabled us to track systematically the progression from a default, 'genderless' determiner towards the target system. It also avoids the problem of having to decide on a determiner when it cannot be heard clearly, potentially over-interpreting the data.

### 7.2 Syntactic roles in code-switching

Our data has a significant amount of both phrase-internal and phrase-external code-switching. CHILDES tools have been used for the study of bi-lingual data (e.g. Sebba http://talkbank.org/data/LIDES/) though we could not locate studies where the morphosyntactic role of both languages has been tracked. Tags for the main line were therefore devised to represent the different syntactic categories when English was used: @d, after first language (L1) nouns, @v after L1 verbs and @a after L1 adjectives and so on. These codes were then written into the MOR programme to produce the following coding:

*21L: pour le skirt@d
%mor: prep|pour det|le&MASC&SING n:eng|skirt

As we were investigating phrase structure in the Linguistic Development project it was important to be able to track the architecture of the learners' phrases regardless of the language used, for example whether verbs had the necessary number of arguments or whether determiners had the necessary noun complement. Additionally we could study the use of determiners with English nouns, e.g. whether predominantly masculine determiners are used with English nouns.

## 8 Current Developments

### 8.1 Digitising and converting the Progression Project data.

A current project is now underway at the University of Southampton to digitise the analogue recordings from the Progression Project and also convert the transcripts from text formant into CHAT format (ESRC funded project R000220070). When completed, together with the Linguistic Development project data, we will be making available a substantial database of French oral interlanguage data produced by classroom learners in Key stages 3 and 4 (ages 11-16). The complete dataset will ultimately be freely available to other researchers on the web, as a set of digital soundfiles and linked transcripts, with various levels of coding.

### 8.2 Integration of datasets from more advanced learners

We have also recognised the potential advantages of extending the dataset beyond the material collected from early classroom learners in the two original Southampton projects. The research team at Southampton has now received funding from the AHRB for a further project to run from 2003-4 (Myles RE-AN9657/APN15456). The aim of this new project is to digitise and reformat existing French L2 oral corpora contributed by other researchers, as well as to create a web interface enabling researchers to interrogate these corpora according to a number of criteria. The corpora to be integrated into the overall database will complement the existing Southampton material, because it will largely be drawn from more advanced learners such as UK undergraduates.

In this way, the considerable research effort involved in data gathering will be made more 'profitable' by making data more easily processed in large amounts, and also by making it more accessible to the wider research community. It is through these methods that we hope to meet the crucial requirement for second language acquisition research that there should be easy access to a wide range of corpora, transcribed and formatted to an internationally agreed standard (Ellis 1999, Rutherford and Thomas 2001). In addition we hope to provide a sound empirical basis for informing curriculum development and policy decisions.

## 9 Conclusion

The rationale behind using CHILDES, to transcribe, store, and analyse the Linguistic Development Project data, was to construct a dataset that was in a format that adhered to agreed procedures and could be analysed using powerful tools, and that it was readily available to the research community. It both meets the needs of our focused research questions and provides an extremely rich dataset that other researchers can interrogate in other ways. The investment into acquiring knowledge of CHILDES has proved extremely worthwhile. Already, analyses are ongoing using the Southampton material which relates to:

- Use of rote-learned chunks by intermediate learners, and their role in ongoing development of the learner language system (Myles forthcoming, Mitchell and Myles 2003)
- The development of the verb phrase in emerging grammars (Myles, in press a)
- Development of negation by early learners (Rule and Marsden 2002)
- Learner development of discourse management and narrative skill (Myles in press b)
- Verb movement and the development of interrogatives
- The role of NS-NNS and NNS-NNS scaffolding in the microgenesis of linguistic structure

We look forward to sharing this data with the wider research community very soon and to seeing the potential of the CHILDES tools for SLA research being explored more fully in relation to further languages and learner levels.

**References**


Ellis, N. 1999, Cognitive approaches to SLA, *Annual Review of Applied Linguistics* 19, pp. 22-42.

Housen, A. (2002) A corpus-based study of the L2-acquisition of the English verb system, in *Computer Learner Corpora, Second Language Acquisition and Foreign Language Learning,* S. Granger, J. Hung & S. Petch-Tyson (eds). John Benjamins.

Jagtman, M. & Bongaerts, T. 1994, Report - COMOLA: a computer system for the anlysis of interlanguage data, *Second Language Research* 10 (1), pp. 49-83.

MacWhinney, B. 1999, The CHILDES System, in *Handbook of Child Language Acquisition*, Academic Press, pp. 457-494.

MacWhinney, B. 2000a, *The CHILDES project: tools for analyzing talk. Volume 1: Transcription format and programs*. *Volume 2: The database*. 3rd ed. Lawrence Erlbaum.

MacWhinney, B. 2002, http://cnts.uia.ac.be/childes/

Malvern, D. & Richards, B. 2002, Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing,* 19(1), pp. 85-104.

Mitchell, R and Myles F, 2003, *Rote learned chunks and interlanguage development*, to be presented at AAAL annual conference 2003, Washington, USA.

Mitchell, R. & Dickson, P. 1997, Progression in Foreign Language Learning*, Centre for Language in Education Occasional Paper No.45*. University of Southampton.

Myles, F. (in press a) The emergence of morpho-syntactic structure in French L2. In J.-M Dewaele (ed), *Focus on French as a foreign language*: Multidisciplinary approaches. Clevedon: Multilingual Matters.

Myles, F. (in press b), The early development of L2 narratives: a longitudinal study. *Marges Linguistiques*, v.5 Saint-Chamas: M.L.M.S.

Myles, F. (forthcoming), From data to theory: the over-representation of linguistic knowledge in SLA. *Transactions of the Philological Society 102*.

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

Paradis, J., Le Corre, M., & Genesee, F. 1998, The emergence of tense and agreement in child L2 French, *Second Language Research* 14(3), pp. 227-257.

Pienemann, M. 1992, COALA -A Computational System for Interlanguage Analysis. *Second Language Research* 8 (1), pp. 59-92

Rule, S. & Marsden, E. 2002, *Expression of negation in French L2 classroom learners*. Paper presented at BAAL / CUP seminar, July 2002, University of Southampton.

Rule, S., Marsden, E. & Myles, F. 2002 "*The acquisition of negatives in the French L2 classroom*", Paper presented at Eurosla, Basel, 2002

Rutherford, W. & Thomas, M. 2001, The Child Language Data Exchange System in research on second language acquisition. *Second Language Research* 17(2), pp. 195-212.