# A Caribbean French Learner Corpus

**Régis Kawecki**

Centre for Language Learning
University of the West Indies
St. Augustine, Trinidad and Tobago
Regis.Kawecki@sta.uwi.edu

Université de Bretagne-Sud
Laboratoire HCTI
Lorient, France

***Abstract***

*Corpus Linguistics is a driving force in the field of language acquisition. Most research focused on Native Speaker Corpora. Academics only quite recently began collecting data derived from learners of foreign languages (FL) in order to build FL Learner Corpora. This was done primarily for EFL. Other languages have followed but are still relatively small in size. Learner Corpora can help greatly in the teaching of foreign languages. They allow researchers to identify the most recurrent linguistic problems faced by specific populations of students. Identifying such problems is important in adapting the teachers' approach to teaching the FL to the particular difficulties their students encounter. It allows for supplementing textbooks that are in general far too broad in their approach. Such corpora can also be made available to students in a self-access mode. This paper describes the French Learner Corpus being built at the Centre for Language Learning (CLL) at the University of the West Indies. It is made up of short essays – collected over several semesters – written by CLL students learning French at all levels (Novice Low to Intermediate High). Some initial analysis shows that Trinidadian learners of French differ significantly from other English-speaking learners of that language.*

*Key words: Corpus Linguistics, Foreign Language, French Language Acquisition, Teaching*

## INTRODUCTION

Thanks to the seminal work of the recently deceased John Sinclair, Corpus Linguistics has now earned its rightful place in the fields of lexicography, language description, and also in language acquisition. Research done in lexicography and language description used native speaker corpora, whether written or oral (Biber 1998). One such corpus would be the International Corpus of English hosted by the University College London under the direction of Professor Nelson and used for comparative studies of English worldwide. Native language corpora have been used in the classroom for some time now for subjects such as teaching English for specific sciences e.g. biology or translation studies (Portington 1998). "Corpora are now part of the resources that more and more teachers expect to have access to" (Sinclair 2004, p. 2).

It is only quite recently that publishing companies and academics alike began collecting data derived from learners of second/foreign languages in order to build learner corpora. The first learner corpora were concerned primarily with the teaching of English as a Foreign Language (EFL). Well-known examples of commercially based English learner corpora include the Cambridge Leaner Corpus and the Longman Learners' Corpus. Sylviane Granger of the University of Louvain la Neuve in Belgium initiated the International Corpus of Learner English (ICLE) made up of essays written by advanced

EFL students in different countries (Granger 1998). Learner corpora of Chinese learner of English have been developed in both Taiwan and Hong Kong.

Learner Corpora about other languages are now being compiled but are still relatively small in size. There exists for instance an International Corpus of Learner Finnish collected at the University of Oulo (Finland) by Jantunen. As for French, the collections are few and are derived mainly from native speakers like novelists, journalists and politicians.

## CORPORA AND SECOND/FOREIGN LANGUAGE ACQUISITION

Learner corpora are electronic collections of texts produced by second/foreign language learners (Granger 1998) that can be of great help in understanding the process of second/foreign language acquisition and in improving the way the languages are taught. They allow for the study of a particular population's learner language either synchronically or diachronically. The synchronic analysis would analyse the peculiar ways in which learners speak or write a second/foreign language at a certain stage of their learning process. Learner corpora also make possible the study of the learners´output over time. Reseachers can therefore pinpoint what linguistic characteristics are typical of a group of learners at a certain point in time and the way in which that ouput evolves over the course of several semesters.

This synchronic/diachronic Saussurian dichotomy (Saussure 1964) applied to second/foreign language learner corpora is linked to the language acquisition concept of Learner Grammar also referred to as Interlanguage. Broadly defined, interlanguage points to the type of language produced by non-native speakers in the process of learning a second or foreign language. Learners develop, for the most part unconsciously, a personal set of rules to explain the way the new idiom they are discovering organizes the world. By essence, this Learner Grammar/Interlanguage is evolutive and changes as the instruction continues. It proceeds by stages and longitudinal studies of learner corpora can help describing this learning process (Barlow 2005).

This synchronic/diachronic description can take different forms. It can rely on frequency data that ¨include the learners´ over use or under use of lexical or grammatical forms¨ (Barlow 2005, p. 335). It can also research the use of correct forms. Interpretation of the frequency of erroneous forms then follows.

To come up with frequency data requires that the learner corpora be annotated with error tags (Granger 1998) in a very consistent manner. This can be very much time-consuming since the computer tools available for annotating native corpora are not really helpful since they have been designed primarily, not to mark up errors, but the standard usage of language (Barlow 2005).

Identifying and analysing such frequency problems are important in understanding the interlanguage associated with a particular learner population in their endeavour to learn a specific second/foreign language. As importantly, it can lead to improving the tutors' pedagogical strategies used to teach this language. The teaching approach can be adapted to suit the particular difficulties students encounter. Finally, textbooks that, for commercial reasons mainly, are in general far too broad in their approach of the second/foreign language can be supplemented by new materials targeted towards the specific learner population. Chinese learners of EFL certainly do not share the same interlanguage as, let´s say, French learners of EFL. A simple reason lies in the fact that different mother tongues lead to different interferences within the learning process.

## DESIGNING AND COLLECTING THE CORPUS

The French Learner Corpus built at the Centre for Language Learning (CLL) of the University of the West Indies (UWI) is part of a doctoral research done in conjunction with the University of Bretagne-Sud in Lorient (France).

It is made up of small written essays – varying in size from 100 words to approximately 250 words – done by CLL students learning French from the Novice Low up to the Intermediate High levels (on the OPI ACTFL scale). The whole range of French courses offered by the Centre for Language learning is

spread over six semesters, which corresponds to the length of the undergraduate studies at UWI. These essays were done during the two tests taken by CLL students each term and were collected during the period October 2007 – April 2009.

Below is an extract from an essay written by a CLL student at the Intermediate Mid level, which was transcribed as it was written, including all the lexically or grammatically incorrect forms:

> Il y a plusiers monuments remarquables dans P.O.S. Mon favorit est le château Stollemmeyer – un édifice grand et gris à côte de la Savanne. La Savanne est grande et elle s'appelle "Queen's Park Savannah". Il y a beaucoup d'arbres et on peut jouer au sports et se promener la. Quelques fois les agents de police se promene au cheval dans les rues de la capitale. L'embouteillage n'arret pas les cheveaux !

Once it is finished being transcribed, the corpus should include approximately 60,000 words in total, corresponding to approximately 400 written essays. A significant number of contributors have volunteered four different pieces of text or more. All essays were used with the express permission of the CLL learners, an authorization that is renewed each semester.

Although the corpus has been rendered anonymous, analyses of the data require that some personal information be recorded in order to, for instance, associate the use of erroneous forms with variables such as proficiency – the number of semesters done – or gender. This personal information was stored in external files, which is linked to the production output through an ID number. Contributors furthermore were given coded names in order to track their different production so as to be able to analyse the development of their interlanguage over the course of their studies at the CLL.

## ANALYSING AND INTERPRETING THE DATA

Through quantitative and qualitative evaluation, the goal of this research is to come up with a list of the lexical and structural difficulties the CLL students at UWI encounter in their acquisition of French as a foreign language. It also aims at suggesting interpretations to explain this particular set of errors, this specific French learner interlanguage.

Variables have mainly to do with proficiency and gender. The CLL learner population is very homogeneous. The majority of CLL students are from the English-speaking Caribbean – mainly from Trinidad and Tobago. The few non-Caribbean learner of French heve been intentionally discarded so as not to bias the results. All students were taught French using the same series of textbooks called *Breakthrough French*.

The production task and setting were remarkably stable. All essays were written in class without any outside help nor reference materials such as textbooks or grammars. Students were only allowed 30 minutes to complete their essay, which were for the most part concerned with giving personal information and views on very general topics related to everyday life.

When compared to a native speaker corpora, initial analysis reveals that the features found in this French learner corpus have mainly to do with:

- The interference of the mother tongue (L1) in the mainly syntactical aspects of the learners' French production such as the undifferentiated use of feminine or masculine forms;
- The interference of a very prevalent second/foreign language (L2), namely Spanish, principally seen in the lexical erroneous forms produced;
- The influence played by the language used in the course books;
- The overgeneralization of some aspects of the L2 syntax such as the use of the auxiliairy 'avoir' to conjugate the ´passé compose' tense.

The different stages linked to the interlanguage development have just started to be examined and initial results are too few yet to be mentioned in this paper.

## CONCLUSION

The results so far are only preliminary and analysis of the corpus is continuing. It is nonetheless hoped that by the end of this research project, learners and tutors alike should be benefiting from the analyses made and the interpretation given. The French learner corpus will help identify the peculiarities of the CLL students' French interlanguage that are often due to the intrusion of their L1 and/or L2 and also to developmental processes. Once made available to a wider audience, it will disseminate useful generalizations related to foreign language acquisition in Trinidad and Tobago.

French tutors will have access to hard evidence with which to evaluate the textbook curently used for French. It will give them a definite sense of the aspects of French syntax or lexicon that require particular attention and reinforcement through the design of specially targeted exercises.

The corpus and the analysis tools associated with it should be accessible to students as well, if possible on line or stored on CD-Roms, to allow learners to get a hands-on approach of the recurrent erroneous forms made by Caribbean learners of French. By being shown numerous instances of the same syntactic or lexical errors, learners should get a precious feedback and be able to improve their internal French learner Grammar which is the cornerstone of the normal foreign language learning process (Hanzeli 1975).

Preliminary analysis and interpretation of the data confirm some initial impressions: Trinidadian learners of French differ significantly from other English-speaking learners of the language due to the distinct variety of English they use, their proximity to South America and also their particular heritage which draw on different European cultures and languages.

## REFERENCES


Barlow, M. 2005. "Computer-based analyses of learner language", in Ellis, R. & Barkhuizen G. (eds.), *Analysing Learner Language*, Oxford University press, Oxford, pp. 335-357.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use*, Cambridge University Press, Cambridge.

"Cambridge Learner Corpus", in *Cambridge International Corpus*, accessed 8 May 2009, from <http://ww.cambridge.org/elt/corpus/learner_corpus.htm>

Ellis, R. & Barkhuizen G. (eds.) 2005. *Analysing Learner Language*, Oxford University Press, Oxford.

Granger, S. (ed.) 1998. *Learner English on Computer*, Longman, Harlow.

Hanzeli, V. E. 1975. "Learner´s Language: Implications of Recent Research for Foreign Language Instruction". *The Modern Language Journal*, vol. 59, no. 8, pp. 426-432.

"International Corpus of Learner English-ICLE", in *Centre for English Corpus Linguistics-CECL*, accessed 8 May 2009, from <http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm>

Jantunen, J. H. "International Corpus of Learner Finnish", in *Corpus Study on Language-Specific and Universal Features in Learner Language*, accessed 8 May 2009, from <http://www.oulu.fi/hutk/sutvi/oppijankieli/en/ICLFI_Corpus.html>

Nelson, G. *International Corpus of English*, accessed 8 may 2009, from <http://www.ucl.ac.uk/english-usage/ice/>

Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*, John Benjamins Publishing Co., Philadelphia.

Saussure de, F. 1964. *Cours de Linguistique Générale*, Payot, Paris.

Sinclair, J. 2004. *How to Use Corpora in Language Teaching*, John Benjamins Publishing Co., Philadelphia.

"The Longman Learners' Corpus", in *Longman Dictionaries*, accessed 8 May 2009, from <http://ww.pearsonlongman.com/dictionaries/corpus/learners.html>