Writing English in Brazil: a computer-aided analysis ...

Kevin John Keys
Universidade Federal de Minas Gerais

Abstract

Este texto relata uma experiência feita dentro do Departamento de Letras Germânicas em 1992. As redações dos alunos de língua inglesa foram lançadas no computador e analisadas em termos de freqüência de ocorrência das palavras. Algumas implicações dos resultados são discutidas. Sugestões para uma futura pesquisa em moldes semelhantes são apresentadas.

WRITING ENGLISH IN BRAZIL: A COMPUTER-AIDED ANALYSIS OF LEARNER OUTPUT IN EFL

During the first semester of 1993, learners on the general English course in the Departamento de Letras Anglo-Germânicas, Faculdade de Letras, UFMG were asked to submit their written work double-spaced and in two copies. The reason this was done will become clear during this article.

Essentially, the aim was to find out **how** the students on this course write in English: the focus of interest was in vocabulary, the use of simple and complex structures and in the typology of mistakes .[1] This was a pilot project which was intended to allow for a better understanding of what would be involved in a full-scale, longer-term study. This paper deals with the process of analysis and looks at some aspects of vocabulary and the use of structures: and at how these change from beginning students to more advanced **within the limits of this narrow sample.**[2]

All the written work submitted to teachers in the department was entered on the computer as an unadorned text file[3] and this was then transformed into an ASCII file readable by the concordance program that was to be used to analyse the subsequent body of data.

The **Oxford Concordance Program** (Oxford University Press, 1989) is a micro-computer version of a text analysis program originally developed for mainframe computers (in which form the author first became familiar with it) and which offers three functions: concordancing, indexing and word frequency counts. The concordance feature has been used extensively in the analysis of literary texts.

The corpus was produced as a single large file, within which four categories were identified: **beginner, lower intermediate, intermediate** and **upper intermediate** and these categories roughly.correspond to the course structure as follows:

CATEGORY	LANGUAGE LEVEL
beginner	I-II
lower intermediate	III
intermediate	IV-V
upper intermediate	VI-VII

The sample was not well-balanced in terms of content: that is, there was not an equal distribution of contributions across the four levels and the input text - the corpus - is in some ways distorted: there are a disproportionate number of contributions from the intermediate level, for example. Therefore, what follows is not **statistically** highly reliable: it is more of an **impression** of the writing skills of learners in this department over a certain period of time. What this pilot project has done is to make clear how a more statistically reliable project could be devised and pursued.

Let us be clear about some of the terminology involved when dealing with lexical corpora. A word frequency count of the following "corpus"

THE CAT SAT ON THE MAT

tells us that the total number of words is 6. However, we also notice that one of these words is repeated, so that a frequency analysis would look like this:

the	2
cat	1
mat	1
on	1
sat	1
total words	6
total vocabulary	5

In fact, there are 6 items in the corpus, but only 5 are different, < the > occurring twice. The terminology applicable here and in any word frequency analysis is as follows:

TYPES	different "types"of words
TOKENS	occurrences of all words

The **TYPE/TOKEN** ratio represents the **total vocabulary** (i.e. the number of **different** words) divided by the **total number** of words in the corpus, It is therefore a measure of the richness of the text in vocabulary terms: the higher the figure, the richer, or more varied, is the vocabulary that was used by the writer of that text.

The total number of words in the corpus is 123, 510, which is a very small number.[4] The distribuition across categories is as follows:

CATEGORY	N° OF WORDS
beginner	5,365
lower intermediate	27,171
intermediate	44,659
upper intermediate	46,315
	123,510

These numbers are not very interesting. What we should be looking at is the **type/token** ratio, and this is more revealing. The type/token ratios for each category and for the whole corpus are shown in the following table:

CATEGORY	TYPE/TOKEN RATIO
beginner	0,216
lower intermediate	0,146
intermediate	0,122
upper intermediate	0,114
whole corpus	0,0797

From this it looks as though beginning students use a richer vocabulary (theirs is the highest ratio) than more advanced students. There is an explanation for this: at the beginning level of language acquisition, the amount of available vocabulary is limited and the range of expression is also constrained. Hence, there is relatively little choice in terms of vocabulary, and texts are often very brief: repetition of lexical items is therefore not common and the type/token ratio gives a high figure. At more advanced levels, texts are longer and the basic functional vocabulary of the language is more under control and begins to repeat itself: the most frequently used word is of course < the > , as it is in English in general. Here are the data:

CATEGORY	MOST FREQUENTLY OCCURRING ITEM	RELATIVE FREQUENCY
beginner	<I>	5,088
lower intermediate	<the>	5,75
intermediate	<the>	4,75
upper intermediate	<the>	4,76

With the beginning group, < the > is not the most frequently occurring lexical item. This is because of the paucity of the input data for this level. 5000 + words is really not enough material to work with. In a general frequency count of spoken and written English, < the > will always be the most frequently occurring word.

These data for the beginning level are therefore an aberration: it is the result of the relationship between the language input at beginning levels and output. At this level, the variable "language input"is much more easily controllable: we can know better the classroom input that leads to the language production that we can see in the data. That is, we are more certain of the kinds of tasks that were set for the written work of this group, and we are clearer about the limits of the language input at this level. The data, for example, for this group, reveal that the writing tasks that were set demanded a large proportion of first person singular constructions: therefore < I > appears as the most frequent word in the data. We begin to see how an analysis of output can give us some clues as to what kind of input has been taking place during the language learning proces.

We can see a further example of this - and it is an example which also cautions us to be wary about making conclusions based on limited data sets.

An analysis of the word < if > for each group gave the following data:

CATEGORY	CONDITIONAL	SENTENCE	TYPES	
	1st	2nd	3rd	?
beginner	(only 3 occurrences of <if>)			
lower interm.	31	2	2	13
intermediate	68	38	4	41
upper interm.	80	13	2	21

What does this tell us? One thing that is obvious is that intermediate students were using second conditional structures very frequently (38 occurrences, compared with an upper intermediate figure of 13). The reason for this is that they were set a task beginning with the phrase “If I knew...”and this second conditional structure was repeated in the texts of a large proportion of learners. If we compare the result with that of the upper intermediate group, who were not specifically invited to use such a structure, we can see that the tendency is **not** to use second conditional forms with <if>. The disposition to use second conditional forms - in these data - is very dependent upon the teaching input, which confirms the suspicion that in order to provoke more ambitious language use by learners, it is esssential that the teaching input be designed in such a way that it requires that learners use structures and vocabulary that would not be their natural instinct to use. That is, the pedagogic input has to be designed to encourage learners to explore the boundaries of their competence in the language.

To conclude: this pilot project has shown what would be necessary if data such as these are to be more useful and statistically generalisable, if only to the sampled community.[5] The data were classified, very broadly, according to language experience, but not according to task; the limiting nature of each task was not considered; the data have not been related clearly to language input; the corpus is extremely small.

Any future development of this project would have to take these factors into account, or explain why it was not necessary to do so. One possibility would be to associate the data with specific learners, especially if our aim were to follow the development of second language acquisition through case studies. Alternatively, an analysis of the types of errors made by this group of learners may provide insights into such matters as first language interference, peer influence (suggesting the notion of a group-defined idiolect) and the influence of teaching materials and methods. We should not forget, too, that this project is dealing with **writing** skills, a very specific language ability that may indicate very little about the general language ability of a specific learner.

NOTAS

[1] The analysis of mistakes/errors has yet to be undertaken, this being a whole project in itself, involving complex methodological questions.

[2] The sampling used here will be discussed in the conclusion. The written tasks completed by students were of every caterogy - free, guided, discursive, narrative, etc. This variable was not controlled in the pre-project.

[3] "Unadorned"meaning without italicisation, underlining ou paragraphing.

[4] The Collins/COBUILD corpus is approaching 200 million items (October 1994).

[5] That is, within the school or department that was used as the basis for the investigation; to generalize away from that base would require the usual excessive limitations on variables, such that the experiment becomes pedagogically artificial. Classroom language learning research is not inherentlly rigorous in strict scientific terms; but it is **science**, in that it represents a seeking after knowledge.