International Conference
"ICT for Language Learning" 4th edition

"The number of words in a Learner Corpus related to the number of errors"
'Be careful, the more words you write, the more mistakes you will make!'

Katerina Florou
University of Athens (Greece)
katiflo29@yahoo.co.uk

Abstract

The purpose of this paper is to prove the application of Computer Aided Error Analysis and Corpus Linguistics on researching cases, which deal with the use of the Italian language by Greek learners. By investigating amount of errors, marks and length of written texts we come to conclusions about the teachers' most common beliefs. Furthermore, there can be additional observations about the kind of errors that a learner makes so that we can give an estimation on what kind of class activities would help these learners.

1.Introduction

Learner Corpora have recently become an important source of data in second language acquisition studies and the main interests of researchers evolve around the differences between native and non-native linguistic system[1]. The research that is being described in this paper refers to the quantity of errors that Greek learners of the Italian language make in writing. The results of such research have important implications for L2 writing instruction.

2. Research question

Most educators believe that there is an interrelation of the size of a text and the number of errors that may occur in that text; fewer mistakes are likely to be found in a short text rather than in a long one. In this paper we aim at examining the validity of this theory so as to accept or reject it in the end. To be more specific, we focus on teaching Italian as a foreign language and we have drawn important information for our research from a written learner corpus.

3.Research method

3.1 Learner Corpus

The first step as to examine the above hypothesis is to use a Learner Corpus constructed with certain design criteria aiming to analyse the learners' interlanguage. The IFLG [2]has been designed primarily to function as a reference corpus for the systematic analysis of the interlanguage of Greek students learning Italian as a foreign language. For this reason the research questions are focused mainly on the quantitative error analysis. It is an ongoing corpus that contains a number of 20.000 words, which covers a range of 150 students.

The coding of learner variables (age, sex, profession, origin, education level ect.) has been done in spreadsheet form so that data can be further analyzed. The text corpus was typed in txt and the manuscripts have already been corrected by the experimenter.

A tagged Learner Corpus also is a source that gives direct access to researchers to formulate hypotheses and make a statistical analysis of errors in foreign language[3]. Despite the existence of an automatic tool the tagging has started by hand as done in other Learner Corpora[4] and in this phase is an attempt by the researcher not only to understand the cause of the error, but also to negotiate correction, especially in cases where it is difficult to understand what the student is really trying to communicate.

Error tagging in this study has been carried out with an error code scheme adapted by the experimenter, based on the Error Editor "Episimiotis" developed in the University of Athens.[5] Some modification was necessary as to adjust the function of this tool to the Italian language.

At the end of the error tagging the spreadsheet that contains all the elements related to error has this form[1]:

ERROR CODE				ERROR	SUGGESTION
WO	WS	P	PRO	gli abbandonare	Abbandonargli
WC	NU	M	ADJ	difficile	Difficili

Table 1: Example of 2 errors in .xls with the error code and the correction.

To this spreadsheet learner's variables had to be added. For example, to the above two lines there was an extension with the writer's personal data in whose text the above errors were indicated. In the first column there is the name of the txt document, in the next two columns there are the first letters of his/her name and surname, and then the age, the profession, the origin, his/her education level, the genre of the written text, the number of words and the grade.

DOC	LN.	FN	AG	PROF	ORIG	ED.L	GENRE	WORDS	MARK
154.txt	K	X	22	STU	SAMOS	UNI	LETTER	89	27/30
154.txt	K	X	22	STU	SAMOS	UNI	LETTER	89	27/30

Table 2: Codified learners data

The above achieves the correspondence of errors with the student, but also creates an overall picture of the public of the Italian language. At this point it is necessary to note that the text was graded by the teacher.

3.2 Analyzing data

The investigation of the research question can be done in two ways: 1) the marks given to students' written work are compared with the size of their texts and 2) the number of errors of each text of the learner corpus is counted and then it is compared with its number of words.

3.2.1Comparing number of words and grade

Considering the two research questions, the texts were grouped according to the number of words, regardless of the text type and then they were again reduced to 100 texts per group in order to achieve comparable results. Making a first comment on learners' evidence, the first impression is that the standard intuition of teachers: -more words mean more mistakes- is not confirmed.

[1] The first four columns describe the kind of error. In the first example WO stands for "wrong order" ,WS stands for "word sequence", P stands for "phrase" and PRO for "pronoun". In the second example, WC stands for "wrong choice", NU stands for "number", M stands for "morphology" and ADJ stands for "adjective".

	Number of words	Number of texts	Number of texts (%)	Average grade
1st group	…-100	27	17,53%	22,3/30
2nd group	101-150	74	48,05%	23,8/30
3rd group	151-200	32	20,77%	24,3/30
4th group	201-…	21	13,63%	25,4/30

Table 3: Comparison of the grade with the length of the texts.

Specifically in the above table there is about half of the students that have produced texts of the average size and only one fifth of them had produced written texts with more than 200 words. However, this last percentage (13.63%), has collected better grades. Of course there is always a percentage of subjectivity in the decision of the teacher. It is therefore likely to appreciate the fact of toil made by the student more than the actual performance.

3.2.2 Comparing number of words and type of errors

Keeping the above grouping of the papers will be possible to make a comparison of the size of text on the type of errors and the quantity.

In this case the categories of errors were used according to which the error tagging was made. Morphological errors, orthographical errors, intralingual errors, complement errors, identifier errors and interlingual errors.In terms of the morphological errors more errors appear in the texts with more words • this is not the case according to the other three categories, so the conclusions that one can be led are likely to be accidental.

	Morphological errors
1st group	23%
2nd group	26%
3rd group	24%
4th group	27%

Table 4: Morphological errors

In the next category of errors, the above risky conclusion is not confirmed, since the category with the most errors is the 2nd group, i.e. texts of moderate size. So neither spelling errors help to confirm the original case.

	Orthographical errors
1st group	24%
2nd group	28%
3rd group	22%
4th group	26%

Table 5: Orthographical errors

In intralingual errors there is a big, mutatis mutandis, difference in the amount of errors in large documents (over 200 words). Similarly the category of small papers (up to 100 words) presents the fewest errors. Apparently the students in their attempt to extend their thoughts not only they are using more words but also more complex speech, and as a result they are confusing forms and meanings of language. By contrast, those who fear to risk more errors they write less and in a simple way. In this way they avoid intralingual errors (but not morphological or spelling as shown above).

	Intralingual errors
1st group	20%
2nd group	23%
3rd group	25%
4th group	32%

Table 6: Intralingual errors

And in the category of complement errors (adverb, molecule, intent, verb) seems that the longer texts have more errors, but without much in the lead, and also there is no following up to the decline of errors depending on the number of words.

	Complement errors
1st group	22%
2nd group	25%
3rd group	24%
4th group	29%

Table 7: Complement errors

In identifier errors it is clear how the extent of writing affects the amount of errors. It is a group of quite frequent errors from all students so that a way to avoid this kind of error is to limit the words.

	Identifier errors
1st group	15%
2nd group	17%
3rd group	23%
4th group	45%

Table 8: Identifier errors

Finally at the interlanguage errors there is a visible "precedence" of the second largest category (101 to 150 words) but it can not lead to a well justified conclusion.

	Interlanguage errors
1st group	22%
2nd group	29%
3rd group	27%
4th group	22%

Table 9: Interlanguage errors

By observing generally and by comparing the categories of the most common errors with the groups of texts divided by size, someone can be sure that the teacher's traditional advise "few words, fewer errors" to the students is not unfounded, since the texts with the minimum of words have the minimum of errors. The above data do not provide the ability to check the type of texts, i.e. if it is about a formal letter or an essay, but given that in the production of essays there is always a limited range of development, someone can easily assume that the texts of the first group (up to 100 words) are letters. This element helps the student even more, because in letters what is asked is specified by the pronunciation and also, the structure of a letter is always the same and due to this the production is somehow checked and the student feels safe during it.

The above opinion is not in accordance with the conclusion that was stated having examined the previous case where according to the mark, the length of the production does not affect the performance of the student. At that point, it must be clarified that the teacher's personal judgment is an element totally subjective and the performance of the student in class affects the mark. Through the mark that is given to a test, basically the student is evaluated in total.

4. The exploitation of errors

A Learner Corpus [6], helps the teacher to prepare the lesson by providing examples of the use of language, of terminology, the tactics for the improvement of communication, or examples of errors in the delivery of information. This last action could be exploited in an Italian class by using the data that comes out from the IFLG as the results of correction and generally of the research of Learner Corpus, it may give directions to the teacher[7].

Consequently, since the teacher has taken into consideration the order of the most common errors, authentic texts could be provided as cloze tests, of multiple choice exercises, that exploit the error and ask from the students to enter the correct type in the phrase. In that way the environment of the exercises must reflect the environment in which errors are made by the students.

In other cases authentic texts could be used when their use is redundant and the students are told to correct them. For even more reliability of the authenticity it is preferable to use the same textual type.

References

[1] Lenko-Szymanska, A. (2006) "Self-mention in argumentative writing". *TaLC 2006 Proceedings*, Paris 1-4 July, Università Paris 7 pp. 87-88

[2] Florou, K. (2006) "The experience of creating a Learner Corpus: The Italian as Foreign Language to Greek students (IFLG)". *TaLC 2006 Proceedings*, Paris 1-4 July, Università Paris 7, pp. 174-176.

[3] Granger, S. (2003) Error-tagged Learner Corpora and CALL: A Promising Synergy. CALICO Journal 20, n.3, pp.1-16

[4] Pravec, N. A. (2002) Survey of Learner Corpora. ICAME Journal No 26, pp.81-114

[5] Koutsis, I., Markopoulos, G., Mikros, G., (2007). Episimiotis. A multilingual tool for hierarchical annotation of texts. In M. Davies, P. Rayson, S. Hunston, & P. Danielsson (eds), *Proceedings of the Corpus Linguistics Conference CL2007*, 27-30 July, 2007, University of Birmingham, UK. Available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/243_Paper.pdf

[6] Valero, C. (2006) An Ad Hoc Corpus in Public Service Interpreting. In Hornero, A.M. , Luzòn M. J. &Murillo, S. (eds) Corpus Linguistics: Applications for the study of English, Bern: Peter Lang, pp.451-462

[7] Granger, S. , Tyson, S. (1996) Connector usage in the English essay writing of native and non-native EFL speakers of English World Englishes, vol 15, No 1, pp.17-27