Journal of Modern Education Review, ISSN 2155-7993, USA
September 2013, Volume 3, No. 9, pp. 695-704
© Academic Star Publishing Company, 2013
http://www.academicstar.us

Comparing Error Tagged Learner Corpora and Learners' Variables

Katerina Florou
(University of Athens, Greece)

Abstract: 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 (Lenko-Szymanska, 2006, p. 87). The research that is being described in this paper refers to the age of the Greek learners of the Italian language and the errors that are made. The results of such research have important implications for L2 writing instruction.

In Greece, Italian is taught as a second or third language; for that reason the students are mostly adults. Most educators believe that there is an interrelation of the age and the number of errors that may occur in written production; fewer mistakes are likely to be found in texts of younger students rather than in those of the older ones. This paper aims to verify or reject this idea by using quantitative and qualitative research methods.

As a primary tool for this investigation the researcher uses a Learner Corpus constructed with certain design criteria aiming to analyse the learners' interlanguage.

The results of the above research, help the teacher to prepare the lesson by providing examples of the use of language and the tactics for the improvement of communication. This action could be exploited in an Italian class by using the data that is produced from the IFLG, as the results of correction, and generally from the research of Learner Corpus

Key words: learner corpus, foreign language acquisition, computer aided error analysis

1. Introduction

This paper first briefly describes a learner corpus of intermediate Greek learners' written Italian. It then refers to some useful tools, automatic and semiautomatic, and ways of analyzing this collection of texts. At the end, it provides results related to errors cross examined with one of the corpus variables. The conclusions of such research can be helpful to educators as well as to some students.

2. Background

The collection and study of learner corpora, or interlanguage corpora, are powerful and necessary and aid in understanding the production and the communicative needs of FL learners (Miton, 1998, p. 186).

The need for better teaching methodologies is especially acute in educational systems such as the Greek. In Greece the majority of the students, due to modern needs, feel obliged to learn at least one foreign language by the age of 15 or 16 years. Most of them manage not only to achieve an advanced level of the first FL, but also to learn

Katerina Florou, Ph.D., University of Athens; research areas: applied linguistcs, learner corpus research, sociolinguistcs, FLA. E-mail: katiflo29@yahoo.co.uk.

at an intermediate or upper level another FL. The prevalent languages are, of course, English as a first choice, followed by German and French. Most of this process is based on private lessons (individual or in institutes). Although the lesson of the first and second language is also offered in primary and secondary public education, the level achieved by the students is never higher than intermediate. Because of that, there are a great number of University students and employees who would like to complete their course of study in public education or to add another language to their requisites. The most popular among the adult students after English is Italian.

Conventional classroom methods are often inadequate for understanding the interlanguage which is formed not only after the mother tongue interference, but also after the interference by one or more other foreign languages.

Many FL teachers can easily attest that there are some mistakes that "insist on been repeated" in most oral or written productions of the students. This was the trigger of the present work. Those same errors making the teachers wonder can guide him/her in finding some answers and explain to himself/herself and, most important, to the student, the nature of the error.

3. Research Hypothesis

Most educators believe that there is an interrelation of the age of a student and the number of errors that may commit writing a text in the foreign language; fewer mistakes are likely to be found in texts written by younger writers rather than in those written by older ones. This paper aims to examine the validity of this theory so as to accept or reject it in the end. To be more specific, we focus on teaching Italian to adults as a foreign language and we have drawn important information for our pedagogical purpose from a written learner corpus.

4. Methodology

In order to prove the above hypothesis we need to use, first of all, a learner corpus, which is presented below, and in this way there is extended data to help in withdrawing valuable results. On the other hand, an automatic tool, an error editor, is crucial for the error analysis and for the categorization and the counting of the errors.

A first kind of comparison can be made as soon as the errors of every age group are counted and classified. Then this first result will be further compared to the marks given to the students' written work.

4.1 IFLG Corpus

Learner corpora are electronic collections of texts produced by foreign/second language learners. This new resource has a lot to offer both to theoreticians interested in the process of second language acquisition and to practitioners keen to produce more efficient teaching and learning tools and classroom methodologies.

Such a collection is IFLG (Italian as Foreign Language for Greek students). It is a Learner corpus, the data of which was collected as part of an ongoing study and currently consists of 20,000 words, which demanded the participation of 150 students to date.

The IFLG 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 (Florou, 2006).

As Atkins and Clear (1992, p. 5) pointed out "a corpus is a body of text assembled according to explicit design criteria for specific purpose". The purposed of IFLG was clearly referred above. As for the design criteria, they were decided in order to cover the needs of some research in the interlanguage sector and also to be clear enough to continue the gathering of the data in the future. The corpus criteria can be described as follows:

- Size: as de Haan (1992) proves, optimum corpus size depends on the specific linguistic needs of the investigation. This study at present has a number of 20,000 words, which covers a range of 150 texts.
- Medium: this was written, as research has shown that most errors occur in written texts. It should be noted here that most Greek learners do not find the pronunciation of Italian particularly difficult, as the two languages are comparable with regard to phonetics/phonology.
- Genre: essays and friendly letters. Those two styles of writing have not been decided at random. They are the kind of texts demanded to be produced by the students during the certification exams.
- Length: this may vary from 120 to 200 words per text, although essays of greater length are not rejected provided they are in accordance with all the other criteria.
- Grade: the grade of each one of the texts was registered in order to see afterwards what the interrelation between grade and errors is.

The learner variables were collected via questionnaires distributed by the experimenter (first named author) or another colleague, and they were completed immediately. The information gathered is the following:

- Age: as explained above, in Greece Italian is learned as the second, third or fourth language, so all participants were adults and they covered an age range of between 18 and 56.

- Gender: the questionnaires were not anonymous. Therefore, it was possible for the experimenter to acquire information on the sex of the participants.

- Education level: with regard to this question, participants had four choices: secondary, higher, university and postgraduate.

- Profession: in the case of university students, the participants were asked to note their department.

- Other foreign languages: participants were also asked to note the level which they had reached.

- Origin of participants: participants were asked to register the region in which they live.

The participants shared some attributes: they were native speakers of Greek. At the time of testing participants had been learning Italian for two years. Thus, they were at an intermediate level. As became obvious in discussion groups, very few of them had visited Italy and only for a short period of time (2 to 10 days). Therefore, there was no need to specify the L2 exposure.

The next step after collecting the information from the questionnaires was coding the learner variables (age, sex, profession, origin, education level etc.). This registration was done in spreadsheet form and the information of every learner was connected to his/her written production so that data can be further analyzed. The form of these spreadsheets is as it shows below:

Table 1 Codified Learners Data

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

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 follow the age, the profession, the origin, his/her education level, the genre of the written text, the number of words and the grade.

4.2 Computer Aided Error Analysis

Computer aided error analysis is a relatively new technique during which traditional error analysis is aided by an error editor, an automatic tool. The error editor uses error tags according to an error classification. An error tagged corpus can be analyzed through standard text retrieval software tools, from simple ones to the most advanced (Granger, 1998, p. 14).

In the case of IFLG not all texts were in electronic format so the experimenter had to type in .txt most of them. Despite the existence of an automatic tool, the error tagging was started by hand, as done in other Learner Corpora, because in this phase the researcher attempts 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.

The error classification was decided during this procedure. It was based on two previous classifications used in computer aided error analysis but not for the Italian language. The first one was the **UCLEE** classification used for English and French (Granger, 2002, p. 18; Granger, 2003) and the second one used for the analysis of the Corpus was the classification that the University of Athens uses for the Greek language through its error editor "**Episimiotis**" (Koutsis et al., 2007) developed at the same University. The second choice was partially obliged mainly by the choice of the error editor. That is why some modification was necessary as to adjust the function of this tool to the Italian language due to its morphosyntactic differences between the two languages and therefore between the two error classifications.

4.3 IFLG Error Taxonomy

The error classification of the learner corpus has to cover all possible cases and at the same time it has to adjust the language and its grammar. For IFLG the previous studies and the already existing error editors were a compass; but since there was no other learner corpus having Italian as target language, new error taxonomy was necessary.

The first division was in two broad taxis according to their communicative result: the errors that really obstruct the communication, the so-called "local" errors, and those that do not create communicative problems, the "global" errors. This distinction, first made by Cattana and Nesci (2004), groups in two large sets the most frequent categories of errors in Italian language and at the same time includes every part of linguistics.

Under the title of "local" one can place:

(1) morphological errors concerning grammar, and

(2) morphological errors concerning form, i.e., orthographical.

Under the title of global one can place five categories of errors:

(1) syntactical errors,

(2) lexical errors,

(3) errors of register,

(4) errors of style,

(5) punctuation errors.

Within the category of syntactical errors belong (1) errors in the phrase: concerning identifiers and complements, (2) errors in the sentence: concerning coordination and subordination between verb and phrases, (3) errors between sentences: concerning linking words.

Lexical errors are divided in two sub categories: intralingual (the learner generalizes norms of the foreign language) and interlingual (the learner transfers norms of his/her mother tongue). The three last categories of global errors have no further division.

Of course the tagging was not limited to the linguistic characterization of errors; other elements were include in every tag; for example the part of speech (verb, noun, pronoun, etc.), the kind of error (overuse, luck, wrong order, etc.) and more particular information (number, gender, ton, etc). At the end of the error tagging the spreadsheet that contains all the elements related to error has this form[1]:

Table 2 Example of Errors in XLS with the Error Code and the Correction

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

To this spreadsheet had to be added learner's variables. For example to the above two lines there was an extension with the writer's personal data in whose text the above errors were found (see Table 1). 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 of the class in which each one of the participants belongs.

5. Cross-examination of the Variables

Considering the two research questions, the texts were grouped according to the age of the participants, regardless of the text type and then they were again reduced to 100 texts per group in order to achieve comparable results

5.1 Comparing Age and Grade

It needs to be noticed that the first three groups (17–21, 22–24, 25–29) are the wider public of the Italian language, maybe because they invest in a better professional future (Bagna, 2004, p. 71). On the other hand, taking into consideration the profession of this part of learners (which is mainly student) it is easy to draw the conclusion that this certain group of students has more spare time and fewer professional or other obligations.

Despite this, making a first comment on learners' evidence, the first impression is that the standard intuition of teachers — older students mean more mistakes — is not confirmed.

Table 3 Comparison of the Grade with the Age of the Learners

	Age	Average grade
1st group	17–21	24,6/30
2nd group	22–24	23,7/30
3rd group	25–29	22,6/30
4th group	30–56	24,2/30

Specifically in the above table the younger seem to have a better competence in written texts but there is only a small difference between those and the fourth group, the oldest. On the other hand, all four percentages/amounts are so close that no safe conclusion can be drawn.

[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".

5.2 Comparing Age and Number of Errors

After recording and counting all errors of all categories the following results were extracted:

Table 4 Comparison of Age and Number of Errors

Age	Number of errors
1st group (17–21)	23%
2nd group (22–24)	25.2%
3rd group (25–29)	25%
4th group (30–56)	26.8%

The first observation is close to the initial thought. Indeed older learners are less competent, but looking more carefully one can see that errors do not increase as the age advances. It is a fact that learners of the first group make fewer mistakes than those of the following groups, but the differences are too small to arrive at a conclusion.

5.3 Comparing Age and Type of Errors

Keeping the above grouping of the participants, it is possible to make a comparison of their age with the type of errors.

As already mentioned, Episimiotis was adjusted to a new error classification with numerous error categories and subcategories; but looking through all errors it is obvious that the largest number of errors was concentrated in some categories. For example, there were two categories relevant to the aspect (local errors), morphological and orthographical, two lexical categories relevant to the language interference: Intralingual and Interlingual errors, and two categories concerned with the lack or the misuse of some elements within the phrase: identifiers and complements. Those are the categories of errors that are used in order to make a further comparison.

It is worth mentioning that the above categories cover a rather large percentage, 88%, while the others share only 12% of the whole of errors of the corpus.

In terms of the morphological errors (24%), more errors appear among the older students. This is not the case according to the other three categories, so the conclusions that one can be led to are likely to be accidental. There is nothing that can lead to the conclusion that one of the groups cannot produce written texts correctly in terms of morphology.

Table 5 Morphological Errors

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

In the next category of errors (23.8% of all errors), the above precarious conclusion is not confirmed, since the first and the fourth group have the lowest number of errors, and the category with the most errors is the 3rd group, i.e., texts of students between the ages of 25 and 29. So neither do spelling errors help to confirm the original case. But it is worth noticing that those who had the fewest morphological errors had the most orthographical. The teacher can easily think that the uncertainty in morphological structure provokes more attention to spelling and vice versa.

Table 6 Orthographical Errors

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

In the intralingual errors (21.3%) one can come across more or less the same numbers in all four groups. Apparently the interference of the language they are learning is, for all students, despite age, equally problematic.

Table 7 Intralingual Errors

	Intralingual errors
1st group	24%
2nd group	26%
3rd group	25%
4th group	25%

And in the category of complement errors (8.55%) (mainly prepositions that follow adverb, molecule, intent, verb) it seems that older learners commit more errors, and in this case with a significant difference, and also there follows a decline in errors depending on the age. As the years pass, it seems to be more difficult, to use correctly or not to forget the complement. This fact can be combined with another (teachers') common belief that prepositions are a considerable obstacle for every student of the Italian language. This phenomenon could be justified saying that the more one uses the mother tongue the less one can accept linguistic structure which differ, in quality and quantity in his/ her language.

Table 8 Complement Errors

	Complement errors
1st group	21%
2nd group	23%
3rd group	25%
4th group	31%

The biggest part in identifier errors (6.64%) is composed of errors in the misuse of articles. The group that gathers the most errors is not the last one (even though there is a consequent increase in errors depending on the age). The learners of the 4th group can obviously use articles correctly and the knowledge of Greek language helps. Therefore the same element that was a disadvantage in the former comparison is, in this one, an advantage. The younger students may be not influenced that much by the Greek language but there is a hint of interference of the other foreign languages (clearly without any proof).

Table 9 Identifier Errors

	Identifier errors
1st group	21%
2nd group	27%
3rd group	35%
4th group	17%

Finally, in the interlanguage errors (4.4%) there is a visible "precedence" of the second largest category but it cannot lead to a well justified conclusion. But in this comparison it is obvious that students' interlanguage is influenced; also, in this case one can assume that they are influenced not by the mother tongue but by other foreign languages. IFLG also provides information about the knowledge of other foreign languages by the students, and this can be helpful in a cross examination in case of intention to prove the above assumption.

Table 10 Interlanguage Errors

	Interlanguage errors
1st group	19%
2nd group	35%
3rd group	31%
4th group	15%

By observing generally and by comparing the categories of the most common errors with the groups of texts divided by age, one can be sure that the teacher's traditional thought "the younger the student is, the fewer errors will be made" is not unfounded, since the texts of the younger have the minimum of errors. But the same belief it is not fully proved taking into consideration the errors of some types (see Tables 9 ad 10). Most important is to clarify that the first group is the one with a slightly better competence, although younger students demonstrate a partial knowledge of the foreign language in categories like orthography or complements and identifiers, which, incidentally, are large categories; that is why they generally seem to have better performance than the older students. As a result, the fourth group seems to have different types of language abilities but is generally weaker.

The above opinion is not in accordance with the conclusion that was stated having examined the previous case where according to the mark, the age of the participant does not affect the performance of the student. At this point, it must be said 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 himself/herself is evaluated and not only the test.

6. Application in FLA/FLT

A Learner Corpus, 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 (Valero, 2006, p. 461). This last action could be exploited in an Italian class by using the data that comes out from the IFLG and applying it in an exercise in which the same students who were objects of an investigation are going to be "correctors" of their own mistakes. This kind of activity is a common practice, but in this case there are already the results of the error analysis so the educator can focus on specific errors and, in addition, those texts the students are working on are "authentic" learner data. Furthermore, the teacher can improve the competence of students who belong in certain age groups with extra input in specific linguistic sectors

Consequently, since the teacher has taken into consideration the order of the most common errors, authentic texts could be provided as cloze tests (in which, for example, some identifiers will be missing), or as multiple choice exercises (in which, for example, they have to choose the correct complement), which exploit the error and ask from the students to realize his own omissions. In this way, the environment of the exercises must reflect the environment in which errors are made by the students. For even more reliability of the authenticity it is preferable to use the same textual type; in this case, friendly letters or texts of description, even academic speech could be adequate.

Additionally, learners can use text retrieval tools to form some lists of concordances of the erroneous forms and then compare them with similar/analogous concordances of native corpora (in this case Italian) in order to notice the gap between their own and the target language forms (Granger, 2002, p. 26). This activity can work for each of the students in a different way; according to the results of the learner corpus research the teacher can delegate different tasks to each of the students (or each group of students).

A learner corpus as IFLG can also be useful as feedback not only for the learners, but also for the educator. After error categorization he/she has a complete picture of the forms in which there is need for insistence and those which are easily acquired by the learners. For example, the curricula of Italian language usually anticipate more time and extra work in syntactic structures (for example passive form, conditionals, indirect speech ect.) than in morphological structure. Looking at the IFLG evidence one can find out that syntactic structures are no problem. Therefore, devoting so much time to syntax seems rather unnecessary.

The results of correction and generally of the research of Learner Corpus may give directions to the teacher (Granger & Tyson, 1996, p. 22); first and foremost is not to take into consideration common beliefs even if they are based on teachers' observation. Instead, the educator can also adopt the role of researcher in order to offer more practical advice in class.

7. Conclusion

Corpus Linguistics has provided a lot in FLA and FLT. Although at the beginning all the research work was oriented to native corpora, even though most native and non-native corpora existing today concern the English language, there is much volition to compile more corpora, native but also learner corpora and enrich the study of all aspects concerning language and language learners. IFLG is a proof of that and these small studies can be a part of bigger project. There is a lot of information that can be derived from this kind of corpora; their use is not exhausted in cross examining the age of the learners and their number of errors. There are all sorts of variables that can be examined, and a great number of ways that this corpus and other corpora can be proved useful.

In particular, teachers of the Italian language in Greece can gain experience and information on developing new pedagogical tools and activities which targets the needs of the learner, by exploiting such learner corpora or by building their own. Most important of all is the contribution of learner corpora to FLT by providing clarifications and explanations to some as yet unresolved questions, such as the one pointed out in this paper.

Acknowledgements

I would like to thank Prof. George Mikros for the encouraging and the constructive guidance. I am also grateful to Sharon Nicholls for much valuable practical help.

References

Atkins S. and Clear J. (1992). "Corpus design criteria", *Literacy and Linguistic Computing*, Vol. 7, No. 1, pp. 1–16.

Bagna C. (2004). *La competenza quasi bilingue/quasi nativa: Le preposizioni in italianoL2*, Edizioni Francoangeli, Milano.

Cattana A. and Nesci M. T. (2004). *Analizzare e correggere gli errori*, Guerra Edizioni, Perugia.

De Haan P. (1992). "The optimum corpus sample size?", in: Leither G. (Ed.), *New Directions in English Language corpora*, Mouton de Gruyter, Berlin and New York, pp. 3–19.

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, Universitè Paris 7, pp. 174–176.

Granger S. (1998). "The computer learner corpus: A versatile new source of data for SLA research", in: Granger S. (Ed.), *Learner English on Computer*, Longman, London and New York, pp. 186–198.

Granger S. and 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.

Granger S. (2002). "A bird's eye view of learner corpus research", in: Granger S., Hung J. and Petch-Tyson S. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, John Benjamins Publishing Company, Amsterdam and Philadelphia, pp. 3–33.

Granger S. (2003). "The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research", *TESOL Quarterly*, pp. 1–14.

Koutsis I., Markopoulos G. and 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 online at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/243_Paper.pdf.

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

Milton J. (1998). "Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment", in: Granger S. (Ed.), *Learner English on Computer*, Longman, London and New York, pp. 186–198.

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*, Peter Lang, Bern, pp. 451–462.