Christopher Waddington[*]

Should translations be assessed holistically or through error analysis?

Abstract

Teachers of translation use a variety of methods to evaluate their students' translations. This paper discusses two kinds of methods typically used at European universities, those based on error analysis and those based on a holistic approach. With the results of the research carried out for his Ph.D. thesis, the author examines the quality of these approaches and suggests possible improvements in the assessment of student translations.

1. Introduction

Teachers of translation use a variety of methods to evaluate their students' translations. A survey of these methods in European faculties of translation studies suggests that these can be broadly grouped into two categories: those based on error analysis and those based on a holistic approach, with some attempts to combine the two. This paper, which is based on research findings included in a Ph.D. thesis (Waddington 1999), examines the quality of these approaches when applied to the correction of a second-year exam of translation into the foreign language (Spanish-English).

2. Evaluation in Faculties of Translation

A questionnaire was sent to 48 European and Canadian universities which offer translation degree studies. The purpose of the survey was to find out about three main aspects of the translation examinations being used at these centres:

[*] *Christopher Waddington*
Universidad Pontificia Comillas
Quintana, 21
E-28008 Madrid

(1) the type of translation exam (whether it involved just translating a text or whether other types of test were used);

(2) the conditions under which the exam was carried out (the time available and whether students had access to reference books, etc.);

(3) how the student translations were corrected.

A total of 52 teachers replied from 20 of these universities and their answers to the third question reflected the following situation:

(1) 19 teachers (36.5%) use a method based on error analysis;

(2) 20 teachers (38,5%) use a holistic method;

(3) 12 teachers (23%) combine error analysis with a holistic appreciation.

3. Evaluation in translation studies

To date, research in the field of translation quality assessment has been mainly theoretical and descriptive, and has concentrated largely on the following themes:

(1) Establishing the criteria for a "good translation" (Darbelnet 1977, Newmark 1991).

(2) The nature of translation errors

· Defining the nature of translation errors as opposed to language errors (House 1981, Gouadec 1989, Nord 1993, Kussmaul 1995);

· Drawing up a catalogue of possible translation errors (Gouadec 1981);

· Establishing the relative, as opposed to absolute, nature of translation errors (Gouadec 1989, Williams 1989, Pym 1992, Kussmaul 1995);

· The need to assess quality not only at the linguistic but also the pragmatic level (Sager 1989, Williams 1989, Hewson 1995, Kussmaul 1995, Nord 1996, Hatim & Mason 1997).

(3) Basing quality assessment on text linguistic analysis (House 1981, Larose 1989).

(4) Establishing various textual levels on a hierarchical basis and linking the importance of mistakes to these levels (Dancette 1989, Larose 1989).

(5) Assessment based on the psycholinguistic theory of "scenes and frames" (Dancette 1989 & 1992, Bensoussan & Rosenhouse 1994, Snell-Hornby 1995).

(6) Attempts to elaborate scales to describe different levels of translation competence (Mahn 1989, Stansfield et al. 1992).

Stansfield et al. (1992) constitute an exception to the general rule of the publications cited above in the sense that it is the only one to present findings based on empirical research into translation quality assessment. Their article aims to "identify the variables that constitute translation ability" (Stansfield et al. 1992:455) and is based on work which the team carried out for the U.S. Federal Bureau of Investigation (FBI) to develop and validate job-related tests of translation ability. The article reports on their initial failure to find any mention in the literature on translation of research which could help them to "understand translation ability either as a psycholinguistic process or as a construct to be measured" (Stansfield et al. 1992:455).

Stansfield et al. claim that there are two basic translation skills: *Accuracy* and *Expression*. *Accuracy* has to do with the transfer of ST content to the TT; *Expression* has to do with form, the quality of linguistic aspects of the TT. Since this article was published, there has been increasing interest in looking beyond the assessment of the quality of a particular translation to the assessment of the underlying translator competence as reflected in the translation test. Campbell (1991) studies the results of a translation exam to see how far they evaluate the translation competence of the candidates and shed light on the translation processes followed. Hatim & Mason (1997) insist on the need to distinguish between translation quality assessment and translator performance assessment, and they proceed to draw up a chart of translation skills which is based on Bachman's (1990) analysis of communicative language ability and divides these into *Source Text Processing Skills*, *Transfer Skills* and *Target Text Processing Skills*. There are similar attempts to base models of translation competence on models of communicative language competence in Beeby Lonsdale (1996) and Bell (1991).

4. Justification of the choice of research area

This brief overview of research carried out in the area of translation quality assessment reveals an almost complete absence of empirical studies and explains why I chose my subject. However, translation quality assessment is a large area and the next question I have to answer is why I chose to research methods of assessment and not some other aspect.

Bachman (1990:40) lays down three steps which should be followed in the development of foreign language tests if the tester wishes to link the putative ability to the observed performance. These steps are as follows:

(1) Identifying and defining the construct theoretically (i.e. deciding exactly *what* is to be tested).

(2) Defining the construct operationally (i.e. deciding *how* it is to be tested).

(3) Establishing the procedure for quantifying observations (i.e. deciding the method of assessing the candidate's performance in the test).

It is difficult to refute the logic behind these three steps and, as they apply equally well to the development of translation tests, they clearly indicate that, if I wanted to improve testing procedures in our field, I should have devoted all my attention to the first one (identifying the construct of translation competence) and followed in the footsteps of Hatim & Mason (1997), Bell (1991) and Beeby Lonsdale (1996). The fact that I chose not to do so is due to two reasons. The first is that, although models of language competence have been widely researched for the last 20 years, there is as yet, to the best of my knowledge, no conclusive empirical evidence as to the precise nature of the various components it involves or their relative importance. The second reason is that, if this has proved so difficult to achieve in the case of linguistic competence, which is clearly composed of different skills, it will certainly prove even more difficult in the case of translation competence, which is also composed of different skills but ones which are interwoven so intimately that, when they are viewed separately, the underlying construct seems to evaporate.

This does not mean that I do not recognise the importance of and the need for research into the nature of the construct of translation competence, and especially its usefulness in syllabus construction and even formative evaluation (Hatim & Mason 1997:206, Beeby Lonsdale 1996:92-93). However, in the case of summative evaluation, which is the subject of this paper, and even in the case of the evaluation of translator competence into the foreign language, I felt that I had no alternative but to consider this competence as essentially unitary, in view of the almost complete lack of empirical research to support any other position.

So I decided to concentrate on Bachman's third step (defining the methods of assessment) and carry out empirical research into the reliability and validity of the different methods employed by university teachers as revealed in the survey reported above.

5. Description of the three methods[1]

5.1. Method A

Method A is the work of Hurtado Albir (1995); she draws up a list of possible errors which are divided into three categories:

(1) Inappropriate renderings which affect the understanding of the source text; these are divided into eight categories: *contresens*, *faux sens*, *nonsens*, addition, omission, unresolved extralinguistic references, loss of meaning, and inappropriate linguistic variation (register, style, dialect, etc.).

(2) Inappropriate renderings which affect expression in the target language; these are divided into five categories: spelling, grammar, lexical items, text, and style.

(3) Inadequate renderings which affect the transmission of either the main function or secondary functions of the source text.

In each of the categories a distinction is made between serious errors (-2 points) and minor errors (-1 point). There is a fourth category which

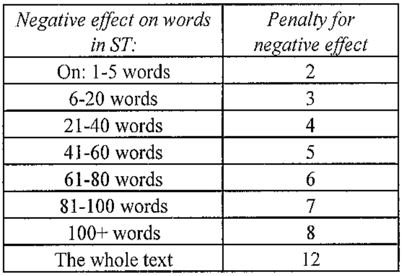
[1] The statistical analyses of the results of the application of Methods A and B, together with a study of the differences between these two error analysis methods, are recorded in Waddington (1999) and also in the paper "Measuring the Effect of Errors on Translation Quality" presented at the Saarbrücker Symposium on Translation and Interpretation: Models in Quality Assessment held at the Universität des Saarlandes 9th-11th March 2000.

describes the plus points to be awarded for good (+1 point) or exceptionally good solutions (+2 points) to translation problems. In the case of the translation exam where this method was used, the sum of the negative points was subtracted from a total of 110 and then divided by 11 to reach a mark from 0 to 10 (which is the normal Spanish system). For example, if a student gets a total of –66 points, his result would be calculated as follows: (110-66=44)/11=4 (which fails to pass; the lowest pass mark is 5).

5.2. Method B

The second analytical method was designed to take into account the negative effect of errors on the overall quality of the translations. The corrector first has to decide whether each mistake is a translation mistake or just a language mistake. This is done by deciding whether or not the mistake affects the transfer of meaning from the source to the target text: if it does not, it is a language error (and is penalised with –1 point); if it does, it is a translation error (and is penalised with –2 points). However, in the case of translation errors, the corrector also has to judge the importance of the negative effect that each one of these errors has on the translation, taking into consideration the objective and the target reader specified in the instructions to the translator in the exam paper. In order to judge this importance, the corrector is given the following table:

Table 1: Typology of errors in Method B



Negative effect on words in ST:	*Penalty for negative effect*
On: 1-5 words	2
6-20 words	3
21-40 words	4
41-60 words	5
61-80 words	6
81-100 words	7
100+ words	8
The whole text	12

The final mark for each translation is calculated in the same way as for Method A: that is to say, the examiner fixes a total number of positive points (in the case of method B, this was 85), and then the corrector subtracts the total number of negative points from this figure, and divides the result by 8.5. For example, if a student is given –30 points, his total mark would be 6.5 (pass): (85-30 = 55)/8.5 = 6.5.

5.3. Method C

Although in the survey mentioned in section 2 above, the teachers who answered were requested to send a brief description of the method of assessment they applied, and I did receive a number of descriptions of error analysis methods, I only received three descriptions of holistic methods. In addition to this, all three methods based their scales on the requirements of professional translation and were consequently of little use for judging the quality of translation into the foreign language. As a result, I had to design the holistic method myself. The design was based on the following principles:

(1) For the reasons laid out in section 3 above, I decided to use a unitary scale which treats the translation competence as a whole, rather than divide it into sub-scales reflecting different sub-competences (such as ST processing skills, transfer skills, TT processing skills).

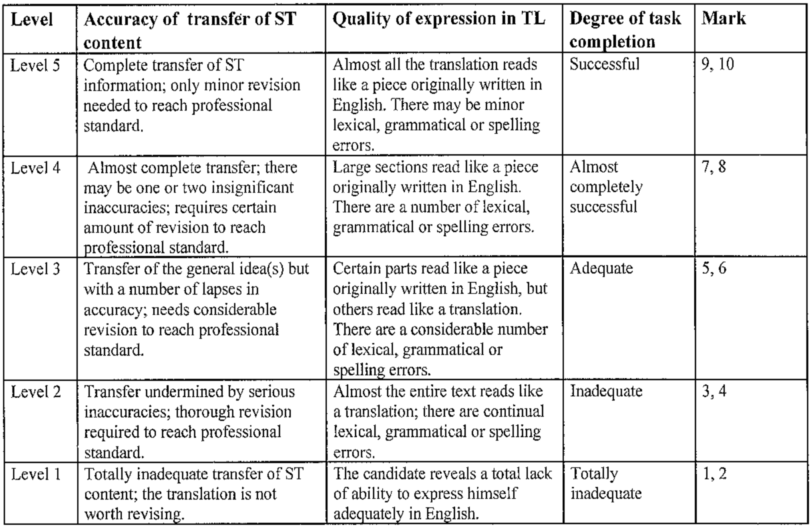
(2) It was important to write the descriptors in clear, simple language and avoid terminology that presupposes specialist knowledge (such as applied linguistics) on the part of the correctors.

(3) To achieve acceptable levels of reliability, it was important to limit the number of levels to a maximum of five[2] . However, in the end it was decided to include two marks within each level (for example 5 and 6), so that the correctors could use the traditional Spanish system of marking (from 0 to 10). And, when it came to applying the method, the correctors themselves asked to use half points (5.5, 6.5), and they were allowed to do so, as it would then prove easier to detect possible differences by their applications of this method.

In accordance with these principles, the following scale was drawn up:

[2] “In judging tests such as we are used to in writing and talking, to claim even 5 reliable bands within the range of ability that we observe, is optimistic.” (Pollitt 1991:90)

Table 2: Description of the five levels of the holistic Method C



Level	Accuracy of transfer of ST content	Quality of expression in TL	Degree of task completion	Mark
Level 5	Complete transfer of ST information; only minor revision needed to reach professional standard.	Almost all the translation reads like a piece originally written in English. There may be minor lexical, grammatical or spelling errors.	Successful	9, 10
Level 4	Almost complete transfer; there may be one or two insignificant inaccuracies; requires certain amount of revision to reach professional standard.	Large sections read like a piece originally written in English. There are a number of lexical, grammatical or spelling errors.	Almost completely successful	7, 8
Level 3	Transfer of the general idea(s) but with a number of lapses in accuracy; needs considerable revision to reach professional standard.	Certain parts read like a piece originally written in English, but others read like a translation. There are a considerable number of lexical, grammatical or spelling errors.	Adequate	5, 6
Level 2	Transfer undermined by serious inaccuracies; thorough revision required to reach professional standard.	Almost the entire text reads like a translation; there are continual lexical, grammatical or spelling errors.	Inadequate	3, 4
Level 1	Totally inadequate transfer of ST content; the translation is not worth revising.	The candidate reveals a total lack of ability to express himself adequately in English.	Totally inadequate	1, 2

Although the above scale is unitary for the reasons already expressed, I preferred to include three different aspects: the accuracy of transfer of ST content to the TT, the quality of expression in the TL and the degree of task completion. In this way, I hoped to help the correctors to judge the translations more consistently by giving them more complete and differentiated descriptors. If a particular student translation only partially fulfilled the requirements laid down by the descriptors at a certain level, then the corrector had to choose between the lowest mark at that level (for example, 7 at level 4) and the highest mark at the lower level (6 at level 3).

I decided to separate *accuracy of transfer* and *quality of expression* in view of the results of the research published by Stansfield et al. (1992), which claims to have empirically validated the existence of these two separate components of overall translation competence. It was also decided to include *degree of task completion* because the translation task used in the exam whose results form the basis of this research included clear instructions to the students in accordance with recommendations made by Nord (1991:164) and Hatim & Mason (1997:201).

6. Description of experiment

6.1. The two hypotheses

6.1.1. The first hypothesis

The first hypothesis was that *Methods of assessment based on error analysis are more reliable and valid than holistic methods*.

The main objective of this hypothesis was to analyse the differences between the marks achieved through the application of these two types of methods of assessing student translations, which, according to our survey, are both widely used by university teachers. I was most interested in possible differences in reliability, and this paper concentrates exclusively on the statistical results obtained in this area. I have omitted the validity studies partly because of lack of space and partly because they did not reveal significant differences between the three methods. The validity study was based on a number of external criteria consisting of:

- marks in other translation exams (both from English into Spanish, and from the other language combinations of the students into and from their mother tongue);
- marks in exams in English language and in Spanish language;
- the results of a questionnaire in which the students were required to evaluate their ability to translate from Spanish into English;
- the results of a survey among teachers (not only of translation but also of other areas in the degree) who had taught the whole group and who were asked to select the best and the worst students.

According to this study, all the methods proved to be equally valid.

6.1.2. The second hypothesis

The second hypothesis was that *The quality of a translation can be assessed more accurately if the method of assessment combines error analysis with a holistic appreciation*.

The argument in support of this hypothesis is the following: methods based on error analysis provide a clear justification of the mark reached, which is greatly appreciated by the students. The system of penalties is clear and its application is apparently objective. However, these methods have two drawbacks:

(1) Methods of error analysis are not as objective as they would appear. A translation error is not so much a question of right or wrong as of degrees of adequacy to the communicative context surrounding a particular communicative act. This means that the corrector's decisions in applying a method based on error analysis are inevitably subjective to a certain extent; the choice between what is appropriate and what is inappropriate depends at least in part on the corrector's personal judgement. As Pym (1992) nicely puts it: he is not distinguishing between black and white but between different shades of grey.

(2) Error analysis only measures the defects in a translation, but it does not measure positive aspects. Methods based on error analysis are founded on a possible fallacy: "the overall quality of a translation is equal to the sum of the defects it contains". There can be no doubt that mistakes undermine the quality of a translation, but it is also true that two translations with the same number of mistakes may vary in terms of overall quality.

To verify the hypothesis, I created Method D, which consisted of combining the marks obtained by the correctors in their application of Methods B and C, in a proportion of 70/30. These results were then compared to those reached with Method A.

6.2. How the methods were applied

To verify the two hypotheses, the three methods were applied to the correction of a second-year translation exam done by 64 students on the undergraduate degree course of Translation and Interpreting at the *Universidad Pontificia Comillas de Madrid*. The text of the exam paper which the students had to translate was an editorial from a Spanish newspaper (the *ABC*), entitled "Diálogo de la lengua", which discussed the present status of the Spanish language. The text was 330 words long and the students had 3 hours to translate it.

The three assessment methods were applied by five correctors; two of these were teachers of translation at Comillas, but the remaining three were teachers of English as a foreign language and had virtually no experience of translation teaching[3] . They applied the methods to the 64 translations in a different order and with at least a month's interval between each method. Before applying the methods to the exams, they had to practice each one on a series of other student translations of a different text (that is different from the text used in the exam, and different in the case of each method); each corrector, individually, then had to show me the results of this preliminary application, which were compared to the results of my own application, and differences and doubts were discussed. If both the corrector and I were satisfied, then I gave him/her the exams to correct; if doubts still persisted, the corrector was asked to apply the method to further translations of yet another different text[4] .

7. Results of the experiment

7.1. Table of results of the application of the three methods

Table 3 below shows the results of the application of the three methods to the student translations that were carried out by five correctors. In the table, the rows are the 64 students and the columns represent the 5 correctors who applied the three different methods. Hence "A1" = "corrector 1 applying Method A". At the foot of each column, the following data have been added: the totals, the means and the standard deviation:

[3] All three had at some time substituted me in my translation classes, but this only involved between four and ten hours' teaching in all in each case.

[4] This in fact only happened on two occasions, and I take this opportunity to mention that the correctors involved in this experiment received no compensation (economic or otherwise) for the considerable effort they had to make both in the training and in the application of the different methods of correction. Such is the life of the researcher and his friends!

Table 3: Results of the application of the three methods by the five correctors

STUDENTS	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
1	4	5.82	5.27	5.91	5.14	6.47	6.71	4.71	7.41	4.71	4.5	6	6	5	6
2	0.7	4.82	3.82	3.86	4.05	4.12	4.35	3.29	5.88	3.53	2.5	4	4	4	4
3	2.7	6.82	5.82	6.45	6.27	6.47	7.18	6.12	7.88	6.24	2.5	7	6.5	5.5	6
4	4.5	5.64	5.27	6.27	6.27	7.29	7.06	3.76	6.94	6.59	4.5	4	6	5.5	5.5
5	2.4	4.45	5.41	4.73	5.91	5.29	5.06	3.65	5.53	5.29	4.5	5	6	7	6
6	3.1	6.91	6.18	6.64	6.09	5.65	5.88	4.94	7.06	6.71	5	4	7	6.5	4
7	3	3.82	4.36	4.32	3.18	5.41	4.24	2.59	5.18	3.18	2.5	2	6	3	4
8	3.5	4.91	6.68	5.77	5.32	5.18	5.76	6.24	7.18	5.88	2.5	6	7	6.5	6
9	3.2	5.82	6.32	8.59	6.14	6.82	7.18	6.35	8	5.88	6	6	8	6.5	7.5
10	4.7	7.27	6.64	7.73	6.45	6.82	6.24	6	8.12	6.47	4.5	5	8.5	7	8
11	4	7.18	7.36	7.5	7.18	6.24	7.53	6	8.47	6.12	4.5	4	8	7	7
12	3.6	6.55	5.64	6.86	6.41	6.35	7.65	5.06	8	5.76	4.5	5	6.5	5	6
13	3.8	5.82	7	7	6.91	7.18	6.82	7.18	8.12	6.94	3	3	9.5	5	7.5
14	3.3	5.55	6.36	6.64	5.41	7.41	7.18	4.71	8	4.94	4.5	5	9	4	6.5
15	3.3	2.55	4.82	5.77	5.45	6.24	5.53	4.82	7.18	5.06	2	4	6	3	5.5
16	5.8	5.73	7.82	7.09	7.09	7.53	7.53	7.29	7.88	7.18	3	6	7.5	3	7.5
17	4.5	5	5.64	7.68	5.05	6.35	7.18	5.41	8.82	5.29	2	7	6.5	4	5.5
18	1.8	4	3.82	2.82	1.73	5.29	3.29	2.71	4	1.18	3	4	4	2	3.5
19	3	5.45	4.86	3.23	3.68	4.94	4.82	4.59	3.06	4.59	2.5	3	5	3	5
20	4.6	7.45	6.55	6	6.64	7.65	7.65	6.35	7.53	6.47	3	5	8.5	6	6.5
21	5.4	6.64	7.55	7.14	6	6.35	6.71	5.18	5.76	6.35	3	4	8.5	8.5	5.5
22	2.9	6	5.73	6.91	5.36	5.41	6.71	4.71	6.94	5.88	5	5	8.5	7	5.5
23	3.8	6.27	6.27	5.91	5.32	6.71	6.24	6	6	6.24	4.5	5	8.5	6	5.5
24	4.2	6.82	6.36	3.64	5.91	3.76	5.41	5.18	3.29	5.65	2.5	5	8.5	5	5
25	1.4	4.09	3.55	3.55	2.82	5.06	4.35	3.88	5.88	4.12	2.5	5	4	4	4.5
26	2.2	4.45	4.64	4.27	4.55	5.88	5.76	3.88	6.71	5.76	2	3	4.5	3	4
27	4.5	5.18	5.64	5.64	5.45	6	5.76	5.06	6.82	4.71	3	5	5	3	4.5
28	5.3	7.36	5.68	7.05	6.82	8.12	7.65	6.35	8.71	7.29	5	6	8.5	7	7.5
29	-0.1	3.82	2.27	4.45	2.5	3.76	4.24	2.71	4.59	2.59	4	5	6	3	4
30	1	4	3.41	3.64	2.36	4	4.94	2.71	4.59	3.65	5	4	4	3	4
31	5.4	8.64	8.36	8.64	8.32	7.65	7.88	7.65	8.59	8.24	5	8	9	8	9
32	3.8	7.82	5.82	6.55	5.45	6.71	6.71	6	8.71	6.47	5	8	8.5	6	6.5
33	5.1	8	6.77	7.5	6.5	7.65	7.88	6.59	8.12	7.41	5	7	9	7	7.5

34	4.7	5.82	4.91	6.36	5.45	6.82	5.18	5.29	6.59	6.24	5	6	6	4	6
35	2.9	5.82	5.27	4.95	5.86	6.94	5.76	4.47	6.71	6.59	5	5	6	6	6.5
36	4.6	6.45	6.09	5.91	6	7.18	6.59	6.47	7.65	6.82	4.5	6	7	5	6
37	4.7	6.09	6.14	5.14	5.05	6.82	6.47	5.29	7.53	4.82	6.5	6	7.5	6	5.5
38	5.8	9.27	7.95	8.09	8.64	7.88	8.35	7.41	9.29	8.71	5	8	9	9	7.5
39	2.7	6.45	5.82	6.68	6.05	6.94	6.94	4.82	7.88	5.65	5	5	8.5	7	5.5
40	3.6	7.18	5.82	6.09	6.36	6.35	7.29	5.76	6.94	6.71	7	6	8	6	6
41	3.2	5.91	6.27	5.64	6.09	6.24	6.71	5.76	6.24	5.88	5	5	8	4.5	5.5
42	6.5	7.64	7.73	6.82	7.09	8.71	7.06	6.82	7.29	7.53	7	7	9	5.5	7
43	2.4	3.91	5.23	3.09	2.55	6	2.82	4.35	2	2.12	5	4	6.5	2	4
44	3.5	5.45	5.91	6.05	5.73	6.12	6.71	4.71	7.65	5.88	7	5	7	4.5	6.5
45	6.1	4.18	5.45	4.05	2.55	5.65	5.41	4.82	5.88	3.76	5	4	6.5	3	5
46	6.4	7.09	6.59	6.82	8.09	7.88	6.94	6	8.82	7.65	5	5	6.5	6	7
47	2	5.45	4.68	5.36	4.45	5.53	5.41	3.65	7.41	4.24	7	6	6	5	5.5
48	3.3	5.82	5.36	6.73	5.82	4.94	6	4.35	6.59	6	5	7	6.5	4	6
49	6	8.82	7.05	8.73	7.55	7.65	8.12	6.47	8.47	7.65	8.8	7	9	9	8
50	5.2	7.82	7.77	8.09	7.64	7.53	7.53	7.29	7.41	7.41	5	7	9	7.5	7.5
51	3.8	7.18	5.68	6.91	6.55	6.24	7.06	4.47	7.88	6.12	5	6	8.5	5	7.5
52	5.5	7.55	6.14	6.59	7.68	7.76	7.76	5.65	7.65	7.76	5	7	9	4.5	7
53	1	4.64	3.64	3.55	4.09	4.59	4.94	2.82	5.88	3.88	4.5	5	7.5	4	5.5
54	4.2	7.55	5.95	7.91	6.59	6.94	7.18	5.53	8.12	6.35	3	6	8.5	8	6.5
55	5.3	6.91	6.68	6.55	6.64	7.18	6.94	6.47	8.12	5.88	4.5	7	6.5	7	6
56	2	5.09	5.14	4.82	4	6.12	5.88	3.65	5.76	4.47	4.5	5	5.5	6	5.5
57	5.6	7	7.18	7.55	7.73	7.41	8.12	6.82	8.47	7.53	5	7	9	8	8.5
58	7.5	8	8	7.18	8.82	8.47	7.06	7.88	8	8.59	6	7	9	9	8.5
59	4.9	6	5.86	5.95	4.64	7.41	6.94	5.65	6.94	4.47	4.5	8	8	4	6
60	2.7	4.45	5.27	5	5.09	5.88	5.18	5.65	6.59	5.76	3	6	6.5	3	5
61	3.5	5.55	6.05	4.73	3.45	5.88	6	4.82	5.76	3.53	3	6	6	3	5.5
62	5.6	3.36	4.82	6.23	4.5	6.47	4.94	4.12	6.12	4.24	3	5	5	4	5
63	2.2	1	3.05	3.36	1.64	3.65	4.94	1.76	3.88	2	3	5	6.5	3	4.5
64	2.6	3.64	3.77	2.82	3.23	3.88	3.88	3.29	0.35	4.47	5	5	4	2	4
TOTAL	244.4	377.7	369	379.5	353.4	404.8	401.2	330	434.8	361.1	279.3	349	453.5	333.5	380.5
MEAN	3.819	5.902	5.765	5.93	5.521	6.325	6.269	5.156	6.794	5.642	4.364	5.453	7.086	5.211	5.945
ST DEV	1.538	1.568	1.265	1.532	1.673	1.199	1.247	1.379	1.709	1.599	1.413	1.322	1.553	1.866	1.287

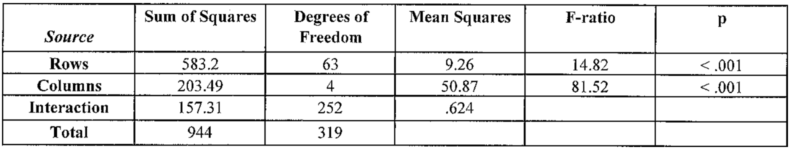
7.2. Analysis of variance (ANOVA) and reliability

An ANOVA for related measures was carried out on the results obtained by the application of the three methods by the five correctors in order to clarify the source of the variance detected. This analysis aimed at determining to what extent the variance could be attributed to differences between the two methods, to the fact that the correctors were applying them differently, or to the fact that the subjects (that is, the students) were different.

7.2.1. Method A

- Columns: 5 correctors assessing with Method A.
- Rows: Subjects (N=64).
- Variables: A1, A2, A3, A4, A5.

Table 4: ANOVA of the results of the application of Method A

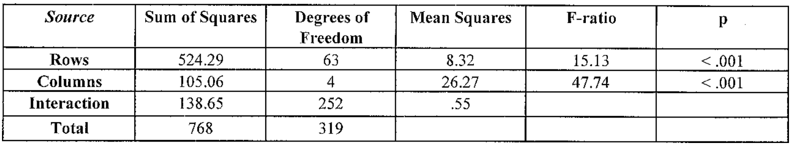


Source	Sum of Squares	Degrees of Freedom	Mean Squares	F-ratio	p
Rows	583.2	63	9.26	14.82	< .001
Columns	203.49	4	50.87	81.52	< .001
Interaction	157.31	252	.624		
Total	944	319			

7.2.2. Method B

- Columns: 5 correctors assessing with Method B.
- Rows: Subjects (N=64).
- Variables: B1, B2, B3, B4, B5.

Table 5: ANOVA of the results of the application of Method B



Source	Sum of Squares	Degrees of Freedom	Mean Squares	F-ratio	p
Rows	524.29	63	8.32	15.13	< .001
Columns	105.06	4	26.27	47.74	< .001
Interaction	138.65	252	.55		
Total	768	319			

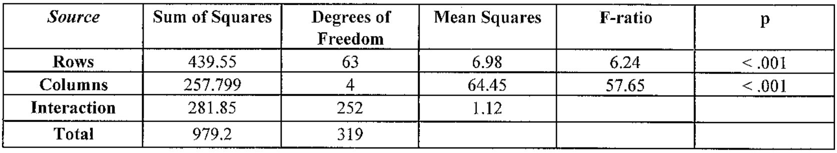
7.2.3. Method C

· Columns: 5 correctors assessing with Method C.

· Rows: Subjects (N=64).

· Variables: C1, C2, C3, C4, C5.

Table 6: ANOVA of the results of the application of Method C



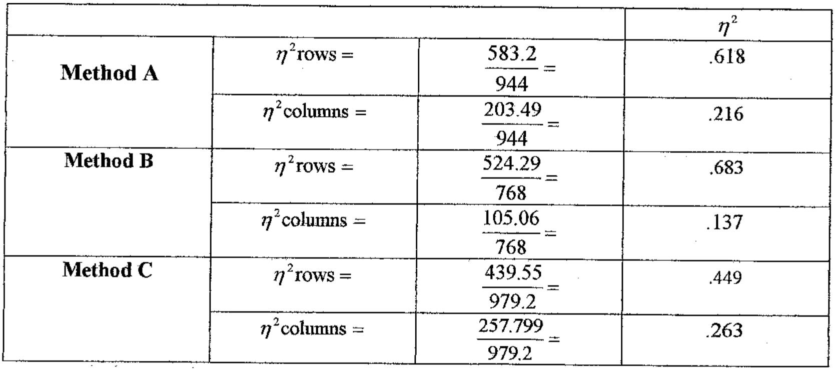
Source	Sum of Squares	Degrees of Freedom	Mean Squares	F-ratio	p
Rows	439.55	63	6.98	6.24	< .001
Columns	257.799	4	64.45	57.65	< .001
Interaction	281.85	252	1.12		
Total	979.2	319			

7.2.4. Interpretation of the results of the ANOVAs

In the case of all three methods, the F-ratio for both the columns (that is, the correctors applying the method) and the rows (that is, the 64 students) is significant (p<.001). This means that the differences that can be observed between the results are not only due to the fact that the students are different, but also to differences between the correctors in their application of each method. However, these significant F-ratios do not tell us the size of these differences. To measure this, we applied the coefficient of eta squared (η^2), which is the result of dividing each partial sum of squares by the total sum of squares, and which shows the proportion of the variance that can be attributed to each of its possible sources. This is a useful complement to the ANOVA as it helps to interpret the results more accurately (Nunnally & Bernstein 1994).

The calculation of the coefficients of eta squared gives us the following results for the three methods:

Table 7: Coefficients of eta squared for Methods A, B and C



			η^2
Method A	η^2rows =	$\frac{583.2}{944} =$	.618
	η^2columns =	$\frac{203.49}{944} =$	.216
Method B	η^2rows =	$\frac{524.29}{768} =$	.683
	η^2columns =	$\frac{105.06}{768} =$	.137
Method C	η^2rows =	$\frac{439.55}{979.2} =$	.449
	η^2columns =	$\frac{257.799}{979.2} =$	.263

According to the above results, the variance in the rows (the students) accounts for 62% of the total variance in the case of Method A, 68% in the case of Method B and 45% in the case of Method C. In contrast to these results, the variance in the columns (the correctors applying the methods) accounts for 22% of the total variance in the case of Method A, 14% in the case of Method B and 26% in the case of Method C. These results lead to the following conclusions:

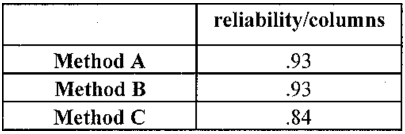
(1) In spite of the variance detected in both the columns and the rows, the differences between the students account for a considerably greater proportion of the total variance than the differences between the correctors.

(2) Although the differences between the students are greater than the differences between the teachers applying the three methods, the comparison of the eta squared coefficients obtained by each method indicates the superiority of the two error analysis methods (A and B) over the holistic one (C).

7.2.5. Results of the inter-rater reliability estimates

The results of the ANOVAs were used to find the inter-rater reliability (reliability of the columns) with the three methods, with the following results:

Table 8: Estimate of reliability of Methods A, B and C



	reliability/columns
Method A	.93
Method B	.93
Method C	.84

The results of the reliability estimates also indicate the superiority of the methods based on error analysis. These results show us that Methods A and B are more consistent than Method C (.94 as against .84), at least with this group of students, these translations and these correctors. Morales (2000) considers that the adequate level of reliability depends above all on the use that is going to be made of the marks obtained. If the marks are going to be used as a basis for decision taking, then Morales recommends that the reliability coefficient should be at least .85. Although he admits that this figure is arbitrary, it is worth noting that, whereas the coefficient for the methods based on error analysis is fully acceptable, the one for the holistic method is just below Morales' minimum[5]. This implies that we should perhaps be careful of basing important academic decisions on marks achieved by means of holistic methods.

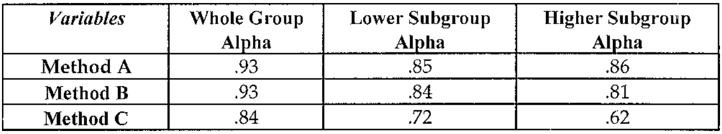
7.2.6. Inter-rater reliability estimates with the best and the worst students

To find out whether the internal consistency of the methods and the correctors varied with the best and the worst students, the total number of 64 students was divided into two subgroups of 32 (lower subgroup and higher subgroup) according to the median of the average mark obtained by the students with the three methods (A, B and C) as applied by the

[5] Hughes (1989:32) comments that in foreign language tests the level of reliability that can be achieved varies according to the nature of the test. He quotes Lado (1961) who claims that good vocabulary, structure and reading tests are usually in the .90 to .99 range, while auditory comprehension tests are more often in the .80 to .89 range, and oral production tests may be in the .70 to .79 range. This means that a reliability coefficient of .85 might be considered high for an oral production test but low for a reading test.

five correctors. A reliability estimate (Cronbach's alpha) was carried out on each of these two groups and the results are laid out in table 9:

Table 9: Reliability estimate with the lower subgroup and the higher subgroup.



Variables	Whole Group Alpha	Lower Subgroup Alpha	Higher Subgroup Alpha
Method A	.93	.85	.86
Method B	.93	.84	.81
Method C	.84	.72	.62

The reliability coefficient is higher for all three methods with the whole group of 64 students than with the two subgroups. This is to be expected, as the reliability coefficient is always lower when it is applied to more homogeneous groups.

What is really worth pointing out in this table, is the fact that the error analysis methods differentiate equally well between the students in the two subgroups, both the lower one and the higher one. However, the holistic method differentiates more clearly between the students in the lower subgroup than between those in the higher subgroup. This indicates that, in this upper subgroup there are differences between the students which are detected by the error analysis methods but not by the holistic one.

In conclusion, the statistical analysis of the results of the application of the three methods by the five correctors to the student translations confirms at least part of our first hypothesis, that methods based on error analysis are more reliable than holistic ones.

8. Verification of the second hypothesis

8.1. ANOVA with Methods A and D

The second hypothesis was that the quality of a translation can be assessed more accurately if the method of assessment combines error analysis with a holistic appreciation. To verify this hypothesis, Method D was created: this consisted of combining the marks obtained by the correctors in their application of Methods B and C, in a proportion of 70/30. These results were then compared to those reached with Method A.

An ANOVA for related measures was carried out on the marks obtained with Method D, in order to ascertain the source of the variance. The results of this ANOVA are set out in table 10:

Method D

- Columns: 5 correctors assessing with Method D.
- Rows: Subjects (N=64).
- Variables: D1, D2, D3, D4, D5.

Table 10: ANOVA of the results of the application of Method D

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Squares**	**F-ratio**	**p**
Rows	460.792	63	7.314	17.125	<.001
Columns	17.427	4	4.357	10.201	<.001
Interaction	107.626	252	.427		
Total	585.846	319			

As was the case with the other three methods, the ANOVA with Method D also shows significant variance in the rows (the students) and in the columns (the correctors). In the same way as before, the coefficient of eta squared was applied to these results to ascertain the size of this variance. In the following table we also repeat the eta squared coefficient for Method A:

Table 11: Coefficients of eta squared for Methods A and D

			η^2
Method A	η^2rows =	$\frac{583.2}{944} =$	.618
	η^2columns =	$\frac{203.49}{944} =$	.216
Method D	η^2rows =	$\frac{460.792}{585.846} =$	.787
	η^2columns =	$\frac{17.427}{585.846} =$	.03

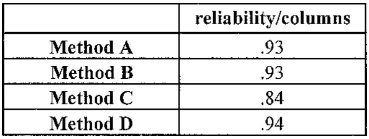
These results provide clear evidence of the superiority of Method D over Method A: in the application of the combined error analysis-holistic Method D, the differences between the correctors are considerably less than with error analysis Method A. Whereas, in the case of Method

A, the differences between the students account for 62% and the differences between the correctors 22% of the total variance, in the case of Method D, the differences between the students account for 79% and the differences between the correctors only 3% of the total variance.

8.2. Results of the inter-rater reliability estimates

The results of the ANOVA were used to calculate the inter-rater reliability (reliability of the columns) with Method D, which is shown in the following table, where we also repeat the reliability coefficients for the other three methods:

Table 12: Estimate of reliability of Methods A, B, C and D



	reliability/columns
Method A	.93
Method B	.93
Method C	.84
Method D	.94

The superiority of Method D over Method A is also reflected in the results of this reliability estimate which shows that, with the combined method, the correctors achieve a coefficient of .94 as against .93 with Method A. Although this is a small difference, it contributes towards the verification of the second hypothesis, especially if we take into account the fact that the combination of the error analysis Method B and the holistic Method C greatly improves the latter’s reliability, which was only .84 when used on its own.

9. Conclusions

In spite of the limited nature of this piece of research, it points to two conclusions:

(1) This research clearly indicates that, provided there is a minimum of coordination between correctors, the results obtained by the application of the type of systems currently used for evaluating student translations achieve a high level of internal consistency between raters.

(2) Although the statistical analysis indicates the superiority of methods based on error analysis over those based on a purely holistic appreciation, it also shows the limitations of error analysis by itself, and the benefits of combining both approaches.

This research was limited to translation into the foreign language, and the findings should also be compared to the application of this experiment to translation into the native language, both in academic and professional spheres. What is clear is the continuing need for closer statistical scrutiny of all aspects of Translation Quality Assessment.

10. Bibliography

Bachman, Lyle F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Beeby Lonsdale, Allison (1996). *Teaching Translation from Spanish to English*. Ottawa: University of Ottawa Press.

Bell, Roger T. (1991). *Translation and Translating: Theory and Practice*. London: Longman.

Bensoussan, Marsha & Rosenhouse, Judith (1994). Evaluating student translations by discourse analysis. In *Babel* 36 (2). 65-84.

Campbell, Stuart J. (1991). Towards a Model of Translation Competence. In *Meta* XXXVI (2/3). 329-243.

Dancette, Jeanne (1989). La faute de sens en traduction. In *Traduction, Terminologie, Rédaction* 2 (2). 83-102.

Dancette, Jeanne (1992). Modèles sémantique et propositionnel de l'analyse de la fidelité en traduction. In *Meta* XXXVII (3). 440-449.

Darbelnet, Jean (1977). Niveaux de la traduction. *Babel* 23 (1): 6-17.

Gouadec, Daniel (1989). Comprendre, évaluer, prévenir. Pratique, enseignement et recherche face à l'erreur et à la faute en traduction. In *TTR (Traduction, Terminologie, Rédaction*) 2 (2). 35-54.

Gouadec, Daniel (1981). Paramètres de l'évaluation des traductions. In *Meta* XXVI (2). 99-116.

Hatim, Basil & Mason, Ian (1997). *The Translator as Communicator.* London: Routledge.

Hewson, Lance (1995). Detecting Cultural Shifts: Some Notes on Translation Assessment. In Mason, Ian & Pagnoulle, Christine (eds.) (1995). *Cross-Words. Issues and Debates in Literary and Non-literary Translating*. Liège: L3 – Liège Language and Literature. 101-108.

House, Juliane (1981). *A Model for Translation Quality Assessment*. Tübingen: Narr.

Hughes, Arthur (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Hurtado Albir, Amparo (1995). La didáctica de la traducción. Evolución y estado actual. In Fernández, P. (ed.) (1995). *Perspectivas de la Traducción*. Universidad de Valladolid. 49-74.

Kussmaul, Paul (1995). *Training the Translator*. Amsterdam: Benjamins.

Lado, Robert (1961). Language Testing. The Construction and Use of Foreign Language Tests. London: Longmans.

Larose, Robert (1989). *Théories Contemporaines de la Traduction*. Montréal: Presses de l'Université de Québec.

Mahn, Gabriela (1989). Standards and Evaluation in Translator Training. In Krawutschke, Peter W. (ed.) (1989). *Translator and Interpreter Training and Foreign Language Pedagogy*. American Translators Association Series, Vol. 3. New York: SUNY Binghamton Press. 100-108.

Morales, Pedro (2000). *Medición de actitudes en Psicología y Educación*. Segunda edición revisada. Madrid: Universidad Pontificia Comillas.

Newmark, Peter (1991). *About Translation*. Clevedon (UK): Multilingual Matters Ltd.

Nord, Christiane (1991). *Text Analysis in Translation*. Amsterdam: Rodopi.

Nord, Christiane (1993). *La evaluación de errores en la enseñanza de traducción*. Summary of postgraduate course given in the Universidad de Las Palmas (Canarias).

Nord, Christiane (1996). El error en la traducción: categorías y evaluación. In Hurtado Albir, A. (ed.) (1996). *La enseñanza de la traducción*. Collecció Estudis sobre la traducció Núm. 3. Castelló de la Plana: Publicacions de la Universitat Jaume I. 91-103.

Nunnally, Jum C. & Bernstein, Ira H. (1994). Psychometric Theory. 3rd edition, New York: McGraw-Hill.

Pollitt, Alastair (1991). Response to Charles Alderson's Paper 'Bands and Scores'. In Alderson, Charles & North, Brian (eds.) (1991). *Language Testing in the 1990s*. London: Macmillan. 87-94.

Pym, Anthony (1992). Translation Error Analysis and the Interface with Language Teaching. In Dollerup C., and Loddegaard, A. (eds.) (1992). *Teaching Translation and Interpreting. Training, Talent and Experience. Papers from the First Language International Conference, Elsinore, Denmark, 31 May - 2 June, 1991.* Amsterdam: Benjamins. 279-288.

Sager, Juan C. (1989). Quality and standards - the evaluation of translations. In Picken, C. (ed.) (1989). *The Translator's Handbook*. London: ASLIB. 91-102.

Snell-Hornby, Mary (1995). On Models and Structures and Target Text Cultures: Methods of Assessing Literary Translations. In Josep Marco Borillo (ed.) (1995). *La Traducció Literària*. Collecció Estudis sobre la traducció Núm. 2. Castelló de la Plana: Publicacions de la Universitat Jaume I. 43-58.

Stansfield, Charles W., Scott, Mary Lee and Kenyon, Dorry Mann (1992). The Measurement of Translation Ability. In *The Modern Language Journal* 76 (iv). 455-67.

Waddington, Christopher (1999). *Estudio comparativo de diferentes métodos de evaluación de traducción general (Inglés-Español).* Madrid: Publicaciones de la Universidad Pontificia Comillas.

Williams, M. (1989). The Assessment of Professional Translation Quality: Creating Credibility out of Chaos. In *TTR (Traduction, terminologie, Rédaction*) 2 (2). 13-33.