

Conde, T. (2011). Translation evaluation on the surface of texts: a preliminary analysis. *The Journal of Specialised Translation*, 15, 69-86. <https://doi.org/10.26034/cm.jostrans.2011.506>

This article is publish under a *Creative Commons Attribution 4.0 International* (CC BY):
<https://creativecommons.org/licenses/by/4.0>



© Tomas Conde, 2011

Translation evaluation on the surface of texts: a preliminary analysis

Tomás Conde, Departament de Traducció i Comunicació, Universitat Jaume I

ABSTRACT

The article introduces the relationship between translation evaluation and certain features on the surface of texts. The process and results from the evaluation of 48 texts by 88 subjects were analysed to explore the following aspects: emphasised words, actions on emphasised words, text length, resemblance to original format, lexical density and readability. Analyses were conducted to examine whether these aspects may have led to an increase in evaluators' attention levels. Data were taken from a previous study and completed with new analyses, some of which were carried out with Wordsmith Tools software. Two content-independent features emerged as significant in the translation evaluation: consistency in the decisions taken and, more especially, productivity, measured by the successful translation of as much of the text as possible. Although the results should be considered preliminary, they may inspire other studies involving control groups and enhanced experimental conditions. Much still remains to be explored in the fields of attention and translation evaluation.

KEYWORDS

Translation evaluation, translation quality, evaluation parameters, attention, cognition, errors, lexical density, readability

1. Introduction

Attention processes—and their effect on individuals—permeate many aspects of everyday life. Marketing experts are all too well aware of this:

[...] location *within* a store can make the difference between success and failure of a product, which is why manufacturers fight so hard for the right spot on supermarket shelves. Typically, the larger and more powerful grocery manufacturers such as Sara Lee, Kellogg, and General Mills get the most visible spots. (Boone et al. 2009: 329)

On the internet, successful bloggers have learned where to place advertising in order to make more money from their readers' clicks. And in television, aspects that most attract the attention of

potential consumers, as well as strategies to captivate their attention, are widely studied.

It has been suggested that when evaluating translations, a latent, somehow unconscious idea of the text quality—which makes it possible to evaluate texts holistically—is as crucial as the more usual and complex analytical systems (Conde 2009, Waddington 1999). In light of this, and because increased attention improves the neural activation of the segments read, it may be assumed that the marks and corrections performed on the emphasised segments have a greater impact on the impression formed by evaluators of the translation's quality. And in turn, this effect could lead to a more or less favourable judgment (depending on whether the event detected is a correct decision or an error).

In order to verify this hypothesis, assessments of the quality of 48 translations made by a group of evaluators were contrasted with certain content-independent aspects, as revealed in the result of their evaluation process. The specific aim was to determine whether these aspects may have an effect, mostly negative,¹ on the result of the evaluation: it could occur, for example, that texts with a higher percentage of emphasised words were given lower grades.

1.1. Emphasised text

Translation evaluation has traditionally been based on error detection. Not surprisingly, analytical systems are still the most widely used, especially within teaching environments. These systems consist of counting and assessing errors in a given translation. That is, errors are *counted*, but also *assessed* or characterised, and the two most common criteria for this characterisation are *nature* and *importance*. This article focuses primarily on the latter.

In Translation Studies, the distinction of an error according to its importance has received much research attention (Ceschin 2004, Darwish 2001, Rosenmund 2001, Vollmar 2001, Koo & Kinds 2000). In all cases, there is a—usually simple—hierarchy with the most serious errors at the top, which are sometimes referred to as *critical*. According to some authors (Cruces 2001: 816, Martinez & Hurtado 2001, Sager 1989 in Waddington 1999: 35-36, Larose 1998: 16), what matters is the error's effect or impact on the entire text.

Other researchers relate error importance directly to its location. Vollmar (2001: 26) considers errors that lead to the misinterpretation of significant portions of the text to be critical.

Also pertinent are the contributions of Hajdú (2002: 249) and House (2001: 151). Hajdú considers the importance of an error to depend on the part of the text in which it is found, and specifically mentions section headings. House regards errors appearing in titles, addresses, phone numbers and indexes as critical errors. In practice, and in many texts, this means the most serious errors are those made in segments that vary typographically from conventional typography, namely, that used in the body text.

This question has also been raised in more specialised fields. Thus in the area of IT, critical errors in localisation products are usually those appearing in the most visible parts of the software (Ceschin 2004: 90-91), which directly links visibility and error importance. All the above comments reflect authors' personal opinions on what they consider to be the most serious errors. Whether these beliefs coincide with reality, and whether translation evaluators pay more attention to—and therefore penalise more heavily—translations that contain a higher percentage of errors in the emphasised segments are questions that need to be addressed.

1.2. Text length and layout

Error location is not the only aspect that may attract evaluators' attention. Other, perhaps more general, features may play a significant role in their assessment of quality.

The length of the translation, particularly during serial evaluation sessions (such as those inevitably faced by lecturers in university translation departments), may affect evaluation either consciously or otherwise. These evaluators may be surprised by the fact that one translation is significantly shorter than the others in a given set, and may then believe that the translator or translators were not efficient enough, or at least, less efficient than their colleagues—who, presumably, would have been given the same time to accomplish the task. The contrasting point of view would be that a translator or group of translators who can translate a longer text in the same time as their peers would surely be considered more efficient, which could result in a higher grade from the evaluator.

Furthermore, an evaluator may notice, perhaps unconsciously, the degree of effectiveness with which the translator reproduces certain aspects of the original format, such as bold type and italics, font colour and size, spacing, indentation, margins, justification and so forth. In an educational psychology study, Mangal (2007: 342-345) defines two types of factors that may have an impact on attention: internal and external. Internal factors are essentially concerned with the subject's predisposition to receive stimuli, while external factors

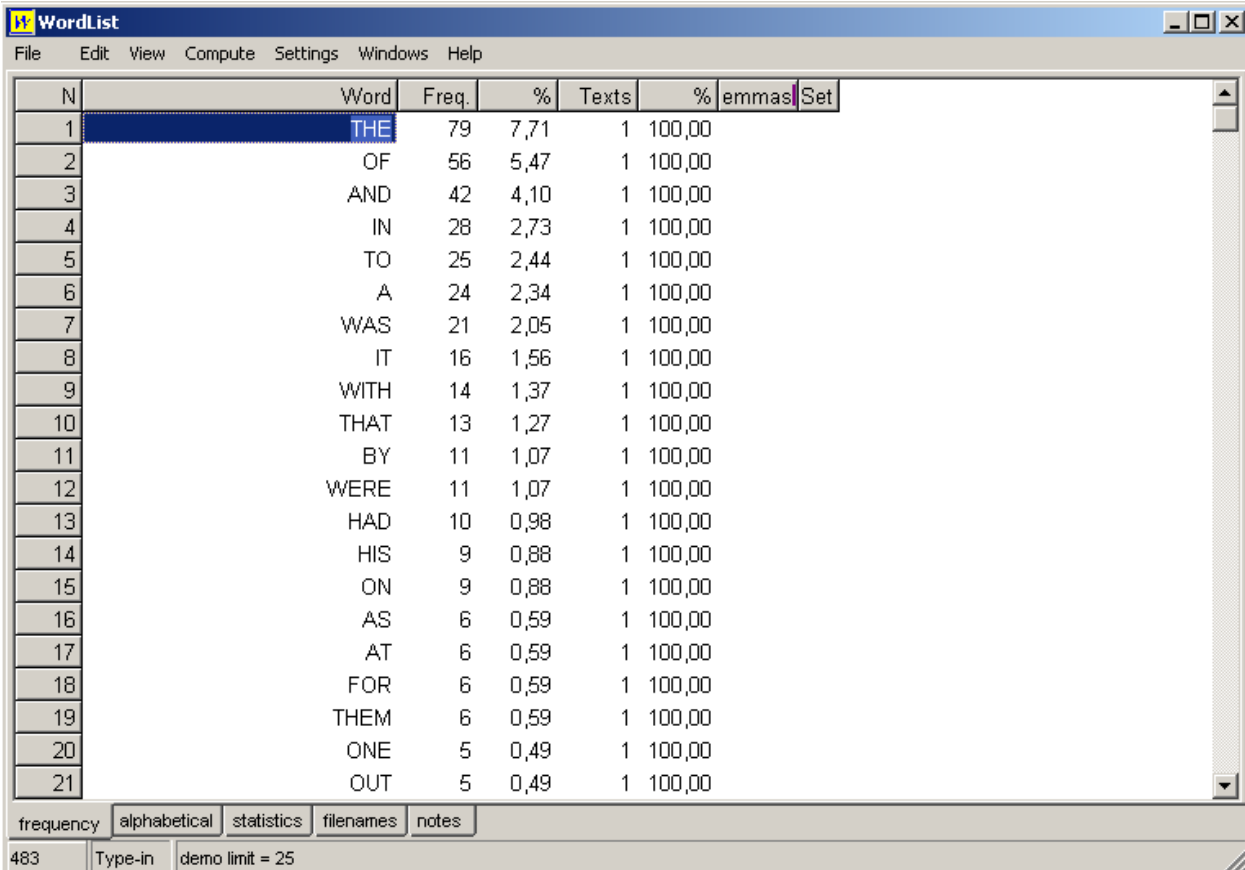
would include the nature of the stimulus, its intensity and size, as well as contrast, change, variety, repetition and movement. Although some of these factors are not directly related to the location or visibility of certain parts of the texts, others—like contrast, change, intensity or size—do show a clear link with typographical emphasis.

Therefore, regardless of the translator's ability to reconstruct the text content, if the original format is not faithfully reproduced, the evaluator may judge the translation's quality less favourably. It is worth noting that not all evaluators use the error analysis method (which usually detects defective resemblance of typography and spelling issues); in the case of holistic evaluators, therefore, this aspect will only be reflected if their evaluations are clearly consistent in this respect.

1.3. Lexical density and readability

A further aspect of translated texts that may attract serial evaluators' attention is lexical density. In order to raise text quality in general (not only translations), Peha (2003) recommends taking particular care, for example, over the choice of vocabulary: texts should include a range of action verbs, as well as creative adjectives and adverbs. In other words, quality relates directly to the variety of registers and vocabulary; in contrast, the repetition of words and structures, and also redundancy, are normally associated with novice writers. The only exception to this may be the field of technical writing, where repeated structures are accepted, since content takes precedence over form.

In recent years, computer tools have gone some way to facilitating analysis of lexical richness. Despite their shortcomings (for example, verb conjugations and variations in gender and number are classified as different words), tools such as Wordsmith Tools have proved useful for quantitative research, not only in translation, but also in different branches of linguistics and the study of language in general. The software automatically generates frequency word lists (see Figure 1) and percentages of both total words (*tokens*) and distinct words (*types*) from the base text file: a high type/token ratio is taken to indicate a wider range of vocabulary (Munday 1998, Baker 1995).



N	Word	Freq.	%	Texts	% emmas	Set
1	THE	79	7,71	1	100,00	
2	OF	56	5,47	1	100,00	
3	AND	42	4,10	1	100,00	
4	IN	28	2,73	1	100,00	
5	TO	25	2,44	1	100,00	
6	A	24	2,34	1	100,00	
7	WAS	21	2,05	1	100,00	
8	IT	16	1,56	1	100,00	
9	WITH	14	1,37	1	100,00	
10	THAT	13	1,27	1	100,00	
11	BY	11	1,07	1	100,00	
12	WERE	11	1,07	1	100,00	
13	HAD	10	0,98	1	100,00	
14	HIS	9	0,88	1	100,00	
15	ON	9	0,88	1	100,00	
16	AS	6	0,59	1	100,00	
17	AT	6	0,59	1	100,00	
18	FOR	6	0,59	1	100,00	
19	THEM	6	0,59	1	100,00	
20	ONE	5	0,49	1	100,00	
21	OUT	5	0,49	1	100,00	

frequency alphabetical statistics filenames notes

483 Type-in demo limit = 25

Figure 1: Wordlist in Wordsmith Tools

Moreover, the software displays the average length of words and sentences. These counts are an objective way of measuring text readability, and improve on readability indexes—such as Flesch Reading and Flesch-Kincaid Grade Level—which are based on arbitrary formulas. In sum, Wordsmith Tools is a valuable application for studies such as the present one, the materials and methods of which are discussed below.

2. Materials and methods

The data for this study were taken from data gathered during research for the PhD thesis *Proceso y resultado de la evaluación de traducciones* (Conde 2009). The analysis of these data has now been extended to carefully examine the extent to which certain aspects on the surface of texts can affect evaluation.

The analysis began with an evaluation carried out by four groups of subjects: a total of 88 evaluators made up of translation students (25), potential addressees of the texts (40), professional translators (13) and translation teachers (10). They were asked to evaluate the quality of 48 Spanish translations, divided into four sets of 12 texts, corresponding to four original English texts: two originals were taken from *The Economist*, one an extensive article (DP3) and the

second a series of brief news items (DP1); the other two originals (CT2 and CT4)—various messages from a specialised internet-based forum—dealt with industrial painting procedures. It should be noted that CT stands for *Comunicación técnica* (technical communication) and DP for *Divulgación política* (political texts for a wide readership). Sets were administered in the following order: DP1, CT2, DP3 and CT4; i.e. alternating sets of different topics, thereby encouraging evaluators to see each set as a complete task².

The texts had been translated by translation students on a previous course and—to stimulate the evaluative activity—were chosen by the lecturer as being representative of low or medium-low quality levels. Evaluators were given just three instructions: (1) to work on each set in one sitting, (2) to follow the set order, and (3) to classify the quality of the translations as very bad, bad, good and very good.³ Apart from these indications, they were free to evaluate as they pleased (for example, on-screen or on printed copies of the translations). In line with previous research (Conde 2009), various behaviours were observed; some evaluators used detailed linguistic analysis while others followed other less thorough, holistic methods.

One of the key concepts of the analysis was the *action*; this refers to any activity the evaluators perform on the text (as reflected in the text, either as a mark, sign or comment), to point out the existence of an error, a correct decision or any other feature present in or absent from the translation. In Conde (2009), formal aspects were only briefly considered, with the exception of analysing actions taken on typographically emphasised segments, which did not appear to affect overall judgments on the texts (Conde 2009: 441).⁴ The percentage of actions carried out on emphasised as compared to non-emphasised segments was only of note among the potential addressee evaluator group (Conde 2009: 342); however, the evaluations made by this group tended to be more concise and include general comments indicating the overall text quality in the titles, i.e. the emphasised segments.

Based on the above arguments, our research hypothesis is as follows:

Certain content-independent aspects of translations affect the outcome of the evaluation they are subject to.

To test the hypothesis, the idea briefly raised in Conde (2009) is further explored, namely, to observe actions performed on typographically emphasised segments. In addition, the fact of including more emphasised segments (regardless of the actions performed on them) might have some effect on the evaluation by virtue of the increased level of attention they can stimulate. Other

content-independent aspects that may affect the evaluators' assessment are the overall length of the translations, the extent to which typographical aspects (bold, spacing, etc) are reproduced, or the lexical density of each text.

The database created for the above-mentioned doctoral thesis was used to test the hypothesis and apply these parameters; in addition, the study was extended with new analyses and procedures conducted primarily with the SPSS v17 (statistics) and MS Excel 2007 (figures) software packages. Finally, data on lexical density were obtained using Wordsmith Tools.

3. Results and discussion

As in § 1, the following results contrast the translations' average quality judgment with the extent to which the texts were emphasised, their length, formal resemblance to the original text and, finally, with their lexical density and readability.

3.1. Emphasised vs non-emphasised text

Unlike body text—which often uses a standard, medium-size font (usually between 10 and 12 points)—titles, captions and footers are usually emphasised by means of boldface, italics, underlining, or other special typographical means. This section discusses the possible effects of this emphasis on the evaluator's overall assessment, and consists of three parts: the impact of the emphasised segments *per se*, the weight of the actions carried out in these segments and, finally, the analysis of a specific example.

3.1.1. Emphasised words

Text evaluation can be affected by the percentage of emphasised words contained in the text. All the translations had a similar percentage of emphasised words with the exception of three texts, which were deemed to be statistically atypical, i.e. outliers that might distort the mean values:

- Two texts had a much lower percentage of emphasised words: T20 (in DP3) and T44 (CT4).
- One text had a significantly higher percentage of emphasised words: T45 (in CT4).

Figure 2 shows the average quality judgments of the texts, according to whether they included a higher (grey bar), average (pink) or lower (blue) percentage of emphasised words.

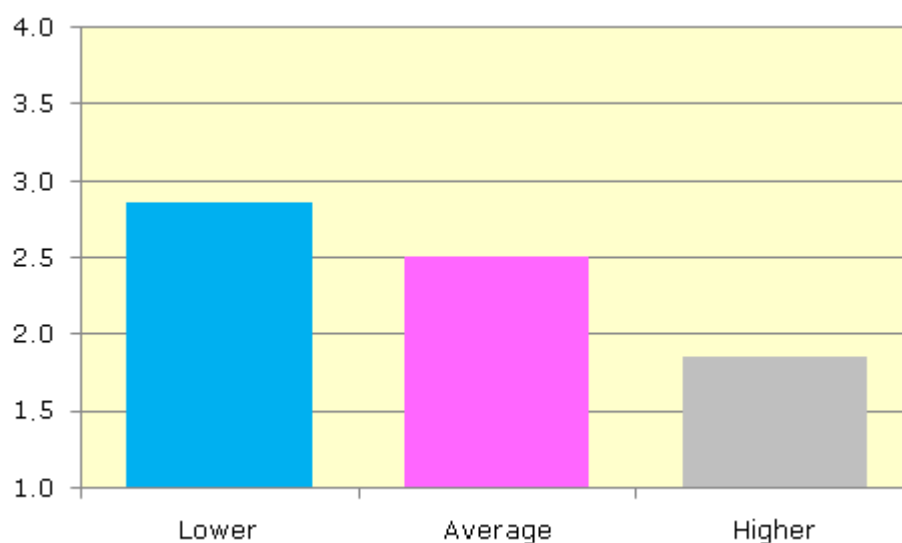


Figure 2: Quality judgment and percentage of emphasised words

Texts with a lower percentage of emphasised words were given the highest grades, whereas those with more emphasised words obtained the lowest scores. This result appears paradoxical since it might be expected that the translations with the most emphasis would most closely reproduce the original format (translations should not have more emphasised text than their originals). One possible explanation may be that a lower number of emphasised segments resulted in a reduction in the evaluators' attention levels and, as a consequence, they had noticed fewer errors. However, the groups of translations that generated these data are very unbalanced in number (1, 2 and 45 translations), and for this reason a new analysis was performed.

In this instance, non-atypical translations were divided into four groups, according to the percentage of emphasised words they contained, and based on the quartiles⁵ of each set. Figure 3 presents the four groups of translations' average judgments (1 represents the translations with the lowest number of emphasised words and 4, those with the highest).

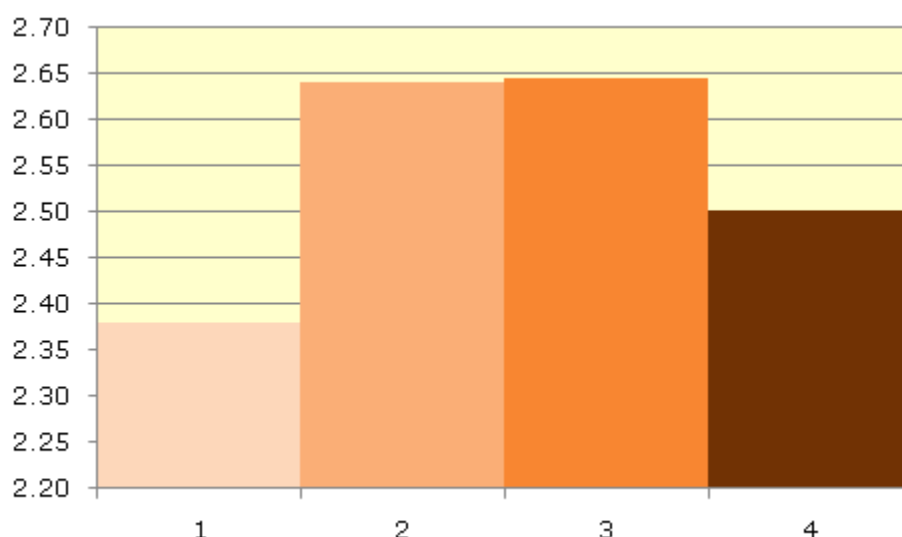


Figure 3: Quality judgment and emphasised words in non-atypical translations

As the two intermediate groups were those with a higher average judgment, no significant effects were noted. The absence of clear phenomena may be explained by the obvious similarity of these four groups of non-atypical translations, which are not representative of most values in the sample.

3.1.2. Actions on emphasised words

Judgments of quality may be affected to some extent by the number of actions carried out on emphasised words, as these words tend to be more salient. Once again, atypical values were useful to distinguish between texts with an average proportion of actions on emphasised words—in contrast to actions carried out on non-emphasised words—and those with atypical numbers (five texts, all above average). Figure 4 shows the average quality judgments of the two groups.

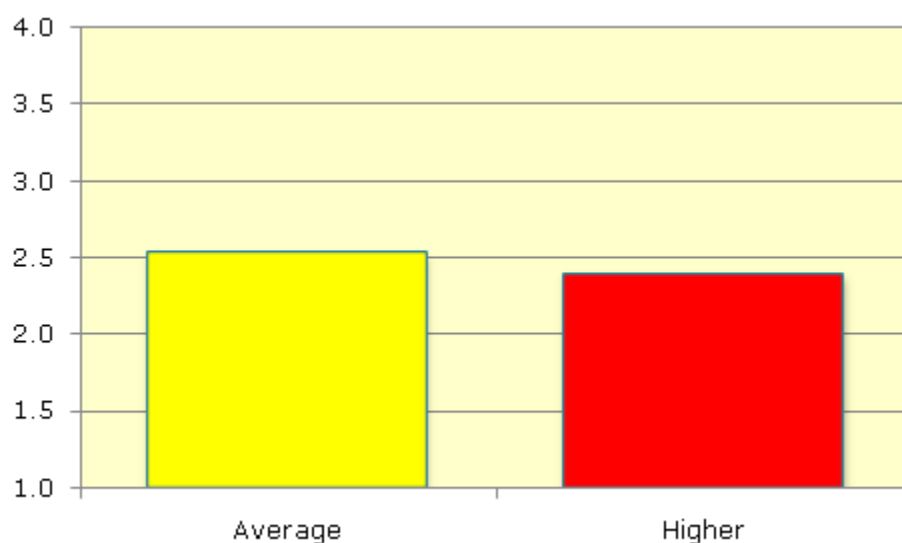


Figure 4: Quality judgment and percentage of actions on emphasised words

Texts with a higher proportion of actions on emphasised words obtained slightly lower grades than the rest of the translations. Contrary to what was expected, differences between the number of actions carried out on emphasised and non-emphasised words did not prove to be relevant. Therefore, it appears that judgments of quality are affected only by a possible increase in attention levels due to a statistically significant disparity in emphasised segments but not to the actions carried out on such segments.

3.1.3. An example

In order to study this phenomenon in greater detail, evaluators' actions on an error were analysed according to whether the same error appeared in typographically emphasised or non-emphasised passages: in CT4, translations of the phrase "aluminium block engines" first appeared in the title and then two sentences below in the body text. In nine out of the twelve texts in the set, the translations of the segment in the title were typographically emphasised (unlike in the body text), a circumstance which could have proved to be significant. Figure 5 illustrates the number of actions carried out by the evaluators on the same string of words, depending on whether the phrase was in the title or in the body text.

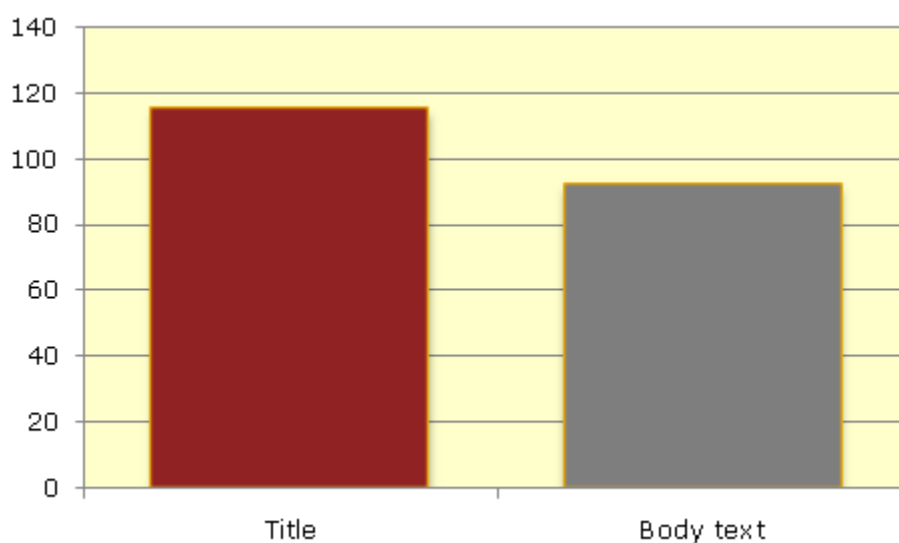


Figure 5: Number of actions on the phrase in the title and in the body text

The same translation of a term merited more actions by the evaluators when it appeared in the title (in bold and colour) than when it appeared in the body text. It could be argued that evaluators correct some errors only the first time they find them, but a thorough analysis of the data points rather to a question of typographical salience and attention. In fact, evaluators who based their evaluations on error analysis endeavoured to mark every error (Conde 2009), regardless of whether the phenomena occurred only once in the same text or on various occasions. What is striking, therefore, is that despite meriting more actions, the increase in the number of actions does not mean the quality is judged to be inferior: errors in emphasised segments are more noticeable but do not lead evaluators to judge the translator more harshly.

3.2. The length of the text

The translations were not all the same length. In some cases, translators did not finish their work and left incomplete texts. In contrast, one translator introduced extra information that was not present in the original text because she had found the original on the internet (which had been subsequently edited by the lecturer). These circumstances may have affected the evaluators' performance. Outliers were used to classify texts as too short (one text), too long (one text) and average (the remaining texts). Figure 6 shows the average quality judgments of the three types of translations.

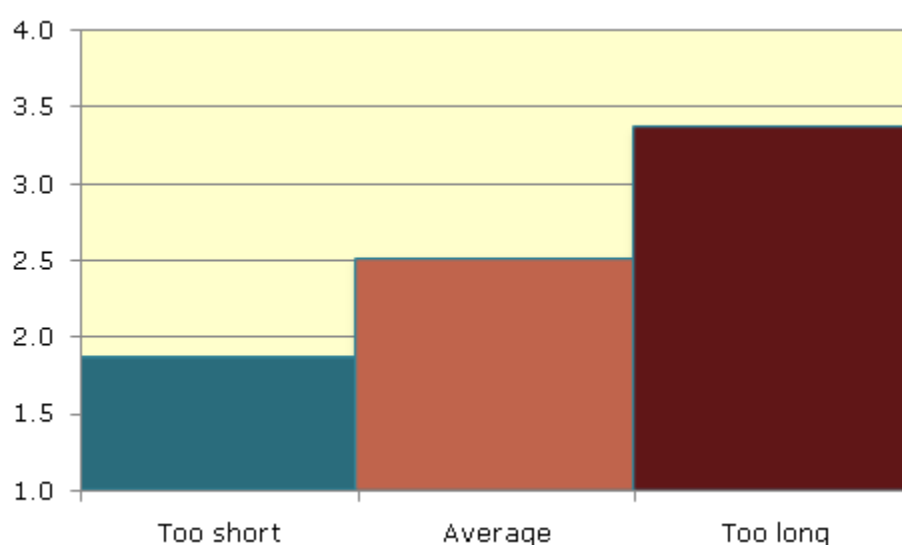


Figure 6: Quality judgment and text length

In contrast to the longest translation, which received a higher grade, the shortest translation was considered to be below average. The data appear to support the hypothesis that evaluators appreciate the fact that translators are able to complete as many words as possible. In any event, as in 3.1.1, the three groups are unbalanced (1, 1 and 46 texts respectively), and therefore the result may easily be due to chance. In order to test this effect, non-atypical translations were classified according to their length. Four groups were created, based on the number of words and taking into account the quartiles. Figure 7 shows the average quality judgment of these four groups.

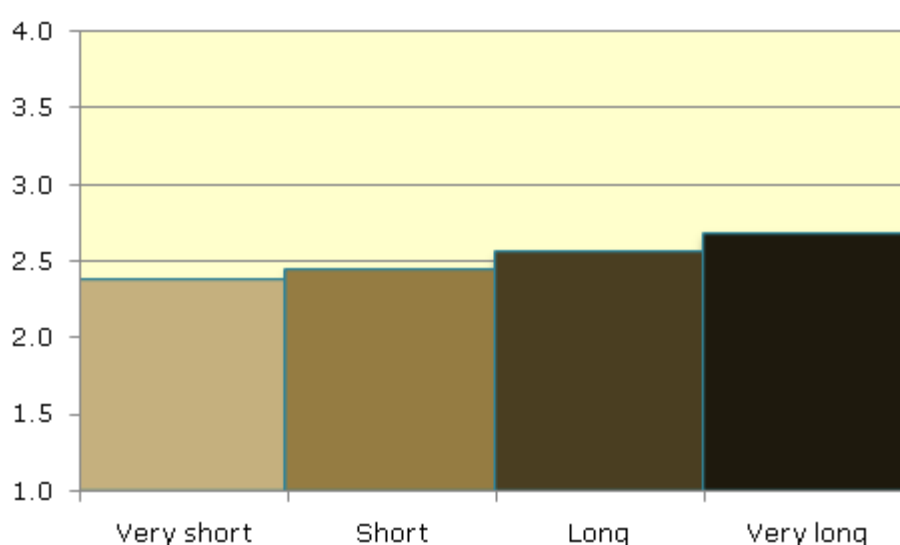


Figure 7: Quality judgment and non-atypical text length

A clear evolution can be observed across the four groups of translations, that is, from the very short translations to the very long ones. This may be taken to indicate that evaluators, either consciously or unconsciously, appreciate the degree to which the translations had been completed.

3.3. Resemblance to the original format

Translations presented different levels of formal resemblance to the four original texts. Regardless of the source texts, translations were classified according to the adequate reproduction of the following aspects: typeface, boldface, uppercase, paragraphs, font colour, pictures and lines of separation. As a result, translations were deemed to show a high, average or low level of resemblance. Of those that reproduced the original format perfectly or almost perfectly, one text obtained a statistically significant lower quality judgment; by contrast, another one obtained a statistically significant higher grade than the rest of the translations in its group. Figure 8 illustrates average quality judgments for the three groups, where outliers were suppressed.

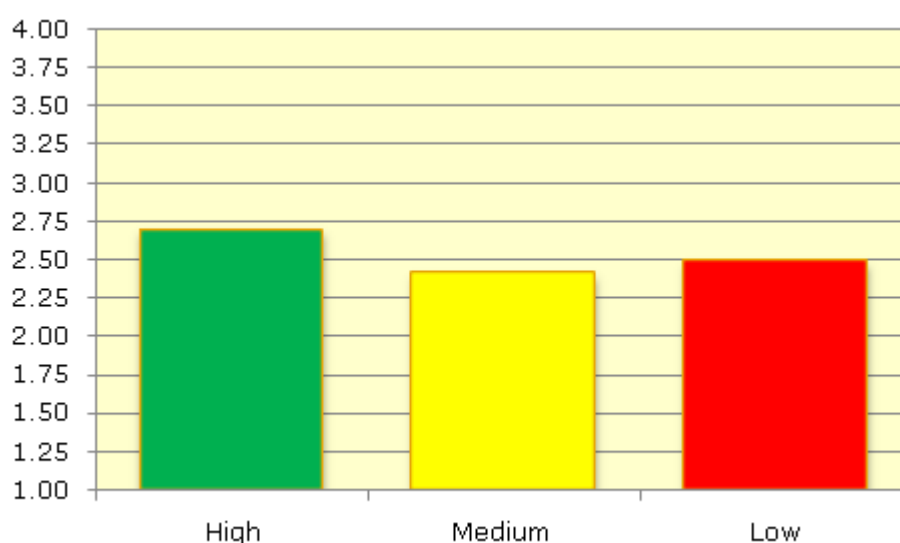


Figure 8: Quality judgment and resemblance to original format

There were no apparent differences among the groups. No pattern of evolution emerged: translations that did not correctly reproduce the original text obtained better grades than the average group. The quality judgment received for the translations that best reproduced the original text format was slightly higher than the rest. This may be attributed to the fact that evaluators appreciate a faithful resemblance to the original text format, but this speculation cannot be considered determinant in the quality judgments issued. The higher grades obtained by the texts with a poor resemblance to

the original (compared with those in the average category) might have a cognitive explanation: when evaluating texts that did not reproduce the original text format in any way, evaluators may have considered that the translators had focused on the text contents, which would be an acceptable translation decision; nevertheless, when evaluating texts that reproduced the original text format only to a certain extent, they may have considered this to be unacceptably negligent on the part of the translators since they had presumably resolved to reproduce the original format but were unsuccessful in doing so. However, the differences are so small that further study is required.

3.4. Lexical density and readability

Finally, lexical density was calculated using the type/token ratio, a measure of vocabulary variation within a written text. Other measures such as word and sentence length were taken into account in order to check for readability. First, the atypical values from each set were suppressed (T20 in DP3 and T45 in CT4): longer texts tend to have a smaller type/token ratio, whereas that of shorter texts tends to be higher. Table 1 shows the correlations found in the task as a whole and within each set.

	STATS	TYPES	TYPE/TOKEN RATIO	WORD LENGTH	SENTENCE LENGTH	Total judgment
Whole task	Pearson	0.150	-0.327*	0.139	0.077	1
	Sig. (bil.)	0.310	0.023	0.347	0.611	
	N	48	48	48	48	48
DP1	Pearson	0.546	-0.345	0.387	0.265	1
	Sig. (bil.)	0.066	0.272	0.214	0.406	
	N	12	12	12	12	12
DP3	Pearson	0.611*	-0.430	0.477	0.014	1
	Sig. (bil.)	0.035	0.163	0.117	0.964	
	N	12	12	12	12	12
CT2	Pearson	1	0.225	0.243	0.643*	0.176
	Sig. (bil.)		0.482	0.447	0.024	0.583
	N	12	12	12	12	12
CT4	Pearson	0.420	-0.379	0.363	-0.119	1
	Sig. (bil.)	0.174	0.224	0.246	0.712	
	N	12	12	12	12	12

Table 1: Lexical density and readability correlations

Three correlations with the quality judgment were found. Firstly, for the task as a whole, the higher the number of types included, the lower the quality judgment of the text. This may have been affected by the order of the tasks: evaluators were more demanding (Conde 2009) in their assessment of the texts in the first set (DP1), which were also longer than the others. Secondly, in DP3, texts with the largest number of words obtained the highest grades. A possible explanation for this is that most translators could not complete the DP3 translation task; evaluators, therefore, would have appreciated the fact that some translators were actually able to finish their task.

And thirdly, in CT2, texts with longer sentences obtained the best grades. CT2 was the first specialised original text that subjects were asked to evaluate. They were likely to be unfamiliar with the scientific language and, having just finished evaluating a non-specialised set, they probably appreciated the syntactic density and sentence length of some of the texts. However, this did not occur in CT4, the second and final specialised set, where subjects had perhaps become accustomed to the more direct, less complex scientific prose.

In summary, neither lexical density nor readability seems to affect the evaluators' average quality judgment since most correlations can be explained by other effects such as serial translation evaluation.

4. Conclusion

Following the discussion of the results, it would be appropriate to reflect on the phenomena that emerged. In general, two content-independent features may be taken into account when evaluating translations: consistency and, above all, productivity.

Neither the percentage of actions on outstanding segments nor lexical density yielded significant data. The first factor shows no clear relationship to quality. Lexical density (measured by the type/token ratio) and readability (by indexes of words and sentence length) are also not directly related to quality, since the relationships found are attributable to other factors, such as serial evaluation order effects.

The parameter "resemblance to original format" has brought to light a small but interesting aspect that may have conditioned the evaluators' assessments in this study and should therefore be considered in future research: consistency. Evaluators may forgive the fact that translators do not reproduce the original format of their translations as long as this appears to be a coherent and deliberate decision; in contrast, translations with a limited resemblance to the original format are penalised more heavily.

In light of the results, however, the clearest effect is that related to productivity: the more complete the translations, the higher the judgments issued by the evaluators. This applied not only to the atypical translations (that is, those significantly longer or shorter than the rest within a given set), but to all translations, since the average quality judgment increases gradually in accordance with the length of the translation.

These results should provide food for thought for translation students, in-house professional translators and all those whose translations are evaluated on a daily basis. In such a dynamic, rapidly changing environment, where speed and deadlines are par for the course, special attention should be paid to more noticeable features since quality controls are sometimes based on rapid, superficial assessments.

However, it should be noted that this study was conducted on the basis of data obtained from the simple observation of the evaluation process; thus, the effect of attention on translation evaluation was not specifically studied. Nevertheless, the present work could pave the way for further research involving specific tests and experimental studies with control groups. Variables should be manipulated, for example, by changing the order of errors: first presenting the highlighted segments in some texts and non-emphasised segments in others, to test the effects suggested in this paper. Future research might use eye-trackers to directly measure error perception in different parts of the text, although this would obviously require greater financial resources, as these devices are still prohibitively expensive.

Meanwhile, we should continue to encourage empirical research on translation evaluation and attention: two processes that are difficult to measure, and perhaps for that reason, endlessly fascinating.

References

- **Baker, Mona** (1995). "Corpora in Translation Studies: An Overview and Suggestions for Future Research." *Target* 7(2), 223-243.
- **Boone, Louise** et al. (2009). *Contemporary marketing*. Florence: Cengage Learning.
- **Ceschin, Adriana** (2004). *Memória de tradução: auxílio ou empecilho?* PhD thesis. Pontifícia Universidade Católica do Rio de Janeiro.
- **Conde, Tomás** (2009): *Proceso y Resultado de la Evaluación de Traducciones*. PhD thesis. Universidad de Granada.
- — (2010). "Tacit technique on the evaluation of technical texts." Lluïsa Gea, Isabel García Izquierdo and María José Esteve (Eds) (2010). *Linguistic and Translation Studies in Scientific Communication*. Oxford: Peter Lang, 197-217.
- — (2008). "Tipos textuales en la evaluación en serie." Luis Pegenaute, Janet DeCesaris & Mercè Tricás (Eds) (2008). *La traducción del futuro: mediación lingüística y cultural en el siglo XXI* (vol. II). Barcelona: Universitat Pompeu Fabra, 33-45.

- **Cruces, Susana** (2001). "El origen de los errores en traducción." Domingo Pujante González et al. (Eds) (2001). *Écrire, traduire et représenter la fête*. Valencia: Universitat de València, 813-822.
- **Darwish, Ali** (2001). *Transmetrics: A Formative Approach to Translator Competence Assessment and Translation Quality Evaluation for the New Millennium*.
www.translocutions.com/translation/transmetrics_2001_revision.pdf
(consulted 27.06.2010)
- **Hajdú, Peter** (2002). "The New Hungarian Translation of Aristotle's Poetics: When Translation and Commentary Disagree." *Across Languages and Cultures* 3(2), 239-250.
- **House, Juliane** (2001). "How do We Know when a Translation is Good?" Erich Steiner & Colin Yallop (Eds) (2001). *Exploring Translation and Multilingual Text Production: Beyond Content*. Berlin/New York: Mouton de Gruyter, 127-160.
- **Koo, Siu Ling** and Harold Kinds (2000). "A Quality-Assurance Model for Language Projects." Robert C. Sprung (Ed.) (2000). *Translation into Success. Cutting-edge strategies for going multilingual in a global age*. Amsterdam: John Benjamins, 147-157.
- **Larose, Robert** (1998). "Méthodologie de l'évaluation des traductions." *Meta* 43(2), 163-186.
- **Mangal, S. K.** (2007). *Essentials of educational psychology*. New Delhi: Prentice-Hall of India.
- **Martínez, Nicole** and Amparo Hurtado (2001). "Assessment in Translation Studies: Research Needs." *Meta* 46(2), 272-287.
- **Munday, Jeremy** (1998). "A computer-assisted approach to the analysis of translation shifts." *Meta* 43(4), 542-556.
- **Muñoz Martín, Ricardo** and José Tomás Conde Ruano (2006). "Effects of Serial Translation Evaluation." Peter A. Schmitt & Heike E. Jüngst (Eds) (2006). *Translationsqualität*. Frankfurt: Peter Lang, 428-444.
- **Peha, Steve** (2003). *Looking for quality in student writing*.
http://www.tms.org/writing_quality/writing_quality.htm (consulted 27.06.2010).
- **Rosenmund, Alain** (2001). "Konstruktive evaluation: Versuch eines Evaluationskonzepts für den Unterricht." *Meta* 46(2), 301-310.
- **Vinay, Jean-Paul** et Jean Louis Darbelnet (1958). *Stylistique comparée du français et de l'anglais : Méthode de traduction*. Paris: Didier.
- **Vollmar, Gabrielle** (2001). "Maintaining Quality in the Flood of Translation Projects: A Model for Practical Quality Assurance." *The ATA Chronicle* 30(9), 24-27.
- **Waddington, Christopher** (1999). *Estudio comparativo de diferentes métodos de evaluación de traducción general (inglés-español)*. PhD thesis. Universidad Pontificia de Comillas.

- **Wolfram MathWorld** (2010). *Quartile*.
<http://mathworld.wolfram.com/Quartile.html> (consulted 22.09.2010).

Biography

Dr Tomás Conde works as a researcher for the GENTT Group (Textual Genres for Translation) at the Universitat Jaume I, Spain. He has studied and completed his PhD in Translation and Interpreting at the Universidad de Granada, Spain, where he has taught and conducted research on translation evaluation for the PETRA Group (Expertise and Environment in Translation). His research interests also include translation quality, professional proofreading and editing of texts.

Contact: jconde@trad.uji.es

¹ Conde (2009: 317) shows that only 3% of the evaluators' actions on translations set out to praise good decisions or to point to the high translation quality.

² The effects of order in this translation corpus are dealt with in depth by Muñoz & Conde (2006). In addition, Conde (2010, 2008) analyses the differences in the evaluation of specialised and non-specialised texts.

³ These levels were then converted into numbers (1, 2, 3 and 4, respectively), to enable statistical treatment of the data and cross-checking of the quality judgment with the other parameters.

⁴ It should be noted, however, that correlations with the quality judgment were calculated by groups in the cited paper, but not in general.

⁵ Wolfram MathWorld defines 'quartiles' as "One of the four divisions of observations which have been grouped into four equal-sized sets based on their statistical rank" (Wolfram MathWorld 2010).