TRANSLATION QUALITY ASSESSMENT FOR RESEARCH PURPOSES: AN EMPIRICAL APPROACH

Rui Rothe-Neves
Universidade Federal de Minas Gerais

1. Introduction

This study was designed to provide an index of translation quality, by means of which several translations of the same text could be compared for research purposes in Translation Studies (TS). It is part of a more comprehensive research on the relationship between some of the translator's cognitive characteristics and various features of the translated text (Rothe-Neves, 2002). Instead of dealing exhaustively with the issue of translation quality assessment from a theoretical point of view, in this article I will concentrate on the question of how to provide empirical information that allows a researcher to compare translations based on quality. I will first discuss some points that seem important to translation quality assessment for the sake of empirical research on the process of translating. The discussion is organized around the perspective that traditional methods of assessment could be improved by refining their data collection techniques. Then, I will deal with the construction of a quality scale using external evaluation and appropriate statistic tools to investigate reliability. Finally, further developments will be suggested.

1.1 Quality assessment in empirical studies on translating

Translation quality assessment is not an undisputed issue in TS. Nonetheless, it is interesting for empirical research about the translating process, since some features that are consistently related to quality could be then systematically investigated. The main problem seems to reside in how to express quality or what measure should be used for the quality of a translation. This question has been typically addressed in two different ways, with many variations. The translated text (TT) may be assessed by experts such as professional translators, translation or language teachers and others, including the researcher. Assessment parameters, that may or may not be clearly stated, are in most cases those used in translation courses and, therefore, it will be referred to here as the "pedagogical approach", although it does not differ considerably from the assessment methods for professional accreditation (ATA, 2000). There are no means to prevent that the evaluator assesses the translation by comparing it to an ideal text she could have produced herself, thus projecting her own individual standards or prejudices onto the actual text. In that way, the evaluator's experience on the subject warrants her opinion about the quality of a TT. Thus, it does not provide an objective measure of quality in translation, but it has been used to investigate the translating process (e.g., Jensen, 1999; Tirkonnen-Condit, 1986).

Alternatively, TTs may be described according to a system that is theoretically motivated, clearly stated and discussed previously to the analysis. That system also serves to describe the source text (ST) so that, through independent analysis, STs and TTs can be compared. This will be referred here as the "scientific approach". In such a way, TTs' quality is normally presented as a degree of similarity of the TT description in relation to that of the ST. The model of quality assessment by House (1987; 2000) is perhaps the most famous example. In her book, a landmark in translation research, House introduces the concern towards a scientific treatment of quality in translation. She also revises empirical studies

directed to the reception of the translated text by the target-culture reader, and brings to the field the very used and still very useful concept of "communicative competence", coined by Hymes (1967). This may be the book's greatest contribution, since the pragmatic background of her model opened a way to further studies that incorporated cultural aspects to the understanding of translation. Nevertheless, her model was directed towards translation as an L2 classroom exercise, and this puts a serious limit to it as a tool to investigate translations as an end.

Some authors have suggested that a comparison between the propositional analysis of STs and TTs should provide an objective measure of quality, namely the proportion of ST propositions that are also present in the TT (Dillinger 1989; Militão 1996; Tommola & Lindholm 1995). Thus, the propositional content figures as a *tertius comparationis*. Such a comparative analysis is, in my opinion, not the path we should strive to. In order to discuss a concrete example, I will next present the work by Militão (1996). It has never been published and deals with written translation, whereas the other above mentioned works deal with simultaneous interpreting.

Militão (1996) asked professional translators to translate a text containing cultural and spatial or "orientational" metaphors. Cultural metaphors relate concepts with other categories that are culture-bound (*She speaks in italics*), while "orientational" metaphors occur when concepts are organized in terms of the more basic system of spatial orientation (*I'm feeling up today*). Her aim was to investigate whether the type of metaphor (cultural vs. "orientational") influences the cognitive processes involved in translating a text. Based on cognitive theories, she hypothesized that as "orientational" metaphors are based on semantic components that could be found in different cultures, they may be preserved in translation. She analyzed all metaphors in terms of their propositions and compared them with the analyses of the translated metaphors. As she had thought, more metaphors of the cultural type turned out to be preserved in the translations, as compared to "orientational" metaphors.

This example highlights the limits of propositional analysis as a research tool. First, it depends on the analysis system used to extract propositions from text (cf. discussion in Tommola & Lindholm 1995). The more detailed a system is, the more difficult it is to apply, and to achieve intersubjective reliability. Most systems are based on the researcher's own interpretation of the propositional content. Second, data interpretation depends on those criteria according to which a certain TT should be assessed. As shown, cultural metaphors tend to be more easily leveled out, e.g. through paraphrasing. Nevertheless, this is a natural process, due to the fact that cultural metaphors, as opposed to "orientational" ones, are generally not bound to language-independent semantic structures. So, they "survive" only after some kind of re-creation. The same fact (leveling out a metaphor) could thus be interpreted either as an error or as a useful strategy, depending on the type of text, audience etc. A similar problem is faced by qualifying the translation according to the reproduced information, as *verbatim*, paraphrase etc., as done by Dillinger (1989). In this case, although there is promising work on systems that automatically extract informational content from texts (Foltz, 1996; Rieger, 1988), a translation is good not only because it shares ST content. Where there is no empirical study on how translations of various types are produced, such a *tertius comparationis* should be affected by the researcher's own notions. In other words, this scientific approach is also in danger of revealing more about the researcher's opinions than about translation quality.

1.2. The question of validity

Having presented and discussed some methodological problems, there is a more far-reaching question to be dealt with. This is the question of validity. A measure is valid only when it really measures what it is supposed to measure. This is not an easy question when it comes to translation quality because, as stated right at the beginning,

there is no consensus on what it means. In the pedagogical approach, it is up to the evaluator with her experience to spur the quality of a given work. In the scientific approach, a common research strategy is to define "quality" in the first place, and then look into the data. This is why House (2000) begins her section on the quality of translation by stating that translation quality assessment requires a theory of translation.

In my opinion, this is not very convincing. First, for epistemological reasons, since a first-order theory based on empirical data always comes ahead of second-order, theoretical formulations (cf. for TS, Königs, 1990). Data about the quality of translations in terms of text characteristics could be of great interest in the investigation of fundamental questions about how translations are produced. A case in point is whether working memory is important for translating as it is for creative writing, where it is known to influence production time and text quality (Ransdell & Levy, 1996). If the answer is positive (as it seems to be, cf. Rothe-Neves, 2002), this piece of information is useful to understand translating under time pressure; an issue that has certainly more to do than only with cognition in translation business. Then, it follows that we should be able to keep track of translation quality *before* theorizing it, or - as it is known - in a theory-independent way. Secondly, it is not very convincing for methodological reasons, with regards of what was previously said about using an interpretative system that is not backed up by actual translations.

As discussed so far, controversies may be raised on whether the scientific approach can fulfill the needs of investigations on translation quality. Probably, those problems derive from the fact that, coming from theoretical linguistics, science is envisaged as consisting of deductive reasoning. In fact, deductive reasoning is quite productive in science, but it helps mostly when there is sufficient empirical knowledge to support it. This is perhaps a good reason for us to return to a pre-scientific status in the area of translation quality which is represented by the first assessment method

presented above. As discussed in the next section, there are methods to extract subjective information in such a way that it can be statistically reliable. So, we could improve the pedagogical assessment of quality in order to generate research-useful, first-order data. As the *momentum* in TS seems to call for more empirical work before we begin with generalizations, how can we deal with the issue of validity as part of this pre-scientific move? In order to be consistent, it seems that the same source of information has to provide evidence for both the validity and the reliability questions, that is, we should be able to collect empirically justifiable data to build valid and reliable answers.

As said, this study was carried out from the perspective that traditional methods that do not use an independent system of assessment could be improved by refining their data collection techniques. The choice here is to skip the researcher's own subjectivity by letting translations be assessed by others. These referees will be called external evaluators because they are not involved in the research process: they are not aware of the hypotheses to be investigated. It is not a new road, on the contrary, it has been proposed quite long time ago by Nida & Taber (1982, p.170 et seq.) in the form of "practical tests". Nida & Taber proposed that normal readers, whom the translation addresses, should read the translations and react to them following standard forms (cloze test, alternative choice etc.). Individual prejudices should be naturally overcome through sampling techniques. In my opinion, the assessment through external evaluators presents at least two advantages. First, it does not require the use by the researcher of a *tertius comparationis*, be it an ideal translation or an analysis system. Secondly, if, contrary to Nida & Taber, the external evaluators are translation professionals (translators, translation teachers etc.) who share similar contextual conditions with the translators who produced the TT to be assessed, assessment data could be taken as a portrait of those quality criteria used at that time and place, provided that subjective data are treated in such a way that it objectively captures whatever intersubjective parameters emerge.

That "objective capture" is likely to be what Tuldava (1995) applied to investigate whether subjective opinion of experts about the quality of literary texts could be traced back to some objective characteristics of the same texts, like sentence length etc. In this way, an assessment procedure could be useful for empirical investigation of translating. By requiring groups with different backgrounds, such a procedure could also be used to empirically investigate the very notions which underlie translation assessment, thus furthering our knowledge about these in a cultural setting. As far as validity is concerned, as much information as possible should be provided to the evaluators when brought to reflect on the assessment itself. That means in the present study the use of scale data for quality and discursive data (people's own words and reflections) for scale data. Thus, by triangulating different research methods, the present study wants to throw, from different angles, some new light onto the same old object of "quality". The next session is devoted to present a standard form to elicit information, and an experiment carried out to test the reliability of the instrument.

2. Method

2.1. The experiment

A set of 12 Portuguese translations of the same English text (the first page of the novel *Emma* by Jane Austen) was assessed for quality by a team of five professional translators, who teach translation courses at three Brazilian universities (henceforth evaluators). The translations were produced by six professional translators and six undergraduate students of a translation course, all of which were native speakers of Brazilian Portuguese. Undergraduates and professionals took part in a previously mentioned research about cognitive characteristics of translators (Rothe-Neves, 2002), for which they provided an informed consent. The DOS version of the program *Translog* (Jakobsen, 1999) was

used to keep a record of the translation process. The subjects were allowed to use only the program's built-in dictionary, so that the point in time when the dictionary was looked up was also registered. No time limits were stipulated. I proofread the translations for Portuguese diacritics left unrecognized by DOS; no other corrections were made. All 12 translations so produced were handed blindly to the evaluators together with an assessment scale and other useful information described below.

2.2. Assessment scale

The objective of the assessment scale was to establish a rank order of translations based on quality. The standard form contained a set of questions (Table 1) about some aspects of the translation to be assessed, and served to direct attention of all evaluators to the same aspects. The questions were extracted from a similar scale for empirical research on writing quality (Ransdell & Levy, 1996), with those items referring to creative work being replaced by questions typically raised by clients of translation services (Stolze, 1997:158). They were chosen for the present study, because the evaluator, as much as the client, is not directly involved with the translation situation.

Table 1 – Questions presented in assessment scale

1.	Does the text read fluently?
2.	Is the translation grammatically correct?
3.	Is the spelling correct?
4.	Are there unjustified inferences?
5.	Is the vocabulary adequate?
6.	Is the vocabulary used consistently throughout the text?
7.	Is the translation performed according to the assignment?
8.	Does the layout correspond to normal standards?
9.	Could the translation be used according to the style norms for this kind of text?
10.	Is the overall result satisfactory?

The evaluators reacted to the questions following a standard procedure (see below). First, they should decide whether or not a question is relevant for that translation. Suppose there was no assignment, then question #7 above is not justified, and gets a zero. When the question is relevant, it should be attributed a value in a 5-points Likert scale (1=Not at all; 2=A bit; 3=Somewhat; 4=Much; 5=Completely). Technically, this is a combination of categorical (0, 1) and ordinal (1-5) data, and the former data serve to decide whether a question should be eliminated or not from the analysis, a point to which I will return later. All questions are positive, that is, a larger value means more quality in the translation for that particular aspect, except question #4, which is the opposite. The sum of points attributed to each question – subtracted the negative value of #4 – formed a Quality Index of each text.

2.3. Procedure

Each evaluator received personally or by mail a presentation letter about the research, along with fill-in and information material. A sheet was included with a standard procedure to be followed for the assessment task. This was intended to direct attention to some important aspects. The evaluator should first read all translations throughout in order to get an impression of their readability (question #1), and only then should the original be read. When responding to a particular question, the evaluator should assess a translation in comparison with all the others, so that each attributed value is relative to the entire text sample. A sheet was also included with a detailed description of each question, partly a translation from the Quality Rating Guidelines presented in Ransdell & Levy (1996:102-105). It was intended to prevent misunderstandings and avoid assessments based in error analysis. The evaluators were unaware of the translators' identities or even that there were two different levels of competence. In order to avoid a systematic influence, they received the translations numbered in two different orders, both with mixed levels of competence.

3. Results and discussion

For analysis, answers from one evaluator were completely excluded because sometimes more than one answer was given. From the others, all answers to question #8 were not considered because an evaluator chose consistently to mark it zero. The assessment scale was tested for concordance among evaluators and for its reliability. Concordance refers to how different the opinions of all evaluators were, so that consensus may be inferred. It is commonly estimated by Kendall's W, which varies from 0 to 1. A text-to-text analysis was not possible, because the evaluator sample comprised less than seven subjects. Taking the median response (Table 2), the concordance coefficient among all texts was significant (W=0.8234), indicating consensus.

Table 2 – Translation assessment by scale item (median response) and Quality Index (QI)

TEXT[1]	SCALE ITEM									QI
	#1	#2	#3	#4	#5	#6	#7	#9	#10	
9	4	4.5	5	1.5	4.5	5	5	4.5	4	35
12	4.5	4.5	5	2	4.5	4.5	5	4	4	34
4	3.5	3.5	4.5	1.5	4	4	4	3.5	3.5	29
10	4.5	4	5	2.5	3.5	3.5	4.5	3	2.5	28
1	3.5	4	4.5	1.5	3.5	4	3	2	2	25
7	3	3.5	3.5	2	3	3.5	4	2.5	2	23
6	2	3	4	3.5	3	3.5	4	3	2	21
5	2	3	2.5	3	3	3.5	4	2.5	2	19.5
11	2	2.5	3.5	3	3.5	3.5	3	2	2	19
3	2	2.5	4	4	2.5	3	4	2	2	18
2	2	2	2	4	2	4	4	1.5	1.5	15
8	1.5	2	2	2	2	2.5	3	1.5	1.5	14

Another way of testing consensus is through the correlation between the values attributed by the evaluator to each question in each text. In this case, correlations were highly significant (Table 3).

Table 3 – Correlation coefficients between evaluators for entire question sample (Pearson's *r*)

(N = 108)	J1	J2	J3	J4
J1	1.000			
J2	0.601	1.000		
J3	0.440	0.443	1.000	
J4	0.596	0.565	0.305	1.000

Put simply, the test of reliability estimates if it is reliable to use a scale to construct a compound index. In this case, it should mean that the sum of all question values could indeed be taken as an indication of translation quality. The most common reliability estimation is, by all means, that of Cronbach's a coefficient. As Kendall's W, Cronbach's a also varies from 0 to 1. The assessment scale (median results) was found to be reliable with a=0.9510, and, considering the internal items variation (Standardized item alpha), a_z=0.9532 was found. It means that the questions used here reliably converge to a Quality Index. Just for comparison, psychological tests are expected to show a between 0.80 e 0.90 (Anastasi & Urbina, 1997).

Reliability analysis also offers a series of statistics to better investigate index composition. In Table 4, the mean column shows the scale mean with all items or excluding items one by one. Through these statistics it is clear that the major contributions to QI came from questions # 7, 3 e 6, in that order, because of how much the general mean decreases when they are excluded. On the other side, 4NEG, representing responses to question #4 multiplied by –1, was the less important to the final index. The second column shows how

much each item contributed to the overall scale variance. The third column presents the correlation coefficient between each item and the rest of the scale taken together; 4NEG is here the item with bears the least relationship with all the others. The squared correlation allows for the estimation of how much each item may be estimated by all the others taken together; it lies above 90% here, except for #6. Finally, the last column shows the amount of change in the scale's a if each item were excluded from computation; excluding 4NEG will make the general index a little better.

Table 4 – Reliability statistics for Quality Index

ITEM	STATISTICS				
	Scale mean if item deleted	Scale variance if item deleted	Corrected Item-Total Correlation	Squared Correlation Coefficient	Alpha if item suppressed
#1	20.5000	35.5000	0.8891	0.9634	0.9409
#2	20.1250	37.2330	0.9375	0.9790	0.9383
#3	19.5833	35.9924	0.8200	0.9546	0.9456
#5	20.1250	37.9602	0.9251	0.9824	0.9395
#6	19.6667	41.4697	0.7500	0.7568	0.9491
#7	19.4167	42.0379	0.6397	0.9860	0.9530
#9	20.7083	36.8390	0.8971	0.9816	0.9401
#10	20.9583	37.3845	0.9117	0.9794	0.9396
#4NEG	25.9167	40.6742	0.5564	0.9677	0.9581
	Mean	Variance	Amplitude	Standard deviation	Variables
Scale	23.3750	48.2330	-	6.9450	9
Items	0.8289	0.0772	0.8182	-	-

The analyses presented here indicate that each question in the assessment scale is related to each other, and that all of them point

to a common subjacent entity. There is no reason to believe that this entity is not the quality of the analysed translations. In this case, the quality of translations can be very well estimated using the sum of responses to each question, or stated in another way, the scale presented here does indeed offer an index of quality. The only technical restriction to Cronbach's a is that the correlation coefficients between all scale items must be positive, otherwise it violates the model, which is no longer valid. But it should not be the problem here, as can be seen in Table 5.

Table 5 – Correlations between scale items

	Q1	Q2	Q3	Q5	Q6	Q7	Q9	Q10	Q4NEG
Q1	1,000								
Q2	0,924	1,000							
Q3	0,835	0,855	1,000						
Q5	0,807	0,880	0,836	1,000					
Q6	0,676	0,719	0,561	0,764	1,000				
Q7	0,628	0,610	0,519	0,530	0,624	1,000			
Q9	0,738	0,822	0,755	0,873	0,733	0,800	1,000		
Q10	0,776	0,792	0,750	0,902	0,764	0,726	0,936	1,000	
Q4NEG	0,615	0,664	0,445	0,619	0,384	0,067	0,486	0,559	1,000

It seems, therefore, reasonable to assume that the rank order by quality produced with the QI (Table 2) allows for inferences about the relationship between quality as a text characteristic and the characteristics of the process and/or the subjects' traits which gave rise to the translation. For the investigation of the translating process it could be quite useful. Nevertheless, for the evaluators the information provided by the scale was not enough. Their opinion about the scale was requested at the end of a questionnaire on their background (taken from Li, 2000). Two questions were answered with the same 5-point Likert scale (1=Not at all; 5=Completely). To the question "Does the assessment scale help focus on text

aspects?" there was a median response of 3.5; the same median arrived at for the question "Is the scale comprising?" This means that concerning those questions the evaluators considered the scale somewhere between 3 (Somewhat) and 4 (Much). The last one was an open-ended question, asking for suggestions on how to improve the whole procedure. Besides, they could contact me using e-mail or telephone as stated in the presentation letter, what indeed happened in one case. In that way, their opinion was discursively informed and is summarized next.

The statements felt on three topics: (a) the scale; (b) suggestions to improve the scale; and (c) the response procedure. Concerning the scale, I reproduce the opinion here for the sake of completeness, but without actually quoting the exact words:

- the scale is too long and, because it is necessary to repeat it to each text, the assessment procedure becomes tiring;
- some important aspects were not covered by the scale, such as cohesion, coherence, and punctuation;
- errors were not systematic approached;
- reading fluency may not be a quality criterion.

In order to improve the scale, it was suggested to revise instructions, present lesser items in the scale, and to combine the scale with objective measurements of the translated text, such as propositions, reader response, and error analysis.

Finally, some words were expressed about the procedure itself. In sum, the evaluators suggested the researcher should present just those questions that make sense for the task at hand. In this case, as the layout was not important neither in the source text nor in the translation assignment question #8 (Does the layout correspond to normal standards?) should be cut off. It was my intention to use a full scale and wait for the responders to choose themselves what

was important. This is why I included a zero, as said before. It makes the scale all-purpose, and allows for the detailed investigation of different response patterns. Theoretically, we may say that for some text types some questions are more or less unimportant, but how important they are is a matter of empirical investigation. The proportion of respondents choosing zero may be used as evidence of the importance attributed to the topic touched upon in a question. Additionally, it is worth mentioning again that only one respondent chose zero consistently for question #8.

The observations made by the evaluators indicate some of the limits of the present study that should be taken into consideration. They do not seem to hinder the entire enterprise, but rather to provoke further developments. Two other aspects are, in my opinion, inherent characteristics of the assessment procedure presented here, and should be discussed in some length. The first concerns the linear relationship that was supposed to hold between the scale items. The Quality Index is a single number with the advantage of being derived from parameters clearly attached to the translated text. In that respect, it is more appropriate to research purposes than a note given by a single evaluator. Nevertheless, it should be applied to other situations so that its length, its parameters, and, more important, its validity can checked. As all evaluators belong to the same cultural system, the responses may reasonably reflect the importance of such text aspects for the quality of translation, while a single note may represent different parameters to different evaluators. This technical strength, however, only holds if quality is a sum of other features, as assumed here. The evaluation team consulted for the purposes of the present research was not large enough to allow for testing other assumptions, say, that each aspect interacts in different ways with each other.

Another far-reaching limit is of course that the scale only allows for conclusions *within* the text sample assessed. In fact, it makes no claim of absolute value. What may be a throttling straitjacket could be taken as an advantage. The IQ fits completely within the

scope of the empirical research it was designed for: it is inexpensive, uncomplicated, and reliable. We could discuss for a decade around an absolute quality standard for translation, without being able to learn from actual translations. As pointed out by Stolze (1997), on whom I draw here, quality parameters in translation are entirely bound to the aim of assessment. Those will differ when the assessment serves the client, who wants to find some cues about the quality of a text that should be related to another, maybe, unreadable text. Should it serve the translator, assessment parameters may function as a quality standard to be attained. Finally, they will be useful to the translation teacher if they indicate the students' competence areas that still need development. A scale that claimed an absolute validity would probably confound those three objectives, and its value would then be no more use. So, it may be a good choice to explicitly ask the evaluators to compare the translated texts with one another, thus restricting the findings to the sample under examination. If it does not allow us to generalize from sample to sample, it is a cost that TS could still afford in its methodological infancy. It may help our researchers to learn how to develop more specific tools, to ask more precise questions, to demand more from results, and finally to begin to untangle, by means of empirical procedures, the quality of a translation.

Nota

1. Texts 1-6 produced by undergraduates and 7-12 by professionals; texts ordered by decreasing QI.

Acknowledgements

This study is part of a thesis done under supervision of Fábio Alves, whose commentaries were most valuable. Thanks to Adriana Pagano, Beatriz Caldas, Carlos Gohn, Heloisa Barbosa, and José Luiz Vila Real Gonçalves for their participation.

References

Anastasi, A. & Urbina, S. *Psychological testing*. 7.ed. New Jersey: Prentice Hall. 1997.

American Translaotrs Association. Framework for standard error marking: an explanation. 2001. URL: http://www.atanet.org/bin/view.pl/12438.html. Accessed: 8/21/02.

Dillinger, M. L. *Component processes of simultaneous interpreting*. Montreal: Department of Educational Psychology, McGill Univ. (unpubl. Doctoral thesis). 1989.

Foltz, P. W. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28/2. 1996. p.197-202.

House, J. *A model for translation quality assessment*. 2.ed. (1.ed., 1977) Tübingen: Narr, 1981.

_____. Quality of translation. In M. Baker (ed.). *Routledge encyclopedia of Translation Studies*. 2.reimpr. (1.ed. 1998). London: Routledge. 2000. p.197-200;

Hymes, D. Why linguistics needs the sociologist. *Social Research*, 34/2. 1967. p.634-647.

Jakobsen, A.L. & Schou, L. Translog documentation; version 1.0. In G. Hansen (ed.). *Probing the process of translation*: methods and results. Copenhagen: Samfundslitteratur. – (Appendix). 1999.

Jensen, A. Time pressure in translation. In G. Hansen (ed.). *Probing the process of translation*: methods and results. Copenhagen: Samfundslitteratur. 1999. p.103-119.

Königs, F. G. Wie theoretisch muß die Übersetzungswissenschaft sein? Gedanken zum Theorie-Praxis-Problem. *Taller de Letras* 18. 1990. p.103-120.

Li, D. Tailoring translation programs to social needs: a survey of professional translators. *Target*, 12/1. 2000. p.127-149.

Militão, J. A. *A significação metafórica e o processo de tradução*: novas perspectivas de uma abordagem integrada. Belo Horizonte: Faculdade de Letras da UFMG. (unpubl. MA dissertation). 1996.

Nida, E & Taber, C. R. *The theory and practice of translating*. 2.reimpr. (1.ed.1969; 1.reimpr.1974). Leiden: Brill. 1982.

Ransdell, S.E. & Levy, M.C. Working memory constraints on writing quality and fluency. In C.M. Levy & S. Ransdell (eds.). *The science of writing*: theories,

methods, individual differences, and applications. Hillsdale/New Jersey: Lawrence Erlbaum. 1996. p.93-101.

Rieger, B. Relevance of meaning, semantic dispositions, and text coherence. Modelling reader expectation from natural language discourse'. In M.E. Conte, J.S. Petöfi & E. Sözer (eds.). *Text and discourse connectedness*. Amsterdam/ Philadelphia: John Benjamins. 1988. p.153-173.

Rothe-Neves. Características cognitivas e desempenho em tradução: investigação em tempo real. Belo Horizonte: Faculdade de Letras da UFMG. (unpubl. Doctoral thesis). 2002.

Stolze, R. Indicadores de qualidade para a avaliação de tradutores no âmbito da didática. *TradTerm* 4/1. 1997. p.157-173.

Tirkonnen-Condit, S. Reader impressions and textlinguistic priorities in translation quality assessment. In S. Tirkonnen-Condit (ed.). *Empirical studies in translation*: textlinguistic and psycholinguistic perspectives, (Studies in Language 8). Joensuu: Fac. Of Arts. 1986. p.49-73.

Tommola, J. & Lindholm, J. Experimental research in interpreting: which dependent variable? In J. Tommola (ed.). *Topics in interpreting research*. Turku: University of Turku/Centre for translation and interpreting. 1995. p.121-133.

Tuldava, J. *Methods in quantitative linguistics*. Trier: WVT. Cap.5: A comparison of subjective and objective characteristics of style. 1995. p.93-108.