2009h

Some thoughts about the evaluation of translation products in empirical translation process research

Gyde Hansen

Abstract

In empirical studies of translation processes with a focus on combinations of methods and a large variety of parameters and data, the parameter "evaluation of the translation product" remains problematic and a source of subjectivity, insecurity and bias. The question is if and how evaluation criteria can be described, applied and documented rigorously, especially in projects where it is claimed that product quality is a main parameter in relation to the final result of the whole study.

Rigorous quality assessment requires transparent evaluation procedures and criteria, and in projects involving process research, the evaluation process easily develops into a complex project on its own.

1. Introduction

A fundamental requirement to empirical research is rigour, which means solid, documented and comprehensible data and analyses that are transparent to other scholars and can perhaps even be replicated. The evaluation of translation products is just like the assessment of texts in general, difficult and often subjective because of evaluators' different individual social and cultural backgrounds, their different ways of thinking, different presuppositions, perceptions, preferences, styles and tastes. Several questions have to be raised with respect to the selection of evaluators, assessment criteria and procedures, and the weighting of results, e.g.:

- Who can be asked to evaluate the translation products? Is it professional translators, colleagues from language and translation departments, other experimenters in process research, students,

bilinguals, native speakers of the target language, or perhaps clients or representatives of the intended target text receivers?

- Is it necessary to take into consideration the evaluators' professional and linguistic backgrounds, their qualifications, experience and theoretical orientation?
- How can misjudgement owing to cognitive constraints such as bad memory, inconsistency or fatigue be avoided?
- Should instructions for evaluation and/or the evaluation criteria be given and discussed in advance?
- Should evaluators be tested and trained in advance?
- Should evaluators focus on errors only or should the evaluation be a holistic one in which good solutions are taken into account as well?
- What are the advantages and disadvantages of weighting errors and grading quality in process research?
- What is the role of the project leader/researcher in the evaluation process, especially if he/she witnesses inconsistent assessment and/or divergent evaluations?

2. Evaluation practices in translation process research

Different approaches have been adopted for the evaluation of translations in translation process research. For research projects which aim at investigating translator behaviour, attitudes or strategies that result in *good/useful* translation products, an evaluation of the product is indispensable, e.g. Krings (1986: 26ff), Jääskeläinen (1999: 110), Pacte (2003: 43), Livbjerg/Mees (2003: 131), Hansen (1997, 1999, 2006a: 15) and Tirkkonen-Condit (1991; 2005: 412). In some projects, because of the issue investigated, an evaluation of the products is regarded as less important, e.g. Jakobsen (2003), who investigated the effects of think aloud on translation speed, revision and segmentation.

The development of Translog by Arnt Lykke Jakobsen and Lasse Schou (1999) has improved process research considerably. With the software, all keystrokes, changes and revisions during a translation process can be registered and replayed. In addition to an evaluation of the final product, the software makes it possible also to evaluate the quality of

changes and revisions *during* the process, e.g. Livbjerg/Mees (1999, 2003) and Hansen (2003, 2006a, 2008). In Hansen (2006a: 106f), this *product evaluation at a micro level* is an important parameter.

In this article, evaluation of products refers to both, the evaluation of final products and the product evaluation at a micro level.

3. Translation quality assessment in Translation Studies

In translation theory, translation practice and translator training, the evaluation of translated texts is one of the most discussed issues (Lee-Jahnke (2001: 259), Gerzymisch-Arbogast (2001: 237). Several issues of journals like *The Translator* 6 (2), 2000, and *Meta* 46 (2), 2001, have been devoted to this theme.

According to the purpose, different approaches to evaluation and translation criticism, methods and theories can be found. Theoretically this issue is discussed by, among others, Koller (1979), Reiß/Vermeer (1984), Neubert/Shreve (1992), Vermeer (1996), Gerzymisch-Arbogast (1997, 2001, 2004), House (1997, 2001) and Kupsch-Losereit (2008). For translation quality assessment for pedagogical purposes and for practical professional translation, see, for example, Nord (1987, 1998), Pym (1992), Hönig (1995), Kußmaul (1995), Stolze (1997), Schmitt (1998), Brunette (2000), Waddington (2001), Ørsted (2001), Schubert (2004), Mertin (2006), Mossop (2007a, 2007b), Künzli (2007), Magris (2009) and Hansen (1999, 2006a, 2009).

Furthermore, a large variety of national and international norms and standards of quality control defining translation processes and product quality have been established, like DIN 2345, the ISO 9000 series, SAE J2450, ÖNORM D1200 and D1201, ASTM F 2575 and the European EN 15038. They primarily apply to translation providers, translation companies and can serve as a basis for certification of the translation processes. Some of them, for instance, SAE J2450 and EN 15038, provide metrics or criteria for assessment of the quality of translation products. Stejskal (2009: 297) gives a description and comparison of existing standards and their applications. With respect to product evaluation, Stejskal (2009: 298) mentions that in industry-specific standards it is the end user who dictates what a "quality translation" is.

Depending on the purpose of the assessment, different requirements have to be met by the evaluators. In professional situations, sometimes the evaluation is restricted to an intuitive assessment by the client or by a representative of the group of end users. They evaluate the target text without any previous information or knowledge, and this seems to be sufficient (Ørsted 2001: 443). In other cases, meticulous translation criticism is asked for and metrics for assessment of the quality are developed (Mertin 2006).

4. Usability and acceptability of texts

As described in Hansen (2007), the purpose of translation quality assessment and the individual perception of quality in a communication situation differ considerably. Schmitt (1998: 394) mentions that quality in translation depends on the receivers' defined expectations and the adequacy of the product or service in relation to these expectations. Expectations, requirements and receivers' needs are stipulated by the translation brief, i.e. the communicative situation and the context in which the target text is used. The acceptability of a target text, however, depends on the target text receiver's perception of what is desirable in a particular situation, and individual tolerance levels regarding errors seem to be quite different. Apart from the term "acceptability", a good term in this connection is *usability* of a text.

We can never be sure that there is a direct relationship between the quality and the perceived usability of a text. There are communicative situations where poor quality seems to be welcome and even turns out to be an advantage, whereas, in other communicative situations, errors are absolutely unacceptable. The following two examples illustrate these two extreme cases.

An empirical investigation of the effect of flawed tourist brochures by Riis Christensen (1997), who asked 83 German tourists on the beaches of Jutland (Denmark) about their opinion of these brochures, showed that the tourists simply loved to read the brochures. They were amused by the deficient language. The flaws even confirmed their expectations, and made them feel being on holiday in a naïve, little, relaxed country. The communication between Danish tourist agencies and the German tourists

served the purpose perfectly – despite the extremely high number of errors in the translated brochures. An example:

(1) « Fisch- und krustentierabend »
In die Saison haben Wir ein ganz spezielles Abend, Mittwoch. Sie können Sich von einem Fischtisch, mit 15 – 20 Spezialiteten bedienen. Dkr 168,00
(Frei/Samstag im Ostern)

An example at the other end of the scale of acceptability and usability is a translation into German of an official Danish website, namely the presentation of the Danish Language Council (Dansk Sprognævn), which shows many linguistic and functional errors. As the sender was an official language institution and the receivers expected a flawless text, the translation was completely unacceptable. The following example, a human translation of the Danish source text, was accessed in 2006:

(2) Art des Instituts: Kulturministerieller Institution. Dänische Kulturministerium. Der Rat ist ein nationales Gremium des Kulturministeriums und hat in der Humanistischen Fakultät, Universität Kopenhagen, zu Hause.

These two examples represent extreme ends of the scale but, depending on the communication situation and receivers' expectations, in between these two opposite sides there are several degrees of usability and acceptability.

Interestingly, the same website accessed in September 2009 offers a machine translation into German which does not make much sense. At present, however, it is this kind of nonsense we *expect* when we touch "Übersetzen".

It is not always easy clearly to draw the borderline between "linguistic errors" or "language errors" (e.g. wrong choices as to grammar and vocabulary) and "translation errors" or "functional errors" (i.e. flaws that may have a negative influence on the reception of the message in the communicative situation, as stated by, among others, Pym (1992: 279ff.), Kußmaul (1995: 129ff), Waddington (2001: 314) and Martínez Melis/Hurtado Albir (2001: 281). The difficulty lies in the fact that grammatical and lexical errors usually have an immediate impact on the acceptability in the communication situation because the readers' attention is automatically directed towards them. Umbreit (2001: 253), for example,

has observed that linguistic errors affect the way a message is received because readers lose confidence in the text and cannot take it seriously.

5. The choice of evaluators

Systematic evaluation of translation products presupposes insight in the evaluators' profiles, i.e. their translation theoretical orientation, cognitive competences, attitudes and taste. The evaluators' orientation also has to be seen in relation to the theoretical orientation and attitudes of the translators/subjects. It is crucial to be aware of the relationship between quality assessment and the assessor's attitudes to functional translation, fidelity, loyalty, ethics, equivalence, norms, acceptability and the usability of translated texts because it is this relationship which has the most important impact on every assessment. For example, in relation to *addition or omission of information*, depending on the theoretical orientation, these strategies are regarded as welcome by some evaluators and criticised as errors by others.

The question as to whether the evaluators are bilingual or monolingual, and whether the target text is written in their mother tongue or in (one of) their foreign language(s) seems to affect their tolerance levels – mainly with respect to linguistic errors, but also with respect to their awareness of functional or pragmatic aspects. There seems to be a difference between evaluations by bilingual and non-bilingual evaluators, an observation, however, that would need to be investigated empirically. Notably, in the case of translations of domain-specific texts, the specialist background of the evaluator is crucial.

The best-case scenario is for several competent evaluators to undertake a consensus-oriented assessment, so that bias as a result of individual subjectivity can be reduced. "Competent" here means having the ability not only to spot errors but also to describe and explain the assessment criteria and to justify the decisions made.

As mentioned earlier in this article, a type of evaluation used by some translation companies is *client assessment* by the end users or by potential target text receivers. An advantage of this approach, in which the target text is often evaluated without access to the source text, is that we obtain the target text receivers' spontaneous, subjective reactions regarding

the usability of the text. However, a study I carried out (Hansen 1999: 57) has shown that these evaluations without the source text are not reliable. Assessments of this type can perhaps be useful as a supplement to other evaluations. In translation process research, they are problematical, not only because they are carried out without the source text but also because the potential receivers often do not have the necessary concepts at their disposal to explain and justify their observations and assessments. Obviously, one has to be trained in the evaluation and revision of (translated) texts (Hansen 2008, 2009).

5.1 Cognitive aspects

As in other situations where texts are evaluated, inconsistency is also a problem in translation quality assessment for translation process research. Evaluators do not always spot identical errors in different products, or they evaluate them differently. It even happens that the same evaluator assesses an identical product (by the same translator) differently a week after the first evaluation (Pavlović 2005). Such inconsistencies can be due to fatigue and/or decreased attention. Some evaluators admit that they get used to the errors and that they become more tolerant the more translations they read. That is why evaluators sometimes insist on reading only one or two products a day, so that they can be as "objective as possible" – as one of my evaluators (Hansen 1999: 51) expressed it. As inconsistency and fatigue seem to be proportional to the number of products evaluated, the texts can perhaps be presented to the evaluators in different orders.

5.2 The role of the researcher in the evaluation process

An important question is if researchers should evaluate the translation products themselves or if they should merely rely on the assessments of an appointed group of evaluators in order to avoid bias from observers' interests. In my experience, it is an advantage if researchers also evaluate the translation products themselves since it is then easier to control what is very often a complicated evaluation process, especially if there are discrepancies between the evaluators' assessments. Observers' influence and bias can be minimized by a precise description of divergences between the evaluations and the criteria applied. Product evaluations at a micro level

which draw on translation process data can be difficult because one has to be trained in reading log files. They should be carried out by the researcher or a competent colleague, but also in accordance with the criteria the evaluators have agreed on.

6. Grading errors and ranking quality

Even if evaluators agree on the quality criteria, they may still have different ideas about errors and their gravity. One problem in this connection is the lack of agreement there may be among evaluators on the units or linguistic entities assessed. This has an impact on a consistent analysis and interpretation of the data. In the case of unwarranted *omissions*, for example, notably in process research under time pressure (Hansen 2006b), it has to be taken into account that evaluators make their judgements using different entities. Some count every missing word as one error while others are more tolerant and work with larger entities. For process research, it is crucial to discuss this problem with the evaluators and to reach an agreement.

A fundamental question is whether it is only errors that should be registered and evaluated or whether *holistic evaluations* are needed in which good translation solutions are taken into consideration (see Martínez Melis/Hurtado Albir 2001). In the latter case, we would need precise criteria for "good" translations since statements like "fluent", "partly fluent" or "easy to read" (e.g. Jääskeläinen 1999: 112) depend on individual, subjective perception.

There are several approaches to holistic descriptions, e.g. Gerzymisch-Arbogast (2001: 230ff; 2004: 69ff) who suggests that network-based parameters of coherence and thematic and isotopic patterns are included in the catalogue of evaluation criteria. Williams (2001) proposes the application of argument macrostructures complementary to other approaches of translation quality assessment based on the theory of argumentation by Toulmin (1964).

However, for translation process research, the situation is still that good translation solutions are assessed differently by different evaluators and that there is a lack of reliable procedures for holistic assessment,

because not only "good" or "successful" would have to be clearly defined but also the weight of good solutions in relation to flaws and errors.

In many projects, in order to get an overview, ranking levels such as "good", "acceptable" or "not acceptable" are used (Hansen/Hönig 2000). However, as Gerzymisch Arbogast (1997: 576) points out, these scales and their ranks also need to be defined and properly described.

7. Instructions, preliminary definitions, draft translations

One way of dealing with the above-mentioned problems and avoiding the risk of bias would perhaps be to give the evaluators predefined criteria and guidelines prior to their assessment. Martínez Melis/Hurtado Albir (2001: 283) mention several kinds of evaluation, "intuitive assessment" without criteria and "reasoned assessment" with objective criteria.

If evaluators are not given the criteria to be applied, they will use their own criteria. However, if they receive predefined criteria and guidelines from the researcher, the results may lack validity because of potential bias from the researchers' interests. Waddington (2001) investigated evaluation processes in which the evaluators not only got precise criteria in advance but were also trained in how to apply them. One of the results of his study was that this method seemed to augment criterion-related validity (2001: 322ff).

A procedure I would like to propose is an interaction between an individual spontaneous assessment according to the evaluators' own norms followed by a description of the criteria applied by them and a discussion of their results. They have to be able to describe problems encountered during the evaluation process and to justify their decisions. As a next step, the individual problems, decisions and results could be compared and an agreement reached on a set of shared quality criteria which are subsequently applied systematically to all the translation products (Hansen 2006a: 112). During the dialogue, different theoretical orientations, attitudes and ideas can be clarified and an agreement can be reached.

The evaluators can also be asked to try to solve the translation task themselves. This increases their sensitivity to special characteristics of the source text and potential translation problems (Hansen 1996: 156). In addition, they become more alert to other translators' good, acceptable or

poor proposals. It is easier to recognize and describe good formulations when they can be compared with other formulations which are either linguistically and stylistically neutral or not successful.

8. Conclusion: transparent and comprehensible evaluation procedures

Because of the general insecurity with respect to the grading of good translations, holistic procedures and especially a weighting of good solutions against errors is not reliable. Consequently, I would like to propose the following two assessment procedures which are based on documented and described errors: a systematic evaluation and a spontaneous evaluation. Criteria and errors are described in both procedures and the results are transparent and comprehensible.

The *systematic evaluation* is an evaluation of the errors in a translation product all the evaluators have agreed on after a discussion of the criteria, problems and divergences. After they have reached agreement, the criteria are applied systematically and identically to all the translation products in the study. Similar errors have to be dealt with in exactly the same manner in all texts wherever they occur. This procedure can be regarded as a relatively precise evaluation method, but it is extremely time-consuming. An advantage of the systematic evaluation in process research with Translog is that the same catalogue of criteria can also be used for product evaluation at a micro level.

The *spontaneous evaluation* is an assessment of quality based only on those errors all evaluators spot and agree upon spontaneously. With this procedure, there is no discussion, but the researcher can ask the evaluators to provide a precise description of the criteria applied. This evaluation seems to be reliable; however as it is based on the means of errors and flaws, errors which are marked by only *one* of the evaluators are neglected. The reliability of the spontaneous method to a high degree depends on the evaluators' background, attitudes and qualifications with respect to quality assessment.

References

Alves, F. (ed.). 2003. *Triangulating Translation*. Amsterdam/Philadelphia: John Benjamins.

ASTM F 2575-06. 2007. *Standard Guide for Quality Assurance in Translation*. West Conshohocken (PA): ASTM International.

Brunette, L. 2000. Towards a terminology for translation quality assessment: a comparison of TQA practices. *The Translator* 6(2): 169–182.

DIN 2345. 1998. *Übersetzungsaufträge*. Berlin: Beuth.

Danish Language Council (Dansk Sprognævn), Website: www.eurfedling.org/dan/dande.htm (accessed 02.02.06 and again 09.09.09).

EN 15038. 2006: *Übersetzungsdienstleistungen – Dienstleistungsanforderungen*. Berlin: Beuth.

Fleischmann, E., Kutz, W. & Schmitt, P. A. (eds). 1997. *Translationsdidaktik*. Tübingen: Narr.

Fleischmann, E., Schmitt, P. A. & Wotjak, G. (eds). 2004. *Translationskompetenz*. Tübingen: Stauffenburg.

Forstner, M., Lee-Jahnke, H. & Schmitt, P. A. (eds) 2009. *CIUTI-Forum 2008. Enhancing Translation Quality: Ways, Means, Methods*. Bern: Peter Lang.

Gerzymisch-Arbogast, H. 1997. Wissenschaftliche Grundlagen für die Evaluierung von Übersetzungsleistungen. In Fleischmann *et al.* (eds) 1997. 573–579.

Gerzymisch-Arbogast, H. 2001. Equivalence parameters and evaluation. *Meta* 46(2): 327–342.

Gerzymisch-Arbogast, H. 2004. Dimensionen textnormativer Äquivalenz. In J. Albrecht, H. Gerzymisch-Arbogast & D. Rothfuß-Bastian (eds). *Übersetzung – Translation – Traduction. Neue Forschungsfragen in der Diskussion*. Tübingen: Narr. 67–79.

Hansen, G. 1996. Übersetzungskritik in der Übersetzerausbildung. In A. G. Kelletat (ed.). *Übersetzerische Kompetenz*. Tübingen: Peter Lang. 151–164.

Hansen, G. 1997. Success in translation. *Perspectives. Studies in Translatology* 5(2): 201–210.

Hansen, G. 1999. Das kritische Bewußtsein beim Übersetzen. *Copenhagen Studies in Language* 24: 43–66.

Hansen, G. 2003. Controlling the process. Theoretical and methodological reflections on research in translation processes. In Alves (ed.). 25–42.

Hansen, G. 2006a. *Erfolgreich Übersetzen – Entdecken und Beheben von Störquellen*. Tübingen: Narr Francke Attempto.

Hansen, G. 2006b. Time pressure in translation teaching and translation studies. In S. Kasar Öztürk (ed.). *Interdisciplinarité en Traduction. Vol. II*. Istanbul: Isis. 71–80.