[*Terminologie et Traduction* 1. 1994; pp.375-415]

Machine Translation Evaluation and Quality Benchmarks

Antonio Torrens

Table of contents

SUMMARY

The present paper has been deliberately written in a way that makes it ideally suited for discussion - following a logical pattern from the most general to the most specific. The subject is complex and largely controversial, so there would be no point in going straight into the discussion of the very tangible concluding proposal for immediate action. The whole range of previous issues leading to that particular end has to be covered first.

In this summary, however, it is practical to follow the opposite order. Nevertheless, before doing it, we should like to put in a word of caution against premature conclusions drawn from such a peculiar fast track. This paper is the result of nearly one year's work, with ample reading and rereading, informal discussions with over half a dozen experts from different areas, and some original thinking. It should therefore deserve some careful consideration.

Starting with the final piece of the proposal, we recommend that action should be undertaken in three stages. First, a newly set up interdisciplinary permanent board should produce *the basic factor* for some language pairs (three, at least). The underlying notions could be borrowed from Diagram 1. Second, mainly *synchronic* intersystem and interpair comparisons could be expressed by a *Formula X* linked to results supplied by proficiency test batteries measuring overall performance, in a way in which *reliability* and *validity* (both terms used with their unambiguous restricted meanings) were similar to the high coefficients reached in so-called *standardized tests* - which educational testing experts utilize with humans. Third, *two further distinct formulas* could be produced: *Y*, for the core of subjective qualified assessment addressed to decision makers (e.g. purchasers); and *Z*, for progress development checks. As Z could be defined more easily, work on Y must be expected to bring results last.

In the *Introduction*, the importance of usefulness is stressed. *The key assumptions* (KAs) are deemed to be inevitable for a fruitful discussion, which cannot take place until basic disagreements disappear. KA 1 proposes *clarity*, KA 2 covers the relevance of *purpose*, and KA 3 gives a realistic *scope*. KA 4 defends the feasibility of *objectivity*, which must be attained with the sort of balanced *interdisciplinary* view defined under KA 5. The limits of *pseudotranslation* are recalled in KA 6. Finally, the methodological requirement of *fair comparison* is backed in KA 7, with KA 8 adding that *empiricism* is the best way to avoid endless disputes based on opinion.

The somewhat 'philosophical' section 2 is followed by the more 'technical' preliminary concepts of section 3. Three different formulas *(X, Y,* Z) must follow pragmatic research confirming and refining the level system of *Diagram 1.* Section 4 presents an emphatic demand that production and administration of testing materials be not attempted without adequate knowledge, understanding and implementation of the *meticulous specifications* that foreign-language testing experts are familiar with. They consist of different types of *reliability* and *validity* concepts and quotients, as well as well-defined *norms.* Additional recommendations of section 5 cover the know-how contributed by testing experts and by several authorities from the MT field.

A glimpse of section 7 could quickly tell demanding readers to what extent some modest kind of 'interdisciplinary wisdom' can be kept within realistic limits and has actually been pursued in this author's writing, particularly influenced by the reduced amount of selected sources recorded in *References.*

1. An introduction to the topic of MT evaluation

MT evaluation is a fascinating topic because it is so confusing, complex and controversial. The term *evaluation* is not very precise by itself, nor is it always clear what different people mean by it. After having read a significant amount of the overwhelmingly large volume of literature claiming some concern with the subject, two striking features become predominant. One is that there is very little consensus. Doubt, hesitation, confusion and chaos seem to prevail. The other striking feature is plain ignorance of the possibility of resorting to the knowledge and experience of the people who should know best - language testing experts. So many MT professionals just do not seem to know that a branch of Applied Linguistics specializing in language testing has existed since the '60s! Language testing covers all forms of linguistic competence and performance, so it can cover translation as well. If it covers translation testing, why should it not cover testing all forms of translation, including translation performed by computer systems?

Let us also recall the need and importance of MT evaluation. MT is quicker and cheaper than human translation, so it should replace it whenever it can be *proved to be good enough for certain limited purposes.* The argument is perfect and no-one will be able to object to it totally. Nevertheless, there *is* resistance. Leaving human and labour problems aside, are there other reasons for MT having only a rather limited success or even being subordinated to some other similarly cheap translation procedures, such as dictating an improvised, approximate translation at interpreter's speed, for example?

In our opinion, there *are* valid reasons, apart from psychological and labour matters. One reason is that the opening argument contains a lot of ill-defined ideas- *good enough,* what for?; *certain purposes,* which ones?; *proved,* how and by whom? All these built-in questions are linked to evaluation, which should provide the corresponding answers and play a crucial role in any promotion activities.

The fact that the advantages of using MT do not always seem to be known or convincing to everyone means that it needs promotion. But promotion involves product description, rather than passionate praising by designers or sellers. Product description, in turn, implies the possibility of being objective, complete and trustworthy. The requirement is already fulfilled for the informatic side (memory size, computer type, communications, input and output devices, etc.) but it has to be fulfilled for the purely linguistic side too - i.e. quality and usefulness of the output text.

The previous remarks about confusion are not new. Not only have evaluation methodologies and approaches of the European Commission changed probably too often in the past, but so have the corresponding policies as well. Pigott (1991) was right in writing that 'perhaps the most difficult thing of all to assess is how long a given development and implementation plan should last before objective results can be obtained'. But what 'objective results' meant should have been defined and still has not.

A somewhat similar complaint was published more recently, in Hutchins and Somers 1992 (p. 161). 'What may be surprising is that despite some forty years of research on MT there is still no generally accepted methodology for the evaluation of systems'.

From the preceding paragraphs we can already draw some decisive preliminary conclusions. Besides the need and importance of MT evaluation and of conventional translation evaluation, which perhaps do not need any further emphasis here, other essential points are to be made from the outset:

(a) Basic definitions are needed to overcome confusion. This is why we shall start our proposal with a presentation of *The key assumptions.*

(b) What seems obvious to some is often different from what seems obvious to others, so almost nothing must be taken for granted now, but discussed and proved (below or elsewhere). This concerns any controversies that may arise once again in the discussion of this paper. We should all bear it in mind as a prerequisite for the only kind of constructive dialogue that can lead to progress. Thus we have another argument in favour of starting with some *key assumptions.*

(c) Expertise in language testing must not continue to be ignored. On the contrary, it is about time it should be exploited. This is why some works providing a formal specialized background are listed under *References.* The necessity of covering this field when setting up a board or a jury must be underlined too. It would be absurd, for example, to undertake a study of the quality of food samples without letting the experts in chemical analyses have their say. For the same reason, we should not try to make progress in the area of quality assessment without resorting to experts in foreign language testing.

The current picture of quality assessment could be compared to a society where a set of moral principles has been accepted but laws still have to be made or adapted. In the case of MT evaluation, ready-made solutions are not available because they are not available for conventional translation either, in spite of the fact that this kind of writing activity has already been performed for some five thousand years. It is therefore inevitable to undertake the long enterprise of supplementing the most widely accepted 'moral principles' (grammars, contrastive studies) with implementation laws (from translation theory and practitioners' experience). An additional difficulty comes from lack of unanimity in the selection and interpretation of the 'moral principles' (the existing knowledge). The obstacle can only be overcome by appointing 'legislators' and 'judges' (a board of evaluates issuing rules and interpreting them), for we have not got the laws and the judicial system yet.

2. The key assumptions

Sharing all of the key assumptions defined below is a requisite for a fruitful discussion leading to the right decisions. This does not mean that it is totally impossible to change this foundation. What *is* impossible is to have a sensible dialogue if there are any basic misunderstandings. This is why we must start by defining the most essential assumptions, even if we feel they are very obvious and unchangeable. Should any changes be made, the rest of the paper ought to be changed accordingly.

Only after agreeing on these or other functionally equivalent decisive assumptions will it make sense to deal with the less philosophical yet still preliminary concepts in section 3, before actually going on to the purely technical crux of the matter in the later pages on *principles and specifications.* Impatient readers may find this long road to the more practical part of the proposal slightly annoying, or may even wonder why some of the starting assumptions are written down, doubting their relevance. This author feels that the problems covered are at the root of the subject matter and, finding them inevitable, prefers to follow what he thinks is the most logical sequence to build up agreement for action efficiently.

2.1 Key assumption No. 1 (KA 1) - CLARITY

The badly needed initial clarification, which has already been proposed in the *Introduction,* involves leaving no room for vagueness or ambiguity anywhere. This belief concerns the overall design of the present proposal (an attempt at providing a rather comprehensive review of ideas, from the most abstract to the most specific), the clearly defined contents of each section of the proposal, and an inventory of relevant pairs of affirmative and negative statements that are recorded immediately below, either because they are fundamental or because they would not fit the logical pattern of later paragraphs. The inventory is open-ended and should be enlarged in the light of formal discussion and the experience with implementation attempts in the future.

FIRST CLARIFYING STATEMENT

Remaining 'neutral' is not possible.

Whenever a controversial point arises, evaluators must take sides, be it after scientific research or after recording an opinion as a working hypothesis.

For example: this proposal *will* take sides in the definitions of overall translation quality and sufficient translation quality.

SECOND CLARIFYING STATEMENT

Testing and *assessing* are not synonymous.

One can assess combined skills and general phenomena but must test only comparatively *simple* skills and *specific* components.

For example: you cannot test a translator's or a system's *competence,* but *performance* only (for particular tasks and components). Assessment comes later, after sufficient testing, always involving some inevitable risk, whereas testing is basically a way of checking facts indisputably.

THIRD CLARIFYING STATEMENT

MT evaluation cannot be seen as a whole without dealing with separate comparisons first.

There are three essential comparisons, with each one requiring distinct techniques: intersystem comparison, interlanguage comparison, interstage comparison (development progress).

For example: testing instruments designed for intersystem comparison (a given kind of performance test for a particular language pair in two different MT systems) are not useful, in principle, for interstage comparison (the same language pair later, as handled by the same MT systems).

FOURTH CLARIFYING STATEMENT

Overall performance evaluation is not equivalent to advice for purchasing purposes.

The latter covers far more ground and should only use the first as a very meaningful piece of information.

For example: purchasing advice usually involves development potential, which is closer to predicting and guessing than it is to testing. Performance evaluation must precede buyer advising and must keep its own features.

FIFTH CLARIFYING STATEMENT

Specifically human features cannot be included as criteria in MT evaluation instruments, no matter how important they are.

In comparing translation performance by computers and by humans the testing instruments must disregard creativity, writing talent, sensitivity needed for register and style recognition and imitation, extralinguistic context, and knowledge of the world.

For example: there is no point in conducting a comparison of machine translation and human translation concerned with style - a mistake that has been made in the past.

SIXTH CLARIFYING STATEMENT

In testing instruments, suitable discriminatory power is not to be mistaken for the validity of criteria.

Efficient measuring devices tend to have a rather limited discrimination range.

For example: good translation tests for competitions among humans are not efficient within the lower range applying to MT intersystem comparisons or to ranking of individuals at the end of a first-year foreign language course, even if such criteria as knowledge of the morphosyntactic codes and of the semantic system are definitely valid for both extremes of the huge factual range.

2.2 Key assumption No. 2 (KA 2) - PURPOSE

Our problem will not be evaluating for the pleasure of it, or for no practical purpose. The activity of evaluating always involves a purpose, which is part of the evaluation concept itself. We must look for a reply to a question, therefore there has to be a question. The clearer the question, the better. In this case the first question is: to what extent does the object of the evaluation fulfil a given task?. Is *X* useful for the purpose it was designed for?. How useful is it (distances to uselessness and to perfect usefulness)?

When dealing with translated texts, the ultimate purpose is synonymous with the perfect attainment of *the translator's ideal: production of a text allowing readers who do not know the source language at all to get the same message they could get if the language barrier did not exist.* But more modest partial purposes can often make sense too. The top standard expectable for book publishing need not be the only translation standard.

What we are trying to underline here is that rather than stating 'this translation is good/bad' or 'this translation is better/worse', evaluation involves initial recognition of the quality level actually needed for a precise goal, which may very well be lower than the quality level usually assumed to be necessary, followed by measurement of the distance between the desirable and the attained levels. The idea will be taken up again in *Diagram 1* (see next section, on *Preliminary concepts),* where the possibility of having 'superfluous' quality is shown implicitly. Such superfluous quality may very well be too expensive to attain, but it may also be exceedingly expensive to measure. Neither objection should be forgotten by evaluation theory or practice.

2.3 Key assumption No. 3 (KA 3) - SCOPE

The desire to make effective progress makes it advisable to restrict the scope of the evaluation problems we shall be concerned with at the beginning of the action period following acceptance of our proposal. If additional ways of exploiting our assumptions, principles and ideas can be found to exist or can be devised later, by producing similar materials with the same source of inspiration, it will certainly bring us satisfaction. But for the moment we must definitely start with a limited realistic scope, no matter how modest.

The restricted evaluation scope deliberately chosen for this paper, in order to let its author make a significant initial contribution that stays within the limits of reasonable feasibility is this:

Measuring, for multi-purpose comparisons, the communicative functional quality of translations provided by MT systems that are operational, fully automatic and suitable for general texts.

As the paragraph above has been worded very concisely, it is probably worth some elaboration. But we prefer to leave it for a later occasion.

There could seem to be some contradiction between the ambition of defining the concepts allowing multi-purpose comparisons and the simultaneous claim that we should have a scope that is limited to texts produced by try-anything, fully-automatic operational MT systems. The explanation is easy. We are aiming at clarity and methodological simplicity while choosing a definition of quality, just as we select a single kind of product for the first few experiments, where only a limited scope can be covered. Nevertheless, this is compatible with the fact that, after becoming capable of measuring a precise concept of quality (worked out via comparison) with a given type of translation product, we want to extend the know-how to all sorts of meaningful comparisons that appear to be compatible with it.

You could also think that we are contradicting our previous Third Clarifying Statement (in KA 1), where we argued that each comparison requires a distinct technique. The fact is that there we were talking about testing materials, while here we are presenting an approach justifying a recommended chronological sequence for action.

Comparisons must always be based on the answer to one main question, which was precisely our previous one (KA 2) - usefulness. Usefulness for a given purpose implies the underlying notions leading to the detailed description of communicative functional quality. With their help we can avoid falling into the trap of endless academic discussions and we can build a system of quality levels consisting of both benchmark definitions and illustrative samples. Our starting attempt will be presented in a workshop in early 1994.

Objective comparisons should thus be made possible for any purposes whatsoever - different MT systems, different language pairs, same pair at different times, same pair for different text types or different semantic areas, same text before and after post-editing, machine translation vs. human translation, human translation by different individuals. We favour a single universal definition of functional quality which is only linked to the very general multi-purpose task of comparison.

But there are all sorts of secondary questions as well, still within the conceptual area of scope of the evaluation activity. Why is a comparison undertaken? How will the results be used? What for? At this point we must not even try to list all the foreseeable questions and their possible answers, for we are merely stating a *key assumption* as a pillar for later discussion. Nevertheless, we ought to admit

already it is all those *secondary* questions that ultimately justify this whole work. Their importance is by all means undeniable and no-one is questioning it. Our point here is that we find it necessary to keep the problem of universal comparable quality levels and the secondary questions totally apart, as a condition for clarity and objectivity.

What we have called secondary questions must be clearly formulated and we shall look for the answers too. Both questions and answers at such secondary level are in any case essential, as admitted before.

The second part of our assumption is therefore that the primary question of quality and any secondary questions are different and the first reply must not be contaminated by the rest. Clarity required for the first reply is equally indispensable for the rest, but they represent separate issues to be dealt with separately (from the beginning, at the conception stage; and at any later stages). In the present paper the primary problem of quality comparison is summed up in *Diagram 1* below, whereas the more specific secondary questions can be found in the section on *Technical principles.*

2.4 Key assumption No. 4 (KA 4) - OBJECTIVITY

Objectivity is possible, somehow. If readers are willing to agree that, once it has been carefully defined, intelligence can be measured (IQ tests developed empirically by Psychometrics), or if they admit that one form of *happiness* can be measured (standards of living), and that even the appreciation of beauty can be fairly unanimous in extreme cases (Saddam Hussein handsomer than King Kong, Lord Byron's poems more moving than this paper), then why should we not all admit that it is possible to judge the quality of translations *objectively? What we have to do is provide the standards and examples that can be used for reference purposes. After that, it will still be necessary to define distance units between standards, ways of measuring such distances and relationships, and finally ways of overcoming the wide gaps of personal interpretation.* It is not a simple task, but it is not a dream either. The knowledge and techniques are already available. What is left to us is getting to know them and trying the best possible combinations and adaptations, as we shall see in the next key assumption.

Anyone could argue that measuring quality against previously set standards is not really objective, philosophically speaking. This has to be granted. But it is the only objectivity we can strive to attain, and it is not worse than any other popular scientific methods, such as measuring speed or temperature.

The main difference is that for speed you can choose between m.p.h. and km.p.h., for temperature you can choose between Celsius and Fahrenheit, for earthquakes you can choose between Mercalli and Richter, but for translation quality there is nothing but vagueness to choose from.

Let us then provide a standard and give it a name - e.g. *'The European Commission's Functional Quality Standards for Translations'*. They could be defined by a board of evaluators on the basis of the present proposal, which in turn is linked to the communicative efficiency approach. After sufficient experimentation and official endorsement by the Commission's Translation Service, the *EC's quality standards* would become a publicly available norm and the world would be free to use it or not to use it, to improve it or to replace it, or to propose additional choices.

The implementation of this idea would represent a leap forward. The sad truth is that we have not reached the stage where one can discuss the advantages and disadvantages of measuring speed in miles or kilometres. We still do not know how long a mile or a kilometre is, nor have we defined *hour* by dividing days into twenty-four equal periods.

With definitions of kilometres and hours we will be able to be objective. The first step in preparation for the leap could be made by improving and completing an existing draft that will be submitted to attendants of the workshop mentioned before (sponsored by DG XIII).

Trying to get official EC standards will prove difficult enough. Let us not make the enterprise even more difficult with a struggle for worldwide cooperation and recognition. Is the Commission's Translation Service not the largest of its kind in the world, anyway?

The draft that this author is working on is entirely based on the belief that objectivity is possible.

2.5 Key assumption No. 5 (KA 5) - INTERDISCIPLINARITY

Machine translation evaluation is an interdisciplinary task involving specialized knowledge from several areas:

- general and applied (contrastive) linguistics
- translation theory and practice
- translation teaching and marking (with learners and translators)
- educational statistics and foreign language testing
- computational linguistics and MT
- functional communication and translation use

It is certainly not possible to find a large amount of all this knowledge and experience in one individual. Besides, it would not be a good policy to rely on one person's judgement only, anyway, no matter how qualified a particular person could be. We shall come back to this when we propose setting up a board, in the *concluding proposal* of section 6.

As it has already been pointed out in the *Introduction*, the theoretical and practical contributions from the language testing field should cease to be ignored - very logically. But this point was emphasized there and we no longer need to isolate it. Instead, let us elaborate briefly on the desirability of benefiting from the expertise accumulated by each of the areas listed above. It is the sum total of this interdisciplinary knowledge that evaluation must rely on.

The process of translation is a very complex task requiring essentially a very good cultural and linguistic background (in both the source and target languages) plus creativity, imagination and writing skills. Evaluating the *product* (a translated text) involves a particularly thorough understanding of the aforementioned process and good critical skills, supplementing it all with the knowledge and know-how of the experts in mental measurements (psychometrics, educational statistics) and in verbal performance (language testing and foreign language testing). The seemingly widespread opinion that a couple of MT experts or translation experts can reach valid conclusions about the quality provided by a system is a mistake, because it is an oversimplification. Valid judgement must be the result of the joint knowledge contributed by different kinds of experts on the basis of an explicit evaluation theory. The theory is still partly missing and the present paper represents only an initial attempt to make some progress towards a consistent collection of well-defined ideas. The subsequent critical work of judging will have to follow the completion of such a collection, but we already know that both stages will require a whole variety of knowledge and techniques that do not fall totally within the domain of a single field among the ones mentioned before. Because of this, if some sort of a *board of examiners* is ever set up, it should not consist of overspecialized experts from one or two areas only, in order to prevent their judgement from being biased, incomplete or insufficiently qualified.

MT experts and specialized informaticians can undoubtedly be useful, for no sensible work is at all possible without their familiarity with the hardware and software needed for any MT system to be operational and successfully run any kind of natural language programs. But observation by translation requesters and communication experts is equally fundamental, since they have the best understanding of what is useful and to what extent. Joint knowledge from those two groups is, however, insufficient, as there is no point in denying that translators and translation theorists are more aware of the language problems involved, as well as of the techniques, limitations, phenomena, formal categories and definitions.

What translators often lack is a formal background in linguistics (general and contrastive), particularly in the case of practitioners specialized in the contents (e.g. medical doctors, engineers). This is a handicap that could show whenever scientific descriptions of facts became inevitable, as in the concepts and terms for tasks and problems to be handled in the mutual understanding needs of the people working on the same assignment.

Not all translation theorists and practitioners are accustomed to objective marking of exercises and many are prejudiced, in different ways. So it would be wise to count on the attitude and sensitivity of people who knew all the tricks and difficulties and shared our belief, confirmed empirically, that objective scores can be obtained, somehow, under carefully controlled circumstances. Satisfactory experience in institutional marking of exams written by translation learners and professionals is not a negligible asset.

Finally, let us point out that objective marking of translation exercises is only one of many aspects covered by foreign language testing, where the technical notion of intermarker reliability has been well studied and researched, as have many other relevant features and concepts, like size reliability, guessing and sampling, statistical refinement, item analysis, validity criteria, influence of extraneous factors, norms, validation processes, etc.

2.6 Key assumption No. 6 (KA 6) - PSEUDOTRANSLATION

Humans can translate; computers only *appear* to translate. The raw result of the process performed by fully-automatic MT systems is not really a *translation*, but a *pseudotranslation.*

Translation theorists (see *References)* tell us, in different ways and with different words, implicitly or explicitly, that the process of translation implies grasping the meaning of a text and then conveying the same meaning in a different language. And yet computers cannot grasp the meaning of anything, so they actually *translate* without understanding, i.e. they do not translate but do something else instead. This alone could very well explain why 'fully automatic high quality translation is not at present possible', as admitted in Hutchins and Somers 1992 (page 161) and by most professionals with a general background in translation theory.

For the comprehension stage, translation implies handling a lot of information that is far beyond the capabilities of central processing units and programming devices (e.g. who is writing something and why, which factual and psychological circumstances influence the meaning, all sorts of unrecorded information resulting in a precise reader's expectable interpretation, etc.). For the writing stage, creativity, talent, humour and emotions are exclusively human.

However, there *are* texts that are so extremely easy to interpret *literally* that a parallel version, drawn up *literally* (again), in a different language can be easily and quickly produced by both humans and computers. A superficial look at the result can create the illusion that the text produced by the computer is also a translation, but we prefer to give it a different name, as a reminder - *pseudotranslation.* It is true that the prefix is often used with a pejorative meaning that could conflict with our purely descriptive intention. But what really matters is the semantic distinction made, which reflects a decisive contrast.

The reason why this fact is underlined here is its relevance for MT performance testing. Good pseudotranslations are based on the luck of finding sentences where the meaning does correspond to what had been foreseen by humans designing an MT system and writing instructions and data within the range of formal specifications comprehensible by the system's CPU. It follows that problem listing must give enough weight to this limitation.

Test writing is based on an initial selection of problems. When you test human translators you produce the usual list of problems involving form recognition, false friends, overall comprehension, writing skills, peculiar difficulties spotted in teaching and in contrastive linguistics. When you test the performance of a computer system the weight and treatment of the previous kinds of problem is of course different, but above all you have to check additional problems that are typical of pseudotranslating - homographs, difference between names and nouns, semantic ambiguity, syntactic ambiguity.

The computer's lack of intelligence also results in the necessity of checking the input text very carefully, to eliminate human input errors (in grammar, spelling, punctuation) but also to confirm that its *translatability* is high and above a peculiar modest threshold (not in the lower translatability levels of contracts involving different national laws, or of political speeches full of subtleties and emotions). A more detailed discussion of extraneous factors and translatability can be found elsewhere, when dealing with the reliability factor (section 4).

2.7 Key assumption No. 7 (KA 7) - FAIR COMPARISON

The justification of MT is mainly economic. Since MT is quicker and cheaper than conventional human translation, it should be preferred - for the core of the translation process, at least whenever its performance can be confidently expected to be good enough for certain limited purposes, such as internal information or informal texts for unusually tolerant readers. The clue is precisely the expression *good enough.* It implies a notion of quality and a notion of quantity, both totally connected with the overall goal of the present pages.

The notion of translation quality is not new and what remains to be done is attaining a precise definition, after which different degrees can be established accordingly and the problem gets solved. But *the task obviously implies measuring degrees of fulfilment of one and the same concept of quality.* Would anyone think that, in order to decide if using a bicycle is preferable to using a car for a given short trip, the comparison of time could be based on different systems of speed measurement? It does sound absurd and useless, but this is exactly what has sometimes been done or proposed in the past, under different circumstances that possibly justified it then and there.

When deciding to use a bicycle instead of a car it must be because the slowness of the bicycle is compensated by the savings in the initial investment and fuel consumption, and anyone will agree that the choice will be made after comparing the speed in km.p.h. or m.p.h. for both vehicles. The comparison will tell anyone exactly how slower the bicycle is and will allow him to be sure that it is not too slow. No-one would consider the possibility of making a decision on the basis of measuring incomparable factors, such as the car's aerodynamic coefficient and the price of the lubricant required for the bicycle, or factors that are comparable but irrelevant, as is the vehicle colour.

When measuring reading time or intelligibility of translations, to give only two of the silliest examples, we are applying a peculiar notion of translation quality that would not make any sense with human translation. Reading time depends mainly on factors outside the quality of the translation process and its result - it depends mainly on the quality of the original and its kind of contents, as well as on the reader's skills, purpose and familiarity with the subject. Intelligibility is an essential requisite and measuring the lack of it is close to ridiculous - if a text is not intelligible, just throw it away, be it original or translated.

Everyone knows bicycles are slower than cars and nevertheless preferable under certain circumstances, but no-one dreams of comparing speeds by using different criteria. In translation, everyone knows that a computer's translation cannot be so good as one produced by a trained mind, but it can nevertheless be preferable, because of its economic advantages, whenever its usefulness is satisfactory. The fulfilment of this condition has to be checked *normally*, not with ad-hoc tests intended to play down the importance of very real shortcomings.

A basic mistake to be avoided is the use of different evaluation criteria for differently produced translations. Comparability requires identical evaluation criteria. The difference between MT and conventional translation can only lie in the amount of tolerance with poor results.

Our previous refusals are not gratuitous.

Measuring irrelevant criteria is reported in Van Slype 1979, for example, when he recalls (p. 159) a case where 33% of the sentences in raw MT samples were found to be unintelligible. This information may interest system designers struggling to move on from initial experiments to operational status, with the latter simply involving the requisite that practically all the sentences of a translation must be as intelligible as the original. The criterion may in any case be useful with embryos of MT systems, if comparison of extremely poor translations necessitates looking for consolation and encouragement by measuring the slow reduction of the very long distance to be covered before attaining satisfactory performance, where intelligibility can be simply taken for granted.

In Vasconcellos 1989a we read that 'there is no point in subjecting it [MT] to the approaches used for evaluating human translation'. Although the statement is acceptable within the context of the introductory paragraph in the first section of the article, it is worth mentioning that this can become an excuse for passionate defenders or attackers of MT to undertake unfair evaluations where criteria selection is used to confirm prejudice on either side of the controversy. The whole problem is a matter of focus. When focusing on the lower end of translation quality, in designing testing tools with the right discrimination power at that level, some criteria are very sensibly left outside - which does not contradict the fact that they are *always* relevant, even if omitted for practical reasons.

It goes without saying that neither of the two preceding paragraphs is intended to be derogatory to the well-known authorities quoted.

2.8 Key assumption No. 8 (KA 8) - EMPIRICISM

In controversial areas like quality, where ideas and emotions often go hand in hand, progress can best be attained with a scientific approach - i.e. an empirical approach. This is how similar problems were partly overcome before by psychology, by foreign language teaching or by language testing. Rather than dealing with opinions, let us try to deal with facts. Finding out all the answers empirically could certainly prove too expensive initially, but a balance must be sought. Shared opinions can wait to be checked, whereas empirical research with scientific methods borrowed from experimental psychology and educational measurements should provide reliable definite answers to the controversial issues right away.

In general, we should work on evaluation according to a simple dichotomy: *for unchallenged plausible hypotheses, postpone empirical research; for disagreements, immediate experimentation should tell us who is right.*

Research will first have to be conducted on the definitions of quality levels presented in *Diagram 1*, to make them complete, reliable, and acceptable (as a reference for any dialogue involving a significant amount of evaluation experts).

Production of evaluation instruments should follow empirical confirmation and, again, it should resort to the empirical approach for refinement and validation of the starting experimental versions.

3. Preliminary concepts

In the previous pages an overall philosophy has been defined through cumulative basic ideas. Now it is time to present practical ways of catering for some of the tangible needs of machine translation evaluation, with evaluation of conventional translation also being covered inasmuch as it is inevitable in solving the more specific problems of our main concern. Both goals are compatible and related to each other, though we are focusing on MT in proposing immediate action.

For the choice between the black box and the glass box approaches, which is mentioned so often in the MT literature, we propose a combined solution based on a chronological sequence and a dialogue, as if operational MT language pairs were a student taking an exam. First, you give the student or the MT pair a performance test (a comprehensive battery on translation proficiency). If he/she or *it* fails but in the ensuing dialogue insists it was bad luck, you may very well decide to move on from the starting black box approach of performance tests to the less opaque diagnostic tests and eventually to the totally open dialogue of a glass box approach where designers and developers try to give explanations, while evaluators try to understand and improvise ad-hoc subtests before eventually taking the risk of reaching personal conclusions. These no longer have the category of scientific testing according to standards and definitely fall within the area of mixed assessments where intuition and opinion play an important role. Mixed personal assessment is less comparable and reliable than scientific testing but is also useful, particularly when faced with the dilemma of having that limited kind of evaluation or nothing at all. The situation can in fact arise with MT prototypes that have neither attained the operational threshold nor claim to have done it, whenever evaluation cannot wait because important development policies or financial investments must be decided upon at a given time. Standardized tests measuring potentials with a high predictive validity checked by an empirical correlation coefficient do not exist and it would be too costly and risky to undertake their design. Good standardized MT proficiency tests do not exist either but can be devised, written and refined fairly easily and quickly.

As far as we can see, there are three different immediate needs in MT evaluation. So we wish to propose three different formulas. It is tempting to call them *magic formulas*, because their definitions do look like short, simple solutions for complex, unsolved problems. Unfortunately, the amount and nature of work involved in their implementation are rather far from being magic, for they involve a considerable investment in man-hours by very qualified staff. For the moment we are beginning to describe what has to be done and how. We shall refer to *who* later, in the section on a *concluding proposal* consisting of three stages.

Our formula identification labels are: *X, Y, Z.*

Formula *X*

For multi-purpose comparable quality evaluation of a translation sample or corpus:

Follow this report's detailed proposal on how to relate the quality of a given text to the six benchmarks represented in the scale shown in *Diagram 1* below (more thoroughly discussed in this author's contribution to a later workshop). Devise and write the required proficiency test batteries.

The results would be particularly useful for intersystem comparisons and interpair comparisons within the same system. Development progress or potential would not be adequately covered by this formula.

Formula *Y*

For purchasing and similar decisions involving comprehensive study of technically and economically relevant features:

Follow advice given in Vasconcellos 1992 and in JEIDA 1992 (see *References,* listed alphabetically in the last section) and supplement it with the functional quality definition supplied in this paper, as well as with the administration of the corresponding ad-hoc subtests. See additional remarks in subsections 5.2 and 5.3.

Formula *Z*

For regular checks of development progress of any given pair of operational MT systems, the prerequisite is to let evaluators supplement their performance measurements with the diagnostic tests needed by developers to plan for systematic remedial and improvement work based on the detected weak areas (master plan for a system, specific plan for each pair). Goal-related progress tests could then easily check the attainment of previously stated well-defined aims.

Usefulness of such selected goals could be confirmed at far wider intervals by the administration of simplified or otherwise adapted versions of the test batteries mentioned under formulas *X* and *Y.*

Development testing must be set up in a circular way, with test design based on planning and goals at one end, and feed-back to planners at the other end. Test results should allow identification of any desirable changes (for an increased efficiency of either the procedures or the goals themselves).

Finally, let us show in a diagram the first sketch of a consistent comprehensive definition of translation quality. It tries to be universal (compatible with any specific purposes), functional (based on communicative efficiency) and easy to understand and recognize. The topic of communicative efficiency deserves being taken up again in a supplementary work. The communicative functions and their peculiarities have been adequately described in the literature of two fields that are not covered in our *References* - (a) foreign language teaching and its communicative approach, (b) communication theories related to social psychology and journalism.

Diagram 1

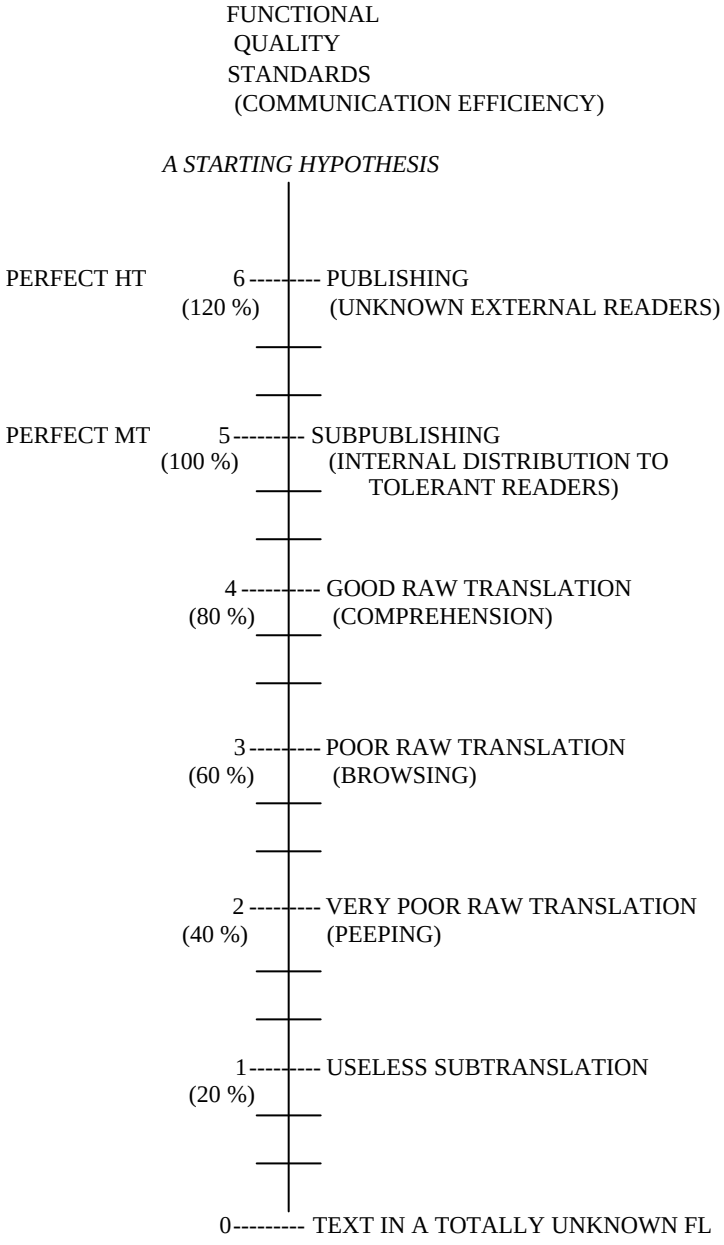
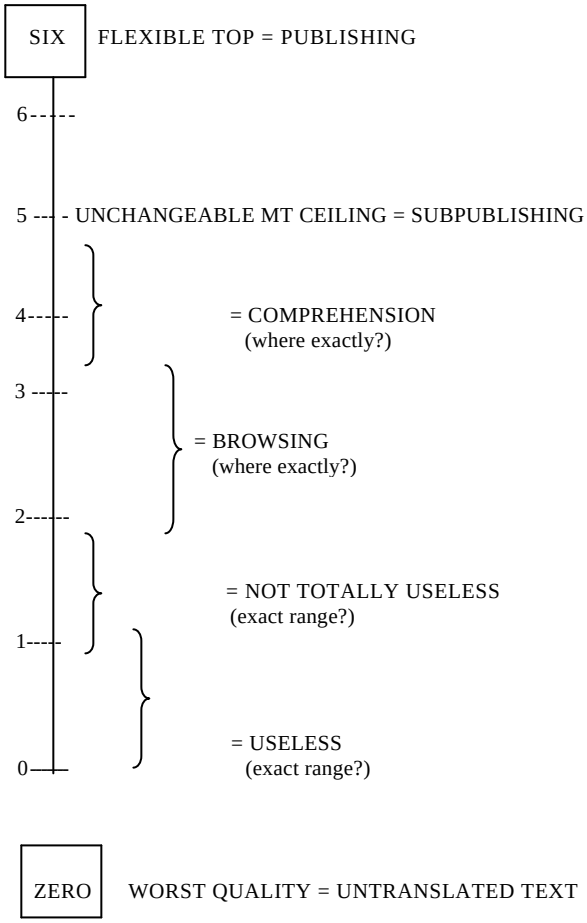

FUNCTIONAL
QUALITY
STANDARDS
(COMMUNICATION EFFICIENCY)
A STARTING HYPOTHESIS
PERFECT HT
6 PUBLISHING
(120 %)
(UNKNOWN EXTERNAL READERS)
PERFECT MT
5 SUBPUBLISHING
(100 %)
(INTERNAL DISTRIBUTION TO
TOLERANT READERS)
4 GOOD RAW TRANSLATION
(80 %)
(COMPREHENSION)
3 POOR RAW TRANSLATION
(60 %)
(BROWSING)
2 VERY POOR RAW TRANSLATION
(40 %)
(PEEPING)
1 USELESS SUBTRANSLATION
(20 %)
0 TEXT IN A TOTALLY UNKNOWN FL

Diagram 2

FUNCTIONAL
QUALITY
STANDARDS
(COMMUNICATION EFFICIENCY)

RESEARCH TO BE UNDERTAKEN



4. Technical principles and specifications

Language testing and educational statistics provide the theoretical background allowing us to list the specifications to be met by any tests or methods deemed suitable for the evaluation work to be done. Concepts and figures are explained in the next few subsections. We shall deal with reliability first, because it is simpler, and then with validity. Result interpretation and norms are covered later, and practical factors are placed at the end of the section.

Further study of testing principles and techniques can be undertaken whenever necessary by resorting to the literature selected in the last section for the area of language testing.

Some readers might wonder why this author did not write a sample test himself, rather than devoting a whole section of the paper to the description of how it should be done. The answer is, unfortunately, quite discouraging. Writing a sample form of a translation proficiency test fulfilling all the requisites that make it really good (by state-of-the-art standards) and presumably reusable (standardizable), just for a single language pair, would probably take a team of at least three co-authors a whole year. The experience of institutions like the Educational Testing Service of New Jersey (USA) or of International English Language Testing Service (IELTS) in the United Kingdom could confirm this hypothesis, based on this author's knowledge of their work and his own experience with 'home-made tests'.

It is useful to point out the difference between a standardized foreign language proficiency test (terms and know-how developed by language testing experts) and a test suite (term and concept often used by computational linguists). *Test suites* usually involve a systematic check of a particular problem or kind of problem (e.g. the morphological system of German) and can therefore be produced and administered quickly and easily. *Standardized translation proficiency tests,* which still do not exist, should follow a pattern similar to foreign language proficiency tests. This involves a lot of work, as will be seen throughout section 4. The inevitable choice that cannot be altered by any clever proposals lies between modest partial attempts (what has been done so far) or the innovative and expensive solution of implementing the less widely known but well tried principles and techniques of specialized language testing.

4.1 Reliability

Our concern is a very precise concept of *reliability,* with the restrictive sense that this term is given in testing. Having devised tests that measure whatever it is in a valid way (see 4.2 below), the next question is how reliable the results are.

Do they vary for no apparent reason (interference from extraneous factors)? Do they depend on *luck* (particular translation samples involved)? Do different people marking the translations give them significantly different scores? The next three subsections deal with each question separately.

4.1.1 Reliability based on absence of extraneous factors

What we need here are source texts that are not beyond an MT system's ceiling (as would be a legal report on comparison of national constitutions, for example) and that involve identical rough translatability levels. These will be described in our supplementary work.

This kind of reliability is a condition, rather than a figure. Fulfilment of the condition can be checked satisfactorily by the concurrence of two different actions. First, the persons choosing the source text do it bearing in mind the requisite of identical translatability within the range of acceptable difficulty. Second, the people correcting the translations point out any anomalous phenomena or problems they may encounter in marking, when problem parts can still be ignored for the score computation.

Feedback from evaluators to users, researchers and developers is desirable. *Washback*, on the contrary, is to be avoided by confidentiality and security - otherwise it would become an extraneous factor seriously affecting reliability of results. *Washback* is a common term amongst testing experts, who use it to refer to the problem of students who, rather than *learning the subject matter*, try to *learn how to pass exams* for which there are similar precedents with common patterns and features.

4.1.2 Reliability based on significant sampling

This is indeed a major problem requiring careful study and research. MT experts tend to agree that large translation corpora (not less than 10,000 words) are needed for evaluation results to be reliable. The reason for this need is obvious. In pseudotranslation, computers do not actually *solve* problems; they *identify* problems and select *ready-made solutions* from a stock filed in their huge memories. It may make sense for the Commission's Translation Service, for example, to have hired over 1,000 translators on the basis of a translation sample of about 900 words (3 pages) in each competition (for one language pair), but it would be impossible to evaluate any MT system reliably with such a minute sample. Having brought up the dramatic gap separating both extremes, understanding the substantial difference between pseudotranslation (by computers) and translation (by humans) provides the relevant arguments needed to justify the similar acceptability of such widely distant solutions.

Comprehension of the conceptual difference is essential for a preliminary perception of the nature of the problem. But the problem remains unsolved. On the one hand, we have to admit that a translation sample that is large enough to evaluate humans is not reliable at all for computer systems, and MT experts can be said to have good reasons for advocating far larger sample sizes. On the other hand, however, marking huge samples of hundreds of pages is unreasonably expensive, in time as in money. When dealing with validity (subsection 4.2) and intermarker reliability (in 4.1.3) it will be proved that marking has to be done carefully - i.e. slowly - by qualified staff. So far, no cheaper alternatives to this solution have been proved to have the same value.

Let us sum all this up by stating that we must have a high reliability but know that it is too expensive.

How high should reliability be? We recommend a coefficient of .9 or above, computed empirically with the *method of equivalent forms.* Can we attain a satisfactory coefficient with small samples? Research with easily markable subtests to be administered before using translation samples should allow a cost reduction that does not involve lowering the coefficient.

In the case of quite bad MT systems or MT pairs, just as with bad students, you can save the time and trouble of conducting the expensive marking process of a large reliable sample by a series of preliminary checks. Unless the initial results are positive, you just stop at that point, as you do with below-average candidates in human competitions. Preliminary checks could consist of two rounds. First, systematic control over separate components, such as morphology, syntax at different levels, dictionary size and appropriateness in assorted semantic areas of the general vocabulary, ad-hoc tests devised to see if expectable flaws related to design or development history are serious. After that you could still have an intermediate step before deciding if it is worthwhile to go on with large authentic text samples. In the intermediate step a specially designed small sample of non-authentic texts involving enough variety (of genre and type, form, style, contents, communicative function) could be used as a second round of the preliminary checks.

Harris 1969 reports a widespread mathematical formula to compute the sample size required to reach a particular higher reliability coefficient. It is also possible to increase reliability without resorting to a larger translation sample by finding suitable ways of computing the effect of the *guessing* factor on small samples.

4.1.3 Intermarker reliability

Some statements concerning the kind of objectivity attainable have already been made under KA 4. If we add more precise definitions and benchmark translations supplied by later research (see *Diagram 2)*, then a high intermarker reliability should turn out to be possible. Our proposal is that all the translation samples selected for a given test administration should be marked by a minimum of two people and the intermarker correlation coefficient should attain a level of at least .9 (divergence of ±5% from the arithmetic mean). This particular coefficient has been found to be both sufficient and possible in educational measurements.

Here is a comprehensive list of the steps and techniques hopefully resulting in such a high coefficient.

First - Markers are selected very carefully, avoiding the common mistake of thinking that people who are capable of doing something (translating, in this case) are by definition equally capable of evaluating the same task when it is performed by others (similar translations by colleagues or by MT systems, for example).

Markers must

(a) have a good background in translation theory and practice, abundant experience in institutional marking of essay-type tests and translation exercises, perfect understanding of descriptive linguistics (terms and concepts related to all the likely occurrences of any linguistic phenomena - morphology, syntax, semantics, style, native language interference, standard dialect, mistake typology, seriousness of mistake occurrences)

(b) *believe* in the kind of attainable objectivity demanded, *love* the challenge involved, *prove* a genuine interest, *show* a positive attitude

(c) be particularly open-minded towards learning while working and becoming team members sharing a common goal (emotional reactions forbidden).

Second - Top-class markers become thoroughly familiar with definitions and examples of each functional quality level.

Third - A member of the board of evaluators proposed elsewhere gives markers formal training and supervises exercises. Formal training includes special techniques, such as exercise classification by quality, case studies etc. Common work goes on until a satisfactory correlation is obtained in simulations.

Fourth - Markers give points to sets of unidentified translations comprising authentic exercises and premarked translations from a confidential corpus of standards.

Fifth - Marks are compared between markers and between authentic and fake exercises. Anomalies are dealt with on a special basis, following some pre-established exceptional method, provided the percentage of anomalies is not above a very low percentage of perhaps 5% of translation units marked. If the percentage were too high, it should be regarded as a clear symptom that something had gone really wrong and the reasons should be found out and analysed. The necessity of going back to the first or second step should not be totally excluded.

Benchmark translations should be of two kinds - public, for MT developers and training of markers; confidential, for actual testing and control of intermarker reliability. Translation banks with all the levels of functional quality could be made up by working on files from different sources, such as revisers' personal files, translations written for competitive exams, raw and annotated or post-edited machine translations used by system developers.

Needless to say that organization, confidentiality and security should be similar to what is customary in official competitions for translators. Otherwise human and other external factors could easily cancel the positive effect of any suitable techniques.

4.2 Validity

Again we shall refer to the restricted meaning the term is given by language testing experts. But there is, of course, the previous starting question of validity as a non-specialized term too. It involves the need to define translation quality in a valid way, which requires cutting down to a few precise notions what otherwise would remain a vague general concept.

In discussing intelligence, for example, the starting step is stating what psychologists mean by *intelligence* - the ability to devise new solutions for new problems. Having accepted a valid definition, checking the existence of any quality and quantifying its importance will imply an acceptance of valid criteria and a refusal of irrelevant features. In the case of intelligence, personality features implicitly regarded as being outside the valid definition of it will include imagination, creativity, meticulousness, wisdom, willpower, or emotional control, to cite only a few of the concepts that the layman might include in his intuitive notion of intelligence. No valid criteria measuring intelligence can consequently be built on any of these examples. Time required for problem solving *is,* on the contrary, a valid criterion, with the advantage of allowing comparison.

In the case of translation quality, let us proceed with the presentation of *our* valid definition. We have chosen *functional* quality in terms of *communicative efficiency*. It is not the only conceivable valid definition, but it serves our purpose. Pros and cons will be discussed elsewhere, and the necessary details will be given. Our choice allows listing related criteria as being valid or irrelevant.

Valid criteria
- Accuracy / fidelity
- Loss of intelligibility
- Undesirable noise

Irrelevant criteria
- Intelligibility of source
- Reading time
- Style
- Register

Style and register are only irrelevant when measuring the kind of modest quality expectable nowadays from MT systems, where comparisons are to be related to the subpublishing level depicted in *Diagrarn 1*.

The criteria selected above are to be included in any testing instruments with *face validity,* which is the kind of obvious transparent validity that can be expected to be undeniable after conscious acceptance of the definition providing its foundation. But face validity is expensive - it takes too many man-hours of highly qualified work - and difficult to measure. This is exactly where the expertise of language testing becomes inevitable, for it is these experts who have devised the two essential kinds of specialized resources: substitutes for face validity (concurrent, predictive, construct, criterion-related, etc.) and measuring techniques (reliable sampling, scales, comparisons, statistical norms, correlation, sample size required, standard deviation, standard error of measurement, etc.).

Our recommendation is to select a large volume of bilingual translation corpora representing a wide, ad-hoc range of text types and contents and then conduct a comprehensive quality study based on the valid criteria proposed. The study would later become the statistical criterion for production of evaluation instruments with an empirical criterion-related validity coefficient. We suggest aiming at .85 as the minimal acceptable correlation.

A closing remark on validity is still advisable. Validity being the common-sense requisite that a test must measure what it claims to be able to measure and not something else, a corollary is that it cannot easily measure more than one thing at a time. Avoiding mixed skills and cutting them into simple components is one of the tricks taught by experience and confirmed by empirical research.

Testing theory tells us, for example, that in a history test, essay-type questions may be inevitable to measure the students' understanding of events and processes, but this particular technique should be avoided to check knowledge of facts, which can be done much more efficiently with a large set of simple questions requiring short objective answers. The second technique would have all these advantages: elimination of an extraneous factor that would otherwise be difficult to eliminate in marking - writing skills; objective marking made easy by comparison with the only right answer to each question; marking made cheaper by the possibility of resorting to secretarial help or even computer scoring; increased reliability due to larger sampling. However, when it is decided that knowledge of facts and understanding of events are equally important, neither technique will be valid. It will become necessary to devise a test consisting of two separate subtests, whose combined results should have construct validity.

Some readers will undoubtedly feel that the preceding remarks are superfluous. Perhaps it seems they are, but the fact is that mistakes are being made all the time. Is it not a mistake for some language teachers to use composition writing as a combined test of grammar and writing skills, ignoring the fact that grammatical knowledge docs not suffice to express thought and that good essay writers will skilfully avoid grammatical forms they are not confident with? Is it not a mistake for some teachers to test foreign language listening comprehension with dictations, where valid conclusions are prevented by too thick a mixture of skills - recognizing grammatical forms and patterns, vocabulary, spelling, aural comprehension, sound discrimination, note making, short-term memory, acoustics?

With the actual performance of operational MT systems valid testing can only be designed and implemented by making several distinctions within the broad purpose of measuring and improving the functional quality of translations. Analogy with testing of human translations will give its full benefit here.

Three different kinds of test are described and recommended (proficiency, diagnostic, progress). A fourth type is rejected (development potential). See again our *formulas X, Y, Z,* recorded when dealing with *Preliminary concepts* (section 3).

For the purpose of evaluating overall translation quality offered by a given language pair at a given time *a full proficiency test battery will be needed.* It will have construct validity if all the key components are present, with the right weight. Translation of sample texts will be included, but so will separate subtests of grammar, lexicon, contrastive difficulties, computer difficulties, writing.

A complete proficiency battery will spot weak points. Further exploration by administration of *separate diagnostic tests* will thus be made possible.

Having a diagnosis, development planning with specific inventories of detailed aims becomes easier. *Progress tests* are then particularly easy to design and interpret. Expectable improvement/deterioration ratios cannot be prescribed beforehand, because of the complexity of factors involved - system's improvability, programmers' skills, point reached in *the learning curve* (near a plateau?), kind of lexical items added, language pair involved. Achievement or progress tests are in any case an indispensable regular feature of any development work deserving systematic attention. But they cannot precede planning, which in turn cannot precede diagnosis, and diagnosis must be founded on previous overall functional quality measurements.

There is one more kind of test that must wait far longer: development potential. Predictions concerning development would require large numbers of MT systems being compared, watching their development over a few years and finally refining the testing materials again several times until predictive validity could have attained a satisfactory level. Too much work for too long, with merely hopes of a very doubtful success. For assessment of development potential there is at present no sensible alternative to subjective judgement by experts, who could always produce a very qualified guess by considering progress measurements reliably recorded and all the past and future relevant factors known to them.

In stating that development potential can only be measured by qualified guessing from experts we do not mean to underestimate or even despise the value of any professional work of this kind. Our purpose here is to make a distinction between scientifically designed prediction tests (which should be infallible, by definition, but do not exist) and reasonable guessing based on facts and experience. Reasonable guessing must be undertaken because it is the only resource immediately available at a reasonable price, but you cannot call that kind of activity a prediction test. Prediction tests would require comprehensive control of all the facts and circumstances, internal or external, and empirical research spread throughout a number of years and allowing for gradual refining of evaluation tools, as written above. In this paper we choose to focus on *performance testing*, leaving outside the probability studies, no matter how good they may be. The same approach was followed in deciding not to relieve the burden of decision makers by providing them with formulae for system selections or purchases. Our modest, more realistic aim is to simply propose better ways of defining and finding out some of the information needed by them.

4.3 Norms

Testing materials should be complete with norms, like prestigious standardized tests, such as TOEFL (of Educational Testing Service - see *References)* or the publicly available ELBA of Edinburgh (Alderson, Krahnke and Stansfield 1987, p. 25), which includes a technical manual and an interpretative guide. There are important reasons for this demand:

(a) If testing materials are not supported by technical specifications recording empirical evidence, the decision to use them is reduced to a matter of personal confidence and subjective judgement.
(b) Result interpretation is too tricky and subjective to produce unbiased reports.

4.4 Practicality

For the sake of completeness, we must also refer to the practical aspects of testing. However, it would be silly to make detailed recommendations at this stage. Let us simply admit that in refusing cheap amateurish oversimplified evaluation by individuals we are actually proposing expensive professional work by several teams - one (a permanent board) for overall design, organization and supervision; other ad-hoc teams for occasional particular assignments, such as test writing, adapting or marking.

The very high costs of such expensive good evaluation make it advisable to try and find ways of reducing them without reducing the value of the results. Even though specific decisions will have to be made later, very often on a day-to-day basis, there is no harm in mentioning a few practical ideas:

(a) Original standardized tests are perhaps the ideal response to our aim, but adapting and supplementing existing ones is far cheaper and hopefully equally effective. 'Home-made' tests should not always be excluded, provided they fulfil all the requisites defined by testing experts.

(b) Evaluate only what is undoubtedly worth the price.

(c) Resort to preliminary checks on presumably weak points systematically. Negative results at that point will be a cheap 'short cut' to a negative conclusion. The experience of foreign language teaching and learning suggests the following detailed and carefully arranged sequence of preliminary checks, before the culmination with overall testing of translation writing: morphology recognition, syntax recognition, general vocabulary comprehension, specialized vocabulary comprehension, reading comprehension. Each of the preliminary checks can have a reduced size if it is filled with the right 'traps', based on language contrasts and computer idiocy.

5. Other recommendations

Before reaching the point of submitting a three-stage proposal in section 6, we still have a few questions left. In the closing proposal, the work is described in chronological order and the emphasis is placed on *what* has to be done, rather than *how*. Ways of doing it are discussed more fully here.

Our approach was defined at the outset, under *The key assumptions* and *Preliminary concepts.* Principles and specifications have just been covered too. The remaining questions are basically related to the problem of who has to produce and administer what kind of testing materials.

As for who, the answer (already advanced under *Practicality)* is: various ad-hoc temporary teams, guided and supervised by a ruling permanent board, as described in section 6. As for the materials, apart from all the requisites and remarks written so far, our final recommendations will be presented below, in a sequence around four nuclei. First, common practice of testing experts will be conveniently summed up (in 5.1). Then we shall pay attention to what three authors have to say (5.2, 5.3, 5.4). The selection will be based on the contributions by Vasconcellos, JEIDA, and Blatt - in that order.

5.1. Production of testing materials

First of all, we must remember the three formulas of the *Preliminary concepts.*

Research and experimentation should give us the necessary laws, rules and jurisprudence, to be added to all we have at present - intuition, rational approach. Until then we cannot obtain the *basic factor* for Formula *X* - i.e. a set of benchmarks that give us a very precise reference, equivalent to what the combined effect of laws, rules and jurisprudence represents in civilized societies, where they all make up a complex framework built on the foundation of accepted preexisting moral and social principles.

Once the cardinal points of quality have been defined and illustrated, with samples, for all the quality levels, a distinction between two different needs must be made. Intersystem or interpair comparison should be made possible by providing a complete Formula *X.* This means using proficiency tests, designed to measure the overall performance. This need is different from purchasing decisions or similar problems involving complex decision making, which must remain outside, at a higher subjective level where quality performance will be an important item of data, but nothing else. Scientific evaluation must not go any farther, leaving the selection and weight of all the other relevant factors in the hands where they should stay, as an unsharable part of their burden and responsibility. Coming back to the example of speed mentioned elsewhere, evaluators of cars must only inform the client about the cruising speed of the different vehicle makes and models, rather than trying to reach definite conclusions concerning the client's course of action. Performance evaluation can certainly be supplemented by advice, but this requires an intelligent flexible dialogue that exceeds the limits of scientific testing and is clearly above the more modest role accepted for Formula *X* here.

Formula *X* must provide: the standard for unambiguous valid comparisons, the benchmarks, and a measure unit for quality. Formula *Y* is intended to facilitate decisions by providing supplementary information that was not revealed before by the measuring instruments concerned with overall performance only. A third problem is the measurement of progress or diachronic contrasts, to be dealt with by Formula *Z.*

Theoretically, progress could also be measured with general proficiency tests *(X),* but the fact is that, after the final plateau of the learning curve (in humans) or of the performance improvement (in computers) has been attained, proficiency testing must be expected not to have enough discrimination power. For this reason, it is obvious that ad-hoc partial tests related to short-term specific goals are preferable. We shall give this custom-made testing its own name - Formula *Z.*

Having finished the review of the three different formulas, we can now focus on *X.* In order to produce testing materials for proficiency measurement, the kind of detailed long process usually followed by the specialized institutions will have to be completed. Here is a quick summary, for readers lacking a formal background in language testing:

(a) Comprehensive list of problem types, from the cognitive point of view, based on contrastive descriptions of the morphosyntactic and semantic systems and system components involved in a specific language pair.

(b) Selection based on peculiar features of testees (contrastive comprehension and writing skills, translation traps, computer idiocy - polysemy, syntactic ambiguity, homographs, names, numbers, neologisms, language mixture, unit cuts, punctuation, *problem words,* etc.).

(c) Overall design of test battery, with the right subtests being given the right weight. Planning, administration, marking, result interpretation.

(d) Seeing the need for having short cuts and preliminary checks, for economic reasons, in the light of the foreseeable high cost of the preceding job.

(e) Producing such short cuts and preliminary checks.

(f) Writing and testing the whole series of subsequent experimental versions of the general proficiency test battery, performing the corresponding statistical analyses and gradually improving each experimental version.

(g) Writing the final version of a cognitive test battery (one for each language pair) and its examiners' norms.

(h) Designing and conducting the equally required second part of the testing process - translation exercises.

Fortunately, adapting and supplementing commercially available testing materials can lower the cost and deadlines of (a) to (g) dramatically, as proposed above (see *Practicality).*

Rinsche's latest proposal (1993) looks like a clever oversimplification of the ideal process. She has probably managed to find and test interesting *short cuts* to scientific proficiency testing, but their value should not be exaggerated or allowed to replace more conscientious work - in principle. Besides, you cannot build *X* without having attained a formal computation, after research, of the *basic factor* involved (quality benchmarks and definitions, with samples), as she has apparently tried to do. None of these remarks, however, are intended to be derogatory. Her work deserves closer scrutiny.

5.2 Vasconcellos

In Muriel Vasconcellos' article about *Perspectives on the Assessment of Machine-translated Output* (1989a) her knowledge and experience are clearly visible everywhere. Although we have already stated as a key assumption that we cannot share her view that 'there is no point in subjecting it [machine translation] to the approaches used for evaluating human translation' - at least not in the simple way it has been worded -, no-one can possibly fail to agree with her emphasis on the relevance of 'what *use* can be made of the machine's output'. As a matter of fact, this idea has inspired the whole definition of the *basic factor* for Formula *X* in the Annex.

The notions of system *serviceability* and 'compatibility with the particular setting envisaged' are also very important, but we regard them as different from Formula *X* and are not dealing with them for the moment. We do propose to buyers, however, to use a grid where all the economic and technical factors are included, in an overview where *X* would be only one of the measures being considered before a decision. As a nearly-comprehensive check-list, adding up and organizing ideas from all the available literature, we propose this:

(a) average performance (X)
(b) dictionary size
(c) supplementary ad-hoc tests *(Y)*
(d) serviceability
(e) fitting into existing setting
(f) hardware features and prices
(g) software features and prices
(h) cost of development
(i) cost of maintenance
(j) suitability of raw output
(k) suitability of quickly edited output
(1) cost of quick human translation
(m) compatibility with pre-editing and post-editing aids

Rather than making additional comments on these and other factors we prefer to suggest direct reading of the selected article as being both worthwhile and sufficient.

5.3 JEIDA

The Japan Electronic Industry Development Association presented a paper in San Diego's MT Evaluation Workshop of 1992, under the title *JEIDA Methodology and Criteria on Machine Translation Evaluation.* Economic factors beyond our scope receive full attention and are freely mixed with linguistic facts and satisfaction feelings. In our opinion, the value of the paper lies with its comprehensiveness from the prospective buyers' point of view and has comparatively little use for the more restricted approach of the present study. This is actually the main reason why it must be recommended here, as a way of covering more ground. IT matters are particularly well covered.

Linguistically speaking, JEIDA's methodology is compatible with our approach and has actually resorted to a proficiency test (TOEFL) devised for human learners. Somehow, it covers almost all the points of our previous check-list (under 5.2), omitting only (1) and (m) and providing a detailed break-down into components for many others.

5.4 Blatt

Achim Blatt has recently produced (December, 1992) a note for an ad-hoc *Evaluation of the LOGOS System* and several short reports (see *References).*

Their common sense and easy practicality are commendable.

6. Concluding with a three-stage proposal

So far we have been piling up a large amount of recommendations, stating pillars and principles, quoting and supporting selected similar or compatible ideas, taking sides in controversial issues. Now we have reached the point where action has to be defined. Specific proposals in chronological order of implementation will try to show the way.

6.1 First stage

After obtaining consensus and interim approval of this or a similar paper, the first stage would be setting up *a permanent evaluation board* of five to seven members and commissioning them *to produce a complete final version of a reference work* based on the present personal attempt or something equivalent. Performing this task together would favour development of a joint personality and ensure genuine common thinking.

Setting up a permanent board is undoubtedly the only sensible way of letting this discussion paper become useful. If the inevitable practical difficulties discouraged decision makers to the point of not following this essential recommendation in spite of having found it desirable, the result might be that the entire paper would have been nearly a waste of time. Let us then try to prove that setting up a board is absolutely necessary. Here are some arguments:

(a) This author's assignment was to propose a methodology. The aim has been fulfilled with the present paper, but only partially. Even after obtaining consensus in the ensuing discussions, the aim will only be partially attained. Why? Because any evaluation methodology can only be equivalent to *a law*, and the law has to be supplemented and interpreted by *legal experts.*

(b) Consensus after general discussion will at least enhance the value of our methodology proposals; and it may also change, replace, delete or add a few things. The process will result in having *a better law* and some *regulations*, but the need for further regulations and interpretation will remain. It will always remain, as it does with legal systems, which are not efficient without *judges* and *courts.*

(c) This paper is perhaps a good personal attempt to put together all the relevant useful ideas known to its author, from any of the selected sources or from his own imagination. Now it has to be *backed by a discussion*, which will produce a final set of acceptable ideas. And then all the following tasks will continue to be pending: rewriting this paper or producing a different one, supplementing it later with ad-hoc secondary rules and minor decisions in view of any of the subsequent implementation problems, performing day-to-day work independently (qualified experienced work; no bias due to developers, users, or personal opinion). Can anyone think of a resource with advantages comparable to those of a permanent board - personal opinions added up and balanced, stability, cumulative knowledge, shared responsibility, minimization of incidents linked to individuals?

(d) Even if the present methodology were totally rejected and replaced, *no good methodology can exist that is simple enough for an individual to implement.* The problem lies with the complexity of the subject. When we compared our problem to that of measuring speed or temperature we were right in claiming the need for standards of distance, time, heat, and cold. With speed or temperature you can then use reliable machines and the whole problem is solved; with verbal behaviour the measuring instrument can only be a trained person, which makes us suggest that we should resort to a board. Only fairly permanent groups or institutions, rather than individuals, can succeed in gaining the credibility that a prestigious name implying joint responsibility of top-class experts can build up after some years.

Having explained why we favour setting up a board, thinking about its desirable composition will tell us why it cannot have three or four members only. It must

consist of enough experts to represent the broad spectrum of knowledge explained in KA 5 *(interdisciplinarity).* Each board member should know as much as possible about the fields related to his own, but should above all be *a good expert* in one or two of the areas mentioned in the corresponding *key assumption,* where MT and computational linguistics were listed as representing only one of the six areas involved.

The previous list in KA 5 was:

- general and applied (contrastive) linguistics
- translation theory and practice
- translation teaching and marking (with learners and translators)
- educational statistics and foreign language testing
- computational linguistics and MT
- functional communication and translation use

Rather than defining exactly the minimum knowledge and experience required from appointable experts, which must be left to people with higher responsibility, we shall simply point out common mistakes to be avoided:

(a) No-one can be expected to be a *good* expert, at the level required for this particular task, in more than one or two of the relevant areas.

(b) Knowing a foreign language does not mean that you are a translator.

(c) Being a translator does not mean that you are a linguist, even if some institutions - like the CEC - decide to call you 'a linguist', very generously, for internal organizational reasons.

(d) Being a teacher or a jury member does not imply that you are an expert in testing, even if your work necessarily involves some testing.

(e) Only a few translators are also experts in machine translation, but the opposite is also true. Psychologically speaking, MT experts and translation experts often underestimate and misunderstand one another. The fact and the reasons are well-known. Can we afford to go on ignoring it?

Appointing board members also inevitably implies having to combine languages - English, French and German is the bare minimum for Europe -, nationality and personality.

The first task of the permanent board should be rewriting some form of the present *'basic law' in a final version.* Then the same individuals could go on with the initial implementation and interpretation, with help form additional people,

appointed by themselves, whose work would remain subject to the board's approval. The starting reference work ought to be fairly wide and complete, possibly along the lines of this proposal, and comprise *a well-developed annex, with experimental benchmark translations representing the main quality levels* of two or three language pairs.

The end of the first stage would be reached with formal definition of the basic factor for formula X.

6.2 Second stage

Applied work could only start after completion of the final step that has just been identified as the culmination of 6.1. Afterwards the board could arrange the *design and development of the translation proficiency testing instruments needed for intersystem and interpair comparisons.* Confidentiality would become one of the board's rights, in order to make reusability of testing materials possible, considering their value. Confidentiality would also minimize the washback effect among system producers and developers and would render resource to *short cuts* effective. *Short cuts* would often prove essential as substitutes for comprehensive testing. They are, on the other hand, the best known way of making evaluation of comparatively small corpora similarly reliable and valid to evaluation of the otherwise recommendable huge corpora, though at a low discriminatory level (widely sufficient with immature systems or language pairs).

Confidentiality is obviously compatible with feedback. Testers should give complete feedback to concerned producers and developers, in the form of results and result interpretation, not in the form of transparent testing enabling interested parties to disguise symptoms with faked 'self-vaccination'.

The end of the second stage would be reached with formal definition of the entire formula X.

6.3 Third stage

Work on devising formulas *Y* and *Z* can best be undertaken after satisfactory completion of the second stage.

For *Z*, the task of the board would consist in producing diagnostic instruments on the basis of the previous experience and in full agreement with development planners. Developers could then be assisted in checking their inevitably slow step-by-step progress. Mutual co-operation of testers and developers would involve information, planning, progress checks and remedial action, in a recurrent process consisting of all four parts each time.

Formula *Y*, the most difficult, important and ambitious of the three, could be handled last. The work of the board might have eventually gained enough know-how and respect to be accepted by both DG XIII (Language Industries) and the Translation Service as a common advisory board assisting each of them in solving many of the practical problems related to quality, in MT at the beginning and in any translation procedure later. Assessment on only quality related MT problems covers a wide range of questions - when can a language pair be said to be *mature* and be consequently transferred to an operator lacking a development team?; what kind of maintenance work would still be necessary?; to what extent can a mature pair be cost-efficient in a hypothetical new translation unit specializing in *rapid service?;* do undeniable time savings due to translation aids in a conventional human process result in a significant loss of quality, and how tolerable is it? On their part, quality related conventional translation problems could welcome some guidance in: selecting translators and revisers, minimizing marking subjectivity, defining ad-hoc translation standards for any special purposes, identifying cases of unaffordable, unjustified translators' perfectionism.

7. References

ALDERSON, C. / KRAHNKE, K.J. / STANSFIELD, C.W., 1987: *Reviews of English Language Proficiency Tests.* TESOL, Washington D.C.

ALDERSON, Charles /NORTH, Brian, 1991: *Language Testing in the 1990s.* Modern English Publications in association with the British Council, MacMillan, London.

ARRARTE, Gerardo, et al., 1992: *A Management Tool for Test Corpora.lBM* Scientific Center, Madrid.

BLATT, Achim, 1992: *Evaluation of the Logos system.* Internal doc. for DG XIII and Translation Service of CEC.

BLATT, Achim, 1993a: *Visit to Logos Germany on April 5-6.* Internal report for DG XIII and Translation Service of CEC.

BLATT, Achim, 1993b: *Visit to Metal on May 5-7.* Internal report for DG XIII and Translation Service of CEC.

EUROS, Oscar Krisen (ed.), 1938 to 1990: *Mental Measurements Yearbooks.* The Gryphon Press, Highland Park, New Jersey (USA).
The various editions of the *Yearbook* constitute a very extensive critical bibliography of standardized tests produced in the English-speaking world. The volumes list thousands of tests, the tests of more general interest being covered by two or more reviews. In addition, hundreds of books on measurement subjects are listed, many entries including excerpts from the reviews which appeared in professional journals.

BYTE magazine, January 1993: *State of the Art - Machine Translation.* A collection of six articles by Vasconcellos, Hovy, Scott, Miller.
An excellent summary of what MT is and where it stands right now.

CARROLL, B.J., 1978: *An English Language Testing Service: Specifications.* British Council, London.

CARROLL, B.J., HALL, P.J., 1985: *Make your own language tests: A practical guide to writing language performance tests.* Pergamon, Oxford.

DATTA, Jean, 1992: T*ranslation: Mission (nearly) Impossible.*
Article in No. 4/5 of 1992 of *Language International*, journal pub. by John Benjamins, Amsterdam.

DURAND, Jacques, 1991: On *evaluating MT systems.*
Paper presented at 'Translating and the Computer 13 - A marriage of convenience?', a conference held at CBI Conference Centre of London in Nov., 1991.
A very good, sensible article. It gives a good overview for an introduction and does not contradict the present proposal in any way, being totally compatible with it. As a matter of fact, the article can be said to be a complete summary of the problems involved in MT evaluation, with very good examples, while our paper tries to provide solutions to all the difficulties listed and described by Durand. Some of the concepts and terms quoted by him give the impression of being new and different from those included in this paper, but careful reading will reveal that we are both talking about the same things, although we occasionally give them different names.

EDUCATIONAL TESTING SERVICE (ETS)
An institution that provides testing services and regularly produces some of the best testing materials in the world, such as the TOEFL test, for foreign students wishing to attend courses in American universities. It is located in Princeton, New Jersey (USA).

FLICKINGER, D. et al., 1987: *Towards Evaluation of NLP Systems.* Forum of the Association for Computational Linguistics - Evaluating Natural Language Processing Systems.
Hewlett-Packard, Palo Alto, California.

GARCIA YEBRA, Valentin, 1984: *Teoria y Practica de la Traduccion.* Editorial Credos, Madrid.
An extremely complete, excellent handbook, consisting of two volumes (873 pages). Full of examples from several European languages (classical and modern). The result of a very rich, long personal experience.

GUILFORD, J.P., 1956 and later editions: *Fundamental Statistics in Psychology and Education.* McGraw-Hill Book Co., New York.

HABERMAN, F.W.A., 1986: *Provision and Use of Raw Machine Translation.*
Paper presented at World Systran Conference (see separate reference).
Together with Vasconcellos 1989a and JEIDA 1992, it makes up a set of three articles providing sufficient inspiration and ideas for proficiency quantitative evaluation. Sections 3, 4 and 5 are essential reading for our concern. Similar statistics should be obtained for the benchmark translations previously accepted by consensus, although input errors are to be disregarded (an extraneous factor damaging to reliability) and intelligibility should not be isolated (taken for granted, rare cases to be mixed with other errors). Figure 2 provides reference of post-editing time that should interest anyone experimenting with economic factors of the translation process.

HRALA, Milan (ed.), 1990: *Miscellany on Translation Criticism - FIT.*
John Benjamins, Amsterdam.

HUTCHINS and SOMMERS, 1992: *An Introduction to Machine Translation.* Academic Press, London.

Rather than an *Introduction,* this book is a very complete and reliable 'encyclopedia'.

It is up to date and looks at things in perspective.

INTERNATIONAL ENGLISH LANGUAGE TESTING SYSTEM (IELTS)

IELTS is a test battery of English language proficiency, accepted for undergraduate and postgraduate entry by British universities and polytechnics. It is jointly managed by the British Council, the University of Cambridge Local Examination Syndicate (UCLES) and the International Development Program of Australian Universities and Colleges (IDP).

JEIDA, 1992: *Methodology and Criteria on Machine Translation Evaluation.* Japan Electronic Industry Development Association (JEIDA), Tokyo.

KNOWLES, Frank, 1979: *Error analysis of Systran output - a suggested criterion for the 'internal' evaluation of translation quality and a possible corrective for system design.*

Translating and the Computer, ed. Barbara M. Snell, pp. 109-134. North-Holland Publishing Company.

LADO, Robert, 1961 and 1964: *Language Testing: The Construction and Use of Foreign Language Tests.* London: Longmans, Green & Co., 1961. New York: McGraw-Hill Book Company, 1964.

Primarily intended for teachers of foreign languages and of English as a second language. Beginning with a discussion of language and language learning, it proceeds to a consideration of how the various language skills may be tested. Techniques for measuring cross-cultural understanding are also proposed. Several chapters at the end deal with fundamental test statistics.

LANGUAGE TESTING

A journal pub. by Edward Arnold, Kent (UK).

LANGUAGE TESTING UPDATE

A journal pub. by Centre for Research in Language Education of Lancaster University, Lancaster (UK).

NEWMARK, Peter, 1988: *A Textbook of Translation.* Prentice Hall, London.

OAKLEY, Brian et al., 1991: *Evaluation of the Commission's Multilingual Action Plan 1976-1991.*

Report commissioned by the CEC (DG XIII), Luxembourg.

PIGOTT, Ian, 1991: *Systran evaluations since 1976.*

Internal report for the Oakley panel (DG XIII of CEC).

PIGOTT, Ian, 1992: *Systran Development at the EC Commission - 1976 to 1992 – A personal account.*

Commission of the European Communities (DG XIII), Luxembourg.

RINSCHE, Adriane, 1991/9: *Towards a System of Benchmarking MT Systems.*

Report for EC Commission. October 1991.

RINSCHE, Adriane, 1993: *The Rinsche MT Evaluation Methodology for the Legal and Economic Applications Domains.*

Second part (Project Description) of a proposal submitted to DG XIII of CEC.

SCHARER, Rolf and NORTH, Brian, 1992: *Towards a Common European Framework for Reporting Language Competency.*

'Occasional paper' published by The National Foreign Language Centre, Washington, D.C.

TELEMATICS MANAGEMENT COMMITTEE (LRE team of DG XIII), 1992: *Assessment and Evaluation of NLP Systems.*
Background document (Ref. TMC/LRE/10/92, 2nd call) produced by the CEC (DG XIII), Luxembourg.

THORNDIKE, Robert L., and HAGEN, Elizabeth, 1961: *Measurement and Evaluation in Psychology and Education.* 2d ed. John Wiley & Sons, Inc. New York.
A sound general survey of the field of testing with informative treatments of such topics as the preparation of objective and essay examinations, the improvement of ratings, the technical aspects of marking and grading, and the use of norms. A wide variety of standardized tests are described briefly. One chapter is devoted to statistical techniques of fundamental use to the *novice.*

TORRENS, Antoni G., 1991: *Primer Plan de Desarrollo Systran del Subequipo Español.*
Internal Systran document including a discussion of crude functional levels and the need to attain and check predefined significant improvement.

TORRENS, Antoni G., 1992: *Six periodical reports on development progress recorded for the Systran French-Spanish pair.*
Informal notes for Systran project leader and subteam members working under the author's co-ordination.

VALETTE, Rebecca M., 1967: *Modern Language Testing: A Handbook.* Harcourt, Brace & World, New York.
Designed for the foreign language teacher, most of the illustrative items being in Spanish, French and German. Beginning with a treatment of basic principles and procedures of a language testing (including a chapter on simple statistical techniques), the manual continues with chapters on the testing of the various language skills, and of culture and literature. An appendix describes the most widely used commercial foreign language tests.

VAN SLYPE, Georges, 1979d: *Critical Study of Methods for Evaluating the Quality of Machine Translation.*
Report commissioned by CEC (DG XIII), Luxembourg.

VAN SLYPE, Georges, 1979: *Première évaluation du système de traduction automatique SYSTRAN français-anglais de la Commission des Communautés européennes.*
Report commissioned by the CEC (DG XIII), Luxembourg.

VASCONCELLOS, Muriel, 1989: *Perspectives on the Assesssment of Machine-Translated Output.*
Included in Hrala's *Miscellany* quoted above.

WORLD SYSTRAN CONFERENCE, 1986: *Special issue (No. 1/86) of Terminologie et Traduction.*
Journal published by the Translation Service of the CEC.

Antonio TORRENS
Commission of the European Communities
Translation Service
Office JECL 6/120A
200, rue de la Loi
B-1049 Bruxelles