Bureau Marcel van Dijk

INGENIEURS - CONSEILS EN METHODES DE DIRECTION

30 November 1979

1050 BRUXELLES,
AVENUE LOUISE 409 B
TEL. (02) 648 66 97

CRITICAL STUDY OF METHODS FOR EVALUATING THE QUALITY OF MACHINE TRANSLATION

FINAL REPORT

BR 19142 By Georges Van Slype

CONTENTS

1. Summary

1.1 Broad lines Of the study.

The Commission of the European Communities has set up a programme aimed at lowering the barriers between the languages of the Community.

Within the scope of this programme, major resources are being utilized for :

-the acquisition and implementation of a first-generation MT system giving rough output : SYSTRAN

-the design of a European second-generation machine translation system : EUROTRA.

To manage these resources under the best conditions, the Commission has to be able to evaluate on an on-going basis the quality of these translation systems, in particular in the light of the successive improvements on the linguistic and the data-processing sides.

After arranging, on 28 February 1978, an international seminar on the problems of evaluation of translation, it asked the Bureau Marcel van Dijk to carry out a critical review of the methods of evaluating machine translation; this review to be based on the presentations made at the seminar and on the studies on evaluation of translation already published.

1.2 Aims of the study.

The present critical study meets two requirements

- to establish the current state of the methodology of evaluation of machine translation

- to make to the Commission a series of recommendations concerning:

  - the methodology to be used to evaluate its translation systems
  - research intended to improve in the long term the efficiency of these evaluations.

1.3 Summary of the study.

1.31 State of the art.

The question of the evaluation of machine translation (MT) and human translation (HT) comprises seven facets:

1.311 Definition of the aims of evaluation.

Evaluation of translation may have two distinct groups of aims:

- Macroevaluation (total evaluation)

  - acceptance of a translation system
  - comparison of the quality of two translation systems or two versions of the same system
  - assessment of the usability of a translation system.

- Microevaluation (detailed evaluation)

  - assessment of the improvability of a translation system
  - establishment of an improvement strategy.

1.312 Definition of translation quality.

Translation quality is not an absolute concept, and has to be assessed

- relatively, applying several distinct criteria illuminating each special aspect of the quality of the translation
- allowing for the specific nature of MT, which is a product quite different from HT and for which a quite different market may open up.

1.313 Text typology.

In the short run, a simple and empirical typology Should make it possible to associate a particular method of translation with each category of texts.

In the medium term, research into the typology of texts might well lead to a more effective classification, which might even be automatic.

1.314 Effectiveness and efficiency of evaluation.

The criteria for evaluating the translation have to be chosen according to:

- their effectiveness in measuring effectively the various facets of translation quality
- their efficiency, or in other words the ratio between their effectiveness and the cost of implementing them.

1.315 Macro evaluation – criteria and methods.

The large number of criteria proposed or applied by the authors quoted can be classified into four groups:

- cognitive level : intelligibility, fidelity, coherence, usability, acceptability

- economic level : reading time, correction time, translation time

- linguistic level : reconstruction of semantic relationships, syntactic and semantic coherence, "absolute" quality, lexical evaluation, syntactic evaluation, analysis of errors

- operational level : automatic language identification, verification of the claims of the manufacturer.

Critical analysis of these criteria leads to the conclusion that those underlined in the list above have the most favourable cost effectiveness ratio.

Among the methods quoted of measuring the first two criteria (intelligibility and fidelity) on the cognitive level (rating on an intelligibility scale, Close test, multiple-choice questionnaire, knowledge test, noise test), the first (rating) appears the most effective from the point of view of an evaluation by or for the Commission.

Acceptability can be effectively measured only by a survey of final users.

The reading and correction times can easily be obtained as functions respectively of the evaluation of intelligibility and the correctness of the texts.

1.316 microeva1uation - criteria and methods.

The methods proposed or applied can be classified into five groups which we have underlined the most effective:

-grammatical symptomatic level : analysis of the grammatical errors detected in the translated texts

-formal symptomatic level : tally of the deletions, additions, modifications, shifts and replacements of words by the revisers and post-editors (i.e. revision and post-edition rates)

-diagnostic level : analysis of the causes of errors input, analysis of the source language, dictionary, etc.

-forecast level : analysis of the improvability of the system

-therapeutic level : detection of the improvements to the system following an upgrading operation.

1.317 Sampling.

The samples of text (5 to 10,000 words) and evaluators must be constituted in such a way as to give both a valid and a cost-effective operation.

The use of texts especially prepared, and identical from one evaluation to the other is an attractive idea, but unfortunately one which must probably be excluded because it would be too easy to adapt a translation system to give excellent results on the standard sample, without any guarantee as to quality for translation of any other texts.

1.32 Recommendations.

Our recommendations, which are intended to apply to the evaluation of MT by or for the Commission of the European Communities, comprise on the one hand the choice of a methodology of evaluation and, on the other, an applied research programme.

1.321 Evaluation methodology.

The evaluation methodology comprises three types of evaluation :

- the first, “superficial evaluation", will be applied each time a new version of an MT system (whether with new linguistic or new data-processing features) has to be approved on delivery. It makes use of criteria characterized by a high degree of effectiveness, low cost and universality of application (to all MT systems and even to HT) : intelligibility, fidelity, reading time, correction time, correction rate

- the second, "in-depth evaluation", will be utilised only at "turning points" in the development of an MT system (decision on experimental or operational implementation, decision on an important improvement contract),and in addition to the criteria already used for the superficial evaluation, it makes use of criteria characterized by a very high effectiveness, but also a high cost and a certain specificity of application (detail of the methodology specific to each MT system) : (acceptability to users and improvability of the system), and a further criterion which is less effective but also less expensive (actual improvement in the system following dictionary updating). The cost of the evaluation, however, has to remain within reasonable limits, and for this reason, it is essential to distinguish clearly between

  - evaluation of acceptability and market research
  - evaluation of improvability and development of the system

the third, "pinpoint evaluation", will be undertaken whenever there is a need to access the impact of certain specific changes to the system. The selection of the evaluation criteria will be a function of the changes concerned, and will thus be specific to each case.

1.322 Applied research programme.

For the applied research programme, we propose:

- on the one hand, a study of the text typology leading to a classification, preferably automatic, of the texts according to the translation process suited to them with or without pre or post-editing, interactive mode, use of specialized dictionaries, etc. (This research should not be started until the various possible methods of translation which EUROTRA will offer have been studied and defined)

- on the other hand, a study of the methodology for evaluating the improvability of an MT system, which should lead to the definition of a strategy for improving the system, making it possible to choose the improvements leading to the best results (under the heading of intelligibility, fidelity and correction rate) at the lowest cost.

This methodology will without doubt vary from tem to another.

1.33 MT evaluation-coverage.

The diagram below shows plainly the relationships between the concepts of:

- evaluation, market research and system development

- macro and microevaluation

- superficial and in depth evaluations.

1.4 Rapid scan of the study.

The reader in a hurry will be able to obtain a summary picture of the results of this study by simply reading:

- § 2 (Introduction), in its entirety

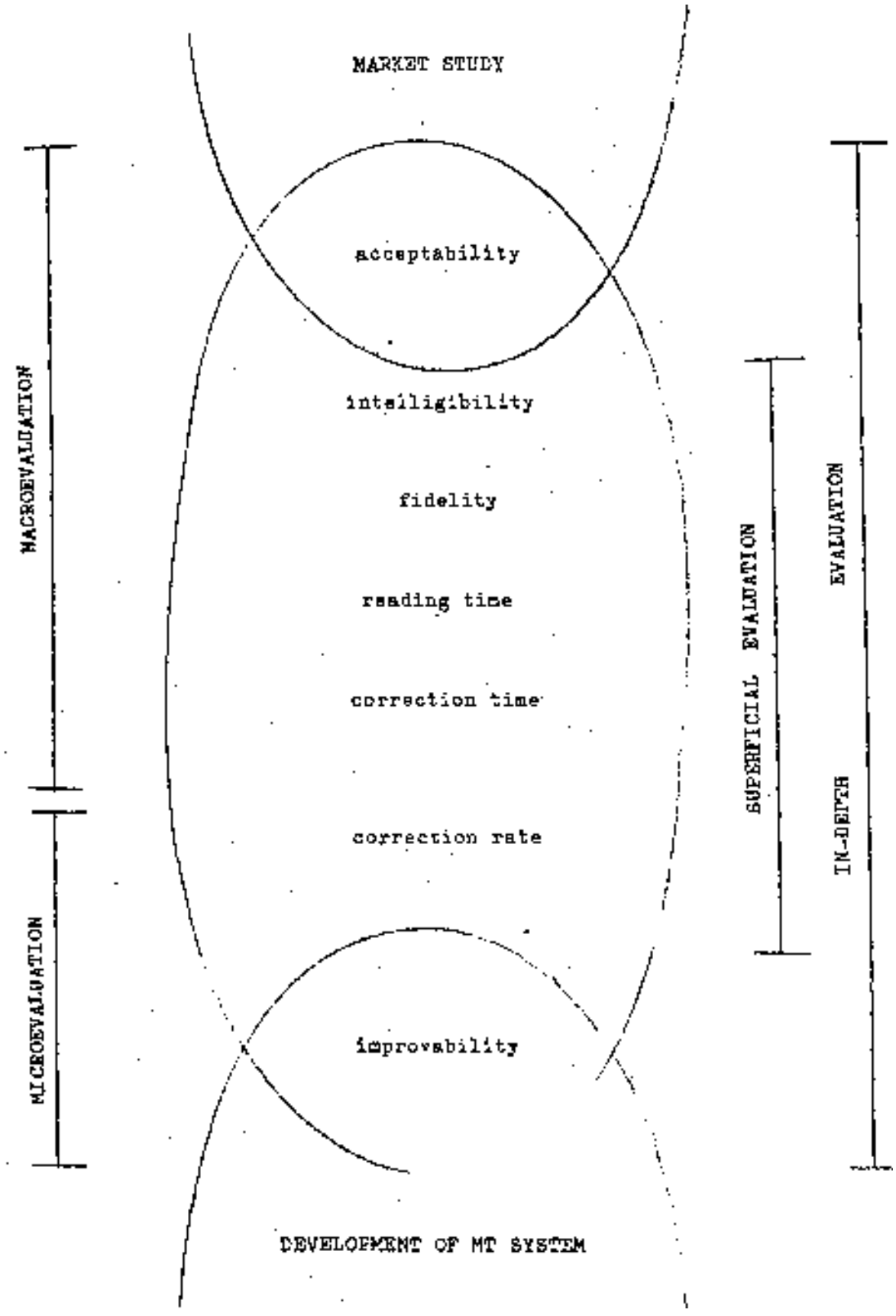
- § 3 (Assessment), only sections:

.3. x 1 (i.e. 3.11, 3.21, ... 3.71 : introduction to each of the seven elements of MT evaluation considered

- 3. x 2 (i.e. 3.12, 3.22, ... 3.72) : summary, tables of the contributions of the various authors quoted (the detailed analysis of which is given in 3. x 3)

- 3. x 4 (i.e. 3.14, 3.24, ... 3.74) : assessment of the contributions of the authors quoted

- § 4 (Summary, conclusions and recommendations), in its entirety.


MARKET STUDY
acceptability
intelligibility
fidelity
reading time
correction time
correction rate
improvability
DEVELOPMENT OF MT SYSTEM
MACROEVALUATION
MICROEVALUATION
SUPERFICIAL EVALUATION
IN-DEPTH EVALUATION

2. Introduction.

2.1 Outline.

The Commission of the European Communities has undertaken a series of long-term actions as regards machine translation, in particular:

- acquisition of a translation system already operational in the United States : the SYSTRAN system. The English-French, French-English and English-Italian versions were bought and the acquisition of other versions is envisaged
- construction of dictionaries comprising several tens of thousands of terms for each of these versions
- a Community programme aimed at developing a European translation system, EUROTRA.

A systematic evaluation of the quality and improvability of these systems has to be carried out:

- on the one hand to enable the decision-takers and managers

  - to carry out technical acceptance tests of the various successive versions delivered by the manufacturers of MT systems
  - to decide on the desirability of asking the manufacturers or other contractors to make improvement's to these versions and/or the total system

- on the other hand, to obtain data useful

  - to the implementation of MT pilot operations within the Commission or other organizations
  - to the development of marketing studies on MT.

A certain number of MT systems have been used in the United States and in Europe for more than ten years, and have been evaluated many times. The majority of methods of evaluation were covered in a short (22 pages) study by the Battelle Institute (T.C. HALLIDAY), at the request of the USAF, which is one of the principal users of MT in the United States.

The Commission of the European Communities, concerned to use the most adequate evaluation methods, sought to gather the maximum information available

- on the one hand by arranging in Luxembourg, on 28 February 1078 a "Seminar on evaluation -problems in machine translation" attended by 35 experts from Germany, Belgium, Canada, Denmark, France, the United Kingdom, Italy, Luxembourg, the Netherlands, Switzerland and the United States. The majority of these experts made a presentation; these being listed in the bibliography of this report (§ 5)
- on the other hand by requesting this critical review.

2.2 Aims.

The aims of this study are threefold:

- to present an outline of the methods of evaluation of MT practised or proposed in the world
- to provide a critical appraisal of these methods
- to recommend to the Commission the adoption of a methodology of evaluation of MT suitable to its specific requirements.

The examination of the evaluation methods was intended to stop at the problem of the appreciation of the quality of MT and not to cover economic evaluation. The literature in this field is very limited, and the Commission in any case already has a methodology for calculating the costs of MT.

2.3 Methodology of the study.

The study comprised three phases, corresponding to each of the aims above:

- collection of the existing literature(cf. bibliography § 5), analysis of the contributions of the various authors, establishment of the contributions of the various authors, establishment of a list of problems, classification of the contributions in terms of these problems, and presentation of the contributions in the form either of extracts from their publications, or of summaries, without comment or value judgements

- critical analysis of these contributions. This criticism was done as impartially as possible; but it must be stressed that it is not neutral, having been undertaken:

  - within the specific context of the requirements of the Commission as regards evaluation of MT. The role of the Commission departments which are interested in MT, DG IX-D (Translation, documentation, reproduction and library), and DG XIII (Scientific and technical information and information management) is to optimize the management of the translation services and to facilitate the transfer of information between Community languages.

  - It is not their function to support linguistic or data-processing research. This, of course, affects the strategic and tactical choices available when it becomes necessary to decide which method of MT evaluation to apply

  - taking into account the experience gained by the author of this report during work on evaluation of MT carried out for the Commission

- study and drafting of recommendations as regards the methodology of MT evaluation to be used by the Commission to assess the quality both of SYSTRAN (currently) and of EUROTRA (later).

2.4 Structure of the report.

The presentation of the experiments and the suggestions of the various authors and the critical analysis of these inputs are grouped in a single chapter (§ 3 : Assessment so that the reader can examine the criticism while the text to which it refers is still familiar.

However, to avoid any confusion, the original contribution of the authors and the criticism thereof are presented in two distinct parts of each of the paragraphs of our assessment.

Each of these paragraphs covers a specific point of the question of the evaluation of MT

- aims of the evaluation (§ 3.1) : a certain number of authors stressed as we ourselves have above, that an evaluation is not a gratuitous operation; it is carried out for a purpose which varies from case to case, using a methodology specifically suitable to each case

- translation quality (§ 3.2)in certain fields of technology, the quality of a product or a method has a precise and unambiguous definition, one agreed to by all concerned (the phrase moreover is then "quality control" and not "quality Evaluation").

As regards translation, whether by the machine or not, things are different. However, to evaluate something, it is at the very least necessary to know that which is to be evaluated; many authors discerned this problem and it is essential to underline their contribution in this field

- text typology (§ 3.3) : in translators' experience, the texts which are presented to them offer very different levels of translation difficulty. Similarly, the quality of MT varies considerably with the types of text submitted for machine translation. It would thus be important to have a typology indicating which texts are machine-translatable or, in the case of EUIROTRA, are candidates for specific MT procedures (pre-editing, interactive mode, post-editing).

Several authors tackled this question, and their contributions are grouped in this paragraph

- effectiveness and efficiency of evaluation (§3.4): when applying a method such as one for MT evaluation, it is useful to consider the effectiveness (the measure to which the method meets the assigned aim) and the efficiency (effectiveness at lowest cost) of the method.

The decision between the various methods available can be made on the basis of these two factors

- macroevaluation -methods and criteria (§ 3.5) : this is the most important part of this review, The part which lists all the criteria and all the methods used or proposed to date to assess the "static" quality of an MT system, i.e. its quality at the moment of evaluation, and regardless of the manner by which this quality has been reached

- microevaluation -methods and criteria (§ 3.6): here we review the methods, unfortunately sti11 too rare, used or proposed to assess the "dynamic" aspect of an MT system i.e. its scope for improvement and the limits thereto

- sampling (§ 3.7): this paragraph covers the methods used to sample the texts for MT and the evaluators who assess the results

Each of these paragraphs in the assessment comprises:

- an introduction clarifying the content of the paragraph and presenting the structure adopted to classify the authors' contributions according to the nature or orientation thereof
- the list of authors quoted
- the extracts or summaries of these authors' contributions, classified according to the structure of the paragraph and alphabetically inside the groups thus delimited *
- our assessment.

The assessment (§ 3) is followed by the conclusions (§ 4) to be drawn from this wide-ranging study of MT evaluation methods, including the methodology we recommend to the Commission for its evaluation work.

The bibliography is in § 5.

* Each author, of course, may have covered several aspects of the of MT evaluation, and extracts or summaries from each author may therefore be found in several different paragraphs of the assessment and in several groups within the same paragraph.

3. Assessment.

3.1 Aims of evaluation.

3.11 introduction.

The writings of ten authors, who had considered the question of the goal to be aimed at when envisaging carrying out an evaluation of translation, were analysed and the significant passages of their text extracted for publication here.

Two groups can be distinguished:

- the authors who sought to count all the possible aims which an evaluation of translation may have:

  - with respect to the recipients of the evaluation
  - with respect to the users of the translation

- the authors who concentrated on a specific aim.

3.12 Table of intended aims.

Aims	Authors
* Overall aims.	
-with respect to the recipients of the evaluation	ANDREEWSKY VAN SLYPE KU-11LEN
-with respect to the users of the evaluation	WEISSENBORN
* Specific aims.	
-effectiveness and usefulness	KNOWLES LENDERS PETIT
-system potential for improvement and establishment of improvement priorities	GREEN PANKOWICZ VEILLON

3.13 Description of intended aims.

3.131 General aims with respect to the recipients of the evaluation.

3.131.1 A. ANDREEWSKY considers that the evaluation of a translation system can vary widely according to the standpoint from which it is viewed:

-that of the user: evaluation regardless of uprating of the system

-that of the manufacturer: evaluation to improve the system.

It is necessary, moreover, to consider the question of defining the moment from which a system may be regarded as having left the initial development stage.

Allowance must also be made, finally, for the point Of view of the post-editor

- acceptability of the task

- quality threshold below which MT is refused.

3.131.2 G. VAN SLYPE thinks that the aims of an evaluation depends on the types of persons concerned and on their motivation.

It is therefore essential, before setting out the aims of the evaluation, to know for whom it is being done and what each of the recipients expects of it.

From this analysis can be deduced the evaluation criteria to be used:

Groups involved	Aims of evaluation	Criteria
Final users of raw MT	- effective transfer of information from one language to another - acceptability - service conditions	- fidelity - intelligibility - legibility - reading time - cost - production time
Post-editors correcting MT	- acceptability (allied to the scale and type of corrections)	- post-edition rate - post-edition time
Decision-makers (responsible for the development of an MT system)	- potential market	- acceptability - cost - improvability (in synthesis)
System technicians (data-processing specialists, linguists, coders)	- error diagnosis (by type elements of the specific MT system concerned) - correctibility	- errors by causes - improvability (analytical)
Linguists	- errors diagnosis (by type grammatical and stylistic)	- errors by linguistic type
Heads of translation services	- number of corrections, perhaps classified by type - comparison of the features of the MT/post-edition circuit with those of the HT/revision circuit	- post-edition rate - cost - production time

3.132 General aims with respect to the users of the translation.

3.132.1 R. KUHLEN enumerates the twelve points which he feels may interest the users of the translation and which affect the evaluation criteria to be used

- interlingual transfer of the words in the text which are essential to comprehension, regardless of syntactic relationships (with subsequent HT once the relevance of the text is established)
- interlingual transfer based on a pre-defined syntactic and semantic standardization of the source text
- interlingual transfer based on a factual syntactic and semantic standardization of the summarized source text
- interlingual transfer of unprepared complete texts in areas with a specialized terminology, to obtain a general idea of the contents of the document
- interlingual transfer of standardized press releases by news agencies or official bodies with post-editing by the recipient
- interlingual transfer of routine texts, the recipients of which will simply scan rapidly
- interlingual transfer of texts for classification by subjects
- interlingual transfer of the semantic and pragmatic information of texts into a logical network incorporating a multilingual question-and-answer system
- conversion of texts into an international structure permitting production of multilingual abstracts
- translation to check the efficiency of linguistic models
- translation as a method of simulating human intelligence in defined situations
- translation in universal fields of application as complete substitute for HT.

3.132.2 J. WEISSENBORN defines the evaluation criteria to be used according to translation types and their qualitative aims

- translation intended for publication :
    - qualitative aim : perfectly correct
    - criterion : cost of post-editing, which in turn is a function of the number of translation errors
- translation intended to inform the specialist of the contents of a text :
    . qualitative aim : errors and gaps permissible
    . criterion : number of translation errors
- translation intended to give an overall picture of the contents of a text :
    - qualitative aim : low quality permissible
    - criterion : number of errors of morphology and syntax.

3.133 Specific aims : effectiveness and usefulness.

3.133.1 F. KNOWLES feels that the cheeks on the quality of an MT system must guarantee a sufficient level to enable a monolingual reader whose mother tongue is the target language, to undertake the necessary post-editing without risk of disaster.

3.133.2 For W. LENDERS, the aim of the evaluation is to assess the practical usefulness of MT rather than its linguistic exactitude.

It is necessary to consider first of all that MT is, generally, defective.

Nevertheless, it can be assumed that these translation can be used with care and rationally, either just as they are, or in a revised form.

It must be possible to ascertain if and when the products of MT can be understood by the users and usefully applied in their daily work.

3.133.3 A.J. PETIT. When evaluating a translation produced by a machine to determine if a system under development meets the requirements or if a system proposed by a supplier is in accordance with the description given of it, it is not a matter of evaluating just a text, but the characteristics of a production tool.

The evaluation method has to be based on a knowledge of the problems and their scale, and has to make it possible to check point by point whether the system comprises all the characteristics necessary to translate effectively. Instead of taking a text and trying to classify the errors, the evaluator will establish requirements corresponding to each of the evaluation criteria and will seek in the translated text all the errors which can be assigned to this criterion. In certain cases, each time the machine successfully resolves a Problem, the cause of this success will be ascertained by means of a check test and if it becomes evident that it can be assigned to a human intervention (coding, for example), the test will be repeated on a similar example which has not been coded.

Any on-the-spot correction has to be regarded as a failure and the use of on-the-spot corrections (or specific coding) will result in the irrevocable refusal of the system.

3.134 Specific aims :capacity and priorities for improvement.

3.134.1 For R. GREEN, one of the essential aims of an evaluation is to detect the errors in translation, and assess their seriousness, so as to be able to decide priorities as regards improvement of the system.

3.134.2 Z.L. PANKOWICZ notes that, up to now, all evaluations of MT have had a political aim.

Their results are consequently dubious, being based as they are on a prior bias, either against MT in general, or in favour of a particular MT system.

All the evaluations of MT carried out in the past aimed at measuring the quality of systems at their level of development at the moment of the evaluation.

However, it is essential, for the users and the purchasers of such systems, to know their capacity for improvement and the limits thereon.

Improvement work can not, in fact, be carried out infinitely, and this work should therefore be directed in such a manner as to optimize its results.

3.134.3 For G. VEILLON, an MT programme has the unfortunate property of never being finally correct, of being in perpetual evolution. It is thus on this "potential" aspect which the evaluation must bear :

a programme has value precisely in its possibilities for enrichment and improvement.

It is necessary consequently to evaluate :

-from the point of view of the user who is not a computer specialist : ease of detection and correction of errors resulting from

- pre-edition and input
- dictionaries *
- grammars *

-from the point of view of the computer specialist : the design of the software, making it possible :

- to integrate it into text-handling, and in particular text-editing system
- to extend the programme with new modules which improve its performance
- to transfer it to other computers

-from the point of view of the cost of the human operators responsible :

- for Pre-editing and post-editing the texts
- for updating the dictionaries and grammars.

* G. VEILLON is evidently considering the hypothesis of a user who is not a computer specialist, but who is a member of a design or maintenance team for an MT system. The normal user is not interested in detecting and correcting this type of error.

3.14 Assessment.

Apart from the contribution of A.J. PETIT, the aim of which seems to be above all to show the impossibility of the contributions of the authors who considered the aims of the evaluation are either convergent, or complementary.

It appears agreed that :

- one essential aim of MT is to be useful to its users (F. KNOWLES and W. LENDERS)

- machine translation can be undertaken with the aim of translating various types of texts, each of these types having a specific qualitative aim and consequently requiring the application of specific evaluation criteria (R. KUHLEN and J. WEISSENBORN)

- evaluation of MT has to be done in the light of the various categories of recipients of the evaluations, each category having one or more specific aims, and the methodology of the evaluation having thus to be specific to each group (A. ANDREEWSKY and G. VAN SLYPE)

- the evaluator of MT has to concern himself not only with the quality of the system, but also with its improvability (Z.L. PANKOWICZ and G. VEILLON) and the selection of the points to be improved (R. GREEN).

3.2 Translation quality.

3.21 Introduction.

Logically, an evaluator has to start by asking himself what is actual object of its activity. It is consequently normal that the majority of the studies on evaluation of MT and the communications submitted to the Luxembourg Conference of February 1978 should include a discussion and/or a proposal for a definition of the quality of translation. In certain cases, a link is made between qualities to be measured and the measuring criteria.

The dozen contributions below on this subject, have been broken down into three groups :

- a definition of translation
- a series of summaries or extracts on translation quality
- contributions on the relation between translation qualities and evaluation criteria.

3.22 Table of proposed definitions.

Definitions	Authors
Concept of translation	J.HOUSE
Quality of translation	ASSOCIATION J. FAVARD H. BRUDERER R.L. JOHNSON R. KUHLEN Z.L. PANKOWICZ A.J. PETIT Y. WILKS
Link between translation qualities and evaluation criteria	G. BOURQUIN M. MASTERMAN A.J. PETIT PHILIPS

3.23 Discussions on the concept of translation and on translation quality.

3.2.31 Definition of translation.

J. HOUSE.

Translation is the replacement of a text written in a source language by a semantically and pragmatically equivalent text written in the target language.
(The translation of oral texts is different activity, namely interpretation).

3.232 Translation quality.

3.232.1 L'Association Jean FAVARD distinguishes

- the intrinsic qualities, which are independent of the reader

- the extrinsic qualities, which are related to the "text-reader" couple.

A text, even badly translated(and thus of low intrinsic quality) can nevertheless, for an informed reader, be as clear as if it had been well translated.

However, beyond a certain deterioration in intrinsic quality, the extrinsic quality becomes very poor.

7.232.2 For H. BRUDERER, quality is a relative concept, i.e. one related to a specific object. Quality can apparently be measured, at least in part, but it remains much more difficult to quantify abstract (conceptual, subjective) phenomena than concrete (perceptible, real, tangible) things.

Quality can be evaluated :

- either positively assessment of merits, advantages

- or negatively : assessment of deficiences, errors, disadvantages

- or totally : assessment of the positive and the negative aspects.

The evaluation of the translation quality - whether human or computerised - has to take into account the following intra-linguistic and inter-linguistic factors: morphology, syntax, content, terminology, style, conformity.

A faithful translation reproduces the sense of the original text, but it does not necessarily, if it is to be considered an intelligent translation, have to be identical to the original text. Given that they partially overlap, content and fidelity should be evaluated on an overall basis. Similarly, it is difficult to differentiate clearly syntax and semantics. Style, on the other hand, influences all levels (morphology, syntax, semantics, terminology).

3.232.3 P.L. JOHNSON defines translation quality by three factors fidelity, intelligibility and elegance. The importance of these three factors may vary with the type of text considered.

Features can be observed :

- superficially, via linguistic elements such as lexical and syntactic exactitude
- indirectly, via the reactions of the users to the translated text.

3.232.4 R. KUHLEN stresses that there is not a universal criterion for MT evaluation :

- on the one hand because it does not seem that MT can ever reach the level of quality of human translation
- on the other hand, because the evaluation criteria have to be chosen according to the aim in view
- finally, because the individual parameters, which taken together permit an assessment of the quality of MT, often contradict each other, with the result that an overall rating would not be significant to the specific -Performance of the components.

3.232.5 Z.L. PANKOWICZ feels that usefulness of MT and HT has to be based on quality, speed and cost. Determination of the optimal balance between these three parameters depends on the environment of each translation activity.

It is necessary to understand, in his view, that the quality of HT and MT is indefinable, at least in any absolute way. The assessment of the quality of HT is traditionally based on its completeness and on stylistic elements.

3.232.6 A.J. PETIT takes the view that the translation should not comprise any misconstruction, but admits however a tolerance of up to 1% of the sentences in the case of translations to be supplied raw to the final user and 2% of the sentences in the case of texts to be revised before submission to the users. This tolerance is intended to allow for normal risks of error or accident.

3.232.7 Y. WILKS thinks that, the purist who feels that the least translation defect nullifies the translation is often mistakes in two of his postulates :

-he exaggerates the attention and comprehension which the average reader achieves with a technical document (consequently, errors of translation do not negate the value of the text)

-he exaggerates the quality of the mass of human translations produced on an enormous scale and at high speed.

3.233 Relationship between translation qualities and evaluation criteria.

3.233.1 According to G. BOURQUIN, the criteria for evaluating a translation will vary according to whether it is produced by a human translator or by the machine

-from the human, "finesse" will be required : open to the ethnoculture and to work on linguistics, the human translates with his sensivity, his intuition, his common sense

-the computer will be expected to offer regularity, precision, infallibility, speed, and encyclopaedic exhaustiveness.

3.233.2 M. MASTERAN notes that our ignorance of the very nature of translation leads to a discordance between the evaluation criteria used or proposed by various authors.

3.233.3 A.J. PETIT.

A product is acceptable only if it meets the requirements of its users. As regards texts (original texts or human or machine translations), the principal requirements are :

- utility technical texts (maintenance or user manuals):
  - no errors
  - homogeneity
  - clarity, without ambiguity or gibberish which might obscur the sense of the message
  - simple correct style, without extravagances or recherche elements
  - use of the terms recognized in the relevant sector
- educational technical texts :
  - no technical errors
  - adaptation of the text to the reader and cultural transposition of any reference or any comparison whose aim is to render comprehensible the material being taught
  - simple correct style
  - introduction to the terms recognized in the relevant sector
- documentary scientific texts
  - clear exposition of theory, without errors
  - flowing style without excessively long sentences incorporating several different ideas
  - use of the basic terminology of the discipline.

These requirements have however to be viewed from a different angle according to whether the translation is intended :

-to be revised : in this case, the translation system (human or machine) has to be aware of its own shortcomings, and indicate by itself all the ambiguities which it was not able to resolve : it delivers an in complete product, but one without serious defects

-to be supplied direct to the final user : the translation must then be complete (experienced human translator or a computerised system producing a complete translation, without any misconstruction) and without serious defects (human error or accident both being normal risks).

3.233.4 The authors of the report presented by PHILIPS distinguish between evaluation of translations with and without comparison with the source text.

In the first case, it is necessary to assess in what measure the translation :

- reproduces which is stated in the original(for example: contractual texts)
- reproduces what the author of the original intends to say, with the certainty that the message is properly understood (for example : translation of manuals).

To assess the quality of a translation, it is necessary to answer the following questions :

- on the aim of the translation :
  - does the translation reproduce the content of the original?
  - does the translation reproduce the formulations of the original?
  - does the translation reproduce the intention of the author?
- on the type of text:
  - is all the information presented?
  - can the translation achieve the desired effect?
  - have the necessary corrections been made in such a way that communication has the best chance of success?

In the second case (evaluation of the translation without reference to the original), the assessment of the quality of the translation has to cover :

- the grammatical correctness
- style and idioms
- the use of current words, expressions and structures in the target language
- the absence of contradictions or ambiguities.

3.24 Assessment.

The concept of the quality of a manufactured product is, in general, unambiguous : the product has to correspond to the specifications and a battery of quality control tests can easily be arranged, and made the responsibility of controllers often relatively unqualified.

The concept of translation quality is much more and the authors' contributions can be summarized fairly briefly :

- the quality has to be assessed, not in the absolute, but according to the aims of the writer of the texts to be translated and by those who decide how it is to be distributed
- the quality achieved by HT can not be expected of MIT, and the latter has thus to be used for more limited aims than the former (which does not mean that, within the scope of these limited aims, there does not exist a major potential demand)
- the evaluation criteria have to be chosen according to these specific aims
- since translation quality can not be measured in the absolute, on the basis of a single criterion, its assessment should combine several criteria.

3.5 Text typology.

3.31 Introduction.

We gathered ten extracts from documents dealing with the problem of text typology.

These extracts can be classified in two different ways, according to consideration either :

- of the criterion or criteria proposed as a basis for the typology of the documents
- or the purpose proposed for this typology.

These two methods of classification are equally useful. Thus, in order not to lengthen this report by covering the extracts twice, we drew up a double-entry grid, indicating, for each author, the type of criteria and the purpose proposed. The extracts themselves are then presented in author alphabetical order.

3.32 Table of proposed typologies.

Note : the typologies whose author's name is underlined have actually been used on an experimental basis by their author.

Purpose Criteria	Evaluation	Assessment of difficulty of texts	Detection of Machine translatable texts	Determination of translation methods
Pragmatics	ROLLING		VAN SLYPE	SAGER JOHNSON
External form				
Functions	HOUSE			
Role of textual units	LEAVITT			
Source language grammar		WEISSENBORN		
Scale of difficulties				PHILIPS
Formal structures			HOFSTETER	
Linguistic characteristics			BOURQUIN	

3.33 Description of proposed typologies.

3.3301 G. BOURQUIN feels that in order to evaluate objectively the fidelity of a translation, it is necessary, as a preliminary, to clarify what in the source product has to reappear in the target product : one can measure the adequacy of B with respect to A only after specifying that with respect to which B is held to be adequate. For these reasons, the way to a definition of methods and criteria of evaluation as regards translation is via construction of a typology discourse.

G. Bourquin proposes to consider four criteria for text classification; stressing, however, that these criteria constitute research topics rather than final answers : operative typologies will be obtained only by successive approximations based on obstacles actually encountered in translation. Text classes and error classes will be progressively set in statistical correlation, which will lead eventually to more realistic criteria. At the end of this process -which is likely to be lengthy and to involve many investigators - it will perhaps be possible to set less subjectively than is the case today the limits of what is translatable and qualitative standards.

The criteria for text classification proposed by Bourquin are :

- according to the referential function :
    - discourse with isolable functions i.e. independent of the mode of expression
    - auto-referential discourse turned in on itself and referring only to its own internal structure
    - mixed discourse, spanning the whole range between these two poles, either simultaneously or successively
- degree of normality (texts with isolable referential function only) :
    - referential function belonging to an existing configuration
    - referential function running counter to known ideas (latest research, epistemic breakdown)
    - immediate (transparent) relation between the vehicular language and the logico-conceptuel referent (1Durely denotative formulation; direct translation)
    - mediate (opaque) relation (connotative formulation; translation by simulation, including use of stylistic methods such as the metaphor)

- information density :

  - when the relation between the language and referent is opaque, the redundancy is integrated into the heuristic -Process,
  - but, otherwise the redundancy is useless, and the translator must eliminate it : in this case a faithful translation is one reproducing what was said, not the way of saying it

- nature of the text-author relationship:

  - utility or technological discourse factual, descriptive, argumentative, explanatory, etc.
  - uniformly impersonal discourse or discourse containing Personalized passages.

On a first analysis, it seems that:

- the human translator is best suited to text which is not conformist, and/or is argumentative and/or is strongly personalized
- the computer is better adapted to translation of text which is conformist (with a predictable and stable phraseology), factual and not personalized.

3.3302 H.HOFSTETTER proposes that the texts to be translated should be characterized not by classifying them in a limited number of extrinsic classes, but by analysing the formal structure number of words per sentence, number of words of less than four characters, number of conjunctions, of prepositions, of subordinate clauses, of noun expressions, etc.

It would then appear to him possible, by means of regression analysis, to determine the weighting of these variables, based on an evaluation of the quality of the translation based in turn on the time necessary for the post-editing a sample. These data once acquired, it should be possible to calculate a priori the machine translability" of a text on the basis of (automatic) detection of its characteristics.

3.3303 J. HOUSE proposes and actually uses (on a sample of eight documents) a typology based on the functions fulfilled by the texts, i.e. on the use made of them by the recipients.

She actually makes use of this typology during an valuation of human translation, and we have therefore classified the summary of her work in the chapter on evaluation criteria (§ 3.5).

3.3340 R.L. JOHNSON feels that a typology based only on the stylistic or linguistic characteristics of the texts would be of less practical utility than one based on the external form; for example : memorandum, scientific paper, technical specifications, etc.
It will in fact be on the basis of this categorization that translation services will decide whether to have a text translated by MT or HT.

3.3305 A.W. LEAVITT advances a classification not of the documents, but of sections within each document, called "textual units", intended for MT evaluation purposes.

The textual unit results from a progressive subdivision of the document up to the point where any additional division would cause the author's intention to be obscured.
A textual unit has the following characteristics :

- taken alone, it retains its capacity to communicate a meaningful item of information

- it expresses a complete thought and may be withdrawn from its context without fully losing its meaning.

- if it is subdivided any, further, it loses its meaning.

List of the textual units.

- Statement of a problem: statement of the conditions which justify establishment of a technical aim or statement of the aim

- method : description of the activities of the investigator and justification of these activities

- conditions : statement of the context of the work, including a description of the surrounding characteristics presumed to influence certain results, definitions, concepts involved by the work, and statement of the constraints showing the limits of the technical work

- proposals : assumptions, axioms, lemmas, theorems and statements of a priori technical specifications

- result : data, derivations, corollaries, proofs and entities arising from the subject or from previous inputs

- conclusion : statement of an interpretation or a conviction concerning the reality, the confidence or the applicability a discovery.

Note : thus defined, the textual units could extend over one or several paragraphs. In reality from the example provided, it appears that they are no longer than one or two sentences. The interest of this internal document typology is that it makes it possible to judge the importance of each category of textual unit with respect to the functions which can be fulfilled hereby, in particular :

- selection of relevant documents
- identification of relevant parts of documents
- expansion or improvement of knowledge.

The experimental implementation of this method by Leavitt has shown up the difficulties of it :

- lack of consistency in the subdivision into textual units of the same documents by several persons

- lack of consistency in the classification (according to the six categories above) of the same textual units by several persons

- lack of consistency in the weighting of the same semantic unit classes with respect to the functions which can be fulfilled thereby.

The idea of textual units, initially established by Leavitt for the evaluation of SYSTRAN Russian-English, was finally not applied.

3.3306 The authors of the PHILIPS report present a table of the scale of difficulty of the translation, based on difficulty factors, taken from of a publication of K. REISS (*)

Difficulty Factors	Scale of difficulty		
	1st grade	2nd grade	3rd grade
Linguistic			
-Language level	Ordinary language Technical and (cultivated and special languages colloquial language)	Technical and special languages	Poetic (artistically shaped) language
-Syntactic semantic structure	Clear, simple method of expression and development of ideas	Hermetical, complex expression and development of ideas	Defective expression and development of ideas
-Translation from - to	From the foreign language into the mother tongue	From the foreign (which is not the mother tongue of the author) into the mother tongue	From the mother tongue into the foreign language
-Function of the text	Informative (primarily referring to the content)	Expressive (primarily referring to the form)	Operative (primarily referring to behaviour)
-Function of the Translation	Expressing the sense	Expressing the sense and adequate reproduction of the form	Expression of the sense, adequate formal and analogous statistic or operative formation
-Content	Field open to general experience	Field can only be dealt with after technical training	Field only to be dealt with if the translator is congenial in his approach
- Content of the text			

(*) REISS (K.).- Zur Bestimmung des Schwierigkeitsgrades von Ubersetzungen.- Mitteilungsblatt fur Ubersetzer und Dolmetscher BDU, May/June 1974.

Difficulty factors	Scale of difficulty		
	1st grade	2nd grade	3rd grade
- Cultural association	Cultural of the source language and the target language are cognate (for example English/German)	Cultural of the source language and the target language are very far apart (for example Japanese/German)	A great differences in the cultural level between the source language and the target language
<u>Technical</u>			
Presentation of the text	Printed or typed text	Manuscript	Recorded text
-Aids for acquiring and extending linguistic and technical competence	Are available	Are scare or inadequate	Are not available

They propose also that texts to be translated should be characterized according to :

* the communication function of the text :
    - mainly descriptive (accent on the content)
    - mainly expressive (accent on the form)
    - mainly appellant (accent on the appeal)
* the presentation of the texts :
    - normal text
    - text with illustrations
    - questionnaire
    - lecture (with adaptation of the syntax to the oral presentation)
    - lecture with slides
    - series of slides with commentary
    - film commentary.

For each of these types of text, the translation method has to be different.

3.3307 L. ROLLING proposes that texts should be characterized by four types of criteria. This characterization will make it possible to evaluate the quality of a translation by comparing it to that of an ideal translation or to that of the source text.
These four criteria are:

- The criterion of precision (P) will make it possible to:
  classify texts into those by which the whole of an item of information or contents of a message can be transmitted to the reader (rating 0), those which do not manage to transmit the information or the message at all(rating 10), and those of an intermediate level, where there are doubts on the information, which comprise ambiguities, which fail to express essential nuances, which have a picturesque or allegorical style or those where the reader has to "read between the lines".

- The criterion of complexity (C) will make it possible to classify texts into those which consist of elementary sentences, comprising only a subject, a verb and possibly a complement (rating 0), those which comprise the most complicated sentence structures, a multiplicity of subjects, verbs and complements of all kinds, which are broken up by mathematical or chemical formulae, brackets and illustrations of all kinds and which have a staccato syntax comprising noun clusters (rating 10), and a complete range of texts of intermediate complexity.

- The criterion of technicality (T) makes it possible to distinguish texts consisting only of words so frequent in use that they may be assumed to be universally known (rating 0) and those comprising a very high number of words from special nomenclatures and known only a number of experts (rating 10).

- The criterion of “correctness” (F*)* makes the distinction between texts free of any kinds of errors (rating 0) and those which comprise many misspellings (due to the author or the transcription), mistakes in syntax and layout, errors (rating 10).

A Published scientific text has a tendency to be precise, fairly complex, highly technical and correct (P = 0, C = 5, T = 10, F = 0).

A legal text is usually precise, highly complex, fairly technical, and correct (P = 0, C = 10, T = 5, F = 0).

A Poetic text will comprise imprecisions, images, intentional ambiguities, will be highly complex, basely technical at all, and correct (P = 10. C = 10, T = 0, F = 0).

The rapid transcription of a journalistic report, dictated to a stenographer, may be fairly precise, not very complex, not very technical, but full of grammatical mistakes and mis-spellings (P = 5, C = 0, T =0, F = 10).

In translation, whether human or computerised, a precise text makes possible and requires a precise translation.

A simple text is easy to translate, while a complex text requires of the inventorof the translation system, as of the human translator, resources of ingeniousness.

A highly technical text does not pose a problem to a system equipped with a complete dictionary, but costs the human translator precious time.

A relatively untechnical text is welcomed by translators but requires complex homograph routines of any system.

Finally, misspellings and errors of syntax are easily corrected by the human translator, but they are beyond the capabilities of a machine translation system.

The art of a translator is measured above all in the skill with which he transposes the nuances and the ambiguities of one language to another. He will be judged on his precision.

The degree of perfection of a translation system, an the other hand, will be defined by its skill in disentangling the syntactic maze of complex texts. It will be judged on its capacity to restore the complexity.

3 .3308 J.C. SAGER refers to the necessity of basing an MT evaluation on a categorization of the texts to be translated : MT should not, be regarded as a group of processes, each applicable to a specific category of texts and of translation, and each requiring development as far as possible within this limited context.

Typology of texts and of translations can not, at present, be based on a linguistic theory, but it is possible, on the other hand, to base it on a pragmatic analysis.

Texts can be classified according to a certain number of characteristics, for example:

-textual :

*semantic : disciplines and special aspects covered:

- application of one subject to another (example administration of education)
- points of view (example : history)
- type of reference (general or special)
- description system : linguistic or non-linguistic (example : mathematical formula)

*syntax, i.e. preponderance of certain structures, sentence length, etc.

*form (example : report, résumé, article, etc.)

*composition (example : sub-heading, list, etc.)

-situational :

- relationship between author and reader (number, social roles, etc.)
- aim of the complete text : informative, directive,
- discursive, etc.
- aim of parts of the text
- conventions
- modes of expression : rigid, strict, advisory, etc.
- use : ephemeral or durable.

Translations too can be classified according to a certain number of categories :

- preliminary or final translation

- simple or multiple translation (one source language and several target languages)

- internal use

- translation with legal force

- working paper

- publication

- educational course.

Knowledge of the volumes to be handled in each of these classes make it possible to decide priorities.

3.3309 G. VAN SLYPE considers it useful to establish a categorization of texts to be translated to determine those among them which lend themselves more to MT and those whose frequency justifies the recourse to this method.

It seems, unfortunately, difficult to achieve, a priori, a categorization which is effective and on which decisions may be based. It is moreover possible that a categorization which proves useful for a given translation system, or for a language couple, or a particular discipline, is not useful in other circumstances.

It seems that there are no studies in this field, and it appears in consequence necessary to start from scratch.

The methodology proposed is as follows :

-establishment of a list of criteria on which a categorization may be based; for example :

* source of the texts to be translated

* length of the sentences

* number of clauses per sentence

* type of message: referential (centred on the underlying sense; example : organizational note), expressive (centred on the author; example : novel, certain political speeches), conative (centred on the recipient; example : publicity, certain political speeches), metalinguistic (centred on the code; example : definitions), phatic (centred on the communication; example : polite formula, certain political speeches),poetic (centred on the form of the message; example certain novels)

* specialized vocabulary / general vocabulary ratio

* ratio of proper names and other words

* type of document : scientific or technical review article, newspaper article, minutes, study report, legal text, legal judgment, instruction for use, market research, bibliography, etc,

* number of authors
* stylistic quality of the original text
* character : descriptive, prescriptive
* redundancy.

It will be necessary, moreover, to take into account the fact that the sentences of the texts to be translated are never entirely homogeneous as regards these criteria. Certain texts will be relatively homogeneous (with a low scatter each side of the average value of the criteria measured for each of the sentences) and others not : this heterogeneity will in fact constitute one of the criteria to be considered

-cross-referencing with the performance ratio obtained following evaluation of a sufficient text sample in each of the categories and search for correlations

-listing of the correlations higher than a certain threshold and, on this basis, the relevant categories in a first analysis

-checking of the universality of these categories by application to other language couples, other disciplines and other translation systems.

In conclusion, it appears that if it is desired to establish a classification of the texts which will be useful to those taking the decisions on MT, the determination of effective classification categories will require a thorough study which will without doubt need a multi-field approach and long-term work.

Pending the results from this study, it will be possible to take account, during evaluation work, only of a limited number of criteria, chosen from those whose relevance seems highest :

- titles and texts
- summaries and texts
- type of document (review articles, working paper, service note, etc.)
- number of words and clauses per sentence
- mother tongue of the authors
- subjects covered.

3.3310 J. WEISSENTBORN proposes classifying texts according to a typology arising out of the grammar of the source language:

-initially, a subset of the grammatical rules is defined which will allow total and unambiguous analysis of the text

- by means of the number and type of rules used, a typology of the texts characterized by these rules (degree of difficulty of the text) is then established.

Other parameters can also be used to characterize the texts

- specificity of the vocabulary
- number and type of ambiguities.

3.31 Assessment

Examination of the literature on typology of texts for translation leads to the following conclusions :

- nobody questions the relevance of such a typology, provided that it is functional and that decisions can be based on it (type of translation to undertake, evaluation criteria to use)
- at the concrete level of the daily round of translation services, a typology of this type is applied informally, but without doubt very effectively.

  The most usual criteria applied are

  - the purpose of the text (working paper, publication, speech, etc.)
  - the subject (scientific, legal, etc.)
  - the deadline requested
  - the acceptable cost (for customers of private services)

- on the scientific level, there exist a large number of proposals, but the few which have been tried out in practice have proved to be inoperative, except perhaps those which are based on very simple criteria : mother tongue of the authors, length of the sentences
- the extremely high cost which without, doubt the development and testing of the proposed methodologies would involve, and the relatively low probability of their success, suggest that the steps needed in this field are a matter for fundamental rather than applied research.

3.4 Effectiveness and efficiency of the evaluation.

3.41 Introduction.

Five authors raise the problem of the effectiveness and the efficiency of the evaluation of translation, and their contributions can be classified in three groups :

- description of the characteristics required of a good evaluation system
- weighting of the evaluation criteria according to their usefulness and their cost
- calculation of the correlation between the results provided by the various evaluation criteria and determination of certain strongly correlated criteria which might be redundant.

3.42 Table of analyses of the effectiveness and the efficiency of evaluation methods.

Analyses	Authors
-characteristics of an evaluation system	JOHNSON LEAVIT
-Efficiency of an evaluation system	VAN SLYPE
-Correlation between evaluation criteria	CAROLL LEICK SINAIKO

3.43 Description of effectiveness and efficiency factors.

3.431 Characteristics of an evaluation system.

3.431.1 R.L. JOHNSON states that, ,when designing a test, the evaluator has to he aware of two conditions which the test has to satisfy :

- validity : a valid test is one which does indeed measure the attribute in question. If this attribute is not directly observable, the evaluator has to choose the characteristics which are observable and which contribute to the property in question, and to design the test in such a way as to exclude any interference from other factors

- reliability : a reliable test which provides, with a high degree of confidence, a result very near to the real value of the attribute being tested.

The sources of non-reliability are :

- biais due to the learning effect (same test applied to the same subject(s) in rapid succession)
- bad selection of the element to be tested (composite attribute)
- variance between evaluators (particularly when one seeks to quantify a value judgement)
- various sources of variation : season, sex, age, etc. of evaluators.

3.471.2 A.W. LEAVITT establishes the following list of characteristics for an effective evaluation system:

-applicability to all translations of scientific and technical documents

-sensitivity to the properties of the translation which facilitate the execution by the use of identifiable tasks

-sensitivity to the parties of the translated documents which are most important for the achievement of identifiable tasks

-ease, economy, and significance of the measurement, within operational constraints

-lowest possible effort for implementation and use.

3.432 Efficiency of an evaluation system.

G. VAN SLYPE suggests that the evaluation of a system Should remain within reasonable financial limits, and therefore believes it useful to limit the number of criteria to be measured to the essential minimum.

For this purpose, he takes a list of criteria and indicates beside each of them, based on the experience gained in the first evaluation of the SYSTRAN English-French system of the Commission of the European Communities :

- a weighting from 1 to 3 measuring the usefulness of the criterion to the recipients : decision-takers, final users, translators and revisers
- an approximate measurement, from 1 to 3, of the cost of taking this criterion into account in an evaluation
- the ratio between the usefulness and the cost
- an indication of the criteria where the ratio is equal to or higher than 1, which are those which he proposes to retain.

CRITERIA	Usefulness	Cost	Ratio	To be retained
- Intelligibility	3	2	1.5	X
- Overall assessment	2	1	2	X
- Correctness or distorsion of the information	2	3	0.66	
- Acceptability	3	2	1.5	X
- Reading speed	1	3	0.33	
- Frequency of consultation of original	1	3	0.33	
- Revision rate	2	2	1	X
- Revision speed	2	2	1	x
- Recognition and reconstruction of sentence structures	1	3	0.33	
- Recognition and reconstruction of parts of speech and agreements between them	1	3	0.33	

3.433 Correlation between criteria.

3.433.1 J.B. CAPROLL, in his evaluation of MT for the ALPAC committee, took the general view that the two principal criteria of quality are intelligibility and fidelity, while theoretically these two criteria are independent of one another.

in practice, he noted at the end of the evaluation a very strong correlation between them.

During the evaluation, he also had the reading times of each of the evaluators taken, sentence by sentence.

This measurement however served to note only that :

- reading times are in linear relationship with the ratings given to the quality of the translation, which leads CARROLL to the conclusion that his rating scale is regularly spaced
- the time spent on their work by bilingual evaluators is appreciably longer than that of their monolingual colleagues, which means that the former use their knowledge of the target language to endeavor the better to understand the translations.

3.433.2 J.M. LEICK discovered the following in the evaluations of SYSTRAN with which he was involved :

-English-French system :

- no correlation (r = 0.10) between revision rate and fidelity

-French-English system :

- very weak correlation (r = 0.32) between revision time and revision rate
- correlation (r = 0.65) between revision time and fidelity.

3.433.3 H.W. SINAIKO uses systematically several criteria to measure the legibility of the translation, but wonders as the relevance of all these measurements : if, in fact, the results obtained from one criterion are in close correlation with those provided by another, only one of the two criteria (the most convenient to use, for example) need be retained for subsequent evaluations.

If, on the other hand, the correlation between the results is weak, this means that the two criteria measure distinct aspects and have thus to be retained.

The correlations calculated by SINAIKO between the various legibility criteria which he uses prove to be of varying significance, depending on the texts being evaluated.

Only in certain cases, is there a correlation between:

- the number of correct answers and the number of answers omitted in the Cloze test
- the results of the Cloze test and the clarity on the Sinaiko scale
- the reading time (but not response time) and the clarity.

3.44 Assessment.

The authors quoted are in agreement in considering the evaluation of MT as a process which can be expensive, requiring as it does the implementation of several distinct criteria, each of them revealing a particular facet of MT quality. Hence the interest in identifying these facets and studying the evaluation method which would be effective and the most efficient for each of them.

Unfortunately, as indicated in paragraph 3.24, the concept of MT quality remains relatively confused, and until its various component parts have been pinned down, the evaluation methods to be used will simply have to be selected empirically.

3.5 Macroevaluation - Criteria and methods.

3.51 Introduction.

The macroevaluation of a system is the operation which consists in assessing the manner in which the system answers to the requirements and the needs of its users, actual or potential, regardless of what occurs inside the "black box". The aim of macroevaluation is to measure the adequacy of the output from the system to its environment, without seeking to diagnose the causes of its inadequacy, if any, and without to pinpoint the component or components that could usefully be modified to improve adequacy.

Macroevaluation is an appreciation of performance as such, not an analysis of possible improvements.

The field of macroevaluation is limited:

- on the one hand, by the marketing, as regards the extent to which a product or a service meets the market demand
- on the other hand, by microevaluation, which is concerned with the diagnosis of errors and with improvability.

It would be possible to envisage establishing a classification of the techniques of macroevaluation on two separate levels

- a list of criteria (example : intelligibility)
- a list of methods of measuring, these criteria (example Cloze test).

In fact, as is underlined by Y. WILKS in his criticism of T.C. HALLIDAY, certain methods can be used to measure the value of several distinct criteria.

H.W. SINAIKO, for his part, points out that it is desirable to use several evaluation methods to improve the power of the evaluation, so as to measure the various aspects of style quality of the translation.

It seemed to us, consequently, convenient to associate the measuring methods with the criteria, and to present, for each criterion, the methods practised or proposed by the various authors, which in certain cases, involves t-he same methods being used to measure different criteria.

We broke down the various criteria into ten classes, assembled in turn into four groups according to the level at which they approach the quality of the translation.

- Cognitive level (effective communication of information and knowledge)

  - intelligibility
  - fidelity
  - coherence
  - usefulness
  - acceptability

- Economic level (excluding costs)

  - reading time
  - correction time
  - translation time

- Linguistic level (conformity with a linguistic model)

- Operational level (effective operation).

3.52 Table of criteria and methods of macroevaluation.

Note : the authors whose name is underlined are those whose method is actually used.

Criteria	Methods	Authors
3.521 Cognitive level 3.521.1 Intelligibility		
- Intelligibility	Rating of sentences read on a 9-point scale	CARROLL
- Intelligibility	Rating of sentences read on a 7-point scale	CROOK & BISHOP
- Readability	Cloze test (every eight/word)	CROOK & BISHOP

Criteria	Methods	Authors
- Readability - Comprehension - Comprehensibility	Clozentropy Noise test Multiple-choice questionnaire	HALIIDAY HALLIDAY LEAVIT
- Intelligibility	Rating of texts read on 9-point scale	LEAVIT
- Comprehension	Multiple-choice questionnaire	LEAVIT
- Clarity	Rating of sentences read on 3-point scale	ORR
- Clarity	Rating of sentences read on 9-point scale	PFAFFLIN
- Comprehension	Knowledge test	SINAIKO
- Readability	Multiple-choice questionnaire+Cloze test (every fifth word) + clarity measurement + time measurement	SINAIKO
- Intelligibility	Rating of sentences read in their context on 4-point scale	VAN SLYPE
- Intelligibility	Rating of sentences read on2-point and a 3-point scale	VAUQUOIS
3.521.2 Fidelity.		
- Informativeness/fidelity	Rating of sentences read on 9-point - point scale	CARROLL
- Fidelity	Rating on 25-point scale	CROOK & BISHOP
- Fidelity	Assessment of the correctness of the information transferred	HALLIDAY

Criteria	Methods	Authors
- Informativeness/fid-elity	Rating of texts units read on 9-point scale	LEAVIT
- fidelity	Rating of texts on 100-point scale	MILLER & BEEBE_CENTER
- fidelity	Shannon measurement of the quantity of information transferred	MILLER & BEEBE_CENTER
- Fidelity	Re-translation	SINAIKO
- Fidelity	Rating of sentences read on 4-point scale	VAN SLYPE
3.521.3 Coherence. - coherence	-	WILKS
3.521.4 Usefulness.		
- Quality	Final user' judgment	DOSTERT
- Quality	Composite measurement of fidelity, intelligibility and elegance	JOHNSON
- Usefulness	Evaluation from the point of view of the user	KROLIMANN
- usefulness or applicability	Task importance rating, And rating of the texts'relative usefulness, on a 9-point scale	LEAVITT
- Usefulness	Assessment of the possibility for actual usage	LENDERS
- Quality	Rating on a 3-point scale	PFAFFLIN
- Adequacy	Rating on a 3-point scale	PFAFFLIN
- Usefulness	Performance test	SINAIKO
- Usefulness	Rating on an 8-point scale	SLANSER

Criteria	Methods	Authors
3.521.5 Acceptability. - Acceptability - Acceptability	Analysis of user motivations Direct questioning of final users -	DOSTERT VAN SLYPE
3.522 Economic level. 3.522.1 Reading time.		CARROLL, DOSTERT, A.D. LITTLE, ORR, PFAFFLIN, SINAIKO, VAN SLYPE
3.522.2 Correction time. - Ease of post-edition - Overall performance - Revision time and post-editing time	Post-editing time Correction time -	ANDREEWSKY HOFSTETTER VAN SLYPE
3.522.3 Production time.	-	DOSTERT PANKOWICZ
3.523 Linguistic level - reconstruction of semantic relationships - Syntactic and semantic coherence	- -	ANDREEWSKY AAOCIATION JEAN FAVARD

Criteria	Methods	Authors
- Absolute translation quality	-	HALLIDAY
- Lexical evaluation		MILER & BEEBE-CENTER
- Syntactic evaluation	-	MILLER & BEEBE-C.
- Power of the MT system	-	WEISSENBORN
- Error analysis	-	WEISSENBORN
3.524 Operational level		
- Automatic language identification	-	HALLIDAY
- Verification of claims		PANKOWICZ
	-	

3.53 Description of criteria and methods of macroevaluation.

3.531 Cognitive level.

3.531.1 Intelligibility

3.531.11 Definitions of the criteria.

We grouped various related criteria under the heading "intelligibility" : intelligibility, clarity, comprehensibility, legibility.

The definitions given of them are as follows:

- Intelligibility

T.C. HALLIDAY.

Ease with which a translation can be understood,
i.e. its to the reader.

G.-VANSLYPE.

Subjective evaluation of the degree of comprehensibility and clarity of the translation.

- Comprehensibility

T.C. HALLIDAY.

Comprehensibility relates to the degree of perfection with which a complete translation can be understood (whereas the intelligibility is based on the general clarity of a translation, whether this is considered in its entirety or by segments out of context).

Note : according to T.C. HALLIDAY, intelligibility and comprehensibility are, in current use, synonymous terms : he differentiates between them only from the point of view of his analysis.

- Readability

T.C.-HALLIDAY.

Measurement of the total contextual coherence.

H.W. SINAIKO.

Comprehensibility of a translation to a representative reader.

- Clarity

H.W. SINAIKO.

Alternative to intelligibility.

3.531.12 Methods of evaluation.

3.531.12.01 J.B. CARROLL : Measurement of intelligibility by rating sentences on a 9-point scale

* Method.

- Reading by a group of readers of translated sentences detached from their context
- Rating of each sentence on a 9-point scale from 1 to 9
- Calculation of the average of the ratings given.

CARROLL scale (based on an adaptation of a psychometric technique known as the method of equal-appearing intervals) :

9: Perfectly clear and intelligible. Reads like ordinary text : has no stylistic infelicities

8: Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or midly unusual word usage that could, nevertheless, be easily "corrected"

7: Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8

6: The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements.
Post-editing could leave this in nearly acceptable form

5: The general idea is intelligible only after considerable study, bull after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible

4: Masquerades as an intelligible sentence, but actu ally it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely appre hended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated

3: Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence

2: Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical

1: Hopelessly unintelligible.

It appears that no amount of study and reflection would reveal the thought of the sentence.

* Applications.

Evaluation of automatic translation for the ALPAC group.

3.531.12.02 CROOK & BISHOP (reported by T.C. HALLIDAY) :
Measurement of intelligibility by rating complete 7-point scale.

* Extraits from the 7-point scale.

1: About as good as comparable material in the target language

7: Only a vague impression of the meaning can be obtained.

3.531.12.03 CROOK & BISHOP (reported by T.C. HALLIDAY): Measurement of readability by the Cloze test.

* Method.

-Translation of a text by HT and MT

-Elimination of one word in 8 in each of the two translations

-Communication of the two texts, each to a group of readers, who are required to fill the blanks with the words which they consider correct

* Advantages.

-Very high coherence of results, as from one group of readers to the other

-Easy to use

* Measurements.

-Number of answers comprising exactly the suppressed original word

-Number of answers with a word close in meaning to the one suppressed

* Application.

Yes.

3.531.12.04 T.C.-HALLIDAY : Measurement of readability by the Clozentropy method.

* Method.

As the CROOK & BISHOP Cloze test, but with establishment of the ratio between the number of correct answers obtained in the MT and HT versions

* Source.

This method was invented by T.C. HALLIDAY from a psychological test of linguistic competence developed by DARNELL, in a field other than translation.

3.531.12.05 T.C. HALLIDAY : Measurement of comprehension by the noise test.

* Method.

The test consists in measuring the intelligibility of sentences expressed orally at a constant voice loudness, but with addition, by electronic means, of noise of a loudness increasing in steps : 1 dB, 4 dB, 7 dB, 10 dB, 50 dB.

T.C. HALLIDAY contemplates the application of this method to MT, by assimilating to the noise applied the distortion of the sense given by MT

* Application.

None; this method was invented by T.C. HALLIDAY from the SPOLSKY psychological linguistic competence test.

3.531.12.06 A.W. LEAVITT : Measurement of comprehensibility by a multiple-choice questionnaire.

* Method.

Use of the multiple-choice Questionnaire conceived by ORR for machine translation of scientific and technical documents, measuring:

-the number of correct answers

-the time elapsed

* Application.

Comparative pilot evaluation on SYSTRAN and the MARK II (Russian-English) system.

3.531.12.07 A.W. LEAVITT (reported by T.C. HALLIDAY) :
Measurement of intelligibility by rating texts on a 9-point scale.

* Method.

Similar to that of J.B. CARROLL, but applied to textual units rather than isolated sentences

* Application.

Experimental

3.531.12.08 ORR (reported by T.C. HALLIDAY) :
Measurement of comprehension by a multiple-choice questionnaire.

* Method.

-Construction of three types of multiple-choice questionnaire :

- direct questions, based on explicit statements in the text
- “equivalent” questions, based on paraphrases of the data in the text
- indirect questions, based on information not explicitly contained in the text

-Execution

-Counting of the number of correct answers from several translations of the same text

* Application.

Only to HT.

3.531.12.09 PFAFFLIN (reported by T.C. HALLIDAY)
Measurement of clarity by rating sentences on a 3-point scale.

* Method.

-Constitution of three sets of sentences detached from their context :

- a set of MT sentences
- a set of HT sentences
- a mixed set including each sentences from of the two preceding groups

- Reading of these sentences by three groups of readers
- Rating of each sentence on a 3-Point scale:

  - clear
  - not clear (because of a bad translation or an ambiguous construction)
  - meaningless

* Application.

Experimental.

3.531.12.10 H.W SINAIKO : Measurement of the clarity by rating sentences on a 9-point scale.

* Method.

Application of the CARROLL intelligibility scale, slightly modified to make it more easily comprehensible to Vietnamese evaluators.

* Extraits from the rating scale.

9: perfectly clear and comprehensible; appears good to the reader

1: not comprehensible at all; no amount of study would be able to help a reader to know what is the principal idea

* Application.

Evaluation of E1glish-Vietnamese MT produced by the LOGOS system.

3.531.12.11 H.W. SINAIKO : Measurement of comprehension by the knowledge test.

* Method.

-Translation of a text from a language A into a language B

-Preparation of a questionnaire suitable for assessing the knowledge which a reader has of the text; the questionnaire being in both language A and language B

-Questioning " open book" method, with the text visible of two groups of readers: one of language A, the other of language B

-Recording and rating of the answers, per individual

-Calculation of the averages by group

-Comparison.

* Advantages.

-Objectivity

-Cheapness.

* Disadvantages.

-Need for readers from each of the two languages

-Need for readers with a certain competence in the field covered by the text.

* H.W. SINAIKO's conclusion.

-A method to be recommended

* Application.

Evaluation of English-Vietnamese MT produced by the LOGOS system.

3.531.12.12 H.W. SINAIKO : Measurement of readability by a combination of various methods.

* Method.

Measurement by a combination of criteria (each criterion measuring a specific aspect of readability).

-Multiple-choice questionnaire, covering material from the texts of the sample, -"open book" method; counting of the number of correct answers

-Cloze procedure deleting one word in five (cf. § 3.531.12.03); counting of the number of correct answers (spelling errors accepted, but synonyms not) and the number of answers omitted

-SINAIKO clarity scale

-Reading time and response time for the various tests

* Application.

Evaluation of English-Vietnamese MT produced by the LOGOS system.

3.534.12.13 G. VAN SLYPE : Measurement of intelligibility by rating sentences on a 4-point scale.

* Method.

-Submission of a text sample in several versions (original text, MT without and with post-editing, human translation with and without revision) to a group of evaluators; the texts being distributed so that each evaluator :

- receives only one of each of the versions of the texts
- receives a series of sentences in sequence (sentences in their context)

-Rating of each sentence according to a 4-point scale

-Calculation of the average of the ratings per text and version, with and without weighting as a function of the number of words in each sentence evaluated.

* Scale of intelligibility.

3: Very intelligible: all the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing, or are badly translated, but close to the target language

2: Fairly intelligible:the major part of the message passes

1: Basely intelligible:a part only of the content is understandable, representing less than 50% of the message

0: Unintelligible: nothing or almost nothing of the message is comprehensible

* Application.

Evaluation of the SYSTRAN English-French MT system acquired by the Commission of the European Communities.

3.531.12.14 B. VAUQUOIS : Measurement of intelligibility of sentences on two scales : 3-point and 2-point.

* Three-point scale.

-Very comprehensible sentences
-Sentences understandable with considerable difficulty
-Indecipherable sentences

* Two-point scale.

-Comprehensible sentences
-Incomprehensible sentences

* Application.

Evaluation of the Grenoble Russian-French MT system.

3.531.2 Fidelity.

3.531.21 Definitions of the criterion.

We have grouped here the criteria known by the authors as "fidelity", “correctness”or “precision”.

Two definitions follow :

T.C. HALLIDAY.

Measurement of the correctness of the information transferred from the source language to the target language.

G. VAN SLYPE.

Subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation.

The fidelity rating should, generally, be equal to or lower than the intelligibility rating, since the unintelligible part of the message is of course not found in the translation. Any variation between the intelligibility rating and the fidelity rating is due to additional distortion of the information, which can arise from :

-a loss of information (silence) (example : word not translated)

-interference (noise) (example : word added by the system)

-a distortion from a combination of loss and interference (example : word badly translated).

Note : detailed analysis of the lack of fidelity of the translation is very difficult to carry out, for each sentence conveys not an item of information or a series of elementary items of information, but rather a message or a series of complex messages whose relative importance in the sentence is not easy to appreciate.

3.531.22 Evaluation methods.

3.531.22.1 J.B. CARROLL : Indirect measurement of fidelity by rating the informativeness of sentences on a 9-point scale.

* Method.

- Machine translation of a text
- Reading of the translation by a group of readers
- Subsequent reading and evaluation of the original text by the same readers
- Rating of each of the sentences of the original text, on the basis of the additional information in it which was not provided by the MT, on a 10-point scale from 9 (“very informative”-which means that the MT is very bad) to 0 ("the original contains less information than the translation - which is thus better and more informative than the original"). If, therefore, the translation is faithful, the informativeness is low, and if it is not, the informativeness is high.

* Scale of informativeness.

9: Extremely informative. Makes "all the difference in the world" in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation)

8: Very informative. Contributes a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the reader’s impression of the meaning intended, although not so much as to change or reverse the meaning completely

7: (Between 6 and 8)

6: Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader "on the right track" as to the meaning intended

5 : (Between 4 and 6)

4: In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words

3: By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different "twist" to the meaning conveyed by the translation. It adds no new information about sentence structure, however

2: No really new meaning is added by the original, either at the world level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended

1: Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced

0: The original contains, if anything, less in formation than the translation. The translator has added certain meanings, apparently to make the passage more understandable

* Application.

Evaluation of Russian-English MT for the ALPAC group.

3.531.22.2 CROOK & BISHOP (reported by T.C. HALLIDAY).

* Application.

Experimental.

3.531.22.3 T.C. HALLIDAY: Measurement of fidelity by assessment of the correctness of the information transferred.

* Method.

-Translation of a text from a language A into a language B

-Comparison of the two versions by a bilingual expert, who judges the correctness of the information transferred from language A to language B

-Value judgement

* Application.

Evaluation of SYSTRAN.

3.531.22.4 A.W. LEAVITT (reported by T.C. HALLIDAY): Indirect measurement of fidelity by rating the informativeness of textual units on a 9-point scale.

* Method.

Similar to the CARROLL method, but dealing not with isolated sentences, but with textual units, i.e. blocks of text fully treating an idea or a concept

* Application.

Experimental.

3.531.22.5 MILLER & BEEBE CENTER (reported by T.C. HALLIDAY):Measurement of fidelity of the translation by rating on 100-point scale.

* Method.

-Comparison of MT with an HT or the original

-Rating of the whole on a 100-point scale

* Application.

None.

3.531.22.6 MILLER & BEEBE-CENTER (reported by T.C. HALLIDAY): Measurement of fidelity by a method based on Shannon's theory of the quantity of information.

* Method.

-MT and HT if a text

-Calculation of the quantity of information (= H (HT)) of the HT version by asking a reader (R 1) to guess in succession each of the letters of the HT text

-Reading of the MT version by a second reader (R 2)

-Calculation of the quantity of information(=H (MT)(HT)) of the HT version when MT is known, by asking the reader R2 to guess in succession each of the letters of the HT text

-Calculation of the total information common to the MT and HT versions (= T) :

$$T = H(MT) - H(MT)(HT).$$

Note: The method measures the transfer of information in probabilistic terms, not in semantic terms

* Application.

None.

3.531.22.7 H.W.SINAIKO : Measurement of fidelity by re.translation.

* Method.

-Translation of text samples from language A into language B

-Re-translation of the texts, from language B back into language A

-Comparison between the original text and translation, analysis of the divergences and more particularly of the errors

* Error criteria.

Any part of the re-translation which is judged not to carry the same significance as the original text is regarded as a translation error

* Measurement scale.

-Addition : an additional word or expression appears in the re-translation

-Minor omission : one or two words of the original are omitted from the re-translation

-Major omission : three words or more of the original are omitted from the re-translation

-Mutilation : three words or more of the re-translation are incomprehensible

-Minor substitution : one or two words of the original do not have an equivalent in the re-translation, but an expression replaces the original words

-Major substitution : three words or more of the original do not have an equivalent in the re-translation, but are replaced by an expression.

-Finally, the re-translation can be regarded as equivalent to the original and marked "OK "

* Advantage.

The evaluator does not have to know the target language

* Disadvantage.

An error can be due:

-either to the translation into the target language

-or to the re-translation back into the source language

* Conclusion.

Method to be used in conjunction with other tests

* Application.

Evaluation of English-Vietnamese MT produced by the LOGOS system.

3.531.22.8 G. VAN SLYPE : Measurement of fidelity by rating on a 4-point scale.

* Method.

-Submission of a sample of original texts, with the corresponding translations, to one or more evaluators

-successive examination of each sentence, in the first place in the translation, then in the original text

-Rating of the fidelity, sentence by sentence

-Calculation of the average of the fidelity ratings

* Scale of fidelity.

3: Completely or almost completely faithful

2: Fairly faithful: more than 50 % of the original information passes in the translation

1: Barely faithful : less than 50 % of the original information passes in the translation

0: Completely or almost completely unfaithful

* Application.

Evaluation of English-French MT produced by the SYSTRAN system of the Commission of the European Communities.

3.531.3 Coherence.

One author only, Y. WILKS, proposes this criterion

* Definition of the criterion.

The quality of a translation can be assessed by its level of coherence without the need to study its correctness as compared to the original text. Once a sufficiently large sample is available, the probability that the translation should be at the same time coherent and totally wrong is very weak

* Advantage.

The assessment of the coherence can thus be done
by a monolingual evaluator, whereas any judgement on the correctness of the translation necessar1y involves making use of a bilingual evaluator

* Method of evaluation.

Y. WILKS does not indicate, unfortunately, how in practice it is possible to rate the coherence of a text. He notes that if an original text may be coherent, this means that any assessment of the coherence of its MT version may not be absolute, based on the MT,but must be relative, as compared to the coherence of the source text.But then one is once again compelled to use bilingual evaluators!

3.531.4 Usability.

3.531.41 Definition of the criterion.

One author, W. LENDERS, defines usability (which he also calls applicability) as the possibility to make use of the translation.

Another, P. ARTHERN, defines usability as far as a translation service is concerned, as revisibility.

3.531.42 Evaluation methods.

3.531.42.1 B.H.DOSTERT: Measurement of the quality by direct questioning of the final users.

* Method.

B.H.DOSTERT, in his questionnaire addressed to users of MT, sought to pin down the concept of quality through a series of questions, many of which bear on the usability of MT:

-Does MT make it possible to judge the importance

-Does MT supply sufficient information on the content of the text ?

.

- Did the MT have to be followed by an HT ?
- is the MT examined simultaneously with the original text
- Quality rating for the MT (3-point scale good, acceptable, poor)
- Reasons affecting comprehensibility (sentence structure, not translated words, lack of diagrams, formulae and figures, badly translated words, other reasons)
- Percentage of MT sentences which are deformed, incomprehensible, comprehensible with difficulty, comprehensible, correct
- Percentage of technical words not translated, incorrectly translated, incomprehensible, deformed, comprehensible with difficulty, correct
- Possibility of mental correction of the style, of mental translation of a not translated word from the context or the original : often, sometimes, never
- Percentage of texts incomprehensible by comparison with the HT
- Does the inadequate translation of words, expressions or sentences result in a complete distortion of the meaning?
- Do distortions in the MT lead to misinterpretations?
- Can a translation of low clarity cause dangerous effects?
- Possibility of getting used to the style of MT

* Application

Evaluation of Russian-English MT systems derived from the Georgetown system, in Ispra and Oak Ridge.

3.531.42.2 R.L.JOHNSON takes the view that the evaluation of translation quality is the result of a group of factors such as fidelity, intelligibility and elegance, which are observable, superficially, from linguistic factors such as lexical and syntactic exactitude, and indirectly through the reactions of the human users of the translated text.

* Method.

In order for an evaluation procedure to be useful, its result should be a small number of values.

The variability inherent in any measurement of quality 11 Q 11 makes it desirable to treat Q statistically, and in practice, it is convenient to consider Q in the form

$$Q = T + e$$ where

T is the true measure of quality and constitutes an invariant property of the system itself

e is the cumulative effect of the error arising out of the test in that it is an imperfect indicator of the true property.

Thus, the variable part of any observed measurement of quality is a function of the test used, and this test only.

The sources of this variation can, in theory, be broken down into :

-a systematic error whose magnitude depends on the degree to which the test is a valid measure

-a random component, which varies non-systematically according to the reliability of the test (cf. § 3.431.1).

The evaluation then comprises two stages:

-construction of a valid and reliable test

-practical application of the test.

The construction of the test has to be based on sound statistical principles, so as to permit generalization and extrapolation of the test procedure to other applications.

The problem of the design of the test has been treated by several behavioural scientists, and approaches to the question have become very sophisticated.

The creation of techniques based on the linear model in particular has enabled designers of experimental tests to exploit all the power of experimental design and analysis which is associated with the analysis of variance (ANOVA).

A particular development of the theory, due to CRONBACH (cf. CRONBACH and a!.- The dependability of behavioural measurements. -Wiley, 1971), called “generalisability” gives very high flexibility of control over influencing factors, and the power to generalize extensively from the test situation.

3.531.42.3 F. KROLLMANN feels that evaluation has to be under taken from the viewpoint of the user, rather than that of the producer of the translation, and therefore proposes that the evaluation should be based on several criteria:

-Diffusion of the information

-Readability, style and diction

-Purely linguistic criteria, for example

- grammatical correctness
- choice of the words
- spelling and punctuations.

3.531.42.4 A.W. LEAVITT : Measurement of usability by Task Importance Rating and the relative usefu1ness of the texts.

* Method.

A.W. LEAVITT has developed a method of evaluation of translation quality, called ASTUTE (Assessor for Translation-User Textual Elements).

ASTUTE is a group of quality measures intended to assess the improvements in translation techniques.

The two most original measures concern the usability of the translation:

-Rating of the relative importance of the textual units as compared to the professional activities of the user of the translation (9-point scale measuring the importance of each of the tasks, at textual unit level : identification of relevant documents and parts of documents)

-Rating of the relative usefulness of the translated texts in providing factual information for the readers, on a 5-point scale.

In addition to this assessment on the usability of the texts, A.W. LEAVITT proposes to use the more traditional criteria of intelligibility (using the CARROLL 9-point intelligibility scale at the level not of sentences, but rather of textual units), and of fidelity (using the CARROLL 10-point scale of informativeness -which is the reverse 0J. fidelity)

* Application.

Only as a test during a pilot evaluation.

3.531.42.5 W.LENDERS : Measurement of usability by assessment if the possibilities_for_actual use.

* Method.

The evaluation has to be carried out on two levels:

-Comparison of the texts produced by MT and HT

-Comparison of summaries and lists of descriptors produced by documentalists from, respectively, MT and HT.

1. Evaluation of the texts.

- First phase.

An external observer questions (by interview or questionnaire) users of MT and HT who have made use of the two types of texts for a fairly long period. The texts concerned must be such as actually arise in these users professional activities.

- Second phase.

A socio-scientific investigation is carried out on subjects who are not necessarily users.

For both groups of evaluators (users and non-users), two series of criteria are used :

- subjective criteria, intended to obtain the general impression and opinion :

  - intelligibility
  - absence of ambiguities
  - syntactic correctness
  - fidelity
  - absence of contradictions
  - stylistic quality
  - acceptability
  - precision

- intelligibility measurement criteria:

- correct reproduction of the conceptual relations in the text by a pattern of conceptual interdependences
- correct filling of the gaps created by the experimenter in the translated text
- questions on the conceptual characteristics of the text which the subject should have understood and retained while reading.

After collection of the values assigned by the subjects to these various criteria, establishment of relationships with certain characteristics of the text :

- applicability/error rate
- complexity/intelligibility
- applicability/competence of the user.

This makes it possible to determine:

- the degree of quality necessary for translated texts to be comprehensible
- the measure to which the errors can be compensated for by the technical competence of the user; this figure permits deduction of the maximum error rate which the system may present to be usable.

2. Evaluation of summaries and lists of descriptors.

Judgement of the quality of summaries and lists of descriptors (established by documentalists from MT on the one hand, from HT on the other) by external specialists thoroughly familiar with the constraints of a documentary information system, on the basis of a series of criteria

- linguistic intelligibility of the texts
- conceptual intelligibility of the texts
- identification of logical relationships
- reproduction of the technical expressions
- time required
- psychological factors (for example : the authority granted to the text if it is not known that it is an MT text)
- determination of identical descriptors.

* Application.

Evaluation of the SYSTRAN Russian-English system by the University of Bonn.

3.531.42.6 J. HOUSE : Measurement of translation quality by the method of analysis of situational dimensions.

* Method.

The evaluation of translation quality is based on a text typology.

This typology is based in turn on the functions fulfilled by the texts, and not on those of the language, since translations of concrete texts are involved.

Furthermore, this typology is based on the functions fulfilled by the texts rather than on the intentions of their authors : the first can in fact be found in the text, whereas intentions can not be established empirically.

The function of a text is defined as the application or use of the text in the specific context of one and only one situation.

To characterize the function of a text, it is then necessary to analyse the "situational dimensions" of this text.

Basing herself on the system of CRYSTAL and DAVY, J. HOUSE proposes the model of multi-dimensional analysis of the texts according to:

-The characteristics of the user of the language:

- geographical origin (regional dialect)
- social class (social dialect)
- time (temporal origin of the text)

-Characteristics of the use of language:

- as a medium

  * simple text written to be read, but not aloud

  * complex text written

    -to be read aloud

    -to be enunciated as if it had not been written

    -to be read as if it was heard

- participation

  * simple: text produced by one person only

  * complex:

    - text produced by two or more persons

    - text produced by one person only but with the participation of the recipient (interrogations, imperatives, etc.)

- social role relation (role relations between the writing of the text and its recipients)

  - ♦ symmetry: solidarity, equality
  - ♦ asymmetry authority

  - ♦ permanent position role (teacher; priest)
  - ♦ transient situation role (visitor)

- social attitude : degree of social distance, characterized by five levels formalism:

  * rigid
  * strict (no participation of the recipient)
  * advisory (with basic information supply)
  * casual (certain elements implicit)
  * intimate (many implications)

- province:

  * occupational and professional activity (example : scientific or advertising languages, etc.)
  * subject of the text, field.

The analysis of the situational dimensions of the text provides a textual profile.

Comparison of the textual profile of the original text and of the translation permits evaluation of the quality of the translation.

Analysis of the functions of the documents forms a basis for the following typology :

- ideational texts : expression of the content the author's vision of the external world, as well as his experience of the interior world of his own consciousness

    - technical (for example : scientific text, commercial text)
    - non technical (for example : journalistic article, tourist brochure)

- interpersonal texts : expression of the relationship between the author and the readers

    - non fiction (for example : religious sermon, political speech)
    - fiction (for example : moral anecdote, comedy dialogue).

In addition to the text typology which she advances J. HOUSE proposes also to consider a typology of translations.

The first typology implies in effect that the quality of the translation is determined by the nature of the source text and that the process of translation is a constant. In fact, it appears to J. HOUSE, after actual analysis of eight, texts, that a more appropriate typology would have to be based on the type of translation required by the various types of texts.

Its classification then becomes as follows:

- overt translation : refers to texts specific to the culture of the source language, the contents of which have only a potential value for other cultural communities

    - with non-specific recipients : the text is not bound to a given historical occasion and is fictional (for example moral anecdote, comedy dialogue)
    - with specific recipients the text is bound to a given historical occasion and is non-fictional (for example : religious sermon, political speech)

- covert translation : refers to texts not specific to the culture of the source language

- with non-specific recipients and text not bound to a given historical occasion (for example : tourist brochure, scientific text, journalistic article)

- with specific recipients and text bound to a given historical occasion (for example commercial text).

This type of typology would have as a consequence a re-examination of the principle of functional equivalence, which has to serve as a criterion for translation quality :

- an overt translation of a source text bound to the source culture, has to fill a similar function (to a second degree) in the culture of the target language

- a covert translation, of a source text not bound to the source culture, has to fill an equivalent function in the two cultures (application of a cultural filter, so that the "effect" of the text on readers in the two languages - source and target- is the same).

Note.

Application of this method leads to essentially qualitative results. For example : evaluation of the quality of the translation of a scientific article (on the application of partial differential equations in physics) is summarized as follows by the author :

> Comparison of the source text and the translation according, to the eight functional dimensions shows a certain number of non-coincidences between the two dimensions ‘social role relation’ and ‘ province’.
>
> The non-coincidences as regards social role relation, which decrease the didactic and instructional nature clearly weaken the interpersonal component of the function of the text.

The non-coincidences as regards the province, which make the translation less coherent, also weaken the interpersonal component, making the text potentially less easily assimilable by novices in the province.

Considered together, the ideational components of the function of the text which consists in transmitting a factual item of information have been preserved at all contributive levels, and also because there is no non-coincidence between the denotative significance of the elements of the source text and of the translation.

However, the interpersonal function of the textual function, i.e. the match between the material and the requirements of its recipients has been violated in certain cases on the two levels which contribute to this component.

* Application.

Eight machine-translated texts.

3.531.42.7 PFAFFLIN (reported by T.C. HALLIDAY):

Measurement of adequacy by rating on a 3-point scale.

* Method.

- Reading of various machine-translated texts, by a group of readers
- Rating of each text on a 3-Point scale

• adequate

• adequate as a guide for deciding whether to ask for a better translation

• useless

* Application.

Yes.

3.531.42.8 H.W. SINAIKO: Measurement of usefulness by performance test.

* Method.

- Choice of a text the content of which describes a process to be carried out by a human being (example: a maintenance manual), instruction by instruction
- Translation from the original language A into a language B
- Submission of the translation to an operator of language B
- Performance by the latter of each of the instructions
- Measurement of the mistakes made in performance

* Advantages.

- Evaluation taking into account many aspects of translation quality
- Objective and effective method

* Disadvantages.

- Expensive
- Slow
- Restricted to a limited number of types of texts

* Conclusion.

Useful method in limited cases

* Performance scale (instruction by instruction).

- No errors
- Minor error
- Major error

* Application.

Evaluation of English-Vietnamese MT produced by the LOGOS system.

3.531.42.9 SZANSER (reported by T.C. HALLIDAY):
Measurement of usefulness by rating on an 8-point scale.

* Method.

- In a first phase, the readers evaluate the MT completely subjectively, without referring to the original
- In a second phase, the persons responsible 'or the test assign a usefulness rating, on a scale from 1 to 8, based on their interpretation of the readers'evaluation

* Extraits from the scale of adequacy.

8 : fully adequate

7 : between 6 and 8

2 : poor

* Application.

Yes

3.531.5 Acceptability.

3.531.51 Definition of the criterion.

Only one author, G. VAN SLYPE, defined the concept of acceptability, as a subjective assessment of the extent to which a translation is acceptable to its final user.

3.531.52 Evaluation methods.

3.531.52.1 B.H. DOSTERT : Measurement of acceptability by analysis of user motivation.

* Method.

Several of the questions asked by B.H. DOSTERT in his survey of users of MT deal with their motivation

- Why do you use MT ?
- How much MT do you request per year ?
- What is the reason for which you use MT (cost, speed, confidence, exactitude) ?
- Do you recommend MT to your colleagues ?

* Application.

Evaluation of Russian-English MT produced by the systems derived from the Georgetown system, in Ispra and Oak Ridge.

3.531.52.2 G.VAN SLYPE : Measurement of acceptability by direct questioning of users.

* Method.

- Submission of a sample of MT with the original texts and the corresponding HTs, to a sample of potential users
- Questions asked (among others)

• Do you consider the translation of these documents to be acceptable, knowing that it comes from a computer and that it can be obtained within a very short time, of the order of half a day ?

* in all cases

* in certain circumstances (to be specified)

* never

* for myself

* for certain of my colleagues

- Would you be interested in having access to a system of machine translation providing texts of the quality of those shown to you ?

* Application.

Evaluation of the English-French SYSTRAN system of the Commission of the European Communities.

3.532 Economic level.

Note : This report is devoted to the evaluation of the quality of translation, regardless of its cost.

In this section we concern ourselves only with economic criteria not directly concerning the cost of MT. That does not mean, of course, that the cost of translation is not an important factor.

B. VAUQUOIS, for example, feels that the homogeneous measure for all types of translation (summaries, working papers, technical literature, normative texts, etc.), whatever translation method is employed (MT with pre-edition, interactive MT, MT with post-editing, human translation), is the total cost including text input and editing of the output.

One can then compare the cost price of a human translation with various methods of automated translation, according to the various standards of quality set for each application. Seen from this angle, the nature of the faults which the reviser has to correct is relevant only as it affects the time necessary to carry out the work.

Similarly, G. VAN SLYPE stresses the fact that the time for MT post-editing constitutes only one of the cost factors, and that when comparing MT with HT, it is necessary to take account of all the human interventions : human translation plus revision of the HT on the one hand. post-editing of MT on the other.

3.532.1 Reading time.

Reading time can be assessed in various ways :

3.532.11 B.H. DOSTERT : by asking final users to state what percentage of additional time they require to read MT, as compared to an original in their own language.

3.532.12 J.B. CARROLL : by timing the time spent by the evaluator in reading each sentence of the sample.

3.532.13 G. VAN SLYPE : by timing the time spent by the evaluator in reading each text of the sample.

3.532.14 PFAFFLIN and ORR (both quoted by T.C. HALLIDAY) by measuring the response time to a multiple-choice questionnaire.

3.532.15 H.W. SINAIKO by measuring the time necessary for the execution of the cloze test.

3.532.2 Correction time.

3.532.21 A.ANDREEWSKY: Measurement of the ease of post-editing by measuring the post-editing time.

* Definition.

Measurement of the ease with which post-editing can becarried out, such ease not necessarily being related to the number of corrections, since a single operation may, in fact, take longer than two or more operations.

* Method.

Measurement of the total post-editing time, text by text, and not sentence by sentence

* Application.

None.

3.532.22 A. HOFSTETTER : Measurement of total performance, by measuring correction_time.

* Method.

-Choice of sample of texts to be translated by
several machine translation systems

-Submission of the translations to several (at least 3) post-editors

-Post-editing of texts:

- until a quality is achieved which is judged by a supervisor or a jury to be homogeneous.
- with timing of the time spent on the correction of each sentence : the total correction time, for each translation system, characterizes overall performance level, but this is true only if each of the systems compared has comparable dictionaries for identical language couples. If this is not the case, the method must go further

- Selection of a number of "structural variables" (which, mathematically speaking, are the independent variables which explain the dependent variable constituting the correction time per sentence); three types of structural variable can be distinguished:

* variables easily calculable by a computer

- number of words
- number of words of less than 4 characters
- number of words of more than 12 characters
- number of commas
- number of characters
- etc.

* variables calculable by computer, on condition that a limited vocabulary is available on the machine :

  • number of conjunctions

  • number of prepositions

  • number of verbal auxiliaries both finite and non finite

  • etc.

* variables calculable by computer, on condition that a wide vocabulary and a syntactic analyzer is available on the machine:

* number of subordinate clauses

* number of coordinate clauses

* number of noun expressions

* number of words constituting verbal expressions etc.

- Automatic calculation (with, if necessary, manual revision) of the value of these variables for each of the sentences of the samples

- Regression analysis, individually for each system tested, to calculate the value of the factors weighting each of the variables in the equation:

$$a1\,x + b1\,y + c1\,z\,. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ T1$$

where x, y, z ..... = the independent variables defined above

a1 1 b1, c1 ...= weightings of these variables in the no 1 system

T 1 = correction time specific to no 1 system

- The value of these factors having thus been determined from text samples specific to each system under test (and established according to the fields covered by each of them), it is possible to calculate the theoretical value of the correction time, for each of the systems under test, from identical text samples, giving comparable correction time values.

* App1ication.

None.

3.532.23 G.VAN SLYPE : Measurement of revision and post-editing time

* Definition.

Time taken in reading through a translation, in examination of the original text as necessary, whether wholly or in part, in terminological research and in correcting the translation.

* Method.

Measurement of the correction time, document by document, by the revisor or post-editor himself.

* Application.

Evaluation of English-French MT produced by the SYSTRAN system of the Commission of the European Communities.

3.532.3 Translation production time.

The translation production time, i.e. the time between a request for a translation and reception thereof has been used as an evaluation criterion by B.H. DOSTERT and by Z.L. PANKOWICZ.

3.533 Linguistic level.

3.533.1 A. ANDREEWSKY : Measurement of the reconstruction of semantic relationships.

* Method.

- Counting of the number of correct (C) semantic relationships in the MT texts

- Counting of the number of wrong (W) semantic relationships in the same texts

- Calculation of the ratio C/W.

* Application.

None.

3.533.2 L'Association Jean FAVARD : Measurement of syntactic and semantic coherence.

* Criteria.

-Translation of the predicates with their agents, with specification of their internal structure : government, complements, incidence

-Translation compound noun groups

-Translation of constant words and expressions

-Correctness of the article

-Punctuation equivalence

-Examination and incorporation of referents.

* Method.

Establishment, for each sentence, of a list of the anomalies detected.

* Application.

None.

3.533.3 T.C. HALLIDAY : Assessment of the absolute quality of the translation.

* Method.

- Translation of a text from a language A into a language B
- Comparison of the two versions by a bilingual expert, who judges the correctness of the information transferred, the correctness of the syntax, and the style
- Value judgement.

* Application.

None.

3.533.4 MILLER & BEEBE-CENTER (reported by T.C. HALLIDAY):

Lexical evaluation.

* Method.

- Translation of a text by HT and MT
- Counting of the total number (= T) of words in the HT version
- Counting of the number of words common (= S) to the HT and MT versions
- Calculation of the evaluation score (= N)

$$N = S/T$$

* Variant.

The method can be refined by including in the “S” batch only the words which are both common and arranged in the same sequence in the two versions, but this has not been tried out on MT.

* Application.

None.

3.533.5 MILLER & BEEBE-CENTER (reported by T.C. HALLIDAY): Syntactic evaluation.

* Method.

- Establishment of an a priori list of syntactic constructions (example : noun-adjective combination)

- Translation of a text by HT and MT

- Counting of the total number (= T) of occurrences of these syntactic constructions in the HT version

- Counting of the number of occurrences (= S) of these constructions common to the MT and HT versions

- Calculation of the evaluation score (= N)

$$N = S/T$$

* Application.

Yes.

* Conclusion.

Inconclusive results.

* Variant.

Inclusion of the immediate constituents in the original version and the MT version; however, this method has not been tried out.

3.533.6 J. WEISSENBORN : Measurement of the power of a translation system.

* Method.

- Enumeration of the number of grammatical rules in the source language, for the type of text to be treated

- Enumeration of the number of the source language analysis grammar rules existing in the MT system of (S)

- Calculation of the ratio S/L.

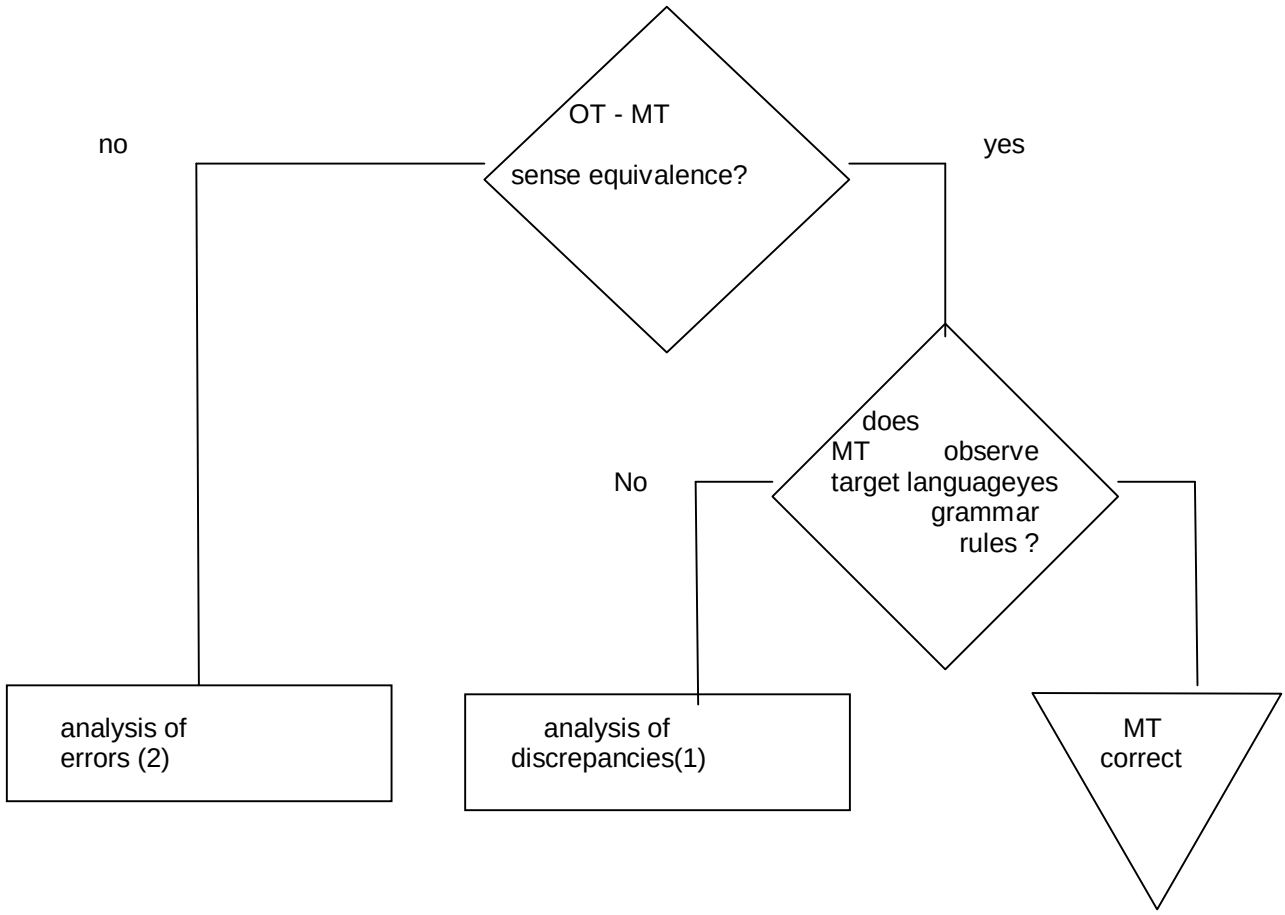
* Application.

None.

3.533.7 J. WEISSENBORN : Analysis of morphological, lexical and syntactic errors.

* Method.

- Selection of a sample of texts to be translated (OT)
- Machine translation (MT)
- Analysis, sentence by sentence, according to the following diagram :



(I) :

- Error types : morphological, lexical and syntactic
- Effects : on readability, but minor on the content of the message and on the cost of post-editing
- Most frequent source : transfer or synthesis component of the MT system.

(2) : Error types.

- Selection of tense, person, number, mode and type
- Choice of lexical unit
- Specification of the syntactic units and type of relationships between them
- Untranslated elements.

Most frequent source of errors.

- Analysis component of the MT system.

* Application.

None.

3.534 Operational level.

One way of evaluating an MT program is to run it and deduce useful conclusions from its o3Deration.

Two authors have undertaken an approach of this type T.C. HALLIDAY and Z.L. PANKOWICZ.

3.534.1 T.C.HALLIDAY : Automatic language identification.

* Method.

Submission of MT sentences to an automatic natural language identification program (NAKAMURA program)

Calculation of the percentage of sentences the language of which was recognized.

* Application.

None.

3.534.2 Z.L. PANKOWICZ : Verification of claims.

Z.L. PANKOWICZ notes that the publicity of translation system is exaggeratedly optimistic, and that the demonstrations arranged by their salesmen are not convincing; since the texts undergoing the demonstration may have beer, used to make the dictionaries and grammar rules include the specific elements which will lead to an excellent quality in the MT of these texts alone.

He therefore proposes that the potential customer for a system, should himself constitute the text sample to be translated (a continuous text of 5,000 words is enough at this stage), list the words in it and provide this list, arranged alphabetically, without translation, to the salesman.

Once this has been done, the text to be translated is given to the salesman, input and translated the same day, so as to avoid any modification of the dictionaries or grammar.

3.54 Assessment.

Let us recall here that our conclusion on the opinions expressed on translation quality (9 3.2), is that this does not constitute a homogeneous element, one which is measurable on one dimension only. All authors agree, on the contrary, that the quality of a translation has to be assessed by combining several different criteria.

It appears also (§ 3.4) that the evaluation of MT is an expensive operation.

It consequently becomes necessary to make use only of the criteria presenting a real relevance to the aims of the recipients of the evaluation (§ 3.1).

The report on MT macroevaluation experiments (§ 3.5) shows that, for the majority of the criteria, a whole range of different methods is available to evaluators. It will be consequently necessary, for each criterion used, to assess the cost/effectiveness ratio, i.e. the efficiency of each method, so as to obtain the most reliable results at the lowest cost.

We will therefore review here the various criteria, and indicate those which appear to measure a significant dimension of translation quality.
Then, for each criterion selected, we will propose a classification of the methods in decreasing order of efficiency.

We must stress yet again that :

-the selection of the significant criteria

-the classification of the methods in order of efficiency

will be based on the specific case of evaluation of MT by the Commission of the European Communities, taking into account the aims of the various departments of the Commission interested in MT.

This selection and this classification in no way seek to be universal, given the conclusions of § 3.2 on the concept of translation quality and § 3.1 on the aims of an MT evaluation.

This being the case, it is nevertheless true that certain criteria and certain methods can be significant and efficient in a large number of evaluation contexts, and that a broad consensus between evaluators ought to be attainable, so as to achieve results at least partially comparable as between evaluations by different teams and covering different translation systems (MT and HT).

3.54.01 Intelligibility.

The criterion of intelligibility, or one of its alternatives (readability, comprehension, comprehensibility, clarity), is the one most used for evaluation of MT, reflecting as it does directly the quality of the translation in the eyes of the reader who receives only the translated version. It is moreover the criterion which is used when it is desired to measure the effectiveness of the drafting of any text, translated or not (for example : maintenance manual) , handbook for training in a specific technique).

It consequently appears useful to make use of this criterion in any evaluation the quality of a text, and in particular of a translated text.

As to the method of measuring intelligibility, there is a wide range of possibilities :

- rating on an intelligibility scale
- Cloze test
- multiple-choice questionnaire
- knowledge test
- noise test.

Apart from the last one, the noise test, these various methods have already been used with success in evaluations of MT.

The features of the various methods are summarized in the following table.

Method	Advantages	Disadvantages	Cost rating (rising)
Rating on intelligibility Scale	-no preparation : the translation being handed as it stands to the evaluators -while the evaluators must have a certain knowledge of the subject, they do not have to be experts in it -direct measurement of intelligibility -the check-test of the source language can be superficial	-subjectivity of the rating (can be countered by using several evaluators and an explicit rating scale)	1
Cloze test	-objectivity of the rating (100 % for selection of the exactly right word, less than 100 % for selection of related concepts) -preparation of the evaluation text can be automated, deleting every x-th word	-the evaluators have to have a greater knowledge of the subject than in the preceding case ne check-test in the source language h be as thorough as the target language because since the "density" differs fro text to another, this distorting factor h be eliminated so as to compare the translation of several text	2

Method	Advantages	Disadvantages	Coat rating (rising)
Multiple-choice questionnaire	· effective measurement of information transfer · objectivity (though not 100%) of the rating of the answers	- subjectivity of the selection of the questions - onerous check-test in the source language - need for access to an expert to put relevant questions - need for the evaluation to have a good knowledge of the subject	3
Knowledge test	-measures both intelligibility and fidelity, thus giving a more complete assessment of the actual transfer of information via the translation	-subjectivity of the choice of questions -subjectivity of the rating of the answers onerous check-test in the source language need for access to an expert to put relevant questions and to rate the answers -need for the valuators to have a good knowledge of the subject -considerable time required for the evaluators to reply to the questions "open book" technique) and for the expert to rate the answers	4

In conclusion, rating on an intelligibility scale provides the best cost/effectiveness ratio and thus appears to be the method which should be selected.

It then remains to:

- decide what should be rated

- choose a rating scale

The method of direct rating on an intelligibility scale has a number of variables, depending on the authors.

These alternatives relate to:

- the element to be rated

  - sentences extracted from their context
  - sentences in their context
  - complete texts

- the rating scale

  - 2-point
  - 3-point
  - 4-point
  - 7-point
  - 9-point.

With regard to the element to be rated

-the method of rating the sentence out of context appears artificial : out of their context, sentences are very often less intelligible than when they are placed in their context. Since in reality, a sentence is almost always read in its context, there is consequently no reason to add a distortion factor additional to that caused a priori by the transfer between languages

the method rating the complete text appears much more subjective than that of rating, sentence by sentence., which in fact brings into Play the evaluator's immediate memory, and proceeds analytically, without requiring of the evaluator that he memorize and integrate judgements covering each part of the text

rating sentence by sentence, in context, appears thus both to correspond better to reading practice and to be surer.

With regard to the rating scale:

-a scale comprising a very low number of points seems insufficiently discriminatory

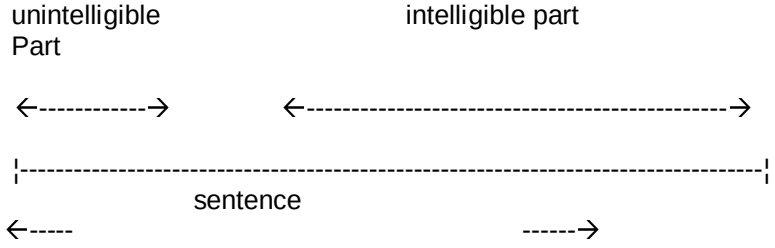
-a scale comprising a high number of points, assessment Of which remains in the final analysis subjective, involves too wide a scatter of the ratings

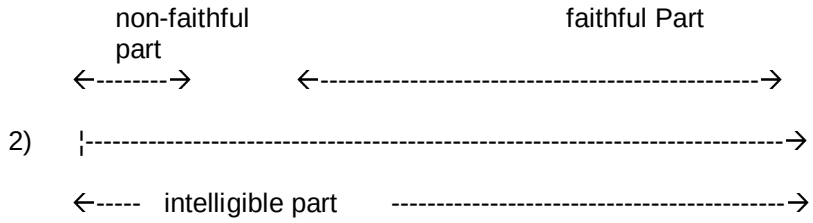
Furthermore, if one seeks, as did J.B. CARROLL, to clarify in detail each of the possible values of the scale, there is a risk of introducing elements not germane to intelligibility : as G. VAN SLYPE showed during the second evaluation of SYSTRAN for the Commission of the European Communities, the Carroll scale, in fact, measures at the same time intelligibility and style, and forces down the rating for sentences at the top of the intelligibility scale a scale comprising a modest number of points -four- appears consequently most adequate, in that it

- measures intelligibility only
- has a low scatter
- is of a sufficiently discriminatory character since the evaluation covers several hundreds of sentences and the average calculated as a percentage is sufficiently precise.

3.54.02 Fidelity.

Fidelity seems to be an excellent criterion for assessing translation quality.
The cost of measuring it is of the same order as that of measuring intelligibility, but its effectiveness appears much less certain.
When the fidelity of a translation is Judged, what is in fact done is measurement of the fidelity of the understandable part of the message transmitted, which combines two elements of subjectivity :





If it appears possible to judge “from outside” what is understandable or not in a sentence, it appears much more difficult to gauge the fidelity of the intelligible Part of the message a sentence in fact usually conveys not one, but several pieces of information, which have a different weight for each reader.

The use of several evaluators and the calculation of an average of their fidelity ratings would provide a meaningless result, incorporating as it would too many distinct elements.

In order to be able to make valid use of the criterion of fidelity, it would be necessary to undertake an analysis of the contents of each sentence of the sample, so as to obtain a complete picture of the elementary items of information transmitted, and it would be then possible to rate the fidelity of the translation of each of these pieces of information.

The cost of the operation would then become very high, without its effectiveness really being guaranteed (due to the subjectivity of the rating, and 'he difficulty in weighting the elementary items of information, which in turn is due to redundancy, prior knowledge in the reader, relative importance).

In conclusion, in spite of this criterion's attractiveness, it does not seem advisable to make use of it.

3.54.03 Coherence.

The reasoning advanced for the exclusion of fidelity applies also to the criterion of coherence, and in any event this criterion has not, been actually used by the author quoted.

3.54.04 Usability.

This criterion also appears at first sight to be very important. As in the case of the two preceding ones, fidelity and coherence, the problem is in effective measurement.

The various methods proposed all present a major disadvantage :

- either evaluators are asked to put, themselves in the place of the final user to assess the relative importance of the various parts of a text or its “situational dimension”, and to judge the usability of the translation of these passages, which is a completely artificial situation, in that only the final user can appreciate the importance of a text and judge the usability of the translation provided
- or else the methods are applied to final users and in order for them to be able to appreciate suitably the usability of the translation, it is necessary to subject them to a whole series of tests which do not provide an overall synthetic measurement, or to a series of performance tests, the result of which is certainly significant, but also very expensive, and in addition these performance tests can apply only when there is a performance to be measured (example: equipment maintenance).

We do not believe, consequently, which this criterion should be used for the evaluations of MT carried out for the Commission of the European Communities.

3.54.05 Acceptability.

Measurement of acceptability of MT by its final users presents many advantages :

-the judgement is made by the one for whom the translation is done

-the criterion is simple : a text is either acceptable or it is not

-the measurement relates to the actual purpose of the operation (acceptance or not of the translated text by the user) and not to an intermediate or partial aspect (intelligibility, fidelity, etc.); although the users’ judgement does include these elements : intelligibility, fidelity, usability.

The disadvantage of the method is that it deals with user/document couples presenting a very wide range of pairs of characteristics : aims of the user/types of document.
To obtain conclusive results, therefore, it is necessary to set a fairly large sample of users and of documents, and the method then becomes very expensive.

Within the framework of a macroevaluation and on a limited budget, this method can cover only a small population and will thus be of indicative value only.

On the other hand, if it is desired to use it on a large scale, the limits of a macroevaluation are thereby surpassed, and the operation becomes one of research.

3.54.06 Reading time.

The speed of reading MT is measured easily during the evaluation of intelligibility. Its cost is thus unimportant, provided that measurement takes place at complete or partial text level, but not sentence by sentence. Measurement of intelligibility implies knowing what is in the text, and thus attentive reading, and therefore the reading speed of an evaluator, on this occasion, is meaningful and can be compared to that of a normal user.

Finally, any variation between the speed of reading MT and HT indicates the time loss suffered by the user because of the lower quality of the MT.

The reading time thus constitutes an efficient criterion for evaluation of one of the aspects of the quality of the translation : its effect on the performance of the reader.

One condition necessary for the measurement to be reliable is that the various versions (HT, MT with or without revision or post-editing) are presented to the evaluators in a homogeneous form : typed or retyped text, with capital letters, small letters and usual diacritic signs, without erasures nor handwritten additions.

3.54.07 Correction time

As in the case of the criterion “reading time”, the correction of MT and HT :

-is measured easily at the level of a whole text (but not sentence by sentence), by the reviser's and the editor's timing themselves

-measures a significant element : the time devoted their work by the reviser and the post-editor.

Correction time, of course, at least in part, is proportional to the scale (nature and number) of the corrections and has thus to be interpreted taking into account the correction rate (cf. microevaluation § 3.6).

3.54.08 Translation production time.

The translation production time is a question more of organisational factors (organization of the translation, revision and post-editing locations; response time by the computer centre to the requests for MT; organization and speed of the transmission circuits; saturation of the work-stations; queues, etc.) than of translation and correction time. This latter in general represents only a tiny part of the time which passes between a request for a translation and the supply of the corrected translation to the requester.

It is consequently barely realistic to consider the use of this criterion to evaluate translation quality (at the level of a market study, the problem is obviously different and translation production time then constitutes an important element in the quality of the service).

3.54.09 Linguistic_criteria.

The evaluation of translation quality on the basis of linguistic criteria is not within the context of the Commission of the European Communities; since it provides useful information, neither to the decision-makers, nor to the translators, nor to the users, nor even to those responsible for maintenance and for development for the system.

That does not mean, of course, that this type of evaluation may not be of use from other points of view.

3.54.10 Operational criteria.

The facility, for MT, of constituting a valid input to an automatic language recognition system, or of meeting a number of points in a specification, certainly constitutes an element by which its quality may be assessed, but unfortunately this element is rather vague and of relatively little interest.

Moreover, the cost of this type of evaluation can be high.

The low efficiency of these criteria consequently causes us to reject them.

3.6 Microevaluation -methods and criteria.

3.61 Introduction.

The microevaluation of a system consists in counting the errors made, in diagnosing the causes with a view to proposing remedies and, finally, in considering how the system can be improved.

Although there are far fewer microevaluation studies in the fields of MT, both in the literature and in practice than there are macroevaluation analyses, some authors stress the importance of them : H.W. SINAIKO, for example, recommends that the evaluations should aim at discovering why the MT system examined is inadequate. Similarly, B. VAUQUOIS recommends analysing the nature of the faults submitted to the reviser, discovering their cause and conducting an assessment of the extent to which simple modifications to the system would enable them to be avoided.

These authors'contributions to the microevaluation discussion can be classified into five groups according to the level of the analysis

- grammatical level : frequency of the errors corrected (by post-editing), by type of grammatical errors (morphological, syntactic, semantic, etc.)

- formal level : error frequency by type of corrections (deletion, addition, displacement, of words) made by the post-editors

- causal level : frequency of the errors corrected (by post-editing), by type of defective sub-functions within the translation system (input, source language analysis, transfer, etc.)

- remedy level (or improvability level) : analysis of the errors corrected (by post-editing) by the type of corrections to be carried out (theoretically) on the system to remedy the error (modification of the dictionaries, modification of various types of routines or instructions at the translation program level

- improvement level (noted) : analysis of the errors corrected (by post-editing), of the corrections (actually) carried out on the system and the improvements noted after modification of the MT system.

Clearly, these five levels range from surface level to deep level; the cause level, the remedy level and the improvement noted level are in effect the only ones which correspond to the definition of microevaluation.

We have nevertheless included the studies at the formal and grammatical levels in this chapter, since they also do not provide an overall evaluation of the quality of the MT, as is this case with all the other criteria listed in the chapter on the macroevaluation.

Predictably, the number of authors who have considered the microevaluation is inversely proportional to the depth of the level of the microevaluation.

3.62 Table of microevaluation methods.

NB. : the authors whose names are underlined are those whose methods have actually been used.

Method	authors
Grammatical level:"error" listing	ASSOCIATION JEAN FAVARD CHAUMIER GREEN KNOWLES MASTERMAN
Formal level : calculation of correction rate	CHAUMIER DEHAVEN VAN SLYPE
Causal level	VAN SLYPE VAUQUOIS
Remedy level	VAN SLYPE
Improvement level (noted) : extent of actual improvement (or dynamic analysis)	HALLIDAY PETIT VAUQOIS

3.63 Description of microevaluation methods.

3.631 Statement of the “errors”

3.631.1 Definition.

Analysis of the errors corrected by post-editing,
classified by grammatical type.

3.631.2 Evaluation methods.

3.631.21 L'Association Jean FAVARD.

* Definitions.

-Translation of the predicates with their agents, with the specification of the internal structure, government level, incidence level

-Translation of the composite noun phrases

-Translation of the constant, words and the expression

-Article

- Examination and identification of referents

* Application.

None.

3.631.22 J.CHAUMIER.

Below can be found the list of the elements analysed with their definition and the enumeration method.

These analyses were successfully completed during the evaluation of the SYSTRAN English-French system of the Commission of the European Communities.

Noun phrase.

* Definition.

The whole set of words (articles, adjectives) relating to a noun or a pronoun and constituting with it a function in the clause : subject, agent phrase, attribute, object phrase, adverbial phrase (of verbs or nouns); the groups are separated by verbs, verbal forms (participles), prepositions, conjunctions, punctuation marks.

* Method.

-Counting the number of noun in phrases in the sentence to be translated (N)

-Counting the number of noun phrases correctly shown in the translation (DEL)

-Counting the number of noun phrases whose internal order, or sequence of the constituent words, is correctly represented in the translation (OIC)

N.B.: The NPs forming part of verb phrases are included

-Calculation of ratios

DEL/N

OIC/N

Agent: subject and agent phrase (in the latter of the passive verbs)

* Definitions.

-“The subject, the starting point of the statement is the word or word group denoting the being or the thing the action or the state of which is stated”(question: who? What?)

-“The agent phrase of the passive verb denotes the being or the thing indicating the originator, the agent of the action that the subject suffers”it is introduced by the prepositions "by".

* Method.

-Counting the total number of subjects and agent complements in the source sentence (N)

- Counting the total number of subjects and agent complements agent correctly recognized as such the translation, and attached to the appropriate_verbs (REC).

N.B. : in the case of a number of verbs governed by the same subject, only the expressed subjects are counted

- Calculation of ratio REC/N.

Noun phrase and adjectival phrase.

* Definition.

- Noun, pronoun, infinitive, adverb or noun-dependent subordinate clause (or the pronoun, or adjective-dependent)which qualify the meaning of the noun, pronoun or adjective)

This includes the comparative : “taller than his father” and the appositions : “the State of Nigeria” or “Alaska peas”.

* Method.

- Counting the number of noun phrases and adjectival phrases in the sentence (N)

- Counting the number of noun phrases and adjectival phrases correctly recognized as such in the translation and attached to the appropriate nouns and adjectives (REC)

- Counting the correct order(OC)

- Calculation of ratios

REC/N

OC/N

Verbal phrase.

* Definition.

The whole set consisting of:

- root of the verb

- the inflexion of the verb

- the verb phrase(s)

- government of the verb.

* Method.

- Counting the number of verb phrases in the source sentence (N)

- Counting the number of verb phrases the tense which was correctly recognized (TPS)

- Counting the number of verb phrases the case of which is correctly governed (RCT)

- Calculation of ratios

TPSIN

RCT/N

Verb phrases (object and adverbial phrases).

* Definitions.

The direct object phrase is the word or word group joined to the verb without a preposition and completes the meaning by showing who or what suffers the action. It can be a noun, a pronoun, an infinitive, a clause, a word acting as a noun (question : which or what?)

The indirect object phrase is the word or word group joined to the verb by a preposition and completes the meaning by showing who or what suffers the action. It can be a noun, a pronoun, an infinitive, a clause, a word acting as noun (question to whom, to what, of which, of what, for whom, for what, against whom, against what ?)

The adverbial phrase is the word or word group which completes the idea expressed by the verb by indicating some external data on the action (time, place, cause, aim, etc.).
It, can be a noun, a pronoun, a word acting as a noun, an infinitive, an adverb, a gerund, a clause.

* Method.

- Counting the total number of object complements and adverbial complements in the sentence (N)

- Counting the total number of object complements and

- adverbial complements correctly recognized as such in the translation and linked to the appropriate verbs (REC)

- Calculation of ratio

REC/N

Attribute.

* Definition

- Word or word group expressing the quality, the nature, the state attributed to the subject or the object complement by means of a verb (to be, or verbs of state or certain verbs of action).

  The attribute can be a noun, a word acting as a noun, a pronoun, an adjective, an adverb, an in
  infinitive, a clause.

* Method.

- Counting the number of attributes in the sentence (N)

- Counting the number of attributes correctly recognized as such in the translation, and linked to the appropriate subject and object complements (REC).

  N.B. : when an adjective is involved, it is also counted under the heading "adjective"

- Calculation of the ratio

  REC/N

Verb.

* Definition.

  Word or word group which expresses the action, the existence or the state of the subject or even the link between the attribute and the subject.

* Method.

  - Counting the number of verb phrases in the sentence (N)

  - Counting the number of verb phrases whether or not translated, but correctly recognized as verbs in the translation, and linked to the appropriate clauses (REC)

  - Counting the number of verb phrases translated whether the conjugation (mood, tense, voice, person, number) is correct or not (T)

  - Counting the number of verb phrases translated with the correct conjugation (CT).

N.B.

1) The infinitives are counted as verbs

2) The present participles of verbs are counted as verbs and their complement as the COMPLEMENT of the verb

- Calculation of ratios

REC/N

T/N
CTIN

Negation.

* Method (actually used).

- Counting the number of negations in the text(N)

- Counting the number of negations correctly attached (REC)

- Counting the number of negations correctly trans1ated (CT)

- Calculation of ratios

REC/N

CT/N

Noun and noun phrase.

* Definition.

- Word or word groups used for indicating beings, things and ideas

* Method.

- Counting the number of nouns and noun phrases, in the sentence to be translated

- Counting the number of nouns and noun phrases, whether translated or not, but recognized as nouns in the translation and located in the appropriate clause

- Counting the number of nouns and noun phrases translated, whether or not the number is correct (T)

- Counting the number of nouns and noun phrases correctly translated, with correct number (CT)

- Calculation of ratios

REC/N
TIN
CT/N

Article.

* Definition.

- Word placed in front of the noun to show that this noun is understood in a fully or partially determined way. A distinction is made between the definite article, the indefinite article and the partitive article.

* Method.

- Counting the number of articles, present or suppressed in the sentence to be translated (including those in titles) (N)

- Counting the number of articles translated, whether or not the agreement (in gender and number), the elision, the contractions are correct (T)

- Counting the number of articles translated with correct agreement, elision and contraction (CT).

  N.B. : a zero English article, which should be and actually was left as a zero article in French, is considered as T and CT.

  An article translated and linked correctly to the noun used by SYSTRAN is considered correct even if the noun is wrong (lexical)

- Calculation of ratios

TIN
CT/N

Adjectives and adjectival phrases.

* Definition.

- Word or word groups linked to the noun (or pronoun) to qualify it or define it. A distinction is made between qualifying adjectives, numerical adjectives, possessive adjectives, demonstrative adjectives, relative adjectives, interrogative adjectives, indefinite adjectives and verbal adjectives

* Method.

- Counting the number of adjectives and adjectival phrases in the sentence to be translated(N)

- Counting the number of adjectives and adjectival phrases translated or not translated. But recognized as adjectives in the translation, and linked to the appropriate nouns and pronouns (REC)

- Counting the number of adjectives and adjectival phrases translated, whether the agreement (in gender and number) is correct or not (T)

- Counting the number of adjectives and adjectival phrases correctly translated and with correct agreement (CT).

N.B.

1) The cardinal number adjectives are constant words

2) Verbal adjective: present participle in agreement

- Calculation of ratios

REC/N
T/N
CT/N

Pronoun and pronoun phrase.

* Definition.

- Word or word groups which, in general, represents a noun, an adjective, an idea, a clause, ... or which plays the role of an unspecified noun.

A distinction is made between personal, possessive, demonstrative, relative, interrogative and indefinite pronouns.

* Method.

- Counting the number of pronouns and of pronoun phrases in the sentence to be translated (N)

- Counting the number of pronouns and pronoun phrases translated or not translated, but recognized as pronouns in the translation and linked to the an appropriate words, at the appropriate place (PEC)

- Counting the number of pronouns and pronoun phrases translated, whether the agreement (in gender and number) and the elision is correct or not (T)

- Counting the number of pronouns and pronoun phrases correctly translated with correct agreement and elision (CT)

- Calculation of ratios

  REC/N
  T/N
  CT/N

Preposition and conjunction.

* Definition.

- The preposition (or prepositional phrase) is an invariable word which is usually used to introduce a complement, which it links by a specific relation- ship to a supplemented word (example : with regard to)

- The conjunction (or conjunctional phrase) is an invariable word which is used to link and establish a relationship between either two clauses, or two words of same function in a clause. A distinction is made between coordinating and subordinating conjunctions (example : since, that).

* Method.

- Counting the number of prepositions and conjunctions present or omitted, in the sentence to be translated (N)

- Counting the number of prepositions and conjunctions translated and correctly positioned (CT)

- Calculation of ratio

  CT/N

Constant words.

* Definition.

- Any invariable word (proper noun, chemical symbol, abbreviation, figures, ...)

- Adverb and adverbial phrase : word or word. invariable group which is attached to a verb, an adjective or another adverb to modify its meaning. A distinction is made between adverbs of manner, quantity, time, place, affirmation, negation and doubt.

* Method.

- Counting the number of constant words in the sentence to be translated (N)

- Counting the number of constant words correctly translated and attached to the appropriate words (CT)

- Calculation of ratio

CT/N

Punctuation.

* Definition.

- All signs **= . ; ! ? () , ...**

* Method.

- Counting of the number of punctuation marks in the original text (N)

- Counting of the number of punctuation marks correctly transcribed (CT).

N.B. : Some punctuation is not necessary in English, but would have to be put in French. Here the same evaluation criteria were applied as for the suppressed article

- Calculation of the ratio : CT/N.

3.631.23 R. GREEN.

* Definitions.

- Structure errors words and expressions in order, incorrect attribution of adjectives, homograph errors, etc, the majority of these errors arise from incorrect analysis of the source text

- Preposition errors : prepositions translated wrongly

- Article errors : failure to restore the articles of the source language (especially English, where many nouns appear without articles) and partitive articles

- Errors in expressions : badly translated expressions

- Translation errors : nouns, verbs and adjectives incorrectly translated, with errors ranging from slight distortion of meaning to complete nonsense

- Miscellaneous errors : errors of number, misprints, superfluous words, foreign words treated as words of the source language, etc.

* Method.

Seriousness rating of the errors on a scale of 1-4 (except the last category - words not in dictionary -where errors are simply counted)

* Rating scale.

1.Very minor error, which does not affect the meaning and is more a matter of style

2.Definite error, but one which does not impair comprehension of the text

3.Error which leads to ambiguity

4.Serious error, which gives either the wrong meaning or no meaning at all.

* Objective.

- Indication of priorities regarding remedial action.

* Application

Internal evaluation of the SYSTRAN English-French system of the Commission of the European Communities.

3.671.24 F. KNOWLES.

* List of errors considered.

- Level of morphology (tense of verbs, etc.)
- Level of morphology and syntagmas
- Level of syntagmas (prepositions, apposition, etc.)
- Order of words and use of articles
- Ellipsis (omission of words in a sentence nevertheless remains comprehensible in the target language)
- Idiomatic expressions
- Syntactic garbage
- Comma placement
- Translation of names
- Semantic level
- Lexicographical level.

* Application.

Experimental, on a text translated from Russian into English by a version of SYSTRAN tested in Bonn.

3.631.25 M. MASTERMAN

* List of errors considered.

In her search for an evaluation program, for MT, M. MASTERMAN proposes that we consider five criteria(four of which are taken from an earlier study by I. RHODES).

- Words not translated
- Incorrect detection of the boundaries of units; by “translation units” : M. M.LSTERMAN means what some authors have called “the brain's Short term memory : units of +/- 7 words”, or “rhythmic punctuation”, or “rhythmic expression” which she proposes to analyse by studying the behaviour of translators, interpreters and teachers of rapid reading
- Expressions whose components were either not combined, or were wrongly combined, and words whose stems and affixes were not combined correctly

- Failed syntactic predictions

- Garbage: “a so-called translation which may or may not appear to be intelligible or acceptable, but which, in fact, leaves the user worse off than if he had been left only with the text in the source language - he not knowing the source language”

* Application.

None.

3.632 Calculation of the correction rate.

3.632.1 Definition.

Analysis of the errors corrected during post-editing, by type of correction.

3.632.2 Evaluation methods.

3.632.21 J. CHAUMIER.

* Definition and method.

- Counting the number of words corrected and comparison with the number of words translated, sentence by sentence.

* Counting rules.

- See next page.

* Application

First evaluation of the SYSTRAN English-French system for the Commission of the European Communities.

Symbol	Feature	Remarks
R	Number of words replaced	- If the replacement of a group of words, and if . the number of replacing words is equal to the number of words replaced, count the latter . the number of replacing words is different from the number of words replaced, count the number of words in the larger group
C	Number of words	- on or more corrections to the same word are to be counted as a single correction - Do not count corrections made to increase legibility of letters
D	Number of words	- If a group of words is moved, count the number of words in the group, as it is . after any deletions . before any additions - If two groups of words are reversed, count the number of words in the smaller group - If a corrected or replaced word is moved, count it also under D
E	Number of words deleted	- To avoid double counting, do not count here those words deleted for replacement or correction purposes
A	Number of words added	- To avoid double counting, do not count here those words added for replacement, correction or movement purposes

3.632.22 P.C.DEHAVEN.

* Definition.

- Analysis of the nature and the frequency of the lexical and syntactic changes made by MT post-editors, by type of syntactic role (or part of speech).

* List of corrections.

- Addition : addition of a word to the translation
- Substitution replacement of a word translated
- Translation replacement, of a source language word not translated by the system, by the appropriate word in the target language
- Deletion : deletion of a word translated
- Suffix: addition, deletion or replacement of the suffix of a word translated
- Capitalization : replacement of a lower-case initial by a capital letter
- Rearrangement : alteration of the position of a word
- Punctuation: addition, deletion or replacement of a punctuation mark.

* List of syntactic roles.

Noun, pronoun, verb, adjective, adverb, preposition, conjunction.

* Application.

Evaluation of the SYSTRAN Russian-English system.

3.632.23 G.VAN SLYPE.

* Definition, method and rules.

Same as J. CHAUMIER.

* Application.

Second evaluation of the SYSTRAN English-French system of the Commission of the European Communities.

3.633 Analysis of causes.

3.633.1 Definition

Analysis of errors corrected during post-editing, by type of causes of errors.

3.633.2 Evaluation methods.

3.633.21 G.VAN SLYPE.

* Method.

- Examination of each post-edited MT sentence, by an evaluator, and counting of the total number of corrections
- Examination of each post-edited MT sentence, by a SYSTRAN specialist
- Identification of probable cause of each of the errors corrected by the post-editor
- Aggregate for the sample
- Calculation of the percentages of errors for each probable cause, compared with the total number of corrections.

* Analysis_grid.

- Number of post-editing changes due to the source text
    - ambiguity
    - incorrect or clumsy style
    - syntactic error
    - spelling mistakes
- Number of input errors
- Number of post-editing chances due to the translation system :
    - dictionary
    - analysis
    - synthesis
    - miscellaneous

- Number of post-editing changes attributable to personal factors :
    - post-editor's stylistic preferences
    - post-editing error

* Application

Second evaluation of the SYSTRAN English-French system of the Commission of the European Communities.

3.633.22 B. VAUQUOIS.

Analysis-grid for sources of difficulties.

-None. The sentence were analysed and generated according to the rules of the various models and yielded the result expected. These results do not necessarily give an "excellent" translation since the models are only approximations. But the sentences thus obtained are very comprehensible

-Exceeding capacity during syntactic analysis

-Errors in the input text

-Errors detected in the coding of words in the dictionary; errors in the tests applied to the grammar rules;(i.e. errors which can be corrected)

-Errors whose origin is dubious or unknown

-Difficulties that are practically insurmountable with the current models.

* Application.

Evaluation of the first Grenoble Russian-French system.

3.634 Improvability.

3.634.1 Definition.

Analysis of errors in terms of the type of remedy required.

3.6,74.2 Evaluation method.

This method is discussed by one author only, G. VAN SLYPE; however it has probably been applied by the majority of manufacturers of MT systems in efforts at improvement.

* Method.

- Examination, by one or more evaluators, of MT sentences not post-edited, and intelligibility rating
- Pinpointing of sentences with intelligibility below 50%
- Examination of these sentences by a translation system specialist to diagnose the main errors (those which if corrected would raise the sentence intelligibility to above 50 %) and to define remedies, predict secondary effects and estimate the time necessary for applying the remedy to the system
- Breakdown of the sentences according to the remedy required
- Calculation of the time required and the number of sentences requiring each type of remedy
- Calculation of percentages.

* Application.

Second evaluation of the SYSTRAN English-French system of the Commission of the European Communities.

Remark :

Since it was impossible to establish a precise
typology of errors, the method was applied only up to the third stage (inclusive) of the method described above: diagnosis of errors, remedies, secondary effects and time required for remedying.

3.635 Measurement of actual improvements, or dynamic analysis.

3.635.1 Definition.

Analysis of the improvements noted, after modification of the system to avoid the errors detected during a post-editing carried out after first MT, or a first series of MT.

3.635.2 Evaluation methods.

3.635.21 T.C. HALLIDAY.

* Method.

-Breakdown of translation system into n subsystems capable of improvement (SYSTRAN example : stem dictionary; expressions dictionary**,** lexical routines, syntactic routines)

-Preparation of a fairly large double sample (2 x 50,000 words per field):

- sample A : control sample, used as source of errors to be corrected in each of the subsystems
- sample B : sample used to measure the effects of the corrections to the subsystems

-Submission of sample A to evaluators with knowledge of source language, target language and translation system, with a view to :

- indicating all translation errors
- attributing each of these errors to one of the subsystems in the system

-Submission of these lists of errors to the system linguists and coders for correction in each subsystem of all errors that can possibly be corrected on the sole basis of the errors arising from sample A

Preparation of (n + 1) versions of the MT system

1) initial system, before the rest

2) the system in which only one subsystem (for example: stem dictionary, as been modified to integrate the above mentioned corrections and the other subsystems have remained in their original form

3) the system in which a second subsystem (for example : expressions dictionary, in addition to the stem dictionary) has been corrected

n + 1) the system in which all the subsystems have been corrected

-Machine translation of the two sample A and B, in succession with the versions 1, 2, ... n + 1 of the MT system

-Comparative evaluation of version 1 translations, on the one hand, and version 2 to (n + 1) translations, on the other hand, to determine the number of sentences where translation has been improved and the percentage of these sentences in each of the samples A and B

-Second run:

- list of errors for each of the translations 2 to (n + 1) of sample B
- correction of subsystems
- preparation of (n + 1) new versions of the MT system
- machine translation of versions 1 to (n+1)
- comparative evaluation and calculation of the percentage of sentences improved in the various versions of samples B and A

- Third run, based on the list of errors

-Determination of the improvement rate

-Continuation of runs until there is no further improvement, or until the corrections made all a certain place lead to new errors at other places and the quality of MT remains the same or even decreases.

* Application.

The method of calculating the improvability of the translation system was hardly applied in T.C. HALLIDAY's evaluation of the Russian-English SYSTRAN'system

- only the first run was carried out, which does not permit determination of the improvement rate
- only the subsystems "stem dictionary" and "expressions dictionary" underwent correction.

Summary of the results.

Sample A (100,000 words) provided 6,4 % errors (6,400 errors), distributed over :

- 1,6 % errors capable of being eliminated by a correction of the stern dictionary
- 2,7 % errors attributable to the expressions dictionary
- 0,2 % errors attributable to the lexical routines
- 0,8 % errors attributable to the syntactic routines
- 0,3 % errors attributable to the input,
- 0,2 % errors impossible to correct in the current design of SYSTRAN
- 0,6 % errors due to unknown causes.

Compared to the initial version of the system *, the correction of the stem dictionary alone permitted an improvement of :

- 50 % of the sentences of sample A
- 40 % of the sentences of sample B.

* Note from the author : The evaluation method used, which consisted of establishing the number of sentences where the quality was improved without rating the quality or quantifying the improvement, provides only general information on improvement : we know that there has been improvement, but not how much.

Compared to the initial version of the system, correction of the stem dictionary and the expressions dictionary permitted an improvement of:

-56% of the sentences in sample A

-41% of the sentences in sample B,

i.e. a gain due to the correction of the expressions dictionary of :

- 6 % for sample A

- 1 % for sample B.

N.B. : The key figures which measures the improvement of the system are obviously those of sample B, since sample A errors were used to correct the dictionaries; it is consequently to be expected that the percentage of sentences actually improved in sample A should be higher than in sample B.

3.635.22 A.J.PETIT.

* Introduction

An evaluation is complete only if it can determine the real qualities, i.e. the general performance aptitude and the possibilities of -improvement, without being limited to an isolated performance.

An MT system can be subdivided into functions and sub-functions which reflect the translation method and, fortunately, evolve in a given order which progresses with increasing difficulties. Each of these functions or sub-functions will be made to correspond with one or more evaluation criteria from which it will be possible to establish the effects of the finished product.

Morphological criteria.

* Definition.

The morphological function of a system is a mainly mechanical function which consists in identifying the words of the character string and, referring to dictionary and the morphological resolution rules, singling out the words which have to be added to the dictionary to make translation possible.

This amounts to a first reading during which the translator underlines the words he does not know; he will obviously recognize proper names, references and what does not have to be translated.

Criterion 1 : The list of unknown words does not have to include proper names, references or any other elements which, generally do not have to be translated.
If the list includes these it will be impossible to recognize them and these terms inevitably be translated when they correspond to an entry in the dictionary. The only possible improvement is a basic change of system. Any solution consisting of adding entries to the dictionary is inadmissible.
It is thus an eliminating criterion.

Criterion 2 : The punctuation and the brackets have to be suitable identified for the purposes of syntactic analysis. Any need to readjust the brackets in the translated text is an inherent defect in the system which distorts the translation.
It is thus an eliminating criterion.

Criterion 3 : The presence, in the list of unknown words, of inflected verb forms or plurals indicates the elementary nature of morphological treatment. This is not an eliminating criterion, although it is necessary to enter all forms of a word into the dictionary, irrespective of what it entails

- additional coding

- multiplication of entries.

The quality of the translation is not compromised by this factor alone.

* Method.

First of all the translation is put through the machine in order to obtain the list of unknown words.

Using the list of unknown words:

1) record an error for the following cases

   - proper name, inscription
   - essential terms
   - inflected form of an essential term.

   Do not add these terms to the dictionary and leave the sentences in the test batch

2) single out all the sentences and phrases containing uncommon words.

   Count them and take them out of the test without deducting them from the initial total.

Machine translation.

Check the morphology sentence by sentence.

Record an error in the following cases

- proper name translated
- inscription translated
- reference translated
- unwarranted insertion of brackets.

Check test. When the proper names and inscriptions or references have been correctly inserted, check whether they are in the dictionary; if they are, replace them by terms which are not in the dictionary and put the offending sentences through the machine. Total the number of errors.

Syntactic criteria.

* Definition.

Syntactic analysis consists of determining the exact role each word in the sentence on the basis of the possibilities described in the dictionary.
The problems are as follows : for a word which can have several functions (noun, verb, adjective, for example it is necessary to determine its real function (homographs).
It is then a question of determining the relations within the sentence : for example, to know whether the noun is the subject or predicate (analysis).

For a machine translation system this is a major problem and homographs are probably the first natural enemies of MT In the majority of cases the which cause such concern to the machine are easily solved an intelligent reader or one who knows the subject but there are natural ambiguities can even the best translator. It will be shown later how one can identify hopeless ambiguities, but, with the exception of these special cases, a machine translation system has to be able to solve the homographs since any error produces an absurd sentence in the translation which has to be completely retranslated.

Criterion 1 : Solution of homographs problem in general : eliminating criterion.

As the solution of homographs in an MT system depends only on its analysis capacity, local adjustments cannot have a lasting effect. This characteristic can be improved only by a radical revision of the system.

Criterion 2 : Homographs relating to a specialized field with the same requirements as homographs in general.

* Method.

Check the syntax sentence by sentence.

Record an error in the following cases

- nouns
- verbs
- prepositions
- isolated forms ending in ing.

Semantic criteria.

* Definition.

Since a word can have several, sometimes very different, even opposite meaning, it must be possible to find the right meaning for the given context: otherwise the result is mistranslation, i.e. failure.

Contrary to widespread and carefully preserved belief polysemy is not simply a question of terminology.

Criterion 1 : Since a word can take on several meaning which are already established (which can thus be entered in the dictionary), each of these meanings has to be identified when the word appears in a context which allows the reader for whom the text is intended to state categorically what is meant.

Criterion 2 : This requirement also applies when the word appears out of context where the reader for whom the text is intended can easily identify the correct meaning.

N.B. : This applies only to cases where syntactic analysis cannot differentiate between meanings. This criterion indicates the possibility of the system's using any form of knowledge whatsoever.

All semantic analysis errors which follow are eliminatory in that the sentence must retranslated and more especially they give rise to mis-translations which, in certain cases, might escape the reviser's attention.

-Errors concerning non-technical critical words where translation is essential for the comprehension of the sentence; these are mainly verbs and abstract nouns

-Errors concerning basic technical words (which can appear alone or in combination) whose meaning can vary within the same field, irrespective of the definition given to “field”

-Failure to attach an isolated word to the complete expression in the preceding sentence.

* Method.

Check the semantics sentence by sentence.

Record an error in the following cases errors due to polysemy:

- common verbs
- common abstract nouns
- technical words.

Check test.

Mark all the successes and check against the dictionary.

1) Whether the successful translation correspond to a lexical routine

2) Whether only the meaning in that context is in the dictionary.

If 1) or 2) applies, replace the subject and predicate by equivalent terms and add the errors to the preceding list.

Reflections on noun clusters.

* Definition.

By "noun clusters" we mean those interminable technical expressions which seen to defy all laws. In a standard technical text there are, on average, at least two clusters per sentence and each cluster presents a problem. Cluster translation errors have serious repercussions since they cover an essential point of the content. Also, correction of these errors requires a good deal of effort and research on the part of the reviser. It is thus logical to consider this criterion as eliminatory.

* Method.

Check the noun clusters sentence by sentence.

Record an error for the following cases : errors of analysis

-error of translation due to polysemy (when all the elements of the cluster are in the dictionary)

-interference of idioms included

-conjunctions.

Check-test. Mark all the successes and check against the expressions dictionary to see whether the cluster is included.

If so, modify it by changing or adding a term.

Add all the errors to the preceding list.

Transfer and generation criteria.

* Definition.

After syntactic analysis and semantic analysis we know the role of the word and its meaning.
Using this information, transfer consists of
syntactic and lexical changes designed to put the text into the target language.

Criterion 1 : The system has to be able to make syntactic changes which are normally indispensable for easy comprehension of the text, especially where the result would be gibberish without these changes.

Criterion 2 : Idiomatic expressions have to be respected if necessary.

Criterion 3 : All the agreement and punctuation rules, etc., of the target language have to be respected.

The faults which can be attributed directly to transfer and generation are limited. The most likely ones in translation from English to French concern the auxiliaries, interrogative and negative forms and the use of articles.

Questions of transfer and generation are handled
by putting into a "miscellaneous bag" category all the errors which could not be directly attributed to one of the preceding eliminating criteria and
by considering them as the only acceptable field of error in a machine translation system.
This comes within the normal scope of revision.
This criterion can thus act as a subsidiary criterion designed to establish the quality of two acceptable translations, but it will not be of any use at the acceptance stage, since it will be automatically subjective and must not interfere with the acceptance process.

* Method.

Take the sentences eliminated from the test after an examination of the 1ist of unknown words (sentences with uncommon words).

- Enter words not in the dictionary and put the sentences through the machine

- Make all the checks on this batch and add the errors by category so as to obtain 'the total number of errors with each item.

Correction test.

Make a list, of all the errors detected during the preceding tests and pass it on for correction to the personnel of the system designer. Note carefully all the corrective measures taken

- coding modifications

- machine instructions (have them explained if necessary).

AIL the factors will have to be monitored during this operation.

Run a second machine translation and compare it with the first.

Count the number of errors per item.

If the results of this test are satisfactory (no errors) that simply proves that the systems' output can be altered locally, i.e. that the system can reproduce an item of information that it has just been provided; it is thus not a correction, but the reproduction of a correction.

In the case of an error all comments are to no avail.

Finishing test.

Correct all the evaluated errors on the text of the first translation, changing the text as little as possible. Have a clean copy typed of the corrected text and give it to a reviser. All the words corrected should be underlined, and the reviser should avoid changing them or the order of the words.
This work will be done in the presence of evaluator to whom the reviser must, justify verbally all his corrections while the evaluators should record his reactions.

The reviser will also Give a personal written assessment, as honest as possible, on the text as a whole.

There will obviously be no justification for this test if the text (system) does not satisfy the acceptance tests. It is given here only as a guide and was conceived as a supplementary factor to ensure that the evaluation concerns primarily objective criteria. Once a system passes the acceptance threshold, a sophisticated form of evaluation method has to be devised.

A method such as that used by DICAL at the Canadian Government translation office in Ottawa could be used as a base especially in a simplified version. In the meantime we will have to be content to measure the correction time and the reviser's level of enthusiasm.

3.635.23 B.VAUQUOIS.

B. Vauquois suggests evaluation of the qualities (at the macroevaluation level) of a translation system, before and after dictionary updating.

Such as “dynamic analysis” would include the following stages:

-preparation of a machine translation system (grammar, dictionaries) in a specific field

-machine translation of sample of texts within this field evaluation of the quality of the translation

-dictionary updating so as to correct errors detected, (but without causing a deterioration elsewhere in the system) machine translation of the same sample of texts

-evaluation of the Quality

-comparison of the quality before and after dictionary updating.

Application : evaluation of the first SYSTRAN English-French version supplied to the Commission.

3.64 Critical assessment.

As in the macroevaluation, our aim here is to assess the efficiency (cost/effectiveness ratio) of the various microevaluation methods.

1.641 Listing of “errors”

This criterion is completely inoperative : semantic and syntactic categories in linguistics do not correspond to the sub-functions carried out by the translation programme (in the case of the SYSTRAN programme at any rate) : these categories are taken into account either by the dictionary, by special routines on words or word groups, or by general subroutines. Thus a simple diagnosis of the erroneous translation of a specific grammatical category is not sufficient to indicate

To determine the remedy, it is necessary to identify the subfunction of the system in question; it is not necessary to specify the relevant grammatical category.

3.642 Calculation of the correction rate.

As indicated above, we are not strictly speaking concerned with criteria of microevaluation of the translation system since they do not enable us to determine how the system can be improved.

They are interesting because they enable us to define the tasks which the post-editor must fulfil in order to improve MT quality . They make possible a diagnosis at the symptomatic rather than at the causal level; microevaluation relates to the correction system and not to the translation system.

it is essentia1 to know their value in order to assess correction time (macroevaluation criterion) in the light of the intelligibility of the translation : experience shows how these three elements are interrelated

-for texts of similar intelligibility correction rates can vary greatly. Correction time is proportional to the correction rate

-it is possible to reduce markedly the and thus also the correction time, without intelligibility to the same extent.
This final point has merely been noted but up to now has not been investigated thoroughly
Such study would be very interesting, because it should lead to post-editing rules in which correction method would be optimized.

Formally, microevaluation is situated on two levels:

- global level : total number of corrections
- analytical level : number of corrections by type (words added, deleted, shifted, etc.).

The global correction rate can be measured without difficulty and gives an idea of the task facing. the post-editor; the correction rate as between systems or versions of a system, or between MT and HT can easily be compared.

Measurement of the analytical correction rate is more costly and gives us a better idea of the nature of the post-editor's basic tasks.

However it is of interest only if the aim is:

-to analyse the work involved in studying the possibility of correction with the aid of text processing equipment rather than a manuscript

-to optimize the correction method (cf. above).

Apart from this its value is more anecdotal than practical.

3.643 Analysis of the causes.

Even if it is not possible to detect the remedies, the diagnosis of how the errors corrected at the post-editing stage come about, provides nevertheless a first approximation of a mico-evaluation because it directs the search for underlying causes and the for appropriate remedies towards the large functions of the system which really matter.

However, this procedure is costly and not very efficient since it does not permit one to diagnose the causes with an accuracy sufficient to propose adequate remedies.

Consequently, it would seem useful to use it in exceptional cases, so as to obtain an overall view of the functions to be improved.

3.644 Improvability.

The patterns which have emerged in this type of microevaluation, as reported in the literature, are particularly interesting, since one can both attempt to develop remedy categories and to estimate the resource (in terms linguist and coder time) to realize these remedies : if it proved possible to determine the remedies and their cost, one would obtain useful data for developing a genuine strategy for improving the MT system with a view to optimizing the quality of the translation while minimizing the cost of improvement.

3.645 Measurement of the improvements made.

The method partially applied by T.C. HALLIDAY and proposed by A.J. PETIT is extremely attractive since it involves exploiting the MT system to its limits.

Unfortunately it is very expensive and this probably explains why T.C. HALLIDAY was not able to carry it through success

Moreover, it is of doubtful reliability : it relates in effect to a closed universe, that of the two selected samples. Although it is specified that the effects of the improvements of the system are analysed on one of the samples, whereas these improvements are realized on the basis of the translation errors recorded in the other sample, the series of successive runs mean in the end that one is working with a finite number of meanings for each word and grammatical construction.

it might be otherwise if, after each improvement of the system, after the initial one, an additional sample was added to be translated automatically. But this would lead to yet a further increase in evaluation costs

In fact, here we are approaching the limits of microevaluation, which shades into maintenance and the continuous improvement, of the system.

The method proposed by B. VAUQUOIS has the same drawbacks but, on the other hand, has the advantage that it can be implemented relatively cheaply (double work of machine translation, post-editing and quality evaluation: before and after updating of the dictionaries alone).

3.7 Sampling.

5.71 introduction.

A large number of authors have treated the problem of sampling in MT evaluation.

Their contributions can be classified in two large groups text sampling and evaluator sampling.

In the first group (text sampling) there are two large classes contributions on sampling methods (bench mark or random sample) and those which relate to the size of the text sample.

To clarify these contributions, we have, following this survey of the authors' contributions, drawn up a table showing the number of evaluators and the volume of the texts to which the evaluations relate wherever these data are available.

3.72 Table of contributions on sampling.

CONTRIBUTIONS	AUTHORS
* Sampling of texts	
-Sampling methods	
• Bench mark	MASTERMAN PANKOWICZ PETIT ZEMB
• Random selection	CARROLL LENDERS SINAIKO VAN SLYPE
-Sample size	LEICK SAGER
* Sampling of users	CARROLL HALLIDAY JOHNSON LENDERS SINAIKO SPILLECOUDT VAN SLYPE
* Table of characteristics for sampling the different translation evaluations	-

3.7 Description Of sampling methods

3.771 Text sampling.

3.731.1 Sampling method.

3.7311.11 Bench mark.

3.7711.11.1 M. MASTERMAN thinks that in future there will be two sampling strategies

-preparation of a sample of "control texts" by random selection from among all the documents which are of relevance to the translation systems

-inclusion of a sequence of simple individual criteria with which all the characteristics of the translation can be evaluated in a machine programme; this programme will be used both to assess and to improve any body of material whatsoever that has been machine-translated.

This second strategy is more difficult to apply, but is certain to become gradually more effective as our understanding of the nature of translation improves.

3.731.11.2 Z.L.PANKOWICZ notes that the results of evaluation of samples of several thousands or tens of thousands of text words are necessarily fragmentary : it is obvious that there is no guarantee that even the bulkiest sample will include all the possible syntactic structures of the source language.

Therefore he proposes a completely different approach. Rather than prepare a random sentence sample which will involve testing only certain grammatical rules of the MT system, he recommends drawing up a complete list of all the grammatical rules of the source language, from the simplest to the most complex, and choosing 20 to 25 sentences in which each of these rules is activated.

A sample drawn up in this way would enable one to test the performance of the system and to prepare a complete list of its gaps.

3.7311.11.3 A.J.-PETIT estimates that the evaluation tests would have to cover 350 to 500 sentences belonging to the kind which the system claims to treat; they should be taken from real texts. No entry in the dictionary would be permitted except when this was called for by the testing method. All the dictionaries would have to be submitted to the evaluators before the test.

All the real difficulties used to test the system would have to be present in the test text.

Each time a sentence contained a fault in respect of the evaluation criterion, an error would be recorded. The number of errors would be established separately for each of the criteria (accordingly, a sentence could present an error for each of the criteria). Wherever a check-test is indicated a success could be recorded only if the check-test was conclusive.

All the error percentages are established by reference to the initial number of sentences in the test batch.

3.731.11.4 J.B. ZEMB estimates that the text sample to be tested should drawn up in vitro, preparing it with the aid of sentences whose structure and syntax would become increasingly complex.

3.731.12 Random choice.

3.731.127.11 J.B.-CAPROLL in his study for the ALPAC Committee took his sample from five different passages in a Russian work. (Machine and Thought).

36 sentences were taken from each of the five passages.

The first of these passages was used to train the evaluators to use the rating grill.

The four other passages were used for the
Evaluation itself, i.e. a total of 144 sentences (no information on length of sentences).

These 144 sentences were mixed at random, so as not to provide them to the evaluators in their normal sequence.

The analysis of the evaluation results showed that the evaluation ratings did not vary greatly from, one passage to another, but differed considerably from one evaluator to another.

J.B. CAPROLL concludes that a sample should contain a “considerable number” of sentences.

3.731.12.2 W. LENDERS, in his evaluation of the Russian-English SYSTRAN prepared a text sample made up of mixtures containing 0%, 33 %, 67 % and 100% of machine-translated passages whereas the rest were human translations.

3.731.12.3 H.W. SINAIKO proceeded as follows

-random selection of documents taken from a large corpus provided by the translation service in order to be sure that the test was based on a typical selection of texts rather than on specially prepared ones

-these texts have to be similar to those which are normally translated in the services of the purchaser of the MT system

-these texts have to include various literary styles statements, abstracts, quantitative data, illustrations, etc.

-there may be good reason to prepare texts which have been intentionally distorted to see if these distortions can be detected.

3.731.12.4 G. VAN SLYPE points out that for a certain number of criteria to be analysed, the sample must consist of complete texts comprising several sentences and not sentences extracted at random from many different texts : the intelligibility of a sentence, for example, can be assessed only if this sentence is in its proper context.

One might contemplate taking only the first sentences of the texts but at the risk of introducing a bias : it seems that the beginning of a text is always simpler and more understandable than the remainder.

A sample of 500 sentences would seem to be sufficient.

3.731.2 Dimension of the samples.

3.731.21 J.M. LEICK calculated the size of the sample necessary to estimate two of the quantitative characteristics of MT within a margin of +/- 0.25 with coefficient of confidence equal to 0.95

- fidelity : n = 173 sentences

- post-editing rate n = 320 sentences.

However, these figures are valid only if the sentences are taken from a batch of very homogeneous documents.

3.731.22 In J.C.SAGER's view, a minimum of 25,000 words is needed when, rather than having to ascertain whether MT is valid or not, the scope offered by MT has to be determined.

3.732 Sampling of evaluators.

3.732.1 J.B. CARROLL in his study for the ALPAC Committee, used :

- 18 monolingual (English)evaluator, undergrates reading sciences, who split up the 144 sentences of the sample between them (in 6 different versions : three human translations and three machine-translations): i.e. 48 different sentences per evaluator, and sentences per version per evaluator

- 18 bilingual (English-Russian) evaluators, having the same educational level, for the same sample.

On the basis of his evaluation, J.B. CARROLL :

- considers it is preferable to use (target language) monolingual evaluators as they are more representative of real users and are not influenced by their knowledge of the source language

- noted that ratings vary little from one evaluator to another, but that the variation is nevertheless sufficient to warrant the use of, at least, 3 or 4 evaluators.

3.732.2. T.C. HALLIDAY's view is that the evaluation of MT has to be based on the theoretical potential of the system, because :

- all data-processing systems are subject to certain limitations inherent in computer design

- my system of translation of natural language contains a certain number of linguistic limitations resulting from the system's design parameters.

Assessment of MT should therefore be carried out by expert who :

- know both source language and target language

- have specialized knowledge enabling them to judge the technical accuracy of the translation

- are sufficiently well acquainted with MT to evaluate the translation taking due account of the system's potential and limitations.

N.B. : the reason for this choice is that T.C. HALLIDAY's appraisal was a microevaluation.

7.732.3 According to R.L.JOHNSON, the most valid, but not necessarily the most reliable, qualitative appraisal will undoubtedly come from user opinion on the adequacy of a translation for a specific application.

This is incompatible with cost minimization which requires that experts he called upon as infrequently as possible for the evaluation; text selection and translation evaluation must therefore be based, as far as possible, on external form rather than content.

Furthermore, the objective criteria for text selection and evaluation will probably be less biased and more reliable respectively but not necessarily more valid.

3.732.4 W. LENDERS feels that comparative evaluation of MT and HT cannot be left to the final user, who will obviously prefer HT every time.

Whence the need for an external observer, who will use a method of investigation and a list of evaluation criteria.

3.732.5 H.W. SINAIKO recommends the following criteria for selecting evaluators:

-the reader-evaluators should be as similar as possible to the usual reader-users of the translated texts

-persons having a financial interest in the MT system to be evaluated must be excluded from all aspects of evaluation.

3.732.6 R. SPILLEBOUDT notes that the severity of evaluators is variable, and can bias evaluations.

It is therefore essential that, quality of translation be appraised by several evaluators.

The homogeneousness of evaluators should be Determined as follows:

Take a sample of n sentences evaluated by m evaluators, with N = n x m (total number of ratings)
Each evaluator gives each sentence an intelligibility rating of X*ij; X bar is the overall* average rating obtained; X*j* bar is the average of the ratings given by all the sentences.

The following table is compiled.

Source of variance	Sum of squares	Range of factors	Mean square
Between evaluators	$S_1 = n \sum_{j=1}^{m} (\bar{x}_j - \bar{\bar{x}})^2$	$m - 1$	$\sigma_1^2 = \frac{S_1}{m-1}$
Between sentences	$S_2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_j)^2$	$N - m$	$\sigma_2^2 = \frac{S_2}{N-m}$

The ratio F= δ1 x δ1 / δ2 x δ2 is then calculated.

The value found for F is then compared to the extreme value found in the table, for the leeway value shown.

If the variation calculated is lower than or equal to the extreme value found in the table,, then the two variances compared are equal and the evaluators are homogeneous.

If it is not, the population of evaluators is heterogeneous.

3.732.7 G.VAN SLYPE points out that some of the criteria applicable to an evaluation of translation quality are objective, e.g. the number of errors of grammatical agreements. The others are more subjective and originate in:

-either the reaction the text elicits in the e.g. the number of corrections made by a reviser each reviser makes corrections which differ in kind and in number on the same basic text

- or the observer's attitude to the text; e.g. his judgement of the of a sentence.

This subjectivity should be eliminated by using not just one, but a team of evaluators and revisers and by statistically arriving at an average evaluation.

This first evaluation of the English-French SYSTRAN system of the Commission of the European Communities covered 506 sentences and 11,200 words.

The evaluation of the intelligibility of the various versions of the documents was carried out by a single person (a socio-psychologist).

The reliability of this evaluation was checked by asking 15 persons to evaluate separately 10 sentences selected from the sample of 506 sentences.

The average intelligibility rating for all 10 sentences was lower than the average rating given by the principal evaluator for the same 10 sentences, with a variation compared to the principal or evaluator

-less than 7% for the original version, the human translation and the post-edited version

-33% for the machine translations.

This comparatively wide variation was put down mainly

- to over-familiarity : the principal evaluator had to evaluate 500 + 10 sentences, whereas the 15 evaluators used in order to check the reliability of evaluation had only 10 sentences each to deal with

- to the fact that the 15 evaluators had to examine their 10 sentences out of context, which was not the case for the principal evaluator.

The method used nevertheless left open the question of statistical reliability.

in his second evaluation (improved version of English-French SYSTRAN), G.VAN SLYPE sought to strike a reasonable balance between the cost of the evaluation and the statistical reliability of the measurements both of which increase proportionately with the extent of the scruting.

For that, a two-stage evaluation was carried Out:

-First stage appraisal of the homogeneousness of the evaluators the two specimen documents, one a human translation, and the other a machine translation, were submitted individually to 9 evaluators in order to assess statistical reliability according to the method developed by R. SPILLEBOUDT

-Second stage : appraisal of the intelligibility of the sample sentences : in order to limit work load, the sample sentences are shared among the evaluators, instead of being submitted en bloc to each and every evaluators for his ratings.

The average rating for the whole sample is obtained by simply by adding the ratings given and dividing the total by the number of sentences : the result is not biased since the evaluators each perform an equal share of the work on the sample texts.

3.733 Table of the sampling characteristics of various translation evaluation

Authors	Systems evaluated. Evaluation criteria	Samples Texts	Samples Evaluators
J.B. CAROLL	several systems (ALPAC report 1966) -intelligibility -reliability	144 sentences	36 science undergra-duates
J. CHAUMIER	English-French SYSTRAN (1976) -intelligibility -listing of errors -calculation of correction rate	506 sentences 11.200 words (field: food science and technology)	1 socio-psychologist
R.C.DEHAVEN	Russian-English SYSTRAN (1972) -calculation of correction rate -intelligibility	24 texts 50.000 words (16 fields) 12 texts (1 field)	- 12 documentalists
D.H. DOSTERT	Georgetown Russian-English system (1972) -acceptability		57 end users
M.GREEN	English-French SYTRAN (1977) -listing of errors	40 abstracts (field: food science and technology)	2 linguistic coders
T.C. HALLIDAY	Russian-English SYSTRAN (1976) -improvability	200.000 words (2 fields)	linguists and systems analysts

Authors	Systems evaluated Evaluation criteria	Samples Texts	Samples Evaluators
E. HOFFMANN	English-French SYSTRAN (1978) -listing of errors	193 sentences (fields: automobile and food)	
F.KNOWLES	Russian-English SYSTRAN (1978)	76 sentences 1.744 words	-
A.W. LEAVIT	Russian-English SYSTRAN and MARK II (1970) -comprehensibility -ASTUTE	 36 articles (3 fields) 20 articles (2 fields)	 36 undergraduates in the 3 relevant fields 13 technical analysts, in the two relevant fields
J.M. LEICK	English-French SYSTRAN (1978) -calculation of the correction rate -fidelity French-English SYSTRAN (1978) -calculation of the correction rate -fidelity	388 sentences (fields : food) 271 sentences (fields: mechanical engineering)	6 linguists coders 6 linguists coders
W.LENDERS	Russian-English SYSTRAN (1971)	-	80 German students of English philology

Auteurs	Systèmes évalués Critères d'évaluation	Echantillons Textes	Echantillons Textes
W.H.SINAIKO	LOGOS anglais-vietnamien -test de performance (1970)	1 texte 1.000 mots (domaine: manuel de maintenance d'hélicoptère)	24 équipes de 3 techniciens vietnamiens et 6 équipes de 3 techniciens américains
	-test de connaissance (1970)	1 texte 2.400 mots (domaine: manuel de maintenance) 10 questions	68 techniciens de maintenance vietnamiens 2 évaluateurs
	-traduction en retour (1970)	3 textes 9.558 mots domaine: manuel de maintenance	88 étudiants pilotes américains 172 étudiants pilotes vietnamiens
	-test de Cloze (1972)	1.617 mots 1 mot éliminé sur 5	58 étudiants officiers de marine américains; étudiants officiers de marine vietnamiens
	-multicritères (1973) (test de connaissance; Cloze; intelligibility)	500 mots	
G.VAN SLYPE	SYSTRAN anglais-français (1978)	20 textes (domaine: administratif; technique et économique en matière agricole et alimentaire)	3 consultants en managment 1 linguiste et 5 analystes-documentalistes
	-intelligibility	656 phrases 2.303 mots	1 consultant en management
	-fidélité		2 consultants en management
	calcul du taux de correction		

Authors	Systems evaluated Evaluation criteria	Samples Texts	Samples texts
G.VAN SLYPE (continued)	-reading time -causes of errors -remedies -acceptability		same team as for the intelligibility evaluation 1 linguist (system specialist) 2 linguists (system specialists) 18 end users
B. VAUQUOIS	Grenoble Russian-French system (1971) -intelligibility -causes of errors	4 texts 15.000 words	

3.74 Assessment.

The sampling method is one of the major problems to be solved when drawing up a quality control system, particularly a system for evaluating the quality of MT.

Compared with conventional quality control of a manufactured product, the evaluation of the quality of translation is somewhat special in that :

- there are no hard-and-fast quality standards (e.g. dimension, weight, physical or chemical properties, etc. to which tolerances can be applied), failure to meet which in full entails rejection of the product concerned. Here, on the contrary, the quality of MT is evaluated without reference to a standard (which it would be very difficult to establish at present), although it is compared with the quality of other translations : HT another MT system or another version of the same system

- certain MT evaluation criteria (in particular intelligibility p are qualitative and the measuring instrument (the evaluator) gives rise to far greater variability than the yardsticks used in industry. In order to take account of this variability and to neutralize its effects, it is necessary to establish, in addition to the product sample (translated texts and the various methods of translation), a sample of measuring instruments (the evaluators).

The problem of the cost of evaluation once again arises at this point

-it is fairly generally accepted that as it stands MT does not and cannot compare favourably with HT from the point of view of quality. However, before being compared with MT, HT had never undergone any quality control of the quantitative type commonly used in manufacturing industry.

It therefore seems somewhat unfair to want to apply to a product (MT) which is less “finished than HT, a sophisticated system of quality evaluation.

This explains why most authors stick to empirical sampling methods.

The fixed bench mark sample method was rejected by the majority of the participants in the seminar held in Luxembourg in February 1978 on the evaluation of translation, for a number of reasons :

- artificial situation
- danger arising from the learning effect
- non-inclusion of different types of texts
- academic model
- difficulty of assessing the representativeness of a sample compiled in this way.

Moreover, a sampling method based on a detailed typology of texts is hardly feasible economically : J.M. LEICK shows that a reliable sample must include several hundreds of sentences, or several thousand homogeneous text' words. An evaluation of texts segmented by homogeneous type, using a typology comprising - because of the need to cross-reference the criteria characteristic of a given text several tens or hundreds of different classes, would require a sample of several tens or hundreds of thousands of words and would cost far too much.

The sample of texts must therefore be compiled on empirical bases :

-volume of text of the order of 5,000 - 10,000 words, i.e. 250 to 500 sentences

-significant passages (5 - 20 sentences) selected from documents belonging to 4 - 6 separate categories.

The same problem of cost arises when the sample of evaluators is drawn up: if unpaid labour (students, officers, etc.) is available, the number (several dozen) and the quality of the evaluators can be selected in such a way as to guarantee maximum statistical reliability; on the other hand, the evaluators have to be remunerated, it, is necessary, in order to keep costs down, to employ a restricted number (a few units), and to measure, a posteriori, the dispersion of their scores.

4. Summary, conclusions and recommendations.

We present below the summary and conclusions of -the analysis carried out in paragraph 3 from the seven viewpoints from which the problem of MT evaluation has been considered

- aims of evaluation
- translation quality
- text typology
- effectiveness and efficiency of evaluation
- macroevaluation - criteria and methods
- microevaluation - criteria and methods
- sampling.

Following this summary, a series of recommendations are made on how MT evaluation should be conducted by the Commission of the European Communities.

4.1 Summary and conclusions.

4.11 Aims of evaluation.

A limited number of evaluation authors have considered the problem of the aim of MT evaluation and have expressly formulated the objectives to be pursued; for the majority of authors, this aim is implicit.

From an examination of these works, and those concerning another field of information science (the evaluation of information recording and retrieval systems), two main approaches emerge, each corresponding to a precise set of aims :

-the macro evaluation, which is designed to measure product quality

-the microevaluation, which seeks to assess the improvability of the system. The macroevaluation makes it possible

-to compare the quality of two translation systems

- MT and HT
- MT produced by various translation softwares
- MT produced by successive versions of the same software

- to take delivery of an MIT system or a new version (linguistic or technological) of an MT system

- to assess the usefulness of MT and, if necessary, the desirability of undertaking:

  . an acceptability study

  . and/or marketing study

  . and/or one or more pilot operations.

The microevaluation makes it possible

- to assess the improvability of a system

- to evaluate the quality of the system "in the limit case"

- to identify the causes of the errors made by the system

- to assess the desirability of implementing a series of improvements to obtain a new technological version of the system

- to set priorities as regards the improvements to be made to the system.

It is immediately clear that the macroevaluation, which concerns the product/user interface, is amore limited and therefore less expensive operation than the microevaluation, which studies the system/product interface.

4.12 Translation quality.

It is difficult to assess the quality of an original text; the evaluation of its translation raises 'he even greater problem of how to define the quality of the translation.

The authors who have dealt with this point agree that there can be no absolute assessment of translation quality : any evaluation must involve several criteria. It is all the more essential that this assessment is made from the point of view of the user, of whom it would be wrong to believe that he always requires a perfect translation : MT is a different product from HT and it has to be able to find its own market.

4.13 Text typology.

It is undeniable that the texts subjected to translation, be it human or machine, are not homogeneous and present varying treatment difficulties, for both man and the machine.

It would therefore be useful to be able:

-to have a text typology

-to assign a single heading from this typology to each text presented for translation

-presented for translation

-to choose the appropriate translation method on the basis of this leading :

. HT with or without revision, by a translator specializing in the subject or a “general practitioner”

. MT with or without pre-editing, preliminary revision of interactive editing during processing, the vocabulary post-editing.

At present, such a typology does not exist; two avenues are open:

- implementation and progressive refinement, in the light of experience, of an empirical typology, based on simple criteria:

.field covered (with which the specialist translator must be familiar or which has to be taken into account by the dictionaries of the MT system)

.accuracy of spelling, vocabulary and syntax : an incorrect text can be rendered more or less satisfactorily by HT whereas it will probably never be well translated by MT

- launching of a fundamental research programme designed to produce an automatic classification of texts according to translation difficulty and the translation methods.

4.14 Effectiveness and efficiency of the evaluation.

The evaluation criteria have to be:

-effective : they have to measure effectively the quality of the translation, which means that they have to be:

. valid
. reliable
. general (i.e. applicable to any MT or HT)
. sensitive (i.e. reveal whether the translation has rendered the spirit of the text, that is the author's intention, and not merely the letter)

- efficient : they have to be effective while minimizing cost of the evaluation.

4.15 Macroevaluation –criteria and methods.

The macroevaluation criteria and methods can be on four levels:

-cognitive level -economic level -linguistic level -operational level.

4.151 The cognitive level

is undoubtedly a fundamental element, insofar as the role of a text is to convey knowledge and the role of the translation is to ensure the faithful rendering of this knowledge in a target language.

The various criteria proposed for measuring the cognitive level of translation quality (intelligibility, fidelity, consistency, usefulness, acceptability) are all valuable since each assesses a specific facet of the complex concept of Quality.

It would in fact appear (§ 4.14) that there is little or no correlation between them.

Consequently, if a completely effective evaluation is required it would seem to be essential to take all these criteria into account.

Si, par contre on insiste plus particulièrement sur le moindre coût de l'evaluation, on sera amené à choisir celui ou ceux de ces critères qui sont les plus efficients: l'intelligibilité est certainement le critère le plus efficient, parce que :

- suffisamment efficace pour mesurer le transfert de l'information en appréciant sa compréhension

- relativement peu couteux à mettre en oeuvre (échantillon de textes originaux, de TA et de TH soumis à un échantillon d'évaluateurs "en chambre" qui n'examinent chacun qu'une version linguistique de tout ou partie de 1'échantillon).

Les quatre autres critères sont moins efficients, essentiellement parce qu'ils sont difficiles (pour ce qui concerne la fidélité et la cohérence) ou impossibles (pour ce qui est de 1'ullilité et de l'acceptabilité) à apprécier par des évaluateurs qui ne sont pas les destinataires réels des textes, c'est-à-dire leurs utilisateurs finals. Or il est difficile, sauf cas particuliers, de mobiliser des utilisateurs finaus pour leur demander d'évaluer de façon suffisamment analytique un échantillon suffisamment important de textes.

Néanmoins, la fidé1ité et la cohérence peuvent être appréciées, d'un point de vue formel, par des évaluateurs qui ne sont pas des utilisateurs finaus.

Parmi les différentes méthodes préconisées pour mesurer effectivement la valeur de ces critères, ceux qui sont à la fois les plus efficaces et les plus performants apparaissent être :

- la cotation sur une échelle en 4 points pour 1' éval- uation de l'intelligibilité et de la fidélité par des évaluateurs "en chambrel"

- le jugement des utilisateurs finaus pour 1'évaluation de 1'utilité et de 1'acceptabilité.

4.152 Le niveau économique est également un critère, essentiel dans le monde réel.

Deux des critères cités apparaissent mesurer cet aspect de la qualité de la traduction de façon suffisamment efficiente :

-le temps de lecture (qui s'obtient en sous-produit de l'évaluation de l'intelligibilité)

-le temps de correction : révision et/ou post-édition (qui s'obtient en sous-produit du travail de correction).

On the other hand, the third criterion proposed (production time), appears to be bound up with organizational factors; it measures the quality of the service rather than the quality of the translation and must therefore be rejected.

4.153 The linguistic level of a translation is of undeniable scientific interest for linguists. For the other parties involved, it has the advantage of a high degree of reliability (thanks to almost 100%- objectivity); its main disadvantage is that it is not valid, since it carries no meaning. It therefore seems that it need not be taken into account in evaluation operations carried out by organizations such as the Commission of the European Communities on systems of the SYSTRAN type, in which the rules for the translation of the same linguistic features are taken into account by different elements of the system : dictionary, limited or conditional dictionary expressions, grammar, etc. It could, of course, be different in other systems where the translation rules for each linguistic feature correspond to a homogeneous element of the system : in such a case, a list of the errors on a linguistic level would provide significant information on the causes of the system's weaknesses and would make for an easy transition from the macroevaluation to the microevaluation.

4.154 The operational level does not appears to be very effective and therefore need not be retained.

4.16 Microevaluation - methods and criteria.

The microevaluation of MT, which is concerned with the causes of and the remedies for the errors in the translation, can be seen on several levels, from the highly theoretical to the highly analytical :

- grammatical level (symptom)
- format level (symptom)
- causes (diagnosis)
- improvability (prognosis)
- actual improvement (therapy).

4.161 The analysis of grammatical errors is an interesting approach it corresponds to a mental pattern which has been in built in all speakers of a language since they were at school; it therefore appears to be the “natural” way of assessing the quality of a text or a translation.

In practice, it has to be admitted that in the case of MT, it is virtually inapplicable : it does not interest any of the parties involved : decision-makers (concerned with the acquisition and/or perfecting of the system), users (translators, end users), post-editors, managers (analysts, linguists and coders in charge of the creation and perfecting of the system).

It thus seems unadvisable to adopt this criterion for the microevaluation of MT.

4.162 The analysis of formal errors (by correction type : words deleted, added, moved to a different position, etc.) is, on the other hand, more interesting, because it makes it possible

- to describe, in a manner which can be objective, the work required to correct the rough MT and give it a similar intelligibility and style to the HT

- to compare, in a quantitative manner (by way of calculation of the correction rate), the work involved in post-editing the MT and revising the HT.

However, when analysing the correction of formal errors, the psychological aspect of post-editing work should not be disregarded : certain post-editors are in fact favorably disposed towards MT and might be tempted to restrict the number of correction to a minimum; others, on the contrary, show true “editorial teal” and make a greater number of corrections to MT than is necessary.

The correction rate must therefore always be assessed in the light of the quality of the post-edited text, and in particular its : when the same texts are submitted to several post-editors, the lowest correction rate giving a degree of intelligibility close to that of the original text or the revised HT, and that rate only, should be taken into account.

4.163 The analysis of causes of errors goes further than the previous two steps, since it permits an initial diagnosis to be wade of the unsatisfactory functions of the system.

However, it should be noted that this diagnosis :

- remains at a rather superficial level, since it covers only the main functions input : analysis of source language, synthesis of target language, etc.

- is relatively expensive, since it requires the intervention of an evaluator who is a specialist in the translation system in question, and a thorough examination of the errors and their origin

- provides information usually “of value”, but cannot serve as a basis for concrete action.

It thus seems, from experience, that this criterion should not be retained for the purposes of the microevaluation.

4.164 The analysis of improvability in fact corresponds exactly to the definition of the microevaluation, it alone makes it possible

- to assess the type of remedies to be made to the translation system to prevent a certain number errors (“effectiveness” aspect of the remedies).

- and to estimate the resources needed to introduce these remedies(“cost” aspect of the remedies).

It alone should serve as a basis for a real strategy for the improvement of the MT system founded on the efficiency of the work to be done, i.e. on the cost/effectiveness of each remedy.

Unfortunately, none of the studies described in the present document has been carried sufficiently far to arrive at a clear statement of a concrete improvement strategy for specific MT system.

It should be noted that this type of microevaluation is very expensive because it' requires considerable time to be devoted to it by specialists of each element of the translation system : systems analysts, linguists in charge of grammar rules, lexicographers, and good coordination of their respective approaches.

4.165 The actual improvement of a translation system, even if it is attempted or proposed in the framework of a microevaluation of MT, seems to us, in its aim and cost, to go far beyond the scope of the evaluation of a system and it therefore seems to us that it need be considered here.

4.17 Sampling.

As regards text sampling:

-the bench mark method seems to be excluded : it is in fact desirable that the author of the system evaluated should be informed of the content of the sample : to proceed otherwise would amount to judging a system, and thus its author, without allowing him a chance to defend himself, and to giving the evaluator absolute authority.

Once the sample has been revealed it can no longer be used for a later evaluation; it would be too easy for the author of the system to change the design of the system so that it could produce an almost perfect translation of the sample

-the size of the sample can reasonably be set at + 10,000 words except in special cases (in particular the actual improvement of the system, which should be measured on a large batch of documents : cf. §4.165).

As regards the sample of evaluators, a large number of works quoted refer to help from students or soldiers, i.e. evaluators who will work without pay.
When this is not the case, as at the Commission, economic considerations lead to a reduction in the number of evaluators.

It is indeed pointless to require greater statistical precision than is necessary to achieve the aims set for the MT evaluation.

4.2 Recommendations.

4.21 Background consideration.

The present state-of-the-art report on the evaluation of machine translation was drawn up at the request of the Commission of the European Communities, with the aim of reviewing the literature existing in this field and of drawing from it practical conclusions in the form of relevant recommendations.

These recommendations must be seen against the specific background of the Commission's objectives, and its various responsibilities with regards to machine translation :

- as regards the aims and the political options :

  .promotion of MT systems with a view to lowering the barriers between the languages of the Community countries

  .improvement in the efficiency of the Commission translation services

- strategic:

  . application and improvement of a certain number of language versions of an MT system already available on the market : SYSTRAN

  . promotion of the development, implementation and improvement of all language versions (of interest to the Community) of an MT system to be created by a European team : EUROTRA

- tactical:

  evaluation of the various language versions of MT systems and successive improvements in such a way as to:

  - measure the Progress achieved and to decide on new improvements to be commissioned
  - compare various systems

  . study of the MT market and analysis of the technical, commercial and organizational conditions of its promotion

Our recommendations are based on:

the results of this study of the state of the art our understanding of the aims of the Commission, based on our participation in the work of CETIL (Committee of Experts for the Transfer of InformatIon between Community languages)and on several evaluations of SYSTRAN carried out for this committee.

These recommendations may, of course, be taken into account by institutions other than the Commission, insofar as they consider it necessary.
A certain harmonization of the methods followed is certainly advisable.
Standardization of these methods, on the other hand, is not desirable : the circumstances governing an evaluation vary from case to case.
For example, the evaluation of an MT system used in a very specific field (e.g. ; an aviation maintenance handbook; meteorological bulletins, etc.) may make use of very specific methods (e.g. : performance testing, consisting in comparing the work carried out respectively on the basis of an original maintenance handbook and of a translated handbook).
On the other hand, when the system to be evaluated has to
be capable of use in a very wide variety of fields, as in the case of the Commission, the evaluation criteria have to be more general in character.

4.22 Orientations

Our recommendations relate:

- on the one hand to the methodology to be applied specifically and in the short term by the Commission in MT evaluation

- on the other hand, to a certain number of lines research which would allow the results of the MT evaluation work to be improved in the medium term,

4.23 Evaluation methodology.

We propose that there should be three types of
evaluation programme :

- a superficial evaluation

- an in-depth evaluation

- a pin-point evaluation.

The first, which would be inexpensive and easy to use, would be applied primarily at the macroevaluation level; it would be applicable when each new version (technologcal and/or linguistic) of an MT system became available; it would permit an overall and comparative appreciation of the quality of each version.

The second, which would be more elaborate and more expensive, would be applied primarily at the microevaluation level; it is purpose would be to evaluate the acceptability and improvability of the system, and the improvements effected by simple updating of the dictionaries on the basis of the sample texts. in general, it would have to be done on delivery of an improved version of a system of which the initial version (for a given language couple) would already have undergone one or more superficial evaluations.

The third type of evaluation would be applied on a case-by-case basis to evaluate an improvement made on a specific feature or a combination of features of the system.

4.231 Superficial evaluation.

4.231.1 Criteria.

Assessments criteria and methods	Effectiv-ness of The criteria	Cost of applying The criteria
- Intelligibility, rated on a four-point scale	Good	Moderate
- Fidelity, rated on a four-point scale	Poor	relatively high
-Reading time(measured during the intelligibility assessment)	Poor	virtually nil
-Correction rate (revision of human translation and post-editing of machine translation)		
. overall	Good	Moderate
. by type of correction	Good	Relatively high
- Correction time	Good	Virtually nil

4.231.2 Text sample.

5,000 to 10,000 words, constituting continuous sentence groups, extracted from 20 to 40 documents; each sentence group must be comprehensible to the reader (the evaluator) who does not have at his disposal the complete document from which it was extracted.

These texts should:

- related to a field or a group of fields covered by the dictionaries of the translation system

- be taken from real documents (not artificially compiled)

- relate to a limited number of categories : complete and summarized texts, scientific, economic and administrative texts

  in general, exclude texts which are known to be unsuited to MT : speeches, legal texts, literary texts, advertising blurbs, etc.

Four versions at least of these texts should undergo comparative treatment

- the original text, in the source language (V1)

- a text translated by the machine and not post-edited (V2)

- a text translated by the machine and post-edited (V3)

- a text translated by a human translator (V 4)

In the cases where a high quality translation is required, a fifth version would have to be examined a text translated by a human translator and revised (V 5).

Each sentence should be given a sequence number. Which should reappear in all the versions.

4.231.3 Sample of evaluators.

The choice of the number of evaluators depends on the following features :

- subjectivity of the criterion to be evaluated :
  - • outstandingly subjective criterion : fidelity
  - • very subjective criterion intelligibility
  - • fairly subjective criteria reading time, correction rate and correction time
- versions to be evaluated :
  - • assessment of the Quality of the original text and of the human translations is merely a subordinate factor which allows a decision to be taken on whether a machine translation is feasible. If the original text is difficult to understand, and/or if the human translation (made under as normal conditions as possible) is poor, the consequence will be that the quality of the machine translation will be more a reflection of the (poor) "translatability" of the text than of the value of the machine translation system
  - • the main purpose of the evaluation is to assess the quality of the machine translation, and the available resources should be directed to this end
- available resources :

  various types of person are involved in an evaluation:
  - • translators and revisers in normal employment with the organization using the machine translation system (this will ensure that the quality of the revised human translation and of the post-edited machine translation will be the same, and that this quality meets the normal standards of that organization)
  - • operators responsible for input of the source texts
  - • computer centre operators
  - • evaluators
  - • project leader (responsible for selecting texts and evaluators, directing and following up operations, and preparing the evaluation report).

On the basis of these elements, and in the light of the experience gained to date it appears advisable:

- to distribute the translation, revision and post-editing work among a number of translators, revisers and post-editors in such a way as to ensure the time taken to do the work, and the standard of correction, are comparable to those found in normal practice

-to use:

. several evaluators (between 4 and 10) to assess the intelligibility of the machine translation (V2)

. one or two evaluators to assess the intelligibility of the original text (V 1), the post-edited machine translation (V3) and the revised human translation (V5)

. a single evaluator to make the overall assessment of the fidelity of the machine translation to the original text. An exact assessment of fidelity is of course virtually impossible in 'the case of scientific, technical or administrative texts : only the the real users of such texts can properly assess the inaccurencies of the translation, and even then such assessments will be subjective since they will depend on the importance which each user attaches to each of the basic messages contained in the text and any distortion of them

. the time taken by the evaluators as a measure of the reading time of a normal reader

. a single evaluator to compile the reading times (noted by the other evaluators) and the correction rates and correction times by the revisers and post-editors.

Remarks.

-The volume of work involved in evaluating the intelligibility of Version 2 (MT not post-edited) can be reduced by giving each of the evaluator a part of the whole text sample on a rotating basis

-It will be possible to calculate the correction rate:

. synthetically (total number of corrections/ number of words of the basic version) for all of the text samples together

. analytically, by type of correction : substitution, correction, alteration of word order, elimination and addition, for half of the text sample, at a rate of one sentence in two.

The evaluators who examine the of the two language versions (intelligibility with rotation of the versions among the evaluators, and fidelity) must have a thorough knowledge of these two languages and a training which enables them to understand the technical content of the documents.

4.232 In-depth evaluation.

4.232.1 Criteria.

In addition to those applicable to the superficial evaluation (intelligibility, fidelity, reading 'Lime, correction rate and correction time) :

-acceptability (effectiveness of criterion good; cost of applying the criterion : rather high)

-improvability (very effective criterion extremely high cost)

-real improvement, before and after dictionary updating (this criterion yields interesting information and is relatively inexpensive to apply).

4.232.2 Text sample.

As for the superficial evaluation (see 9 4.231.2), but in the upper range, i.e. 10,000 words.

4.232.3 Sample of evaluators.

-Superficial evaluation criteria : see § 4.231.3

-Acceptability : the evaluators must be regular readers of the texts of the sample.

A sample of around twenty readers, from at least three different organizations, constitutes a minimum capable of providing very general assessments.
To obtain statistically reliable data, it would be necessary to question several hundreds of users. This, however, would mean leaving the field of evaluation for that of market research

-Improvability : the assessments can be drawn up only by specialists with knowledge of all the functions of the translation system. Insofar as they exist, it will be necessary to recruit such specialists from teams charged with managing and improving the machine translation system and located within the Commission, or working directly for the Commission. Recourse to specialists provided by the authors of the system is obviously excluded,(i.e. for evaluation purposes; it will obviously be necessary for work on improving the system)

-Improvement : the same evaluators as those who assess the intelligibility.

4.24 Main lines of research.

In the context of the MT evaluation methodology recommended above, two types of research should be carried out which would have an important impact on the effectiveness of the operation:

-Research into the typology of the texts subjected to MT (cf. § 4.13); it should be noted, however, that this type logy should aid the choice of translation method best adapted to each category (editing a priori, a continue, a posteriori, types of person involved, types of dictionaries, etc.)It appears that EUROTRA will offer, in this connection, much greater possibilities than SYSTRAN. It is consequently desirable that the conditions of cooperation between EUROTRA and the translation editors should be defined with sufficient precision before the text typology study is undertaken

-Research into the methodology of analysing the improvability of an MT system (cf. § 4.164), This research would have to relate, less to the identification of the system features capable of improvement than to the possibility of formulating a true improvement strategy : e.g. a list of the elements to be improved, individual cost of the improvements to be carried out, probable individual effects on the criteria of intelligibility, fidelity and correction rate.

5. Bibliography.

Remarks.

The references preceded by an asterisk relate to documents presented during the Workshop, held in Luxembourg on 28 February 1978, on the problems of machine translation evaluation.

This bibliography contains 45 references, 29 of which are papers read to the February 1978 Workshop.

* ANDREEWSKY (A.).- Le problème de l'évaluation d'une traduction automatique.- Luxembourg, CEC, memorandum, February 1978, 4 p.

* ARTHERN- The usefulness of machine translation.- in : Minutes of the Workshop on evaluation problems in machine translation.-CETIL Doc. 38/78, p. 3

Association Jean FAVARD.- Définition de la notion de qualité de traduction et m6thodes d'évaluation.- Luxembourg, CEC, memorandum, January 1977, 13 + 5 P.

* BOURQUIN (G.).- Observations on the problems of assessing human and machine translation.- Luxembourg, CEC, memorandum, February 1978, 4 p.

* BRUDERER (H.).- The qualitv of translations.- Luxembourg, CEC, memorandum, Doc. nr 2226/78f, January 1978, 5 p.

* CARROLL (J.B.).- An experiment in evaluating the quality of translations. -Washington DC, National Academy of Sciences, 1966, ALPAC report, appendices 10 and 11, pp. 67-78.

CHAUMIER (J.), MALLEN (M.C.) and VAN SLYPE (G.).- Evaluation du système de traduction automatique SYSTRAN; évaluation de la qualité de traduction.-Luxembourg, CEC, Report no 4, June 1977, 51 p. (a part of this report was distributed during the Workshop held in Luxembourg on 2-8 February 1978).

DEHAVEN (R.C.) et al. (Synectics Corp.).- SYSTRAN machine translation evaluation- Springfield, NTIS, November 1972, 153 P., RADC-TP-72-293/AD 753.676

DOSTERT (B.H.).- Users' evaluation of machine translation- New York., Rome Air Development Center, RADC-TR-73-239, August 1q973, 10-5 + 6 p.~ (IT)art of this report was made available to the Luxembourg February 1978 Workshop).

* GILLESPIE (P.).- Comments on machine translation evaluation problems.-Luxembourg. CEC , memorandum, February 1078, 2 p.

* GREEN (R.).- Analysis of errors- Luxembourg, CEC, memorandum, October 1977, 5 + 5 p.

HALLIDAY (T.C.) and BRISS (E.A.).- The evaluation and systems analysis of the SYSTRAN machine translation system.-NTIS, ADA 036.070
January 1977, 70 + 6 p. (part of thiS report was made available to the Luxembourg February 1978 Workshop).

HOFFMANN (E.).- Tests on SYSTRAN (English-French) .-Luxembourg, CEC, CETIL memorandum 68/78, June 1978, 7 p.

* HOFSTETTER (A.).-Methodological concept for the evaluation of translation quality- Luxembourg, CEC, memorandum, January 1978, 5 P.

HOUSE (j.), A model for translation quality assessment, Tubigen, TBL Veriag Gunter Narr, 1977, 344 p., ISBN 3-87808-088-3.

* JOHNSON (R.L.).- Evaluation of machine translation systems : some thoughts.- Luxembourg, CEC, memorandum, February 1978, 4 p.

KLARE (G.R.), SINAIKO (H.W.) and STOLUROW (L.M.).- The Cloze procedure : a convenient readability test for training materials and translations.- International Review of Applied Psychology, 1972, Vol. 21 , no 2, pp. 77-106.

* KNOWLES (F.), A brief error analysis of SYSTRAN output.- Luxembourg, CEC, memorandum, February 1978, 3 p.

* KROLLMANN (F.).-Kurze Stellungnahme.-Luxembourg, CEC, February 1978,2 p.

* KUHLEN (R.).- Einige vorbereitende Bemerkungen zur Voraussetzung einer Bewertung von Systemen zur automatischen Ubersetzung.-Luxembourg, CEC, January 1978, 3 p.

LEAVITT (A.W.) et al. (Synectics Corp.).- Machine translation quality and production process evaluation, Springfield, NTIS, October 1971, 146 p., RADC-TR-71-2o6, AD 732.886

LEICK W.M.) and SCHRAEN (D.).- Some statistical results of a brief evaluation of the SYSTRAN machine translation system.Luxembourg, CEC, CETIL memorandum 77/78, 13 + 3 P.

* LENDERS (W.), Bewertungskriterien fur maschinelle Sprachubersetzungsysteme.- Luxembourg, CEC, February 1,978, 13 P.

* MASTERMAN (M.).- Can we progress in determining criteria for evaluating machine translation.-Luxembourg, CEC, memorandum, February 1978, 9 P.

* PANKOWICZ (Z.L.), Evaluation of machine translation; a position paper.- Luxembourg, CEC, memorandum, 1978, 6 + 3 P.

* PETIT (A.J.).- Notes sur 1'évaluation d'un système de traduction automatisée.- Ottawa, Bureau of Translations, January 1978, 31 P.

* PHILIPS- Report on a survey of quality criteria pertaining to technical translations.- Philips Technology Bulletin, Vol. 6, no 2, 1077, pp. 9-13.

* ROLLING (L.), Essai de typologie des textes.- Luxembourg, CEC, memorandum, January 1978, 2 p.

* SAGER (J.C.).- An investigation of quantity and type of translation material, Luxembourg, CEC, CETIL memorandum 18/77, December 1977, 3 P.

* SAGER (J.C.).- Criteria for MT evaluation.- Luxembourg, CEC, memorandum, February 1978, 7 p.

* SINAIKO (H.W.), Some thought about evaluating language translation.- Luxembourg, CEC, memorandum, February 1978, 7 p.

SINAIKO (H.W.) and BRISLIN (R.W.)- Evaluating language translations experiments on three assessment methods- Journal of Applied Psychology, 1973, Vol. 57, no 3, pp. 328-334.

SINAIKO (H.W.) and KLARE (G.R.), Further experiments in language translation : readability of computer translations. - Louvain, institute of Applied Linguistics, 1972, 29 p., ref. ITT, 15/1072.

SINAIKO (H.W.) and KLARE (G.R.), Further experiments in language translation : a second evaluation of the readability of computer translations.- Louvain, Institute of Applied Linguistics, 1973, pp. 29-52, ref. ITL 19/1973.

SPILLEBOUDT (R.), Method for measuring intelligibility. Luxembourg, CEC, memorandum, March 1978, 4 p.

* VAN SLYPE (G.).- Colloque du 28 février 1978 sur l'évaluation des systèmes de TA; Option Paper.- Luxembourg, CEC, memorandum, January 1978, 11 p.

*VAN SLYPE (G.)- Second evaluation of the English-French SYSTRAN machine translation system of the Commission of the European Communities.-Luxembourg, CEC, Final report, November 1978, 179 p.

VAUQUOIS (B.)- La traduction automatique A Grenoble-Paris, Dunod, 19751 179 p., ISBN 2-0L-009956-5.

* VAUQUOIS (B.).- Some thoughts on evaluation criteria for computer aided translation systems.- Luxembourg, CEC, memorandum, February 1978, 2 p.

* VAUQUOIS (B.).- Dynamic analysis.- in: Minutes of the Workshop on evaluation problems in machine translation, CETIL Doc. 38/78, p. 4.

* VEILLON (G.).- Thoughts on the evaluation of a machine translation programme.- Luxembourg, CEC, memorandum, February 1978, 5 P.

* WEISSENBORN (J.).- Skizze eines Bewertungsverfahrens fur Ubersetzungen und einer maschinenUbersetzungsbezogenen Texttypologie.- Luxembourg, CEC, February 1978, 5 P.

* WILKS (Y.).- The value of the monolingual component in MT evaluation and its role in the Battelle report on SYSTRAN.Luxembourg, CEC, memorandum, February 1978, 7 p.

* ZEMB (J.M.)- Sampling in vitro.- in : Minutes of the Workshop on evaluation problems in machine translation, CETIL Doc. 38/78, p.4.