9
Evaluation of MT systems

The different modes of use of MT which were discussed in the last chapter are motivated to a considerable extent by the inadequacies of the translations produced by systems. Fully automatic high quality translation is not at present possible; in general, the 'raw' translations produced by MT systems must be revised or post-edited. Alternatively, the input to MT systems must be pre-edited or adapted to the limitations of systems. A major question to be asked about any MT system is, therefore, how good are its raw translations, what is the potential for improvement, and how may it be best and most cost-effectively used in practice?

What may be surprising is that despite some 40 years of research on MT there is still no generally accepted methodology for the evaluation of systems. Although the ALPAC report did include some evaluations of then existing systems it is only since the initial assessments of the Systran system for the European Communities in the late 1970s that the topic has received much attention. Most evaluations take place under contract and often under confidentiality agreements. Consequently there is little constructive criticism of methodology. A major deficiency is that many evaluations are undertaken by people with little or no expertise in MT techniques, unable to judge what is possible and what is unrealistic, unable to estimate the potential rather than current performance. On the other hand, 'evaluations' made by MT researchers are often minimal and misleading: the demonstration of a system with a carefully selected set of sentences or sentence types is not the basis for claims about a large-scale system. In view of the misconceptions and misunderstandings concerning nearly all aspects of MT, one role of evaluation must be to introduce realism in public discussions of what MT systems can and cannot do and what they may be able to do in the future.

With no generally accepted methodology all that can be done here is to indicate the principal areas in which evaluation can take place, what aspects should be taken into consideration, and some of the methods which may be employed.

9.1 Types and stages of evaluation

At the outset we need to distinguish various stages in the development, installation and operation of MT systems. At each stage systems can be evaluated for performance and efficiency. The first stage is the development of a prototype system, the design of the basic system, writing programs, compiling dictionaries, etc. Evaluation will be restricted to the testing of processes alone, without consideration of potential operational environments. It will be the concern primarily of the system developer, who wants to know whether the programs written for the system are in fact performing in the ways intended and who wants to discover whether the output is acceptable as a 'translation'. Such evaluations can be, in fact normally will be, continuous during the development stages of a system, i.e. the programmer or the linguist will want to know the effects of any changes of coding or of a grammar rule. However, from time to time during development the designers of the system will undertake more extensive tests, e.g. running texts of some size against the system and testing the adequacy of the dictionary information and the grammar. These tests may be called **prototype evaluations.**

The second stage is the development of a system which can operate in the intended environment, the design and provision of facilities for inputting text, for compiling and updating dictionaries, for revising (post-editing) output, for interacting with the computer, etc. This stage will include also improvements in the robustness and computational efficiency of programs, the integration of the system in a particular computer environment, e.g. adaptation to particular operating systems or particular computer hardware. These **development evaluations** are concerned not only with the linguistic capabilities of the system but also with operational capabilities. Before offering systems to potential users, the developers will want to be assured that the system does what it is claimed to do, whatever type of system it may be, and that modifications and improvements can be made without radical changes to programs and facilities. In the case of systems to be marketed, they will also, of course, want to evaluate the economic viability of the system as a commercial product, potential sales and leasing agreements, servicing costs, etc.

The third stage is the evaluation of a system by its potential purchasers. They will want to know not only whether the system performs as the designers or vendors claim but also whether it can be used cost-effectively in their particular circumstances. These **operational evaluations** will include assessments of how much and what kind of human input is required to produce acceptable translations, what working conditions are required and what qualifications are needed by operators, what technical facilities are required, how compatible the system is with existing equipment, how improvements can be introduced, etc.

Closely related to these concerns are those of the actual users of the system, particularly professional translators. They will want to know how much work will be involved in pre-editing, post-editing or interactive operation, whether

productivity will be increased, how much time and effort is saved, and the impact on working practices. Typically, these **translator evaluations** are included within operational evaluations, but they do represent a distinct area of concern, since the individual attitudes of translators can have major impact on the success or otherwise of MT systems in organisations.

The final stage of evaluation is that of the recipients of translations. They will be concerned primarily with quality, cost and speed. Inevitably these **recipient evaluations** will be comparative: whether the introduction of the MT system produces translations faster and/or more cheaply than human translation while maintaining quality. What they want is good, fast and inexpensive translation.

9.2 Linguistic evaluation of 'raw' output

Common to all stages is the testing of the linguistic quality of the output, the quality of 'raw' translations. This is an evaluation of the basic computer processes. Here there are important differences between **glass-box** evaluations and **black-box** evaluations, between assessment by those who have access to all the workings of the system and assessment by those who can work only with inputs and outputs. The former is available generally only to the researchers and developers of prototype systems, while potential purchasers and users are restricted to black-box evaluations. In examinations of systems we may further distinguish between overall assessments of 'quality' and more detailed identifications of 'errors'. The former tend to produce more subjective evaluations, while the latter provide more objective practical data.

9.2.1 Quality assessment

The most obvious tests of the quality of a translation are: (a) its fidelity or accuracy, the extent to which the translated text contains the 'same' information as the original; (b) its intelligibility or clarity, the ease with which a reader can understand the translation; and (c) its style, the extent to which the translation uses the language appropriate to its content and intention. Each factor can be independent: a translation which is faithful to the original may be difficult to understand; a translation easy to read may have distorted the original message; and an intelligible rendition may be in a quite inappropriate style.

Various tests have been proposed and implemented to provide measurements of **fidelity**. As part of the ALPAC investigation (Chapter 1), people were asked to read the output of an MT system and then judge how much more 'informative' the original was. This procedure can obviously be criticised as being excessively subjective. Rather more objective tests of fidelity have been proposed in more recent years. In the case of instruction manuals, a practical **performance evaluation** is feasible: can someone using the translation carry out the instructions as well as someone using the original? Another proposal involves **back-translation**: the MT output is translated back into the original language and the result compared with the original text, though of course there is a risk that any shortcomings are magnified by the double process.

For **intelligibility** (or clarity) a number of evaluations have asked readers to rank output from, e.g. 'perfectly intelligible' to 'hopelessly unintelligible'. Generally, only individual sentences are evaluated. But, by isolating sentences from their contexts, such tests are made even more subjective and uncertain than they might be. More objective tests have employed readability scales, such as the well-known Flesch scales, use of the Cloze technique, and comprehension tests. The Flesch scales are based on average sentence lengths, use of complex nominalizations, etc. The Cloze technique involves the masking of words in sentences and texts and asking readers to suggest words to fill the blanks: the correlation between suggested and original words is an index of 'readability'. Comprehension tests (well developed for educational purposes) can be employed to evaluate the intelligibility of the translated text as a whole, by testing readers' understanding of its content.

Measurements of **style** are as subjective as the global rankings of intelligibility. Nevertheless, the appropriateness of a particular style is an important factor. Determiners and copulas may be legitimately omitted in some contexts e.g. newspaper headlines and chapter headings. The translation of an English imperative (1a) by a French imperative would not be appropriate in an instruction manual, where the acceptable style is the infinitive (1b).

(1a) Open the control valve.

(1b) *Ouvrir le régulateur.*

9.2.2 Error analysis

While performance tests, comprehension tests, Flesch readability scales, etc. are certainly valuable and reasonably reliable global evaluations of translations, in most instances the most useful practical information is obtained from **error counting**. It is an index of the amount of work required to correct 'raw' MT output to a standard considered acceptable as a translation. In a typical case, the reviser (post-editor) counts each addition or deletion of a word, each substitution of one word by another, each instance of the transposition of words in phrases, and calculates the percentage of corrected words (errors) in the whole text. The method cannot be completely 'objective' for a number of reasons. Firstly, revisers differ in what they consider to be errors; some will ignore stylistic infelicities if they do not affect intelligibility and accuracy. Secondly, there are different levels of acceptability which are dependent on the particular circumstances in which the revision is taking place. But since operational evaluations (see below) are made for specific situations the estimation of error correction has a practical value.

For many purposes, however, the simple counting of errors is insufficient. What is needed is a **classification of errors** by types of linguistic phenomenon and by relative difficulty of correction. Some lexical errors are easily resolved by simple changes to dictionaries, while others may have implications for grammatical rules and for a whole range of vocabulary items. Some grammatical mistakes may be corrected by simple adjustments to a few lexical entries, others might involve alterations to the basic design of whole translation modules. This kind of detailed evaluation of errors is clearly essential for developers of systems, but it is also

of value to potential purchasers who want to know how and whether quality can be improved.

All the above tests can be and have been used to compare different MT systems. An additional method, particularly for research systems, would be the application of **bench-mark** tests. The performance of systems would be compared in the translation of a corpus of texts covering the whole range of linguistic phenomena which an MT system should be expected to deal with. Performance could be evaluated in terms of quality (overall fidelity and intelligibility), error rates and production speed, thus also providing potential purchasers with an indication of basic achievement. However, as yet, no bench-mark tests for MT systems exist and there will be obvious difficulties in reaching agreement about the selection of texts, what should be tested and how performance should be measured. The difficulty of defining a suitable bench-mark is exacerbated by the fact that MT systems differ widely in languages treated, types of texts, involvement of users in pre-editing, or interaction, and many other factors.

9.3 Evaluation by researchers

The continuous monitoring of performance is obviously crucial in the development of prototype systems. The principal tool is the **diagnostic trace**, a record of the stages through which a program goes to produce output. Typically, a text fragment (often a single sentence) is submitted to the system for processing and the results of each stage are displayed (on screen or in print) to enable researchers to examine the actual operations taking place. On this basis they can discover whether the program is doing what is intended, and, if it is not, where the mistakes are occurring. In many cases researchers will be testing a single part of the system, e.g. an analysis module or just the transformation of structures in a transfer module. In other cases they may want to test the whole system.

The errors identified combine problems of computational treatment and linguistic analysis. In theory, programming errors are relatively easily rectified, since it should already be known exactly what results are intended. Linguistic errors involve closer examination, distinguishing between omissions, incompleteness and mistakes in lexical data or grammatical rules and unsatisfactory application of data and rules in procedures of analysis, transfer and generation. In the identification and analysis of errors the researchers are guided by most of the approaches outlined above, from subjective evaluations of fidelity and intelligibility to classifications of error types. In effect, they see translation errors in the same way that post-editors do: what needs to be changed in order to produce 'acceptable' output. However, researchers are usually concerned with particular types of errors connected with the specific problematic linguistic phenomenon which they are dealing with at the time.

At various times during development, MT systems are evaluated as whole systems. This is when researchers test programs against larger fragments of texts in prototype evaluations. At this stage they look for more than erroneous treatments of specific constructions and vocabulary items. Evaluation of errors should be more exhaustive, with the aim of determining which errors can be corrected only by revision of basic procedures and which errors are simply a matter of correcting

dictionary data and can therefore be left for later development stages. Typically a system will pass through a series of such prototype evaluations.

9.4 Evaluation by developers

At some point, an MT system passes from a research project to a development project: changes in the basic design are no longer undertaken and the emphasis is on the creation of a practical system. The provision of facilities for users to input text, to compile and update dictionaries and to edit output texts entails the introduction of further types of evaluation. The linguistic evaluation of translations continues, but now the main concern is to identify errors which can be corrected within the capabilities of the system. Features which cannot be corrected will have to be dealt with by the provision of other software facilities.

During the development phase, it will be decided precisely what the limitations of the system are, what types of texts it can deal with satisfactorily, what subject domains should be covered in the dictionaries and the grammar rules, and what facilities must be provided to users, if these aspects were not already part of the original systems. The objective of development evaluations is to ensure that before being offered to potential users the system performs as intended and produces translations which are believed to be acceptable in known circumstances. The developers should discover precisely what improvements are feasible within the basic design, and how and at what cost they may be achieved.

Evaluations during development involve typically translations of substantial text corpora (covering the intended subject domain of the system) and the identification and classification of errors. The perspective will approximate that of potential users. Normally, the most frequent or least complex errors will be tackled first. After the dictionaries have been updated and some rules changed as appropriate, the corpora are tested again. The cycle is repeated until the results are 'acceptable', or it is considered that further investment of development time cannot be justified.

9.5 Evaluation by potential users

In many cases, system developers and potential users work together closely. Often, systems are in fact designed for specific users, who may participate in the development, may select corpora to be tested, and may be involved in evaluation. Such users will often know as much about the internal workings of the system, its limitations and potential, as the developers themselves.

More typical, however, is the potential user or purchaser who can evaluate a system only by its performance, as a 'black box' to be assessed by its results. Whereas the principal focus of evaluation by researchers and developers is the linguistic quality of the translation, the potential purchasers are concerned primarily with the capabilities, acceptability and cost-effectiveness of the system in their particular environments. From this perspective (Figure 9.1), the MT system is just one component in a broader configuration of systems.

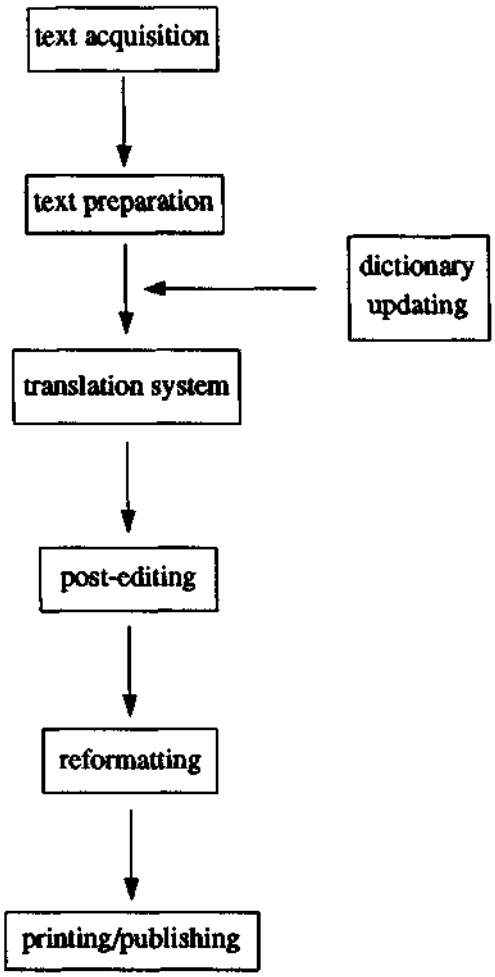

Figure 9.1 MT system within broader context

Text acquisition includes the production or receipt of a machine-readable text: word-processing, text transmission, optical character recognition (OCR) of printed or typed texts. Text preparation involves at least the separation of graphs and figures from the alphabetic text. In the case of 'restricted language' systems (section 8.3.1), preparation obviously embraces all the operations involved in checking and controlling the vocabulary and grammar of input texts. An important stage is often a preliminary examination of the text for unknown words and the consequent updating of dictionaries. After translation comes the revision of texts (post-editing), the reformatting of the text by insertion of graphs and figures, and finally stages in the preparation of translated text for printing and, possibly, publishing. In this context, translation is by no means the most expensive part of

the operation: text input and preparation and text post-editing and printing may be more significant in total costs.

The capabilities of a system are to be assessed in the light of the user's particular needs. These should be precisely established. What types of texts does the user intend to translate (e.g. patents, abstracts, manuals, reports)? What subject areas do they cover? Are they intended for specialists or for the general public, for information purposes only or for publication? What volume of texts (of each type and/or in each subject domain) are anticipated? And are they all to be dealt with by the system? Are there any time constraints on the production of translations?

9.5.1 Linguistic assessment

When evaluating the linguistic capabilities of the system, the user should also examine the linguistic characteristics of the texts to be translated. Particularly important are vocabulary considerations. If texts cover a wide range of subjects there may be problems of homography which the system cannot deal with satisfactorily. If the vocabulary of texts is well covered by existing terminology standards the problems of dictionary compilation may be lessened; otherwise the user may have to undertake terminological research. A high proportion of compounds and nominalizations may also have an impact. Other linguistic factors to be considered include the occurrence of structures or features which the system may not cover (e.g. interrogatives, imperatives, subjunctives, particular tenses). Certain styles may cause problems, e.g. the system may not be able to parse 'telegraphic' constructions (2a) for (2b), long noun strings (3) or sentences containing parentheses (4).

(2a) Arrive London Tuesday.

(2b) I will arrive in London on Tuesday.

(3) mains cable socket switch indicator

(4) The new system will (if installed correctly) increase production.

The submission of representative texts for translation should allow users to assess the limitations of the system and its potential for improvement and extension. Users do not, however, have access to information about grammars and programs, do not know the linguistic model on which it is based (if any), and may not be able to judge these aspects in any case. They are limited to examining output and analysing errors in order to determine how much human revision is required if translations are to be brought to an acceptable standard. Their difficulty is in identifying those errors which can be easily rectified (e.g. by changes to dictionary entries) and in recognising those errors which reveal major deficiencies of the basic design. Furthermore, some errors (e.g. pronouns) which look important may in fact be easy to correct in post-editing while they are very difficult to correct in the program. Users need to be aware that certain linguistic phenomena cause problems for all MT systems, e.g. complex sentences, coordination, ellipsis, fragmentary sentences, telegraphic phrasing, complex noun phrases.

In many cases, only discussion with developers can establish which errors can be rectified easily and which cannot. Nevertheless, a proper evaluation should involve careful selection of texts, the making and testing of hypotheses about the

way the system operates, by choosing suitable examples and counter-examples, and the detailed analysis of results. From such evaluations potential users should have an inventory of linguistic phenomena that are satisfactorily processed, and of those phenomena not processed at all; they should have available an explanation for good translations as well as poor translations, for good grammatical analyses as well as poor ones; they should know, in brief, why some kinds of texts are suitable for the system and why others are not. Ideally, they should also understand the basic nature and content of the linguistic components of the system, e.g. whether it analyses deep structure as well as surface structure, whether it uses an intermediate representation, what kind of information is taken from the dictionary, and when and how it is used, etc. Potential users should aim to be capable of making informed judgments of the true possibilities of improving the quality of output and extending the system to other domains and languages.

9.5.2 Evaluation of limitations, improvability, and extendibility

No MT system can ever be regarded by its designers as complete, least of all 'perfect', in its performance. Both developers and potential purchasers need to know what the limitations of the system are, and how it may be improved. Some limitations are obviously inherent in the basic design: most systems are limited to sentence analysis, thus no treatment of cross-sentence anaphora (section 5.4) can ever be possible. Many systems provide little or no semantic feature analysis of nouns, and thus the treatment of multiple noun–noun compounds such as (5) is severely restricted. However, many limitations are less obvious and can only be established with effort.

(5) fan nozzle discharge static pressure water manometer

To a great extent, assessment of the potential improvability of a system depends on detailed knowledge of the basic design and internal workings of all components of the system. Some failings of a system may be easier to overcome than others, and these phenomena may differ from system to system: what one system can achieve easily may be difficult in another. Changes in dictionary information may result in improvements but may also have an impact on other components; changes of grammar rules or of other procedures relating to linguistic structures carry the risk of a 'ripple effect', where the desired output is achieved but the change causes a new error elsewhere. While many commercial systems permit users to make changes to dictionaries of technical vocabulary, they prohibit alterations to basic vocabulary (i.e. common verbs and nouns) and forbid any changes of basic grammar and structural analysis. Improvements of basic design can be undertaken only by linguists and programmers employed by the vendor.

It is particularly difficult to assess a system's ultimate potential improvability. Many researchers have argued that MT systems based on the direct translation approach have much lower improvability potential than transfer systems. Their argument appears to be borne out by experience with the early Georgetown system, whose computational complexity (Chapter 1) precluded any changes after installation, and with the Systran Russian–English system at the US Air Force,

where for over twenty years expansions and changes to dictionaries improved output quality, but are now found to be more likely to degrade it, due to the ripple effect mentioned above. However, nobody is able to assess these limits in advance, least of all for systems based on the transfer or interlingua approaches where operational experience is very much more recent.

Equally important is the potential extendibility of systems to other subject domains and other languages. Systems designed for one particular subject area are inherently more difficult to extend to other areas, not only because of differences in vocabulary but also because differences of grammar and style constitute different sublanguages (section 8.4). It should be noted that many systems which claim to be applicable to all subject domains have in fact been based on a particular subject area; at present many Japanese microcomputer systems are intended for translation of texts in computer science or electrical engineering. Extendibility to related subject areas could well be simply a matter of extending the dictionaries, although the dangers of homography should not be underestimated. Extendibility to less similar domains not only entails the creation of necessary dictionaries, but it may also require more fundamental changes to grammar rules and basic features of components. It is therefore important to establish the modularity of a system and the support given by developers and vendors for extending and improving systems for their users.

The extendibility of a system is highly dependent on the modularity of its analysis, transfer and generation components. A basic question is whether components developed for one language pair can be applied or easily adapted for another language pair. In theory, transfer and interlingua systems should be more easily extendible than direct systems. However, the claim has yet to be substantiated. A number of basically direct systems have added new pairs of languages without great difficulty (e.g. Systran and Weidner); as yet it is not known whether indirect systems have found it any easier.

9.5.3 Technical assessment

The computational features of the system may be easier to determine than the linguistic ones, and may often have more influence when evaluating MT systems. The particular hardware and software requirements of a system may inhibit some potential purchasers (although the usual advice is that it ought not to). What ought to be more important are the computational limitations of the system itself. There may, for example, be limits on the size of dictionary entries, on the character set, on the number of lexical items in dictionaries, on the length of sentences that can be analysed, on the size of texts that can be processed at any one time. There may be software or hardware features which determine the average processing times, which restrict access to dictionaries, and which reduce compatibility with other computer equipment such as OCRs, word processors or printers. Questions of software reliability and robustness are more difficult to answer, but at least purchasers should be able to reach agreement with developers or vendors about future maintenance of software and hardware.

9.5.4 Assessment of personnel requirements and ease of use

MT systems are usually installed in organisations employing professional translators. Their attitudes to computers, to MT and to changes in work practices are crucial to any successful introduction of an MT system. An important part of any evaluation must therefore be an assessment of how the system will be integrated with other components of the translation process (Figure 9.1 above).

Ideally, the operation should involve persons with computer expertise in natural language processing, linguists with knowledge of computational linguistics, translators with knowledge of MT, and terminologists familiar with computational tools. In practice, systems are run by translators who have to learn its capabilities and limitations by trial and error. Adaptability and a positive attitude are essential. But so too is adequate training. If translators are to become post-editors they have to learn new skills; the revision of MT output differs from the revision of human translation.

How much revision is required is therefore an important consideration; the linguistic evaluation of a system should provide some estimate of the types and frequencies of errors which need correction for the texts to be translated. An excessive amount of post-editing (or interactive correction) lowers morale and increases any antagonism towards MT. The low quality of some systems is known to have been a considerable irritant. If some texts are to be only lightly revised (e.g. if they are needed for information only) then the reluctance of professional translators to put out low-quality work has to be overcome. Whatever procedures are adopted, revisers should be positively involved in improvements to the system, by expanding and updating dictionaries, by proposing changes in grammar rules, etc. Above all, they need good facilities, e.g. split-screen editing, 'user-friendly' tools for correcting and replacing words, for transposing and transforming phrases and sentences, and for checking spelling

The adequacy of the basic 'core' dictionary supplied with the system is a major factor. Translators are likely to be antagonistic if they have to supply dictionary information for large quantities of common vocabulary. Special dictionaries are another matter; but the facilities for entering and updating data must not be complex: in particular, users should not be expected to know linguistic terminology, and differences between, e.g. 'transitive' and 'intransitive', should be illustrated by clear examples. Particularly desirable features are mnemonic codes, automatic generation of verb and noun paradigms, and checks for consistency.

Other facilities which enhance the practical acceptability of systems include the statistical monitoring of usage (sizes of texts, amounts of correction, frequency counts of vocabulary, etc.), and the comprehensiveness and readability of the documentation for the system.

9.5.5 Evaluation of costs and benefits

No MT system will be used, however good its quality, if it cannot compete in terms of cost with human translation. Inevitably, a major feature of any evaluation by potential users and purchasers has to be an assessment of the costs of installing and running a system and of the expected savings to be obtained in comparison

with existing operations. A major overhead with many commercial systems is the need to build up collections of data such as customised dictionaries, but also, for example in one case, corpora of previously translated texts for use as examples: in any case, the efficiency of the system increases with use, so initial cost estimates may be considerably misleading, and a true estimate of MT costs may be impossible to achieve until the decision to invest time and money has already been taken. Costings are, of course, based not on the translation system itself in isolation but on the total operational environment (Figure 9.1).

Typically, the potential purchaser will submit texts for translation, and monitor and measure all production costs up to the final versions of translated texts, identifying all direct and indirect costs. These should include all costs of transcription and transmission (e.g. OCR, word processing), of preparing texts for input (checking, correcting, reformatting, perhaps controlling vocabulary and style), of updating dictionaries, of human interaction (if relevant), of recovering and post-editing texts, and of production and printing. A comparison will be made with the costs of producing the same quality of output by human translators, including all costs of acquisition, preparation, transcription, preliminary reading, dictionary consultation, typing, retyping, revision, proof-reading and printing. It should also be remembered that human translators need breaks, and cannot work intensively for long periods. Ideally the comparisons should involve the same texts with both human translation and human revision, and MT and human revision.

Against the costs should be set the relative benefits, which vary from system to system. Major benefits could include: faster production of translations, higher rates of productivity (which are particularly marked if large volumes of text in one specific subject domain are being translated), greater degrees of consistency in terminology (the importance of this factor may outweigh many perceived drawbacks), simultaneous output in many languages (which may well compensate for high costs of text preparation, e.g. in a 'restricted language' system, section 8.3.1). To these should be added expected benefits from future improvements in the translation system and in other facilities, particularly those for pre-editing, post-editing and formatting. Furthermore, the costs of maintaining dictionaries may well diminish: initially users have to enter large numbers of lexical entries; after some time the number of new items declines (unless, of course, new subject domains are undertaken).

The final stage of operational evaluation should always be a trial of the system in the actual working environment for a reasonable period of time (e.g. six months). Only then should the final decision be made.

9.6 Evaluation by translators

Professional translators are almost invariably the direct users of MT systems; few have yet been developed for non-translators, although more can be expected in the future. It has been made clear already that translators are likely to be closely involved in evaluations of systems for purchase and installation, even if in large organisations they may not make the final decision, which may be taken primarily on grounds of costs and benefits.

Translators' concerns are predominantly with the amount of work involved in pre- and post-editing, and the facilities available for substituting words, transposing phrases and changing texts. The quality of the MT system's output is obviously important; if translators consider that too often there will be more work in revising than in translating from scratch, then they are going to reject the system and oppose its introduction. Gains in productivity have to be substantial, and the revision work must not be tedious. Judgement is however often clouded by translators' attitudes; they have been trained to produce high quality work, sometimes even to improve the clarity and style of original texts. Translators are naturally reluctant to be responsible for what they consider an inferior product. Their instinct is to revise MT output to a quality expected from human translators, and they are as concerned with 'stylistic' quality as with accuracy and intelligibility.

In assessing MT they need to adopt a different attitude, to acknowledge that perfectionism is neither always desirable nor always appreciated, particularly if it results in higher charges. An MT system gives them the option of adjusting 'stylistic' quality to users' needs without sacrificing accuracy and consistency. In evaluating MT, therefore, translators should be less concerned with the fluency of the output than with the amount of revision required to produce documents at agreed levels of quality.

Some translators will undoubtedly decide that the changes in working practices which MT systems bring are not tolerable and they will continue with non-automated methods: the demand for good quality translation means they will not lack employment. Those translators who are prepared to be adaptable and to learn new skills and techniques will constitute a new 'breed' of translators. Nevertheless, adaptability should not be all in one direction: the attitudes and evaluations of translators ought to have much greater impact on the design of MT systems than has often been the case in the past. Too frequently, systems have been developed with little or no consultation with translators, who as principal users could make significant practical contributions.

9.7 Evaluation by recipients

Recipients of translations are interested in how much they have to pay for them, how quickly they have been done, and how acceptable they are in terms of quality and readability. Judgements by recipients ought to be included in evaluations by users and purchasers since, after all, those receiving translations have to see benefits as well.

Cost and speed are the only criteria in the case of MT output which has been revised to the standard of human translation. In theory, an MT system should permit both lower costs and faster production. It is, however, possible that the quality may not be as high as that produced wholly by human translation. The operators of the MT system may decide that the costs of high quality revision cannot be justified. Perhaps the recipient wanted the translation quickly and sufficient time could not be devoted to revision. It is for the recipient to decide whether the advantages of speed outweigh the disadvantages of lower quality.

What levels of quality are acceptable depend on the readership and use of the translation. Accuracy, style and readability may be crucial for instruction manuals and for text intended for general publication; and these will be assessed on the lines indicated earlier (section 9.2.1 above). However, if the recipient is an expert in the subject and only wants to scan the text for basic information a low level may be perfectly adequate, e.g. 'raw' output or lightly edited texts. Omission of articles, wrong prepositions, stylistic awkwardness, and many other types of 'errors' may all be accepted as long as the text is comprehensible and the information being sought can be extracted.

9.8 Further reading

Lehrberger and Bourbeau (1988) provide a thorough discussion of the linguistic evaluation by users, with a check-list of other factors. Dyson and Hannah (1987) give a useful summary of what potential users should look for when evaluating systems. The collection edited by Vasconcellos (1988) includes a number of general discussions of methods for evaluating systems, and the economic aspects are well covered by van Slype (1982).

For the evaluation of particular systems a good source is the proceedings of Aslib conferences, e.g. Lawson (1982), Picken (1985, 1986, 1987, 1988, 1990), Mayorcas (1990). Important evaluations are those of Systran by van Slype (1979) and of Logos by Sinaiko and Klare (1972, 1973). Contributions towards a methodology for MT evaluation have also been made by Melby (1988) and King and Falkedal (1990). Karlgren (1987) discusses the value of raw MT output for various purposes.