Chapter 9

Evaluating MT Systems

9.1 Introduction

How can you tell if an MT system is 'good'? How can you tell which of two systems is 'better'? What do 'good' and 'better' mean in this context? These are the questions that this chapter tries to answer.

In a practical domain like MT, such questions reduce to questions of suitability to users' needs: what is the best and most economical way to deal with the user's translation requirements? In the ideal case, it should be possible to give a simple and straightforward answer to this question in a consumers' magazine. An article in such a magazine would discuss the most important issues with a comparison table displaying the achievements of different MT systems on tests of important aspects such as speed and quality. Unfortunately, the information necessary to make informed judgements is not so readily available, partly because the methods for investigating suitability are not well developed. In reality, MT users can spend quite a lot of money finding out what a system can and cannot do for them. In this chapter we will look at the kind of thing that should matter to potential users of MT systems, and then discuss some existing methods for assessing MT system performance.

As we pointed out in the Introduction (Chapter 1), we think that, in the short term, MT is likely to be of most benefit to largish corporate organizations doing a lot of translation. So we adopt this perspective here. However, most of the considerations apply to any potential user.

9.2 Some Central Issues

The evaluation of MT systems is a complex task. This is not only because many different factors are involved, but because measuring translation performance is itself difficult. The first important step for a potential buyer is to determine the translational needs of her

organization. Therefore she needs to draw up a complete overview of the translational process, in all its different aspects. This involves establishing the size of the translation task, the text type of the material and its form (is it machine readable and if so, according to which standards). It also involves considering organizational issues, e.g. the tasks of each member of staff concerned in some way with translation. With that information at hand she can start to investigate what the consequences of the purchase of an MT system would be. These are some of the factors to keep in mind:

Organizational Changes Incorporating an MT system into the translation process will impact upon both the process and the personnel involved. There will be consequences for system administrators and support staff, but above all for the translators themselves, whose tasks will change significantly. Whereas before they will probably have spent the major part of their time actually translating or editing human translations, they will now find themselves spending a lot of time updating the system's dictionaries and post-editing the results of machine translation. There may also be a need to build automatic termbanks. Translators will need to receive training in order to perform these new tasks adequately.

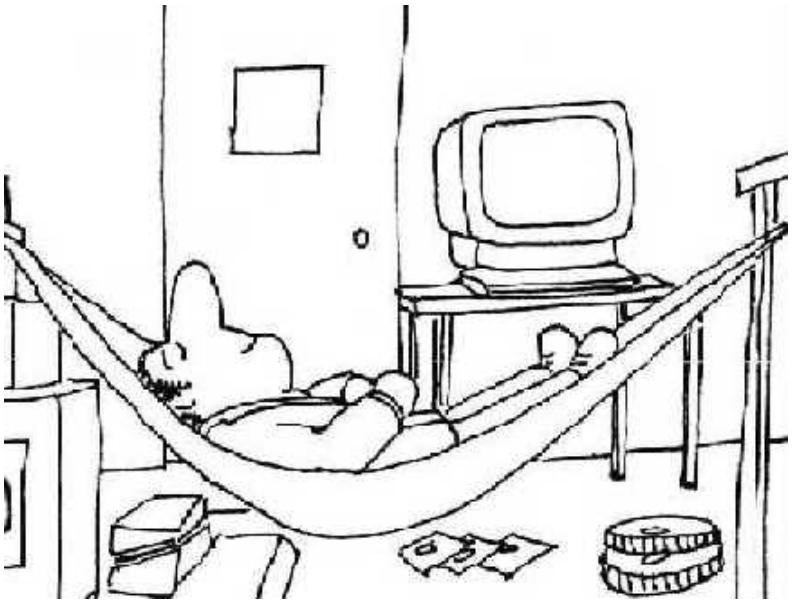
It is important that the personnel support the changeover to MT. They may not always be aware of the fact that MT can lead to more job satisfaction among translators since MT systems are particularly efficient at tedious, repetitive tasks whereas more challenging translation work often still needs to be done by the human translators. If translators in an organization have decided for some reason or other that they do not want to work with MT, imposing it on them is *guaranteed* to produce poor results.

Technical environment We have emphasised right from the start that success depends in part on MT being effectively incorporated as part of a wider document preparation process inside an organization. Smooth handling of text throughout the whole process will prevent unnecessary delays. The MT engine and the document system may well come from different suppliers but they must adhere to the same standards and formats for textual material.

Bear in mind that good document preparation facilities in themselves can improve translator productivity. A decade or so ago much of the productivity increase claimed by some vendors of smaller MT systems could be attributed to their providing rather good multi-lingual word processing facilities, at a time when many translators used only an electric typewriter. Some MT vendors still supply a whole MT system package where the engine is inextricably wrapped up with some specialised word processing and text-handling tool unique to that particular system. This is undesirable on two counts: first, if you are already familiar with a good multi-lingual word processor, little is gained by having to learn another which does much the same things; second, it is likely that an MT vendor's home-grown text-processing facilities will be inferior to the best independent products, because most of the effort will have gone into developing the translation engine.

Status of Vendor Buying an MT system is a considerable investment, and the stability and future solvency of the vendor is an important consideration. After all, contact with the vendor is ideally not just limited to the initial purchase of the system. A

solvent vendor can provide installation support and training in the early stages, and general support and updates later, which may improve performance considerably (e.g. specialized dictionaries, or new language pairs which can be integrated into the existing MT set-up).



Key Issues in the Evaluation of MT Systems:
The Importance of After Sales Support

Engine Performance: Speed In some circumstances, the speed at which the engine churns out raw translated text won't actually be crucial. If the system requires interaction with the translator whilst it is translating, then of course it should not amble along so slowly as to to keep the translator waiting all the time. But if it is functioning without direct interaction, it can proceed at its own pace in the background whilst the translator gets on with other jobs such as post-editing or hand translation of difficult material. This aspect also depends on the user's translational needs: if the user's material requires 15 hours daily on a fast MT system and 20 on a slower one, no one will notice the difference if the system is running overnight. Of course, there are situations where the quick delivery of translation output is essential. (The agronomist in Chapter 2, who wants to process very large quantities of material to a low level may be an example.) But in general, slow speed is the one component of MT performance of which upgrading is relatively easy: by buying some faster hardware for it to run on.

Engine Performance: Quality This is a major determinant of success. Current general purpose commercial MT systems cannot translate all texts reliably. Output can sometimes be of very poor quality indeed. We have already mentioned that the post-editing task (and with it the cost) increases as translation quality gets poorer. In

the worst case, using MT could actually increase translation costs by tying up translators in editing and maintenance tasks, ultimately taking up more time than would have been required to produce translations entirely by hand. Because of its enormous influence on the overall translation cost, translation quality is a major aspect in MT evaluation.

9.3 Evaluation of Engine Performance

Substantial long-term experience with particular MT systems in particular circumstances shows that productivity improvements and cost-savings actually achieved can be very variable. Not all companies can apply MT as successfully as the following:

> In the 1980s, Perkins Engines was achieving reported cost savings of around £4000 for each diesel engine user manual translated on a PC-based WEIDNER MT system. Moreover, overall translation time per manual was more than halved from around 26 weeks to 9-12 weeks. Manuals were written in Perkins Approved Clear English (cf. Chapter 8).(Pym, 1990, pages 91-2)

Different organizations experience different results with MT. The above examples indicate that the kind of input text is one of the important factors for getting good results. A sound system evaluation is therefore one which is executed within the company itself. An MT vendor might provide you with translated material which shows what their system can do. There is, however, no guarantee that the system will do the same in a different company setting, with different texts. Only a company specific evaluation will provide the client with the feedback she ultimately wants. Information provided by the MT vendor can be useful though, e.g. if system specifications indicate what sort of text type it can or cannot handle or what sort of language constructions are problematic for their system.

In evaluating MT systems one should also take into account the fact that system performance will normally improve considerably during the first few months after its installation, as the system is tuned to the source materials, as discussed in Chapter 2. It follows that performance on an initial trial with a sample of the sort of material to be translated can only be broadly indicative of the translation quality that might ultimately be achieved after several months or years of work.

Something similar holds for those stages of the translation process which involve the translator, like dictionary updating and post-editing of the output. Times needed for these tasks will reduce as translators gain experience.

So how do we evaluate a system? Early evaluation studies were mainly concerned with

the quality of MT. Of course, assessing translation quality is not just a problem for MT: it is a practical problem that human translators face, and one which translation theorists have puzzled over. For human translators, the problem is that there are typically many possible translations, some of them faithful to the original in some respects (e.g. literal meaning), while others try to preserve other properties (e.g. style, or emotional impact).[1]

In MT, the traditional transformer architecture introduces additional difficulties, since its output sentences often display structures and grammar that are unknown to the target language. It is the translator's task to find out what the correct equivalent is for the input sentence and its ill-formed translation. And, in turn, the evaluator's task is to find out how difficult the translator's task is.

In the rest of this chapter we will describe the most common evaluation methods that have been used to date and discuss their advantages and disadvantages.

9.3.1 Intelligibility

A traditional way of assessing the quality of translation is to assign scores to output sentences. A common aspect to score for is **Intelligibility**, where the intelligibility of a translated sentence is affected by grammatical errors, mistranslations and untranslated words. Some studies also take style into account, even though it does not really affect the intelligibility of a sentence. Scoring scales reflect top marks for those sentences that look like perfect target language sentences and bottom marks for those that are so badly degraded as to prevent the average translator/evaluator from guessing what a reasonable sentence might be in the context. In between these two extremes, output sentences are assigned higher or lower scores depending on their degree of awfulness — for example, slightly fluffed word order ("*... in an interview referred Major to the economic situation...*" will probably get a better score than something where mistranslation of words has rendered a sentence almost uninterpretable ("*...the peace contract should take off the peace agreement....*). Thus scoring for intelligibility reflects directly the quality judgment of the user; the less she understands, the lower the intelligibility score. Therefore it might seem a useful measure of translation quality.

Is there any principled way of constructing an intelligibility scoring system? Or rather is there any generally agreed, and well motivated scoring system? We do not know of any. The major MT evaluation studies which have been published report on different scoring systems; the number of points on the scoring scales ranging from 2 (intelligible, unintelligible) to 9. The 9 point scale featured in the famous ALPAC Report and was not just used to score the intelligibility of MT, but also of human translation. As a consequence the scale included judgments on fairly subtle differences in e.g. style. This scale is relatively well-defined and well-tested. Nevertheless we think that it is too fine-grained for MT evaluation and leads to an undesirable dispersion of scoring results. Also, we think that style should not be included because it does not affect the intelligibility of a text. On the other hand, a two point scale does not give us enough information on the seriousness of those

[1] For an excellent discussion of the range of aspects that a good translation may need to take into account, see Hatim and Mason Hatim and Mason (1990).

errors which affect the intelligibility. (A two point scale would not allow a distinction to be drawn between the examples in the previous paragraph, and complete garbage, (or something completely untranslated) and a fully correct translation.) Perhaps a four point scale like the one below would be more appropriate.

An Example Intelligibility Scale

1 The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.

2 The sentence is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means.

3 The general idea of the sentence is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choices.

4 The sentence is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable.

Once devised, scoring scales need to be tested, to make sure that scale descriptions are clear and do not contain any expression that can be interpreted differently by different evaluators. The test procedure should be repeated until the scale descriptions are uniformly interpreted by evaluators.

A reasonable size group of evaluators/scorers must be used to score the MT output. Four scorers is the minimum; a bigger group would make the results more reliable. The scorers should be familiar with the subject area of the text they will score and their knowledge of the source language of the translation should also be good. Before an official scoring session is held the scorers participate in a training session in which they can become acquainted with the scale description. This training session should be similar for all scorers. During scoring it should be impossible to refer to the source language text.

9.3.2 Accuracy

By measuring intelligibility we get only a partial view of translation quality. A highly intelligible output sentence need not be a correct translation of the source sentence. It is important to check whether the meaning of the source language sentence is preserved in the translation. This property is called **Accuracy** or **Fidelity**. Scoring for accuracy is normally done in combination with (but after) scoring for intelligibility.

As with intelligibility, some sort of scoring scheme for accuracy must be devised. Whilst it might initially seem tempting to just have simple ‘Accurate’ and ‘Inaccurate’ labels, this

could be somewhat unfair to an MT system which routinely produces translations which are only slightly deviant in meaning. Such a system would be deemed just as inaccurate as an automated 'Monty Python' phrasebook which turns the innocent request *Please line my pockets with chamois* [2] into the target language statement *My hovercraft is full of eels*. Obviously enough, if the output sentence is complete gobbledegook (deserving of the lowest score for intelligibility) then it is impossible to assign a meaning, and so the question of whether the translation means the same as the original cannot really be answered. (Hence accuracy testing follows intelligibility rating).

The evaluation procedure is fairly similar to the one used for the scoring of intelligibility. However the scorers obviously have to be able to refer to the source language text (or a high quality translation of it in case they cannot speak the source language), so that they can compare the meaning of input and output sentences.

As it happens, in the sort of evaluation considered here, accuracy scores are much less interesting than intelligibility scores. This is because accuracy scores are often closely related to the intelligibility scores; high intelligibility normally means high accuracy. Most of the time most systems don't exhibit surreal or Monty Python properties. For some purposes it might be worth dispensing with accuracy scoring altogether and simply counting cases where the output looks silly (leading one to suppose something has gone wrong).

It should be apparent from the above that devising and assigning quality scores for MT output — what is sometimes called 'Static' or 'Declarative Evaluation'[3] — is not straightforward. Interpreting the resultant scores is also problematic.

It is virtually impossible — even for the evaluator — to decide what a set of intelligibility and accuracy scores for a single MT system might mean in terms of cost-effectiveness as a 'gisting' device or as a factor in producing high quality translation. To see this, consider the sort of quality profile you might get as a result of evaluation (Figure 9.1), which indicates that most sentences received a score of 3 or 4, hence of middling intelligibility. Does that mean that you can use the system to successfully gist agricultural reports? One cannot say.

Turning to the high-quality translation case, it is clear that substantial post-editing will be required. But it is not clear — without further information about the relationship between measured quality and post-editing times — what effect on overall translator productivity the system will have. Whilst it is presumably true that increasingly unintelligible sentences will tend to be increasingly difficult to post-edit, the relationship may not be linear. For example, it may be that sorting out minor problems (which don't affect intelligibility very much) is just as much of an editing problem as correcting mistranslations of words (which affect intelligibility a great deal). We could for example imagine the following two sentences to be part of our sample text in Chapter 2. The first one is more intelligible than the

[2] This comes from the section on 'Talking to the Tailor' in an English-Italian phrasebook of the 1920s.

[3] 'Declarative' here is to be contrasted with 'procedural'. A declarative specification of a program states what the program should do, without considering the order in which it must be done. A procedural specification would specify both what is to be done, and when. Properties like Accuracy and Intelligibility are properties of a system which are independent of the dynamics of the system, or the way the system operates at all — hence 'non-procedural', or 'declarative'.

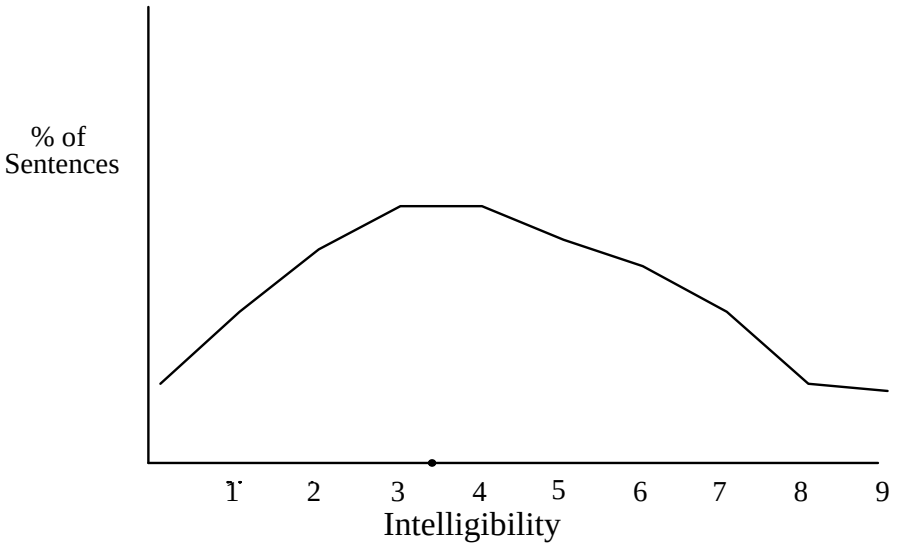


Figure 9.1 Typical Quality Profile for an MT System

second, yet more time will be needed to fix the errors in it:

(1) a. The print□ page should be from excell ing quality,
b. The printed page should his excellent quality.

It is true that a comparative evaluation of a number of different MT systems might demonstrate that one system is in all respects better than the others. The information however does not tell us whether buying the better MT system will improve the total translation process — the system could still be unprofitable. And even if two particular systems have different performance profiles, it may not always be clear whether one profile is likely to be better matched to the task in hand than the other. For example, look at the intelligibility ratings for systems A and B in Figure 9.2. For system A the majority of sentences are neither very good nor bad (rating 3 or 4). System B, by comparison, tends to do either quite well (scores of 7 are common) or quite badly (scores 1, and 2 are frequent). Which system will be better in practice? It is not possible to say.

9.3.3 Error Analysis

Rather than using broad indicators as guides to score assignments, you could focus on the errors the MT system makes. The technique of error analysis tries to establish how seriously errors affect the translation output.

The method is this. To start off, write down a large list of all the types of errors you think the MT system might make. During the evaluation, all the errors in the translated text are counted up. Because you consider some errors more serious than others, each type of error will be multiplied by some *weighting factor* which you assign to it. The score then for each individual sentence or the whole text will be the sum of all the weighted errors. So, if we

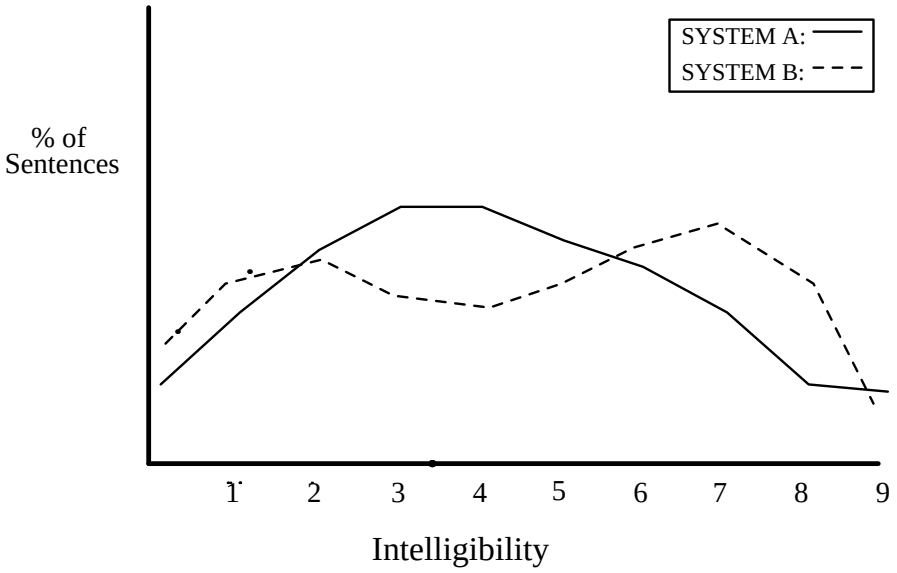


Figure 9.2 Which Performance Curve is Better?

take the raw translation we were using in the scenario in Chapter 2 as an example, error analysis might work as follows.

For the example three sorts of error are identified. These three sorts are errors involving selection of *a* vs *one* as the translation of German *ein*, errors in number agreement (e.g. **a computers*), and errors in the selection of prepositions. Using some short codes for each error type, each error occurrence is marked up in the raw output. The resulting marked text is given below.

To calculate the seriousness of the errors, weights in the range 0 to 1 are assigned to the three error types. The weight for an error in preposition selection is higher than that for incorrect number because the person responsible considers that incorrect number is relatively less serious. This is summarized in the following table.

ERROR TYPE	WEIGHT
a/*one* selection	0.4
Number	0.2
Preposition	0.6

On the basis of this the total error score can be calculated. There are two errors in NUMber agreement, two involving PREPositions, and one involving A/ONE selection, so the score is: $(2 \times 0.2) + (2 \times 0.6) + (1 \times 0.4) = 2$

Although this method gives more direct information on the usefulness of an MT system, there are immediate problems with using detailed error analysis. The first is practical: it

Markup of Errors

Adjustment of the print density:

- Turn the button an A/ONE or two positions in direction of the dark indicator.
- Switch off the printer for a moment and then again a PREP, so that the test page is printed.
- Repeat the two previous steps as long as, until you see Gray on the background of the page, similarly like at PREP easily unclean copies of a photocopier.
- Turn back the button a position.

Now you can connect the printer to the computer.

If you connect the printer to a Macintosh computers NUM, continue with the instructions in the chapter 3. If you use an other computer, continue with chapters NUM 4.

will usually require considerable time and effort to train scorers to identify instances of particular errors — and they will also need to spend more time analysing each output sentence. Second, is there any good basis for choosing a particular weighting scheme? Not obviously. The weighting is in some cases related to the consequences an error has for post-editing: how much time it will take to correct that particular mistake. In some other cases it merely reflects how badly an error affects the intelligibility of the sentence. Consequently, the result will either indicate the size of the post-editing task or the intelligibility of the text, with its relative usefulness. In both cases devising a weighting scheme will be a difficult task.

There is, however, a third problem and perhaps this is the most serious one: for some MT systems, many output sentences are so corrupted with respect to natural language correlates that detailed analysis of errors is not meaningful. Error types are not independent of each other: failure to supply any number inflection for a main verb will often mean that the subject and verb do not agree in number as required. It will be difficult to specify where one error starts and another ends and thus there is the risk of ending up with a general error scale of the form *one, two, ... lots*. The assignment of a weighting to such complex errors is thus a tricky business.

9.3.4 The Test Suite

As we noted before, for some years the trend (at least in research circles) has been towards Translation Engines with substantial linguistic knowledge in the form of grammars. LK Engines have a different performance profile from Transformer Engines in that their output will tend to contain rather fewer badly degraded sentences. (Perhaps at the price of failing to produce *anything* in some cases).

Although the use of linguistic-knowledge based techniques tends to promote higher Intelligibility (and Accuracy) output, it is possible that the linguistic knowledge embedded in the system is defective or incomplete. Sometimes a certain grammar rule is too strict or too general to apply correctly in all circumstances; sometimes the rules that handle one phenomenon (e.g. modal verbs like *may* in *The printer may fail*) and the rules that handle another phenomenon (eg. negation) fail to work correctly together when the two phenomena co-occur or interact in a sentence. (For example, imagine the problems that will result if *The printer can not be cleaned* (i.e. can be left uncleaned), and *The printer cannot be cleaned* (i.e. must not be cleaned) are confused.)

Keeping track of these sorts of constructional errors and deficits has become rather a severe problem for developers of MT systems and other large NLP systems. For example, while running the system on a corpus of test texts will reveal many problems, many potential areas of difficulty are hidden because the statistics are such that even quite large corpora will lack even a single example of particular grammatical combinations of linguistic phenomena.

Rather than churning through increasingly large 'natural' text corpora, developers have recently turned their attention to the use of suites of specially constructed test sentences. Each sentence in the suite contains either one linguistic construction of interest or a combination thereof. Thus part of an English test suite might look as follows.

This fragment just churns through all combinations of modal verbs like *can*, *may* together with optional *not*. In practice, one would expect test suites to run to very many thousands of sentences, because of the many different combinations of grammatical phenomena that can occur. Suites may include grammatically unacceptable sentences (e.g. **John not run*) which the parser should recognize as incorrect. In systems which use the same linguistic knowledge for both analysing and synthesising text, the fact that an ill-formed sentence is rejected in analysis suggests that it is unlikely to be constructed in synthesis either.

Nobody knows for sure how test suites should be constructed and used in MT. A bi-directional system (a system that not only translates from German to English *and* from English to German) will certainly need test suites for both languages. Thus success in correctly translating all the sentences in a German test suite into English and all the sentences in an English test suite into German would definitely be encouraging. However, standard test suites are rather blunt instruments for probing translation performance in the sense that they tend to ignore typical differences between the languages involved in translation.

Extract from a Test Suite

John runs.	
John will run.	*modal auxiliaries*
John can run.	
John may run.	
John should run.	
John could run.	
John does not run.	*negation (with do-support)*
John not run.	
John will not run	*negation and modal auxiliaries*.
John can not run.	
John may not run.	
John should not run.	
John could not run.	
....	

We can look at an example. In English the perfect tense is expressed with the auxiliary verb *have*, like in *He has phoned*. In German however there are two auxiliary verbs for perfect tense: *haben* and *sein*. Which verb is used depends on the main verb of the sentence: most require the first, some require the second. So an English and a German test suite designed to check the handling of perfect tense will look different.

The German test suite thus tests the perfect tense for verbs that take *sein* and verbs that take *haben* and therefore have to test twice the number of sentences to test the same phenomenon. However, if *He has phoned* is correctly translated into German *Er hat angerufen*, then we still can not be sure that all perfect tenses are translated correctly. For testing of the English grammar alone, there is no reason to include a sentence like *He has gone* into the English test suite, since the perfect tense has already been tested. For translation into German however it would be interesting to see whether the auxiliary verb sein is selected by the main verb *gehen*, giving the correct translation *Er ist gegangen*.

Given this sort of problem, it is clear that monolingual test suites should be supplemented with further sentences in each language designed to probe specific language pair differences. They could probably be constructed by studying data which has traditionally been presented in books on comparative grammar.[4]

In a bi-directional system, we need test suites for both languages involved *and* test suites probing known translational problems between the two languages. Constructing test suites is a very complicated task, since they need to be complete with regard to the phenomena

[4]It would be nice to try to find possible problem areas by some sort of automatic scanning of bilingual texts but the tools and techniques are not available to date.

Part of English-German Test Suite

English:

....
He has phoned.
He had phoned.
...

German:

....

Er ist gegangen.	*sein*
Er hat angerufen.	*haben*
Er war gegangen.	*sein*
Er hatte angerufen.	*haben*

...

occurring in the present and future input texts of the MT user. Thus one should first check whether there are any existing test suites for the languages that need to be tested. (There are several monolingual test suites around). Such a suite can be modified by adding material and removing restrictions that are irrelevant in the texts for which the system is intended (eg. the texts to be translated might not contain any questions). As far as we know there are no readily available test suites for translational problems between two languages; to test for this, the evaluator will have to adapt existing monolingual ones.

Once the test suites have been devised they are run through the system and an inventory of errors is compiled. Clearly the test suite is an important tool in MT system development. How useful will it be for a *user* of MT systems?

It is of course possible for the user to run an MT system on a test suite of her own devising and, in some cases, this may be perfectly appropriate. It is especially useful to measure improvements in a system when the MT vendor provides a system update. However, the test suite approach does entail some drawbacks when used to assess system performance in comparison with competing systems. The problem is familiar by now: how are the evaluation results to be interpreted? Suppose System A and System B both produce acceptable translations for 40% of the test sentences and that they actually fail on different, or only partially overlapping, subsets of sentences. Which one is better? If System B (but not System A) fails on test sentences which embody phenomena with very low frequencies in the user's type of text material, then clearly System B is the better choice. But users typically do not have reliable information on the relative frequencies of various types of constructions in their material, and it is a complex task to retrieve such information by going through texts manually (automated tools to do the job are not yet widely available).

The same problem of interpretability holds when MT systems are evaluated by an indepen-

dent agency using some sort of standard set of test suites. Published test suite information certainly gives a much better insight into expected performance than the vague promisory notes offered with current systems; but it doesn't immediately translate into information about likely performance in practice, or about cost effectiveness.

On top of this there is the problem of how to design a test suite, and the cost of actually constructing it. Research is ongoing to determine what sort of sentences should go into a test suite: which grammatical phenomena should be tested and to what extent should one include co-occurrence of grammatical phenomena, should a test suite contain sentences to test semantic phenomena and how does one test translation problems? These and additional problems might be solved in the future, resulting in proper guidelines for test suite construction.

9.4 Operational Evaluation

In the previous sections we have discussed various types of quality assessment. One mayor disadvantage of quality assessment for MT evaluation purposes, however, is the fact the overall performance of an MT system has to be judged on more aspects than translation quality only. The most complete and direct way to determine whether MT performs well in a given set of circumstances is to carry out an operational evaluation on site comparing the combined MT and post-editing costs with those associated with pure human translation. The requirement here is that the vendor allows the potential buyer to test the MT system in her particular translation environment. Because of the enormous investment that buying a system often represents, vendors should allow a certain test period. During an operational evaluation a record is kept of all the user's costs, the translation times and other relevant aspects. This evaluation technique is ideal in the sense that it gives the user direct information on how MT would fit in and change the existing translation environment and whether it would be profitable.

Before starting up the MT evaluation the user should have a clear picture of the costs that are involved in the current set-up with human translation. When this information on the cost of the current translation service is available the MT experiment can begin.

In an operational evaluation of MT time plays an important role. Translators need to be paid and the more time they spend on post-editing MT output and updating the system's dictionaries, the less profitable MT will be. In order to get a realistic idea of the time needed for such translator tasks they need to receive proper training prior to the experiment. Also, the MT system needs to be tuned towards the texts it is supposed to deal with.

During an evaluation period lasting several months it should be possible to fully cost the use of MT, and at the end of the period, comparison with the costs of human translation should indicate whether, in the particular circumstances, MT would be profitable in financial terms or not.

One problem is that though one can compare cost in this way, one does not necessarily hold quality constant. For example, it is sometimes suspected that post-edited MT translations tend to be of inferior quality to pure human translations because there is some temptation to post-edit only up to that point where a correct (rather than good) translation is realised. This would mean that cost benefits of MT might have to be set against a fall in quality of translation. There are several ways to deal with this. One could e.g. use the quality measurement scales described above (Section 9.3.1). In this case we would need a fine-grained scale like in the ALPAC Report, since the differences between post-edited MT and HT will be small. But what does this quality measurement mean in practice? Do we have to worry about slight differences in quality if after all an 'acceptable' translation is produced? Maybe a better solution would be to ask an acceptability judgment from the customer. If the customer notices a quality decrease which worries him, then clearly post-editing quality needs to be improved. In most cases, however, the experienced translator/post-editor is more critical towards translation quality than the customer is.

In general it seems an operational evaluation conducted by a user will be extremely expensive, requiring 12 personmonths or more of translator time. An attractive approach is to integrate the evaluation process in the normal production process, the only difference being that records are kept on the number of input words, the turnaround time and the costs in terms of time spent in post-editing. The cost of such an integrated operational evaluation is obviously less. After all, if the system is really good the translation costs will have been reduced and will compensate for some of the costs of the evaluation method. (On the other hand, if the system is not an improvement for the company, the money spent on its evaluation will be lost of course.)

9.5 Summary

The purchase of an MT system is in many cases a costly affair and requires careful consideration. It is important to understand the organizational consequences and to be aware of the system's capacities. Unfortunately, it is not possible to draw up a comparison table for MT systems on the basis of which MT buyers could choose their system. Although system specifications can provide us with some useful information there are too many aspects which influence the performance of MT that cannot be included in such a table. Furthermore, MT will perform differently in different translation environments, depending mainly on the character of the typical input texts. Without having the necessary information of the kind of input texts the user has in mind, it is not possible to make a reliable prediction about the cost effectiveness of an MT system. The consequences are that if we want information about an MT system we have to evaluate it, and that this evaluation has to be specifically for the user's translational needs.

The evaluation strategies discussed in this chapter are strategies that a buyer might want to pursue when considering the purchase of an MT system. Although they will provide the client with a certain amount of useful information, each method has some drawbacks, which we have tried to point out in our discussion.

9.6 Further Reading

Useful discussion of evaluation methods can be found in van Slype (1982), and Lehrberger and Bourbeau (1987). Practical discussion of many different aspects of MT evaluation can be found in King and Falkedal (1990), Guida and Mauri (July 1986), and Balkan et al. (1991).

A special issue of the Journal *Machine Translation* is dedicated to issues of evaluation of MT (and other NLP) systems. The introduction to the issue, Arnold et al. (in press b), gives an overview of the state of the issues involved, going into more detail about some issues glossed over here. Several of the articles which appear in this issue report practical experience of evaluation, and suggest techniques (for example, Albisser (in press); Flank et al. (in press); Jordan (in press); Neal et al. (in press).)

The problems of focusing evaluation on the MT engine itself (i.e. apart from surrounding peripherals) are discussed in Krauwer (in press).

As things stand, evaluating an MT system (or other NLP system) involves a great deal of human activity, in checking output, for example. A method for automating part of the evaluation process is described in Shiwen (in press).

Some of the issues involved in construction of test suites are discussed in Arnold et al. (in press a), and Nerbonne et al. (in press).

In this chapter, we have generally taken the users' perspective. However, evaluation is also an essential for system developers (who have to be able to guage whether, and how much, their efforts are improving a system). How evaluation technique can be applied so as to aid developers discussed in Minnis (in press).

One of the best examples of MT evaluation in terms of rigour was that which formed the basis of the ALPAC report Pierce and Carroll (1966), which we mentioned in Chapter 1 (it is normal to be rude about the conclusions of the ALPAC report, but this should not reflect on the evaluation on which the report was based: the evaluation itself was a model of care and rigour — it is the interpretation of the results for the potential of MT which was regrettable).

See (Nagao, 1986, page 59) for more detailed scales and criteria for evaluating fidelity and ease of understanding.

As usual, Hutchins and Somers Hutchins and Somers (1992) contains a useful discussion of evaluation issues (Chapter 9).