

Machine Translation: Matching Reality to Expectations

• JOANN P. RYAN

Although operational machine translation (MT) has been in existence for over 20 years, few persons have had any practical, day-to-day experience with MT. Thus the decision to introduce MT into an organization may be made without the careful planning that the introduction of a new technology requires, and the process of successfully integrating the new MT system into the office environment may be delayed because of the unrealistic expectations of management and other personnel.

While management may think that MT is a quick and automatic solution to all of their translation problems, translators often feel that this new technology will never work and may even make their lives more difficult. Both over-optimism and over-pessimism can be fatal to the success of an MT project. On the one hand, the overly optimistic expectations of managers may mean that insufficient personnel and planning time are allotted to the project; they may also lead to premature cancellation of a project before it has had a chance to prove itself. On the other hand, the low expectations of translators may mean a lack of interest on their part at the time when their input and commitment are crucial to the success of the project.

Maria Russo of the Xerox Corporation, which has been using the Systran MT system since 1978 for the translation of technical manuals, has prepared the following chart which illustrates this point:

Fortunately in the case of Xerox, over time and with growing experience, the expectations of managers and linguists have converged in the direction of reality, as MT has been successfully integrated into the company's day-to-day operations. In the following pages you will find some practical suggestions that will help you and your organization to develop realistic expectations about what MT can and should do for you and to make sure that the reality of your MT application matches and even exceeds these expectations.

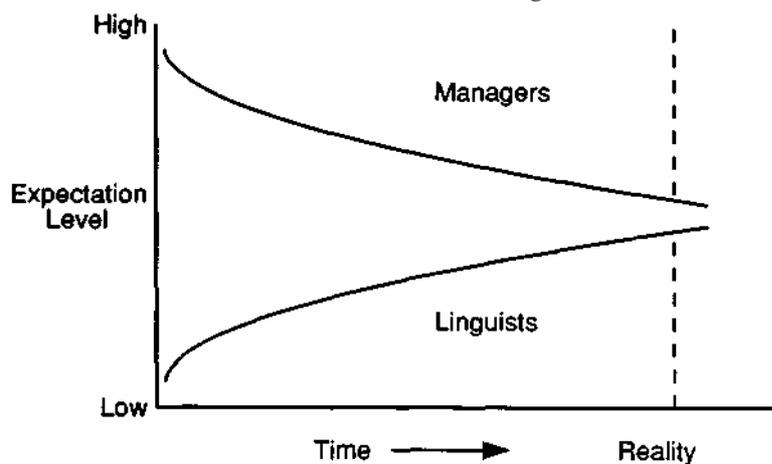


Figure 1: Expectation versus reality of computer translation. (Source: Maria Russo)

1 Six Steps to Developing Realistic Expectations

These six steps will help you and your organization to develop a realistic approach to MT utilization.

1. Ask Yourself the Right Questions.
2. Explore All the Options for Using MT.
3. Ask the MT Vendors the Right Questions.
4. Choose the Option That Best Fits Your Needs.
5. Plan to Customize the System to Meet Your Needs.
6. Involve In-House Personnel at Every Stage of the Process.

2 What to Ask Yourself

Taking the time to ask yourself some important questions will help you to determine whether MT will be useful to your organization. If you have answered all of these questions carefully, you will have a good basis for deciding whether or not MT is the answer to your translation problems. At the same time, you will have taken the first step toward finding the type of MT system that matches your needs.

1. **Does your organization have information needs that are not being met because translations cannot be produced in a timely manner?** This is the key to determining whether MT will be useful to your organization, because timely translation is one of the biggest advantages that a well-designed and well-managed MT system can offer. In today's increasingly fast-paced world, information value quickly plummets over time. Scientific and technical information can be outdated in a matter of weeks, and many librarians and information specialists report that requests for translations have decreased

dramatically because they are unable to obtain the translations within a reasonable time period. Companies that experience delays in getting their products into foreign markets because of problems in translating documentation may find themselves losing out to competitors who have learned to manage their translation needs more effectively.

- 2. What types of documents does your organization need to translate?** It is commonly believed that only long, repetitive technical documents are successfully translated by machine translation. While it is true that such documents are the ones that are translated most cost-effectively, translation of technical manuals is only one of the uses of MT. Of increasing importance, both to governments and businesses, is the translation of foreign documents to keep abreast of advances in science and technology in other countries. In general, any document that is reasonably well-written in a non-conversational style is a possible input to MT. However, don't plan on translating everything that comes into your office with MT. Start by identifying those documents that are most suitable for MT. Your MT vendor can help you with this. In general, look for machine-readable or machine-scannable documents that contain repetitive terminology, are well-defined by subject field, and, above all, are clearly written in a simple style with few grammatical and spelling errors. The important point is to determine your needs and then find out if there is a system or approach that will meet them in a cost-effective way.
- 3. What style of output do your end users require?** Clearly this depends upon the purpose of the translation. A user who must scan large volumes of text very quickly to determine its information content will find MT to be a very cost-effective way to do this, as use of raw, unedited MT can mean as much as an 80% savings in cost, compared to roughly 50% for post-edited MT. The Foreign Technology Division (FTD) of the U.S. Air Force has been using this approach to MT for over 20 years. The resulting raw translation may be used as is, sent to a post-editor for polishing, or sent to a human translator, depending upon the needs of the user and his knowledge of the subject field. At the other end of the spectrum, many users have documents that must be of publishable quality, and here you must be sure to take into account the costs of post-editing and possibly pre-editing. In between are users who require complete accuracy but do not demand an elegant prose style, and in this case the technique of rapid post-editing, as used by large organizations like the U.S. Air Force and the CEC, can save both time and money.
- 4. Are your documents in machine-readable form?** This is a big plus for anyone planning to use MT. Otherwise, if you have a reasonably large volume of text, you will want to purchase a high-quality scanner, which can double or even triple the speed of the document input process for documents written in the major European languages. Even if your documents are already in machine-readable form, you will still have to make sure that the MT system that you plan to use has an interface for your document format. Otherwise, you may have to convert your documents to an ASCII format and lose all of your formatting codes. A more desirable alternative for an organization with large volumes of text is to contract with the MT vendor to write a customized interface if one is not already available for your documents. Large organizations may find it advantageous to look into the Xerox line of document processing software, which includes the DocuTran interface between their GlobalView publishing software and the Systran translation system, and to look for MT software that will interface with SGML, the government-mandated standard markup language for compliance with

CALS. Remember that preserving the graphics and formatting codes of your documents throughout the translation process is the area where a large organization will realize the biggest savings through MT!

5. **Are you willing to make a long-term commitment to MT?** As with the introduction of any new computer technology, you must be prepared to put a lot of initial effort into customizing the MT system to meet your needs and integrating the new process into your office environment. But if you have planned your MT project carefully, your MT system can bring you benefits from the first day of implementation and, with close attention to the care and feeding of the system, you can be sure that these benefits will continue to increase over time.
6. **Are you willing to provide feedback to MT developer?** This may not be an issue if you have purchased an inexpensive software package, but it's very important if you are using one of the large, general-purpose MT systems like Systran or Logos. It will take some extra time to do this, but the more you make your specific needs known to the developer of your MT software, the more likely it is that future versions of the system will incorporate the features that you need. To be successful, a commercial MT developer must market a system that meets the needs of real-world users, so input from your organization will be welcomed.
7. **Do you have personnel who are excited about the prospect of working with MT?** This is one of the most important questions to ask yourself when deciding whether or not to introduce MT into your organization. The manager who imagines that MT is going to replace a significant percentage of his work force is making the mistake of unrealistic expectations mentioned earlier. MT should be viewed as a tool to produce cost-effective translations by enhancing translator productivity, and thus the importance of the human factor in making MT work for you cannot be overestimated. You are going to need people to perform various MT-related tasks, such as dictionary building, post-editing, providing feedback to the developer, and possibly pre-editing, and whether or not these people are in-house personnel or outside contractors, their enthusiasm for and dedication to working with this new technology will be the key to the success of your application.

3 Options for Using MT

Here is an outline of the options currently available to an organization that is interested in using MT:

1. Send Document to Translation Bureau That Uses MT

- No initial investment
- Good for occasional long document
- Translation bureau will use MT only if appropriate
- Translation bureau handles all input and post-editing, and has post-editors experienced in editing MT
- User usually provides terminology, but translation bureau creates dictionary

- Several MT vendors offer a translation service, but with no or limited post-editing

2. Purchase PC-Based Software

- Low one-time cost
- No major hardware investment
- Widening range of products and prices
- Cheaper packages not suitable for longer documents
- Cost varies with degree of linguistic sophistication
- Tradeoff between linguistic sophistication and amount of post-editing required

3. Access MT System Online

- No major hardware investment
- User pays only cost per word plus access charge
- Good for lower volume and shorter texts
- User has access to large dictionaries but can still customize system with own terminology
- Relatively cheap way to access multiple language pairs
- Good way to try out a general-purpose MT system

4. Lease MT System

- Major hardware investment unless hardware already in place
- Cost-effective for large volumes of translation (over 300 pages per month)
- Better customer support
- User control over entire translation process
- Maximum savings from automation of entire document production process

4 What to Ask the MT Vendor

Now that you have an idea of the various options available to you, your next step is to approach several MT vendors and to start asking a lot of questions. Ideally you should give the vendor several pages representative of the type of text you would like to translate and ask to see a raw translation run without any special dictionary update. This will give you an idea of the suitability of the raw MT output for your purposes and how much post-editing will be required. Since you or the vendor will almost certainly be updating the system dictionaries with your specialized terminology, you should then ask that the same text be re-run after the dictionary has been updated with a few of your terms; if you will be creating your own user-specific dictionary, ask to see how this process is done and make sure that you see the text after a user-specific dictionary update (as opposed to a system dictionary update).

The following list of questions is meant to serve as suggestions to help you in your fact-finding session. Some questions may not be appropriate for the type of system that you are looking into or the type of application you are interested in. You should not expect every

vendor to offer every type of service or to answer yes to every question. The idea is to focus on the items that are of greatest interest to you and to make sure that you get the system that best meets your needs.

Type of System: Is your system a general-purpose, “try-anything” system or is it best suited to translating specific domains of interest?

Hardware Requirements: Does your software run on a microcomputer or does it require a mainframe or large workstation? If it does not run on a microcomputer (and most major MT systems sold today still don't), can it be accessed from a PC or other microcomputer via an electronic network? What type of operating system does it run under? How much memory is required for the basic system? How much additional storage for each language pair, including dictionaries?

Services Offered: Is the system leased or purchased? Can the system be accessed online via an electronic network at a cost-per-word charge? Do you offer a translation service for a machine-readable document sent to you on disk? Does this service include post-editing?

Operational Language Pairs: Which language pairs are currently operational? How long has each of these pairs been operational? Which of them currently have satisfied users? What is your standard for determining when a language pair becomes operational? What language pairs are currently under development, and when do you estimate that they will become operational?

Cost: What is the cost for each type of service offered? What is the cost per language pair? Is there a discount on additional language pairs and/or the reverse direction of an initial language pair purchased or leased (e.g., French-English if English-French is purchased). What is the cost of installing the system on a second CPU or at an additional site? Is there an annual maintenance fee? Is there a discount based on number of words translated?

System Dictionaries: Do you provide your customers with a general dictionary? What is the size of this dictionary in single words and in multi-word expressions? What subject fields do your dictionaries cover, and which of these have the best coverage? (Be sure to ask these questions for the specific language pairs in which you are interested, as the answer may differ according to language pair). Are the microglossaries for specific subject fields contained in the system dictionary, or do you offer special technical dictionaries that must be purchased at additional cost? What is the cost of additional specialized dictionaries? Does the expression dictionary include more complex linguistic rules or does it consist only of simple collocations?

User-Specific Dictionaries: Are users able to create their own supplementary dictionaries, or can they update the general dictionary? How user-friendly is this coding process, and how much linguistic information can be added per word? (Note that there is some tradeoff between the complexity of the information that can be added and the user-friendliness of the process.) If you offer user-specific dictionaries, are they for single terms only, or can expressions be added? Does the user have a mechanism for adding multiple meanings for different subject areas and selecting the desired meanings for a particular document? If expressions can be added, can these expressions contain

simple linguistic rules? Will your company do dictionary updates for the customer? If so, what is the cost per entry? How many additional entries might be needed for this particular subject field and text type?

Integration into User Environment: Which word processing packages and publishing packages is your software compatible with? Will your company write a customized interface for our document format, if necessary? How long will this take and what is the cost? Is your MT software currently integrated into any other applications, such as an e-mail system, a document retrieval system, or a full-scale document management system?

Training: How much training is included in the leasing or purchase cost? Is additional training available, if necessary, and at what additional cost? Where is the training done?

Customer Support: How do you handle user requests for specific system improvements? Is there an organized method for accepting customer feedback and incorporating it into the system? What type of telephone assistance do you provide, and what hours is it available?

New Releases: How often do you provide new releases to your users? Is there any additional cost for new releases? Do these new releases include improvements made for all users?

Pre-Editing: Would you recommend pre-editing for this type of text and the purpose for which it will be used? If so, what form of pre-editing would you recommend? Does your company ever develop specific tools to aid users with pre- or post-editing?

References: Who are your principal users? Are any translation bureaus using your system as a translation aid? Would any of these users be willing to serve as references?

5 Customizing the MT System to Meet Your Needs

- 1. Decide on the type of documents to be translated.** If your organization has several different types of documents, start with those that are well-written, with repetitive terminology in a clearly defined subject field. Choose documents with a simple, more formal style, for example, product repair manuals or scientific reports. When you have achieved success with your first application, you can expand your dictionary to include more subject fields and text types. Keep in mind that certain text types, like advertising brochures and training manuals written in a chatty conversational style, will never be good candidates for machine translation.
- 2. Decide on the number and type of personnel needed.** You'll need to think about the volume of text that you will be translating, keeping in mind that the demand for translation in your organization is likely to expand as timely translations become available. You will also need to make a decision about the quality of output required. You may need to hire persons to be trained in the art of post-editing MT unless you have translators who are willing to learn this new skill.
- 3. Involve in-house translators at the earliest planning stages.** If you have in-house translators, these are the people who best understand the steps involved in the translation process, and their input will be invaluable. Since they are also likely to be the persons

who will be working with the MT system, their enthusiasm, commitment and creativity are the determining factors in making your system work for you. If you will be hiring a translator or bilingual subject field specialist to create dictionaries and work with the system output, then be sure to bring them on board as soon as possible so that they can be part of the project from its planning stages.

- 4. Determine how your documents will be input.** If your documents are already in machine-readable form, make sure that the MT software is compatible with your word processing or publishing software. If not, find out if the developer will write a customized interface for you. If you have lots of graphics and complex formatting in your documents, MT will be far more cost-effective for you if you can maintain this formatting throughout the document translation process. You may also want to make sure that the MT system will interface with SGML, the industry standard for CALS-compliant publishing software.

If you will be accessing documents from a machine-readable database, consider the peculiarities of certain database formats, such as the lack of two spaces after a period to differentiate the end of a sentence from the end of an abbreviation, the lack of diacritical markings in accented languages, and the lack of differentiation between upper and lower case letters. Any one of these things can cause big problems in machine translation, so be sure to find out how the MT system would handle such problems if you have them.

If your documents are not already in machine-readable form and you will not be creating them in machine-readable form, then consider purchasing a good scanner. Some high-end scanners on the market today can increase your document input rate by a factor of 3 if your documents are clearly printed on good-quality paper with no columns, tables or graphics. Even if your documents do not meet this standard, you can still double your input speed for longer documents if you purchase a high-quality scanner with good learning capability. You'll still need someone who is responsible for proofreading the input text, however, and it helps if this person has some familiarity with the language being input. Today's scanners work well for most of the European languages, but scanners for languages like Japanese and Russian await advances in optical character recognition to be cost-effective.

Finally, keying in documents manually is an option if machine translation, even with manual input, will be more timely than human translation. The U.S. Air Force has been using this method of inputting Russian for over 20 years and has trained numerous data entry personnel to perform the task skillfully. In this case, their Russian-English MT system, with its enormous dictionaries, is complete enough to serve as a spell-checker to catch many input errors. As machine translation gains a wider following, the market demand for better OCR devices will drive the development of the equipment needed to meet user needs.

- 5. Prepare your terminology list.** This is one of the most important steps in customization of an MT system to meet your special needs. Even if an MT dictionary contains a lot of terminology in the subject field of interest, you cannot assume that the terminology used by another member of your industry will be the terminology that your company uses. Experience has shown that every company has its own terminology preferences, but often these are not found in any company glossaries or terminology lists but rather in the heads of company translators. If you have approved company lists, by all means

plan to input this terminology into the system before you do any translating. If not, there are several options:

You or your MT vendor should run a translation of a representative sample of your text and get a list of not-found words.

or Ask your vendor if they can run an index concordance of a sample of your text that will display key words and phrases in context.

or Use a translator support package like INK TextTools to build a company glossary.

In all of the above cases, have your company translators supply the necessary terminology. If your company does not have any translators or terminology lists, then you will have to ask your MT vendor to supply some non-company-specific terminology for that subject field. You or the MT vendor should then input this terminology into your user-specific dictionary and re-run the translation of your sample text. At this point, it's important to give this output to a translator whom you have selected as the final authority and have them mark up the document and return this feedback for input into your dictionaries. If you have only supplied the meaning for a list of not-found words, you will find that there are a number of terms in the system dictionaries which will not have the meaning that you want as well as a large number of subject-specific technical expressions whose meaning cannot be formed by the sum of the meaning of the individual words, and these will have to be entered in your user-specific dictionary over time for best results.

- 6. Plan to create your own dictionaries.** Creating your own dictionaries is the second step in customizing an MT system to meet your needs. For this reason, you will need to create your own user-specific dictionary, or add to the system dictionaries that you have received, or have the MT vendor create a dictionary for you. If at all possible, you should assign one of your linguists the responsibility for dictionary update and maintenance. This person would be responsible for all terminology that is entered into the system. If a number of translators are allowed to modify the user-specific dictionary at will, then you tend to lose one of the chief advantages of MT, which is consistency of terminology. Also plan to update your dictionary on an ongoing basis, because this is the best way to cut down on post-editing and ensure that your system will become increasingly cost-effective over time.
- 7. Provide sufficient training for your personnel.** Make sure that everyone who will be involved with the MT system - systems programmers, data entry personnel, dictionary specialists, and post-editors - has sufficient training from the beginning of your project. If necessary, contract with the MT vendor to obtain any additional training needed. If your operation is large enough, you might want to ask if the vendor can give your people a special course in dictionary coding or post-editing. They will be able to give you useful tips for working with MT that can be gained only through experience. For example, your dictionary specialist should be trained to recognize when changing the basic meaning of a general term could cause problems; there is an art to recognizing when changes should be made to the dictionary and when changes should be left to post-editing.

8. **Include feedback to the MT developer in your planning.** Find out in what form you can supply your feedback and how it is incorporated into the system. You may be given improvement request forms to fill out and, if so, it's worthwhile to take the extra time to do this for the problems that occur most frequently in post-editing, as it will focus the developer's attention on areas where changes can bring the biggest increase in customer satisfaction. At least send back some post-edited copies of documents from time to time, and keep in touch with the developer to point out what you like best about the system and where you feel it needs the most improvement. This is not to imply that all requests for improvement can be easily incorporated into the system, but at least the developer will be able to explain the reason for a problem and tell you whether or not it can be easily fixed. Often the problem will be something that your personnel can correct by modifying the dictionary or other parameters of the system, and then your feedback becomes a valuable part of the learning process for your personnel.
9. **Decide whether or not you will need pre-editing.** Pre-editing is a term that is used to cover a variety of ways of manipulating an input text to produce input that is more suitable for machine translation. In general, it can be said that pre-editing becomes more valuable when you have an input text that must be translated into multiple target languages and/or you have a text that is subject to revision and re-translation. If you do decide that you need pre-editing, there are a variety of ways to go about it, including the following:
 - Use a form of controlled or restricted input. Xerox, which by early 1992 will be translating from English into 10 target languages, developed Multinational Customized English (MCE) as a sublanguage which restricts the number of words, homographs, and grammatical structures used in the input text. It is important to note that this style of writing is suggested to but not imposed upon their technical writers and that they use a computer program to automatically flag the areas of an input text that violate the language rules. It is also important to note that in recent years they have begun to use MT to translate additional types of documents that are not written in MCE.
 - Use an automatic text preprocessor, such as the Smart Expert Editor, an interactive text editor which converts ordinary English text into text more suitable for machine translation with the help of a human editor. This option is becoming increasingly popular as a front end to machine translations systems.
 - Contract an MT developer to write a customized preprocessor, which will convert your input text into input more appropriate for the MT system that you are using. As with any pre-editor, there is always an advantage in tailoring the pre-editing rules to a particular MT system.

Even the addition of more careful human editing at the document preparation stage could prove very useful, especially if done by a person who has some familiarity with the MT system in use. In all cases, a pre-editing program or controlled language will be more successful if it is written with some knowledge of the MT system for which the input is being prepared.

10. **Determine the amount of post-editing needed.** Decide whether your documents will require complete post-editing or a form of rapid post-editing. This should be clear from

the purpose of the document. Post-editing rates will vary with the type of text being translated, the quality of the input text, and the quality of the dictionary. Generally you can expect post-editing rates to vary from 2 to 6 pages per hour, which is about 2 to 6 times faster than human translation. This rate can be expected to improve over time as the dictionary continues to be updated, the system is improved from user feedback, and the post-editors gain familiarity with the MT system.

Be sure that your post-editors are willing to post-edit on screen; most of the advantages of MT are lost if translators insist on using the paper-and-pencil method. If your documents contain a lot of formatting codes and graphics, look into the possibility of completely automating the document production process. Using a method which allows you to post-edit an exact copy of the finished document on screen, complete with graphics and formatting, in a side-by-side display of the input and output text can speed up document production by another factor of 2 over the use of machine translation of text alone. Inquire whether or not the MT developer might offer training for post-editors or even consider writing a program to semi-automate the post-editing process. The U.S. Air Force has successfully used such a program for years. Because the program flags only the 20% of the text that needs special attention and because the Air Force uses a form of rapid post-editing that emphasizes information content over style, they have been able to achieve the remarkable post-editing rate of 15 to 20 pages per hour.

6 Making MT Work for You

Careful planning, customization, and commitment: If you keep these three C's in mind, you will be well on your way to having an MT application that meets your expectations. *Careful planning* will help you find the MT system that best meets your organization's needs and to integrate this system smoothly into your office environment. *Customization* over time, especially of the MT dictionaries, will mean increasing accuracy of translation and a steadily decreasing need for post-editing of texts of the same style and subject field. Finally, the importance of *Commitment*, especially on the part of the personnel who will be working with the system on a daily basis, cannot be overemphasized, because it is the enthusiasm, dedication and creativity of the MT system's human users that is the key to making MT work for you.