The IMO terminology databank

Nicole Walls

International Maritime Organization, London, UK

I came here at the kind invitation of Aslib to talk about the terminology database which IMO started some two years ago. It may be helpful if I begin by saying a few words about this organisation and its objectives. But before I start, I would like to specify that throughout the text I will refer to translators in the masculine gender, purely for reasons of convenience.

IMO, the International Maritime Organization, is the specialised agency of the United Nations dealing only with maritime affairs and its main concern is to improve safety of life at sea and to protect the marine environment. It has been in existence since 1958 and, after several moves, its headquarters are now on the Albert Embankment, opposite the Tate Gallery, on the south side of the Thames. IMO is the only United Nations specialised agency with its headquarters in the United Kingdom. It is a technical organisation and most of its work is carried out through a number of committees, attended by representatives of the Member States, the most important being the Maritime Safety Committee and the Marine Environment Protection Committee. These committees in turn work through a dozen or so specialised sub-committees, such as the Sub-Committee on Safety of Navigation, the Sub-Committee on Fire Protection, the Sub-Committee on the Carriage of Dangerous Goods, and so on.

The Secretariat of IMO, which employs some 290 permanent members of staff and a number of temporary employees whenever necessary, is divided into six main divisions corresponding to the main committees of the Organization, except for the Conference Division which is in charge of servicing the meetings. This involves interpretation, translation, production of documents and publications, typing and printing. Out of the 290 permanent staff, 150 are in the

Conference Division; this figure gives an idea of the relative importance of the Organization's programme of meetings and its publishing activities.

The Translation Services consist of six sections: Arabic, Chinese, English, French, Russian and Spanish, comprising 33 translators in all. These sections vary in size from one translator in the English Section to 13 in the Spanish one. Although IMO has six official languages, that is to say languages for which interpretation services are provided at meetings, it only has three working languages: English, French and Spanish. All documents are issued in these three working languages. Most of them are originally drafted in English (either by the Secretariat or by a Member State or outside body) and are translated by the Translation Sections into French and Spanish. Similarly, documents that are received in French or Spanish are also translated into the other two languages. A certain number of texts such as conventions, resolutions and committee reports, are also translated into Arabic, Chinese and Russian.

IMO prints about 20 million pages per year, which represents 75,000 pages per working day, and translates approximately 40,000 pages. To cope with this workload, it was felt that the introduction of new technology in the Conference Division would facilitate the monitoring of documents along the production line, that is to say translation, word processing and printing. It could also help to improve the sales operation in the Publications Section, that is to say invoicing, accounting, stock control, etc., and allow for the setting up of a terminology database in the Translation Services.

Why was a terminology database felt necessary? Part of the reason was that there are very few maritime dictionaries in existence and the available ones are not really satisfactory or are too incomplete. The *IMO glossary of technical terms* produced in 1982 was nearly out of print by 1986 and in need of updating and, in any case, it was only an English-French or English-Spanish glossary, supplemented by a cumbersome card index system. It was only in 1985 that Spanish became a full working language of the Organization and it was thought that a computerised terminology glossary giving immediate answers in all the language combinations was the best available solution to this new problem. It was also in 1986 that the growing translation sections which had previously worked independently of each other, were placed under the leadership of a newly-appointed chief. Cooperation between sections was to be reinforced and a common venture in the terminology field was to serve that purpose.

Self-revision, which was also introduced in IMO as in most other UN agencies, was another factor. Under this system, the work of senior translators was no longer revised and they took full responsibility for their own final translations. This system was expected to save time and improve output without any significant loss of quality. However, with self-revision, there was an inherent risk of lack of harmonisation in the translated texts and in the terminology used. It was thought that a database accessible to all would provide a means of alleviating that problem.

For all these main reasons and since funds for the acquisition of a computer were available in the Printing Fund, which is financed by the sales of IMO publications, the Director of the Conference Division decided to go ahead with the purchase of a computer system.

The next question then was: which computer?

Other UN agencies already had their own computerised terminology databases and the first thought was to draw on their experience. So, IMO organised a fact-finding mission to the International Monetary Fund (IMF) and the World Bank in Washington and to the United Nations in New York. It appeared that the IMF used a program called MINISIS which gave entire satisfaction. Moreover, it could be obtained free of charge from the IDRC (International Development Research Centre), a public corporation created in 1970 and funded by the Canadian Government, which makes the MINISIS program available without charge to all developing countries and to non-profit-making organisations. MINISIS was also used by the International Labour Organisation (ILO) and possibly other agencies based in Geneva. IMO therefore decided to opt for MINISIS since it had given excellent results and performed well for terminology purposes. Before giving you more information on MINISIS, I would like to add that the choice of MINISIS also solved the problem of the selection of the computer in so far as MINISIS can only run on the Hewlett-Packard 3000 series of computers.

MINISIS is an information management system written in Hewlett-Packard System Programming Language (SPL). It is a member of the ISIS family of information systems. ISIS stands for Integrated Set of Information Systems. The ISIS systems were developed to permit the handling of bibliographic databases using large mainframe computers. MINISIS fills the same need, but operates on smaller, less expensive minicomputers and its design is general enough to allow it to be used in many other applications, namely terminology. It has a multilingual and multi-character set capability and can handle Roman, Cyrillic, Arabic and Chinese characters.

As for the hardware, a Hewlett-Packard computer (series 3000, model 52) with a total disk space capacity of just over 1,000 megabytes, was purchased in 1987. As mentioned earlier, it is shared by three Sections: the Documents Section, the Publications Section and the Translation Services. A dozen or so terminals were initially bought along with line printers and a few individual printers. The number of terminals has now increased to 38, of which 28 are in the Translation sections. Each translator in the English, French and Spanish sections has one. There is also one terminal in each of the other sections (Arabic, Chinese and Russian).

This just about covers the software and hardware. After getting this equipment, we had to set up the database. It was an entirely new field and the necessary know-how was not available at IMO. So, there again, we had to draw on others' experience, namely IMF and IDRC. A three-week training course

was organised, at IMO's premises, by an IDRC representative. During this course, she adapted for IMO's purposes the data definition which had been kindly provided by IMF. The data definition is the structure of the database. It regulates the way in which data is entered, stored, retrieved and, lastly, presented on the screen. From the start, we decided that retrieval of data should be made as simple as possible so that by typing a term in English, French or Spanish, the translator would immediately be able to retrieve the term in all three languages. For example, the database includes the term 'fail-safe automatic closing fire-damper'. If the user of the database types in the term the record including it can immediately be displayed on the screen. In fact it is not usually necessary to type the whole term. If the user types 'fire damper' he will find then that those words occur together in two records. ('Fire', on its own, occurs in 93 records). If he wants more information on 'fail-safe' he can type those words and immediately consult the three records in which the term occurs, one of them including a long note defining what is meant by 'fail-safe' and how it is used. This is made possible by an inverted file, which provides a high-speed index to the information contained in the database. The inverted file is created by extracting words or terms from the data in each record, and then organising it so it can be read quickly. Like the index to a book, the inverted file tells MINISIS in which records to find the specified information. Entries in inverted files are created or changed when a record is added to the database or when its contents are changed. The IMF's data definition was considerably simplified for IMO's purposes, the chief differences being that, at IMO, we have fewer classification headings and no reliability code. Also, it was felt that the database would grow more quickly if there were fewer prompts and categories to fill in and that it would become operational sooner.

It was essential to get the cooperation of every translator and to show them the benefit of the whole exercise as soon as possible. This project did not have the full support of all translators at first. As IMO did not have a terminologist, it was necessary for the translators to do the additional work of preparing terms for inclusion in the database.

Once the data definition was ready, input of data could begin. We started with the *IMO glossary of technical terms* which existed both as English-French and English-Spanish. These two versions were combined after thorough revision by the Heads of the English, French and Spanish Sections.

Another substantial way of obtaining data was by the introduction of a terminology form which all translators have to fill in whenever they experience some difficulty in translating a word or expression. This form requests certain items of information, in particular the initials of the translator submitting it, the date of submission, the classification of the term. The classification field in the database is repeatable and more than one code may be entered if appropriate, in any order. A list of 40 abbreviations for classification headings was drawn up. The subjects covered include fishing, hydrography, navigation, environment,

ship stability, tonnage measurement, etc. With experience, it was found useful to add another category 'SP' (which stands for 'special'). Terms classified under this heading should not appear in any glossary, but may be useful, i.e. theme for World Maritime Day, tentative translations, titles of guidelines or resolutions, awkward expressions used by delegates during meetings or found between quotation marks in reports.

Then comes the English term followed if need be by its abbreviation. It is repeatable; this means that synonyms may be entered. As a general rule, the preferred term comes first. The same applies to the French and Spanish terms. Then comes the 'reference': this is the document symbol, title of document or publication in which the term was found. The 'source' indicates the name of the organisation originating the document or publication cited under 'reference'. And lastly, 'notes': this is the place for any comments by the translator which he thinks should appear in the database. Definitions and explanations are also placed there if necessary as well as an indication of the degree of reliability in a particular language, if the translation is not considered to be 100 per cent, reliable. This is because IMO's data definition, unlike that of IMF, has no reliability code.

When a translator has finished a translation, he hands it in with one or several terminology forms. These forms have to be approved by the Head of section. They are then sent to me in the office of the Chief, Translation Services, where they are recorded before being forwarded to the other translation sections for completion and approval. When they have been approved by the Heads of the English, French and Spanish Sections, they are given to the Chief, Translation Services, for final approval and then the data is entered into the database.

Data is also collected by screening existing or new publications, resolutions and recommendations from IMO and other organisations and by scanning specialised journals, drawing from them material relevant to IMO's activities.

To date, the translators have produced about 3,000 of these forms. It is very gratifying to note that even though some of them were somewhat reluctant at the beginning, all of them are now convinced of the usefulness of the database and readily participate in this terminological work. The database contains approximately 18,000 records and 26,500 terms at present.

The hit rate, that is to say the rate of successful queries by users is now very satisfactory and is of the order of 80 per cent. for temporary translators and 65 per cent. for permanent translators. The difference is explained by the fact that temporary translators tend to query terms which permanent translators know by heart and who therefore interrogate the computer for more obscure terms.

As a general rule, we do not produce printouts. The only time this was done was last year on the occasion of a diplomatic conference which took place in Rome. We extracted all the legal terms from the database, sorted them into alphabetical order with different source and target languages, and thus printed specialised glossaries which the translators took with them to Rome. They were also handed over to interpreters.

And now, I would like to discuss a few problems which we have encountered over the last couple of years. Most of these have been satisfactorily resolved. The first one was that of staff training and staff availability. Nobody in the Translation Services had any real knowledge of computers and several courses had to be organised to remedy this, first of all, courses at Hewlett-Packard's offices and then a course on MINISIS given at IMO by an IDRC representative who came over from Canada. As I mentioned before, there were no staff assigned purely to terminology and to the running of the database. The Secretary-General agreed to create a post to fill that gap in 1987.

The majority of day-to-day problems are connected with entering terms into the database and modifying them and also with other factors which are not of direct concern to the users of the database (the members of the translation sections) whose main interest is in reading the terms contained in the database in order to solve their own problems of translation. Some decisions that are made at the inputting stage do, however, affect the way in which the database responds to questioning and I shall concentrate mainly on these.

As I said earlier, when a translator types in a term and presses carriage return the computer searches not on the main database, which would be a rather slow process, but on a kind of index called an inverted file. The selection of terms to be included in the inverted file or excluded from it affects the way in which data is retrieved. The more words are excluded, the faster the search will be. The general principle is to exclude words, referred to as 'stopwords' or 'noisewords', such as the definite and indefinite article, and all prepositions and so forth, which appear frequently in terms but have no substantive content. These terms are therefore not, as a rule, included in the inverted file.

However, we have encountered some problems because of the absence of these words which meant that expressions like 'stand by', 'below deck', 'full ahead', 'full away', and so forth, could not be retrieved easily. For example, we currently have about 120 terms which include the word 'deck'. If the word 'below', in accordance with the general principle, were eliminated as a stopword that would mean that in order to find a suggested translation for the term 'below deck' the translator would have to read 120 terms. Similarly, a word or letter in brackets is interpreted by the computer as a separate word. Thus, when we enter a term such as 'activités relatives au(x) programme(s)', which is one of the suggested French translations for 'programme activities', making allowances for 'programme' in French to be either singular or plural, in other words adding an x in brackets for the plural of the article and an s in brackets for the plural of 'programme', we might have decided that the 'x' was irrelevant to the substantive meaning of the term and have made it a stopword, like the article 'au'. However, this would have meant that terms such as x-ray would not have appeared and we therefore decided to live with the minor inconvenience of having 'x' as a term with the result that a translator looking for x-ray and putting in just 'x' would have to look through 12 terms some of which contain an x of the 'plural option' type which is of no interest. The letter 's' has 140 postings which

include the terms where the 's' appears in brackets as an optional plural or after an apostrophe such as 'ship's papers'. However, the letter 's' is not a stopword because that would make it impossible to retrieve the term 'raccord en S' which is one of the French translations offered for the term 'gooseneck' which, as you may know, is a kind of pipe having a curved shape resembling the neck of a goose. If 's' were a stopword, the translator would not be able to find out whether the term 'raccord en S' was in the database and would have to look through all the postings of the word 'raccord'. As it is he merely types in 'raccord S' (omitting 'en' which is a stopword) and will be able to display the record on the screen immediately.

This brings me to another problem, which is that for the computer a word in the singular and the same word in the plural are completely different words. For this reason we always try to enter all data in the singular but sometimes this is not possible since some words or some expressions are inherently plural. If translators forget this they may fail to retrieve information which is available in the database. For example, if a translator has to translate a term including the word 'alarms' and asks for that word, he will find that it appears only twice in the database. If he asks for the singular he will find that there are 49 terms which include the word 'alarm' and this considerably increases the chance of finding a satisfactory equivalent in French or Spanish for the particular type of alarm mentioned in the document being translated. In the same way the database regards words joined by hyphen as being separate words.

One convenient short cut which provides a way of finding a term which can be written in several ways is to use the command 'at' which is the symbol on the keyboard consisting of a small 'a' surrounded by a circle which is used to mean 'at' in certain accounting and other contexts. This facility is known as 'right truncation' or 'left truncation'. For right truncation, which is the most commonly used, a word or part of a word is typed in and the 'at' symbol is then typed. When carriage return is pressed the computer will give a listing of all words which begin with the letters typed. For example, if the translator has to translate 'foremast' and it appears, unusually, in the text in two words 'fore mast' and he tries to retrieve the word in that form from the computer he will find that it is not there. Alternatively he might realise that it could be written as one word (as it usually is, in fact) in which case he would find that it has one posting. If he is in doubt he can type in 'fore' and 'at' and press carriage return and receive a listing of all the words in the database which begin with those letters, including, of course 'foremast'.

This facility is also useful when the translator is looking for a translation of a term that is not directly available in the database but which is cognate with terms that are to be found. For example, if a translator wanted to translate the French term 'canaliser' and typed that word, in the infinitive, into the database he would receive the message 'illegal term' which means that the term is not found. If, however, he types 'canali at' and presses carriage return he will receive a list of

words which begin with those letters namely 'canalisation' (in the singular and in the plural), 'canalisé' (past participle) and two Spanish terms 'canalización' and 'canalizado'; if the translator then reads those terms (13 in all) by typing in 'browse' it is very likely, by consulting terms cognate with the term wanted, that he will find useful guidance on how to translate the term by analogy with the terms that are actually in the database.

Left truncation works in exactly the same way as right truncation but is less commonly used. If the translator types in 'at' followed by a number of letters and presses carriage return he will receive a list of words ending with the letters typed. This may be useful in searching for compound terms.

I mentioned earlier that all terms are assigned to a classification which differentiates between terms used in different contexts. It is possible to narrow, and also speed up, the search by specifically requesting the display of a term in a particular context. For example, a translator working on a text relating to radiocommunications may come across the term 'valve' and wish to find some suggestions as to how to translate it. If he types in the word 'valve' and presses carriage return he will find that he has an apparent embarrassment of riches because there are 104 terms which include this word. When he begins to browse through the terms, however, he will soon see that many of the translations of 'valve' are quite irrelevant to his present requirements since they relate to a mechanical device rather than to a radio valve. French terms such as 'soupape', 'clapet', 'purgeur' are therefore of no use to him. Instead of reading through all 104 terms he can repeat the request for the word 'valve' specifying that the classification is electricity. He will then be able to display the three terms which are in the database and are relevant to that context. Maybe this is not a very good example as the only radio valves to be found are probably in museums and not aboard ships!

When scanning inverted files, MINISIS does not differentiate between different languages. For example if one looks for a translation of the English word 'file', the records retrieved would also contain the word 'file' in French, most of which are unrelated in meaning. It is, however, possible to use a different search method which restricts the search to a particular word in a specified language. This is slower since the search is made on the database itself rather than on the inverted file but in spite of the relative slowness of this method of searching it may be worthwhile using it as an alternative to reading a long sequence of records, many of which are inevitably irrelevant to the query in hand.

These are some of the problems which we have found and solved, in most cases, in a satisfactory manner. Finally, a few words about future developments. The most obvious development would be the extension of the database to the other three official languages. This is technically possible but due to a recent financial crisis and to staff shortage and turnover, especially in the Russian and Chinese sections, we are not able to envisage this extension in the near future. As

an intermediate measure, the Arabic, Chinese and Russian Sections have been given a terminal so that they can benefit from the explanations and definitions in the notes. Also, some of these translators speak either French or Spanish or both, so they can benefit from the translations of terms. The question of machine-assisted translations has been raised but no development in that field is foreseen for the time being.

AUTHOR

Nicole Walls, International Maritime Organization, 4 Albert Embankment, London SE1 7SR, UK.