[*Terminologie et Traduction*, no.2, 1985]

MACHINE TRANSLATION DICTIONARIES

Ian M. Pigott,
Commission of the European Communities

Introduction

Machine translation dictionaries, unlike dictionaries, glossaries or terminology banks designed essentially for use by human beings, need to contain extremely full and precise lexical descriptions of all the semantico-syntactic features required for source language analysis, bilingual transfer and target language generation in the translation process. In addition, as the vast majority of current translation requirements concern written texts, MT dictionaries have been compiled first and foremost on the basis of the level of language used in fairly formal written communication which differs substantially from the wider sphere of conversation, poetry, literature, slang and rhetoric covered by many paper dictionaries.

Furthermore, as the computer cannot a priori distinguish between general and technical language, MT dictionaries must provide extremely detailed dictionary entries on basic vocabulary and terminology (e.g. the 5,000 to 10,000 most common words or terms in a source language) as without this base, no matter how well more technical terms are represented, the result of any machine translation will leave much to be desired.

Direct experience of operational machine translation systems at the Commission has been principally concerned with Systran (under development since 1976) and Logos (tested for a nine-month period). However, we have also had the opportunity to monitor progress over the years on other systems such as Spanam/Engspan, Weidner, Smart and Alps. All these systems use very similar approaches to dictionary compilation and coding. Those differences which do occur are normally a result of the maturity of the system and the corresponding development effort. For the purposes of this paper, special reference will therefore be made to Systran, any important differences found in other systems being clearly referenced.

Compilation

The most widely used approach to dictionary compilation is via the establishment of a representative corpus of texts covering the preferred subject field or text type chosen as a basis for development. Such a corpus varies typically between 100,000 and 1 million words and is frequently based on the selection of a large number of text excerpts of up to 500 words each.

By running word frequency counts and key-word-in-context (KWIC) listings, it is possible to identify the basic vocabulary required and undertake objective dictionary coding on the basis of the actual occurrence of words, terms or idioms in context. At this stage, words with very low frequencies are usually bypassed.

Once a basic dictionary (essentially one-word entries and frequently recurring two- or three-word string expressions) has been coded up, test translations can be run and additional dictionary enhancements can be made, more often than not on the basis of translation errors occurring in the tests.

It is interesting to note in this context that developers who have made use of corpora covering too wide a range of language (e.g. the Mannheim corpus) have often found it necessary to undertake quite radical modifications to the initial basic dictionary as the semantics of terms in the real world of translation may well not coincide with usage in literature or, for example, in the popular press. Furthermore, everyday words such as uncle, hello, and stupid may never occur in texts submitted for machine translation.

Subsequent enhancements to the dictionary can be made on the basis of experience, more often than not as a result of feedback from translators and end users. This is an on-going process which may continue for many years.

Finally it should be noted that some suppliers (e.g. Smart and Weidner) frequently market systems with a relatively small dictionary (up to 6000 words) and expect further dictionary enhancement to be handled by the user. In our experience, this can lead to substandard coding as the user is often unable to appreciate the way in which the MT software will interpret his preferences. On the other hand, the larger systems (Systran, Spanam, Logos) usually attempt to maintain centralized control on dictionary enhancement despite pressure from users. In this way, they can provide bigger and better dictionaries, typically of over 100,000 entries, to an ever wider group of users as time goes by.

Source-language one-word dictionary files

The source-language dictionaries are at the very basis of any MT system. Indeed, an appreciation of the actual or potential performance of a system can usually be made in relation to the types and complexity of the coding features and markers available at this level. The less sophisticated systems are often limited to morphology and what may be referred to as "basic grammar" while more highly developed systems will also make provision for quite complex series of syntactic and semantic markers.

Let us attempt to categorize the types of information which may be found in the source-language one-word dictionary:

1. Morphology. Traditionally (and even now as far as internal format is concerned) the morphology of inflections has been handled by series of codes which point to tables of regular and irregular roots and endings, principally for nouns, verbs and adjectives. Depending on the source language, inflected forms may be created physically as full forms by the dictionary updating programs (for the less highly inflected languages such as English) or (for the more highly inflected languages such as French) be resolved on the basis of an analysis of the possible

endings. *The* more advanced systems provide for automatic recognition of regular and irregular forms based on comprehensive tables of all inflections for a given language which are accessed from the base form (e.g. infinitive for verbs, singular for nouns, masculine singular for adjectives, etc.).

2. Part of speech information. Each entry in the source language dictionary will contain extremely precise information on part of speech which will either be coded manually or created by the type of automated updating programs described under 1 above. Many systems provide for two levels of part of speech coding, the first being the basic part of speech (noun, verb, etc.), the second being the subcategory of the actual entry (proper noun, modal verb, comparative adjective, etc.). Many special categories of "part of speech" are found in MT dictionaries which rarely occur in paper dictionaries, e.g. categorizations of numerals (Roman numeral, fraction, decimal, year), of pronouns (subject, object or both) and even of punctuation marks (left or right curved parentheses or square brackets).

3. Basic grammar. Depending on the individual characteristics of the source language, information on gender, number, case, person and tense appears as required with nouns, verbs, adjectives, pronouns, articles, conjunctions, prepositions, etc.

4. Capitalization. It frequently happens that information on capitalization can be of enormous assistance in resolving MT problems. For example, proper nouns or abbreviations (Nice, May, As) can often be distinguished from other words with the same spelling (nice, may, as) with reference to this information. Moreover, the fact that German grammar requires the capitalization of nouns provides a means of distinguishing between part-of-speech homographs (Lesen vs lesen).

5. Homograph type. In the field of machine translation, a homograph is defined as a word with a given spelling which can act as two or more different parts of speech, e.g. light as noun, verb or adjective. Homograph resolution is usually the most complicated part of MT analysis and it is therefore vital that homograph types be correctly coded. In English, there are some 80 types varying between verb / noun and comparative conjunction / adverb / subordinate conjunction / preposition / relative pronoun. In some cases, even if paper dictionaries list various possible parts of speech, the MT dictionary will be more restrictive as certain of these rarely occur in written texts. Thus buy would probably not be recorded as a noun - verb homograph but only as a verb.

6. Syntactic codes. The precise functions of syntactic codes varies considerably from system to system. The smaller systems may make do with as few as six basic codes such as transitive, intransitive, governs infinitive or impersonal adjective, while more sophisticated systems may have as many as 60, including such intricate characteristics as past participle adjective preceded by "more", to be considered comparative or noun followed by "of" plus present participle requires gerund rather than verbal adjective. Paper dictionaries are usually lacking in such information which is, however, extremely useful for natural language analysis.

7. Simple semantic codes. Most MT systems make use of semantic codes such as concrete, abstract or animate which, in conjunction with similar semantico-syntactic codes such as animate subject, are widely used in establishing elementary sentence analysis.

8. Semantic primitives. This area of coding varies widely from system to system. Some systems make little or no provision for semantic codes while others have sophisticated lists of codes, essentially for nouns (as in Systran) or for verbs (as in Logos). Such codes may include features such as DEVICE (for all types of equipment), PROCESS (e.g. for verbal nouns) or SCIENCE (for areas of knowledge). Typically, up to 40 such codes can be used systematically in an MT dictionary, and as many as 500 may be used for resolving special cases. In Systran, they are arranged in five major taxonomies. In Logos and to some extent in Weidner and Alps, they are accessed via thesaurus mechanisms. Similar codes often occur in paper dictionaries but they are rarely adapted to MT requirements and can often be misleading.

9. Potential preposition government. Mainly as a result of the inordinate amount of work carried out in this area by pure linguists, most systems make provision for the inclusion of the prepositions which are likely to be governed by nouns, verbs or adjectives. As better methods of parsing prepositional structures are developed, the usefulness of this type of coding is reduced. However, it may still be useful to include reliable information indicating, for example, that if the verb fight occurs in a clause containing the preposition against, it is probable that there is a direct relationship between these two words.

Source language multi-word and contextual dictionaries

Multi-word MT dictionaries exist in all systems, at least at the elementary levels. The more developed systems also provide for the coding of so-called contextual entries which supply meanings in the target language to be entered on the basis of the syntactic or semantic relationships already established by the system parser. At least one system (Systran) also provides contextual dictionary facilities for re-establishing the parsing of non-typical contexts in the source text.

Contextual dictionaries have two main functions: to enhance the analysis of the source text and/or to provide correct target meanings for given contexts.

The various levels of contextual coding will be considered in order of complexity.

1. String idioms are frequently used to unite two or more words in the source text so as to enable them to be treated subsequently as the equivalent of a one-word entry with a given part of speech. For example "in order to" can be reduced to a one-word string (in*order*to) and then be coded as an infinitive particle, or "by dint of" can be regarded as a preposition. At this stage, no target meaning needs to be given as this will be provided by the standard one-word dictionary or by other contextual dictionaries.

2. Noun strings can be entered at the source language level as a means of establishing the fact that the head word (in English usually the last of the group) is to be analysed as a noun rather than another part of speech. Examples here would be electric light or air traffic control where light and control would be firmly established as nouns rather than verbs or adjectives.

3. Straight idioms are usually entered in bilingual files as a means of ensuring that whenever a given string of words occurs in the source text, it will be translated by a given target meaning. For example, in this file such strings as in a clockwise manner or according to company regulations might be listed. It is only fair to point out, however, that this type of rather unsophisticated entry is rapidly disappearing from the more mature systems as the inherent translation problems can be better handled in other ways.

4. Noun phrases requiring a special translation in the target language are perhaps the most frequent of all entries in contextual MT dictionaries. Frequently they are technical (fast breeder reactor, optical character reader) but they can equally well be general (railway station, letter head, old age). As an indication of the potential of this type of entry, some 200 expressions containing the word oil have been coded in this file in English-French Systran. Such coding also facilitates the analysis of the strings themselves by a process of crystallisation.

5. Contextual relationships based on the results of source language parsing constitute an extremely powerful dictionary tool in some systems but are not usually to be found in the smaller systems. Systran has extremely wide and powerful capabilities here while Logos provides a subset of about a dozen typical sequences.

   In some cases meanings may be ascribed to a number of words (e.g. when company is the subject of employ special translations of both words are required in French). More often, generalized semantic relationships are entered (employ with the semantic category PROFESSION as object) requires a special translation. Some entries might be extremely complex when a complicated structure requires careful treatment (in as a preposition governing that followed by a plural noun which is itself the subject of an intransitive verb requires a special translation).

   In this type of file, some 90% of the entries are unlikely to be found in paper dictionaries as they document processes of translation which rarely present problems to human beings.

6. Parsing contextual entries are used in Systran to deal with structures which are exceptions to more general rules of grammar. For example, the verb go followed by to followed in turn by a noun-verb homograph usually requires to to be analysed as an infinitive and the homograph to be analysed as a verb, as in "they went to help as quickly as possible"; however, when the noun-verb homograph is school, noun resolution is required. The correct result can be obtained by means of a parsing entry.

Target language coding

Whatever the system, target language coding is much simpler than source coding. The main purposes of target coding are to provide the correct meaning, to ensure that target morphology is fully catered for and to add codes indicating any special grammatical features or word-order requirements the translated term may have.

1. Meanings, depending on the system, are either added on the basis of subject field parameters (see below) or are chosen as the most generally acceptable equivalent for one-word entries and on the basis of specific contexts or strings for contextual entries.
2. Morphology can be dealt with in any of the following three ways: by physical reference to tables of regular or irregular endings, by comparison with other words in a given class, or (as in the present Systran) by fully automatic recognition of the word type on the basis of an exhaustive set of previously coded tables.
3. Grammatical requirements for the target language are ensured by the introduction of special codes indicating such features as when requiring infinitive government introduce "de" before the infinitive or replace this adjective before the noun. Some thirty codes of this type usually suffice.

Subject field parameters

One of the most difficult and controversial subjects for discussion in the management of target language MT dictionaries is the way in which terminology for different subject fields should be introduced.

All systems provide for subject field coding (TGs in Systran, SMCs in Logos, etc.). However, the definition of subject field is a problem in itself as no widely accepted categories exist. Some systems, for example, provide for very generalized spheres such as "general" or "technical", some provide for particular areas such as "medical", "electrical" or "nuclear", while some go as far as to provide for client-specific dictionary entries (e.g. Ford or General Motors in the automobile industry). Finally, some user or development groups have sought to create dictionaries which are as generally applicable as possible, reducing subject field coding to an absolute minimum.

The initial attractions of being able to use subject field parameters are obvious. The smaller systems, in particular, often produce a large number of incorrect translations when they are first installed at *a* user site dealing with specific areas of science or technology. It may well happen that the word "file" in a computer maintenance manual is translated as the tool (Fr. lime) whereas the intended meaning is that of an area of data storage (Fr. fichier). To correct the error, the user is often encouraged to define a subject field (say computers) and to enter the new meaning in a corresponding glossary. Once this is done, providing the correct parameter is set for the rerun of the translation, the right translation will be accessed.

Complications occur as new subject fields or subsectors are identified. For example, in aeronautical texts a "line" may require different translations depending on whether a text is dealing with hydraulic or electrical systems (pipe vs cable). This will require further definition of subject sectors, new parameters, new dictionary entries and so on. In addition, more often than not, documents cover more than one subject area. A text emanating from an aircraft corporation may well involve consideration of fields as widely different as economics, meteorology, physics, materials, seating arrangements, staffing, meals or statistics. Each will require special treatment and subject field parameters will offer no reliable solution.

However, some users are admittedly able to obtain fully usable results on the basis of subject field parameters, providing the majority of their translation work is concerned with one particular sector. For this reason, quite large specialized target dictionaries have been built up for fields of major interest such as computers, aerospace, commerce, etc. The target dictionary entry may then have one basic general translation and a list of more specialized translations by subject area. The translation of "power" could thus give "puissance" (general), "capacité" (computers), "accéleration" (aerospace) and "pouvoir" (commerce). Exceptions to the general meaning will however only be added to the dictionary if the general meaning does not apply to a given field. The lists by subject field are therefore far from exhaustive.

At the Commission, where it was obvious from the start that any large-scale use of machine translation would depend on high-quality output for a large number of quite different subject fields, the strategy has been to allocate widely-acceptable general default meanings and to obtain specialized translation through the compilation of contextual dictionary files. In this way, quite satisfactory results can be obtained for texts dealing with any area of activity without the necessity of subject field coding. This approach also avoids problems of definition of the source text subject field by the user.

In the Commission's system, the subject field parameter is only used for dealing with the few cases where a single word frequently occurs out of context (in titles, tables, etc.) and requires a fundamentally different meaning in a particular sector. For example, the default translation of "réacteur" is "reactor" which is perfectly adequate for general texts and for some areas of technology such as nuclear science, but would be quite incorrect and misleading in aerospace texts where the meaning is "engine".

To sum up, the target dictionary depends not only on development strategy but on the specific translation problems encountered in one or more subject fields. Thus, here too, MT dictionaries are very different from bilingual paper dictionaries, even those to be found for specialized subject areas.

Suggestions for lexicographers

From the above, it is obvious that until now, generally available dictionary resources have been of very limited use to those compiling MT dictionaries who have been forced to develop other tools to assist them.

One of the main difficulties appears to be that lexicographers have never paid any particular attention to the major sources of translation work such as technical reports, maintenance manuals, informative texts, documentary data bases, minutes of meetings, annual reviews, etc. The level of language in all these is quite different from that to be found in conversation, literature or debate.

A first suggestion would therefore be that lexicographers clearly identify the area of discourse addressed, either by listing all areas which apply (e.g. general, conversation, literature, informative texts) or by listing exceptions (e.g. not conversation, not literature). This would obviously have to be done on the basis of large corpora covering the various categories.

If this work were to be properly coordinated, it might also be possible for lexicographers to cater for the other detailed levels of dictionary coding necessary for machine translation, e.g. syntactic government, semantic primitives, homograph types based on frequency of occurrence and, in bilingual dictionaries, basic general default translations.

Finally, it would be gratifying if dictionary publishers were to take a more active and direct interest in the requirements of machine translation specialists now that MT dictionaries are expanding rapidly and are covering more and more language pairs and subject fields. Collaboration on lexicography in this area might well be of benefit to both parties.

Conclusion

Until now, lexicographers have not taken account of machine translation and related applications. As a result, most MT dictionaries have been created from scratch on the basis of the special requirements of the field.

MT dictionaries are constructed in formats and files which appear very different to those used in other areas of lexicography, although in many cases the types of data they require could be usefully documented in other types of monolingual or bilingual dictionary.

At the very least, more detailed study by lexicographers of the field of machine translation could lead to concrete proposals for providing the types of information required.

Unless such steps are taken, it is probable that the relationship between MT dictionaries and other lexicographic sources will remain extremely superficial and, in many cases, misleading.

This paper was originally prepared for the Workshop on Automating the Lexicon, Marina di Grosseto, Italy (May 1986).

1805/87