[From: Paul L.Garvin (ed.) *Natural Language and the Computer* (New York: McGraw Hill, 1963)]

KENNETH E. HARPER

Dictionary problems in machine translation

The machine-translation process is customarily divided into four stages: input, analysis of input-language units (normally, the sentence unit), synthesis of the output-language unit, and output. The problems connected with the first and last stages are primarily technological, whereas the two middle stages are the province of the linguist. The problems of analysis and synthesis must be solved in terms of morphology, syntax, lexicography, and semantics. Although technology can certainly contribute to a more accurate formulation of these linguistic problems and to more specific answers to them, it is safe to say that an adequate linguistic theory is still the prerequisite for successful translation.

The role of the MT dictionary in the central translation process has frequently been misunderstood. This misunderstanding has to do with the content and function of the dictionary. A view still prevalent is that the dictionary is a word list, however obtained, and that its exclusive function is the identification of items in the input text. It is only in recent years that a broader concept of the MT dictionary has evolved. The thesis of this discussion is that this broader concept must be more generally accepted if MT research is to advance beyond its presently primitive state.

The chief reasons for the original view of the MT dictionary are reliance on tradition and a misunderstanding of the role of technology. Traditionally, dictionaries have been word lists; as inventories of the words in a language, they contain more or less complete information about the sound, meaning, and grammatical properties of the items recorded. The science has its own name: lexicography. Grammar and semantics are separate fields of study. On this traditional foundation, the early MT researchers proceeded to build word lists by, and for, machines.

There was some disagreement on the manner of selecting entries for the dictionary, and the MT linguist frequently insisted that the dictionary contain more information of a grammatical nature than is customary in published word lists. Usually, however, this excess of grammar was

viewed as a necessary evil, admissible as an aid in the fulfilling of the primary function of the dictionary—text lookup with a stem dictionary. Beyond this, the MT linguist has customarily been content to go along with the traditional division of labor; if he studies syntactic structure, he thinks of resolving such problems by means of separate routines that operate on the discrete items in the dictionary. The lexicon and the grammar are as distinct as they have always been, and new knowledge obtained in one area is not seen as important for the other.

Students of MT have in general clung to the conventional division of grammar and lexicography, without recognizing that technology has provided a means of fusing the two. To put it differently, technology has at best been applied unimaginatively. The trend might be defended if existing dictionaries and grammars were adequate to the tasks posed by MT. They are, in fact, woefully inadequate, as an intelligent analysis of the failures of machine translation will reveal. It should be noted that when failures are encountered in test translations, the tendency of MT workers has been simply to improve the technology; when the translation difficulty is "resolved" by the expedient of placing the problem and its solution in storage, the "technical improvement" essentially amounts to an increase in the size of the store. The utter impracticality of such "solutions" can be readily demonstrated. The whole situation has evolved, somewhat naturally, out of the linguist's overestimation of the powers of the technologist and the technologist's overestimation of the achievements of lexicography and linguistics.

The following discussion of the chief problems in building and operating MT dictionaries is not a critical survey of past work; the literature reveals that much of this work is tentative and exploratory. The purpose of the discussion is to point out the areas in which progress has been made and those in which work has lagged. Throughout, the orientation is linguistic.

PHYSICAL CHARACTERISTICS OF MT DICTIONARIES

As text enters the MT system, the text items are compared with the items in storage. The two obvious requirements of the machine store are a large capacity and rapid access, although what is meant by "large" and "rapid" is still indefinite. It is clear, however, that the dictionary-lookup process should result in the identification of all but a very small fraction of the items in a given text, at a rate that is at least equal to the translation process itself.

These minimum requirements apparently rule out certain types of stores that are otherwise promising, such as punched cards, magnetic drums, and discs. The advantages of punched cards include unlimited capacity, variable length of record, and relative ease of modification; the decisive disadvantage, as with magnetic discs, is slow and cumbersome

handling. The chief drawback to the magnetic drum is limited capacity. Other types of stores, including magnetic tape and the photoscopic store, avoid these difficulties but introduce new problems. Magnetic tape, which provides an unlimited store, has been used as an instrument for text lookup in at least three ways: (1) an alphabetized text tape is compared with an alphabetized dictionary tape; (2) assuming core storage as the locus of input text, a text occurrence is transformed into a memory address, the same transformation rule is applied to dictionary items, and the assignment of a text occurrence to the same address in memory as a dictionary item constitutes a match; and (3) a character-by-character table lookup is performed for the input occurrence, assuming that a considerable portion of the dictionary is present in core storage. The chief drawbacks to the tape dictionary are serial access and the necessity of writing a new tape when modification of individual records is required. The latter difficulty is most apparent in the case of the photoscopic store, which otherwise answers quite well the initial requirements of size and speed. See "Programming for Natural Language" by L. C. Ray, earlier in this volume.

It is not yet clear that a special-purpose device will be required for an MT dictionary. A number of factors are involved; it would make a difference, for example, whether the store is used for the sole purpose of the MT dictionary, or whether it is to be shared with nonlinguistic users. At any rate, it is obvious that the thinking of most MT researchers has been dominated by the theoretical requirements of capacity and speed. It is also obvious now, if not always so, that this cart-before-the-horse approach is the result of a naive faith in the efficacy of the dictionary as a word list, and of pressures to get MT on a "production" basis. The further requirements of an MT dictionary discussed below have been either ignored or attended to in a piecemeal fashion.

CONTENTS OF THE DICTIONARY

A bilingual MT dictionary consists of a series of "records" encoded in the machine language of the storage unit. These records bear only a general resemblance to the entries in ordinary dictionaries, and their ordering may be quite different. The main elements in a record are representations of the input-language lexical item, its correspondent (s) in the output language, and attached codes relating to grammatic and semantic properties of one or both of the lexical items. For purposes of discussion, we shall treat these elements separately, despite the fact that this separation may not always be preserved in practice.

REPRESENTATION OF THE INPUT-LANGUAGE LEXICAL ITEM

In ordinary dictionaries, a "record" is headed by a canonical form of a given word, i.e., by a form arbitrarily chosen as the "name" of the word.

This convention is rarely observed in MT dictionaries, since the input-language item must be recognized in terms of graphemes (letters, spaces, and punctuation). Lexical items have generally been represented in one of two ways: by listing members of the paradigm for inflected words (the form dictionary), or by listing a segment of inflected words (the stem dictionary). A variation employed in the photoscopic store is the listing of certain word combinations as main entries. For example, *once in a while* may be listed as a single entry, because the combination may be considered as a lexical unit, or because it is idiomatic or difficult to translate. A more common practice is to list the four elements separately in the dictionary; the combination and its translation are stored in a separate table, to which reference is made by codes attached to the separate lexical items.

The question of form versus stem dictionaries is still unsettled. The form dictionary has certain advantages in MT research. Operation and maintenance are relatively simple; as text is processed, phenomena such as form frequency and frequency of equivalent (corresponding output-language word or words) can be conveniently studied. The chief disadvantage is, of course, the size of store required. In a highly inflected language such as Russian, the required storage is some fifteen times greater than for a stem dictionary with the same number of lexical items. In English, the ratio is approximately three to one. This discrepancy has influenced a number of MT researchers toward a stem dictionary.

There are really two types of stem dictionaries: the *stem-affix* type, in which *affix* signifies only inflectional endings; and the *root* type, which provides for the storage of roots, inflectional endings, derivational suffixes, and prefixes. The latter type obviously represents a saving in storage over the former and has as an advantage the potential capability of recognizing and translating lexical items in text that are not contained in the dictionary. It would not be practicable, for example, to list in storage all the English nouns and verbs to which the prefix *re-* can be added. A number of such words have never appeared in published dictionaries, and indeed have never appeared in print. Normally, if the root is stored, even neologisms of this type can be identified in text and translated.

The saving in storage in all types of stem dictionaries is, of course, accompanied by a far greater complexity of operation than is characteristic of the form dictionary. The degree of complexity is proportionate to the degree of segmentation. The main problems are occasioned by multiple stems, stem homographs,[1] and the use of morphological codes that will permit the correct juxtaposition of stem and affix and the correct reference to the output lexical item. The bookkeeping problem is enormous, and generalized solutions are hard to find. Although sub-

[1] Homographs are different words that are spelled alike, such as *lead.*

stantial efforts have been made, it has not yet been demonstrated that the inherent problems can be solved through the application of non-linguistic techniques. It can well be argued that large-scale stem dictionaries depend on a much greater knowledge of lexicology, grammar, and semantics than we now possess. To put it differently, a strategic, rather than a tactical, assault on the problem may be required; the efficacy of the tactical assault, in any case, has not been proved.

One other aspect of the input-language item should be mentioned: its source. Two general selection procedures have been used: (1) only those items appearing in processed text are stored; and (2) the items are taken from published dictionaries. A combination of the two methods has also been used. The text-based dictionary has the advantage of greater precision: its items are endowed with reality (they have really appeared in given texts), and their frequency, forms, and translation in these texts is known. It can be argued that only a text-based dictionary can provide a true idioglossary;[2] certainly it is an excellent instrument for automatic language-data processing. Its disadvantage is incompleteness, short of the processing of a very large body of text. Experience at one MT study center has shown that after the first few hundred pages of text were glossarized (in the general field of physics), one new word was encountered for every sixty running words of new text. This proportion is likely to decrease very slowly.

MT dictionaries compiled from existing dictionaries (and supplemented by text processing) are larger and provide more complete "coverage" for new text. They are also imprecise. Some of the entries will always be superfluous, since they will never be encountered in text ("never" is used here in a relative sense). Similarly, many items contained in the dictionary are of dubious value until also found in text: the information attached to the item will be incomplete and/or inaccurate for purposes of syntactic analysis and translation.

GRAMMAR CODE

A minimum of morphological information is encoded in all MT dictionaries, indicating the part-of-speech and inflectional characteristics of each lexical item. As noted above, the amount of detail in these codes depends on the degree to which the lexical items are segmented. Codes specifying the syntactic properties of the items may also be included, either in combination with the morphological codes, or separately. The existence or detail of syntactic codes depends on the degree to which syntactic analysis is applied to the input sentence. If syntactic analysis is not attempted, these codes will clearly be unnecessary; if the aim is a complete structural description of the input sentence, detailed information about the

[2] An idioglossary is a dictionary of terms used in a single field of interest.

combinatorial properties of each item is highly desirable. There has been a trend in recent years toward more detailed syntactic analysis.

The main problem of the grammar code is one of content: What grammatic information should be coded for each item in the dictionary? Most MT research groups have assumed that since morphology is well described and the basic facts of syntax (government and agreement) are well known, the only problem was to imbed these facts in the grammar code. The involuntary consequence of this reliance on existing grammars has been a built-in intractability—grammar codes are difficult to modify. In some instances, this is due to the compactness of the code (the use of a single symbol to represent two or more grammatic properties). The obvious virtues of compactness (saving in storage space) are more than offset by the considerable complexities that arise when a multiple-purpose symbol must be changed. In other instances, the problem is simply one of economics: The physical characteristics of the store make alteration of any item expensive. At the root of all such difficulties is the original assumption that the contents of the grammar code could be fixed in advance.

It is by now widely recognized that as we process more and more texts, we learn a great deal more about grammar. Although automatic language-data processing has so far taught us little that is really new, it has enabled us to organize and codify the thousands of facts, particularly about syntax, that "everyone knows." Nor do we need to rely on text processing to obtain this necessary detail. The point is simply that there is an enormous advantage in a greatly enriched "grammar code" for every lexical item in the dictionary. A Russian grammar code, for example, should contain information about the nonexistence of a short form for a given adjective, or about the absence of an adjective-adverb pairing for a given word. This information should be stored in the dictionary, in the same way that part-of-speech codes are stored. The clear implication is that MT dictionaries of the future must have ample storage space for this information and must easily accept alteration. The grammar code of any entry will be altered from time to time, as new information is gathered; it can be said that an MT dictionary that does not possess the capability of easy modification is already obsolescent.

"SEMANTIC CODES"

The use of semantic codes has been proposed when an input-language occurrence or phrase is ambiguous in the output language. These codes, attached to the input-language items in the dictionary, are intended to apply to the residue of problems that cannot be solved with other techniques (e.g., the idioglossary). Many problems of ambiguity are connected with high-frequency general-language words, translatable only on the basis of specific word combinations (if at all). Decisions must be

made in terms of context, which means, for the most part, on the basis of syntactic combinations. For example, the translation of a multiple-equivalent noun used as subject of the sentence will often depend on the kind of verb for which it is subject. The kind of verb will be indicated by the semantic code. The study of syntactic combinations for purposes of establishing semantic codes is only beginning. Again, considerable storage space in the dictionary will be required, and the dictionary must have the capability of easy modification as new evidence is acquired.

REPRESENTATION OF OUTPUT-LANGUAGE ITEM

The output-language item may be stored in MT dictionaries in one of two locations: as a part of the record that contains the input-language item or in a separate dictionary. In the latter instance, the lexical item is represented by a code which serves as the address of the item in the separate dictionary. In either instance, we have the same problem as with the input-language item: Shall the item be represented by full forms or by stems? For English output, where the inflectional problem is less severe, some groups have proposed a form dictionary. Others have stored an English canonical form with a code which will permit inflection of the item in accordance with the rules of English grammar and orthography. The problem is a minor one, and should be decided on the basis of appropriateness to the whole MT system. It may be difficult, for example, in a given context to decide whether to add the ending *-ed* or *-ing* to an English verb; the decision must take into account the input sentence and the requirements of English grammar. The implementation of this decision, however, is a simple matter. The techniques for carrying out such a decision are well known and widely applied.

The chief problem of the output-language item in a bilingual MT dictionary is again one of content: Which equivalents should be stored for the input lexical item? Experience has shown the danger of placing too much reliance on published dictionaries, even "technical" dictionaries, as a source for the output items. The problem may be illustrated by citing two entries from Callaham's *Russian-English Technical and Scientific Dictionary:*

Probeg m. run, mileage; race; *ispytanie -om* road test.

Prodolzh/at' v. continue, go on, go ahead, proceed; carry on, pursue, persist; resume; prolong, extend; elongate, lengthen, broaden; *-at'sya* v. continue, last, be prolonged, be extended.

In the translation of Russian physics texts at the RAND Corporation, the noun *probeg* occurred 157 times; the equivalents chosen were either *range* or *path.* The verb *prodolzhat'* occurred 16 times, always with the translation *continue.* In the case of *probeg,* it is clear that the published dictionary is not an idioglossary for physics (which, of course, it does not

pretend to be). In the case of *prodolzhat',* we have "too much" information: Although the wealth of synonyms and nuances may be helpful to the human translator, they will be useless to a machine-translation system unless a selection procedure is provided. The customary procedure in such instances is editing—i.e., the arbitrary selection, by the MT researcher, of a minimum number of English equivalents. The selection process may be more or less successful, depending on the experience, judgment, intuitive powers, and luck of the editor. In any event, his choices must be subjected to verification (text processing) before they can be declared valid, or before any remaining problem of ambiguity can be studied. There are good grounds for arguing that the assistance of the best editors in the world does more harm than good in building of MT dictionaries.

The sad truth is that very little progress has been made in this area of MT research in the past ten years, despite the fact that it is an easy and obvious place to begin. A survey of the literature indicates that little has been done in the building of idioglossaries since the early work of Oswald and Fletcher.[3] Although a substantial amount of text has been processed, no one has yet produced a good idioglossary, nor made a systematic study of the coverage provided by glossaries of different size and composition, for texts of a specified nature.

In summary, consider the following important points brought out in the preceding discussions.

1 A great deal of work has gone into the preparation of MT dictionaries. Reasons for the concentration of effort in this area include the following: *(a)* the fact that dictionaries are a prerequisite for text processing, whether for purposes of translation or research; *(b)* the assumption that existing dictionaries and grammars are adequate and suitable for MT dictionaries; *(c)* the belief that machine translation is essentially an engineering problem, solvable, at least in part, by the application of engineering techniques.

2 The chief developments in MT dictionaries have been in terms of the text-lookup function: Is the dictionary big enough and fast enough for the job?

3 The desirability of performing syntactic analysis on the input sentence is now generally accepted. Dictionaries that serve only the lookup function have limited usefulness in the area of language processing.

4 Present and future MT dictionaries must have the capability of easy and cheap modification. As more grammatic and semantic information is acquired (from whatever source) about the items in storage, it will be stored in the dictionary. In essence, the dictionary is a tool for research.

[3] *Cf.* V. A. Oswald, Jr., and R. H. Lawson, "An Idioglossary for Mechanical Translation," *Modern Language Forum,* vol. 38, nos. 3/4, pp. 1-11, 1953.