[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

DICTIONARIES FOR MACHINE AND MACHINE-AIDED TRANSLATION (MT & MAT)

Prof. Frank Knowles
(Aston University)

The task of MT/MAT is to take input, i.e. source-language, text (usually — regrettably — in the form of individual sentences) and to turn this material into target-language utterances which preserve sense. No more, no less!

The MT/MAT process has two basic facets: textocentric and lexicocentric, sometimes referred to as being, respectively, dynamic and static. The fundamental feature of life here is to carry out the following actions in the following order:

Look up what you can.
Compute, i.e. "calculate", what you can.
Interact with a human "bystander" to resolve cruces.

The division of labour among the above three options varies from one system to another and, for certain types of problem, two-pronged strategies are not precluded. However, the bigger and more cross-referenced and structured the MT/MAT dictionary — or lexical database (LDB) — is, the better.

One crucially important parameter is the nature or type of the text being processed: is the textual material syntactically, stylistically and lexically varied or is it homogeneous and "stationary", as is nearly always the case with text composed on a particular and narrowly professional — engineering, science, medicine, economics — theme?

How are LDB's for MT/MAT created, configured, enhanced, calibrated and deployed? Firstly, a LDB's structure and configuration for English, say, as source-language is not the same as for English as target-language. Hence, much depends on the linguistic structure of the languages involved and on whether they are sources or targets.

There are only three methods of structuring syntactic meaning: inflection, the use of function words, and element order. Various languages employ different "product mixes" of these methods and the particular structures of all natural languages, in terms of both form and contact, display asymmetries and mutual non-orthogonality: this means that they are distinctly suboptimal, intrinsically and notably when confronted with the exigencies of the translation process. If this process is to be performed entirely, or even largely, by machine the complexity factor looms very large. One of the chief difficulties revolves round the segmentation of the input text: algorithms which carry out this stage work through their segmentation task on the basis of orthographic words encountered in the text. This sounds reasonable enough but it is only a beginning: what must be striven for is the isolation of all **cognitive units,** having a properly referential meaning outside the particular text, indeed outside any text. Many of these cognitive units will be bigger/longer than one orthographic word, especially in the case of English. The hope is that such cognitive units will also turn out to be **translation units** as well! It is hence the duty of the LDB to identify the cognitive units in the source text and to make available their translation equivalences for insertion in the target text.

The act of configuring LDB's for MT/MAT in this way and style is a problem of no mean order because the subtleties involved in differentiating between extra-textual meaning(s) and sentential sense are always present. This difficulty has led to increased concentration on the idea of constructing and then analysing extensive translation corpora — this, of course, "off

line" from any active MT/MAT system but, rather, a necessary preparatory step in setting up an appropriately structured and replete LDB.

The configuration of this LDB would normally be bipartite: a monolingual module for the source-language and a transfer dictionary for the systematic replacement/insertion of the chosen target-language elements. Occasionally, other smaller modules might be present: it is always necessary during the generation phase of the MT/MAT process to massage the chosen target-language lexical units into a form appropriate for their syntactic environment. It is possible to separate out this function from the transfer dictionary. It is also possible to introduce special sub-language dictionaries or "topical glossaries" in case of need — this happens frequently. It is moreover worth pointing out that the "listing structures" of MT/MAT LDB's are mostly of little concern, although frequency considerations sometimes have an influence on matters. What is important, and vitally so, is total traversability, or achieving the maximum possible number of search pathways so that the information resident in the LDB is retrievable, as and when it is needed, without any topographic problems.

The types of information recorded in LDB's for MT/MAT are quite numerous: in the analysis dictionary, clearly, each lexical unit needs to have appended to it a multitude of data. In terms of morphological information a set of accompanying relevant variables needs to be present with appropriate values set, e.g. "count noun" v. "mass noun". Valency information needs to be slotted in for all verbs and also for many nouns and adjectives etc. Semantically-based distinctive features (DF's) can also play a vitally useful disambiguating role in the translation process and hence belong in the LDB too. Statistical information appertaining to word frequency, both grosso modo and micro-environmentally conditioned, is also recognised as a powerful feature. Each headword also needs to "cross-refer" to any additional or alternate forms it may have , e.g. US v. British spelling. One special feature of LDB's for MT/MAT is the need to "carry" extensive lists of proper names, personal, geographical, plus commercial trademarks and product brand names. Common and not-so-common abbreviations need to be fitted in as well. Occasionally, encyclopaedic information is also helpful, particularly where the organisational infrastructure — and the nomenclature that goes with it — between two countries does not coincide.

If the MT/MAT system in question is involved in translating special subject material — and this is the predominant case — then a terminology module needs to be incorporated into the LDB. The entries in this module may represent material developed in house or acquired from standards bodies — be that as it may, it is hardly likely to remain static. Each translation transaction is likely to leave behind in a special file a list of "words not found": many of these will be technical terms and they must be entered into the LDB along with their counterparts in the target language, but only after the most rigorous check on quality control. It is common for further information to be recorded — for human personnel this time, rather than for the machine — relating to these terms: their source, validity, date of acquisition, coder concerned etc. etc.

Of very great significance are the benefits to be derived from lodging in the LDB copious information, cross-referenced, relating to derivations, compounds (especially verb + particle, in English), multi-word units, idioms and even proverbs. For humans, this open the eyes — and the door — to the linguistic truth that the greater the "expert knowledge" available to the LDB, about statistical "bonding" of lexis, about matters such as the colligation and collocation of words, the more reliable and authentic translations produced by machine are going to be. Here the name of the game is to conduct a very fine and extensive juxtapositional analysis of actual texts, preferably — perhaps — those actually being translated "in anger", so to speak. In this way, and in this way only, can the quality of a LDB be rapidly enhanced — but only if the organisation involved is wise enough to have lexicographers-cum-terminologists on its payroll!