A MULTILINGUAL COMPUTERISED DICTIONARY FOR MACHINE TRANSLATION

J Douglas Clarke

Department of Applied Computing and Mathematics
Cranfield Institute of Technology

When working in Machine Translation (MT), one becomes increasingly aware of the importance of a good dictionary (in addition to good MT software itself) to ensure the best possible quality of translated text. The quality, i.e. accuracy, of the machine-translated text can be no better than the quality of the computerised dictionary being used for the translation. The importance of the design of the dictionary can therefore never be under-estimated. The importance of a bilingual dictionary for good quality machine-translation is, no doubt, matched by the importance of a monolingual dictionary for good-quality monolingual text-processing. Many characteristics required for a monolingual dictionary will undoubtedly also be required for each monolingual component of a bilingual or a multi-lingual dictionary.

When constructing a computerised dictionary, we do not necessarily merely have to convert a conventional printer (paper) dictionary into a computerised version with the same format or the same layout. This would be a distinct disadvantage, in view of the facilities available in a computer which are not available if use is made of the medium of the printed page.

The medium of the printed page is too restrictive and allows us, for example, to prepare a dictionary e.g. only as a linearly arranged (alphabetical) sequence of items, without being able to incorporate other advantageous arrangements, as can be done in the case of a computerised dictionary.

One example of a way we can break away from the restriction of a 'printed-paper' layout is in the broad structure of a bilingual dictionary.

In a conventional printed bilingual dictionary, the entries, in alphabetical order, of language L_1 are mapped across to the entries of language L_2 (the mapping being either one-to-one, or one-to-many or many-to-one). This printed dictionary will also incorporate, in a subsequent section, the entries, in alphabetical order, of language L_2 mapped across to the corresponding entries of language L_1.

In a computerised bilingual dictionary, the entries of L_1 and L_2 need only occur once, the mapping, i.e. cross-references, between them being accommodated by the internal structure of the computer.

Such a computerised bilingual dictionary can be extended to a multilingual dictionary, e.g. by incorporating an alphabetical list of entries for another language L_3, which are cross-referenced to the corresponding entries for each languages L_1 and L_2.

Convention and human psychology doubtless require that, at the person-computer interface, a monolingual dictionary is also seen, or represented, as a linearly structured alphabetical sequence of entries.

Otherwise, no restrictions (need) apply against using other structures where appropriate. Thus a monolingual dictionary can operate as the 'front-end' to other structures better able to represent linguistic, logical and real-world relationships, and thereby realise improved quality in both text-processing and machine translation.

These structures should also be properly regarded as constituent parts of the dictionary. They are nevertheless usually considered as being behind the interface, and not necessarily seen by the human user.

A dictionary for machine translation

A procedure for machine translation is described in an earlier paper. In the case of simple sentences this procedure involves, for each source sentence,

(i) the determination, from the source dictionary, of the grammatical categories of the constituent words in the sentence;

(ii) the syntactic analysis of the sentence, using the stored production rules representing the grammar of the source language;

(iii) the semantic analysis of the sentence;

(iv) the stored representation of the tree of the sentence as a data-structure in the computer;

(v) the application of the transfer rules to form the tree of the target (translated) sentence;

(vi) the determination of the target words, in the target sentence, from the target dictionary.

Such a procedure requires, as minimum information about each word in a monolingual dictionary,

(a) the grammatical category corresponding to that word;

(b) the set of semantic features

representing the formal definition of that word;

(c) the reference pointer to the target entry in the target dictionary;

in addition to other information, including e.g. the stem or root of the word entry, and the informal definition of the word.

The procedure described for machine-translation can be extended to cases where word-for-word translation does not apply, e.g. by the incorporation of phase-trees in the sentence-trees created in the computer data-files.

The incorporation of phases as entries in each monolingual dictionary is thus advantageous, indeed necessary, particularly in any case where a phrase represents a unit of meaning. The above list of minimum requirements should accordingly be extended to satisfy phrase entries, in addition to word entries, in each monolingual dictionary.

Although we may merely follow the same format as in a printed dictionary for including phrases, there is nevertheless again no restriction for doing so in a computerised dictionary.

Phase trees

Having used tree structures to represent sentences in Machine Translation, it was thought to be interesting to explore the possibility of using tree-structures as components of the dictionary to represent phrases - with the object of improving the quality, i.e. accuracy of the translation.

In this scheme, each word (and associated meaning and grammatical category) occurs once in the dictionary, in its correct word entry; it does not occur once in each quoted phrase containing it (as in a printed dictionary).

To achieve this, each word entry, occurring once, occurs as the 'leaf' of one or more trees - as many trees as there are phrases recorded with that word in it.

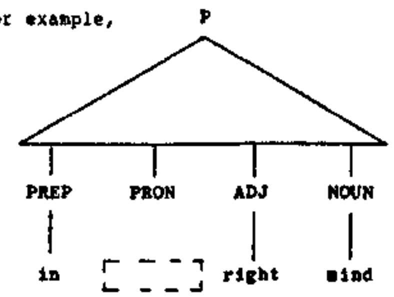
These tree structures may be in data files, or perhaps represented as Prolog statements.

<u>Phrase categories</u>

By extending some branches of phrase-trees only as far as grammatical categories, rather than on to individual word entries, we may represent 'phrase categories' in the dictionary.

Each phrase category represents a whole class of phrases, delimited only by each of the grammatical category leaves of the tree.

For example,

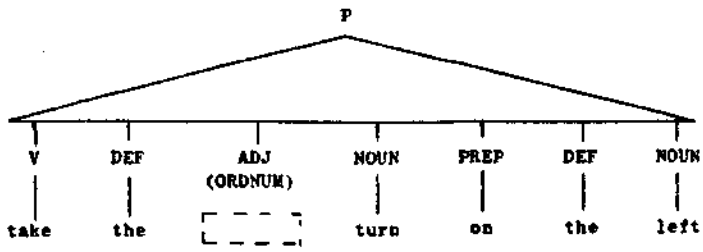


(Schematic representation)

(where a personal possessive pronoun can be inserted in [ ]) represents a category of phrases including

in his right mind,
in my right mind,
in her right mind
etc,

Similarly,



(where any ordinal number, e.g. "first", "second", "third", ..., can be inserted in [ ]) also represents a phrase category tree.

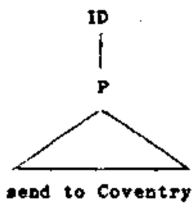
<u>Properties of, and relations between, linguistic units</u>

We may represent the properties of a linguistic unit (e.g. word (W), phrase (P) or sentence (S)), and the relations between such units, by the ue of meta-linguistic operators or functions. Among the functions proposed and used here are:

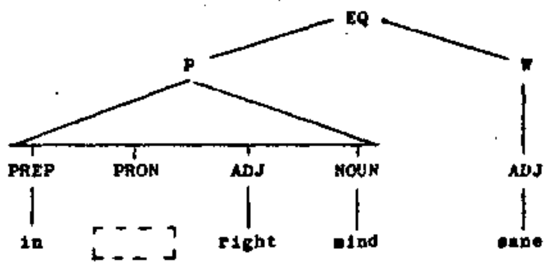
EQ (equivalent)
NEQ (not equivalent)
CONV (converse (not NEG since this is already used as a grammatical category))
NCONV (not converse)
ID (idiom)
PR (proverb)

Properties of linguistic units, and relations between units, may be represented in the computerised dictionary by embedding the (tree) representations of these units in extended tree-structures.

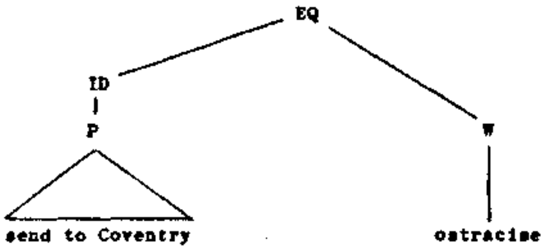
For example, a phrase-tree P representing an idiom may be subtended from a further stem 'ID' as in



The equivalence of two units may be represented by subtending these units, e.g. as (sub-) trees, in a tree whose main stem is an item EQ for that particular equivalence relation, for example:



Another example is:



Word-for-word translation of

in his right mind

may lead to a poor, or bad, translation in the target text. On the other hand, preliminary scanning of the corresponding EQ tree by the MT software procedures, in the pre-translation stage, can isolate the equivalent term 'sane' which may lead to a safer, and more accurate, translation.

The idiom 'send to Coventry' is similarly safer to translate it replaced by the equivalent term 'ostracise', similarly located by a scanning procedure.

In some cases, however, it may not be necessary, for purely translation purposes, to link an expression to an equivalent, or near-equivalent, expression whose translation is known. For example, the English proverb

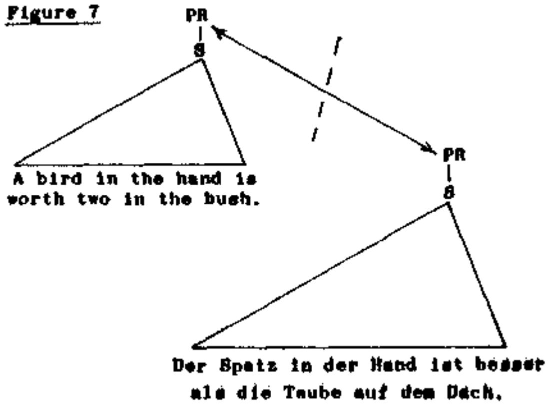
A bird in the hand is worth two in the bush.

can be assumed to be so close in meaning to the corresponding German proverb

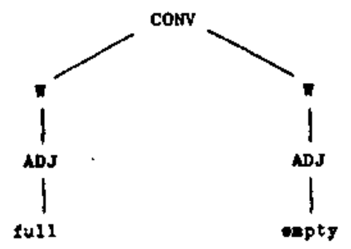
Der Spatz in der Hand ist besser als die Taube auf dem Dach.

that these expressions, i.e. their tree stems, may be cross-referenced directly between the corresponding English and German dictionaries in the computer.

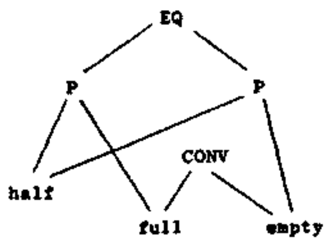
Figure 7



An example in which CONV occurs is:



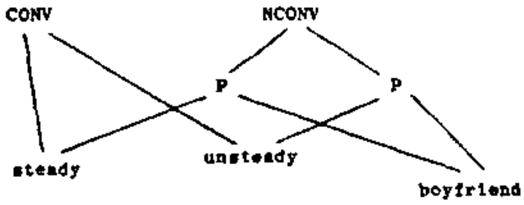
The structure showing the relation between 'half full' and 'half empty' can be incorporated in the above structure:



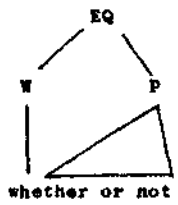
This last structure reflects the real-world situation that "half-full" and "half-empty" are (factually) the same. Nevertheless, this does not take into account nuances of meaning, occasioned by the view of that world as seen by the speaker or writer who, at one time states e.g. that a bottle is half-full, and, on another, states that it is half empty.

These nuances may perhaps be determined by accessing the informal definitions of "half-full" and "half-empty", using the operation "DEF"

Another interesting example is one of 'opposites which aren't' (:):



Another example is:



The MT pre-translation procedure, scanning this structure in the dictionary, will be able to replace the phrase

whether or not

by the equivalent word 'whether', which is more likely to be more safely and accurately translated into the target text.

Such structures can be interrogated mechanically, e.g. in the MT procedure as exemplified in the last 2 or 3 pages.

Alternatively, they can be interrogated by a user, in interactive mode, by keying in such questions as:

IS "send to Coventry" ID?
EQ "send to Coventry"?
CONV "full"?

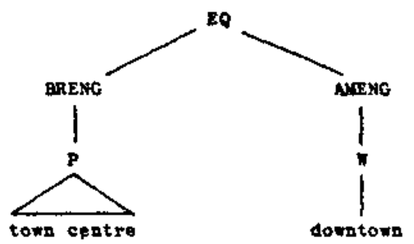
IS "full", "empty" EQ?
DEF "full"?

the corresponding answers being output on the user's terminal screen.

The operation DEF is intended to output the (informal) definition of the linguistic unit " " requested. The stem of the definition tree is in the entry for that linguistic unit.

The operation "DEF" could be used in cascaded mode, to elucidate the definition already given. Although this could be a considerable facility for the user, undoubtedly the cascade would eventually be a "circular" cascade.

Other operations or functions which could be employed are those showing national variations in a language, e.g. differences (or similarities) between American English (AMENG) and British English (BRENG) or differences between Castilian Spanish (CASP) and American Spanish (AMSP), e.g.



With this last facility, the user may initially opt, in any machine translation run, for the translation to be between specified national versions of source and/or target languages. With such an option, or options, specified to the computer, the machine translation software can seek out, from the dictionary, the appropriate national variants - where appropriate and where they occur.

<u>Some advantages of the multi-lingual dictionary</u>

Not all aspects of the dictionary design can be covered in one paper. Nevertheless, the dictionary, as described, has the advantages of being

modifiable
updatable
extendible

In addition, further structures can be built (into) it, e.g. those which enable it to be used as a rhyming dictionary.

Also, further components can be appended to each entry, e.g. the codified phonemic features, which allow the translated text to be output in spoken form.

Although the front-end of each monolingual dictionary may be an alphabetical list of (word-) entries, additionally other front-ends may be built onto the same data-structure of the dictionary, e.g. where entries are required (to be accessed) on a category and sub-category basis.

The system is flexible. The dictionary could be used in automatic mode for machine translation. Alternatively, it could be used 'manually', where the user, e.g. human translator, can access the dictionary via a terminal by keying in a query using one of the operator or function codes described above. The appropriate response would be output on the same terminal.

In each constituent monolingual dictionary, all phrases containing a given (key-) word may be found via that word. In many cases, this feature probably does not occur in a printed dictionary, where each phrase will occur once in the dictionary, in an entry under just one or another of the (key-) words in the phrase. Thus user-access to the phrases containing a given word is more readily obtained on the computerised dictionary.

For a similar reason, a somewhat similar feature is that a phrase may be accessed via any constituent (key-) word of that phrase, again offering to the user ready access to the phrase.

The dictionary described, interacting closely with the processing and translating software, is designed to give not only high quality of text in monolingual processing but also high accuracy of translation in bilingual processing.