A Semantic Knowledge-based Computational Dictionary

Mohammed Y Al-Hafez♣, Douglas Clarke♦

& Alfred Vella♠

♣Higher Institute of Applied Science and Technology, Damascus. Syria,
♦Cranfield University, UK,
♠The University of Luton, UK

Abstract

The Computational Dictionary, described in this paper, is structured on a knowledge base.

The semantic features of each word, in a relevant grammatical category, can be determined through a hierarchical tree structure. Semantic knowledge of verbs is represented using predicate calculus definitions. This allows each expression, e.g. sentence or command, to be tested in order to determine whether it is meaningful and, if meaningful, what its meaning is or indeed whether it is ambiguous.

The dictionary is designed on a modular basis so that it can be used

1) in automated mode, in conjunction with any one of a variety of Natural Language Processing systems;

2) in interactive mode, for manual access even by a naive user, i.e. in dictionary look-up;

3) in output mode, available for printing machine produced dictionaries in any one of the variety of chosen formats and combination of features.

One of the applications of the dictionary in automated mode is in Machine Translation. This application is discussed in this paper.

1 Introduction

A computational dictionary is not limited to the restrictions of a traditional printed dictionary. Two clear advantages come immediately to mind:

1. it is updatable, even interactively by the user to his/her own copy;

2. its structure can lend itself to any one of a choice of multi-access modes; not only alphabetic access but alternatively using access by semantic category; or by grammatical category or using phonetic access (Clarke).

The term “computational dictionary” (CD) rather than computerised dictionary is used advisedly; it is intended to indicate that it is a system that forms part of the computational process rather than being a purely ‘static’ knowledge base.

The research described is a comprehensive approach to building a semantic knowledge base as a global and local hierarchical tree. The knowledge base is analogous to Roget’s plan of classification.

In this knowledge base, one may use symbols instead of words at the nodes of this tree. In the system described, the symbols are represented by English words. These symbols may be changed according to the language used, e.g. into Arabic words if the system is specifically designed for use in Arabic. This one-for-one replacement nevertheless only applies in the simplest of cases. More involved relationships are discussed in section 5.1 below.

2 Semantics

The importance of semantics in Natural Language Processing (NLP) cannot be stressed enough. It is this consideration that has led to the structured design of this semantic knowledge-based dictionary. This will then cater for the implementation of semantic knowledge - an essential element of every NLP application(Omar).

The system also incorporates a morphology sub-system - necessary when catering for a highly inflected language such as Arabic (figures 6 -1 and 6 - 2).

At least six types of meaning are generally recognised (figure 6 - 3). Of these, the lexical meaning can be found in a traditional printed dictionary. Another type of meaning, the conceptual meaning is concerned with communication as transaction of thought. Nida defines this type of meaning as the only one which is related to the lexical unit.

e.g. woman [+ human - male + adult]

As will be described below (section 4.2) a traditional dictionary definition can be derived, i.e. generated, from the cognitive meaning.

2.1 Semantic Field Theory (SFT)

A semantic field is a group of words which are related by their semantic meaning as illustrated in figure 6 -4, and classified under one word (Ullman, Lyons).

SFT encourages the building of a comprehensive dictionary which incorporates all existing semantic fields (SFs) in a sequenced hierarchical structure, depending on the hyponymy relation. Examples are Roget's Thesaurus and, more ambitiously, the dictionary presented by Warlburg and Hally in 1952 (Ullman).

Difficulties arise in deciding the headwords and in classifying headwords in an SF. Difficulties also arise in:

1. specifying the language fields;

2. distinguishing between headwords and marginal words inside one field;

and 3. defining the relation(s) between words inside each field.

The depth of meaning of each word, and its hyponymy to other words, suggests a scale to govern the importance and influence of each word in a field and to determine which is the headword.

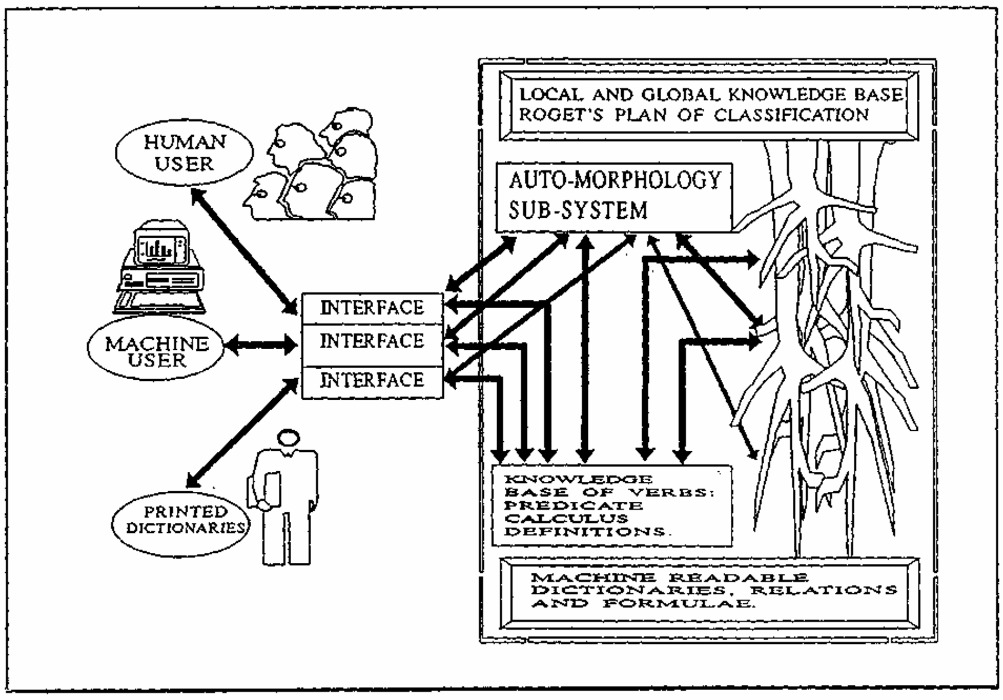
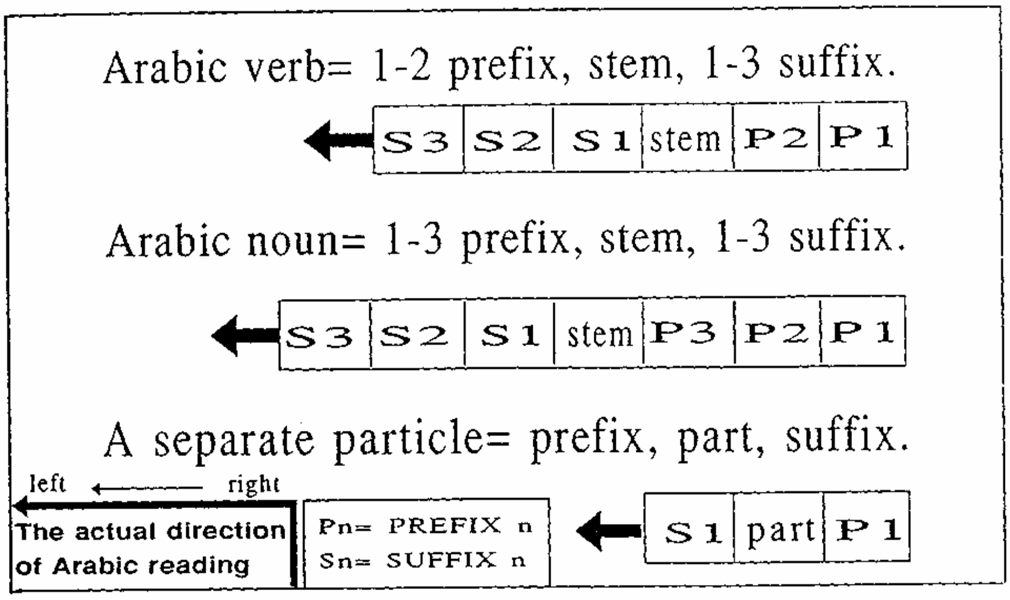

LOCAL AND GLOBAL KNOWLEDGE BASE
ROGET'S PLAN OF CLASSIFICATION
HUMAN USER
AUTO-MORPHOLOGY SUB-SYSTEM
INTERFACE
INTERFACE
INTERFACE
MACHINE USER
KNOWLEDGE BASE OF VERBS: PREDICATE CALCULUS DEFINITIONS.
PRINTED DICTIONARIES
MACHINE READABLE DICTIONARIES, RELATIONS AND FORMULAE.

Figure 6 – 1: A semantic knowledge based computational dictionary


Arabic verb= 1-2 prefix, stem, 1-3 suffix.

Arabic noun= 1-3 prefix, stem, 1-3 suffix.

A separate particle= prefix, part, suffix.

Pn= PREFIX n
Sn= SUFFIX n

Figure 6 – 2: The components of an Arabic word

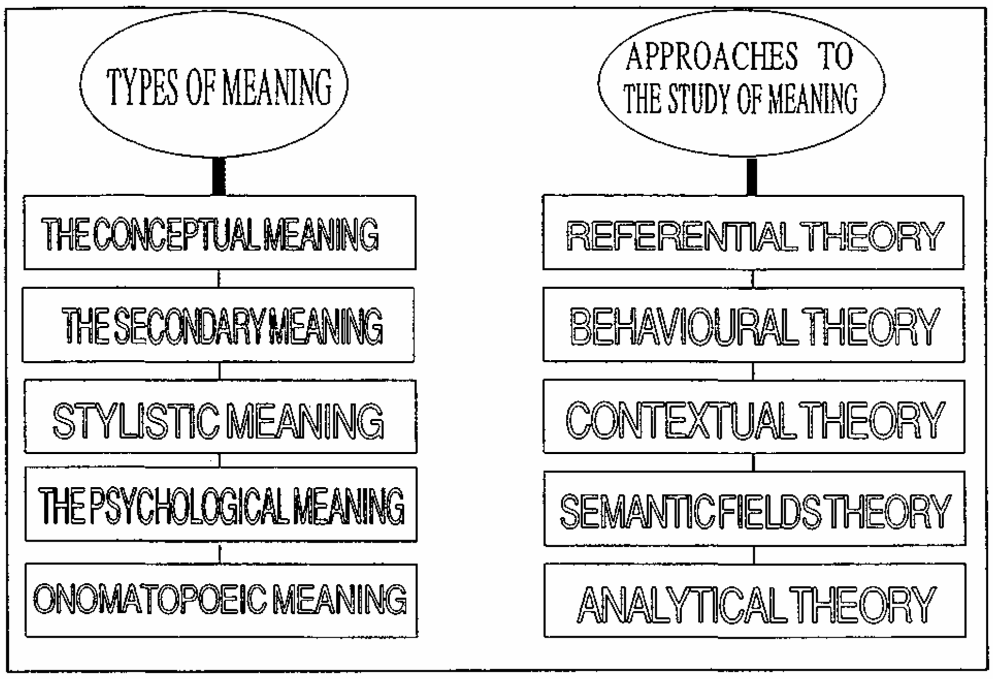


Figure 6 – 3: The theoretical bases of semantics

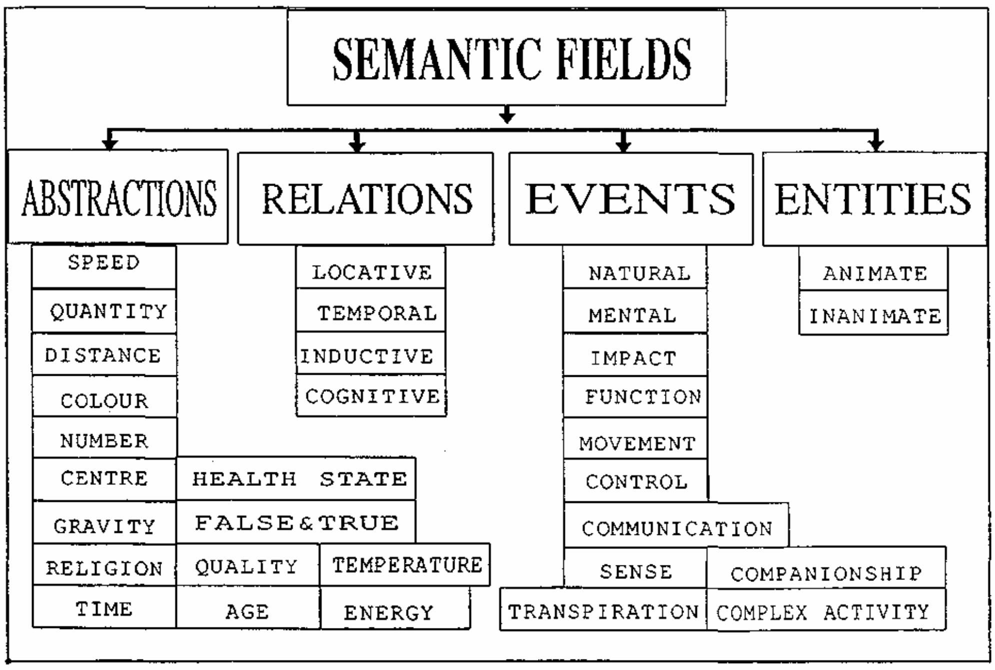


Figure 6 – 4: Semantic fields

The criteria used in determining whether a word is a headword include the Kay and Berlin criterion and the Montague and Batting pattern (Leech GN).

SFs may be conveniently divided into

1. the continuous
2. the discontinuous

and 3. the abstract .

2.2 Importance of SFT

This importance arises from consideration of a number of features including:

1. the relations, similarities and even contradictions between the headword and other words of one SF;

2. a list of words for each subject;

3. the structured representation of language vocabulary;

4. the distinction between homonymy and polysemy;

5. the facility of being represented in a hierarchical tree.

2.3 Analytical Theory (AT)

The analytical approach examines the meaning of a word in successive stages:

1. Analysis of all words of an SF and definition of the interrelations of their meanings;

2. Analysis of words which retain a common utterance (polysemy) to their constituents, and determination of the difference between their meanings;

3. Analysis of each meaning reduced to its absolute constituents.

AT is considered as an extension of SFT.

The principles of the two theories (SFT and AT) are used to construct the CD, which is arranged according to SF's. Accordingly, lexical entries will be searched according to the headwords.

2.4 Lexical-Semantic (LS) relations (LSRs)

The relations between lexical elements and their meanings provide an effective and compact structure of the CD. These relations will overcome the lack of association between semantic and syntactic information, which would otherwise be the case.

Word-relation-word triples (predicates) can be used to build this CD. Such LSRs include those of synonymy and antonymy and such relations as 'part of' and 'is a'.

LSRs are represented in this knowledge-base, by 'vertical' (hierarchical) and 'horizontal' linkages. Vertical linkages represent such relations as 'part-whole', 'part-of' and 'kind of' relations, and they represent the hyponymy relations between words in that SF. Other relations such as 'typical relation', 'typical result', 'characteristic sound' and 'made of' are horizontal linkages.

These relations represent the link between syntagmatic fields between different semantic fields.

The vertical and horizontal linkages are two 'overlays' that form the semantic network and are available for the purpose of retrieval. Antomony and synonymy relations are supplied adjacent to the appropriate lexical entries (Evans).

3. Further Techniques

Other techniques incorporated in the design of this dictionary include:

1. methodology of knowledge bases (MKB);
2. representation of predicate calculus definitions of verbs (PCDV);
3. Roget's plan of classification (RPC);

and 4. the analysis of machine readable dictionaries.

3.1 Declarative schemes; Semantic networks

In a semantic net, the program can start at a node of interest and (then) follow links to related nodes and through to more distant nodes.

Inferences drawn by examination of the net are not necessarily valid. The interpretation (semantics) depends solely upon the program that manipulates the network. This requires a strong organising principle.

3.2 Semantic primitives

One of the objectives of this research has been to reduce the relations to a minimum number of semantic primitives, i.e. terms which do not overlap in meaning.

3.3 Machine readable dictionaries

Machine readable dictionaries (MRD's) such as Webster's Seventh New Collegiate Dictionary and Longman's Dictionary of Contemporary English (Bran Boguraev) exist as computer files from which it is possible to analyse and extract lexical knowledge by computer (Uri Zernick).

The computational dictionary being designed incorporates all the features and uses of an MRD, but has the advantages of using symbols and primitives in the saving of memory. So, for example, no definitions are stored in this dictionary, but these can be generated from the semantic net as described below. Also, the dictionary can be used for inference and decision-making.

3.4 Roget's plan of classification

In Roget's plan, the items are classified according to objects and concepts (semantic fields). For this, a compact methodology is required such that knowledge is incorporated without contradiction, redundancy or circularity of definition.

3.5 Predicate calculus definitions of verbs (PCDVs)

A PCDV specifies, for each verb, its generic form and the categories of the agent, instrument etc.

More specifically the PCDV includes:

I. The verb itself as the first argument.
II. A list of synonyms and list of antonyms.
III. A list of relevant semantic primitives.
IV. A list of verb features.
V. The domain in which the verb is used.
VI. The number of the verb's definition formula.
VII. The case structure of the verb.
VIII. The linkage between the verb and its original node in the knowledge base.
IX. The generic verb to be used in the verb definition.
X. The instrument used to fulfil the act.
XI. The selection restrictions of the deep structure components of the sentence with such a verb, i.e. the agent, the first recipient and the second recipient if they are applicable.

The agent, instrument etc. are each the most superordinate entity i.e. they are the hyponymy words which can be compatible with the verb.

To determine whether a sentence is meaningful (which is usually the case) and, if so, what its meaning is, the agent and instrument of its constituent verb must be found (such that they are) compatible with those specified by the PCDV of that verb (Otany Miyo), (Aoe Jim-ichi).

3.6 The choice of programming tool

The programming language used was Turbo-Prolog 2-0 from Borland International Inc. This most easily expresses operations on the data-structures involved in this dictionary.

4 The Dictionary

Relevant aspects of the design of this dictionary are:

1. the structure of the knowledge base;
2. the data structure
3. the generation of definitions

and 4. the testing of the meanings of sentences.

The dictionary can be used in three modes:

1. interactive mode, i.e. for manual access by a human user
2. fully automated mode, i.e. supporting an NLP system

and 3. in output mode, i.e. to produce printed machine readable dictionaries.

In fully automated mode, the dictionary (SKBD) can support one of a range of NLP systems including a machine translation system.

This semantic knowledge based dictionary (SKBD) incorporates all the usual features, found in MRDs, for each entry,

i.e. its pronunciation

its grammatical category (ies)

and, for each meaning,

its definition(s)

its significance

its context(s)

its domain(s)

its semantic features

its semantic primitives

its synonyms and antonyms

and, in the case of each verb, the selection restriction and the case structure of events.

4.1 A semantically comprehensive dictionary

The dictionary provides, for a concrete noun as symbol, a full hierarchical tree, constructed from nodes and branches, and designed to depict classified knowledge of the world.

Each node carries a symbol which refers to a (corresponding) noun. The nodes have additional features as primitives. These primitive features are inherited from the node itself and from its superior nodes in the same branch.

For such noun symbols, the dictionary also provides relevant relations with other symbols in horizontal linkages. Also provided are lists of stylistic selections.

The dictionary provides, for a verb, a predicate-calculus definition of the verb (PCDV). For example, the PCDV of the verb 'hit' is:

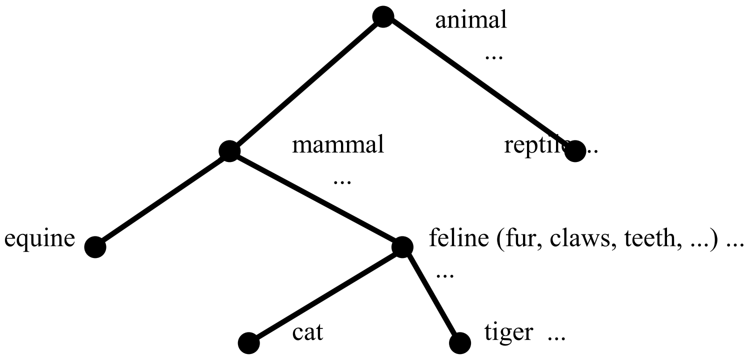
```
2.1 verb {
hit
link word to the knowledge base : impulsion
Generic verb      : strike *
Agent             : X
1st Object        : Y
2nd Object        : no
Instrument        : Z
Domain  : O
Synonyms:[strike, knock off, shoot, slug, crash, bat, crash, smash, beat]
Antonyms          : []
Semantic feature:[location - alteration]
Verb attributes:[attentive, volitional, implemental]
Domain, global or local:[punishment]
Destination       : [object]
is_ a (X, animate, [vital])
is _ a (Y, existence, [object])
is _ a (Z,[Vital_part_of_body, stick, rod, whip, pelt, hammer, natural forces]),
```

case structure: AGENT, RECIPIENT, INSTRUMENT. AGENT, RECIPIENT, INSTRUMENT, RECIPIENT INSTRUMENT in passive voice}

4.2 Generation of definitions

With the standardised convention for the tree-structure, described elsewhere in this paper, it is possible to use a fairly simple algorithm to generate a 'natural language' definition of a word, if required by a user.

For example, the animal cat occurs in its part of the tree structure (in simplified representation) as:



On request, by the user, for the definition of the word 'cat', the system will produce the definition

A cat is a feline mammal with fur, claws and teeth.

This definition compares well with a conventional dictionary definition of 'cat'.

4.3 Implementation of PCDVs

Consider, for example, the three sentences

- The cat drank a saucer of milk.
- The captain hit the sailor with the cat.

and - The crew sailed the cat out of the harbour.

The entries for 'cat' in the dictionary are:

* feline; (animate)

* whip; (instrument for flogging, tool for punishment)

* boat; (transportation on water)

The second entry is the shortened form of 'cat-o'-nine-tails', as used in the (originally nautical) expression "There's no room to swing a cat". The third entry is the name of a kind of boat.

Further, more recent, entries could include:

* shortened form, used for example in the Antarctic, of 'snow cat', a means of conveyance

* shortened form, used in the motor-trade, of 'catalytic converter'.

For the first sentence, the PCD of the verb ‘drank’ requires the agent to be animate and the objective to be liquid and consumable. Only the first entry of ‘cat’ above is consistent with this requirement for agent. Also, the requirement for the objective can be satisfied by ‘milk’ but not by ‘saucer’. So access to the dictionary can only allow for this sentence to be interpreted as

- The cat drank the milk from a saucer.

In the case of the second sentence quoted, the PCD of the verb ‘hit’ has, as one of its arguments, an instrument of hitting. Only the second entry of ‘cat’ above is consistent with this requirement. So this will be the entry assumed in interpreting this sentence.

In the third sentence, the verb ‘sail’ would be found from the dictionary to have a PCD including the objective ‘boat’. It would be found, from the corresponding hierarchical tree, that only the third entry of ‘cat’ above would satisfy this requirement.

Thus, in each of these three sentences, use of the CD will elicit the corresponding appropriate meaning of ‘cat’, uniquely and unambiguously.

There are situations however, in which a meaning cannot be determined uniquely from the sentence itself.

Consider now, for example, this fourth sentence.

- The sailors took the cat out of the harbour.

which is like the third sentence above but with the verb ‘took’ replacing ‘sailed’. Here there is a choice of case of direction in relation to harbour, i.e. taking out through the harbour entrance or taking out from the water in the harbour. This choice of case of direction should be catered for by the hierarchical tree for ‘harbour’ in the dictionary.

The former possibility will result in the same meaning of ‘cat’ as that in the third sentence above which included the verb ‘sailed’.

The latter possibility would allow for almost anything to be taken out of the water, including e.g. anything from feline cat to catalytic converter cat. Thus here the PCD of the verb ‘took’ would allow for any of the meanings of the word ‘cat’ to be extracted from the CD.

In such a case the analysis would have to go beyond the sentence level, e.g. by determining, whether reference to the same item ‘cat’ occurs in the preceding or succeeding sentence(s) and thereby using the CD to determine its meaning.

In addition to nouns and verbs, the dictionary also provides appropriately for entries having other grammatical categories, e.g. for adjectives, adverbs and prepositions.

Prepositions, as Margaret Masterman of CLRU has pointed out, present a major problem. Considerable care has to be taken in determining the meaning of each preposition occurring in a text. In a typical English dictionary there are, for example, nineteen different meanings of ‘by’ and seventeen meanings of ‘to’. If ‘by’ occurs somewhere in a text, which of the nineteen meanings is intended?

Consider, for example, the sentences

- The dead body was found by the front door;

and - The dead body was found by the policeman.

In the CD, the PCD of the verb “was found” requires that its agent is animate. The phrase “by the front door” in the first sentence contains no animate noun. Its case is therefore deduced to be location, with ‘by’ meaning ‘near to’.

In the second sentence, ‘by’ introduces the animate noun ‘policeman’, which can accordingly be regarded as the agent. The horizontal links will show this to be the most likely meaning of this sentence.

Consider now the sentence

- The dead body was found by the sleeping policeman.

In the CD, the semantic feature ‘awake’, normally associated with the noun ‘policeman’, will be negated in virtue of the qualifying adjective “sleeping”. This exemplifies, incidentally, how information found in this computational dictionary is modifiable in virtue of the current computational analysis. It follows from this analysis that the sleeping policeman cannot be the agent of “was found”.

If the phrase takes on the case of location, with ‘by’ meaning ‘near to’ the sentence takes on a meaningful interpretation. Horizontal links in the dictionary equate the (idiomatic) phrase “sleeping policeman” with road hump.

5 Machine Translation

The dictionary is designed as a general-purpose system and, as such, it can be used in a wide variety of NLP applications, e.g. in information retrieval, CALL or robot control, in addition to a range of different modes of access.

One important application of NLP is Machine Translation in which interpreting meaning and resolving ambiguity in the source text are important aspects. Machine Translation is the specific application considered in this paper.

In the most simplistic approach, one would expect the tree-structure of the target language dictionary to match the tree-structure of the source -language dictionary, and therefore be able to (re)label each node in the target-language dictionary with the corresponding word in the target language.

However, there is not always an exact correspondence in meaning between words in the two languages. Also, there is not always a one-to-one match between words in the two languages; such a match could be e.g. two-to-one or one-to-three. Sometimes, in one family, when sitting down to a meal, someone may ask:, ‘Is the dinner hot?’. To this, someone else may resort to Spanish in the response ‘Do you mean ‘hot caliente’ or ‘hot piquente’?’ , the choice of translation corresponding to two different meanings of the English word “hot”.

There are numerous other examples. Thus ‘verbal’ could mean ‘in words’ (written or spoken) or, more specifically, just ‘spoken’ (i.e. oral). The preposition ‘on’ could correspond in French to ‘dans’ or ‘sur’, depending on the context. As a last example here, ‘country’ could correspond with ‘nation’, i.e. ‘realm’, or with the natural, non-urban areas.

It is clear that the links between matching words, i.e. in the new dimension across the inter-dictionary space, are not necessarily all 'horizontal', so to speak.

Furthermore, in order to determine the correct or , rather, best translation in the target text, an analysis of both the source and the target PCDV, and of the possible chouces of entries in the target language, will be required.

With such analysis as previously described, using the relevant PCDVs etc. in the CD, the appropriate translations of the above-quoted English sentences can be obtained. For example, these English sentences and their corresponding translations in French would be :

1 The cat drank a saucer of milk.

Le chat a bu une soucoupe de lait.

2 The captain hit the sailor with a cat.

Le capitaine a frappé le marin avec un martinet à neuf cordes.

3 The crew sailed the cat out of the harbour.

L'equipage a sorti le capon du port.

4 The sailors took a cat out of the harbour.

Les marins ont sorti un capon du port.

<u>or</u>

Les marins ont repéché un chat de l'eau du port.

5 The dead body was found by the front door.

Le cadavre a été trouvé près de la porte d'entrée.

6 The dead body was found by the policeman.

Le cadavre a été trouvé par le policier.

7 The dead body was found by the sleeping policeman.

Le cadavre a été trouvé près du ralentisseur.

These examples indicate how the design of the computational dictionary can ensure close preservation of meaning in translation. Thus the use of such a dictionary is necessary if a high degree of accuracy of translation is to be achieved.

References

Al-Hafez, M Y. A semantic knowledge-based computational dictionary for support of natural language processing systems, Ph D thesis, Cranfield University, 1993.

Al-Hafez, M Y, Morayati, M, Vella, A and Clarke J D. Design of an Arabic language knowledge-base as a lexicon for NLP, Proc 3rd Int Conf on multi-lingual computing (Arabic and Roman script), December 1992.

Aoe, Jun-ichi. A method for building knowledge bases with morphological semantics, Dept of Information Science and Intelligent Systems, University of Tokushima, Japan

Boguraev, Bran and Briscoe Ted. Computational lexicography for natural language processing, Longmans, UK, 1989.

Clarke, J D. A multi-lingual computerised dictionary for machine translation, ICDBHSS/85, Grinnell College, Grinnell, Iowa, USA, June 1985

Evans, Martha et al. Lexical-Semantic Relations in Information Retrieval. In S Williams (ed), Humans and Machines: The Interface Through Language, Ablex.

Leech, G N. Semantics, Harmondsworth; Penguin, 1974

Lyons, J. Semantics, Cambridge University Press, 1977

Nida, E A. Computational Analysis of Meaning, Mouton, Hague, 1975

Omar, A M. Semantics science, 'Ealm Al-Delaleh', Kuwait University, Dar Al-Uroba lebnacher oa Altaouzie

Otani, Miyo and Lancel, Jean-Marie. Sentence generation: from semantic representations to sentences throughout linguistic definitions and lexicon-grammar Proc. 8th European Conference on Artificial Intelligence, Aug 1988, Pitman, London UK

Ullman, S. Meaning and Style, in The Meaning of Meaning, by Ogden C K and Richards I A, Oxford, 1973

Zernik, Uri Lexical acquisition: where is the semantics? Artificial Intelligence program, General Electric Research and Development Centre.