[From: Internat.conf. Methodology & Techniques of Machine Translation, British Computer Society, Cranfield, 13-15 February 1984]

A New Dictionary Structure for Bi-directional MT system

by

Shiu-chang LOH, Luan KONG and Hing-sum HUNG

Hung On-To Research Centre for Machine Translation
The Chinese University of Hong Kong

ABSTRACT

The importance and structure of MT-dictionary were discussed extensively by many researchers in machine, translation in the past. These structures were mainly concerned with MT-dictionaries for one-way translation systems. In the present paper, a new dictionary structure for bi-directional machine translation is being introduced. The new structure is being tested for Chinese-English as well as English-Chinese machine translation.

INTRODUCTION

The importance and structure of MT-dictionary were discussed extensively by many researchers in machine translation in the past (Knowles 1982, Lamb and Jacobsen 1966, Liu 1982, Loh 1975, Oettinger 1960, Wang 1982, Wang, T'sou and Chan 1971). These structures were mainly concerned with MT-dictionaries for one-way translation systems. Dictionary structures for bi-directional or multi-language machine translation systems were rarely discussed. The aim of this paper is to introduce a dictionary structure suitable for multi-language translation system. The dictionary structure was designed in conjunction with the Dual Language Translator (DLT) developed at the Chinese University of Hong Kong in 1978 (Loh, Hung and Kong 1978).

2. BASIC STRUCTURE

It is generally agreed that for translation from one language L1 into another language L2, the MT-dictionary D_{12} must contain the following informations:

$$D_{12} = \{ IC_{L1} , GI_{L1} , IC_{L2} , GI_{L2} \}$$

where

IC_{L1} is a set of internal codings of the source lexical items in L1,

GI_{L1} is a set of grammatical informations of these items,

IC_{L2} is a set of target equivalences (the target lexical items in L2) of these items,

GI_{L2} is a set of grammatical informations of these target equivalences.

Similarly, for translation from language L2 into language L1, the MT-dictionary D_{21} must also contain these types of informations.

For a one-way translation system, this kind of dictionary structure may seem to be quite suitable. However, for a bi-directional language translation system, this kind of dictionary structure requires almost identical information to be kept in two different storages which is redundant and undesirable. A new structure for the dictionary is thus required.

In the course of designing the Dual Language Translator (DLT) we had foreseen this problem and realized that only two types of information are stored in D_{12} or D_{21}, namely, the source language information and the target language information. Thus these two dictionaries D_{12} and D_{21} for the bi-directional translation system between the languages L1 and L2 may then be replaced by the dictionaries

$$D_1 = \{ IC_{L1} , GI_{L1} \} \text{ and } D_2 = \{ IC_{L2} , GI_{L2} \}$$

together with some relations between D_1 and D_2.

Bearing these points in mind, we proposed a structure for the MT—dictionary of a multi-language translation system. The basic organization of the proposed MT-dictionary is illustrated in Fig. 2.1. The two main components of the dictionary are the DICTIONARY ADMINSTRATOR and the n SUB-DICTIONARIES.

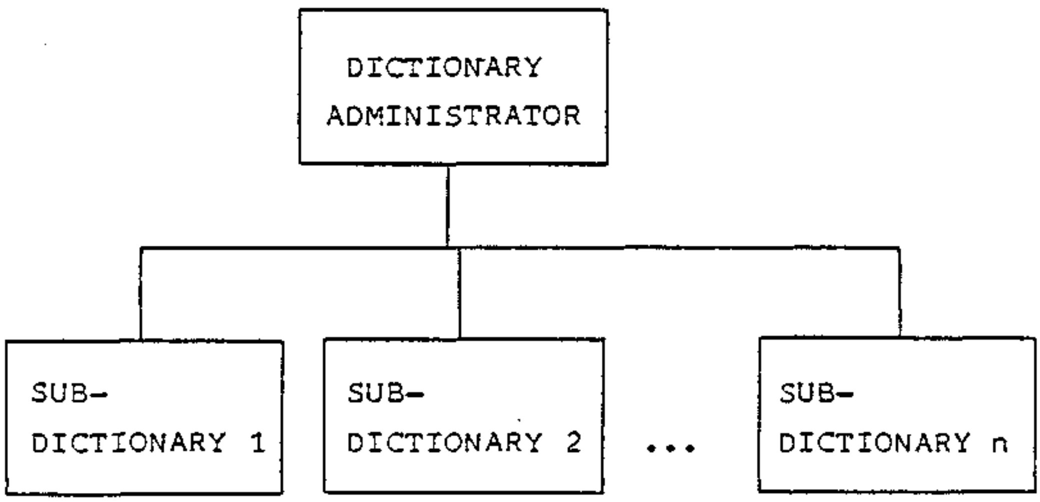


Fig. 2.1 The dictionary structure for a multi-language translation system

The DICTIONARY ADMINISTRATOR is a software responsible for the housekeeping of the n SUB-DICTIONARIES. This includes creation, deletion of a SUB-DICTIONARY and updating, insertion, deletion and listing of the contents of a SUB-DICTIONARY.

Each of the n SUB-DICTIONARIES contains the information on the lexical items of one of the n languages concerned. Each entry of items of a SUB-DICTIONARY specifies the codings of a lexical item, its grammatical information and (n-1) pointers which point to the entries of the other (n-1) SUB-DICTIONARIES where the target equivalences of the lexical item can be found respectively (Fig. 2.2).

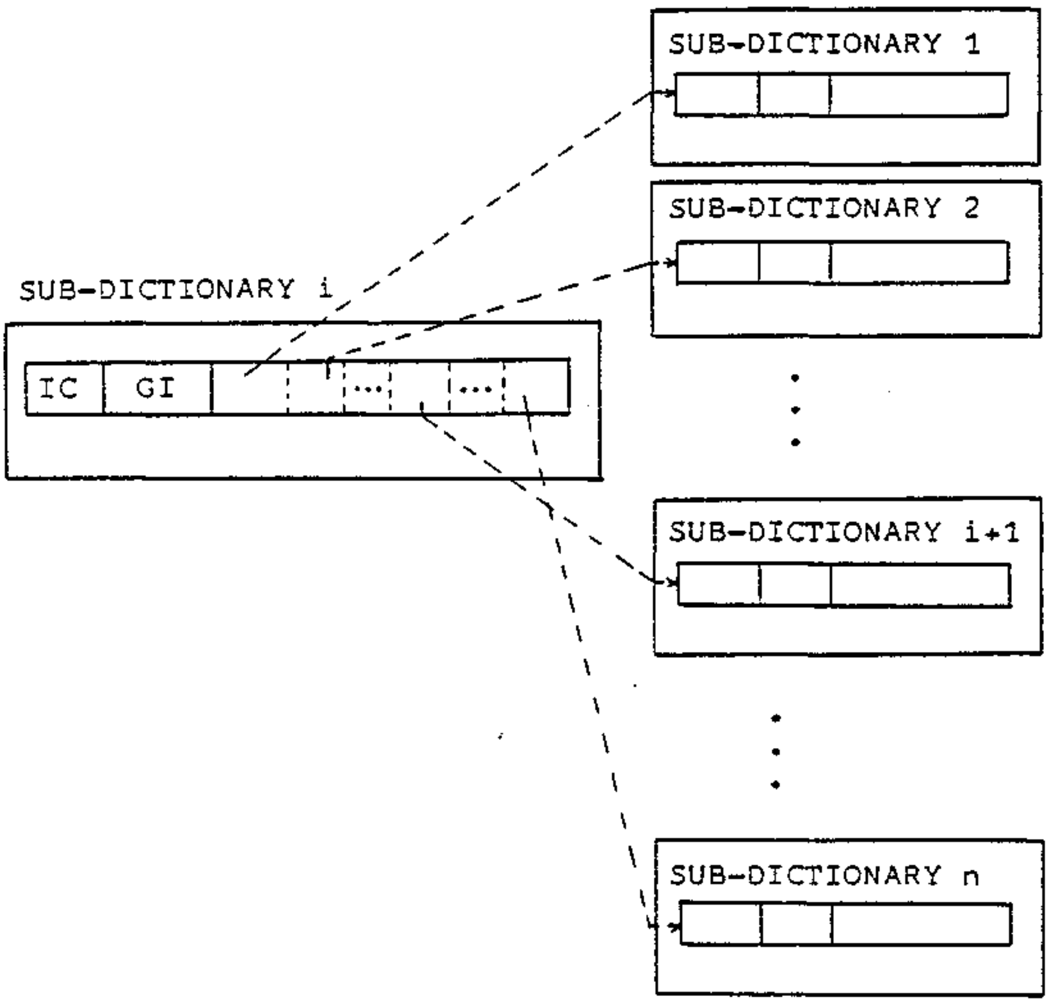


Fig. 2.2 The relationship of the SUB-DICTIONARIES

Using this approach, the number of MT-dictionaries required for a multi-language translation system of n languages can be reduced from n(n-l) to n. Duplication of information and redundancy are eliminated. This certainly minimizes the storage spaces required.

An implementation of this dictionary structure is the dictionary of the Dual Language Translator (DLT) for the translation between Chinese and English. This dictionary consists of a DICTIONARY ADMINISTRATOR, a Chinese SUB-DICTIONARY and an English SUB-DICTIONARY (Fig.2.3). The actual organization of a SUB-DICTIONARY will be discussed in the following section.

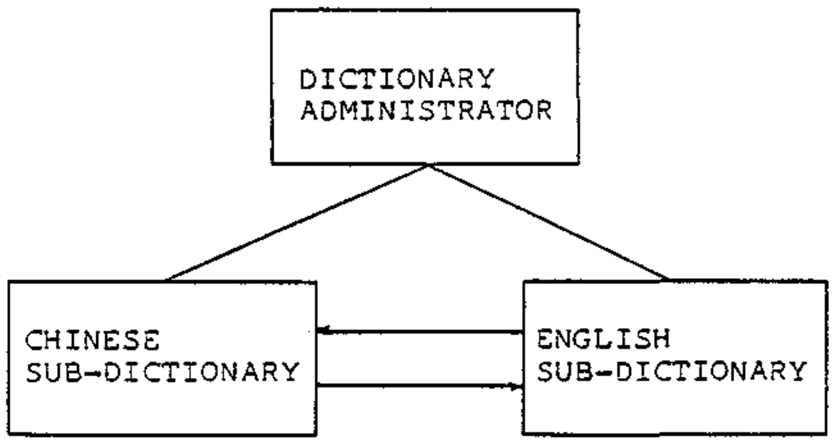


Fig. 2.3   Dictionary structure for the DLT

3. A SUB-DICTIONARY

Although the actual contents of the Chinese SUB-DICTIONARY and English SUB-DICTIONARY for the Dual Language Translator (DLT) are different, their organization are the same. Basically, there are three main types of information in a SUB-DICTIONARY, namely, CONTROL INFORMATION, SYNTACTIC/ SEMANTIC ITEMS and a COMMON DATA-POOL (Fig. 3.1).

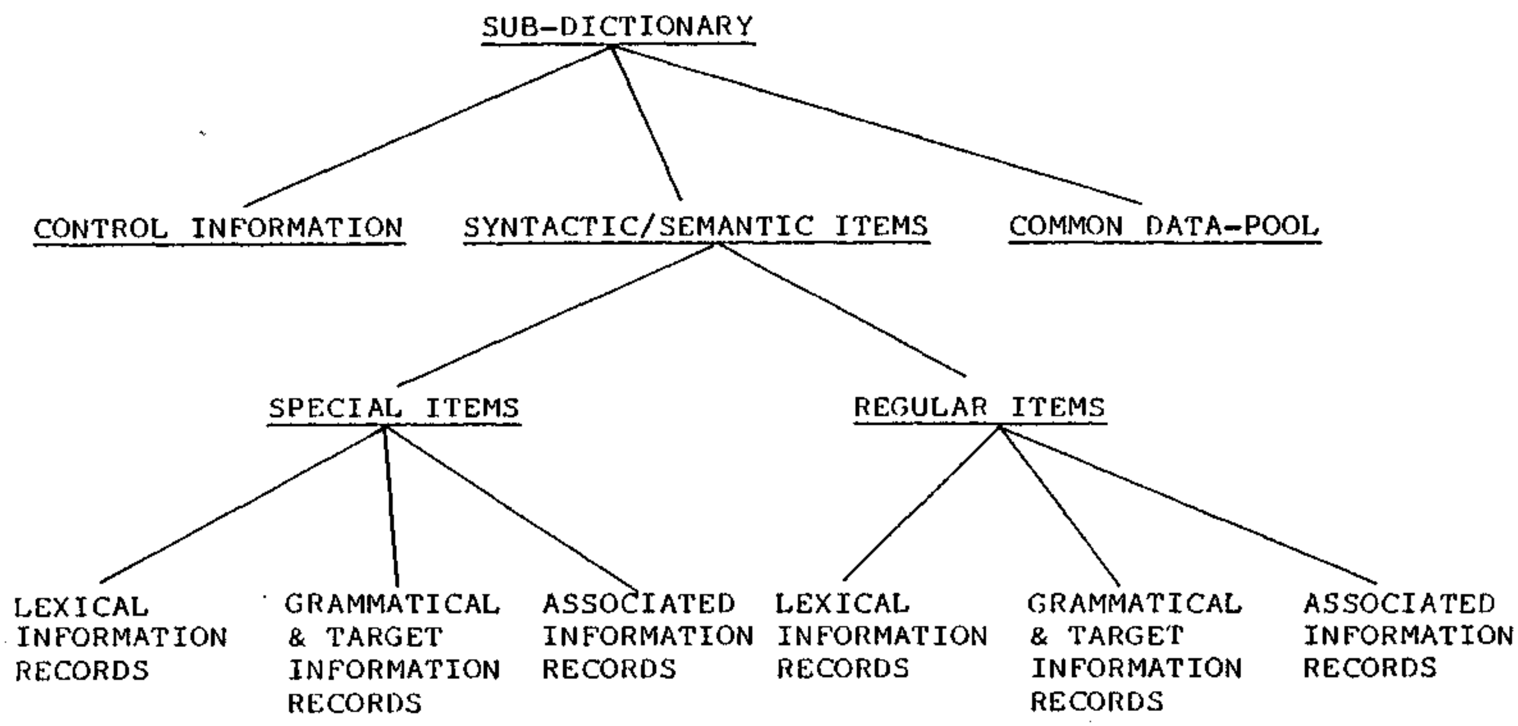


Fig. 3.1 Organization of a SUB-DICTIONARY

CONTROL INFORMATION

The CONTROL INFORMATION specifies the identification of the SUB-DICTIONARY and some control data relevant to the organization of the SYNTACTIC/SEMANTIC ITEMS (or simply items).

SYNTACTIC/SEMANTIC ITEMS

SYNTACTIC/SEMANTIC ITEMS may further be sub-divided into special items and regular items. Special items, in contrast to regular items, are those most frequently used items. The reason for separating these items from the others is to speed up the translation process. During the translation process, these items will be kept in computer main memory. Thus dictionary consultation for these items will not be necessary, consequently reduce the times for lexical analysis. Due to the limitation of the size of computer main memory, the number of these special items is limited.

The information attached to each item, either special or regular, may be grouped into three different types of records:

(a). Lexical information records,

(b). Grammatical and target information records, and

(c). Associated information records.

These three types of records are linked together internally by using pointers.

Lexical information records are used for lexical analysis. A traditional method for representing lexical information is by means of a linear list such as illustrated in Fig. 3.2. The disadvantages of this method are that duplication of lexical information exists and the search for a lexical item may have to be carried out linearly. We can rewrite the same list in Fig. 3.2 into a tree (Fig. 3.3), and a linked list representation of such a tree is given in Fig. 3.4. It is by linked list that the lexical information in the SUB-DICTIONARIES are represented. The format of the lexical information records is illustrated in Fig. 3.5.

1	正
2	正 比
2	正 則
4	正 則 曲 綫
4	正 則 曲 面
3	正 則 域
2	正 交
3	正 交 極
3	正 交 补
3	正 交 集
2	正 規
3	正 規 化
4	正 規 化 子
4	正 規 化 基
2	正 态

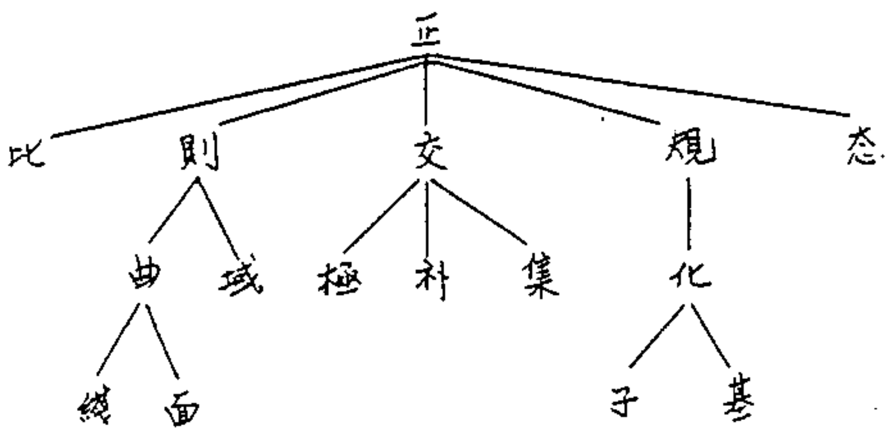
<u>Format:</u>

n #### #### ... ####

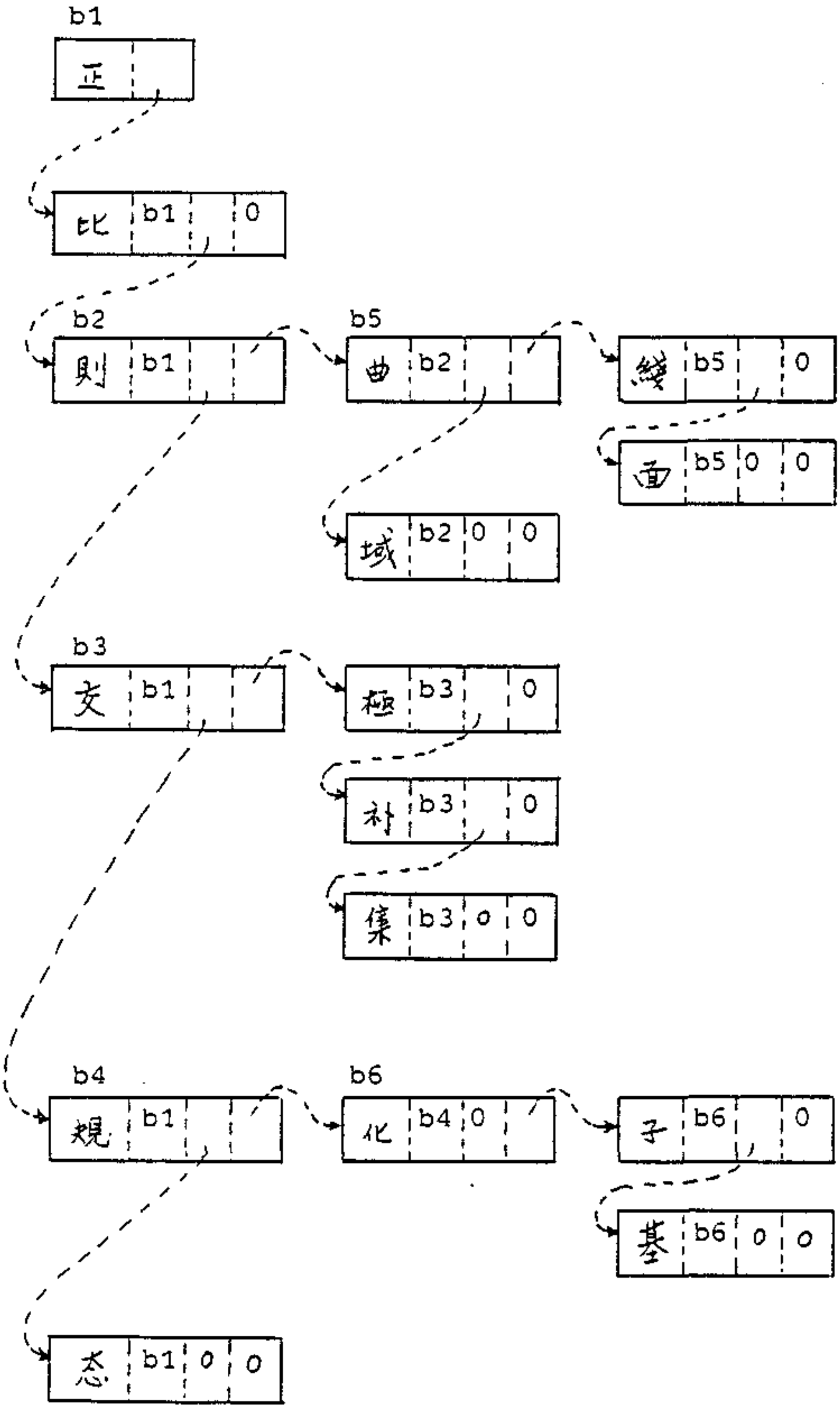
n=no. of codes

= Internal codings(or representation) of lexical items

Fig. 3.2 The traditional organization or
<u>lexical information</u>



<u>Fig. 3.3 A tree representation of lexical information</u>



bi = backward pointers

Fig. 3.4 A linked list representation of lexical information

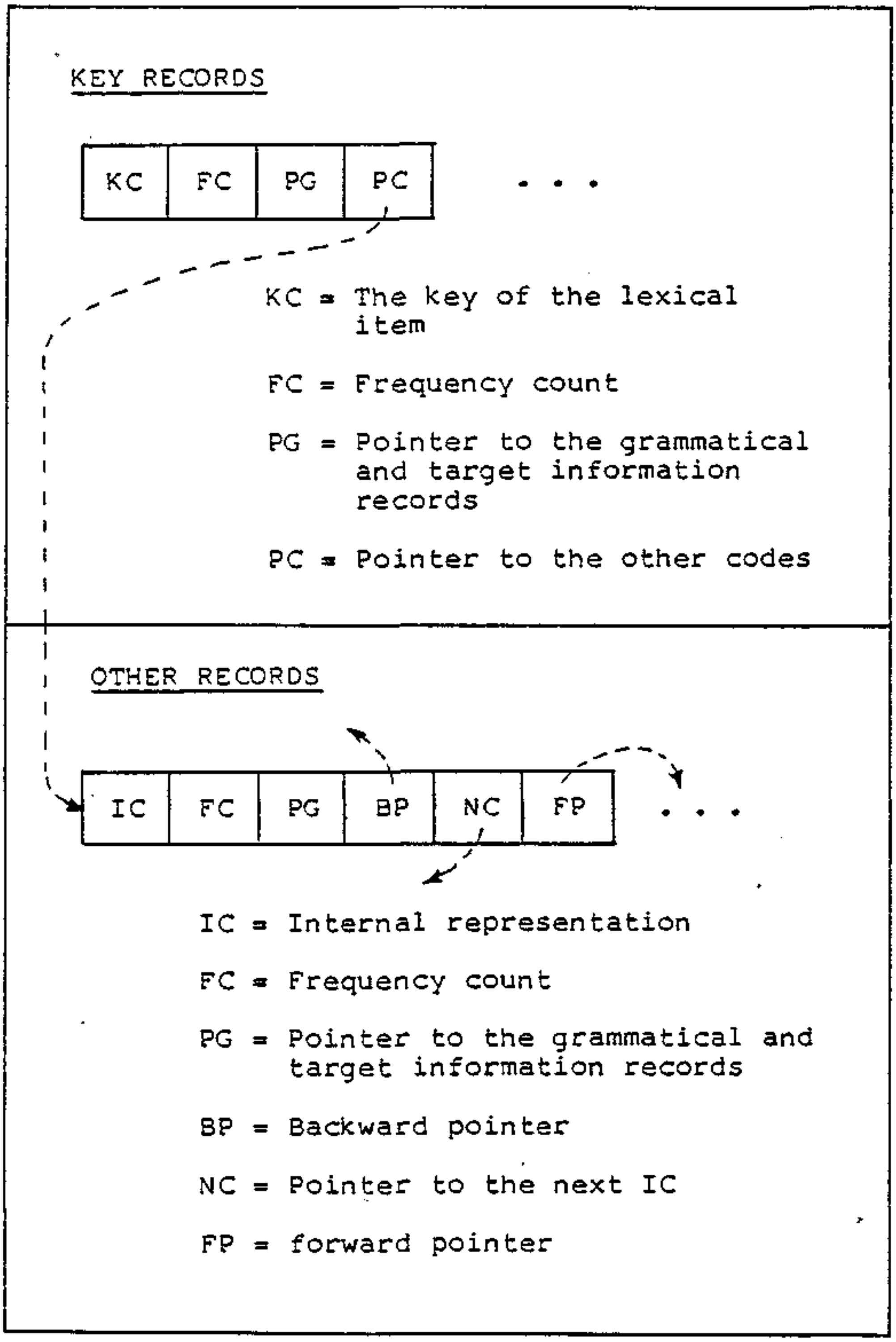


Fig. 3.5 Format of the lexical information records

The grammatical and target information records are used for syntactical/semantical analysis and target determination. The format of this type of records is illustrated in Fig. 3.6.

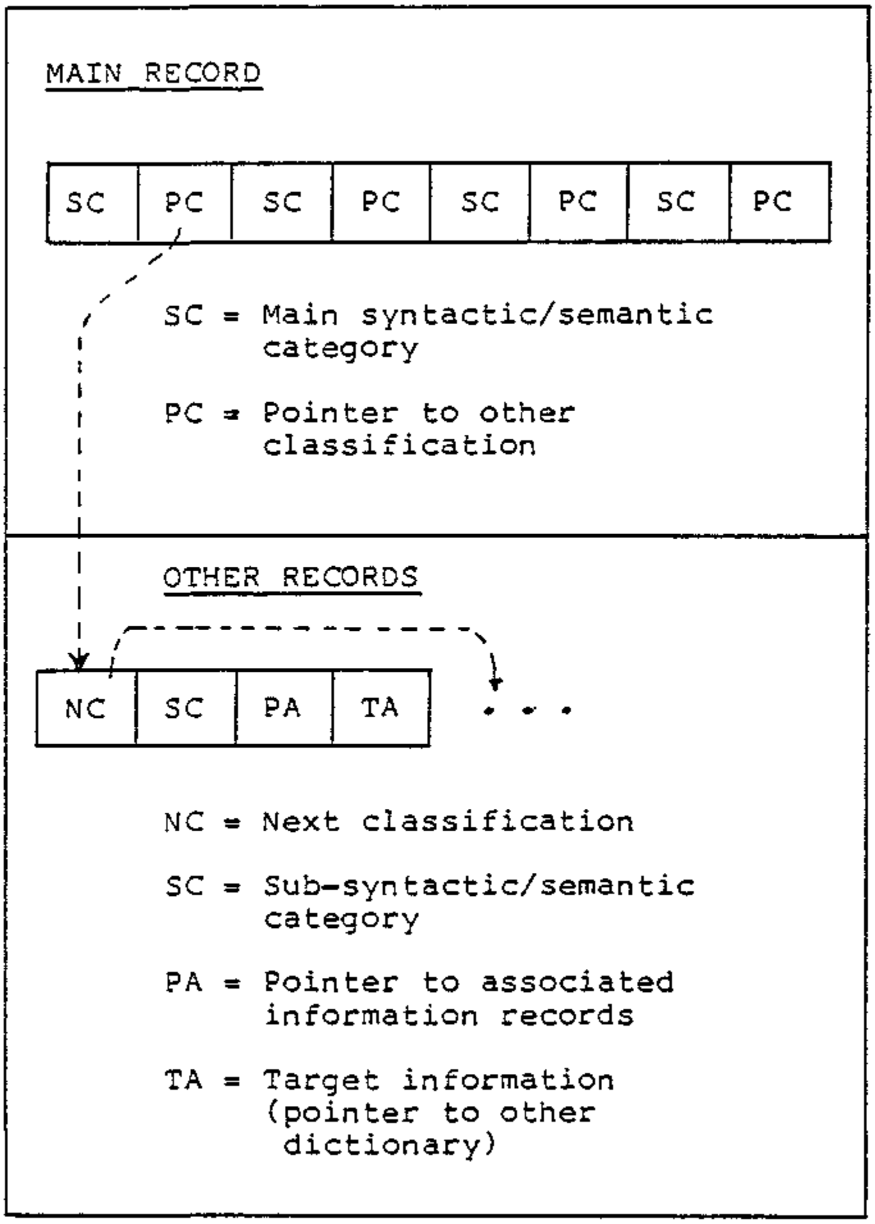


Fig. 3.6 Grammatical and target information records

The associated information records are used for determining the particular properties of the items, such as special article or measure word required, etc. and their format is illustrated in Fig. 3.7.

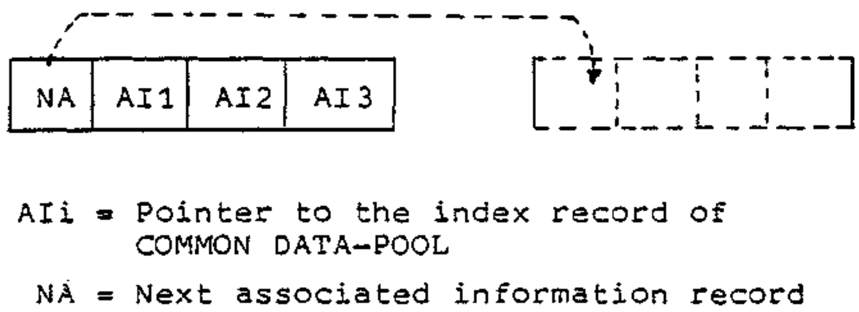


Fig. 3.7 Associated information records

COMMON DATA-POOL

The COMMON DATA-POOL is a set of data which will be used by all items or a subset of items. For example, articles, measure words, prefix and postfix etc. Fig. 3.8 illustrates the record format of COMMON DATA-POOL.

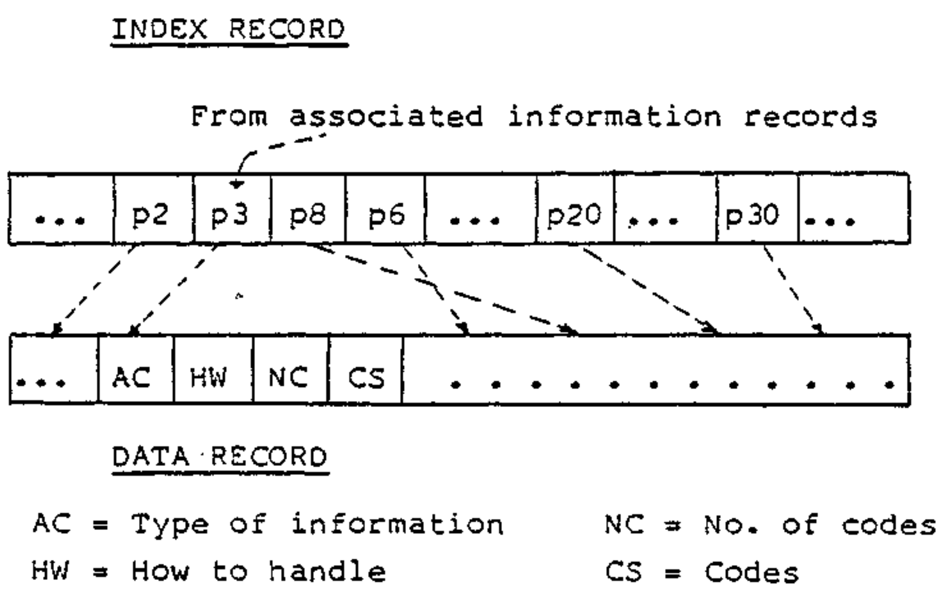


Fig. 3.8 COMMON DATA-POOL records

4. EXAMPLES

Example 1.

Consider the Chinese lexical item "研究(4282 4496)".

(1). This item can be noun or a verb.

(2). If it is a noun then it has one meaning, and can be assigned the semantic category NA (non-animate). For this particular meaning, the item has an associated information which specifies that it requires the particular measure word "項 (7309)".

(3). If it is a verb then it has one meaning, and can be assigned the semantic category HA (humanized action). For this particular meaning, the item does not have an associated information.

According to the above specification, the lexical information record, grammatical and target information record and the associated information record of the item will be as shown in Fig. 4.1(a).

The COMMON DATA-POOL of the Chinese SUB-DICTIONARY may be a set of Chinese characters which might be necessary to be inserted into the Chinese sentence. For example, the Chinese character "一(0001)", "個 (0020)", measure word "項 (7309)" etc.

Example 2.

Consider the English lexical item "STUDY".

(1). This item can be a noun or a verb.

(2). If it is a noun then it has one meaning, and can be assigned the semantic category NA (non-animate).

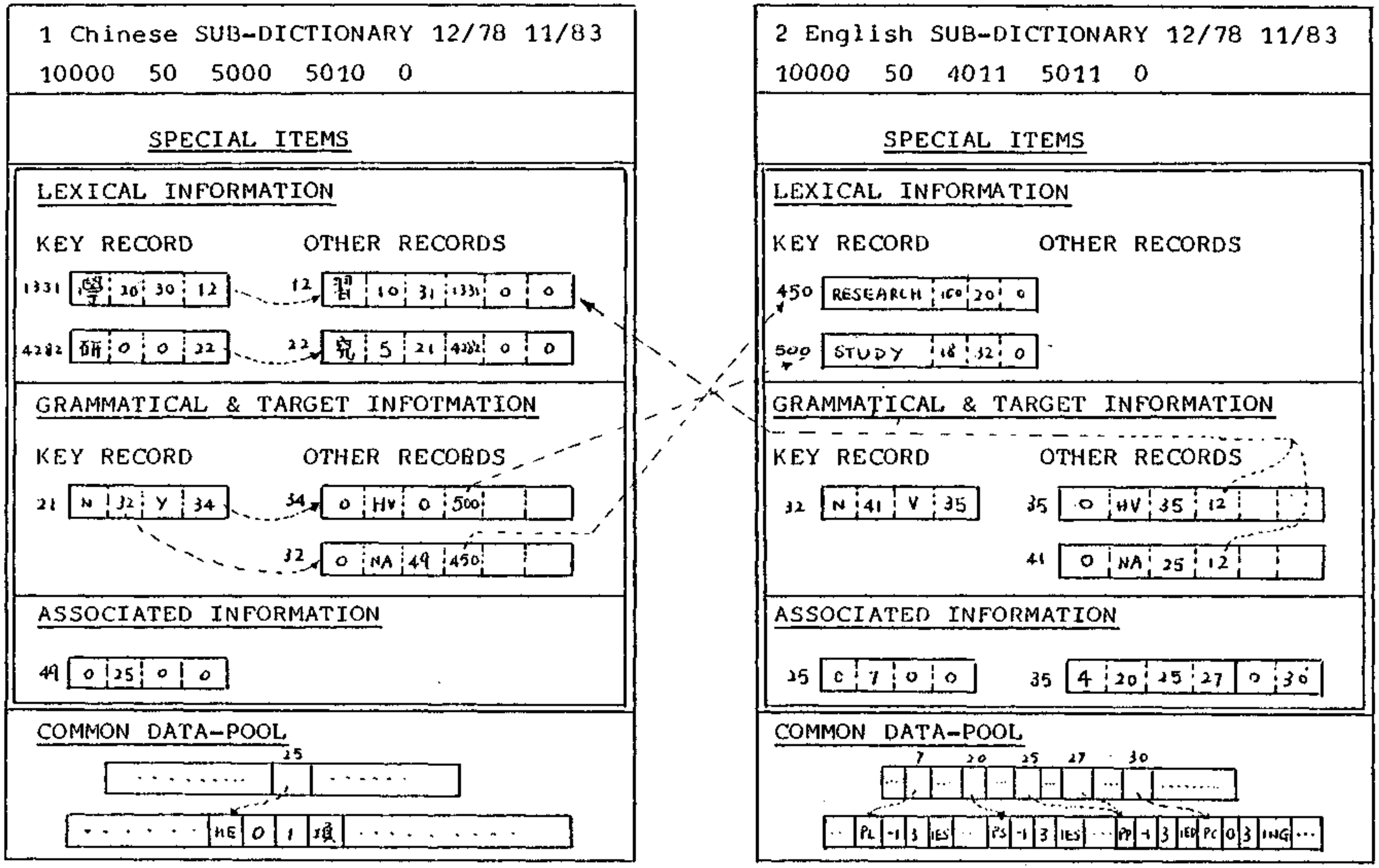
For this particular meaning, the item has an associated information which specifies that the plural form of the item is "STUD + IES".

(3). If it is a verb then it has one meaning, and can be assigned the semantic category HA (humanized action). For this particular meaning, the item has a, set of associated information which specify how the various form of the item can be constructed.

According to the above specification, the lexical information record, grammatical and target information record and the associated information record of the item will be as shown in Fig 4.1(b).

The COMMON DATA-POOL of the English SUB-DICTIONARY may be a set of characters which might be necessary to be inserted into the English words or sentence. For example, " S, ISS, D, ED, ING " etc.

Fig. 4.1(a) and Fig. 4.1(b) together show the relationships of the records specified in Example 1 and Example 2, and thus give a general idea how the two SUB-DICTIONARIES — the Chinese SUB-DICTIONARY and English SUB-DICTIONARY are linked together. The SUB-DICTIONARIES are linked together by the lexical information record and the grammatical and target information records. The relations between the items of the two SUB-DICTIONARIES are not necessary one to one, that is, an item in one SUB-DICTIONARY may have one and more than one equivalences in another SUB-DICTIONARY.



(a) (b)

Fig. 4.1 Example of records of the SUB-DICTIONARIES

REFERENCES

Knowles,F.E. (1982) The pivotal role of the various dictionaries in an MT system. Practical Experience of Machine Translation. North Holland Publishing Company.

Lamb, S.M. and Jacobsen, W.H. (1966) A high-speed large - capacity dictionary system. Readings in automatic language processing. American Elsevier

Liu, Z. (1982) Experiments on English-Chinese machine translation of titles. Language and computer. Academia Sinica, Peking.

Loh, S.C. (1975) Final report on machine translation. Machine translation project, CUHK.

Loh, S.C., Hung, H.S. and Kong, L.(1978) A dual language translator. Advances in computer-aided literary and linguistic research.

Oettinger, G. (1960) Automatic language translation. Harvard University Press.

Wang, G.Y. (1982) On the fixed phrases in machine translation. Language and computer. Academia Sinica, Peking.

Wang, S.Y., T'sou K. and Chan W. (1971) Research in Chinese—English machine translation. University of California.