Chapter 5

Dictionaries

5.1 Introduction

This Chapter is about the role played by dictionaries in MT. Our decision to devote a whole chapter to this discussion reflects the importance of dictionaries in MT:

- Dictionaries are the largest components of an MT system in terms of the amount of information they hold. If they are more then simple word lists (and they should be, if a system is to perform well), then they may well be the most expensive components to construct.
- More than any other component, the size and quality of the dictionary limits the scope and coverage of a system, and the quality of translation that can be expected.
- The dictionaries are where the end user can expect to be able to contribute most to a system — in fact, an end user can expect to have to make some additions to system dictionaries to make a system really useful. While MT suppliers rarely make it possible for users to modify other components, they normally expect them to make additions to the dictionary. Thus, from the point of view of a user, a basic understanding of dictionary construction and sensitivity to the issues involved in 'describing words' is an important asset.
- In discussing dictionaries here, we include also some discussion of terminology — it is with respect to the treatment of terminology that MT provides some of its most useful benefits.

We shall approach the question of dictionaries in MT obliquely, by considering in some detail the information contained in, and issues raised by, the paper dictionaries with which we are all familiar. There are a number of reasons for this, but the most important is that the dictionaries in existing MT systems are diverse in terms of formats, coverage, level

of detail and precise formalism for lexical description. This diversity should not be a surprise. Different theories of linguistic representation can give rise to different views of the dictionary, and different implementation strategies can make even fundamentally similar views of the dictionary look very different in detail. Moreover, the different kinds of MT engine obviously put quite different requirements on the contents of the dictionary. For example, dictionaries in an interlingual system need not contain any translation information *per se*, all that is necessary is to associate words with the appropriate (collections of) interlingual concepts. By contrast, transformer systems will typically give information about source language items, and their translations, including perhaps information that is really about the target language, and which is necessary to trigger certain transformations (e.g. to do with the placement of particles like *up* in *look it up*, and *look up the answer*). Since transfer systems typically use more abstract levels of representation, the associated dictionaries have to contain information about these levels. Moreover, in a transfer system, especially one which is intended to deal with several languages, it is common to separate monolingual dictionaries for source and target languages (which give information about the various levels of representation involved in analysis and synthesis), from bilingual dictionaries which are involved in transfer (which normally relate source and target lexical items, and which normally contain information only about the levels of representation that are involved in transfer).

We would like to abstract away from these divergences and points of detail in order to focus on the main issues. Accordingly, we will begin with a brief discussion of typical entries that one might find in a good monolingual 'paper' dictionary, and a good bilingual 'paper' dictionary.[1] We will then briefly discuss the sort of information about words that one typically finds in MT dictionaries, outlining some of the different ways such information can be represented. As we have said, a simple view is that a dictionary is a list of words. However, it is impractical, and perhaps impossible to provide an exhaustive list of words for most languages. This is because of the possibility of forming new words out of existing ones, by various morphological processes. In Section 5.4 we will look briefly at these, and provide some discussion of how they can be dealt with, and the problems they raise in an MT context. In Section 5.5 we will briefly describe the difference between terminology and general vocabulary.

5.2 Paper Dictionaries

The best place to start our discussion is by looking at typical entries that one might find in a monolingual English dictionary (cf. page 89), and a bilingual dictionary (cf. page 90). [2]

We will start by looking at the layout of the first half of the monolingual entry. The entry

[1] 'Paper' here is intended to convey 'intended for human readers', as opposed to 'electronic' meaning 'intended for use by computers'. Of course, it is possible for a paper dictionary to be stored on a computer like any other document, and our use of 'paper' here is not supposed to exclude this. If one were being precise, one should distinguish 'paper' dictionaries, 'machine readable' dictionaries (conventional dictionaries which are stored on, and can therefore be accessed automatically by computer), and 'machine *usable* dictionaries'.

[2] The form of the monolingual entry is based on that used in the *Oxford Advanced Learner's Dictionary* (OALD); the bilingual entry is similar to what one finds in *Collins-Robert English-French dictionary.*

A Monolingual Dictionary Entry

but.ton /'bʌtn/ *n* **1** knob or disc made of wood, metal, etc sewn onto a garment as a fastener or as an ornament: *a coat, jacket, shirt, trouser button* ∘ *lose a button* ∘ *sew on a new button* ∘ *do one's buttons up* ⇒ illus at JACKET. **2** small knob that is pressed to operate a doorbell, a switch on a machine, etc: *Which button do I press to turn the radio on?* **3**(idm) **bright as a button** ⇒ BRIGHT. **on the 'button** (*US infml*) precisely: *You've got it on the button!*
▷ **but.ton** *v* **1**(**a**)[Tn,Tn.p] ∼**sth**(**up**) fasten sth with buttons: *button (up) one's coat, jacket, shirt, etc.* (**b**)[I,Ip] ∼(**up**) be fastened with buttons: *This dress buttons at the back.* **2**(idm) **button (up) one's lip** (*US sl*) be silent. **3**(phr v) **button sth up** (*infml*) complete sth successfully: *The deal should be buttoned up by tomorrow.*
□ **,buttoned 'up** silent and reserved; shy: *I've never met anyone so buttoned up.*
,button-down 'collar collar with ends that are fastened to the shirt with buttons.
'buttonhole *n* **1** slit through which a button is passed to fasten clothing. ⇒ illus at JACKET. **2** flower worn in the buttonhole of the lapel of a coat or jacket. - *v*[Tn] make (sb) stop and listen, often reluctantly, to what one wants to say.
'buttonhook *n* hook for pulling a button into place through a buttonhole.
,button 'mushroom small unopened mushroom.

for *button* starts off with the word itself in bold print. This is called the *head word*. The dot in the word indicates where the word may be broken off (e.g. for hyphenation). After that there is a phonetic transcription of the word's pronunciation. Then the entry falls into two main parts, describing first the noun and then the verb *button*. Definitions identify two different meanings, or *readings* of the noun *button*, with examples of usage given in italics. The ⇒ refers the reader to a related entry. Idiomatic expressions are given under **3**. As for the verb, the code [Tn,Tn.p] indicates that the verb is transitive, i.e. appears in a sentence with a subject and an object (Tn), or is transitive with an adverbial particle (Tn.p). In this case the adverbial particle is the preposition *up*. Under **b** another usage is described where *button* is an intransitive verb and thus takes only a subject (I), or a subject plus the preposition *up* (Ip). Idioms appear under **2**. The box halfway through the entry signals the start of a list of complex forms, a phrasal verb (*button up*), and several compounds, which we will discuss later in this chapter. The verb, and noun, phrasal verbs and compounds are given in a standard form (the *citation* form), with information about stress (given by raised or lowered apostrophes). By convention, this is normally the singular form of nouns, and the infinitive form of verbs (i.e. the form that one finds after *to*, as in *to button, to be*, etc.)

The bilingual entry for the noun *printer* begins with the head word, its pronunciation and

Two Bilingual Dictionary Entries

button ['bʌtn] **1** *n* **(a)** (*garment, door, bell, lamp, fencing foil*) bouton *m*. **chocolate ∼s** pastilles *fpl* de chocolate. **2** *vt* (*also* ∼ **up**) *garment* boutonner. **3** *vi* (*garment*) se boutonner. **4** *cpd* **buttonhook** tirebouton *m*; **button mushroom** (petit) champignon *m* de couche *or* de Paris.

printer ['prIntə] *n* **(a)** imprimeur *m*; (*typographer*) typographe *mf*, imprimeur. **the text has gone to the** ∼ le texte est chez l'imprimeur; **∼'s devil** apprenti imprimeur; **∼'s error** faute *f* d'impression, coquille *f*; **∼'s ink** encre *f* d'imprimerie; **∼'s reader** correcteur *m*, -trice *f* (d'épreuves). **(b)** (*Comput*) imprimante *f*. **(c)** (*Phot*) tifeuse *f*.

word class, in this case noun. Logically, the entry then divides into three component parts **(a)**, **(b)**, and **(c)**, essentially distinguishing three different uses or meaning of the noun in English which have distinct translations into French. Where a particular meaning can be identified by reference to a subject field, this information is given (bracketed, in italics) — here computation and photography are identified as subject fields. If the context of use is other than these two fields, then the translation given under **(a)** is assumed to be appropriate. For each reading, the gender of the translation is given: *m* or *f* (for *masculine* or *feminine*, *mf* indicates either is possible; where the masculine and feminine forms differ, both are indicated — *printer's reader* is thus either *correcteur* or *correctrice*). If two different translations are possible they are both given, separated by a comma (thus, either *typographe*, or *imprimeur* are possible 'general' translations). The entry also contains some examples of idioms, or other usages, again with the appropriate translations.

Normal, 'paper' dictionaries, are collections of entries such as these. That is, they are basically lists of words, with information about the various properties. While grammar rules define all the possible linguistic structures in a language, the descriptions of individual words that are found in the dictionary or dictionaries state which words can appear in which of the different structures. A common (though not completely correct) view is that dictionaries contain all the 'idiosyncratic', 'irregular', or unpredictable information about words, while grammars provide general rules about classes of word, and phrases (this is only true if one excludes morphological rules and idioms from the dictionary — the former can be viewed as dealing with classes of word, and the latter are phrases).

One can get an idea of the sheer volume of information of this kind that may be needed by considering that for commercial purposes a lexicon with 20 000 entries is often considered as the minimum. This however is still only a modest percentage of existing words — the *Oxford English Dictionary* contains about 250 000 entries without being exhaustive even of general usage.[3] In fact, no dictionary can ever be really complete. Not only do dic-

[3]One can also get some idea of the *cost* of dictionary construction from this. Even if one were able to write four entries an hour, and keep this up for 8 hours a day every working day, it would still take over three

tionaries generally restrict themselves to either general, or specialist technical vocabulary (but not both), in addition, new words are constantly being coined, borrowed, used in new senses, and formed by normal morphological processes.[4]

5.3 Types of Word Information

We have already observed that dictionaries are a, perhaps *the*, central component of MT systems. In earlier Chapters, we have presented a highly simplified view of dictionaries — for example, in Chapter 3 the dictionary was sometimes little more than a list of rules such as `v` $\rightarrow$ `walk`, which only allows information about part of speech to be represented, and in Chapter 4 we gave translation rules which simply paired up the citation forms of source and target words (e.g. `temperature` $\leftrightarrow$ `temperatur`). However, though some of the information that is found in a typical paper dictionary is of limited value in MT (e.g. information about pronunciation is only useful in speech to speech systems), in general the quality and detail of the information one needs for MT is at least equal to that which one finds in paper dictionaries. In this section we discuss the various pieces of information about words that a good MT system must contain, basing ourselves on the dictionary entries above. An issue we will not address in this Chapter is the treatment of idioms, which one typically finds in paper dictionary entries. We discuss the treatment of idioms in Chapter 6.

It is useful to make a distinction between the characteristics of a word itself (its inherent properties) and the restrictions it places on other words in its grammatical environment. Although this distinction is not explicitly drawn in paper dictionaries, information of both types is available in them. Information about grammatical properties includes the indication of gender in the French part of the bilingual dictionary entry, and the indication of number on nouns (typically, the citation form of nouns is the singular form, and information about number is only explicitly given for nouns which have only plural forms, such as *scissors*, and *trousers*).

Information about the grammatical environment a word can appear in is normally thought of as dividing into two kinds: **subcategorization** information, which indicates the syntactic environments that a word can occur in, and **selectional restrictions** which describe semantic properties of the environment. Typical information about subcategorization is the information that *button* is a transitive verb. This is expressed in the verb code [Tn] in the dictionary entry on page 89. More precisely, this indicates that it is a verb that appears as the HEAD of sentences with a (noun phrase) SUBJECT and a (noun phrase) OBJECT. The following gives some examples, together with the appropriate verb codes from OALD:

years to construct even a small size dictionary. Of course, the time it takes to write a dictionary entry is very variable, depending on how much of the work has already been done by other lexicographers.

[4] In fact, it is arguable that the vocabulary of a language like English, with relatively productive morphological processes, is infinite, in the sense that there is no longest word of the language. Even the supposedly longest word *antidisestablishmentarianism* can be made longer by adding a prefix such as *crypto-*, or a suffix such as *-ist*. The result may not be pretty, but it is arguably a possible word of English. The point is even clearer when one considers compound words (see Section 5.4.3.

(1) a. The president died. [I]
b. The Romans destroyed the city. [Tn]
c. Sam gave roses to Kim. [Dn.pr]
d. Sam gave Kim roses. [Dn.n]
e. Sam persuaded Kim to stay at home. [Cn.t]
f. Kim believed that the library was closed. [Tf]
g. The quality is low. [La]
h. Sam appeared the best man for the job. [Ln]

Note that [I] refers to intransitive verbs that only need a subject to form a grammatical sentence, [Tn] to transitive verbs (like *button*) that need a subject and an object, [Dn.pr] to ditransitive verbs which take a subject and two objects, where the second one is introduced by the preposition *to*, [Dn.n] to ditransitive verbs that take a subject plus two object nouns, [Cn.t] to complex transitive verbs which require a subject, object and an infinitival (non-tensed) clause introduced by *to*, [Tf] to transitive verbs taking a subject, object and a finite (tensed) sentence introduced by *that*, [La] to linking verbs which link an adjectival phrase (which describes in some way the subject), to the subject, and [Ln] refers to linking verbs which link a noun phrase to the subject.

Verbs are not the only word categories that subcategorize for certain elements in their environment. Nouns exhibit the same phenomenon, like those nouns that have been derived from verbs (deverbal nouns).

(2) a. *The death of the president* shocked everybody.
b. *The destruction of the city by the Romans* was thorough.

Similarly, there are some adjectives that subcategorize for certain complements. Note that in the examples below we find three different types of complements, and that 3b and 3c differ from each other because in 3b the subject of the main clause is also the understood subject of the subclause, whereas in 3c the subject of the main clause is the understood object of the subclause.

(3) a. Mary was *proud of her performance*.
b. He was *eager to unwrap his present*.
c. That matter is *easy to deal with*.

An adequate dictionary of English would probably have to recognize at least twenty different subcategorization classes of verb, and a similar number for adjectives and nouns.

The reason one cannot be precise about the number of different subcategorization classes is that it depends (a) on how fine the distinctions are that one wants to draw, and (b) on how far one relies on rules or general principles to capture regularities. For example, probably all verbs allow coordinated subjects such as *Sam and Leslie*, but there are some, like *meet*, where this is equivalent to an ordinary transitive SUBJECT-VERB-OBJECT construction (cf. (4a), and (4b) mean the same, but (4c) and (4d) do not). One could decide to recognise this distinction by creating a separate subcategorization class, thus extending the number

of classes. But one could also argue that this fact about *meet* and similar verbs is probably related to their semantics (they describe symmetric relations, in the sense that if A meets B, then B meets A), and is thus regular and predictable. The appropriate approach could then be to treat it by means of a general linguistic rule (perhaps one that transforms structures like (4a) into ones of the form (4b)) Of course, unless one can rely on semantic information to pick out verbs like *meet*, one will have to introduce some mark on such verbs to ensure that they, and only they, undergo this rule. However, this is not necessarily the same as introducing a subcategorization class.

(4) a. Sam met Mary
b. Sam and Mary met
c. Sam saw Mary
d. ∗Sam and Mary saw

Subcategorization information indicates that, for example, the verb *button* occurs with a noun phrase OBJECT. In fact, we know much more about the verb than this — the OBJECT, or in terms of semantic roles, the PATIENT, of the verb has to be a 'buttonable' thing, such as a piece of clothing, and that the SUBJECT (more precisely AGENT) of the verb is normally animate.[5] Such information is commonly referred to as the **selectional restrictions** that words place on items that appear in constructions where they are the HEAD. This information is implicit in the paper dictionary entry above — the information that the object of *button* is inanimate, and normally an item of clothing has to be worked out from the use of *sth* (= 'some*thing*') in the definition, and the example, which gives *coat*, *jacket*, *shirt* as possibilities. The entry nowhere says the SUBJECT of the verb has to be an animate entity (probably human), since no other entity can perform the action of 'buttoning'. It is assumed (rightly) that the human reader can work this sort of thing out for herself. This information has to be made explicit if it is to be used in analysis, transfer or synthesis, of course.

Basic inherent information and information about subcategorization and selectional restrictions can be represented straightforwardly for MT purposes. Essentially, entries in an MT dictionary will be equivalent to collections of attributes and values (i.e. features). For example, one might have something like the following for the noun *button*, indicating that its base, or citation form is `button`, that it is a common noun, which is `concrete` (rather than `abstract`, like *happiness*, or *sincerity*)

```
lex = button
cat = n
ntype = common
number =
human = no
concrete = yes
```

[5] The restriction applying on the OBJECT of the verb actually concerns the thing which is *buttoned* whether that appears as the OBJECT of a active sentence or the SUBJECT of a passive sentence.

An obvious way to implement such things is as records in a database, with attributes naming fields (e.g. `cat`), and values as the contents of the fields (e.g. `n`). But it is not always necessary to name the field — one could, for example, adopt a convention that the first field in a record always contains the citation form (in this case the value of the feature `lex`), that the second field indicates the category, and that the third field some sort of subdivision of the category.

Looking at the dictionary entry for the noun *button* it becomes clear that different parts of speech will have a different collection of attributes. For example, verbs will have a `vtype`, rather than an `ntype` feature, and while verbs might have fields for indications of number, person and tense, one would not expect to find such fields for prepositions. In the entry we have given we also find one attribute — `number` — without a value. The idea here is to indicate that a value for this attribute is possible, but is not inherent to the word *button*, which may have different number values on different occasions (unlike e.g. *trousers*, which is always plural). Of course, this sort of blank field is essential if fields are indicated by position, rather than name. In systems which name attribute fields it might simply be equivalent to omitting the attribute, but maintaining the field is still useful because it helps someone who has to modify the dictionary to understand the information in the dictionary. An alternative to giving a blank value, is to follow the practice of some paper dictionaries and fill in the default, or (in some sense) normal value. For an attribute like `number`, this would presumably be singular. This alternative, however, is unfashionable these days, since it goes against the generally accepted idea that in the best case linguistic processing only *adds*, and never changes information. The attraction of such an approach is that it makes the order in which things are done less critical (cf. our remarks about the desirability of separating declarative and procedural information in Chapter 4).

In order to include information about subcategorization and selectional restrictions, one has two options. The first is to encode it via sets of attributes with atomic values such as those above. In practice, this would mean that one might have features such as `subcat=subj_obj`, and `sem_patient=clothing`. As regards subcategorization information, this is essentially the approach used in the monolingual paper dictionary above. The resulting dictionary entry could then look something like the following:

```
lex = button
cat = v
vtype = main
finite =
person =
number =
subcat = subj_obj
sem_agent = human
sem_patient = clothing
```

In some systems this may be the only option. However, some systems may allow values

to be sets, or lists, in which case one has more flexibility. For example, one might represent subcategorization information by means of a list of categories, for example `subcat = [np,np,np]` might indicate a verb that allows three NPs (such as *give*), and `[np,np,pp]` might indicate a verb that takes two NPs and a PP (again like *give*).

(5) a. Sam gave roses to Kim. (`subcat = [np,np,pp]`)
b. Sam gave Kim roses. (`subcat = [np,np,np]`)

A further refinement would be to indicate the actual grammatical relations involved, perhaps as in `subcat = [SUBJ:np, OBJ:np, IOBJ:pp]`. A notation which allows the lexicographer to indicate other properties of the items would be still more expressive. For example, it would be useful to indicate that with *give*, the preposition in the PP has to be *to*. This would mean that instead of '`pp`' and '`np`' one would have collections of features, and perhaps even pieces of syntactic structure. (A current trend in computational linguistics involves the development of formalisms that allow such very detailed lexical entries, and we will say a little more about them in Chapter 10).

Turning now to the treatment of translation information in MT dictionaries, one possibility is to attempt to represent all the relevant information by means of attributes and values. Thus, as an addition to the dictionary entry for *button* given above, a transformer system could specify a 'translation' feature which has as its value the appropriate target language word; e.g. `trans = bouton` for translation into French. One might also include features which trigger certain transformations (for example for changing word order for certain words). However, this is not a particularly attractive view. For one thing, it is clearly oriented in one direction, and it will be difficult to produce entries relating to the other direction of translation from such entries. More generally, one wants a bilingual dictionary to allow the replacement of certain source language oriented information with corresponding target language information — i.e. replace the information one derives from the source dictionary by information derived from the target dictionary. This suggests the usage of translation rules which relate head words to head words. That is, rules of the type we introduced in Chapter 4, like `temperature` $\leftrightarrow$ `temperatur`.

As we noted before, not all translation rules can be a simple mapping of source language words onto their target language equivalents. One will have to put conditions on the rules. For example, one might like to be able to describe in the bilingual entry that deals with *like* and *plaire*, the change in grammatical relations that occurs if one is working with relatively shallow levels of representation,. In effect, the transfer rule that we gave for this example in Chapter 4 might be seen as a bilingual lexical entry. Other translation rules that may require more than just a simple pairing of source and target words are those that treat phenomena like idioms and compounds, and some cases of lexical holes (cf. Chapter 6). To deal with such phenomena bilingual dictionary entries may have a single lexical item on the side of one language, whereas the other side describes a (possibly quite complex) linguistic structure.

The entry for *button* taken from a paper dictionary at the beginning of this Chapter illustrates an issue of major importance to the automatic processing of some languages, in-

cluding English. This is the very widespread occurrence of **homography** in the language. Loosely speaking, homographs are words that are written in the same way. However, it is important to distinguish several different cases (sometimes the term homography is restricted to only one of them).

1 The case where what is intuitively a single noun (for example) has several different readings. This can be seen with the entry for *button* on page 89, where a reading relating to clothing is distinguished from a 'knob' reading.

2 The case where one has related items of different categories which are written alike. For example, *button* can be either a noun or a verb.

3 The case where one has what appears to be unrelated items which happen to be written alike. The classic example of this is the noun *bank*, which can designate either the side of a river, or a financial institution.

These distinctions have practical significance when one is writing (creating, extending, or modifying) a dictionary, since they relate to the question of when one should create a new entry (by defining a new headword). The issues involved are rather different when one is creating a 'paper' dictionary (where issues of readability are paramount) or a dictionary for MT, but it is in any case very much a pragmatic decision. One good guiding principle one might adopt is to group entries hierarchically in terms of amounts of shared information. For example, there is relatively little that the two senses of *bank* share apart from their citation form and the fact that they are both common nouns, so one may as well associate them with different entries. In a computational setting where one has to give unique names to different entries, this will involve creating headwords such as `bank_1` and `bank_2`, or (`bank_finance`, and `bank_river`). As regards the noun and verb *button*, though one might want to have some way of indicating that they are related, they do not share much information, and can therefore be treated as separate entries. For multiple readings of a word, for example, the two readings of the noun *button*, on the other hand, most information is shared — they differ mainly in their semantics. In this case, it might be useful to impose an organization in the lexicon in which information can be inherited from an entry into sub-entries (or more generally, from one entry to another), or to see them as subentries of an abstract 'protoentry' of some sort. This will certainly save time and effort in dictionary construction — though the savings one makes may look small in one case, it becomes significant when multiplied by the number items that have different readings (this is certainly in the thousands, perhaps the hundreds of thousands, since most words listed in normal dictionaries have at least two readings). The issues this raises are complex and we cannot do them justice here, however, the following will give a flavour of what is involved.

More generally, what one is talking about here is **inheritance** of properties between entries (or from entries into subentries). This is illustrated in Figure 5.1. One could imagine extending this, introducing abstract entries expressing information true of *classes* of (real) entry. For example, one might want to specify certain facts about all nouns (all noun readings) just once, rather than stating them separately in each entry. The entry for a

typical noun might then be very simple, saying no more than 'this is a typical noun', and giving the citation form (and semantics, and translation, if appropriate). One allows for subregularities (that is lexical elements which are regular in some but not all properties), by allowing elements to inherit some information while expressing the special or irregular information directly in the entry itself. In many cases, the optimal organization can turn out to be quite complicated, with entries inheriting from a number of different sources. Such an approach becomes even more attractive if default inheritance is possible. That is, that information is inherited, unless it is explicitly contradicted in an entry/reading — it would then be possible to say, for example, 'this is a typical noun, except for the way it forms its plural'.

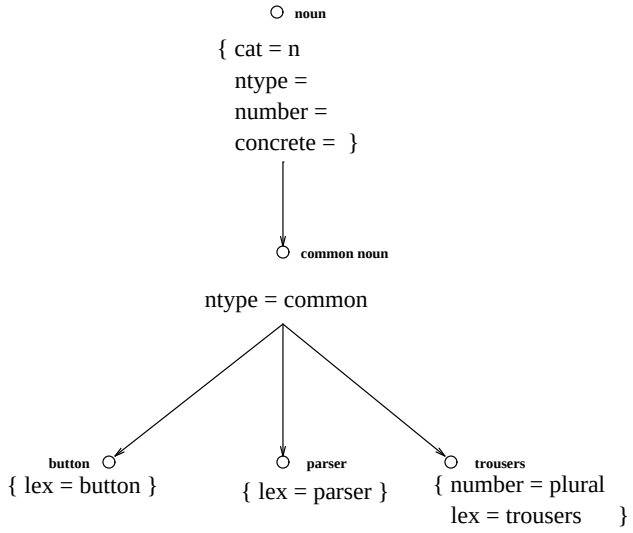

Figure 5.1 Inheritance

One final and important component of an MT dictionary, which is entirely lacking in paper dictionaries (at least in their printed, public form) is *documentation*. Apart from general documentation describing design decisions, and terminology, and providing lists and definitions (including operational tests) for the attributes and values that are used in the dictionary (it is, obviously, essential that such terms are used consistently — and consistency is a problem since creating and maintaining a dictionary is not a task that can be performed by a single individual), it is important that each entry include some lexicographers' comments — information about who created the entry, when it was last revised, the kinds of example it is based on, what problems there are with it, and the sorts of improvement that are required. Such information is vital if a dictionary is to be maintained and extended. In general, though the quality and quantity of such documentation has no effect on the actual performance of the dictionary, it is critical if a dictionary is to be modified or extended.

5.4 Dictionaries and Morphology

Morphology is concerned with the internal structure of words, and how words can be formed. It is usual to recognize three different word formation processes.

1 **Inflectional** processes, by means of which a word is derived from another word form, acquiring certain grammatical features but maintaining the same part of speech or category (e.g. *walk, walks*);

2 **Derivational** processes in which a word of a different category is derived from another word or word stem by the application of some process (e.g. *grammar* $\mapsto$ *grammatical, grammatical* $\mapsto$ *grammaticality*);

3 **Compounding**, in which independent words come together in some way to form a new unit (*buttonhole*).

In English, inflectional and derivational processes involve **prefixes** (as in *undo*) and **suffixes** (as in *stupidity*), and what is called **conversion**, or **zero-affixation** where there is a change of category, but no change of form (and example would be the process that relates the noun *button* to the verb). In other languages, a range of devices such as changes in the vowel patterns of words, doubling or reduplication of syllables, etc., are also found. Clearly, these prefixes and suffixes (collectively known as **affixes**) cannot 'stand alone' as words. Compounding is quite different in that the parts can each occur as individual words. Compounding is a very productive phenomenon in the Germanic languages, and poses some particular problems in MT, which we will discuss later.

5.4.1 Inflection

As a rule, paper dictionaries abstract away from inflection. Head words are generally uninflected, that is, nouns appear in singular form and verbs have the base (or infinitival) form. There are a number of reasons for this. The first is that inflection is a relatively regular process, and once the exceptional cases have been separated out, inflectional processes apply to all members of a given category. For example, to form the third person singular of the present tense of verbs one simply suffixes *s* (or its variant *es*) to the citation form of the verb. There are very few exceptions to this rule. Since it is a regular process, the dictionary user can be relied upon to form regularly inflected words from the citation forms given in the dictionary at will. Of course, irregularities, such as irregular plurals (*sheep, oxen, phenomena*, etc.) and plural only nouns (*trousers*) must be stated explicitly. A second important reason is eminently practical — it saves space, time and effort in constructing entries. Since English inflectional morphology is rather impoverished, these savings are not enormous. But Spanish, for example, has six different verb forms for the present tense, and if we add those for the past tense (either imperfecto or pretérito in Spanish) it amounts to 16 different verb forms. Other languages make even more use of inflections, like, for example, Finnish where there are said to be in the region of 2000 forms for most nouns, and 12 000 forms for each verb. It will be obvious that the need to describe inflectional variation by means of rules is pressing in such cases.

Within the context of MT, it is clearly desirable to have a similar approach, where monolingual and transfer dictionaries only contain the head words and no inflected words. In order to achieve this a system must be capable of capturing the regular patterns of inflection. This can be done by adding a morphological component to the system, which describes

all the regular inflections in general rules, with additional explicit rules for irregular inflection, thus allowing dictionary writers to abstract away from inflected forms as much as possible. The morphological component will be able to map inflected words onto the appropriate head words and will retain the information provided by the inflectional affix by adding the relevant features.

Let us consider again the verb *affects* in the simple sentence *Temperature affects density*. First, we want our morphological component to recognize *affects* as an inflected form of *affect*. Secondly, we want to retain the information carried by the affix so we can use it later when generating the output sentence. In the case of *affects* this means we want to state that the verb is finite, or tensed (in fact, present tense). This is important since it allows the verb to occur as the only verb of a main clause. The tense also prevents the verb from occurring behind auxiliary verbs like *will*. Other information that we gather from the inflection is the fact that the verb is third person (as opposed to first person, occurring with *I* or *we*, and as opposed with second person, occurring with *you*), and that it is singular (rather than third person plural, which occurs with *they*, or with a plural noun).

There are various ways of describing this, but perhaps the simplest is to use rules of the following form:[6]

```
(lex=V,cat=v,+finite,person=3rd,number=sing,tense=pres)
↔ V+ s
```

Here we have introduced a rule which says that finite verbs which are third person singular and have present tense (`cat=v, +finite, person=3rd, number=sing, tense=pres`) can be formed by adding `s` to the base form (the base form is represented as the value of the attribute `lex`). The rule can also be read in the opposite direction: if a word can be divided into a string of characters and *s*, then it may be a finite verb with third person singular in present tense. Other rules would have to be given to indicate that the `+s` ending can be added to all verbs, except for those that end in *+s*, themselves,[7] in which case `es` is added (cf. *kiss*, *kisses*).

Whether something is indeed the base form of the verb can be verified in the monolingual dictionary. So, if the morphological analyser encounters a word like *affects*, it will check whether the monolingual dictionary contains an entry with the features `cat = v, lex = affect`. Since it does, *affects* can be represented by means of the lexical entry, with some of the information supplied by the rule.The result of morphological analysis then is a representation which consists of both the information provided by the dictionary and the information contributed by the affix.

[6] In this rule we write `+finite` for `finite=+`. We also ignore some issues about datatypes, in particular, the fact that on the right-hand-side `V` stands for a string of characters, while on the lefthand (lexical) side it stands for the value of an attribute, which is probably an atom, rather than a string.

[7] More precisely, the rule is that the third person singular form is the base form plus *s*, except (i) when the base form ends in *s*, *ch*, *sh*, *o*, *x*, *z*, in which case *+es* is added (for example, *poach-poaches*, *push-pushes*), and (ii) when the base form ends in *y*, when *ies* is added to the base minus *y*.

```
lex = affect
cat = v
vtype = main
subcat = subj_obj
sem_agent = ?
sem_patient = ?
vform = finite
person = 3rdsing
tense = pres
```

In order to recognize irregular forms the morphological component has to contain explicit rules. One approach here is to try to normalise the spelling, so that the ordinary morphological rules can deal with the result. For example, one might have rules like the following to deal with the irregular third person singular forms of *be* and *have*.

```
be+s → is

have+s → has
```

Under this approach, morphological analysis for *is* and *has* is a two stage process.

The alternative is to state the relationship between the forms *is* and *has* directly, via rules like the following:

```
(lex=be,cat=v,+finite,person=3rd,number=sing,tense=pres)
  ↔ is

(lex=have,cat=v,+finite,person=3rd,number=sing,tense=pres)
  ↔ has
```

A graphic interpretation of the two alternative approaches is given in Figure 5.2.

Notice that we must ensure that these rules apply in the right cases. For example, *dies* should not be analysed as *di+es*. This is not problematic, providing we ensure that the analyses we produce contain actual lexical items.[8]

In synthesis, there is a related problem of making sure that the regular rules do not produce **bes*, and **haves*. One approach to this is to try to divide rules into exceptional and default groups, and to make sure that no default rule applies if a an exceptional rule can apply. Thus, for example, the fact that there is a special rule for the third person singular form of

[8]Notice, however, that we still cannot expect morphological analysis and lexical lookup to come up with a single right answer straight away. Apart from anything else, a form like *affects* could be a noun rather than a verb. For another thing, just looking at the word form in isolation will not tell us which of several readings of a word is involved.

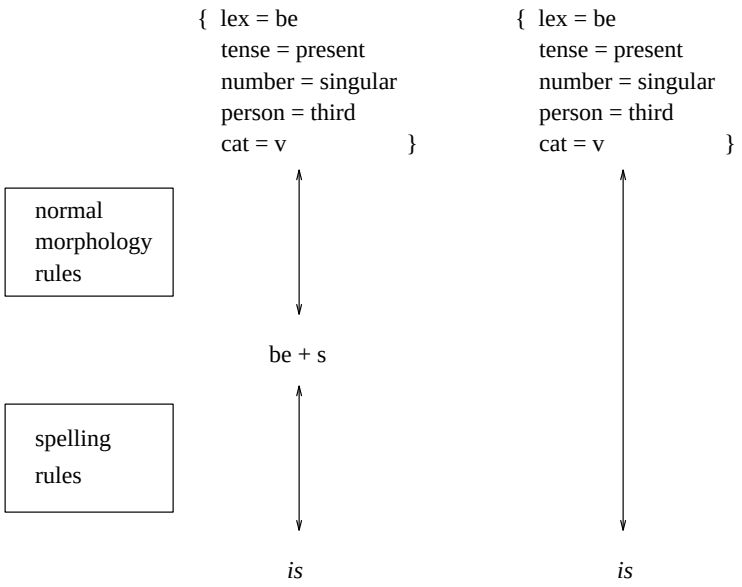

Figure 5.2 Treatment of Irregular Verbs

is would prevent the application of the normal or default rule that simply adds *s* to the base form of the verb.

Alternatively, one could add features to control which rules apply to lexical entries, and have the morphological rules check for the presence of the particular feature. This approach is particularly attractive in cases where a language has a number of conjugation or declension classes — lexical items can contain features indicating their conjugation/declension class, which the morphological rules can check.

So far, we have talked about morphological rules as things that actually apply as a sentence is being analysed. Another way in which one could use them is to compile out a full form dictionary from a dictionary of uninflected words, essentially by running the morphological rules over the dictionary of uninflected forms. Note, however, that this strategy would build a monolingual dictionary of an enormous size for languages like Spanish, or Finnish.

5.4.2 Derivation

Derivation processes form new words (generally of a different category) from existing words, in English this is mainly done by adding affixes. For example, *industrialization*, and *destruction* can be thought of as being derived in the way illustrated below. As one can see from *destruction*, it is not necessarily the citation form of a word that appears in derivations, for this reason it is common to talk of derivational processes involving stems and affixes (rather than words and affixes).

(6) a. $[_N\ [_V\ [_{ADJ}\ [_N$ industry $]$+ial $]$+ize $]$+ation $]$
b. $[_N\ [_V$ destroy $]$+ion $]$

In a paper dictionary, some derived words are listed, under the relevant head word. This is partly because affixes differ in their productivity and in the regularity of the effect they have on the words or stems that they combine with. For example, there seems to be no real basis on which to predict which of the noun-forming affixes produce nouns from particular verbs. This is illustrated below by the verbs *arrive*, *destroy*, and *deport*:

(7)

Verb	+al	+uction	+ation
arrive	arrival	*arruction	*arrivation
destroy	*destroyal	destruction	*destroyation
deport	*deportal	*depuction	deportation

However, some derivational processes are quite regular and can be described by means of a bf word grammar. This involves: (i) entering the affix in the dictionary; (ii) allowing it to subcategorize for what it combines with (e.g. *-able* combines with transitive verbs: witness *read-readable*) — this is just like normal syntactic subcategorization; (iii) making sure that the rules to combine words and affixes give the derived word the correct features for the result, and take care of any spelling changes in word or affix; (iv) finding some way of specifying the meaning in terms of the meanings of the word and affix.

As with inflection, the rules must be set up so as to produce only genuine lexical items. For example, we can ensure that the rules that analyse *cordiality* as *cordial+-ity* do not produce *qual+-ity* from *quality*, because there is no lexical item **qual*.

One approach to handling derivational morphology in MT is to simply list all derived words, and for some derived words (e.g. *landing*, in the sense of area at the top of stairs), this is clearly the right approach, because their meaning is unpredictable. But not all derivational morphology is unpredictable. Some affixes almost always have just one sense, like the prefix *un* which (when combined with an adjective) normally means 'not X' (*un-happy* means *not happy*)[9], and for others there are certain tendencies or regularities: with the examples in (8) the addition of the suffix *-ing* to the verb stem seems to have the same, regular consequence for the meaning of the word, so the derived word denotes the action or process associated with the verb (the act of Xing). Speakers exploit this fact by creating new words which they expect hearers to understand.

(8) a. The killing of elephants is forbidden.
b. Driving off went without any problems.
c. The painting of still lives never appealed to me.

In contrast with the examples in (8), one should consider the nouns in (9), where the meaning, although common, is not predictable from the suffix *-ing*:

(9) a. Painting: a picture produced with paint
b. Covering: something which covers something

[9]Note that the category of the stem word is important, since there is another prefix *un* which combines with verbs to give verbs which mean 'perform the reverse action to X' — to *unbutton* is to reverse the effect of buttoning.

c. Cutting: something which has been cut out
d. Crossing: a place were e.g. roads cross

We see here that a verb+ing noun can refer to a product (9a), a thing which performs an action (9b), a thing which undergoes an action (9c), or a place (9d). At the same time, however, it is true that in most cases the regular interpretation 'the act of Xing' is *also* available. What this means is that there is almost always a problem of ambiguity with derived words.

Moreover, there are cases where one can translate derived words by translating the stem, and translating the affix. For example, the French translation of English adverbs formed from an adjective plus *-ly* is often made up of the translation of the adjective plus *-ment* (e.g. *quick+ly* $\rightarrow$ *rapide+ment*, *easy+ly* $\rightarrow$ *facile+ment*), etc. But this is only possible for some affixes, and only when the interpretation of the derived word is predictable. The difficulties of translating derived words by translating stems and affixes can can be seen from the translation of the previous examples into Dutch.

(10) a. killing $\Rightarrow$ doden
b. driving off $\Rightarrow$ wegrijden
c. painting (the act) $\Rightarrow$ schilderen

d. painting (the product) $\neq$ schilderen, but $\Rightarrow$ schilderij
e. covering $\neq$ bedekken, but $\Rightarrow$ bedekking
f. cutting $\neq$ knippen, but $\Rightarrow$ knipsel
g. crossing $\neq$ kruisen, but $\Rightarrow$ kruispunt

Thus, though the idea of providing rules for translating derived words may seem attractive, it raises many problems and so it is currently more of a research goal for MT than a practical possibility.

5.4.3 Compounds

A compound is a combination of two or more words which functions as a single word. In English, the most common type of compound is probably a compound made up of two nouns (noun-noun compounds), such as those in the dictionary entry for *button*:

(11) a. buttonhole:
$[_N\ [_N$ button $][_N$ hole $]]$
b. buttonhook:
$[_N\ [_N$ button $][_N$ hook $]]$
c. button mushroom:
$[_N\ [_N$ button $][_N$ mushroom $]]$

In Spanish, for example, other types of compounds are equally important, including adjective-adjective compounds:

(12) a. guardacostas (‘coastguard’):
$[_N\ [_N$ guarda $][_N$ costas $]]$
b. rojiblanco (‘red and white’):
$[_A\ [_A$ roji $][_A$ blanco $]]$

Orthographically, different languages follow different conventions for compounds. For example, in German compounds are generally written as one word, but in English some are written as one word (as *buttonhole* and *buttonhook* above), some as hyphenated words (e.g. *small-scale*) and some as juxtaposed words (e.g. *button mushroom*).

As with derivations, it is possible to describe the range of possible compounds by means of a word grammar, and as with derivations the possibility that one might be able to translate compounds by translating the component parts is very attractive — especially since it is in principle not possible to list all English compounds, because compounding can give rise to words that are arbitrarily long. To see this, consider that one can form, in addition to *film society*:

(13) a. student film
b. student film society
c. student film society committee
d. student film society committee scandal
e. student film society committee scandal inquiry

Unfortunately, though there are cases where decomposing a compound and translating its parts gives correct results (e.g. the German compound *Wassersportverein* translates as *water sport club*), the problems of interpretation and translation are even worse for compounds than for derivations. Apart from the fact that some compounds have completely idiosyncratic interpretations (e.g. a *redhead* is a person with ginger coloured hair), there are problems of ambiguity. For example, *student film society* could have either of the structures indicated, with different interpretations (the first might denote a society for student films, the second a film society for students):[10]

(14) a. $[_N\ [_N$ student film$]$society$]$
b. $[_N$ student $[_N$ film society$]]$

A different type of ambiguity can be illustrated by giving an example: *satellite observation* may on one occasion of use mean *observation by satellite*, while on another occasion of use it might mean *observation of satellites*. Most of the time humans are able to rely on either our world knowledge or on the context to unravel a compound’s meaning. Moreover, it is frequently important for translation purposes to work out the exact relation expressed by a compound. In Romance languages, for example, this relation may be explicitly re-

[10]Where words have been fused together to form a compound, as is prototypically the case in German, an additional problem presents itself in the analysis of the compound, namely to decide exactly which words the compound consists of. The German word *Wachtraum*, for example, could have been formed by joining *Wach* and *Traum* giving a composite meaning of *day-dream*. On the other hand, it could have been formed by joining *Wacht* to *Raum*, in which case the compound would mean *guard-room*.

alised by a preposition. For example, *research infrastructure* in Spanish translates as *infraestructura para la investigación* (literally, 'infrastructure *for* research'). Nor can we happily assume that an ambiguity in one language will be preserved in another. Thus *satellite observation* has two possible translations in Spanish, depending on its meaning: *observación por satelite* ('observation by satellite') and *observación de satelites* ('observation of satellites').

A further problem with compounds is that a wide variety of relations are possible between the elements of a compound. Thus *buttonhole* is a hole for buttons, but *button mushroom* is a mushroom that resembles a button. It is not clear how to capture these relations.

Thus, as with derivations, a really general approach to the treatment of compounds remains a research goal for MT.

5.5 Terminology

The discussion so far has been about issues relating to general vocabulary. However, a slightly different, and somewhat less troublesome, set of issues arise when one turns to the specialist vocabulary that one finds in certain types of text in certain subject fields (the vocabulary of weather reports is an extreme example, other examples might be the vocabulary of reports on trials for medical reports, reports of tests of pharmaceutical drugs, or reports of particular kinds of sporting event). Such fields often have a relatively well-defined terminology, which is sometimes even codified, and given official recognition by professional bodies. What this codification involves is settling on a collection of concepts, and assigning each a name (or perhaps several names, one in each of several languages). When a word (or collection of words in several languages) designate a single concept in this way, it is called a term. Examples of terms include the names for material objects, but also the abstract entities (processes, properties, functions, etc). Concepts, and hence the associated terms, can be organized into conceptual structures, based on the relationship between them. For example tables, chairs, cupboards, etc. can be grouped together as *furniture*, with a possible subdivision into *household furniture* and *office furniture*.

Terms may be simple words or multiword expressions. Syntactically, there is nothing to distinguish terms from ordinary language, although there is a strong tendency for terms to be nouns, often compound nouns.

Terms are potentially more tractable for MT systems than general language vocabulary, since for the most part they tend to be less ambiguous. While a general language word may represent more than one concept in a system of concepts, there is frequently a one-to-one mapping between terms and the concepts they represent. Take for example the word *graduation*, which in machine tool terminology has the very precise meaning: "distribution of divisions on the scale of an apparatus (linear, logarithmic, quadratic, etc)" The general language word *graduation*, on the other hand, has many more meanings, including "the ceremony at which degrees are conferred". What this means, of course, is that one can in principle adopt an interlingual approach to terminology. For example, even in a transfer

system, one need not deal with terms on a language pair basis — all one may need is to have analysis and synthesis rules which relate the words for individual terms to an interlingual name for the associated concept (this could be an arbitrary numerical code, a collection of features, or even the actual term used in one of the languages, of course).

It is not always the case that a term represents one and only one concept — there are examples of terms which are ambiguous. For example, in machine tool terminology the term *screw* is defined as follows: "a machine thread whose essential element is a screw thread. A screw is either an external screw or an internal screw." (Likewise, synonymy amongst terms occurs, though much less frequent than in general language. In machine tool terminology, for example, *cramp* and *clamp* appear to designate the same concept.) However, the problems of ambiguity are small when compared to the problems one has with general vocabulary.

There are still some translational problems with terminology, however. In particular, there are problems whenever there is a mismatch between the conceptual systems of the two languages to be translated. An example of a concept mismatch from wine-making terminology is the difference between the English *acid* and the French *acide* which are defined as follows:

(15) a. **acid**: term applied to wine containing an excessive amount of acid, usually a wine made from grapes not completely ripe.
b. **acide**: caractère d'un vin dont la teneur elevée en acides organiques provient généralement de raisins incomplètement mûrs.

While the French definition speaks of *acides organiques* ('organic acids'), the English speaks only of *acids*. If the mismatch is considered significant enough, the term may need to be paraphrased in the other language. In such cases translating terminology raises the same problems as dealing with general vocabulary.

Fortunately, problem cases in terminology translation are much less frequent than in general vocabulary.

From the point of view of the human translator, and more particularly, groups of human translators collaborating on the translation of documents, terminology poses other sorts of problem. First, there is the problem of size — the sheer number of terms there are to deal with. Second, there is the problem of consistency.

With respect to the second problem, MT offers a considerable advantage. This is because once a term has been translated, it is possible to store the term and its translation, and ensure that the term is translated consistently throughout texts.

Of course, this is partly a solution to the problem of size also, because it ensures that the research and effort that goes into finding a translation for a term is not duplicated by other translators working with the same system. However, it is only a partial solution, because there is a seemingly inexorable increase in terminology in many subject areas. Many

hours of research are put into the recognition and documentation of new terms and their translational equivalents in other languages. To alleviate this problem, many translators and translation bureaux make use of **termbanks**, either pre-existing, or constructed in-house.

Termbanks are basically databases which contain many thousands of entries, one for every term. These entries consist, just like dictionary entries, of several fields, but the type of information given in these fields is rather different from that which one finds in an ordinary dictionary. Partly, this is because the proper documentation of a term requires specific information about the provenance of the entry, and about when it was created, and when modified (of course, one would expect to find information of this kind available to the builders of a properly documented dictionary too). Other information will typically concern related terms (synonyms, antonyms, abbreviations, superordinate terms and hyponyms), subject area (e.g. pharmaceutical products vs. sports goods), and sources of further information (e.g. specialist dictionaries or reference books). On the other hand, information about grammatical properties and pronunciation is typically rather scant. This is partly because terms are very often new words, or loan words, and typically follow the regular morphological rules of a language. Similarly, the lack of phonological information is partly because the entries are oriented towards written material, but also because it is expected that the terms will be phonologically regular (i.e. they will follow the normal rules for the language, or the normal rules that apply to loans words).

Apart from in-house termbanks which are local to a single organization, there are a number of large termbanks which offer open access (sometimes at a small charge). Examples are *Eurodicautom* (European Commission), *Termium* (Canadian Government), *Normaterm* (the French standards organization) and *Frantext* (National Institute of the French Language), which offer a range of terminology areas including science, technology, administration, agriculture, medicine, law and economics.

It should be evident from even this brief discussion that ensuring clear and consistent use and translation of terminology is a significant factor in the translation process, which in most technical domains necessitates the creation and maintenance of termbanks — itself a costly and time-consuming endeavour. It is not surprising, therefore, that with the increasing availability of large amounts of on-line texts, researchers have begun to experiment with the automatic extraction of terms from running text, using a variety of statistical methods to determine the likelihood that a word, or string of words, constitutes a term. Of course, lists of (putative) terms cannot be made to emerge magically from a corpus of texts - the process takes into account the frequency of items in the texts and is often guided by some information provided by the user, such as a thesaurus of concepts or concept hierarchy or a list of already identified terms, or a list of typical syntactic patterns for terms. There is no reason to expect such techniques to be limited to the extraction of monolingual terminology, and in fact the idea of automating to some degree the compilation of bilingual and multilingual termbanks is also gaining ground.

5.6 Summary

This Chapter has dealt with a number of issues concerning dictionaries in MT, including issues relating to various kinds of word structure (morphology), and terminology. Apart from stressing their importance, we have outlined the main sorts of information that one typically finds in dictionaries, and raised some questions about how this information should be represented.

5.7 Further Reading

A readable account of what is involved in producing a dictionary can be found in Sinclair (1987) — in this case the dictionary is monolingual, and intended for human readers, but many of the issues are similar. A general discussion of what are taken to be the main theoretical issues in the design and construction of dictionaries for NLP purposes is given in Ritchie (1987).

On morphology, Spencer (1991) provides an excellent up-to-date description of current linguistic theory. For a more extensive discussion of compounding see Bauer (1983). A detailed description of the state of the art as regards computational treatments of morphological phenomena is given in Ritchie et al. (1992). Almost the only discussion of morphology which is specifically related to MT is Bennett (1993).

For a general introduction to the study of terminology, see Sager (1990), on termbanks, see Bennett et al. (1986); McNaught (1988b, forthcoming, 1988a). For discussion of computerized termbanks and translation, see Thomas (1992). Experience of using a terminological database in the translation process is reported in Paillet (1990).

These days, many paper dictionaries exist in machine readable form (i.e. they have been created as 'electronic documents' in the sense of Chapter 8, below). OALD, the Oxford Advanced Learners' Dictionary Hornby et al. (1974), from which the monolingual entry on page 89 is taken, and LDOCE, Longman's Dictionary of Contemporary English Proctor (1978), are typical in this respect. They are sufficiently consistent and explicit to have been used in a number of experiments which try to take 'paper' dictionaries (or rather the machine readable versions of them), and convert them into a form which can be used directly in NLP systems. Some of this work is reported in Boguraev and Briscoe (1989).

The representation and use of lexical information in NLP is the focus of a great deal of research currently. Some idea of the range of this can be obtained from Evens (1988) and Pustejovsky and Bergler (1992). The idea of structuring a dictionary hierarchically so that individual entries can inherit information (and so be simplified), which we mentioned briefly, is particularly important in this research. A clearer idea of what is involved can be gained from (Pollard and Sag, 1987, Chapter 8).