Chapter 6

Translation Problems

6.1 Introduction

In this chapter we will consider some particular problems which the task of translation poses for the builder of MT systems — some of the reasons why MT is hard. It is useful to think of these problems under two headings: (i) Problems of *ambiguity*, (ii) problems that arise from *structural* and *lexical differences* between languages and (iii) multiword units like idioms and collocations. We will discuss typical problems of ambiguity in Section 6.2, lexical and structural mismatches in Section 6.3, and multiword units in Section 6.4.

Of course, these sorts of problem are not the only reasons why MT is hard. Other problems include the sheer size of the undertaking, as indicated by the number of rules and dictionary entries that a realistic system will need, and the fact that there are many constructions whose grammar is poorly understood, in the sense that it is not clear how they should be represented, or what rules should be used to describe them. This is the case even for English, which has been extensively studied, and for which there are detailed descriptions – both traditional 'descriptive' and theoretically sophisticated – some of which are written with computational usability in mind. It is an even worse problem for other languages. Moreover, even where there is a reasonable description of a phenomenon or construction, producing a description which is sufficiently precise to be used by an automatic system raises non-trivial problems.

6.2 Ambiguity

In the best of all possible worlds (as far as most Natural Language Processing is concerned, anyway) every word would have one and only one meaning. But, as we all know, this is not the case. When a word has more than one meaning, it is said to be **lexically ambiguous**. When a phrase or sentence can have more than one structure it is said to be **structurally ambiguous**.

Ambiguity is a pervasive phenomenon in human languages. It is very hard to find words that are not at least two ways ambiguous, and sentences which are (out of context) several ways ambiguous are the rule, not the exception. This is not only problematic because some of the alternatives are unintended (i.e. represent wrong interpretations), but because ambiguities 'multiply'. In the worst case, a sentence containing two words, each of which is two ways ambiguous may be four ways ambiguous (2×2), one with three such words may be $2 \times 2 \times 2 = 2^3 = 8$, ways ambiguous etc. One can, in this way, get very large numbers indeed. For example, a sentence consisting of ten words, each two ways ambiguous, and with just two possible structural analyses could have $2^{9+2} = 2^{11} = 2048$ different analyses. The number of analyses can be problematic, since one may have to consider all of them, rejecting all but one.

Fortunately, however, things are not always so bad. In the rest of this section we will look at the problem in more detail, and consider some partial solutions.

Imagine that we are trying to translate these two sentences into French:

(1) a. You must not use abrasive cleaners on the printer casing.
b. The use of abrasive cleaners on the printer casing is not recommended.

In the first sentence *use* is a verb, and in the second a noun, that is, we have a case of lexical ambiguity. An English-French dictionary will say that the verb can be translated by (inter alia) *se servir de* and *employer*, whereas the noun is translated as *emploi* or *utilisation*. One way a reader or an automatic parser can find out whether the noun or verb form of *use* is being employed in a sentence is by working out whether it is grammatically possible to have a noun or a verb in the place where it occurs. For example, in English, there is no grammatical sequence of words which consists of *the* + V + PP — so of the two possible parts of speech to which *use* can belong, only the noun is possible in the second sentence (1b).

As we have noted in Chapter 4, we can give translation engines such information about grammar, in the form of grammar rules. This is useful in that it allows them to filter out some wrong analyses. However, giving our system knowledge about syntax will not allow us to determine the meaning of all ambiguous words. This is because words can have several meanings even within the same part of speech. Take for example the word *button*.Like the word *use*, it can be either a verb or a noun. As a noun, it can mean both the familiar small round object used to fasten clothes, as well as a knob on a piece of apparatus. To get the machine to pick out the right interpretation we have to give it information about meaning.

In fact, arming a computer with knowledge about syntax, without at the same time telling it something about meaning can be a dangerous thing. This is because applying a grammar to a sentence can produce a number of different analyses, depending on how the rules have applied, and we may end up with a large number of alternative analyses for a single sentence. Now syntactic ambiguity may coincide with genuine meaning ambiguity, but very often it does not, and it is the cases where it does not that we want to eliminate by

applying knowledge about meaning.

We can illustrate this with some examples. First, let us show how grammar rules, differently applied, can produce more than one syntactic analysis for a sentence. One way this can occur is where a word is assigned to more than one category in the grammar. For example, assume that the word *cleaning* is both an adjective and a verb in our grammar. This will allow us to assign two different analyses to the following sentence.

(2) Cleaning fluids can be dangerous.

One of these analyses will have *cleaning* as a verb, and one will have it as an adjective. In the former (less plausible) case the sense is 'to clean a fluid may be dangerous', i.e. it is about an activity being dangerous. In the latter case the sense is that fluids used for cleaning can be dangerous. Choosing between these alternative syntactic analyses requires knowledge about meaning.

It may be worth noting, in passing, that this ambiguity disappears when *can* is replaced by a verb which shows number agreement by having different forms for third person singular and plural. For example, the following are not ambiguous in this way: (3a) has only the sense that the action is dangerous, (3b) has only the sense that the fluids are dangerous.

(3) a. Cleaning fluids is dangerous.
b. Cleaning fluids are dangerous.

We have seen that syntactic analysis is useful in ruling out some wrong analyses, and this is another such case, since, by checking for agreement of subject and object, it is possible to find the correct interpretations. A system which ignored such syntactic facts would have to consider all these examples ambiguous, and would have to find some other way of working out which sense was intended, running the risk of making the wrong choice. For a system with proper syntactic analysis, this problem would arise only in the case of verbs like *can* which do not show number agreement.

Another source of syntactic ambiguity is where whole phrases, typically prepositional phrases, can attach to more than one position in a sentence. For example, in the following example, the prepositional phrase *with a Postscript interface* can attach either to the NP *the word processor package*, meaning "the word-processor which is fitted or supplied with a Postscript interface", or to the verb *connect*, in which case the sense is that the Postscript

interface is to be used to make the connection.

(4) Connect the printer to a word processor package with a Postscript interface.

Notice, however, that this example is not genuinely ambiguous at all, knowledge of what a Postscript interface is (in particular, the fact that it is a piece of software, not a piece of hardware that could be used for making a physical connection between a printer to an office computer) serves to disambiguate. Similar problems arise with (5), which could mean that the printer and the word processor both need Postscript interfaces, or that only the word processor needs them.

(5) You will require a printer and a word processor with Postscript interfaces.

This kind of real world knowledge is also an essential component in disambiguating the pronoun *it* in examples such as the following

(6) Put the paper in the printer. Then switch it on.

In order to work out that *it* is the printer that is to be switched on, rather than the paper, one needs to use the knowledge of the world that printers (and not paper) are the sort of thing one is likely to switch on.

There are other cases where real world knowledge, though necessary, does not seem to be sufficient. The following, where two people are re-assembling a printer, seems to be such an example:

(7) A: Now insert the cartridge at the back.
B: Okay.
A: By the way, did you order more toner today?
B: Yes, I got some when I picked up the new paper.
A: OK, how far have you got?
A: Did you get it fixed?

It is not clear that any kind of real world knowledge will be enough to work out that *it* in the last sentence refers to the cartridge, rather than the new paper, or toner. All are probably equally reasonable candidates for fixing. What strongly suggests that *it* should be interpreted as the cartridge is the structure of the conversation — the discussion of the toner and new paper occurs in a digression, which has ended by the time *it* occurs. Here what one needs is knowledge of the way language is used. This is knowledge which is usually thought of as pragmatic in nature. Analysing the meaning of texts like the above example is important in dialogue translation, which is a long term goal for MT research, but similar problems occur in other sorts of text.

Another sort of pragmatic knowledge is involved in cases where the translation of a sentence depends on the communicative intention of the speaker — on the sort of action (the

speech act) that the speaker intends to perform with the sentence. For example, (8) could be a request for action, or a request for information, and this might make a difference to the translation.

(8) Can you reprogram the printer interface on this printer?

In some cases, working out which is intended will depend on the non-linguistic situation, but it could also depend on the kind of discourse that is going on — for example, is it a discourse where requests for action are expected, and is the speaker in a position to make such a request of the hearer? In dialogues, such pragmatic information about the discourse can be important for translating the simplest expressions. For example, the right translation of *Thank you* into French depends on what sort of speech act it follows. Normally, one would expect the translation to be *merci*. However, if it is uttered in response to an offer, the right translation would be *s'il vous plaît* ('please').

6.3 Lexical and Structural Mismatches

At the start of the previous section we said that, in the best of all possible worlds for NLP, every word would have exactly one sense. While this is true for most NLP, it is an exaggeration as regards MT. It would be a *better* world, but not the best of all possible worlds, because we would still be faced with difficult translation problems. Some of these problems are to do with lexical differences between languages — differences in the ways in which languages seem to classify the world, what concepts they choose to express by single words, and which they choose not to lexicalize. We will look at some of these directly. Other problems arise because different languages use different structures for the same purpose, and the same structure for different purposes. In either case, the result is that we have to complicate the translation process. In this section we will look at some representative examples.

Examples like the ones in (9) below are familiar to translators, but the examples of colours (9c), and the Japanese examples in (9d) are particularly striking. The latter because they show how languages need differ not only with respect to the fineness or 'granularity' of the distinctions they make, but also with respect to the basis for the distinction: English chooses different verbs for the action/event of putting on, and the action/state of wearing. Japanese does not make this distinction, but differentiates according to the object that is worn. In the case of English to Japanese, a fairly simple test on the semantics of the NPs that accompany a verb may be sufficient to decide on the right translation. Some of the colour examples are similar, but more generally, investigation of colour vocabulary indicates that languages actually carve up the spectrum in rather different ways, and that deciding on the best translation may require knowledge that goes well beyond what is in the text, and may even be undecidable. In this sense, the translation of colour terminology begins to resemble the translation of terms for cultural artifacts (e.g. words like English *cottage*, Russian *dacha*, French *château*, etc. for which no adequate translation exists, and for which the human translator must decide between straight borrowing, neologism, and

providing an explanation). In this area, translation is a genuinely creative act[1], which is well beyond the capacity of current computers.

(9)	a.	know (V)	savoir (a fact)
			connaître (a thing)
	b.	leg (N)	patte (of an animal)
			jambe (of a human)
			pied (of a table)
	c.	brown (A)	brun
			châtain (of hair)
			marron (of shoes/leather)
	d.	wear/put on (V)	kiku
			haku (shoes)
			kakeru (glasses)
			kaburu (hats)
			hameru (gloves, etc. i.e. on hands)
			haoru (coat)
			shimeru (scarves, etc. i.e. round the neck)

Calling cases such as those above lexical mismatches is not controversial. However, when one turns to cases of structural mismatch, classification is not so easy. This is because one may often think that the reason one language uses one construction, where another uses another is because of the stock of lexical items the two languages have. Thus, the distinction is to some extent a matter of taste and convenience.

A particularly obvious example of this involves problems arising from what are sometimes called **lexical holes** — that is, cases where one language has to use a phrase to express what another language expresses in a single word. Examples of this include the 'hole' that exists in English with respect to French *ignorer* ('to not know', 'to be ignorant of'), and *se suicider* ('to suicide', i.e. 'to commit suicide', 'to kill oneself'). The problems raised by such lexical holes have a certain similarity to those raised by idioms: in both cases, one has phrases translating as single words. We will therefore postpone discussion of these until Section 6.4.

One kind of structural mismatch occurs where two languages use the same construction for different purposes, or use different constructions for what appears to be the same purpose.

Cases where the same structure is used for different purposes include the use of passive constructions in English, and Japanese. In the example below, the Japanese particle *wa*, which we have glossed as 'TOP' here marks the 'topic' of the sentence — intuitively, what the sentence is about.

(10) a. Satoo-san wa shyushoo ni erabaremashita.

[1]Creative in the sense of 'genuine invention which is not governed by rules', rather than the sense of 'creating new things by following rules' — computers have no problem with creating new things by following rules, of course.

Satoo-hon TOP Prime Minister in was-elected
b. Mr. Satoh was elected Prime Minister.

Example (10) indicates that Japanese has a passive-like construction, i.e. a construction where the PATIENT, which is normally realized as an OBJECT, is realized as SUBJECT. It is different from the English passive in the sense that in Japanese this construction tends to have an extra adversive nuance which might make (10a) rather odd, since it suggests an interpretation where Mr Satoh did not want to be elected, or where election is somehow bad for him. This is not suggested by the English translation, of course. The translation problem from Japanese to English is one of those that looks unsolvable for MT, though one might try to convey the intended sense by adding an adverb such as *unfortunately*. The translation problem from English to Japanese is on the other hand within the scope of MT, since one must just choose another form. This is possible, since Japanese allows SUBJECTs to be omitted freely, so one can say the equivalent of *elected Mr Satoh*, and thus avoid having to mention an AGENT [2]. However, in general, the result of this is that one cannot have simple rules like those described in Chapter 4 for passives. In fact, unless one uses a very abstract structure indeed, the rules will be rather complicated.

We can see different constructions used for the same effect in cases like the following:

(11) a. He is called Sam.
b. Er heißt Sam.
'He is-named Sam'
c. Il s'appelle Sam.
'He calls himself Sam'

(12) a. Sam has just seen Kim.
b. Sam vient de voir Kim.
'Sam comes of see Kim'

(13) a. Sam likes to swim.
b. Sam zwemt graag.
'Sam swims likingly'

The first example shows how English, German and French choose different methods for expressing 'naming'. The other two examples show one language using an adverbial ADJUNCT (*just*, or *graag*(Dutch) 'likingly' or 'with pleasure'), where another uses a verbal construction. This is actually one of the most discussed problems in current MT, and it is worth examining why it is problematic. This can be seen by looking at the representations for (12) in Figure 6.1.

These representations are relatively abstract (e.g. the information about tense and aspect conveyed by the auxiliary verb *have* has been expressed in a feature), but they are still

[2]This discussion of the Japanese passive is a slight simplification. The construction does sometimes occur without the adversive sense, but this is usually regarded as a 'europeanism', showing the influence of European languages.

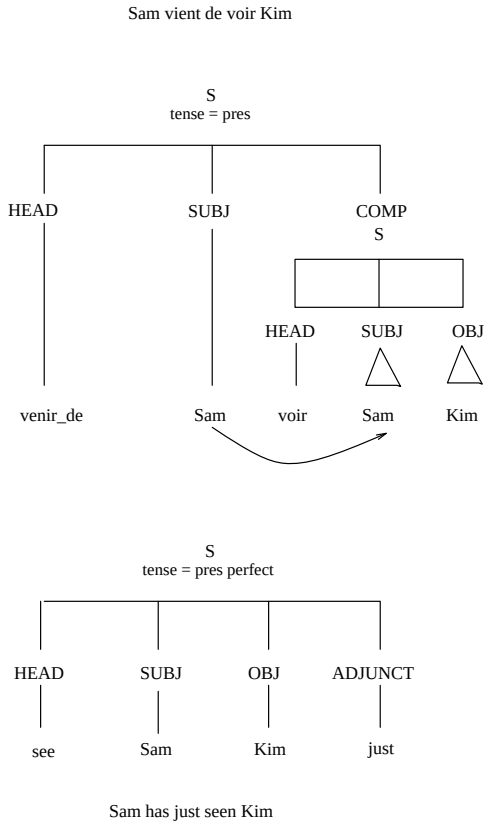

Figure 6.1 *venir-de* and *have-just*

rather different. In particular, notice that while the main verb of (12a) is *see*, the main verb of (12b) is *venir-de*. Now notice what is involved in writing rules which relate these structures (we will look at the direction English $\rightarrow$ French).

1 The adverb *just* must be translated as the verb *venir-de* (perhaps this is not the best way to think about it — the point is that the French structure must contain *venir-de*, and *just* must not be translated in any other way).

2 *Sam*, the SUBJECT of *see*, must become the SUBJECT of *venir-de*.

3 Some information about tense, etc. must be taken from the S node of which *see* is the HEAD, and put on the S node of which *venir-de* is the HEAD. This is a complication, because normally one would expect such information to go on the node of which the translation of *see*, *voir*, is the HEAD.

4 Other parts of the English sentence should go into the corresponding parts of the sentence HEADed by *voir*. This is simple enough here, because in both cases *Kim* is an OBJECT, but it is not always the case that OBJECTs translate as OBJECTs, of course.

5 The link between the SUBJECT of *venir-de* and the SUBJECT of *voir* must be established — but this can perhaps be left to French synthesis.

All this is summarized in Figure 6.2 and Figure 6.3.

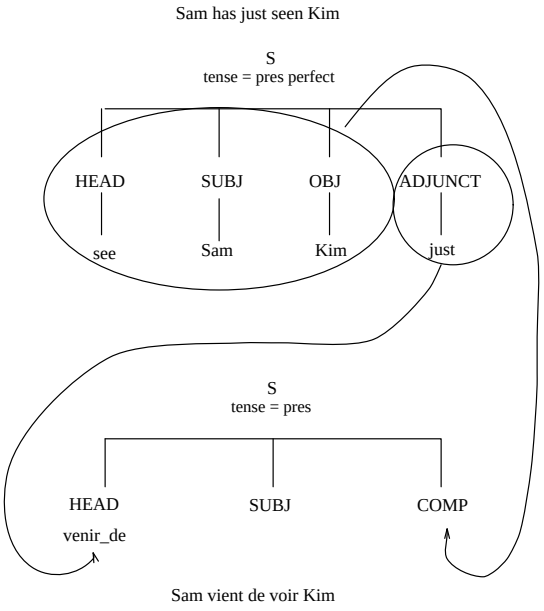

Figure 6.2 Translating *have-just* into *venir-de*

Of course, given a complicated enough rule, all this can be stated. However, there will still be problems because writing a rule in isolation is not enough. One must also consider how the rule interacts with other rules. For example, there will be a rule somewhere that tells the system how *see* is to be translated, and what one should do with its SUBJECT and OBJECT. One must make sure that this rule still works (e.g. its application is not blocked by the fact that the SUBJECT is dealt with by the special rule above; or that it does not insert an extra SUBJECT into the translation, which would give **Sam vient de Sam voir Kim*). One must also make sure that the rule works when there are other problematic phenomena around. For example, one might like to make sure the system produces (14b) as the translation of (14a).

(14) a. Sam has probably just seen Kim.
b. Il est probable que Sam vient de voir Kim.
'It is probable that Sam comes of see Kim'

We said above that everything except the SUBJECT, and some of the tense information goes into the 'lower' sentence in French. But this is clearly not true, since here the translation of *probably* actually becomes part of the main sentence, with the translation of (12a) as its COMPLEMENT.

Of course, one could try to argue that the difference between English *just* and French *venir de* is only superficial. The argument could, for example, say that *just* should be treated as a verb at the semantic level. However, this is not very plausible. There are other cases where this does not seem possible. Examples like the following show that where English uses a

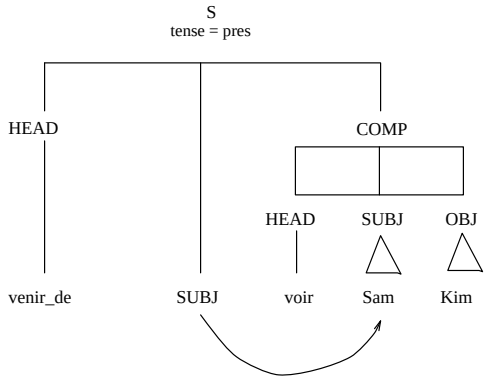

Figure 6.3 The Representation of *venir-de*

'manner' verb and a directional adverb/prepositional phrase, French (and other Romance languages) use a directional verb and a manner adverb. That is where English classifies the event described as 'running', French classifies it as an 'entering':

(15) a. She ran into the room.
b. Elle entra dans la salle en courant.
'She entered into the room in/while running'

The syntactic structures of these examples are very different, and it is hard to see how one can naturally reduce them to similar structures without using very abstract representations indeed.

A slightly different sort of structural mismatch occurs where two languages have 'the same' construction (more precisely, similar constructions, with equivalent interpretations), but where different restrictions on the constructions mean that it is not always possible to translate in the most obvious way. The following is a relatively simple example of this.

(16) a. These are the letters which I have already replied to.
b. *Ce sont les lettres lesquelles j'ai déjà répondu à.
c. These are the letters to which I have already replied.
d. Ce sont les lettres auxquelles j'ai déjà répondu.

What this shows is that English and French differ in that English permits prepositions to be 'stranded' (i.e. to appear without their objects, like in 16a). French normally requires the preposition and its object to appear together, as in (16d) — of course, English allows this too. This will make translating (16a) into French difficult for many sorts of system (in particular, for systems that try to manage without fairly abstract syntactic representations). However, the general solution is fairly clear — what one wants is to build a structure where (16a) is represented in the same way as (16c), since this will eliminate the translation problem. The most obvious representation would probably be something along the lines of (17a), or perhaps (17b).

(17) a. These are the letters $[_S$ I have already replied $[_{PP}$ to which $]]$
b. These are the letters $[_S$ I have already replied $[_{PP}$ to the letters $]]$

While by no means a complete solution to the treatment of relative clause constructions, such an approach probably overcomes this particular translation problem. There are other cases which pose worse problems, however.

In general, relative clause constructions in English consist of a head noun (*letters* in the previous example), a relative pronoun (such as *which*), and a sentence with a 'gap' in it. The relative pronoun (and hence the head noun) is understood as if it filled the gap — this is the idea behind the representations in (17). In English, there are restrictions on where the 'gap' can occur. In particular, it cannot occur inside an indirect question, or a 'reason' ADJUNCT. Thus, (18b), and (18d) are both ungrammatical. However, these restrictions are not exactly paralleled in other languages. For example, Italian allows the former, as in (18a), and Japanese the latter, as in (18c). These sorts of problem are beyond the scope of current MT systems — in fact, they are difficult even for human translators.

(18) a. Sinda node minna ga kanasinda hito wa yumei desita.
'died hence everyone SUBJ distressed-was man TOP famous was'
b. *The man who everyone was distressed because (he) died was famous.
c. L'uomo che mi domando chi abbia visto fu arrestato.
d. *The man that I wonder who (he) has seen was arrested.

6.4 Multiword units: Idioms and Collocations

Roughly speaking, **idioms** are expressions whose meaning cannot be completely understood from the meanings of the component parts. For example, whereas it is possible to work out the meaning of (19a) on the basis of knowledge of English grammar and the meaning of words, this would not be sufficient to work out that (19b) can mean something like 'If Sam dies, her children will be rich'. This is because *kick the bucket* is an idiom.

(19) a. If Sam mends the bucket, her children will be rich.
b. If Sam kicks the bucket, her children will be rich.

The problem with idioms, in an MT context, is that it is not usually possible to translate them using the normal rules. There are exceptions, for example *take the bull by the horns* (meaning 'face and tackle a difficulty without shirking') can be translated literally into French as *prendre le taureau par les cornes*, which has the same meaning. But, for the most part, the use of normal rules in order to translate idioms will result in nonsense. Instead, one has to treat idioms as single units in translation.

In many cases, a natural translation for an idiom will be a single word — for example, the French word *mourir* ('die') is a possible translation for *kick the bucket*. This brings out the similarity, which we noted above, with lexical holes of the kind shown in (20).

(20) a. J'ignore la solution.
b. I do not know the solution.
c. se suicider.
d. commit suicide.

Lexical holes and idioms are frequently instances of word ↔ phrase translation. The difference is that with lexical holes, the problem typically arises when one translates from the language with the word into the language that uses the phrase, whereas with idioms, one usually gets the problem in translating from the language that has the idiom (i.e. the phrase) into the language which uses a single word. For example, there is no problem in translating *I do not know the solution* literally into French — the result is perfectly understandable. Similarly, there is no problem in translating *mourir* 'literally' into English (as *die*) — one is not forced to use the idiom *kick the bucket*.

In general, there are two approaches one can take to the treatment of idioms. The first is to try to represent them as single units in the monolingual dictionaries. What this means is that one will have lexical entries such as `kick_the_bucket`. One might try to construct special morphological rules to produce these representations before performing any syntactic analysis — this would amount to treating idioms as a special kind of word, which just happens to have spaces in it. As will become clear, this is not a workable solution in general. A more reasonable idea is not to regard lexical lookup as a single process that occurs just once, before any syntactic or semantic processing, but to allow analysis rules to replace pieces of structure by information which is held in the lexicon at different stages of processing, just as they are allowed to change structures in other ways. This would mean that *kick the bucket* and the non-idiomatic *kick the table* would be represented alike (apart from the difference between *bucket* and *table*) at one level of analysis, but that at a later, more abstract representation *kick the bucket* would be replaced with a single node, with the information at this node coming from the lexical entry `kick_the_bucket`. This information would probably be similar to the information one would find in the entry for *die*.

In any event, this approach will lead to translation rules saying something like the following, in a transformer or transfer system (in an interlingual system, idioms will correspond to collections of concepts, or single concepts in the same way as normal words).

```
in_fact => en_fait

in_view_of => étant_donné

kick_the_bucket => mourir

kick_the_bucket => casser_sa_pipe
```

The final example shows that one might, in this way, be able to translate the idiom *kick the bucket* into the equivalent French idiom *casser sa pipe* — literally 'break his/her pipe'. The overall translation process is illustrated in Figure 6.4.

The second approach to idioms is to treat them with special rules that change the idiomatic source structure into an appropriate target structure. This would mean that *kick the bucket* and *kick the table* would have similar representations all through analysis. Clearly, this approach is only applicable in transfer or transformer systems, and even here, it is not very different from the first approach — in the case where an idiom translates as a single word, it is simply a question of where one carries out the replacement of a structure by a single lexical item, and whether the item in question is an abstract source language word such as *kick_the_bucket* or a normal target language word (such as *mourir*).

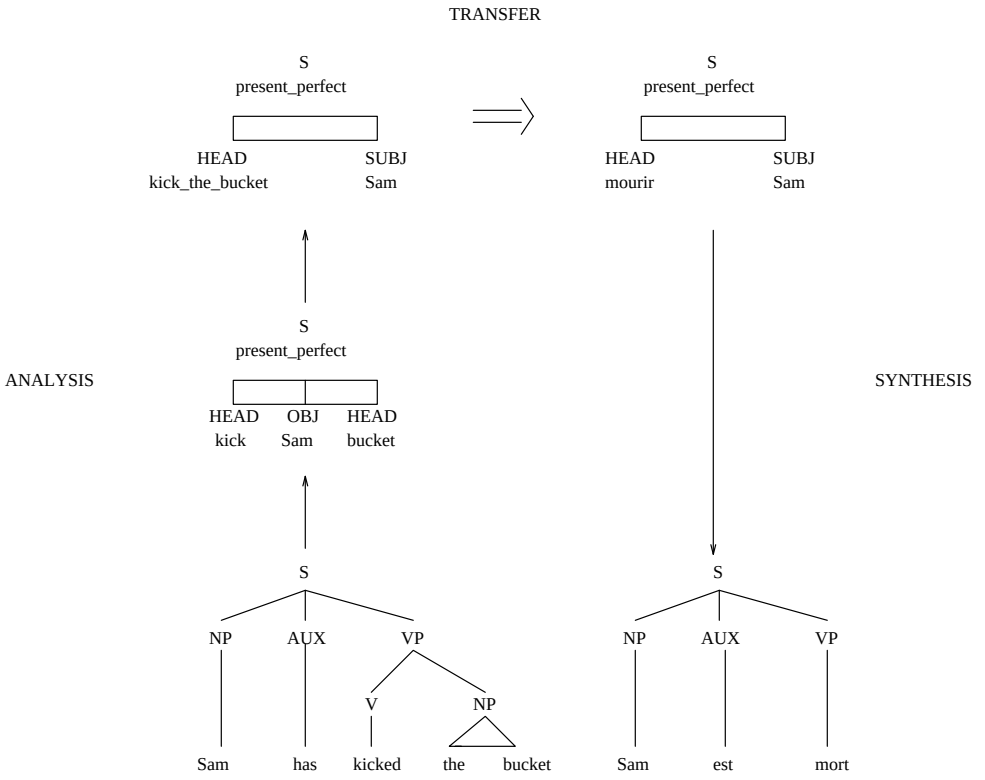

Figure 6.4 Dealing with Idioms 1

One problem with sentences which contain idioms is that they are typically ambiguous, in the sense that either a literal or idiomatic interpretation is generally possible (i.e. the phrase *kick the bucket* can really be about buckets and kicking). However, the possibility of having a variety of interpretations does not really distinguish them from other sorts of expression. Another problem is that they need special rules (such as those above, perhaps), in addition to the normal rules for ordinary words and constructions. However, in this they are no different from ordinary words, for which one also needs special rules. The real problem with idioms is that they are not generally fixed in their form, and that the variation of forms is not limited to variations in inflection (as it is with ordinary words). Thus, there is a serious problem in recognising idioms.

This problem does not arise with all idioms. Some are completely frozen forms whose parts always appear in the same form and in the same order. Examples are phrases like *in*

fact, or *in view of*. However, such idioms are by far the exception. A typical way in which idioms can vary is in the form of the verb, which changes according to tense, as well as person and number. For example, with *bury the hatchet* ('to cease hostilities and becomes reconciled', one gets *He buries/buried/will bury the hatchet*, and *They bury/buried/shall bury the hatchet*. Notice that variation in the form one gets here is exactly what one would get if no idiomatic interpretation was involved — i.e. by and large idioms are syntactically and morphologically regular — it is only their interpretations that are surprising.

A second common form of variation is in the form of the possessive pronoun in expressions like *to burn one's bridges* (meaning 'to proceed in such a way as to eliminate all alternative courses of action'). This varies in a regular way with the subject of the verb:

(21) a. He has burned [his] bridges.
b. She has burned [her] bridges.

In other cases, only the syntactic category of an element in an idiom can be predicted. Thus, the idiom *pull X's leg* ('tease') contains a genitive NP, such as *Sam's*, or *the king of England's*. Another common form of variation arises because some idioms allow adjectival modifiers. Thus in addition to *keep tabs on* (meaning *observe*) one has *keep* **close** *tabs on* ('observe closely'), or *put a* **political** *cat among the pigeons* (meaning 'do or say something that causes a lot of argument politically'). Some idioms appear in different syntactic configurations, just like regular non-idiomatic expressions. Thus, *bury the hatchet* appears in the passive, as well as the active voice.

(22) a. He buried the hatchet
b. The hatchet seems to have been buried

Of course, not all idioms allow these variations (e.g. one cannot passivize *kick the bucket* meaning 'die'), and, as noted, some do not allow any variation in form. But where variation in form is allowed, there is clearly a problem. In particular, notice that it will not be possible to recognise idioms simply by looking for sequences of particular words in the input. Recognising some of these idioms will require a rather detailed syntactic analysis. For example, despite the variation in form for *bury the hatchet*, the idiomatic interpretation only occurs when *the hatchet* is always DEEP OBJECT of *bury*. Moreover, the rules that translate idioms or which replace them by single lexical items may have to be rather complex. Some idea of this can be gained from considering what must happen to *pull Sam's leg* in order to produce something like equivalent to *tease Sam*, or the French translation involving *taquiner* ('tease'), cf. Figure 6.5. This figure assumes the input and output of transfer are representations of grammatical relations, but the principles are the same if semantic representations are involved, or if the process involves reducing *pull X's leg* to a single word occurs in English analysis.

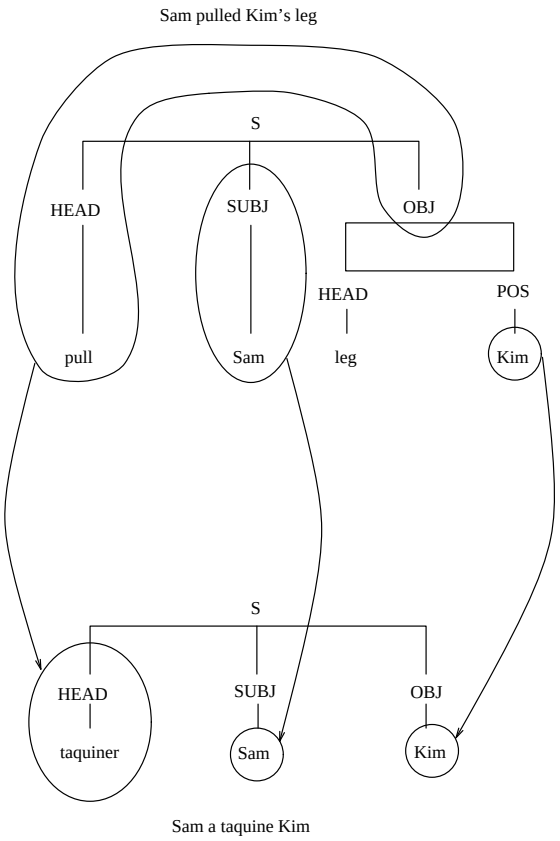

Figure 6.5 Dealing with Idioms 2

Rather different from idioms are expressions like those in (23), which are usually referred to as **collocations**. Here the meaning can be guessed from the meanings of the parts. What is not predictable is the particular words that are used.

(23) a. This butter is rancid (*sour, *rotten, *stale).
b. This cream is sour (*rancid, *rotten, *stale).
c. They took (*made) a walk.
d. They made (*took) an attempt.
e. They had (*made, *took) a talk.

For example, the fact that we say *rancid* butter, but not **sour butter*, and *sour cream*, but not **rancid cream* does not seem to be completely predictable from the meaning of *butter* or *cream*, and the various adjectives. Similarly the choice of *take* as the verb for *walk* is not simply a matter of the meaning of *walk* (for example, one can either *make* or *take* a journey).

In what we have called linguistic knowledge (LK) systems, at least, collocations can potentially be treated differently from idioms. This is because for collocations one can often think of one part of the expression as being dependent on, and predictable from the other. For example, one may think that *make*, in *make an attempt* has little meaning of its own, and serves merely to 'support' the noun (such verbs are often called **light verbs**, or **sup-**

port verbs). This suggests one can simply ignore the verb in translation, and have the generation or synthesis component supply the appropriate verb. For example, in Dutch, this would be *doen*, since the Dutch for *make an attempt* is *een poging doen* ('do an attempt').

One way of doing this is to have analysis replace the lexical verb (e.g. *make*) with a 'dummy verb' (e.g. `VSUP`). This can be treated as a sort of interlingual lexical item, and replaced by the appropriate verb in synthesis (the identity of the appropriate verb has to be included in the lexical entry of nouns, of course — for example, the entry for *poging* might include the feature `support_verb=doen`. The advantage is that support verb constructions can be handled without recourse to the sort of rules required for idioms (one also avoids having rules that appear to translate *make* into *poging* 'do').

Of course, what one is doing here is simply recording, in each lexical entry, the identity of the words that are associated with it, for various purposes — e.g. the fact that the verb that goes with *attempt* is *make* (for some purposes, anyway). An interesting generalisation of this is found in the idea of **lexical functions**. Lexical functions express a relation between two words. Take the case of *heavy smoker*, for example. The relationship between *heavy* and *smoker* is that of intensification, which could be expressed by the lexical function `Magn` as follows, indicating that the appropriate adjective for English *smoker* is *heavy*, whereas that for the corresponding French word *fumeur* is *grand* ('large') and that for the German word *Raucher* is *stark* ('strong').

```
(English) Magn(smoker) = heavy

(French) Magn(fumeur) = grand

(German) Magn(Raucher) = stark
```

If one wants to translate *heavy smoker* into French, one needs to map *smoker* into *fumeur*, together with the information that *fumeur* has the lexical function `Magn` applied to it, as in English. It would be left to the French synthesis module to work out that the value `Magn(fumeur)` = *grand*, and insert this adjective appropriately. Translation into German is done in the same way.

6.5 Summary

This chapter looks at some problems which face the builder of MT systems. We characterized them as problems of ambiguity (lexical and syntactic) and problems of lexical and structural mismatches. We saw how different types of linguistic and non-linguistic knowledge are necessary to resolve problems of ambiguity, and in the next chapter we examine in more detail how to represent this knowledge. In this chapter we discussed instances of lexical and structural mismatches and the problem of non-compositionality (as exemplified by idioms and collocations) and looked at some strategies for dealing with them in MT systems.

6.6 Further Reading

The problem of ambiguity is pervasive in NLP, and is discussed extensively in the introductions to the subject such as those mentioned in the Further Reading section of Chapter 3.

Examples of lexical and structural mismatches are discussed in (Hutchins and Somers, 1992, Chapter 6). Problems of the *venir-de*/*have just* sort are discussed extensively in the MT literature. A detailed discussion of the problem can be found in Arnold et al. (1988), and in Sadler (1993). On light verbs or support verbs, see Danlos and Samvelian (1992); Danlos (1992).

Treatments of idioms in MT are given in Arnold and Sadler (1989), and Schenk (1986). On collocations, see for example Allerton (1984), Benson et al. (1986a), Benson et al. (1986b) and Hanks and Church (1989). The notion of **lexical functions** is due to Mel'čuk, see for example Mel'čuk and Polguere (1987); Mel'čuk and Zholkovsky (1988).

A classic discussion of translation problems is Vinay and Darbelnet (1977). This is concerned with translation problems as faced by humans, rather than machines, but it points out several of the problems mentioned here.

The discussion in this chapter touches on two issues of general linguistic and philosophical interest: to what extent human languages really do carve the world up differently, and whether there are some sentences in some languages which cannot be translated into other languages. As regards the first question, it seems as though there are some limits. For example, though languages carve the colour spectrum up rather differently, so there can be rather large differences between colour words in terms of their extensions, there seems to be a high level of agreement about 'best instances'. That is, though the extension of English *red*, and Japanese *akai* is different, nevertheless, the colour which is regarded as the best instance of *red* by English speakers is the colour which is regarded as the best instance of *akai* by Japanese speakers. The seminal work on this topic is Berlin and Kay (1969), and see the title essay of Pullum (1991). The second question is sometimes referred to as the question of **effability**, see Katz (1978); Keenan (1978) for relevant discussion.