5
Analysis

This chapter is concerned with the specifically linguistic (rather than computational) problems which have to be faced in MT system design. It does not make reference to particular MT systems. Instead it discusses the issues in terms of an ideal mode] or a check-list to examine and evaluate particular systems.

The first general point to be made is that 'state-of-the-art' MT systems are not in general based on one single linguistic theory. As outlined in Chapter 2, a number of approaches to the linguistic description of natural languages have been influential in the design of MT systems. However, there have been very few systems based on a single linguistic model, and these have been predominantly experimental projects undertaking basic research using MT as a test-bed for computational linguistic theories. The great majority of MT systems are amalgams of different approaches and models, or even occasionally (particularly in the early years of MT research, cf. Chapter 1) with no discernible theoretical basis at all. Most commonly, systems are vaguely based on a general theory, such as transformational grammar or dependency theory, greatly modified by borrowings from other theories and by the demands of computational implementation.

MT research has often been criticised for ignoring developments in linguistic theory. There would appear to be a wide communication gap between theoretical linguistics and practical MT research. Some observers believe that there are good reasons for this situation: until recently, linguistic theories had not provided adequate accounts of all aspects of language use; a good linguistic theory may have given a convincing analysis of, say, quantifiers or coordination, but not explained all the peculiarities of actual usage in the coverage required for MT.

However, recent theories such as Lexical Functional Grammar or Generalized Phrase-Structure Grammar and their various derivatives have set out explicitly to cover as broad a range as possible, not only within one specific language, but also for different types of languages. In the past, and unfortunately it is still generally true today, much of linguistic theory was based on phenomena observed in English, the language of the majority of theoretical linguists. This neglect of other languages has been a further reason why linguistic theory has had less impact on MT than some observers might have expected. In other words, linguistic theories have rarely addressed questions of contrastive linguistics, i.e. the way in which different languages use different means to express similar meanings and intentions. Such questions are of course at the heart of MT.

Other reasons for the 'neglect' of linguistic theory by MT researchers are more practical: MT research is sometimes regarded as an 'engineering' task, a search for computational methods that work with the facts of language. The aims of many theoretical linguists are more abstract, concerned with investigations of human faculties in general, the nature of language itself, and the psychological foundations of language acquisition and use.

The result has been somewhat cynically described by Yorick Wilks (1989:59):

"... the history of MT shows, to me at least, the truth of two (barely compatible) principles that could be put crudely as 'Virtually any theory, no matter how silly, can be the basis of some effective MT' and 'Successful MT systems rarely work with the theory they claim to.'"

The search for solutions that work, whatever their theoretical status and whether or not they fit the alleged principles of the project, has meant that MT systems inevitably present a confusing picture of disparate methodologies and that researchers have been obliged to take much more pragmatic attitudes to theoretical issues than their colleagues in computational linguistics and in linguistic theory.

5.1 Morphology problems

It is generally accepted in MT system design that the first and last tasks of analysis and generation respectively are the treatment of morphology. In the context of MT, morphological analysis (and generation) is frequently regarded not as problematic in itself but rather as a means of helping to simplify the more difficult problems of lexical, syntactic and semantic analysis and generation. For example, by including a model of morphological analysis, it is often possible to reduce the size of the dictionaries considerably, and thus the human effort in creating them, and the computer time in accessing them. Although English, for example, has a relatively impoverished inflectional morphology, the size of an MT lexicon can be almost halved by treating singular-plural alternations systematically and by limiting verb entries to 'root' forms (e.g. *like* as the origin of *likes, liking* and *liked)*. With other languages, such as French and German, which have richer inflectional morphology, greater savings can be derived. Nevertheless, some older MT systems, and some small-scale experimental projects make use of **full-form dictionaries**, i.e. lexicons in which all inflected variants of the words appear. Two reasons are usually put forward for omitting morphological analysis: the perceived difficulties of dealing with irregular forms and the wide range of possible inflectional paradigms lead to

complex and time-consuming components; and the ever-decreasing access times of modem computers outweigh the problems previously caused by very large dictionaries.

However, morphological analysis has other benefits, principally in the recognition of **unknown words**, i.e. word forms not present in the system's dictionary. From the identification of grammatical inflections, it is often possible to infer syntactic functions, even if the root word itself cannot be translated directly.

As mentioned in section 2.4, there are three main areas of morphology: inflection, derivation and compounding. In many respects the morphological systems of many languages are well understood and systematically documented, even for languages not as intensively studied as English and other 'major' languages of the world; nevertheless, the computational implementation of morphological analysis is not entirely straightforward. Even the relatively mundane task of recognising and stripping affixes requires an elegant solution. The potential complexities of morphological routines can be illustrated as follows: it would be easy to imagine a simple rule for English which recognised a word ending in *-ing* as the present participle of a verb. To get the root form of the verb, all that would be needed is a procedure to strip off the ending. It would work well for *foaming, trying* and *testing,* but not for *having, hopping* and *tying,* the root forms of which are not **hav, *hopp* and **ty* but *have, hop* and *tie* respectively. These examples are well-known regular alternatives to the simple stripping rule, which must obviously be made more complex. And the rule must also ensure that *bring* and *fling* are not analysed as present participles of the 'verbs' **br* and **fl*! Although the solution is not computationally complex, this type of problem is quite common, and calls for some attention on the part of linguists.

Morphological analysis alone is sometimes sufficient to identify grammatical categories and structural functions, e.g. in English the suffix *-ize* usually indicates a verb, and German words ending in *-ung* are usually nouns. Frequently, however, morphological analysis cannot be divorced from syntactic analysis. Although German has a relatively rich inflectional morphology, providing indicators of subject-verb and adjective-noun agreement and marking case relations explicitly, it employs relatively few suffixes for these purposes, and many of the suffixes have multiple functions. A suffix such as *-en* can be highly ambiguous out of context, as Table 5.1 illustrates.

Most examples so far have been of inflectional morphology. Many languages have rich systems of **derivational** morphology and regularities in this area can also be exploited to reduce dictionary sizes. In English for example the recognition of negative prefixes such as *un-* and *non-,* the formation of adverbs from adjectives by the suffix *-ly* (though not all words with this ending are adverbs, e.g. *silly);* in French the suffix *-ment* is used to form adverbs (but again many words ending in *-ment* are nouns); and in German the suffix *-heit* can be used to form nouns from adjectives. However, derivational morphology has to be applied with care: for example, the English *-er* suffix frequently indicates a human agentive noun derived from a verb (a *dancer* is someone who dances, a *walker* someone who is walking), but there are often further meaning shifts involved: for example, a *computer* is a machine which computes, and a *revolver* is a weapon which has a mechanism which revolves.

noun plural
noun dative plural
weak noun singular non-nominative
strong declension* masculine singular accusative
strong declension* dative plural
adjective non-nominative masculine singular after definite article
adjective dative or genitive feminine or neuter singular after definite article
adjective accusative or genitive masculine singular without article
adjective genitive neuter singular without article
adjective dative plural without article
verb infinitive
verb 1st or 3rd person plural
verb past participle (strong verb)
a word which happens to end in *-en*

* articles, possessive and demonstrative adjective, etc.

Table 5.1 Possible interpretations of *-en* in German

The third aspect of morphological analysis involves the treatment of compounds. In English, terms are often coined by the simple juxtaposition of nouns (e.g. *steam engine, steam hammer),* and each component noun could appear in the dictionary; in the course of time, some juxtaposed nouns may be fused and become a single noun (e.g. *steamship)* and would be accommodated by inclusion of the full form in the dictionary. However, in languages such as German and Dutch, fusion is more common than juxtaposition (e.g. German *Dampfmaschine, Dampfhammer, Dampfschiff),* and novel original compounds are readily created (e.g. *Dampfschiffahrtsgesellschaft* 'steamship company'). A novel compound creates a problem for morphological analysis in an MT system: to treat it as an unknown word is unrealistic, since its meaning and correct translation can often be derived from its component parts; the difficulties lie in the multitude of feasible alternative segmentations, as exemplified in Table 5.2.

To illustrate the complexities that might arise with English examples, a morphological analysis routine might incorrectly segment *coincide* as *coin+cide, cooperate* as *cooper+ate, extradition* as *ex+tradition* or *extra+dition, mandate* as *man+date.*

Some of the possibilities ought to be excluded by the analysis program on semantic grounds if possible, but in other cases there may be a truly plausible ambiguity (e.g. *Wachtraum* as *Wacht+raum* or *Wach+traum, deferment* as *defer+ment* or *de+ferment).*

Alleinvernehmen	*Allein + Vernehmen*	'lone perception'
	All + Einvernehmen	'global agreement'
Beileid	*Bei + Leid*	'condolence'
	Beil + Eid	'hatchet oath'
Kulturgeschichte	*Kultur + Geschichte*	'history of culture'
	Kult + Urgeschtchte	'pre-history of worship'
Uranbrenner	*Uran + Brenner*	'uranium pile'
	Ur + Anbrenner	'primitive kindler'
Wachtraum	*Wach + Traum*	'day-dream'
	Wacht + Raum	'guard-room'

Table *5.2* Examples of ambiguous compounds in German

5.2 Lexical ambiguity

Some of the morphological problems discussed above involve ambiguity, that is to say, there are potentially two or more ways in which a word can be analysed. More familiar, and more complex, are lexical ambiguities, where one word can be interpreted in more than one way. Lexical ambiguities are of three basic types: category ambiguities, homographs and polysemes, and transfer (or translational) ambiguities.

5.2.1 Category ambiguity

The most straightforward type of lexical ambiguity is that of category ambiguity: a given word may be assigned to more than one grammatical or syntactic category (e.g. noun, verb or adjective) according to the context There are many examples of this in English: *light* can be a noun, verb or adjective; *control* can be a noun or verb. Especially common are noun-verb pairs, since almost any noun can function as a verb (examples in this sentence include *pair* and *function);* and almost any noun can function as an adjective (modifying another noun), hence the facility to produce new compound nouns in English, mentioned above. As an extreme but not rare example, *round* in the following sentences is a noun (la), verb (1b), adjective (1c), preposition (1d), particle (1e) and adverb (1f).

    (1a) Liverpool were eliminated in the first round.
    (1b) The cowboy started to round up the cattle.
    (1c) I want to buy a round table.
    (1d) We are going on a cruise round the world,
    (1e) A bucket of cold water soon brought him round.
    (1f) The tree measured six feet round.

   Category ambiguities can often be resolved by morphological inflection: e.g. *rounding* must be a verb form, although *rounds* could be a plural noun or a

third-person singular present tense verb. More frequently the ambiguity can be resolved by syntactic parsing. Thus in (1b) *round* could only be analysed as a verb because the syntactic context determines that, of the available options, only a verb would be appropriate here.

However, the problems increase when several categorially ambiguous words occur in the same sentence, each requiring to be resolved syntactically, as in (1c), where *want* could be a noun, *table* a verb. A good example is sentence (2).

(2) Gas pump prices rose last time oil stocks fell.

Bach of the words in this sentence has at least a two-way category ambiguity: all can be nouns or verbs, and *last* can be noun, verb, adjective or adverb. (As a noun it is in addition semantically ambiguous: see 5.2.2. below.) Despite the individual ambiguities, there is in fact only one correct analysis of this sentence, although it requires more than simple knowledge of possible category combinations to arrive at it: for a simple parser (which looks only at categories, and has information about possible legal combinations of them), this sentence is massively ambiguous. Although this example has been deliberately concocted, numerous examples — perhaps less extreme — can be found in real life, especially in the headlines of English newspapers, which typically omit function words which would radicate categories. Two actual examples are shown in (3).

(3a) Foot heads arms body.
(3b) British left waffles on Falklands.

To understand (3a) demands the 'real-world knowledge' (see 5.3.2.3 below) that a British politician named Foot was made chairman of a committee investigating arms control. In the case of (3b), the sentence makes sense for British readers familiar with the use of *waffle* as a verb (where *left* is then interpreted as a noun indicating political tendency). The interpretation of *waffle* as a noun (with *left* as a verb) is dismissed on grounds of implausibility (see again 5.3.2.3 below); but for American readers who know *waffle* only as a noun (a kind of doughnut) the sentence is unambiguous, but surprising.

5.2.2 Homography and polysemy

The second type of lexical ambiguity occurs when a word can have two or more different meanings. Linguists distinguish between homographs, homophones and polysemes. **Homographs** are two (or more) 'words' with quite different meanings which have the same spelling: examples are *club* (weapon and social gathering), *bank* (riverside and financial institution), *light* (not dark and not heavy). Homophones are words which are pronounced the same but spelled differently (e.g. *hair* and *hare),* but since MT is concerned essentially with written texts, they need not concern us here. **Polysemes** are words which exhibit a range of meanings related in some way to each other, e.g. by metaphorical extension or transference *(mouth* of a river, *branch* of a bank, *flow* of ideas, *channel* of communication, *tide* of opinion, etc.). When the extension becomes too distant from the origin then the polysemes effectively become homographs: for example, a *crane* as a lifting device may have been a metaphorical transfer from the bird sense of the word, but the words are no longer thought of as polysemes. Conversely, the meaning of *ear*

as in *ear of corn* might be thought to be a case of polysemy because of a physical similarity between the two concepts; but in fact the two words have different historical derivations (Latin *auris* and *acus)* and so technically are homographs. This demonstrates the fluidity of the difference, which in any case for MT (and computational linguistics) may be largely irrelevant

Sometimes one use of a homograph set may be more prevalent than the others, as for example with *rail* which is the name of a genus of wading bird with a harsh cry, as well as something a train runs on. In this case, the homograph could be disambiguated according to text-type, so that the unusual usage is simply excluded from the dictionary unless it is appropriate to the subject field of the texts to be translated.

In MT analysis homography and polysemy can often be treated alike, since it is a question of identifying the sense in context of a particular written 'word'. Homographs of different grammatical categories can be resolved as described above, but for homographs of the same category syntactic analysis alone is insufficient: semantic information must be used. One common approach is to assign **semantic features** such as 'human', 'female', 'liquid', etc. and to specify which features are compatible in given syntactic constructions, via **selection restrictions**. For example it might be specified that the verb *drink* must have an 'animate' subject.

There are difficulties in devising a consistently applicable set of semantic features and in specifying the selection restrictions of nouns and verbs in terms of such features. Nevertheless they are widely used in MT systems, often in conjunction with case roles (see 5.3.2.1 below). But semantic features do not solve all problems, even in situations for which they have been devised. For example, consider the homograph *ball* with one meaning 'ball$_1$ the spherical object and another, 'ball$_2$' the dance party. The two could easily be distinguished in the following sentences by defining appropriate feature co-occurrence restrictions.

(4a) The ball rolled down the hill.

(4b) The ball lasted until midnight.

In (4a), *roll* would require a round object as its subject, i.e. 'ball$_1$'. In (4b), the verb *last* would expect a subject with temporal duration, which a 'ball$_2$' does have. However, in a sentence beginning as in (4c),

(4c) When you hold a ball, ...

the word *hold* is also ambiguous, requiring as its direct object in one case ('grasp') a physical object and in the other ('organize') an event. Since *ball* can be either a physical object ('ball$_1$') or an event ('ball$_2$'), there is no way out of the dilemma until, hopefully, the rest of the sentence provides further linguistic or contextual information.

5.2.3 Transfer ambiguity

Category ambiguities, homography and polysemy are all examples of lexical ambiguities which cause problems primarily in the analysis of the source language text. They are monolingual ambiguities. In MT there are also transfer ambiguities (or translational ambiguities) which arise when a single source language word

can potentially be translated by a number of different target language words or expressions: several examples are given in the next chapter. The source language word itself is not ambiguous, or rather it is not perceived by native speakers of the language to be ambiguous; it is 'ambiguous' only from the perspective of another language. It is, therefore, a problem of translation — certainly a difficult one — but not a problem of linguistic analysis *per se.* Lexical transfer problems will be dealt with alongside other problems relating to transfer in the next chapter.

5.3 Structural ambiguity

5.3.1 Types of structural ambiguity

Whereas lexical ambiguities involve problems of analysing individual words and transferring their meanings, structural ambiguity involves problems with the syntactic structures and representations of sentences. Ambiguity arises when there is more than one way of analysing the underlying structure of a sentence according to the grammar used in the system. The two qualifications — "of a sentence" and "according to the grammar used in the system" — are important. The majority of MT systems are restricted to the sequential analysis of single sentences; they do not generally deal with larger units of translation such as paragraphs, and although some attempts are made to deal with features of the ways in which sentences are linked, e.g. pronouns, theme-rheme structure and so on, most such systems remain experimental. The second qualification is a reminder that no parser can go beyond the limitations of the grammar being implemented. If the grammar does not make distinctions which a human reader would make, then the parser will not be able to decide between alternative analyses. In the following sections there are a number of illustrations of this point: it is the grammar which determines whether a particular structure has more than one 'legal' interpretation and is therefore ambiguous. It is consequently valid to distinguish between 'real' ambiguities, for which a human might find several interpretations, and 'system' ambiguities which the human reader would not necessarily recognise.

5.3.1.1 Real structural ambiguity

Linguists in general are fond of writing about **real structural ambiguities**, in order to illuminate alternative syntactic interpretations revealed by formal analysis. Examples such as the following are well known (8)-(10):

(8) Flying planes can be dangerous.
(9) Time flies like an arrow.
(10) The man saw the girl with the telescope.

For each of these sentences it is possible for human readers to find more than a single interpretation, as shown by the following paraphrases.

(8a) It can be dangerous to fly planes.
(8b) Planes which are flying can be dangerous.

(9a)   The passage of time is as quick as an arrow.
(9b)   A species of flies called 'time flies' enjoy an arrow.
(9c)   Measure the speed of flies like you would an arrow.
(10a) The man saw the girl who possessed the telescope.
(10b) The man saw the girl with the aid of the telescope.

The recognition (by humans) of the ambiguities in these sentences may not always be easy (especially (9)); and **in context** — in a particular situation or in a specific text — the sentences may be quite unambiguous. For example, if (10) were to appear in a story, it would probably be obvious from the story-line who had the telescope, the man or the girl. However, since MT systems cannot make use of these contextual clues, except in a very limited way (see below), the sentences are effectively ambiguous. In these cases, the 'system' ambiguities coincide with 'real' ambiguities.

5.3.1.2 Accidental structural ambiguity

However, the generalization of such examples by the substitution of other verbs, nouns, prepositions, etc., leads to further structural ambiguities which would not be regarded as ambiguous by the human reader. Since the system lacks the information required to disambiguate, it treats real ambiguities and system ambiguities in the same way. We call such cases of ambiguity **accidental** (or 'system') **structural ambiguities**, and part of the interest in studying them is that in this way linguists can attempt to clarify what information should be added to the grammar so that the false reading is rejected.

Accidental structural ambiguities occur due to an accidental combination of words having category ambiguities, due to alternative grammatical uses for syntactic constituents, or due to different possible combinations of syntactic constituents. The types of structural ambiguities that occur differ from language to language, and, importantly, from grammar to grammar. However, it is useful to discuss here some of the types of ambiguity that a reasonably broad-coverage grammar of English might produce (or have to cope with), by way of exemplification.

Many ambiguities arise from the fact that a single word may serve in a different function within the same syntactic context. This is one possible consequence of the category ambiguities discussed in 5.2.1 above. Some examples of this have already been given: in (8), *flying* can either be a gerund governing a complement noun with the interpretation (8a), or an adjective modifying a noun (8b); in (9), *time* can be a subject noun, an imperative verb, or a modifying noun. These ambiguities are reflected in different structural interpretations of the same sentences; in terms of the Chomskyan model (see Chapter 2) they are **deep structure ambiguities**, because there are different 'deep structures' for the same 'surface structures'. As before, we can illustrate with real ambiguities: in (11a) the present participle *shaking* can function as an adjective (like *tiny* in (11b)) or as a gerundive verb (like *drinking* in (11c)).

(11a) He noticed her shaking hands,
(11b) He noticed her tiny feet.
(11c) He noticed her drinking beer.

The present participle can also be involved in an ambiguity between its use as a (verbal) noun and as a gerundive verb: compare the sentences in (12):

(12a) I like swimming.
(12b) I like tennis.
(12c) I like getting up late.

The word *that* in (13a) can be either a relative pronoun (equivalent to *whom* in (13b)) or a complementizer as in (13c):

(13a) They complained to the guide that they could not hear.
(13b) They complained to the guide whom they could not hear.
(13c) They complained to the guide that they could not hear him.

A final example of a deep structure ambiguity comes from the interpretation of a string of nouns either as a single constituent, i.e. as a compound noun, or with a constituent boundary in the middle. It is a particular problem in English which permits the omission of relative pronouns: consider the noun sequence *mathematics students* in (14a) and (14b):

(14a) The mathematics students sat their examinations.
(14b) The mathematics students study today is very complex.

In English the prevalence of noun compounds and the frequency of category ambiguities (particularly between nouns and verbs: see 5.2,1 above) means that this type of structural ambiguity is very common, and may accumulate to produce sentences like (2) above (reproduced here), which baffle most readers, until they have identified which words are functioning as verbs *(rose* and *fell)*.

(2) Gas pump prices rose last time oil stocks fell.

Another type of structural ambiguity arises from lack of information about the attachment of (typically) prepositional phrases and relative clauses to preceding constituents of sentences. These are **surface structure ambiguities** in contrast to the deep structure ambiguities discussed above. We have already seen examples of this attachment ambiguity in (10) and (13a) above (reproduced here).

(10)   The man saw the girl with the telescope.
(13a) They complained to the guide that they could not hear.

Sentence (10) is an example of **prepositional-phrase attachment a**mbiguity. The prepositional-phrase *with the telescope* may modify either the preceding noun *the girl* or the main verb of the sentence *saw.* In (13a) the relative clause *that they could not hear* might refer to the guide or might be the content of the complaint. Further examples show other types of attachment ambiguity: in (15), the ambiguity lies in attachment of the *in*-phrase to either the first preceding noun *story* or the second *aircrash;* in (16a) the prepositional-phrase *to Susan* could attach either to the main verb *mentioned* as also in (16b) or to the embedded verb *sent* as in (16c); in (17a) the phrase *about the strike* could be attached either to the modifying adjective *concerned* as in (17b) or to the main verb *told* as in (17c).

(15a) Have you seen the story about the aircrash in the paper?
(15b) Have you seen the story about the aircrash in the jungle?
(16a) John mentioned the book I sent to Susan.
(16b) John mentioned the book I sent him to Susan.
(16c) John mentioned the book I sent to Susan to his brother.
(17a) I told everyone concerned about the strike,
(17b) I told everyone concerned about the strike not to worry.
(17c) I told everyone about the strike.

5.3.2 Resolution of structural ambiguity

When syntactic analysis produces more than one possible interpretation for a sentence, it is necessary to find a way of choosing the correct one. The reason for this is that often (though not always) the translation into the target language will be different depending on the interpretation chosen. For example, the Japanese translation of (10) has to be unambiguous: either (18a) or (18b); in German, the translation of *that* in (13a) differs from one interpretation to the other (19).

(18a) *Otoko wa BŌENKYŌ wo MOTTE IRU onnanoko wo mita.*

MAN subj TELESCOPE obj HOLDING GIRL obj saw

'The man saw the girl who was holding the telescope'.

(18b) *Otoko wa BŌENKYŌ DE onnanoko wo mita.*

MAN subj TELESCOPE inst GIRL obj SAW

'The man, using the telescope, saw the girl.'

(19a) *Sie beklagten sich bei dem Reiseführer, DEN sie nicht hören konnten.*
(19b) *Sie beklagten sich bei dem Reiseführer, DAß sie nicht hören konnten.*

There are a number of options available for ambiguity resolution: the use of semantic or other linguistic information, the use of contextual clues, the use of non-linguistic 'real world knowledge', and interactive consultation. Other options include ignoring the ambiguity, using a 'best guess' default strategy or hoping for a 'free ride'.

5.3.2.1 Use of linguistic knowledge

Often, potentially ambiguous sentences can be disambiguated by reference to what might be called **linguistic knowledge**. There are various types of linguistic knowledge, but they all have in common that they make use of information about words and the way they combine, rather than the events in real life that sentences describe.

One such method is to provide parsers with information about co-occurrence restrictions, that is indications of how the presence of certain elements in a structure influences the likely presence of other elements. The clearest example is the use of **subcategorization frames** for verbs (cf. section 2.9.1). These indicate what types of complements are 'expected' by a particular verb. A verb such as *give* for example, expects to have a noun referring to a 'giver' as its subject, a noun referring to the thing 'given' as its direct object, and a noun referring to

a 'recipient' as indirect object. Furthermore, we can specify to a certain extent what types of nouns fill these syntactic roles, by assigning **semantic features** to them, for example the 'giver' should be animate and so on. In this way, given a potentially ambiguous pair of sentences like those in (20) the parser may produce correct interpretations *if* it knows that *read* may be modified by a prepositional phrase introduced by *in* when the noun concerned belongs to the set of nouns marked as 'readable' (e.g. *book, magazine, newspaper,* etc.)

(20a) I read about the aircrash in France.
(20b) I read about the aircrash in the paper.

In a similar way, the sentences in (15) can be analysed correctly if similar information for *story* and the *in*-phrase is coded.

This kind of information can be handled also at a more general level, in terms of Valency and Case grammar (cf. sections 2.9.5 and 2.9.6). In Valency, verbs are characterised by the number and type of complements which either must or might be associated with them. In Case grammar, the roles of dependent complements are identified, such as Agent, Patient (or Object), Instrument, Manner, Accompanier, Location, etc. A typical generalization is that the subjects of transitive verbs are Agents, which are typically 'animate', 'potent', etc. For *give,* the expected case roles may be Agent, Recipient and Patient. The example sentences in (21) wilt be correctly analysed if the parser has access to the information that *write* takes an Instrument *with*-phrase and that the Instrument should be a physical object (21a); that an Accompanier of *visit* and similar verbs should be animate (21b); and that *tell* takes a Manner *with*-phrase, with a restricted range of possible fillers *(stutter, accent,* etc.).

(21a)   He wrote the letter with a fountain-pen.
         cf. He wrote the letter with the parcel.
(21b)   He visited the museum with Ms brother.
         cf. He visited the museum with the Elgin marbles.
(21c)   He told the story with a funny accent.
         cf. He told the story with a funny ending.

5.3.2.2 Contextual knowledge

In practice, very few sentences are truly ambiguous: if nothing else serves to disambiguate, then usually the context in which the sentence occurs will suggest that one or other reading is to be preferred. In discussing example (10) above it was suggested that the story-line might indicate which of the two characters had the telescope. It might have been mentioned in the previous sentence, or a previous paragraph, or perhaps even some chapters earlier. But it was also stated that very few MT systems are able to make use of such **contextual knowledge** for precisely the reason that there is no hard and fast rule about where to look for the piece of 'knowledge' which will help disambiguate in a particular case. From the other side, even assuming an efficient way of storing text-derived knowledge such as 'the man has a telescope', it would be difficult to know which pieces of knowledge were likely to be useful later, or how long they should be stored on the off-chance that they would be needed: for it would be clearly impracticable to

extract and store every fact that could be inferred from every sentence of a given text, just in case it was needed to disambiguate something.

5.3.2.3 Real world knowledge

The third approach to ambiguity resolution when syntactic analysis is insufficient is to have recourse to what is generally called **real world knowledge**. An example of a structural ambiguity that is resolved by applying real world knowledge is sentence (22).

(22) The man saw the horse with the telescope.

In the similar earlier example (10) there was an ambiguity about the attachment of *with the telescope.* The same structural ambiguity is present in (22) too. But we know that *with the telescope* must modify *saw* because our knowledge of the world, and in particular of horses, makes the alternative improbable.

It should be noted that the line between real world knowledge and the linguistic knowledge discussed above is by no means clear cut. For example, in sentences like those in (23)

(23a) We will meet the man you told us about yesterday.
(23b) We will meet the man you told us about tomorrow.

there is no doubt that the adverbial *yesterday* attaches to the verb *told* in (23a) and that *tomorrow* attaches to *will meet* in (23b), rather than vice versa. It is a linguistic observation that the future tense of *tomorrow* is in agreement with the future tense of *will meet* and that the past tense of *yesterday* is in agreement with the past tense of *told.* But it is also a fact of reality that you cannot now make plans to do something in the past, and that you cannot already have spoken about something in the future. In other words, linguistic knowledge often correlates (not surprisingly) with real world knowledge.

Another example is a sentence like (24).

(24) The man saw the girl with red hair.

There is DO ambiguity here for human readers because they know that hair cannot be used to aid vision. In linguistic terms, *hair* is not a permissible Instrument for the verb *see.* The semantic feature coding of potential Instruments for the verb is in effect a way of defining the physical act of seeing. Again, linguistic knowledge and real world knowledge coincide to a targe extent.

The problem for MT systems is that it is at present impossible in practice to code and incorporate all the potential (real world) knowledge that might be required to resolve all possible ambiguities in a particular system, even in systems restricted to relatively narrow ranges of contexts and applications. Despite advances in Artificial Intelligence and in computing technology, the situation is unlikely to improve in the near future: the sheer complexity and intractability of real world knowledge are the principal impediments to quick solutions.

5.3.2.4 Other strategies

There are always going to be a residue of unresolved ambiguities (whether 'real' or 'system' ambiguities). In such situations the MT system could adopt strategies like those a human translator might apply in the same circumstances. The first might be to select the analysis which seems most 'natural' or most plausible. There has been ample psycholinguistic research into the question of which of competing structures are more easily understood, and this may be applied to build parsers which behave like humans when faced with these ambiguities.

The second strategy might be to ask the author if possible. Some interactive MT systems take this approach: they ask human operators to select the analysis which conforms with their knowledge of the subject and their understanding of the intention of the author. Consider an example like (25):

> (25) investigation of techniques of stress analysis by thermal emission

The interpretation of this example would demand specialised scientific knowledge in order to determine whether it is *analysis by thermal emission,* or *investigation by thermal emission,* and whether it should be *techniques for analysis of stress* or *analysis of techniques of stress.* Interaction with a human operator is discussed in more detail in section 8.3.3.

The third strategy would be to make a **best guess**, based on the relative likelihood of one analysis over another, based only on which structures are more or less common, regardless of the words involved. For example, in complex noun-phrases like (25) a 'best guess' might be to assume that each prepositional phrase modifies the noun preceding it; or that adverbial phrases always modify the most recent verb. Obviously, this 'blind' approach will sometimes lead to wrong analyses, but if the 'guesses' are well motivated, the correct analysis may be made more often than not, and at very little cost.

Finally, it may be hoped that the ambiguity does not even have to be resolved because it can be retained in the target language. This option, sometimes called a **free ride**, is available usually only for languages of similar structure and vocabulary. Taking the familiar example (10) again, the ambiguity of the prepositional-phrase attachment can be retained in French and German (26) though not in Japanese (see (18) above).

> (26a) *L'homme a vu la jeune fille avec le téléscope,*
> (26b) *Der Mann sah das Mädchen mit dem Teleskop.*

One important point must be stressed. The distinctions that have been made among different types of ambiguities are irrelevant from the perspective of the system itself. What matters is not whether an ambiguity requires linguistic, contextual or real world knowledge but whether the relevant data are available which permit disambiguation. If the system can recognise that an ambiguity exists and it has the means to resolve it, then the system will obviously proceed to resolve it.

5.4 Anaphora resolution

Besides ambiguity resolution, which is often seen as the most important linguistic problem for MT, another major difficulty is the resolution of pronoun references or **anaphora resolution**. 'Anaphora' is the term linguists use to refer to an oblique reference being made to an entity mentioned explicitly elsewhere in a text. The most frequent linguistic device for this is the use of pronouns like *it, he, they,* etc., demonstratives like *this,* and phrases like *the latter.* The pronoun *she* for example may refer to a female already described as *woman,* and the pronoun *it* to a previously mentioned object, *a hat,* as in (27).

(27) The woman bought a hat and she paid for it by cheque.

The identification of pronouns involves, therefore, the identification of the earlier noun to which they refer, called the pronoun's **antecedent**. In (27) there is no problem since both *she* and *woman* are female, and are presumably assigned semantic features to reflect the fact. Likewise *hat* has no sex, and so *it* is the appropriate pronoun.

The establishment of the antecedents of anaphora is very often of crucial importance for correct translation. When translating into languages which mark the gender of pronouns for example, it is essential to resolve the anaphoric relations. Consider a set of sentences like those in (28).

(28a) The monkey ate the banana because it was hungry.
(28b) The monkey ate the banana because it was ripe.
(28c) The monkey ate the banana because it was tea-time.

In each case, the pronoun *it* refers to something different: in (28a) the monkey, in (28b) the banana, and in (28c) it refers to the abstract notion of the time of the action. For translation into German, where pronouns take the same grammatical gender as their antecedents, the appropriate connections must be made, as illustrated in (29). Since *Affe* ('monkey') is masculine, the corresponding pronoun is *er; Banane* ('banana') is feminine, and so requires the pronoun *sie;* and the appropriate pronoun for (28c) is *es.*

(29a) *Der Affe hat die Banane gefressen, da er Hunger hatte.*
(29b) *Der Affe hat die Banane gefressen, da sie reif war,*
(29c) *Der Affe hat die Banane gefressen, da es die Teestunde war.*

Making the correct anaphora resolution involves much the same types of knowledge — linguistic or otherwise — as for ambiguity resolution. In fact anaphora can be thought of as a sort of ambiguity, in that the antecedent of a given pronoun might be uncertain. In (28), where there is a choice between *monkey* and *banana* as antecedents, it is possible to invoke linguistic knowledge of co-occurrence restrictions to indicate that the former is more plausible in (28a), because *be hungry* requires an animate subject, and the latter is more plausible in (28b) because a fruit is more likely to be ripe than an animal is. In (28c) neither antecedent is appropriate, but a 'time-of-day' phrase is acceptable as the complement of *it was.*

However, as in ambiguity resolution, linguistic knowledge is sometimes not sufficient for the correct analysis. When translating sentences like those in (30) into French,

(30a) The soldiers shot at the women and some of them fell.

(30b) The soldiers shot at the women and some of them missed.

it is necessary to know whether *some of them* refers to *the soldiers* (masculine) or *the women* (feminine), not only for the pronoun gender *(quelques-uns* vs. *quelques-anes)* but also for agreement of the past participle in the case of *tombé(e)s* ('fell'). Here, to get the correct translations (31) it is the application of contextual or real world knowledge which gives an 'understanding' of the sequence of events being described.

(31a) *Les soldats ont tiré sur les femmes et quelques-unes sont tombées.*

(31b) *Les soldats ont tiré sur les femmes et quelques-uns ont raté.*

5.5 Quantifier scope ambiguity

A final problem to be discussed is that of **quantifier scope ambiguity**. This problem has been well studied in the field of natural language question-answering systems, where it is perhaps more crucial than in MT. This type of ambiguity occurs when the scope or range of a quantifier like *some, all, none* is unclear (cf. section 2.8.4). English in particular freely allows a syntactic phenomenon called 'quantifier raising', so that a sentence like (32a) has the true meaning expressed in (32b).

(32a) I don't think he'll come.

(32b) I think he won't come.

Sometimes quantifier raising can introduce an ambiguity as in (33a) and (34a), which can each be interpreted in two different ways (33b,c) and (34b,c).

(33a) All women don't like fur coats.

(33b) Not all women like fur coats, only some do.

(33c) There are no women who like fur coats.

(34a) All wires are attached to a pin.

(34b) There is one (large) pin to which all wires are attached.

(34c) Each wire is attached to its own pin.

Consider finally this notice, which was spotted in Japan. If you know that in Japan non-smokers are still somewhat in the minority, you will not so quickly assume that reading (35b) is the correct one.

(35a) No smoking seats are available on domestic flights.

(35b) There are no seats where you may smoke on domestic flights.

(35c) There are "no smoking" sections on domestic flights.

Many languages of course have similar (though not necessarily identical) quantifier scope ambiguities: with luck the same ambiguity may be present in both languages, in which case it may be possible to retain it (i.e. as a 'free ride'). But if not, the ambiguity must somehow be resolved. Since this often involves contextual or real-world knowledge, it may well be very difficult.

There are a large range of other linguistic problems that have to be solved, and in the next chapter we will be discussing several of those that concern contrastive

aspects, that is problems that arise from specific differences between the language pairs involved. These are problems such as differing tense systems, modality, determination in noun phrases, thematisation, levels of politeness and formality, and so on.

5.6 Further reading

Discussion in the literature on the relationship between MT and theoretical linguistics has been quite heated and quite a few MT practitioners have taken the view that MT is an 'engineering' rather than theoretical problem.

In general, relatively little is written about morphological analysis in MT. Nevertheless, certain existing MT systems have significant morphological modules, notably the WOBUSU module in SUSY (Maas, 1980 and 1987:212-217; see also Chapter 11), in the work of the TAUM group (Lehrberger and Bourbeau, 1988:80-89; see Chapter 12), and GETA's Ariane system (Guilbaud, 1987:280-294; see Chapter 13).

For a discussion of compounding, see Selkirk (1982) and Lieber (1983).

Lexical ambiguity is discussed at length in Hirst (1987). Example (3b) was given to us by Scott Bennett. The discussion of homography and polysemy is based on Leech (1974:228ff).

Johnson (1983) has a brief discussion of parsing from the point of view of MT. Nirenburg (1987b) includes an interesting discussion of the question of linguistic vs. contextual or real-world knowledge.

Beyond the realm of MT, and particularly in the context of computational models of human parsing, syntactic ambiguity is even more widely studied. A good starting point might be Kurtzman (1984), Frazier (1985) and Pulman (1987). Disambiguation of attachment ambiguities is particularly spotlighted in Wilks *et al* (1985) and Hirst (1987).

Anaphora resolution is discussed in Reinhart (1983), though not specifically from the point of view of Computational Linguistics or MT. For a discussion from the CL point of view, see Barwise and Cooper (1981), Hobbs (1983) or Hobbs and Shieber (1987).