Chapter 10

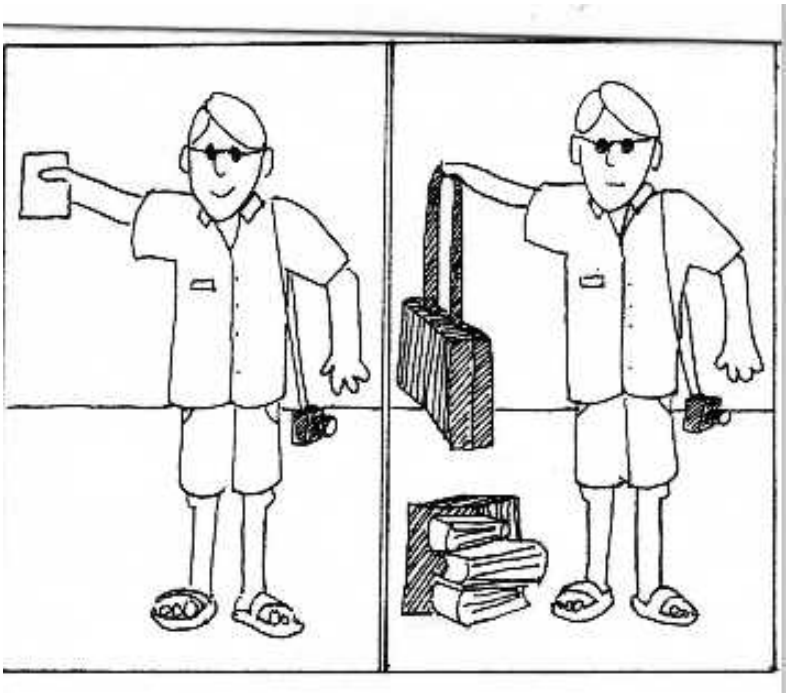
New Directions in MT

10.1 Introduction

In the previous chapters, we have tried to give an idea of what is currently possible in MT. In this chapter, we look to the future. Our aim is to give a flavour of current research in MT, indicating what issues are receiving attention and what techniques are thought to be promising.

Of course, not all the ideas that are currently important are really *new* ones. A great deal of current research is directed at how familiar techniques can be improved — for example, how standard 'Linguistic Knowledge'approaches can be improved by using better linguistic analyses (analyses based on better linguistic theories, or a better understanding of existing theories), and developing or adapting more efficient processing methods, and better tools for use in constructing and modifying systems. Likewise, an important feature of current research involves work on sublanguage MT (cf. Chapter 8), but though the design of tools to aid sublanguage analysis is an increasingly important area, it is hardly a new development. Other currently important work is concerned with *integration*, which can relate either to the integration of MT with other Natural Language Processing technologies, or to the (non-trivial) problems of integration of MT into general document processing technology that arise as one tries to make a practically and commercially usable system out of a research prototype MT system. A particularly important example of the former is research on 'speech-to-speech' MT systems — that is, systems that can take spoken input, and produce spoken output (e.g. for more or less simultaneous interpreting of telephone conversations). Such work is clearly important, and often throws up interesting differences of emphasis (for example, in speech-to-speech work, there is an emphasis on speed, and on dealing with sentence fragments, since one would like to be able to translate each utterance as it is spoken, without waiting for the end. This gives importance to 'bottom up' methods of analysis, and severe restrictions on the input in terms of text-type, etc). However, there is an obvious sense in which such work it is 'more of the same' — it involves improving one aspect of an existing idea, rather than presenting a genuinely new direction,

and would be accessible on the basis of the earlier chapters of this book. In this chapter, we will concentrate on what we think may turn out to be more radical ideas.



The Impact of Technology No. 58: Machine Translation and Tourism.
The Super Mini Etrans Tourist Translation System replaces the old fashioned Phrase Book. It comes complete with integrated laptop computer, carrying case, power pack, and 3 volumes of documentation.

The chapter has three main sections. In Section 10.2, we outline some current issues and trends in the design of sets of linguistic rules for MT, that is, work within the established 'Linguistic Knowledge', or 'Rule-Based' paradigm. The next section (10.3) gives an overview of some of the corpus and machine readable dictionary resources which have recently become available. These resources have stimulated a great deal of research within the traditional LK/rule-based paradigm, and have also been of key importance in the trend towards so-called *empirical* approaches to MT, which are sketched in Section 10.4.

10.2 Rule-Based MT

10.2.1 Flexible or Multi-level MT

Most transfer or interlingual rule-based systems are based on the idea that success in practical MT involves defining a level of representations for texts which is abstract enough to make translation itself straightforward, but which is at the same time superficial enough to permit sentences in the various source and target languages to be successfully mapped

into that level of representation. That is, successful MT involves a compromise between depth of analysis or understanding of the source text, and the need to actually compute the abstract representation. In this sense, transfer systems are less ambitious than interlingual systems, because they accept the need for (often quite complex) mapping rules between the most abstract representations of source and target sentences. As our linguistic knowledge increases, so too MT systems based on linguistic rules encoding that knowledge should improve. This position is based on the fundamental assumption that finding a sufficiently abstract level of representation for MT is an attainable goal. However, some researchers have suggested that it is not always the case that the deepest level of representation is necessarily the best level for translation.

This can be illustrated easily by thinking about translation between closely related languages such as Norwegian and Swedish.

(1) a. Min nya bil är blå(Swedish)
'my new car is blue'
b. Den nye bilen min er blå(Norwegian)
'the new car mine is blue'

(2) a. Var har du hittat en såful slips? (Swedish)
'Where did you find a such ugly tie'
b. Hvor har du funnet et såstygt slips? (Norwegian)
'Where did you find a such ugly tie'

In the second example here, both languages have exactly the same word order, although the words themselves and their grammatical features differ. In the first example, we see that Swedish (like English) does not allow the use of an article together with a possessive pronoun, which Norwegian (like, say, Italian) does. These are certainly minimal differences, and it would be a serious case of overkill to subject the source language sentences to 'in depth' analysis, when essentially all that is required to deal with this structural difference is to express a correspondence between the structures described by the following syntactic rules (here 'Poss' stands for 'Possessive pronoun').

(Swedish) NP $\rightarrow$ Poss Adj N

(Norwegian) NP $\rightarrow$ Det Adj N Poss

Of course, it would be straightforward to design a special purpose MT system which was equipped only with the sort of linguistic rules required to perform this type of superficial manipulation of syntactic structures. But a number of considerations, not least economic considerations, militate against this. Instead one could conclude that what is required is an approach to rule-based translation which is sufficiently *flexible* to carry out deep analysis only when required, so that the same MT engine can be used for dealing with pairs of closely related languages and pairs of languages which differ greatly. Such ideas lie behind attempts to design flexible systems which can operate in a variety of modes,

according to the depth of analysis required for the language pair, or even the particular examples in hand.

There are other reasons for the current interest in flexible systems. In the example above, we have tried to show that what is the 'appropriate level' of analysis for one language pair might be quite inappropriate for another pair. But some researchers have pointed out that a similar situation obtains within one and the same language pair. Though really convincing arguments are hard to find, the idea is that translation seems to depend on information about different levels of linguistic information at the same time. For example, for most translation purposes, as we have noted previously, a representation in terms of semantic relations (AGENT, PATIENT, etc.) is attractive. However, such a representation will probably not distinguish between (2a), (2b) and (2c). This means they will be translated alike, if this is the representation that is produced by analysis. But in many cases this would not produce a very good translation.

(3) a. Sam broke the printer.
b. It was the printer that Sam broke.
c. It was Sam that broke the printer

Ideally, what one wants is a semantic account of the differences between these examples. This has to do with the difference between what is presupposed, and what is asserted, or what is treated as 'given', and what as new information (e.g. in (3b) it is presupposed that Sam broke something, and stated that the thing in question was the printer). Producing such an account is not impossible, and may indeed produce a better MT system in the long run. However, it is by no means easy, and, at least in the short term, it would be nice if one could use information about semantic relations where that is useful, and information about surface syntactic form where that was useful. This would be possible if one had a way of allowing information from a variety of levels to be referred to in transfer. Of course, the difficulty then would be to allow this flexibility while still ensuring that the pieces of information can be correctly combined to give a suitable target translation.

There are various proposals in the MT literature concerning flexible MT. Some researchers working within the paradigm of example-based MT, which we discuss below, have proposed architectures which are flexible with respect to the level at which translation occurs. Another rather radical idea depends on the fact that several contemporary linguistic theories provide a 'multidimensional' characterisation of a linguistic string. One can get a flavour of what is involved by looking at the following representation.

This representation of the sentence *Kim walks* is multidimensional, in the sense that it contains information about several levels, or dimensions, of structure at the same time: information about ORTHography, SYNtax, SEMantics, and constituent structure (the DaughTeRs feature). Such multidimensional representations are known as **signs**. Identity of values is indicated by tags, boxed indices like $\boxed{1}$, $\boxed{2}$.

If we look first at the DTRS value, we can see that there are two daughters, the first an NP (i.e. whose SYNtax contains an attribute CAT with value NP), and the second a

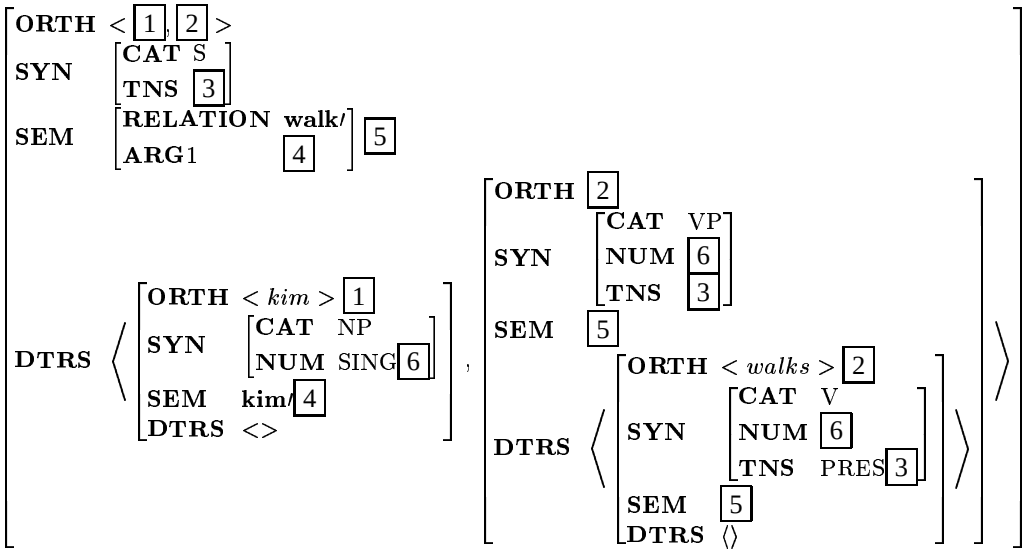


Figure 10.1 A Multidimensional Representation

VP. The NP has no daughters, and the VP has one daughter, whose category is V. The ORTHography of the whole S is made up of 1, the ORTHography of the NP, i.e. *mary*, and the ORTHography of the VP, which is identical to the ORTHography of the V, tagged 2. The TNS (TeNSe) of S, VP, and V are identical, and the NP, VP, and V have the same NUMber value.

The semantics of the S indicates that the argument of the predicate **walk/** is the value tagged 4, that is, the semantics of the NP, **mary/**.

We have seen that representation carries information about ORTHography, SYNTax, SEMantics and daughters (DTRS) at the same time (a fuller representation would include information about morphology too). Formally, it is just a collection of features (i.e. attributes and values) of the kind we have seen before, with the difference that the value of some of the attributes can be an entire structure (collection of features), and we allow different attributes to have the same value (indicated by means of a **tag**, a number written in a box). This is sometimes called a re-entrance.[1]

The syntactic information is essentially equivalent to the sorts of category label we have seen before, and the value of the DTRS attribute simply gives the values of the daughters a node would have in a normal consituent structure tree of the kind that were given in Chapter 3. One interesting point to note is that there is a value for SEMantics given for the mother sign, and for every one of the daughter signs. (In fact, the SEM value of the S is

[1]Here 'same value' is to be interpreted strongly, as *token* identity — in a sentence with two nouns, there would be two objects with the 'same' category value, namely, the two nouns. This is often called 'type' identity. In everyday usage, when we speak of two people having the 'same' shirt, we normally mean type identity. Token identity would involve them sharing one piece of clothing. On the other hand, when we speak of people having the same father, we mean token identity.

identical to the SEM value of the VP, and the V, and the SEM value of the AGENT of the S is identical to the SEM value of the NP *Kim*.)

One way one could use such a structure would be just to take the value of the SEM attribute for the mother sign in the output of analysis, and input this value to transfer (in a transfer system) or synthesis (in an interlingual system). This would involve only adapting the techniques we described in earlier chapters for transfer and synthesis to deal with complex attribute-value structures, rather than trees (this is not very difficult). Of course, this would mean that one was losing any benefit of multidimensionality for translation (though one might be able to exploit it in analysis).

If one is to exploit multidimensionality in transfer or synthesis (which was the aim) the only possible part of the sign to recurse through, applying rules, is the structure of the DTRS attribute. However, as we noted, this is just the surface phrase structure, enhanced with some information about semantics and orthography. If this is so, then one might wonder whether any advantage has been gained at all.

The solution is not to think in terms of applying rules to representations or structures at all, but to focus on the attribute-value structure as simply a convenient graphic representation of the solution to a set of constraints. For example, for the representation on page 177, one such constraint would be that the CATegory value of the mother sign is S. More precisely, the value of SYN on the mother sign is an attribute-value structure which contains an attribute CAT, with value S. That is, if we give names like X0, X1, X2, etc. to the various attribute-value structures, with X0 the name of the mother sign, then the value of SYN in X0 is a structure X1, and the value of CAT in X1 is S:

```
X0:SYN = X1
X1:CAT = S
```

If we name the attribute-value structure of the VP X4, and that of the V X5, we also have the following, indicating that S, VP, and V all have the same SEM values.

```
X0:SEM = X4:SEM
X4:SEM = X5:SEM
```

The value of the ORTHography attribute in X0 is the concatenation of the values in the NP (X6) and the VP (X5):

```
X0:ORTH = concatenation(X6:ORTH, X5,ORTH)
```

One can think of a representation like that on page 177 as simply a graphic representation of the solution to a set of such equations, and one can use the equations as the basis for

translation, in the following way. First, it is the task of analysis to produce the equation set. This is not, in fact, difficult — we have already seen, in Chapter 3 how one can add instructions to grammar rules to create different kinds of representations. Using them to create sets of equations is a simple extension of this idea. This set of constraints describes a source structure. The translation problem is now to produce a set of constraints whose solution will yield a target language structure. Ultimately, of course, one is interested in the ORTH value in such a structure, but in the meantime, one can state constraints such as: "the SEM of the source structure, and the SEM of the target structure must be identical" (this assumes that the SEM values are 'interlingual'), or "the SEM of the target structure must be the result of applying some 'transfer' function to the SEM of the source structure". But one can easily state constraints in terms of other attributes, for example, "in the case of proper nouns, the value of ORTH in the source structure and the value of ORTH in the target structure must be the same". Similarly, if we add attributes and values giving information about grammatical relations such as subject, etc. into the constraints, we can state constraints in terms of these.

Of course, we cannot, in this way, guarantee that we will deal with all of the source structure (we may leave parts untranslated by failing to produce appropriate target language constraints), or that solving the target language constraints will produce a single target structure, or even any structure at all (the constraints may be inconsistent). Nor have we indicated *how* the constraints are to be solved. Moreover, one will often not want such constraints to be observed absolutely, but only by default. For example, proper names should only keep the same orthography form if there is no constraint that says otherwise (in translating English into French, one would like to ensure that *London* translates as *Londres*). There are a number of serious difficulties and open research questions here. However, one can get a feeling for a partial solution to some of these problems by considering the following rather simple approach.

Recall that the constraints we gave above made the SEMantics of the S equal to the SEMantics of the VP, and the V. One may immediately think of this as involving the V contributing its SEMantics to the S, but one can also see it the other way round, as putting the semantics of the whole S 'into' the V. What this means, of course, is that all the semantic information conveyed by the sentence is represented (somewhat redundantly) in the representations of the words. Now suppose that we have translation constraints which say, for example, that the translation of the word *walk* must be the word *marcher*, with the same semantics, and that the translation of *Sam* must be *Sam*, again with the same semantics. What we must do now is produce a target structure. The problem we have is interestingly like the problem we have when we try to parse a sentence: then we typically know what the words are, and what order they are in, but not what the sentence as a whole means; here we know what the words are, and what the sentence as a whole means (it is represented 'in the words'), but not what the word order should be. One possibility is simply to use the target grammar to parse *Sam*, and *marcher* in all possible orders. To take a slightly more interesting case, suppose the source sentence is (3):

(4) Sam sees London.

If the target language is French, the target grammar will be asked to parse the strings in (4):

(5) a. *voit Sam Londres.
b. ?Londres voit Sam.
c. *Sam Londres voit.
d. Sam voit Londres.

One can expect the target grammar to reject (5a), and (5c). It would accept (5b), but only with the meaning that is different from that of the source sentence, which we have carried over in the constraints linking *see* to *voir*. This leaves only the correct solution (5d).

10.2.2 Knowledge-Based MT

The term *knowledge-based MT* has come to describe a rule-based system displaying extensive semantic and pragmatic knowledge of a domain, including an ability to reason, to some limited extent, about concepts in the domain (the components, installation and operation of a particular brand of laser printer could constitute a domain). We noted the appeal of such an approach as a way of solving some basic MT problems in earlier chapters. Essentially, the premise is that high quality translation requires in-depth understanding of the text, and the development of the *domain model* would seem to be necessary to that sort of deep understanding. One of the important considerations driving this work is an appreciation that post-editing is time-consuming and very expensive, and therefore that efforts made to produce high quality output will pay off in the long run. Since this may well turn out to be of great utility, in this section we concentrate on an approach which attempts some degree of text understanding on the basis of detailed domain knowledge, developed at the Center for Machine Translation at Carnegie Mellon University in Pittsburgh.

Subclasses	personal-computer mini mainframe super
is-a	independent device
has-as-part	software computer-keyboard input-device disk-drive
	output-device CD-Rom card computer-hardware-card cpu
	memory-expansion-card monitor printer system unit
max-users	(<>1 200)
make	Plus AT XT 750 780
token	"The basic IBM Personal Computer consists of a system
	unit and keyboard"
Part-of	airport-check-in-facility security-check-device
operational	yes no
manufactured-by	intentional-agent
configuration	minimal regular extra
theme-of	device-event spatial-event

Table 10.1 Example Frame for the concept `computer`

To give some idea of what is at stake here, the prototype systems developed for English $\leftrightarrow$ Japanese translation during the late 1980s at CMU, dealing with the translation of instruction manuals for personal computers, contained the following components:

- an ontology of concepts
- analysis lexica and grammars for English and Japanese
- generation lexica and grammars for English and Japanese
- mapping rules between the Interlingua and English/Japanese syntax

For a small vocabulary (around 900 words), some 1500 concepts were defined in detail. The ontology dealt solely with the interaction between personal computers and their users. Nouns in the interlingua correspond to 'object concepts' in the ontology, which also contains 'event concepts', such as the event `remove`, corresponding to the English verb *remove* and the Japanese verb *torinozoku* (by no means are all mappings from the interlingua into natural language as straightforward as this, for example, the concept `to-press-button` must be divided into subevents corresponding to pressing, holding down and releasing the button). Concepts are represented in a form of frame representation language, familiar from work in Artificial Intelligence and Natural Language Processing, in which frames (providing an intrinsic characterisation of concepts) are linked in a hierarchical network. To give an idea of the amount of detailed knowledge about concepts that one might want to encode, Table 10.1 gives by way of example a frame for the concept `computer`.

Knowledge-based MT is still pursued today at CMU in the KANT system, but is much more modest in terms of its goals for domain knowledge, which is limited to that which

is necessary for stylistically adequate, accurate translation, as opposed to deep textual understanding. Thus the domain model simply represents all the concepts relevant in the domain, but does not support any further reasoning or inference about the concepts in the domain, other than that which is directly encoded (e.g. hierarchical information such as the fact that personal computers and mainframes are types of computer). The essential role of the domain model is to support full disambiguation of the text. An important part of this is specifying, for every event concept in the domain, what restrictions it places on the object concepts which constitute its arguments (e.g. only living things can die, only humans can think, in a literal sense) or the 'fillers' of 'slots' in its (frame-based) representation.

Once you start adding detailed knowledge in the pursuit of high quality translation through text understanding, it is tempting to add more and more sources of knowledge. It is quite clear that anaphora resolution and the resolution of other referential ambiguities requires reference to a level of structure above sentential syntax and semantics (see e.g. the examples in Chapter 6). Likewise, for stylistic reasons, to increase the cohesiveness of the text, one might need to keep some working representation of the paragraph structure. Achieving a really high quality translation, especially with some sorts of text, might require treatment of metaphor, metonymy, indirect speech acts, speaker/hearer attitudes and so on. Over the last few years a variety of groups in different parts of the world have begun experimenting with prototypes intended to work with explicit knowledge or rule components dealing with a wide variety of different types of information. All of these approaches can be viewed as examples, of one form or another, of knowledge-based MT.

10.2.3 Feasibility of General Purpose Rule-Based MT Systems

The approaches to MT that we have discussed so far in this chapter can be distinguished from each other mainly in terms of the various knowledge sources which are used in translation. They are all straightforward rule-based approaches, as most work in MT has been until the last few years. However it is widely recognised that there are serious challenges in building a robust, general purpose, high quality rule-based MT system, given the current state of linguistic knowledge. As we shall see, these problems and the increasing availability of raw materials in the form of on-line dictionaries, termbanks and corpus resources have led to a number of new developments in recent years which rely on empirical methods of various sorts, seeking to minimize or at least make more tractable the linguistic knowledge engineering problem.

One of the most serious problems, and probably *the* most serious problem, for linguistic knowledge MT is the development of appropriate large-scale grammatical and lexical resources. There are really a number of closely related problems here. The first is simply the scale of the undertaking, in terms of numbers of linguistic rules and lexical entries needed for fully automatic, high quality MT for general purpose and specialised language usage. Even assuming that our current state of linguistic knowledge is sophisticated enough, the effort involved is awesome, if all such information must be manually coded. It is generally accepted, then, that techniques must be adopted which favour the introduction of semi-automatic and automatic acquisition of linguistic knowledge.

The second concerns the difficulties of manipulating and managing such knowledge within a working system. The experience of linguists developing a wide variety of natural language processing systems shows that it is all too easy to add ad hoc, specially crafted rules to deal with problem cases, with the result that the system soon becomes difficult to understand, upgrade and maintain. In the worst case, the addition of a new rule to bring about some intended improvement, may cause the entire edifice to topple and performance to degrade. To a certain extent, these familiar problems can be avoided by adopting up to date formalisms, and restricting the use of special devices as much as possible. It is also very important to do everything possible to ensure that different grammar writers adopt essentially the same or consistent approaches and document everything they do in detail.

The third issue is one of quality and concerns the level of linguistic detail required to make the various discriminations which are necessary to ensure high quality output, at least for general texts. This problem shows up in a number of different areas, most notably in discriminating between different senses of a word, but also in relating pronouns to their antecedents.

Some consider that this third aspect is so serious as to effectively undermine the possibility of building large scale robust general purpose MT systems with a reasonably high quality output, arguing that given the current state of our understanding of (especially) sense differences, we are at the limits of what is possible for the time being in terms of the explicit encoding of linguistic distinctions. An extremely radical approach to this problem is to try to do away with explicitly formulated linguistic knowledge completely. This extreme form of the 'empirical' approach to MT is found in the work carried out by an MT group at IBM Yorktown Heights and will be discussed in the section below on Statistical Approaches.

One interesting development is now evident which receives its impetus from the appreciation of the difficulty and costliness of linguistic knowledge engineering. This is the growth of research into the reusability of resources (from application to application and from project to project) and the eventual development of standards for common resources. One of the reasons why this is happening now is that there is undoubtedly a set of core techniques and approaches which are widely known and accepted within the Natural Language Processing research community. In this sense a partial consensus is emerging on the treatment of some linguistic phenomena. A second important motivation is a growing appreciation of the fact that sharing tools, techniques and the grammatical and lexical resources between projects, for the areas where there is a consensus, allows one to direct research more appropriately at those issues which pose challenges.

As well as the various difficulties in developing linguistic resources, there are other issues which must be addressed in the development of a working MT system. If a system is to be used on free text, then it must be robust. That is, it must have mechanisms for dealing with unknown words and ill-formed output (simply answering 'no' and refusing to proceed would not be cooperative behaviour). In a similar way, it must have a way of dealing with unresolved ambiguities, that is, cases in which the grammar rules, in the light of all available information, still permit a number of different analyses. This is likely to happen

in terms of both lexical choice (for example, where there are a number of alternatives for a given word in translation) and structural choice. For example, taken in isolation (and in all likelihood, even in many contexts) the following string is ambiguous as shown:

(6) a. Sam told Kim that Jo had died last week.
b. Sam told Kim [that Jo had died] last week.
c. Sam told Kim [that Jo had died last week].

Such attachment ambiguities with adverbial phrases (such as *last week*) and prepositional phrases (*on Tuesday*) occur quite frequently in a language like English in which PPs and ADVPs typically occur at the end of phrases. In many cases, they are strictly structurally ambiguous, but can be disambiguated in context by the hearer by using real-word knowledge. For example, the following *is* ambiguous, but the hearer of such a sentence would have enough shared knowledge with the speaker to chose the intended interpretation (and perhaps would not even be aware of the ambiguity):

(7) a. Joe bought the book that I had been trying to obtain for Susan.
b. [Joe bought [the book that I had been trying to obtain for Susan]].
c. [Joe bought [the book that I had been trying to obtain] for Susan].

Consideration of issues such as these underlies work in integrating core MT engines with spelling checkers, fail-safe routines for what to do when a word in the input is not in the dictionary and adding preference mechanisms which chose an analysis in cases of true ambiguity, but an appreciation of the serious nature of these issues has also provided an motivation for the current interest in empirical, corpus or statistical-based MT, to which we return after discussing the question of resources for MT.

10.3 Resources for MT

As researchers begin to consider the implications of developing their systems beyond the level of proof-of-concept research prototypes with very restricted coverage, considerable attention is being paid to the role that existing bilingual and monolingual corpus and lexical resources can play. A corpus is essentially a large collection of texts, but for our purposes we are interested only in such texts stored on computers in a standard format (e.g. extended ASCII). Such texts may often contain standard markup (e.g. in SGML) and for most practical purposes one needs a set of corpus access tools for retrieving data at will.

Various research centres throughout the world have been developing monolingual corpus resources for many years, and there has been a growing awareness throughout the eighties of their importance to linguistic and lexicographic work. A number of sites hold substantial corpus resources (several millions of words), an example being the Unit for Computer Research on the English Language at the University of Lancaster which currently holds in excess of 5 million words of corpus material, of which 4M words have been tagged with part-of-speech information. Such collections are a rich repository of information about actual language usage. Efforts are underway at different centres to (automatically or semi-

automatically) annotate corpus resources with various types of linguistic information, in addition to grammatical (POS) tagging, prosodic annotation (indicating features of stress and annotation), syntactic tagging (indicating phrasal groups of words, i.e. parsing or partial (skeleton) parsing); semantic tagging and discourse level tagging (indicating anaphoric and other similar links). To give some idea of scale, the planned British National Corpus will contain around 100M words of grammatically tagged corpus material, with standard SGML markup. The following example text has been tagged with the CLAWS tagset developed at UCREL, University of Lancaster — in cases where multiple tags are possible, the tag chosen by the probabilistic tagger is shown in square brackets, with the alternatives following after commas.

Excerpt from a Tagged Corpus

Satellite_[JJ], NN1 communications_NN2 have_VH0 been_VBN used_[VVN], VVD, JJ for_[IF], CF, RP almost_RR two_MC decades_NNT2 to_TO provide_VVI intercontinental_[JJ], NN1 traffic_[NN1], VV0 through_[II], RP, JB the_AT INTELSAT_[NNJ], VV0, NN1 ,_, INTERSPUTNIK_[NN1], NNJ and_CC INMARSAT_[VV0], NN1,NNJ systems_NN2 ._. INTELSAT_VVC, now_[RT], CS also_RR provides_VVZ regional_JJ traffic_[NN1], VV and_CC leases_[NN2], VVZ transponders_[VVZ], NN2 to_[II], TO, RP several_DA2 countries_NNL2 for_[IF], CF, RP domestic_[JJ], NN1 use_[NN1], VV0 ._.

These tags, which it must be stressed are assigned completely automatically and with a high level of accuracy, provide a detailed parts of speech analysis of the text, distinguishing between some 40 different subcategories of Noun (the tags for Nouns begin with N for Noun or P for pronoun) and some 30 different subcategories of Verb, and so on.

Over the last few years there has been an increasing awareness of the importance of corpus resources in MT research. Tools for extracting information automatically from texts are being increasingly used, and new techniques developed. At the simplest level, a monolingual corpus is a crucial tool for the linguist in determining language usage in a given domain, and a bilingual corpus for determining the facts of translation. In developing MT systems, bilingual texts are an extremely important resource, and they are most useful if organized in such a way that the user can view translation 'chunks' or 'units'. In **bitext** (or 'multitext') the text is aligned so that within each bilingual (or multilingual) chunk the texts are translations of each other. The most common form of alignment takes the sentence to be the organizing unit for chunking and techniques exist for performing this alignment of bitext automatically with a high level of accuracy (96% or higher). Of course alignment does not need to stop at the sentence level and it is possible to apply simple probability measures to a sentence aligned bitext to extract automatically the most probable word pair alignments, and given some skeleton or phrasal parsing, to attempt to extract useful information about phrasal alignment. A caveat is of course in order — the success of

techniques such as probabilistic word pair alignment depends on the size and quality of the corpus resource, and minimum size is probably 2M words of clean text. The availability of bilingual or multilingual corpus resources of a decent size is currently a limiting factor. Despite the fact that many international institutions and companies have large bilingual or multilingual resources in appropriate formats, they have been slow to appreciate the value of releasing these to the research community, although there are indications that this situation is now changing (the Canadian English-French Hansard record of parliamentary proceedings is a notable exception, see the extract on page 187).

Much of the interest in corpus resources and machine-readable dictionaries comes not from their value as static knowledge banks, which the grammar writer can consult but in the possibilities of using the information they contain directly in the MT system, thus providing some solution to the knowledge acquisition problem we noted above. One way this can be achieved is by investigating procedures for automatically or semi-automatically deriving linguistic rules for the MT system from the various sources of information. Ideas currently under investigation include the use of monolingual corpus of sufficient size for automatic sense disambiguation in context.[2] As a further example, a part of speech tagged sentence aligned bilingual text together with some probabilistic model, could be used to automatically provide equivalent terms in the two languages which could then be automatically compiled into the relevant formalism for lexical entries in an MT system.

A further resource which is now beginning to be adequately exploited is the machine-readable dictionary (cf. Chapter 5). Monolingual lexical entries can be constructed semi-automatically from machine-readable dictionaries, and research is underway into semi-automatically deriving a bilingual lexicon from these monolingual lexica by statistical comparison of the lexical structures associated with various word senses. Another possibility is that of automatically deriving subcategorization and semantic selectional information for lexical entries and grammatical rules from corpus resources and machine-readable dictionaries. In all of these applications, the knowledge banks can be used to ease the formulation of large amounts of detailed linguistic information in a rule-based system. A number of other approaches, to which we now turn, attempt to use the information implicit in bilingual corpora, dictionaries and thesauri much more directly, as a component in the MT system.

10.4 Empirical Approaches to MT

Given the questions that have been raised about the feasibility of 'rule-based' approaches, the increasing availability of large amounts of machine readable textual material has been seen by a number of research groups as opening possibilities for rather different MT architectures — in particular, so called 'empirical' architectures which apply relatively 'low-level' statistical or pattern matching techniques either directly to texts, or to texts that have been subject to only rather superficial analysis. The reasoning behind the term empirical is that in such approaches, whatever linguistic knowledge the system uses is derived em-

[2]This may use the measure of Mutual Information, taking into account (roughly) the amount of mutual context elements share

Extract from Bilingual Hansard

French

Score 24 Que la Chambre blâme le gouvernement pour son inaction dans les dossiers de la grande région de Montréal, comprenant l' Agence spatiale, le développement du Vieux-Port, l' aménagement du Port, le projet Soligaz, les chantiers maritimes , la relance économique de l' est de Montréal, ainsi que la détérioration de l' économie du sud-ouest de la région.
Score 52 Monsieur le Président, je pense qu' il est important de rappeler pourquoi aujourd'hui, nous, du parti libéral, déposons une telle motion de blâme à l' endroit de ce gouvernement, après trois ans et demi de pouvoir, concernant les dossiers de Montréal, principal centre du Québec et aussi du Canada, un des principaux centres.
Score 8 Pourquoi il y a tant de dossiers pour qu' aujourd'hui on en arrive à une motion de blâme à l' endroit du gouvernement?
Score 86 Il est tout simplement important de se rappeler qu' après les élections de 1984, et suite à de multiple promesses faites par ce gouvernement à la population montréalaise, aux autorités municipales, aux gens de tout le Québec, dès 1985, malgré une représentation de 56 ou 57 députés, huit députés conservateurs sur l' île de Montréal, le milieu des affaires commence à se plaindre.

English

Score 24 That this House condemns the government for its failure to act in matters of interest to the region of Greater Montreal, including the space agency, the development of the Vieux-Port, the planning and development of Montreal Harbour, the Soligaz project, the shipyards and the economic renewal of East Montreal as well as the economic deterioration of the southwestern part of the region.
Score 52 He said : Mr. Speaker, I think it is important to recall why today, we in the Liberal Party move this motion to condemn a Government that has been in power for three and half years, a motion that concerns matters of interest to Montreal, the main urban centre of Quebec and one of the major urban centres in this country.
Score 8 Why has the number of issues outstanding increased to the point that today, we moved a motion condemning the Government?
Score 86 We must remember that after the election in 1984, following the many promises made by this Government to the people of Montreal, the municipal authorities and Quebecers as a whole, that in 1985, despite strong representation consisting of fifty-six or fifty-seven Members, including eight Conservative Members on Montreal Island, the business community started to complain.

pirically, by examination of real texts, rather than being reasoned out by linguists. We will look at two such approaches: the so called 'example' or 'analogy' based approach, and the 'statistical' approach.

10.4.1 Example-Based Translation

Throughout most of this book, we have assumed a model of the translation machine which involves explicit mapping rules of various sorts. In the 'translation by analogy', or 'example-based' approach, such mapping rules are dispensed with in favour of a procedure which involves matching against stored example translations. The basic idea is to collect a bilingual corpus of translation pairs and then use a best match algorithm to find the closest example to the source phrase in question. This gives a translation template, which can then be filled in by word-for-word translation.

This idea is sometimes thought to be reminiscent of how human translators proceed when using a bilingual dictionary: looking at the examples given to find the source language example that best approximates what they are trying to translate, and constructing a translation on the basis of the target language example that is given. For example, the bilingual dictionary entry for *printer* which we discussed in Chapter 5 gave the following as examples.

(8) a. ∼**'s error** faute *f* d'impression, coquille *f*;
b. ∼**'s reader** correcteur *m*, -trice *f* (d'épreuves).

Given a sentence like (8) to translate, a human translator would certainly choose *faute d'impression* or *coquille* as the translation, on the basis that a mistake is much more like an error than it is like a reader.

(9) This seems to be a printer's mistake.

The distance calculation, to find the best match for the source phrase, can involve calculating the closeness of items in a hierarchy of terms and concepts provided by a thesaurus. To give a flavour of the idea, and the sort of problem it addresses, consider the problem of translating Japanese phrases of the form *A no B* (*no* is a particle indicating the relation between A and B) into English. Among the forms to choose from are *AB*, *A's B*, *B of A*, *B on A*, *B in A*, and *B for A*, cf Table 10.2 which gives English paraphrases of examples involving *no*, together with the correct translations for these different patterns. The problem is certainly not an esoteric one, since the expression is claimed to occur in around 50% of Japanese sentences.

For a given input, the system will then calculate how close it is to various stored example translations based on the distance of the input from the example in terms of the thesaurus hierarchy (this involves finding the 'Most Specific Common Abstraction' for the input and the alternative translations — i.e. 'closest' concept in the thesaurus hierarchy) and how 'likely' the various translations are on the basis of frequency ratings for elements in the database of examples. (Notice this means we assume that the database of examples is

B of A	8th *no* afternoon	the afternoon of the 8th
B for A	conference *no* application fee	the application fee for the conference
B in A	Kyoto *no* conference	the conference in Kyoto
A's B	a week *no* holiday	a week's holiday
AB	hotel *no* reservation	the hotel reservation
AB	three *no* hotel	three hotels

Table 10.2 Alternative Translations for the Particle *no*

representative of the texts we intend to translate.)

The following is an extension to this basic idea: pairs of equivalent source and target language expression are given, along with example translations, written in parentheses, and interpreted as stating 'conditions' under which the given equivalence holds. For example, the rule for the Japanese word *sochira* ('this', or 'this person' — i.e. the addressee, *you*), given below, indicates that *sochira* translates as *this* when the example involves *desu*, (translating as *be*), and as *you*, when the input involves something like *okuru* (translating as *send*). In translating an input like *sochira ni tsutaeru*, the English pronoun *you* would be selected as the translation of *sochira*, because *tsutaeru* (convey) is closest to *okuru* (send) in the thesaurus.

```
sochira
→
this (( desu {be}),...)
you (( okuru {send}),...)
this (( miru {see}),...)
```

This rule uses only information about the surrounding string, but one could imagine other sorts of example, where information is given in terms of patterns of strings, or of grammatical information. An example involving string patterns is given below, which would be involved in translating examples involving the expression *o-negaishimasu* along the lines of (9) (*o-negaishimasu* ('please') is a general expression indicating that a request is being made, or a favour requested, *o* indicates that the preceding noun phrase is an OBJECT).

(10) a. jinjika o o-negaishimasu.
personnel section OBJ please
May I speak to the personnel section?

b. daimei o o-negaishimasu.
title OBJ please
Please give me the title.

To deal with this, rules like the following use information about surrounding string *patterns*:

```
X o o-negaishimasu
→
May I speak to X' ((jimukyoku {office}),...)
Please give me X' ((bangou {number}),...)
```

It should be evident that the feasibility of the approach depends crucially on the collection of good data. However, one of the advantages of the approach is that the quality of translation will improve incrementally as the example set becomes more complete, without the need to update and improve detailed grammatical and lexical descriptions. Moreover, the approach can be (in principle) very efficient, since in the best case there is no complex rule application to perform — all one has to do is find the appropriate example and (sometimes) calculate distances. However, there are some complications. For example, one problem arises when one has a number of different examples each of which matches part of the string, but where the parts they match overlap, and/or do not cover the whole string. In such cases, calculating the best match can involve considering a large number of possibilities.

A pure example-based approach would use no grammar rules at all, only example phrases. However, one could also imagine a role for some normal linguistic analysis, producing a standard linguistic representation. If, instead of being given in simple 'string' form, examples were stated in terms of such representations (i.e. given as fragments of linguistic representations), one would expect to be able to deal with many more variations in sentence pattern, and allow for a certain amount of restructuring in generation. In this way, one would have something that looked more like a standard LK architecture. The chief difference would be in the level of specificity of the rules. In particular, where in a traditional transfer system the rules are stated in as general a form as possible, to cover entire classes of case, what one would have here is a system where the rules are stated in highly particular forms (each one for essentially one case), but there is a general procedure for estimating, for each case, which rule is most appropriate (i.e. by estimating which example is closest). Of course, what this suggests is that there is no radical incompatibility between example-based, and rule-based approaches, so that the real challenge lies in finding the best combination of techniques from each. Here one obvious possibility is to use traditional rule-based transfer as a fall back, to be used only if there is no complete example-based translation.

10.4.2 Statistical MT

Over the last few years there has been a growing interest in the research community in statistical approaches to Natural Language Processing. With respect to MT, the term 'statistical approaches' can be understood in a narrow sense to refer to approaches which try to do away with explicitly formulating linguistic knowledge, or in a broad sense to denote the application of statistically or probablistically based techniques to parts of the MT task

(e.g. as a word sense disambiguation component). We will give a flavour of this work by describing a pure statistical-based approach to MT.

The approach can be thought of as trying to apply to MT techniques which have been highly successful in Speech Recognition, and though the details require a reasonable amount of statistical sophistication, the basic idea can be grasped quite simply. The two key notions involved are those of the **language model** and the **translation model**. The language model provides us with probabilities for strings of words (in fact sentences), which we can denote by $\texttt{Pr}(\texttt{S})$ (for a source sentence S) and $\texttt{Pr}(\texttt{T})$ (for any given target sentence T). Intuitively, $\texttt{Pr}(\texttt{S})$ is the probability of a string of source words S occurring, and likewise for $\texttt{Pr}(\texttt{T})$. The translation model also provides us with probabilities — $\texttt{Pr}(\texttt{T}|\texttt{S})$ is the conditional probability that a target sentence T will occur in a target text which translates a text containing the source sentence S. The product of this and the probability of S itself, that is $\texttt{Pr}(\texttt{S}) \times \texttt{Pr}(\texttt{T}|\texttt{S})$ gives the the probability of source-target pairs of sentences occurring, written $\texttt{Pr}(\texttt{S},\texttt{T})$.

One task, then, is to find out the probability of a source string (or sentence) occurring (i.e. $\texttt{Pr}(\texttt{S})$). This can be decomposed into the probability of the first word, multiplied by the conditional probabilities of the succeeding words, as follows.

$$\texttt{Pr}(\texttt{s1}) \times \texttt{Pr}(\texttt{s2}|\texttt{s1}) \times \texttt{Pr}(\texttt{s3}|\texttt{s1},\texttt{s2}),\ \texttt{etc...}$$

Intuitively, the conditional probability $\texttt{Pr}(\texttt{s2}|\texttt{s1})$ is the probability that s2 will occur, given that s1 has occurred; for example, the probability that *am* and *are* occur in a text might be approximately the same, but the probability of *am* occurring after *I* is quite high, while that of *are* is much lower). To keep things within manageable limits, it is common practice to take into account only the preceding one or two words in calculating these conditional probabilities (these are known respectively as 'bigram' and 'trigram' models). In order to calculate these source language probabilities (producing the source language model by estimating the parameters), a large amount of monolingual data is required, since of course the validity, usefulness or accuracy of the model will depend mainly on the size of the corpus.

The second task requiring large amounts of data is specifying the parameters of the translation model, which requires a large bilingual aligned corpus. As we observed above, there are rather few such resources, however, the research group at IBM which has been mainly responsible for developing this approach had access to three million sentence pairs from the Canadian (French-English) Hansard — the official record of proceedings in the Canadian Parliament (cf. the extract given above), from which they have developed a (sentence-) aligned corpus, where each source sentence is paired with its translation in the target language, as can be seen on page 192.

It is worth noting in passing that the usefulness of corpus resources depends very much on the state in which they are available to the researcher. Corpus clean-up and especially the correction of errors is a time-consuming and expensive business, and some would

argue that it detracts from the 'purity' of the data. But the extract given here illustrates a potential source of problems if a corpus is not cleaned up in some ways — the penultimate French sentence contains a false start, followed by . . . , while the English text (presumably produced by a human translator) contains just a complete sentence. This sort of divergence could in principle effect the statistics for word-level alignment.

In order to get some idea of how the translation model works, it is useful to introduce some further notions. In a word-aligned sentence-pair, it is indicated which target words correspond to each source word. An example of this (which takes French as the source language) is given in the second extract.

A Sentence-Aligned Corpus

Often, in the textile industry, businesses close their plant in Montreal to move to the Eastern Townships.
Dans le domaine du textile souvent, dans Montréal, on ferme et on va s' installer dans les Cantons de l' Est.

There is no legislation to prevent them from doing so, for it is a matter of internal economy.
Il n' y a aucune loi pour empêcher cela, c' est de la régie interne.

But then, in the case of the Gulf refinery it is different : first of all, the Federal Government asked Petro-Canada to buy everything, except in Quebec.
Mais là, la différence entre la Gulf... c' est différent parce que la vente de la raffinerie Gulf: premièrement, le gouvernement fédéral a demandé à Petro-Canada de tout acheter, sauf le Québec.

That is serious.
C'est grave.

Word Aligned Corpus

The Federal Government asked Petro-Canada to buy everything.
Le(1) gouvernement(3) fédéral(2) a demandé(4) à Petro-Canada(5) de(6) tout(8) acheter(7).

The numbers after the source words indicate the string position of the corresponding target word or words. If there is no target correspondence, then no bracketted numbers appear after the source word (e.g. *a* in *a demandé*). If more than one word in the target corre-

sponds, then this is also indicated. The **fertility** of a source word is the number of words corresponding to it in the target string. For example, the fertility of *asked* with English as source language is 2, since it aligns with *a demandé*. A third notion is that of **distortion** which refers to the fact that source words and their target correspondences do not necessarily appear in the same string position (compare *tout acheter* and *buy everything*, for example).

The parameters which must be calculated from the bilingual sentence aligned corpus are then (i) the fertility probabilities for each source word (i.e. the likelihood of it translating as one, two, three, etc, words respectively), (ii) the word-pair or translation possibilities for each word in each language and (iii) the set of distortion probabilities for each source and target position. With this information (which is extracted automatically from the corpus), the translation model can, for a given S, calculate $\mathtt{Pr}(\mathtt{T}|\mathtt{S})$ (that is, the probability of T, given S). This is the essence of the approach to statistically-based MT, although the procedure is itself slightly more complicated in involving search through possible source language sentences for the one which maximises $\mathtt{Pr}(\mathtt{S}) \times \mathtt{Pr}(\mathtt{T}|\mathtt{S})$, translation being essentially viewed as the problem of finding the S that is most probable given T — i.e. one wants to maximise $\mathtt{Pr}(\mathtt{S}|\mathtt{T})$. Given that

$$\mathtt{Pr}(\mathtt{S}|\mathtt{T}) = \frac{\mathtt{Pr}(\mathtt{S})\mathtt{Pr}(\mathtt{T}|\mathtt{S})}{\mathtt{Pr}(\mathtt{T})}$$

then one just needs to choose S that maximizes the product of $\mathtt{Pr}(\mathtt{S})$ and $\mathtt{Pr}(\mathtt{T}|\mathtt{S})$.

It should be clear that in an approach such as this there is no role whatsoever for the explicit encoding of linguistic information, and thus the knowledge acquisition problem is solved. On the other hand, the general applicability of the method might be doubted, since as we observed above, it is heavily dependent on the availability of good quality bilingual or multilingual data in very large proportions, something which is currently lacking for most languages.

Results to date in terms of accuracy have not been overly impressive, with a 39% rate of correct translation reported on a set of 100 short test sentences. A defect of this approach is that morphologically related words are treated as completely separate from each other, so that, for example, distributional information about *sees* cannot contribute to the calculation of parameters for *see* and *saw*, etc. In an attempt to remedy this defect, researchers at IBM have started to add low level grammatical information piecemeal to their system, moving in essence towards an analysis-transfer-synthesis model of statistically-based translation. The information in question includes morphological information, the neutralisation of case distinctions (upper and lower case) and minor transformations to input sentences (such as the movement of adverbs) to create a more canonical form. The currently reported success rate with 100 test sentences is a quite respectable 60%. A major criticism of this move is of course precisely that linguistic information is being added piecemeal, without a real view of its appropriacy or completeness, and there must be serious doubts about how far the approach can be extended without further additions of explicit linguistic knowledge, i.e. a more systematic notion of grammar. Putting the matter more positively, it seems clear that

there is a useful role for information about probabilities. However, the poor success rate for the 'pure' approach without any linguistic knowledge (less than 40%) suggests that the real question is how one can best combine statistical and rule-based approaches.

10.5 Summary

We have tried in this chapter to give a brief overview of some of the issues and techniques which are being actively researched today in MT. Of course, there is not enough room in one chapter to do justice to the field, and we have of necessity omitted much work that is of interest. In particular, we have restricted our discussion to MT itself and have said nothing at all about recent work in the development of translators aids, multilingual authoring packages and terminological systems of various sorts. Nonetheless we have identified three important trends in current research in MT. The first is the exploitation of current techniques from computational linguistics to permit a multidimensional view of the translational relation between two texts. The second in the increasing orientation of the research community towards the use of existing resources of various sorts, either to extract useful information or directly as components in systems. The third, related, trend is towards statistical or empirical models of translation. Though we have dwelt in some detail in this short survey on 'pure' statistical and simple pattern matching methods, in fact much recent work advocates a mixture of techniques, for example with statistical methods supplementing rule-based methods in various ways.

10.6 Further Reading

Our discussion of flexible translation between Swedish and Norwegian is based on unpublished work by Dyvik (1992). The standard references on sign-based approaches to linguistic representation are Pollard and Sag (1987, 1993). The view of constraint based translation that we describe is loosely modelled on that used in 'Shake and Bake' Whitelock (1992); Beaven (1992). See Kaplan et al. (1989), Sadler (1991) and Sadler (1993) for a slightly different approach. General discussion of how multi-dimensional representations can be used in MT can be found in Sadler and Arnold (1993).

On knowledge-based MT see Goodman and Nirenburg (1991), and the special issue of the journal *Machine Translation*, Goodman (1989).

On the processing of corpora, and their use in linguistics generally, see Garside et al. (1987), and Aijmer and Altenberg (1991).

The idea of example-based MT was first discussed in a paper by Nagao Nagao (1984). For a review of more recent work along these lines, see Somers (1992).

The pure statistical approach to MT is based on the work of a team at IBM, see for example Brown et al. (1990). As regards aligned, bilingual corpora, the most common form of alignment takes the sentence to be the organizing unit for chunking, see Brown et al. (1991) and Gale and Church (1991b) for relevant discussion. On automatic extraction of

word correspondences across bitext, see Gale and Church (1991a). Techniques involving the use of corpus resources for automatic sense disambiguation have also been explored within the DLT project, see Sadler (1989).

The translation of *no*, which was described around page 188 above, is discussed by Sumita et al. (1990). The discussion of *o-negaishimasu* is from Furuse and Iida (1992b), see also Furuse and Iida (1992a), and Sumita and Iida (1991).

The frame for `computer` on page 181 above is taken from (Goodman and Nirenburg, 1991, page 25).

For up to date reports on research in the field of MT, there are several journals, and several major international conferences. The specialist Journal is *Machine Translation*, edited by Sergei Nirenburg, from Carnegie Mellon University in Pittsburg, USA, and published by Kluwer Academic Publishers. However, the journal *Computational Linguistics*, published by the MIT Press for the Association for Computational Linguistics (ACL), also publishes research which is directly about MT.

The specialist conference for research on MT is called *TMI* — for 'Theoretical and Methodological Issues (in Machine Translation)'. This has been held every two years since 1986, and proceedings are published (TMI1,TMI2TMI3,TMI4). Many of the papers in the last of these are directly or indirectly about the issue of 'rationalist' (i.e. rule-based) vs. empirical approaches to MT. The proceedings of the main Computational Linguistics conferences, namely (*COLING*), the conferences of the Association for Computational Linguistics (ACL) and the conferences of the European Chapters of the ACL, also contain a high percentage of papers about MT. ACL conferences are held annually in the USA (for example, ACL28; ACL29; ACL30). The EACL conferences are held biennially, EACL1; EACL2; EACL3; EACL4; EACL5, as is COLING: *Coling 84* Coling84 was held in Stanford, California, *COLING 86* Coling86 in Bonn, *Coling 88* Coling88in Budapest, *Coling 90* Coling90 in Helsinki, and *Coling 92* Coling92 was held in Nantes.