18
Some other systems and directions of research

In this chapter we attempt to indicate major current lines of MT research and to predict some future directions. In doing so we shall be describing briefly some MT projects which have not been given the full treatment dedicated to systems in the last eight chapters. This chapter will necessarily be more suggestive and speculative than preceding ones and readers should regard it as a gateway into the somewhat confusing picture of contemporary MT research activity.

18.1 AI and Knowledge-based MT at CMU

For many observers of MT development it has been the conventional wisdom that the most likely source of techniques for improving MT quality is the research on natural language processing within the context of **Artificial Intelligence** (AI). The involvement of AI researchers in MT-related projects began in the early 1970s with Yorick Wilks' work at Stanford University and the research of Roger Schank and his colleagues at Yale University. This was after the ALPAC report had highlighted the inadequacies of current approaches to MT. A major deficiency, obvious to many at the time, was their impotence in face of what was called the 'semantic barrier'.

The basic justification for AI approaches is the argument that since translation is concerned primarily with conveying the content or 'meaning' of a text in one language into a text in another language any MT system must be able to

'understand' the meanings of texts, as Bar-Hillel, Yngve and others were arguing already in the early 1960s. Without understanding, it is contended, no system can be expected to be able to decide which of possible target language expressions correspond most closely to the meaning of the original text. AI research claims to tackle this problem directly and is thus seen as likely to improve the quality of MT output. Characteristic of AI approaches is the adoption of primarily semantics-oriented parsing, the interpretation of texts by reference to knowledge bases and the use of inference mechanisms, and language-independent representations of the 'meaning' of texts.

The 1980s saw continued and increasing activity in research on AI approaches to translation, in Europe (some in relation to the Eurotra project), in Japan (notably at the Electro-Technical Laboratory), and in particular in North America. Much of this AI-inspired research has been on a small-scale, but a major centre has for some years been located at the Carnegie-Mellon University (CMU) in Pittsburgh.

The research at the CMU Center for Machine Translation under Jaime Carbonell and Sergei Nirenburg continues work which began initially in 1983 at Colgate University. The experimental systems are based on a methodology described as 'meaning-oriented MT in an interlingua paradigm'. Most attention is paid to the creation of appropriate and efficient software and the acquisition of the knowledge bases, giving the research theme its name **Knowledge-based MT** (KBMT). The systems developed are seen as gradual approximations of an ideal interlingua-based MT system. Some parts of the system are relatively complete, others are still experimental, including domain knowledge and lexicons and many areas of linguistic processing.

The working prototype for English and Japanese in both directions is designed for translation of personal computer manuals. It has a small 'domain model' of 1,500 concepts, and analysis and generation lexicons for both languages, each of nearly 900 items. The system is written in CommonLisp, and the grammar formalism is based on Lexical Functional Grammar (LFG). The basic modules are (Figure 18.1): syntactic parser with semantic constraints, a semantic mapper (for semantic interpretation), an interactive 'augmentor' for remaining ambiguities, a semantic generator producing syntactic structures with lexical selection, and a syntactic generator for producing target strings. The language-specific databases are analysis and generation grammars, and analysis and generation lexicons providing syntactic information. The concept lexicon and the semantic information in the analysis and generation lexicons (i.e. defining some semantic constraints) are language-independent but specific to the subject domain. The mapping rules, which convert f-structures into interlingua texts are both language- and domain-dependent. The CMU system is supported by software for creating concept lexicons (the 'knowledge acquisition tool' ONTOS), for compiling grammars and for testing modules and components.

The Analyzer consists of two components, a syntactic parser and a semantic interpreter, the 'mapping rule interpreter'. The syntactic parser uses an LFG-type grammar and produces an LFG-type 'f-structure' (see section 2.10.1). For example, the sentence (1) is represented as in (2).

(1) Remove the diskette from the drive.

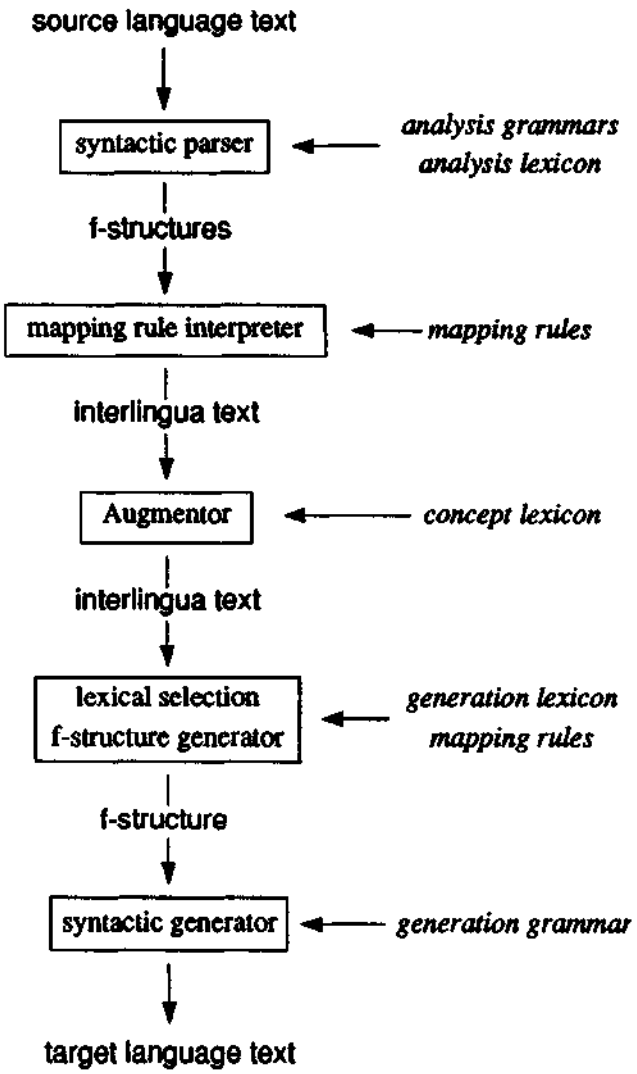


Figure 18.1 KBMT basic modules

(2)
```
((OBJ ((CASE ACC) (REF DEFINITE)
       (DET ((ROOT THE) (REF DEFINITE)))
       (ROOT DISKETTE) (PERSON 3) (NUMBER SINGULAR)
       (COUNT YES) (PROPER NO))
  (PPADJUNCT ((PREP FROM) (REF DEFINITE)
       (DET ((ROOT THE) (REF DEFINITE)))
       (ROOT DRIVE) (PERSON 3) (NUMBER SINGULAR)
       (COUNT YES) (PROPER NO)))
  (VALENCY TRANS)
  (MOOD IMPERATIVE)
```

```
(TENSE PRESENT)
(FORM INF)
(COMP-TYPE NO)
(ROOT REMOVE))
```

The semantic interpreter tests any f-structure component for ambiguities and substitutes interlingual units for source language lexical items and constructions, e.g. surface subject–predicate structures are replaced by case frames ('agent', 'theme', 'experiencer', etc.)

The central core of the system is the representation of interlingua texts. These are representations of 'actual events' as reported in the source texts. They are in the form of networks of propositions, i.e. events or states with their arguments and causal, temporal, spatial, etc. links to other events or states. The representations are produced as instantiations of concepts (events, individuals, etc.) from the 'concept lexicon'. The latter is a database of knowledge about the events and entities in the domain (subject field of the system). There is thus a clear distinction between the static knowledge database (network of relations independent of particular texts) and the dynamic interlingua text representations. It is argued that to ensure adequate understanding of input texts, the knowledge required must go beyond propositional knowledge, it must cover pragmatic and discourse meaning, i.e. the attitudes of speakers and hearers to the propositions expressed, speech acts, thematic structures, and the ways in which separate utterances are combined as coherent texts (cf. section 2.7).

The interlingua text is represented as a non-linear network of frames with slot values. Typically the values correspond to those listed in the concept lexicon; e.g. a textual *rose* will have as its 'color' value one of the list '(white red yellow blue ...)' attached to the concept 'flower' (of which *rose* is identified as an instance). The representation of non-propositional information from input texts is regarded as an innovative feature of the CMU project. The same knowledge structure is used in the concept lexicon and in interlingua texts as for propositional information. Thus, the lexicon indicates the potential values for speech acts ('statement', 'definition', 'request-info', 'request-action'), one of which is relevant for a particular sentence (or clause) in a text. Similarly, the lexicon offers sets of values for markers of text cohesion: 'expansion', 'similar', 'generalization', 'contrastive', 'digression', etc.

The result of a mapping rule application on the f-structure above (2) is a candidate interlingua text (3).

(3)
```
[*REMOVE
      (THEME (*DISKETTE (NUMBER SINGULAR)
                        (REFERENCE DEFINITE)))
      (SOURCE (*DISKETTE-DRIVE (NUMBER SINGULAR)
                        (REFERENCE DEFINITE)))
      (TENSE PRESENT)
      (MOOD IMPERATIVE)]
```

As the example shows, the knowledge base has enabled the identification of *drive* as referring in this subject domain to a 'diskette drive'.

The task of the Augmentor is to produce a single unambiguous interlingua text for input to the Generator. Since the output from the Analyzer still reflects

to some extent the syntactic configurations of the source language it must first be reformatted by the Augmentor into a language-independent form. Secondly, the Augmentor has to disambiguate genuinely ambiguous candidate interlingua texts. For example, *tape* might refer to 'adhesive tape' or 'magnetic tape' in this domain, and the semantic processing of the Analyzer is unable to resolve the ambiguity in a sentence such as (4).

(4) Remove the tape from the diskette drive.

It is here that the 'knowledge' of the subject matter as embodied in the concept lexicon is to be called upon. As yet, however, the CMU project has automated only one part of the Augmentor's disambiguation operations, namely the task of identifying referents of pronominal anaphors across sentences; consequently, most of the Augmentor tasks are at present performed interactively with users.

The Generator produces first a target language f-structure by the selection of lexical items and the application of mapping rules similar to those in the Analyzer; and then produces a surface structure and output text. Whereas the result of analysis is multiple output of possible interpretations, generation stops as soon as one valid target string has been produced (i.e. the first and not necessarily either the 'best' or even one expressing the complete input text). The same grammars are used for both analysis and generation, but only in the case of Japanese is there some partial reversibility of rule application.

As this brief description indicates, the development of the CMU system is still at an early stage. Its importance lies in the investigation of MT via representations in a conceptual ('meaning') interlingua, specific to a domain but independent of particular languages. It remains to be seen whether the CMU project will confound the arguments of those sceptical about interlingual representations (section 6.8) or indeed about the need for 'understanding' of the AI-type in translation at all. It is pointed out that human translators often do not need to fully understand what they are translating (in fact it would be unusual if they understood scientific reports as well as the researchers themselves). A secondary objection that AI-type meaning-oriented systems produce 'paraphrases' rather than translations (because the surface presentation of the content in the source is lost) may be answered in the CMU project by the incorporation of textual (pragmatic) information in 'interlingua texts'. The main reservations, however, are practical: whether the construction of language-independent knowledge bases is feasible for other than highly restricted subject domains — although this is a question for any system designed for a particular sublanguage which is intended to be applicable in general — and whether the potential improvements in quality will be commensurate with the greatly increased computation involved.

18.2 Example-based MT at BSO

In section 6.9 we briefly described **example-based** methods as alternatives to knowledge-based approaches and as supplementary to traditional rule-based methods. Various researchers have been investigating their potential, including members of the Japanese ATR project (see section 18.6 below). To illustrate what is involved we describe a proposal from the DLT project to use a 'Bilingual

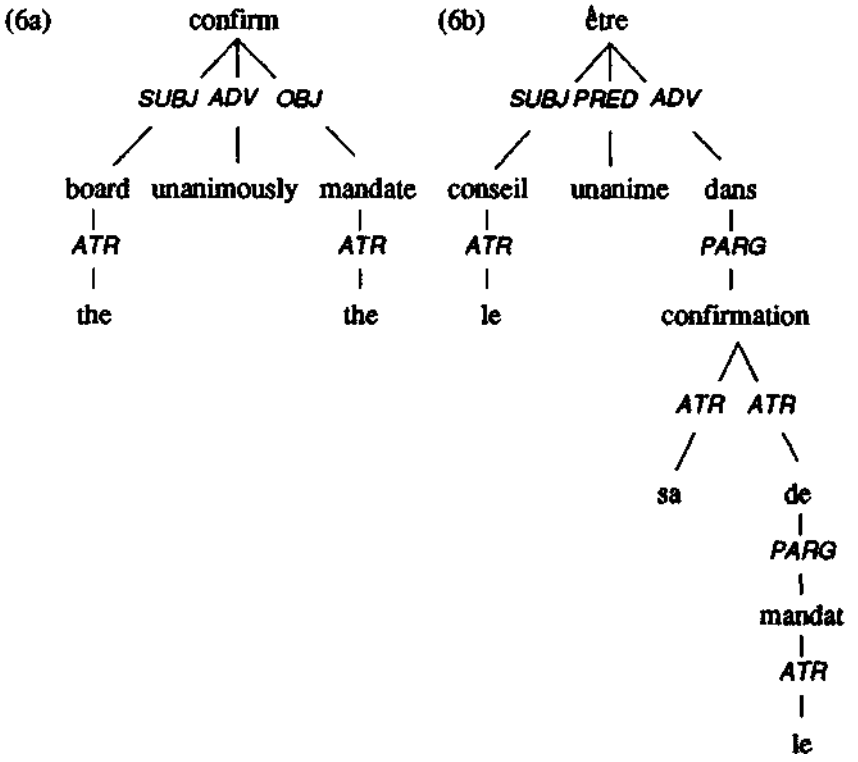
Knowledge Bank' (BKB) as a translation tool. The earlier DLT research is described in Chapter 17.

The purpose of the BKB is to serve as the primary source of linguistic 'knowledge' for all modules in the translation process. It consists of a corpus of equivalent texts in two languages which have been structurally analysed (by the same type of parser) into 'translation units' and which have been aligned to each other. Translation units are fragments of text in the two languages, which a translator would consider equivalent and mutually substitutable, and as far as possible structurally autonomous. Such aligned 'bitexts' have previously been proposed as aids for professional translators, giving them access to their own or others' previous practice in translation. The DLT researchers intend to use the structured bilingual corpus for the automatic resolution of source language ambiguities, problems of lexical and structural transfer and difficulties in target language selection.

As an illustration of alignment, the sentences (5a) and (5b) in an English–French BKB might be analysed, by a dependency parser, as (6a) and (6b).

(5a) The board unanimously confirms the mandate.

(5b) *Le conseil est unanime dans sa confirmation du mandat.*



The aligned translation units for the structure would be derived by rules relating bilingual structural equivalents (7).

(7) the board ↔ *le conseil*
the ↔ *le*
unanimously confirm ↔ *être unanime dans sa confirmation de*
unanimously ↔ *unanime*
the mandate ↔ *le mandat*

As a result of the full comparative analysis and alignment of translation units in a large bilingual corpus, the data in (8) might be available for English expressions of the form *have... effect on....*

(8) have a direct effect on ↔ *ont une influence directe à*
have a direct effect on ↔ *intéressent directement*
had a direct effect on ↔ *ont eu une répercussion directe sur*
has had a marked effect on ↔ *a largement influencé*
had a positive effect on ↔ *s'est avérée positive dans*
had a highly negative effect on ↔ *en auraient été gravement affectés*
will have a decisive effect on ↔ *influencera de façon déterminante*

The stages of transfer from source to target remain as in the earlier DLT model (Figure 17.1). The difference lies in the type of information applied at each stage. Modules operate on a common data structure, rather than passing on tree structures sequentially. The various databases (dictionaries and other knowledge bases) are integrated into the BKB. There are four mechanisms which operate interactively: Parser, Text Expert, Metataxor and Examiner. The Parser and Metataxor correspond roughly to the mechanisms for syntactic analysis and transformation described in sections 17.4 and 17.5, except that neither is now restricted to rule-based analyses and to unordered (uninterpreted) output; they can use frequency information in the BKB about which structural patterns are most likely, and the Metataxor can also assess potential target structures with evidence from aligned bilingual subtrees.

The Text Expert is a new module which is proposed to deal with referential relations across sentences. Given, for example, the sentence sequence (9), the Text Expert would seek potential antecedents of the pronoun *it*: in this case, either *translation* or *screen* but not *translator* because of semantic incompatibilities.

(9) The translator may see the translation on his screen. It will be displayed in a special format.

A decision is needed when translating into French or German: *il* or *er* would refer to the screen (*écran* or *Bildschirm*), *elle* or *sie* to the translation (*traduction* or *Übersetzung*).

It is the task of the Examiner (which replaces SWESIL of the earlier DLT model, section 17.6) to select the best analyses from the Parser, to choose the best transformations from the Metataxor and to resolve text–grammatical uncertainties from the Text Expert. It assesses the semantic and pragmatic plausibility of interpretations and proposed translations by reference to the BKB. Thus, to test whether in a particular instance, *around* is to be translated as *autour de* or as *vers*, *sharp* as *aigu*, *pointu*, *vif* or *aigre*, the Examiner would look for examples of similar contexts in either source or target language texts (or both). The procedure should

be able to tackle problems of lexical transfer often characterised as 'stylistic' (cf. section 7.4), such as the choice between *big* and *large* (as translations of French *grand* and German *gross*), the subtle differences between *fast, swift, rapid* and *quick* when translating French *rapide*, and so forth. The procedure potentially identifies implicit relations, e.g. that in *radiation protection* something is being protected against radiation, while in *wildlife protection* it is the wildlife which is being protected against something or someone. Relevant texts might include phrases such as: *steps taken to protect the inhabitants against radiation hazards* and *reservations for the protection of wild birds and animals* (plus other fragments linking *birds* and *animals* to *wildlife*). The approach could also in principle tackle the inter-sentential problems illustrated in (9), e.g. by finding examples linking *translation* and *text* and fragments such as *the text was displayed....* In all these cases, it should be stressed the examples could come from either texts in the source language or texts in the target language; the bilingual corpus as a whole represents a database of extra-linguistic knowledge, as well as a source of linguistic knowledge about the two languages concerned.

The advantages of the BKB approach are claimed to be: (a) the system is in principle reversible; the same procedures and textual information are applied in both translation directions; (b) no dictionaries and no knowledge bases have to be compiled with great expenditure of specialist effort: the dangers of inconsistency and insufficiency are avoided; (c) analysis can be as deep or as shallow as desired and can be adjusted with experience; with a full-text database no information is lost; (d) text corpora (and thus lexical coverage) can be selected to suit the needs of specific users; (e) decisions about homography and polysemy are superfluous; the 'sublanguages' of text corpora determine correspondences and differences between languages; (f) databases can be easily updated to deal with neologisms by adding new texts and by 'learning' from texts while the system is translating; (g) translation expertise is acquired by 'imitating' the best human translators, i.e. using the wealth of complex contextual information usually absent from dictionaries.

The concept of example-based MT is also found in the UMIST–ATR project, described below (18.5).

18.3 Statistics-based MT at IBM

Most of the systems and methods described in this book involve linguistic analysis and generation of sentences and texts. This is true not only for the traditional 'linguistics-based' systems (Systran, SUSY, Ariane, Eurotra, METAL) but also for the less typical Météo, Rosetta and DLT systems and for the knowledge-based and example-based approaches in the last two sections. A more radical departure from the assumptions which have dominated MT research since the early 1960s is represented by the investigation at the IBM Research Laboratories at Yorktown Heights, NY, of a translation system based almost exclusively on statistical techniques.

The use of statistical analysis was not uncommon in the early years of MT research. It was employed primarily as a tool for the automatic classification of linguistic data, on the assumption that contemporary knowledge about language

was insufficient for computational processing. This application has continued to the present time: many research projects make use of statistical data to guide rule writing and the formulation of routines; some go further, as we have seen (Chapter 17), and use statistical information in lexical selection and disambiguation. What is unique about the IBM project is the use of statistical techniques as the sole tool for analysis and generation. It has been made possible by the increasingly sophisticated application of statistics-based approaches in speech recognition and parsing.

The IBM research is based on a large corpus of the Canadian *Hansard*, which records parliamentary debates in both English and French. The corpus for the experiment was 40,000 pairs of sentences (totalling some 800,000 words) in each language. The essence of the method is the alignment of sentences in the two languages and the calculation of the probabilities that any one word in a sentence of one language corresponds to two, one or zero words in the translated sentence in the other language. Alignment was established by a technique widely used in speech recognition. The probabilities were estimated by matching bigrams (two consecutive words) in each English sentence against bigrams in 'equivalent' French sentences.

Two sets of probabilities were calculated. First, for each individual English word the probabilities of its correspondences to a set of French words; e.g. *the* corresponds to French *le* with probability .610, to *la* with .178, to *l'* with .083, to *les* with .023, to *ce* with .013, to *il* with .012, etc. Second, the probabilities that two, one or zero French words correspond to a single English word, e.g. *the* corresponds to one French word with .871 probability, to zero with .124, and to two with .004. In the case of *not* there is a .758 probability of correspondence with two French words, and these are most likely to be *ne* (.460) and *pas* (.469), with *plus* (.002) and *jamais* (.002) much less likely; other correspondences of *not* are *non* (.024), *pas du tout* (.003), and *faux* (.003), etc.

The effectiveness of the approach was evaluated by translating from French into English. The vocabulary was limited to the 1,000 most frequent English words and their corresponding most frequent 1,700 French words. The translation model was tested on 73 new French sentences from elsewhere in the *Hansard* corpus. The results were classified as: (a) exact (same as *Hansard* translation), (b) alternative (same meaning but in slightly different words), (c) different (legitimate translation but not conveying the same meaning as the *Hansard* translation), (d) wrong (intelligible result but not a translation of the French), and (e) ungrammatical (no sense conveyed). Some examples are shown in Table 18.1.

Although only 5% came into the 'exact' category, translations were considered 'reasonable' if they came into any of the first three categories (exact, alternate, different). On this criterion, the system performed with 48% success. Improvements are expected with a larger corpus (only 10% of the *Hansard* was used), by probabilistic segmentation of sentences into phrases, by using trigrams as well as bigrams, and by including data on inflectional morphology to group together, for example *tall, taller, tallest* and *va, vais, vont*.

The proposed incorporation of segmentation and morphology data suggests that statistical approaches have inherent limitations that even their advocates

exact	*Ces amendements sont certainement nécessaires*
Hansard	These amendments are certainly necessary
IBM	These amendments are certainly necessary
alternative	*C'est pourtant très simple*
Hansard	Yet it is very simple
IBM	It is still very simple
different	*J'ai reçu cette demande en effet*
Hansard	Such a request was made
IBM	I have received this request in effect
wrong	*Permettez que je donne un exemple à la Chambre*
Hansard	Let me give the House one example
IBM	Let me give an example in the House
ungrammatical	*Vous avez besoin de toute l'aide disponible*
Hansard	You need all the help you can get
IBM	You need of the whole benefits available

Table 18.1 Examples of translations

acknowledge. Nevertheless, the importance of this research is that it demonstrates how far it is possible to go in bilingual uni-directional MT without recourse to linguistic analysis. However, it must be remembered that the results are biased to a particular corpus. For example, in these texts the translation of *hear* would invariably be *bravo* (a probability of .992) — i.e. in the Canadian parliament *Hear, hear!* corresponds to *Bravo!* — while the 'normal' translation *entendre* has a very low probability of just .005.

There is little doubt that statistics-based techniques will be a feature of many future MT projects, although whether many will follow the exclusivity of the IBM team is uncertain. At the present time, the assumption is that linguistic data and methodology will remain at the centre of any practical MT system.

18.4 Sublanguage translation: Titus

The restriction of an MT system to one particular text corpus may be regarded as an extreme variant of the **sublanguage** approach (section 8.4). It is justifiable only if the corpus is particularly large. This was certainly the case with the aviation manuals for which the TAUM team were asked to develop an MT system. However, as we have already mentioned (Chapter 12), their success with the sublanguage system Météo could not be repeated.

In practice, nearly all MT systems, whether experimental or commercial, are limited to particular subject fields. This can be seen in many implementations of Systran, in the Pan American Health Organization systems (medical and public health documents), in the Eurotra prototype (information technology), in METAL (technical documents), in the CMU system (personal computer manuals),

and in many microcomputer-based systems from Japanese computer companies, which have concentrated on translations of computer technology and electronic engineering. Some further examples come later in this chapter. However, these systems are not classifiable as 'sublanguage systems' as such, because they are neither designed for nor intended to be restricted to particular subjects. Their present limitations are regarded as temporary: extension to other subject areas is anticipated. By contrast, sublanguage MT systems are developed specifically for one particular subject or text type, in order to minimise problems of homography, translational ambiguity and structural variety (section 8.4). The archetypal system, as we have seen, is Météo (Chapter 12); another example is TITUS.

The TITUS system was designed by the Institut Textile de France for the multilingual treatment of abstracts in an on-line database for the textile industry. First installed in 1970, the system is now in its fourth version. Abstracts are stored in representations from which texts can be generated in any one of four languages: English, French, German, and Spanish. Abstracts can be entered in any of the languages; they are formulated in a controlled syntax, in a controlled basic vocabulary, and in the standardised terminology of the textile industry sublanguage. The controlled syntax determines the order of basic phrase types: subject NP, circumstantial NP, verb phrase, complement NP, prepositional NP (where some of these NP types may be coordinate phrases, and the VP and the complement NP are optional). The syntax defines also the structure of noun phrases and of verb phrases, in terms of the order and optionality of constituents. At a pre-editing stage, words which are potentially ambiguous are distinguished, e.g. the French verb form *a* is distinguished from the preposition *a* (not entered with the grave accent) by a following slash. Furthermore, structural ambiguities (e.g. antecedents of prepositional phrases) are eliminated by the insertion of punctuation marks. Abstracts are entered interactively; the system requests reformulations of phrases and sentences not conforming to the controlled syntax and asks for clarification of homographs and structural ambiguities. Despite the constraints, the range of permitted structures is nevertheless impressive, for example (10).

(10) *L'analyse du fluage des fibres de polyéthylène après irridiation sous vide montre ; qu'un processus de pontage survient dans la région amorphe ; tandis qu'une coupure de la chaîne moléculaire de la région cristalline est observée.*

After entry, the system produces a version in the input language to check that analysis is correct, and then generates versions in each of the other languages. For example, the English entry corresponding to (10) would be (11).

(11) The polyethylene fibre creep analysis after irridiation under vacuum shows that a cross-linking process occurs in the amorphous region whereas a molecular chain scission of crystalline region is observed.

The quality of the translations is ensured by combining control of the input, extensive pre-editing, interactive feedback during and after analysis, and restriction to a regulated sublanguage. With such constraints, some may doubt whether TITUS should be considered a real MT system at all; what is clear is that it illustrates, as well as Météo, what can be achieved in well-defined environments.

It is perhaps surprising that there have not been more sublanguage systems. It does appear, however, that there must be very few situations where translation can be restricted to a relatively self-contained lexical and syntactic domain. Since Météo, researchers from the TAUM group have been searching for an ideal area for applying sublanguage methods, and at present some are working on MT for livestock market reports (see 18.7 below). As we shall see, other sublanguages under investigation include business correspondence, hotel and conference reservations and police communication.

18.5 MT for monolingual users

As discussed already, most MT systems are intended (explicitly or implicitly) for users knowing both source and target languages, who are able to make good the deficiencies of current MT systems in various ways (Chapter 9). However, the use of MT by monolinguals not familiar with one of the source or target languages is also possible. The output from batch systems, such as Systran, can be valuable unrevised to specialists expert in the subject field, but since these experts do not necessarily know the source language, they could not be expected to be of assistance in the translation process itself. But when we consider users at the other end of the process, at the input of text, there are a number of interesting possibilities, as we have already mentioned (section 8.3.4). Some of these are now under active investigation; in particular, systems for the interactive composition of texts to be translated into a language that the user does not know.

At UMIST (University of Manchester Institute of Science and Technology), pioneering work on MT systems for monolingual users was undertaken in the development of Ntran. This was an English–Japanese system which involved interactive disambiguation of the English source text and, where further ambiguities arose during lexical transfer, a stage of interaction involving choices that a user with no knowledge of Japanese could nevertheless be expected to make (i.e. based on English paraphrases of the target language distinctions). The Japanese generation was entirely automatic. Inspired by the relative success of this prototype system, researchers at UMIST have become involved in several further projects having as a common theme the development of MT for the monolingual user.

One such project, funded by British Telecom, aims at developing an MT system which will help users to compose business letters in an unfamiliar foreign language, by guiding them through a menu of choices. The system is based on the idea of 'pro-forma' texts corresponding to certain types of business letter, e.g. complaint, offer, enquiry, and so on. The pro-forma texts are templates with slots for set phrases, names, dates, addresses, etc., which have to be entered by users in their own language. Once the pro-forma is complete, the system is able to generate multilingual equivalents of the text by comparing the information provided by the user with a database of 'pre-translated' text fragments. The attraction of the system is that, because the target langauge templates are based on stylistically appropriate texts written by human translators (rather than built up by largely literal translation), high quality output can be guaranteed, as long as the text remains strictly within

the given domain. A similar approach is found in a system being developed in Malaysia which helps writers draft official letters in Malay language.

Another system, also being developed at UMIST, takes the idea of system–user interaction in place of a 'source text' a little further. In a project being carried out together with the Japanese ATR (Advanced Telecommunications Research) Laboratories, a **dialogue translation** system is being developed. The domain is that of an on-line conversation (ultimately, by telephone, but keyboard conversation is the medium) between a conference office in Japan and an English-speaking enquirer. The idea is that the system works as an intermediary between the two conversation partners, translating their dialogue between English and Japanese. Dialogue translation is a particularly difficult task because of the high frequency of elliptical, partial or ill-formed utterances, as well as the use of anaphora and deictic reference (cf. section 2.7). Furthermore, in comparison with the kinds of texts normally translated by MT systems, dialogue contains a much larger proportion of language where literal translation is inappropriate: identical utterances can have completely different discourse functions — and hence translations — from one moment to the next. For example, *OK* in a dialogue might mean any of the paraphrases in (12).

(12a) I agree with you.

(12b) I can still hear you.

(12c) Let's change the subject now.

(12d) That is good.

As in the UMIST–British Telecom research, the system has some expectations of the kinds of things the user might want to say: the system has a bilingual 'dialogue model', and interacts with the user to try to match the user's input to the range of possible utterances in the model, which of course the system can be confident of translating correctly. There are two possible scenarios. In one, the user takes the initiative, typing in proposed utterances. In the other, the system might take the intiative, making proposals about what the next utterance might be, on the basis of its dialogue model.

In so far as these systems attempt to match input phrases against pre-translated **text segments** in a database, they belong to the genre of Example-based MT systems (cf. 18.2 above). A rather more ambitious project, known as LIDIA ('Large Internationalisation of Documents through Interaction with Authors'), has been proposed from the University of Grenoble. For its linguistic techniques it is founded on GETA's well tested Ariane system (described in Chapter 13). The aim is an interactive system enabling researchers to translate their own texts from their own language (French in this case) into another unknown language (German or Russian), providing also as a check a translation of the target text back into the original. Unlike the UMIST projects it will not be based on pre-translated text fragments but will translate in the familiar MT manner from largely already written texts.

18.6 Speech translation: British Telecom and ATR

Advances in speech technology in recent years have encouraged a number of researchers to investigate the integration of translation systems with speech recognition and synthesis. We have seen already (section 18.3 above) the application of methods from speech research in a statistics-based translation system. Here, we describe briefly two projects where input and output is spoken but where translation itself is based on more traditional linguistics-based approaches.

The experiment at the British Telecom Research Laboratories was based on the matching of spoken words against standard phrases in the highly restricted domain of telephonic business communication. The restriction was necessitated by the severe performance limitations of current speech recognisers. The researchers isolated the distinctive words which uniquely identified phrases. For example, given the three phrases in (13), the recognition of the three keywords *you, speak* and *I* would be sufficient to ensure unique identification.

(13a) Whom do you want to speak to?

(13b) I cannot hear you.

(13c) May I speak to Mr Smith please?

With a phrasebook of 400 sentences it was found that the ten most useful English keywords were *the, a, I, you, to, room, is, hotel, for, of*. Similar sets were found for French, German and Spanish. Obviously, variable elements such as times and prices could not be treated in the same way; these had to be identified and translated individually. The system involved three stages: the telephone caller input a phrase to the speech recogniser, enunciating each word clearly and pausing between each word; the computer selected a phrase and processed any variables; the chosen phrase was displayed on a screen; if acceptable as a paraphrase, the stored translation was transmitted; and the final message was output in synthetic speech (and/or displayed on a screen) at the receiver's end. The BT researchers are now investigating a bilingual spoken communication system for use by the French and British police forces, involving research on the sublanguage of police messages ('Policespeak').

A much more ambitious project for telephone translation is underway, at the ATR Laboratories in Japan. The long-term aim is a system for translating unrestrained spoken dialogue between English and Japanese. It involves basic research on speech recognition, speech synthesis and automatic translation of dialogue. Current speech recognisers usually require pauses between words and are adapted to particular individuals. Speaker-independent recognition of continuous conversation presents considerable challenges, involving the incorporation of prosodic information such as pitch, stress and duration, the recognition of syllable boundaries, the integration with a parser for identifying phonemes and word boundaries, the use of semantic and pragmatic information to narrow down potential interpretations, and much else. The requirements for automatic speech synthesis are almost equally challenging: in comparison with current synthesisers the demand is for higher quality, more 'natural' output which takes account of prosody and inflexion and which is more 'personalised'.

As for the translation of dialogue, this is itself a new area for MT (cf. 18.5 above): spoken dialogue differs from written text in vocabulary and grammar, while incomplete utterances, false starts, ellipses, unstated implications, etc. are commonplace. The ATR project is undertaking fundamental linguistic research on speech acts, dialogue switching, Japanese honorifics, inferring function words and omitted subjects, and producing idiomatic output (including use of example-based methods). As stated above, the research focus at present is the communication environment of an international conference office. Whether an operational prototype emerges or not, the basic research at ATR has been contributing substantially to MT in general and will continue to do so in the future.

18.7 Reversible grammars

The linguistic foundations of MT continue to be explored in a number of directions. There is much attention to the application and implementation of recent developments in theoretical linguistics, such as LFG, GPSG, GB, Systemic–Functional grammar, Categorial grammar, Situation Semantics, Logic grammars (some of which have been mentioned in passing). Other prominent topics at present include compositionality and reversibility, as we saw in Chapter 16 on Rosetta. Reversibility of grammars, in particular, has been an objective of a number of projects; just one fairly typical example is given here: the CRITTER project in Canada. (See also Chapter 7 for further references.)

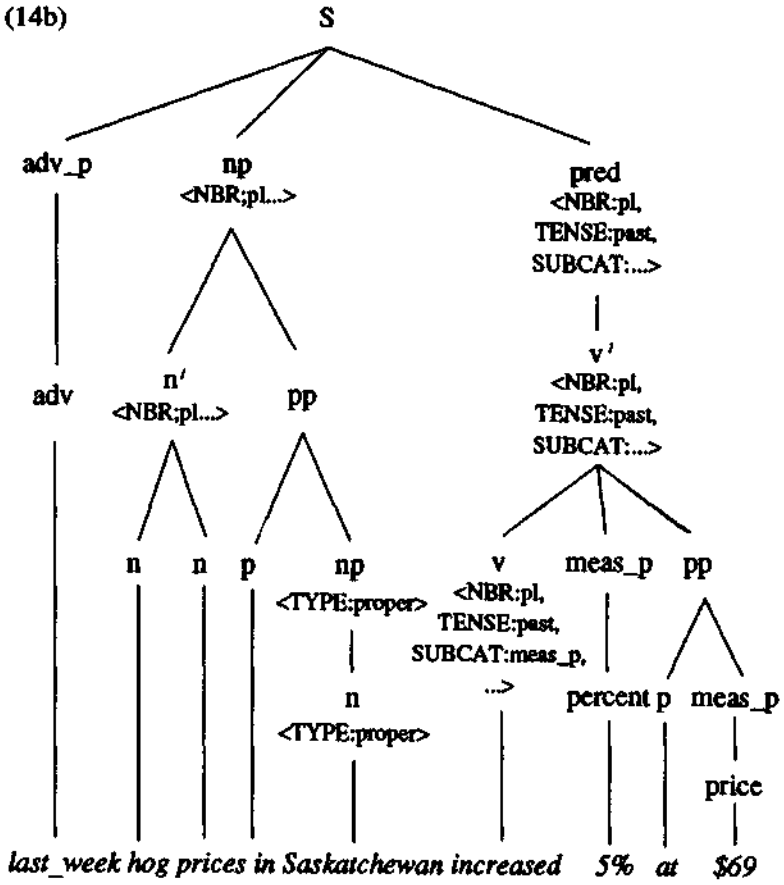
CRITTER is a transfer-based system for translating between English and French. It has been developed at the Centre Canadien de Recherche sur l'Informatisation du Travail (CCRIT), the Canadian Workplace Automation Research Centre, by researchers originally attached to the TAUM project in Montreal. It is designed as a general experimental MT model applied at present to the specific sublanguage of the weekly reports produced by the Canadian Department of Agriculture which describe the situation in the livestock and meat trade markets of Canadian provinces. The principal consequence is that the lexicon is restricted to the vocabulary of the reports, although the grammatical formalism and the computational implementation are not constrained.

CRITTER exemplifies many typical features of the current philosophy in linguistics-oriented MT. Syntactic representations are fairly standard surface structure dependency trees, labelled with syntactic and semantic features, reflecting $\bar{X}$ conventions, and marking gaps and traces in the normal way. For (14a) the syntactic tree is (14b).

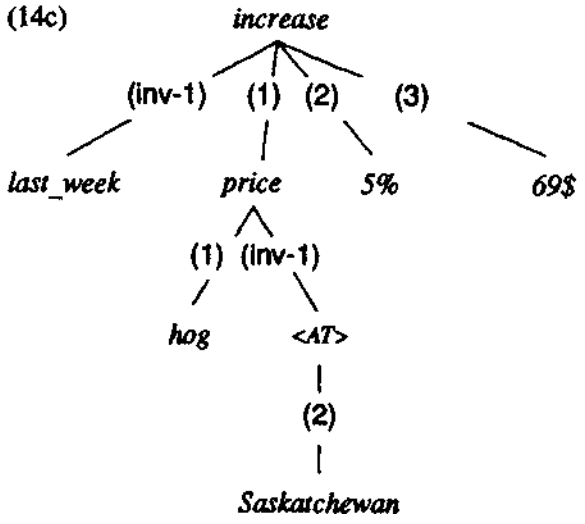
(14a) Last week hog prices in Saskatchewan increased 5% at $69.

In this particular tree, the only feature reflecting the sublanguage type is the inclusion of an idiosyncratic 'meas_p' as a subcategorisation of the verb *increase*. The intermediate semantic representation for the same sentence (14c) is an abstract dependency tree with branches labelled by argument numbers and with lexical items as nodes. The treatment of *last week* as an unanalysed unit is justified by its status as a fixed 'idiom' in the sublanguage.

(14b)


S
adv_p
np
<NBR;pl...>
pred
<NBR:pl,
TENSE:past,
SUBCAT:...>
v'
<NBR:pl,
TENSE:past,
SUBCAT:...>
adv
n'
<NBR;pl...>
pp
n
n
p
np
<TYPE:proper>
n
<TYPE:proper>
v
<NBR:pl,
TENSE:past,
SUBCAT:meas_p,
...>
meas_p
pp
percent
p
meas_p
price
last_week hog prices in Saskatchewan increased 5% at $69

(14c)


increase
(inv-1)
(1)
(2)
(3)
last_week
price
5%
69$
(1)
(inv-1)
hog
<AT>
(2)
Saskatchewan

The use of 'inverted' argument numbers ('inv-1') enables the indication of two types of relationship: predicate–argument relations and syntactic dependency. Thus the dependent *last week* is a predicate with *increase* as argument, and *prices* is in first argument position relative to the 'abstract' item <AT> (the second argument being *Saskatchewan*).

Before transfer, the semantic structure is checked for consistency of predicate and argument nodes. A semantic lexicon associates semantic 'types' to nodes (e.g. MOVEMENT to *increase*, MEASURE-FUNCTION to *price*, INCREMENT to *5%*, etc.) and validates predicate–argument structures against 'schemas' of semantic types, such as (15).

(15) MOVEMENT(MEASURE-FUNCTION,INCREMENT,MEASURE)

Transfer rules are all associated with lexical entries. They are mostly straightforward, as in (16a), but may deal with 'structural' differences, such as argument conversion (16b), or more complex transfer (16c).

(16a) eat ↔ *manger*

(16b) miss(1:X,2:Y) ↔ *manquer*(1:Y',2:X')

(16c) walk(inv-1:across(2:X)) ↔ *traverser*(2:X',inv-1:$manner(2:à_pied))

CRITTER's grammars are the same for both analysis and generation, and are written in a Definite Clause Grammar formalism suitable for implementation in Prolog. From them are derived the parsers and generators, which differ primarily in the order in which rules are applied. In so far as the grammars used for syntactic analysis are also used for generation, and as the transfer rules are reversible, CRITTER exemplifies the reversibility methodology. In fact, its developers argue that CRITTER is as yet the only MT system with fully reversible grammars; Rosetta, for example (Chapter 16), includes a surface syntactic component in analysis which is absent on the generation side. On the other hand, CRITTER is itself, in one respect, not a truly symmetric system, since the semantic checking that is part of the analysis is not necessary for generation.

18.8 Computational developments

One of the most significant advances in the computational modelling of cognition, perception and learning has been research on parallel computation, neural networks and connectionist models. There is a widespread belief that the higher mental functions of language understanding, logical inferencing, and memory can only be modelled with brain-like mechanisms which compute massive amounts of information in parallel. The representation of knowledge demands likewise highly complex networks of interrelated 'concepts'. The neural network of the brain is assumed, with much experimental evidence, to be able to access and activate nodes simultaneously. The 'spreading activation' model of neural networks can be computationally modelled by the connectionist approach to computer design.

The relevance of the connectionist model to natural language processing is clear enough. The traditional stratificational approach to parsing and generation (morphology, syntax, semantics), while conceptually and computationally tractable, is not seriously accepted by linguists or computer scientists as a psychologically

real model of how humans understand and communicate. The frequently expressed aspiration to integrate syntactic, semantic and pragmatic operations is an acknowledgement of the intrinsic attraction of connectionist processing. It is indeed difficult to envisage how the ambitious ATR speech translation system can possibly be realistically implemented without massively parallel computation. From a 'lower level' perspective it is obvious that parallel parsing of syntactic structures could bring marked improvements in present speeds of computer processing.

Speculation about the impact of parallel computation is complicated by the dearth of programming experience with neural networks, most of which have to be simulated on existing sequential computers. Evidence so far, however, suggests that connectionist networks can successfully 'learn' (or be 'trained') to parse previously unseen sentences with a high degree of accuracy. The idea of an MT system learning from past mistakes and from the corrections of post-editors has been put forward frequently. Indeed 'learning' routines have been implemented (e.g. on the Japanese MAPTRAN system and on the commercial Tovna system), but what is meant usually is that changes are suggested by the system on the basis of statistics about errors and corrections, and confirmed or rejected by developers or users. In a true learning system, changes would be initiated automatically by a complex feedback mechanism and constantly tested against new input. The connectionist model arguably offers the prospect of MT systems which really learn.

Less speculative developments are the possible integration of MT with other natural language processes. There have already been successful links with **information retrieval** systems, i.e. systems which enable users to search for titles and abstracts of documents on subjects of interest. Users are able to search remote databases containing abstracts in unknown languages and request translations of the abstracts or the full documents into their own language. Either the queries are translated and searches carried out in the other language, or the titles (or abstracts) have been already translated (e.g. by an MT system) and are searched in the user's language. At present the information retrieval systems and the MT systems are separated, but it is not difficult to envisage a future integrated multilingual information retrieval system.

A further step would be the automatic production of **abstracts** or summaries of texts for users unfamiliar with the original language. Summaries of foreign language documents would certainly be more attractive to most administrators, business people and scientists than even the roughest translations of full texts. However, it is evident from small-scale experiments on summarisation in restricted domains that the complexities of the task are at least equal to those of MT itself.

18.9 Concluding comments

The focus of this book has been the problems and difficulties of programming computers to translate and the methods which have been developed to tackle and overcome them. Except in two chapters (8 and 9), we have not discussed the practical aspects of MT operations, for reasons stated in the Preface. However, research on MT is not 'pure' research: it is directed towards the provision of tools for practical use, and it should be motivated by clearly defined objectives.

This has often been forgotten or ignored. Until the late 1970s most MT research activity was undertaken in academic environments with relatively little regard for immediate or even potential long-term applications. During the 1980s much basic research has been undertaken by independent companies, mainly in the electronics and computer business, for short- or long-term commercial interests. The impact has been two-fold: on the range of languages covered and on the types of systems developed.

For the first two decades of the history of MT, systems were developed primarily for the use of scientists to keep abreast of technological activity. Research concentrated on translation from Russian, or — in the case of Soviet MT research — from English. In the 1970s systems were designed for the pressing needs of bilingual Canada and the multilingual European Communities. The emphasis was on systems producing translations in bulk for post-editing by translators. In the 1980s the demand has been for systems covering the major commercial languages of the world (chiefly English, French, German, Spanish, and Japanese), and the need was for high-quality output: systems where the input could be controlled or systems involving considerable intervention by translators. Now, additional demands are emerging: systems translating other commercially important languages (e.g. Arabic, Chinese, Korean), systems for translating documents, textbooks and manuals from the 'major' languages of the developed world into the languages of the less developed countries, and systems for business people and researchers to translate messages and documents into languages they do not know. At the same time, the traditional expectations remain: there is a growing demand for rough translations for information-gathering and review purposes; users want improved translation aids (not just automatic dictionaries, multilingual word-processing and the like, but provisional pre-translations produced automatically); and companies are looking for systems to tackle multilingual documentation of various kinds and levels of quality (correspondence, technical manuals, marketing literature, etc.)

What can be safely predicted is that in the future we will see MT systems serving many varieties of purposes and users: systems for free unedited text input, for guided input, for texts in controlled languages, for pre-edited texts; systems for spoken communication, for on-line dialogue, for access to databases; systems for specific sublanguages, for specific text types (e.g. patents, abstracts), for broad subject areas (technical documents), even for 'any' subject; systems producing rough translations, good quality translations, preliminary drafts for translators; systems demanding human interaction during analysis, transfer or generation and systems operating in batch modes; systems for monolingual users, for professional translators, for occasional translators, for users ignorant of the source language, for users not knowing the target language; systems for scientists, for business people, for travellers, for administrators, for diplomats, for language learners, etc.; expensive systems for large multinational companies, systems for freelance translators, cheap desktop systems, hand-held systems. The permutations begin to seem endless, but what we can also predict is that there will not, in the foreseeable future, be an 'ideal' system capable of accepting all types of texts in all or most subjects producing output to the standard of the best human translators, and that there will not be MT systems capable of literary translation.

As for future methods and techniques, we can predict an equal variety of permutations: linguistics-based, knowledge-based, example-based, and statistics-based methods; direct, transfer and interlingua systems, and hybrids of various types; bilingual and multilingual configurations; facilities for non-interactive text input, dialogue-based composition of texts, spoken input; reversible grammars, rule-based systems, learning systems; and no doubt many others. Approaches and methods will be developed in response to purposes and goals, and in response to developments in linguistics, in computer science and technology, in cognitive science, in telecommunications, and no doubt elsewhere. Which approaches and techniques (or rather combinations of techniques) will lead to substantial improvements in the quality of MT output cannot be foreseen.

Machine Translation is one of the most challenging research activities, involving the application of complex theoretical knowledge to the building of systems whose successes and failures can be judged by laymen in the simplest of terms. We hope that this book has shown the nature and difficulties of the task and will inspire others to take up the challenge.

18.10 Sources and further reading

Surveys of current research appear regularly; recent ones by the authors include Hutchins (1988, 1990) and Somers (1990, 1991);

The most complete description of the CMU system is found in a special issue of *Machine Translation* edited by Goodman (1989); see especially Nirenburg (1989). The theoretical arguments for the CMU approach are given by Nirenburg and Goodman (1990). For details of earlier AI approaches, see Chapter 15 of Hutchins (1986).

The Bilingual Knowledge Bank of the DLT project is described most fully by Sadler (1989). The notion of 'bitext' corpora as aids for translators was proposed by Harris (1988). References to other experiments in Example-based MT are given in Chapter 6.

The statistical methods employed by the IBM system are described by Brown *et al.* (1990). For TITUS see Ducrot (1989).

The Ntran system is described in Whitelock *et al.* (1986) and in Wood and Chandler (1988). For the UMIST–British Telecom research see Jones and Tsujii (1990); the Malaysian work is described in Zaki and Muhayat (1991). For the UMIST–ATR work see Somers *et al.* (1990). The LIDIA project is described in Boitet (1990).

The discussion of CRITTER is based on Isabelle *et al.* (1988)

The BT 'phrasebook' project is described by Steer and Stentiford (1990), and the plans for Policespeak by Jackson (1990). There are numerous papers describing the ATR speech project, including Kakigahara and Aizawa (1988), Yoshimoto (1988), Kume *et al.* (1989), Kogure *et al.* (1990).

The 'learning' mechanism (PECOF) for the English–Japanese MAPTRAN system is described by Nishida and Takamatsu (1990). The investigation of the use of various Japanese–English MT systems in conjunction with Japanese databases of

scientific and technical abstracts has been described by Sigurdson and Greatrex (1987).

For those readers who want to keep up to date with MT research, the proceedings of conferences are essential, notably the MT Summit conferences, the series of International Conferences on Theoretical and Methodological Issues in Machine Translation of Natural Language and the series of biennial Coling conferences. The main journal in the field is *Machine Translation* published by Kluwer (Dordrecht, The Netherlands), but articles on this topic appear in a wide range of other journals.