Chapter 8

Multilinguality

8.1 Overview

Martin Kay

Xerox Palo Alto Research Center, Palo Alto, California, USA

Multilinguality is a characteristic of tasks that involve the use of more than one natural language. In the modern world, it is a characteristic of a rapidly increasing class of tasks. This fact is most apparent in an increased need for translations and a consequent interest in alternatives to the traditional ways of producing them. The principal alternatives that have been proposed include partially or fully automatic translation, machine aids for translators, and fully or partially automated production of original parallel texts in several languages. But multilinguality is more than just the preparation of parallel texts. Before anything nontrivial can be done with a text—before it can be filed, or sent to the appropriate person, or even responsibly destroyed—the language in which it written must be identified. This so called *Language ID* problem is therefore a pressing one, and one on which technology has fruitfully been brought to bear. In working environments where more than one language is in use, the problem of storing and retrieving information acquires a multilingual dimension. These problems, as well as that of processing spoken material in a multilingual environment, will be reviewed in this chapter.

Where only one language is involved, a great deal of useful processing can be done on the basis of a model that sees texts as essentially sequences of characters. This is the view that most word processors embody. Words are recognized as having properties beyond the characters that make them up for the purposes of detecting and correcting spelling errors and in information retrieval. However, of the multilingual problems just identified, the only one that might possibly be treated with a character-oriented model is that of language identification. The remainder trade in an essential way on equivalences, or near equivalences, among words, sentences, and texts mediated through their meaning. Language processing of this kind is notoriously difficult and it behooves us to start by considering, however cursorily, why this is. We will do this in the context of translation, though what we say is true for the most part of the other tasks mentioned.

The question of why translation should have been so successful in resisting the most determined efforts to automate it for close to forty years is complex and sometimes quite technical. But it is not a mystery. The basic problems have long been known and, the most important thing that has been learnt about them recently is that they are more severe and more widespread than was first thought. Perhaps the most important problem concerns a distinction between meaning and interpretation. Another has to do with the more classical distinction between meaning and reference.

One example must suffice to give a sense of the problem concerning meaning and reference. The French sentence *Où voulez-vous que je me mette?* means, more or less literally, *Where do you want me to put myself*? Colloquially translated into English, however, it would have to be *Where do you want me to sit / stand / park / tie up (my boat) / sign my name, etc.* Information must be added which is not in the original to make the English sound natural. The sentence *Where do you want me to put myself?* means what the French means, but we claim a translator would deliberately choose a rendering that would change the meaning. In this case, it may also be possible to get the right effect by deleting information, as in *Where do you want me?* but this also changes the meaning. What remains invariant under translation is not the meaning, but the interpretation, that is, the response the text is intended to evoke in a reader. Interpretation depends on context, in there lies the principal source of the difficulty.

The distinction between meaning and interpretation, to the extent that it was recognized at all, has generally been thought too subtle to be of practical interest. The belief has been that, in sufficiently restricted or sufficiently technical domains of discourse, it essentially disappears. However, in the massive speech-to-speech translation project recently initiated by the German government (Kay, Gawron, et al., 1991), the universe of discourse is limited to discussions between a pair of individuals on the time and place of their next meeting. In one of the first simulated dialogs examined, the sentence *Geht es bei Ihnen?* occurs. There are two interpretations, which can be captured in English by *Is that alright with you?* and *Can we meet at your place?*. The domain of the discourse is already restricted to an extreme degree and it is clear that nothing but an analysis of the context will decide the interpretation. Restriction to a technical domain can help, but it can also hinder. When I order ice cream, I may be asked if I want two scoops or three—in French *Deux boules ou trois?* and, in German, *Zwei Kugeln oder drei?*. But *boule* and *Kugel* mean *ball*, not *scoop*. At first the problem seems easy to resolve. The words *scoop*, *boule* and *Kugel* are classifiers for ice cream in their respective languages, just as *loaf* is classifier for bread in English. But this fails utterly in a technical document, say a patent application, describing an ice cream scoop the very point of which is that it delivers ice cream in different shaped pieces. To handle these words right in any context, one must understand the relationship of the scoop to the shape it imposes on the material it is used to manipulate. More importantly, one must understand from the context when the usual assumptions about this relationship no longer hold.

The question concerning meaning and reference has to do with a philosophical problem that is far beyond our present scope, namely the extent to which meaning is parasitic on reference. To many, it seems unlikely that a baby could learn a language in any useful sense without ever establishing an association between the words in it and objects in the environment. Essentially all computers lack anything that could reasonably be called a perceptual system—they have never seen, heard, felt, or smelt anything. Just how much processing of meaning and interpretation is possible for such a device is open to considerable doubt. Language processing is done, at worst, with characters and, at best, with meanings. Occasionally, programs have been written that manipulate toy blocks or the pieces on a chess board, and which talk about these objects using natural language, but these experiments have been too small to tell us anything about the importance of a genuine ability to refer to things. This will be taken up again in section 8.2.

Workers in artificial intelligence and computational linguistics are often at odds on the extent to which computer programs intended to mimic human performance for practical reasons need to use human methods. On the one hand, computers have quite different properties from humans; we usually do not know what methods humans use in any case; and airplanes do not flap their wings. On the other hand, divining the intended interpretation of a text requires second-guessing the intentions of its author in the given context, a task that seems essentially to require a human point of view.

There is an essentially *bottom-up* quality to the translation problem as usually seen by those that attempt to automate it. It starts with words, phrases, and sentences and rarely takes any account of larger structures. This adds greatly to its difficulty both for people and machines. The point is simply that the translator must attempt to reproduce the intention of the author, whatever it might be, in the large and in the small. To the extent that the translator can permit himself any assumptions about these matters, the problem assumes some top-down properties which make it, to however small an extent, more tractable. This is why the results reported in the recent ARPA Message Understanding Conferences (MUC) are so much more encouraging. The aim here was to extract information about terrorist incidents from newspaper material, ignoring all else, and attending only to certain facts about the incidents. For the same reason, some of the early experiments of Roger Shank and his students on translation also seemed encouraging, because they allowed themselves to make strong assumptions about the texts they were working with. They allowed themselves assumptions not only about the overall subject matter, but also about the structure of the texts themselves. For similar reasons, there is reason to hope for more positive results in multilingual information retrieval.

Three responses to the problems of context and interpretation suggest themselves. First, in the long run, there is no alternative to continuing to build more faithful models of human behavior. The second alternative is to design systems involving both people and machines, assigning to each those parts of the task to which they are best suited. The third is to seek ways of modifying the task so that the machine will naturally have greater control over the context. Section 8.4 explores the second of these alternatives. The third, we discuss briefly now.

The METEO machine-translation system translates Canadian meteorological bulletins between English and French. Realizing that METEO's spectacular sucess was due to the remarkably restricted nature of the texts it worked on, workers at the University of Montreal reflected on the possibility of eliminating the input text altogether in favor of data gathered directly from weather stations. This line of thought led to a system that produces parallel English and French marine weather bulletins for the Canadian eastern seaboard. The planning of what will be said and in what order is done once for both languages. It is only towards the end that the processes diverge (Chandioux, 1989). The same approach is being taken with reports based on Canadian labor statistics. The TECHDOC project at the University of Ulm aims to produce parallel technical documentation in multiple languages on the basis of a language-independent database (Rösner & Stede, 1994); and the Information Technology Research Institute at the University of Brighton has a group working on the automatic drafting of multilingual instructional technical texts in the context of GIST (Generating InStructional Text), part of the European Union's LRE program (Delin, Hartley, et al., 1994; Paris & Scott, 1994). These projects eliminate the problem of determining the intended interpretation of a piece of input text in differing degrees. In the second and third cases, there is still intentional material in the input but the idea in each case is to shift the emphasis from determining the intentions behind a given document to creating intentions for a new set of parallel documents.

8.2 Machine Translation: The Disappointing Past and Present

Martin Kay

Xerox Palo Alto Research Center, Palo Alto, California, USA

The field of machine translation has changed remarkably little since its earliest days in the fifties. The issues that divided researchers then remain the principal bones of contention today. The first of these concerns the distinction between that so-called interlingual and the transfer approach to the problem. The second concerns the relative importance of linguistic matters as opposed to common sense and general knowledge. The only major new lines of investigation that have emerged in recent years have involved the use of existing translations as a prime source of information for the production of new ones. One form that this takes is that of example-based machine translation (Furuse & Iida, 1992; Iida & Iida, 1991; Nagao, 1992; Sato, 1992) in which a system of otherwise fairly conventional design is able to refer to a collection of existing translations. A much more radical approach, championed by IBM (Brown, Cocke, et al., 1990), is the one in which virtually the entire body of knowledge that the system uses is acquired automatically from statistical properties of a very large body of existing translation.

In recent years, work on machine translation has been most vigorously pursued in Japan and it is also there that the greatest diversity of approaches is to be found. By and large, the Japanese share the general perception that the transfer approach offers the best chance for early success.

Two principal advantages have always been claimed for the interlingual approach. First, the method is taken as a move towards robustness and overall economy in that translation between all pairs of a set of languages in principle requires only translation to and from the interlingua for each member of the set. If there are n languages, n components are therefore required to be translated into the interlingua and n to translate from it, for a total of $2n$. To provided the same facilities, the transfer approach, according to which a major part of the translation system for a given pair of languages is specific to that pair, requires a separate device to translate in each direction for every pair of languages for a total of $n(n-1)$.

The PIVOT system of NEC (Okumura, Muraki, et al., 1991; Muraki, 1989) and ATLAS II of Fujitsu (Uchida, 1989) are commercial systems among a number of research systems based on the two-step method according to which texts are translated from the source language to an artificial interlingual representation and then into the target language. The Rosetta system at Phillips (Landsbergen, 1987), and the DLT system at

BSO (Witkam, 1988; Schubert, 1988) in the Netherlands also adopt this approach. In the latter, the interlingua is not a language especially designed for this purpose, but Esperanto.

According to the majority transfer view of machine translation, a certain amount of analysis of the source text is done in the context of the source language alone and a certain amount of work on the translated text is done in the context of the target language, but the bulk of the work relies on comparative information about the specific pair languages. This is argued for on the basis of the sheer difficulty of designing a single interlingua that can be all things for all languages and on the view that translation is, by its very nature, an exercise in comparative linguistics. The massive Eurotra system (Schutz, Thurmair, et al., 1991; Arnold & des Tombes, 1987; King & Perschke, 1987; Perschke, 1989), in which groups from all the countries of the European Union participated, was a transfer system, as is the current Verbmobil system sponsored by the German Federal Ministry for Research and Technology (BMFT).

A transfer system in which the analysis and generation components are large relative to the transfer component and where transfer is therefore conducted in terms of quite abstract entities takes on much of the flavor of an interlingual system while not making the commitment to linguistic universality that many see as the hallmark of the interlingual approach. Such semantic transfer systems are attracting quite a lot of attention. Fujitsu's ATLAS I (Uchida, 1986) was an example, and Sharp's DUET system is another. The approach taken by SRI (Cambridge) with the Core Language Engine (Alshawi, Carter, et al., 1991) also falls in this category.

Just as these systems constitute something of an intermediate position between interlingua and transfer, they can also be seen to some extent as a compromise between the mainly linguistically based approaches we have been considering up to now and the so-called knowledge-based systems pursued most notably at Carnegie Mellon University (Nirenburg, Raskin, et al., 1986; Carbonell & Tomita, 1987), and at the Center for Research in Language at New Mexico State University (Farwell & Wilks, 1990). The view that informs these efforts, whose most forceful champion was Roger Shank, is that translation relies heavily on information and abilities that are not specifically linguistic. If it is their linguistic knowledge that we often think of as characterizing human translators, it is only because we take their common sense and knowledge of the everyday world for granted in a way we clearly cannot do for machines.

Few informed people still see the original ideal of fully automatic high-quality translation of arbitrary texts as a realistic goal for the foreseeable future. Many systems require texts to be preedited to put them in a form suitable for treatment by the system, and post-editing of the machine's output is generally taken for granted. The most successful systems have been those that have relied on their input being in a sublanguage (Kittredge, 1987), either naturally occurring, as in that case of weather reports, or deliberately controlled. The spectacular success of the METEO system (Chevalier, Dansereau, et al., 1978) working on Canadian weather reports encouraged the view that sublanguages might be designed for a number of different applications, but the principles on which such languages should be designed have failed to emerge and progress has been very limited.

Future Directions

Research in machine translation has developed traditional patterns which will clearly have to be broken if any real progress is to be made. The traditional view that the problem is principally a linguistic one is clearly not tenable but the alternative that require a translation system to have a substantial part of the general knowledge and common sense that humans have seems also to be unworkable. Compromises must presumably be found where knowledge of restricted domains can facilitate the translation of texts in those domains. The most obvious gains will come from giving up, at least for the time being, the idea of machine translation as a fully automatic batch process in favor of one in which the task is apportioned between people and machines. The proposal made in Kay (1980), according to which the translation machine would consult with a human speaker of the source language with detailed knowledge of the subject matter, has attracted more attention in recent times. A major objection to this approach, namely that the cost of operating such a system would come close to that of doing the whole job in the traditional way, will probably not hold up in the special, but widespread situation in which a single document has to be translated into a large number of languages.

8.3 (Human-Aided) Machine Translation: A Better Future?

Christian Boitet

Université Joseph Fourier, Grenoble, France

As the term *translation* covers many activities, it is useful to distinguish, at least, between:

- **re-creation**, e.g., the translation of poetry or publicity, which aims above all at transmitting the subjective aspect of a text, even if its objective meaning is somewhat altered;
- **localization**, practised on a large scale nowadays on computer manuals for end users, where it is important to adapt certain parts of the content, and perhaps the style of the presentation, to a certain cultural and linguistic environment;
- **diffusion translation**, in particular the translation of technical documentation, where the objective content must be strictly rendered in another language, without addition and omission, even if the style *smells translation*;
- **screening translation**, which covers translation of written material for gathering information as well as simultaneous interpretation of oral presentations.

8.3.1 Types of MAT Systems Available in 1994

It is impossible to envisage an automation of re-creation translation and of localization which would go beyond machine aids for human translators for many years to come. By contrast, the *translating function* may be automated in the case of diffusion-translation and screening-translation. To fix our vocabulary, we would like to take the term machine assisted translation (MAT) as covering all techniques for automating the translation activity. The term human-aided machine translation (HAMT) should be reserved for the techniques which rely on a real automation of the translating function, with some human intervention in preedition, postedition or interaction. The term machine-aided human translation (MAHT) concerns machine aids for translators or revisors and is the topic of the section 8.4.

MT for Screening Purposes

Around 1949, MT projects were launched first in the US, and soon thereafter in the USSR. They were motivated by the growing needs for intelligence gathering. They gave rise to the first MT screening systems. The goal of such systems is to produce automatically, quickly and cheaply large volumes of *rough* translations. The quality of the rough translations obtained is not essential. The output can be used to get an idea of the content. If the user wants a good translation of a part which looks interesting, he simply asks a human translator (who in general will judge the machine output to be too bad to bother with revision).

What is essential is that in order to keep costs low, no professional translator or revisor should be used. Preedition should be reduced to confirming system proposals for separating figures, formulae, or sentences. Postedition, if any, should consist only in formatting operations. The need for *screening* MT is still actual. However, civil uses (gathering technological, economical and financial information) are now predominant over military uses. Examples of working systems are SYSTRAN (Russian-English in the US and several language pairs at the EC), ATLAS-II (Japanese-English for the EC), and CAT from Bravice, used to access Japanese data bases in English (Sigurdson & Greatex, 1987).

Users can get access to these systems from terminals (even minitels), standard PCs or Macintoshes connected to a network. In the last few years, stand alone configurations have appeared on PCs and workstations. We describe briefly the different access modes:

Access to a Server: In France, Systran SA commercializes an MT server via the minitel network (6–7 million of these relatively dumb terminals are installed in French homes). This service gives access to several Systran *language pairs*. This system can meet users expectations if used for screening purposes (translation into the mother tongue). At the European Commission, Systran has also been used since the end of 1976. These translations are now distributed as they stand to interested readers, instead of being revised by human translators. With that change, the amount of texts going through MT has suddenly increased from 2,000 pages in 1988 to 40,000 in 1989 to 100,000 in 1993 (the total number of pages translated varying from 800,000 to 1,000,000 to 1,500,000). We should also mention the growing use of PC's connected to computer networks for getting access to rough MT translations of textual data bases (economical for NHK, scientific and technical at JICST, etc.), sometimes transcontinentally (Sigurdson & Greatex, 1987).

Integrated Stations: Hardware has become powerful and cheap enough to run some MT systems on a PC, possibly coupled with an OCR. These systems include very restricted systems for diffusion, such as METEO on PC, and some systems for screening, such as Translator by Catena on Macintosh. However, at this point, the size of the dictionaries and the sophistication (and associated computational cost) of the underlying tools make workstations mandatory for the majority of currently available commercial systems but this is bound to change soon.

MT for Diffusion Purposes

Work on diffusion MT or MT for the revisor began when the first interactive systems appeared. The aim is to automate the production of professional quality translations by letting the computer produce the *first draft.* Hence, the MT system must be designed to produce *raw* translations good enough so that professional revisors will accept to postedit them, and that overall costs and delays are reduced. That is possible only if the system is specialized to texts of a certain style and domain ("suboptimization approach" in L. Bourbeau's terminology Bourbeau, Carcagno, et al., 1990; Lehrberger & Bourbeau, 1988). Political, scientific and industrial decision makers, as well as the public at large, often envisage that arrangement (pure MT followed by postedition) as the only possible.

About twenty systems are now commercially available. About fifteen of them are Japanese (AS-Transac by Toshiba, ATLAS-II by Fujitsu, PIVOT by NEC, HICAT by Hitachi, SHALT-J by IBM-Japan, PENSÉ by OKI, DUET by Sharp, MAJESTIC by JICST, etc.), and handle almost exclusively the language pairs Japanese / English. Other systems come from the U.S. (LOGOS, METAL, SPANAM), France (Ariane/aéro/F-E by SITE-B'VITAL, based on GETA's computer tools and linguistic methodology), or Germany (SUSY by IAI in Saarbruecken), and center on English, German or French, although mockups and prototypes exist for many other languages. Still others are large and operational, but not (yet ?) commercially offered (JETS by IBM-Japan, LMT by IBM-US, ALT/JE by NTT, etc.).

What can be expected from these systems? Essentially, to answer growing needs in technical translation. In the average, a 250-word page is translated in 1 hour and revised in 20 min. Hence, 4 persons produce a finished translation at a rate of 3 pages per hour (p/h). Ideally, then, some translators could become revisors and 6 persons should produce 12 p/h. As it is, that is only an upper limit, and a more realistic figure is 8 p/h, if one counts a heavier revision rate of 30 mn/p (after adequate training). Several users report overall gains of 40 to 50%. An extreme case is the METEO system (Chandioux, 1989), which is so specialized that it can produce very high quality raw translations, needing only 3 text processor operations per 100 words translated. Another way of looking at the economics of MT is in terms of human effort: according to figures given by producers of MT systems (JEIDA, 1989), the creation of a new (operational) system from scratch costs between 200 and 300 man-years with highly specialized developers. Also, the cost to adapt an existing system to a new domain and a new typology of texts is in the order of 5 to 10 man-years, which makes it impractical for less than 10,000 pages to translate. All counted, the breakeven point lies between 9,000 and 10,000 pages, an already large amount.

This approach, then, is at present only envisageable for large flows of homogeneous and computerized texts, such as user or maintenance manuals. An essential condition of success is that the team in charge of developing and maintaining the lingware (dictionaries, grammars) be in constant touch with the revisors, and if possible with the authors of the documents to be translated. A good example in this respect is Pan American Health Organization (PAHO) (Vasconcellos & Len, 1988), with its systems ENGSPAN and SPANAM.

Users should consider this kind of MT systems in the same way they consider expert systems. Expert systems can be developed by third parties, but it is essential for users to master them in order to let them evolve satisfactorily and to use them best.

As the MT systems designed for diffusion purposes are computationally very heavy, they have been developed on mainframes. The situation is changing rapidly, however. Since powerful PCs are becoming widely available, they are now replacing terminals. Although many vendors offer specialized editors, on terminals or on PCs, there is a trend to let revisors work directly with their favorite text processor (such as Word, WordPerfect, WordStar, FrameMaker, Interleaf, Ventura, etc.) and to add specific functionalities as *tools* (such as Mercury/Termex or WinTool). But this technique is not yet able to offer all functionalities of specialized editors (such as showing corresponding source and target phrases in inverse video, or doing linguistic alignment, etc.). For example, the METAL system commercialized by Siemens runs on a LISP machine, while revision is done on a kind of PC. It seems also that the ATLAS II, PIVOT, and HICAT systems are still running on mainframes when used in house for the translation of technical documentation, or out house by translation offices submitting possibly preedited material. In France, SITE-B'Vital has ported the Ariane-G5 MT system generator (not yet the development environment) on Unix-based workstations, but the current use is from a PC under Word accessing an MT server running on an IBM 9221 minicomputer. Finally, there is now a commercial offer for diffusion MT systems on workstations (Toshiba, Sharp, Fujitsu, Nec). About 3,000 machines in total had been sold in Japan by April 1992. Systems used for diffusion MT are characterized, of course, by their specialization for certain kinds of texts (grammatical heuristics, terminological lexicons), but also by the richness of the tools they offer for preediting, postediting and stylistic system control (that is possible because intended users are bilingual specialists). They all include facilities to build terminological *user dictionaries*.

8.3.2 Four Main Situations in the Future

We anticipate that users of MT systems will increasingly be non-professionals, that is occasional translators or monolingual readers. According to the linguistic competence of the user and to whether he works in a team or alone, we envisage four types of situations in the middle term future, say, by the year 2000.

Individual Screening Translation Workstations: Servers should continue to coexist with integrated solutions on PCs or workstations. Servers look appropriate for all situations where the same information is likely to be required by many persons, and is already available in computer-readable form (textual data bases, flow of short lived messages such as weather bulletins or stock exchange notices, computerized libraries, etc.). Translation may be performed once, possibly in advance, and some amount of quick revision may even be performed. It is also possible to analyze the text typology and to use corresponding specialized versions of the MT system. Large-spectrum systems will no doubt be ported to the more powerful PCs which will soon be available.

In each case, we can expect environments to be generic. The only difference between the two solutions will be the required computer power. For accessing a server, basic PCs already suffice. But running MT systems requires more power, simply because small improvements in output quality and ergonomy will continue to require a lot of computational resources, and because the basic software tools are also continuously requiring more computer resources.

Occasional Translation: Current tools will no doubt be improved, in terms of speed, ergonomy and functionalities. As far as ergonomy is concerned, we envisage that the translator's aids will work in background and continuously offer help in windows associated with windows of the current application (text processor, spreadsheet, etc.). This begins to be possible, at least on Macintoshes, where different applications can communicate.

New functionalities should include more aids concerning the target language, in particular paraphrasing facilities and better tools for checking spelling, terminology, grammar, and style. They may even include some MT helps, not aiming at translating whole paragraphs or sentences, but rather at proposing translations for simple fragments, perhaps in several grammatical forms that seem possible in the context (case, number, person, time, etc.).

Individual Professional Translation: It can be envisaged that free lance translators will make increasing use of communication facilities, to retrieve terminology, to communicate with authors, or to submit parts of their workload to some MT system. Perhaps they will even have computer tools to help them determine which MT system accessible over the network would be most suitable for the text currently at hand, if any. Current research in *example-based MT* will perhaps lead to much better tools for accessing previous translations of similar passages. As far as hardware is concerned, professional free lance translators should increasingly equip themselves with comfortable, but not too expensive configurations, such as middle-range PCs with large screens, CD-ROMs, and lots of disk space.

Industrial Professional Translation: Industrial translation aims at a very high quality of fairly long documents. That is why the raw translation job (first draft) is usually divided among several translators, and why there is often more than one revision step. If MT is introduced, the revision job still has to be divided among several persons. There is a need for managing this collective effort. Hence, we can anticipate that this kind of translation will be organized around a local network, each translator/revisor working on a powerful PC, and accessing one or more MT servers, a terminology server, an example server (access to available parallel texts), etc., all being controlled by a senior translator using reserved managing facilities on his PC.

8.3.3 Future Directions

From the four types of users (screener, occasional translator, free lance translator, industrial translator), only the first and fourth can already use existing MT technology in a cost-effective way. The third will probably also be able to use it by the year 2000. But there is still a fifth possibility, which is now at the research stage, that of MT for monolingual writers, or *personal MT*. See e.g., Boitet (1986); Boitet and Blanchon (1993); Chandler, Holden, et al. (1987); Huang (1990); Maruyama, Watanabe, et al. (1990); Sadler (1989); Somers, Tsujii, et al. (1990); Tomita (1986); Wehrli (1992); Whitelock, Wood, et al. (1986); Wood and Chandler (1988).

There is actually a growing need to translate masses of documents, notes, letters, etc., in several languages, especially in the global market. People are very conscious that they waste a lot of time and precision when they read or write texts in another language, even if they master it quite well. To take one language like English as the unique language of communication is not cost-effective. There is a strong desire to use one's own language, while of course trying to learn a few others for personal communication and cultural enrichment.

The idea behind this new kind of MT system is that users will accept to spend a lot of time interacting with the machine to get their texts translated into one or more languages, with a guaranteed high quality of the raw output. Engineers or researchers accustomed to painfully (try to) translate their prose into a foreign language (very often English, of course) would perhaps prefer to spend about the same time in such interaction, that is 60 to 90 mn per page, and get their text translated into all the languages of their correspondents. The system would negotiate the text with the author, in order to normalize it according to changeable parameters (style, terminology, etc.), and get a correct abstract representation of it (a so-called *deep* or *intermediate* structure) by asking questions to remove all ambiguities. Then, current technology could be applied to produce quality texts, needing no revision as far as grammaticality is concerned (the content is guaranteed to be correct because of the indirect preedition performed by the author himself, but the form and style would certainly be improvable).

This is of course another version of the old idea of interactive translation, proposed time and again since the first experiments by Kay and Kaplan in the sixties at the Rand Corporation (MIND system, Kay, 1973). We attribute the relative failure of this approach to the fact that the user felt a *slave* of the machine, that the texts were supposed to be *sacred*, unchangeable, and that the questions asked were at the same time very specialized and quite unsettling. We hope that the time is now ripe for yet another attempt, using latest advances in ergonomy, AI methods for designing intelligent dialogues, and improved linguistic technology. One of the most challenging aspects of that approach is actually the need to express very sophisticated linguistic notions (such as modality, aspect, etc.) in a way understandable by users with no particular training in linguistics or translatology, and no knowledge of the target language(s). Some computer firms are already working on that concept, and may propose products well before the year 2000. But it will be a long time until it is possible to buy off-the-shelf multilingual systems of that kind, because of the tremendous amount of lexical and grammatical variety which is necessary if one does not want to restrict the domain and typology.

It will of course be possible to put a whole system of that kind on a very powerful PC. But an essential ingredient of success, we think, is that the user be never forced to wait, or to answer a question before being allowed to proceed with what he is doing. In other words, the system should simply tell (or better show) that there are some questions waiting to be answered before translation can proceed on some fragments of the text (or hypertext). Then, an attractive solution is to use a comparatively cheap PC as workstation, with a periodic connexion to an MT server (exactly as is done nowadays by e-mail environments).

8.4 Machine-aided Human Translation

Christian Boitet

Université Joseph Fourier, Grenoble, France

Section 8.3 has covered Machine Translation (MT), where translation proper is performed by a computer, even if the human helps by preediting, postediting, or answering questions to disambiguate the source text. In Computer-Aided Translation, or more precisely Machine-Aided Human Translation (MAHT), by contrast, translation is performed by a human, and the computer offers supporting tools.

8.4.1 State of the Art

We can distinguish three types of MAHT systems, corresponding to three types of users, and offering different sets of functionalities.

Specific Software Environments Designed for Professional Translators Working in Teams

Existing products now are those of Trados (MultiTerm), IBM (Translation Manager), and SITE-EuroLang (EuroLang Optimizer). They are available on PC/Windows, PS/OS2, or Unix-based workstations.

The intended users are competent translators working in teams and linked through a local network. Each translator's workstation offers tools to:

- access a bilingual terminology.
- access a *translation memory.*
- submit parts ot the text to an MT server.

These tools have to be completely integrated in the text processor. The software automatically analyzes the source text, and attaches keyboard shortcuts to the terms and sentences found in the terminogical data base and in the translation memory. One very important design decision is whether to offer a specific text processor, as in IBM's Translation Manager, or whether to use directly one or more text processors produced by third parties, as in EuroLang Optimizer.

The server supports tools to:

- manage the common multilingual lexical data base (MLDB), often a multilingual terminological data base (MTDB), and the common translation memory, where previous translations are recorded. Here, concurrent access and strict validation procedures are crucial.
- manage the translation tasks (not always offered).

Let us take the case of the most recent product, EuroLang Optimizer. One instance is available on Sun workstations under Unix. The server uses a standard DBMS (data base management system) (Oracle or Sybase) to support the terminological data base and the translation memory. The translator's workstations use Interleaf or Framemaker as text processors, while their data base functions are degraded versions of those of the servers, and are implemented directly in C++. In the other instance, the server runs on a PC under Windows NT, again with Oracle or Sybase, while the translator's workstations use Word 6 on PCs under Windows 3. Source languages currently include English, French, German, Italian and Spanish. There are 17 target languages (almost all languages written with the Latin character set).

When a document has to be translated, it is preprocessed on the server, and sent to a translator's workstation with an associated *kit*, which contains the corresponding subsets of the dictionary and of the translation memory, as well as (optionally) translation proposals coming from a batch MT system. MAHT-related functionalities are accessible through a supplementary menu (in the case of Word 6) and keyboard shortcuts dynamically associated with terms or full sentences. The translator may enrich the kit's lexicon. When translation is completed, the document is sent back to the server with its updated kit. On the server, the new translation pairs are added to the translation memory, and updates or additions to the dictionary are handled by the (human) manager of the MTDB. The overall productivity of the translators is said to be increased by up to 30% or 40%.

Environments for Independent Professional Translators

These environments are usually less powerful, quite cheaper, and callable from all or at least many commercial text and document processors. This is because free lance translators are usually required to deliver their translations in the same formats as the source documents, and those vary from one customer to the next.

As far as dictionaries are concerned, the situation is different from the preceding case. There is no central MLDB to manage, but it is very important for independent translators to be able to easily create, access, modify, export and import terminological files.

Examples are Mercury/Termex (Melby, 1982) by LinguaTech, a resident program for PCs, and WinTool (Winsoft, 1987), a desk accessory for Macintoshes. In 1992, MicroMATER, an SGML-based standard for PC-oriented terminological dictionaries, has been adopted in relation with ongoing efforts to devise standards for the encoding of more complex dictionary structures within the TEI initiative and in cooperation with InfoTerm (Vienna) and other organizations working on terminology.

Tools for Occasional Translators

An occasional translator may be competent in both languages, or only in the source language! As a matter of fact, there exist tools to help monolinguals produce parametrizable *canned* text in two languages. For example, Ambassador by Language Engineering runs on Macintosh and PC, is available in English-Japanese, English-French, English-Spanish and French-Japanese, and offers about 200 *templates* of letters and forms, and 450 textual *forms* (of sentence or paragraph size).

In the other context, the translator is at least bilingual, but is not a professional, and does not necessarily translate into his native tongue. Even if s/he does, s/he often does not know certain specific terms s/he has learned in the source language (take for example English-Malay or French-Arabic). Tools for bilinguals, such as SISKEP (Tong, 1987), are designed for such users. All are implemented on micros.

These tools offer different functionalities from those for professionals:

- There is no translation memory.
- The dictionaries must contain general terms, and there are usually three dictionary levels: personal and temporary terms, terminology, general vocabulary.
- There are aids concerning the target language (thesaurus, conjugator, style checker, etc.).

Again, it is possible to propose a specific editor, with filters to and from standard word processors, as is done in SISKEP, or to interface the tools directly with one or several word processors. That second course was impractical until a recent past, because developers had to obtain access to the source code of the text processors. This has changed since 1991, when Apple launched version 7 of MacOS, which offers the possibility to let applications communicate through special *events*. The PC world is following with Windows.

8.4.2 Limitations in Current Technology

Serious problems in current technology concern the unavailability of truly multilingual support tools, the engineering of multilingual lexical data bases, the sacred character of the source text, and the limitation to handling only one language pair at a time.

Unavailability of Truly Multilingual Support Tools

MacOS 7.1, available since mid-1992, is still the only operating system supporting any number of writing systems at the same time. With a text processor based on Apple's Script Manager, such as WinText, it is possible to include English, Arabic, Chinese, Japanese, Thai, Russian, etc., in the same document, and to use the writing system as a distinctive feature in search-and-replace actions, or for checking the spelling or the grammar. But, in practice, the size of the OS grows considerably, because it is necessary to include a variety of fonts and input methods. With the languages above, MacOS 7.1 takes 4 to 5 Mbytes of RAM. Input methods and fonts must also often be purchased from third parties.

For other environments, the situation is still very unsatisfactory. At best, it is possible to find localized versions, which handle one *exotic* writing system besides the English one.

Engineering of Multilingual Lexical Data Bases

The multilingual lexical data bases (MLDB) found on MAHT servers are often nothing more than collections of bilingual dictionaries. In the case of terminology proper, MTDBs do exist, but are not yet integrated with MAHT environments. Such MTDBs include, for example, EuroDicautom at the EU (European Commission), Aquila by SITE-Sonovision, and MultiTerm by Trados. In the current state of EuroLang Optimizer, the MTDBs are planned to be monosource and multitarget, but are still bilingual, although the same company continues to market the fully multilingual Aquila on PC LANs.

As far as MLTBs are concerned, then, the problems concern more the management of the data bases than their design. That is because the MTDBs have to evolve constantly, taking into account possibly contradictory or incomplete contributions by many translators. In the case of MLDBs of general terms, there are still many design problems, and available solutions, such as that of EDR in Tokyo, are still too heavy to be used in MAHT systems.

Sacred Character of the Source Text and Limitation to Handling One Language Pair at a Time

Very often, translation is more difficult than it should be because the source text is not well written. If translation has to be performed into several languages, which is often the case, for example for technical manuals, it would make sense to prepare the source text, possibly annotating or rewriting parts of it. That possibility is however not offered in current MAHT systems, and the source texts remain *sacred.*

8.4.3 Future Directions

Current tools will no doubt be improved, in terms of speed, ergonomy and functionalities. Key research issues concern ergonomy, progress in Example-Based MT (EBMT), and integration with Dialogue-Based MT (DBMT).

Ergonomy

It must be realized that accessing large MLDBs and translation memories are very computer intensive operations. To identify complex terms requires full morphological analysis and partial syntactic analysis. Matching a sentence against a large set of sentences and producing a meaningful set of exact or *near* matches is not feasible in real time. The current answers to that problem is to preprocess the documents on a server (or on the workstations, in the background), or, in the case of PC-oriented stand-alone tools for occasional translators, where real time behavior is required, to simplify the morphological analysis and to suppress the translation memory.

The increase of computing power and the object orientation of future operating systems should make it possible to drastically improve the ergonomy and power of MAHT tools, by searching the terminological data base and the translation memory in the background, and dynamically updating MAHT suggestions for the current part of the document being translated, and possibly modified in the source form. These suggestions might appear in MAHT windows logically attached to the windows of the main applications (text processor, spreadsheet, etc.), or, if tighter integration is possible, in its application windows themselves. The main point here is that it would not be necessary to modify the code of the main applications.

Progress in EBMT

Example-Based MT (EBMT) goes one step further than the retrieval of identical or similar sentences. It aims at producing translation proposals by combining the translations of similar chunks of texts making up the sentence and previously identified as possible translation units in the translation memory. It is not yet clear whether the intensive efforts going into that direction will succeed to the point where EBMT could be included in MAHT tools in a cost-effective way.

8.5 Multilingual Information Retrieval

Christian Fluhr

CEA-INSTN, Saclay, France

8.5.1 State of the Art

The problem of multilingual access to text databases can be seen as an extension of the general information retrieval (IR) problem corresponding to paraphrase. How does one retrieve documents containing expressions which do not exactly match those found in the query?

The most traditional approach to IR in general and to multilingual retrieval in particular, uses a controlled vocabulary for indexing and retrieval. In this approach, a documentalist (or a computer program) selects for each document a few descriptors taken from a closed list of authorized terms. Semantic relations (synonyms, related terms, narrower terms, broader terms) can be used to help choose the right descriptors, and solve the sense problems of synonyms and homographs. The list of authorized terms and semantic relations between them are contained in a thesaurus.

To implement multilingual querying using this approach, it is necessary to give the corresponding translation of each thesaural term for each new language recognized. This work is facilitated by the fact each descriptor is not chosen randomly but in order to express a precise unambiguous concept. The CIENTEC term bank (Velho Lopes, 1989) is one of many multilingual projects adopting this approach.

A problem remains, however, since concepts expressed by one single term in one language sometime are expressed by distinct terms in another. For example, the common language term *mouton* in French is distinguished into two different concepts in English, *mutton* and *sheep*. One solution to this problem, given that these distinctions are known between the languages implemented is to create pseudo-words such as *mouton (alimentation)*—mutton, and *mouton (animal)*—sheep. These domain semantic tags (such as *animal* and *alimentation*) as well as the choice of transfer terms depend on the final use of the multilingual thesaurus, and it is therefore sometimes easier to build a multilingual thesaurus from scratch rather than to adapt a monolingual one.

This controlled vocabulary approach gives acceptable results but prohibits precise queries that cannot be expressed with these authorized keywords. It is however a common approach in well-delimited fields for which multilingual thesauri already exist (legal domain, energy, etc.) as well as in multinational organizations or countries with

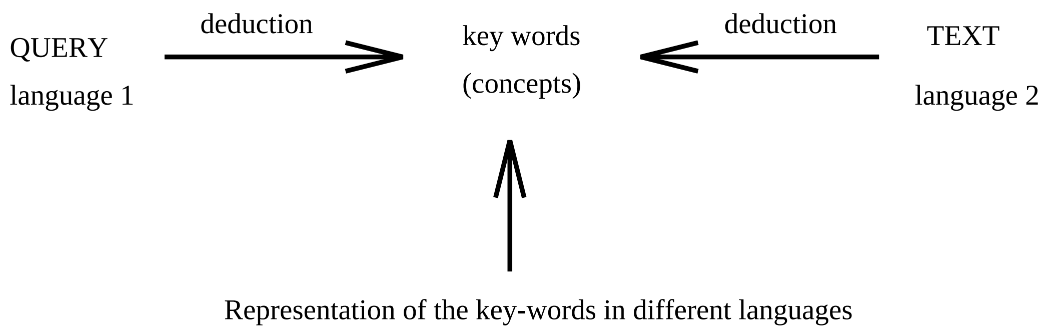

Figure 8.1: Multilingual Interrogation using interlingual pivot concepts.

several official languages, which contain lexicographical units familiar with problems of terminological translation.

Automatization of such methods consists in deducing, during indexing, the key-words that would be supplied for a text from the terms contained in the full-text or summary. Links between full-text words and controlled descriptors can be constructed either manually or by an automatic learning process from previously indexed documents. During interrogation, the same process can deduce the key-words from the terms used in the query to produce a search request. If links between text words and key-words are pre-established using different languages, it is possible to interrogate texts that are not in the same language as the query using the key-words as a pivot language. See figure 8.1.

Generally, the controlled vocabulary approach means that queries can only be as precise as the predefined key-words (i.e., concepts) present in the thesaurus, posing an upper limit on query precision.

A third approach to multilingual interrogation is to use existing machine translation (MT) systems to automatically translate the queries, or even the entire textual database from one language to another. When only queries are translated from a source to target language, text can be searched in the target language and results can be dynamically translated back to the source language as they are displayed after the search.

This kind of method would be satisfactory if current MT systems did not make errors.

A certain amount of syntactic error can be accepted without perturbing results of information retrieval systems, but MT errors in translating concepts can prevent relevant documents, indexed on the missing concepts, from being found. For example, if the word *traitement* in French is translated by *processing* instead of *salary,* the retrieval process would yield wrong results.

This drawback is limited in MT systems that use huge transfer lexicons of noun phrases like the RETRANS system developed by Belogonov, Khoroshilov, et al. (1993) in the VINITI, Moscow. But in any collection of text, ambiguous nouns will still appear as isolated nouns phrases untouched by this approach.

A fourth approach to multilingual information retrieval is based on the Salton's vector space model (Salton & McGill, 1983). This model represents documents in a n-dimensional space (n being the number of different words in the text database). If some documents are translated into a second language, these documents can be observed both in the subspace related to the first language and the subspace related to the second one. Using a query expressed in the second language, the most relevant documents in the translated subset are extracted (usually using a cosine measure of proximity). These relevant documents are in turn used to extract close untranslated documents in the subspace of the first language.

An improvement to this approach using existing translations of a part of the database has been investigated by a team in Bellcore (Landauer & Littman, 1990). Their information retrieval is based on latent semantic indexing. They approximate the full word-document matrix by a product of three lower dimensionality matrices of orthogonal factors derived by singular value decomposition. This transformation enables them to make a comparison not using individual words but taking into account sets of semantically related words. This approach use implicit dependency links and co-occurrences that better approximate the notion of concept.

The method has been tested with some success on the English-French language pair using a sample of the Canadian Parliament bilingual corpus. 2482 paragraphs were selected. 900 were used for training, using both the English and French words in the documents to build the matrices. The 1582 remaining documents were add to the matrices in their French version only. The English versions of these 1582 documents were then used as queries using the 900 English documents of the training set to relate the French and English words in the latent semantic indexing. For 92% of the English text documents the closest document returned by the method was its correct French translation.

Such an approach presupposes that the sample used for training is really representative of the full database. Translation of the sample remains a huge undertaking that must be done for each new database.

Still another approach consists combining machine translation methods with information retrieval methods. This approach has been developed by a European ESPRIT consortium (French, Belgian, German) in the project EMIR (European Multilingual Information Retrieval) (EMIR, 1994). Experiments have been performed on French, English and German. This system uses 3 main tools:

- linguistic processors (morphological and syntactic analysis) which perform grammatical tagging, identify dependency relations (especially within noun phrases), and normalize the representation of uniterms and compounds;
- a statistical model which is used to weight the query-document intersection;
- a monolingual and multilingual reformulation system whose aim is to infer, from the original natural language query words, all possible expressions of the same concept that can occur in the document whatever the language.

The EMIR (1994) system uses large monolingual and bilingual dictionaries enabling it to process full-text databases in any domain. That means that all possible ambiguity in the language from both the syntactic and the semantic point of view are taken into account. A few additions are needed for unseen technical domains in the monolingual and bilingual dictionaries, especially in the bilingual dictionaries of multiterms.

Database texts are processed by linguistic processors which normalize single words and compounds. A weight is computed for all normalized words using a statistical model (Debili, Fluhr, et al., 1989). During the interrogation the text which is used as a query undergoes the same linguistic processing. The result of this processing is passed to the reformulation process which infers new terms using monolingual reformulation rules (on source language and/or target language) and bilingual reformulation rules (transfer) (Fluhr, 1990). Compounds that are translated word for word are restructured by transformational rules. It can be seen that this approach differs significantly to the MT approach where only one translation of each query word is used. EMIR uses all possible translations in its database search.

In such an approach training for each database is not needed. Experiments on different databases have shown that, in most cases, the translation ambiguities (often more than 10 for each word) are solved by a comparison with the database lexicon and the co-occurrence with the translations of the other concepts of the query. Implicit semantic information contained in the database text is used as semantic filter to find the right translation in cases where current MT systems would not succeed.

In the framework of EMIR, tests have been been performed on the English CRANFIELD information retrieval testbed. First the original English language queries were translated into French by domain experts. Then two approaches were tested. Querying using the French-to-English SYSTRAN translation followed by a monolingual search was compared to querying using the first bilingual EMIR prototype to access English text by expanding the French queries into English possibilities. The multilingual EMIR interrogation was 8% better than using SYSTRAN followed by monolingual interrogation. On an other hand monolingual interrogation using the original English queries with monolingual EMIR was 12% better than the bilingual interrogation.

8.5.2 Future Directions

To continue research in the domain of multilingual information retrieval it is necessary to develop tools and textual data resources whose construction will be costly. Apart from the need for tools that are needed in all or most areas of natural language research, we see the need for the following:

Large bilingual test corpora are urgently needed in order to evaluate and compare methods in an objective manner. Existing test databases are monolingual, mainly in English. Large-scale test databases which are truly multilingual (i.e., with texts which are strict translations of each other) are needed. It will then be necessary to elaborate a set of queries in the various languages tested as well as to find all the relevant document for each query. This is a huge task. Such an undertaking for English textual database has begun in the TREC (Text Retrieval Evaluation Conference) project (Harman, 1993). A similar process needs to be put in motion for multilingual test databases.

Databases of lexical semantic relations as general as possible are needed in a variety of languages for monolingual reformulation using classical relations like synonyms, narrower terms, broader terms and also more precise relations like *part of*, *kind of*, *actor of the action*, *instrument of the action*, etc., such as is being created for English in WordNet (Miller, 1990). Bilingual transfer dictionaries should also be as general as possible (general language as well as various specific domains).

To accelerate the construction of such lexicons, tools are needed for extracting terminology and for automatic construction of the semantic relations from corpora of texts. If bilingual corpus of texts are available in a domain, tools for computer aided building of transfer dictionaries should be developed. This extraction is specially needed for recognizing translations of compounds.

8.6 Multilingual Speech Processing

Alexander Waibel

Carnegie-Mellon University, Pittsburgh, Pennsylvania, USA
and Universität des Karlsruhe, Germany

Multilinguality need not be textual only, but will take on spoken form, when information services are to extend beyond national boundaries, or across language groups. Database access by speech will need to handle multiple languages to service customers from different language groups within a country or travelers from abroad. Public service operators (emergency, police, department of transportation, telephone operators, and others) in the US, Japan and the EU frequently receive requests from foreigners unable to speak the national language (see also section 8.7.1).

Multilingual spoken language services is a growing industry, but so far these services rely exclusively on human operators. Telephone companies in the United States (e.g., AT&T Language Line), Europe and Japan now offer language translation services over the telephone, provided by human operators. Movies and foreign television broadcasts are routinely translated and delivered either by lipsynchronous speech (dubbing), subtitles or multilingual transcripts. The drive to automate information services, therefore, produces a growing need for automated multilingual speech processing.

The difficulties of speech processing are compounded with multilingual systems, and few if any commercial multilingual speech services exist to date. Yet intense research activity in areas of potential commercial interest are underway. These are aiming at:

- **Spoken Language Identification** By determining a speaker's language automatically, callers could be routed to human translation services. This is of particular interest to public services such as police, government offices (immigration service, drivers license offices, etc.) and experiments are underway in Japan and some regions of the US. The technical state of the art will be reviewed in the next section;

- **Multilingual Speech Recognition and Understanding** Future Spoken Language Services could be provided in multiple languages. Dictation systems and spoken language database access systems, for example, could operate in multiple languages, and deliver text or information in the language of the input speech.

- **Speech Translation** This ambitious possibility is still very much a research area, but could eventually lead to communication assistance in the form of portable voice activated dictionaries, phrase books or spoken language translators, telephone based speech translation services and/or automatic translation of foreign broadcasts and speeches. There is a wide spectrum of possibilities, but their full realization as commercial products still requires considerable research well into the next decade and beyond.

8.6.1 Multilingual Speech Recognition and Understanding

The last decade has seen much progress in performance of speech recognition systems from cumbersome small vocabulary isolated word systems to large vocabulary continuous speech recognition (LV-CSR) over essentially unlimited vocabularies (50,000 words and more). Similarly, spoken language understanding systems now exist that process spontaneously spoken queries, although only in limited task domains under benign recording conditions (high quality, single speaker, no noise). A number of researchers have been encouraged by this state of affairs to extend these systems to other languages. They have studied similarities as well as differences across languages and improved the universality of current speech technologies.

Large Vocabulary Continuous Speech Recognition (LV-CSR)

A number of LV-CSR systems developed originally for one language have now been extended to several languages, including systems developed by IBM (Cerf-Danon, DeGennaro, et al., 1991), Dragon Systems (Bamberg, Demedts, et al., 1991), Philips and Olivetti (Ney & Billi, 1991) and LIMSI. The extension of these systems to English, German, French, Italian, Spanish, Dutch and Greek illustrates that current speech technology does generalize to different languages, provided sufficiently large transcribed speech databases are available. The research results show that similar modeling assumptions hold across languages with a few interesting exceptions. Differences in recognition performance are observed across languages, partially due to greater acoustic confusability (e.g., English), greater number of homonyms (e.g., French) and greater number of compound nouns and inflections (e.g., German). Such differences place a different burden on acoustic modeling vs. language modeling, vs. the dictionary, or increase confusability, respectively. Also, a recognition vocabulary is not as easily defined as a unit for processing in languages such as Japanese and Korean, where pictographs, the absence of spaces, and large numbers of particles complicate matters.

Multilingual Spoken Language Systems

While LV-CSR systems tackle large vocabularies, but assume benign speaking styles (read speech), spoken language systems currently assume smaller domains and vocabularies, but require unrestricted speaking style. Spontaneous speech significantly degrades performance over read speech as it is more poorly articulated, grammatically ill-formed and garbled by noise. ARPA's Spoken Language projects have attacked this problem by focusing increasingly on the extraction of the semantic content of an utterance rather than accurate transcription. One such system, that has recently been extended to other languages is MIT's Voyager system (Glass, Goodine, et al., 1993). It was designed to handle information delivery tasks and can provide directions to nearby restaurants in Cambridge and also for airline travel information (ATIS). It has recently been extended to provide output in languages other than English. Researchers at LIMSI have developed a similar system for French (also airline travel information), thereby providing extension to French on the input side as well. Availability of recognition capabilities in multiple languages have also recently led to interesting new language, speaker and gender identification strategies (Gauvain & Lamel, 1993). Transparent language identification could enhance the application of multilingual spoken language systems (see also section 8.7.1).

Despite the encouraging beginnings, multilingual spoken language systems still have to be improved before they can be deployed on a broad commercially feasible scale. Prototype systems have so far only been tested in benign recording situations, on very limited domains, with cooperative users, and without significant noise. Extending this technology to field situations will require increases in robustness as well as consideration of the human factors aspects of multilingual interface design.

8.6.2 Speech Translation Systems

There are no commercial speech translation systems in operation to date, but a number of industrial and government projects are exploring their feasibility. The feasibility of speech translation depends largely on the scope of the application, and ranges from applications that are well within range -such as voice activated dictionaries- to those that will remain impossible for the foreseeable future (e.g., unrestricted simultaneous translation.) Current research therefore aims at milestones between these extremes, namely limited domain speech translation. Such systems restrict the user in what he/she can talk about, and hence constrain the otherwise daunting task of modeling the world of discourse. Nevertheless such systems could be of practical and commercial interest, as they could be used to provide language assistance in common yet critical situations, such as registration for conferences, booking hotels, airlines, car rentals and theater tickets, ordering food, getting directions, scheduling meetings or in medical doctor-patient situations. If successful, it may also be possible to combine such domains to achieve translation in a class of domains (say, travel).

To be sure, spoken language translation—even in limited domains—still presents considerable challenges, which are the object of research in several large research undertakings around the world. Translation of spoken language (unlike text) is complicated by syntactically ill-formed speech, human (cough, laughter, etc.) and non-human (door-slams, telephone rings, etc.) noise, and has to contend with speech recognition errors. The spoken utterance does not provide unambiguous markers indicating the beginning or end of a sentence or phrase, and it frequently contains irrelevant information, that need not or should not be translated. Even simple concepts are expressed in quite different ways in different languages. A successful system must therefore interpret the speaker's intent -instead of translating his/her words- and deliver an appropriate message in the target language. For the speech processing components of a speech recognition system high accuracy is not the primary or only area of concern, but understanding, and understanding may be achieved by selectively extracting words of interest, and/or by occasionally prompting the user for important information. Researchers are now exploring solutions to the problem as a whole without expecting each separate part to function perfectly.

A speech translation system can also not be categorized uniquely as either "translation for assimilation" nor as "translation for dissemination", as textual translation systems are frequently described. It has some of the characteristics of both. Aiming at the interpretation of a speaker's intent, some research avenues in speech translation are attempting to extract (assimilate) the key information to interpret the gist of an utterance. Yet spoken language in many of the targeted application scenarios involves the interaction between two cooperative speakers, who can control to some extent the input to produce the desired result. This may allow for some limited domain systems to interact with the speaker of the source language until the correct interpretation can be transmitted (disseminated) in the target language(s).

A further complicating factor currently under investigation is that speech translation involves aspects of both human-to-human, as well as human-machine (the interpreting system) dialogues. This may require a system to distinguish between utterances and meta-level utterances, and to deal with code switching (change of language) in case of speakers with partial knowledge of each others' language or when making reference to objects, names or items in the other language. Experiments over several speech databases in several languages indicate that human-to-human speech contains more disfluencies, more speaking rate variations and more coarticulation resulting in lower recognition rates (Levin, Suhm, et al., 1994) than human-machine interaction. These difficulties require further technological advances, a rethinking of common speech and language processing strategies, and a closer coupling between the acoustic and linguistic levels of processing.

Early Systems: Speech Translation research today is being developed against the background of early systems implemented in the eighties to demonstrate the feasibility of the concept. In addition to domain limitations, these early systems had also fixed speaking style, grammatical coverage and vocabulary size and were therefore too limited to be of practical value. Their system architecture is usually strictly sequentially, involving speech recognition, language analysis and generation, and speech synthesis in the target language. Developed at industrial and academic institutions and consortia, they represented a modest but significant first step and proof of concept that multilingual communication by speech might be possible. Systems include research prototypes developed by NEC, AT&T, ATR, Carnegie Mellon University, Siemens AG, University of Karlsruhe, and SRI. Most have arisen or been made possible through international collaborations that provide the cross-linguistic expertise.

Among these international cooperations, the Consortium for Speech TrAnslation Research (C-STAR) was formed as a voluntary group of institutions committed to building speech translation systems. Its early members, ATR Interpreting Telephony Laboratories (now "Interpreting Telephony Laboratories") in Kyoto, Japan, Siemens AG in Munich, Germany, Carnegie Mellon University (CMU) in Pittsburgh, USA, and University of Karslruhe (UKA) in Karlsruhe, Germany, developed early systems, that accepted speech in each of the members' languages (i.e., English, German and Japanese) and produced output text in all the others (Morimoto, Takezawa, et al., 1993; Waibel, Jain, et al., 1991; Woszczyna, Aoki-Waibel, et al., 1994). The system modules allowed for continuous speaker-independent (or adaptive) input from a 500 word vocabulary in the domain of conference registration. The systems' modules operated strictly sequential, did not allow for feedback, and only accepted syntactically well formed utterances. After speech recognition, language analysis and generation, output text could then be transmitted to each of the partners sites for synthesis there. Translation was performed by an Interlingua approach in JANUS, the CMU/UKA system, while a transfer approach was used in ATR's ASURA and Siemens's systems. In early '93, they were shown to the public in a joint demonstration using video conferencing. Given the restrictions on speaking style and vocabulary, the systems performed well and provided good translation accuracy.

Early industrial speech-translation efforts are illustrated by AT&T's VEST (Roe, Pereira, et al., 1992) and NEC's Intertalker systems. VEST resulted from a collaboration between AT&T and Telefonica in Spain and translated English and Spanish utterances about currency exchange. It uses a dictionary of 374 morphological entries and an augmented phrase structure grammar that is compiled into a finite state grammar used for both language modeling and translation. The system was demonstrated at EXPO'92 in Seville, Spain. NEC's Intertalker system also used finite state grammars to decode input sentences in terms of prescribed sentence patterns. The system ran on two separate tasks: reservation of concert tickets and travel information, and was successfully demonstrated at GlobCom'92. SRI in collaboration with Swedish Telecom recently reported on another system (Rayner et al., 1993), that is based on previously developed system components from SRI's air travel information system. The ATIS speech understanding component is interfaced with a generation component. The system's input language is English and it produces output in Swedish. It represents an early attempt at extending spontaneous multilingual human-machine dialogues to translation.

Translation of Spontaneous Speech: To develop more practical, usable speech translation, greater robustness in the face of spontaneous ill-formed speech has to be achieved. A number of research activities aiming at the translation of spontaneous speech have since been launched. Several industrial and academic institutions, as well as large national research efforts in Germany and in Japan are now working on this problem. Virtually all of these efforts aim at restricted domains, but now remove the limitation of a fixed vocabulary and size, and also no longer require the user to speak in syntactically well-formed sentences (an impossibility in practice, given stuttering, hesitations, false starts and other disfluencies found in spontaneous speech).

The C-STAR consortium was extended to translate spontaneous speech. In addition to the partners of the first phase, it includes presently ETRI (Korea), IRST (Italy), LIMSI (France), SRI (UK), IIT (India), Lincoln Labs (USA), MIT (USA), and AT&T (USA). Each C-STAR partner builds a complete system that at the very least accepts input in the language of this partner and produces output in one other language of the consortium. In a multinational consortium, building full systems thereby maximizes the technical exchange between the partners while minimizing costly software/hardware interfacing work. C-STAR continues to operate in a fairly loose and informal organizational style. Present activity has shifted toward a greater emphasis on interpretation of spoken language, i.e., the systems ability to extract the intent of a speakers utterance. Several institutions involved in C-STAR therefore stress semantic parsers and an interlingual representation (CMU, UKA, MIT, ATT, ETRI, IRST), more in line with message extraction than with traditional text translation. Other approaches under investigation include Example Based Translation (ATR), with its potential for improved portability and reduced development cost through the use of large parallel corpora. Robust Transfer Approaches (ATR, Siemens) are also explored, with robust and stochastic analysis to account for fragmentary input. System architectures under investigation are no-longer strictly sequential, but begin to involve clarification or paraphrase in the speaker's language as first attempts at the machine's feedback of its understanding. At the time of this writing, such feedback is still very rudimentary and does not yet involve more elaborate confirmatory meta-level dialogues or repair mechanisms. Current research also begins to actively exploit discourse and domain knowledge, as well as prosodic information during turn taking, for more robust interpretation of ambiguous utterances.

Verbmobil is a large new research effort sponsored by the BMFT, the German Ministry for Science and Technology (Wahlster, 1993). Launched in 1993 the program sponsors over 30 German industrial and academic partners who work on different aspects of the speech translation problem and are delivering system components for a complete speech translation system. The system components (e.g., speech recognition components, analysis, generation, synthesis, etc.) are integrated into a research prototype, available to all. The initial task is appointment scheduling with possible extensions to other domains. Verbmobil is aimed at face-to-face negotiations, rather than telecommunication applications and assumes that two conversants have some passive knowledge of a common language, English. It is to provide translation on demand for speakers of German and Japanese, when they request assistance in an otherwise English conversation. Verbmobil is therefore concerned with code switching and the translation of sentence fragments in a dialog. Verbmobil is an eight-year project with an initial four-year phase.

8.6.3 Future Directions

To meet the challenges in developing multilingual technology, an environment and infrastructure must be developed. Contrary to research fostered and supported at the national level, multilingual research tends to involve cooperations across national boundaries. It is important to define and support efficient, international consortia, that agree to jointly develop such mutually beneficial technologies. An organizational style of cooperation with little or no overhead is crucial, involving groups who are in a position to build complete speech translation systems for their own language. There is a need for common multilingual databases and data involving foreign accents. Moreover, better evaluation methodology over common databases is needed to assess the performance of a speech translation systems in terms of accuracy and usability. Research in this direction needs to be supported more aggressively across national boundaries.

Beyond improvements in component technologies (speech and language processing), innovations in language acquisition are badly needed to achieve greater portability across domains. While acoustic models can be reused to a certain extent (or at least adapted) across domains, most language work still requires inordinate amounts of resources. Grammar development requires considerable development work for each domain. Language models have to be retrained and require large amounts of transcribed data within each domain. Continued research on language acquisition may provide better domain adaptation, and/or incrementally improving language models, grammars and dictionaries.

The limitation to restricted domains of discourse must be lifted, if broader usage is to be guaranteed. Short of universal and reliable speech translation (as could be needed for example, for automatically translated captions in movies, or simultaneous translation), intermediate goals might be given by large domains of discourse, that involve several subdomains. Integration of subdomains will need to be studied.

Last, but not least, better human-computer interaction strategies have to be developed, as multilingual spoken language translation becomes a tool to broker an understanding between two humans rather than a black box that tries to translate every utterance. A useful speech translation system should be able to notice misunderstandings and negotiate alternatives. Such ability requires better modeling of out of domain utterances, better generation of meta-level dialogues and handling of interactive repair.

8.7 Automatic Language Identification[1]

Yeshwant K. Muthusamy[a] **& Lawrence Spitz**[b]

[a] Texas Instruments Incorporated, Dallas, Texas, USA
[b] Daimler Benz Research and Technology Center, Palo Alto, California, USA

8.7.1 Spoken Language

The importance of spoken language ID in the global community cannot be ignored. Telephone companies would like to quickly identify the language of foreign callers and route their calls to operators who can speak the language. A multilanguage translation system dealing with more than two or three languages needs a language identification front-end that will route the speech to the appropriate translation system. And, of course, governments around the world have long been interested in spoken language ID for monitoring purposes.

Despite twenty-odd years of research, the field of spoken language ID has suffered from the lack of (i) a common, public-domain multilingual speech corpus that could be used to evaluate different approaches to the problem, and (ii) basic research. The recent public availability of the OGI Multilanguage Telephone Speech Corpus (OGI_TS) (Muthusamy, Cole, et al., 1992), designed specifically for language ID, has led to renewed interest in the field and fueled a proliferation of different approaches to the problem. This corpus currently contains spontaneous and fixed vocabulary speech from 11 languages. The National Institute of Standards and Technology (NIST) conducts an annual common evaluation of spoken language ID algorithms using the OGI_TS corpus. At the time of writing, eight research sites from the U.S. and Europe participate in this evaluation. There are now papers on spoken language ID appearing in major conference proceedings (Berkling & Barnard, 1994; Dalsgaard & Andersen, 1994; Hazen & Zue, 1994; Kadambe & Hieronymus, 1994; Lamel & Gauvain, 1994; Li, 1994; Ramesh & Roe, 1994; Reyes, Seino, et al., 1994; Zissman & Singer, 1994). See Muthusamy, Barnard, et al. (1994) for a more detailed account of the recent studies in spoken language ID.

Many of the approaches to spoken language ID have adopted techniques used in current speaker-independent speech recognition systems. A popular approach to language ID consists of variants of the following two basic steps: (i) develop a phonemic/phonetic recognizer for each language, and (ii) combine the acoustic likelihood scores from the

[1] Automatic language identification (language ID for short) can be defined as the problem of identifying the language from a sample of speech or text. Researchers have been working on spoken and written language ID for the past two decades.

recognizers to determine the highest scoring language. Step (i) consists of an acoustic modeling phase and a language modeling phase. Trained acoustic models of phones in each language are used to estimate a stochastic grammar for each language. The models can be trained using either HMMs (Lamel & Gauvain, 1994; Zissman & Singer, 1994) or neural networks (Berkling & Barnard, 1994). The grammars used are usually bigram or trigram grammars. The likelihood scores for the phones resulting from step (i) incorporate both acoustic and phonotactic information. In step (ii), these scores are accumulated to determine the language with the largest likelihood. Zissman and Singer (1994) have achieved the best results to date on OGI_TS using a slight variant of this approach: The exploits the fact that a stochastic grammar for one language can be developed based on the acoustic models of a different language. This has the advantage that phonetic recognizers need not be developed for all the target languages. This system achieves 79% accuracy on the 11-language task using 50-second utterances and 70% accuracy using 10-second utterances.

Li (1994) has applied speaker recognition techniques to language ID with tremendous success. His basic idea is to classify an incoming utterance based on the similarity of the speaker of that utterance with the most similar speakers of the target languages. His similarity measure is based on spectral features extracted from experimentally determined syllabic nuclei within the utterances. His results on the 11-language task: 78% on 50-second utterances, and 63% on 10-second utterances.

The importance of prosodic information such as pitch and duration in recognizing speech or in discriminating between languages has long been acknowledged. However, this information has not yet been fully exploited in language ID systems. Muthusamy (1993) examined pitch variation within and across broad phonetic segments with marginal success. He found other prosodic information such as duration and syllabic rate to be more useful, as did Hazen and Zue (1994).

While the progress of language ID research in the last two years has been heartening, there is much to do. It is clear that there is no "preferred approach" as yet to spoken language ID; very different systems perform comparably on the 11-language task. Moreover, the level of performance is nowhere near acceptability in a real-world environment. Present systems perform much better on 50-second utterances than 10-second ones. The fact that human identification performance asymptotes for much shorter durations of speech (Muthusamy, Jain, et al., 1994) indicates that there are some important sources of information that are not being exploited in current systems.

8.7.2 Written Language

Written language identification has received less attention than spoken language recognition. House and Neuberg (1977) demonstrated the feasibility of written language ID using just broad phonetic information. They trained statistical (Markov) models on sequences of broad phonetic categories derived from phonetic transcriptions of text in eight languages. Perfect discrimination of the eight languages was obtained. Most methods rely on input in the form of character codes. Techniques then use information about short words (Kulikowski, 1991; Ingle, 1991); the independent probability of letters and the joint probability of various letter combinations (Rau, 1974 who used English and Spanish text, to devise an identification system for the two languages); n-grams of words (Batchelder, 1992); n-grams of characters (Beesley, 1988; Cavner & Trenkle, 1994); diacritics and special characters (Newman, 1987); syllable characteristics (Mustonen, 1965), morphology and syntax (Ziegler, 1991).

More specifically, Heinrich (1989) evaluated two language ID approaches (one using statistics of letter combinations and the other using word rules) to help him convert French and English words to German in a German text-to-speech system. He found that the approach based on word-boundary rules, position independent rules (e.g., 'sch' does not occur in French) and exception word lists was more suited to the conversion task and performed better than the one based on statistics of letters, bigrams and trigrams. His experiments, however, did not use an independent test set.

Schmitt (1991) patented a trigram-based method of written language ID. He compared the successive trigrams derived from a body of text with a database of trigram sets generated for each language. The language for which the greatest number of trigram matches were obtained, and for which the frequencies of occurrence of the trigrams exceeded a language-specific threshold, was chosen the winner. No results were specified.

Ueda and Nakagawa (1990) evaluated multi-state ergodic (i.e., fully connected) HMMs, bigrams and trigrams to model letter sequences using text from six languages. Their experiments revealed that the HMMs had better entropy than bigrams but were comparable to the computationally expensive trigrams. A 7-state ergodic HMM, in which any state can be visited from any other state, provided 99.2% identification accuracy on a 50-letter test sequence.

Judging by the results, it appears that language ID from character codes is a less hard problem than that from speech input. This makes intuitive sense: text does not exhibit the variability associated with speech (e.g., speech habits, speaker emotions, mispronunciations, dialects, channel differences, etc.) that contributes to the problems in speech recognition and spoken language ID.

More and more text is, however, only available as images, to be converted into possible character sequences by OCR. However, for OCR it is desirable to know the language of the document before trying the decoding. More recent techniques try to determine the language of the text before doing the conversions. The Fuji Xerox Palo Alto Laboratory (Spitz, 1993) developed a method of encoding characters into a small number of basic character shape codes (CSC), based largely on the number of connected components and their position with respect to the baseline and x-height. Thus characters with ascenders are represented differently from those with descenders and in turn from those which are entirely contained between the baseline and x-line. A total of 8 CSCs represent the 52 basic characters and their diacritic forms.

On the basis of different agglomerations of CSCs, a number of techniques for determining the language of a document have been developed. Early work used word shape tokens (WSTs) formed by one-to-one mappings of character positions within a word to character shape codes. Analysis of the most frequently occurring WSTs yields a highly reliable determination of which of 23 languages, all set in Roman type, is present (Sibun & Spitz, 1994). More recent work uses the statistics of n-grams of CSCs (Nakayama, 1994).

8.7.3 Future Directions

A number of fundamental issues need to be addressed if progress is to be made in spoken language ID (Cole, Hirschman, et al., 1995). Despite the flattering results on OGI_TS, current studies have not yet addressed an important question: what are the fundamental acoustic, perceptual, and linguistic differences among languages? An investigation of these differences with a view to incorporating them into current systems is essential. Further, is it possible to define language-independent acoustic/phonetic models, perhaps in terms of an interlingual acoustic/phonetic feature set? An investigation of language-specific versus language-independent properties across languages might yield answers to that question. As for written language ID, languages using non-Latin and more general non-alphabetical scripts are the next challenge.

8.8 Chapter References

ACL (1991). *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California. Association for Computational Linguistics.

Alshawi, H., Carter, D., et al. (1991). Translation by quasi logical form transfer. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California. Association for Computational Linguistics.

ANLP (1994). *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Stuttgart, Germany. ACL, Morgan Kaufmann.

Arnold, D. and des Tombes, L. (1987). Basic theory and methodology in EUROTRA. In Nirenburg, S., editor, *Machine Translation: Theoretical and Methodological Issues*, pages 114–135. Cambridge University Press.

ARPA (1993). *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Bamberg, P., Demedts, A., Elder, J., Huang, C., Ingold, C., Mandel, M., Manganaro, L., and van Even, S. (1991). Phonem-based training for large-vocabulary recogntition in six european languages. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, volume 1, pages 175–181, Genova, Italy. European Speech Communication Association.

Batchelder, E. O. (1992). A learning experience: Training an artificial neural network to discriminate languages. Technical Report.

Beesley, K. R. (1988). Language identifier: A computer program for automatic natural-language identification on on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, pages 47–54.

Belogonov, G., Khoroshilov, A., Khoroshilov, A., Kuznetsov, B., Novoselov, A., Pashchenko, N., and Zelenkov, Y. (1993). An interactive system of Russian-English and English-Russian machine translation of polythematic scientific and technical texts. Technical report, VINITI internal Report, Moscow.

Berkling, K. M. and Barnard, E. (1994). Language identification of six languages based on a common set of broad phonemes. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1891–1894, Yokohama, Japan.

Boitet, C. (1986). The French national MT-project: technical organization and translation results of CALLIOPE-AERO. *Computers and Translation*, 1:281.

Boitet, C. and Blanchon, H. (1993). Dialogue-based machine translation for monolingual authors and the LIDIA project. In Nomura, H., editor, *Proceedings of the 1993 Natural Language Processing Rim Symposium*, pages 208–222, Fukuoka. Kyushu Institute of Technology.

Bourbeau, L., Carcagno, D., Goldberg, E., Kittredge, R., and Polguere, A. (1990). Bilingual generation of wheather forecasts in an operations environment. In Karlgren, H., editor, *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 318–320, Helsinki. ACL.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Carbonell, J. G. and Tomita, M. (1987). Knowledge-based machine translation, the CMU approach. In Nirenburg, S., editor, *Machine Translation: Theoretical and Methodological Issues*, pages 68–89. Cambridge University Press.

Cavner, W. B. and Trenkle, J. M. (1994). N-gram based text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 261–169.

Cerf-Danon, H., DeGennaro, S., Ferreti, M., Gonzalez, J., and Keppel, E. (1991). Tangora—a large vocabulary speech recognition system for five languages. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, volume 1, pages 183–192, Genova, Italy. European Speech Communication Association.

Chandioux, J. (1989). Meteo: 1000 million words later. In Hammond, D. L., editor, *American Translators Association Conference 1989: Coming of Age*, pages 449–453. Learned Information, Medford, New Jersey.

Chandler, B., Holden, N., Horsfall, H., Pollard, E., and McGee, M. W. (1987). N-tran final report, Alvey project. Technical Report 87/9, CCL/UMIST, Manchester.

Chevalier, M., Dansereau, J., et al. (1978). *TAUM-METEO: Description du Système.* Université de Montréal.

Cole, R. A., Hirschman, L., Atlas, L., Beckman, M., Bierman, A., Bush, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D., Ostendorf, M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., and Zue, V. (1995). The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21.

COLING (1986). *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn. ACL.

COLING (1988). *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest.

COLING (1992). *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France. ACL.

Dalsgaard, P. and Andersen, O. (1994). Application of inter-language phoneme similarities for language identification. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1903–1906, Yokohama, Japan.

Debili, F., Fluhr, C., and Radasao, P. (1989). About reformulation in full-text IRS. *Information Processing and Management*, 25:647–657.

Delin, J., Hartley, A., Paris, C., Scott, D., and Vander Linden, K. (1994). Expressing procedural relationships in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 61–70, Kennebunkport, Maine. Springer-Verlag, Berlin.

EMIR (1994). Final report of the EMIR project number 5312. Technical report, European Multilingual Information Retrieval Consortium For the Commission of the European Union, Brussels.

Eurospeech (1991). *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, Genova, Italy. European Speech Communication Association.

Farwell, D. and Wilks, Y. (1990). *Ultra: A Multi-lingual Machine Translator*. New Mexico State University.

Fluhr, C. (1990). Multilingual information. In *AI and Large-Scale Information*, Nagoya.

Furuse, O. and Iida, H. (1992). Cooperation between transfer and analysis in example-based framework. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France. ACL.

Gauvain, J.-L. and Lamel, L. F. (1993). Identification of non-linguistic speech features. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, page Session 6, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Glass, J., Goodine, D., Phillips, M., Sakai, S., Seneff, S., and Zue, V. (1993). A bilingual voyager system. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann. Session 6.

Harman, D., editor (1993). *National Institute of Standards and Technology Special Publication No. 500-207 on the The First Text REtrieval Conference (TREC-1)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Hazen, T. J. and Zue, V. W. (1994). Recent improvements in an approach to segment-based automatic language identification. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1883–1886, Yokohama, Japan.

Heinrich, P. (1989). Language identification for automatic grapheme-to-phoneme conversion of foreign words in a german text-to-speech system. In *Speech-89*, pages 220–223.

House, A. S. and Neuberg, E. P. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3):708–713.

Huang, X. M. (1990). A machine translation system for the target language inexpert. In Karlgren, H., editor, *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 364–367, Helsinki. ACL.

ICASSP (1994). *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

ICSLP (1994). *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, Japan.

Iida, E. S. and Iida, H. (1991). Experiments and prospects of example-based machine translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 185–192, Berkeley, California. Association for Computational Linguistics.

Ingle, N. C. (1991). A language identification table. *The Incorporated Linguist*, 15(4):98–101.

IWNLG (1994). *Proceedings of the Seventh International Workshop on Natural Language Generation*, Kennebunkport, Maine. Springer-Verlag, Berlin.

JEIDA (1989). A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, USA. Technical report, Japanese Electronic Industry Development Association, Tokyo.

Kadambe, S. and Hieronymus, J. L. (1994). Spontaneous speech language identification with a knowledge of linguistics. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1879–1882, Yokohama, Japan.

Karlgren, H., editor (1990). *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki. ACL.

Kay, M. (1973). The MIND system. In Rustin, R., editor, *Courant Computer Science Symposium 8: Natural Language Processing*, pages 155–188. Algorithmics Press, New York.

Kay, M. (1980). *The Proper Place of Men and Machines in Language Translation.* Xerox Palo Alto Research Center, Palo Alto, California.

Kay, M., Gawron, J. M., and Norvig, P. (1991). Verbmobil: A translation system for face-to-face dialog. Technical report, Stanford University.

King, M. and Perschke, S. (1987). *Machine Translation Today: The State of the Art.* Edinburgh University Press. EUROTRA.

Kittredge, R. I. (1987). The significance of sublanguage for automatic translation. In Nirenburg, S., editor, *Machine Translation: Theoretical and Methodological Issues*, pages 59–67. Cambridge University Press.

Kulikowski, S. (1991). Using short words: a language identification algorithm. Unpublished technical report.

Lamel, L. F. and Gauvain, J.-L. S. (1994). Language identification using phone-based acoustic likelihoods. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 293–296, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Landauer, T. K. and Littman, M. L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual*

Conference of the UW Centre for the New Oxford English Dictionary and Text Research, UW Centre for the New OED and Text Research, Waterloo Ontario.

Landsbergen, J. (1987). Isomorphic grammars and their use in the ROSETTA translation system. In *Machine Translation Today: The State of the Art.* Edinburgh University Press, Edinburgh.

Lehrberger, J. and Bourbeau, L. (1988). *Machine translation: linguistic characteristics of MT systems and general methodology of evaluation.* John Benjamins, Amsterdam, Philadelphia.

Levin, L., Suhm, B., Coccaro, N., Carbonell, J., Horiguchi, K., Isotani, R., Lavie, A., Mayfield, L., Rose, C. P., Van Ess-Dykema, C., and Waibel, A. (1994). Speech–language integration in a multi-lingual speech translation system. In *Proceedings of the 1994 AAAI Conference*, Seattle. American Association for Artificial Intelligence.

Li, K.-P. (1994). Automatic language identification using syllabic features. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 297–300, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Maruyama, H., Watanabe, H., and Ogino, S. (1990). An interactive Japanese parser for machine translation. In Karlgren, H., editor, *Proceedings of the 13th International Conference on Computational Linguistics*, volume 2, pages 257–262, Helsinki. ACL.

Melby, A. K. (1982). Multi-level translation aids in a distributed system. In *Proceedings of the 9th International Conference on Computational Linguistics*, volume 1 of *Ling. series 47*, pages 215–220, Prague. ACL.

Miller, G. (1990). Wordnet: An on-line lexical database. *International journal of Lexicography*, 3(4):235–312.

Morimoto, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M., and Kurematsu, A. (1993). ATR's speech translation system: ASURA. In *Proceedings of the Third Conference on Speech Communication and Technology*, pages 1295–1298, Berlin, Germany.

MTS (1989). *Proceedings of the Second Machine Translation Summit*, Tokyo. Omsha Ltd.

MTS (1991). *Proceedings of the Third Machine Translation Summit*, Carnegie Mellon University.

Muraki, K. (1989). PIVOT: Two-phase machine translation system. In *Proceedings of the Second Machine Translation Summit*, Tokyo. Omsha Ltd.

Mustonen, S. (1965). Multiple discriminant analysis in linguistic problems. In *Statistical Methods in Linguistics*. Skriptor Fack, Stockholm. Number 4.

Muthusamy, Y. K. (1993). *A Segmental Approach to Automatic Language Identification*. PhD thesis, Oregon Graduate Institute of Science & Technology, P.O.Box 91000, Portland, OR 97291-1000 USA.

Muthusamy, Y. K., Barnard, E., and Cole, R. A. (1994). Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41.

Muthusamy, Y. K., Cole, R. A., and Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 2, pages 895–898, Banff, Alberta, Canada. University of Alberta.

Muthusamy, Y. K., Jain, N., and Cole, R. A. (1994). Perceptual benchmarks for automatic language identification. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 333–336, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Nagao, M. (1992). Some rationales and methodologies for example-based approach. In *Fifth Generation Natural Language Processing*. Publisher Unknown.

Nakayama (1994). Modeling content identification from document images. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 22–27, Stuttgart, Germany. ACL, Morgan Kaufmann.

Newman, P. (1987). Foreign language identification: First step in the translation process. In *Proceedings of the 28th Annual Conference of the American Translators Accociation*, pages 509–516.

Ney, H. and Billi, R. (1991). Prototype systems for large-vocabulary speech recognition: Polyglot and Spicos. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, volume 1, pages 193–200, Genova, Italy. European Speech Communication Association.

Nirenburg, S., editor (1987). *Machine Translation: Theoretical and Methodological Issues*. Cambridge University Press.

Nirenburg, S., Raskin, V., et al. (1986). On knowledge-based machine translation. In *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn. ACL.

Okumura, A., Muraki, K., and Akamine, S. (1991). Multi-lingual sentence generation from the PIVOT interlingua. In *Proceedings of the Third Machine Translation Summit*, Carnegie Mellon University.

Paris, C. and Scott, D. (1994). Stylistic variation in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 45–52, Kennebunkport, Maine. Springer-Verlag, Berlin.

Perschke, S. (1989). EUROTRA project. In *Proceedings of the Second Machine Translation Summit*, Tokyo. Omsha Ltd.

Ramesh, P. and Roe, D. B. (1994). Language identification with embedded word models. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1887–1890, Yokohama, Japan.

Rau, M. D. (1974). Language identification by statistical analysis. Master's thesis, Naval Postgraduate School.

Rayner, M. et al. (1993). A speech to speech translation system built from standard components. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Reyes, A. A., Seino, T., and Nakagawa, S. (1994). Three language identification methods based on HMMs. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1895–1898, Yokohama, Japan.

Roe, D. B., Pereira, F. C., Sproat, R. W., and Riley, M. D. (1992). Efficient grammar processing for a spoken language translation system. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 213–216, San Francisco. Institute of Electrical and Electronic Engineers.

Rösner, D. and Stede, M. (1994). Techdoc: Multilingual generation of online and offline instructional text. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 209–210, Stuttgart, Germany. ACL, Morgan Kaufmann.

Sadler, V. (1989). Working with analogical semantics: Disambiguation technics in DLT. In Witkam, T., editor, *Distributed Language Translation (BSO/Research)*. Floris Publications, Dordrecht, Holland.

Salton, G. and McGill, M. (1983). *An Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Sato, S. (1992). CTM: An example-based translation aid system using the character-based best match retrieval method. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France. ACL.

Schmitt, J. C. (1991). Trigram-based method of language identification. U.S. Patent number: 5062143.

Schubert, K. (1988). The architectre of DLT—interlingual or double direct. In *New Directions in Machine Translation*. Floris Publications, Dordrecht, Holland.

Schutz, J., Thurmair, G., et al. (1991). An architecture sketch of Eurotra-II. In *Proceedings of the Third Machine Translation Summit*, Carnegie Mellon University.

Sibun, P. and Spitz, L. A. (1994). Language determination: Natural language processing from scanned document images. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 15–21, Stuttgart, Germany. ACL, Morgan Kaufmann.

Sigurdson, J. and Greatex, R. (1987). *Machine Translation of on-line searches in Japanese Data Bases*. RPI, Lund University.

Somers, H. L., Tsujii, J.-I., and Jones, D. (1990). Machine translation without a source text. In Karlgren, H., editor, *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 271–276, Helsinki. ACL.

Spitz, L. A. (1993). Generalized line word and character finding. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 686–690.

Tomita, M. (1986). Sentence disambiguation by asking. *Computers and Translation*, 1(1):39–51.

Tong, L. C. (1987). The engineering of a translator workstation. *Computers and Translation*, 2(4):263–273.

Uchida, H. (1986). Fujitsu machine translation system: ATLAS. In *Future Generations Computer Systems 2*, pages 95–100. Publisher Unknown.

Uchida, H. (1989). ATLAS-II: A machine translation system using conceptual structure as an interlingua. In *Proceedings of the Second Machine Translation Summit*, Tokyo. Publisher Unknown.

Ueda, Y. and Nakagawa, S. (1990). Prediction for phoneme/syllable/word-category and identification of language using HMM. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, volume 2, pages 1209–1212, Kobe, Japan.

Vasconcellos, M. and Len, M. (1988). SPANAM and ENGSPAM: Machine translation at the Pan American Health Organization. In Slocum, J., editor, *Machine Translation systems*, pages 187–236. Cambridge University Press.

Velho Lopes, R. R. (1989). Automated access to multilingual information: a Brazilian case study. *Information Development*, 5(3).

Wahlster, W. (1993). Verbmobil, translation of face-to-face dialogs. In *Proceedings of the Fourth Machine Translation Summit*, pages 127–135, Kobe, Japan.

Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., and Tebelskis, J. (1991). JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 793–796, Toronto. Institute of Electrical and Electronic Engineers.

Wehrli, E. (1992). The IPS system. In *Proceedings of the 14th International Conference on Computational Linguistics*, volume 3, pages 870–874, Nantes, France. ACL.

Whitelock, P. J., Wood, M. M., Chandler, B. J., Holden, N., and Horsfall, H. J. (1986). Strategies for interactive machine translation: The experience and implications of the UMIST Japanese project. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 25–29, Bonn. ACL.

Winsoft (1987). *Manuel d'utilisation de WinTool.* Winsoft Inc., Grenoble. Version 1.1.

Witkam, T. (1988). DLT—an industrial R&D project for multilingual machine translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest.

Wood, M. M. and Chandler, B. (1988). Machine translation for monolinguals. In *Proceedings of the 12th International Conference on Computational Linguistics*, pages 760–763, Budapest.

Woszczyna, M., Aoki-Waibel, N., Buo, F. D., Coccaro, N., Horiguchi, K., Kemp, T., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C. P., Schultz, T., Suhm, B., Tomita, M., and Waibel, A. (1994). Towards spontaneous speech translation. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 345–349, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Ziegler, D. V. (1991). *The automatic identification of languages using linguistic recognition signals.* PhD thesis, SUNY Buffalo.

Zissman, M. A. and Singer, E. (1994). Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 305–308, Adelaide, Australia. Institute of Electrical and Electronic Engineers.