Chapter prepared in 1994 and ultimately published as a re-worked version in Wright, S.E., and G. Budin [eds.] Handbook for Terminology Management: *vol. 2:* Application-Oriented Terminology Management*, Amsterdam/Philadelphia: John Benjamins, 2001, pp. 697-723)*

TERMINOLOGY AND MACHINE TRANSLATION

Muriel Vasconcellos, Brian Avey, Claudia Gdaniec, Laurie Gerber,
Marjorie León, and Teruko Mitamura

1. Introduction

Machine translation will be defined here as *the technology whereby computers attempt to model the human process of translating between natural languages*. The computer, rather than a person, generates the "output" though it is only a rough draft, not yet fit for most types of consumption. The draft is usually polished into final form by a translator or a bilingual editor, though in some cases it may be used directly by a technical expert who is gathering data for ongoing research.

If the right terminology has been supplied to a machine translation (MT) system, the target-language equivalents are retrieved not only automatically but also in their correct place in the output document. This is one of the advantages of MT: it dispenses with the need to look up terms, whether in hard-copy dictionaries or on-line.

With large projects and multi-translator teams, fully utilized MT has proven to be very effective for keeping terminology uniform (e.g., Brace 1993). Indeed, in a group of recent reports from 36 MT users (Vasconcellos 1993), this was the advantage most often cited. Alternatively, even if a translator prefers to create translation in the traditional way, an MT printout can be a valuable guide, assuming that the system's database is well stocked with the required vocabulary. It's worth the spelling assistance alone. To cite a few well-known bugbears in medical terminology, it makes the right choice between *gluco-* and *glyco-*; it saves looking up *erysipelas* and *paracoccidioidomycosis*; and it never gives **military tuberculosis* for *tuberculosis miliar* – a common translator error. The question is not so much *whether* MT is useful for banking and retrieving terminology, but rather what is involved in building up the terminological reserve so that it can produce needed terms on demand.

The "opposite" of MT, so to speak, is M*A*T, or machine-*aided* (also called machine-*assisted*) translation, in which people do the translation and computers stand at the ready to look up isolated terms on command. The MT-MAT distinction is not water-tight, however, and recent developments are blurring it even more. Innovative approaches are making MAT more MT-like and vice versa. In the case of MAT, the computer is taking on an increasingly active role: systems have been developed that monitor human translation input and identify identical or similar strings of text in a database of previously stored documents, which are then flashed on the screen as candidates to be pasted into the ongoing translation. Meanwhile, in the MT world, interactive systems are finding ways in which to capture needed human reactions in midstream in order to produce output of better quality.

For present purposes, however, the distinction between MAT and MT is quite useful. It defines the difference between terminology management software, which is covered elsewhere in this volume, and full-text MT, dealt with here, in which surrounding context interacts with terminology and, as a result, terms cease to be isolated entities.

2. Getting the Right Translation

When terms are part of a larger text, even in the most limited of subject areas, their translation is often affected by how they are being used. Take, for example, the wholly ambiguous sentence

> (1) Because the clock did not work properly, the conductor failed to make the connection and the bus had to operate with just a driver.

It could have two entirely different readings, one in the general sense referring to public transportation, and the other a technical description of the failure of a piece of electrical equipment, depending on the meanings assigned to *clock*, *conductor*, *connection*, *bus*, and *driver* (Wheeler 1983).

This problem is relatively easy to solve, however, compared with the difficulty of getting the right translation when a given word has more than one meaning within the same domain. For example, in a Spanish text on atomic energy the word *núcleo* could be translated as *nucleus* or *core*; in a legal text, *derecho* could be *law* or *right*—and the words that would make the meaning clear are not always nearby.

MT systems use various techniques to generate the right translation for the context—in other words, a *context-sensitive* translation. Some of them have a broad range of strategies and can process many kinds of linguistic information. Linguistic detail helps a system to make subtle distinctions based on context and to produce high-resolution translations, especially in limited subject areas. A number of systems can cope with complex syntactic formulations, and some respond to pragmatic criteria and make use of world knowledge drawn from intricate databases.

While such capabilities make for powerful MT systems, the truth is that building in the required knowledge is not an easy task. The terminologist or other person working on the system's "dictionary" has to have a fair amount of specific training as well as some appreciation for the principles of general linguistics. Most of the commercial systems have developed user-friendly software for updating their dictionaries (Logos Corporation of Mt. Arlington, New Jersey, was a pioneer in this area), but even so, the price of such refinement can be quite high, and some users do not have the time or the human resources that are needed for elaborate customization. Then there are systems that are not so linguistically complex, and they offer a tradeoff that may be worthwhile for certain applications: easy dictionary-building in exchange for less linguistic specificity. The sections that follow will describe the range of MT systems that are available and give examples of the kinds of problems they are equipped to handle.

3. Types of MT Systems

MT systems are usually classified as *direct*, *transfer*, or *interlingual*. We now also have *corpus-based* MT. This classification is useful when it comes to understanding what is involved in entering terminology in the dictionaries.

3.1 *Direct Systems*

One of the main characteristics of a "direct" MT system is that it does not analyze full sentences.

Rather, it builds the sentence phrase by phrase using local routines. Because it doesn't parse, it can make only limited use of syntactic, semantic, and pragmatic information in the translation process. Even if the dictionary allows for such codes and the system has rules that act on them, the lack of a full analysis of the sentence will limit the ways in which the information can be used.

To take a very simple example, if we wanted to say *female rats* in Spanish, in order to get the desired translation *ratas hembras* instead of the default translation **ratas femeninas*, in the simplest of the direct systems we would have to enter a hard-wired phrase on each side of the equation, and we would have to repeat the exercise for each animal name that we expected to encounter. If, on the other hand, the direct system is in transition and gearing up to parse, as in the case of the MicroTac products of San Diego, California, we might be able to enter a code 'minus human' [-HUM] on animal names. This would make it possible to trigger *hembras* for all the animal names modified by *female*, though without the capacity to parse, the usefulness of this code would be quite localized. In other words, the system would probably *not* be able to handle a sentence like

(2) The technician separated the rats, placing the females in a cage by themselves.

In such a sentence, a direct system would also miss the fact that the translation of *themselves* requires a feminine ending in Spanish.

For many translation tasks, it may never be necessary to generate *hembras*. It simply doesn't come up. The possibility of the term occurring is predictable if we are dealing with a particular subject matter, or *domain*, such as computer manuals or weather reports. With general translation, on the other hand, there is no way of anticipating when the need will arise.

Of course, it may not really *matter* whether the correct translation or the feminine ending are generated. For a scientist scanning a database to find material that might be of interest, the sense is clear. The translation has served its purpose. On the other hand, it may be feared that the wrong translation will be offensive to the reader or undermine the client's image. In such cases there is usually a posteditor who mediates by correcting the machine's output with a few simple keystrokes. Here the operative words are *few* and *simple*. Too many required changes, or operations that are difficult to execute,[1] may ultimately outweigh the advantages of MT, including its potential to produce uniform terminology. The user may decide that MT is not worth the trouble.

Most of the commercial systems running on personal computers at the beginning of the 1990s were direct systems, but that picture has started to change. Systems that have traditionally run on mainframes and high-end workstations are in the process of downsizing to 386 and 486 machines, and the earlier PC systems are maturing. It is therefore unsafe to generalize about the linguistic approach that a system uses based on its platform or its price alone.

While it may be difficult to judge the linguistic prowess of an MT system by its platform and its price, one thing is certain: with increased public awareness of MT, people are buying large volumes of software packages off the shelf and through the mail. MicroTac Software, the market leader in number of units sold, reported in May 1993 that all-time total sales of its four bidirectional packages had reached the whopping figure of 150,000.[2]

---

[1]Some maneuvers may be difficult because of the operator's level of word-processing skills; other difficulties may be inherent in the word-processing software, in which case one solution might be to develop customized macros that perform repetitive operations more efficiently.

[2]Source: Michael Tacelosky, President, MicroTac Software (figure does not include upgrades).

3.2 *Transfer Systems*

While direct systems are far ahead of any other kind of MT in terms of number of packages sold and installed, it is safe to say that the bulk of practical machine translation is performed by *transfer systems*. These systems have been in everyday use since the 1970s and continue to be the mainstay of large, heavy-duty translation operations.

Transfer systems have three basic components: *analysis*, *transfer*, and *synthesis*. After analyzing the source text, they convert the result of the analysis, or *parse*, into an intermediate representation, which the *transfer component* then interprets in order to produce a more abstract representation. They then take the output of the transfer component and feed it into the *synthesis component*, which stitches together the final translation.

This modularized approach emerged in response to several different pressures, mainly generated by the demand for translation of English into foreign languages. First, the need to parse the source text was becoming increasingly evident. Direct systems were challenged by the frequent part-of-speech homographs and syntactic ambiguities of English. It often happens that several of the words in an English sentence can function as more than one part of speech, and decisions have to be made in order to work through the labyrinth of possibilities. For example, in the sentence

(3) Slow sand filters produce more even results

every word has at least two potential functions. The possible combinations are astronomical. In addition, words of the same part of speech can have multiple translations, as terminologists well know. A parser sets a series of conditions that must be fulfilled in order for a given structure or word choice to be valid. In this way, illegal readings of the sentence are ruled out.

The second pressure for developing this kind of system was the demand for MT from a single source language into multiple targets. Manufacturers in the electronics industry, in particular, began to look to MT as a means of speeding up the process of localization—i.e., translating customer support manuals and related material into many languages simultaneously for introduction into foreign markets. The intermediate representation going into the transfer component provides an efficient starting point for translation into more than one target language: once the analysis has been performed, the results are captured in a concise and systematic representation to which MT developers can readily attach target modules in a number of different languages. The lion's share of the MT task has been accomplished. While it is true that direct systems allow developers to add multiple targets to a single source language, and to update them simultaneously, the main point about the transfer system is that it analyzes the complete sentence and, in order to do so, organizes massive amounts of in-depth linguistic information in a more efficient way.

To summarize, then, transfer systems make it possible not only to translate between texts that have very different structures but also to use a single analysis component to capture sophisticated linguistic data for multiple target languages. In addition, they have the advantage that they are open-ended in terms of room for linguistic growth. They can accommodate a wealth of semantic criteria.

They are also well-suited to storing information from a pre-existing terminological entry—although what they actually retrieve, and how they are made to do this, is another story. Aspects of this challenge will be explored in Section 5.

3.3 *Interlingual Systems*

Despite all the capabilities of transfer systems, there are some kinds of ambiguity that they don't handle. They also fail to capture many of the discourse-related features of language that are important for maximum communication. From the developer's standpoint, another disadvantage of the transfer system is that it is top-heavy in the analysis module and, in ways, inefficient. These problems have been addressed with *interlingual systems*, in which the transfer component is replaced by an *interlingua*.

An interlingua is a comprehensive intermediate representation containing universal linguistic, ontological, logical, and pragmatic knowledge which (ideally) applies to all languages—or at least to all the languages being translated. The ontological knowledge base, or ontology, can be used for generalization and inferencing, two functions that cannot be performed by a transfer system. Since the target translation is rebuilt "from scratch" on the basis of deep-level information returned by the interlingua, the final phase is called *generation*. The idea is for the interlingual component to serve as a pivot from which translations can be produced in several or many directions with much less investment in language-specific modules at both the input and the output ends. The fact of having an ontological knowledge base also means that it should be possible to build up the lexicon dynamically—and to some extent automatically.

While interlingual systems are costly to build, the economies to be gained from reduced effort in the development of specific languages make this approach attractive when there is a demand for translation in a number of different directions—especially both from and into one or several pairs. They are particularly interesting in limited subject areas, where their carefully constructed domain models can be relied on to produce richly detailed knowledge about the world to which the translation refers.

Some very interesting work has been done on the use of domain models for the development of terminology (see Meyer, Eck, and Skuce in this volume). Knowledge-based terminology systems—or more precisely, terminological knowledge bases (TKBs)—dovetail nicely with interlingual MT systems. In principle, at least, a TKB could serve as one of the knowledge sources accessed in the course of an automatic translation process. Section 6 below looks at an effort currently under way to marry one such TKB to MT, as well as at terminology development for a particular interlingual system.

3.4 *Corpus-based Systems*

A summary of MT types would not be complete without some reference to recent efforts, still in the laboratory, to develop *corpus-based* systems. The idea behind these systems is to dispense with labor-intensive linguistic rules and deeply coded knowledge sources and instead produce MT by automatically extracting data from corpora of existing translations. In their purest form, there are two basic types of corpus-based systems: (1) *statistical*, in which the translation is derived directly from corpora of matched pre-existing translations and decisions are based on statistical probabilities, and (2) *example-based*, in which the corpora serve as resources for the automatic compilation of dictionaries and other knowledge sources, which are then stored and later invoked by a translation program.

A statistical system—for example, CANDIDE at IBM (Brown et al. 1990)—builds up its knowledge sources by "training" on immense corpora of existing matched human translations in a particular subject area. The dictionaries are developed automatically, solely on the basis of statistics, and the "terminology" is merely the reflection of the parallel source and target texts that happened to be available in the domain in question. These systems have at least two advantages: the output tends to sound like natural text, and they do away with manual linguistic coding, thus saving thousands of person-hours. Nevertheless, in their strict form they are inherently incompatible with terminology work and the incorporation of existing glossaries, knowledge bases, or other external collections of terms.

There is talk of finding ways in which the advantages of the statistical design can be blended with more classic, controllable approaches.

The example-based strategy—which relies on a bilingual database of example phrases extracted from a corpus of existing human translations—has been around somewhat longer (e.g., the DLT project in Utrecht and work at Kyoto University, Nagao 1984). The news is that it is gaining favor. The concept is predicated on the belief that translation consists largely of recalling or finding analogous examples (Hutchins 1993:17). Like statistical systems, example-based systems tend to produce natural-sounding translation. Their strength is their greater flexibility: they may be combined with knowledge sources, grammar rules, concept definitions, mapping rules, etc. There is no reason why terminology databases could not be incorporated into such systems as well.

4. Terminology Work with Direct Systems

It will be recalled that the example of *hembra* was used to illustrate some of the characteristics of a direct system. To continue with the discussion, it might be possible and worthwhile to tailor such a system to elicit *hembra* in a limited domain, especially if the number of animals to which it applies was relatively small. There are a couple of ways of handling this. If the more usual translation *femenina* will never be needed, the most rudimentary solution is to assume that *hembra* is the dictionary's default translation for *female*. If it is the *only* option in the dictionary, this is known in the trade as "protective undercoding"—the dictionary simply does not offer the usual choices. But that solution would be quite risky. Many systems offer a safer alternative: the user can enter a customer-specific translation for *hembra* in an easily modifiable dictionary that overrides the general-purpose lexicon supplied by the vendor. In this way the user has flexibility and the basic translation is preserved. Yet another approach, if there are only a few cases in which *hembra* is needed, is to provide hard-wired translations for such fixed strings as *female rats*, *female guinea pigs*, and *female hamsters*. These are all common strategies that are easy for the user to implement. However, they quickly break down if there is any variation at all in the word string, as for example in the phrase *female rats, guinea pigs, and hamsters* or the sentence *The rats, guinea pigs, and hamsters were all females.*

An example of building in hard-wired phrases can be seen in selections from a list of English-Spanish dictionary entries prepared for PC-TRANSLATOR by the vendor, Linguistic Products of The Woodlands, Texas. (In these examples, the symbol "X" at the end of each line means that the phrase is entered in the dictionary as an uninflected unit.) The following list of selected examples shows the kind of terminology that would be useful to have stored in an MT system (though not necessarily the recommended translation equivalents):

```
FLOW CELL ALIGNMENT ,ALINEACIÓN DE LA CÉLULA DE FLUJO,X
FLOW CYTOMETRY ,CITOMETRÍA DE FLUJOS,X
LASER LIGHT SCATTER ,DISPERSIÓN DE LUZ LASER,X
LEUKOCYTE DIFFERENTIAL COUNTER ,CONTADOR DIFERENCIAL DE LEUCOCITOS,X
LIGHT SCATTER ANALYSIS ,ANÁLISIS POR LA DISPERSIÓN DE LA LUZ,X
LYTIC REAGENT ,AGENTE REACTIVO, LITICO,X [NB: should be lítico]
MEAN CORPUSCULAR HEMOGLOBIN CONCENTRATION ,CONCENTRACION MEDIANO DE
HEMOLOBINA CORPUSCULAR,X
```

Other examples, however, give an idea how labor-intensive the task really is:

```
BLOOD CELL COUNT ,CONTEO DE CÉLULAS DE SANGRE,X
BLOOD CELLS ,CÉLULAS DE SANGRE,X
CELL ANALYSIS ,ANÁLISIS CELULAR,X
CELL COMPONENTS ,LOS COMPONENTES DE LAS CÉLULAS,X
CELL SIZE DISTRIBUTION ,LA DISTRIBUCIÓN DEL TAMAÑO DE LAS CÉLULAS,X
CELL WALLS ,LOS PAREDES DE LAS CÉLULAS,X [NB: should be las paredes]
DIFFERENTIAL WHITE CELL COUNT ,CONTEO DIFERENCIAL DE CÉLULAS BLANCAS,X
RED BLOOD CELL ,CÉLULA DE SANGRE ROJA,X
RED BLOOD CELL COUNT ,CONTEO DE CÉLULAS DE SANGRE ROJAS,X
RED BLOOD CELLS ,CÉLULAS DE SANGRE ROJAS,X
RED CELL DISTRIBUTION WIDTH ,ANCHO DE DISTRIBUCIÓN DE CÉLULAS ROJAS,X
SIZE OF THE CELL ,TAMAÑO DE LA CÉLULA,X
```

Apparently separate entries were required in some cases in order to get *celular* as opposed to *de las células*. However, if the system had been able to parse it would have had the capability of generating these distinctions automatically.

Other entries hinted at the system's strategy for triggering context-sensitive translations (the questions in square brackets are suggested reasons why the string was entered as a phrase):

```
A SMALL ORIFICE ,UN ORIFICIO PEQUEÑO,X [To get the right word order?]
DISCUSSED THE POSSIBILITY ,DISCUTIERON LA POSIBILIDAD,X
      [To get the plural inflection of the verb?]
DAILY INSTRUMENT CHECKS ,CHEQUEOS DIARIOS DEL INSTRUMENTO,X
      [To block the default translation cheque?]
IT HAS A ,TIENE UNA,X [To block the auxiliary verb ha?]
ENGLAND ET AL ,ENGLAND ET AL,X
      [To block the default translation Inglaterra?]
```

It can be seen that the listing from a batch dictionary update tells quite a bit about a system's linguistic capability. However, it's not too easy to have access to this kind of information. Another way to find out how linguistically sensitive a system will be in translating user-supplied terminology is to have a look at the set of dictionary codes that the user is allowed to define—bearing in mind, of course, that even if a system is capable of analysis at the sentence level, it cannot act on codes that are not attached to specific entries. A system that expects only a small amount of linguistic coding is likely to perform in a direct *manner*, even though it might have the potential of performing more heroic feats.

In the fall of 1992 the American Translators Association conducted a survey of all the PC-based systems then on the market in the United States (Miller, for ATA 1992). As part of this survey, the consultant, L. Chris Miller, tabulated for each system the types of grammatical information that users are allowed to supply for their dictionary entries. At the time of the survey, the spread of permissible dictionary features for six of the systems looked like this:

Dictionary Codes Available to Users of PC Systems

User dictionary codes	Finalsoft	Globalink	Linguistic Products	MicroTac	Socatra	Toltran
Part of Speech	6	8	6	12	5	5
Gender	Yes	Yes	Yes	Yes	Yes	Yes
Number	Yes	Yes	Yes	Yes	Yes	Yes
Inflection	No	Yes	No	Yes	Yes	No
Attributes	No	No	No	Yes	Yes	No
Multiple parts of speech for a term	Yes	Yes	Yes	Yes	Yes	No
Multiple translations within same POS	Yes	8	No	15	Unlimited	No

Most of the systems above are capable of making certain localized decisions based on the part of speech that has been specified. They all allow for the identification of nouns, verbs, and adjectives. However, the part-of-speech heading is not necessarily limited to traditional grammatical categories. As with any MT system, this is the first cut in syntactic analysis—the branching point after which separate sets of rules are going to apply throughout the translation process, and a useful point at which to make some other basic assumptions as well.

For example, in the following list of categories used by MicroTac, the first nine of them are traditional and the last three are system-specific: Noun, Pronoun, Personal Pronoun, Verb, Adjective, Adverb, Preposition, Article, and Conjunction, on the one hand, and, on the other, Phrase, Miscellaneous, and Unknown. The Phrase category sets aside a "part of speech" for the purpose of triggering multi-word phrases. Many other systems do this as well. Indeed, using one tactic or another, they all offer the user the possibility of forming hard-wired phrases. This capability is sometimes enhanced by the possibility of using a *wild card*—a free variable that functions in a multi-word phrase much the same way as a joker does in a game of canasta.

In addition, five of the six systems allow the user to specify more than one part of speech for the same word, and some of them also provide for multiple translations within the same part of speech.

Attributes are the other capability of interest for terminology. These are the characteristics of the concept that tie it to the context and help to specify choices within the same part of speech when more than one translation is possible. To some extent they are system-specific. MicroTac, a system in transition, offers 17 attributes that can be applied to nouns (Animate, Human, Family, Nationality, Occupation, Place, City, Country, Proper, etc.), 12 attributes for verbs, six for adjectives, five for adverbs, and two for conjunctions.

Another way of obtaining specialized translations is through topic-specific dictionaries. When such a dictionary is called in at translation time, it is consulted first, and the translation that is retrieved takes precedence over anything in the general dictionary. This approach would be useful for translating the sentence cited at the beginning:

(1) Because the clock did not work properly, the conductor failed to make the connection and the bus had to operate with just a driver.

It would not work, of course, if the usual meanings of *bus*, *driver*, etc. came up elsewhere in the same translation project.

The premise of this discussion has been that a system which uses more linguistic information has greater flexibility in providing context-sensitive translations. However, as pointed out before, there is a tradeoff. Complex entries tend to be more difficult to add to the existing database—although many systems have user-friendly software that greatly facilitates the task. Still, the direct systems that do not require a great deal of grammatical information are easier to learn and to manage. They may be more labor-intensive and more limited in their capacity to handle technical terminology, but their updatability is an important factor in their appeal to the public.

5. Terminology Work with Transfer Systems

The first line of defense in a system that parses is *syntax*. Transfer systems contain massive sets of lexical rules that combine syntactic features with other information about a word, often semantic features, to trigger the desired translation. Three examples of such systems, all of which have extensive resources for eliciting context-sensitive translations, are SYSTRAN, LOGOS, and ENGSPAN.

5.1 *SYSTRAN*[3]

Of all the transfer systems, SYSTRAN probably reflects the longest and most extensive work on lexical development. The company, based in La Jolla, California, began the arduous process of building its mammoth dictionaries in 1964. Today it has dictionaries in 27 language combinations.

For the handling of words with multiple meanings, or words that can have both general and technical meanings, SYSTRAN relies on a combination of syntactic and semantic analysis plus such strategies as topical glossaries and dictionaries of customer-specific translations. A jolly review of its capabilities was presented by Wheeler in his article "The Errant Avocado" (1983), in which he steps the reader through a series of exercises, including the correct coding of the French phrase *poste d'avocat general* so that it translates as *post of advocate-general* rather than what SYSTRAN first came up with—namely, *general avocado station*!

The semantic classification system includes over 500 categories organized in a hierarchical structure. Each word in any of the dictionaries can be encoded with the categories that apply. The system uses this information on both the program and dictionary levels to select the right translation in the target language. An example is the following semantic rule for the English word *frozen*:

[3]Brian Avey and Laurie Gerber of Systran Translation Systems, Inc., contributed to this section.

If word is "frozen" functioning as a modifier, check noun modified
for the semantic property [FOOD PRODUCT].

This rule would distinguish the sense of 'a frozen food' from that of 'an immobilized mechanism'.

For handling more complex and more generalized problems of ambiguity, SYSTRAN has specialized programs at the transfer stage of the translation process. One example is a routine which determines whether a word is animate or inanimate. This program is called in whenever the system encounters a word coded in the dictionary as [+ANIMATE/INANIMATE AMBIGUOUS] for example, *seal*, *parent*, *acquaintance*.

In the topical glossaries, each word or expression in the source dictionaries is assigned target translations in as many fields as necessary to cover the range of meanings and uses for that word. For example, *stem* in English has a variety of senses when it is used in different domains. The target translations for this word would likely include a default translation for general use plus ones for the sense of 'plant stem' in botany and 'brain stem' in anatomy. Before the translation is run, a user may indicate preference for up to 10 fields. For the words that have meanings encoded for the fields indicated, only those specialized translations will be chosen.

The customer can also choose specific translations that override all the others. Companies with special technical or proprietary terms may have expressions encoded so that their translation for the expression will be selected only when a translation is run with their particular code.

Another resource is the set of conditional and nonconditional string replacement features whereby fixed expressions such as *brake shoes* can be handled as a single item with a unique meaning in order to elicit a technical translation in the target language rather than a mere word-for-word replacement.

In addition to strategies for assigning appropriate meanings to frequently occurring noun phrases, SYSTRAN emphasizes the preventive medicine of removing ambiguity, when possible, at a very early stage. The development team forestalls problems by examining large corpora and bringing in new terminology with new rule sets for resolving ambiguity in an ongoing effort to expand dictionary coverage and the system's overall scope. Such an exercise yielded the following readings of the Japanese word *ippou*:

(4) rajio ya terebi no *ippou* tsuukou no jiyouhou wo . . .
'the *one-way* information of radio and television . . .'

(5) *ippou* ni ion bi-mu wo shousha shite iru . . .
'shining an ion beam on *one side* . . .'

(6) sangyou wo sasaeru *ippou*, . . .
'*While* industry is supported, . . .'

(7) *ippou* LPG ni tsuite ha, . . .
'*On the one hand*, concerning LPG, . . .'

(8) Kyousou ha hageshiku naru *ippou* da.
'Competition will become *more and more* intense'

For capturing the differences in meaning in these examples, SYSTRAN is able to encode not only the knowledge on which they depend but also the rules for selecting one lexical item over another. The expandability of the routines that handle homographs allows for new lexicon growth without an unwieldy proliferation of ambiguities.

SYSTRAN has not only addressed the question of ambiguity in depth, it has also dealt with the issues of batch updating and mass encoding of terminology files and integrating terminology so that it produces the desired translations.

5.2 *LOGOS*[4]

Logos Corporation has also gone to great lengths to facilitate the development of large dictionaries of context-sensitive technical terminology in an effort that dates back nearly as long as SYSTRAN's.

The LOGOS system offers a variety of ways to differentiate terminology depending on the *company* the translator works for, the *country* of destination, the *subject matter* of the translation, and *general vs. technical usage.*

Clients such as AT&T, Microsoft, and IBM have their own company-coded dictionaries that their terminologists are free to customize. If they need to differentiate, for example, between terms used by IBM/England and IBM/USA, they can create country-specific dictionaries to accommodate spelling variations and capture such cultural idiosyncrasies as *boot* vs. *trunk* for the German *Kofferraum*. At translation time, LOGOS allows five company codes to be specified, and the system looks the words up in the order in which the user lists the codes.

To deal with semantic ambiguity, LOGOS has up to 500 subdictionaries for different subject areas. The word *Ausleger*, for example, could have the translation *extension arm* in the general dictionary, *boom* in the automotive dictionary, and *jib* in a specific dictionary for cranes. The user can specify up to five domain-specific subdictionaries when submitting a translation. These subdictionaries are logically linked through generic codes. This means that pointers in the translation program map to several related fields—for example, electrical engineering, electricity, and electromagnetism. If a word is not found in the subdictionary that was specified, the system will look for it in one of the others that are related. The user does not have to specify this connection.

Another way to differentiate the meaning of semantically ambiguous words is on the basis of their syntactic and semantic context. For this purpose, LOGOS has a special database which is consulted at various stages of parsing the input sentence. For example, the German adjective *fest* may have the English translation *fixed* in the general dictionary, but it can also elicit *hard* in the context of *Platte* or *Währung*, and it can be translated as *heavy* in the phrase *heavy blow* and *lasting* in the expression *lasting peace*. The tool that does this job is SEMANTHA, an easy-to-use interactive program for creating and updating customized context-sensitive rules. In the example of *female* applying to rats and other animals, SEMANTHA would enable the user to write a single rule specifying that the English adjective *female* should be rendered as *hembra* in Spanish whenever it modifies a noun that is coded as a nonhuman animal. This guarantees that LOGOS will consistently translate female animals as *hembra*. (It should be noted that this rule is too powerful, since it would

[4]Claudia Gdaniec of Logos Corporation contributed to this section.

translate *female cat* as **gato hembra* instead of *gata*; exceptions would have to be handled individually.)

The context rules in the semantic database apply not only to contiguous strings at the surface level but to all types of transformations based on a single deeper structure. Thus, for example, a context rule written for the abstract-level construction *Tafel (case=accusative) löschen <I> clean N* will apply to very different surface-structure transformations at different stages of the syntactic parse:

(9) Die Lehrerin löschte die Tafel
'The teacher cleaned the blackboard'

(10) Die Tafel wurde von der Lehrerin gelöscht
'The blackboard was cleaned by the teacher'

(11) Die regelmässig von der Lehrerin gelöschte Tafel . . .
'the blackboard that was cleaned regularly by the teacher . . .'

(12) Mit Löschen der Tafel hat die Lehrerin . . .
'In cleaning the blackboard, the teacher . . .'

The user first has to decide whether to update an expression (1) as a context rule, through SEMANTHA, or (2) as an entry in the dictionary. The next decision is which subdictionary to update, i.e., under which company code and/or subject matter code. The tool ALEX is used for updating the dictionaries. This interactive program asks the user to give minimum semantic, syntactic, and morphological information about the source and target entries. LOGOS has 130 hierarchically ordered semantic categories for the coding of nouns alone, but the user does not have to learn all these categories. If the basic entry (or head noun in the case of a compound), is already in the dictionary, ALEX offers a choice of categories based on the coding already found in the dictionary. If the entry or the head are entirely new to the system, the user can enter a word that has similar syntactic and semantic properties and ALEX will propose possible semantic categories based on that "synonym."

5.3 *ENGSPAN*

ENGSPAN is a transfer system developed by the Pan American Health Organization (PAHO) in Washington, D.C., for the translation of English into Spanish. While it has a robust capacity to cope with general text, it also has highly developed lexical resources in fields related to medicine, public health, and agriculture.

On more than one occasion the developers of ENGSPAN chose to enrich their system through the infusion of large paired lists—including a list of about 30,000 medical terms. When these entries were being augmented with the information required for ENGSPAN's parser, a program was written to generate semantic codes for English words that ended in typical medical suffixes such as *-itis* 'inflammation', *-osis* 'condition of', *-otomy* 'incision into', etc. In this way, a portion of the semantic coding was added to the system automatically. Similar logic is also used in the regular translation program to make educated guesses about words not found in the course of translation.

A more labor-intensive exercise was involved in the introduction of 8,697 English and

matching Spanish descriptors from AGROVOC, a multilingual thesaurus developed by the Food and Agriculture Organization of the United Nations (FAO) and the Commission of the European Communities (CEC). In 1987, the ENGSPAN team undertook a terminology project in the field of agriculture to prepare the system for pilot outplacement at two agricultural research centers. The experience was also intended to yield a methodology that could be applied to the incorporation of other specialized terminology in the future. This exercise was reported in detail in *TermNet News* (León, Santangelo, and Vasconcellos 1987).

It was expected that a sizable proportion of the terms would already be present in ENGSPAN's dictionaries. However, given the difference in purpose between a thesaurus of indexing terms and a dictionary for the translation of running text, it was also recognized that some of the AGROVOC Spanish equivalents might not be appropriate for ENGSPAN. Indeed, it was soon found that a number of considerations prevented any large-scale automatic entry of these terms.

This following set of pairs exemplifies the many problems that called for the discerning judgment of a human linguist:

FLAVOUR	AROMA
FLAVOUR ENHANCERS	REFORZADORES DE AROMA
FLAVOURED MILKS	LECHE AROMATIZADA
FLAVOURING	AROMATIZACION
FLAVOURING CROPS	PLANTAS DE CONDIMENTO
FLAVOURINGS	AROMATIZANTES

- AGROVOC uses British spelling, whereas ENGSPAN uses American spelling (e.g., *flavour* vs. *flavor*. While ENGSPAN's lookup procedure does have logic for recognizing British spelling at translation time, entries in the dictionary must be in the American style.
- Many AGROVOC descriptors are given in the plural form; ENGSPAN's entries, on the other hand, need to be in the singular in order to generate the system's full range of possibilities.
- The AGROVOC descriptors are in full caps, while the distinction between upper and lowercase needs to be made in ENGSPAN's entries so that the system will know how to write abbreviations, proper names, scientific words, the names of proprietary products, etc.
- The Spanish words in AGROVOC do not have orthographic accents (acute accent, diaeresis, tilde), whereas these diacritics are required by ENGSPAN.
- Some of the words that are used as count nouns in AGROVOC (*milks* and *flavourings*) would normally be treated as a bulk noun (*milk*) or a verb nominalization (*flavouring*) in ENGSPAN.
- Many terms used only as nouns in AGROVOC (*flavour* and *milk*) would have to be coded for other parts of speech (in this case, verb and noun) in ENGSPAN.
- Many of the AGROVOC terms include inflected verb forms (*flavoured*) that would be unnecessary additions to the ENGSPAN dictionary.
- Some of AGROVOC's Spanish equivalents (for example, *aroma*) conflicted with existing ENGSPAN translations (in this case, *sabor*).
- Within AGROVOC, the same word is sometimes translated differently in different contexts. In these six entries alone, *flavouring* has been rendered as *aromatizacion*, *de condimento*, and *aromatizante*.

These problems were solved on a case-by-case basis, beginning with the single words. In the

event of conflicts between AGROVOC and an existing ENGSPAN entry, the linguist had three basic options: the PAHO term could be retained, the existing gloss could be replaced, or a second term could be added. Most often both were needed and special coding was introduced to trigger different choices on the basis of context. For example, the distinction between *drinker* [+HUM] and *drinker* [+DEVICE] is an analysis problem that has to be handled in the source entry, whereas the choice between *cuello* [+HUM][+BODY PART] and pescuezo [-HUM][+BODY PART] gets handled on the target (synthesis) side. When the context had to be specified, the single words were treated as expressions.

The next step was to review the multi-word expressions and decide whether they should be entered as substitution units (SU), analysis units (AU), or transfer units (TU). The criteria for choosing between these three types of units are as follows:[5]

SU There is a high degree of certainty that the particular string will always have the same translation; it is not likely that other words will interrupt the phrase or that one of its elements will form part of another syntactic structure.

AU The words are likely to occur together but may belong to another syntactic structure.

TU The words may not be contiguous; the input string must be analyzed syntactically in order to decide whether or not the phrase exists; a rule can be formulated that will apply to a class of lexical items.

The following list shows cases in which these various approaches would be used:

English term	Type of unit	Spanish gloss
FISH	(Main entry)	PESCAR (verb), PEZ (noun)
FISH	TU	PESCADO
FISH CULTURE	SU	PISCICULTURA
FISH DETECTION	None	DETECCIÓN DE PECES
FISH EXTRACTS	TU	EXTRACTOS DE PESCADO
FISH FEEDING	None	ALIMENTACIÓN DE PECES
FISH LARVAE	SU	ALEVINES
FISH OILS	Same TU	ACEITES DE PESCADO
FISH POISONING	SU	PESCA CON VENENO
FISH PONDS	AU	ESTANQUES PISCÍCOLAS
FISH WASTES	AU	DESECHOS DE(L) PESCADO

The AGROVOC exercise showed that the incorporation of specialized vocabularies into the ENGSPAN dictionaries could only be automated to a certain extent and that many decisions had to be made. While the MT program itself, its supporting software, and word-processing capabilities can be used to speed up some parts of the work, human decisions have to be made at almost every turn. Each set of terms will present problems of its own. Of course, the result of all this painstaking work is a better MT system with increased subject-area coverage and overall improved quality.

[5]A full explanation of ENGSPAN's dictionary options is given in León and Schwartz (1986).

6. Terminology Work with Interlingual Systems[6]

As noted above, interlingual MT relies at least in part on concept systems with elaborate networks expressing relationships between the nodes—relationships such as 'is a', 'is a property of', 'composed of', 'precedes', 'follows', 'for the purpose of', 'subtype of', 'supertype of', 'sufficient condition for', 'necessary input for', 'opposite of', 'also known as', etc. Although on-line terminological knowledge bases (TKBs) have been developed which show conceptual links of this kind—for example, COGNITERM (see Meyer, Eck, and Skuce in this volume)—until quite recently the work has been carried on independently of MT. Many of the elements in a traditional terminology record are "excess baggage" for machine translation, while on the other hand MT requires specific syntactic and semantic features and attributes that are usually not included in a terminological entry. In any case, it seems a rather daunting task to map the information in a TKB so that it can be captured usefully in an automatic translation process that must also consider syntax, case frames, schemata, pragmatics, and the overall fabric of discourse.

The ET10/66 Consortium in Europe is now embarked on a project which, among other aims, seeks to overcome the presumed incompatibility between terminology banks, knowledge bases, and machine translation systems. The ET10/66 team, which includes investigators in Dublin, Luxembourg, Lisbon, Saarbrücken, and Bonn, is working with EIRETERM in Dublin and building a reusable terminological resource that will provide extralinguistic domain knowledge within the advanced linguistic engineering platform (ALEP) designed under the auspices of the European Commission. Working in the domain of telecommunications, they have developed an ontology by formalizing the definition of each term and establishing the conceptual links between them (Pearson [ed.] 1992). Since the ALEP framework allows for encoding not only domain knowledge but also linguistic knowledge, they are also able to incorporate syntactic and semantic information based on the EUROTRA design. All this information, taken together, may be consulted in a parsing process for machine translation (Schütz and Ripplinger 1993). This model has been used to show how ontological knowledge helps to solve ambiguities in the attachment of prepositional phrases, the identification of collocations, and the selection of appropriate target translations from among several possibilities (Ripplinger [ed.] 1993). The importance of this work is that a common formal device is used to express and represent different types of knowledge, thus avoiding the need to build an interface between them (Schütz and Ripplinger 1993).

A more practical example of terminology development for interlingual MT is the work being done with KANT, a system developed and now being scaled up by Carnegie Mellon University in Pittsburgh, Pennsylvania, to translate heavy equipment manuals into a number of target languages for Caterpillar, Inc. The developers are counting on KANT's rich knowledge-base capabilities to generate machine translation that will require no postediting. Experiments have

[6]This section draws on Pearson (1993), Ripplinger (1993), Schütz and Ripplinger (1993), and Mitamura, Nyberg, and Carbonell (1993). The author is grateful to Ingrid Meyer and Teruko Mitamura for supplying advance copies of their work and to Bärbel Ripplinger for sending a full set of reports of the ET10/66 Project.

shown that systems like KANT are capable of producing good results in technical domains as long as the domain semantics are "tractable" (Mitamura, Nyberg, and Carbonell 1993).

In the past, efforts to scale up knowledge-based MT have entailed a large amount of hand coding to build the necessary lexical resources, ontology, grammatical and other rule sets, etc. This work represents a sizable investment. Accordingly, the developers of KANT's dictionaries for Caterpillar decided to reduce the labor-intensive aspects wherever possible and take advantage of on-line textual resources and corpus analysis software to at least partially automate the process, thereby maximizing human productivity.

The first step was to put together a comprehensive corpus of documents covering the entire domain of Caterpillar heavy equipment. The words in the corpus were "tagged" with their parts of speech by matching it against the Brown Corpus (a widely used linguistic research tool), which yielded the skeleton of a lexicon containing both single words and phrases. Candidate phrases were identified on the basis of part-of-speech patterns. In the next step, sets of corresponding source-target manuals were aligned automatically, following which the actual translations of terms were identified manually—the initial round being limited to French. The result was a bilingual English-French lexicon.

A special model was then constructed for the domain. Having a model restricted to a single tightly delimited domain enables the system to make more precise choices; it does not have to allow for the many contingencies that would be spurious for this purpose. This step involved constructing both *concept frames* and a *concept hierarchy*.

Other on-line tools were used to fine-tune KANT's grammar and mapping rules for both the source and target languages. On the source side, *lexical mapping rules* establish the link between a given word (tagged for one of its possible parts of speech) and a set of domain concepts. Lexical mappings are pointers to concept nodes, which are created automatically through the application of relevant grammatical rules. *Argument mapping rules*, in turn, are constructed from semantic roles identified by the lexical mapping rules and the semantic features present in the interlingua. On the generation side, *target language mapping rules* map the interlingua output onto target *f-structures* (a formalism used in lexical-functional grammar). Next the lexical mapping rules, created automatically from the lexicon, generate one-to-many target mappings. These are further refined by a target-language linguist, who then adds contextual patterns for use in selecting the desired target equivalents.

The team found that the use of automatic knowledge acquisition considerably reduced the human effort required to build up the system's knowledge sources for a particular application. This strategy, coupled with similar approaches using other tools that they intend to try, will undoubtedly reduce costs and therefore make a big difference in user acceptance of interlingual MT for commercial production purposes.

7. Final Thoughts

In the 1980s, transfer-based systems were far and away the most widely used workhorses of machine translation, although MT on PCs was emerging as a viable option for production translation. Knowledge-based interlingual MT was still in the laboratory. In 1989 FUJITSU debuted its ATLAS-II based on the interlingual approach. In the user survey mentioned at the beginning of this chapter (Vasconcellos 1993), all of the 15 sites in operation prior to 1990 used transfer systems. In the total

population of 36 MT users, transfer systems accounted for 14 of the 16 systems employed; of the other two, one was direct and the other was an entry-level interlingual system.

The news in the 1990s is that people are now experimenting at both ends of the spectrum. Direct systems are popular because they are affordable and easy to learn to manage, while interlingual systems, just starting to be used, are at least as costly as transfer systems and at least as difficult to update and maintain, if not more so—though they hold out the promise of less human intervention.

One point that was not mentioned before is that terminology cannot be ported easily between systems. Portability between different levels may well be out of the question. If terminology management is one of the main concerns, as it is for many MT users, the choice of a system, once it has been tailored to meet their needs, is nearly irrevocable. Some user sites have switched, but customization of the new system has proven to be a major adjustment not unlike from starting afresh.

For transfer and interlingual systems, multi-word entries are more difficult to encode than single words, as was seen above particularly in the case of ENGSPAN. This becomes a consideration for the management of terminology, since many technical terms have more than one word. At the same time, however, it is probably in the area of multi-word terms that MT makes its most valuable contribution to the maintenance of uniformity in large projects. In the survey of users, the advantage of MT that was cited most often was consistency of terminology. Without exception, these were high-volume users. There are sites that use MT for 17, 25, 30, and even 45 million words a year, and a number of them average around 10 million. In all, the 24 users that reported statistics produce 170 million words a year. With the exception of METEO, which does weather bulletins, and the operation at Wright-Patterson Air Force Base, the high-volume users are all engaged in some aspect of localization—i.e., translation to support the speedy marketing of products in other countries, ranging from product manuals to computer screen displays and even actual computer code.

These users are forging the link between MT and terminology, which will clearly be getting stronger in the coming years.

REFERENCES

American Translators Association. 1992. Report: PC-based MT products. ATA Committee on Machine Translation. Prepared by L. Chris Miller.

Brace, Colin. 1993. Making MT work. *Language Industry Monitor*, no. 13 (1-4), Jan-Feb.

Brown, Peter, et al. 1990. A statistical approach to language translation. *Computational Linguistics* 16:79-85.

Hutchins, John. 1993. Latest developments in machine translation technology. *Proc MT Summit IV* (Kobe, 19-22 July 1993). pp. 11-34.

León, Marjorie, and Lee A. Schwartz. 1986. Integrated development of English-Spanish machine translation: from pilot to full operational capability. Technical Report. Grant DPE-5543-G-SS-3048-00 from the U.S. Agency for International Development. Washington, D.C., Pan American Health Organization.

León, Marjorie, Susana Santangelo, and Muriel Vasconcellos. 1987. Terminology work and automatic translation Systems: A case study at the Pan American Health Organization. *TermNet News* (Vienna) 18:21-25, 1987.

Mitamura, Teruko, Eric H. Nyberg, 3rd, and Jaime G. Carbonell. 1993. Automated corpus analysis and the acquisition of large, multi-lingual knowledge bases for MT. *Proc Fifth International Conference on Theoretical and Methodological Issues in Machine Translation* (Kyoto, 14-16 July 1993). pp. 312-328.

Nagao, Makoto. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and Human Intelligence*, ed. A. Elithorn and R. Banerji, (Amsterdam: North-Holland).

Pearson, Jennifer, ed. 1992. ET10/66: Terminology and extra-linguistic knowledge. Reports 1 and 2 of the ET10/66 Consortium. Dublin: Dublin City University, National Center for Language Technology.

Ripplinger, Bärbel, ed. 1993. ET10/66: Terminology and extra-linguistic knowledge. Report 3, ver. 2, of the ET10/66 Consortium. Luxembourg: Centre Universitaire, Centre de Recherche Public.

Schütz, Jörg, and Bärbel Ripplinger. 1993. Machine translation supported by terminological information. *Proc Fifth International Conference on Theoretical and Methodological Issues in Machine Translation* (Kyoto, 14-16 July 1993). pp. 102-116.

Vasconcellos, Muriel. 1993. The present state of machine translation usage technology, or How do I use thee? Let me count the ways. *Proc MT Summit IV* (Kobe, 19-22 July 1993). pp. 35-46.

Wheeler, Peter. 1988. The translator and the dictionary experience. In *Technology as Translation Strategy*, ed. Muriel Vasconcellos, Binghamton (N.Y.): State University Press. American Translators Association Scholarly Monograph Series, 2. pp. 149-158.

_____. 1983. The errant avocado. *Newsletter of the British Computer Society Natural Language Translations Specialist Group*, no. 13, 1-6.