The Bilingual Knowledge Bank

A new conceptual basis for MT

Dr. Victor Sadler
BSO/Research

March 1989

THE BILINGUAL KNOWLEDGE BANK (BKB)

A new conceptual basis for machine translation

Victor Sadler
BSO/Research

SUMMARY

This paper proposes a way of integrating translation expertise, language-specific knowledge (monolingual and bilingual dictionaries and text representation), and extra-linguistic knowledge (general and specialised "knowledge of the world"), into a single, dynamic knowledge bank which can be constructed and updated semi-automatically from corpora and automatically from machine translation throughput.

1 INTRODUCTION

The concept of the Bilingual Knowledge Bank (henceforth "BKB") has grown out of ongoing research at the BSO software house in the Netherlands, on a system of semi-automatic machine translation called Distributed Language Translation (DLT).[1] Although it was not a part of the DLT (English to French) prototype system first demonstrated in 1987, and has not yet been implemented in any form, feasibility studies are already well advanced.

There are two major obstacles determining the speed and cost of development of a practical MT (machine translation) system. The first is the need to build large bilingual dictionaries. The second is the need to incorporate extra-linguistic knowledge into the system.

The degree to which extra-linguistic knowledge is really necessary is a matter on which not all MT researchers are agreed. But the need for a **large and detailed bilingual dictionary** is inescapable. Boitet (1987: 31) notes that

> *Ultimately, the cost of MT systems lies essentially in their dictionaries, which are quite difficult to construct and to maintain.*

Conventional dictionaries, however large, are no solution. Even if they can be automatically converted into machine-readable form, they rely heavily on human understanding for the interpretation of their entries. Information to be used by an MT system has to be far more explicit. Typically, conventional bilingual dictionaries contain lists of possible translations for each entry word, with little or no indication of the conditions under which one or other of those alternatives is to be selected – and certainly nothing which a computer could base a decision on. This example from an English-French technical dictionary (Ernst, 1984) illustrates the problem:

[1] distance (between points)/ distance f, écart m, écartement m, éloignement m, espace m, intervalle m.

[1] For a general description of the DLT project see Hutchins (1986: 287-291; 1988: 39-40) and Schubert (1986).

The computer requires precise indications as to when to choose one translation, and when to prefer another. The addition of selection cues such as the ever-popular semantic features, besides being highly labour-intensive, is often inadequate to ensure an appropriate choice of expression in the target language. Sometimes, as in the case of [1], the criteria are much more subtle.

Another deficiency of virtually all conventional dictionaries – both from the MT viewpoint and from that of the professional translator – is the limited cover they provide of the kind of structural transformations which the translator needs in nearly every sentence, e.g.:

[2] The board *unanimously confirms* this interpretation.
Le conseil *est unanime dans sa confirmation de* cette interprétation.[2]

If the computer is to produce high-quality translations, it has to know all the tricks of the translator's trade – and these are rarely to be found in existing dictionaries. Somehow, the expertise of the professional translator has to find its way into the machine.

Developing a workable bilingual dictionary for MT is a daunting task which requires an enormous investment in specialised human labour, since it cannot as yet be performed automatically within the state of the art in computational linguistics (CMT, 1988: 2). What's more, each language pair demands two bilingual dictionaries, since probably all existing dictionary structures for MT are one-way only. Sooner or later, a way of automating the dictionary-building process has to be found:

> *It has become clear that the construction of computer systems that process natural language requires the creation of large computerized lexicons with extensive and accurate syntactic and semantic information about words [...] . It is also clear that it will be impossible to build these lexicons in the number and sizes required with only the manual labor of individual computer scientists, linguists, or lexicographers. There are too many systems requiring too much information about too many words for the manual approach to succeed.*[3]

The first question, then, is how to automate the construction of large bilingual dictionaries, including extensive contextual cues for the selection of appropriate TL (target language) equivalents and an abundance of structural transformation rules.

As for extra-linguistic knowledge, it is generally acknowledged that **"understanding"** must play some part in any successful machine translation system. The question is only how large a part it should play (Hutchins, 1988: 12). Some problems can be solved by knowledge derived from the current text, as in

[3] He could not agree with the amendments to the draft resolution proposed by the delegation of India.[4]

where a correct translation into French, for example, is only possible when the attachment ambiguity has been resolved, i.e. if the translator (or MT system) knows whether India proposed the amendments, or the resolution. In other cases, general knowledge from outside the current text is required for ambiguity resolution, as in the notorious

[4] pregnant women and children

where, again, a French translation requires a decision as to whether the children are likely to be pregnant as well as the women. In either case – whether the knowledge required is

[2] Example adapted from Harris (1988c)

[3] Byrd *et al.* (1987)

[4] Example from Piron (1988)

available in the current text or only from other sources – it will only be accessible to the MT system when it has been stored in a suitable form or representation.

Research into knowledge representation for the purposes of machine translation has mainly concentrated on techniques of decomposition: building "deep" abstractions of meaning out of some arbitrary[5] set of semantic primitives, as independent as possible from the actual words of any specific human language. (See review in Hutchins, 1986: 272-284.) Yet many aspects of knowledge which are extremely relevant to translation – e.g. questions of time/tense, aspect, emphasis and focus – are delicately entwined with the form in which they are expressed (Tsujii, 1986: 659). For this reason, any knowledge representation which fails to preserve all the information expressed or implied in human language is of itself inadequate for the purposes of machine translation. Moreover, such decompositional methods are even more labour-intensive than the building of computer dictionaries has proved to be, and it is safe to say that no-one has yet developed a representation which is even remotely practicable for a large-scale system:

> *[...] the thought of writing complex models of even one complete technical domain is staggering: one set of manuals we have worked with [...] is part of a document collection that is expected to comprise some 100,000 pages. A typical NLP research group would not even be able to read that volume of manual, much less write the necessary semantic models, in any reasonable amount of time.*[6]

Another aspect of understanding which needs to be built into an MT system is the possibility of breaking out of the knowledge base and looking elsewhere for information. Just as a human translator is frequently obliged to turn to external information sources (encyclopaedias, colleagues, newspapers, the author of the text being translated, etc.) in order to arrive at a correct understanding of the text, so the computer too must have a means of accessing external knowledge, e.g. via a dialogue with the operator. This principle implies that the system must also have the means to explain the problem to the operator, and building this capacity into an MT system is by no means trivial.

It is from these two enormous and fundamental problems – of building huge dictionaries and constructing a comprehensive and open-ended knowledge bank – that the concept of a Bilingual Knowledge Bank was born: a structure which can function, at one and the same time, as a powerful, two-way bilingual dictionary and as a representation for all the various levels of knowledge relevant to translation, from the purely linguistic to the purely extra-linguistic or encyclopaedic, and which can to a large extent be constructed automatically.

2 BILINGUAL CORPORA AS KNOWLEDGE SOURCES

2.1 Linguistic knowledge

Given the aim of building a bilingual dictionary for an MT system by largely automatic means, and given the inadequacy of conventional dictionaries as source material, the problem now shifts to that of obtaining suitable input material for the dictionary-building program. Fortunately, such material is available in abundance. In most expert systems, the central problem is that of getting the human expert to formalize his or her intuition. The expert translator stands out among other experts by the simple fact that the application of the translator's expertise –

[5] Wilks (1972: 105) pointed out that there cannot be a *right* set of semantic primitives, only better and worse sets.

[6] Bennett & Slocum (1985)

unlike that of the surgeon or mechanic – always leaves a readable, and very often machine-readable, trace. In principle, it should be possible to devise a system to infer lexical equivalences and other translation rules from an analysis of the translator's actual output.

The idea of using bilingual text as an aid to dictionary construction is not entirely new. A recent experiment in this direction was reported by Brown *et al.* (1988), who applied statistical methods to a bilingual corpus (proceedings of the Canadian parliament) to extract a tentative glossary of lexical equivalences, using the basic assumption that the words of each English sentence correspond, in some unknown order, to the words in the corresponding French sentence. They recognized, however, that future methods should incorporate "the use of appropriate syntactic structure information".

In what Hutchins (1986: 319) qualifies as "speculative suggestions", Nagao (1984) proposed a system of automatic translation based on a set of example sentences:

> *We have to see as wide a scope as possible in a sentence, and the translation must be from a block of words to a block of words. To realize this we have to store varieties of example sentences in the dictionary and to have a mechanism to find out analogical example sentences for the given one.*

This amounts to using a kind of bilingual corpus as a dictionary of lexical transformation rules, or lexical "metataxis" rules in DLT terminology (Schubert 1987). Nagao suggests that this technique of translating by drawing an analogy between the phrase to be translated and some example phrase already encountered, is close to what the human language learner actually does when using dictionary examples to generate original sentences.

Nagao's proposal has since been implemented in a limited fashion by Sumita & Tsutsumi (1988) as a computer aid to the human translator. Their system uses a data base of equivalent example sentences in Japanese and English. Although no full syntactic parsing is done, the system maintains an index of function words appearing in the example sentences. At runtime, the pattern of function words appearing in the Japanese sentence to be translated is matched against the indexed patterns, and those example sentences which give the best match are retrieved and displayed for the operator, together with their English equivalents. The operator can then select whichever example is felt to be closest to the input structure, and edit the English version, replacing the content words as necessary.

Although Sumita & Tsutsumi list among their future tasks that of "enhancing this mechanism in order to generate target sentences automatically, as suggested by Nagao", it is still a far cry from such a CAT (computer-aided translation) implementation to the type of fully-automatic MT system envisaged by Nagao. There remains the unavoidable stumbling-block of lexical transfer, for one thing. (For this, Nagao proposed using a thesaurus to check on the similarity of the content words to be translated to those in the example sentences.)

A somewhat similar software support for human translation has recently been proposed by Harris (1988abc), under the name of "bi-text". Bi-text consists of a bilingual corpus, normally comprising the translator's own previous work, in which the source text and its translation are coupled together in parallel, unit by unit, using one or other hypertext system. The concept of "translation units" as applied here is defined by Harris (1988a) as follows:

> "*The translator's working segments of text are called* translation units *in the writings on the subject. We can say, using this term, that retrieval of a translation unit of ST [source text] from a bi-text will always bring with it the corresponding unit of TT [target text]. People who do not know much about translation tend to think the translation units are individual words, but in fact they mostly consist of whole phrases and even whole clauses or sentences. Bi-text therefore binds together not the individual words of ST and TT but those somewhat longer segments.*"

The translator delineates these "working segments of text" in such a way that it is possible to output one segment of translation in its more or less definitive form before starting on the next segment. Suitably indexed, the bi-text corpus would enable the translator held up by a particular expression or technical term, to check whether the same expression has turned up before and, if so, how it was translated on that occasion. From the above definition of a translation unit it is clear that Harris's proposal is aimed primarily at multi-word expressions, since it is these complex translation units, rather than simple word-for-word equivalences, which cost the professional translator time and trouble. For an MT system, on the other hand, the problem of dictionary building is much broader, as emphasized in the Introduction.

Other researchers who appear to be moving in a similar direction are the group at Pisa (Calzolari, 1988), although it is not clear whether their work actually extends to bilingual corpora.

The present proposal for a Bilingual Knowledge Bank for MT contrasts with the suggestions of Nagao and Harris, and with the experiments of Brown *et al.* and Sumita & Tsutsumi, firstly in its insistence on full syntactic analysis of the bilingual corpus. As Boitet (1987: 31) also emphasizes:

> *The study of* parallel corpuses *of texts and their translations into one or several languages should lead to interesting results, but they should be based (at least) on structured representations of the texts.*

Where the BKB concept also breaks new ground is in its combination of two separate dimensions: the horizontal dimension of cross-language equivalence and the vertical dimension of text coherence. This two-dimensional structure allows the BKB to represent not just lexical and sentence-level linguistic knowledge, as in Nagao's database of example sentences, but the intersentential relations of discourse structure as well. Instead of an arbitrary collection of example sentences, the proposed BKB structure consists of large amounts of continuous text, or bi-text, in which textual coherence is made explicit by the analysis and tagging of all forms of reference, and which automatically and progressively incorporates the text currently being translated. By the formal definition and coding of translation units, it allows for linguistic knowledge to be accessed at any level from the morpheme to the overall text structure, thus doing away with the need for separate dictionaries of word-level equivalences or verbal case-frames.

As compared with traditional methods of lexicography and the writing of conventional metataxis rules, this corpus-based approach takes advantage of the fact that vast amounts of human translation expertise are already available in a highly accessible form - namely as texts and their translations. What grammars, dictionaries and formal translation theory tell us to do, and what the expert translator actually does, are two very different things. A musical analogy may help to underline the point.

> *[...] at IBM there is now a computer that composes Bach chorales. Well, almost. [...] For the computer to harmonize a 20-bar piece of music, it needs [...] 350 separate rules, all drawn from analysis of the 300 chorales the German composer actually wrote in his lifetime. [...] Kemal Ebcioglu [...] complains that when he programmed a computer with only the harmonization rules from orthodox music theory treatises, he got tunes with a mechanical, computer-loop sound. The additional couple of hundred rules - which Mr. Ebcioglu then wrote based on study of the chorales - come out of the gap between what Bach was taught to do and what he intuitively did.*
>
> - Washington Post, 31 August 1988

The Bilingual Knowledge Bank is a device for getting the human translator's intuition into the computer. May we hope that it will prove to be the tool needed to get the "mechanical, computer-loop" quality out of machine translations?

2.2 Extra-linguistic knowledge

Having established the aim of using a kind of structured bi-text as a bilingual dictionary for MT, let us now turn to the second major developmental headache: the acquisition and representation of extra-linguistic knowledge. In the Introduction, I already suggested that the knowledge base must be open-ended to allow for interaction with the operator whenever the system's own knowledge proves inadequate to resolve a particular ambiguity.

But there is still another sense in which the knowledge base needs to be open-ended. Boitet (1987: 32) has put the problem in a nut-shell:

> *Even if a big knowledge base is available, no machine analysis of a text can be 100% correct, because* new *knowledge is usually introduced by the translated text. But no adequate learning method is yet able to dynamically modify and enrich the knowledge base.*

During translation, it is necessary to build up a structured representation of the text which has already been translated, in order to cope with problems of text coherence - in particular, deixis, reference and theme/rheme (Papegaaij & Schubert, 1988: 196-197). I shall refer to this structured representation as the **text representation**. Now it can be argued that this text representation has much in common with the representation of "encyclopaedic" or "hard" knowledge, in that it has to deal both with **specific concepts** such as *President Bush*, and with **generic concepts** such as *heads of state*, and has to establish various kinds of relations between the concepts identified.

Now consider the sentence

[5] This stops the motor and applies the electromagnetic brake.

It is clear from the use of definite noun syntagmata (*the motor* and *the electromagnetic brake*) that these are being used to refer to concepts already familiar to the reader. Familiarity exists, in this particular case, by virtue of an earlier specification in the same body of text (an aircraft maintenance manual). For example, *the motor* in question had already been specified (some 200 lines earlier) by:

> 2. Component Description
> A. Electric Motor (Refer to Fig. 3)
> (1) The electric motor is a dc motor which is a part of the flap-power drive-unit in the LH nacelle.

However, it should not be assumed that the original specification of a given referent is necessarily to be found in the recent context. A definite noun may well refer back to a specification introduced several chapters earlier, and of course this may be explicitly indicated, e.g. *(See Chapter 2, Section A)*. Or consider the techniques applied by literary writers (e.g. Wouk in *The Winds of War*), where the narrative may switch between chapters from one country to another, taking up the threads of separate stories again and again. The reader is assumed to be capable of immediately retrieving the referents from earlier chapters, without any explicit help from the author.

On the other hand, of course, many definite noun phrases refer back, not to the recent context, but to the general knowledge the reader is presumed to possess. Thus a text which begins with

[6] The world is getting smaller.

assumes that the reader will understand which specific world is indicated.

In the case of a computer system, knowledge is necessarily textual. The computer has no experience of outside reality and can construct a picture of that reality only from digital data fed in. It follows that if we expect a computer system to be capable of "understanding" a reference to general knowledge, we are assuming that the general knowledge required has been fed into the system in digital form.

This raises the question of what exactly constitutes a "text", as far as machine understanding is concerned. If all previous experience is basically textual in nature, as it must be in the case of the computer, where do we put the borderline between the "current" text and the rest of the material which has been fed to the computer in the past? Maintaining text coherence in translation and identifying referents in the text representation, can certainly not be achieved only on the basis of the current paragraph or even the last chapter, as the above examples have shown. How far back should the system search in its (textual) experience in order to instantiate a reference? The last 10,000 words? The text accumulated since the start of the current translation session? Everything since the same time last week?

Of course, we can always define an arbitrary limit. But the point being made here is that it is **arbitrary**. Whereas for humans, there is a clear division between text and non-text, between a piece of writing and a piece of pizza, for the computer this division is non-existent. This suggests that the representation of the "current" text (whatever its limits may be) and the representation of "general knowledge" (which amounts to "non-current" text) should be similar. There is probably no good reason for building different types of structure to represent the meaning of these two blocks of text, the "old" and the "new". We may want to store the older material in a more compact, less redundant form, but this need not imply a basic difference in structure.

These considerations lead us to an important conclusion. This is, that the best available means of representing knowledge in the machine, just as for human beings, may be human language. Attempts at building some kind of abstract, non-linguistic knowledge representation may be misguided.

> *Ever since Descartes, it has been assumed that real knowledge must be mathematical in nature: either mathematics itself or the so-called exact sciences that mathematics supports. Concomitantly, it has also been assumed that so-called verbal or language-based knowledge must be in some way inferior, since language does not easily lend itself to mathematical precision. But now, inadvertently, unexpectedly, and with unforeseeable consequences, through such concepts as hypertext and its inevitable spinoffs, language may at last be in a position to make a comeback on the knowledge ladder.*[7]

The next important conclusion is the following. If the representation of the translated text and the representation of general knowledge share a common structure, and if the former can be built up semi-automatically during the translation process, then surely general knowledge can also be acquired in the same fashion.[8]

[7] Gross (1989: 44)

[8] I say "semi-automatically" because it is a basic feature of the DLT translation strategy to have the computer consult the operator whenever automatic procedures fail to resolve an ambiguity. This computer-initiated dialogue, already implemented in the prototype system, takes place in the source language only and thus does not require the operator to possess any knowledge of the target language.

But what of the obvious pitfalls to be expected if human language is to be used as a knowledge representation? What of structural, referential and lexical ambiguity? I shall return to this question in section 4 below, but the quick answer is this. If the bilingual dictionary for MT can be replaced with a structured bilingual corpus, and if extra-linguistic knowledge is also represented by a structured text corpus, then the two structures can be integrated into one. The dictionary and knowledge bank (and text representation) are conceptually one and the same. The consequence is that the representation of extra-linguistic knowledge is also a syntactically structured body of bi-text. As such, it contains no structural ambiguity, since this is required to be eliminated during parsing; no referential ambiguity, since this must be resolved during BKB construction, just as it is during translation; and little lexical ambiguity, since every lexical unit in one language is tied to an equivalent unit in the other language: monolingual lexical ambiguity is greatly reduced by the constraints of the other language in the BKB. For example, the highly ambiguous English word *line* will always be coupled, in the BKB conception, with a specific translation in the other language. If the other language is Esperanto (which forms the intermediate language or pivot in DLT's multilingual architecture), and if the translation is *tubo*, for example, then the concept it represents is restricted to that of a pipeline, eliminating all the other meanings of both *line* and *tubo*. Of course, some shared ambiguity may still remain, but it can be argued that any further disambiguation beyond this point is largely irrelevant to the requirements of the translation.

In sum, the BKB is **unambiguous**, at least for the practical purposes of translation.

3 THE STRUCTURE OF THE BKB

In this section we are only concerned with devising an appropriate structure. The question of how such a structure can be built up semi-automatically will be answered later, under 4.

3.1 Translation units

The raw material from which a BKB is constructed is bilingual text, which we can define as two bodies of text which are asserted to be equivalent in meaning. Whether one of the texts is a translation from the other, or they are both translations from a third language, is unimportant. For the sake of illustration, suppose the corpus consists of the following sentence, a modified version of [2] above:

[7] The board of PAC unanimously confirms the mandate.
Le conseil du PAC est unanime dans sa confirmation du mandat.

The first requirement is that the text be assigned a syntactic structure. Figure 1 shows dependency trees for this example.[9] The choice of dependency syntax for DLT has been abundantly motivated elsewhere (e.g. Schubert, 1987: 193-194). Schubert's argument that constituency syntax is at first hand concerned with syntactic **form** and dependency syntax with syntactic **function**, and that the latter is therefore more suitable for the purposes of translation, is obviously equally applicable to the purposes of a bilingual dictionary. But it can also be argued that this emphasis on syntactic function, which implies relations between words, also favours dependency syntax for knowledge representation, where relations between concepts are of vital importance. An additional point in favour of dependency is the smaller number of tree nodes required in comparison with constituency analysis. Where very large corpora are concerned, this compaction is significant.

[9] For an explanation of the syntactic function labels, see Appendix C.

The next step is to divide the syntactic structure into translation units. A translation unit, as the term was used by Harris, consists of two fragments of text in different languages, which the translator considers equivalent. The essence of a unit is that it is autonomous. It can be used without necessarily causing alterations in the surrounding context. It may very well, of course, be sensitive to context, in that the choice of one TU (translation unit) or another will usually depend on the context in which it appears. But it will not, when selected, necessitate changes in the context, in particular, that part of the text which has already been translated.

Figure 1: Dependency trees for example [7]

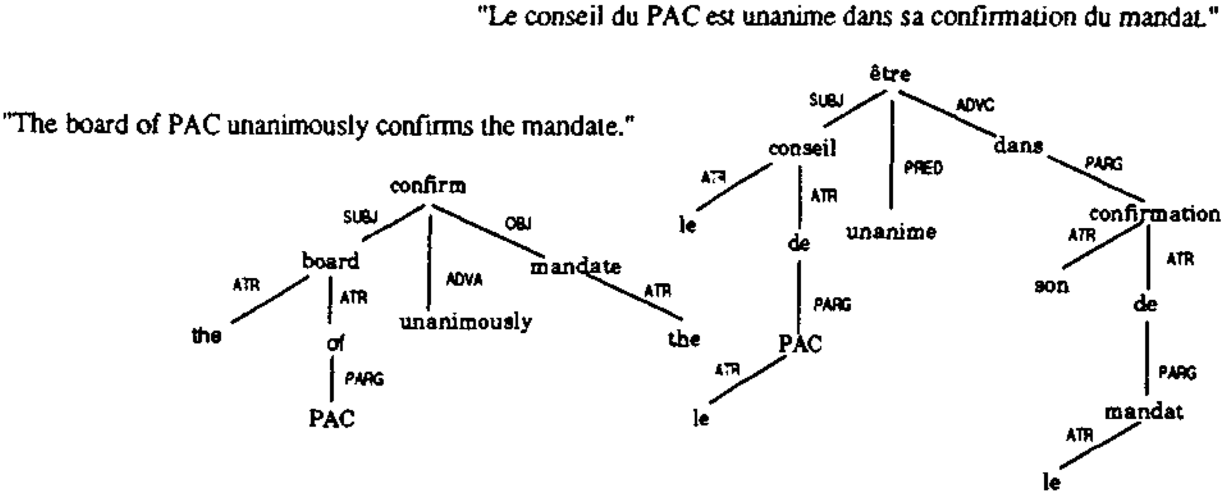
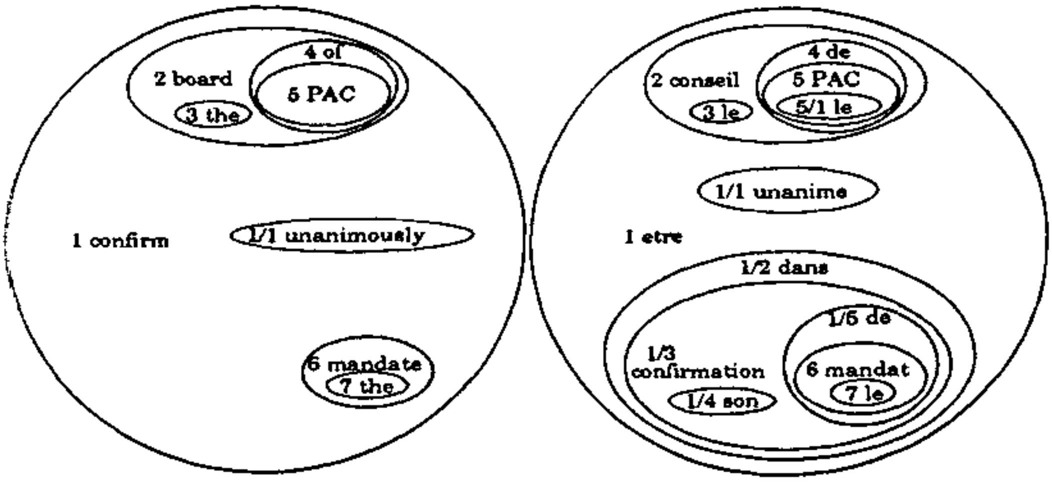


Figure 2 provides another view of example [7], this time in terms of translation units. This view is isomorphic with the more conventional tree diagram, with each ellipse in Figure 2 corresponding to a sub-tree in Figure 1. Each of the seven identifiable translation units has been assigned an identification number (ID).

Figure 2: English-French translation units for example [7]



The first part of Table 1 lists the TU numbers with the corresponding bilingual equivalences. For example, TU 1 identifies the complete sentence, governed by the verb *confirm* in English and *être* in French, and enclosed within the largest circle. TU 2 is the subject noun phrase, 3 the determiner, 4 the prepositional phrase, etc. From these examples it should be immediately clear that every translation unit corresponds to a (sub-)tree. On the other hand, it is not necessarily the case that every sub-tree corresponds to a translation unit. The French sub-tree governed by *dans*, for instance, does not constitute a translation unit. There is no sub-tree in the English sentence which can translate *dans sa confirmation du mandat*. In the BKB coding, this is shown by the ID "1/2" attached to *dans*, which indicates that this sub-tree is the second bound dependent of TU 1.

Table 1: Translation units inherent in an example sentence

English: "The board of PAC unanimously confirms the mandate."
French: "Le conseil du PAC est unanime dans sa confirmation du mandat."

TU coding	English phrase	French phrase
1	The board ... mandate.	Le conseil ... du mandat.
2	the board of PAC	le conseil du PAC
3	the	le
4	of PAC	du PAC
5	PAC	le PAC
6	the mandate	le mandat
7	the	le
1-2-6	unanimously confirm	être unanime dans sa confirmation de
2-3	board of PAC	conseil du PAC
2-4	the board	le conseil
2-3-4	board	conseil
4-5	of	de
6-7	mandate	mandat

Harris's statement that "translation units [...] mostly consist of whole phrases and even whole clause or sentences" is borne out by this diagram. But of course, translation units can also consist of individual words. Word-for-word correspondences are not as infrequent as the quotation might suggest. Their frequency depends in part on the type of text being translated (e.g. technical or literary) and the demands made on style in the target language. In technical writing, where terms in various languages are usually intended to refer to identical real-world objects or processes, one-to-one lexical equivalences are quite common. Moreover, the kind of stylistic somersaults performed by literary translators are usually avoided in the down-to-earth style of technical translation.

Harris's bi-text proposal is primarily concerned with literal equivalences, but he also recognizes the need for non-literal TUs, based on similarity. The translator cannot rely on always finding exactly the same literal expression in the bi-text base. The BKB structure allows for the replacement of sub-trees within an existing TU by a process of tree subtraction, which amounts to a kind of generalization. This permits the translation system to make productive use of all the equivalences in the BKB, even if they do not constitute independent sub-trees. For example, subtracting TU 2 from TU 1 in Figure 2 yields the equivalence of *to unanimously confirm the mandate* with *être unanime dans sa confirmation du mandat*. Further subtracting TU 6 generalizes the verbal construction to *to unanimously confirm* and *être unanime dans sa confirmation de*. The lower part of Table 1 lists the remaining possibilities and the corresponding coding. In this way an expression such as *to tamper with (something)*, which may very well never occur in the corpus without a dependent, can still be accessed as a (generalized) TU.

One further restriction can be added to our definition of the BKB structure in terms of translation units: every word in each monolingual text should be a part of at least one translation unit. It will often be the case that a word in one language is not translated explicitly in the other. But we will not permit translation units of the type

[word] = []

The motive behind this is a strong one: we want to build a two-way translation tool. Rules of this kind would not be workable in the reverse direction, since they would always apply!

One element of Figure 1 was omitted from Figure 2 for the sake of simplicity: the syntactic labels on the tree branches. In Figure 2, these should be visualised as located on the ellipses. Table 2 shows the TU-coded structure in full.

Table 2: Dependency trees for example [7], coded for translation units

```
[GOV 1,confirm                [GOV 1,être
 [ADVA 1/1,unanimously ]       [PRED 1/1,unanime ]
                               [ADVC 1/2,dans
                                [PARG 1/3,confirmation
                                 [ATR 1/4,son ]
                                 [ATR 1/5,de
 [OBJ 6,mandate                   [PARG 6,mandat
  [ATR 7,the ]]                    [ATR 7,le ]]]]]
 [SUBJ 2,board                 [SUBJ 2,conseil
  [ATR 3,the ]                  [ATR 3,le ]
  [ATR 4,of                     [ATR 4,de
   [PARG 5,PAC ]]]]              [PARG 5,PAC
                                  [ATR 5/1,le ]]]]]
```

Here again, a translation unit consists of a head word and all its dependents. Thus TU 2 consists of the head words *board* and *conseil*, respectively, and all the remaining dependents, namely TUs 3 and 4, which in turn consist of the head word *of* and its dependent TU 5. Note that a certain amount of normalization has been applied to the words on the nodes: the verbs have been reduced to their basic forms, the French *du* has been split into its constituent preposition and article, and the possessive pronoun *sa* has been normalized to *son*.

A larger sample of text coded for translation units is to be found in Appendix B, which represents the parallel dependency structure for the text in Appendix A. This contains an example text for writers using Simplified English as defined in the international aircraft industry (AECMA, 1984), together with a translation in Esperanto. In the Esperanto version, the morpheme structure has also been made explicit to some extent, in order to illustrate the possibility of coding morphemes as translation units, e.g.:

```
[3,tank                 [3,((3.1,al)(fuel)ujo)]
 [ATR 3.1,wing ]]
```

TU 3.1 (decimal numbers have no special significance) consists of the word *wing* in English and of the morpheme *al* in Esperanto, which is part of the word *alfuelujo*.

Two TUs can differ in one language while being identical in the other. For example:

```
[GOV 1,test            [GOV 1,testo
 [ATR 2,3,tank          [ATR 2,de
  [ATR 3.1,wing ]]]      [PARG 3,((3.1,al)(fuel)ujo) [la] ]]]
```

In Esperanto TU 2 is headed by the preposition *de* and includes all its dependent elements. In English, on the other hand, TU 2 has no distinctive governor but simply consists of TU 3.

3.2 Text coherence and extra-linguistic knowledge

I have claimed in section 2.2 above that the text representation required for the analysis of the text being translated is adequate to represent extra-linguistic knowledge as well, at least for the purposes of MT. What kind of additions to the BKB structure, as described so far, are necessary for knowledge representation?

Undoubtedly the most important relation which has to be added to the structure is that of reference. We need to be able to follow the various items mentioned and the events relating to them, throughout the text and throughout the knowledge base. Different expressions referring to the same concept must be linked via pointers. Besides identity, other reference relations such as PART-OF, MEMBER-OF etc. can also be used. (For the interactive identification of such relations see section 4 below.)

This augmentation should be language-specific, because one language may be able to refer back to a concept which has not been explicitly introduced in the other. For example:

[8] *English:*
I often had headaches.
They were worst during rehearsals.
Esperanto:
Mia kapo ofte doloris. ('My head often ached.')
Ĝi pleje doloris dum provludoj. ('It ached most during rehearsals.')

In order to illustrate the combination of reference identification and coding with that of translation units, a larger text sample is required. In Appendix C, the same structures already shown in Appendix B are augmented with vertical links between the TUs to identify the references they contain. At the same time, identical surface forms are replaced by their original TU code for the sake of compaction. A few concrete instances may help to make this clear.

The first examples concern conceptual references, i.e. they identify those translation units whose meanings overlap. The form of such co-referent TUs may or may not be identical. Conceptual references are shown between curly brackets directly following the TU code. The sample text illustrates two kinds of conceptual reference: the identity and MEMBER-OF relations.

Identity between the concepts represented by two translation units is marked by an equals sign, e.g.:

```
[107{=83}:83-85            [107{=83}:83-85
 [109,valve [the]            [109,((109.1:70.1)valvo) [la]
  [ATR 109.1:70.1 ]
  [ATR 110:66]]]]             [ATR 110:66]]]]
```

Here, TU 107 (*the light for the right-hand shutoff valve*, in English) is identified with TU 83 (*the light for the shutoff switch of the right-hand outer wing tank*), because these two different forms in fact refer to the same entity. It does not follow that the TUs are interchangeable in translation. One may be more appropriate than the other in a particular context. This referential identification is important in order to impart an explicit structure to the knowledge of the world implicit in the text. The system can make use of this structure for simple inference procedures. (See examples under 5.2 below.)

The MEMBER-OF relation is marked by a "<":

```
[SUBJ 83{<27}:18-19        [SUBJ 83{<27}:18-19
 [84:43                     [84:43
  [85:70 ]]]                 [85:70 ]]]
```

The coding "83{<27}" means that the entity referred to by TU 83 (in English, *the light for the shutoff switch of the right-hand outer wing tank*), is a member of the set referred to by TU 27 (*the shutoff valve lights*). From this identification the system can infer that TU 107, which has previously been identified with TU 83, is also a member of the set referred to by TU 27, a fact which had not been given explicitly. (All these extra-linguistic relations are, of course, equally applicable when the same concepts are referred to by the corresponding terms in the other language.) In this way the system can automatically check and improve the consistency of the knowledge base.

Given that various surface forms of reference can be projected onto the same extra-linguistic entity, the question arises of whether it is useful or necessary to preserve the surface variety. Pronouns, for example, cannot be translated directly between say, English and Turkish, or between English and Japanese, without reference to the entity they represent, and even then quite complex choices may have to be made on the basis of broader knowledge of the discourse context – questions of physical proximity in the case of Turkish, or of presupposition in the case of Japanese (Tsujii, 1988: 161). So why not discard the surface forms from the BKB, preserving only the code reference?

The answer is that part of the surface reference may need to be preserved because it **adds** information to the original description. Given, for example, the text

[9] My secretary will arrive at three.
Please pick him up at the airport.

we could replace the pronoun *him* with the ID for *my secretary*, but the feature "sex: male" would first have to be added to the referent. Paraphrases, too, may contain information which can enrich the original description. In the example

[10] There was a girl sitting on a beach-mat.
They could see the young woman with the binoculars.

the expression *the young woman*, even if known to refer to the same entity as *a girl sitting on a beach-mat*, adds to the original description the fact that the girl in question could also be considered a woman. An additional desideratum for the BKB is the possibility of regenerating the original text in its literal form. This would not be possible if reference forms were discarded.

The remaining examples illustrate cross-references concerning form, rather than content. A TU code followed by a colon and a second code indicates that the TU identified by the first code is **identical in form** to that identified by the second code, e.g.:

```
[ATR 97:6]    [ATR 97:6]
```

This notation means that TU 97 has exactly the same form (in both languages) as TU 6. Considerable compaction can be achieved by this kind of coding. In this example, TU 6 has the following structure:

```
[ADVA 6,on                [ADVA 6,sur
 [PARG 7,panel [the]       [PARG 7,((komand)panelo) [la]
  [ATR 7/1,control
   [ATR 8,fueling ]]]]       [ATR 8,por
                              [PARG 8/1,((fuel)izado) ]]]]
```

A repeated structure can, of course, accept new dependents. In this case they are attached by default to the head word. Thus:

```
[PARG-C 65:57.1           [PARG-C 65:57.1
 [ATR 66,right-hand ]]     [ATR 66,dekstra ]]
```

TU 65 is a new unit with exactly the same form as TU 57.1, except that the new TU 66 is attached as an attribute to the head word of 57.1.

A new TU may be only partly identical with a previous structure. In this case the coding shows those dependents which are not repeated, as subtracted from the main TU:

```
[GOV 37:13-17-21-26    [GOV 37:13-17-21-26
   [41,come                [41-as,((ek)lumi)
    .....]                  .....]
   [49,stay                [49-as,(lum)adi
    .....]                  .....]
   [54,flow                [54-as,flui
    .....]]                 .....]]
```

The figures "37:13-17-21-26" indicate that the new TU 37 has the form of TU 13, after subtraction of TUs 17, 21 and 26. The new dependents (41, 49 and 54) are attached to the points of subtraction in the same order.

Where the number of subtracted dependents is not equal to that of the new additions, the attachment points have to be made explicit. In such cases the new TU code is followed, between round brackets, by the code of the dependent it replaces, e.g.:

```
[90:54-54.1-57       [90:54-54.1-57
 [93(57):65 ]]        [93(57):65]]
```

This coding means that TU 90 has the form of TU 54, after subtraction of TUs 54.1 and 57, and that TU 57 is replaced by TU 93, which happens to have the same form as TU 65. All of these relations obtain for both languages.

4 COMPILING THE BILINGUAL KNOWLEDGE BANK

The building of a Bilingual Knowledge Bank entails a great deal of interactive text processing. Even if a suitable corpus of bilingual text is available and the text in each language has been parsed with the aid of an appropriate dependency parser, the conversion of the parallel dependency trees to the proposed BKB structure cannot be performed automatically. However, it does appear that a great deal of the work can become automatic. There are two reasons for this. First, the information contained in one language version can be applied to the other, and the addition of further languages to the system can reinforce this effect. Second, the BKB itself can provide more and more support, the larger it becomes.

The human-aided processing required can be described under three separate headings: structure, translation and reference.

4.1 Structural disambiguation

The dependency parser may very well produce multiple parses. Such structural ambiguities must be resolved. The same applies to structural ambiguities within words. A word grammar module should identify the same structural ambiguity in the Dutch *stoomtreinmuseum* as the English parser should recognize in *steam train museum*.

However, not all structural ambiguities are likely to be common to both languages in the corpus. On the contrary, a small experiment on paper has shown that of 20 structural ambiguities identified in a short passage in Esperanto[10] for which translations in eight other languages were available, between 8 and 14 could in principle be resolved automatically by comparison with one or other of the eight translations. Comparison of the possible parses of the Esperanto text with those of both the English and the German version together resulted in the elimination of no less than 16 of the 20 ambiguities. The remaining 4 cases were ambiguous in all of the nine languages and would therefore have to be solved by the operator.

These observations suggest the following strategy. Given the aim of building, not a bilingual, but ultimately a multilingual translation system, it may be profitable to put off the interactive resolution of ambiguities in the construction of the knowledge bank for as long as possible. Those structural alternatives which cannot be eliminated automatically by comparison of the possible structures for the first language pair may be left until a third language is processed for the BKB, and so on.

The larger the BKB becomes, the better it can support the disambiguation process. This is not to say that certain alternative parses can be automatically excluded at this stage on the basis of the existing knowledge, but it does mean that the possible structures can be ordered on the basis of likelihood, thus easing the job of the operator. Existing knowledge of the death of Maxwell Madondo can help to resolve the attachment ambiguity in:

[11] The girl lived in the same street as Maxwell Madondo, one of the bodyguards of Mrs Mandela, who last week was stoned and stabbed to death.[11]

If it is not decided to put off the disambiguation dialogue for as long as possible, and all structural ambiguities are resolved for the first language pair, whether automatically or interactively, then there should be very little of this kind of work required from the third language onwards, because each new language can be disambiguated by comparing the alternative parses with the known structures for the languages already processed. Residual structural ambiguity will be found mainly in sentences which strongly deviate from the other language versions (i.e. where the translation is very "free"), and in idiomatic expressions. For example, choosing the correct attachment for the prepositional phrase in *to pull the wool over someone's eyes* is unlikely to be helped by a comparison with the equivalent idioms in other languages.

4.2 Identifying translation units

The second dimension of BKB construction in which human support is inevitable is the cross-coupling of the parallel structures by means of translation units. At the beginning, the operator

[10] Part of the Preface to Munniksma (1975)

[11] Translation from the equally ambiguous Dutch original *Het meisje woonde in dezelfde straat als Maxwell Madondo, een van de lijfwachten van mevrouw Mandela, die vorige week is gestenigd en doodgestoken.* (Utrechts Nieuwsblad, 26 Feb. 1989).

obviously has to do most of the work. (Moreover, in contrast to the task of structural disambiguation described above, TU identification demands bilingual knowledge from the operator.) Gradually, however, the growing knowledge in the BKB under construction makes it increasingly easy for the system to suggest the correct equivalences. This can be demonstrated by reference to Appendix C. In this sample text, consisting of ten sentences, roughly 50% of all translation units are repetitions. At the beginning, all the expressions are new, but towards the end of the text, very few new concepts are introduced. It seems likely that in a large corpus a high proportion of all sentences could be automatically analysed into translation units – that is to say that the system could recognize that a given sentence and its equivalent in the other language can be put together from the building bricks of known TUs, without remainder and in a unique fashion.

If a given sentence can be put together in this way, then it might be thought that it adds nothing new to the BKB and could therefore be discarded. This will never be the case, however, unless the same corpus text is fed in in duplicate. Even if a sentence can be constituted from known translation units, their combination may form new, more complex units. The relations between the TUs in the sentence provide contextual information which is relevant to the choice of translations in context. Finally, even if the whole sentence is identical, both in form and in referential content, to an earlier sentence, its links to other sentences in the text add new information at the level of discourse analysis.

Experiments on paper suggest that the identification of translation units can to a high degree be regarded as a **transitive** process. That is to say that, given the equivalence of expression α in language A with expression β in language B, and further given that β is equivalent to γ in language C, then it follows that a translation unit can be established between α in language A and γ in language C in the same context. An important implication of this principle for the development of a multilingual system is that given the BKBs for the language pairs A-B and B-C, the knowledge base for the language pair A-C can be derived automatically. The only disadvantage to this procedure is that some TUs in the automatically generated BKB may be unnecessarily large. For example, the Spanish *deudor hipotecario* is equivalent to the English *mortgagor*, which in turn can be translated into Esperanto as *hipoteka debitoro*. If the Spanish term is now cross-coupled, using the transitivity principle, to the Esperanto term, the result is a translation unit which could in fact be further subdivided, since the dependent attributes are also equivalent. Although this failure to subdivide will not necessarily cause problems during translation, an additional interactive process could identify such cases and thus improve the productivity of the new BKB.

4.3 Identifying referents

The third dimension of BKB construction requiring interactive processing is that of text coherence: the vertical linking of translation units which refer to the same entities or events. This is essential for the use of the BKB for artificial intelligence processes, as well as for the generation of appropriate surface forms of reference in the target language. The first two dimensions described under 4.1 and 4.2 above (structural disambiguation and identification of translation units) were necessary in order to convert the bilingual corpus into a bilingual dictionary. This third dimension promotes the bilingual dictionary to the rank of knowledge base.

In establishing referential links between TUs, the degree of conceptual specificity is very important. In many languages such distinctions are cued by determiners. Just as in the case of structural disambiguation, contrastive analysis can aid this classification and identification of concepts. In

[12] English: Resolutions have become a necessity. (generic)
French: Les résolutions sont devenues une nécessité.

English: Resolutions have been adopted. (specific)
French: Des résolutions ont été adoptées.

the French determiners, when contrasted with the lack of determiners in the English version, make the generic/specific distinction clear.

The identification of specific references can itself be supported by contrastive analysis. E.g.:

[13] English: This dictionary is the fruit of more than nine years of international collaboration. In planning it, ...

German: Dieses Wörterbuch ist die Frucht einer mehr als neunjährigen Arbeit. Bei seiner Ausarbeitung ...

where the German pronoun *seiner* makes it clear that the English *it* in the second sentence refers back to *dictionary* and not to *fruit* or *collaboration*. What proportion of such referential ambiguities can be resolved by such contrastive means is difficult to estimate, in particular because it is bound to be dependent on the specific language pair concerned. In combination with a set of text coherence rules for each language, contrastive analysis should, however, significantly reduce the burden on the operator.

Just as in the case of structural disambiguation, the referential identification completed for the first language pair should virtually eliminate this aspect of the operator's task from the third language onwards, unless it is decided to put off the interactive processing until a number of languages have been processed automatically. There will, of course, always be a residue of monolingual references to be resolved. (See example [8] under 3.2 above.)

4.4 Example of procedure

Let us look now at the sample text in Appendix A and its BKB representation in Appendix C. Suppose that each version (English and Esperanto) has been parsed. Assume also that we are starting with an empty BKB. The basic problem is now: how do we identify the translation units?

Let us suppose, for the sake of simplification, that there is a one-to-one correspondence between sentences. Then the first TUs the BKB constructor module can identify are complete sentences. In the sample text, the title has also been treated as a sentence. We can assign the first ID (TU identification number) to the tree governed by *test* in English and *testo* in Esperanto. Note that this ID relates to the whole tree. It does not tell us that *test* = *testo*.

Matching should now proceed bottom-up. This is because the linking of any pair of sub-trees between the parallel tree structures requires prior matching of its parts. In order to avoid repeating this matching, it is better to start with the parts.

The procedure now is to start with one language and look for correspondences in the other. Taking the last sub-tree in the title, which consists of the single word *outer*, the constructor searches its knowledge bank for prior occurrences of this governor. As the BKB is still empty, the search is unsuccessful. The process is repeated for the other nodes, all in vain. The next step is to start guessing. The word *outer* is an adjective under an ATR (attribute) label. Looking through the Esperanto tree we find only one word which has these same attributes: *ekstera*. The constructor can now consult the operator as to the validity of the TU, with the query:

[outer] = [ekstera] ?

The reply will be positive, so an ID can be assigned to these two sub-trees. (The sequence of numbers in the Appendix is not intended to represent the order in which they are assigned.)

The identification of

[wing] = (al)

may require a little trial and error, since *fuel* is also a possible candidate, but then the only sub-tree left to identify is

[tank [ATR wing] [ATR outer]]

Not all potential equivalences need be identified explicitly. The equivalence of

[test [ATR ...]] = [testo [ATR ...]]
[tank [ATR ...]] = [((fuel)ujo) [ATR ...]]

can be inferred by a process of tree subtraction.

A word about syntactic labels. The labels used in the sample structure have intentionally been made as symmetrical as possible. This is not of course necessary. Even if we use the same literal label, e.g. 'SUBJ', in two different syntaxes, the meaning of each is defined by the relevant syntax, and they are not necessarily the same. In practice, however, when the constructor is searching for possible equivalents in the two languages, it will use the default rules for label transfer which must always exist at the lowest level of metataxis (Schubert, 1987: 148).

Turning to the first complete sentence of the sample structure, note that the difference in structures in sub-tree 7 prevents any independent translation of *fueling control*. The code 7/1 cannot refer to an independent TU.

In sentence (1)(a), we encounter reference phenomena for the first time. Working bottom-up from the first sub-tree, the constructor searches for the word *power* in the BKB and retrieves TU 10.1 from the previous sentence. This enables it to link the new occurrence to the translation *la alimento* in the Esperanto version. Here, the definite article in Esperanto suggests that a definite reference is involved: we are probably talking about the same 'power' previously mentioned in *power switch*. Checking reference identity with the operator involves slightly more than just presenting equivalences. The question that needs to be asked here is something like

Is **power** in "power light" the same **power** as in "set the power switch to on" ?

After operator confirmation, the constructor then augments the new TU by setting a pointer to the referent.

Finally, some clarification is called for at the very last reference in Appendix C, where *the fuel flow* is identified with the earlier TU 90, *fuel flows into the right-hand tank*. The definite article indicates that the fuel flow concerned has already been introduced. Attempting to match earlier sub-tree governors retrieves TU 90, headed by the word *flow*, and leads to the (interactive) recognition that the noun phrase refers back to a whole clause and can be treated as a simple case of event reference.

4.5 Summary of principles

- A translation unit consists of a pair of sub-trees labelled with, and thus linked by, the same ID.
- A TU is headed by a content word or morpheme, not by a syntactic label.
- Translation units are sought by scanning each tree bottom-up and left-to-right, trying to match sub-trees in one language with existing TUs and identify the equivalent in the other language within the current tree.
- Tree matching itself proceeds top-down.
- TU identification requires confirmation from the operator, at least in the early stages of

BKB construction.

– When no existing unit can be matched, the constructor offers a guess based on structural similarity or subtraction of sub-trees.

– Reference forms are semi-automatically identified with their referent and augmented with its original ID.

– Compaction is achieved by replacing repeated surface structures with their IDs.

It will already be obvious that the construction of the BKB makes heavy demands on the operator's help in the early stages, but that gradually the growing BKB itself makes the processing of new material a semi-automatic process.

4.6 Dreaming

The most efficient organization of the BKB may not always be the one created during construction. For example, the first text processed might include several occurrences of the concept 'shutoff valve' with the appropriate Esperanto translation, without the concept 'valve' or 'shutoff' appearing independently. The translation unit will be coded as a whole, and will not be recognized as consisting of separable concepts. If now a later text does introduce 'valve' and 'shutoff' independently, a BKB search will not retrieve their occurrences in the 'shutoff valve' combination, because the independent TU codes had not been registered in the compound term.

This situation can be rectified by an automatic process of recoding the BKB, where complex concepts can be reduced to their component TUs with the aid of more recent knowledge. Recoding could normally be carried out in off-peak hours, in a process reminiscent of human dreaming (which may well perform a similar function).

5 TRANSLATING WITH THE BKB

What are the practical consequences of the BKB concept for actual translation? These can be considered under three headings: metataxis, text coherence and disambiguation.

5.1 Metataxis

Metataxis, or structural transformation, can be guided by rules which are implicit in the whole BKB structure. The subtraction of translation units from each other is a powerful device equivalent to complex lexical metataxis rules containing variables, such as those included in the DLT prototype dictionaries. Generating translations with the BKB is a jigsaw-like process in which translation units associated with the words or morphemes of the input sentence are retrieved from the BKB and put together in such ways as will reproduce one or more of the possible source language structures and at the same time produce an internally consistent structure in the target language. Of course, the problem of selecting among alternative translations remains. (See 5.3 below.)

N.B.: Given the English-Dutch TUs

[14] 1. John has kicked the bucket. = John is doodgegaan.
2. bucket of milk = emmer melk
3. at last = eindelijk

the sentence

[15] John has kicked the bucket of milk.

will not be automatically divided into TUs 1 and 2, because the attachment point 'bucket' is not accessible in TU 1, i.e. it does not itself head a TU. On the other hand,

[16] John has kicked the bucket at last.

can be composed of TUs 1 and 3, because 'at last' is governed by the head word of TU 1.

5.2 Text coherence

The "backbone" of text coherence consists in **reference and deixis** (Papegaaij & Schubert, 1988: 199). As already shown for the sample text in the appendix, the BKB structure provides for, and even demands, the systematic identification of the items and events mentioned in the text. Knowledge of a particular entity can be accumulated over a number of references.

The sentence

[17] The Mayor has resigned.

may occur more than once in a corpus, but the separate occurrences may or may not be cross-linked, even if the words are identical in both languages, depending on whether they refer to the same mayor or not. This is necessary, for one thing, because the specific knowledge available about the mayor in question can determine the surface form of future references: for example, whether the appropriate pronoun is *he* or *she*. Translation is geared to concept IDs, leaving the way open for TL-specific generation of references. The fact that the SL (source language) originally used a pronoun, for example, in no way constrains the TL reference. This may also take the form of a pronoun, but it may equally well be a repetition of the original form, or a generic term or synonym used earlier for the same entity. The appropriateness and possible ambiguity of the chosen form can be reliably checked within the TL half of the text representation, because this includes not only bilingual concepts, but also concepts introduced in the TL text only, and which might well lead to a misidentification of the entity referred to. A notable consequence of all this is that pronouns, pro-verbs etc. are never translated, but must be generated, where required, by the TL part of the metataxis process.

Consider the following example:

[18] *English:*
Snow was falling.
It had been doing so for hours.

Esperanto:
Neĝis. ('It-was-snowing.')
Neĝis jam dum horoj. ('It-was-snowing already for hours.')

Here, the concept of falling, which is present in the English text, is only implicit in the Esperanto. (The verb *neĝi*, 'to snow', implies, of course, that snow is falling.) The same is true of the nominal concept 'snow', which again is only implied in the Esperanto verb. So where the second English sentence uses a pronoun *it* to refer back to 'snow' and a pro-verb (*doing so*) to refer back to the action of falling, the Esperanto version has no pro-forms at all. It simply repeats the verb *neĝis*. Of course, every word in each monolingual text must be included in one or other translation unit, since by our definition the bi-text consists of translation units and nothing else. Thus the monolingual concepts expressed by *snow* and *falling* are included in the TU (bilingual concept)

[fall [SUBJ snow]] = [neĝi].

Instantiating the English pronoun and pro-verb (in the second sentence of [18]) with their referents automatically ensures that they will be translated by a repetition of *neĝis*, since the English sub-tree now matches the left-hand side of the TU shown above (and since Esperanto has no pro-verbs). If the BKB also contains a literal translation in the form of the TU

[fall [SUBJ snow]] = [fali [SUBJ neĝo]]

then of course we can also obtain the alternative translation

[19] Neĝo falis. ('Snow was-falling.')
Ĝi falis jam de horoj. ('It was-falling already for hours.')

in which the use of the pronoun *ĝi* echos the English reference.

What this example shows is that the monolingual identification of concepts plays a vital role in the resolution of reference, which cannot be achieved on the basis of translation units alone.

Apart from handling reference and deixis, the BKB can preserve the SL order of syntagmata: word order information, though omitted in the examples in the appendix, should also have its place in the structure. This can serve the purposes of the **theme-rheme** distinction. For example, there is different theme/rheme information in the Esperanto sentences

[20] La prezidanto malfermis la kongreson.
La kongreson malfermis la prezidanto.
('The president opened the congress.')

although syntactically and lexically they are identical. Moreover, the identification of the unmarked order is simply a matter of performing a frequency count across the BKB for the structure concerned, for the SL or the TL independently.

In principle, the BKB could be adapted to the requirements of **text level** analysis (discourse structure). Each sentence or clause is labelled with an ID, and rhetorical relations between sentences could easily be inserted, if these could be identified. Recent work on automated abstracting (Thürmer, forthc.) has focussed on identifying the essential parts of the text by first finding the most frequent nouns and then checking each sentence for the frequency of their appearance. Sentences showing a high frequency for the most frequent nouns tend to be the most central to the text. This method could be refined on the basis of the BKB reference structure by checking the occurrence of the most frequent **concepts** rather than mere nouns (which may refer to different concepts or entities).

Finally, the BKB structure has already shown its value for text coherence as a vital element in natural language understanding in general, and in **controlled language** in particular. The analysis of the sample text in Appendix A highlights two points where an interactive system could have aided the writer of Simplified English to avoid the possibility of misunderstanding:

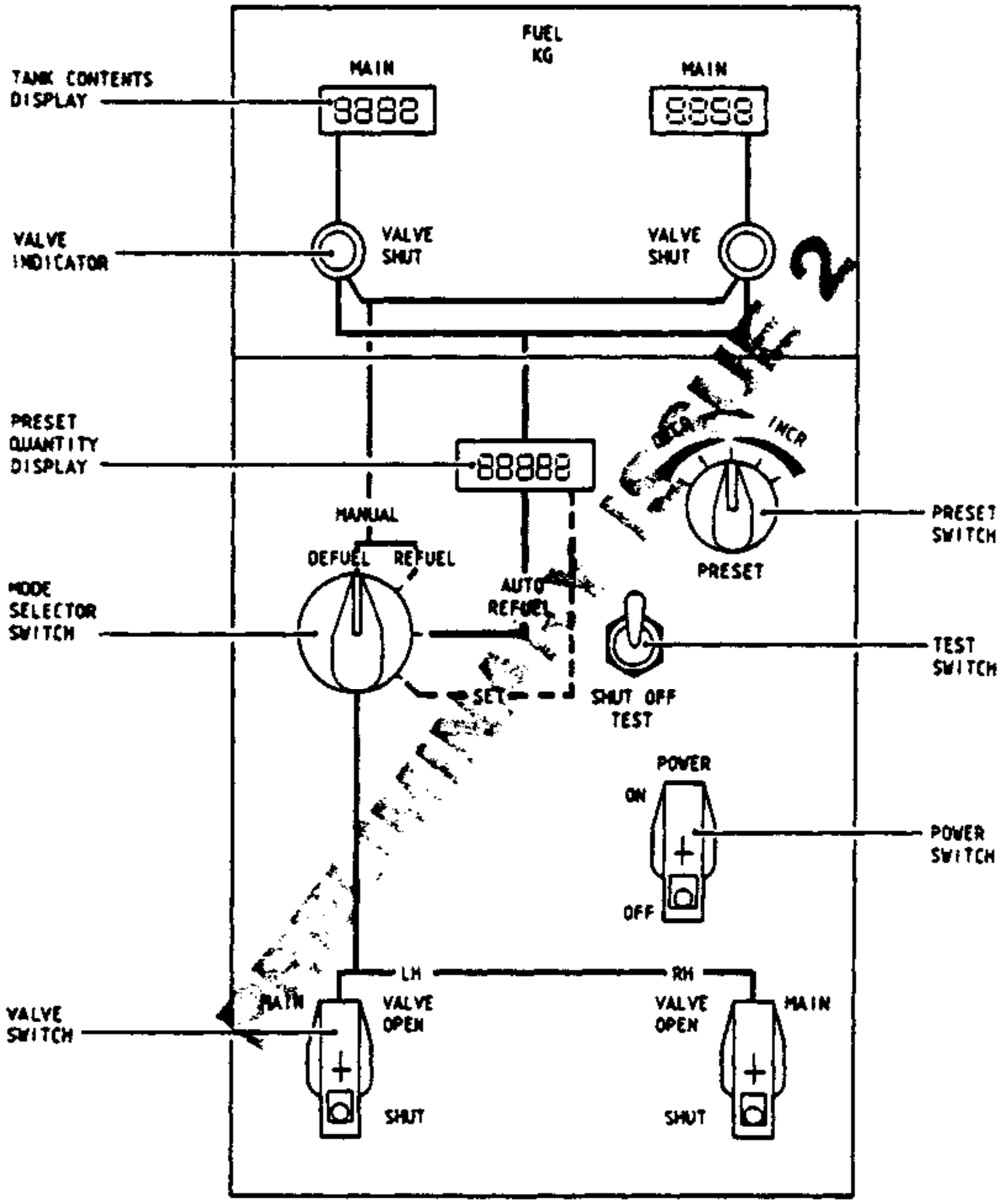
(1) Instruction 5(a) tells the reader to make sure that *the light for the right-hand shutoff valve* comes on. Checking our information on this entity during the first attempt at encoding, however, we found that the shutoff valve lights (of which this is one) were already on! The system should be able to recognize here that there is a contradiction in the instructions. The explanation is that the following expressions apparently refer to the same entity:

    the light for the shutoff switch of the right-hand outer wing tank
    the light for the right-hand shutoff valve

    The former of these went off according to instruction 4(a), so if they refer to the same entity the contradiction is removed. In the context of an aircraft maintenance manual this kind of confusion should be corrected.

(2) Instruction 5 tells the reader to hold *the switch on the fueling control panel* to TEST. Here, the definite reference suggests that this switch has been already introduced. The only likely candidate referent is the power switch in instruction 1, which by fairly simple inference can be understood to be located on the fueling control panel. Reference to a diagram (Fig. 3), however, shows that this is not the correct referent, because the power switch has no TEST position. Actually, the original text, before "translation" to

Figure 3



FKFSMM121102301AAM

Fueling Control Panel
Fig. 301

Printed in France

Simplified English, referred in instruction 5 to the test switch on the fueling control panel. Here again, a routine query from the DLT system as to the intended referent could have prevented the omission of this useful epithet.

This example, incidentally, underlines the importance of integrating diagram legend with the knowledge base. It should be quite feasible to devise a program to enter this information interactively, thus adding the knowledge of what switches are to be found on the control panel, what their possible settings are, etc.

5.3 Disambiguation

The BKB concept offers significant advantages for disambiguation, both automatic and interactive (via a dialogue with the user).

5.3.1 Automatic:

The main advantages are the following:

(1) There is no need to perform consistency checks between the dictionaries and the lexical knowledge base (LKB) used by the semantic module, since any lexical item included in the dictionary is *ipso facto* represented in the LKB.

(2) The semantic module can handle **multi-word units**. Welding together two parallel versions of the same text provides an operational criterion for the definition of multi-word concepts. The BKB can be compared to a network of molecules consisting of translation units, and the choice between alternative translations (i.e. between alternative TUs) will be determined by contextual probabilities. A major difference is that contextual patterns no longer consist merely of word-to-word relations, but of TU-to-TU relations. The context of a given unit consists of other units, which may be simple words but may also be complex sub-trees.

(3) The problem of identifying and tagging individual **word meanings** in the LKB (e.g. Esperanto *akso* as 'axis' or 'axle'), is essentially solved in the BKB, which equates meanings or "concepts" with bilingual equivalences. Since all contextual information is now tied to bilingual equivalences, the contextual patterns of *akso* as 'axis' and as 'axle' are clearly separated, in an English-Esperanto BKB, by the distinctive translations to which they are attached. If, on the other hand, they both happened to have the same translation in the other language, then the distinction might be considered irrelevant for translation purposes.

(4) The semantic module can access **content** morphemes in the BKB, which uses word grammar to structure polymorphemic words. In most cases this amounts to making generic terms available, because, for example, a *fuelujo* ('fuel container') is a type of *ujo* ('container'). The principle of representing words as morpheme trees in the BKB means that the semantic module can apply morpheme-level knowledge, immediately recognizing the plausibility of *skribi per skribilo* ('to write with a writing-instrument') or the contradiction in *senakva akvo* ('waterless water').

(5) The BKB structure allows the semantic module to match the **total input pattern** against the selected total patterns in the BKB. There is no dismemberment of the original structures in the knowledge sources (corpora) which constitute the BKB.

(6) The semantic module is supported by **text coherence** mechanisms. These can greatly reduce the number of alternative translations under consideration. Definite references will normally exclude translations which cannot apply to the entity referred to. A comparable aid is the recency factor. When comparing the input expression with alternative TUs in the BKB, the module can also take their recency into account. If the text representation

shows that the word *power* has repeatedly been used in the recent text in the electrical sense, then the other senses can already be considered less likely on that score alone. Text coherence can also reduce structural ambiguity, in that repeated word patterns are likely to retain the same structure. This can be used to establish a preference ranking between alternative structural interpretations.

(7) The use of a bilingual knowledge bank means that it is possible to compute the semantic proximity of any given pair of concepts in either language (Sadler, forthc.). If, for example, the system has to choose a French translation for the Esperanto *glata* ('smooth') but, after rejecting various other alternatives, is unable at first sight to choose between *lisse* and *poli*, it is possible to compute the extent to which these two words are synonymous in French, so that a decision can be taken as to how important it is to make an informed choice between them and whether it is worthwhile making further efforts to establish a preference.

(8) There must always be a default preference for the case where the semantic module is unable to make a reasoned choice between alternative translations. In the DLT prototype, lexical alternatives were assigned a default preference order in the bilingual dictionary, which proved a difficult and time-consuming task for the lexicographers. In the BKB, much more objective criteria are available: frequency and recency. Default preference can be based on the relative frequencies of the alternative TUs in the BKB, weighted for recency. This now becomes a dynamic index, influenced by every new addition to the BKB, including the text being translated.

5.3.2 The dialogue:

In the 1988 DLT prototype, a computer-initiated dialogue was used to allow the user to confirm or override the interpretations selected by the semantic module. To support this dialogue, the bilingual English-Esperanto dictionary was equipped with English paraphrases of all the alternative Esperanto translations. Entering these paraphrases proved to be one of the most time-consuming tasks of the lexicographers and one of the least satisfactory. It often proved virtually impossible to paraphrase the meaning of a given word in a way that is reasonably concise and at the same time sufficiently distinctive when compared with the paraphrases of alternative translations.

In the BKB conception, based as it is on corpus analysis, there is no place for arbitrary paraphrases. So what are the alternatives? Somehow, lexical ambiguities have to be presented to the operator in a clear manner.

The solution proposed is to replace paraphrases with examples. Every time a translation is selected, the semantic module can be assumed to have found a translation unit in the BKB which matches best the current context. Since the TU thus pinpointed is also embedded in a broader BKB context, this context can be used to provide the example.

The examples in the following illustrations are taken from a corpus. Given the input sentence

[21] What is the subject of the question?

the system could offer:

```
Interpretations as in:
  [1] the SUBJECT of the verb
  [2] aspects of the QUESTION of aging
```

If the operator disagrees with any of the interpretations offered, the mouse can be used to click up an alternative. In this case, clicking on both [1] and [2] might produce a revised display like

Interpretations as in:
[1] the SUBJECT of very detailed study
[2] some of the QUESTIONs raised

If none of the interpretations offered satisfies the operator (e.g. s/he has clicked through the whole cycle of possible translations for the word SUBJECT) the system could, if requested, search the BKB for other examples of each interpretation which may prove more acceptable.

An obvious question which now arises is: How does the system know where to set the limits of each example? The examples should not consist of whole sentences, which would take too long to read, but on the other hand they should contain enough context to make the meaning plain. The answer can probably be found in the pattern matching process in the semantic module itself. When this module makes a (provisional) choice, it does so on the basis of a match score consisting of various components (Sadler, forthc.). Each contextual relation of the word in question contributes something to the total word score. This means that it should always be possible, not only to select a particular translation unit, but also to select those parts of its context which had the greatest influence on its selection. For example, if the TU selected for *subject* is embedded in the sentence

[22] We must first identify the subject of the verb.

it is likely that the most influential contextual relation is the prepositional phrase *of the verb*, and this is also the relation which makes the meaning of *subject* most explicit.

This approach to the disambiguation dialogue has two important advantages. First and foremost, it requires no effort whatever on the part of the lexicographer. The almost impossible task of thinking up suitable synonyms or paraphrases is eliminated altogether. Second, the method can easily cope with pseudo-structural ambiguities such as word-class ambiguity, which in the DLT prototype had to be presented in such unsatisfactory terms as

"second" must be interpreted as adjective/noun
"as" must have label E-PREC/E-CIRC.

6 SUMMARY OF ADVANTAGES OF A BILINGUAL KNOWLEDGE BANK

The advantages this new conception offers can be summarized as follows:

(1) Linguistic and extra-linguistic knowledge can be stored in retrievable form with relatively little human effort. The BKB is strongly oriented towards **machine learning from textual input**. The system is self-improving, because its application for machine translation automatically produces new bilingual structures which can be used to further enrich the knowledge bank. Complex rules of syntactic transformation, such as are frequently required in translation, can be kept implicit in the BKB but can nevertheless be automatically accessed and applied by the machine translation system. They do not need to be formulated explicitly.

(2) The **translation** expertise needed in a machine translation system can be acquired by "digesting" the work of highly qualified human translators. A computer system can translate by imitating the performance of the human translator, without first requiring the expert to explain and formalise the rules he intuitively applies. It is no longer necessary to rely on such often inadequate sources as conventional dictionaries and grammars.

(3) The BKB is a **dynamic system**, because new material can be added (and old material discarded) in such a way that changes in usage, new terminology etc. can be reflected in

the output of the translation system. Provided up-to-date human translations are available, it is not necessary to wait for these changes or new terms to be first recorded by linguists or terminologists, a process which often takes years (Shaikevich & Oubine, 1988: 10).

(4) The BKB is a symmetrical construction, in which no distinction is made between source language and target language. It is immaterial which of the texts was the source text, or whether both are translations from some original in a third language. Consequently, all the information in the BKB can be used in either direction. The BKB thus comprises a dictionary and rule system which is **100% reversible**.

A consequence of this general conception is that the BKB, consisting as it does of translation units, is necessarily language pair-specific. This is not to say, however, that the knowledge it contains need be different, in a broad sense, from language pair to language pair. Both general knowledge and domain-specific knowledge can be built up for each language pair on the basis of a comparable corpus, provided translations are available in the languages concerned. This consideration strongly favours the development of a multilingual corpus.

LITERATURE

AECMA (1984): Writing Rules for AECMA Simplified English. Association of European Aerospace Manufacturers.

Bennett, W.S. & Slocum, J. (1985): The LRC Machine Translation System. Computational Linguistics, 11, No. 2-3.

Boitet, Ch. (1987): Current state and future outlook of the research at GETA.
In: MT Summit, manuscripts and program. Hakone: Machine Translation Summit, pp. 26-35.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roossin, P. (1988): A statistical approach to language translation.
In: 12th International Conference on Computational Linguistics. Proceedings of Coling '88. Budapest: John von Neumann Society for Computing Sciences, pp. 71-76.

Byrd, R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S., Rizk, O.A. (1987): Tools and Methods for Computational Lexicology. Yorktown Heights: T.J.Watson Research Center. IBM Research Report RC 12642.

Calzolari, N. (1988): The Pisa Lexical Database: perspectives and new developments. Mitt. Autom. Sprachverarb. Sept. 1988, p.5-7.

CMT [Center for Machine Translation] (1988): Carnegie Mellon University: Site Reports. The Finite String, 14, 2, p.2.

Ernst, R. (1984): Comprehensive dictionary of engineering and technology. Dictionnaire général de la technique industrielle. Wiesbaden: Brandstetter.

Gross, A. (1989): A New Addition To The Translator's Toolbox. Language Technology No.12, p. 42-45.

Harris, Brian (1988a): Bi-text, a new concept in translation theory. Language Monthly, 54, p.8-10.

Harris, Brian (1988b): Are you bitextual? Language Technology May/June 1988, 7, p.41.

Harris, Brian (1988c): Interlinear bitext. Language Technology Nov/Dec 1988, 10, p.12.

Hutchins, W.J. (1986): Machine Translation: Past, Present, Future. Chichester: Horwood.

Hutchins, W.J. (1988): Recent Developments in Machine Translation: A Review of the Last Five Years. In: New Directions in Machine Translation. Maxwell, D., Schubert, K. & Witkam, T. (eds.). Dordrecht/Providence: Foris, pp. 7-64.

Munniksma, F. (1975): International Business Dictionary in nine languages. Deventer-Antwerp: Kluwer.

Nagao, M. (1984): A framework of a mechanical translation between Japanese and English by analogy principle.
In: Artificial and human intelligence. A.Elithorn & R. Banerji (eds.). Elsevier. Pp. 173-180.

Papegaaij, B.C. & Schubert, K. (1988): Text coherence in translation. Dordrecht/Providence: Foris. Distributed Language Translation 3.

Piron, C. (1988): Learning from Translation Mistakes.
In: New Directions in Machine Translation. Maxwell, D., Schubert, K. & Witkam, T. (eds.). Dordrecht/Providence: Foris. Pp. 233-242.

Sadler, V. (forthcoming): Working with analogical semantics: An assessment of current disambiguation techniques in DLT. Dordrecht/Providence: Foris. Distributed Language Translation 6.

Schubert, K. (1986): Linguistic and extra-linguistic knowledge. Computers and Translation, Vol. 1, No. 3. Osprey (Florida): Paradigm Press.

Schubert, K. (1987): Metataxis. Contrastive dependency syntax for machine translation. Dordrecht/Providence: Foris. Distributed Language Translation 2.

Shaikevich, A. & Oubine, I. (1988): Translators and researchers look at bilingual terminological dictionaries. Babel 34, 1, pp. 10-16.

Sumita, E. & Tsutsumi, Y. (1988): A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching.
In: Proceedings Supplement, Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Pittsburgh: Carnegie Mellon University Center for Machine Translation.

Thürmer, Robert & Uta (forthc.): THEA – ein Programmsystem zur Analyse, Bewertung und Adaptation von Texten.
In: Proceedings of the XIV International Congress of Linguists, Berlin, 1987.

Tsujii, J. (1986): Future directions of machine translation.
In: 11th International Conference on Computational Linguistics. Proceedings of Coling '86. Bonn: Institut fuer angewandte Kommunikations- und Sprachforschung, pp. 655-668.

Tsujii, J. (1988): What Is a Cross-Linguistically Valid Interpretation of Discourse?
In: New Directions in Machine Translation. Maxwell, D., Schubert, K. & Witkam, T. (eds.). Dordrecht/Providence: Foris. Pp. 157-166.

Wilks, Y. (1972): Grammar, Meaning and the Machine Analysis of Language. London: Routledge.

Appendix A: Sample English text with Esperanto translation

English: Outer Wing Tank Test

(1) On the fueling control panel, set the power switch to ON.

  (a) Make sure that:
   - the power light is off;
   - the overflow valve lights are off;
   - the shutoff valve lights are on.

(2) Apply pressure to the refueling system.

  (a) Make sure that:
   - the lights for the overflow valves of the outer wing tanks come on;
   - the shutoff valve lights stay on;
   - fuel does not flow into the tanks.

(3) Make sure there is no leakage from the refueling lines between the right-hand tank and the left-hand tank.

(4) Set the shutoff switch of the right-hand outer wing tank to OPEN.

  (a) Make sure that:
   - the light for the shutoff switch of the right-hand outer wing tank goes off;
   - fuel flows into the right-hand tank.

(5) Hold the switch on the fueling control panel to TEST.

  (a) Make sure that:
   - the light for the right-hand shutoff valve comes on;
   - the fuel flow stops.

Esperanto: Testo de la eksteraj alfuelujoj

(1) Sur la komandpanelo por fuelizado, movu la alimentsxaltilon al "ON".

  (a) Kontrolu, ke:
   - la signallampo de la alimento ne lumas;
   - la signallampoj de la superversxaj valvoj ne lumas;
   - la signallampoj de la baraj valvoj lumas.

(2) Apliku premon al la sistemo de refuelizado.

  (a) Kontrolu, ke:
   - la signallampoj de la superversxaj valvoj de la eksteraj alfuelujoj eklumas;
   - la signallampoj de la baraj valvoj lumadas;
   - fuelo ne fluas en la fuelujojn.

(3) Kontrolu, ke ne likas la refuelizaj tuboj inter la dekstra fuelujo kaj la maldekstra fuelujo.

(4) Movu la barsxaltilon de la dekstra ekstera alfuelujo al "OPEN".

  (a) Kontrolu, ke:
   - la signallampo de la barsxaltilo de la dekstra ekstera alfuelujo cxesas lumi;
   - fuelo fluas en la dekstran fuelujon.

(5) Tenu la sxaltilon sur la komandpanelo por fuelizado cxe "TEST".

  (a) Kontrolu, ke:
   - la signallampo de la dekstra barvalvo eklumas;
   - la fuelfluo cxesas.

Appendix B: Coding of translation units between the syntactic structures in English and Esperanto

```
[GOV 1,test                             [GOV 1,testo
 [ATR 2,3,tank                           [ATR 2,de
  [ATR 3.1,wing ]                         [PARG 2/1,j,3,((3.1,al)(fuel)ujo) [la]
  [ATR 4,outer ]]]                         [ATR 4,ekstera ]]]]

["(1)"                                  ["(1)"
 [GOV 5,set                              [GOV 5-u,movi
  [ADVA 6,on                              [ADVA 6,sur
   [PARG 7,panel [the]                     [PARG 7,((komand)panelo) [la]
    [ATR 7/1,control
     [ATR 8,fueling ]]]]                    [ATR 8,por
                                             [PARG 8/1,((fuel)izado) ]]]]
  [OBJ 10,switch [the]                    [OBJ 10,(((10.1,aliment)sxalt)ilo) [la] ]
   [ATR 10.1,power ]]
  [ADVC 11,to                             [ADVC 11,al
   [PARG 12,"ON" ]]]]                      [PARG 12,"ON" ]]]]

["(1)(a)"                               ["(1)(a)"
 [GOV 13,make                            [GOV 13-u,kontroli
  [PRED 13/1,sure ]
  [OBJ 14,that                            [OBJ 14,ke
   [SUBC 15,"; -"                          [SUBC 15,"; -"
    [SUBC-C 16,"; -"                        [SUBC-C 16,"; -"
     [SUBC-C 17,be                           [SUBC-C 17-as,lumi
      [PRED 17/1,off ]                        [ADVA 17/1,ne ]
      [SUBJ 18,light [the]                    [SUBJ 18,((signal)lampo) [la]
       [ATR 19,20,power ]]]                    [ATR 19,de
                                                [PARG 20,alimento [la] ]]]]
     [SUBC-C 21,be                           [SUBC-C 21-as,lumi
      [PRED 21/1,off ]                        [ADVA, 21/1,ne]
      [SUBJ 22,s,22.1,light [the]             [SUBJ 22,j,22.1,((signal)lampo) [la]
       [ATR 23,24,valve                        [ATR 23,de
                                                [PARG 23/1,j,24,valvo [la]
        [ATR 25,overflow ]]]]                    [ATR 25,((super)versxa) ]]]]]]
    [SUBC-C 26,be                           [SUBC-C 26-as,lumi
     [PRED 26/1,on ]
     [SUBJ 27,s,27.1,light [the]             [SUBJ 27,j,27.1,((signal)lampo) [la]
      [ATR 28,29,valve                        [ATR 28,de
                                               [PARG 28/1,j,29,valvo [la]
       [ATR 30,shutoff ]]]]]]]]                 [ATR 30,bara ]]]]]]]]]

["(2)"                                  ["(2)"
 [GOV 31,apply                           [GOV 31-u,apliki
  [OBJ 32,pressure ]                      [OBJ 32,premo ]
  [ADVC 33,to                             [ADVC 33,al
   [PARG 34,system [the]                   [PARG 34,sistemo [la]
    [ATR 35,refueling ]]]]]                 [ATR 35,de
                                             [PARG 35/1,((re)(fuel)izado) ]]]]]]

["(2)(a)"                               ["(2)(a)"
```

```
[GOV 37,make                                    [GOV 37-u,kontroli
 [PRED 37/1,sure ]
 [OBJ 38,that                                    [OBJ 38,ke
  [SUBC 39,"; -"                                  [SUBC 39,"; -"
   [SUBC-C 40,"; -"                                [SUBC-C 40,"; -"
    [SUBC-C 41,come                                 [SUBC-C 41-as,((ek)lumi)
     [PRED 41/1,on ]
     [SUBJ 42,s,42.1,light [the]                     [SUBJ 42,j,42.1,((signal)lampo) [la]
      [ATR 43,for                                     [ATR 43,de
       [PARG 44,s,44.1,valve [the]                     [PARG 44,j,44.1,valvo [la]
        [ATR 45,overflow ]                              [ATR 45,((super)versxa) ]
        [ATR 46,of                                      [ATR 46,de
         [PARG 47,s,47.1,tank [the]                      [PARG 47,j,47.1,((47.2,al)(fuel)ujo) [la]
          [ATR 47.1,wing ]
          [ATR 48,outer ]]]]]]]                           [ATR 48,ekstera ]]]]]]]
    [SUBC-C 49,stay                                 [SUBC-C 49-as,(lum)adi
     [PRED 49/1,on ]
     [SUBJ 50,s,50.1,light [the]                     [SUBJ 50,j,50.1,((signal)lampo) [la]
      [ATR 51,52,valve                                [ATR 51,de
                                                       [PARG 51/1,j,52,valvo [la]
       [ATR 53,shutoff ]]]]                             [ATR 53,bara ]]]]]
   [SUBC-C 54,flow                                 [SUBC-C 54-as,flui
    [ADVA, 54.1,not]                                [ADVA, 54.1,ne]
    [SUBJ 55,fuel ]                                 [SUBJ 55,fuelo ]
    [ADVC 56,into                                   [ADVC 56,alen
     [PARG 57,s,57.1,tank [the] ]]]]]]]              [PARG 57,j,57.1,((fuel)ujo) [la] ]]]]]]]

["(3)"                                          ["(3)"
 [GOV 58,make                                    [GOV 58-u,kontroli
  [PRED 58/1,sure ]
  [OBJ 59,60,is                                   [OBJ 59,ke
   [ADVC 60/1,there ]                              [SUBC 60-as,liki
   [SUBJ 60/2,leakage                               [ADVA 60/1,ne]
    [ATR, 60/3,no]
    [ATR 60/4,from
     [PARG 61,s,61.1,line [the]                      [SUBJ 61,j,61.1,tubo [la]
      [ATR 62,refueling ]                             [ATR 62,((re)(fuel)iza) ]
      [ATR 63,between                                 [ATR 63,inter
       [PARG 64,and                                    [PARG 64,kaj
        [PARG-C 65,tank [the]                           [PARG-C 65,((fuel)ujo) [la]
         [ATR 66,right-hand ]]                           [ATR 66,dekstra ]]
        [PARG-C 67,tank [the]                           [PARG-C 67,((fuel)ujo) [la]
         [ATR 68,left-hand ]]]]]]]]]]                    [ATR 68,((mal)dekstra) ]]]]]]]]]

["(4)"                                          ["(4)"
 [GOV 69,set                                     [GOV 69-u,movi
  [OBJ 70,switch [the]                            [OBJ 70,(((70.1,bar)sxalt)ilo) [la]
   [ATR 70.1,shutoff ]
   [ATR 71,of                                      [ATR 71,de
    [PARG 72,tank [the]                             [PARG 72,((72.1,al)(fuel)ujo) [la]
     [ATR 72.1,wing ]
     [ATR 73,outer ]                                 [ATR 73,ekstera ]
     [ATR 74,right-hand ]]]]                         [ATR 74,dekstra ]]]]
  [ADVC 75,to                                     [ADVC 75,al
  [PARG 76,"OPEN" ]]]]                  [PARG 76,"OPEN" ]]]]

["(4)(a)"                              ["(4)(a)"
 [GOV 78,make                           [GOV 78-u,kontroli
  [PRED 78/1,sure ]
  [OBJ 79,that                            [OBJ 79,ke
   [SUBC 80,"; -"                          [SUBC 80,"; -"
    [SUBC-C 81,go                           [SUBC-C 81-as,cxesi
     [PRED 81/1,off ]                        [INFC 81/1,lumi ]
     [SUBJ 83,light [the]                    [SUBJ 83,((signal)lampo) [la]
      [ATR 84,for                             [ATR 84,de
       [PARG 85,switch [the]                   [PARG 85,(((85.1,bar)sxalt)ilo) [la]
        [ATR 85.1,shutoff ]
        [ATR 86,of                               [ATR 86,de
         [PARG 87,tank [the]                      [PARG 87,((87.1,al)(fuel)ujo) [la]
          [ATR 87.1,wing ]
          [ATR 88,outer ]                           [ATR 88,ekstera ]
          [ATR 89,right-hand ]]]]]]]                [ATR 89,dekstra ]]]]]]]
    [SUBC-C 90,flow                         [SUBC-C 90-as,flui
     [SUBJ 91,fuel ]                         [SUBJ 91,fuelo ]
     [ADVC 92,into                           [ADVC 92,alen
      [PARG 93,tank [the]                     [PARG 93,((fuel)ujo) [la]
       [ATR 94,right-hand ]]]]]]]]             [ATR 94,dekstra ]]]]]]]]

["(5)"                                 ["(5)"
 [GOV 95,hold                           [GOV 95-u,teni
  [OBJ 96,switch [the]                   [OBJ 96,((sxalt)ilo) [la]
   [ATR 97,on                             [ATR 97,sur
    [PARG 98,panel [the]                   [PARG 98,((komand)panelo) [la]
     [ATR 98/1,control
      [ATR 99,fueling ]]]]]                 [ATR 99,por
                                             [PARG 99/1,((fuel)izado) ]]]]]
  [ADVC 101,to                           [ADVC 101,cxe
   [PARG 102,"TEST" ]]]]                  [PARG 102,"TEST" ]]]]

["(5)(a)"                              ["(5)(a)"
 [GOV 103,make                          [GOV 103-u,kontroli
  [PRED 103/1,sure ]
  [OBJ 104,that                           [OBJ 104,ke
   [SUBC 105,"; -"                         [SUBC 105,"; -"
    [SUBC-C 106,come                        [SUBC-C 106-as,((ek)lumi)
     [PRED 106/1,on ]
     [SUBJ 107,light [the]                     [SUBJ 107,((signal)lampo) [la]
      [ATR 108,for                              [ATR 108,de
       [PARG 109,valve [the]                     [PARG 109,((109.1,bar)valvo) [la]
        [ATR 109.1,shutoff ]
        [ATR 110,right-hand ]]]]]                  [ATR 110,dekstra ]]]]]
    [SUBC-C 111,stop                          [SUBC-C 111-as,cxesi
     [SUBJ 112,flow [the]                      [SUBJ 112,((112.1,fuel)fluo) [la] ]]]]]]
      [ATR 112.1,fuel ]]]]]]]
```

Appendix C: Knowledge bank structure for the same text in English and Esperanto, with coding of translation units and reference

```
 [GOV 1,test                                [GOV 1,testo
  [ATR 2,3,tank                              [ATR 2,de
   [ATR 3.1,wing ]                            [PARG 2/1,j,3,((3.1,al)(fuel)ujo) [la]
   [ATR 4,outer ]]]                            [ATR 4,ekstera ]]]]

["(1)"                                     ["(1)"
 [GOV 5,set                                 [GOV 5-u,movi
  [ADVA 6,on                                 [ADVA 6,sur
   [PARG 7,panel [the]                        [PARG 7,((komand)panelo) [la]
    [ATR 7/1,control
     [ATR 8,fueling ]]]]                        [ATR 8,por
                                                 [PARG 8/1,((fuel)izado) ]]]]
  [OBJ 10,switch [the]                       [OBJ 10,(((10.1,aliment)sxalt)ilo) [la] ]
   [ATR 10.1,power ]]
  [ADVC 11,to                                [ADVC 11,al
   [PARG 12,"ON" ]]]]                         [PARG 12,"ON" ]]]]

["(1)(a)"                                  ["(1)(a)"
 [GOV 13,make                               [GOV 13-u,kontroli
  [PRED 13/1,sure ]
  [OBJ 14,that                               [OBJ 14,ke
   [SUBC 15,"; -"                             [SUBC 15,"; -"
    [SUBC-C 16,"; -"                           [SUBC-C 16,"; -"
     [SUBC-C 17,be                              [SUBC-C 17-as,lumi
      [PRED 17/1,off ]                           [ADVA 17/1,ne ]
      [SUBJ 18,light [the]                       [SUBJ 18,((signal)lampo) [la]
       [ATR 19,20{=10.1},power ]]]                [ATR 19,de
                                                   [PARG 20{=10.1},alimento [la] ]]]]
     [SUBC-C 21:17-18                           [SUBC-C 21:17-18
      [22,s,22.1:18-19                           [22,j,22.1:18-19
       [23,24,valve                               [23,de
                                                   [PARG 23/1,j,24,valvo [la]
        [ATR 25,overflow ]]]]                       [ATR 25,((super)versxa) ]]]]]
    [SUBC-C 26,be                              [SUBC-C 26-as,lumi
     [PRED 26/1,on ]
     [SUBJ 27:22-25                             [SUBJ 27:22-25
       [30,shutoff ]]]]]]]]                        [30,bara ]]]]]]]]

["(2)"                                     ["(2)"
 [GOV 31,apply                              [GOV 31-u,apliki
  [OBJ 32,pressure ]                         [OBJ 32,premo ]
  [ADVC 33,to                                [ADVC 33,al
   [PARG 34,system [the]                      [PARG 34,sistemo [la]
    [ATR 35,refueling ]]]]]                    [ATR 35,de
                                                [PARG 35/1,((re)(fuel)izado) ]]]]]]

["(2)(a)"                                  ["(2)(a)"
 [GOV 37:13-17-21-26                        [GOV 37:13-17-21-26
    [41,come                                   [41-as,((ek)lumi)
```

```
      [PRED 41/1,on ]
      [SUBJ 42{<22}:22-23                                 [SUBJ 42{<22}:22-23
       [43,for                                             [43,de
        [PARG 44,s,44.1:24                                  [PARG 44{<23/1},j,44.1:24
         [ATR 46,of                                          [ATR 46,de
          [PARG 47,s,47.1:3 ]]]]]]                            [PARG 47,j,47.1:3 ]]]]]]
     [49,stay                                            [49-as,(lum)adi
      [PRED 49/1,on ]
      [SUBJ 50{=27}:27 ]]                                 [SUBJ 50{=27}:27 ]]
     [54,flow                                            [54-as,flui
      [ADVA, 54.1,not]                                    [ADVA, 54.1,ne]
      [SUBJ 55,fuel ]                                     [SUBJ 55,fuelo ]
      [ADVC 56,into                                       [ADVC 56,alen
       [PARG 57{=47},s,57.1,tank [the] ]]]]]               [PARG 57{=47},j,57.1,((fuel)ujo) [la] ]]]]]

["(3)"                                               ["(3)"
 [GOV 58:13-14                                        [GOV 58:13-14
  [59,60,is                                            [59,ke
   [ADVC 60/1,there ]                                   [SUBC 60-as,liki
   [SUBJ 60/2,leakage                                    [ADVA 60/1,ne]
    [ATR, 60/3,no]
    [ATR 60/4,from
     [PARG 61,s,61.1,line [the]                          [SUBJ 61,j,61.1,tubo [la]
      [ATR 62,refueling ]                                 [ATR 62,((re)(fuel)iza) ]
      [ATR 63,between                                     [ATR 63,inter
       [PARG 64,and                                        [PARG 64,kaj
        [PARG-C 65{<47}:57.1                                [PARG-C 65{<47}:57.1
         [ATR 66,right-hand ]]                               [ATR 66,dekstra ]]
        [PARG-C 67{<47}:57.1                                [PARG-C 67{<47}:57.1
         [ATR 68,left-hand ]]]]]]]]]                         [ATR 68,((mal)dekstra) ]]]]]]]]

["(4)"                                               ["(4)"
 [GOV 69:5-6-10-12                                    [GOV 69:5-6-10-12
  [70(10),switch [the]                                 [70(10),(((70.1,bar)sxalt)ilo) [la]
   [ATR 70.1,shutoff ]
   [ATR 71,of                                            [ATR 71,de
    [PARG 72{=65}:3                                       [PARG 72{=65}:3
     [ATR 74:66 ]]]]                                       [ATR 74:66 ]]]]
   [76(12),"OPEN" ]]]                                    [76(12),"OPEN" ]]]

["(4)(a)"                                            ["(4)(a)"
 [GOV 78:13-16-21                                     [GOV 78:13-16-21
   [81,go                                               [81-as,cxesi
    [PRED 81/1,off ]                                     [INFC 81/1,lumi ]
    [SUBJ 83{<27}:18-19                                  [SUBJ 83{<27}:18-19
     [84:43                                               [84:43
      [85:70 ]]]]                                         [85:70 ]]]]
   [90:54-54.1-57                                       [90:54-54.1-57
     [93(57):65 ]]]]                                      [93(57):65]]]]

["(5)"                                               ["(5)"
 [GOV 95,hold                                         [GOV 95-u,teni
  [OBJ 96,switch [the]                                 [OBJ 96,((sxalt)ilo) [la]
   [ATR 97:6 ]]                                         [ATR 97:6 ]]
  [ADVC 101,to
   [PARG 102,"TEST" ]]]]

 ["(5)(a)"
  [GOV 103:13-16-21
     [106:41-42
      [107{=83}:83-85
        [109,valve [the]
         [ATR 109.1:70.1 ]
         [ATR 110:66]]]]
     [111,stop
      [SUBJ 112{=90},flow [the]
       [ATR 112.1,fuel ]]]]]
```

```
  [ADVC 101,cxe
   [PARG 102,"TEST" ]]]]

 ["(5)(a)"
  [GOV 103:13-16-21
     [106:41-42
      [107{=83}:83-85
        [109,((109.1:70.1)valvo) [la]

         [ATR 110:66]]]]
     [111-as,cxesi
      [SUBJ 112{=90},((112.1,fuel)fluo) [la] ]]]]
```

Explanation of syntactic labels

ADVA	adverbial adjunct
ADVC	adverbial complement
ATR	attribute
GOV	governor
INFC	infinitival complement
OBJ	direct object
PARG	prepositional argument
PARG-C	coordinated prepositional argument
PRED	predicative
SUBC	subordinate clause
SUBC-C	coordinated subordinate clause
SUBJ	subject
PROA	propositional adjunct
LIA	linking adjunct
PREA	prepositional adjunct