VS

Paper submitted for inclusion in "Language Technology in the Netherlands" (ed. Bunt). 1990-06-22.

THE BILINGUAL KNOWLEDGE BANK (BKB)

A new foundation for machine translation

Victor Sadler
BSO/Research
P.O. Box 8348
NL-3503 RH Utrecht
The Netherlands

email: sadler@dlt1.uucp

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. After an introduction to the general problem of knowledge sources for machine translation (MT), section 2 considers the merits of bilingual corpora for this purpose. The structure needed to convert a bilingual corpus to a knowledge bank is discussed in section 3, and its actual construction in section 4. Section 5 covers the application of the knowledge bank for machine translation and includes an extensive simulation for a sample sentence. Finally, sections 6 and 7 summarize the advantages of this approach and compare it with the work of other researchers.

1 INTRODUCTION

The concept of the Bilingual Knowledge Bank (henceforth "BKB") has grown out of 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 is not a part of the DLT English to French prototype system first demonstrated in 1987, it is central to the present design for a future production-scale DLT system. It was first described in an internal DLT report (Sadler 1989a). A pilot implementation of the BKB has been completed (Sadler / Vendelmans forthc.), and wider feasibility studies are already well advanced (see also Sadler 1989b, Part II).

There are two major obstacles determining the speed and cost of development of a practical MT 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

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

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 hand-held 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.*

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 too subtle to be captured in terms of semantic features.

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] *Enter the document title* you want *the converted document* to have. =
Indiquez le titre à attribuer au *document converti.*

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 <u>two</u> 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.*[2]

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

---

2 Byrd *et al.* (1987).

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.*[3]

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 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[4] 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 <u>all</u> the information expressed or implied in human language is of itself inadequate for the purposes of machine translation. Moreover, the decompositional methods mentioned above 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.*[5]

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 <u>explain</u> 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

---

[3] Example from Piron (1988).

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

[5] Bennett & Slocum (1985).

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 – unlike that of the surgeon or mechanic – always leaves a readable, and very often machine-readable, trace. True, the translated text does not explain how or why the translator came to choose the words it contains. But at least it is concrete evidence. In principle, it should be possible to devise a computer system – not necessarily an "expert" system – to infer lexical equivalences and other local 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 was 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. 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.

Kjærsgaard (1987, 1989) has implemented a prototype of a corpus-based tool for translators under the name of REFTEX. This consists of a program which generates a concordance of the words in a bilingual text and allows the operator to scroll through the occurrences of any given expression in either language. The two versions of the corpus are aligned paragraph by paragraph. For each occurrence, the system displays both the enclosing paragraph and the translation of that paragraph. It is up to the user to decide whether the occurrence displayed has the intended meaning; what, if any, is its translation in the corresponding paragraph in the other language; and whether that translation is appropriate to the job in hand.

A somewhat more sophisticated software support for human translation has recently been proposed by Harris (1988a, 1988b, 1988c) 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 hyper-text system.[6] 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 Harris's definition of a translation unit it is clear that his proposal primarily concerns multi-word expressions and more complex translation units, since his aim is to supplement the word-for-word equivalences provided by standard dictionaries.

For a machine translation project, the problem of dictionary building is much broader than that of supplementing existing dictionaries for translators or of providing statistical tools for the lexicographer. And the concept of a translation unit needs to be defined rather more rigorously than it does for the purposes of an interactive translation aid. The present proposal for a Bilingual Knowledge Bank for MT contrasts with all of the research described above, 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 three separate dimensions: the bilingual dimension of cross-language equivalence, and the monolingual dimensions of syntactic structure and text coherence (deixis, reference and the like). This three-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 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

[6] Melby (1988: 413) appears to have used the WordCruncher concordance tool to similar effect.

the morpheme to the overall text structure, thus doing away with the need for separate dictionaries of word-level equivalences, verbal case-frames etc. Instead of the dictionary being derived from the corpus, in the BKB approach the dictionary is the corpus.

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. As already pointed out in the Introduction, existing knowledge representation techniques are far too labour-intensive to be useful for large-scale knowledge banks.

Now I have already suggested above that a knowledge base for MT 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]

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 must 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 <u>unambiguous</u>, at least for the practical purposes of translation.

3 THE STRUCTURE OF THE BKB

In this section I am only concerned with devising an appropriate structure. The question of <u>how</u> such a structure can be built up semi-automatically will be answered later, under 4.

---

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.

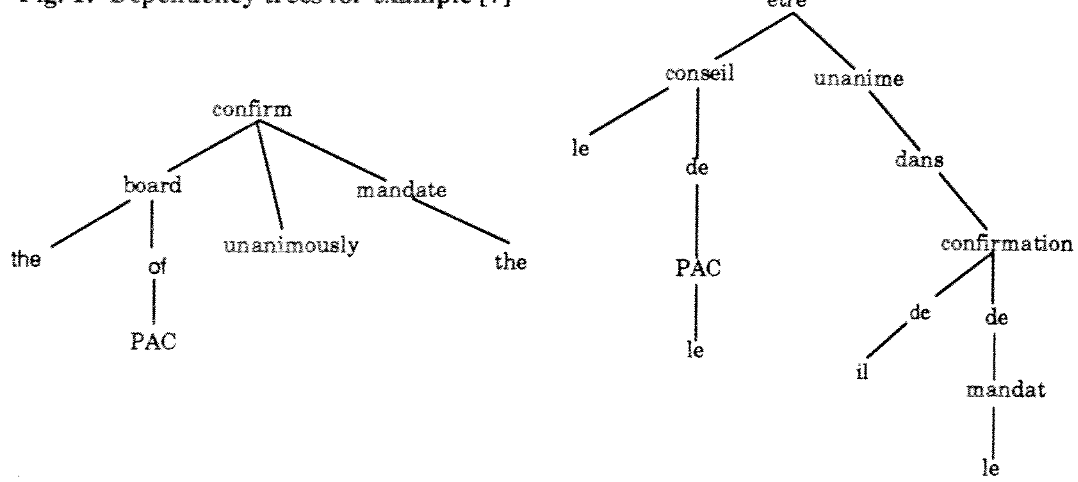
3.1 Dependency syntax

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:[9]

[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.[10]

Fig. 1: Dependency trees for example [7]



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. It is no coincidence that dependency trees bear a strong resemblance to semantic networks. An additional point in favour of dependency is the smaller number of tree nodes required in comparison with constituency analysis. For very large corpora, this compaction is significant.

3.2 Translation units

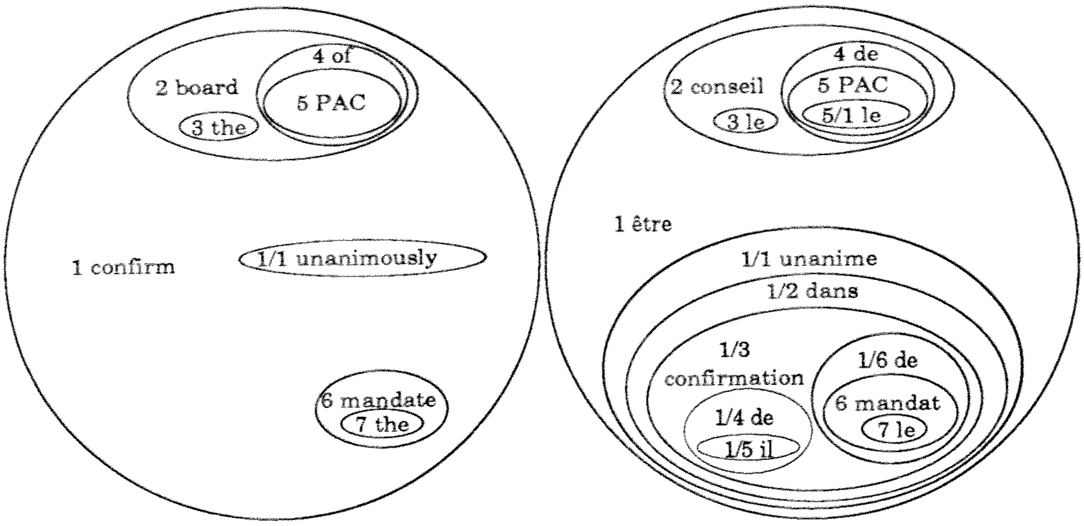
The next step is to divide the syntactic structure into translation units. A translation unit, as the term was used by Harris, consists of two fragments of text in different languages, which the translator considers equivalent. The essence of a unit is that it is autonomous. It can be used without necessarily 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.

[9] Adapted from Harris (1988c).

[10] Here, word forms have been normalized. Information on syntactic features and functions is omitted for the sake of clarity.

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 subtree in figure 1. Each of the seven identifiable translation units has been assigned an identification number (ID).

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



"The board of PAC unanimously confirms the mandate."

"Le conseil du PAC est unanime dans sa confirmation du mandat."

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 ellipse. TU 2 is the subject noun phrase, 3 the determiner, 4 the prepositional phrase, etc. From these examples it should be immediately clear that each of the primary translation units corresponds to a (sub)tree. On the other hand, it is not necessarily the case that every subtree corresponds to a translation unit. The French subtree governed by *dans*, for instance, does not constitute a translation unit. There is no subtree 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 is the second subtree in the bound dependent of TU 1.

Table 1: Translation units identified in Fig. 2

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

Harris's statement that "translation units [...] mostly consist of whole phrases and even whole clause or sentences" is true enough. 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 subtrees 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 subtrees. For example, subtracting TU 2 from TU 1 in figure 2 yields the equivalence of *unanimously confirm the mandate* with *être unanime dans sa confirmation du mandat*. Further subtracting TU 6 generalizes the verbal construction to *unanimously confirm* and *être unanime dans sa confirmation de*. Table 2 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 (or indeed in the language at large) without a dependent, can still be accessed as a (generalized) TU.

Table 2: Translation units derived from Table 1

TU coding	English phrase	French phrase
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

Box 1 shows the TU-coded structure of figure 2 in text form. 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 words *of* and *de* and their dependent TU 5.

Box 1: Dependency trees for example [7]

```
[1,confirm                  [1,être
   [1/1,unanimously ]          [1/1,unanime ]
                                  [1/2,dans
                                     [1/3,confirmation
                                        [1/4,de ]
                                           [1/5,il ]
                                        [1/6,de
  [6,mandate                               [6,mandat
     [7,the ]]                                [7,le ]]]]]
  [2,board                     [2,conseil
     [3,the ]                     [3,le ]
     [4,of                        [4,de
        [5,PAC]]]]                   [5,PAC
                                        [5/1,le ]]]]]
```

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 *sa* has been normalized to *de il*.

3.3 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 (in the broad sense, including deixis). 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 inclusion (PART-OF, MEMBER-OF etc.) and exclusion can also be used. (For the interactive identification of such relations see section 4.3 below.)

Although the network of reference relations created in this way must be approximately the same in each half of the BKB, the two networks will not correspond exactly, because one language may make references which are omitted in the other version. Example [7] contains an illustration of this, where the French possessive form *sa* refers back to *conseil* – a link which is not reflected in the English construction.

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

The reason why *sa* should be normalized to *de il*, as noted earlier, is that this allows the personal pronoun to be identified as entirely co-referent with *le conseil du PAC*.

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

[8] *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

[9] *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.

3.4 Extended example

In order to illustrate the combination of reference identification and coding with that of translation units, a larger text sample is required. The following example is based on a model text for writers using Simplified English as defined in the international aircraft industry (AECMA 1984), together with an ad hoc translation in Esperanto. Each sentence is shown in parallel, in English and Esperanto, first as text and then in the kind of BKB structure already illustrated in Box 1 – but with the addition of the third dimension: that of reference. The structure below includes vertical links between TUs which are co-referential. Translation units are identified by corresponding numbers on each side. The syntactic function labels which have been inserted are explained in table 3 at the end of this section. Additional comments are interpolated between the sentence units.

Outer Wing Tank Test	Testo de la eksteraj alfuelujoj
[GOV 1,test [ATR 2,tank [ATR 3,wing] [ATR 4,outer]]]	[GOV 1,testo [ATR 2,de [PARG 2/1,j,2/2,((3,al)(fuel)ujo) [la] [ATR 4,ekstera]]]]

Note that in the Esperanto version, morpheme structure has also been made explicit to some extent, illustrating the possibility of coding morphemes as translation units. Thus TU 3 consists of the word *wing* in English and of the morpheme *al* in Esperanto, which is part of the word *alfuelujo*.

(1) On the fueling control panel, set the power switch to ON.	(1) Sur la komandpanelo por fuelizado, movu la alimentsxaltilon al "ON".
["(1)" [GOV 5,set [ADVA 6,on [PARG 7{=103},panel [the] [ATR 7/1,control] [ATR 8,fueling]]] [OBJ 9,switch [the] [ATR 10,power]] [ADVC 11,to [PARG 12,"ON"]]]]	["(1)" [GOV 5,movi [ADVA 6,sur [PARG 7{=103},((komand)panelo) [la] [ATR 8,por [PARG 8/1,((fuel)izado)]]]] [OBJ 9,(((10,aliment)sxalt)ilo) [la]] [ADVC 11,al [PARG 12,"ON"]]]]

In this first complete sentence of the sample structure, note that the cross-coding of the dependent definite articles *the* and *la* has been left out, merely to save space. Note also that no independent translation of *fueling control* is possible on the basis of this text, since there is no corresponding TU.

TU 7 introduces the first reference link. The same control panel will be referred to again in instruction (5), by TU 103. The referential link between TUs 7 and 103 is shown here by the coding between braces directly following the TU code: "{=103}". Identity between the concepts represented by two translation units is marked by an equals sign. TUs 7 and 103 are identical not only in their referential content, but also in their form (in both languages). Of course, the form of such co-referent TUs need not be identical. In the case of pronouns, for example, the forms will be completely different.

(1a) Make sure that: - the power light is off; - the overflow valve lights	(1a) Kontrolu, ke: - la signallampo de la alimento ne lumas; - la signallampoj de la superversxaj valvoj

are off;
- the shutoff valve lights are on.

```
["(1)(a)"
 [GOV 13,make
  [PRED 13/1,sure ]
  [OBJ 14,that
   [SUBC 15,"; -"
    [SUBC-C 16,"; -"
     [SUBC-C 17,be
      [PRED 17/1,off ]
      [SUBJ 18,light [the]
       [ATR 19,power ]]]
     [SUBC-C 20,be
      [PRED 20/1,off ]
      [SUBJ 21{>40},s,22,light [the]
       [ATR 23,valve
        [ATR 24,overflow ]]]]
    [SUBC-C 25,be
     [PRED 25/1,on ]
     [SUBJ 26{=52},s,27,light [the]
      [ATR 28,valve
       [ATR 29,shutoff ]]]]]]]]
```

ne lumas;
- la signallampoj de la baraj valvoj lumas.

```
["(1)(a)"
 [GOV 13,kontroli
  [OBJ 14,ke
   [SUBC 15,"; -"
    [SUBC-C 16,"; -"
     [SUBC-C 17,lumi
      [ADVA 17/1,ne ]
      [SUBJ 18,((signal)lampo) [la]
       [ATR 19,de
        [PARG 19/1,alimento [la] ]]]]
     [SUBC-C 20,lumi
      [ADVA, 20/1,ne]
      [SUBJ 21{>40},j,22,((signal)lampo) [la]
       [ATR 23,de
        [PARG 23/1,j,valvo [la]
         [ATR 24,((super)versxa) ]]]]]]
    [SUBC-C 25,lumi
     [SUBJ 26{=52},j,27,((signal)lampo) [la]
      [ATR 28,de
       [PARG 28/1,j,valvo [la]
        [ATR 29,bara ]]]]]]]]
```

TU 17 in the above structure equates the English verbal construction *be off* with the Esperanto *ne lumi* ('not shine'). Note that 17/1 *off* on the English side is not translatable by 17/1 on the Esperanto side, *ne*, because the slash indicates that each of them is a bound dependent.

As in the previous sentence, the equals sign at TU 26 marks its referential identity with TU 52. But TU 21 illustrates another kind of conceptual reference: the inclusion relation, shown here by a ">" sign between braces. The coding at TU 21, *the overflow valve lights*, indicates that this expression refers to a concept which includes the concept referred to by TU 40, *the lights for the overflow valves of the outer wing tanks*.

(2) Apply pressure to the refueling system.

```
["(2)"
 [GOV 30,apply
  [OBJ 31,pressure ]
  [ADVC 32,to
   [PARG 33,system [the]
    [ATR 34,refueling ]]]]]
```

(2) Apliku premon al la sistemo de refuelizado.

```
["(2)"
 [GOV 30,apliki
  [OBJ 31,premo ]
  [ADVC 32,al
   [PARG 33,sistemo [la]
    [ATR 34,de
     [PARG 34/1,((re)(fuel)izado) ]]]]]]
```

(2a) 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.

```
["(2)(a)"
 [GOV 35,make
  [PRED 35/1,sure ]
  [OBJ 36,that
   [SUBC 37,"; -"
```

(2a) 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.

```
["(2)(a)"
 [GOV 35,kontroli
  [OBJ 36,ke
   [SUBC 37,"; -"
    [SUBC-C 38,"; -"
```

```
[SUBC-C 38,"; -"
 [SUBC-C 39,come
  [PRED 39/1,on ]
  [SUBJ 40{<21},s,41,light [the]
   [ATR 42,for
    [PARG 43,s,44,valve [the]
     [ATR 45,overflow ]
     [ATR 46,of
      [PARG 47{=59},s,48,tank [the]
       [ATR 49,wing ]
       [ATR 50,outer ]]]]]]
 [SUBC-C 51,stay
  [PRED 51/1,on ]
  [SUBJ 52{=26>86},s,53,light [the]
   [ATR 53/1,valve
    [ATR 54,shutoff ]]]]
[SUBC-C 55,flow
 [ADVA, 56,not]
 [SUBJ 57,fuel ]
 [ADVC 58,into
  [PARG 59{=47=67},s,60,tank [the] ]]]]]]]
```

```
 [SUBC-C 39,((ek)lumi)
  [SUBJ 40{<21},j,41,((signal)lampo) [la]
   [ATR 42,de
    [PARG 43,j,44,valvo [la]
     [ATR 45,((super)versxa) ]
     [ATR 46,de
      [PARG 47{=59},j,48,((49,al)(fuel)ujo) [la]
       [ATR 50,ekstera ]]]]]]
 [SUBC-C 51,(lum)adi
  [SUBJ 52{=26>86},j,53,((signal)lampo) [la]
   [ATR 53/1,de
    [PARG 53/2,j,valvo [la]
     [ATR 54,bara ]]]]]
[SUBC-C 55,flui
 [ADVA, 56,ne]
 [SUBJ 57,fuelo ]
 [ADVC 58,en
  [PARG 59{=47=67},j,60,((fuel)ujo) [la] ]]]]]]
```

In instruction (2a) we see the converse of the relation noted earlier at TU 21: since TU 21 includes TU 40, this link can also be read the other way round: TU 40 is included in (is a part or member of) TU 21.

The combination of reference links at TU 52 indicates that this item, *the shutoff valve lights*, already identified with TU 26, includes the concept referred to by TU 86 in instruction (4a) below: *the light for the shutoff switch of the right-hand outer wing tank*.

TU 59, *the tanks*, is identified as co-referent with both TU 47, *the outer wing tanks* and the coordinated phrase which constitutes TU 67, *the right-hand tank and the left-hand tank*.

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

```
["(3)"
 [GOV 61,make
  [PRED 61/1,sure ]
  [OBJ 62,be
   [ADVC 62/1,there ]
   [SUBJ 62/2,leakage
    [ATR, 62/3,no]
    [ATR 62/4,from
     [PARG 63,s,64,line [the]
      [ATR 65,refueling ]
      [ATR 66,between
       [PARG 67{=59},and
        [PARG-C 68{=76},tank [the]
         [ATR 69,right-hand ]]
        [PARG-C 70,tank [the]
         [ATR 71,left-hand ]]]]]]]]]]
```

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

```
["(3)"
 [GOV 61,kontroli
  [OBJ 61/1,ke
   [SUBC 62,liki
    [ADVA 62/1,ne]
    [SUBJ 63,j,64,tubo [la]
     [ATR 65,((re)(fuel)iza) ]
     [ATR 66,inter
      [PARG 67{=59},kaj
       [PARG-C 68{=76},((fuel)ujo) [la]
        [ATR 69,dekstra ]]
       [PARG-C 70,((fuel)ujo) [la]
        [ATR 71,((mal)dekstra) ]]]]]]]]
```

Instruction (3) contains a nice example of a lexical metataxis (structural transformation) rule in TU 62, which after the subtraction of its dependent TU 63 can be generalized as [be [there] [leakage [no] [from [X]]]] = [liki [ne] [X]]

Since TU 67, *the right-hand tank and the left-hand tank*, has been identified with TU 59, *the tanks*, and since the concept referred to by each member of a coordination such as TU 67 is necessarily included in the concept represented by the whole syntagma, the system can automatically conclude that TU 68, *the right-hand tank*, and TU 70, *the left-hand tank*, are included in TUs 47 and 59, i.e. that they are members of *the outer wing tanks*.

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

```
["(4)"
 [GOV 72,set
  [OBJ 73,switch [the]
   [ATR 74,shutoff ]
   [ATR 75,of
    [PARG 76{=68=91},tank [the]
     [ATR 77,wing ]
     [ATR 78,outer ]
     [ATR 79,right-hand ]]]]
  [ADVC 80,to
   [PARG 81,"OPEN" ]]]]
```

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

```
["(4)"
 [GOV 72,movi
  [OBJ 73,(((74,bar)sxalt)ilo) [la]
   [ATR 75,de
    [PARG 76{=68=91},((77,al)(fuel)ujo) [la]
     [ATR 78,ekstera ]
     [ATR 79,dekstra ]]]]
  [ADVC 80,al
   [PARG 81,"OPEN" ]]]]
```

(4a) 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.

```
["(4)(a)"
 [GOV 82,make
  [PRED 82/1,sure ]
  [OBJ 83,that
   [SUBC 84,"; -"
    [SUBC-C 85,go
     [PRED 85/1,off ]
     [SUBJ 86{=111<52},light [the]
      [ATR 87,for
       [PARG 88,switch [the]
        [ATR 89,shutoff ]
        [ATR 90,of
         [PARG 91{=76=98},tank [the]
          [ATR 92,wing ]
          [ATR 93,outer ]
          [ATR 94,right-hand ]]]]]]
    [SUBC-C 95{=117},flow
     [SUBJ 96,fuel ]
     [ADVC 97,into
      [PARG 98{=91},tank [the]
       [ATR 99,right-hand ]]]]]]]
```

(4a) Kontrolu, ke:
- la signallampo de la barsxaltilo de la dekstra ekstera alfuelujo cxesas lumi;
- fuelo fluas en la dekstran fuelujon.

```
["(4)(a)"
 [GOV 82,kontroli
  [OBJ 83,ke
   [SUBC 84,"; -"
    [SUBC-C 85,cxesi
     [INFC 85/1,lumi ]
     [SUBJ 86{=111<52},((signal)lampo) [la]
      [ATR 87,de
       [PARG 88,(((89,bar)sxalt)ilo) [la]
        [ATR 90,de
         [PARG 91{=76=98},((92,al)(fuel)ujo) [la]
          [ATR 93,ekstera ]
          [ATR 94,dekstra ]]]]]]
    [SUBC-C 95{=117},flui
     [SUBJ 96,fuelo ]
     [ADVC 97,en
      [PARG 98{=91},((fuel)ujo) [la]
       [ATR 99,dekstra ]]]]]]]
```

Here, TU 86, *the light for the shutoff switch of the right-hand outer wing tank*, is identified with TU 111, *the light for the right-hand shutoff valve*, because these two different forms in fact refer to the same entity. It does not follow, of course, that the TUs are interchangeable in translation. One may be more appropriate than the other in a particular context. But 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 the example under 5.2 below.)

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

```
["(5)"
 [GOV 100,hold
  [OBJ 101,switch [the]
   [ATR 102,on
    [PARG 103{=7},panel [the]
     [ATR 103/1,control]
     [ATR 104,fueling ]]]]
  [ADVC 105,to
   [PARG 106,"TEST" ]]]]
```

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

```
["(5)"
 [GOV 100,teni
  [OBJ 101,((sxalt)ilo) [la]
   [ATR 102,sur
    [PARG 103{=7},((komand)panelo) [la]
     [ATR 104,por
      [PARG 104/1,((fuel)izado) ]]]]]
  [ADVC 105,cxe
   [PARG 106,"TEST" ]]]]
```

(5a) Make sure that:
- the light for the right-hand shutoff valve comes on;
- the fuel flow stops.

```
["(5)(a)"
 [GOV 107,make
  [PRED 107/1,sure ]
  [OBJ 108,that
   [SUBC 109,"; -"
    [SUBC-C 110,come
     [PRED 110/1,on ]
     [SUBJ 111{=86},light [the]
      [ATR 112,for
       [PARG 113,valve [the]
        [ATR 114,shutoff ]
        [ATR 115,right-hand ]]]]]
    [SUBC-C 116,stop
     [SUBJ 117{=95},flow [the]
      [ATR 118,fuel ]]]]]]
```

(5a) Kontrolu, ke:
- la signallampo de la dekstra barvalvo eklumas;
- la fuelfluo cxesas.

```
["(5)(a)"
 [GOV 107,kontroli
  [OBJ 108,ke
   [SUBC 109,"; -"
    [SUBC-C 110,((ek)lumi)
     [SUBJ 111{=86},((signal)lampo) [la]
      [ATR 112,de
       [PARG 113,((114,bar)valvo) [la]
        [ATR 115,dekstra ]]]]]
    [SUBC-C 116,cxesi
     [SUBJ 117{=95},((118,fuel)fluo) [la] ]]]]]
```

The sample structure shows only such reference links as will normally be inserted during the construction of the BKB. They can, of course, be further extended using the principles of transitivity. For example, given that TU 111, *the light for the right-hand shutoff valve*, has been identified with TU 86, which in turn is known to be included in TU 52, which has further been identified with TU 26, the system can infer that TU 111 is also included in the concept referred to by TU 26, *the shutoff valve lights*, a fact which had not been given explicitly. In this way the system can automatically check the consistency of the knowledge base and improve its coverage. All such referential relations are, of course, equally applicable when the same concepts are referred to by the corresponding terms in the other language, i.e. whenever references link autonomous translation units and not bound dependents.

Finally, some clarification is called for at the very last reference in the sample, where TU 117, *the fuel flow*, is identified with the earlier TU 95, *fuel flows into the right-hand tank*. In spite of the difference in syntactic categories between the head words of these units (a deverbal noun in TU 117, and a verb in TU 95), a reference link can be set to show that the noun phrase refers back to a whole clause and that, conceptually, TU 117 therefore represents an event.

One element which is not visible in the above example is information on the original word order. Dependency trees are not projective. Nevertheless, such information can and should be included in the BKB structure because it is an important part of the linguistic knowledge the text contains, and also because it is needed if the original text is to be reconstituted. It was omitted from the example only for the sake of simplicity.

Table 3: Explanation of syntactic labels

ADVA	adverbial adjunct
ADVC	adverbial complement
ATR	attribute
DET	determiner
GOV	governor
INFC	infinitival complement
INFC-C	coordinated 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

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 the syntaxes of two different languages, the meaning of each is defined by the relevant syntax, and they do not necessarily mean the same.

4 BUILDING 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 after 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 BKB itself can provide more and more support, the larger it becomes. Second, the information contained in one language version can support the processing of the other version, and the addition of further languages to the system can reinforce this effect.

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

4.1 Parsing

To guarantee that the syntactic structures in the BKB are correct, any parser used must be interactive or allow for post-editing of the output structures. In the pilot implementation (see section 1 above), a simple, fast, category-based parser is used which displays only one possible structure per sentence. The operator can correct the tree, if necessary, by exchanging nodes or moving branches with the aid of a mouse, before storing it in the BKB.

The pilot system parser is BKB-supported in that it extracts information on word categories, syntactic functions and the like from the growing BKB – i.e. from that part of the corpus which has already been processed. In future the parser will also match the input string against the structures already stored in the BKB and decide attachments on a probabilistic basis. It will also suggest alternative parses if the first proposal is rejected.

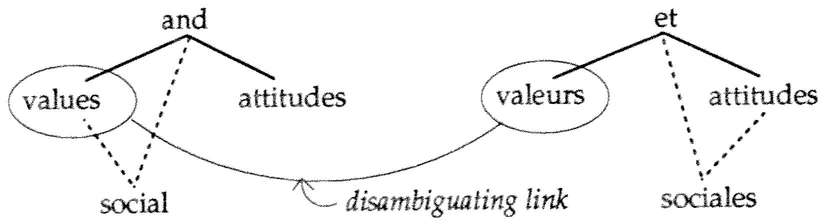
The larger the BKB becomes, the better it can support the parser. This is not to say that certain parses can be automatically excluded on the basis of the existing knowledge bank, but it does mean that the most likely structure can be displayed first, thus considerably easing the job of the operator. The knowledge applied will include not only syntactic probabilities, but the extra-linguistic information stored in the BKB as well. For example, existing knowledge of the death of Maxwell Madondo – or of the fact that Mrs Mandela is still alive – can help to resolve the attachment ambiguity in:

[10] *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]

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[12] 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, once the translation units have been identified. 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.

These observations suggest the following strategy. The parsing of both language versions should be carried out in parallel and interleaved with the identification of translation units. In this way the system can avoid generating parses which may be incompatible in terms of translation units. Figure 3 illustrates this approach. The English adjective can be attached either to the first noun, or else to the coordination. The French adjective can be attached either to the coordination or to the second noun. Therefore, as soon as one of the translation units *values* = *valeurs* and *attitudes* = *attitudes* has been identified, the dependency of *social* on a single noun can be eliminated from consideration, because no dependent is possible in the other language.

Fig. 3: Contrastive structural disambiguation



social attitudes and values = les valeurs et les attitudes sociales

In this example, disambiguation is likely to be automatic as soon as the BKB is large enough to contain examples of the basic TUs involved. In other cases, where human intervention may be required to decide the most likely structural interpretation of the text, the operator too can be greatly aided by being able to compare the two different language versions.

If the aim is to build knowledge banks for a multilingual translation system, it may be profitable for the system to compare each new version of the text with the versions already

---

[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).

[12] Part of the Preface to Munniksma (1975).

processed to improve the parser's performance. Once all structural ambiguities have been resolved (whether automatically or interactively) for the first language pair, there should be very little of this kind of work required from the third language onwards, because each new language can be structurally disambiguated by establishing translation units with 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 obviously has to do most of the work. Gradually, however, the growing amount of 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 the extended example in section 3.4 above. 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. In the pilot implementation, the system attempts to identify all the translation units automatically and presents the results to the operator graphically. After confirmation or correction, the results are stored and are then used to influence subsequent analyses. The system is thus self-improving. Experience with this implementation shows that even with only a few thousand words in the BKB, the system is frequently able to correctly identify all the translation units in quite complex sentences. It seems likely that in a large corpus a high proportion of all sentences could be analysed fully automatically 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 <u>transitive</u> 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 *hipotecario* and *hipoteka* 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. Alternatively, the situation could be rectified by an automatic process in which complex concepts are reduced to their component TUs with the aid of knowledge available elsewhere in the BKB. This reorganization of the knowledge bank need not be restricted to BKBs generated on the transitivity principle, but can be applied on a routine basis to maximize productivity.

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 inferencing over the BKB, 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 (syntactic parsing and identification of translation units) were necessary in order to convert the bilingual corpus into a bilingual dictionary. This third dimension augments the bilingual dictionary with extra-linguistic knowledge.

Just as in the cases of syntactic structure and translation units, the identification of referential links can be strongly supported by the BKB itself. First, expressions such as pro-forms and definite noun phrases which are used to refer to concepts introduced elsewhere in the text, can increasingly be recognized automatically by the fact that previous occurrences of those expressions have been assigned reference links in the BKB. Second, existing links in the BKB can help the system to identify the most likely antecedent or other referent for the expression in question, as well as the type of link involved (e.g. identity, inclusion or exclusion).

The identification of references, like syntactic structures, can also be supported by contrastive analysis. E.g.:

[11] 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 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 grammatical rules for each language, contrastive analysis should, however, significantly reduce the burden on the operator.

Just as for structural disambiguation, the referential identification completed for the first language pair should largely eliminate this aspect of the operator's task from the third language onwards. There will, of course, always be a residue of monolingual references to be resolved. (See example [7] under 3.3 above.)

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.

5 TRANSLATING WITH THE BKB

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

5.1 Syntax

Syntactic analysis and generation are traditionally rule-driven in machine translation systems. The BKB, containing as it does large numbers of syntactic structures, provides a potential substitute for rules. In a BKB-based parser, the analysis of the input string can be based on analogy: a process of matching the input sequence against patterns stored in the BKB structures and selecting those which provide the closest analogy. Van Zuijlen (1989a, 1989b) has explored the potentialities of this approach, and a BKB-based parser is currently (June 1990) under construction. Some suggestions as to the working of such a parser are also contained in section 5.5 below.

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 section 3.4, the BKB structure provides for, and even demands, the systematic identification of the items and events mentioned in the text via the setting of referential links of various types. And this knowledge can be applied during the analysis of a source text to suggest the most likely referent for any referential expression, just as it can during BKB construction (see section 4.3).

Knowledge of a particular entity can be accumulated over a number of references. The sentence

[12] *The Mayor has resigned.*

may occur more than once in a corpus, but the separate occurrences of *the Mayor* 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 only be reliably checked within the TL half of the text representation, because this half 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 simply translated, but must be generated, where required, by the TL part of the metataxis process.

Consider the following example:

[13] 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 [13]) with their referents automatically ensures that they will be translated by a repetition of *neĝis*, since the English subtree 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

[14] *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.

Besides handling reference and deixis, the BKB can preserve the order of syntagmata: although dependency trees are not projective, word order information 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

[15] *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 (Schubert 1987: 181) 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 can 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 can easily be inserted, if these can be identified.

Finally, it is easy to demonstrate the value of the BKB structure 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 section 3.4 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 an apparent contradiction in the instructions. The explanation is that the following expressions actually 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 may have been already introduced, or else may be the only switch on the control panel. The only likely candidate referent in the preceding text 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, however, shows that this referent is not correct, because the power switch has no TEST position. Nor can the definite noun phrase refer to a unique entity, because there are several switches on the panel. Actually, the original text, before "translation" to Simplified English, referred in instruction 5 to the test switch on the fueling control panel, which is a unique entity. Here again, a routine query from the 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 to the knowledge base the knowledge of what switches are to be found on the control panel, what their possible settings are, etc.

5.3 Metataxis

Metataxis, or structural transformation, can be guided by rules which are implicit in the whole BKB structure. As its simplest, this means using the BKB as a dictionary of word-for-word equivalences. But any level of structural complexity can be handled. 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. For example:

[16] *The contents are being deleted. =*
La destruction du contenu est en cours.

When these two sentences have been processed for a BKB, the resulting TUs include

[be [X] [being [deleted]]] =
[être [destruction [la] [de [X]]] [en [cours]]]

where X is a variable obtained by subtracting the TU

[contents [the]] = [contenu [le]]

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 5.4 below.) But the definition of the translation unit, including bound or untranslatable dependents, prevents their abuse in many cases. For example, given the English-Dutch TUs

[17] *1. John has kicked the bucket. = John is doodgegaan.* ('John has died.')
2. bucket of milk = emmer melk
3. at last = eindelijk

the sentence

[18] *John has kicked the bucket of milk.*

cannot be translated by the conjunction of 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,

[19] *John has kicked the bucket at last.*

can be composed of TUs 1 and 3, because *at last* can be attached to the head of TU 1.

5.4 Semantics

The BKB concept offers significant advantages for semantic control, both automatic and interactive (via a dialogue with the user).[13]

5.4.1 Automatic:

The main advantages are the following:

(1) 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 semantic molecules consisting of translation units, and the choice between alternative translations (i.e. between alternative TUs) will be determined by contextual probabilities. The context of a given unit consists of other units, which may be simple words but may also be complex subtrees.

(2) On the other hand, the semantic module can access **content morphemes** below the word level in the BKB, because word grammar is used to structure polymorphemic words.

(3) The vexing problem of identifying and tagging individual **word meanings** in the lexicon (how many senses should we attribute to a word like English *take*?) is operationally 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 an ambiguous word such as the Esperanto *akso* ('axis' or '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 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.

(5) The semantic module can be supported by **text coherence** mechanisms which derive from the BKB their knowledge of textual patterns (discourse structures) and of patterns of reference and deixis. These constraints can greatly reduce the number of alternative translations under consideration. Reference links such as that between *young woman* and *girl* in example [9] above provide a built-in network of semantic relations which can be used to match input phrases with BKB examples.

(6) The use of a bilingual knowledge bank means that it is possible to compute **semantic proximity** on the basis of contextual pattern (Sadler, forthc.; 1989b: 55-58) for any given pair of concepts (i.e. translation units), rather than for mere words, and to apply this measure to either language. If, for example, a system based on an English/Esperanto BKB needs to determine the degree of semantic proximity between the materials *wood* and *iron*, it can do this by comparing the contextual patterns of the translation units

wood = ligno
iron = fero

without clouding the picture with the unrelated senses of the words (e.g. the 'group of trees' sense of *wood* or the 'smoothing instrument' sense of *iron*).

[13] See Sadler (1989b: 149-233) for a more extensive discussion.

(7) **Default choices** can be based on the relative frequencies in the BKB. This is a dynamic criterion, influenced by every new addition to the BKB, including the text being translated. Moreover, since the BKB consists of a number of different texts, the translation process can be made to keep track of the frequency with which units from any given BKB text are being accessed. In this way, the subject matter of the text being translated can be implicitly identified with that of certain parts of the BKB, which can then be given priority over others in the determination of default choices. Text priorities will shift in a fluid manner as the input shifts from one topic to another. Since the text being translated is constantly added to the BKB, this mechanism also automatically weights frequency-based preferences towards those choices already made for the current text.

5.4.2 Interactive:

In the 1988 DLT prototype, a computer-initiated dialogue allowed 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 <u>examples</u>. Every time a translation is selected, the semantic module can be assumed to have found a translation unit in the BKB which best matches 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

[20] *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 is judged satisfactory (e.g. the operator has clicked through the whole cycle of possible translations for the word SUBJECT) the system could allow the operator to scroll through other examples of each interpretation in order to find one more acceptable.

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

5.5 An example of (simulated) BKB-based translation

The following analysis represents a simulation of how a machine translation system based on a BKB as its primary knowledge source might go to work on a sample sentence. The simulation makes use of the model BKBs built as a pilot implementation. The contents were derived from part of a software manual in English, with translations in French and Esperanto. The text in each language amounted to roughly 20,000 words. From these three language versions, two BKBs were produced: one English/Esperanto, and the other Esperanto/French. Although both can be used in either direction, this example translation goes from English to Esperanto and from Esperanto to French. The test sentence chosen is the last-but-one[14] sentence in the BKB:

[21] *You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to allow other users to copy that document.*

This sentence is now considered to have been deleted from the BKB.

The process of translating with a BKB consists basically in identifying the same three types of relation already described in section 4 above for the building of a BKB: syntactic links, translation units and references. These three operations should be seen as interleaved, with structural analysis, transfer and synthesis proceeding in parallel and coming into play intermittently. The procedure is to match input patterns against patterns in the knowledge base. The general strategy suggested for a non-parallel implementation is an incremental, depth-first one (Sadler 1989b: 149ff.): a process of selecting the most likely solution at each step, and backtracking only when forced to by some inconsistency. The most likely solution I define as the one consistent with the best available match with the BKB. I will have more to say later about what constitutes the "goodness" of a match. In the extreme case, however, if the whole of the input sentence happened to be literally present in the BKB, then the translation should be identical to that of the BKB sentence, unless some extra-sentential factor dictates otherwise. As to what constitutes an inconsistency in this approach, backtracking will be indicated whenever the most probable interpretation of the current word (in its input context) conflicts with earlier choices, i.e. with the information being added by the processes of analysis and translation.

The following analysis proceeds word by word, left to right. I will try to formulate the main procedural rules as we come to them.

5.5.1 English to Esperanto translation

Word 1: You...

Where this word constitutes a self-contained translation unit, its Esperanto equivalent is almost always (96%) *vi*.

(TU1) [you] = [vi]

[14] The last sentence of all was too short to be interesting.

> *Procedural rule 1:* If a given input pattern matches more than one structure in the BKB, select the most frequent structure.

Word 2: You **can...**

Now that more than one word of the sentence is available, a search can be made for a possible syntactic link between the input words.[15] The BKB contains 181 occurrences of the pattern *you can*, i.e. of a syntactic dependency link between these two words where *you* precedes *can* in the linear string. In all 181 cases of this link in the BKB, *you* is governed by *can* and functions as the subject of the verb. (The word *you* never has any dependents.) In 83% of these cases, the word *can* governs the whole sentence. So, by rule 1, this structure is selected.

> *Procedural rule 2:* If a structure in the BKB has been selected as matching a given input pattern, augment the SL structure with any syntactic information specific to the selected structure. This information can then be used to constrain further selections.

Having matched the input pattern, we can augment the SL structure with syntactic function labels from the BKB as follows:[16]

[GOV can [SUBJ -you!]]

(In this representation, a '+' or '-' indicates a word's position in the linear string, relative to its governor. A '!' after a word means that on BKB evidence no further dependents are to be expected at that point.) This structure appears 151 times in the BKB. Of these 151 examples, all are translatable, i.e. they either constitute a complete translation unit, or else they form the head of such a unit, but do not form a bound, inseparable dependent of some larger unit. The most frequent TU (58%) is

(TU2) [can [-you]] = [povi [-vi]]

which also comprises, and therefore replaces, TU 1.[17]

Word 3: You can **also...**

The word *also*, like *you*, never has a dependent in this BKB. In a working translation system, its governor would normally be determined by referential factors, because the word *also*, like a number of other so-called "floaters", serves to flag a disjunctive reference, or exclusion relation. For example, in

[22] *If the document you delete is the last in a folder, PC ALL-IN-1 also deletes the folder.*

the word *also* marks the fact that certain concepts are being contrasted: in this case *document* and *folder*. (Neither of these need be contiguous with *also*, of course!) Such referential aspects cannot, however, be taken into account here, because we are processing the example sentence in isolation and the referent would need to be identified in a previous sentence. So iden-

---

15 The interface for the model BKB allows the user to retrieve, for any given pattern *X Y ...*, all examples in which X, Y, ..., or their normalized forms, appear in the order given (though not necessarily as a string) and occupy adjacent nodes in the dependency tree. For example, the search pattern *send to* will retrieve, among others, the example *For information on sending messages with attachments to users on other systems...* because *to* depends on *sending*, which is a derivative of *send*.

16 See table 3 at the end of section 3.4 for an explanation of the function labels.

17 A TU also contains syntactic function labels, of course, but I omit these here, wherever possible, for the sake of simplicity.

tifying the governor of *also* will have to depend on other criteria.

A search for the pattern *you can also* turns up one occurrence[18] which matches the provisional structure on the SL side. The example sentence in the BKB is

[23] *You can also change the drawer and folder.*

and, under rule 2, the SL structure can be augmented to

[GOV can [SUBJ -you!] [ADVA +also!]]

Now, in the introduction to this section I suggested that the translation of an input sentence which happens to appear in its entirety in the BKB should normally be identical to the BKB translation. This claim implies a matching principle which I will call the "principle of maximal match".

> *Procedural rule 3:* If two different input patterns, one of which is a subset of the other, both match the BKB, preference should be given to the larger pattern. This rule has precedence over rule 1 (frequency criterion).
> Corollary: If there is more than one possible way of matching an input pattern against the BKB by combining one or more partial matches, preference should be given to that solution which involves the smallest number of partial matches.

This rule implies that, other things being equal, the translation of any given input string should consist of as few translation units as possible.

If sentence [23] above contained a translation unit corresponding to the pattern *you can also*, this unit would have priority, according to the principle of maximal match, over any compositional translation made up of smaller units. As it happens, however, the three-word structure does not constitute a TU in sentence [23], nor does the pattern *can also*. (The expression *you can also change* is translated as a single unit.) The match with sentence [23] has provided some additional syntactic information on the SL side, but it cannot provide an extension of the translation.

The next possibility is to look for a match with either *you also* or *can also* (including the syntactic functions already selected, of course). The first of these fails, but a search restricted to the latter pattern turns up two sentences, in both of which *can also* does constitute a translation unit, namely:

(TU3) [can [+also]] = [povi [+ankaŭ]]

Since TU 3 overlaps TU 2, the output translation can be extended by merging the two units, thus:

[povi [-vi] [+ankaŭ]]

Word 4: You can also **copy...**

This pattern is not available in the BKB. Given the tentative parse established so far, the only likely attachment for *copy* is to the main verb *can*. Looking for the pattern *can copy* yields 4 examples, all of which also cover the broader pattern *you can copy*, including the syntactic functions already identified on the SL side. In all four cases, *copy* is an infinitival complement of the main verb. So the provisional input structure can be extended to:

[18] The string *you can also* occurs 17 times, but in 16 of these examples the three words are not linked syntactically.

[GOV can [SUBJ -you!] [ADVA +also!] [INFC +copy]]

As for the translations, three of the four examples contain the unit

(TU4) [can [-you] [+copy]] = [povi [-vi] [+kopii]]

TU 4 overlaps TU 3 and contains TU 2, which can now be discarded. The translation is now:

[povi [-vi] [+ankaŭ] [+kopii]]

Word 5: You can also copy **a...**

The indefinite article *a* appears 667 times. It always has a governor to its right, and never has any dependents. Hence no link to the provisional structure can be made for the time being. Only 51 of the 667 occurrences in the BKB are translated. This means that the default translation of this word takes the form of a collocation with its governor. What's more, the translation of the collocation is most commonly identical with the translation of the governor alone, so that the default equivalence is a null string:

(TU5) [a] = [ ]

(In other words, Esperanto does not use indefinite articles).

Word 6: You can also copy a **document...**

There are three words in the provisional structure to which the word *document* could conceivably be linked: *can, copy* and *a*.

The search pattern *can document* fails, but *copy document* turns up 23 examples, of which the best match is provided, under rule 3, by two sentences which cover the broader pattern *you can copy document*. Both examples show *document* as the direct object of *copy*, so that the provisional input structure is now

[GOV can [SUBJ -you!] [ADVA +also!] [INFC +copy [OBJ +document]]]

The translation unit is:

(TU6) [can [-you] [+copy [+document]]] = [povi [-vi] [+kopii [+dokumento]]]

which contains and replaces TU 4.

The pattern *a document* appears 197 times, always with the same syntactic relation, so the previously unattached article can now be made to depend on the noun:

[GOV can [SUBJ -you!] [ADVA +also!] [INFC +copy [OBJ +document [DET -a!]]]]

The default (62%) translation unit is simply

(TU7) [document [-a]] = [dokumento]

which makes TU 5 superfluous. The output translation now consists of TUs 3, 6 and 7:

[povi [-vi] [+ankaŭ] [+kopii [+dokumento]]]

Word 7: You can also copy a document **to...**

There are three possible candidates to which the word *to* might be attached: *can, copy* and *document*. The first produces nothing. The second search pattern, *copy to*, produces 8 examples, one of which was also selected at word 6:

[24] *When you are finished editing you can copy the document back to your SERVER DRAWER.*

This example matches the largest portion of the input pattern, namely *you can copy document to*, and is therefore given preference by rule 3. It has *to* as a prepositional adjunct to *copy*. The third search, for *document to*, returns only one example, in which the noun is plural and depends on the preposition (in a relative clause). Since this conflicts with the SL structure (where *document* already has a governor), this match fails. Consequently we can extend the SL structure to:

> [GOV can [SUBJ -you!] [ADVA +also!]
> [INFC +copy [OBJ +document [DET -a!]] [PREA +to]]]

The translation unit is:

(TU8) [can [-you] [+copy [+document] [+to]]] = [povi [-vi] [+kopii [+dokumento] [+en]]]

which fits the output structure and can now replace TU 6 in the translation:

> [povas [-vi] [+ankaŭ] [+kopii [+dokumenton] [+en]]]

Word 8: You can also copy a document to **your...**

The word *your* occurs 252 times in the BKB. It always has a governor to its right and never has any dependents. For the time being, then, it can only be matched on its own. In 201 cases it heads a translation unit, the commonest (63%) being

(TU9) [your] = [via]

Word 9: You can also copy a document to your **SERVER...**

The word *SERVER* appears 14 times, always with a governor to its right. This is always *DRAWER* and always the next word. It never appears in the BKB with dependents of its own, so a link with *your* is unlikely. It is usually (57%) translated in the collocation, but does form a separate TU on 6 occasions, the commonest being

(TU10) [SERVER] = [SERVILO]

Word 10: You can also copy a document to your SERVER **DRAWER...**

There are 14 examples of the pattern *SERVER DRAWER*. One of the 14 – example [24] quoted under Word 7 – also matches a much broader pattern in the input: *you can copy document to your SERVER DRAWER*, and the structure also fits the provisional SL structure. This is clearly the best available match. The SL structure now becomes:

> [GOV can [SUBJ -you!] [ADVA +also!]
> [INFC +copy [OBJ +document [DET -a!]]
> [PREA +to [PARG +DRAWER [DET -your!] [ATR -SERVER!]]]]]

This structure also constitutes a translation unit in the BKB:

(TU11) [can [-you] [+copy [+document] [+to [+DRAWER [-your] [-SERVER]]]]] =
[povi [-vi] [+kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]]

This unit includes TUs 8, 9 and 10, which can therefore be discarded. The complete translation so far is now:

> [povi [-vi] [+ankaŭ] [+kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]]

Word 11: You can also copy a document to your SERVER DRAWER **and...**

The word *and* occurs 287 times. It always has exactly two coordinated dependents, one on its left and one on its right. Its most frequent role (20%) is that of sentence governor. This role would also be possible here. After all, *can* is the only word encountered so far which has not been assigned a governor, and which could therefore fill the left dependent slot for the coordinator. This would be a premature conclusion, however, because it would ignore the fact that coordinators are very special words requiring special treatment.

First, they can coordinate virtually anything at all, provided the items coordinated play a similar syntactic role in the sentence. So, in many cases, neither a computer system nor a human reader or listener can form a sensible idea of what a conjunction is going to coordinate until the continuation of the sentence is known. In the example sentence, the words *can, copy* and *DRAWER* are all plausible left-hand coordinates, and the choice between them will be more reliable when the possible right-hand coordinates are known.

Second, in the model of dependency grammar adopted for the DLT project a coordinator governs the coordinated items (Schubert 1987: 114). This means that adding a coordinator on the right of an existing string can reverse the direction of one of the existing dependency links, converting a right-hand dependent into a left-hand one. In other words, some rearrangement of the structure built up before the appearance of the coordinator must be regarded as normal and should not have a negative influence in the process of deciding the most likely extension of the tree.

> *Procedural rule 4:* The appearance of a coordinator in the input pattern can cause earlier selections to be overridden, in that one dependency link is broken and the coordinator is inserted between the former governor and dependent. The appearance of a coordinator in the input string is therefore a signal to relax the syntactic constraints of the provisional SL structure in searching for patterns which include the coordinator.

Checking the possible links between *and* and the words to its left, we quickly discover that the example in the BKB which matches the broadest pattern from the input string, *you can copy document to and*, is:

[25] *you can copy the document to a local drawer and edit it there*

This is the best available match by rule 3. But problems arise when we try to fit it into the existing SL structure. In the existing structure, *copy* depends on *can*; in the new example, on the other hand, *copy* depends on *and*, which in turn depends on *can*. To combine the two structures, then, the coordinator needs to be inserted between the two verbs in the dependency tree. Applying rule 4, the SL structure becomes

```
[GOV can [SUBJ -you!] [ADVA +also!]
 [INFC +and
  [INFC-C -copy [OBJ +document [DET -a!]]
   [PREA +to [PARG +DRAWER [DET -your!] [ATR -SERVER!]]]]]]
```

where the labels above and below the coordinator are taken from the new example [25].

An important consequence of this reconstruction in the SL is that any previously selected TU which bridges the point of insertion of the coordinator can no longer be maintained and must be replaced by other units. This is the case now with TU 11, which must be discarded. However, it can be replaced by a smaller TU from the same example sentence:

(TU12) [+copy [+document] [+to [+DRAWER [-your] [-SERVER]]]] =
[+kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]

The new example [25] contains the TU

(TU13) [can [-you] [+and [-copy [+document] [+to]]]] =
[povi [-vi] [+kaj [-kopii [+dokumento] [+en]]]]

The output translation can now be derived from TUs 3, 7, 12 and 13:

[povi [-vi] [+ankaŭ] [+kaj [-kopii [+dokumento]
[+en [+TIRKESTO [-via] [+SERVILO]]]]]]

Word 12: You can also copy a document to your SERVER DRAWER and **use...**

Checking possible left links for this word including the provisional syntactic structure shows that the broadest match is with the pattern *you can and use* in the following example:

[26] *Then you can edit and use it as you would any WPS-PLUS document.*

This example extends the structure to:

[GOV can [SUBJ -you!] [ADVA +also!]
[INFC +and
[INFC-C -copy [OBJ +document [DET -a!]]
[PREA +to [PARG +DRAWER [DET -your!] [ATR -SERVER!]]]]
[INFC-C +use]]]

The TU corresponding to the maximal match is

(TU14) [can [-you] [+and [+use]]] = [povi [-vi] [+kaj [+uzi]]]

which fits the translation to date, giving:

[povi [-vi] [+ankaŭ] [+kaj
[-kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]
[+uzi]]]

Word 13: You can also copy a document to your SERVER DRAWER and use **a...**

Further input awaited (see remarks at Word 5).

Word 14: You can also copy a document to your SERVER DRAWER and use a **File...**

There is only one example in the BKB of the word *File* (with a capital 'F') linked to any of the accessible[19] words to its left in the input string:

[27] *Refer to a VMS File (RVF).*

which, by rule 2, allows a link to be laid to the previously unattached article:

[File [DET -a!]]

> *Procedural rule 5:* If the BKB structures selected by different parts of the input pattern are discontinuous, attempt to connect the existing selections by generalizing, first to basic word forms and then to the level of syntactic categories and functions.

Since no literal link has been found with the body of the SL structure, the first step is to generalize to basic word forms. However, replacing *File* with *file* still fails to produce a link.

19 By "accessible" I mean other than the left-hand dependents of the coordinator in the provisional structure, which can be ruled out. It appears to be a general rule, at least for English, that the left and right dependents of a coordinator never have direct syntactic links between them.

The next step is to generalize to the syntactic level. The verb *use* in example [26] (Word 12), which is the only obvious candidate for a possible link because of its unfilled valency slots, has both a direct object and an adjunct. Can the current word, *File*, play either of these roles? The BKB shows that it does appear (once) with an object, but never with an adjunct label. So, by rule 5, a very tentative extension can be made to the SL structure:

[INFC-C +use [OBJ? +File [DET -a!]]]

This extension has some semantic support from the fact that the pronoun *it* in example [26] has a referential identity link in the BKB with a subtree headed by *file*.

The link

[File [DET -a!]]

does not correspond to a translation unit (in example [27], *a VMS File* is translated as a single unit). So a compositional translation is called for. But the BKB has no translation available for the word *File* alone as a direct object. The only available translations appear with the attribute label ATR, and all (6) of them take the form of a prepositional phrase headed by *de*. This is an implausible extension of the Esperanto structure, because *de* (with 847 occurrences) never functions as a direct object in the BKB. This inconsistency throws doubt on the direct object link tentatively established above by rule 5. If the pre-attribute function proves more plausible, then the OBJ link will have to be broken. For the time being, *File* must remain untranslated.

Word 15: You can also copy a document to your SERVER DRAWER and use a File **Cabinet...**

This word appears 160 times, of which 133 times (83%) as the governor of the preattribute *File*, the broadest match being found in two identical examples of

[28] *Using the File Cabinet (FC)*

where *Cabinet* appears as the direct object of *using*, and this in turn can be matched, at the level of its basic form, with the input word *use*. On the strength of these examples, and given the earlier doubts about the function of *File* as object of *use*, the SL structure can be revised, using rule 6, as follows:

[INFC-C +use [OBJ +Cabinet [ATR -File [DET -a!]]]]

> *Procedural rule 6:* If the most probable link between the current word and the provisional SL structure conflicts with an earlier link, and if the evidence for the new link is stronger than that for the earlier one, then break the earlier link.

There is, however, an inconsistency in the above structure, and a higher-level type of pattern matching is required in order to detect it. While the lower-level pattern matching process looks for literal matches with words or morphemes in the BKB, a second process is required to check the plausibility of the structures being built. At this level, the question is not whether, for example, the article *a* occurs as a dependent of *File*, but whether a word with a DET label occurs as a dependent of a word with an ATR label. The answer, as far as the BKB is concerned, is no. Given the high frequencies of these labels, the failure to find even one example of this structure in the BKB is reason enough to call it implausible and to reconsider the previous steps (rule 7).

Procedural rule 7: An SL structure will be rejected if it proves implausible at the syntactic level, e.g. because a given pattern of syntactic categories or functions is not represented in the BKB.

The choice between breaking the ATR link or breaking the DET link is easily decided: the latter is supported by only one BKB example for the words concerned, whereas the former occurs 133 times. Since the article always has a governor to its right, the only alternative attachment is to *Cabinet*. This link is not supported by the BKB at the literal level: the pattern *a Cabinet* does not occur. The new attachment is acceptable, however, at the functional level (rule 7), since the pattern

[OBJ * [DET -a]]

is very common (265 occurrences). There are also 98 BKB examples of the pattern

[OBJ * [DET -a] [ATR -*]]

The revised structure thus becomes:

[INFC-C +use [OBJ +Cabinet [DET? -a!] [ATR -File]]]

Turning to the translation, we find that example [28] does not contain *File Cabinet* as a translation unit. Where the term appears in the BKB as a direct object, the commonest (40%) translation is

(TU15) [Cabinet [-File]] = [arkivo].

Using the default (null) translation for the article (TU 5), the output is now:

[povi [-vi] [+ankaŭ] [+kaj
[-kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]
[+uzi [+arkivo]]]]

Word 16: You can also copy a document to your SERVER DRAWER and use a File Cabinet **Menu...**

This next word occurs 200 times, and the broadest retrievable pattern is *File Cabinet Menu*, which matches 50 occurrences, with the structure

[Menu [ATR -Cabinet [ATR -File]]]

Again there are structural contradictions here with the previous analysis, and rules 6 and 7 will be invoked to revise the SL structure accordingly:

[INFC-C +use [OBJ +Menu [DET? -a!] [ATR -Cabinet [ATR -File]]]]

The *Cabinet Menu* link, with 50 occurrences, is much stronger than the *use Cabinet* link. The attachment of *Menu* to *use* is supported by one example in the BKB:

[29] *The Document Transfer (DT) option lets you use the Document Transfer Menu to copy marked documents...*

and the reattachment of the indefinite article to *Menu* is the only possibility for the time being, although it does not appear literally in the BKB.

As to the translations, the structures matched yield the following most frequent TUs:

(TU16) [Menu [-Cabinet [-File]]] = [menuo [+Arkivadministrado]].

(TU17) [use [+Menu]] = [uzi [+menuo]]

TU 15 can now be discarded, and TU 17 overlaps both 14 and 16 to produce:

[povi [-vi] [+ankaŭ] [+kaj
[-kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]
[+uzi [+menuo [+Arkivadministrado]]]]]

Word 17: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu **option...**

The most probable link from the word *option* to the preceding structure is represented by no less than 125 BKB examples of

[use [OBJ +option]]

(with the noun in the singular) which account for 50% of its 248 occurrences. This selection conflicts with the provisional attachment of *Menu* as the object of *use*, but as this had only one example to support it the new link prevails (by rule 6). (The verb *use* never has more than one object, on BKB evidence.) The word *Menu* now remains unattached, but the BKB has 4 examples where it is a dependent (always a pre-attribute) of *option*, which is therefore its probable governor. Once more shifting the indefinite article because of rule 7, we obtain:

[INFC-C +use [OBJ +option [DET -a!] [ATR -Menu [ATR -Cabinet [ATR -File]]]]]

The word *option* occurs twice with a dependent indefinite article. A search for *option* in the context of the revised structure turns up the following maximal match:

[30] *You can use File Cabinet Menu options to: ...*

in which *use File Cabinet Menu options* matches the new structure except for the noun plural.

This example provides the TU

(TU18) [use [+option [-Menu [-Cabinet [-File]]]]] =
[uzi [+opcio [-la] [+de [+menuo [-la] [+Arkivadministrado]]]]]

which supersedes TUs 16 and 17.

An interesting point here is that the translation derived from example [30] changes the indefinite to a definite noun phrase. Where the translation introduces potentially referential expressions such as this definite noun phrase, a mechanism is needed to check whether appropriate referents can be found in the context. In the present case, the lack of any referential expression on the SL side is a signal for caution. The definite article in Esperanto would only be justified, in these circumstances, if the *opcio de la menuo Arkivadministrado* were a unique entity. However, the BKB has three examples of this concept in the plural, so that the article *la* can be discarded as inappropriate. By way of contrast, the definite NP *la menuo Arkivadministrado* in TU 18 can be justified by the singularity of this concept in the BKB. The extended TL structure now becomes:

[povi [-vi] [+ankaŭ] [+kaj
[-kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]
[+uzi [+opcio [+de [+menuo [-la] [+Arkivadministrado]]]]]]]

Word 18: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option **to...**

The maximal match is provided by example [30] above, which extends the provisional structure to

[INFC-C +use [OBJ +option [DET -a!] [ATR -Menu [ATR -Cabinet [ATR -File]]]]
[PREA +to]]

The BKB interface shows this to be the commonest valency pattern for the verb *use*.

TU 18 can also be extended and replaced by

(TU19) [use [+option [-Menu [-Cabinet [-File]]]] [+to]] =
[uzi [+opcio [+de [+menuo [-la] [+Arkivadministrado]]]] [+por]]

and the output becomes

[+uzi [+opcio [+de [+menuo [-la] [+Arkivadministrado]]]] [+por]]

Word 19: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to **allow...**

The word *allow* occurs only three times in the whole BKB. The only word in the preceding input with which it has a literal BKB link is the preposition *to*, in the sentence

[31] *For example if you do not want to allow access to a document...*

Using this example, the provisional structure can be tentatively extended to

[INFC-C +use [OBJ +option [DET -a!] [ATR -Menu [ATR -Cabinet [ATR -File]]]]
[PREA +to [INFC +allow]]]

None of the BKB examples of *allow* by itself constitutes a translation unit. There is only one example which could still constitute a TU for *allow* in the present context, and this contains:

(TU20) [allow [+you]] = [ebligi]

Look-ahead shows that the word *you* does not appear anywhere to the right in the input string, so TU 20 is discarded and the conclusion must be that the word *allow* is not translatable in this context with the existing BKB. There is simply insufficient information about its behaviour.

At this point the system must request help from the SL operator, for example by asking for a synonym or paraphrase of *allow*. The most obvious synonym, *permit*, is also missing from the BKB. The user's next suggestion might be *let*, but this requires some (interactive) readjustment of the input sentence, because the second complement of *let*, unlike that of *allow*, cannot be headed by *to*. The revised input sentence becomes:

[32] *You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to let other users copy that document.*

There are 56 occurrences of *let* in the BKB, two of which match the pattern *to let*, and the provisional structure becomes:

[INFC-C +use [OBJ +option [DET -a!] [ATR -Menu [ATR -Cabinet [ATR -File]]]]
[PREA +to [INFC +let]]]

Unfortunately, neither of these examples translates *to let* or *let* as a unit. For the word on its own, the commonest translation is provided by:

(TU21) [let] = [ebligi]

With this adjustment, the output structure becomes:

[+uzi [+opcio [+de [+menuo [-la] [+Arkivadministrado]]]] [+por [+ebligi]]]

Word 20: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to let **other...**

The word *other* appears 40 times and rarely has a dependent. It has a 36/40 probability of being a pre-attribute. The 4 exceptions (and only they) are all followed by *than* and governed by a noun to the left. None of the nouns to the left in the input string has *other* as a post-attribute in the BKB. And a little look-ahead confirms that it is not followed by *than*. Consequently the pre-attribute function is virtually certain, and attachment must await further input.

There are 35 occurrences where the word constitutes an independent translation unit, namely

(TU22) [other] = [alia]

Word 21: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to let other **users...**

The most likely link for the current word is given by the pattern *other users* which appears 11 times, always with *other* in the role of pre-attribute. This construction is one which triggers a search for a suitable referent, because the word *other* is one which contrasts the concept it qualifies, with some other concept in the text. Using the referential links built into the BKB (see section 4.3 above), the most probable contrast is found to be with *you*, since this pronoun has an exclusion relation with *user* in 20 of its 22 referential links. On the TL side, too, there is similar evidence for the contrastive use of *alia uzanto* with *vi*. The referential link (not shown here) can therefore be added to the provisional structures.

The pattern *let other users* is not available, but *let user* occurs 9 times, with *user* as the direct object of the verb. These are the best available matches, and they justify extending the SL structure as follows:

[PREA +to [INFC +let [OBJ +user [ATR -other!]]]]

The translation unit selected by rule 1 is:

(TU23) [let [+user]] = [permesi [+al [+uzanto]]]

where the English object is replaced by an adjunct in Esperanto. This TU is represented by 9 BKB examples and replaces TU 21 by rule 3, changing the lexical equivalent of *let*.

There is only one available translation unit for *other users*:

(TU24) [user [-other]] = [uzanto [-alia]]

This overlaps TU 23 and now replaces TU 22, extending the output to:

[+uzi [+opcio [+de [+menuo [-la] [+Arkivadministrado]]]]
[+por [+permesi [+al [+uzanto [-alia]]]]]]

Although the gap between TUs 19 and 23 (*to ... let*) was bridged at the syntactic level in the SL, any such monolingual connection remains weak unless supported by semantic evidence. This principle is summarized in rule 8.

> *Procedural rule 8:* An output structure which is not literally represented in the BKB but has been built out of smaller units can be challenged if it proves implausible at the semantic level, i.e. if a semantic connection between the known context of the smaller units and the context provided by the new structure appears unlikely. A challenge at the semantic level requires choices at other levels to be reconsidered.

Looking at the known contexts of the Esperanto half of TU 23, we encounter one example,

[33] *Kiam vi kundividas dosieron vi povas permesi aliron al ĝi al aliaj uzantoj...*
which contains the structure

[povi [-vi] [+permesi [+al [+uzanto [-alia]]]]]

which covers both TU 23 and TU 24 and links them, not to TU 19, but to the beginning of the output structure. The question now is: Does this bridge to an earlier part of the TL structure imply that the link to TU 19 is semantically justified? The answer requires a little linguistic inference and presupposes the existence of a set of inference rules which may or may not be derived from the BKB.

In any construction such as *X can use Y to Z*, or, more precisely,

[can [SUBJ X] [INFC use [OBJ Y] [PREA to [INFC Z]]]]

the implication is that *X can Z (by means of Y)*. In other words,

[can [SUBJ X] [INFC use [OBJ Y] [PREA to [INFC Z]]]] ⇒
[can [SUBJ X] [INFC Z]]

This inference is supported by a strong correlation between the two constructions in the BKB. On the Esperanto side of the case in point, 9 out of 12 verbs which appear both in the pattern *uzi opcion por Y* and in the pattern *Y per X* ('to Y by means of X') have *X = opcio*. Skipping over the coordinator in the output structure justifies the conclusion that example [33] correlates well with the broader output structure at the semantic level and does therefore confirm the plausibility of the connection between TU 23 and TU 19.

Word 22: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to let other users **copy...**
We have already seen that the verb *let* expects an infinitival complement as well as a direct object. The pattern *to let copy* with *copy* as infinitival complement matches 2 BKB examples, as well as the existing SL structure, which thus becomes:

[PREA +to [INFC +let [OBJ +users [ATR -other!]] [INFC +copy]]]

There are two translation units corresponding to *let copy*, but neither of them fits the TL output so far by overlapping TU 23. This inconsistency will trigger a comparison of the probabilities involved, which seem to favour TU 23, with its 9 occurrences.[20] Consequently the next step is to look for an independent translation of *copy* as an infinitival complement. Four out of five cases favour the translation

(TU25) [copy] = [kopii]

and the output becomes

[+por [+permesi [+al [+uzanto [-alia]]] [+kopii]]]

Semantically, this coupling is supported (rule 8) by the fact that in 6 (67%) of the examples for TU 23, the verb *kopii* is a part of (a coordination in the role of) the infinitival complement.

---

[20] Comparison of the frequencies of different patterns should also take into account the frequencies of the words themselves and the respective probabilities of the observed combinations occurring by chance in a randomly constituted BKB. For example, on the basis of chance alone, the frequency of the *let user* pattern should be somewhat higher than that of the *let copy* pattern, simply because *user* occurs 126 times in the BKB, as against only 75 occurrences for the word *copy*.

Word 23: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to let other users copy **that...**

The word *that* appears 142 times, most commonly as a determiner with a governor to its right. No literal link can be found with any of the possible partners to its left, so rule 5 is applied. All the cases of *copy* in TU 25 expect a direct object, and this is the second commonest function of *that*. The SL structure can be extended to:

[PREA +to [INFC +let [OBJ +users [ATR -other!]] [INFC +copy [OBJ +that]]]]

The default TL interpretation on the grounds of frequency (semantically improbable, of course) is

(TU26) [that] = [ke]

where *that* is a conjunction heading a noun clause. This gives:

[+por [+permesi [+al [+uzanto [-alia]]] [+kopii [+ke]]]]

Word 24: You can also copy a document to your SERVER DRAWER and use a File Cabinet Menu option to let other users copy that **document.**

The broadest pattern which includes this last word is *to let copy that document* which at the level of the basic word forms matches the example

[34] *DATE – displays the Copy Date form to let you copy only those documents created or modified before a specific date.*

This example causes (by rule 3) a revision of the last extension to the SL structure:

[PREA +to [INFC +let [OBJ +users [ATR -other!]]
[INFC +copy [OBJ +document [DET -that]]]]]

The whole pattern *to let copy that document* does not constitute a TU, but *copy that document* does. Of three different translations, two match the TL structure so far, the choice being:

(TU27) [copy [+document [-that]]] = [kopii [+dokumento [-la]]]

which of course displaces TUs 25 and 26.

The fact that five examples of *that document* in the BKB have a referential link with another unit suggests a possible referential function for that expression. Intersecting the set of BKB referents for this expression with the units in the SL structure suggests only one candidate: *a document*. Given that the BKB supports a similar link on the TL side between *la dokumento* and *dokumento* (also 5 examples), a tentative identity link (not shown here) can be added to both output structures.

There remains the problem of confirming the link between TUs 23 and 27 on the Esperanto side. A search for a pattern which bridges this gap turns up 7 examples of the structure

[permesi [+al [+uzanto]] [* [kopii] [+dokumento]]]

('let user copy document'), where the asterisk stands for one or more coordinators. At the semantic level, coordinators can be by-passed, so that these 7 examples can be taken as strong evidence that the TL structure is semantically plausible.

Thus the final SL structure becomes:

[GOV can [SUBJ -you!] [ADVA +also!] [INFC +and
  [INFC-C -copy [OBJ +document [DET -a!]]
    [PREA +to [PARG +DRAWER [DET -your!] [ATR -SERVER!]]]]
  [INFC-C +use [OBJ +option [DET -a!] [ATR -Menu [ATR -Cabinet [ATR -File]]]]
    [PREA +to [INFC +let [OBJ +users [ATR -other! *(THAN 'you')*]]
      [INFC +copy [OBJ +document *(= 'a document')* [DET -that]]]]]]]]

and the final TL version, composed of TUs 3, (5), 7, 12-14, 19, 23-24 and 27, becomes:

[povi [-vi] [+ankaŭ] [+kaj
  [-kopii [+dokumento] [+en [+TIRKESTO [-via] [+SERVILO]]]]
  [+uzi [+opcio [+de [+menuo [-la] [+Arkivadministrado]]]]
    [+por [+permesi [+al [+uzanto [-alia *(OL 'vi')*]]]
      [+kopii [+dokumento *(= 'dokumento')* [-la]]]]]]]]

In the DLT design, a dialogue module consults the SL operator as to whether the system's interpretations of the input are correct (see section 5.4.2 above). Assuming the operator has opted for a dialogue at the end of each sentence (rather than each paragraph or whatever), the interpretation of the test sentence will at this point be put to the user. Now the whole translation process as simulated above has built up a bilingual text representation comprising syntactic links, functions and features, bilingual equivalences (translation units) and monolingual reference links. Once this representation has been approved by the operator it can be integrated with the BKB, where it will influence subsequent translations. In the meantime, however, the structure should be seen as a patchwork in which some pieces are clear and strong, while others are weaker. In the above account of a BKB-based translation, I repeatedly referred to the 'strength' of various links and often used such words as 'tentative'. The process of translation can be imagined as one in which a structure is built up bit by bit, with various modules or processes going to work on it in parallel or in alternation, each attaching a probability estimate to the elements it adds or else changing the probabilities associated with elements added earlier. When automatic processing is finished, then, it should be easy enough to identify the weakest points in the structure. They are those to which the lowest probabilities are attached or for which the BKB evidence was least abundant. These are the points – whether they are syntactic links, or lexical choices, or reference links – which call for confirmation (or rejection) by the user.

Once the interpretation has been approved, morphological generation rules and tree-to-string rules can convert the tree representation shown above, including the hidden syntactic functions and features such as number, to a string:

[35] *Vi povas ankaŭ kopii dokumenton en vian TIRKESTOn SERVILO kaj uzi opcion de la menuo Arkivadministrado por permesi al aliaj uzantoj kopii la dokumenton.*

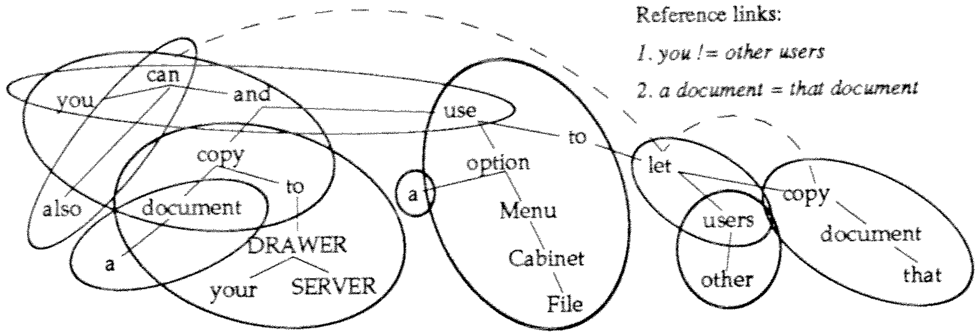
For comparison, the original Esperanto version of this sentence read:

[36] *Vi povas ankaŭ kopii dokumenton en vian tirkeston SERVILO kaj, per opcio de la menuo Arkivadministrado, permesi al aliaj uzantoj kopii ĝin.*

The two versions are essentially equivalent in content, though different in structure. The simulated translation is more literal than the human version, but this is not a necessary consequence of the method. A BKB-based translation can perfectly well contain structural transformations and more idiomatic translations, provided these are available in the BKB and are matched by the input.

Figure 4 shows the SL structure arrived at, including translation units and referential links. Dotted lines indicate the semantic links examined, although these are not part of the formal structure such as it would later be added to the user's BKB.

Fig. 4: SL structure for the test sentence, with TUs and semantic links



Three points of special interest thrown up by this simple experiment are the treatment of unknown or untranslatable words via a dialogue with the user (Word 19), the need to check for semantic coherence between translation units (Words 14, 21 and 24), and the selection of appropriate deictic or other referential forms in the target language (Words 17 and 24). The fact that a common verb such as *allow* is only inadequately represented in this small corpus underlines the need for much larger corpora as a basis for a BKB. Of course, a domain-specific corpus such as that chosen for the model implementation can always be backed up by a general-purpose corpus to provide better coverage of the general vocabulary.

5.5.2 Esperanto to French translation

Having produced a plausible Esperanto translation [35] of test sentence [21], the obvious next step is to try to complete the double translation by going from Esperanto to French, using the sister BKB of the one used for English to Esperanto translation. I will not describe this process in detail, as the principles involved have been adequately illustrated in the previous section.

The French version ultimately obtained by an analogous process of simulated translation is:

[37] *Vous pouvez aussi copier un document dans le TIROIR SERVEUR et utiliser une option du menu Gestion des archives pour autoriser la copie du document à d'autres utilisateurs.*

This BKB-based version can now be compared with the human translation of the original English sentence, which was

[38] *Vous pouvez également copier un document dans votre TIROIR SERVEUR et à l'aide d'une option du menu Gestion des archives autoriser d'autres utilisateurs à le copier.*

Apart from several minor differences, there are two major structural points of interest. First, the simulated version sticks more literally to the pattern of the English *use an option to...* The result is a potentially tedious repetition of the root *utilis-*. Second, the simulation has produced a striking departure from the human version by applying an alternative valency pattern with the verb *autoriser*, which in turn forces the nominalization of *copier* to *la copie (du document)*. However, the output appears to be correct and readable, and the (double) translation, despite the very limited size of the BKB, essentially successful.

6 SUMMARY OF ADVANTAGES OF A BILINGUAL KNOWLEDGE BANK

The advantages the BKB concept 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.

(2) The **translation expertise** needed in a machine translation system can be acquired by "digesting" the work of 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 or she intuitively applies. 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 a machine translation system. They do not need to be formulated explicitly. It is no longer necessary to rely on such often inadequate sources as conventional dictionaries and grammars.

(3) **Extra-linguistic knowledge** can be acquired from ordinary (informative) text input. The BKB structure is sufficiently unambiguous to allow the application of basic inferencing procedures. The BKB, consisting as it does of translation units, is necessarily language pair-specific. This is not to say, however, that the extra-linguistic 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.

(4) 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).

(5) 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**.

(6) In view of the considerable storage requirements for a corpus-based knowledge bank, **compaction** is obviously important. Large-scale compaction – much better than that provided by conventional string-based compression techniques – can be achieved by coding translation units. In the sample text in section 3.4, for example, a repeated term such as *the shutoff switch of the right-hand outer wing tank* need only be stored once in its literal form.

7 COMPARISON WITH OTHER RECENT RESEARCH

Probably the closest approach to the BKB concept already implemented elsewhere is the linguistic database at ATR (Sumita *et al.* forthc.). This consists of a bilingual English/Japanese corpus of some 100,000 words with syntactic structure superimposed and with equivalent expressions in the two languages cross-coded. It is not clear whether all translation units are coded, as in the BKB design, or only word-for-word equivalences. The third BKB dimension, that of referential links, is apparently not included. The ATR group have experimented with corpus-based translation of Japanese noun phrases into English. They are more inclined to regard the corpus as a supplementary knowledge source, and to use what they call Example-Based Machine Translation (EBMT) as a complementary tool to Rule-Based MT (RBMT), whereas I have proposed the BKB as an all-purpose knowledge source which can replace rule systems at all but the highest levels. Following Nagao's (1984) proposal they use a thesaurus to check on the similarity of the content words to be translated to those in the example sentences.

Sato and Nagao (forthc.) have attempted to construct a general model of example-based translation, which they regard as "the new wave of machine translation". However, their study concentrates on modelling the translation process, using a small number of example sentences as the knowledge source, rather than building a large-scale knowledge bank. Although elsewhere (Sato and Nagao 1989) they refer to the possibility of measuring semantic proximity by comparing contextual patterns (as in the DLT prototype), they still consider the construction of large thesauri for this purpose both necessary and problematic.

Neither of these groups, however, appears to make use of referential information in the database to detect semantic relations, nor do they envisage a corpus-based approach to source text analysis, as suggested above.

LITERATURE

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

Bennett, W.S. / J. Slocum (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. / J. Cocke / S. Della Pietra / V. Della Pietra / F. Jelinek / R. Mercer / P. Roossin (1988): A statistical approach to language translation.
In: Coling '88, pp. 71-76.

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

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

Coling '86: *11th International Conference on Computational Linguistics. Proceedings of Coling '86.* Bonn: Institut für angewandte Kommunikations- und Sprachforschung.

Coling '88: *12th International Conference on Computational Linguistics. Proceedings of Coling '88.* Budapest: John von Neumann Society for Computing Sciences.

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, pp. 42-45.

Harris, Brian (1988a): Bi-text, a new concept in translation theory. *Language Monthly* No. 54, pp. 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: Maxwell, D. / K. Schubert / T. Witkam (eds.) (1988), pp. 7-64.

Kjærsgaard, Poul Søren (1987): REFTEX – A context-based translation aid.
In: *Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, 1-3 Apr. 1987, pp. 109-112.

Kjærsgaard, Poul Søren (1989): REFTEX – un progiciel pour la traduction assistée par ordinateur. *Meta* **34**, 3, pp. 496-501.

Maxwell, D. / K. Schubert / T. Witkam (eds.) (1988): *New Directions in Machine Translation*. Dordrecht/Providence: Foris. Distributed Language Translation 4.

Melby, A.K. (1988): Lexical Transfer: Between a Source Rock and a Hard Target.
In: Coling '88, pp. 411-413.

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: A. Elithorn / R. Banerji (eds.): *Artificial and human intelligence*. Elsevier, pp. 173-180.

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

Piron, C. (1988): Learning from Translation Mistakes.
In: Maxwell, D. / K. Schubert / T. Witkam (eds.) (1988), pp. 233-242.

Sadler, V. (1989a): *The Bilingual Knowledge Bank, a new conceptual basis for MT*. Utrecht: BSO/Research. DLT report.

Sadler, V. (1989b): *Working with analogical semantics: Disambiguation techniques in DLT*. Dordrecht/Providence: Foris. Distributed Language Translation 5.

Sadler, V. (forthc.) *A corpus-based measure of semantic proximity*.
In: [Proceedings of the Maastricht-Lodz Colloquium on "Translation and Meaning", Maastricht, 4-6 Jan. 1990]

Sadler, Victor / Ronald Vendelmans (forthc.): Pilot implementation of a Bilingual Knowledge Bank.
In: [Proceedings of the Coling conference, Helsinki 1990]

Sato, Satoshi / Makoto Nagao (1989): *Memory-based Translation*. Reprint of WGNL70-9, IPSJ (in Japanese).

Sato, Satoshi / Makoto Nagao (forthc.): Toward Memory-based Translation.
In: [Proceedings of the Coling conference, Helsinki 1990]

Schubert, K. (1986): Linguistic and extra-linguistic knowledge. *Computers and Translation* **1**, 3, pp. 125-152.

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

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

Sumita, E. / Y. Tsutsumi (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.

Sumita, Eiichiro / Hitoshi Iida / Hideo Kohyama (forthc.): Translating with Examples: A New Approach to Machine Translation.
In: [Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, June 1990.]

Tsujii, J. (1986): Future directions of machine translation.
In: Coling '86, pp. 655-668.

Tsujii, J. (1988): What Is a Cross-Linguistically Valid Interpretation of Discourse?
In: Maxwell, D. / K. Schubert / T. Witkam (eds.) (1988), pp. 157-166.

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

Witkam, Toon (1988): *DLT – an industrial R&D project for multilingual MT.*
In: Coling '88, pp. 756-759.

Zuijlen, J. van (1989a): *A comprehensive parser for DLT.* Utrecht: BSO/Research. DLT report.

Zuijlen, J. van (1989b): Probabilistic methods in dependency grammar parsing.
In: *Proceedings of the International Workshop on Parsing Technologies, Carnegie Mellon University, August 1989.* Pittsburgh: CMU, pp. 142-151.