17
DLT

17.1 Background

Distributed Language Translation (DLT) is the name of an MT project at the Utrecht software company Buro voor Systemontwikkeling (BSO). Preliminary investigations by A.P.M. (Toon) Witkam began in 1979, a feasibility study was supported during 1982–83 by a grant from the Commission of the European Communities, and in 1985 the project received a six-year contract from the Netherlands Ministry of Economic Affairs, with the initial task of producing a prototype English to French system in 1987 and a commercial version in 1993. The long-term aim of the DLT project is to build a system for translating between European languages of the Community (French, German, English, Italian) with eventual extensions to other languages. The prototype, written in Prolog, is designed to translate from a restricted form of English (Simplified English) into French, and it was demonstrated on schedule in December 1987. Since that date, the DLT group has been working towards a commercial system based on somewhat different principles than those in the prototype.

DLT is intended as an interactive multilingual system for use on personal computers in data communication networks, not as a tool for translators but primarily as a tool for monolingual users in the interlingual communication of 'informative' literature (abstracts, reports, manuals) or commercial messages. Translation is 'distributed' in the sense that processes of analysis and generation take place at different terminals. A monolingual user will enter a text (in English, for example) at one terminal where it is immediately translated into an interlingua, (or 'intermediate language'), which is based on Esperanto. Analysis and translation

take place in 'real time'; the system attempts to translate text as it is entered, whether sentences are complete or not. Problems which the system itself cannot resolve are submitted in interactive dialogue to the user, who does not need to know the interlingua or any target language or even to know which languages the text is going to be translated into. In some cases, monolingual interactions may lead to rephrasings of original texts in order to simplify and remove translation problems. The text (in the interlingua) is then transmitted to another terminal where another user initiates translation from the interlingua into the language required (e.g. French). No dialogue with the system is possible by the recipient, and no post-editing of texts is expected.

DLT is a modular system in that source and target language modules can in principle be created for analysis into and generation from the interlingua without affecting existing modules. It is intended to be easily extendible to any other languages, whether typologically similar or not. The requirements presuppose a fully expressive interlingua which is clear and unambiguous enough to enable fully automatic translation into any target language.

17.2 The interlingua

In most interlingua-based MT systems, intermediate representations are not genuinely interlingual. Usually the structural representation is language-independent, e.g. a predicate–argument structure, but lexical items are not. In most cases, lexical transfer is based on bilingual dictionaries; there are no interlingual (language-independent) lexical items. In DLT, by contrast, the Esperanto interlingua is more like a 'natural language' with its own independent structures and lexical items.

It is usually argued that an interlingua has to be more explicit than any individual human language; it must be capable of representing all linguistic relationships (including implicit relations) within texts and within the language system. It is assumed that this means (quasi) logical representations involving semantic primitives which are universally valid (see section 6.7). The DLT researchers argue that this assumption is mistaken. Analysis into primitives, even if practically feasible, would lead to 'unimaginably huge dictionaries and a never-ending, but largely useless disambiguating process'. They contend that no artificial language can be more explicit than the human language(s) on which it is based and with which it is defined; it cannot go beyond the capabilities of human language. Thus, while an artificial language can always be translated into a human language, they argue that it is not always possible to translate fully from a human language into an artificial one. Thus, in the opinion of the DLT researchers, an interlingua can be as explicit and expressive as a human language only if it has the character of a 'human language'.

For its interlingua DLT adopted Esperanto, which it is argued combines the expressiveness of 'natural' languages and the desired regularity and consistency. Esperanto is an 'a posteriori' planned language, taking elements from existing languages and arranging them in its own (autonomous) way. But it is a language with its own speech community, it has developed into a genuine 'natural' language, which can be said to be no more 'artificial' than standardised, normalised, official

languages. It has had a 'test phase' of over a hundred years, acquiring the expressivity of a human language not from the designer's drawingboard but from actual usage.

The advantages of Esperanto as an interlingua are, therefore, claimed to be: (a) as a (semi) 'natural' language it has an expressiveness, richness and flexibility which surpass constructed logical interlinguas, (b) it provides a ready-made standardised vocabulary based on common Indo-European roots, (c) it is regular and consistent, (d) it is autonomous and independent of other languages, and (e) it can be learned and understood like any other human language. Learnability is of practical importance for an interlingua, since it has to be used consistently by those working on the system, not only consistently with each other but also consistently as individuals. This consistency is difficult to achieve with an artificially constructed vocabulary of primitives and abstract structure rules. The 'naturalness' of interlingua representations (as linear strings) means also that developers of the system can more easily check whether analyses and interpretations are performing as expected; more easily than checking complex formal–linguistic tree-structure representations with multiple labels and features.

Esperanto does have some disadvantages as an MT interlingua. In use for over a hundred years, it has acquired 'naturally' some homographs, structural imprecisions and lexical ambiguities. It has no standard procedures for creating new terminology, and there are known weaknesses in technical and scientific vocabulary. DLT has sought to make modifications as necessary and to expand the lexical base.

17.3 System design

Since the interlingua (modified Esperanto) is not an abstract representation but a regularised language, the analysis of source texts and the generation of target texts represent in effect two 'translation systems': from source to Esperanto and from Esperanto to target. As a result, the DLT system may be regarded as a network of bilingual MT systems with modified Esperanto at its centre.

In the DLT system developed until 1988 for the prototype, these two halves were not fully independent bilingual translation systems; as we shall see, all semantic and pragmatic processing takes place in the kernel interlingual component whether translation is to the interlingua or from the interlingua. Since 1988, the project has been working on a different conception (involving the Bilingual Knowledge Bank) where the two halves operate in the same way.

In this section and in the following descriptions of syntactic and semantic processing (sections 17.4 to 17.6) we shall outline the basic model of the prototype, which was developed for translation of manuals from English (in fact 'Simplified English') into French. The later Bilingual Knowledge Bank approach proposed for the commercial system represents in effect a new system and for this reason will be described separately in section 18.2. This chapter is therefore devoted to the original Esperanto-based model.

The basic processing stages of the prototype English–French system are as follows:

1. Source language parsing. The parser, an Augmented Transition Network (ATN), recognises English words, their morphological and syntactic features, identifies dependency relations (subject, object, attribute, etc.) and produces a dependency tree, delivering alternatives where there is syntactic ambiguity. No semantics are involved; it generates all possible analyses regardless of semantic plausibility or statistical probability.
2. Monolingual (source language) tree transformations. At this stage, monolingual variants are reduced to common forms, e.g. *can not*, *cannot* and *can't* all become *can not*; and auxiliary–verb constructions are reduced to single verbs with labelled features, e.g. *has been eaten* becomes *eat* [present perfect, passive].
3. Bilingual tree transformations. This is the task of the Metataxor (section 17.4), which replaces English words by Esperanto equivalents and replaces English syntactic dependency labels by Esperanto ones. It may entail rearrangements of the tree and the insertion of function words (as explicit indicators of relationships). Because there are usually several translation alternatives for each single English item (i.e. because of the lexical ambiguities of English and the lexical transfer ambiguities of English to Esperanto), there will be a large number of alternative Esperanto trees. As in the first stage, no semantic or pragmatic selection is performed; all possibilities are produced.
4. Semantic–pragmatic word choice. From the alternative interlingual trees presented, the most likely in the given context is selected on the basis of Esperanto word patterns encoded in the 'Lexical Knowledge Bank' (LKB). This is the operation of the SWESIL component (Semantic Word Expert System for the Intermediate Language), which is described in 17.5 below. The result is a plausibility ranking of alternative interpretations.
5. Disambiguation dialogue. If no clear preference can be determined by SWESIL, the problems of interpretation are presented to the operator (normally the original author) in an interactive computer-initiated dialogue. For this dialogue the fragments of interlingua representations requiring disambiguation are expressed in the source language. After this stage there should be only a single interlingual representation of the input sentence.
6. Monolingual (interlingua) tree transformations. This involves the regularisation and determination of the morphological features for 'correct' interlingua representations, including government and agreement indicators.
7. Tree linearisation of the interlingua. The transformation of interlingua tree into the linear form of Esperanto involves the determination of the correct word order and the removal of labels and feature lists. The result is a plain Esperanto text which can be read by humans.
8. Correctness check. For security, each sentence is passed through a parser to check for syntactic well-formedness; any rejected sentences are sent back to step 6.
9. Coding and network transmission. Accepted sentences are converted for electronic transmission.
10. Decoding. Esperanto text is received at another terminal.

11. Esperanto parser. The decoded string is transformed into a dependency tree; because input is relatively unambiguous this process is fast.
12. Monolingual (interlingua) tree transformations. This stage mirrors stage 2, this time for Esperanto.
13. Bilingual tree transformations. The Metataxor generates from the single Esperanto tree a set of alternative French dependency trees; e.g. lexical transfer may indicate more than one French equivalent for a single Esperanto word.
14. Semantic–pragmatic word choice. In this stage, the correct French words are selected on the basis of word pattern information in the LKB (i.e. based on interlingua information), as illustrated in section 17.6 below. As there can be no interaction with users (they are passive receivers of the information with no knowledge of the source language), selection must be fully automatic.
15. Monolingual (target) tree transformations. This involves adjustments of incorrect French dependency trees and the insertion of government and agreement relations and features.
16. Tree linearisation of the target language. The French tree is linearised, with any necessary contractions and elisions, e.g. *de le* → *du*, *je ai* → *j'ai*.

The two halves of the translation process are distinct but not congruent 'translation systems': the conversion of source text into interlingua text (stages 1 to 8) is not the same as the conversion of interlingua text into target text (stages 11 to 16). As far as morphological and syntactic aspects are concerned, there are close parallels: (a) dependency parsing (1 and 11), (b) source tree transformations (2 and 12), (c) bilingual tree transformations (3 and 13), (d) target tree transformations (6 and 15), and (e) tree linearisation (7 and 16). The major differences lie in the stages of semantic–pragmatic interpretation and disambiguation; no processes are carried out in the source and target languages but all are performed in the Esperanto kernel of the system. In the first half, the heavy load of semantic processing is performed at the target language end (stage 4); in the second half, the processing takes place at the source language end (stage 14). As a result the modules which are specific to source and target languages can concentrate on language-specific manipulations of morphological and syntactic forms and structures. All the content (meaning) analysis and transfer takes place in the interlingua component of the overall system.

The basic processes, the modules and the data used at each stage may be illustrated schematically as in Figure 17.1, where SL is 'source language', IL is 'intermediate language' or 'interlingua', and TL is 'target language'.

It is evident from the description and from the diagram that simplification could be introduced by omitting the stages of generating the Esperanto string, its coding and decoding and the reanalysis as a tree (stages 7 to 12); and it is in fact accepted as a variant of the basic design that intermediate representations could be distributed over the network as Esperanto trees. However, this option would have the disadvantage that developers of the system would not have easy access to Esperanto texts in order to check the validity of interlingual interpretations, which, as we have seen, was claimed to be one of the primary benefits of a 'natural' language interlingua.

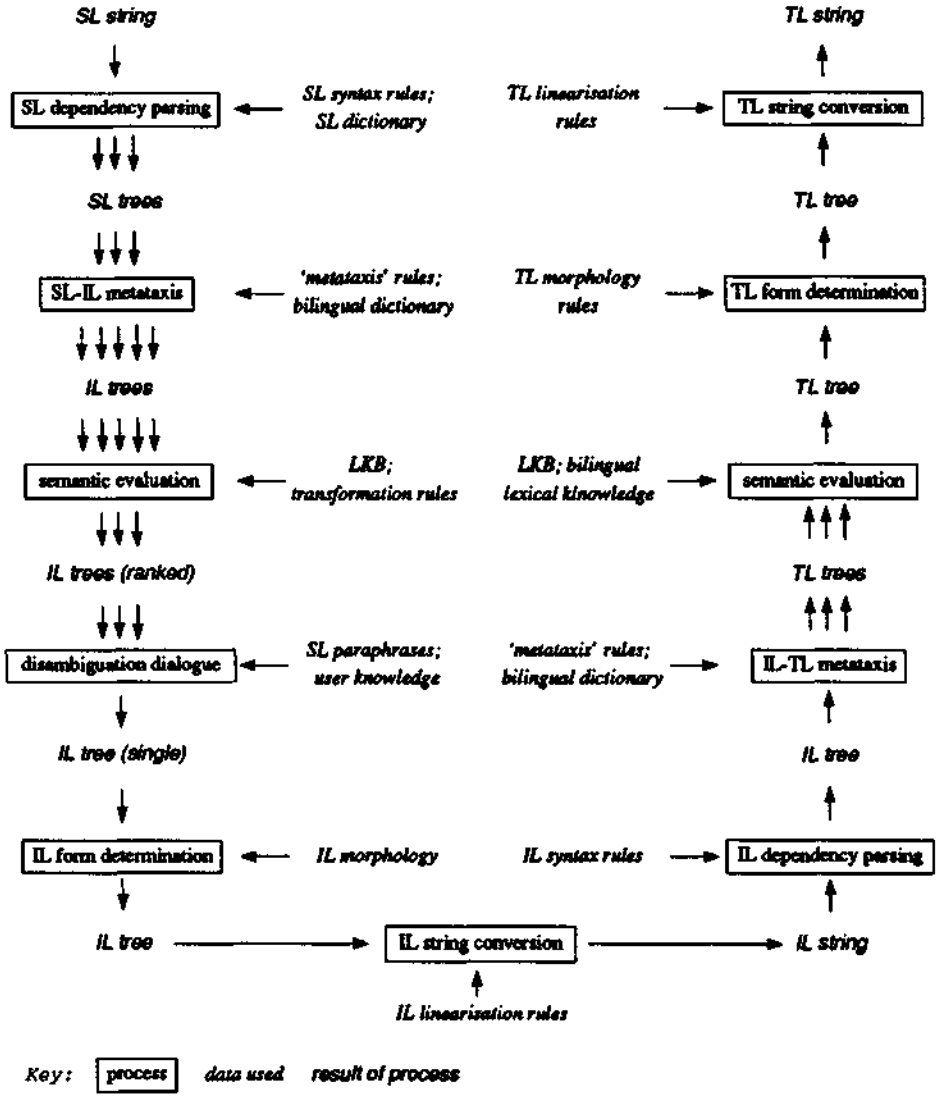


Figure 17.1 DLT basic processes

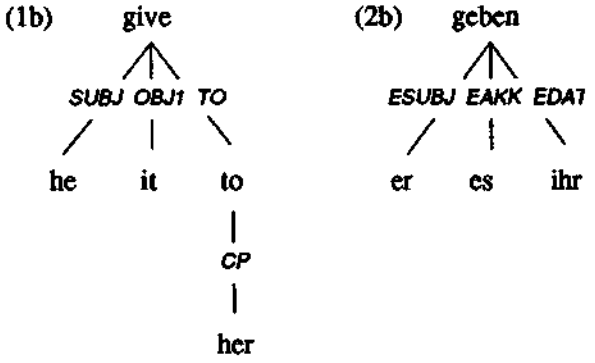
17.4 Dependency parsing

The formalism used in DLT for the representation of structures at all stages is based on dependency syntax (section 2.8.1) The basic arguments for using dependency trees are that structural relations (subject, object, attribute, etc.) are more central to analysis and transfer than constituency structures (noun phrase, prepositional phrase) and that dependency analyses are more suitable for languages with 'free' word orders than most constituency analyses. In DLT, the syntax of each language is developed with no reference to meanings (i.e. purely on distributional grounds) and quite independent of all other languages in the system. Thus, even for

similar structures in closely related languages there are differences of structure and labelling; compare the representations of English (1) and German (2).

(1a) He gives it to her.

(2a) *Er gibt es ihr.*



The DLT parser is based on a version of the ATN parser. In the analysis of source strings, the parser operates on information located in the 'syntactic dictionary' entries of the source language lexical items. These entries are in uninflected root forms; therefore, as in most MT systems morphological analysis precedes syntactic (dependency) analysis. Multiple parses are passed to the next stage of structural transfer.

17.5 Metataxis

Metataxis is the name given in DLT for the rule systems which link the dependency syntaxes of two languages, and the mechanism which transforms structures is called the 'Metataxor'. Rule systems are specific to one pair of languages and in one translation direction only. Thus, for the English to French prototype two metataxis rule systems have been developed: from English to Esperanto and from Esperanto to French. (Outlines of metataxis systems have also been described for other languages, including German, Danish, Polish, Bangla, Finnish, Hungarian and Japanese, and several others not published.)

The 'Metataxor' transforms dependency trees of source texts into dependency trees of target texts. Rules are effective at all levels: word, sentence and text. Metataxis rules may (a) change the syntactic category of a word, (b) change its morphological form, (c) change its syntactic function, (d) change a configuration of dependency relations, (e) add or remove words or items, and (f) merge or split dependency trees. Constraints on metataxis are imposed by the well-formedness conditions of the syntaxes of the source and target languages.

Metataxis operates on both lexical items and on trees, in two closely interlinked processes: the replacement of source words by one or more target language equivalents taken from bilingual dictionary, and the formation of tentative, syntactically correct target trees. Choices of different equivalents trigger changes

in tree structures, and some tree transformations necessitate changes in lexical items. It follows that conditions on lexical items are also formulated as tree fragments, e.g. (3) for English.

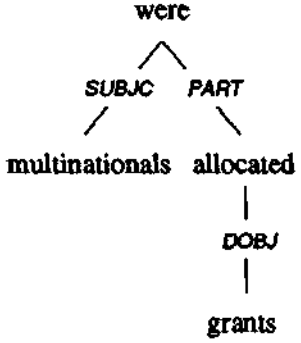
(3) sell — to — [ ]

Metataxis is not concerned with resolving ambiguities: if the bilingual dictionary contains more than one translation, metataxis will produce more than one structural representation. The objective of metataxis is to ensure that the representations are syntactically valid. Choice between alternative representations is made on semantic and pragmatic grounds by the 'expert system' SWESIL (section 17.6 below).

To illustrate metataxis from English into Esperanto (stage 3), consider the translation of (4a), with the dependency parse (4b), into the Esperanto (5), where the English subject has been transformed into a prepositional complement (with the preposition *al* 'to') and the English direct object into a subject. The theme–rheme word order can be preserved because the nominative case of *subvencioj* indicates it is a subject.

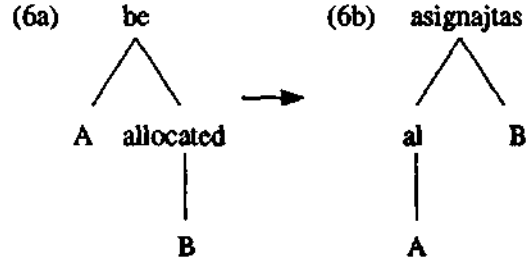
(4a) Multinationals were allocated grants.

(4b)



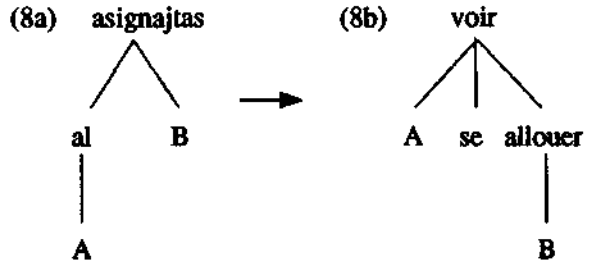
(5) *Al multnaciaj entreprenoj asignajtis subvencioj.*

The relevant metataxis rule is shown in (6), transforming the English passive subtree (6a) into the Esperanto tree (6b).



A similar metataxis rule would apply in the second half of the prototype (stage 13) to convert Esperanto trees into French trees. For example, to produce the French equivalent of (4a) and (5), namely (7), the rule might be as in (8).

(7) *Les multinationales se sont vu allouer des subventions.*



Such rules are clearly specific to particular structure types and particular sets of lexical items. In addition, the Metataxor has available general default rules in order to ensure that everything gets converted if no specific rules have applied. One would be that a source subject dependency becomes a target subject dependency.

17.6 Interlingual data and SWESIL

Originally conceived simply as a means of compact intermediate representation, Esperanto acquired greater importance as the project developed. By concentrating routines for semantic interpretation in the intermediate stages of the system, so that they do not have to be repeated for every additional language, the interlingua represents a linguistic 'knowledge bank' which is referred to in semantic and pragmatic processing. In the prototype system there were two databases: the LKB of Esperanto texts and the bilingual Esperanto–French dictionary. The LKB comprised pairs of 'content' words linked by a function or relator (e.g. Table 17.1), extracted from a dependency analysis of a 500,000-word corpus of Esperanto texts. The frequency of each pair was, however, not recorded, for reasons of computational simplicity. A typical example is given in Table 17.1 for the Esperanto word *ĉambro* ('room'), where *a* is the attribute relator.

ĉambro a apuda	'adjoining room'
ĉambro a bela	'beautiful room'
ĉambro a granda	'large room'
ĉambro a hela	'light room'
ĉambro a komforta	'comfortable room'
ĉambro a komuna	'common room'
ĉambro a nuda	'bare room'
etc.	

Table 17.1 Word pairs for *ĉambro*

In the bilingual dictionary the pairing of Esperanto and French lexical items was accompanied by contextual clues indicating the circumstances in which choices were to be made during lexical transfer. These clues were also given in Esperanto, e.g. Table 17.2 for the word *akra*.

akra a (doloro, malvarmo, vortoj, ...)	→	*vif*
akra a (nazo, oreloj, turo)	→	*pointu*
akra a (spico, pipro, brando)	→	*fort*
akra a (disputo, batalo, krizo, ...)	→	*violent*
akra a (ironio)	→	*mordant*

Table 17.2 Contextual clues for bilingual pairings

The table shows that Esperanto *akra* should be translated as *vif* if it is an attribute of *doloro* ('pain'), *malvarmo* ('cold'), *vortoj* ('words'); as *pointu* if an attribute of 'nose', 'ears' or 'tower'; as *fort* in the context of 'spice', 'pepper' and 'brandy'; etc.

The two databases have been the sources for the 'expert system' SWESIL at two stages of semantic processing: during translation from English into the interlingua and during translation of the interlingua representation into French.

17.6.1 English to Esperanto semantic processing

Tables such as 17.1 have been employed in the prototype for checking the acceptability of Esperanto lexical relationships. Checks (obviously performed on root forms) would confirm, for example, that *hela* is the correct translation of English *light* when modifying *ĉambro* and that *malpeza* ('light in weight') is incorrect. The problem comes, of course, when the database (the LKB) does not contain the sought pattern.

In the initial phase of the prototype design new and unknown combinations were dealt with by reference to hierarchies of words (i.e. in a thesaurus structure). For example, English *wood* can be either *ligno* ('wood as material') or *arbaro* ('wood as group of trees'); in translating (9) the Metataxor (stage 3; section 17.5) would present both as possibilities (10).

(9) The wood has been ordered.

(10a) *La ligno menditas.*

(10b) *La arbaro menditas.*

On consulting the hierarchies for these two possibilities, SWESIL would find *materialo* and *vegetaĵo* as their respective superordinates. It would then find that 'order' (Esperanto *mendi*) can have 'material' objects but not 'vegetable' ones. Thus *wood* must be translated as *ligno*.

However, the building of large and consistent hierarchies of Esperanto words proved to be extremely time-consuming and yet still insufficient for some types of contextual analysis; e.g. *eat* and *cook* are related not hierarchically but procedurally

(as in an AI-type 'script' of event sequences). It was decided, therefore, to check the plausibilities of unknown patterns on the basis of word-pair comparisons alone. (In all cases, SWESIL worked with word pairs rather than triples or larger groups — again for reasons of computational simplicity.)

For every pair of words the comparison is done in two stages, an examination of the known contexts of the first word and an examination the known contexts of the second word. The process may again be illustrated with sentence (9). From the two possibilities generated by the Metataxor (10a) and (10b) are derived two standardised pairs of content words plus relator (11), where *n* indicates an 'object' relation.

(11a) *mendi n ligno* ('to order wood')

(11b) *mendi n arbaro* ('to order a wood')

SWESIL first calculates the 'semantic proximities' of *ligno* and *arbaro* to each of the words given in the database as objects of *mendi*. The semantic proximity of two words is a simple calculation of the number of their identical word-pair contexts as a proportion of the total word-pairs in which the two occur. Thus *nigra* ('black') occurs in 50 word-pairs, *blanka* ('white') in 51 word-pairs, and on 7 occasions the word context is the same ('hair', 'skin', 'colour', etc.); their semantic proximity is calculated as 0.829; by contrast, *lando* ('country') occurring in 346 word-pairs and *filo* ('son') in 109 word-pairs have only 7 in common, and a lower proximity score of 0.202.

In the database (LKB) SWESIL would find the following words occurring as objects of *mendi*: *armilo* ('weapon'), *automobilo* ('car'), *bileto* ('ticket'), *glaso* ('glass'), *materialo* ('material'), *servo* ('service'), *telefono* ('telephone'), *varo* ('ware'). The proximity scores for each of these and *ligno* or *arbaro* are given in Table 17.3 (next page).

The results show that the best available analogy for *ligno* is *materialo* (0.752) and for *arbaro* it is *varo* (0.137); and that, therefore, the clear preference is for the interpretation in (10a) and (11a), with *ligno*.

In the second stage, SWESIL calculates the plausibility of *mendi* ('order') as an action applied to both *ligno* and *arbaro*. For the *ligno* ('material') the LKB lists as actions: *bori* ('drill'), *eksporti* ('export'), *konsumi* ('consume'), *poluri* ('polish'), *rompi* ('break'), *vendi* ('sell'), etc.; for *arbaro* it lists: *ataki* ('attack') and *planti* ('plant'). The highest values for semantic proximities were *mendi* and *vendi* (0.817) and *mendi* and *konsumi* (0.777) — both found with *ligno* as object; whereas the highest proximity scores for *mendi* with *ataki* and *planti* are 0.379 and −0.078 respectively. The clear preference is again for the 'material' translation (*ligno*) of *wood*.

The same procedure is applied to relational words. For example *with* can be translated in (at least) two ways: *kun* (association) and *per* (instrument). Either are possible with the verb *look at* (12).

(12a) to look at with love

(12b) to look at with a microscope

The Metataxor generates two alternatives, *kun* and *per*, in both cases. SWESIL would calculate the word-pair match scores as in Table 17.4. and therefore accept (13) as preferred translations.

ligno	*materialo*	→	0.752
ligno	*armilo*	→	0.542
ligno	*bileto*	→	0.537
ligno	*varo*	→	0.526
ligno	*automobilo*	→	0.495
ligno	*servo*	→	0.242
arbaro	*varo*	→	0.137
arbaro	*materialo*	→	-0.047
arbaro	*armilo*	→	-0.089
arbaro	*glaso*	→	-0.096
arbaro	*telefono*	→	-0.154
arbaro	*bileto*	→	-0.181
arbaro	*automobilo*	→	-0.210
arbaro	*servo*	→	-0.249
ligno	*glaso*	→	-0.290
ligno	*telefono*	→	-0.411

Table 17.3 Proximity scores for *ligno* and *arbaro*

rigardi	*amo*	*mikroskopo*
kun	0.570	-0.116
per	0.015	0.732

Table 17.4 Word-pair match scores

(13a) *rigardi kun amo*

(13b) *rigardi per mikroskopo*

Not all source language (English) ambiguities can be solved by SWESIL and so the DLT prototype includes a stage of 'Dialogue disambiguation' (stage 5). Essentially this involves the presentation to the user of translation choices in English. For example, SWESIL might not have been able to decide between *devii disde* and *foriri de* as translations of *depart from*. It would consequently select two paraphrases for the user to choose between: *deviate from* and *leave*, respectively. The selection of appropriate paraphrases remains, however, a major difficulty (as it is in any MT system with interactive disambiguation, section 8.3.3) and is not fully solved in the prototype.

17.6.2 Esperanto to French semantic processing

Whereas in the first half of translation (from source to interlingua, stage 4), disambiguation involves only monolingual knowledge of word-pair matches in the interlingua LKB, in the second half (from interlingua to target, stage 13) disambiguation involves bilingual knowledge and a different type of matching

procedure is applied. Lexical choice in the target language (French) is determined, not by the monolingual plausibility of the possible combinations of the French words but by matching contextual clues against the input (Esperanto) context.

We illustrate this with the problem of translating Esperanto *fako*. This may appear in French as *branche, case, compartiment, division, discipline, section, spécialité*, etc. according to context. The possibilities are listed in the dictionary as Esperanto pairs (with relators) and their French translations (Table 17.5).

fako a blanka ('empty square')	*case*
fako a libera ('free square')	*case*
fako de dungitaro ('personnel')	*division*
fako de ekonomio	*branche*
fako de financo	*section*
fako de matematico	*section*
fako de medicino	*spécialité*
fako de industrio	*branche*
fako de scienco	*spécialité*
fako de sporto	*discipline*
fako en bretaro ('shelving')	*rayon*
fako en dokumentujo ('filing cabinet')	*compartiment*
fako en ekonomio	*branche*
fako en kofro ('case')	*compartiment*
fako en magazeno ('store')	*rayon*
fako en medicino	*spécialité*
fako en scienco	*spécialité*

Table 17.5 Translation into French of *fako*

If there is a straight match, as in (14a), there is no problem with the translation (14b); but if the particular pair does not occur, e.g. (15a), then proximity scores of *industrio* and the known conjuncts of *fako* must be calculated. The closest analogy is *ekonomio*, based on such examples in LKB as: *industrio en lando* ('industry in a country'), *ekonomio en lando* ('economy in a country'), *industrio de agrikultura* ('industry of agriculture'), *ekonomio de agrikulturo* ('economy of agriculture'), *kreski as industrio* ('industry grows'), *kreski as ekonomio* ('economy grows'), etc. As a result the French output is (15b).

(14a) *fako de medicino*

(14b) *spécialité de la médecine*

(15a) *fako de industrio*

(15b) *branche d'industrie*

As in the first half of translation (stage 4, section 17.6.1 above), the same method is applied also to relational ambiguity. The Esperanto *ĉirkaŭ* could be

autour de, vers or *aux environs de*. The highest word-pair match scores for each French version in the contexts of phrases (16a,b) are given in Table 17.6.

(16a) *flugi ĉirkaŭ planedo* ('fly around a planet')

(16b) *veni ĉirkaŭ tempo* ('come around a time')

	autour de	*vers*	*aux environs de*
flugi	*rotacii* 'rotate'	*alveni* 'arrive'	*viziti* 'visit'
Score	0.604	-0.218	-0.293
planedo	*tero* 'earth'	*dato* 'date'	*dato* 'date'
Score	0.761	0.320	0.320
Average	0.683	0.051	0.014
veni	*instali* 'install'	*alveni* 'arrive'	*okazii* 'happen'
Score	0.530	0.820	0.789
tempo	*jaro* 'year'	*horo* 'time of day'	*dato* 'date'
Score	0.677	0.801	0.753
Average	0.604	0.811	0.771

Table 17.6 Word-pair scores for *ĉirkaŭ*

The clear preference in the case of (16a) is *autour de* (17a), and the preference (but by not so great a margin) in the case of (16b) is *vers* (17b).

(17a) *voler autour d'une planète*

(17b) *venir vers le temps [de la vendage]*

17.6.3 Evaluation of SWESIL

In early 1988 an evaluation was made of the effectiveness of SWESIL in lexical transfer, with test passages amounting to some 600 words restricted to the vocabulary of 'Simplified English.' The conclusion was that performance suffered from (a) the lack of frequency information in the LKB, both of lexical collocations and of structural plausibilities (i.e. neither the Metataxor nor SWESIL gave probability rankings to different analyses); (b) the lack of source language information in the first (English–Esperanto) stage, in particular the relational contexts of any pairs of words being examined; and (c) the deficiencies and inconsistencies of the databases. In addition, the model took no account of morphological structure (e.g. it did not know that the suffix *-isto* refers to agentive nouns); it had no access to phrase-level relations; it handled structural ambiguity poorly; and it could not handle inter-sentence relations or other text grammatical features.

It was decided, therefore, in 1989 that for the commercial system to be developed by 1993 a new translation model should be adopted by DLT based on the concept of the Bilingual Knowledge Bank (BKB). A major motive was the recognition that the large databases required could not be constructed in the way they were in the prototype; lexical information was henceforth to be derived

from actual texts, not built by human effort in dictionaries largely subjectively. In general, rule-driven processing was to be reduced and replaced by example-based processing using data from structured corpora of parallel bilingual texts. There would be no rigid distinction between syntactic and semantic analysis, disambiguation and transfer would be based on comparative bilingual examples, frequency information and dynamic updating would be provided, and in particular texts themselves would be the actual databases.

17.7 Conclusions

The BKB concept represents the final stage of a gradual move in DLT from the traditional rule-based approach to MT to an example-based approach. It is also a departure from the initial Esperanto-centred model: there is no longer an Esperanto 'knowledge base' (LKB) as the sole source of information for source language disambiguation. Esperanto is effectively seen as just one of the languages which may be present in BKBs. In theory, the new DLT model (described in section 18.2) need not involve translation from and into Esperanto at all, although it is argued that an intermediate language is still essential in the kind of multilingual configuration envisaged at the outset, namely a system for monolingual users to communicate in a communications network (section 17.1 above).

The main points of interest in the DLT prototype project can be summarized, then, as: the thorough exploration of a 'natural language' interlingua as the basis for an MT system; the full commitment to autonomous dependency syntax (i.e. not attuned as in Rosetta, cf. Chapter 16); and the use of text-based information in disambiguation and transfer (as opposed to rule-based dictionaries with semantic features, etc.). Although Esperanto may be more regular and consistent lexically and structurally than 'unplanned' natural languages such as English, French, Russian and Japanese, it presents similar problems of lexical and structural transfer, as we saw in the *akra*, *fako* and *ĉirkaŭ* examples above, and it is probably the distinctive approach to these problems which gives the DLT investigations significance beyond this particular project. However, it may well be the later example-based BKB model which, from a future prespective, proves more influential than the Esperanto-based prototype design.

17.8 Sources and further reading

The overall conception of the basic DLT prototype is discussed in the report of the feasibilty study by Witkam (1983) and in later accounts by Schubert (1986, 1988). The primary source for details of the metataxis is Schubert (1987) and Maxwell and Schubert (1989), and for the prototype version of SWESIL the source is Papegaaij *et al.* (1986). SWESIL is also described in Sadler (1989), which is devoted primarily, however, to an outline of the BKB model (see section 18.2).