Computers and Translation 1 (1986)

Linguistic And Extra-Linguistic Knowledge

A catalogue of language-related rules and their computational application in machine translation

Klaus Schubert

1. Rules and knowledge

Machine translation has inspired the hopes and inventiveness of scholars. When computers were developed, success seemed to be within reach. Disillusion followed about a decade later. The failure was due to "the number and difficulty of the linguistic problems" (Slocum 1985: 1).

Now new attempts are being made. This paper is intended to give an overview of the problems encountered in translating a text. Language is a rule-governed system. Language science is the discovery, translation and application of these rules. But while a human translator can use the rules intuitively, the application of a computer involves the necessity of formulating the rules explicitly.

Rules can only work when well-defined knowledge is accessible. Thus translation is based on both **rules** and **knowledge**. Although a rule-governed system, language is by no means free from outside influence. Describing the system of a language therefore requires reference to extra-linguistic influences. Accordingly, translation requires rules about both inside and outside influences on language; these rules in turn presume knowledge about those language-related influences.

After a look at the theoretical basis of this view in section 2, I describe in section 3 the practical details of the DLT machine translation system starting from the search for rules and knowledge. In 4, I sum up the rule systems accounted for so far and relate them to the types of knowledge they require. This concordance of rules and knowledge leads into a discussion in 5 of three characteristic features of the DLT system which might seem controversial, but which can then be shown to be strictly related to the rules and knowledge needed for machine translation. Finally, in 6, the question of priority for either language-specific or extra-linguistic rules and knowledge is taken up.

2. Computers and translation

Before turning to the practical details of making a computer translate, it is useful to have a clear idea about what translation is and what a computer really does.

To translate means to render in another language the content of a given text. This definition implies that the content can be preserved while the language is changed. A language is a system of signs of **form** and **content** (Saussure 1916:32). Thus, to translate means to change the form of the signs, preserving their content. Given a linguistic sign in a source language, the goal of translation then is to find its content and to determine a linguistic sign in the target language which renders that content. The term "linguistic sign" may refer to various units, such as morpheme, word, phrase, clause, sentence, paragraph, and text.

A computer manipulates symbols. If translation means replacing the form of signs, it comes down to the manipulation of symbols. A computer ought to lend itself to

---

Klaus Schubert is chief linguistic designer in the DLT team at Buro voor Systeemontwikkeling in The Netherlands.

this task straightforwardly. Manipulating forms, however, is not enough. The crucial goal is to preserve the meaning of the text to be translated. The computer is well-suited to manipulate the formal side of the linguistic sign, but there is no simple link to the content side.

Before asking whether, and eventually how, content can be handled in a computer, it is worth asking whether this is really necessary, and if so, whether there are limitations in the degree of depth and detail to which meaning must be handled.

Correspondences of meaning between two languages can sometimes be found at the word level, but most often they must be sought on the phrase, clause, sentence or text level or in some cross-relation.

If meanings are to be analyzed and explicitly described, some sort of meta-representation is required. I return to a suitable form for such a represention (5.2).

Machine translation, therefore, requires the capacity to handle in a computer the content side of the linguistic sign. This does not mean that meaning analysis should go into indefinite depth. Meaning analysis need not go deeper than to establish correspondences between forms sufficient to fulfil that function. For a machine translation system, there is no reason to imitate the way a human translator works. A computer system relies much more than a human does on formal rules, rather than on rules of content.

3. DLT--the practical details of machine translation

In section 2, I outlined what translation is and how it can be done by means of a computer. The answers given so far have been so abstract that many questions have remained unanswered. What about morphology, word formation, syntax, text coherence, word and phrase semantics, semantic deixis, theme and rheme? What about parsers, tree structures, transducers, word experts, meta-representations, on-line dictionaries, semantic networks, knowledge banks, world models, artificial intelligence?

These questions can be subsumed under the overall question asked above: What kinds of rules are necessary to "replace the form, preserving the content", and what kinds of **knowledge** do these rules require?

In spite of this admittedly abstract question, in this section I address machine translation from a very practical starting point. I describe how a sentence is translated in one machine translation system, take up the problems in the order in which they are encountered and outline the rule systems required for resolving them.

In a description of the complex network of rules, modules and systems involved in translating by computer, a sharp distinction between a number of different levels is required. I distinguish three levels:

- language-related rules,
- formalisms, and
- programs.

These three levels can be defined as follows: Language-related rules concern language without any link to a computational application; programs are purely computational without any link to the content of the problems processed; formalisms are an interface between the two. The main concern of this paper is language-related rules and the knowledge they assume. I describe them in this section and its subsections. In section 4, I link these rules to the formalisms in which they are applied, e.g., parsing formalisms.

The machine translation system I describe is called "Distributed Language Translation" (DLT). It is being worked on as a major research and development project at the Dutch software house *Buro voor Systeemontwikkeling* (BSO/Research) at Utrecht. My description covers the existing grammars, modules etc., as well as future developments which are still in the design phase. The overall linguistic and computa-

tional architecture of the system has been accounted for by Witkam (1983). To illustrate the current stage of development (September 1986), it may suffice to name a number of deadlines ahead. DLT is designed as a multilingual system, which is one of the major reasons for using an intermediate language (IL). In the present period, 1985-1990, a prototype version is being developed for the full IL kernel and one pair of source and target languages, English and French. Development started from the IL kernel, with English and French following. The DLT system is scheduled to translate Simplified English (a restricted-writing grammar) into French in early 1988 and ordinary informative (non-literary) English without domain restrictions into French by the end of 1990. Earlier still, by the end of 1986 the semantic translation precision of the DLT system will be tested by an external expert (cf. Melby 1986: 105).

3.1. Monolingual source language processing

The DLT system is designed for data communication networks among desk computers. Suppose an author (or typist) enters the sentence:

> Many multinationals were allocated grants for the study of capital development strategies for Third World member states, which will be of increasing importance in the future.

Typing such a sentence takes quite some time. This is the time the DLT system uses as its main processing time. It reads the first letter as soon as it has been entered and begins the syntactic analysis as soon as the first word boundary has been reached.

The first rule system needed is therefore a sentence syntax of English [1]. (The rule systems are here given numbers in square brackets and will be summed up in 4.) The syntactic analysis (language-related level, cf. 3) is carried out by a parser (formalism level). The parser [2] which uses the English dependency syntax in DLT is an Augmented Transition Network (ATN) for Simplified English (see 3). DLT's ATN formalism is an extended version of Bates's (1978); details of the DLT application are discussed by Doedens (to appear). Alternative parsing formalisms are under experimentation. The parser finally selected will be extended to cover not only Simplified, but normal informative English as well.

A **syntax** is a description of how words combine. Words combine to form meaningful texts in a systematic way. Syntax can thus be described as a **system** which, like any system, consists of **elements** and the **relations** among them. A traditional way classifies elements, such as verbs, nouns, etc. There are, however, two different theoretically possible approaches to a description of the relations among words: **constituency** and **dependency** (Klein 1971: 14; Mel'čuk 1979: 4; Schubert 1986a: 13). Generative syntax and other well-known syntactic schools apply the constituency approach, whereas DLT has opted for dependency, which is especially suited for a contrastive application such as translation (Schubert 1986a: 14). This will be argued in more detail in 5.1.

When a syntactic analysis is carried out, access is required to syntactic information about words; this is found in syntactic dictionary entries [3]. In order not to extend the dictionary with inflected forms, compounds and the like, an analysis as to morphology [4] and word formation [5] should be carried out. Thus a syntactic analysis does not consider words in sentences only, but also morphemes within words. Morphology and word formation can be seen in close connection with the dictionary, which therefore consists of word entries with syntactic information, and is supplemented with lexical redundancy rules on morphology and word formation.

What the DLT parser for English does is to analyze the sentence in accordance with

an English **dependency tree** structure. In a dependency tree, the words are displayed on the nodes (optionally with labels for word class and morphological features). The branches represent the relation "depends on", so that every word has one and only one immediate governor, which stands above it. The only exception is the main governor of the whole sentence, usually a finite verb, which itself has no governor on the sentence level. Dependency relations can be classified on distributional grounds, without reference to meaning. The type of dependency relation between the word and its governor is given as a branch label between the governor and the dependent. A dependency tree for the example sentence is given in Figure 1.

The tree shown in Figure 1 is the "correct" one. Because the analysis is purely syntactical, i.e., formal, and is thus performed without access to semantic information, the sentence appears to be syntactically ambiguous. If the syntactic dictionary entries contained only word class information, a preposition, for example, could depend on virtually any preceding verb, adjective or noun. The ambiguity that would result can be restricted by means of valency information, which is either directly entered in the dictionary or rendered in lexical redundancy rules. **Valency** is subclass-specific government capacity (Engel 1982: 110; Schubert 1986a: 18). Superficially seen, it resembles strict subcategorization of generative grammar.

A parser equipped with an English dependency syntax with full valency information in the dictionary nevertheless leaves two of the dependencies in the example sentence unresolved: The second *for*, with *Third World member states* depending on it, can depend either on *strategies* or on *study* (*strategies for member states* vs. *study for member states*), and the relative clause (*which will. . .*) could depend on *grants*, *study* or *strategies* - probably not on *states*, because of the comma. In addition, it can also depend on the whole preceding sentence (thus on *were* as the sentence governor). On the sentence level, this leaves eight parallel successful parse trails and eight alternative dependency trees.

It is a question of efficiency in effort, run time and storage space, whether these eight trees should really be generated as eight different pieces of output or in the form of one complex structure, e.g., a "packed shared forest" of Tomita's type or a similar structure (Tomita 1986a:20, cf. 1986b:43). Such compaction becomes more desirable when syntactic ambiguity and word choice parallelism combine in the subsequent bilingual steps (cf. 3.2). For the sake of clarity in the present discussion, however, I speak of distinct trees. These trees are conceptually equivalent to a more compact representation.

How can the syntactic ambiguity be resolved? There are four possibilities which supplement each other: **text syntax, sentence semantics, text semantics** and a **disambiguating dialogue** with the user. Text syntax is mainly used for analyzing (and translating) text coherence as found in theme-rheme relations. The syntactic ambiguities of the example sentence can probably not be resolved by text-syntactic means. Neither can sentence semantics contribute much here; its main role is in contrastive word choice (see 3.2.2). Text semantics, describing the semantic coherence of a text and its components, can resolve at least the question about a governor for the relative clause, if useful cues are found in the preceding text (DLT's speed-oriented approach makes it desirable not to wait for the subsequent text). The *for* phrase is a typical representative of the virtually unresolvable type of ambiguity. It might go to the dialogue. Details about the dialogue will be discussed below in 3.2.3.

In the DLT approach, no attempt is made to resolve any not purely syntactic problems

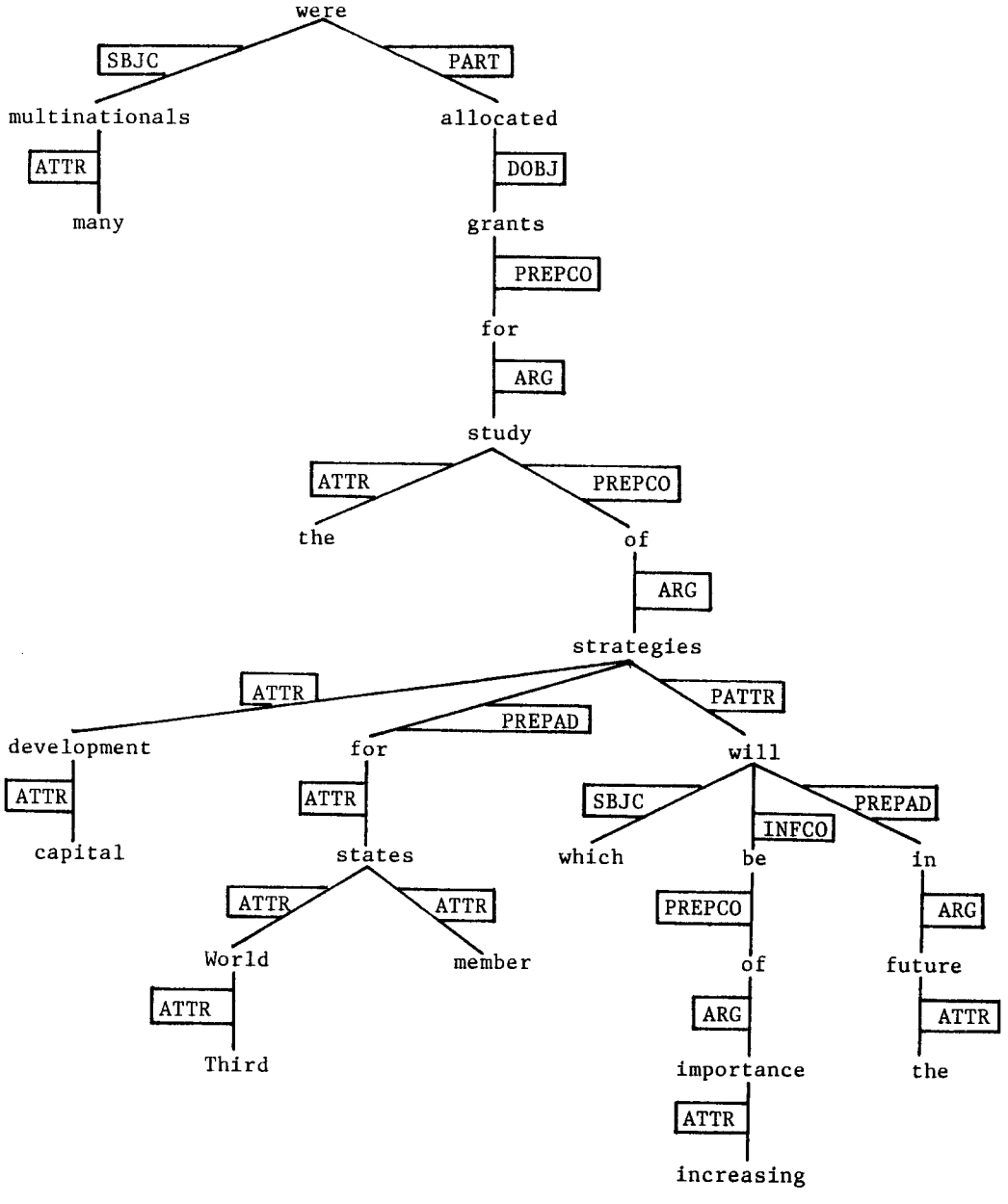

Figure 1: Dependency tree for the example sentence. The labels are tentative.

on the monolingual level of the source language (in the example, English). Whenever meaning is involved, processing takes place in DLT's intermediate language (IL). Of the four rule systems mentioned above which might aid disambiguation, the sentence has so far encountered only one: a text syntax of English [6]. If text syntax cannot reduce the degree of ambiguity, the sentence is passed on from the monolingual stage in eightfold ambiguity and is in this form translated into the intermediate language.

3.2. Source language into intermediate language

DLT not only starts reading and parsing a source language text as soon as the first letter has been entered, but also tries to begin translating it as early as possible, virtually from the first word.

Translating means replacing the form of the linguistic sign while preserving the content. The form in this respect is not only the sum of the words, but also the syntactic (i.e., formal) relations among them, the dependencies. In other words, the form of a sentence is its dependency tree.

Accordingly, transfer is performed essentially by two rule systems: **word translation rules** and **dependency transfer rules**.

The system of word translation rules in turn consists of two distinct systems: first, there is an English-IL dictionary [7] which specifies which English word to replace with which IL words, as any conventional bilingual dictionary does. Because there generally are no one-to-one correspondences, the dictionary gives several alternative translations for each word, among which a choice must be made by means of word choice rules (see 3.2.2).

Word choice is the major semantic problem in translation. A human translator usually understands immediately what is meant from the context. What, then, is context, and can it be processed in a computer? **Context** is co-text plus situation. This clear-cut definition, taken from Nikula (1986:41), points directly to the types of knowledge needed for resolving the word choice problem: knowledge from the co-text, i.e., the rest of the text within which the word occurs, and knowledge which cannot be found in the text, but which the human translator tacitly applies, referring to sources other than the text itself. These sources can either be other texts (which for a computer application could be simulated by a very large text corpus), or the translator's knowledge about the objects and phenomena of the world, which may never have been expressed explicitly. In other words, knowledge from both inside and outside the language system is needed, i.e., **grammatical** and **pragmatical** knowledge, or: linguistic and extra-linguistic knowledge.

The necessity of handling both linguistic and extra-linguistic knowledge having been stated, the question remains how these sources of knowledge can be made accessible for processing. The linguistic knowledge needed can be inferred from the text by means of grammar, whereas the pragmatical factors require knowledge of the world.

A grammar is a description of the internal rules of a language about the form and content of the signs (words, etc.) that make up the language system. Grammar, as needed for machine translation purposes, thus comprises syntax and semantics. When I avoided the word "semantics" in section 2 above, speaking only about meaning or content there, I did so in order to emphasize that meaning is more than what semantics deals with. **Semantics** is concerned with the language-internal part of the overall concept of meaning, i.e., meaning as part of the system of one language. There is also, however, language-independent meaning, often termed knowledge of the world, which comprises encyclopedic knowledge, logic, etc.

The interaction of language-independent meaning (the situation, in Nikula's definition, above) with what is meant by language utterances, is the subject of **pragmatics**.

In DLT, **knowledge of the world** is stored in a lexical knowledge bank and made accessible in a word expert system. What the expert system does is to figure out which of the alternative IL translations of an English word semantically and pragmatically best fits the other words in the text. Before discussing (in 3.2.2) how this is done in detail, it is worth spending a few words on the question of what exactly is the input to meaning processing in the word expert system.

It would of course be useless to process every possible two-word combination within a text or even a sentence. What could be won from a check on meaning compatibility between *many* and *future* in the example sentence, not to mention comparing the two *of*'s? Thus it is important first to find out which words can influence each other's interpretation. Before applying extra-linguistic knowledge, we should attempt to resolve as much of the problem as possible on linguistic grounds. This is not only the most efficient way of doing it, it is also a theoretical requirement. Only when the syntactic, and, as far as possible, the semantic contribution to the translation task has been made does it become clear how to apply semantic-pragmatic word expert knowledge.

At the beginning of this section, two rule systems were mentioned for performing the transfer from source language to intermediate language: word translation and **dependency transfer**. So far, I have dealt only with word translation. Before returning to the details of this intricate question (in 3.2.2), it is essential to have a look at dependency transfer and see how much it can prepare and select the input to the word expert system, i.e., to word translation.

The idea is not to feed into the word translation modules every random word pair from the sentence or, still worse, from the whole text, but to proceed along the lines of structural relations, i.e., applying the word translation rules basically only to a word and its syntactic governor. This opens a restricted number of directions in which to look for meaning compatibility: upwards in the syntactic tree, only one direction, to the governor (if any); downwards, as many directions as there are words for which the current word is the governor, normally not more than a handful of dependents. Of course, word translation ultimately requires access to the whole text, but the analysis should not proceed from one word to another along other than syntactic dependency lines.

The syntactic analysis can contribute more to translation than just establishing which word depends on which other word. It also provides a classification into various types of dependency relations. Because the goal of word choice is to select one out of several IL words given in the dictionary as tentative translations of an English word, meaning processing deals with IL words. If the syntactic analysis of the input sentence (or text) is to be used for selecting the input word pairs for meaning processing, it is necessary to transfer from English into the IL not only the words but also their mutual dependencies. This comprises two tasks: the structure of the English dependency tree must be transformed into an IL tree, and the English dependent labels must be replaced by IL labels.

The rule system by means of which these two tasks of dependency transfer are performed is called the English-IL metataxis system [8]. **Metataxis** is a name for translation-oriented contrastive-syntactic rules for a given language pair (from Tesnière's term *métataxe*. 1959:283). Metataxis rules for a given language pair can only be written if there are consistent syntactic descriptions of both languages on sentence and text level (English [1], [6] and

IL [9], [10]) with corresponding dictionaries ([3], [11]).

Metataxis and word choice are not so distinct as they might seem. Metataxis is syntactic, and syntactic rules are based on the syntactic features of words such as word class and valency. Therefore the choice between two different IL words translating an English word can also imply the choice between different ways of combining the chosen word with the rest of the sentence. Word choice can trigger rearrangements of the sentence structure. Metataxis rules, consequently, occur on several levels of generality, and the less general rules are found - either explicitly stated or in the form of **lexical redundancy rules** (cf. Schubert 1986a: 184) - in the bilingual dictionary, which I mentioned earlier as one of the rule systems for word translation. The more general metataxis rules are triggered by word class and need not appear in the dictionary.

It is in the bilingual steps that DLT's intermediate language appears. Essentially, this language is Esperanto. The reasons for this choice will be presented in 5.3. A number of adaptations have been necessary for Esperanto to fulfil the role of an intermediate language in machine translation. In order to make that difference clear, DLT's intermediate language is called IL as opposed to common Esperanto.

In the IL, there is no word class ambiguity: each word belongs to one word class only. Furthermore, in the productive word classes (verb, noun, adjective, adverb), the word class can be read from the word itself, so that dictionary look-up is only needed for the enumerable classes (some 300 words). The word *grand'a'j* 'big', for example, consists of three morphemes, where *grand* means 'big', *a* is the adjective identifier and *j* denotes the plural (concerning IL morphemes, Schubert 1986a: 23). The IL is completely agglutinative. Within the words, the morpheme boundaries are marked by the token *(')*, so that no ambiguities can arise from morpheme boundary uncertainty, although morpheme combination is productive in DLT's IL-generating modules (cf. 3.2.1).

In the following paragraphs, the English-IL translation steps of the DLT system will be illustrated. As the discussion so far has shown, the order of working steps will be:

- replacing an English word with tentative IL translations in parallel that come from the English-IL dictionary (rule system: bilingual dictionary);
- arranging the tentative IL translations in syntactically correct IL trees (rule system: metataxis);
- selecting the semantically and pragmatically best tree (rule systems to be discussed in 3.2.2).

The first of these steps is easily done: The English word is replaced with all its IL translations from the bilingual dictionary. In the following sections, these word translations are taken from DLT's dictionary, and where words happen not to have been entered as yet, from Wells (1985) and Munniksma et al. (1975). These replacements, as well as the rearrangements of the whole tree structure brought about by the rules to be discussed below, are carried out by means of a tree transducer [12] which can be found either on the formalism or on the program level.

The second and third steps will be illustrated in 3.2.1 and 3.2.2. I do not show all the steps of translation for every word of the example sentence, but only a few typical problems with their solutions.

3.2.1. Metataxis

Metataxis comprises rules for correspondences in terms of syntactic form, i.e., word class, inflection and dependent type. Furthermore, the level of syntactic unit can be changed, so that a word can be translated as a phrase or as one word, or even as a morpheme within a complex word.

As nothing may stay untranslated, it is advisable to use metataxis rules for **unmarked cases** (default rules). For each English word class there should be a rule which specifies the word class of the IL translation, and for each English dependent type there should be an unmarked dependent type correspondence in the IL. As far as word class is concerned, the unmarked rules are rather implicit, because lexicographers usually render nouns as nouns, verbs as verbs, etc., so that the IL translations will come from the bilingual dictionary in the form triggered by the unmarked rule.

These remarks have direct consequences for the processing of the example sentence. If the number of words in the English and the IL form of a sentence do not coincide, replacing the words implies rearranging the syntactic tree structure. This is necessary in the case of *multinationals*, for which the dictionary gives not one IL word, but a two-word phrase, *mult 'naci 'a 'j entrepren 'o 'j*, with an internal syntactic structure in which the adjective *mult 'naci 'a 'j* is a pre-attribute of the noun *entrepren 'o 'j*.

The typical English construction *multinationals were allocated grants* (subject, passive, direct object) is impossible in the IL where passive forms (as in many other languages) are always intransitive. There is a metataxis rule which inserts the preposition *al* 'to' and transforms the construction into *al mult 'naci 'a 'j entrepren 'o 'j asign 'ajt 'is subvenci 'o 'j*, i.e., it transforms the English subject into a prepositional complement in the IL, the passive into a passive, and the direct object into a subject. The constituent order is preserved. In the IL, the subject is not identified as in English by constituent order, but by case (nominative as denoted by that fact that *subvenci 'o 'j* lacks the accusative morpheme *n*).

Metataxis rules are formalized in the form of tree structures which in the system itself will be represented in a list notation (Schubert 1986a: 189), but which are more perspicuous in a graphic form as displayed in Figure 2.

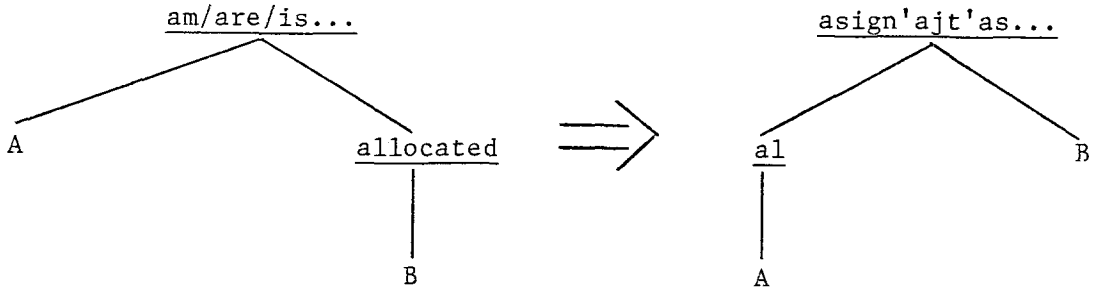

Figure 2: Graphic representation of one of the English-IL metataxis rules needed for the example sentence.

As was shown in 3.1, the relative clause of the example sentence can depend on four different governors. The choice of governor determines which word is the correlate of the relative pronoun *which*. (The correlate, sometimes denoted by the more word order-directed term 'antecedent', is the word to which a relative pronoun or another deictic element semantically refers. It is not normally the syntactic governor of the deictic word itself cf. Schubert 1986a: 76.) The correlate triggers some features of the form of the IL corrrespondence (whereas others are triggered by the governor). The first question is whether the correlate is a noun or a verb. For the former, the translation is *kiu*, for the latter *kio*. For *kiu*, which depends on nouns much like an adjective, number agreement must be observed. If the correlate is *study/stud`o, which* becomes *kiu* (singular), if it is *grants/subvenci`o`j* or *strategies/strategi`o`j*, the translation is *kiu`j* (plural). In this way, the IL translation becomes syntactically more precise than the original sentence. This is just one small example of DLT's **information enrichment**, the result of which is stored in IL form.

When the nouns of the example sentence are translated as IL nouns in accordance with the appropriate unmarked metataxis rule, it turns out that a number of IL nouns occur in positions where they cannot remain as nouns. This is usually true when an English noun complex is translated: *Capital development* cannot become **kapital`o evolu`ig`o* (*o* denotes nouns). *Capital* is an attribute (ATTR) of its governor *development*, and the unmarked metataxis rule for attributes transforms it into a pre-attribute (ATR1) in the IL. An IL noun, however, cannot be a pre-attribute. There are two ways out: Either the pre-attribute label must be removed, e.g., by merging the two words into a one-word compound: *kapital`o`evolu`ig`o*, or the word must be adapted to its function as a pre-attribute and be turned into an adjective: *kapital`a evolu`ig`o*. These two forms differ slightly in meaning, so that the choice must be made on semantic grounds. Furthermore, both options can be still enriched, so that they render more precisely the semantic relation between 'capital' and 'development':

- Is the capital being developed? Then: *kapital`o`n`evolu`ig`o*; here, the accusative morpheme *n* denotes the 'capital' as the object of 'develop';
- Does the capital develop, as it were, self-initiated? Then: *kapital`o`evolu`o*; IL *evolu`i* is intransitive, and the transitivizing *ig* is missing;
- Does something develop by means of capital? Then: *per`kapital`a evolu`o*; the preposition *per* 'by means of' is added;
- Is something being developed by means of capital? Then: *per-`kapital`a evolu`ig`o*.

The possibility of combining words and morphemes to form new words *(kapital`o`evolu`ig`o)* allows for **productive generation** of IL words which may not be found in any of the IL or bilingual dictionaries within the system. Because a computer cannot invent totally new morphemes, but only combine known ones in a new way, all unknown words are analyzable. This analysis of multi-morpheme words is carried out according to a grammar covering the syntactic and semantic relations of morphemes within a word, the IL **word grammar [13]**. It comprises morphology and word formation and was given the name 'word grammar', because in a totally agglutinative language, there is no need to separate morphology and word formation in the traditional way. The output of the analysis is a dependency tree of morphemes, from which a paraphrase in known words

(semantic analysis) and valency features (syntactic analysis) are inferred. Since this word grammar module is needed, it can also be used as a set of lexical redundancy rules for complex words and additionally even as a tool for lexicographers (Schubert, to appear).

The translation of English noun clusters is another example of information enrichment, which leads to semantic or, rather, pragmatic questions.

3.2.2. Word choice

Before refinements such as semantic information enrichment can be considered, however, the problem of word choice must be solved. The word choice modules are used to select one out of a number of parallel tentative IL trees. When this selection begins, the metataxis rules have already performed the structural transfer and have, during that process, continuously checked the IL tree for its syntactic coherence. The result of that syntactic transfer is syntactically correct trees of IL words with dependent labels.

These trees are now (in a tree transducer [12]) chopped up into **word pairs**, the meaning compatibility of which will be compared in the word expert system ("Semantic Word Expert System for the Intermediate Language" = SWESIL [14]) with that of parallel pairs with alternative word or phrase translations. The input pairs for the word choice module are not formed mechanically from any two words which happen to occur adjacent to each other in the tree, but they are formed of content words only. Two content words are linked by a **relator** which can be derived either from the dependency relation that holds between the two words, or from a function word occurring between the two content words. From one of the alternative IL trees which tentatively translate the example sentence, among others, the following pairs are derived:

asign`ajt`is	al	entrepren`o`j
subvenci`o`j	`is	asign`ajt`is

In the first example, the relator is the preposition that links the two words in the tree. In the second, the relator *`is* shows *subvenci`o`j* as the subject of *asign`ajt`is* (*`is* is the past tense morpheme from the finite verb itself; thanks to a purposeful choice of existing morphemes for the relators, the relator pairs can be read roughly as rudimentary Esperanto sentences, which much facilitates the work of the DLT semanticians. For details, cf. Papegaaij 1986: 96).

The alternative IL translations are fed into the word expert system in the form of such related pairs. Let us take as an illustration the words *study*, *future* and *capital* from the example sentence.

The three words have the following IL correspondences:

study	stud`o	'studying'
	stud`aĵ`o	'a study of something'
	stud`ej`o	'room for study'
	kabinet`o	'room for study, consultation, etc.'
	etud`o	'musical study'
future	futur`o	'future tense'
	est`ont`ec`o	'future time'
capital	kapital`o	'finance'
	ĉef`urb`o	'capital city'
	majuskl`o	'upper-case letter'

In these translations, the unmarked word class metataxis rules have already been applied, so that the translations for the verb *study* and the adjectives *future* and *capital*

are not taken into consideration, because the three words have been recognized as nouns in the given English sentence.

The syntactic form provides more information than simply the word class. The dependency tree displays a map of syntactic relations among the words of the sentence. These relations define the lines along which to look for semantic compatibility within the sentence. As for *study*, the goal is to find out whether an activity *(stud`o)*, a result of the activity *(stud`aj`o)* or a room for such an activity *(stud`ej`o, kabinet`o)* is meant. The search for cues begins in the immediate co-text as defined by the dependency tree. The tentative translations of *study* are paired off with those of the closest content words in the tree, while the prepositions occurring in between determine the semantic relation among the words. For *study*, this is especially easy, as *grants* and *strategies* have only one translation each, so that they can be taken as islands of certainty which are useful starting points for semantic decision taking (Papegaaij 1986: 187). The pairs to be entered into the word choice process are thus:

subvenci`o`j	por	stud`o
subvenci`o`j	por	stud`aj`o
subvenci`o`j	por	stud`ej`o
subvenci`o`j	por	kabinet`o
stud`o	de	strategi`o`j
stud`aj`o	de	strategi`o`j
stud`ej`o	de	strategi`o`j
kabinet`o	de	strategi`o`j

There is not only one possible translation for such a vague word as the English preposition *of*. Therefore, there are alternatives also for the relator:

stud`o	de-`n	strategi`o`j
*stud`aj`o	de-`n	strategi`o`j
*stud `ej`o	de-`n	strategi`o`j
*kabinet`o	de-`n	strategi`o`j

The relator *de* denotes a possessive relation, the relator *de-`n* with the object case morpheme *`n* marks a relation in which the right-hand word is the object of the (nominalized) event expressed in the left-hand word. (This is a disambiguation of *the shooting of the hunters* problem.) Of the last four pairs, only the first one is possible. The other three are ruled out as soon as they occur, because the left-hand words do not express nominalized events and are therefore incompatible with the relator. Does this compatibility check require another rule system? Fortunately not: as the relators are IL morphemes, their compatibility with IL words can be found in the syntactic dictionary entries of those words. The question whether a noun expresses an event (and not an object or a phenomenon), can be resolved by means of the IL word grammar [13]. This rule system is not needed for this purpose only. The compatibility check for relator pairs is a secondary application of it. (For its main use, cf. 3.2.1.)

The technical particulars of how the meaning compatibility of the words of different pairs is established and compared will not be discussed here in detail. They are described extensively by Papegaaij (1986: Part II; for a short overview cf. Papegaaij/Sadler/Witkam 1986). It may suffice here to describe what exactly meaning compatibility is and what knowledge is required for establishing and measuring it.

What does it mean, exactly, to compare the meaning compatibility of alternative relator pairs? Speaking in somewhat more concrete terms, it means that a score is com-

puted which renders the probability of the two words of the pair co-occurring in the syntactic relation found in the sentence. The comparison (carried out in an evaluator [15]) is then made among the scores computed for various alternative pairs. How can such scores be established?

This question implicitly contains its answer. Given that the goal of applying the word expert system is to compute probability scores, two things are needed in order to reach this goal: knowledge and rules. Knowledge is needed about the words which are semantically and pragmatically compatible in certain dependency relations, and rules are needed in order to infer scores from that knowledge.

The required knowledge is stored in a **lexical knowledge bank [16]** which contains both language-specific (semantic) knowledge about meaning, and language-independent knowledge, the influence of which on language is **pragmatic**. In DLT, the knowledge bank is entirely in the IL. A lexical knowledge bank essentially is nothing but a common monolingual explaining dictionary or encyclopedia, in a form suited for an **artificial-intelligence** application. In a normal monolingual printed dictionary for human users, definitions of the head words are given in (parts of) sentences in the same language. The DLT knowledge bank is also arranged in this way, defining IL words by IL words, but the explanatory text is divided up into relator pairs. These can be pairs containing the head word itself, but also additional pairs containing words which themselves occur in a pair with the head word. Such an entry may look like Figure 3.

`kapital`o (`sum`o) ((`mon`kvant`o))

FIRST ARGUMENT	RELATOR	SECOND ARGUMENT
`kapital`o (`sum`o)	`as	`don`i, (mal)kresk`i
`kapital`o (`sum`o)	`ies-de	`person`o, `rent`ul`o
`kapital`o (`sum`o)	`da	`mon`o
`kapital`o (`sum`o)	`por	`invest`ad`o, `viv`asekur`o
`invest`i, `hered`i, `donac`i	`io-n	`kapital`o (`sum`o)
`grand`a, `baz`a, `ĉef`a	`a	`kapital`o (`sum`o)
`viv`i	`de	`kapital`o (`sum`o)
`spekulaci`i, `viv`ten`i	`per	`kapital`o (`sum`o)
`don`i	`io-n	`interez`o, `rent`o

Figure 3: An entry from the IL lexical knowledge bank.

These relator pairs are not the only form of knowledge stored in the word expert system. Using them from the knowledge bank (called SOLL pairs; Papegaaij 1986:95), the only thing one would be able to say about the pairs derived from the sentence (the IST pairs) is whether or not they coincide. This is not sufficient; such an approach would only be applicable if the knowledge bank were complete, i.e., if it listed all possible combinations of words, which is, of course, impossible to achieve. It is therefore necessary to express in some way the **degree** to which the SOLL and IST pairs coincide in meaning. This degree of coincidence is the probability score needed as a selection aid.

In order to express **meaning proximity** in figures, however, yet another type of knowledge is required, other than that expressed by the SOLL pairs. It is not enough to know how words can combine; we also need to know how they can replace each other--in other words, how they are conceptually linked. Transferring the distinction of syntagmatic and paradigmatic relations to semantics and pragmatics, one finds the "syntagmatic" meaning relations--in a form intended to be complete for the words involved--in the relator pairs. The "paradigmatic" meaning relations, however, cannot be enumerated exhaustively, not even if the number of words is limited. The paradigmatic relations (which state whether a word denotes a part, an instance or a generalization of another word) are expressed by arranging the entry head words in a taxonomic tree structure. Each entry word for this purpose is assigned a **superordinate** word (a hypernym). Within the resulting structure, proximity can be computed in terms of how many superordinate levels have to be crossed upward and downward in order to link the SOLL word with an IST word (Papegaaij 1986: 119).

If such an analysis mechanism is at hand, the three example words *study, future* and *capital* show that the scope of the word expert analysis required for choosing a translation is quite different for different words. *Study* can be found to be *stud'o*, within the immediate sentence context, thus the activity of studying. For *future*, no cue is found in the sentence, but the two translations are so different in frequency, that cues ought to be present if the correct interpretation were to be *futur'o*, a grammatical tense. The lack of any such cue is sufficient reason to opt for *est'ont'ec 'o* 'future time'. *Capital*, however, is more difficult to resolve. At least the two interpretations *kapital'o* (finance) and *cêf 'urb'o* (capital city) make sense. Here, the entire preceding text should be scanned in search of cues to either the finance or the city reading.

3.2.3. A disambiguating dialogue

The problems of syntactic ambiguity and word choice are resolved by means of semantic and pragmatic knowledge about meaning inside and outside the language system. If all these efforts fail to lead to a safe conclusion, there must be a way out. This is essential to the DLT system, because the environment of communication networks among desk computers **does not allow for any form of post-editing**.

It is generally agreed that fully automatic high quality translation is ultimately unattainable, though it remains the ambition of many researchers. A translating computer system needs human help. In the DLT architecture, this help is provided in the form of a **system-initiated interactive dialogue** with the user who enters a text. The advantage compared to post-editing is that the DLT user does not need to know any other language but the language of the text. The user need not be a translator or even have the slightest knowledge about the language into which the text will eventually be translated, nor about the IL, which is invisible to the user.

DLT's interactive dialogue has been outlined by Witkam (1983: III-91). Others have

described similar ideas (e.g., Tomita 1986b). Before the dialogue can be started, the processing has passed from the monolingual analysis stage (English) into the bilingual transfer stage (English to IL). The rule systems controlling the dialogue must therefore be **bilingual** as well.

The function of the dialogue is to resolve syntactic ambiguities in the source text and problems of word choice, if they cannot be resolved by other means within the system. It is thus the task of the user to choose among a number of alternative IL trees which differ in words or labels or structure. This choice must be requested without bothering the user with either the IL or tree structures. The user will see only questions in plain English. The rule systems needed are:

- a tree matcher which separates out those nodes and branches at which the alternative trees differ [17];
- a paraphrase generator which formulates the crucial parts of the sentences in alternative IL phrases or sentences [18];
- a question generator which transfers the IL paraphrases into the source language (English). This is not "normal" translation with English as the target language, but rather a sort of retranslation, where appropriate words of the source sentence can and should be used, while the crucial ones are paraphrased [19].

3.3. Monolingual IL processing before transmission

As soon as one IL tree has been chosen as the correct translation of a source sentence (or text), the bilingual transfer goal is reached. In principle, the only thing that remains to be done is to send this IL tree to the receiver and there to translate it into the target language.

Before a text is ready for transmission, however, two other objectives have to be reached: The text for transmission should be as **compact** as possible, and it should be **inspectable**.

Compaction is achieved in two ways: first by converting the tree into a string (a sentence), second by coding the sentence for transmission. At first sight, it might seem superfluous and even disadvantageous to convert a syntactically analyzed tree into an unlabelled string. If this is done, it should happen in such a way that no information is thrown away. An IL tree is **syntactically unambiguous**, so an IL string should be syntactically unambiguous as well.

The modifications which distinguish the IL from common Esperanto are designed to achieve syntactic unambiguity. In accordance with one of DLT's basic principles, this has been done by **linguistic means** only (Witkam 1983: IV), without using indices, numbers, brackets, labels, flags, tags or whatever. The only exception to this principle is required when converting a tree into a string. Here, an extra, disambiguating space is used in order to indicate that the dependent following the space does not depend on the immediately preceding possible governor, but on the next one to the left. Several extra spaces in succession denote a corresponding number of possible governors to be skipped. Using the space for this purpose has the advantage that this (strictly speaking non-linguistic) sign is as invisible as possible, so that the sentence remains a readable IL sentence.

The rule system needed here is a tree-to-string converter [20]. In order to insert the disambiguating spaces it must have access to syntactic information about what is a possible governor of what. This information is found in the IL syntax [9].

The IL tree is thus transformed into the following IL sentence. (In order to make the

disambiguating spaces more visible for the reader here, I render them as underscores.)

Al mult`a`j mult`naci`a`j entrepren`o`j

asign`ajt`is subvenci`o`j por stud`o de

strategi`o`j`n por per`kapital`a evolu`ig` o _por

tri`a`mond`a`j ŝtat`o`j-membr`o`j_, kiu` j hav`os

kresk`ant`a`n grav`ec`o`n iam-en la

est`ont`ec`o.

The second step of compaction is achieved by a binary coding [21] of this text. DLT here profits from the total agglutination of its IL, which allows coding not of letters, but of morphemes.

It should be emphasized that tree-to-string conversion (which requires us to parse the IL string later cf. 3.4) is not only performed for the sake of compaction, nor even mainly for that reason. Compaction should be seen rather as a welcome but secondary effect. The main need for strings is that the string form of a text is not a complicated structure, but just an ordinary IL text. If the morpheme tokens and the extra spaces are suppressed, it looks hardly different from a normal Esperanto text and can be read as such. For trouble shooting and quality checks this possibility of half-way inspection is extremely important. It is especially useful during the development stage, but certainly not only then.

In addition, the possibility of rendering the text being translated in a machine translation system, at the intermediate level, as a normal text without recourse to non-linguistic signs, seems to be an indispensable theoretical requirement. This will be argued in greater detail in 5.2 below. Only if the intermediate representation is equivalent to a text in a human language can it render the full content of the input text.

As a last security check, the IL text is examined for syntactic correctness and unambiguity. This is done by an IL recognizer [22] which makes use of the IL syntax [9]. Any sentence which does not have one and only one syntactic interpretation is rejected and returned to the disambiguating modules (word expert system and dialogue). The IL recognizer is an ATN, an improved version of the ATN skeleton shown by Wilkam (1983: IV-89, Appendix 7-15).

3.4. Monolingual IL processing after transmission

When the SL text has been translated into an IL text, it leaves the sending computer and is transmitted through a communication network to one or more receivers. When the text arrives in that form at the receiver, it is decoded [23] and is then ready to be translated into any available target language. During the translation steps described above, the text is not prepared for translation into any specific target language.

The input for the following bilingual steps is an IL tree. The first module to go to work in the receiver is therefore an IL parser [24] equipped with the IL syntax [9]. This parser works faster than the recognizer (cf. 3.3), because the first path completed must be the proper and only solution.

3.5. Intermediate language into target language

The bilingual steps on the receiver side resemble the steps performed at the sender, with one important difference: at the receiver, no human aid is possible. In DLT, translation from the IL into the target language is **fully automatic**, and **no post-editing** is possible.

The bilingual steps from the IL to the target language, however, are not merely a copy of what was done at the sender, but rather a mirror-image. Here again, the problem of word choice is resolved in the IL, which is the source language for this part of the overall translation process. The meaning differences among alternative target language (French) translations of one IL word

must be rendered in IL paraphrases or hypernyms. These extended IL alternatives can then be entered into the **word choice** procedure. The same word expert system with the same lexical knowledge bank [16] as in the step from source to intermediate language (cf. 3.2.2) is used, which is possible because the operation is in both instances performed entirely in the IL. What is different is the bilingual dictionary [25]. (Here, I take the translations either from DLT's own dictionary or, where it is still incomplete, from Waringhien 1976 or Munniksma et al. 1975.) In addition, a different evaluator [26] is needed.

While the right French words are being chosen, they must be brought into an appropriate mutual relation. This goal is achieved by IL-French **metataxis** rules [27], which assume a dependency syntax not only for the IL but also for French (sentence syntax [28], text syntax [29] and syntactic dictionary [30]). The arrangement of dependents under the main governor of the sentence can again be taken as an example of how metataxis works. In 3.2.1, it was shown how the English construction of subject-passive-direct object (*multinationals were allocated grants*) was changed by a metataxis rule into the IL construction of prepositional complement-passive-subject *(al mult 'naci 'a 'j entrepren 'o 'j asign 'ajt 'is subvenci 'o 'j).* If the order of principal constituents (but not the word order within constituents) is considered to be controlled by text coherence requirements on the text-grammatical level, this IL construction should again be changed into another form for French, which preserves the principal constituent order: *multinationales se sont vu allouer des subventions,* i.e., subject-active-object. Here, *se sont vu allouer* is a complex verb construction whose form is triggered by tense-translating metataxis rules. The important feature here is that it is an active voice form.

3.6. Monolingual target language processing

Interleaved with the metatactic changes and transformations in the resulting French dependency tree, the generated forms are checked as to their compatibility with French morphology [31], word formation [32] and syntax on sentence [28] and on text level [29]. The output of this combined step with both bilingual and monolingual rules is a correct French **dependency tree.**

The last monolingual step is French tree-to-string conversion, carried out by an appropriate module [33]. The example sentence is along these lines translated into

> De nombreuses multinationales se sont vu allouer des subventions en vue de l´étude de stratégies de développement au moyen de capitaux pour des États membres appartenant au tiers monde, stratégies dont l´importance s´accroîtra dans l´avenir.

Here, there is no need to aim at the same strict syntactic unambiguity required of the IL (cf. 3.3), but the desirable degree of clarity for human readers should of course be achieved. For this purpose, the word *stratégies* is repeated in the relative clause. This insertion fulfils a similar function to that of the extra space in the IL.

4. A catalogue of knowledge and rule systems

After all the details presented in section 3, can the question about rules and knowledge now be answered? In brief, two types of language-related rule systems were mentioned: **grammar** and **pragmatics.**

Three grammars are needed, describing the internal rules of the three languages involved (source, intermediate and target). But there is only one pragmatic rule system. It applies world knowledge which is language-independent but is required for translation only inasmuch as it controls the choice among linguistic forms. As a consequence of the insight that world

knowledge is language-independent, the DLT architecture avoids repeating the encyclopedic knowledge of the world model for every source or target language, and stores it only in the one language which is always involved, the IL.

The grammars require grammatical, i.e., syntactic and semantic knowledge. There is, however, an important difference in the extent to which the various grammars make use of this knowledge. Syntax is language-specific, whereas semantics describes that meaning level on which a translation interface between languages must be sought. It is a distinctive feature of the DLT system that as many of the semantic rules as possible are written in the IL, so that while the DLT grammars of English and French (and other languages to be added) give a full account of syntax, they try to transfer the semantic problems as soon as ever possible into the IL. Returning to the discussion of section 2, this means that meaning is used primarily as a tertium comparationis between a source or target language and the IL, and whenever more detailed analysis and description of meaning is required, this is given on the IL side.

Therefore, the monolingual grammatical rule systems in DLT are almost entirely syntactic, that is, they concern all levels of **formal** analysis: word, sentence and text. On word level, a word formation analysis is required, which, however, for most languages cannot be done without reference to semantics. Most of the semantic information about the source and target languages is found in the bilingual dictionaries with the IL.

To sum up the language-related rule systems, it can be said that there are **three** blocks of monolingual syntax, **two** blocks of bilingual syntax and semantics and **one** block of monolingual semantics and pragmatics.

As a matter of fact, the three monolingual blocks contain syntax with a little area of semantics. For English, this block consists of:

- text syntax [6] which contains
  - sentence syntax [11] which contains
    - word syntax which contains
      - morphology [4] and
      - word formation [5] which exceeds syntax.

All these rule systems make use of the syntactic knowledge about words stored in

- a syntactic dictionary [3].

For French, exactly the same structure is found (with [29], [28], [32], [31] and [30], respectively). The structure of the syntax blocks for English and French is shown in Figure 4. For the IL, the syntax block has a similar structure, with text syntax [10], sentence syntax [9] and a syntactic dictionary [11], but instead of morphology and word formation, there is just one complex word grammar [13] which exceeds the limits of syntax.

The two blocks of bilingual rule systems each comprise a syntactic and a semantic part. These are the English-IL [8] and the IL-French metataxis [27]. They make use of the syntactic blocks of the two languages involved and in this way apply the syntactic knowledge found in the monolingual syntactic dictionaries. In addition, they interact with the bilingual dictionaries for English-IL [7] and IL-French [25], making use of the semantic knowledge stored there.

The only block of semantics and pragmatics is the world model expressed in the lexical knowledge bank in the IL [16]. It makes use of knowledge about the internal meaning system of that language, as well as

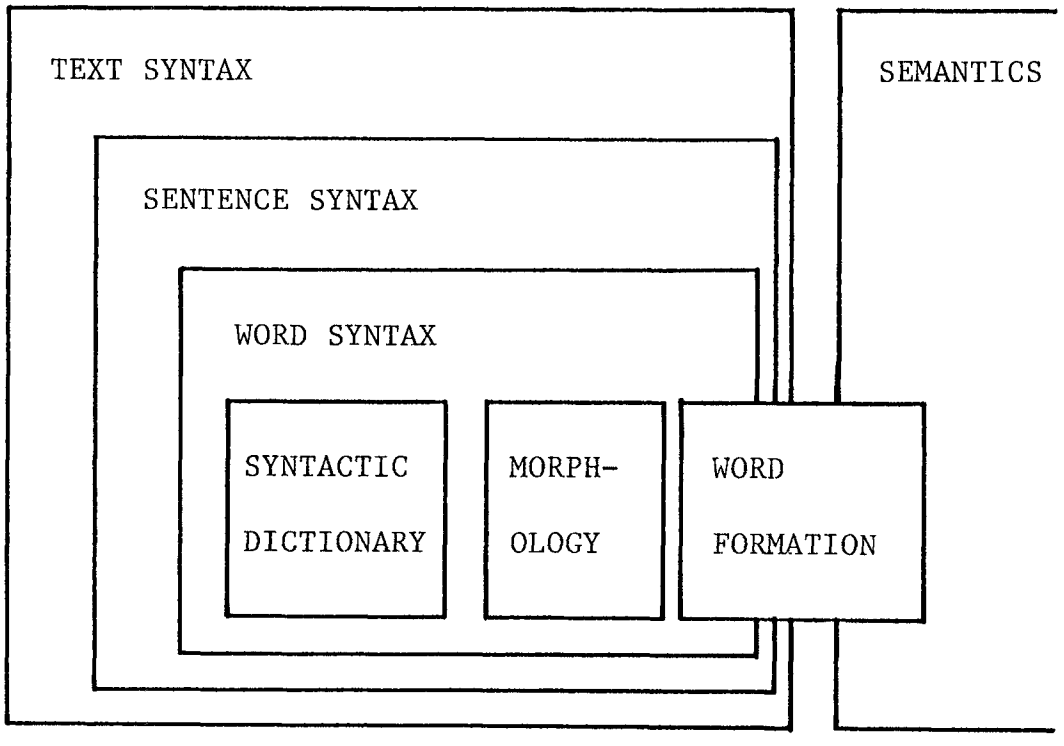

Figure 4: Relations among the syntactic systems of rules and knowledge.

of knowledge of the world expressed by means of the IL.

In order to show how these rule systems with their knowledge sources are put to work, it is necessary to look at some of the levels of computational application. On the levels of formalism and program (cf. 3), these rule systems are used in a way which is displayed in Figure 5.

5. Controversial features of the DLT system

In the preceding sections, I have postponed further argumentation about three characteristic features of DLT which reveal an undoubtedly somewhat unusual, and perhaps even controversial, approach to machine translation. These are:

- dependency syntax,
- the meta-representation of meaning,
- the intermediate language.

These topics are taken up in the following three sections.

5.1. Dependency syntax

As Mel'čuk (1979: 4) points out, there are only two ways of describing a syntactic system: Either the sentence (or text) is divided up into smaller pieces until the level of words is reached (constituency), or the relations among the words are studied until all the words in the sentence (or text) are linked together (dependency). (On the formalism level, these two approaches resemble the top-down and the bottom-up parsing technique, respectively, but there need not be any firm link between either of the syntaxes and either of the techniques).

DLT, as opposed to many other systems, adopts the dependency approach. This is not the place to give either a history of the dependency school (cf. Helbig 1973: 198; Schubert 1986a: 14; Nikula 1986: 14) or a list of previous applications of dependency grammars in natural-language processing. What can be given here is a selection of reasons why a dependency syntax lends itself to (automatic) translation especially well.

Both constituency and dependency syntax describe the same object, and they should, therefore, if all goes well, also provide the same information about their object, though in incompatible forms. If a constituency description of a sentence is complete, the dependency relations can be inferred, and vice versa. For a practical application such as machine translation, the question to ask is therefore not so much which of the two is better in an absolute sense, but rather which one arrives with less effort at the information required for the purpose.

In machine translation, syntactic analysis is performed with a very specific goal: to find an interface between the two languages. Translation-directed syntax should thus be suited to **contrastive comparison** and need not describe features which are irrelevant for this purpose. When translating syntactically arranged words (a text) from one language into another, the syntactic structures of the source language are transferred into those of the target language while words are replaced. It can be shown that the rules which cover this structural transfer, whatever the syntactic school by which they are written, must use a level of syntactic **relations** and cannot make do with the syntactic **form** of words alone.

For example, if what is traditionally known as an object of a finite verb in a sentence is translated from English into French, information is primarily needed whether the word group concerned is related to the verb as an object or as a subject or as something else, and only secondarily whether this object has the syntactic form of a noun phrase, an infinitival clause or the like. In order to **recognize** the word group as an object, it may be interesting to study word order, morphological form, etc., but these are language-specific, and the purpose of studying them is to establish the syntactic

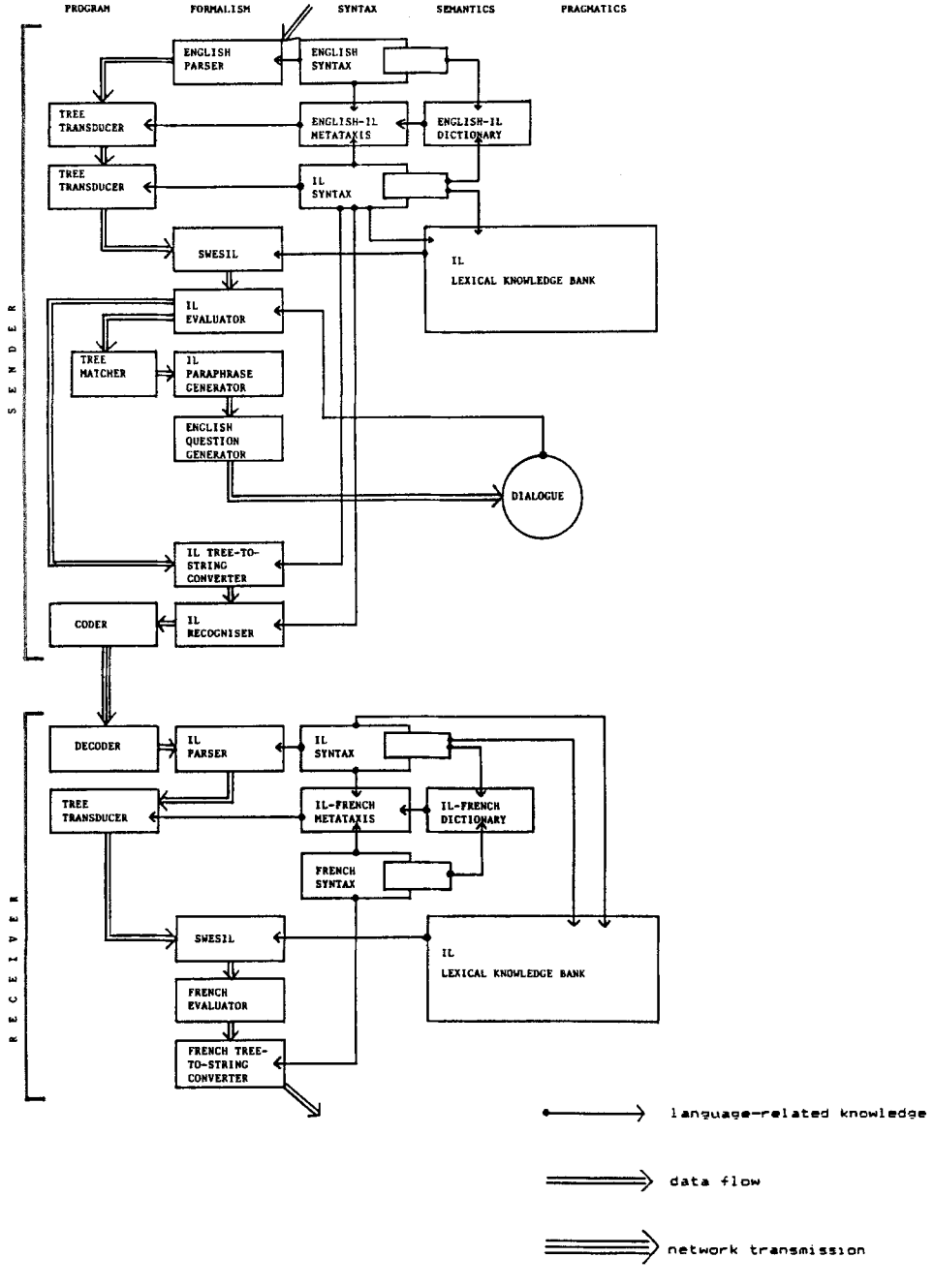

Figure 5: Overview of the translation process in the DLT system.

relation. In order to translate the English object into French, the primary information required is that it is an object, and second, how an object of the current sort of verb is transferred into French.

Not before the syntactic relation has been defined for the target language, can the morphological form, the word order and other features of word form be determined. They are language-specific in the target language as well, and are thus not inferable directly from the source language forms, but can only be found via an explicit or implicit description of the syntactic relation plus a transfer of relations plus target language-specific rules for syntactic word form.

This is only one example, but it shows perhaps that contrastive syntactic rules have their most straightforward interface on the level of syntactic relations. These relations are the primary object of dependency syntax, whereas in constituency approaches they are at best second-hand products of an analysis which has to proceed through consideration of many language-specific features, not all of which may be needed for the purpose of machine translation.

In addition, most of the current constituency approaches are too much committed to word order and turn out to be less applicable to the description of languages with "free" word order.

The concept of metataxis which describes the variant of contrastive syntax applied in DLT not only implies that preference is given to a dependency rather than to a constituency approach to syntax, but it also has another consequence which perhaps distinguishes DLT still more from other machine translation systems. The term metataxis seems also to mean that the main interface between two languages is sought on a syntactic level (and in the semantic information from a bilingual dictionary), rather than on the level of deep cases, semantic or thematic roles and the like.

What is so difficult about **semantic roles** is (among other problems) the impossibility of describing them explicitly in an inter-subjective way. In addition to this difficulty, they are not as language-independent as is usually maintained. This is an insight which is slowly gaining ground is general linguistics (cf. Pleines 1978: 372) and which is now turning up in the machine translation context too (e.g., Tsujii 1986: 656). The solution appears again to lie in avoiding making semantic roles of this type explicit. When a bilingual dictionary translates English *elephant* as German *Elefant*, it does not need to explain explicitly what an elephant is. For words this avoidance strategy is insufficient, but for semantic roles it is reasonable. Metataxis thus consists of rules which relate two syntactic relations of two different languages, stating implicitly that they denote the same semantic role. What this semantic role is, and whether the same or a slightly different role is found with other words, is not stated and need not be specified. Once the correspondence between two dependency relations of two languages is established, the semantic roles have played their part of tertium comparationis and are no longer relevant for the purposes of translation. This is the central idea behind Tesnière's (1959: 283) concept of metataxis.

5.2. A meta-representation of meaning

In section 2, above, I discussed how far meaning has to be analyzed for the needs of translation, concluding that some meta-representation of meaning is indispensable. Where is that meta-representation in DLT? Where are the predicators and arguments, modifiers and quantifiers, semantic primitives and speaker attitudes? The reader who has been waiting for an intricate system of tables and labels, flags and tags to occur, will be disappointed. There is nothing like that in DLT. Nevertheless, there are predicators and arguments, there are modifiers and quantifiers and there is much more. In DLT, however, they are not given

as symbols and abbreviations, but as morphemes and words with syntactic and semantic relations among them. In a word, **DLT's meta-representation is a human language:** Esperanto. (DLT's IL is Esperanto with some minor modifications in syntactic form). As Esperanto is quite commonly denoted as artificial, in 5.3 below I motivate my claim that Esperanto is a human language and should be considered non-artificial with respect to the criteria relevant here.

First, however, a few words on the question of whether a human language is suitable as a meta-representation of meaning. The question is not, of course, whether human languages are suited for representing meaning, because they are the best and by far the most prevalent means of expressing meaning man has. The question is whether a **meta**-representation of meaning for translation purposes should have other features than a human language. Neijt (1986: 12) claims that an intermediate language should be "more like a universal grammar - its elements being primitive notions - than like Esperanto, which closely resembles a natural language". The idea that research into language universals might become fruitful for machine translation is certainly interesting, but her assertion than an intermediate language should consist of "primitive notions" leads astray. Neijt borrows the title of her article intentionally from Andreev, who, however, demonstrates just the opposite of her opinion: "Human languages are much nearer to each other, than to symbols of any variation of a logical system, and consequently an effective IL must be sufficiently similar to spoken human PL's" (Andreev 1967: 5, PL = input and output languages, "paralanguages"). Andreev's opinion is based on long-term experience with intermediate structures. Ten years earlier he had still been arguing in much the way that Neijt does today, and he had then been much more optimistic about what man-made codes could perform in this field (Andreev 1957).

The main objection to Neijt's version of an intermediate language is that there are no "primitive notions". Every notion and every concept can always be easily divided into still more "primitive" ones. Moreover, even if it were feasible to divide up in advance every possible meaning into primitives, universally valid for all languages, an intermediate language made up of these primitives would **explode**, resulting in unimaginably huge dictionaries and a never-ending, but largely useless disambiguating process. Andreev has already pointed out the reason for this. Neijt (1986: 12) nevertheless demands excessive explicitness and maintains that "At present, the grammar of Esperanto, like all grammars of second language learners, does not provide the level of explicitness required for MT".

As far as the grammar of Esperanto is concerned, these words bear witness only to a deficient acquaintance with the findings of interlinguistics (cf. 5.3). However, they point to one important feature an intermediate language must not lack: it should be learnable. Machine translation systems are built by humans. The people who work with the meta-representation - syntax writers, semanticians, lexicographers, terminographers, etc. - should have full command of it. Furthermore, it is not sufficient that each of these persons (probably with different mother tongues and different cultural backgrounds) use the meta-representation consistently within his/her own work, but it is essential for the meaning-preserving goal of translation that they apply it in an inter-subjectively consistent way.

Such consistency can only be guaranteed by a language community which by convention defines and maintains the meaning of the arbitrary signs. It is because of this conventional link between the two sides of the linguistic sign that every artificial meta-

representation is inseparably bound to the human language in which it is defined (cf. Schubert 1986b:451).

Thanks to this bond of definition, an artificial symbol system can express meaning, but for the very same reason it cannot convey anything which could not be said in its human reference language. As Hjelmslev (1963:101) puts it, an artificial system can always be translated into a human language, **but not the other way round**. If Hjelmslev is right with regard to the latter restriction, all artificial symbol systems are inherently insufficient for application as an intermediate representation in (machine) translation, because they fall short of its essential function.

What, then, is the essential function of an intermediate (meta-)representation? It is not only to provide a rough classification of events, objects and phenomena, but, as Tucker and Nirenburg (1984:132) emphasize, to express the **full meaning** of the text to be translated. If translation is attempted with unrestricted meaning domain and from a variety of source languages not sharply delimited in advance, the intermediate representation must be a **language into which every human language can be translated**. This is, almost literally, Hjelmslev's definition of a **human language** (*dagligsprog*: 1963:101).

As the arguments quoted suggest, a meta-representation in a machine translation system should not be less, and cannot be essentially more, explicit than a human language. Ultimately, in short, only a human language can function as a sufficiently rich intermediate meta-representation of meaning.

5.3. The intermediate language

Even if the arguments in 5.2 are accepted as answering the question about the kind of intermediate language required, there are still two questions to be answered: Should there be an intermediate language at all; and, if so, is Esperanto a suitable choice for the purpose?

It has often been argued that an intermediate language makes a machine translation system **modular**. Adding a new source language means adding just one module, source language into intermediate, and no changes whatsoever need be made to the existing modules for other source and target languages. This is, however, only true if the intermediate language is grammatically fully independent, so that the intermediate form does not carry any features which restrict its translatability to certain target languages; nor must the translation into the target language require the intermediate form to bear any sign of which source language it comes from. In order to provide for modularity, an intermediate language must be **autonomous** (Schubert 1986a:12, 179). If it is not, the grammars of all source and target languages have to be "attuned" each time another language is added (Appelo 1986: 41).

In DLT, however, modularity is not the only use of the IL. As the previous presentation (cf. 3) has shown, the whole architecture of the system is centered around the idea of resolving in the IL kernel all problems of a **language-independent** nature. Knowledge of the world is a typical example of this. As far as lexical knowledge banks contain language-independent information, it is not efficient to build up a knowledge bank for each language separately.

In the long run, DLT's knowledge bank is to become a **learning system** which evaluates the information gathered from the texts processed and from the dialogue answers given (Papegaaij 1986:217). This is more efficient if there is only one knowledge bank at a place in the system through which all texts pass, rather than separate banks which learn only from the texts in one language.

Translating first into an intermediate language and then, in a second step, into the target language, might seem to double the

problem of automatic translation, which is already difficult enough without this requirement - so difficult indeed that it cannot be done fully automatically. However, the DLT environment necessitates fully automatic translation from the IL into the target language. For this reason, the intermediate language cannot be just any language, as Tucker and Nirenburg claim (1984: 132). It is obvious from the discussion in 5.2 that nothing less expressive than a human language will do, but on the other hand the intermediate language must necessarily **facilitate translation** in a decisive way.

This is the reason why DLT has chosen Esperanto. It has the full semantic **expressivity** of a human language, and it has such syntactic and semantic **regularity** and such **reliability** in its clearly structured **productive rules** that fully automatic translation **from** it is possible. (This is argued in more detail in Schubert 1986b.)

Is Esperanto a human or an artificial language? In the literature, one finds quite a number of severe judgements from both laymen and linguists about what Esperanto is not or cannot do. Neijt's words about the grammar of Esperanto (cf. 5.2) are only one, rather mild sample from the collection. With Esperanto standing on the threshold of its second century of human use, however, one can no longer discuss it in the same way as in the 1890's. After a century of experience with Esperanto, the serious linguist of today may be expected not to speculate about Esperanto, but to argue on the basis of actual knowledge. There is an abundance of literature available in both esperantology and general interlinguistics (cf. the annual interlinguistics chapter in the *MLA International Bibliography* and, for a very rich bibliographical overview, Blanke 1985: 296-381), and above all, of course, the language itself is open to study.

In the terminology of interlinguistics, Esperanto is an a posteriori planned language, which means that it takes elements (morphemes) from existing languages and arranges them in its own, autonomous way. It has thus come into existence in a way which can only in a very special sense be described as artificial, in the same sense, essentially, in which ethnic languages too have experienced periods of artificial language planning and standardization. Since the first publications in 1887, however, Esperanto has developed a great deal. As Sakaguchi (1983: 336) has shown, this process has followed closely the lines of **natural language development**. Esperanto is the only language project which has developed from a project into a **language** (Blanke 1985: 108). Such development requires a language community, and in the case of Esperanto, this community exists. Indeed, it is an unusual community, which deserves to be an attractive subject of sociolinguistic study rather than of unqualified judgements. This same speech community has enabled Esperanto to go on developing as a normal language. Analyses by Sakaguchi (1983: 347), Lo Jacomo (1981: 339) and others show how Esperanto, although developing in a community which lacks any controlling authority, has preserved an extremely high degree of standardization, which provides the special **translation-friendly** quality required for machine translation.

For DLT's application of Esperanto as the basis of its IL, it might seem irrelevant whether there is a language community or not. Human language, however, acquires its full expressivity and semantic precision not on the desk of a planner, but in a community. DLT can profit from a "test phase" of a hundred years for its IL: for Esperanto, there are dictionaries, terminological lists, grammars, linguistic literature, and, perhaps even more important, there are linguists, lexicographers and others who have full and long-standing command of the language and do not need to begin learning it the day they start work on DLT.

In the light of these considerations, DLT can be said to have as its IL a fully expressive and at the same time highly regular **human language**.

6. Grammar and world knowledge

Machine translation requires, as I have tried to demonstrate in this paper, language-related rules about linguistic and extra-linguistic knowledge. In section 2, I entered the discussion about these two types of knowledge from a somewhat different dividing line, the distinction of form and content. With the practical design requirements of a machine translation system in mind, it may be worth asking in this final section whether form or content rules should take **priority**.

The presentation in section 3 considered grammatical form, as described by **syntax**, as a means and **instrument of meaning processing**. The main problem proved to be word choice, and only when a solution to this problem had been found, could the subsequent rules about how to arrange the chosen words in a text, be applied. Syntax appeared there to be a preparatory tool for the processing of meaning, sorting out the words to arrange in pairs as the input to the word expert system. Syntactic redundancy rules were presented as a means of reducing the lexicographic and lexicon-handling workloads.

Another view is, however, possible. When meaning processing is supported by a preliminary syntactic analysis, by contrastive-syntactic metataxis rules, etc., these **syntactic rules** can also be seen as **prevailing over the semantic and pragmatic rules**. They are applied with the highest priority, and semantic-pragmatic choices are only taken into consideration where the syntactic form allows for them. From this point of view, syntax is the controlling rule system which can call upon meaning processing for solving well-defined residual problems. This order of rules is important when nonsense is to be translated. This is not at all an academic pastime, but a normal requirement for any type of text. Any text, of whatever style, can contain false claims, lies, unrealistic predictions and illogical utterances. The syntactic form, analyzed before any call to the knowledge bank, forces the semantic interpretation to go along the lines intended by the author and prevents the word expert system from destroying the text by rewording it according to some pre-formulated world knowledge.

Because the two sides of the linguistic sign cannot be separated, the question whether rules about the one or the other side are more important, is ultimately unresolvable. For the practical purposes of machine translation, the consequence should be an architecture of ordered and ranked rule applications, carefully designed on the basis of thorough theoretical insight into the relations between linguistic and extra-linguistic knowledge.

NOTE: The work reported here has been and still is being carried out by the entire DLT team. I thank them for their comments on this paper and for the loan of their native command of English and French.

REFERENCES

Andreev, N. D. 1957. Mašinnyj perevod i problema jazyka-posrednika. *Voprosy jazykoznanija* 6 [5]:117-121.

[Andreev] Andreyev, N.D. (1967): The intermediary language as the focal point of machine translation. In *Machine translation*. A. D. Booth (ed.). Amsterdam: North-Holland, pp. 1-27.

Appelo, Lisette. 1986. The machine translation system Rosetta. In *I. International Conference on the State of the Art in Machine Translation in America, Asia and Europe. Proceedings of IAI-MT86.* Tom C. Gerhardt (ed.). Saarbrücken: IAI/Eurotra-D, pp. 34-50.

Bates, Madeleine. 1978. The theory and practice of augmented transition network gram-

mars. In *Natural language communication with computers*. Leonard Bolc (ed.). Berlin/Heidelberg/New York: Springer, pp. 191-259.

Blanke, Detlev. 1985. *Internationale Plansprachen*. Berlin: Akademie-Verlag.

Doedens, Crist-Jan. To appear. On-screen ATN parsing. In *Proceedings of the International Conference on Machine and Machine-Aided Translation, Birmingham, April 1986.*

Engel, Ulrich. 1982. *Syntax der deutschen Gegenwartssprache*. Berlin: Schmidt [2nd ed.].

Helbig, Gerhard. 1973. *Geschichte der neueren Sprachwissenschaft*. München: Hueber [2nd ed.].

Hjelmslev, Louis. 1963. *Sproget*. København: Berlingske forlag [2nd ed.]

Klein, Wolfgang. 1971. *Parsing. Studien zur maschinellen Satzanalyse mit Abhängigkeitsgrammatiken und Transformationsgrammatiken*. Frankfurt a.M.: Athenäum.

Lo Jacomo, François. 1981. *Liberté ou autorité dans l'évolution de l'espéranto*. Pisa: Edistudio [doct. diss. Paris 1981].

Melby, Alan K. 1986. Lexical transfer: a missing element in linguistics theories. In *11th International Conference on Computational Linguistics. Proceedings of Coling '86.* Bonn: Institut für angewandte Kommunikations- und Sprachforschung, pp. 104-106.

Mel'čuk, Igor A. 1979. Dependency syntax. In Mel'čuk: *Studies in dependency syntax*. Ann Arbor: Karoma, pp. 3-21.

Munniksma, F. et al. 1975. *International business dictionary in nine languages*. Deventer/Antwerp: Kluwer.

Neijt, A. 1986. Esperanto as the focal point of machine translation. *Multilingua* 5:9-13.

Nikula, Henrik. 1986. *Dependensgrammatik*. Malmö: Liber.

1984 MLA international bibliography of books and articles on the modern languages and literatures. Vol. 3: *Linguistics*. New York: Modern Language Association 1985.

Papegaaij, B. C. 1986. *Word expert semantics: an interlingual knowledge-based approach*. V. Sadler and A. P. M. Witkam (eds.). Dordrecht/Riverton: Foris.

Papegaaij, B. C., V. Sadler, and A. P. M. Witkam. 1986. Experiments with an MT-directed lexical knowledge bank. In *11th International Conference on Computational Linguistics. Proceedings of Coling '86.* Bonn: Institut für angewandte Kommunikations- und Sprachforschung, pp. 432-434.

Pleines, Jochen. 1978. Ist der Universalitätsanspruch der Kasusgrammatik berechtigt? In *Valence, semantic case, and grammatical relations*. Werner Abraham (ed.). Amsterdam: Benjamins, pp. 355-376.

Sakaguchi, Alicja. 1983. Plansprachen zwischen Spontaneität und Standardisierung. Semiotik und Interlinguistik. In *Zeitschrift für Semiotik* 5:331-351.

Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Paris: Payot [new ed. 1969].

Schubert, Klaus. 1986a. *Syntactic tree structures in DLT*. Utrecht: BSO/Research.

_____. 1986b. Wo die Syntax im Wörterbuch steht. In *Pragmantax*. Armin Burkhardt/Karl-Hermann Körner (eds.). Tübingen: Niemeyer, pp. 449-458.

_____. To appear. Interlingual terminologies and compounds in the DLT project. In *Proceedings of the International Conference on Machine and Machine-Aided Translation, Birmingham, April 1986.*

Slocum, Jonathan. 1985. A survey of machine translation: its history, current status, and future prospects. In *Computational Linguistics* 11:1-17.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck [2nd ed. 1982].

Tomita, Masaru. 1986a. *Efficient parsing for natural language*. Boston/Dordrecht/Lancaster: Kluwer.

_____. 1986b. Sentence disambiguation by asking. *Computers and Translation* 1:39-51.

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

Tucker, Allen B. and Sergei Nirenburg. 1984. Machine translation: a contemporary view. In *Annual Review of Information Science and Technology* 19:129-160.

Waringhien, G. 1976. *Grand dictionnaire espéranto-français*. Paris: SAT-Amikaro.

Wells, J. C. 1985. *Concise Esperanto and English Dictionary*. Sevenoaks: Hodder and Stoughton.

Witkam, A. P. M. 1983. *Distributed language translation*. Utrecht: BSO.