DLT: a multilingual translation system [1]

Dan Maxwell
BSO/Research, Utrecht (Netherlands)

Translation by computer is a goal that has been pursued by scholars in various fields throughout the world since the end of World War II. But it has turned out to be a lot more difficult than most people thought. This is not just because languages have considerable differences in sentence structure, word structure, vocabulary, and sometimes writing systems; even in a single language, the proper interpretation of a given sentence or word often depends on context. But transforming this context dependence into a computer program is no easy matter.

In this article, I will describe a machine translation project which within a relatively short period of time has produced a prototype which translates texts dealing with aircraft maintenance from English to French. Plans have been made to extend this program to not only a variety of other kinds of informative texts, but also a wide range of other languages. I will be particularly concerned with the nature of the problems found in dealing with non-European languages, which are usually very different from those in more familiar languages that most Europeans have been exposed to - if only superficially - in school or as tourists.

1. An overview of the project

The name of the project is "Distributed Language Translation" (DLT) and the research has so far been carried out in the Dutch software firm BSO, mentioned in the heading of this paper. After a considerable amount of preparatory investigating in the early 1980's, mainly by the BSO engineer Toon Witkam, a total of 17 million guilders in funding was made available for a prototype project from the Netherlands department of economic affairs and BSO itself. Since 1985, a variety of linguists and computer scientists have been working at BSO to implement Witkam's ideas.

The word "distributed" in the title of the project refers to the intention of creating a program which can be used in personal computers in different parts of the world linked together by an electronic net. In the scenario made possible by such a program, a text is entered at one terminal of this net in the user's language and emerges in some language specified by a different user at another terminal. The only requirement is that both of the languages concerned must be part of the DLT system in the appropriate way.

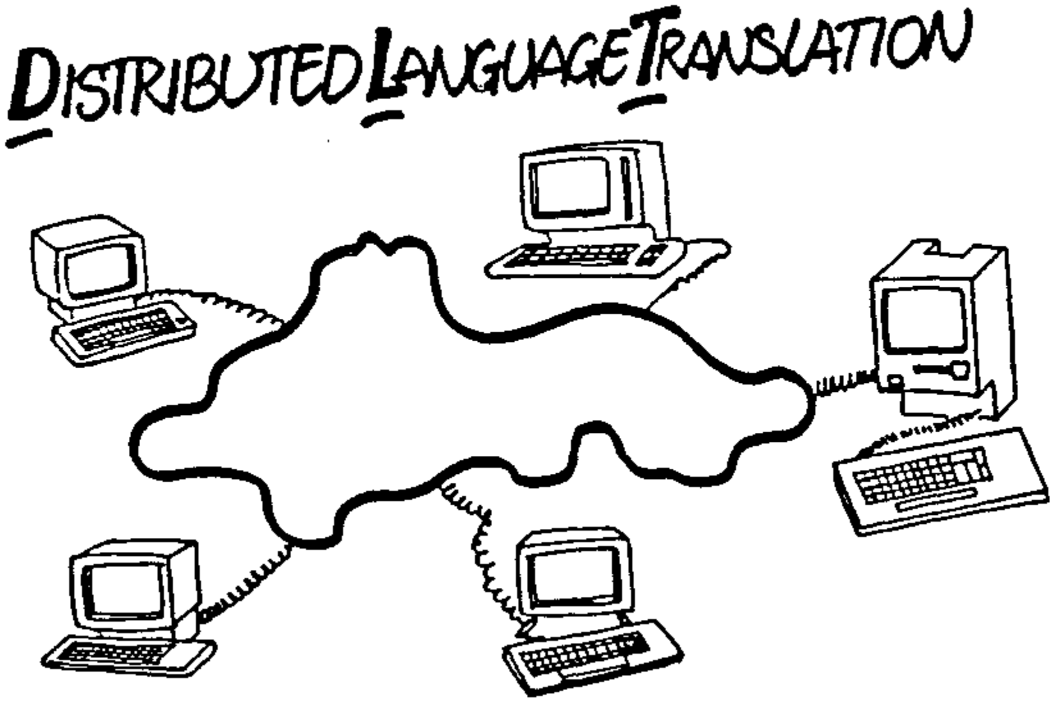
Let us use the common terms *source language* and *target* language, respectively, to refer to the two languages in the above scenario. In DLT, programs will be available to translate all source languages into the *intermediate language*, and to translate the intermediate language into every target language. In a translation system with a large number of languages, the use of such a link between the source languages and the target languages reduces the total number of programs necessary to translate every language into every other language.

In DLT, the intermediate language is the planned language Esperanto, except for a few fairly insignificant modifications. Esperanto has now been in existence for slightly more than a century and is now spoken by people all over the world, although the number in any one country is probably never higher than about 10,000. As an intermediate language in machine translation, it has the advantages of being systematic in structure, autonomous, and fully expressive[1].

After being checked for spelling and elementary errors of grammar, the input sentence in the source language is given to the *parser* for analysis. The result of this program is a *tree structure* which shows the relationships between the words in the sentence. There is often more than one such tree, because the purely grammatical information available to the parser may make more than one analysis possible. DLT trees are based on "dependency grammar", which was initially developed by the French linguist Tesnière[2] and consist of nodes and branches. With certain systematic exceptions (see section 2), a word is located at each node, and a label showing the relationship between each pair of connected words is found on the connecting branch, as in examples later in this article.

The resulting trees are passed to the *metataxis*, which performs the actual translation to Esperanto. The word metataxis is taken from Tesnière and means "structural change in translation." Metataxis makes use of grammatical information concerning both languages as well as a bilingual dictionary. Languages can differ from each other not only with respect to the meanings of individual words, but also with respect to the structure of individual words, phrases, and sentences. There will accordingly be a set of rules which deal with these individual differences between any given source language and Esperanto, and some subset of them will be necessary to translate any given sentence.

The metataxis will often increase the number of trees created by the parser, since there are often several acceptable ways of translating a given sentence, or at least choosing between them has to be done on the basis of context. Here is an example, which serves at the same time to demonstrate the use of dependency grammar as a tool for analysing a sentence.



When the bell rings, the students put their examinations in the boxes made available by the professors.

```
                              put<present>

       E-PROA         E-SUBJ         E-OBJ       E-PREC
        /               |              |            \
     when            students       examinations    in
       |                |              |             \
     E-SUB            E-DET          E-DET          E-PARG
       |                |              |               \
     ring<present>    the            their            boxes
       |                               |              /    \
     E-SUBJ                          E-DET           /    E-ATR2
       |                                            /        |
     bell                                        the       make<pap>
       |                                                   /    \
     E-DET                                           E-PRED    E-PREC
       |                                               /          \
     the                                         available         by
                                                                   |
                                                                 E-PARG
                                                                   |
                                                               professors
                                                                   |
                                                                 E-DET
                                                                   |
                                                                  the
```

<pap> = 'passive participle'

The most likely translation for this in Esperanto is:

Kiam la sonor'il'o son'as, la student'o'j met'as si'a'j'n ekzamen'o'j'n en la kest'o'j'n dispon'ig'it'a'j'n far'e-de la profesor'o'j.

Note the use of ' to show internal divisions of words. This is undoubtedly the most visible difference between everyday Esperanto and our use of Esperanto as an intermediate language.

The given translation of *their* is correct if it refers to the students, as would normally be expected. If a somewhat unusual context makes it clear that someone else has written the examinations, then *their* should be translated as *ili'a'j*. The metataxis would make both translations available.

The above translation into Esperanto involves mostly word for word translation, requiring no structural changes. The one exception is *made available*, which becomes the single word *dispon'ig'it'a'j'n* in Esperanto. This means that the Esperanto tree has no branch or node corresponding to E-PRED and the word below it in the English tree.

A further problem is that any given source language bilingual dictionary often provides several Esperanto words or phrases to translate a specific source language word or phrase. The metataxis simply puts all the possibilities consisting of one word at the corresponding node of the developing Esperanto tree. A new tree is created if a single word is translated by a phrase, or vice-versa, since this necessarily involves adding or eliminating structure.

It is clear that the result of the metataxis may be many Esperanto trees, any of which may have several alternative words at some of the nodes. The next step is to reduce the number of trees and the number of words at any given node, with the ideal result being a single tree with just one word at each node. This task is carried out by SWESIL (Semantic Word Expert System in the Intermediate Language). SWESIL makes use of information in the Lexical Knowledge Bank (LKB) to pick the most suitable tree and the most suitable word at any given node. See Papegaaij[3] for a detailed (though now slightly out-of-date) description of SWESIL.

It is often the case that SWESIL does not eliminate all the alternative possibilities, largely because there is not yet enough information in the LKB to make a reliable choice possible. Another reason might be that the choices are sometimes dependent on context, but DLT does not yet have a module which deals with levels of text above the sentence. If choices still need to be made after application of SWESIL, the dialogue is activated. For each of the remaining Esperanto alternatives, this module provides paraphrases in the source language and requests the user (the author of the text) to choose the most appropriate one. Assuming that the user cooperates and understands what he or she has written, the result will be a single tree with one word at each node.

At this point a second metataxis module is activated, this one designed to translate Esperanto to the target language. There may of course be several target languages, and in that case all of the appropriate metataxis modules are activated. Again, more than one tree may be created, and there may be several words at some of the nodes. With the help of contextual clues in the bilingual dictionary and the fact that Esperanto by its nature provides a relatively clear representation of meaning, SWESIL this time eliminates all but one possibility. A linearisation algorithm transforms this target language tree to a string of words.

2. Application of DLT to non-European languages

One initial caveat must be granted: the DLT developers themselves make no attempt to deal with writing systems other than the Latin alphabet. The subsequent discussion of languages which use such systems presupposes that someone else can work out a way of dealing with these and that such a solution can be linked to DLT, perhaps as a provisional step which transforms a text in such a writing system into the Latin alphabet. Even languages which do not use this alphabet in everyday written communication often have such an alternative available, often created by western linguists who wished to make information about the language available to a larger audience. But automating such a transformation is of course not easy, especially for languages such as Chinese, in which the symbols do not represent sounds and accordingly have no direct correspondence with the Latin alphabet. There are nevertheless many specialists working on such problems, and there is even software on the market for this purpose.

Assuming that the problem of scripts will be solved, DLT can be applied to non-European languages in the same way as to other languages. The first step is to produce a dependency grammar which covers a reasonable portion of the language concerned. With this goal in mind, we have – largely through the Esperanto movement – established contact with linguists who either speak these languages as their mother tongue or have studied them intensively and have lived in a country where they are spoken. A dependency syntax which suits our needs is already available for Arabic, Bengali, Finnish, Hungarian, and Japanese, which are all either spoken outside of Europe or are of non-European origin:

Progress has also been made towards a dependency syntax of Chinese, and initial contacts have been established concerning a few other non-European languages. The information in the dependency syntax is used to construct the parser, which in turn builds the tree structures mentioned above.

What are the particular difficulties in dealing with such languages, and how does the DLT system help make it possible to deal with them? My experience in working with non-European languages suggests that the differences which do exist probably lead to a greater amount of ambiguity when translating them to and from Esperanto, but this difference appears to be quantitative rather than qualitative.

Here are several concrete ways in which they may differ from more familiar languages, even if we are concerned only with written language, as in DLT:

(i) They have a somewhat different set of syntactic categories. Although all of these languages have nouns, verbs, adjectives, and adverbs, their other categories tend to be quite different from what westerners are used to. They generally lack definite and indefinite articles and an explicit future tense, for example. On the other hand, Japanese has a number of verbal suffixes which indicate the "probability" of the proposition expressed. Chinese and Bengali have "classifiers" with nouns.

(ii) The order of words is different. The direct object generally precedes the verb instead of following it in Japanese and Bengali; their prepositions precede the noun they go with instead of following it (for this reason they are called postpositions); their helping verbs (similar to *have* and *be* in English) follow the main verb instead of preceding it. In Japanese, particles equivalent to subordinating conjunctions come at the end of a clause instead of at its beginning.

(iii) Morphemes are put together as a single word in some languages, but treated as separate words in others. We will see interesting examples of this in Hungarian.

(iv) Partly as a result of variations in culture, the packaging of concepts in the form of words is often very different. One example of this phenomenon often cited by linguists is the case of the concept of snow. Eskimo languages have many words with *snow* as an essential component of the meaning. These are distinct in ways that Europeans would consider not essential enough to justify a separate word. Languages in central Africa, on the other hand, may have no word for snow at all.

In dealing with (i), it is necessary to recognize that a category which is found in language A but in absent in language B, may nevertheless be translatable in language B by some other category. Or, if it is not, it may not require an overt translation, instead being implicit in some feature of the context. Assuming that the text is translatable at all, then within the DLT framework it is in principle not difficult to write metataxis rules which change the structure provided by the parser in the source language to the corresponding structure found in Esperanto or similarly from Esperanto to the target language.

Suppose, for example, that it is necessary to translate a specific verb-object combination in the source language as a single verb in Esperanto. This would be the case if the source language is English and the text to be translated includes the phrase *give a hand*, which in its idiomatic meaning must be translated as *help'i* in Esperanto. In that case, the necessary change in terms of dependency tree structures can be represented as follows:

```
give
 |
E-OBJ                          help'i
 |
hand
 |
E-ATR
 |
 a
```

This much of the structure will be the same, regardless of whether the actual sentence has the phrase *gave a hand*, *will give a hand*, *has given a hand*, etc. Other rules dealing with these different tenses will ensure the proper form in Esperanto. The infinitival ending -i actually given in the above example is simply the ending in the dictionary. Another point is that the above structure will be used even if the words in the actual sentence are not adjacent to each other, as for example, in *give your friend next door a hand*. The basic relationship between the words *give* and *hand* remains constant, and it is accordingly the job of the parser to construct the above English tree fragment in all these cases. A separate linearisation algorithm is used when it is necessary to create a string of words from the tree.

This brings us to point (ii) above. The essential elements of a transitive sentence are subject (S), verb (V), and object (O). The order of these elements varies from one language to another, but 99% of the world's languages[4] have one of three of these orders as the default order: S-V-O (e.g. Jack saw Jill), S-O-V (Jack Jill saw), or V-S-O (saw Jack Jill). All of these are represented in the languages mentioned at the beginning of this section, as the following table shows:

S-V-O	S-O-V	V-S-O
Chinese	Japanese	Arabic
Finnish	Hungarian	
Russian	Bengali	

But whatever the default word order, the part of the dependency tree showing the syntactic relationships between verb subject and object is the same. This separation of functions — trees to show syntactic relationships and separate linearisation algorithms to deal with word order — makes the actual translation step dealt with in the metataxis considerably easier. In the above example, the simplest kind of metataxis rules is sufficient: the subject is translated as the subject, the verb as the verb, and so on, but the linearisation rules, based on the syntactic functions mentioned in the tree, will produce the differences shown in the above chart.

The following example from Hungarian[5] demonstrates how point (iii) is dealt with in DLT: sequences of morphemes within a single word are often placed on separate nodes (dependency labels are omitted here and in subsequent examples).

beszélhetek a könyvemről.

'I can speak about my book.'

```
       -het-                              can
       /   \                             /   \
   -ek    beszél                        I    speak
             \                                 |
            -röl                             about
               \                               |
               könyv                          book
                / \                          /    \
               a  -em                      the    of-mine
```

The hyphens in the tree serve to indicate which morphemes must be attached to another morpheme and on which side this attachment must take place. The corresponding English tree (used for expository purposes, since the Esperanto tree has the same structure) makes it evident how such an arrangement makes direct morpheme by morpheme translation possible in many cases, although this would not be the case, if we always avoided multi-node treatments of complex words.

The above example shows a case of a single word being spread over more than one node. The converse of this situation, in which a single node lacks anything at all, is also possible. There is a class of sentences in some non-European languages which lack a verb. These are predicative sentences in the present tense such as "This is Alī" In Arabic, for example, is translated as *haadaa alī*, literally "this Alī"[6]. We treat this by leaving the top node of the tree, normally reserved for the finite verb, entirely empty:

```
             0
            / \
      hadaa     Alī
```

This is not a strictly necessary change, but does regularize the paradigm for such sentences and simplifies the metataxis rules to some extent. The following sentences in Bengali [7] serve to demonstrate how an apparently subtle distinction in one language can in fact be the source of a major structural difference in another.

(a)	hiteS	upenke	khObor	Ta	bujhie	di	lo
	Hitesh	Upen-to	news	Cl	explain-conj	give/let	past

'Hitesh explained the news to Upen'

(b)	hiteS	Upenke	khobor	Ta	bojhate	di	lo
	Hitesh	Upen-to	news	Cl	explain-inf	give/let	past

'Hitesh permitted Upen to explain the news'

The only difference between these two sentences is the form of the verb meaning 'explain'. Our syntactically-based analysis makes the provision that if this verb is in the so-called conjunctive form, as in (a), it governs the complements (Subject, Object, Indirect Object), i.e. it in effect becomes the main verb of the sentence. If it is not in this form, then the verb which governs it (in this case *dD* governs whichever complements are in its valency, except that the converbial itself takes the place of the direct object. The difference in terms of dependency trees is as follows:

```
(a)             di               (b)             di
              /    \                     /     / |  \
            lo   bojhie               lo  HiteSh ke  bojhate
                /  |  \                          |    |
            Hitje  ke  ta                      Upen   ta
                   |    \                             |
                 Upen   khObor                       khObor
```

Once this difference in structure is established by the parser, it is an easy matter for the metataxis to transfer this difference from Bengali to Esperanto or vice-versa.

This sentence simultaneously illustrates some of the other problems discussed previously: the word order is very different from that found in many European languages, though not so different from, for example, subordinate clauses in German. The classifier *ta* is simply left untranslated in the intermediate language and in European languages generally. If Bengali is the source module, our still to be developed text grammar will have to face the problem of determining when to insert a definite article.

Finally we are ready to deal with point (iv). While no-one would claim that Esperanto has words covering all the concepts of all the languages to be dealt with by DLT, the ones that it does have make it possible to represent the meaning of the sentence relatively clearly. This is because Esperanto comes closer than most other languages to being *compositional.* The principle of compositionality was postulated by the German philosopher Frege* and asserts that the meaning of any unit of speech is some function of the meaning of its individual parts. Our ability to make up and understand sentences that we have never heard before is the best evidence of this property, but linguists have been long aware that natural languages also have idioms, in which the meaning of the whole unit can not generally be composed on the basis of the meanings of its parts, and words like *huckleberry*, in which one part of the word has no meaning at all outside of the word itself. Esperanto as a planned language achieves a greater degree of compositionality than other languages by being generally free of such properties. The greater degree of compositionality also means that Esperanto can more freely combine its units of meaning than other languages to create new words. This makes it a relatively powerful tool for the purpose of expressing the nuances of meaning which seem significant in a given text as well as for dealing with new concepts that are constantly being developed in all languages.

3.Conclusion

Although it is not possible at this stage to foresee with absolute certainty how well DLT is able to deal with the problems of automatic translation in non-European languages, the issues addressed here suggest that the modularity of the DLT system makes it possible to break up the large problems which arise in dealing with such languages into a series of smaller problems. I have also argued that two of the tools used in DLT – dependency syntax and Esperanto – have intrinsic properties which provide a relatively natural solution to at least a large number of these problems.

References:

[1] SCHUBERT, Klaus Ausdruckskraft und Regelmäßigkeit: was Esperanto fur automatische Übersetzung geeignet macht. *Language Problems and Language Planning* 12 1988; 130-147.

[2] TESNIÈRE, Lucien *Éléments de syntaxe structurale.* [2nd ed.] Paris: Klincksieck, 1962

[3] PAPEGAAIJ, Bart. *Word Expert Semantics: an Interlingual Knowledge-Based Approach.* Dordrecht: Foris, 1986.

[4] GREENBERG, Joseph H. Some universals of grammar, with particular attention to the order of meaningful units. in *Universals of Language* ed. by Joseph H. Greenberg Cambridge, Mass.: MIT Press, 1966; 73-113.

[5] PRÓSZÉKY, Gabor, KOUTNY, Ilona and VACHA, Balász *A dependency grammar of Hungarian.* Budapest, 1987. [Manuscript]

[6] TOSCO, Mauro *A Dependency Syntax of Arabic.* Ms., Genova, 1988. [Manuscript]

[7] DASGUPTA, Probal *A dependency syntax of Bangla.* Pune, 1988. [Manuscript]

[8] FREGE, Gottlob *Über Sinn und Bedeutung.* 1892; translated as: On sense and reference in *Translations from the philosophical writings of Gottlob Frege,* ed. by Peter Geach and Max Black, Oxford, 1952.