A PROTOTYPE MACHINE TRANSLATION BASED ON EXTRACTS FROM DATA PROCESSING MANUALS

E. Luctkens Ph. Fermont
Department of Information Science and Documentation
Free University of Brussels
Belgium

The following article presents a prototype for the machine translation of English into French. The study was carried out over a period of nine months, following a six months preliminary study, under contract with the Burroughs Company and using a micro-computer of the B20 series.

The prototype aims to provide a diagnostic study that lays the foundations for further development rather than immediately producing an accurate but limited realisation.

By way of experiment, the corpus for translation was based on selected extracts from computer systems manuals. After studying the basic material, as well as assessing the various decision criteria, it was decided to construct a prototype made up of three components : analysis, transfer and generation.

Although the prototype was designed with multilingual applications in mind, it appeared preferable at this stage not to set up a system with interlingua since the elaboration of the interlingua alone would have taken up a disproportionate amount of time (King, Perschke, 1984), thus handicapping the development of the prototype itself.

1. General outline of the prototype

General outline — Prototype

SL text — Analysis

Preprocessing...formating of text with a view to further processing

+Morph. anal.....not envisaged for the moment

+Synt. anal.....ATN to produce a deep structure

+Disambiguation..not envisaged for the moment

Transfer

+Lex. transfer...morphemic translation

Str. transfer...adaptation of the parse tree to generation in the TL

Generation

Synt. synth....generation of surface structures linked with SL

+Morph. synth...rules of agreement, conjugation,...

TL text — Post-editing...in the first stage, use of the B20 text processor

+ : sub-components with dictionary look-up

2. "Analysis" component

In the prototype, the "analysis" component uses only three of the above sub-components: preprocessing, source-language dictionary and syntactical parser. Reasons for not using morphological analysis and desambiguation are given below.

2.1. Preprocessing

The preprocessing sub-component recognizes which sentences to analyse, a sentence being considered as a series of signs which are themselves grouped together in words, and ending in a full stop. The latter is the only special sign which is taken into account. Moreover, all the capital letters placed at the beginning of sentences are converted to the lower case before analysis and are reintroduced during generation.
One could envisage allowing for punctuation signs when parsing, since these sometimes help to root out ambiguities of certain sentences. A study is currently considering this.

2.2. Morphological analysis

As the prototype was being realised based on and for a limited corpus, the SL dictionary was made up of complete forms : the working out of a morphological parser is simpler than that of a syntactical parser.

2.3. Syntactical analysis

The Augmented Transition Network (ATN) was selected for the analysis : it had successfully been used in many previous systems : LUNAR, SHRDLU, INTELLECT and, more recently, ENGSPAN (Leon, 1984). T. Winograd presents three networks in great detail in his book 'Language as a Cognitive Process' (Winograd, 1983). These were taken as the basis for the four (Sentence, Noun Phrase, Prepositional Phrase and Adjectival Phrase) of the prototype, thus making it possible to speed up the development of a parser which had already proved itself in other respects.

The majority of the modifications made to the Winograd's ATN were aimed at increasing its performance (especially by dealing with the most common cases of coordination) as well as its determinist capacities thereby ensuring the accuracy of the initial analysis supplied by the system (it is in fact on this analysis that the transfer operates because the micro-computer's memory was satura-

ted before it had managed to supply all possible analysis).

2.4. Disambiguation

Within the prototype framework, the creation of a disambiguation sub-component would have taken up too much time and would not have been useful particularly that this research is deliberately designed to apply to only a limited corpus in which most of the ambiguities concern the Prepositional Phrase attachment and need not be solved for the translation if English into French.

2.5. Source-Language dictionary

For the various reasons explained above, the dictionary includes only complete forms.

All variable words are characterised by different syntactical features, certain of which concern their form, others do not. All of these are treated by the analysis component. Semantic features could easily be added at a later stage.

Words forming certain 'traditional' classes may belong to various categories of the prototype dictionary. This is notably the case with cardinal adjectives, which are at once classified as determiners and substantives.

At present, the only compounds that the prototype dictionary accepts are locutions with a maximum of two consecutive words. Longer locutions, compound verbs and other discontinuous compounds, quite rare in the corpus, will be treated as follows at a later stage: all words liable to appear in compounds will be tagged with a pointer to this effect, to enable the preprocessing sub-component to determine whether a compound or simple form is present in a given text.

Numbers were not introduced into the prototype dictionary. The parser would accept them if a routine were created that would automatically attribute noun and determiner categories to them.

3. "Transfer" component

The transfer component deals with the results obtained by the analysis component.

3.1. Structural transfer

By dealing with the structural transfer first, one is saved, notably, from having to waste time translating forms that will duly be dropped (such as 'will'), since the adaptation to tense in French is done along with the structural transfer.

The structural transfer operates on the sentence as a whole, on various levels. It only saves those results of the analysis that are pertinent for the generation.

3.1.1. Sentence

The various constituent elements of the clause are rewritten so as to conform to the following sequence :

(Passive) + (Negative) + Role + NP1 + Auxiliaries + Verb + (NP2) + (NP3) + PP

NP1 is the deep subject of the clause, NP2 is the direct object (the attribute or even nothing at all if the main verb is of the 'be' type) and NP3 is the indirect object.

All passive clauses are put into the active voice during the analysis and structural transfer. These are the transformations that, where necessary, regain the passive voice in the process of generation into French.

3.1.2. Noun Phrase

Three rewrites are possible for the noun phrase :

- Number + Pronoun
- DNP ('dummy NP')
- Number + ((Determiner) + Noun + (Adjective) + (Noun) + (PP) + (S))

The rewrite elements are derived from various registers of the analysis result.

3.1.3. Verb Phrase

By Verb Phrase is understood here the Auxiliary together with the Main Verb. This involves 'Auxiliary' in its widest sense, that is comprising all that precedes the verb : tense (present, infinitive and/or imperfect), modality and even person. It should be noted that only third person forms appear in the corpus studied.
The verb phrase rewrites itself extensively in the following manner :

(Infinitive) + Present/Imperfect + 3rd.p + (Avoir/Etre + Past Participle) + (Modal) + (Avoir/Etre + Past Participle) + Verb

To arrive at this rewrite, many rules that combine together are brought into play for various reasons concerning, notably, the multiple feature categories, the treatment of 'be', 'dummy be' and 'dummy modal'.

3.2. Transfer dictionary

In English as in other languages, a word may belong to several grammatical categories ('all' is at once adverb, determiner and pronoun) or, indeed, the same form may have various dimensions ('read' has the features of infinitive, present (except for the 3rd person in the singular), and past as well as past participle). Besides, one word in English may have several possible translations in French. For these reasons, it seemed convenient to create a transfer dictionary situated in between source and target language dictionaries in order to avoid excessive multiplication of relationships and also to facilitate the extension of the system to other language pairs.
Unlike the English terms which are in the

dictionary in a complete form, their French translations are presented in canonical form.

3.3. Lexical transfer

Lexical transfer operates directly after the structural transfer. At the moment, it is always the first translation (when there are several possibilities) that is chosen.

One could envisage adapting various means of selecting the best translation, ranging from the human operator to the style index.

4. "Generation" component

The generation or synthesis takes place in two stages : the syntactical generation is followed by the morphological generation. Both of these stages refer to data from the target-language dictionary as well as from the common data pool.

The generation in French is inspired by the rules of Chomskian generative and transformational grammar, specifically as presented in the work of C. Nique (Nique, 1978).
Most of the other grammatical theories currently in vogue (Montagovian Grammar, Generalized Phrase Structure Grammar, ...) make wide use of semantics and thus necessitate far more powerful computer resources than those available on micro-computers at present.

4.1. Target-Language dictionary

In the target-language dictionary, the different features allowing for the agreement of the canonical forms must be added to the various grammatical categories.

A common data pool is associated with this dictionary. This enables one to conjugate the verbs correctly (root table and conjugation table). It also contains the different forms of the determiners and their conditions of usage.

4.2. Syntactical generation

The generation is carried out by means of transformations. Below are presented those transformations that have a fundamental role in the elaboration of the structure of the sentence in French and in the ordering of its terms. Others directly concern the morphology of the words, and are outlined briefly later on.

In accordance with the theory of generative and transformational grammar, transformations occur in an orderly manner in an ascending cycle, that is to say from the inside, outwards, starting with the most subordinate clauses.

Passive Transformation :
e.g. : The entire field of booleans can be treated - active deep structure - Le champ entier de booléens peut être traité.

Transformation of Negation :
e.g. : Each name is an identifier which cannot be allocated - positive deep structure - Chaque nom est un identifieur qui ne peut pas être alloué.

Transformation of Subordination, which correctly inserts the subordinate clauses :
e.g. : Each bit may be used to store a logical value - Chaque bit peut être employé pour mémoriser une valeur logique.

Auxiliary Transformation :
- if, in the rewrite of the verbal phrase, Avoir/Etre occur, the appropriate auxiliary is chosen depending on the feature specified in the target-language dictionary.

Transformation Movement of the Adverb :
e.g. : A virtual field item always occupies an integral number of 4-bit digits - Un article virtuel du champ occupe toujours un nombre entier de chiffres de quatre bits.

4.3. Morphological generation

The morphological generation is made up of the following transformations : subject-verb agreement, conjugation, noun qualifier (which inserts 'de le' between a noun and its complement), insertion of determiner, noun agreement, determiner agreement, adjective agreement, placement of adjective, elision and contraction.

5. Conclusion

The results obtained over a relatively brief period by a team of two researchers may be considered as encouraging and tend to be optimistic as to the future of machine translation or machine-aided translation on small systems.

References

KING (M.), PERSCHKE (S.). - Eurotra. - Lugano, April 1984.

LEON (M.). - Development of English-Spanish Machine Translation. - Cranfield, 1984.

NIQUE (C.). - Initiation à la grammaire générative. - Paris, Colin, 1978. - 176 p.

NIQUE (C.). - Grammaire générative : hypothèses et argumentations. - Paris, Colin, 1978. - 207 p.

WINOGRAD (T.). - Language as a Cognitive Process, Syntax. - London, Addison-Wesley, 1983. - 640 p.