INTERFACE. Journal of Applied Linguistics 3.1 (1988), 45-56

THE TREATMENT OF FORM DETERMINATION FOR FRENCH IN DLT

Dorine TAMIS
BSO/Research, Utrecht

ABSTRACT

A word in a sentence can determine the syntactic (mostly morphological) form of another word in the same sentence or even beyond the limit of the sentence. A machine translation system should deal with this phenomenon of form determination. The aim of the present paper is to give an overview of the machine translation system DLT and to discuss the treatment of form determination for French, the first target language of DLT, within the framework of dependency grammar.

1. INTRODUCTION

Using correct and grammatical forms of words and endings is an essential "detail" in translation. Words in languages can have a wide variety of inflected forms and the human writer or translator should correctly recognize and use them. In this paper I would like to discuss the place of these word forms within a machine translation system and the functions stems and endings can fulfill.

A word in a sentence can determine the syntactic (mostly morphological) form of another word in the same sentence or even in another sentence. This phenomenon of form determination has two separate functions in a machine translation system. In the analysis part of a machine translation system, form determination suggests itself as a guideline for recognizing the syntactic functions of words in order to analyze the sentence to be translated. In the synthesis part of a machine translation system, it is the other way round. The syntactic functions of the words are already known and are presented in a syntactic tree structure. Now, the correct syntactic form of the words has to be generated in order to express their syntactic function in the linear sentence.

DLT's approach to translation is based on dependency grammar (Tesnière, 1959). Dependency grammar tries for each language to make a classification of word classes and a list of possible syntactic relations or dependencies between these syntactic categories. The result of sentence analysis in dependency grammar consists of a dependency tree, a structure

which shows the syntactic relations between the words of the sentence. The syntactic form of words is used to establish dependency relations between them. Form determination can be an important criterion (but not the only one) for establishing co-occurrence lines between the words of a sentence in order to create a syntactic dependency representation of the sentence. Usually, the syntactic form of words concerns not only those elements of words which are expressed by means of certain morphemes, but also others that can be traced in a certain dependency or in form determination. These characteristics and morphologically expressed syntactic form can be called syntactic features (Schubert 1987: 152). Form determination has a determining and a determined word (or several determined words), syntactically related words. A distinction is made between two types of form determination:

1. Form government: Words receive their syntactic form as a result of their syntactic functions. For example, in German, verbs and prepositions can "govern" certain cases in the nouns and pronouns that co-occur with them (Schubert 1987: 31):

   Example 1: form government

Sie	hilft	ihm.	Sie	unterstützt	ihn.
nominative		dative	nominative		accusative
'she	helps	him'	'she	supports	him'

2. In addition to form government, traditional grammar makes use of the notion of agreement, i.e. formal correspondence. In French, for example, the determiner, adjective and noun agree in number and gender in the same noun phrase: they share the same syntactic features. The function of the gender in the syntax here is to obtain a specific syntactic form of the word itself and of the related words.

   Example 2: agreement

une	grande	maison
[+singular]	[+singular]	[+singular]
[+feminine]	[+feminine]	[+feminine]
'a	large	house'

Before discussing in more detail the treatment of syntactic form during the synthesis part of the DLT system and in particular for French, I will give an overview of the whole system.

2. THE DLT SYSTEM

2.1. A general overview

DLT (Distributed Language Translation) is a machine translation system

under development at the research department of the Dutch software house BSO. DLT is intended to become a system for multilingual semi-automatic machine translation with a monolingual interactive dialogue with the user. It is designed for use in personal computers in a network environment and is therefore set up to work without post-editing. The first source language (SL) under development is English and the first target language (TL) is French, but the system is easily extensible to other languages. The translation process is called distributed because it is split up into two independent translation processes. This approach is based on the network idea which makes communication between computers at various places in the world possible, so the system can be linked to any number of different places. The first translation process of the system translates from a given source language into the intermediate language (IL). This is performed at the sender site of the network. The intermediate translation in the interlingua is sent over the network to the receiver site where the second translation process takes place: the translation from the intermediate language into the intended target language. The intermediate form of a text can be translated into any target language and is thus in no way dependent on the source language: the send and receive processes are independent of each other. The intermediate language has to be fully expressive, clear and autonomous with regard to the possible source and target languages. The interlingua used in the DLT system is not an abstract formal language but a slightly modified form of Esperanto. The IL has no syntactic ambiguity, it has regular morphology and reduced lexical ambiguity. Problems that can not be solved by the system are submitted to the user by means of an interactive dialogue. In this way, the user decides about the correct interpretation of the input text The dialogue is exclusively in the source language so that the user does not have to know either the intermediate language or any of the possible target languages. The second part of the translation - from intermediate language to target language - is performed fully automatically without any human intervention and accordingly lacks the disambiguation dialogue that is present during the first process. Let's now have a closer look at the two translation processes.

2.2. DLT step by step

The two translation stages are both built from a number of separate modules, each of which performs a well-defined task in the translation process. In the following overview, all the steps from the translation of a source language sentence into a target language sentence are made by describing the modules and their function.

Sender

1. SOURCE LANGUAGE PARSER
   The input is a SL sentence entered by the operator. For the current DLT prototype the source language is English. A syntactic analysis of the sentence is performed. The output consists of one or more possible

syntactic interpretations of the sentence in the form of dependency trees.

2. METATAXIS: ENGLISH-IL
   Metataxis, contrastive dependency grammar, concerns the translation of the English tree(s) into the IL tree(s). Because of lexical ambiguities, which means that one English word can have several translation alternatives in the IL, a large number of dependency trees can be created for each English tree.
3. SWESIL (Semantic Word Expert System for the Intermediate Language)
   From all the syntactically correct translations, SWESIL, the semantic part, chooses the most likely translation within the given context: the result is an ordered set of possible translations, each with a specific score attached to it. During this semantic evaluation of the sentence, SWESIL uses the knowledge of the world which is encoded in the Lexical Knowledge Bank.
4. DISAMBIGUATION DIALOGUE
   The system presents the choices of SWESIL to the user in the form of a dialogue. The dialogue consists of multiple-choice questions and the user must make a selection between a number of alternatives which are presented in decreasing order of plausibility as established by the semantic part. Finally, one IL tree is retained by the DLT system.
5. IL TREE TRANSDUCER
   Some tree adjustments, such as form determination and morphology, transform the selected IL tree into the final IL tree.
6. TREE-TO-STRING CONVERTER
   The final IL tree is transformed into an IL string. This final IL translation is sent over the communication network to the receivers).

Receiver

7. INTERMEDIATE LANGUAGE PARSER
   The IL string is parsed into a dependency tree. As opposed to the English parser, the IL parser generates only one possible tree structure, because the IL lacks structural ambiguity.
8. METATAXIS : IL-FRENCH
   This metataxis concerns the structural translation from the IL tree into one or more target language trees. For the current DLT prototype the target language is French. Again a large number of alternative French trees may be generated.

9. SWESIL
   French semantic evaluation and selection: the semantic part has to make the semantic and pragmatic choices between the alternative French trees, but now fully automatically. SWESIL orders the trees and makes an automatic selection of one target language tree, the tree with the highest probability score, making complete use of Artificial Intelligence methods.

10. FRENCH TREE TRANSDUCER
    Some monolingual tree transformations are performed in the target language.
11. TREE-TO-STRING CONVERTER
    The French tree is linearized, and there follows a contraction and elision program which operates on the French string. The final French translation is presented to the receiver.

3. FORM DETERMINATION FOR FRENCH

3.1. Function

In this chapter, I will describe in more detail the treatment of syntactic form in the synthesis part of DLT, giving examples from the first target language French for which the first form determination program has been developed. The role of this form determination program and the interaction between features and morphology however, will be similar for any other target language of the system (assuming that the language has inflection) and for this reason, the description has a more general scope. A similar part already is included for the IL of the current system (it can be regarded in some ways as the target language of the first DLT part), but the French counterpart is more elaborated and complicated.

The French form determination part takes place in the French Tree Transducer (see 2.2). The input tree is the French dependency tree selected by the semantic module which has to choose between a number of alternative trees generated by the IL-French metataxis. All the words that should be in the final sentence are on the correct nodes of this selected tree. The correct dependency relations are all present in the form of labels at the branches. Each tree node contains a word and some information about the word:

1. the rank number of the word in the IL sentence
2. the basic form of the word
3. the word class of the word
4. a list of features : each feature has a name and a value.

Example 3 shows a French node (in PROLOG format) with the verb [VRB] *voir* ('to see') having the present-tense-value [pr] of the tense feature [f_tns] and the indicative value [ind] of the mood feature [f_mood].

Example 3 : French node

[3,voir,VRB',[f_mood,ind,f_tns,pr]]

The French words still have their basic form, for example the French adjective is represented in its singular masculine form, the French verb has its infinitive form etc. In order to obtain the correct syntactic form, the feature lists of the words must be adapted: all the words must receive their relevant features. Then, the morphology program makes use of this basic

form, the category and the adapted feature list of the word to generate the correct syntactic form. In languages where the words do not change that much because of inflection, it is also possible and feasible to put all word forms in the lexicon. But in languages like French, every verb, for example, has many inflected forms. The lexicon would become very large if all these forms were included. It is accordingly understandable that we want to have only basic forms in the lexicon and to obtain all the other information with the help of features.

The first operation to be carried out now is the distribution of syntactic features through the tree. As the input tree is a syntactic structure in accordance with the dependency model, the words with common features can be expected to have a direct or indirect dependency relationship or a common governor. According to form government and agreement rules, the relevant syntactic features are distributed through the tree, from determining words to determined words. The rules scan the tree and adjust syntactic features. For example, a French article depending on a noun receives the gender and number feature of that noun and adds them to its own feature list. When no rule can be applied, the feature list of the word remains unchanged. The output tree of the French form determination program is a French dependency tree with words having a correct feature list containing all necessary syntactic features.

3.2. Sources of the syntactic features

The syntactic features in the final feature lists of the words necessarily originate from different sources. The same is valid for the DLT translation process from source language to intermediate language. In order to have a more general view of the feature distribution process, it is necessary to distinguish the four linguistic knowledge sources which are able to deliver syntactic features and which will be treated in the following sections:

1. source language (in this case : the IL)
2. lexicon
3. target language (in this case : French)
4. text grammar.

3.2.1. *Source language*

The features of a determined word are as a matter of fact redundant in a labeled tree structure. They do contribute to the recognition task of syntactic analysis, but when the dependency structure is established they do not add anything further to the translation. On the other hand, there are features which certainly must be transferred from source language to target language during a translation process. The question rises, what are directly translation relevant features and what are not? Which source language knowledge has to be transported to the target language? The parser splits the English or IL words into their basic form, which can be found in the lexicon, and into a set of syntactic features. Generally, the features that are transported from

one language to the other are the number features. For example, the English sentence *the computer translates a manual* is translated into the French *l'ordinateur traduit un manuel*. Both nouns, *computer* and *manual*, keep the same number feature in English and French. Some words or constructions, however, have a fixed number translation and are only used in the singular or plural form. For example, the IL word *mono* ('money') together with a singular number feature [+sg] can be translated into the French *devise* which, in this meaning, must carry a plural feature [+pl]. A possible general rule such as *preserve the same number feature from source language to target language* must be overruled by a more specific rule which treats this particular case. An entry in the bilingual dictionary indicating that the feature [+sg] has to be present in the feature list of the word *mono* for this particular translation into French can solve the problem. In addition, it is not always possible to throw the number features of the determined words away. For example, adjectives which normally receive their features from the determining nouns do have translation relevant features in some cases. Example 4 shows an IL-French translation where the number features of the adjectives ([+sg]) already differ in the source language from those of the determining noun ([+pl]). In French, the best translation is obtained if noun and adjectives have a singular number feature. The French translation which preserves the value of the source number features is *la grande et la petite maisons,* but this is rather rare in French.

Example 4 : translation relevant number feature

la	granda	kaj	la	malgranda	domoj
	[+sg]			[+sg]	[+pl]
la	grande	et	la	petite	maison
	[+sg]			[+sg]	[+sg]
'the large		and the small			house'

3.2.2. *Lexicon*

In addition to features received from the source language, determining words can be distinguished by means of inherent features. These features are language specific and are part of the lexical information, for example the gender of the French noun [f_gen]. They must be stated explicitly in the lexicon, which in this way becomes the second feature source. This static information about French words, including the category, is put in the part of the lexicon called the French syntactic dictionary. Sometimes, however, the gender feature still can play a significant role in translation. A French noun, for example *critique* (example 5) can have two gender features [f_gen,2], masculine [+m] and feminine [+f], both with a different meaning and in most cases a different translation in other languages. In this case, the translation depends on the source language word. The general rule *preserve the gender feature from the syntactic dictionary* is insufficient. The word has to be put in the bilingual part of the lexicon, in this case the IL-French dictionary, together with a specific feature, so that these features immedia-

tely can be included in the feature list of the translated word in the tree. The same solution has been adopted for the fixed number features (see 3.2.1) and can be adopted for French words whose gender features depend on their number features, such as *délice* [+sg,+m] and *délices* [+pl,+f].

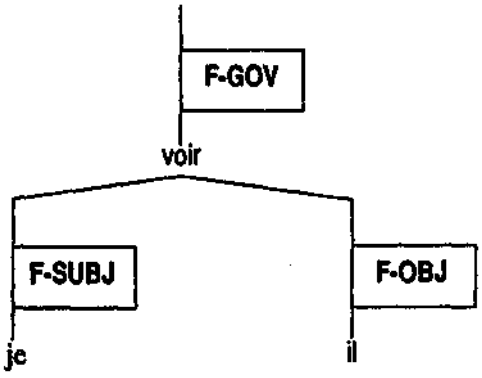
Example 5 : translation relevant gender feature

syntactic French	dictionary	bilingual IL-French	lexicon		
critique noun	[f_gen,2]	kritikisto = kritiko =	critique critique	[+m] [+f]	'critic' 'criticism'

3.2.3. *Target language*

To distribute features from determining words to other words in the tree, target language specific knowledge is required about the phenomena of form determination and agreement in the target language. The rules distribute syntactic features on purely language-specific grounds and so they have to be equipped with explicit form government and agreement rules of the target language. In case of form government for French, the words will receive syntactic features by virtue of their syntactic function in the French tree. For example, as shown in example 6, in the sentence *je le vois* ('I see him') the pronoun in basic form *il* gets the accusative [acc] case [f_case] in object position (F-OBJ). After the entire form determination process, the French morphology program transforms *il* [f_cas,acc] into the correct form *le.*

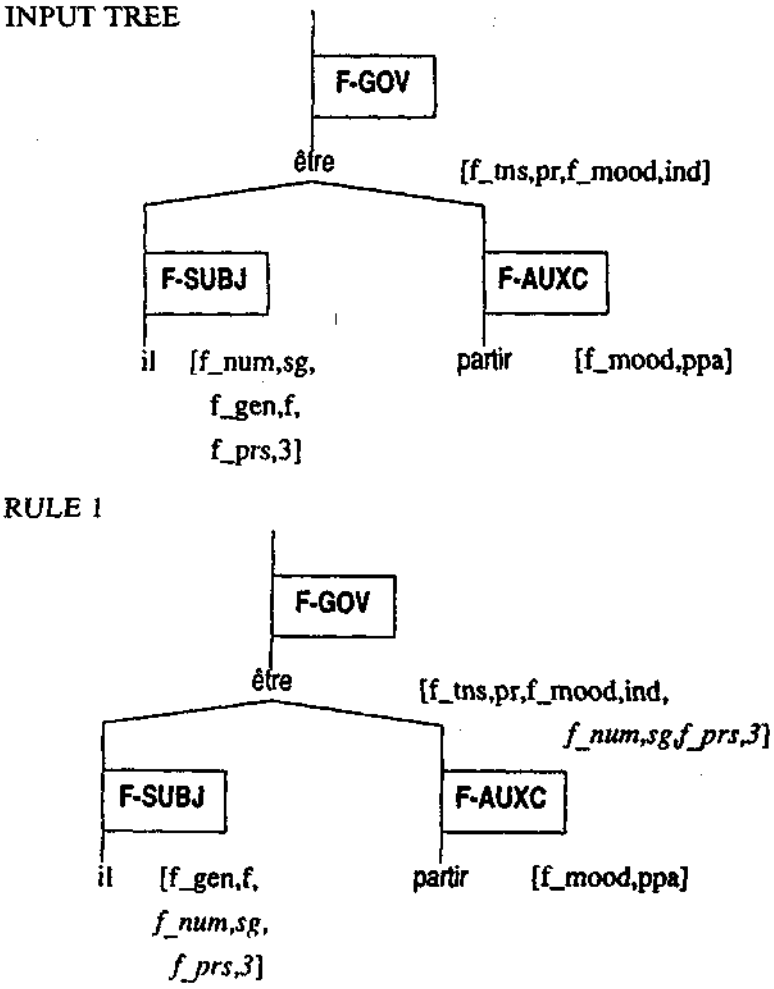
Example 6: form government rule



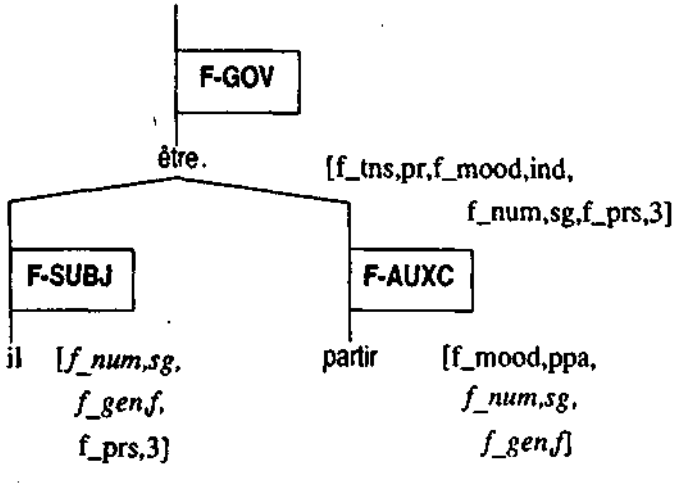
RULE → il + [f_case,acc]

In case of agreement, the rules distribute redundant syntactic features on purely language-specific grounds. The rules must first identify determining words and read their feature lists, and then find and examine determined words, adapting their feature lists if necessary. According to these rules of the target language, redundant features are distributed to the determined words in the target language tree. The French sentence *elle est partie* ('she has left') is represented in the tree as *il être partir* with syntactic features. The main verb *être* with its present tense feature [+pr] and its indicative mood feature [+ind], however, agrees in number [+sg] and person [+3] with the subject and rule 1 modifies its feature list. *Partir* in its past participle form [+ppa] also receives the number [+sg] and gender [+f] feature from the subject by means of rule 2.

Example 7: agreement rules



RULE 2



3.2.4. *Text grammar*

Until now, DLT translates sentences, but a lot of research on implementation in the field of text cohesion and larger text units than the sentence has been done. The less elaborated part of the French form determination is the fourth and last feature source: text grammar. Sometimes, agreement crosses the sentence boundary and effects other parts of the text. In this case, information from outside the sentence is required in order to distribute features through the whole tree. To illustrate this kind of agreement, let's take as an example the French translation for the English sentences *Take the machine. It is here,* that is as follows: *Prenez la machine. Elle se trouve ici.* The antecedent of *Elle* is located in the preceding sentence, namely *machine* having the inherent gender feature [+f] and the translation relevant number feature [+sg]. Both features have to be transported to the pronoun *Elle.* A machine translation system translating only sentences can not handle this problem of deixis and anaphora. The same phenomenon shows up when the pronoun belongs to the same sentence, such as the second *la* in the following French sentence *Prenez la machine et nettoyez-la* ('Take the machine and clean it').

Another phenomenon of the same order concerns tense and aspect. Adding tense features in a translation process presupposes knowledge of the context and in particular of tenses, adverbials and conjunctions in the preceding and even in the following sentences of the same text. A general rule retains the tense form of the source text in most cases. An exception is the French subjunctive for which a rule adds subjunctive features to verbs that follow a number of clearly defined French conjunctions, such as *bien que,* which always require a subjunctive mood of the verb.

4. DIFFICULTIES IN THE TARGET LANGUAGE

4.1. Semantic form determination

A general principle according to which the DLT system has been developed is the maintenance of the distinction between syntax and semantics. In grammars, form determination often is explained according to semantic criteria. A difference in form determination is explained on the basis of a different interpretation of the sentence. A form determination program based on a syntactic dependency tree has to search for well-defined criteria within the scope of this problem. In a verb construction whose subject is *une foule de* ('a crowd of'), the verb agrees with the complement when the subject noun phrase is regarded as a plurality, "... 1'idee de nombre predomine..." (Grevisse 1986: 695). For example, *une foule* [+sg] *de gens* [+pl] *diront* [+pl]... But when there is emphasis on the totality of the entity, the verb agrees with the subject, for example *une foule* [+sg] *de gens* [+pl] *accourait* [+sg]. A solution could be found by using some semantic features which give information about the interpretation of the sentence, and then give a certain code or feature to *foule* or to the main verb. Another example constitutes the French construction *un nombre de* ('a number of') which is treated in accordance with the same above-mentioned semantic criteria. In this case however, the source language IL simplifies the procedure: for each different interpretation, it has a different preposition, *nombro da* which means 'plurality' and *nombro de* which means 'totality'.

4.2. Long-distance form determination

Sometimes it is hard to establish the real source of form determination in a sentence, especially when dealing with long-distance agreement in complicated structures in which it is not always clear whether agreement is required or not. And if so, with which other word in the sentence? Let's take as an example the French sentence *Ils ont pour règle de ne jamais être contents* ('they have the rule to never be satisfied'). How can the form determination program be aware of this long-distance agreement between the subject *Ils* and the word *contents?* Another solution would be not to change the basic form of the adjective *content.* In this way, the word would be related to some vague term or concept, such as *on* ('one', 'people').

5. CONCLUSION

I have tried to show how a specific grammatical phenomenon, form government and agreement, is approached and treated in the machine translation system DLT. The languages I used for the examples, English, IL and French are the first languages involved in the system. French is the first target language for which a form determination program has been developed. Some further developments definitely concern a number of refinements and additions to the existing program for French with a more general scope in the field of text cohesion and its influence.

ACKNOWLEDGEMENTS

I would like to thank Job van Zuijlen, Klaus Schubert and Dan Maxwell (members of DLT research team) for their constructive comments on this paper.

REFERENCES

DANLOS Laurence, NAMER Fiammetta (1988), Morphology and cross dependencies in the synthesis of personal pronouns in Roman languages, in *Proceedings of Coling '88,* Budapest.

GREVISSE M. (1986), *Le bon usage,* grammaire française, Paris-Gembloux: Duculot.

PAULUS D. (1986), *Ein Programmpaket zur morphologischen Analyse,* Universität Erlangen-Nurnberg, RRZE, Diplomarbeit, RRZE-IAB-259.

SCHUBERT Klaus (1987), *Metataxis. Contrastive dependency syntax for machine translation*, Dordrecht/Providence: Foris.

TAMIS Dorine (1987), *Contribution à la synthèse automatique du français.* Unpublished report, Utrecht: BSO/Research/Unpublished doctorandus thesis, Amsterdam: Vrije Universiteit.

TESNIÈRE Lucien (1959), *Elements de syntaxe structurale*, Paris : Klincksieck [2. ed., 4. print 1982].

Address: Buro voor Systeemontwikkeling,
Kon. Wilhelminalaan 3, Postbus 8348, N-3503 RH Utrecht (Holland).

Received: 18 September 1988.