Transfer InterStructure:
Designing an 'Interlingua' for Transfer-based MT Systems

Juan Alberto Alonso
Siemens, C.D.S.
Cornelia, Barcelona - Spain
May 1990

CONTENTS

1.- Introduction.

This paper tries to analyze and present some relevant factors on the design, role and use of a "pivot" syntactic structure between the analysis and the synthesis phase in Transfer-based multilingual Full Automatic Machine Translation (FAMT) systems. Throughout this paper, this universal pivot syntactic structure will be called "Transfer Interstructure" (henceforth TIS, for short).

While the use of a normalized pivot structure between analysis and synthesis is a fact for many current transfer-based FAMT systems, this structure is usually tailored for one particular language-pair. The question for a "universal" pivot structure arises when the FAMT is intended to be multilingual. Such universal pivot structure (i.e., the TIS) deviates from the Interlingua approach in that it does not include a universal representation for lexical items nor for semantic relations (discourse analysis).

A type of TIS is already being used by EUROTRA (cf. EUROTRA IS level) and is currently being incorporated into the METAL system (cf. METAL MIR).

2.- The Aims of the Transfer InterStructure (TIS).

Current work with actual non multilingual MT systems shows that there is a strong tendency for the analysis grammar to be written with the generation to a particular target language in mind, and the same applies for the synthesis grammar with respect to a given analysis. This is so, even if there is a pivot normalized structure defined between analysis and transfer; the problem is that this pivot structure is usually defined for the specific language pair being handled. When a new source or target language comes into play, this bilingual pivot structure is of little use. While this is not so serious for MT systems handling one or two language pairs (which is the case for the vast amount of current MT systems; cf. [Hutchins86]), it represents a severe drawback for multilingual systems.

Following the traditional Transfer approach there must be n(n - 1) transfer modules (where n is the number of source/target languages handled by the system). For each possible language-pair, these transfer modules typically consist not only of a bilingual lexicon, but also of a grammar component which "tunes" the analysis of the source language to the synthesis of the target language.

The existence of the TIS guarantees the independence of analysis and synthesis grammars, which is a basic requirement for practical multilingual MT systems, and at the same time, minimizes the size and complexity of the transfer modules, reducing them, in the ideal case, to a bilingual lexicon.

With the use of TIS the following aims are intended to be achieved:

Reduce the transfer module for each language-pair to the bilingual lexicon. Therefore, no grammar transfer module will exist anymore (see 4.3 for some remarks on this subject).

Each source language will only have one single analysis module which is target-language independent. Each analysis module should deliver well-formed TIS trees as output.

Each target language will only have one single synthesis module which is source-language independent. Each synthesis module takes well-formed TIS trees as input.

A model for a TIS-based MT system for three languages (i.e., six language-pairs) is sketched in figure 1.

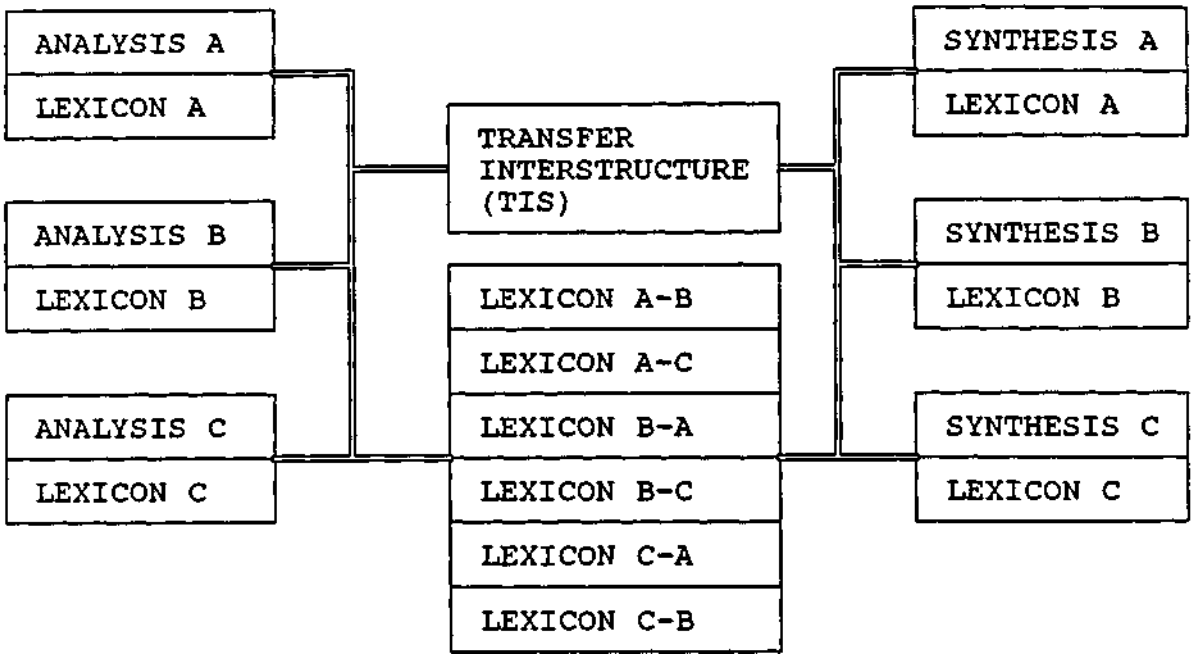


Fig. 1: TIS-based MT system model languages (6 language-pairs)

3.- Notational Tools for a Formal Description of the TIS.

In this section, a notation intended to be used for formally describing TIS trees is presented. Basically, a TIS tree is a set of nodes characterized by the following conditions:

* Each node is identified by a category label.
* Each node consists of a bundle of feature-value pairs.
* There is a dominance dependence between root and son nodes.
* There is a linear order dependence among nodes having the same level of dominance sharing a common root node (sibling nodes).

Thus, the description of a TIS consists of a declaration of all the possible feature-value pairs which can be found in a TIS tree. There follows a declaration of all the possible category labels for a node, together with the obligatory features for each of these labels. Finally, there is a declaration of the dominance and lineal order dependences among the possible TIS nodes (the TIS legal tree structures).

In the following description of the TIS declaration language we will use the following metalanguage conventions:

```
<non_terminal_symbols>
'terminal_symbols'
{terminal_symbol, terminal_symbol, .. . ,  terminal_symbol}
```

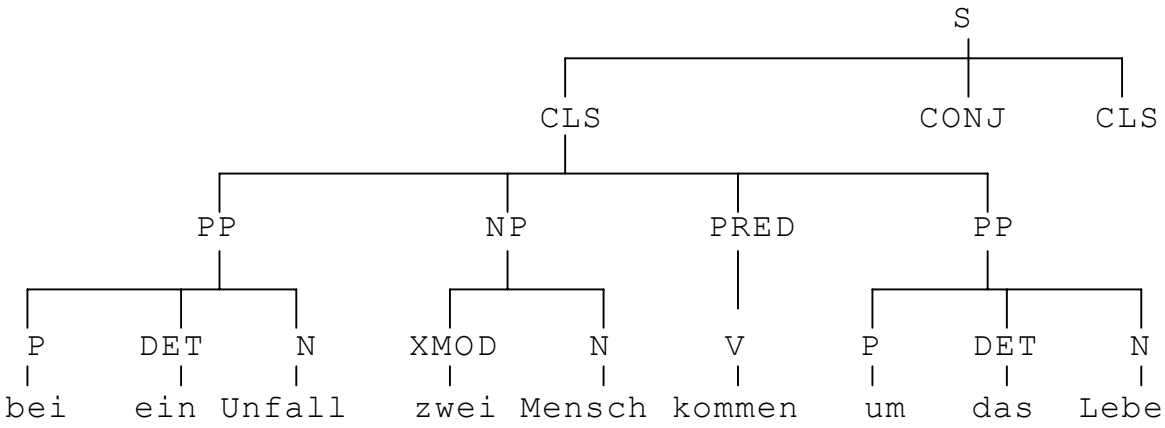
"+" indicates one or more occurrences of the following symbol.
"-" indicates cero or one occurrence of the following symbol.
"*" indicates cero or more occurrences of the following symbol.

```
<tis_description> ::

     'tis' ': '
     <tis_fv_descriptor>
     <tis_nodes_descriptor>
     <tis_tree_descriptor>.

<tis_fv_descriptor> ::  'tis_fv' ' : '  +<fv_descriptor>.

     <fv_descriptor> :: <feature_label> '=' <value_set>.

          <feature_label> :: {GD, NU, ROL, ...}.
          <value_set> :: '['
                         +{M, F, N, SG, PL, SUBJ, DOBJ, . . . }
                         <string_value>
                         ']'.
          <string_value> :: String.

<tis_nodes_descriptor> :: 'tis_nodes' ' : '  +<node_descriptor>.

     <node_descriptor> :: <node_label> '=' <feature_set>.

          <node_label> :: {S, CLS, CLS-SUB, NP, PP, N, . . . } .
          <feature_set> :: + {GD, NU, ROL, . . . } .

<tis_tree_descriptor> :: 'tis_tree' ' : '  +<tree_descriptor>.

     <tree_descriptor> :: <root_node_label>
                          '->'
                          +<son_nodes_list>.

          <root_node_label> :: {S, CLS, CLS-SUB, NP, PP, N, . . . } .
          <son_nodes_list> :: '(' -<quantifier><node_label> ')'.
          <quantifier> :: {-, +, * } .
          <node_label> :: {S, CLS, CLS-SUB, NP, N, . . . } .
```

A node N can be unambiguously identified by stating its Label, its Path and (optionally) its Sequence number. The Label corresponds to the syntactic category of N (CLS, NP, P, etc.); the Path indicates the sequence of node labels which must be traversed in order to go from the upmost root node of the tree to N; the Sequence number is the left-to-right order number of

the node in case there are more nodes with the same Path and Label under the same immediate root node.

So, for instance, in the tree



the node N corresponding to "Unfall" can be described as s:clsl:ppl:N, while the PP "um das Leben" is s:clsl:PP2.

4.- Constraints on the Structure and Information of the TIS.

4.1.- Basic Structure Constraints and Definitions.

Given the following figure:

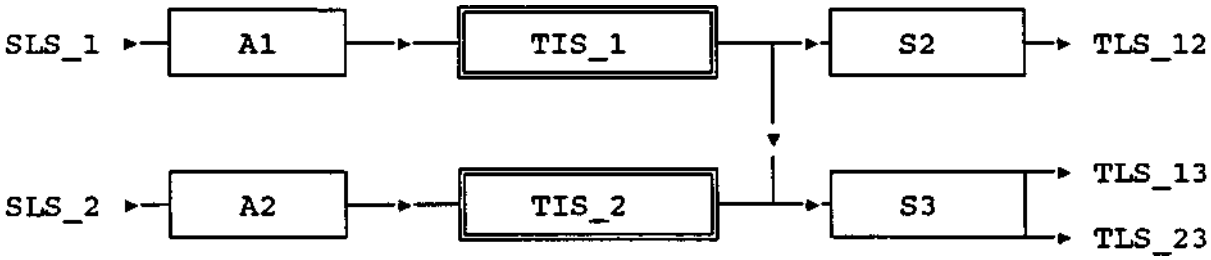


Fig. 2

and having into account the aims stated in section 2, we can derive the following general constraints operating on TIS trees:

Analysis Equivalence Constraint:

Two different analysis modules for two different source languages should deliver two TIS trees which are strongly equivalent for two input sentences with exactly the same meaning in both languages.

(SLS_1 <=> SLS_2) ==> (TIS_1 <> TIS_2) in Fig. 2

Note: <=> indicates "has the same meaning"
<> indicates "is strongly equivalent to"
>< indicates "is weakly equivalent to"
==> indicates "implies"

Meaning Preservation Constraint:

Two different synthesis modules for two different target languages should generate two sentences with exactly the same meaning if they have either the same TIS tree or two strongly equivalent TIS trees as input.

(TIS_1 <> TIS_2) ==> (TLS_12 <=> TLS_23) in Fig. 2

Strong Equivalence Definition:

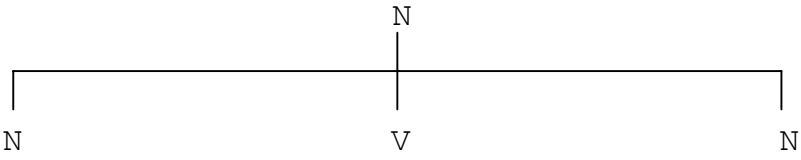
Two TIS trees T and T' are strongly equivalent (T <> T') if for every node N belonging to T there is a corresponding node N' which has the same path, label and sequence as N and the same set of feature-values as N (disregarding lexical features, i.e., string-valued features).

4.2.- Relaxing the Constraints.

Up to now, we already have the aims which the TIS must satisfy, the constraints which the TIS must meet and the notational tools which allow us to describe it. However, any implementation of a TIS following these guidelines for a practical MT system will soon be fated to failure because of an obstacle which is impossible to overcome: the Analysis Equivalence Constraint is practically impossible to satisfy in the terms it was stated in 4.1.

There is a number of linguistic phenomena in different languages which are extremely difficult to reduce to common syntactic representations. One example can be the compound constructions. German, for instance, makes extensive use of compound nominal constructions, like, for example, "Zugbremsvorrichtung" ("train braking mechanism"). Romance languages, on the other hand, usually express the same idea with a noun complemented by one or more prepositional phrases: "dispositivo de frenado para trenes". It is extremely difficult to be able to come to a common intermediate representation for both constructions. First, let us propose the following TIS structure for "Zugbremsvorrichtung":

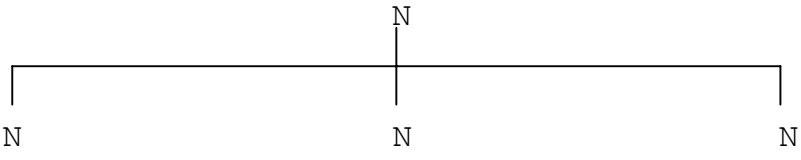
Tl_Ger:



Zug	brems	Vorrichtung
FUNC SPEC	FUNC SPEC	FUNC HEAD
HEAD Vorrichtung	HEAD Vorrichtung	--

and the following TIS structure for "dispositivo de frenado para trenes":

Tl__Spa:

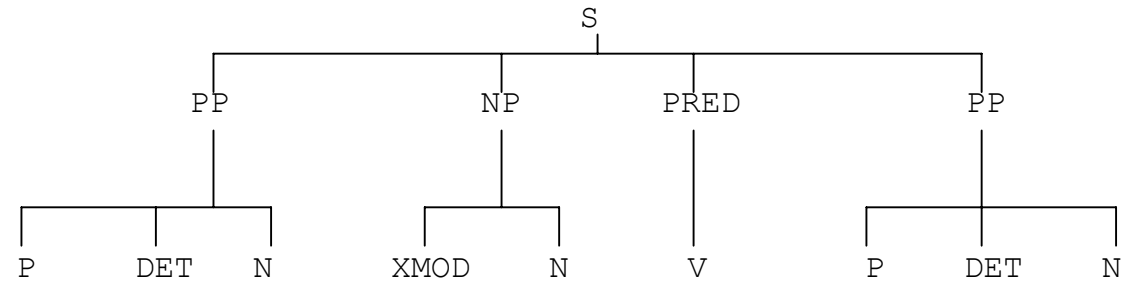


dispositivo	frenado	trenes
FUNC HEAD	FUNC SPEC	FUNC SPEC
--	HEAD dispositivo	HEAD dispositivo

We already "skipped" the fact that "de frenado" and "para trenes" are prepositional phrases (PP). We can choose to represent them in TIS as nominal structures, preserving the nominal head of the PP and supposing that the preposition is something which is Spanish-specific. Even then, it is evident that Tl Ger and Tl Spa are not strongly equivalent.

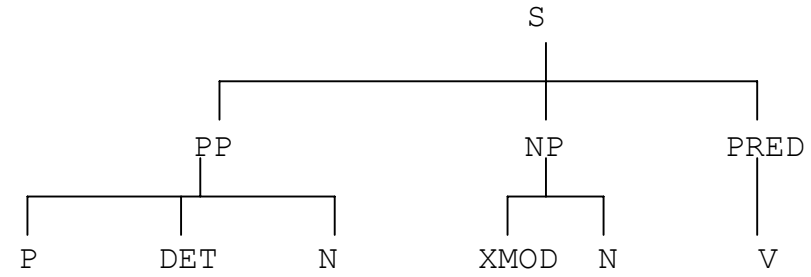
There are even more extreme cases. Let us consider the phrase "ums Leben kommen" in German, which translates as "to die" in English and "morir" in Spanish. Let us give the corresponding TIS representations for "Bei einem Unfall kamen zwei Menschen ums Leben" and the corresponding English "Two persons died in an accident":

T2_Ger:



bei ein Unfall zwei Mensch kommen um das Leben

T2_Eng:



The lack of strong equivalence is even more evident in this case, a whole PP branch present in the German TIS tree being missing in the English TIS tree. A possible solution would be to delete the "ums Leben" PP and 'abstract' it in form of feature-value pairs in the PRED node (for example, as [PP_COMP_HEAD Leben] and [PP_COMP_PREP um]). However, we cannot do this before accessing the transfer bilingual lexicon. It is in the transfer lexicon where it is stated that whenever "kommen" has a prepositional complement with "Leben" as head it translates as "to die" and the "Leben" PP must be pruned (see 4.3) . Pruning the PP from the TIS structure before accessing the transfer lexicon would give wrong results for sentences like "Bei einem Unfall kamen zwei Menschen um die Ecke" ("In an accident two persons came round the corner"), where the 'um' PP need not be pruned.

In general, constructions involving compounds, multiwords, function verbs, ellipses and coordination are typical cases where the Analysis Equivalence constraint cannot be satisfied.

Since it seems impossible to ensure the Analysis Equivalence Constraint, let us try to relax somehow this constraint:

Weak Analysis Equivalence Constraint:

> Two different analysis modules for two different source languages should deliver two TIS trees which are weakly equivalent for two input sentences with exactly the same meaning in both languages.
>
> (SLS_1 <=> SLS_2) ==> (TIS_1 >< TIS_2) in Fig. 2

We must now define what the "weakly equivalent" dependence is.

Weak Equivalence Dependence Definition

> Two TIS trees Tl and T2 coming from analysis modules of languages L1 and L2 respectively are weakly equivalent to each other (Tl >< T2) if either

T1' is strongly equivalent to T2 or
T2' is strongly equivalent to T1,

where T1' and T2' are the result of submitting T1 and T2, respectively, to the lexicon-driven structural transformations specified by the transfer lexical entries corresponding to the lexical material contained in the trees (see 4.3).

Thus, if T2_Ger undergoes the changes specified in the entry for "kommen" (if it has a prepositional object with head "Leben", translate it by "to die" and prune the prepositional object) we get a tree T2'_Ger which is strongly equivalent to T2_Eng.

In fact, the above Weak Equivalence definition implies the following synthesis constraint on weak equivalent TIS trees:

Weak Equivalent Trees Synthesis Constraint:

Every synthesis module should produce sentences with the same meaning when coming from different TIS trees which are weakly equivalent.

(TIS_1 >< TIS_2) ==> (TLS_13 <=> TLS_23) in Fig. 2

Now, we have to re-enunciate the Meaning Preservation Constraint for weakly equivalent TIS trees:

Weak Meaning Preservation Constraint:

Two different synthesis modules for two different target languages should generate two sentences with exactly the same meaning if they have either the same TIS tree or two weakly equivalent TIS trees as input.

(TIS_1 >< TIS_2) ==> (TLS_12 <=> TLS_13)
(TLS_12 <=> TLS_23) in Fig. 2

4.3.- Lexicon-driven Transformations.

It is important to clearly define what kind of transformations carried out by the synthesis module may be triggered by the bilingual lexicon entries. Three basic types of such lexicon-driven transformations exist:

Adding Sub-structures to the TIS tree.

The transfer lexicon entry specifies, together with the target language translation for the source language entry, a node descriptor to be added to the current TIS tree. This node descriptor is defined by the node label, the node path and the minimal set of feature-value pairs to be contained in the added node.

Example of ADD transfer lexical entry:

betrachten V —> **tomar** V
Test: none.
Do: ADD S:CLS:PP([ROL POBJ] [PREP en] [HEAD consideración]).
Comment: ("etwas betrachten" = "tomar algo en consideración").

Pruning Sub-structures from the TIS tree.

The transfer lexicon entry specifies, together with the target language translation for the current source language entry, a node descriptor to be deleted from the current TIS tree. This node descriptor contains the path and label of the root node dominating the sub-tree to be deleted, together with either the sequence number or a set of feature-value pairs which uniquely identify this root node.

Example of PRUNE transfer lexical entry:

machine N —> **lavadora** N
Test: EXISTS S:CLS:NP:PP([PREP à] [HEAD laver])
Do: PRUNE S:CLS:NP:PP([PREP à] [HEAD laver])
Comment: ("machine à laver" = "lavadora").

Mapping Sub-structures in the TIS-tree.

The transfer lexicon entry specifies, together with the target language translation for the current source language entry, a node descriptor to be mapped into another node descriptor in the current TIS tree. Both node descriptors contain the path and label of the source and target root nodes dominating the sub-tree to be mapped, together with either the sequence number or a set of feature-value pairs which uniquely identify these root nodes.

Example of MAP transfer lexical entry:

like V —> gustar V
Test: none.
Do: MAP S:CLS:NP([ROL SUBJ]) S:CLS:NP([ROL IOBJ])
MAP S:CLS:NP([ROL DOBJ]) S:CLS:NP([ROL SUBJ])
Comment: ("The boy likes the game" —> "El juego gusta al niño")

5.- Designing and Implementing the TIS.

The design and implementation of an actual TIS can be separated into two different parts: the design of an adequate set of normalized tree structures which the analysis phase must yield and the choosing of an adequate set of Feature-Value pairs to represent the relevant information extracted during the analysis.

The specification of the normalized tree structures implies choosing a set of syntactic categories which will identify the nodes of the TIS trees, as well as a series of dominance and linear order dependences among these categories which are both simple enough to be handled easily by the synthesis modules and sufficient to express the structural information extracted during the analysis.

On the other hand, the decision on which information should be present and in what form should it be stored in the TIS after the analysis phase is one of the most crucial factors in the design of the InterStructure Feature-Value set. The TIS must gather all the relevant information from every possible analysis module in such a way that it is ready for use for any synthesis module at generation time.

The following general constraints regarding the TIS design and implementation must be considered:

Features instead of nodes.

It is better to featurize information than to have it in structural form (i.e., in nodes). In this way, the normalized tree structures become simpler and easier to be handled by the synthesis modules.

Many nodes which are present in the parse tree can be eliminated in the TIS and the information they represent expressed in form of feature-value pairs contained by higher level nodes. This is the case with morphological affixes (prefixes, infixes and suffixes) , case particles, some types of determiners (articles), auxiliary and modal verbs, some types of adverbs conveying verbal aspect or time information, etc. This type of nodes usually represent the surface language-dependent structure, which is of no interest at all for the TIS.

Deep information instead of surface information.

The information conveyed by the Feature-Value set of the TIS must not refer to surface phenomena of the source language; instead, the Feature-Value set should convey deeper level information which is common to all the languages in question.

Thus, for example the TIS must not have information on the grammatical Gender of a source language noun, but on its Sex (natural gender) . Case information for Noun Phrases is also irrelevant once Role functional information (subject, direct object, etc.) is present in the corresponding TIS nodes. The same could be said for Tense vs. Time, Predicate Form vs. Aspect, grammatical Voice vs. Diathesis, etc.

No language-specific information must be present.

The TIS should not contain any information which is language-specific. Only that information which is essential to preserve the source sentence meaning should be yielded by the analysis in the TIS.

The decision of what is and what is not language-specific is indeed one of the major problems in the task of designing a TIS. For example, if in a FAMT system handling indoeuropean languages a language like Japanese is included, the 'politeness' information contained in the Japanese verbal forms could be considered "Japanese-specific" an does not need to be included in the TIS, since no indoeuropean will make use of it. However, whenever another Asian language (cf. Korean or Chinese) is also added, this politeness information will be no more language specific and, thus, must be represented by dome feature in the TIS representation.

6.- Conclusions.

If a common syntactic pivot interstructure in the terms presented in this paper can successfully be implemented for practical MT systems, this would imply a big step towards the attainment of a multilingual MT system easier to design and maintain.

In fact, when multilinguality comes into play in FAMT systems, there is no other way out than either taking the "strong" Interlingua approach or taking this "syntactic" Interlingua approach. While the traditional Interlingua approach seems to be still a subject for laboratory research due to the complexity of finding a common abstract representation not only for syntactic structures, but also for lexical items and meaning relationships, the more humble (but handier) Interstructure method seems to be a good attempt to overcome the difficulties of multilingual FAMT.

Nevertheless, there are still a number of open problems in the design and implementation of a fully operative Interstructure:

Finding the exact information to be present in the TIS and the adequate set of feature-value pairs which represent it.

Dealing with some linguistic phenomena which seem reluctant to be "interstructured".

The inclusion of a new target language may require re-designing the current TIS used until then, since new information could be necessary for the new target language to be generated which is not yet present in the current TIS.

The Interstructure approach becomes more difficult to apply when translation is carried out between languages of different linguistic families. Nowadays, a variety of TIS is being used at least in two big FAMT multilingual systems (EUROTRA and METAL), both of them handling indoeuropean languages. The inclusion of a non-indoeuropean language (Basque, Hungarian, Finnish, Japanese, Chinese, Arabic, etc.) would imply a re-design of the current TIS being used, with unpredictable results.

7.- Literature.

[Bourbeau88]: BOURBEAU, L & LEHRBERGER, J. Machine Translation: Linguistic Characteristics of MT System and General Methodology of Evaluation. John Benjamins Publishing Company. 1988 Amsterdam/Philadelphia.

[Hutchins86]: HUTCHINS, W.J.; Machine Translation: past, present and future. Ellis Horwood Ltd. 1986 Chichester, England.

[Nirenburg87]: NIRENBURG, S. Ed. Machine Translation. Cambridge University Press. 1987 Cambridge, UK.