About and Around the French Enconverter and the French Deconverter

Etienne Blanc

GETA, CLIPS-IMAG
BP53, F-38041 Grenoble Cedex 09
etienne.blanc@imag.fr

Abstract. We briefly describe the French Enconverter and the French Deconverter. We discuss then a few general points concerning the possibility of designing dependency trees equivalent to UNL graphs, the treatment of the ambiguity and anaphora resolution, and the structure of the compound nodes.

1 Introduction

In a previous paper [1], we described the basic principle of our French Enconverter, in which the UNL input graph is processed into an equivalent Dependency Tree, which is in turn applied to the entry of a rule-based French generator. We developed similarly a French enconverter, in which a French Analyser provides a representation of the text meaning as a Dependency Tree, which is further processed into an equivalent UNL graph.

In this paper, we will first briefly present the structure of the French Deconverter and Enconverter. We will then recall and discuss a little further than in our previous paper the general problem of the equivalency between UNL graph and dependency tree. And finally briefly comment on three topics we had to deal with when devising our Enconverter and Deconverter : Ambiguity and Anaphora Resolution, Processing of the Unknown Word, the exact structure of the Compound Node of a UNL graph.

2 Overall Structure of the French Deconverter and Enconverter

The French Enconverter and the French Deconverter are written on ARIANE-G5.

ARIANE-G5 is a generator of MT systems, which is an integrated environment designed to facilitate the development of MT systems. These MT systems are written by a linguist using specialised languages for linguistic programming. ARIANE is not devoted to a particular linguistic theory. The only strong constraint is that the structure representing the unit of translation (sentence or paragraph) must be a decorated tree.

Fig.1 shows an overview of a classical transfer MT system using the ARIANE environment. The processing is performed through the three classical steps: analysis, transfer and generation. An interactive disambiguation module may be inserted after

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 157–166.

the analysis step. Deconversion or Enconversion cannot be performed straightforward by ARIANE, whose inputs and outputs are texts or trees. Thus additional external modules are necessary, transforming a graph into an equivalent tree, or inversely.

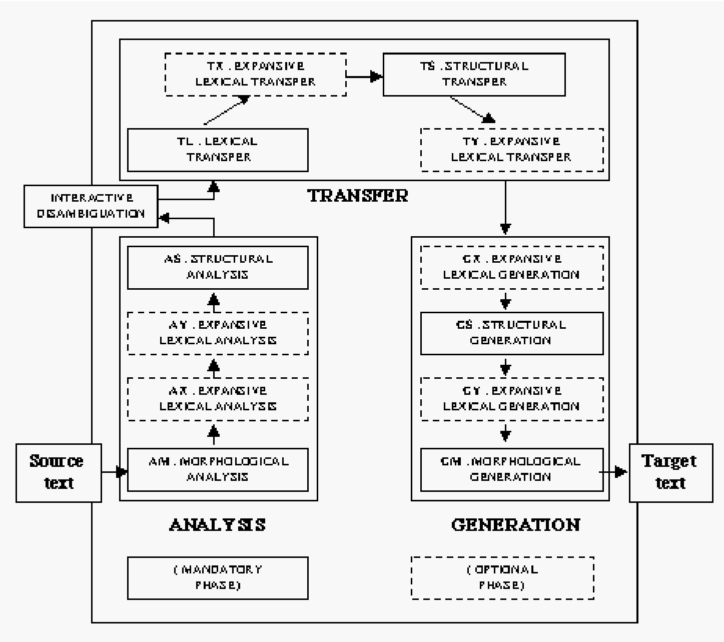

TX . EXPANSIVE LEXICAL TRANSFER
TS . STRUCTURAL TRANSFER
TL . LEXICAL TRANSFER
TY . EXPANSIVE LEXICAL TRANSFER
INTERACTIVE DISAMBIGUATION
TRANSFER
AS . STRUCTURAL ANALYSIS
AY . EXPANSIVE LEXICAL ANALYSIS
AX . EXPANSIVE LEXICAL ANALYSIS
AM . MORPHOLOGICAL ANALYSIS
GX . EXPANSIVE LEXICAL GENERATION
GS . STRUCTURAL GENERATION
GY . EXPANSIVE LEXICAL GENERATION
GM . MORPHOLOGICAL GENERATION
Source text
Target text
ANALYSIS
GENERATION
(MANDATORY PHASE)
(OPTIONAL PHASE)

Fig. 1. The Ariane-G5 environment as used for generating a transfer MT system

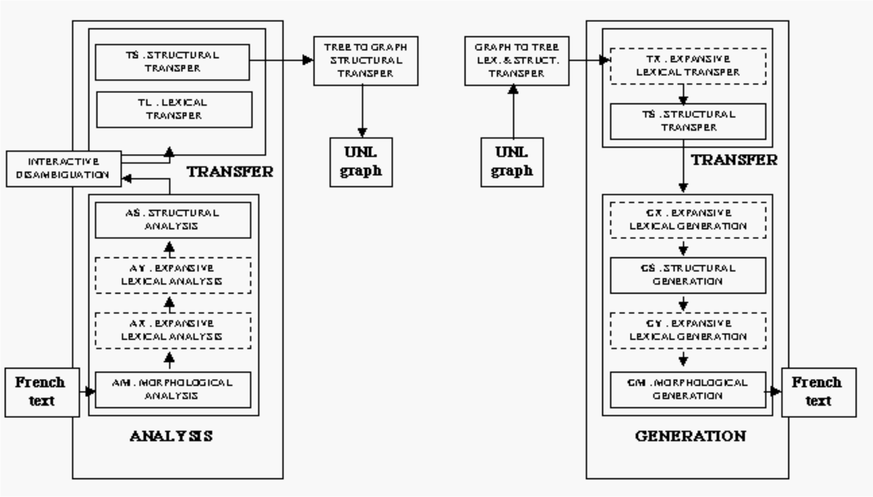

TS . STRUCTURAL TRANSFER
TL . LEXICAL TRANSFER
TREE TO GRAPH STRUCTURAL TRANSFER
UNL graph
GRAPH TO TREE LEX. & STRUCT. TRANSFER
UNL graph
TX . EXPANSIVE LEXICAL TRANSFER
TS . STRUCTURAL TRANSFER
INTERACTIVE DISAMBIGUATION
TRANSFER
TRANSFER
AS . STRUCTURAL ANALYSIS
AY . EXPANSIVE LEXICAL ANALYSIS
AX . EXPANSIVE LEXICAL ANALYSIS
AM . MORPHOLOGICAL ANALYSIS
GX . EXPANSIVE LEXICAL GENERATION
GS . STRUCTURAL GENERATION
GY . EXPANSIVE LEXICAL GENERATION
GM . MORPHOLOGICAL GENERATION
French text
French text
ANALYSIS
GENERATION

Fig. 2. French Enconverter (left) and French Deconverter (right) using Ariane-G5

Fig. 2 shows the overall structure of the French enconverter (left) and of the French Deconverter (right).

Enconversion takes place in two steps :

- Analysis of the French text producing a representation of its meaning in the form of a dependency tree (ARIANE Analyser).
- Lexical (from the French lemmas to the Universal Words) and structural (from tree to graph) transfer from the dependency tree to an equivalent UNL graph (External Transfer module).

Similarly, Deconversion takes place in the two following steps:

- Lexical (from the Universal Words to the French lemmas) and structural (from graph to tree) transfer from the UNL graph to an equivalent dependency tree (External Transfer module).
- Generation of the French sentence (ARIANE Generator).

3 Dependency trees equivalent to UNL graphs

Several cases are to be considered.

3.1 Graph with tree structure

The simplest case is when the graph has in fact a tree structure. The only difference between the graph and the tree is then that the semantic relations are attached to the arcs for the graph, to the target nodes for the tree. This is shown on Fig. 3.

For the sake of clarity, in this figure as in the following ones, restrictions and attributes are omitted. The entry node of a graph (or of a compound node) is indicated by a bold border.

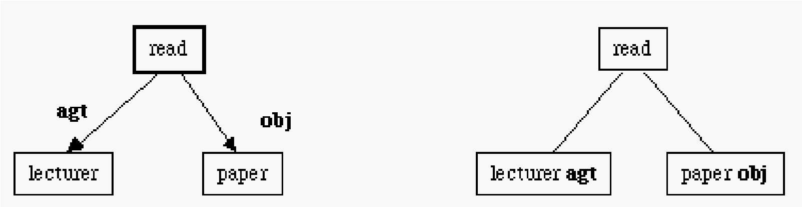


Fig. 3. A graph with tree structure (left), and its equivalent tree (right). *The lecturer reads a paper*

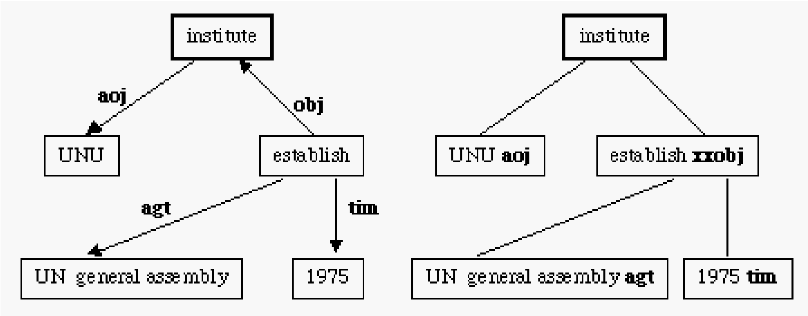


Fig. 4. A graph where the entry node has a mother node. The **obj** relation is inverted in the tree (**xxobj**). *UNU is an Institute which was established by the UN General Assembly in 1975*

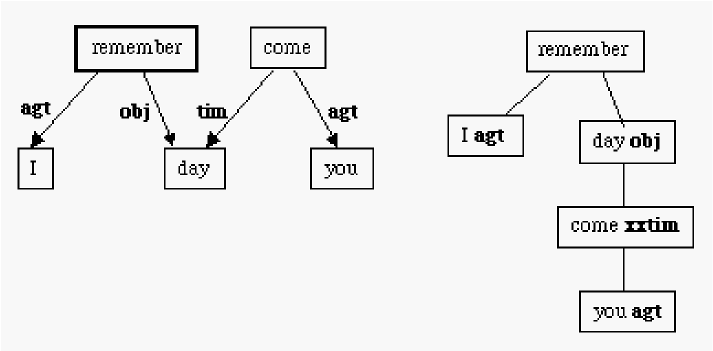


Fig. 5. The node "day" has two mother nodes in the graph. The **tim** relation is inverted in the tree (**xxtim**). *I remember the day where you came.*

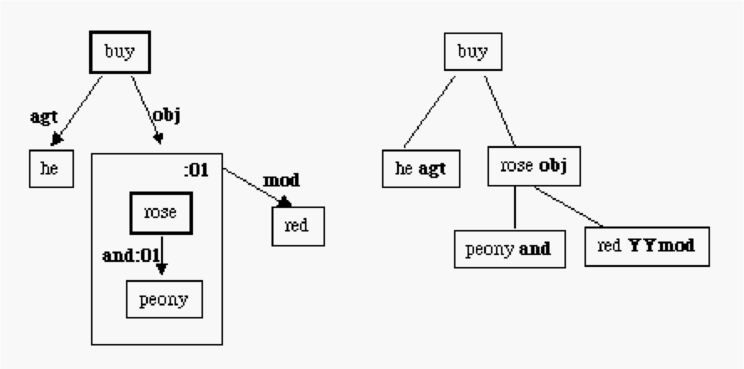


Fig. 6. A graph with a closed circuit and its equivalent tree. *The lecturer reads his paper.*

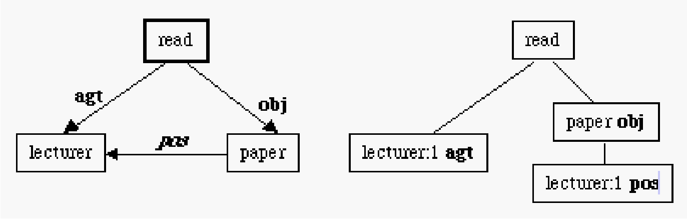


Fig. 7. A graph with a compound node. In the tree, the **yymod** relation of the "red" node indicates that the **mod** relation applies to the dependants of its mother node "rose" as a whole. *He buys red roses and red peonies.*

3.2 Graphs containing nodes having more than one mother node, or an entry node having a mother node

In a tree, the root node has no mother node, and the other nodes have only one mother node. This is of course generally not the case for a graph, where all the nodes (including the entry one) may have several mother nodes.

Let’s for instance consider the graph of fig. 4, representing the meaning of the sentence "The University of the United Nations is an Institute founded by the United Nations General Assembly in 1975". In this graph, the entry node (« institute ») has a mother node (« establish »), and the arc joining both nodes bears the *obj* relation: In order to get a tree structure, the direction of this arc is inverted, and the *obj* relation replaced by an "inverted *obj* relation" we denote by *xxobj*. The transfer into an equivalent tree is then straightforward. In the original graph, « institute » is the *obj* of establish, whereas what is expressed in the tree by the *xxobj* relation is that « establish » has « institute » as *obj*. Such an "inverted relation" is usually deconverted into French as a relative clause. The deconverted French text reads *"L'université des Nations Unies est un institut que l'Assemblée Générale des Nations Unies a fondé en 1975".*

Fig. 5 shows a graph where a node has two mother nodes. In the same manner, one of the arcs is inverted, and a *xxtim* relation replaces the *tim* relation. And again a relative clause will appear in the deconverted sentence *"Je me souviens du jour où tu es venu" ("I remember the day when you came")*

3.3 Graph containing a closed circuit

An equivalent tree structure of a graph containing closed circuits may be obtained by opening the circuits, splitting one of their nodes as shown on fig.6, where the node "lecturer" has been split into two nodes.

The new created node bears the same id number as the original one, indicating that it refers to the same object. In this example, this new node will be translated in French by the possessive "son" (its).

3.4 Graph containing compound nodes (scopes)

Fig. 7 shows a graph containing a compound node. The head :01 of the compound node does not appear in the corresponding tree. But the attributes and the dependants of the compound node as a whole are distinguished from the dependants and attributes of the entry node by specific variables, like *yymod* (for a node dependant of the scope) to be compared with *mod* (for a node dependant of the entry node).

4 Ambiguity and Anaphora Resolution

The problems of ambiguity and anaphora resolution have in principle not to be considered in the Deconversion process, a correct graph being unambiguous, and without any anaphoric pronouns.

They are on the other hand of course essential in Enconversion.

In the French Enconverter we develop, disambiguation is realised automatically, but we plan to introduce in the future interactive disambiguation using the methodology developed at Geta [2], and complete if necessary by a revision of the graph using a graph editor [3].

Anaphora resolution is interactive, as shown by the example of Fig. 8

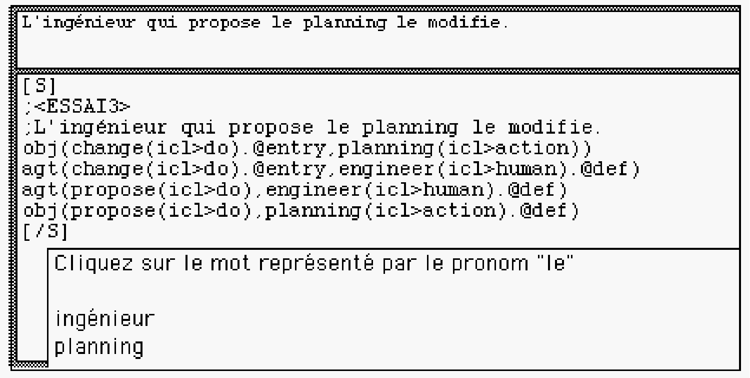
5 The Unknown Word

The problem of the unknown word arises as well in Enconversion as in Deconversion, but is generally more important for deconversion, where the user should use the deconverter as a black box providing the best result without any intervention.

Fortunately, in the case of deconversion, the very principle of UNL offers two means for deducing the part of speech of the target word associated to the UW of a given node. The first and most straightforward one is to deduce it from the UW restriction. The second one is to look at the relations in which the node is involved. For instance a node related to a daughter node by the *agt* relation, is very probably a predicate.

Both methods are implemented in our French Deconverter. The result is illustrated by the following example. The graph of fig. 9 contains 5 UWs, 4 of them corresponding to chemical terms unknown by our dictionaries. The structure of the deconverted sentence " Le? <<alcoylbenzene>> est obtenu en <<reduce>> le? <<group>> <<carbonyle>>." is nevertheless correct, and the sentence is comprehensible (The correct sentence would read "L'alcoylbenzène est obtenu en réduisant le groupe carbonyle"). The unknown words are represented by the headwords of the corresponding UWs put between <<>> marks. The question marks indicate that the articles could not be correctly calculated due to the lack of information about the gender.

Such a processing is particularly effective for technical texts where words are often similar in many languages.



```
L'ingénieur qui propose le planning le modifie.
```

```
[S]
;<ESSAI3>
;L'ingénieur qui propose le planning le modifie.
obj(change(icl>do).@entry,planning(icl>action))
agt(change(icl>do).@entry,engineer(icl>human).@def)
agt(propose(icl>do),engineer(icl>human).@def)
obj(propose(icl>do),planning(icl>action).@def)
[/S]
```

Cliquez sur le mot représenté par le pronom "le"

ingénieur
planning

Fig. 8. Anaphora resolution in the French Enconverter. The upper field contains the input text. The field in the middle the output graph. The lower field is the dialog window appearing during the Enconversion process.

```
  [S]
  obj(obtain(icl>do).@entry,alcoylbenzene.@def)
  met(obtain(icl>do).@entry,reduce(icl>do,field>chemistry))
  obj(reduce(icl>do,field>chemistry),group(icl>thing,field>chem
istry).@def)
  nam(group(icl>thing,field>chemistry).@def,carbonyle)
  [/S]
  Le? <<alcoylbenzene>> est obtenu en <<reduce>> le? <<group>>
<<carbonyle>>.
```

Fig. 9. An example of processing an unknown word.

6 Connection between nodes internal and external to a compound node

The question of the possibility of relating nodes external and internal to a given compound node seems not to be settled yet. There appears to be cases where this possibility would be very useful, if not necessary.

Let's consider for instance the left graph of figure 10. This graph is obviously ambiguous, it may express "The cat eats the mouse it caught" as well as "The cat which caught the mouse eats it". The ambiguity may be solved by introducing a compound node, but with the necessity of having an arc relating the predicate inside the compound node to its object or its agent outside the compound node.

Another possibility, avoiding arcs relating nodes inside and outside a compound node, is illustrated fig.11: the outer node is duplicated in the compound node, with the attribute @anaf indicating the peculiar nature of this duplicated node (it will often be deconverted into a pronoun).

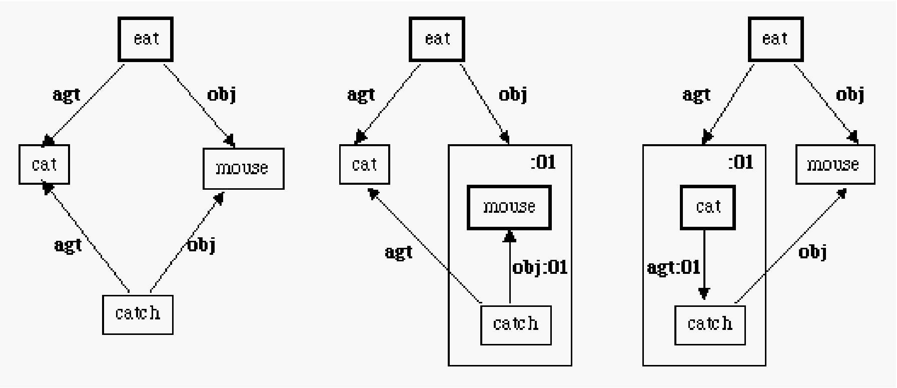


Fig. 10. The left graph is ambiguous: *The cat eats the mouse it caught / The cat which caught a mouse eats it.* The ambiguity may be solved using a scope with an arc emerging from it. The second graph expresses the meaning *The cat eats the mouse it caught*, the third one *The cat which caught a mouse eats it*

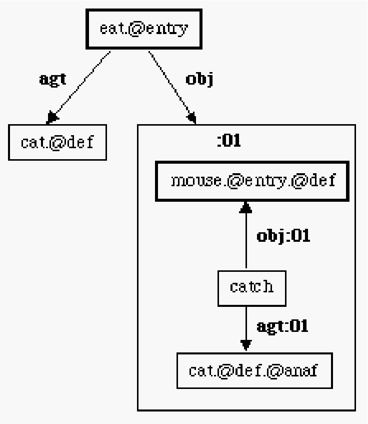


Fig. 11. Avoiding arcs connecting nodes internal and external to a compound node.

7 Evaluation and Conclusion

Evaluating the performances of a Deconverter or of an Enconverter is more difficult than evaluating a Natural Language MT system, which itself is well known to be a not so easy task.

The difficulty of evaluating an Deconverter lies in the fact that one has not only to devise the content of the test corpus, but to ensure the "linguistic" quality of this test corpus, which is of course not a problem for Natural Languages. The same applies for the evaluation of the output of an Enconverter.

As a result of several years of common work of the various UNL teams, an agreement about the correct use of the language is emerging. Nevertheless remaining discrepancies may influence the quality of the processing.

For instance, we had recently to deconvert two sets of graphs corresponding to the same source text, but encoded by two different teams. For a number of graphs, the quality of the output was not the same depending on their origin. Let's take two examples :

Example 1:

Source sentence :*The general conference adopts the universal declaration on cultural diversity*

Graph 1 :

```
agt(adopt(agt>thing,obj>thing).@entry,conference(icl>meeting))
obj(adopt(agt>thing,obj>thing.@entry,declaration(icl>information))
mod(conference(icl>meeting),general(mod<thing))
aoj(universal(aoj>thing),declaration(icl>information))
```

Graph 2:

```
agt(adopt(icl>accept(icl>do)).@entry.@present,conference(
icl>meeting).@def)
obj(adopt(icl>accept(icl>do)).@entry.@present,
declaration(icl>document).@def)
mod(conference(icl>meeting).@def, general(mod<thing))
mod(declaration(icl>document).@def, universal(aoj>thing))
```

Deconversion of graph 1: *Une conférence générale adopte une déclaration qui est universelle sur une diversité qui est culturelle.*

Deconversion of graph 2: *La conférence générale adopte la déclaration universelle sur la diversité culturelle.*

Comment : The quality of the deconversion of the second graph is quite better. The main problem lies here in the choice between the aoj relation (graph 1) and the mod relation (graph 2). We consider that the mod relation corresponds to an attributive use of the adjective, whereas the aoj relation corresponds to a predicative one. But the agreement seems to be not quite complete on this topic.

Example 2:

Source sentence *The general conference is aware of the specific mandate which has been* ***entrusted to UNESCO****, within the United Nations system, to ensure the preservation and promotion of the fruitful diversity of cultures*

Graph 1:

```
obj(aware(aoj>person,obj>thing).@entry,mandate(icl>authority).@def)
obj(entrust(agt>thing,gol>person,obj>thing).@complete,mandate(icl>author
ity).@def)
mod(mandate(icl>authority).@def,specific(mod<thing))
pur(entrust(agt>thing,gol>person,obj>thing).@complete,ensure(agt>thing,o
bj>thing))
man(entrust(agt>thing,gol>person,obj>thing).@complete,within(icl>how(obj
>thing)))
gol(entrust(agt>thing,gol>person,obj>thing).@complete,UNESCO(equ>United
Nations Educational, Scientific, and Cultural Organization))
obj(within(icl>how(obj>thing)),system(icl>functional thing).@def)
mod(system(icl>functional thing).@def,United Nations)
obj(ensure(agt>thing,obj>thing),:01.@def)
mod(:01.@def,diversity(icl>property).@def)
```

```
and:01(promotion(icl>activity).@entry,preservation(icl>state))
mod(diversity(icl>property).@def,culture(icl>abstract thing).@pl)
aoj(fruitful(aoj>thing),diversity(icl>property).@def)
```

Graph 2:

```
aoj(aware(mod<thing).@entry, conference(icl>meeting).@def)
mod(conference(icl>meeting).@def, general(mod<thing))
obj(aware(mod<thing).@entry, mandate(icl>authority).@def)
mod(mandate(icl>authority).@def, specific(mod<thing))
obj(entrust(icl>do).@complete, mandate(icl>authority).@def)
gol(entrust(icl>do).@complete, UNESCO(iof>institution).@def)
scn(entrust(icl>do).@complete, bosom(icl>abstract thing).@def)
pof(bosom(icl>abstract thing).@def, system(icl>abstract thing).@def)
pos(system(icl>abstract thing).@def, United
Nations(iof>institution).@def)
aoj(consist(icl>be), mandate(icl>authority).@def)
obj(consist(icl>be), ensure(icl>do))
obj(ensure(icl>do), promotion(icl>action).@def)
and(promotion(icl>action).@def, preservation(icl>action).@def)
obj(promotion(icl>action).@def, diversity(icl>abstract thing).@def)
mod(diversity(icl>abstract thing).@def, culture(icl>abstract
thing).@def.@pl)
mod(diversity(icl>abstract thing).@def, fruitful(mod<thing))
```

Deconversion of graph 1:

*La conférence générale est consciente du mandat spécifique qui est **confié à l'unesco** pour assurer la préservation et une promotion de la diversité de cultures qui est <fructueux> dans le système des <nations_unies>*

Deconversion of graph 2:

*La conférence générale est consciente du mandat spécifique qui est **confié pour l'unesco** dans le sein du système des <nations_unies> qui consiste que la préservation et la promotion de la diversité <fructueux> des cultures sont assurées*

Comment : We will only comment on the part of the graphs we printed in bold, corresponding to the words "***entrusted to UNESCO***" of the source text. Here the deconversion of graph 1 "***confié à l'Unesco"*** is more satisfactory than the deconversion of graph 2 "***confié pour l'Unesco"*** . The reason is that the restriction of the uw **`entrust(agt>thing,gol>person,obj>thing)`** indicates that a *gol* relation corresponds with a high probability to an argument of entrust, and not to a mere circumstantial. This allowed the enconverter to choose the right preposition. The uw **`entrust(icl>do)`** used in graph 2 didn't allow to choose the correct preposition.

But no doubt further cooperative work will soon smooth out the remaining difficulties in the use of the Universal Networking Language.

References

1. Blanc E, Sérasset G, Tsai W.J. *Structural and Lexical Transfer from a UNL Graph to an equivalent NL Dependency tree.* Proceedings of the First International Workshop on UNL, other Interlinguas and their Applications, LREC Conference. (2002)
2. Boitet, C. & Blanchon, H. *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup.* in Machine Translation. (1995). Vol. 9(2) : pp 99-132.
3. Tsai W.J. *La coédition Langue <-> UNL pour partager la révision entre langues d'un document multilingue.* Thèse d'Université. Université Joseph Fourier, Grenoble (2004).