On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter

Gilles SÉRASSET
GETA, CLIPS, IMAG
385, av. de la bibliothèque, BP 53
F-38041 Grenoble cedex 9, France
Gilles.Serasset@imag.fr

Christian BOITET
GETA, CLIPS, IMAG
385, av. de la bibliothèque, BP 53
F-38041 Grenoble cedex 9, France
Christian.Boitet@imag.fr

Abstract

After 3 years of specifying the UNL (Universal Networking Language) language and prototyping deconverters[1] from more than 12 languages and enconverters for about 4, the UNL project has opened to the community by publishing the specifications (v2.0) of the UNL language, intended to encode the meaning of NL utterances as semantic hypergraphs and to be used as a "pivot" representation in multilingual information and communication systems.

A UNL document is an html document with special tags to delimit the utterances and their rendering in UNL and in all natural languages currently handled. UNL can be viewed as the future "html of the linguistic content". It is only an interface format, leading as well to the reuse of existing NLP components as to the development of original tools in a variety of possible applications, from automatic rough enconversion for information retrieval and information gathering translation to partially interactive enconversion or deconversion for higher quality.

We illustrate these points by describing an UNL-French deconverter organized as a specific "localizer" followed by a classical MT transfer and an existing generator.

Keywords

UNL, interlingua, pivot, deconversion, UNL-French localization, transfer, generation.

Introduction

The UNL project of network-oriented multilingual communication has proposed a standard for encoding the meaning of natural language utterances as semantic hypergraphs intended to be used as pivots in multilingual information and communication systems. In the first phase (1997-1999), more than 16 partners representing 14 languages have worked to build deconverters transforming an (interlingual) UNL hypergraph into a natural language utterance.

In this project, the strategy used to achieve this initial objective is free. The UNL-French deconverter under development first performs a "localization" operation within the UNL format, and then classical transfer and generation steps, using the Ariane-G5 environment and some UNL-specific tools.

The use of classical transfer and generation steps in the context of an interlingual project may sound surprising. But it reflects many interesting issues about the status of the UNL language, designed as an interlingua, but diversely used as a linguistic pivot (disambiguated abstract English), or as a purely semantic pivot.

After introducing the UNL language, we present the architecture of the UNL-French deconverter, which "generates" from the UNL interlingua by first "localizing" the UNL form for French, within UNL, and then applying slightly adapted but classical transfer and generation techniques, implemented in the Ariane-G5 environment, supplemented by some UNL-specific tools. Then, we discuss the use of the UNL language as a linguistic or semantic pivot for highly multilingual information systems.

1 The UNL project and language

1.1 The project

UNL is a project of multilingual personal networking communication initiated by the University of United Nations based in Tokyo. The pivot paradigm is used: the representation

[1] The terms « deconversion » and « enconversion » are specific to the UNL project and are defined at paragraph 2.

of an utterance in the UNL interlingua (UNL stands for "Universal Networking Language") is a hypergraph where normal nodes bear UWs ("Universal Words", or interlingual acceptions) with semantic attributes, and arcs bear semantic relations (deep cases, such as agt, obj, goal, etc.). Hypernodes group a subgraph defined by a set of connected arcs. A UW denotes a set of interlingual acceptions (word senses), although we often loosely speak of "the" word sense denoted by a UW.
Because English is known by all UNL developers, the syntax of a normal UW is: "<English word or compound> (<list of restrictions>)", e.g. "look for (icl>action, agt>human, obj>thing)".
Going from a text to the corresponding "UNL text" or interactively constructing a UNL text is called "enconversion", while producing a text from a sequence of UNL graphs is called "deconversion".
This departure from the standard terms of analysis and generation is used to stress that this is not a classical MT project, but that UNL is planned to be the source format preferred for representing textual information in the envisaged multilingual network environment. The schedule of the project, beginning with deconversion rather than enconversion, also reflects that difference.
14 languages have been tackled during the first 3-year phase of the project (1997-1999), while many more are to be added in the second phase. Each group is free to reuse its own software tools and/or lingware resources, or to develop directly with tools provided by the UNL Center (UNU/IAS).
Emphasis is on a very large lexical coverage, so that all groups spend most of their time on the UNL-NL lexicons, and develop tools and methods for efficient lexical development. By contrast, grammars have been initially limited to those necessary for deconversion, and will then be gradually expanded to allow for more naturalness in formulating text to be enconverted.

1.2 The UNL components

1.2.1 Universal Words

The nodes of a UNL utterance are called Universal Words (or Uws). The syntax of a normal UW consists of 2 parts :

- a headword,
- a list of restrictions

Because English is known by all UNL developers, the headword is an English word or compound. The restrictions are given as an attribute value pair where attributes are semantic relation labels (as the ones used in the graphs) and values are other UWs (restricted or not).
A UW denotes a collection of interlingual acceptions (word senses), although we often loosely speak of "the" word sense denoted by an UW. For example, the unrestricted UW "look for" denotes all the word-senses associated to the English compound word "look for". The restricted UW "look for(icl>action, agt>human, obj>thing)" represents all the word senses of the English word "look for" that are an action, performed by a human that affects a thing. In this case this leads to the word sense: "look for – to try to find".

1.2.2 UNL hypergraph

A UNL expression is a hypergraph (a graph where a node is simple or recursively contains a hypergraph). The arcs bear semantic relation labels (deep cases, such as agt, obj, goal, etc.).

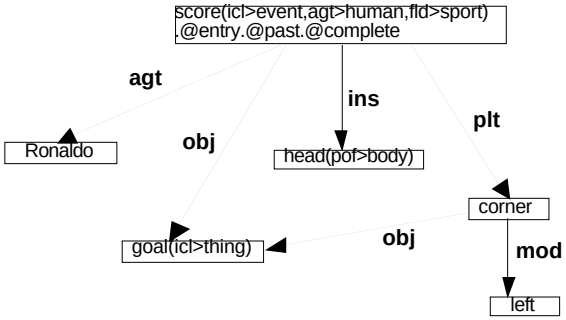

Figure 1.1: A UNL graph deconvertible as "Ronaldo has headed the ball into the left corner of the net"

In a UNL graph, UWs appear with attributes describing what is said from the speaker's point of view. This includes phenomena like speech acts, truth values, time, etc.

Hypernodes may also be used in UNL expressions.

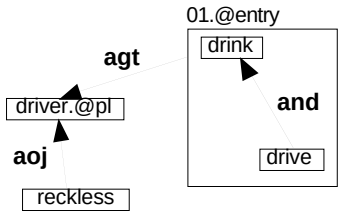

Figure 1.2: A UNL hypergraph that may be deconverted as "Reckless drivers drink and drive"

Graphs and subgraphs must contain one special node, called the entry of the graph.

1.2.3 Denoting a UNL graph

These hypergraphs are denoted using the UNL language per se. In the UNL language, an

expression consists in a set of arcs, connecting the different nodes. As an example, the graph presented in figure 1.1 will be denoted as:

```
agt(score(…).@entry.@past.@complete,
    Ronaldo)
obj(score(…).@entry.@past.@complete,
    goal(icl>thing))
ins(score(…).@entry.@past.@complete,
    head(pof>body))
plt(score(…).@entry.@past.@complete,
    corner)
obj(corner, goal(icl>thing))
mod(corner, left)
```

Hypernodes are denoted by numbers. The graph contained by a hypernode is denoted as a set of arcs colored by this number as in:

```
agt(:01.@entry, driver.@pl)
aoj(reckless, driver.@pl)
and:01(drive, drink.@entry)
```

Entries of the graph and subgraphs are denoted with the ".@entry" attribute.

2 Inside the French deconverter

2.1 Overview

Deconversion is the process of transforming a UNL graph into one (or possibly several) utterance in a natural language. Any means may be used to achieve this task. Many UNL project partners use a specialized tool called DeCo but, like several other partners, we choose to use our own tools for this purpose.
One reason is that DeCo realizes the deconversion in one step, as in some transfer-based MT systems such as METAL [17]. We prefer to use a more modular architecture and to split deconversion into 2 steps, transfer and generation, each divided into several phases, most of them written in Ariane-G5.
Another reason for not using DeCo is that it is not well suited for the morphological generation of inflected languages (several thousands rules are needed for Italian, tens of thousands for Russian, but only about 20 rules and 350 affixes suffice to build an exhaustive GM for French in Sygmor). Last, but not least, this choice allows us to reuse modules already developed for French generation.
This strategy is illustrated by figure 2.1.

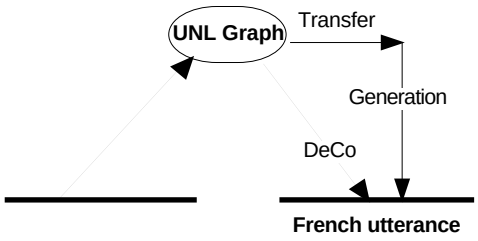

Fig. 2.1: 2 possible deconversion strategies

Using this approach, we segment the deconversion process into 7 phases, as illustrated by figure 2.2.
The third phase (graph-to-tree) produces a decorated tree which is fed into an Ariane-G5 TS (structural transfer).

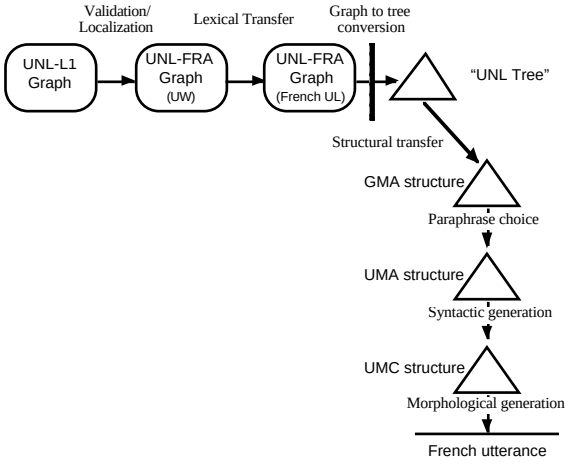

Fig. 2.2: architecture of the French deconverter

2.2 Transfer

2.2.1 Validation

When we receive a UNL Graph for deconversion, we first check it for correctness. A UNL graph has to be connected, and the different features handled by the nodes have to be defined in UNL.
If the graph proves incorrect, an explicit error message is sent back. This validation has to be performed to improve robustness of the deconverter, as there is no hypothesis on the way a graph is created. When a graph proves valid, it is accepted for deconversion.

2.2.2 Localization

In order to be correctly deconverted, the graph has to be slightly modified.
2.2.2.1 Lexical localization
Some lexical units used in the graph may not be present in the French deconversion dictionary.
This problem may appear under different circumstances. First, the French dictionary (which is still under development) may be incomplete. Second, the UW may use an unknown notation to represent a known French word sense, and third, the UW may represent a non-French word sense.
We solve these problems with the same method : Let w be a UW in the graph G. Let D be the French dictionary (a set of UWs). We substitute w in G by w' such that : $w' \in D$ and $\forall x \in D\ d(w, w', G) = d(w, x, G)$. where d is a pseudo-distance function.

If different French UWs are at the same pseudo-distance of w, w' is chosen at random among these UWs (default in non-interactive mode).

2.2.2.2 "Cultural" localization

Some crucial information may be missing, depending on the language of the source utterance (sex, modality, number, determination, politeness, kinship…).

It is in general impossible to solve this problem fully automatically in a perfect manner, as we do not know anything about the document, its context, and its intended usage: FAHQDC[2] is no more possible than FAHQMT on arbitrary texts. We have to rely on necessarily imperfect heuristics.

However, we can specialize the general French deconverter to produce specialized servers for different tasks and different (target) sublanguages. It is possible to assign priorities not only to various parts of the dictionaries (e.g., specialized vs. general), but also to equivalents of the same UW within a given dictionary. We can then define several user profiles. It is also possible to build a memory of deconverted and possibly postedited utterances for each specialized French deconversion server.

2.2.3 Lexical Transfer

After the localization phase, we have to perform the lexical transfer. It would seem natural to do it within Ariane-G5, after converting the graph into a tree. But lexical transfer is context-sensitive, and we want to avoid the possibility of transferring differently two tree nodes corresponding to one and the same graph node.

Each graph node is replaced by a French lexical unit (LU), along with some variables. A lexical unit used in the French dictionary denotes a derivational family (e.g. in English: destroy denotes destroy, destruction, destructible, destructive…, in French: détruire for détruire, destruction, destructible, indestructible, destructif, destructeur).

There may be several possible lexical units for one UW. This happens when there is a real synonymy or when different terms are used in different domains to denote the same word sense[3]. In that case, we currently choose the lexical unit at random as we do not have any information on the task the deconverter is used for.

The same problem also appears because of the strategy used to build the French dictionary. In order to obtain a good coverage from the beginning, we have underspecified the UWs and linked them to different lexical units. This way, we considered a UW as the denotation of a set of word senses in French.

Hence, we were able to reuse previous dictionaries and we can use the dictionary even if it is still under development and incomplete. In our first version, we also solve this problem by a random selection of a lexical unit.

2.2.4 Graph to tree conversion

The subsequent deconversion phases are performed in Ariane-G5. Hence, it is necessary to convert the UNL hypergraph into an Ariane-G5 decorated tree.

The UNL graph is directed. Each arc is labelled by a semantic relation (agt, obj, ben, con…) and each node is decorated by a UW and a set of features, or is a hypernode. One node is distinguished as the "entry" of the graph.

An ARIANE tree is a general (non binary) tree with decorations on its nodes. Each decoration is a set of variable-value pairs.

The graph-to-tree conversion algorithm has to maintain the direction and labelling of the graph along with the decoration of the nodes.

Our algorithm splits the nodes that are the target of more than one arc, and reverses the direction of as few arcs as possible. An example of such a conversion is shown in figure 2.3.

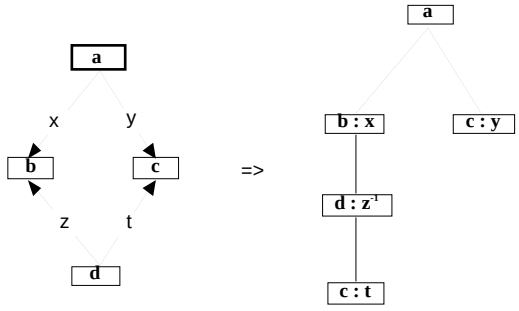

Fig. 2.3: example graph to tree conversion

Let Σ be the set of nodes of G, Λ the set of labels, T the created tree, and N is the set of nodes of T.

The graph $G = \{ (a,b,l) \mid a \in \Sigma, b \in \Sigma, l \in \Lambda \}$ is defined as a set of directed labelled arcs. We use an association list $A = \{ (n_G,n_T) \mid n_G \in \Sigma, n_T \in N \}$, where we memorize the correspondence between nodes of the tree and nodes of the graph.

---

[2] fully automatic high quality deconversion.

[3] strictly speaking, the same collection of interlingual word senses (acceptions).

```
let  e_G ∈ Σ such that e is the entry of G
     e_T ← new tree-node(e_G,entry)
in T ← e_T(); N ← {e_T}; A ← {(e_G,e_T)}
   while G ≠ ∅ do
      if there is (a,b,l) in G such that (a,a_T) ∈ A then
         G ← G\(a,b,l);
         b_T ← new tree-node(b,l);
         A ← A ∪ {(b,b_T)};
         let  a_T ∈ N such that (a,a_T) ∈ A
         in   add b_T to the daughters of a_T;
      else if there is (a,b,l) in G such that (b,b_T) ∈ A then
         G ← G\(a,b,l);
         a_T ← new tree-node(a,l^-1);
         A ← A ∪ {(a,a_T)};
         let  b_T∈ N such that (b,b_T) ∈ A
         in   add a_T to the daughters of b_T;
      else exit on error ("non connected graph");
```

2.2.5 Structural transfer

The purpose of the structural transfer is to transform the tree obtained so far into a Generating Multilevel Abstract (GMA) structure [4].
In this structure, non-interlingual linguistic levels (syntactic functions, syntagmatic categories…) are underspecified, and (if present), are used only as a set of hints for the generation stage.

2.3 Generation

2.3.1 Paraphrase choice

The next phase is in charge of the paraphrase choice. During this phase, decisions are taken regarding the derivation applied to each lexical unit in order to obtain the correct syntagmatic category for each node. During this phase, the order of appearance and the syntactic functions of each parts of the utterance is also decided. The resulting structure is called Unique Multilevel Abstract (UMA) structure.

2.3.2 Syntactic and morphological generation

The UMA structure is still lacking the syntactic sugar used in French to realize the choices made in the previous phase by generating articles, auxiliaries, and non connected compunds such as ne…pas, etc.
The role of this phase is to create a Unique Multilevel Concrete (UMC) structure. By concrete, we mean that the structure is projective, hence the corresponding French text may be obtained by a standard left to right traversal of the leaves and simple morphological and graphemic rules. The result of these phases is a surface French utterance.

3 Different uses of the UNL language

3.1 Hypergraphs vs colored graphs

As presented in section 1.2.3, the syntax of the UNL language is based on the description of a graph, arc by arc. Some of these arcs are "coloured" by a number. This colouring is currently interpreted as hypernodes (nodes containing a graph, rather than a classical UW). This interpretation is arbitrary and imposes semantic constraints on a UNL utterance:

- the subgraph (the set of arcs labeled with the same colour) is connected,
- arcs with different colours cannot be connected to the same node.

However, even if one uses the UNL language for a particular kind of application, a different interpretation may be chosen. By adding new semantic constraints to UNL expressions, one may restrict to the use of trees. On the contrary, by loosening semantic constraint, one may use colored graphs instead of the more restrictive hypergraphs.
This flexibility of UNL may lead to uses that differ from the computer science point of view (different structures leading to different kinds of methods and applications) as well as from the linguistic point of view (different ways to represent the linguistic content of a utterance).
This kind of structure is very useful to represent some utterances like “Christian pulls Gilles’ leg”. Using a colored graph, one can represent the utterance with the graph shown in figure 3.1, which is not a hypergraph.

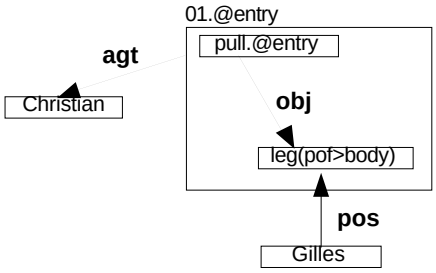

Figure 3.1: this graph is not an hypergraph, it can however be represented in UNL language

When using normal hypergraphs, one could only represent the utterance as shown in figure 3.2.

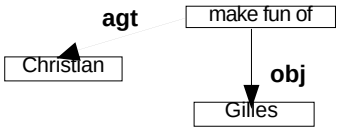

Figure 3.2: this graph is a valid hypergraph

Hence, keeping backward compatibility with other UNL based systems, one may develop an entirely new and more powerful kind of application.

3.2 Linguistic vs semantic pivot

The UNL language defines the interface structure to be used by applications (either a hypergraph or a colored graph). However, it does not restrict the choice of the data to be encoded.

Since the beginning, two possible and valid approaches has been mentioned. During the kickoff meeting of the UNL project, Pr. Tsujii promoted the use of UNL as a linguistic pivot. With this approach, a UNL utterance should be the encoding of the deep structure of a valid English utterance that reflects the meaning of the source utterance. With this approach, the German sentence "Hans schwimt sehr gern" should be encoded as shown in figure 3.3.

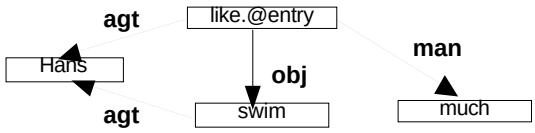

Figure 3.3: a linguistic encoding of "Hans schwimt sehr gern"

On the opposite, Hiroshi Uchida promotes the use of UNL as a semantic pivot. With this second approach, the same sentence should be encoded as shown in figure 3.4.

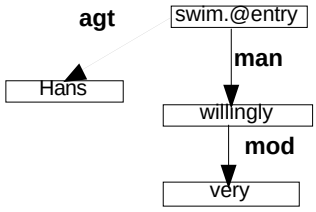

Figure 3.4: a semantic encoding of "Hans schwimt sehr gern"

Each approach has its advantages and drawbacks and the choice between them can only be made with an application in mind. The linguistic approach leads to a better quality in the produced results and is an answer to highly multilingual machine translation projects. With this approach, the UNL graphs can only be produced by people mastering English or by (partially) automatic enconverters.

With the semantic approach, subtle differences in source utterances (indefinite, reflexivity…) can not be expressed, leading to a lower quality. However, using this approach, the UNL encoding is much more natural and easy to perform by a non English speaker (as the semantic relations and UWs are expressed at the source level). Hence, this approach is to be used for multilingual casual communication where users may express themselves by directly encoding UNL expressions with an appropriate editing tool.

Conclusion

Working on the French deconverter has led to an interesting architecture where deconversion, in principle a "generation from interlingua", is implemented as transfer + generation from an abstract structure (UNL hypergraph) produced from a NL utterance. The idea to use UNL for directly creating documents gets here an indirect and perhaps paradoxical support, although it is clear that considerable progress and innovative interface design will be needed to make it practical.

However, the UNL language proves flexible enough to be used by very different projects. Moreover, with deconverters currently developed for 14 languages, joining the UNL project is really attractive. Let's hope that this effort will help breaking the language barriers.

Acknowledgements

We would like to thank the sponsors of the UNL project, especially UNU/IAS (T. Della Senta) & ASCII (K.Nishi) and of the UNL-FR subproject, especially UJF (C. Feuerstein), IMAG (J. Voiron), CLIPS (Y. Chiaramella), and the

French Ministery of Foreign Affairs (Ph. Perez), as well as the members of UNL Center, especially project leader H. Uchida, M. L. Zhu, and K. Sakai. Last but not least, other members of GETA have contributed in many ways to the research reported here, in particular N. Nédeau, E. Blanc, M. Mangeot, J. Sitko, L. Fischer, M. Tomokiyo, and K. Fort.

References

[1] **Blanc É. & Guillaume P. (1997)** *Developing MT lingware through Internet : ARIANE and the CASH interface.* Proc. Pacific Association for Computational Linguistics 1997 Conference (PACLING'97), Ohme, Japon, 2-5 September 1997, vol. 1/1, pp. 15-22.

[2] **Blanchon H. (1994)** *Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mockup.* Proc. 15th International Conference on Computational Linguistics, COLING-94, 5-9 Aug. 1994, vol. 1/2, pp. 115—119.

[3] **Boitet C., Réd. (1982)** *"DSE-1"— Le point sur ARIANE-78 début 1982.* Contrat ADI/CAP-Sogeti/Champollion (3 vol.), GETA, Grenoble, février 1982, 400 p.

[4] **Boitet C. (1994)** *Dialogue-Based MT and self-explaining documents as an alternative to MAHT and MT of controlled languages.* Proc. Machine Translation 10 Years On, 11-14 Nov. 1994, Cranfield University Press, pp. 22.1—9.

[5] **Boitet C. (1997)** *GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects.* Proc. PACLING-97, Ohme, 2-5 September 1997, Meisei University, pp. 23-57. (invited communication)

[6] **Boitet C. & Blanchon H. (1994)** *Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup.* Machine Translation, **9**/2, pp. 99—132.

[7] **Boitet C., Guillaume P. & Quézel-Ambrunaz M. (1982)** *ARIANE-78, an integrated environment for automated translation and human revision.* Proc. COLING-82, Prague, July 1982, pp. 19—27.

[8] **Brown R. D. (1989)** *Augmentation.* Machine Translation, 4, pp. 1299-1347.

[9] **Ducrot J.-M. (1982)** *TITUS IV.* In "Information research in Europe. Proc. of the EURIM 5 conf. (Versailles)", P. J. Taylor, ed., ASLIB, London.

[10] **Kay M. (1973)** *The MIND system.* In "Courant Computer Science Symposium 8: Natural Language Processing", R. Rustin, ed., Algorithmics Press, Inc., New York, pp. 155-188.

[11] **Maruyama H., Watanabe H. & Ogino S. (1990)** *An Interactive Japanese Parser for Machine Translation.* Proc. COLING-90, Helsinki, 20-25 août 1990, ACL, vol. 2/3, pp. 257-262.

[12] **Melby A. K., Smith M. R. & Peterson J. (1980)** *ITS : An Interactive Translation System.* Proc. COLING-80, Tokyo, 30/9-4/10/80, pp. 424—429.

[13] **Moneimne W. (1989)** (159 p. +annexes)*TAO vers l'arabe. Spécification d'une génération standard de l'arabe. Réalisation d'un prototype anglais-arabe à partir d'un analyseur existant.* Nouvelle thèse, UJF.

[14] **Nirenburg S. & al. (1989)** *KBMT-89 Project Report.* Center for Machine Translation, Carnegie Mellon University, Pittsburg, April 1989, 286 p.

[15] **Nyberg E. H. & Mitamura T. (1992)** *The KANT system: Fast, Accurate, High-Quality Translation in Practical Domains.* Proc. COLING-92, Nantes, 23-28 July 92, ACL, vol. 3/4, pp. 1069—1073.

[16] **Quézel-Ambrunaz M. (1990)** *Ariane-G5 v.3 - Le moniteur.* GETA, IMAG, juin 1990, 206 p.

[17] **Slocum J. (1984)** *METAL: the LRC Machine Translation system.* In "Machine Translation today: the state of the art (Proc. third Lugano Tutorial, 2–7 April 1984)", M. King, ed., Edinburgh University Press (1987).

[18] **Wehrli E. (1992)** *The IPS System.* Proc. COLING-92, Nantes, 23-28 July 1992, vol. 3/4, pp. 870-874.