

Une solution au problème de la désambiguïsation interactive

Hervé BLANCHON

GETA, Institut IMAG (UJF & CNRS)
BP 53X, 38041 Grenoble Cedex, France
e-mail : blanchon@imag.fr

Résumé

Dans le cadre du projet de TAO personnelle LIDIA [1,3,4], un dialogue entre le système et le rédacteur permet d'obtenir une traduction de qualité. Ce dialogue a deux objectifs : standardiser et désambiguïser le texte. La standardisation concerne l'orthographe, le vocabulaire, les formules figées et quelques mesures stylistiques. La clarification concerne toutes les ambiguïtés que l'on peut rencontrer lorsque l'on fait de la traduction automatique.

Ici, nous ne nous intéressons qu'à une partie des traitements de désambiguïsation. Nous ne considérons que les problèmes repérables lors de l'examen de la structure MMC produite par l'analyseur, ambiguïté de classe, ambiguïté de géométrie et ambiguïté de décoration syntaxique, logique et sémantique.

Nous examinons le type de structure d'où nous partons. Nous définissons alors les trois grandes classes d'ambiguïtés que nous proposons. Nous décrivons ensuite notre méthode en fournissant trois principes, une stratégie et les traitements qui nous permettent, en raffinant les classes d'ambiguïtés, d'obtenir finalement le dialogue. Cette méthode sera illustrée sur quelques exemples.

Mots Clé

Traduction Automatique, TAO personnelle, Désambiguïsation Interactive, Production de Dialogue

I. La structure MMC

L'analyseur utilisé dans le cadre du projet est un analyseur Multisolution et la structure qu'il produit a deux caractéristiques : elle est Multiniveau et Concrète.

1. Multisolution

La structure produite est dite "multisolution" car, pour chaque phrase, on produit toutes les analyses vérifiant le modèle syntagmatique, syntaxique et logico-sémantique des grammaires utilisées. En effet, comme on veut obtenir l'analyse qui correspond aux intentions de l'auteur, il nous faut produire toutes les analyses vérifiant le modèle pour que le rédacteur aide le système à choisir la bonne.

2. Multiniveau

La structure produite est dite "multiniveau" car les nœuds portent des décorations complexes qui

représentent divers niveaux d'interprétation : le niveau des classes syntaxiques et des classes syntagmatiques, le niveau des fonctions syntaxiques et enfin le niveau des relations logiques et sémantiques.

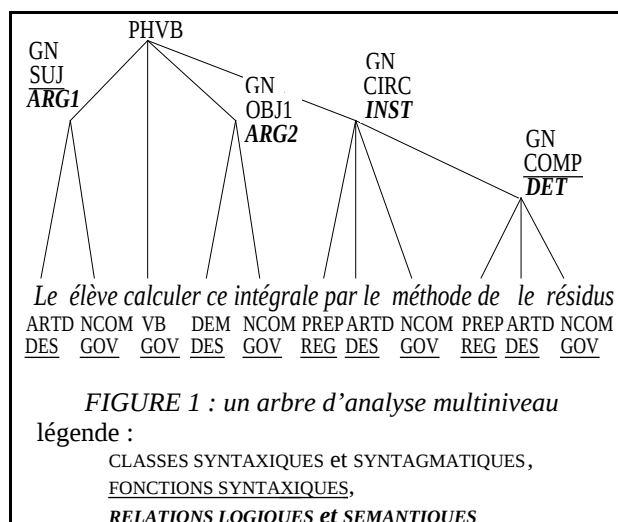
Pour préciser les choses, voici une brève description de ces divers niveaux d'interprétation.

- Le niveau des classes syntaxiques (classes terminales telles que : "NOM COMMUN", "ADJECTIF", "PRÉPOSITION", etc.) et des classes syntagmatiques (classes non terminales telles que : "GROUPE NOMINAL", "PHRASE INFINITIVE", "PHRASE VERBALE", etc.) donne le parenthésage de la phrase en groupes syntagmatiques. Les feuilles de l'arbre portent les classes syntaxiques et les nœuds non terminaux portent les classes syntagmatiques.

- Le niveau des fonctions syntaxiques donne le rôle syntaxique de chaque nœud dans le groupe auquel il est attaché directement ("GOUVERNEUR" ou "TÊTE", "RÉGISSEUR", "SUJET", etc.). Avec ce niveau, on peut très simplement transformer l'arborescence en une arborescence de "dépendance syntaxique".

- Le niveau des relations logiques et sémantiques peut être considéré comme celui de la construction logique de la phrase, mettant en évidence les places des arguments attachés aux unités lexicales prédicatives et donnant l'interprétation sémantique des compléments et circonstanciels ("DÉTERMINATION", "PREMIER ARGUMENT D'UNE UNITÉ LEXICALE PRÉDICATIVE", "CIRCONSTANCIEL INSTRUMENTAL", etc.).

Voici l'arbre d'analyse de la phrase "l'élève calcule cette intégrale par la méthode des résidus".



3. Concrète

La structure produite est dite “concrète”, car on retrouve directement le texte analysé (*en italique dans la figure 1*) via le mot des feuilles de l’arbre. Cette propriété permet d’avoir un processus de génération des questions plus simple à mettre en œuvre.

Pour produire le dialogue à partir de la structure MMC, nous sommes d’abord amenés à définir les types d’ambiguïté qui apparaissent dans la structure.

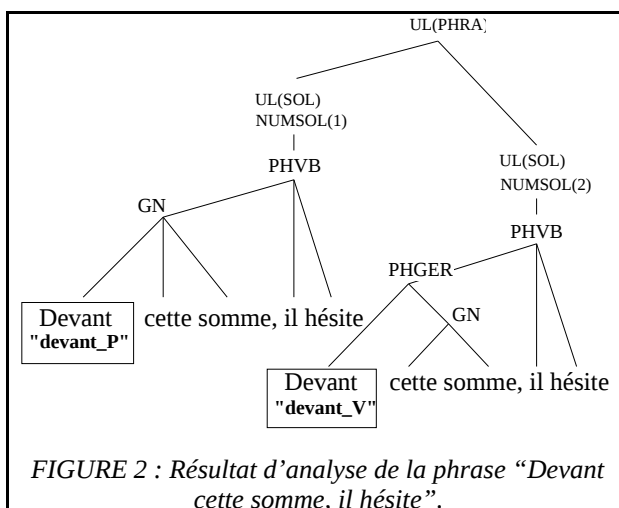
II. Les ambiguïtés dans la structure MMC

Nous considérons trois types d’ambiguïté, que nous traiterons différemment.

1. Ambiguïté de classe

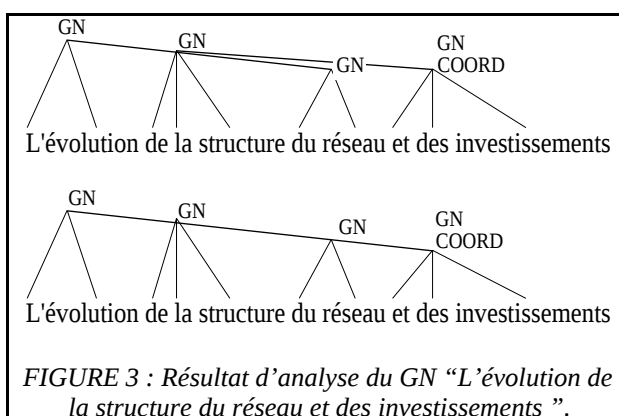
Il y a ambiguïté de classe si une même occurrence est étiquetée par deux classes morpho-syntaxiques différentes dans deux arbres d’analyse distincts.

Le schéma suivant (figure 2) montre les deux arbres produits par l’analyse de la phrase “Devant cette somme, il hésite” dans laquelle ‘devant’ peut être interprété comme gérondif du verbe devoir ou bien comme préposition.



2. Ambiguïté de géométrie

Il y a ambiguïté de géométrie si deux arbres d’analyse différents ont des graphes distincts.

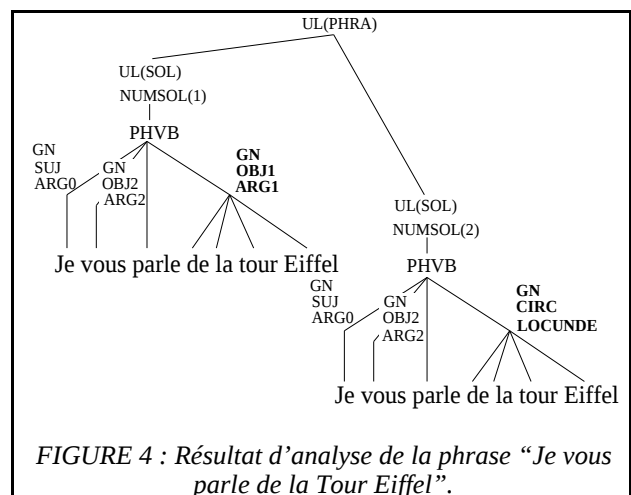


La figure 3 montre les deux arbres produits pour le groupe “L’évolution de la structure du réseau et des investissements” où l’on peut comprendre *l’évolution de la structure du réseau* **et** *l’évolution des investissements* (NUMSOL(1)) ou bien *l’évolution de la structure du réseau* **et** *l’évolution de la structure des investissements* (NUMSOL(2)).

3. Ambiguïté de décoration syntaxique, logique et sémantique

Il y a ambiguïté de décoration syntaxique, logique et sémantique si pour des arbres ayant tous les mêmes groupes simple¹, l’étiquetage des nœuds non terminaux en fonction syntaxique et/ou en relation logique et sémantique est différent.

La figure 4 montre les deux arbres produits par l’analyse de la phrase “Je vous parle de la tour Eiffel” où l’on peut comprendre qu’on nous parle depuis la tour Eiffel (NUMSOL(2)) ou qu’on nous parle au sujet de la tour Eiffel (NUMSOL(1)).



III. Méthodologie

1. Principes

Nous avons défini trois principes qui fondent la stratégie que nous allons décrire plus bas. Les voici, par ordre d’importance :

- 1) trouver d’abord les bons groupes simples
- 2) trouver ensuite le sujet, les arguments et les circonstants de tous les prédicats
- 3) trouver enfin la structure du sujet, des arguments et des circonstants des prédicats

Ces trois principes ont été guidés par des raisons pragmatiques. Nous voulons en effet, et avant toute autre chose, trouver le bon découpage en groupes simples de

¹ Un groupe simple est constitué de tout gouverneur accompagné de sa complémentation à gauche. Exemple “le chien”, “le tout petit chien”. Un gouverneur est un lexème qui n’est pas un mot outil. Exemple : “le” et “tout” sont des mots outil.

la phrase. Ceux-ci forment, en effet, les éléments de base qui permettent de produire le sens. Nous voulons ensuite choisir la construction de la phrase au plus haut niveau, c'est-à-dire la construction des prédicats. C'est finalement, lorsque l'on a construit correctement les prédicats que l'on peut s'intéresser à la structure interne des éléments qui les composent.

2. Stratégie

Comme plusieurs propriétés peuvent apparaître dans un résultat d'analyse, il faut lever les ambiguïtés dans un ordre qui respecte les trois principes que nous avons posés.

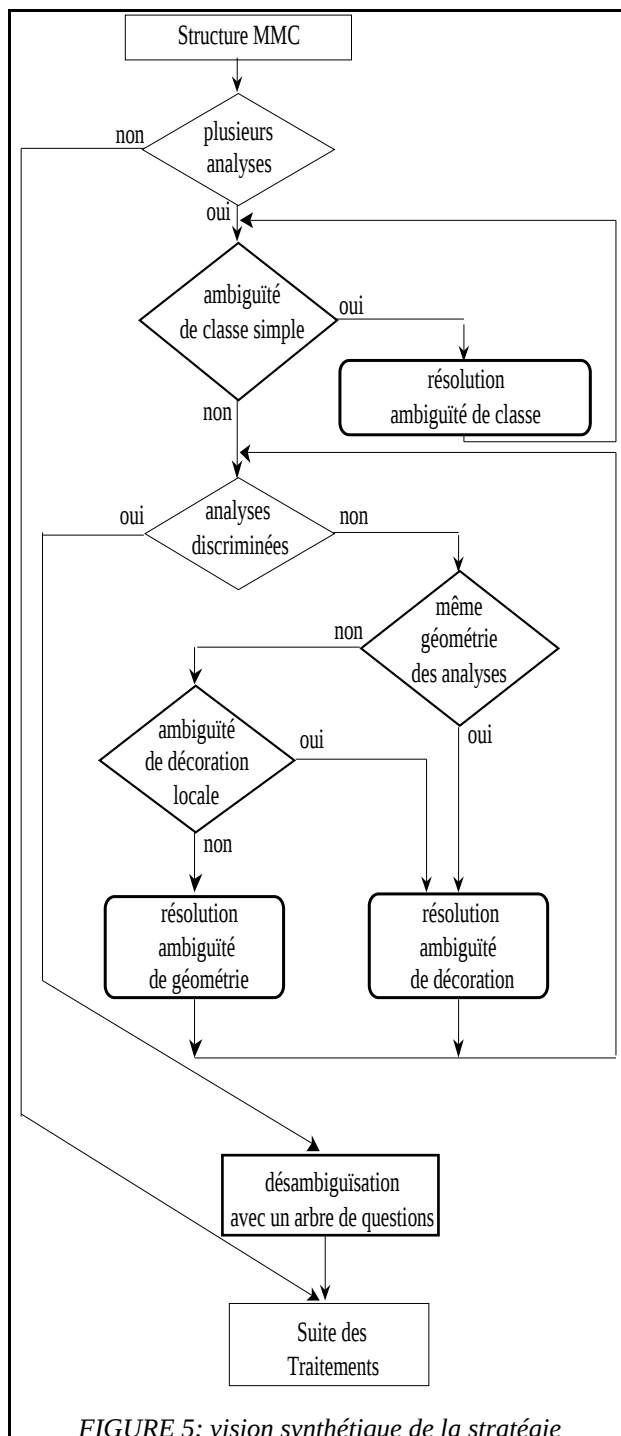


FIGURE 5: vision synthétique de la stratégie

Les groupes simples étant pour nous les briques de la phrase, il est nécessaire de les obtenir le plus rapidement possible. Si, pour toutes les occurrences d'une phrase, les classes syntaxiques associées à chacune d'elles dans les diverses solutions d'analyse produites sont les mêmes, alors ces solutions sont composées des mêmes groupes simples. Si les classes sont différentes, les occurrences qui présentent cette propriété appartiennent ou non à un groupe coordonné de la phrase. Si elles n'appartiennent pas à un groupe coordonné, nous résoudrons une ambiguïté de classe. Nous appellerons ce cas ambiguïté de classe simple. Si elles appartiennent à un groupe coordonné, nous résoudrons le problème par une autre méthode. Nous appellerons ce dernier cas ambiguïté de classe avec occurrences coordonnées.

Lorsqu'on dispose des bons groupes simples, on peut chercher à construire le sujet, les objets et les circonstants de tous les prédicats. Pour ce faire, on localise la définition de l'ambiguïté de décoration en imposant que la propriété soit vérifiée pour une PHrase VerBale (PHVB) et non vérifiée par les groupes qui lui sont coordonnés. Cette ambiguïté sera appelée ambiguïté de décoration locale.

Nous allons maintenant itérer une série de tests et d'actions jusqu'à produire un arbre de questions permettant de désambiguïser totalement la phrase. Chaque action a pour rôle de produire une partition de l'ensemble des arbres qu'elle reçoit et une question qui permet de choisir la bonne parmi les classes créées. Chaque classe ainsi créée est une nouvelle donnée de l'itération qui fabriquera si nécessaire une nouvelle question et de nouvelles classes. L'arbre de questions sera totalement produit lorsque la partition de l'ensemble des analyses initialement produites ne comportera que des classes à un seul élément.

La stratégie de production de l'arbre de questions est résumé dans la figure 5.

3. Traitement

Nous allons examiner les trois types d'ambiguïté que nous avons définis et montrer comment on produit le dialogue pour chacun d'eux. Nous ne différencions pas ici l'ambiguïté de décoration et l'ambiguïté de décoration locale.

3.1 Ambiguïté de classe

Pour résoudre une ambiguïté de classe, on projette la partie de la phrase pour laquelle les classes diffèrent selon deux composantes, l'occurrence et la classe qui lui est associée. On pourra aussi proposer un dialogue plus détaillé.

3.2 Ambiguïté de géométrie

Pour résoudre une ambiguïté de géométrie en respectant le second et le troisième principe, nous distinguons hiérarchiquement les problèmes suivants :

I. coordination verbale : problème de coordination pour lequel un même groupe peut être une PHrase VerBale coordonnée ou non. C'est le cas d'ambiguïté de

classe avec occurrences coordonnées que nous avons écarté tout à l'heure. Pour résoudre un tel problème, nous faisons un paraphrasage, par permutation, de chaque analyse qui met en valeur tout ce qui a pu être éliminé, car nous disposons d'un modèle unique de phrase pour ce cas.

II. structure argumentaire du verbe : le verbe peut être complété de différentes façons. Par exemple, il peut être doté d'un sujet et d'un circonstant, ou bien, d'un sujet et d'un objet, ou enfin, d'un sujet, d'un objet, et d'un circonstant. Ici on projette la partie de la phrase qui permet de distinguer les différentes possibilités en remplaçant la préposition ambiguë qui introduit les compléments par des prépositions non ambiguës.

III. coordination non verbale : problème de coordination qui n'est pas lié à un problème de classe, le groupe coordonné n'est jamais une PHVB. Ici, on distribue ce qui peut l'être pour retrouver les interprétations sous une forme complètement explicite.

IV. subordination : un groupe établit différents liens de subordination avec les groupes qui le précèdent. Par exemple, il peut être subordonné au groupe qui lui est contigu immédiatement à gauche ou à un groupe non contigu qui le précède. Dans ce cas, on rapproche par permutation les groupes subordonnés qui ont été séparés par un groupe intermédiaire. On se sert de parenthèses pour bien montrer les mots que l'on a regroupés.

3.3 Ambiguïté de décoration

Résoudre une ambiguïté de décoration ou une ambiguïté de décoration locale nécessite les mêmes traitements. Dans le premier cas, on travaille sur un arbre tout entier, alors que dans l'autre on travaille sur un sous-arbre. On distingue les problèmes suivants :

I. étiquetage logico-sémantique : seules les relations logiques et sémantiques diffèrent entre les analyses. Il suffit de remplacer la préposition introduisant le groupe à problème par les prépositions introduisant le groupe à problème par les prépositions non ambiguës qui représentent chacune des relations possibles.

II. ordre des arguments d'un verbe transitif direct : à une permutation près, tous les arbres ont le même étiquetage. On reconstruit par permutation toutes les interprétations selon un ordre "standard" sous une forme normalisée. On ne conserve que le gouverneur de chaque groupe sujet ou objet, et on le munit du bon article avant de faire les projections dans l'ordre sujet + verbe + objet 1 + objet 2.

III. étiquetage syntaxique : il y a ambiguïté sur le rôle d'un groupe subordonné au verbe ou à un autre groupe simple. Il suffit de projeter la phrase en remplaçant la préposition ambiguë introduisant le groupe qui pose problème par des prépositions ou des groupes sémantiquement équivalents non ambiguës.

Les informations nécessaires au remplacement d'une occurrence se trouvent dans la base lexicale.

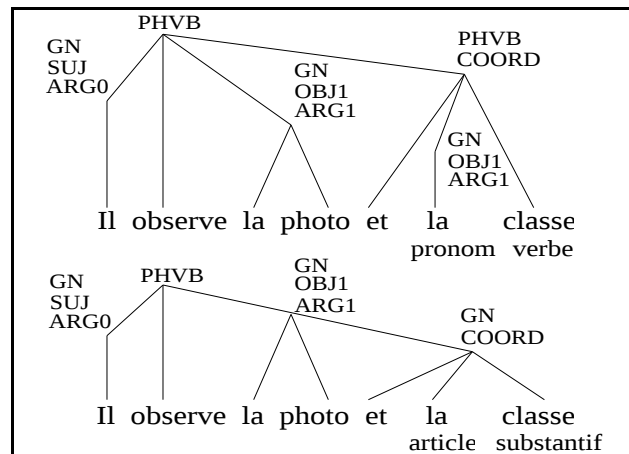
IV. Exemples

Voici deux exemples de dialogues que l'on peut produire en utilisant la méthode décrite ci-dessus.

La coordination verbale

Phrase : Il observe la photo et la classe.

Arbres :



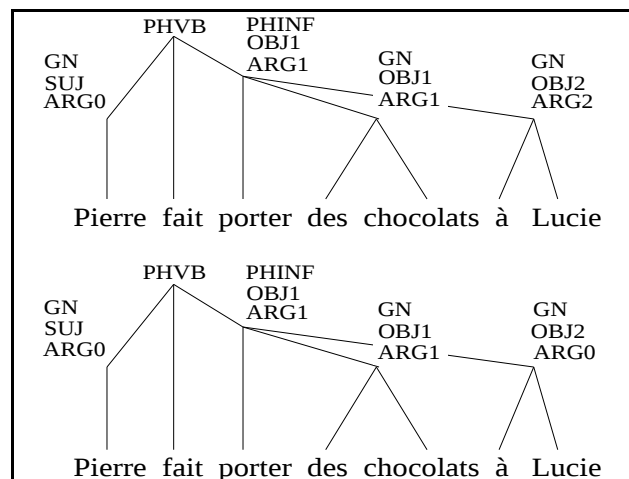
Dialogue :

ambiguïté
l'analyseur a rencontré un problème de classes pour la phrase : <i>Il observe la photo et la classe.</i>
Que faut-il comprendre ?
☒ il observe la classe ☐ il classe la photo
OK

L'étiquetage logico-sémantique

Phrase : Pierre fait porter des chocolats à Lucie.

Arbres :



Dialogue :

ambiguïté
L'analyseur a rencontré un problème de décoration pour la phrase : <i>Pierre fait porter des chocolats à Lucie.</i>
Que faut-il comprendre ?
☒ Pierre fait porter des chocolats pour Lucie.
☐ Pierre fait porter des chocolats par Lucie.

La coordination non verbale

Phrase : l'évolution de la structure du réseau et des investissements est étudiée

Arbre : cf figure 3

Dialogue :

ambiguïté
L'analyseur a rencontré un problème de coordination pour la phrase : <i>L'évolution de la structure du réseau et des investissements est étudiée.</i>
Que faut-il comprendre ?
☐ L'évolution de la structure du réseau et l'évolution des investissements
☐ L'évolution de la structure du réseau et l'évolution de la structure des investissements
OK

Conclusion

Vis à vis de la structure MMC que nous utilisons, une phrase est désambiguïsée lorsqu'on a trouvé les bons groupes simples, le bon rattachement de ceux-ci, les bonnes qualités de ces rattachements. La méthode que nous proposons traite tous ces problèmes. Pour trouver les bons groupes simples, il suffit de trouver la classe syntaxique de chaque occurrence. Pour trouver le bon rattachement des groupes simples, il faut gérer les problèmes de subordination et de coordination. Nous avons pris en compte le cas où ces problèmes influencent la construction des prédicats, ici intervient la qualité des rattachements. Enfin, nous résolvons le cas où se pose uniquement un problème de qualité des rattachements en différenciant les trois types de problèmes qui peuvent se poser.

Afin de diminuer le nombre des analyses produites et de réduire ainsi le temps à accorder au processus de désambiguïsation, nous réfléchissons aux moyens dont pourrait disposer l'utilisateur pour fournir au système des informations sur ce qu'il vient de rédiger. Ces informations devraient permettre de restreindre le modèle syntagmatique et/ou syntaxique et/ou logico-sémantique de l'analyseur utilisé.

La méthode actuelle est figée. Nous désirons aller vers plus de généricité en offrant au linguiste un environnement avec lequel il puisse définir d'autres processus de désambiguïsation. Cela veut dire qu'il faut des outils qui permettent de définir, de façon explicite ou non, des types d'ambiguïté, des principes, une stratégie et des méthodes de production du dialogue, comme nous l'avons fait ici. Dans notre approche, les principes sont implicitement pris en compte dans la stratégie et dans les traitements. On peut par exemple imaginer que le linguiste décide de faire appel à des processus de génération qu'il aura développés sur le serveur de traduction.

Remerciements

Je tiens à remercier Jean-Philippe Guilbaud pour son assistance linguistique. Je remercie également Christian Boitet, Mathieu Lafourcade, Kiki Levenbach, Gilles Sérasset qui ont corrigé ce document dans son fond et dans sa forme ; je suis bien sûr seul responsable des éventuelles imperfections qui subsisteraient.

Bibliographie

- [1] Blanchon, H., (1990). *LIDIA-1 : Un prototype de TAO personnelle pour rédacteur unilingue*. Proc. Avignon-90 : Le traitement du langage naturel et ses applications. Avignon. 28 mai-1 juin, 1990.
- [2] Blanchon, H., (1992). *Désambiguïsation interactive en TAO personnelle : poser les bonnes questions!*. Proc. Avignon-92 : Le traitement du langage naturel et ses applications. Avignon. 3-5 juin, 1992.
- [3] Boitet, C., (1989). *Motivation and Architecture of the Lidia Project*. Proc. MTS-89. Munich. 16-18 août, 1989.
- [4] Boitet, C., (1990). *Towards Personal MT : on some aspects of the LIDIA project*. Proc. Coling-90. Helsinki. 20-25 août, 1990. vol. 3/3 : pp. 30-35.
- [5] Brown, R. D. & Nirenburg, S., (1990). *Human-Computer Interaction for Semantic Disambiguation*. Proc. Coling-90. Helsinki. August 20-25, 1990. vol. 3/3 : pp. 42-47.
- [6] Chandler, B. & Wood, M., (1988). *Machine Translation For Monolinguals*. Proc. Coling-88. Budapest. 1988 : pp. 760-763.
- [7] Huang, X., (1990). *A Machine Translation System for the Target Language Inexpert*. Proc. Coling-90. Helsinki. August 20-25, 1990. vol. 3/3 : pp. 364-367.
- [8] Maruyama, H., Watanabe, H. & Ogino, S., (1990). *An interactive Japanese Parser for Machine Translation*. Proc. Coling-90. Helsinki. August 20-25, 1990. vol. 2/3 : pp. 257-262.
- [9] Melby, A. K., (1982). *Multi-Level Translation Aids in a Distributed System*. Proc. Coling-82. Prague. July 5-10, 1982. vol. 1/1 : pp. 215-220.
- [10] Wehrli, E., (1990). *STS : An Experimental Sentence Translation System*. Proc. Coling-90. Helsinki. August 20-25, 1990. vol. 1/3 : pp. 76-78.
- [11] Zajac, R., (1988). *Interactive Translation : a new approach*. Proc. Coling-88. Budapest. August 22-27, 1988.