

Interagir pour traduire : la TAO personnelle pour rédacteur monolingue

Hervé Blanchon

GETA, Institut IMAG (UJF & CNRS), BP 53, 38041 Grenoble Cedex 9

Notre recherche actuelle concerne un nouveau concept de TAO (Traduction Automatisée par Ordinateur) : la TAO personnelle ou TA Fondée sur le Dialogue pour rédacteur monolingue. Il s'agit de permettre à un rédacteur de traduire ses documents dans des langues qu'il ne connaît pas. Il faut donc produire des traductions de haute qualité. Cet objectif n'est pas réalisable si le système est complètement autonome : il sera donc interactif. Ainsi, les énoncés ambigus seront clarifiés au moyen de questions. Pour que le système soit personnel, il ne faut pas que celles-ci fassent appel à des connaissances linguistiques d'expert. Seul le niveau du baccalauréat peut être requis. Poussant plus loin des expériences déjà conduites dans le domaine des systèmes de TAO interactifs, nous souhaitons montrer la faisabilité d'un système personnel grand public. Pour produire le dialogue, nous nous appuyons sur la structure multi-solutions, multi-niveaux et concrète produite par un analyseur de pré-traitement. Une fois cette structure désambiguïsée à tous les niveaux, l'analyse continue et donne, automatiquement, une structure abstraite multiniveau. Le transfert et la génération sont classiques (et entièrement automatiques). En définissant trois types d'ambiguïtés remarquables et en posant trois principes qui doivent guider le processus de génération des questions, nous montrons ici une méthode qui semble adéquate.

1 Introduction

C'est en 1975 que commencent les recherches autour des systèmes de TAO (Traduction Automatisée par Ordinateur) interactive avec le projet ITS au Translation Sciences Institute à l'université Brigham Young (Provo, Utah) [1, 2, 3]. Les recherches se poursuivent de 1985 à 1987 au Centre for Computational Linguistics de l'université de Manchester autour du projet N-tran [4, 5, 6] qui visait à permettre à des

anglophones de traduire leurs documents vers le japonais sans connaître cette langue. Le dernier projet de ce type fut le projet KBMT-89 [7, 8] au Centre for Machine Translation à l'université Carnegie Mellon. Le projet consistait à montrer la faisabilité d'un système de traduction doté d'une ontologie du domaine à traiter et interagissant avec le rédacteur pour les cas où la connaissance du domaine ne suffit pas. Des travaux analogues sont actuellement en cours sur de tels systèmes de TAO, en Australie au Department for Computer Science [9], à l'université de Genève [10, 11, 12] et au Japon au laboratoire d'IBM à Tokyo [13, 14].

Suivant la trace de ces projets et dépassant leurs limitations, la TAO pour le rédacteur monolingue, ou "TAO personnelle", est maintenant envisageable tant par l'expérience acquise en "TAO lourde" (pour le veilleur ou pour le réviseur) que par l'évolution récente de la bureautique vers des outils très interactifs et multimédia (hypertextes) disponibles sur des postes de travail bon marché, connectables à des serveurs puissants. Au lieu de réviser ("postéditer") les traductions brutes produites en langue(s) cible(s), l'idée est de prééditer indirectement le texte source grâce à un dialogue entre le système et l'auteur. Ce dialogue vise, entre autres, à clarifier l'entrée, c'est-à-dire à supprimer toutes les ambiguïtés que les connaissances grammaticales "fortes" et les connaissances sémantiques "faibles" du système ne permettent pas de résoudre automatiquement. La structure profonde ainsi obtenue, non ambiguë sur les plans morphologique, syntaxique, sémantique et pragmatique, doit permettre de produire des traductions de grande qualité.

Pour que le système soit vraiment personnel et utilisable par un large public, il faut qu'il s'insère naturellement dans le processus de rédaction et qu'il ne requière aucune compétence spéciale dans les domaines de la linguistique, de l'informatique et de la traduction. Cela fixe donc des contraintes fortes pour la production du dialogue. Si l'on ne veut pas faire appel à des connaissances linguistiques particulières, on est tout de même obligé de demander à l'utilisateur de savoir, par exemple, ce qu'est un verbe, un substantif, un pronom, un article..., ce qui correspond à un niveau d'expertise ne dépassant pas le baccalauréat. Nous sommes aussi amenés à proposer des questions sous forme de paraphrasage des segments ambigus, de façon que les questions que l'on produit ne soient pas ambiguës. Le projet LIDIA (Large Internationalisation des Documents par Interaction avec leur Auteur) [15, 16, 17] cherche à valider ce concept en réalisant un prototype intégrant plusieurs innovations liées à cette nouvelle architecture et au support multimédia.

2 Structure MMC support de la production du dialogue

L'analyseur utilisé dans la partie des traitements qui précède la phase interactive est un analyseur Multi-solutions et la structure produite pour chaque solution est Multi-niveaux et Concrète. D'où l'abréviation MMC.

Le résultat est dit "multi-solutions" 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 faut produire toutes les analyses vérifiant le modèle pour que le rédacteur aide le système à choisir la bonne.

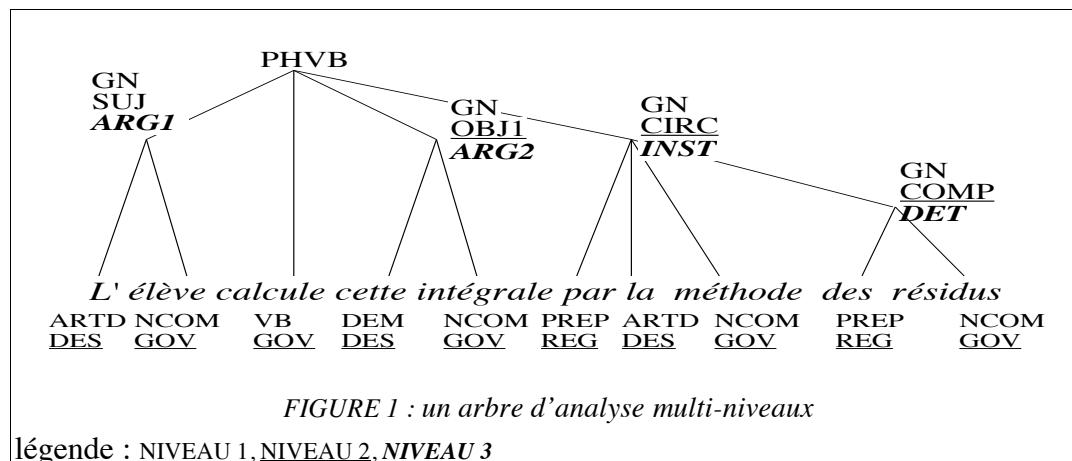
La structure produite est dite “multi-niveaux” car les nœuds portent, entre autres, des décorations complexes qui représentent trois niveaux d’interprétation : le niveau des classes syntaxiques et des classes syntagmatiques, le niveau des fonctions syntaxiques, et enfin le niveau des relations logiques et sémantiques.

- Le niveau des classes syntaxiques (classes terminales telles que : “NOM COMMUN”, “ADJECTIF”, “PREPOSITION”, etc.) et des classes syntagmatiques (classes non terminales telles que : “GROUPE NOMINAL”, “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 (par exemple : “GOUVERNEUR”, “REGISSEUR”, “SUJET”, etc.).

- Le niveau des relations logiques et sémantiques est celui où l’on décrit la construction logique de la phrase, en mettant en évidence les places des arguments attachés aux unités lexicales prédicatives et où l’on donne l’interprétation sémantique des circonstanciels (par exemple : “DETERMINATION”, “CIRCONSTANCIEL INSTRUMENTAL”, “PREMIER ARGUMENT D’UNE UNITE LEXICALE PREDICATIVE”, etc.).

Voici l’arbre d’analyse de la phrase “l’élève calcule cette intégrale par la méthode des résidus” tel qu’il est classiquement calculé au GETA.



La structure produite est dite “concrète”, car on retrouve directement le texte analysé (*en italiques dans la figure 1*) en lisant les feuilles de l’arbre de gauche à droite.

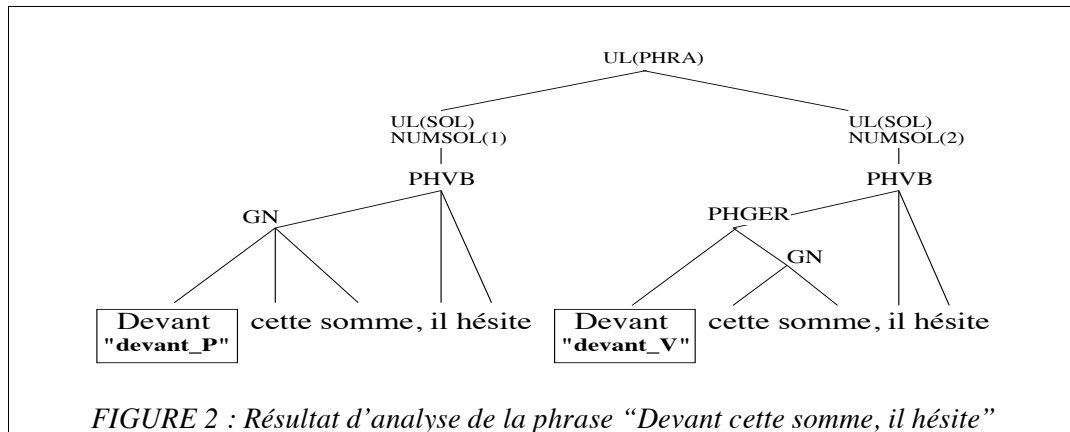
Pour produire le dialogue à partir de la structure MMC, nous sommes d’abord amenés à définir les types d’ambiguïté qui apparaissent dans la structure.

3 Ambiguïtés propres à la structure MMC

Nous considérons trois types d’ambiguïtés, qui permettent de définir des modèles généraux de traitements : ambiguïté de classe, ambiguïté de géométrie et ambiguïté de décoration syntaxique, logique et sémantique.

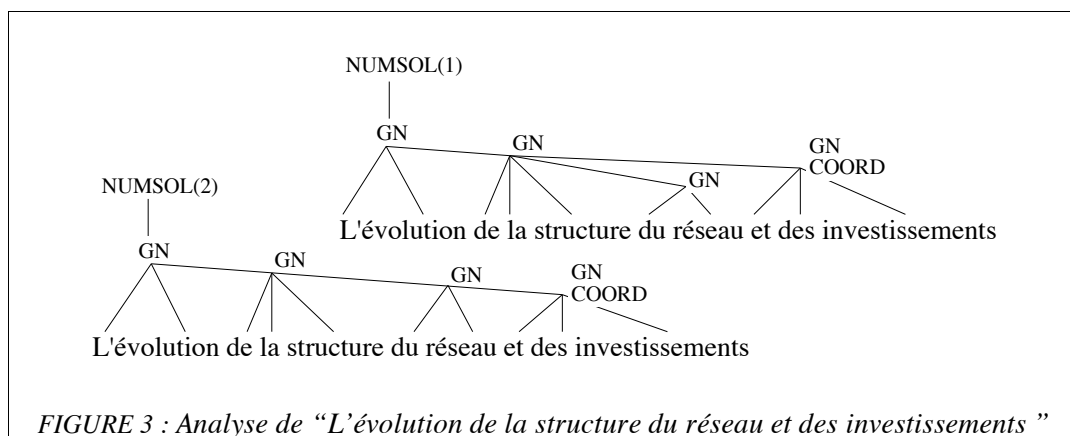
Il y a ambiguïté de classe si une même occurrence est étiquetée par deux classes morpho-syntaxiques différentes dans deux arbres d’analyse distincts.

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.



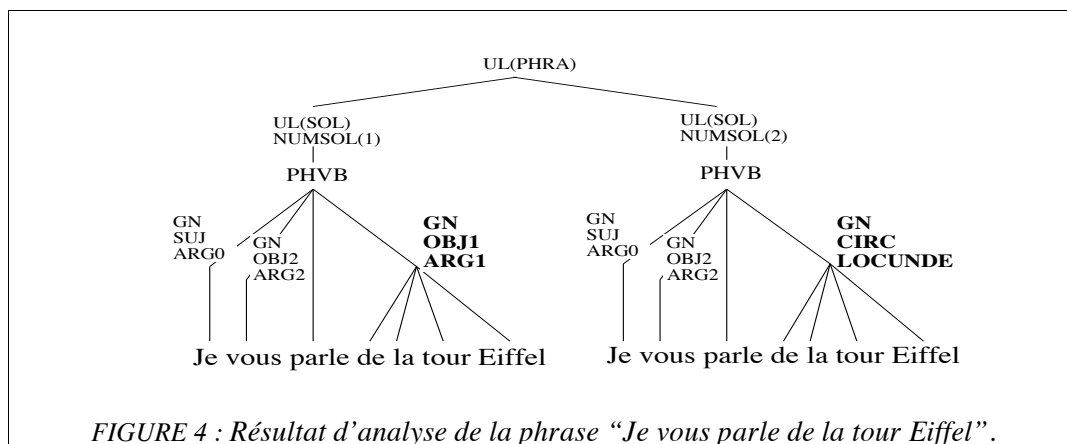
Il y a ambiguïté de géométrie si deux arbres d'analyse différents ont des graphes distincts.

Prenons l'exemple : "L'évolution de la structure du réseau et des investissements". Le schéma suivant (figure 3) montre les différences entre les deux arbres produits, correspondant aux deux lectures de ce groupe nominal : 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)).



Il y a ambiguïté de décoration syntaxique, logique et sémantique si pour des arbres ayant tous la même géométrie, l'étiquetage des nœuds non terminaux en fonction syntaxique et/ou en relation logique et sémantique est différent.

Le schéma suivant (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 bien qu'on nous parle au sujet de la Tour Eiffel (NUMSOL(1)).



4 Méthodologie pour la production du dialogue

4.1 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 objets et les circonstants de tous les prédicats,
- 3) trouver enfin la structure du sujet, des objets et des circonstants des prédicats.

Ces trois principes proviennent de considérations pragmatiques. Nous voulons avant toute autre chose trouver le bon découpage de la phrase en groupes simples. 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 leur sont rattachés.

4.2 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.

Les groupes simples étant pour nous les briques de la phrase, il est nécessaire de les obtenir le plus vite possible. Ainsi, si les groupes simples de toutes les analyses ne sont pas les mêmes, le système résoudra itérativement toutes les ambiguïtés de classes indépendantes. Par contre, avec cette méthode, nous ne traiterons pas toutes les ambiguïtés de classe. En effet, on rencontre des phrases ne présentant pas de problème de découpage en groupes simples et qui présentent une ambiguïté de classe. Cela signifie que la phrase comporte une ambiguïté de géométrie et plus particulièrement de coordination. Pour des raisons ergonomiques, nous avons préféré traiter cette ambiguïté pour elle-même.

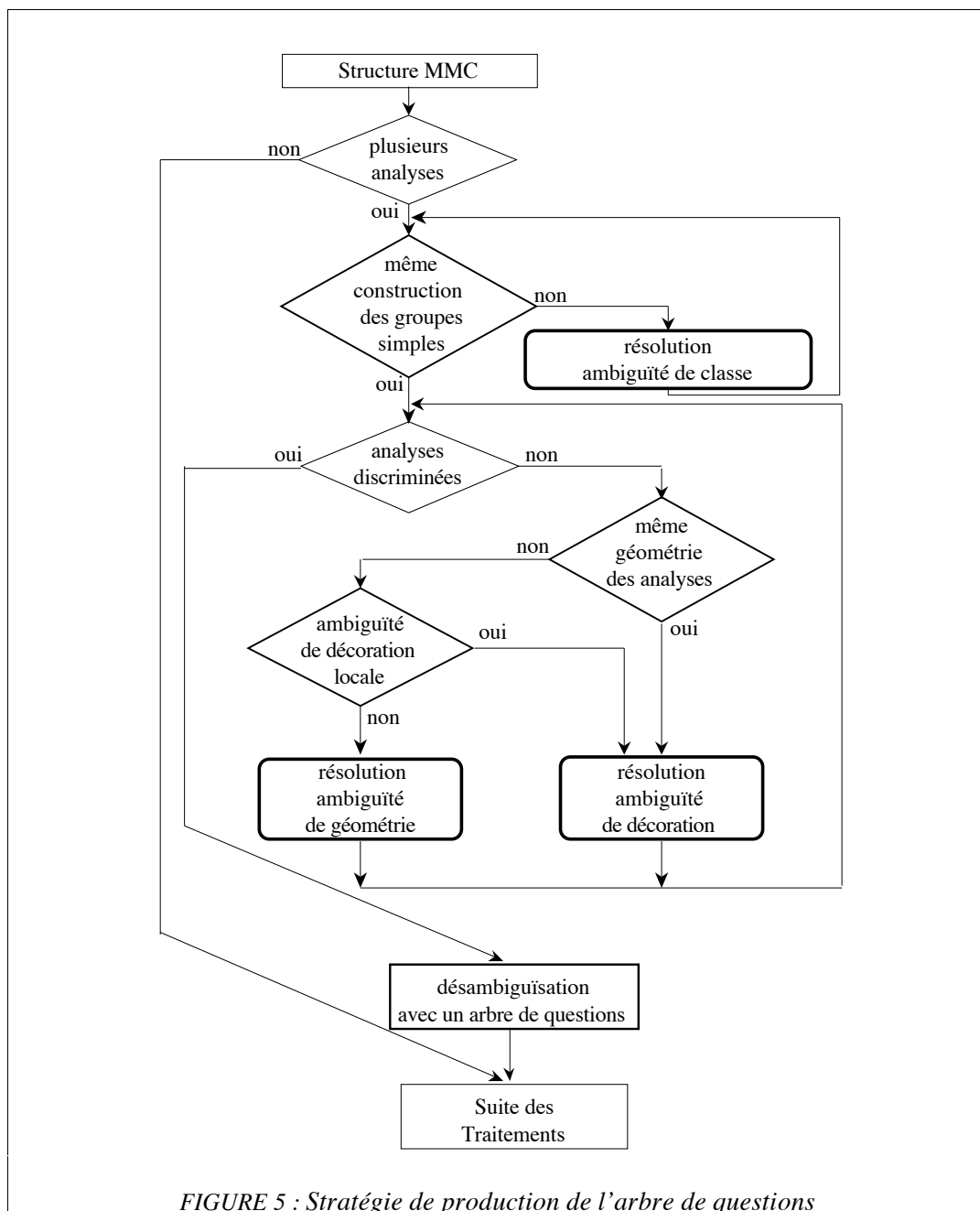


FIGURE 5 : Stratégie de production de l'arbre de questions

Lorsque l'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.

Pour répondre au second principe, nous avons inclus les contraintes nécessaires dans le module de résolution de l'ambiguïté de géométrie (voir plus bas).

4.3 Traitement

Nous allons maintenant examiner les trois types d'ambiguïtés 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.

- Ambiguïté de classe

Pour résoudre une ambiguïté de classe, on projetera la partie de la phrase pour laquelle les classes diffèrent selon deux composantes, l'occurrence et la classe qui lui est associée. On pourra si nécessaire proposer un dialogue plus détaillé [15].

- 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 que nous avons écarté tout à l'heure. Pour résoudre un tel problème, nous faisons un paraphrasage 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 : des groupes subordonnés au verbe le sont avec des relations différentes suivant les analyses. Ici on projette la partie de la phrase qui permet de distinguer les différentes possibilités en remplaçant les prépositions ambiguës qui introduisaient ces groupes par des prépositions qui ne le sont plus.

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 non verbale : un groupe peut établir différents liens de subordination avec les groupes qui le précèdent sauf un verbe. 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 les groupes subordonnés qui ont été séparés par un groupe intermédiaire. On se sert des parenthèses pour bien montrer les mots que l'on a regroupés.

V. subordination verbale : même chose que précédemment, sauf que le groupe qui pose problème peut se raccorder à un verbe. On procède comme pour le cas précédent, mais le verbe est mis à l'infinitif.

- Ambiguïté de décoration

Résoudre une ambiguïté de décoration ou une ambiguïté de décoration locale nécessite les mêmes traitements. Dans un 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. relation logique et sémantique : seules les relations logiques et sémantiques diffèrent entre les analyses. Il suffit de remplacer la préposition qui introduit le groupe qui pose problème par les prépositions non ambiguës qui représentent chacune des relations possibles.

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

III. fonction syntaxique : il y a ambiguïté sur le rôle d'un groupe subordonné au verbe. Il suffit de projeter la phrase en remplaçant la préposition ambiguë qui introduit le groupe qui pose problème par des prépositions non ambiguës.

5 Exemples

Arbre d'analyse et dialogue produits pour la phrase : Il observe la photo et la classe.

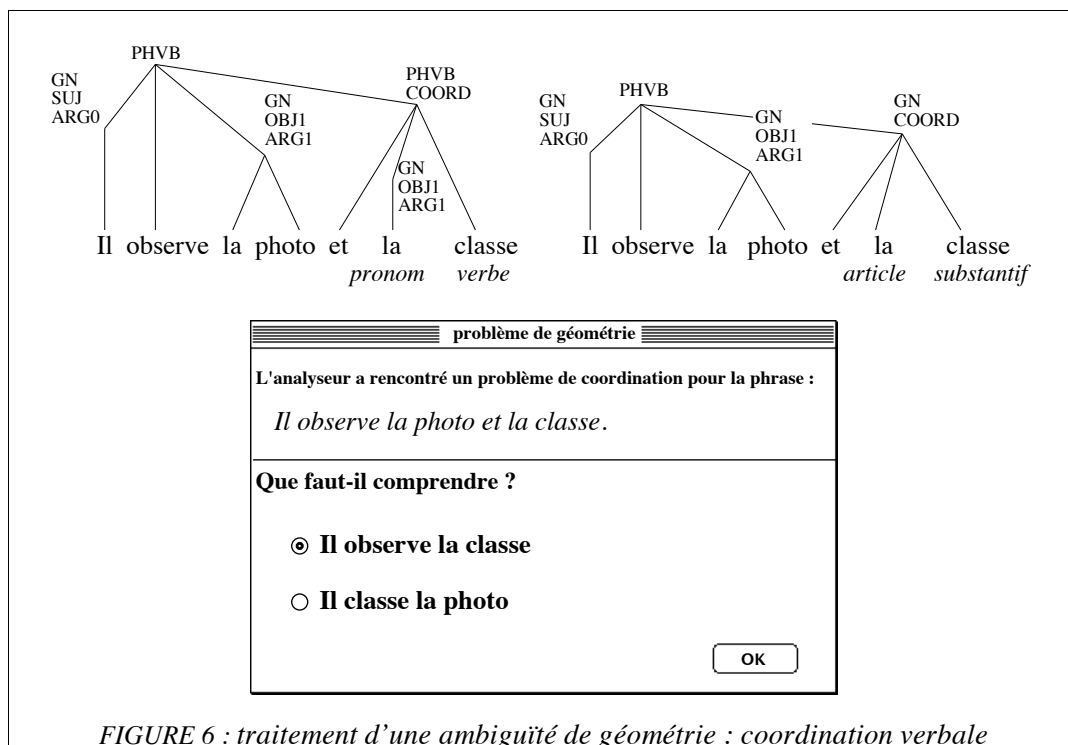
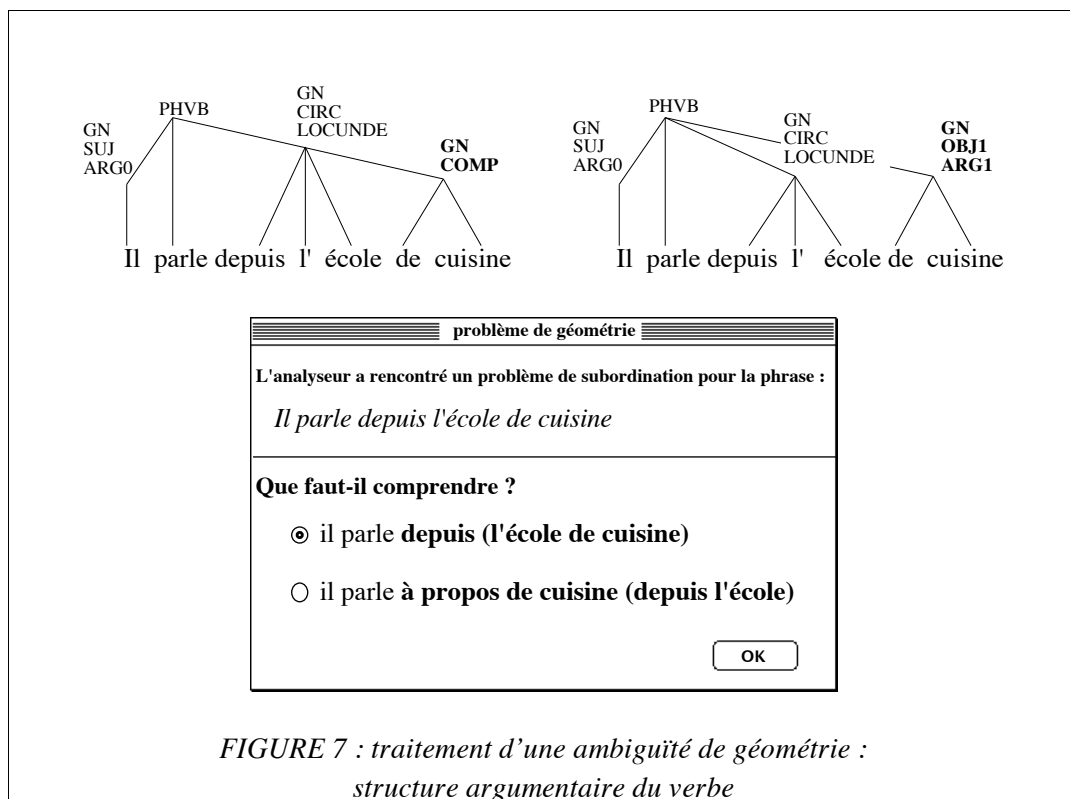
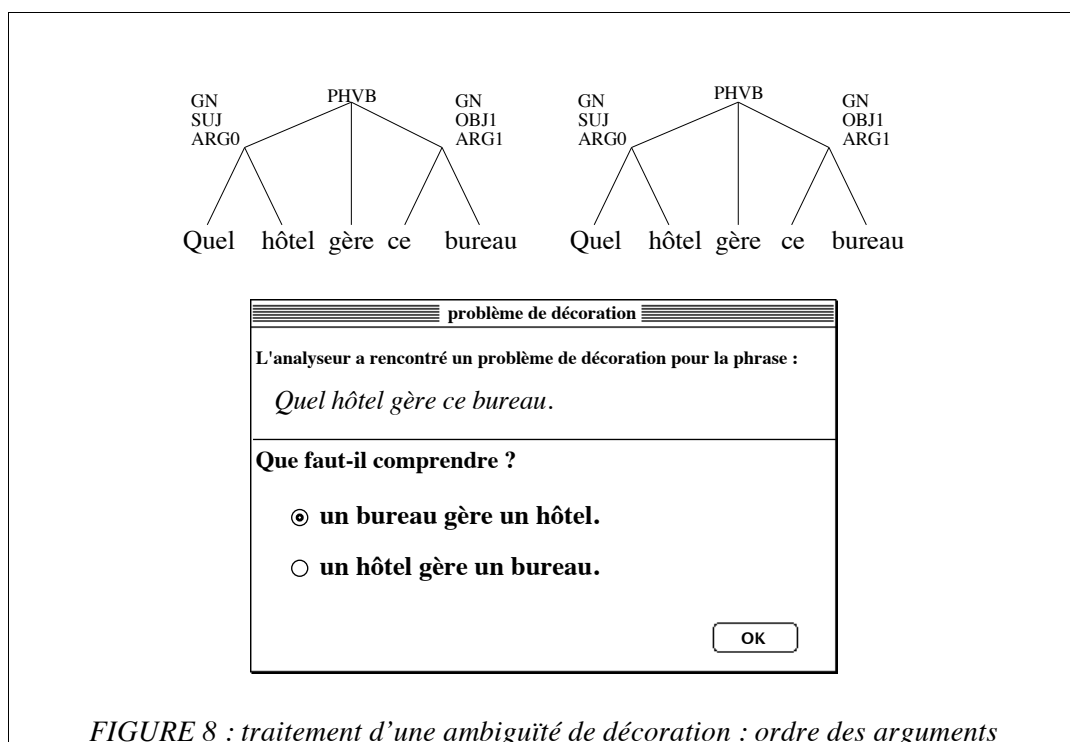


FIGURE 6 : traitement d'une ambiguïté de géométrie : coordination verbale

Arbre d'analyse et dialogue produit pour la phrase : Il parle depuis l'école de cuisine.



Arbre d'analyse et dialogue produit pour la phrase : Quel hôtel gère ce bureau.



6 Discussion

Nous avons décrit ici une partie du processus de désambiguïsation. En effet, nous n'avons pas désambiguïté totalement un énoncé en ayant obtenu le bon arbre d'analyse. C'est ainsi qu'il faut aussi résoudre les problèmes de polysémies (pures et induites par la (ou les) langue(s) cible(s)), les problèmes de référence pronominale ou adjectivale et les problèmes d'ellipse, soit pure (toutes les fois qu'un terme de phrase ou de groupe n'est pas repris, souvent dans la coordination ou la comparaison), soit dans les termes composés (lorsque l'on ne reprend pas complètement un terme composé pour abrégé ou alléger le texte, le terme est remplacé par son gouverneur).

Par rapport aux problèmes pour lesquels nous n'avons pas donné ici une méthode de résolution, notre approche, du point de vue de sa réalisation, sera probablement moins originale que la méthode que nous avons décrite ici. Avec ITS, les questions posées pour résoudre une ambiguïté utilisent des termes relatifs au modèle linguistique sous-jacent. Avec N-tran, le problème se pose aussi. Par contre, les auteurs de KBMT-89 ont veillé à avoir un bon niveau de lisibilité des questions. Le système présente plusieurs questions en même temps, et la réponse à une question modifie, si nécessaire, le libellé des autres questions. Cela permet au rédacteur de fournir d'abord les réponses aux questions qu'il comprend le mieux, les questions plus difficiles peuvent alors être amenées à disparaître.

La solution que nous avons imaginée permet de rédiger toutes les questions en langue naturelle. Elle est fondée sur la distinction entre trois grandes classes de problèmes pour lesquels nous sommes en mesure de proposer un modèle de résolution peu complexe. Nous évitons aussi l'utilisation d'un vrai processus de génération. Comme nous avons raffiné avec précision les modèles de résolutions, nous sommes assurés de ne pas produire une question ambiguë. L'utilisation d'un processus de génération ne nous mettait pas, au contraire, à l'abri d'un tel phénomène. Comme les questions utilisent toujours des remises en forme de ce qui a été rédigé par le rédacteur, nous sommes assurés de bien nous faire comprendre (cf exemples ci-dessus). Le simple paraphrasage permet de ne pas faire intervenir de connaissances linguistiques.

On pourrait nous reprocher de calculer toutes les analyses licites, ce qui est coûteux en temps de calcul (donc en temps de réponse) et en espace mémoire. Nous ne considérons pas que cela soit rédhibitoire. En effet, il ne s'agit pas de traduction en temps réel. Pendant que le serveur de traduction travaille de façon transparente sur un segment de document transmis depuis le système hypertexte, le rédacteur peut continuer son travail. Ce n'est que lorsqu'une analyse s'est révélée ambiguë que le système signale au rédacteur qu'il a besoin de son aide. Le rédacteur peut alors décider ou non d'interrompre sa tâche pour répondre à une ou plusieurs questions. D'une session de travail à l'autre, les questions en suspens seront mémorisées.

7 Conclusion

Nous avons décrit ici une solution que nous croyons simple et élégante pour résoudre certains des problèmes auxquels nous sommes confrontés si nous voulons, à terme,

proposer des systèmes de TAO personnelle à des rédacteurs monolingues. Par l'intermédiaire d'une maquette, nous voulons essayer de montrer que de tels systèmes sont envisageables.

Il ne faut cependant pas confondre une étape d'étude et de maquettage avec la réalisation d'un produit. D'abord, il n'est pas exclu que la recherche bute sur des obstacles non prévus ou sous-estimés. De plus, pour la TAO personnelle grand public, les bases lexicales devront être de très grande taille, puisqu'on ne peut restreindre le domaine des documents à traiter. Il faudra donc, aussi, que la recherche dans ce domaine progresse.

Références

- [1] **Melby, A. K.**, (1982). *Multi-Level Translation Aids in a Distributed System*. Proc. Coling-82. Prague. 5-10 juillet 1982, vol. **1/1** : pp. 215-220.
- [2] **Melby, A. K.**, (1981). *Translators and Machines - Can they cooperate ?* in META. vol. **26**(1) : pp. 23-34.
- [3] **Melby, A. K., et al.**, (1980). *ITS : An Interactive Translation System*. Proc. Coling-80. Tokyo. 30 septembre-4 octobre 1980, vol. **1/1** : pp. 424-429.
- [4] **Tomita, M.**, (1985). *Feasibility Study of Personal/Interactive Machine Translation System*. Proc. Conference on Theoretical and methodological Issues in Machine Translation of Natural Language. Colgate University, Hamilton, New York. 14-16 août 1985, vol. **1/1** : pp. 289-297.
- [5] **Chandler, B., et al.**, (1987). *N-Tran Final Report. Alvey Project*. Rap. CCL/UMIST, Manchester. n° 87/9. 23 p.
- [6] **Wood, M. M.**, (1989). *Japanese for speakers of English: The UMIST/Sheffield Machine Translation Project*. in Recent Developments and Applications of Natural Language Processing. Kogan Page Limited. London. pp. 56-64.
- [7] **Brown, R. D.**, (1989). *Augmentation*. in Machine Translation. vol. **4** : pp. 129-147.
- [8] **Goodman, K. & Nirenburg, S.** (ed.), (1991). *The KBMT Project: A case study in knowledge-based machine translation*. Morgan Kaufmann. San Mateo, California. 331 p.
- [9] **Huang, X.**, (1990). *A Machine Translation System for the Target Language Inexpert*. Proc. Coling-90. Helsinki. 20-25 août 1990, vol. **3/3** : pp. 364-367.
- [10] **Wehrli, É.**, (1993). *Vers un système de traduction interactif*. in La traductique. Les presses de l'Université de Montréal, AUPELF/UREF. pp. 423-432.
- [11] **Wehrli, É.**, (1992). *The IPS System*. Proc. Coling-92. Nantes, France. 23-28 juillet 1992, vol. **3/4** : pp. 870-874.
- [12] **Wehrli, É.**, (1990). *STS: An Experimental Sentence Translation System*. Proc. Coling-90. Helsinki. 20-25 août 1990, vol. **1/3** : pp. 76-78.
- [13] **Maruyama, H., et al.**, (1990). *An Interactive Japanese Parser for Machine Translation*. Proc. Coling-90. Helsinki. 20-25 août 1990, vol. **2/3** : pp. 257-262.
- [14] **Tsutsumi, T., et al.**, (1993). *Example-Based Approach to Machine Translation*. Proc. Premières journées franco-japonaise sur la traduction assistée par ordinateur.

Ambassade de France au Japon, Tokyo, Japon. 15-16 mars 1993, vol. **1/1** : pp. 161-169.

- [15] **Boitet, C.**, (1990). *Towards Personal MT : general design, dialogue structure, potential role of speech*. Proc. Coling-90. Helsinki. 20-25 Août 1990, vol. **3/3** : pp. 30-35.
- [16] **Blanchon, H.**, (1992). *A Solution to the Problem of Interactive Disambiguation*. Proc. Coling-92. Nantes, France. 23-28 juillet 1992., vol. **4/4** : pp. 1233-1238.
- [17] **Boitet, C. & Blanchon, H.**, (1993). *Dialogue-based MT for monolingual authors and the LIDIA project*. Proc. NLPRS'93. Fukuoka, Japon. 6-7 décembre 1993, vol. **1/1** : pp. 208-222.