HAL
open science

Restrictions sémantiques apportées à l'étude des groupes nominaux en 'NdeN' : application à la machine à dicter

Joëlle Klein, Kamel Smaïli, Laurent Romary, François Charpillet

To cite this version:

Joëlle Klein, Kamel Smaïli, Laurent Romary, François Charpillet. Restrictions sémantiques apportées à l'étude des groupes nominaux en 'NdeN' : application à la machine à dicter. 19èmes Journées d'Etudes sur la Parole, 1992, Bruxelles, Belgique. hal-00403955

HAL Id: hal-00403955

https://hal.science/hal-00403955

Submitted on 15 Jul 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESTRICTIONS SÉMANTIQUES APPORTÉES À L'ÉTUDE DES GROUPES NOMINAUX EN 'NDEN' : APPLICATION À LA MACHINE À DICTER

J. Klein, K. Smaïli, L. Romary, F. Charpillet

Crin Inria Lorraine
BP 239 54506 Vandœuvre-lès-Nancy

Dans cet article, nous présentons comment l'intégration d'un analyseur sémantique permet d'améliorer les résultats de la machine à dicter. L'analyseur sémantique ne traite actuellement que des expressions de la forme *Nom* de *Nom.* La machine à dicter quant à elle fonctionne à l'aide de modèles biclasses et triclasses essentiellement syntaxiques et, vu les incertitudes de la reconnaissance de la parole, fournit en résultat un ensemble de phrases syntaxiquement correctes mais qui peuvent être totalement dépourvues de sens. l'analyseur sémantique permet de sélectionner parmi les expressions en NdeN celles qui sont porteuses de sens et ainsi de diminuer le nombre de solutions possibles, rendant l'outil plus performant et donc plus ergonomique.

INTRODUCTION

Les systèmes de traitement automatique du langage naturel peuvent être classés en deux grandes catégories : les systèmes syntaxiques tels que la machine à dicter, où la compréhension n'est pas nécessaire, et les systèmes où la compréhension, donc l'analyse sémantique, est indispensable pour accomplir une tâche par exemple comme dans les systèmes de commande, de dialogue ou d'interrogation de bases de données. L'interprétation du sens d'expressions en langage naturel dans de tels systèmes est, en général, restreinte par le domaine d'application.

Il existe actuellement peu de travaux portant sur la sémantique hors contexte et hors domaine d'application. Notre étude a pour but d'analyser les relations de sens existant entre les mots de groupes nominaux complexes de la forme 'Nom de Nom'[1] de manière à obtenir une couverture maximale de la langue dans un contexte syntaxique restreint. Les différents sens de telles expressions peuvent être mis en évidence par leur traduction dans une autre langue. Par exemple, 'les livres de Sartre' se traduira en Allemand par 'Sartres Bücher' si la relation sémantique est une relation de possession et par 'Die Bücher von Sartre' s'il s'agit d'une relation de production (auteur).

L'application d'un tel analyseur sémantique à la machine à dicter qui fonctionne à base de modèles biclasses ou triclasses essentiellement syntaxiques, permet de sélectionner parmi les expressions reconnues, celles qui sont porteuses de sens et celles qui ne le sont pas.

Dans la première partie, nous allons présenter l'analyseur sémantique et plus particulièrement, ses fondements théoriques, son application spécifique aux expressions à valeur prédicative et son implémentation. Dans le deuxième paragraphe, nous présenterons les fonctionnalités de la machine à dicter et enfin, dans le troisième, les résultats obtenus par l'intégration de l'analyseur sémantique à la machine à dicter.

1. L'ANALYSEUR SÉMANTIQUE

1.1. Fondement théorique

La première étape de notre travail a consisté en l'analyse d'une typologie des expressions de la forme NdeN [Lejosne-91], qui a montré la nécessité de décrire

1 Nous utiliserons indifféremment les notations NdeN, Nom de Nom ou N1 de N2 pour décrire de telles expressions.

sémantiquement les mots d'une manière très fine. C'est pourquoi nous avons fondé notre étude sur la théorie componentionnelle de F. Rastier [Rastier-87, Rastier-89]. Cette théorie s'appuie sur les notions de sémème et de sème. Un sémème est le contenu sémantique d'un morphème et un sème est caractérisé comme étant l'extrémité d'une relation fonctionnelle binaire entre sémèmes. Ou encore d'après Tutescu [Tutescu-74] cité dans [Rastier-87] : " l'unité minimale de sens, le trait pertinent du contenu sémantique, l'invariant de sens s'appelle marque sémantique, marqueur sémique ou sème…".

On peut distinguer deux classes de sèmes : les sèmes inhérents (relevant du système fonctionnel de la langue) et les sèmes afférents (relevant de normes socialisées voire idiolectales). Les sèmes sont différenciés par leur niveau de généricité, on distingue : les sèmes macrogénériques (ex : /humain/, /animé/), les sèmes mésogénériques (ex : /alimentation/), les sèmes microgénériques (ex : /partie-du-corps/) et enfin les sèmes spécifiques (ex : /fonctionnel/). La description sémantique d'un mot est obtenue par la définition de trois groupes de sèmes :

- le **taxème** contient les sèmes spécifiques et microgénériques. Les sèmes microgénériques servent à regrouper au sein d'un même taxème des éléments voisins alors que les sèmes spécifiques servent à les différencier ;
- le **domaine** est un groupe de taxèmes (caractérisé par un sème mésogénérique) tel que dans un domaine il n'existe pas de polysémie ;
- la **dimension** est une classe de généralité supérieure. Elle inclut des sémèmes comportant un même trait générique (sème macrogénérique). Les dimensions peuvent être articulées entre elles par des relations de disjonction exclusive (ex : /animé/ vs /inanimé/).

EXEMPLE :

'cuiller' est caractérisé par l' ensemble de sèmes génériques suivants (d'après [Rastier-87]):

- taxème : /couvert/
- domaine : /alimentation/
- dimension : /concret/ /inanimé/

Une telle représentation permet d'obtenir une description fine du sens et de définir un ensemble de traits spécifiques à l'étude envisagée.

1.2. Application aux prédicatifs

Dans un premier temps, nous avons focalisé notre étude sur le traitement des groupements fonctionnels N1deN2 tels que le N1 se comporte de manière prédicative. Les exemples suivants illustrent les trois cas envisagés :

EXEMPLE :

- l'achat de Pierre
  - -> action : *Pierre achète quelque chose*
  - -> résultat : *ce que Pierre achète*
- le conducteur du camion
  - -> agent : *celui qui conduit le camion*

Dans la typologie, il existe trois grandes classes mettant en relation un N1 prédicatif et un N2 argumentatif. Ces trois classes correspondent au cas où N2 est agent (ex : l'achat de Pierre), au cas où N1 est un prédicat et N2 est objet (ex : l'achat de la voiture) et enfin le cas où N1 est un agent et N2 un objet (ex : le conducteur du camion).

Il nous faut déterminer la classe d'appartenance de chaque groupement dont le N1 est de type prédicatif. Pour cela il est nécessaire de connaître la structure sémantique associée au prédicat concerné (i.e. ses arguments) ainsi que les contraintes qui leur sont rattachées. A cette fin, nous avons défini un ensemble de cadres sémantiques, chaque cadre décrivant une classe de prédicats ayant le même comportement sémantique. Comme les seuls arguments que peuvent instancier le N1 et le N2 sont : le prédicat, le nominatif et l'accusatif, les descriptions des cadres sémantiques seront restreintes à ces trois cas (au sens d'une grammaire de cas).

EXEMPLE :

Cadre_achat est un cadre décrivant les arguments des prédicats ayant le même comportement que **acheter**.

Ce cadre possède deux arguments :

- un nominatif dont la dimension doit prendre la valeur : ***humain***
- un accusatif dont la dimension doit prendre la valeur : ***concret et non humain***

Dans cet exemple, dans un souci de simplification, nous ne nous plaçons pas dans le contexte historique où l'on achetait des personne (ex : l'achat de l'esclave) qui relève de la culture mais pas du modèle fonctionnel de la langue[2]. De même, nous n'envisageons pas les cas métaphoriques, tels que ***l'achat du maire par la Mafia***, ou ***l'achat de son silence***.

Les prédicatifs *don, vente, apport…* appartiennent à la classe décrite par le cadre *Cadre_achat*..

En plus du cadre sémantique associé au prédicat, il est indispensable de posséder des informations concernant le comportement de celui-ci dans le contexte d'utilisation que nous étudions. En effet, des prédicats possédant les mêmes arguments nominatif et accusatif, pourront accepter en position de N2, pour l'un, indifféremment le nominatif ou l'accusatif (ex : l'achat de Pierre, l'achat de la voiture), pour un autre n'autoriser que l'accusatif (ex : l'abolition de l'esclavage). Nous avons recensé trois comportements différents :

- **PLEIN :** n'importe quel argument peut être instancié par le N2 ;
- **REDUIT** : l'instanciation est réduite à l'accusatif ;

2 Ce qui implique en substance qu'il est encore difficile d'envisager une étude qui aurait une couverture totale de la lengue française actuelle.

- **PLEIN_REFL** : le prédicat a à la fois le comportement d'un plein et d'un réfléchi, dans le cas du réfléchi, l'accusatif est égal au nominatif.
  ex : - l'abonnement de Pierre par Marie
  - l'abonnement de Pierre.

Nous pouvons considérer, dans notre application, que les traits /plein/, /réduit/ et /plein_refl/ permettent de regrouper au sein d'un même taxème les mots ayant un même comportement. La gestion des prédicats ayant un aspect résultatif au sens où nous l'avons défini, passe par la définition d'un trait **spécifique** /résultatif/, en effet, il n'existe pas de classes correspondant aux résultatifs et aux non-résultatifs, mais dans chaque classe que nous avons définie peut se trouver des mots comportant ce trait et d'autres ne l'ayant pas.

Les taxèmes et les dimensions vont nous permettre de donner des définitions sémantiques des mots dans le lexique (les domaines ne sont pas utilisés pour les prédicatifs). Pour certains lexèmes, en particulier pour les agentifs, il nous a paru nécessaire d'introduire deux noyaux sémantiques dans la définition du mot, un noyau principal (NP) spécifique à l'agent et un noyau secondaire (NS) spécifique au prédicat qui lui est associé. La figure 1 nous donne la définition du mot 'conducteur'. Nous remarquerons que le lien avec le cadre sémantique associé au prédicat se fait par l'intermédiaire de son nom.

CONDUCTEUR

Noyau Principal : /conducteur/
taxème : comptable
dimension : +humain
Noyau Secondaire : /conduire/
taxème : plein
dimension : notion
Cadre Associé : cadre_achat =>
contraintes : sème spécifique (ACC) = /véhicule/

Figure 1. description sémantique de 'conducteur'

1.3. Implémentation

Les sources de connaissance sont :

- les cadres sémantiques des prédicats qui permettent de définir les contraintes sémantiques portant sur chaque argument, nous en donnons un exemple à la figure 2 ;

cadre_achat
Nominatif ->dimension : +humain
Prédicat ->taxème : prédicat
Accusatif ->dimension : concret et non +humain

Figure 2. Cadre sémantique du prédicat 'achat'

- le lexique qui fournit la description sémantique des mots telle que nous l'avons décrite dans la figure 1 ;
- un ensemble de règles qui oriente l'analyse du N1deN2 en fonction des caractéristiques propres aux deux mots et de valider l'interprétation au moyen d'une fonction d'appariement qui teste si les contraintes sémantiques apportées par la règle et par le cadre sont vérifiées. La figure 3. nous montre un exemple de règle utilisée dans la recherche de la relation sémantique existant entre les deux groupes N1 et N2.

Origine :
cadres_associés au N1
Condition d'application
Taxème (noyau secondaire de N1)
a la valeur /PLEIN/
Actions possibles :
instanciation :
- prédicat = Noyau Secondaire de N1
- nominatif = Noyau Principal de N1
- accusatif = Noyau Principal de N2

schéma associé dans la typologie :
schéma subjectif (N1 sujet)
résultat de l'analyse : Noyau Principal de N1

Figure 3. Exemple de règle d'interprétation

La figure 4. schématise l'analyse du groupement "le conducteur du camion" grâce à la règle et au cadre définis ci-dessus ainsi qu'à la description des mots *conducteur* et *camion*.

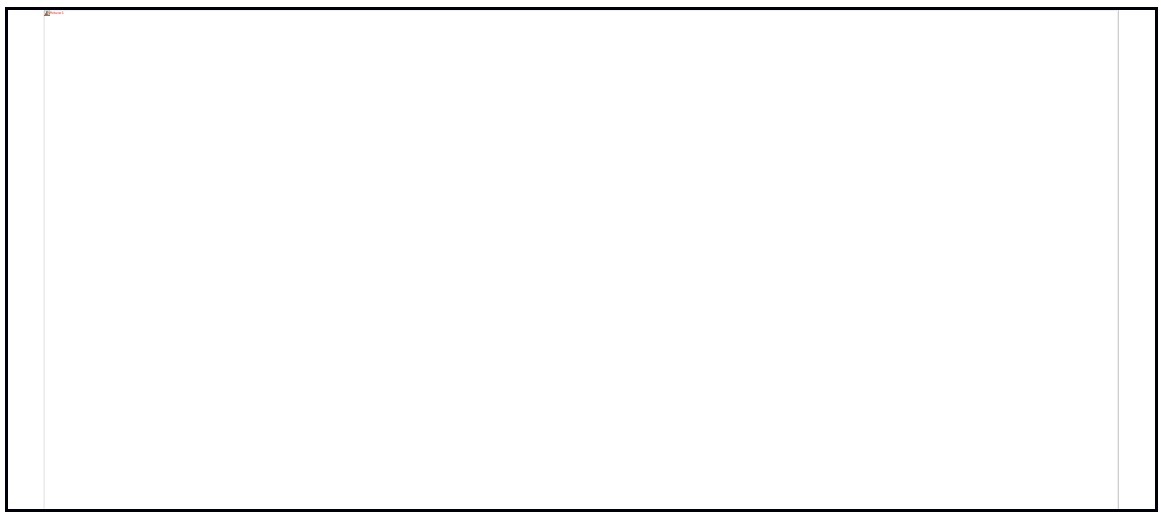

Figure 4. Schéma d'interprétation du groupe "le conducteur du camion"

2. LA MACHINE À DICTER

On confond souvent système de reconnaissance de la parole (SRP) et machine à dicter (MAD). S'il est vrai qu'une machine à dicter est fortement dépendante de son SRP, elle a cependant besoin d'un certain nombre d'outils pour pouvoir fonctionner en tant que telle. Un simple SRP n'est pas en mesure d'apporter des modifications sur le texte produit par la MAD. D'où l'intérêt d'un éditeur. Cet éditeur peut être vocal ou non.

MAUD (Machine AUtomatique à Dicter) est un prototype de machine à dicter acceptant en entrée le langage naturel (aucune restriction syntaxique n'est imposée). Très souvent l'utilisation du langage naturel s'accompagne par l'utilisation d'un grand vocabulaire. En effet, les systèmes de reconnaissance actuels (et pour un bon nombre d'années à venir) ne peuvent reconnaître un mot si celui-ci n'appartient pas au lexique. La réalisation d'une telle machine a nécessité l'utilisation de quatre composantes : la comosante acoustico-phonétique, la composante lexicale, la composante linguistique, et l'éditeur associé à la MAD [smaïli 91b]. On retrouve ces quatre composantes dans le schéma de l'architecture générale de MAUD de la figure 5.

Figure 5 : Architecture générale de MAUD.

Nous détaillons dans ce qui suit les composantes : acoustico-phonétique, lexicale, et syntaxico-sémantique et nous montrerons à quel niveau intervient l'analyseur sémantique 'NdeN' dont il est question dans cet article.

2.1 La composante acoustico-phonetique de maud

Cette composante est fondée sur le décodeur acoustico-phonétique APHODEX [François 90]. La première tâche de MAUD est de fournir deux treillis phonétiques à partir du treillis phonétique d'APHODEX. Le premier est appelé treillis phonétique d'acceptation (TPA). Ce treillis est construit à partir des nœuds de chaque segment du treillis d'origine, en gardant les étiquettes qui ont un coefficient de vraissemblance supérieur à un certain seuil. Le deuxième est dit treillis phonétique de rejet (TPR). Il est composé des étiquettes phonétiques qui ont été reconnues avec un mauvais score. Ce treillis est utilisé pour rejeter des hypothèses de substitution émises par les niveaux supérieurs.

2.2 La composante lexicale

La composante lexicale joue un rôle central dans MAUD puisqu'elle s'articule avec la composante linguistique et la composante acoustico-phonétique avec lesquelles elle interagit pour identifier dans le continuum de parole les mots pouvant être mis en correspondance avec le signal vocal.

Le lexique de MAUD est composé de 37000 entrées lexicales. Il ne faut pas considérer ce lexique comme une simple liste de mots, mais comme une base de données de laquelle on peut extraire un grand nombre d'informations concernant chacune de ces entrées. La tâche d'identification ne peut s'effectuer sans une organisation efficace du lexique. Les fonctions d'accès doivent en faciliter la mise en œuvre de filtres dont la combinaison doit permettre d'extraire du lexique des sous-ensembles aussi restreints que possible. Les informations lexicales du lexique de MAUD sont réparties en deux catégories permettant à la composante lexicale de MAUD d'agir sur deux niveaux : infra-lexical, et supra-lexical. La composante supra-lexicale permet les interactions avec les niveaux syntaxico-sémantiques, alors que la composante infra-lexicale facilite la communication avec le niveau acoustico-phonétique.

La mise en place de procédures d'accès au lexique doit être aussi efficace que possible tant les accès sont nombreux pendant la reconnaissance. Pour ce faire, un certain nombre de filtres très complexes sont mis en œuvre dans MAUD permettant une efficacité, à la fois sur la rapidité et sur la capacité à extraire des sous-vocabulaires restreints [Smaïli 92].

2.3 LA COMPOSANTE SYNTAXICO-SEMANTIQUE

Le rôle de la syntaxe en reconnaissance de la parole est de participer au choix du prochain mot à reconnaître et à l'élimination d'un certain nombre d'hypothèses. La constitution de phrases d'une langue n'est pas une simple combinaison de mots, pris dans n'importe quel ordre, mais un mécanisme de construction de phrases très précis. En traitement de langue naturelle, on ne sait toujours pas fournir un modèle linguistique permettant de traiter automatiquement la langue. C'est pour cette raison, que les informaticiens partent du principe que la probabilité de production d'un mot dépend conditionnellement de toute la première partie de la phrase, pour proposer un modèle permettant de traiter la langue. D'après cette constatation, il est naturel de penser à l'utilisation d'un modèle probabiliste. En effet, la composante syntaxico-sémantique de MAUD est composée d'un modèle markovien comprenant 6000 états et 37000 transitions. Ce modèle est augmenté d'un certain nombre de règles grammaticales et phonologiques permettant de prendre en compte les phénomènes linguistques qui ne peuvent l'être par le modèle probabiliste. MAUD est composée de sept modules syntaxico-sémantiques : le préprocesseur syntaxico-sémantique, le processeur stochastique, le filtre des patrons syntaxiques [Smaïli 91a], le générateur de phrases, le filtre grammatical, et le filtre phonologique.

Ces sept modules agissent de manière pyramidale. Autrement dit, lorsque les solutions proposées à un certain niveau arrivent au niveau supérieur, elles sont filtrées (donc réduites) et envoyées de nouveau au niveau immédiatement supérieur. Ce filtrage multi-niveaux assure une bonne réduction de l'espace de solutions. Cependant, et ce malgré l'existence de ces sept modules syntaxico-sémantiques, à cause de l'imperfection du décodage acoustico-phonétique et du non recouvrement total de la langue du modèle probabiliste, les solutions proposées sont en nombre important. Pour réduire le nombre de propositions, une première solution consiste à introduire un analyseur sémantique permettant de traiter les groupes 'NdeN' prédicatifs. L'apport d'un tel analyseur est très important comme le montre les résultats des prochains paragraphes.

3. INTÉGRATION DE L'ANALYSEUR SÉMANTIQUE À LA MACHINE À DICTER

Nous avons utilisé l'analyseur sémantique des expressions en NdeN comme un filtre agissant sur les résultats obtenus par la machine à dicter. Nous avons analysé quatre phrases contenant les expressions suivantes : 'l'achat de Chantal', 'la chute de l'enfant', 'le chauffeur de taxi' et 'l'achat du livre'. Nous avons extrait des résultats fournis ceux ayant une valeur prédicative et nous les avons soumis à l'analyseur sémantique. Sur 51 expressions proposées, 20 possèdent une valeur prédicative parmi celles-ci 9 ont été validées et correspondent aux expressions qui intuitivement paraissent correctes. L'inconvénient principal de la restriction aux expressions à valeur prédicative, est que l'on ne peut pas déterminer si une expression n'a pas été validée parce qu'elle n'est pas porteuse de sens ou parce qu'elle n'est pas prédicative. L'intérêt dans ce cas de figure est d'utiliser la validation sémantique comme un facteur intervenant sur le score de reconnaissance, ce qui augmente ainsi la convivialité de l'interface avec l'utilisateur qui aura l'avantage de trouver les meilleures expressions en tête de liste.

Ceci n'étant malgré tout pas entièrement satisfaisant, nous avons décidé d'étendre l'analyse à toutes les expressions en NdeN. De la même manière que nous avons associé un cadre à un prédicat, nous avons défini un ensemble de cadres associés aux différentes classes de la typologie (ex : appartenance, propriété …). Certaines relations n'ont cependant pas pu être explicitées car elles ne dépendent pas de la sémantique mais de connaissances encyclopédiques pour lesquelles il faudra envisager un traitement particulier (ex : relation de production : les livres de Sartre, relations de partie-tout : le pied de la table). Certains cadres sont quant à eux associés directement aux mots même si ceux-ci ne sont pas des prédicatifs (ex : spécialiste -> cadre connaitre). Après l'intégration de ces cadres, nous avons pu constater que la typologie était incomplète d'où le rejet de certaines expressions pourtant valides. Sur les 51 expressions de départ, seules 16 d'entre elles ont été validées, et toutes sont porteuses de sens. Parmi les expressions rejetées, il y en a 6 qui font partie des expressions qui peuvent entrer dans la catégorie partie-tout et qui n'ont donc pas été traitées, la seule solution envisagée actuellement est de ne pas les rejeter même si l'on est incapable de déterminer si elles ont un sens ou non. Sur les 51 expressions de départ,

nous nous retrouvons donc avec 22 expressions conservées et 29 rejetées, ce qui fait un gain de plus de 50%. Parmi les expression validées, nous trouvons par exemple : *l'achat du fauteuil, l'achat du sapin* (pour la phrase contenant *l'achat de Chantal)*, ou encore *la chute de l'avocat, la chute de l'assassin* (pour *la chute de l'enfant*), et parmi celles rejetées nous avons: *l'agent de chagrin, l'enfant de sapin, le chauffeur de pêcher…*

CONCLUSION

L'étude qui a été faite montre l'utilité d'une analyse sémantique comme filtre des résultats de la machine à dicter. Les jeux d'essais effectués ne permettent pas d'établir le gain moyen obtenu par cette analyse sur un grand corpus mais les résultats obtenus sont tout de même pertinents et montrent la validité des traitements. L'application à la machine à dicter a permis de mettre en évidence un certain nombre de classes ne figurant pas dans la typologie et qu'il faudra prendre en compte dans la suite des travaux. Une extension de l'étude faite sur les expressions prédicatives consistant en la définition des modules casuels complets associés aux prédicats (et non plus limités au nominatif et à l'accusatif) permettra d'élargir l'analyse sémantique à des expressions englobant des NdeN et de résoudre ainsi un certain nombre d'ambiguïtés d'interprétation possibles (la conduite de Paul est sportive, la conduite de Paul à la gare m'a pris deux heures).

Dans la version actuelle, l'analyseur sémantique sert uniquement à valider ou invalider des expressions reconnues par la machine à dicter. il serait intéressant maintenant de l'intégrer à celle-ci de manière à pouvoir faire des prédictions lors de la reconnaissance de phrases et ainsi, d'une part, ne construire que des groupes corrects, et d'autre part, de gagner du temps par la réduction par contraintes sémantiques du lexique à consulter.

BIBLIOGRAPHIE

[François-90] D.François, D.Fohr " Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de la parole continue", XVIII Journées d'Etudes sur la parole, 1990.

[Lejosne-91] Lejosne J.-C., Klein J., Lauvray J., Romary L., "Typologie des groupes nominaux complexes", Colloque *Lexique et Inférence*, Metz, 1991.

[Rastier-87] Rastier F. *Sémantique interprétative*, puf , Formes sémiotiques(1987).

[Rastier-89] Rastier F. *Sens et Textualité*, Hachette, Paris (1989)

[Smaïli-91a] K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton " A continuous speech recognition approach for the design of a dictation machine", European Conference on SSpeech Technology, pp953-956, Genova 1991.

[Smaïli-91b] K.Smaïli " Conception et réalisation d'une machine à dicter à entrée vocale destinée aux grands vocabulaires: Le système MAUD", Thèse de Doctorat de l'université de Nancy I, 1991.

[Smaïli-92] K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton " La composante lexicale de la machine à dicter MAUD", Séminaire lexique. Communication Homme-Machine Pôle langage naturel, pp46-57, 1992