

## **Résolution des anaphores pronominales : quelques postulats du TALN mis à l'épreuve du dialogue oral finalisé**

Antoine Jean-Yves

VALORIA (EA 2593), Université de Bretagne Sud  
BP 573, F-56017 Vannes, France – Mel : Jean-Yves.Antoine@univ-ubs.fr

### **Résumé – Abstract**

Cet article étudie l'adaptation au dialogue oral homme-machine des techniques de résolution des anaphores pronominales qui ont été développées par le TALN pour les documents écrits. A partir d'une étude de corpus de dialogue oral, il étudie la faisabilité de ce portage de l'écrit vers l'oral. Les résultats de cette étude montrent que certains indices utilisés à l'écrit (accord en nombre, distance entre le pronom et son antécédent) sont plus friables en dialogue oral finalisé. Les techniques développées pour l'écrit ne peuvent donc pas être réutilisées directement à l'oral.

In this paper, we present a corpus analysis on pronominal anaphora that investigate the adaptation of anaphora resolution techniques which have been developed for written language processing and that should apply to spoken man-machine dialogue. Unfortunately, this corpus study shows that the criteria that are used on written texts (gender and number agreement, distance between the pronoun and its antecedent) seems to lack robustness on interactive spoken language.

### **Keywords – Mots Clés**

Référence, anaphore pronominale, dialogue oral homme-machine, analyse des usages sur corpus.  
Reference, pronominal anaphora, spoken man-machine dialogue, corpus analysis.

## **1 Résolution automatique des anaphores : du DHM oral au TALN**

Le calcul de la référence est une problème de première importance pour le dialogue oral homme-machine (Pierrel et Romary, 2000). Dans le cas d'un serveur vocal d'information, il est par exemple essentiel d'accéder à l'ensemble des objets de l'application auxquels réfère l'utilisateur dans sa requête avant de pouvoir interroger la base de données.

L'expression de la référence repose sur des procédés complexes et variés qui mettent en jeu de multiples sources de connaissances et impliquent souvent des inférences extra-linguistiques. Dans le cadre de cet article, nous nous intéresserons à un type de référence qui reste ancré avant

tout dans le linguistique : il s'agit de l'anaphore, qui repose sur le rappel d'un élément du discours. Plus précisément, nous allons nous intéresser aux anaphores du français qui mettent en jeu des pronoms de la 3<sup>ème</sup> personne (*il(s), elle(s), on, le, la, les, lui, eux, etc...*).

L'anaphore a fait l'objet de nombreuses recherches aussi bien pour le traitement des documents écrits qu'en compréhension de parole. En dialogue oral homme-machine, la résolution des anaphores repose généralement sur la consultation de l'historique des éléments de l'application auxquels le dialogue a déjà fait référence (Allen, 1995). Cette approche *ad hoc* donne des résultats très acceptables du fait du caractère finalisé de l'interaction : le nombre d'antécédents candidats reste en effet relativement limité, et le caractère restreint de l'univers du discours modère les possibilités d'inférence référentielles. Pour autant, cette approche à ses limites, même dans le cadre relativement balisé de l'anaphore pronominale :

- on peut tout d'abord s'interroger sur l'efficacité de ces techniques relativement simples sur des domaines applicatifs moins restreints. Or, une des évolutions du domaine consiste précisément à s'attaquer à des tâches plus complexes,
- même dans le cas d'un dialogue finalisé sur une tâche très précise, la résolution de certaines anaphores indirectes risque de mettre en défaut ces approches.

Dans le cadre de recherches plus générales sur la référence en dialogue homme-machine, de nombreux chercheurs ont déjà tenté de dépasser ces limitations (Salmon-Alt, 2001). Leurs travaux relèvent essentiellement d'un niveau d'analyse pragmatique du dialogue. Compte tenu de la complexité des phénomènes étudiés, il est cependant envisageable d'intégrer d'autres informations dans la résolution des co-références anaphoriques. C'est pourquoi nous nous sommes interrogés sur l'adaptation au dialogue oral homme-machine de techniques de résolution développées par le TALN pour la compréhension des documents écrits. Ces approches mettent en effet en jeu des connaissances essentiellement syntaxiques. Cet article dresse un premier bilan de la faisabilité de cette approche, en se basant sur l'étude de plusieurs corpus de dialogue oral.

## 2 Résolution des anaphores pronominales : les postulats du TALN

Avec son « algorithme naïf », Hobbs a montré dès 1978 qu'un ensemble de règles assez simple pouvait permettre de résoudre plus de 88% des anaphores pronominales sur un texte technique. En première approximation, cet algorithme se fonde sur des heuristiques relativement intuitives. D'une part, il cherche prioritairement les antécédents dans l'énoncé courant avant d'aller étudier la phrase précédente : Hobbs a en effet remarqué, dans une étude de corpus écrits, que 98 % des antécédents des pronoms non déictiques se trouvaient dans l'une de ces deux phrases. Ensuite, l'algorithme recherche les antécédents compatibles en genre et nombre avec le pronom, dans une stratégie de recherche gauche droite et en largeur d'abord (préférence aux constituants immédiats plutôt qu'aux sous-constituants enchâssés). Le parcours gauche droite est justifié par un postulat de préférence de rattachement au sujet puis à l'objet sur les autres compléments. Cette stratégie s'applique donc aux langages à ordre fixe SVO.

Cet algorithme a connu plusieurs tentatives d'amélioration — voir par exemple (Lappin et Leass, 1994) — qui ont relativisé la prédisposition du sujet comme antécédent. Il constitue cependant toujours un étalon pertinent de performance. Les améliorations de robustesse apportées par les approches les plus récentes (Mitkov, 1998) ne sont ainsi que légèrement significatives d'un point

de vue statistique. Par contre, une recherche des antécédents reposant sur une analyse syntaxique complète peut rencontrer des difficultés sur un langage aussi peu normé que la parole spontanée :

- d'une part, les réparations, incises et autres procédés de l'oral spontané cassent fréquemment la structure syntaxique des énoncés, ou du moins rendent difficile leur analyse automatique.
- d'autre part, la variabilité significative d'ordonnement linéaire qui caractérise le français parlé (Antoine & Goulian, 2001) risque de mettre à mal une stratégie d'analyse gauche droite, même si nos études de corpus ont montré que le français parlé restait un langage de type SVO.

Cette dépendance à une analyse syntaxique préalable a été relevée par d'autres auteurs (Kennedy & Boguraev, 1996) et a conduit certains chercheurs à étudier des approches alternatives. Globalement, les recherches sur la référence ont pris deux directions opposées au cours de la dernière décennie :

- certains chercheurs ont mis en œuvre des modèles théoriques qui permettent une description plus fine de la référence mais qui ont donné lieu à peu de réalisations opérationnelles, comme le note (Salmon-Alt, 2001). C'est le cas par exemple de la théorie du centrage (*centering theory*) qui propose un modèle local basé sur les deux derniers énoncés (Grosz *et al.* 1995).
- à l'opposé, d'autres travaux suivent une démarche proche du TAL robuste (*shallow parsing*). Certains chercheurs ont ainsi proposé des approches simples et assez robustes exigeant un minimum de connaissances.

Les travaux de Mitkov sont représentatifs de cette seconde démarche (Mitkov, 1998). Mitkov considère comme antécédents possibles tous les groupes nominaux des deux énoncés précédents qui s'accordent en genre et nombre avec le pronom. L'algorithme qu'il a proposé procède ensuite à un classement des candidats suivant des règles de préférence *ad hoc* sans réelle justification linguistique (avantage aux groupes définis sur les indéfinis, désavantage des groupes nominaux introduisant un groupe prépositionnel, etc.). On peut s'interroger sur les fondements et la généralité de ce type d'approche. Les performances obtenues par Mitkov soutiennent néanmoins largement la comparaison avec les techniques alternatives évoquées plus haut.

En dépit de leur diversité, toutes ces approches semblent s'accorder sur un certain nombre de contraintes ou heuristiques que respecte l'anaphore pronominale :

- accord en genre et en nombre entre le pronom et son antécédent,
- localisation de l'antécédent dans l'énoncé portant le pronom, ou à défaut dans le précédent,
- priorité au sujet puis au complément d'objet pour le rattachement (Hobbs, 1978) ou d'une manière plus fine parallélisme fonctionnel entre les fonctions et structures syntaxiques.

Ces techniques n'ont généralement été appliquées qu'à des documents écrits. Afin d'étudier leur transposition en français parlé, nous avons cherché à valider ces hypothèses communes à toutes les travaux du domaine sur différents corpus de dialogue oral.

### 3 Etude de corpus : méthodologie

Cette analyse des usages est menée sur cinq corpus (Air France, CIO<sup>1</sup>, Murol, OTG<sup>2</sup> et MEDIA) de dialogue oral finalisé qui correspondent à des contextes applicatifs ou interactifs (dialogue homme-homme ou homme-machine simulé par la technique du magicien d'Oz) différents, ce qui nous permettra d'atteindre à terme une certaine généralité de diagnostic. Cet article présente les résultats obtenus sur trois premiers corpus : Air France, Murol MEDIA (tableau 1)

**Tableau 1.** Description des corpus ou sous-corpus étudiés.

Corpus	Nombre de dialogues	Nombre de tours de parole	Nombre de mots	Type de dialogue	Domaine applicatif
<b>Air France</b>	103	5 149	49 703	H-H	rens. aérien
<b>Murol</b>	9	1 078	13 500	H-H simulé	rens. touristique
<b>MEDIA</b>	153	4934	—	H-M simulé	réservation hôtel

- Le corpus **Air France** (AF), recueilli par l'équipe de Marie-Annick Morel à l'Université de la Sorbonne Nouvelle, puis retravaillé par Pierre Nerzic dans le cadre du projet DALI (Sabah, 1994) réunit des conversations téléphoniques entre un centre de réservation aérienne et ses clients, qui peuvent être des particuliers ou des personnels d'agence de voyage. Notons que le corpus qui a été étudié dans le cadre de DALI, et sur lequel nous avons travaillé, ne comporte que des dialogues comportant une reformulation.
- Le corpus **Murol** du laboratoire CLIPS-IMAG (Bessac et Caelen, 1995) réunit des conversations téléphoniques simulées entre deux compères jouant respectivement le rôle d'un touriste et d'un employé d'un syndicat d'initiative (renseignement touristique). Pour chaque dialogue, un scénario a été conçu afin de favoriser l'apparition de situations de négociation.
- Le corpus **MEDIA** (Maynard *et al.* 2004) est un corpus simulé de dialogue oral homme-machine. La tâche retenue est la réservation d'hôtel sur un serveur vocal interactif. Suivant la technique du magicien d'Oz, un compère simule le fonctionnement du serveur. Le corpus a été réalisé dans le cadre du programme d'évaluation EVALDA-MEDIA de l'action TECHNOLOGUE du Ministère de la Recherche. Seule une partie du corpus sera accessible avant le déroulement de la campagne d'évaluation. Les travaux présentés ici ont été réalisés sur un extrait significatif (près de 3000 tours de parole) de cette partie.

Nous avons analysé toutes les pronoms personnels sujets, conjoints ou disjoints de la troisième personne présents dans les corpus. Chaque occurrence a fait l'objet d'une annotation détaillée suivant des caractéristiques qui seront reprises dans la synthèse des observations effectuées :

- **présence ou non d'anaphore.** Il n'est pas inutile de rappeler que ces pronoms personnels ne réfèrent pas toujours à un élément du discours (usages impersonnels, références indirectes).

<sup>1</sup> Le corpus CIO a également été réalisé par l'équipe de Marie-Annick Morel (Morel, 1989).

<sup>2</sup> Voir (Antoine 2002 ; Antoine *et al.*, 2002)

- **distance entre le pronom et son antécédent**, mais également de la dernière référence à l'antécédent, dans le cas d'anaphores multiples. Cette distance est mesurée en nombre de tours de parole<sup>3</sup>. Cette métrique est le pendant des distances exprimées en nombre de phrases dans les travaux sur l'écrit. Les cataphores sont également décrites par cet attribut.
- **fonction syntaxique** (prédicat, sujet, objet, autres compléments, modificateurs), **genre** et **nombre** du pronom, de son antécédent et éventuellement de la dernière référence à ce dernier

## 4 Etude de corpus : résultats

### 4.1 Fréquence d'apparition et distribution des anaphores pronominales

Les chercheurs qui ont travaillé sur la référence ont souvent eu l'occasion d'observer que l'usage de l'anaphore est relativement variable d'un corpus à un autre. L'analyse des fréquences d'apparition des anaphores pronominales (tableau 1) révèle ainsi des différences d'usages significatives entre le corpus MEDIA d'une part, et les corpus Murol et Air France d'autre part. La situation d'interaction influe sur le recours à l'anaphore. Son usage est moins fréquent en dialogue homme-machine, sans doute parce que l'ajustement de l'interaction est plus difficile dans ce cas (le locuteur semble vouloir éviter au système des erreurs d'interprétation) mais aussi parce que les dialogues directifs de MEDIA ne favorisent pas l'apparition d'anaphores.

**Tableau 2** — Fréquence d'apparition des anaphores pronominales (3<sup>e</sup> personne)

Corpus	Murol	Air France	MEDIA
% de tours de parole avec anaphore	5,9 %	4,5 %	0,9 %

Cette variation de fréquence n'a par contre que peu d'influence sur les usages en eux-mêmes. Par exemple, la distribution des occurrences des pronoms suivant leurs utilisations témoigne d'une stabilité structurelle significative (tableau 3). Nous avons distingué trois types d'usages :

- anaphore (ou cataphore) directe ou indirecte.
- anaphore implicite, forme complexe d'anaphore inférentielle pour laquelle il n'existe aucune trace lexicale dans le dialogue permettant d'amorcer le calcul de la référence :

(1) *à 12h10 attendez je l'appelle sur une autre ligne* (AF I-13 :C7)

Dans cet exemple, il n'a jamais été fait (et ne sera jamais fait) mention explicite au cours du dialogue du client concerné qui est visiblement l'objet de la référence.

<sup>3</sup> Cette unité de segmentation peut avoir une portée différente suivant la définition qu'on lui donne. Dans cette étude, nous avons repris la définition opératoire utilisé dans les conventions de transcription du projet Parole Publique (Antoine *et al.* 2002) : on définit un nouveau tour de parole à chaque fois qu'un nouveau locuteur prend la parole, que ce soit à la suite d'un interlocuteur ou dans le cas d'un chevauchement (Antoine, 2002).

- usage impersonnel figé (*il y a*) ou non figé : *il est certain qu'il y a* (AF I-64 :O10), marque de certitude générale à comparer à : *Jean, il est certain qu'il y a eu de la vie sur Mars.*

Nous ne nous intéresserons pas ici aux impersonnels figés qui peuvent être aisément détectés par l'analyse automatique. Le tableau 3 donne la distribution des pronoms dans les autres situations.

**Tableau 3**— Distribution des pronoms suivant leurs usages dans les corpus Air France et Murol

<b>Corpus</b>	<b>anaphores / cataphores</b>	<b>anaphores implicites</b>	<b>impersonnels non figés</b>
<b>Air France</b>	80 %	<b>12 %</b>	<b>8 %</b>
<b>Murol</b>	91 %	<b>3 %</b>	<b>6 %</b>
<b>MEDIA</b>	87 %	<b>4 %</b>	<b>9 %</b>

On constate une nette prédominance des anaphores directes ou indirectes sur les trois corpus. Deux remarques complémentaires peuvent néanmoins être faites. D'une part, on observe une sur-représentation des anaphores implicites sur le corpus Air France. Ce résultat traduit l'influence du contexte applicatif sur la référence. Une part importante de ces observations correspond à une mention implicite du client, comme dans l'exemple (1) étudié plus haut. Du fait du caractère extrêmement ciblé de la tâche (réserver un billet pour un client), les inférences nécessaires à la résolution de cette référence implicite sont réduites, d'où un recours facilité à ce procédé.

D'autre part, le tableau 3 montre que dans 9% (Murol) à 20 % (Air France) des cas, le pronom ne réfère pas ou n'a pas d'antécédent exprimé dans le dialogue. Certains auteurs intègrent une détection des impersonnels dans leur analyse (Kennedy et Boguraev, 1996). Mais à ma connaissance, aucune approche ne propose de caractériser les usages implicites. Il s'agit d'une limitation importante (et peu relevée jusqu'ici) pour le dialogue oral homme-machine finalisé.

## 4.2 Le postulat de l'accord en genre et en nombre

Le tableau 4 fait la synthèse du pourcentage d'accord en genre et en nombre entre, d'une part, le pronom et son antécédent, et d'autre part le pronom et la dernière référence à cet antécédent

**Tableau 4** — Respect de l'accord en genre et en nombre entre le pronom et son antécédent

<b>Corpus</b>	<b>accord pronom / antécédent</b>		<b>accord pronom / dernier référent</b>	
	<b>genre</b>	<b>Nombre</b>	<b>genre</b>	<b>nombre</b>
<b>Air France</b>	94 %	82 %	96 %	91 %
<b>Murol</b>	100 %	97 %	100 %	99 %
<b>MEDIA</b>	95 %	88 %	100 %	88 %

Pour le corpus Murol, le non respect de l'accord reste relativement l'exception. A l'opposé, la

situation est plus partagée pour ce qui concerne l'accord en nombre dans les corpus MEDIA et Air France. Pour la plupart, ces cas correspondent à des emplois métonymiques du pronom *ils*. L'exemple (2a) issu du corpus Air France est représentatif de ce type d'usage :

(2a) *voilà e pouvez vous me dire s'il y a eu une réponse de l'hôtel* (AF I-89 :C5)

[...] *oui mais // mais est ce est ce qu'ils vous ont déjà répondu* (AF I-89 :C8)

On le retrouve sous la même forme dans le corpus MEDIA, pour qui l'accord en nombre est problématique dans 12 % des cas :

(2b) *le Mercure centre voulez-vous réserver dans cet hôtel* (MEDIA 08730-531 :C9)

[...] *euh s'ils acceptent les chiens* (MEDIA 08730-531 :S11)

Ce type d'anaphore indirecte entre une classe et ses éléments particuliers ou constitutifs a déjà été décrite (Charolles 2002, Berrendonner & Reichler-Béguelin 1995) mais n'a pas fait l'objet d'étude distributionnelle. Il remet pourtant en question, à l'oral du moins, le postulat d'accord sur lequel s'appuient la plupart des techniques actuelles. On notera par ailleurs que ce phénomène peut être encore plus marqué :

- d'une part, la différence singulier / pluriel n'est pas toujours marquée phonétiquement
- d'autre part, les études en cours sur le corpus OTG (Antoine 2002, Antoine *et al.* 2002) tendent vers une équiprobabilité entre accord et non accord en nombre !

Les variabilités que l'on observe entre MUROL et les autres corpus semblent traduire une fois encore l'influence de l'univers de la tâche : un cadre applicatif peut en effet être plus favorable à la mise en place de ces métonymies. Ces observations révèlent en tout cas un réel problème, même si le phénomène est circonscrit aux pronoms pluriels *ils* et *elles*.

Notons enfin que l'accord est un peu plus respecté dans le cas d'une mise en correspondance entre le pronom est la dernière référence à l'antécédent : une fois la métonymie mise en place (et acceptée par l'interlocuteur), rien ne s'oppose à la poursuite de cet usage. Par exemple, le dialogue (2a) se poursuit comme suit :

(3) [...] *si on ne les relance pas* [...] (AF I-89 :C5)

### 4.3 La préférence aux antécédents les plus proches

Une autre hypothèse semble également moins fiable en dialogue oral. Il s'agit de la distance, en nombre de tours de parole, entre le pronom et son antécédent (ou également avec la dernière référence à ce dernier). Le tableau 5 (page suivante) qui présente les distributions correspondantes, ne tient pas compte des cataphores, qui ne représentent que 1% à 2% des anaphores observées sur nos corpus. Notons cependant les techniques actuelles de résolution ne gèrent pas les cataphores.

Hobbs avait montré dans ses études sur des textes techniques que 90% des anaphores pronominales sont interphrastiques, et que 98% des antécédents se situent dans la phrase

courante ou la précédente. Comme bien d'autres depuis, le système de résolution proposé par (Mitkov 1998) limite lui aussi la recherche de l'antécédent aux deux derniers énoncés. En dialogue oral, il paraît cohérent de transposer cette caractérisation à l'énoncé courant et au tour de parole précédent de chacun des interlocuteurs (distances 0, 1 ou 2 dans le tableau 5).

**Tableau 5** — Distribution des anaphores pronominales suivant la distance en nombre de tours de parole entre le pronom et l'antécédent (même tour de parole: dist. = 0 ; tour précédent: dist. = 1)

Corpus	distance pronom / antécédent				distance pronom / dernier référent			
	0	1	2	> 2	0	1	2	>2
<b>Air France</b>	35%	15%	17%	<b>33%</b>	46%	19%	19%	<b>16%</b>
<b>Murol</b>	25%	18%	32%	<b>25%</b>	27%	24%	33%	<b>16 %</b>
<b>MEDIA</b>	30%	38%	8%	<b>14%</b>	38%	46%	8%	<b>6 %</b>

Dans tous les cas, on observe une dégradation de ce critère de récence sur nos corpus oraux. Si le corpus MEDIA est celui qui s'approche le plus des distributions observées à l'écrit, 14 % des antécédents sont antérieurs aux deux derniers tours de parole. Cette proportion monte à plus d'un quart sur les corpus Air France et Murol. Ces proportions restent encore élevées si on s'intéresse à la distance entre le pronom et la dernière référence à l'antécédent. De par son ancrage interactif et pragmatique, le dialogue oral finalisé semble ainsi pouvoir référer à des éléments beaucoup plus lointains que le langage écrit.

Lorsque l'ajustement dynamique et interactif de l'interprétation des pronoms est moins assuré (dialogue personne-système), la prise de risque semble plus limitée. Cela explique sans doute les distributions plus encourageantes, en terme d'analyse automatique, observées sur le corpus MEDIA. Si la récence d'information reste pertinente en dialogue homme-machine, la barrière des « deux derniers énoncés » doit donc être interprétée d'une manière moins stricte qu'à l'écrit.

#### 4.4 L'influence de la fonction syntaxique

L'algorithme naïf de Hobbs (1978) reposait sur le postulat d'une prédisposition du sujet comme antécédent. Cette hypothèse a été depuis raffinée par les chercheurs du domaine (Lappin et Leas 1994 ; Rolland 2001). Considérée brutalement, elle est d'ailleurs totalement infirmée par nos études de corpus oraux (tableau 6). En dépit de variations des distributions entre les corpus, on ne relève en effet aucune prédominance du sujet.

**Tableau 6** — Distribution de la fonction syntaxique des antécédents (ou du dernier référent)

Corpus	sujet	objet	autres	dislocation orale
<b>Air France</b>	28 %	33 %	14 %	23 %
<b>Murol</b>	13 %	54 %	11 %	22 %
<b>MEDIA</b>	12 %	20 %	45 %	23 %



Plus intéressant est le relevé des antécédents situés dans une construction de l'oral spontané (dislocation, incise...). Nos résultats (tableau 6) montrent que près d'un antécédent sur cinq se trouve en moyenne dans ces constructions qui cassent la structure syntaxique de l'énoncé. En dialogue oral, il semble donc difficile de baser la résolution des anaphores sur la fonction syntaxique.

Afin d'affiner cette conclusion, nous avons cherché à établir s'il existait une corrélation entre les rôles syntaxiques du pronom et de l'antécédent, comme le suggère l'algorithme RAP (Lappin et Leas, 1994) ou encore le filtre SYR défini par (Rolland 2001). En dépit d'une variabilité sensible des observations, le tableau 7 montre malheureusement que cette corrélation n'existe pas à l'oral.

**Tableau 7** — Corrélation entre la fonction syntaxique de l'antécédent et le pronom

Corpus	Air France	Murol	MEDIA
% accord sur la fonction	42 %	42 %	15 %

## 5 Conclusion

Si ce n'était pas le but premier de cette étude, les analyses variationnelles qui ont été menées sur ces corpus permettent de caractériser différentes sources d'influence sur la réalisation des anaphores pronominales en dialogue oral finalisé :

- influence du domaine de la tâche ou de l'univers du discours sur les emplois métonymiques ou implicites, avec comme corollaire une influence sur le respect de l'accord en nombre entre le pronom et son antécédent,
- influence du type d'interaction (homme-homme ou homme-machine) sur la fréquence et dans une moindre mesure sur la portée (distance avec l'antécédent) des anaphores pronominales.

Par delà ces variations, la cohérence des résultats obtenus montre que les heuristiques utilisées à l'écrit ne peuvent être reprises strictement en dialogue oral. Certains indices (accord en nombre, portée de l'anaphore) s'avèrent moins fiables sur l'oral spontané voire totalement inopérants (fonction syntaxique). Cette étude laisse toutefois un espoir pour l'adaptation de ces techniques au dialogue oral homme-machine. C'est désormais notre objectif dans le cadre de nos travaux en compréhension de la parole.

## Références

ALLEN J. (1995) Local Discourse Context and Reference. *In* Allen J. *Natural language understanding*. Benjamins / Cummings, Redwood City, CA. 429-461.

ANTOINE J.-Y. (2002) Corpus OTG : présentation générale. Rapport technique VALORIA-CORAIL-2002-2. [http://www-valoria.univ-ubs.fr/antoine/rapports\\_CORAIL.htm](http://www-valoria.univ-ubs.fr/antoine/rapports_CORAIL.htm).

ANTOINE J.-Y. (2003) Pour une ingénierie des langues plus linguistique. Habilitation à Diriger les Recherches. Université de Bretagne Sud, Vannes, France. 28 novembre 2003.

- ANTOINE J.-Y., GOULIAN J. (2001) Etude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé. Application à la communication orale homme – machine, *TAL*, 42(2), 413-440.
- ANTOINE J.-Y., LETELLIER-ZARSHENAS S., SCHADLE I., NICOLAS P., CAELEN J. (2002) Corpus OTG et ECOLE\_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement. Actes *TALN'2002*. Nancy, France
- BERRENDONNER A., REICHLER-BEGUELIN M.J. (1995) Accords associatifs, *Cahiers de praxématique*, 24, 14-21.
- BESSAC M. ET CAELEN J. (1995) Analyses pragmatiques, prosodiques et lexicales d'un corpus de dialogue oral homme-machine, Actes *JADT'95*, Rome, Italie. 363-370.
- CHAROLLES M. (2002) La référence et les expressions référentielles en français. Ophrys, Gap.
- GROSZ B.J., JOSHI A.K., WEINSTEIN S. (1995) Centering : a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2), 203-225.
- HOBBS J. R. (1978) Resolving pronoun references. *Lingua*, 44, 331-338.
- LAPPIN S., LEASS H.J. (1994) An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 535-561
- KENNEDY C., BOGURAEV B. (1996) Anaphora for everyone : pronominal anaphora resolution without a parser. Actes *COLING-96*, Copenhague, Danemark. 113-118
- MAYNARD H. ET AL. (2004) Constitution d'un corpus de dialogue oral pour l'évaluation automatique de la compréhension hors et en contexte de dialogue. Actes *JEP'2004*, Fés, Maroc.
- MITKOV R. (1998) Robust pronoun resolution with limited knowledge. Actes *ACL-COLING'98*, Montréal, Canada, 869-875.
- MOREL M.A. (1989) Analyse linguistique d'un corpus ; 2° corpus : centre d'information et d'orientation de l'Université Paris V. Publications de la Sorbonne Nouvelle, Paris, France.
- PIERREL J.-M., ROMARY L. (2000) Dialogue oral homme-machine. In Pierrel J.-M. (Dir.) *Ingénierie des langues*. Hermès, Paris, France. 331-349.
- ROLLAND S. (2001) Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm. *Computational Linguistics*, 27 (4), 479-506
- SABAH G. (1994) Projet DALI. *rapport d'activité GDR-PRC CHM*, CNRS. 71-88.
- SALMON-ALT S. (2001) Référence et dialogue finalisé: de la linguistique à un modèle opérationnel. Thèse U. Henri Poincaré, Nancy, France. 16 mai 2001.