Le magicien d'Oz pour du dialogue oral : expérience avec un assistant virtuel en entreprise.

Y. Fouquet

CLIPS-IMAG Laboratory, Joseph Fourier University
B.P. 53, 38041 Grenoble cedex 9, France
Yannick.Fouquet@imag.fr

ABSTRACT

This paper presents a platform for testing human-computer spoken dialog system, exploring future applications of automatic speech dialog in enterprise domain. The purpose is to simulate by wizard of Oz a virtual secretary assistant which could respond to different task actually reserved to real secretaries. This paper presents a general architecture of the test platform which would build a human-computer spoken dialog corpus. Wizard of Oz for understanding and for dialog management is then presented. First results obtained show a credible way of capturing and annotate spoken dialogs.

RESUME

A travers cet article, nous présenterons nos travaux sur une plate-forme de test de dialogue homme-machine, explorant le dialogue oral avec pour exemple un assistant électronique dans le domaine de l'entreprise. Nous présenterons d'abord l'architecture générale de la plate-forme de test nous permettant de recueillir un corpus de dialogue oral homme-machine suivant certaines tâches dévolues aux secrétaires. Nous proposerons ensuite des magiciens d'Oz pour simuler la compréhension et le contrôle de dialogue. Enfin, nous aborderons les premiers résultats et les perspectives du projet.

1. INTRODUCTION

Les interfaces utilisant le canal parole sont souvent plus appropriées que celles uniquement graphiques. En effet, elles sembles plus naturelles et permettent leur accès à certains individus non-voyants, handicapés, utilisant leurs yeux ou leurs mains pour d'autres tâches, etc. Pour ces raisons, ces interfaces devraient rapidement devenir un paradigme de communication plus utilisé.

Pour le développement de plates-formes de dialogue oral avec le souhait d'une génération automatique de tels systèmes de dialogue, nous avons suivi la méthodologie montrée par Rouillard & Caelen dans [Rou00] et [Rou98]. Nous avons opté pour un contexte d'entreprise dans l'idée d'améliorer les échanges d'information au sein de l'entreprise. Dans un premier temps, nous avons enregistré plus de 800 dialogues entre des secrétaires et leurs interlocuteurs. Nous avons ensuite repéré les dialogues les plus représentatifs des tâches habituelles. Ceux-ci ont ensuite été analysés dans le but de mettre en exergue leurs particularités et leurs parties génériques. Cette analyse, première pierre de la méthodologie, nous a permis de développer un premier système de dialogue. En revanche, le dialogue inter-humains ne semble pas suffisant pour créer un système dans lequel la machine remplacerait l'un des deux interlocuteurs. C'est pourquoi nous avons développé un premier magicien d'Oz afin de simuler le comportement du système et ainsi élaborer un corpus de dialogue oral homme-machine. L'analyse de ce corpus a permit de montrer certaines différences entre du dialogue oral inter-humains et du dialogue oral homme-machine. Le magicien simule le contrôleur de dialogue d'un assistant en entreprise. Le client appelle ce secrétaire virtuel pour différentes tâches (les plus représentatives de celles de secrétaires) comme la mise en relation avec une personne de l'entreprise, la prise de rendez-vous, la réservation de salle, la gestion d'un agenda partagé, l'envoi de documents ou la réception d'information. Il communique actuellement via un micro-casque.

Cet article présentera d'abord l'architecture générale de la plate-forme de test considérant le magicien d'Oz comme une solution acceptable pour capturer des dialogues oraux homme-machine et tester la première version du système (les phrases servant de réponse). Nous présenterons ensuite le corpus enregistré pendant le projet et les différents modules de reconnaissance vocale, interprétation en actes, contrôle de dialogue et synthèse de parole.

2. ARCHITECTURE GENERALE DU SYSTEME

Le système de dialogue est composé de quatre parties dont la reconnaissance vocale en début et la synthèse text-to-speech en fin. Les deux autres sont simulés dans notre approche par des magiciens d'Oz. L'un simule la partie interprétation en actes par l'annotation en actes de dialogue des hypothèses de reconnaissance. L'autre s'occupe du contrôle de dialogue en recevant les hypothèses de reconnaissance et les actes de dialogues associés, choisissant la réponse la plus appropriée et la renvoyer à la synthèse text-to-speech du client. La Figure 1 montre l'architecture du système de dialogue. Dans lequel chacune des quatre parties est implantée comme un serveur qui communique avec les autres parties.

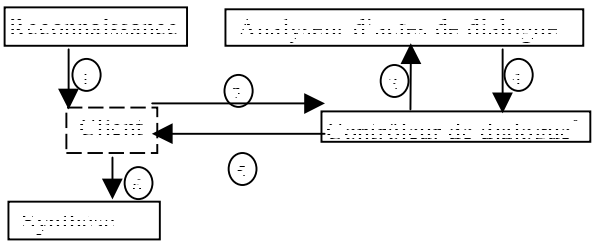

Figure 1: Architecture du système de dialogue.

Le scénario est le suivant :

1. Le client reçoit l'**hypothèse** de reconnaissance d'un serveur de reconnaissance local.
2. Le client envoie cette **hypothèse** au contrôleur de dialogue.
3. Le contrôleur de dialogue envoie cette **hypothèse** à l'analyseur d'actes de dialogue.
4. L'analyseur d'actes envoie en retour l'**acte de dialogue** correspondant à l'hypothèse reçue.
5. Le contrôleur de dialogue choisit la **réponse** la plus appropriée et l'envoie au client.
6. Le client envoie la **réponse** reçue au serveur local de synthèse

3. COLLECTE DE DONNEES

Afin de définir les spécifications du modèle de dialogue, les tâches habituelles et le vocabulaire de l'application, nous avons enregistré plus de 800 vrais dialogues entre des secrétaires et leurs interlocuteurs.

Nous avons repéré 3 tâches très fréquentes : joindre une personne, prendre un rendez-vous et réserver une salle. Nous avons ensuite ajouté 3 autres tâches considérées comme importantes pour un assistant virtuel : gérer un agenda partagé, recevoir une information et envoyer un document.

Nous avons ensuite transcrit 44 dialogues choisit pour leur adéquation avec les tâches sélectionnées. Ces transcriptions ont permit de définir le vocabulaire de l'application et des classes propres à l'application.

L'analyse de ces transcriptions a été faite alors en vue de développer le model de dialogue et le premier magicien d'Oz. Elle a permit de déterminer 122 phrases typiques (comprenant des variantes et des paramètres) que le système devra pouvoir dire. Des variantes déduites de l'analyse ont été introduites afin d'obtenir des dialogues plus naturels. En outre, les phrases contiennent des paramètres dépendants de la tâche : le nom et patronyme de l'utilisateur, les paramètres de date (jour de la semaine, jour du mois, mois, année) et d'heure (heure, minutes). Le contrôleur de dialogue choisit les paramètres appropriés dans un annuaire, un calendrier et d'autres menus (pour les paramètres particuliers tels l'heure, les minutes et le moment de la journée).

4. COMPOSANTS DE DIALOGUE ORAL

4.1 Reconnaissance de parole française, un modèle de langage spécifique.

Le système de reconnaissance vocale en parole continue française utilise la boite à outils Janus-III toolkit [Wos93] du CMU. Le modèle acoustique dépendant du contexte issu de [Bes01] a été appris sur un corpus, extrait de Bref80 [Lam91], de 12 heures de parole continue prononcée par 72 locuteurs. Le module de reconnaissance vocale, composant très important du système complet, a été adapté aux tâches choisies en créant un modèle de langage spécifique à ces tâches.

Corpus & vocabulaire pour l'apprentissage d'un modèle de langage.

Dans un premier temps, de vrais dialogues oraux ont été enregistrés mettant en relation des secrétaires et leur interlocuteur. Les dialogues les plus représentatifs des tâches les plus fréquentes d'un assistant ont été sélectionné puis transcrits pour analyse. Ces transcription nous ont également servi à élaborer un corpus adapté aux tâches auxquelles le système doit pouvoir répondre. Ce corpus étant fait de dialogues inter-humains, il ne peut être suffisant pour l'adapter à du dialogue homme-machine. Nous avons donc collecté de nouveaux dialogues oraux par le biais d'un magicien d'Oz afin d'avoir du dialogue homme-machine. Ces dialogues ont été rajoutés au premier corpus afin d'obtenir un corpus final de 1110 phrases composées d'un vocabulaire de 1119 mots.

La difficulté majeure du processus de reconnaissance vocale consiste à élaborer un bon modèle de langage. Pour obtenir le meilleur, nous avons élaboré et testé de nombreux modèles. Les premiers tests combinent le corpus du journal Le Monde (1997-2001) avec le vocabulaire de notre application (1119 mots) enrichi des X mots les plus fréquents de Le Monde où X varie de 0 à 25000. Considérant l'amélioration du web comme corpus [Vau996], nous avons testé également des modèles appris sur un corpus collecté sur le web. Nous avons fait les derniers tests sur des modèles appris sur notre corpus spécifique (les transcriptions des dialogues réels) et avec la boîte à outils CMU Statistical Language Modeling (<http://www.speech.cs.cmu.edu/SLM_info.html>) pour le premier et la boîte à outils SRI-LM Language Modeling (http://www.speech.sri.com/projects/srilm/) pour le deuxième. Ainsi la comparaison de 8 modèles a été faite afin de conserver le modèle le plus adapté à un système de dialogue : notre corpus spécifique et la boîte à outils SRI-LM.

Evaluation

Pour évaluer les performances du système de reconnaissance vocale, nous avons essayé avec celui-ci de

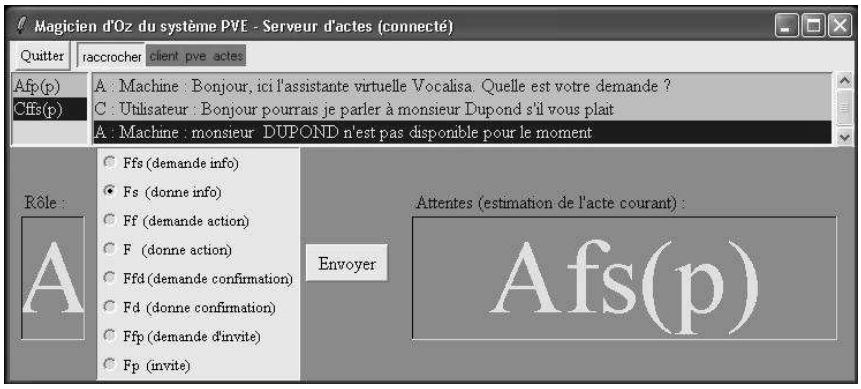

Figure 2 : Un magicien d'Oz pour la compréhension en actes de dialogues

reconnaître 63 phrases prononcées pendant un premier beta-test du système.

Ces phrases sont issues de situations de dialogues réels dans lesquelles les utilisateurs avaient à parler, avec des requêtes spécifiques définies par avance, au système simulé par magicien d'Oz. Les meilleures performances ont été trouvées avec notre corpus spécifique et la boîte à outils SRI-LM. Les autres solutions sont trop mauvaises pour les considérer acceptables. Le Monde est trop loin des spécificités du dialogue tandis que le web ignore les spécificités des phrases courtes trouvées dans nos dialogues oraux. Le taux d'erreur de mot obtenu dans nos meilleures condition est de 39% ce qui semble tout à fait acceptable pour les magiciens qui n'ont demandé que trois fois à l'utilisateur de répéter dans les six dialogues ayant servis au test.

4.2 Analyseur d'actes de dialogue, un magicien d'Oz

Afin d'être le plus générique possible, le modèle de dialogue est fondé sur la théorie des actes de langage introduite par Austin [Aus62] & Searle [Sea72]. Ainsi, considérant que la compréhension du texte en actes est correctement faite, travailler à un niveau pragmatique nous permet d'aborder des combinaisons d'actes de langage et stratégies de dialogue afin de trouver la réponse la plus adéquate.

Ainsi, l'analyse en actes de dialogues est un point crucial du projet car de la correcte identification de ces actes dépend tout les reste du modèle. Nous avons trouvé quelques travaux sur cette identification comme la thèse de Colineau [Col97] qui utilise des réseaux de neurones mais pas d'applications génériques.

C'est pourquoi un magicien d'Oz a été développé pour annoter l'hypothèse de reconnaissance vocale en actes de dialogue. Le magicien reçoit donc la phrase reconnue par le système. Il peut garder en mémoire des marqueurs du discours de Colineau comme « je veux », « je voudrais », etc. Il choisit dans une liste d'actes, le plus approprié. Pour faire ce choix, il est aussi aidé des attentes, introduites dans [Fou01], déduites de l'acte précédent [Fou02]. La première attente, c'est-à-dire celle qui maximise la probabilité de prédiction, est présélectionnée. Lorsque le magicien envoie l'acte correspondant à la phrase, les attentes sont mises à jour de sorte que les probabilités sont apprises dans le contexte de l'application.

Dans l'exemple Figure 2, Le magicien reçoit une première phrase de demande d'information : « Bonjour, pourrais-je parler à monsieur Dupond s'il vous plaît ? ». Il l'annote $F^{fs}_{C}(p)$ qui représente la demande d'information du client. L'attente la plus probable pour cet acte $F^{fs}_{C}(p)$ est une réponse de l'ordinateur $F^{s}_{A}(p)$ avec une probabilité de 77%. Lorsque le magicien recevra la phrase suivante (« Monsieur Dupond n'est pas disponible pour le moment ») du système, $F^{s}_{A}(p)$ sera présélectionné de sorte qu'il lui suffit de confirmer si cet acte est le bon. Dans ce cas, la probabilité précédente est incrémentée considérant cette nouvelle occurrence. Dans le cas contraire (par exemple l'attente suivant, $F^{fs}_{A}(p)$, correspondant ici à une demande de précision avec la phrase « avez-vous essayé son poste direct ? », le magicien sélectionne l'acte $F^{fs}_{A}(p)$ et le confirme pour l'envoyer au système et les probabilités seront mises à jour avec cette nouvelle occurrence.

Cette aide à l'annotation par la prédiction de l'acte suivant nous semble d'ailleurs une issue intéressante de notre théorie des attentes. Des travaux futurs viseront à l'évaluation de cette prédiction d'acte de dialogue en utilisant la première attente.

4.3 Contrôleur de dialogue, un autre magicien d'Oz

Le but ici est de collecter et tester des dialogues homme-machine dans lesquels le système doit pouvoir répondre aux tâches les plus répandues d'un secrétaire. Le système final sera complètement automatique mais nous voulions d'abord un corpus de dialogue homme-machine spécifique aux tâches choisies avec l'hypothèse que les dialogues homme-machine sont différents de ceux inter-humains

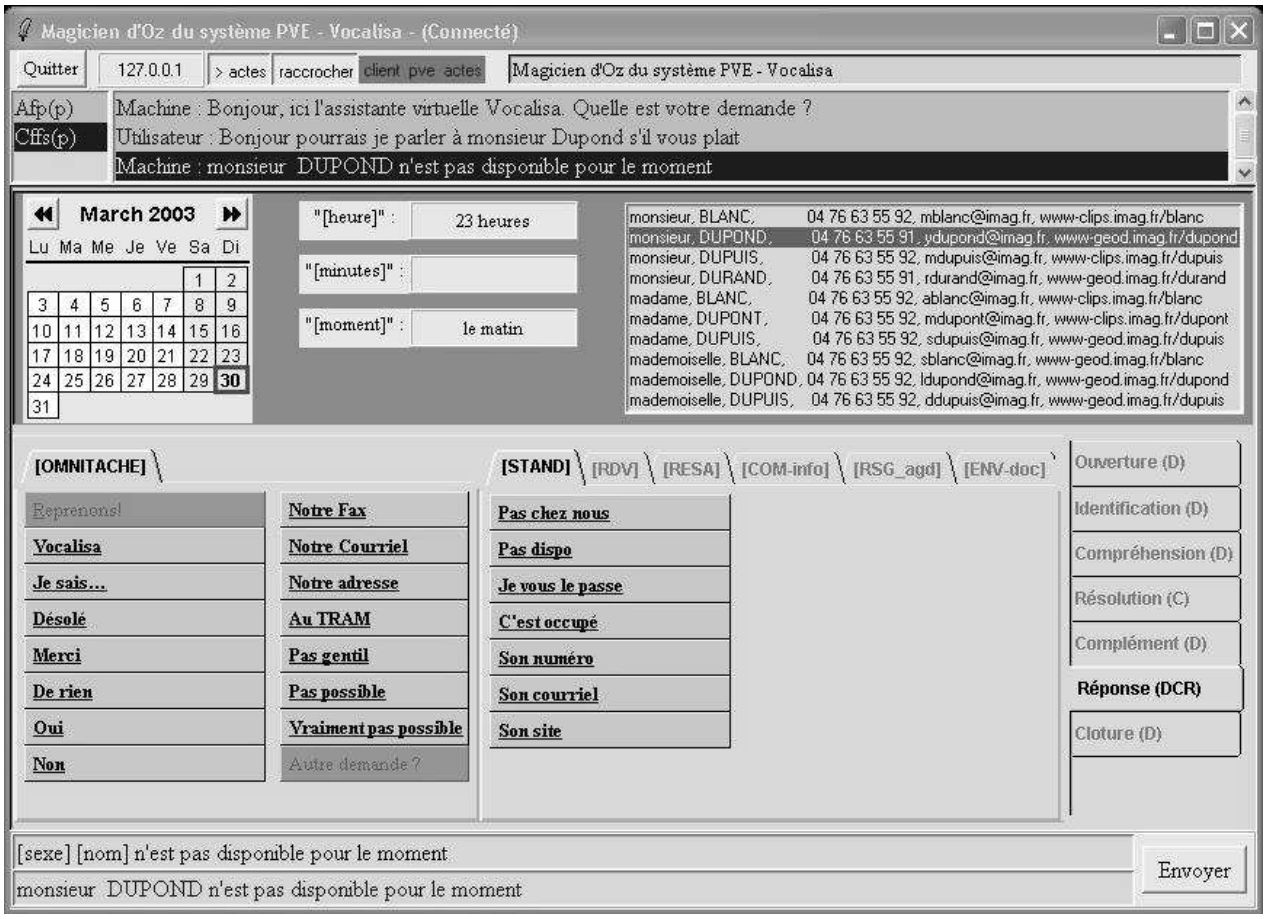

Figure 3 : Un magicien d'Oz pour le contrôleur de dialogue

particulièrement en dialogue oral où les conventions sociales sont omniprésentes.

Nous avons développé un magicien d'Oz afin de simuler complètement le modèle de dialogue. Celui-ci reçoit l'hypothèse de reconnaissance vocale de l'utilisateur (afin de tenir compte des erreurs de reconnaissance). Il reçoit également l'acte de dialogue, correspondant à cette hypothèse, du premier magicien (Figure 2). Il a alors le rôle de contrôleur de dialogue, contrôleur de tâche et générateur de réponse en choisissant les paramètres adéquats dans l'interface de la tâche et une réponse appropriée dans la liste des 122 phrases obtenues à partir des transcriptions manuelles des dialogues réels (Figure 3).

Lorsqu'il a choisit la meilleure réponse et ses paramètres associés, il l'envoie à la synthèse du client. Par exemple, lorsque l'utilisateur dit « Bonjour, pourrais-je parler à monsieur Dupond s'il vous plaît ? », le magicien peut sélectionner monsieur Dupond dans son annuaire et répondre « Monsieur Dupond n'est pas disponible pour le moment » ou « Je vous passe monsieur Dupond » en sélectionnant l'étiquette correspondante puis la phrase correspondante dans le menu lié à cette étiquette.

En outre, ce magicien utilise la première attente déduite de l'acte reçu. Celle-ci est mise en valeur (l'étiquette correspondante est soulignée) afin de faciliter le choix mais le magicien peut choisir une autre réponse.

D'autres aides sont apportées comme des valeurs par défaut aux paramètres, des onglets pour séparer les différentes tâches et une partie omnitâche regroupant toutes les phrases valables dans toutes les tâches. En outre, l'analyse du corpus inter-humains a permit d'élaborer un diagramme de l'architecture de dialogue. Nous avons suivi ce diagramme en classant chacune de ces 122 phrases en fonction des tâches et des phases, mettant chaque phase soulignée en onglets. Le magicien choisit la tâche en début de dialogue mais peut en changer en cours de dialogue. Il suit ensuite le diagramme en choisissant la phase (par exemple introduction, compréhension du but ou réponse) et choisit dans les phrases de la phase sélectionnée.

La dernière aide consiste en la coloration de coopération. En plus des réponses, la stratégie correspondante (directive, réactive ou coopérative) est montrée avec un code de couleur de sorte que le magicien puisse choisir sa stratégie lorsqu'il répond. Il peut donc être directif puis réactif ou coopératif en choisissant la phrases par sa

couleur lorsqu'il a le choix de différentes stratégies. Pour l'exemple précédent, choisir « Monsieur Dupond n'est pas disponible pour le moment » ou « Je vous passe monsieur Dupond » correspond à une stratégie réactive tandis que « Avez-vous essayé son poste direct ? » semble être davantage directif.

4.4 Synthèse de parole française

La dernière partie de notre système de dialogue oral consiste en une synthèse Text-to-speech (TTS) qui produit un signal audio à partir du texte.

La synthèse a été développée en tant que serveur qui reçoit le texte de l'utilisateur et répond en créant le fichier son correspondant et en le jouant. Il a été intégré au serveur de reconnaissance vocale parce qu'ils utilisaient tous les deux le son et que la reconnaissance vocale doit être stoppée pendant le processus de synthèse.

Afin de minimiser le temps d'attente de l'utilisateur, la synthèse doit se faire le plus vite possible. Le système Mbrola TTS [Bag] de la faculté Polytechnique de Mons (<http://tcts.fpms.ac.be/synthesis/euler>) nous semble suffisamment rapide. Il nécessite environ 30% du temps de la phrase pour produire le fichier son correspondant. La qualité de synthèse vocale nous a semblé juste acceptable pour notre application.

5. CONCLUSIONS ET PERSPECTIVES

Nous avons présenté dans ce papier les différents modules (dont deux simulés par magicien d'Oz) développés pour l'élaboration d'un corpus de dialogues oraux homme-machine et l'annotation de parole. Pour ces deux magiciens, nous avons montré que la théorie des attentes semble être une perspective intéressante pour améliorer la compréhension en actes de dialogues mais aussi le contrôle du dialogue en prédisant l'acte de dialogue le plus attendu et les autres moins attendus mais possibles.

Ces modules nous ont servi à enregistrer 12 dialogues pour chacune des 6 tâches retenues considérant les tâches les plus communes pour des secrétaires. L'analyse de ces 72 dialogues nous permettra de tester et de mettre à jour le système de dialogue. Elle servira ensuite à l'automatisation de l'interprétation pour une tâche spécifique et un contrôleur de dialogue générique utilisant tous deux les attentes en dialogue.

REMERCIEMENTS

Cet article s'appuie sur les travaux d'Anne-Claire Descalle qui a transcrit, analysé et montré l'architecture des dialogues pour les tâches retenues. Les phrases du contrôleur de dialogue ont été formées par Anne-Claire Descalle et Solange Hollard. Les modèles de langage pour le système de reconnaissance n'auraient pas pû être faits sans l'aide de Brigitte Bigi. Grand merci à elles donc.

BIBLIOGRAPHIE

[Rou00] Rouillard J. (2000), Hyperdialogue Sur Internet ; Le Système Halpin, Thèse De Doctorat Informatique, 2000.

[Rou98] Rouillard, J. & Caelen, J., *Etude Du Dialogue Homme-Machine En Langue Naturelle Sur Le Web Pour Une Recherche Documentaire*, Deuxième Colloque International Sur L'apprentissage Personne-Système, Caps'98, Caen, Juillet 98.

[Wos93] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "Recent Advances in JANUS : A Speech Translation System". *Eurospeech*, 1993, volume 2, pp. 1295-1298.

[Bes01] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz (2001), Speech Translation For French In The Nespole! European Project, Eurospeech 2001.

[Lam91] Lamel, L.F., Gauvain, J.L., Eskénazi, M. "BREF, a Large Vocabulary Spoken Corpus for French", *Eurospeech*, Gênes, Italy, Vol 2, pp. 505-508, 24-26 September 1991.

[Vau99] Vaufreydaz, D., Akbar, M., Rouillard, J., Internet documents : a rich source for spoken language modeling, *ASRU'99 Workshop*, Keystone Colorado, pp. 277-280.

[Aus62] Austin, J.L. How To Do Things With Words, Oxford U.P., 1962.

[Sea72] Searle, J.R. Les Actes De Langage : Essai De Philosophie Du Langage (Trad. Française Par H. Pauchard). Paris Hermann, 1972.

[Col97] Colineau, N., Etude Des Marqueurs Discursifs Dans Le Dialogue Finalisé, Thèse De Doctorat De Sciences Cognitives, Université Grenoble 1, 1997.

[Fou01] Fouquet Y. (2001), De L'étude De Dialogues Oraux Dans Une Langue Non Maternelle, Rencontre Jeunes chercheurs en Parole, Mons, septembre 2001.

[Fou02] Fouquet Y. (2002), Une modèle de dialogue par les attentes du locuteur, TALN, Nancy, juin 2002.

[Bag] Bagein, M., Dutoit, T., Malfrere, F., Pagel, V., Ruelle, A., Tounsi, N., Wynsberghe, D. "An Open, Generic, Multi-lingual and Multi-platform Text-to-Speech System".