SIMDIAL

Un paradigme pour évaluer automatiquement des systèmes de dialogue homme-machine en simulant un utilisateur de façon déterministe

Joseph Allemandou[✢] — Laurent Charnay[❖] — Laurence Devillers[✦] Muriel Lauvergne[✢] — Joseph Mariani[✦]

[✢] *France Telecom R&D – 2, av. Pierre Marzin – 22300 LANNION*

[❖] *France Telecom – 27 rue Médéric – 75017 PARIS*

{joseph.allemandou,laurent.charnay,muriel.lauvergne}@orange-ftgroup.com

[✦] *LIMSI - CNRS (UPR 3251) – B.P. 133 - 91403 ORSAY CEDEX*

{devil,mariani}@limsi.fr

RÉSUMÉ. *Cet article présente* SIMDIAL, *un paradigme d'évaluation des comportements de Systèmes de Dialogue Homme-Machine (SDHM). Ce paradigme est fondé sur la simulation déterministe d'utilisateurs et permet une évaluation automatique ou semi automatique des SDHM face à des tâches précises. Il utilise le langage naturel pour interagir dynamiquement avec les systèmes évalués et est aisément transposable à différentes tâches de recherche d'information.*

ABSTRACT. *In this paper we present* SIMDIAL, *a paradigm to assess Spoken Language Dialogue Systems (SLDSs) using deterministic user simulation. The paradigm allows automatic or semi-automatic evaluation on different tasks and uses natural langage to interact with evaluated SLDSs.*

MOTS-CLÉS : *Paradigme d'évaluation, systèmes de dialogue homme-machine, simulation déterministe d'utilisateurs, évaluation automatique, évaluation semi-automatique.*

KEYWORDS: *Evaluation paradigm, dialog systems, deterministic user simulation, automatic evaluation, semi-automatic evaluation.*

1. Introduction

Les récentes campagnes et projets tels que DARPA MADCOW (Hirschman *et al.*, 1990), Aupelf B2 (Mariani, 1998), DARPA Communicator (Walker *et al.*, 2000), Technolangue MEDIA (Devillers *et al.*, 2004), témoignent de l'intérêt croissant pour l'évaluation des systèmes de dialogue et de leurs composants. Ce sujet est d'importance tant pour le monde industriel que pour la communauté de recherche sur le traitement automatique des langues. À l'heure actuelle il n'existe toujours pas de paradigme ni de mesures standard pour évaluer les systèmes de dialogue homme-machine.

L'évaluation des systèmes de dialogue est un sujet difficile par essence vu l'aspect dynamique et co-construit du dialogue. Il est difficile aussi pour des questions de généricité et de pouvoir diagnostique des paradigmes d'évaluation. Enfin, un système de dialogue finalisé ayant pour objectif premier de réaliser une tâche *via* une interaction avec un humain, l'étude des usages en situation de dialogue prend une place importante dans la problématique de l'évaluation de Systèmes de Dialogue Homme-Machine (SDHM) (Chaudiron, 2001).

Ce document présente SIMDIAL, un paradigme d'évaluation des comportements dialogiques de systèmes de dialogue homme-machine qui a pour objectif d'être le plus indépendant possible des systèmes à évaluer. Il permet la comparaison de l'efficacité des systèmes évalués, d'une part en terme de résolution de la tâche et d'autre part en terme de nombre de tours de parole. Pour ce faire un simulateur d'utilisateur interagit en langue naturelle avec les systèmes à évaluer, comme pourrait le faire un évaluateur humain. Ce simulateur a été réalisé de façon indépendante de tout SDHM existant pour éviter au maximum les biais associés. Il exploite différents modules constituants un nouveau système de dialogue à même de mener une interaction avec un SDHM à évaluer sur une tâche précise. Les critères permettant de juger les systèmes évalués sont, outre la complétude de réalisation de la tâche et les performances en terme de nombre de tours de parole, des paramètres mesurés à des fins de diagnostic, à savoir deux stratégies d'utilisateurs (directif ou non-directif) ainsi que deux phénomènes perturbateurs (hésitations et désambiguïsations). Ces critères sont observables automatiquement, ce qui permet de focaliser l'évaluation humaine sur les dialogues n'ayant pas correctement abouti.

Dans la suite de cette section seront analysées succinctement la problématique de l'évaluation de SDHM ainsi que les paradigmes existants. Les choix méthodologiques qui ont découlé de cette analyse et ont mené à la réalisation de ce travail seront ensuite exposés en section 2. Une présentation plus détaillée du paradigme SIMDIAL ainsi que les résultats de son application à une tâche de recherche de restaurants suivront en sections 3 et 4. Une autre évaluation dont les résultats détaillés ne sont pas présentés ici a été réalisée face à l'application DIALOGUEBOURSE, SDHM dont la tâche est la recherche d'informations boursières. Enfin les conclusions et perspectives seront exposées en section 5.

1.1. *Systèmes de dialogue homme-machine*

Le défi de voir une machine dialoguer avec un humain comme un humain a été proposé par Turing en 1950 (Turing, 1950). Le « premier » SDHM, proposé dans les années soixante par Weizenbaum, fut Eliza (Weizenbaum, 1966). Depuis, les progrès accomplis dans les différents domaines tant de la reconnaissance vocale que de la représentation du dialogue ou des connaissances déclaratives ont mené à des systèmes capables de réaliser des tâches *via* une interaction langagière avec un utilisateur. Par exemple, ARTIMIS (Sadek *et al.*, 1997) est une technologie de SDHM issue des laboratoires de France Télécom R&D fondée sur une théorie de l'interaction rationnelle (Sadek, 1991).

À l'heure actuelle il est courant de représenter un SDHM de façon modulaire en distinguant la reconnaissance vocale, l'interprétation sémantique, la gestion de l'interaction, la génération en langue naturelle et la synthèse vocale (voir figure 1).

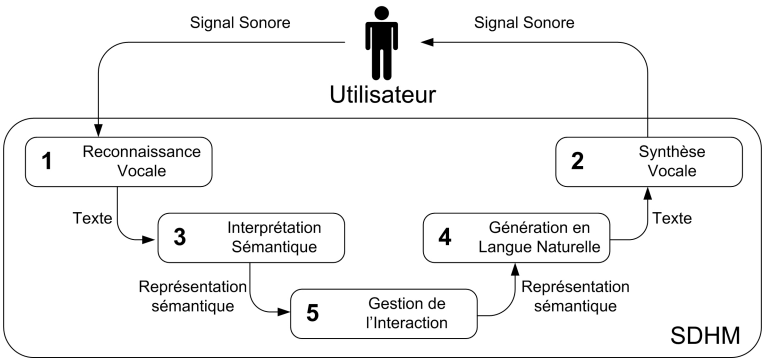

Figure 1. *Représentation modulaire standard d'un SDHM*

Le rôle du module de reconnaissance vocale (1) est de transcrire le signal de parole de l'utilisateur en une chaîne de caractères correspondant à ce que l'utilisateur a dit[1]. Le module de synthèse vocale (2) a le rôle symétrique, i.e. transformer une chaîne de caractères en signal sonore compréhensible par un utilisateur.

Le module d'interprétation sémantique (3) peut-être considéré comme un traducteur automatique. Il reçoit en entrée un énoncé en langue naturelle au format texte et doit fournir en sortie le même énoncé traduit dans un langage formel. Comme la synthèse vocale face à la reconnaissance vocale, le module de génération en langue naturelle (4) a la fonction symétrique du module d'interprétation sémantique, à savoir transformer un énoncé produit dans un formalisme en un énoncé en langue naturelle.

Le module de gestion de l'interaction (5) est la pièce centrale du SDHM. Il doit fournir le contenu de la réponse du système en fonction du contenu de l'énoncé de

1. Les transcriptions sont parfois enrichies d'annotations de phénomènes non verbaux (rire, souffle) ou d'informations prosodiques.

l'utilisateur déterminé par l'interprétation sémantique et du contexte d'interaction (énoncés déjà produits, connaissances disponibles etc).

1.2. *Un problème complexe : la généricité*

Aujourd'hui le nombre de formalismes de représentation, tant au niveau sémantique que dialogique, est approximativement aussi élevé que le nombre de SDHM. Des efforts sont faits pour créer des cadres suffisamment généraux pour permettre une homogénéisation de ces formalismes. Par exemple, le projet TrindiKit (Larsson et Traum, 2000), fondé sur la théorie de l'« Information State » (Traum et Larsson, 2003), propose un cadre outillé pour la création de gestionnaires d'interactions. Parallèlement le projet Technolangue MEDIA (Devillers *et al.*, 2003) propose d'évaluer l'interprétation sémantique d'énoncés à partir d'un formalisme de représentation sémantique commun. Il n'en reste pas moins vrai que cette diversité de formalismes est à l'origine des difficultés à trouver des paradigmes d'évaluation généraux (Devillers *et al.*, 2002).

Trois axes de recherches sont aujourd'hui explorés :

– des paradigmes indépendants du système qu'ils évaluent, notés génériques par rapport au système ;

– des paradigmes indépendants de la tâche que les systèmes évalués cherchent à résoudre, notés génériques par rapport à la tâche ;

– des paradigmes indépendants des deux et notés génériques par rapport au système et à la tâche.

Parmi les deux types de généricité dont il est question ci-dessus, la généricité par rapport au système est plus facile à obtenir que celle par rapport à la tâche.

Les difficultés pour être générique par rapport au système sont de l'ordre des formats d'entrées/sorties et de l'accord sur la définition de ce qui est évalué, ici un SDHM et/ou ses composants.

La généricité par rapport à la tâche objet du discours implique soit (1) de pouvoir évaluer sans avoir à représenter cette tâche, soit (2) d'avoir un formalisme de représentation des tâches qui permettent d'y représenter différentes tâches.

1) Ce premier cas est assez peu exploité car d'une part la tâche à réaliser joue un rôle prépondérant dans les dialogues (Walker, 1994) et que d'autre part même avec des critères qui ne seraient pas impactés par la tâche, l'évaluation nécessite l'observation et donc la réalisation d'une tâche.

2) La majorité des paradigmes qui se disent génériques par rapport à la tâche sont à classer dans cette catégorie du fait qu'il n'existe pas encore de formalisme de représentation d'une tâche qui soit standard dans la communauté. En fonction des paradigmes, leurs degrés d'expressivité permet de représenter plus ou moins de tâches à différents niveaux de précision.

1.3. *Évaluation de SDHM : aperçu de l'existant*

L'évaluation empirique de SDHM a naturellement débuté conjointement à la création de SDHM. Depuis une quinzaine d'années, une communauté de recherche s'est créée suite à la première campagne d'évaluation multi-système MADCOW (Hirschman *et al.*, 1990), qui avait pour objectif de fournir des résultats sur l'interprétation sémantique d'énoncés utilisateurs pris hors contexte. Cette communauté s'est intéressée tant à la définition de standards pour la caractérisation d'évaluations au travers des *cadres d'évaluation* EAGLES I et II (King *et al.*, 1996; Blasband *et al.*, 1999) et DISC (DISC Consortium, 1999), qu'à la création de paradigmes plus ou moins génériques pour évaluer des systèmes de dialogue et/ou leurs composants.

Parmi ces derniers certains évaluent automatiquement[2] l'interprétation sémantique, comme MADCOW (Hirschman *et al.*, 1990), DCR (Antoine et Caelen, 1999) ou plus récemment PEACE (Devillers *et al.*, 2003) utilisé notamment dans le cadre du projet européen TechnoLangue/MEDIA.

Concernant l'évaluation de la gestion de l'interaction, un type de paradigme met en jeu des utilisateurs réels qui mènent des interactions avec le système évalué et fournissent des informations concernant leur expérience avec ce système. PARADISE (Walker *et al.*, 1997) est le paradigme de ce type le plus connu, utilisé notamment pour la campagne d'évaluation multi-système COMMUNICATOR (Walker *et al.*, 2001). Les évaluations sont alors réalisées avec mesures. Une autre méthode consiste à observer des interactions où les utilisateurs sont simulés par un système. Dans ce cas les utilisateurs peuvent être simulés soit de façon stochastique (Eckert *et al.*, 1997; Scheffler et Young, 2000; Pietquin, 2004; Georgila *et al.*, 2005), soit de façon déterministe (Lin et Lee, 2001; López-Cózar *et al.*, 2003).

Du point de vue de la généricité :

– tous les paradigmes présentés ci-dessus sont génériques par rapport au système ;

– la généricité des paradigmes par rapport à la tâche dépend de la façon dont la tâche y est modélisée, car même les plus génériques nécessitent une représentation de la tâche à résoudre par le système qu'elles évaluent (ou sa contraposée, voir 3.2.1).

2. Choix méthodologiques

L'objectif initial de ce travail est d'évaluer de façon automatique et quantitative les comportements interactionnels de SDHM. Les analyses présentées ci-dessus montrent différentes façons de procéder. Cette partie expose les choix réalisés ainsi que leurs motivations pour atteindre l'objectif fixé.

2.1. *Boîte noire ou transparente ?*

Pour un besoin de généricité face aux systèmes évalués, une approche où le sys-

2. Sans interaction entre le système et un utilisateur réel.

tème est perçu en boîte noire a été choisie. L'évaluation étant focalisée sur la gestion de l'interaction, une approche en boîte transparente aurait nécessité pour observer le composant correspondant :

– soit que le formalisme de représentation sémantique de tous les systèmes évaluables soit identique, ce qui aujourd'hui est loin d'être le cas ;

– soit qu'un formalisme de représentation sémantique commun, sur lequel auraient été projetées les représentations sémantiques de chacun des systèmes évalués, soit défini comme proposé par le paradigme d'évaluation PEACE. Cette solution est envisageable mais présente le biais d'évaluer à la fois l'interprétation sémantique et la capacité de transformation des énoncés vers le formalisme commun.

2.2. *Utilisateurs réels ou simulés ?*

Les évaluations avec des utilisateurs réels sont plus coûteuses que des évaluations où l'utilisateur est simulé, tant en temps passé qu'en moyens à mettre en œuvre. En revanche les évaluations par simulation d'utilisateurs ne fournissent pas de résultats sur la satisfaction réelle des utilisateurs à se servir du système évalué.

Le choix a été fait de se placer dans le cadre de la simulation d'utilisateurs. Les évaluations par simulation d'utilisateurs peuvent permettre de réduire le nombre d'évaluations avec des utilisateurs réels, ce qui permet de réduire le coût de développement du système. Par ailleurs la simulation d'utilisateur n'a pas vocation à remplacer les évaluations face à des utilisateurs réels. En effet, à l'heure actuelle les comportements simulés ne sont pas aussi variables et précis que ceux d'utilisateurs réels, de même que les modèles de la satisfaction des utilisateurs ne sont pas représentatifs de celle obtenue par des sondage sur des utilisateurs réels. En revanche l'utilisation de simulateurs d'utilisateurs pour observer et évaluer des comportements et phénomènes prédéfinis est bien adaptée.

Reste alors à choisir entre simulation d'utilisateurs de façon stochastique ou de façon déterministe.

2.3. *Simulation stochastique ou déterministe ?*

Là encore les deux possibilités ont été explorées (voir 1.3). Une évaluation quantitative des méthodologies stochastiques a été proposée dans (Schatzmann *et al.*, 2005). L'argumentation qui suit s'inspire de cet article.

L'argument majeur de la méthode stochastique est qu'elle permet de simuler des comportements généralement plus proches de ceux d'un utilisateur réel que dans le cas d'une simulation déterministe. En revanche la simulation stochastique s'appuyant sur la notion de répétitivité[3], ne permet pas aisément de simuler des comportements peu ou pas observés dans les corpus d'apprentissage. De façon plus générale l'appren-

3. L'apprentissage automatique de comportements se base sur des comportements récurrents.

tissage de comportements étant fondé sur des corpus d'apprentissage, la représentativité de ces derniers face aux comportements de simulation souhaités joue un rôle très important. Enfin, bien que les récents progrès fassent augmenter la qualité des simulations, des métriques statistiques simples permettent toujours de différencier des dialogues réels de dialogues générés par simulation, comme, par exemple, l'uniformité des distributions du nombre de tours de parole pour résoudre une tâche.

À l'opposé, la méthode déterministe génère des comportements maîtrisés et potentiellement aussi précis que souhaités, mais il est nécessaire de les spécifier. La difficulté réside alors dans la modélisation des comportements souhaités. Il est par ailleurs intéressant de noter que cette méthodologie est bien adaptée pour l'observation de corrélations entre les stratégies de dialogue, les comportements de l'utilisateur simulé et les performances du système évalué. Enfin, cette méthode n'a pas pour objectif de reproduire précisément les comportements d'utilisateurs réels comme la simulation stochastique. Elle vise plus à générer des ensembles de comportements prédéfinis dans des cas précis.

Le choix de simuler les utilisateurs de façon déterministe est motivé notamment par le besoin d'expérimentation de certains types de comportements qu'il n'aurait pas été aisé (1) d'observer dans des corpus et (2) d'expliciter *via* un modèle dans ces mêmes corpus.

2.4. *Conclusion*

Le paradigme SIMDIAL propose d'évaluer un SDHM en interagissant dynamiquement avec lui grâce à un utilisateur simulé dont le comportement est défini par un modèle déterministe de l'interaction (voir 3.2). La maîtrise des situations proposées en entrée du système permet notamment la définition des comportements attendus[4] face à ces situations, tant à un niveau général qu'au niveau de ses composants.

3. SIMDIAL, un paradigme d'évaluation automatique

3.1. *Présentation générale*

Le paradigme SIMDIAL a pour objectif principal **l'évaluation automatique des comportements interactionnels d'un SDHM perçu en boîte noire en contexte de dialogue**. Il offre aussi la possibilité d'évaluer en boîte transparente les composants de reconnaissance vocale et d'interprétation sémantique par comparaison à des transcriptions/annotations humaines[5]. Enfin ce paradigme est plus novateur par la modélisation de l'utilisateur simulé qu'il utilise que pour l'approche par simulation déterministe d'utilisateur.

Une possibilité offerte par le simulateur et non encore exploitée dans la littérature

4. Il est aussi possible de définir les comportements non attendus.
5. Avec les problèmes de formalismes de représentation associés.

est l'évaluation semi-automatique. Lors d'une évaluation automatique par simulation d'utilisateur, le système évalué génère un énoncé et le simulateur lui répond. L'évaluation semi-automatique proposée ici est une évaluation automatique où les énoncés fournis en entrée du système évalué sont contrôlés par un humain[6] avant d'être proposés au système évalué. Cette approche permet notamment d'être sûr que les énoncés générés par le simulateur ressemblent à ceux d'utilisateurs réels.

L'architecture du simulateur est très semblable à celle d'un SDHM. Il est développé dans le langage de programmation Java pour plus de généricité face aux plates-formes sur lesquelles il doit être utilisé. Il est composé des modules classiques de SDHM pour l'interprétation sémantique , la gestion de l'interaction et la verbalisation des énoncés (voir les choix technologiques qui suivent). Du point de vue des résultats d'évaluation, le paradigme SIMDIAL dispose de mécanismes de diagnostic automatique qui identifient d'une part les dialogues n'ayant pas permis au simulateur de résoudre sa tâche et, d'autre part, les dialogues où ont eu lieu des désambiguisations ou des relaxations sur les critères de recherche (voir 3.4). Il est à noter que, pour plus de généricité, ces diagnostics sont réalisés sans utiliser de traces des systèmes évalués.

Enfin, ce paradigme est théoriquement **générique face aux systèmes évalués pour une tâche fixée**. En effet, le simulateur reçoit et génère des énoncés en langue naturelle et doit pouvoir s'interfacer avec n'importe quel SDHM[7], la définition de sa tâche à résoudre étant la contraposée de celle que cherchent à résoudre les systèmes évalués (voir 3.2.1).

Des choix technologiques concernant la mise en œuvre d'un tel simulateur sont présentés ci-dessous.

– **Formalisme de représentation du contenu des énoncés :** bien que n'étant pas un module à part entière, ce formalisme est à la base du fonctionnement du simulateur. En effet le gestionnaire d'interaction, qui est la pièce maîtresse du simulateur, reçoit en entrée et fournit en sortie des données dans ce formalisme. Le choix a été fait d'utiliser le formalisme de représentation de la campagne d'évaluation Technolangue/MEDIA pour représenter le contenu des énoncés[8]. Son pouvoir de représentation a été évalué dans l'article (Bonneau-Maynard *et al.*, 2005) où il a été démontré que ce formalisme permettait de représenter la tâche de réservation d'hôtels et de recherche touristique. Une description en est fournie en section 3.2.3.

– **Interprétation des énoncés du système évalué :** ces derniers étant générés par un système informatique, il est supposé que leur variabilité est suffisamment faible pour qu'il soit possible de les interpréter dans le formalisme précédemment cité. Un accord inter-annotateur élevé lors de l'annotation manuelle d'énoncés (voir 4.1) confirme l'homogénéité du sens associé aux annotations. La réalisation du module d'interprétation sera présentée dans la partie 3.3.1.

Le choix d'utiliser un module d'interprétation sémantique plutôt que d'exploiter

6. Une liste d'énoncés est proposée et l'évaluateur humain choisit, ou propose un nouvel énoncé correspondant au contenu.
7. Aux formats d'entrées/sorties près.
8. Au dictionnaire des concepts et valeurs près.

directement les représentations sémantiques du système évalué a pour fondement l'objectif de généricité face aux systèmes à évaluer. En effet, il est moins difficile d'annoter les quelques verbalisations types des systèmes à évaluer plutôt que de réaliser des traducteurs représentations sémantiques pour chacun des formalismes utilisés par ces systèmes.

– **Génération des énoncés du simulateur :** l'idée exploitée ici est qu'il est possible de générer des énoncés similaires voire même équivalents à ceux produits par des humains à partir d'un corpus annoté dans le formalisme déjà cité. Une stratégie de génération à partir de corpus a été développée dans ce but et sera présentée dans la partie 3.3.2. Enfin des tests perceptifs ont été réalisés pour contrôler la validité de cette hypothèse, voir 4.2.3. Là encore la verbalisation en langue naturelle vise la généricité face aux systèmes évalués.

– **Modélisation de l'interaction :** enfin, le paradigme SIMDIAL se veut suffisant pour permettre de juger de la capacité du système évalué à résoudre une tâche par le dialogue face à des utilisateurs dont le comportement est connu. Il n'a pas pour objectif de fournir des résultats d'évaluation similaires à ceux obtenus face à des utilisateurs réels. En effet, il serait illusoire de penser pouvoir simuler de façon à la fois suffisamment précise et variée des comportements d'utilisateurs réels, et, d'autre part, une évaluation étant toujours relative à une norme, même implicite, il est normal de ne pouvoir évaluer que face à des notions qui sont déjà modélisées.

À l'*évaluation minimale* proposée ci-dessus peuvent être ajoutés des phénomènes perturbateurs qui viennent bruiter l'interaction aux niveaux linguistique et dialogique. Ils permettent ainsi d'évaluer de façon diagnostique les comportements du système face à des situations non générables par le modèle de l'interaction tel qu'il a été défini (voir 3.2.5).

Ces choix technologiques visent à maximiser la généricité du paradigme SIMDIAL face aux systèmes à évaluer. L'évaluation des comportements dialogiques des systèmes est ainsi réalisée comme le ferait un évaluateur humain, en dialoguant avec le système. En effet, le simulateur instaure lui aussi avec le système évalué une véritable dynamique de dialogue.

3.2. *Le gestionnaire d'interaction*

Le gestionnaire d'interaction produit le contenu des énoncés à verbaliser. La génération de ce contenu dépend 1) du contenu des énoncés produits par le système évalué et 2) de l'état interne du gestionnaire d'interaction. Cet état interne exploite plusieurs modèles statiques, à savoir un modèle de la tâche, un modèle du dialogue, un formalisme de représentation du contenu des énoncés, un modèle de l'utilisateur simulé et un modèle de phénomènes perturbateurs.

3.2.1. *Modèle de la tâche*

Dans le cadre du dialogue finalisé, le « bon » déroulement d'une interaction est dépendant de la définition de la tâche à résoudre ainsi que de la réussite de cette dernière (Walker, 1994).

Le modèle de la tâche défini pour le simulateur représente la tâche contraposée de celle que cherche à résoudre le système évalué. La représentation de la tâche du simulateur joue donc un rôle prépondérant dans le « bon » déroulement des interactions, au sens de la référence face à laquelle le système est évalué. Les évaluations réalisées sont éminemment relatives à cette représentation.

Le formalisme choisi pour représenter la tâche est un modèle à base de *frames* (Minsky, 1975). Une tâche de recherche d'information y est représentée par :

– l'ensemble de buts à atteindre ;

– l'ensemble hiérarchisé[9] des critères de recherche utilisables ;

– l'ensemble des capacités de niveau méta comme par exemple la gestion des répétitions ou des listes de propositions.

La figure 2 présente la modélisation de la tâche contraposée à celle du système évalué en section 4, soit la recherche de restaurants à Paris.

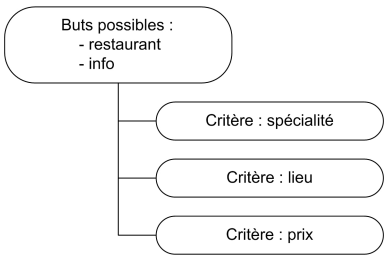

Figure 2. *Représentation d'une tâche de recherche de restaurants à Paris*

3.2.2. *Modèle du dialogue*

Il s'inspire de la philosophie du langage et notamment des actes de langage d'Austin (Austin, 1969) et Searle (Searle, 1972). Les actes utilisés sont les six actes principaux extraits du formalisme FIPA ACL (FIPA Consortium, 2002) par le consortium MEDIA. Cette sous-liste a été extraite pour des raisons d'accord inter-annotateurs, la liste complète ne permettant pas des annotations suffisamment homogènes. Le tableau qui suit présente ces six actes ainsi que leur signification.

9. La notion de hiérarchisation s'applique ici dans la définition de sous-critères et non dans l'ordonnancement des critères.

Inform(X)	Acte assertant l'information X
Request(X)	Acte de requête de l'information X[10]
Accept	Acceptation de l'assertion de l'énoncé précédent
Reject	Refus de l'assertion de l'énoncé précédent
Open	Acte d'ouverture de dialogue
Close	Acte de fermeture de dialogue

Tableau 1. *Liste des six actes de dialogue*

3.2.3. *Formalisme de représentation du contenu des énoncés*

Fondé sur celui défini pour la campagne d'évaluation MEDIA de l'action interministérielle Technolangue (Devillers *et al.*, 2004), il est structuré sur deux niveaux.

Le premier représente le niveau dialogique des énoncés de l'interaction. Il utilise les actes de dialogue définis précédemment dans le modèle du dialogue (voir 3.2.2).

Le second définit le contenu sémantique relatif à la tâche (voir 3.2.1). Chaque information est représentée par un triplet contenant un mode, un attribut et une valeur. Le mode définit si l'information est affirmée ou niée et l'attribut fournit le type de l'information relativement à la tâche. La valeur est, comme son nom l'indique, le contenu informationnel véhiculé par la portion d'énoncé, par exemple le nom d'une ville ou un horaire. C'est le modèle de la tâche qui définit la précision de la représentation et donc des valeurs, par exemple un horaire peut être vu comme une heure et des minutes agrégées ou non. Le contenu d'un énoncé est alors une liste d'actes de dialogue à chacun desquels est associée une liste de triplets sémantiques.

Un exemple d'un énoncé dont le contenu est annoté dans le formalisme de représentation est présenté ci-dessous.

Auriez-vous un restaurant antillais pour moins de vingt euros ?
Request (< + / spécialité / antillais >, < + / prix inférieur / 20 >)

3.2.4. *Modèle de l'utilisateur simulé*

Le modèle présenté dans cette partie ne doit pas être confondu avec le « modèle utilisateur » tel qu'il est généralement entendu. Ce dernier est d'après Fisher (Fischer, 2001) le modèle que le système a d'un utilisateur, qu'il soit obtenu *a priori* par spécification du concepteur ou dynamiquement par inférences du système. Il est alors un moyen d'être plus sensible aux besoins d'un utilisateur et aux caractéristiques d'une situation donnée.

Le modèle de l'utilisateur simulé décrit ici un ensemble de paramètres qui permettent de faire varier le comportement du simulateur autour de la tâche prédéfinie. La liste qui suit présente ces différents paramètres.

10. La notion de requête est ici rapportée à l'acte de surface et non à la pragmatique du discours.

– **Nombre d'informations relatives à la tâche par tour de parole :** c'est le nombre maximum d'informations que le simulateur fournit en un énoncé.

– **Ordre de présentation des critères :** ordre dans lequel le simulateur verbalisera ses critères s'il n'est pas contraint par sa stratégie et une question du système, voir ci-après.

– **Ordre dans lequel les critères doivent être relâchés :** en cas de requêtes sans solution, le simulateur essaie de relâcher un des critères de recherche, *i.e.* de le transformer en un autre. Si la stratégie du simulateur est directive (voir ci-dessous), le critère à relâcher est défini par l'ordre dont il est question ici, sinon le premier critère proposé par le système évalué est accepté.

– **Utilisateur simulé directif ou non directif :** si sa stratégie est directive, le simulateur ne cherche pas à répondre aux propositions du système évalué, il prend l'initiative dans le discours à tous les tours de parole pour résoudre la tâche selon ses paramètres. Dans le cas de la stratégie non directive, le simulateur cherche au contraire à répondre aux questions du système[11] à tous les tours de parole.

3.2.5. *Phénomènes perturbateurs*

L'ajout de phénomènes perturbateurs à une interaction est équivalent à un bruitage du comportement du simulateur dans le but d'obtenir des informations prédictives sur les phénomènes considérés, comme les hésitations ou les ambiguïtés (voir ci-après).

Il aurait été possible d'ajouter la notion de phénomène perturbateur aux modèles déjà définis. En effet les phénomènes perturbateurs impactent le comportement du simulateur comme les modèles déjà présentés. L'intérêt de les distinguer réside dans le fait de n'avoir, dans les modèles déjà définis, que des paramètres dont la définition est nécessaire pour générer une interaction. Les phénomènes perturbateurs ne sont pas nécessaires à la génération d'une interaction, ils étendent le pouvoir diagnostique[12] du paradigme SIMDIAL. De plus cette distinction permet de ne pas modifier les différents modèles lors de l'ajout d'un phénomène perturbateur, ce qui rend l'approche plus générique face à ces derniers.

Deux phénomènes perturbateurs ont été modélisés pour valider l'approche. L'un est de niveau linguistique et permet l'ajout d'hésitations aux verbalisations du simulateur, l'autre de niveau dialogique vise la génération d'ambiguïtés sur les critères de recherche. Ce dernier phénomène est observable lorsque deux valeurs sont verbalisées pour un même critère lors d'un dialogue. Il peut être intra-énoncé lorsque les deux valeurs sont présentes dans le même énoncé ou inter-énoncé lorsque les deux valeurs apparaissent à des tours de parole différents. Par exemple l'énoncé *« Je veux un restaurant chinois, un japonais »* contient une ambiguïté intra-énoncé.

L'ajout d'hésitations aux énoncés verbalisés par le simulateur est réalisé en ajoutant au contenu de l'énoncé un concept spécifique pour représenter l'hésitation. Dès lors la stratégie de verbalisation générera un énoncé contenant une hésitation, comme,

11. Sans pour autant se répéter.
12. Au sens prédictif.

par exemple, *« Pour moins de **euh** dix euros »*. Cette méthode simple a pour avantage d'être applicable à la majorité des disfluences linguistiques observables dans les dialogues.

Les ambiguïtés sur les critères de recherches sont modélisées en ajoutant un segment sémantique au contenu d'un énoncé. Pour une ambiguïté intra-énoncé le segment ajouté porte sur le même concept qu'un de ceux déjà verbalisés, mais véhicule une valeur différente de celle normalement produite. Pour une ambiguïté inter-énoncé, le segment ajouté porte sur un des critère non déjà verbalisé dans l'énoncé courant, et véhicule, là encore, une valeur différente de celle normalement attendue dans le scénario.

3.2.6. *Algorithme de génération du contenu de la réponse du simulateur*

Cet algorithme exploite les modèles décrits précédemment et le contenu de l'énoncé produit par le système évalué. Il dispose, pour chacun des critères de recherche et des buts du modèle de l'utilisateur simulé, de variables décrivant si les critères ou les buts ont déjà été énoncés et s'ils ont été compris. Le calcul du contenu de la réponse se déroule alors comme présenté par la figure 3.

Sur ce schéma les entités représentées par des rectangles sont des procédures et celles représentées par des ellipses les résultats possibles des procédures. Les rectangles grisés sont des procédures terminales alors que les blancs sont des procédures intermédiaires.

Une précision concernant l'algorithme du simulateur est qu'il dispose d'un mécanisme mettant fin au dialogue après un nombre prédéfini de tours de parole. Ce mécanisme a pour objet de mettre fin aux boucles d'interaction infinies et de restreindre la résolution de la tâche en nombre de tours de parole. Il est aussi impliqué dans le diagnostic automatique d'aboutissement des dialogues, voir 3.4.

3.3. *Les autres modules*

3.3.1. *Interprétation sémantique*

Ce module transforme les énoncés du système évalué en leur contenu dialogique et sémantique dans le formalisme décrit en 3.2.3. Cette transformation est réalisée par deux méthodes appliquées successivement et nécessite un thésaurus des concepts associés à la tâche.

Un premier niveau d'analyse est basé sur le principe d'îlots sémantiques et d'encrage sur les concepts de la tâche. Il exploite le thésaurus en associant par égalité textuelle des portions d'énoncés du SDHM évalué à des concepts relatifs à la tâche. Des améliorations pourraient y être apportées, notamment en incluant des variantes flexionnelles.

La deuxième passe associe des actes de dialogue aux portions d'énoncés. Cette association a lieu par analyse de la ponctuation et de marqueurs linguistiques. Par

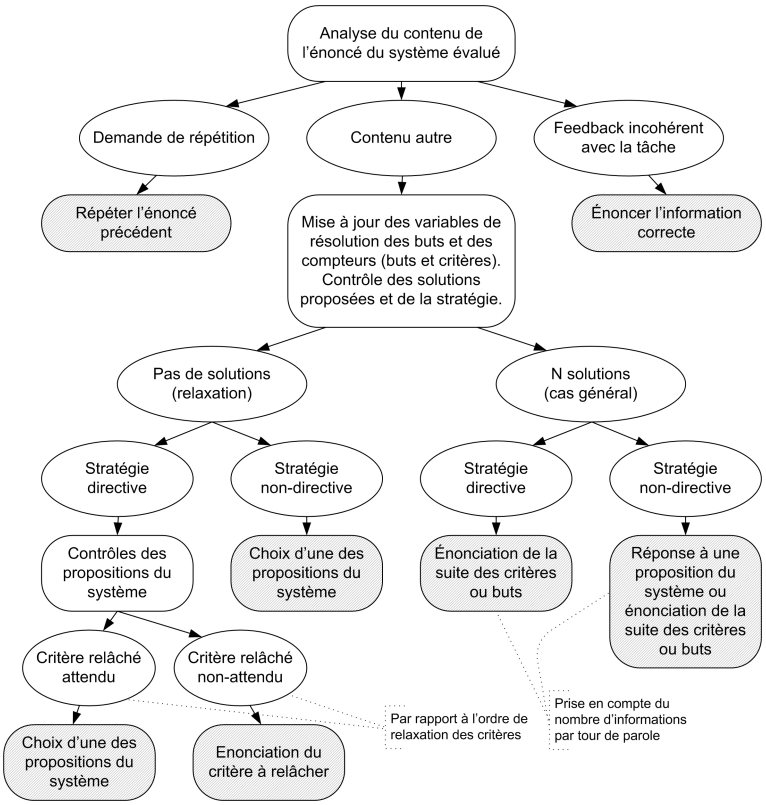

Figure 3. *Schéma de représentation de l'algorithme de génération du contenu de la réponse*

exemple, une portion d'énoncé clôturée par un point (resp. un point d'interrogation) est considérée comme une assertion (resp. une requête).

3.3.2. *Verbalisation des énoncés*

Ce module transforme les énoncés produits par le module de gestion de l'interaction du format de représentation du contenu vers des énoncés en langue naturelle. Il permet une génération d'énoncé totalement automatique ou semi-automatique en proposant une liste d'énoncés à un évaluateur humain. Ce dernier choisira alors dans la liste ou entrera l'énoncé à la main, voir ci-après.

– **Stratégie automatique :**le passage d'un énoncé du formalisme de représentation vers la langue naturelle est réalisé par appariement de contenu sémantique et/ou dialogique avec d'une part des énoncés ou des portions d'énoncés d'utilisateurs annotés et, d'autre part, un thésaurus des concepts de la tâche. Cette méthode requiert

donc un corpus annoté ainsi qu'un thésaurus de la tâche à partir desquels les énoncés du simulateur seront générés.

L'algorithme de génération de l'énoncé exploite une succession de recherches sur les annotations des énoncés du corpus et du thésaurus. Si une recherche ne donne pas de résultat, la suivante est essayée. À chaque recherche, les contraintes sur l'ensemble des portions d'énoncés recherchées sont diminuées, d'où un plus grand choix à chaque étape. Lorsque tout le contenu de l'énoncé a été trouvé, sa verbalisation est reconstruite.

Cette stratégie permet notamment de générer des variations de verbalisations pour des contenus identiques, d'où un intérêt pour l'évaluation de l'interprétation sémantique, bien que ce ne soit pas le sujet traité dans cet article.

– **Stratégie semi-automatique :** elle est la même que la stratégie automatique si ce n'est qu'au lieu de choisir une verbalisation parmi les possibilités, elle les garde toutes. Les énoncés reconstruits à partir des portions de contenus sont ensuite proposés à un évaluateur humain. Ce dernier a alors le choix d'entrer un nouvel énoncé ou d'en prendre un dans la liste proposée par le simulateur. Une fois le choix réalisé, l'énoncé est envoyé au système évalué.

Un point à ajouter concernant ces stratégies de génération d'énoncés est qu'elles permettent de générer des énoncés pour toutes les valeurs des différents attributs à condition de disposer d'une façon de les verbaliser. L'utilisation du thésaurus de concepts relatifs à la tâche pour verbaliser les portions d'énoncés dont le contenu n'est pas présent dans le corpus annoté permet ainsi de produire toutes les valeurs des différents attributs définis dans le thésaurus. Il est alors possible de « doper » les corpus, par exemple en testant le système sur toutes les valeurs de tous les attributs disponibles. Le dopage est ici considéré comme un moyen artificiel de faire augmenter la complétude d'un corpus par rapport à son domaine d'application.

3.4. *Diagnostic automatique*

Le diagnostic automatique permet d'obtenir des informations sur les dialogues auxquels a pris part le simulateur. Deux phénomènes sont diagnostiqués par ce module, à savoir l'aboutissement des dialogues et la relaxation de critères de recherches. Il est à noter que pour des raisons de généricité face aux systèmes évalués, les diagnostics réalisés n'exploitent pas de fichiers de traces générés par ces derniers.

– **Abouti vs. non abouti :** un dialogue est dit abouti quand le simulateur a réussi à réaliser sa tâche ou, en d'autres termes, à atteindre tous les buts de sa liste de buts. S'il n'y arrive pas, le dialogue est dit non abouti. La méthode de diagnostic de l'aboutissement repose sur le fait que le simulateur ne met fin à l'interaction que s'il a réussi à réaliser sa tâche ou si le nombre de tours de parole qu'il a dû générer dépasse un seuil (voir 3.2.6). Un dialogue abouti est donc un dialogue dont le nombre de tours de parole est inférieur au nombre de tours de parole seuil.

– **Relaxation de critères :** comme expliqué dans la partie 3.2.4, le simulateur permet qu'un de ses critères soit transformé en un autre en cas de requêtes sans solution. Ce phénomène nommé relaxation de critère est diagnostiqué grâce à la différence de comportements qu'il induit pour le simulateur. Cette mesure du nombre de tours de parole où le simulateur relâche des critères permet d'expliquer certaines différences en terme de nombre de tours de parole pour résoudre la tâche.

– **Désambiguïsation :** la désambiguïsation de critères de recherche correspond à une demande de choix parmi un ensemble de valeurs proposées par le système évalué. Par exemple, l'énoncé « *Voulez-vous un restaurant chinois ou japonais ?* » est considéré comme une désambiguïsation. Le diagnostic de ce phénomène est intéressant face à la génération d'ambiguïtés sur les critères de recherche en tant que phénomènes perturbateurs. En effet, ces ambiguïtés telles qu'elles sont modélisées peuvent mener à différentes stratégies de résolution de la part du système. Le fait de diagnostiquer les désambiguïsations permet d'analyser la ou les différentes stratégies mises en œuvre par le système face à la génération d'une ambiguïté.

4. Évaluation de l'application PLANRESTO

Afin de valider expérimentalement le paradigme SIMDIAL, un démonstrateur qui évalue les système de recherche d'information PLANRESTOa été réalisé. Ce système de dialogue, construit avec la technologie ARTIMIS, permet de chercher un restaurant à Paris. Un thésaurus de concepts relatif à la tâche était disponible ainsi que des corpus de dialogues réalisés avec ce système. La technologie ARTIMIS visant un dialogue naturel entre le système et l'utilisateur, les corpus exploités contenaient peu d'énoncés où les utilisateurs étaient contraints par le système. Cette évaluation a porté sur les comportements dialogiques du système, mais pas sur son composant d'interprétation sémantique.

4.1. *Corpus d'interprétation et de verbalisation*

Pour l'interprétation des énoncés de l'application PLANRESTO, un thésaurus relatif à la tâche de recherche de restaurants comprenant 6 302 valeurs est disponible.

Une étape de validation de l'interprétation sémantique a été menée par une double annotation. Cette dernière a été réalisée sur 10 dialogues, soit 66 énoncés contenant 67 annotations dialogiques et 87 annotations sémantiques. Les Kappas de Cohen (Cohen, 1960) obtenus sont de 0,88 (resp. 0,84) pour les annotations dialogiques (resp. sémantiques), ce qui valide l'homogénéité des annotations du corpus relativement au formalisme d'annotation.

Pour la verbalisation d'énoncés par le simulateur, un corpus de 110 dialogues, dont seuls les énoncés utilisateurs sont annotés, est utilisé. Ces énoncés bien que collectés à partir de l'application PLANRESTO sont représentatifs de la tâche de recherche de restaurants en ce que la technologie ARTIMIS permet des interactions naturelles à

initiative mixte entre utilisateurs et systèmes.

Le nombre d'énoncés utilisateurs est de 790, soit une moyenne d'environ 7 énoncés par dialogue. À ces 790 énoncés sont associées 1 757 annotations dont 800 dialogiques et 957 sémantiques.

4.2. *Résultats*

L'évaluation a été effectuée en observant et en analysant des dialogues entre le simulateur et PLANRESTO.

4.2.1. *Exemple de dialogue généré*

La figure 4 est un dialogue ayant eu lieu entre le système PLANRESTO et le simulateur d'utilisateur. Il est comparable à notre avis à un dialogue entre un utilisateur réel et le système tant du point de vue de la résolution de la tâche que des verbalisations.

PlanResto :	Bienvenue sur PlanResto. Ce service vous permet de trouver un restaurant à Paris et d'être mis en relation téléphonique avec lui. Vous pouvez à tout moment obtenir une aide en disant *mode d'emploi général*. Que désirez-vous ?
Simulateur :	Je cherche un restaurant dans le cinquième arrondissement.
PlanResto :	Vous recherchez un restaurant dans le 5e arrondissement. J'en ai trouvé plus de 10. Vous pouvez donner plus de précisions, par exemple la spécialité, consulter les solutions ou effectuer une nouvelle recherche.
Simulateur :	Je recherche un restaurant antillais.
PlanResto :	J'ai trouvé deux restaurants antillais dans le 5e arrondissement. Le premier restaurant est le restaurant « La Barbacane ». Souhaitez-vous plus d'informations, le restaurant suivant ou effectuer une nouvelle recherche ?
Simulateur :	Plus d'informations.
PlanResto :	Le restaurant « La Barbacane » est situé 13, rue Cardinal-Lemoine. Son numéro de téléphone est le 01 43 26 37 01. Voulez-vous d'autres informations, être mis en relation avec ce restaurant, le restaurant suivant ou effectuer une nouvelle recherche ?
Simulateur :	Quitter.
PlanResto :	Voulez-vous réellement quitter PlanResto ?
Simulateur :	Oui.

Figure 4. *Exemple de dialogue entre* PLANRESTO *et le simulateur*

4.2.2. *Observations*

Lors de l'évaluation, les comportements du simulateur variaient d'un ensemble de dialogues à l'autre. Ces variations portaient sur :

– le nombre d'informations relatives à la tâche par énoncé ;

– la stratégie directive ou non de l'utilisateur simulé ;

– l'ajout d'hésitations et d'ambiguïtés en tant que phénomènes perturbateurs.

L'observation des dialogues auxquels a pris part le simulateur a porté sur quatre observables objectifs :

– l'aboutissement ou non des dialogues ;

– le nombre de tours de parole ;

– les relaxations de critères de recherche ;

– les désambiguïsations.

Enfin, une étape de validation de la qualité des énoncés produits par le simulateur a été menée par des tests perceptifs. Ont été présentés aux sujets 50 énoncés, dont 25 étaient produits par la stratégie de verbalisation et 25 étaient des transcriptions automatiques d'énoncés humains. L'ordre de présentation des énoncés était aléatoire et pour chacun d'eux il était demandé si l'énoncé semblait naturel (produit par un humain) ou non naturel (produit par une machine), si le sens des énoncés était compréhensible et si la syntaxe des énoncés paraissait correcte ou incorrecte. Ce test, dont les résultats sont présentés dans la section 4.2.3, a été présenté à 25 sujets, dont 7 femmes et 18 hommes. L'âge des participants était majoritairement (70 %) compris entre 20 et 30 ans et tous disaient utiliser un ordinateur plusieurs fois par semaine.

Dans le tableau 2 les termes « dialogues », « tours de parole » « relaxation de critère », « désambiguïsation », « hésitation » et « ambiguïté » ont été abrégés respectivement en « Dial », « TdP » « relax », « désamb », « H » et « A ».

			Total	**Aboutis**	**Avec relax.**	**Avec désamb.**
	H	**A**	Dial – TdP	Dial – TdP	Dial – TdP	Dial – TdP
Directif	–	–	300 –1 901	277 –1 533	122 – 232	64 – 83
	✔	–	300 –1 965	278 –1 613	125 – 274	65 – 87
	–	✔	300 –2 432	240 –1 472	114 – 217	103 – 134
	✔	✔	300 –2 155	261 –1 531	119 – 208	106 – 128
Non dir.	–	–	300 –1 745	278 –1 393	118 – 144	43 – 49
	✔	–	300 –1 617	286 –1 393	112 – 143	41 – 45
	–	✔	300 –2 182	250 –1 382	116 – 164	84 – 102
	✔	✔	300 –2 176	243 –1 264	114 – 141	70 – 78
Total			2 400–16 173	2 113–11 581	940 –1 523	576 – 706

Tableau 2. *Dialogues générés et tours de parole associés*

4.2.3. *Analyses*

1) **Sur l'aboutissement des dialogues :** le premier résultat est que 2 113 dialogues sur 2 400 ont abouti, soit 88 %. Le diagnostic automatique d'aboutissement des dialogues permet notamment de n'analyser manuellement que les dialogues n'ayant pas abouti, soit 287 par rapport à 2 400. Un contrôle manuel effectué sur 100 dialogues a permis de valider la qualité du diagnostic d'aboutissement puisqu'aucun faux-positif ni faux-négatif n'a été relevé. Un faux-positif est ici un dialogue qui aurait été classé comme ayant abouti alors que la tâche n'aurait pas été réalisée et un faux-négatif un dialogue qui aurait été classé comme non abouti alors que la tâche y aurait été réalisée.

L'analyse manuelle des dialogues non aboutis permet de distinguer deux types d'erreurs. L'un est relatif à la gestion de l'initiative mixte dans l'application PLANRESTO et l'autre est dû à des erreurs d'interprétation des énoncés du système évalué. Le premier représente environ 40 % des dialogues non aboutis et le second 60 %, soit environ 170 dialogues sur 2 400.

L'erreur de gestion d'initiative mixte a lieu lors de la relaxation de critère quand le comportement du simulateur est directif. Le simulateur cherche dans ce cas à imposer le critère à relâcher tandis que l'application PLANRESTO propose des solutions avec un autre relâchement de critère. L'application PLANRESTO attend une réponse à ses propositions et ne comprend pas que le simulateur lui demande autre chose.

2) **Sur l'efficacité en fonction de la stratégie :** Il apparaît que le nombre de tours de parole par dialogue abouti est plus faible de 11,68 % lorsque la stratégie de l'utilisateur simulé est non directive, soit 5,82 tours de parole en moyene pour la stratégie directive et 5,14 pour la stratégie non directive. Bien que faible, ce résultat est observable sur l'ensemble des dialogues, quel que soit le nombre d'informations fournies par tour de parole et que les dialogues contiennent des relaxations ou pas. Un point intéressant à noter est que les résultats obtenus en évaluant l'application DIALOGUEBOURSE sont opposés à ceux présentés ci dessus, c'est-à-dire que la stratégie directive du simulateur a obtenu de meilleurs résultats en terme d'efficacité que la stratégie non directive. Ce résultat contre-intuitif est dû au fait que l'application DIALOGUEBOURSE ne relâche pas de contraintes sur les critères de recherche (voir plus bas). Il explicite bien le besoin de contrôle des comportements des systèmes face aux stratégies des utilisateurs.

Il est aussi intéressant de préciser que la différence entre stratégie directive et non directive sur le nombre de tours de parole contenant des relaxations est de 33,50%, ce qui est beaucoup plus significatif. Enfin, la corrélation de ces deux données montre qu'environ 50 % des 5 % de différence vus plus haut sont dus à la relaxation de critère. En d'autres termes la moitié de la différence du nombre de tours de parole entre un utilisateur directif et non directif est expliquée par la relaxation de critère.

L'interprétation de ces résultats est intuitivement aisée. En effet, un utilisateur directif impose son critère à relâcher et il y a assez peu de chance pour que ce dernier soit celui proposé par le système. S'ensuivent alors des tours de parole pour la désambiguïsation du critère choisi par le simulateur, ce qui peut même parfois mener à d'autres

relaxations. Un utilisateur non directif se laissant guider, il accepte directement une des solutions qui lui sont proposées, sans besoin de tours de parole supplémentaires.

La faible différence observée entre les stratégies sur le nombre de tours de parole montre que l'application PLANRESTO gère plutôt bien l'initiative mixte et tolère généralement que l'utilisateur simulé ne réponde pas à ses propositions.

3) **Sur l'ambiguïté et l'hésitation :** les résultats obtenus sur l'hésitation démontrent que l'application PLANRESTO est robuste face à ce phénomène. La différence du nombre de tours de parole pour les dialogues auxquels ont été ajoutées des hésitations par rapport à ceux auxquels il n'en a pas été ajouté est de moins de 2 %, soit 5,43 tours de parole pour les dialogues ne contenant pas d'hésitations et 5,53 pour ceux qui en contenaient.

Sur l'ambiguïté, les résultats sont plus intéressants. Il est notable que le nombre de dialogues contenant des désambiguïsations, pour les dialogues où étaient générées des ambiguïtés, n'est que de 442 dialogues sur 1 200. Ce nombre représente plus du double du nombre de dialogues contenant des désambiguïsations sans ajout d'ambiguïtés, mais reste faible cependant. L'analyse manuelle de dialogues auxquels avaient été ajoutées des ambiguïtés permet de fournir deux explications au fait que peu de désambiguïsations aient été générées. La première est que l'application PLANRESTO permet de demander plusieurs spécialités pour une même recherche, ce qui transforme certaines des ambiguïtés générées sur le critère spécialité en requêtes sur plusieurs spécialités. La seconde explication provient du fait que l'application PLANRESTO n'applique pas toujours les mêmes stratégies en fonction des critères ambigus et du fait que l'ambiguité soit intra- ou inter-énoncé.

4) **Sur la vraisemblance des verbalisations du simulateur :** les résultats du test perceptif présenté ci-dessus montrent que la différence entre les énoncés produits par la stratégie de verbalisation et les transcriptions automatiques d'énoncés humains n'est pas perceptible, tant du point de vue de l'aspect naturel ou pas, que de la correction du sens ou de la syntaxe. En effet, que les énoncés aient réellement été produits par des humains ou par la stratégie de verbalisation, les réponses concernant l'aspect, le sens et la syntaxe sont identiques à moins de cinq points près.

De plus, des accords inter-annotateur ont été calculés en appliquant le Kappa de Fleiss (Fleiss, 1971) aux résultats sur l'aspect et la syntaxe. Les coefficients de 0,25 pour l'aspect et 0,52 pour la syntaxe peuvent être interprétés d'après (Landis et Koch, 1977) comme « passable » et « bon », ce qui valide les résultats sur ces deux points. Pour le sens des énoncés, la cohérence des résultats a été mesurée à l'aide de l'Alpha de Cronbach (Cronbach, 1951), puisque l'échelle des valeurs possibles était continue. Là encore, la cohérence du résultat est confirmée par un coefficient de 0,95.

5. Conclusions

Le travail présenté dans ce document a pour vocation d'être utilisé tant par la communauté de recherche sur le traitement automatique des langues naturelles que par le monde industriel.

Sur les méthodes et modèles employés pour la réalisation du paradigme SIMDIAL, le premier point à noter est que l'approche utilisée pour modéliser les comportements dialogiques du simulateur est déterministe, autrement dit définie par expertise. Comme avec une approche stochastique, l'approche choisie permet d'évaluer les systèmes dans les cas d'utilisation standard, pour peu qu'ils aient été définis. Ce qu'elle permet en plus est l'évaluation face à des comportements marginaux peu représentés dans les corpus, notamment grâce à la notion de phénomène perturbateur. Enfin, elle est plus aisément transposable à d'autres tâches, voire d'autres modalités d'interaction de par sa conception basée sur des modèles (voir plus bas).

Concernant la verbalisation des énoncés du simulateur, l'approche développée est hybride : par expertise et avec corpus. L'avantage majeur de cette approche est de fournir de bons résultats de verbalisations sans nécessiter de gros corpus. En effet, le premier corpus de tests ne contenait que vingt-huit dialogues et les énoncés générés étaient déjà de bonne qualité. Enfin, la scission entre verbalisation et choix du contenu sémantique de la réponse autorise la modification des contenus indépendamment de la verbalisation, ce qui permet de réaliser des évaluations semi-automatiques, d'ajouter des phénomènes perturbateurs, et trouve un intérêt particulier pour l'évaluation automatique des composants d'interprétation sémantique.

Enfin, sur la modélisation des phénomènes perturbateurs, le fait de les décorréler de l'algorithme de génération de la réponse semble un bon compromis entre, d'une part la possibilité de générer « tous » les phénomènes perturbateurs, et d'autre part la difficulté d'adaptation des comportements du simulateur à ces phénomènes. Il est clair que les modèles utilisés pour le simulateur ne permettent pas de représenter correctement tous les phénomènes perturbateurs. Par exemple, la définition d'un critère de recherche par négation, comme dans l'énoncé « *Je veux un restaurant mais pas un japonais* », n'est pas réalisable sans modifier les modèles de la tâche et du dialogue. En revanche l'ajout des phénomènes perturbateurs de l'oral spontané définis par le GDR-I3 (Antoine *et al.*, 2002) est aisée, ainsi que des phénomènes d'ordre dialogique comme par exemple la génération de contestations ou de demandes de reformulation.

Concernant les propriétés du paradigme SIMDIAL, l'idée de simuler un utilisateur pour obtenir des corpus de dialogues ouvre des perspectives quant aux méthodes de développement des SDHM. En effet, l'approche proposée permet théoriquement de comparer différents systèmes sur une même tâche en les mettant dans les mêmes conditions contrôlées : scénarios, phénomènes perturbateur, critères d'évaluation. L'évaluateur humain a alors toute latitude de mettre en œuvre sa propre métrique d'évaluation comparative.

Par ailleurs, il devient possible d'accéder à de gros ensembles de données générées à partir du SDHM en développement. Évidemment les dialogues obtenus ne sont pas de « vrais » dialogues homme-machine et les corpus générés sont principa-

lement utilisables pour des tests. Un apprentissage sur ces données amènerait le biais de n'apprendre que les stratégies spécifiées de façon déterministe. En revanche ces corpus permettent d'obtenir des données à observer et analyser sans avoir à réaliser de tests avec des utilisateurs réels. En d'autres termes cette méthode donne la possibilité d'automatiser certaines étapes de tests du cycle de développement d'un SDHM. Il est clair que des résultats relatifs aux utilisateurs comme la satisfaction ou l'utilisabilité ne peuvent être obtenus avec cette méthode.

Toujours dans l'idée de réduire le coût de la phase d'évaluation d'un SDHM, le diagnostic automatique de phénomènes restreint significativement les analyses manuelles. Les tests réalisés avec l'application PLANRESTO ont permis de diagnostiquer le non-aboutissement et ainsi de n'analyser manuellement que 217 dialogues sur 2 400. Ont aussi été diagnostiqués les dialogues et tours de parole contenant des relaxations de critères et des désambiguisations, offrant ainsi la possibilité d'obtenir des informations d'ordre prédictif sur ces phénomènes.

Une troisième capacité du paradigme SIMDIAL est l'observation de corrélations entre mesures de performance sur les dialogues, stratégies d'utilisateurs simulés et stratégies des systèmes. Comme observé dans (Schatzmann *et al.*, 2005), cette propriété est générale à toutes les méthodes d'évaluation par simulation déterministe d'utilisateurs. Les limites de cette propriété sont celles imposées par les modèles utilisés, tant au niveau des stratégies d'utilisateurs simulés que des systèmes évalués. En d'autres termes, il n'est possible d'obtenir d'informations que sur les stratégies modélisées.

La dernière propriété à noter est que le paradigme SIMDIAL est modifiable à faible coût pour être appliqué à d'autres tâches. En effet, le simulateur est fondé sur l'analyse de capacités généralement admises pour les SDHM et non précisément celles de SDHM en particulier. Le transfert du simulateur de l'application PLANRESTO vers DIALOGUEBOURSE a été effectué en onze jours par une personne, soit une journée d'annotation pour vingt dialogues, quatre pour modifier les modules d'interprétation et de verbalisation, trois pour implémenter la tâche et trois pour affiner le simulateur dans son ensemble.

Le paradigme SIMDIAL permet une évaluation automatique des comportements dialogiques de systèmes de dialogue. Les critères d'évaluation exploités sont la résolution de la tâche et la durée des dialogues en tours de parole, associés à différentes stratégies et phénomènes perturbateurs à des fins de diagnostic. Par ailleurs, le paradigme utilise le langage naturel pour interagir avec les systèmes évalués, ce qui le rend théoriquement interfaçable avec différents SDHM sur une tâche précise et permet ainsi des évaluations comparatives. Enfin, il est aisément transposable à différentes tâches.

6. Perspectives

Les comportements générés par le simulateur pourraient être définis plus finement, ce qui demanderait de complexifier l'algorithme de génération de contenu de la réponse ainsi que les modèles associés. Outre le coût associé, il faudrait alors choisir

sur quelles notions affiner les comportements, comme par exemple l'implémentation d'autres stratégies utilisateurs ou la modélisation de phénomènes dialogiques complexes comme la négation ou les critères disjoints.

Par ailleurs, il serait alors intéressant de chercher le bon niveau de granularité des modèles, afin d'obtenir un niveau d'évaluation suffisamment fin tout en préservant leurs capacités de réutilisation. En effet, la « généricité » par rapport à la tâche et aux modes d'interaction est dépendante des possibilités de généralisation des modèles aux nouveaux cas à traiter.

Enfin, restent les difficultés propres à l'évaluation à savoir la modélisation des phénomènes dont l'interprétation nous est encore inconnue aujourd'hui.

Les SDHM sont fondés sur la définition d'un « bon » comportement face aux utilisateurs. Or, la définition de comportements de référence pour une tâche comme le dialogue naturel est éminemment complexe. En effet, le « bon » comportement dialogique n'est pas nécessairement unique et il est fortement lié au contexte de déroulement du dialogue. Le « bon » comportement face à un utilisateur lambda ne sera pas le même que face à un utilisateur epsilon, de même que le « bon » comportement face à un utilisateur se servant d'un téléphone fixe ne sera pas le même que celui face à un utilisateur en situation de mobilité. Une perspective intéressante serait donc de s'appuyer sur la notion d'instance de modèle utilisateur plutôt que sur un modèle unique de tous les utilisateurs. Dans cette optique, la distinction en *utilisateur* et *usager* proposée par Stéphane Chaudiron semble une bonne dichotomie entre les côtés sociaux et donc généraux des usages, et les capacités individuelles des utilisateurs.

7. Bibliographie

Antoine J.-Y., Bousquet C., Goulian J., Jamoussi S., Kurdi M., Rosset S., Vigouroux N., Villaneau J., Quelques problèmes posés à la compréhension de parole : typologie de phénomènes étudiés dans le cadre des campagnes d'évaluation par défi du GDR-I3 du CNRS, Technical report, GT 5.5 - GDR I3 - CNRS, 2002.

Antoine J.-Y., Caelen J., « Pour une évaluation objective, prédictive et générique de la compréhension de CHM orale : le paradigme DCR (Demande, Contrôle, Résultat) », *Langues*, vol. 2, p. 130-139, 1999.

Austin J. L., *How to do Things with Words*, Harvard University Press, 1962.

Austin J. L., *Quand dire c'est faire*, Éditions du Seuil, 1969. Traduction de (Austin, 1962).

Blasband M., Bevan N., King M., Maegaard B., des Tombe L., Krauwer S., Manzi S., Underwood N., « Expert Advisory Group on Language Engineering Standards / Evaluation Working Group Final Report 2 », 1999.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D., the MEDIA consortium, « Semantic annotation of the French MEDIA dialog corpus », *Proceedings of the 9th European Conference on Speech Communication and Technology*, p. 3457-3460, 2005.

Chaudiron S., Évaluation des systèmes de traitement de l'information textuelle, Habilitation à diriger des recherches, Université Paris X, 2001.

Cohen J., « A coefficient of agreement for nominal scales », *Education and Psychological Measurement*, vol. 20, p. 37-46, 1960.

Cronbach L. J., « Coefficient alpha and the internal structure of tests », *Psychometrika*, vol. 16, p. 297-334, 1951.

Devillers L., Bonneau-Maynard H., Paroubek P., « Méthodologie d'évaluation des systèmes de dialogue parlé : réflexions et expériences autour de la compréhension », *Traitement Automatique des Langues*, vol. 43, n° 2, p. 155-184, 2002.

Devillers L., Maynard H., Paroubek P., Rosset S., « The PEACE SLDS understanding evaluation paradigm of the French MEDIA campaign », *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics Workshop on Evaluation Initiatives in NLP*, 2003.

Devillers L., Maynard H., Rosset S., Paroubek P., McTait K., Mostefa D., Choukri K., Charnay L., Bousquet C., Vigouroux N., Béchet F., Romary L., Antoine J.-Y., Villaneau J., Vergnes M., Goulian J., « The French MEDIA/EVALDA project : the evaluation of the understanding capability of Spoken Language Dialogue Systems », *Proceedings of the 4th Language Ressource and Evaluation Conference*, p. 2131-2134, 2004.

DISC Consortium, « DISC - Final Report », 1999.

Eckert W., Levin E., Pieraccini R., « User modeling for spoken dialogue system evaluation », *Proceedings of the IEEE Automatic Speech Recognition Worshop*, p. 80-87, 1997.

FIPA Consortium, « FIPA Communicative Act Library Specification », 2002.

Fischer G., « User Modeling in Human-Computer Interaction », *User Modeling and User Adapted Interaction*, vol. 11, p. 65-86, 2001.

Fleiss J., « Measuring nominal scale agreement among many raters », *Psychological Bulletin*, vol. 76, p. 378-382, 1971.

Georgila K., Henderson J., Lemon O., « Learning user simulations for information state update dialogue systems », *Proceedings of the 9th European Conference on Speech Communication and Technology*, p. 893-896, 2005.

Hirschman L., Dahl D., McKay D., Norton L., Linebarger M., « Beyond Class A : A proposal for Automatic Evaluation of Discourse », *Proceedings of the DARPA Speech and Natural Language Workshop*, p. 109-113, 1990.

King M., Maegaard B., Schütz J., des Tombe L., Bech A., Neville A., Arppe A., Balkan L., colin Brace, Bunt H., Carlson L., Douglas S., Höge M., Krauwer S., Manzi S., Mazzi C., Sielemann A. J., Steenbakkers R., « Expert Advisory Group on Language Engineering Standards / Evaluation Working Group Final Report I », 1996.

Landis J. R., Koch G. G., « The measurement of observer agreement for categorical data », *Biometrics*, vol. 33, p. 159-174, 1977.

Larsson S., Traum D., « Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit », *Natural Language Engineering Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*, Cambridge University Press, p. 323-340, 2000.

Lin B.-S., Lee L.-S., « Computer-aided analysis and design for spoken dialogue systems based on quantitative simulations », *IEEE Transactions on Speech and Audio Processing*, vol. 9, p. 534-548, 2001.

López-Cózar R., De La Torre A., Segura J., Rubio A., « Assessment of dialogue systems by means of a new simulation technique », *Speech Communication*, vol. 40, p. 387-407, 2003.

Mariani J., « The Aupelf-Uref Evaluation-Based Language Engineering Action and Related Projects », *Proceedings of the 1st Language Ressource and Evaluation Conference*, vol. 1, p. 123-128, 1998.

Minsky M., « A Framework for Representing Knowledge », *in* P. Winston (ed.), *Psychology of Computer Vision*, McGraw-Hill, p. 211-277, 1975.

Pietquin O., A Framework for Unsupervised Learning of Dialogue Strategies, Thèse de doctorat, Faculté Polytechnique de Mons, 2004.

Sadek D., Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication, Thèse de doctorat, Rennes I, 1991.

Sadek D., Bretier P., Panaget F., « ARTIMIS : Natural dialogue meets rational agency », *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, p. 1030-1035, 1997.

Schatzmann J., Geogila K., Young S., « Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems », *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, 2005.

Scheffler K., Young S., « Probabilistic simulation of human - machine dialogues », *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 1217-1220, 2000.

Searle J., *Speech Act, an Essay in the Philosophy of Language*, Cambridge University Press, 1969.

Searle J., *Les actes de langage*, Éditions Hermann, 1972. Traduction de (Searle, 1969).

Traum D., Larsson S., « The Information State Approach to Dialogue Management », *in* Smith, Kuppevelt (eds), *Current and New Directions in Discourse & Dialogue*, Kluwer Academic Publishers, p. 325-353, 2003.

Turing A. M., « Computing machinery and intelligence », *Mind*, vol. 49, p. 433-460, 1950.

Walker M. A., « Experimentally Evaluating Communicative Strategies : The Effect of the Task », *Proceedings of the 12th National Conference on Artificial Intelligence*, vol. 1, p. 86-93, 1994.

Walker M., Hirschman L., Aberdeen J., « Evaluation for DARPA COMMUNICATOR Spoken Dialogue Systems », *Proceedings of the 2nd Language Ressource and Evaluation Conference*, 2000.

Walker M., Litman D., Kamm C., Abella A., « PARADISE : A Framework for Evaluating Spoken Dialogue Agents », *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, p. 271-280, 1997.

Walker M., Passonneau R., Boland J., « Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems », *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 512-522, 2001.

Weizenbaum J., « Eliza – a computer program for the study of natural language communication between man and machine », *Communications of the Association for Computing Machinery*, vol. 9, p. 26-45, 1966.