Étude d'un corpus de requêtes en langue naturelle pour des agents assistants

François Bouchet — Jean-Paul Sansonnet

LIMSI-CNRS
Université Paris-Sud XI
BP 133, F-91403 Orsay Cedex
{bouchet,jps}@limsi.fr

RÉSUMÉ. *Un agent conversationnel peut être utilisé pour assister des utilisateurs novices d'applications informatiques. Pour réaliser un système adaptable, nous proposons dans cet article de le fonder sur un corpus spécifiquement constitué pour cet objectif. Nous justifions de la nécessité de construire un tel corpus, qui se distingue des corpus classiques, et détaillons la manière dont nous l'avons construit. Après avoir vérifié son adéquation par rapport à nos objectifs, nous cherchons à en dégager les spécificités en vue de concevoir un langage de requêtes formelles adapté à la fonction d'assistance dans les agents conversationnels.*

ABSTRACT. *A conversational agent can be used to assist ordinary people using softwares. To devise an adaptable system, we suggest in this paper to base it on a corpus built up in this objective. We justify the need for such a corpus, different from conventionnal corpora, and detail the way it has been built. After a check of its appropriateness to our objectives, we try to draw its particularities in order to create a fitting formal requests language for the assistance function in conversationnal agents.*

MOTS-CLÉS : *Agents assistants, fonction d'assistance, corpus de requêtes, actes de dialogue*

KEYWORDS: *Assisting agents, assistance function, corpora of requests, speech acts*

1. Problématique

1.1. *Contexte de l'étude : Agent Conversationnel Assistant DAFT*

Le projet DAFT du LIMSI-CNRS [SAN 05] vise à développer des agents conversationnels animés (ACA), spécialement dédiés à la *fonction d'assistance*, pour des petits composants logiciels (applications à part entière ou services dans des pages web). Dans ce projet, nous développons des outils de traitement automatique de la langue naturelle (TALN) permettant une analyse fine de requêtes en langue naturelle issues d'usagers novices (par exemple de nouveaux usagers d'internet n'ayant pas ou peu d'expérience de l'informatique) et de raisonner sur la structure et le fonctionnement des applications pour fournir de l'assistance en contexte à la manière des systèmes d'aide contextuelle [CAR 93][JAN 05], avec en plus tout l'apport des personnages virtuels [LES 97][CAS 99].

Dans ce cadre, nous avons été amenés à réaliser des travaux de recueil de données empiriques afin d'essayer de cerner les propriétés et d'exhiber les besoins liés à la fonction d'assistance. Ce corpus a pour objectif de servir de base à la réalisation d'un langage de requêtes formelles ancré dans la réalité empirique, car élaboré à partir d'actes de dialogue émis par des sujets en situation d'assistance.
Nous nous appuyons ainsi sur un corpus de requêtes d'assistance en langue naturelle, recueilli au moyen d'expérimentations avec des sujets humains. Par exemple, dans le cadre du site du GT ACA (`http://www.limsi.fr/aca/`) nous avons installé un agent assistant (Marco) avec lequel les visiteurs peuvent interagir en langue naturelle qui nous a permis d'enregistrer des demandes d'assistance en situation de "clavardage" ("chat" en anglais). L'expérience Marco fait partie d'une campagne de recueil beaucoup plus large détaillée dans la section 2.1.

1.2. *Contribution de l'étude aux ACA*

Se pose alors la question de la contribution de telles données empiriques au domaine des Agents Conversationnels. Dans la mesure où nous considérons que la fonction d'assistance est un des champs applicatifs importants des agents conversationnels, il nous paraît essentiel de l'étudier à travers son expression naturelle que constituent des demandes d'aide réellement attestées. Nous présentons donc dans cet article notre démarche, en s'intéressant en particulier à deux problèmes :

1) Constituer un corpus :

a) **nécessité du corpus** : il nous faut d'abord montrer l'existence de spécificités fortes de ce corpus par rapport à d'autres disponibles et potentiellement similaires.

b) **suffisance du corpus** : il reste ensuite à s'assurer de la faisabilité de recueillir un corpus couvrant effectivement la fonction d'assistance dans son ensemble.

2) Établir une typologie des activités :

a) **catégoriser le corpus** : il est nécessaire de réaliser une typologie des activités parce que le corpus recueilli contient des requêtes qui ne relèvent manifestement pas directement de l'assistance.

b) **caractériser les activités** : il faut chercher les moyens de relier les catégories établies à des phénomènes linguistiques pour pouvoir classifier automatiquement les phrases et ainsi les traiter différemment en fonction de l'activité dont elles relèvent.

Dans la section 2, nous présentons l'aspect général du corpus DAFT et la manière dont il a été constitué. La section 3 expose une étude comparative basée sur les actes de dialogue (car ils sont fortement liés à l'activité) entre ce corpus et d'autres corpus de dialogue orienté tâche pour justifier sa nécessité et sa représentativité de la fonction d'assistance. Enfin, la section 4 s'intéresse à la catégorisation du corpus en termes d'activités et tente de caractériser celles-ci.

2. Constitution du corpus DAFT

2.1. *Recueil des phrases*

Le corpus DAFT est composé de plus de 8 000 requêtes recueillies entre juin 2004 et septembre 2006 par différents ACA intégrés dans deux types d'applications :

1) deux applications de type Java [LEG 04][LER 05] : un simple compteur temps réel (thread Java) dont l'utilisateur contrôle le démarrage et la vitesse et un jeu de tours de Hanoi fonctionnant de manière modale (ie n'évoluant que si l'utilisateur agit).

2) deux sites web : une version active du site du groupe AMI du LIMSI permettant l'édition de contenu et le site du GT ACA en libre accès sur internet à des utilisateurs réels.

En dépit de ces différentes origines, on ne constate pas en pratique de différences dans la structure des requêtes et seul le vocabulaire spécifique à chaque application permet parfois de les distinguer.

Pour constituer ce corpus, nous avons eu recours à deux méthodes complémentaires garantissant à la fois l'empirisme et la bonne couverture du corpus :

1) Recueillir des requêtes réelles produites par des utilisateurs placés devant des applications intégrant un ACA de type LEA[1] (2/3 du corpus final).

2) Utiliser des structures dialogiques génériques issues de classifications figurant dans des thésaurus anglais (600 structures [MOL 94]) ou bilingue (300 structures [DUV 96]). Ces structures ont été adaptées pour les employer dans des requêtes d'assistance formulées dans le contexte des applications mentionnées dans le point précédent afin d'assurer l'homogénéisation lexicale du corpus (1/3 du corpus final).

1. LIMSI Embodied Agent, développé par J-C Martin dans le cadre du projet NICE [BUI 05]

Phrases courtes	Phrases de longueur intermédiaire
a plus	A quoi sers-tu ?
a+	alors **lâ** t'es completement paumé !
ah	appelle moi simplement Sylvie
Allez, bye	as tu des **informtion** ?
Allez ciao.	as tu entendu parler d'une expérimentation en cours ? ?
Alors ?	au sujet de cette page, que **peut tu** dire ?
As-tu des amis ?	avec ce corpus, tu sauras ce qu'est une anaphore ...
auf viedersen	à quoi penses tu ?
avec quoi ?	Bah tu viens de dire que tu pouvais remonter le moral !
à l'aide !	**be ouais** tu comprends pas
bah !	ben alors reponds
barre toi de là	bon j'en ai marre je me tire ...
bidule	Bon je me casse. Bye.
bon à rien !	bon y a rien **â** tirer de toi ! !
bon week end	bon, ça va, bonne année 2006
Bon.	Bon, dis-moi plutôt ce que tu sais faire
Bonjour, Marco	bon, reviens à **l apage** d'accueil
bonsoir	**bonjourmon** vieux

Tableau 1. *Exemples de phrases du corpus DAFT de différentes longueurs : en gras sont soulignées quelques sources de confusion pour l'analyseur de requêtes.*

L'utilisation de la seconde méthode compense la difficulté à obtenir un corpus de taille supérieure avec le panel réduit de sujets dont nous disposions[2]. Il y a recouvrement partiel avec les phrases recueillies par la première méthode, mais des formulations moins fréquentes et absentes du corpus recueilli y sont ainsi intégrées, augmentant la couverture générale tout en restant dans le cadre de la fonction d'assistance.

2.2. *Première vue du corpus*

Dans la table 1, nous présentons un extrait du corpus DAFT qui met en relief certaines de ses caractéristiques :

– il contient beaucoup de phrases *bruitées* (expressions orales, fautes d'orthographe, de syntaxe et de grammaire, langage SMS…) non triviales à traiter avec des outils classiques de TALN.

– il ne se présente pas comme une succession de dialogues homme-machine mais comme une liste de phrases employées par les usagers (questions, ordres ou remarques

2. Environ 50 sujets ayant réalisé une session représentant une dizaine de requêtes en environnement contrôlé sur les 3 premières applications décrites, et 30 personnes ayant participé à la campagne Marco ouverte à tous sur le web entre janvier et mars 2006

à l'agent...). En effet, on constate que dans le cadre de la fonction d'assistance, les interactions dialogiques se limitent essentiellement à un seul tour de parole (commande-action, question-réponse...), et peuvent donc être traitées de manière *isolée*.

3. Étude comparative de corpus de dialogue homme-machine orientés tâche

Afin de justifier la nécessité et la représentativité du corpus construit, on peut suivre la méthode de comparaison statistique de corpus établie par G. Ripoche [RIP 06] afin de comparer DAFT à quelques corpus de dialogues homme-homme orientés tâche par l'étude de leurs profils interactionnels.
On appelle profil interactionnel d'un corpus une représentation sous forme d'histogrammes de la répartition des différents actes de dialogue (au sens de Searle [SEA 69]) au sein de celui-ci (cf fig. 1). Les trois corpus de référence pour cette comparaison sont Switchboard [JUR 98] (200 000 énoncés de conversations téléphoniques orientées tâche annotés manuellement), MapTask [CAR 96] (128 dialogues visant à reconstruire une carte par placement de points de repère) et Bugzilla [RIP 06] (1 200 000 commentaires issus de 128 000 rapports de défauts établis lors du développement de la suite logicielle de la Fondation Mozilla).[3]

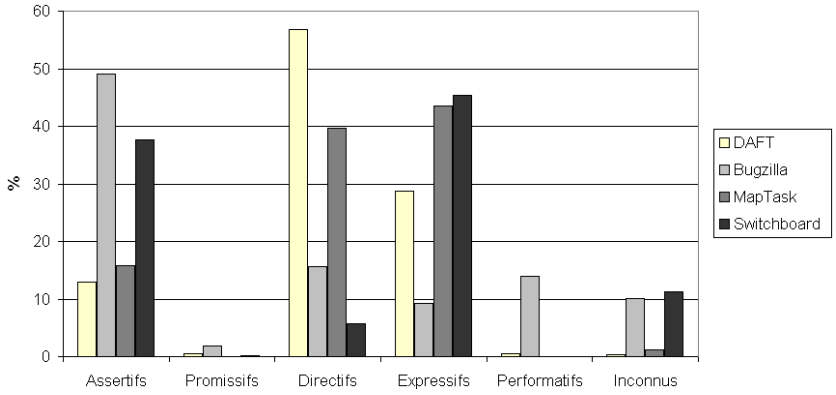

Figure 1. *Comparaison des profils interactionnels de corpus d'assistance dialogique*

Les taxonomies d'actes de dialogue à partir desquelles ces corpus ont été étudiés sont différentes, et la conversion dans une taxonomie commune en 5 actes n'étant pas parfaite, cette comparaison doit être considérée avec précaution. Toutefois, on observe assez nettement certaines caractéristiques distinguant le corpus DAFT des trois autres :

3. Comme signalé par un des critiques, Switchboard et MapTask de part leur nature orale sont naturellement plus riches en nombre de mots que des corpus écrits [KEL 77], mais la proximité des activités a primé sur cette différence de nature dans notre décision de les employer pour cette comparaison (en l'absence à notre connaissance de corpus écrits équivalents).

– une présence majoritaire (57 %) de *directifs*, s'expliquant par un nombre élevé d'ordres directs ou de questions à l'agent. Bien qu'orientés tâche, les autres corpus mettent en jeu uniquement des interlocuteurs humains, et il est vraisemblable que le fait de s'adresser à la machine (même via un agent) tend à rendre les requêtes plus directes car on suppose l'agent incapable des mêmes inférences qu'un être humain.

– un nombre assez faible d'*assertifs* (13 %), l'utilisateur exprimant bien plus son état d'esprit (29 %) par rapport à des faits que ces mêmes faits de manière neutre et "objective" comme c'est le cas par exemple dans le corpus Bugzilla.

– quelques *promissifs* sont présents (1 %) mais marginaux, ce qui s'explique par la nature de la relation entre l'agent assistant et l'utilisateur, car si l'agent doit régulièrement se soumettre à l'humain, l'inverse n'est pas vrai.

Ces divergences entre corpus portant sur un même thème (l'assistance dialogique à une tâche) justifient le besoin de disposer d'un corpus propre à notre champ applicatif.

Conclusion : le corpus est *nécessaire*, car il concrétise les spécificités de la fonction d'assistance médiée par un agent, et *de taille suffisante* car il couvre relativement bien le domaine considéré.

4. Études de catégorisation et de caractérisation du corpus DAFT

4.1. *Catégorisation des activités conversationnelles*

En dépit des conditions de recueil du corpus dans lesquelles les sujets humains *savaient* qu'ils s'adressaient à un agent assistant, de nombreuses phrases ne relèvent pas directement du domaine de l'assistance. Nous nous sommes donc intéressés à identifier les différentes activités conversationnelles réellement présentes dans le corpus.

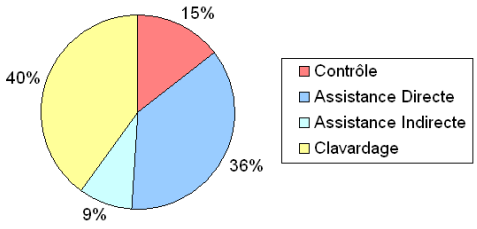

Figure 2. *Répartition du corpus en sous-corpus par activités*

Des requêtes issues de deux sous-ensembles au 1/10e du corpus ont été regroupées manuellement par activités similaires. Ces deux regroupements indépendants ont donné des résultats très proches où l'on a pu distinguer 4 catégories de tailles inégales (cf fig. 2) qui constituent 4 "sous-corpus" correspondant à 4 types d'activités distinctes :

1) **activité de contrôle :** constitué de *commandes*, afin que l'agent agisse lui-même sur l'application.

2) **activité d'assistance directe :** regroupant des *demandes d'aide* explicitement formulées comme telles par l'utilisateur.

3) **activité d'assistance indirecte :** formé d'*opinions* sur l'application qui constituent des demandes d'aide sous-entendues, probablement perceptibles uniquement au niveau de la pragmatique.

4) **activité de clavardage :** réunissant le reste des interactions essentiellement centrées sur l'agent ainsi que des expressions métalinguistiques, phatiques[4] et de back-channeling[5].

L'existence des sous-corpus de contrôle et de clavardage démontre que l'utilisateur attend également d'un ACA dédié à l'assistance des capacités d'action sur les applications où il est intégré et des capacités de dialogue hors assistance (explicables par la présence d'une représentation visuelle).

4.2. *Méthodes de caractérisation des sous-corpus*

Les quatre sous-corpus ont été catégorisés précédemment uniquement par une annotation *manuelle*, mais il serait souhaitable de pouvoir automatiser cette classification. Ceci pourrait alors constituer une étape de prétraitement permettant d'analyser de manière spécifique les activités propres à chaque sous-corpus. On envisage alors trois méthodes de caractérisation possibles de ceux-ci :

– l'étude de la distribution des phrases des sous-corpus en fonction de leur longueur (nombre de mots).

– l'étude des profils interactionnels des sous-corpus, tels que définis en 3.

– l'étude de la sémantique des phrases par analyse de leur retranscription sous forme de requêtes formelles : celle-ci ne sera pas présentée ici car elle nécessiterait l'introduction des spécifications du langage de requêtes formelles[6].

4.2.1. *Caractérisation par la longueur des phrases*

On observe une certaine disparité de longueur des requêtes, les requêtes de contrôle semblant globalement assez courtes comparées aux requêtes d'assistance (cf. table 2), et on peut approximer les répartitions des sous-corpus de contrôle et d'assistance indirecte par une loi normale (test de χ^2 avec un seuil de tolérance de 1%). Néanmoins, les écart-types trop importants ($\sigma \approx 3,5$) disqualifient cette méthode de classification.

4.2.2. *Caractérisation par l'analyse des profils interactionnels*

On distingue sur la figure 3 certaines différences de profils interactionnels assez nettes entre les sous-corpus et par rapport au profil générique du corpus DAFT (rappelé en gris foncé), notamment pour distinguer l'assistance directe (avec une forte

4. Pour maintenir le contact communicatif avec l'agent : « pas vrai lea ? », « tu dors ou quoi ? »...
5. Pour marquer son accord aux propos du locuteur et l'inciter à continuer : « Bon. », « ok ok »...
6. Disponible sur : `http://fbouchet.vorty.net/doc/Spécifications DAFT 2.0.pdf`

Sous-corpus	Contrôle	Assist. directe	Assist. indirecte	Clavardage
Moyenne	5,44	8,01	9,90	6,01
Écart-type	3,36	3,54	3,30	3,62

Tableau 2. *Répartition des phrases par longueur (mots) dans les sous-corpus*

majorité de directifs et quelques expressifs) de l'assistance indirecte (une majorité d'assertifs et des expressifs). En revanche, les profils interactionnels des sous-corpus de contrôle et d'assistance directe sont assez similaires. Cette méthode présente donc un certain intérêt mais ne peut être utilisée de manière unique. En outre, en pratique, l'automatisation de détection de ces actes de dialogue n'est pas non plus triviale.

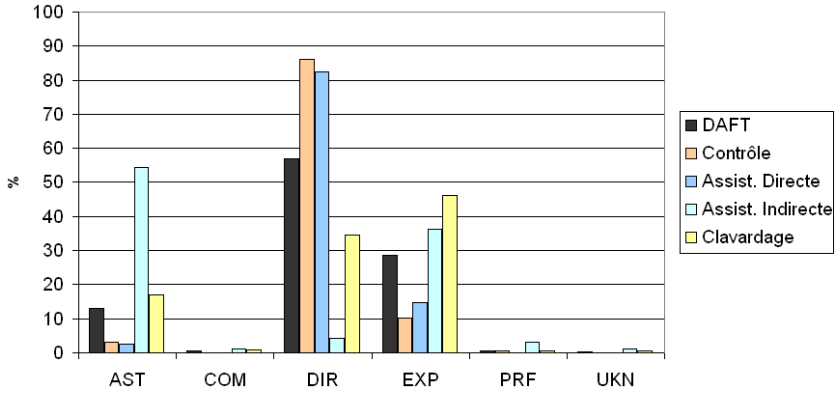

Figure 3. *Répartition des actes de dialogue searliens dans les différents sous-corpus*

Conclusion : Ni l'étude de la longueur des phrases, ni celle de leurs profils interactionnels ne sont suffisantes pour une classification automatique des requêtes en fonction de l'activité dont elle relève : il faut donc envisager une analyse sémantique plus profonde sortant du cadre de cet article mais réalisée dans [BOU 06].

5. Conclusion

Nous avons montré que la fonction d'assistance constitue un domaine particulier des interfaces en langue naturelle, doté d'un vocabulaire spécifique et assez restreint la distinguant d'activités connexes comme le dialogue homme-homme orienté tâche. Pour l'étudier, nous avons constitué un corpus mêlant des requêtes recueillies et construites en situation qui se révèle représentatif de l'assistance. Enfin, nous avons montré que ce corpus recouvre en réalité quatre activités distinctes non facilement distinguables par des méthodes statistiques classiques, mais qui apparaissent plus clairement lors d'une analyse des actes de dialogue réalisée après transcription des requêtes dans un langage formel.

6. Bibliographie

[BOU 06] BOUCHET F., « Conception d'un langage de requêtes pour un agent conversationnel assistant », Master's thesis, Univ. Paris XI, septembre 2006.

[BUI 05] BUISINE S., MARTIN J.-C., « Children's and Adults' Multimodal Interaction with 2D Conversational Agents », *Proceedings of the SIGCHI conference on Human factors in computing systems*, Portland, Oregon, USA, avril 2005, ACM Press, p. 1240-1243.

[CAR 93] CARENINI G., MOORE J. D., « Generating Explanations In Context », *IUI'93 : Proceedings of the 1st international conference on Intelligent user interfaces*, New York, NY, USA, 1993, ACM Press, p. 175-182.

[CAR 96] CARLETTA J., ISARD A., ISARD S., KOWTKO J., DOHERTY-SNEDDON G., ANDERSON A., « HCRC dialogue structure coding manual », rapport, juin 1996, HCRC, University of Edinburgh.

[CAS 99] CASSELL J., BICKMORE T., BILLINGHURST M., CAMPBELL L., CHANG K., VILHJLÁMSSON H., YAN H., « Embodiment in conversational interfaces : Rea », *CHI '99 : Proceedings of the SIGCHI conf. on Human factors in comp. syst.*, New York, NY, USA, 1999, ACM Press, p. 520-527.

[DUV 96] DUVAL A., MARR V., AL., « *Dictionnaire Français-Anglais Robert & Collins* », chapitre Grammaire active, Dictionnaires Le Robert, 4e édition, 1996.

[JAN 05] JANSEN B. J., « Seeking and implementing automated assistance during the search process », *Information Processing and Management*, vol. 41, n° 4, 2005, p. 909-928.

[JUR 98] JURAFSKY D., BATES R., COCCARO N., MARTIN R., METEER M., RIES K., SHRIBERG E., STOLCKE A., TAYLOR P., VAN ESS-DYKEMA C., « Switchboard Discourse Language Modeling Project Final Report », rapport, 1998, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, USA.

[KEL 77] KELLY M. J., CHAPANIS A., « Limited Vocabulary Natural Language Dialogue », *International Journal of Man-Machine Studies*, vol. 9, n° 4, 1977, p. 479-501.

[LEG 04] LE GUERN K., « Définition d'une architecture de médiateur pour des agents conversationnels animés », Master's thesis, Univ. Paris XI, septembre 2004.

[LER 05] LERAY D., « Daft Swing : une famille de composants dialogiques pour un agent conversationnel assistant », Master's thesis, Univ. Paris XI, septembre 2005.

[LES 97] LESTER J. C., CONVERSE S. A., KAHLER S. H., BARLOW S. T., STONE B. A., BHOGAL R. S., « The Persona Effect : Affective Impact of Animated Pedagogical Agents », *CHI '97 : Proceedings of the SIGCHI conf. on Human factors in comp. syst.*, New York, NY, USA, mars 1997, ACM Press, p. 359-366.

[MOL 94] MOLINSKY S. J., BLISS B., « *Inventory of functions and conversation strategies* », p. 177-187, Prentice Hall, janvier 1994.

[RIP 06] RIPOCHE G., « Sur les traces de Bugzilla », PhD thesis, Univ. Paris XI, juin 2006.

[SAN 05] SANSONNET J.-P., LE GUERN K., MARTIN J.-C., « Une architecture médiateur pour des agents conversationnels animés », *WACA'01 : Actes du Premier Workshop Francophone sur les Agents Conversationnels Animés*, juin 2005, p. 31-39.

[SEA 69] SEARLE J. R., *Speech Acts : An essay in the Philosophy of language*, Cambridge University Press, new édition, janvier 1969.