HAL
archives-ouvertes.fr

Quelques spécificités du dialogue Homme-Machine sur Internet

José Rouillard

▶ **To cite this version:**

José Rouillard. Quelques spécificités du dialogue Homme-Machine sur Internet. Bulletin de linguistique appliquée et générale, Presses Universitaires de Franche-Comté, 2001, Traitement Automatique des Langues et Internet, 26, pp.16. hal-02439942

HAL Id: hal-02439942
https://hal.inria.fr/hal-02439942

Submitted on 14 Jan 2020

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

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

QUELQUES SPÉCIFICITÉS DU DIALOGUE HOMME-MACHINE SUR INTERNET

José Rouillard
Laboratoire Trigone - CUEEP Bât B6
Université des Sciences et Technologies de Lille
59655 Villeneuve d'Ascq Cedex – France
E-mail : Jose.Rouillard@univ-lille1.fr

Résumé

Nous nous intéressons à l'étude de nouvelles formes d'interactions pour la navigation et la recherche d'informations sur le Web. Cela nous amène à explorer des pistes dans le domaine du Dialogue Homme-Machine (DHM) écrit et oral, sur Internet. Dans le présent article, nous présentons des éléments du corpus de DHM recueilli grâce au système Halpin-Documentaire, qui nous paraissent être spécifiques au support Internet. Ce sont notamment des contractions de mots, absences volontaires d'accentuation ou encore l'intégration de smileys au sein des répliques des utilisateurs. Les conséquences de ces observations pour le TALN dans le cadre d'une utilisation sur Internet sont ici présentées et discutées.

Mots clefs

Hyperdialogue ; Dialogue Homme-Machine ; TALN ; Spécificité ; Internet.

Abstract

We are interested in new forms of interaction for navigation and research information on the Web. That leads us to explore tracks in the field of the written and oral Man-Machine Dialogue, on Internet. In this article, we present elements of our Man-Machine Dialogue corpus collected thanks to the Halpin-Documentaire system, which appear to be specific to the Internet support. They are in particular contractions of words, voluntary absences of accentuation or integration of smiley within the sentences of the users. The consequences of these observations for the automatic treatment of the natural language within the use on Internet are presented and discussed here.

Key-words

Hyperdialogue ; Man-Machine Dialogue ; Automatic Treatment of the Natural Language ; Specificity ; Internet.

1 Introduction

Dans le cadre du Traitement Automatique de la Langue Naturelle (TALN) sur le réseau Internet, nous nous intéressons à certaines formes d'interactions pour la navigation et la recherche d'informations basées sur le Dialogue Homme-Machine (DHM) écrit et oral. Dans un rapport du groupe « Internet du Futur » demandé par le Réseau National de

Recherche en Télécommunications (RNRT)[1], on pouvait noter que « *Trois applicatifs principaux sont à l'origine de l'adoption d'Internet par le grand public : le Web, la messagerie électronique et l'IRC (Internet Relay Chat, « discussions » sur Internet). Ils répondaient à un besoin social important de communication inter-personnelle asynchrone (la messagerie électronique) ou presque synchrone (IRC ou les forums de discussion) et à un besoin de services en ligne (le Web). Ces éléments restent prépondérants aujourd'hui encore.* ».

Or, si les messages électroniques et l'IRC sont effectivement des communications interpersonnelles, la recherche d'information sur le Web, elle, est très souvent une communication entre un homme et une machine. Malgré la convivialité des interfaces proposées, les outils actuellement disponibles pour la recherche d'information sur le Web demandent à l'utilisateur de fournir un effort cognitif pour traduire une question formulée en langue naturelle, en une suite de mots-clés combinées à des opérateurs logiques booléens.

Violaine Prince (dans un document non publié – 2000) écrivait, à propos du DHM en langue naturelle (LN) pour la recherche d'informations : « *Mais il ne faut pas oublier que le problème premier reste celui de la communication entre l'usager et les environnements logiciels. Bien que les recherches par mots-clés aient beaucoup évolué vers une élaboration et une souplesse qu'elles n'avaient pas précédemment, elles demeurent assujetties à la question suivante : le « mot » (ou l'expression) demandé(e) est-il (elle) ou n'est-il (elle) pas dans le texte, alors que le point de vue de l'utilisateur n'est justement pas un problème de mot, c'est-à-dire de chaîne de caractères, mais un problème de thème, d'idée, de notion, parfois vague, ou parfois non pertinent. C'est là tout le rôle fort important du dialogue entre un usager et un système médiateur, qui consiste à expliciter, à conseiller, à choisir.* ».

De plus, dans la plupart des moteurs de recherche étudiés, chaque requête que formule l'usager est indépendante de la précédente, et aucun de ces systèmes ne s'adapte vraiment à son interlocuteur pour lui fournir un meilleur service. Sur ces constats et sur la base d'une enquête d'usage des outils informatiques pour la recherche d'informations, réalisée dans une médiathèque, nous avons décidé de concevoir un système de DHM utilisant la langue française écrite et orale, pour rechercher de l'information sur une partie du Web, et capable de s'adapter à son interlocuteur de manière dynamique, au cours du dialogue [ROUILLARD, J. (2000)].

Nous avons mis au point un premier système informatique (Halpin-Recueil) grâce auquel nous avons obtenu plus d'un millier de fichiers, fournissant un corpus de dialogues que nous avons jugé suffisant. Cela nous a permis d'étudier les paramètres linguistiques, dialogiques et socio-cognitifs que l'on pouvait déduire de ce corpus [ROUILLARD, J. et CAELEN, J., (1998)]. Puis, les différentes composantes nécessaires pour un DHM oral sur Internet ont été intégrées : reconnaissance vocale, compréhension de la LN par identification de concepts, et synthèse vocale. Enfin, nous avons éprouvé notre modèle, grâce à une réalisation logicielle complète de DHM, avec entrées et sorties vocales via le Web : le système Halpin-Documentaire. Ce système, multiplateforme, qui permet d'interroger une base de données d'Internet, est utilisable avec n'importe quel navigateur compatible Java. Il a permis de montrer qu'un tel hyperdialogue est apprécié par une partie des usagers, notamment les débutants. Son utilisation, basée sur la langue naturelle et l'emploi d'un dialogue oral, permet de réduire la surcharge cognitive de l'utilisateur, et dépasse, dans certains cas, les performances que l'on peut atteindre avec une interface traditionnelle.

En annexe du rapport du RNRT cité précédemment, on pouvait encore lire « *Internet est affaire essentiellement de contenus. Un contenu ne vaut que par le sens qu'il possède pour un utilisateur ou un lecteur, par la valeur d'usage qu'il représente pour ce dernier. Dans ce contexte, un des enjeux liés à l'Internet du futur est de considérer dans quelle mesure la technologie influence la valeur d'usage et la signification des contenus pour ceux qui les abordent via internet.* ».

Nous voulons montrer dans le présent article, que certains éléments du corpus que nous avons recueilli nous paraissent être spécifiques au support Internet : nous présenterons la contraction de mots, l'absence volontaire d'accentuation ou encore l'intégration de smileys au sein des répliques des utilisateurs.

[1] www.telecom.gouv.fr/rnrt et www.education.gouv.fr/rnrt

Certains phénomènes ou comportements langagiers rencontrés dans nos corpus nous semblent être typiquement imputables au type de support utilisé dans cette communication. Le fait d'être connecté au réseau Internet influence, consciemment ou non, les comportements des usagers qui dialoguent avec un ordinateur distant. Contrairement aux expérimentations dans lesquelles les personnes se sentent observées, ici, on note bien que les usagers dialoguent librement avec la machine. On observe ainsi des insultes (« *gros nul, cherche encore* »), des mises en doute des informations données par l'ordinateur (« *non, tu mens* », « *vérifie ta base de données* » , « *vous êtes sûr ?* »). L'exemple du Tableau 1 ci-après est intéressant dans la mesure où l'utilisateur ne se contente pas de « commander » la machine ; il profite pleinement du dialogue en LN pour suggérer la manière dont il faudrait, selon lui, effectuer la recherche : « *H8=alors prend seulement ceux qui ne contiennent pas le mot 'the'* » (H étant mis pour Homme et M pour Machine).

212.198.231.233 (d233.paris-231.cybercable.fr) SUR Mozilla/4.0 (compatible ; MSIE 4.01 ; Windows 98) LE 22/04/99
18 :08 :36 :M0=Que souhaitez-vous faire ? 1 : Retrouver un document que vous connaissez, 2 : Rechercher un ensemble de documents, ou 3 : Découvrir ce site ?
18 :12 :56 :H4=2
18 :13 :09 :M5=Rechercher un ensemble de documents. D'accord. Souhaitez-vous que je vous aide à formuler votre requête ?
18 :13 :25 :H5=oui
18 :13 :30 :M6=Pouvez-vous me donner le thème recherché, entre guillemets s'il vous plaît ?
18 :13 :52 :H6= « intelligence artificielle »
18 :14 :10 :M7=J'ai trouvé 1215 documents ayant comme thème intelligence artificielle. Voulez-vous affiner votre requête ?
18 :14 :47 :H7=je cherche des documents en français seulement
18 :14 :59 :M8=Je n'ai pas d'informations à propos de la langue des documents.
18 :16 :05 :H8=alors prend seulement ceux qui ne contiennent pas le mot « the »
(…)

Tableau 1 : Exemple de DHM avec le système Halpin-Documentaire

2 Contractions et abréviations

On voit donc que les utilisateurs de ce système de dialogue sur le Web profitent du fait que l'échange soit distant et non surveillé pour, semble-t-il, s'exprimer plus « librement », sans contraintes grammaticales ou syntaxiques. D'autres phénomènes de ce genre sont observables dans notre corpus. On y trouve, par exemple la contraction ou l'abréviation de mots.

Mendeleiev.cybercable.fr SUR Mozilla/4.0 (compatible ; MSIE 4.01 ; Windows 95) LE Tue Oct 26 12 :23 :06 MET 1999
12 :23 :07 :M0=Que souhaitez-vous faire ? 1 : Retrouver un document que vous connaissez, 2 : Rechercher un ensemble de documents, ou 3 : Découvrir ce site ?
12 :23 :07 :H0=2
12 :23 :20 :M1=Rechercher un ensemble de documents. D'accord. Souhaitez-vous que je vous aide à formuler votre requête ?
12 :23 :27 :H1=ok
12 :23 :37 :M2=Pouvez-vous me donner le thème recherché, entre guillemets s'il vous plaît ?
12 :23 :59 :H2=cherche « java »
12 :24 :16 :M3=J'ai trouvé 81 documents ayant comme thème java. Voulez-vous affiner votre requête ?
12 :24 :50 :H3=sur koi j'peux affiner
12 :25 :06 :M4=Afin d'affiner, vous devez choisir parmi les critères : Auteur, Titre, Thème, Année, ou Type de document.

Tableau 2 : Utilisation de mots contractés dans un DHM sur Internet

Le Tableau 2 ci-dessus montre comment un internaute contracte le mot « quoi » dans la phrase « *H3=sur koi j'peux affiner* ». Il est assez facile pour un humain de comprendre ce genre de contraction, même si on les rencontre pour la première fois, parce que ces abréviations sont principalement d'ordre phonétique. Ainsi, dans des discussions entre connectés sur Internet, on peut rencontrer par exemple : « T sur ? » (pour « t'es sûr ? »), « C koi ton pb ? » (pour « c'est quoi ton problème ? »), « Alors, koi 2 9 ? » (pour « alors, quoi de neuf ? »), « A+ » (pour « à plus tard »), etc. En anglais, les usagers utilisent souvent des acronymes pour éviter d'avoir à taper toute une expression, comme le montre le Tableau 3 ci-après.

Abréviation	Correspondance	Traduction française
IMHO	In My Humble Opinion	A mon humble avis
J/K	Just Kidding	Je plaisante
AFK	Away from keyboard	Loin du clavier
BAK	Back at keyboard	De retour au clavier
BBL	Be Back Later	Je reviens plus tard
CUL8er	See you Later	A bientôt
OIC	Oh I See	Oh, je vois…
BTW	By The Way…	Au fait…
GTRM	Going To Read Mail	Je vais lire mes messages

Tableau 3 : Quelques abréviations anglaises courantes pour les dialogues sur Internet

C'est à la fois un moyen de gagner du temps, et une façon de revendiquer une appartenance à un groupe d'individus (seuls les anciens et les initiés comprendront ces messages).

Le moteur de recherche Lycos [LYCOS (2001)] donne des informations aux internautes à ce propos : *Sur Internet dans un dialogue en direct ou dans un courrier électronique, ou pour des messages SMS*[2]*, pour gagner du temps, de nombreuses abréviations sont utilisées ; ces abréviations sont d'une grande utilité car elles traduisent une humeur qui se décrit difficilement lors de message court. En voici les plus courantes :*

A+ : à plus tard.
Ad'taleur : Abréviation phonétique de « A tout à l'heure »
Arffffff : Onomatopée exprimant le rire, le nombre de f exprime la force du rire.
ASAP : (As soon as possible) Le plus vite possible.
ASV : Age, Sexe, Ville ? Expression laconique mais qui vous permet de trouver le bon interlocuteur.
Bof : exprime une lassitude ou un léger désaccord.
C : c'est beaucoup plus rapide d'écrire « c » que « c'est ».
CV : abréviation de carte de visite.
G : pour dire « j'ai ».
GRRRR : Exprime le mécontentement.
F.A.Q. : signifie « questions fréquemment posées » (Frequently Asked Questions) ou « Foire Aux Questions ».
Koa ? : abréviation phonétique de « quoi ? », mais certains préfèrent écrire « koua ? »[3].

[2] Short Message Service : envoi de courts messages écrits sur téléphones portables
[3] Notons que l'on ne gagne rien ici à abréger « quoi » en « koua », statistiquement parlant.

Kiécki ? : abréviation phonétique de « qui est ce qui ? » ; variantes possibles : kiélà ? Tkitoa ?
LOL : (lot of loughing) Rire aux éclats.
MDR : Mort de rire.
Slt : abréviation de Salut.
Vi : dérivé phonétique de « Oui » .

Tableau 4 : Quelques abréviations utilisables sur Internet

3 Absence d'accentuation

Sur le Web, il est très fréquent de voir apparaître des mots mal accentués, comme par exemple dans la phrase « *20:03:38:Q=qu'y a-t-il marque sur la page de garde ?* ». Nous pensons que cela n'est pas une erreur de frappe, mais un oubli volontaire de la part de l'utilisateur. En effet, nous savons que pour des raisons d'incompatibilités entre certains types d'ordinateurs (PC vs Mac, par exemple), les caractères accentués sont quelquefois mal retranscrits d'une machine à l'autre. Le caractère diacritique est alors remplacé par un code. Une partie des utilisateurs du courrier électronique (E-mail) et de certains sites de « chat » (conversation/bavardage) ont pris volontairement l'habitude d'omettre l'accentuation dans leurs communications via Internet.
Ce phénomène semble donc ici se retrouver dans les dialogues homme-machine que nous avons recueillis. Si pour l'homme, il est facile de corriger, sans même parfois s'en rendre compte, de tels défauts d'accentuation, il n'en va pas de même pour une machine, et cela risque de compliquer sa tâche de compréhension de l'énoncé formulé par l'utilisateur.
Dans l'exemple de DHM ci-dessous (Tableau 5), l'utilisateur n'accentue pas le mot « Année » en H7. La machine n'arrivant pas à déterminer le critère de recherche choisi, lui demande de reformuler. L'usager rectifie l'accentuation en H9 et le dialogue se poursuit convenablement.

128.93.20.3 (vercoanne1p.inria.fr) SUR Mozilla/4.04 [en] (WinNT; I) LE 30:3:99
14:26:46:M0=Que souhaitez-vous faire ? 1 : Retrouver un document que vous connaissez, 2 : Rechercher un ensemble de documents, ou 3 : Découvrir ce site ?
14:26:46:H0=1
14:26:59:M1=Retrouver un document que vous connaissez. D'accord. Souhaitez-vous que je vous aide à formuler votre requête ?
14:28:30:H1=oui
14:28:34:M2=Connaissez-vous le nom de l'auteur du document que vous recherchez ?
14:28:47:H2=Abiteboul
14:29:00:M3=J'ai trouvé 17 documents avec comme auteur Abiteboul. Voulez-vous affiner votre requête ?
14:29:16:H3=oui
14:29:24:M4=Afin d'affiner, vous devez choisir parmi les critères : Auteur, Titre, Thème, Année, ou Type de document.
14:29:47:H4=auteur
14:30:01:M5=Donnez-moi son nom avec une majuscule. Vous pouvez aussi effacer cet auteur. Ou bien voulez-vous que je cherche les auteurs proches de Abiteboul ?
14:30:57:H6=non
14:31:06:M7=Faites un choix parmi les critères : Auteur, Titre, Thème, Année, ou Type de document.
14:31:34:H7=Annee
14:31:42:M8=Vous ne voulez pas choisir de critère ?
14:32:07:H8=si
14:32:21:M9=Faites un choix parmi les critères : Auteur, Titre, Thème, Année, ou Type de document.
14:32:41:H9=Année
14:32:55:M10=Vous pouvez soit me donner une année, soit effacer l'année en cours. Ou bien voulez-vous que je recherche les documents les plus récents ?

14:33:08:H10=1999 14:33:22:M11=Je n'ai pas trouvé de document de 1999 avec comme auteur Abiteboul. Voulez-vous modifier votre requête ?

Tableau 5 : L'usager rectifie une mauvaise accentuation pour se faire comprendre de la machine

L'exemple précédent montre que pour le traitement automatique de la LN, le système doit prendre garde à l'accentuation des éléments propres au dialogue. Des outils de lemmatisation arrivent d'ailleurs à rétablir les accentuations erronées [GAUSSIER, E., GREFENSTETTE, G., SCHULZE, M. (1997)]. Cependant l'ordinateur ne pourra pas grand chose lorsque c'est l'objet même de la recherche qui est en cause. Par exemple, dans le Tableau 6 ci-après, en H11 et H12, les mots clefs de la recherche ne sont pas accentués. Même si nous savons que la plupart des moteurs de recherche réaccentuent certains mots clefs, il est probable que la machine aurait trouvé plus d'un seul document si l'usager avait tapé "réalité augmentée" au lieu de "realite augmentee".

129.88.32.156 (jmenpet.imag.fr) SUR Mozilla/4.5 [en] (WinNT; I) LE 26:3:99 (…) 18:45:49:M10=Je n'ai pas trouvé de document de 1901 avec comme auteur Dubois. Voulez-vous modifier votre requête ? 18:46:03:H10=je veux tous recommencer 18:46:18:M11=D'accord, tous les critères sont vides à présent. Faites un choix parmi les critères : Auteur, Titre, Thème, Année, ou Type de document. 18:46:36:H11=theme realite augmentee 18:46:47:M12=Vous pouvez soit m'indiquer le thème, entre guillemets, soit effacer le thème en cours. 18:47:09:H12="realite augmentee" 18:47:24:M13=J'ai trouvé un document ayant comme thème realite augmentee. Consultez-le sur la page de droite. Est-ce bien celui que vous cherchiez ? 18:48:00:H13=oui 18:48:05:M14=Très bien, je suis heureux d'avoir pu vous aider dans votre recherche.

Tableau 6 : Importance de l'accentuation dans les mots clés de la requête

4 Smileys

Nous avons vu que le langage employé dans ces extraits de dialogues recueillis sur le Web prend une forme originale et s'apparente plus à de l'oral qu'à de l'écrit. Cependant, les communications demeurent le plus souvent écrites sur Internet, et de ce fait, les intonations, humeurs, et autres états d'esprit ne sont pas évidents à percevoir. Les internautes utilisent donc des signes graphiques pour restituer leurs émotions lors des dialogues.

Ainsi, les smileys (« émoticônes » en français), sont des signes permettant de faire passer une humeur, un sentiment ou une émotion de manière écrite, sur Internet. C'est souvent un moyen de remplacer le ton de la voix. Il faut pencher la tête sur le côté gauche pour comprendre la signification du smileys. Quelques émoticônes sont donnés à titre indicatif dans le Tableau 7.

`:-)`	Le smiley le plus connu, souriant, utilisé pour indiquer un passage humoristique.
`;-)`	Le smiley clin d'œil, pour les remarques ironiques, par exemple.
`:-(`	Le smiley triste : on n'a pas aimé ce qui vient de se dire, ou l'on est pessimiste, ou mélancolique.
`:-o`	Le smiley surpris, ou choqué.

Tableau 7 : Les smileys les plus utilisés sur Internet

Le moteur de recherche Lycos [*LYCOS (2001)*] donne à nouveau des éléments pour comprendre ces pratiques : *Lors des discours sur Internet les phrases sont souvent courtes et peu expressives .Pour mieux se faire comprendre, il est recommandé d'utiliser des smileys qui donneront le ton de vos phrases. Par exemple " Tu m'énerves :-) " et " Tu*

m'énerves :-(" ont deux sens totalement différents : alors que la première phrase est ironique la seconde exprime un fort énervement. Il n'y a pas de copyrights sur les smileys alors utilisez les aussi souvent que nécessaire et n'hésitez pas à en inventer. En voici une liste non exhaustive :

:-) Sourire :-D Rire :-* Bisou ! ;-) Être complice :-X Motus & bouche cousue :-| Être indifférent :-(Ne pas être content :-> Sourire sarcastique (-: Être gaucher 8-) Avoir des lunettes de soleil B:-) Avoir ses lunettes sur le front 8:-) Être une petite fille	:-[Être un Vampire :'-) Pleurer de rire :-# Censuré O:-) Être un ange :-S Avoir dit quelque chose d'incohérent [:-) Mettre son baladeur @:-) Porter un turban =:-) Avoir une coupe porc-épic --:-) Entre un punk-rocker <|-) Être un chinois. *:o) Être un clown .-)N'avoir qu'un oeil

Tableau 8 : Quelques smileys utilisables sur Internet et SMS

C'est ainsi que dans un dialogue de notre corpus, l'utilisateur dit à la machine « H5=*tu es specialiste java on dirait :-)* ». Le smiley à la fin de la phrase est donc ici présent pour signifier une plaisanterie. De même, dans le dialogue du Tableau 9 ci-après, l'utilisateur répond un laconique « *m'en fou :-)* » lorsque la machine lui propose son aide. Certes, le verbe 'foutre' est ici mal conjugué[4] par l'usager, néanmoins un système capable de détecter

[4] Dans l'annexe « Nouvelles technologies de recherche » du rapport sur « l'Internet du futur » déjà cité dans ce document, il est dit : *« Enfin, il faut bien reconnaître que la majorité des utilisateurs font des fautes d'orthographe, sans parler des fautes de grammaire, qui rendent utopique toute interprétation purement linguistique des requêtes.»*

une boutade de ce style ferait preuve d'une remarquable adaptation au type de langage rencontré. Notre système n'étant pas programmé pour interpréter les smileys et les contractions de mots, il ne pouvait pas tenir compte de ces éléments dans son interprétation. Il serait intéressant, dans le cadre d'un dialogue sur le Web, comme le nôtre, qu'une machine puisse interpréter ces formes de communications, et en générer elle-même, si nécessaire.

62.161.240.65 (mailhost.qualitech-info.fr) SUR Mozilla/4.0 (compatible; MSIE 5.0b2; Windows 95) LE 26:3:99
15:37:05:M0=Que souhaitez-vous faire ? 1 : Retrouver un document que vous connaissez, 2 : Rechercher un ensemble de documents, ou 3 : Découvrir ce site ?
15:37:05:H0=2
15:37:17:M1=Rechercher un ensemble de documents. D'accord. Souhaitez-vous que je vous aide à formuler votre requête ?
15:37:40:H1=oui
15:37:45:M2=Pouvez-vous me donner le thème recherché, entre guillemets s'il vous plaît ?
15:38:11:H2="dialogue en direct sur internet"
15:38:26:M3=Je n'ai pas trouvé de document ayant comme thème dialogue en direct sur internet. Voulez-vous modifier votre requête ?
15:38:48:H3=oui
15:38:53:M4=Vous souhaitez modifier votre requête. Selon quel critère je vous prie ?
15:39:20:H4=theme
15:39:31:M5=Vous pouvez soit m'indiquer le thème, entre guillemets, soit effacer le thème en cours.
15:39:49:H5="dialogue"
15:40:09:M6=J'ai trouvé 29 documents ayant comme thème dialogue. Voulez-vous affiner votre requête ?
15:40:26:H6=non

15:40:30:M7=Dites moi si ces résultats vous satisfont.
15:40:50:H7=bof
15:41:00:M8=Souhaitez-vous plutôt affiner, modifier ou annuler votre requête ?
15:41:17:H8=m'en fou :-)
15:41:25:M9=fou est un mot que je ne connais pas.
(…)

Tableau 9 : Utilisation d'émoticônes dans un DHM

Enfin, pour terminer avec les smileys, précisons que certains systèmes de *chat*, à l'image des serveurs [CARAMAIL (2001)] ou [SPRAY (2001)] utilisent des « traducteurs » de smiley, de sorte que les usagers puissent percevoir à l'écran des formes graphiques de smileys tapés sous forme textuelle par leurs interlocuteurs.

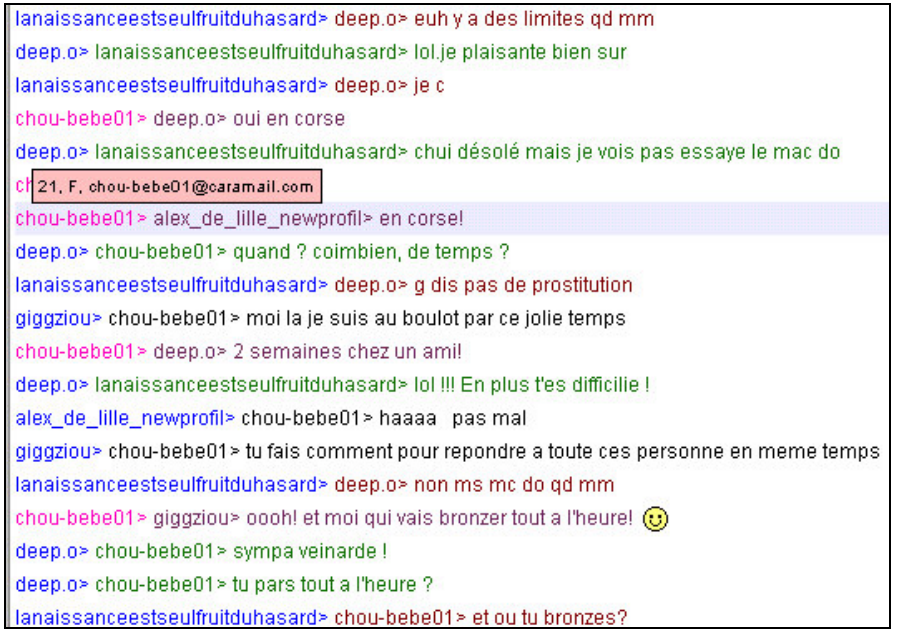

Tableau 10 : Exemple de dialogues en direct sur caramail.com

Dans l'exemple du Tableau 10, on observe bon nombre des phénomènes décrits jusqu'à présent. Tout d'abord, pour bien comprendre qui parle à qui, il faut noter que chaque énoncé est précédé de deux pseudonymes séparés par un signe supérieur. Cela signifie pseudoA parle à pseudoB. Ainsi dans la quatrième phrase, « chou-bebe01 » répond à « deep.o » en lui disant « *oui en corse* ».

Bien entendu, les forums de discussions synchrones sont multi-utilisateurs, et il y a souvent plus de deux interlocuteurs connectés simultanément. Pour analyser un dialogue entre deux internautes, nous avons isolé leur conversation, et remplacé leur pseudonymes respectifs (A mis pour « lanaissanceestseulfruitduhasard », et B mis pour « deep.o »).

A1 : euh y a des limites qd mm
B1 : lol.je plaisante bien sur
A2 : je c
B2 : chui désolé mais je vois pas essaye le mac do
A3 : g dis pas de prostitution
B3 : lol !!! En plus t'es difficile !
A4 : non ms mc do qd mm
B4 : et ou tu bronzes ?

Tableau 11 : Dialogue (isolé) entre de deux internautes

Dans ces huit répliques, les interlocuteurs utilisent des nombreuses abréviations et contractions : « qd mm » remplace « quand même », « lol » remplace « rire aux éclats », « je c » remplace « je sais », « chui » remplace « je suis », « g dis » remplace « j'ai dis », et « ms » remplace « mais ». On remarque également que certains mots ne sont pas accentués : « sûr » ; « essayé » ; « où ». Notons l'ambiguïté de la phrase B2. Plusieurs interprétations sont possibles :

- Soit l'usager a voulu dire : « Je suis désolé mais je ne me vois pas essayer le Mac Do ». Dans ce cas, « ne me » a été oublié, et « essayer » a été remplacé par « essaye » non accentué et mal accordé.
- Soit l'usager a voulu dire : « Je suis désolé mais je ne vais pas essayer le Mac Do ». Dans ce cas, une erreur de frappe fait que le verbe « aller » a été remplacé par le verbe « voir ».

- Soit il a voulu dire : « Je suis désolé mais je ne vois pas. Essaye le Mac Do. ». Cette dernière hypothèse, moins probable, montrerait alors un défaut de ponctuation.

Notons enfin que la dernière réplique (B4) n'est pas directement liée à cette conversation. Le pseudonyme « lanaissanceestseulfruitduhasard » a fait cette remarque en réponse à l'énoncé de « chou-bebe01 » à « giggziou » : « *oooh ! et moi qui vais bronzer tout a l'heure ! :-)* » (voir Tableau 10).

Nous ne nous attarderons pas sur les fautes de français des internautes (« coimbien », « jolie temps », « toutes ces personne », « tout a l'heure ») mais nous dirons seulement qu'il reste encore beaucoup de labeur avant que nos machines puissent interagir en langue naturelle avec nous, si nous nous exprimons de la sorte. Nous avons mis ici en évidence quelques spécificités de la langue française telle quelle se parle sur Internet. Il est possible que dans le futur, l'oral égale voire supplante l'écrit pour la majeure partie des communications homme-homme mais aussi homme-machine. La communication orale sur le réseau Internet, encore plus sujette à être malmenée, sera sans nul doute un vaste sujet d'étude.

5 Conclusion

Il semble donc, au vu de ces résultats, que certaines règles ou coutumes applicables sur Internet, lorsque l'on dialogue entre hommes, se retrouvent dans une situation de dialogue avec une machine. Par exemple, la règle qui conseille d'éviter d'écrire un message tout en majuscule (car CELA DONNE L'IMPRESSION QUE L'ON CRIE) semble être relativement bien respectée.

Nous nous sommes particulièrement intéressé à trois phénomènes observés dans nos corpus de DHM recueillis sur le Web grâce au système Halpin-Documentaire :

(a) la contraction de mots : c'est le fait d'écrire un mot ou une expression de manière abrégée. Les causes peuvent en être diverses : parler rapidement pour tenter de réduire les coûts de communication, pour parler un langage particulier propre à un groupe d'individus, pour coder certains messages, etc. Ces contractions sont parfois compréhensibles, même par des néophytes, car elles sont phonétiquement proches du mot ou de la locution source. Dans d'autres cas, il est impossible de déduire le sens des

abréviations rencontrées pour la première fois. Les débutants sont donc obligés, et cela devient paradoxal, de demander à quoi correspond tel acronyme ou tel sigle, censé améliorer les communications interpersonnelles. Consciemment ou non, les usagers du Web utilisent également ces contractions de mots lorsqu'ils dialoguent avec des machines. Cela nous amène à réfléchir à de possibles dictionnaires d'abréviations que nos systèmes pourraient utiliser aussi bien pour recevoir que pour émettre de l'information.

(b) l'absence d'accentuation : c'est le fait, volontaire ou non, de ne pas accentuer les caractères diacritiques qui théoriquement le sont. Ce phénomène peut avoir plusieurs origines. La compatibilité des jeux de caractères des différentes plate-formes connectées au réseau mondial Internet semble cependant en être la cause première. D'autres éléments nous amènent à penser que l'accentuation peut parfois faire défaut lorsque l'usager ne sait pas exactement où accentuer et quel type d'accent utiliser. Les outils morphosyntaxiques pour le TALN permettent de rétablir automatiquement les erreurs les plus fréquentes. Dans la plupart des cas, ces erreurs ne gênent en rien la compréhension des énoncés. Par exemple le mot 'television' n'est pas équivoque, et l'humain sait parfaitement lui substituer son équivalent proprement accentué. Il le fait d'ailleurs certainement sans en être conscient. Cela devient plus problématique dans des phrases où les mots peuvent être confondus, et altérer le sens de l'énoncé : « livre » versus « livré », « titre » versus « titré », « marque » versus « marqué ». C'est aussi le cas lorsque la négation n'est pas correctement utilisée. Ainsi, dans la phrase « j'en veux plus », l'usager peut vouloir dire « je n'en veux plus (du tout) » ou bien « j'en veux (encore) plus ».

(c) l'utilisation de smileys : nous avons observés que les manques engendrés par les communications écrites, aussi bien synchrones qu'asynchrones, sur le réseau Internet, se sont vus compensés par des astuces textuelles ou graphiques. Cela permet aux intéressés (les initiés dans un premier temps, puis petit à petit, une plus large communauté d'internautes) de palier ces manques.

Des travaux proches des nôtres [LIFE, A., SALTER, I., TEMEM, J.N., BERNARD, F., ROSSET, S., BENNACEF, S., LAMEL, L. (1996), LEHUEN J. (1997), GERARD F., NICOLLE, A. (1998), LEMEUNIER, T. (2000)], montrent que le dialogue entre l'ordinateur et l'humain devient utile et utilisable. Des informations

riches de sens peuvent être extraites des corpus de DHM obtenus en situation de conversations réelles. On peut alors juger de la pertinence des modèles mis au point par les chercheurs des communautés du TALN et des IHM (Interfaces Homme-Machine), et mesurer ainsi l'étendue des perspectives scientifiques du domaine, en matière de DHM finalisés [PIERREL, J.M. (1987)] ou non.

Références

CARAMAIL (2001), http://www.caramail.com

GAUSSIER, E., GREFENSTETTE, G., SCHULZE, M. (1997), « Traitement du langage naturel et recherche d'informations : quelques expériences sur le français. ». *Premières Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, Avignon.

GERARD F., NICOLLE, A. (1998), « Bistro, un modèle de dialogue intégrant la manipulation de concepts », *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, RECTAL'98,* Le Mans.

LYCOS (2001), http://www.lycos.fr

LEHUEN J. (1997), « Un modèle de dialogue dynamique et générique intégrant l'acquisition de sa compétence. Le système COALA ». Thèse de doctorat, Université de Caen.

LEMEUNIER, T. (2000), « L'intentionnalité communicative dans le dialogue homme-machine en langue naturelle », Thèse de doctorat d'informatique, Université du Maine.

LIFE, A., SALTER, I., TEMEM, J.N., BERNARD, F., ROSSET, S., BENNACEF, S., LAMEL, L. (1996), « Data collection for the Mask kiosk: WOz vs prototype system » In *International Conference on Speech and Language Processing*, pp. 1672-1675, Philadelphia.

PIERREL, J.M. (1987), « Dialogue oral homme-machine », Hermès, Paris, 1987.

ROUILLARD, J. (2000), « Hyperdialogue sur Internet. Le système HALPIN », Thèse de doctorat d'informatique, Université Grenoble I.

ROUILLARD, J. et CAELEN, J., (1998) « Étude 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.

SPRAY (2001), http://www.spray.fr