TALN 2010, Montréal, 19–23 juillet 2010

Comment formule-t-on une réponse en langue naturelle ?

Anne Garcia-Fernandez[1,2], Sophie Rosset[1], Anne Vilnat[1,2]
(1) LIMSI-CNRS B.P. 133 91403 Orsay Cedex
(2) Université Paris Sud 11 Orsay
annegf, vilnat, rosset @limsi.fr

Résumé. Cet article présente l'étude d'un corpus de réponses formulées par des humains à des questions factuelles. Des observations qualitatives et quantitatives sur la reprise d'éléments de la question dans les réponses sont exposées. La notion d'information-réponse est introduite et une étude de la présence de cet élément dans le corpus est proposée. Enfin, les formulations des réponses sont étudiées.

Abstract. This paper presents the study of a corpus of human answers to factual questions. Observations of how and how much question elements are used in the answer are done. We define the concept of "information-answer" and study its presence in the corpus. Finally, answer formulations are shown.

Mots-clés : systèmes de réponse à une question, variations linguistiques, réponse en langue naturelle, oral et écrit.

Keywords: question-answering systems, linguistics variations, natural language answer, oral and written.

1 Introduction

La problématique de la génération des réponses dans le cadre des systèmes de question-réponse en domaine ouvert a été étudiée essentiellement pour améliorer la précision de l'extraction de la réponse ou encore de la sélection des documents ((Plamondon *et al.*, 2002), (Soubbotin & Soubbotin, 2001)). Des études sur naturalité des réponses ont été menées sur des domaines limités ((Green & Carberry, 1999), (Higashinaka *et al.*, 2006)). La production de réponses interactives multimodales a été étudiée (van Hooijdonk *et al.*, 2008) mais toujours en domaine limité. Le système ARISE (Lamel *et al.*, 2000) génère des réponses en langue naturelle et produit des phrases qui, si elles sont imitées par les utilisateurs, sont simples à reconnaître par le système de reconnaissance et à comprendre par le système de compréhension. Il intègre des stratégies employées par des opérateurs humains ce qui a abouti à un meilleur taux de satisfaction et une meilleure note sur la naturalite lors d'évaluation utilisateur.
Cependant, les formulations des réponses semblent toujours avoir été déterminées de façon *ad hoc*. Pourtant au vue des exemples suivants, il est évident que plusieurs formulations, différents choix lexicaux mais aussi différents éléments de contenu sont possibles. Cet article se centre sur des choix lexicaux et sémantique (quel contenu pour la réponse) et laisse de côté la question de la syntaxe.

Question : *Où est-ce que se trouve la Joconde ?*
Réponse A : *au Louvre*
Réponse B : *La Joconde se trouve au Louvre.*

Réponse C : *Elle se trouve au Louvre.*
Réponse D : *La Joconde est au Louvre.*
Réponse E : *C'est au Louvre que la Joconde se trouve.*
Réponse F : *La Joconde est une oeuvre de Léonard de Vinci exposée au Louvre.*

Dans le cadre d'une campagne d'évaluation des systèmes de question-réponse, la réponse A sera à privilégier (TREC, site Web). Dans le cas de systèmes coopératifs, ce sera la réponse F qui fait preuve de complétion. Mais, si le but est de proposer une réponse en langue naturelle, pour faciliter l'interaction par exemple, les mêmes interrogations que dans n'importe quelle tâche de génération en langue naturelle demeurent : faut-il construire une phrase (réponse B) ou bien une réponse minimale telle que A suffit-elle ou bien ? Doit-on réutiliser les termes de la question ou bien des pronoms (réponse C), des synonymes (réponse D) ? Quelle structure syntaxique choisir (réponses B *vs* E) ?

Dans cet article, nous présentons une étude de corpus permettant de mettre en valeur les caractéristiques de réponses formulées par des locuteurs français. L'application visée est de produire des réponses en langue naturelle au sein d'un système de question-réponse multimodal (oral ou écrit) et en domaine ouvert. Il ne s'agit pas de donner de bonnes réponses ou bien des réponses dont la forme serait la plus claire ou la plus correcte possible mais de donner une réponse qui peut sembler la plus "naturelle" possible à un utilisateur humain. Nous supposons qu'ainsi, il sera plus satisfait et aura tendance plus facilement à utiliser notre système. Après avoir présenté le corpus utilisé, nous étudions la façon dont les questions sont reprises dans les réponses. Nous observons qualitativement et quantitativement ce que nous définissons comme *l'information-réponse*. Puis les patrons des réponses les plus fréquents du corpus seront exposés.

2 MACAQ : Un corpus de réponses produites par des humains

Nous disposons d'un corpus de réponses humaines à des questions construites. Nous les avons collectées au cours d'expériences (Garcia-Fernandez *et al.*, 2009) où un système posait des questions aux participants qui devaient y répondre.
Le corpus décrit dans (Garcia-Fernandez *et al.*, 2010) est constitué de réponses écrites à des questions posées au travers d'une interface web et de réponses orales transcrites à des questions posées par téléphone et synthétisées vocalement. Les questions posées pour obtenir ces réponses sont des questions de quantité, lieu et temps qui portent sur des thèmes variés relevant de la culture générale française (œuvres d'art, fleuves, évènements nationaux, feuilles "A4", grossesse,...). Pour chaque question, différentes formulations ont été proposées (l'élaboration des variations d'une même question repose sur la grammaire interactive (Luzzati, 2006)).
Les exemples de la figure 1 montrent, pour une même question, quelques unes des reformulations utilisées lors de l'expérience.

Q116 : *Combien mesure une feuille A4 ?*
Q122 : *Une feuille A4 mesure combien de centimètres ?*
Q129 : *Je voudrais savoir quelle taille mesure une feuille A4.*
Q137 : *Est-ce qu'une feuille A4 mesure 21 centimètres de largeur par 29,7 centimètres de longueur ?*

FIG. 1 – Exemples de reformulations d'une question

Dans ces exemples, les questions portent toutes sur le même objet "une feuille A4". On observe que les trois premières formulations sont des questions ouvertes, qui attendent un nombre (accompagné ou non

d'une unité) comme réponse tandis que la dernière question est fermée et attend une réponse *fermée* telle que "oui", "peut-être",...

Nous proposons ici le tableau des caractéristiques principales du corpus global et des sous-parties correspondant d'une part à la modalité d'interaction (orale ou écrite), d'autre part à la nature ouverte ou fermée de la question.

	Tout	Oral	Écrit	Q. ouvertes	Q. fermées
nb de réponses	3132	1044	2088	2119	1025
nb de mots total	25104	7128	17976	18101	7111
nb de mots différents	4574	1634	3363	3331	1794
nb de mots moy par réponse	8,01	6,82	8,60	8,54	6,93
nb de questions différentes	507	496	507	336	171
nb de participants	152	53	99	152	152
nb de réponses/question	6,17	2,39	4,11	6,30	5,99

TAB. 1 – Caractéristiques générales du corpus

Le nombre de participants (152 au total) est de 53 à l'oral et 99 à l'écrit ; les deux phases de l'expérience se sont déroulées indépendamment. Le corpus est constitué de 3132 réponses dont deux tiers de réponses écrites et un tiers de réponses orales transcrites. Il correspond à plus de 25 000 mots pour de près de 4 600 mots distincts. La taille du lexique est 2 fois plus importante à l'écrit qu'à l'oral alors qu'on pourrait s'attendre à un plafonnement du nombre de formes étant donné que les questions posées à l'écrit et à l'oral sont les mêmes. Il s'agit en fait d'un décompte du nombre de formes différentes sans aucun traitement du corpus initial. À l'écrit, les réponses contiennent des fautes d'orthographe, erreurs et coquilles qui élargissent la taille du lexique. À l'oral, le corpus est constitué des transcriptions des réponses des participants, toutes effectuées par un même transcripteur. Les formes sont donc plus régulières et normalisées. Le nombre de mots total est 2,5 fois plus important à l'écrit qu'à l'oral. L'information sur le nombre de mots moyens par réponse indique que les réponses sont en moyenne plus longues (de presque deux mots) à l'écrit qu'à l'oral. Plus précisement, dans deux tiers des cas, les réponses à une même question sont plus longues à l'écrit qu'à l'oral.
Contrairement au caractère écrit ou oral de l'interaction qui sont deux modalités qui s'excluent mutuellement, la nature ouverte ou fermée des questions a été testée au cours de même passations : tous les participants (152) ont répondu à des questions ouvertes et à des questions fermées. Les formulations de questions fermées selon le modèle de variation de la question utilisé sont moins nombreuses que les formulations de questions ouvertes ce qui explique qu'on ait plus du double de questions ouvertes que de fermées (et donc plus du double de réponses dans le sous-corpus "ouvert" que dans le sous-corpus "fermé"). On note que les réponses aux questions ouvertes sont en moyenne plus longues que celles aux questions fermées : un nombre de mot moyen de 8,54 pour les réponses à des questions ouvertes et de 6,93 pour les réponses à des questions fermées (voir tableau 1).

3 Étude des réponses

L'étude que nous avons menée sur ce corpus s'articule en 3 axes.
À travers l'observation des couples question-réponse, nous cherchons tout d'abord à savoir si les réponses réutilisent les termes de la question ou pas. Les réponses répètent-elles le focus, le verbe, ... de la question ou bien les reprennent-elles par un pronom, un synonyme ? Pour un système de question-réponse, il s'agit

de savoir comment un feed-back sur la compréhension de la question peut être donné et plus précisément de travailler sur la réalisation lexicale de la réponse à générer.
Notre second axe concerne l'information répondue en elle-même : les systèmes de question-réponse sont évalués sur l'exactitude de la réponse extraite. Cette réponse, nous l'appelons *information-réponse*. Dans le corpus, nous avons ainsi observé la partie de la *phrase-réponse* (nous utiliserons le mot *réponse*) qui correspond à cette *information-réponse*. La section 3.2 traite plus particulièrement de cet élément dans la réponse.
Un dernier axe d'observations concerne les différentes formulations de réponses possibles. Nous nous attachons aux éléments de la question repris dans la réponse et aux réponses qui ne proposent qu'une seule information-réponse.

3.1 Reprise de la question

Nous avons observé quantitativement et qualitativement la façon dont les réponses reprennent les questions. Nos interrogations sont : reprendre des éléments est-il courant ou au contraire rare ? En cas de reprise, qu'est-ce qui est repris et à quelle fréquence ? La question est-elle reprise de la même façon à l'écrit et à l'oral ? Quelles formes prennent ces reprises ?
Pour répondre à ces différentes questions, nous avons annoté manuellement à la fois le corpus de questions et le corpus de réponses.

Annotation du corpus de questions Le corpus de questions a été annoté de façon à identifier les différents éléments qui les composent. L'annotation porte sur : *l'objet de la question* ou focus, *le verbe principal*, *le type* de réponse attendu s'il est explicitement précisé (comme c'est le cas dans la question Q122 (figure 2)), *les informations complémentaires* qui permettent de restreindre l'objet de la question ou le cadre de validité de la réponse (dans la question Q022 est précisé que l'on recherche le poids d'un bébé *à sa naissance*) ainsi que *l'information-réponse* à confirmer dans le cas d'une question fermée (exemple Q022). Nous entendons par *objet de la question* l'élément sur lequel elle porte. Ainsi dans l'exemple Q116, l'objet est "une feuille A4".

Les questions posées sont des questions factuelles simples construites en faisant varier les caractéristiques morphosyntaxiques de questions prototypiques de la forme "<interrogatif> <verbe> <complément d'objet> (<information complémentaire>" ((Garcia-Fernandez *et al.*, 2009)). L'objet de la question est donc systématiquement, dans l'étude présentée ici, son complément d'objet.

Q116 : *Combien <verbe> mesure </verbe> <objet> une feuille A4 </objet> ?*
Q122 : *<objet> Une feuille A4 </objet> <verbe> mesure </verbe> combien de <type> centimètres </type> ?*
Q129 : *Je voudrais savoir quelle <type> taille </type> <verbe> mesure </verbe> <objet> une feuille A4 </objet>.*
Q022 : *Est-ce qu'<objet> un bébé </objet> <verbe> pèse </verbe> <reprise-information-réponse> 3,2 </reprise-information-réponse> <type> kilos </type> <Qcompletion> à la naissance </Qcompletion> ?*

FIG. 2 – Exemples d'annotation des questions

Comme nous pouvons le voir dans les exemples de la figure 2, les marqueurs interrogatifs n'ont pas été annotés : ce ne sont *a priori* pas des éléments susceptibles d'être repris dans la réponse.

Comment formule-t-on une réponse en langue naturelle ?

Annotation du corpus des réponses Le corpus des réponses a lui aussi été annoté afin de détecter les éléments de la réponse qui correspondent à une reprise. Pour la reprise de l'objet, trois cas ont été distingués : les reprises "exactes", les reprises avec modification et les reprises par pronom. Sont considérées comme reprises exactes, les reprises qui comportent tous les termes de l'élément dans la question y compris si la forme est erronée, abréviée, mal accentuée,... ("bebe", "Bb" et "béb é" sont considérés comme des reprises exactes de "bébé"). Les reprises avec modification peuvent être, comme nous le voyons dans l'exemple A45 de la figure 3, des reprises de l'objet de façon réduite : "une brique" au lieu de "une brique de lait". Les reprises par pronom ("*elle* pèse 1kg", "*cela* dure 5 ans") ont été annotées spécifiquement. Nous soulignons le fait qu'une réponse peut reprendre plusieurs fois un élément de la question, notamment son objet, et ce de différentes façons. Dans la réponse A2280 (figure 3), les trois types de reprise de l'objet sont présents simultanément : reprise exacte, modifiée et par pronom.

A45 : <*objet_modifié*> *Une brique* </*objet_modifié*> *de 1L doit* <*verbe*> *peser* </*verbe*> *environ 1kg (je ne suis pas sûre).*
A2280 : <*objet*> *Une bouteille d'eau* </*objet*> *contient du liquide. (...) Si* <*objet_modifié*> *la bouteille* </*objet_modifié*> *contient 1 litre,* <*objet_pronom*> *elle* </*objet_pronom*> <*verbe*> *pèsera* </*verbe*> *un kilo et ainsi de suite.*

FIG. 3 – Exemples d'annotation de la reprise de la question dans les réponses

Concernant le *type*, les différentes formes des unités de mesure (par exemple "centimètre", "cm") sont considérées comme équivalentes.
On considère qu'il y a reprise du verbe quelle que soit la forme qu'il prend (personne, temps, accompagné d'un verbe modal...)

Taux de reprise Le tableau 2 donne les pourcentages de reprise d'éléments de la question en fonction du nombre effectif d'éléments présents dans les questions. Nous parlons de taux effectifs car si toutes les questions comportent les éléments *objet* et *verbe*, seules certaines précisent un *type*, une *information complémentaire* ou une *information-réponse*. Les pourcentages sont présentés pour l'ensemble du corpus et pour deux partitions de celui-ci : un découpage en fonction de la modalité d'interaction (oral ou bien écrit) et un découpage en fonction de la nature ouverte ou fermée de la question posée.
Le tableau 3 détaille des différents types de reprises de l'objet de la question sur l'ensemble du corpus et les deux partitions oral/écrit, ouvert/fermé.

Élément repris	Tout	Oral	Écrit	Q. ouvertes	Q. fermées
Tous	9.90%	11.69%	9.02%	12.61%	4.31%
Au moins 1 élément	27.39%	29.23%	26.50%	30.87%	20.41%
Objet	22.54%	22.31%	22.65%	24.95%	17.76%
Verbe	15.79%	17.83%	14.78%	18.35%	10.59%
Type	14.51%	19.20%	12.08%	19.52%	5.60%
Information complémentaire	11.79%	12.59%	11.27%	13.05%	9.36%
Information-réponse	9.58%	13.26%	7.87%	N.A.	9.58%

TAB. 2 – Taux de reprise des éléments de la question dans les réponses

Presque 10% des réponses reprennent l'ensemble des éléments de la question et près de 30% en reprennent au moins un élément. On observe que les éléments les plus repris sont l'objet, le verbe et le type de la question. Près d'une réponse sur cinq reprend au moins une fois l'objet de la question de façon modifiée ou non. Parmi ces reprises (tableau 3), 63% reprennent l'objet de la question exactement tandis que

	Tout	Oral	Écrit	Q. ouvertes	Q. fermées
Parmi les réponses qui reprennent l'objet...	22.54%	22.31%	22.65%	24.95%	17.76%
ont au moins une reprise exacte	62.48%	67.24%	60.16%	66.28%	51.38%
ont au moins une reprise avec modification	16.11%	14.41%	16.94%	16.66%	14.36%
ne le reprennent que par un pronom	23.39%	18.77%	25.63%	19.15%	35.91%

TAB. 3 – Détails des reprises de l'objet de la question

19% reprennent l'objet de la question uniquement par un pronom. On observe que les reprises modifiées consistent à réduire l'objet de la question à sa tête de syntagme, les éventuels modificateurs étant omis (exemple A2691 où "une brique de lait" est repris par "la brique") ou au terme sémantiquement le plus important (exemple A862 où "Le mois de février" est repris par "Février").

A2691 : *Cela dépend de la contenance de <objet_modifié> la brique </objet_modifié> (...)*
A862 : *<objet_modifié> Février </objet_modifié> <verbe> dure </verbe> en général 28 jours. (...)*

FIG. 4 – Exemples de reprise avec modification de l'objet de la question

Parmi les réponses qui ne reprennent aucun élément de la question, nous avons comptabilisé les réponses qui ne sont constituées que d'une information-réponse à la question : elles sont au nombre de 1413 soit 60% des réponses qui ne reprennent aucun élément de la question.
En observant les différents sous-corpus, on note qu'il y a plus de reprise à l'oral qu'à l'écrit et en réponse aux questions ouvertes qu'aux questions fermées La proportion de réponses orales est la même dans les corpus *fermé* et *ouvert*. De même, la proportion de réponses aux questions ouvertes est identique dans les corpus *oral* et *écrit* (environ 32%).
C'est en réponse aux questions fermées qu'il y a le moins de reprise. Notamment, peu de réponses reprennent le type précisé dans la question. La reprise de l'objet de la question par un pronom se fait plus souvent à l'écrit qu'à l'oral et pour les réponses à des questions fermées que celles à des questions ouvertes.
Il est intéressant d'observer que bien qu'une question fermée n'attend *a priori* qu'une courte réponse qui infirme ou valide, l'objet de la question et le verbe sont repris dans respectivement 17 et 10 % des réponses.
Concernant l'opposition oral/écrit, on voit qu'en général il y a plus de reprises à l'oral qu'à l'écrit. Nous l'expliquons de deux façons. D'une part, le canal de communication est plus facilement bruité à l'oral qu'à l'écrit. D'autre part la question n'est jamais répétée à l'oral (alors qu'à l'écrit, elle reste affichée à l'écran). On peut interpréter ces reprises comme une confirmation de sa propre compréhension de la question et/ou comme le fait de rendre la réponse donnée plus compréhensible à son interlocuteur.

Les informations complémentaires des questions sont des éléments restrictifs qui servent à limiter l'ambiguïté de la question, à préciser le focus de la question. Par exemple, lorsqu'une question porte sur le poids d'un bébé, il est précisé qu'il s'agit de son poids *à la naissance*. Nous observons que dans les réponses du corpus, ces éléments sont moins repris que les autres éléments de la question.

Synthèse L'un des éléments de la question est repris dans une réponse sur trois et 10% des réponses reprennent entièrement la question : ce phénomène n'est donc absolument pas marginal.
On observe que l'objet, le verbe et le type de la question sont les éléments les plus repris et ce quelle que soit la modalité d'interaction. La plupart des reprises de l'objet sont des reprises exactes mais les reprises par pronom sont aussi possibles (notamment à l'écrit et en réponse aux questions fermées). Les reprises

avec modification consistent en une réduction de l'objet au mot le plus porteur de sens.
Les réponses aux questions fermées se distinguent par leur plus faibles taux de reprise : l'objet est repris dans 17% des réponses mais tous les autres éléments ne sont repris que dans une réponse sur dix.

Tous les systèmes de question-réponse analysent d'une façon ou d'une autre la question. Ils peuvent détecter entre autres le focus, le verbe principal et le type, s'il est précisé, de la question. Une réponse-phrase possible peut donc être composée de ces trois éléments dès lors que la réponse est ouverte. Pour les questions fermées, l'objet de la question peut constituer une base pour la formulation d'une phrase-réponse. Pour produire de telles réponses, il faut bien évidemment déterminer d'autres éléments : quelle structure syntaxique ? quelle information-réponse ? Dans la section suivante, nous nous intéressons plus particulièrement à cette information-réponse.

3.2 L'information-réponse

Donner une réponse à une question va beaucoup plus loin que répondre une nouvelle information qui correspond à ce qui est mis en cause par la question. Comme nous l'avons vu dans la section précédente, reprendre des éléments de la question est un phénomène fréquent, ce qui laisse supposer qu'un retour d'information sur ce qui a été compris de la question est essentiel.

Du point de vue des systèmes de question-réponse, l'un des éléments évalué est l'information extraite des documents. Il peut s'agir d'une réponse précise de peu de mots, souvent une entité nommée ou assimilée, ou encore une liste de mots[1] (TREC, site Web). Face à des réponses spontanées comme celles de notre corpus, la définition de cette *information-réponse* est primordiale.

Ici, notre but est d'annoter la ou les informations qui sont une information nouvelle (par rapport à celles contenues dans la question) et correspondent soit au type attendu de la réponse, soit à un aveu d'incompétence (nous parlerons de *réponse à valence négative* : "je ne sais pas" par exemple).
Les informations supplémentaires sur le cadre de validité de la réponse (par exemple "les années bissextiles" dans l'exemple A60) ne font pas partie de l'information-réponse.
Dans le cas des réponses à des questions fermées, l'élément annoté est le premier qui exprime une infirmation ou une confirmation de la question.

La figure 5 présente des exemples d'annotation que nous discutons à présent.
La valeur des réponses et le fait qu'il s'agissent bel et bien d'une bonne réponse ou pas dépend de la connaissance du monde du participant. Nous ne nous attacherons pas à les commenter.
Certaines réponses du corpus contiennent plus d'une information-réponse (exemple A2766 figure 5).
Que ce soit une réponse à une question ouverte ou fermée, dans le cas des réponses à valence négative (exemple A2480 figure 5), l'information-réponse se réalise en général par une expression figée ou semi-figée telle que "je n'en sais pas grand chose", "je sais pas", "je ne connais pas",...
Dans le cas des réponses à valence positive, les réponses aux questions ouvertes et fermées ne prennent pas la même forme. Pour les réponses à des questions ouvertes, l'information-réponse est une entité nommée : nom de ville, musée, aéroport,... pour les questions de lieu, date, mois, année... pour les questions de temps et valeur accompagnée ou non d'une unité de mesure pour les questions de quantité. Ces informations-réponse peuvent aussi contenir des modifieurs comme on peut le voir dans l'exemple A1938 de la figure 5.
Une réponse à valence positive à une question fermée peut être un adverbe ou une expression : "oui","en

[1] Nous parlons ici de réponses à des questions factuelles uniquement

effet", "non",... (exemple A665 figure 5). Mais il peut aussi s'agir d'une reformulation sous forme déclarative de la question (exemple A2499). Plus précisément, on distingue différents cas. On observe des affirmations qui reprennent l'information-réponse proposée dans la question et confirment ainsi la question (exemple A2499). On observe des phrases négatives qui elles aussi reprennent l'information-réponse de la question, infirmant ainsi la question (exemple A2608 figure 5).

On observe aussi des affirmations au sein desquelles l'information-réponse de la question est substituée par une autre information-réponse (exemple A1353 figure 5).

A2766 : *la Vénus de Milo se trouve <information-réponse> en France </information-réponse> notamment <information-réponse> à paris </information-réponse>,<information-réponse> au Louvre </information-réponse>.*
A2480 : *<information-réponse> je n'en ai aucune idée </information-réponse>*
A1938 : *un mois de février dure <information-réponse> à peu près 28 jours </information-réponse>*
A60 : *<information-réponse> 28 jours </information-réponse> les années bissextiles, <information-réponse> 29 jours </information-réponse> les autres*
A665 : *<information-réponse> oui </information-réponse>, il dure effectivement 5 ans en France*
A2499 : *Elle <information-réponse> a bien été signée </information-réponse> à cette date*
A2608 : *Le Rhône <information-réponse> ne fini pas </information-réponse> dans le delta de la Camargue*
Q061 : *Le poids d'une bouteille d'eau est de 2 kilos ?*
A1353 : *le poids d'une bouteille d'eau de 1 litre est de <information-réponse> 1 kilo environ </information-réponse> (...)*

FIG. 5 – Exemples d'annotation de l'information-réponse dans les réponses

3.3 Formulations des réponses

Nous avons observé dans la section précédente la forme que peut prendre une information-réponse. Nous avions vu dans la section 3.1 comment les réponses reprennent des éléments de la question. À présent, nous proposons d'étudier comment ces éléments s'articulent entre eux.

Nous avons créé pour chaque réponse son patron. Il s'agit d'une forme réduite de la réponse dans laquelle les éléments de reprise de la question sont remplacés par le nom de l'élément repris, l'information-réponse est substituée par le terme "information-réponse" et les autres éléments de la réponse sont supprimés. L'exemple 6 montre en quoi consiste cette réduction.

A2212 : *Si tout se passe bien, <objet>une grossesse</objet> <verbe>dure</verbe> <information-réponse>aux environ de 9 mois</information-réponse>. Il se peut qu'elle soit plus courte.*
A2212 *réduite* : *objet verbe information-reponse*

FIG. 6 – Exemple de patron de réponse

Nous avons observé les différents patrons de réponse du corpus et des sous-corpus oral/écrit, ouvert/fermé pour les réponses qui ne proposent qu'une unique information-réponse. Le tableau 4 présente les patrons de réponse les plus fréquents dans l'ensemble du corpus et leurs effectifs dans chacun des sous-corpus.

Comment formule-t-on une réponse en langue naturelle ?

	Tout	Oral	Écrit	Q.ouvertes	Q.fermées
info-rep	2100	711	1389	1315	785
objet verbe info-rep	147	54	93	143	4
objet_pronom verbe info-rep	56	20	36	52	4
objet info-rep	44	11	33	40	4
objet_modifié verbe info-rep	23	8	15	22	1
objet verbe info-rep Qcompletion	22	11	11	20	2
objet_pronom info-rep	17	2	15	16	1
objet_modifié info-rep	10	2	8	8	2
verbe info-rep	9	3	6	7	2
objet_pronom info-rep reprise-info-rep	7	3	4	0	7
objet info-rep reprise-info-rep	7	2	5	0	7
objet Qcompletion verbe info-rep	7	3	4	7	0
info-rep reprise-info-rep	6	1	5	0	6
Qcompletion objet verbe info-rep	6	3	3	6	0
objet objet_pronom verbe info-rep	4	1	3	4	0
info-rep objet reprise-info-rep	3	0	3	0	3

TAB. 4 – Patrons de réponses les plus fréquents. Avec *info-rep* pour information-réponse et *reprise-info-rep* pour reprise-information-réponse.

Nous observons les différentes formulations de réponse présentes dans le corpus. L'ordre des éléments est assez figé : *objet verbe information-réponse*. Seule la reprise des informations complémentaires de la question peut se faire tant avant l'information-réponse qu'après. On observe que certains patrons ne sont valables que pour les questions ouvertes et d'autres uniquement pour les questions fermées. En revanche, on n'observe pas de telle différence entre l'oral et l'écrit.
Si ces patrons donnent une idée sur l'ordre d'apparition des éléments de reprise de la question et l'information-réponse, ils sont peu représentatifs de la forme syntaxique de la réponse qui comporte des marques d'hésitation, des connecteurs,... Cependant, ces patrons sont intéressants dans l'optique de donner une réponse en langue naturelle minimale qui ne demande pas de traitement ou de connaissances supplémentaires à celles qu'ont déjà les systèmes de question-réponse : d'une part une analyse de la question mettant en valeur son focus, son verbe principal et les éventuels éléments complémentaires, d'autre part la recherche d'une information-réponse. Un système de question-réponse peut s'inspirer des différents patrons ici mis en évidence pour construire une "phrase-réponse" sans se limiter à retourner l'information-réponse seule.

4 Discussion et conclusion

Nous avons montré que les réponses du corpus reprennent les éléments de la question et en particulier l'objet, le verbe et le type et ce, que l'interaction soit orale ou écrite. D'une manière générale, il y a moins de reprise en réponse aux questions fermées qu'aux questions ouvertes. D'autre part, l'objet est plus souvent repris par un pronom à l'écrit qu'à l'oral.
Nous avons défini l'information-réponse comme l'élément qui soit apporte une réponse à la question, soit exprime un aveu d'incompétence. Cette information-réponse revêt différentes formes en fonction de la nature fermée ou ouverte de la question.
Nous avons réduit les réponses à leurs éléments de reprise de la question et à l'information-réponse et avons montré les patrons de réponse présents dans le corpus. Il ne s'agit pas de patrons syntaxiques mais

ils peuvent être utilisés pour déterminer le *Quoi dire* et effectuer certains choix lexicaux dans la phase de génération de réponse en langue naturelle et ce, quel que soit le système de question-réponse.
Nous avons mis de côté les réponses comportant plus d'une information-réponse. Il serait intéressant d'observer ces cas particuliers (19% des réponses) et de déterminer le rapport qui existe entre ces informations-réponse : modification de la granularité, reformulation, validité différente,... D'autre part, nous avons ignoré, dans les réponses, les éléments autres que les reprises de la question et l'information-réponse. Une étude plus fine est d'ores et déjà en cours. nous cherchons à mettre en valeur les structures syntaxiques des réponses du corpus.
Du point de vue de l'intégration à un système de question-réponse, la construction d'une phrase-réponse minimale suivant les patrons mis en valeur dans la section 6 peut constituer les bases d'un premier module de génération de réponses en interaction en langue naturelle.
Enfin nous envisageons une évaluation type "satisfaction utilisateur" qui permettra une observation, non plus de la production de réponses, mais de leur perception.

Remerciements Ce travail a été partiellement financé par OSEO par le biais du programme Quaero.

Références

GARCIA-FERNANDEZ A., ROSSET S. & VILNAT A. (2009). Collecte et analyses de réponses naturelles pour les systèmes de questions-réponses. In *Actes de TALN 2009*.

GARCIA-FERNANDEZ A., ROSSET S. & VILNAT A. (2010). MACAQ : A multi annotated corpus to study how we adapt answersto various questions. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2010/.

GREEN N. & CARBERRY S. (1999). A computational mechanism for initiative in answer generation. *User Modeling and User-Adapted Interaction*, **9**(1), 93–132.

HIGASHINAKA R., PRASAD R. & WALKER M. A. (2006). Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In *ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 265–272, Morristown, NJ, USA : Association for Computational Linguistics.

LAMEL L., ROSSET S., GAUVAIN J.-L., BENNACEF S., GARNIER-RIZET M. & PROUTS B. (2000). The limsi arise system. *Speech Communication*, **31**(4), 339–354.

LUZZATI D. (2006). Essai de description interactive : l'exemple des questions quantificatrices. *Colloque La quantification*, **1**, 15.

PLAMONDON L., LAPALME G. & KOSSEIM L. (2002). The quantum question answering system at trec-11. In E. M. VORHEES & D. K. HARMAN, Eds., *Proceedings of the Eleventh Text Retrieval Conference (TREC-2002)*, p. 750–757, Gaithersburg, Maryland : NIST.

SOUBBOTIN M. M. & SOUBBOTIN S. M. (2001). Patterns of potential answer expressions as clues to the right answers. In *In Proceedings of the Tenth Text REtrieval Conference (TREC 10)*, p. 293–302.

TREC (site Web). Text REtrieval Conference. http ://trec.nist.gov/.

VAN HOOIJDONK C., KRAHMER E., MAES A., THEUNE M. & BOSMA W. (2008). Production and evaluation of (multimodal) answers to medical questions. In *EARLI SIG2 - 2008 Conference on Comprehension of Text and Graphics*, p. 72–76, Tilburg, The Netherlands.