

## Collecte et analyses de réponses naturelles pour les systèmes de questions-réponses

Anne Garcia-Fernandez, Sophie Rosset, Anne Vilnat  
LIMSI-CNRS / Université Paris Sud 11 Orsay  
annegf, vilnat, rosset @limsi.fr

**Résumé.** Notre travail se situe dans le cadre des systèmes de réponse à une question et à pour but de fournir une réponse en langue naturelle aux questions posées en langue naturelle. Cet article présente une expérience permettant d’analyser les réponses de locuteurs du français à des questions que nous leur posons. L’expérience se déroule à l’écrit comme à l’oral et propose à des locuteurs français des questions relevant de différents types sémantiques et syntaxiques. Nous mettons en valeur une large variabilité dans les formes de réponses possibles en langue française. D’autre part nous établissons un certain nombre de liens entre formulation de question et formulation de réponse. Nous proposons d’autre part une comparaison des réponses selon la modalité oral / écrit. Ces résultats peuvent être intégrés à des systèmes existants pour produire une réponse en langue naturelle de façon dynamique.

**Abstract.** Situated within the domain of interactive question-answering, our work is to increase the naturalness of natural language answers. This paper presents an experiment aiming at observing the formulation of answers by French speakers. The system asked simple questions to which the humans had to answer. Two modalities were used : text (web) and speech (phone). We present and analyze the collected corpus. Within the large variability of answer forms in French, we point some links between the answer form and the question form. Moreover we present a preliminary study on the observed variation between modalities. We expect these results to be integrable in existing systems to dynamically produce adapted natural language answers.

**Mots-clés :** systèmes de réponse à une question, expérience, variations linguistiques, réponse en langue naturelle.

**Keywords:** question-answering systems, experimentation, linguistics variations, natural language answer.

## 1 Introduction

Les systèmes de question-réponse ont pour objectif de permettre l’accès à une information. À partir d’une question en langue naturelle et d’un ensemble de documents, le système fournit une réponse à la question. Actuellement la plupart de ces systèmes propose une réponse composée de l’*information réponse* à laquelle s’ajoute une éventuelle *justification* (extrait du (des) document(s) dans le(s)quel(s) la réponse a été trouvée). Le travail présenté ici s’inscrit dans l’objectif de fournir une réponse complètement formulée en langue naturelle à la question posée. Les enjeux sont multiples. Il s’agit d’offrir une interface plus *naturelle* à l’utilisateur : à une question

en langue naturelle, une réponse en langue naturelle se doit d'être proposée. Nous supposons qu'il s'agit d'un pré-requis pour créer l'échange interactif qui mène d'une première question posée par l'utilisateur à la réponse finale donnée par le système et qui satisfera l'utilisateur.

Au delà de répondre la bonne information, nous nous centrons ainsi sur le niveau phrastique de la réponse et nous attachons au fait que la réponse peut prendre plusieurs formes linguistiques comme le montre l'exemple ci-dessous :

**Question :** *Où est-ce que se trouve la Joconde ?*

**Réponse A :** *La Joconde est au Louvre.*

**Réponse B :** *C'est au Louvre que la Joconde se trouve.*

Par opposition à un système qui répondrait :

**Réponse :** *musée du Louvre*

**Passage :** *La Joconde (ou Portrait de Mona Lisa) est un tableau de Léonard de Vinci, réalisé entre 1503 et 1506 d'une femme qui s'appelait à l'époque Mona Lisa del Giocondo. Huile sur panneau de bois de peuplier de 77 x 53 cm, il est exposé au musée du Louvre à Paris.*

Notre intuition est qu'en fonction des caractéristiques linguistiques de la question (sa syntaxe, mais aussi sa sémantique (le thème et le rhème de la question), le lexique sélectionné par le locuteur et la granularité qu'elle sous-entend pour la réponse), la ou les réponses possibles sont différentes toujours d'un point de vue linguistique. Il nous semble donc primordial de prendre en compte ces possibles variations. Être capable de produire une réponse adaptée à la forme linguistique de la question, c'est s'éloigner des systèmes de dialogue où les réponses prennent des formes figées fixées par un ensemble de patrons syntaxiques prévus à l'avance et aller vers un système de dialogue plus naturel et surtout plus dynamique. Ainsi pour déterminer la ou les formulation(s) de réponses possibles pour une question donnée, nous nous appuyons d'une part sur un modèle linguistique que nous utiliserons comme un générateur de variations de question, d'autre part sur l'observation et l'analyse de réponses formulées par des locuteurs du français.

Des travaux récents en linguistique (Luzzati, 2006) proposent un modèle de la question en interaction mettant en exergue que la formulation d'une question traduit d'une part le focus exact que le questionneur vise, d'autre part son intention d'obtenir une réponse plus ou moins précise, de remettre en cause ou non la capacité de l'interlocuteur à répondre,... L'idée sous-jacente est que la forme morphosyntaxique d'une question traduit la volonté du questionnant et peut donc être un indice de la réponse qu'il escompte et de sa forme. Nous ne supposons pas qu'à une forme de question proposée par le modèle corresponde une forme de réponse unique. Mais les différentes classes de questions qu'il détermine nous servent de point de départ pour la mise en place de notre étude. Nous souhaitons donc l'utiliser afin de déterminer la ou les formes des réponses en langue naturelle qu'un système de question-réponse pourrait fournir.

La problématique des réponses et de leur génération complète dans le cadre des systèmes de question-réponse en domaine ouvert a été essentiellement étudiée pour améliorer, à terme, la précision de l'extraction de la réponse et de la sélection des documents et passages (par exemple (Plamondon *et al.*, 2002), (Soubotin, 2006) ou encore (de Chalendar *et al.*, 2002)). Des travaux existent également sur la forme de la réponse dans le cadre de systèmes de question-réponse interactifs mais portent soit sur l'aspect multimodal (par exemple (van Hooijdonk *et al.*, 2008)) soit sur des systèmes en domaine limité pour lesquels il existe des modèles de tâche ou de domaine permettant une adaptation naturelle fine de la réponse (par exemple (?)).

Nous nous intéressons à la production naturelle de réponses complètes dans le cadre de systèmes interactifs de question-réponse en domaine ouvert. Observer comment des humains produisent des réponses à des questions paraît être un préalable nécessaire.

C'est pourquoi, à l'instar de (Anaya & Kosseim, 2003), nous observons des réponses produites par des locuteurs natifs du français afin de mettre en valeur les liens qui existent entre différentes formes linguistiques de question et les formes linguistiques des réponses récoltées au cours des expériences. Les réponses obtenues seront analysées en veillant à ce que les liens entre questions et réponses mis en exergue puissent être exploitables quel que soit le système de question-réponse considéré.

La section 2 détaille la mise en place des expériences. La section 3 présente le corpus collecté et des analyses préliminaires obtenues. Enfin la section 4 discute cette expérience.

## **2 Protocole d'observation de réponses humaines**

Afin d'observer les réponses formulées par des humains, nous avons mis en place une expérience. Dans celle-ci, plutôt qu'un système de question-réponse réponde à des questions posées par des utilisateurs (ou données dans un fichier d'entrée dans le cas des campagnes d'évaluation), nous avons demandé à des locuteurs natifs du français de répondre à un ensemble de questions posées par un système. Il s'agit d'un protocole unique et même si certains travaux observent des réponses de locuteurs humains, aucun ne permet d'observer des réponses sur différentes modalités (orale et écrite dans notre cas) pour les mêmes questions et/ou des questions dont la variance linguistique est contrôlée et avec un panel de sujet aussi large que le nôtre (40 chez (Anaya & Kosseim, 2003), 148 chez nous).

L'expérience a pour but de permettre l'observation de réponses dites humaines. Pour cadrer cette observation et comme notre hypothèse première est qu'il y a un lien entre la forme linguistique de la question et celle de la réponse, nous avons mis en place une expérimentation permettant d'observer comment des locuteurs natifs du français répondent à une question ayant une forme linguistique bien déterminée. Pour cela nous avons créé un panel de questions à soumettre aux sujets. La sous-section suivante expose la construction. Nous présentons par la suite la mise en place de l'expérience sur deux modalités : l'écrit et l'oral, ainsi que l'organisation des passations.

### **2.1 Panel de question**

Le choix des questions à soumettre aux sujets est crucial. Il doit nous permettre de faire une observation la moins biaisée possible et d'établir un modèle qui soit le plus complet possible. Concrètement, il doit nous permettre de déterminer si pour une forme de question donnée, il y a une forme de réponse privilégiée par les sujets. Et ce, en prenant en compte un maximum de variation linguistique possible pour une question.

Étant donné le cadre de ce travail, nous aurions pu, à l'instar d'autres études qui nous ont précédé, sélectionner des questions issues de campagnes d'évaluation. Cependant cette approche ne permet d'obtenir qu'un éventail de variations linguistiques faible ((Anaya & Kosseim, 2003)).

Nous avons à notre disposition un modèle, la Grammaire Interactive (ou GRINT, (Luzzati,

2006) et (Garcia-Fernandez & Lailier, 2007)) qui peut nous servir d'outil de génération de variations autour d'une question prototypique. Il s'agit d'une modélisation des questions de la langue française selon trois dimensions. La dimension dite ontologique peut être mise en correspondance avec le type attendu de la réponse, information sur laquelle se base, par ailleurs, les systèmes de question-réponse.

En se positionnant sur la dimension ontologique (en profondeur dans le tableau 1), on peut considérer une question prototypique dont les variations sont contrôlées selon les deux autres dimensions. Ces deux dernières représentent des variations de type morphosyntaxique. La dimension dite paradigmatique (en vertical dans le tableau 1) correspond à une variation de l'interrogatif utilisé (pronom interrogatif, locution introductrice de question indirecte, pas d'interrogatif,...). La variation dite syntagmatique (en horizontal) concerne la structure syntaxique globale de la question (inversion du sujet, question directe, indirecte,...).

	adverbiale	déterminative	déterminative	déterminative	confirmative	
	adverbiale	numérale	déterminative	nominale	confirmative	
	adverbiale	déterminative neutre	déterminative grain fin	déterminative gros grain	confirmative	
prototypique	Où est la Joconde ?	A quel endroit est la Joconde ?	Dans quel musée est la Joconde ?	Dans quel pays est la Joconde ?	La Joconde est-elle au musée du Louvre ?	LIEU
tonique	La Joconde est où ?	La Joconde est à quel endroit ?	La Joconde est dans quel musée ?	La Joconde est dans quel pays ?	C'est bien au Louvre qu'est la Joconde ?	TEMPS
renforcée	Où est-ce qu'est la Joconde ?	A quel endroit est-ce qu'est la Joconde ?	Dans quel musée est-ce qu'est la Joconde ?	Dans quel pays est-ce qu'est la Joconde ?	Est-ce la Joconde est au Louvre ?	QUANTITÉ
périphrastique	Je voudrais savoir où est la Joconde.	Je voudrais savoir à quel endroit est la Joconde.	Je voudrais savoir dans quel musée est la Joconde.	Je voudrais savoir dans quel pays est la Joconde.	Je voudrais savoir si la Joconde est au Louvre.	
assertive	La Joconde est à Paris, au musée du Louvre ?	La Joconde est dans un musée ?	La Joconde est au Louvre ?	La Joconde est en France ?	C'est bien au Louvre qu'est la Joconde ?	

TAB. 1 – Illustration des variations tri-dimensionnelles proposées par la GRINT

Nous avons ainsi déterminé un ensemble de question prototypiques dont nous avons construit l'ensemble des variantes telles que définies par la GRINT.

Les questions prototypiques ont été choisies selon différents critères. Tout d'abord notre choix a été de ne pas balayer l'ensemble des classes ontologiques définies par la GRINT. En effet le total des combinaisons possibles selon les 3 dimensions s'élève à plus de 200 (15 classes ontologiques et en moyenne 4 classes paradigmatiques et 4 classes syntagmatiques) ce qui représente un échantillon bien trop important de tests à réaliser. Un corpus de requêtes utilisateurs collecté au LIMSI (Galibert *et al.*, 2005), (Rosset & Petel, 2006)) en situation de Dialogue Oral Homme Machine (corpus RITEL) a été annoté selon le type ontologique défini par la GRINT. Le tableau 2 présente le taux de présence de ces types dans le corpus Ritel. Nous avons observé ces taux de présence et choisi de mener notre étude sur 3 des plus fréquentes : *Lieu*, *Temps* et *Quantité*.

Nous supposons cependant que choisir une unique question prototypique par classe ontologique est insuffisant. Notamment, nous supposons que l'objet de la question ou *focus* joue un rôle sur la formulation de la réponse. Nous avons donc créé un panel de questions prototypiques dont le choix suit un certains nombre de variations autour de facteurs de type sémantiques.

Nombre de type concernés	% de présence dans le corpus Ritel
5	> 8 %
0	< 8 % et > 3 %
6	< 3 %

TAB. 2 – Présence des questions relevant des classes ontologiques de la GRINT dans le corpus RITEL

Pour les questions *quantité* nous avons pris en compte deux facteurs : le type de quantification/mesure (mesure de l'espace, du poids,...) et le type de réponse (fixe, variable - selon l'objet du monde considéré) ou dépendante d'un paramètre externe). Les détails de ces choix sont présentés dans (Garcia-Fernandez & Lailler, 2007).

Pour les questions *lieu* et *temps* nous avons ajouté un facteur concernant le verbe utilisé dans la question. Ainsi pour les questions *lieu*, les facteurs sont : le type de localisation (*où est, d'où part, où va,...*), le type de réponse (un lieu fixe ou plusieurs lieux) et le type de verbe (localisant ou bien neutre). Les questions prototypiques *Temps* sont présentées dans le tableau 3. Quant aux questions *temps*, les 3 facteurs se réalisent par le type de temporalité (quand est, depuis quand, jusqu'à quand,...), le type de réponse (fixe ou multiple) et le type de verbe (temporalisant ou bien neutre).

		Type de localisation	
		lieu où l'on est	autre
Type de réponse	fixe	Où est/se trouve la Joconde ?	Où est l'arrivée/arrive le Tour de France ?
	multiple	Où est/se trouve le Rhin ?	Où est l'estuaire/fini le Rhône ?

TAB. 3 – Questions prototypiques de *Lieu*

Les questions proposées sont simples, leurs réponses possibles supposées connues du plus grand nombre, afin de maximiser les réponses à valence positive (le sujet donne en effet une information réponse). En effet, les réponses du type *je ne sais pas* (ou à valence négative) sont moins intéressantes compte-tenu du but de l'étude.

## 2.2 Protocole multimodal

L'expérimentation a eu lieu à l'oral et à l'écrit. Une partie des questions a été posée aux mêmes sujets à la fois à l'écrit et à l'oral permettant ainsi d'observer les différences entre ces deux modalités dialogiques.

Malgré les différences inhérentes aux deux modalités, nous avons voulu proposer un protocole variant le moins possible entre les deux phases expérimentales. L'expérience à l'écrit a été mise en place sur une page web. Nous avons créé une interface *php* pour l'occasion. A l'oral, nous avons utilisé un dispositif déjà existant au sein du laboratoire (mise en place de lignes téléphoniques, système de détection de la parole - notamment pour détecter quand le locuteur a fini de répondre, synthèse de la parole - même modèle vocal pour l'ensemble des tests (Toney *et al.*,



2008). Les réponses des utilisateurs ont été transcrites automatiquement grâce à ce système. Ces transcriptions ont été corrigées manuellement.

Dans les deux cas, les sujets ont été exposés au même nombre de questions proposées dans un même contexte expérimental (cf. section 2.3).

## 2.3 Contexte expérimental et passation

Notre but n'est pas d'obtenir de vraies réponses aux questions. Par contre nous cherchons à observer la formulation de cette *information réponse*. Nous souhaitons inciter les sujets à formuler des phrases tout en suggérant un contexte expérimental cohérent avec la "simplicité" des questions soumises. Nous avons ainsi fait appel à des locuteurs natifs du français en leur demandant de répondre à des questions supposées posées par des enfants de CM2 préparant des exposés.

La passation des expériences a eu lieu en trois grandes phases. D'abord, nous avons monté une expérience portant uniquement sur les questions de quantité posées à l'écrit. Chaque sujet s'est vu poser un total de 18 questions. Et chaque forme variationnelle de question a été posée à au moins 3 sujets différents. Cette expérience nous a permis de tester et valider le principe expérimental tant dans son déroulement que dans ses résultats (Garcia-Fernandez & Lailler, 2007). Puis nous avons mené l'expérience correspondante à l'oral. Comme nous avons observé que l'expérience pouvait être allongée tant à l'oral qu'à l'écrit, nous avons décidé d'augmenter le nombre de question posées à chaque locuteur (passant ainsi de 18 à 25). Ceci a permis de diminuer notablement le nombre de sujets nécessaires et donc d'accélérer d'autant le processus de collecte de ces données. Enfin une phase concernant les questions *Lieu et Temps* à l'oral et à l'écrit a eu lieu.

Nous avons fait appel à près de 150 sujets. La passation a été organisée de façon à avoir au moins 4 sujets par *question test*. Ceci équivaut à une moyenne de 50 questions posées pour une forme morphosyntaxique donnée à l'écrit et autant à l'oral.

## 3 Analyses du corpus recueilli

Dans cette section, nous commençons par décrire le corpus d'un point de vue général puis nous présentons quelques analyses plus fines sur certains aspects.

### 3.1 Description générale

Le tableau 4 présente les caractéristiques principales du corpus global et des sous-parties orale et écrite.

La première constatation que l'on peut faire est que les réponses écrites sont nettement plus longues que les réponses orales. Ceci se remarque d'une part par le nombre total de mots présents dans les réponses (environ 17000 pour les réponses écrites contre environ 5000 pour les réponses orales) et par le nombre moyen de mots par réponse (8,18 à l'écrit contre 5,53 à l'oral). La taille du lexique des réponses écrites est 2,5 fois plus élevée que pour les réponses orales.

Le tableau 5 présente le nombre moyen de mot par réponse en distinguant selon la classe de

	Écrit	Oral
nb. mots total	<b>17395</b>	<b>5255</b>
taille lexicque	1536	609
moy. mot/réponses	<b>8,18</b>	<b>5,83</b>
nb. questions	2125	900
nb. questions différentes	507	434
nb. sujets	<b>101</b>	<b>47</b>
nb. sujets/question	<b>4,19</b>	<b>2,17</b>

TAB. 4 – Caractéristiques générales du corpus

	Quantité			Lieu			Temps		
	Écrit	Oral	Total	Écrit	Oral	Total	Écrit	Oral	Total
moy. mot/reponses	9,81	7,08	8,67	<b>6,81</b>	<b>2,70</b>	6,09	<b>7,46</b>	<b>3,11</b>	6,69

TAB. 5 – Taille moyenne (en nombre de mots) des réponses considérées selon la classe de la question et la modalité

la question (*Temps*, *Lieu* et *Quantité*). On peut faire la même remarque en ce qui concerne la différence écrit vs oral que précédemment : la longueur moyenne des réponses (en nombre de mots) est nettement plus importante pour l'écrit que pour l'oral et ce dans toutes les classes. Cette différence est particulièrement importante pour les questions *Lieu* et *Temps* (plus de deux fois plus de mots à l'écrit qu'à l'oral).

Les différences dans le nombre total de sujets (101 à l'écrit et 47 à l'oral) ainsi que le nombre de sujets par question (4,19 à l'écrit et 2,17 à l'oral) s'expliquent par un phénomène beaucoup plus important de rejet dans l'expérience orale que dans l'expérience écrite. Pour l'expérience orale, il a été fait appel à des sujets non contrôlés <sup>1</sup>). Ceci explique peut-être en partie pourquoi nous avons dû rejeter plus de données dans ces expériences. Le nombre de sujets nécessaires pour obtenir un corpus raisonnable a été choisi en fonction d'une première expérience à l'écrit. Ce nombre, les conditions de recrutement des sujets ayant changé, n'était en fait pas suffisant.

### 3.2 Analyse du lexique

Le tableau 6 présente les différents corpus en fonction du lexique général et des lexiques pour chaque modalité. L'objet de ce tableau est de nous permettre de répondre aux questions suivantes : (1) *Etant donné un lexique, quelle est la part de ce lexique utilisé dans les réponses des différentes classes ?*, (2) *Etant donné un lexique associé à une modalité, quelle est la part de ce lexique utilisé dans les réponses des différentes classes ?* et (3) *Le vocabulaire des réponses est-il varié ?*

Une première observation que l'on peut faire est que le lexique des réponses *Quantité* (1158) est presque deux fois plus grand que celui des réponses *Lieu* (708) et *Temps* (538). En ce qui concerne la part des lexiques effectivement utilisée dans les réponses, on constate tout d'abord que les réponses orales utilisent très peu des mots du lexique comparativement aux réponses

<sup>1</sup>Nous remercions le RISC et le site <http://faites-la-science.risc.cnrs.fr/> pour la base de données de volontaire qu'ils proposent

	Écrit			Oral			Total		
	Q	L	T	Q	L	T	Q	L	T
nb. mots	8674	5662	3059	4515	476	274	13189	6138	3333
taille lexique	932	682	518	554	122	97	<b>1160</b>	<b>708</b>	<b>538</b>
% du Lexique L	43.4	31.7	24.1	25.8	<b>5.7</b>	<b>4.5</b>	54.0	33.0	25.0
% du lexique LM	<b>60.7</b>	44.4	<b>33.7</b>	<b>90.6</b>	20.0	<b>15.8</b>	54.0	33.0	25.0
fréq moy rel.L	<b>9.3</b>	<b>8.3</b>	<b>5.9</b>	<b>8.1</b>	<b>3.9</b>	<b>2.8</b>	11.4	8.7	6.2
fréq moy abs L	2.3	3.1	4.1	3.9	17.5	22.1	1.8	3.0	4.0
fréq moy abs LM	1.6	2.3	2.9	1.1	5.0	6.2	1.8	3.0	4.0

TAB. 6 – Répartition du lexique (L) suivant les classes de questions et les modalités (LM)

écrites. En particulier les réponses des classes *Temps* et *Lieu* couvrent moins de 6% du lexique général. De façon systématique, pour toutes les classes, les réponses écrites couvrent une plus grande part du lexique général. Si on regarde le comportement selon les lexiques spécifiques à chaque modalité, on observe que les questions *Quantité* représentent une part importante de ces lexiques (60,7% pour les questions écrites et 90,6% pour les questions orales). Il n'en va pas de même pour les questions *Lieu* et *Temps*, en particulier ces dernières couvrent moins d'un tiers du lexique de la modalité. On constate aussi qu'il y a une moins grande réutilisation des mots à l'oral qu'à l'écrit. La ligne *fréquence moyenne relative au lexique L* du tableau 6 donne une indication sur le nombre de fois où un mot est utilisé. On peut remarquer que les mots sont plus souvent réutilisés à l'écrit (de 6 à 9 fois selon la classe) qu'à l'oral (de 3 à 8 fois). Il est intéressant de constater qu'à l'oral il y a une grande différence selon les classes (8,1 pour *Quantité*, 3,9 pour *Lieu* et 2,8 pour *Lieu*).

On peut également constater que les réponses de la classe *Quantité* forment un groupe, en terme de comportement, nettement différent des réponses des classes *Lieu* et *Temps*. Cette particularité va dans le même sens que l'observation selon laquelle la taille des réponses moyennes pour les questions *Quantité* est supérieure (1 à 3 mots de plus) à celle des deux autres classes de question. Nous expliquons ce phénomène par le fait qu'une quantité peut faire l'objet d'approximation, d'intervalle beaucoup plus facilement qu'un lieu par exemple. Les quantités (le poids, la distance,...) sont des éléments relativement difficiles à déterminer et l'estimation est toujours permise. Par contre, les questions *Lieu* et *Temps* portent sur des informations plus facilement déterminables. Et même si l'estimation est possible, il existe souvent un terme permettant de désigner l'estimation en question.

**Question :** *Où est le Rhin ?*

**Réponse précise :** *En France, en Allemagne,...*

**Réponse estimation :** *En Europe de l'ouest.*

	Écrit			Oral			Total		
	Q	L	T	Q	L	T	Q	L	T
taille lexique	932	682	518	554	122	97	1160	708	538
taille lexique commun	159			45			185		

TAB. 7 – Recouvrement des lexiques suivant les questions et les modalités

Le tableau 7 présente les tailles des différents lexiques de mots communs aux différentes classes de questions, à l'écrit, à l'oral et en général. On observe que seule une faible partie des diffé-



rents lexiques est commun quelle que soit la classe de la question. Ceci tient à la diversité des thèmes abordés par les trois classes de question. Nous avons observé plus en détails chacun de ces lexiques communs. On retrouve dans chacun une grande part de mots-outils (préposition, conjonction,...). Mais on y trouve aussi des verbes dont des auxiliaires, des verbes modaux et des verbes tels que “croire”, “penser”, “savoir” et ce quelle que soit la modalité considérée.

Le tableau 8 présente le nombre moyen des parties du discours (substantif, verbe et nombre) pour les sous-parties écrit et oral du corpus et détaille les moyennes selon la catégorie de la question. Nous avons ici sélectionné les 3 parties du discours qui permettent de distinguer au mieux les 6 classes (les 6 colonnes du tableau) au travers d’une analyse en composantes principales. On constate que les questions *Quantité* se distinguent encore une fois des deux autres catégories de question. Le nombre de chiffres moyen par réponse y est supérieur : il y a en moyenne plus d’une valeur numérique pour de telles questions (1,22 à l’écrit et 1.13 à l’oral) alors qu’il y en a moins d’une réponse sur deux pour les questions de *Lieu* et de *Temps* (respectivement 0.47 et 0.14 à l’écrit et 0.26 et 0.10 à l’oral). Cette différence doit être attribuée directement aux types des questions : une question de *Quantité* attend plus naturellement une valeur numérique qu’une question relevant de l’un des deux autres types de question. De plus le nombre de noms et de verbes y est lui aussi supérieur tant à l’écrit qu’à l’oral. Cependant, il est intéressant d’observer que le taux de verbe des questions *Quantité* est supérieur à l’oral qu’à l’écrit alors que la tendance est inverse concernant les deux autres types de question. D’autre part les taux des deux autres parties du discours, toujours pour les questions *Quantité*, restent stables alors même que la taille moyenne de la réponse est plus faible à l’oral (7 mots en moyenne, voir tableau 5). Il semble donc que les réponses à l’oral soient plus centrée vers un transfert précis de l’information (utilisation plus importante de mots pleins) plutôt que vers une formulation détaillée (utilisation de connecteurs, etc.).

	Écrit			Oral		
	L	Q	T	L	Q	T
Substantifs.	1.47	1.93	1.80	0.67	1.97	1.00
Verbes	0.71	0.95	0.89	0.41	1.34	0.57
Valeurs numériques	0.47	1.22	0.14	0.26	1.13	0.10

TAB. 8 – Effectifs moyens de présence de catégories : substantifs, verbes et valeurs numériques dans les réponses

## 4 Discussion et conclusion

Dans le cadre de nos travaux sur la génération complète en langue naturelle de réponses pour les systèmes interactifs de question-réponse, nous avons mis en place un protocole d’acquisition d’un corpus de production naturelle de réponse. Ce protocole a été décrit ainsi que le corpus. Le corpus collecté est à notre connaissance le seul de ce genre et en tout cas le seul de cette importance. Une première analyse a permis de mettre en évidence la diversité des réponses obtenues. Cependant nous avons mis en valeur des caractéristiques plus particulières dépendantes d’une part de la modalité de l’interaction (oral vs. écrit) d’autre part du type de la réponse. En particulier, aux questions *Quantité* une réponse plus détaillée, plus longue est fournie.

Le corpus collecté est extrêmement riche tant par sa taille que par les possibles comparaisons qui peuvent être mise en place. Il nous semble à présent essentiel de poursuivre l’analyse des

réponses et notamment en se centrant sur une analyse de la correspondance syntaxique entre les questions et les réponses.

## Références

- ANAYA G. & KOSSEIM L. (2003). Generation of natural responses through syntactic patterns. In *TALN*, p. 297.
- DE CHALENDAR G., DALMAS T., ELKATEB-GARA F., FERRET O., GRAU B., HURAUULT-PLANTET M., ILLOUZ G., MONCEAUX L., ROBBA I. & VILNAT A. (2002). The question answering system QALC at LIMSI, experiments in using web and wordnet. In *TREC*.
- GALIBERT O., ILLOUZ G. & ROSSET S. (2005). An open-domain, human-computer dialog system. In *InterSpeech*.
- GARCIA-FERNANDEZ A. & LAILLER C. (2007). Morphosyntaxe de l'interrogation pour le système de question-réponse RITEL. In *RECITAL*.
- LUZZATI D. (2006). Essai de description interactive : l'exemple des questions quantificatrices. In *Colloque La quantification*.
- PLAMONDON L., LAPALME G. & KOSSEIM L. (2002). The quantum question answering system at TREC 11. In *TREC*.
- ROSSET S. & PETEL S. (2006). The Ritel corpus - an annotated human-machine open-domain question answering spoken dialog corpus. In *LREC 2006*, Genoa, Italy.
- SOUBBOTIN M. (2006). Patterns of potential answer expressions as clues to the right answer. In *The 10th TREC Conference (TREC 2001)*.
- TONEY D., ROSSET S., MAX A., GALIBERT O. & BILINSKI E. (2008). An Evaluation of Spoken and Textual Interaction in the RITEL Interactive Question Answering System. In E. L. R. A. (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.
- VAN HOOIJDONK C., KRAHMER E., A.MAES, 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.