Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert

Anne Garcia-Fernandez

UNIVERSITÉ PARIS SUD 11 ORSAY
LIMSI-CNRS

THÈSE

présentée et soutenue publiquement en vu d'obtenir le grade de

Docteur spécialité « Informatique »

par

Anne GARCIA-FERNANDEZ

GÉNÉRATION DE RÉPONSES EN LANGUE NATURELLE ORALES ET ÉCRITES POUR LES SYSTÈMES DE QUESTION-RÉPONSE EN DOMAINE OUVERT

Thèse soutenue le 10 décembre 2010 devant le jury composé de :

M.	FRÉDÉRIC BÉCHET	Université Aix-Marseille et LIF	(Rapporteur)
M.	GUY LAPALME	Université de Montréal et RALI	(Rapporteur)
M.	DANIEL LUZZATI	Université du Maine et LIUM	(Examinateur)
M.	BERNARDO MAGNINI	Université de Trento, Italie et FBK	(Examinateur)
Mme	SOPHIE ROSSET	LIMSI-CNRS	(Directeur)
Mme	ANNE VILNAT	Université Paris Sud Orsay et LIMSI-CNRS	(Directeur)

Non importa quante anse ha un fiume, alla fine di un giusto corso, ogni acqua trova il suo mare.

Prediction is very difficult, especially about the future.
Niels Bohr
Danish physicist (1885 - 1962)

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

DANS un monde où l'accès à l'information est de plus en plus facilité et où la quantité d'information disponible s'accroît sans cesse (Internet, télécommunications,...), *trouver* l'information que l'on cherche vraiment est un souci à ne pas négliger. L'un des moyens d'accéder à l'information le plus répandu sur la Toile, est le moteur de recherche. Ces moteurs sont d'une efficacité incomparable, cependant la finesse des résultats obtenus n'est pas toujours idéale. Il faut bien souvent fouiller parmi les résultats renvoyés voire même reformuler sa requête. On parle bien de *requête*. Il s'agit pour la plupart du temps d'un ensemble de mots-clefs. Trois étapes qui peuvent se renouveler sont alors nécessaires à un internaute pour trouver l'information qu'il cherche : (1) la conception d'une requête, (2) la sélection d'un document, d'après son titre ou encore l'extrait proposé (3) la fouille du document pour trouver ou non l'information recherchée. Si le document ne contient pas l'information, alors un autre document peut être sélectionné et fouillé. Si aucun document ne contient d'information pertinente, il faut revenir à la première étape et reformuler sa requête. Dans cette démarche, nous voyons au moins deux éléments majeurs qui ralentissent l'accès à l'information. Le premier est que la réponse du moteur de recherche n'est pas précise : il faut fouiller par soi-même parmi les documents pour trouver ce que l'on cherche. Le second est que la "requête" n'est pas précise, de ce fait les documents ne correspondent pas toujours à ce qui est recherché, ce qui nécessite de la reformuler, c'est-à-dire de rédiger une autre requête.

CONTEXTE DE TRAVAIL

En réponse à ces limites, un autre moyen de faire de la recherche dans des documents a été proposée : **la réponse à une question**. Il s'agit d'une interaction permettant à l'internaute (ou plus généralement, à l'utilisateur) de poser une question précise à laquelle la réponse apportée est rendue directement visible. C'est ce type d'interaction que les systèmes dits de *question-réponse* se proposent de gérer. L'utilisateur ne doit plus formuler une requête mais peut poser une question. Le système quant à lui renvoie une information précise (un mot, un groupe de mots) accompagnée de l'extrait du document au sein duquel l'information a été trouvée ainsi qu'un lien vers ce même document. L'utilisateur peut donc toujours vérifier, valider l'information. Mais, comme dans le cas des moteurs de recherche, la bonne information n'est pas toujours trouvée "du premier coup". Une reformulation de la question peut être nécessaire. Le système peut aussi avoir fait une erreur d'*interprétation* de la question et renvoyer ainsi une information non pertinente. En effet, le système de question-réponse devant traiter une question et non pas une requête, sa tâche est plus complexe que celle du moteur de recherche : il doit *comprendre* la question. Par exemple, il doit identitifer dans la négation que l'util-

isateur recherche un hôtel qui *n'*est *pas* dans Paris plutôt qu'un hôtel *à* Paris. C'est bien sûr là que réside un des intérêts majeurs des systèmes de question-réponse mais aussi la complexité de la tâche qu'ils doivent accomplir. En cas de problème d'interprétation, si la réponse du système se limite à un mot ou groupe de mot, l'utilisateur n'a que peu de matière pour identifier le problème, aider le système à se corriger ou encore lui proposer une autre question plus adaptée, plus précise.
Comme les moteurs de recherche, certains systèmes de question-réponse interagissent **à l'écrit** mais, au sein des recherches autour de ces systèmes, un autre aspect semble plus intéressant encore : l'interaction orale. **À l'oral**, l'expression d'une requête est peu concevable. Il serait difficile de demander à un utilisateur de dire "hôtel Paris pas cher" alors qu'une demande telle que "je cherche un hôtel pas cher à Paris" est naturelle et facile à formuler. Des systèmes sont développés pour reconnaître la parole et transcrire ce que l'utilisateur dit. Ensuite, le travail de "question-réponse" se déroule comme à l'écrit. Finalement, la réponse est synthétisée oralement et dans ce cas, si elle peut éventuellement être accompagnée de la lecture de l'extrait du document dans lequel l'information a été trouvée, elle ne peut *pas* être accompagnée d'un lien vers le document. L'utilisateur a alors très peu d'information à sa disposition pour juger de la validité de la réponse. Ceci est d'autant plus important compte-tenu du fait que des erreurs de reconnaissance vocale s'ajoutent aux erreurs éventuelles d'interprétation de la question (voire même peuvent en provoquer).

PROBLÉMATIQUE

Pour palier ces limitations, une évolution possible des systèmes de question-réponse serait de proposer une réponse plus complète : sous forme de phrase par exemple. Une phrase peut permettre de présenter l'information sous forme plus naturelle tout en donnant à voir la compréhension du système : elle peut reprendre certains mots de la question ou reformuler la question exactement comme lorsqu'on demande à un enfant de reformuler la question qu'on lui a posée avant d'y répondre afin d'observer si il a bien compris ce qu'on lui demande. Nous voyons le fait de répondre à une question, non plus comme la tâche de trouver l'information qui correspond à la bonne réponse mais comme la tâche de produire une réponse en interaction avec l'utilisateur.

PRINCIPALES CONTRIBUTIONS

Notre travail a ainsi comme thématique principale la génération de réponse en langue naturelle à l'oral et à l'écrit pour les systèmes de question-réponse. Après avoir étudié ce qu'est une interaction de type question-réponse, nous avons observé les réponses données par les systèmes de question-réponse et par des humains (sur des sites Internet collaboratif de question-réponse). Nous avons défini les notions de *réponse*, d'*information-réponse* et de *réponse en interaction avec la question* et avons constaté l'absence de corpus de réponses en interaction qui puissent être reproduites par des systèmes de question-réponse. Nous avons ainsi mis en place une expérience permettant la collecte d'un corpus de réponses de la façon suivante :

- des locuteurs humains répondent à une série de questions, le contexte proposé aux participants les incitant à être spontané dans leur façon de répondre ;

– les questions posées ont des caractéristiques linguistiques précises : nous inspirant de travaux montrant le lien qui existe entre le forme d'une question et la forme de sa réponse, nous avons voulu avoir un contrôle fin sur les questions utilisées ;

Dans l'optique de fournir aux travaux menés dans le domaine des systèmes de question-réponse des indications sur la forme que peut prendre une réponse en interaction avec la question, nous avons analysé le corpus collecté en répondant aux questions suivantes :

– est-il judicieux de réutiliser des éléments de la question ?
– quelle information faut-il donner en réponse ?
– comment construire une réponse minimale ?
– la modalité (oral/écrit) influe-t-elle sur la formulation de réponse à générer ?
– euh... et si on hésite ?
– que répondre quand on n'a pas de réponse ?

PLAN DE LA THÈSE

Nous présentons tout d'abord la réponse à une question comme une interaction en langue naturelle (chapitre 1). Puis nous nous attachons, par une observation des réponses fournies par des humains et par des systèmes, à définir ce qu'est une réponse en interaction avec la question (chapitre 2). Ces observations mettent en valeur d'une part le lien qui existe entre la forme d'une question et celle de la réponse, d'autre part le manque de corpus de réponses en interaction qui soient reproductibles par les systèmes de question-réponse. Nous arrivons ainsi à la constatation qu'il nous faut collecter notre propre corpus et que les réponses collectées doivent correspondre à des questions dont la forme est contrôlée. Nous présentons ensuite un ensemble de recherches sur la classification des questions (chapitre 3) ce qui nous permet d'identifier des paramètres de variations de la question et donc de décrire la forme d'une question de façon très précise. Ceci nous permet de construire un corpus de question dont les caractéristiques sont contrôlées. Ces questions ont été soumises à des humains au cours d'une expérience de collecte de corpus. Cette collecte ainsi que l'évaluation du protocole expérimental mis en œuvre sont présentés (chapitre 4). Dans une dernière partie, différentes analyses du corpus sont exposées (chapitre 5). Elles répondent à un ensemble de questions que l'on doit se poser si l'on veut adapter un système de question-réponse afin qu'il fournisse une réponse en interaction.

Première partie

De la réponse précise à la réponse en interaction

1

RÉPONDRE À UNE QUESTION, UNE INTERACTION EN LANGUE NATURELLE

SOMMAIRE

Ce chapitre présente une vue d'ensemble des travaux scientifiques existants dans le domaine de la réponse à une question (ou *question-réponse*). Il permet d'introduire au lecteur l'objet d'étude des travaux présentés et conclut sur la thèse que nous défendons dans le présent document. Le chapitre est organisé en trois sections. À partir de la notion d'interaction (section 1.2), on spécialise l'objet d'étude à l'interaction en langue naturelle (section 1.3) puis à l'interaction de type question-réponse (section 1.4). On cherche à montrer au lecteur ce que sous-entend travailler avec la langue naturelle par un survol des thématiques et enjeux principaux du domaine. On expose aussi, de façon grossière, ce que sont les systèmes de question-réponse, leur fonctionnement et plus particulièrement, leurs limites actuelles du point de vue de l'interaction. Une comparaison entre oral (ou parole) et écrit (ou texte) est proposée tout au long du chapitre. L'exposé conclut sur l'importance et l'intérêt pour les systèmes de question-réponse de produire une réponse en langue naturelle.
Ce chapitre correspond à un travail d'état de l'art et de positionnement scientifique.

This chapter provides an overview of scientific work in the domain of question answering system. It is useful to introduce the domain, in particular to the out-of-domain reader and concludes with the thesis put foward in this document. The chapter is organised in three sections. From the notion of interaction (section 1.2), the object of study is specified to the interaction in natural language (section 1.3) and then to the question answering interaction type (section 1.4). We show what is working with the natural language by a survey of the domain and present the question answering systems. A comparison between speech and writing is done all along the chapter. We concludes on the importance and usefulness for question-answering systems of producing answers in natural language.

1.1 INTRODUCTION

Nombre de travaux tentent de fournir de façon automatique une réponse à une question. Parmi eux, la plupart s'intéressent au problème de trouver une réponse à la question du point de vue de la recherche de l'information qui répond en effet à la question. Le défi étant, étant donné une question et un ensemble de documents, d'extraire des documents la *bonne réponse* à la question. Nous adoptons un point de vue différent. Nous considérons le fait de répondre à une question comme une interaction entre deux locuteurs (en l'occurrence, un homme et une machine). Ce chapitre présente donc la réponse à une question de façon originale en s'attachant d'abord à l'interaction entre l'homme et la machine (1.2), puis à l'interaction en langue (1.3) et enfin en se centrant sur l'interaction du type "répondre à une question" (1.4).

1.2 INTERACTION ENTRE L'HOMME ET LA MACHINE

Actuellement les machines sont partout autour de nous. Interagir avec elles peut se faire de différentes façons. Pour ne citer que quelques exemples, on peut penser à des interactions qui vont de l'appui sur le bouton du réveil-matin à l'utilisation du tout nouveau téléphone portable "toutes options comprises" en passant par l'utilisation des machines tactiles d'une gare pour la réservation d'un billet ou encore une centrale de réservation automatique par téléphone. Ces façons d'interagir divergent par le but de l'interaction, la forme qu'elle prend, sa modalité... Cette dernière peut être tactile, visuelle, auditive, ou encore passer par un périphérique spécifique (tel qu'un clavier, une souris, ...). [Bellik et al., 1995] traite en particulier d'interface de réalité virtuelle. Certaines interactions peuvent aussi être multimodales [Karsenty, 2006; Horchani, 2007]. Ainsi les agents conversationnels animés que l'on trouve sur certains sites Internet [Chatbot.org, site web] proposent une interaction par le clavier (écrite) et visuelle puisqu'ils s'animent et proposent une réponse à lire sur l'écran. On remarque d'ailleurs que les modalités d'interactions sont différentes en entrée et en sortie.
Au delà de la modalité, la façon dont une modalité est utilisée diffère aussi. Par exemple, un agent conversationnel "répond" par un affichage textuel, l'affichage d'un graphique ou encore la navigation vers une autre page du site Internet. Plus finement encore, le texte affiché peut être un lien hypertexte, un mot ou une liste de mots ou encore une phrase. Ce texte peut être dans différentes langues, relever de différents niveaux de langue mais aussi être mis en forme de différentes façons. L'application pour laquelle l'interaction est mise en place joue un rôle crucial dans le choix de la modalité d'interaction et la forme qu'elle prend. Par exemple, un agent virtuel sur un site Internet commercial utilisera un niveau de langue relativement soutenu en naviguant vers les pages des produits proposés alors que s'il a des visées pédagogiques, il serait certainement plus judicieux qu'il forme des phrases complètes et simples en utilisant un vocabulaire relevant de la langue courante.

Nous n'entrerons pas dans les détails de l'ensemble des possibilités d'interaction évoquées ici. Le travail de thèse présenté dans ce document s'attache à un type d'interaction bien particulier puisqu'il s'agit d'une interaction qui met en jeu la langue.

1.3 INTERACTION EN LANGUE NATURELLE

Qu'est-ce qu'une langue ? Et pourquoi parle-t-on de "langue naturelle" ? Parce que ce terme ne relève pas de la langue courante, nous en proposons ici une définition correspondant à l'acception que nous en avons :

Définition 1.1 ***Langue naturelle*** *: les langues ou langues naturelles sont à opposer aux langages (ou langues artificielles). Les langues du monde tel que le français, l'anglais, la langue des signes ou encore les créoles sont considérés comme des langues naturelles. Les langages sont quant à eux des langues créées par l'homme comme par exemple les langages informatiques.*
Dans la communauté du traitement automatique des langues naturelles (en anglais, Natural Language Processing), une requête (par exemple, un ensemble de mots-clefs donné à un moteur de recherche par exemple) n'est pas considérée comme étant en langue naturelle.

Un système capable d'interagir en langue naturelle est un système qui traite de la langue naturelle en entrée et/ou produit de la langue naturelle en sortie comme l'illustre la figure 1.1. Il doit alors effectuer des traitements spécifiques à la langue naturelle.

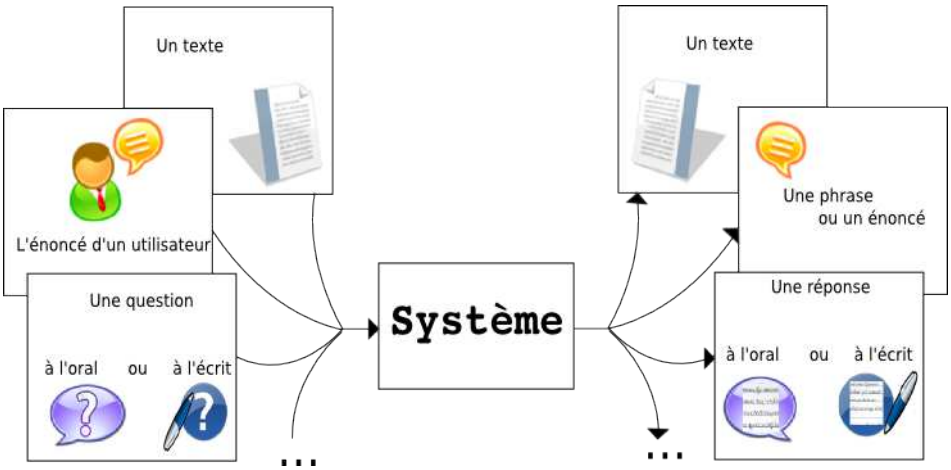

FIGURE 1.1 – *Exemples d'entrées et sorties possibles pour un système d'interaction en langue naturelle*

Les intérêts d'interagir en langue naturelle sont multiples. D'abord, il s'agit d'utiliser une langue que nous, humains, nous connaissons déjà. Nous n'avons pas besoin de nous adapter à un système ou d'apprendre un nouveau language ou une nouvelle façon d'interagir. C'est le système qui s'adapte à nous. D'autre part, les langues naturelles sont extrèmement puissantes et expressives. Lorsque nous interrogeons un moteur de recherche avec des mots-clefs, nous utilisons bel et bien des mots de la langue mais nous ne produisons pas de la langue naturelle, nous nous contentons d'aligner des mots les uns à la suite des autres. Ceci n'est ni aussi expressifs que de formuler une phrase, ni naturel (nous ne nous exprimons pas de façon spontanée ainsi).
Pourtant "apprendre" à une machine à utiliser la langue naturelle n'est pas im-

médiat. Les sections suivantes présentent l'enjeu de traiter une entrée en langue naturelle et celui de produire en sortie de la langue naturelle.

1.3.1 Entrée en langue naturelle

Prenons l'exemple d'un des premiers systèmes de dialogue en langue naturelle entre l'homme et la machine : Eliza. Ce système, développé par Joseph Weizenbaum dès 1964 [Weizenbaum, 1966], simule un psychothérapeute en reformulant la plupart des énoncés de l'utilisateur (le "patient") sous forme de questions et en les lui posant. L'interaction se fait à l'écrit. Le dialogue 1.2 en est un exemple. Les phrases en entrée sont décomposées en fonction de règles déclenchées par des mots-clefs. La sortie est produite en recomposant certains éléments de la phrase en entrée à des patrons prédéfinis. La figure 1.3 présente un exemple d'une telle décomposition/recomposition.

Men are all alike.
IN WHAT WAY ?
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE ?
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
...

FIGURE 1.2 – *Exemple de dialogue avec le système ELIZA, extrait de [Weizenbaum, 1966] - Les lignes capitalisées sont les réponses de la machine - Une traduction vers le français de ce dialogue est disponible en annexe A.2*

Analyser le texte entré par l'utilisateur signifie, dans le cas de ce système, repérer les mots importants ou mots-clefs et identifier le contexte minimal dans lequel les différents mots-clefs apparaissent. En fait, ce système a très peu de connaissances sur la langue. Il connaît un ensemble de mots-clefs. Ce qu'il ne connaît pas, il l'utilisera tel quel dans sa réponse sans même chercher à comprendre ce dont il s'agit. Par exemple, si l'utilisateur écrit la phrase "Je suis déprimé en ce moment", le système fonctionne de la façon suivante : les mots "je" et "suis" sont connus du système. Il utilisera des règles de transformation pour traiter ces mots. En revanche, "déprimé en ce moment" n'est pas une suite de mots connus. Cette séquence sera donc laissée telle quelle. Une réponse du système peut ainsi être :

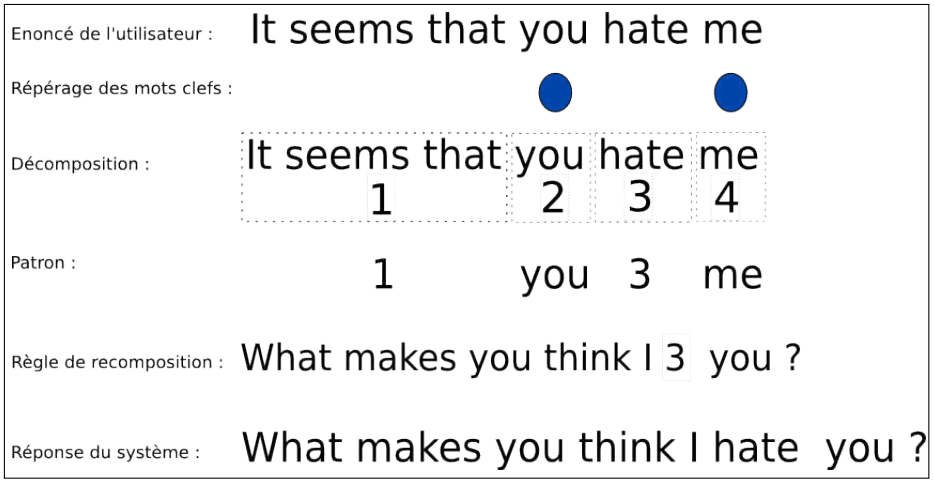

FIGURE 1.3 – *Exemple de décomposition puis recomposition au sein du système ELIZA basé sur l'exemple donné dans [Weizenbaum, 1966]*

"Je suis désolé de savoir que vous êtes déprimé en ce moment" dans laquelle "Je suis désolé de savoir que" appartient à un patron de réponse prédéfini. "vous êtes" résulte d'une règle de transformation à partir de la séquence de mots connus "je suis". Et "déprimé en ce moment" est simplement une réutilisation telle quelle de la séquence de mots inconnus. Mais cette règle est aussi valable si l'utilisateur dit "je suis content en ce moment" et alors, dans ce cas, le système peut, en utilisant la même décomposition/recomposition que précédemment, répondre : "je suis désolé de savoir que vous être content en ce moment".

Cet exemple montre l'importance d'avoir des connaissances sur la langue. Ces connaissances sont à la fois lexicales (besoin de connaître les mots de la langue) et syntaxiques (ou grammaticales) notamment pour transformer la séquence "je suis" en "vous êtes". Des connaissances sémantiques sont aussi utiles permettant ainsi de produire l'une ou l'autre des réponses suivantes le cas échéant : "je suis désolé de savoir que vous êtes déprimé en ce moment" ou bien "je suis content de savoir que vous êtes bien en ce moment".

Eliza est l'un des premiers systèmes traitant de la langue naturelle en interaction. Il n'a que très peu de connaissances, n'est disponible qu'à l'écrit et n'est pas capable de coopération. À présent, les systèmes ont à leur disposition des connaissances bien plus larges : des dictionnaires incluant toutes les formes que peuvent prendre les mots de la langue (par exemple, l'ensemble des formes conjuguées d'un verbe), des connaissances syntaxiques leur permettant d'analyser des phrases complexes (comprenant une proposition subordonnée par exemple) ou encore d'identifier les relations entre les différents groupes d'une phrase. Certains systèmes sont aussi capables de traiter de la langue orale. Par exemple, un système de reconnaissance vocale utilisé au LIMSI [Gauvain et al., 2005] est capable de transcrire de la parole spontanée c'est-à-dire une production orale qui n'a pas été préparée ([Lamel et al., 2005] en décrit une utilisation pour la parole spontanée, [Galibert et al., 2005c] en décrit une utilisation pour la question-réponse en domaine ouvert et [Lamel et al., 2000] en décrit une utilisation en domaine limité).

Spécificités liées à une entrée orale

Lorsque l'interaction est orale, les premiers traitements ont pour but de transformer un flux audio en une suite de mots. Il s'agit de reconnaître les mots prononcés par l'utilisateur. On parle de *reconnaissance vocale* [Jurafsky & Martin, 2006; Rabiner & Schafer, 1978; Huang et al., 2001]. Une première étape peut être simplement de détecter quand l'utilisateur parle, notamment dans un environnement bruyant. On parle alors de *détection de la parole* [Gauvain et al., 2000]. Des travaux portent aussi sur la séparation des différents flux audio afin d'isoler une voix parmi d'autres ou bien de l'isoler d'un fond sonore (musique,...). On parle alors de l'*effet cocktail party* [Darrell et al., 2000]. Une seconde étape est de détecter les mots prononcés, les reconnaître et ainsi de proposer une transcription du message audio.

Lorsqu'un utilisateur parle au système, il n'a pas forcément (et bien souvent c'est le cas) préparé sa phrase à l'avance, il n'est pas non plus en train de lire un document. Il parle "comme ça lui vient". On parle alors de *parole spontanée* [Shriberg, 2005]. Ce type de discours comporte des hésitations ("euh", "ben",...), des reprises ("je voudrais euh j'aimerais savoir quand..."). Les phrases ou les mots peuvent être coupés, parce que le locuteur hésite, choisit une autre formulation ou bien parce qu'il a un hoquet, mal à la gorge,... On parle de disfluences [1].

L'*intonation* change d'une phrase à l'autre, d'un locuteur à l'autre [Beaugendre, 1994; Morel & Danon-Boileau, 1998; Rossi, 1999] de même que le rythme et la vitesse de parole aussi. Les locuteurs ont des accents différents et prononcent les mots de façons différentes [Woehrling, 2009]. Des *modèles de la langue parlée*, qui peuvent éventuellement être adaptés au locuteur ou encore à la tâche , sont utilisés pour obtenir à partir du signal sonore une suite de mots, et ce en faisant le moins d'erreurs possible. La plupart des logiciels commerciaux de reconnaissance vocale demandent une phase de configuration pendant laquelle l'utilisateur doit lire un certain nombre de textes qui lui sont imposés. C'est à partir de ces données d'entraînement que le modèle de langage est adapté au locuteur. L'adaptation à la tâche consiste par exemple à *favoriser* certains champs lexicaux par rapport à d'autres en entraînant un système d'apprentissage automatique sur des données d'un domaine thématique particulier. Dans le cas d'un système d'aide à la réservation de billets de train, par exemple, des termes comme "prix", "matin" ou encore des noms de ville seront *favorisés* dans le modèle de langage par rapport à des termes comme "histoire", "rouge" ou encore des noms d'objets domestiques qui eux ont très eu de chance d'être prononcés par l'utilisateur au cours de l'interaction [2].

Spécificités liées à une entrée écrite

Prenons le cas où l'utilisateur - parce qu'il a saisi son texte rapidement, sans le relire - écrit "J esuis déprimé". Nous voyons ici qu'il a commis une faute de frappe. Pour l'humain, il est facile de repérer la faute, l'expliquer et même la corriger. Le système quant à lui doit avoir à sa disposition un lexique des mots du français et peut alors repérer que les deux "mots" n'appartiennent pas à son lexique. Des règles de *corrections typographiques et orthographiques* peuvent alors être utilisées afin de corriger l'erreur. Dans ce cas, la règle à utiliser sera d'échanger l'espace avec l'un des deux caractères qui l'entoure et de tester l'existence des mots ainsi

1. [Guénot, 2005] en propose une typologie précise.

2. En réalité, les modèles sont plus complexes et notamment se basent sur des probabilités de succession de mots. Pour des raisons de clarté, nous ne détaillons pas plus ce point.

formés dans le dictionnaire. D'autres règles existent pour résoudre d'autres types d'erreurs typographiques. Les erreurs peuvent être aussi orthographiques ou grammaticales [Naber, 2003; Souque, 2008]. La langue écrite suit en principe des règles de grammaire strictes. Elles ne sont cependant pas forcément respectées. D'autre part, le niveau de langue utilisé joue un rôle dans la qualité du message écrit par l'utilisateur [Authier & Meunier, 1972]. Un cas bien connu est celui de la langue utilisée pour rédiger des textos (souvent dénommé *langage sms*) : l'orthographe et la construction des phrases suivent des règles spécifiques (voir [Fairon et al., 2006]) bien différentes de celles qui régissent la langue écrite.

Traitements communs aux deux modalités en entrée

Qu'il s'agisse de parole ou de texte, une première étape est donc d'obtenir une suite de mots "propre", sans erreur ni coquille et, pour le cas de l'oral, la plus proche possible de ce qui a été prononcé par l'utilisateur. Mais le message en langue naturelle contient des ambiguïtés et les systèmes doivent les prendre en compte. Une phrase bien connue des linguistes est la suivante : "La belle ferme le voile". Elle peut être interprétée de deux façons : ou bien une jolie ferme cache (ou "voile") quelque chose, ou bien une femme ferme un rideau (ou "un voile"). Il y a une *ambiguïté sémantique* du sens de la phrase mais aussi de la façon dont cette phrase est construite (*ambiguïté syntaxique*), et des mots dont elle est composée (*ambiguïté lexicale*) : "belle" est-il un adjectif ou un nom ? "ferme" est-il un nom ou un verbe ? Et ainsi de suite. Nous voyons ici qu'au delà de reconnaître un mot comme appartenant au dictionnaire du système, il est nécessaire de le caractériser : quelle est sa catégorie grammaticale [3] ? Quel est son rôle dans la phrase ? S'il relève de la partie du discours (ou *part-of-speech* en anglais) "déterminant", à quel nom se rapporte-t-il ? Le groupe nominal ainsi formé est-il le sujet de la phrase, un complément,... ? Le découpage en groupe (nominal, verbal, prépositionnel, etc.) s'appelle l'analyse en constituant. La détection du rôle syntaxique des différents groupes et leur mise en relation s'appelle l'*analyse syntaxique* et cherche à construire un arbre syntaxique des phrases. La figure 1.4 en montre des exemples et en particulier montre les deux arbres syntaxiques possibles pour la phrase "La belle ferme le voile". On observe qu'en fonction de la détection des parties du discours, le chunking et l'analyse syntaxique diffèrent.

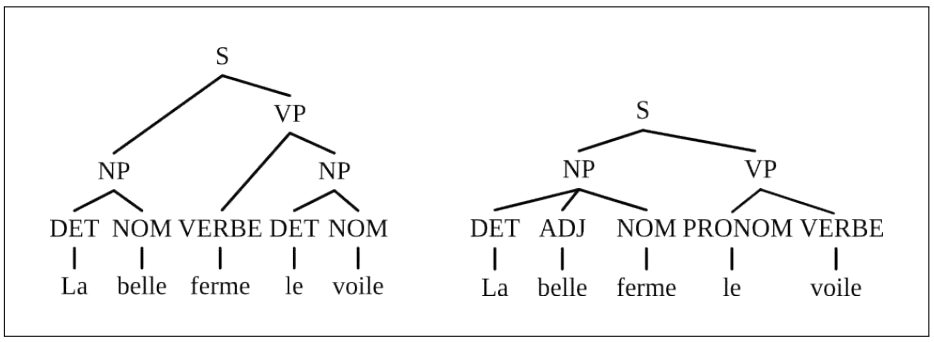

FIGURE 1.4 – *Deux arbres syntaxiques possibles pour la phrase La belle ferme le voile*

À partir d'une telle représentation, on peut construire une représentation sé-

3. On parle en linguistique de partie du discours.

mantique de la phrase [4]. La figure 1.5 montre deux représentations sémantiques possibles de la phrase “la belle ferme le voile” sous forme d’arbre de dépendance sémantique.

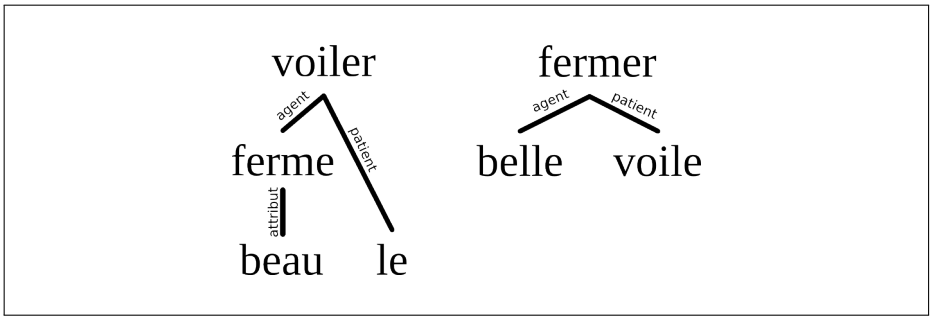

FIGURE 1.5 – *Deux arbres sémantiques possibles pour la phrase La belle ferme le voile*

Dans cet exemple, l’ambiguïté est syntaxique et sémantique. Mais elle peut être uniquement sémantique. Prenons une autre phrase bien connue des linguistes : “L’avocat est bon”. Est-ce qu’il s’agit d’un fruit goûtu ou bien d’un professionnel de qualité ? Et pourtant les deux phrases ont des syntaxes équivalentes. L’ambiguïté vient du fait que les mots “avocats” et “bon” peuvent renvoyer à plusieurs éléments du monde (ou *référent*). On parle de polysémie dans le cas de “bon” car, qu’il s’agisse du sens “goûtu” ou “de qualité”, le mot est le même. En revanche, l’avocat (fruit) et l’avocat (homme de loi) sont des homonymes. L’arbitraire de la langue a fait que ces deux mots bel et bien différents s’écrivent et se prononcent de la même façon.

Résoudre ces ambiguïtés peut se faire de différentes façons. Nous n’aborderons pas ici ces différentes techniques mais l’exposé des différents cas d’ambiguïté permet de mettre en évidence que des processus apparemment simples pour l’humain tel que corriger, interpréter, résoudre des ambiguïtés peuvent être autant d’obstacles pour des systèmes de *Traitement Automatique de la Langue* (TAL).

Ces traitements permettent une meilleure “compréhension” de la langue et offrent ainsi la possibilité d’un traitement plus adapté par le système. L’éventail des possibilités de traitement s’agrandissant, les systèmes proposés aux utilisateurs peuvent être de plus en plus intéressants. Jusqu’à concevoir, par exemple, de véritables systèmes de dialogue entre l’homme et la machine. Les systèmes doivent non seulement être capables de comprendre l’entrée en langue naturelle donnée par l’utilisateur et de la traiter (en extraire les informations importantes par exemple) mais aussi de produire une sortie en langue naturelle. La section suivante aborde ce point particulier.

1.3.2 Sortie en langue naturelle

Pour produire de la langue naturelle en sortie, on utilise le terme de *génération* (ou *génération automatique de texte*, GAT ; en anglais *Natural Language Generation*, NLG). En fonction du système et du contexte la sortie peut être un mot, une suite de mots, un *syntagme* (groupe grammatical), une phrase ou un texte. Par exemple, s’il s’agit d’un système de traduction automatique et si l’entrée est un texte, on

4. Il existe différents formalismes tant pour la représentation syntaxique que sémantique. Les arbres proposés ici ne sont que des exemples.

s'attend aussi à un texte en sortie. Des exemples d'application sont la génération de questions, la traduction automatique, la production de résumés automatiques ou encore la génération de réponses.

Le travail de génération en langue naturelle peut, à première vue, sembler bien plus abordable que celui du traitement d'une entrée en langue naturelle. En effet, on n'a plus à traiter des erreurs commises par l'utilisateur, des cas d'ambiguïté, des phrases complexes, etc. Mais suffit-il de connaître des règles de grammaire pour produire une phrase ? Si l'on cherche à produire une phrase grammaticale, peut-être. Mais si l'on cherche à produire une phrase ayant du sens, il faut savoir comment s'utilisent les mots de la langue et éviter de dire, par exemple, "amener un gâteau" au profit de "apporter un gâteau". Dans cet exemple, ceci implique d'avoir des connaissances sémantiques tant (1) au niveau de l'utilisation des verbes "amener" et "apporter" (le premier doit avoir comme complément d'objet direct un être vivant tandis que le second doit être complété par un objet) que (2) sur la nature du référent du monde désigné par "gâteau" : il s'agit d'un objet et non pas d'un être vivant.

[Abeillé et al., 2000, Chapitre 14] souligne le fait que la tâche de génération en langue naturelle implique de produire des énoncés "grammaticalement corrects, sémantiquement cohérents et pragmatiquement pertinents". Il faut donc à la fois connaître la grammaire de la langue, disposer d'un vocabulaire et du sens des mots de langue mais aussi être pertinent en fonction du contexte au sein duquel la sortie doit être produite.

En fait, le travail de génération de texte est généralement scindé en deux étapes [Danlos, 1989a]. Le "quoi dire ?" consiste à déterminer le contenu sémantique du texte à générer tandis que le "comment le dire ?" consiste à l'exprimer en langue naturelle. Nous présentons ici un survol de la production en langue naturelle en montrant les spécificités pour les deux modalités d'interaction oral et écrit.

Produire ou générer ?

Produire une sortie en langue naturelle ne signifie pas forcément la *construire*, la créer. Il peut s'agir seulement de renvoyer du texte préexistant. Par exemple, lorsqu'un utilisateur se déplace sur une page Internet qui n'existe pas, alors s'affiche un message indiquant que la page n'existe pas et précisant le numéro d'erreur "404". Le texte du message préexiste et est renvoyé car il correspond au message à afficher dans le cas où l'erreur numéro 404 survient. Ce type de traitement existe dans de nombreux programmes et permet de donner un retour compréhensible à l'utilisateur, même non expert, lorsqu'un problème ou une erreur survient au cours du déroulement de ce programme. Il s'agit simplement de mettre en correspondance un code d'erreur (un numéro qui désigne l'erreur ou le type d'erreur) et un *message pré-écrit* (ou "canned-text" en anglais). Ce dernier est renvoyé à l'utilisateur.

Un cas similaire est celui des serveurs vocaux par téléphone. L'utilisateur interagit en saisissant des numéros sur son clavier de téléphone. L'ensemble des interventions du système sont quant à elles pré-enregistrées.

Dans ces deux cas, le fait que le retour du système soit en langue naturelle est *presque* un détail [5]. Il pourrait s'agir, dans le cas du serveur vocal, de musique, de bruit,... ou encore, dans le cas de la page d'erreur, d'image, de symbole, de tableau.

5. Nous verrons dans les sous-sections suivantes que ceci implique toutefois des "précautions" particulières à la langue naturelle.

La sortie en langue naturelle n'est pas construite, ni créée par le système. C'est un expert humain (celui qui a créé le programme ou encore celui qui l'a commandé) qui décide du contenu du message et de sa formulation. Concernant l'erreur 404, l'administrateur du site peut choisir d'informer l'internaute de l'inexistence de la page ou bien de le rediriger vers une autre page : il choisit *quoi dire*. Ensuite, l'administrateur peut choisir la façon dont le message est exprimé, formulé : "La page désirée n'est pas accessible", "Cette page n'existe pas. Cliquez ici pour revenir à la page précédente", ... Il choisit ainsi *comment dire* son message.
Dans le cas de l'utilisation de messages pré-écrits ou pré-enregistrés, nous parlons de production en langue naturelle et non pas de génération. Le système ne décide ni du *quoi dire* (le contenu), ni du *comment dire* (la formulation en langue naturelle). Il ne fait que renvoyer un message en langue naturelle.
Les messages pré-enregistrés doivent être prévus à l'avance et constituent une liste fermée (non infinie) qui est élaborée en fonction du contexte. Plus l'application est précise et limitée, plus restreinte sera la liste des messages. Mais si le système est face à un grand nombre de cas de figures, constituer une liste exhaustive peut s'avérer fastidieux voire impossible.

Un système peut produire de la langue naturelle de façon "hybride" à la fois en utilisant du texte pré-enregistré et en insérant de la variation au sein de ces textes. C'est le cas par exemple de systèmes tels qu'une messagerie automatique qui utilise des schémas de réponses pré-fabriqués et les adapte au contexte. Un tel schéma est appelé "patron" (ou "templates" en anglais). Les exemples 1.1 à 1.4 montrent que ces patrons peuvent prévoir un ou plusieurs "emplacements libres" ou *variables* qui peuvent soit désigner un élément précis (le nombre de messages reçus, le numéro du message en cours) soit contraindre le type d'élément qui peuvent s'insérer dans le patron. Les contraintes peuvent être syntaxiques (un adjectif, un nom), sémantiques (une expression de temps). De même, plusieurs contraintes peuvent s'appliquer à la même variable. En effet, comme le montre l'exemple 1.3, la variable `STATUT_MESSAGE` ne peut correspondre qu'à un adjectif masculin singulier. D'autre part, une même variable peut entrer en jeu dans différents patrons. Lors de l'utilisation d'un patron, les variables sont remplacées par leur valeur comme le montrent les exemples 1.5 et 1.6.

Exemple 1.1 *Vous avez `NB_MESSAGES` message(s).*

Exemple 1.2 *Message `NUMERO_MESSAGE` sur `NOMBRE_MESSAGE` reçus à `HEURE_MESSAGE`.*

Exemple 1.3 *Le message est `STATUT_MESSAGE:adjectif,masculin,singulier`.*

Exemple 1.4 *Il vous reste `DURÉE_RESTANTE` pour valider votre inscription.*

Exemple 1.5 *Il vous reste `8 jours` pour valider votre inscription.*

Exemple 1.6 *Il vous reste `1 mois` pour valider votre inscription.*

L'avantage d'une telle approche est qu'un patron peut donner lieu à de nombreux messages différents. Il suffit alors d'étendre la liste des valeurs que prennent les variables pour augmenter le nombre de messages productibles.

Nous avons décrit deux façons de produire de la langue naturelle sans construction du message final. Les messages pré-enregistrés sont fixés par avance par un expert humain. Une telle approche ne peut prendre place que dans le cas d'un domaine limité. Il y a peu ou pas de dynamisme dans la sortie du système c'est-à-dire que le système réagira toujours de la même façon. De plus, la qualité dépend du travail de l'expert et de la finesse avec laquelle il aura prévu les différents messages. Les patrons de réponse offrent une plus grande souplesse mais restent très contraints. De plus, leur élaboration peut s'avérer fastidieuse (trouver le juste milieu entre la généricité d'un patron et le nombre de patron à créer). Cette approche implique elle aussi que l'expert détermine le *quoi dire* et le *comment dire*.

Certaines applications ne peuvent pas se limiter à l'utilisation de textes pré-enregistrés, voire même, *ont pour objet* de produire du texte en langue naturelle. Ce qui était décidé par un expert doit alors être calculé par le système. Ainsi ces systèmes sont architecturés autour de deux grandes tâches : déterminer le *quoi dire* et déterminer le *comment dire*. C'est ce travail automatique effectué par un système que nous appelons la *génération* en langue naturelle. Nous la distinguons de la production de langue naturelle et c'est à cette tâche que nous nous attachons dans la suite de ce document. Dans la section suivante, nous présentons quelques unes de ces applications afin de mettre en valeur les enjeux du travail de génération automatique de texte. Nous verrons notamment que dans le cas de certaines applications, seule l'une des deux tâches, *comment dire* et *quoi dire*, est effectuée par le système. Puis nous aborderons les différents traitements relatifs à une sortie en langue naturelle tant à l'écrit qu'à l'oral.

Applications à la génération de texte

Le choix du *Quoi dire* dépend de l'application et/ou du contexte d'interaction. Les exemples 1.8 à 1.11 sont des réponses à la question 1.7. On comprend bien qu'elles correspondent à des contextes bien différents (oral/écrit, information sur les transports en communs, "dans la rue", sur Internet...)

Exemple 1.7 *Où est le musée du Louvre ?*

Exemple 1.8 *Ben euh à Paris en France*

Exemple 1.9 *Station "Palais Royal musée du Louvre" sur la ligne 1.*

Exemple 1.10 *Continuez tout droit sur la "rue de Rivoli" et l'entrée sera à votre gauche.*

Exemple 1.11 *1 place André Malraux, 75001 Paris, France. Cliquez ici pour voir la carte.*

Au niveau des applications, on peut citer par exemple la *traduction automatique*. Cette tâche consiste à traduire du texte (une phrase, un paragraphe, un document) d'une langue source à une langue cible. On s'attend bien évidement à avoir,

en langue cible, un texte en langue naturelle. Dans ce cas, le *quoi dire* est contraint par le texte source et ce n'est pas au système de le déterminer. En revanche, le système doit gérer la sous-tâche du *comment dire*. Une bonne traduction n'est pas une traduction mot à mot, elle ne doit pas consister en un simple remplacement des mots du texte source par les mots correspondant dans la langue cible. Les formulations et "façons de dire" changent d'une langue à l'autre, les expressions sont différentes et certains termes n'ont pas d'équivalent et doivent donc être paraphrasés (voir l'exemple 1.12).

Exemple 1.12

Langue source	*anglais*
Texte source	*It's up to you !*
Langue cible	*français*
Texte cible	**C'est en haut jusqu'à toi !*
	C'est à toi de décider !

Les Systèmes de dialogue homme-machine (SDHM) quant à eux permettent d'établir une interaction de type dialogique entre un utilisateur et un système. On en trouve de nombreux tant grand public que dans le monde de la recherche dont par exemple le système académique HALPIN [Rouillard, 1998], les agents conversationnels de la société VirtuOz [VirtuOz, site web], le système de dialogue *oral* homme-machine ARISE [Lamel et al., 2000]. Notamment sont apparus depuis quelques années des agents virtuels intégrés à des sites Internet. Ces personnages sont bien souvent animés. Ils répondent aux questions que les utilisateurs saisissent dans une zone de texte prévue à cet effet. En général sur des sites commerciaux, ils permettent à l'utilisateur de trouver un produit ou encore d'obtenir une information [6]. Ils doivent répondre en fonction de ce que l'utilisateur a écrit mais aussi de ce qui a été dit précédemment au cours du dialogue (ou *historique du dialogue*). C'est à partir de ces deux éléments et des connaissances internes que possède le système qu'une réponse va être élaborée. Prenons en exemple un agent intégré à un site commercial de transport. Observons l'échange 1.13. "S" indique les interventions du système et "U" celles de l'utilisateur. La seconde réponse du système montre que le système prend en compte la toute première intervention de l'utilisateur : il a mémorisé le trajet auquel l'utilisateur s'intéresse. Elle montre aussi que le système répond en fonction de ses connaissances : il indique pour chaque train les informations dont il dispose. De plus, on note que le système a sélectionné "quoi dire" : lister tous les trains qui partent le jour même et annoncer cette liste par une phrase introductive. Il y a aussi une sélection du "comment dire" au niveau de la formulation de la phrase introductive et de la présentation d'une liste des trains sur une page (alors que le système aurait pu, par exemple, énoncer les trains un à un). Dans un tel exemple, le *quoi dire* est guidé par le contexte (historique du dialogue et connaissances du système) et la "compréhension" de l'intervention de l'utilisa-

6. On en trouve sur les sites commerciaux de voyage-sncf.com et IKEA par exemple.

teur. Le “comment dire” est quant à lui déterminé à l’avance lors de la conception du système.

Exemple 1.13 U_1 *Je cherche une information sur les trains Paris-Lyon.*
S_1 *Quelle information cherchez vous ?*
U_2 *Quand est le prochain train ?*
S_2 *Voici la liste des trains au départ de Paris et à destination de Lyon aujourd’hui.*
+ affichage d’une page listant les trains concernés : heure de départ, heure d’arrivée, type de train, prix

Un autre exemple d’application est celle du *résumé automatique de texte*. Cette tâche consiste d’une part à extraire les informations essentielles d’un texte (détermination du *quoi dire*), d’autre part à rédiger le résumé (détermination du *comment dire*) [Minel, 2002; Mani & Maybury, 1999].

D’autres applications peuvent être citées comme la production automatique de rapport sur les données numériques ou encore celle de manuels d’instructions, la description automatique d’itinéraires, ...

Traitements communs aux deux modalités en sortie

Comme nous venons de le voir, selon l’application visée, le *quoi dire* et le *comment dire* ne sont pas construits de la même façon. Quoi qu’il en soit, un système de génération en langue naturelle est en général organisé en deux grands modules qui réalisent les deux tâches principales que sont la détermination du *quoi dire* et du *comment dire* [Danlos, 1989b]. Un premier module est nommé composant stratégique, générateur profond ou encore conceptualisateur. Il s’occupe à la fois de sélectionner le contenu (profond) de ce qui va être généré et d’organiser la structure rhétorique c’est-à-dire l’ordre des phrases et des idées. On parle ainsi aussi de macroplanificateur ou encore de planificateur de texte. Le second module est nommé composant tactique, générateur de surface ou encore le formulateur. C’est un module linguistique qui s’occupe de la lexicalisation et du choix de la syntaxe à utiliser. On parle ainsi aussi de microplanificateur ou encore de planificateur linguistique.
D’autres approches découpent la tâche en trois parties qui sont (1) la détermination du contenu (relevant du *quoi dire*), (2) la plannification de phrase (relevant à la fois du *quoi dire* et du *comment dire*) et (3) la réalisation de surface (relevant du *comment dire*). De telles approches prennent en compte le fait qu’il y a un lien entre les concepts, les idées qui sont à exprimer et la façon dont celles-ci sont exprimées c’est-à-dire leur réalisation linguistique. La détermination du contenu consiste à construire une représentation conceptuelle de ce qui doit être dit étant donné le contexte et les connaissances du système. À partir de cette représentation, il faut déterminer la structure rhétorique du texte à générer. Cette étape permet d’obtenir

un plan de texte sur lequel le système se base pour déterminer la structure syntaxique des différentes phrases à générer. Cette structure peut, par exemple, prendre la forme des arbres syntaxiques que nous avons vus précédemment dans la section consacrée à l'entrée de langue naturelle (section 1.3.1, page 11). Pour pouvoir y insérer les mots (les feuilles de l'arbre syntaxique), il faut passer par une étape de lexicalisation c'est-à-dire de choix du lexique, des mots, qui vont être utilisés. On obtient alors une suite de mots *hors-contexte*. En effet, ces mots doivent être adaptés à leur contexte. Ainsi l'article (ou *déterminant*) "le" et l'adjectif "beau" précèdent le nom "ferme" qui est féminin, se réaliseront sous les formes respectives "la" et "belle". Il s'agit de choisir la bonne forme. On parle d' "ajustement morphologique". D'autres ajustements doivent être faits comme par exemple la gestion des phénomènes d'élision et de contraction. Par exemple, l'article "le" précédant un nom commençant par une voyelle comme "écran" doit être élidé et transformé en " l' ". De même, la suite "à les" doit être contractée en "aux". On obtient alors une suite de mots *en contexte* ou bien une phrase. S'il s'agit de générer un texte, alors on a une suite de phrases qui ont été ordonnées à l'étape de structuration rhétorique. La lexicalisation se fait en revanche phrase par phrase. On a donc des phrases hors-contexte. Si le même mot est utilisé dans toutes les phrases, il peut cependant être judicieux de ne pas le répéter et d'utiliser un pronom ou bien une ellipse. On parle de façon générale d'*expressions référentielles*. Dans la même idée, pour éviter certaines répétitions malvenues ou bien peu naturelles, on peut aussi opérer des regroupements (ou *agrégations*). Il peut s'agir par exemple de l'ajout de conjonctions de coordination.

Ces traitements ne suffisent pas pour produire une sortie satisfaisante. En effet, le texte donné en exemple ci-après n'est satisfaisant ni à l'écrit, ni à l'oral (il s'agit de texte et non pas d'audio). Nous abordons à présent les traits spécifiques à chacune des deux modalités orale et écrite.

Exemple 1.14 *les présentations auront lieu dans la salle de conférence*
elle est située au premier étage
boissons chaudes et encas seront disponibles à la pause dans la cafétéria

Spécificités liées à une sortie écrite

La question de savoir quels sont les traitements spécifiques à l'oral et à l'écrit n'est pas anodine. En effet, si les traitements peuvent être les mêmes, ils peuvent se réaliser de façons très différentes. Ce n'est pas parce que dans les deux cas, il y a une étape de lexicalisation que le résultat escompté est le même quel que soit la modalité. Ainsi par exemple, on dit "par contre" mais il est plus correct d'écrire "en revanche". Peu de travaux se sont intéressés à ce type de différences, il est donc difficile d'identifier les différences réelles.

On peut tout de même identifier quelques traitements spécifiques à une sortie écrite. En particulier, il faut ajouter la ponctuation et les majuscules adéquates.

De plus, l'encodage des caractères doit être contrôlé en fonction de l'interface au sein de laquelle le message sera présenté. Ceci afin d'éviter un affichage illisible comme "les notes des Ã©lÃ"ves sont Ã rendre au plus tÃ't" au lieu de "les notes élèves sont à rendre au plus tôt". Enfin la question de la mise en forme (couleur, police, mise en page) doit être réglée. Notamment, selon l'espace disponible pour afficher le texte (page Internet, écran de téléphone portable,...), la quantité de texte et sa mise en forme ne sera pas la même. D'autres spécificité à l'écrit peuvent être citées. Notamment la mise en valeur d'un texte, qui se fait par la mise en gras des caractères ou un changement de la taille de la police par exemple. D'autre part, la quantité d'information que l'on peut donner à voir est relativement importante. En effet, à l'écrit, on est pas limité par la mémoire de travail puisqu'il est toujours possible de revenir sur un message afficher, le lire et le relire. Ceci n'est pas le cas à l'oral.

Spécificités liées à une sortie orale

En ce qui concerne une sortie orale, l'objectif est à partir d'une suite de mots, d'obtenir un fichier audio. Cette étape est appelée la synthèse vocale (ou "text to speech", en anglais). Certains synthétiseurs proposent plusieurs voix (d'homme, de femme ou encore d'enfant). De plus le débit de parole et l'intonation peuvent être paramétrés. De plus, un synthétiseur est spécifique à la langue parlée. En effet, le texte "bravo" ne doit pas être synthétisé de la même façon pour l'italien ([bravo] [7]), l'anglais ([ˌbrɑ ːvəʊ]) ou encore le français ([bʁavo]). Le texte doit aussi être formaté de telle façon à ce que le synthétiseur sache qu'il faut épeler le mot dans le cas d'un acronyme comme "CNRS". Enfin des indications prosodiques doivent y être ajoutée. Notamment, une forme de ponctuation (pas nécessairement celle de la langue écrite) doit indiquer au synthétiseur les temps de pause mais aussi la prosodie à produire (par exemple, dans le cas d'une question, l'intonation doit monter sur la fin de l'énoncé).

1.3.3 En résumé

Nous avons vu que de nombreuses applications nécessitent de traiter de la langue naturelle que ce soit en entrée, en sortie ou bien les deux. La modalité orale/écrite d'interaction doit être prise en compte lors de la construction de ces systèmes. En effet, non seulement certains traitements sont réservés à chacune de ces modalités mais de plus même les traitements communs aux deux modalités doivent y être adaptés. En ce qui concerne la sortie en langue naturelle, nous avons distingué la production en langue naturelle de la génération automatique de texte. Cette dernière étant la seule qui consiste réellement en la construction d'une sortie en langue naturelle.

7. Il s'agit d'une notation phonétique utilisant l'alphabet phonétique international (International Phonetic Association, [Association, 1999]).

1.4 LA QUESTION-RÉPONSE

1.4.1 Les systèmes de réponse à une question précise

Les systèmes de réponse à une question précise ou systèmes de question-réponse gèrent des interactions dont le but est de répondre une information précise à une question en langue naturelle. Pour réaliser cette tâche, ces systèmes analysent la question et déterminent ce qu'il faut chercher comme information. Ensuite, à partir d'un ensemble (ou collection) de documents à leur disposition, ils cherchent la réponse à la question.

Le schéma 1.6 extrait de [Ligozat, 2006] présente l'architecture générale d'un système de question-réponse.

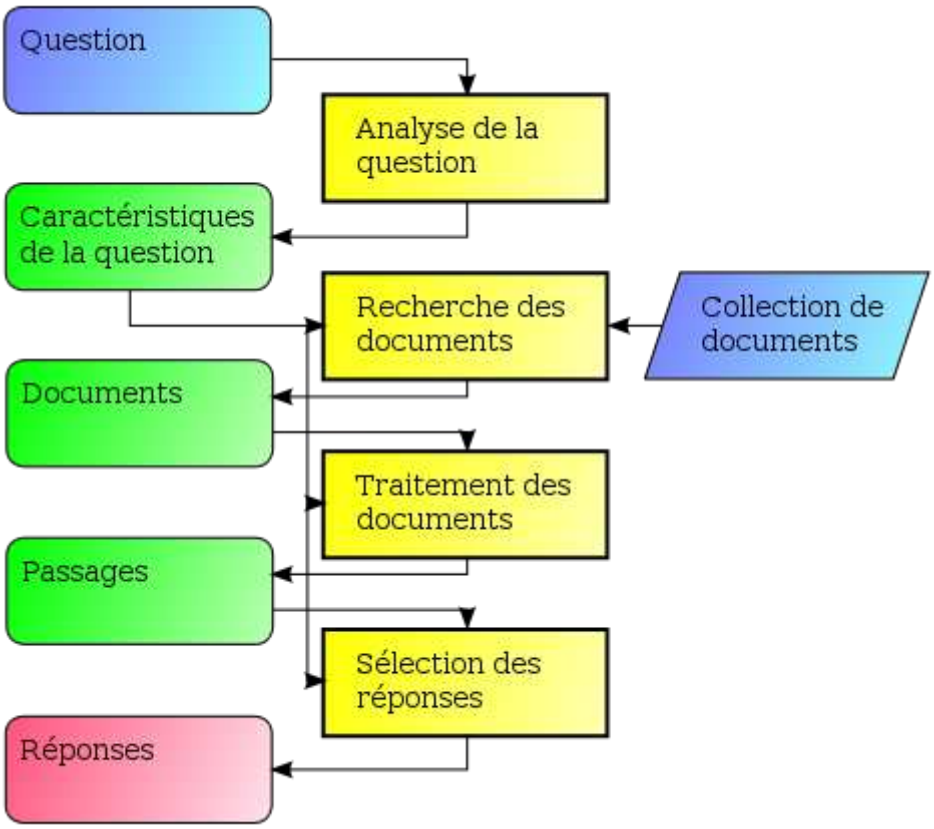

FIGURE 1.6 – *Représentation schématique d'un système de question-réponse, d'après [Ligozat, 2006]*

Ce que ces systèmes ne sont pas : des moteurs de recherche

Nous avons tous déjà utilisé un moteur de recherche sur la Toile, comme "Google" par exemple. Dans ce cas, pour chercher une information, nous donnons un ensemble de mots-clefs. La réponse du système est un ensemble de liens vers des pages Internet (bien souvent accompagnés d'un extrait du document correspondant).

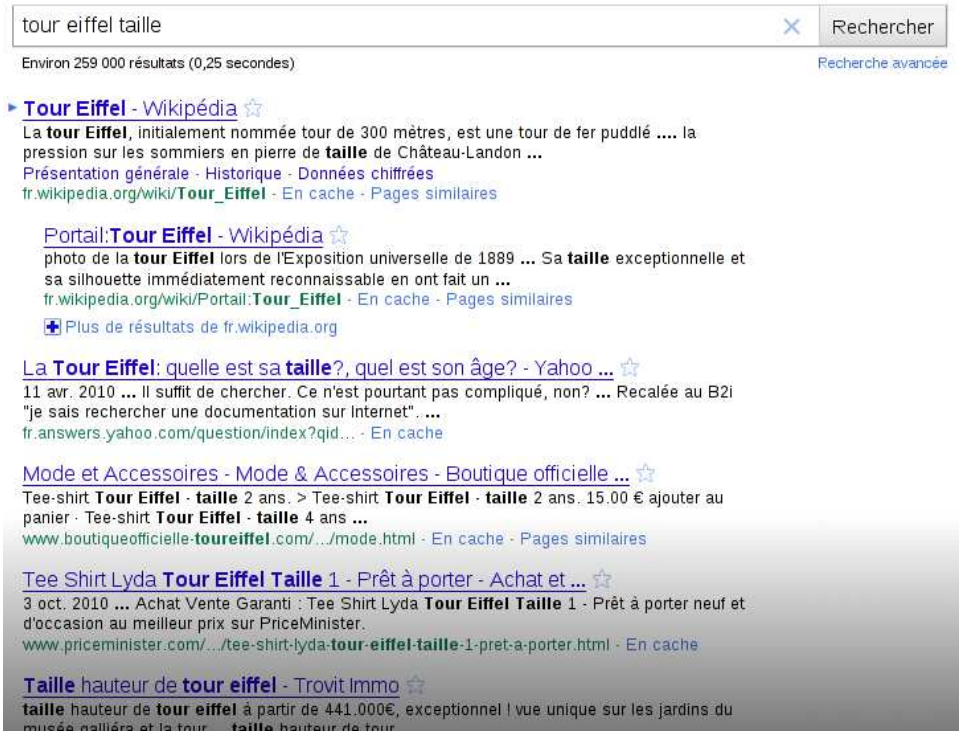

FIGURE 1.7 – *Page du moteur de recherche "Google" pour la requête "tour eiffel taille"*

La figure 1.7 montre, à travers l'exemple du moteur de recherche Google, deux zones principales : la requête de l'utilisateur et le retour du moteur de recherche. La requête est constituée de mots-clefs choisis par l'utilisateur. Ces mots sont les mots qu'il aura jugé pertinents pour obtenir l'information qu'il recherche (ou bien l'accès vers le site qu'il vise). L'utilisateur doit donc se faire une idée du document qu'il recherche et des mots qu'il contient. Il doit aussi penser à utiliser des mots suffisamment spécifiques pour que le moteur ne lui renvoie pas des documents non pertinents. Ceci implique qu'il ait une connaissance *a priori* du domaine, du thème, de l'objet de sa recherche. Ce qui n'est bien sûr pas toujours le cas. Le retour du moteur de recherche est en fait un ensemble de réponses constituées principalement de : (1) le titre du document concerné qui est un lien hypertexte (cliquable) vers la page Internet où il est hébergé (2) un extrait du document qui est en texte brut et contenant certains ou tous les mots-clefs de la requête. D'autres éléments peuvent accompagner une réponse. Selon le moteur de recherche et/ou selon les préférences de l'utilisateur, les réponses peuvent être accompagnées d'une prévisualisation de la page Internet, de liens supplémentaires vers d'autres pages du site Internet hébergeur, etc. Une fois ce retour donné par le moteur de recherche, l'utilisateur doit encore naviguer vers une ou plusieurs des pages proposées en réponse et fouiller dans ces documents pour obtenir l'information qu'il recherche, en supposant que le document soit pertinent par rapport à la requête de l'utilisateur. Ceci est un avantage car si l'utilisateur ne cherche pas une information précise, il accède à un document entier traitant *a priori* du thème qui l'intéresse. Parce que le

moteur propose plus d'une réponse, l'utilisateur peut même accéder à beaucoup d'informations différentes autour du thème défini par ses mots-clefs. Cependant, si les documents sont peu ou pas pertinents, par exemple parce que certains mots de la recherche sont polysémiques, l'utilisateur doit soit fouiller dans de nombreux documents pour obtenir ce qu'il cherche soit reformuler sa requête. Ceci est coûteux en temps et bien souvent fastidieux. Dans le cas où l'utilisateur cherche une information précise, celle-ci peut se trouver dans l'extrait du document fourni dans la réponse. Dans le cas contraire, l'utilisateur doit fouiller par lui-même dans le document afin de trouver l'information qu'il recherche.

Les moteurs de recherche sont des outils extrêmement puissants. Parce qu'ils sont très utilisés, leurs développeurs disposent de quantités très importantes de données afin de les rendre encore plus performants et dans bien des cas, ils sont des outils très efficaces et satisfaisants pour les internautes. Cependant, dans certains cas, il peut être très coûteux de les utiliser. Les cas principaux sont : (1) lorsque l'utilisateur n'a que peu de connaissances du domaine dans lequel il fait sa recherche ; alors il ne sait pas choisir les mots-clefs pertinents pour formuler sa requête, (2) lorsque le moteur de recherche ne renvoie pas de document pertinent ; alors l'utilisateur doit reformuler sa requête, (3) lorsque l'utilisateur cherche une information précise ; alors il doit fouiller par lui-même dans un ou plusieurs documents pour découvrir sa réponse.

Si les systèmes de question-réponse ne résolvent pas tous ces problèmes, il peuvent cependant constituer une alternative intéressante dans la plupart des cas.

Ce que ces systèmes sont

Un système de question-réponse permet à l'utilisateur de poser une question précise. En effet, il ne s'agit plus d'une requête, mais d'une phrase : une question plus précisément. On parle de question en langue naturelle. Elle est formulée dans la langue de l'utilisateur sans qu'il ait besoin de la transformer pour la rendre compréhensible par le système. Elle peut être émise à l'écrit (saisie au clavier par exemple) ou bien à l'oral (dans le cas d'un système par téléphone par exemple).

Définition 1.2 ***Système de question-réponse (ou SQR)*** *: il gère la tâche de répondre à une question précise exprimée en langue naturelle. Ces systèmes peuvent être en domaine fermé (les questions portent sur un domaine précis comme la médecine, la politique,...) et prendre appui sur des connaissances particulières à ce domaine ou en domaine ouvert (les questions peuvent concerner n'importe quel domaine et disposent de sources d'information bien plus larges qu'en domaine fermé).*

Dans les grandes lignes, un SQR fonctionne en enchaînant différentes étapes de traitements. D'abord la question est analysée afin d'en extraire les éléments pertinents pour la suite des traitements. Les documents à disposition du système sont traités (en général au préalable à toute question). Une sélection des documents

pertinents par rapport à la question est faite. Ensuite, la réponse est localisée puis extraite des documents sélectionnés.

Ces systèmes sont évalués principalement à l'occasion de campagnes nationales et internationales. Ainsi, un système peut se confronter à d'autres, ses performances peuvent être testées et il peut ainsi être amélioré.

Ces systèmes ont été développés tant pour des interactions écrites qu'orales.

Architecture générale

Les systèmes de question-réponse fonctionnent en deux phases. La première phase est si possible effectuée à l'avance, alors même qu'aucune question n'a été posée au système. Il s'agit d'un pré-traitement de la collection de documents dont il dispose. La seconde phase est effectuée lorsque l'on soumet une question au système.

Les différentes étapes de traitement sont les suivantes :

Pré-traitement des documents On distingue deux types de pré-traitements sur les documents. Les premiers relèvent de la normalisation de ces documents, les seconds s'occupent d'indexer les documents afin de rendre leur sélection plus facile et rapide. La normalisation des documents consiste par exemple en une transformation de page Internet, au format html, en document textuel. L'indexation des documents consiste à repérer les mots particulièrement importants (ou significatifs) pour créer un index des documents. L'indexation peut aussi se faire sur la base d'une analyse des documents comme c'est le cas dans les systèmes Qristal [Laurent & Séguéla, 2005] et RITEL [Rosset et al., 2006]. On pourra se référer à [Manning et al., 2008] pour un état de l'art sur ce thème.

Traitement de la question Cette étape consiste en l'extraction des mots importants de la question dont notamment l'objet de la question c'est-à-dire ce sur quoi elle porte, nommé aussi *focus* [El Ayari, 2007] et les entités nommées (qui seront définies dans la section 3.2.1). Elle consiste aussi en la détermination du type de la question et du type de réponse attendu à partir de la forme interrogative.

Sélection des documents Il s'agit d'une première sélection qui permet de ne garder en lice que les documents les plus susceptibles de contenir la réponse.

Sélection des passages ou *snippet* Seconde sélection : il s'agit de fouiller dans les documents et de sélectionner des passages susceptibles de contenir la réponse. La taille de ces passages varie d'un système à l'autre.

Extraction des réponses candidates Détection fine au cours de laquelle est extraite la réponse elle-même. Il peut y avoir plusieurs réponses trouvées dans différents passages.

Sélection des *n* meilleures réponses En fonction de différents indices élaborés au cours du travail de sélection des réponses, les réponses sont ordonnées de la plus probable à la moins probable. Ensuite, en fonction du contexte et du système, seule la ou les premières réponses sont conservées. Certains systèmes sont capables de fournir plus d'une réponse à la fois par exemple en renvoyant une phrase qui indique le nombre de réponses et les citant. Nous verrons plus en détail le retour des systèmes dans la section 1.4.3.

Les campagnes d'évaluation

Les SQR sont évalués au cours de campagnes auxquelles participent bon nombre de systèmes. Ces campagnes sont nationales (EQueR, QUAERO,...) ou bien internationales (TREC, NTCIR, CLEF,...). Lors de ces campagnes, différentes tâches (ou "tracks" en anglais) sont proposées aux systèmes participants. Par exemple, il existe des tâches inter-langues (ou "Cross-Language" en anglais) ou encore des tâches en domaine spécialisé. Chaque tâche propose une série de questions (quelques centaines en général) et un ensemble de documents parmi lesquel les réponses doivent être recherchées. Les questions peuvent être de différents types. Par exemple il peut s'agir de questions factuelles (sur un fait précis), de définitions ou encore de questions "liste" qui demandent une liste d'information. Certaines tâches proposent des questions enchaînées. Dans ce cas, les questions sont regroupées par thème et ne sont pas indépendantes les unes des autres. Une question pouvant par exemple faire référence à une question antérieure du même thème [Séjourné, 2009]. [Galibert, 2009] propose un tableau détaillé très précis résumant les principales caractéristiques des différentes campagnes d'évaluation). Les systèmes sont évalués sur les résultats qu'ils fournissent et non pas de façon interne. Le résultat attendu est une "réponse" précise à la question, le document au sein duquel la réponse a été trouvée, éventuellement le passage au sein duquel elle a été trouvée et éventuellement un score de confiance que le système accorde à sa propre réponse. L'évaluation porte sur ces différents éléments. En général, les concurrents sont invités à proposer plus d'une réponse à chaque question. Ces réponses sont ordonnées. Un système est meilleur que les autres s'il a plus de bonnes réponses positionnées en première position et si les passages associés justifient bien cette réponse. Un passage qui contient bien la réponse proposée par le système mais qui n'a aucun rapport avec la question ne sera pas considéré comme justifiant la réponse.

1.4.2 La question-réponse, une interaction ?

La tâche question-réponse n'est en fait pas à proprement parler une interaction puisqu'il n'y a pas d'échange réel. Actuellement, les systèmes de question-réponse se donnent pour but de donner une réponse précise à une question. Cette dernière

est supposée être posée par un utilisateur à qui on n'offre pas la possibilité de reformuler sa question, ou encore d'en poser une autre sur le même thème. La réponse se présente sous la forme d'un mot ou d'un groupe de mots. Pourquoi une telle limitation ?

Pensons un instant la question-réponse comme une interaction. Alors on peut la voir comme un sous-type de dialogue dans lequel les interventions de l'utilisateur seraient des questions et auxquelles le système devrait répondre. Le but du dialogue est donc de donner l'information voulue à l'utilisateur. La particularité des systèmes de question-réponse en domaine ouvert est qu'ils sont indépendants du domaine d'application : on peut poser des questions sur *tout et rien*. En revanche, le type de questions posées est bien souvent limité. Si les questions factuelles (sur des faits comme par exemple le nom d'un lieu ou d'une personne) ont fait l'objet de nombreuses études, d'autres types de questions ont été moins étudiés, comme par exemple les questions dites procédurales (question "comment" qui interrogent sur la façon dont faire quelque chose par exemple).

1.4.3 La réponse des systèmes

Ce sont les campagnes d'évaluation des SQR qui définissent la forme que doit prendre la réponse des SQR. La contrainte majeure étant le besoin d'évaluer de façon automatique la qualité d'une réponse. Il s'agit ainsi d'une réponse courte, limitée, précise. Elle est accompagnée du document et du passage dont elle a été extraite. Dans le cadre des campagnes QAST, une réponse correcte est définie comme "*une suite de caractères qui consiste en une information pertinente (la réponse exacte) et la réponse est confirmée par le document retourné*" (traduit depuis [Turmo et al., 2007]). Les guides de travail des campagnes TREC précisent qu'une réponse sera considérée comme correcte si la suite de caractères répondue est exactement la bonne réponse et est confirmée par le document retourné. Une réponse non-exacte quant à elle contient la bonne réponse mais la suite de caractères répondue contient soit d'avantage que simplement la réponse, soit ne contient pas l'intégralité de cette réponse[QA Track, site web].

Ces définitions sont assez vagues (le terme "réponse" étant utilisé pour référer à différents éléments) et peuvent sembler surprenantes voire complexes à première vue. Cependant, définir ce qu'est une bonne réponse précise n'est pas simple. [Voorhees, 1999] remarque en particulier que le jugement d'une réponse comme bonne ou mauvaise est subjectif. Quoi qu'il en soit, la réponse est définie comme un élément court, restreint. Elle doit être pertinente par rapport à la question c'est-à-dire fournir une information cohérente par rapport à l'information recherchée à travers la question.

1.4.4 Limites de ces réponses

Les réponses fournies par les systèmes sont les réponses attendues dans le cadre des campagnes d'évaluation. En ce sens, elles ont l'avantage de rendre possible une comparaison des performances de différents systèmes. Et c'est précisément le rôle que se donnent les campagnes d'évaluation décrites précédemment. Cependant, ces réponses ne prennent pas en compte l'hypothétique utilisateur qui pose la question et qui est à la recherche d'une information. Ces réponses ne sont pas élaborée pour être intégrées au sein d'une interaction mais pour correspondre aux exigeances des campanges d'évaluation et plus précisement d'une évaluation qui doit se faire de façon automatique. L'évaluation étant un travail indispensable pour permettre des avancées dans ce domaine de recherche expérimental, ce choix se justifie largement. Le résultat cependant est que la réponse doit se réduire à une information accompagnée d'un document qui justifie la véracité de l'information.

CONCLUSION DU CHAPITRE

Nous avons présenté dans ce chapitre une vision très générale des enjeux de l'interaction en langue naturelle. Bien qu'il ne s'agisse que d'un survol, il permet d'une part d'introduire les problématiques du domaine et d'autre part de contextualiser le travail de thèse exposé dans le présent document.

Plus précisément, nous nous sommes centrée sur l'interaction du type "question-réponse" et sur les deux modalités d'interaction que sont l'écrit et l'oral. Nous avons montré qu'actuellement les systèmes de question-réponse fournissent des réponses très succinctes qui sont adaptées aux exigences des campagnes d'évaluation. Ces campagnes permettent une comparaison des résultats obtenus par les systèmes de question-réponses ce qui est primordial pour les évaluer et déterminer les techniques les plus efficaces pour trouver la réponse à une question précise. Cependant, parce que nous faisons le choix de voir l'échange question-réponse comme une interaction, il nous semble important que ces systèmes soient aussi en mesure de produire des réponses en interaction. Notre point de vue est ainsi de considérer la question et la réponse en lien l'une avec l'autre existant dans l'interaction. Le chapitre suivant traite plus précisemment du cas de la réponse vue comme en interaction avec la question.

RÉPONSE EN INTERACTION 2

SOMMAIRE

Ce chapitre vise à montrer que répondre n'est pas uniquement présenter une information mais fait partie d'une interaction entre deux locuteurs (même dans le cas d'une interaction homme-machine). Nous proposons une comparaison de réponses d'humains et de systèmes puis exposons des travaux montrant les caractéristiques d'une réponse prise comme appartenant à une interaction. Nous en déduisons les définitions d'*énoncé-réponse*, d'*information-réponse*. Puis nous cherchons à définir ce que pourrait être une *réponse en interaction* pour les systèmes de question-réponse et constatons l'absence de corpus constitué de telles

réponses. Le chapitre conclut sur la nécessité de constituer notre propre corpus.
Ce chapitre correspond à un travail d'état de l'art, de caractérisation et de définition de notre objet d'étude "la réponse" et à la justification de notre choix qui est de créer notre propre corpus de réponse à une question.

This chapter aims to show that answering is more than presenting an information but is part of an interaction. We propose a comparison between human and system answers and present works on the caracteristic of an answer in interaction with its question. We deduce the definitions of *answersing uterance* and *answering information*. Then we address the case of defining what could be an answer *in interaction* for question-answering systems and note the lake of such answer corpus.

2.1 INTRODUCTION

Répondre à une question est souvent considéré d'un point de vue "informatif" dans la communauté des systèmes de question-réponse. Les problèmes principaux que l'on cherche à résoudre sont la détermination du type d'information à donner en réponse et la recherche de cette information. Cependant, répondre à une question ne consiste pas seulement à donner une information. Il s'agit aussi de produire un acte de langage qui correspond à ce que l'interlocuteur a exprimé vouloir dans sa question et de répondre en fonction du contexte et de l'interaction. Les questions à se poser sont alors par exemple : faut-il réutiliser les mots de la question ou bien y faire référence par un pronom ou une anaphore ? Faut-il prendre en compte ce qu'il s'est passé précédemment au cours de l'interaction ? Faut-il préparer l'interlocuteur à sa réponse en l'introduisant, par exemple, par une explication ? Est-il important de se faire une idée des connaissances de son interlocuteur pour formuler une réponse qui lui soit compréhensible ?... Actuellement, les systèmes de question-réponse ne prennent pas en compte toutes ces questions : ils ne fournissent pas une réponse en interaction avec la question. L'un de nos buts, au travers de ce travail, est de montrer les avantages et l'intérêt qu'ont les systèmes de question-réponse (SQR) à produire des réponses en interaction plutôt que strictement des réponses informationnelles que nous appelons d'*information-réponse*.

2.2 EXEMPLES DE RÉPONSES

Qu'est-ce que répondre ? Comme nous l'avons vu, cela dépend du type d'interaction. Dans ce travail, nous nous attachons uniquement à la réponse dans le cadre d'une interaction homme-machine de type "réponse à une question précise" : un interlocuteur humain pose une question, un système répond à cette question. Notre but est que l'interlocuteur questionnant puisse être un humain et que ce soit un système automatique qui lui réponde. Nous voulons proposer une réponse qui soit plus naturelle que celles proposées actuellement par les SQR et, surtout, qui satisfasse l'utilisateur. Déterminer ce que peut être une telle réponse fait partie des problèmes que nous nous posons. Nous supposons qu'une telle réponse ne se limite pas à *donner une information*. Notre intuition est qu'elle doit au moins montrer au questionneur que sa requête a été bien comprise en plus de donner une information qui répond à la question.

Nous proposons ici une observation de différentes réponses et plus précisément de différents corpus de réponses [1] que nous pourrions utiliser pour observer et déterminer les caractéristiques d'une réponse donnée dans une interaction de type réponse à une question. Nous avons observé à la fois des réponses produites par des humains, et des réponses fournies par des systèmes. Les réponses d'humains peuvent être considérées en quelque sorte comme des références tandis que

1. Par "corpus de réponses", nous entendons un ensemble, une collection de réponses.

les réponses des systèmes nous indiquent ce que, actuellement, les systèmes fournissent.

2.2.1 Réponses d'humains

En réalité, nous, humains, passons notre temps à répondre à des questions. Nous y répondons dans la rue, au travail, à la maison,... Nous répondons à l'oral, à l'écrit, en face-à-face, sur un répondeur téléphonique, par mail, sur un bout de papier,... Dans ce travail, nous nous intéressons aux réponses d'humains pour pouvoir les étudier. De plus, nous travaillons dans le cadre des systèmes de réponse à une question. Nous nous limitons ainsi ici à donner des exemples de réponses humaines dans des cas d'interactions similaires.

Nous avons observé les réponses données sur des sites de question-réponse en domaine ouvert (c'est-à-dire quel que soit le thème abordé par les questions). Ces sites permettent aux utilisateurs de poser une question, de répondre à une question, d'observer les questions et les réponses existantes et éventuellement de désigner deux questions comme équivalentes.

Pour la même question "Combien mesure la Tour Eiffel ?", voici les résultats proposés sur deux sites de question-réponse francophones [2].

FIGURE 2.1 – *Réponse donnée sur le site "Answers.com Français" à la question "Combien mesure la Tour Eiffel ?"*

Sur le site "Answers.com Français" [Answers.com , fr], la question "Combien mesure la Tour Eiffel ?" a été considérée comme équivalente à la question "Quelle est la hauteur de la Tour Eiffel ?" par des humains (des contributeurs du site). Sur ce site, on a directement une réponse qui s'affiche. Il s'agit d'un ensemble de deux paragraphes au cours desquels sont donnés des informations autres que celle visée

2. Réponses valables le 06/07/2010.

par la question ainsi que l'information qui répond à la question. Cette information est donnée sous deux formes différentes (deux unités de mesure sont utilisées).

FIGURE 2.2 – *Résultat affiché sur le site "Yahoo ! France Question réponses" pour la question "Combien mesure la Tour Eiffel ?"*

Le second résultat (figure 2.2) est celui du site "Yahoo ! France Question réponses" [Yahoo QR, site web]. On observe que l'on n'accède pas directement à une réponse à la question bien que cette même question ait déjà été posée (à quelques différences de casse et espace près). En cliquant sur cette première question (Q1 figure 2.3), on accède à ses réponses. Pour une question donnée, il peut ne pas y avoir de réponse, en avoir une ou encore plusieurs. La personne qui a posé la question, mais aussi les internautes qui accèdent au site, peuvent aussi choisir la réponse qu'ils estiment être la meilleure réponse. Dans le cas de la question Q1, il n'y a qu'une seule réponse. Cette réponse ne donne pas directement la taille de la Tour Eiffel. Elle liste d'abord les hauteurs des différents étages puis donne la taille totale. Ces différentes informations sont introduites par un titre. Ensuite est indiquée une question en relation avec Q1 et à laquelle est jointe une réponse. Enfin, la source est indiquée par un lien hypertexte. De plus une indication sur la source de la réponse est fournie. Il s'agit ici d'une part de justifier la réponse, de donner une indication sur sa validité et d'autre part de suggérer au questionnant un moyen d'obtenir d'autres informations sur le même thème.

D'autres sites similaires existent, qu'ils soient francophones (par exemple [QuébecTop, site web]) ou non (par exemple [Answers.com, site web]). Ils

FIGURE 2.3 – *Unique réponse et "Meilleure réponse" présentée sur le site "Yahoo ! France Question réponses" pour la question "combien mesure la tour eiffel ?" (Q1 sur la figure 2.2)*

peuvent être aussi spécialisés dans un domaine donné. Par exemple, [The Correct, site web] ne propose des questions et des réponses que dans le domaine médical.

Les FAQ (Foires Aux Questions ou Frequently Asked Question, en anglais) fourmillent aussi de réponses d'humains. Ces regroupements de questions et de réponses traitent d'un thème particulier. Il s'agit de domaine limité.

2.2.2 Réponses de systèmes

Les systèmes répondent en fonction de leurs domaines d'application. La plupart des systèmes existants ne sont pas commercialisés mais sont développés au sein d'équipes de recherche et participent à des campagnes d'évaluation. Une liste de ces systèmes est difficilement exhaustive étant donné leur nombre cependant on peut se référer par exemple à [TREC, site web; NTCIR, site web; QA@CLEF, site web] pour la liste des systèmes participant aux campagnes d'évaluation et à [Galibert, 2009] (tableau 10.1, page 144) pour une synthèse des campagnes d'évaluation et de leur caractéristiques. Il faut noter que les réponses que demandent ces campagnes sont l'information-réponse éventuellement accompagnée d'autres éléments tels que le document dans lequel la réponse a été trouvée.

Nous prenons ici l'exemple de deux systèmes disponibles au sein de notre laboratoire : les systèmes RITEL et FIDJI.

Le système FIDJI (Finding In Documents Justifications and Inferences) est un système de question-réponse en domaine ouvert. Il se donne pour but de produire

des réponses qui sont entièrement validées par des extraits de textes [3] et ce en utilisant des mécanismes de compréhension reposant sur des inférences. Ce système est décrit plus précisément dans [Moriceau et al., 2009]. Si nous lui soumettons une question telle que "Où est la Joconde", le retour du système (voir la figure 2.4) consiste en un ensemble de trois réponses ordonnées et accompagnées chacune par un extrait de texte (une cinquantaine de mots). Cet extrait est un justificatif de la réponse : il s'agit d'un passage du document au sein duquel figure la réponse et qui, en principe, soutient la réponse en montrant qu'elle correspond bien à la question posée : le passage permet d'inférer la réponse.

<**question**>where is the Mona Lisa ?</**question**>
<**answer_1**>Paris</**answer**>
<**justification**>Even more than in Paris or Vienna or Rome , it can be hard to see anything worthwhile because of the crowds who flock to see it . Much of the National Gallery is afflicted by the kind of congestion which in the vaster Louvre concentrates on a few five - star items like the Mona Lisa .</**justification**>
<**answer_2**>America</**answer**>
<**justification**>For this , he assured John F Kennedy and his wife Jackie , " may both of you be praised , in the name of all the unnamed artists who thank you , perhaps from the depths of funereal night ' ' . By this , I suppose that the artists he had in mind were dead , though why dead anonymous artists should thank the Kennedys for bringing the Mona Lisa to America is not , to this reviewer at least , altogether clear .</**justification**>
<**answer_3**>Chinese Theatre</**answer**>
<**justification**>The game takes place in the year 2011 at a high - stakes auction of some of the world 's most treasured artworks and curios , including the Mona Lisa , the Pieta , the missing 18 minutes of Richard Nixon 's White House audiotape , The Scream , Mann 's Chinese Theatre (the last movie theater in the country) and a robotic sex partner .</**justification**>

FIGURE 2.4 – *Retour du système FIDJI à la question "where is the Mona Lisa ?" (en français, "Où est la Joconde ?").*

On note que ce système ne propose pas de réponse en langue naturelle : il se limite à un terme. La justification ne constitue pas non plus une réponse "convenable" qui puisse être présentée telle quelle à un utilisateur.

Le système RITEL quant à lui est destiné à la fois à une interaction orale et à une interaction écrite. Il existe deux versions du système : avec ou sans l'utilisation du module qui se charge de formuler une réponse. Ce module dit de génération en langue naturelle est nommé *RIGENTEL*. Une première version ne propose que

3. Ceci correspond aux conditions imposées par les campagnes d'évaluation en question-réponse.

la réponse extraite au cours de la phase de recherche d'information et de fouille dans les documents. La figure 2.5 montre la réponse du système à la question "*Combien mesure la Tour Eiffel ?*" sans le module RIGENTEL. Comme dans le cas du système FIDJI, il s'agit d'une réponse courte, qui consiste uniquement à fournir l'information qui répond à la question. La réponse illustrée par la figure 2.6 est élaborée par un module de génération en langue naturelle. Ce module décrit dans [Rosset et al., 2006] a été élaboré dans l'optique de donner un retour à l'utilisateur sur la compréhension du système. D'autre part, parce que l'utilisateur a tendance lui-même à réutiliser les formes utilisées par le système, les énoncés construits sont des énoncés faciles à comprendre par le module de reconnaissance vocale. Guidant l'utilisateur dans ses choix lexicaux, on minimise les erreurs de reconnaissance et la qualité de l'échange s'en voit améliorée. La réponse donnée par le système n'est donc pas prévue pour être plus agréable ou naturelle à l'utilisateur. Elle ne propose pas non plus d'information complémentaire à la réponse en elle-même.

id : 1
question : combien mesure la tour Eiffel
type : factuelle
class : nombre
answers :
324 mètres
mille pieds
(...)
answers2 :
324 mètres
<_det> la </_det> <K p="2 :1"> <_subs> tour </_subs> </K> <K p="1 :1"> <_org> <_org_com> < Eiffel> Eiffel </ Eiffel> (...)
mille pieds
<_pval> <_val> <4> 4 </4> </_val> </_pval> <_punct> , </_punct> <_np> La </_np> <_subs> tour </_subs> <K p="1 :1"> <_org> <_org_com> <Ẽiffel> Eiffel </Ẽiffel> </_org_com> (...)
(...)

FIGURE 2.5 – *Retour du système RITEL sans le module RIGENTEL à la question "combien mesure la tour eiffel ?" (version simplifiée et ne présentant que les deux premières réponses).*

ritel : Bonjour, c'est Ritel. Quelle question souhaitez-vous me poser ?
user : combien mesure la tour Eiffel ?
ritel : Je cherche ... tour Eiffel et taille ... C'est 324 mètres. J'écoute votre question.

FIGURE 2.6 – *Échange entre un utilisateur et le système RITEL utilisant le module RIGENTEL (version corrigée car le système fait une erreur d'analyse de la question).*

On observe que les réponses des systèmes sont bien plus limitées que celles des humains. Ce n'est pas une surprise : les humains ont une maîtrise de la langue

plus accrue (utilisation de phrases complexes, de liste,...), une connaissance du monde plus large (un édifice est construit par quelqu'un, à un moment du temps, éventuellement pour une occasion particulière,...) et sont capables d'interprétation (par exemple, dire que deux questions sont équivalentes). Ainsi dans les deux exemples de réponses d'humains, nous voyons qu'il y a une élaboration du message, qui est organisé et qui ne contient pas uniquement l'information visée par la question. Les réponses de systèmes que nous avons prises en exemple sont bien différentes. Elles sont plus limitées et même quand un système produit un énoncé réponse, ne se limitant pas à un mot-réponse, celui-ci ne contient aucune information complémentaire. La réponse reste centrée sur l'information qui correspond à une réponse à la question.

2.2.3 Réponses de systèmes plus élaborées

Certains travaux ont porté et portent encore sur l'élaboration de réponses plus élaborées qu'une réponse qui ne donnerait que l'information réponse dont notamment des travaux visant à proposer une réponse coopérative ([Cuppens & Demolombe, 1989; Gaasterland et al., 1992; Chu et al., 1996; Benamara, 2004; Moriceau, 2007]) ou encore des travaux à visée pédagogique, dans le cadre du e-learning principalement ([Gurevych et al., 2009] par exemple).

Nous ne citerons pas tous les travaux existants mais présentons les travaux de Véronique Moriceau à titre d'exemple pour montrer au lecteur le positionnement de notre travail par rapport à de tels travaux.

Les travaux de [Moriceau, 2007] portent sur les réponses coopératives. Il s'agit d'intégrer au sein d'une même réponse plusieurs informations-réponse potentielles. Ces travaux traitent notamment de la présentation de la réponse. La forme des réponses générées est `sujet verbe (objet) réponse` et est choisie de façon "intuitive".

Quoi qu'il en soit, les travaux ne parlent pas du choix de la forme linguistique de surface. Ce *point de détail* ne semble pas attirer l'intérêt des chercheurs et n'est que rarement mentionné au lecteur dans littérature.

À travers l'observation de réponses produites par des humains, de réponses produites par des systèmes et de réponses coopératives, nous avons présenté ce que peut être une réponse à une question actuellement. Nous avons pris un angle de vue applicatif, nous intéressant principalement à la réalisation concrète d'une tâche : répondre à une question. Nous nous intéressons maintenant à détailler quelques travaux plus théoriques visant à modéliser ce qu'est une interaction puis ce qu'est une réponse en interaction.

2.3 THÉORIES DE L'INTERACTION

Interagir à l'écrit ou à l'oral, c'est prendre en compte les différents acteurs de l'interaction dans un contexte particulier. On communique à quelqu'un, pour dire quelque chose et pour une raison donnée. Le linguiste Jakobson définit le schéma de la communication verbale comme composé d'un *message* qu'un *destinateur* envoie au *destinataire* dans un *contexte* particulier [Jakobson, 1963]. Le message est exprimé par un *code* (dans notre cas, la langue) et requiert qu'un *contact* existe entre destinateur et destinataire (c'est le canal de communication, la parole, par exemple).

Á partir de ces travaux, d'autres travaux ont considéré la communication comme un acte en interaction. Notamment, [Grice, 1975] énonce le principe de coopération suivant :

> ❝ Que votre contribution à la conversation soit, au moment oú elle intervient, telle que le requiert l'objectif ou la direction acceptée de l'échange verbal dans lequel vous êtes engagé. ❞

On observe que par définition, Grice considère l'acte d'énonciation comme appartenant à l'échange avec autrui. Il décrit un ensemble de principes décrivant le comportement coopératif des humains aux cours d'un dialogue [Grice, 1975]. On parle des maximes (conversationnelles) de Grice :

> ❝ **Maximes de quantité**
> – Que votre contribution contienne autant d'informations qu'il est requis.
> – Que votre contribution ne contienne pas plus d'informations qu'il n'est requis.
>
> **Maximes de qualité (ou de véridicité)**
> – Que votre contribution soit véridique :
> – N'affirmez pas ce que vous croyez être faux.
> – N'affirmez pas ce pour quoi vous manquez de preuves.
>
> **Maxime de relation (de pertinence)**
> – Parlez à propos (soyez pertinent)
>
> **Maximes de manière**
> – Soyez clair :
> – Évitez de vous exprimer avec obscurité.
> – Évitez d'être ambigu.
> – Soyez bref (évitez toute prolixité inutile)
> – Soyez ordonné.❞

Dans ces maximes, Grice exprime l'importance du choix de l'information à transmettre (maximes de quantité, de qualité et de relation) mais aussi l'importance de la façon dont cette information est transmise (maxime de manière).

D'autres travaux, notamment traitant du dialogue (homme-homme ou homme-machine), utilisent la notion de plan partagé pour décrire le fait qu'à l'occasion

d'un dialogue, les interlocuteurs construisent et exécutent des plans pour atteindre des buts [Lehuen, 1997; Grosz & L., 1990].

La plupart de ces travaux se rapportent en fait aux notions de cohérence et de pertinence car quelle que soit l'interaction, il est primordial de produire des énoncés qui correspondent à la fois au contexte d'énonciation et à la tâche visée.

Si ces théories ne s'appliquent pas directement à l'interaction de type question-réponse, elles mettent toutes en valeur que l'information fournie en réponse à l'utilisateur n'est pas l'unique élément à avoir son importance au cours d'une interaction.

2.4 THÉORIE DE LA RÉPONSE EN INTERACTION

2.4.1 Une grammaire dite interactive

Des travaux se sont intéressés plus spécifiquement au cas de la réponse en interaction et plus précisément de la réponse dans son interaction avec la question. Ces travaux partent du principe qu'un locuteur posant une question, non seulement cherche une information, mais a aussi une intention bien particulière qu'il exprime à travers la formulation de sa question. En choisissant une formulation plutôt qu'une autre, le questionneur se positionne de façon précise quant à l'information qu'il recherche et quant aux connaissances qu'il présuppose de son interlocuteur. En d'autres termes, la forme de la question traduit la réponse à laquelle s'attend un locuteur lorsqu'il pose sa question.

[Luzzati, 2001] définit une grammaire dite Grammaire Interactive, ou GRINT, qui constitue non seulement un éventail de formulation pour les questions en français mais aussi indique pour chaque formulation de question *l'intention* du questionneur et ce qu'il *met en cause* dans sa question. Enfin, pour chaque question, deux exemples types de réponses attendues sont proposées : le premier dans le cas où l'interlocuteur répond en effet à la question, le second illustre les cas où il ne répond pas (parce qu'il ne sait pas ou ne veut pas répondre ou bien parce qu'il demande de plus amples informations,...). On parle de la valence d'une réponse qui peut être positive (il y a en effet une réponse à la question) ou bien négative (l'énoncé répondu ne contient pas de réponse à la question).

2.4.2 Trois dimensions de variation des questions

La figure 2.1 montre le modèle de la GRINT comme un modèle de variation de la question selon trois dimensions : la dimension paradigmatique, syntagmatique et ontologique.

Si l'on se positionne dans la case "initiale", en se déplaçant dans la profondeur du cube, on fait varier la question selon son type ontologique : question portant sur une quantité, un lieu, un temps, une définition, l'explication d'une procédure

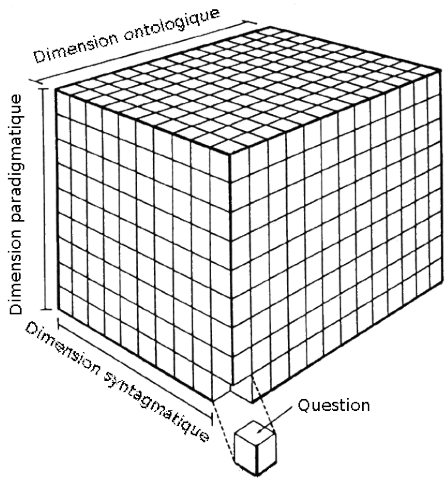

TABLE 2.1 – *Illustration des variations tridimensionnelles proposées par la GRINT.*

(comment faire),... Il s'agit d'une variation du sens de la question. On parle de variation sémantique.

En se déplaçant verticalement, c'est une variation de l'interrogatif utilisé dans la question qui a lieu. Cette variation traduit, selon le modèle, l'intention du locuteur questionnant. Il s'agit d'une variation de la forme syntaxique, autrement dit d'une variation morphosyntaxique.

En se déplaçant horizontalement, c'est une variation de la structure syntaxique de la question qui a lieu. Cette variation traduit, selon le modèle, ce que le locuteur questionnant met en cause en posant sa question. Il s'agit d'une variation morphosyntaxique.

Ainsi, chaque coupe du cube correspond à un modèle de variation morphosyntaxique pour un type ontologique donné.

Les différents types ontologiques identifiés par la GRINT sont hiérarchisés ainsi :

Questions en substantif localisant
- quantité (exemple 2.1)
- lieu (exemple 2.2)
- temps (exemple 2.3)

Questions amont/aval
- pourquoi (exemple 2.9)
- définition (exemple 2.10)
- comment (exemple 2.11)
- procédurales (exemple 2.12)

Questions entités nommées
- personne (exemple 2.4)
- lieu (exemple 2.5)
- objet (exemple 2.6)
- évènement (exemple 2.7)
- nombre (exemple 2.8)

Questions interlocutives
- interlocution (exemple 2.13)
- confirmation (exemple 2.14)

Pour illustrer ces catégories, nous proposons des exemples de réponses avérées issues d'un corpus annoté en fonction de la hiérarchie ontologique de la GRINT. Il

s'agit de questions orales transcrites d'où l'absence de ponctuation. Plus de détails sur ce corpus et son annotation sont donnés dans la section 2.4.5.

Exemple 2.1 *j' aimerais savoir combien il y a d' îles en Grèce*

Exemple 2.2 *où est le fleuve Amour*

Exemple 2.3 *à quelle époque a lieu le carnaval de Venise*

Exemple 2.4 *qui a réalisé le film le dernier milliardaire*

Exemple 2.5 *je voudrais savoir quelle est la capitale de la Mauritanie*

Exemple 2.6 *je voudrais la langue officielle de la Tanzanie*

Exemple 2.7 *je voudrais savoir quel exploit à réaliser Gérard d' Aboville (...)*

Exemple 2.8 *j' aimerais savoir quel est le numéro de téléphone de police secours en France*

Exemple 2.9 *pourquoi parle -t-on d' un pays neutre en ce qui concerne la Suisse*

Exemple 2.10 *aux Etats-Unis on parle de l' aigle bleu de quoi s' agit -il*

Exemple 2.11 *non je voudrais savoir comment James Dean est mort*

Exemple 2.12 *comment est élu le président français*

Exemple 2.13 *est -ce que vous pourriez la répéter s'il-vous-plaît*

Exemple 2.14 *j' aimerais savoir si Nanterre est une préfecture*

2.4.3 Exemples du modèle : questions de lieu, de temps et de quantité

Nous présentons ici trois coupes de cube. En plus d'illustrer le modèle qu'est la GRINT, ils permettent au lecteur une première approche de certaines variations que nous avons nous-même utilisées dans nos travaux. Jusqu'à présent ce sont les seules grilles de variation de la question qui ont été développées.

Le tableau 2.2 représente à travers l'exemple de la question *de base* "Combien pèse un bébé à la naissance ?", les variations paradigmatiques et syntagmatiques des questions dont le type ontologique est "quantité". On observe qu'il y a cinq types de variations paradigmatiques (les cinq colonnes du tableau) et cinq types de variations syntagmatiques (les cinq lignes).

Les tableaux 2.3 et 2.4 représentent ces mêmes variations à partir des question *de base* respectives "Où se trouve la Joconde ?" et "Quand ont lieu les Jeux Olympiques d'été ?". Pour ces deux types ontologiques sont définis cinq types de variations paradigmatiques (colonnes) et trois types de variations syntagmatiques (lignes).

	assertive	périphrastique	prototypique	renforcée	tonique
adverbiale	Un bébé pèse environ 3,2 kilos à la naissance ?	Je voudrais savoir combien un bébé pèse à la naissance.	**Combien pèse un bébé à la naissance ?**	Combien est-ce que pèse un bébé à la naissance ?	Un bébé pèse combien à la naissance ?
confirmative	–	Je voudrais savoir si un bébé pèse 3,2 kilos à la naissance.	Un bébé pèse-t-il environ 3,2 kilos à la naissance ?	Est-ce qu'un bébé pèse 3,2 kilos à la naissance ?	–
déterminative	Le poids d'un bébé à sa naissance est d'environ 3,2 kilos ?	Je voudrais savoir quel poids pèse un bébé à la naissance.	Quel poids pèse un bébé à la naissance ?	Quel poids est-ce que pèse un bébé à la naissance ?	Un bébé pèse quel poids à la naissance ?
nominale	Un bébé pèse quelque chose à sa naissance ?	Je voudrais savoir qu'est-ce qu'un bébé pèse à la naissance.	Que pèse un bébé à la naissance ?	Qu'est-ce que pèse un bébé à la naissance ?	Un bébé pèse quoi à la naissance ?
numérale	Un bébé pèse environ 3,2 en kilos à sa naissance ?	Je voudrais savoir combien de kilos pèse un bébé à la naissance.	Combien de kilos pèse un bébé à la naissance ?	Combien de kilos est-ce que pèse un bébé à la naissance ?	Un bébé pèse combien de kilos à la naissance ?

TABLE 2.2 – *Modèle de la GRINT pour les questions de quantité*

	assertive	périphrastique	prototypique	renforcée	tonique
adverbiale	La Joconde se trouve au Louvre ?	Je voudrais savoir où se trouve la Joconde	**Où se trouve la Joconde ?**	Où est-ce que se trouve la Joconde ?	La Joconde se trouve où ?
confirmative	C'est bien au Louvre que se trouve la Joconde ?	Je voudrais savoir si la Joconde se trouve au Louvre ou non	La Joconde se trouve-t-elle au Louvre ?	Est-ce que la Joconde se trouve au Louvre ?	Au Louvre, c'est bien là que la Joconde se trouve ?
déterminative	L'endroit où se trouve la Joconde, c'est bien le Louvre ?	Je voudrais savoir à quel endroit la Joconde se trouve	A quel endroit se trouve la Joconde ?	A quel endroit est-ce que se trouve la Joconde ?	La Joconde se trouve à quel endroit ?

TABLE 2.3 – *Modèle de la GRINT pour les questions de lieu*

	assertive	périphrastique	prototypique	renforcée	tonique
adverbiale	Les jeux olympiques d'été ont lieu en été, tous les 4 ans ?	Je voudrais savoir quand ont lieu les jeux olympiques d'été	**Quand ont lieu les jeux olympiques d'été ?**	Quand est-ce qu'ont lieu les jeux olympiques d'été ?	Les jeux olympiques d'été ont lieu quand ?
confirmative	C'est bien en été et tous les 4 ans qu'ont lieu les jeux olympiques d'été ?	Je voudrais savoir si les jeux olympiques d'été ont lieu en été, tous les 4 ans ou non	Les jeux olympiques d'été ont-il lieu en été, tous les 4 ans ?	Est-ce que les jeux olympiques d'été ont lieu en été, tous les 4 ans ?	En été et tous les 4 ans, c'est bien à ce moment qu'ont lieu les jeux olympiques d'été ?
déterminative	Le moment où ont lieu les jeux olympiques d'été, c'est bien en été, tous les 4 ans ?	Je voudrais savoir à quel moment les jeux olympiques d'été ont lieu	A quel moment ont lieu les jeux olympiques d'été ?	A quel moment est-ce qu'ont lieu les jeux olympiques d'été ?	Les jeux olympiques d'été ont lieu à quel moment ?

TABLE 2.4 – *Modèle de la GRINT pour les questions de temps*

2.4.4 À chaque variation sa réponse escomptée

La GRINT associe à chaque forme linguistique de question, la forme de la réponse escomptée. La variation de la question selon la dimension syntagmatique traduit ce que le locuteur met en cause à travers sa question tandis que la variation selon la dimension paradigmatique traduit l'intention du locuteur. La mise en cause du locuteur questionnant n'impose pas clairement de forme de réponse au locuteur répondant. En revanche, elle laisse plus ou moins la possibilité au répondant de ne pas savoir répondre.

Le tableau 2.5 montre, pour les questions de quantité, l'intention du locuteur selon la variation de la question considérée et un exemple de réponse escomptée pour une demande de prix.

Type paradigmatique	**Intention**	**Réponse escomptée**
adverbiale (Combien...)	neutralité	son prix est de X
numérale (Combien de ...)	numérisation	c'est X
déterminative (Quel ...)	estimation	pas cher
nominale (Que ...)	estimation	environ X
confirmative (*inversion du sujet*)	confirmation/infirmation	oui

TABLE 2.5 – *Lien entre l'intention du locuteur et la variation paradigmatique de la question - Exemple des questions de quantité pour une demande de prix.*

Le tableau 2.6 montre, pour les questions de quantité, l'élément mis en cause.

Type syntagmatique	**Mise en cause**
prototypique (Combien...)	repérage quantitatif
tonique (... combien)	réalité de la valeur
renforcée (... est-ce que...)	quantification
périphrastique (... je vd savoir ...)	connaissance de la réponse
assertive (∅)	confirmation/infirmation

TABLE 2.6 – *Lien entre la mise en cause de la question et la variation syntagmatique de celle-ci - Exemple des questions de quantité.*

2.4.5 Un modèle testé sur des données avérées

Le modèle qu'est la GRINT a été confronté à des données réelles. En effet, un corpus de questions a été annoté pour situer les questions parmi les classes ontologiques de la GRINT. Ce travail réalisé au LIUM par Carole Lailler a été présenté dans [Lailler, 2009]. Le corpus annoté, dit corpus RITEL, est un corpus de requêtes utilisateurs du LIMSI [Galibert et al., 2005a; Rosset & Petel, 2006] en situation de Dialogue Oral Homme Machine (DOHM) collecté de septembre 2004 à janvier 2005 dans lequel 13 personnes qui ont reçu, à titre d'exemple, une

liste différente d'environ 300 questions ont appelé un serveur, ce qui a donné lieu à 582 dialogues qui comportent 5362 énoncés. Les locuteurs obéissaient à une consigne simple : "le système a beaucoup de mal à reconnaître et à comprendre ce que vous lui dites, mais il a besoin de données pour apprendre : essayez d'obtenir des réponses et de lui expliquer ce que vous cherchez".

Type ontologique	**Taux de couverture**
Quantité	8,55%
Lieu	3,39%
Temps	7,52%
Pourquoi	0,13%
Définition	1,08%
Comment	0,2%
Procédurales	0,34%
EN personne	13,30%
EN lieu	13,13%
EN objet	7,25%
EN évènement	0,61%
EN nombre	0,51%
Interlocution	0,51%
Confirmation	2,81%
Autre	40,67%

TABLE 2.7 – *Taux de couverture des classes ontologiques de la GRINT dans le corpus RITEL (avec EN pour Entité Nommée).*

Nous voyons dans cette théorie une réelle volonté de ne plus voir la réponse comme un énoncé qui apporte une nouvelle information, mais de la considérer dans sa relation avec la question qui lui a été posée. Nous l'avons présentée dans une section s'intitulant "Théorie de la réponse en interaction". En fait, il s'agit plus d'une théorie de la question en interaction, mais nous l'avons compris, question et réponse sont définitivement indissociables.

2.5 DÉFINITION *des* RÉPONSES

Nous avons vu la question telle qu'elle est considérée dans le domaine des systèmes de question-réponse et d'un point de vue interactif. Pour chacun de ces point de vue, le même terme de "réponse" est utilisé alors qu'il ne désigne pas la même chose. Nous proposons ainsi d'utiliser deux termes distincts pour désigner ces deux "réponses". Ces termes seront utilisés selon les définitions données ci-dessous dans l'ensemble du présent document.

Définition 2.1 ***Énoncé-réponse*** *: (en anglais, answering-utterance) ensemble de la production lan-*

gagière fournie pour répondre à une question. Il peut s'agir d'un mot, d'un ensemble de mots, d'une phrase ou d'un ensemble de phrases.

Nous utilisons à présent le terme de *réponse* pour désigner un énoncé-réponse.

Par exemple, les réponses A à G présentées dans les exemples ci-après sont toutes des réponses.

Exemple 2.15 *[Question] Où est-ce que se trouve la Joconde ?*

Exemple 2.16 *[Réponse A] au Louvre*

Exemple 2.17 *[Réponse B] La Joconde se trouve au Louvre.*

Exemple 2.18 *[Réponse C] Elle se trouve au Louvre.*

Exemple 2.19 *[Réponse D] La Joconde est au Louvre.*

Exemple 2.20 *[Réponse E] C'est au Louvre que la Joconde se trouve.*

Exemple 2.21 *[Réponse F] La Joconde est une œuvre de Léonard de Vinci exposée au Louvre.*

Exemple 2.22 *[Réponse G] La Joconde est au Louvre depuis 1945. Elle a été exposée au palais des Tuileries entre 1800 et 1804, (...)*

Les énoncés-réponses proposés dans les exemples 2.16 à 2.22 montrent la diversité de forme que peut prendre un énoncé-réponse.

Dans le cadre d'une campagne d'évaluation des systèmes de question-réponse, la réponse A sera à privilégier [TREC, site web]. Cette réponse n'est constituée que du "strict minimum" : l'information qui répond à la question. Quand, dans le cadre de campagne d'évaluation ou de développement de systèmes de question-réponse, on parle de "réponse", on se réfère en fait à cet élément. Nous utiliserons le terme spécifique d'*information-réponse* pour le désigner et le définissons comme suit :

Définition 2.2 ***Information-réponse*** *: (en anglais, answering-information) nouvelle information (par rapport à celles contenues dans la question) qui correspond soit au type attendu de la réponse, soit à un aveu d'incompétence ("je ne sais pas" par exemple).*

2.6 Produire une réponse en interaction

2.6.1 Définition de réponse en interaction pour les systèmes de question-réponse

Nous avons vu que les réponses données par des humains et par des systèmes de question-réponse sont bien différentes. Si il est actuellement difficilement imaginable qu'un système puisse produire des réponses telles que celles d'humains, il n'est en revanche pas dénué de sens de penser que ces systèmes puissent renvoyer

des réponses plus complexes qu'une information-réponse seule.
Nous avons vu, au travers de la notion d'interaction et de réponse en interaction, que répondre à une question ne se limite pas à donner une information. Comme dans toute interaction, la situation d'énonciation et la nécessité d'être pertinent jouent un rôle dans la production langagière. Dans le cas plus particulier de l'interaction question-réponse, un lien existe entre question et réponse et a été mis en valeur dans les travaux portant sur la Grammaire Interactive. En effet, celle-ci propose une modélisation de la question dans son interaction avec la réponse et établit un lien entre la forme linguistique de la question et la réponse qu'elle présuppose.
Notre objectif est de permettre à un système de question-réponse quel qu'il soit de fournir une réponse en interaction. Notre objectif n'est pas de permettre de fournir des réponses telles que celles proposées par les humains. Il s'agit plutôt de rendre les réponses des systèmes plus agréables à l'utilisateur, plus naturelles. Nous pensons que non seulement une réponse qui se limite à une information est lacunaire mais qu'en plus elle n'est pas à sa place au sein d'un échange communicationnel tel que ceux dont nous avons parlé dans ce chapitre. Ceci est une première motivation à nos travaux. Une seconde vient du fait que les utilisateurs produisent des énoncés en langue naturelle pour répondre à une question, que le contexte soit oral ou écrit, formel ou informel. Si les utilisateurs peuvent poser leur question en langue naturelle, il semble cohérent de fournir une réponse du même type.
La question que nous nous posons alors est celle de la définition d'une telle réponse. Nous avons observé des réponses de systèmes et des réponses d'humains pour savoir si de telles réponses sont disponibles auquel cas nous pourrions prendre exemple sur celles-ci.

2.6.2 Production de réponse dans le domaine de la question-réponse

Les premières sections de ce chapitre ont montré que les réponses des systèmes de question-réponse sont composées uniquement d'une information-réponse (ou plusieurs dans le cas des questions listes). Nous définissons de tels énoncés-réponse (ou réponse) comme des réponses succinctes :

Définition 2.3 ***Réponse succincte*** *: réponse qui est constituée uniquement d'une information-réponse.*

Certains systèmes, comme le système RITEL, proposent des réponses plus élaborées contenant aussi un retour sur les mots-clefs compris par le système. Mais nous avons pu voir que ces réponses ne sont pas formulées de façon naturelle.

Des travaux traitent de la transformation de la question sous forme de réponse [Mendes & Moriceau, 2004; Anaya & Kosseim, 2003]. Mais le but est d'améliorer la phase de recherche de l'information-réponse dans les documents et la forme linguistique des réponses n'est pas étudiée. De plus, il ne s'agit pas de voir la

réponse dans une interaction mais de savoir comment l'information-réponse peut se présenter à l'intérieur de documents.

2.6.3 Réponses d'humain : y a-t-il des corpus disponibles ?

Nous avons vu section 2.2.1 que les réponses d'humains disponibles sur des sites de question-réponse collaboratifs ne sont pas utilisables car les formes des réponses sont bien trop diverses pour en tirer des généralités. De plus, bon nombre de réponses contiennent des informations qu'un système de question-réponse ne peut pas produire.

2.6.4 Besoin d'un corpus de réponses

Puisqu'il n'existe actuellement pas de corpus de réponses en langue naturelle qui puissent être produites par des systèmes de question-réponse ni sur lesquelles nous pourrions nous appuyer pour définir ce que peut être un réponse en langue naturelle, en interaction, nous avons opté pour la construction d'un corpus de réponses. Notre idée est d'avoir des réponses qui ne contiennent pas "trop d'informations" de telle sorte que les systèmes de question-réponse puissent les engendrer et qui soit produites par des humains donc en principe, naturelle.

Notre idée est ainsi de poser des questions à des humains et de collecter leur réponse. Si l'on s'appuie sur la GRINT, alors il est indéniable que la forme linguistique des questions posées jouent un rôle sur la forme que peuvent prendre les réponses. Nous nous sommes ainsi posé comme base qu'il faut poser des questions de formes linguistiques variées afin de collecter un corpus de réponses non biaisé par la forme de la question. Notre travail s'est donc déroulé en trois phases principales. D'une part, nous avons travaillé à la construction d'un corpus de question de formes linguistiques variées et contrôlées. Puis nous avons soumis ces questions à des locuteurs natifs du français. Enfin, une fois le corpus collecté, nous avons observé et analysé ces réponses.

Conclusion du chapitre

Nous avons vu au cours de ce chapitre qu'une réponse ne peut être envisagée hors de son interaction avec la question et qu'une réponse en interaction ne se limite pas à donner une information-réponse seule comme le font actuellement la plupart des systèmes de question-réponse. Parce qu'il n'existe aucun corpus de réponses en interaction qui non seulement ne soient pas succinctes mais qui puissent également être produites par des systèmes de question-réponse, nous avons choisi de constituer notre propre corpus de réponse. Pour ce faire, suivant les principes exposés par le modèle de la GRammaire INTeractive, nous avons montré l'intérêt d'avoir des réponses à des questions de formes linguistiques variées. Le chapitre

suivant s'attache ainsi à établir un état de l'art des variations linguistiques des questions. C'est à partir de ces connaissances qu'un corpus de questions a été élaboré et soumis à des locuteurs du français afin de collecter leurs réponses.

3 UNE RÉPONSE POUR QUELLE QUESTION ?

SOMMAIRE

Ce chapitre présente une étude de l'état de l'art sur les variations linguistiques des questions. Nous identifions les différents éléments qui peuvent varier dans une question intervenant dans une interaction avec une réponse. Ces éléments correspondent à des caractéristiques de la question elle-même (son type sémantique, sa syntaxe, l'interrogatif utilisé, le type de son verbe principal et l'objet sur

lequel elle porte) mais aussi à des caractéristiques de la réponse attendue par la question (son type, sa nature unique ou multiple et sa valence). Ce travail sert de base pour la construction du corpus de questions que nous présenterons dans le chapitre suivant.
Ce chapitre correspond à un travail d'état de l'art sur la variation de la question.

This chapter presents a study of works on question linguistic variations. We identify different elements which can variate in a question considerated in interaction with the answer. Those elements concern the question itself (its semantic type, its syntactic form, the principal verb and the interrogative pronoun used and its focus) or the answer expected by the question (its semantic type, its *nature*, single or multiple, and its *valence*). This work is the base used for building the question corpus presented in the next chapter.

3.1 INTRODUCTION

Afin d'obtenir un corpus de réponses, nous avons choisi de collecter un corpus en demandant à des locuteurs de répondre à des questions. Parce que nous voyons la réponse en interaction avec la question, nous pensons que la forme linguistique de la question joue un rôle sur la forme linguistique de la réponse. Notre idée est donc d'utiliser un corpus de questions dont les formes linguistiques sont variées. Nous voulons être en mesure de contrôler les différentes caractéristiques linguistiques des questions. Pour cela, nous nous sommes intéressée aux travaux portant sur les variations de la question et en particulier sur la classification des questions. À travers cet état de l'art, nous cherchons à identifier différents paramètres de variation des questions et à les utiliser pour la construction de notre corpus de question.

Qu'est-ce qu'une question ? Une question est une suite de mots. Prenons comme exemple la question 3.1. Elle est composée de du groupe interrogatif "Dans quelle ville", du verbe principal "est" et du groupe nominal "la Tour Eiffel" qui est l'objet sur lequel la question porte. L'ordre et la réalisation de ses groupes peut changer comme on le voit dans l'exemple 3.2. Si l'on change la locution interrogative, le verbe ou encore l'objet de la question, le sens de cette question change lui aussi. En remplaçant "Dans quelle ville" par "De quelle couleur" (exemples 3.1 et 3.3), la question ne porte plus sur le lieu où l'objet de la question se situe mais sur l'une de ses caractéristiques : sa couleur. C'est le *type sémantique* de la question qui a changé. Posant l'une et l'autre des deux questions 3.1 et 3.4, on observe que le type de réponse auquel on s'attend change lui aussi. Si l'on change l'objet de la question "la Tour Eiffel" par "le Rhin" (exemple 3.5), le type sémantique de la question reste le même, sa formulation aussi. En revanche, l'objet de la question est de nature différente. En effet, on ne peut donner une unique réponse cette dernière question puisque le Rhin traverse plusieurs villes.

Exemple 3.1 *Dans quelle ville est la Tour Eiffel ?*

Exemple 3.2 *La Tour Eiffel se trouve dans quelle ville ?*

Exemple 3.3 *De quelle couleur est la Tour Eiffel ?*

Exemple 3.4 *Dans quel musée est la Tour Eiffel ?*

Exemple 3.5 *Dans quelle ville est le Rhin ?*

Ce que nous montrons dans ce chapitre n'est pas une vision de la question comme un énoncé isolé. Nous considérons la question comme une production langagière "incomplète", en quelque sorte, puisqu'elle attend d'être complétée par une réponse. C'est ce que traduit ce chapitre qui traite à la fois des variations de la question en elle-même et des variations de la réponse avec laquelle elle est inexorablement en relation. Les sections suivantes sont ainsi consacrées tant à la

variation de la question et notamment à son type sémantique (section 3.3), à sa réalisation lexicale (section 3.4) et à sa forme morphosyntaxique (section 3.5), qu'à la variation de la réponse et notamment à son type attendu (section 3.2), à sa nature (section 3.6) et à sa valence (section 3.7).

3.2 TYPE ATTENDU DE LA RÉPONSE

Dans le domaine des systèmes de question-réponse, l'un des indices détectés au sein de la question est le type attendu de la réponse ([Monceaux & Robba, 2002], [Ligozat, 2004]). Il s'agit de déterminer si la réponse attendue est une entité nommée ou non, le cas échéant de quel type d'entité nommée il s'agit ou encore de combien d'éléments la réponse devrait être composée. En d'autres termes, on cherche à prédire la forme que la réponse prendra.

Nous nous intéressons aux travaux relatifs au type attendu de la réponse pour les classifications qui y sont proposées. Nous supposons en effet que la réponse à une question prendra des formes différentes en fonction de ce facteur. Par exemple, pour les deux questions 3.6 et 3.7 qui se ressemblent d'un point de vue surfacique (`interrogatif verbe être groupe nominal`), on identifie des types attendus de la réponse différents et les réponses peuvent en effet prendre des formes très différentes.

Exemple 3.6 *Qui est le Président de la République ?*

Exemple 3.7 *Quelles sont les règles du jeu de pétanque ?*

À la question 3.6, une réponse minimale serait *Nicolas Sarkozy* alors qu'à la question 3.7, la réponse attendue est au minimum un paragraphe constitué de plusieurs phrases.

3.2.1 Réponses *élémentaires*

Assez naturellement les recherches se sont d'abord portées sur les questions dont le type attendu de la réponse est une information "simple", courte, précise, et parmi ces réponses courtes, les entités nommées (EN). De nombreux travaux se sont développés autour de la notion d'EN : projets scientifiques, conférences dédiées, campagnes d'évaluations dont une tâche traite des EN,...

[Ligozat, 2004] définit une EN comme :

> ❝ une expression qui, dans un contexte particulier, identifie de façon unique une entité telle qu'une personne, un lieu, une organisation, une date ❞

Les noms propres, par exemple, correspondent à une sous-classe importante d'EN mais aussi les noms collectifs (*les Italiens*), les noms de maladies (*la grippe*), le

personnages (*Tintin*), les sigles (*CNRS*) ou encore les noms de religion (*Bouddhisme*)... Les EN peuvent aussi être imbriquées. Par exemple, *Université Paris Sud* est une EN qui contient l'EN *Paris*.

Il existe de nombreux travaux sur la détection d'EN et notamment des campagnes d'évaluation consacrées à leur détection ([Tjong Kim Sang & De Meulder, 2003]). D'autres travaux portent sur la classification des questions en fonction du type d'EN attendu. Nous nous intéressons plus particulièrement à ces classifications. On distingue différentes approches. Certaines se basent sur des corpus pour élaborer une classification. [Paik et al., 1996] propose une classification élaborée à partir d'une étude de corpus (Wall Street Journal) composée de 9 classes et 30 sous-classes. [Wolinski et al., 1995] définit une classification en 50 thèmes pour un corpus de dépêches de l'Agence France Presse. D'autres approches se basent sur des modèles linguistiques qui correspondent en fait à des ensembles de règles. [Grass, 2000] propose une classification pour la traduction qui se limite aux noms propres. Cette classification est basée sur les travaux en linguistique décrits dans [Bauer, 1985]. Toujours dans le cadre de la traduction automatique, [Fourour & Morin, 2003a] proposent différentes classifications. Notamment ils proposent une classification basée sur celle de [Grass, 2000] qu'ils étendent en ajoutant des catégories. À cette catégorisation dite référentielle, ils ajoutent une classification graphique prenant en compte la casse et le nombre d'unités lexicales dont l'EN est constituée et distinguant les sigles des autres EN.

Nous observons qu'il existe bon nombre de classifications qui dépendent notamment de la tâche à accomplir et du domaine d'application et bien qu'il n'existe pas de consensus, une classification largement répandue est celle proposée par MUC (Message Understanding Conference [MUC, site web]). En 1997 apparaît la tâche *entités nommées* au sein de cette campagne d'évaluation. Une classification est alors proposée [Grishman & Sundheim, 1996]. Elle distingue 3 grands types d'entités nommées : (1) ENAMEX qui désignent des noms de personnes, de villes,... (2) TIMEX pour les dates, l'heure,... (3) NUMEX pour les montants, pourcentages,... [Ferret et al., 2001] reprend cette classification, l'affine, hiérarchise les étiquettes et propose l'arbre d'étiquettes repris figure 3.1.

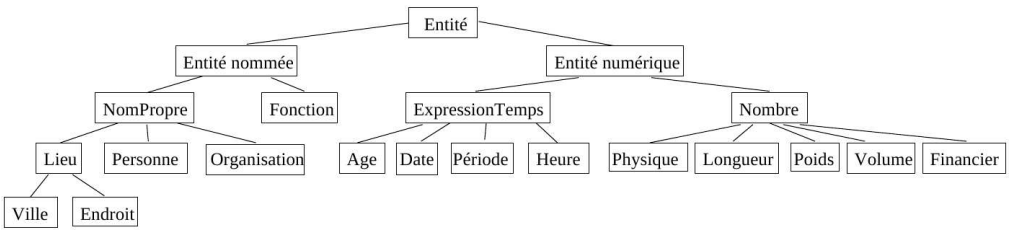

FIGURE 3.1 – *Hiérarchie de types de réponses et de catégories sémantiques proposée par [Ferret et al., 2001]*

Cette classification est intéressante car elle couvre l'ensemble des types d'entités nommées, est plus fine que la classification MUC tout en restant relativement simple (20 classes pour 4 niveaux de hiérarchie). De plus elle n'a pas été

élaborée sur la base d'un corpus et n'est donc pas biaisée par celui-ci alors que chez [Paik et al., 1996], par exemple, on observe la catégorie `Scientifique` subdivisée en `maladies`, `drogues` et `médicaments` mais pas de sous-catégorie `molécule` par exemple. Pourtant elle est considérée comme EN dans des travaux par exemple de question-réponse en domaine médical (par exemple chez []).

Nous notons que les questions dont nous parlons attendent comme réponse une information précise : il n'y a pas forcément une unique réponse à la question mais la question attend un seul type d'information et porte sur une information unique et précise. Par exemple, la question 3.8 attend une unique information, d'un seul type précis alors que la question 3.9 attend plusieurs informations en réponse et que la question 3.10 attend différents types d'informations.

Exemple 3.8 *Où est la Joconde ?*

Exemple 3.9 *Quels sont les pays membres de l'Union Européenne ?*

Exemple 3.10 *Quand et où est né le Dalai Lama ?*

Nous définissons alors les réponses *élémentaires* comme suit :

Définition 3.1 ***Réponse élémentaire*** *: réponse qui ne consiste qu'en un unique type d'information (un lieu, un date, un nom... mais pas des horaires, des coordonnées,...).*

Les types de questions proposées par [Ferret et al., 2001] ne sont pas les seuls qui attendent une réponse élémentaire. Les questions fermées ou booléennes en font aussi partie. En effet, ce sont des questions auxquelles on peut répondre par "oui", "non" comme on peut le voir dans les exemples de réponse R1 à R3 à la question 3.11.

Exemple 3.11 *Q : Le Louvre est-il à Paris ?*

Exemple 3.12 *R1 : oui*

Exemple 3.13 *R2 : en effet*

Exemple 3.14 *R3 : non*

[Benamara, 2004] propose une classification des réponses dites "atomiques". Nous reprenons par la figure 3.2 sa figure 1.3 en y ajoutant des exemples illustratifs. On observe que [Benamara, 2004] inclut dans les réponses atomiques, les réponses correspondant à des listes de valeurs (entité ou numérique). Ceci signifie que les réponses atomiques de Benamara ne correspondent pas toutes à des questions qui attendent des réponses élémentaires. La différence de point de vue vient du fait que nous nous centrons sur la quantité d'information véhiculée par la réponse. Notre définition de réponse élémentaire implique qu'on s'attend à une et une seule information réponse. Les questions "liste" sont donc pour nous des réponses non

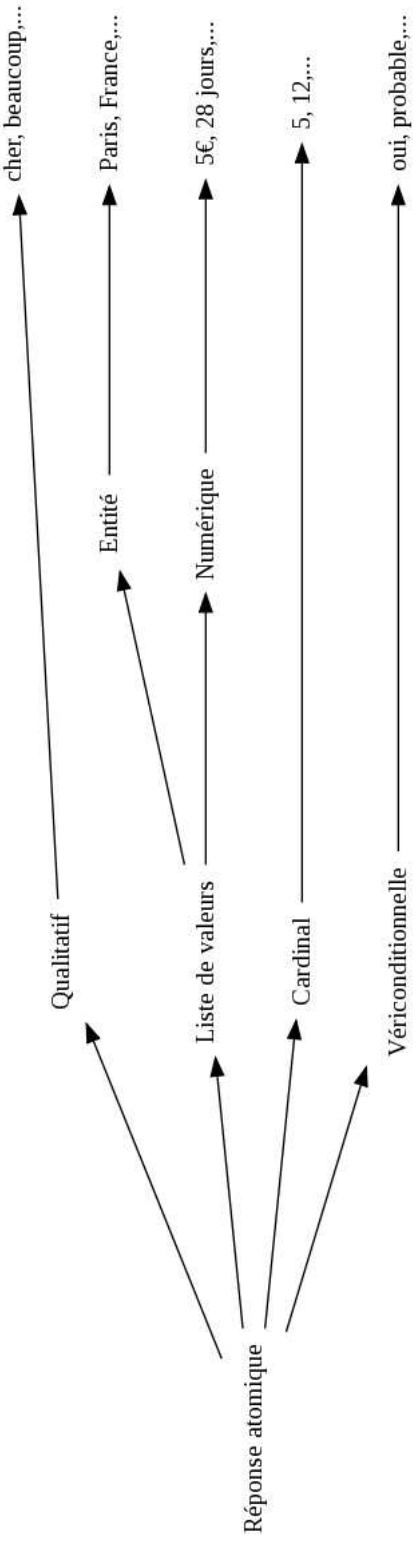

FIGURE 3.2 – *Classification des réponses atomiques proposée par [Benamara, 2004]*

élémentaires sur lesquelles nous reviendrons dans la section suivante dédiée aux réponses non élémentaires.

D'autre part, il est intéressant d'observer la distinction faite chez [Benamara, 2004] entre les réponses numériques, qualitatives et cardinales. Bien qu'elles correspondent à des questions qui portent sur une quantité, elles sont en effet de natures différentes.

En résumé, on peut considérer qu'il existe différents types de réponses élémentaires : (1) les entités nommées, elles-mêmes subdivisables en différentes catégories, (2) les réponses fermées (ou booléennes), (3) les réponses quantitatives (dont les réponses numériques, qualitatives et cardinales).

3.2.2 Réponses *non élémentaires*

De plus en plus, les recherches portent sur des questions dont la réponse n'est pas aussi précise et circonscrite que celles dont nous venons de parler. Il peut s'agir de listes (d'entités nommées ou pas) comme le montrent les exemples 3.15 et 3.16. Dans ce cas, on s'attend à *des* informations réponses et non plus à *une* information réponse. Il peut aussi s'agir de procédure ou de définition comme le montrent les exemples 3.17 et 3.18. Dans ce cas la réponse prend la forme d'une ou plusieurs phrases. [Benamara, 2004] nomme ces réponses les réponses narratives parmi lesquelles sont distinguées les réponses "procédure" (exemple 3.17), définition (exemple 3.18), description (exemple 3.19), cause/raison (exemple 3.20), but (exemple 3.21) et évaluation/comparaison (exemple 3.22).

Exemple 3.15 *Quelles sont les sept merveilles du monde ?*

Exemple 3.16 *Quels livres a écrit Eric-Emmanuel Schmitt ?*

Exemple 3.17 *Comment préparer une bûche de Noël ?*

Exemple 3.18 *Qu'est-ce qu'un carreau à la pétanque ?*

Exemple 3.19 *À quoi ressemble la fusée Arianne ?*

Exemple 3.20 *Pourquoi le ciel est bleu ?*

Exemple 3.21 *Pourquoi l'âge de la retraite a-t-il été repoussé ?*

Exemple 3.22 *Quelle est la différence entre tirer et pointer ?*

De notre point de vue, les réponses non élémentaires sont les réponses liste et les réponses narratives. Ces questions sont largement traitées et définies par les études s'intéressant aux systèmes coopératifs. De tels systèmes se veulent proposer

des informations supplémentaires en marge de la réponse (exemples pour la question 3.23 des réponses 3.24 et 3.25).

Exemple 3.23 *Q : Qui est le président de la république ?*

Exemple 3.24 *R : Nicolas Sarkozy. Il a été élu en 2007.*

Exemple 3.25 *R : Nicolas Sarkozy. Avant il s'agissait de Jacques Chirac.*

Il ne s'agit pas de répondre en donnant plusieurs informations du même type (comme c'est le cas pour les *questions liste*) ou encore une information par nature complexe (comme dans le cas d'une définition) mais d'ajouter des informations à la réponse stricte (une information d'un autre type dans le cas de la réponse 3.24, une information du même type pour la réponse 3.25).

Si des travaux portent sur les réponses coopératives, aucun travail, à notre connaissance, ne traite de questions qui impliqueraient plus particulièrement une réponse coopérative. Pourtant on pourrait supposer qu'une question telle que la question 3.26 demande une réponse coopérative tandis que ce n'est le cas ni de la question 3.23 donnée précédemment, ni de la question 3.27 suivante.

Exemple 3.26 *Je voudrais en savoir plus sur qui est le président de la république.*

Exemple 3.27 *Le président de la république, c'est qui ?*

Le *type attendu de la réponse* est lié à d'autres caractéristiques de la question et notamment à son type sémantique. En effet, une question de type sémantique "quantité" porte nécessairement sur un objet mesurable, "pesable" (exemple 3.29). De même, une question dont le type attendu de la réponse est une entité numérique "poids" est une question de type sémantique "quantité" (exemple 3.28). L'implication est vraie aussi dans l'autre sens : les questions dont le type attendu de la réponse est un poids, un nombre, une distance,... sont des questions de quantité. De même les questions dont le type attendu de la réponse est une date (complète ou bien limitée à l'un de ses composants, par exemple, un mois, une année,...) sont des questions de temps.

Exemple 3.28 *Combien pèse une brique de lait ?*

Exemple 3.29 **Combien pèse la pétanque ?*

Nous nous intéressons à présent plus précisément au type sémantique de la réponse.

3.3 TYPE SÉMANTIQUE DE LA QUESTION

[Vicedo et al., 2002], dans le cadre de leur participation à la campagne d'évaluation TREC-10, considère sept types sémantiques de questions (`personne`,

`groupe`, `quantité`, `lieu`, `manière`, `raison`) auquel il ajoute la catégorie `none` pour toutes les questions d'un *autre* type sémantique (par exemple les questions de définition, 3.30)

Exemple 3.30 *Qu'est-ce que faire un bec à la pétanque ?*

Pour une question donnée, [Vicedo et al., 2002] attribue un ou plusieurs types en fonction de l'interrogatif utilisé. C'est l'interrogatif qui permet de classer la question. Mais alors qu'en est-il des questions fermées ? Elles ne contiennent pas d'interrogatif, le marqueur de l'interrogation étant soit la prosodie soit l'inversion du sujet. Or ces marqueurs interrogatifs ne contiennent pas d'indices sémantiques.

3.3.1 Le cas des questions fermées

Le type sémantique de la question peut être déterminé par l'interrogatif utilisé, sauf dans le cas des questions fermées. Le choix fait dans [Vicedo et al., 2002] est qu'une question fermée telle que 3.31 doit être classée comme une question `none` puisqu'il n'y a pas d'interrogatif (l'interrogation se réalisant au travers d'une inversion du sujet). Cependant, ce point de vue ne prend pas en compte le fait que cette question évalue une quantité et qu'on peut y répondre par *Oui, environ* ce qui n'est pas le cas pour la question 3.33.

Exemple 3.31 *Q : Une boule de pétanque de compétition pèse-t-elle 700 grammes ?*

Exemple 3.32 *R : Oui, environ.*

Exemple 3.33 *Q : Une boule de pétanque de compétition est-elle creuse ?*

Exemple 3.34 *R : * Oui, environ.*

Une comparaison de ces deux questions fermées permet de mettre en évidence que le type sémantique de la question joue un rôle sur la réponse que ces questions acceptent. Il est donc intéressant de classer les questions dans des catégories sémantiques différentes. Une classification selon le type sémantique de la question doit donc, selon nous, considérer la sémantique de la question sans prendre en compte la réalisation linguistique de celle-ci.

Comme le type attendu de la réponse, le type sémantique de la question est fortement lié à l'objet de la question. On ne peut poser une question de lieu à propos d'un objet conceptuel tel que la liberté (cf. exemple 3.35). Le type sémantique de la question joue aussi un rôle au niveau lexical et notamment sur le verbe utilisé dans la question.

Exemple 3.35 **Combien pèse la liberté ?*

3.4 TYPE DU VERBE PRINCIPAL DE LA QUESTION ET AUTRES VARIATIONS LEXICALES

On peut noter que certains verbes sont en forte relation avec le type sémantique de la question parce qu'ils sont porteurs de sens alors que d'autres sont plus neutres sémantiquement. Par exemple, l'auxiliaire `être` peut être utilisé pour les questions de lieu et de temps (exemples 3.36 et 3.37) mais pas pour une question de quantité (exemple 3.38) tandis que le verbe `se trouver` utilisable dans une question de lieu (exemple 3.39) est incompatible avec les questions de temps et de quantité (exemples 3.40 et 3.41).

Exemple 3.36 *Où est le salon du chocolat ?*

Exemple 3.37 *Quand est le salon du chocolat ?*

Exemple 3.38 **Combien est le salon du chocolat ?*

Exemple 3.39 *Où se trouve le salon du chocolat ?*

Exemple 3.40 **Quand se trouve le salon du chocolat ?*

Exemple 3.41 **Combien se trouve le salon du chocolat ?*

Pour la question de quantité utilisée dans nos exemples, un verbe possible est le verbe `coûter` (exemple 3.42.

Exemple 3.42 *Combien coûte le salon du chocolat ?*

Pour la question de temps utilisée dans nos exemples, un verbe possible est le verbe `avoir lieu` (exemple 3.43.

Exemple 3.43 *Quand a lieu le salon du chocolat ?*

En d'autres termes, pour une question portant par exemple sur le lieu où se trouve le salon du chocolat, un locuteur questionnant peut sélectionner soit un verbe spécifique à la localisation d'un objet du monde tel que "se trouver" ou encore "prend place", soit l'auxiliaire "être". Il y a donc possibilité de faire varier une question simplement en changeant son verbe principal qui peut être un verbe "spécifique" au type sémantique ou bien un verbe neutre tel qu'un auxiliaire. Cette variation n'est cependant pas possible pour toutes les questions. En effet, la question de quantité prise en exemple n'accepte pas l'utilisation d'un auxiliaire.

Le verbe n'est pas le seul élément lexical qui peut varier et qui est en lien avec le type sémantique de la question. Les questions 3.44 et 3.45 par exemple, diffèrent

d'un terme et font partie de deux classes sémantiques différentes, respectivement quantité et lieu.

Exemple 3.44 *À quel endroit se déroulent les Jeux Olympiques de 2012 ?*

Exemple 3.45 *À quelle date se déroulent les Jeux Olympiques de 2012 ?*

Le pronom interrogatif ou locution interrogative et le type sémantique de la question sont aussi liés. Le choix de l'interrogatif diffère pour les questions de quantité et pour les questions de lieu par exemple (exemples 3.46 et 3.47).

Exemple 3.46 *Combien pèse une boule de pétanque ?*

Exemple 3.47 **Combien a lieu le prochain championnat du monde ?*

Il peut même être le seul indice de la catégorie sémantique de la question comme le montre le couple d'exemples 3.48 et 3.49.

Exemple 3.48 *Quand a lieu le prochain championnat du monde ?*

Exemple 3.49 *Où a lieu le prochain championnat du monde ?*

Les travaux de Daniel Luzzati sur la GRammaire INTeractive (GRINT) dont nous avons déjà parlé dans la chapitre 2 proposent un modèle de variation de la question dans lequel une dimension de variation concerne justement le pronom interrogatif de celle-ci. Nous présentons à présent la variation de la question du point de vue de la GRINT.

3.5 Variations morphologiques et syntaxiques de la question

En français, une interrogation peut se réaliser sous différentes formes. Si le questionneur cherche à obtenir une information telle que le nom de la ville où se trouve la Tour Eiffel, plusieurs formulations sont possibles comme on le voit dans les exemples 3.50, 3.51 et 3.52 suivants.

Exemple 3.50 *Dans quelle ville se trouve la Tour Eiffel ?*

Exemple 3.51 *Où est la Tour Eiffel ?*

Exemple 3.52 *Quelle est la ville où se trouve la Tour Eiffel ?*

Si certaines formulations peuvent paraître plus lourdes que d'autres, elles n'en demeurent pas moins correctes tant d'un point de vue syntaxique que sémantique.

[Luzzati, 2001] propose un modèle de variation de la question. (la GRammaire INTeractive ou GRINT). Ce modèle est un modèle descriptif du lien qui existe entre la forme morphosyntaxique d'une question et celle de la réponse escomp-

tée. Ce modèle dont nous avons déjà parlé section 2.4.1 est celui sur lequel nous nous appuyons pour émettre l'hypothèse de l'existence d'un lien entre les formes linguistiques de surface d'une question et de sa réponse. Ici, nous regardons la GRINT d'un autre point de vue. Nous considérons les grilles de variation morphosyntaxique qu'elle établit et proposons d'observer les types qu'elle décrit.

La GRINT propose une grille de variation des questions selon 3 dimensions : (1) la *dimension ontologique* qui peut être mise en correspondance avec le type sémantique de la question et le type escompté de la réponse, (2) la *dimension paradigmatique* qui décrit une variation de l'interrogatif de la question, (3) et la *dimension syntagmatique* qui concerne la structure syntaxique globale de la question.

3.5.1 La dimension ontologique

La dimension ontologique est décrite comme ayant le focus sur l'objet visé de la question. Quatre groupes ontologiques sont proposés : les questions dont le but communicatif est phatique et interactif (exemple 3.53), les questions dont le but est de questionner sur un objet du monde (exemples 3.54, 3.55 et 3.56), les questions dont le but est le nommage d'un objet du monde (3.57) et les questions dont le but est explicatif (3.58 et 3.59).

Exemple 3.53 *Tu as compris ?*

Exemple 3.54 *Combien de satellites a Jupiter ?*

Exemple 3.55 *Quand est a été crée l'ATALA ?*

Exemple 3.56 *Où est la Joconde ?*

Exemple 3.57 *Qui est l'architecte de la Tour Eiffel ?*

Exemple 3.58 *Pourquoi le soleil est-il jaune ?*

Exemple 3.59 *Comment les points sont-ils comptés à la pétanque ?*

Parmi ces quatre groupes, les questions *phatiques* sont des questions relevant d'un domaine assez fermé. Il s'agit de meta-dialogue : questions de politesse, de vérification de la compréhension de son interlocuteur ou bien de sa propre compréhension. Ces questions ne demandent pas une recherche d'information dans une base documentaires mais bel et bien des capacités d'auto-évaluation, d'auto-observation. Répondre à une telle question n'est pas prévu au sein des systèmes de question-réponse, ce n'est pas leur but. Un agent conversationnel ou dialogique en revanche sera plus souvent doté de telles capacités.

Les questions *explicatives* comprennent aussi les questions dites procédurales (3.60). En fait, ce groupe rassemble toutes les questions "pourquoi" et "comment".

Exemple 3.60 *Comment faire de la mayonnaise ?*

Ces questions sont des questions non élémentaires (on observe là le lien fort qui existe entre la syntaxe d'une question : son pronom interrogatif, et la réponse attendue : une phrase au minimum) dont nous parlions à la section 3.2.

Les questions de nommage ne sont autres que les questions de type "entité nommée". Ces questions ont été très largement étudiées dans la littérature du domaine de la question-réponse.

Enfin les questions dont le but est d'interroger sur un objet du monde regroupent les questions dont le type sémantique (tel que nous l'avons décrit dans la section 3.3) est une quantité, un lieu ou un temps.

3.5.2 Type de l'interrogatif de la question

En se positionnant sur un type de question donné au niveau de la dimension ontologique définie par la GRINT, deux autres variations sont modélisées. La dimension paradigmatique est l'une d'elle. Elle reflète la possibilité de formuler une même question sous différentes formes en ne faisant varier que l'interrogatif sélectionné. D'un point de vue linguistique, ceci reflète une variation dans l'intention du locuteur questionnant [1]. Par exemple, la question 3.61 met en avant l'intention du locuteur d'obtenir une valeur numérique à la question alors que la question 3.62 traduit l'intention du locuteur d'obtenir une estimation.

Exemple 3.61 *Combien de grammes pèse une boule de pétanque ?*

Exemple 3.62 *Quel est le poids d'une boule de pétanque ?*

Les tableaux 3.1, 3.2 et 3.3 donnent des exemples de variations de questions selon la dimension paradigmatique pour les questions de type ontologique respectif quantité, lieu et temps. On observe qu'à chaque type correspond un positionnement particulier du locuteur par rapport à sa propre requête. Dans ces tableaux les questions sont construites à partir d'un unique interrogatif. Il est cependant envisageable, notamment à l'oral, de combiner plusieurs interrogatifs en une même question. C'est le cas de l'exemple 3.63.

Exemple 3.63 *Quel est le musée où est la Joconde ?*

Nous notons que les questions de type entité nommée (EN) sont des questions déterminatives dont l'interrogatif est un type d'entité et non pas un type indéfini tel que "l'endroit" pour les questions de lieu et "le moment" pour les questions de temps. Par exemple, les questions 3.65 et 3.67 sont des variantes des questions 3.64

1. Nous avons déjà parlé de ce point dans la section 2.4.4 et nous conseillons au lecteur de ce référer au tableau 2.5 page 46 s'il souhaite revenir sur le lien entre interrogatif utilisé dans la question et intention du locuteur. Nous redonnons deux exemples ici pour plus de clarté.

Type paradigmatique	Intention	Exemple de question
adverbiale	neutralité	*Combien pèse un bébé à la naissance ?*
numérale	numérisation	*Combien de kilos pèse un bébé à la naissance ?*
déterminative	estimation	*Quel poids pèse un bébé à la naissance ?*
nominale	estimation	*Que pèse un bébé à la naissance ?*
confirmative	confirmation	*Un bébé pèse-t-il environ 3,2 kilos à la naissance ?*

TABLE 3.1 – *Variations paradigmatiques des questions de quantité*

Type paradigmatique	Intention	Exemple de question
adverbiale	neutralité	*Où est la Joconde ?*
déterminative	estimation	*À quel endroit est la Joconde ?*
confirmative	confirmation	*La Joconde est-elle au Louvre ?*

TABLE 3.2 – *Variations paradigmatiques des questions de lieu*

Type paradigmatique	Intention	Exemple de question
adverbiale	neutralité	*Quand sont les Jeux Olympiques d'été ?*
déterminative	estimation	*À quel moment sont les Jeux Olympiques d'été ?*
confirmative	confirmation	*Les jeux olympiques d'été sont-il en été, tous les 4 ans ?*

TABLE 3.3 – *Variations paradigmatiques des questions de temps*

et 3.66 qui précisent le type attendu de la réponse comme étant respectivement une `EN ville` et une `EN année`.

Exemple 3.64 *À quel endroit est la Joconde ?*

Exemple 3.65 *Dans quelle ville est la Joconde ?*

Exemple 3.66 *À quel moment sont les Jeux Olympiques d'été ?*

Exemple 3.67 *Quels mois sont les Jeux Olympiques d'été ?*

Concernant les questions de quantité, c'est la variante numérale ("Combien de kilos") qui correspond à une question de type entité nommée.

Le type d'entité attendu, quand il est précisé dans la question, influe sur la granularité de la réponse obtenue. La question peut être neutre quand à la granularité attendue de la réponse ("À quel moment", par exemple) ou bien peut imposer un type d'une granularité plus ou moins fine ("À quelle date", "Quel jour",...). La section suivante est consacrée à la granularité.

Les questions déterminatives peuvent ainsi être vues comme un continuum de questions en fonction de la granularité du type indiqué dans la question.

3.5.3 Différentes granularités pour les questions *entité nommée*

La granularité d'une question est un facteur qui entre en jeu uniquement si un type de réponse est précisé dans la question elle-même. La question 3.68 par exemple ne contraint en rien la granularité de la réponse attendue. C'est au locuteur répondant de choisir la granularité de sa propre réponse. On parlera de questions à granularité neutre. Les exemples 3.69 et 3.70 montrent comment une question peut contraindre la granularité de la réponse apportée.

La granularité d'une question (et d'une réponse) est difficile à identifier dans l'absolu. La ville où se trouve la Joconde peut être considérée comme de grosse granularité face à une question portant sur le musée où elle se trouve, mais comme de granularité fine face si on demande le pays dans lequel l'œuvre est exposée.

Exemple 3.68 *Dans quel endroit est la Joconde ?*

Exemple 3.69 *Dans quel musée est la Joconde ?*

Exemple 3.70 *Dans quelle ville est la Joconde ?*

Nous considérons donc que la granularité d'une question ne peut être définie qu'à travers une comparaison avec une autre question. Dans les deux exemples ci-dessus, nous définissons ainsi la granularité de la question 3.69 comme fine en regard de la question 3.70 qui elle a une grosse granularité.

3.5.4 Type syntaxique de la question

La troisième dimension de variation proposée par la GRINT est dite dimension syntagmatique. Elle reflète la possibilité pour un locuteur questionnant de choisir la structure syntaxique de la question qu'il émet. En effet, on peut être plus ou moins direct dans son interrogation en laissant plus ou moins l'opportunité au locuteur qui répond de ne pas savoir par exemple. Les 3 tableaux 3.4, 3.5 et 3.6 représentent les variations des questions de type ontologique respectif quantité, lieu et temps.

Type syntagmatique	**Mise en cause**	**Exemple de question**
prototypique	repérage quantitatif	*Combien pèse un bébé à la naissance ?*
tonique	réalité de la valeur	*Un bébé pèse combien à la naissance ?*
renforcée	quantification	*Combien est-ce que pèse un bébé à la naissance ?*
périphrastique	connaissance de la réponse	*Je voudrais savoir combien un bébé pèse à la naissance ?*
assertive	confirmation	*Un bébé pèse-t-il environ 3,2 kilos à la naissance ?*

TABLE 3.4 – *Variations syntagmatiques des questions de quantité*

Type syntagmatique	**Mise en cause**	**Exemple de question**
prototypique	repérage quantitatif	*Où se trouve la Joconde ?*
tonique	réalité de la valeur	*La Joconde se trouve où ?*
renforcée	quantification	*Où est-ce que se trouve la Joconde ?*
périphrastique	connaissance de la réponse	*Je voudrais savoir où se trouve la Joconde ?*
assertive	confirmation	*La Joconde se trouve au Louvre ?*

TABLE 3.5 – *Variations syntagmatiques des questions de lieu*

Type syntagmatique	**Mise en cause**	**Exemple de question**
prototypique	repérage quantitatif	*Quand ont lieu les Jeux Olympiques d'été ?*
tonique	réalité de la valeur	*Les Jeux Olympiques d'été ont lieu quand ?*
renforcée	quantification	*Quand est-ce qu'ont lieu les Jeux Olympiques d'été ?*
périphrastique	connaissance de la réponse	*Je voudrais savoir quand ont lieu les Jeux Olympiques d'été ?*
assertive	confirmation	*Les Jeux Olympiques d'été ont lieu en été tous les 4 ans ?*

TABLE 3.6 – *Variations syntagmatiques des questions de temps*

On observe qu'à chaque type correspond une intentionnalité particulière du questionnant par rapport à son interlocuteur. Dans ces tableaux sont présentées des questions dites prototypiques (forme neutre d'un point de vue intentionnel et construction typique d'un point de vue grammatical) dont on modifie la structure syntaxique pour obtenir les questions dites toniques, renforcées, périphrastiques et assertives. L'ensemble de ces variations correspondent à des phrases simples mais une combinaison des structures syntaxiques est possible (exemples 3.71 et 3.72) bien qu'elle donne souvent lieu à des structures lourdes.

Exemple 3.71 *Je voudrais savoir où est-ce qu'est la Joconde ?*

Exemple 3.72 *Je voudrais savoir dans quel musée est-ce qu'est la Joconde ?*

3.6 NATURE DE LA RÉPONSE

Nous avons déjà vu que la réponse peut être élémentaire ou non. Si elle l'est, il peut s'agir notamment d'une réponse fermée ou d'une entité nommée. Quel que soit le type de réponse attendue à une question, ce que nous appellerons la nature de la réponse peut quant à elle varier. Nous nous centrons ici sur la réponse du point de vue de l'"information" dont il s'agit.

Dans le cas d'une question portant sur le lieu d'exposition de la Joconde, il existe une unique réponse, même si, en fonction de la granularité choisie elle peut revêtir diverses formes (exemple 3.73).

Exemple 3.73 *Q : Où est la Joconde ?*

Exemple 3.74 *R : Au Louvre, à Paris, en France,...*

D'autre part, une réponse peut varier en fonction du temps (exemple 3.76) ou bien en fonction d'un critère qui n'aurait pas été mentionné dans la question (exemple 3.78). Ces variations sont dues à une ambiguïté ou à une imprécision dans la question. Un facteur "extérieur" à l'objet de la question n'a pas été fixé.

Exemple 3.75 *Q : Où a lieu TALN ?*

Exemple 3.76 *R : ça dépend des années. Cette année, c'était à Montréal !*

Exemple 3.77 *Q : Combien d'articles ont été acceptés à TALN en 2010 ?*

Exemple 3.78 *R : ça dépend, tu veux parler des articles longs ou des articles courts ?*

Une réponse peut aussi varier selon la nature même de l'objet considéré. Par exemple, si on se questionne sur le poids d'un bébé à sa naissance, la réponse dépend du nouveau né considéré (exemple 3.80). En fait, il n'y a pas une réponse mais plusieurs réponses. (à moins de considérer le poids moyen des bébés à la naissance mais alors la question n'est plus la même). Un facteur "propre" à l'objet de la question existe car le référent qu'il désigne est ambigu ou encore multiple.

Exemple 3.79 *Q : Combien pèse un bébé ?*

Exemple 3.80 *R : ça dépend du bébé. Mon frère pesait (..)*

Prenons à présent l'exemple du Rhin (exemple 3.81). Dans ce cas, il est fort peu probable que son lieu change tant du point de vue du Rhin lui même, de sa trajectoire, que du point de vue des noms des zones, régions, pays qu'il traverse. Mais justement, il est délicat de ne citer qu'un seul lieu quand il s'agit de ce fleuve car il traverse plusieurs zones géographiques. Répondre "en Europe" est une réponse juste et unique. Cependant on peut penser qu'elle ne sera pas satisfaisante, au moins pour un européen, car trop peu précise. Que répondre alors ? Il est possible de délimiter, décrire par extension la zone que ce cours d'eau traverse, en citant

l'ensemble des zones concernées (exemple 3.82) ou encore en circonscrivant la zone (exemple 3.83).

Exemple 3.81 *Où est le Rhin ?*

Exemple 3.82 *Le Rhin traverse la Suisse, le Liechtenstien, l'Autriche, l'Allemagne, la France, les Pays-Bas.*

Exemple 3.83 *Le Rhin est en Europe de l'ouest, entre la France et l'Allemagne et prend sa source dans les Alpes pour se jeter dans la mer du Nord.*

Il existe donc des cas de réponse *unique* et des cas de réponse *multiple* dus par exemple à une ambiguïté ou imprécision de la question, de l'objet de la question ou encore du référent du monde auquel il se rapporte.

Nous rappelons ici que nous nous intéressons à la variation de la question dans le but de construire un corpus de questions. Celles-ci vont être posées à des humains et nous allons collecter ainsi un corpus de réponses. Dans cette optique, nous cherchons à obtenir des réponses en langue naturelle pas ou peu complexe. Si les réponses ont des formes très différentes, il sera difficile de les étudier et de déduire des traits communs aux réponses collectées. C'est d'ailleurs pour cette raison, entre autres, que nous n'avons pas envisagé d'utiliser des réponses issues de sites collaboratifs de question-réponse. Les questions de notre corpus sont donc des questions dont la réponse est de nature unique. Dans la même idée, il est important que les questions ne soient pas trop compliquées (au risque d'obtenir beaucoup de réponses du type "je ne sais pas"). La section suivant aborde le point du thème de la question en introduisant la notion de valence d'une réponse.

3.7 Valence d'une réponse et type thématique de la question

Nous avons bon nombre de caractéristiques à prendre en compte pour la création du panel de question. Ces caractéristiques sont linguistiques et concernent tant la question que la réponse.

Une dernière caractéristique se doit d'être ajoutée pour les besoins de l'expérience, plus précisément il s'agit d'une contrainte de résultat : nous voulons poser des questions *faciles* pour maximiser le nombre de réponses à *valence positive*.

3.7.1 Valence d'une réponse

Pour toute question, on peut classer les réponses en deux catégories. Les réponses qui répondent en effet à la question, nous entendons par là, les réponses qui proposent une "information" en réponse (ou information-réponse) à la question comme dans l'exemple 3.85). Nous parlerons de *réponse à valence positive*.

D'autres réponses ne répondent pas au sens qu'elles constituent par exemple un aveu de non connaissance de la réponse (3.86), un renvoi vers un autre moyen d'obtenir la réponse (3.87) ou encore une demande de précision de la question (3.88). Nous parlerons de *réponses à valence négative*.

Exemple 3.84 *Q : Combien pèse un bébé à sa naissance ?*

Exemple 3.85 *R : Environ 3 kg 500.*

Exemple 3.86 *R : Je ne sais pas.*

Exemple 3.87 *R : Heu... Regarde sur un site médical.*

Exemple 3.88 *R : Tu veux dire le poids moyen ?*

Si, à première vue, il s'agit d'une vision de la réponse dans le dialogue, dans l'interaction uniquement, cette différence existe aussi dans la cadre de réponse à une question unique. Il s'agit des cas où le système de question-réponse ne trouve pas de réponse.

3.7.2 Type thématique de la question

En conséquence, nous privilégierons des questions "faciles" au sens que nous espérons que la plupart de nos participants en connaîtrons la réponse. L'intérêt est double. D'une part, les réponses à valence positive sont beaucoup plus intéressantes compte-tenu du but de l'étude. D'autre part nous supposons que les réponses à valence négative sont moins variées que les réponses à valence positive. Une plus faible quantité nous semble donc nécessaire.

Les questions seront ainsi des questions de culture générale. Le vivier de participants potentiels auxquels nous pouvons faire appel sont des contacts personnels (majoritairement parisiens), des étudiants (tant de l'université qu'au sein du laboratoire), des chercheurs mais aussi des personnes "intéressées par la recherche en général" [2].

CONCLUSION DU CHAPITRE

Nous avons identifié nombre de paramètres de variation tant concernant la question en elle-même que la réponse attendue à cette question. Les paramètres de variation de la question sont son type sémantique, le type de son verbe principal, le type de l'interrogatif utilisé ou type paradigmatique, son type syntaxique ou type syntagmatique, son type de granularité et son thème correspondant à l'objet sur lequel la question porte. Le tableau 3.7 résume ces paramètres.

2. Ceci sera détaillé dans la section 4.5.1.

Paramètre	**Exemples de valeurs possibles**
Type sémantique	*quantité, lieu, temps, personne,...*
Type du verbe principal	*neutre, spécifique*
Type paradigmatique	*adverbiale, déterminative, confirmative,...*
Type syntagmatique	*prototypique, tonique, renforcée,...*
Type de granularité	*fin ... moyen ... gros, neutre*
Thème	*pétanque, Tour Eiffel, Rhin,...*

TABLE 3.7 – *Résumé des paramètres de variation de la question en elle-même*

Les paramètres que nous avons identifiés peuvent aussi correspondre à une variation de la réponse attendue à la question et sont le type attendu de la réponse, la nature de la réponse et sa valence. Le tableau 3.8 résume ces paramètres.

Paramètre	**Exemples de valeurs possibles**	
Type attendu	*réponse élémentaire*	entité nommée, valeur numérique,...
	réponse non élémentaire	liste de réponse élémentaire, définition,...
Nature	unique, multiple	
Valence	positive, négative	

TABLE 3.8 – *Résumé des paramètres de variation de la réponse attendue*

Ces paramètres ne sont pas indépendants les uns des autres. Par exemple, le type sémantique de la question contraint le type du verbe principal, le type paradigmatique, le type syntaxique mais aussi le thème de la question.

À présent que nous avons identifié cet ensemble de paramètres, nous sommes en mesure de positionner une question donnée en fonction de ceux-ci. Prenons la question 3.89 comme exemple.

Exemple 3.89 *Où est la Joconde ?*

Elle prend les valeurs suivantes pour nos différents paramètres :

- `Type sémantique :` *`lieu`*
- `Type du verbe principal :` *`neutre`*
- `Type syntagmatique :` *`prototypique`*
- `Type paradigmatique :` *`adverbiale`*
- `Type de granularité :` *`neutre`*
- `Thème :` *`La Joconde`*
- `Type attendu de la réponse :` *`lieu`*
- `Nature de la réponse :` *`unique`*
- `Valence :` *`inconnue`*

La valence de la réponse est inconnue puisqu'on considère la question seule.

Puisque nous pouvons décrire ainsi les questions, nous pouvons construire un corpus de questions pour lequel nous pouvons contrôler les formes des questions tant au niveau lexical que syntaxique et sémantique. Nous avons constitué un tel corpus et mené une expérience permettant de collecter des réponses d'humains. Le chapitre suivant expose cette collecte.

Deuxième partie

Réponses humaines ?

4 COLLECTE D'UN CORPUS DE RÉPONSES HUMAINES

SOMMAIRE

Ce chapitre détaille le protocole utilisé pour collecter un corpus de réponses humaines, orales et écrites. Il s'agit d'un protocole par lequel des participants se sont vu poser une série d'une vingtaine de questions soit à l'oral, soit à l'écrit. Le chapitre s'organise comme suit. L'objectif général de la collecte de corpus est rappelé. Une description technique de la mise en œuvre de l'expérience est donnée, à savoir la mise en place de deux plates-formes expérimentales pour l'oral et pour l'écrit, le choix et la rédaction d'un contexte expérimental, la présentation d'un scénario-type. Les questions soumises aux participants sont exposées. Des informations sur le déroulement des passations et notamment le recrutement des participants sont exposées. Enfin une évaluation du protocole est faite. Ce chapitre correspond au cœur du travail de la thèse et présente la création d'un protocole expérimental et d'un corpus de question construites, l'organisation d'une collecte de corpus auprès de plus de 150 participants et l'évaluation du protocole mis en œuvre.

This chapter presents the experimental protocol used to collect a corpus of oral and written human answers to questions. In this protocol, participants have to answers about twenty questions. The chapiter is organized as follow. The general goal of the collect is presented. A technical description of the protocol is given including the presentation of two experimental platforms (for oral and written collect), the experimental context and an example of the procedure. The question corpus used is presented. Information about the collect, the obtained corpus and the

participant recruitement is given. Finally, an evaluation of the exprimental protocol is presented.
This chapter presents a first main part of the work done during the Ph.D : the creation of an experimental protocol and a corpus of questions, the collect of a corpus of answers involving more than 150 participants and the experimental protocol evaluation.

4.1 INTRODUCTION

Nous avons, dans les chapitres précédents, exposé notre point de vue sur la réponse à une question qui est celui de voir cet échange comme une interaction. Nous avons mis en évidence le fait que les réponses que fournissent actuellement les systèmes de question-réponse ne sont pas des réponses en interaction. De plus, nous avons cherché à définir ce que pourraient être des telles réponses, cependant aucun corpus n'est disponible pour les observer. Nous avons ainsi choisi de constituer un corpus de réponses à des questions. Parce que question et réponse sont liées, nous ne pouvons envisager de poser des questions prises "au hasard" pour collecter un corpus de réponse. Le choix des questions utilisées est en effet primordial. Inspirée par les travaux exposés par la GRINT, nous sommes convaincue du lien entre la forme de la question et celle de la réponse. Nous avons ainsi recherché dans les travaux existants des paramètres permettant de décrire de façon précise la forme d'une question. Nous sommes ainsi en mesure de construire un corpus de questions dont les caractéristiques sont contrôlées. Parce qu'il s'agit d'obtenir des réponses qu'un système de question-réponse pourrait reproduire, la façon dont la collecte des réponses est faite est elle aussi primordiale. Ce chapitre décrit en détail les objectifs que nous nous sommes fixés en collectant des réponses d'humains (section 4.2), la mise en place du protocole expérimental (section 4.3), la construction du corpus de questions soumises aux participants (section 4.4), le déroulement des passations (section 4.5) et une évaluation du protocole mis en œuvre (section 4.6). Après trois chapitre introductifs et portant sur des travaux du domaine, ce chapitre constitue l'exposé d'un premier travail que nous avons effectué au cours de la thèse.

4.2 OBJECTIF ET PRINCIPE GÉNÉRAL

Comme nous l'avons vu précédemment (section 2.6.4 du chapitre 2, la réutilisation d'un corpus existant ou bien l'extraction d'un corpus à partir de question-réponse existantes n'est pas adapté aux objectifs que nous nous fixons. En effet, les corpus existants proposent des réponses très succinctes, composées uniquement de quelques mots correspondant à l'information qui répond à la question ou au contraire des réponses collaboratives qui contiennent des informations dont un système de question-réponse ne dispose pas. Certains corpus, notamment ceux que l'on pourrait extraire de site Internet de question-réponse, sont trop disparates : les réponses variant énormément par leur forme et leur contenu, une étude systématique serait problématique. Enfin, les questions proposées ne forment jamais un tout cohérent : si une annotation pourrait permettre de connaître les caractéristiques des questions, il est fort peu probable qu'un corpus contienne un échantillon représentatif de différentes variations linguistiques de question possibles en français.

Nous avons donc choisi de constituer notre propre corpus. Nous avons mis en place une expérience pour collecter des réponses à des questions. Les objectifs principaux sont multiples. Nous voulons obtenir un corpus de question-réponse dont les réponses ne se réduisent pas à l'information qui répond à la question contrairement aux réponses de certains systèmes. Nous définissons ces réponses comme les réponses "succinctes". De plus, nous voulons des réponses qui soient comparables les unes aux autres contrairement aux réponses existant sur des sites de question-réponse collaboratifs. Enfin, nous souhaitons obtenir des réponses qui ne contiennent pas d'informations supplémentaires contrairement aux réponses des systèmes coopératifs. L'idée est que les réponses doivent être reproductibles par les systèmes de question-réponse quels qu'ils soient. Elles ne doivent donc pas contenir d'informations autres que celles dont disposent les systèmes de question-réponse c'est-à-dire la question et la ou les réponses à retourner. Les réponses contenant des éléments autres que ceux-ci sont définies comme "complémentaires". Le choix des questions soumises, du protocole et des participants découle ainsi de plusieurs objectifs. Pour chaque objectif, nous avons déterminé des contraintes à mettre en œuvre au travers du protocole expérimental et une méthode pour évaluer si l'objectif est atteint.

Un premier objectif est d'obtenir des réponses spontanées (ou au moins, non préparées) afin de s'assurer d'une certaine naturalité de ces réponses. Le protocole expérimental mis en place (et détaillé section 4.3) nous assure ainsi un certain dynamisme de l'expérience : une "histoire" motivante, des séries de questions pas trop longues, des questions simples,... Nous évaluons si cet objectif est atteint en observant le temps de réponse et la présence d'interventions humoristiques, complexes ou encore contenant des sous-entendus que nous interprétons comme des signes de spontanéité dans les réponses.

Un deuxième objectif est celui de minimiser le nombre de réponses à valence négative [1] c'est-à-dire de réponses qui ne contiennent pas d'information-réponse (comme par exemple "je ne sais pas"). Les questions posées sont ainsi supposées "faciles" (à répondre) compte-tenu des participants auxquels nous avons fait appel. La section 4.5.1 détaille le recrutement des participants et la section 4.14 détaille le choix des questions. Cet aspect est évalué par l'observation du nombre de réponses à valence négative et de non-réponse/absence de réponse,

Notre troisième objectif est d'obtenir des réponses non succinctes [2] c'est-à-dire des réponses qui ne sont pas constituées uniquement d'une information-réponse. Le contexte expérimental, qui a été rédigé pour mettre les participants en situation, suggère que les participants doivent répondre à des enfants. Ceci devrait les inciter à donner des réponses sous forme de "phrases" plutôt que des réponses très courtes. Le contexte expérimental est présenté dans la section 4.3.1. La taille des réponses

1. Voir la définition d'une réponse à valence négative page 71.
2. Voir la définition d'une réponse succincte page 49.

et le nombre de réponses minimales sont des observations permettant d'évaluer que cet objectif est atteint.

Un quatrième objectif est celui d'obtenir des réponses qui ne contiennent pas d'information complémentaire. Le contexte expérimental proposé et la formulation des questions tend à inciter les participants à répondre à la question sans développer leur réponse. En effet, d'une part le contexte indique de répondre aux questions mais ne suggère nullement de développer les réponses, d'autre part les questions sont factuelles et très précises. La section 4.4 détaille le choix des questions et la façon dont elles ont été élaborées pour limiter leur ambiguïté. L'évaluation porte sur l'observation du nombre de réponses "complétives".

Un cinquième objectif est celui d'obtenir des réponses tant à l'oral qu'à l'écrit afin de pouvoir comparer les deux modalités et éventuellement de répondre à la question de savoir s'il est envisageable d'utiliser pour ces deux modalités un unique module de génération en langue naturelle. L'expérience a été mise en place pour l'oral et pour l'écrit en minimisant les variations entre les deux protocoles expérimentaux. La section 4.3.2 traite de ce point. Nous évaluons ce point en observant le nombre de réponses collectées à l'oral et à l'écrit.

Un autre de nos objectifs est celui de poser différentes formulations d'une même question pour éviter le biais dû à la formulation de la question. Nous avons ainsi créé un ensemble des variations à partir de questions de base. Les différentes variantes d'une même question ont été créées selon des facteurs de variation précis. Ces facteurs sont détaillés dans la section 4.4.2. L'ensemble des variations autour d'une même question est appelé "groupe de questions". Le nombre de réponses pour chaque groupe de questions permet d'évaluer cet aspect.

Enfin, un septième objectif est celui d'éviter l'ennui des participants ou encore des réponses trop "systématiques" (qui correspondraient à notre sens en une perte de la naturalité). Ainsi, pour un participant donné, les questions proposées sont diverses tant dans leur formulation que dans leur thème. La grille de passation qui assure cette diversité est présentée section 4.4.3. La lecture des remarques libres laissées par les participants en fin d'expérience permet d'évaluer cet aspect.

4.3 PROTOCOLE EXPÉRIMENTAL

Le protocole mis en place est original. Aucune collecte de corpus telle que nous l'entendons n'ayant été menée jusqu'à présent. Dans notre protocole, plutôt qu'un système réponde à des questions posées par des utilisateurs (ou données dans un fichier d'entrée dans le cas des campagnes d'évaluation des systèmes de question-réponse), nous avons demandé à des locuteurs natifs du français de répondre à un ensemble de questions posées par notre 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 participants aussi large que le nôtre (40 chez [Anaya & Kosseim, 2003], 152 chez nous).

Au cours de cette section, nous présentons le contexte expérimental, la mise en place sur deux modalités d'interaction, l'équipement technique utilisé et un scénario type de passation pour chacune des deux modalités d'interaction. Le recrutement des participants est quant à lui présenté dans la section 4.5.1 consacrée au déroulement des passations.

4.3.1 Contexte expérimental

Notre but n'est pas d'obtenir des bonnes réponses c'est-à-dire des réponses qui soient justes, exactes. En revanche, nous cherchons à observer la formulation de l'*information-réponse*. Nous souhaitons inciter les participants 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 dont on leur disait qu'elle avaient été posées par des enfants de CM2 préparant des exposés.

Le contexte proposé était présenté par le texte suivant :

> ❝ *La classe de CM2 de l'école primaire de Château-la-Vallière se propose de créer des affiches pour la fête de l'école. Ces posters portent sur des participants variés et pour les réaliser, les élèves ont besoin de quelques informations. Vous êtes là pour répondre à leurs questions ! Ne soyez pas surpris s'il vous semble avoir déjà répondu à une question. Certaines se ressemblent en effet mais ont été posées par différents enfants.* ❞

FIGURE 4.1 – *Consigne donnée aux participants et présentant le contexte expérimental*

Nous espérons en présentant l'expérience ainsi motiver les participants à répondre mais aussi les préparer à avoir à répondre à plusieurs questions similaires et à des questions faciles.

4.3.2 Protocole multimodal

L'expérimentation a eu lieu à l'oral et à l'écrit. Malgré les différences inhérentes aux deux modalités, nous avons voulu proposer un protocole variant le moins possible selon la modalité. L'expérience à l'écrit a été mise en place sur une page Web. À l'oral, l'expérience se déroule par téléphone. Les réponses obtenues ont par la suite été transcrites. Pour les deux modalités, le nombre de questions posées à chaque participant, les questions et leur ordre de présentation, ainsi que le contexte expérimental sont les mêmes. Seules changent les consignes données aux utilisateurs avant de débuter l'expérience puisque l'utilisation des deux plates-formes d'expérimentation n'est pas la même. Le schéma 4.2 représente l'organisation de ce protocole.

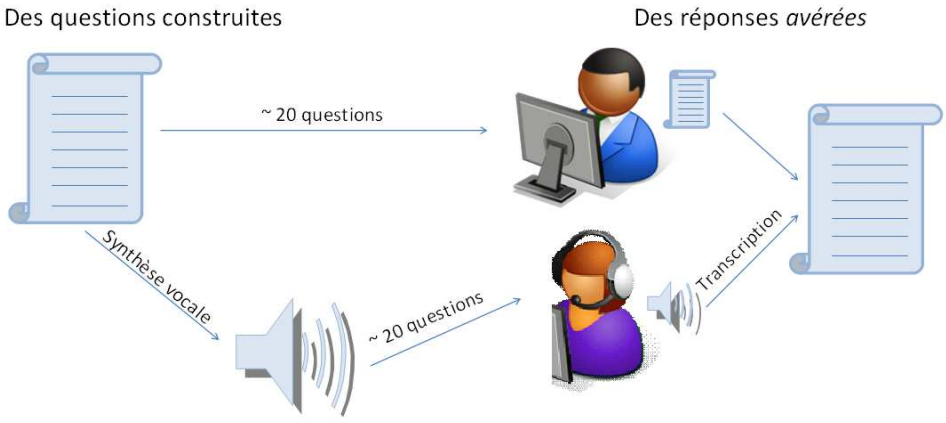

FIGURE 4.2 – *Schéma général du protocole expérimental multimodal*

En plus de ce souci d'équivalence des deux versions de l'expérience, nous avons souhaité pouvoir comparer des réponses à l'oral et à l'écrit obtenu dans des conditions similaires. Une partie des questions a été ainsi posée aux mêmes participants à la fois à l'écrit et à l'oral.

4.3.3 Équipement technique

À l'écrit

À l'écrit, nous avons mis en place une page Web dynamique (php) permettant grâce à l'utilisation de formulaires de récupérer et vérifier l'identifiant du participant et de récupérer ses réponses. À chaque fois que le participant répond à une question, il clique sur le bouton "Question suivante". Les informations envoyées par le formulaire sont alors à la fois la réponse à la question et la date de validation du bouton (à la seconde près). Ceci nous permet de calculer le temps de réponse du participant.

L'information du numéro de la question courante et du nombre de question total est présente à chaque question sous la forme : `Question n° # / ##` .

Les figures 4.3 et 4.4 sont des captures d'écran respectives de la page d'accueil et d'une page de question-réponse.

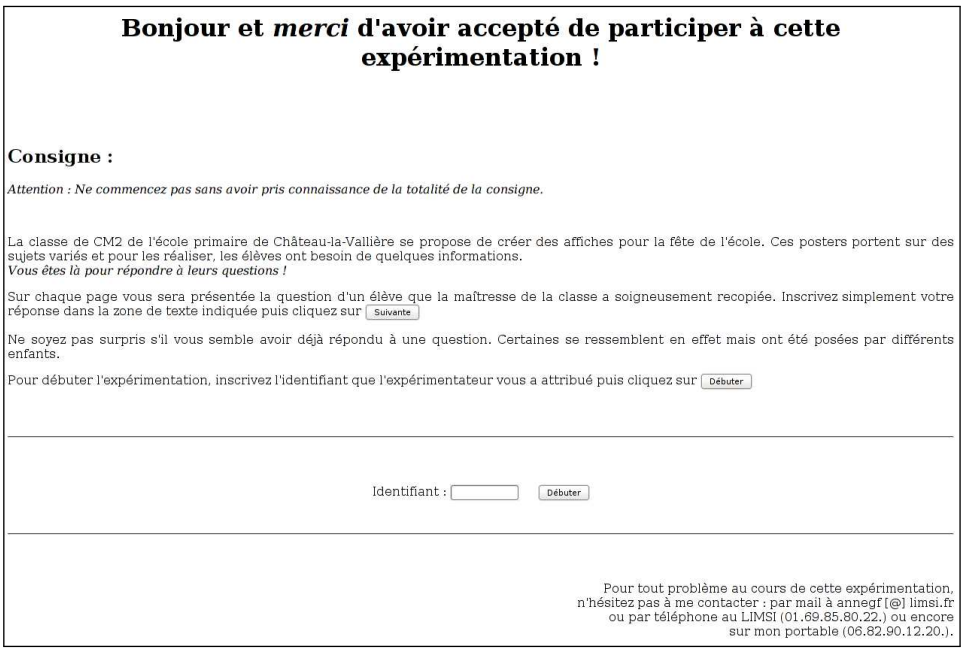

FIGURE 4.3 – *Capture d'écran de la page d'accueil de la version écrite de l'expérience*

Nous avons construit l'interface de façon a ce qu'elle soit réutilisable pour d'autres expériences similaires. Il suffit de placer sur un serveur Web un ensemble de fichiers et de dossier contenant les données de l'expérience :

- Un fichier contenant pour chaque question son identifiant et la question en elle-même
- Un fichier contenant pour chaque identifiant d'accès à la plate-forme la liste ordonnée des identifiants de question à poser
- Un fichier de paramètre (.php) peut être modifié en fonction du serveur et des choix de l'expérimentateur pour "personnaliser" l'interface (nom et adresse mail de l'expérimentateur, noms des fichiers de sortie,...)
- Des pages .htm contenants la consigne à afficher en début d'expérience, les

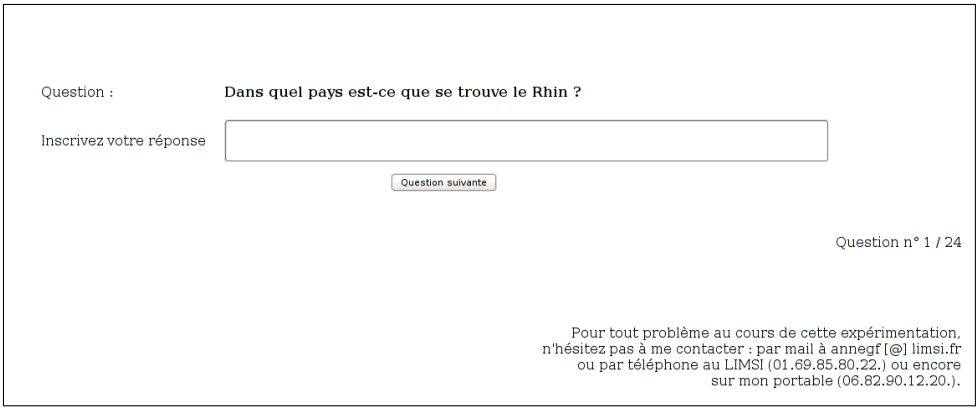

FIGURE 4.4 – *Capture d'écran d'une page question-réponse de la version écrite de l'expérience*

messages d'erreur à afficher le cas échéant, un message s'affichant en bas de page (contenant dans notre cas les informations pour contacter l'expérimentateur) et le message de remerciement final.

Le fichier index.php permet l'affichage html des pages. Les fichiers de sortie sont les suivants :

- Un fichier par passation contenant pour chaque réponse du participant son identifiant et l'identifiant de la question posée.
- Un fichier par passation contenant une version "lisible" de celle-ci. L'identifiant du participant est indiqué, pour chaque question posée, l'identifiant et la question sont rappelés, la réponse et le temps de réponse sont indiqués.
- Un fichier listant l'ensemble des connexions à la plate-forme. Une connexion correspond à la date et l'heure, l'ip et l'identifiant de connexion utilisé.
- Un fichier listant l'ensemble des identifiants de connexion pour lesquels une passation a été effectuée. Une participation est considérée comme "effectuée" lorsque le participant a atteint la dernière page de l'expérience contenant un message de remerciement et la possibilité de laisser une remarque.
- Un fichier contenant l'ensemble des remarques des participants. À chacune est associé l'identifiant du participant ayant laissé la remarque.

À l'oral

Nous avons utilisé une plate-forme déjà existante au sein du laboratoire. Il s'agit du même dispositif que celui décrit dans [Toney et al., 2008]. Des lignes téléphoniques (y compris des numéros *verts* permettant aux utilisateurs d'appeler gratuitement) ont été activées et reliées à un système gérant les interactions par téléphone. Ce système peut être relié à une ou plusieurs lignes téléphoniques. En entrée, le système permet la reconnaissance de la numérotation saisie sur les touches des combinés téléphoniques afin de pouvoir utiliser des identifiants d'accès, la détection des interventions du locuteur humain et leur enregistrement. En sortie, il prévoit le lancement de fichiers audio et produit un fichier audio qui contient l'enregistrement

de l'ensemble de l'interaction ainsi qu'un fichier textuel qui est la transcription de cette interaction. La transcription des interventions du participant est obtenue grâce à un système de reconnaissance vocale [Gauvain et al., 2005]. Celle des interventions du système doit être donnée au système. Au niveau de la gestion de l'interaction, il prévoit la formulation d'un message d'accueil, de messages intermédiaires entre chaque question et d'un message final.
Il suffit de donner au système :

- l'ensemble des fichiers audios qui devront être utilisés par le système,
- un fichier contenant pour chaque fichier audio, le texte auquel il correspond,
- un fichier contenant les identifiants utilisateur valides,
- un fichier de correspondance entre identifiant utilisateur et liste des questions à poser pour cet identifiant.

Chaque fichier contient une intervention du système et le nom du fichier sert d'identifiant à cette intervention. Nous avons différents types d'interventions :

- le message d'accueil,
- les messages de retour après saisie d'un identifiant : ils annoncent si l'identifiant est correct ou incorrect et propose jusqu'au troisième identifiant erroné, un nouvel essai,
- le message de présentation du contexte expérimental tel que décrit dans la section 4.3.1,
- les transitions qui sont prononcées après chaque réponse du participant. Elles le remercient et l'informent de la progression de l'expérience en annonçant le numéro de la question courante ou encore le nombre de questions restantes,
- les questions,
- le message de clôture.

Chaque intervention est identifiée par le nom du fichier audio auquel elle correspond.
Tous ces fichiers correspondent à la vocalisation d'une intervention du système. Le lien entre une intervention (version écrite du message) et le nom du fichier audio (version orale du message) doit être donnée au système au sein d'un fichier de correspondance. Ainsi il peut produire le fichier de transcription de l'intégralité de l'interaction. Dans ce fichier, pour chaque intervention, on sait par quel canal (entrant ou sortant) elle a été émise (c'est-à-dire si l'intervention a été prononcé par le participant ou par le système), à quel moment précis elle a débuté et terminée.

Le système utilise le détecteur de la parole (ou *Speech Activity Detection component*, décrit par [van Schooten et al., 2007]). Cet outil permet de détecter, si un son est détecté, s'il s'agit de parole ou non (bruit, musique,...). Cela nous permet de détecter quand le participant commence à parler mais surtout quand il s'arrête de parler et donc quand le système peut reprendre son tour de parole et passer à la question suivante. Étant donné qu'un participant peut faire une pause au cours de sa réponse, nous avons dû paramétrer le temps d'attente avant de considérer que

l'utilisateur avait fini de parler. Il s'agit de ne pas lui couper la parole (donc attendre suffisamment longtemps) tout en ne laissant pas un silence trop long (auquel cas le participant pourrait s'impatienter ou bien penser que le système est bloqué). Après quelques essais auprès de collègues du LIMSI n'ayant pas participé à l'expérience par la suite, nous avons fixé ce seuil à 3 secondes. Choisir un seuil est délicat. Il n'est pas dépendant de l'intonation ou d'autres indices paralinguistiques mais figé, fixé par avance. Le temps d'attente avant la réaction du système peut être ressenti comme trop long ou trop court selon le participant mais aussi par un même participant au cours de l'expérience. Il n'est donc jamais vraiment adapté. Nous avions, dans notre cas, quelques interventions coupées et un bon nombre de participants qui exprimaient leur impatience ou doute quand au fonctionnement du système (le temps d'attente étant trop long).
Nous avons choisi de laisser le seuil à 3 secondes pour minimiser le nombre de coupure mais d'ajouter une indication dans les instructions données en début de passations expliquant qu'un temps d'attente entre la fin de la réponse et la question suivante est normal. Ceci a été satisfaisant puisque les participants n'ont dès lors plus exprimé d'impatience.

Les fichiers audio correspondent à la synthèse vocale automatique des différentes interventions. Nous avons choisi d'utiliser une voix de synthèse afin d'éviter un biais dû à l'intonation dans le cas où les questions auraient été formulées par un humain. Le même modèle vocal et la même voix ont été utilisés pour l'ensemble des interventions. Le synthétiseur vocal utilisé est la version 2006 de Sayso, commercialisé par la société Acapela[Acapella, site web]. La voix sélectionnée est "claire08s".

Guide de transcription Les réponses des utilisateurs ont été transcrites automatiquement grâce à un module de reconnaissance de la parole (ou *Text-to-Speech component*, [Galibert et al., 2005a]). Les transcriptions ont été corrigées manuellement. La transcription finale correspond aux règles de transcription suivantes [3] :

Règle 1 : Sont transcrits l'ensemble des mots prononcés, sans ajout de la ponctuation et sans modification. Ainsi, même si l'intervention n'est pas grammaticale, par exemple si une négation n'est pas prononcée, elle n'est pas ajoutée dans la transcription,

Règle 2 : Les mots transcrits doivent être correctement orthographiés,

Règle 3 : On ne rend pas compte des réductions ou élisions présentes dans la parole (par exemple, la prononciation "*s'iou plaît*", doit être transcrite "`s'il-vous-plaît`")

3. Les règles de transcription données sont issues du guide d'annotation produit au cours du projet RITEL [RITEL, site web].

Règle 4 : Les mots répétés, dans le cas d'hésitations notamment, sont transcrits autant de fois qu'ils sont répétés.

Règle 5 : Les mots tronqués sont écrits partiellement sauf si le mot complet peut être déduit du contexte. On utilise les parenthèses pour signaler le phénomène. Par exemple, on note "`vou()`" si la partie finale du mot est tronquée et "`vou(drais)`" si cette partie finale peut être déduite du contexte.

Règle 6 : Les erreurs de prononciations sont indiquées par la notation étoile (*). Par exemple, on note "`*spectacle`" pour un mot prononcé "*pestacle*".

Règle 7 : Les hésitations "*euh*" sont transcrites par "`{fw}`".

Règle 8 : Pour la parole incompréhensible et les bruits (sons qui ne correspondent pas à de la parole), l'ensemble des notations suivantes doit être utilisée :

$[pi]$	parole incompréhensible
$[pif]$	parole inaudible et/ou faible
$[r]$	respiration (inspiration, expiration, reniflement, souffle,...)
$[bb]$	bruit de bouche (toux, raclement de gorge, éternuement, bruit de gorge, rire, sifflement, chuchotement,...)
$[conv]$	conversation et bruits de l'environnement (conversation de fond, froissement de papier, bruits du micro,...)

4.3.4 Scénario type

Le scénario d'une participation suit les étapes suivantes :

1	Une personne répond à l'appel à participation	
2	Nous lui communiquons un identifiant et les informations pour accéder à l'expérience.	
	puis	
	À l'oral	**À l'écrit**
3	Le participant accède à l'expérience par téléphone.	Le participant accède à l'expérience sur le site Internet.
4	Le système demande l'identifiant d'accès.	Le participant lit la consigne. Elle se compose du texte présentant le contexte expérimental et d'indications d'utilisation.
5	Le participant saisit sur le clavier du téléphone l'identifiant qui lui a été donné.	Sur la page un formulaire demande au participant son identifiant.
6	Le système donne la consigne. Elle se compose du texte présentant le contexte expérimental et d'indications d'utilisation.	Le participant indique son identifiant et clique sur le bouton "suivant".
7	L'échange de question et réponse débute : – La 1ère question est présentée au participant. – Le participant répond. – Lorsque le détecteur de la parole détecte que l'utilisateur a fini de parler la question suivante est posée. – Et ainsi de suite jusqu'à la dernière question.	L'échange de question et réponse débute : – La 1ère question est présentée au participant. – Le participant répond. – Lorsqu'il clique sur "Valider", la question suivante s'affiche. – Et ainsi de suite jusqu'à la dernière question.
8	Le système donne le message final de remerciement et indique l'adresse électronique à utiliser pour transmettre d'éventuelles remarques. Éventuellement, le participant envoie ses remarques par mail.	Le message final s'affiche sur la page. Il remercie le participant et propose de laisser des remarques libres dans une zone de texte. Éventuellement, le participant laisse une remarque et la valide.

4.4 LES QUESTIONS SOUMISES AUX PARTICIPANTS

L'un des objectifs de cette collecte de réponse est d'obtenir des réponses à valence positive quelle que soit d'ailleurs leur valeur vériconditionnelle. En d'autre terme, on cherche à éviter les réponses qui ne contiennent pas d'information-réponse, du type "je ne sais pas", mais on ne cherche pas nécessairement à obtenir des réponses contenant une information-réponse correcte, des *bonnes réponses*. Nous avons ainsi choisi de poser des questions *simples*, auxquelles une majorité des participants pourront donner une réponse.

Un autre objectif est d'obtenir des réponses pour différentes formulations de la même question. Nous fondant sur l'état de l'art concernant la variation des questions, nous avons déterminé un ensemble de facteurs de variation qui ont servi à créer différentes variantes de questions à partir d'une question de base.

Nous avons ainsi construit le corpus des questions soumises aux participants en deux étapes : (1) nous avons créé un ensemble de *questions de base*, puis (2) à partir de celles-ci nous avons créé l'ensemble des variantes de questions selon nos facteurs de variation. Nous parlerons désormais de *groupe de questions* pour désigner l'ensemble des variantes d'une question (dont la question de base).

Les questions de base n'ont pas été choisies au hasard. Elles correspondent à un ensemble de contraintes que nous nous sommes fixées. Nous exposons ces contraintes (section 4.4.1 avant de donner la liste des questions de base utilisées. Ensuite nous expliciterons la création des variantes de question en exposant les facteurs de variations utilisés pour créer les groupes de questions (section 4.4.2). Nous présentons une vue d'ensemble du corpus ainsi créé (section 4.4.4). La liste complète des questions soumises au cours de la collecte constitue l'annexe A.1.

4.4.1 Questions de base

Le choix des questions de base dérive d'un certain nombre d'objectifs que nous avons pour la collecte de corpus. Nous les avons exposés en début de chapitre, page 81. Ils sont repris ici afin d'expliciter nos choix.

Questions dont les réponses sont élémentaires Nous avons mis en valeur dans le chapitre 3 l'existence de deux classes principales de réponses à une question. Les réponses élémentaires et les réponses qui ne le sont pas. Les questions factuelles attendent plus particulièrement des réponses élémentaires. Les réponses non élémentaires sont dues à un choix d'interaction (être coopératif) ou correspondent à des questions liste, procédurales, définition,...

Dans le cas de notre étude, nous nous intéressons à l'observation de la réponse en fonction de la question et de ses caractéristiques. Nous cherchons à savoir comment elle est présentée quelle que soit l'information-réponse présentée, c'est-à-dire que nous nous attachons plus à la forme qu'au contenu. Nous avons décidé de limiter l'étude aux questions dont le type attendu de la réponse est élémentaire. Ces

réponses sont a priori plus simples que des réponses à des questions non élémentaires et nous évitons ainsi le risque d'obtenir des réponses trop complexes.

Lors de notre expérience, nous avons ainsi testé des questions factuelles dont la réponse attendue est une entité nommée. Parmi les classes proposées par [Benamara, 2004], nous avons construit des questions de base relevant de l'ensemble des classes de la hiérarchie hormis* les questions listes.

Des questions ayant beaucoup de variations possibles Nous voulons poser des questions qui puissent être formulées de façons très différentes. Nous évitons ainsi le biais qui pourrait être dû au fait de n'avoir que des questions qui prennent la même forme. Nous évitons aussi l'ennui du participant : il sera face à des questions diverses, qui ne se ressemblent pas toutes. Le modèle proposé par [Luzzati, 2001] et dont nous avons déjà parlé section 2.4.1 propose un ensemble de variations pour les questions de lieu, temps et quantité. Notre choix se porte donc sur ces trois types sémantiques de question.

Les exemples 4.1, 4.2 et 4.3 ont respectivement pour type sémantique quantité, lieu et temps.

Exemple 4.1 *Combien dure un mandat présidentiel en France ?*

Exemple 4.2 *Où se trouvent les Alpes ?*

Exemple 4.3 *Quand ont lieu les jeux olympiques d'été ?*

Différentes natures de réponses attendues Nous pensons que la nature de la réponse attendue (que nous avons traité section 3.6) joue un rôle sur la formulation de la réponse. Du point de vue des systèmes de question-réponse, il s'agit non pas de rechercher plusieurs réponses mais d'exploiter le fait que plusieurs réponses sont extraites lors de la phase de recherche d'information. Il peut bien sûr s'agir de réponses erronées. Mais il peut aussi s'agir de cas où une unique réponse n'existe pas.

Ainsi, nous avons des questions pour lesquelles la nature de la réponse est unique et des questions pour lesquelles la nature de la réponse est multiple.

Les exemples 4.4 à 4.6 sont des questions pour lesquelles la réponse est unique.

Exemple 4.4 *Combien mesure un double décimètre ?*

Exemple 4.5 *Dans quel musée se trouve la Joconde ?*

Exemple 4.6 *En quelle année a eu lieu l'armistice de la 1ère guerre mondiale ?*

Les exemples 4.7 à 4.9 sont des questions pour lesquelles la réponse est multiple.

Exemple 4.7 *Combien mesure un bébé à la naissance ?*

Exemple 4.8 *Où se trouvent les Alpes ?*

Exemple 4.9 *Quand ont lieu les jeux olympiques d'été ?*

Différents types sémantiques de réponses attendues Nous pensons aussi que le type sémantique de la réponse attendue joue un rôle sur la formulation de la réponse.

Pour les questions de quantité, nous avons créé des questions portant sur différents type de quantification : un poids (exemple 4.10), une longueur (exemple 4.11) et une durée (exemple 4.12).

Exemple 4.10 *Combien pèse un bébé à la naissance ?*

Exemple 4.11 *Combien mesure une feuille A4 ?*

Exemple 4.12 *Combien dure un mois de février ?*

Pour les questions de lieu, nous avons créé des questions portant sur différents types de localisation : le lieu où est un objet du monde (exemple 4.13), un lieu autre comme par exemple le lieu d'où vient un objet du monde ou le lieu par où il passe (exemple 4.14).

Exemple 4.13 *Où se trouve le Rhin ?*

Exemple 4.14 *Où peut-on traverser la Seine ?*

Pour les questions de temps, en revanche, les différents types de temporalité sont le moment auquel a lieu un évènement du monde et la durée d'un évènement du monde. Or la durée est une question considérée comme quantité. Les questions de durée étant déjà traitées parmi les questions de quantité, nous n'avons créé qu'une seule série de questions de temps (exemples 4.6 et 4.9 présentées précédemment).

Des questions "faciles" et non ambiguës Comme nous l'avons déjà précisé, les questions doivent être "faciles à répondre". En effet nous voulons éviter au maximum les réponses à valence négative (par exemple, "je ne sais pas"). Les thèmes abordés par les questions doivent être évidents, "connus de tous". Évidemment, tout le monde n'a pas les mêmes connaissances. Nous avons donc dû imaginer des questions à propos desquelles "tous le monde peut avoir une idée". Nous avons fait appel à des personnes dont la langue maternelle est le français et d'un niveau scolaire au moins post-baccalauréat. Nous avons ainsi choisi des questions relevant de la culture générale et en particulier de la culture française.
L'ensemble des thèmes abordés est le suivant :

- grossesse
- mandat présidentiel
- bébé à la naissance
- double décimètre
- feuille A4
- bouteille d'eau
- brique de lait
- avions à Paris
- Tour de France
- Rhône
- la Seine
- la Joconde
- la Vénus de Milo
- le Rhin
- les Alpes
- l'armistice de la première guerre mondiale
- les jeux olympiques d'été

Nous avons cherché à limiter au maximum l'ambiguïté des questions. Quand l'objet d'une question n'est pas clair, nous avons ajouté un complément d'information. Par exemple, dans la question portant sur les jeux olympiques, nous avons précisé qu'il s'agissait de ceux qui ont lieu en été.

Chaque question de base a été créé pour sa simplicité, son thème, les caractéristiques de l'objet sur lequel elle porte, sa non-ambiguïté et son type sémantique. Nous avons au total 15 questions de base auxquelles s'ajoutent 4 questions permettant de créer des équivalences de questions déjà dans notre liste. Ce point sera détaillé plus tard dans la section 4.4.3. Les 19 questions de base dans le tableau sont présentées dans le tableau suivant :

4.4.2 Facteurs de variation des questions

À partir de chacune des 19 questions de base, nous avons créé différentes variations en nous appuyant sur la GRINT. Ces variations dépendent des deux facteurs suivants :

- la structure syntaxique de la question (définie par la dimension syntagmatique de la GRINT)
- l'interrogatif utilisé dans la question (dimension interrogative)

Type de structure syntaxique Nous avons balayé l'ensemble des variations proposées par la GRINT. Les questions entité nommées ont été insérées dans ces varia-

Type sémantique de la question	**Questions de base**
Quantité	Combien dure un mois de février ? Combien dure une grossesse ? Combien dure un mandat présidentiel en France ? Combien mesure un bébé à la naissance ? Combien mesure un double décimètre ? Combien mesure une feuille A4 ? Combien pèse un bébé à la naissance ? Combien pèse une bouteille d'eau ? Combien pèse une brique de lait ?
Lieu	Où arrivent les avions à Paris ? * Où peut-on traverser la Seine ? Où arrive le Tour de France ? * Où finit le Rhône ? Où se trouve la Joconde ? * Où se trouve la Vénus de Milo ? Où se trouvent les Alpes ? * Où se trouve le Rhin ?
Temps	Quand a eu lieu l'armistice de la 1ère guerre mondiale ? Quand ont lieu les jeux olympiques d'été ?

TABLE 4.1 – *Liste des questions de base triées par type sémantique. Les questions précédées d'un * sont des questions "équivalentes".*

tions et relèvent systématiquement de deux types de granularités. Nous n'avons pas testé d'autres types d'entités nommées que celles dont le type sémantique relève des 3 types sélectionnés.

L'ensemble des variations pour lesquelles seule la structure syntaxique de la question de base "Combien dure un mois de février ?" change est présenté dans les exemples 4.15 à 4.19 :

Exemple 4.15 *Combien dure un mois de février ?*

Exemple 4.16 *Un mois de février dure combien ?*

Exemple 4.17 *Combien est-ce que dure un mois de février ?*

Exemple 4.18 *Je voudrais savoir combien un mois de février dure.*

Exemple 4.19 *Un mois de février dure 28 ou 29 jours selon les années ?*

Type de l'interrogatif De même que pour le type de l'interrogatif de la question, l'ensemble des variations proposées par la GRINT sera représenté.

Pour la même question de base que précédemment, voici l'ensemble des variations pour lesquelles seul le type d'interrogatif utilisé change :

Exemple 4.20 *Combien dure un mois de février ?*

Exemple 4.21 *Combien de jours dure un mois de février ?*

Exemple 4.22 *Quel temps dure un mois de février ?*

Exemple 4.23 *Que dure un mois de février ?*

Exemple 4.24 *Un mois de février dure-t-il 28 ou 29 jours ?*

Type du verbe principal Pour les questions de lieu et de temps, nous avons utilisé tantôt un verbe spécifique au type de question, tantôt un verbe neutre (un auxiliaire). Les exemples 4.25 et 4.26 utilisent des verbes neutres tandis que les exemples 4.27 et 4.28 sont les questions équivalentes utilisant un verbe spécifique.

Exemple 4.25 *Où est le Rhin ?*

Exemple 4.26 *Quand sont les jeux olympiques d'été ?*

Exemple 4.27 *Où se trouve le Rhin ?*

Exemple 4.28 *Quand ont lieu les jeux olympiques d'été ?*

4.4.3 Grille de passation

Une fois la liste de facteurs de variation établi, nous avons un total de 507 variations de question à soumettre aux participants dont 207 questions de quantité, 200 questions de lieu et 100 questions de temps.

Les variations de question sont construites suivant le modèle proposé par la GRINT. Or, tous les modèles n'étaient pas disponibles au moment où nous avons débuté la collecte de corpus. Le modèle des questions de quantité a été le premier finalisé. Nous avons donc débuté la collecte en ne posant que des question de quantité. Par la suite, les modèles des questions de lieu et de temps ayant eux aussi été établis, nous avons pu mener une seconde phase de collecte.

Nous avons créé des listes de questions à soumettre à chaque participant. Nous appelons l'ensemble des listes de question la grille de passation. Les listes de questions ont été créées de façon à présenter des formulations de question aussi variées que possible. Ainsi, un participant ne se verra pas poser deux questions de forme syntaxique identique l'une après l'autre. De même les thèmes des questions ont été répartis : on ne présente pas toutes les questions à propos du poids d'un bébé, puis toutes les questions à propos de la taille d'un double décimètre, etc. De même nous avons, pour un type thème de question donné, une vingtaine de formes syntaxiques de questions construites à partir de deux facteurs de variations créant des formulations de questions plus ou moins similaires. Nous avons espacé les formulations de questions les plus similaires de façon à éviter les redondances.

Pour les questions de quantité, nous avons 9 thèmes. Nous avons donc choisi d'établir des listes de 18 questions ne revenant ainsi que deux fois sur le même thème au cours de la passation. Les formulations de questions (23 pour les questions de quantité) autour d'un même thème ont elles aussi été espacées dans les listes de questions. Pour les 207 questions de quantité, il nous faut 207 / 18 = 11,5 listes de questions pour couvrir l'ensemble des formulations de question. Nous voulions avoir exactement 18 questions dans chaque liste, nous avions donc besoin de 12 listes de questions. Ces listes couvrent l'ensemble des 207 questions au moins une fois et certaines sont présentes deux fois.

Après la première phase de collecte, nous avons conclu que les listes de questions pouvaient être plus longues. Ce point est détaillé dans la section 4.5 traitant du déroulement des passations. Le nombre de thèmes abordés par les questions de temps et de lieu réunis est de 6 (4 thèmes pour les questions de lieu et 2 pour les questions de temps). Les listes de questions de la seconde phase de passation contiennent ainsi 24 questions (multiple de 6, légèrement supérieur à 18). L'avantage de créer des listes plus longues est celui de minimiser le nombre de participants nécessaire pour collecter suffisamment de réponses. Ce point est primordial car comme nous le verrons dans la section 4.5.1, trouver des participants n'est pas chose facile. Lors de cette seconde phase, nous avons soumis des questions de temps et de lieu. Elles sont au nombre de 300. Nous avions donc initialement besoin de 13 listes de

questions (300 / 24 = 12,5). Cependant de telles listes reviennent quatre fois sur le même thème. Pour minimiser cette forte redondance, nous avons décidé de dédoubler certaines questions en créant des questions équivalentes (selon nos facteurs de variation) aux questions déjà existantes. Par exemple, à la question “Où est la Joconde ?”, un équivalent est “Où est la Vénus de Milo ?”. Puisqu'il y a deux fois plus de questions de lieu que de questions de temps, nous avons choisi de dédoubler uniquement les questions de lieu portant ainsi le nombre de thèmes à dix. Les listes de questions ainsi établies alternent deux questions de lieu et une question de temps afin de soumettre aux participants l'ensemble des 300 formulations de question.

4.4.4 Corpus des questions

L'ensemble du corpus de questions est constitué de 707 questions dont 400 sont équivalentes deux à deux. Elles sont présentées dans l'annexeA.1.

4.5 Déroulement des passations

La passation des expériences a eu lieu en quatre phases. D'abord, nous avons monté l'expérience en utilisant uniquement les questions de quantité [4], chaque participant devant répondre à 18 questions et chaque forme variationnelle de question étant posée à au moins 3 participants différents. Cette expérience nous a permis de tester et valider le principe expérimental tant dans son déroulement que dans les résultats produits. Puis nous avons mené l'expérience équivalente à l'oral. Nous avons pu observer que le nombre de questions pouvait être augmenté, une passation durant à peine 10 minutes en moyenne et les retours sur l'expérience étant positifs (voir la section 4.6.7 pour plus de détails). Nous avons décidé d'augmenter le nombre de question posées à chaque participant. Nous avons ainsi mis en place l'expérience à l'écrit en utilisant les question de temps et lieu, en posant 24 questions à chaque participants. Enfin, l'expérience équivalente à l'oral a été mise en place.
L'augmentation du nombre de question par passation (passant de 18 à 24) a permis de diminuer notablement le nombre de participants nécessaires et donc d'accélérer d'autant le processus de collecte de ces données.

Nous avons fait appel à près de 150 participants. La passation a été organisée de façon à soumettre chaque question à au moins 3 participants rendant ainsi possible la comparaison des différentes réponses à une même question.

Le tableau suivant résume en quelques chiffres les caractéristiques des différentes phases de la collecte ainsi que le nombre de participants volontaires nécessaires pour obtenir ce minimum de 3 réponses par question.

4. Ceci résulte simplement du fait que la modélisation des questions de quantité selon la GRINT a été élaborée avant la modélisation des deux autres types de question qui au moment où nous avons débuté notre collecte n'étaient pas encore prêtes.

Phase	1	2	3	4
Type sémantique des questions	Quantité	Quantité	Temps et lieu	Temps et lieu
Modalité	Écrit	Oral	Écrit	Oral
Nb de questions	207	207	300	300
Nb de questions par passation	18	18	24	24
Nb de participants minimum nécessaire	35	35	38	38

TABLE 4.2 – *Nombre de participants nécessaire par phase de collecte.*

4.5.1 Recrutement des participants

Nous avons d'abord recruté des participants au sein de notre entourage (amis, famille, collègues, mais aussi étudiants de notre université, ...) en ne nous adressant pas à ceux ayant connaissance de nos travaux. Si ce mode de recrutement est très efficace (le taux de participation par rapport au nombre de personnes contactées est très élevé), il reste limité. Nous avons donc utilisé la base de contact proposées par le site "Faites la science" [Faites-la-Science, site web] mis en œuvre et hébergé par le Relais d'Information en Sciences Cognitives (RISC, [Faites-la-Science, site web]). Ce site propose à des volontaires de s'inscrire sur une liste de diffusion et à des expérimentateurs de publier des annonces de recherche de participants. La base est actuellement d'environ 1600 volontaires [5]. Elle en comptait près de 1200 lors du dépôt de nos annonces (en mars 2009). Ce mode de recrutement est très efficace car il permet de contacter un grand nombre de personnes *volontaires* facilement. Cependant, le taux de participation par rapport au nombre de personnes contactées est bien moins intéressant. Il faut noter que la plupart des participants répondent dans les quelques jours après la publication de l'annonce.

4.5.2 Taux de participation et explication des non-passations

Le tableau 4.3 indique, pour chaque étape du recrutement, le nombre de participants ayant atteint cette étape et le pourcentage de participants qui se sont arrêtés à l'étape précédente. On observe que la plupart des gens ne répondent pas à l'appel à participation (en moyenne 88%). On observe clairement une différence d'implication entre les deux modes de recrutement puisque pour un recrutement dans l'entourage, toutes les personnes qui ont répondu à l'appel se connectent en effet à la plate-forme d'expérimentation et seulement 16% ne font finalement pas l'expérience. En revanche, dans le cas du recrutement *anonyme* proposé par la site "Faites-la-Science", 12% des personnes ayant accepté de participer ne se con-

5. 1607 en septembre 2010.

nectent pas à la plate-forme et 66% de celles qui se sont connectées au moins une fois ne font pas l'expérience.

	Total	Entourage	Faite-la-Science
Nb de personnes contactées	2164	964	1200
Nb de personnes ayant répondu et accepté de participer	257 (-88 %)	138 (-85 %)	119 (-90 %)
Nb de personnes ayant accédé à l'une des plates-formes (participation valide ou non)	242 (-5 %)	138 (-0 %)	104 (-12 %)
Nb de passations réelles effectuées	152 (-37 %)	116 (-16 %)	35 (-66 %)
Pourcentage de personnes contactées ayant participé et dont la participation a été validée	7 %	12 %	3 %

TABLE 4.3 – *Évolution du nombre de contacts en fonction des étapes de participation à l'expérience.*

Quel que soit le mode de recrutement, nous observons que le taux de participation par rapport au nombre de personnes contactées initialement est faible (7% en moyenne). Il semble donc primordial de réduire au maximum le nombre d'étapes de recrutement. Dans notre cas, parce que nous utilisions des identifiants de connexion, nous avons dû envoyer un mail d'appel à participant, puis, pour chaque acceptation, un mail donnant les informations d'accès à la plate-forme. Un participant devait donc (1) recevoir et lire l'appel (2) envoyer un mail d'acceptation (3) recevoir et lire les information d'accès à la plate-forme (4) aller sur la plate-forme et se connecter (5) faire l'expérience.
En revanche, il aurait été préférable de donner directement l'adresse de la plate-forme en utilisant un système de gestion des identifiants automatisé. Dans un tel cas, nous supposons que le taux de participants par rapport au nombre de personnes contactées peut être largement augmenté. Un autre avantage de rendre ce processus automatique est le gain de temps puisque, à chaque acceptation de l'appel, nous avons dû attribuer un identifiant et répondre au mail reçu.
Malgré les difficultés de recrutement, plus de 150 participants ont répondu à nos questions. Nous proposons à présent de quantifier le corpus collecté.

4.5.3 MACAQ : Le corpus collecté

L'ensemble du corpus collecté a été baptisé MACAQ : Multi-Annotated Corpus of Answers to Questions (Corpus Multi-Annoté de Réponses à des Questions). En effet, en plus d'être un corpus de réponse à des questions, MACAQ a été plusieurs fois annoté de façon à repérer différents éléments composant tant les questions que les réponses. Le tableau 4.4 présente une description générale du corpus obtenu

et des deux sous-corpus correspondant aux réponses obtenues à l'oral et celles obtenues à l'écrit.

	Total	Oral	Écrit
Nb de réponses	3 132	1 044	2 088
Nb questions différentes	507	436	507
Nb de participants	152	53	99
Nb de question par participants	18 ou 24		
Nb réponses par question	6,17	2,39	4,11
Nb mots total	25 104	7 128	17 976
Taille du lexique	4 574	1 634	3 363
Nb mots moyen par réponse	8,01	6,82	8,60

TABLE 4.4 – *Caractéristiques générales du corpus*

Le nombre de participants (152 au total) est de 53 à l'oral et 99 à l'écrit. En effet moins de participants ont accepté de participer à l'expérience par téléphone qu'à celle par Internet.
Le corpus est constitué de 3 132 réponses dont deux tiers de réponses écrites et un tiers de réponses orales transcrites.

Il correspond à plus de 25 000 mots pour près de 4 600 mots distincts. La taille du lexique est deux 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 artificiellement 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 précisément, dans deux tiers des cas, les réponses à une même question sont plus longues à l'écrit qu'à l'oral.

4.6 ÉVALUATION DU PROTOCOLE EXPÉRIMENTAL

4.6.1 Spontanéité des réponses des locuteurs

Nous entendons par réponse spontanée, une intervention qui n'a pas été préparée à l'avance et qui a été produite par le participant lui-même au moment de l'expéri-

ence. À l'écrit, le fait qu'une réponse soit écrite puis modifiée et réécrite n'est pas, à notre sens, un signe de non-spontanéité.

Observation du temps de réponse

Nous avons observé le temps de réponse à l'écrit et l'oral. Nous rappelons que le temps de réponse à l'écrit est mesuré entre deux validations de question. Il s'agit donc du temps que met le participant à valider sa réponse à partir du moment où une question s'affiche sur la page Internet. Cette mesure ne correspond donc pas au temps de réponse mais englobe un certain nombre d'actions de la part du participant : lecture de la question ; saisie la réponse dans la zone de texte prévue à cet effet ; éventuellement, relecture de la réponse saisie ; validation du bouton permettant de passer à la question suivante [6]. À l'oral, en revanche, cette mesure correspond au temps entre la fin de la formulation de la question par le système et la fin de la formulation de la réponse par le participant. On ne prend donc pas en compte le temps d'écoute de la question.

Le tableau 4.5 donne des indications sur la durée des réponses pour l'ensemble du corpus de réponse et chacun des deux sous-corpus constitués uniquement des réponses à l'oral ou à l'écrit. Nous notons que les durées sont bien différentes selon la modalité. On peut supposer que leur distribution aussi, étant donné que les moyennes et écarts-types des deux sous-corpus sont bien différents.

	Tout	Oral	Écrit
Moyenne	22.48	3.28	34.11
Minimum	0.73	0.73	1.00
Maximum	1379.00	23.82	1379.00
Écart-type	45.00	2.46	54.38

TABLE 4.5 – *Durée des réponses en seconde sur l'ensemble du corpus, à l'oral et à l'écrit*

Les boîtes à moustaches [7] de la figure 4.5 montrent clairement que non seulement les durées des réponses à l'oral et l'écrit ne sont pas comparables mais aussi que leur distribution est bien différente.

Nous observons donc les durées des réponses de façon indépendante à l'oral et à l'écrit.

À l'écrit, comme nous le voyons dans la boîte à moustache supérieure de la figure 4.5, la plupart des réponses ont une durée située entre 40 et 132 secondes et quasiment toutes ont une durée comprise entre 1 et 266 secondes.

6. On ne tient pas compte des évènements extérieurs à l'expérience comme par exemple les cas où le participant est dérangé, etc.

7. Une boîte à moustache est construite de telle façon que le rectangle central représente 50% des effectifs, la ligne centrale représente la médiane et l'espace entre le trait à l'extrême gauche et celui à l'extrême droite représente 90% des effectifs.

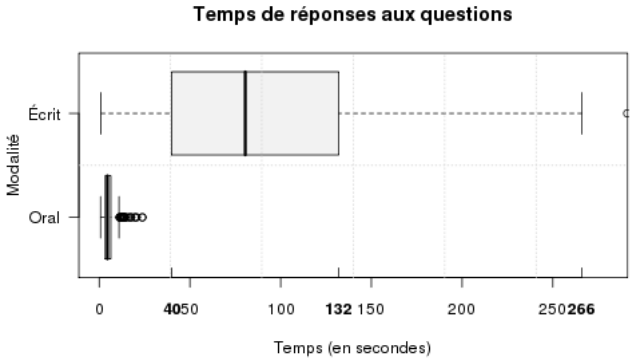

FIGURE 4.5 – *Boîtes à moustache des durées des réponses à l'oral et à l'écrit.*

Saisir une réponse au clavier peut prendre un temps variable en fonction de la vitesse de frappe du participant. Ceci peut sans doute expliquer la variabilité du temps de réponse à l'écrit. Il serait hasardeux de supposer de l'habitude de nos participants à écrire au clavier. De plus, ils ont pu être dérangés ou même déconcentrés au cours de leur passation. Il est donc délicat de se prononcer sur la spontanéité des réponses à partir de l'observation de la durée des réponses. Notre intuition est qu'il est peu probable que les participants aient recherché une réponse toute faite (sur Internet par exemple) au cours des passations puisque la durée des réponses dépasse peu souvent les 2 minutes et très rarement les 4 minutes. Si pendant une telle durée, il est envisageable d'écrire et réécrire une réponse, il est plus difficile d'avoir le temps de faire une recherche *fructueuse* de la réponse et de la recopier.

À l'oral, la figure 4.6 zoome sur la boîte à moustache concernant ce sous-corpus de réponses. On observe que la plupart des réponses ont une durée située entre 3,3 et 6,4 secondes et quasiment toutes ont une durée comprise entre 0,7 et 10,9 secondes. Il s'agit donc de réponses très rapides.

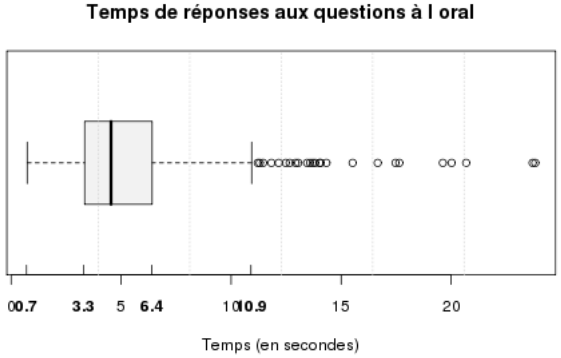

FIGURE 4.6 – *Boîtes à moustache des durées des réponses à l'oral et à l'écrit.*

Observation des hésitations et marques de doute

Hésitations à l'oral Comme nous l'avons vu page 89 au cours de la section détaillant le guide de transcription suivi lors de la transcription des réponses orales, l'hésitation *euh* a été transcrite par le symbole "`{fw}`" (de l'anglais, *filler word*).

Un simple décompte des hésitations montre que 23% des réponses contiennent une hésitation.

Marques de doute L'annotation d'expressions de doute ("il me semble que", "je dirais",...) a été faite manuellement sur les premières réponses collectées à savoir celles collectées lors de la première phase de l'expérience (donc des questions de quantité, posées à l'écrit).

Cette annotation montre que 5% des réponses du sous-corpus concerné contiennent au moins une expression de doute.

Observation qualitative de réponses spontanées

Certaines réponses peuvent paraître gênantes et ne pourraient être considérées que comme du bruit ou des réponses de mauvaise qualité au sein du corpus. Notre but est d'obtenir des réponses "sérieuses" puisque nous voulons les observer, les étudier, voire même les faire simuler par un système automatique. Les systèmes de question-réponse actuels ne sont pas capables de faire de l'humour par exemple. Cependant, dans le cadre de l'évaluation de notre protocole expérimental, nous estimons que la présence de réponse faisant preuve d'humour ou encore faisant référence à des éléments externes éventuellement complexes sont des signes indiquant une certaine aisance de la part du participant et peuvent donc être interprétés comme des indicateurs de spontanéité.

Nous avons relevé cinq types de signes :

- *l'humour*, les jeux de mots et autres processus mentaux "complexes" (exemple 4.29),
- *les références à l'interlocuteur* par un pronom ou nom le désignant (exemples 4.30 et 4.31),
- la *complexité linguistique* qui se traduit par exemple par l'utilisation d'expression idiomatique ou de phrases complexes (exemple 4.32),
- les *fautes linguistiques* à l'écrit, qui peuvent être des fautes d'orthographe, de frappe ou encore des abréviations (exemple 4.33)
- et les hésitations à l'oral (exemple 4.34).

Voici des exemples, issus du corpus, pour chacun de ces types de signes :

Exemple 4.29 *[A1582] Helas... Si seulement on pouvait se debarrasser de cette crapule plus tot...*

Exemple 4.30 *[A278] Va regarder à la fin de ton agenda, cela doit être marqué.*

Exemple 4.31 *[A1360] et bien comme je le disait à ton camarade entre [e] 3 et 4 kilos*

Exemple 4.32 *[A79] Les Alpes se trouve à cheval sur plusieurs pays : la France, l'Italie, l'Autriche, l'Allemagne, et sur l'ensemble du territoire Suisse.*

Exemple 4.33 *[A521] Dans plusieurs, l'allemagne, la france et ptetre la belgique*

Exemple 4.34 *[A1138] {fw} oui c' est à dire que {fw} la section graduée mesure 20 centimètres*

Nous n'avons pas quantifié ces signes au sein du corpus car ceci demande une annotation préalable de l'intégralité du corpus. Or nos cinq types de signes ont été choisis de façon intuitive et l'annotation d'un corpus est coûteuse (notamment car nous sommes persuadé que nous aurions été confronté à bon nombre de cas "litigieux" comme dans tous les cas de travaux d'annotation) pour un apport minime.

La présence de certaines réponses contenant des signes de spontanéité nous suffit à penser que l'expérience n'a pas été fastidieuse pour nos participants et qu'au moins une partie d'entre eux se sont pris au jeu : ils ont formulé des réponses à destination d'enfants.

D'autre part, il est à noter qu'aucune réponse n'est hors-sujet. Les interventions des participants, que ce soit à l'oral ou à l'écrit, sont bien cohérentes par rapport à la question posée.

Conclusion

L'observation de la durée des réponses et notamment de leur durée maximale, nous permet de supposer que les réponses ont été construites et produites par les participants eux-mêmes, au moment de l'expérience. La présence d'hésitations et d'expressions de doute mais aussi de réponses contenant des signes de spontanéité est aussi un indice en faveur de la conclusion de la spontanéité des réponses.

4.6.2 Facilité de donner une réponse aux questions

Nous considérons une réponse comme "facile", si il est aisé d'y donner une réponse. La véracité de la réponse n'est dans, notre cas pas importante, et nous n'observons pas cette caractéristique. L'important est d'avoir une information répondant à la question. Ce point de vue est à distinguer de la facilité de trouver la réponse à la question. Le fait de trouver (en mémoire) la réponse à une question relève plus d'une facilité d'accès cognitif, ou bien d'une facilité d'évaluation de

la réponse. Par exemple, demander son nom à quelqu'un est une question "facile" tant du point de vue informatif (on sait qu'il faut répondre un nom) que cognitif (il s'agit de son propre nom).
À l'opposé, demander la définition d'un concept, par exemple, est difficile tant du point informatif (quelle doit être la nature de la réponse ?) que du point de vue cognitif (une telle question demande un travail de synthèse et de réflexion).
Une question portant sur le poids d'un bébé à la naissance est facile d'un point de vue informatif (même si on peut hésiter, donner une réponse n'est pas compliqué et on peut estimer que nous avons tous une idée de la valeur à répondre) mais relativement difficile d'un point de vue cognitif (car elle demande de faire une estimation notamment).

Observation du nombre de *réponses à valence positive*

Nous définissons comme "réponse à valence positive", une réponse qui contient (au moins) une information-réponse répondant à la question.
Une "réponse à valence négative" au contraire n'en contient aucune. Il peut s'agir d'un aveu d'incompétence ("je ne sais pas"), d'une suggestion pour trouver la réponse autrement ("va regarder à la fin de ton agenda"), d'une correction de la question (voir l'exemple 4.36 [8], réponse à la question Q154 [9]).

Exemple 4.35 *[Q154] Que dure une grossesse ?*

Exemple 4.36 *[A732] Combien de temps dure une grossesse ? Environ 9 mois chez l'homme.*

Ces réponses sont à distinguer des réponses ne contenant pas d'information répondant à la question. Nous les nommons les "non-réponses".

Nous avons repéré et annoté les informations-réponse au sein du corpus. La valence positive ou négative de ces informations-réponse a été annotée. Cette annotation manuelle a plusieurs intérêts. Elle permet par exemple d'observer les informations-réponse de façon qualitative, mais aussi de les quantifier, de repérer les réponses comportant plus d'une information-réponse et de comparer ces différentes informations-réponse. Dans le cas présent, elle permet de comptabiliser les réponses qui ne contiennent pas d'information-réponse. Plus de détail sur cette annotation sera donné dans la section 5.3.1 dédiée à la description de cette annotation dans le cadre de l'étude de la nature et de la quantification des informations-réponse dans le corpus.

Dans le tableau 4.6 nous pouvons voir que seules 2,6% des réponses ne contiennent pas d'information-réponse (non-réponses) et parmi celles qui en contiennent, 3% ne contiennent pas d'information-réponse à valence positive (réponse à valence négative). Au final, plus de 94% des réponses du corpus contiennent au

8. Les exemples commençant par [A…] sont des exemples de réponse issues de notre corpus.
9. Les exemples commençant par [Q…] sont des exemples de question issues de notre corpus.

moins une information-réponse à valence positive. Ce sont les réponses que nous recherchons à collecter.

	Tout	Oral	Écrit
Nb de réponses	3132	1044	2088
Nb de non-réponses	81	43	37
% de non-réponses	2.59	4.12	1.77
Nb de réponses à valence négative	95	22	73
% des réponses à valence négative	3.03	2.11	3.5
% des réponses à valence positive	94.38	93.77	94.73

TABLE 4.6 – *Quantification des réponses ne contenant pas d'information-réponse*

Conclusion

Plus de 94% des réponses du corpus contiennent au moins une information-réponse à valence positive. Ce sont précisément les réponses que nous recherchons à collecter. Nous pouvons donc considérer que les questions sélectionnées sont suffisamment "simples". Ce résultat est très positif car le nombre de réponse à valence positive est très important. Nous avons donc une large partie du corpus que nous pouvons utiliser pour observer comment sont formulées ces informations-réponse positives. Le fait que 3% des réponses soient des réponses à valence négative est aussi intéressant car nous avons à notre disposition, près de 100 réponses pour étudier ce cas particulier [10].

4.6.3 Obtention de *réponses non succinctes*

Le contexte expérimental, parce qu'il suggère que les réponses sont destinées à des enfants de CM2, devrait nous avoir permis d'obtenir non pas des réponses très courtes contenant uniquement l'information répondant à la question (ou information-réponse) mais plutôt des phrases ou des énoncés. Nous rappelons ici la définition de ces réponses (déjà donnée section 2.3 du chapitre 2) :

Définition 4.1 ***Réponse succincte*** *: réponse qui est constituée uniquement d'une information-réponse.*

L'exemple 4.38 présente réponse succincte à la question Q098 de notre corpus de question.

Exemple 4.37 *[Q098] Combien de centimètres mesure un double décimètre ?*

Exemple 4.38 *[A2300] 20*

10. La section 5.7 s'attache à l'étude de ces réponses.

Observation de la taille des réponses

	Tout	Oral	Écrit
Moyenne	7,05	5,85	7,66
Minimum	1	1	1
Maximum	98	62	98
Écart-type	8,42	7,23	8,9

TABLE 4.7 – *Taille des réponses en nombre de mots*

Le tableau 4.7 nous montre qu'en moyenne les réponses sont constituées de 7 mots [11] avec un écart-type très important de plus de 8 mots. À partir de la figure 4.7, nous pouvons observer qu'un grand nombre de réponses (Plus de 1 000) ne sont en fait constituées que d'un à deux mots. On peut supposer que ces réponses sont en fait constituées uniquement d'une information-réponse et sont donc des réponses succinctes. Cependant le nombre de mots des réponses n'est pas une information suffisante pour en être sûr. La section suivante propose une observation plus détaillée des réponses basée sur une annotation du corpus.

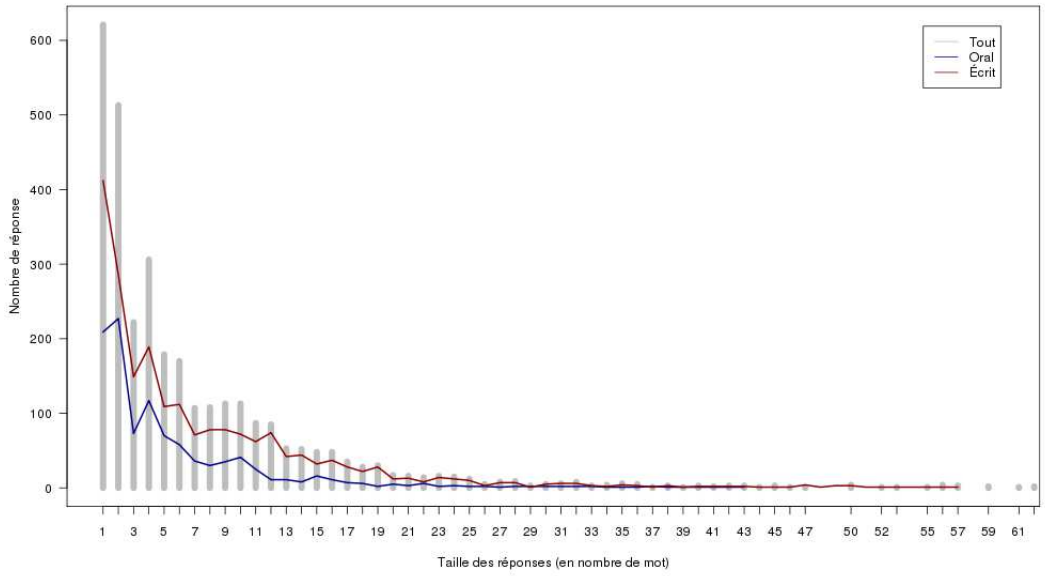

FIGURE 4.7 – *Histogramme et courbes de fréquence des longueurs des réponses en nombre de mots*

11. Nous considérons un mot comme une suite de caractère entre deux séparateurs : espace ou ponctuation.

Observation du nombre de réponses succinctes

Nous avons annoté manuellement l'information-réponse au sein des réponses du corpus. Cette annotation sera détaillée plus en avant dans une section dédiée (voir page 129 du chapitre 5). L'annotation nous permet de repérer les réponses succinctes c'est-à-dire les réponses qui ne sont constituées que d'une information-réponse. Le tableau 4.8 montre que plus de 40% des réponses sont des réponses succinctes. Le taux de réponses succinctes est comparable à l'oral et à l'écrit.

	Tout	Oral	Écrit
Nb de réponses	3132	1044	2088
Nb de réponses succinctes	1370	477	893
Taux de réponses succinctes	43,74%	45,73%	42,79%

TABLE 4.8 – *Quantification des réponses succinctes au sein du corpus.*

Conclusion

Le protocole n'a pas permis d'obtenir systématiquement des réponses non succinctes. Plus de 55% des réponses ne sont pas succinctes, ce qui n'est pas négligeable. Nous aurions pu préciser clairement que nous nous attendions à des "phrases" et non pas "la réponse seule" ou encore donner un exemple de réponse. Cependant, ceci aurait pu être un biais sur la façon de répondre de nos participants. Peut-être que d'autres contextes expérimentaux auraient permis de minimiser le taux de réponses succinctes. Par exemple, en proposant que les réponses sont destinées à des apprenants du français. Le contexte joue un rôle important sur le choix des questions : il faut qu'il y ait une certaine cohérence entre les thèmes abordés par les questions et ce contexte expérimental. Changer le contexte reviendrait donc aussi à changer les questions et peut-être avoir des résultats moins satisfaisants concernant d'autres critères que nous évaluons sur le protocole notamment concernant la facilité de répondre aux questions.

4.6.4 Obtention de réponses sans information complémentaire

Nous avons cherché, en mettant en place notre protocole expérimental, à obtenir des réponses que les systèmes de question-réponse pourraient eux aussi produire. Nous entendons ceci quel que soit le système c'est-à-dire que nous n'entendons pas que ces systèmes soient capables de faire mieux que de trouver une information-réponse à la question. Nous avons ainsi cherché à obtenir des réponses contenant uniquement des éléments dont les systèmes de question-réponse disposent : une information-réponse et la question. Auxquelles peuvent s'ajouter des expressions figées (pour introduire la réponse par exemple), des mots outils,... Une annotation manuelle du corpus permet d'identifier trois grandes classes d'éléments :

- les éléments qui correspondent à une nouvelle information : il peut s'agir de l'*information-réponse* ou bien d'informations autres non demandées dans la question. Nous parlons d'*informations complémentaires*,
- ceux qui correspondent à une ancienne information : il s'agit d'informations présentes dans la question et sont donc des *réutilisations d'éléments de la question*
- et ceux qui sont en rapport avec la situation d'énonciation : il s'agit de *marqueurs dialogiques ou pragmatiques*. Ces éléments ne contiennent pas une nouvelle information mais relève de l'utilisation de connaissances linguistiques. Par exemple, il peut s'agir de l'utilisation de connecteurs ou encore d'expressions ou locutions introduisant la réponse.

Ce que les systèmes de question-réponse utilisent déjà sont les informations anciennes et l'information nouvelle qu'est la réponse. En effet, ces systèmes analysent la question et sont capables d'identifier les différents éléments qui composent une question et ils ont pour but de fournir une information-réponse. Les marqueurs dialogiques et pragmatiques sont des connaissances linguistiques en particulier des expressions figées ou semi-figées. Elles peuvent être par exemple intégrées au module de génération de la réponse en langue naturelle et ce, sans modifier l'architecture du système de question-réponse. En revanche, les nouvelles informations autres que l'information-réponse ne sont pas disponibles. Nous ne pouvons nous baser sur des réponses contenant ce type d'élément pour proposer des schémas de réponse en langue naturelle quel que soit le système de question-réponse. Notre protocole a ainsi été construit de façon à minimiser le nombre de réponses contenant des informations complémentaires. Nous appelons les réponses contenant ces éléments des réponses complétives et les définissons ainsi :

Définition 4.2 ***Réponse complétive*** *: réponse qui contient au moins un élément de complétion ou de suggestion, c'est-à-dire un élément qui apporte une nouvelle information mais qui ne correspond pas à l'information-réponse. Les mots outils, les marqueurs du dialogue, les hésitations, les marqueurs pragmatiques ne sont pas considérés comme des éléments de complétion puisqu'ils n'apportent pas de nouvelle information.*

Observation du nombre de réponses complétives

Le tableau 4.9 donne le nombre de réponses complétives et leur taux dans l'ensemble du corpus et les sous-corpus oral et écrit. On observe que très peu de réponses sont complétives à l'oral (moins de 6%) et à peine plus de 11% le sont à l'écrit.

Conclusion

Le protocole expérimental et notamment, la situation proposée et les instructions données aux participants, les ont incités à se contenter de répondre à la question sans donner, la plupart du temps, d'information nouvelle autre que la réponse.

	Tout	Oral	Écrit
Nb de réponses	3132	1044	2088
Nb de réponses complétives	288	56	232
Taux de réponses complétives	9,19%	5,36%	11,11%

TABLE 4.9 – *Quantification des réponses complétives au sein du corpus.*

4.6.5 Possibilité de comparaison entre oral et écrit

Observation du nombre de réponses à l'oral et à l'écrit

Le tableau 4.10 montre à la fois le nombre de réponses par sous-corpus et le nombre de participants. Nous observons que le corpus de réponses orales est moitié moindre que celui des réponses écrites. En effet, moins de participants ont accepté de participer à l'expérience par téléphone que par Internet. Nous expliquons ceci par l'accessibilité plus immédiate de l'interface en ligne. En effet, même si par téléphone, nous avions mis en place 3 lignes téléphoniques dont une à numéro vert (gratuit), l'accès à Internet est plus répandu que celui à un téléphone. De plus, certains opérateurs de téléphonie semblent faire payer les communications y compris sur des numéros verts. Or le coût d'un appel, pour des étudiants notamment [12] peut être rebutant, l'expérience durant environ 10 minutes. D'autres part, participer à cette expérience orale demandait de pouvoir téléphoner dans des conditions calmes et non bruitées.

	Tout	Oral	Écrit
Nb de réponses	3 132	1 044	2 088
Nb questions différentes	507	436	507
Nb de participants	152	53	99
Nb réponses/question	6,17	2,39	4,11

TABLE 4.10 – *Tableau comparatif des sous-corpus oral et écrit*

Conclusion

Bien que les sous-corpus des réponses obtenues à l'oral et à l'écrit ne soient pas de la même taille, nous avons pu obtenir par deux versions d'un même protocole expérimental [13] des réponses orales et écrites. Les conditions d'accès à la plateforme orale rendent la collecte d'un corpus oral plus difficile que celle d'un corpus écrit. Une alternative pourrait être de proposer une expérience orale par Internet. Mais ceci demanderait aux participants d'avoir à leur disposition un microphone et des haut-parleurs. D'autre part, ceci n'évacuerait pas la difficulté que peut être celle de trouver un lieu calme et isolé (sans bruit de fond) pour participer à l'expérience.

12. Une partie non négligeable de nos participants sont en effet étudiants.
13. La section 4.3.2 a détaillé les différences entre les deux versions du protocole expérimental.

4.6.6 Obtention de diverses réponses pour une question

Observation du nombre de réponses par *groupe de questions*

Comme nous l'avons vu dans la section 4.4.2 à partir d'une question de base, nous avons créé différentes formulations de question. Ici, nous observons le nombre de réponses obtenues pour l'ensemble des variations d'une même question de base ou *groupe de questions*. Le tableau 4.11 donne la moyenne, le maximum, le minimum et l'écart-type du nombre de réponses obtenus par *groupe de questions* [14].

	Tout	Oral	Écrit
Nb de réponses au total	3132	1044	2088
Nb de groupe de questions	18	18	18
Nb minimum de réponses par groupe	132	30	95
Nb maximum de réponses par groupe	345	153	204
Nb moyen de réponses par groupe	173,89	57,94	115,94
Écart-type du nb de réponses par groupe	59,21	32,09	38,86

TABLE 4.11 – *Nombre de réponses par groupe de questions*

On observe qu'il y a au minimum 132 réponses par groupe de questions pour une moyenne de 174 réponses par groupe. À l'oral, nous avons obtenu moins de réponses mais le minimum de réponses par groupe est de 30.

Conclusion

Avec au minimum 30 réponses par groupe de questions à l'oral, 95 à l'écrit et 132 toutes modalités confondues des comparaisons entre les réponses à une même question sont possibles. Le nombre important de facteur de variation à permis la création de grands groupes de questions et donc la collecte d'un nombre important de réponses par groupe.

4.6.7 Analyse qualitative des remarques faites par les participants

Pour la version écrite de l'expérience, les participants ont pu laisser des remarques libres après avoir répondu à l'ensemble des questions. Pour sa version orale, un message audio leur indiquait qu'ils pouvaient envoyer leur remarques par mail. Nous avons reçu peu de remarques (environ 50 pour l'expérience écrite et 30 pour l'expérience orale).

14. Il faut noter que les groupes de questions ayant obtenu le minimum (respectivement le maximum) de réponses à l'écrit et à l'oral ne correspondent pas forcément. Ainsi la somme du nombre minimum (resp. du nombre maximum) de réponses par groupe à l'oral et de ce nombre concernant l'écrit n'est pas nécessairement égale au nombre minimum (resp. maximum) de réponses par groupe pour l'ensemble du corpus.

Observation des remarques données par les participants

Les remarques peuvent être classées grossièrement dans les différentes classes suivantes :

- remarques d'encouragement,
- doute du participant sur la qualité de sa participation,
- description de l'impression générale,
- critiques (positives ou négatives) sur l'expérience,
- demande de retour sur les résultats,
- récapitulatif des réponses données ou bien correction sur une ou plusieurs réponses données [15],
- explications ou justifications sur certaines réponses données ou encore sur le temps mis pour répondre à une question,
- tentative de découverte du but de l'expérience.

L'ensemble des remarques montre un intérêt certain pour l'expérience. Aucune remarque n'exprime un désaccord ou une critique négative forte. En revanche, nous avons nombre de remarques exprimant que l'expérience a été vécue comme amusante, souhaitant même bon courage aux enfants de CM2 dans leur travail.

En revanche, quelques remarques indiquent la difficulté de comprendre certaines questions "mal formulées" ou encore des erreurs (coquilles ou problèmes d'affichage à l'écrit ; mauvaise prononciation à l'oral). Concernant la version téléphonique de l'expérience, des critiques portent sur l'impossibilité de réécouter la question. Enfin, des critiques négatives ont porté sur la redondance des questions (une concernant une passation à l'oral et deux concernant une passation à l'écrit).

Conclusion

L'expérience semble avoir été appréciée par les participants et même si certains (3 parmi les 80 participants ayant donné un retour et parmi 150 participants au total) ont trouvé les questions redondantes, ceci ne semble pas avoir été un problème majeur. Nous rappelons qu'au maximum 4 questions du même groupe (c'est-à-dire portant sur le même thème) ont été posées au cours d'une unique passation.

4.6.8 Conclusion sur l'évaluation du protocole

Le protocole est satisfaisant dans l'ensemble car il permet d'obtenir un corpus de réponses à nombre de variations autour de questions de base et ce sur deux modalités d'interaction. Les réponses peuvent être supposées spontanées et l'expérience a été jugée agréable par les participants.

Cependant, la mise en place de cette collecte est lourde du point de vue de la création des listes de question à soumettre à chaque passation (élément crucial

15. Nous n'avons dans ce cas pas pris en compte les corrections afin de préserver la spontanéité des réponses.

qui a permis d'avoir un maximum de variations dans les questions au cours d'une même passation), du point de vue du recrutement des participants mais aussi du point de vue des réponses obtenues puisque 43% des réponses obtenues sont des réponses succinctes.

CONCLUSION DU CHAPITRE

Nous avons présenté un protocole expérimental unique qui permet la collecte de réponses à des questions. Notre but était d'obtenir diverses réponses pour la même question quelle que soit sa variante. Le protocole mis en place est satisfaisant selon les critères d'évaluation que nous avons définis et respectent les contraintes que nous nous sommes fixées.

Nous disposons à présent d'un corpus unique parce qu'il est oral et écrit, par sa taille (plus de 150 participants et de 3100 réponses), par sa nature (les réponses sont des réponses spontanées de locuteurs natifs du français) et enfin par le contrôle extrêmement précis des variations de questions auxquelles les participants ont répondu.

Ce corpus est disponible [16] à la communauté scientifique sous forme brute et sous forme annotée. Seules les enregistrements audios des participations orales ne sont pas disponibles et ce, pour des raisons de droits.

Nous avons analysé ce corpus suivant plusieurs axes dans l'optique de répondre à certaines questions qui nous semblent clefs dans le cadre d'un travail de recherche mettant en place un module de génération de réponse en langue naturelle au sein d'un système de question-réponse. C'est ce que présente le chapitre suivant.

16. S'addressez à moi-même ou bien l'une de mes directrices de thèse : {annegf, sophie.rosset, anne.vilnat}@limsi.fr

5 QUELS CRITÈRES POUR GÉNÉRER UNE RÉPONSE EN LANGUE NATURELLE ?

SOMMAIRE

Ce chapitre présente un ensemble d'analyses et d'annotations faites sur le corpus de réponses collecté. Le point de vue adopté est d'identifier les questions à se poser pour mettre en place un module de génération de réponses en langue naturelle dans un système de question-réponse, puis de répondre à ces questions. Chacune des sept questions constitue une section du chapitre. Pour chacune d'elles, une étude du corpus est présentée. L'ensemble des analyses du corpus consiste en un ensemble d'annotations, la définition de notions clefs, l'utilisation d'outils et de mesures statistiques, l'utilisation d'outils d'analyse de texte, la présentation des résultats sous forme de tableaux et de graphiques et, pour chaque question, une synthèse des résultats obtenus.

Ce chapitre correspond à une seconde grande phase du travail de thèse.

This chapter presents annotations and analyses done on the collected corpus. We identify and answer questions one should answer before implementing a module of natural language answer generation in a question-answering system. To each of those seven questions, we dedicate a section presenting analyses and annotations of the corpus, definitions of key notions, use of tools and statistical measures, graphical or tabular presentations of the results and a synthesis.

5.1 INTRODUCTION

Comme nous l'avons vu dans l'introduction de ce document, notre but est de donner la possibilité à des systèmes de question-réponse de générer des réponses en langue naturelle et ce, "à moindre coût". Notre idée est que, quel que soit le système considéré, à partir du moment où il ne s'agit pas d'un système purement statistique, alors introduire un module de génération de la réponse n'implique pas nécessairement d'effectuer des analyses ou traitements supplémentaires ni sur les documents, ni sur les questions. Pour cela, nous avons vu que les éléments dont tous les systèmes disposent sont la question en elle-même (analysée) et la réponse trouvée par le système. Notre idée est d'étudier la possibilité pour des systèmes de question-réponse de répondre des phrases construites à partir de ces deux éléments.

Quelles sont les questions à se poser lorsque l'on cherche à formuler une phrase en langue naturelle ?

Une première question est de savoir quelle information-réponse on doit fournir. De quel type d'information-réponse s'agit-il ? Un lieu, une date, un nom,... ? Il s'agit d'un travail qu'effectuent déjà les systèmes de question-réponse. L'observation des information-réponse du corpus est présentée dans la section 5.3.

Ensuite il est important de savoir si, réutiliser des informations issues de la questions peut permettre de générer de bonnes réponses et est judicieux. Notre idée est que si des humains réutilisent en effet des éléments issus de la question, alors le fait qu'un système le fasse aussi ne sera pas jugé comme "bizarre" et les réponses seront "bien perçues" par l'utilisateur. Pour ce faire, nous avons observé, au sein de notre corpus de réponses, les éléments (les suites de mots) qui sont identiques à ceux de la question. L'annotation du corpus en ce sens et l'observation de la composition des réponses sont présentées dans la section 5.2.

Une autre question est celle de savoir quelle forme pourrait prendre une réponse qui ne serait constituée que d'éléments repris de la question et de l'information-réponse. Nous appelons de telles réponses des *réponses minimales*. La section 5.4 traite de la définition du terme "réponse minimale", la construction de patrons de réponse et leur observation. Nous montrons au cours de cette section que les patrons de réponse ne sont pas les mêmes à l'écrit et à l'oral.

Ceci nous amène à une nouvelle question : celle de savoir si les réponses sont les mêmes à l'oral et à l'écrit. La section 5.5 s'attache à la question de savoir si les réponses sont les mêmes (dans leur contenu et dans leur forme) à l'oral qu'à l'écrit. Par la comparaison des réponses des deux sous-corpus *oral* et *écrit*, nous observons les différences et les points communs entre les réponses du corpus.

Nous présentons une étude menée sur le sous-corpus des réponses orales et qui vise à observer et caractériser l'utilisation des disfluences que sont les hésitations (en l'occurrence le "euh"). Nous montrons dans la section 5.6 que ces hésitations ont un but communicatif particulier et peuvent être judicieusement utilisées par des systèmes de question-réponse à l'oral.

Enfin nous observons d'une part les réponses qui ne proposent pas d'information-réponse et d'autre part les réponses qui contiennent plus qu'une information-réponse. Les premières, que nous appelons *réponses à valence négative*, constituent des exemples de formulation dans le cas où le système ne trouve pas de réponse et sont traitées dans la section 5.7. Les réponses contenant plusieurs information-réponse seront quant à elles étudiées dans la section 5.8.

5.2 Est-il judicieux de réutiliser des éléments de la question ?

Ce travail a été présenté (en français) dans [Garcia-Fernandez et al., 2010a]. Les annotations ont aussi été décrites (en anglais) dans [Garcia-Fernandez et al., 2010b].

Cette section définit ce que nous entendons par *reprise de la question*, présente différents types de reprises observés dans le corpus MACAQ, traite du lien entre reprise de la question et justification dans le domaine des systèmes de question-réponse, propose une série d'observations, analyses et conclusions sur les reprises dans le corpus et la façon dont, à partir de cela, on peut faire des choix de justification au sein d'un système de question-réponse et produire des justifications.

Nous cherchons à observer les cas où une réponse réutilise la question, en partie ou en totalité. Il ne s'agit pas de correction de la question mais d'utilisation de termes issus de la question. Nous observons d'une façon générale (1) le nombre de réponses qui reprennent la question mais aussi (2) ce qu'elles reprennent (l'objet de la question, le verbe principal,...) et (3) la façon dont ces reprises se réalisent.

Les conclusions et applications que nous pourrons éventuellement en tirer sont les suivantes : (1) si il y a un fort taux de reprise, nous pourrons estimer que le fait qu'un système de question-réponse propose des réponses reprennant la question n'est pas "étrange" pour un humain puisqu'il pratique lui aussi cette stratégie ; (2) nous pourrons éventuellement décider de calquer les éléments qu'un système de question-réponse justifie sur les éléments justifiés par nos participants, (3) nous pourrons extraire les reprises, les traiter afin de déterminer les différentes formulations de réponses utilisées et élaborer différents patrons de reprise qui pourront être utilisés pour construire des justifications au sein d'un système de question-réponse.

5.2.1 Identification des types de reprise de la question

Au sein du corpus, on observe que certaines parties des questions sont réutilisées à l'identique ou presque au sein même de la réponse. Nous appelons ceci une *reprise de la question*. Il y a, dans notre corpus, différentes façons de reprendre

la question. Les exemples 5.2 à 5.5 issus de notre corpus présentent certaines des réponses correspondant à la question Q001 :

Exemple 5.1 *[Q001] Combien pèse un bébé à la naissance ?*

Exemple 5.2 *[A577] 3 kilos*

Exemple 5.3 *[A1765] Ça dépend du bébé.*

Exemple 5.4 *[A1932] un bébé pèse environ 3 kilos*

Exemple 5.5 *[A1332] un bébé pèse environ 2 kilos à la naissance*

On observe que la réponse A577 ne reprend pas la question : aucun terme de la question n'apparaît dans la réponse. La réponse A1765 réutilise *l'objet de la question* "bébé". La réponse A1932 réutilise *le verbe* "pèse" et l'objet de la question. Quant à la réponse A1332, elle réutilise ces éléments ainsi que l' *information complémentaire* "à la naissance" [1].

Les exemples 5.7 et 5.8 qui présentent d'autres types de reprise de la question correspondent à certaines des réponses à la question Q122 :

Exemple 5.6 *[Q122] Une feuille A4 mesure combien de centimètres ?*

Exemple 5.7 *[A1554] 21cm par 29.7cm*

Exemple 5.8 *[A1094] 21 29 7*

On observe que la question Q122 contient une indication sur le *type précis de réponse attendu* "centimètres". Le réponse A1554 le réutilise (sous une forme abrégée) tandis que la réponse A1094 (réponse orale) ne le réutilise pas.

Les exemples 5.10 et 5.11 sont des réponses à la question 5.9 :

Exemple 5.9 *[Q212] La Joconde se trouve au Louvre ?*

Exemple 5.10 *[A2389] La Joconde est en effet actuellement exposée au Louvre.*

Exemple 5.11 *[A2971] Oui*

La question Q212 est une question fermée. Pour cette question, nous adoptons le point de vue que l'information "au Louvre", qui doit être validée ou invalidée par le locuteur répondant à la question, est en fait l'information-réponse qui aurait pu être donnée en réponse à une question ouverte portant exactement sur le même objet. On considère donc une question fermée comme une question qui propose une information-réponse. Pour distinguer l'information-réponse proposée

1. Nous rappelons que toutes les questions ne contiennent pas une information complémentaire. Par exemple, la question "Combien pèse un bébé *à la naissance ?* contient une information complémentaire alors que *Combien dure un mois de février ?* n'en contient pas.

dans une question fermée et l'information-réponse de la réponse (telle que nous l'avons défini page 48), nous utiliserons le terme *information-proposée*. Celle-ci peut être réutilisée dans la réponse (c'est le cas de la réponse A2389) ou pas (réponse 2971).

Nous distinguons ainsi les reprises suivantes :

- reprise de *l'objet* de la question (présent dans toutes les questions)
- reprise du *verbe* principal de la question (présent dans toutes les questions)
- reprise des *éléments complémentaires* (éventuels de la question)
- reprise du *type* attendu de la réponse (éventuellement précisé par la question)
- reprise de *l'information-proposée* (dans le cas de questions fermées)

Une réponse peut ne pas contenir de reprise, en contenir une ou en contenir plusieurs.

5.2.2 Repérage des différents éléments dans les questions du corpus

Le corpus de questions a été annoté de façon à identifier les différents éléments qui les composent. L'annotation porte sur les éléments listés précédemment :

- l'objet (exemples 5.12 à 5.15)
- le verbe principal (exemples 5.12 à 5.15)
- les éléments complémentaires (exemple 5.15)
- le type précis de la réponse attendu (exemples 5.13 et 5.14)
- l'information-proposée (exemple 5.15)

Exemple 5.12 *[Q116] Combien <verbe> mesure </verbe> <objet> une feuille A4 </objet> ?*

Exemple 5.13 *[Q122] <objet> Une feuille A4 </objet> <verbe> mesure </verbe> combien de <type> centimètres </type> ?*

Exemple 5.14 *[Q129] Je voudrais savoir quelle <type> taille </type> <verbe> mesure </verbe> <objet> une feuille A4 </objet>.*

Exemple 5.15 *[Q022] Est-ce qu'<objet> un bébé </objet> <verbe> pèse </verbe> <information-proposée> 3,2 </information-proposée> <type> kilos </type> <information-complémentaire> à la naissance </information-complémentaire> ?*

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

Les questions posées sont des questions factuelles simples construites en faisant varier les caractéristiques morphosyntaxiques de questions de base de la forme "<interrogatif> <verbe> <complément d'objet> (<information complémentaire>)". Dans l'étude présentée ici, l'objet de la question est donc systématiquement le complément d'objet.

Une fois ces éléments annotés dans les questions, nous avons annoté manuellement l'ensemble du corpus de réponses.

5.2.3 Annotation des reprises de la question

L'objectif de cette annotation est de repérer, dans les réponses, les éléments qui correspondent à la réutilisation d'un élément présent dans la question correspondante. 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, abrégée, mal accentuée,... (exemples 5.16 à 5.18)).

Exemple 5.16 *[A2496] <objet type="exact">L'amristice de la 1ère Guerre Mondiale</objet> a été signée le 11 novembre 1918*

Exemple 5.17 *[A2607] <objet type="exact">Les Jo d'été</objet> ont lieu en Aout.*

Exemple 5.18 *[A446] <objet type="exact"> Un bebe </objet> mesure environ 50cm a la naissance.*

Les reprises avec modification peuvent être, comme nous le voyons dans la réponse A45 (exemple 5.19), des reprises de l'objet de façon réduite : "une brique" au lieu de "une brique de lait".

Exemple 5.19 *[A45] <objet type="modifié"> Une brique de 1L </objet> doit peser environ 1kg (je ne suis pas sûre).*

Les reprises par pronom ont été annotées spécifiquement :

Exemple 5.20 *[A579] <objet type="pronom"> Elle </objet> pèse 1 kilo*

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 (exemple 5.21), les trois types de reprise de l'objet sont présents simultanément : reprise exacte, modifiée et par pronom.

Exemple 5.21 *[A2280] <objet type="exact"> Une bouteille d'eau </objet> contient du liquide. (...) Si <objet type="modifié"> la bouteille </objet> contient 1 litre, <objet type="pronom"> elle </objet> pèsera un kilo et ainsi de suite.*

Concernant le *type*, les différentes formes d'unité de mesure sont considérées

comme équivalentes et donc comme des reprises exactes :

Exemple 5.22 *[A829] 3.5 <type>kgs</type> environ*

Exemple 5.23 *[A613] 3 <type>kg</type> environ*

Exemple 5.24 *[A711] autour de 3 <type>kilos</type>*

Exemple 5.25 *[A2260] Environ 3 <type>kilogrammes</type>*

On considère qu'il y a reprise du verbe quelle que soit la forme qu'il prend (personne, temps, accompagné d'un verbe modal...) :

Exemple 5.26 *[A579] Elle <verbe>pèse</verbe> 1 kilo*

Exemple 5.27 *[A2279] Si la brique contient 1 litre alors elle <verbe>pèsera</verbe> 1 kilo. Si elle contient 2 libres, elle <verbe>pèsera</verbe> 2 kilos et ainsi de suite.*

Exemple 5.28 *[A1099] une brique de lait <verbe>peut peser</verbe> un kilo [mic] oui*

Comme pour les reprises de tous les autres éléments, la reprise de l'information-proposée et de l'information-complémentaire sont annotées et ce, quelles que soit les erreurs typographiques ou orthographiques :

Exemple 5.29 *[Q184] Je voudrais savoir si un mandat présidentiel dure 5 ans en France.*

Exemple 5.30 *[A16] Oui, un mandat présidentiel dure <information-proposée>5 ans</information-proposée> <information-complémentaire>en France</information-complémentaire>*

5.2.4 Taux de reprise

Le tableau 5.1 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-proposée*. 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. Ainsi, par exemple, le taux de reprise de l'information-proposée sur l'ensemble du corpus équivaut au nombre de réponses reprennant au moins une fois l'information-proposée de la question divisé par le nombre de questions contenant une information-proposée c'est-à-dire divisé par le nombre de questions fermées. Les taux de reprise donnés dans le tableau sont donc indépendants les uns des autres.

Le tableau 5.2 détaille les différents types de reprises de l'objet de la question.

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%**
Aucun élément	**72,61%**	**70,77%**	**73,50%**	69,13%	**79,59%**
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-proposée	**9,58%**	13,26%	7,87%	N.A.	9,58%
Nombre de réponses	3132	1044	2088	2110	1022

TABLE 5.1 – *Taux de réponses reprenant les différents éléments de la question. N.A. est utilisé pour une donnée non disponible.*

Les réponses peuvent contenir plusieurs reprises de l'objet. Nous indiquons dans ce tableau le taux de réponses qui réutilisent au moins une fois de façon exacte, modifié ou par un pronom l'objet de la question. Nous avons ajouté les taux de reprise de l'objet par un pronom lorsqu'aucun autre type de reprise de l'objet n'a lieu dans la réponse considérée. Il s'agit de la ligne du tableau "ne reprennent *que* par un pronom".

	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%**

TABLE 5.2 – *Détails des taux de reprises de l'objet de la question dans les réponses.*

Réponses avec reprise

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. Plus d'une réponse sur cinq (22,54%) reprend au moins une fois l'objet de la question de façon modifiée ou non. Parmi ces reprises (tableau 5.2), 62,48% reprennent l'objet de la question exactement tandis que 23,39% reprennent l'objet de la question uniquement par un pronom. En observant les reprises avec modification, on voit qu'elles consistent en

une réduction de l'objet de la question soit (1) à sa tête de syntagme, les éventuels modificateurs étant omis (exemple 5.31 où "une brique de lait" est repris par "la brique"), soit (2) au terme sémantiquement le plus important (exemple 5.32 où "Le mois de février" est repris par "Février").

Exemple 5.31 *[A2691] Cela dépend de la contenance de <objet type="modifié"> la brique </objet> (...)*

Exemple 5.32 *[A862] <objet type="modifié"> Février </objet> dure en général 28 jours. (...)*

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. Il est à noter que 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%).

Comparaison des réponses à des question ouvertes et fermées

C'est en réponse aux questions fermées qu'il y a le moins de reprise : 20,41% seulement reprennent au moins un élément et 4,31% reprennent tous les éléments (voir tableau 5.1). Notamment, peu de réponses reprennent le type précisé dans la question (5,60% contre 19,52% en ce qui concerne les questions ouvertes). D'après le tableau 5.2, on peut dire que la reprise de l'objet de la question par un pronom uniquement se fait plus souvent en réponse à une question fermée (35,91%) qu'en réponse à une question ouverte (19,15%).
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 17,76 et 10,59 % des réponses.

Comparaison des réponses orales et écrites

Concernant l'opposition oral/écrit, on voit qu'en général il y a plus de reprises à l'oral (29,23% reprennent au moins un élément, 11,69% les reprennent tous) qu'à l'écrit (respectivement 26,50% et 9,02%). 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.
La reprise de l'objet de la question par un pronom uniquement se fait plus souvent à l'oral (25,63%) qu'à l'écrit (18,77%).

Réponses sans reprise

Nous avons quantifié les réponses qui ne reprennent aucun élément de la question. Il s'agit plus de 70% des réponses. Parmi ces réponses, on dénombre les réponses succinctes. Les réponses succinctes sont les réponses qui contiennent uniquement une information-réponse [2]. Ces réponses correspondent a 61,32% des réponses qui ne reprennent pas la question. Si on ignore les réponses succinctes, réponses qui nous intéressent peu dans le cadre de notre étude puisque notre but est justement que les systèmes de question-réponse puissent fournir une réponse plus élaborée que celles qu'ils fournissent actuellement, le taux de réponses du corpus qui reprennent au moins un élément de la question est de 50,96%.

5.2.5 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 d'autant moins si on ignore les réponses succinctes qui sont le type de réposes que fournissent actuellement les système de question-réponse.
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 utilisées (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 leurs plus faibles taux de reprise et notamment l'objet est repris dans 17,8% des réponses mais tous les autres éléments ne sont repris que dans une réponse sur dix.

La plupart des 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 attendu de la réponse, s'il est précisé dans 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 réponse-phrase. 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.

5.3 Quelle information-réponse ?

Ce travail a été présenté (en français) dans [Garcia-Fernandez et al., 2010a]. Les annotations ont aussi été décrites (en anglais) dans [Garcia-Fernandez et al., 2010b].

Donner une réponse à une question va beaucoup plus loin que de fournir en réponse une nouvelle information qui correspond à ce qui est mis en cause par

2. Voir la définition d'une réponse succincte page 49.

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 [3].

Nous avons annoté les informations-réponse dans le corpus de réponse. Face à des réponses spontanées comme celles de notre corpus, la définition de cette *information-réponse* est primordiale.

5.3.1 Définition de l'information-réponse et choix d'annotation

Nous avons défini (chapitre 2, page 48) l'information comme :

Définition 5.1 ***Information-réponse*** *: nouvelle information (par rapport à celles contenues dans la question) qui correspond soit au type attendu de la réponse, soit à un aveu d'incompétence ("je ne sais pas" par exemple).*

Ici, notre but est d'annoter la ou les informations-réponse présentes dans les réponses du corpus. Les informations supplémentaires sur le cadre de validité de la réponse (par exemple "les années bissextiles" dans l'exemple 5.33) ne font donc pas partie de l'information-réponse.

Exemple 5.33 *A60 <information-réponse> 28 jours </information-réponse> les années bissextiles, <information-réponse> 29 jours </information-réponse> les autres*

Dans le cas des réponses à des questions fermées, l'identification de l'information-réponse est plus délicate. En effet, il ne s'agit pas à proprement parler d'une nouvelle information. Dans l'exemple 5.35, c'est "oui" qui correspond à l'information-réponse. Certains cas sont moins évidents. Prenons l'exemple 5.35. On observe que l'élément qui répond à la question est "Oui". C'est un élément qui nécessite d'avoir une réponse à la question pour être formulé. Nous annoterons systématiquement selon ce point de vue. En ce sens, nous verrons que l'information-réponse peut consister en un verbe utilisé à la forme négative ou encore en un adverbe.

Exemple 5.34 *[Q587] L'arrivée des avions à Paris est soit à Orly soit à Roissy ?*

Exemple 5.35 *[A2924] <information-réponse> Oui </information-réponse>, Paris possède deux aéroports civils : à Orly et Roissy*

La valeur des réponses et le fait qu'il s'agisse bel et bien d'une bonne réponse ou pas dépend de la connaissance du monde du participant. Nous n'avons pas pris en compte la valeur de la réponse lors de l'annotation. Nous ne nous attacherons

3. Nous parlons ici de réponses à des questions factuelles uniquement.

pas non plus à la commenter.

Si une réponse contient plusieurs informations-réponse, elles sont toutes annotées.

5.3.2 Observation des informations-réponse

Il y a quatre types de réponses : d'une part les réponses à valence positive à des questions ouvertes et à des questions fermées d'autre part les réponses à valence négative à des questions ouvertes et à des questions fermées.

Réponses à valence négative Que ce soit une réponse à une question ouverte ou fermée, dans le cas des réponses à valence négative (exemple 5.36), 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",...

Exemple 5.36 *[A2480] <information-réponse> je n'en ai aucune idée </information-réponse>*

Réponses à valence positive à des questions ouvertes Les réponses à valence positive aux questions ouvertes et fermées ne prennent pas la même forme dans notre corpus. Pour les réponses à des questions ouvertes, l'information-réponse est une entité nommée : nom de ville, musée, aéroport,... pour les questions de lieu (exemple 5.37) ; date, mois, année... pour les questions de temps (exemple 5.38) ; et valeur accompagnée ou non d'une unité de mesure pour les questions de quantité ((exemple 5.39). Ces informations-réponse peuvent aussi contenir des modifieurs comme on peut le voir dans l'exemple 5.40.

Exemple 5.37 *[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>.*

Exemple 5.38 *[A465] Elle a été signée <information-réponse> le 11 novembre 1918 </information-réponse>*

Exemple 5.39 *[A442] un brique d'un litre pese <information-réponse> environ 1kg </information-réponse>*

Exemple 5.40 *[A1938] un mois de février dure <information-réponse> à peu près 28 jours </information-réponse>*

Réponses à valence positive à des questions fermées Une réponse à valence positive à une question fermée peut être un adverbe ou une expression : “oui”,“en effet”, “non”,... (exemple 5.41).

Exemple 5.41 *[A408] <information-réponse> Tout à fait ! </information-réponse>*

Il peut aussi s’agir d’une reformulation sous forme déclarative de la question (exemple 5.42).

Exemple 5.42 *[A665] <information-réponse> oui </information-réponse>, il <information-réponse> dure effectivement </information-réponse> 5 ans en France*

Plus précisément, on peut distinguer différents cas. On observe des affirmations qui reprennent l’information-réponse proposée dans la question et ainsi la confirment (exemple 5.44).

Exemple 5.43 *[Q612] L’armistice de la 1ère guerre mondiale a eu lieu le 11 novembre 1918 ?*

Exemple 5.44 *[A2499] Elle <information-réponse> a bien été signée </information-réponse> à cette date*

On observe des phrases négatives qui elles aussi reprennent l’information-réponse de la question, infirmant ainsi la question (exemple 5.46).

Exemple 5.45 *[Q512] Le Rhône finit dans le delta de Camargue ?*

Exemple 5.46 *[A2608] Le Rhône <information-réponse> ne fini pas </information-réponse> dans le delta de la Camargue*

On observe aussi des affirmations au sein desquelles l’information-réponse de la question est substituée par une autre information-réponse. C’est la nouvelle information qui constitue alors l’information-réponse.

Exemple 5.47 *[Q061] Le poids d’une bouteille d’eau est de 2 kilos ?*

Exemple 5.48 *[A1353] le poids d’une bouteille d’eau de 1 litre est de <information-réponse> 1 kilo environ </information-réponse> (...)*

On observe et ce quel que soit le type (ouvert ou fermé) de la question, qu’une réponse peut contenir plus d’une information-réponse (exemples 5.49 et 5.50). Nous observons à présent ce point.

Exemple 5.49 *[A3097] <information-réponse>en France</information-réponse>, <information-réponse>à Paris</information-réponse>, <information-réponse>au musée du Louvre</information-réponse>*

Exemple 5.50 *[A1749] <information-réponse>ouais</information-réponse> <information-réponse>c’ est ça</information-réponse> <information-réponse>ouais</information-réponse>*

5.3.3 Nombre d'informations-réponse

Nous avons observé le nombre d'informations-réponse proposées dans les réponses du corpus. Le tableau 5.3 résume cette observation. Hormis les réponses qui ne proposent pas d'information-réponse (2,77% du corpus), près de 80% des réponses proposent une et une seule information-réponse et une réponse sur cinq (18,92%) présente plus d'une information-réponse.

	Tout	Oral	Écrit	Q. ouvertes	Q. fermées
0	2,52%	4,21%	1,68%	2,60%	2,45%
1	78,56%	82,07%	76,80%	78,18%	79,33%
≥ 2	18,92%	13,72%	21,52%	19,22%	18,22%

TABLE 5.3 – *Nombre d'informations-réponse proposées dans les réponses*

Certaines réponses (moins de 3%) ne proposent pas d'information-réponse. Il peut aussi s'agir de corrections de la question ou de suggestion pour pouvoir trouver la réponse autrement/ailleurs (exemples 5.51 et 5.52). D'autres sont des demandes de précision voire même des demandes de reformulation (exemples 5.53 et 5.54).

Exemple 5.51 *[A812] <suggestion> pourquoi ne pas la peser toi même ? </suggestion>*

Exemple 5.52 *[A2849] Je ne suis pas sur, <suggestion> il faut chercher dans un dictionnaire </suggestion>.*

Exemple 5.53 *[A2093] dans quelle unité ? la question est mal posée*

Exemple 5.54 *[A1497] {fw} quelle est la question*

Quoi qu'il en soit la plupart des réponses proposent une et une seule information-réponse (près de 80%). Une étude des 593 réponses qui proposent plus d'une information-réponse est présentée dans la section 5.8.

5.3.4 Temps avant information-réponse

Qu'un système de question-réponse fournisse des énoncés-réponse et non pas une information-réponse seule comporte selon nous différents avantages. Comme nous l'avons déjà vu, il s'agit de naturalité dans l'interaction et de confirmation de la compréhension de la question par le système.

Mais en plus, à l'oral notamment, si on commence par produire des éléments de reprise de la question, le temps de prononciation avant d'exprimer l'information-réponse peut, selon nous, être mis à profit par exemple pour rechercher des informations complémentaires ou affiner la sélection de la réponse.

Nous cherchons ici à quantifier le temps qui pourrait ainsi être mis à profit en produisant des énoncés-réponse. Les réponses ont été découpées afin de n'en garder que la partie précédant la première information-réponse. Les portions de réponse obtenues ont été synthétisées [4] et la durée des fichiers audio produits a été observée. Nous avons réduit le corpus aux réponses dont l'information-réponse n'est pas en première position dans la réponse. Le corpus utilisé pour cette étude correspond donc à 744 réponses, soit 23,75% du corpus total.

Cette méthode permet de prévoir le temps qu'un système de question-réponse peut *mettre à profit* dans le cas où il répondrait des réponses-phrase et non plus des informations-réponse seules.

Le tableau 5.4 présente la moyenne des temps avant l'information-réponse en secondes.

(en secondes)	Oral	Écrit	Toutes
Questions ouvertes	10,74	9,23	8,47
Questions fermées	11,57	8,12	11,28
Toutes les questions	9,64	8,78	9,02

TABLE 5.4 – *Position moyenne en secondes de l'information-réponse. Seules les réponses possédant une information-réponse qui ne soit pas en tout début d'énoncé sont prises en compte.*

Nous observons que le temps calculé est quel que soit le sous corpus considéré supérieur à 8 secondes.

5.3.5 Synthèse

Dans cette section, nous nous sommes intéressée aux informations-réponse proposées par nos participants. On distingue trois types d'informations-réponse. Les réponses à valence positive à des questions ouvertes sont des entités nommées qui correspondent au type attendu par la question. Les réponses à valence positive à des questions fermées peuvent être des phrases affirmatives ou négatives, accompagnées ou non d'un adverbe et reprenant l'information proposée dans la question ou en proposant une nouvelle. Elles peuvent aussi consister en un adverbe modal tel que "oui", "non", "peut-être". Le troisième type d'information-réponse correspond à des réponses à valence négative (quel que soit le type ouvert ou fermé de la question). Il s'agit d'expressions correspondant à un aveu d'incompétence.

D'autre part, nous avons observé que la plupart des réponses proposent une et une seule information-réponse (près de 80% des réponses). Presque 20% en fournissent plusieurs.

4. Nous avons utilisé l'infrastructure du système RITEL, un système de question-réponse en domaine ouvert qui gère aussi les interactions orales [Toney et al., 2008]. Il s'agit de la même synthèse que celle utilisée pour créer les versions audio des questions dans le cadre de l'expérience de collecte du corpus de réponses (voir section 4.3.3 page 87).

Enfin, nous avons calculé le temps avant l'information-réponse dans le cas où les réponses seraient synthétisées oralement. Cette simulation, nous permet de supposer qu'un système de question-réponse oral peut "profiter" d'en moyenne 8 secondes avant de fournir sa réponse à l'utilisateur. Nous voyons deux applications majeures au sein d'un système de question-réponse. Qu'il s'agisse d'une interaction orale ou écrite, détecter le temps ou le nombre de mots avant l'apparition de l'information-réponse peut permettre de proposer une réponse en deux temps. D'abord des éléments reprenant la question sont proposés puis l'information réponse est présentée. Le système peut alors laisser l'opportunité à l'utilisateur d'intervenir entre ces deux phases. De plus, le système peut, au cours de la production et de l'affichage ou synthèse (selon la modalité), continuer à chercher une information-réponse.

Nous avons vu que les réponses du corpus réutilisent des éléments de la question et nous avons observé l'information-réponse proposée dans ces réponses. À présent, nous voulons savoir comment ces éléments s'articulent entre eux. Dans la section suivante, nous définissons ce que nous entendons par "réponse minimale" et montrons des patrons de réponses minimales construits à partir des réponses de notre corpus.

5.4 Comment construire une réponse minimale ?

Ce travail a été présenté (en français) dans [Garcia-Fernandez et al., 2010a]. Les annotations ont aussi été décrites (en anglais) dans [Garcia-Fernandez et al., 2010b].

Nous avons observé dans la section précédente la forme que peut prendre une information-réponse. Nous avons vu dans la section 5.2 comment les réponses reprennent des éléments de la question. À présent, nous proposons d'étudier comment ces informations-réponse et reprises de la question s'articulent entre elles.

5.4.1 Une réponse que tous les systèmes pourraient produire

L'objectif que nous visons est de pouvoir permettre à un système de question-réponse quel qu'il soit de produire des réponses en langue naturelle. Pour cela, nous supposons que tous effectuent une analyse, même minimale, de la question et proposons donc que ces systèmes puissent construire des réponses en langue naturelle basées sur (1) l'information-réponse extraite par le système lui-même (c'est son but) et (2) des éléments réutilisés de la question. Nous définissons de telles réponses comme *minimales* :

Définition 5.2 ***Réponse minimale*** *: réponse composée d'éléments repris de la question, d'une information-réponse et éventuellement de marqueurs dialogiques et pragmatiques.*

L'idée principale est qu'une telle réponse ne contient qu'une unique nouvelle information (par rapport à celles contenues dans la question) : l'information-

réponse. Elle ne consiste donc ni en de la complétion, ni en de la suggestion. Il s'agit d'une réponse qui n'est ni succincte, ni complétive.

Pour repérer les réponses minimales dans notre corpus, nous avons besoin d'avoir une annotation de l'information-réponse, des reprises de la question mais aussi des autres éléments qui composent la question. En particulier, nous avons besoin de savoir si une question contient une *nouvelle information* (par rapport à celles contenues dans la question) autre que l'information-réponse.

5.4.2 Annotation des informations "supplémentaires"

L'annotation a été menée de façon à détecter les informations additionnelles à celles reprises de la question et à l'information-réponse. Nous en avons considéré deux types : les complétions et les suggestions. Les complétions sont des informations supplémentaires qui ne correspondent pas à l'information qui répond à la question (exemple 5.55). Les suggestions, quant à elles, donnent une information sur un moyen autre d'obtenir la réponse (exemple 5.56).

Exemple 5.55 *[A55] <rep>9 mois</rep> si <complétion> tout se passe normalement </completion>*

Exemple 5.56 *[A2849] Je ne suis pas sur, <suggestion>il faut chercher dans un dictionnaire</suggestion>.*

Le tableau 5.5 permet de quantifier les réponses contenant l'un ou l'autre de ces éléments.

	Tout	Oral	Écrit
Toutes les réponses	3132	1044	2088
Nb de réponses avec complétion	278	55	223
Nb de réponses avec suggestion	12	1	11
Total des réponses avec information supplémentaire	288	56	232
Taux de réponses avec information supplémentaire	9,19%	5,36%	10,68%

TABLE 5.5 – *Quantification des réponses complétives et suggestives*

On observe qu'il y a très peu de suggestions dans les réponses du corpus. Nous parlerons donc de façon générale de réponse complétives pour parler des réponses qui contiennent au moins une suggestion ou une complétion.

L'annotation des informations supplémentaires dans les réponses permet d'écarter les réponses complétives et de n'observer que les réponses minimales.

5.4.3 Les réponses minimales dans le corpus

Le tableau 5.6 présente le nombre et le taux de réponses minimales dans l'ensemble du corpus et pour les deux sous-corpus oral et écrit. Ces chiffres ont été calculés à partir d'une version annotée des réponses dans laquelle l'information-réponse et les informations supplémentaires ont été repérées. Les réponses minimales sont celles qui ne contiennent pas d'information supplémentaire (réponses non complétives) et qui ne sont pas composées uniquement d'une information-réponse (réponses non succinctes).

	Tout	Oral	Écrit
Toutes les réponses	3132	1044	2088
Nb de réponses minimales	1472	510	962
Taux de réponses minimales	47,00%	48,85%	46,07%

TABLE 5.6 – *Quantification des réponses minimales*

On observe que près de la moitié des réponses du corpus sont des réponses minimales.

5.4.4 Création de *patron de réponse*

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 5.58 montre en quoi consiste la réduction de la réponse A2212.

Exemple 5.57 *[A2212] Si tout se passe bien, <objet>une grossesse</objet> <verbe>dure</verbe> <information-réponse>aux environ de 9 mois</information-réponse>. (…)*

Exemple 5.58 *[A2212 patron] objet verbe information-réponse*

5.4.5 Observations des *patrons de réponse*

Nous avons observé les différents patrons de réponse du corpus et des sous-corpus oral/écrit, ouvert/fermé uniquement pour les réponses qui ne proposent qu'une unique information-réponse afin d'obtenir un nombre plus restreint de patrons différents. Le tableau 5.7 présente les patrons de réponse les plus fréquents dans l'ensemble du corpus et leurs effectifs dans chacun des sous-corpus de réponses orales, écrites, à des questions ouvertes et à des questions fermées.

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 aussi bien

	Tout	Oral	Écrit	*Q.ouvertes*	*Q.fermées*
Tous les patrons	2617	911	1706	1722	895
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 Qcomplétion	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 info-proposée	7	3	4	0	7
objet info-rep info-proposée	7	2	5	0	7
objet Qcomplétion verbe info-rep	7	3	4	7	0
info-rep info-proposée	6	1	5	0	6
Qcomplétion objet verbe info-rep	6	3	3	6	0
objet objet_pronom verbe info-rep	4	1	3	4	0
info-rep objet info-proposée	3	0	3	0	3

TABLE 5.7 – *Patrons de réponses les plus fréquents. Avec info-rep pour information-réponse, info-proposée pour l'information-proposée et Qcomplétion pour l'information-complémentaire.*

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.

5.4.6 Synthèse

Nous avons observé que les réponses minimales représentent presque la moitié du corpus de réponse. Nous avons créé des patrons de réponse à partir des annotations faites sur les réponses. 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, des signes de ponctuation,... Le corpus est cependant trop diversifié pour permettre une telle observation. En effet, les formes des réponses sont bien trop différentes pour permettre de faire des regroupements. Cependant, ces patrons sont intéressants dans l'optique de donner une réponse 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 l'objet sur lequel elle porte, 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 mis en évidence ici pour

construire une “phrase-réponse” sans se limiter à retourner l'information-réponse seule.

Tout au long de cette section, nous avons mis en valeur les différences entre les réponses obtenues à l'oral et celle obtenues à l'écrit. Une question que l'on peut alors se poser est celle de savoir si les réponses de ces deux sous-corpus sont équivalentes ou non. C'est à cette observation qu'est dédiée la section suivante.

5.5 La modalité influe-t-elle sur la formulation de réponse à générer ?

Nous avons d'ores et déjà comparé sur de nombreux points les réponses orales et écrites. Notamment, au cours de la section 5.2, nous avons noté qu'il y a plus reprises à l'oral qu'à l'écrit et que les réponses écrites reprennent plus souvent l'objet de la question par un pronom que ne le font les réponses orales. En observant les informations-réponse (section 5.3), nous n'avons pas noté de différence entre le nombre d'information-réponse contenues à l'oral et à l'écrit. L'observation qualitative de ces informations-réponse ne montre pas non plus de différences. C'est en revanche la valence de la réponse et le type (ouvert ou fermé) de la question qui jouent un rôle sur la forme que prend l'informattion-réponse. Les observations du temps avant information-réponse, du nombre de réponses minimales et des patrons de réponse n'ont pas non plus mis en valeur de différences entre les deux sous-corpus oral et écrit.

Nous avons en revanche noté qu'il y a plus de réponses complétives à l'écrit (10% des réponses écrites le sont) qu'à l'oral (5%) et qu'il y a plus de reprise de l'objet de la question par un pronom à l'écrit (25% des reprise de l'objet se font par un pronom) qu'à l'oral (19%).

Nous verrons au cours de la section 5.7 que le taux de réponses qui ne contiennent pas d'information-réponse à valence positive (elles seront définies comme les réponses non informatives) est similaire à l'oral et à l'écrit. Dans la section 5.8 consacrée aux réponses contenant plusieurs informations-réponse, nous verrons que les réponses à l'écrit contiennent plus souvent que les réponses orales plus d'une information-réponse.

Nous proposons ici de compléter ses comparaisons par l'étude de deux types d'observations. D'abord nous avons observé d'une façon globale les différences entre les réponses obtenues à l'oral et à l'écrit, sans observer leur structure interne. Nous nous sommes appuyé sur deux indices quantitatifs : le temps que les participants ont mis à répondre (ou temps de réponse) et la taille (en nombre de mots) des réponses. Ensuite, nous avons observé plus précisément le contenu des réponses en terme de lexique et de partie du discours.

5.5.1 Opposition quantitative oral / écrit

Nous avons tout d'abord observé les réponses orales et écrites de façon globale. Les deux indices sur lesquels nous nous basons sont le temps de réponse (en seconde) et la taille des réponses (en nombre de mots).

Concernant les réponses orales, le temps de réponse a été calculé grâce à un module de détection de la parole. Le temps de réponse des réponses orales est donc le temps qu'un participant a mis pour prononcer sa réponse.

Concernant les réponses écrites, le temps de réponse est l'intervalle de temps entre l'affichage de la question sur la page Internet de la plate-forme et le moment où le participant clique sur le bouton "Question suivante".

Le tableau 5.8 présente les caractéristiques de taille et de durée du corpus et des sous-corpus des réponses orales et écrites.

	Écrit	Oral
Nb total de mots	**17395**	**5255**
Nb moyen de mots par réponse	**8,18**	**5,83**
Temps moyen de réponse	33,15	3,62

TABLE 5.8 – *Caractéristiques de taille et de durée du corpus*

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 17 000 pour les réponses écrites contre environ 5 000 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. Ceci peut s'expliquer par le fait que nous avons choisi de conserver le corpus des réponses écrites brut. Ainsi, ce sous-corpus contient éventuellement des fautes d'orthographe, erreurs et coquilles qui élargissent artificiellement la taille du lexique.

Le tableau 5.9 présente le nombre moyen de mots par réponse en distinguant selon la classe de la question (*Temps*, *Lieu* et *Quantité*). On peut faire la même remarque en ce qui concerne la différence écrit *vs* oral que précédemment : la taille 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).

La figure 5.1 montre une représentation en deux dimensions des réponses orales (o) et écrite (w).

On observe que les deux dimensions permettent de différencier quasiment parfaitement les deux sous-corpus (les points bleus "o" et rouges "w" sont très peu mélangés. On observe que les réponses écrites sont plus longues en temps de

	Quantité			Lieu			Temps		
	Écrit	Oral	Total	Écrit	Oral	Total	Écrit	Oral	Total
Nb moyen mot/reponses	9,81	7,08	8,67	**6,81**	**2,70**	6,09	**7,46**	**3,11**	6,69

TABLE 5.9 – *Taille moyenne (en nombre de mots) des réponses considérées selon la classe de la question et la modalité*

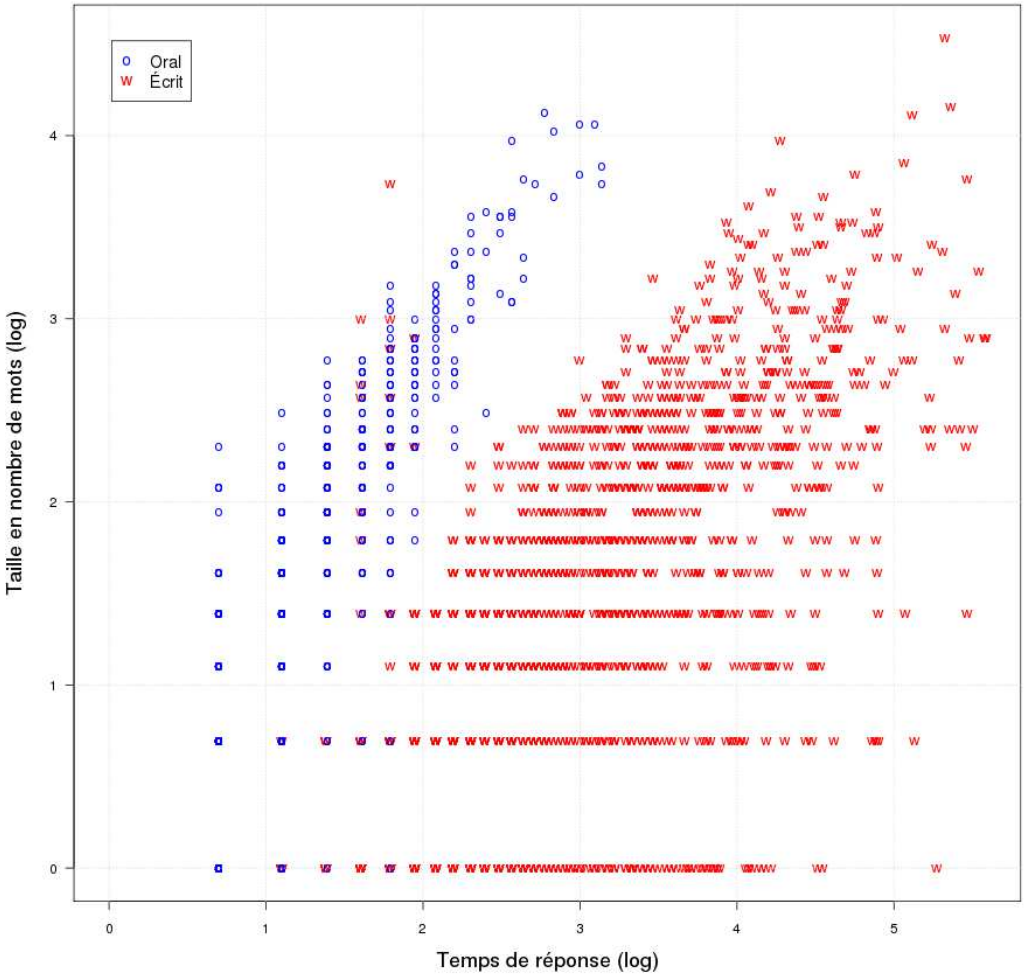

FIGURE 5.1 – *Représentation des réponses orales et écrites en fonction du temps de réponse et de leur taille en nombre de mots.*

réponse que les réponses orales. Nous expliquons ceci par le fait que nous n'avons pas mesuré le même phénomène à l'oral et à l'écrit. En effet, à l'écrit le temps de réponse contient aussi le temps de lecture de la question.

La figure nous permet aussi de voir que les réponses écrites sont plus dispersées que les réponses orales notamment selon la dimension du temps de réponse (le nuage de points rouges "w" est plus étalé horizontalement que le nuage de points

bleus “o”). En revanche, en terme de taille, la dispersion est similaire (les deux nuages s’étalent autant verticalement).

Un test statistique pour vérifier si la différence de distribution en terme de temps de réponse est significative, nous indique que les deux distributions des réponses orales et écrites sont significativement différentes [5] ($p<0.00001$).
De même concernant la taille des réponses : les distributions des tailles des réponses en nombre de mot sont significativement différentes [6]($p<0.0004$).

Nous expliquons la différence de distribution en terme de temps de réponse par le fait que les participants peuvent avoir des niveaux de compétence d’utilisation du clavier très différents (certains l’utilisent certainement tous les jours, d’autres moins souvent) alors que la parole est une compétence acquise *en principe* par tous [7]. La différence concernant la taille des réponses en nombre de mots montre quant à elle que les participants ont produit des réponses plus longues à l’écrit qu’à l’oral mais aussi que les comportement de réponse sont plus diversifiés (certaines réponses sont plus longues que d’autres à l’écrit, alors qu’à l’oral toutes les réponses font *en gros* la même taille).

5.5.2 Opposition lexicale

Nous avons observé les lexiques utilisés dans le corpus. Nous considérons le lexique de l’ensemble des réponses, quelle que soit leur modalité, comme le lexique général (L) et les lexiques des réponses orales d’une part et écrites d’autre part, comme des lexiques par modalité (LM).

Le tableau 5.10 présente des indices généraux sur le lexique de différents sous-corpus du corpus de réponse 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) *Étant donné un lexique, quelle est la part de ce lexique utilisé dans les réponses des différentes classes ?*, (2) *Étant 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 ?*

Ce tableau propose de différencier les réponses en fonction du type sémantique de la question (quantité, lieu, temps) car nous supposons que le type sémantique peut influer sur le vocabulaire utilisé dans les réponses.

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 en comparaison avec les réponses écrites. En particulier, les

5. Nous avons utilisé le test de Kolmogorov-Smirnov avec la variable “temps de réponse” comme facteur et la modalité comme variable nominale. L’hypothèse nulle a été validée avec $p<0.00001$.

6. Même test que précédemment mais avec la variable “taille en nombre de mots” comme facteur. L’hypothèse nulle a été validée avec $p<0.0004$.

7. Nous n’avons fait appel qu’à des adultes locuteurs natifs du français.

	Écrit			Oral			Total		
	Quantité	*Lieu*	*Temps*	*Quantité*	*Lieu*	*Temps*	*Quantité*	*Lieu*	*Temps*
Nb de mots	8 674	5 662	3 059	4 515	476	274	13 189	6 138	3 333
Taille lexique	932	682	518	554	122	97	**1 160**	**708**	**538**
% du lexique L	43,4	31,7	24,1	25,8	**5,7**	**4,5**	54,0	33,0	25,0
% du lexique LM	**60,7**	44,4	**33,7**	**90,6**	20,0	**15,8**	54,0	33,0	25,0

TABLE 5.10 – *Répartition du lexique (L) suivant les classes de questions et les modalités (LM)*

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.

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. En revanche, 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 (exemples 5.60 et 5.61).

Exemple 5.59 *[Q283] Où est le Rhin ?*

Exemple 5.60 *[A2357] Entre la France et l'Allemagne*

Exemple 5.61 *[A2957] En Europe, une partie passe en France, dans le Nord-Est.*

Le tableau 5.11 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 différents 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

	Écrit			Oral			Total		
	Quantité	*Lieu*	*Temps*	*Quantité*	*Lieu*	*Temps*	*Quantité*	*Lieu*	*Temps*
Taille lexique	932	682	518	554	122	97	1160	708	538
Taille du lexique commun	159			45			185		

TABLE 5.11 – *Recouvrement des lexiques suivant les questions et les modalités*

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. On observe aussi que le lexique commun à l’écrit est plus important qu’à l’oral. Ceci peut venir du fait que le lexique à l’écrit est plus important qu’à l’oral.

5.5.3 Opposition en terme de partie du discours

Le tableau 5.12 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. Les chiffres donnés dans ce tableau ont été calculés en comptant le nombre de substantifs, verbe et valeur numériques contenus dans les réponses puis en divisant par le nombre total de réponses du corpus et des sous-corpus considérés.

	Écrit			Oral		
	Lieu	*Quantité*	*Temps*	*Lieu*	*Quantité*	*Temps*
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

TABLE 5.12 – *Moyennes du nombre d’éléments des catégories substantifs, verbes et valeurs numériques dans les réponses*

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 moyen de chiffres 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 5.9). Il semble donc que les réponses à l'oral soient plus centrées vers un transfert précis de l'information (utilisation plus importante de mots pleins) plutôt que vers une formulation détaillée (utilisation de connecteurs, etc.).

5.5.4 Synthèse

Les réponses orales et écrites, comme nous l'avons vu dans les sections précédentes, ne comportent pas de différences majeures. Nous notons cependant qu'il y a plus de réponses complétives à l'écrit (10% des réponses écrites le sont) qu'à l'oral (5%). Dans le cas de la modalité écrite, la réponse reste affichée à l'écran et peut être lue et relue par celui qui a posé la question. On peut alors envisager de donner une quantité importante d'information. À l'oral en revanche donner trop d'information peut se révéler fastidieux. Elles ne seront pas toutes retenues par le destinataire. Une autre différence observée concerne la reprise de l'objet de la question par un pronom. À l'écrit, ce type de reprise est moins utilisé qu'à l'oral. Nous pensons que ceci peut s'expliquer par le fait que répéter l'un des mots-clefs de la question permet, à l'oral, de confirmer ce qui a été compris dans la question. Ce retour est moins nécessaire à l'écrit où, lorsque nos participants ont saisi leur réponses, la question était toujours affichée à l'écran. De façon globale, nous avons aussi observé que les temps de réponses sont plus importants à l'écrit qu'à l'oral et notamment que les temps de réponse à l'écrit sont plus variables d'une réponse à l'autre. Enfin, nous avons observé que le lexique utilisé à l'écrit est plus large que celui utilisé à l'oral. Ceci peut s'expliquer simplement par le fait que le sous-corpus écrit est plus important (deux fois plus de réponses) que le sous-corpus oral.

5.6 EUH... ET SI ON HÉSITE ?

Ce travail présenté (en anglais) dans [Garcia-Fernandez et al., 2010c] a été fait en collaboration avec Ioana Vasilescu et Martine Adda-Decker.

Nous avons effectué une étude de l'hésitation *euh* sur le sous-corpus des réponses orales. Notre idée est d'une part de quantifier ces hésitations, d'autre part d'observer le positionnement de ces hésitations dans les réponses et plus particulièrement le contexte dans lequel elles apparaissent.

5.6.1 Quantification des hésitations dans le corpus

Le tableau suivant montre le nombre de réponses contenant au moins une hésitation *euh*.

Durée totale des réponses	1h10
Nb de réponses	1 044
Nb de mots	6 472
Nb de réponses contenant au moins un "*euh*"	244
Taux de réponses contenant au moins un "*euh*"	23,4%

TABLE 5.13 – *L'hésitation "euh" dans le corpus des réponses orales.*

Plus de 23% des réponses contiennent une telle hésitation. Elles peuvent correspondre soit à un doute de la part du locuteur, soit au besoin de plus de temps pour formuler la réponse [Brennan & Williams, 1995].

5.6.2 Annotation du corpus

Pour observer les contextes dans lesquels apparaissent les hésitations, nous avons utilisé une version annotée du corpus contenant un repérage de l'information-réponse, des éléments repris de la question, des informations additionnelles (complétion et suggestion) mais aussi des marqueurs pragmatiques et dialogiques.

Les reprises des questions décrites dans la section 5.2.2 ont été normalisées par `QUE`. Elles correspondent à la réutilisation d'une information.
L'information-réponse annotée `REP` et les informations additionnelles annotées `ADD` correspondent à de nouvelles informations.
Les marqueurs pragmatiques et dialogiques ont été annotés respectivement `PM` (pragmatic marker) et `DM` (dialogic marker).
Nous rappelons que {fw} (de l'anglais, "filler word") correspond à la transcription de l'hésitation *euh*.

Exemple 5.62 *alors je dirais qu'un bébé mesure {fw} 50 centimètres*

Exemple 5.63 *<DM>alors</DM> <PM>je dirais qu'</PM> <QUE>un bébé </QUE> <QUE> mesure </QUE> euh <REP> 50 centimètres </REP>.*

Exemple 5.64 *{fw} 28 jours ou 29 pendant les années bissextiles*

Exemple 5.65 *euh <REP> 28 jours </REP> <PM> ou </PM> <REP> 29 </REP> <ADD> pendant les années bissextiles </ADD>*

Le tableau suivant quantifie la présence des différents éléments annotés au sein du corpus. Dans celui-ci les occurrences de l'hésitation *euh* ont été considérées en fonction de la position de l'hésitation dans l'énoncé-réponse : en position initiale,

c'est-à-dire en début d'énoncé (INIT) ; en position intermédiaire, c'est-à-dire au milieu de l'énoncé (MID) ; en position finale (END).

Étiquette	#occ.	%
REP	340	27,6
QUE	225	18,2
PM	189	15,2
EUH_INIT	172	13,9
ADD	149	11,8
EUH_MID	94	7,6
DM	68	5,5
EUH_END	4	0,3
Total	1 218	100

TABLE 5.14 – *Distribution des étiquettes d'annotation au sein du corpus de réponses orale*

Nous observons que reprises de la question et information-réponse correspondent à plus de 45,8% des étiquettes attribuées. D'autre part, au niveau des hésitations, nous observons qu'elles apparaissent majoritairement en début (172 occurrences) et milieu (94) d'énoncé et exceptionnellement en fin d'énoncé (4).

5.6.3 Analyse des contextes

Nous avons observé les contexte droit et gauche de l'hésitation *euh*. Le tableau 5.15 présente les bigrammes contenant *euh* en fonction de la position de cette hésitation dans la phrase.

Nous observons que le cas le plus fréquent est la réalisation d'une hésitation juste avant l'information-réponse (REP) et ce quelle que soit la position de *euh* dans la phrase (INIT ou MID).

En observant les contextes gauche et droit de *euh* en position intermédiaire, on note que cette hésitation est majoritairement précédée d'un marqueur pragmatique (PM) ou d'une reprise de la question (QUE) et suivie de l'information-réponse (REP). Ceci correspond à des cas où l'information réponse est introduite soit par une locution dédiée ("la réponse est"), soit par une amorce construite par réutilisation d'éléments de la question ("le poids d'un bébé est"), puis le locuteur hésite (réfléchi ?) à la réponse à fournir avant de donner sa réponse.

5.6.4 Synthèse

Nous avons étudié les hésitations au sein des réponses orales de notre corpus. Nous avons observé que plus d'une réponse sur cinq contient au moins une hésitation : le phénomène est donc loin d'être marginal. De plus, nous avons déterminé en observant les contextes d'occurrence de *euh* que cette hésitation a lieu en général avant l'information-réponse. L'information-réponse est une information nouvelle

	Étiquette	Nb d'occurrences	%
INIT(+1)	REP	126	73,3
	QUE	19	11,0
	PM	17	9,9
	DM	7	4,0
	ADD	2	1,2
	EUH_MID	1	0,6
	TOTAL	172	100
MID(+1)	REP	42	44,7
	PM	20	21,3
	QUE	14	14,9
	ADD	12	12,8
	CDM	5	5,2
	EUH_MID	1	1,1
	TOTAL	94	100
MID(-1)	QUE	34	36,2
	PM	25	26,6
	REP	18	19,1
	DM	7	9,5
	ADD	6	6,4
	EUH_MID	1	1,1
	EUH_INIT	1	1,1
	TOTAL	94	100
END(-1)	REP	2	50,0
	PM	1	25,0
	ADD	1	25,0
	TOTAL	4	100

TABLE 5.15 – *Distribution des bigrammes contenant l'hésitation euh en fonction de sa position dans l'énoncé-réponse.*

cruciale puisqu'il s'agit à proprement de l'information attendue par le locuteur ayant posé la question. Si l'on s'appuie sur les travaux de [Brennan & Williams, 1995], on peut interpréter ces hésitations comme l'expression de la confiance qu'à un locuteur en sa propre réponse. De plus, la présence d'une hésitation joue un rôle sur la perception de la réponse du locuteur ayant posé la question. En effet, des études en psycholinguistiques sur le ressenti du locuteur et de l'auditeur envers le message reçu/émis montrent que la présence de traits para-linguistiques tels que les pauses, l'intonation, les hésitations et les interjections dans les énoncés jouent un rôle convainquant [Smith & Clark, 1993; Brennan & Williams, 1995].

Cette analyse, spécifique au sous-corpus oral, peut trouver son application notamment dans le cas de systèmes de question-réponse établissant un score de con-

fiance. En fonction de ce score, la présence ou l'absence d'hésitation peut être modulée.

Si tous les systèmes de question-réponse ne proposent pas un score de confiance, un cas qu'ils peuvent tous rencontrer est celui de ne pas trouver de réponse à la question. Nous observons à présent les cas de *non-réponse* au sein de notre corpus.

5.7 QUE RÉPONDRE QUAND ON N'A PAS DE RÉPONSE ?

5.7.1 Information-réponses à valence négative et réponses *non informatives*

L'étude présentée ici observe au sein du corpus des cas de "non réponse". Il s'agit des réponses contenant une information-réponse à valence négative (exemple 5.66).

Exemple 5.66 *[A1054] <information-réponse valence="négative">je ne sais pas exactement</information-réponse>*

Une réponse *non informative* est une réponse qui ne contient aucune information-réponse à valence positive. nous précisons "qui ne contient aucun" car une réponse peut contenir plus d'une information-réponse, les valences de ces informations-réponse pouvant différer comme nous pouvons le voir dans l'exemple 5.67.

Exemple 5.67 *[A1011] <information-réponse valence="négative"> je ne sais pas</information-réponse>, <information-réponse valence="positive"> environ 1 kilo </information-réponse> peut-être.*

Le nombre d'information-réponse à valence négative dans notre corpus de réponse est de 113 tandis que le nombre de réponses non informatives est de 95. Le tableau suivant donne le nombre de telles réponses au sein du corpus.

	Tout	Oral	Écrit	Q. ouvertes	Q. fermées
Nb de réponses	3132	1044	2088	2110	1022
Nb de réponses non informatives	95	22	73	61	34
Taux de réponses non informatives	3,03%	2,11%	3,50%	2,90%	3,33%

TABLE 5.16 – *Nombre de réponses non informative.*

On observe qu'il y a très peu de réponses non informatives au sein du corpus et ce quel que soit le sous-corpus (oral, écrit, en réponse à des questions ouvertes ou fermées) considéré.

5.7.2 Observation des réponses à valence négatives

Les réponses à valence négative correspondent à des expressions très figées telles que "je ne sais pas", "je ne connais pas" et ce quel que ce soit le type (ouvert ou fermé) de la question et quel que soit la modalité. On dénombre au total 67 informations-réponse équivalentes à "je ne sais pas" (quelle que soit la casse). Cette expression peut être complétée comme on peut le voir dans les exemples 5.68 à 5.69. Portant ainsi le nombre de réponses à valence négative basée sur l'expression "je ne sais pas" à 95 c'est-à-dire quasiment toutes les réponses à valence négative (84%).

Exemple 5.68 *[A1054] je ne sais pas exactement*

Exemple 5.69 *[A793] J'avoue que je ne sais pas du tout.*

Les autres expressions utilisées sont de la forme "je ne connais pas la réponse" (7 occurrences), "Je ne suis pas sûr" (6), "je n'en ai aucune idée" (1), "je n' en sais pas grand chose" (1), "Je ne comprend pas ta question" (1), "je pourrais pas te répondre" (1), "Joker" (1)

5.7.3 Synthèse

Le nombre d'informations-réponse à valence négative est peu élevé mais, sur les 113 occurrences, nous observons peu de variations dans les expressions utilisées. Ce qu'exprime ces réponses est une incapacité à donner une réponse parce que les locuteurs ne connaissent pas la réponse. Le cas des systèmes de question-réponse est selon nous différent. En effet, dans notre expérience, les participants devaient répondre aux questions sans *a priori* chercher l'information. Les participants expriment ainsi qu'ils *ne savent pas*. En revanche, ce que fait un système de question-réponse est de chercher la réponse dans des documents et non pas dans ses propres connaissances. Nous pensons donc que ces réponses à valence négative ne sont pas toutes adaptées au cas de l'interaction question-réponse au sein de laquelle les systèmes *ne trouvent pas* une information. En effet, il semble étrange qu'un système fournissent un réponse telle que "je ne sais" puisqu'en réalité son travail est de chercher et trouver la réponse. Une réponse possible qui nous semble plus juste sera "je n'ai pas trouvé la réponse". Parmis les expressions utilisées dans le corpus, les expressions "je ne connais pas la réponse", "je ne suis pas sûr", "je ne comprend pas ta question", "je ne pourrais pas te répondre" voire même "joker" sont en revanche, selon nous, utilisables.

5.8 COMMENT RÉPONDRE PLUSIEURS INFORMATIONS-RÉPONSE ?

Comme nous venons de le voir, parmi les réponses du corpus, certaines ne proposent pas d'information-réponse. D'autres en proposent une et une seule et certaines en proposent plus d'une. Nous avons observé plus en détail ces dernières pour identifier la nature des différentes informations-réponse proposées. Notre principale question est de savoir si les informations sont concurrentes ou bien ont des granularités différentes. Nous avons aussi observé l'ordre d'apparition des informations-réponse dans les réponses.

Les réponses concernées correspondent à 18% du corpus des réponses comme nous pouvons le voir dans le tableau 5.17 ci-après.

	Tout	Oral	Écrit
Nb de réponses	3132	1044	2088
Nb de réponses proposant plus d'une information-réponse	592	143	449
Taux de réponses proposant plus d'une information-réponse	18,9%	13,7%	21,5%

TABLE 5.17 – *Nombre de réponses proposant plus d'une information-réponse*

Plus précisément, le graphique 5.2 nous montre que la plupart des réponses contenant plus d'une information-réponse en contiennent deux exactement.

Nous avons observé les suites de deux informations-réponse et nous avons identifié le lien existant entre ces couples d'information-réponse.

5.8.1 Annotation du lien entre deux informations-réponse successives

Nous avons construit des couples (infoRep1, infoRep2) dans lesquels infoRep1 est une information-réponse suivie de l'information-réponse infoRep2. Pour chaque couple, nous avons annoté le lien existant entre les deux informations-réponse. La liste des liens utilisés a été établie à la volée en cours d'annotation et est la suivante :

- Réponses équivalentes (exemples 5.70 et 5.71)
- Affinage de la granularité dans la seconde réponse (exemples 5.72 et 5.73)
- Élargissement de la granularité (exemples 5.74 et 5.75)
- Type différent (exemples 5.76 et 5.77)
- Réponses identiques (exemple 5.78)

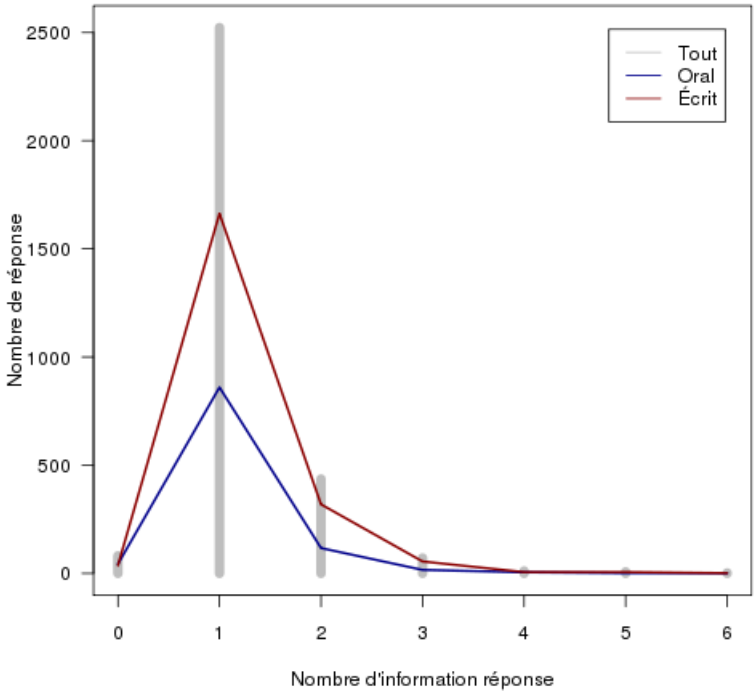

FIGURE 5.2 – *Histogramme et courbes de fréquence du nombre d'informations-réponse dans les réponses du corpus.*

Exemple 5.70 *[A2705] Le mois de février dure <information-réponse> 28 jours </information-réponse> les années non bissextiles et <information-réponse> 29 jours </information-réponse> les années bissextiles.*

Exemple 5.71 *[A512] <information-réponse> Roissy </information-réponse> ou <information-réponse> Orly </information-réponse>*

Exemple 5.72 *[A3089] <information-réponse> en France </information-réponse>, <information-réponse> à Paris </information-réponse>, <information-réponse> au musée du Louvre </information-réponse>*

Exemple 5.73 *[A3117] <information-réponse> 1918 </information-réponse>, <information-réponse> le 11 novembre </information-réponse> pour être précis.*

Exemple 5.74 *[A2474] <information-réponse> Sur les Champs Elysées</information-réponse> <information-réponse>à Paris</information-réponse>.*

Exemple 5.75 *[A394] Je ne suis pas à Paris depuis peu... Mais la Seine passe également dans d'autres villes et on peut la traverser <information-réponse> sur le*

pont Corneille</information-réponse> par exemple <information-réponse>dans la ville de Rouen</information-réponse>.

Exemple 5.76 *[A3096] Ils ont eu lieu <information-réponse> cet été </information-réponse> donc <information-réponse> en 2008 </information-réponse>, puis <information-réponse> tous les 4 ans </information-réponse> : <information-réponse> 2012, 2016, 2020 ... </information-réponse>*

Exemple 5.77 *[A1542] ça <information-réponse> peut </information-réponse> arriver <information-réponse> oui </information-réponse>*

Exemple 5.78 *[A1177] {fw} <information-réponse>oui</information-réponse> depuis 2002 <information-réponse>oui</information-réponse>*

5.8.2 Observation et quantification des liens entre deux information-réponse

Le tableau 5.18 montre la distribution de ces différents cas.

Type de lien	# occurrences	Taux
Réponses équivalentes	298	55,2%
Changement de type	110	20,4%
Affinage de la granularité	71	13,1%
Élargissement de la granularité	54	10,0%
Réponses identiques	7	1,3%

TABLE 5.18 – *Distribution des annotations du lien entre couple d'information-réponse*

On observe que dans la plupart des cas (55%), la deuxième information-réponse correspond à une autre proposition de réponse sans aucun changement ni de la granularité, ni du type de la réponse. En général, les informations-réponse sont alors accompagnées d'une information-complémentaire indiquant une condition de validité de la réponse. Par exemple, dans la réponse A2705 (exemple 5.70), on remarque que la première information-réponse “28 jours” est accompagnée de l'information complémentaire “les années non bissextiles” qui indique le cadre dans lequel l'information-réponse est valide. La seconde information-réponse (“29 jours”) est quant à elle accompagnée d'une autre information complémentaire (“les années bissextiles”) indiquant un cadre de validité différent. L'exemple 5.71 est une réponse qui montre deux informations-réponse concurrentes, sans préciser de cadre de validité. Ceci est dû à la nature même de la réponse. En effet, il n'y a pas une unique réponse à la question de savoir où les avions atterrissent à Paris puisque la ville dispose de plusieurs aéroports.

On observe aussi que plus de 40% des liens entre informations-réponse correspondent à une modification du type ou de la granularité. Une modification de

granularité peut autant être un élargissement (10%) qu'un affinage (13%). On observe quelques rares cas de répétition à l'identique d'une information-réponse.

5.8.3 Synthèse

L'observation des réponses contenant plusieurs informations-réponse permet une analyse du lien entre les différentes informations-réponse proposées. Au sein de notre corpus, nous observons deux cas principaux de formulation de plusieurs informations-réponse. D'une part, le participant peut donner plusieurs réponses à la question, d'autre part, il peut donner des réponses de type ou de granularité différente. Ces réponses multiples peuvent s'expliquer par la nature même de la réponse à la question (il n'y a pas qu'une seule réponse à la question, le participant propose donc plusieurs réponses), par un manque de précision de la question (par exemple, le type de réponse attendue n'est pas précisé) ou encore par une volonté de compléter la réponse par des informations supplémentaires (le participant fait alors preuve de coopération).

CONCLUSION DU CHAPITRE

À partir de sept questions générales, nous avons effectué différentes analyses du corpus.
L'observation de la réutilisation d'éléments de la question au sein de la réponse, qui a impliqué une annotation du corpus des questions et du corpus des réponses, permet d'identifier les différents éléments repris, la façon dont ils sont repris et en quelle proportion.
La définition et l'observation de l'information-réponse dans le corpus des réponses, a impliqué une annotation de ce corpus, permet de qualifier et de quantifier cet élément dans le corpus.
La création de patrons de réponses minimales, c'est-à-dire des réponses réduites aux éléments de reprise de la question et à l'information-réponse permet d'observer l'ordre et l'enchaînement de ces éléments.
La comparaison des réponses selon la modalité (orale ou écrite) permet de déterminer si les réponses sont les mêmes à l'oral et à l'écrit.
Le cas des hésitations à l'oral a lui aussi été étudié. Il met en valeur la présence d'hésistation juste avant l'information-réponse.
Une observation des réponses non informatives, c'est-à-dire qui ne proposent pas d'information-réponse ou bien qui ne contiennent pas d'information-réponse à valence positive, permet de lister des exemples de réponses qu'un système pourrait réutiliser dans le cas où il ne trouve pas de réponse à la question.
Enfin, une étude des réponses contenant plusieurs informations-réponse permet d'identifier les différents liens existant entre celles-ci.

CONCLUSION GÉNÉRALE

BILAN

Les travaux présentés dans ce mémoire se situent dans le contexte de la réponse à une question. Contrairement à de nombreux travaux traitant de la recherche de l'information à fournir en réponse à une question, nous avons considéré la question-réponse comme une interaction. Notre problématique principale a été de caractériser la forme que peut prendre une réponse en interaction avec la question qui puisse être produite par des systèmes de question-réponse.

Le chapitre 1 de ce mémoire a exposé les enjeux de l'interaction en langue naturelle, puis, plus précisement, de l'interaction du type "réponse à une question" et ce, considérant deux modalités d'interaction : l'oral et l'écrit.

Nous avons ensuite étudié dans le chapitre 2, d'une part des réponses fournies par des utilisateurs sur des sites collaboratifs de question-réponse et par deux systèmes de question-réponse développés au LIMSI (les systèmes FIDJI et RITEL), et d'autre part des théories de l'interaction et plus précisement de la réponse en interaction. Ceci nous a permis de redéfinir le terme *réponse* et de définir l'*information-réponse*. Nous avons aisni pu mettre en valeur que les réponses que fournissent les systèmes ne sont pas des réponses en interaction avec la question et que les réponses données par des humains sur des sites de question-réponses ne sont pas reproductibles par les systèmes. Pour observer et analyser notre objet d'étude, la réponse en interaction avec la question, nous avons ainsi choisi de constituer un corpus de telles réponses. Une dernière constation faite au cours de ce chapitre a été celle du lien entre la forme de la question et celle de la réponse, mis en évidence par le modèle de la GRammaire INTeractive.

Le chapitre 3 a consisté en l'étude des variations possibles d'une question. Ce travail découle de la dernière constation faite dans le chapitre 2 et prépare la constitution du corpus présentée dans le chapitre 4. En effet, si nous voulons collecter un corpus de réponses à des questions, la forme des questions utilisées lors de la collecte est primordiale. Cet état de l'art a permis de mettre en valeur des paramètres de variation de la question en fonction de caractéristiques de la question elle-même (son type sémantique, le type de son verbe principal, le type de l'interrogatif utilisé ou type paradigmatique, son type syntaxique ou type syntagmatique, son type de granularité et son thème) et de la réponse qu'attend la question (le type attendu de la réponse, sa nature et sa valence).

La collecte du corpus de réponses, présentée dans le chapitre 4, a été organisée

à l'écrit et à l'oral, auprès de plus de 150 locuteurs natifs du français. Les questions soumises aux participants ont été construites à partir des paramètres mis en valeur dans le chapitre 3. Les objectifs fixés lors de la collecte du corpus ont été explicités. Le protocole mis en œuvre, la construction du corpus de question et le déroulement des passations ont été décrits. Une évaluation du protocole a été faite concernant les points suivants :

- la spontanéité des réponses : l'observation d'indices tels que la durée maximale de production d'une réponse et la présence d'hésitations et d'expressions de doute sont en faveur de la spontanéité des réponses ;
- la facilité de donner une réponse aux questions : plus de 94% des réponses contiennent une information-réponse ;
- l'obtention de réponses non succinctes : une réponse succincte est une réponse qui est constituée uniquement d'une information-réponse. Nous avons constaté que 45% des réponses sont succinctes et qu'ajouter une contrainte aux participants aurait pu éviter ce résultat ;
- l'obtention de réponses sans information complémentaire : une information complémentaire a été définie comme une information nouvelle (par rapport à celles contenues dans la question) qui n'est pas une information-réponse (c'est-à-dire qui n'est pas l'information attendue par la question). Parce que les systèmes ne sont pas forcément capables de gérer des informations complémentaires et parce que nous voulions collecter un corpus de réponses reproductibles par des systèmes, il était intéressant de collecter un corpus qui contienne peu ou pas d'information complémentaire. Nous avons constaté qu'en effet, seules 10% des réponses du corpus contiennent de telles informations.
- la possibilité de comparer des réponses orales et écrites : les corpus de réponses orales et écrites, bien que de tailles différentes, ont été collectés par deux versions très similaires d'un même protocole.
- la collecte de plusieurs réponses pour une même question : avec au minimum 132 réponses par *groupe de questions*, c'est-à-dire par ensemble de variations linguistiques autour d'une même question de base, des comparaisons entre les réponses à une même question sont possibles.
- une expérience *agréable* pour les participants malgré la redondance des questions qui leur sont posées : l'observation des remarques laissées par les participants en fin de passation laisse à penser que l'expérience a été appréciée par les participants.

Le chapitre 5 a présenté une analyse du corpus collecté. Il a cherché à répondre à un ensemble de questions à se poser pour mettre en place un module de génération de réponses en langue naturelle dans un système de question-réponse.

Est-il judicieux de réutiliser des éléments de la question ? Nous avons observé que le fait de réutiliser des éléments de la question n'est absolument pas marginal au sein du corpus. L'objet de la question , son verbe principal et son type sont les

éléments les plus repris. Les réponses aux questions fermées se distinguent par rapport aux réponses aux questions ouvertes par leurs plus faibles taux de reprise. Une réponse possible peut être composée des trois éléments *objet de la question*, *verbe principal* et *type* 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 réponse.

Quelle information faut-il répondre ? Nous avons distingué trois types d'information-réponse au sein du corpus. Les réponses à valence positive à des questions ouvertes sont des entités nommées qui correspondent au type attendu par la question. Les réponses à valence positive à des questions fermées peuvent être des phrases affirmatives ou négatives, accompagnées ou non d'un adverbe et reprenant l'information proposée dans la question ou en proposant une nouvelle. Elles peuvent aussi consister en un adverbe modal tel que "oui", "non", "peut-être". Le troisième type d'information-réponse correspond à des réponses à valence négative (quel que soit le type ouvert ou fermé de la question). Il s'agit d'expressions correspondant à un aveu d'incompétence. La plupart des réponses proposent une et une seule information-réponse. L'ordre des éléments constituant les réponses du corpus est tel que d'abord des éléments reprenant la question sont proposés puis l'information réponse est présentée. Nous avons observé que le temps avant que l'information-réponse soit donnée est d'environ 8 secondes.

Comment construire une réponse minimale ? Nous avons construit des patrons de réponses minimales qui consistent en une réduction de la réponse aux éléments reprennant la question et à l'information-réponse. Ils donnent une idée sur l'ordre d'apparition des éléments de reprise de la question et de l'information-réponse, mais sont peu représentatifs de la forme syntaxique de la réponse qui comporte des marques d'hésitation, des connecteurs, des signes de ponctuation,... Un système de question-réponse peut s'inspirer des différents patrons mis en évidence pour construire une réponse sans se limiter à retourner l'information-réponse seule.

La modalité (oral/écrit) influe-t-elle sur la formulation de réponse à générer ? Nous avons comparé les réponses orales et écrites du corpus et avons montré qu'il y a plus de réponses complétives à l'écrit qu'à l'oral et plus de reprises de l'objet de la question par pronom à l'écrit qu'à l'oral.

Euh... et si on hésite ? Nous avons montré que plus d'une réponse sur cinq du corpus contient au moins une hésitation et que cette hésitation a lieu en général juste avant l'information-réponse. Nous avons vu que la présence d'une hésitation jouent un rôle sur la perception de la réponse du locuteur ayant posé la question. Cette analyse, spécifique au sous-corpus oral, peut trouver son application notamment dans le cas de systèmes de question-réponse établissant un score de confiance. En fonction de ce score, la présence ou l'absence d'hésitation peut être modulée.

Que répondre quand on n'a pas de réponse ? Il y a, au sein du corpus, peu de cas de non réponse à la question. Les réponses qui ne contiennent aucune information-réponse à valence positive ont été définies comme des réponses non informatives. On observe que celles-ci expriment une incapacité à répondre à la question, le plus

souvent due à un manque de connaissance. Il y a peu de variations dans les formes linguistiques utilisées qui correspondent pour la plupart à des expressions figées telles que “je ne sais pas”.

Comment répondre plusieurs informations-réponse ? Nous avons analysé le lien entre les différentes informations-réponse fournies au sein d’une même réponse. Deux cas principaux de formulation de plusieurs informations-réponse ont été identifiés : l’expression d’informations-réponse différentes (par exemple, parce qu’il n’existe pas une unique réponse à la question) et l’expression d’informations-réponse de granularité différentes.

PERSPECTIVES

Les perspectives de notre travail concernent d’abord la poursuite de la collecte d’un corpus de réponse. Nous souhaitons compléter le corpus des questions en y incluant des questions relevant d’autres types sémantiques. Une nouvelle phase de collecte devra cependant ajouter une contrainte aux participants leur précisant de “ne pas répondre juste la réponse” afin de limiter le nombre de réponses succinctes. La formulation de cette contrainte reste encore à définir mais ne doit, à notre avis, pas consister en un exemple sur lequel les participants pourraient calquer leur réponse.

En ce qui concerne les analyses du corpus, nous envisageons de travailler sur des patrons de réponses moins restrictifs que ceux présentés dans ce document. Nous voudrions qu’ils puissent représenter aussi la forme syntaxique de la réponse, même de façon très schématique.

En s’inspirant des résultats obtenus par l’analyse du corpus, nous souhaiterions implémenter un module de génération de réponse en interaction. Les deux systèmes disponibles au LIMSI, FIDJI et RITEL, effectuent une analyse linguistique de la question sur laquelle nous pouvons nous appuyer pour construire des réponses reprennant la question.

Reste bien évidemment la question de l’évaluation de la production de réponses en interaction. Nous savons combien l’évaluation d’un système est importante, notamment pour l’améliorer et le comparer à d’autres systèmes similaires, et une réflexion sur des méthodes et des mesures d’évaluation nous semble primordiale. Ceci pourrait donner lieu à une nouvelle tâche au sein de campagnes d’évaluation existantes. Une piste de réflexion est selon nous d’observer les travaux autour de l’évaluation du dialogue homme-machine. Une première évaluation pourrait consister en une évaluation de la satisfaction des utilisateurs d’un système de question-réponse fournissant des réponses en interaction. Il pourrait s’agir par exemple d’implémenter deux versions d’un même système : l’une fournissant des réponses succinctes, l’autre fournissant des réponses en interaction. Puis nous proposerions à des participants d’utiliser un système de question-réponse, certains utilisateurs utilisant la première version du système, d’autre utilisant la seconde version et de leur faire remplir un questionnaire de satisfaction. Une autre façon de

faire peut être d'utiliser un questionnaire dans lequel pour une question donnée, le participant doit choisir la réponse qu'il *préfère*.

ANNEXES

A

SOMMAIRE

A.1 LISTE DES QUESTIONS BRUTES

TABLE A.1 – *Liste des questions posées aux participants*

Identifiant	Question
Q001	Combien pèse un bébé à la naissance ?
Q002	Un bébé pèse combien à la naissance ?
Q003	Combien est-ce que pèse un bébé à la naissance ?
Q004	Je voudrais savoir combien un bébé pèse à la naissance.
Q005	Un bébé pèse environ 3,2 kilos à la naissance ?
Q006	Combien de kilos pèse un bébé à la naissance ?
Q007	Un bébé pèse combien de kilos à la naissance ?
Q008	Combien de kilos est-ce que pèse un bébé à la naissance ?
Q009	Je voudrais savoir combien de kilos pèse un bébé à la naissance.
Q010	Un bébé pèse environ 3,2 en kilos à sa naissance ?
Q011	Quel poids pèse un bébé à la naissance ?
Q012	Un bébé pèse quel poids à la naissance ?
Q013	Quel poids est-ce que pèse un bébé à la naissance ?
Q014	Je voudrais savoir quel poids pèse un bébé à la naissance.
Q015	Le poids d'un bébé à sa naissance est d'environ 3,2 kilos ?
Q016	Que pèse un bébé à la naissance ?
Q017	Un bébé pèse quoi à la naissance ?
Q018	Qu'est-ce que pèse un bébé à la naissance ?
Q019	Je voudrais savoir qu'est-ce qu'un bébé pèse à la naissance.
Q020	Un bébé pèse quelque chose à sa naissance ?
Q021	Un bébé pèse-t-il environ 3,2 kilos à la naissance ?
Q022	Est-ce qu'un bébé pèse 3,2 kilos à la naissance ?
Q023	Je voudrais savoir si un bébé pèse 3,2 kilos à la naissance.
Q024	Combien pèse une brique de lait ?
Q025	Une brique de lait pèse combien ?
Q026	Combien est-ce que pèse une brique de lait ?
Q027	Je voudrais savoir combien une brique de lait pèse.
Q028	Une brique de lait pèse 1 kilo ?
Q029	Combien de kilos pèse une brique de lait ?
Q030	Une brique de lait pèse combien de kilos ?
Q031	Combien de kilos est-ce que pèse une brique de lait ?
Q032	Je voudrais savoir combien de kilos pèse une brique de lait.
Q033	Une brique de lait pèse 1 en kilo ?
Q034	Quel poids pèse une brique de lait ?
Q035	Une brique de lait pèse quel poids ?

Suite page suivante

Identifiant	Question
Q036	Quel poids est-ce que pèse une brique de lait ?
Q037	Je voudrais savoir quel poids pèse une brique de lait.
Q038	Le poids d'une brique de lait est de 1 kilo ?
Q039	Que pèse une brique de lait ?
Q040	Une brique de lait pèse quoi ?
Q041	Qu'est-ce que pèse une brique de lait ?
Q042	Je voudrais savoir qu'est-ce qu'une brique de lait pèse.
Q043	Une brique de lait pèse quelque chose ?
Q044	Une brique de lait pèse-t-elle 1 kilo ?
Q045	Est-ce qu'une brique de lait pèse 1 kilo ?
Q046	Je voudrais savoir si une brique de lait pèse 1 kilo.
Q047	Combien pèse une bouteille d'eau ?
Q048	Une bouteille d'eau pèse combien ?
Q049	Combien est-ce que pèse une bouteille d'eau ?
Q050	Je voudrais savoir combien une bouteille d'eau pèse.
Q051	Une bouteille d'eau pèse 2 kilos ?
Q052	Combien de kilos pèse une bouteille d'eau ?
Q053	Une bouteille d'eau pèse combien de kilos ?
Q054	Combien de kilos est-ce que pèse une bouteille d'eau ?
Q055	Je voudrais savoir combien de kilos pèse une bouteille d'eau.
Q056	Une bouteille d'eau pèse 2 en kilos ?
Q057	Quel poids pèse une bouteille d'eau ?
Q058	Une bouteille d'eau pèse quel poids ?
Q059	Quel poids est-ce que pèse une bouteille d'eau ?
Q060	Je voudrais savoir quel poids pèse une bouteille d'eau.
Q061	Le poids d'une bouteille d'eau est de 2 kilos ?
Q062	Que pèse une bouteille d'eau ?
Q063	Une bouteille d'eau pèse quoi ?
Q064	Qu'est-ce que pèse une bouteille d'eau ?
Q065	Je voudrais savoir qu'est-ce qu'une bouteille d'eau pèse.
Q066	Une bouteille d'eau pèse quelque chose ?
Q067	Une bouteille d'eau pèse-t-elle 2 kilos ?
Q068	Est-ce qu'une bouteille d'eau pèse 2 kilos ?
Q069	Je voudrais savoir si une bouteille d'eau pèse 2 kilos.
Q070	Combien mesure un bébé à la naissance ?
Q071	Un bébé mesure combien à la naissance ?
Q072	Combien est-ce que mesure un bébé à la naissance ?
Q073	Je voudrais savoir combien un bébé mesure à la naissance.

Suite page suivante

Identifiant	Question
Q074	Un bébé mesure environ 50 centimètres à la naissance ?
Q075	Combien de centimètres mesure un bébé à la naissance ?
Q076	Un bébé mesure combien de centimètres à la naissance ?
Q077	Combien de centimètres est-ce que mesure un bébé à la naissance ?
Q078	Je voudrais savoir combien de centimètres mesure un bébé à la naissance.
Q079	Un bébé mesure environ 50 en centimètres à la naissance ?
Q080	Quelle taille mesure un bébé à la naissance ?
Q081	Un bébé mesure quelle taille à la naissance ?
Q082	Quelle taille est-ce que mesure un bébé à la naissance ?
Q083	Je voudrais savoir quelle taille mesure un bébé à la naissance.
Q084	La taille d'un bébé est d'environ 50 centimètres à la naissance ?
Q085	Que mesure un bébé à la naissance ?
Q086	Un bébé mesure quoi à la naissance ?
Q087	Qu'est-ce que mesure un bébé à la naissance ?
Q088	Je voudrais savoir qu'est-ce qu'un bébé mesure à la naissance.
Q089	Un bébé mesure quelque chose à la naissance ?
Q090	Un bébé mesure-t-il environ 50 centimètres à la naissance ?
Q091	Est-ce qu'un bébé mesure environ 50 centimètres à la naissance ?
Q092	Je voudrais savoir si un bébé mesure environ 50 centimètres à la naissance.
Q093	Combien mesure un double décimètre ?
Q094	Un double décimètre mesure combien ?
Q095	Combien est-ce que mesure un double décimètre ?
Q096	Je voudrais savoir combien un double décimètre mesure.
Q097	Un double décimètre mesure 20 centimètres ?
Q098	Combien de centimètres mesure un double décimètre ?
Q099	Un double décimètre mesure combien de centimètres ?
Q100	Combien de centimètres est-ce que mesure un double décimètre ?
Q101	Je voudrais savoir combien de centimètres mesure un double décimètre.
Q102	Un double décimètre mesure 20 en centimètres ?
Q103	Quelle taille mesure un double décimètre ?
Q104	Un double décimètre mesure quelle taille ?
Q105	Quelle taille est-ce que mesure un double décimètre ?
Q106	Je voudrais savoir quelle taille mesure un double décimètre.
Q107	La taille d'un double décimètre est de 20 centimètres ?
Q108	Que mesure un double décimètre ?
Q109	Un double décimètre mesure quoi ?
Q110	Qu'est-ce que mesure un double décimètre ?
Q111	Je voudrais savoir qu'est-ce qu'un double décimètre mesure.

Suite page suivante

Identifiant	Question
Q112	Un double décimètre mesure quelque chose ?
Q113	Un double décimètre mesure-t-il 20 centimètres ?
Q114	Est-ce qu'un double décimètre mesure 20 centimètres ?
Q115	Je voudrais savoir si un double décimètre mesure 20 centimètres.
Q116	Combien mesure une feuille A4 ?
Q117	Une feuille A4 mesure combien ?
Q118	Combien est-ce que mesure une feuille A4 ?
Q119	Je voudrais savoir combien une feuille A4 mesure.
Q120	Une feuille A4 mesure 21 centimètres de largeur et 29,7 centimètres de longueur ?
Q121	Combien de centimètres mesure une feuille A4 ?
Q122	Une feuille A4 mesure combien de centimètres ?
Q123	Combien de centimètres est-ce que mesure une feuille A4 ?
Q124	Je voudrais savoir combien de centimètres mesure une feuille A4.
Q125	Une feuille A4 mesure 21 de largeur par 29,7 de longueur en centimètres ?
Q126	Quelle taille mesure une feuille A4 ?
Q127	Une feuille A4 mesure quelle taille ?
Q128	Quelle taille est-ce que mesure une feuille A4 ?
Q129	Je voudrais savoir quelle taille mesure une feuille A4.
Q130	La taille d'une feuille A4 est de 21 centimètres de largeur par 29,7 centimètres de longueur ?
Q131	Que mesure une feuille A4 ?
Q132	Une feuille A4 mesure quoi ?
Q133	Qu'est-ce que mesure une feuille A4 ?
Q134	Je voudrais savoir qu'est-ce qu'une feuille A4 mesure.
Q135	Une feuille A4 mesure quelque chose ?
Q136	Une feuille A4 mesure-t-elle 21 centimètres de largeur par 29,7 centimètres de longueur ?
Q137	Est-ce qu'une feuille A4 mesure 21 centimètres de largeur par 29,7 centimètres de longueur ?
Q138	Je voudrais savoir si une feuille A4 mesure 21 centimètres de largeur par 29,7 centimètres de longueur.
Q139	Combien dure une grossesse ?
Q140	Une grossesse dure combien ?
Q141	Combien est-ce que dure une grossesse ?
Q142	Je voudrais savoir combien une grossesse dure.
Q143	Une grossesse dure environ 9 mois ?
Q144	Combien de mois dure une grossesse ?
Q145	Une grossesse dure combien de mois ?
Q146	Combien de mois est-ce que dure une grossesse ?
Q147	Je voudrais savoir combien de mois dure une grossesse.

Suite page suivante

Identifiant	Question
Q148	Une grossesse dure 9 en mois ?
Q149	Quel temps dure une grossesse ?
Q150	Une grossesse dure quel temps ?
Q151	Quel temps est-ce que dure une grossesse ?
Q152	Je voudrais savoir quel temps dure une grossesse.
Q153	La durée d'une grossesse est d'environ 9 mois ?
Q154	Que dure une grossesse ?
Q155	Une grossesse dure quoi ?
Q156	Qu'est-ce que dure une grossesse ?
Q157	Je voudrais savoir qu'est-ce qu'une grossesse dure.
Q158	Une grossesse dure quelque chose ?
Q159	Une grossesse dure-t-elle environ 9 mois ?
Q160	Est-ce qu'une grossesse dure environ 9 mois ?
Q161	Je voudrais savoir si une grossesse dure environ 9 mois.
Q162	Combien dure un mandat présidentiel en France ?
Q163	Un mandat présidentiel dure combien en France ?
Q164	Combien est-ce que dure un mandat présidentiel en France ?
Q165	Je voudrais savoir combien un mandat présidentiel dure en France.
Q166	Un mandat présidentiel dure 5 ans en France ?
Q167	Combien d'année dure un mandat présidentiel en France ?
Q168	Un mandat présidentiel dure combien d'années en France ?
Q169	Combien d'année est-ce que dure un mandat présidentiel en France ?
Q170	Je voudrais savoir combien d'années dure un mandat présidentiel en France.
Q171	Un mandat présidentiel dure 5 en années en France ?
Q172	Quel temps dure un mandat présidentiel en France ?
Q173	Un mandat présidentiel dure quel temps en France ?
Q174	Quel temps est-ce que dure un mandat présidentiel en France ?
Q175	Je voudrais savoir quel temps dure un mandat présidentiel en France.
Q176	La durée d'un mandat présidentiel est de 5 ans en France ?
Q177	Que dure un mandat présidentiel en France ?
Q178	Un mandat présidentiel dure quoi en France ?
Q179	Qu'est-ce que dure un mandat présidentiel en France ?
Q180	Je voudrais savoir qu'est-ce qu'un mandat présidentiel dure en France.
Q181	Un mandat présidentiel dure quelque chose en France ?
Q182	Un mandat présidentiel dure-t-il 5 ans en France ?
Q183	Est-ce que un mandat présidentiel dure 5 ans en France ?
Q184	Je voudrais savoir si un mandat présidentiel dure 5 ans en France.
Q185	Combien dure un mois de février ?

Suite page suivante

Identifiant	Question
Q186	Un mois de février dure combien ?
Q187	Combien est-ce que dure un mois de février ?
Q188	Je voudrais savoir combien un mois de février dure.
Q189	Un mois de février dure 28 ou 29 jours selon les années ?
Q190	Combien de jours dure un mois de février ?
Q191	Un mois de février dure combien de jours ?
Q192	Combien de jours est-ce que dure un mois de février ?
Q193	Je voudrais savoir combien de jours dure un mois de février
Q194	Un mois de février dure 28 ou 29 en jours selon les années ?
Q195	Quel temps dure un mois de février ?
Q196	Un mois de février dure quel temps ?
Q197	Quel temps est-ce que dure un mois de février ?
Q198	Je voudrais savoir quel temps dure un mois de février.
Q199	La durée d'un mois de février est de 28 ou 29 jours ?
Q200	Que dure un mois de février ?
Q201	Un mois de février dure quoi ?
Q202	Qu'est-ce que dure un mois de février ?
Q203	Je voudrais savoir qu'est-ce qu'un mois de février dure.
Q204	Un mois de février dure quelque chose ?
Q205	Un mois de février dure-t-il 28 ou 29 jours ?
Q206	Est-ce qu'un mois de février dure 28 ou 29 jours ?
Q207	Je voudrais savoir si un mois de février dure 28 ou 29 jours.
Q208	Où se trouve la Joconde ?
Q209	La Joconde se trouve où ?
Q210	Où est-ce que se trouve la Joconde ?
Q211	Je voudrais savoir où se trouve la Joconde
Q212	La Joconde se trouve au Louvre ?
Q213	A quel endroit se trouve la Joconde ?
Q214	La Joconde se trouve à quel endroit ?
Q215	A quel endroit est-ce que se trouve la Joconde ?
Q216	Je voudrais savoir à quel endroit la Joconde se trouve
Q217	L'endroit où se trouve la Joconde, c'est bien le Louvre ?
Q218	La Joconde se trouve-t-elle au Louvre ?
Q219	Au Louvre, c'est bien là que la Joconde se trouve ?
Q220	Est-ce que la Joconde se trouve au Louvre ?
Q221	Je voudrais savoir si la Joconde se trouve au Louvre ou non
Q222	C'est bien au Louvre que se trouve la Joconde ?
Q223	Dans quel musée se trouve la Joconde ?

Suite page suivante

Identifiant	Question
Q224	La Joconde se trouve dans quel musée ?
Q225	Dans quel musée est-ce que se trouve la Joconde ?
Q226	Je voudrais savoir dans quel musée la Joconde se trouve
Q227	Le musée où se trouve la Joconde, c'est bien le Louvre ?
Q228	Dans quel pays se trouve la Joconde ?
Q229	La Joconde se trouve dans quel pays ?
Q230	Dans quel pays est-ce que se trouve la Joconde ?
Q231	Je voudrais savoir dans quel pays la Joconde se trouve
Q232	Le pays où se trouve la Joconde, c'est bien la France ?
Q233	Le Rhin se trouve où ?
Q234	Où se trouve le Rhin ?
Q235	Où est-ce que se trouve le Rhin ?
Q236	Je voudrais savoir où se trouve le Rhin
Q237	A quel endroit se trouve le Rhin ?
Q238	Le Rhin se trouve à quel endroit ?
Q239	A quel endroit est-ce que se trouve le Rhin ?
Q240	Je voudrais savoir à quel endroit le Rhin se trouve
Q241	L'endroit où se trouve le Rhin, c'est bien l'Europe ?
Q242	Le Rhin se trouve-t-il en Europe de l'Ouest ?
Q243	En Europe de l'Ouest, c'est bien là que le Rhin se trouve ?
Q244	Est-ce que le Rhin se trouve en Europe de l'Ouest ?
Q245	Je voudrais savoir si le Rhin se trouve en Europe de l'Ouest ou non
Q246	C'est bien en Europe de l'Ouest que se trouve le Rhin ?
Q247	Dans quel pays se trouve le Rhin ?
Q248	Le Rhin se trouve dans quel pays ?
Q249	Dans quel pays est-ce que se trouve le Rhin ?
Q250	Je voudrais savoir dans quel pays le Rhin se trouve
Q251	Le pays où se trouve le Rhin, c'est bien la France ?
Q252	Sur quel continent se trouve le Rhin ?
Q253	Le Rhin se trouve sur quel continent ?
Q254	Sur quel continent est-ce que se trouve le Rhin ?
Q255	Je voudrais savoir sur quel continent le Rhin se trouve
Q256	Le continent où se trouve le Rhin, c'est bien l'Europe ?
Q257	Le Rhin se trouve en Europe ?
Q258	Où est la Joconde ?
Q259	La Joconde est où ?
Q260	Où est-ce qu'est la Joconde ?
Q261	Je voudrais savoir où est la Joconde

Suite page suivante

Identifiant	Question
Q262	La Joconde est au Louvre ?
Q263	A quel endroit est la Joconde ?
Q264	La Joconde est à quel endroit ?
Q265	A quel endroit est-ce qu'est la Joconde ?
Q266	Je voudrais savoir à quel endroit la Joconde est
Q267	L'endroit où est la Joconde, c'est bien le Louvre ?
Q268	La Joconde est-elle au Louvre ?
Q269	Au Louvre, c'est bien là que la Joconde est ?
Q270	Est-ce que la Joconde est au Louvre ?
Q271	Je voudrais savoir si la Joconde est au Louvre ou non
Q272	C'est bien au Louvre qu'est la Joconde ?
Q273	Dans quel musée est la Joconde ?
Q274	La Joconde est dans quel musée ?
Q275	Dans quel musée est-ce qu'est la Joconde ?
Q276	Je voudrais savoir dans quel musée la Joconde est
Q277	Le musée où est la Joconde, c'est bien le Louvre ?
Q278	Dans quel pays est la Joconde ?
Q279	La Joconde est dans quel pays ?
Q280	Dans quel pays est-ce qu'est la Joconde ?
Q281	Je voudrais savoir dans quel pays la Joconde est
Q282	Le pays où est la Joconde, c'est bien la France ?
Q283	Où est le Rhin ?
Q284	Le Rhin est où ?
Q285	Où est-ce qu'est le Rhin ?
Q286	Je voudrais savoir où est le Rhin
Q287	Le Rhin est en France ?
Q288	A quel endroit est le Rhin ?
Q289	Le Rhin est à quel endroit ?
Q290	A quel endroit est-ce qu'est le Rhin ?
Q291	Je voudrais savoir à quel endroit le Rhin est
Q292	L'endroit où est le Rhin, c'est bien en Europe de l'Ouest ?
Q293	Le Rhin est-t-il en Europe de l'Ouest ?
Q294	En Europe de l'Ouest, c'est bien là que le Rhin est ?
Q295	Est-ce que le Rhin est en Europe de l'Ouest ?
Q296	Je voudrais savoir si le Rhin est en Europe de l'Ouest ou non
Q297	C'est bien en Europe de l'Ouest qu'est le Rhin ?
Q298	Dans quel pays est le Rhin ?
Q299	Le Rhin est dans quel pays ?

Suite page suivante

Identifiant	Question
Q300	Dans quel pays est-ce qu'est le Rhin ?
Q301	Je voudrais savoir dans quel pays le Rhin est
Q302	Le pays où est le Rhin, c'est bien la France ?
Q303	Sur quel continent est le Rhin ?
Q304	Le Rhin est sur quel continent ?
Q305	Sur quel continent est-ce qu'est le Rhin ?
Q306	Je voudrais savoir sur quel continent le Rhin est
Q307	Le continent où est le Rhin, c'est bien l'Europe ?
Q308	Le Tour de France arrive où ?
Q309	Où arrive le Tour de France ?
Q310	Où est-ce qu'arrive le Tour de France ?
Q311	Je voudrais savoir où arrive le Tour de France
Q312	Le Tour de France arrive à Paris ?
Q313	A quel endroit arrive le Tour de France ?
Q314	Le Tour de France arrive à quel endroit ?
Q315	A quel endroit est-ce qu'arrive le Tour de France
Q316	Je voudrais savoir à quel endroit le Tour de France arrive
Q317	L'endroit où arrive le Tour de France, c'est bien Paris ?
Q318	Le Tour de France arrive-t-il à Paris ?
Q319	A Paris, c'est bien là que le Tour de France arrive ?
Q320	Est-ce que le Tour de France arrive à Paris ?
Q321	Je voudrais savoir si le Tour de France arrive à Paris ou non
Q322	C'est bien à Paris qu'arrive le Tour de France ?
Q323	Sur quelle avenue arrive le Tour de France ?
Q324	Le Tour de France arrive sur quelle avenue ?
Q325	Sur quelle avenue est-ce qu'arrive le Tour de France ?
Q326	Je voudrais savoir sur quelle avenue le Tour de France arrive
Q327	L'avenue où arrive le Tour de France, c'est bien l'avenue des Champs Elysées ?
Q328	Dans quelle ville arrive le Tour de France ?
Q329	Le Tour de France arrive dans quelle ville ?
Q330	Dans quelle ville est-ce qu'arrive le Tour de France ?
Q331	Je voudrais savoir dans quelle ville le Tour de France arrive
Q332	La ville où arrive le Tour de France, c'est bien Paris ?
Q333	Où peut-on traverser la Seine ?
Q334	La Seine peut être traversée où ?
Q335	Où est-ce que peut être traversée la Seine ?
Q336	Je voudrais savoir où peut être traversée la Seine
Q337	La Seine peut être traversée à Paris ?

Suite page suivante

Identifiant	Question
Q338	A quel endroit peut-on traverser la Seine ?
Q339	La Seine peut être traversée à quel endroit ?
Q340	A quel endroit est-ce que peut être traversée la Seine ?
Q341	Je voudrais savoir à quel endroit la Seine peut être traversée
Q342	Un endroit où peut être traversée la Seine est bien Paris ?
Q343	La Seine peut-elle être traversée à Paris ?
Q344	A Paris, c'est bien là que la Seine peut être traversée ?
Q345	Est-ce que la Seine peut être traversée à Paris ?
Q346	Je voudrais savoir si la Seine peut être traversée à Paris ou non
Q347	C'est bien à Paris que peut être traversée la Seine ?
Q348	A quel pont peut-on traverser la Seine ?
Q349	La Seine peut être traversée à quel pont ?
Q350	A quel pont est-ce que peut être traversée la Seine ?
Q351	Je voudrais savoir à quel pont la Seine peut être traversée
Q352	Un pont où la Seine peut être traversée est bien le pont des Arts ?
Q353	Dans quelle ville peut être traversée la Seine ?
Q354	La Seine peut être traversée dans quelle ville ?
Q355	Dans quelle ville est-ce que peut être traversée la Seine ?
Q356	Je voudrais savoir dans quelle ville la Seine peut être traversée
Q357	Une ville où la Seine peut être traversée, c'est bien Paris ?
Q358	Où est l'arrivée du Tour de France ?
Q359	L'arrivée du Tour de France est où ?
Q360	Où est-ce qu'est l'arrivée du Tour de France ?
Q361	Je voudrais savoir où est l'arrivée du Tour de France
Q362	L'arrivée du Tour de France est à Paris ?
Q363	A quel endroit est l'arrivée du Tour de France ?
Q364	L'arrivée du Tour de France est à quel endroit ?
Q365	A quel endroit est-ce qu'est l'arrivée du Tour de France
Q366	Je voudrais savoir à quel endroit l'arrivée du Tour de France est
Q367	L'endroit où est l'arrivée du Tour de France, c'est bien Paris ?
Q368	L'arrivée du Tour de France est-elle à Paris ?
Q369	A Paris, c'est bien là que l'arrivée du Tour de France est ?
Q370	Est-ce que l'arrivée du Tour de France est à Paris ?
Q371	Je voudrais savoir si l'arrivée du Tour de France est à Paris ou non
Q372	C'est bien à Paris qu'est l'arrivée du Tour de France ?
Q373	Sur quelle avenue est l'arrivée du Tour de France ?
Q374	L'arrivée du Tour de France est sur quelle avenue ?
Q375	Sur quelle avenue est-ce qu'est l'arrivée du Tour de France ?

Suite page suivante

Identifiant	Question
Q376	Je voudrais savoir sur quelle avenue l'arrivée du Tour de France est
Q377	L'avenue où est l'arrivée du Tour de France, c'est bien l'avenue des Champs Elysées ?
Q378	Dans quelle ville est l'arrivée du Tour de France ?
Q379	L'arrivée du Tour de France est dans quelle ville ?
Q380	Dans quelle ville est-ce qu'est l'arrivée du Tour de France ?
Q381	Je voudrais savoir dans quelle ville l'arrivée du Tour de France est
Q382	La ville où est l'arrivée du Tour de France, c'est bien Paris ?
Q383	Où est possible la traversée de la Seine ?
Q384	La traversée de la Seine est possible où ?
Q385	Où est-ce qu'est possible la traversée de la Seine ?
Q386	Je voudrais savoir où est possible la traversée de la Seine
Q387	La traversée de la Seine est possible à Paris ?
Q388	A quel endroit est possible la traversée de la Seine ?
Q389	La traversée de la Seine est possible à quel endroit ?
Q390	A quel endroit est-ce qu'est possible la traversée de la Seine ?
Q391	Je voudrais savoir à quel endroit la traversée de la Seine est possible
Q392	Un endroit où est possible la traversée de la Seine est bien Paris ?
Q393	La traversée de la Seine est-elle possible à Paris ?
Q394	A Paris, c'est bien là que la traversée de la Seine est possible ?
Q395	Est-ce que la traversée de la Seine est possible à Paris ?
Q396	Je voudrais savoir si la traversée de la Seine est possible à Paris ou non
Q397	C'est bien à Paris qu'est possible la traversée de la Seine ?
Q398	A quel pont est possible la traversée de la Seine ?
Q399	La traversée de la Seine est possible à quel pont ?
Q400	A quel pont est-ce qu'est possible la traversée de la Seine ?
Q401	Je voudrais savoir à quel pont la traversée de la Seine est possible
Q402	Un pont où la traversée de la Seine est possible est bien le pont des Arts ?
Q403	Dans quelle ville est possible la traversée de la Seine ?
Q404	La traversée de la Seine est possible dans quelle ville ?
Q405	Dans quelle ville est-ce qu'est possible la traversée de la Seine ?
Q406	Je voudrais savoir dans quelle ville la traversée de la Seine est possible
Q407	Une ville où la traversée de la Seine est possible, c'est bien Paris ?
Q408	Où se trouve la Vénus de Milo ?
Q409	La Vénus de Milo se trouve où ?
Q410	Où est-ce que se trouve la Vénus de Milo ?
Q411	Je voudrais savoir où se trouve la Vénus de Milo
Q412	La Vénus de Milo se trouve au Louvre ?
Q413	A quel endroit se trouve la Vénus de Milo ?

Suite page suivante

Identifiant	Question
Q414	La Vénus de Milo se trouve à quel endroit ?
Q415	A quel endroit est-ce que se trouve la Vénus de Milo ?
Q416	Je voudrais savoir à quel endroit la Vénus de Milo se trouve
Q417	L'endroit où se trouve la Vénus de Milo, c'est bien le Louvre ?
Q418	La Vénus de Milo se trouve-t-elle au Louvre ?
Q419	Au Louvre, c'est bien là que la Vénus de Milo se trouve ?
Q420	Est-ce que la Vénus de Milo se trouve au Louvre ?
Q421	Je voudrais savoir si la Vénus de Milo se trouve au Louvre ou non
Q422	C'est bien au Louvre que se trouve la Vénus de Milo ?
Q423	Dans quel musée se trouve la Vénus de Milo ?
Q424	La Vénus de Milo se trouve dans quel musée ?
Q425	Dans quel musée est-ce que se trouve la Vénus de Milo ?
Q426	Je voudrais savoir dans quel musée la Vénus de Milo se trouve
Q427	Le musée où se trouve la Vénus de Milo, c'est bien le Louvre ?
Q428	Dans quel pays se trouve la Vénus de Milo ?
Q429	La Vénus de Milo se trouve dans quel pays ?
Q430	Dans quel pays est-ce que se trouve la Vénus de Milo ?
Q431	Je voudrais savoir dans quel pays la Vénus de Milo se trouve
Q432	Le pays où se trouve la Vénus de Milo, c'est bien la France ?
Q433	Où se trouvent les Alpes ?
Q434	Les Alpes se trouvent où ?
Q435	Où est-ce que se trouvent les Alpes ?
Q436	Je voudrais savoir où se trouvent les Alpes
Q437	Les Alpes se trouvent en Europe ?
Q438	A quel endroit se trouvent les Alpes ?
Q439	Les Alpes se trouvent à quel endroit ?
Q440	A quel endroit est-ce que se trouvent les Alpes ?
Q441	Je voudrais savoir à quel endroit les Alpes se trouvent
Q442	L'endroit où se trouvent les Alpes, c'est bien l'Europe ?
Q443	Les Alpes se trouvent-t-elles en Europe ?
Q444	En Europe, c'est bien là que les Alpes se trouvent ?
Q445	Est-ce que les Alpes se trouvent en Europe ?
Q446	Je voudrais savoir si les Alpes se trouvent en Europe ou non
Q447	C'est bien en Europe que se trouvent les Alpes ?
Q448	Dans quel pays se trouvent les Alpes ?
Q449	Les Alpes se trouvent dans quel pays ?
Q450	Dans quel pays est-ce que se trouvent les Alpes ?
Q451	Je voudrais savoir dans quel pays les Alpes se trouvent

Suite page suivante

Identifiant	Question
Q452	Le pays où se trouvent les Alpes, c'est bien la France ?
Q453	Sur quel continent se trouvent les Alpes ?
Q454	Les Alpes se trouvent sur quel continent ?
Q455	Sur quel continent est-ce que se trouvent les Alpes ?
Q456	Je voudrais savoir sur quel continent les Alpes se trouvent
Q457	Le continent où se trouvent les Alpes, c'est bien l'Europe ?
Q458	Où est la Vénus de Milo ?
Q459	La Vénus de Milo est où
Q460	Où est-ce qu'est la Vénus de Milo ?
Q461	Je voudrais savoir où est la Vénus de Milo
Q462	La Vénus de Milo est au Louvre ?
Q463	A quel endroit est la Vénus de Milo ?
Q464	La Vénus de Milo est à quel endroit ?
Q465	A quel endroit est-ce qu'est la Vénus de Milo ?
Q466	Je voudrais savoir à quel endroit la Vénus de Milo est
Q467	L'endroit où est la Vénus de Milo, c'est bien le Louvre ?
Q468	La Vénus de Milo est-elle au Louvre ?
Q469	Au Louvre, c'est bien là que la Vénus de Milo est ?
Q470	Est-ce que la Vénus de Milo est au Louvre ?
Q471	Je voudrais savoir si la Vénus de Milo est au Louvre ou non
Q472	C'est bien au Louvre qu'est la Vénus de Milo ?
Q473	Dans quel musée est la Vénus de Milo ?
Q474	La Vénus de Milo est dans quel musée ?
Q475	Dans quel musée est-ce qu'est la Vénus de Milo ?
Q476	Je voudrais savoir dans quel musée la Vénus de Milo est
Q477	Le musée où est la Vénus de Milo, c'est bien le Louvre ?
Q478	Dans quel pays est la Vénus de Milo ?
Q479	La Vénus de Milo est dans quel pays ?
Q480	Dans quel pays est-ce qu'est la Vénus de Milo ?
Q481	Je voudrais savoir dans quel pays la Vénus de Milo est
Q482	Le pays où est la Vénus de Milo, c'est bien la France ?
Q483	Où sont les Alpes ?
Q484	Les Alpes sont en Europe ?
Q485	Les Alpes sont où ?
Q486	Où est-ce que sont les Alpes ?
Q487	Je voudrais savoir où sont les Alpes
Q488	A quel endroit sont les Alpes ?
Q489	Les Alpes sont à quel endroit ?

Suite page suivante

Identifiant	Question
Q490	A quel endroit est-ce que sont les Alpes ?
Q491	Je voudrais savoir à quel endroit les Alpes sont
Q492	L'endroit où sont les Alpes, c'est bien l'Europe ?
Q493	Les Alpes sont-elles en Europe ?
Q494	En Europe, c'est bien là que les Alpes sont ?
Q495	Est-ce que les Alpes sont en Europe ?
Q496	Je voudrais savoir si les Alpes sont en Europe ou non
Q497	C'est bien en Europe que sont les Alpes ?
Q498	Dans quel pays sont les Alpes ?
Q499	Les Alpes sont dans quel pays ?
Q500	Dans quel pays est-ce que sont les Alpes ?
Q501	Je voudrais savoir dans quel pays les Alpes sont
Q502	Le pays où sont les Alpes, c'est bien la France ?
Q503	Sur quel continent sont les Alpes ?
Q504	Les Alpes sont sur quel continent ?
Q505	Sur quel continent est-ce que sont les Alpes ?
Q506	Je voudrais savoir sur quel continent les Alpes sont
Q507	Le continent où sont les Alpes, c'est bien l'Europe ?
Q508	Où finit le Rhône ?
Q509	Le Rhône finit où ?
Q510	Où est-ce que finit le Rhône ?
Q511	Je voudrais savoir où finit le Rhône
Q512	Le Rhône finit dans le delta de Camargue ?
Q513	A quel endroit finit le Rhône ?
Q514	Le Rhône finit à quel endroit ?
Q515	A quel endroit est-ce que finit le Rhône ?
Q516	Je voudrais savoir à quel endroit le Rhône finit
Q517	L'endroit où finit le Rhône, c'est bien la Camargue ?
Q518	Le Rhône fini-t-il en Camargue ?
Q519	En Camargue, c'est bien là que le Rhône finit ?
Q520	Est-ce que le Rhône finit en Camargue ?
Q521	Je voudrais savoir si le Rhône finit en Camargue ou non
Q522	C'est bien en Camargue que finit le Rhône ?
Q523	A quel delta finit le Rhône ?
Q524	Le Rhône finit à quel delta ?
Q525	A quel delta est-ce que finit le Rhône ?
Q526	Je voudrais savoir a quel delta finit le Rhône
Q527	Le delta où finit le Rhône, c'est bien le delta de Camargue ?

Suite page suivante

Identifiant	Question
Q528	Dans quelle partie de France finit le Rhône ?
Q529	Le Rhône finit dans quelle partie de France ?
Q530	Dans quelle partie de France est-ce que finit le Rhône ?
Q531	Je voudrais savoir dans quelle partie de France le Rhône finit
Q532	La partie de France où finit le Rhône, c'est bien le Sud-Est ?
Q533	Où arrivent les avions à Paris ?
Q534	Les avions arrivent où à Paris ?
Q535	Où est-ce qu'arrivent les avions à Paris ?
Q536	Je voudrais savoir où arrivent les avions à Paris
Q537	Les avions à Paris arrivent à soit à Orly, soit à Roissy ?
Q538	A quel endroit arrivent les avions à Paris ?
Q539	Les avions à Paris arrivent à quel endroit ?
Q540	A quel endroit est-ce qu'arrivent les avions à Paris ?
Q541	Je voudrais savoir à quel endroit les avions arrivent à Paris
Q542	Un endroit où arrivent les avions à Paris est bien soit Orly soit Roissy ?
Q543	Les avions arrivent-ils à Paris soit à Orly soit à Roissy ?
Q544	Soit à Orly, soit à roissy, c'est bien là que les avions arrivent à Paris ?
Q545	Est-ce que les avions arrivent à Paris soit à Orly soit à Roissy ?
Q546	Je voudrais savoir si les avions arrivent à Paris soit à Orly soit à Roissy ou non
Q547	C'est bien soit à Orly soit à Roissy qu'arrivent les avions à Paris ?
Q548	A quel terminal arrivent les avions à Paris ?
Q549	Les avions à Paris arrivent à quel terminal ?
Q550	A quel terminal est-ce qu'arrivent les avions à Paris ?
Q551	Je voudrais savoir à quel terminal les avions arrivent à Paris
Q552	Un terminal où les avions arrivent à Paris est le terminal 3 à Orly ?
Q553	Dans quel aéroport arrivent les avions à Paris ?
Q554	Les avions à Paris arrivent dans quel aéroport ?
Q555	Dans quel aéroport est-ce qu'arrivent les avions à Paris ?
Q556	Je voudrais savoir dans quel aéroport les avions arrivent à Paris
Q557	Un aéroport où les avions arrivent à Paris, c'est bien Orly ?
Q558	Où est l'estuaire du Rhône ?
Q559	L'estuaire du Rhône est où ?
Q560	Où est-ce qu'est l'estuaire du Rhône ?
Q561	Je voudrais savoir où est l'estuaire du Rhône
Q562	L'estuaire du Rhône est en Camargue ?
Q563	A quel endroit est l'estuaire du Rhône ?
Q564	L'estuaire du Rhône est à quel endroit ?
Q565	A quel endroit est-ce qu'est l'estuaire du Rhône ?

Suite page suivante

Identifiant	Question
Q566	Je voudrais savoir à quel endroit l'estuaire du Rhône est
Q567	L'endroit où est l'estuaire du Rhône, c'est bien la Camargue ?
Q568	L'estuaire du Rhône est-il en Camargue ?
Q569	En Camargue, c'est bien là que l'estuaire du Rhône est ?
Q570	Est-ce que l'estuaire du Rhône est en Camargue ?
Q571	Je voudrais savoir si l'estuaire du Rhône est en Camargue ou non
Q572	C'est bien en Camargue qu'est l'estuaire du Rhône ?
Q573	A quel delta est l'estuaire du Rhône ?
Q574	L'estuaire du Rhône est à quel delta ?
Q575	A quel delta est-ce qu'est l'estuaire du Rhône ?
Q576	Je voudrais savoir à quel delta l'estuaire du Rhône est
Q577	Le delta où est l'estuaire du Rhône, c'est bien l'estuaire de Camargue ?
Q578	Dans quelle partie de France est l'estuaire du Rhône ?
Q579	L'estuaire du Rhône est dans quelle partie de France ?
Q580	Dans quelle partie de France est-ce qu'est l'estuaire du Rhône ?
Q581	Je voudrais savoir dans quelle partie de France est l'estuaire du Rhône
Q582	La partie de France où est l'estuaire du Rhône, c'est bien le sud-ouest ?
Q583	Où est l'arrivée des avions à Paris ?
Q584	L'arrivée des avions à Paris est où ?
Q585	Où est-ce qu'est l'arrivée des avions à Paris ?
Q586	Je voudrais savoir où est l'arrivée des avions à Paris
Q587	L'arrivée des avions à Paris est soit à Orly soit à Roissy ?
Q588	A quel endroit est l'arrivée des avions à Paris ?
Q589	L'arrivée des avions à Paris est à quel endroit ?
Q590	A quel endroit est-ce qu'est l'arrivée des avions à Paris ?
Q591	Je voudrais savoir à quel endroit l'arrivée des avions à Paris est
Q592	L'endroit où est l'arrivée des avions à Paris est bien soit Orly soit Roissy ?
Q593	L'arrivée des avions à Paris est-elle soit à Orly soit à Roissy ?
Q594	Soit à Orly, soit à Roissy, c'est bien là que l'arrivée des avions à Paris est ?
Q595	Est-ce que l'arrivée des avions à Paris est soit à Orly soit à Roissy ?
Q596	Je voudrais savoir si l'arrivée des avions à Paris est soit à Orly soit à Roissy ou non
Q597	C'est bien soit à Roissy soit à Orly qu'est l'arrivée des avions à Paris ?
Q598	A quel terminal est l'arrivée des avions à Paris ?
Q599	L'arrivée des avions à Paris est à quel terminal ?
Q600	A quel terminal est-ce qu'est l'arrivée des avions à Paris ?
Q601	Je voudrais savoir à quel terminal l'arrivée des avions à Paris est
Q602	Un terminal où l'arrivée des avions à Paris est, c'est bien le terminal 3 à Orly ?
Q603	Dans quel aéroport est l'arrivée des avions à Paris ?

Suite page suivante

Identifiant	Question
Q604	L'arrivée des avions à Paris est dans quel aéroport ?
Q605	Dans quel aéroport est-ce qu'est l'arrivée des avions à Paris ?
Q606	Je voudrais savoir dans quel aéroport l'arrivée des avions à Paris est
Q607	Un aéroport où l'arrivée des avions à Paris est, c'est bien Orly ?
Q608	Quand a eu lieu l'armistice de la 1ère guerre mondiale ?
Q609	L'armistice de la 1ère guerre mondiale a eu lieu quand ?
Q610	Quand est-ce qu'a eu lieu l'armistice de la 1ère guerre mondiale ?
Q611	Je voudrais savoir quand a eu lieu l'armistice de la 1ère guerre mondiale
Q612	L'armistice de la 1ère guerre mondiale a eu lieu le 11 novembre 1918 ?
Q613	A quel moment a eu lieu l'armistice de la 1ère guerre mondiale ?
Q614	L'armistice de la 1ère guerre mondiale a eu lieu à quel moment ?
Q615	A quel moment est-ce qu'a eu lieu l'armistice de la 1ère guerre mondiale ?
Q616	Je voudrais savoir à quel moment l'armistice de la 1ère guerre mondiale a eu lieu
Q617	Le moment où a eu lieu l'armistice de la 1ère guerre mondiale, c'est bien le 11 novembre 1918 ?
Q618	L'armistice de la 1ère guerre mondiale a-t-il eu lieu le 11 novembre 1918 ?
Q619	Le 11 novembre 1918, c'est bien à ce moment qu'a eu lieu l'armistice de la 1ère guerre mondiale ?
Q620	Est-ce que l'armistice de la 1ère guerre mondiale a eu lieu le 11 novembre 1918 ?
Q621	Je voudrais savoir si l'armistice de la 1ère guerre mondiale a eu lieu le 11 novembre 1918
Q622	C'est bien le 11 novembre 1918 qu'a eu lieu l'armistice de la 1ère guerre mondiale ?
Q623	A quelle date a eu lieu l'armistice de la 1ère guerre mondiale ?
Q624	L'armistice de la 1ère guerre mondiale a eu lieu à quelle date ?
Q625	A quelle date est-ce qu'a eu lieu l'armistice de la 1ère guerre mondiale ?
Q626	Je voudrais savoir à quelle date l'armistice de la 1ère guerre mondiale a eu lieu
Q627	La date à laquelle a eu lieu l'armistice de la 1ère guerre mondiale, c'est bien le 11 novembre 1918 ?
Q628	En quelle année a eu lieu l'armistice de la 1ère guerre mondiale ?
Q629	L'armistice de la 1ère guerre mondiale a eu lieu en quelle année ?
Q630	En quelle année est-ce qu'a eu lieu l'armistice de la 1ère guerre mondiale ?
Q631	Je voudrais savoir en quelle année l'armistice de la 1ère guerre mondiale a eu lieu
Q632	L'année pendant laquelle a eu lieu l'armistice de la 1ère guerre mondiale, c'est bien 1918 ?
Q633	Quand ont lieu les jeux olympiques d'été ?
Q634	Les jeux olympiques d'été ont lieu quand ?
Q635	Quand est-ce qu'ont lieu les jeux olympiques d'été ?
Q636	Je voudrais savoir quand ont lieu les jeux olympiques d'été
Q637	Les jeux olympiques d'été ont lieu en été, tous les 4 ans ?
Q638	A quel moment ont lieu les jeux olympiques d'été ?
Q639	Les jeux olympiques d'été ont lieu à quel moment ?

Suite page suivante

Identifiant	Question
Q640	A quel moment est-ce qu'ont lieu les jeux olympiques d'été ?
Q641	Je voudrais savoir à quel moment les jeux olympiques d'été ont lieu
Q642	Le moment où ont lieu les jeux olympiques d'été, c'est bien en été, tous les 4 ans ?
Q643	Les jeux olympiques d'été ont-il lieu en été, tous les 4 ans ?
Q644	En été et tous les 4 ans, c'est bien à ce moment qu'ont lieu les jeux olympiques d'été ?
Q645	Est-ce que les jeux olympiques d'été ont lieu en été, tous les 4 ans ?
Q646	Je voudrais savoir si les jeux olympiques d'été ont lieu en été, tous les 4 ans ou non
Q647	C'est bien en été et tous les 4 ans qu'ont lieu les jeux olympiques d'été ?
Q648	Quels mois ont lieu les jeux olympiques d'été ?
Q649	Les jeux olympiques d'été ont lieu quels mois ?
Q650	Quels mois est-ce qu'ont lieu les jeux olympiques d'été ?
Q651	Je voudrais savoir quels mois les jeux olympiques d'été ont lieu
Q652	Les mois où ont lieu les jeux olympiques d'été, c'est bien juillet, août ou septembre ?
Q653	Quelles années ont lieu les jeux olympiques d'été ?
Q654	Les jeux olympiques d'été ont lieu quelles années ?
Q655	Quelles années est-ce qu'ont lieu les jeux olympiques d'été ?
Q656	Je voudrais savoir quelles années les jeux olympiques d'été ont lieu
Q657	Les années où ont lieu les jeux olympiques d'été, c'est bien tous les 4 ans, en 2004, en 2008, en 2012...
Q658	Quand a été l'armistice de la 1ère guerre mondiale ?
Q659	L'armistice de la 1ère guerre mondiale a été quand ?
Q660	Quand est-ce qu'a été l'armistice de la 1ère guerre mondiale ?
Q661	Je voudrais savoir quand a été l'armistice de la 1ère guerre mondiale
Q662	L'armistice de la 1ère guerre mondiale a été le 11 novembre 1918 ?
Q663	A quel moment a été l'armistice de la 1ère guerre mondiale ?
Q664	L'armistice de la 1ère guerre mondiale a été à quel moment ?
Q665	A quel moment est-ce qu'a été l'armistice de la 1ère guerre mondiale ?
Q666	Je voudrais savoir à quel moment l'armistice de la 1ère guerre mondiale a été
Q667	Le moment où a été l'armistice de la 1ère guerre mondiale, c'est bien le 11 novembre 1918 ?
Q668	L'armistice de la 1ère guerre mondiale a-t-il été le 11 novembre 1918 ?
Q669	Le 11 novembre 1918, c'est bien à ce moment qu'a été l'armistice de la 1ère guerre mondiale ?
Q670	Est-ce que l'armistice de la 1ère guerre mondiale a été le 11 novembre 1918 ?
Q671	Je voudrais savoir si l'armistice de la 1ère guerre mondiale a été le 11 novembre 1918
Q672	C'est bien le 11 novembre 1918 qu'a été l'armistice de la 1ère guerre mondiale ?
Q673	A quelle date a été l'armistice de la 1ère guerre mondiale ?
Q674	L'armistice de la 1ère guerre mondiale a été à quelle date ?
Q675	A quelle date est-ce qu'a été l'armistice de la 1ère guerre mondiale ?

Suite page suivante

Identifiant	Question
Q676	Je voudrais savoir à quelle date l'armistice de la 1ère guerre mondiale a été
Q677	La date à laquelle a été l'armistice de la 1ère guerre mondiale, c'est bien le 11 novembre 1918 ?
Q678	En quelle année a été l'armistice de la 1ère guerre mondiale ?
Q679	L'armistice de la 1ère guerre mondiale a été en quelle année ?
Q680	En quelle année est-ce qu'a été l'armistice de la 1ère guerre mondiale ?
Q681	Je voudrais savoir en quelle année l'armistice de la 1ère guerre mondiale a été
Q682	L'année pendant laquelle a été l'armistice de la 1ère guerre mondiale, c'est bien 1918 ?
Q683	Quand sont les jeux olympiques d'été ?
Q684	Les jeux olympiques d'été sont quand ?
Q685	Quand est-ce que sont les jeux olympiques d'été ?
Q686	Je voudrais savoir quand sont les jeux olympiques d'été
Q687	Les jeux olympiques d'été sont en été, tous les 4 ans ?
Q688	A quel moment sont les jeux olympiques d'été ?
Q689	Les jeux olympiques d'été sont à quel moment ?
Q690	A quel moment est-ce que sont les jeux olympiques d'été ?
Q691	Je voudrais savoir à quel moment les jeux olympiques d'été sont
Q692	Le moment où sont les jeux olympiques d'été, c'est bien en été, tous les 4 ans ?
Q693	Les jeux olympiques d'été sont-il en été, tous les 4 ans ?
Q694	En été et tous les 4 ans, c'est bien à ce moment que sont les jeux olympiques d'été ?
Q695	Est-ce que les jeux olympiques d'été sont en été, tous les 4 ans ?
Q696	Je voudrais savoir si les jeux olympiques d'été sont en été, tous les 4 ans ou non
Q697	C'est bien en été et tous les 4 ans que sont les jeux olympiques d'été ?
Q698	Quels mois sont les jeux olympiques d'été ?
Q699	Les jeux olympiques d'été sont quels mois ?
Q700	Quels mois est-ce que sont les jeux olympiques d'été ?
Q701	Je voudrais savoir quels mois les jeux olympiques d'été sont
Q702	Les mois où sont les jeux olympiques d'été, c'est bien juillet, août ou septembre ?
Q703	Quelles années sont les jeux olympiques d'été ?
Q704	Les jeux olympiques d'été sont quelles années ?
Q705	Quelles années est-ce que sont les jeux olympiques d'été ?
Q706	Je voudrais savoir quelles années les jeux olympiques d'été sont
Q707	Les années où sont les jeux olympiques d'été, c'est bien tous les 4 ans, en 2004, en 2008, en 2012...

Suite page suivante

A.2 Exemple de dialogue avec le système ELIZA - Version française

Les hommes sont tous les mêmes.
EN QUEL SENS ?
Ils nous embêtent toujours à propos de tout et n'importe quoi.
PENSEZ-VOUS À UN EXEMPLE PRECIS ?
Et bien, mon petit copain m'a fait venir ici.
VOTRE PETIT COPAIN VOUS A FAIT VENIR ICI
Il dit que je suis souvent déprimée.
JE SUIS DÉSOLÉ D'ENTENDRE QUE VOUS ÊTES DÉPRIMÉE
C'est vrai. Je suis malheureuse
PENSEZ-VOUS QUE VENIR ICI VOUS AIDERA À NE PAS ÊTRE MALHEUREUSE
J'ai besoin d'aide, ça me semble évident.
QU'EST-CE QUE ÇA SIGNIFIE POUR VOUS AVOIR DE L'AIDE
Peut-être que je pourrais apprendre à m'entendre avec ma mère.
DITES M'EN PLUS SUR VOTRE FAMILLE
...

FIGURE A.1 – *Traduction de l'exemple de dialogue avec le système ELIZA présenté page 11 et extrait de [Weizenbaum, 1966] - Les lignes capitalisées sont les réponses de la machine*

? ? ? ? ? ?

BIBLIOGRAPHIE

Anne Abeillé, et al. (2000). *Ingénierie des langues*. Jean-Marie Pierrel.

Acapella (site web). http ://www.acapela-group.com/index.php, dernière consultation 06/2010.

Martine Adda-Decker, et al. (2003). 'A disfluency study for cleaning spontaneous speech automatic transcriptsand improving speech language models'. Dans *Proc. Disfluency in Spontaneous Speech*, pp. 67–70.

Peter Adolphs, et al. (2010). 'Question Answering Biographic Information and Social Network Powered by the Semantic Web'. Dans *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/.

Salah Ait-Mokhtar, et al. (2002). 'Robustness beyond shallowness : incremental deep parsing'. *Natural Language Engineering* **8**(2-3) :121–144.

Glenda B. Anaya & Leila Kosseim (2003). 'Generation of natural responses through syntactic patterns'. Dans *Actes de la Dixième Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN 2003)*, pp. 297–302.

Answers.com (site web). 'WikiAnswers'. http ://wiki.answers.com/, dernière consultation 09/2010.

Answers.com (fr) (site web). 'Français WikiAnswers'. http ://fr.answers.com/, dernière consultation 09/2010.

International Phonetic Association (1999). *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet*. Cambridge : Cambridge University Press.

J. Authier & A. Meunier (1972). 'Norme, grammaticalité et niveaux de langue'. *Langue française* **16**(1) :49–62.

Gerhard Bauer (1985). *Namenkunde des deutschen*. P. Lang, Bern ; New York.

Frédéric Beaugendre (1994). *Une étude perceptive de l'intonation du français : Développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du texte*. Thèse de doctorat, Université Paris Sud Orsay.

Yacine Bellik, et al. (1995). ‘Interaction Multimodale : Concepts et Architecture’. Montpellier, France.

Farah Benamara (2004). *WebCoop : Un systèmes de réponses coopératives sur le Web*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Farah Benamara & Patrick Saint Dizier (2003). ‘Construction de réponses coopératives : du corpus à la modélisation informatique’. *Revue québécoise de linguistique* **32**(1) :199–234.

Farah Benamara, et al. (2004). ‘COOPML : a Language for Annotating Cooperative Discourse’. Dans *ACL- Discourse annotation, Barcelona, 22/07/04-26/07/04*, pp. 9–16, http ://www.aclweb.org. Association for Computational Linguistics (ACL).

Farah Benamara & Patrick Saint-Dizier (2004). ‘Lexicalisation Strategies in Cooperative Question- Answering Systems’. Dans *COLING 04, Genève, 22/08/04-26/08/04*, pp. 345–352. MIT Press.

Susan E. Brennan & Maurice Williams (1995). ‘The Feeling of Another’s Knowing : Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers’. *Journal of Memory and Language* **34**(3) :383–398.

Eric Brill, et al. (2002). ‘An Analysis of the AskMSR Question-Answering System’. Dans *In Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP*, pp. 257–264.

Eric Brill, et al. (2001). ‘Data-Intensive Question Answering’. Dans *In Proceedings of the Tenth Text REtrieval Conference (TREC*, pp. 393–400.

Gaël De Chalendar, et al. (2002). ‘The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet’. Dans *In Proceedings of the Eleventh Text REtrieval Conference (TREC)*, pp. 407–416.

Chatbot.org (site web). http ://www.Chatbot.org, dernière consultation 06/2010.

Wesley W. Chu, et al. (1996). ‘CoBase : a scalable and extensible cooperative information system’. *J. Intell. Inf. Syst.* **6**(2-3) :223–259.

Frederic Cuppens & Robert Demolombe (1989). ‘How to recognize interesting topics to provide cooperative answering’. *Inf. Syst.* **14**(2) :163–173.

Laurence Danlos (1989a). ‘Génération automatique de textes en langue naturelle’. *Annals of Telecommunications* **44** :94–100. 10.1007/BF02999881.

Laurence Danlos (1989b). ‘Génération automatique de textes en langue naturelle’. *Annals of Telecommunications* **44** :94–100.

T. Darrell, et al. (2000). ‘Audio-visual Segmentation and "The Cocktail Party Effect"’. *Advances in Multimodal Interfaces-ICMI 2000* pp. 32–40.

Florence Duclaye, et al. (2002). ‘Using the Web as a Linguistic Resource for Learning Reformulations’. Dans *Proceedings of the third international conference on language resources and evaluation (LREC’02*, pp. 390–396.

Sarra El Ayari (2007). ‘Évaluation transparente de systèmes de questions-réponses : application au focus’. Dans *RECITAL*.

Marc El-Bèze (2006). *Systèmes de questions-réponses in Compréhension des langues et interaction*, chap. Chap 10 : Systèmes de questions-réponses, pp. 277–297. Sabah, Gérard.

C. Fairon, et al. (2006). *Le langage SMS : étude d’un corpus informatisé à partir de l’enquête Faites don de vos SMS à la science*. Cental.

Faites-la-Science (site web). http ://faites-la-science.risc.cnrs.fr/, dernière consultation 07/2010.

Olivier Ferret, et al. (2001). ‘Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse’. Dans *TALN*, Tours.

Olivier Ferret, et al. (2002). ‘How NLP can improve question answering’. *Knowledge Organization* **29**(3-4) :135–155.

N. Fourour & E. Morin (2003a). ‘Apport du Web dans la reconnaissance des entités nommées’. *Revue québécoise de linguistique* **32**(1) :41–60.

Nordine Fourour & Emmanuel Morin (2003b). ‘Apport du Web dans la reconnaissance des entités nommées.’. Dans *Revue Québécoise de Linguistique (RQL)*, vol. 32, pp. 41–60.

Terry Gaasterland, et al. (1992). ‘An Overview of Cooperative Answering’. *Journal of Intelligent Information Systems* **1** :123–157.

Olivier Galibert (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. Thèse de doctorat, Université Paris Sud, Orsay.

Olivier Galibert, et al. (2005a). ‘An Open-Domain, Human-Computer Dialog System’. Dans *InterSpeech*.

Olivier Galibert, et al. (2005b). ‘Ritel : an open-domain, human-computer dialog system’. Dans *Ninth European Conference on Speech Communication and Technology*.

Olivier Galibert, et al. (2005c). ‘Ritel : dialogue hommemachine à domaine ouvert’. *les actes de Traitement Automatique des Langues (TALN)* **1** :439.

Anne Garcia-Fernandez & Carole Lailler (2008). ‘Morphosyntaxe de l’interrogation pour le système de question-réponse RITEL’. Dans *RECITAL*, p. 10.

Anne Garcia-Fernandez, et al. (2009). ‘Collecte et analyses de réponses naturelles pour les systèmes de questions-réponses’. Dans *Actes de TALN 2009*.

Anne Garcia-Fernandez, et al. (2010a). ‘Comment formule-t-on une réponse en langue naturelle ?’. Dans *Actes de TALN 2010*.

Anne Garcia-Fernandez, et al. (2010b). ‘MACAQ : A Multi Annotated Corpus to study how we adapt Answers to various Questions’. Dans *Proceedings of the Seventh International Language Resources and Evaluation (LREC’10)*, p. 7, Valletta, Malta. European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2010/.

Anne Garcia-Fernandez, et al. (2010c). ‘Euh as cue for speaker confidence and word searching in human spoken answers in French’. Dans *DiSS-LPSS Joint Workshop*.

Jean-Luc Gauvain, et al. (2005). ‘Where Are We in Transcribing French Broadcast News ?’. Dans *InterSpeech*, Lisbon.

Jean-Luc Gauvain, et al. (2000). ‘An Overview of Speech Recognition Activities at LIMSI’. Dans *Sino-French Symposium on Speech and Language processing*, Beijing.

Thierry Grass (2000). ‘Typologie et traductibilité des noms propres de l’allemand vers le français’. *Traitement Automatique des Langues* **41**(3) :643–670.

Nancy Green & Sandra Carberry (1999). ‘A computational mechanism for initiative in answer generation’. *User Modeling and User-Adapted Interaction* **9**(1) :93–132.

Herbert P. Grice (1975). *Syntax and Semantics*, vol. 3 of *Speech Acts*, chap. Logic and conversation, pp. 41–58. Academic Press, New York.

R. Grishman & B. Sundheim (1996). ‘Message understanding conference-6 : A brief history’. Dans *Proceedings of the 16th conference on Computational linguistics-Volume 1*, p. 471. Association for Computational Linguistics.

Barbara J. Grosz & Sidner Candace L. (1990). *Intentions in COMMUNICATION*, chap. Plans for discourse, pp. 417–444. MIT Press.

M.L. Guénot (2005). ‘Parsing de l’oral : traiter les disfluences’. *Les actes de Traitement Automatique des Langues (TALN)* **1** :439.

Iryna Gurevych, et al. (2009). 'Educational Question Answering based on Social Media Content'. Dans *Proceeding of the 2009 conference on Artificial Intelligence in Education*, pp. 133–140, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Sanda Harabagiu, et al. (2001). 'The role of lexico-semantic feedback in open-domain textual question-answering'. Dans *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 282–289. Association for Computational Linguistics Morristown, NJ, USA.

Ryuichiro Higashinaka, et al. (2006). 'Learning to generate naturalistic utterances using reviews in spoken dialogue systems'. Dans *ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 265–272, Morristown, NJ, USA. Association for Computational Linguistics.

Meriam Horchani (2007). *Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales*. Thèse de doctorat, Docteur de l'Université Joseph Fourier - Grenoble 1.

Xuedong Huang, et al. (2001). *Spoken language processing : A guide to theory, algorithm, and system development*. Prentice Hall PTR Upper Saddle River, NJ, USA.

Roman Jakobson (1963). *Éssais de linguistique générale*. Ed. de Minuit.

Daniel Jurafsky & James H. Martin (2006). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

L. Karsenty (2006). 'Les déterminants du choix d'une modalité d'interaction avec une interface multimodale'. Dans *Actes de la conférence ERGO-IA' O*, pp. 11–13.

Leila Kosseim (2000). 'Systemes de réponse automatique : Etat de l'art'. -.

Carole Lailler (2008). 'Validation d'un modèle théorique via des corpora'. Dans *Actes des Xèmes RJC ED268 "Langage et Langues" intitulé "Objets, outils et concepts"*, Paris.

Carole Lailler (2009). 'Une grammaire à l'épreuve de corpora : la Grammaire Interactive'. Dans *Actes du 3ème colloque international de l'AFLiCo : grammaire(s) en construction*, Nanterre.

Lori Lamel, et al. (2005). 'Transcribing Lectures and Seminars'. Dans *in InterSpeech'05*, Lisbon, Portugal.

Lori Lamel, et al. (2000). 'The limsi arise system'. *Speech Communication* **31**(4) :339–354.

Dominique Laurent & Patrick Séguéla (2005). 'QRISTAL, système de Questions-Réponses'. pp. 53–62.

Wendy G. Lehnert (1978). *The process of question answering : a computer simulation of cognition / Wendy G. Lehnert*. L. Erlbaum Associates ; distributed by Halsted Press, Hillsdale, N.J. : New York :.

Jérome Lehuen (1997). *Un modèle de dialogue dynamique et générique intégrant l'acquisition de sa compétence linguistique, Le système COALA*. Thèse de doctorat, Université du Maine, Le Mans.

Anne-Laure Ligozat (2004). 'Système de Question-Réponse : Apport de l'Analyse Syntaxique à l'Extraction de la Réponse.'. Dans *RECITAL*.

Anne-Laure Ligozat (2006). *Exploitation et fusion de connaissances locales pour la recherche d'informations précises*. Thèse de doctorat, Université Paris Sud, Orsay.

Daniel Luzzati (2001). 'Pour une grammaire interactive : l'exemple de l'interrogation'. 2001.

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

I. Mani & M.T. Maybury (1999). *Advances in automatic text summarization*. MIT Press.

Christopher D. Manning, et al. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

David D. McDonald (2003). 'Producing dialog at MERL : problems in generation engineering'. Dans A. Spring (ed.), *Proceedings of Natural Language Generation in Spoken and Written Dialogue*, pp. 104–111.

Sara Mendes & Véronique Moriceau (2004). 'L'analyse des questions : intérêt pour la génération des réponses'. Dans *TALN*.

Jean-Luc Minel (2002). *Filtrage sémantique : du résumé automatique à la fouille de textes*. Hermès-Lavoisier.

Laura Monceaux & Isabelle Robba (2002). 'Les analyseurs syntaxiques : atouts pour une analyse des questions dans un système de question-réponse ?'. Dans *TALN*.

Mary-Annick Morel & Laurent Danon-Boileau (1998). *Grammaire de l'intonation l'exemple du français*. Editions Ophrys.

Véronique Moriceau (2006). 'Numerical data integration for cooperative question-answering'. Dans *KRAQ'06 : Proceedings of the Workshop KRAQ'06 on Knowledge and Reasoning for Language Processing*, pp. 42–49, Morristown, NJ, USA. Association for Computational Linguistics.

Véronique Moriceau (2007). *Intégration de données dans un système question-réponse sur le Web*. Thèse de doctorat.

Véronique Moriceau, et al. (2009). 'Utilisation de la syntaxe pour valider les réponses à des questions par plusieurs documents'. Dans *Actes de la 6ème COnférence en Recherche d'Information et Applications (CORIA)*.

Amihai Motro (1994). 'Intensional Answers to Database Queries'. *IEEE Trans. on Knowl. and Data Eng.* **6**(3) :444–454.

MUC (site web). 'Evaluations website'. http ://www-nlpir.nist.gov/related_projects/muc/index.html, dernière consultation 07/2010.

D Naber (2003). 'A rule-based style and grammar checker'. Master's thesis, Technische Fakultät Universität Bielefeld.

Adeline Nazarenko, et al. (2006). *Compréhension des langues et interaction*. Gérard Sabah.

NTCIR (site web). http ://research.nii.ac.jp/ntcir/index-en.html, dernière consultation 10/2010.

A.H. Oh & A.I. Rudnicky (2002). 'Stochastic natural language generation for spoken dialog systems'. *Computer Speech and Language* **16**(3) :387–407.

T. Paek (2001). 'Empirical methods for evaluating dialog systems'. Dans *ACL 2001 workshop on evaluation methodologies for language and dialogue systems*, pp. 3–10.

Woojin Paik, et al. (1996). *Corpus Processing for Lexical Acquisition*, chap. Categorizing and standardizing proper nouns for efficient information retrieval, pp. 44–54. MIT Press, Cambridge, MA, USA.

Luc Plamondon, et al. (2002). 'The QUANTUM Question Answering System at TREC-11'. Dans E. M. Vorhees & D. K. Harman (eds.), *Proceedings of the Eleventh Text Retrieval Conference (TREC-2002)*, pp. 750–757, Gaithersburg, Maryland. NIST.

QA Track (site web). 'The TREC 2007 Question Answering Track Guidelines'. http ://trec.nist.gov/data/qa/2007_qadata/qa.07.guidelines.html, dernière consultation 05/2010.

QA@CLEF (site web). http ://celct.isti.cnr.it/ResPubliQA/index.php, dernière consultation 10/2010.

QuébecTop (site web). http ://www.qctop.com/, dernière consultation 08/2010.

Lawrence R. Rabiner & Ronald W. Schafer (1978). *Digital processing of speech signals*. Prentice-Hall, Englewood Cliffs, N.J.

RITEL (site web). 'Recherche d'Information par Téléphone'. http ://ritel.limsi.fr/, dernière consultation 10/2010.

Sophie Rosset, et al. (2007). 'The LIMSI participation to the QAst track'. Dans A. Nardi & C. Peters (eds.), *Working Notes of CLEF Workshop, ECDL conference*, Budapest, Hungary. Springer.

Sophie Rosset, et al. (2006). 'Interaction et recherche d'information : le projet RITEL'. *Traitement Automatique des Langues* **46**(3) :155–179.

Sophie Rosset & Sandra Petel (2006). 'The Ritel Corpus - An annotated Human-Machine open-domain question answering spoken dialog corpus'. Dans *LREC 2006*, Genoa, Italy.

Mario Rossi (1999). *L'intonation : le système du français : description et modélisation*. Editions Ophrys.

José Rouillard (1998). 'Hyperdialogue Homme-Machine sur le World Wide Web : Le système HALPIN'. Dans *Colloque International Ergonomie et Informatique, ERGO'IA*.

David Sadek (1996). 'Le dialogue homme-machine : de l'ergonomie des interfaces à l'agent intelligent dialoguant'. *Nouvelles interfaces hommes-machine in Lavoisier Editeur, Paris, série ARAGO* **18** :277–321.

Helmut Schmid (1994). 'Probabilistic part-of-speech tagging using decision trees'. Dans *Proceedings of International Conference on New Methods in Language Processing*, vol. 12. Manchester, UK.

Kévin Séjourné (2009). *Questions-réponses et interactions*. Thèse de doctorat, Université Paris Sud, Orsay.

Elizabeth Shriberg (2005). 'Spontaneous speech : How people really talk and why engineers should care'. Dans *Ninth European Conference on Speech Communication and Technology*.

V. L. Smith & H. H. Clark (1993). 'On the course of answering questions'. *Journal of Memory and Language* **32** :25–38.

Martin M. Soubbotin & Sergei M. Soubbotin (2001). ‘Patterns of Potential Answer Expressions as Clues to the Right Answers’. Dans *In Proceedings of the Tenth Text REtrieval Conference (TREC 10)*, pp. 293–302.

Agnès Souque (2008). ‘Vers une nouvelle approche de la correction grammaticale automatique’. Dans *Rencontre des Étudiants Chercheurs en Informatique pour le TAL, RECITAL*, pp. 121–130.

Adam J. Sporka, et al. (2009). ‘Can machines call people ? : user experience while answering telephone calls initiated by machine’. Dans *CHI EA ’09 : Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pp. 3625–3630, New York, NY, USA. ACM.

The Correct (site web). ‘Correct Medical Questions & Answers’. http ://www.thecorrect.com/, dernière consultation 08/2010.

Erik F. Tjong Kim Sang & Fien De Meulder (2003). ‘Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition’. Dans W. Daelemans & M. Osborne (eds.), *Proceedings of CoNLL-2003*, pp. 142–147. Edmonton, Canada.

Dave Toney, et al. (2008). ‘An Evaluation of Spoken and Textual Interaction in the RITEL Interactive Question Answering System’. Dans E. L. R. A. (ELRA) (ed.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco.

TREC (site web). ‘Text REtrieval Conference’. http ://trec.nist.gov/.

TREC ciQA task (site web). ‘The TREC complex, interactive QA task’. http ://www.umiacs.umd.edu/ jimmylin/ciqa/, dernière consultation 03/2010.

Jordi Turmo, et al. (2007). ‘Overview of QAST 2007’. Dans *Working notes of Cross Language Evaluation Forum (CLEF)*.

Jordi Turmo, et al. (2008). ‘Overview of QAST 2007’. Dans *Lecture Notes in Computer Science*.

Charlotte van Hooijdonk, et al. (2008). ‘Production and evaluation of (multimodal) answers to medical questions’. Dans *EARLI SIG2 - 2008 Conference on Comprehension of Text and Graphics*, pp. 72–76, Tilburg, The Netherlands.

Boris van Schooten, et al. (2007). ‘Handling speech input in the Ritel QA dialogue system’. Dans *Proceedings of Interspeech’07*.

José Luis Vicedo & Antonio Ferrández (2001). ‘A semantic approach to Question Answering systems’. *NIST SPECIAL PUBLICATION SP* **SP** :511–516.

José Luis Vicedo, et al. (2002). ‘University of Alicante at TREC-10’. *NIST SPECIAL PUBLICATION* **SP** :510–518.

VirtuOz (site web). http ://www.virtuoz.com, dernière consultation 06/2010.

Ellen M. Voorhees (1999). ‘The TREC-8 Question Answering Track Report’. Dans *TREC*.

Joseph Weizenbaum (1966). ‘ELIZA - a computer program for the study of natural language communication between man and machine’. *Commun. ACM* **9**(1) :36–45.

Andi Winterboer, et al. (2008). ‘Do discourse cues facilitate recall in information presentation messages ?’. Dans *InterSpeech*.

Cécile Woehrling (2009). *Accents régionaux en français : perception, analyse et modélisation à partir de grands corpus*. Thèse de doctorat, Université Paris-Sud Orsay.

F. Wolinski, et al. (1995). ‘Automatic Processing of Proper Names in Texts’. Dans *Proceedings of EACL’95*.

Anne Xuereb & Jean Caelen (2004). ‘Un modèle d’interprétation pragmatique en dialogue homme-machine basé sur la SDRT’. Dans *Journées scientifiques Sémantique et Modélisation*, Lyon.

Yahoo QR (site web). ‘Yahoo ! France question réponses’. http ://fr.answers.yahoo.com/, dernière consultation 06/2010.

NOTATIONS

EXEMPLES

Les exemples numérotés [A###] sont des réponses issues du corpus MACAQ.
Les exemples numérotés [Q###] sont des questions issues du corpus MACAQ.

ACRONYMES

QR	Question-Réponse
SQR	Système de réponse à une question, plus communément système de question-réponse
TAL	Traitement automatique des langues
POS	Part-of-speech, partie du discours
MUC	Message Understanding Conference
FAQ	Foire Aux Questions ou Frequently Asked Questions

LEXIQUE FRANÇAIS-ANGLAIS

Ce court lexique français-anglais ne recherce pas l'exhaustivité. Il regroupe un ensemble de mots ou expressions que j'ai rencontré au cours de ma thèse (dans des articles, lors de conférences,...). Ils ne font pas tous strictement partie du domaine dans lequel cette thèse s'inscrit. Certains m'ont donné du fil à retordre (que peut bien signifier LVASR ?), d'autres sont ici car on en cherche toujours une traduction français alors même qu'on passe notre temps à les utiliser en anglais. Bref, l'objectif de ce petit lexique, avant tout, est de donner un coup de pouce à de jeunes doctorants perdus entre deux langues et un nouveau domaine à découvrir.

Anne GF

actes (de conférence)	proceedings
apprentissage automatique	machine learning
campagne d'évaluation	evaluation campaign
compréhension de la langue naturelle	natural language understanding (NLU)
détection de la parole	speech activity detection (SAD)
données brutes	raw data
énoncé-réponse	answering-utterance
entité nommée (EN)	named entity (NE)
étiquette (d'annotation)	tag, label
foire aux questions (FAQ)	frequently asked questions (FAQ)
forme (mot)	lexical type, distinct lexical items (word)
génération automatique de texte (GAT)	natural language generation (NLG)
hésitation	filler word
implication textuelle	textual entailment
information-réponse	answering-information
inter-langue	cross-language
interuption, coupure	barge-in
langue peu dotée	low-ressourced language
mot outil	function word
mot plein	content word
occurrence d'une forme	lexical item, token

parole spontannée	spontaneous speech
partie du discours	part-of-speech (POS)
patrons	templates
question-réponse (QR)	question-answering (QA)
reconnaissance automatique de la parole	automatic speech recognition (ASR)
reconnaissance de la parole à large vocabulaire	large vocabulary automatic speech recognition (LVASR)
segmentation, analyse en constituants	chunking
stage	internship
synthèse de la parole ou synthèse vocale (TTS)	text-to-speech synthesis (TTS)
système de dialogue (oral) homme-machine (SDHM, SDOHM)	(spoken) dialogue system (DS, SDS)
système de question-réponse (SQR)	question-answering system (QAS)
tâche (d'une campagne d'évaluation)	track (of an evaluation campaign)
texte pré-écrit	canned-text
traitement automatique des langues naturelles (TAL)	natural language processing (NLP)

Ce document a été préparé à l'aide de l'éditeur de texte GNU Emacs et du logiciel de composition typographique LaTeX 2_ε.

Titre Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert

Résumé Les travaux présentés dans ce mémoire se situent dans le contexte de la réponse à une question. Contrairement à de nombreux travaux traitant de la recherche de l'information à fournir en réponse à une question, notre problématique principale a été de caractériser la forme que peut prendre une réponse en interaction avec une question qui puisse être produite par des systèmes de question-réponse.

Nous exposons les enjeux de l'interaction du type "réponse à une question" considérant deux modalités d'interaction : l'oral et l'écrit. Nous montrons que répondre n'est pas uniquement présenter une information mais fait partie d'une interaction entre deux locuteurs. Cherchant à définir ce que pourrait être une *réponse en interaction* pour les systèmes de question-réponse, nous constatons l'absence de corpus constitué de telles réponses. Dans l'optique de constituer un tel corpus, la forme des questions utilisées lors de la collecte est primordiale. Une étude de l'état de l'art sur les variations linguistiques des questions est ainsi présentée.

Nous exposons ensuite la constitution des questions ainsi que la collecte du corpus de réponses à l'oral et à l'écrit, et effectuée auprès de plus de 150 locuteurs natifs du français. Une évaluation du protocole utilisé est ensuite effectuée.

Enfin, nous présentons une analyse du corpus collecté en répondant à un ensemble de questions préalables à création d'un module de génération de réponses en langue naturelle dans un système de question-réponse.

Mots-cléfs Question-réponse, génération de réponse en langue naturelle, acquisition de corpus

Title Generation of Speech and Written Answers in Natural Language for Question-Answering Systems in Open-Domain

Abstract The work presented in this thesis is in the context of question answering. Unlike many works dealing with information retrieval, our main problem is to characterize the possible form of an answer in interaction with the question and which could be produced by question answering systems. We present the case of answering a question considering two interaction modes : orally and in writing. Trying to define what could be an "in interaction answer" for question answering systems, we note the absence of a corpus of such answers. In order to constitute our own answers corpus, the question linguistic form used is important. A state-of-the-art study of questions linguistic variations is presented so. We present the corpus collection, both on oral and written modalities and performed with over 150 french native speakers, and an evaluation of the used protocol. Finally, we present an analysis of the collected corpus answering questions to be asked to implement a module for generating in natural language answers in a question-answering system.

Keywords Question-answering, natural language generation, corpus acquisition