Évaluation transparente du traitement des éléments de réponse à une question factuelle

Sarra El Ayari

Thèse soutenue le 23 novembre 2009
pour l'obtention du Doctorat de l'Université Paris-Sud 11
(spécialité informatique)

Composition du jury :

Président	Joseph Mariani
Directeurs	Brigitte Grau
	Benoît Habert
Rapporteurs	Patrice Bellot
	Guy Lapalme
Examinateurs	Andrei Popescu-Belis
	Sophie Rosset

Résumé

Les systèmes de questions-réponses permettent à un utilisateur de rechercher une information précise dans un corpus de données massif, comme le *Web*. Ce sont des systèmes complexes qui utilisent des techniques de traitement automatique des langues (TAL). Des campagnes d'évaluation sont organisées pour évaluer leur performance finale, mais les équipes de recherche doivent ensuite réaliser des évaluation de diagnostic pour savoir les raisons de leurs succès et de leurs échecs. Or, il n'existe ni outil, ni méthode pour réaliser des évaluations systématiques de critères linguistiques pour de tels systèmes.

L'objectif de ce travail est de proposer une méthodologie d'évaluation transparente des résultats intermédiaires produits par les systèmes de questions-réponses, en combinant à la fois une évaluation de performance et une analyse de corpus. Ainsi, nous discuterons de l'évaluation telle qu'elle est menée pour ces systèmes, et des limites rencontrées pour une évaluation de diagnostic.

Dans un premier temps, nous nous interrogerons sur les pratiques d'évaluation, qu'elles portent sur les résultats finaux d'un système ou bien sur ceux que produisent les différents composants dans l'optique de dégager les tenants et les aboutissants d'une évaluation plus fine des stratégies linguistiques mise en œuvre au sein des systèmes de questions-réponses. Cette étude nous permettra de dégager les principes d'une méthodologie d'évaluation de diagnostic transparente pour les systèmes de questions-réponses.

Dans un deuxième temps, nous nous sommes intéressée aux erreurs classiques d'un système de questions-réponses afin de détailler les fonctionnalités nécessaires à un outil de diagnostic systématique de ces erreurs. Ceci nous a conduit à la création d'un outil d'évaluation, REVISE (*Recherche, Extraction, VISualisation et Evaluation*), qui permet de stocker les résultats intermédiaires d'un système de façon à en disposer pour les annoter, les modifier, les visualiser et les évaluer. Nous avons également discuté la généricité de cet outil à l'aide des résultats du système de questions-réponses RITEL.

Enfin, nous avons mené à l'aide de notre outil deux types d'études sur les systèmes de questions-réponses FRASQUES et QALC, l'une portant sur le critère linguistique focus extrait lors de l'analyse des questions et sur ses variations en contexte dans les phrases réponses sélectionnées par le système ; l'autre sur l'application des règles d'extraction de réponses précises.

Sommaire

Table des figures

Liste des tableaux

Introduction

L'évaluation transparente de systèmes de recherche d'information

Les systèmes de recherche d'information sont des outils qui permettent de rechercher une information dans un flot de données non structuré tel le Web. Leur fonctionnement diffère en fonction du matériau sur lequel ils travaillent ainsi que sur leur utilité. Nous pouvons distinguer trois types d'outils :

- les moteurs de recherche, qui travaillent sur l'Internet principalement et permettent de rechercher de l'information liée aux mots clés saisis par l'utilisateur,
- les systèmes de questions-réponses, qui permettent de rechercher de l'information précise à partir d'une question posée à l'écrit en langage naturel,
- les systèmes de dialogue, qui recherchent une réponse précise à une question posée à l'oral en langage naturel.

Évaluer un système contenant plusieurs composants est une tâche complexe, qui nécessite de prendre en compte les stratégies utilisées ainsi que l'architecture. Or, il n'y pas d'architecture standard pour les systèmes de recherche d'information en général, pas plus que de consensus en ce qui concerne les méthodes à utiliser. Des campagnes d'évaluation sont organisées chaque année et permettent aux participants de les mesurer sur une tâche donnée afin de les classer les uns par rapport aux autres. C'est ce que l'on appelle l'évaluation de type *boîte noire*.

De telles campagnes remplissent leur fonction d'évaluation et de comparaison des systèmes entre eux, mais ne permettent pas aux équipes de recherche d'obtenir un savoir réel sur les défauts et les qualités de leurs systèmes. Il s'agit d'une vision globale de l'ensemble qui, par définition, n'est pas précise. Or pour améliorer un système modulaire, il faut savoir ce qui fonctionne et ce qui pose problème précisément.

Cette précision lors de l'évaluation des résultats est d'autant plus importante pour les systèmes qui incorporent des stratégies linguistiques et qui reposent sur des techniques de

traitement automatique des langues (TAL), plutôt que statistiques : ils nécessitent une analyse précise des phénomènes linguistiques en jeu ainsi que de leurs traitements afin d'affiner les stratégies, ce que l'on appelle évaluation de type *boîte transparente*.

Nous nous intéressons plus particulièrement aux systèmes de questions-réponses qui sont des systèmes complexes de recherche de réponses précises utilisant des techniques de traitement automatique des langues. Notre position ne consiste pas uniquement à avoir accès aux différents modules qui constituent ces systèmes pour les évaluer. Nous nous situons dans une démarche d'analyse de corpus basée sur la visualisation, l'annotation et la modification de résultats intermédiaires produits par ces systèmes.

Problématique

Un système de questions-réponses permet de répondre à une question posée en langage naturel (*Quel est le premier homme à avoir marché sur la lune ?*) par une réponse précise (*Neil Armstrong*).

Si les campagnes d'évaluation menées à grande échelle sur ces systèmes s'intéressent essentiellement aux résultats finaux produits pour les classer, il n'y a pas de méthodes pour évaluer les critères linguistiques qui sont mis en œuvre pour rechercher des réponses. De plus, il n'y a pas non plus de corpus dédiés à certains phénomènes au sortir des campagnes d'évaluation.

Pour répondre à ces manques, nous avons développé une méthodologie d'évaluation fine de critères linguistiques qui se base sur l'observation de sous-corpus représentatifs de phénomènes intéressants. Pour ce faire, nous avons développé un outil qui permet à la fois la création de sous-corpus d'étude, la visualisation des données produites et l'évaluation les résultats à différentes étapes de la chaîne de traitement.

Principales contributions

Nous avons développé une méthodologie d'évaluation des systèmes de questions-réponses, basée sur l'analyse de leurs résultats. Cette méthodologie se veut transparente puisqu'il s'agit de réaliser une analyse de corpus des résultats de systèmes de questions-réponses associée à une évaluation de performance des systèmes.

Pour ce faire, nous avons créé un outil de constitution de sous-corpus, d'annotation et de visualisation de résultats afin de faciliter l'évaluation transparente des systèmes de questions-réponses. REVISE est l'acronyme de *Recherche, Extraction, VISualisation et Evaluation* et permet d'évaluer des résultats de systèmes de questions-réponses dans un cadre d'utilisation de techniques de traitement automatique des langues. Il est conçu autour d'une base de données relationnelle, laquelle est interrogée pour sélection finement les données que l'utilisateur veut observer. De façon à justifier la conception de notre outil et démontrer son adaptabilité, nous avons présenté les principes d'une démarche d'évaluation générique, basée sur les résultats d'un autre système de questions-réponses que celui sur lequel l'outil a été développé.

Nous avons également présenté différentes méthodes d'évaluation transparente de problèmes qui relèvent du domaine de la linguistiques : le critère focus et les règles d'extraction de réponses précises, à l'aide de REVISE. Nous avons défini et validé la notion de focus (événement exprimé dans une question) sur le système de questions-réponses FRASQUES et nous avons réalisé une étude de la variation linguistique de ce focus, afin de déterminer l'importance de la prise en compte des variations de façon automatique.

Une deuxième étude a été menée sur les règles d'extraction de réponses précises du système QALC, qui reposent sur les critères linguistiques extraits de la question (focus, type général, verbe principal) avec un souci particulier accordé à la visualisation des ces phénomènes pour mesurer l'application effective des règles créées.

Plan de la thèse

Dans la première partie de ce travail, nous discutons des pratiques d'évaluation pour les systèmes modulaires selon différents paradigmes afin de proposer une redéfinition de l'évaluation transparente. Nous décrivons ensuite le fonctionnement du système de questions-réponses FRASQUES sur lequel nous avons basé nos études et montrons l'importance du traitement de la variation en langue pour de tels systèmes. Enfin, nous proposons un état de l'art des évaluation transparentes appliquées aux systèmes de questions-réponses, afin de mieux situer notre approche.

Dans la deuxième partie, nous présentons une analyse des erreurs effectuée sur le système FRASQUES, de façon à montrer les points importants à évaluer. Nous avons pour ce faire développé un outil, REVISE (acronyme de Recherche, Extraction, VISualisation

et Évaluation), de visualisation et d'évaluation fines des résultats intermédiaires produits par un système de questions-réponses. Nous décrivons en détail son architecture et son fonctionnement, et discutons des aspects liés à la généricité et à la spécificité de l'outil. Afin de mesurer efficacement la généricité de notre outil, nous avons réalisé une étude un système de questions-réponses différent : RITEL.

La troisième partie de ce travail rassemble les différentes études linguistiques que nous avons menées grâce à notre outil. La première étude porte sur la notion de focus, que nous avons définie et validée en corpus, tout en s'intéressant à la variation de sa forme de surface au sein des phrases réponses . Dans un deuxième temps, nous nous sommes intéressée à l'extraction de réponses précises, afin d'affiner les règles suite à l'étude de corpus réalisée.

Chapitre 1

Comment évaluer un système modulaire ?

Sommaire

1.1 Les pratiques d'évaluation

La recherche, quels que soient les domaines, nécessite d'évaluer les techniques employées, de mesurer les performances des systèmes, les avancées et ce qu'il reste à accomplir. De

façon à encourager les progrès, des campagnes d'évaluation sont organisées qui mettent en compétition différents systèmes et comparent les résultats qu'ils obtiennent. C'est monnaie courante en ce qui concerne les systèmes de traitement de l'information (moteurs de recherche, résumés automatiques, traduction automatique, systèmes de questions-réponses, etc.).

1.1.1 Historique

Si l'on veut dresser un historique des pratiques d'évaluation pour les systèmes de traitement de l'information, il faut remonter au milieu du vingtième siècle. Des campagnes d'évaluation ont commencé à voir le jour afin de comparer les systèmes existants sur une même tâche [Sparck Jones et Galliers, 1996]. La naissance des systèmes de recherche d'information (*information retrieval*) coïncide avec la profusion des informations sur Internet et, de ce fait, l'apparition de la nécessité de disposer d'outils performants pour avoir accès à l'information recherchée. Dans le même temps, l'évaluation des performances de ces nouveaux systèmes est devenue un enjeu important.

Si nous parlons d'enjeux ici, c'est que l'évaluation est au centre des préoccupations des institutions politiques, scientifiques et industrielles [Timim, 2006] et que les intérêts liés au développement des systèmes de recherche d'information (SRI) sont réels. Les premières campagnes d'évaluation ont été organisées par des infrastructures comme le NIST (*National Institute of Standards and Technology*), la DARPA (*Defense Advances Research Projects Agency*) ou encore le LDC (Linguistic Data Consortium), principalement aux États-Unis.

Nous présentons deux campagnent qui illustrent les pratiques d'évaluation de système de recherche d'information, puis centrons notre propos autour de l'évaluation de systèmes plus complexes.

Les débuts de l'évaluation de systèmes de traitement de l'information

Le projet Cranfield est le premier projet d'évaluation des systèmes d'indexation des documents (processus clé des moteurs de recherche). Il a été organisé en 1957 et son rôle a été déterminant en ce qui concerne la façon d'évaluer des systèmes de recherche d'information. La tâche consistait à comparer l'indexation et la recherche de documents entre différents systèmes.

En terme méthodologique, la tâche organisée par le projet Cranfield consistait à réaliser une comparaison entre les réponses attendues (1 200 requêtes en tout) au sein

des articles (18 000 articles) et celles extraites par les systèmes. Le critère le plus important était celui de la pertinence, qui permet de mesurer en quoi le système satisfait au besoin informationnel de l'utilisateur. Ce critère repose sur deux mesures : le rappel et la précision, le rappel indiquant la capacité du système à sélectionner les documents pertinents et la précision sa faculté à rejeter les documents non pertinents. Les mêmes principes d'évaluation sont encore utilisés pour les systèmes de RI aujourd'hui [Voorhees, 2002].

TREC est l'acronyme de *Text REtrieval Conference*. La campagne débute en 1992 et prend la suite du projet Cranfield. Il s'agit également d'une plateforme de tests basée sur l'indexation, mais elle comporte des données plus importantes avec plus de 500 mégaoctets de corpus à indexer dès la première campagne [Voorhees et Harman, 2005]. Le cadre d'évaluation (anglophone) proposé a le mérite de disposer de requêtes ainsi que des documents pertinents par rapport à ces requêtes. Les mesures d'évaluation utilisées sont les mêmes que pour Cranfield : rappel et précision, auxquelles d'autres se sont ajoutées au fil du temps pour rendre compte des différentes capacités des systèmes.

L'utilisation des mesures de performance permettent de mesurer les résultats finaux des systèmes. Ce que ces deux campagnes montrent autant au niveau du succès qu'elles obtiennent en terme de participants qu'au niveau des défis qui poussent les avancées, c'est que l'on a besoin de structures compétentes avec des paradigmes d'évaluation clairs pour dresser une cartographie de la recherche appliquée qui permette d'avancer sur ces problématiques en dehors des campagnes.

Le deuxième aspect intéressant de ces campagnes d'évaluation est qu'elles ont permis de constituer des corpus de taille importante. Il s'agit de corpus de référence qui sont représentatifs de la tâche et permettent de tester les systèmes en dehors des campagnes. C'est un apport considérable pour la recherche dans le domaine pour le développement des systèmes.

Par la suite, l'apparition de systèmes plus complexes en terme d'architecture a fait émerger de nouvelles questions concernant les évaluations que nous avons présentées. Il s'agit de systèmes composés de plusieurs modules de traitement séquentiels, que nous allons présenter.

Évaluation de systèmes modulaires

L'évaluation de systèmes modulaires doit prendre en compte des traitements multiples de données, à des niveaux différents. Nous nous intéressons particulièrement aux systèmes de traitement automatique des langues (TAL). L'évaluation de ces systèmes est une tâche malaisée qui ne cesse de poser question à la communauté concernée. Quel que soit le domaine d'application (résumé et traduction automatiques, correction orthographique, synthèse de la parole, reconnaissance vocale ou encore classification et catégorisation de textes), la question de l'évaluation de la pertinence des méthodes linguistiques mises en œuvre dans ces systèmes se pose.

« Le traitement automatique des langues (TAL) relève à la fois de la démarche scientifique et de la démarche technologique. Dans les deux cas, l'évaluation des systèmes informatiques implémentés est indispensable pour estimer le succès d'une recherche » [Popescu-Belis, 2007]. Ici, l'auteur met en avant la difficulté de l'évaluation en TAL, qui réside notamment dans l'impossibilité de créer des spécifications formelles permettant l'évaluation de tâches qui utilisent des données linguistiques. Ainsi, les seules évaluations possibles font appel à des métriques traditionnelles, qui évaluent l'adéquation du système par rapport à la tâche demandée.

Certaines campagnes d'évaluation sont organisées pour évaluer des techniques de traitement automatique des langues. Nous en décrivons trois, qui portent sur des domaines différents : MUC[1] (*Message Understanding Conferences* pour l'extraction d'informations, EASY[2] (*Évaluation des Analyseurs SYntaxiques du français*), devenu PASSAGE[3], pour l'analyse syntaxique et GRACE[4] (*Grammaire et Ressources pour les Analyseurs de Corpus et leur Évaluation*) pour l'annotation morpho-syntaxique.

MUC est une campagne organisée par le DARPA autour de l'extraction d'information. Le souci principal des organisateurs était de développer des composants informatiques d'extraction d'information indépendants d'une tâche de façon à ce qu'ils puissent être réutilisables et adaptables à différents systèmes de recherche d'information, de façon générique, en développant des sous-tâches spécifiques à certains types de problèmes. La tâche entité nommée, créée lors de la campagne MUC-6, témoigne de cette volonté et des besoins pour les systèmes de traitement de l'information de reconnaître les

[1] Voir http ://www.itl.nist.gov/iad/894.02/related_projects/muc/
[2] Voir http ://www.limsi.fr/Recherche/CORVAL/easy/
[3] Voir http ://atoll.inria.fr/passage/
[4] Voir http ://www.limsi.fr/RS99FF/CHM99FF/TLP99FF/tlp10/

entités dans les textes [Poibeau, 2005]. On pourra citer également l'analyse de la co-référence. En ce qui concerne les métriques utilisées, on retrouve le rappel et la précision pour évaluer les techniques mises au point.

EASY est un projet d'évaluation d'analyseurs syntaxiques pour le français. Les principaux problèmes liés à leur évaluation est qu'en général les systèmes disposent de sorties de formats différents et qu'il est difficile d'obtenir des métriques qui soient satisfaisantes et réellement équitables [Vilnat *et al.*, 2004]. Cette évaluation est constituée de quatre phases : la constitution du corpus, l'annotation manuelle d'un échantillon qui est la référence, la définition d'un formalisme unique des résultats et l'adoption des métriques d'évaluation de rappel et de précision.

Ce qui est intéressant en terme d'outils, c'est qu'une plate-forme de visualisation et d'annotation manuelle a été créée, de même qu'un outil de validation permettant une visualisation précise des relations étiquetées. Ce besoin d'outils pour revenir à une dimension observable des phénomènes nous paraît indispensable pour évaluer finement les qualités et les défauts d'un système.

La campagne GRACE a commencé en 1994 avec pour objectif l'évaluation d'analyseurs morpho-syntaxiques du français. Afin de comparer différents analyseurs, il est nécessaire de constituer un jeu d'étiquettes de référence que chacun des participants doit respecter. Les mesures adoptées sont la précision (capacité du système à fournir une étiquette morpho-syntaxique correcte à une forme donnée) et la décision, qui est « la plus ou moins grande capacité d'un système à fournir, dans le formalisme de référence, une étiquette totalement désambiguïsée (des réponses partiellement ambiguës sont admises) » [Adda *et al.*, 1998].

Les campagnes d'évaluation présentées montrent le besoin de créer des paradigmes réalistes d'évaluation pour des systèmes utilisant des techniques liées au TAL. Néanmoins, il est difficile d'associer analyses poussées des résultats et évaluations finales : « en dehors des campagnes d'évaluation, le développement d'un système implique des phases d'évaluation individuelle qui lui sont spécifiques dans la mesure où elles visent à donner des indications précises sur la validité des principes que le système met en œuvre ainsi que sur ses perspectives d'évolution » [Ozdowska, 2007]. Voilà ce que ne permettent pas les campagnes d'évaluation et ce que devrait fournir une évaluation idéale.

Nous présentons maintenant différents types d'évaluation qu'il est possible de réaliser sur un système, afin de préciser notre position.

1.1.2 Différents paradigmes d'évaluation

Évaluer un système suppose de savoir ce que l'on veut évaluer et de quel point de vue. En effet, les campagnes organisées dans les domaines de la recherche d'information privilégient une évaluation des résultats finaux obtenus par le système (évaluation *boîte noire*) de façon à pouvoir comparer des systèmes entre eux sur une base commune (les techniques utilisées varient en fonction des systèmes).

Néanmoins, une évaluation peut porter sur autre chose que sur la notion de performance du système. Nous reprenons ici la distinction faite par [Hirschman et Thompson, 1997], reprise notamment par [Chaudiron, 2004a], en terme de typologie d'évaluation : l'évaluation de progression, l'évaluation de diagnostic et l'évaluation de mise en adéquation. Ces trois notions permettent de balayer différents enjeux de l'évaluation.

- Évaluation de progression (*performance evaluation*) : mesure de la performance du système pour un domaine donné.

  Elle consiste à mesurer la performance du système, c'est-à-dire l'aspect quantitatif des résultats obtenus. C'est le type d'évaluations réalisées par la campagne TREC par exemple, qui mesure un système en calculant la précision des réponses qu'il a fournies (nombre de réponses correctes obtenues par rapport au nombre de réponses totales). Il s'agit essentiellement d'établir un score en fonction des résultats finaux obtenus par le système.
- Évaluation de mise en adéquation (*adequacy evaluation*) : adéquation des résultats obtenus par le système en fonction de la tâche elle-même.

  Elle est liée à la satisfaction de l'utilisateur, ainsi que la pertinence des résultats du système par rapport à la tâche. Cela correspond aux évaluations centrées sur la satisfaction de l'utilisateur ou encore concernant l'ergonomie des interfaces présentées. On pourra lire [Chaudiron, 2004b] pour plus de précisions dans ce sens. Si les évaluations se sont centrées sur la performance globale, il est nécessaire pour des systèmes interactifs de prendre en compte la gestion de l'interaction avec l'utilisateur. Dans ce cadre, de nouveaux paradigmes d'évaluation ont émergé, notamment pour les systèmes de dialogue [Devillers *et al.*, 2003]. Ces systèmes sont comparables à des systèmes de questions-réponses dans le sens où ils permettent d'effectuer une recherche d'information précise en partant d'une question et en fournissant une réponse concise à l'utilisateur. En revanche, ils travaillent sur des données orales et incluent un dialogue avec l'utilisateur afin d'affiner sa requête, ou bien d'en créer une autre. Cette discussion, qu'on appelle un scénario, suppose que le système soit capable de

faire face aux questions de l'utilisateur et donc à l'expression libre de son besoin d'information.

Ainsi, évaluer un tel système sans prendre en compte sa capacité de dialogue ne comporte pas de sens véritable. On touche alors à la complexité de l'évaluation automatique : quels référents utiliser, quels corpus, comment créer artificiellement une situation de dialogue qui permet d'évaluer les stratégies de plusieurs systèmes afin de les comparer ?

– Évaluation de diagnostic (*diagnostic evaluation*) : évaluation de l'état du système.
Elle permet de mesurer les performances du système en interne, ainsi que d'analyser les erreurs présentes. Le diagnostic d'un système est réalisé par les équipes de recherche pour améliorer leurs systèmes et déceler les problèmes. Néanmoins, il n'existe pas de méthodologie d'évaluation des systèmes, ni d'outil pour faciliter le diagnostic.

On notera aussi une distinction entre évaluation d'un système dans sa globalité (*whole-system evaluation*) ou bien de ses composants (*component-wise evaluation*), qui va de pair avec une approche qualitative et/ou descriptive (*Comment le système fait ce qu'il fait ?*) s'opposant à une approche quantitative et/ou analytique (*Quelle est la performance réalisée par le système en faisant ce qu'il fait ?*)

Dans [Saracevic, 1995], l'auteur effectue une analyse critique et historique des évaluations portant sur les systèmes de recherche d'information. Il soulève un problème majeur de l'évaluation : le fait qu'elle n'opère que sur un seul niveau donné. Or, il différencie pas moins de six niveaux à prendre en compte lors de l'évaluation d'un système :

– Niveau d'ingénierie (*engineering level*) : erreurs, vitesse, maintenance, flexibilité, etc.
– Niveau des flux d'entrée (*input level*) : représentativité du corpus.
– Niveau du processus (*processing level*) : performance, techniques et approches utilisées.
– Niveau des flux de sortie (*output level*) : interactions avec l'usager, retours.
– Niveau des usages (*use and user level*) : ergonomie de l'interface.
– Niveau social (*social level*) : recherche, productivité, prise de décision.

En effet, si la plupart des évaluations et des études sont organisées au niveau du processus, il paraît essentiel de prendre en compte toute la dimension d'un système (représentée par ces six niveaux) pour réellement réaliser une évaluation complète. Les trois types d'évaluation explicités précédemment se retrouvent plus ou moins sur ces niveaux. En effet, une évaluation de performance se tiendra au niveau du processus, une évaluation de mise en adéquation est liée au niveau des flux d'entrée et de sortie, ainsi que des usages. L'évaluation de diagnostic se situerait plus au niveau processus également, c'est-à-dire en amont des

métriques et des scores. Le découpage par niveaux est intéressant, et présente l'ensemble des plans à évaluer si l'on veut obtenir un diagnostic complet. Néanmoins, ces niveaux manquent de précision à notre sens et mettre l'analyse des erreurs dans le niveau ingénierie montre une approche en terme d'évaluation de performance des erreurs. Ce modèle ne semble donc pas représenter l'analyse de diagnostic.

Dans notre travail, nous nous intéressons plus précisément à l'évaluation de diagnostic appliquée à des systèmes modulaires. Notre position diffère des évaluations habituelles, qui évaluent les systèmes « par rapport à un résultat de référence plutôt que sur la dissection et l'analyse détaillée du comportement de chaque composant de chaque application » [Habert et Zweigenbaum, 2002]. Une évaluation de diagnostic suppose une évaluation différente de celles qui sont menées lors des campagnes d'évaluation. Nous allons discuter des points à prendre en compte.

1.1.3 Quelques réflexions autour des campagnes

Ces différentes facettes de l'évaluation montrent l'importance d'évaluer les systèmes de TAL de façon à améliorer leurs performances et leur adéquation avec ce à quoi ils servent. La mise en œuvre d'évaluations pour ces systèmes est une tâche qui nécessite des ressources adaptées et des mesures précises pour évaluer les résultats.

L'exhaustivité de la tâche

Une question importante concerne le jeu de données, le corpus, utilisé pour évaluer les systèmes. Un des enjeux des campagnes d'évaluation réside dans la volonté de coller au plus proche du contexte d'application réel des systèmes de façon à ne pas biaiser l'évaluation réalisée. Le corpus doit être représentatif de la tâche effective, ce qui pose des problèmes réels. Cette exhaustivité demande de créer le corpus, en veillant à la présence des phénomènes que l'on veut étudier. Par exemple, le corpus ne doit pas être trop petit afin que la variabilité de la langue soit présente et assez fourni pour que sa prise en compte constitue un enjeu de recherche. La constitution d'un jeu de données de référence permet la comparaison de différentes techniques, mais est-ce que la qualité d'un corpus est tributaire de sa taille ?

La littérature sur ce qu'est un corpus et comment le constituer est très présente, notamment chez les linguistes. On pourra regarder [Péry-Woodley, 1995] et [Mellet, 2003] dans

ce sens. Ils se sont posé la question de l'échantillonnage, et de savoir ce qu'on échantillonne en terme d'exhaustivité et de représentativité du corpus. Si la question reste complexe, il est clair à présent qu'avec la disponibilité de corpus comme le Web ou bien la *Wikipedia*, les corpus de grande taille attirent la recherche. Ainsi se pose désormais le problème de la validité des approches mises en œuvre pour les traiter de façon automatique et conséquemment du besoin d'avoir accès à des sous-parties de façon à visualiser les phénomènes en jeu et leur traitement.

De plus, la constitution de corpus pour une évaluation ciblée pose de réels problèmes. En effet, comment créer un corpus adéquat pour une tâche donnée ? Nous reviendrons sur ce point, qui est fondamental, de constitution de sous-corpus créés en fonction des difficultés qu'ils présentent.

Quel diagnostic pour quelles évaluations ?

Si les campagnes d'évaluation permettent de faire avancer les systèmes, elles sont également sources de publications. Chaque participant écrit un article qui rend compte des stratégies adoptées et des gains et des pertes qu'elles engendrent. Ces réflexions sont très intéressantes et permettent de mesurer les difficultés qui se posent quand on traite automatiquement le langage.

Ainsi, une fois un système classé lors d'une évaluation, si les développeurs veulent connaître les raisons de leurs échecs, ils doivent passer par une phase de diagnostic et d'évaluation précise de leur système, d'analyse des erreurs. Il s'agit d'identifier dans les résultats obtenus, au sein des différents modules ce qui n'a pas été traité correctement, les phénomènes linguistiques posant problème par exemple (variations). C'est ce type de savoir précis, obtenu par un diagnostic des résultats eux-mêmes en relation avec les composants qui les ont produits et ceux qui les ont précédés, qui va permettre de changer, d'affiner une stratégie ou bien de modéliser un phénomène.

Si la chose est valable pour n'importe quel système, cette observation minutieuse des résultats l'est encore plus pour les systèmes qui utilisent des techniques de traitement de la langue, lesquels se basent sur des règles fines de traitement du langage et souffrent le plus souvent des irrégularités rencontrées.

Les campagnes d'évaluation sont de précieux outils pour dessiner une carte de la recherche et des avancées des systèmes de recherche et d'extraction d'informations. Néanmoins, si elles offrent une estimation des capacités des systèmes, elles ne permettent pas de

dégager un savoir réel sur les obstacles que les systèmes rencontrent. Elles ne suffisent pas à déceler les problèmes ; il faut regarder à l'intérieur des systèmes pour cela, notamment en analysant plus finement les résultats obtenus.

Nous allons maintenant nous intéresser plus particulièrement aux systèmes de questions-réponses, qui sont des systèmes modulaires utilisant des techniques de traitement automatique des langues.

1.2 Les systèmes de questions-réponses

Les systèmes de questions-réponses sont des systèmes modulaires qui mettent en jeu des phénomènes linguistiques variés, notamment pour l'appariement entre questions et réponses [Grau, 2004a]. Si l'on veut améliorer un système de questions-réponses une fois un niveau de performance atteint, il faut passer par une évaluation de diagnostique poussée. Afin de montrer la nécessité de ce type d'évaluation, nous décrivons le fonctionnement d'un système de questions-réponses et présentons comment les campagnes d'évaluation cherchent à évaluer ces systèmes.

1.2.1 Description

Un système de questions-réponses (SQR) permet à un utilisateur de poser une question en langage naturel et de lui fournir en retour une réponse précise : *Quel est le menu préféré de Gaston Lagaffe ?* attendra la réponse *morue aux fraises avec du cabillaud à l'ananas et des crêpes montgolfières.*

En effet, un SQR est composé au minimum de quatre modules, qui sont l'analyse des questions, la recherche et la sélection des documents, la sélection des phrases et l'extraction de la réponse. Chaque module renvoie à une discipline particulière. Par exemple, l'analyse des questions repose sur des techniques liées au traitement automatique des langues (règles syntaxiques et sémantiques). Les SQR sont des systèmes modulaires qui utilisent des techniques relevant du domaine de la recherche d'information pour retrouver des documents pertinents par rapport à la question ainsi qu'au domaine de l'extraction d'information pour extraire la réponse précise attendue. La figure 1.1 montre les domaines auxquels les SQR sont liés.

Généralement, les pratiques d'évaluation mises en œuvre sont du même type que ceux des campagnes présentées précédemment, avec la possibilité de réaliser des évaluations de

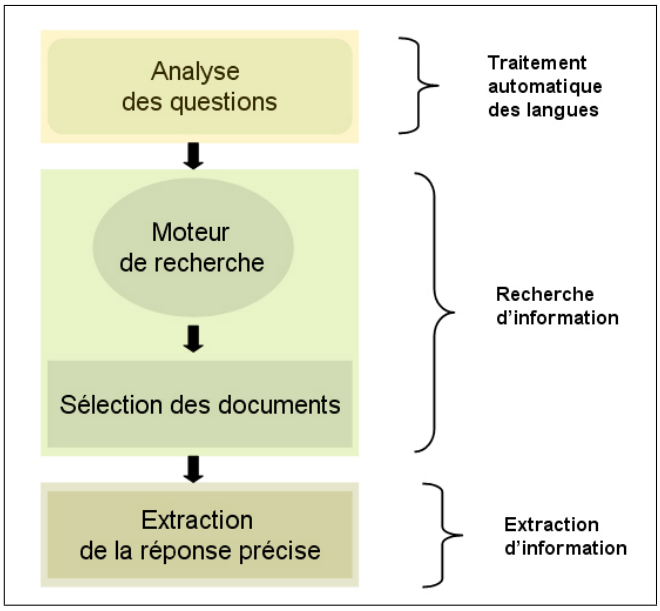


Fig. 1.1 SQR, entre recherche et extraction d'information

performance sur les résultats produits par le système dans son ensemble ou bien sur chacun des modules.

En revanche, les évaluations possibles ne permettent pas de réaliser d'études fines sur la variabilité qui est en jeu, notamment pour l'application de techniques TAL.

La variabilité en langue

La problématique inhérente aux systèmes de questions-réponses réside dans la difficulté de trouver des réponses en partant d'une question. En effet, on se heurte ici assez rapidement à la variabilité de la langue et au fait qu'il y a de multiples façons d'exprimer une même information.

Le principe des systèmes de questions-réponses est d'identifier les éléments importants présents dans la question afin de rechercher au mieux les phrases réponses potentielles. Néanmoins, il est rare de trouver une phrase réponse qui soit une reformulation déclarative de la question. Les phénomènes de variation linguistique rendent l'appariement d'une question avec différentes phrases qui contiennent la réponse plus ou moins complexe.

Nous allons présenter l'exemple d'une question (issue de la campagne d'évaluation CLEF07) afin de bien mesurer ce problème. La question est *Quel évêque fut suspendu par le Vatican le 13 janvier 1995 ?* et voici une liste non exhaustive de passages qui contiennent

la réponse[5] :

1. **Jacques Gaillot**, 59 ans, avait été démis le [13 janvier] par le [Vatican] de la charge d'[évêque] d'Évreux dont il était titulaire depuis 1982.
2. Une délégation de la conférence épiscopale française sera reçue au [Vatican] par le pape, vendredi 3 mars, pour faire le point sur les réactions qui ont suivi en France la destitution, le [13 janvier], de **Mg Jacques Gaillot**, ancien [évêque] d'Évreux.
3. Le [Vatican] justifie sa décision de [suspendre] **Mgr Gaillot**.
4. De son côté, le [Vatican] a expliqué vendredi avoir démis **Mgr Gaillot** en raison de son manque d'orthodoxie sur des sujets comme le sida ou les droits des travailleurs.
5. Dans la délégation du [Vatican] qui suit le pape à Manille, rapporte notre envoyé spécial Henri Tincq, personne ne se risquait à un commentaire, samedi 14 janvier, sur la révocation de **Mgr Gaillot**.
6. L'éviction d'Évreux de **Mgr Jacques Gaillot**, le [13 janvier], a, en effet, réveillé le complexe anti-romain et a creusé le fossé entre l'Église et la société.
7. Trente minutes d'entretien en tête à tête avec le pape et quarante minutes pour le raconter à la presse au complet : la visite au [Vatican] de **Mgr Jacques Gaillot**, déchu de ses fonctions d'[évêque] d'Évreux en [janvier], ne sera pas passée inaperçue, jeudi 21 décembre à Rome.
8. La mise à pied du « libéral » [évêque] français **Jacques Gaillot**, le [13 janvier], a toutefois ravivé ces craintes.

Nous avons indiqué entre crochets les mots de la question présents dans les phrases réponses sans aucune variation. Si les entités nommées de la question varient peu (date, organisation), le verbe est presque toujours absent sous sa forme initiale. Regardons maintenant le verbe, comment il varie dans les phrases et quels phénomènes précis sont en jeu.

Les exemples de variations contenus dans le tableau 1.1 illustrent les phénomènes à gérer quand on veut partir d'une formulation pour aboutir à une autre. Les variations peuvent être de types différents. On distingue notamment celles qui ont trait à la sémantique (synonymie, polysémie, périphrases, expressions figées, mots en plusieurs mots) et donc à l'ambiguïté de la langue et celles liées à la syntaxe (valence, voix, anaphore, propositions

[5]La réponse est indiquée en gras, et les mots de la question entre crochets quand il s'agit de la même forme canonique (lemme).

Phrases	Types de variation	Variations
1	synonymie	démettre
2, 5	nominalisation + synonymie	révocation
3	nominalisation	destitution
4	synonymie	démettre
6	nominalisation + synonymie	éviction
7	synonymie	déchoir
8	locution	mise à pied

Tab. 1.1 Exemples de variations entre question et réponses

relatives, nominalisation, verbalisation). Nous reviendrons plus en détail sur les problèmes de variabilité de la langue.

A l'opposé, voici les passages candidats qui contiennent des mots de la question mais pas la réponse attendue :

1. Devenu [évêque] du diocèse jusque à sa révocation le [13 janvier] dernier par le [Vatican], **ce rebelle** a ainsi souligné le vide épiscopal en s'abstenant de s'asseoir sur le siège officiel jusqu'alors réservé à sa fonction.
2. Sa révocation le [13 janvier] a suscité de multiples protestations en France.
3. Une délégation de la conférence épiscopale française sera reçue au [Vatican] par le pape, vendredi 3 mars, pour faire le point sur les réactions qui ont suivi en France la destitution, le [13 janvier], de l'ancien [évêque] de Evreux.
4. Pour calmer la situation, le [Vatican] a nommé le [13] avril dernier l['évêque] auxiliaire de Vienne, Mgr Christoph Schonborn, 50 ans, comme coadjuteur pour seconder le cardinal.
5. Le [Vatican] a [suspendu] l'[évêque] paraguayen à la retraite Fernando Armindo Lugo Mendez, qui a déclaré son intention de se présenter à la présidence de son pays, l'un des plus pauvres d'Amérique latine.
6. Le grand rabbinat d'Israël, qui avait [suspendu] le dialogue avec le [Vatican] à la suite de la levée de l'excommunication d'un [évêque] négationniste, a décidé de le reprendre.

Ces exemples montrent qu'il ne suffit pas de chercher des passages contenant les mots de la réponse à l'identique pour trouver une réponse correcte. Il faut approfondir leur analyse pour extraire la réponse. Cet approfondissement passe en général par la nécessité de prendre en compte la variabilité de la langue. Une évaluation des critères linguistiques consistera à mesurer l'impact des critères choisis par le processus pour le reste du système (verbe

de la question, noms propres, etc.). Ainsi une évaluation de type diagnostic appliquée à un système de questions-réponses devra tracer les phénomènes linguistiques rencontrés et déterminer en quoi leur prise en compte permettra d'améliorer le système.

Le système FRASQUES

Nous détaillons un système de questions-réponses afin de montrer les stratégies linguistiques mises en œuvre et sur lequel nous prenons appui pour préciser notre propos. La figure 1.2 montre l'architecture du système FRASQUES [Grau *et al.*, 2006], développé au LIMSI, système sur lequel nous avons appliqué certaines de nos évaluations.

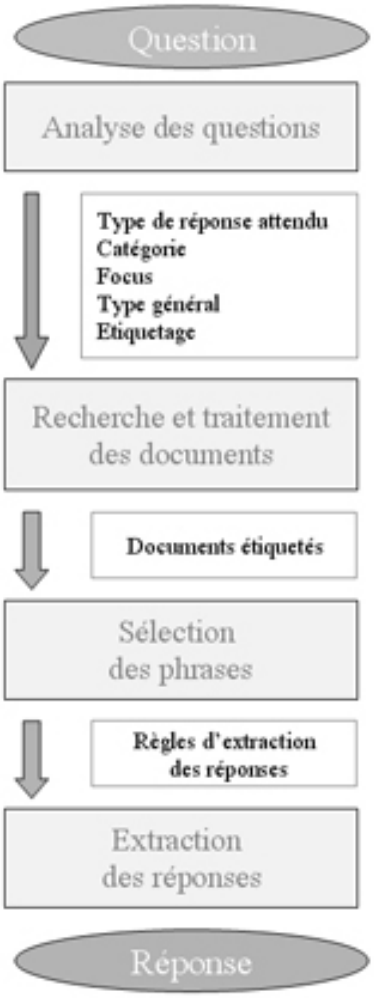


Fig. 1.2 Architecture de FRASQUES

L'analyse des questions permet d'extraire les informations essentielles à l'identification de la réponse correspondant à la question. En effet, la question constituant l'unique

source d'information disponible, il est indispensable d'en extraire les informations à traiter. Si celles-ci sont erronées, il va sans dire que le système n'arrivera pas à trouver de réponse. Nous présentons ici différents éléments qui importent pour le repérage de la réponse et pour lesquels des traitements linguistiques sont réalisés.

- La catégorie de la question (qui permettra de spécifier le traitement effectué par le système).
- Le type de réponse attendu, dans le cas où la réponse est une entité nommée (c'est-à-dire une unité d'un élément discursif qui fait référence à une personne, un lieu, une organisation, etc.). Ces entités sont repérées comme telles dans les documents extraits par le moteur de recherche. Le type est déterminé par des critères syntaxiques et sémantiques : le pronom *qui* indique que la réponse devra être une personne, *où* indique un lieu, *quand* indique une date, etc. S'il s'agit d'une question de type *Quel président français a été élu deux fois ?* le système attendra une entité nommée de type *personne*.
- Le type sémantique de la réponse, lorsque celui-ci est explicite, est déterminé par un terme hyperonyme de la réponse. Pour *Quelle est la capitale du Togo ?* le type sémantique de la réponse sera *capitale* et le type de sémantique sera **LOCATION-CITY**.
- Le focus, critère sur lequel nous reviendrons plus longuement, correspond à l'entité sur laquelle porte la question.
- Les noms propres présents dans la question, constituants les moins sensibles à variation.
- Une extension sémantique de la question, qui consiste en la recherche de synonymes.

Pour la question *Quel célèbre ferry a fait naufrage en mer Baltique ?* les informations présentes dans le tableau 1.2 seront extraites lors de son analyse.

Critère	**Exemple**
Catégorie	quel
Type sémantique	ferry
Focus	naufrage
Nom propre	Baltique
Extension sémantique	bateau/couler/etc.

Tab. 1.2 Exemple d'analyse d'une question

C'est également lors de l'analyse de la question qu'a lieu son étiquetage morpho-

syntaxique et son analyse syntaxique. Il est effectué par le Tree-Tagger[6], outil multilingue développé par Helmut Schmid.

La recherche et le traitement des documents sont effectués dans le système FRASQUES par le moteur de recherche Lucène[7], outil libre écrit en Java. Les paramètres donnés au moteur de recherche sont issus de l'analyse de la question : mots de la question et leurs variations.

La sélection des phrases réponses est effectuée par un module qui analyse les documents sélectionnés par le moteur de recherche afin de prendre en compte au mieux les variations entre la formulation de la questions et les diverses formes de réponses possibles. L'utilisation de l'outil FASTR [Jacquemin, 1996] permet de « reconnaître des variantes d'expressions contenant un ou plusieurs mots, à partir de lexiques de variantes morphologiques et sémantiques, et de règles de reformulation syntaxique » [Ferret *et al.*, 2001a]. Ceci permet une certaine souplesse pour reconnaître des variantes quelques fois éloignées. Ensuite, le système découpe les documents en phrases (les phrases sont pondérées en fonction des termes communs à la question et des variations qu'elle contient) et étiquette les entités nommées qui s'y trouvent. Les entités nommées désignent des noms de personne, de lieux, d'organisation ou encore des dates ou des unités monétaires. Ce sont des éléments très importants pour l'extraction des réponses, notamment quand la question attend une réponse qui est une entité nommée. La campagne d'évaluation MUC-7 (*Message Understanding Conference*) en 1998 a permis des avancées sur l'identification et le typage des ces entités [Poibeau, 2005]. La classification hiérarchique de ces entités utilisées dans FRASQUES est présentée sur la figure 1.3, issue de [Ferret *et al.*, 2001b].

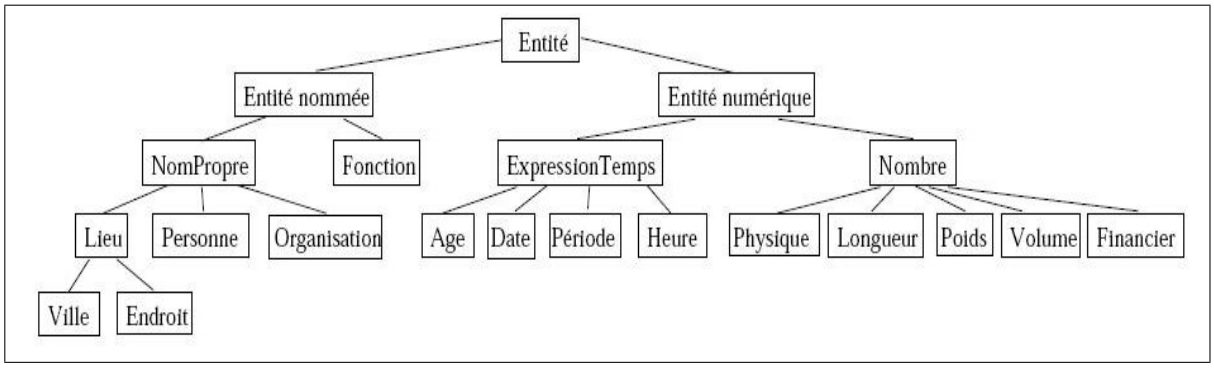


Fig. 1.3 Typologie des entités nommées

[6]http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[7]http ://lucene.apache.org/

Ainsi, pour la question *Où se situent les îles Marquises* qui attend en réponse une entité nommée de type **LIEU**, le système privilégiera la phrase réponse suivante[8] en fonction de sa pondération :

*Je suppose que les dames des îles Marquises, dans l' <**enamex type="LOCATION-PLACE"**>**océan Pacifique**</**enamex**>, ont déterminé beaucoup de vocations ethnographiques.*

L'extraction de la réponse se fait en attribuant un score aux réponses en fonction du poids de la phrase réponse dont elle est extraite. Le processus diffère en fonction du type de réponse attendu (identifié lors de l'analyse de la question), c'est-à-dire s'il s'agit d'une entité nommée ou non. Dans le premier cas, on extrait l'entité nommée la plus proche du barycentre des variantes de mots [de Chalendar *et al.*, 2002].

Sinon, on applique des patrons d'extraction écrits sous la forme de règles. Ils reposent sur le focus extrait de la question (et éventuellement ses synonymes), le type général ainsi que sur le verbe principal. Ces patrons permettent, à partir d'un point d'ancrage dans la phrase, d'extraire la réponse en fonction de son lien syntaxique avec celui-ci.

Par exemple, l'application de patrons d'extraction sur des termes issus de la question, ici le focus, sur la question *De quelle organisation Javier Solana était-il secrétaire général ?* donnera[9] :

- Focus : Javier Solana

  Phrase : *Le sénat américain s'est dit vendredi "préoccupé" par la nomination attendue du ministre espagnol des Affaires étrangères **Javier Solana** au poste de secretaire général de l'[OTAN].*

Néanmoins, si les éléments de la question ne sont pas reconnus correctement, il reste peu de chances de trouver la réponse. Différents problèmes peuvent survenir :

- L'analyse de la question n'a pas été effectuée correctement,
- L'étiquetage morpho-syntaxique n'est pas correct,
- Le patron d'extraction est trop large ou trop restrictif.

Nous reviendrons sur l'impact de ces erreurs lors de l'analyse des erreurs typiques des systèmes de questions-réponses (2.1).

La description des différents composants d'un système de questions-réponses permet de mettre en lumière certains points cruciaux reposant sur des processus d'analyse linguistique qu'il faut analyser : l'extraction des informations de la question, la validité des

[8] *Enamex* indique qu'il s'agit d'une entité nommée, qui est de type LIEU. Baliser les entités facilitent leur extraction a posteriori.

[9] La réponse est indiquée entre crochets dans les phrases.

critères choisis pour l'analyse, l'étiquetage morpho-syntaxique, l'application des patrons d'extraction des réponses précises. Ces informations nécessitent d'être évaluées si l'on veut effectuer un diagnostic complet du système, pas seulement en sachant quel composant doit être amélioré, mais comment.

1.2.2 L'évaluation en question-réponse

Nous présentons quelques campagnes d'évaluation pour les systèmes de questions-réponses afin de mesurer leur apport pour notre démarche d'évaluation.

Les campagnes d'évaluation

Afin de mettre en évidence les performances possibles des systèmes de questions-réponses, des campagnes d'évaluation centrées sur ces systèmes sont organisées chaque année. Le principe de ces campagnes consiste à évaluer les systèmes de questions-réponses en proposant un jeu de questions ainsi qu'un corpus donné au sein duquel rechercher les réponses. Chaque système doit renvoyer une réponse précise pour chacune des questions, réponses sur lesquelles il sera jugé. Nous décrivons trois campagnes d'évaluation, mais on peut également citer NTCIR[10], campagne d'évaluation pour les SQR dédiée aux langues asiatiques.

TREC [11] est la première campagne d'évaluation en questions-réponses a eu lieu en 1999 : il s'agit de la huitième édition de TREC, qui ne portait jusqu'alors que sur les systèmes de RI. Le déroulement de la tâche s'est complexifié au fil des années : les premières campagnes proposaient 200 questions, pour lesquelles les participants pouvaient proposer jusqu'à cinq réponses par question (les questions étaient uniquement factuelles). 500 questions étaient proposées, factuelles mais également des questions définitionnelles, des questions qui attendent plusieurs réponses (listes) et des scénarios (questions dont la réponse est liée à celle de la précédente). Enfin, certaines réponses ne sont pas présentes dans le corpus de documents et on attend du système qu'il l'indique. Pour la validation des résultats du système, une seule réponse courte est exigée. Elle doit être accompagnée du document qui justifie la réponse proposée par le système.

[10] Voir http ://research.nii.ac.jp/ntcir/
[11] Voir http ://trec.nist.gov/

La taille du corpus de documents a augmenté de 528 000 documents en 1999 à 1 033 000 en 2005 (près du double), ce qui amène les systèmes a être de plus en plus puissants et précis dans leurs traitements.

CLEF [12] est l'acronyme de *Cross Language Evaluation Forum* [Voorhees, 2002]. Il s'agit d'une campagne d'évaluation qui permet à des systèmes de travailler en monolingue, mais aussi sur deux langues voir sur des corpus multilingues. L'idée est de favoriser les recherches sur des langues autres que l'anglais afin de permettre à différentes équipes de travailler sur leurs langues et de pouvoir s'évaluer et progresser.

EQUER est un projet français, organisée par TECHNOLANGUE[13] de façon à créer un paradigme d'évaluation pour les systèmes de questions-réponses français [Ayache *et al.*, 2005]. Les tâches proposées sont analogues à celles de TREC. 500 questions, un corpus journalistique (environ 1,5 gigaoctets) et un corpus médical ont été mis à disposition des participants. L'intérêt de cette évaluation était d'obtenir un savoir sur les différents types de questions et leur résolution pour le français, afin d'affiner les stratégies employées par les SQR. Le deuxième intérêt de cette campagne était de permettre ensuite à n'importe quel industriel ou académique d'évaluer son système lui-même en l'utilisant dans des conditions identiques.

Ces structures d'évaluation permettent de stimuler la recherche par la création de nouveaux enjeux chaque année. De ce fait, nous pouvons faire un constat mesuré de l'avancé technologique de ces systèmes chaque année. Un autre aspect important consiste en la mise au point d'étalons de référence afin d'évaluer ces systèmes à l'aide de processus bien définis et de méthodes de comparaisons qui se veulent objectives.

Néanmoins, si la nécessité de ces campagnes d'évaluation n'est plus à démontrer, on peut tout de même avancer quelques limites quant aux évaluations qui sont réalisées : seul le résultat final obtenu par chacun des systèmes est pris en compte (classement selon les résultats obtenus). Or, de la sorte, on n'évalue pas ce qui se passe à l'intérieur des systèmes-participants de façon précise. Le participant ne sait donc pas pourquoi il a obtenu tel ou tel résultat erroné [Grau, 2004b].

De plus, ces campagnes ne permettent pas d'évaluer les critères linguistiques mis en œuvre par ces systèmes, ni de constituer des corpus dédiés aux certains phénomènes. En effet, comme le soulignent Guy Lapalme et Karine Lavenus dans [Lapalme et Lavenus,

[12] http ://www.clef-campaign.org/
[13] Voir http ://www.technolangue.net/

2002], « la compétition pourrait être basée sur un ensemble de questions plus variées, plus proches des questions d'utilisateurs et comprenant des difficultés linguistiques graduelles pour le TAL ». En effet, la création de jeux de questions basés sur les difficultés de résolution plutôt que sur le type de questions permettrait des évaluations plus significatives en terme de difficultés intrinsèquement liées à la tâche.

Nous présentons des évaluations sur des sous-tâches plus ciblées pour évaluer des stratégies linguistiques, utilisables dans les systèmes de questions-réponses.

L'évaluation de sous-tâches

Des campagnes d'évaluation sont organisées uniquement sur des sous-tâches [Habert et Zweigenbaum, 2002], c'est-à-dire des phénomènes linguistiques rencontrés par les systèmes, en dehors de leur cadre d'application. C'est le cas pour trois outils utilisés pour les systèmes de questions-réponses :

- La reconnaissance des entités nommées (MUC, ACE (Automatic Content Extraction), DUC, Ester, Quaero) ;
- L'analyse syntaxique (EASY, PASSAGE) ;
- L'étiquetage morpho-syntaxique (GRACE, Parseval).

Ces évaluations sont permises par un jeu d'étiquette standard que les systèmes de tous les participants doivent adopter ainsi que par des données communes, afin d'évaluer les performances avec un référentiel commun. Si ces évaluations sont nécessaires afin de mesurer la pertinence de ces outils, peu d'erreurs peuvent engendrer beaucoup de mauvaises réponses en contexte. En effet, l'utilisation de ces outils en contexte complexifie les traitements à effectuer, ce qui implique que les performances sont moins bonnes. Une évaluation hors contexte est donc insuffisante, C'est pourquoi il est intéressant de pouvoir réaliser de telles évaluations en contexte, c'est-à-dire d'isoler les résultats produits par ces outils dans le cadre d'une tâche qui les utilise (ceux d'un système de questions-réponses par exemple) afin de mesurer leur impact réel.

Les catégories de questions

Il s'agit ici de se poser la question de savoir si la catégorisation des questions peut répondre à l'évaluation de difficultés linguistiques ? En effet, les campagnes de questions-réponses fournissent des jeux de questions aux participants, lesquelles sont réparties en

différentes catégories : Les principales catégories de questions que l'on retrouve lors des campagnes sont :

- Définition : *Qu'est-ce que l'accélération centrifuge ?*
- Combien : *Combien y a-t-il eu de mariages en Grande-Bretagne en 1993 ?*
- Quand : *Quand a eu lieu la chute du régime communiste en Afghanistan ?*
- Où : *Où se situent les îles Marquises ?*
- Quel : *Quelle était la nationalité d'Ayrton Senna ?*
- Qui : *Qui est Michael Jackson ?*
- Instance : *Citer le nom d'un corps céleste.*

Or, ces questions ne sont pas définies en fonction de la difficulté de leur résolution. Elles sont censées représenter le besoin réel des utilisateurs, tout en étant représentatives des différents phénomènes en langue qu'il est possible de rencontrer dans leur formulation. Étant donné que les réponses et leur contexte d'apparition ne sont pas pris en compte, on ne connaît pas les phénomènes a priori qu'il sera nécessaire de traiter pour leur résolution puisque la difficulté de résolution d'une question est liée à la formulation de sa réponse. En effet, les corpus sont créés pour des évaluations de performance en comptant sur le nombre de questions proposées pour avoir un corpus représentatif des phénomènes en jeu.

Le seul effort de traitement d'un phénomène sémantique particulier a eu lieu lors des campagnes CLEF, avec l'introduction de questions portant sur le temps. Quant aux phénomènes de variations en langue, les campagnes TREC et EQUER ont introduit, en plus des questions de catégories habituelles, des reformulations de questions. Cette piste n'a pas été concluante : les reformulations ont généralement amené des réponses différentes de celles des questions originales.

Ainsi, si l'on veut constituer un corpus ciblé et réaliser une évaluation ciblée, il faut mettre en relation question et réponse et être capable de retrouver les phénomènes que l'on veut étudier. C'est le type d'étude menée dans [Cohen *et al.*, 2004], appliquée à l'extraction d'information dans le domaine bio-médical. Il s'agit de d'une part de constituer des corpus selon des critères linguistiques fins pour les noms de gènes (majuscule au début du mot, longueur du mot, etc.) et d'autre part de produire des corpus qui respectent certaines propriétés. C'est la démarche que celle que nous adoptons, en partant des résultats d'un système de questions-réponses.

Ceci nous amène à poser les principes d'une évaluation transparente des systèmes de questions-réponses, qui permettrait une analyse fine des systèmes. Les évaluations classiques s'arrêtent aux modules informatiques qui composent la chaîne de traitement, or nous avons

besoin d'un niveau de granularité plus fin pour tracer les phénomènes.

En effet, la granularité des composants ne permet pas d'aller au plus près de la granularité des phénomènes linguistiques et l'on ne peut se limiter à une évaluation de composant. De plus, mesurer l'impact d'un critère linguistique sur le résultat global du système nécessite une évaluation transversale du système car celui-ci transite d'un module à l'autre.

1.3 L'évaluation transparente

1.3.1 Terminologie

Il est courant de définir un système comme une boîte : on a une entrée, un processus et une sortie, comme le montre la figure 1.4.

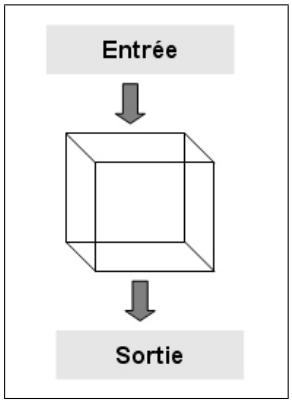


Fig. 1.4 Système schématique

À partir de cette configuration, nous allons pouvoir présenter la notion d'évaluation transparente. Comme nous l'avons vu lors de la présentation des différentes campagnes d'évaluation, l'évaluation standard qui permet de comparer des SQR est l'évaluation de performance, ou évaluation **boîte noire** (*black box evaluation*). Cette idée fait echo au schéma précédent, avec une évaluation des résultats produits en sortie par le système : l'évaluation est opaque puisqu'elle ne tient pas compte de ce qui se passe à l'intérieur du système.

À l'inverse, une évaluation de type **boîte transparente** [Hurault-Plantet et Monceaux, 2002] évalue non plus les résultats finaux du système, mais prend en compte ce qu'il y a à l'intérieur des systèmes. Il s'agit d'évaluer les performances des composants eux-mêmes.

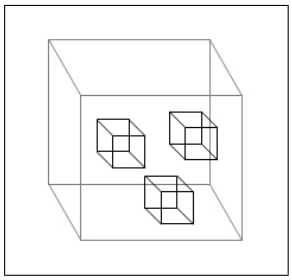


Fig. 1.5 Evaluation de type boîte transparente

Intérêt pour les systèmes modulaires

Les systèmes de questions-réponses sont des systèmes modulaires, dont les composants fonctionnent généralement de manière linéaire. On peut schématiser le processus à l'aide de rouages qui représentent les différents modules (figure 1.6).

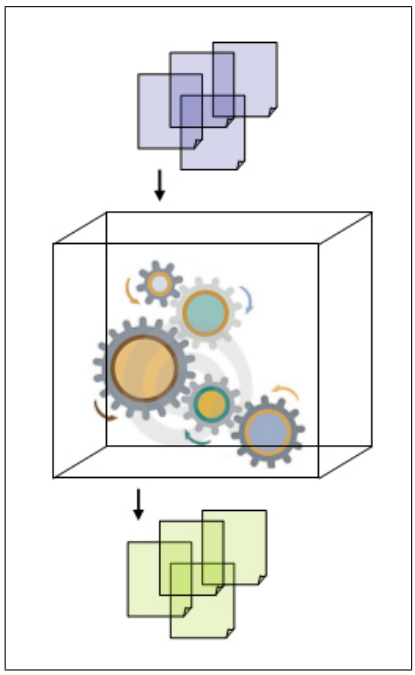


Fig. 1.6 Enchaînement des modules

Ainsi, si le premier rouage ne fonctionne pas, le reste du processus est voué à l'échec. Selon [Gillard *et al.*, 2006], l'état de l'art des évaluations réalisées sur des systèmes de questions-réponses montre le manque de lisibilité des apports des composants sur les résultats finaux et la nécessité d'une étude plus approfondie de chacun des composants. On voit alors la pertinence d'évaluations de type boîte transparente : le résultat d'un module est pris en entrée par celui qui suit, et le résultat obtenu en aval dépend alors de la qualité des

traitements effectués. Ainsi, dans un souci d'amélioration du système, ces nouvelles formes d'évaluation deviennent nécessaires [Sparck Jones, 2001].

Selon [Nyberg et Mitamura, 2002], il est également important de prendre en compte l'évaluation de type boîte transparente qui apporte deux avantages :

- la possibilité d'évaluer la performance des modules pris individuellement peut aider les développeurs à localiser rapidement les problèmes liés aux fonctionnalités, à la performance, etc. ;
- la compréhension de la facilité de modifier, développer et maintenir le système.

Il s'agit donc d'observer les résultats produits par les différents modules isolément, sans avoir à lancer le processus dans sa globalité, lequel ne sera pas forcément révélateur des problèmes existants. Une évaluation sélective et partitionnée permet de mesurer l'efficacité réelle de tel ou tel processus, afin de maximiser l'équilibre entre temps de traitement et efficacité. En effet, certains traitements peuvent être longs à s'exécuter sans pour autant être significatifs au niveau des résultats. C'est un des points sur lesquels une évaluation interne permet d'améliorer un système. Évaluer de façon pointue permet de visualiser de façon précise ce qui n'est pas correct, mais aussi d'évaluer la rentabilité de certains traitements. Plus précisément, dans le cadre qui nous intéresse, analyser les résultats obtenus par le système va permettre la redéfinition de certaines notions (comme celle du focus, point que sera explicité en aval), dont les définitions seraient trop larges ou bien trop strictes par rapport à l'utilisation qui en est faite. De plus, la mise en place d'un procédé d'évaluation transparente permet de modifier certains traitements, d'en ajouter ou bien d'en enlever certains et de tester la pertinence de ces modifications.

Granularité de l'évaluation transparente des composants

Néanmoins, si l'on parle d'évaluation transparente, il s'agit plus ou moins d'une évaluation boîte noire locale. On évalue non plus la sortie finale du système, mais la performance de chacun des modules indépendamment les uns des autres et les mêmes manques sont présents. Nous nous situons plus dans une perspective d'observation et d'analyse de corpus avec l'idée de tracer le comportement d'un critère linguistique tout au long de la chaîne de traitement afin de mesurer comment les stratégies implémentées fonctionnent quand elles sont confrontées à différents phénomènes linguistiques. Nous pourrions définir notre approche comme de l'évaluation de type boîte transparente des composants eux-mêmes : on regarde dedans et on analyse de façon à obtenir un diagnostic précis de ce qui est et ce qui n'est pas effectué correctement.

Ce savoir précis ne peut être obtenu sans avoir accès à ce qui se passe à l'intérieur du système, mais également à l'intérieur des composants. De la sorte, nous proposons une évaluation non plus transparente du système, mais une évaluation transparente des composants eux-mêmes, avec un accès aux données produites à chacun des niveaux de la chaîne de traitement.

Afin de généraliser et de montrer cette nécessité d'évaluation transparente beaucoup de systèmes de questions-réponses, nous allons présenter des travaux qui montrent les points importants à évaluer.

1.3.2 Évaluation spécifique en question-réponse

On dénombre un certain nombre d'articles en lien avec les campagnes d'évaluation organisées présentant les retours d'évaluation des équipes sur leurs systèmes. Certains présentent l'évaluation de nouvelles stratégies au sein de SQR, et permettent d'entre-apercevoir comment les équipes évaluent leurs systèmes en interne afin de tracer les erreurs et d'y remédier. Nous présentons ici trois articles qui portent sur différents systèmes et donnent à voir des indices sur les points clés à évaluer dans un SQR.

QRISTAL

QRISTAL est un système de questions-réponses développé par la société Synapse, qui utilise des stratégies linguistiques. Il s'agit du SQR qui obtient les meilleurs résultats à l'heure actuelle sur le français. Ce système nous intéresse particulièrement car il y est fait « une utilisation intensive des outils TAL, entre autre l'analyse syntaxique, la désambiguïsation sémantique, la recherche des référents des anaphores, la détection des métaphores, la prise en compte des converses, le repérage des entités nommées ou encore l'analyse conceptuelle et thématique » [Laurent *et al.*, 2006]. Dans ce même article, les auteurs présentent une étude fine de leur système en en déroulant le traitement d'une question, avec en filigrane la nécessité du retour au texte et à l'analyse de façon à vérifier le traitement effectué.

Le système QRISTAL dispose de quatre étapes principales qui sont :

1. l'analyse la question (forme syntaxique, extraction des mots importants et recherche de leurs synonymes) ;
2. la recherche des blocs de documents contenant les mots extraits de la question et attribution d'un score ;

3. l'analyse les blocs extraits (syntaxe, sémantique, recherche d'entités nommées et d'anaphores) ;
4. l'extraction des réponses précises (entité nommée correcte proche des mots importants de la question) et l'attribution d'un score.

Nous nous intéressons essentiellement à l'analyse des questions, module primordial pour la suite du traitement. Si l'on veut évaluer cette étape, il est nécessaire de regarder à l'intérieur du système avec accès aux données produites. C'est ce qu'ont fait les auteurs de l'article ; nous reprenons sur la figure 1.7 l'exemple qu'ils présentent.

N°	MOT	LEMME	P	Fonction	Groupe	Ss-Gr.	Type	Type d...	Sémantique
1	Qui	qui	1	Sujet	Groupe pronominal	1/1	PRI	PRON....	
2	était	être	1	Verbe		2/2	VINDI3S	Indicati...	existence/événement/vérité
3	le	le	1	Attribut du sujet	Groupe nominal	4/4	DETDMS	ART.Dé...	
4	président	président	1	Attribut du sujet	Groupe nominal	4/4	NCMS	NOM M...	exécutive
5	de	de	1	Attribut du sujet	Groupe nominal prépositionnel	4/7	PREP	PRÉPO...	
6	la	le	1	Attribut du sujet	Groupe nominal prépositionnel	4/7	DETDFS	ART.Dé...	
7	France	France	1	Attribut du sujet	Groupe nominal prépositionnel	4/7	NCFS	NOM F...	
8	au moment des	au moment des	1	Complément circonstanciel de temps	Groupe nominal prépositionnel	9/9	PREP	PRÉPO...	moment
9	essais	essai	1	Complément circonstanciel de temps	Groupe nominal prépositionnel	9/9	NCMP	NOM M...	polysémique : tentative/experiment,attempt,try
10	nucléaires	nucléaire	1	Complément circonstanciel de temps	Groupe nominal prépositionnel	9/9	ADJPIG	ADJ.Plu...	atome; domaine = physique
11	dans	dans	1	Complément circonstanciel de lieu	Groupe nominal prépositionnel	13/13	PREP	PRÉPO...	
12	le	le	1	Complément circonstanciel de lieu	Groupe nominal prépositionnel	13/13	DETDMS	ART.Dé...	
13	Pacifique	Pacifique	1	Complément circonstanciel de lieu	Groupe nominal prépositionnel	13/13	NCMS	NOM M...	
14	Sud	Sud	1	Complément circonstanciel de lieu	Groupe nominal prépositionnel	13/13	NCHSIG	NOM H ...	
15	?	?	1	ponctuation forte					

Fig. 1.7 Analyse d'une question par QRISTAL

Bien évidemment, il est possible d'effectuer des mesures en sortie du module afin d'obtenir un pourcentage de bon fonctionnement. Mais on obtiendra alors une évaluation des performances du système et non pas un savoir précis sur les difficultés que rencontre le module (évaluation de diagnostic). Comme on peut le voir sur la figure 1.7, l'analyse comporte différents traitements et c'est en regardant ce qui a été produit que les auteurs ont pu vérifier que les stratégies employées fonctionnent. Si ce retour aux sources n'est pas mentionné directement, c'est qu'il s'agit d'une pratique courante qui ne fait pas appel à une ingénierie quelconque : chaque équipe dépouille ses résultats pour orienter les améliorations.

Néanmoins, la quantité d'informations liées à l'analyse de la question montre l'intérêt d'un accès visuel rapide pour s'assurer du bon fonctionnement de l'analyse. Toutefois, dans le cadre du traçage d'une erreur particulière, comment avoir accès à différentes erreurs du même type ? Ou bien à l'analyse de questions d'une catégorie donnée ?

De plus, aucune méthodologie d'évaluation de ces informations n'apparaît, et nous ne pouvons pas mesurer l'impact de la bonne ou mauvaise reconnaissance des informations sur le résultat final. Nous allons décrire l'article [Moldovan *et al.*, 2003] qui propose quelques pistes de réflexion sur d'autres aspects importants à évaluer pour les systèmes de questions-réponses.

Évaluation modulaire de Moldovan et al.

[Moldovan *et al.*, 2003] montre la nécessité d'évaluer l'apport de chacun des composants d'un système modulaire, afin de percevoir la part de chaque module pour les faiblesses globales. Il s'agit de comprendre où le système se trompe.

Les auteurs identifient différents points qui leur semblent importants d'évaluer. Les modules qui génèrent le plus d'erreurs sont : la construction d'une représentation de la question, la dérivation du type de la réponse, la sélection de mots clés et leur expansion et l'identification de réponses candidates. De plus, de certains modules dépendent la suite de l'analyse, ce qui rend leur performance d'autant plus déterminante. Par exemple si l'analyse des questions est erronée, il sera presque impossible de trouver la réponse précise attendue. La question est le seul contexte dont dispose le système, tous les indices sont là. Leur repérage est extrêmement important.

De plus, on voit que certains modules sont plus importants que d'autres, plus fondamentaux et qu'il est essentiel de les reconnaître si l'on veut y consacrer toute l'attention qu'il faut.

Ce que cet article montre également, c'est l'apport concret des techniques de traitement automatique des langues en association avec les stratégies liées à la recherche d'information. Les auteurs indiquent que leur combinaison améliore la précision des réponses obtenus, notamment à l'aide de ressources lexicales (ils citent WordNet[14] pour la langue anglaise).

Néanmoins, même si cet article apporte un éclairage en terme de modules et non pas de processus global, les techniques TAL utilisées sont mesurées en calculant la précision obtenue par le module avec ou sans telle ou telle stratégie. Or, cela ne permet pas d'obtenir une connaissance précise des phénomènes linguistiques rencontrés et de la qualité du traitement qui en est réalisé.

Les deux articles présentés sont intéressants dans leur démarche d'évaluation, mais ne présentent ni méthodologie, ni outil pour réaliser des évaluations systématiques, autre que la suppression de composants.

1.3.3 Évaluation transparente appliquée aux systèmes de questions-réponses

Évaluer finement l'apport des composants d'un système suppose de mesurer la contribution de chacun des modules par rapport aux résultats finaux obtenus par le système. De ce

[14]Plus d'informations sur : http ://wordnet.princeton.edu/

point de vue, l'évaluation de type boîte transparente n'est pas en opposition avec une évaluation boîte noire, mais complémentaire pour obtenir un diagnostic complet [Sparck Jones, 2001]. Ces deux méthodes dépendent avant tout de ce que l'on veut évaluer.

Néanmoins, l'évaluation transparente n'est pas très présente dans la littérature du domaine. En effet, peu de chercheurs se sont réellement intéressés à cette thématique de façon générique. Les références sont pauvres et datent de quelques années. On trouve néanmoins deux courants liés à l'évaluation des différents modules d'un système.

Modification de l'architecture du système

Le premier consiste à enlever un composant et à mesurer les résultats finaux obtenus. C'est ce qui est développé dans [Costa et Sarmento, 2006] sur le système ESFINGE, qui travaille essentiellement sur le portugais. Leur problématique est la suivante : évaluer la performance de chacun des composants d'un système modulaire est primordial pour mesurer leur impact sur la performance globale et pour identifier quand tel composant est nécessaire, doit être amélioré ou bien remplacé. Trois évaluations différentes sont menées : une analyse des erreurs, une évaluation en enlevant un constituant et une autre en remplaçant un constituant par un autre. L'avantage qu'ils observent concernant la modification des *logs d'évaluation* est qu'il s'agit d'une solution alternative pour les composants que nous nommerons porteurs, c'est-à-dire ceux qui ne peuvent être déconnectés du système sans en bloquer totalement le fonctionnement. De la même façon, la limite des modifications des *logs d'évaluation* est qu'elles sont insuffisantes pour déceler les contributions mineures. Leur méthodologie d'évaluation consiste à comparer les résultats de différents runs (avec remplacement de composants différents) par rapport au nombre final de bonnes réponses obtenues (évaluation de type *black box*). C'est la même démarche dans [Moriceau et Tannier, 2009] sur le système FIDJI (*Finding In Documents Justifications and Inferences*), où les auteurs cherchent à évaluer l'apport des modules d'analyse syntaxique, notamment pour l'extraction de réponses précises. Si l'on voit bien que la déconnection du module d'analyse syntaxique fait perdre des résultats de façon significative, il serait intéressant d'évaluer précisément l'analyse elle-même, ce qui n'est pas réalisable sans avoir accès aux données produites par le système.

Un autre modèle qui tend à faciliter l'évaluation de composants de systèmes de questions-réponses, plutôt qu'à la permettre réellement, est illustré dans l'article [Tomas *et al.*, 2005]. Il s'agit d'un outil basée sur la technologie XML qui permet plus facilement

l'intégration, la combinaison et l'évaluation des composants d'un système construit sur des approches différentes. Ils s'intéressent notamment à faciliter des développements à venir du système, en terme d'architecture. La centralisation des informations et résultats produits par le système permet deux choses :

- interchanger des modules sans modifier le système (tout le processus est indiqué dans un fichier XML)
- tester un module indépendamment du reste du processus (les données nécessaires à son bon fonctionnement sont stockées dans un fichier)

Néanmoins, l'article ne fait pas état d'exemples qui pourraient venir illustrer l'apport et la nécessité de cet accès à différents états de la chaîne de traitement. Le propos se situe plus à un niveau d'ingénierie des systèmes, mais demeure intéressant dans le fait qu'il rend possible différents traitements avec une interopérabilité des résultats.

Contrôle de l'exécution

Le deuxième courant, assez novateur dans le domaine, est illustré par le système JAVELIN (*Justification-based Answer Valuation through Language Interpretation*) [Nyberg *et al.*, 2003]. JAVELIN est un système de questions-réponses qui intègre un module permettant de contrôler l'exécution du processus, ainsi que les informations qui sont utilisées. Il a été conçu de façon à permettre une évaluation de type boîte transparente : l'architecture a été créée pour intégrer une évaluation des composants, de façon à comparer des stratégies différentes sur la base de critères de performance [Nyberg *et al.*, 2002]. Il permet également de tester différentes stratégies qui peuvent ensuite être intégrées au système. Les données, propres à l'exécution et celles produites par le système, sont stockées grâce à l'*Execution Manager*. Ils utilisent également un module de planification (*the Planner*), lequel permet de sélectionner de façon dynamique entre différentes versions des composants du système et de générer différentes stratégies en même temps, de façon à identifier par la suite la meilleure. Néanmoins, la question d'un outil générique permettant de réaliser de l'évaluation transparente d'autres systèmes de questions-réponses demeure non résolue.

Cependant, les travaux réalisés sur *Xtrieval*[15] [Kursten *et al.*, 2008] décrivent une interface d'évaluation pour systèmes de recherche d'information. Il s'agit d'une plate-forme flexible qui permet de tester et d'évaluer différents aspects de systèmes de recherche d'information, tout en mettant en avant l'extensibilité et la flexibilité de l'outil. Si cette étude

[15] Ce système a participé à CLEF07 et CLEF08 sur la tâche *Domain-specific*.

est plutôt axée sur le développement et l'amélioration de systèmes de recherche d'information en intégrant des graphiques du rappel et de la précision des résultats obtenus par un système, elle nous intéresse néanmoins dans son approche. En effet, la démarche d'intégrer des outils permettant une évaluation des composants est proche de ce que nous avons mis en oeuvre pour évaluer les systèmes de questions-réponses. Cette étude légitime le besoin d'interfaces pour évaluer des systèmes modulaires, notamment en permettant de stocker et de relancer des tests : la fonctionnalité la plus importante d'une évaluation des composants est sa capacité à stocker les tests et de relancer les expérimentations, ce qui permet de reprendre ces expériences à un autre moment.

1.3.4 Synthèse

Comme nous l'avons vu, les campagnes d'évaluation n'évaluent que des scores de performance. Il incombe aux participants d'évaluer en interne les résultats de leurs systèmes, et de parcourir les résultats obtenus pour traquer les erreurs.

Si certains auteurs se sont penchés sur la problématique d'une évaluation interne des systèmes, les techniques proposées proposent une évaluation des composants du système, ce qui est plus précis d'une évaluation globale, mais ce qui ne permet pas réellement d'obtenir une connaissance des obstacles que le système rencontre. Notre démarche consiste en une évaluation locale de type *boîte noire* plutôt qu'à une évaluation de type *boîte transparente*. De la sorte, nous appelons évaluation de type *boîte transparente* celle que nous proposons. Cette dernière a pour but l'évaluation des résultats produits par un système de questions-réponses donné, de façon à isoler les erreurs qu'il rencontre (démarche qui nous semble être « transparente »). Cette évaluation est réalisée de façon transversale, c'est-à-dire qu'elle prend en compte toutes les dimensions du système (et non pas un seul composant à la fois).

Nous avons montré les insuffisances des campagnes d'évaluation quant au diagnostic d'un système de questions-réponses et présenté différents types d'évaluation possibles à mettre en œuvre. L'architecture de forme modulaire de ce type de systèmes rend plus difficile l'accès aux données ainsi que la recherche systématique d'erreurs. Nous avons présenté différentes façons d'évaluer un système de questions-réponses, tout en relatant les écueils des évaluations existantes pour tracer les erreurs.

Plus précisément, la stratégie que nous adoptons consiste à étudier et éventuellement à modifier les résultats intermédiaires créés par les composants et insérer ces nouveaux

résultats dans le processus de traitement sans modifier le système lui-même. L'outil que nous avons conçu, REVISE (*Recherche, Extraction, VISualisation et Evaluation*), permet de stocker les données produites par le système de façon organisée (base de données relationnelle), de les observer en filtrant sur les informations qui nous intéressent (requêtes SQL) et de modifier ces données afin d'évaluer l'impact de ces modifications sur le système. Il s'agit de mettre en oeuvre un espace de classement des résultats, de proposer une visualisation intelligente de l'information, c'est-à-dire qui donne du sens aux données observées (jeux de couleurs, format dédié), dans le but d'une analyse fine des erreurs, mais également de permettre facilement de modifier les données afin de les réinjecter dans le processus au point de relance adéquat, dans le but de tester sans avoir à intervenir dans la mécanique du système. Ceci dans le but d'évaluer le taux de présence et l'impact d'un phénomène linguistique dans la tâche ainsi que son traitement informatique.

Nous proposons une méthode d'analyse transparente des systèmes de questions-réponses qui prend en compte un principe d'évaluation basé sur l'extraction de critères dans les questions pour les rechercher dans les phrases réponses. Nous avons établi le plan d'évaluation transparente pour d'un SQR :

1. Évaluer la présence et la pertinence de critères linguistiques fins, en particulier extraits des questions pour trouver des réponses ;
2. Mesurer la présence et la pertinence des éléments déclencheurs de réponse ;
3. Évaluer l'impact des outils linguistiques utilisés (étiquetage, analyse syntaxique en contexte) ;
4. Évaluer les résultats : pertinence des phrases réponses et précision des réponses précises ;

Ce sont ces points que nous allons définir plus précisément et systématiser par le développement d'un outil d'évaluation transparente des systèmes de questions-réponses, en discutant des aspects génériques liés à la démarche d'évaluation (chapitre 2).

Chapitre 2

Méthodologie d'évaluation transparente pour les systèmes de questions-réponses

Sommaire

Dans ce chapitre, nous présentons d'abord différents types d'erreurs fréquentes pour un système de questions-réponses afin de montrer les points à évaluer et d'en tirer une méthodologie d'évaluation. Nous proposons ensuite des solutions pour réaliser ces études à l'aide de l'outil que nous avons développé : REVISE.

2.1 Analyse des erreurs classiques d'un système de questions-réponses

Afin de bien cerner le problème, nous présentons les erreurs classiques qu'un système de questions-réponses peut rencontrer. Nous prenons appui sur le système de questions-réponses FRASQUES, développé au LIMSI sur le français. Mais on retrouve les mêmes problèmes dans tous les systèmes de questions-réponses. C'est ce que montrent les analyses des erreurs des systèmes de l'Université d'Amsterdam [Jijkoun *et al.*, 2003], Javelin [Shima *et al.*, 2006], AskMSR [Brill *et al.*, 2002] et SMART IR [Abney *et al.*, 2000]. Dans ces systèmes, les analyses d'erreurs sont réalisées en examinant les fichiers de logs ou bien en corrigeant leurs données de façon à obtenir un jeu d'analyse correct. Des comparaisons sont ensuite menées en comparant les résultats corrects à ceux produits par le système, de façon à compter le nombre d'erreurs décelées. Notons au passage le problème récurrent des problèmes en chaîne, où si la question n'a pas été analysée correctement, les systèmes ont d'énormes difficultés à trouver une réponse. Néanmoins, si ces analyses visent à montrer quel composant du système est responsable de quel type d'erreur, il n'est pas fait mention d'analyse fine des critères linguistiques utilisés et de l'impact sur le système d'erreurs dans leur reconnaissance.

Nous allons essayer de montrer le type d'erreurs précises qui peuvent survenir dans un système de questions-réponses utilisant des techniques de traitement automatique des langues, tout en esquissant des pistes pour diagnostiquer ces erreurs de façon systématique. Nous présenterons des solutions à mettre en œuvre pour chacun des problèmes que nous détaillons dans des encadrés.

2.1.1 Présentation des modules concernés

Notre étude est centrée sur le premier et le dernier module du système FRASQUES, qui sont l'analyse des questions et l'extraction des réponses précises. Il s'agit des deux modules du système dits linguistiques, le moteur de recherche et la sélection de passages reposant essentiellement sur des pondérations de termes.

L'analyse de la question

L'analyse des questions est le module qui permet d'extraire toutes les informations pertinentes de la question, qu'il s'agisse d'informations morpho-syntaxiques intéressantes pour la recherche de passages (les mots significatifs, les noms propres) ou bien des critères ayant un impact pour l'extraction de la réponse (le type de la réponse, le focus, le verbe principal, la catégorie de la question). L'hypothèse qui sous-tend l'extraction de ces informations est qu'une phrase qui contient ces termes (ou bien leurs synonymes) a plus de chances d'être pertinente qu'une autre car elle est susceptible de contenir une reformulation de question et donc de contenir la réponse.

De ce fait, cette étape d'extraction d'éléments est essentielle pour la suite du processus, étant donné qu'ils définissent tout le contexte de la recherche. C'est pourquoi des erreurs dans leur reconnaissance peut poser des problèmes pour l'extraction correcte d'une réponse précise. Ces problèmes peuvent être de plusieurs ordres dans le système FRASQUES [Ligozat *et al.*, 2006] :

- étiquetage morpho-syntaxique incorrect (problème pour l'analyse syntaxique) ;
- analyse syntaxique défaillante (problème pour l'extraction de critères) ;
- extraction des critères erronée (problème pour l'extraction des réponses précises) ;
- problème de reconnaissance d'un nom propre (problème pour l'identifier) ;
- mauvaise catégorisation de la question (processus de résolution non adéquat).

Il faut bien voir que la suite des traitements effectués vont s'appuyer sur les résultats de l'analyse de la question. Les informations que l'on doit extraire pour une question sont présentées dans le tableau 2.1. Nous observons la question *De quelle organisation Javier Solana était-il secrétaire général ?*

Ces informations vont être recherchées dans les passages, et la réponse extraite devra être étiquetée ORGANISATION.

Critères	Informations
Catégorie	quel
Type de réponse attendu	ORGANISATION
Type sémantique	organisation
Verbe principal	être
Focus	Javier Solana
Nom propre	Javier Solana
Extension sémantique	ordre (SYN d'organisation)

Tab. 2.1 Analyse de la question par FRASQUES

L'extraction des réponses

L'extraction de réponses précises étant le dernier module de la chaîne, c'est également celui qui pâtit le plus des erreurs commises lors de l'analyse des questions. En effet, les répercussions sont immédiates : en général, une analyse de la question erronée ne permettra pas que l'extraction de la réponse soit possible.

Voici un exemple de couple question et phrase-réponse, où les mots de la question sont indiqués dans des cadres de façon à bien voir les principes mis en œuvre :

- Question : *Qu'est-il arrivé à l'[ancien parlement de Bretagne] ?*
- Phrase réponse : *Des milliers de Rennais ont défilé dimanche 6 février devant l'[ancien parlement de Bretagne],* ***ravagé par l'incendie*** *qui s'est déclaré dans la nuit du 4 au 5 février.*

Dans cet exemple, on notera la proximité du groupe nominal de la question (*ancien parlement de Bretagne*) avec la réponse au sein de la phrase-réponse. On voit alors l'intérêt d'un patron d'extraction qui s'appliquerait à partir de ce groupe, en s'aidant de l'apposition qui le suit dans la phrase. Les patrons d'extraction sont fondés sur des règles de syntaxe locale qui prennent appui sur des critères extraits de la question. Ils ne peuvent pas s'appliquer si :

- l'étiquetage morpho-syntaxique des phrases candidates est incorrect ;
- les patrons d'extraction sont trop larges ou trop précis ;
- le terme pivot (focus, verbe ou type général) est mal identifié.

Afin d'extraire une réponse, FRASQUES utilise des règles d'extraction qui s'appliquent sur des distances de mots réduites. Le choix des patrons à appliquer est fonction de la catégorie de la question. Si les distances de mots réduites sont garantes de réussite quand les règles s'appliquent, il va sans dire que si les termes pivots de la question sont trop loin (en terme de distance de mots) de la réponse effective, les règles ne serviront à rien. Cette

stratégie nécessite donc une proximité des mots pivots avec la réponse, ce qui dépend de la formulation des phrases candidates.

De façon à bien voir l'importance du module d'extraction des réponses et à mesurer ses performances, il est intéressant de comparer le potentiel de réponses contenues dans les phrases sélectionnées par le système par rapport aux réponses précises effectives extraites. Sur le jeu de questions de Clef07, FRASQUES obtient des phrases qui contiennent la réponse pour 100 questions sur 122 en ce qui concerne le corpus journalistique. Or, il n'extrait que 40 réponses précises, ce qui est moins de la moitié des potentialités du système. Une première observation des résultats montre que la raison principale de la perte de résultats entre les phrases réponses potentielles contenant la réponse et l'extraction de la réponse précise vient du fait que les éléments des phrases sont mal étiquetés. De ce fait, les règles définies ne peuvent s'appliquer.

Nous allons maintenant détailler les principes de l'analyse syntaxique.

2.1.2 L'analyse syntaxique

L'analyse syntaxique de la question sert essentiellement à déterminer les constituants de la question et leurs fonctions syntaxiques. La forme syntaxique obtenue permet par la suite d'identifier les critères à extraire : le type sémantique de la réponse, la catégorie de la question, le focus, etc. Une erreur à ce niveau de l'analyse ne pourra pas être récupérée par la suite.

Les erreurs sont souvent dues à un problème de reconnaissance des locutions. Nous donnons un exemple d'erreur de ce type issue de l'analyse du système FRASQUES à la figure 2.1

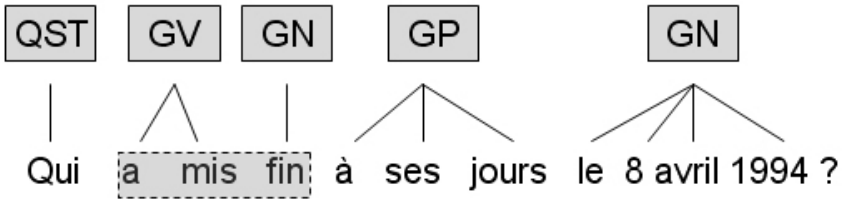


Fig. 2.1 Erreurs d'analyse syntaxique des questions

La locution *mettre fin à ses jours* n'est pas reconnue comme telle. Étant donné que les critères de la question sont identifiés en fonction de cette analyse, le mauvais traitement

de ce phénomène génère ici l'extraction du terme *fin* en focus, ce qui rend inutilisables les règles d'extraction de réponses s'appuyant sur le critère focus.

Mener une étude systématique sur l'impact des erreurs de l'analyse syntaxique supposerait de :
- observer l'analyse syntaxique de la question (filtrer sur ce critère) en affichant les groupes différents par des couleurs différentes ;
- corriger les analyses syntaxiques erronées ;
- relancer le système avec les nouvelles données ;
- comparer les résultats obtenus de façon à mesurer l'impact d'une analyse syntaxique correcte sur le reste du processus ;
- quantifier les résultats (précision) : est-ce que le système extrait plus de bonnes réponses ?

2.1.3 L'étiquetage morpho-syntaxique

L'étiquetage morpho-syntaxique est une étape nécessaire à l'application de règles syntaxiques pour extraire des constituants en fonction de leur place et de leur nature dans les phrases. Les étiquettes permettent ensuite de définir des cas types de formes syntaxiques où l'on ne s'intéressera qu'à certains éléments. Ces erreurs peuvent avoir lieu lors de l'analyse des questions comme lors de l'analyse des phrases candidates extraites par le système. Le problème majeur est que pour apparier les mots de la question avec les mots des phrases-réponses, il faut une cohérence sur l'étiquetage effectué.

Erreurs lors de l'analyse de la question

La figure 2.2 montre des étiquetages morpho-syntaxiques erronés.

Nous avons une erreur sur l'adjectif *chinois*, repéré comme issu du verbe *chinoiser*, de même que pour le verbe *mandaté* reconnu comme un nom. De fait, les seuls verbes reconnus comme tels sont les copules *avoir* et *être*. Ce type d'erreurs a des répercussions pour le processus de résolution de la question, notamment pour apparier des phrases-réponses à la question.

Forme	Etiq	Lemme
Dans	IN	dans
quelle	DT	quel
ville	NN	ville
chinoise	VVZ	chinoiser
le	DT	le
secrétaire	NN	secrétaire
d'	IN	de
Etat	NN	état
américain	JJ	américain
Warren	NP	Warren
Christopher	NP	Christopher
a	VHZ	avoir
-t-il	PP	il
été	VBN	être
mandaté	NN	mandaté
par	IN	par
Bill	NP	Bill
Clinton	NP	Clinton
?	SENT	?

Fig. 2.2 Erreurs d'étiquetage de questions

Mener une étude systématique sur l'étiquetage morpho-syntaxique des questions afin d'en mesurer l'impact supposerait de :
- observer l'analyse de questions en ciblant sur l'étiquetage morpho-syntaxique (jeux de couleurs sur les composants de même nature) ;
- annoter les erreurs d'étiquetage ;
- modifier l'étiquetage ;
- relancer la chaîne de traitement ;
- évaluer l'apport des modifications (performance avec étiquetage correct) et mesurer l'impact des erreurs d'étiquetage.

Erreurs ayant un impact sur l'extraction de réponses

Des problèmes liés à l'étiquetage des formes gênent le processus de reconnaissance des critères de la question (focus, type général, verbe) au sein des phrases-réponses. Ainsi, alors que la correspondance entre question et réponses existe, et que le système possède les règles adaptées au traitement de la réponse courte, il est fréquent que les problèmes d'étiquetage empêchent la résolution.

Les types de problèmes engendrés sont :

- la non reconnaissance des critères de la question (problème d'application des règles d'extraction) ;
- le mauvais étiquetage de la phrase (problème d'appariement question/réponse) ;

Un autre problème concerne l'orthographe et de ce fait la reconnaissance d'un mot sous différentes variantes orthographiques : *Quel évèque fut suspendu par le Vatican le 13 janvier 1995 ?* où les phrases qui contiennent *évêque* avec un accent circonflexe plutôt qu'un accent grave ne verront pas ce mot étiqueté comme le type général extrait de la question.

Mener une étude systématique sur l'étiquetage morpho-syntaxique des phrases réponses supposerait de :
– filtrer les phrases réponses pour lesquelles on n'obtient pas de réponse correcte et observer leur étiquetage ;
– mesurer l'impact de cet étiquetage erroné sur l'extraction des réponses : non-application des règles d'extraction ;
– modifier l'étiquetage ;
– relancer la chaîne de traitement ;
– évaluer l'apport des modifications.

2.1.4 L'extraction des critères et réponses précises

L'extraction de termes, qu'il s'agisse de ceux de la question ou bien de la réponse précise recherchée, se base sur l'analyse syntaxique ainsi que sur l'étiquetage morpho-syntaxique des phrases, qu'elles soient interrogatives ou déclaratives.

Erreurs lors de l'extraction de critères

Les critères sont déterminants pour l'extraction de réponses précises, notamment les termes pivots utilisés lors de l'application des règles d'extraction : le type sémantique de la réponse, le verbe principal et le focus. Nous présentons quelques erreurs existantes à cette étape dans le tableau 2.2.

L'exemple 1 a déjà été expliqué dans le paragraphe précédent. Il s'agit du même phénomène pour l'exemple 2 : le focus *A* n'est pas pertinent pour l'application des règles d'extraction. Dans l'exemple 3, nous perdons l'information qu'apporte l'adjectif *chinois* et le verbe principal réel *mandater* n'est pas pris en compte. L'exemple 4 possède un verbe erroné, qui devrait être *fermer*, et pour l'exemple 5, le type sémantique de la réponse devrait

Questions	**Erreurs**
1) Qui a mis fin à ses jours le 8 avril 1994 ?	Focus = fin
2) Où l'A 340 a-t-il établi le record du plus long vol sans escale ?	Focus = A
3) Dans quelle ville chinoise W. C. a-t-il été mandaté par B. Clinton ?	Verbe = chinoiser
4) Combien de puits sont fermés en Sibérie ?	Verbe = être
5) Avec quel groupe Alcatel annonça sa fusion en mars 2006 ?	Type sémantique = Alcatel

Tab. 2.2 Erreurs d'extraction de critères

être *groupe* (hyperonyme de la réponse), ce qui empêche l'utilisation des entités nommées (on aurait dû rechercher une entité de type ORGANISATION).

Mener une étude systématique sur l'extraction des critères supposerait de :
- observer l'extraction des critères dans l'analyse de la question ;
- modifier les critères erronées ;
- relancer le système avec les nouveaux résultats ;
- comparer les deux versions (avant et après modification) ;
- quantifier l'apport en terme de précision avec une extraction correcte.

Erreurs liées aux règles d'extraction de réponses précises

Les règles d'extraction[1] sont basées sur de la syntaxe locale et nécessitent une proximité syntaxique entre les termes déclencheurs de règles que sont les critères extraits de la question (focus, verbe, type général) et la réponse à extraire. La principale difficulté de ces règles est d'ajuster leur application afin qu'elles ne soient pas trop lâches (ce qui génère du bruit) ni trop précises (extraction de réponse incomplète). Nous présentons un exemple de non-application d'une règle d'extraction sur un exemple tiré de l'évaluation QUAERO 2009 sur l'anglais, obtenu avec le système QALC [Ferret *et al.*, 2001a]. Ce système a le même fonctionnement que FRASQUES, seule la langue change.

- Question : *Who is [Gilbert Charles Stuart] ?*
- Réponse : *In North Kingstown is the [Gilbert Stuart] Memorial, built in 1751, which preserves the birthplace of the famous* ***American portrait painter****.*

Le focus de la question, *Gilbert Charles Stuart*, et la réponse attendue, *American portrait painter*, sont séparés par une incise. La règle d'extraction ne peut pas être conçue sur

[1] Le fonctionnement des règles est décrit plus loin (3.3.1).

ce modèle : cela génèrerait trop de bruit dans les résultats. La distance entre le focus et la réponse est trop grande pour utiliser un patron d'extraction : on sort du domaine d'application de la syntaxe locale.

Mener une étude systématique sur les règles d'extraction de réponses précises supposerait de :
- observer les phrases réponses qui contiennent la réponse attendue et l'application des règles en indiquant celles qui se sont appliquées ;
- étiqueter les phrases qui posent problème ;
- affiner les règles d'extraction en fonction de l'observation précédente ;
- relancer le système ;
- observer l'application des règles sur les phrases étiquetées précédemment ;
- quantifier les résultats en terme de précision.

Cette analyse des erreurs a mis en évidence les étapes et les types d'analyse importants pour mener une évaluation transparente de résultats basée à la fois sur l'analyse de corpus et l'évaluation de performance :

1. Analyse de corpus
   - sélection des données à analyser
   - observation de propriétés
   - étiquetage des données
   - modification des données
2. Évaluation de performance
   - relance du processus
   - comparaison de résultats

Étant donné les erreurs commises par le système, nous avons élaboré une méthodologie d'évaluation fine des résultats de façon à tracer les obstacles à la bonne résolution des questions, notamment en ce qui concerne les résultats de l'analyse des questions et ceux de l'extraction des réponses (tout en surveillant les étiquetages fournis par l'étiqueteur) dont nous avons montré les failles principales. Nous présentons maintenant la démarche à appliquer pour réaliser une évaluation de ce type ainsi que des fonctionnalités que devra posséder notre outil d'évaluation pour la mettre en application.

2.2 Définition de la méthodologie d'évaluation

La première partie de ce manuscrit a montré les besoins d'une évaluation qui porterait sur la dimension linguistique d'un système, et non plus uniquement des résultats obtenus. Ce type d'évaluation suppose d'avoir accès aux différents modules qui composent la chaîne de traitement ainsi qu'aux résultats qu'ils produisent. Cela implique différents savoir-faire qui permettent de suivre une certaine transversalité du système. Plusieurs études sont alors envisageables : tester la pertinence de l'utilisation d'un critère linguistique, regarder précisément à quoi les erreurs du système sont dues, analyser à quel point elles sont améliorables, etc.

2.2.1 Principes

On peut s'interroger sur les différentes façons de tracer une dimension linguistique dans un système de questions-réponses. Si les composants informatiques se décrivent sous la forme d'étapes de traitement (l'analyse des questions, la recherche des documents, la sélection de passages et l'extraction de réponses précises), ils ne reflètent pas les éléments linguistiques à traiter, lesquels sont omniprésents. C'est cette transversalité que nous cherchons à évaluer, indépendamment des modules qui segmentent le processus linguistique. Nous cherchons à réaliser un retour au texte qui reflète les traitements effectués et va permettre de saisir au mieux les phénomènes linguistiques présents.

Il s'agit véritablement d'identifier les points de blocages qui nuisent au bon déroulement du processus, tout en gardant à l'esprit qu'une performance qui atteindrait les 100% n'est tout simplement pas envisageable. Cette idée nous renvoie aux campagnes d'évaluation réalisées sur des sous-tâches de systèmes complexes, l'étiquetage morpho-syntaxique par exemple qui, s'il atteint jusqu'à 95% de réussite, a un taux de réussite en contexte bien inférieur si l'on regarde les résultats obtenus par les SQR. Les phénomènes linguistiques qui transitent tout au long de la chaîne de traitement nécessitent essentiellement d'être repérés afin, dans un deuxième temps, de pouvoir regarder de plus près comment le système les traite automatiquement.

Une évaluation linguistique nous renvoie à deux aspects distincts de l'évaluation : l'analyse de corpus et l'évaluation de performance du système. Ces deux pôles permettent de prendre en compte ce qu'est un système de QR : un système qui traite la langue de façon performante. De ce fait, il n'est pas possible de prendre en compte le traitement des phénomènes linguistiques sans s'assurer que la performance globale ne diminue pas.

Nous allons présenter ce qui nous intéresse dans ces deux champs que sont l'évaluation de performance d'une part et l'analyse de corpus d'autre part afin de dégager les éléments essentiels à une méthodologie d'évaluation de dimension linguistique au travers des résultats d'un système de questions-réponses.

2.2.2 Évaluation de performance

La performance est le critère qui entre en jeu lorsque l'on conçoit un système, quel qu'il soit. En effet, ce sont les résultats qu'il obtient qui permettent de le comparer à d'autres et de mesurer les progrès qui sont à faire. Ainsi, on ne peut pas évaluer un système sans s'assurer que les modifications effectuées n'altèrent pas sa performance globale.

Nous devons prendre en compte le fait d'avoir accès à cette performance lors de l'utilisation de notre outil d'évaluation afin de garantir une cohérence entre prise en charge des phénomènes linguistiques et performance globale du système. Il est fréquent que le changement, même infime, d'une règle permette de traiter correctement par exemple la variation sémantique d'un terme mais qu'elle génère du bruit par rapport à l'ensemble des données à traiter. Il est donc fondamental de vérifier si les changements effectués ne détériorent pas la somme des traitements.

Dans le domaine des systèmes de questions-réponses, la garantie de résultats passe par le calcul de la précision du système, ce qui revient au nombre de bonnes réponses trouvées. Une modification sera jugée acceptable selon les critères d'acceptation de l'utilisateur, étant donné qu'un changement dans le processus pourra à la fois amener de nouvelles réponses et en perdre d'autres.

De façon plus générale, un système de questions-réponses dispose de plusieurs informations : celles qui concernent la question et celles qui concernent la ou les réponses proposées. De plus, on aura besoin de connaissances liées à l'évaluation de la pertinence des résultats obtenus. Ces trois types d'information différentes permettent de classer les informations de la sorte :

- les informations liées aux questions
- les informations liées aux réponses (phrases et réponses précises)
- les informations liées à l'évaluation des résultats obtenus (par modules)

De plus, pour observer les résultats produits par le système, nous aurons également besoin de stocker des informations liées à l'évaluation fine de certains résultats.

Il s'agit donc, à la lumière de ces préoccupations, de rendre compte des aspects qualitatifs d'un système, et, dans le cas d'une modification des traitements, d'un retour au quantitatif des résultats, de façon à motiver les modifications et tracer les différences.

2.2.3 Analyse de corpus

La rencontre de la linguistique et de l'informatique a fait naître des outils de TAL, qui ont permis un certain outillage de la linguistique [Habert, 2005]. Le développement d'outils d'aide au traitement de corpus nous intéresse ici puisqu'il s'agit de fouiller dans le véritable corpus que forment les résultats obtenus par des systèmes de questions-réponses pour en extraire des savoirs sur les systèmes, tout d'abord, mais aussi dans une perspective d'analyse fine pour améliorer la chaîne de traitement.

Si l'on observe les résultats produits comme un corpus à part entière, nous avons besoin d'avoir accès aux données, et cet accès doit être le plus riche possible. En effet, il faut pouvoir annoter ces données, les catégoriser si besoin en fonction des phénomènes rencontrés, les visualiser, etc. Il s'agit de prendre en compte l'aspect qualitatif de la langue et pour pouvoir en rendre compte, il faut être outillé. Nous cherchons à constituer des sous-corpus de données « en contexte » en sélectionnant des sous-corpus contenant les phénomènes que nous voulons étudier.

En terme d'outils d'aide au traitement, il s'agit d'une part de faciliter la visualisation des résultats, en prenant en compte les critères inhérents aux systèmes de questions-réponses et d'autre part de pouvoir les annoter. Comme nous l'avons décrit un peu plus haut, un système de questions-réponses analyse la question pour extraire les éléments qui semblent pertinents à rechercher dans les phrases réponses. On peut d'emblée percevoir l'intérêt d'une évaluation de ces informations, ainsi que de leur présence au sein des phrases réponses. Il faut développer des outils de visualisation et d'annotation pour mener à bien des études fines de données.

Différentes stratégies s'ouvrent alors pour la constitution de corpus :

- évaluer les stratégies mises en œuvre par le système en s'appuyant sur des phrases réponses qui contiennent la ou les réponses attendues en maximisant le nombre de phrases contenant la réponse recherchée (sélectionner les documents pertinents en indiquant au moteur de recherche les mots de la question ainsi que la réponse attendue) ;
- sélectionner les phrases selon l'étude à mener ;

– annoter les phénomènes rencontrés dans les questions et les phrases-réponses pour constituer des sous-corpus dédiés à certains phénomènes linguistiques.

On rejoint ici l'approche mise en œuvre dans [Cohen *et al.*, 2004] mais nous ne pouvons garantir l'exhaustivité du corpus forcé.

Nous allons maintenant discuter des fonctionnalités nécessaires d'un outil qui permettrait de réaliser de telles études.

2.2.4 Synthèse des fonctionnalités requises

Les fonctionnalités requises d'un outil qui permet une évaluation qualitative et quantitative des résultats d'un système de questions-réponses sont les suivantes :

1. Analyse de corpus
   – constitution de sous-corpus d'étude de phénomènes ;
   – sélection des données à analyser dans ces sous-corpus sur des critères fins contenus dans les données ou provenant d'annotations ;
   – observation de propriétés par une mise en relief des différents phénomènes étudiés ;
   – étiquetage des données ;
   – modification des données.
2. Évaluation de performance
   – relance du processus ;
   – réalisation de comptages des phénomènes et des résultats ;
   – comparaison de résultats.

Il s'agit de permettre un accès au corpus de résultats selon ce que l'on veut évaluer, de permettre de mesurer l'impact d'un critère en le modifiant et en testant ce qu'il apporte si l'on relance le SQR avec les résultats modifiés, de comparer les deux versions d'une chaîne, avant et après modification pour mesurer l'apport du nouveau critère tout en ayant accès à la mesure de précision des résultats. L'étude de corpus avec étiquetage des phénomènes linguistiques qui posent problème est une fonctionnalité essentielle pour effectuer un diagnostic des problèmes du système.

Ainsi, pour chaque système évalué, l'idée est d'aboutir à une classification précise des problèmes, de façon à savoir :

1. Ce qui doit être fait (priorité) ;
2. Ce qui coûte trop cher (ressources, temps de calcul, etc.) ;

3. Ce que l'on ne sait pas traiter ;
4. Ce qui n'est pas urgent ;
5. Ce qui ne nous intéresse pas.

Il s'agit précisément d'une prise de conscience des problèmes à dépasser, des potentialités du système ainsi que des obstacles durables pour lesquels il faudrait appliquer d'autres stratégies, tout en étant conscient que la notion de performance pour des systèmes modulaires, où plusieurs traitements s'appliquent sur un même objet, ne peut dépasser un certain seuil de réussite dû à la combinaison de traitements successifs.

L'application de ce type de méthodologie suppose un outil qui permette un stockage des résultats, une visualisation adaptée ainsi qu'un export dans un format adéquat pour relancer le processus de résolution des questions. Nous allons maintenant présenter notre outil, REVISE, qui a été conçu pour répondre à ces besoins.

2.3 REVISE, un outil d'évaluation transparente pour les systèmes de questions-réponses

REVISE est l'acronyme de Recherche, Extraction, VISualisation et Évaluation. Nous allons montrer son application au système de questions-réponses FRASQUES, puis nous détaillerons ses fonctionnalités avant de dérouler un exemple d'utilisation qui permettra d'illustrer le type d'études pouvant être menées avec cet outil. Enfin, nous discuterons des choix effectués ainsi que des limites rencontrées dans ce cadre.

2.3.1 Application à FRASQUES

Nous appliquons la méthodologie d'évaluation définie au système FRASQUES. Le schéma 2.3 montre les points d'évaluation qui nous intéressent au sein des composants du système. En parallèle de ces informations, nous indiquons où les endroits d'intervention de l'outil : les différents modules du système FRASQUES sont indiqués dans des rectangles (analyse de la question, recherche des documents, traitement des documents, pondération des phrases et extraction des réponses).

Les outils d'observations nécessaires à une évaluation transparente des résultats sont présents à droite. Le schéma met également en évidence le type d'informations véhiculées

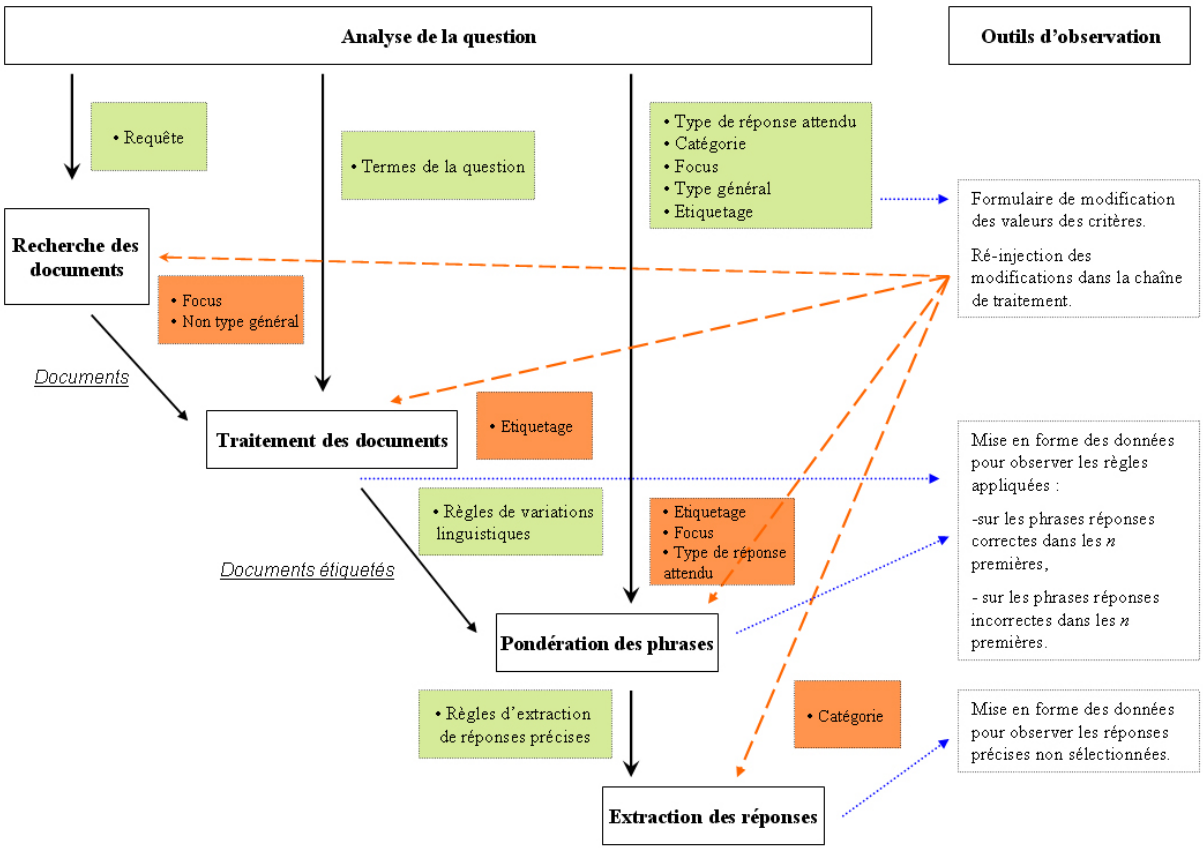


FIG. 2.3 Schéma global de l'évaluation transparente

par le système, et notamment la transmission de critères entre le module d'analyse des questions et les autres. Ces informations sont des résultats intermédiaires extraits du système FRASQUES qui sont sauvegardés dans un fichier XML. Les flèches en pointillés indiquent les informations que nous extrayons et les flèches à tirets montrent les données que nous pouvons ré-injecter dans le système après modification. Ce sont ces informations que nous visualisons de façon précise et qui rendent possible la création d'une fenêtre ouverte sur le comportement du système.

En fonction de l'information véhiculée par le système, REVISE intervient pour la stocker et la modifier si besoin, quel que soit le module informatique concerné. Cet outil permet un accès aux données issues d'un système de questions-réponses tout en conservant la structure ainsi que la granularité de l'information stockée. Cette information devient accessible à l'utilisateur, lequel peut sélectionner les critères de la visualisation qui l'intéresse. Cette visualisation précise permet à l'utilisateur d'intervenir sur l'affichage de l'information à l'aide de jeux de couleurs. De la même façon, l'annotation des information visualisées

permet de constituer des sous-corpus d'étude en fonction des phénomènes annotés.

Nous utilisons également des techniques d'évaluation de type boîte noire, qui permettent de s'assurer que les modifications effectuées ne perturbent pas les résultats de façon globale (précision des réponses obtenues).

Par ailleurs, la sélection des données peut se faire de façon libre ou bien guidée, quand l'utilisateur suit la méthodologie que nous avons implémentée.

En terme de processus, il y a différentes façons de traiter les données tout en gardant le sens des données elles-mêmes. En effet, stocker des informations doit permettre de dégager un savoir sur les informations elles-mêmes, ce qu'on pourrait nommer des métadonnées. Il nous faut garder les informations liées à la chronologie :

- du système qui peut avoir différentes versions (après modifications notamment) ;
- du corpus de documents qui peut avoir changé (*Le Monde*, *Wikipedia*, le *Web*) ;
- des questions (et de la campagne d'évaluation qui peut être liée).

Il faut prendre en compte ces facteurs pour réfléchir au stockage efficace des résultats, tout en sachant que l'outil devra également permettre de comparer différents états de la chaîne de traitement après modification du système.

En effet, l'évaluation d'un processus tel qu'un système de questions-réponses suppose une gestion des flux d'information internes au système qui pourra les dériver (sortir un flux) de façon à permettre une transformation ainsi qu'une réentrance des données vers le système. La mesure des résultats observés fournira matière à une évaluation fine de l'apport des modifications effectuées. De plus, les critères de filtrage des résultats devront être pertinents par rapport aux traitements effectués par le système. Par exemple, pour le système FRASQUES, il devra être possible de filtrer sur la présence ou l'absence des critères extraits lors de l'analyse de la question. De la même façon, il est essentiel de pouvoir modifier certains résultats obtenus par le système si l'on veut relancer par la suite la chaîne de traitement avec les nouvelles valeurs qui auront été ajoutées. Dans l'optique d'une relance de la chaîne de traitement, il faut fournir un export de fichier analysable par le système qui prend en compte les modifications de l'utilisateur sur les résultats initiaux, et stocker à nouveau les résultats obtenus qui prennent en compte ces modifications si l'on veut comparer les choses.

Nous présentons maintenant les choix d'implémentation effectués pour créer les fonctionnalités nécessaires à une évaluation transparente et les discutons.

2.3.2 Implémentation

Nous décrivons l'implémentation de REVISE, qui a été conçu sur le modèle du système de questions-réponses FRASQUES [El Ayari *et al.*, 2009]. Le schéma 2.4 récapitule les technologies utilisées et montre les différentes interactions entre les trois composants de REVISE : le stockage des données (base de données relationnelle), la visualisation des données (XML, XSLT) ainsi que la modification (SQL, PHP) et la génération des données modifiées pour les ré-injecter dans le système (SQL, PERL).

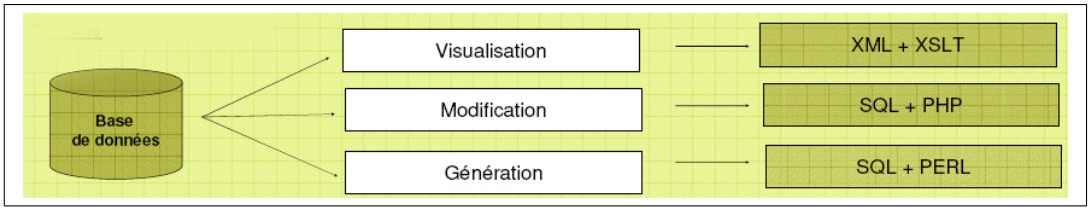


Fig. 2.4 Technologies utilisées

La base de données

REVISE possède une base de données relationnelle où les résultats produits par la chaîne de traitement sont stockés. Les tables ont été créées en fonction des résultats produits par le système FRASQUES, avec des tables différentes en fonction du niveau d'analyse (l'analyse des questions, l'analyse des phrases, les réponses précises, les phrases sélectionnées et leurs scores, les données liées à l'évaluation des résultats). La structure des tables réalisées est indiquée sur la figure 2.5.

Chacune des tables contient l'ensemble des critères extraits lors du traitement des données par FRASQUES. L'analyse des questions extrait des critères comme le focus, le verbe principal, les noms propres, etc. et la sélection des passages réponses extrait des variations, l'étiquetage de chacun des mots, un score, etc. Ces données issues de FRASQUES ont été exportées au format XML (figure 2.6).

Si l'on met en parallèle les figures 2.6 et 2.5, on voit l'adéquation entre le fichier XML et le schéma XML lié. Cette structure de table avec la corrélation critères/attributs permet des sélections fines de données et différents niveaux d'analyse, ainsi que la recherche de ces critères au sein des phrases réponses. Il est important d'avoir la notion de « run » présente dans cette base, de façon à pouvoir stocker différentes versions des résultats (autre campagne d'évaluation, version du système différente, etc.) pour les comparer.

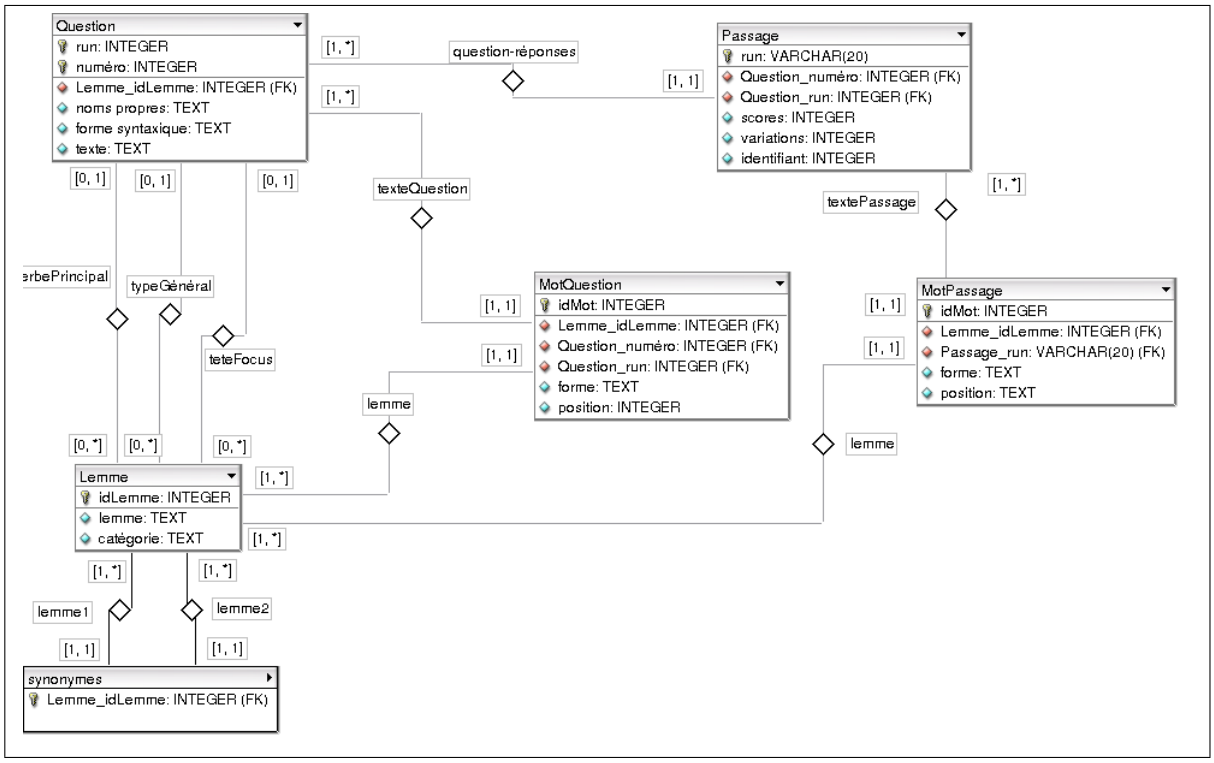


Fig. 2.5 Structure des tables relationnelles

Les données stockées pour nos études représentent 5 jeux de questions différents, que nous avons rassemblés dans le tableau 2.3.

Jeux de questions	Questions	Langue	Système utilisé
CLEF05	200	Français	FRASQUES
CLEF07	200	Français	FRASQUES
QUAERO 2008	500	Anglais	QALC
QUAERO 2009 (dev)	500	Anglais	QALC
QUAERO 2009 (test)	500	Anglais	QALC

Tab. 2.3 Données présentes dans la base de données de REVISE

Le système garde environ 70 phrases réponses pour chaque question, que nous stockons, ainsi que l'analyse de chacune des questions. De la sorte, nous avons pu utiliser REVISE pour diagnostiquer les erreurs du système QALC avant la campagne d'évaluation QUAERO 2009. Les campagnes CLEF consistent en un jeu de 200 questions en langue française et en un corpus de documents de type journalistique (nous l'avons utilisé pour l'étude du focus). En ce qui concerne le projet QUAERO [Quintard, 2008], il s'agit de 500 questions en anglais avec un corpus de documents issus du Web.

La base de données est organisée et permet de filtrer les informations que l'on veut

```
<QUESTION NumQuest="10">
  <ANALYSE>
    <TEXTE_QUESTION lang="F">Avec qui Michael Jackson s'est marié en 1994 ?</TEXTE_QUESTION>
    <CONSTITUANTS>
      <f id="F1" lemme="avec" tag="IN">Avec</f>
      <f id="F2" lemme="qui" tag="WP">qui</f>
      <f id="F3" lemme="Michael" tag="NP">Michael</f>
      <f id="F4" lemme="Jackson" tag="NP">Jackson</f>
      <f id="F5" lemme="se" tag="PP">s'</f>
      <f id="F6" lemme="être" tag="VBZ">est</f>
      <f id="F8" lemme="marier" tag="VVN">marié</f>
      <f id="F9" lemme="en" tag="IN">en</f>
      <f id="F10" lemme="1994" tag="CD">1994</f>
      <f id="F11" lemme="?" tag="SENT">?</f>
    </CONSTITUANTS>
    <FORME_SYNTAXIQUE>GP1quiNP3etreGN5VerbeGP7</FORME_SYNTAXIQUE>
    <CATEGORIE>qui</CATEGORIE>
    <TYPE_GEN/>
    <FOCUS>
      <LEMMES>Michael Jackson</LEMMES>
      <TETE leNum="F4" forme="Jackson" eti="NP">Jackson</TETE>
    </FOCUS>
    <TYPE_EN>PERSON</TYPE_EN>
    <VERBE_PRINCIPAL>marier</VERBE_PRINCIPAL>
    <LISTE_NP>Michael Jackson</LISTE_NP>
    <LEMMES_QUESTION>
      <LEMME literal="marier" cat="VVN" id="F8">
        <SENS>assortir épouser unir</SENS>
      </LEMME>
    </LEMMES_QUESTION>
  </ANALYSE>
  <PHRASES_REPONSES NumQuest="10">
    <PHRASE_REPONSE ID="1">
      <DOCNO NumPhrase="4">ATS.940711.0123.0</DOCNO>
      <POIDS_PHRASE>1081</POIDS_PHRASE>
      <PHRASE>
        <f id="F1" lemme="il" tag="PP">Il</f>
        <f id="F2" lemme="préciser" tag="VVZ">précise</f>
        [...]
      </PHRASE>
      <EN>
        <enamex type="PERSON"> Michael Joseph Jackson </enamex>
        <enamex type="PERSON"> Lisa Marie Presley Keough </enamex>
      </EN>
      <TERMES>
        <mot tag="NP">Michael</mot>
        <mot tag="CD">1994</mot>
        <mot tag="N" variation="XtoX">mariage</mot>
        <mot tag="NP">jackson</mot>
      </TERMES>
    </PHRASE_REPONSE>
  </PHRASES_REPONSES>
</QUESTION>
```

Fig. 2.6 Structure du fichier d'entrée au format XML

visualiser ou compter, en recherchant par catégorie, présence de critères, réponse correcte, etc. Les champs de tables créées permettent d'interroger la base selon certains critères liés aux informations présentes : avec les tables présentées à la figure 2.5, il est possible de sélectionner les données liées à une catégorie en particulier, de rechercher les passages qui contiennent un mot particulier ou encore de ne sélectionner que les passages contenant des variations des mots de la question. Le choix d'une base de données relationnelle per-

met, à l'aide des fonctions du langage SQL, une puissance d'interrogation des données à l'utilisateur.

L'outil permet, en plus de la visualisation, de générer les sorties des requêtes effectuées au format XML également, de façon à rendre exploitables par d'autres outils les données produites. En ce qui concerne les formats d'entrée et de sortie, les résultats du système de questions-réponses sont transformés en XML, puis réécrits sous la forme de fichiers tabulés qui sont insérés dans la base sous la forme de tables SQL (ce qui permet de garder la structure de l'information). La figure 2.7 montre le type de structure du fichier XML.

```
<questions>
 <meta>
  <date>22-09-2009</date>
  <requete>
   SELECT * FROM questions WHERE focus = '' ORDER BY Num
  </requete>
 </meta>
 <question>
  <Num>1</Num>
  <Run>CLEF07</Run>
  <Categorie>où</Categorie>
  <Type_EN>LIEU</Type_EN>
  <Vb_principal>situer</Vb_principal>
  <Liste_NP>Marquises</Liste_NP>
  <Focus>île</Focus>
  <Type_gen/>
  <Texte>Où se situent les îles Marquises ?</Texte>
  <Erreur/>
 </question>
 <[...]>
</questions>
```

Fig. 2.7 Export au format XML

Nous avons opté pour une base de données relationnelle pour sa capacité de stockage, la puissance et l'efficacité de son langage de manipulation SQL et son interaction avec PHP pour créer une interface Web. PHP est un langage qui permet d'interagir de façon simple avec la base de données et avec un temps de traitement minimal, ce qui est un critère important pour ce type d'outil.

La capacité de la base à accueillir différents runs, les résultats de différentes campagnes d'évaluation et d'autant de tests réalisés pour l'amélioration du système est indispensable, de même que la facilité et la rapidité de mise en correspondance des informations stockées dans des tables différentes en fonction du run et des critères à observer. Le choix d'une base de données XML n'aurait pas permis cette efficacité d'utilisation sur de gros volumes. Néanmoins, le fait que la base de données ne soit pas en XML complique l'adéquation du contenu de ce fichier avec le schéma relationnel de la base de données.

Par ailleurs, nous avons fait le choix de n'engendrer avec PHP que du XHTML valide, c'est-à-dire qui respecte les standards d'accessibilité du W3C (*Wold Wide Web Consor-*

tium). Les styles sont définis à l'aide de feuilles de style CSS pour plus de lisibilité.

Sélection des données

L'interface de cet outil suppose d'avoir à disposition différentes options liées à l'observation et à l'évaluation de résultats. Nous sommes partie du principe que l'utilisateur de cette interface doit être guidé tout en gardant une part de liberté dans les choix de données à visualiser. Nous discuterons des aspects concernant la généricité de l'outil, il nous a paru important de permettre à l'utilisateur d'effectuer ses propres requêtes sur la base de données contenant les résultats. En parallèle, l'intérêt de ce travail est de donner les clés d'une évaluation de critères linguistiques, ce pourquoi nous guidons l'utilisateur par la méthodologie que nous avons adoptée et lui offrons des choix prédéfinis pour sélectionner les données.

Les choix prédéfinis sont liés à la visualisation de résultats (complets ou partiels, en fonction de ce que l'on veut observer), avec la possibilité de se focaliser sur les questions, et les résultats obtenus, pour lesquels le système ne trouve pas de réponse correcte. La figure 2.8 montre l'interface développée, avec un menu central permettant de naviguer parmi les différents types d'observation proposés.

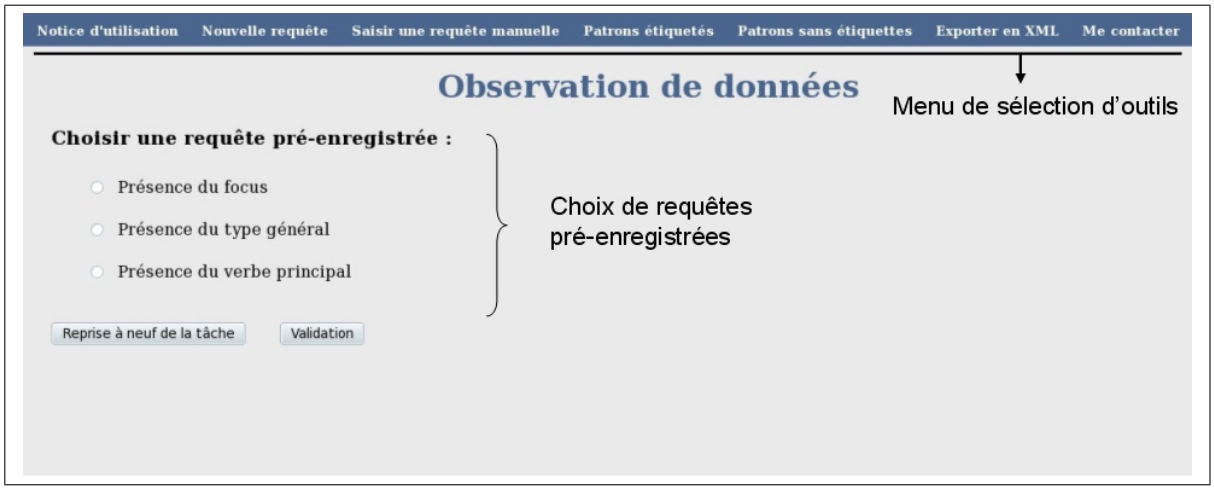


Fig. 2.8 Interface de sélection de données

L'idée ici est de prendre en compte l'évaluation des résultats pour affiner l'observation. S'il peut être utile de savoir quels sont les cas où le système réussit à trouver la réponse, ce qui nous importe le plus est bien évidemment ceux qui présentent des difficultés, donc de filtrer les résultats en fonction de l'évaluation boîte noire : ne sélectionner les données

liées uniquement aux questions pour lesquelles le système n'a pas trouvé la réponse précise attendue.

L'utilisateur peut se laisser guider par les types d'évaluations proposées mais également faire des sélections libres de données. Dans ce cas, les nouvelles requêtes sont stockées à leur tour dans la base de données pour l'enrichir et permettre à l'utilisateur de relancer ces requêtes sans avoir à les refaire.

La visualisation de données

L'aspect visualisation est très important si l'on veut observer les résultats en prenant en compte les éléments pertinents pour faire émerger les problèmes. En comparaison avec des fichiers texte qui contiendraient les traces de l'exécution du système, ce que nous cherchons à faire est de dégager des informations grâce à la mise en relief de certains éléments. Par exemple, une coloration des mots en fonction de leur étiquetage morpho-syntaxique, ou bien des synonymes des mots importants extraits de la question permettra d'emblée de voir si la non-application d'une règle est due à la présence d'un mauvais étiquetage.

La figure 2.9 montre un exemple de visualisation de phrases-réponses pour lesquelles on a indiqué en couleurs les éléments de la question.

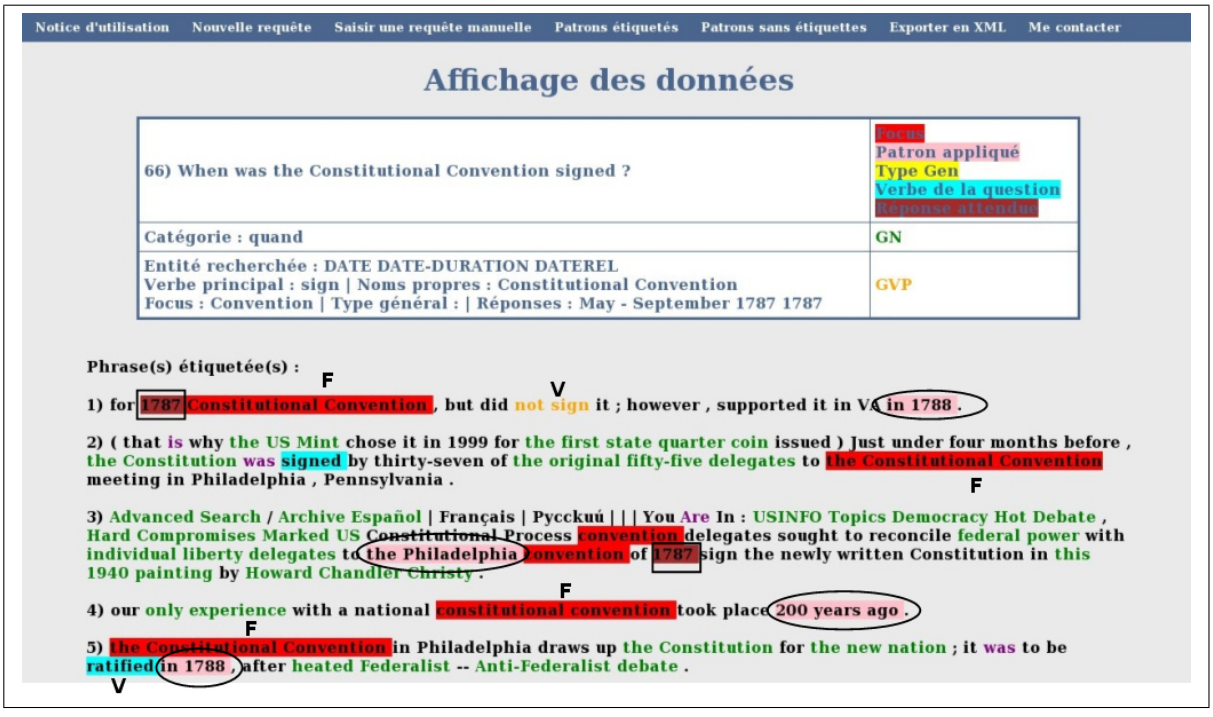


Fig. 2.9 Exemple de visualisation avec jeux de couleurs

Nous faisons apparaître en couleur au sein des phrases réponses sélectionnées par le système les mots de la question qui déclenchentl'application des règles d'extraction : le focus ou bien son synonyme (indiqués par la lettre **F** sur la figure), le verbe principal de la question ou son synonyme également(indiqués par la lettre **V**), puis les règles d'extraction qui se sont appliquées en colorant les mots extraits, ici insérées dans des formes ovales. Ce type d'affichage fournit de lui-même des informations quant aux applications effectives des règles ainsi qu'à propos de la pertinence des verbes comme déclencheur de règles d'extraction (ils ne sont pas encore utilisés dans la définition de règles dans le système FRASQUES). Une version avec les informations morpho-syntaxiques est également disponible pour vérifier si la non-application des règles est liée à des problèmes d'étiquetage. De la sorte, il sera rapide de déceler une règle non appliquée basée sur le terme focus de la question si ce terme n'a pas été étiqueté comme focus dans les phrases-réponses (problème d'accents, de majuscules, etc.).

Pour ce faire, nous extrayons les critères à mettre en évidence ainsi que les couleurs à associer par un formulaire PHP : les visualisations sont paramétrables. De la sorte, il devient possible de travailler sur la visualisation des données de manière à mettre en évidence certains traitements ou bien certains phénomènes. Pour certaines fonctionnalités prédéfinies, comme la visualisation de l'application des règles d'extraction, les critères et leurs couleurs sont prédéfinis également : nous savons d'emblée que l'observation de l'application des patrons passe par la mise en couleur des mots de la question ainsi que de leurs variations et des mots sur lesquels les règles se sont appliquées.

La question de comment trouver une représentation des données satisfaisante n'est néanmoins pas triviale, et semble dépendre à la fois des données (mettre en valeur ce qui est rare/fréquent ?) que du système (quels obstacles semblent primordiaux et réalistes à traiter ?). Il s'agit alors de définir des filtres à appliquer sur les données afin de cibler le plus possible ce qui nous intéresse.

De plus, certains calculs pourront être automatisés de façon à vérifier la précision des résultats (nombre de bonnes réponses trouvées par le système), ou bien pour étayer les choix stratégiques mis en œuvre, comme le calcul de la distance entre la réponse précise et les mots de la question dans les phrases-réponses.

En ce qui concerne la visualisation de données, l'application de jeux de couleurs a été réalisée en PHP avec un script qui récupère le ou les critères choisis ainsi que les couleurs à appliquer. C'est essentiellement par commodité et manque de temps ; en effet, ce choix de style aurait pu être défini avec des feuilles de style CSS.

L'annotation de données

Dans le cadre d'une observation de corpus, nous avons intégré un export des résultats observés dans un format XML, qui garantit une certaine interopérabilité avec d'autres outils afin d'aller encore plus loin dans l'analyse. Il peut s'agir de concordanciers, d'outils d'analyse lexicale (Lexico3[2]), etc.

Il nous a également paru important de laisser l'utilisateur libre d'étiqueter ses données, notamment dans le but de créer des groupes de questions qui pourraient représenter une difficulté particulière, ou bien un degré différent de difficulté pouvant se résoudre à l'aide d'un processus particulier. Bien souvent, il est utile de créer des paquets homogènes afin de tester des hypothèses de traitement. Pour reprendre la discussion entamée un peu plus haut dans la rédaction, s'il n'est pas possible de présager de la difficulté d'une question a priori, l'étiquetage des phénomènes rencontrés dans les phrases-réponses peut alors permettre la création de groupes de questions représentatives d'un phénomène bien précis à traiter. La classification des résultats obtenus est essentielle en terme de méthodologie liée à l'amélioration du système en se focalisant sur les difficultés une par une. Pour ce faire, l'utilisateur peut à tout moment choisir d'étiqueter ses données en cliquant sur un bouton de formulaire pendant qu'il observe le corpus. Ces annotations sont de deux types : soit l'utilisateur annote un phénomène (de façon à créer par la suite un sous-corpus de phénomènes précis à observer), soit il annote une erreur (de façon à mesurer sa portée et à la modifier par la suite). Une table ANNOTATION permet de stocker le phénomène ou l'erreur en lien avec ses critères : « run », numéro de question, etc.

La modification de données

Comme nous venons de le voir, REVISE permet de stocker les résultats grâce à une base de données SQL, d'effectuer des requêtes qui permettent ensuite de filtrer ces résultats en fonction des phénomènes qui nous intéressent.

Outre la sélection de données, REVISE offre la possibilité de modifier les résultats obtenus, de façon à relancer le système en aval de ces modifications. Ce processus permet plusieurs choses : tester par exemple le potentiel d'un module avec une analyse des questions parfaite, mais aussi évaluer la pertinence d'un critère stratégique si on change sa définition, sans avoir à modifier le système lui-même. Les données modifiées ainsi que les critères liés

[2]Disponible sur : http ://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/

à la modification sont stocker dans une table MODIFICATION de même structure que la table d'origine des données modifiées.

Enfin, dans un souci de traçage de ce qui est effectué, nous offrons à l'utilisateur la possibilité d'insérer des commentaires concernant ses observations, de façon à nourrir la base de données. Ces observations pourront permettre de garder trace des études menées par soi-même ou bien par d'autres, de pouvoir les consulter, les approfondir ainsi que mutualiser les réflexions.

Enfin, une fois le système relancé avec les modifications, il devient à nouveau possible d'observer les nouveaux résultats et de les comparer aux résultats antérieurs.

De la sorte, nous proposons un outil ainsi qu'une méthodologie d'évaluation des résultats de systèmes de questions-réponses dans un cadre d'utilisation de techniques de traitement automatique des langues. Il est évident qu'un SQR basé sur des techniques statistiques n'aura pas le même intérêt d'évaluation de phénomènes linguistiques ponctuels qui posent problème, étant donné que leur résolution de questions passe par des raisonnements quantitatifs et non qualitatifs. Toutefois, notre outil pourra servir à déceler et quantifier les phénomènes linguistiques à traiter.

Après avoir présenté l'application de REVISE au système de questions-réponses FRASQUES, nous en donnons des exemples d'utilisation afin de montrer le type d'études pouvant être réalisées avec cet outil sur les résultats d'un système de questions-réponses.

2.3.3 Exemple d'utilisation

REVISE a été conçu pour permettre d'effectuer certains types d'analyses des résultats obtenus par un système de questions-réponses. Ainsi, il est possible de réaliser des observations, par exemple, en :

- filtrant les phrases réponses qui ont une certaine taille ;
- s'intéressant uniquement à une catégorie donnée, etc. ;
- se focalisant sur les critères extraits de la question comme le focus, le type général, le verbe ;
- analysant les variations sémantiques des termes de la question au sein des phrases réponses ;
- comptant le nombre de phrases réponses qui contiennent tel élément de la question.

Nous allons détailler un problème particulier afin de montrer comment utiliser REVISE sur des cas précis :

– Comment améliorer le traitement des questions de catégorie COMBIEN ?

Ce type d'étude sera réalisé par étapes, notamment en vérifiant d'abord la pertinence de l'analyse des questions de ce type pour ensuite observer les phrases-réponses.

L'interrogation de la base de données

Il s'agit tout d'abord de filtrer les questions de la catégorie qui nous intéresse (COMBIEN). Il est possible :

– soit de se laisser guider par les requêtes SQL pré-existantes ;

– soit d'en constituer une.

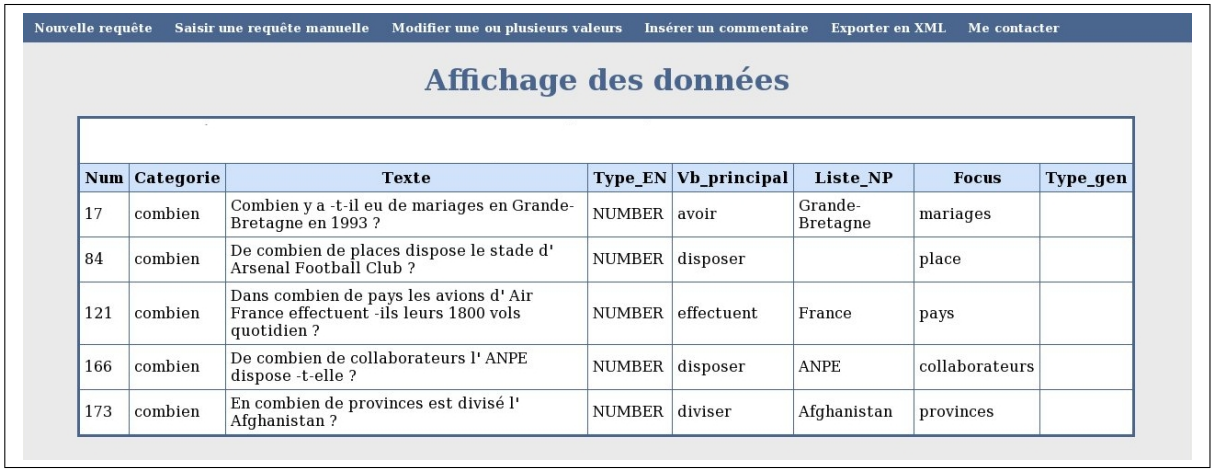
Les deux possibilités sont présentes sur la figure 2.10.

Fig. 2.10 Requêtes sur la base de données

La requête SQL suivante :

`SELECT * FROM questions WHERE Categorie='Combien'`

insérée dans REVISE produira l'affichage de la figure 2.11.



Num	Categorie	Texte	Type_EN	Vb_principal	Liste_NP	Focus	Type_gen
17	combien	Combien y a -t-il eu de mariages en Grande-Bretagne en 1993 ?	NUMBER	avoir	Grande-Bretagne	mariages	
84	combien	De combien de places dispose le stade d' Arsenal Football Club ?	NUMBER	disposer		place	
121	combien	Dans combien de pays les avions d' Air France effectuent -ils leurs 1800 vols quotidien ?	NUMBER	effectuent	France	pays	
166	combien	De combien de collaborateurs l' ANPE dispose -t-elle ?	NUMBER	disposer	ANPE	collaborateurs	
173	combien	En combien de provinces est divisé l' Afghanistan ?	NUMBER	diviser	Afghanistan	provinces	

Fig. 2.11 Filtrage et affichage des questions de catégorie COMBIEN

Cette requête montre l'analyse des questions effectuée pour chacune des questions de catégorie COMBIEN. L'extraction des critères est correcte, toutes les questions attendent une entité de type NOMBRE, et disposent d'un focus et d'un verbe corrects. On voit également que pour les questions de ce type, le focus représente l'unité car la réponse attendue est de la forme <NOMBRE> suivi du focus.

L'affichage des résultats

Nous avons mené ensuite une étude question par question. Nous regardons les phrases-réponses qui ont été extraites par le système pour la question *En combien de provinces est divisé l'Afghanistan ?* de façon à mesurer l'importance de l'extraction des mots de la question pour sélectionner une phase-réponse valide. Pour ce faire, la requête[3] effectuée est :

```
SELECT questions.Num, Texte, Id, Phrase FROM questions, phrases NJ Run, Num WHERE Run='Clef7_2' AND questions.Num=173
```

Les résultats présentés à la figure 2.12.

Num	Texte	Id	Doc	Phrase
173	En combien de provinces est divisé l' **Afghanistan** ?	1	ATS.950312.0078.1	Ils avaient volé de victoire en victoire ces six derniers mois , prenant le contrôle de neuf provinces de le **Afghanistan** (sur **trente**) avant de parvenir à la mi-février aux portes de Kaboul .
173	En combien de provinces est divisé l' **Afghanistan** ?	2	ATS.950210.0055.0	Les Talibans , qui ont surgi sur la scène politique afghane à le automne dernier , avec la prise de Kandahar , le ancienne capitale royale , ne ont cessé depuis lors de gagner du terrain dans le sud de le **Afghanistan** où ils contrôlent huit provinces .
173	En combien de provinces est divisé l' **Afghanistan** ?	3	ATS.950212.0028.0	Si la chute du Logar se confirmait , le mouvement , surgi à le automne dernier à la frontière pakistanaise , au sud , contrôlerait à présent neuf provinces de le **Afghanistan** sur **30** .
173	En combien de provinces est divisé l' **Afghanistan** ?	4	ATS.940103.0037.0	D'abord limités à Kaboul , les combats se sont étendus lundi à plusieurs provinces du nord de le **Afghanistan** , ont affirmé lundi à Islamabad des diplomates afghans et des opposants .
173	En combien de provinces est divisé l' **Afghanistan** ?	5	ATS.950312.0078.1	Contre lui , il a les Ouzbeks du général Rashid Dostam , qui occupent plusieurs provinces du nord de le **Afghanistan** , les Hazaras chiites , défaits à Kaboul mais irréductibles dans leurs bastion du Centre , et surtout un " pays pachtoune " unifié dans sa plus grande partie derrière les taliban .
173	En combien de provinces est divisé l' **Afghanistan** ?	6	LEMONDE94-001001-19940609.0	De violents combats ont lieu dans plusieurs provinces à le ouest et à la est de le **Afghanistan** entre troupes loyales au président Rabbani et forces du premier ministre intégriste Hekmatyar , alliées à celles du général ex-communiste Dostom , ont rapporté des voyageurs arrivés mardi 7 juin à Kaboul .
173	En combien de provinces est divisé l' **Afghanistan** ?	7	ATS.950220.0064.0	Deuxièmement que seuls les " bons musulmans " puissent participer au processsus de paix , et enfin que des représentants des **30** provinces afghanes soient invités dans le nouveau conseil dirigeant .

Fig. 2.12 Phrases-réponses obtenues

La figure 2.12 montre les phrases réponses trouvées par le SQR où le focus et la réponse apparaissent en couleur. Cet affichage permet de voir si les mots de la question (affichés en couleur) sont présents en même temps que la réponse, de façon à s'assurer que les termes extraits sont pertinents pour la recherche de phrases réponses.

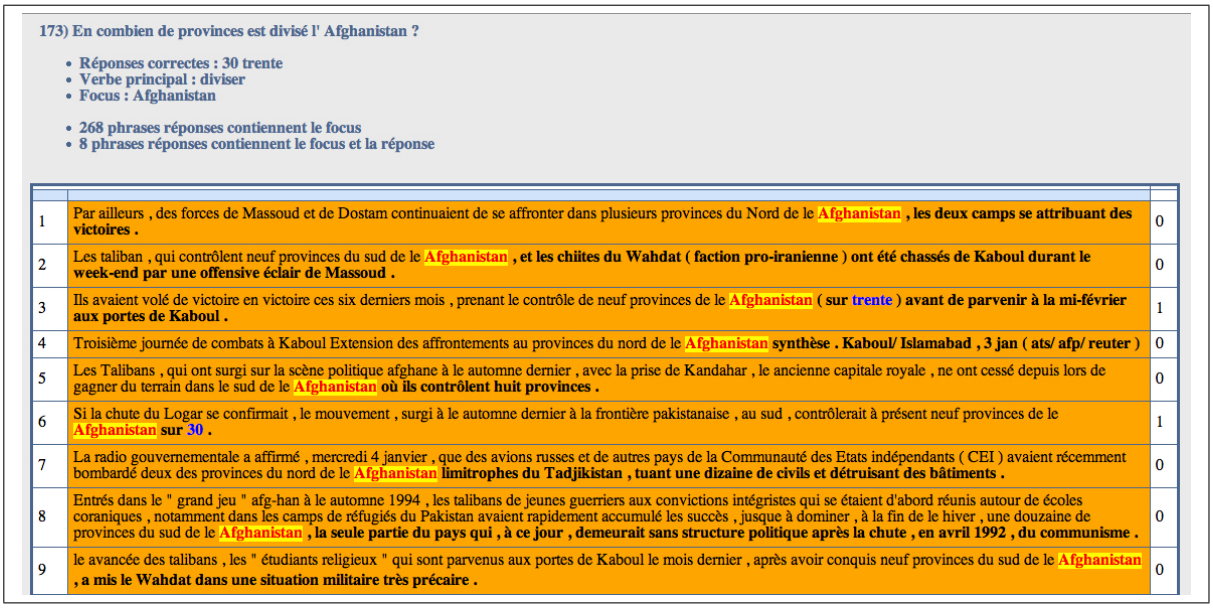
[3]NJ est l'abréviation de NATURAL JOIN, qui permet de réaliser des jointures entre tables avec SQL.

Compter les phénomènes

Il est également possible de compter les phrases qui contiennent la réponse et le focus, avec la requête :

```
SELECT COUNT(Num) FROM questions, phrases, reponses NJ Run, Num WHERE Phrase LIKE questions.Focus AND reponses.Reponse
```

C'est ce que montre la figure 2.13, où sont affichées toutes les phrases qui contiennent le focus déterminé par le système. Des comptes sont réalisés de façon automatique pour déterminer le nombre de réponses qui contiennent le focus et la réponse attendue de la question. Ces comptes permettent de juger efficacement de la pertinence de la présence du focus pour la recherche des réponses précises.



173) En combien de provinces est divisé l' Afghanistan ?

- **Réponses correctes : 30 trente**
- **Verbe principal : diviser**
- **Focus : Afghanistan**

- **268 phrases réponses contiennent le focus**
- **8 phrases réponses contiennent le focus et la réponse**

1	Par ailleurs , des forces de Massoud et de Dostam continuaient de se affronter dans plusieurs provinces du Nord de le Afghanistan **, les deux camps se attribuant des victoires .**	0
2	Les taliban , qui contrôlent neuf provinces du sud de le Afghanistan **, et les chiites du Wahdat (faction pro-iranienne) ont été chassés de Kaboul durant le week-end par une offensive éclair de Massoud .**	0
3	Ils avaient volé de victoire en victoire ces six derniers mois , prenant le contrôle de neuf provinces de le Afghanistan **(sur trente) avant de parvenir à la mi-février aux portes de Kaboul .**	1
4	Troisième journée de combats à Kaboul Extension des affrontements au provinces du nord de le Afghanistan **synthèse . Kaboul/ Islamabad , 3 jan (ats/ afp/ reuter)**	0
5	Les Talibans , qui ont surgi sur la scène politique afghane à le automne dernier , avec la prise de Kandahar , le ancienne capitale royale , ne ont cessé depuis lors de gagner du terrain dans le sud de le Afghanistan **où ils contrôlent huit provinces .**	0
6	Si la chute du Logar se confirmait , le mouvement , surgi à le automne dernier à la frontière pakistanaise , au sud , contrôlerait à présent neuf provinces de le Afghanistan **sur 30 .**	1
7	La radio gouvernementale a affirmé , mercredi 4 janvier , que des avions russes et de autres pays de la Communauté des Etats indépendants (CEI) avaient récemment bombardé deux des provinces du nord de le Afghanistan **limitrophes du Tadjikistan , tuant une dizaine de civils et détruisant des bâtiments .**	0
8	Entrés dans le " grand jeu " afg-han à le automne 1994 , les talibans de jeunes guerriers aux convictions intégristes qui se étaient d'abord réunis autour de écoles coraniques , notamment dans les camps de réfugiés du Pakistan avaient rapidement accumulé les succès , jusque à dominer , à la fin de le hiver , une douzaine de provinces du sud de le Afghanistan **, la seule partie du pays qui , à ce jour , demeurait sans structure politique après la chute , en avril 1992 , du communisme .**	0
9	le avancée des talibans , les " étudiants religieux " qui sont parvenus aux portes de Kaboul le mois dernier , après avoir conquis neuf provinces du sud de le Afghanistan **, a mis le Wahdat dans une situation militaire très précaire .**	0

Fig. 2.13 Observation et comptages des critères de résolution au sein des phrases-réponses

Néanmoins, le problème du biais induit par le système pour les phrases-réponses obtenues rend difficile la généralisation des résultats obtenus. Il est difficile de disposer d'un corpus avec des phrases qui contiennent la réponse à notre question et qui n'aient pas été extraites par notre système, de façon à évaluer la portée d'un critère sans prendre en compte notre stratégie d'interrogation du corpus de test. Nous essaierons de remédier à cette lacune au chapitre 3, de façon à rendre notre étude linguistique un peu plus objective.

Exporter les résultats

Enfin, il est possible d'exporter les résultats obtenus, ou les modifications effectuées le cas échéant, sous la forme d'un fichier XML. Ainsi, si nous souhaitons conserver les résultats obtenus pour la catégorie COMBIEN, il suffit de cliquer sur le bouton du menu prévu à cet effet. Le résultat obtenu est présenté à la figure 2.14.

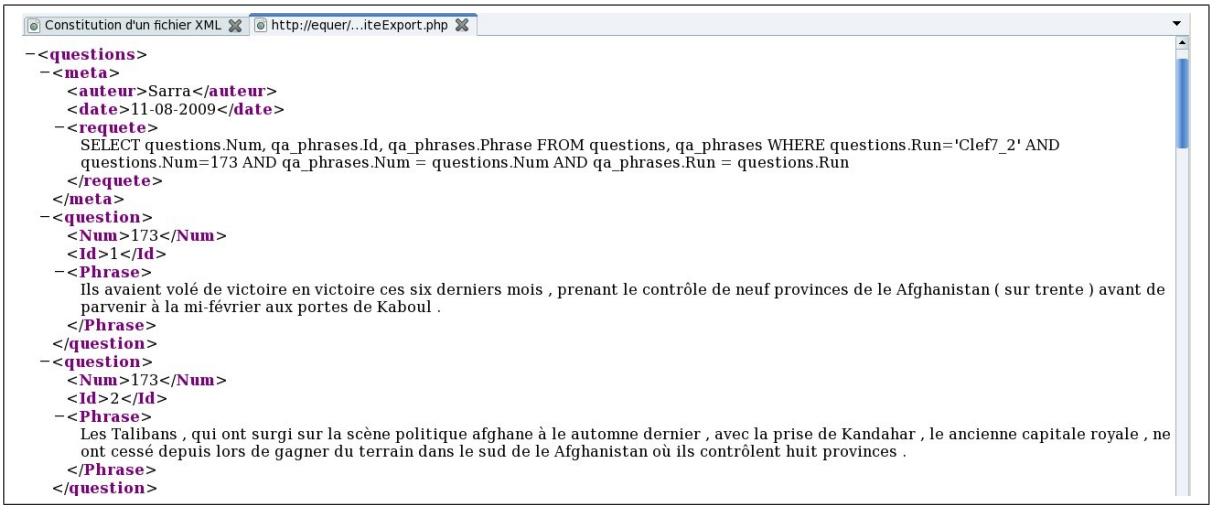


Fig. 2.14 Export des résultats en format XML

Cet export donnera lieu à la relance du système de questions-réponses afin de prendre en compte les modifications effectuées. La ré-injection des résultats obtenus dans la base de données permettra la réalisation d'une étude comparée des résultats obtenus avant et après modification.

Après avoir illustré le fonctionnement de REVISE, nous discutons du choix des techniques choisies pour la conception de l'outil, ainsi que des limites observées.

2.3.4 Discussion

Nous ne permettons pas à l'utilisateur de choisir le format de données à insérer dans REVISE : il doit fournir un fichier de format XML bien formé pour les fichiers d'entrée et de sortie de la base de données. XML permet une certaine interopérabilité des données ainsi que la garantie d'une bonne formation du document (pas d'erreurs de saisie), ainsi que la possibilité de conserver la structure arborescente liée aux traitements effectués (par exemple, balise mère <analyse_question> qui englobe tous les traitements effectués lors de cette étape).

Certains aspects n'ont pas été développés, essentiellement par manque de temps, notamment la création de session utilisateurs dans l'optique d'une utilisation de l'outil par plusieurs personnes. L'idée est de permettre aux utilisateurs de bénéficier de l'expérience des autres, en mettant en commun des requêtes de travail, voire des résultats d'étude tout en préservant les observations et modifications de chacun.

Un autre point intéressant que nous n'avons pas mené à son terme consiste à permettre une visualisation parallèle de données issue de deux versions différentes du système afin de comparer les résultats. Le but serait de pouvoir choisir la meilleure des versions et par la suite de valider telle ou telle modification du système en ayant préalablement vérifié les apports. Il sera important d'avoir accès aux résultats qui diffèrent mais aussi au nombre de réponses correctes obtenues (précision) afin de pouvoir s'y référer.

Nous avons esquissé les différentes possibilités qu'offre REVISE, en termes de modularité de traitement des résultats issus d'un système de questions-réponses de façon à montrer ce qu'il permet de réaliser. Comme nous l'avons montré précédemment, la problématique de l'évaluation transparente touche tous les systèmes. C'est pourquoi nous allons discuter la mise en place de la généricité de cet outil, de façon à ce qu'il soit utilisable pour d'autres systèmes de questions-réponses.

2.4 Développement de la généricité

L'évaluation transparente dépend essentiellement de l'architecture du système analysé ainsi que des stratégies mises en oeuvre. REVISE permet de prendre en entrée les fichiers de données qui sont produits par un système donné pour en systématiser l'évaluation. Néanmoins, si un système de questions-réponses diffère d'un autre quant aux stratégies utilisées, les résultats produits seront également de structure différente. De la même façon, l'évaluation pourra porter sur d'autres éléments que ceux que nous avons approfondis pour FRASQUES.

Nous présentons les principes d'application de la généricité sur REVISE, généricité dont nous expliquerons l'application sur les données d'un autre système de questions-réponses, RITEL, avant de discuter des perspectives d'amélioration de cet outil.

2.4.1 Principes d'application de la généricité

L'idée ici est de montrer que cette méthodologie et notre outil sont applicables à d'autres systèmes de questions-réponses que ceux sur lesquels nous nous sommes basée lors de notre étude. Nous partons du principe que n'importe quel système de questions-réponses qui dispose de résultats intermédiaires peut explorer ses données à l'aide de REVISE. Pour ce faire, le fichier doit comporter les éléments extraits lors de l'analyse des questions ainsi que les phrases-réponses sélectionnées. De la sorte, il est possible de projeter les critères extraits des questions dans les phrases-réponses afin d'évaluer leur portée. La méthodologie mise en place permet soit de guider l'utilisateur dans ses évaluations en lui proposant notre méthodologie d'évaluation, soit de le laisser libre.

Quels sont les éléments à prendre en compte pour rendre REVISE plus générique et donc s'abstraire du format et des données des systèmes étudiés ? Voici deux questions que nous nous sommes posées afin d'étendre notre réflexion sur l'évaluation des systèmes de questions-réponses.

2.4.2 Mise en œuvre de la généricité

Pour resituer notre propos, nous disposons d'une base de données dans laquelle stocker des résultats de SQR, de scripts PHP qui permettent d'interroger cette base et de visualiser les résultats, de modifier les données et de les exporter dans le format de départ (fichier XML) et enfin d'annoter les résultats pour constituer des sous-corpus d'étude. Nous allons déterminer les questions qui se posent pour ces points.

REVISE est un outil qui permet de :

1. stocker des résultats dans une base de données relationnelle ;
   Le schéma de la base étant lié à la structure et à la nature des résultats, se posent les questions suivantes :
   - Comment créer une base de données automatiquement ?
   - Comment concevoir un schéma relationnel correct selon le format XML d'entrée ?
   - Comment remplir les tables automatiquement ?
2. effectuer des requêtes SQL sur la base de données ;
   Se posent les questions suivantes :
   - Comment créer des requêtes génériques ?
   - De quelle manière guider l'utilisateur pour explorer ses données ?

3. visualiser les résultats, à l'aide de jeux de couleur ;
   Se posent les questions suivantes :
   - Qu'est-ce qu'un utilisateur doit pouvoir visualiser ?
   - Sous quelle forme ?
4. annoter et modifier les données.
   Se posent les questions suivantes :
   - Quel format adopter pour le stockage ?
   - Comment avoir accès aux annotations dans un deuxième temps ?

Conception de la base de données

Comment concevoir une base de données de façon automatique à partir d'un fichier de résultats ? C'est la question que nous nous sommes posée, notamment dans le cadre du stage d'Alice Gio [Gio, 2009]. L'idée retenue consiste à générer de façon automatique le schéma des tables de la base de données en fonction d'un fichier XML de résultats que l'utilisateur devra fournir. Si l'on part du principe qu'un système de questions-réponses produit des données lors de l'analyse de la question (critères importants extraits) ainsi que lors de la sélection et de l'extraction des documents (passages ou réponses, scores, réponses précises, etc.), la structure de la base de données sera produite selon un algorithme qui crée une table pour chaque élément XML qui contient des sous-éléments et fait de chaque sous-élément un attribut de cette table (voir la figure 2.15). Si cette solution n'est pas optimale, elle est automatique et permet de ré-engendrer le fichier XML après modifications.

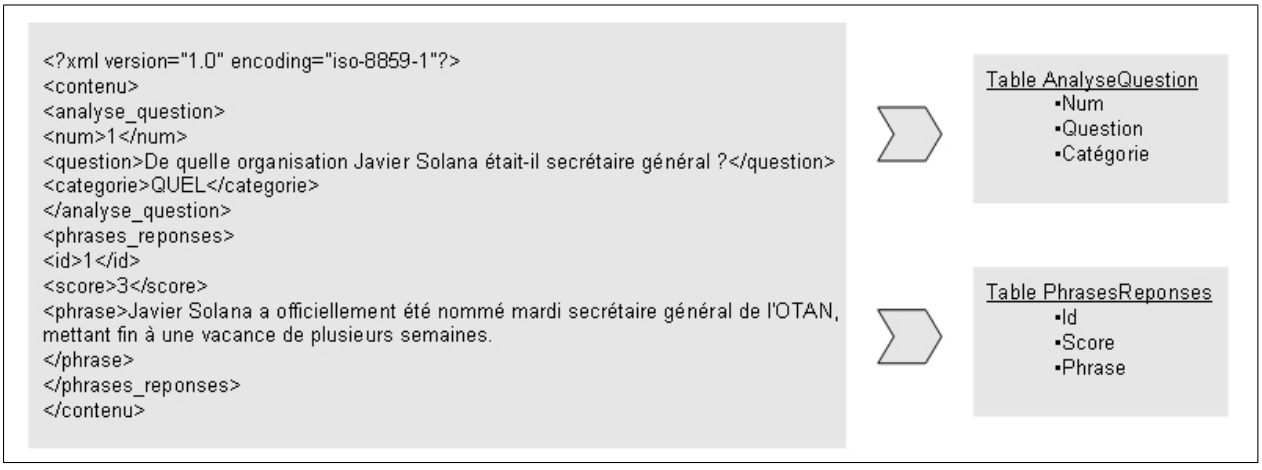


Fig. 2.15 Génération automatique d'une base de données relationnelle

L'utilisateur doit pouvoir modifier le schéma relationnel engendré ou bien concevoir lui-même le schéma (solution optimale pour la base, mais pose un problème pour ré-engendrer le fichier), notamment dans le cas où un attribut aurait été transformé en une table (cas d'un attribut composé lui-même de sous-éléments, comme la forme, le lemme et l'étiquette morpho-syntaxique par exemple).

Stratégie d'interrogation de la base de données

Pour l'évaluation des résultats produits par les systèmes FRASQUES et QALC, nous avions conçu les requêtes en fonction des besoins. Ici, il s'agit de guider l'utilisateur et de lui proposer :

1. de sélectionner les critères qu'il veut observer ;
2. de sélectionner les données qu'il veut observer ;
3. de spécifier la nature des critères (absence, présence, etc.).

Si l'on prend une requête SQL de base, il est possible de l'écrire sous la forme d'une équation, que l'utilisateur va nous aider à résoudre : SELECT (A) FROM (B) WHERE (C) où A représente les critères à observer, B les données sélectionnées et C les conditions de sélection. De la sorte, il faut proposer à l'utilisateur de remplir ces trois champs, à l'aide de formulaires où les choix proposés seront instanciés en allant chercher les attributs des tables concernées. Les requêtes obtenues sont stockées dans la base de données pour faciliter leur lancement à nouveau mais aussi pour garder trace des études menées.

Visualisation de données

Comment permettre à un utilisateur de choisir des jeux de couleurs sur les données qu'il a choisi de visualiser ? Pour ce faire, nous avons mis en place un formulaire qui permet de sélectionner les données une zone de définition des jeux de couleurs lors de la sélection des données. L'utilisateur choisit dans les critères qu'il a sélectionné auparavant ceux qu'il veut mettre en relief et de quelle manière. On s'appuie pour cela sur les critères définis par la base de données (soit ceux extraits des questions lors de leur analyse, soit ceux extraits des phrases réponses) en fonction de ce qui doit être visualisé.

Annotation et modification

L'annotation et la modification de données nous semblent deux points très importants pour une démarche d'évaluation. L'annotation de corpus nous semble être un moyen simple de constituer des sous-corpus de travail, de façon à tester son système sur ce qu'il traite avec difficultés. L'autre point est complémentaire, et permet soit d'évaluer le fonctionnement d'un module (par exemple l'extraction des réponses) avec une analyse des questions correcte (puisque validée manuellement), soit de tester une autre stratégie de résolution sans avoir à modifier le système pour autant.

L'interface est constituée de deux accès : l'un à la visualisation et à l'étude des résultats, avec possibilité d'annotation (et de ce fait, possibilité de création de corpus de questions en fonction des difficultés et des phénomènes linguistiques rencontrés) ; l'autre à la modification des données (dont le but est de ré-engendrer le fichier XML de départ avec les nouvelles valeurs pour que l'utilisateur puisse relancer son système de questions-réponses et tester l'impact des modifications effectuées).

Après avoir présenté les principes généraux de la généricité, nous décrivons son application sur les résultats produits par un autre système de questions-réponses, RITEL. Si la constitution de la base de données a été réalisée, l'étude des résultats est à l'état de spécifications.

2.4.3 Application au système de questions-réponses RITEL

Présentation du système RITEL

RITEL, acronyme de Recherche d'Informations par TELéphone, est un système de dialogue, c'est-à-dire un système qui traite des questions posées oralement [Rosset *et al.*, 2006]. Ce système est développé au LIMSI, au sein du groupe *Traitement du Langage Parlé* (TLP). En revanche, si l'on déconnecte le module de reconnaissance vocale, nous obtenons un système de questions-réponses prêt à fonctionner sur des données écrites.

RITEL utilise des techniques de traitement automatique des langues, notamment en ce qui concerne l'analyse des questions. Les éléments identifiés lors de cette analyse sont regroupés sous la forme de descripteurs de recherche (DDR) [Galibert, 2009]. Ces descripteurs contiennent :

– les éléments de la question à rechercher dans les documents ;

- les types de réponses attendus à rechercher ;
- des indices de pondération des éléments, en fonction de leurs variations (un nom complet aura plus de poids que le prénom ou le nom seuls, de même que le terme exact de la question sera privilégié par rapport à un synonyme).

Par ailleurs, RITEL utilise une reconnaissance très fine des entités nommées, lesquelles ont été étendues au maximum. Ce sont ces entités qui permettent l'extraction des réponses précises : il s'agit de sélectionner l'entité du type attendu par la question qui obtient le score le plus élevé. Pas d'utilisation de syntaxe locale pour l'extraction ici : la stratégie repose sur l'étiquetage des entités. Nous montrons sur la figure 2.16 issue de [Rosset *et al.*, 2006] les types d'entités reconnus par le système.

Entités nommées	<_org> NIST </> <_eve> festival de Cannes de 2006 </> qui a dit <_cit> veni vidi vici </>
Entités non précises	<_Eve> festival de Cannes </> le <_Pers> president </> a déclaré ...
Entités étendues multi-niveaux	Fonctions, titres (président, professeur, évêque...) couleurs, animaux...
Super classes hiérarchiques	évêque → fonction religieuse → fonction
Marqueurs thématiques	Je m' intéresse aux <_litterature> romans </> de ... qui a gagné le <_sport> Mondial </> de 1998
Marqueurs interrogatifs	<_Qqui> qui </< a écrit ce livre <_Qmesure> de combien </> d' heures dure ...
Marqueurs d'interaction	<_DA_close> au-revoir </> <_DA_yes> oui s'il vous plaît </>
Mots composés	les <_NN> logiciels de base de données </> sont ... les <_NN> élections multi-raciales </>
Chunks verbaux	il a <_action> gagné</> ... ils <_action> prendront part à </> ...
Entités linguistiques	<_adj_comp> le plus gros </> exportateur... cela se produit <_adv> souvent </> quand ...

Fig. 2.16 Types d'entités reconnues par le système RITEL

On voit ici l'intérêt d'un outil qui permette de visualiser à la fois le contenu des descripteurs de recherche, la présence de ces éléments dans les phrases réponses ainsi que le typage des entités.

Conception de la base de données

Pour constituer la base de données, nous avons appliqué la méthode explicitée plus haut qui consiste à générer les tables automatiquement en partant du fichier XML de résultats engendré par RITEL (figure 2.17) pour la création des tables. Nous avons donc créé deux tables, l'une qui contient l'ensemble de l'analyse des questions et l'autre les éléments liés à la sélection de phrases réponses.


```
<ritel>
 <analyseQ>
  <id>001</id>
  <question>qui a écrit Au Pays des poissons captifs ?</question>
  <etiq>
   <_Qpers> <_Qqui> qui </_Qqui> </_Qpers>
   <_pers_act> <_Aauteur> <_aux> a </_aux> <_action> écrit </_action> </_Aauteur> </_pers_act>
   <_loc> <_ville> Au </_ville> </_loc>
   <_pers> <_nom> Pays </_nom> </_pers>
   <_det> des </_det>
   <_type_animal> poissons </_type_animal>
   <_adjectif> captifs </_adjectif>
   <_punct> ? </_punct>
  </etiq>
  <classe>pers</classe>
  <methode>ddr</methode>
  <ddr>
   <poids>1</poids>
   <type>critique</type>
   <element>
    <poids>1</poids>
    <trans>identite</trans>
    <type>pers_act</type>
    <texte>a écrit</texte>
   </element>
   <element>
    [...]
   </element>
  </ddr>
 </analyseQ>
 <reponses>
  <score>1.66306</score>
  <type>pers_comp</type>
  <texte>Nedim Gursel</texte>
  <phrase>
   <_det> le </_det> <_Organisation> parti </_Organisation> <_acronym> AKP </_acronym> <_action> a </_action>
   <_adv> très </_adv> <_adv> vite </_adv> <_adjectif> compris </_adjectif> <_conjs> qu' </_conjs> <_pronom> il </_pronom>
   <_adv> ne </_adv> <_action> pouvait </_action> <_DDnon> pas </_DDnon> <_action> réaliser </_action> <_det> son </_det>
   <_subs> programme </_subs> <_pronom> s' </_pronom> <_pronom> il </_pronom> <_action> commençait </_action> <_prep> par </_prep>
   <_action> remettre </_action> <_prep> en </_prep> <_subs> cause </_subs> <1> un </1> <_det> des </_det> <_subs> piliers </_subs>
   <_action> disons </_action> <_prep> de </_prep> <_det> la </_det> <_subs> république </_subs> <_prep> de </_prep> <_loc>
   <_pays> Turquie </_pays> </_loc> <_conjc> et </_conjc> <_det> la </_det> <_subs> laïcité </_subs> <_conjc> mais </_conjc>
   <_punct> , </_punct> <_loc_adv> en même temps </_loc_adv> <_det> ce </_det> <_Organisation> parti </_Organisation> <_aux> a </_aux>
   <_action> rapproché </_action> <_det> la </_det> <_loc> <_pays> Turquie </_pays> </_loc> <_punct> . </_punct>
  </phrase>
 </reponses>
</ritel>
```

Fig. 2.17 Fichier XML fournit par RITEL

Nous disposons de deux tables : l'une pour l'analyse des questions, l'autre pour la sélection des phrases réponses et l'extraction de réponses précises. Les attributs des tables, qui sont les critères extraits (balises encadrées sur la figure 2.17), vont permettre l'évaluation de ces critères par la suite.

Stratégie d'interrogation de la base de données

Sur les données du système RITEL, en reprenant les éléments mis à jour lors de la description du système, il est intéressant d'évaluer le contenu des descripteurs de recherche,

d'observer leur présence au sein des phrases-réponses ainsi que de vérifier les étiquetages d'entités fines. Les requêtes pourront être guidées, en proposant la méthodologie que nous avons développée, ou bien libres.

L'interrogation de la base de données est fondée sur la structure du schéma relationnel défini lors de la conception de la base. Par exemple, si l'on veut observer l'analyse des questions qui recherche une entité de type PERSONNE, il suffira de sélectionner les différents critères qui nous intéressent (A), sur les données de l'analyse des questions (B) en spécifiant que la classe doit être de type PERSONNE (C). Ceci nous donne la requête suivante :

```
SELECT Id, Question, Classe, Methode FROM table_question WHERE Class='pers'
```

Les requêtes effectuées (les attributs seront proposés en parcourant le schéma des tables de la base) seront stockées dans la base de données, de façon à pouvoir les relancer sans repasser par le formulaire de sélection de critères et de corpus. De la sorte, il sera aussi possible de réaliser une méthodologie pour un système donné en fonction des requêtes effectuées.

Dans le cas de l'observation des descripteurs de recherche, il suffira de modeler les requêtes sur les critères présents dans la table liée à l'analyse des questions. Pour ce qui est de leur visualisation au sein des phrases réponses, il s'agira de projeter ces critères dans les phrases-réponses et de les mettre en valeur. Enfin, la vérification de l'étiquetage des entités pourra se faire en affichant les étiquettes en plus des termes dans les phrases réponses.

Visualisation de données

En ce qui concerne la visualisation de données, celle qui est proposée permet l'affichage du résultat de la requête effectuée sous la forme d'un tableau XHTML, avec une ligne pour chaque entrée et une colonne pour chaque champ sélectionné. La visualisation permet également la mise au point de jeux de couleurs laissant à l'utilisateur le choix des informations qu'il veut mettre en relief et de quelle façon. Dans ce cas précis, il s'agira de lister les critères liées aux question et aux phrases réponses et de leur appliquer un style que l'utilisateur pourra définir par un formulaire. La visualisation prendra ces éléments en compte, tout en laissant la possibilité de mettre en relief d'autres éléments sur la page de visualisation elle-même. Il est important que l'utilisateur puisse modifier la visualisation en fonction de ce qu'il observe.

Annotation et modification

L'annotation consiste en la création d'une table ANNOTATION, dans laquelle les commentaires sont stockés. L'utilisateur peut choisir la nature de son annotation (donnée erronée, phénomène particulier, etc.) ainsi que les données sur lesquelles porte son commentaire. Cette annotation doit permettre dans une deuxième temps la création de sous-ensembles de corpus, afin d'affiner l'étude à des phénomènes particuliers. Dans le cas de RITEL, il sera possible d'annoter les questions unes par unes, afin de constituer un sous-corpus d'analyse erronée (descripteurs de recherche incorrects).

La modification des résultats observés est effectuée en modifiant le schéma relationnel initial : une nouvelle table est crée qui contient les données modifiées. Ce sont ces données qui remplaceront les données initiales lors de la génération d'un fichier pour relancer le système et le tester avec les nouvelles valeurs. La modification de données est effectuée en créant un champ supplémentaire correspondant au critère modifié. Sa valeur est instanciée par ce que l'utilisateur veut y mettre. L'outil sélectionnera les valeurs non nulles du champ adéquat de cette table lors de l'export de fichier. La modification du sous-corpus de questions dont l'analyse est erronée pourra permettre la génération du fichier initial, de façon à relancer RITEL avec des descripteurs de recherche corrigés.

En ce qui concerne l'annotation de données, nous distinguons trois types d'annotation :

- annotation d'une étude, qui correspond au stockage d'une requête ;
- annotation d'un phénomène, qui correspond à une phrase ou une question ;
- annotation d'une erreur, qui correspond à un critère.

De la sorte, l'utilisateur peut annoter différents objets en fonction de ce qu'il veut en faire. Le stockage d'une requête permettra de relancer une étude particulière qui correspond à certains phénomènes auxquels l'utilisateur s'intéresse, le stockage d'une phrase ou d'une question (qui revient à identifier une ou plusieurs lignes particulières) permettra de constituer un corpus d'éléments relevant d'une particularité commune. Enfin, l'annotation d'erreurs liées à un critère particulier (un descripteur de recherche par exemple) permettra la constitution d'un corpus d'éléments à modifier. Ces annotations pourront également être accompagnées de commentaires de façon à garder trace des motivations de l'utilisateur et pourront donner lieu à des recherches d'études par ce biais.

Nous allons maintenant discuter des problèmes qui restent dans REVISE.

2.4.4 Conclusion et discussion

Le premier problème concerne la conception de la base de données. Nous avons proposé une solution pour que l'outil la conçoive automatiquement en partant du fichier XML contenant les résultats produits par un système de question-réponses. Or, cette solution a le défaut de créer des tables pour chaque élément XML qui contient des sous-éléments. Il est possible d'y remédier en laissant à l'utilisateur la possibilité de détruire ces tables inutiles et d'ajouter les critères dans une table plus appropriée. Néanmoins, le fait que la création de la base de données dépendent uniquement du contenu du fichier XML pose la question suivante : comment concevoir un fichier XML contenant des résultats imbriqués de façon satisfaisante pour créer un schéma relationnel satisfaisant également ? Cela demande une réelle réflexion de l'utilisateur sur les informations auxquelles il veut avoir accès pour évaluer son système. En effet, si l'étiquetage des mots de la question n'est pas présent au sein du fichier, une étude de l'impact de l'étiquetage ne pourra pas être menée. L'utilisation de REVISE nécessite donc d'être au clair avec le type d'études à mener lors de la conception du fichier de données.

Une deuxième question, qui reste un peu en suspend, concerne le moyen optimal d'offrir à l'utilisateur toutes les possibilités qu'offre REVISE sur une seule page, c'est-à-dire comment lui permettre, au fil de ses observations, d'annoter un phénomène, de préciser sa sélection de données en fonction, de revenir à un niveau plus général d'observation, etc. Par exemple, l'utilisateur s'intéresse aux questions de catégorie COMBIEN, veut tracer le traitement d'une seule de ces questions, annoter le phénomène rencontré (par exemple, le verbe n'est pas bien reconnu) puis revenir à l'observation de toutes les questions de catégorie COMBIEN. Cette navigation pose question quant au développement de l'outil et concerne son ergonomie.

De la même façon, il est important que l'utilisateur aie une trace des différentes études qu'il a réalisées : la création d'un historique des études menées serait un plus en terme de méthodologie d'évaluation et de traçage, non plus des phénomènes eux-mêmes, mais de la méthode. C'est par ce biais que nous pourrons constituer de nouvelles méthodes d'observation de phénomènes, et leur stockage dans la base de données sous la forme de métadonnées liées aux résultats pourra enrichir fortement la pratique de l'évaluation de sous-corpus de résultats à proprement parler.

Nous avons présenté dans cette section les erreurs classiques rencontrées par des systèmes qui utilisent des techniques de traitement automatique de la langue, et mis en évi-

dence les fonctionnalités nécessaires à une étude systématique des systèmes en prenant ces difficultés en compte : c'est-à-dire disposer d'un outil qui puisse réaliser des études quantitatives et qualitatives à la fois, alliant visualisation précise des données, étiquetage des phénomènes rencontrés et modification des résultats produits par le système. Ainsi, il est possible de tracer un critère sur l'ensemble de la chaîne de traitement, ainsi que de systématiser son étude.

Nous allons maintenant décrire l'application de notre méthodologie aux modules d'analyse des questions et d'extraction des réponses précises, menée sur les systèmes de questions-réponses FRASQUES et QALC.

Chapitre 3

Étude d'enjeux linguistiques

Sommaire

Nous présentons ici différentes méthodes d'évaluation transparente de problèmes relevant du domaine de la linguistique, que sont le critère focus et les règles d'extraction de réponses précises, à l'aide de notre outil, REVISE. Nous nous sommes intéressé de près à la notion de focus, pré-existante à notre étude dans le système FRASQUES [El Ayari, 2007], de façon à préciser sa portée et mesurer son importance pour toutes les catégories de questions. Nous présentons une définition augmentée du focus, puis menons une étude basée sur cette redéfinition de façon à en mesurer la pertinence. Enfin, nous étudierons de plus près les règles d'extraction de réponses en contexte.

3.1 Etude d'un paramètre : le focus

L'approche d'extraction de réponses précises dans le système FRASQUES repose sur le focus : il s'agit d'un élément de la question qui doit être présent dans la phrase-réponse en relation syntaxique avec la réponse précise.

3.1.1 Définition du terme focus

En linguistique

La notion de focus prend ses origines dans la linguistique et fait écho en phonétique où il est défini comme un constituant qui porte l'accent saillant de la phrase (élément accentué à l'oral par le biais de l'intonation) ou encore en syntaxe : constituant mis en exergue dans une phrase clivée. Une phrase clivée correspond à la dérivation d'une phrase simple à l'aide des éléments *C'est ... que/qui*. Par exemple :

- Phrase initiale : *Jean a cassé le pot de fleur.*
- Phrase clivée : *C'est* ***Jean*** *qui a cassé le pot de fleur.*

Jean est ici le focus de la phrase, c'est-à-dire l'élément nouveau important, celui dont on parle. Ces disciplines considèrent que le focus est un élément informationnel important de la phrase. Par contre, il n'y a pas trace d'études sur la notion de focus appliquée à une forme interrogative

Pour les systèmes de questions-réponses

Wendy Lehnert a été la première à appliquer ce concept de focus à l'étude des questions pour les systèmes de questions-réponses [Lehnert, 1978]. Elle définit alors le focus comme le concept de la question qui représente le besoin d'information exprimé par la question. C'est-à-dire qu'il désigne l'entité réponse.

Plusieurs systèmes de questions-réponses ont intégré la reconnaissance du focus à leur traitement de la question, et en ont donné une définition différente : c'est l'objet à propos duquel on cherche une information, cette information étant la réponse. Cette définition entraîne les propriétés suivantes :

- Pour [Ferret *et al.*, 2002] : « l'élément important de la question, qui devra se trouver à proximité de la réponse »

- Pour [Plamondon *et al.*, 2002] : « une portion de la question qui doit obligatoirement figurer près du candidat-réponse [...]. Par exemple, le focus de la question *What was the monetary value of the Nobel Peace Prize in 1989 ?* serait *Nobel Peace Prize* car l'hypothèse est faite que la réponse correcte devrait se trouver la proximité de l'expression Nobel Peace Prize ou d'une expression sémantiquement apparentée ». Leur système est XR3 7, premier système de question-réponse développé à l'Université de Montréal.
- Pour [Mendes et Moriceau, 2004] : « l'élément le plus important de la question i.e. le focus ».
- Dans [Ravichandran et Hovy, 2002], les auteurs ne discutent pas d'un focus mais ont déterminé des *Qtargets* qui sont des classes définies soit par des types de réponses attendues (hyperonyme de la réponse), soit par des relations entre question et réponses. Le principe de sélection d'une information dans la réponse qui va pouvoir servir de pivot lors de l'extraction de la réponse précise au sein d'une phrase réponse se retrouve ici également.

Ces définitions, qui viennent de différentes équipes de recherche, mettent en relief l'intérêt de la reconnaissance d'un terme particulier qui doit se trouver dans la phrase réponse : le focus. Si ce terme est présent, il est intéressant de regarder la relation syntaxique et la proximité « physique » qui peut exister au sein de la phrase réponse entre ce terme focus et la réponse à la question.

Le focus dans le système FRASQUES

Au sein du système FRASQUES, le focus est habituellement défini comme « un mot de la question qui devrait idéalement être présent dans la phrase réponse »[Ferret *et al.*, 2002]. Sa reconnaissance automatique est basée sur la forme de la question, et il correspond le plus souvent au sujet de la question (ou bien à son complément dans le cas d'une forme passive). Ainsi, la question *Quel slalom Alberto Tomba a-t-il remporté le 6 février 1994 ?* aura comme focus le terme *Alberto Tomba*.

De façon à compléter cette définition, au niveau du module d'extraction des réponses, des règles ont été créées pour extraire la réponse précise attendue en se basant sur la position du terme focus dans la phrases. La prise en compte des verbes est gérée à ce niveau, par la constitution de règles sur le modèle : FOCUS + CONNECTEUR + REPONSE, où le connecteur peut être un signe de ponctuation, une préposition ou bien un verbe.

Nous proposons aujourd'hui une définition plus linguistique du terme focus pour les systèmes de questions-réponses en unifiant la dichotomie nom/verbe pré-existante et de façon à donner une définition claire de ce terme quelles que soient les catégories de questions observées.

3.1.2 Réalisations linguistiques

Notre définition

Une phrase est définie par Z. Harris comme l'ensemble du prédicat et de ses arguments : le sujet et les éventuels compléments [Harris, 1976]. Cette théorie est reprise par M. Gross, pour qui « la formalisation des phrases en terme de fonctions ou prédicats, et de variables ou arguments, est une activité courante en linguistique. [...] Ces descriptions reposent toutes sur l'hypothèse que le verbe est une fonction, et que les termes qui en dépendent sont des variables » [Gross, 1981]. L'idée défendue ici est que le prédicat verbal est l'élément essentiel de la phrase, celui sur lequel repose le propos de la phrase. Les arguments du verbe (quand ils sont exprimés) sont essentiels à l'expression du sens du prédicat : « un prédicat est un sens qui a des « trous » pour recevoir d'autres sens » [Mel'cuk *et al.*, 1995].

Appliqué aux systèmes de questions-réponses, et donc à des questions de type factuel (c'est-à-dire qui portent sur une entité souvent instanciée ou sur un événement particulier) nous définissons le focus, en fonction du contexte, comme :

- soit **l'entité sur laquelle porte la question** ;
- soit **le procès exprimé dans la question**.

Le choix de la nature du focus est fonction de la formulation de la question, laquelle pourra chercher à obtenir de l'information sur une entité (personne, organisation, etc.) ou bien à propos d'un événement (mariage, création, etc.).

Nous établissons la typologie illustrée à la figure 3.1 pour les questions, les focus sont indiqués entre crochets.

Une **entité** peut être exprimée en intension ou en extension. Par exemple, la question *De quelle organisation Javier Solana était-il secrétaire général ?* illustre les deux aspects : une entité exprimée en extension, c'est-à-dire nommée : *Javier Solana* et l'autre en intension, c'est-à-dire qu'on fait référence à quelqu'un par une propriété : *secrétaire général*. Dans cet exemple, le focus sera l'entité exprimée en intension (*secrétaire général*) car c'est sur la fonction que la question est posée. La phrase réponse *Javier Solana a officiellement*

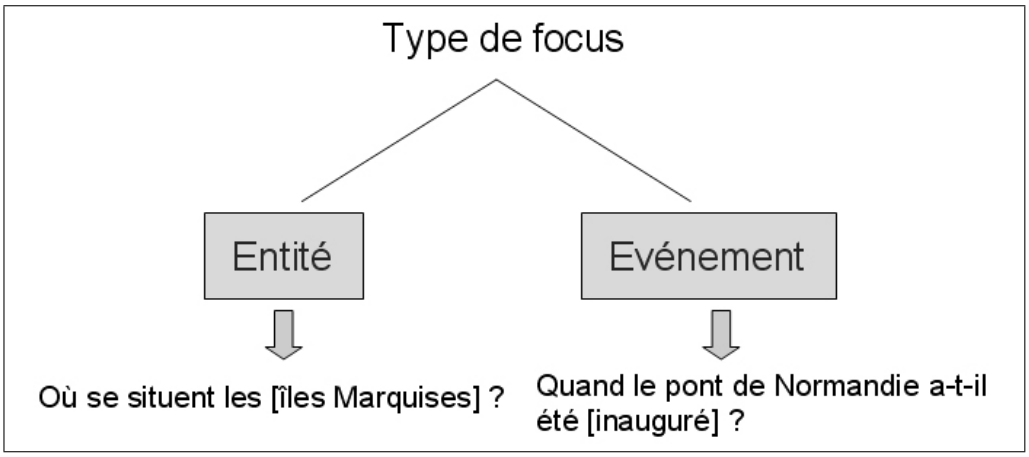


Fig. 3.1 Typologie de questions pour le focus

été nommé mardi [secrétaire général de l'OTAN], mettant fin à une vacance de plusieurs semaines. montre que la réponse est liée à ce terme focus dans la phrase réponse.

Un deuxième exemple, *A quel parti appartient Thérèse Aillaud ?* où il n'y a pas de verbe porteur d'événement. Le focus sera *Thérèse Aillaud*, entité à propos de laquelle la question est posée.

Dans le cas de l'expression du focus sous la forme d'un **procès** (événement exprimé), une question contient une prédication à plusieurs places, dont l'une n'est pas remplie : c'est la place manquante qui constitue la réponse recherchée. Nous allons dérouler un exemple afin d'illustrer notre propos.

– Avec qui Michael Jackson s'est-il marié en mai 1994 ?

Nous avons ici le procès suivant : **SE MARIER AVEC(Michael Jackson, x)**, où la variable **x** représente la réponse recherchée. Nous cherchons à identifier le procès de la question pour en extraire l'argument manquant dans la réponse, en suivant l'hypothèse selon laquelle la réponse sera syntaxiquement liée au procès. Si une question comprend un prédicat à plusieurs places dont l'une est vide, il nous est alors possible de déduire des relations entre ce prédicat et l'argument manquant. Ainsi, pour la question *Avec qui Michael Jackson s'est-il marié en mai 1994 ?* on peut noter la relation entre le verbe et ses arguments de la sorte : **SE MARIER AVEC**(Per1, Pers2).

A. Polguère se pose la question d'une représentation satisfaisante du contenu d'une phrase, et opte pour le réseau sémantique. Il identifie le prédicat central, qui est celui autour duquel gravite le message exprimé, qui sera le nœud racine de l'arbre créé [Polguère, 2008].

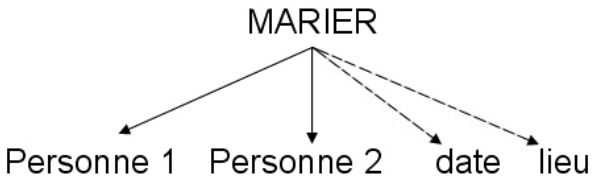


Fig. 3.2 Représentation du procès

Comme l'illustre la figure 3.2[1], nous avons un verbe : **SE MARIER AVEC** qui peut avoir un lieu, une date, et différents acteurs (ici au nombre de deux). Nous nous intéressons à la question *Avec qui Michael Jackson s'est-il marié en mai 1994 ?* qui renseigne déjà un des deux acteurs (*Michael Jackson*) ainsi que la date (*mai 1994*). L'argument qui nous intéresse *Pers2* est manquant et constitue la réponse que nous recherchons.

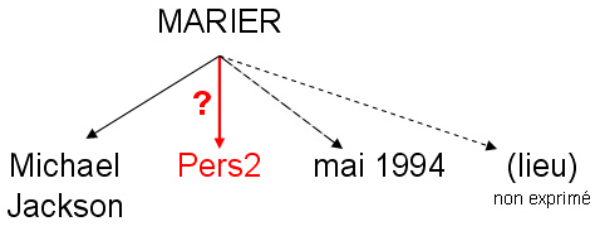


Fig. 3.3 Équation du procès

La figure 3.3 représente l'équation du procès telle qu'elle est formulée dans la question. L'argument manquant est une des deux personnes. Nous avons indiqué la date, qui est renseignée, ainsi que le lieu, qui n'est pas exprimé. Ces deux éléments sont arguments optionnels (indiqués en pointillés sur la figure) qui rentrent dans la catégorie des circonstants (leur présence n'est pas nécessaire à la complétude du sens du prédicat). Néanmoins, ces éléments précisent l'événement dont il est question, et peuvent aider à la désambiguïsation des réponses, si par exemple un même événement a eu lieu à des endroits différents et/ou à des temps différents (pour l'exemple qui nous intéresse, la date permet de préciser de quel mariage il s'agit).

Nous allons maintenant affiner la définition du focus en fonction des différentes catégories de questions, lesquelles reflètent des besoins d'information différents, et de ce fait des formulations différentes également.

[1] Les flèches en pointillés indiquent les circonstants.

Particularités en fonction des catégories de questions

Dans nos systèmes de questions-réponses, la catégorie de la question correspond au focus exprimé, au type de réponse attendu et à sa relation syntaxique avec le focus. Nous avons défini des catégories en fonction du type de focus exprimé. Le focus sera exprimé de deux façons : soit sous la forme d'un procès, soit sous la forme d'une entité. Nous discutons des particularités de la reconnaissance du focus en fonction des catégories.

Les questions de type DÉFINITION ne posent pas de question par rapport à la définition du focus que nous avons donnée. Il s'agit le plus souvent de définir un objet, et c'est donc cet objet qui sera le focus de la question. Dans la question *Qu'est-ce que l'accélération centrifuge ?* le focus est *accélération centrifuge*.

Pour les questions de type COMBIEN nous extrayons deux valeurs de la question : l'unité du chiffre recherché (euro, pourcentage, mariages, etc.) et le terme focus extrait de la question. Nous montrons un exemple, où l'unité est indiquée en italique et le focus en gras :

- Combien de *puits* ont dû être **fermés** suite à la rupture d'un oléoduc en Sibérie ?
- La rupture d'un oléoduc dans le gisement de Samotlor, dans l'ouest de la Sibérie, a contraint les autorités à **fermer** 52 *puits* pour empêcher toute extension de la pollution, a rapporté lundi la télévision russe.

On voit que l'unité de la réponse recherchée (ici *puits*) constitue un indice fort pour extraire la réponse attendue, de même que l'événement (ici *fermeture*) dont il est question. La réponse attendue est encadrée par ces deux éléments dans la phrase réponse, ce qui tend à justifier la prise en compte de l'unité et du focus. Dans ce cas-là, des patrons d'extraction sont constitués autour de ces deux éléments et par défaut autour de l'unité.

Les questions de type INSTANCE expriment le type sémantique de la réponse (la réponse est un hyponyme). De la sorte, elles ne contiennent pas de focus et une stratégie différente pour extraire la réponse axée sur la vérification de ce type, est mise en place. Nous donnons deux exemples de ce type de questions : *Quelle est la plus grande banque du Japon ?* ou *Que s'est-il produit en Algérie dans la nuit du 17 au 18 août 1994 ?* On voit que *la plus grande banque du Japon* est une description de la réponse cherchée, de même qu'un événement qui aura pour justification de s'être *produit en Algérie dans la nuit du 17 au 18 août 1994* correspondra à la réponse de la deuxième question.

Les questions qui attendent un complément du verbe en réponse sont celles qui commencent par les interrogatifs *qui*, *où* et *quand*, ou bien *que* suivi d'un verbe. Nous utiliserons notre définition augmentée du focus pour ce type de questions, de même que pour les questions de type QUEL (*Quel évêque fut suspendu par le Vatican le 13 janvier 1993 ?*. Il s'agira de reconnaître si le verbe exprime un événement (verbe de sens plein) ou bien si c'est l'entité qui est questionnée, qu'elle soit exprimée en intension ou en extension.

Si cette nouvelle définition permet de traduire la notion d'événement, elle pose également la question de la reconnaissance du procès au sein de la question et de sa formulation du procès de la question au sein des phrases réponses, avec des variations de la forme de surface beaucoup plus importantes que pour les entités. Nous allons discuter de ces phénomènes de variation en contexte.

Variations en langue

La représentation de surface du procès, c'est-à-dire sa formulation, peut varier dans la phrase réponse par rapport à celle que nous identifions dans la question. Par exemple, la phrase-réponse *Or voilà que son président, Sir Richard Greenbury, invite Tony Blair, le leader du Parti travailliste, intronisé le 21 juillet, à déjeuner avec lui dans le courant du mois de septembre.* exprime le même événement que celui de la question *Qui a pris la tête du Parti Travailliste au Royaume Uni le 21 juillet 1994 ?*, bien que la formulation varie.

Le prédicat et ses éventuels arguments identifiés dans la question pourront être exprimés, au sein de la réponse, sous une forme nominale ou adjectivale directe (sans variation sémantique) ou bien sous une forme variante (synonyme, conversif, locution). On dit d'un mot qu'il est le conversif d'un autre « si et seulement si leur sémantisme est identique et les actants sont inversés » [Mel'cuk *et al.*, 1995].

Nous reprenons notre exemple premier *Avec qui Michael Jackson s'est-il marié en mai 1994 ?* et allons maintenant observer les phrases réponses qui contiennent la réponse (elle est indiquée en gras). Le procès est indiqué entre crochets.

1. Le chanteur américain Michael Jackson et [son épouse], **Lisa-Marie Presley**, sont arrivés vendredi soir en jet privé à Budapest.
2. La star américaine Michael Jackson [s'est mariée] le 26 mai dernier en République dominicaine avec **la fille de Elvis Presley**.

3. Le 26e jour du mois de mai 1994 ont contracté [le mariage] civil Michael Joseph Jackson et **Lisa Marie Presley Keough**.
4. **Lisa Presley** confirme avoir [épousé] Michael Jackson.

Dans le contexte d'un système de questions-réponses, on voit qu'il est nécessaire de tenir compte de ces types de variations, qui correspondent à la modification de la forme de surface du procès identifié dans la question. En ce qui nous concerne, si nous voulons utiliser la notion de focus pour rechercher la réponse précise attendue, il est essentiel de pouvoir la repérer quelle que soit sa forme. Nous citons des exemples de variation pour le procès **MARIER** :

- forme nominale : *mariage*
- forme adjectivale : *marié*
- synonyme : *épouser*
- conversif : -
- locution : *passer la bague au doigt*

Afin de nous rendre compte de la présence de variations du focus dans les phrases-réponses, nous avons mené une étude des différentes variations du focus au sein des phrases qui contiennent la réponse dans le but de mesurer l'importance de cette notion pour l'extraction de réponses précises.

3.1.3 Validation de la définition du focus

Nous avons mené une expérimentation à l'aide de REVISE, l'outil que nous avons présenté au chapitre précédent. L'objectif est de tester notre nouvelle définition du focus et son applicabilité, en comparaison avec la définition existante, c'est-à-dire de mesurer en contexte la présence du focus dans les phrases réponses ainsi que sa proximité avec la réponse précise. La première définition du focus portait sur un objet, mais non pas sur l'événement en entier. Il s'agissait le plus souvent de l'entité sur laquelle la question est posée (sujet ou objet de la question). Nous élargissons cette définition en incluant le verbe à ses compléments.

Nous rappelons que le terme focus sert essentiellement à extraire les réponses précises. Pour tester la pertinence de cette nouvelle définition, nous avons comparé l'ancienne définition du focus à la nouvelle, en terme de proximité avec la réponse. Les différentes étapes de notre méthodologie ont été les suivantes :

- sélection manuelle du focus des questions en fonction de la définition et des critères donnés,

- sélection des phrases réponses qui contiennent le focus et une réponse possible,
- calcul de la distance (en mots) entre focus et réponse (plus les deux éléments sont proches, plus les patrons de syntaxe locale seront fiables).

De cette façon, nous avons la possibilité de voir si une des définitions est plus pertinente que l'autre pour extraire la réponse précise.

Sélection du focus des questions

Dans un premier temps, nous avons donc étudié chaque question, et extrait le focus qui répondait à la définition donnée précédemment : soit une entité, soit un procès.

Nous citons des exemples de choix de focus :

- Question : *En quelle année est né Richard Nixon ?*
- Focus : *né*
- Réponse : *Né le* ***9 janvier 1913*** *à Yorba Linda (Californie), Richard Nixon avait été élu à la Maison Blanche en 1968, puis réélu en novembre 1972.*

Le focus est ici l'événement que représente la naissance de Richard Nixon. La question *En quelle année est mort Richard Nixon ?* montre l'intérêt de se centrer sur l'événement plutôt que l'entité liée.

- Question : *Où se situent les îles Marquises ?*
- Focus : *îles Marquises*
- Réponse : *Je suppose que les dames des îles Marquises, dans* ***l'océan Pacifique****, ont déterminé beaucoup de vocations ethnographiques.*

Ici, le verbe *situer* fait doublon avec le pronom interrogatif *où*, c'est donc l'entité qui sera étiquetée comme focus de la phrase.

Nous avons modifié manuellement les focus extraits par le système, données préalablement insérées dans la base à l'aide d'un formulaire PHP.

Sélection des phrases réponses

Nous avons sélectionné les phrases réponses contenant le focus grâce à REVISE. En effet, la sélection de données correspond à une requête SQL qui spécifie les critères que l'on veut observer (phrases qui contiennent le focus) ainsi que le corpus (phrases-réponses sélectionnées par le système qui contiennent la réponse).

Nous avons ensuite sélectionné les phrases réponses grâce à une requête pré-enregistrée, et calculé automatiquement la distance en mots entre le focus et la réponse (voir 3.4). La sélection du focus, qu'il s'agisse de la première ou de la deuxième, a été faite en fonction des mots de la question uniquement. C'est-à-dire qu'aucune variation syntaxique ni sémantique n'a été prise en compte : uniquement la forme canonique. Étant donné que notre nouvelle définition du focus repose essentiellement sur la notion d'événement, c'est le verbe de la question qui est sélectionné. Il est fréquent que l'événement exprimé par le verbe dans la question soit sous forme nominale dans la réponse. Les synonymes sont autant de termes que nous n'avons pas comptabilisés ici. Nous avons voulu étudier le focus sans variations d'abord.

Calcul de la distance entre focus et réponse

Notre base de données permet d'effectuer certains calculs de façon automatique : compter les occurrences d'un mot, le nombre de questions qui appartiennent à une catégorie spécifique, etc. Nous avons calculé la distance moyenne en mots entre un focus et une réponse pour chacune des définitions, à l'aide d'un script PHP qui utilise le focus défini ainsi que la ou les réponses possibles pour une question, et classé par catégorie de questions (figure 3.4).

27) En quelle année Richard Nixon est -il né ?
Focus : né | Type général : année | Réponses : 1913

Distances

Num	Texte	Id	Doc	Phrase	
27	En quelle année Richard Nixon est -il né ?	1	LEMONDE94-003533-19940429.0	La ville est plus peuplée que en 1913 , lorsque Richard Nixon est né dans une famille quaker , mais elle a conservé ses églises presbytérienne , méthodiste et baptiste .	4
27	En quelle année Richard Nixon est -il né ?	2	ATS.940423.0068.0	Richard Nixon est né dans une famille modeste le 9 janvier 1913 à Yorba Linda , en Californie .	7

Fig. 3.4 Distance entre focus et réponses précises

Notre base de données a également permis d'effectuer des calculs, comme la moyenne des distances, de façon à pouvoir faire notre évaluation.

Résultats

Cette étude a été réalisée sur 34 questions. Nous avons observé 588 phrases réponses qui contiennent le focus **à l'identique** dans les phrases réponses. Les résultats présentés dans le tableau ci-dessous ont été produits par des requêtes sur la base de données, après annotation des focus.

Aspect	Ancien focus	Nouveau focus
Phrases réponses contenant le focus	403	390
Phrases réponses sans le focus	185	198
Questions pour lesquelles une phrase au moins contient le focus	27	30
Distance moyenne entre focus et réponse	8 mots	4 mots

Tab. 3.1 Comparaison des deux versions du focus

Cette étude comparative des définitions du focus montre que l'ancienne définition est plus présente dans les phrases contenant la réponse, pour un focus identique à celui qui est extrait de la question. Néanmoins, étant donné que la nouvelle définition repose sur les verbes, et qu'il s'agit d'un composant très variable, nous présenterons un peu plus loin une étude sur les variations observées en corpus.

390 phrases réponses sur les données de Clef07 contiennent une réponse et le focus (selon la nouvelle définition) de la question à laquelle cette phrase répond. Nous nous sommes arrêtés à une distance maximale de quatre mots, une distance syntaxique supérieure rend inutilisable les patrons d'extraction. Nous présentons les distances obtenus enfonction des catégories de questions.

Distance (mots)	Nb phrases	Catégories concernées
0	129	Combien, Définition
1	48	Quel, Quand
2	39	Quel
3	22	Quel
4	24	Quel

Tab. 3.2 Distances les plus fréquentes

Les distances 0 et 1 comptabilisent un peu plus de la moitié des phrases réponses (177). Si l'on rajoute les distances 2, 3 et 4 on obtient un score de 75% des phrases réponses, ce qui est très encourageant pour l'extraction des réponses.

REVISE a permis ici de vérifier notre nouvelle définition du terme focus, grâce à l'observation et la modification des résultats de l'analyse des questions, la sélection des phrases réponses extraites de la base de données qui nous intéressaient ainsi que le calcul des distances de mots entre focus et réponses.

Cette étude a été faite en recherchant le focus à l'identique, mais il peut y avoir des variations. C'est pourquoi nous avons mené une étude de corpus uniquement sur les verbes, composants plus fortement soumis à variation.

3.2 Évaluation de l'impact de la variation linguistique

3.2.1 Étude des variations des focus de typé événement en corpus

Objectif

Les variations peuvent être de types différents. Nous séparons notamment celles qui ont trait à la sémantique (synonymie, périphrases, expressions figées) et donc à l'ambiguïté de la langue, de celles liées à la syntaxe (valence, voix, anaphore, propositions relatives, nominalisation, verbalisation). Dans le système FRASQUES, les variations de termes complexes sont reconnues par l'outil FASTR créé par Christian Jacquemin [Jacquemin, 1996]. Il s'agit d'un analyseur de surface pour la reconnaissance de variantes terminologiques [Ferret *et al.*, 2001b], uniquement sur les termes nominaux. Les variations liées aux verbes ne sont pas traitées, même si la verbalisation d'un nom est reconnue. Par ailleurs, les synonymes de termes simples sont aussi repérés à partir de listes de synonymes.

Nous cherchons à mesurer l'impact des variations des focus de type événement sur les phrases réponses. C'est le même type d'étude menée dans [Cohen *et al.*, 2008], où les auteurs cherchent à observer et annoter les variations des verbes les plus représentatifs du domaine bio-médical de façon à évaluer l'intérêt de leur traitement automatique. Seulement, en domaine ouvert, nous ne pouvons pas n'étudier qu'un sous-ensemble de verbes.

Pour réaliser cette étude, nous n'avons sélectionné à l'aide de notre base de données que des phrases réponses qui contiennent la réponse. Parmi la centaine de questions candidates issues de la campagnes d'évaluation Clef07[2], nous avons retenu un sous-ensemble de 34 questions satisfaisant les critères fixés : une question où le focus est exprimé sous la forme d'un procès et pour laquelle le système extrait des phrases réponses valides (où la réponse est justifiée), ce qui représente 587 phrases réponses au total. Cet échantillonnage devrait nous permettre d'avoir une idée des variations les plus fréquentes de l'expression du procès, et de fait d'étudier les stratégies à mettre en place pour les traiter automatiquement. Dans le cas contraire, il faudrait former un nouveau corpus de questions que l'on soumettrait à FRASQUES afin de compléter cette étude.

[2]http ://clef-qa.itc.it/

Méthodologie

Cette observation est effectuée à l'aide de REVISE. Cet outil permet d'afficher les phrases réponses obtenues pour chaque question, de colorer un terme particulier dans ces phrases (focus ou autre). Afin de faciliter l'observation des phénomènes, le terme focus s'il est présent est coloré dans la phrase réponse, de même que la réponse présente comme le montre la figure 3.5. Ainsi, lorsque le focus n'est pas reconnu, nous pouvons le chercher en priorité à proximité de la réponse.

100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	1	ATS.940930.0021.0	Le ferry **Estonia** avait fait **naufrage** avec plus de un millier de personnes à bord dans la nuit de mardi à mercredi en mer Baltique .	2
100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	2	ATS.950927.0163.0	Un an après le **naufrage** du ferry **Estonia** en mer Baltique , une longue bataille technique et juridique se est engagée sur les responsabilités , encore mal éclaircies , de le accident .	2
100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	3	ATS.940928.0020.0	Plus de 820 personnes ont trouvé la mort dans le **naufrage** du ferry **Estonia** , survenu dans la nuit de mardi à mercredi en mer Baltique .	2
100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	4	LEMONDE94-003580-19940930.0	Les opérations de secours entamées après le **naufrage** du ferry **Estonia** dans la mer Baltique ont repris jeudi matin 29 septembre .	2
100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	5	ATS.940929.0047.0	Les enquêteurs chargés de déterminer les causes du **naufrage** du ferry " **Estonia** " en mer Baltique sont devant un puzzle compliqué : peu ou pas de indices , témoignages divers et fragiles , parfois contradictoires , de survivants .	2
100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	6	LEMONDE94-003472-19940929.0	Plus de 700 personnes étaient portées disparues , mercredi 28 septembre en fin de matinée , après le **naufrage** du ferry " **Estonia** " , dans la nuit précédente en mer Baltique .	2
100	Quel célèbre ferry a fait **naufrage** en mer Baltique ?	7	ATS.950302.0027.0	le épave du ferry **Estonia** , qui gît au fond de la mer Baltique depuis son **naufrage** ayant fait 912 morts en septembre , sera recouverte de une coque de béton , a annoncé jeudi le ministère suédois des Transports .	10

Fig. 3.5 Coloration du focus et des réponses-précises

Les types de variation du focus ont été annotées dans la base de données.

3.2.2 Observations et résultats

Nous avons mis en place une typologie des variations, après observation du corpus :

- identique (présence du focus tel que dans la question)
- nominalisation (forme nominale du procès)
- synonymie
- locution (mot en plusieurs mots)
- préposition (expression de la notion de possession ou bien de durée par une préposition)

Voici des exemples des différents types de variations annotés :

– Nominalisation

Question : *Avec qui s'est* ***marié***[3] *Michael Jackson en mai 1994 ?*

Réponse : *Il précise que " le 26e jour du mois de mai 1994 , devant moi-même , Francisco Alvarez Perez , avocat et officier de l'état civil de la seconde circonscription de La Vega , République dominicaine , ont contracté le* ***mariage*** *civil Michael Joseph Jackson , 35 ans , citoyen américain , chanteur , et Lisa Marie Presley Keough , 26 ans , actrice , citoyenne américaine fille de Elvis Aaron Presley et de Priscila Presley"*

– Synonymie

Question : *Quel évêque fut* ***suspendu*** *par le Vatican le 13 janvier 1995 ?*

Réponse : *De son côté, le Vatican a expliqué vendredi avoir* ***démis*** *Mgr Gaillot en raison de son manque d'orthodoxie sur des sujets comme le sida ou les droits des travailleurs.*

– Locution

Question : *Quand l' Organisation Mondiale du Commerce est-elle* ***entrée en vigueur*** *?*

Réponse : *La Russie a déposé vendredi sa demande de adhésion à l'Organisation mondiale du commerce (OMC), qui* ***verra le jour*** *le 1er janvier 1995.*

– Préposition

Question : *Qui a* ***reçu*** *le prix Goncourt en 1995 ?*

Réponse : *Le Goncourt* ***à*** *Andreï Makine .*

Voici un tableau récapitulatif du pourcentage de variations présentes dans le corpus de travail :

Phénomènes	Nominalisation	Synonymie	Locution	Préposition	Identique
Présence	13%	14%	3%	2%	68%

Tab. 3.3 Taux de variations du focus de type procès

Pour 68% des phrases-réponses, le focus est présent à l'identique : il s'agit du même lemme. Pour 13% des phrases, ce focus se retrouve sous la forme d'un substantif. Nous avons pu observer que le focus tel que nous l'avons défini est présent dans toutes les phrases réponses contenant une réponse attendue. Cette observation tend à valider notre définition.

[3]Le focus et ses variations sont indiqués en gras.

Ici, un tableau récapitulatif des variations par questions :

Phénomènes	Nominalisation	Synonymie	Locution	Préposition	Identique
Nb questions	19	21	7	4	29

Tab. 3.4 Variations du focus de type procès par questions

L'organisation des variations par questions permet de voir le nombre de questions pour lesquelles il est indispensable de traiter les variations pour extraire la réponse en contexte.

Néanmoins, s'appuyer sur le focus pour sélectionner les passages candidats et extraire la réponse nécessite de pouvoir l'identifier de façon automatique. Or les variations possibles en compliquent la reconnaissance. Cette étude montre les performances que le système pourrait atteindre selon les choix de problèmes à résoudre.

3.2.3 Discussion

Après avoir montré l'importance de la prise en compte des variations pour identifier le focus événement au sein des phrases-réponses, nous discutons des ressources qu'il est possible d'utiliser et proposons une synthèse des points abordés.

En ce qui concerne les nominalisations du procès, des ressources lexicales pourraient permettre de les identifier. Des études sur la constitution de dictionnaires électroniques d'unités lexicales ont été menées dans ce sens au Laboratoire d'Automatique Documentaire et Linguistique (LADL), notamment le projet Unitex[4] ; on pourra également regarder les travaux de Nabil Hathout au CLLE-ERSS[5] dans ce sens. Identifier des synonymes du focus dans les phrases réponses est également réalisable à l'aide de ressources adaptées, comme le dictionnaire des synonymes élaboré au Centre de Recherche Inter-langues sur la Signification en COntexte (CRISCO[6]). Les locutions figées sont plus difficiles à repérer étant donné leur construction en plusieurs mots mais avec une seule unité de sens. Des lexiques existent pour aider à leur repérage, notamment [Dubois et Dubois-Charlier, 2004]. Néanmoins, les ressources, lors de leur utilisation en contexte, peuvent s'avérer insuffisantes pour reconnaître les variations notamment à cause des erreurs d'étiquetage des formes qui peuvent survenir.

Des ressources complètes indiquant les arguments du verbe et leur classification ne sont pas disponibles pour le français. Aussi faudra-t-il s'appuyer sur la syntaxe pour reconnaître

[4]Voir http ://www-igm.univ-mlv.fr/ unitex/
[5]http ://w3.erss.univ-tlse2.fr/
[6]http ://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

les arguments présents et cherchés. Il est néanmoins possible de définir des listes de verbes n'exprimant pas un procès. Par exemple, les verbes suivants : *avoir lieu*, *causer*, *citer*, *trouver*, *s'appeler*, etc.

Nous avons également examiné la présence des entités nommées de la question de type PERSONNE, dans les phrases réponses contenant le focus et la réponse, de façon à comparer la présence de chacun. En effet, l'on pourrait imaginer des stratégies d'extraction s'appuyant sur ces entités qui subissent peu de variations et sont donc facilement identifiables. Or les entités nommées, constituants pas ou peu soumis à variation, sont présents de façon moins systématique dans les phrases que le focus. En effet, le focus est présent dans toutes les phrases réponses contenant la réponse précise attendue, alors que les noms propres ne le sont pas toujours. L'explication de ce phénomène est simple : il est fréquent que les entités apparaissent dans une phrase antérieure, puis qu'elles soient remplacées par une anaphore, qu'elle soit lexicale ou grammaticale.

- *Quand* ***débuta*** *le procès de Paul Touvier ?*
- *Contrairement à la rumeur, le procès de l'ancien chef milicien a bien* ***commencé*** *le 17 mars.*

Cet exemple illustre le remplacement de l'entité Paul Touvier par la description d'une de ses fonctions.

Cette observation conforte notre choix de typer l'événement exprimé dans la question comme un élément focus présent dans la phrase réponse et et non de choisir un complément du verbe ou de s'appuyer sur les entités nommées de type PERSONNE. Cette étude nous permet aussi de voir qu'il est nécessaire de disposer de processus d'extraction s'appuyant moins sur la syntaxe : quand les patrons d'extraction de la réponse ne s'appliquent pas et que le focus est présent, on peut alors envisager de sélectionner comme réponse une entité du type cherchée syntaxiquement proche du terme focus. Cette stratégie est aussi évaluable à l'aide de notre outil.

Pour résumer, le focus est utilisé comme un terme pivot sur lequel appliquer des patrons d'extraction de la réponse précise qui correspondent à des règles simples de syntaxe locale. Or, cette stratégie n'est possible que si ce pivot est proche en distance de la réponse dans la phrase. Afin de vérifier à la fois la définition du focus que nous avons donnée et sa pertinence dans la stratégie d'extraction, nous avons calculé la distance en mots entre le focus et la réponse, ce qui nous a permis de fixer la fenêtre syntaxique moyenne comprenant à la fois le focus et la réponse à extraire. Dans le cas contraire, cela signifierait qu'il faut mettre en oeuvre une analyse syntaxique plus poussée des phrases candidates.

3.3 Étude des règles d'extraction de réponses précises

L'étape finale d'un système de questions-réponses consiste à extraire une réponse précise répondant à une question donnée. Or, en moyenne sur différents jeux de questions, l'évaluation comparée des phrases candidates qui contiennent la réponse attendue avec les réponses précises qui sont extraites de ces phrases montre une perte de 50% de réponses correctes sur les systèmes FRASQUES et QALC du LIMSI [El Ayari, 2009]. Il nous a paru intéressant d'évaluer les règles d'extraction appliquées ainsi que de mesurer l'impact de l'analyse de la question sur l'application de ces règles.

Nous présentons une étude menée sur les règles d'extraction de réponses précises appliquées sur les phrases candidates sélectionnées par le système de questions-réponses QALC sur des données anglaises. Nous détaillons le fonctionnement de ces règles, ainsi que la méthodologie d'évaluation que nous avons mise en place et discutons des limites de l'application de règles pour extraire la réponse attendue.

3.3.1 Principe d'utilisation des règles d'extraction

Il s'agit du dernier module du système de questions-réponses, où l'on applique des règles d'extraction sur les phrases candidates fondées sur le type d'entité attendu (personne, organisation, date, lieu, montant, etc.) ou bien le cas échéant sur les critères extraits de la question (type général, focus, verbe principal).

En ce qui concerne le premier cas de figure (questions qui attendent une entité nommée en réponse), le succès de la résolution repose sur le module de reconnaissance des entités nommées. Nous ne nous y intéresserons pas ici. En revanche, pour le deuxième cas de figure, ce sont bien les règles d'extraction qui vont permettre l'extraction de la ou des réponses précises attendues.

Les règles sont écrites sous la forme d'une grammaire au format CASS (*Cascaded Analysis of Syntactic Structure*) [Abney, 1996]. CASS est un analyseur syntaxique robuste développé par Steven Abney, qui repose sur les concepts de *chunk* et de cascade [Bourigaut, 2007]. Un *chunk* est créé au moyen de têtes sémantiques correspondant à des groupes syntaxiques, comme NP (nom propre), VP (verbe principal), PP (participe passé), etc. Ainsi, une phrase est constituée de différents *chunks* qui entretiennent des relations les uns avec les autres.

Une règle contient plusieurs niveaux, notamment la définition de groupes syntaxiques (ou *chunks*) et le marquage des réponses pour les règles développées au sein du système QALC [Ligozat, 2006]. Par exemple, pour la question *What does Knight Reader publish ?* qui est de structure syntaxique What-do-GN-VB, le système détermine le focus *Knight Reader* et l'un des patrons d'extraction appliqué sera GNfocus + Verbe + GNréponse [Ferret *et al.*, 2002]. Ainsi, la réponse *daily newspapers* issue de la phrase *Knight Reader publishes 30 daily newspapers, including the Miami Herald and the Philadelphia Inquirer* [...] sera extraite.

Nous allons définir la méthodologie d'évaluation appliquée au module d'extraction de réponses précises du système de questions-réponses QALC.

3.3.2 Méthodologie d'évaluation

Contexte de travail

L'étude que nous avons menée sur l'évaluation des patrons d'extraction est liée à la campagne d'évaluation QUAERO organisée sur la langue anglaise et à laquelle le système QALC[7] a participé. Cette étude consiste en l'évaluation des stratégies linguistiques mises en œuvre pour l'application de règles d'extraction de réponses précises.

Nous avions plusieurs objectifs : améliorer l'extraction de réponses pour les questions de catégorie COMBIEN et améliorer l'extraction de réponses pour les questions de catégorie QUE. Par ailleurs, la campagne QUAERO ayant introduit des questions dites complexes de type POURQUOI et COMMENT, nous avons utilisé notre outil pour concevoir des règles d'extraction qui leur sont adaptées.

Méthode appliquée

Nous rappelons la méthodologie que nous avons suivie, telle qu'elle a été définie auparavant :

- la sélection du corpus de travail (filtrage par catégories) ;
- l'observation de corpus (étiquetage des erreurs) ;
- l'affinement des règles (modifications au sein du système) ;
- l'évaluation des nouveaux résultats et réitéraction sur la modification des règles.

[7] QALC est un système de questions-réponses développé pour la langue anglaise au LIMSI.

L'étude menée comporte deux volets : l'amélioration des règles liées à certaines catégories et la création de règles pour les nouvelles catégories proposées par la campagne QUAERO. Pour ces deux études, il est nécessaire d'avoir accès au corpus, c'est-à-dire de pouvoir observer les phrases réponses qui contiennent la réponse ainsi que les réponses elles-mêmes au sein de ces phrases.

Nous avons tout d'abord sélectionné les questions qui nous intéressaient en fonction des règles à appliquer, c'est-à-dire que nous les avons triées par catégories. En effet, les règles varient en fonction des catégories de question, c'est-à-dire en fonction de l'objet que l'on cherche et des informations dont on dispose. Par exemple, les règles pour la catégorie POURQUOI que nous avons explicitées un peu plus haut ne s'appliquent pas aux questions de catégorie COMMENT, qui s'appuieront sur d'autres déclencheurs. Cette première étape a été effectuée à l'aide de REVISE, en filtrant les résultats en fonction du champ *Catégorie*.

Nous sommes tributaires des phrases réponses sélectionnées par le système de questions-réponses sur lequel nous travaillons. Afin de dépasser ce biais, nous avons ré-interrogé le corpus en ajoutant la réponse précise recherchée dans la requête, de façon à obtenir plus de phrases réponses contenant la réponse. De la sorte, nous avons augmenté les phrases pertinentes et constitué un corpus de phrases correctes pour chacune des catégories de questions.

L'intérêt de REVISE ici est de permettre une visualisation dédiée aux règles d'extraction : les phrases réponses qui contiennent la réponse sont sélectionnées grâce à la base de données, et affichées. De plus, les critères de la question (focus, type général, verbe principal et leurs variations), s'ils sont présents dans la phrase, sont affichés en couleur. Cette première étape nous a permis d'étudier les structures syntaxiques des phrases réponses sous la forme d'une représentation plate, et de voir les liens qu'entretiennent les critères de la question avec la réponse effective. Cette phase d'observation a été fort utile pour déterminer les nouvelles règles d'extraction des catégories POURQUOI et COMMENT.

Dans un deuxième temps, nous avons ajouté l'affichage en couleur également sur les phrases réponses des règles d'extraction appliquées par le système. Cette visualisation permet de mesurer les patrons d'extraction qui s'appliquent et ceux qui ne s'appliquent pas (alors qu'ils le devraient), dans l'optique de l'amélioration des règles de catégories déjà existantes.

Par ailleurs, nous avons modifié l'interface de REVISE de façon spécifique pour le système sur lequel nous travaillons afin de permettre un accès direct aux règles dans l'ar-

chitecture même de QALC [Gio, 2009]. L'outil permet d'afficher les règles d'extraction liées aux questions, puis de les modifier en fonction de l'observation de corpus. De la sorte, il est possible de relancer le système (uniquement au niveau de l'application des règles) pour que le système prenne en compte les modifications effectuées. L'interface permet également de nourrir la base de données avec les nouveaux résultats obtenus de façon automatique, et d'observer à nouveau le corpus de résultats fraîchement constitué.

Enfin, nous évaluons les nouveaux résultats obtenus, en terme de précision (nombre de réponses précises extraites), mais aussi en terme de précision effective de la réponse : ne pas sélectionner trop de mots ou pas assez. Par exemple, à la question *Which document provides information about classroom methods ?* la réponse *Comprehensive Positive* provenant de la phrase *For more specific information on classroom methods, see Comprehensive Positive Behavior Supports (CPBS) Pre-Service Training Program, a paper presented by Emma Martin and Tary Tobin in San Diego at the annual conference of the Teacher Education Division of the Council for Exceptional Children in 2006.* sera jugée incomplète (la réponse étant *Comprehensive Positive Behavior Supports*). De la même façon, la réponse *Comprehensive Positive Behavior Supports (CPBS) Pre-Service Training Program* sera trop longue.

Visualisation ciblée

Nous avons développé, au sein de notre outil, une visualisation adaptée qui permet d'observer précisément les phrases réponses obtenues par le système, avec une interrogation des documents à l'aide des réponses.

Il s'agit de produire un corpus adapté à la tâche : observer les phrases contenant la réponse et mesurer l'application des règles d'extraction de réponses précises. Il est intéressant de voir que la dimension linguistique passe également par un corpus adéquat sur lequel nous utilisons uniquement un composant de la chaîne de traitement. On peut rapprocher cette démarche d'une évaluation de sous-tâche, qui consisterait à évaluer des patrons d'extraction de réponses basées sur les éléments d'une question, mais qui dans ce cas précis serait réalisée en contexte (ce qui semble faire défaut aux évaluation de sous-tâches classiques).

La figure 3.6 montre un exemple de visualisation de l'application des règles pour la question *What do ballet dancers wear ?*[8]

[8] Quelle tenue portent les danseurs de ballet ?

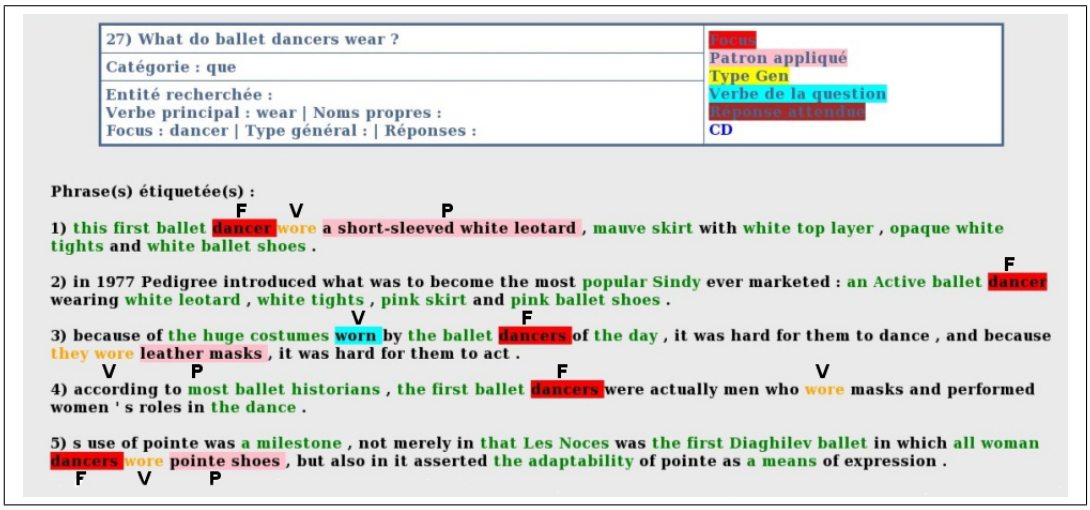


Fig. 3.6 Exemple de visualisation de l'application des règles d'extraction

En haut de la figure sont récapitulés les différents éléments extraits lors de l'analyse de la question (catégorie, verbe, focus, type général, noms propres). Ce sont ces éléments qui sont mis en relief par des jeux de couleurs lors de l'affichage des phrases réponses sélectionnées par le système. La légende en haut à droite montre les couleurs utilisées.

Ensuite viennent les phrases étiquetées, avec mise en couleur des éléments de la question ainsi que des réponses précises extraites par le système. La lettre **F** indique les focus, **V** les verbes et **P** les mots extraits par les règles appliquées. On voit rapidement les liens qui peuvent exister entre mots de la question et réponse précise, de même que les cas où le système repère la réponse (phrases 1, 3 et 5), des phrases où les règles ne s'appliquent pas. La phrase 2 montre que malgré la présence du focus, le système de reconnaît pas le réponse qui commence par un participe présent *wearing* qu'il n'identifie pas comme une forme du verbe principal. La figure 3.7 montre les règles d'extraction basées sur le verbe principal définies pour les questions de ce type.

```
Niveau 1 :
      # le GV du verbe principal - VP
      GVP ->   PP? RB* VP RB* | PP? (AUX|AUXPourPassif) RB* VPPP RB* ADJ? ;

      # le GV du verbe principal synonyme - VS
      GVPS ->   PP?  RB* VS RB*| PP? (AUX|AUXPourPassif) RB* VSPP RB* ADJ?  ;

Niveau 2 :

      VbRep -> (GVP|GVPS) a=(GN|GNavecGP|GNavecPoss) ;

      RepVb -> a=(GN|GNavecGP|GNavecPoss) (GVP|GVPS) ;
```

Fig. 3.7 Règles au format CASS autour du verbe

En ce qui concerne la phrase 4, nous allons visualiser les résultats étiquetés afin de voir

le problème qui se pose (figure 3.8).

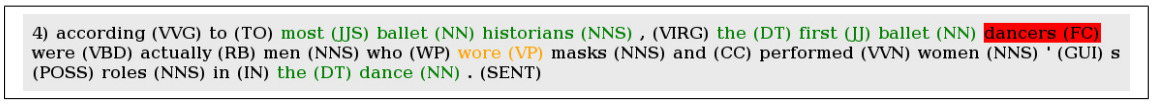


Fig. 3.8 Visualisation avec étiquettes morpho-syntaxiques

Tout d'abord sont définis les constituants à utiliser : un groupe verbal principal qui comprend le verbe extrait de la question (VP) et possiblement un pronom (PP) qui est optionnel ou bien un groupe verbal qui comprend un auxiliaire (AUX) suivi du verbe principal, pour gérer les cas de participes passés et de voix passive. La même règle est appliquée pour définir un groupe verbal synonyme : constitué à l'aide d'un synonyme du verbe. Le 2e niveau permet d'établier les règles d'extraction de réponses autour de ces groupes verbaux : on récupère comme réponse le groupe nominal qui suit un groupe verbal ou bien un groupe nominal qui précède le groupe verbal. La première des règles sera privilégiée par rapport à l'autre.

Le contexte situé à la gauche du verbe ne satisfait pas celui spécifié dans les règles (trop restrictif dans ce cas), ce qui fait la règle d'extraction ne se déclenche pas. Nous pourrions élargir la règle, mais il est risqué, en terme de bruit généré sur l'ensemble du corpus, d'autoriser un contexte plus large. Nous rappelons ici que le but recherché dans les campagnes de questions-réponses n'est pas d'extraire toutes les réponses possibles pour une question mais bien de proposer une réponse par question. Il s'agit de maximiser la précision du système.

3.3.3 Étude des règles d'extraction

Nous présentons les modifications effectuées sur les règles d'extraction pour les affiner et discuterons des limites rencontrées.

Modifications

Les règles sont définies en fonction de la catégorie de la question : suivant la catégorie, la structure de la règle d'extraction diffère. Nous nous intéressons particulièrement aux questions de type QUE, QUEL, COMMENT et POURQUOI, pour lesquelles nous donnons un exemple de question dans le tableau 3.5. Les questions de type QUI et QUAND diffèrent quant à leur résolution, qui dépend principalement de la reconnaissance d'entités nommées

dans les phrases réponses. Néanmoins, il pourra être intéressant de mesurer si notre étude des patrons peut affiner le choix des entités pour ces deux types de questions.

Catégorie	Exemple	Nb de questions
QUE	What did the first Little Pig use to build his house ?	40
QUEL	What kind of public transportation did Rosa Parks use ?	112
COMMENT	How is cereal made ?	37
POURQUOI	Why did afternoon tea originate ?	38

Tab. 3.5 Catégories de questions étudiées

Les règles ont été complétées pour prendre en compte le focus extrait de la question, mais également le verbe principal et les noms propres. Nous avons étendu de la sorte l'impact des mots de la question pour l'extraction de la réponse précise recherchée. AInsi, pour les quesions de type QUE et QUEL, deux règles avec comme termes pivots le focus et le verbe a été ajoutée : GNfocus + Verbe + GNréponse et GNréponse + Verbe + GNfocus. Ces patrons permettent l'extraction de réponse dans des phrases telles que :

- Catégorie QUEL : *Johannesburg South Africa's former deputy president,* ***Jacob Zuma****, was [acquitted] of rape on Wednesday, a verdict met with wild celebrations by supporters outside the Johannesburg High Court and dismay by women's rights activists.* pour la question *Which politician was acquitted of rape charges in 2006 ?*
- Catégorie QUE : *Klara Hitler [died]* ***from cancer*** *when Adolf was nineteen.* pour la question *What did Klara Hitler die of ?*

Des règles ont été créées pour les questions de type POURQUOI et COMMENT. Pour les règles de type POURQUOI 3.9

```
Niveau 1 :
     # le GN focus - FC
     GNFoc -> DT? RB? (ADJ (CC ADJ)?)? NPA* FC ;

     # le GN focus avec synonyme - FS
     GNFocS -> DT? RB? (ADJ (CC ADJ)?)? NPA* FS ;

Niveau 2 :

     FBecauseR -> (GNFoc|GNFocS) (BECAUSE) a=Tout+ ;

     FtoR -> (GNFoc|GNFocS) (TO|AS) b=Tout+ ;
```

Fig. 3.9 Règles d'extraction pour les questions de catégorie POURQUOI

Les règles FBecauseR et FtoR sont définies pour se déclencher lors de la présence du GN focus (ou d'un synonyme) suivi de la conjonction *because* (règle prioritaire de rang **a**,

to ou bien *as* (règle secondaire, rang **b**). Elles extrairont tout ce qui se trouve après l'un ou l'autre des termes déclencheurs que sont les conjonctions.

Nous illustrons l'application de cette règle avec une question, tirée de la campagne QUAERO 2009 : *Why did afternoon tea originate ?*[9] Nous rappelons que c'est l'ancienne définition du focus qui est appliquée ici.

- *The custom of [afternoon tea] is said to have originated with Anna, 7th Duchess,* ***to bridge the gap between luncheon and dinner.***
- *The tradition of [afternoon tea] began in England in the 1700's* ***as a working class effort to ward off hunger before the main dinner meal.***
- *Join us at afternoon tea* ***to take part in this historic, cultural tradition !***

Les éléments mis en gras sont des éléments de réponses plausibles à la question posée. On voit bien ici que ces phrases contiennent le *GN focus* de la question (*afternoon tea*), et que la réponse se situe effectivement après les conjonctions *to* et *as*. Le verbe principal est présent dans la première phrase sous une forme identique à celle de la question, sous la forme d'un synonyme dans la deuxième phrase (*began*). Par contre, la troisième phrase ne comporte pas la réponse correcte, ni le verbe principal.

Pour les règles liées aux questions de type COMMENT, un principe similaire aux questions de type POURQUOI a été mis en place : nous avons défini des marqueurs de réponse : BY et WITH, après observation de corpus, et nous avons défini les règles autour de ces marqueurs. Elles sont présentées sur la figure 3.10.

```
Niveau 1
    # le GN focus - FC
    GNFoc -> DT? RB? (ADJ (CC ADJ)?)? NPA* FC ;

    # le GN focus avec synonyme - FS
    GNFocS -> DT? RB? (ADJ (CC ADJ)?)? NPA* FS ;

     # le GV du verbe principal - VP
    GVP ->   PP? RB* VP RB* | PP? (AUX|AUXPourPassif) RB* VPPP RB* ADJ? ;

    # le GV du verbe principal synonyme - VS
    GVPS ->   PP?  RB* VS RB*| PP? (AUX|AUXPourPassif) RB* VSPP RB* ADJ?  ;

Niveau 2 :
    FVerbeR -> (GNFoc|GNFocS) (GVP|GVPS) a=Tout+ ;

    VerbeR -> (GVP|GVPS) b=Tout+ ;
```

Fig. 3.10 Règles d'extraction pour les questions de catégorie COMMENT

Sont définis des groupes nominaux qui contiennent le focus (GNFoc) ou bien son synonyme (GNFocS) ainsi que les groupes verbaux contenant le verbe principal (GVP) ou un

[9] Traduction : *Pourquoi le thé de quatre heures est-t-il né ?*

synonyme (GVPS). Ainsi, la règle qui s'appliquera pour les questions de catégorie COMMENT sera de récupérer tout ce qui suit la séquence focus + verbe (règle plus fortement pondérée) ou bien, en deuxième choix, ce qui le verbe principal.

De la sorte, nous extrayons correctement la réponse *with tidbit savouries, and lotsa pastry things*[10] de la phrase *Afternoon Tea is served around 4ish, with tidbit savouries, and lotsa pastry things ; scones and clotted cream included.* à la question *How is afternoon tea served ?*

Étude de la non-application des règles

Certaines questions occasionnent des erreurs en chaîne. C'est ce que nous allons montrer pour la question *What saint slew a dragon ?* Si l'on regarde l'analyse de la question effectuée, on s'aperçoit que la question a pour catégorie QUE, ce qui est étonnant : la catégorie est attribuée en fonction de l'analyse syntaxique de la question et dans ce cas, il devrait s'agir d'une question QUEL (on recherche un type de saint). Si l'on regarde de plus près l'analyse syntaxique effectuée, *saint* est un verbe et *slew* un nom. On peut alors remonter à l'étiquetage de la question qui est erroné : le verbe et le nom, qui sont ambigus en anglais ont été mal étiquetés. Or, la catégorie conditionne la règle d'extraction qui est appliquée et le système, alors qu'il sélectionne une phrase réponse correcte, n'en extrait pas la réponse précise.

De façon plus précise, nous avons répertorié les différents types d'erreurs liées à l'analyse des questions sur les données de la campagne d'évaluation QUAERO 2009, sur 336 questions en anglais.

Erreurs	**Nb de questions**
Verbe principal non reconnu	20
NP non reconnus	8
Focus erroné	95
Type général erroné	38
Analyse syntaxique erronée	46
EN attendue erronée	4

Tab. 3.6 Erreurs lors de l'analyse des questions

Comme le montre le tableau 3.6, concernant la reconnaissance des verbes par le système, il y a 141 questions formulées avec une copule (*be* ou *have*). Nous disposons de 195 questions

[10] Cette évaluation a été menée sur un corpus issu du Web, ce qui explique la formulation *lotsa*.

formulées avec un verbe, et sur ces 195 questions, 46 analyses syntaxiques sont erronées. De plus, il peut y avoir plusieurs erreurs par questions, notamment en ce qui concerne l'analyse syntaxique qui génère le plus souvent des problèmes au niveau de l'extraction des critères. Nous expliquons plus en détail les causes de ces erreurs :

- Verbe principal non reconnu : cela est dû notamment à une erreur d'étiquetage morpho-syntaxique. En effet, le verbe n'est pas reconnu comme tel.
- Nom propre non reconnu : il s'agit essentiellement d'un problème lié aux noms propres composés qui ne sont pas reconnus par le système. C'est surtout un problème de couverture des noms propres existants.
- Focus erroné : la mauvaise reconnaissance du focus est liée à des problèmes d'analyse syntaxique. En effet, sa reconnaissance est basée sur les composants de la phrase.
- Type général erroné : il s'agit du même phénomène que le focus.
- Analyse syntaxique erronée : elle est due à des problèmes d'étiquetage, notamment du verbe principal lequel, s'il n'est pas reconnu comme tel, influe sur le reste.
- Entité nommée attendue erronée : c'est un cas plutôt rare, dû à une mauvaise reconnaissance du type général.

Les problèmes rencontrés lors de l'extraction de réponses précises sont causés essentiellement par des erreurs d'étiquetage qui entraînent des erreurs d'analyse syntaxique puis de non reconnaissance de critères. Les problèmes liés à l'analyse syntaxique dépendent de l'étiquetage des questions, de la reconnaissance des verbes. Il est fréquent que l'auxiliaire *do* soit reconnu comme verbe principal de la phrase, et que le verbe véritable soit ignoré comme, par exemple, pour la question *How many pieces of jewelry does Queen Elizabeth own ?* On peut aussi en conclure que sans un effort pour améliorer cet étiquetage, l'amélioration des processus qui en dépendent ne permettra pas d'augmenter les performances globales du système de manière si significative.

D'autres erreurs sont récurrentes, notamment sur des phrases réponses avec des structures syntaxiques plus complexes. Pour la question *Why did the Swiss authorities deny political asylum during the Second World War ?* le système sélectionne la phrase *the Swiss authorities denied him a political asylum , because he was one of 32 persons whose name appeared on the country's persona non grata list.* qui contient effectivement la réponse attendue mais la structure de la phrase rend l'extraction difficile à effectuer avec des règles autour du focus ou du verbe.

Un autre problème demeure : celui des formulations trop éloignées de la formulation de départ (la question). A la question *What major political event allowed Hitler to become*

an important character in Germany ?, la phrase *The political turning point for Hitler came when the Great Depression hit Germany in 1930.* contient bien la réponse attendue, mais il est impossible de l'extraire à l'aide de règles, lesquelles sont trop dépendantes de la forme de la question.

Enfin, nous rappelons que cette étude a été menée sur un corpus issu du Web. il est fréquent d'utiliser des corpus journalistiques comme *Le Monde* pour rechercher de l'information, notamment parce qu'ils sont garants d'une certaine normalisation et structuration des propos qui y sont formulés. Ce n'est absolument pas le cas sur Internet, qui regorge de blogs, forum et autre sites à tendance pornographique. Ces phénomènes favorisent une certaine « pollution » des données, qui sont plus difficiles à traiter que les évaluations classiques comme CLEF ou TREC et génèrent un certain bruit et de ce fait une baisse globale des résultats obtenus par les participants pour la tâche QUAERO organisée en 2008.

Conclusion et discussion des résultats

Les erreurs répertoriées qui demeurent présentent un enjeu pour les systèmes de questions-réponses, enjeu qui présente de réelles difficultés. En effet, si l'on revient sur la notion de performance pour des systèmes modulaires, on se rend compte que celle-ci ne peut être de 100%, mais aussi que plus on enchaîne de traitements séquentiels, plus il y a de risques d'erreurs. Nous avons montré l'importance de l'étiquetage morpho-syntaxique qui détermine l'analyse de la question mais aussi celle des phrases-réponses. Un mauvais étiquetage rend les chances d'extraire une réponse précise relativement ténue. De la même façon, l'ambiguïté de la langue crée de la difficulté et des difficultés supplémentaires pour apparier questions et réponses de façon automatique.

Néanmoins, si ces erreurs présentent une réelle difficulté, notre étude montre que des enjeux linguistiques sont encore surmontables, et qu'une observation fine et méthodique des résultats permet d'endiguer certains blocages et d'obtenir un savoir précis des améliorations possibles.

De plus, au-delà de l'amélioration du traitement de certains phénomènes, ces études fines devraient permettre de mettre en œuvre des stratégies choisies en fonction des caractéristiques reconnues, ou non, dans les questions et dans les phrases réponses au lieu d'appliquer la même stratégie sans tenir compte des capacités du système.

Enfin, la visualisation de phénomènes linguistiques, comme l'application de patrons d'extraction basés sur des critères issus de la question, est un outil précieux pour réellement

observer les relations syntaxiques entre composants de la phrase réponse, ainsi que pour s'assurer de l'application effective de règles d'extraction. La mise en relief des critères au sein des phrases réponses permet d'en mesurer l'application en contexte.

Après nous être intéressée aux aspects quantitatifs du focus ainsi qu'à la vérification de la stratégie employée pour sa reconnaissance, nous avons détaillé une étude menée sur des règles d'extraction de réponses précises en contexte, laquelle aurait été difficile à mener sans notre outil d'observation « intelligente » des critères.

Nous avons développé une évaluation transparente des résultats de systèmes de questions-réponses et montré comment REVISE a permis de faciliter le diagnostic des systèmes FRASQUES et QALC sur les problèmes récurrents rencontrés par le système (étiquetage, analyse syntaxique, extraction de critères, règles d'extraction) avec une approche « trans-composants ».

En effet, si l'accès aux données permet d'observer les difficultés ainsi que des les mesurer, il est également possible de tracer l'impact d'un critère tout au long de la chaîne de traitement, comme nous l'avons vu avec le focus. Or précisément, c'est cet aspect de traçabilité des phénomènes à différents niveaux qui manquent aux analyses d'erreurs classiques que nous avons présentées (2.1).

Conclusion

Bilan

Dans la première partie de ce travail, nous avons présenté l'évaluation telle qu'elle est pratiquée sur les systèmes de recherche ou d'extraction d'information en opposant une démarche globale (*boîte noire*) à une démarche plus fine (*boîte transparente*). Si les évaluation de composants sont plus fines que les évaluations globales, elles ne permettent pas de diagnostiquer précisément les phénomènes linguistiques qui posent problème à un système donné. Nous avons montré les limites de l'évaluation de type *boîte transparente* telle qu'elle est menée sur les systèmes de questions-réponses pour une évaluation de diagnostic. Nous défendons la thèse selon laquelle une évaluation de diagnostic doit comporter à la fois une évaluation de performance couplée à une analyse du corpus que forment les résultats intermédiaires produits par les systèmes.

Dans la deuxième partie, nous avons mis au point une méthodologie d'évaluation des systèmes de questions-réponses, en partant du principe qu'un système dispose d'une analyse des questions de laquelle on extraie des critères et d'une analyse des phrases-réponses sélectionnées par le système. Ainsi, il nous a semblé intéressant de discuter d'une démarche générique d'évaluation pour n'importe quel système de questions-réponses qui possède des résultats intermédiaires.

Pour ce faire, nous avons développé un outil, REVISE *Recherche, Extraction, VISualisation et Evaluation*, qui permet d'explorer les résultats intermédiaires produits par un système de questions-réponses de façon à réaliser une évaluation transparente des phénomènes linguistiques rencontrés par l'étude du corpus de résultats, la visualisation et l'étiquetage des phénomènes non traités, la modification des données ainsi que la possibilité de ré-engendrer un fichier avec les nouvelles valeurs pour relancer la chaîne de traitement et mesurer les résultats ainsi obtenus. De plus, cet outil permet, grâce à l'annotation des données, de créer des sous-corpus de phénomènes linguistiques à traiter.

Enfin, dans la troisième partie, nous avons mené différentes études sur les systèmes FRASQUES et QALC, notamment sur le critère focus ainsi que sur les règles d'extraction de réponses précises, qui mettent en application une évaluation transparente à l'aide de REVISE. Ces études montrent la faisabilité et l'intérêt de la constitution de sous-corpus d'études représentatifs de certaines difficultés rencontrées ainsi que de la nécessité de l'observation fine des résultats pour réaliser une évaluation de diagnostic efficace des systèmes.

Perspectives

La création de jeux de questions de difficulté calculée en fonction de la difficulté de résolution d'un question serait un plus pour la création de sous-tâches dédiées à certains types de problèmes fréquents en questions-réponses. Nous avons soulevé la limite des campagnes d'évaluation par rapport à l'aspect aléatoire de la difficulté des questions proposées. L'étiquetage des phénomènes qui posent problème au sein des phrases réponses permettrait la constitution de corpus « intelligents » pour faire progresser la résolution de certains types de problèmes.

Il serait également intéressant de pouvoir réaliser des études comparées de différents outils, par exemple de changer l'étiqueteur d'un système de façon à les évaluer. Ces études pourraient permettre de se diriger vers des schémas de résolution de questions : on aurait différentes façons de traiter les questions en fonction de certains critères : telle résolution pour tel phénomène linguistique présent.

De plus, la mise en commun des méthodologies d'évaluation réalisées sur différents systèmes de questions-réponses (telle combinaison de critères) permettrait d'avancer sur les techniques d'évaluation et de dresser une typologie d'études à mener pour tel ou tel problème.

Enfin, REVISE pourrait être utilisé pour construire des corpus dédiés, en prenant les résultats d'un système de questions-réponses comme une proposition d'annotation, que l'on peut modifier et compléter. L'utilisateur pourrait décider des caractéristiques et des propriétés que doit avoir le corpus à construire, caractéristiques qui donneraient lieu à une sélection de questions pour lesquelles des phénomènes auraient été annotés au sein des phrases réponses.

Il serait alors possible d'envisager de choisir des questions en fonction de leurs difficultés de résolution, lesquelles seraient préalablement annotées, et donc disponibles pour

sélectionner le corpus de questions.

Bibliographie

[Abney, 1996]Abney, S. (1996). Partial Parsing via Finite-State Cascades. *In Natural Language Engineering*, pages 337–344. Cambridge University Press.

[Abney *et al.*, 2000]Abney, S. P., Collins, M. et Singhal, A. (2000). Answer extraction. *In Applied Natural Language Processing Conference (ANLP)*, pages 296–301.

[Adda *et al.*, 1998]Adda, G., Lecomte, J., Mariani, J., Paroubek, P. et Rajman, M. (1998). The GRACE French Part-of-Speech Tagging Evaluation Task. *In First International Conference on Language Resources and Evaluation*.

[Ayache *et al.*, 2005]Ayache, C., Grau, B. et Vilnat, A. (2005). Campagne d'évaluation EQueR-EVALDA, évaluation en question-réponse. *TALN 2005*, pages 63–72.

[Bourigaut, 2007]Bourigaut, D. (2007). Un analyseur syntaxique opérationnel : Syntex. Habilitation à diriger des recherches. ERSS, Université de Toulouse le Mirail.

[Brill *et al.*, 2002]Brill, E., Dumais, S. et Banko, M. (2002). An analysis of the askmsr question-answering system. *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 257–264. Association for Computational Linguistics (ACL).

[Chaudiron, 2004a]Chaudiron, S. (2004a). Introduction. *In* Chaudiron, S., éditeur : *Évaluation des systèmes de traitement de l'information*, pages 17–24. Hermès.

[Chaudiron, 2004b]Chaudiron, S. (2004b). La place de l'usager dans l'évaluation des systèmes de recherche d'informations. *In* Chaudiron, S., éditeur : *Évaluation des systèmes de traitement de l'information*, chapitre 12, pages 287–310. Hermès.

[Cohen *et al.*, 2008]Cohen, K. B., Palmer, M. et Hunter, L. (2008). Nominalization and Alternations in Biomedical Language. *PLoS ONE*.

[Cohen *et al.*, 2004]Cohen, K. B., Tanabe, L., Kinoshita, S., et Hunter, L. (2004). A resource for constructing customized test suites for molecular biology entity identification systems. *In Association for Computational Linguistics (ACL)*, pages 1–8.

[Costa et Sarmento, 2006]Costa, L. F. et Sarmento, L. (2006). Component Evaluation in a Question Answering System. *LREC*.

[de Chalendar *et al.*, 2002]de Chalendar, G., Dalmas, T., Elkateb-Gara, F., Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I. et Vilnat, A. (2002). The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet. *Text retrieval conference (TREC)*.

[Devillers *et al.*, 2003]Devillers, L., Maynard, H., Paroubek, P. et Rosset, S. (2003). The PEACE SLDS understanding evaluation paradigm of the French MEDIA campaign. *European Chapter of the Association for Computational Linguistics*.

[Dubois et Dubois-Charlier, 2004]Dubois, J. et Dubois-Charlier, F. (2004). *Locutions en français*. Larousse.

[El Ayari, 2007]El Ayari, S. (2007). Évaluation transparente de systèmes de questions-réponses : application au focus. *Actes de ReciTAL*.

[El Ayari, 2009]El Ayari, S. (2009). A framework of evaluation for question-answering systems. *European Conference on Information Retrieval (ECIR)*, pages 744–748.

[El Ayari *et al.*, 2009]El Ayari, S., Grau, B. et Ligozat, A.-L. (2009). REVISE, un outil d'évaluation précise des systèmes de questions-réponses. *Actes de Coria*.

[Ferret *et al.*, 2001a]Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G. et Jacquemin, C. (2001a). Document selection refinement based on linguistic features for QALC, a Question Answering system. *RANLP*.

[Ferret *et al.*, 2001b]Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G. et Jacquemin, C. (2001b). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *Actes de TALN*.

[Ferret *et al.*, 2002]Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I. et Vilnat, A. (2002). Recherche de la réponse fondée sur la reconnaissance du focus de la question. *Actes de TALN*.

[Galibert, 2009]Galibert, O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. Thèse de doctorat, Université de Paris-Sud 11.

[Gillard *et al.*, 2006]Gillard, L., Bellot, P. et El-Beze, M. (2006). Question answering evaluation survey. *Language Resources and Evaluation Conference*.

[Gio, 2009]Gio, A. (2009). Évaluation transparente de systèmes de questions-réponses - aspect générique. Rapport de stage, École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE).

[Grau, 2004a]Grau, B. (2004a). Les systèmes de question-réponse. *In* Ihadjadène, M., éditeur : *Méthodes avancées pour les systèmes de recherche d'informations*, chapitre 10, pages 189–218. Hermès.

[Grau, 2004b]Grau, B. (2004b). Évaluation des systèmes de question-réponse. *In* Chaudiron, S., éditeur : *Évaluation des systèmes de traitement de l'information*, chapitre 3, pages 77–98. Hermès.

[Grau *et al.*, 2006]Grau, B., Ligozat, A.-L., Robba, I., Vilnat, A. et Monceaux, L. (2006). FRASQUES : A Question-Answering System in the EQueR Evaluation Campaign. *Language Resources and Evaluation Conference*.

[Gross, 1981]Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *In Languages*, volume 63, pages 7–52.

[Habert, 2005]Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Ophrys, L'essentiel français.

[Habert et Zweigenbaum, 2002]Habert, B. et Zweigenbaum, P. (2002). Régler les règles. *In Traitement automatique des langues*, volume 43, pages 83–105.

[Harris, 1976]Harris, Z. (1976). *Notes du cours de syntaxe*. Le Seuil.

[Hirschman et Thompson, 1997]Hirschman, L. et Thompson, H. S. (1997). Overview of evaluation in speech and natural language processing.

[Hurault-Plantet et Monceaux, 2002]Hurault-Plantet, M. et Monceaux, L. (2002). Cooperation between black box and glass box approaches for the evaluation of a question answering system. *LREC*.

[Jacquemin, 1996]Jacquemin, C. (1996). A symbolic and surgical acquisition of terms through variation. *In* Wermter, S., Riloff, E. et Scheler, G., éditeurs : *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438. Springer.

[Jijkoun *et al.*, 2003]Jijkoun, V., Mishne, G. et de Rijke, M. (2003). How frogs built the berlin wall : A detailed error analysis of a question answering system for dutch. *In CLEF*, pages 523–534.

[Kursten *et al.*, 2008]Kursten, J., Wilhelm, T. et Eibl, M. (2008). Extensible retrieval and evaluation framework : Xtrieval. *In Proceedings of Lernen - Wissen - Adaption (LWA-2008)*.

[Lapalme et Lavenus, 2002]Lapalme, G. et Lavenus, K. (2002). Évaluation des systèmes de question-réponse. Aspects méthodologiques. *Traitement automatique des langues*, 43: 181–208.

[Laurent *et al.*, 2006]Laurent, D., Nègre, S. et Séguéla, P. (2006). QRISTAL, le QR à l'épreuve du public. *In Actes de Traitement automatique du langage naturel (TALN)*.

[Lehnert, 1978]Lehnert, W. (1978). *The Process of Question Answering : A Computer Simulation of Cognition*. John Wiley & Sons Inc.

[Ligozat, 2006]Ligozat, A.-L. (2006). *Exploitation et fusion de connaissances locales pour la recherche d'informations précises*. Thèse de doctorat, Université de Paris-Sud 11.

[Ligozat *et al.*, 2006]Ligozat, A.-L., Grau, B., Robba, I. et Vilnat, A. (2006). L'extraction des réponses dans un système de question-réponse. *TALN 2006*.

[Mel'cuk *et al.*, 1995]Mel'cuk, I., Clas, A. et Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot.

[Mellet, 2003]Mellet, S. (2003). Corpus et recherches linguistiques. http ://corpus.revues.org/index7.html.

[Mendes et Moriceau, 2004]Mendes, S. et Moriceau, V. (2004). L'analyse des questions : intérêt pour la génération des réponses. *Workshop Question-Réponse*.

[Moldovan *et al.*, 2003]Moldovan, D., Paca, M., Harabagiu, S. et Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *In Actes de ACM Transactions on Informations Systems*.

[Moriceau et Tannier, 2009]Moriceau, V. et Tannier, X. (2009). Apport de la syntaxe dans un système de question-réponse : étude du système FIDJI. *TALN*.

[Nyberg et Mitamura, 2002]Nyberg, E. et Mitamura, T. (2002). Evaluating QA Systems on Multiple Dimensions. *In Proceedings of the Workshop on QA Strategy and Resources*.

[Nyberg *et al.*, 2003]Nyberg, E., Mitamura, T., Callan, J., Carbonell, J., Frederking, R., Collins-Thompson, K., Hiyakumoto, L., Huang, Y., Huttenhower, C., Judy, S., Ko, J., Kupse, A., Lita, L., Pedro, V., Svoboda, D. et Van Durme, B. (2003). The JAVELIN Question-Answering System at TREC 2003 : A Multi-Strategy Approach with Dynamic Planning. *In Proceedings of the Text Retrieval Conference*.

[Nyberg *et al.*, 2002]Nyberg, E., Mitamura, T., Carbonell, J., Callan, J. et Collins-Thopson, K. (2002). The JAVELIN Question-Answering System at TREC 2002. *In Proceedings of the Text Retrieval Conference*.

[Ozdowska, 2007]Ozdowska, S. (2007). Trois expériences d'évaluation dans le cadre du développement d'un système d'alignement sous-phrastique. *In T.A.L. : Traitement automatique de la langue*, volume 48.

[Plamondon *et al.*, 2002]Plamondon, L., Kosseim, L. et Lapalme, G. (2002). The quantum question answering system at trec-11. *In Proceedings of the Eleventh Text Retrieval Conference (TREC-2002)*, pages 750–757.

[Poibeau, 2005]Poibeau, T. (2005). Sur le statut référentiel des entités nommées. *Actes de TALN*.

[Polguère, 2008]Polguère, A. (2008). *Lexicologie et sémantique lexicale*. Presses de l'Université de Montréal.

[Popescu-Belis, 2007]Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *In T.A.L. : Traitement automatique de la langue*, volume 48, pages 67–91.

[Péry-Woodley, 1995]Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques ? *In T.A.L. : Traitement automatique de la langue*, volume 36, pages 213–232.

[Quintard, 2008]Quintard, L. (2008). Plan d'évaluation de référence des systèmes de question-réponse. Rapport interne QUAERO, QPR.

[Ravichandran et Hovy, 2002]Ravichandran, D. et Hovy, E. (2002). Learning Surface Text Patterns for a Question Answering System. *ACL conference*.

[Rosset *et al.*, 2006]Rosset, S., Galibert, O., Illouz, G. et Max, A. (2006). Interaction et recherche d'information : le projet Ritel. *Traitement automatique des langues*, 46:155–179.

[Saracevic, 1995]Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–146.

[Shima *et al.*, 2006]Shima, H., Wang, M., Lin, F. et Mitamura, T. (2006). Modular approach to error analysis and evaluation for multilingual question answering. *LREC*.

[Sparck Jones, 2001]Sparck Jones, K. (2001). Automatic language and information processing : rethinking evaluation. *In Natural Language Engineering*, pages 1–18.

[Sparck Jones et Galliers, 1996]Sparck Jones, K. et Galliers, J. R. (1996). *Evaluating Natural Language Processing Systems : An Analysis and Review*. Springer-Verlag.

[Timim, 2006]Timim, I. (2006). Évaluation des systèmes d'acquisition de terminologie : nouvelles pratiques, nouvelles métriques. *In Actes des 8e journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*.

[Tomas *et al.*, 2005]Tomas, D., Vicedo, J. L., Saiz, M. et Izquierdo, R. (2005). Building an xml framework for question answering. *Cross Language Evaluation Forum*.

[Vilnat *et al.*, 2004]Vilnat, A., Monceaux, L., Paroubek, P., Robba, I., Gendner, V., Illouz, G. et Jardino, M. (2004). Annoter en constituants pour évaluer des analyseurs syntaxiques. *TALN 2004*.

[Voorhees, 2002]Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. *In CLEF '01 : Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK. Springer-Verlag.

[Voorhees et Harman, 2005]Voorhees, E. M. et Harman, D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. MIT Press.