Évaluation d'un système d'extraction de réponses multiples sur le Web par comparaison à des humains

Mathieu-Henri Falco Véronique Moriceau Anne Vilnat
LIMSI-CNRS, Université Paris-Sud, 91405 Orsay, France
prenom.nom@limsi.fr

Résumé. Dans cet article, nous proposons une évaluation dans un cadre utilisateur de Citron, un système de question-réponse en français capable d'extraire des réponses à des questions à réponses multiples (questions possédant plusieurs réponses correctes différentes) en domaine ouvert à partir de documents provenant du Web. Nous présentons ici le protocole expérimental et les résultats pour nos deux expériences utilisateurs qui visent à (1) comparer les performances de Citron par rapport à celles d'un être humain pour la tâche d'extraction de réponses multiples et (2) connaître la satisfaction d'un utilisateur devant différents formats de présentation de réponses.

Abstract. In this paper, we propose a user evaluation of Citron, a question-answering system in French which extracts answers for multiple answer questions (expecting different correct answers) in open domain from Web documents. We present here our experimental protocol and results for user evaluations which aim at (1) comparing multiple answer extraction performances of Citron and users, and (2) knowing user preferences about multiple answer presentation.

Mots-clés : système de question-réponse, réponses multiples, évaluation utilisateur.

Keywords: question-answering system, multiple answers, user evaluation.

1 Introduction

Nous nous intéressons à l'évaluation de Citron, un système de question-réponse (SQR) en français capable de répondre à des questions à réponses multiples (*question-ARM*) en domaine ouvert à partir de documents provenant du Web. Les questions-ARM sont des questions possédant plusieurs réponses correctes différentes (par exemple *Quand le PSG a-t-il remporté la Coupe de France de football ?*), les questions de type liste (*question-liste*) en sont un exemple (*Quelles sont les planètes du système solaire ?*). Nous avons constaté en étudiant les données des campagnes d'évaluations des SQR pour le français comportant des questions-listes (Ayache, 2005), (Quintard *et al.*, 2010) que toutes les réponses à une question-liste se trouvaient quasiment toujours à l'intérieur d'un même document et étaient même concentrées majoritairement dans une seule phrase (Falco, 2012). Ce constat va de pair avec l'évaluation faite par ces campagnes qui impose un seul support justificatif par réponse, limité en caractères et continu, rendant alors impossible des recoupements multi-documents pour justifier une réponse. Ces contraintes sont d'autant plus pénalisantes lorsque les documents utilisés pour extraire les réponses sont issus du Web. En effet, ces documents sont des sources non négligeables de réponses aux questions-ARM grâce aux nombreuses structures (énumérations, tableaux) qu'ils contiennent. Pourtant les formats de réponse imposés par les campagnes d'évaluation ne permettent pas d'exploiter ces structures entièrement.

Citron exploite les structures (énumérations, tableaux) propres à ce type de documents pour mieux extraire les réponses et éventuellement les agréger pour faire apparaître des critères variants comme la date ou le lieu (Falco, 2012). Nous avons déjà évalué notre système selon les critères des campagnes d'évaluation (précision, rappel) et les résultats obtenus sont encourageants (Falco *et al.*, 2013). Pour la tâche d'extraction de réponses à des questions-ARM, nous souhaitons comparer les performances de Citron à celles d'un humain. L'hypothèse intuitive est qu'un système automatique est plus rapide mais peut-être un peu moins bon qu'un humain pour extraire les réponses. Nous avons surtout souhaité évaluer un aspect original de Citron qui n'est pas pris en compte durant les campagnes d'évaluation à savoir la présentation des réponses extraites. En effet, lorsque plusieurs réponses sont correctes, nous pensons qu'apporter le critère variant expliquant pourquoi il existe plusieurs réponses correctes est important pour la compréhension d'un utilisateur. Par une deuxième expérience, nous avons donc souhaité connaître la satisfaction des utilisateurs concernant le format de présentation des ré-

ponses imposé par les campagnes et celui de Citron. Dans cet article, nous présentons donc rapidement Citron et détaillons les protocoles de nos deux expériences en cadre utilisateur. Enfin, nous discutons leurs résultats.

2 Le système de question-réponse Citron

Citron est un SQR spécialisé dans l'extraction de réponses à des questions-ARM en français. Il peut travailler sur des collections de documents de différents types ; nous nous intéressons ici à une collection de documents issus du Web prétraités par le programme Kitten (Falco *et al.*, 2012) qui permet notamment de repérer les éléments structuraux susceptibles de contenir des réponses multiples (structures énumératives (SE) et tableaux) des documents Web et de les formater de manière exploitable pour une analyse syntaxique. Citron utilise l'analyseur XIP (Aït-Mokhtar *et al.*, 2002) qui permet d'obtenir une analyse en dépendances et intègre également un détecteur d'entités nommées. L'analyse de la question avec XIP vise à extraire les termes importants de la question et permet de générer une requête soumise à Lucene (Hatcher *et al.*, 2010) pour trouver des snippets susceptibles de contenir la réponse. Cette recherche se fait dans 3 index différents : l'index des parties "texte" des documents, l'index des SE et celui des tableaux. XIP analyse ensuite les snippets en dépendances. Citron utilise les règles syntaxiques et lexiques de XIP définis pour le SQR FIDJI (Moriceau & Tannier, 2010) développé précédemment que nous avons complétés notamment pour analyser les éléments structuraux prétraités par Kitten. La validation du type des candidats-réponses utilise Wikipédia dans une approche similaire à (Grappy, 2011) mais nous n'utilisons que l'introduction de l'article correspondant au candidat à valider pour y rechercher des définitions. Finalement, l'agrégation de réponses a pour but de regrouper des réponses désignant une même entité. Citron utilise une approche surfacique et une normalisation des dates contenues dans les supports des réponses effectuée à l'aide de HeidelTime (Strötgen & Gertz, 2013). Cette agrégation permet également de regrouper les réponses selon un critère variant temporel (Falco, 2012) (voir partie gauche de la figure 1). Nous ne détaillons pas plus ici le fonctionnement de Citron qui est décrit précisément dans (Falco *et al.*, 2013).

3 Expériences utilisateur : présentation du protocole

Nous présentons dans cette section le protocole expérimental utilisé pour nos deux expériences utilisateurs réalisées consécutivement : (1) l'extraction de réponses (*extraction*) : son objectif est de comparer les performances de Citron par rapport à celles d'un être humain pour la tâche d'extraction de réponses multiples depuis une collection de documents imposés issus du Web. Cette première expérience permet surtout d'étudier ce que les utilisateurs considèrent comme des réponses correctes et comment ils les rédigent/présentent ; et (2) la satisfaction devant la présentation de réponses (*satisfaction*) : son objectif est de connaître la satisfaction d'un utilisateur devant des réponses extraites et formatées de deux façons différentes, l'une au format d'une campagne d'évaluation et l'autre telle que Citron les présenterait.

3.1 Données d'évaluation et outils

Nous avons généré deux jeux (*jeuA* et *jeuB*) de 10 questions-ARM homogènes au niveau des types de questions proposées : 5 questions dont le type attendu de la réponse est général (une entité nommée) et 5 dont le type est spécifique. Chaque question-ARM d'un jeu de questions a son miroir syntaxique dans l'autre jeu avec un changement de focus :
– type attendu général : *Qui a joué Knocking on Heaven's Door ?* (A), *Qui a joué Where Did you Sleep last Night ?* (B)
– type attendu spécifique : *Quels sont les **signes** du zodiaque ?* (A), *Quels sont les **péchés** capitaux ?* (B)
Pour chaque utilisateur, un jeu sert pour l'expérience *extraction* et l'autre pour *satisfaction*. Pour chaque question, nous avons indexé grâce à Lucene les 30 premiers documents renvoyés par Google pour la requête générée par Citron et prétraités par Kitten. À l'aide de cette même requête sur les 3 index (texte, SE et tableaux), on obtient une liste de snippets pour chaque question. Nous avons ensuite utilisé WebAnnotator (Tannier, 2012) afin d'annoter toutes les occurrences de réponses jugées correctes, soit au total 1084 occurrences. Puis nous les avons agrégées lorsque nécessaire (par exemple, *Olympique de Marseille* et *OM* désignent une même entité). Ces 1084 réponses correctes correspondent à 280 entités différentes.

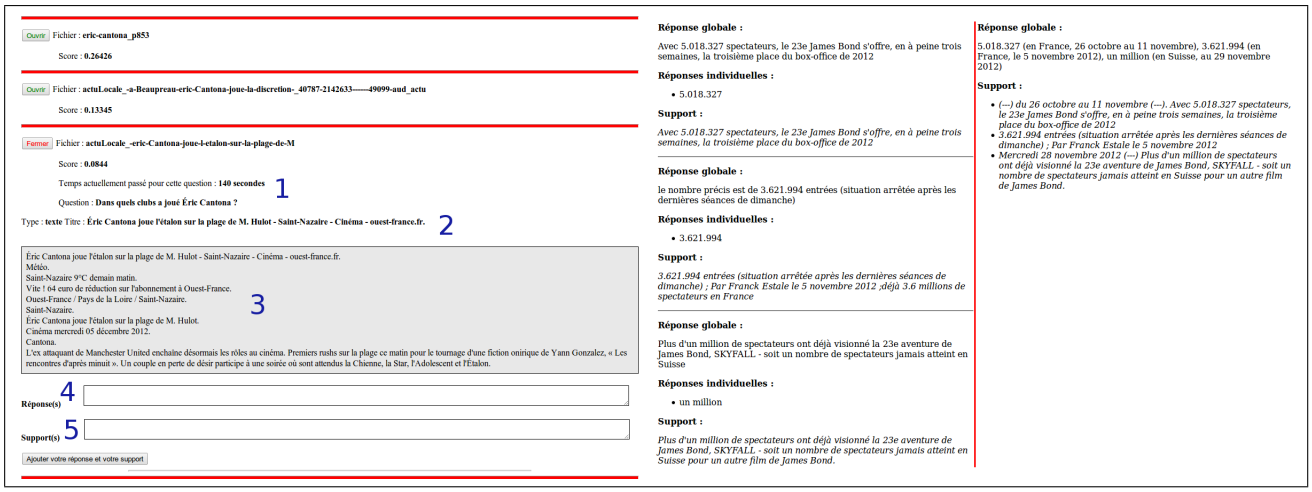

FIGURE 1 – À gauche : zone de saisie de réponses pour l'expérience d'*extraction*. À droite : présentation des réponses en deux colonnes pour l'expérience *satisfaction* sur la question *Combien de spectateurs ont vu Skyfall ?* (Formatage campagne d'évaluation à gauche, formatage Citron à droite).

3.2 Interface d'utilisation

Les deux expériences d'*extraction* et de *satisfaction* se font par l'intermédiaire d'une interface Web que nous avons développée avec le framework Django [1]. Un tutoriel était proposé avant le début de chaque expérience. La figure 1 présente à gauche l'interface pour l'évaluation de l'extraction de réponses et à droite celle pour la présentation des réponses.

Dans l'interface pour l'extraction des réponses, pour chaque question, les documents sont présentés les uns en dessous des autres et fermés au début. L'utilisateur choisit les documents à ouvrir et une fois un document ouvert, s'affichent le temps écoulé depuis le début de la question et le type du snippet (texte, tableau ou SE). Si l'utilisateur pense que le snippet contient une réponse correcte, il peut l'extraire avec un support (défini par *ce que vous pensez être le texte justifiant la ou les réponses extraites*) dans la zone de formulaire. Si un snippet contient plusieurs réponses correctes, l'utilisateur peut saisir une réponse à la fois ou plusieurs réponses en même temps. Les contraintes sont que la réponse doit être extraite depuis le snippet et qu'elle doit être accompagnée d'un support non vide. L'utilisateur est informé qu'un snippet ne contient pas forcément de réponse, qu'il y a au moins deux réponses/entités différentes pour chaque question et qu'il peut passer à la question suivante dès qu'il le souhaite, y compris sans fournir la moindre réponse.

Les données présentées dans la seconde interface sont les réponses de deux utilisateurs durant l'expérience *extraction* formatées de deux façons : une selon les critères de la campagne Quaero 2009 et l'autre tel que Citron les présenterait :
– format Quaero : de 1 à 3 bloc-réponses. Un bloc-réponse se compose d'une réponse globale continue (250 caractère maximum), d'un support continu (8 000 caractères maximum) et de n réponses individuelles provenant du support ;
– format Citron : une réponse globale composée des n réponses individuelles choisies par les deux utilisateurs et une liste de supports (pas forcément continus). Chaque réponse individuelle peut être accompagnée de critères variants.

Chaque formatage est présenté aléatoirement cinq fois à droite et cinq fois à gauche. Des questions sont posées à l'utilisateur pour lui demander quel format de réponse il préfère.

3.3 Les utilisateurs

32 utilisateurs ont passé les 2 expériences : 46,88 % ont passé l'expérience *extraction* sur le *jeuA* et *satisfaction* sur le *jeuB*. Avant de commencer l'expérience, l'utilisateur répond à 6 questions dans le but de définir son profil : sa tranche d'âge, s'il travaille dans la recherche, dans le TAL, combien d'articles scientifiques portant sur les SQR il a lus [2], combien de requêtes il effectue sur un moteur de recherche quotidiennement et son niveau d'informatique [3].

1. `http://djangoproject.com`
2. Nous avons considéré qu'une personne était sensibilisée à la thématique question-réponse si elle a lu au moins un article.
3. Le niveau *normal* correspond à une utilisation quotidienne d'Internet et d'un logiciel de bureautique, celui *intermédiaire* implique l'utilisation d'un programme plus poussé (retouche d'image, musique).

10-25 ans	26-40 ans	41 ans et +	Dans la recherche	Dans le TAL	Lecture SQR	+ de 10 requêtes	niveau normal	niveau intermédiaire	niveau avancé
28,13 %	53,12 %	18,75 %	59,38 %	37,50 %	53,13 %	51,61 %	31,25 %	12,50 %	56,25 %

TABLE 1 – Répartition des profils utilisateurs.

4 Résultats de l'extraction des réponses

4.1 Annotations des réponses et mesures d'évaluation

Les utilisateurs étaient libres de saisir dans le formulaire une seule réponse à la fois ou plusieurs en même temps tant qu'elles provenaient du même snippet. Nous avons donc d'abord segmenté toutes les réponses fournies en réponses individuelles et sommes arrivés à un total de 2319 réponses/entités individuelles, soit une moyenne de 72,47 réponses individuelles par utilisateur et de 7,27 réponses par question. Ces 2319 réponses proviennent pour 33,72 % de structures énumératives, 24,28 % de tableaux et 42 % de texte. Nous les avons annotées manuellement avec les statuts de réponses de la campagne Quaero que nous avons complétés ainsi (le statut Quaero est donné entre parenthèses) :

– (full) **Correct full** : la réponse est correcte, le support valide entièrement la réponse ;
– (full) **Correct OK** : la réponse est correcte, le support valide presque la réponse. Par exemple, pour *Quelles villes ont été la capitale des États-Unis ?*, la réponse *Philadelphie* associée au support *Philadelphie, capitale originelle* ne permet pas de savoir qu'il est question des États-Unis sans regarder le snippet ;
– (right) **Correct** : la réponse est correcte, provient du snippet mais le support ne valide pas la réponse ;
– (unsupported) **Correct absente extrait** : la réponse est correcte mais n'a pas été extraite depuis le snippet ;
– (inexact) **Correct non segmentée** : la réponse est correcte mais mal segmentée. C'est le cas par exemple quand on fournit une phrase complète sans extraire ce qui est considérée comme la réponse ;
– (inexact) **Correct pb orthographe** : la réponse est correcte mais contient une faute d'orthographe ;
– (inexact) **Correct rédigée** : la réponse est correcte mais l'utilisateur l'a rédigée au lieu de simplement l'extraire. Par exemple : *C'est NIRVANA qui a joué "Where did you sleep last night"* ;
– (supported) **Incorrect mais support correct** : la réponse est incorrecte mais le support contient une réponse correcte ;
– (false) **Incorrect contradiction** : la réponse est incorrecte et il était possible de s'en rendre compte par recoupement avec un autre extrait proposé (aucune occurrence dans cette évaluation) ;
– (false) **Incorrect** : la réponse est incorrecte.

La campagne Quaero ne comptait comme réponse valide que les réponses obtenant le statut *right* ou *full*. De notre côté, nous avons réalisé deux évaluations : une évaluation *QR* où seules les réponses ayant les statuts *Correct full*, *Correct OK* et *Correct* sont considérées comme valides (nous mesurons ainsi la tâche réelle d'extraction correcte de réponse ainsi que sa justification par le support) et une évaluation *humaine* où toute réponse dont le statut commence par *Correct* est considérée comme valide (ce sont les réponses qui peuvent être jugées acceptables par des utilisateurs).

Le tableau 2 présente les statuts des réponses extraites par Citron et par les 32 utilisateurs. Nous y constatons que par une évaluation *humaine*, il n'y a quasiment pas de réponses incorrectes parmi celles extraites par les utilisateurs (3,11 %). On constate également que quelques réponses ont été rédigées, principalement pour apporter des précisions, notamment en terme de critère variant. Par exemple pour le nombre de spectateurs ayant vu Skyfall, des utilisateurs ont ajouté la date ou le lieu comme Citron le fait (figure 1). Ce comportement est le résultat le plus important de cette évaluation puisque pour 65 % des réponses rédigées (11 sur 17), les utilisateurs ont ajouté naturellement un critère variant tout comme le propose Citron. Enfin, du point de vue de l'évaluation *QR*, les utilisateurs ont extrait 86,11 % de réponses valides uniquement avec leur support, ce qui est au-dessus des performances de Citron (72,34 %). Contrairement aux utilisateurs, Citron fournit une proportion importante de réponses incorrecte (27,66 %), les causes de ces erreurs sont détaillées juste après.

Nous avons utilisé la F-mesure pour évaluer la qualité de l'extraction des réponses. Nous avons aussi souhaité comparer les performances de Citron en terme de vitesse d'exécution par rapport à celles humaines. Pour cela nous avons mesuré les vitesses de Citron et ceux des utilisateurs pour répondre à chaque question. Pour chaque utilisateur, nous comparons sa vitesse par rapport au plus rapide des utilisateurs à fournir une réponse correcte : $V = min(\frac{n_{rapide}}{n_{utilisateur}}, 1)$ où, pour une question, n_{rapide} est le nombre de secondes le plus petit nécessaire pour fournir une réponse correcte parmi tous les utilisateurs humains et $n_{utilisateur}$ le nombre de secondes passées par l'utilisateur (le minimum sert pour Citron qui est très souvent plus rapide que le plus rapide des utilisateurs).

	Moyenne des deux jeux (% (#))		***JeuA*** (% (#))		***JeuB*** (% (#))	
Source / **Statut**	Citron	Utilisateurs	Citron	Utilisateurs	Citron	Utilisateurs
Correct full	72,36 (17)	46,74 (1084)	72 (18)	53,77 (556)	72,72 (16)	41,09 (528)
Correct OK	0	22,42 (520)	0	12,00 (124)	0	30,82 (396)
Correct	0	16,95 (393)	0	25,53 (264)	0	10,04 (129)
Correct absente extrait	0	0,26 (6)	0	0,19 (2)	0	0,31 (4)
Correct non segmentée	0	7,98 (185)	0	4,93 (51)	0	10,42 (134)
Correct pb orthographe	0	0,69 (16)	0	0,29 (3)	0	1,01 (13)
Correct rédigée	0	1,85 (43)	0	0,10 (1)	0	3,27 (42)
Incorrect mais support correct	16,82 (4)	1,34 (31)	20 (5)	0,58 (6)	13,64 (3)	1,95 (25)
Incorrect	10,82 (2,5)	1,77 (41)	8 (2)	2.61 (27)	13,64 (3)	1,09 (14)

TABLE 2 – Répartition des réponses individuelles des 32 utilisateurs et de Citron.

4.2 Résultats

Profil	$\mathbf{P_{hum}}$	$\mathbf{R_{hum}}$	$\mathbf{F_{hum}}$	$\mathbf{P_{QR}}$	$\mathbf{R_{QR}}$	$\mathbf{F_{QR}}$	$\mathbf{V_{QR}}$
TAL	0,91	0,61	0,68	0,87	0,59	**0,65**	**0,22**
Non TAL	0,86	0,55	0,62	0,77	0,51	0,57	0,21
Recherche	0,94	0,67	0,73	0,87	0,64	**0,69**	0,19
Non Recherche	0,81	0,45	0,52	0,71	0,42	0,48	**0,22**
Nb requête quotidienne <=10	0,83	0,47	0,54	0,73	0,43	0,49	**0,21**
Nb requête quotidienne > 10	0,93	0,67	0,73	0,88	0,64	**0,70**	0,20
Niveau informatique normal	0,80	0,47	0,53	0,76	0,45	0,52	**0,24**
Niveau informatique intermédiaire	0,86	0,38	0,47	0,68	0,33	0,40	0,21
Niveau informatique avancé	0,95	0,68	0,75	0,87	0,65	**0,70**	0,18
Âge 10-25 ans	0,89	0,59	0,65	0,76	0,54	0,59	**0,22**
Âge 26-40 ans	0,93	0,64	0,71	0,88	0,62	**0,68**	0,21
Âge 41 ans et plus	0,76	0,42	0,47	0,69	0,40	0,45	0,16
Aucun article QR lu	0,81	0,51	0,57	0,75	0,49	0,54	**0,23**
De 1 à 5 articles QR lus	0,91	0,63	0,69	0,82	0,59	0,65	0,20
>= 6 articles QR lus	0,95	0,66	0,73	0,92	0,65	**0,72**	0,19
Moyenne humaine (sur la totalité des questions)	0,88	0,58	0,65	0,81	0,55	**0,61**	0,21
Moyenne Citron (sur la totalité des questions)	0,47	0,36	0,37	0,47	0,36	0,37	**0,89**
Moyenne humaine (sur les questions auxquelles Citron répond)	0,91	0,64	0,70	0,85	0,61	**0,66**	0,17
Moyenne Citron (sur les questions auxquelles Citron répond)	0,72	0,56	0,56	0,72	0,56	0,56	**0,88**

TABLE 3 – Moyennes de *jeuA* et *jeuB* (*hum* pour l'évaluation des réponses sur des critères humains et *QR* selon les critères d'une campagne d'évaluation). P pour précision, R pour rappel, F pour F-mesure, V pour vitesse

.

Le tableau 3 présente les résultats pour les deux jeux de question. La **rapidité** se lit à l'aide de la mesure V_{QR} et montre que plus les utilisateurs sont jeunes, plus ils ont été rapides (+31 et +38 % pour les 2 tranches d'âge les plus jeunes). Il faut noter que la quasi-totalité des utilisateurs a choisi d'extraire le maximum de réponses différentes possibles pour chaque question avec une durée moyenne de 238,37 secondes pour une médiane de 239,77. Par exemple, seul 9 % des utilisateurs n'ont pas répondu à une question de leur jeu. Citron a été conçu pour répondre de façon rapide aux questions et se retrouve sans surprise le plus rapide, Citron passant en moyenne 22 secondes [4] par question. Même pour les questions auxquelles Citron décide de ne pas répondre, le temps consommé pour extraire les candidats-réponses et finalement ne pas les proposer ne le désavantage pas au niveau temporel. Le coefficient de corrélation de Bravais-Pearson montre qu'il n'y a pas de corrélation entre le temps passé et la qualité de l'extraction.

Pour la **qualité de l'extraction des réponses**, on constate que la F-mesure F est naturellement plus élevée selon les critères d'évaluation *humain* plutôt que *QR*. Citron est nettement moins performant principalement parce qu'il ne répond

4. 16 processeurs INTEL QUAD 5570, 50 Go de RAM, Ubuntu SMP

pas à 7 questions sur les 20. Pour ces questions, Citron a notamment rencontré des problèmes d'absence de candidats-réponses du type attendu et de non application de nos règles XIP (problème de conversion HTML, question en anglais). Citron a une forte précision : si on ne prend en compte que les questions auxquelles Citron a répondu, il n'est alors en moyenne que de 10 points en F-mesure derrière les utilisateurs, obtenant même un meilleur score pour le *jeuB*.

5 Résultats de la satisfaction des utilisateurs

Le tableau 4 montre que 71,39 % des utilisateurs ont trouvé que la présentation des réponses *à la* Citron répondait mieux à la question. La présentation *à la* Citron est vue comme justifiant le mieux ses réponses à 61,80 %, ce qui semble confirmer l'intérêt d'ajouter le critère variant aux réponses extraites quand il existe.

En ne prenant en compte que les réponses globales, quelle colonne répond le mieux à la question posée ?				
	très nettement la gauche	plutôt la gauche	plutôt la droite	très nettement la droite
Moyenne (jeux A et B)	7,27	21,34	**30,0**	**41,39**
Quelle colonne justifie le mieux ses réponses ?				
	très nettement la gauche	plutôt la gauche	plutôt la droite	très nettement la droite
Moyenne (jeux A et B)	8,46	29,74	**42,73**	**19,07**

TABLE 4 – Préférences de présentation des réponses : *gauche* pour le format campagne d'évaluation et *droite* pour Citron.

6 Conclusion

Citron est un système de question-réponse en français capable d'extraire des réponses à des questions à réponses multiples en domaine ouvert à partir de documents provenant du Web. La particularité de Citron est qu'il exploite les structures (énumérations, tableaux) propres à ces documents pour mieux extraire les réponses et éventuellement les agréger pour faire apparaître des critères variants comme la date ou le lieu. Nous nous sommes intéressés ici à la comparaison de performances entre des humains et notre système de question- réponse Citron pour la tâche d'extraction de réponses à des questions à réponses multiples. Nos différentes expériences nous ont permis de confirmer l'hypothèse intuitive qu'un être humain est plus performant que notre système mais à un coût temporel très fort. Elles ont aussi confirmé l'importance pour un système de pouvoir traiter les éléments structuraux comme les tableaux. Mais ces évaluations ont surtout permis de montrer que les utilisateurs produisent eux-aussi des réponses présentant des critères variants et préfèrent la présentation des réponses offerte par Citron, notamment parce qu'elle aide mieux à comprendre la justification des réponses et leur multiplicité.

Références

AÏT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2002). Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, **8**, 121–144.

AYACHE C. (2005). Evaluation en question-réponse, rapport final de la campagne EVALDA. In *Campagne EQUER*.

FALCO M.-H. (2012). Typologie des questions à réponses multiples pour un système de question-réponse. In *RECITAL*, p. 191–204, Grenoble, France.

FALCO M.-H., MORICEAU V. & VILNAT A. (2012). Kitten : a tool for normalizing HTML and extracting its textual content. In *Proceedings of the Eight International Conference on (LREC'12)*, Istanbul, Turkey.

FALCO M.-H., MORICEAU V. & VILNAT A. (2013). Répondre à des questions à réponses multiples : premières expérimentations. In *Conférence francophone en Recherche d'Information et Applications (CORIA)*.

GRAPPY A. (2011). *Validation de réponse dans un système de question-réponse*. PhD thesis, Université Paris-Sud.

HATCHER E., GOSPODNETIC O. & MCCANDLESS M. (2010). *Lucene in action, Second Edition*. Manning.

MORICEAU V. & TANNIER X. (2010). FIDJI : Using Syntax for Validating Answers in Multiple Documents. *Information Retrieval, Special Issue on Focused Information Retrieval*, **13**(5), 507–533.

QUINTARD L., GALIBERT O., ADDA G., GRAU B., LAURENT D., MORICEAU V., ROSSET S., TANNIER X. & VILNAT A. (2010). Question Answering on Web Data : The QA Evaluation in Quaero. In *LREC*, Valletta, Malta.

STRÖTGEN J. & GERTZ M. (2013). Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, **47**(2), 269–298.

TANNIER X. (2012). WebAnnotator, an Annotation Tool for Web Pages. In *LREC 2012*, Istanbul, Turkey.