

Projet ANR blanc CONIQUE (2005-2008)

Inférences en contexte pour trouver, justifier et présenter des réponses à des questions en domaine ouvert.

Partenaires : LIMSI, CEA LIST, MoDyCo

Dans le domaine de la recherche d'information, l'un des défis actuels porte sur la détermination de l'information précise cherchée par un utilisateur. L'objectif est de dépasser le paradigme de la recherche documentaire, dans lequel le système laisse à la charge de l'utilisateur le soin d'explorer une liste de documents pour y trouver l'information qu'il cherche, pour reporter la plus grande partie de ce travail sur le système de recherche d'information. Cette focalisation sur la recherche précise d'information s'est concrétisée ces dernières années par un intérêt porté aux systèmes de question-réponse en domaine ouvert. L'objectif de ces systèmes est de fournir une réponse à une question factuelle exprimée en langage naturel en trouvant cette réponse dans un ensemble de documents. La plupart des systèmes sont à même d'extraire la réponse à une question lorsqu'elle est explicitement présente dans les textes mais dans le cas contraire, ils ne sont pas capables d'agencer différents morceaux d'information dans le cadre d'un raisonnement pour produire une réponse complètement justifiée. Ainsi, pour nous, une réponse est la réponse exacte associée aux extraits de textes qui ont permis de la trouver, extraits permettant de justifier cette réponse exacte aux yeux d'un utilisateur.

Le projet CONIQUE a pour objectif d'atteindre ce but. Le premier axe de notre projet a pour objectif non pas de constituer ou d'exploiter une base de connaissances a priori permettant de répondre aux questions et de les justifier mais de modéliser l'extraction de ces connaissances à partir de différents textes en fonction des besoins nécessaires à la construction d'un chemin inférentiel entre les éléments trouvés dans les textes et l'information cherchée, telle qu'elle est spécifiée par une question. Outre le fait de combler l'inévitable incomplétude des bases de connaissances face à un travail en domaine ouvert, l'extraction des connaissances à partir de textes présente l'intérêt de pouvoir disposer en parallèle du contexte d'usage et de validité de ces connaissances. Ce contexte est particulièrement important pour contrôler l'enchaînement des inférences et leur validité dans le processus de recherche d'une réponse mais il est aussi très intéressant pour la présentation des réponses. Le second axe concerne ainsi la présentation des réponses possibles à une question en les accompagnant de leur contexte afin de permettre à l'utilisateur de comprendre l'origine de leurs différences.

1. Recherche de réponses par inférences

L'objectif est de produire des réponses intégralement validées (justifiées) par un extrait de texte issu de la collection de documents. La difficulté principale est que la réponse (ou une partie de l'information portée par la réponse) peut être validée par plusieurs documents. Par exemple :

Question : *Quel premier ministre français s'est suicidé ?*

Réponse : *Pierre Bérégovoy*

Passage justificatif 1 : *Il y a deux ans, Pierre Bérégovoy s'est suicidé à la suite de son implication...*

Passage justificatif 2 : *Le premier ministre français Pierre Bérégovoy a prévenu Mr. Clinton contre...*

On voit grâce à cet exemple que la question peut être décomposée en deux sous-questions successives : *Qui s'est suicidé ?* et pour chacune des réponses possibles, *cette personne est-elle premier ministre français ?* et qu'il faut combiner les deux résultats pour trouver et valider la réponse correcte. C'est ce type de raisonnement qui a été étudié et mis en œuvre dans CONIQUE par le système FIDJI, qui se fonde sur la représentation d'une demande d'information précise par l'ensemble des relations entre les entités données dans la question, relations qui doivent être vérifiées dans des textes.

FIDJI utilise l'information syntaxique produite par l'analyseur de phrase Syntex¹ pour construire une représentation uniforme des questions et des documents. Le système détecte, pour une question donnée, si les relations de la question peuvent être retrouvées dans un ou plusieurs documents, en réalisant cet appariement si nécessaire avec différentes réécritures des relations. Ainsi, si les dépendances de la question sont

¹ D. Bourigault and C. Fabre. 2000. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25.

entièrement retrouvées dans une phrase, la réponse peut en être extraite directement. Nous verrons ci-dessous les relations que nous avons plus spécifiquement étudiées dans le cadre de CONIQUE.

Afin de reconnaître les types de réponse recherchés, les textes sont étiquetés par environ 160 types d'entités nommées. Cet étiquetage, combiné à l'analyse de la question, permet de vérifier la concordance entre le type attendu et celui de la réponse candidate.

2. Relations étudiées

Les problèmes que nous avons plus spécifiquement étudiés concernent :

- l'identification de relations dans les textes sous leurs différentes formes linguistiques. Les relations présentes dans les questions que nous avons retenues dans le projet sont les suivantes :
 - informations temporelles
 - relations prédicatives entre deux entités
 - relation d'hyponymie
- la réinterrogation pour vérifier certaines relations attendues dans le passage réponse

2.1. Relations temporelles

Dans les systèmes actuels, le calcul de la cohérence temporelle s'appuie d'une part sur la reconnaissance, dans la question et la réponse, d'expressions calendaires absolues et d'autre part sur une comparaison entre ces expressions calendaires reconnues. Ce calcul ne prend généralement pas en compte les prépositions ou locutions prépositionnelles (en, depuis, au cours, etc.) qui précèdent l'expression d'une date ou d'une durée. Conceptuellement, nous avons proposé dans le cadre de ce projet d'améliorer ce calcul en prenant en compte le type de la préposition qui figure dans la question (PrQ) et dans la réponse (PrR) : les réponses proposées par le système de Q/R sont annotées par les transducteurs qui identifient les expressions calendaires, les prépositions employées et leurs positions (PosPrR et PosPrQ) dans le graphe (cf. Fig. 1). Le système recherche s'il existe un chemin entre PosPrR et PosPrQ et détermine le type de relation de façon plus ou moins précise en fonction des valeurs instanciées (précédence ou chevauchement) puis dans le cas du chevauchement, (inclusion ou non), et dans le cas de l'inclusion (début, intérieur, fin).

A ce jour, le prototype effectue une comparaison entre l'expression calendaire de la question et celle du passage candidat en permettant de dire si les références se chevauchent ou sont en relation de précédence selon la figure 1.

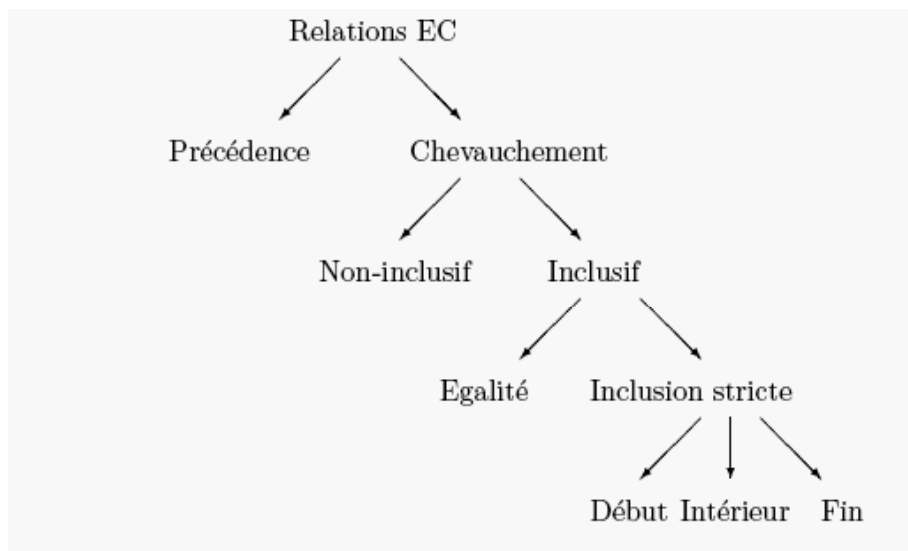


Figure 1 : Graphe des relations temporelles entre expressions calendaires

Plus précisément, nous avons évalué la situation de la référence temporelle de la réponse proposée par rapport à la référence temporelle de la question selon les valeurs suivantes :

- identique (cas d'égalité),
- incluse (cas d'inclusion stricte de la référence temporelle de la réponse par rapport à celle de la question),
- contient (cas d'inclusion stricte de la référence temporelle de la question par rapport à celle de la réponse),

- disjointe (cas de précédence).

Cette dernière valeur est un indice fort de non adéquation de la réponse par rapport à la question. Les autres valeurs sont de simples indices de pertinence que les autres relations étudiées doivent permettre de confirmer ou d'infirmer.

Une expérimentation visant à valider notre approche a été effectuée sur le corpus AVE 2006 (corpus qui sera présenté en section 3).

2.2. *Relations prédicatives*

Comme nous avons pu le voir précédemment, la vérification de l'identité entre le réseau de relations d'une question et le réseau de relations d'un passage candidat peut reposer sur l'exploitation de relations issues d'une analyse syntaxique. La vérification de relations de type paradigmatique (synonymie, hyperonymie, etc.) ou de relations syntagmatiques plus spécialisées demande des moyens différents. Parmi ceux-ci, l'utilisation de patrons lexico-syntaxiques occupe une place importante depuis le travail fondateur de Hearst².

Dans le cadre de CONIQUE, nous avons choisi de développer un mécanisme de vérification de relations fondé sur des patrons lexico-syntaxiques. L'utilisation d'un apprentissage supervisé permet de plus de prendre en compte facilement de nouveaux types de relation, tout en offrant des capacités d'expression importantes au niveau des patrons par l'intégration de différents niveaux d'information. Plus précisément, ces patrons peuvent faire référence à la forme de surface, à la catégorie morpho-syntaxique ou à la forme lemmatisée des mots. Ils sont appris grâce à un algorithme fondé sur la distance d'édition appliqué à un ensemble de phrases exemples de la relation considérée au sein desquelles sont identifiés les deux termes de cette relation.

Nous avons appliqué ce mécanisme de vérification dans le domaine médical pour des relations syntagmatiques de type [*Médicament*] – (*traiter*) – [*Maladie*] ou [*Examen*] – (*détecter*) – [*Maladie*]. Un patron appris tel que <*médicament*> utilisé pour VERBE_PRINC_INFINIT * <*maladie*> permet ainsi de valider la présence de la relation [*Médicament*] – (*traiter*) – [*Maladie*] dans la phrase *Le vaccin utilisé pour prévenir la fièvre aphteuse ...* Appliqué à l'extraction de relations, ce mécanisme de vérification a permis d'obtenir des performances moyennes sur quatre types de relations de 91% en précision et 64% en rappel (Embarek et Ferret, 2008).

Ce même mécanisme a été utilisé pour la validation de réponses dans le cadre du système de question-réponse Esculape, dédié au domaine médical (Embarek, 2008). Dans ce cas, les patrons de relations permettent de valider que les réponses trouvées se trouvent bien au cœur de la relation sous-jacente à la question. Le patron <*maladie*>, *se manifester* * par DET_ART_IND <*symptôme*> valide ainsi la réponse *sécheresse de la bouche* dans la phrase *... Botulisme, se manifeste par une sécheresse de la bouche ...* pour la question *Quels symptômes accompagnent le Botulisme ?*, imposant à la réponse d'entretenir la relation [*Symptôme*] – (*être_signe*) – [*Maladie*] avec la maladie *Botulisme*. Ce mécanisme de validation des réponses trouvées a montré son efficacité en permettant au système Esculape d'atteindre un niveau de performances de plus de 30% alors que le système Œdipe sur lequel il s'appuie présente des performances inférieures à 10%.

2.3. *Réinterrogation*

La stratégie utilisée dans FIDJI pour décider de réinterroger la base de connaissances que sont les documents analysés est guidée par la forme de la question. En effet, de nombreuses questions attendent une réponse d'un type particulier, type souvent absent du passage contenant la réponse :

- une entité nommée (EN : PERSONNE, LIEU...), comme dans *Qui est le président de la Russie ?*
- un type plus général comme dans *Quel président russe a assisté au sommet du G7 ? où on attend un "président russe" en plus de l'entité nommée PERSONNE*

L'analyse des questions consiste à identifier :

- les dépendances syntaxiques à rechercher dans un passage contenant la réponse
- le type de la question (factuelle, définition, booléenne, complexe...)

²

Marti Hearst (1992) Automatic acquisition of hyponyms from large text corpora. 14th International Conference on Computational Linguistics (COLING'92), Nantes, France.

- le type de réponse attendu (EN et/ou type général)

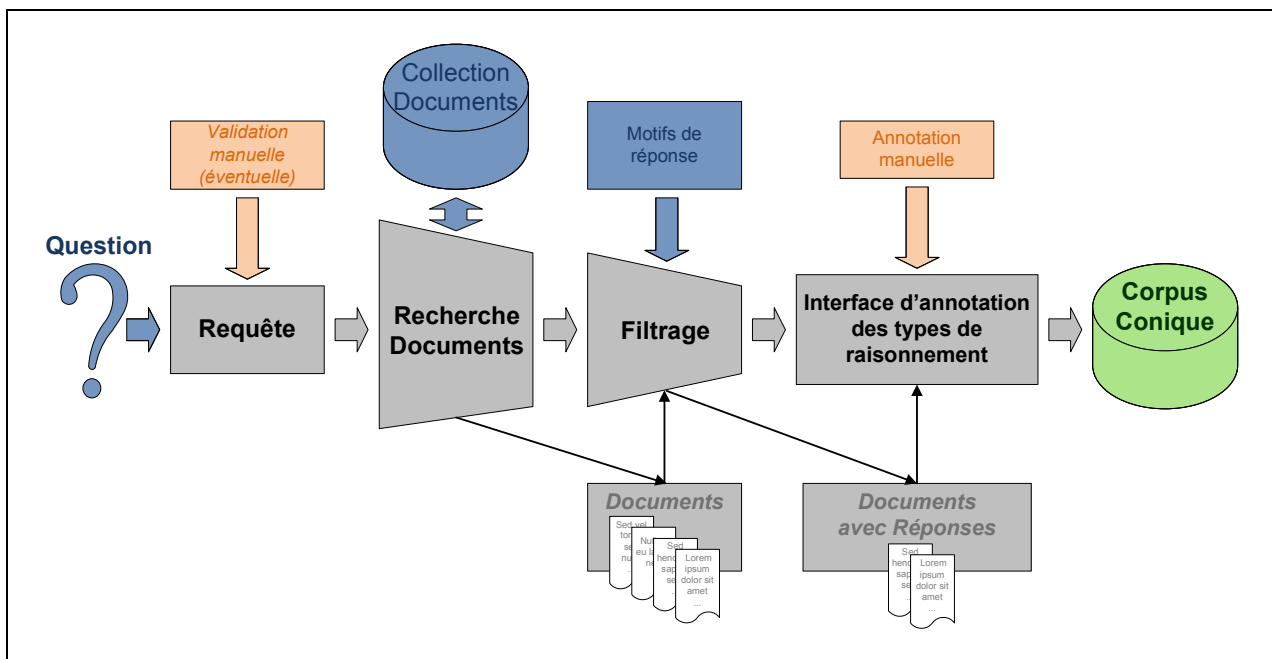
L'hypothèse est que le type de réponse, qui introduit une relation d'hyperonymie avec la réponse, peut être validé dans différents documents. Si le type de la réponse trouvée n'est pas vérifié dans le passage sélectionné, le système cherche alors la vérification des relations associées au type de la réponse dans d'autres documents de la collection.

L'application de ce type de stratégie à d'autres types de relations est à l'étude, ce qui implique de définir plus précisément le contexte de recherche à associer à ces autres types de relation.

3. Etude des justifications

L'évaluation de la problématique de la justification dans les systèmes de question-réponse n'est pas nouvelle puisqu'elle a déjà été abordée dans le cadre des campagnes AVE (Answer Validation Exercise) de CLEF-QA³. Néanmoins, il nous est apparu nécessaire de développer un corpus spécifique dans le cadre du projet CONIQUE car les corpus d'évaluation d'AVE présentent certaines insuffisances du fait de leur mode de constitution :

- ces corpus s'appuient sur les résultats des participants à la campagne CLEF-QA. De ce fait, ils limitent les phénomènes de justification couverts en privilégiant des cas simples puisque les systèmes cherchent prioritairement des réponses proches des questions ;
- les extraits de texte proposés en justification sont continus, généralement de la taille d'une phrase ;
- aucune typologie des problèmes de justification n'est proposée dans le cadre d'AVE ;
- le faible nombre de participants pour le français en tant que langue cible dans le cadre de CLEF-QA limite l'intérêt des corpus AVE pour le français.



Le schéma ci-dessus présente les étapes de la constitution du corpus CONIQUE.

Les éléments en entrée de ce processus sont les suivants : un jeu de questions, pour chacune de ces questions, ses réponses connues exprimées sous la forme de motifs de réponses, et un corpus de documents dans lequel la recherche en contexte va être effectuée.

Les questions sont d'abord transformées en requêtes à destination d'une étape de recherche documentaire conventionnelle. Ces requêtes couvrent des sous-ensembles des mots des questions. En simplifiant ainsi les questions, l'objectif est d'augmenter le rappel et de minimiser la perte des documents « lointains », i.e. contenant les informations de la question mais sous forme de variations complexes, susceptibles de contenir

³

Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo (2007) Testing the Reasoning for Question Answering Validation. Journal of Logic and Computation.

une réponse. Les documents retrouvés à cette étape sont ensuite filtrés grâce à la présence ou l'absence des motifs de réponses connues. À l'étape suivante d'annotation manuelle des types de raisonnements, seuls les documents contenant au moins une réponse sont examinés. Il est possible de filtrer davantage les documents retournés suivant un critère de distance entre les termes de la requête et ceux de la réponse.

Exemple : *En quelle année est né Jacques Carral ?*

Une requête possible n'est constituée que de (*Jacques#N*) et (*Carral#N*). « *né/nâître* » ainsi que « *année* » sont écartés afin de pas restreindre la recherche. La recherche reste ainsi ciblée sur le sujet principal, ce qui permet, par exemple, d'étudier un document qui contiendrait l'expression « *voir le jour le* » et qui aurait été éloigné du point de vue d'un score de pertinence par rapport à ceux contenant {*année, nâître, Jacques et Carral*}.

Un découpage des documents en phrases ayant été réalisé, une interface d'annotation met en évidence les mots de la requête utilisés pour trouver le document en cours d'examen, les entités nommées et les réponses connues dans ce document. Elle permet à l'annotateur de visualiser à la fois les indices permettant de répondre et les éléments de réponse.

En outre, l'annotation manuelle inclut le choix d'un type de justification, chaque type renvoyant à la nature des éléments à vérifier pour que la justification soit acceptable. Nous distinguons quatre types de justification :

- **Type de réponse.** Pour la question *Dans quel musée se trouve « Les noces de Cana » ?* la phrase *Les Noces de Cana de Véronèse, toile trop vaste pour être rapatriée sans dommage, sont demeurées au Louvre* contient une réponse valide sans que le type attendu de la réponse ne soit présent. Cette phrase ne permet pas en effet de vérifier la contrainte le *Louvre* est un *musée*. Cette information doit par conséquent être recherchée par ailleurs afin de compléter la chaîne inférentielle ;
- **Reformulation sémantique ou paraphrastique et inférence.** Dans l'exemple précédent, il est nécessaire de rapprocher « *se trouver dans* » avec « *demeurer* » pour être capable de justifier complètement la réponse par rapport à la question ;
- **Résolution de coréférence.** Dans le cas de la question *Quel groupe a racheté « Terre sauvage » ?* et du passage réponse *Le mensuel Terre sauvage... Le tribunal de commerce a accepté la proposition de Bayard Presse de reprendre le mensuel*, la justification de la réponse demande entre autres éléments à vérifier d'effectuer une résolution d'anaphore entre « *le mensuel* » et « *Terre Sauvage* » ;
- **Éléments de contexte à vérifier ou manquants.** Pour la question *Qui est le ministre algérien de la justice ?*, ni la phrase réponse *Le président a démis de leur fonction les ministres de la justice et de la communication, Mohamed Téguia et Benamar Zerhouni*, ni le document qui l'abrite ne font apparaître que les ministres mentionnés sont algériens. Cette contrainte de nature géographique doit être vérifiée par ailleurs.

Chacune des informations utilisées le long du processus est sauvegardée dans un fichier XML (question, étiquettes morphosyntaxiques de la question, requête, identifiant de documents, documents étiquetés avec morphosyntaxes et entités nommées, éléments de justifications ajoutés manuellement). L'ensemble de ces fichiers constitue le corpus CONIQUE. Ce corpus est élaboré dans le but de le mettre à disposition de la communauté.

4. Visualisation et navigation

Nous avons proposé de décrire formellement, pour le français, les expressions calendaires (notées EC), définies comme les expressions qui font référence directement à des unités de temps relatives aux divisions courantes des calendriers (Battistelli et al. 2006). Cette formalisation repose sur la distinction explicite entre différentes classes de marqueurs linguistiques qui apparaissent dans les EC selon les types d'opérations qu'elles mettent en œuvre. Cette approche nous a amenés à proposer des critères fins pour d'une part l'annotation automatique de ces expressions et d'autre part la navigation temporelle dans un texte ou un corpus de textes. Nous nous distinguons des approches classiques sur au moins deux points cruciaux : (i) notre but n'est pas de relier une EC ramenée à sa valeur – selon la norme ISO – à un événement dans un texte mais de relier qualitativement les EC d'un texte entre elles, c'est-à-dire établir leurs relations de positionnement relatif (l'ensemble de ces relations correspond à ce que nous appelons le calendrier propre au

texte) ; (ii) nous modélisons la sémantique d'une EC simple comme une expression fonctionnelle fondée sur une approche formelle dans laquelle les EC sont construites incrémentalement à partir d'un référent calendaire, transformé en EC par une première fonction, suivie de plusieurs fonctions opérant uniquement sur des EC (ce qui permet de parler d'une algèbre d'opérateurs). Un projet d'application qui consiste à aider à la lecture d'un texte biographique sur la base de cette algèbre a été mis en place. Un prototype logiciel avec des interfaces de visualisation a été développé (Battistelli et al. 2008a, 2008b).

Un travail est en cours concernant la présentation des réponses en fonction des caractéristiques présentes dans les passages réponses.

5. Evaluation

Le système FIDJI a été évalué *a posteriori* sur les collections des campagnes CLEF 2005 et 2006 et obtient respectivement 59,5 et 48,5 % de bonnes réponses, ce qui le place parmi les meilleurs systèmes de l'état de l'art.

Nous avons aussi participé à la campagne AVE (Answer Validation Exercise) en 2008 pour le français. Dans cette tâche, les systèmes considèrent des triplets (question, réponse, passage justificatif) et doivent décider si la réponse à la question est non seulement correcte, mais également bien justifiée par le passage.

Deux approches ont été testées lors de cette campagne :

- Une stratégie utilisant uniquement le système FIDJI, décrit ci-dessus ;
- Une stratégie par apprentissage, où plusieurs traits sont combinés pour valider les réponses (Ligozat et al., 2007a, 2007b ; Grappy *et al.*, 2008) : présence de mots communs entre la question et le texte, distance entre les mots, réponse donnée par FRASQUES (autre système de question-réponse du groupe LIR) ainsi que la réponse fournie par FIDJI.

Le premier système a obtenu une très bonne précision (88 %) mais un rappel relativement faible (42 %), tandis que le second système améliore le rappel et parvient à une F-mesure de 61 %. Ce dernier système s'est classé premier pour le français et second pour l'ensemble des langues (Moriceau *et al.*, 2008).

6. Bibliographie du projet

- Battistelli Delphine, Jean-Luc Minel, Sylviane Schwer * (2006) – « Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies », TAL vol. 47/2, 26 pages, 2006.
- Battistelli Delphine, Javier Couto, Jean-Luc Minel, Sylviane Schwer *(2008a) - « Representing and visualizing calendar expressions in texts », in actes STEP'08 (Symposium on Semantics in Systems for Text Processing), 22-24 septembre 2008, 10 pages, Venise.
- Battistelli Delphine, Javier Couto, Jean-Luc Minel, Sylviane Schwer *(2008b) - « Représentation algébrique des expressions calendaires et vue calendaire d'un texte », in actes TALN'08 (Traitement automatique du langage naturel 2008), 8-12 juin 2008, 10 pages, Avignon.
- Embarek Mehdi (2008) Un système de question-réponse dans le domaine médical – Le système Esculape. Thèse de l'Université Paris Est, Saclay, France.
- Embarek Mehdi et Olivier Ferret (2008) Learning patterns for building resources about semantic relations in the medical domain. 6th Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Maroc.
- Grappy Arnaud, Anne-Laure Ligozat, Brigitte Grau, Évaluation de la réponse d'un système de question-réponse et de sa justification, Coria 2008
- Ligozat Anne-Laure, Brigitte Grau, Anne Vilnat, Isabelle Robba, Arnaud Grappy (2007a) Towards an automatic validation of answers in Question Answering, 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI).
- Ligozat Anne-Laure, Brigitte Grau, Anne Vilnat, Isabelle Robba, Arnaud Grappy (2007b) Lexical validation of answers in Question Answering, The 2007 IEEE / WIC / ACM international conference on Web Intelligence (WI 07).
- Moriceau Véronique, Xavier Tannier, Arnaud Grappy, Brigitte Grau (2008) Justification of Answers by Verification of Dependency Relations - The French AVE Task, 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008).

* ordre alphabétique des auteurs