



HAL
open science

Terminologie et accès à l'information thématique et précise

Olivier Ferret, Brigitte Grau

► **To cite this version:**

Olivier Ferret, Brigitte Grau. Terminologie et accès à l'information thématique et précise. Mustapha el Hadi, Widad. Terminologie et accès à l'information, Hermes - Lavoisier, 2006. hal-02307367

HAL Id: hal-02307367

<https://hal.science/hal-02307367>

Submitted on 7 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chapitre 8

Terminologie et accès à l'information thématique et précise

1 Introduction

Lorsque l'on cherche à répondre à une question portant sur le nom d'un pays ou la date de naissance d'une personne, l'unité textuelle d'information qui s'impose naturellement est la réponse à la question, même si la présentation de cette réponse dans un contexte significatif est également à prendre en compte. Le problème de la définition de l'unité d'information à présenter à l'utilisateur ne se pose pas explicitement car il est en pratique confondu avec celui de la recherche de l'information. Dans l'accès à l'information thématique, cette question intervient de manière plus explicite car la définition des unités informationnelles, dans ce cas de nature thématique, a une plus grande autonomie vis-à-vis du processus de recherche. Dans un tel contexte, l'unité d'information textuelle se définit plus précisément comme un segment textuel homogène du point de vue thématique, ce que l'on désigne généralement sous le vocable de segment thématique. La seconde caractérisation de l'information thématique est constituée par l'identification du thème de chaque segment thématique d'un document, quel que soit son niveau de granularité.

Selon l'information recherchée dans un document, les approches envisagées ne s'appuieront pas sur les mêmes propriétés des textes. La délimitation des segments thématiques reposera en grande partie sur les propriétés lexicales du texte : quels termes sont réitérés dans le texte, comment ces termes se répartissent-ils dans l'ensemble du texte, et quelles relations partagent-ils les uns avec les autres ?

En ce qui concerne la recherche d'une information précise, factuelle, les systèmes de question-réponse extraient des documents la réponse à la question, avec sa justification, consistant en général à donner le passage qui en parle. Dans ce cas, ce n'est pas la thématique du document qui compte, mais le fait qu'il contienne un passage sémantiquement proche de la question, susceptible de contenir la réponse. L'évaluation de l'adéquation d'un passage ne relève pas des mêmes approches quand la question porte sur un domaine de spécialité connu a priori et quand la question peut porter sur n'importe quel domaine. En domaine de spécialité, le raisonnement peut se fonder sur une représentation conceptuelle riche du domaine, telle qu'une ontologie ou une terminologie pré-établie [MOL 00] et [RIN 03], et la recherche de réponses peut alors reposer sur un processus de déduction logique. En domaine ouvert, le processus de résolution s'appuiera plus sur les informations linguistiques et lexicales, et les termes sur lesquels vont se fonder les processus seront les termes issus des questions dont une équivalence sera recherchée dans les textes. Dans la section suivante, nous développerons essentiellement les approches relevant de la recherche de réponses en domaine ouvert, réponses qui sont extraites de textes ; cette thématique a donné lieu à de nombreux travaux et figure dans de nombreuses évaluations.

2. Recherche d'informations précises

Les systèmes de question-réponse connaissent un essor de plus en plus important. Pour preuve, ils font maintenant partie de toutes les conférences d'évaluation en recherche d'information. Certains sont maintenant des produits vendus dans le commerce (par exemple les systèmes Qristal de Synapse Développement¹, PowerAnswer de LCC²). Leur force réside en ce qu'ils proposent la réponse exacte ou un court extrait contenant la réponse cherchée, et évite ainsi le parcours de documents. Ces systèmes mettent en œuvre des techniques linguistiques généralement absentes des systèmes de recherche d'information classiques, mais qui se trouvent être nécessaires pour sélectionner et extraire l'information cherchée : on se doit d'adopter des techniques approfondies d'analyse de la langue si on veut augmenter la précision et la fiabilité des réponses extraites des textes. Précision signifie ici que l'on cherche à donner l'information exacte cherchée, en omettant les informations qui ne répondent pas au besoin de l'utilisateur.

Les composants classiquement présents dans un système de question-réponse sont (cf. figure 1) :

¹ <http://www.qristal.fr>

² <http://www.languagecomputer.com>

- l'analyse des questions en langue naturelle afin de déduire tous les indices permettant ensuite de guider la recherche et de vérifier la fiabilité des réponses proposées,
- la recherche de documents ou de passages, réalisée à partir de la requête construite selon les termes clefs de la question
- la sélection de passages, en fonction des caractéristiques de la question qui y sont présentes
- l'extraction des réponses et leur justification.

Les informations extraites des questions varient selon les systèmes, mais ils s'accordent tous pour essayer de typer le plus finement possible la réponse attendue. Ensuite, certains se fonderont plutôt sur une analyse sémantique [MOL 03], [MOL 02] afin d'extraire des relations entre les éléments de la questions, ou sur l'extraction de contraintes syntaxiques entre ceux-ci [HOV 01], [BUC 01]. Pour une description détaillée des systèmes de question-réponse, le lecteur pourra se reporter à [GRA 04].

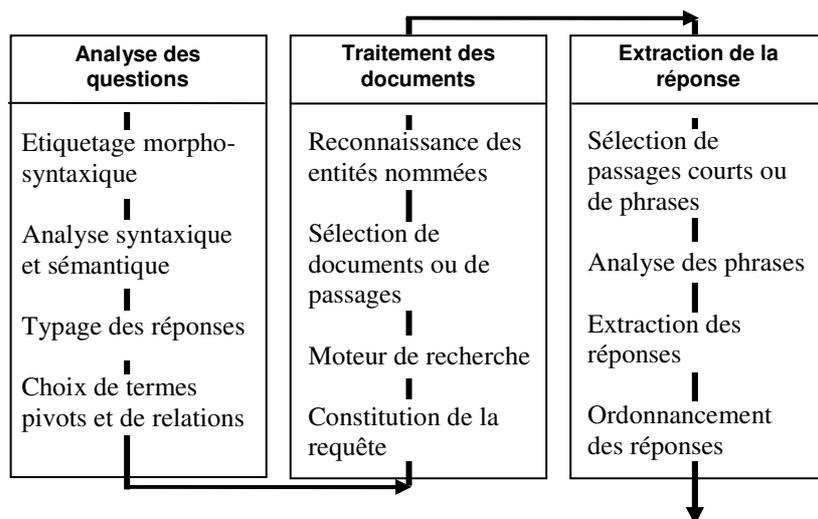


Figure 1. Schéma général d'un système de question réponse

Si le système est multilingue et part d'une question dans une langue source pour trouver une réponse dans une langue cible différente, il faut ajouter un module de traduction : soit pour traduire la question et le problème se ramène alors à un système monolingue sur la langue cible, soit pour traduire des termes importants de la question et l'analyse de la question s'effectue sur la langue source alors que la suite des processus travaillent sur la langue cible.

Chacun des composants doit considérer la variation linguistique : une réponse ne sera pas toujours formulée dans les mêmes termes que ceux de la question, ni sous la même forme syntaxique. Voici quelques exemples de telles variations :

– variation flexionnelle sur le verbe, reprise anaphorique et variations syntaxiques :

Q : Dans Pierre et Le Loup, qu'est-ce que mange le loup ?
 R : Rappelons que si, au dénouement, le loup est amené au zoo par les chasseurs alertés par Pierre, c'est qu'il a mangé tout le canard.

– variations sémantiques :

Q : Quelle maladie a tué Carlos d'Alessio ?
 R : LE DIMANCHE 14 JUIN 1992 DECES DE CARLOS D'ALESSIO Le compositeur est mort du sida le 14 juin à Paris.
 Q : En quelle année les éditions Ramsay ont-elles vu le jour ?
 R : C'est la quatrième fois depuis leur fondation par M. Jean-Pierre Ramsay, en 1976, que les éditions Ramsay changent de propriétaire.

Dans les sections suivantes, nous allons voir comment la variation des termes est traitée dans les différents composants d'un système de question-réponse.

Sélection de passages

La sélection d'un sous-ensemble de documents pertinents repose sur la reconnaissance des termes de la question ou de leurs variantes dans les documents sélectionnés par un moteur de recherche. Les systèmes de question-réponse utilisent souvent un moteur booléen pour la sélection des documents de la base. Il a en effet été montré que les moteurs vectoriels et booléens ont des performances comparables [TEL 03], mais les requêtes booléennes autorisent des manipulations plus fines. Les variations gérées à l'interrogation portent sur les mots simples de la question et sont les dérivations morphologiques et les variations sémantiques.

Les variations morphologiques représentent toutes les formes fléchies d'un mot (les différentes conjugaisons pour un verbe, la variation en nombre pour les noms et les flexions en genre et en nombre pour les adjectifs), ainsi que ses dérivations par ajout d'affixes (composer, composition, recomposer, etc.). Deux solutions existent pour rendre compte de ces dérivations. Le stemming permet de réduire tous les mots d'une même famille morphologique à une racine commune, par exemple : *compos*. Cette technique apporte une amélioration en question-réponse [MON 03], mais reste génératrice de bruits. Aussi, Bilotti et al. [BIL 04] ont comparé stemming et expansion de requêtes par ajout des différentes flexions et ont constaté un meilleur rappel avec la seconde technique.

La *famille sémantique* d'un mot simple m est l'union $S(m)$ des synonymes de ce mot. Par exemple, la famille sémantique de *tremblement* se compose de nombreux noms : *chevrotement, bêlement, frisson, soubresaut, frémissement, etc.*

Les patrons de variations qui reposent sur des familles morphologiques et sémantiques sont engendrés au moyen de métrarègles. Le patron suivant³, dénommé NtoV, extrait l'occurrence *terre a tremblé* comme variante de *tremblement de terre* :

VM ('tremblement') (PREP ? ART | PREP) JJ ? NN) N('terre')

où VM('tremblement') est tout verbe de la famille morphologique du nom *tremblement* est N('terre') le nom *terre*.

Les systèmes QALC et FRASQUES se limitent aux bi-termes, sachant que si un terme est plus long, il sera décomposé en plusieurs bi-termes. La recherche de termes complexes permet de mieux classer les passages contenant ces formes. Par exemple, il sera toujours préférable de rechercher un nom propre complet, en plus de ses composants (nom et prénom), plutôt que rechercher uniquement ses composants. Ainsi, *Harlem Désir* dans la question « *De quelle association Harlem Désir était-il le président ?* » permet de ne pas privilégier les documents comportant *Harlem* d'une part et *désir* d'autre part. Il en est de même pour la question « *Qui est président du Rwanda ?* », où les passages les mieux classés sont ceux qui comportent le terme « *président du Rwanda* », même si dans certains *président* et *Rwanda* sont proches. Par ailleurs, l'utilisation de FASTR permet de gérer par une même technique l'ensemble des phénomènes relatifs à la variation.

La recherche multilingue

La recherche de réponse dans un contexte multilingue, lorsque la stratégie choisie consiste à analyser la question dans la langue source et traduire les termes dans la langue cible, introduit une variation supplémentaire : passer d'un terme dans une langue à son équivalent dans une autre. Les traductions se font généralement en utilisant des dictionnaires bilingues puis en essayant de valider les traductions. [TAN 04] valident leurs traductions dans deux grands corpus. [SUT 05] combinent deux systèmes de traduction automatique et un dictionnaire pour traduire des expressions. MUSQAT [GRA 05b], le système bilingue issu de FRASQUES et QALC, traduit les bi-termes mot à mot à partir de deux dictionnaires bilingues, quand ceux-ci sont absents des dictionnaires utilisés, et ils sont ensuite recherchés par FASTR dans les documents sélectionnés. L'ensemble des termes donnés à FASTR conserve toutes les traductions possibles, sans lever d'ambiguïté. L'hypothèse faite dans MUSQAT est que si une traduction ou une variante de bi-terme se trouve dans un document, alors la traduction doit être correcte, et les

³ Nx sont les noms, JJx les adjectifs, RP les adverbes, IN les prépositions, DT les déterminants et Vx les verbes.

documents la contenant seront favorisés. Cette hypothèse est en général vérifiée. Dans [GRA 05], les auteurs ont effectué une validation manuelle des bi-termes, qui a été rapprochée des termes retrouvés par FASTR. Seuls 17% (19/110) des termes dont la traduction est fautive sont retrouvés dans des textes quand 45% des termes trouvés (105/233) peuvent être considérés comme des traductions unitaires valides, ceux-ci étant majoritairement constitués de bi-termes bien traduits (83/120). Cette validation de traduction en corpus devra permettre d'effectuer des choix de traduction quand des bi-termes sont présents dans les questions.

Types de réponse et types de termes

La principale caractéristique issue de l'analyse des questions concerne le type de réponse attendu, qui correspond soit à un type de réponse identifiable dans les textes hors contexte des questions, une entité nommée, ou à un type présent dans une base de connaissances, souvent une classe de WordNet [FEL 98] ou d'une base qui en est dérivée. L'étendue de la réponse dérive de ce type : une entité nommée donne lieu à un marquage dans les textes et y est explicitement délimitée. Une réponse d'un type plus général correspondra souvent à un terme exprimé par un groupe nominal, mais peut aussi amener à extraire des sous-phrases.

– Réponses de type Entité Nommée :

Quel est le nom du parti politique de M. Vajpayee ? → ORGANIZATION Qui a écrit le Comte de Monte Christo ? → PERSON Où se trouve le siège d'Adidas en France ? → LOCATION

– Réponses d'un type sémantique connu :

Quelle maladie a tué Carlos d'Alessio ? → maladie Quel métier exerçait Richard Scarry ? → métier

– Réponses de type général :

Qu'est-ce que le maire de Toulon a fait raser place Besagne ? → GN Comment est habillée Barbie dans le modèle « Solo in the Spotlight » ? → GP Citez un livre publié chez L'Harmattan en 1990 ? → Titre

Entité nommée

Selon les systèmes le niveau de granularité des types d'entité nommée définis varie énormément. Des catégories générales définies dans MUC6 [GRI 95] PERSONNE, ORGANISATION, LIEU, DATE, HEURE, MONTANT FINANCIER et POURCENTAGE, les systèmes de question-réponse ont dérivé ou ajouté des catégories plus précises. Les lieux, par exemple, peuvent être décomposés en pays, ville, cours d'eau, régions, montagnes, etc. de manière à pouvoir répondre à des questions telles que :

Dans quel pays se trouve Ljubljana ?
 Citez le nom d'une île grecque.
 Citez le nom d'une région azerbaïdjanaise peuplée majoritairement d'Arméniens.

Il en va de même pour les dates, les heures ainsi que pour toutes les quantités numériques. Certains systèmes ne font pas la distinction, au niveau du typage des réponses entre type d'entité nommée et autre type. Ainsi, Prager et al [PRA 00] ont défini 50 types de réponses, le système Webclopedia [HOV 01] en recense 122, appelées qtargets, constituées de noms d'entités nommées classiques et de catégories sémantiques venant d'une base de connaissances. Une classification très large est encore plus développée dans [HAR 00] qui s'appuie essentiellement sur WordNet.

Dans QALC et FRASQUES, le typage de la question concerne des types d'entités nommées classiques (figure 2); seules sont détaillées les entités numériques avec les différentes mesures (vitesse, montant financier, quantité physique, pourcentage, etc.) et les différents types de date (date comportant l'année, date ne donnant que le jour de l'année ou le mois, etc.). Pour les types correspondant à des classes sémantiques (ex. animal, fleur, etc.), ceux-ci sont considérés comme le type général de la question, et ne seront pas utilisés de la même manière, dans la mesure où les textes ne seront pas étiquetés préalablement à la sélection de la réponse. Ils jouent un rôle pour la validation sémantique de la réponse.

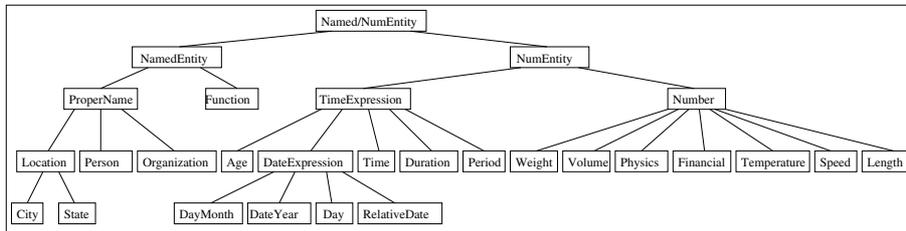


Figure 2. Hiérarchie des entités nommées de QALC

Le terme est un groupe nominal dont le type de la tête peut être connu

Une réponse factuelle est souvent exprimée par un groupe nominal. Afin de l'extraire, la plupart des systèmes ont une approche par patrons d'extraction qui vont tenter de décrire toutes les formulations possibles permettant d'exprimer l'information cherchée. Ces patrons sont en général associés à des types de question ou plus exactement à des types d'information cherchés et vont contenir certains mots ou termes pivots de la question. L'exemple le plus simple est celui d'une demande

de définition, que ce soit d'un terme, d'une organisation ou d'une personne, ou de la signification d'un acronyme.

<p>Q. Qu'est-ce que la Camorra ? R. The Camorra (Naples Mafia) members ... Q. Qui est Javier Solana ? R. Spanish Foreign Minister Javier Solana ... / The Spanish Foreign Minister, Javier Solana, ...</p>

Les patrons de définition sont de la forme suivante :

<p><Terme à définir> (définition) ou <Terme à définir> est <définition> <définition>, <NP à définir>, ou <définition><NP à définir></p>
--

Les définitions elles-mêmes sont décrites par un groupe nominal. Les autres types de question résolues par patron d'extraction peuvent porter, de manière non exhaustive, sur la recherche :

- d'une entité d'un type connu, un hyponyme, exemple *Quel grade occupe Juan-Carlos Rolon dans la marine ?*
- d'une caractéristique d'une entité, exemple *Quelle est la nationalité de Leyla Gencer ?*

Si on considère qu'une réponse est donnée par une phrase équivalente à la formulation à l'affirmative de la question, le patron décrit une sous-séquence de cette phrase, celle qui relie le sujet de la question à la réponse. Ainsi, la conception de patrons consiste à la fois à déterminer l'élément sujet de la question et les différentes formulations possibles des réponses. Pour QALC et FRASQUES, le sujet de la question, dénommé focus, correspond à l'entité à propos de laquelle une information doit être donnée, et est constitué par l'un des groupes nominaux de la question. Comme dans [ITT 01], il constitue un point d'ancrage qui permet de trouver la réponse dans la mesure où la formulation de celle-ci reprendra souvent ce terme (ou un équivalent).

Pour conclure en se ramenant à la notion de terme, une réponse, lorsqu'elle donne la description d'une entité, est la recherche d'un terme complexe, pouvant figurer sous différentes formes, donc différentes variations, dans les textes. La reconnaissance de ce terme peut-être réalisée indépendamment de la question, c'est le cas pour les entités nommées, ou dans le contexte d'une question, et dans ce cas on extrait un terme correspondant souvent à la description d'un type sémantique général.

3. Conclusion

Lorsque l'on parle de terminologie, on évoque plutôt le choix de termes représentatifs et leur organisation en vue de décrire un domaine. La problématique de recherche d'information sur laquelle s'est centrée ce chapitre, qu'elle soit précise avec les systèmes de question-réponse ou thématique, concerne des approches applicables en domaine ouvert. Aussi, ce ne sont pas ces caractéristiques qui sont apparues comme primordiales.

Dans un système de question-réponse, le choix des termes ne s'effectue pas a priori, mais ceux-ci découlent des questions produites par l'utilisateur. Ce sont ces termes et surtout leurs variantes sous toutes leurs formes qui vont être recherchés et serviront de base pour sélectionner des passages courts répondant aux questions. Toutes les variations doivent être prises en compte : variation dérivationnelles, flexionnelles, sémantiques mais aussi syntaxiques ainsi que toutes les combinaisons autorisées de ces variations. Une autre application de ces mêmes caractéristiques terminologiques se retrouve pour sélectionner et extraire des réponses des textes. Il s'agit à nouveau de reconnaître des variantes, mais cette fois la définition du terme à rechercher est plus lâche, car elle correspond à un type sémantique. Aussi, des contraintes de dérivations sont-elles étudiées afin de limiter les candidats possibles : contraintes sémantiques et syntaxiques liées au contexte de la question. Ce deuxième point n'est généralement pas résolu par des techniques propres à la reconnaissance de termes, mais est traité par des approches relevant de l'analyse de phrases et de définition de patrons d'extraction, que l'on peut assimiler à des règles de réécritures de termes.

En ce qui concerne la segmentation thématique, ce sont les propriétés sur l'usage des termes dans un texte pour respecter un principe de cohésion qui sont exploitées. Ces termes, souvent des termes simples, sont sélectionnés en fonction de leur catégorie syntaxique et non parce qu'ils sont représentatifs d'un domaine ; rares sont les systèmes de segmentation thématique qui s'appuient sur une représentation des domaines, car cela signifie qu'ils doivent tous être connus et représentés a priori. La propriété des termes essentielle dans ce cadre concerne les relations de collocations existant entre eux. Ces relations sont trouvées dans des lexiques structurés, des thésaurus, ou non, réseau de co-occurrences. Par ailleurs, ce sont des mécanismes de répétition, de disposition et répartition des termes dans le texte qui permettent de définir et délimiter des segments. L'identification de thème relève plus de techniques relevant de la terminologie classique en ce qui concerne l'extraction de termes descriptifs pertinents.

4. Bibliographie

- [BIL 04] Bilotti M. W., Katz B., Lin J., “What Works Better for Question Answering: Stemming or Morphological Query Expansion?”, *SIGIR Workshop in Information Retrieval for Question Answering (SIGIR 2004)*, 2004
- [BUC 01] Buchholz S., Daelemans W., “Complex answers: a case study using a WWW question answering system”, *Journal of Natural Language Engineering*, 2001.
- [FEL 98] Fellbaum C., “*WordNet: An Electronic Lexical Database*”, MIT Press, Cambridge, MA, 1998.
- [FER 02a] Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., « Quand la réponse se trouve dans un grand corpus », *revue Ingénierie des Systèmes d'Information, n°spécial Recherche et filtrage d'information*, sous la direction de Catherine Berrut et Mohand Boughanem, Volume 7 - n°1-2, pp. 95-123, 2002
- [GRA 04] Grau B., 2004, chap. 10, « Systèmes de question-réponse », dans *Méthodes avancées pour les systèmes de recherche d'information*, dir. Ihadjadène, Hermes, pp. 189-218
- [GRA 05a] Grau B., Illouz I., Monceaux L., Paroubek P., Pons O., Robba I. et Vilnat A., « FRASQUES, le système du groupe LIR, LIMSI », *Atelier EQueR de TALN 05*, Dourdan, 2005.
- [GRA 05b] Grau B., Ligozat A.L., Robba I., Sialeu M., Vilnat A., “Term Translation Validation by Retrieving Bi-terms”, *Working Notes of CLEF Workshop, ECDL conference*, septembre 2005
- [GRI 95] Grishman R., Sundheim B., “Design of the MUC6 evaluation”, *Proceedings of MUC-6*, Morgan Kauffmann Publisher, Columbia, MD, 1995.
- [HAR 00] Harabagiu, S., Pasca, M., Maiorano, J., “Experiments with Open-Domain Textual Question Answering”. *Proceedings of Coling'2000*, Saarbrücken, Germany, 2000.
- [HOV 01] Hovy, E. , Hermjacob, U. & Lin C-Y., Ravichandran, D. , “Towards Semantics-Based Answer Pinpointing”, *DARPA Human Technology Conference (HLT)*, San Diego, 2001.
- [ITT 01] Ittycheriah, A., Franz, M. & Roukos, S., “IBM’s Statistical Question Answering System – TREC-10”. *Proceedings of the Text retrieval conference, TREC 10*, Gaithersburg, MD. NIST Eds., 2001
- [JAC 99] Jacquemin C., “Syntagmatic and paradigmatic representations of term variation”, *proceedings of ACL'99*, p. 341-348, 1999.
- [MOL 02] Moldovan, D., Harabagiu S., Girju R., Morrarescu P., Lacatusu F., Novishi A., Badulescu A., Bolohan O., “LCC Tools for Question Answering”, *Proceedings of the Text retrieval conference, TREC11*, Gaithersburg, MD. NIST Eds, 2002
- [MOL 03] Moldovan D., Clark C., , Harabagiu S., Maiorano S., “COGEX: A Logic Prover for Question Answering”, *proceedings of HLT-NAACL*, Edmonton, pp. 87-93, 2003
- [MOL 00] Mollá Aliod D., Schwitter R., Hess M., Fournier R., « Extrans, an answer extraction system », *Traitement Automatique des Langues*, vol. 41, n°2, p. 496-522, 2000.

- [MON 03] Monz C., « Document Retrieval in the Context of Question Answering », *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR-03)*, *Lecture Notes in Computer Science 2633*, Springer, pp. 571-579, 2003.
- [PRA 00] Prager J., Brown E., Radev D. R., Czuba K., “One Search Engine or two for Question-Answering”, *proceedings of TREC9*, Gaithersburg, MD, p 235-240, 2000.
- [RIN 03] Rinaldi F., Dowdall J., Hess J., Kaljurand K., Karlsson M., “The Role of Technical Terminology in Question Answering”, *Actes de la conférence TIA*, avril 2003
- [SUT 05] Sutcliffe R.F., Mulcahy M., Gabbay I., O’Gorman A., White K., Slattery D., “Cross-Language French-English Question Answering using the DLT System at CLEF 2005”, *Working Notes, CLEF Cross-Language Evaluation Forum*, 2004
- [TAN04] H. Tanev, M. Negri, B. Magnini et M. Kouylekov, 2004, The DIOGENE Question Answering System at CLEF-2004, *Working Notes de CLEF Cross-Language Evaluation Forum*, Bath UK, pp.325-333
- [TEL 03] Tellex, S., Katz B., Lin J., Marton G., Fernandes A., “Quantitative evaluation of passage retrieval algorithms for question answering”, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, 2003
- [WHI 04] White K., Sutcliffe R. F. E., “Seeking an Upper Bound to Sentence Level Retrieval in Question Answering”, *Workshop on Information Retrieval for Question Answering, SIGIR*, 2004