Détecter le potentiel d'ambiguïté d'une requête - le cas des recherches portant sur l'actualité

Fanny Lalleman, Cécile Fabre, Johannes Heinecke

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

Détecter le potentiel d'ambiguïté d'une requête – le cas des recherches portant sur l'actualité

Lalleman, Fanny[1 2], & Fabre, Cécile[1]

[1] CLLE, Université de Toulouse & CNRS
[2] Orange Labs
{fanny.lalleman et cecile.fabre}@univ-tlse2.fr

Heinecke, Johannes

Orange Labs
Johannes.heinecke@orange.com

1 Introduction : réexaminer la question de l'ambiguïté des requêtes[i]

Le traitement de l'ambiguïté est considéré comme un enjeu important pour l'amélioration des performances d'un système : l'apport d'une phase de désambiguïsation a été démontré à plusieurs reprises, par exemple par (Schütze et Pederson, 1995) ou plus récemment (Stokoe et al., 2003). De fait, de nombreux travaux ont été consacrés à cette question depuis les années 1990. Les solutions proposées pour la résoudre se sont d'abord focalisées sur le traitement de l'ambiguïté lexicale par recours à des dictionnaires ou, en recherche d'information multilingue, à des corpus alignés (Krovetz et Croft, 1992 ; Sanderson, 2000 ; Stokoe, 2005). Ces travaux étaient fondés sur une double hypothèse : la polysémie et l'homonymie des mots sont la source principale de l'ambiguïté, et les mots de la requête fournissent des indices contextuels pour permettre une désambiguïsation mutuelle. On se situerait donc dans le cas d'une tâche classique de désambiguïsation lexicale (*word sense disambiguation*). Cette conception de l'ambiguïté dans le contexte de la recherche d'information (RI) est aujourd'hui remise en cause.

Tout d'abord, l'inadéquation de dictionnaires génériques (typiquement Wordnet) pour le traitement sémantique de requêtes s'est avérée patente, pour plusieurs raisons : les emplois recensés dans des ressources lexicales externes peuvent ne pas être représentatifs de ceux qui apparaissent dans la base de textes ; en particulier, la présence massive d'entités nommées, non recensées dans ces dictionnaires, pose des problèmes d'ambiguïté spécifiques (Ehrmann, 2008 ; Sanderson, 2008). Ensuite, l'examen de requêtes issues de contextes opérationnels de RI – par opposition aux données d'évaluation artificielles longtemps pratiquées dans les campagnes TREC[ii] – a montré que beaucoup de requêtes sont constituées d'un seul mot et ne fournissent donc pas d'indices contextuels pour mener à bien cette tâche. Enfin, d'autres types d'ambiguïté ont été mis en évidence. Spärck-Jones et al. (2007) en identifient trois, relatifs au sens du mot, aux différents « aspects » de l'information considérée, au type de la requête :

> The ambiguity may be of the word sense, or of reference aspect. The request "house" may mean 'building', 'home', or 'firm', and the request "house prices" may refer to actual prices or economic factors. There is also the issue of request type e.g. topic vs home-page seeking (…)

Song et al. (2009) distinguent quant à eux des requêtes réellement ambiguës, dont les termes ont plusieurs sens (*Giant* réfère à un film ou une équipe sportive), et des requêtes larges qui couvrent plusieurs sous-thèmes (ex : *songs*).

La question de l'ambiguïté des requêtes s'est donc complexifiée. Et plus encore si on considère que l'intention de l'utilisateur est elle-même souvent vague : non seulement son expression linguistique l'est nécessairement, mais il peut être difficile de considérer qu'un besoin informationnel précis et parfaitement prédéterminé prévaut à l'expression de la requête. En réponse à ce problème, des traitements

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100286

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

non supervisés de l'ambiguïté sont proposés comme alternative aux traitements de désambiguïsation classique : des procédures de clusterisation des textes visent à faire émerger les différents emplois des mots de la requête représentés dans la base documentaire interrogée et à fournir à l'utilisateur les moyens de percevoir l'ambiguïté latente (Navarro et al., 2011 ; Zhai et al., 2003).

Dans ce contexte de redéfinition de la nature et du traitement de l'ambiguïté en RI, l'objectif du travail que nous présentons ici est d'examiner cette notion à travers l'étude des requêtes produites dans un système de RI, le site 2424actu.fr d'Orange, opérationnel du 1/10/2009 au 1/09/2011. Celui-ci vise le traitement d'une base de documents relatifs à l'actualité française, domaine particulièrement mouvant et par conséquent propice à l'examen de la question de l'ambiguïté. Nous cherchons à déterminer la nature de l'ambiguïté des requêtes en examinant les *logs* de requêtes disponibles et en les confrontant à différents indices contextuels qui enrichissent la perception de la variabilité sémantique des termes de la requête. Nous commençons par présenter les données (requêtes et bases de textes) sur lesquelles nous avons travaillé avant de détailler ces indices et de les appliquer à nos données.

2 Présentation des données issues d'un site d'actualités

Les données utilisées dans cette expérience proviennent d'une plateforme d'actualités développée dans un contexte industriel. Ce site permet de consulter l'actualité française en temps réel, et propose différents modes d'accès à l'information : l'utilisateur peut exprimer une requête en utilisant une barre de recherche traditionnelle, il peut également naviguer dans la base en utilisant des entrées thématiques ou en explorant des clusters de documents. Nous étudions ici les requêtes qui ont permis d'interroger la base de documents. Les données que nous avons constituées associent ces requêtes et les documents, en les organisant temporellement.

2.1 Le corpus : requêtes et base de textes

Le corpus de requêtes choisi pour cette étude correspond à une période temporelle de huit mois, de mai à décembre 2010. Pendant cette période, les utilisateurs ont produit près d'1/2 million de requêtes (487 231 exactement), contenant 30 668 requêtes différentes. Cet ensemble de requêtes est partitionné par mois. Dans le cadre de cette étude, nous nous sommes concentrés sur les 400 requêtes les plus souvent formulées en sélectionnant les 50 requêtes les plus fréquentes de chaque partition temporelle, dont le Tableau *1* montre un extrait.

Tableau 1 : Extrait des requêtes utilisateurs – année 2010

Mai	*sport foot* (8344) - *réforme retraites* (6459) - *festival de cannes* (6444) - *johnny hallyday* (4627) – *éruption islande* (4409)
Juin	*sport foot* (9884) – *apéro facebook* (4415) – *international* (4275) – *réforme retraites* (3840) – *afghanistan* (3323)
Juillet	*sport foot* (7068) - *réforme retraites* (3498) - *international* (3399) - *inondation var* (3387) - *afghanistan* (2984)
Août	*sport foot* (8049) - *incendie russie* (6570) - *haïti* (3125) - *afghanistan* (3109) - *sortie cinéma* (2927)
Sept	*grève rer* (2812) - *sport foot* (2600) - *afghanistan* (1023) - *delarue* (962) - *johnny hallyday* (888)
Oct	*grève* (14358) - *grève rer* (12179) - *mineurs chili fr* (5549) - *afghanistan* (3978) - *larry clark* (3801)
Nov	*grève* (9412) - *réforme retraites* (5650) - *éruption* (2573) - *international* (2389) - *afghanistan* (2355)

Déc	*wikileaks* (19796) - *côte d'ivoire* (8574) - *neiges* (5914) - *sortie cinéma* (5370) - *grève* (3223)

Dans le même temps, nous avons collecté un corpus de documents. Ces documents correspondent aux actualités disponibles sur le site de mai à décembre 2010, soit la base de textes vers laquelle les requêtes des utilisateurs ont été émises. Le corpus est donc partitionné de la même manière que les requêtes (Tableau *2*). Cette collection de documents, de nature exclusivement textuelle, est constituée de sources hétérogènes : audio ou journaux télévisés retranscrits, dépêches AFP, articles de journaux. Les documents proviennent des différents partenaires du site (AFP, Le Monde, Le Point, L'Express, France Télévision, Paris Match, etc.).

Tableau 2 : Corpus de documents (nombre de documents par corpus)

Mai 2010	Juin 2010	Juillet 2010	Aout 2010	Sept 2010	Oct 2010	Nov 2010	Déc 2010
23521	26782	15773	19543	17634	22822	16015	11096

2.2 Premiers éléments de caractérisation des requêtes

Les caractéristiques formelles des requêtes offrent d'emblée des éléments d'appréciation du potentiel d'ambiguïté d'une requête. Leur taille moyenne est un premier critère susceptible de nous renseigner sur le degré de spécificité de l'information exprimée. L'étude de Spink et al. (2002) a ainsi montré que la longueur des requêtes en anglais (provenant du moteur *Excite Web*) était de 2,6 mots et qu'elle variait peu dans le temps. Dans notre corpus, la longueur moyenne d'une requête est inférieure, puisqu'elle est de 1,73 mots. Les requêtes multi-mots sont donc minoritaires dans notre corpus. On trouve principalement :

- des termes nominaux complexes ou des noms propres composés (*assemblée nationale*, *côte d'ivoire*) ;

- des termes avec effacement du joncteur grammatical (*réforme retraites, grève RER*) ;

- un terme associé à une spécification d'ordre temporel ou spatial (*boue Hongrie*, *marée noire Etats-Unis*) ;

- des termes juxtaposés selon des associations sémantiques variées (*proxénétisme équipe de France*, *carla bruni photos*).

Le deuxième critère de caractérisation des requêtes susceptible de conditionner leur degré d'ambiguïté est la nature des termes qui les composent, et plus particulièrement la proportion de noms propres, ou plus largement d'entités nommées (désormais EN). Barr et al. (2008) signalent que des requêtes en anglais (issues du site *Yahoo!*) contiennent 40% de noms propres et 30% de noms communs. Plus spécifiquement, différentes études ont montré l'importance des lieux et des personnes dans l'expression de la requête : (Spink et al., 2004) décomptent de 11 à 17% de noms de personnes, (Gan et al., 2008) près de 38% de requêtes contenant des termes de type « géographique ». Cette tendance se retrouve plus fortement encore dans notre corpus de requêtes. Une annotation manuelle de 392 requêtes nous a permis de dénombrer 70% de requêtes contenant une entité nommée. Or le potentiel d'ambiguïté de ces unités a été étudié aussi bien en linguistique qu'en TAL, qu'il s'agisse d'homonymie dans le cas des noms de personne (Artiles et al., 2007) ou de métonymie dans le cas des toponymes (Lecolle, 2007 ; Ehrmann, 2008).

D'autres critères seraient également intéressants à mobiliser dans cette étude préliminaire, comme le degré de généricité des termes, la complexité morphologique, ou le nombre de sens recensés, indices utilisés par (Mothe et Tanguy, 2005) pour prédire la difficulté d'une requête. En l'état, on peut déjà concevoir que les caractéristiques de ces requêtes – souvent réduites à un seul mot et comportant beaucoup d'entités nommées – sont des indicateurs d'ambiguïté. S'ajoutent à cela les caractéristiques du

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

champ de la collection de documents : l'actualité, par nature changeante, caractérisée par la transformation du contexte référentiel, fait rapidement évoluer les informations associées aux termes de la recherche.

Dans l'étude, nous nous focalisons sur les requêtes comportant un seul terme, qu'il s'agisse d'un mot isolé ou d'un terme complexe correspondant au premier type de requêtes à plusieurs mots présenté précédemment (*assemblée nationale*, *côte d'ivoire*) : cela nous permet à la fois de nous concentrer sur un seul cas de figure, qui maximise les risques d'ambiguïté et qui est le plus fréquent dans notre corpus, et de simplifier la tâche d'analyse. L'ensemble de l'analyse qui suit porte donc sur 247 requêtes (soit 62% du corpus de requêtes au complet).

3 Analyse des requêtes

3.1 Les éléments de contextualisation de la recherche

Notre objectif étant de détecter l'ambiguïté d'une requête, nous cherchons à identifier les indices susceptibles de mettre au jour la diversité des sens ou, pour reprendre le terme de Spärck-Jones et al. (2007), des aspects référentiels qu'elle recouvre. Le contexte applicatif qui est le nôtre fournit des éléments de contextualisation de la recherche qui constituent un ensemble de points de vue complémentaires sur la requête (Hearst, 2009, Allan et al. 2003). Ces éléments peuvent avoir trois origines : ils peuvent concerner l'utilisateur (éléments de spécification de son besoin ou de son profil), les documents de la base (caractéristiques externes ou linguistiques), ou provenir de ressources externes de type lexicographique et encyclopédique susceptibles de fournir des connaissances générales relatives aux termes employés.

Du point de vue de l'utilisateur, nous manquons d'informations stratégiques concernant son profil, et même d'informations anonymisées permettant de l'identifier (adresse IP) : il n'est pas possible de savoir quelles requêtes ont été formulées par un même utilisateur. Nous pouvons néanmoins étudier d'un point de vue global les stratégies de reformulation qui ont été utilisées, ou exploiter des informations temporelles pour regrouper des requêtes proches. Ainsi il est intéressant de savoir que deux requêtes proches comme *grève* et *grève rer* coexistent au mois d'octobre 2010, l'une étant donc potentiellement la spécification de l'autre (sans que les deux requêtes aient été nécessairement formulées successivement par un même utilisateur).

Du côté de la base de textes, nous disposons de plusieurs éléments de contextualisation de la recherche – informations temporelles et caractéristiques portant sur la source dont est issu le document. Parmi celles-ci, nous utilisons le fait que les documents sont catégorisés sur le plan thématique, selon un classement hérité de l'AFP. On compte six thématiques : *ECONOMIE* (questions économiques), *INTERNATIONAL* (actualités hors de France), *SOCIETE*, *POLITIQUE*, *CULTURES* (musique, sciences, art, people) et *SPORT*. Par ailleurs, nous pouvons mobiliser des indices liées au co-texte d'apparition des termes de la requête (cooccurrences).

Nous avons enfin utilisé une ressource externe, Wikipédia, comme étalon pour estimer le potentiel d'ambiguïté des requêtes. L'utilisation de cette encyclopédie en ligne plutôt que d'un dictionnaire se justifie par la possibilité de pouvoir ainsi prendre en compte les entités nommées.

Outre ces trois dimensions d'étude, nous avons considéré la temporalité des documents et des requêtes comme une dimension essentielle de notre étude. L'actualité est rythmée par le temps. La recherche d'information dans ce contexte particulier hérite de cette contrainte, et permet d'accéder à des indices invisibles sans la trame temporelle. Il est donc apparu pertinent de regarder les requêtes utilisateurs du point de vue diachronique. On voit alors apparaître différents profils de requêtes qui témoignent de la fluctuation des thèmes d'actualité dans des temporalités même courtes et offrent un autre point de vue sur la diversité des facettes auxquelles une requête est susceptible de renvoyer.

Dans ce qui suit, les corpus de requêtes et de textes sont appréhendés à partir des différents modes de caractérisation que nous venons d'évoquer.

3.2 Catégorisation thématique

La catégorisation thématique des documents de la base fournit une grille macroscopique de découpage de l'information en six thèmes (*INTERNATIONAL, SOCIETE, POLITIQUE, CULTURES, ECONOMIE, SPORT*). Ce nombre de catégories est très limité. Elles ont néanmoins une pertinence car elles sont utilisées par l'AFP et les journalistes pour classer et typer les flux d'information. Elles représentent donc les domaines principaux de l'actualité. Cette information est disponible pour chaque texte de la base[iii]. Précisons qu'un texte est rattaché à une seule catégorie thématique. On dispose ainsi d'un mode de classification certes grossier mais adapté aux spécificités thématiques de la base de textes, qu'il est donc intéressant d'utiliser pour catégoriser les requêtes et observer leur distribution éventuelle sur plusieurs domaines, indice potentiel d'ambigüité ou tout au moins d'ambivalence.

Pour réaliser cette catégorisation, nous avons projeté les requêtes d'une période temporelle donnée sur la base textuelle correspondant à la même période. Dans la mesure où il s'agit de requêtes à un seul terme, la procédure d'identification du sous-ensemble de textes pertinent est extrêmement simple : nous retenons les textes qui contiennent la requête. Chaque fois que le terme de la requête apparaît dans un document, on incrémente le compteur de la catégorie thématique du document cible. A la fin du calcul, pour limiter les catégories résiduelles on ne retient que celles qui représentent plus de 10% des textes liés à la requête.

Les résultats montrent que 54% des requêtes sont mono-catégorielles : tous les textes auxquels elles sont associées relèvent de la même catégorie. 46% sont donc pluri-catégorielles, le nombre de catégories pouvant aller de 2 à 6. La répartition des requêtes qui donnent lieu à un classement thématique unique peut varier fortement par période (environ 67% pour le sous-corpus de décembre contre 27% dans le sous-corpus de mai).

La pluri-catégorisation fournit-elle un premier indice de l'ambiguïté de la requête ? La comparaison des deux types de requêtes ainsi dégagées (mono vs pluri-catégories) fournit quelques éléments d'analyse. On constate que les requêtes mono-catégorisées contiennent massivement des EN (80%) comme par exemple *miss france* ou *audrey pulvar* (catégorisées en *CULTURES*). Quelques requêtes contiennent des noms communs comme *neiges* (*SOCIETE*) ou *agriculture* (*ECONOMIQUE*). Ces deux derniers exemples montrent qu'on peut avoir affaire à des requêtes larges, sous-spécifiées, mais dont la portée s'inscrit dans une seule thématique. La part des EN dans les requêtes pluri-catégorisées est de 60%, c'est donc moins que dans le cas des requêtes mono-thématiques.

L'étude des requêtes pluri-catégorisées montre plusieurs cas d'ambiguïté. Pour une requête comme *royal*, le renvoi à plusieurs thématiques recouvre un cas clair d'homonymie. La requête est catégorisée comme suit: *INTERNATIONAL* (27), *POLITIQUE* (63), *SPORT* (40). La catégorie *POLITIQUE* renvoie à Ségolène Royal alors que la catégorie *SPORT* correspond à l'adjectif « royal ». D'autres requêtes pluri-catégorielles comme *obama* ou *sarkozy* correspondent à un autre cas de figure. Ainsi la requête *sarkozy* pointe vers trois catégories : *INTERNATIONAL* (196), *POLITIQUE* (544), *SOCIETE* (182). Ce type de requête désigne une personnalité présente sur différents sujets et qui endosse différents rôles : il intervient sur des problèmes de nature politique qui peuvent se poser hors de France (*INTERNATIONAL*), mais il est également présent sur des terrains sociétaux en lien avec une série de suspicions d'affaires ou de faits divers (*SOCIETE*). Enfin, on détecte également des requêtes pluri-catégorisées qui manifestent une ambiguïté référentielle réelle : il s'agit de requêtes comme *otages*, *éruption* ou *ministre*. En effet, ces requêtes montrent que l'utilisateur peut avoir tendance à désigner de façon très vague et implicite des événements qui dominent l'actualité au moment où il exprime la requête. A l'échelle de la base de textes, ce type de requête peut par contre être associé à une information très éparpillée. Par exemple, la requête *intempéries* paraît peu ambigüe dans un contexte d'actualité, pourtant la catégorisation faire ressortir deux thématiques : *INTERNATIONAL* et *SOCIETE,* signifiant que la France n'est pas la seule touchée (catégorie *SOCIETE*) mais que plusieurs endroits dans le monde ont été victimes d'intempéries (catégorie *INTERNATIONAL*).

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

Les thématiques offrent donc un premier point de vue sur le potentiel d'ambiguïté d'une requête, et révèlent la diversité des formes d'ambiguïté à l'œuvre, depuis une réelle ambiguïté lexicale jusqu'à une ambiguïté référentielle relative à une pluralité d'événements associés à un terme. Mais elles montrent des limites évidentes. Du fait du petit nombre de catégories thématiques très générales, le classement d'une requête dans une seule catégorie ne permet pas de déduire qu'elle est univoque. En particulier, une catégorie comme *INTERNATIONAL* recouvre une multitude de sujets et de thèmes dans l'actualité, elle permet surtout de localiser les *news* (hors de France en l'occurrence). La conséquence directe est qu'elle capte énormément de requêtes, et par exemple 40% des requêtes mono-catégorielles sont étiquetées *INTERNATIONAL*. Le problème est similaire pour la catégorie *CULTURES,* peu précise.

3.3 Consultation d'une ressource externe

L'utilisation d'une ressource externe pour rendre compte de l'ambiguïté des requêtes est la démarche la plus couramment employée pour apprécier l'ambiguïté des requêtes, comme nous l'avons signalé en introduction (Sanderson, 2008). Nous avons utilisé l'encyclopédie en ligne Wikipédia qui fournit une indication de la diversité des notions associées à un terme, à travers les pages dites d'homonymie ou de désambiguïsation. Ces pages répertorient les différents sujets et articles correspondant à une même forme. Par exemple, une des requêtes fréquentes dans notre corpus, *éruption*, renvoie vers une page qui recense les différents domaines d'emploi du mot *éruption* – éruption volcanique, cutanée, solaire – assortis de la mention du nom d'un groupe de musique et d'une chanson.

Nous avons procédé à une annotation manuelle de deux de nos sous-corpus de requêtes, soit 91 requêtes. Elle a consisté tout d'abord à déterminer si le terme de la requête était présent dans Wikipédia. C'est le cas de 67 % des requêtes. Sur les 61 requêtes trouvées, 35 pointent vers une seule entrée dans Wikipédia (57%), 26 correspondent à des entrées multiples (43%). S'agit-il d'un autre regard sur l'ambiguïté ou bien y a-t-il un recouvrement avec l'analyse précédente ?

L'utilisation de Wikipédia pour repérer l'ambiguïté montre immédiatement ses limites : l'ambiguïté qui est décrite dans l'encyclopédie ne correspond que rarement à une ambiguïté réelle dans la base de textes. En général, Wikipédia recense plus d'acceptions que celles qui apparaissent dans le contexte de l'actualité. C'est particulièrement le cas des noms de personne. Par exemple, si *Johnny Hallyday* pointe vers une page homonymique, c'est qu'il peut être un cascadeur ou un chanteur ; l'actualité ne connait bien sûr que le chanteur. La situation inverse est également fréquente : des termes univoques selon Wikipédia se déclinent selon plusieurs emplois ou sous-domaines dans la base de textes, comme illustré dans le paragraphe suivant dans le cas des noms de pays.

Nous avons confronté cette catégorisation de nature encyclopédique à la catégorisation thématique réalisée précédemment. On remarque que les proportions entre requêtes univoques et potentiellement ambiguës sont assez similaires : il existe 54 à 57% de requêtes pluri-catégorisées selon les deux points de vue. On constate néanmoins que le désaccord est important. En effet, en calculant l'accord inter-annotateurs (entre les deux types de classification) grâce à la mesure du coefficient du Kappa de Cohen[iv], nous obtenons un coefficient proche de 0,15 signifiant un accord très faible. A titre d'exemple, un désaccord intéressant concerne le terme *haïti*, univoque dans Wikipédia mais pluri-catégorisé par la classification thématique en *SOCIETE*, *INTERNATIONAL* et *CULTURES*. Dans l'actualité, *haïti* désigne le pays mais également une série d'évènements consécutifs au tremblement de terre de janvier 2010. De façon générale, les noms de pays ne sont pas ambigus au sens de Wikipédia, qui, en tant qu'encyclopédie, ne considère pas les effets métonymiques des toponymes ; ces termes sont par contre très mouvants au regard de la catégorisation thématique. De même, la requête *afghanistan* (une seule entrée dans Wikipédia) renvoie à la fois à la guerre en Afghanistan (catégorie *INTERNATIONAL*) et aux otages français (catégorie *SOCIETE*). La double catégorisation *SOCIETE* et *INTERNATIONAL*, bien que très grossière, en rend mieux compte. On peut enfin citer la requête *facebook* qui dans le cadre de l'actualité ne concerne pas le site (seul décrit dans Wikipédia), mais l'entreprise (*CULTURES*) ou les évènements liés au site comme les *apéros facebook* (*SOCIETE*). La nature de l'ambiguïté repérée est donc bien différente selon les méthodes

utilisées, et les exemples de décalage que nous avons examinés montrent que la catégorisation thématique capte mieux la réalité des emplois du corpus.

3.4 Cooccurrence

L'analyse que nous venons de faire de l'utilisation de Wikipédia illustre le décalage d'une ressource externe avec l'utilisation qui est faite des termes de la requête dans la base de textes interrogée. Cela confirme que c'est bien la base de textes qui fournit la grille la plus pertinente pour apprécier l'ambiguïté des requêtes. Nous avons poursuivi l'analyse en recherchant dans les textes eux-mêmes des éléments d'information sur le comportement sémantique des requêtes dans la base textuelle. Nous avons opté pour une procédure très simple d'analyse, consistant à examiner la cooccurrence des termes des requêtes dans les textes. Ce procédé est couramment utilisé pour étudier la variation sémantique en diachronie (Picton, 2009) ou la polysémie (Yarowsky, 1995 ; Turney, 2004 ; Audibert, 2003). Nous procédons ici à une analyse des cooccurrents de surface à l'aide de l'outil Antconc (Anthony, 2011) sur le corpus non lemmatisé, en utilisant la mesure d'information mutuelle (notée IM). Le contexte considéré est une fenêtre de 3 mots avant et après l'unité étudiée. L'intérêt de cette analyse distributionnelle dans notre contexte d'étude est de pouvoir identifier des liens forts entre la requête et des cooccurrents qui permettent de dégager un ou des comportements sémantiques de cette requête. Cette approche est néanmoins plus exploratoire : elle s'appuie sur un examen manuel et ne permet pas de déboucher comme dans les deux cas précédents sur un score d'ambiguïté potentielle.

Nous nous focalisons sur l'analyse des cooccurrents des requêtes qui ont été pluri-catégorisées par la catégorisation thématique, de manière à observer le lien entre la pluri-catégorisation des textes et la diversité des contextes d'apparition des termes. Cette analyse montre que la diversité thématique des requêtes se traduit effectivement dans les contextes d'apparition. C'est ce que montre l'exemple *royal* (Tableau 3). Le terme de la requête a pour cooccurrents des noms qui fonctionnent avec l'adjectif *royal* (*Stadium royal*, *Royal Navy, etc.*). Certains d'entre eux commencent par une majuscule, signe qu'ils appartiennent à une EN. *Ségolène* apparaît également comme cooccurrent, et les autres termes renvoient à des thématiques qui concernent la femme politique (*Tempête*, *Xynthia*).

Tableau 3 : Collocats du mot « royal »

« royal » : collocats mots-pleins (Fenêtre de 3 mots à gauche et à droite) [Corpus Juin]	
	Stadium (IM:14), Wever (IM:14), Tempête (IM:14), Navy (IM:14), Recours (IM:13), proposition (IM:13), hospital (IM:13), Bibliothèque (IM:13), Bart (IM:13), Xynthia (IM:12), Bangkok (IM:12) , Ségolène (IM:12) (….)

Un autre exemple intéressant concerne la requête *tabac*, qui a été catégorisée en ECONOMIE (27) et en SOCIETE (25). L'analyse de ses cooccurrents (Tableau 4) montre à la fois la dimension marchande du terme (ECONOMIE) : *débits*, *transformation*, *multinationale* et la dimension santé publique (SOCIETE) avec *sensibiliser* et *hypertension*.

Tableau 4 : Collocats du mot « tabac »

« tabac » : collocats mots-pleins (Fenêtre de 3 mots à gauche et à droite) [Corpus Nov]	
	Barman (IM:16), débits (IM:16), coopérative (IM:15), transformation (IM:15), Sensibiliser (IM:14), kilo (IM:14), Dordogne (IM:13), plantations (IM:13), multinationales (IM:13), hypertension (IM:13)

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

Enfin, la requête *ministre* (Tableau 5), pluri-catégorielle et ambiguë du point de vue référentiel, est fortement associée à des cooccurrents qui spécifient le terme *ministre* comme *vietnamien*, *yéménite*, ou de nombreux noms propres tels que *Cowen*, *Ouattara* et *Gillard*.

Tableau 5 : Collocats du mot « ministre »

« ministre » : collocats mots-pleins (Fenêtre de 3 mots à gauche et à droite) [Corpus Nov]	
	Cowen (IM:15), Socrates (IM:14), Brian (IM:13), Ouattara (IM:11), Gillard (IM:11), dominicain (IM:11), Alassane (IM:11), Hubert (IM: 11,64), hyper (IM:11), vietnamien (IM : 11), promue (IM: 11), dédouaner (IM: 11), relance (IM:11), Manmohan (IM:11), yéménite (IM: 10)

Grâce à la cooccurrence, nous accédons aux évènements décrits dans les documents de notre base dessinant un contexte précis d'utilisation des termes de la requête dans un texte d'actualité. Ils nous informent sur la diversité des facettes référentielles attachées à la requête.

3.5 Reformulation

Bien que les informations concernant l'utilisateur soient très incomplètes, les requêtes sont en elles-mêmes des traces de la diversité des modes d'expression qui sont utilisés au fil des recherches (Song et al., 2009 ; Jansen et al., 2009). La façon dont une requête a été reformulée, ou en tous cas (en l'absence d'information sur l'identité de l'utilisateur), les différentes manières dont un terme a été utilisé dans une série de requêtes, fournissent des indices sur la diversité des points de vue exprimés par l'utilisateur. La version « étendue » d'une requête courte peut nous informer sur les types de spécification possible d'une requête sémantiquement large. En observant les requêtes pluri-catégorisées telles que *royal*, *sarkozy* ou *ministre* on voit se dessiner différents types d'extensions de ces requêtes. Il s'agit bien entendu dans tous les cas de formulations moins fréquentes puisque plus spécifiques.

L'étude de la reformulation des requêtes paraît donc être une piste intéressante pour discriminer les requêtes pluri-catégorisées par la catégorisation thématique. En effet, les observations laissent apparaître deux modes de reformulation différents.

Le premier type de reformulation produit des extensions qui complètent la requête et qui entretiennent une relation de spécification avec celle-ci. Cette procédure permet d'identifier un sens et de lever une ambiguïté lexicale ou référentielle. Par exemple, pour la requête *royal*, deux sens apparaissent : Ségolène Royal (*ségolène royal*, *segolène royal tf1*) et l'adjectif « royal » (*mariage royal*, *royal emirat*, *mariage royal suède*). Les requêtes qui manifestent une ambiguïté référentielle comme *ministre* ont des extensions qui spécifient le mot ambigu. Ainsi, les mots associés à la requête permettent de créer une unité où le référent est identifié comme *juppé ministre*, *borloo premier ministre*, *démission 1er ministre belge*. Le deuxième type de reformulation consiste principalement à ajouter un autre terme juxtaposé permettant de resserrer une thématique ou de préciser une requête « large ». Par exemple, une requête pluri-catégorisée comme *sarkozy* présente les deux types de reformulations :

- soit on retrouve un fonctionnement similaire à une requête comme *royal* avec *sarkzoy/ guillaume sarkozy*, où l'on voit apparaître un phénomène d'homonymie sur la requête *sarkozy*.
- soit un autre terme juxtaposé est ajouté, qu'il s'agisse d'un autre nom propre (*sarkozy merkel*) ou d'un nom commun (*sarkozy criminalité*), ce qui permet de réduire la recherche d'information à un seul aspect de la requête *sarkozy*.

L'analyse des reformulations opérées par les utilisateurs montre la complexité de certaines requêtes qui peuvent potentiellement manifester plusieurs types d'ambiguïté. Cependant, cet indice doit être exploité

avec précaution, dans la mesure où les spécifications sont peu fréquentes, et n'épuisent certainement pas l'éventail des sens que recouvre le terme.

3.6 Diachronie

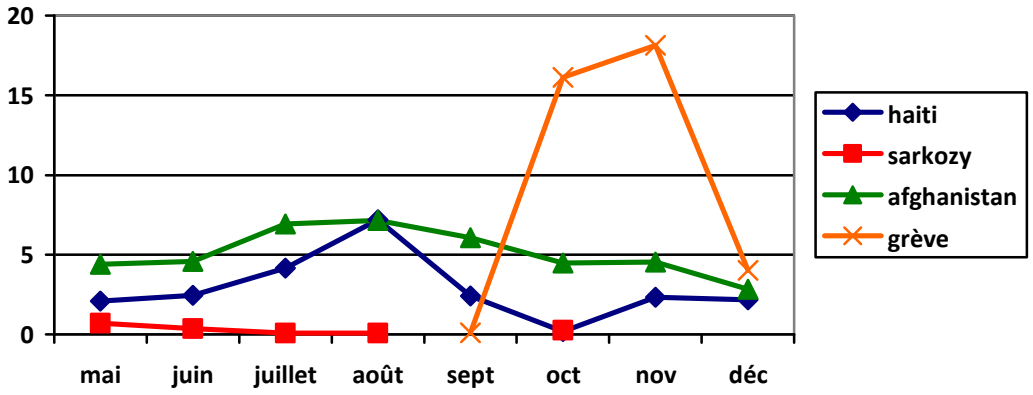
La dernière dimension que nous étudions est la dimension temporelle. Elle va nous amener à croiser la plupart des dimensions d'analyse que nous venons de présenter. Les requêtes n'ont pas toutes le même comportement dans le temps. Nous distinguons deux types de requêtes :

- Les requêtes « durables » : ce sont des requêtes qui apparaissent fréquemment tout au long des 8 mois d'actualité que couvre notre corpus. Sur le Graphique 1, ce cas est illustré par les requêtes *haïti* et *afghanistan*.
- Les requêtes « ponctuelles » : leur durée de vie est plus réduite, elles présentent des fluctuations beaucoup plus marquées selon les événements qui sont apparus sur cette période. C'est le cas des requêtes *sarkozy* ou *grève* sur ce même graphique.

On peut faire l'hypothèse que cette différence de comportement a un impact sur l'ambiguïté potentielle. Nous pouvons par exemple supposer que des requêtes très ponctuelles correspondront à un événement spécifique et seront plus univoques.

Notre attention se porte sur les requêtes « durables » : exprimées de manière très récurrente par les utilisateurs pour accéder à l'information, sont-elles pour autant porteuses du même type d'information au fil du temps ? Nous nous concentrons ici sur l'exemple d'une requête très fréquente sur toute la période, *haïti*. Nous croisons la dimension temporelle avec les autres points de vue dont nous disposons : catégorisation thématique, cooccurrence et reformulations. Rappelons que ce terme est univoque selon Wikipédia (c'est un pays).

Graphique 1 : Fréquences relatives d'apparition en %



La requête *haïti* a un comportement intéressant vis-à-vis de la catégorisation thématique. Cette requête est la plupart du temps catégorisée en *INTERNATIONAL* (au mois de juin, août, octobre et novembre). Malgré une tendance forte à la mono-catégorisation, on observe plusieurs changements:

- Au mois de juin, la requête est catégorisée en *CULTURES* et en *SPORT*.
- Au mois de décembre, la requête est catégorisée à la fois en *SOCIETE* et en *INTERNATIONAL*.

Nous savons également que la thématique *INTERNATIONAL* est très englobante. On peut donc supposer que cette requête a une capacité de variation forte.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

Dans un deuxième temps, nous observons si cette variation mise en évidence par la catégorisation thématique se retrouve dans les documents de la base textuelle, pour cela nous réalisons une analyse des cooccurrents fréquents et fortement liés au terme *haïti* sur plusieurs périodes temporelles (mai, août, octobre, novembre, décembre). L'hypothèse est que la variation se traduit par la présence de cooccurrents différents selon la période temporelle.

Dans le Tableau 6, sont présentés les cooccurrents les plus fréquents de *haïti* à différentes époques temporelles. Nous constatons qu'ils sont effectivement très différents. Nous discernons plusieurs évènements importants comme l'annonce des élections au mois d'août 2010 : *élection*, *invalidée, rappeur* (candidature du rappeur Wyclef Jean). Au mois d'octobre c'est l'épidémie de choléra qui apparaît (*éradiqué*, *ralentie*, *kits*, *choléra*) suivie de l'ouragan Tomas au mois de novembre. Le seul cooccurrent de *haïti* présent sur plusieurs mois est *Minustah* qui désigne la mission des Nations Unies pour la stabilisation du pays, qui constitue de fait un arrière-plan stable. On détecte donc une variation des évènements associés au terme *haïti*, induite par cette actualité mouvementée.

Tableau 6 : « haïti », cooccurrents en diachronie.

« haïti » : 5 plus fréquents collocats mots-pleins (Fenêtre de 3 mots à gauche et à droite)	
Mai	Crowe (IM:12), Etats (IM:11), unis (IM: 9), Forte (Im: 9), Russell (IM: 8)
Août	quittait (IM:14), invalidée (IM:13), Minustah (IM:13), élection (IM:13), rappeur (IM 13)
Oct	Éradiqué (IM:13), ralentie (IM:13), kits (IM:13), adoptions (IM: 13), choléra (IM:12)
Nov	Tomas (IM:17), Ouragan (IM:15), Casques (IM:14), Minustah (IM:14), évêques (IM: 14)
Dec	adoptés (IM:14), passeraient (IM:14), Minustah (IM:13), secouent (IM:13), recomptage (IM:13)

L'analyse des reformulations de la requête *haïti* tout au long du corpus (Tableau 7) vient conforter les observations faites sur les cooccurrents de *haïti* en contexte. En effet, les utilisateurs procèdent à des reformulations qui permettent d'identifier des thèmes associés à l'actualité d'Haïti. On retrouve en particulier les thématiques des élections et de l'adoption.

Tableau 7 : Reformulations de la requête « haïti »

Mai	*haïti adoption (26) - haïti séisme (2) - adoption haïti (1) - haïti cacao (1)*
Juin	*haïti adoption (26) - haïti actu (2) - images du séisme en haïti (2)*
Août	*haïti adoption (8) - haïti 17 (3) - haïti aujourd'hui (2) - officiellement candidat à haïti (2) - tremblement de terre à haïti (2) – haïti séisme (2) - élection en haïti (1)*
Sept	*haïti adoption (4) - radio d'haïti (2) - haïti reconstruction (1)*
Oct	*haïti adoption (11) - élection haïti (4) - adoption haïti (2)*
Nov	*haïti adoption (16) - haïti choléra (9) - bilan d'octobre à aujourd'hui haïti choléra (4) - haïti 18 novembre (2) - haïti association l'ile aux enfants (2)*

Déc	*haïti adoption (68) - haïti élection (11) - haïti actualités 11 12 2010 (3) - haïti choléra (3) - haïti jude célestin (3) - élections haïti (2)*

L'analyse que nous avons effectuée sur la requête *haïti* fait intervenir trois faisceaux d'information : la catégorisation thématique, la distribution de la requête dans les textes et les reformulations de cette requête. La catégorisation nous a surtout montré que les thématiques liées à cette requête pouvaient évoluer, malgré un ancrage fort dans la catégorie INTERNATIONAL. Nous savons que cette thématique est difficile à interpréter et qu'elle cache potentiellement une diversité plus grande. L'analyse des cooccurrents de la requête *haïti* dans les documents et les processus de reformulation ont effectivement confirmé cette diversité. En effet, on peut observer au moins deux emplois possibles du mot *haïti* : comme référence au tremblement de terre ou dans un sens locatif. *Haïti* semble manifester une polyvalence de base (le lieu, le pays et les habitants) et des variations contextuelles propres à l'actualité. Ainsi lorsque le pays Haïti a été touché par une épidémie de choléra, les cooccurrents de mot « Haïti » dans le document ont changé. La requête a pris une signification différente, touchée par une variation contextuelle.

Cet exemple manifeste une source de variation particulière, décrite par (Lecolle, 2007) sous le terme de « polysignifiance ». Etudiée dans le cadre des noms de lieux par Lecolle (2007), la polysignifiance renvoie au fait qu'un nom de lieu habité peut présenter des valeurs sémantico-référentielles différentes, désignant à la fois au lieu, mais aussi les habitants et l'institution qui le gouverne. Ces glissements sémantiques peuvent amener certains noms de lieu à revêtir un sens évènementiel comme par exemple *Outreau*, étudié par Lecolle, qui a pris la valeur d'erreur judiciaire en supplément de sa valeur locative, ou *Tchernobyl*, qui désigne désormais une catastrophe nucléaire. Cette malléabilité du nom de lieu décrite par (Lecolle, 2007) ouvre une gamme large de possibilités, et suscite des problèmes évidents si les différentes valeurs ne peuvent être discriminées et apparaissent dans des contextes identiques. La polysignifiance ne peut pas être appréhendée par le biais de ressources lexicographiques ou de bases de connaissances qui ne rendent pas compte de ces différentes valeurs, la fonction de localisation étant généralement la seule à être retenue dans le cas des noms de lieu.

4 Conclusion

Dans cette étude, nous avons examiné les formes que prend l'ambigüité dans un contexte opérationnel de recherche d'information, en considérant les termes employés dans les requêtes à un seul mot, majoritaires dans le corpus que nous avons constitué. Nous avons établi et testé un ensemble de critères permettant d'apprécier l'ambivalence référentielle de ces termes en croisant diverses sources d'informations, en l'absence d'indices contextuels contenus dans la requête elle-même. Ces différents indices nous ont amené à étudier l'ambiguïté des requêtes dans toute sa complexité, en tenant compte à la fois d'informations issues de la base de textes (catégorisation, cooccurrences) et des trajets de recherche des utilisateurs (reformulations).

Nous avons montré les limites de l'utilisation d'une ressource externe de type encyclopédique : celle-ci a tendance à surestimer la dimension homonymique de certains termes qui sont monoréférentiels dans la base de textes, et également à sous-estimer la polysémie de certaines unités, particulièrement les toponymes, qui se sont avérés particulièrement mouvants dans le corpus. Le recours à d'autres moyens d'observation, adaptés cette fois aux particularités du contexte de recherche d'information qui est analysé, constitue une piste plus intéressante d'analyse. L'existence d'une catégorisation thématique des textes fournit un premier critère de tri facile à mettre en œuvre, mais son pouvoir de discrimination est limité en l'état, du fait de catégories très générales. L'observation conjointe des cooccurrents des termes dans les textes et de leurs reformulations fait en effet ressortir de nombreux sujets ou évènements dont la granularité est beaucoup plus fine. Enfin, la prise en compte de ces différents niveaux d'analyse dans une perspective diachronique révèle le caractère mouvant de l'information évènementielle attachée au même terme sur une période de quelques mois.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100286

Ces différents indices nous ont permis de mettre au jour plusieurs sources d'ambiguïté: la polysémie (*tabac*) et l'homonymie (*royal*), classiquement étudiées dans les travaux en recherche d'information, se combinent avec d'autres formes d'ambiguïté. Ainsi, on trouve parmi les requêtes fréquentes des termes auxquels manque une spécification (*éruption*, *otages* ou *ministre*), probablement parce qu'elle est considérée par l'utilisateur comme suffisamment saillante dans l'actualité pour ne pas nécessiter d'être mentionnée. Dans ce cas, l'ambiguïté surgit de la confrontation avec les textes, dans lesquels ces termes peuvent recevoir des spécifications diverses. Mais c'est surtout l'utilisation prédominante de noms propres qui constitue une source d'ambivalence majeure sur le plan référentiel. Des requêtes « larges » sont constituées de termes qui présentent différentes facettes, qu'il s'agisse de noms de personne occupant certains rôles (*sarkozy*, *obama*) ou de noms de lieux (*haïti*, *afghanistan*), dont on a vu la propension à évoquer des évènements de nature diverse au fil du temps. Ces dimensions de l'ambiguïté en recherche d'information ne sont pas prises en compte dans les ressources externes susceptibles d'être utilisées.

Dans cette perspective, l'enjeu sur le plan applicatif semble moins de résoudre l'ambiguïté potentielle de la requête (comme c'est le cas lorsqu'on a affaire à une ambiguïté lexicale classique) que de trouver des modalités pour présenter à l'utilisateur les différentes facettes du terme qu'il emploie. La présentation de reformulations ou de cooccurrents fréquents, ou le classement des textes par catégorie thématique constituent des pistes possibles. Notre prochain objectif est d'évaluer l'intérêt sur le plan ergonomique de ce dernier critère.

Références bibliographiques

Allan, J. et al. (2003). Challenges in Information Retrieval and Language Modeling. *SIGIR Forum*, *vol. 37, 1*, 31–47.

Anthony, L. (2011). AntConc (Version 3.2.2) [Computer Software]. Tokyo, Japan :Waseda University. Disponible à partir de http://www.antlab.sci.waseda.ac.jp/

Artiles, J., Gonzalo, J. and Sekine, S. (2007). The semeval-2007 weps evaluation : Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (Semeval-2007)*, 64–69.

Audibert, L. (2003). Etude des critères de désambigüisation sémantique automatique : résultats sur les cooccurrences. In *Actes de la 10ème conférence Traitement Automatique des Langues Naturelles (TALN 2003)*, 35–44.

Barr, C., Jones, R. and Regelson, M. (2008). The linguistic structure of English web-search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 1021–1030.

Erhmann, M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.* Thèse de doctorat, Paris VII.

Gan, Q., Attenberg, J., Markowetz, A. and Suel, T. (2008). Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web, LOCWEB '08*, 49–56.

Hearst, M. A. (2009). *Search User Interfaces.* Cambridge University Press.

Jansen, B., Booth, D., Spink A. (2009). Patterns of query reformulation during web searching. *Journal of the American Society for Information Science and Information Technology, 60(7)*, 1358–1371.

Krovetz, R. and Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst., 10*, 115–41.

Lecolle, M. (2007). Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas Outreau. *Corpus, (6),* 101–125.

Mothe, J. et Tanguy, L. (2005). Linguistic features to predict query difficulty, *ACM SIGIR Workshop: Predicting Query Difficulty - Methods and Applications*, Salvador - Bahia – Brazil.

Navarro, E., Chudy, Y., Gaume, B., Cabanac, G., Pinel-Sauvagnat, K. (2011). Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? In *actes de la conférence CORIA*, Avignon, 25–40.

Picton, A. (2009). *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial.* Thèse de doctorat en Sciences du Langage, Université Toulouse 2.

Sanderson, M. (2000). Retrieving with good sense. *Information Retrieval, 2(1)*, 45–65.

Sanderson, M. (2008). Ambiguous queries: test collections need more sense. In *SIGIR*, 499–506.

Schutze, H. and Pedersen, J.O. (1995). Information retrieval based on word senses. In *Symposium on Document Analysis and Information Retrieval*.

Song, R., Luo, Z., Nie, J.-Y., Yu, Y. and Hon, H.-W. (2009). Identification of ambiguous queries in web search. *Information Processing and Management, 45(2),* 216–229.

Spärck-Jones, K., Robertson, S. E. and Sanderson, M. (2007). Ambiguous requests: implications for retrieval tests, systems and theories. *SIGIR Forum 41, 2*, 8–17.

Spink, A., Jansen, B. J., Wolfram, D., and Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer, 35(3),*107–109.

Spink, A., Jansen, B. J. and Pedersen, J. (2004). Searching for people on web search engine. *Journal of Documentation, 60(3)*, 266–278.

Stokoe, C., Oakes, M. P. and Tait, J. (2003). Word sense disambiguation in information retrieval revisited. In *SIGIR*, 59–166.

Stokoe, C. (2005). Automated word sense disambiguation for web information retrieval. In *Proceedings of SIGIR Forum*, 68.

Turney, P. (2004). Word Sense Disambiguation by Web Mining for Word Co-Occurrence Probabilities. *In Actes de la 3eme conference internationale « Evaluation of Systems for the Semantic Analysis of Text » (SENSEVAL-3 2004)*, Barcelone, Espagne, 25–26 juillet 2004.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *In Proceedings of Association for Computational Linguistics (ACL 1995)* (Ed.), Cambridge, MA, 189–196.

Zhai, C. X., Cohen, W. W. and Lafferty, J. (2003). Beyond independent relevance : methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '03*, 10–17.

---

[i] Nous remercions Michelle Lecolle (Université de Metz & CELTED) pour les suggestions qu'elle nous a apportées à la lecture de l'article.

[ii] *Text Retrieval Conferences* : http://trec.nist.gov/.

[iii] La catégorisation est disponible soit parce qu'elle a été annotée manuellement par des professionnels sur certains types de textes (dépêches AFP, presse écrite), soit parce qu'elle a été calculée par des procédés de clusterisation des documents.

[iv] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educ. Psychol. Meas*., 20, 27-46.