# La détection automatique des citations et des locuteurs dans les textes informatifs

**Emmanuel Giguet**
Lattice CNRS – UMR 8094
Ecole Normale Supérieure
1 rue Maurice Arnoux
F92120 Montrouge

**Nadine Lucas**
Greyc CNRS – UMR 6072
Université de Caen
Bd Maréchal Juin
F14032 Caen

## Introduction

La détection automatique des citations et des locuteurs est une problématique qui suscite un intérêt tout particulier chez les industriels et institutionnels soucieux des divers courants d'opinion. Dans le milieu de la veille économique et de l'intelligence stratégique par exemple, lorsqu'une crise surgit, il est essentiel de pouvoir répondre très rapidement à des questions telles que "*Qui sont les principaux leaders d'opinion? ", "Quand se sont-ils exprimés ?" "Quels ont été les propos tenus ?", "Quels sont les journaux ayant relayé ces propos ?".*

Sans répondre à la question spécifique de la surveillance des courants d'opinion, le traitement automatique des langues propose dans un cadre général des réalisations concrètes dont l'objet est la détection des citations (par exemple, Mourad & Minel, 2000). Le problème est alors formulé comme l'identification d'une source d'information et d'un discours rapporté. A titre d'illustration, dans l'extrait journalistique suivant :

Exemple 1. *Dans un communiqué commun, Jean Glavany, Bernard Kouchner et François Patriat, qui disent avoir pris connaissance du rapport de la commission d'enquête, soulignent jeudi que "la sécurité sanitaire des aliments, la lutte contre l'encéphalopathie spongiforme bovine (ESB) et sa prévention, constituent une priorité majeure de l'action gouvernementale".*

une telle application est censée identifier *Jean Glavany, Bernard Kouchner et François Patriat* comme source d'information et *"la sécurité sanitaire des aliments, la lutte contre l'encéphalopathie spongiforme bovine (ESB) et sa prévention, constituent une priorité majeure de l'action gouvernementale"* comme discours rapporté.

Dans cet article, nous commencerons par présenter l'état de l'art, qui consiste en la mise en œuvre d'une solution principalement lexicale, avec un cadre de travail implicite qui est la phrase. Nous proposerons ensuite une alternative, appelée stratégie syntaxique, dans laquelle nous utilisons non plus des lexiques mais des constantes morpho-syntaxiques intra-phrastiques. Enfin, nous présenterons une version plus élaborée, appelée stratégie trans-phrastique, intégrant des constantes textuelles pour améliorer la détection. Nous présenterons notamment des premiers résultats sur une résolution non lexicale de la co-référence.

# Contexte

En informatique linguistique et spécifiquement en extraction automatique de contenu, la question "qui dit quoi" revient dans de très nombreuses applications. C'est ainsi en effet qu'est exprimée la problématique du discours rapporté, faute de métalangage linguistique, dans la communauté concernée, les utilisateurs et concepteurs de logiciels.

La détection automatique s'appuie sur des formes et exploite la typographie. Le problème en informatique linguistique est ordonné par degré de difficulté croissante, en fonction des éléments mémorisés et reconnaissables : discours rapporté marqué par des guillemets et source identifiable ; discours rapporté marqué par des guillemets et source non identifiable ; discours rapporté non marqué et source non identifiable ; enfin, la résolution de la coréférence, considérée comme difficile. La source est assimilée à un nom propre (entité nommée) et celui-ci à un mot ou un couple de mots capitalisé (avec initiale en majuscule) ou encore à une entité référencée et mémorisée. On est donc loin des concepts linguistiques de *discours* (ou *style) direct* et *indirect*, basés sur l'observation plus fine des traits de textualité ou de reformulation, par le biais des pronoms notamment (voir par exemple Banfield, 1973 ; Coulmas, 1986 ; Authier-Revuz 1992-93 ; Dendale, 2001).

## 1.1 La stratégie lexicale

La stratégie la plus communément employée repose sur l'appel à des listes en mémoire pour identifier des éléments au sein d'un cadre de travail implicite qui est la phrase. Les ressources sont principalement et massivement lexicales et vont permettre un ciblage très précis et très

directif de la source et du verbe. Dans notre exemple, ci-dessous annoté manuellement, l'appel aux lexiques permettrait la détection des noms propres (en gras), des verbes citatifs (en gras souligné), et du discours rapporté parce qu'il est entre guillemets (en gras et en droit).

Exemple 1'. *Dans un communiqué commun,* ***Jean Glavany****,* ***Bernard Kouchner*** *et* ***François Patriat****, qui* ***disent*** *avoir pris connaissance du rapport de la commission d'enquête,* ***soulignent*** *jeudi que* **''la sécurité sanitaire des aliments, la lutte contre l'encéphalopathie spongiforme bovine (ESB) et sa prévention, constituent une priorité majeure de l'action gouvernementale''**.

On suppose ici que les verbes *dire* et *souligner* ont été tous deux répertoriés comme verbes citatifs. Dans le cas contraire, au moins une source ne serait pas détectée. L'algorithme sous-jacent à de telles approches peut se décomposer en deux étapes : la catégorisation des mots ou groupes de mots de la phrase présents dans les lexiques ; l'identification des motifs stockés à partir des mots catégorisés.

Les écueils rencontrés sont bien connus. Il s'agit d'une part de bruit (des citations détectées à tort) et de silence (des citations non identifiées), d'autre part d'erreurs de segmentation ou de caractérisation des éléments, en particulier de la difficulté à établir la fin du discours rapporté indirect (Charolles, 2000). Les sources d'amélioration classiquement envisagées sont l'ajout dans le lexique de nouvelles entrées, pour les entités sources (e.g., *Putin*) ou pour les titres et fonctions, ainsi que l'ajout de verbes citatifs (e.g., *to caution*) (voir par exemple Mourad, 2000).

## 1.2 Discussion

Bien que la stratégie lexicale permette la reconnaissance d'un grand nombre de citations, la grande variabilité formelle des sources (les entités nommées, titres et fonctions) et des verbes citatifs est en quelque sorte contradictoire avec l'objectif d'un système de détection fiable et robuste des citations, pouvant s'adapter à l'occurrence de verbes non attendus, tels que *stigmatiser, s'alarmer*.

Dans l'approche que nous allons maintenant présenter, nous cherchons à caractériser ces formes extrêmement variables, non pas par la gestion d'un lexique, mais par la recherche de

constantes. En effet, la diversité des formes peut être grande, la liste des verbes introduisant le discours rapporté n'est jamais close. De plus, ajouter des verbes présente des inconvénients, car on relèvera beaucoup de verbes de mouvement comme *jeter* ou *lancer*, ou de verbes d'expression émotive comme *ricaner*, *s'énerver*, qui peuvent créer du bruit, il faudra donc éliminer des sources de confusion. Sur le plan pratique, la recherche d'un mot a un certain coût, or on constate que les verbes-outils comme *dire* sont très fréquents, il est donc légitime de les rechercher. *A contrario* les verbes rares qui vont alourdir la liste des items recherchés seront peu productifs dans la détection de citations. Il est en fait irréaliste de considérer qu'il suffirait d'ajouter toujours plus de "verbes citatifs" ou de noms propres pour une détection fiable, à partir de listes dont l'exhaustivité s'avère être illusoire.

Un autre argument en défaveur de la stratégie lexicale est le coût du passage dans une autre langue, car il faudra constituer des listes qui peuvent être très conséquentes. Enfin, les considérations sur la maintenance et la supervision du système automatique militent en faveur de la conception d’un système autonome. On ne peut en effet dédier un analyste pour superviser les résultats fournis par un logiciel : autant il est facile de s'apercevoir ponctuellement qu'un passage de texte surligné ou extrait en tant que citation est mal analysé, autant il est difficile sans tout lire de repérer les oublis (on mesure sous le terme de silence les items présents non détectés). Or les quantités de texte traité défient justement les capacités de lecture d'un individu, et il ne s'agirait jamais que de corrections.

Dans les conditions réelles, cependant, la stratégie lexicale donne des résultats jugés corrects, pour ce qui est notamment de la "veille d'image". Il s'agit de repérer la reprise par les médias des propos tenus par une entreprise, et de relever ceux tenus par les concurrents. On pourra sans grand risque utiliser un lexique clos, réputé de couverture suffisante, pour traiter d'un corpus dont on connaît les caractéristiques stylistiques ainsi que le domaine. Les résultats de la détection automatique à base lexicale sont satisfaisants dans ces conditions.

## La stratégie syntaxique

Au vu des critiques formulées à l'égard de la stratégie purement lexicale, et dans la perspective d'une veille non ciblée dite "veille stratégique", nous avons défini comme objectif la mise en œuvre d'un système de détection léger basé sur des marques linguistiques de surface. Les études linguistiques que nous avons menées pour réaliser notre logiciel d'annotation ont porté sur un corpus de dépêches en anglais issues de diverses agences de presse, d'une part, et d'articles en français de divers journaux, d'autre part.

La stratégie que nous proposons a pour ambition de caractériser les citations à l'aide d'invariants, qui jouent le même rôle que des constantes dans une résolution d'équations. Il s'agit de localiser, sans recourir à des listes exhaustives de formes, les trois éléments inconnus qui nous intéressent : la source, le discours rapporté et un *relateur*. Le *relateur* est défini comme le segment établissant la relation entre la source et le discours rapporté (par exemple "*soulignent… que*" dans notre première illustration exemple 1). Il peut être verbal, conjonctif (verbe et conjonction) ou prépositionnel.

Selon les principes généraux de notre méthode, la résolution peut se faire à deux niveaux, syntagmatique ou paradigmatique, c'est-à-dire qu'une résolution immédiate de la séquence est basée sur des critères observables dans un cadre donné, le contexte minimal, et qu'une résolution différée utilise un contexte plus large. Pour la détection de citations, le contexte minimal est celui de la phrase, aussi les critères "internes" sont intra-phrastiques et les critères "externes" sont dits trans-phrastiques.

### 1.3 Critères phrastiques

La séquence canonique de notre modèle de la citation est (**source** + ***relateur*** + *discours rapporté*).

L'analyse distributionnelle menée dans le cadre phrastique a montré qu'il était possible d'obtenir une caractérisation pertinente basée sur trois classes : des indices typographiques (*i.e.*, ponctuations, capitalisation) ; des indices morpho-syntaxiques (*i.e.*, morphèmes grammaticaux) et enfin des indices positionnels (*i.e.*, début, fin). Dans l'exemple 1 repris ci-

dessous, nous avons annoté manuellement, en gras et en souligné, les principaux indices que nous utilisons :

Exemple 1". *Dans un communiqué commun**,** **J**ean **G**lavany, **B**ernard **K**ouchner et **F**rançois **P**atriat, qui dis**ent** avoir pris connaissance du rapport de la commission d'enquête, soulign**ent** jeudi **que** ''la sécurité sanitaire des aliments, la lutte contre l'encéphalopathie spongiforme bovine (ESB) et sa prévention, constitu**en**t une priorité majeure de l'action gouvernementale**''.***

On y voit des morphèmes grammaticaux (le suffixe *ent***,** la conjonction *que*) et des indices typographiques (les guillemets, la virgule, le point, les mots capitalisés), l'indice positionnel étant la position finale du couple guillemet + point.

**Le rôle de l'indice**

Il est important de noter que la valeur affectée à la source ou au relateur est le résultat d'un calcul dépendant du contexte, la valeur n'est jamais affectée *a priori* ou hors-contexte. Autrement dit, un indice est un élément qui entre en compte dans le calcul d'une valeur, il ne la détermine pas à lui seul. Dans l'exemple ci-dessus, certaines capitalisations (*ESB*) ou certaines finales en *ent* sont détectées, puis abandonnées. En effet, le suffixe *ent* entre dans le calcul de l'identification du relateur et participe à la caractérisation d'un relateur verbal. Sa présence en fin de mot ne suffit pourtant pas à décider d'emblée que le mot est effectivement un relateur verbal.

De même, pour le segment *Jean Glavany*, l'identification d'une suite de mots capitalisés participe à l'identification de la source de la citation (la source pouvant être un locuteur). Cependant l'attribution de la valeur *source* ne sera effective que lorsque des indices formels externes à ce segment (le contexte) co-détermineront cette attribution.

**Le calcul de la valeur**

Le calcul de la valeur des trois variables, source, relateur et discours rapporté exploite le faisceau d'indices intra-phrastiques, c'est-à-dire l'ensemble des indices présents dans le cadre de travail. Comme nous l'avons vu précédemment, c'est la co-présence et la position relative des indices qui déclenche la résolution du problème donc l'affectation des 3 variables. L'algorithme utilisé évalue les variables du motif et le contexte minimal : si le faisceau

d’indices permet de déterminer immédiatement trois valeurs cohérentes, alors l’instanciation des 3 variables est effectuée ; si le faisceau d’indices permet d’identifier seulement 2 variables alors une déduction sur critère positionnel permet de trouver la valeur de la 3ème variable.

Avant d’illustrer ces principes généraux, nous présentons plus avant la nature de la déduction lorsque seulement 2 variables sont identifiées par le faisceau d’indices. La déduction repose sur le fait que l’ordre des éléments de la citation peut être utilisé. Pour cela, un modèle est nécessaire. Pour le français, nous avons établi le modèle suivant, admettant deux motifs (comprenant chacun 3 variables et 3 positions) :

— l’ordre normal : (source + relateur + discours rapporté), qui constitue le motif canonique ;

— l’ordre inversé : (discours rapporté + relateur + source).

Pour l’anglais, nous utilisons un modèle où l’ordre normal est (source + relateur + discours rapporté) mais où l’ordre inversé est un peu différent (discours rapporté + source + relateur).

Lorsque deux variables sont identifiées et que le motif obtenu est compatible avec l’une des séquences du modèle, alors la déduction est possible puisqu’il ne reste qu’une position disponible pour une variable non déterminée. Par exemple, si le contexte permet de construire le motif incomplet suivant : (? + relateur + discours rapporté), le recours au modèle permet de diagnostiquer que ? doit correspondre à la valeur de la source.

Dans le cas où une seule variable est calculable à partir d’un faisceau d’indices, aucune déduction fiable sur la position des deux autres n’est possible. Le problème n’est donc pas soluble, du moins dans le contexte immédiat.

**Illustrations**

Examinons comment notre détecteur automatique a calculé la solution pour l'exemple 3 en ordre inversé.

Exemple 3. *"Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux",* ***a estimé*** **le ministre de l'Agriculture à l'Assemblée nationale**.

La recherche des indices met en évidence 3 marques à exploiter :

**" / position début** marque la possibilité d'une ouverture de discours rapporté en tête et **",** marque la fin potentielle d'un discours rapporté. **a …é** marque un relateur verbal potentiel.

L'exploitation des indices est la suivante :

— la co-présence des marques d'ouverture et de fin de discours rapporté direct, ainsi que leur position relative (la marque de début précède la marque de fin) indique qu'un discours rapporté est envisageable en tête ;

— la possibilité d'un discours rapporté en tête, marqué en fin par le couple **",** indique qu'un ordre inversé peut être attendu ;

— la co-présence des marques de discours rapporté potentiel et de la marque de relateur verbal potentiel, ainsi que leur position relative, est cohérente avec l'ordre inversé attendu.

La résolution par le système se fait en deux temps. Le faisceau d'indices permet de déterminer 2 variables : le discours rapporté et le relateur. Le motif en ordre inversé étant identifié, à savoir (discours rapporté + relateur + ?), le recours au modèle permet d'identifier que l'inconnue en position 3 correspond à la source. Sur cet exemple réel, on constate que le processus permet une identification des trois variables, sans recours à des ressources lexicales, alors qu'une variable, en l'occurrence la source, n'est pas formellement marquée.

Détaillons le processus du détecteur automatique sur un second exemple.Exemple 4. **Il a prévenu que** *la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.*

La recherche des indices met en évidence 3 marques à exploiter : *Il* / position début, marque la possibilité d'un locuteur en tête ; *a …u*, marque un relateur verbal potentiel ; *que*, marque un potentiel de discours rapporté indirect à suivre.

L'exploitation des indices en co-présence et en position relative dans la phrase conduit à l'identification immédiate du motif en ordre normal (source + relateur + discours rapporté).

Au travers de ces deux exemples, on note que les marques d'élément verbal au passé composé (*a …é* et *a …u*) ne déterminent pas immédiatement l'affectation de la variable relateur. C'est le contexte et notamment l'identification simultanée d'au moins une des deux autres variables

qui déclenche l’affectation. En situation réelle, beaucoup d'indices sont relevés sans aboutir à une affectation effective, spécialement dans le cas de la variable source.

**Bilan**

Les écueils de cette approche en contexte immédiat sont centrés sur les citations indirectes, puisque le discours rapporté n’y est pas marqué. On note du silence, c’est-à-dire des citations non repérées, lorsque la source est un titre ou une fonction car les indices à exploiter manquent. On note également du bruit (citations abusivement détectées), suite à l’*a priori* sur le statut de pronom personnel de *il* et de l'attribution potentielle de rôle de source alors qu’il peut s’agir du pronom impersonnel.

La source de la citation peut être détectée avec certaines "scories", comme on le voit par exemple dans le segment "*le ministre de l'Agriculture à l’Assemblée nationale*" identifié globalement comme source dans l'exemple 1". Nous avons laissé ce problème en suspens, aucune solution satisfaisante n’ayant été trouvée.

Par ailleurs, lorsque le discours rapporté est isolé typographiquement (comme c’est le cas pour le second paragraphe de l’exemple suivant), la source et le relateur sont absents du cadre de travail, on ne peut alors les identifier.

Exemple 4. *Prof Jane Wardle, of the Imperial Cancer Research Fund, said: "Women who felt they had poorer knowledge about breast cancer were more likely to report that they were worried about it and were more pessimistic about surviving breast cancer.*

*"The results of this survey show there is considerable room for improvement in women's knowledge about breast cancer. Knowing more about the disease may also reduce anxiety about breast cancer."*

Les indices du cadre phrastique ayant tous été utilisés, les sources d’amélioration ne sont pas difficiles à identifier : soit l’on reste dans le cadre intra-phrastiques et la seule option est alors de caractériser la source et le relateur à l’aide de ressources lexicales. Soit l’on a recours a des invariants, cette fois-ci, extra-phrastiques. C’est la seconde stratégie que nous allons retenir ici, d’une part pour prolonger une approche syntaxique donnant des résultats satisfaisants, et d’autre part pour ne pas mélanger deux approches par nature opposées.

## 1.4 Critères trans-phrastiques

Les critères trans-phrastiques ont pour objectif d'améliorer les capacités de détection de la stratégie syntaxique par l'apport d'informations externes au cadre phrastique. Pour assurer une continuité de la démarche, la nature des informations complémentaires reste cependant similaire, à savoir la co-présence et la position relative d'indices typographiques, morpho-syntaxiques et positionnels. On observe des régularités dans le cadre d'une construction plus large. Ces régularités, qu'une recherche exclusivement centrée sur la phrase ignore, ont été observées sur notre corpus d'articles journalistiques et de dépêches ; nous ne présenterons ici que celles qui sont nécessaires à la compréhension de l'implémentation proposée.

**Variations formelle et positionnelle de la source**

Nous savons qu'en contexte élargi à un discours écrit, ici l'article de presse, une forme significative est rarement répétée à l'identique : des procédés tels que l'anaphore, la co-référence ou la réduction lexicale (Kocourek, 1982) interviennent régulièrement pour soutenir la structuration discursive. On peut aussi évoquer le principe d'économie de langage. Par exemple, dans la dépêche suivante, le journaliste n'a pas répété continuellement *Jean Glavany* dans son texte, pour y faire référence : il a modifié la forme significative tout au long de son article en ayant recours au procédé de réduction lexicale (*Glavany*), de co-référence nominale (*le ministre de l'Agriculture*), et d'anaphore pronominale (*il*). Ces différents procédés permettant de référencer une entité unique sont utilisés en alternance.

Exemple 6.

31Janv2001 FRANCE: **Glavany** plaide pour une nouvelle PAC où l'on produirait "mieux".

PARIS, 31 janvier (Reuters) - Préconisant une rupture avec le modèle productiviste de l'après-guerre, **Jean Glavany** a plaidé pour une politique agricole européenne s'appuyant sur la qualité et le respect de l'environnement.

"Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux", a estimé **le ministre de l'Agriculture** à l'Assemblée nationale.

Mettant l'accent sur le respect de l'environnement, la sécurité sanitaire et la qualité des

produits, **Jean Glavany** a estimé que la PAC devait "être refondée en profondeur".

"Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", a relevé **le ministre de l'Agriculture**.

**Il** a prévenu que la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.

"Maintenant devant le débat public européen qui est posé, nous devons essayer d'aller plus loin au niveau de l'Europe, notamment pour tirer les leçons de la crise bovine", a-t-**il** lancé.

"La détermination du gouvernement français est de tirer les leçons de cette crise et d'acter le pas dans cette reconversion de l'agriculture (...) vers ce modèle qualitatif que nous attendons", a conclu **Jean Glavany**.



Lorsque l'alternance met en scène la co-référence nominale, le procédé fait l'objet d'une annonce rapide, par une apposition intra-phrastique du qualifiant à la forme première lors de sa première occurrence. On peut également trouver, comme dans l'exemple cité, une quasi-citation (*plaider pour*) dans l'introduction thématique. Suivant ce principe, l'anaphore pronominale apparaît, quant à elle, généralement après la co-référence nominale. Les transformations successives au cours d'un développement se présentent souvent comme des dégradations ou des simplifications des formes précédentes (notamment de la forme première et de la première co-référence nominale). Ainsi, les réductions lexicales et les anaphores auront tendance à apparaître après des formes très déterminées, parfois même sur-déterminées.

Pour conclure cette partie, nous mentionnerons que l'identification de la source d'un discours rapporté ne coïncide pas avec l'identification de la source locale intra-phrastique. Il convient de gérer les différents procédés de co-référence pour associer à un discours rapporté sa

véritable source. Pour cela, notre étude a montré qu'il convenait de chercher cette source de référence dans l'introduction thématique ou en début de chaîne de citations.

**Constantes textuelles pour la détection automatique**

Dans le corpus journalistique traité, nous constatons que la plupart des articles courts sont mono-thématiques et mono-locuteur. Pour ces articles, il est possible de faire une adéquation simplificatrice entre texte et développement thématique. Plus précisément, la source peut être assimilée au thème principal. Dans cette configuration bien particulière mais assez fréquente pour ne pas être négligée, deux constantes textuelles ont été observées et ont donné lieu à modélisation dans notre système de détection automatique :

— La préservation de la marque aspectuo-temporelle du relateur : on note en effet, que malgré la grande variété des verbes citatifs utilisés dans une chaîne de citations, l'auteur a toujours tendance à conserver une constance du temps et de l'aspect (Weinrich, 1973), par exemple le passé composé (en gras et souligné dans l'exemple suivant) ;

— La constance du trait de définitude de la source : on constate que, dans une chaîne de citations, les références à une même entité source conservent toujours leur trait "défini" (en gras dans l'exemple suivant) ou "indéfini". Le caractère indéfini est marqué par l'emploi du pronom *on*, par exemple "*précise-t-on*", ou de l'article indéfini *un*, dans "*ajoute un analyste*.".

Exemple 6'

31Janv2001 France: Glavany plaide pour une nouvelle PAC où l'on produirait "mieux".

PARIS, 31 janvier (Reuters) - Préconisant une rupture avec le modèle productiviste de l'après-guerre, **Jean Glavany a** plaid**é** pour une politique agricole européenne s'appuyant sur la qualité et le respect de l'environnement.

"Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux", **a** estim**é le** ministre de l'Agriculture à l'Assemblée nationale.

Mettant l'accent sur le respect de l'environnement, la sécurité sanitaire et la qualité des produits, **Jean Glavany a** estim**é** que la PAC devait "être refondée en profondeur".

"Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de

produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", **a** relev**é le** ministre de l'Agriculture.

**Il a** préven**u** que la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.

"Maintenant devant le débat public européen qui est posé, nous devons essayer d'aller plus loin au niveau de l'Europe, notamment pour tirer les leçons de la crise bovine", **a**-t-**il** lanc**é**.

"La détermination du gouvernement français est de tirer les leçons de cette crise et d'acter le pas dans cette reconversion de l'agriculture (...) vers ce modèle qualitatif que nous attendons", **a** concl**u Jean Glavany**.



**Mise en œuvre des critères textuels retenus**

L'intégration des critères trans-phrastiques au sein du système de détection a nécessité la prise en compte de la chronologie du discours, pour pouvoir vérifier au fil du texte la régularité des critères rencontrés.

Le principe de l'algorithme consiste à dérouler le texte phrase à phrase tout en recherchant des début de chaînes de citations (apparition d'un premier locuteur ou changement manifeste de locuteur). Lorsqu'une nouvelle chaîne est détectée, la source de la première citation devient la source de référence. Un mécanisme d'oubli de la source de référence se déclenche en cas d'absence prolongée de référence à cette entité. La marque aspectuo-temporelle est quant à elle mémorisée lors de la première occurrence d'un relateur verbal dans la chaîne de citation. En effet, le premier relateur de la source peut être un introducteur prépositionnel tel que *pour* ou *selon*.

Le système de détection des citations basé sur la stratégie syntaxique a été complété de manière à déclencher la résolution lors de la co-présence des trois phénomènes suivants : existence d'une entité source de référence, préservation de la marque aspectuo-temporelle

mémorisée, caractère défini de la source potentielle. La position relative de ces marques doit bien entendu être en cohérence avec le modèle.

Dans l'exemple ci-dessous, nous montrons le résultat d'une analyse automatique avec contraintes trans-phrastiques, calculée selon le principe décrit. Les sources et relateurs détectés sont mis en valeur, avec, pour les citations intégrant une co-référence nominale ou une anaphore pronominale, l'entité source de référence calculée. Cette entité est placée entre les signes **[@ …@]** pour les co-références nominales et entre **[#…#]** pour les anaphores pronominales.

Exemple 6".

31Janv2001 France: Glavany plaide pour une nouvelle PAC où l'on produirait "mieux".

PARIS, 31 janvier (Reuters) - Préconisant une rupture avec le modèle productiviste de l'après-guerre, **Jean Glavany a plaidé** pour une politique agricole européenne s'appuyant sur la qualité et le respect de l'environnement.

"Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux", **a estimé** **[@Jean Glavany@]le ministre de l'Agriculture à l'Assemblée nationale**.

Mettant l'accent sur le respect de l'environnement, la sécurité sanitaire et la qualité des produits, **Jean Glavany a estimé que** la PAC devait "être refondée en profondeur".

"Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", **a relevé** **[@Jean Glavany@]le ministre de l'Agriculture**.

**[#Jean Glavany#]Il a prévenu que** la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.

"Maintenant devant le débat public européen qui est posé, nous devons essayer d'aller plus loin au niveau de l'Europe, notamment pour tirer les leçons de la crise bovine", **[#Jean Glavany#]a-t-il lancé**.

"La détermination du gouvernement français est de tirer les leçons de cette crise et d'acter le pas dans cette reconversion de l'agriculture (...) vers ce modèle qualitatif que nous attendons", **a conclu Jean Glavany**.

(c) Reuters Limited 2001.

**Bilan**

De même que la stratégie syntaxique exploite des invariants internes aux phrases citatives, la stratégie trans-phrastique exploite des invariants internes aux chaînes de citations. Les invariants recherchés sont de même nature : il s'agit de co-présence et de position relative d'indices typographiques, morpho-syntaxiques et positionnels. Les deux stratégies sont donc basées sur une même méthode ; seule varie l'unité sur laquelle est appliquée cette méthode.

En alliant les deux stratégies au sein d'un unique système de détection automatique, on constate que la prise en compte des critères trans-phrastiques permet de traiter de nombreux cas insolubles dans un cadre purement phrastique. Il y a notamment une réduction importante du bruit et du silence sur les citations indirectes. Le bruit (détection abusive de citations) provenait principalement de l'impossibilité de distinguer les pronoms impersonnels des pronoms personnels. Les critères complémentaires, à savoir la présence d'une source de référence et la préservation de la marque aspectuo-temporelle, ont permis de distinguer ces deux cas. Le silence (non détection de citations) était lié à l'absence de marques, en présence d'une source locale identifiée par une mention de titre ou de fonction. Les critères complémentaires ont permis d'améliorer sensiblement la détection avec cependant une limite intrinsèque au modèle développé : les chaînes de citations sont supposées mono-locuteur. On peut toutefois noter que la modélisation proposée permet non seulement de traiter les articles courts où l'unique locuteur est assimilé au thème principal, mais également de fournir des analyses correctes sur des articles plus longs, organisés en prises de position successives, avec locuteurs alternés.

L'écueil restant s'avère être la récupération de la source de référence dans des chaînes de citations multi-locuteurs avec discours croisés ou enchâssés. Bien qu'une solution lexicale puisse être envisagée en intégrant des verbes tels que *répondre*, *répliquer*, *rétorquer*, nous ne

souhaitons pas nous engager dans cette voie mais tentons plutôt de remplacer la stratégie trans-phrastique par une stratégie rhétorique plus ambitieuse. La modélisation de nouvelles combinaisons de critères textuels, qui ont fait l'objet d'un travail linguistique de fond (Lucas, 2000) devrait nous permettre de résoudre la difficulté sans déroger à nos principes méthodologiques.

Avant de poser les jalons d'une telle approche, il convient d'achever le bilan de la stratégie trans-phrastique. Il est intéressant de noter que pour les cas où le discours direct est isolé et où, par conséquent, aucune source n'est mentionnée dans le cadre de la phrase, il existe désormais une solution élégante : la source de référence mémorisée en début de chaîne de citations permet tout simplement de récupérer l'information manquante et de l'associer au discours rapporté.

Enfin, alors qu'en traitement automatique des langues, la résolution de la co-référence est toujours abordée dans une perspective lexicale, la méthode que nous proposons permet une résolution non lexicale de la co-référence pour les sources de citations, que cette co-référence soit pronominale ou nominale. A ce titre, on notera que les problèmes déjà mentionnés de localisation précise de la source (dans le segment le *ministre de l'Agriculture à l'Assemblée nationale*) n'empêche pas le calcul de l'entité de référence associée (*Jean Glavany*).

# Conclusion

Dans cet article, nous avons présenté une méthode de détection automatique des citations dans les textes informatifs de type dépêches et articles journalistiques. Contrairement à la pratique courante faisant un usage systématique de grands lexiques pour repérer les citations, nous avons souhaité montrer qu'une approche alternative, légère et performante, était envisageable. Cette approche est basée sur la recherche de constantes caractéristiques.

Les constantes caractéristiques sont exprimées exclusivement en terme d'indices typographiques, morpho-syntaxiques et positionnels. Nous avons vu que pris isolément, ces marques n'ont aucune valeur, ce n'est qu'en co-présence et par leur position relative qu'elles deviennent informatives et peuvent permettre l'affectation fiable de valeurs. La méthode que

nous proposons gère la prise en compte simultanée de critères à la fois internes et externes au segment à identifier. Cette méthode n'est en fait pas nouvelle puisqu'elle est dans la droite lignée de nos travaux sur la reconnaissance des structures formelles (Giguet, 2000). Dans le cadre de la détection de citations, les critères internes sont intra-phrastiques et les critères externes sont puisés dans les chaînes de citation.

Les avantages d'une telle méthode sont nombreux. Bien entendu, la légèreté de la solution est appréciable mais également, l'abandon des ressources lexicales permet une plus grande robustesse, ou capacité d'adaptation, que ce soit face à l'apparition de nouveaux locuteurs (e.g., *Ben Laden*), à l'apparition de nouveaux titres ou fonctions (e.g., *Ministre de la Vertu* ou *président du Parti Communiste*), ou à l'emploi de relateurs verbaux inattendus (e.g., *stigmatiser*). Sur le plan économique, il est aussi intéressant de noter que la transposition multilingue d'un système reposant sur de tels critères est peu coûteuse : aucun dictionnaire n'est en effet à traduire ou à spécialiser. Dans une approche lexicale classique, le coût en traduction concerne également les noms propres lorsqu'ils sont transcrits (e.g. existence des graphies *Putin* et *Poutine* ou *bin Laden* et *ben Laden*). Les noms propres de l'actualité sont le principal casse-tête en filtrage d'information.

Enfin, nous avons montré les premiers résultats d'une résolution non lexicale de la co-référence pour les sources de citations. Ces résultats sont très prometteurs et nous incitent à poursuivre plus avant notre recherche, tout en systématisant la méthode. Ainsi, la gestion des chaînes de citations avec discours croisés et enchâssés, qui restent le véritable écueil du système actuel, est en cours de modélisation informatique. Nos travaux nous invitent à penser que cette difficulté peut être surmontée sans déroger à nos principes méthodologiques. Ils devraient déboucher sur une nouvelle stratégie de détection automatique dite stratégie rhétorique.

## Références


Authier-Revuz, J. (1992-93). Repères dans le champ du discours rapporté *L'information grammaticale*, 55 et 56, pp. 10-15.

Banfield A. (1973). Le style narratif et la grammaire des discours direct et indirect. *Change*, n° 16-17, pp. 190-226.

Charolles, M. (2000). *Les expressions introductrices de cadres de discours et leur portée textuelle*, séminaire de recherche. Paris III, Sorbonne Nouvelle – (Censier).

Coulmas, F. (1986) (sous la dir. de). *Direct and indirect Speech*. Berlin, New York, Mouton de Gruyter.

Dendale, P. (2001). "Le discours rapporté;. Quelques réflexions à propos de son traitement dans les grammaires en général et dans la *Grammaire 2000* en particulier, *In* : Van Huffel, B. & Segers, W. (éds), *Mélanges. Vertalers en verwanten*, Antwerpen, Lessius Hogeschool pp. 55-73.

Desclés, J.-P. et Minel, J.-L. (2000). "Résumé automatique et filtrage sémantique de textes", In *Ingénierie des langues*, sous la direction de J.-M. Pierrel, Paris, Hermès, pp. 253-268.

Giguet, E. (1998*). Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de doctorat. Université de Caen.

Kocourek, R. (1982). *La langue française de la technique et de la science*. Wiesbaden,, Brandstetter Verlag ; distrib. Paris: Documentation française. 259 p.

Lucas, N. (2000). "Le rôle de la citation dans la structuration des articles de presse", In:*Actes du premier colloque d'études japonaises de l'Université Marc Bloch: Strasbourg, 5 et 6 mai 2000*, sous la dir. de S. Murakami-Giroux, et C. Séguy. Strasbourg, Université Marc Bloch Strasbourg, pp. 215-244.

Mourad, G. (2000). "Présentation de connaissances linguistiques pour le repérage et l'extraction de citations". *TALN, 7ème Conférence Annuelle sur leTraitement Automatique des Langues Naturelles* Lausanne, Suisse 16 - 18 octobre 2000. pp. 495-501.

Mourad, G. et Minel J.-L (2000). "Filtrage sémantique du texte, le cas de la citation", *3ème Colloque International sur le Document Electronique, CIDE'2000, Lyon, France* sous la direction de Gaio et Trotoux, Caen, pp. 41-56.

Vergne, J. (1990). A parser without a dictionary as a tool for research into French syntax. *Proceedings 13th International Conference on Computational Linguistics, CoLing 90*, 70-72.

Vergne, J. (2001). Analyse syntaxique automatique de langue : du combinatoire au calculatoire, Tours, *TALN 2001*.

Weinrich, H. (1973). *Le Temps : le récit et le commentaire*. Paris, Seuil.