RÉSOLUTION ANAPHORIQUE : ÉTAT D'UNE REFLEXION COLLECTIVE

PASCAL AMSILI, FRÉDÉRIC LANDRAGIN, ALEJANDRO ACOSTA, AND ANDRÉ BITTAR

Les objectifs de cette communication sont d'une part de présenter à la communauté les travaux menés au sein d'un groupe de travail « anaphores » dans le laboratoire Lattice, et en particulier certains choix issus de notre réflexion, et d'autre part de faire office de bilan intermédiaire de nos activités, concernant en particulier l'annotation des relations anaphoriques.

L'objectif principal de ce groupe de travail est la mise en commun de travaux de recherche menés dans notre laboratoire, portant sur de nombreux aspects de la résolution anaphorique (résolution automatique d'anaphore pronominale en anglais et en français, repérage automatique des emplois non anaphoriques de certains pronoms (*il*, *ce*...), traitement automatique des anaphores abstraites, et en particulier événementielles, etc.).

Plus précisément, le travail collectif du groupe vise à produire des ressources pertinentes pour l'ensemble de ces recherches, en particulier, des ressources bibliographiques, et surtout un **corpus de référence** annoté, et, ce qui va avec, un **guide d'annotation**, ces deux aspects étant cruciaux pour permettre une validation des modules réalisés. Un dernier point important : les modules réalisés sont d'ores et déjà intégrés dans une chaîne de traitement développée depuis plusieurs années dans le laboratoire, appelée *macaon* [Nasr et Acosta, 2007].

Les activités du groupe s'étant pour le moment concentrées sur l'analyse de corpus et l'annotation collaborative des relations anaphoriques, la présente communication fait le point sur les phénomènes initialement retenus, sur les problèmes rencontrés lors de leur annotation dans des textes réels, et sur les choix représentationnels qui en ont résulté. Nous ne présentons pas de schéma d'annotation finalisé, mais nous soulevons un certain nombre de problèmes qui nous semblent avoir été quelquefois ignorés et s'avérer pourtant dignes d'intérêt pour les systèmes de traitement automatique des langues.

1. Hypothèses privilégiées, choix méthodologiques

Dans le but de préciser notre démarche, nous adoptons la terminologie suivante :

Anaphore: expression référentiellement non autonome au sens de [Milner, 1982, p. 19]. Le point essentiel est que ce manque d'autonomie référentielle est marqué linguistiquement (pronoms, groupes nominaux possessifs, groupes nominaux démonstratifs). Selon cette définition, les descriptions définies ne sont pas inclues dans les anaphores. C'est un choix de départ qui nous permet de restreindre l'ensemble des phénomènes auxquels nous nous intéressons, et en particulier de laisser de côté (pour le moment) les anaphores associatives et autres phénomènes dont l'identification automatique est plus que délicate. Cette définition, opérationnelle mais sans doute un peu brutale, est donc liée aux formes linguistiques plus qu'aux phénomènes sémantiques.

Antécédent: portion de texte, potentiellement discontinue, dont l'interprétation est nécessaire pour calculer le sens de l'anaphore.

Nos réflexions, qui se sont inspirées de plusieurs projets antérieurs rapportés dans la littérature (en particulier [MUC-7, 1997, Salmon-Alt, 2002]), ainsi que de nos premières expérimentations, ont mené aux observations ou positionnements suivants :

– Les schémas d'annotation existants sont surtout prévus pour l'annotation des liens de coréférence et non des liens anaphoriques, avec souvent une confusion entre les deux, confusion qui tend à simplifier certains phénomènes linguistiques (MUC se restreint par exemple à la seule relation d'identité 'Ident'). Face à ce problème nous avons choisi de nous intéresser pour l'instant aux seuls liens anaphoriques, afin de mieux les comprendre et les cerner.
– Il est parfois difficile de distinguer d'une part une relation anaphorique ou de coréférence, et d'autre part une relation ontologique. Plusieurs schémas d'annotation incluent de ce fait des connaissances ontologiques, par exemple du domaine dont le texte relève [Gasperin *et al.*, 2007]. Pour notre part, nous visons une approche générique, indépendante du domaine, et nous n'inclurons aucune information ontologique, ce qui ne nous permettra pas de repérer toutes les relations, mais nous incitera à rester dans les limites de faisabilité d'un résolveur destiné à traiter du texte tout-venant.
– Même quand l'objectif initial est la résolution des pronoms, il est peu fait mention de certaines catégories de pronoms et les efforts sont souvent concentrés sur les seuls pronoms personnels de troisième personne. Nos exemples (avec les commentaires ci-dessous) montrent que l'annotation de tous les pronoms est une tâche délicate soulevant de nombreux problèmes qui ne nous semblent pas assez débattus.
– Comme le soulignent [van Deemter et Kibble, 2000], on observe des lacunes dans la méthodologie, ce qui tend à argumenter pour un retour aux sources. C'est ainsi que nous avons décidé de démarrer les activités de notre groupe

Date: 30 avril 2007.

Nous remercions les autres participants au groupe « anaphores » à Lattice : Francois-Régis Chaumartin, Elisabeth Chenevois, Benoît Crabbé, Laurence Danlos, Elzbieta Gryglicka, Sylvain Kahane, Frédéric Laurens, Ricardo Minhoto, Alexis Nasr, Maud Pironneau, Marie-Pierre Sales Benjamin Surma, Grégoire Winterstein ; ainsi que Susanne Salmon-Alt.

de travail en ne prenant pas pour point de départ un schéma d'annotation existant (ce qui ne nous empêche pas de consulter les schémas décrits dans la littérature).
– La notion de **chaîne anaphorique** (ou de coréférence) [Corblin, 2002] est considérée comme seconde dans notre approche : résultat d'un calcul sur les relations anaphoriques (voir plus loin).

2. Phénomènes étudiés

Une des tâches de notre groupe de travail consiste donc à discuter de la nature des antécédents, dans divers cas de figure extraits de plusieurs corpus (dont les archives du journal « Le Monde » et « Frantext »). Les exemples qui suivent ont suscité des discussions particulièrement longues. Ils sont annotés avec les anaphores en gras, les antécédents identifiés en souligné, et des numéros de correspondance en indice pour les antécédents et en exposant pour les anaphores. Le double soulignement est utilisé lorsque deux antécédents sont imbriqués.

(1) « Tu$_4$ es mariée$_1$, et je$_2$ n'ai pu **l'**^{1}empêcher ; mais si **mon**^{2}$_3$ bonheur, si **mes**^{3}$_5$ jours **te**^{4}$_7$ sont chers, il faut qu'avant **mon**^{5}$_6$ départ **je**^{6} **te**^{7} voie, ne fût-ce que cinq minutes ».
Frantext, N253 (Scribe, *Le mariage de raison*).

Dans l'exemple (1), nous avons choisi d'annoter tous les pronoms à l'exception du pronom impersonnel « il » dans « il faut qu'avant mon départ [...] », du pronom « ce » dans « ne fût-ce » et des premières mentions des pronoms personnels « je » et « tu », dans la mesure où il ne s'agit pas d'anaphores. Les mentions ultérieures de « je » et « tu » sont en revanche annotées comme des anaphores, d'où les antécédents 2 et 4. Il en va de même pour les adjectifs possessifs, et nous suivons le principe consistant à choisir comme antécédent la mention la plus récente. C'est pourquoi « mes » dans « mes jours » reprend « mon » dans « mon bonheur » qui lui-même reprend la première mention de « je ».

(2) « Autrefois **j'**^{3}avais une servante$_1$ — jeune et jolie ; **c'**^{1}était la fille d'une pauvre femme ; — mais on$_2$ jasait dans la maison, et quand **on**^{2}$_4$ rencontrait **ma**^{3}$_6$ domestique$_5$, **on**^{4} **lui**^{5} chantait sur l'escalier : "allons, Babet, un peu de complaisance."$_7$ **J'**^{6}$_9$ai entendu **ça**^{7} un jour$_8$ — et **ça**^{8} **m'**^{9}a fâché ».
Frantext, L863 (Murger, *Scènes de la vie de jeunesse*).

Dans l'exemple (2), nous suivons le même principe pour l'annotation des occurrences de « on », de « je » et des adjectifs possessifs. Une première remarque concerne le personnage de Babet. Elle est successivement mentionnée comme « une servante [jeune et jolie] », « c' », « la fille d'une pauvre femme », « ma domestique », « lui » et « Babet », mais seuls les pronoms activent la recherche d'un antécédent et sont annotés. D'où les anaphores 1 et 5, chacune de ces anaphores reprenant l'antécédent le plus récent, à savoir « une servante » dans le premier cas et « ma domestique » dans le second. Autrement dit, aucun rapprochement n'est fait entre toutes les occurrences de ce personnage. La raison en est que, comme nous l'avons dit, nous cherchons les relations anaphoriques et non les relations de coréférence. Si tel avait été notre objectif, il nous aurait fallu mettre en œuvre des principes d'annotation pour les groupes nominaux indéfinis et définis (ce qui n'entre pas dans le cadre de ce travail, du moins pour le moment). Une deuxième remarque concerne l'annotation des références abstraites, avec l'apparition de deux « ça » dans cet exemple. Le premier « ça » est quelque chose que le narrateur a entendu, et nous avons choisi comme antécédent le discours rapporté. Or nous aurions pu inclure le fait que cette phrase était chantée, le verbe chanter apparaissant dans la narration et relevant bien de ce que le narrateur a entendu. Il aurait alors été tentant d'inclure la phrase « on lui chantait sur l'escalier », ce qui, par contre, relève de la situation et donc dépasse l'objet entendu. Une alternative à notre solution aurait ainsi été l'antécédent discontinu « chantait [...] : "allons, Babet, un peu de complaisance." », qui nous a paru pour le moins délicat à défendre. Dans des cas semblables à celui-ci, nous choisirons la solution la plus courte.

(3) « Et ce$_1$ **que**^{1}$_2$ je$_3$ n'oublierai jamais de la vie, voyez-vous$_5$, **c'**^{2}est lorsque M. Cauche, là-bas, sur le quai$_4$, est venu arrêter aussi M. Roubaud$_6$. **J'**^{3}**y**^{4} étais. **Vous**^{5} savez que **ça**^{6} **s'**$^?$est passé huit jours après seulement, lorsque M. Roubaud$_7$, au lendemain de l'enterrement de **sa**^{7}$_8$ femme, avait repris **son**^{8} service d'un air tranquille ».
Frantext, L857 (Zola, *La bête humaine*).

L'exemple (3) pose tout d'abord le problème de la cataphore, avec le deuxième mot de l'extrait, « ce ». Si nous avions voulu rendre compte des cataphores, il nous aurait fallu commencer par identifier l'antécédent de « ce », à savoir l'antécédent événementiel qui est étiqueté par ailleurs avec le nombre 6. Dans ce cas, l'annotation des anaphores concernant cet événement aurait posé un problème : « que » (étiqueté 1) reprend « ce », puis « c' » (étiqueté 2) reprend « que », puis « ça » (étiqueté 6) aurait repris soit « c' » soit l'antécédent 6 avec une ré-identification complète de celui-ci. Or si nous appliquons le principe consistant à étiqueter l'antécédent le plus récent (disons le plus proche dans le cas des cataphores), l'antécédent de « ce » aurait été « que », lui-même étant ensuite identifié comme une anaphore reprenant l'antécédent le plus récent, donc « ce »... Autrement dit, le principe que nous suivons depuis le début s'avère incompatible avec les cataphores. Face à cette limitation, nous avons choisi temporairement d'ignorer les cataphores, et c'est pourquoi l'exemple (3) est ainsi annoté. Nous noterons au passage que cet exemple inclut un pronom relatif, ce que nous n'avions pas vu jusqu'à présent. Le cas particulier des *relatives périphrastiques* [Riegel *et al.*, 1994, p. 487] en *ce que*, *celui qui*... demande vraisemblablement un traitement spécifique.

(4) « En mobilisant les associations et les enfants, les minéraliers$_1$ **se**^{1} donnent une image$_2$ encore plus écolo que **celle**^{2} de les verriers$_3$, **qui**^{3}$_4$ **se**^{4} contentent de racheter les bouteilles recueillies dans les conteneurs municipaux ».
Extrait du journal Le Monde.

L'exemple (4) illustre l'annotation des pronoms démonstratifs et réfléchis, sans difficulté particulière, mais avec des situations qui montrent à quel point l'identification des relations anaphoriques est un problème différent de celle des relations de coréférence : l'image des minéraliers n'est pas coréférente à celle des verriers, mais la relation anaphorique permet de mettre les deux en rapport. Bien sûr, il arrive aussi qu'une chaîne anaphorique puisse s'assimiler à une chaîne de coréférence, comme avec les antécédents 3 et 4.

(5) « Pendant des années, les constructeurs automobiles français$_1$ ont refusé de **s'**[1]engager dans la "voiture propre"$_2$, sous prétexte qu'**elle**[2] faisait la part trop belle à l'électronique allemande (Bosch) et à les techniques américaines (Du Pont de Nemours et Corning Glass pour les pots d'échappement catalytiques). A **cet**[?] égard, le président de PSA (Peugeot), Jacques Calvet, aura lutté jusqu'au bout pour refuser le catalyseur, jugé d'un prix prohibitif pour les petites cylindrées ». *Extrait du journal Le Monde.*

Dans l'exemple (5), « à cet égard » est difficile à rattacher à un antécédent : quand on lit l'ensemble de la phrase, on peut comprendre soit que le sujet est de faire la part trop belle aux techniques américaines pour les pots d'échappement catalytiques, soit que « à cet égard » vaut pour « en ce qui concerne les pots d'échappement catalytiques ». Dans les deux cas, il semblerait plus raisonnable de dire que l'antécédent est quelque part dans le début du paragraphe, sans préciser exactement où. Face à la difficulté de cet exemple, nous avons choisi de ne pas déterminer d'antécédent et d'utiliser le marquage « ? » déjà présent dans (3). Une autre stratégie possible serait de considérer l'expression comme partiellement figée (mais ce type de construction est productif : *à ce sujet*, *à ce propos*...).

(6) « **Le marché**[1] est myope ; livré à **lui**[1]-même$_2$, **il**[2] ne donne pas spontanément un prix à les ressources écologiques$_3$ **dont**[3] la "gratuité" entraîne la surexploitation, le gaspillage et la dévastation ». *Extrait du journal Le Monde.*

Enfin, l'exemple (6) montre un cas peut-être peu pertinent mais intéressant de dissociation de « lui-même ». La solution simple aurait été de considérer « lui-même » et non « lui » comme une reprise anaphorique de « le marché », puis comme antécédent de « il ». En fait, l'une ou l'autre solution dépend du découpage en mots réalisé, et les deux sont acceptées.

3. Vers un guide d'annotation

Au final, l'annotation des phénomènes retenus pose les principaux problèmes suivants :

(1) **Le non marquage des pronoms impersonnels**. Nous supposons que ceux-ci ont été détectés et mis de côté au préalable, comme c'est fait dans la plupart des systèmes, et, à LaTTICe, par [Danlos, 2005] (un module est en cours de développement pour *ce*).

(2) **Le marquage des antécédents longs**, comme des phrases voire des paragraphes complets dans le cas des anaphores abstraites. L'exemple (5) fait intervenir « à cet égard », mais nous avons également rencontré « pour ce faire » ou encore « dans ce contexte ». Pour ce problème, nous retenons le principe d'un antécédent minimum (« entendu » ou « ai entendu » dans le cas de l'anaphore 8 de l'exemple 2), qu'un système doit identifier à défaut d'identifier l'antécédent complet. On s'inspire là de la stratégie retenue dans des travaux de résolution automatique des anaphores événementielle [Eckert et Strube, 2001, Byron et Tetreault, 1999].

(3) **Le marquage des antécédents discontinus** (cf. la discussion autour de l'exemple 2).

(4) **Le marquage des antécédents de type groupe nominal prépositionnel**, avec les exemples : « des verriers » et « aux ressources écologiques ». Faut-il inclure la préposition ou au contraire l'exclure en excluant corrélativement le déterminant ? Nous choisissons de réaliser le marquage anaphorique après le passage d'un analyseur morphosyntaxique chargé de dissocier ces amalgames, ce qui explique l'apparition dans nos exemples de « de les » et de « à les » à certains endroits (la chaîne de traitement *macaon* intervient en amont).

(5) **Le marquage des possessifs et des démonstratifs**, avec le choix que nous avons fait de n'étiqueter que l'adjectif, choix qui conduit parfois à des annotations peu lisibles mais qui a l'avantage, en particulier pour les possessifs, de permettre facilement à la fois le marquage du possesseur (adjectif) et celui du possédé (le groupe nominal complet), quand celui-ci est identifié par ailleurs comme antécédent. C'est d'ailleurs le choix fait lors des conférences MUC [MUC-7, 1997].

(6) **Le marquage de l'antécédent le plus récent**, qui va parfois à l'encontre du calcul intuitif du sens, du fait de la pauvreté sémantique de certains antécédents retenus par rapport à d'autres moins récents. Ce point est essentiel pour plusieurs raisons : d'une part parce que c'est un principe très strict qui permet d'éviter aux annotateurs certaines interrogations (on les imagine tentés d'étiqueter comme antécédent le groupe nominal le plus complet, c'est-à-dire le plus autonome référentiellement) ; d'autre part parce que ce principe est dirigé par des préoccupations computationnelles. Autrement dit nous cherchons à spécifier une méthode d'annotation qui soit proche des processus calculatoires.

Les principes d'annotation ainsi choisis présentent plusieurs avantages :

– Les anaphores sont faciles à détecter, même pour un annotateur non spécialiste, puisqu'elles consistent en un ensemble clair et fini de formes linguistiques (pronoms, possessifs, démonstratifs).
– Pour chaque anaphore, la recherche de l'antécédent par l'annotateur se fait en remontant le texte pas à pas, toujours dans le sens inverse de la lecture, et s'arrête au premier antécédent trouvé. Si le principe est simple, sa contrepartie est que l'antécédent peut avoir un contenu sémantique faible (et qu'il faille vérifier sa justesse en remontant plus haut dans le texte).

– Pour les anaphores abstraites, la priorité doit être donnée à l'empan textuel le plus court, d'une part parce que c'est plus facile pour l'annotateur, d'autre part parce que cela permet d'évaluer de manière relativement souple les résolveurs d'anaphores : ceux qui auront identifié un empan plus large (avec inclusion complète de l'antécédent déterminé) ne seront pas pénalisés.
– Egalement à propos de l'évaluation de résolveurs d'anaphores, ces principes d'annotation permettent de mettre en place quelques heuristiques simples, en particulier une heuristique capable de remonter les **chaînes anaphoriques** pour ne pas (trop) pénaliser un système qui aurait identifié un antécédent moins récent que celui déterminé manuellement. Avec l'exemple (1), un système qui aura attribué « je » comme antécédent aux différents possessifs de première personne sera ainsi (presque) aussi bien noté qu'un système qui aura retrouvé l'annotation donnée plus haut. D'autre part, si nous reprenons la chaîne anaphorique « les verriers qui se contentent [...] » de l'exemple (4), il est clair que le système qui aura identifié les deux anaphores sera mieux noté que celui qui n'aura pas détecté le « qui » et qui aura mis un lien direct entre « se » et son antécédent « les verriers ». Mais ce dernier lien pourra être compté comme juste, même s'il ne correspond pas exactement à l'annotation manuelle. Nous prévoyons donc un algorithme chargé de donner un sens à la notion de chaîne anaphorique, afin de comparer de manière raisonnable une chaîne complète à une chaîne partielle.

4. Conclusion

Les quelques observations que nous avons voulu partager dans cette communication constituent les bases d'un guide d'annotation qui n'est pour le moment qu'à l'état d'ébauche, tant la confrontation avec la réalité complexe des corpus, même en travaillant de manière collective pour privilégier l'accord inter-annotateurs, reste une source importante de reflexion et de découvertes.

Cette ébauche nous semble cependant pleine de promesses : la méthode de travail choisie, basée à la fois sur les expériences passées, leurs critiques, et la constitution d'un groupe varié et relativement important, devraient nous permettre de proposer rapidement à la communauté concernée des ressources de bonne qualité.

Indépendamment du corpus annoté et de son guide d'annotation, nous sommes engagés actuellement dans le développement de plusieurs autres projets : résolution d'anaphores abstraites (essentiellement, pronom *ça/cela*), résolution d'anaphores pronominales, marquage des occurrences non anaphoriques de pronoms, marquage des évènements dans un texte, traitement des chaînes de co-références dans des textes biographiques.

Notre travail a aussi des perspectives plus lointaines, avec plusieurs dimensions théoriques. D'une part, nous envisageons bien entendu un retour vers les chaînes de co-référence, après ce passage par la relation anaphorique stricte. Cela permet de considérer entre autres le lien entre ces chaînes et la structure discursive. D'autre part, inspirés par les travaux de [van der Sandt, 1992], nous chercherons à voir dans quelle mesure la thèse selon laquelle la présupposition est fondamentalement un mécanisme anaphorique peut être mise à l'épreuve en corpus, et même exploitée dans le cadre d'applications de traitement de la langue.

References

[Amsili *et al.*, 2002] Pascal Amsili, Claire Beyssade, Anne Garreta, et Laurent Roussarie, éditeurs. *Workshop "Chaînes de références et résolveurs d'anaphores"*, Nancy, Juin 2002. Workshop associé à TALN'02.

[Byron et Tetreault, 1999] Donna K. Byron et Joel R. Tetreault. A flexible architecture for reference resolution. In *Proceedings of the 9^{th} Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, 1999. Student paper.

[Corblin, 2002] Francis Corblin. Les chaînes de la conversation et les autres. Talk given at the workshop "Théorie de la référence et théories du personnage", Paris 7, organized by Claire Beyssade, Anne Garreta and Christine Montalbetti, Juin 2002.

[Danlos, 2005] Laurence Danlos. ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*. In *Actes de TALN'05*, pages 123–132, Dourdan, France, 2005.

[Eckert et Strube, 2001] Miriam Eckert et Michael Strube. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17:1:51–89, 2001.

[Gasperin *et al.*, 2007] Caroline Gasperin, Nikiforos Karamanis, et Ruth Seal. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC 2007*, pages 19–24, Lagos, Portugal, 2007.

[Milner, 1982] Jean-Claude Milner. *Ordres et raisons de langue*. Seuil, Paris, 1982.

[MUC-7, 1997] *7th Message Understanding Conference*, 1997. `http://www-nlpir.nist.gov/related_projects/muc/index.html`.

[Nasr et Acosta, 2007] Alexis Nasr et Alejandro Acosta. *macaon*, une architecture de développement de modules de tal. `http://code.google.com/p/macaon/`, 2007.

[Riegel *et al.*, 1994] Martin Riegel, Jean-Christophe Pellat, et René Rioul. *Grammaire méthodique du français*. Linguistique nouvelle. PUF, Paris, 1994.

[Salmon-Alt, 2002] Susanne Salmon-Alt. Le projet ananas : Annotation anaphorique pour l'analyse sémantique de corpus. In Amsili et al. [2002].

[van Deemter et Kibble, 2000] Kees van Deemter et Rodger Kibble. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(2):629–637, 2000.

[van der Sandt, 1992] Rob A. van der Sandt. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–378, 1992.

Lattice (CNRS) & Université Paris-Diderot
E-mail address: `prenom.nom@linguist.jussieu.fr`