

LA RESOLUTION D'ANAPHORE A PARTIR D'UN LEXIQUE-GRAMMAIRE DES VERBES ANAPHORIQUES

Blandine GELAIN & Célestin SEDOGBO
26 place Ovale BULL cediag
94230 Cachan 78430 Louveciennes
(France)

Abstract

This paper presents a system which intends to resolve anaphora in the framework of the Discourse Representation Theory, and using a lexicon-grammar of anaphoric verbs, through the application of selection criteria for assignment of a referent to an anaphora.

From a semantic representation of text provided by a DRT system implemented in Prolog, the system uses several criteria of selection of referent. One of these criteria is the anaphoric conditions of verbs described as a lexicon-grammar of anaphoric verbs.

The present paper investigates a transformational analysis of verbs related to their anaphoric behaviour, and the adequacy of extension of the lexicon-grammar of M.GROSS to anaphoric conditions on verbs.

1. Introduction

La Théorie de la Représentation du Discours (ou DRT¹) propose une approche unifiée de phénomènes du langage naturel tels que le temps, l'évènement, l'anaphore. Elle se caractérise par sa filiation avec la sémantique logique, et sa distance d'avec les niveaux de représentations basées sur la logique des prédicats et ses extensions. Ainsi les notions de conséquence logique et de validité ([SEDO 87]) peuvent s'appliquer naturellement aux structures manipulées par la DRT.

Cette théorie propose une explication de la formation de l'anaphore, sans en proposer la résolution. Celle-ci passe en général par l'application de critères de sélection syntaxiques, sémantiques et pragmatiques, qui lèvent les ambiguïtés engendrées par l'usage de l'anaphore. [GUIN 85], [CARB 88], [RICH 88] entre autres ont introduit les notions de *foyer* et de *contraintes* successives à activer. Cependant ces critères sont parfois insuffisants pour établir une relation évidente entre un nom et un pronom.

Dans sa théorie du Gouvernement-liage,

N.CHOMSKY² explique l'anaphore à partir d'un mécanisme de liage et s'appuie sur la remarque que le liage d'une anaphore à son référent dépend aussi des propriétés anaphoriques du verbe. Aucune étude empirique des propriétés anaphoriques n'a été faite à ce jour, alors que toute approche de résolution d'anaphore devrait être basée sur un lexique des verbes et leurs propriétés anaphoriques associées. Le présent article décrit une approche de résolution d'anaphore qui repose sur:

- la mise en œuvre de la DRT pour représenter la sémantique d'un discours;
- l'élaboration d'un lexique-grammaire des verbes anaphoriques³;
- un système de filtre basé sur différents critères de sélection.

Ce système illustre la résolution automatique de certaines anaphores en partant de la représentation sémantique d'un texte obtenue d'après la DRT. Les pronoms que nous avons étudiés font partie d'un type d'anaphore qui représente une relation pouvant s'établir entre deux phrases sans mettre forcément en jeu une règle syntaxique (le pronom peut identifier un référent dans le discours précédent): *Jean croit que Marie lui offre un livre.*

Cet article est divisé en cinq parties: introduction à notre travail, présentation de la DRT, puis de son implémentation en Prolog, description des lexique-grammaires et leur extension aux propriétés anaphoriques, présentation générale de l'architecture de notre système de résolution, et enfin perspectives de création systématique d'un lexique-grammaire.

2. La représentation sémantique

La DRT se fonde donc sur un ensemble de règles de construction traduisant un discours en une représentation sémantique formelle: **la Structure de Représentation du Discours**. Pour chaque partie de discours, une DRS est construite, boîte pouvant en contenir d'autres,

2 "Government Binding", notée GB.

3 Extension des tables de verbes développées au LADL, aux propriétés anaphoriques (cf [GELA 92]).

1 H.KAMP "A Theory of Truth and Semantic Interpretation" Groenendijk Amsterdam 1981

qui représente le contenu significatif de cette partie. Une DRS complète est l'ensemble de plusieurs DRSs apparaissant à mesure que le discours continue. La DRT étudie donc les contraintes sur cette continuation.

La forme d'une DRS consiste en une paire **<U, Con>** constituant deux zones, où **U** (univers) est un ensemble des **référénts du discours** représentant les entités du discours, et **Con** un ensemble de **conditions** que doivent satisfaire ces référénts. Celles-ci sont des prédicats et des relations de référénts du discours mais peuvent être plus complexes. Notées comme en logique des prédicats, les conditions de vérité sont définies par rapport à la possibilité d'inclure la DRS dans un modèle (pour plus de détails, consulter [GUEN 85]). D'autre part, l'extension d'une DRS ne peut changer les valeurs déjà assignées: tout ce qui était vrai auparavant restera vrai par la suite:

Un camion transporte une charge

U.K=[X1,X2],

Con.K=[camion(X1),charge(X2),transporte(X1,X2)].

Tout camion transporte une charge

U.K=[],

Con.K = [=>,K1,K2];

U.K1=[X1],

Con.K1= [camion(X1)];

U.K2=[X2],

Con.K2=[charge(X2),transporte(X1,X2)].

2.1 La notion d'accessibilité

On peut représenter des restrictions configurationnelles sur les relations anaphoriques possibles entre les pronoms et leurs antécédents. Ces restrictions sont obtenues en réduisant **l'accessibilité** des référénts. L'accessibilité permet donc de déterminer les liens anaphoriques entre un marqueur pronominal et un marqueur de discours. *Toute DRS est accessible d'elle-même; son univers de marqueurs accessibles est l'univers du discours de la DRS.*

Dans une DRS implicative, la DRS antécédente est accessible de la DRS conséquente. Enfin, la relation d'accessibilité est transitive. Donc pour *"tout camion qui transporte une charge la déclare"*, l'antécédent du pronom objet *la* est *une charge*. Ici, U.K1 est accessible à K2.

Les conditions de continuité d'un discours sont aussi fonction de l'accessibilité: la phrase: *"il va à Berlin"* ne peut continuer la précédente puisqu'aucun marqueur de discours n'est accessible de la DRS K (voir à ce sujet [KASP 86]). Ceci explique pourquoi

une phrase comme: *"Chaque chauffeur possède un camion. Il le conduit"*: doit être formulée: *"Chaque chauffeur qui possède un camion le conduit"* (avec *"qui possède un camion"* comme extension de *"chaque chauffeur"*) pour être représentable.

Mais la DRT a des limites; elle n'explique pas la bonne formation de ce texte, par exemple:

Chaque chauffeur transporte une charge.

Il met plusieurs jours à la livrer.

Elle ne sera livrée qu'au bout de 3 jours

Les deux dernières phrases, selon la DRT, ne peuvent suivre la première, à cause de la portée du quantificateur *chaque*. Par contre la phrase *"Chaque chauffeur transporte une charge qui ne sera livrée qu'au bout de 3 jours"* sera parfaitement représentée par la DRT. Ces règles obligeraient donc le locuteur à décrire une situation en une seule phrase!

2.2 Implémentation de la DRT

Notre analyseur sémantique démarre avec des arbres syntaxiques résultant d'une grammaire de type GPSG, programmée en Prolog. La grammaire sémantique est basée sur les mêmes principes: unification de structures, augmentation de listes ordonnées, présentée sous forme de règles de réécriture suivies de contraintes, de type:

ph -> gn gv

<ph drs_courante>=<gn drs_courante>

<gn suite_drs>=<gv drs_courante>

<gn argument>=<gv arg_sujet>

<ph arg_sujet>=<gv arg_sujet>

<ph arg_objet>=<gv arg_objet>.

Ceci donne, à partir de règles DCG issues de la compilation des précédentes:

traduire(ph(GN,GV),P) :-

traduire(GV,P2),

traduire(GN,P1),

imerge(P,[courante],P1,[courante]),

imerge(P1,[suite],P2,[courante]),

imerge(P1,[arg],P2,[arg_sujet]),

imerge(P,[arg_sujet],P2,[arg_sujet]),

imerge(P,[arg_objet],P2,[arg_objet]).

soit la formule sémantique:

drs (arg_sujet(X1),arg_objet(X2), cour(cond([imp(drs(cond([camion(X1])),univ([X1])), drs(cond([charge(X2),transporte(X1,X2])),univ([X2]))]))) correspondant à la phrase *"tout camion transporte une charge"*.

C'est à partir d'une telle formule que commence la résolution anaphorique.

3.Le lexique-grammaire

Le lexique-grammaire, représenté sous forme de tables (matrices composées de colonnes

et de rangées), contient les phrases simples, les différents emplois verbaux et les propriétés qui leur sont associées: nature sémantique des arguments, transformations possibles et leurs conditions, nombre et structure des compléments, type de la préposition associée etc.

Il considère la nominalisation comme la transformation d'une phrase contenant un opérateur verbal, en une autre phrase contenant un opérateur nominal. On y introduit un verbe prädicativement vide -*verbe support*- dont le rôle est d'actualiser le substantif qui n'a pas de marques morphologiques susceptibles de le faire:

Luc complimente les acteurs

= *Luc fait des compliments aux acteurs*

3.1 L'entourage lexical

Nous nous limiterons dans cet article aux possessifs, sur lesquels [GUIL 81], [DANL 80], [GROS 89] et [VIVE 83] notamment nous ont fourni des informations fort utiles.

L'examen des mots voisins du possessif est important, e.g. *Luc_i donne à Léa_j son_{i,j} argent*

Luc_i donne à Léa_j son_i amour

Dans le premier cas, *donner* est un verbe ordinaire (plein) alors que dans le second, il est un verbe support (*V-sup*). Seul le substantif *N2* change ([*NO donne à N1 N2_{poss}*]). Pourtant dans la première phrase, *son* peut référer à trois personnes:

- si l'argent est à Léa, *son* est relié à Léa;
- si l'argent est à Luc, *son* est relié à Luc;
- si l'argent n'est ni à l'un ni à l'autre, *son* est relié à une tierce personne du discours précédent.

Dans la seconde phrase par contre il n'y a qu'un référent: *Luc*, à cause du terme *amour* qui appartient aux mots "abstraits" ou de sentiments, pour lesquels on ne peut pas trouver, dans ce type de structure, d'autre relation à *son* que le sujet, ici *Luc*. Il s'agit de corréférence **obligatoire au sujet**.

Dans "*Luc_i choque Léa_j par ses_j idées*" et "*Luc_i choque Léa_j dans ses_j idées*", il y a le verbe *choquer*. Pourtant dans la première phrase *ses* est forcément relié à *Luc*. Il s'agit d'un cas de corréférence **obligatoire au sujet**, induit par la préposition *par*. Alors que dans la seconde, la corréférence est **obligatoire au complément d'objet**: *ses* est relié à Léa, à cause de la préposition *dans*.

3.2 Phrases construites autour d'un V-sup

On trouve des expressions verbales figées et d'autres mettant en jeu la paire *Vsup/Npred* mais sans être des expressions figées.

Les **expressions verbales figées** sont de la forme [*NO V N1-poss*], construites autour de variantes aspectuelles et d'opérateur causatif du verbe *avoir*, dont le *N1* est toujours "partie du corps" ou "abstrait" ("*Luc_i retient ses_j larmes*"), et dont la transformation en gn est impossible. Elles peuvent aussi être complétées par un troisième argument ([*NO V N1 Prép N2*] avec *V support* ou non. Là encore, le nom (*N1* ou *N2*) déterminé par le possessif est "partie du corps" ou "abstrait", et aucune restructuration n'est possible: *Luc_i jette son_i dévolu sur Léa*

Luc_i garde Léa sous sa_j protection

Si le possessif détermine l'objet direct dans une structure [*NO V N1-poss Prép N2*], [*Prép N2*] peut être remplacé par une complétive l'infinitive, donnant [*NO V N1-poss Prép Vinf*]:

Luc_i passe son_i temps au travail

Luc_i passe son_i temps à travailler

Dans toutes ces phrases, la corréférence est toujours **obligatoire au sujet**.

Les **expressions non figées** sont construites autour de verbes support ou de variantes aspectuelles [*NO V N1-poss*] ou [*NO V N1-poss Prép N2*]: *Luc_i a perdu ses_j illusions (sur Léa)*.

On ne peut pas transformer [*NO V Npred (Prép N1)*] en [*NO V Npred de N2 (Prép N1)*] (**Luc a perdu les illusions de Paul sur Léa*). En d'autres termes, l'argument *NO* de *Nprep* est le sujet du verbe, mais ce verbe ne peut prendre une expression [*Npred Prép NO*] comme complément (ici: *les illusions de Paul*).

La présence ou non du complément d'objet indirect n'a pas d'incidence sur la relation qui lie le possessif au sujet. Ces phrases ne se construisent pas avec une infinitive.

Certaines expressions non figées se construisent avec le possessif comme déterminant de *N3*; dans ce cas, les trois arguments sont obligatoires et la corréférence n'est pas obligatoire au sujet, Il s'agit de cas de **non-corréférence obligatoire à l'objet** ("*Luc_i met Léa à sa_{i,j} disposition*").

On peut restructurer en reliant les compléments par *être*: *Léa est à la disposition de (Luc, Paul)*

En résumé, parmi les phrases construites autour de *Vsup*, l'adjectif possessif qu'elles contiennent n'est jamais corréférent à l'objet,

mais toujours **au moins coréférent au sujet**. En reconnaissant d'emblée ces phrases et leur verbe comme *Vsup*, on pourra résoudre automatiquement l'anaphore. Pour cela nous proposons de repérer les autres phrases, puis de considérer les phrases non reconnues comme étant de cette catégorie

3.3 Phrases construites autour de verbes ordinaires

Elles s'articulent autour d'un verbe à un ou deux arguments, élément prédicatif de la phrase puisqu'il définit la structure des arguments. Il est déterminé par :

- le nombre d'arguments
- l'articulation syntaxique de ces arguments
- les traits sémantiques de ces arguments

Dans la structure [*NO V N1-poss*] que l'on ne peut poursuivre avec [*à/de N2*], la coréférence est obligatoire à un **autre nom que le sujet** *NO* (du discours antérieur).

Luc approuve son choix

Luc approuve le choix de Léa

Par contre, si une phrase a deux arguments et qu'elle peut être complétée par un troisième, la présence ou non de ce dernier fait varier la coréférence, ou tout au moins la préférence entre les antécédents possibles :

Luc_i avoue son_i dépit.

Luc_j avoue son_{i+j} dépit à Léa_k

La relation de coréférence existe toujours entre le possessif et le sujet dans les phrases simples [*NO V N1*], ou complétées par [*Prép N2*] où la relation peut aussi exister entre le *poss* et un référent du discours antérieur. On a donc là un cas de **non-coréférence obligatoire à l'objet**. Tous les verbes qui donnent ces résultats dans une telle structure appartiennent à la table 9.

Quand le syntagme prépositionnel est obligatoire, on distingue les phrases où l'adjectif possessif détermine le *N1* et celles où il détermine le *N2*. Parmi les premières, on trouve une **coréférence obligatoire au sujet** lorsqu'il y a possibilité de pronominalisation :

Luc_i consacre sa_i vie à la peinture.

Luc se consacre à la peinture.

ou de verbalisation simple :

Luc_i accorde son_i pardon à Léa_k

Luc pardonne à Léa

Mais pour celles dont la transformation donne une complétive infinitive [*nO V N1 Vinf*], la coréférence est **obligatoire à l'objet**

Luc_i motive Léa_j dans son_j travail

Luc motive Léa à travailler

Tous ces verbes appartiennent à la table 11

Parmi les phrases de structure [*NO V N1 Prép N2-poss*], la relation est obligatoire au **sujet**, ou obligatoire à l'**objet**. Dans les exemples suivants (verbe de la table 4), où la restructuration est possible en [*N2 de NO V N1*], la relation est établie entre le possessif et le **sujet**, lorsque la préposition est **par** ou **avec** : *Luc_i choque Léa_j par/avec ses_j idées.*

Les idées de Luc choquent Léa.

Par contre, la même phrase avec **dans** ou **pour**, par exemple, donne la transformation [*NO V N2 de N1*], et on établit alors la relation entre le possessif et le **complément d'objet**.

Luc_i choque Léa_j dans ses_j idées.

Luc choque les idées de Léa.

4. Architecture générale du système

Notre système se compose donc d'un **analyseur syntaxique** qui donne des arbres à partir desquels un **analyseur sémantique** produit des DRS. C'est sur elles qu'opérera le programme de résolution anaphorique. Ce système se complète d'un **fichier de verbes par tables**, et d'un **fichier "fonctionnel"**, constitué à mesure de l'analyse sémantique, où sont stockés tous les noms et pronoms, et leur fonction grammaticale.

La procédure de résolution, après repérage des pronoms, commence par une recherche des verbes (à chacun est associé un trait pour sa table d'appartenance) et de leur structure, dans le fichier lexique-grammaire. Si cela est trouvé, on cherche si la coréférence est obligatoire. Si oui, le traitement est terminé. Sinon, il faut activer **d'autres filtres** syntaxico-sémantiques : en partant des listes ordonnées de pronoms, univers et conditions, on vérifie la compatibilité de fonction⁴, de genre et de nombre, sémantique⁵. Il faut parfois chercher le verbe dans plusieurs tables. Si l'identité de structure entre le texte et les tables n'est pas établie, on examine **l'entourage** substantival du verbe : si l'objet est concret, il faut activer les autres filtres. S'il est abstrait ou "partie du corps", on a affaire à une phrase à verbe support (peut-être figée) dont le statut induit la (non) coréférence obligatoire.

5. Conclusion et perspectives

Remarques sur ces travaux :

4 Un candidat sujet est préféré à un autre pour être relié à un pronom sujet, dans deux phrases dites parallèles.

5 Des traits sémantiques sont associés aux mots lexicaux.

- Tous les verbes donnant lieu à un type de construction particulier, appartiennent à la **même table**. Sur chaque table, on peut contraindre la relation de (non) coréférence entre le possessif et un argument du verbe.

-Tous les compléments d'objet déterminé par le possessif sont "abstrait" ou "partie du corps". Avec d'autres noms, non prédicatifs, les verbes sont ordinaires et on ne peut résoudre l'ambiguïté anaphorique.

Donc en ajoutant ces caractéristiques dans la table en question on peut résoudre automatiquement l'anaphore. Ceci confirme l'hypothèse que toutes les tables peuvent, *a priori*, être ainsi complétées par les spécificités liées à l'emploi d'adjectifs possessifs et permettre ainsi au système d'éviter d'autres filtres plus coûteux en calcul et pas toujours fiables. Il faut donc établir un lexique-grammaire des verbes anaphoriques, (pris dans une structure mettant en jeu un pronom ou, ici, un adjectif possessif).

Dans la table 4 du lexique-grammaire, par exemple, nous répartissons les verbes en :

groupe 1: verbe+par (ou avec et parfois de), et coréférence obligatoire au **sujet** (N0): *Luc, déprime Léa, par son, attitude.*

groupe 2: verbe+dans, et coréférence **obligatoire à l'objet**. *Luc, déforme Léa, dans ses, propos.*

Nous avons ajouté à la table une colonne concernant la présence ou non de [prép N2], divisée en deux colonnes: les deux cas de coréférence obligatoire.

Les verbes de la table 11 sont répartis en :

Groupe 1: V N1 dans [N0 V N1 à V0inf] peut être remplacé par **pronominalisation** du verbe. Le possessif est forcément coréférent au **sujet**.

Luc consacre sa vie (à dessiner) au dessin

Luc se consacre (à dessiner) au dessin

Si N1 n'est pas abstrait la pronominalisation est impossible: (**Luc consacre son hangar au dessin*):

Groupe 2: Même structure sans pronominalisation. Il peut y avoir simple **verbalisation**. Le possessif est toujours coréférent au **sujet**. N1 est toujours abstrait et permet la verbalisation:

Luc apporte son soutien à cette affaire

Luc soutient cette entreprise

Groupe 3: Verbes qui, dans la transformation de [N0 V N1 à/pour V0inf] en [N0 V N1 à/dans N2] où N2 est déterminé par un possessif, induisent obligatoirement une

coréférence à N1 (*sujet*). Il s'agit là de **nominalisations**.

Luc stimule Léa pour travailler

Luc stimule Léa dans son travail

Nous avons étendu cette table en y ajoutant une caractéristique: existence ou non d'une structure pronominale, subdivisée en trois, pour les structures phrastiques étudiées et l'obligation de coréférence qui leur correspond.

Bibliographie:

- [CARB 88] J. CARBONELL, R. BROWN
Anaphora Resolution: a multi-Strategy Approach, COLING, Budapest, Aug 1988
- [CHAN 87] T. CHANIER, D. GELAIN, C. SEDOGBO
Une Approche Unifiée de la Syntaxe et de la Sémantique, Congrès AFCEI, Sofia Antipolis, Nov 1987
- [DANL 80] L. DANLOS
Représentations d'informations linguistiques: constructions N être prep X, Thèse de 3ème cycle, LADL Paris 7, 1980
- [GELA 91] B. GELAIN
Thèse de doctorat (à paraître) Paris 7, 1992
- [GROS 89] G. GROSS
Désambiguïsation sémantique à l'aide d'un lexique grammair, LADL et Univ. Paris 7, SEMANTICA, Paris Juin 1989
- [GROS 75] M. GROSS
Méthodes en syntaxe, Ed. Hermann, Paris 1975
- [GUEN 85] F. GUENTHER, P. SABATIER
Formal Semantics and Knowledge Representation, FNS, Université Tubingen, Déc 1985
- [GUIN 85] R. GUINDON
Anaphora Resolution: Short-Term Memory and Focusing, MCC Austin, A.C.L. 1985
- [GUIL 81] A. GUILLET, C. LECLERE
Formes syntaxiques et prédicats sémantiques, Langages n°63, Ed Larousse, Paris 1981
- [KASP 85] W. KASPER
Montague Grammar, Situation semantics and Discourse Representation Theory, Rapport ACORD, Univ. Stuttgart, Mai 1986
- [RICH 88] E. RICH, S. LUPERFOY
An Architecture Program for Anaphora Resolution, MCC Austin, A.C.L., fév 1988
- [SEDO 87] C. SEDOGBO
De la Grammaire en Chaîne du Français à un Système Question-Réponse, Thèse de Doctorat d'Etat, Université de Marseille, 1987
- [VIVE 83] R. VIVES
Avoir, vendre, partir: constructions à verbe support et extensions aspectuelles, Thèse de 3ème cycle, LADL Paris 7 1983
- [WADA 87] H. WADA, N. ASHIER
A Computational Account of Syntactic, Semantic and Discourse Principles for Anaphora Resolution, University Texas, Austin, Jan 1987