DES HEURISTIQUES POUR LA RECHERCHE DU THEME D'UN DISCOURS ET DE L'ANTECEDENT D'UN PRONOM

Monique Rolbert

Groupe d'Intelligence Artificielle
Faculté de Luminy case 901
route Léon Lachamp
13009 Marseille
FRANCE
tel : 91. 26. 90. 70

Résumé

Un des problèmes résiduels pour le traitement des références dans les interfaces en langage naturel est le grand nombre d'ambiguités que génère un pronom du point de vue de la recherche d'antécédent. Dans cet article, nous allons montrer comment l'utilisation de critères issus d'études de psychologie expérimentale sur des méthodes de construction d'un discours par des locuteurs peut apporter un plus pour résoudre ce problème. Nous présentons tout d'abord des résultats de tests faits par des psychologues autour de la notion de thème et de représentation interne du discours ; puis, nous utilisons ces résultats pour énoncer un certain nombre de critères pragmatiques concernant la recherche d'antécédents. Nous montrons enfin que ces critères, tout étant concis et facilement programmables, sont assez généraux au regard de ceux présentés dans des cadres similaires.

I - Introduction

Identifier l'antécédent d'un pronom dans un texte est un processus complexe à réaliser de manière automatique. Un des problèmes résiduels pour sa résolution dans un système informatique d'interface en langage naturel est le grand nombre d'ambiguïtés que génère l'emploi d'un pronom. Dans un premier temps, on peut être tenté de n'utiliser que des critères syntaxiques et sémantiques, car ils ont l'avantage d'être rigoureux. Cependant, ils sont en général insuffisants pour identifier de manière unique un référent. Les critères syntaxiques (du type accord en genre et en nombre ou c-commande (voir [Reinhart 81])) sont des filtres, c'est à dire qu'ils éliminent des candidats plutôt qu'ils ne désignent précisément le syntagme nominal antécédent. Ces filtres sont fixés pour un sous-ensemble d'une langue donnée et ne s'appliquent pas à tous les types de pronom (voir pour une étude plus précise [Rolbert 89]); Ils ont l'avantage d'être rigoureux mais ne sont pas extensibles. Les critères sémantiques sont des règles d'accord de traits ou des règles basées sur des notions de présuppositions - le choix d'un antécédent qui donne une solution sémantique absurde est rejeté. Leur efficacité dépend étroitement de la précision de la sémantique du domaine d'application. Pour des domaines sémantiquement pauvres, ils seront assez inefficaces.

Il est donc nécessaire d'introduire d'autres critères, indépendants du domaine d'application du système, permettant de choisir l'antécédent d'un pronom parmi les syntagmes nominaux que la syntaxe et la sémantique n'ont pas écartés. C'est pourquoi, dans la plupart des systèmes permettant l'emploi de pronoms, des critères pragmatiques sont aussi mis en œuvre pour tenter de résoudre ce problème ([Guenthner & Lehmann 83], [Danlos 85], [St Dizier 86], [Asher & Wada 87], [Sedogbo 87]...). Cependant, au delà de la justification de l'emploi de règles pragmatiques, il nous a semblé important de faire le point sur leur contenu, celui-ci étant rarement justifié. Pour cela, nous nous sommes basés sur des études faites sur la construction d'un discours chez un locuteur.Celles-ci mettent en évidence l'importance de la notion de *thème* pour notre approche et nous permettent d'énoncer 3 critères pour la recherche d'un thème dans un discours. Nous les appliquons ensuite, complétés par deux autres, à la recherche de l'antécédent de pronoms puis les comparons à ceux habituellement utilisés dans les systèmes existants et montrons que l'ensemble de ces 5 critères forme un tout complet et général.

II - Des expériences sur le discours

De nombreuses expériences sont faites par des psychologues dans le domaine de la caractérisation de modèles mentaux du discours. La problèmatique générale de ce type d'expérimentation est de comprendre comment un texte est analysé, compris et représenté dans la mémoire du lecteur. Les résultats s'appuyent principalement sur l'analyse de temps de réponse concernant la résolution d'anaphore suivant la prédominance du référent dans le discours et le type de critères mis en jeu ([Corbett & Chang 83], [Morrow 85], [Murphy 84] etc...).

On trouve plusieurs résultats mettant en relation le thème et la coréférence. [Corbette & Chang 83] avancent l'hypothèse qu'un syntagme nominal est plus accessible s'il a été par ailleurs 'mis en valeur dans le texte'. [Garrod & Sanford 85] indiquent également que la résolution d'une référence liant un personnage principal du texte est toujours plus rapide que des références similaires liant des personnages secondaires. On voit donc apparaître l'importance du thème. Comment détecter le thème d'un discours ? En fait, on voit que ce qu'il faut repérer n'est pas vraiment le thème du discours, mais un thème local, propre au lecteur (dans notre cas, au locuteur) mis en valeur dans sa propre représentation du discours. Des résultats d'expériences ([Mckoon & Ratcliff 80] cités dans [Corbett & Chang 83]) ont conduit à la conclusion qu'un nom qui a été repris par un pronom dans la dernière phrase d'un texte est plus accessible dans le modèle du discours d'un lecteur qu'un nom qui n'a pas été référencé. [Morrow 85] va plus loin en disant que la prédominance d'un nom pour un lecteur (donc son thème) est déterminée par des propriétés de surface telles que l'ordre d'apparition ou la fréquence de mention d'un objet coréférent à ce nom.

Dans une analyse presque similaire, la notion de sujet est jugée pertinente du point de vue du thème. A partir d'expériences faites sur l'évaluation d'un temps de réponse à propos de personnes citées dans un texte, après lecture des deux phrases formant ce texte, la deuxième contenant un pronom référent à un élément variable de la première ([Corbett & Chang 83]), les auteurs déduisent que le sujet (en temps que rôle fonctionnel tenu par un syntagme) de la première phrase est plus accessible à la fin du texte que le nom prédicat (ici, un complèment d'objet direct) de cette même phrase, indépendemment des effets de la référence. Le sujet d'une phrase serait donc prédominant dans la représentation du discours. Ils donnent comme explication possible de ce phénomène que le sujet fonctionne comme position topicalisée de la phrase (c'est-à-dire qu'il représente ce sur quoi parle la phrase).

En ce qui concerne la résolution de référence propement dite, on retrouve dans de nombreuses études une hypothèse basée sur la proximité entre le pronom et son référent. Ces expériences ont montré qu'un antécédent est plus accessible lorsqu'il apparaît dans la phrase précédent immédiatement celle contenant le pronom. Nous verrons l'intérêt de ce résultat indépendant de la notion de thème en III.2.

En conclusion, nous pouvons dire que bien que les buts de ces expériences soient différents du nôtre, les résultats qu'elles permettent d'obtenir sont intéressants pour plusieurs raisons :

- la notion de thème est assez ancienne en linguistique (voir les thèses de Mathésius et de l'école de Prague dans [Firbas 64]) mais cette approche expérimentale, si elle ne se soucie pas des problèmes essentiellement linguistiques qui entourent cette notion (entre autre un problème de définition - voir en particulier à ce sujet et sur la place du thème en intelligence artificielle [Maradin 88]), aborde une analyse plus pragmatique des faits.

- la plupart des expériences que nous citons sont basées sur des problèmes de coréférence. Leurs conclusions finales sont orientées vers la modélisation d'une représentation du discours ; cependant, des résultats intermédiaires présentent des hypothèses sur des processus inférentiels concernant la coréférence.

- ces expériences portent sur le comportement d'un lecteur, mais il nous semble que nous pouvons appliquer leurs résultats à celui d'un locuteur : un locuteur construit un discours pour se faire comprendre. Il applique donc des règles de bonne formation du discours lors de la construction. Par exemple, il va employer un pronom uniquement s'il lui semble que celui-ci reprend de manière non (ou peu) ambigüe un élément prépondérant de sa propre représentation interne du discours.

- certains concepts introduits pour expliquer les mécanismes décrits (du type distance entre mots) ne sont pas très complexes : ce sont des notions informatisables sans trop de difficultés.

- enfin, ces résultats sont basés sur des expérimentations. Ils ont donc toute la puissance de 'données collectées sur le terrain'. Ces expériences peuvent remplacer des études sur des corpus que les informations ont du mal à faire (peu de systèmes en Langage Naturel en ligne).

III - Utilisation de ces résultats pour la recherche automatique de référent

Nous nous situons donc dans le cadre de l'analyse automatique d'un texte contenant des pronoms ayant pour antécédents des syntagmes nominaux (nous écartons dans un premier temps d'autres types de référence - emploi déictique des pronoms, référence à des propositions, référence temporelle, etc...). Le problème que nous nous posons est de déterminer quel est l'antécédent d'un pronom parmi les syntagmes nominaux présents dans le texte.

A partir des données expérimentales que nous venons de présenter, nous allons élaborer tout d'abord des critères permettant de retrouver les thèmes d'un discours.

III.1 - Des critères pour la recherche d'un thème dans un discours

Un premier critère que nous allons utiliser apparait à plusieurs reprises à travers les expérimentations que nous avons évoquées. Il semble, en effet, que la répétition d'un syntagme nominal (que ce soit une simple répétition ou une reprise par un syntagme pronominal) joue un rôle très important dans la mise en valeur d'un objet dans un discours. On remarque en particulier que la répétition d'un nom propre peut être une manière de changer de thème (puisqu'en général, elle est motivée - par rapport à l'emploi d'un pronom- par le fait que le personnage nommé n'est pas le thème courant ([Grosz & al 87])). Nous énoncerons donc un premier critère :

> Critère 1 : *Un syntagme nominal est mis en valeur dans un discours à travers les reprises dont il fait l'objet.*

Ce rapport entre la densité des liens anaphoriques et la manifestation du thème du discours a d'ailleurs été souligné dans de nombreux travaux de linguistes ([Beaugrande & Dressler 1981], [Sidner 83]...)

D'autre part, à l'intérieur d'une phrase certains syntagmes sont plus prépondérants que d'autres ; en linguistique, on parle d'emphase, d'insistance. La position d'un syntagme à l'intérieur d'une phrase semble aussi être un critère pour catégoriser celui-ci comme thème de la phrase. Nous reprenons ici les idées développées autour de la fonction prédominante du sujet dans les expériences citées, en remarquant que cette notion est adéquate si l'on ne considère que des phrases déclaratives sans forme de focus marquée. Nous allons la généraliser à des phrases interrogatives et des phrases comportant des syntagmes en position de focus. Notre idée est la suivante :

"un objet est mis en valeur dans une phrase lorsqu'il apparaît en tête de cette phrase et, de manière intuitive, au niveau le plus haut."

Ceci est vrai pour des interrogatives : l'objet sur lequel porte une interrogative simple est en général en tête de la phrase.

Exemple : De quel homme Max parle-t-il?
(Which man is Max speaking about ?)

On retrouve ce phènomène pour les formes syntaxiques de focus qui sont souvent des 'montées' dans l'argument TOP.

Exemple : C'est de la fille de Lili que Max parle.
(It's about Lili's daughter that Max is speaking)

Il nous semble donc que cette approche tient réellement compte de ce que l'on peut appeler le 'topic' de la phrase. Certaines différences plus subtiles entre deux positions syntaxiques sont également prises en compte en fonction de la profondeur (différence d'importance entre un syntagme et son complément de nom, moins grande importance des informations introduites au niveau d'une relative etc...). De plus,

pour des phrases déclaratives, le syntagme nominal sujet de la phrase est bien le syntagme de tête le plus 'haut' de la phrase. Notre critère est donc bien une généralisation de ce phénomène.

En se basant sur une représentation du type S-structure ([Chomsky 82]) de la phrase, ces considérations peuvent se résumer en :

Critère 2 : *les syntagmes nominaux mis en valeur dans une phrase sont ceux dont les nœuds correspondant dans la S-structure de la phrase sont situés le plus haut et le plus à gauche de l'arbre.*

Pour permettre un changement de thème dans un discours (ce qui peut arriver couramment), nous tiendrons également compte de l'ancienneté de la phrase dans laquelle le syntagme apparait par rapport au développement du discours.

Critère 3 : *Un syntagme nominal récemment introduit est prédominant par rapport aux syntagmes nominaux plus anciens dans le discours.*

Les critères 1, 2 et 3 permettent donc de déterminer le thème d'un discours. Voyons leur utilisation possible dans la recherche de l'antécédent d'un pronom.

III.2 - Application à la levée d'ambiguïté dans le traitement des références

Pour choisir le référent d'un pronom parmi les solutions non écartées par la syntaxe et la sémantique, on peut se baser sur le principe suivant : 'Un pronom a pour référent de préférence le thème du discours'. On peut alors directement utiliser les critères 1-3. Cependant, d'autres critères spécifiques à la recherche d'antécédent et indépendants de la notion de thème doivent être également pris en compte. Nous allons en introduire deux en faisant le point sur ce qui est proposé habituellement dans les systèmes informatiques s'intéressant à ce problème.

Certains systèmes informatiques intègrent déjà des heuristiques concernant la recherche d'antécédent. Parmi les règles qui reviennent le plus souvent, on peut citer :

a) Les syntagmes nominaux dans une proposition principale sont préférés à ceux inclus dans des subordonnées ([St Dizier 86], [Guenthner & Lehmann 83])

b) Le syntagme nominal sujet est plus souvent repris que les autres (correspondant au Grammatical Function Filter de [Asher & Wada 87], cité également dans [Guenthner & Lehmann 83])

c) Un antécédent est plus probable pour une occurence de pronom donné si le pronom et l'antécédent ont le même rôle syntaxique (correspondant au Parallelism Filter de [Asher & Wada 87]. Utilisé également dans [Sedogbo 87]. Correspond au critères de relations formelles entre phrases utilisé pour la synthèse dans [Danlos 85])

d) L'antécédent d'un pronom le plus probable est celui qui est le plus près ([St Dizier 86] et Principe de Proximité dans [Guenthner & Lehmann 83])

Remarquons d'abord que notre critère 2 est une généralisation de a) et b). En ce qui concerne le critère de parallélisme c), il ne nous semble justifié que dans des phrases dont les structures sont très marquées et dont la similarité est mise en évidence par des mots tels que 'aussi', 'également'. Nous ne l'utiliserons pas ici. Seul donc d) nous donne un principe supplémentaire, détaché de la notion de thème (donc non redondant avec les hypothèses que nous avons faites précédemment) et spécifique à la recherche d'antécédent. Nous l'avons de plus rencontré dans certains résultats des expériences que nous avons exposées (voir page 3). Nous posons donc un quatrième critère :

critère 4 (Principe de Proximité) : *Un syntagme nominal proche d'un pronom est prédominant comme référent par rapport aux syntagmes nominaux plus éloignés dans le discours.*

Ce critère n'est pas redondant avec le critère 3 puisqu'il est spécifique à la recherche d'antécédent et lié à la position de chaque pronom.

Pour le cas particulier de la cataphore, on peut utiliser une hypothèse présentée dans [Guenthner & Lehmann 83] :

Critère 5 : *Les syntagmes nominaux précédant le pronom sont prédominant comme référent par rapport à ceux suivant le pronom (la référence en arrière est préférée à la cataphore)*

IV - Conclusion

A partir de ces critères, il est donc possible de choisir le syntagme nominal antécédent d'un pronom parmi ceux que la syntaxe et la sémantique n'ont pas écartés. Remarquons que le critère 1 est totalement original et que le critère 2 est une généralisation justifiée de ce qui est utilisé en général. On peut dire que par l'ensemble des informations qu'ils prennent en compte et par leur modularité, les critères que nous proposons peuvent jouer un rôle intéressant dans un système informatique.

Il reste cependant à évaluer quel est le poids de chacun d'eux par rapport aux autres. Sont-ils tous de même importance ? Par exemple, le critère 5 semble plus déterminant que le Principe de Proximité : entre un syntagme nominal très proche du pronom mais qui se trouve après et un syntagme qui se trouve plus loin mais avant, on préférera sans doute celui qui se trouve avant. Leur rôle peut dépendre également du type du pronom (par exemple, le pronom 'celui-ci' a un comportement particulier). En fait, il ne nous semble possible de répondre à cette question qu'après avoir réellement testé ces critères sur des textes en faisant varier leur influence réciproque.

Remarquons également que les solutions que nous mentionnons ici pourraient être intégrées dans un traitement plus élaboré du discours. Citons les recherches faites sur des modèles informatiques de représentation du discours, approche qui semble intéressante bien que la notion de modèle du discours ne soit pas encore vraiment définie. Parmi ces modèles, la DRT ([Kamp 84]) qui grâce à la notion de liste d'accessibilité et de sous-discours est un pas vers l'intégration de nouveaux critères dans la recherche d'antécédent. Il faut également citer les travaux de linguistes sur les phénomène d'"empathie" d'un discours et leur corrélation avec des phénomènes syntaxiques et la forme de surface d'une phrase. Les concepts introduits dans [Kuno & Kaburaki 77] ('point de vue' du locuteur à travers la manière dont il s'exprime, verbe orienté sujet et verbe orienté objet, hiérarchie dans l'"empathie" suivant des critères syntaxiques etc.) nous semblent intéressants.

BIBLIOGRAPHIE

[Adjémian 78] Adjémian C. "Theme, Rheme and Word Order. From Weil to Present Day Theories" *Historiographia Linguistica* Vol 5 N° 3 (1978) pp253-273

[Asher & Wada 87] Asher N. et Wada H. "A computational Account of Syntactic, Semantic and Discourse Principles for Anaphora Resolution" Preliminary Draft

[Austin 62] Austin J.L "How to Do Things with Words" ed. by J.O Urmson.
N.Y. Oxford University Press (1962)

[Beaugrande & Dressler 81] Beaugrande (de) et Dressler W *Introduction to Text-Linguistics* Longman London (1981)

[Chafe 76] Chafe W.L "Giveness, Contrastiveness, Definitness, Subjects, Topics and Point of View" in *Subjecct and Topic* Li Ch. Ed. Academic Press (1976) pp 27-55

[Chomsky 82] Chomsky N. "Some Concepts and Consequences of the Theory of Government and Binding" *Linguistic Inquiry monograph*, N° 6
Cambridge, Mass., MIT Press (1982)

[Corbett & Chang 83] Corbett A.T et Chang F.R "Pronoun Disambiguation : Accessing potentials antecedents" *Memory And Cognition* 11(3), pp 283-294 (1983)

[Danlos 85] Danlos L. *Génération automatique de Texte en Langage Naturel* Masson (1985)

[Ducrot 72] Ducrot O. *Dire et ne pas Dire* Hermann (1972)

[Firbas 64] Firbas J "on defining the theme in functionnal sentence analysis" Travaux linguistiques de Prague, 1 Prague 64

[Garrod & Sanford 85] Garrod S. et Sanford A.J "On the Real-time Character of Interpretation during Reading" *Language and Cognitive Processes* Vol N°1, pp 43-59 (1985)

[Grosz, Joshi & Weinstein 87] Grosz B., Joshi A.K et Weinstein S. "Towards a Computational Theory of Discourse Interpretation" Preliminary Draft

[Guenthner & Lehmann 83] Guenthner F. et Lehmann H. "Rules for Pronominalisation" *ACL European Chapter* (1983)

[Guéron 79] Guéron J. "Relation de Coréférence dans la Phrase et le Discours" *Langue Française* 44 pp 42-79 (1979)

[Hagège 78] Hagège C. "Du Thème au Thème en Passant par le Sujet. Pour une Théorie Cyclique" *La Linguistique* Vol 14 N° 2 pp 3-38 (1978)

[Jackendoff 72] Jackendoff R.S *Semantic Interpretation in Generative Grammar* MIT Press Cambridge (1972)

[Kamp 84] Kamp H. "A Theory of Truth and Semantic Representation" *Truth Interpretation and Information*, Groenendijk et als (eds), Foris (1984)

[Kuno & Kaburaki 77] Kuno S. et Kaburaki E. "Empathy and Syntax" *Linguistic Inquiry* Vol. 8 N°4 (fall 1977) pp 627-672

[Marandin 88] Marandin J.F "A Propos de la Notion de Thème de Discours. Eléments d'Analyse dans le Récit" *Langue Française* 78 pp 67-128 (1988)

[McKoon & Ratcliff 80] McKoon G. et Ratcliff R. "The Comprehension processes and Memory Structures involved in Anaphoric Reference" *Journal of Verbal Learning and Verbal Behavior* 19 pp 668-682 (1980)

[Morrow 85] Morrow D.G "Prépositions and Verbe Aspects in Narrative Understanding" *Journal of Memory and Language* Vol 24, pp 390-404 (1985)

[Murphy 84] Murphy G.L "Establishing and Accessing Referents in Discourse" *Memory and Cognition* 12(5), pp 489-497 (1984)

[Reinhart 83] Reinhart T. "Coreference and Bound Anaphora : A Restatement of the Anaphora Questions" *Linguistics and Philosophy* Vol. 6 (1983) pp 47-88

[Rolbert 89] Rolbert M. "Résolution de formes pronominales dans l'interface d'interogation d'une base de données" Thèse de doctorat. Faculté des Sciences de Luminy (1989)

[Saint-Dizier 86] Saint-Dizier P. "Résolution des anaphores et Programmation en Logique" Papier Préliminaire

[Sedogbo 87] Sedogbo C. "SYLOG : A DRT System in Prolog" *Second International Workshop on Natural Language and Logic Programming* Simon Fraser University, Vancouver, B.C Canada (1987)

[Sidner 83] Sidner C. "Focusing in the comprehension of definite Anaphora" in *Computational Models of Discourse* Brady & Berwick eds. MIT Press 1983 pp 267-329