L'anaphore à antécédent flou : une caractérisation et ses conséquences sur l'annotation des relations anaphoriques

Frédéric Landragin – CNRS, Laboratoire LaTTICe, Paris

La résolution des anaphores dans les systèmes de traitement automatique des langues passe par l'identification des expressions anaphoriques, celle de leurs antécédents potentiels, et l'attribution de relations entre les entités du discours ou les segments textuels retenus. Or il arrive que l'identification précise de l'antécédent soit difficile, voire impossible. C'est typiquement le cas des anaphores abstraites : dans une narration qui se termine par « *ça arrive* », « *ça* » reprend-il la phrase précédente, seulement sa tête verbale, ou tout le début du paragraphe ? L'entité du discours et le segment textuel qui la décrit peuvent ainsi varier grandement, sans qu'on sache exactement quels sont les éléments sémantiques couverts. Il arrive également que l'identification précise de l'antécédent soit inutile à la compréhension du sens global, et qu'il soit judicieux de laisser actifs plusieurs antécédents potentiels. C'est le cas de certaines anaphores pronominales telles que « *le N_1 de le N_2 […] il […]* » où « *il* » peut aussi bien reprendre « *le N_1 de le N_2* » que « *le N_2* ». Nous voulons montrer ici que choisir entre N_1 et N_2 n'est pas forcément pertinent, et qu'il existe au contraire des situations où le doute doit être laissé, non seulement parce que le lecteur ou interlocuteur a du mal à choisir, mais aussi et surtout parce que ce choix n'est pas nécessaire à une bonne compréhension du texte. Autrement dit, il existe des anaphores à antécédent « sous-spécifié », « vague » ou encore « flou », et nous tenterons de les caractériser puis de montrer comment les représenter en utilisant et affinant les principes de l'annotation des relations anaphoriques. L'objectif de ce travail est de mieux rendre compte des phénomènes d'interprétation partielle dans le discours, au niveau des anaphores.

Caractérisation des situations impliquant un antécédent flou

Nous nous focaliserons sur les pronoms personnels et nous laisserons de côté les phénomènes plus complexes telles que les anaphores abstraites ou associatives pour lesquelles la notion d'antécédent flou est peut-être encore plus pertinente. Le but est de montrer que même le classique et déjà très étudié « *il* » (Kleiber, 1994) peut impliquer un antécédent flou :

(1) « Je lui ai dit sur un ton de plaisanterie que son idée était intéressante, qu'**elle** montrait les choses sous un angle auquel en effet on n'est pas habitué […] » Frantext, K899 (Romains, *La douceur de la vie*, p. 213).

(2) « Marcello devient révolutionnaire, quitte un beau jour à craquer inexorablement, tant il est clair que son engagement relève du pur volontarisme et qu'**il** est miné par un manque de conviction intime. » (tiré de la préface de Moravia, *Le conformiste*, p. 38).

(3) « […] dans le comportement de tout public quel qu'**il** soit » Frantext, R875 (Gracq, *Préférences*, p. 34).

(4) « Le travail de l'intellect naît d'une insatisfaction profonde devant le monde, mais **il** s'avère incapable d'y porter remède […] » (adapté de la préface de Moravia, *Le conformiste*, p. 37).

(5) « […] les rapports des rapporteurs et les documents auxquels **ils** se réfèrent, pour autant qu'**ils** n'ont pas été communiqués antérieurement, sont […] » Frantext, P659 (*Le Conseil Société des Nations*, p. 107).

(6) « […] cette faille, cette malédiction qui est comme une résurgence laïque du péché originel, est d'autant plus grave qu'**elle** laisse un vide […] » (tiré de la préface de Moravia, *Le conformiste*, p. 38).

(7) « Voulez-vous Jean Dupont comme époux ? – Oui, je **le** veux ».

(8) « Jean a coupé tout le bois. Il **l**'a fait en petits morceaux » (exemple que M. Prandi interprète comme une anaphore individuelle, en avançant l'argument que « faire » est le verbe principal et non une proforme).

(9) « L'homme doit choisir mais il ne parvient pas à **le** faire ».

Ces exemples mettent l'accent non seulement sur les groupes nominaux de type « *le N_1 de le N_2* », mais aussi sur les possessifs, les coordinations et les juxtapositions. A chaque fois, l'attribution précise de l'antécédent n'a pas une grande importance. Dans (1), la personne autant que son idée peut montrer les choses sous un angle particulier : même si les deux représentations sémantiques diffèrent, elles transmettent le même sens global car dans ce contexte phrastique on peut assimiler la personne à son idée. Dans (2), on peut également assimiler le personnage à son engagement : les deux peuvent être minés par un manque de conviction. Dans (3) ou (5), une analyse fine peut être réalisée pour identifier le bon antécédent, mais son résultat reste peu sûr et peut-être vaut-il mieux laisser le doute. Dans (6), la précision et la correction sont deux hypothèses qui peuvent être maintenues. Cette imprécision des juxtapositions peut d'ailleurs amener à des exemples où c'est le premier constituant et non le second qui est conservé comme antécédent potentiel : « le rêve, l'évasion est nécessaire. **Il** permet de mieux supporter les tracas de la vie quotidienne ». Dans tous les cas, la résolution exacte de l'anaphore n'est pas nécessaire pour continuer à lire à et comprendre le texte. La remarque de (Prandi, 1987) à propos des ambiguïtés syntaxiques liées à la portée des prépositions s'applique donc parfaitement à notre problème : « ce genre d'ambiguïté structurale, très fréquent dans l'expression nominale, reste d'ailleurs sans conséquences sur l'interprétation, du fait, probablement, de son caractère systématique et non annulable » (p. 135). Nous noterons également que les processus classiques de résolution des anaphores, en particulier la prise en compte de

la saillance pour classer les antécédents potentiels, sont ici inutiles : cf. (Miltsakaki, 2007) pour la saillance en général et (Kister, 1995) pour le classement de N_1 et N_2 dans les groupes nominaux de type « *<det> N_1 de <det> N_2* ».

En confrontant ces exemples avec d'autres tirés de la littérature scientifique, du corpus « Frantext », ou construits à partir de ceux-ci, nous proposons une classification des antécédents potentiellement flous :

1. **Les possessifs**, quand le possesseur est un être animé et que le possédé est un trait de personnalité auquel il peut s'assimiler, ou quand il s'agit d'un objet et de sa fonction principale (« *l'intellect et son travail* »). Plus généralement, et c'est là un point essentiel de notre caractérisation, les deux participants doivent être reliés par une relation de type « l'un est une facette de l'autre », ou encore, « l'un fait partie de l'ensemble des propriétés de l'autre ». C'est le cas dans les exemples (1) et (2), mais aussi dans (3) et (4).
2. **Les groupes complexes « *le N_1 de le N_2* »**, avec les mêmes participants que dans la situation précédente.
3. **Les coordinations**, dans le même cas que précédemment et dans les cas où les coordonnés aussi bien que le pronom anaphorique sont au pluriel : « *les N_1 et les N_2 […] ils […]* ».
4. **Les juxtapositions**, en particulier quand il est difficile voire impossible de distinguer une simple énumération d'une reformulation, et, dans ce dernier cas, de distinguer une précision d'une correction (cf. exemple 6). C'est aussi le cas des mentions telles que « *le premier secrétaire, M. Martin et sa femme* » quand le texte ne permet pas de comprendre si M. Martin est le premier secrétaire ou une tierce personne.
5. **Les référents évolutifs** : dans des exemples tels que « *Marcello adulte […] il […]* », « *il* » reprend-il « *Marcello* » ou « *Marcello adulte* » ? Il est parfois impossible (et inutile) de choisir.

A ces cas s'ajoutent ceux où l'antécédent n'est tout simplement pas mentionné. On pensera par exemple à « *ils ont encore augmenté les impôts* » (Kleiber, 1994), aux usages attributifs et non référentiels, ou encore aux exophores (anaphores sans antécédent linguistique). Concernant les anaphores abstraites, nous noterons deux cas supplémentaires : d'une part les ambiguïtés entre une référence individuelle et une référence à un événement (ou à un état, cf. exemples 7 et 8), d'autre part les ambiguïtés entre deux références événementielles (cf. exemple 9).

Conséquences sur l'annotation des anaphores à antécédent flou

Un moyen d'appréhender l'identification des antécédents est l'annotation de textes, en tant que mode de représentation et de confrontation des phénomènes (et non en tant que mode d'évaluation de systèmes de TAL). Or l'annotation des phénomènes décrits dans nos exemples pose plusieurs problèmes que nous proposons de gérer comme suit :

1. La détermination d'un antécédent (flou) est indispensable, sinon l'interprétation est amoindrie.
2. **Principe des alternatives** : dans la plupart des cas, plusieurs alternatives se confrontent, le choix restant indifférent : la personne ou son idée, l'ensemble des documents ou juste les rapports, etc. Il s'agira donc de spécifier quelles sont les alternatives, et de les regrouper en tant qu'antécédent. Autrement dit, l'antécédent est dans l'ensemble des alternatives, sans qu'on sache laquelle correspond à l'intention du locuteur (mais toutes sont plausibles). Nous noterons que ce principe d'alternative doit être clairement distingué du **principe de groupement** qui peut être mis en œuvre lors d'un pluriel (« *Jean dormait. Marie lisait.* ***Ils*** *étaient heureux* »).
3. **Principe des possibles** : dans certains cas, les alternatives sont difficiles à déterminer, par exemple concernant les juxtapositions. Il s'agira alors d'identifier toutes les possibilités et de les étiqueter comme alternatives potentielles. Autrement dit, l'antécédent est dans l'ensemble des possibilités, sans qu'on sache laquelle lui correspond exactement, et toutes n'étant pas forcément plausibles. Ainsi, dans l'exemple avec M. Martin, on aura un « possible » correspondant au groupement de « *le premier secrétaire, M. Martin* » (ou plus simplement « *M. Martin* ») et de « *sa femme* », et un autre « possible » correspondant au groupement de « *le premier secrétaire* », de « *M. Martin* » et de « *sa femme* ».
4. **Principe de double balisage** : pour les anaphores abstraites telles que « *ça arrive* », les alternatives et les possibles ne suffisent plus. Il faudrait pouvoir gérer des antécédents à limites floues ou progressives, ce qui s'avère incompatible avec les principes de balisage mettant en œuvre une balise ouvrante et une balise fermante. Nous proposons dans ce cas un double balisage, le premier en tant que limite inférieure de l'empan textuel, et le second en tant que limite supérieure. A ces limites correspondent des entités de discours, et toute entité à l'intérieur est considérée comme acceptable. Nous noterons que ce principe se distingue de l'attribut « MIN » proposé dans MUC-7 : celui-ci a été prévu pour les mentions de personnes afin de ramener toute expression référentielle à une mention minimale comparable au nom propre, et ajoute une caractéristique à une entité de discours, sans remettre en cause l'identité de celle-ci. Ce sont au contraire deux entités de discours distinctes qui sont en jeu dans notre anaphore à antécédent flou, chacune pouvant être caractérisée par une mention minimale indiquée par « MIN ». Le principe de double balisage est donc indépendant de l'attribut « MIN ».

Références bibliographiques

Kister, L. (1995) Accessibilité pronominale des *dét. N1 de (dét.) N2* : le rôle de la détermination. *Linguisticae Investigationes*, XIX (1), pp. 107-121.

Kleiber, G. (1994) *Anaphores et pronoms*. Duculot, Louvain-la-Neuve.

Miltsakaki, E. (2007) A rethink of the relationship between salience and anaphora resolution. *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium*, Lagos, Portugal, pp. 91-96.

Mitkov, R. *et al*. (2007) Anaphora resolution: to what extent does it help NLP applications? In: *Anaphora: Analysis, Algorithms and Applications*, Springer-Verlag, Berlin Heidelberg, pp. 179-190.

Prandi, M. (1987) *Sémantique du contresens. Essai sur la forme interne du contenu des phrases.* Minuit, Paris.