

Génération de descriptions définies et démonstratives

Hélène Manuélian

LORIA, France

helene.manuelian@loria.fr

1. Introduction

Les algorithmes de génération d'expressions référentielles génèrent essentiellement des descriptions définies (Dale et Reiter, 1995, Van Deemter, 2000, Van Deemter, 2001, Khramer et al., 2001). Le but de ces algorithmes est de produire, étant donné un référent présent dans le contexte, une expression référentielle suffisamment informative pour que l'auditeur puisse identifier son référent. Dans une perspective différente, Gardent et Striegnitz (Gardent et Striegnitz, 2000) présentent un algorithme pour la génération d'anaphores associatives basé sur un modèle structuré du contexte et la prise en compte de phénomènes inférentiels pour la génération.

Tous ces algorithmes génèrent des descriptions définies et pour la plupart, ils manipulent l'opposition entre description définie et pronom. En revanche, ils ne tiennent pas compte de la distinction entre les descriptions définies et les descriptions démonstratives. Les deux types de descriptions pouvant référer à des objets déjà mentionnés dans le contexte (Kleiber, 1986 et 1988, Corblin, 1987), il semble pourtant nécessaire de savoir comment on choisit le déterminant d'une expression référentielle anaphorique, et d'introduire la distinction entre les emplois des descriptions définies et des descriptions démonstratives dans les algorithmes de génération d'expressions référentielles.

Notre article constitue un premier pas dans cette direction : nous présenterons dans un premier temps les problèmes posés par la distinction entre le défini et le démonstratif. Nous présenterons ensuite une classification des descriptions définies et démonstratives, fondée sur une étude de corpus dont nous présenterons les résultats. Pour finir, nous donnerons des indications pour la conception d'algorithmes de génération tenant compte des données issues de notre étude.

2. Descriptions définies ou démonstratives et apport d'information

2.1. Données linguistiques

Le défini est généralement reconnu comme présupposant la familiarité et l'unicité du référent dans le contexte (Russel, 1905, Heim, 1982). Cette propriété est d'ailleurs utilisée en génération d'expressions référentielles : pour pouvoir produire une description définie, la description doit être distinguante, c'est à dire qu'elle doit permettre d'identifier de façon unique et non ambiguë le référent dans le contexte (Dale et Reiter, 1995).

Le démonstratif est considéré comme déterminant des descriptions dont le référent est saillant, et comme ayant la capacité de reclasser le référent de la description, c'est à dire de changer de point de vue sur le référent en employant un groupe nominal sans relation lexicale avec l'antécédent (Corblin, 1987).

Considérons maintenant les reprises anaphoriques suivantes :

- (1) *Paul₁ a été agressé. La victime₁ se rendait à son bureau lorsqu'un malfaiteur a surgit.*
- (2) *Paul₂ a été agressé en allant au cinéma. Ce professeur de mathématiques₂ est un cinéphile.*

Ces deux phrases illustrent bien les données connues sur les emplois du défini et du démonstratif (Corblin, 1987, Kleiber, 1986, Kleiber, 1988). Dans le premier texte, le syntagme *Paul* est repris par un groupe nominal défini dont le contenu sémantique permet d'identifier le référent sans ambiguïté dans la mesure où il est le seul à correspondre à la description dénotée par le syntagme (le nom *victime* étant un moyen de référer à une personne ayant subi une agression).

Dans le second texte, le syntagme *Paul* est repris par un groupe nominal démonstratif dont le contenu sémantique illustre la propriété "reclassifiante" du démonstratif, qui permet la coréférence entre deux syntagmes sans qu'ils entretiennent une relation lexicale particulière (nous considérons que *Paul* et *la victime* entretiennent une relation lexicale par le biais du verbe dans le premier texte.).

Ces deux exemples, en plus d'illustrer les données issues de la littérature sur les déterminants, nous montrent par ailleurs que deux syntagmes coréférents peuvent ne pas avoir le même contenu sémantique. Dans un cas cependant, l'anaphore n'apporte pas d'information nouvelle sur le référent (le fait que Paul soit une victime est inféré à partir du contexte), tandis que dans l'autre, l'anaphore sert de support à de l'information nouvelle (rien ne peut laisser deviner que Paul est professeur de mathématiques). On pourrait penser que la possibilité d'ajouter de l'information sur le référent est une propriété du syntagme démonstratif, mais si on considère les exemples suivants, on constate que ce n'est pas le cas :

- (1) *Paul₁ a été agressé. Le professeur de mathématiques de ma fille₁ se rendait à son bureau lorsque le malfaiteur a surgit.*
- (2) *Paul₂ a été agressé en allant au cinéma. Cet homme₂ a toujours des ennuis.*

Cette constatation est importante pour la génération d'expressions référentielles. En effet, la préoccupation principale des algorithmes de génération était jusqu'à présent de produire des descriptions définies suffisamment informatives pour identifier le référent dans le contexte et éventuellement de décider de l'utilisation d'une description définie ou d'un pronom dans les mentions subséquentes du référent (Dale et Reiter, 1995). De plus, ces algorithmes présupposent que l'information contenue dans la reprise anaphorique est de l'information déjà connue des locuteurs. Il est donc possible d'étendre ces algorithmes de génération de deux façons :

- en permettant le choix entre description définie et description démonstrative,
- en tenant compte du contenu de la description en termes d'information donnée et de distinguer les anaphores répétant de l'information donnée (désormais ARID) des anaphores apportant de l'information nouvelle (désormais AAIN).

2.2. Implications pour la génération

Il est généralement admis qu'une tâche de génération de texte se divise en (au moins) deux phases : la première est la phase où on décide du contenu du texte à générer en fonction du but de la communication (module *Quoi dire ?*) et la seconde est la phase où l'on décide de la façon dont on va formuler le contenu (module *Comment le dire ?*) (Danlos, 1985).

2.2.1. Quoi dire ?

Après avoir identifié le but communicatif à atteindre par la reprise anaphorique, c'est à dire après avoir répondu à la question : "Quelle est la fonction de la reprise : ARID ou AAIN ?", une des premières tâches d'un algorithme de génération d'expressions référentielles sera donc, de répondre à la question "Quoi dire?" et de décider du contenu exact de la description à générer :

- **Si la fonction est anaphorique**, on doit savoir quelle information reprendre et d'où elle provient. Nos exemples montrent que dans le cas de la reprise par "la victime", l'information vient du contexte, tandis que dans le cas de la reprise par *cet homme*, l'information provient de nos connaissances encyclopédiques à la fois sur le prénom *Paul* qui en français désigne généralement un humain de sexe masculin, et aussi sur le fait qu'il aille au cinéma qui nous laisse penser qu'il s'agit d'une personne d'âge adulte.

- **Si l'anaphore ajoute de l'information**, le module *quoi dire ?* doit alors déterminer l'information supplémentaire à donner.

2.2.2. Comment le dire ?

Quelle que soit la fonction de la reprise anaphorique, il faut ensuite répondre à la question "comment le dire ?". Pour cela, il est nécessaire de connaître les moyens linguistiques couramment utilisés :

- pour faire une reprise simplement anaphorique n'ajoutant pas d'information
- pour ajouter de l'information

2.2.3. Résumé

Les éléments à connaître pour parvenir à générer des reprises démonstratives et définies sont donc les suivants :

Pour les Anaphores Répétant de l'Information Donnée (ARID)

- Les provenances possibles de l'information contenue dans l'anaphore, afin de pouvoir faire des inférences et de ne pas se contenter de générer uniquement des reprises directes.
 - Le type de relation (lexicale ou autre) entretenue par l'antécédent et l'anaphore
- Pour les Anaphores Ajoutant de l'Information Nouvelle (AAIN)
- Le type de lien entre l'antécédent et l'anaphore
 - Le moyen linguistique par lequel on apporte l'information nouvelle.

3. Mise au point d'une nouvelle classification des descriptions définies et démonstratives

Partant des observations données dans la section 2, nous avons mis au point une classification des reprises définies ou démonstratives indirectes, afin d'apporter des éléments de réponses aux questions posées précédemment à partir de données issues d'un corpus (Nous présentons ce corpus dans la section 3.2). Nous avons donc annoté un corpus en fonction de la nouvelle classification établie. Dans cette section, nous présentons brièvement la classification, puis l'étude de corpus et les résultats obtenus.

3.1. Classification

Les classes établies sont présentées dans le tableau suivant (les exemples sont issus du corpus de démonstratifs, mais des phénomènes semblables apparaissent dans le corpus des descriptions définies).

ARID : l'information vient de l'antécédent	Celle-ci, (...) aurait en effet tissé un réseau de liens ambigus dans la gendarmerie, la sûreté de l'Etat, les clubs de tir . Le procès, (...), avait permis de mettre ces liens en relief.
ARID: L'information vient de l'antécédent et du contexte	Le patronat avait très sensiblement modifié son comportement . (...) La clé de ce nouveau comportement tient en deux chiffres.
ARID : L'information vient de la relation lexicale entre antécédent et anaphore	L'Inde paie un tribut sans cesse plus lourd à la sécheresse (...). Ce phénomène a été accentué par des choix économiques erronés.
ARID : L'information vient d'une relation lexicale et du contexte	La municipalité s'est dotée récemment d' un somptueux Palais des concerts . C'est dans ce bâtiment confortable et flambant neuf qu'a eu lieu l'inauguration.
ARID : L'information est inférée à partir des connaissances du monde	"Les journalistes ne feront pas de reportage sur la visite de M. Honecker au cimetière de Neunkirchen, dans la Sarre, où sont enterrés ses parents . (...) après que le chef d'Etat eut requis la "tranquillité" pour cette partie "privée" de son voyage en République fédérale .
AAIN - L'information vient d'une relation lexicale spécifiante	Ce document souligne la gravité croissante des conséquences médicales de la consommation de tabac, (...). Les auteurs de ce rapport formulent une série de propositions (...).
AAIN - L'information vient des modifieurs	L'aviation israélienne a effectué (...) un raid sur le camp de réfugiés palestiniens d'Ain-el-Heloue, dans les faubourgs de Saida, chef-lieu du Liban-sud , (...). Les chasseurs-bombardiers israéliens ont effectué (...) plusieurs attaques sur ce camp qui compte soixante-mille habitants , (...).
AAIN - L'information vient d'une relation lexicale spécifiante et des modifieurs	(...), le personnel a l'impression de seulement compter les points. (...) Pour ces ouvrières du bassin houillier dont quelques-unes ont déjà trois heures de transport par jour , la nouvelle (...)
AAIN - l'information nouvelle est apportée par un syntagme sans relation lexicale avec l'antécédent	(...) je tombe sur un article intitulé " Pourquoi les maris prennent le large". Je me dis : cherche pas, ils se débînent pendant que tu t'échînes à faire des pompes et des flexions, ces salauds-là .

3.1.1. Intérêt d'une telle classification

Economie de moyens : L'avantage à utiliser l'anaphore comme support à de l'information nouvelle est le suivant : cela évite de produire un texte long peu naturel. Apporter de l'information nouvelle sur un référent est un moyen de faire des textes plus économiques.

Par ailleurs, cette utilisation des groupes nominaux donne une solution alternative à l'apposition qui peut parfois être lourde et qu'on ne sait pas générer. Bien entendu, pour pouvoir permettre au générateur d'apporter l'information dans le groupe nominal, nous avons besoin d'un module de planification du texte qui autorise ce type de tournure. Ceci n'entrant pas directement dans notre domaine, nous supposons que le planificateur de document utilisé le permet.

Le modèle de l'interlocuteur : Nous souhaitons ajouter à cette présentation un commentaire concernant deux catégories difficiles à annoter : la catégorie des ARID passant par les connaissances lexicales et la catégorie des AAIN apportant de l'information grâce à un syntagme nominal sans relation avec l'antécédent peut poser un problème lorsque l'antécédent est un nom propre : en effet, l'annotation peut varier d'une personne à une autre, en fonction de son âge et de ses propres connaissances. En génération, ce problème va se poser en termes de représentation des connaissances de l'interlocuteur. Cependant, nous pensons que l'utilisation de l'anaphore comme support à de l'information nouvelle permet aussi une conception plus souple des connaissances de l'interlocuteur. On ne préjuge plus complètement de ses connaissances, puisque deux cas de figure sont possibles : Soit il sait qui est la personne ou l'entité dont il est question dans le texte, auquel cas il résoud l'anaphore grâce à ses connaissances du monde, soit il ne sait pas, et accomode l'anaphore grâce à ses connaissances linguistiques sur l'utilisation du déterminant défini (et apprend une information sur le référent).

3.2. Analyse de corpus

3.2.1. Présentation du corpus

Le corpus dont nous disposons est un extrait du corpus PAROLE¹, comprenant 65 000 mots, 545 syntagmes nominaux démonstratifs et 9000 syntagmes nominaux définis. Il est composé d'articles du journal Le Monde datant de septembre 1987 et tirés de toutes les rubriques (Politique, Economie, Sport, Culture et Loisirs). Le corpus est annoté au niveau morphosyntaxique suite au projet PAROLE selon le schéma d'annotation Multext / Multitag de l'action GRACE (Lecomte et al. 1997, Beaumont et al., 1998). Afin d'automatiser au maximum le travail sur corpus, nous avons réalisé une série de pré-traitements avant l'annotation, que nous détaillons dans Manuélian (2003).

3.2.2. Résultats et discussion

La classification présentée dans la section précédente nous a permis d'annoter un corpus contenant 1402 utilisations coréférentielles définies et 369 reprises coréférentielles démonstratives. Les descriptions sélectionnées pour cette annotation parmi les presque 10 000 du corpus ont les caractéristiques suivantes : elles sont coréférentielles, et leur antécédent n'est jamais propositionnel. L'antécédent peut parfois être un verbe dans le cas des reprises par le nom de rôle thématique (Manuélian, 2002).

Les résultats de l'annotation sont présentés dans les tableaux suivants : Le premier présente de façon générale la proportion d'ARID et d'AAIN en fonction de chaque déterminant, le deuxième présente le détail des résultats pour les anaphores répétant de l'information, et le troisième le détail pour les anaphores ajoutant de l'information.

	Démonstratif	Démonstratif	Defini	Defini
ARID	299	81%	1314	93%
AAIN	70	19%	88	7%
TOTAL	369	100%	1402	100%

¹ Ce corpus a été fourni par le laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française, UMR 7118, Nancy) dans le cadre de la collaboration LORIA-ATILF dans le Contrat de Plan Etat-Région

ARID	Démonstratif	Démonstratif	Défini	Défini
Antécédent seul	92	31%	642	50%
Antécédent et Contexte	24	8%	131	10%
Relation Lexicale	71	24%	226	17%
Relation Lexicale et Contexte	25	8%	57	4%
Connaissances encyclopédiques	87	30%	258	20%
Total	299	81%	1314	94%

AAIN	Démonstratif	Démonstratif	Défini	Défini
Relation Lexicale	2	3%	7	8%
Modificateurs	19	27%	22	25%
Relation Lexicale et modificateurs	3	4%	2	2%
SN sans relation	46	65%	52	59%
Total	70	19%	88	7%

Le premier constat que nous pouvons faire concerne la différence d'emploi des deux déterminants dans l'ajout ou la répétition d'information. Si l'ajout d'information en utilisant une description définie est possible, il est clairement moins important qu'avec le démonstratif, pour lequel on constate 19% de cas où la reprise ajoute de l'information sur le référent.

Le second constat que nous pouvons faire concerne la catégorie des ARID : on observe que la moitié des descriptions définies, contre un tiers pour les démonstratives. Par ailleurs, on constate que 20% des descriptions définies anaphoriques impliquent des connaissances du monde, contre 30% des démonstratives. Ceci confirme clairement la capacité de reclassification attribuée au démonstratif, même quand reclassification ne signifie pas apport d'information sur le référent. Par ailleurs, nous pouvons noter que les cas où l'information connue est inférée d'informations explicitement données dans le cotexte sont relativement rares pour le défini comme pour le démonstratif, et qu'il ne semble pas y avoir de différence significative entre les deux déterminants dans la possibilité de faire une reprise impliquant une relation lexicale.

Concernant les AAIN, le moyen d'apport d'information le plus utilisé pour les deux déterminants est l'utilisation d'un syntagme nominal sans relation avec l'antécédent. Cette information reste cependant à nuancer : on ne peut en effet pas dire qu'une description définie ou démonstrative reprenant un nom propre entretient une relation lexicale avec son antécédent. Aussi, si on retire de nos statistiques toutes les anaphores dont les antécédents sont des noms propres, la proportion d'AAIN par un syntagme nominal sans relation lexicale avec l'antécédent tombe à 22% pour le défini, contre 37% pour le démonstratif. Une fois encore, on constate la capacité reclassifiante du démonstratif quand l'antécédent est un nom commun. Notons aussi que l'ajout d'information par un hyponyme est un phénomène tout à fait marginal. Enfin, nous pouvons terminer en affirmant que l'utilisation des modificateurs est courante avec les deux déterminants.

4. Conclusions

4.1. Nécessité des inférences en génération d'expressions référentielles

Notre étude montre l'importance des connaissances du monde et des inférences présentes lors de la résolution et de la production des expressions référentielles. Le problème principal va donc être de les modéliser, et de les structurer, de façon à ce que le générateur puisse les construire. D'un autre côté, comme nous l'avons noté en section 2, cette vision des reprises par des descriptions définies et démonstratives peut permettre de simplifier le problème du modèle de l'interlocuteur, crucial en génération.

4.2. Nécessité d'un modèle structuré du contexte

L'autre point important de notre travail concerne le modèle du contexte. Si on veut pouvoir autoriser le générateur à faire des inférences pour générer des expressions référentielles, il est nécessaire d'identifier clairement les sources de l'inférence. De plus, il nous semble que les sources utilisées pour faire les inférences vont varier en fonction de paramètres linguistiques, il nous faudra

alors contraindre l'utilisation des diverses sources d'inférences, et pour cela, une structuration nette des connaissances impliquées dans la production d'expressions référentielles est indispensable en entrée de l'algorithme.

4.3. Propositions pour une extension des algorithmes de génération existant :

Au vu des conclusions que nous avons tirées de notre étude de corpus, nous proposons alors de structurer le modèle du contexte de la façon suivante en entrée de l'algorithme :

- Les connaissances du monde
- Le modèle du locuteur, contenant des informations à communiquer.
- Les connaissances lexicales intégrant des relations lexicales standards – hyponymier, hypéronymie, et synonymie, encodées dans un outil comme Wordnet (Fellbaum, 1998), par exemple.
- Le modèle du discours, contenant les informations données à propos du référent dans l'antécédent de l'expression référentielle, mais aussi dans le reste du contexte.

Nous proposons ensuite d'intégrer aux algorithmes de génération des priorités de choix entre les diverses sources d'inférences afin de pouvoir construire des expressions coréférentielles distinguantes, utilisant toutes les ressources possible. L'ordre est déterminé par la fréquence d'apparition des phénomènes dans le corpus, et est le même pour le défini et le démonstratif. Ceci peut être exprimé de la façon suivante :

Si l'anaphore n'ajoute pas d'information, construire une description distinguante en recherchant une propriété provenant :

- (1) de l'antécédent
- (2) des connaissances du monde
- (3) des connaissances lexicales
- (4) de l'antécédent et du reste du modèle de discours
- (5) des connaissances lexicales et du modèle de discours.

Si l'anaphore ajoute de l'information, l'information contenue dans la reprise doit provenir du modèle du locuteur, et ne pas se trouver dans les autres bases de connaissances. Pour la réaliser, utiliser dans l'ordre de préférence suivant :

- (1) un syntagme nominal complet n'entretenant pas de lien lexical avec l'antécédent
- (2) les modifieurs
- (3) un hyponyme de l'antécédent.

4.4. Perspectives

A moyen terme et étant donné les besoins identifiés pour la génération, nous souhaitons corrélérer ces résultats avec des données syntaxiques (distance avec l'antécédent, fonction grammaticale de l'antécédent) et discursives (saillance du référent, position dans la chaîne anaphorique), afin de :

- vérifier les données concernant la saillance de l'antécédent d'un groupe nominal : il est admis qu'un démonstratif reprend un élément saillant dans le contexte, tandis qu'un défini reprend un élément identifiable uniquement dans le contexte. Nous souhaitons voir si cette théorie est vérifiée aussi bien pour la catégorie des ARID que pour les AAIN.
- affiner nos directives pour la génération et parvenir à un algorithme tenant compte des paramètres informationnels, sémantiques, syntaxiques et discursifs afin de générer des reprises anaphoriques réalistes et variées.

5. Bibliographie

Beaumont C., Lecomte J., et Hatout N., (1998) *Etiquetage morpho-syntaxique du corpus "Le Monde" pour les besoins du projet PAROLE*, Technical Report, INALF, Nancy.

- Corblin F. (1987), *Indéfini, Défini et Démonstratif*, Genève, Paris, Droz.
- Dale R. Reiter E., (1995), Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions *Cognitive Sciences*, 19(2), pp233-263.
- Danlos, L. (1985), *Génération automatique de textes en langues naturelles*, Paris, Masson.
- Fellbaum C., 1998, *Wordnet. An electronic lexical database*, MIT Press, Cambridge, Mass.
- Gardent C., Striegnitz K., (2000), Generating Indirect Anaphora, proceedings of *IWCS'00 (International Workshop on Computational Semantics)*.
- Heim I., (1982), *The Semantics of Definite and Indefinite Noun Phrases*, Ph. D. University of Massachusetts-Amherst.
- Kleiber G., (1986), Pour une explication du paradoxe de la reprise immédiate un N → le N / un N → Ce N, *Langue Française*, 72, pp 54-79.
- Kleiber G., (1988), Reprise immédiate et théorie des contrastes, *Studia Romanica Posnaniensa*, 13, pp 67-83.
- Krahmer E., van Erk S., Verleg A., (2001), A Meta-Algorithm for the Generation of Referring Expressions, proceedings of *8th European Workshop on Natural Language Generation* pp 29-39, Toulouse, France.
- Lecomte J., (1997) *Codage Multext - GRACE pour l'action GRACE / Multitag*, Technical Report, INALF, Nancy.
- Manuélian H., (2002) Annotation des descriptions définies : le cas des reprises par les rôles thématiques, *RECITAL 2002*, Nancy, France.
- Manuélian H. (2003) Une analyse des emplois du démonstratif en corpus, *Actes de TALN 2003*, Batz sur Mer, France.
- Russell, B. (1905), On denoting, *Mind*, 14 pp. 479-493.
- Van Deemter K., (2000), Generating Vague Descriptions, proceedings of *First International Conference on Natural Language Generation*, pp 179-186.
- Van Deemter K., (2001), Logical Form Equivalence : the Case of Referring Expressions Generation, proceedings of *8th European Workshop on Natural Language Generation* pp 21-29, Toulouse, France.