Le calcul de la référence

Bernard Victorri

Le calcul de la référence[1]

Bernard Victorri

1. Présentation

1.1 *Notion de chaîne de coréférence*

La *référence* est la relation qui associe certaines expressions linguistiques avec ce dont elles parlent, qu'il s'agisse d'un individu, d'un objet, d'un événement, etc. On appelle *expression référentielle* une expression linguistique qui a ce pouvoir d'évocation, et le *référent* l'élément extralinguistique ainsi évoqué, qui peut tout aussi bien appartenir au monde réel ou être le pur produit de l'imagination du locuteur. Dans un texte comme dans un discours, un même élément extralinguistique peut être évoqué à plusieurs reprises, et, le plus souvent, par des expressions référentielles différentes. Ainsi une personne peut être désignée de temps en temps par son nom (ex. : *Jacques Chirac*), dans d'autres passages par une description qui le caractérise et que l'on appelle une *description définie* (ex : *le Président de la République, le chef de l'Etat*, etc.), ou encore, plus brièvement, par un simple pronom (*il, celui-ci*, etc.). Les expressions qui réfèrent à un même élément sont dites *coréférentielles*, et elles forment dans un texte ce que l'on appelle une *chaîne de coréférence*. En traitement automatique des langues, le calcul de la référence consiste d'abord et avant tout à déterminer ces chaînes de coréférence. En effet, pour toute application qui nécessite un traitement sémantique, même très léger, il est important de repérer les différentes occurrences dans le texte d'un même référent. Pour certaines applications, comme l'indexation automatique, ce simple repérage peut suffire. Pour d'autres, qui visent une compréhension plus riche et complète, une représentation des référents (comme entités extralinguistiques), accompagnée de raisonnements appropriés, doit être mise en œuvre mais le calcul de coréférence constitue une étape indispensable du traitement de ces référents.

Pour illustrer concrètement ce que sont ces chaînes de coréférence, et les principaux problèmes qui se posent quand on cherche à les déterminer automatiquement, considérons le constat d'accident suivant, dans lequel on s'intéressera uniquement aux véhicules évoqués dans le texte[2] :

(1) J'étais (véhicule A) à l'arrêt au feu rouge derrière ***une 4L de la gendarmerie****. (2) Lorsque le feu est passé au vert,* ***la 4L*** *a calé. (3) J'attendais qu'****elle*** *redémarre lorsque j'ai été heurtée à l'arrière par* ***le véhicule B****. (4) Conduisant un véhicule aménagé pour la conduite manuelle (?), j'ai sans doute accentué la pression sur le frein principal que je tenais dans la main droite ce qui a causé une douleur du poignet mais qui m'a évité de heurter* ***l'autre véhicule****. (5) Bien entendu ma tête a fait un hochement involontaire*.

Il est question de trois véhicules dans cette petite histoire : le véhicule de l'auteur, celui de son adversaire, et un troisième véhicule qui n'est pas impliqué dans l'accident. Commençons par ce dernier. Il est introduit dans la première phrase par le syntagme nominal complexe *une 4L de la gendarmerie*. De manière typique, le déterminant de ce syntagme est l'article indéfini *un*, qui est en général un bon indice de l'introduction d'une nouvelle entité. Dans la deuxième phrase, cette entité est à nouveau évoquée, cette fois-ci par un groupe nominal plus simple *la 4L*. Cette reprise d'une partie du contenu lexical de la première évocation avec l'article défini *le* s'appelle une *anaphore nominale* (ç'aurait pu aussi être à l'aide d'un démonstratif : *cette 4L*, ou encore avec un hyperonyme : *cette voiture* ou *ce véhicule*). Dans la troisième phrase, de manière tout aussi

1 Ce texte est, pour l'essentiel, un résumé de la thèse de doctorat de Michel DUPONT [DUP 03].

2. Nous numérotons les phrases pour une bonne compréhension des explications qui suivent. Trois chaînes de coréférence sont matérialisées par un jeu de police gras et/ou souligné. Le ' ?' note un statut particulier, discuté infra, dans la chaîne qui le contient.

classique, l'entité est maintenant exprimée par le simple pronom *elle* : c'est une *anaphore pronominale*. Enfin dans la quatrième phrase, pour des raisons sur lesquelles nous reviendrons ci-dessous, c'est à nouveau à une anaphore nominale à laquelle nous avons affaire avec *l'autre voiture*. Ainsi la chaîne de coréférence pour cette entité comporte quatre expressions linguistiques. Si l'on se tourne maintenant vers le véhicule de l'auteur, sa chaîne de coréférence comporte deux mentions explicites : *véhicule A* dans la première phrase, et *un véhicule aménagé pour la conduite manuelle* dans la quatrième. Mais en fait, ce véhicule est évoqué implicitement à plusieurs autres reprises, notamment par *j'ai été heurtée à l'arrière* (phrase 3). C'est en effet le véhicule de l'auteur qui a été heurté à l'arrière et non sa conductrice : il s'agit d'une *métonymie intégrée* (cf. chap. 3, § 2.3) très fréquente dans ce type de texte. C'est aussi pour cette raison que la première mention, *véhicule A*, est mise entre parenthèses dans la première phrase : *j'étais à l'arrêt au feu rouge* parle déjà de la voiture de l'auteur, et l'indication entre parenthèses ne fait qu'apporter une précision sur la manière de nommer cette voiture. La chaîne de coréférence pour cette entité est donc plus imprécise. Si l'on s'en tient aux mentions explicites, elle ne comporte que deux évocations, mais ce nombre passe à quatre, voire cinq, si l'on inclut les métonymies. Quant au véhicule de l'adversaire, il n'est nommé qu'une fois, à la troisième phrase, par le groupe nominal *le véhicule B*. La chaîne de coréférence se réduit donc dans ce cas à un seul élément.

Il existe donc une très grande diversité dans la manière de référer à une entité, dont ce petit texte ne donne d'ailleurs qu'un aperçu très partiel. Généralement, plus une entité est au centre du discours, plus elle est évoquée avec concision. C'est ainsi que le véhicule des gendarmes est indiqué par des formes de plus en plus brèves dans les deuxième et troisième phrases (*la 4L*, puis *elle*) parce qu'elle occupe le devant de la scène, pour ainsi dire. En revanche, après l'introduction du véhicule B et du choc à la fin de la troisième phrase, elle passe à l'arrière-plan, ce qui explique qu'il faille revenir à un syntagme nominal plein (*l'autre véhicule*), avec une marque (*l'autre*) qui indique explicitement qu'il ne s'agit pas de l'entité la plus saillante de la classe désignée (les véhicules). La multiplicité des formes d'expression de la référence permet donc un ajustement assez subtil avec la position occupée dans le discours par l'entité dont on veut parler au moment où on veut en parler.

Les règles qui régissent cet ajustement ne sont pas simples. Par exemple, nous avons dit qu'en général une entité était introduite la première fois par l'article indéfini *un*. Or on aura pu noter que cela n'est pas le cas pour les premières références aux véhicules impliqués dans l'accident, introduits l'un par une métonymie accompagnée d'une précision (*véhicule A*), et le second par l'article défini : *le véhicule B*. Cela s'explique par le contexte de ces petites narrations : elles font partie d'un formulaire de constat d'accident dans lequel les dénominations *véhicule A* et *véhicule B* sont omniprésentes. Autrement dit, ces entités sont pré-introduites par les conditions d'énonciation avant même que ne débute le texte. De même, on aura remarqué que l'indéfini *un* est utilisé plus tard, à propos du véhicule de l'auteur du texte à la quatrième phrase, bien après que celui-ci ait été introduit. Il s'agit là d'un phénomène différent : ce n'est pas à une *désignation* du véhicule à laquelle on a affaire, mais à une *qualification*, comme dans la formulation : *ce véhicule est un véhicule aménagé pour la conduite manuelle*. En fait la valeur de l'article *un* est alors proche d'un emploi *générique*, c'est-à-dire un emploi où le groupe nominal ne désigne plus une entité mais une catégorie d'entités. Aussi bien *le* que *un* sont susceptibles d'un emploi générique, comme le montre la phrase : *La 4L est une voiture fabriquée par Renault*.

Du point de vue du traitement automatique, la difficulté du calcul de la référence est très variable. Par exemple, il est assez facile de déterminer à quoi réfère le pronom *elle* de la troisième phrase de notre texte. D'abord il n'y a qu'un nom féminin dans la phrase précédente qui puisse lui servir d'antécédent[3] (*la 4L*). De plus il s'agit d'une entité très présente dans le texte, puisqu'elle a déjà été évoquée dans les deux phrases précédentes. Enfin ce *elle* est sujet d'un verbe, *redémarrer*, qui accepte volontiers comme actant une entité de type véhicule. Ces trois indices concordent donc

3. L'antécédent d'un pronom est une expression, située en amont dans le texte, que « reprend » le pronom (avec laquelle il est en relation de coréférence).

à conclure que *elle* évoque bien la 4L de la gendarmerie introduite au début du texte, et ces indices, en tout cas les deux premiers, sont assez faciles à obtenir automatiquement. On pourrait même penser que l'on prend trop de précautions, et que le premier indice devrait suffire : c'est effectivement ainsi que fonctionnent généralement les analyseurs, mais il faut tout de même noter que la règle de l'accord en genre et un nombre du pronom avec son antécédent n'est pas toujours vérifiée, elle non plus. Ainsi dans *Le ministre a remercié la foule venue l'accueillir : elle était visiblement très émue*, ce n'est sans doute pas la foule qui est émue, mais le ministre qui doit être une femme. De même, dans *Les Martin ont acheté un chihuahua ; il paraît qu'ils sont très affectueux et faciles à vivre*, ce ne sont sûrement pas les Martin mais les chihuahuas dont on vante la convivialité. Mais ces cas sont suffisamment rares pour que les analyseurs puissent les ignorer sans trop dégrader leurs performances.

En revanche, il y a beaucoup de cas, très banals, où la résolution des anaphores s'avère très difficile pour un système informatique, parce qu'elle réclame une compréhension en profondeur du texte. Nous en avons un exemple dans le texte de constat que nous avons commenté : il s'agit de *l'autre véhicule*. En effet, pour trouver qu'il s'agit bien de la 4L de la gendarmerie et non pas du véhicule B (que l'auteur aurait pu aussi appeler *l'autre véhicule* par opposition à son propre véhicule), il faut effectuer un véritable raisonnement que l'on peut gloser ainsi : l'auteur dit qu'elle a pu éviter de heurter l'autre véhicule ; or elle a dit auparavant que le véhicule B l'avait heurtée ; donc l'autre véhicule ne peut pas être le véhicule B. On le voit : cette inférence, bien que de forme très simple, exige une représentation assez complète du sens de l'ensemble du texte, et des connaissances adéquates sur les mouvements des voitures, les chocs, etc.

1.2. *Cadre de cette étude*

Ainsi la détermination des chaînes de coréférence doit s'appuyer sur des données de différents niveaux, depuis des indices morpho-syntaxiques (les accords) jusqu'à la compréhension du sens. Nous allons passer en revue les différentes approches qui ont été utilisées pour ces calculs, puis nous présenterons plus en détail le modèle que Michel Dupont a conçu et implémenté ([DUP 96], [DUP 98], [DUP 03]). Mais auparavant, quelques remarques sont nécessaires pour délimiter plus précisément le champ d'étude que nous allons couvrir.

D'abord, nous allons nous en tenir à l'identification des *entités*, terme qui englobe les personnes, les objets, etc., par opposition aux autres éléments faisant partie de la représentation du sens : les procès, les qualités, les repères spatiaux et temporels, etc. C'est en effet sur les entités qu'ont porté l'essentiel des efforts, à juste titre puisque ce sont les chaînes de coréférence d'entités qui sont les plus fréquentes et les plus utiles dans la plupart des applications. Mais il faut tout de même noter que les chaînes de coréférence de procès peuvent aussi jouer un rôle primordial dans la compréhension de certains textes. Ainsi, dans l'exemple suivant :

Pierre est parti vivre au Québec. Cela n'a pas plus à ses parents mais il l'a fait quand même : il en avait trop envie.

le texte est structuré par une chaîne de coréférence du procès présenté dans la première phrase, et qui est repris à trois reprises par des marques très variées : un démonstratif (*cela*), un groupe verbal (*il l'a fait*), et un pronom « personnel » (*en*). De même, comme on le verra au chapitre 5, toute la cohérence temporelle des textes repose sur les relations anaphoriques par lesquelles sont déterminés les *intervalles de référence*, sortes de fenêtres temporelles au travers desquelles le locuteur donne à voir ce dont il parle.

Une autre limite de cette présentation, c'est que nous n'allons pas nous préoccuper de la manière dont on repère les marques linguistiques qui réfèrent effectivement à une entité. Ces marques sont essentiellement des syntagmes nominaux pleins (c'est-à-dire à tête nominale ; ex. : *la voiture blanche*), des syntagmes nominaux incomplets (ex. : *la blanche*) et des pronoms (personnels,

démonstratifs, possessifs, etc.)[4]. Mais cela ne veut pas dire que tous les syntagmes nominaux et tous les pronoms désignent des entités, loin de là. Du côté des pronoms, il faut éliminer les pronoms dits « impersonnels » : *il pleut, il reste deux cent francs, il arrive que l'on se trompe*, etc., de même que des emplois génériques des pronoms, comme le *on* de notre dernier exemple. Les syntagmes nominaux, quant à eux, peuvent aussi exprimer des procès (*le choc, le départ de Pierre, le transport des marchandises par camion*, etc.), des qualités (*la splendeur du paysage, le bleu du ciel*, etc.), ou, plus délicat, des éléments dont on peut se demander si on doit les traiter comme des entités. Ainsi les lieux (*la place du village, la Côte d'Azur*, etc.) doivent-ils être traités comme des entités ou relever uniquement d'une représentation spatiale (cf. chapitre 6) ? Nous avons déjà signalé aussi (chap. 2 § 2.2) le cas de syntagmes comme *la partie droite de la chaussée* pour lesquels on peut hésiter à créer une entité 'partie' en plus de l'entité 'chaussée'. On peut rapprocher ce dernier cas des *anaphores associatives*, qui posent aussi des problèmes d'identification d'entités qui font partie d'autres entités évoquées précédemment. Ainsi, dans le texte de constat d'accident suivant :

Je m'engageais (véhicule A) dans une file de station-service. La pompe étant en panne, je reculais pour repartir lorsque j'ai heurté le véhicule B qui s'était engagé également dans la même file pour prendre de l'essence.

le fait que le syntagme nominal *la pompe* soit introduit par un article défini s'explique par le fait que la pompe en question fait partie de la station-service évoquée dans la phrase précédente. C'est ce que l'on appelle une *anaphore associative* (cf. [SCH 94]), et il faut bien entendu en rendre compte dans la représentation des deux entités évoquées. De même, dans la phrase *J'ai freiné mais mon pied a glissé sur la pédale*, il est crucial, pour interpréter correctement *la pédale* comme 'la pédale de frein', de prendre en compte l'anaphore associative, qui est cette fois plus complexe puisqu'elle lie le syntagme nominal *la pédale* au groupe verbal *j'ai freiné*.

Enfin une dernière remarque sur l'intérêt applicatif du calcul de la référence. S'il est clair qu'un module d'identification d'entités et de procès le plus complet possible est indispensable pour des tâches de compréhension automatique de textes (cf. chapitre 7) ou d'extraction d'information (cf. chapitre 8, voir aussi [DUP 02]), des techniques plus légères de construction de chaînes de coréférence peuvent être d'une grande utilité pour des tâches qui ne réclament pas une compréhension en profondeur des textes, notamment dans le domaine de l'information documentaire (cf. chapitre 9). En effet, mêmes si elles contiennent des erreurs, ces chaînes de coréférence pourraient permettre, par exemple, de mieux calculer les fréquences de certains descripteurs de documents. Une entité qui est le thème central d'un texte va être évoquée par une longue chaîne de coréférence comprenant essentiellement des pronoms : comptabiliser ces occurrences pour calculer le poids du descripteur en question permettrait de gagner en précision dans l'indexation pondérée de documents (cf. le modèle vectoriel d'indexation chap. 9, § 2.2). Autre exemple[5] : la recherche d'information par requête structurée (chap. 9 §3), qui devra, pour améliorer ses performances, être capable de repérer les phrases contenant l'information recherchée même quand l'une des entités pertinentes est désignée par un pronom.

2. Les différentes approches

2.1. *Heuristiques, contraintes syntaxiques et sémantiques*

Les premiers travaux en traitement automatique des langues s'étaient déjà préoccupés de la résolution des anaphores. On peut ainsi citer SHRDLU, le système conçu par Winograd [WIN 72]. L'univers de SHRDLU consiste en un ensemble de blocs de couleurs, de formes et de tailles variées manipulables par un robot. Le système dialogue avec un utilisateur qui demande, en langage complètement libre, que certains déplacements des objets soient effectués par le robot. Ce système

4. Auxquels il faut aussi ajouter les adjectifs possessifs, qui posent aussi un problème d'anaphore pronominale : *sa voiture = la voiture de lui/elle*. Les pronoms possessifs sont donc doublement anaphoriques : *la sienne = celle de lui/elle*.

5. On peut aussi citer certaines approches récentes en résumé automatique fondées sur la saillance (cf. chap. 9, § 4.3).

est alors capable de « comprendre » la consigne et, à partir de la connaissance de la configuration initiale, de planifier et d'exécuter les tâches à effectuer pour arriver à la configuration demandée. De plus, il peut dialoguer avec l'utilisateur pour répondre à des questions sur les commandes qu'il effectue. Un tel système doit résoudre des anaphores, pour pouvoir interpréter correctement des commandes telles que : *Prends le cône qui se trouve sur l'un des gros cubes rouges, pose-le à côté de l'autre cône et déplace le cube dans le coin droit.*

Les approches de la première période consistent à grouper dans un algorithme des règles de filtrage et de préférence, très simples, portant sur des critères de différents niveaux, morphologique, syntaxique ou sémantique. Elles ont par ailleurs aussi recours à des heuristiques comme celle-ci : quand il reste plusieurs candidats, prendre le dernier rencontré. Lorsqu'un tel algorithme opère sur un micro-monde fermé, comme celui de SHRDLU, il résout pratiquement toutes les anaphores rencontrées. Sur d'autres types de texte, les performances sont forcément plus limitées. Nous reproduisons ci-dessous (tableau 1) la description que donne Sabah ([SAB 89], p. 222-223) d'un tel algorithme de résolution des anaphores pronominales. Celui-ci est utilisé pour l'identification des personnages dans un système de compréhension d'histoires simples.

Au regard de sa relative simplicité, un tel algorithme est très efficace. Il prend en compte de multiples facteurs : catégorie sémantique de l'entité compatible avec les contraintes imposées par le prédicat, accords entre antécédent et pronom, contraintes syntaxiques concernant le domaine de recrutement de l'antécédent, parallélisme des fonctions et aussi la récence qui intervient en dernier recours.

- A l'apparition d'un pronom, le système construit une liste des candidats possibles en éliminant de la liste des personnages ceux qui ne conviennent pas en genre, nombre et personne.
- Si la liste restante n'a qu'un élément, on considère qu'il s'agit de la bonne interprétation.
- Si la liste contient plusieurs éléments, le système applique deux types de règles afin d'exclure certains candidats : les règles strictes et les règles de préférence. Donnons deux exemples des premières :
  - Un pronom à la troisième personne ne peut remplacer un personnage représenté par des références absolues (je, tu) (comparer, par exemple, les interprétations de « L'**auteur** voit un lion, **il** se cache » et de « je vois un **lion**, **il** se cache »).
  - Un pronom non réfléchi ne peut remplacer un personnage apparu au même niveau syntaxique (phrase ou sous-phrase) que le pronom (dans « Un renard **le** salue » ou dans « Je dis qu'un renard **le** salue », le personnage salué ne peut être le renard).
- Si des ambiguïtés subsistent, on applique des règles de préférence comme le parallélisme syntaxique : si le pronom sujet peut se référer à plusieurs objets de la phrase précédente, préférer le sujet (normalement on ne dira pas « Un lapin voit un lion. **Il** le mange » si l'on souhaite que « il » se réfère au lion).
- Enfin, en dernier recours, on peut choisir le dernier mentionné, parmi les candidats restants.
- Dans tous les cas, dès que l'on a trouvé une solution potentielle, on vérifie que les contraintes sémantiques issues du verbe sont respectées sinon le candidat est éliminé. Cette élimination est effectuée systématiquement, après chaque étape intermédiaire, et en particulier avant l'application des règles de préférence.
- Si à la fin du processus, il ne reste plus aucun candidat, il y a incompréhension et le système demande au narrateur des éclaircissements sur l'antécédent du pronom.

Tableau 1 *Un algorithme simple de résolution d'anaphores pronominales*

2.2. *Processus inférentiels et représentation de la connaissance*

A la suite de ces premiers travaux, un effort considérable a été fourni dans différents courants de recherche de l'Intelligence Artificielle pour tenter d'améliorer les performances de ces systèmes en

effectuant des raisonnements prenant en compte des connaissances de sens commun. Nous avons vu en effet que des inférences complexes sont parfois nécessaires pour trouver l'entité réalisée par une forme linguistique et que celles-ci doivent prendre en compte des données extérieures au texte (cf. l'exemple donné au début de ce chapitre où l'on a vu que seul un raisonnement s'appuyant sur des connaissances sur les mouvements des voitures et des chocs permettait de résoudre l'anaphore nominale *l'autre voiture*).

Les recherches menées pour intégrer ces processus inférentiels dans les systèmes de TAL ont participé pour une grande part à l'évolution de la représentation des connaissances en I.A., laquelle a abouti à l'élaboration de langages spécialisés, tels que les langages de *frames* et les systèmes hybrides associant les paradigmes de la programmation logique et de la programmation par objets. Le problème étant de structurer la masse de connaissances, pour une part déclarative et pour une part procédurale, nécessaire pour effectuer ces raisonnements. On peut notamment citer les travaux de Charniak ([CHA 72] [CHA 78a], [CHA 78b]) qui a utilisé, pour traiter de petits textes narratifs, des techniques associant des règles d'inférence spécifiques à des classes d'entités particulières. Ces règles se déclenchent quand une entité de la classe associée est évoquée dans le texte, simulant ainsi, en quelque sorte, des processus de lecture « active » : par exemple, quand on parle de tel type de personnage ou de situation, le système active les scénarios correspondants et interprète la suite en fonction de ses attentes.

2.3. *La notion de focus*

Constatant que les approches fondées sur les processus inférentiels n'exploitent pas les contraintes que les formes linguistiques imposent au processus cognitif[6], Sidner [SID 83], en s'appuyant sur les travaux de Grosz [GRO 77], propose une approche dite par *focus*[7]. En guise de définition de la notion de focus, on peut faire cette observation : il existe dans la plupart des énoncés une entité qui, plus facilement que les autres, peut être évoquée par la suite par un pronom à la troisième personne. C'est ce qu'on appelle le focus. Le principe de l'algorithme consiste d'abord à identifier un *focus prévisible* (*expected focus*) à partir de certaines marques syntaxiques. S'il y a un pronom dans la phrase suivante, le système examine si l'entité reconnue dans la phrase précédente comme étant le focus prévisible est compatible avec les marques morpho-syntaxiques de la nouvelle réalisation, puis un processus inférentiel intervient pour détecter une éventuelle contradiction avec les connaissances de sens commun. Si une contradiction est trouvée, le focus candidat est rejeté et un autre candidat est choisi. Dans tous les cas, un focus prévisible est à nouveau choisi pour la phrase suivante. Des critères comme la position de sujet syntaxique ou des marques de mises en relief permettent de repérer les changements de focus. Les processus inférentiels n'interviennent que pour détecter une éventuelle contradiction entre la proposition faite et les connaissances de sens commun. Cela apparaît clairement avec ces deux exemples extraits de Kleiber ([KLE 94a], p. 108) :

Paul enleva son manteau. Il avait trop chaud.

Paul enleva son manteau. Il était élimé.

Dans les deux cas, les algorithmes basés sur le focus proposeront *Paul* comme référent de *Il*, mais, dans le deuxième exemple, un processus inférentiel viendra invalider cette proposition. Il peut d'ailleurs arriver que la proposition soit invalidée beaucoup plus tôt par détection d'incompatibilités d'ordre morpho-syntaxique, comme dans l'exemple : *Paul enleva son écharpe. Elle était élimée*.

---

6. On trouve dans Grosz et Sidner [GRO 98] cette observation : « Sidner's work was in marked contrast to contemporaneous computational research by Winograd, Charniak, Rieger, and Hobbs in which open-ended inference was applied to logical-form like representations in which pronouns were represented by free variables. These approaches provided no constraints from linguistic information on the cognitive processing required »

7. Parmi les algorithmes ayant utilisé cette approche, il faut citer RAFT/RAPR (*Revised Algorithms for Focus Tracking / Revised Algorithms for Pronoun Resolution*) de Suri et McCoy [SUR 94].

Examinons sur un exemple simple comment l'algorithme peut fonctionner. Soit le petit texte suivant, inspiré de [SID 83] :

La semaine dernière, il y avait des bonnes fraises dans le réfrigérateur. Elles venaient de notre coopérative et étaient particulièrement fraîches. J'ai voulu les prendre pour dîner mais quelqu'un les avait toutes mangées. Plus tard, j'ai découvert que c'était Marc qui les avait mangées. Il était affamé et il est impossible de conserver de la nourriture près de lui quand son estomac crie famine.

- PHRASE 1 : L'analyse de la première phrase permet au système de choisir 'fraises' comme focus prévisible à partir de la marque de présentatif *il y a*. Les autres groupes nominaux sont des compléments secondaires.
- PHRASE 2 : La règle du focus permet de résoudre l'anaphore *Elles*, qui est donc interprétée comme une réalisation de l'entité 'fraises'. Un processus inférentiel confirme qu'il n'y a pas de contradiction avec les connaissances de sens commun dont le système dispose puisque les fraises peuvent venir d'une coopérative et être fraîches. Ensuite, le focus prévisible est mis à jour. Ici 'fraises' est maintenu comme focus puisque cette entité a été à nouveau réalisée (contrairement aux deux autres entités de la première phrase, 'la semaine dernière' et 'réfrigérateur') et, de plus, en tant que sujet syntaxique de la phrase.
- PHRASE 3 : Le même processus est répété et permet d'identifier 'fraises' comme étant l'entité réalisée par les deux occurrences de *les*.
- PHRASE 4 : Le même processus est répété. Mais, au moment de la mise à jour du focus prévisible, la marque d'emphase donnée par la construction clivée *c'était... qui* indique un changement de focus. A partir de ce moment l'entité 'Marc' devient le focus attendu.
- PHRASE 5 : Le même processus permet de résoudre correctement les anaphores pour *Il*, *lui* et *son*.

Si le modèle semble fonctionner correctement sur des exemples simples tels que celui que nous venons de voir, il faut lui ajouter des extensions pour traiter des cas plus complexes. Sidner [SID 83] est ainsi amenée elle-même à introduire plusieurs focus. Elle distingue le *focus acteur* et le *focus du discours*. Le focus du discours est l'entité « dont il est question » ; « ce dont on parle ». Le focus acteur est l'entité qui a le rôle d'acteur dans la relation actancielle (sujet du verbe à la voix active). Le focus acteur a la priorité sur le focus du discours dans le traitement des pronoms personnels sujets alors que c'est l'inverse pour les autres pronoms.

La définition du focus de discours correspond à la définition du *thème* dans l'opposition thème/rhème, telle qu'elle est décrite par exemple par Ducrot et Todorov [DUC 72] : « (...) la distinction du thème et du propos est d'ordre psychologique. Le thème (anglais : *topic*) d'un acte d'énonciation, c'est ce dont parle le locuteur, c'est l'objet du discours, ou, comme disaient les linguistes du début du siècle, le sujet psychologique ; le propos, ou encore rhème (anglais : *comment*), c'est l'information qu'il entend apporter relativement à ce thème - ce qu'on appelait autrefois le prédicat psychologique (...) ». Il a souvent été souligné que ces définitions restaient trop vagues pour être utilisables. Kleiber, par exemple, parle du « caractère flou, intuitif, non conceptualisé de la notion de thème, topique ou encore focus » ([KLE 94a], p. 112)[8]. On peut donc considérer que les travaux de Sidner constituent une des premières tentatives sérieuses de rendre opérationnelle cette notion de thème en exhibant un algorithme capable de calculer concrètement un focus. Il faut aussi signaler les travaux d'Hajicova *et al.* [HAJ 95] qui proposent un algorithme d'identification du thème et du rhème en serrant au plus près le point de vue de l'école de Prague sur ces concepts, et ceux de Tabuteau [TAB 96] qui a réalisé une implémentation dans un système de dialogue oral homme-machine inspirée de ces principes.

8. On peut aussi citer Azzam *et al.* [AZZ 98] : « The term focus, along with its many relations such as theme, topic, center, etc., reflects an intuitive notion that utterances in discourse are 'about' something. This notion has been in accounts of numerous linguistic phenomena, but it has rarely been given a firm enough definition to allow its use to be evaluated ». Voir aussi Sabah ([SAB 89], p. 232).

2.4. *La théorie du centrage*

A la suite de leurs travaux sur le focus, Grosz et Sidner [GRO 86] se sont intéressées à la structure du discours en distinguant une cohérence locale à chaque segment du discours et une cohérence globale à l'ensemble du discours. Ce travail a été à l'origine de la théorie du centrage (*centering theory*) qui est définie comme une « théorie qui relie le focus d'attention, le choix d'une expression référentielle et la cohérence des énoncés à l'intérieur d'un segment du discours » ([GRO 95], p. 204, repris dans [WAL 00], p. 31-32. Voir aussi [GOR 93], [WAL 94], et l'ouvrage collectif [WAL 98]).

Dans cette théorie, l'accent est mis sur les *transitions* entre énoncés d'un même segment de discours. On va s'intéresser aux entités évoquées par un énoncé de deux points de vue : par rapport à l'énoncé précédent (*backward-looking*), et par rapport à l'énoncé suivant (*forward-looking*). Du point de vue de l'énoncé précédent, une des entités évoquées par l'énoncé joue un rôle privilégié : c'est ce qui correspond au focus ou au thème de l'énoncé, que l'on appelle ici le *centre rétroactif*, noté Cb (*backward-looking center*)[9]. Du point de vue de l'énoncé suivant, ce sont toutes les entités évoquées par l'énoncé qui ont plus ou moins vocation à devenir le thème futur : elles sont donc ordonnées dans une *liste des centres anticipateurs*, notée Cf (*forward-looking centers*). L'entité qui est en tête de cette liste s'appelle le *centre préféré*, noté Cp (*preferred center*) : on s'attend donc à ce que ce soit cette entité qui devienne le thème de l'énoncé suivant. Cette prédiction peut ne pas se réaliser, car l'énoncé suivant peut ne pas évoquer du tout cette entité. Mais dans tous les cas, le thème de l'énoncé suivant est l'entité de plus haut rang dans Cf qui est effectivement évoquée dans cet énoncé. Pour illustrer ces notions, reprenons l'exemple du petit texte de la section précédente, en numérotant les différents énoncés pour pouvoir utiliser des notations plus précises :

(E_1) *La semaine dernière, il y avait des bonnes fraises dans le réfrigérateur.*

(E_2) *Elles venaient de notre coopérative et étaient particulièrement fraîches.*

(E_3) *J'ai voulu les prendre pour dîner mais quelqu'un les avait toutes mangées.*

(E_4) *Plus tard, j'ai découvert que c'était Marc qui les avait mangées.*

(E_5) *Il était affamé et il est impossible de conserver de la nourriture près de lui quand son estomac crie famine.*

La liste des entités évoquées par le premier énoncé, notée $Cf(E_1)$, comporte 'fraises' et 'réfrigérateur' (peut-être aussi 'semaine dernière'), l'entité en tête de liste, $Cp(E_1)$, étant 'fraises', qui est donc le meilleur candidat comme thème du deuxième énoncé. Dans celui-ci, le thème est bien 'fraises', évoqué par le pronom *Elles*. On a donc $Cb(E_2) = Cp(E_1)$. Par ailleurs, la liste $Cf(E_2)$ comporte deux éléments, 'fraises' et 'coopérative', dans cet ordre : on a donc $Cp(E_2) = Cb(E_2)$, ce qui signifie que l'on s'attend à conserver le même thème dans l'énoncé suivant. La situation est analogue dans ce troisième énoncé. En revanche, dans l'énoncé E_4 on peut voir l'amorce d'un changement de thème. En effet, si $Cb(E_4)$, le thème de cet énoncé est toujours 'fraises' (évoqué ici par *les*), la liste $Cf(E_4)$ comporte une nouvelle entité en tête de liste : $Cp(E_4)$ = 'Marc'. On a donc $Cb(E_4) \neq Cp(E_4)$, ce qui veut dire que l'on s'attend à un changement de thème dans l'énoncé suivant. C'est bien ce qui se produit : $Cb(E_5)$ est bien 'Marc' (évoqué par *Il*, *lui* et *son*) : $Cb(E_5) \neq Cb(E_4)$. Dans ce petit texte, on a donc des exemples de trois types de transition possibles (cf. tableau 2) : continuité complète (appelée 'CONTINUE' par les auteurs) pour E_2 et E_3, annonce de changement ('RETAIN') pour E_4, et réalisation de changement ('SMOOTH-SHIFT') pour E_5. La quatrième transition prévue par le tableau ('ROUGH-SHIFT') n'apparaît qu'exceptionnellement : en effet, les auteurs énoncent une règle de cohérence du discours qui revient à ordonner préférentiellement les transitions dans l'ordre :

CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT

9. On notera que chaque énoncé a donc un Cb et un seul, à l'exception du premier énoncé d'un segment de discours, pour lequel ce « regard arrière » n'a pas lieu d'être.

	$Cb(E_n) = Cb(E_{n-1})$	$Cb(E_n) \neq Cb(E_{n-1})$
$Cp(E_n) = Cb(E_n)$	CONTINUE (continuité et prévision de maintien de la continuité)	SMOOTH-SHIFT (changement et prévision de continuité)
$Cp(E_n) \neq Cb(E_n)$	RETAIN (continuité et prévision de changement)	ROUGH-SHIFT (changement et prévision de nouveau changement)

Tableau 2 *Les quatre transitions de la théorie du centrage*

L'un des principaux intérêts de la théorie du centrage est de permettre d'exprimer des règles d'emploi de marqueurs en fonction des caractéristiques thématiques de l'entité. C'est ainsi que la première règle énoncée par les auteurs peut se formuler de la manière suivante : si certaines entités sont évoquées par des pronoms dans un énoncé, alors le thème (Cb) en fait forcément partie.

Mais c'est surtout sur les règles d'ordonnancement des entités dans la liste Cf des candidats à la position de thème dans l'énoncé suivant que les efforts ont porté, pour un nombre assez important de langues très diverses. En effet, il faut connaître ces règles pour prédire le thème de l'énoncé suivant[10], ce qui est nécessaire pour rendre opérationnel le modèle. Ces recherches ont été d'un intérêt capital puisqu'elles ont permis, dans un cadre unifié, d'obtenir pour de nombreuses langues des règles de résolution d'anaphores, non seulement pour les pronoms mais aussi pour des expressions anaphoriques nominales, ainsi que pour certaines formes elliptiques.

2.5. *Modèle du contexte et saillance*

Parallèlement à la théorie du centrage et de manière complètement indépendante, Hiyan Alshawi [ALS 87] a développé une approche plus cognitive qui, selon nous, représente une avancée considérable. Il ne s'agit plus de repérer quelques entités pour tâcher de prédire laquelle sera le focus dans la suite du texte, mais de prendre en considération toutes les entités qui ont été introduites depuis le début du texte et d'évaluer pour chacune sa *saillance*, qui indique dans quelle mesure cette entité est présente et disponible dans la mémoire du lecteur, mise de l'avant dans la représentation cognitive qu'il construit au cours de sa lecture. Pour cela, Alshawi propose de gérer un *modèle du contexte* dans lequel est associée à chaque entité du discours une valeur numérique qui mesure cette saillance, appelée *activation contextuelle*, et calculée en fonction des différents facteurs contextuels qui contribuent à mettre de l'avant l'entité en question. Comme le dit Alshawi ([ALS 87], p. 16) : « We can now give a precise definition for the notion of *context activation*, a measure of salience in context. The context activation of a memory entity is the sum of the current significance weights of the context factors within the scope of which the entity lies. Thus, at any point in processing, the relative importance of each entity is determined by the context factors which contribute to its activation score ».

Les facteurs pris en considération pour déterminer la saillance sont les mêmes que ceux qui servaient dans les algorithmes précédents à classer les entités dans la liste des focus potentiels dans l'approche par focus ou dans la liste Cf des candidats à la position de thème dans la théorie du centrage. Les algorithmes fondés sur ces théories ont donc en commun de chercher à faire une prédiction sur ce qui va apparaître dans le prochain énoncé en se servant de la partie déjà traitée du

10. Rappelons que le thème de l'énoncé suivant est l'entité de Cf de plus haut rang parmi celles qui sont évoquées dans cet énoncé. Il faut aussi noter que la théorie prévoit que l'ensemble Cf peut n'être que *partiellement* ordonné, ce qui implique que la connaissance de cet ordre ne suffit pas à elle seule à prévoir le thème (il peut y avoir plusieurs entités maximales pour cet ordre partiel qui sont en compétition pour le rôle de thème).

texte et en prenant en compte pour cela plusieurs facteurs. Ce qui fait l'originalité de l'approche d'Alshawi, c'est d'une part l'utilisation de valeurs numériques, et d'autre part le fait que l'analyse n'est plus centrée sur les transitions d'une phrase à la suivante : elle prend en considération toutes les entités qui sont dans le contexte du discours et elle affecte à chacune d'elles une saillance qui se modifie tout au long du texte, que l'entité soit ou non évoquée par la phrase en cours. On ordonne ainsi, à chaque moment de la lecture, l'ensemble des entités du contexte du discours.

Un mécanisme joue un rôle très important dans le modèle : la *dégradation*. A chaque facteur contextuel est associé une fonction de dégradation qui fait décroître sa valeur initiale au fur et à mesure que l'on s'éloigne dans le texte. Ainsi une entité qui n'est plus évoquée dans le texte va voir sa saillance décroître progressivement jusqu'à être complètement « oubliée » (saillance nulle). Par contre, si cette entité apparaît dans une nouvelle phrase, on ajoute à sa saillance dégradée la valeur correspondant à la nouvelle marque linguistique qui réfère à cette entité : plus une entité est évoquée, plus sa saillance sera grande.

Pour illustrer concrètement le modèle, nous allons prendre un exemple très simplifié, en ne faisant jouer qu'un seul facteur contextuel, à savoir la fonction syntaxique de la marque linguistique qui évoque chaque entité. Nous attribuerons la valeur de 100 à une fonction sujet, de 60 à un complément d'objet (direct ou indirect), et de 20 à toutes les autres fonctions syntaxiques, et nous appliquerons à ce facteur contextuel une fonction de dégradation qui consiste à diviser par 2 sa valeur après chaque phrase. Soit le petit texte suivant :

Pierre est allé au restaurant hier. Il était très indécis. Tout lui faisait envie. Il a changé trois fois sa commande. Ça n'a pas plu au serveur. Il a laissé un très gros pourboire pour se faire pardonner.

Calculons la saillance S de l'entité 'Pierre' après chaque phrase :

- PHRASE 1 : *Pierre* en position sujet : $S_1 = 100$.
- PHRASE 2 : On résout d'abord l'anaphore pronominale : *Il* évoque l'entité la plus saillante, à savoir 'Pierre'. Comme il est en position sujet, on ajoute 100 à la saillance initiale de 'Pierre' à laquelle on applique la fonction de dégradation : on a donc $S_2 = S_1 / 2 + 100 = 150$.
- PHRASE 3 : Même traitement ; *lui* évoque 'Pierre', et il est complément essentiel. On ajoute donc 60 à S_2 après dégradation : $S_3 = S_2 / 2 + 60 = 135$.
- PHRASE 4 : L'entité 'Pierre' est évoquée deux fois, par *Il* (+100) et par *sa* (+20). Donc $S_4 = S_3 / 2 + 120 \approx 188$.
- PHRASE 5 : 'Pierre' n'est pas évoqué du tout, on applique donc simplement la fonction de dégradation : $S_5 = S_4 / 2 = 94$.

Ainsi, quand arrive la sixième et dernière phrase, 'Pierre', le thème principal du texte, garde une forte activation, bien que cette entité n'ait pas été évoquée dans la phrase précédente. Sa saillance est plus forte que celle de l'entité 'serveur', qui est de 60 puisqu'elle vient d'être évoquée pour la première fois comme complément essentiel. Le modèle peut donc facilement rendre compte du fait que le pronom *Il* dans la dernière phrase réfère à Pierre et non au serveur. Notons que pour la théorie du centrage, cette anaphore pose un problème insurmontable puisque, selon les règles que nous avons présentées, le thème (Cb) de la dernière phrase doit d'une part avoir été évoqué dans la phrase précédente, et d'autre part doit être l'entité évoquée par le pronom *Il* : on en déduit que celui-ci doit obligatoirement référer au serveur…

Le modèle d'Alshawi a été implémenté dans des systèmes de TAL, notamment par Lappin et Leass [LAP 94] qui ont réalisé le système RAP (*Resolution of Anaphora Procedure*). Les performances de RAP sont tout à fait remarquables : sur du texte tout venant (manuels techniques), contenant 560 reprises anaphoriques par des pronoms, 85% des cas sont résolus alors qu'aucune connaissance de niveau sémantique n'est utilisée.

3. Un nouveau modèle de l'identification des entités

Le modèle d'identification des entités développé par Michel Dupont (cf. [DUP 96], [DUP 98], [DUP 03]) que nous allons maintenant présenter s'appuie d'une part sur une approche cognitive de la construction du sens qui a conduit à l'élaboration d'un *modèle des attentes*, comparable au modèle du contexte d'Alshawi, et d'autre part sur une classification des différentes formes linguistiques permettant de référer à des entités qui s'inspire de la *théorie de l'accessibilité* développée par Mira Ariel [ARI 90]. Comme on le verra, le fait de bien séparer ces deux aspects a pour corollaire de distinguer aussi plus nettement deux calculs de nature différente, qui sont souvent mêlés dans les approches que nous venons de présenter : le *calcul de saillance*, qui ne porte que sur les entités, qu'il s'agit de classer dans le modèle des attentes, et l'*identification des entités*, qui consiste à apparier formes linguistiques et entités. En particulier, pour mener à bien cet appariement, ce modèle introduit deux paramètres (qui n'existent ni chez Alshawi ni chez Ariel), la *plage de saillances admissibles* et *l'importance de la concordance*, qui jouent un rôle essentiel à l'interface des deux calculs.

3.1. *Le modèle des attentes*

Le modèle des attentes s'inscrit dans une théorie cognitive de l'interprétation, qui confère une place essentielle au contexte au sens large, c'est-à-dire incluant les conditions d'énonciation. Comme le dit Kleiber [KLE 94b] à propos des approches cognitives en général : « Le contexte n'est plus considéré comme un bouche-trou interprétatif, qui n'entre en ligne de compte que si l'on en a besoin. Il accède au rang d'élément décisif dans le processus d'interprétation de toute forme verbale ; c'est-à-dire qu'il intervient pendant et pour chaque interprétation (…) ». L'idée générale que nous défendons, c'est que le lecteur/auditeur ne reçoit pas passivement un message qu'il doit décoder, mais qu'il le fait entrer dans une boucle interprétative : avant la réception d'un énoncé, il anticipe ce qui va être dit/écrit ; le message est donc confronté à ses attentes qu'il valide ou infirme. Et c'est en tenant compte des attentes du lecteur/auditeur que le scripteur/locuteur formule son discours. C'est d'ailleurs ce qui explique que la plupart des ambiguïtés relevées par l'analyse linguistique ne soient pas une gêne pour les interlocuteurs en situation. Par exemple, l'ambiguïté (hors contexte) d'un énoncé comme *Dix vols par jour* ne sera même pas perçue par le lecteur d'un dépliant publicitaire d'une compagnie aérienne, comme elle ne l'a sans doute pas été par l'auteur de ce dépliant.

Ainsi, aussi bien le rédacteur que le lecteur d'un texte (pour nous en tenir ici aux écrits), prennent en compte les différents éléments contextuels qui « entourent » chaque énoncé, et qui proviennent de trois sources principales :

- le *contexte physique* du texte, dans lequel nous incluons le support (livre, prospectus, panneau, etc.), son environnement (qui peut aller du présentoir éventuel auquel il est destiné jusqu'à la mise en page dans laquelle il s'insère, les images qui l'accompagnent, etc.), et ce que l'on appelle les conditions d'énonciation : qui écrit ce texte, quand, dans quelle intention, etc.
- le *co-texte* de l'énoncé, constitué par le texte qui le précède, et qui est donc déjà censé avoir déjà été écrit/lu au moment où on aborde cet énoncé.
- les connaissances supposées du lecteur qui peuvent jouer un rôle essentiel dans la compréhension de certains textes. C'est ainsi qu'un énoncé tel que *L'idéal de l'anneau se plonge facilement dans un corps* ne peut être produit que par un rédacteur supposant que ses lecteurs auront les connaissances mathématiques adéquates pour lui donner un sens.

Reprenons l'exemple des constats d'accident pour illustrer l'importance de la première source d'information contextuelle, le contexte physique, qui n'est que rarement pris en compte dans les différentes approches que nous avons présentées. Le petit texte qui « raconte » l'accident n'est qu'une partie d'un formulaire rempli par l'assuré qui comporte de nombreuses rubriques :

informations sur les deux adversaires et leurs véhicules (les véhicules A et B), indication de la date, de l'heure et du lieu de l'accident, croquis de ce lieu incluant la position des véhicules au moment du choc, indication des points d'impacts sur un dessin représentant chaque véhicule, etc. Il est évident que l'auteur du texte prend en compte cet environnement quand il écrit son texte. Cela explique notamment que beaucoup de textes commencent par une phrase à l'imparfait, sans autre indication temporelle, comme le montrent les exemples suivants :

Nous roulions sur une route à 90 km/h. Un véhicule (…).

Je circulais à bord de mon véhicule A sur la file de droite réservée aux véhicules allant tout droit. Le véhicule B (…).

Je circulais à environ 45 km/h dans une petite rue à sens unique où stationnaient des voitures de chaque côté. Surgissant brusquement (…).

J'allais vers St Brice. Je venais de doubler (…).

Je circulais sur la voie de droite. Dans le virage, (…).

Les deux véhicules étaient en stationnement perpendiculairement, l'un A dans la direction de la voie, l'autre B perpendiculairement à la chaussée. En l'absence des conducteurs, (…).

Je roulais sur la voie à droite. Devant moi (…).

Comme on le verra au chapitre 5, l'imparfait pose un problème de référence temporelle, dans la mesure où il faut trouver un ancrage pour l'intervalle temporel de référence (l'instant dont on parle), qui ne peut pas être le procès qui porte cette marque d'imparfait. En début de texte, les phrases à l'imparfait sont donc généralement accompagnées d'une marque temporelle susceptible de servir d'antécédent à cet intervalle de référence (*Le 1er mai 1998, Lundi dernier, Ce jour-là,* etc.). Si ces textes semblent ignorer cette règle, c'est tout simplement parce que le lecteur comme le scripteur savent que ce texte va parler d'un accident qui s'est produit quelque temps avant l'écriture du texte en question, et que d'ailleurs la date et l'heure précise peuvent être trouvées dans son environnement immédiat.

De même, c'est le contexte qui explique que, comme nous l'avons déjà remarqué (section 4.1), la métonymie intégrée conducteur → véhicule est assez systématique dans les premières phrases de ces textes, en parlant du locuteur (cf. ci-dessus : *Nous roulions, Je circulais, J'allais*). En effet, c'est normalement en tant que conducteur d'un véhicule accidenté que le locuteur rédige ce texte. C'est pour la même raison que les ambiguïtés potentielles d'un verbe comme *rouler*, souvent présent dans ces premières phrases (cf. ci-dessus : *Nous roulions, Je roulais*), ne sont absolument perçues par le lecteur, comme nous avons déjà eu l'occasion d'en discuter (chap. 2, § 2.1, cf. aussi chap. 3, § 1 et § 2.4).

En ce qui concerne l'identification des entités dans ces petits textes, il est clair qu'avant même d'avoir commencé à lire la première phrase, le lecteur s'attend à ce que plusieurs entités soient évoquées (directement ou non) au cours du texte : le locuteur et son véhicule, l'adversaire et son véhicule, et au moins une entité de type 'lieu de circulation' (rue, route, carrefour, parking…). L'expression de ces entités va être facilitée relativement à d'autres, moins attendues, comme le montre l'exemple suivant :

Je sortais du parking de la résidence des fleurs pour me diriger vers le centre d'Ecouen sur ma gauche. Il y avait un autre véhicule devant le véhicule A qui roulait à plus de 45 km/h et pour lequel j'ai dû reculer un peu pour le laisser passer. Cet autre véhicule étant passé j'ai cru que le véhicule A me laisserait le passage mais il n'en a rien été et il m'a percuté au niveau de mon phare avant gauche.

Pour introduire « cet autre véhicule », l'auteur utilise une formulation assez lourdes, le présentatif : *Il y avait un autre véhicule*, que l'on ne trouve pratiquement jamais pour l'évocation des véhicules A et B.

D'une manière générale, dans le modèle des attentes, on considère donc qu'un certain nombre d'entités, qu'on appelle *entités préconstruites*, sont déjà présentes avant le début du texte[11]. Ensuite, tout au long du texte, la liste des entités du modèle des attentes varie en fonction de la confirmation ou non de la présence de ces entités préconstruites, et bien sûr surtout de la construction de nouvelles entités évoquées par le texte lui-même. A tout instant, chacune de ces entités est dans un état d'activation que l'on peut définir comme une mesure de l'intensité de l'attente du lecteur concernant l'évocation de cette entité dans la suite du texte. Dans le modèle, ce que l'on appelle *saillance de l'entité en un point donné du texte*, c'est une mesure de cet état d'activation. Pour utiliser une métaphore scénographique[12], plus les entités occupent le devant de la scène, plus leur saillance est forte.

La saillance est a priori une valeur numérique continue, définie sur une échelle arbitraire, puisque seules les valeurs relatives des saillances ont une pertinence dans le modèle. Mais, comme on le verra, il peut aussi être intéressant de discrétiser cette notion, en qualifiant (à l'aide de seuils à déterminer) les saillances de fortes, moyennes et faibles, de manière à ce que ces trois catégories correspondent aux cas de figure suivants :

- Les entités de *forte saillance* doivent être celles que l'on vient de mentionner très récemment (dans les deux ou trois phrases précédentes).
- Les entités de *saillance moyenne* peuvent être de trois sortes :
  a) les entités préconstruites (par l'environnement contextuel du texte), au cours des premières phrases du texte, tant qu'elles n'ont pas été encore évoquées
  b) les entités qui, après avoir été évoquées, n'ont plus été mentionnées depuis plusieurs phrases (leur saillance a été forte et s'est dégradée)
  c) les entités qui n'ont pas été explicitement évoquées, mais qui sont directement liées à une entité de forte saillance (elles sont donc susceptibles d'être l'objet d'une anaphore associative).
- Enfin, une *saillance faible* doit être attribuée à toutes les autres entités présentes dans le modèle des attentes. En pratique on n'y trouvera que les entités dont la saillance a été moyenne et s'est dégradée parce qu'elles n'ont pas été mentionnées au cours des phrases précédentes ; mais en théorie, on doit considérer que sont dans ce cas toutes les entités qui font partie des connaissances partagées entre le scripteur et ses lecteurs présumés (depuis la lune et le soleil jusqu'à Jésus-Christ ou Napoléon).

3.2. *Le calcul de la saillance*

Le principe du calcul de la saillance est repris du modèle d'Alshawi. Il est régi par un double mécanisme que l'on appellera *persistance-dégradation*. Il consiste d'une part à augmenter la saillance d'une entité chaque fois qu'elle est mentionnée dans le texte, et d'autre part à faire baisser les saillances de toutes les entités de manière régulière au fil du texte.

Commençons par la dégradation. Plus on s'éloigne de la dernière mention d'une entité, plus sa saillance diminue. La « distance » à prendre en compte ne doit pas être mesurée en nombre de phonèmes ou de mots, mais plutôt en nombre d'unités proprement textuelles telles que les phrases, les paragraphes et les chapitres. Ce sont en effet ces unités qui se révèlent pertinentes. Cela n'est pas étonnant. On sait qu'un certain nombre de traitements cognitifs d'intégration et de mémorisation ont lieu en fin de phrase, et ces traitements jouent naturellement un rôle dans la saillance des entités. De même, la mise en forme d'un texte en paragraphes, chapitres, etc. reflète en grande partie l'organisation thématique du discours, qui influence grandement les saillances. Le processus de dégradation consistera donc à diminuer les saillances à chaque frontière de phrase, paragraphe, chapitre, ou tout autre division du texte du même genre. La diminution pourra être plus

11. La prise en compte de ces entités préconstruites est l'une des principales améliorations que le modèle des attentes apporte au modèle du contexte d'Alshawi.
12. Cf. la notion de *scène verbale* ([VIC 96], p. 200 et [VIC 99]), qui peut être considérée comme une élaboration théorique de cette métaphore.

forte pour les ruptures les plus importantes : typiquement, on utilisera un facteur multiplicatif de l'ordre de 0,75 pour chaque frontière de phrase (on diminue donc du quart la saillance après chaque phrase), et de l'ordre de 0,25 ou moins pour les frontières de plus haut niveau. Le fait d'utiliser un facteur multiplicatif assure de toute façon une décroissance suffisamment rapide (de forme exponentielle) pour les entités qui ne sont plus mentionnées dans le texte : en les éliminant quand leur saillance descend en dessous d'un seuil, on évite d'encombrer inutilement la liste des entités actives dans le modèle des attentes.

Venons-en maintenant au mécanisme d'augmentation de la saillance. Il intervient à chaque mention d'une entité : on ajoute à la saillance existante une valeur qui dépend de la forme linguistique utilisée. Plusieurs facteurs entrent en ligne de compte avec plus ou moins de poids. En voici les trois principaux :

- *les marques de mise en relief* : nous regroupons sous ce terme deux types d'opérations linguistiques : la *thématisation* et la *focalisation*. La thématisation consiste à introduire un nouveau thème dans le discours. En français, cette opération est effectuée notamment par des présentatifs tels que *il y a*, *voici*, etc., ainsi que par des introducteurs comme *en ce qui concerne*, *quant à*, etc. Souvent, les entités mises en relief de cette façon n'ont pas été évoquées précédemment, mais elles sont mises d'emblée sur le devant de la scène, avec une saillance très forte. La focalisation correspond plutôt, dans la métaphore scénographique, à un mouvement de projecteur ou à un zoom de caméra sur une entité déjà présente en arrière-plan. En français, les marques les plus évidentes de focalisation sont les constructions clivées en *c'est ... qui*, *c'est ... que*. Là encore, ces formes linguistiques confèrent une très grande saillance aux entités auxquelles elles réfèrent ;

- *la hiérarchie des fonctions grammaticales et des relations actancielles* : suivant le rôle du groupe nominal ou du pronom dans la phrase, l'entité correspondante verra sa saillance augmenter d'une valeur plus ou moins grande. Deux hiérarchies de ce type ont été mises en évidence. D'une part, il y a, au niveau syntaxique, les fonctions grammaticales (sujet, objet, etc.) : en particulier, la position sujet confère une plus grande saillance que la position objet direct. D'autre part, il y a, au niveau sémantique, les relations actancielles, avec une plus grande saillance pour le rôle d'agent. La plupart du temps, ces deux hiérarchies se recoupent, l'agent occupant la position sujet, et le patient la position objet, ce qui pourrait laisser penser qu'elles sont redondantes. Mais une étude plus approfondie (notamment des énoncés à la voix passive, où le patient est en position sujet et l'agent, quand il est présent, occupe une fonction grammaticale de complément indirect) montre qu'il faut tenir compte des deux (en accordant sans doute, en français, un poids plus grand à la fonction grammaticale, le sujet étant généralement aussi le thème de l'énoncé). C'est ainsi que Givón, par exemple, a proposé d'utiliser conjointement les deux hiérarchies suivantes ([GIV 89], p. 212) :

  sujet > objet direct > objet indirect > autres compléments

  agent > bénéficiaire > patient > autres rôles

- *la distance syntaxique* : en plus de la fonction grammaticale, il faut prendre en compte un autre facteur, plus rarement mentionné dans la littérature : la position plus ou moins « enfouie » du groupe nominal dans la structure syntaxique de la phrase. Ainsi, à fonction grammaticale identique, le sujet d'une subordonnée confère moins de saillance que le sujet de la principale, et plus la subordonnée est à un niveau de récursion élevé, moins la saillance de son sujet devra être augmentée. De même, dans une cascade de compléments de noms telle que *le jardin de la maison du frère de Paul*, on a l'ordre de saillance suivant pour les entités correspondantes :

  'le jardin' > 'la maison' >'le frère' > 'Paul'

Nous n'avons donné ici que des règles qualitatives pour le calcul de la saillance. Une des difficultés de cette approche réside dans la nécessité de quantifier ces différents facteurs[13]. Comme nous le verrons plus bas (§4.1), nous préconisons de résoudre ce problème par une méthode de type expérimental.

13. Sans oublier que nous n'avons présenté que les principaux facteurs à prendre en compte : il en existe sans doute d'autres, plus spécifiques, dont les effets, même s'ils sont moins massifs, doivent être aussi quantifiés.

3.3. *Plage de saillance et concordance*

Résoudre le problème de l'identification des entités consiste à associer la bonne entité à chaque marque linguistique pertinente, syntagme nominal « plein » ou pronom[14]. Dans certains cas, il faut créer une nouvelle entité dans le modèle des attentes que l'on associe à la forme linguistique en question. C'est le cas notamment de la plupart des groupes nominaux introduits par l'indéfini *un*. Dans tous les autres cas, l'entité est considérée comme étant déjà présente dans le modèle des attentes, soit parce qu'elle a déjà été évoquée explicitement dans le texte, soit parce qu'elle fait partie des entités préconstruites, ou des entités susceptibles d'anaphores associatives, ou encore des connaissances partagées. L'identification revient alors à retrouver dans le modèle des attentes l'entité correspondant à la forme linguistique analysée. Comme chaque entité du modèle des attentes possède une saillance, il est important de savoir dans quelle fourchette de saillances on doit chercher cette entité. En effet, les formes linguistiques ne sont pas équivalentes de ce point de vue : un pronom réfère en général à une entité de plus forte saillance qu'un groupe nominal défini. Ainsi, suivant son degré de saillance, un même véhicule pourra être mentionné dans un constat par des expressions aussi diverses que :

La voiture immatriculée 356 XP 93
La voiture qui arrivait sur la droite
La voiture rouge
La voiture
Cette voiture
Celle-ci
Elle

Il est donc intéressant de *classer* les formes linguistiques en fonction des valeurs de saillance pour lesquelles elles sont utilisables. Wallace Chafe [CHA 87] propose ainsi de distinguer trois classes de formes linguistiques capables d'évoquer respectivement les « concepts » actifs, semi-actifs et inactifs, qui correspondent, grosso modo, à ce que nous appelons les entités de forte, moyenne et faible saillance. Mais c'est sans aucun doute la théorie de l'accessibilité de Mira Ariel [ARI 90] qui a le plus contribué à cette question. Ariel propose d'ordonner les expressions référentielles sur une *échelle d'accessibilité* : « I (will) argue that all referring expressions in all languages are arranged on a scale of Accessibility. Although actual marking systems are to some extent language-specific, for the most part they are all based on a principled connection between marker form and degree of Accessibility. The more informative, rigid (unambiguous), and unattenuated the marker, the lower the Accessibility it is specialized for, and vice versa » ([ARI 90], p. 29).

Les trois critères que prend en compte Ariel pour classer les expressions référentielles sont donc :

- *l'informativité*, c'est-à-dire la quantité d'information véhiculée par la forme linguistique. Par exemple, *la route nationale qui relie Cherbourg à Paris* et *la nationale 13* sont deux manières non ambiguës de désigner la même route, mais la première expression est plus informative que la seconde.

- *la sélectivité*, qui mesure le degré de spécificité de l'expression employée : ainsi a-t-on par degré de sélectivité décroissant *la Toyota rouge* > *la Toyota* > *la voiture* > *le véhicule*. De même un nom de famille est généralement plus sélectif qu'un prénom.

- *la légèreté*, qui regroupe plusieurs facteurs : la taille de l'expression (plus une expression contient de phonèmes, moins elle est légère), les aspects intonatifs (comme l'accentuation : ce facteur joue plus en anglais qu'en français), et le caractère plus ou moins marqué de la forme : un pronom démonstratif (*celui-ci, celle-ci, ceux-ci)* est moins léger qu'un pronom personnel plein (*lui, elle, eux*), qui est lui-même moins léger qu'un pronom clitique (*il, elle, ils*).

14. Rappelons (cf. § 1) que tous les syntagmes nominaux ne réfèrent pas à des entités (il en est de même pour les pronoms).

L'échelle d'accessibilité permet donc de spécifier une caractéristique importante des formes linguistiques, qui doit jouer un rôle essentiel dans l'identification des entités. Mais la sémantique des marques linguistiques de la référence ne se réduit pas à cette caractéristique. Par exemple, une expression comme *ce dernier* réfère de manière systématique à la « dernière » entité évoquée, et cela quelle que soit sa saillance, ou presque : en fait, puisque c'est une entité qui vient d'être évoquée, cette saillance est forcément assez forte, mais ce n'est qu'une conséquence du sémantisme du marqueur, qui doit seul être pris en compte. De même, si l'on est en présence du couple de marqueurs *celui-ci* et *celui-là*, c'est l'opposition proximal/distal qui doit cette fois être d'abord prise en compte : on doit rechercher deux entités de même type susceptibles de répondre à ce critère, dans la dimension temporelle (on a parlé de l'une avant l'autre), ou spatiale (en terme de distance par rapport au locuteur ou au centre de la scène), ou encore thématique (l'une est plus « proche » que l'autre du centre d'intérêt). Autre exemple, encore plus net : les pronoms personnels de première et deuxième personne ne peuvent désigner que des personnes qui joue un rôle précis dans l'interlocution, et l'on n'a aucunement besoin d'une échelle d'accessibilité pour les identifier[15].

Beaucoup de travaux de sémanticiens mettent l'accent sur les propriétés sémantiques spécifiques des différentes formes d'expressions référentielles (cf., entre autres, [COR 95], [CHA 02], [KLE 94a], [SCH 97], et [KLE 97]). Kleiber insiste à juste titre sur l'importance de ces facteurs sémantiques dans le calcul de la référence : « D'autres facteurs, essentiellement des facteurs cognitifs (accessibilité, pertinence, etc.) entrent en ligne de compte, qui militent pour des approches "moins" linguistiques. La légitimité de telles approches n'est pas à mettre en doute, bien au contraire. Elles comportent toutefois un risque, celui de céder à l'excès contraire en minimisant le rôle sémantique propre de chaque marqueur. Or, une telle occultation, partielle ou totale du sens, fréquente dans les travaux cognitivistes, est à notre avis dommageable aux études sur le fonctionnement des marqueurs référentiels » ([KLE 94a], p. 11).

Deux modifications à la théorie de l'accessibilité permettent de prendre en compte cet équilibre nécessaire entre le facteur d'accessibilité et les autres propriétés des formes linguistiques référentielles [DUP 03] :

- d'une part, on définit pour chaque forme linguistique une *plage des saillances admissibles*. Autrement dit, au lieu d'occuper un point sur l'échelle d'accessibilité, chaque marque est associée à un intervalle de cette échelle, qui peut être plus ou moins large suivant que la marque détermine plus ou moins précisément la saillance de l'entité à laquelle elle réfère.
- d'autre part, on définit un deuxième paramètre, *l'importance de la concordance*, qui indique le poids qu'il faut accorder dans le calcul d'identification des entités au paramètre précédent pour chaque forme linguistique.

Ainsi pour le pronom personnel *il*, la plage de saillance sera restreinte aux fortes saillances et l'importance de la concordance sera très élevée, ce qui revient à dire que la condition de forte saillance est impérative pour qu'une entité puisse être évoquée par ce pronom. En revanche, pour une forme comme *ce dernier*, l'importance de la concordance sera très faible, ce qui signifie que ce ne sont pas des contraintes sur la saillance qui comptent pour l'utilisation de cette forme, mais d'autres conditions sémantiques. En l'occurrence, pour *ce dernier*, c'est la *récence* de l'entité qui compte avant tout : pour pouvoir être reprise par *ce dernier*, une entité doit être la plus récente de sa classe à avoir été évoquée dans le texte précédent.

Ces deux paramètres rendent le modèle beaucoup plus souple que celui d'Ariel. Ils permettent d'adapter assez finement le calcul à la forme linguistique à analyser. En premier lieu, si

15. Ce qui ne veut pas dire que cette identification soit toujours aisée, notamment dans les textes de fiction, où il faut démêler le jeu de rôle, souvent subtil et complexe, entre le *je* de l'*auteur*, qui a écrit le texte, celui du *narrateur*, qui produit le récit à l'intérieur de l'univers raconté sans en être forcément partie prenante, et celui de l'*énonciateur*, à qui doit être attribué ce qui est dit et qui peut éventuellement être l'un des personnages du récit

l'importance de la concordance est faible, le calcul d'identification s'appuiera sur les autres conditions sémantiques associées à cette forme, la saillance des entités n'entrant en jeu que de manière secondaire. En second lieu, plus la plage de saillance d'une forme linguistique est large, moins cette concordance est contraignante puisqu'elle ne peut contribuer à éliminer que les entités dont la saillance est en dehors de cette plage.

Cette souplesse est essentielle notamment à cause d'un phénomène, appelé *compétition* par Ariel ([ARI 90], p. 28), qui joue un rôle très important dans le choix d'une expression référentielle par un locuteur. Quand plusieurs entités ont des degrés de saillance élevés, elles ont vocation à être désignées par des pronoms, ce qui peut poser des problèmes de compréhension. Prenons le cas d'une histoire dans laquelle trois personnages sont très saillants, une femme et deux hommes. Il n'y aura aucun problème pour utiliser le pronom *elle* pour le personnage féminin. En revanche les deux autres personnages sont en compétition pour le pronom *il*, et celui-ci sera donc assez systématiquement évité. Deux types de solutions sont possibles. On peut utiliser des marqueurs pour lesquels l'importance de la concordance est faible, comme *ce dernier*, *celui-ci*, *le premier*, *l'autre*, etc. (première forme de souplesse). On peut aussi utiliser des groupes nominaux plus complets (*L'homme au pardessus*, *le petit blond*, *le concierge*, etc.), qui sont habituellement employés pour des entités de saillance moyenne, ce qui signifie que la plage de saillance de ces formes linguistiques doit être très large pour englober à la fois les saillances moyennes et les saillances fortes (deuxième forme de souplesse). Notons qu'il est difficile de prendre en compte ce dernier cas dans un modèle qui fait correspondre un niveau de saillance précis à chaque marqueur. Ariel est ainsi amenée à considérer que la saillance des deux personnages masculins est plus faible que celle du personnage féminin à cause de la compétition, ce qui est manifestement peu satisfaisant.

3.4. *L'identification des entités*

Comme nous l'avons annoncé, nous séparons donc très nettement le calcul de la saillance et l'identification des entités. La saillance des entités entre bien sûr en ligne de compte dans l'identification des entités par l'intermédiaire de la concordance entre la saillance des entités et la plage des saillances admissibles de l'expression linguistique analysée. Mais ce n'est que l'un des facteurs intervenant dans le processus, et, comme on vient de le voir, il est plus ou moins important selon la forme linguistique concernée[16]. L'ensemble des facteurs à prendre en compte est alors le suivant.

- *facteurs morpho-syntaxiques* : il s'agit là des accords, en genre, nombre et personne, qui sont effectivement des règles simples sur lesquelles on peut s'appuyer. Comme nous l'avons déjà signalé, il existe des contre-exemples, qui empêchent de les considérer comme des règles intangibles. Nous avons déjà cité le cas d'un ministre, qui peut être repris par un féminin. C'est encore plus souvent le cas pour un mannequin, alors qu'une vigie ou une sentinelle sont généralement de sexe masculin. Signalons aussi le pluriel bien singulier du *nous* de modestie utilisé dans la rédaction d'une thèse par exemple. Cependant, ces cas sont rares, et pour la plupart, plutôt faciles à repérer[17].

- *règles syntaxiques* : un certain nombre de règles syntaxiques régentent, de manière intangible cette fois, les anaphores pronominales au sein d'une même phrase. Les plus célèbres sont sans aucun doute les règles issues de la théorie du liage [CHO 81], qui expliquent notamment que *elle* ne peut pas avoir pour antécédent *la concierge* dans les phrases suivantes :

16. Il est très important pour les pronoms personnels, ce qui explique sans doute que son rôle ait été quelque peu surévalué par les études centrées sur la résolution des anaphores pronominales.

17. Nous ne prenons pas en compte ici certains cas de métonymie intégrée dont les effets ressemblent, superficiellement, à des violations d'accord. On peut citer par exemple le passage de constat d'accident suivant : *La motocyclette a été perturbée par un véhicule qui circulait devant lui et a heurté ma voiture à l'arrêt…* Les problèmes posés par ces métonymies, dont nous avons déjà eu l'occasion de parler, sont bien entendu d'une autre nature : ils doivent être traités en tant que tels et non pas comme des violations d'accord.

Elle a dit que la concierge était fatiguée

Elle a vu le frère de la concierge

On peut aussi mettre dans cette catégorie les règles qui régissent l'emploi des pronoms réfléchis (et qui explique, cette fois, que *la* ne puisse pas co-référer avec *la concierge* dans la phrase : *La concierge la voit dans la glace*), ainsi que des règles portant sur les déterminants possessifs (cette fois, c'est *son* qui ne peut pas co-référer avec *la concierge* dans *Elle a vu la concierge de son immeuble*).

- *fonctions grammaticales* : les fonctions grammaticales ont aussi une influence sur le phénomène de l'anaphore pronominale. En effet le parallélisme des fonctions grammaticales entre un pronom et son antécédent est privilégié. C'est ainsi que, en l'absence d'indications contextuelles contradictoires, on associera plutôt *Il* avec *Jean* et *lui* avec *Paul* dans l'exemple suivant :

Paul a rencontré Jean. Il lui a raconté tous ses ennuis.

Ce parallélisme n'est pas à proprement parler une règle, mais plutôt une préférence, donc un indice à n'utiliser qu'en dernier lieu, si rien d'autre ne départage deux entités en compétition. En effet, il suffit d'une indication contextuelle contraire pour que la préférence tombe, comme le montre ce nouvel exemple :

Paul a raconté tous ses ennuis à son professeur. Il lui avait donné rendez-vous dans son bureau.

- *anaphores et cataphores* : on parle de *cataphore* au lieu d'anaphore quand l'antécédent se situe après le pronom, comme dans l'exemple suivant : *Quand il est rentré chez lui, Paul était tout essoufflé.* La cataphore réclame un effort cognitif plus important de la part de l'auditeur/lecteur puisque celui-ci doit attendre d'avoir rencontré l'antécédent avant de pouvoir identifier le référent du pronom. Il est donc logique de penser que le locuteur évitera un tel procédé s'il y a compétition avec un autre candidat déjà disponible au moment où l'on rencontre le pronom. En cas de compétition, le modèle privilégie donc un antécédent en position anaphorique à un antécédent en position cataphorique.

- *sémantique des marqueurs grammaticaux* : à chaque marque grammaticale sont associées des instructions plus ou moins contraignantes. Comme nous avons déjà eu l'occasion d'en discuter, pour un certain nombre d'entre elles, telles que *celui-ci, celui-là, ce dernier, l'un, l'autre,* etc., ces instructions sont suffisamment impératives pour guider à elles seules la recherche de l'entité qu'elles évoquent. Dans le modèle, cela revient notamment à attribuer à ces marques une très faible valeur au paramètre 'importance de la concordance'.

- *sémantique lexicale* : on peut distinguer deux groupes de règles. Les premières s'appliquent aux reprises par un groupe à tête nominale (par opposition à un pronom). La reprise se fait le plus souvent en reprenant une partie du groupe nominal qui a servi à introduire l'entité la première fois ou l'un de ses synonymes ou hyperomymes. Ainsi *un camion semi-remorque transportant du gravier* sera-t-il repris par *le camion, le semi-remorque, le poids lourd, le véhicule, l'engin,* etc. La connaissance des relations sémantiques lexicales de synonymie et d'hyperonymie est donc très utile pour résoudre ce type d'anaphore. Le deuxième groupe de règles concerne les anaphores pronominales : il s'agit des restrictions de sélection opérées notamment par les verbes sur leurs actants. Par exemple le verbe *manger* réclame généralement un sujet humain ou animal et un objet comestible. Cette règle permet de résoudre correctement les anaphores du petit texte suivant (alors que la règle du parallélisme des fonctions grammaticales, vue ci-dessus, est prise en défaut :

Cette tarte n'est pas terrible, si on écoute Marie. Elle en a quand même mangé la moitié à elle toute seule hier soir, remarque !

Ce qu'il faut pour pouvoir exploiter ces indices, c'est donc disposer de classes sémantiques lexicales. Il faut cependant noter que ces règles ne sont pas très simples à utiliser : outre la polysémie verbale (la rouille mange le fer, et la petite vérole peut manger un visage…), les phénomènes de coercition de type [PUS 95] doivent aussi être pris en compte (des tables peuvent aussi manger, comme le montre l'exemple -attesté- suivant : *Les tables qui ont fini de manger se mettent en rang par deux et sortent en silence !*).

- *connaissances de sens commun* : enfin, comme nous l'avons fait remarquer dès l'introduction de ce chapitre, des connaissances extralinguistiques sont nécessaires dans un certains nombre de cas. D'abord pour traiter des anaphores associatives, qui reposent très massivement sur des connaissances encyclopédiques (nous avions vu les exemples de la pompe d'une station service et

celle d'une pédale de frein dans une voiture). Elles sont aussi indispensables pour guider les inférences dont nous avions aussi vu la nécessité. Soit les deux exemples suivants :

Jean a emprunté une voiture à son garagiste après s'être fait voler la sienne. Il l'a rendue ce matin.

Jean a emprunté une voiture à son garagiste après s'être fait voler la sienne. Il l'a retrouvée ce matin.

Il est clair que seules des connaissances extralinguistiques sur les « scénarios » associés aux vols, aux emprunts, etc. peuvent permettre de retrouver dans chaque cas de quelle voiture on parle dans la deuxième phrase. Enfin elles servent à déterminer les référents de certains marqueurs, comme *nous*, dont l'interprétation précise repose souvent sur ces connaissances de bon sens. Ainsi, dans les constats d'accidents, dans une phrase du type *Nous roulions à faible allure*, le pronom désigne le locuteur et les autres occupants de son véhicule, alors que dans *Nous avons rempli ce constat* il désigne le locuteur et son adversaire.

Ainsi, un système d'identification des entités doit travailler sur tous les niveaux de la linguistique, depuis la morpho-syntaxe jusqu'à la pragmatique. Comme on l'a vu, il existe très peu de règles sûres, vérifiées dans 100% des cas. La plupart expriment plutôt des « tendances », des « préférences », du moins au degré de généralité dans lequel nous les avons formulées. Il serait sans doute possible de les complexifier afin de les rendre plus précises, plus fines et donc plus fiables, mais elles seraient alors par ce fait même plus difficilement implémentables dans un système automatique. Si l'on a cet objectif d'implémentation, la question qu'il vaut mieux se poser est la suivante : quel poids respectif doit-on attribuer aux règles de niveaux différents, de manière à optimiser les performances, sachant que le système fera toujours des erreurs dans la mise en œuvre de ces règles ? En particulier y a-t-il suffisamment de *redondance* dans le système linguistique pour que l'on puisse attribuer des poids nuls, c'est-à-dire se passer complètement des règles de certains niveaux (les plus difficiles à automatiser, dont l'implémentation sera donc peu fiable), sans trop dégrader les performances ?

La réponse à cette question, comme on va maintenant le voir, passe par la mise en place d'un système capable de mener à bien des études expérimentales.

4. Un système d'expérimentation : CALCOREF

4.1. *Un logiciel voué à l'expérimentation*

Michel Dupont a conçu et réalisé un logiciel d'étude, CALCOREF, qui implémente les principes théoriques présentés ci-dessus [DUP 03]. Ce système est capable de traiter des textes tout venant. Il a pour finalité première de permettre d'expérimenter le modèle, de le valider, en ajustant une certain nombre de paramètres, depuis le choix d'un algorithme général faisant alterner la mise à jour des saillances dans le modèle des attentes avec le calcul d'identification des entités proprement dit, jusqu'aux ajustement de divers paramètres quantitatifs : poids à attribuer aux différents facteurs intervenant dans ces deux processus , seuils à mettre en place pour définir différentes classes de saillance, ou encore valeurs à donner au processus de persistance-dégradation.

Il faut d'ailleurs noter que ce type d'étude expérimentale a pris de plus en plus d'importance au cours de la dernière décennie pour le traitement automatique de phénomènes de haut niveau en linguistique informatique. Comme l'ont mesuré Walker et Moore [WAL 97] : « In recent years, there has clearly been a groundswell of interest in empirical methods for analysing discourse. A survey of recent ACL papers shows that the percentage of empirical papers in semantics, pragmatics, and discourse hovered between 8% and 20% until 1993 when it increased to 40%. In 1995 and 1996, 75% of the ACL papers in semantics, pragmatics, and discourse used empirical

methods ». Ces méthodes empiriques, au-delà de leur intérêt général (comparaison de différentes approches et de différents paramétrages sur une base objective, analyse des erreurs pour guider les améliorations ultérieures en les ordonnant par degré d'importance suivant leur impact sur les performances, etc.), sont particulièrement pertinentes dans le cas de notre modèle. En effet, ce dernier se caractérise par la très grande diversité des facteurs impliqués, allant d'indices morpho-syntaxiques, assez faciles à calculer, jusqu'à des connaissances de sens commun encore largement hors de portée d'un système informatique, excepté dans des domaines très restreints et bien délimités. Il est donc décisif de pouvoir tester la redondance entre ces différents facteurs pour savoir dans quelle mesure on peut se passer des indices de plus haut niveau, et seule l'expérimentation permet de le faire.

Pour pouvoir mener à bien de tels tests, il convient de résoudre un double problème de *spécification* et *d'évaluation*. La spécification concerne le but à atteindre : il s'agit de déterminer quel est le résultat qui sera considéré comme une bonne réponse du système, autrement dit, dans notre cas, quelles sont précisément les chaînes de coréférence à détecter. Cette question est moins simple qu'il n'y paraît, à cause de problèmes de fond dont nous allons simplement mentionner ici quelques exemples, mais qui mériteraient de plus longs développements[18]. D'abord, comme nous l'avons déjà signalé, les métonymies et autres phénomènes du même genre posent le problème de l'évocation implicite d'entités, qu'il est délicat de faire figurer dans une chaîne de coréférence. Autre problème épineux : une entité peut changer de qualification au cours du temps. Ainsi, un même individu, Jacques Chirac, a été Maire de Paris avant de devenir Président de la République. Si on parle de ces deux périodes dans un même texte, doit-on considérer que les mentions *le Maire de Paris* et *Le Président de la République* sont coréférentielles, parce que c'est le même individu qui est désigné ainsi, ou doit-on faire deux chaînes différentes, en distinguant un Chirac_1 et un Chirac_2 qui se sont succédés dans le temps ? Dernier exemple : les entités collectives. Supposons qu'un texte parle d'un groupe d'individus en alternant les références au groupe dans son ensemble (*Ils se sont avancés, Ils nous ont menacés,* etc.) et à chacun des individus (*Le premier portait des lunettes noires, L'un d'entre eux a sorti un couteau,* etc.), doit-on construire des chaînes séparées pour le groupe et pour chacun des individus ou faut-il considérer que chaque chaîne individuelle doit inclure, d'une façon ou d'une autre, les références au groupe ? Les choix que l'on fait sur ces différentes questions, et qui vont comporter inévitablement une part d'arbitraire, doivent être explicités le plus soigneusement possible, si l'on veut comparer sur une même base les résultats obtenus par différents systèmes. Les deux dernières conférences MUC[19], qui comportaient une tâche de calcul de coréférence, ont grandement contribué à faire émerger ces problèmes de spécification.

Quant à l'évaluation, elle consiste à définir des mesures de l'écart entre le résultat désiré et le résultat réellement obtenu par le système. On définit habituellement (cf. chap. 8, §2.4) une mesure de *précision* qui cherche à évaluer le nombre de réponses correctes par rapport au nombre de réponses du système, et une mesure de *rappel*, pour le nombre de réponses correctes par rapport au nombre de réponses attendues. Dans le cas des chaînes de coréférence, ces mesures doivent être adaptées puisqu'il s'agit de comparer deux partitions différentes de l'ensemble des expressions linguistiques référentielles d'un texte. Là encore les conférences MUC ont permis de définir des mesures appropriées. Sans entrer dans les détails ici (on pourra se reporter à [VIL 95] et [DEE 00]), le principe consiste à définir chaque chaîne de manière canonique par un nombre minimal de liens (*n*-1 pour une chaîne de longueur *n*), puis à compter le nombre d'erreurs de type *bruit* (coréférences en trop dans une chaîne), noté N_{bruit}, et de type *silence* (coréférences manquantes dans une chaîne), noté $N_{silence}$. La précision et le rappel sont alors donnés par les formules suivantes :

18. Voir à ce sujet le travail de Salmon-Alt [SAL 01] qui a cherché à définir un format général d'annotation de chaînes de coréférence pour des corpus électroniques.
19. *Message Understanding Conference* : on trouvera une présentation de ces conférences au chapitre 8 (§1).

Précision = $1 - (N_{bruit} / Total_{obt})$

Rappel = $1 - (N_{silence} / Total_{dés})$

où $Total_{obt}$ et $Total_{dés}$ désignent respectivement le nombre total de liens dans les résultats obtenus et dans les résultats désirés.

Même si ces mesures ne sont pas parfaites (certaines erreurs de résolution d'anaphores sont comptées deux fois, une fois comme silence pour une chaîne et une fois comme bruit pour une autre, alors que d'autres erreurs ne sont comptées qu'une fois), elles sont suffisantes pour comparer les performances de plusieurs systèmes ou de plusieurs paramétrages d'un même système.

4.2. *L'architecture générale*

CALCOREF est constitué de deux modules organisés de manière séquentielle :

- un module syntaxico-sémantique qui prend en entrée un texte tout venant et qui construit une représentation de ce texte contenant toutes les informations syntaxiques et sémantiques qui seront utilisées dans le calcul de la référence ;
- le module de calcul de coréférence proprement dit, qui détermine, à partir de cette représentation, les chaînes de coréférence du texte de longueur au moins égale à deux (les chaînes réduites à une seule expression linguistique sont écartées). A titre d'illustration, on trouvera ci-dessous dans le tableau 3 la sortie de CALCOREF sur le premier texte sur lequel il a été testé : les trois premiers paragraphes de *La peste* d'Albert Camus.

Le module syntaxico-sémantique utilise l'analyseur syntaxique développé par Jacques Vergnes [VER 01] qui a été adapté pour les besoins de cette application. Notamment ont été ajouté un traitement spécifique de certaines entités nommées (noms de personne, dates, etc. : cf. chapitre 8, § 1.2 et § 2.1), un étiquetage sémantique des syntagmes nominaux de types spatial et temporel (pour éviter de les prendre en compte dans l'identification des entités), une intégration au groupe verbal de certains autres groupes nominaux (comme par exemple *le passage* dans *laisser le passage* ou *garde* dans *prendre garde*), et la détection d'un certain nombre de pronoms impersonnels (*il y a, il faut, il s'agit,* etc.).

C'est le module de calcul de coréférence qui implémente le modèle d'identification des entités à proprement parler. On retrouve les deux aspects mentionnés ci-dessus, qu'il convient de bien distinguer (cf. § 3.4) : d'une part, la gestion de la liste des entités déjà évoquées avec leur saillance, conformément au modèle des attentes ; et d'autre part l'appariement entre les expressions linguistiques référentielles et les entités (deux expressions font partie de la même chaîne de coréférence si elles sont associées à la même entité).

L'algorithme fonctionne phrase par phrase. Il commence par dégrader les saillances des entités déjà existantes. Puis il dresse la liste des expressions linguistiques référentielles de la nouvelle phrase. Il traite alors ces expressions « au fil du texte », au sens où il parcourt la liste des expressions linguistiques dans l'ordre en associant au fur et à mesure chacune d'entre elles à une entité et en mettant à jour aussitôt la saillance de cette entité[20]. Ainsi, si la même entité est évoquée à plusieurs reprises dans la même phrase, sa saillance croît au gré de ces occurrences.

L'appariement entre expressions linguistiques et entités utilise toute une série de critères en plus de la concordance[21]. Ainsi, pour les groupes nominaux définis, l'appariement se fait d'abord sur la base de la reprise d'unité lexicales du groupe nominal qui a servi à introduire l'entité (par exemple,

20. En fait, pour pouvoir traiter correctement les cataphores, l'algorithme comporte deux passes, mais la première passe ne sert, fondamentalement, qu'à la création de nouvelles entités.

21 i.e., rappelons-le, l'adéquation entre la saillance de l'entité et la plage de saillances admissibles de l'expression.

si une entité a été introduite par l'expression *une voiture blanche*, on considèrera que *la voiture* et *la blanche* peuvent être des reprises de cette entité). Pour les pronoms, les accords et les règles syntaxiques (comme l'impossibilité pour les pronoms *le* et *lui* de coréférer avec le sujet) servent d'abord de filtre. D'une manière générale, pour chaque type d'expression linguistique, un certain nombre de règles spécifiques éliminent une partie des candidats, et, dans un deuxième temps, s'il reste plusieurs compétiteurs, d'autres critères interviennent dans un *calcul de score*, où chacun de ces facteurs intervient avec un certain poids. Ce poids peut d'ailleurs dépendre du type d'expression considérée. Par exemple, dans la première version de CALCOREF, le calcul de score prend en compte le parallélisme de fonctions uniquement pour les pronoms personnels, et une valeur de « récence »[22] uniquement pour les adjectifs possessifs.

L'interface utilisateur de CALCOREF joue un rôle très important. Elle permet d'abord d'afficher les différentes étapes de traitement, en détaillant à la demande tel ou tel aspect. Mais surtout elle permet de modifier très facilement les différents paramètres de l'algorithme : calcul de saillance (pourcentage de dégradation après chaque phrase et chaque paragraphe ; valeur ajoutée à chaque mention selon sa position syntaxique dans la phrase, etc.), seuils de définition des classes de saillance (forte, moyenne, faible, éliminatoire), et poids des différents facteurs dans le calcul de score. Ainsi l'utilisateur peut immédiatement observer l'effet de telle ou telle modification de paramètre sur les résultats du calcul, et donc mener à bien le type d'expérimentation pour lequel le système a été conçu.

Le matin du 16 avril, le docteur Bernard Rieux$_1$ sortit de son$_1$ cabinet et buta sur un rat mort$_2$ au milieu du palier$_3$. Sur le moment, il$_1$ écarta la bête sans y prendre garde et descendit l'escalier. Mais, arrivé dans la rue, la pensée lui$_1$ vint que ce rat$_2$ n'était pas à sa$_2$ place et il$_1$ retourna sur ses$_1$ pas pour avertir le concierge$_4$. Devant la réaction du vieux M. Michel$_5$, il$_1$ sentit mieux ce que sa$_1$ découverte avait d'insolite. La présence$_6$ de ce rat mort$_2$ lui$_1$ avait paru seulement bizarre tandis que, pour le concierge$_4$, elle$_6$ constituait un scandale$_7$. La position de ce dernier$_7$ était d'ailleurs catégorique : il n'y avait pas de rats$_8$ dans la maison$_9$. Le docteur$_1$ eut beau l$_7$'assurer qu'il y en avait un sur le palier$_3$ du premier étage, et probablement mort, la conviction de M. Michel$_5$ restait entière. Il n'y avait pas de rats$_8$ dans la maison$_9$, il fallait donc qu'on eût apporté celui-ci$_1$ du dehors. Bref il s'agissait d'une farce.

Le soir même, Bernard Rieux$_1$, debout dans le couloir de l'immeuble, cherchait ses$_1$ clefs avant de monter chez lui$_1$, lorsqu'il$_1$ vit surgir, du fond obscur du corridor, un gros rat$_{10}$ à la démarche incertaine et au pelage mouillé. La bête$_{11}$ s$_{11}$'arrêta, sembla chercher un équilibre, prit sa$_{11}$ course vers le docteur$_1$, s$_{11}$'arrêta encore, tourna sur elle-même$_{11}$ avec un petit cri et tomba enfin en rejetant du sang$_{12}$ par les babines entrouvertes. Le docteur$_1$ la$_{11}$ contempla un moment et remonta chez lui$_1$.

Ce n'était pas au rat$_{10}$ qu'il$_1$ pensait. Ce sang rejeté$_{12}$ le$_1$ ramenait à sa$_1$ préoccupation. Sa$_1$ femme$_{13}$, malade depuis un an, devait partir le lendemain pour une station de montagne. Il$_1$ la$_{13}$ trouva couchée dans leur chambre, comme il$_1$ lui$_{13}$ avait demandé de le faire. Ainsi se$_{13}$ préparait-elle$_{13}$ à la fatigue du déplacement. Elle$_{13}$ souriait.

Tableau 3. *La sortie de CALCOREF sur un extrait de* La peste.

4.3. *Premiers résultats*

Les premières expérimentations du système ont été menées en opérant un certain nombre de simplifications. C'est ainsi qu'il n'y a que deux classes de saillance, les saillances actives, qui

22. Plus précisément, cette récence est donnée par une distance entre l'expression traitée et la dernière évocation de l'entité, calculée en nombre d'expressions référentielles qui séparent ces deux expressions. Dans *Il a percuté le mur avec sa moto*, la distance entre *sa* et *le mur* est nulle, et la distance entre *sa* et *Il* est de 1.

regroupent, de fait, les saillances fortes, moyennes et faibles, et les saillances inactives, qui correspondent aux entités éliminées du modèle des attentes. Il faut noter cependant que la valeur quantitative des saillances actives est prise en compte dans le calcul de score pour les expressions pronominales (y compris les adjectifs possessifs). Par ailleurs, si les accords et les principales règles syntaxiques sont implémentés, aucune connaissance lexicale n'est utilisée, ni les relations de synonymie et d'hyperonymie, ni les restrictions de sélection. En revanche, les fonctions syntaxiques sont utilisées, à la fois pour le calcul de la saillance et pour le calcul de score des pronoms personnels (parallélisme de fonctions). Enfin, la sémantique des marqueurs grammaticaux n'est que très partiellement implémentée. Malgré ces limites, les performances de cette première version sont tout à fait encourageantes, comme on pourra le constater en observant les résultats obtenus sur l'extrait de *La peste*, présentés ci-dessus (tableau 3).[23]

D'un point de vue qualitatif, plus intéressant à ce stade de développement du projet, on peut classer les erreurs en plusieurs catégorie,s que nous allons illustrer en analysant les sorties du système sur l'extrait de La Peste. D'abord des erreurs facilement évitables si l'on ajoute des connaissances lexicales de base, comme la relation d'hyperonymie entre *rat* et *bête*, qui permettrait de récupérer deux anaphores nominales, l'une dans le premier paragraphe (*il écarta la bête*), et l'autre dans le second (*La bête s'arrêta*). Il en est de même pour les relations entre *maison* et *immeuble*, et entre *couloir* et *corridor*.

En revanche, le lien entre *M. Michel* et *le concierge*, qui n'est pas fait non plus par le système semble beaucoup plus difficile à établir. A priori, seule une compréhension en profondeur du texte permettrait d'éviter l'erreur. Il faut noter que le moyen le plus efficace de ne pas faire ce genre d'erreurs serait de disposer de bases de données, basées sur des listes d'entités nommées, qui fournissent les équivalences entre les différentes manières de parler d'une entité, par exemple entre une personne et sa fonction. Il est clair que cette technique n'est pas applicable aux romans (impossible de savoir que M. Michel est le concierge avant d'avoir lu le début de La peste), mais elle est tout à fait envisageable dans d'autres domaines : par exemple, on peut disposer de l'information que *Jacques Chirac* et *le Président de la République* sont une seule et même entité, du moins si certaines conditions sont remplis (par exemple, pour des journaux français à telle période).

Une autre erreur dont la correction réclame un calcul du sens assez approfondi concerne la résolution de l'anaphore *ce dernier*, dans le premier paragraphe, auquel le système a donné comme antécédent *un scandale*. En fait, pour obtenir le bon référent (le concierge), on peut s'appuyer sur une restriction de sélection qui ne porte pas uniquement sur *position*, qui est bien trop polysémique, mais sur *position catégorique* : seul un humain est susceptible d'avoir des positions catégoriques, mais l'on conçoit la difficulté d'implémenter ce type de subtilité. Là encore l'erreur semble inévitable à moins d'une compréhension de l'ensemble de la scène. Enfin les deux dernières erreurs réclament toutes les deux de prendre en compte de manière plus approfondie le fonctionnement de marqueurs grammaticaux particuliers. La première semble plus à la portée d'un système automatique : elle concerne la proposition *il y en avait un sur le palier du premier étage*, où une bonne analyse de cet emploi de *en* permet de retrouver l'antécédent *rat* dans la phrase précédente (dans la construction parallèle *il n'y avait pas de rat dans cette maison*). Quant à la seconde erreur, elle est à double détente, si l'on peut dire. Elle porte sur *celui-ci*, à la fin du premier paragraphe, que le système a attribué au docteur Rieux. En codant de manière plus fine l'instruction sémantique associée à *celui-ci*, on peut sans aucun doute écarter cet antécédent, mais sans doute au profit de l'autre protagoniste de cette petite scène, M. Michel, ce qui est tout aussi faux ! En effet, un dans son emploi le plus fréquent, *celui-ci* sert à sélectionner, parmi deux entités en compétition, celle dont on vient de parler en dernier lieu. Or dans la liste des entités les plus saillantes, deux entités

23 D'un point de vue quantitatif, les mesures de précision et de rappel, calculées à partir des formules données plus haut (§ 4.1), sont respectivement de 96% et de 81% sur cet extrait. Ces indications devraient évidemment être confortées par une expérimentation sur un large corpus.

sont bien en compétition : le docteur et le concierge, et, des deux, c'est le concierge qui a été évoqué en dernier. Le vrai référent de *celui-ci*, en l'occurrence le rat mort, est d'autant plus difficile à trouver qu'il n'est évoqué qu'indirectement dans la proposition précédente par le pluriel *rats*, qui ne fait pas, à juste titre, partie de la chaîne de coréférence du rat mort… Comme on le voit, une erreur, facile à corriger, peut en cacher une autre, bien plus épineuse !

Ainsi l'analyse qualitative des erreurs de cette première version de CALCOREF montre que diverses améliorations peuvent être envisagées . La plus « payante » dans l'immédiat semble être la prise en compte de relations sémantiques lexicales : elles sont relativement faciles à implémenter, étant donné la disponibilité croissante de ressources électroniques de ce type (dictionnaires de synonymes et autres), et elles peuvent éviter une proportion importante des erreurs. Bien entendu, l'analyse de ce court extrait ne saurait suffire : il faut traiter un grand nombre de textes, les plus variés possible, avant de tirer ce type de conclusion. C'est précisément l'objectif de ce « logiciel d'étude » que de permettre ce type d'expérimentation.

5. Conclusion

Les premiers résultats obtenus avec CALCOREF montrent que l'on peut réaliser des implémentations suffisamment « légères » du modèle pour qu'elles puissent être intégrées dans des systèmes devant traiter des corpus textuels de grande taille. Comme nous l'avons dit en introduction, de nombreuses applications, notamment en informatique documentaire, pourraient bénéficier de l'apport de ce type de traitement. Nous avions cité l'indexation automatique de documents et la recherche d'information (cf. chap 8 resp. § 2 et § 3). Il faut noter que dans ces deux cas, les performances du module relativement « minimaliste » que nous avons présenté semblent déjà bien suffisantes pour apporter une amélioration sensible des résultats, surtout s'il se confirme que les erreurs de CALCOREF sont plus massives en *rappel* qu'en *précision* (cf. les définitions données § 4.1). En effet, pour de tels systèmes, seules les erreurs en précision constituent des erreurs pénalisantes, puisqu'elles reviennent à attribuer une forme linguistique à une autre entité que l'entité pertinente. Les erreurs en rappel (qui provoquent la coupure d'une chaîne de coréférence en deux chaînes distinctes) ne font que passer sous silence une partie des occurrences de l'entité pertinente : elles limitent donc l'apport du module, mais ne peuvent en aucun cas détériorer les résultats du système. Il y a donc là, nous en sommes convaincus, une voie extrêmement prometteuse, en particulier pour les systèmes de veille documentaire qui reposent aujourd'hui en grande partie sur l'utilisation de grosses bases de données d'entités nommées : un module de type CALCOREF les rendrait plus efficaces en repérant la plupart des reprises anaphoriques des entité nommées cibles de la recherche.

En ce qui concerne les applications qui réclament des analyses sémantiques plus importantes, il faut sans aucun doute une implémentation plus complète du modèle d'identification des entités. Notamment, dans le domaine de l'Extraction d'Information, un certain nombre de caractéristiques du modèle, non encore implémentées, devraient s'avérer très utiles [DUP 02], comme le recours aux entités pré-construites (§ 3.1) limitées au domaine visé, et la mise en œuvre de traitements de certaines anaphores associatives et de certaines métonymies, limitées là aussi à celles qui sont les plus fréquentes dans le domaine dans lequel on opère[24].

Enfin il faut souligner qu'il reste un grand nombre de problèmes de modélisation proprement dit que nous avons volontairement écartés de cette présentation (cf. les problèmes de spécification à peine évoqués § 4.1) parce qu'ils sont pour l'heure encore loin d'être résolus, et qui constituent donc autant de thèmes de recherche à approfondir. On peut ainsi citer le problème des collectifs d'entités, qui forcent à complexifier les chaînes de coréférence : une entité donnée pouvant de fait

24 Le chapitre 8, § 3.4, donne un exemple significatif de situation de ce type.

appartenir à plusieurs chaînes de coréférence, l'une pour sa représentation individuelle, les autres pour la représentation de collectifs de différents niveaux dont elle peut faire partie. On a un exemple de ce phénomène dans le petit extrait de *La Peste* que nous avons analysé, à l'avant-dernière phrase (voir tableau 3) : *Il la trouva couchée dans leur chambre, comme il lui avait demandé de le faire*. Il s'agit du *leur* de *leur chambre*, dont l'antécédent est une entité collective, le couple formé par Bernard Rieux et sa femme, qui n'a pas été introduit explicitement dans le texte auparavant, et qu'il faut donc construire d'une manière ou d'une autre[25]. On peut aussi sans doute traiter dans le même cadre le problème des reprises par des pronoms génériques (ex. : *J'ai acheté une Toyota. Elles sont très robustes*), qui peuvent d'ailleurs être très subtils [KLE 91], comme dans l'exemple suivant : *le train de Paris est déjà arrivé ? Pourtant, d'habitude, il a au moins 10 minutes de retard*, où le pronom *il* réfère à un train générique dont l'exemplaire déjà mentionné est visiblement très peu représentatif. Autres problèmes, que nous citons en vrac et sans prétention d'exhaustivité : les pronoms *ils* dits *sans antécédent*, à l'œuvre dans un exemple tel que *A Paris, ils roulent comme des fous* [KLE 94a] ; les référents dits *évolutifs* ([CHA 93], [SCH 93]), qui changent de forme, de statut, voire perdent même leur intégrité physique, tout en restant dans la même chaîne anaphorique (ex. : *Prenez un poulet, tuez-le, plumez-le, découpez-le, désossez-le, hachez-le menu, faites-le cuire mélangé à des tomates et des oignons et servez-le tout chaud !*) ; les problèmes redoutables d'*opacité référentielle*, qui conduisent à des reprises superficiellement contradictoires, comme dans *J'ai rencontré la suédoise que Paul a épousée : en fait, elle est danoise !* ; etc. Comme on peut le constater, il reste fort à faire dans ce domaine de recherche, et l'intérêt de ces travaux de modélisation ne concerne pas que les applications : ils peuvent aussi aider à faire avancer, au plan théorique, notre compréhension de ces phénomènes référentiels qui sont au cœur même de l'activité de langage. Un système, suffisamment général et souple, tel que CALCOREF peut permettre de conduire des expérimentations à l'appui de ces études.

Bibliographie

[ALS 87] ALSHAWI H., *Memory and context for language interpretation*, Cambridge Unversity Press, 1987.

[ARI 90] ARIEL M., *Accessing Noun Phrases Antecedents*, Londres, Routledge, 1990.

[AZZ 98] AZZAM S., HUMPHREYS K., GAISAUSKAS R., « Evaluationg a focus-based approach to anaphora resolution », *Proceedings COLING-ACL'98*, Montreal, 1998.

[CHA 72] CHARNIAK E., Towards a model of children's story comprehension, AI memo 266, AI Lab, MIT.

[CHA 78a] CHARNIAK E., « On the use of framed knowledge in language comprehension », *Artificial Intelligence*, 11, 1978, p. 225-265.

[CHA 78b] CHARNIAK E. , « With spoon in hand this must be the eating frame », in D.L. Waltz (ed.), *Theoretical issues in natural language processing*, 2, 1978, p. 187-198.

[CHA 87] CHAFE W., « Cognitive constraints on information flow », in R. Tomlin (ed.), *Coherence and grounding in discourse*, Amsterdam, Benjamins, p. 21-52, 1987.

[CHA 93] CHAROLLES M., SCHNEDEKER C., « Coréférence et identité », *Langages*, 112, 106-127, 1993.

[CHA 02] CHAROLLES M., *La référence et les expressions référentielles en français*, Gap, Ophrys, 2002.

[CHO 81] CHOMSKY N., *Lectures on government and binding*, Dordrecht, Foris, 1981.

[COR 95] CORBLIN F., *Les formes de reprise dans le discours. Anaphores et chaînes de référence*, Rennes, Presses Universitaires de Rennes, 1995.

[DEE 00] DEEMTER K., KIBBLE R., « On corefering in MUC and related annotation schemes », *Computational Linguistics*, 26(4), p. 629-637, 2000.

25. Sur cette question et sur les problèmes connexes : liens d'*altérité* (ex. : *Prends un cube. Prends-en un autre*), et liens d'*extraction* (ex. : *Prends deux cubes. Mets le premier sur la table*), voir les travaux de Salmon-Alt [SAL 01] et notamment ses propositions en terme de codage.

[DUC 72] DUCROT O., TODOROV T., *Dictionnaire encyclopédique des sciences du langage*, Editions du Seuil, 1972.

[DUP 96] DUPONT M, « Le modèle des attentes », *Actes de ILN'96*, Nantes, 1996, p. 219-229.

[DUP 98] DUPONT M., « Le calcul de la référence dans un système de compréhension automatique limité de corpus homogènes », *Travaux linguistiques du CERLICO*, 11, Presses Universitaires de Rennes, 1998, p. 237-259.

[DUP 02] DUPONT M., VUILLAUME J.M., VICTORRI B., ENJALBERT P., MATHET Y., MALANDAIN N., « Nouvelles perspectives en extraction d'information », *TSI*, 21(1), p. 37-63, 2002.

[DUP 03] DUPONT M., Une approche cognitive du calcul de la référence, thèse de l'Université de Caen, 308 p., 2003.

[GIV 89] GIVÓN T., *Mind, code and context: essays in pragmatics*, Laurence Elbaum Ass., 1989.

[GOR 93] GORDON P., GROSZ B, GILLIOM L., « Pronouns, Names, and the centering of attention in discourse », *Cognitive Science*, 17, 1993, p. 311-347.

[GRO 77] Grosz B., « The representation and use of focus in a system for understanding dialogues », *Proceedings 5th IJCAI*, Cambridge, 1977.

[GRO 86] Grosz B, Sidner C., « Attention, intentions, and the structure of discourse », *American Journal of Computational Linguistics*, 12(3), 1986, p. 175-204.

[GRO 95] GROSZ B, JOSHI A., WEINSTEIN S., « Centering: a framework for modeling the local coherence of discourse », *Computational Linguistics*, 21(2), 1995, p. 203-225.

[GRO 98] Grosz B, Sidner C., « Lost intuitions and forgotten intentions », in M. Walker *et al.* (eds), *Centering theory in discourse*, Oxford, Clarendon Press, 1998.

[HAJ 95] Hajicova E., Skoumalova H., Sgall P., « An automatic procedure for topic-focus identification », *Computational Linguistics*, 21(1), 1995, p. 81-94

[KLE 91] KLEIBER G., « Anaphore-déixis : où en sommes-nous ? », *L'information grammaticale*, 51, p. 3-16, 1991.

[KLE 94a] KLEIBER G., *Anaphores et pronoms*, Louvain-la-Neuve, Duculot, 1994.

[KLE 94b] KLEIBER G., « Contexte, interprétation et mémoire : approche standard vs approche cognitive », *Langue Française*, 103, p. 9-22, 1994.

[KLE 97] KLEIBER G., SCHNEDECKER C., TYVAERT J.E., *La continuité référentielle*, Paris, Klincksieck, 1997.

[LAP 94] Lappin S., Leass H., « An algorithm for pronominal anaphora resolution », *Computational Linguistics*, 20(4), p. 535-561, 1994.

[PUS 95] PUSTEJOVSKY J., *The generative lexicon*, Cambridge, MIT Press, 1995.

[SAB 89] SABAH G., *L'intelligence artificielle et le langage*, vol. 2, Hermès, 1989.

[SAL 01] Salmon-Alt S. « Du corpus à la théorie : l'annotation (co-)référentielle », *Traitement Automatique des Langues (T.A.L.)*, 42(2), Hermès, Paris, 2002.

[SCH 93] SCHNEDEKER C., CHAROLLES M., « Les référents évolutifs: points de vue ontologique et phénoménologique », *Cahiers de Linguistique Française*, 14, 97-227, 1993.

[SCH 94] SCHNEDECKER C., CHAROLES M., KLEIBER G., DAVID J. (éds.), *L'anaphore associative. Aspects linguistiques, psycholinguistiques et automatiques*, Paris, Klincksiesk, 1994.

[SCH 97] SCHNEDECKER C., *Nom propre et chaînes de référence*, Paris, Klincksieck, 1997.

[SID 83] SIDNER C., « Focusing in the comprehension of definite anaphora », in B. Grosz *et al.* (eds), *Readings in natural language processing*, Morgan Kaufmann, p. 362-394, 1983.

[SUR 94] SURI L., MCCOY K., « RAFT/RAPR and centering: a comparison and discussion of problems related to processing complex sentences », *Computational Linguistics*, 20(2), p. 301-317, 1994.

[TAB 96] TABUTEAU G., Modélisation du thème dans le dialogue oral homme-machine, thèse de l'Université de Rennes, 1996.

[VER 01] VERGNE J., « Analyse syntaxique automatique de langues : du combinatoire au caculatoire », *Actes de TALN 2001*, p. 15-29, 2001.

[VIC 96] VICTORRI B.,FUCHS F., *La polysémie, construction dynamique du sens*, Paris, Hermès, 1996.

[VIC 99] VICTORRI B., « Le sens grammatical », *Langages*, 136, p. 85-105, 1999.

[VIL 95] VILAIN M., BURGER J., ABERDEEN J., CONNOLY D., HIRSCHMAN L., « A model theoretic coreference scoring scheme », *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann, 1995.

[WAL 94] WALKER M., IIDA M., COTES S., « Japonese discourse and the process of centering », *Computational Linguistics*, 20(2), p. 193-232, 1994.

[WAL 97] WALKER M., MOORE J., « Empirical studies in discourse », *Computational Linguistics*, 23(1), p. 1-12, 1997.

[WAL 98] WALKER M., JOSHI A., PRINCE E., *Centering theory in discourse*, Oxford, Clarendon Press, 1998.

[WAL 00] WALKER M., « Vers un modèle de l'intégration du centrage avec la structure globale du discours », *Verbum*, 22(1), p. 31-58, 2000.

[WIN 72] WINOGRAD T., *Understanding Natural Language*, Academic Press, 1972.