Revue, Informatique et Statistique dans les Sciences humaines
XXII, 1-4, 1986, pp. 191-211.

Analyse automatique de textes en langage naturel

G.P. ZARRI

1. Introduction

Le travail que je présente ici[1] se situe dans le contexte des recherches du type "Intelligence Artificielle" (I.A.) qui visent à effectuer l'extraction et le codage par ordinateur du "contenu sémantique d'un message en langage naturel (Silva *et al.* 1979, Schank *et al.* 1980, Hirschmann et Sager 1981, etc.); le but de ces recherches est d'arriver à constituer, *de façon automatique ou semi-automatique*, des "bases de faits" où les informations seront représentées dans un format "riche" permettant un traitement de type avancé (recherche de l'information implicite, mise en relation par inférence de faits *a priori* disjoints, etc.). Dans notre cas, la représentation-cible est celle du système RESEDA, système "intelligent" d'informatique documentaire biographique, mais les techniques utilisées pour la transposition automatique dans ce type de "métalangage" ont, en fait, une portée suffisamment générale.

Je rappellerai que le système RESEDA (Zarri 1979, 1981, 1983) se présente comme un "système expert" en informatique documentaire s'appliquant à la représentation et à l'exploitation par des techniques I.A. de données biographiques complexes. Le

1. Cette recherche est financée conjointement par l'Agence de l'Informatique (A.D.I.) – Convention CNRS/ADI n° 507568 – et par le C.N.R.S. (Action Thématique Programmée n° 955045).

terme "données biographiques" doit être compris ici dans son sens le plus large, soit toute information, qu'elle relève de la vie privée ou publique, des caractéristiques physiques ou intellectuelles, etc., d'un personnage donné.

Comme tout système expert, RESEDA est formé de deux composantes principales, une "base de connaissances" et un "moteur d'inférences". La base de connaissances comprend une "base de faits", où sont stockées les données propres au domaine considéré, et une "base de règles" qui contient les schémas de raisonnement qui s'appliquent à ces données; le moteur d'inférence se charge d'interpréter les informations contenues dans la base de règles pour en déduire des procédures de recherche à mettre en oeuvre dans la base de faits.

La base de faits a une importance particulière dans RESEDA. Cette base est en effet *permanente* et d'une *extension considérable*, contrairement à ce qui se passe dans la plupart des systèmes experts où les informations à interpréter sont introduites au moment même de leur traitement. Les connaissances qu'elle contient peuvent être réparties en deux classes. La première concerne les informations biographiques proprement dites; pour les décrire, a été mis au point un langage de représentation des connaissances ("métalangage") particulièrement souple et riche, fondé sur une approche du type "grammaire de cas", dans l'acception "Intelligence Artificielle" de ce terme. Cette forme de codage permet entre autre de représenter les informations fausses ou douteuses par rapport à un certain consensus, les propositions intentionnelles, conditionnelles, etc. La deuxième classe contient les connaissances générales du système à propos du domaine spécifique qu'il est censé traiter; ces données définissent les termes du "lexique" propre au métalangage utilisé.

La base de règles contient deux types de règles d'inférence, "transformation" et "hypothèses".

Les premières opèrent "par extension", leur but étant de récupérer et d'utiliser toute connaissance existante sous forme latente dans la base de faits. Par exemple, on pourra présumer avec un certain degré de fiabilité que quelqu'un a suivi des cours universitaires, même si cette information n'est pas explicitement attestée dans la base, en retrouvant l'information que cette personne est titulaire d'un diplôme correspondant au cours qu'elle est censée avoir suivi; normalement, la deuxième

information implique en effet la première. Il s'agit donc de "transformer" le "modèle de recherche" d'origine (suivre des cours) en lui substituant automatiquement un nouveau modèle (avoir un diplôme) de telle façon que les informations récupérées dans la base de faits par ce dernier "impliquent" celles que l'on a ou l'on aurait retrouvées avec le modèle d'origine.

Les hypothèses opèrent "par spécialisation", car leur fonction est de construire *ex nihilo* des connaissances nouvelles dans le domaine spécifique des relations de causalité. Plus exactement, les modèles de recherche générés automatiquement par les hypothèses essaieront d'apparier, à l'intérieur de la base de faits, un certain ensemble d'"événements", qui seront proposés à l'utilisateur en tant qu'explication *plausible* d'un autre événement connu; entre les deux groupes, on pourra ainsi instaurer une relation de type causal. On proposera donc d'expliquer, par exemple, certains changements de personnes au sommet de l'administration de l'Etat en fonction d'un changement de pouvoir politique.

Le "moteur d'inférence" est constitué par un ensemble de routines qui s'appliquent à trouver, dans la base de faits, les éventuelles réalisations des schémas explicatifs, dans le cas des hypothèses, ou des schémas d'équivalence sémantique, dans le cas des transformations, contenus dans la base de règles. D'un point de vue algorithmique, cette recherche de solutions se traduit par l'exploration d'un "arbre de choix", selon des modalités du type "en profondeur d'abord" mais comportant "un retour en arrière" (*backtracking*) systématique. La nécessité d'effectuer le *backtracking* est principalement lié au fait que les schémas d'inférence à vérifier sont généralement complexes, constitués de plusieurs composantes interdépendantes; on expliquera par exemple une nomination en fonction du fait qu'un parti politique s'est emparé du pouvoir *et* que le nouveau nommé a des relations étroites avec les chefs de ce parti. Les valeurs affectées aux "variables" dans l'une des composantes doivent être réutilisées dans les autres, d'où la nécessité de revenir en arrière pour changer d'affectation, en cas d'échec, dans la suite du déroulement de l'inférence : le chef de parti retenu dans un premier moment n'aura peut-être aucun rapport avec le nouveau fonctionnaire nommé. "Retour en arrière systématique" signifie que le moteur ne s'arrête pas à la première solution positive trouvée mais qu'il explore l'arbre des choix de façon exhaustive, de façon à avoir un panorama complet des résultats de la procédure d'inférence utilisée.

Dans la suite de l'article, j'illustrerai la stratégie que nous avons mise au point pour la constitution automatique d'une base de faits, en utilisant des exemples en langage naturel qui se réfèrent au contexte d'une application RESEDA bien connue, dédiée au traitement des biographies de certains personnages de l'Histoire de France appartenant à la période 1350-1450.

2. Le métalangage RESEDA

2.1. *Catégories du métalangage*

Les informations biographiques de la base de données sont organisées sous forme d'unités appelées "épisodes codés" ou "plans". Il y a différents types de plans, le "plan prédicatif" étant le plus important; chacun correspond à un "flash" qui illustre un moment particulier de la vie d'un ou plusieurs personnages. Un plan prédicatif est formé autour de cinq "prédicats" reconnus par le système : DEPLACER, ETRE-AFFECTE, ETRE-PRESENT, PRODUIRE, SE-COMPORTER; au prédicat, peuvent être reliés un ou plusieurs "modulateurs". Les modulateurs ont la tâche de préciser et de délimiter le rôle sémantique du prédicat. Dans un plan, le prédicat est accompagné de l'indication des "cas" qui introduisent ses arguments; les plans prédicatifs comportent aussi l'indication des coordonnées spatio-temporelles et de la source bibliographique de l'information. Les plans prédicatifs peuvent être reliés entre eux de plusieurs façons, entre autres, en utilisant des liens explicites (plans parenthétiques) de "coordination", "alternative", "causalité", "finalité", etc. Le mode de représentation des données utilisé dans le projet RESEDA est donc essentiellement du type "grammaire de cas", dans le sens où on l'entend en I.A. : voir Bruce (1975), Rosner et Somers (1980), Charniak (1981).

L'information "André Marchant est nommé prévôt de Paris par le Conseil du roi le 22 septembre 1413; il perd cet office le 23 octobre 1414, au profit de Tanguy du Châtel qui a impétré l'office" sera représentée en trois épisodes : la nomination d'André Marchant, sa destitution et la nomination de Tanguy du Châtel. Le codage de l'information doit s'effectuer selon deux niveaux différents : un codage "externe", réalisé jusqu'à maintenant manuellement par l'analyste, fait l'objet d'un premier

type de représentation formalisée dans les termes des catégories du métalangage de RESEDA. Une seconde étape, automatique, aboutit à un codage numérique "interne". Par exemple, le codage externe "manuel" des trois épisodes ci-dessus donnera lieu aux trois plans prédicatifs de la figure 1.

```
1)  incep+soc+ETRE-AFFECTE   SUJ      André-Marchant
                             OBJ      prévôt : Paris
                             SOURCE   conseil-du-roi
                             date1 : 22-septembre-1413
                             date2 :
                             bibl. : Demurger1,273

2)  fin+ETRE-AFFECTE         SUJ      André-Marchant
                             OBJ      prévôt : Paris
                             date1 : 22-octobre-1414
                             date2 :
                             bibl. : Demurger1,273

3)  incep+ETRE-AFFECTE       SUJ      Tanguy-du-Châtel
                             OBJ      prévôt : Paris
                             MODAL    lettres-de-don
                             date1 : 23-octobre-1414
                             date2 :
                             bibl. : Demurger1,273
```

fig. 1

Les codes en lettres majuscules indiquent le prédicat et les cas qui lui sont associés. A chaque plan prédicatif correspond toujours un couple de repères temporels "date1-date2" qui donne la durée de l'épisode en question. Dans les trois plans précédents, "date2" est vide, parce que les modulateurs associés aux prédicats (incep, fin) impliquent un changement d'état pris d'un point de vue ponctuel (début d'un état, fin d'un état). "André-Marchant" et "Tanguy-du-Châtel" sont des "vedettes", c'est-à-dire des personnages historiques connus par le système : "prévôt", "conseil-du-roi" et "lettres-de-don" font partie du "lexique" de RESEDA. Les classifications associées aux termes du lexique fournissent l'essentiel des connaissances du système à propos du contexte socio-historique de la période. "Paris" est le "lieu de l'objet". Si les sources historiques dépouillées nous donnaient, par exemple, les causes précises de la nomination d'André Marchant, on introduirait dans la base les plans correspondants et on leur associerait, à l'intérieur d'un plan parenthétique, le plan 1 par un lien explicite du type "CAUSE".

En plus des plans prédicatifs et des plans parenthétiques de liaison, le métalangage RESEDA utilise un troisième type, dit plan de relation. Ces plans servent à noter les relations familiales ou sociales qui existent entre les personnages. Ainsi, l'information "Guy d'Aigreville laisse un fils, Philippe d'Aigreville" sera représentée dans le plan de relation de la figure 2.

fig. 2

La principale caractéristique de ce type de plans est de ne pas avoir de prédicat explicitement exprimé, voir Zarri *et al.* (1977). Les "acteurs" de la relation sont inscrits dans une liste COORD (pour "COORDination") et le type de relation qui existe entre eux est précisé dans une liste SPECIF (pour "SPECIFication").

2.2. *Inconvénients du codage manuel*

Cette procédure manuelle de transformation de l'information en langue naturelle en un ou plusieurs plans présente au moins deux inconvénients majeurs auxquels l'étude envisagée se propose d'apporter des solutions :

- La représentation des informations biographiques dans les termes du métalangage ne peut être effectuée que par un personnel très spécialisé. Il est actuellement accompli par les chercheurs mêmes qui construisent la base de connaissances. Conserver une telle procédure est évidemment hors de question dans la perspective d'une utilisation quotidienne du système par une clientèle non-initiée, d'autant plus que RESEDA a été conçu comme un système alimenté *en permanence* d'informations biographiques extraites des sources les plus diverses.
- Bien que la syntaxe du métalangage de RESEDA impose de nombreuses contraintes sur la formation des schémas prédicatifs acceptés par le système et que ceux-ci soient en principe complètement recensés, on ne peut exclure la possibilité que deux codeurs traduisent la même information de deux façons différentes.

3. Cadre théorique de l'étude

Dans le domaine du traitement du langage naturel, les recherches en Intelligence Artificielle ont, dès l'abord, mis l'accent sur la prédominance des composantes sémantiques et pragmatiques. Dans ce cadre, l'établissement d'une représentation formelle du message véhiculé par l'expression de surface s'effectue généralement selon deux types de démarche.

3.1. *Approche "intégrée"*

L'une, la plus traditionnelle, respecte la progression habituelle des trois niveaux d'analyse, morphologique, syntaxique et sémantique, tout en intégrant leurs résultats dans l'interprétation finale. Pour discussion, voir par exemple Winograd (1972), Woods (1970, 1973), Marcus (1979), etc.

Dans cette optique, une technique a suscité beaucoup d'intérêt, celle des *"Augmented Transition Networks"* (ATN) introduite par Woods; une caractéristique des analyseurs de type ATN est de se fonder surtout sur une approche syntaxique, tout en testant à l'occasion certaines propriétés sémantiques simples des unités de l'expression de surface. L'analyseur de Winograd, s'il participe de cette même philosophie, permet de plus aux règles de grammaire de faire appel à des fonctions d'interprétation de type sémantique. Chaque fois qu'une entité syntaxique – par exemple un groupe nominal – est identifiée, elle est immédiatement soumise à l'interpréteur sémantique; en cas d'échec, le groupe syntaxique est invalidé et le texte ré-analysé. Il n'y a jamais passage réel d'information entre les deux composantes, sémantique et syntaxique, en dehors de cette évaluation du type "vrai ou faux" du résultat de l'analyse syntaxique. Une généralisation de cette approche est proposée par Marcus; dans son analyseur, un véritable échange de communications existe entre la composante syntaxique et les modules sémantique et pragmatique. Toutefois, ces derniers ne sont appelés qu'à la demande de l'analyseur syntaxique qui garde encore l'initiative.

3.2. *Approche "sémantique"*

Schank a, au contraire, mis en avant l'idée, reprise ensuite par de nombreux chercheurs, qu'une analyse essentiellement sémantique refusant à la syntaxe son rôle directeur traditionnel était

possible. On utilise alors la "représentation profonde" même que l'on est en train de construire pour faire des "prévisions" appropriées sur la fonction logico-sémantique des éléments qui seront rencontrés au fur et à mesure de l'exploration de l'expression de surface, voir Schank (1975), Wilks (1975), etc.

Cette idée a été particulièrement développée par "l'école de Yale": le premier analyseur opérationnel à concrétiser ces vues a été, en effet, le système bien connu de Riesbeck (1975) qui "comprend" une phrase en la transformant en sa représentation sémantique profonde selon la théorie des "dépendances conceptuelles" développée par Schank. DeJong (1979a, 1979b) définit ainsi son système: *"The proposed system organization has two modules: a PREDICTOR and a SUBSTANTIATOR. One module predicts constraints on what might happen next. The other module tries to justify and give substance to these predicted characterizations on the next possible events. The communication between the two modules is comparatively unconstrained"* (DeJong 1979b: 255-256). A remarquer que la syntaxe a encore ici sa place: elle sert à l'analyseur pour repérer dans le texte de surface les actualisations des prévisions faites au niveau profond (DeJong 1979b: 259). Rappelons enfin les travaux de Wilensky (1980, 1982), qui relèvent de la même école de pensée.

Une attitude plus radicale encore est exprimée dans les travaux récents de Rieger et Small (Rieger et Small 1979, Small 1980), où l'analyse est réalisée par une série d'interactions entre les "experts" associés à chaque mot de l'expression de surface qui *coopèrent* à la compréhension globale du texte par l'échange de deux types d'information, les "structures conceptuelles" et les "signaux de contrôle".

3.3. *Approche RESEDA*

L'hypothèse que nous avons adoptée pour ce projet relève plutôt de la deuxième option, dans la mesure où les structures de la représentation interne propre à RESEDA fournissent, par avance, un cadre très complet des prévisions qui doivent guider la lecture et la compréhension des textes à traduire dans le métalangage du système.

Pour présenter notre méthode de traduction, nous l'appliquerons sur l'exemple donné en 2.1. Le texte de départ en langage

naturel est d'abord (pré-)traité pour en dégager les constituants grammaticaux; pour ce faire nous utilisons, à titre expérimental, la Grammaire française de surface du logiciel DEREDEC, mis au point par Pierre Plante à l'Université du Québec à Montréal (Plante 1980a, 1980b). Ce système, qui se rapproche des analyseurs ATN, permet de dégager d'un texte écrit en français naturel ses constituants syntaxiques sous forme arborescente; il permet également d'établir entre ces constituants les liens syntagmatiques de type "thème-propos", "déterminant-déterminé", "coordonnés", qui les unissent. Cette analyse préliminaire fournit une base pour les traitements (informatiques) ultérieurs, sans que toutes les ambiguïtés soient nécessairement levées pour autant.

Les instruments que nous entendons mettre au point dans cette étude sont de deux types : une procédure générale, que l'on peut assimiler à une sorte de *"parsing"* sémantique, et un ensemble de règles heuristiques.

4. Description d'ensemble de la solution envisagée

4.1. *"Parsing" sémantique*

La procédure générale comprend plusieurs étapes.

4.1.1.

Une première étape consiste à repérer les "déclencheurs", définis comme des unités lexicales qui appellent un ou plusieurs des schémas prédicatifs prévus par le métalangage de RESEDA. Nous ne prenons donc pas en considération chacun des items lexicaux rencontrés dans le texte de surface mais nous retenons seulement ceux qui sont directement pertinents pour la "traduction" à effectuer.

Nous ne nous limitons pas, toutefois, à une simple approche par mot-clé. Dans un contexte donné, certains items lexicaux peuvent être considérés comme d'éventuels déclencheurs, mais ils ne déclencheront un schéma prédicatif que s'ils répondent à certaines

contraintes dictées tant par l'analyse morpho-syntaxique fournie par DEREDEC que par les informations socio-historiques enregistrées dans le système RESEDA.

Le DEREDEC fournit, entre autre, une sorte de lemmatisation qui permet de ramener les formes de surface d'un texte à une forme canonique, par exemple l'infinitif pour les verbes. Par la suite, ces formes canoniques sont comparées avec une liste de déclencheurs potentiels, stockée de façon permanente dans le système. Pour notre exemple, cette procédure permet de construire la liste de déclencheurs suivante : formes verbales : *"nommer"*, *"perdre"*, *"impétrer"*; termes appartenant au métalangage, ou termes qui correspondent directement à un élément lexical du métalangage : *"office"*, synonyme de "emploi" dans RESEDA ("emploi" étant un générique, une tête de sous-arborescence dans le lexique de RESEDA), et son spécifique *"prévôt"*. Les résultats de la pré-analyse effectuée par DEREDEC permet d'éliminer parmi les schémas potentiels associés au déclencheur *"nommer"* celui qui correspond à une construction "active" en surface, comme par exemple dans : "Le duc d'Orléans a nommé André Marchant prévôt de Paris ...". Les schémas retenus sont, par conséquent, ceux de la figure 3.

On remarquera que, pour le déclencheur *"nommer (forme active)"*, la vedette assumant le rôle d'"objet de surface" viendrait remplir la position associée au cas SUJet du prédicat ETRE-AFFECTE, tandis que le sujet de surface se trouverait associé au cas SOURCE.

4.1.2.

En réalité, les structures prédicatives retenues ne se limitent pas à celles de la figure 3. Elles sont en effet doublées d'autres schémas prédicatifs du type ETRE- AFFECTE, mais dont la position correspondant au cas SUJet doit être remplie par un "représentant de vedette" (un organisme, l'Université de Paris par exemple, qui comporte comme un individu, une "vedette") et celle du cas OBJet par une "vedette" suivie d'une spécification ("SPECIF") de "emploi". Ces dernières correspondent à une information du type: "Une vedette reçoit un poste dans un organisme (un organisme, SUJet, est "augmenté", ETRE-AFFECTE, d'un nouveau membre, OBJet, en rapport, SPECIF, avec un certain emploi)". L'équivalent en langage naturel serait par exemple : "André Marchant (vedette) est nommé secrétaire

```
nommer = incep+(soc+)ETRE-AFFECTE  SUJ   <vedette>-sujet de surface
(forme passive)                                  du déclencheur
                                    OBJ   <emploi> -complément de
                                                   surface
                                 (SOURCE  <vedette>|<représentant-
                                                   de-vedette>-
                                           complément d'agent de
                                           surface du déclencheur)
                                    date1 : obligatoire
                                    date2 : interdit
                                    bibl. : obligatoire

prévôt = (soc+)ETRE-AFFECTE  SUJ   <vedette>
                             OBJ   <emploi>-déclencheur
                          (SOURCE  <vedette>|<représentant-de-vedette>)
                             date1 : obligatoire
                             date2 : facultatif
                             bibl. : obligatoire

perdre = fin+ETRE-AFFECTE  SUJ <vedette>-sujet de surface du déclencheur
(forme active)             OBJ <terme-du-lexique>-objet de surface du
                                  déclencheur
                           date1 : obligatoire
                           date2 : interdit
                           bibl. : obligatoire

office = (soc+)ETRE-AFFECTE  SUJ  <vedette>
                             OBJ  <emploi>
                           (SOURCE <vedette>|<représentant-de-vedette>)
                             date1 : obligatoire
                             date2 : facultatif
                             bibl. : obligatoire

impétrer = incep+(soc+)ETRE-AFFECTE  SUJ <vedette>-sujet de surface du
(forme active)                             déclencheur
                                    OBJ <emploi>-complément de surface
                                 (SOURCE  <vedette>|<représentant-de-
                                       vedette>-complément d'agent de
                                       surface du déclencheur)
                                    MODAL   lettres-de-don
                                    date1 : obligatoire
                                    date2 : interdit
                                    bibl. : obligatoire

office = (soc+)ETRE-AFFECTE  SUJ  <vedette>
                             OBJ  <emploi>
                          (SOURCE <vedette>|<représentant-de-vedette>)
                             date1 : obligatoire
                             date2 : facultatif
                             bibl. : obligatoire
```

fig. 3

(emploi) de la papauté (représentant de vedette)". Ainsi, par exemple, on doit associer le schéma prédicatif de la figure 4 au déclencheur *"nommer (forme passive)"*. A la fin de la procédure

de construction, on éliminera les schémas prédicatifs de la seconde série puisqu'il n'est pas possible de trouver en surface une actualisation du concept "représentant une vedette" pour remplir la position SUJet et que, de ce fait, les structures prédicatives obtenues sont incomplètes.

```
nommer = incep+(soc+)ETRE-AFFECTE  SUJ  <représentant-de-vedette>
(forme passive)                    OBJ  <vedette>-sujet de surface
                                          du déclencheur  SPECIF <emploi>
                                          -complément de surface
                                (SOURCE   <vedette>|<représentant-de-
                                          vedette>)
                                     date1 : obligatoire
                                     date2 : interdit
                                     bibl. : obligatoire
```

fig. 4

4.1.3.

La dernière étape de la procédure générale consiste à examiner les déclencheurs appartenant aux mêmes environnements morpho-syntaxiques, définis par l'analyse DEREDEC.

S'il y a plusieurs déclencheurs dans le même environnement et s'il y a une possibilité de coïncidence des divers schémas prédicatifs déclenchés – même prédicat, même jeu de cas, modulateurs et thématique compatibles – on peut considérer que ces déclencheurs se rapportent à une même situation. En conséquence, les schémas prédicatifs sont combinés de manière à obtenir la description la plus complète possible, et les "prévisions" portant sur le remplissage des positions associées aux cas des différents schémas prédicatifs dirigent ensemble la recherche dans l'expression de surface.

```
incep+(soc+)ETRE-AFFECTE  SUJ <vedette>-sujet de surface de "est nommé"
                          OBJ <emploi>-"prévôt"
                     (SOURCE  <vedette>|<représentant-de-vedette>
                              -complément d'agent de "est nommé")
                          date1 : obligatoire
                          date2 : interdit
                          bibl. : obligatoire
```

fig. 5

Ainsi les deux premiers déclencheurs de notre exemple, reconnus comme relevant d'un même environnement, se combinent dans la formule de la figure 5, qui donne le cadre général du plan 1 de la figure 1.

4.1.4

L'exemple que nous venons de proposer illustre une situation particulièrement simple, où il n'a pas été nécessaire d'établir des liens entre les schémas déclenchés. A l'opposé, si on avait pris la phrase "Philibert de St Léger est nommé sénéchal de Lyon le 30 juillet 1412, à la place de A. de Viry, décédé", trois plans auraient été déclenchés : le premier pour couvrir la nomination de Philibert de St Léger, le second pour la mort de A. de Viry, et le troisième pour établir un lien de causalité faible (CONFER, dans le métalangage de RESEDA) entre les deux premiers plans. Les éléments du langage naturel tel que les conjonctions, les prépositions et les adverbes modificateurs de phrase peuvent être utilisés pour déclencher des liens de causalité, de finalité, de coordination, etc., entre plans. Plus concrètement, dans le dernier exemple, "à la place de" est identifié comme déclencheur potentiel en vertu de la règle suivante : si le groupe nominal principal du syntagme prépositionnel de surface comporte un déclencheur, l'ensemble du syntagme constitue l'environnement d'un plan qui sera lié à un autre par CONFER.

4.2. *Règles heuristiques*

La procédure que nous venons d'exposer exige un corpus de règles heuristiques – exprimées en termes de contraintes associées aux schémas prédicatifs du métalangage RESEDA. Ces règles permettront de remplir les positions correspondant aux cas des schémas déclenchés par l'utilisation des informations de surface qui concordent avec les "prédictions" associées à ces cas. Pour le schéma de la figure 5, le remplissage ne pose pas de problème, puisque les éléments de surface "André Marchant", "prévôt", "conseil du roi" et "22 septembre 1413" – normalisés conformément aux règles du métalangage RESEDA, voir figure 1 – vont remplir les positions "SUJet", "OBJet", "SOURCE" et la position "date1" directement. Mais les procédures de remplissage sont souvent beaucoup plus compliquées et peuvent nécessiter le recours à des règles d'inférences complexes, en particulier lorsque le

texte de surface contient des expressions anaphoriques, qui doivent être résolues avant de pouvoir remplir les positions des schémas déclenchés.

Un certain nombre de "théories" pour la résolution automatique des expressions anaphoriques ont été mises au point dans le cadre des recherches en intelligence artificielle; on rappellera à ce propos les travaux de Grosz (1977), Webber (1978) et Sidner (1979). Etant données les finalités éminemment pratiques du projet que nous présentons ici, il est évident que, en affrontant le problème des anaphores, il n'était pas question pour nous de développer une "théorie générale"; les considérations qui suivent doivent être rapprochées plutôt de certains travaux heuristiques développés récemment dans un contexte I.A. pour la résolution des références pronominales à l'intérieur d'univers de discours très limités, voir Hayes (1981), Walker et Hobbs (1981), par exemple.

A l'aide du fragment de biographie qui suit, nous allons illustrer le type de règles que nous utilisons pour résoudre les cas d'anaphores.

> "*Guy d'Aigreville.*
> (...)
> *Il* est mort avant le 12 juillet 1414, date à laquelle *sa veuve*, Isabelle d'Aigreville, rapporte au Parlement les sceaux du bailliage de Sens. *Il* laisse un fils, Philippe d'Aigreville.
>
> *Sa veuve* doit répondre, devant le Parlement, à plusieurs accusations portées contre *son mari*."

Dans cet exemple, les anaphores sont marquées en surface par des pronoms personnels et par des déterminants possessifs; leur résolution relève de deux mécanismes légèrement différents. Les unes peuvent être résolues sur des bases purement morpho-syntaxiques (*il*); les autres nécessitent, en plus, le recours à des règles d'inférence sémantique (*sa veuve, son mari*). Dans le cas, il s'agit de retrouver le personnage représenté par l'expression anaphorique, pour pouvoir ensuite remplir les positions des schémas déclenchés. Ceci est une contrainte très stricte dans la construction de la base de faits RESEDA. Ainsi, pour une biographie donnée, si aucun nom propre ne peut être identifié comme co-référent d'une expression anaphorique donnée (comme cela aurait pu se vérifier pour *sa veuve* dans le fragment ci-dessus), un numéro d'identification temporaire doit être généré pour remplacer le nom propre manquant.

4.2.1. *Résolution des pronoms de la troisième personne du singulier*

Il convient de souligner au préalable que les biographies que nous pensons traiter de façon presque entièrement automatique doivent satisfaire un certain nombre de contraintes, qui ne sont pas trop limitatives. Elle seront donc écrites à la troisième personne, dans un style relativement figé : le personnage qui fait l'objet de la biographie est donné en tête de texte; au début de chaque paragraphe, le pronom personnel sujet de la troisième personne du singulier (*il*) renvoie systématiquement à ce personnage. Cette règle d'écriture des textes de type biographique, réutilisée dans la procédure d'interprétation, permet de lever un certain nombre d'ambiguïtés possibles des pronoms personnels. En réalité, cette "règle du paragraphe" peut être étendue également aux déterminants possessifs de la troisième personne du singulier : voir dans le fragment ci-dessus *"il"* et *"sa veuve"*, tous deux en position sujet de surface en tête de paragraphe, où *"il"* et *"sa"* renvoient à l'en-tête de la biographie, *"Guy d'Aigreville"*. On peut donc poser la règle suivante :

> *Règle du paragraphe*
> *Tout pronom personnel et tout déterminant possessif de la troisième personne du singulier en position sujet en tête de paragraphe renvoie à l'en-tête de la biographie, à condition, bien sûr, que les contraintes de genre soient respectées.*

A l'intérieur du paragraphe, les pronoms personnels et les déterminants possessifs de la troisième personne du singulier en position sujet renvoient (si les contraintes de genre sont respectées) au focus courant, qui est défini comme étant le dernier personnage mentionné en position de sujet de surface dans une proposition principale (les propositions subordonnées, voir par exemple la relative "... date à laquelle *sa veuve*, Isabelle d'Aigreville, rapporte ...", n'ont donc pas d'influence dans l'établissement du focus). Ainsi, la seconde occurrence du pronom *"il"* dans notre exemple renvoie à *"Guy d'Aigreville"*, identifié comme co-référent de la première occurrence de *"il"*.

4.2.2. *Résolution des anaphores nominales définies*

Ce second type d'anaphores nécessite le recours à des règles d'inférence. Sidner (1979), entre autres, utilise des *frames* associés à des réseaux sémantiques pour effectuer ces inférences.

Nous utilisons, pour notre part, le système de représentation de connaissances de RESEDA pour effectuer les inférences nécessaires.

Ainsi, les occurrences de *"sa veuve"* et de *"son mari"*, dans l'exemple précédent, déclenchent une série de schémas, dont les positions doivent être remplies par un nom de personnage masculin (le mari décédé) et un nom de personnage féminin (sa femme), voir figure 6. Etant donné que la structure interne du métalangage RESEDA exige que les positions associées aux cas des schémas déclenchés soient remplies par des noms propres représentant des personnages, les référents de *"veuve"* et *"mari"* doivent être trouvés. Pour ce faire, on doit d'abord identifier le type du pré-déterminant (ici, le possessif *"sa"*, *"son"*) et reconnaître l'arborescence du lexique RESEDA correspondant au nom de l'expression anaphorique (soit l'arborescence des termes de parenté pour *"veuve"* et *"mari"*). Le possessif guidera la recherche vers l'un des personnages nécessaires au remplissage des plans déclenchés, et le nom vers l'autre personnage, par la mise en oeuvre conjuguée de règles morpho-syntaxiques et sémantiques.

```
veuve  =   const+ETRE-AFFECTE   SUJ   <vedette-fém>
                                OBJ   veuve-déclencheur

"Il y a un personnage féminin qui se trouve dans l'état de
'veuve'"

        (COORD  <vedette-fém>   (SPECID épouse)
                <vedette-masc>  (SPECID époux))

"Le personnage féminin est marié à un personnage masculin"

          fin+abs+ETRE-PRESENT   SUJ <vedette-masc>

"Le personnage masculin est mort"


mari  =  (COORD  <vedette-masc> (SPECID époux)
                 <vedette-fém>  (SPECID épouse))

"Il y a un personnage masculin qui est marié à un personnage
féminin"
```

fig. 6

Comme nous l'avons vu précédemment, tout déterminant possessif de la troisième personne du singulier en position sujet en tête de paragraphe renvoie au personnage qui fait l'objet de la biographie (autrement dit, à l'en-tête du texte). Dans toute autre situation, le déterminant possessif renvoie au dernier personnage mentionné dans le rôle de sujet de surface dans la proposition principale. D'où la règle du possessif :

Règle du possessif
Tout déterminant possessif de la troisième personne du singulier qui n'est pas en position sujet en tête de paragraphe co-réfère avec le dernier personnage mentionné dans le rôle de sujet de surface dans la proposition principale.

Pour la première occurrence de *"sa veuve"*, la "règle du possessif", appliquée à *"sa"* permet de remonter à *"il"*, qui a déjà été reconnu comme co-référant avec *"Guy d'Aigreville"* : c'est le personnage masculin requis pour remplir les positions adéquates des schémas déclenchés par *"veuve"*. Le nom du personnage féminin se trouve en surface, en apposition au terme déclencheur; il est récupéré par l'application d'un modèle de recherche sur l'analyse syntaxique fournie par DEREDEC. Les trois schémas déclenchés par *"veuve"* peuvent alors être remplis (voir figure 7), puis insérés dans la base RESEDA, où ils pourront être retrouvés facilement. Parallèlement, la résolution de l'expression anaphorique ("Isabelle d'Aigreville" pour *"sa veuve"*) sera attachée à l'expression de surface.

```
const+ETRE-AFFECTE   SUJ  Isabelle-d'Aigreville
                     OBJ  veuve
                     date1 : 12-juillet-1414

      (COORD  Isabelle-d'Aigreville (SPECIF épouse)
              Guy-d'Aigreville (SPECIF époux))

fin+abs+ETRE-PRESENT  SUJ  Guy-d'Aigreville
                      date1 : avant-le-12-juillet-1414
```

fig. 7

Pour la seconde occurrence de *"sa veuve"*, le possessif, suivant la "règle du paragraphe", est reconnu comme co-référent du personnage central de la biographie, *"Guy d'Aigreville"*. Les inférences nécessaires à la résolution de cette expression anaphorique sont désormais possibles grâce aux plans qui viennent

d'être stockés dans la base. Ainsi le nom de la veuve est récupéré dans les plans associés à la vedette *"Guy d'Aigreville"* de la façon suivante : on cherche dans l'index un plan représentant le mariage de cette vedette, voir le deuxième plan de la figure 7; le nom féminin contenu dans ce plan en est extrait. Les positions des schémas déclenchés par le reste de la phrase sont alors remplies. Enfin, la résolution de l'expression anaphorique ("Isabelle d'Aigreville") est attachée à l'expression de surface (*"sa veuve"*), comme précédemment.

Pour la dernière anaphore nominale *"son mari"*, le possessif *"son"* est identifié comme renvoyant à la dernière occurrence de *"sa veuve"*, suivant la règle du possessif. L'expression *"sa veuve"* ayant déjà été identifiée comme représentant "Isabelle d'Aigreville", le nom du mari sera retrouvé dans les plans associés à la vedette "Isabelle d'Aigreville", par une inférence similaire au cas précédent.

La procédure que nous venons d'exposer illustre assez bien la complexité du problème de la résolution des expressions anaphoriques. Encore faut-il noter que nous n'avons traité que des expressions anaphoriques marquées par des pronoms. La complexité du système de représentation de connaissances générales à mettre en place et celle des règles d'inférences à développer serait encore plus grande pour les expressions anaphoriques non-marquées, c'est-à-dire les syntagmes nominaux déterminés par un simple article, voir Sidner (1979, 1982). On remarquera cependant que l'organisation même du système RESEDA facilite la mise en oeuvre de ces procédures d'inférence et fournit déjà un système de représentation.

5. Conclusion

Le travail que nous venons de présenter a été développé dans le but de générer automatiquement une représentation sémantique du type RESEDA pour le "message" véhiculé par des expressions françaises de surface. La première étape de la procédure envisagée concerne le repérage des "déclencheurs", définis comme des unités lexicales qui appellent un ou plusieurs des schémas prédicatifs prévus dans le métalangage de RESEDA. Les schémas obtenus sont ensuite fusionnés, et les positions qui sont associées

aux cas des schémas en question sont remplies avec des éléments retrouvés en surface en utilisant les "prévisions" relatives aux différents cas. Ces procédures de remplissage peuvent nécessiter le recours à des règles d'inférence complexes, en particulier quand le texte de surface contient des expressions anaphoriques, qui doivent être résolues avant de pouvoir remplir les positions des schémas déclenchés.

Bibliographie

Bruce, B. (1975). "Case Systems for Natural Language", *Artificial Intelligence*, VI, 327-360.

Charniak, E. (1981). The Case-Slot Identity Theory", *Cognitive Science*, V, 285-292.

De Jong, G. (1979a). *Skimming Stories in Real Time* (Ph. D. thesis). New Haven, Yale University Computer Science Department.

De Jong, G. (1979b). "Prediction and Substantiation : A New Approach to Natural Language Processing", *Cognitive Science*, III, 251-273.

Grosz, Barbara (1977). "The Representation and Use of Focus in a System for Understanding Dialogs", in *Proceedings of the 5th International Joint Conference on Artificial Intelligence – IJCAI/5.* Los Altos, William Kaufmann, Inc.

Hayes, Philip J. (1981). "Anaphora for Limited Domain Systems", in *Proceedings of the 7th International Joint Conference on Artificial Intelligence – IJCAI/7.* Los Altos, William Kaufmann, Inc.

Hirschmann, Lynette et Sager, Naomi (1981). "Automatic Informatting of a Medical Sublanguage", in *Sublanguage: Studies of Language in Restricted Semantic Domains*, Kittredge, R. et Lehrberger, J. (eds), Berlin, de Gruyter.

Marcus, M. (1979). *A Theory of Syntactic Recognition for Natural Language*, Cambridge (Mass.), MIT Press.

Plante, P. (1980a). *Une grammaire DEREDEC des structures de surface du français, appliquée à l'analyse de contenu des textes*, Montréal, Université du Québec à Montréal.

Plante, P. (1980b). *DEREDEC – Logiciel pour le traitement linguistique et l'analyse de contenu des textes, manuel de l'usager*, Montréal, Université du Québec à Montréal.

Rieger, C. et SMALL, S. (1979). "Word Expert Parsing", in *Proceedings of the 6th International Joint Conference on Artificial Intelligence – IJCAI/6*. Los Altos, William Kaufmann, Inc.

Riesbeck, C.K. (1975). "Conceptual Analysis", in *Conceptual Information Processing*, Schank, R.C., ed. (Fundamental Studies in Computer Science, 3), Amsterdam, North-Holland.

Rosner, M. et Somers, H.L. (1980). *Case in Linguistics and Cognitive Science* (Working paper n° 40). Genève, Institut pour les études sémantiques et cognitives.

Schank, R.C., ed. (1975). *Conceptual Information Processing* (Fundamental Studies in Computer Science, 3). Amsterdam, North-Holland.

Schank, R.C., Lebowitz, M. et Birnbaum, L. (1980). "An Integrated Understander", *American Journal of Computational Linguistics*, VI, 13-30.

Sidner, Candace L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse* (Ph. D. thesis), Cambridge (Mass.), MIT Artificial Intelligence Laboratory.

Sidner, Candace L. (1982). *The Pragmatics of Non-Anaphoric Noun Phrases* (BBN Report). Cambridge (Mass.), Bolt, Beranek and Newman.

Silva, G.M.T., Dwiggins, D.L., Busby, S.G. et Kuhns, J.L. (1979). *A Knowledge-Based Automated Message Understanding Methodology for an Advanced Indications System* (OSI Report R79-006). Woodland Hills, Operating Systems, Inc.

Small, S. (1980). *Word Expert Parsing : A Theory of Distributed Word-Based Natural Language Understanding* (TR n° 954). College Park, University of Maryland Department of Computer Science.

Walker, D.E. et Hobbs, J.R. (1981). *Natural Language Access to Medical Text* (SRI Technical Note n° 240). Menlo Park, SRI International.

Webber, Bonnie L. (1978). *A Formal Approach to Discourse Anaphora* (BBN Report n° 3761). Cambridge (Mass.), Bolt, Beranek and Newman.

Wilensky, R. et Arens, Y. (1980). *PHRAN : A Knowledge Approach to Natural Language Analysis* (Memo VCB/ERL M80/4). Berkeley, Electronics Research Laboratory.

Wilensky, R. (1982). *Planning and Understanding*. Reading (Mass.), Addison-Wesley.

Wilks, Y. (1975). "A Preferential, Pattern-Seeking Semantics for Natural Language Inference", *Artificial Intelligence*, VI, 53-74.

Winograd, T. (1972). *Understanding Natural Language*. New York, Academic Press.

Woods, W.A. (1970). "Transition Network Grammars for Natural Language Analysis", *Communications of the ACM*, XIII, 591-606.

Woods, W.A. (1973). "An Experimental Parsing System for Transition Network Grammars", in *Natural Language Processing*, Rustin, R. (ed.), New York, Algorithmics Press.

Zarri, G.P. (1979). "What Can Artificial Intelligence Offer to Computational Linguistics ? The Experience of the RESEDA Project", in *Advances in Computer-Aided Literary and Linguistic Research*, D.E. Ager *et al.* (eds), Birmingham, The University of Aston in Birmingham.

Zarri, G.P. (1981). "Building the Inference Component of an Historical Information Retrieval System", in *Proceedings of the 7th International Joint Conference on Artificial Intelligence – IJCAI/7*. Los Altos, William Kaufmann, Inc.

Zarri, G.P. (1983). "An Outline of the Representation and Use of Temporal Data in the RESEDA System", *Information Technology : Research and Development*, II, 89-108.

Zarri, G.P., Ornato, Monique, King, Margaret, Zwiebel, Anne et Zarri-Baldi, Lucia (1977). *Projet RESEDA/0 : Rapport Final*. Paris, Equipe Recherche Humanisme Français.

Zarri, G.P., Ornato, Monique, Lee, G., Meissonnier, V., Cridford, Lesley, Nedemec, Y. et Zarri-Baldi, Lucia (1979). *Projet RESEDA/1/: Rapport sur les recherches effectuées du 1er octobre 1978 au 1er avril 1979* (Rapp. LISH/129). Paris, Laboratoire d'Informatique pour les Sciences de l'Homme.