Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes

Delphine Battistelli, Jean-Luc Minel

▶ **To cite this version:**

Delphine Battistelli, Jean-Luc Minel. Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes. Gérard Sabah. Compréhension des langues et interaction, Hermes Science Publications, 2024, Cognition et Traitement de l'Information, 2746212560. halshs-00096816

HAL Id: halshs-00096816
https://shs.hal.science/halshs-00096816

Submitted on 26 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes

4.1. Comprendre un texte : avec ou sans représentations ?

S'agissant de textes, le terme de « compréhension » peut leur être appliqué dans des cadres très divers, ce qui reflète sans doute la manière dont nous envisageons la compréhension, ou encore ce que nous nous figurons être la « démonstration » qu'il y a eu compréhension : en pouvant résumer un texte, en tirer des conclusions, le traduire dans une autre langue, le paraphraser, ou encore en donner une illustration figurative. À ces différentes activités correspondent probablement différentes stratégies cognitives et si l'on se place dans l'éventualité selon laquelle l'esprit humain se représenterait l'information sous une forme unique, il faut alors supposer que les représentations sont d'un degré élevé d'abstraction, les rendant aptes à exprimer des notions et des opérations impliquées dans des activités cognitives différentes. Une autre hypothèse (parfois qualifiée de « multi-modale ») est celle selon laquelle l'information serait représentée sous des formes différenciées, avec des propriétés différentes, des modes d'organisation différents, des conditions d'utilisation différentes ; on distingue alors en particulier les modes de représentation propositionnel et imagé (voir par exemple[DEN 93]).

Il reste que ces différentes activités dénotent toutes une certaine capacité à accéder au « sens », qui peut donc être testée de manières très diverses, les exercices scolaires faisant d'ailleurs souvent appel à ces différentes stratégies d'évaluation de la compréhension. De fait, la compréhension de textes reste un sujet fort complexe

---

Chapitre rédigé par Delphine BATTISTELLI et Jean-Luc MINEL.

sur lequel on sait relativement peu de choses, que ce soit quant à la nature des processus mis en œuvre (les mécanismes inférentiels en particulier) ou des représentations « internes » manipulées (aucun format précis pour ces représentations n'a jamais été finalement proposé) [COI & al. 96]. La diversité des travaux en cours dans ce domaine illustre la difficulté théorique à aborder le thème de la compréhension de manière unifiée.

Nombre d'individus déclarent, lorsqu'ils se trouvent en situation de lecture, former des représentations visuelles des personnages, des objets, des évènements décrits. Le langage ordinaire utilise d'ailleurs des termes liés à la perception, et tout particulièrement à la vision, pour parler de la compréhension linguistique, de l'accès au sens, par exemple dans des expressions comme « je *vois*/*ne vois pas ce que vous voulez dire* », ou encore « c'est *clair*, *lumineux fumeux*, *opaque*, *obscur* ». Souvent, « *montrer* » constitue un équivalent métaphorique de « *faire comprendre* ». Cette conception de la compréhension – appréhendée ici dans son corrélat perceptif – repose sur le concept de *représentation*[1], et, qu'elles soient envisagées ou non dans leur rapport à l'image, il s'avère que les représentations sont au centre des réflexions sur la compréhension, en Intelligence Artificielle (IA), en psychologie comme en linguistique. À ces trois disciplines correspondent en fait trois conceptions de la compréhension et trois manières d'aborder les représentations, selon que l'on privilégie les opérations sur ces représentations (conception IA), la description des processus intégratifs à l'œuvre dans la construction de ces représentations (conception psychologique) et l'explicitation de ces représentations construites à partir des formes langagières en prenant en compte des phénomènes tels que l'ambiguïté ou la polysémie (conception linguistique).

Rappelons quelques éléments liés à ces trois points de vue, déjà évoqués dans l'introduction.

En intelligence artificielle, l'étude du raisonnement, dit « naturel », occupe une place importante ; elle repose sur une description des formes de raisonnement qui s'expriment dans et par le langage. le jugement de l'accomplissement d'un tel objectif étant envisagé dans le cadre d'une confrontation entre les conclusions que le système est capable de tirer d'un texte donné et celles qu'un sujet humain est amené à tirer de la lecture de ce même texte [PIT 85, SAB 88]. Les capacités déductives – ou plus généralement inférentielles – des modèles formels se vérifient en confrontant les résultats de certaines opérations logiques exécutées à partir des représentations – formelles — des contenus sémantiques, avec les conclusions des raisonnements correspondants que conclurait un sujet humain informé par un même message linguistique.

1. Les sciences cognitives mettent même l'accent sur la nécessité de prendre en compte différents niveaux de représentation.

Les travaux de psychologie cognitive mettent en exergue le rôle de la construction de représentations. *cf.* [COI et al. 96]. La théorie des modèles mentaux de Johnson-Laird [JOH 83] par exemple a permis de renouveler l'approche de diverses questions, même si certains chercheurs, soulignant le caractère vague du terme de « modèle mental », ont préféré utiliser d'autres expressions, celui de « modèle de situation » [KIN 78] par exemple. « Quand les gens comprennent un discours, ils construisent ainsi un modèle schématique de la situation décrite dans le discours » [JOH 83]. Ainsi, depuis la fin des années soixante-dix, l'analyse des processus mis en œuvre dans le traitement sémantique des textes et la mise en mémoire des produits de ce traitement a connu un développement considérable [DEN 84, DIJ & KIN 83]. Dans les travaux de ce courant de recherche, l'analyse des unités de traitement et celle des processus opérant sur ces unités ont été menées à un niveau « propositionnel » : les unités constitutives de la « base de texte » sont des structures sémantiques prédicatives (des « propositions ») regroupées le plus souvent en structures de niveau supérieur (appelées « macrostructures »).

La problématique de la compréhension en linguistique peut aussi être posée dans les termes mêmes de la sémantique linguistique : les langues naturelles constituent une médiation entre un référentiel externe perçu et les représentations internes que nous construisons. L'approche de la compréhension adoptée en linguistique s'illustre ainsi au niveau de l'analyse sémantique (mais parfois aussi pragmatique) et pose alors quant à elle directement le problème du passage marqueurs linguistiques (dits « de surface ») et représentations (dites « sémantiques »)[2]. La plupart des théories linguistiques traitent de la sémantique de phrases isolées, ce n'est que dans ces trente dernières années qu'un courant de recherche important, celui de la « linguistique textuelle », s'est attaché à l'étude systématique des structures discursives.

De ces trois types d'approches de la compréhension en découle en réalité une quatrième qui est celle qui nous occupe, celle adoptée dans le domaine de la linguistique informatique. Visant à proposer des solutions opératoires pour des analyses de textes appartenant à tout domaine, elle a en fait émerger une autre conception de la compréhension : une conception qui fait reposer le processus de compréhension sur le principe du repérage et de l'extraction de certaines unités linguistiques sans recours nécessaire, *a priori*, à des représentations. On a souvent qualifié cette approche de la compréhension de textes d'approche « en surface » par opposition à l'approche qui postulait la nécessité de recourir à des représentations pour aborder la sémantique des textes, approche souvent qualifiée pour sa part d'approche « en profondeur ».

[2] Voir [FRA 98] pour une présentation des conceptualisations linguistiques sous-jacentes aux modes de référenciation du langage au monde.

Le résumé automatique fait partie de ces applications qui ont historiquement vu leur développement s'appuyer d'abord sur la notion de représentation (sémantique) puis sur celle d'étiquette (sémantique).

Aussi, c'est à travers cette application que nous nous proposons d'aborder le thème de la compréhension. La cohérence d'un résumé étant un des éléments essentiels de son intelligibilité, nous nous pencherons (cf. 4.3.2.) sur un des éléments fondamentaux qui participent à la cohérence d'un texte, celui qui concerne sa structure temporelle. Cette composante de la signification d'un texte (la composante temporelle), pratiquement absente dans les premiers travaux sur le résumé automatique, fait aujourd'hui partie, à côté des problèmes liés à la résolution des anaphores, des principaux thèmes de recherche pour les systèmes de résumé automatique et de filtrage d'informations. Nous verrons que ce thème de la temporalité peut lui aussi être appréhendé au travers de la double perspective approche de surface *vs.* approche en profondeur, cette dernière étant par définition la seule à même de permettre le calcul et la représentation de la structure temporelle d'un texte. En abordant enfin une nouvelle perspective de recherche en ce qui concerne les modes d'accès au texte (cf. 4.3.3), nous chercherons à montrer que la compréhension passe peut être par encore une troisième voie après celle en surface et en profondeur, la voie qui fait de la navigation textuelle et du parcours de lecture assisté, le vecteur d'une nouvelle forme de compréhension d'un texte.

4.2. Résumer un texte

4.2.1. Les besoins

La rédaction d'un résumé de texte, le résumé scolaire, constitue dans le système éducatif français un exercice traditionnel permettant de vérifier qu'un élève a compris le texte étudié. Ce type d'exercice obéit à des contraintes précises, comme l'interdiction d'utiliser des extraits du texte original ou l'obligation de reformuler les points essentiels développés par l'auteur. Les enseignants considèrent que comprendre un texte écrit peut être contrôlé en vérifiant que le texte produit par un élève est cohérent, synthétique, informatif, etc. C'est à partir de cette conception du résumé et de la compréhension, exprimée plus ou moins explicitement dans leurs travaux, que des chercheurs en psychologie cognitive [KIN 78] ou en IA [SCH & ABE 77, SAB 78] se sont attaqués à la construction de modèles de compréhension ou de systèmes automatiques aptes à simuler ou à reproduire cette activité humaine.

La production de résumés est depuis longtemps une activité sociale et plus particulièrement économique, puisqu'elle est apparue dès le début de l'écriture comme l'atteste la découverte de tablettes produites par la civilisation sumérienne,

soit 3600 ans avant le début de l'ère chrétienne. Plus récemment et jusqu'à la fin des années 70, la plupart des centres de documentation scientifique, comme par exemple le Centre de Documentation Scientifique et Technique du CNRS, employaient des documentalistes dont la tâche était de rédiger des résumés d'articles scientifiques. Il convient de remarquer que ces documentalistes, bien que spécialisés par discipline, ne pouvaient disposer ni des connaissances nécessaires à la compréhension de certains articles particulièrement novateurs [END 98] ni du temps que nécessite la lecture approfondie d'un article scientifique d'une dizaine de pages (puisqu'en moyenne un résumeur professionnel y consacre entre 12 à 20 minutes). Les résumés, distribués aux abonnés des centres de documentation sous forme papier, constituaient des instruments essentiels qui devaient permettre aux chercheurs de prendre rapidement connaissance des thèmes importants de l'article en question, et de décider de s'engager ou non dans une lecture approfondie. De ce fait, les résumés se devaient d'être le plus général possible puisque destinés à des lecteurs dont les besoins pouvaient être très différents. À l'heure actuelle, les bureaux d'étude des grands groupes industriels, qui ont une production écrite importante sous forme de rapports internes, (documents dont la longueur excède souvent la centaine de page), exigent de leurs rédacteurs d'y adjoindre une synthèse d'environ une page [LER & al. 94], afin de fournir aux membres de l'entreprise un premier accès à l'information. Ce type de résumé s'apparente ainsi à un instrument d'aide à la décision pour le lecteur qui, à la lecture de ces résumés, doit pouvoir décider de lire ou non le texte.

La revue de presse ou plus généralement la note d'intelligence économique, dont l'objectif peut être par exemple de surveiller quotidiennement l'image d'une entreprise dans la presse, représente un autre exemple de production de résumés. Il s'agit là de produire un texte très court, éventuellement réduit à la mention d'un évènement très factuel, en vue de réagir ou d'anticiper une situation de crise. Dans ce cas, le contenu du résumé n'est pas déterminé par l'auteur du texte mais doit répondre à des besoins spécifiques d'un lecteur. Ce type de résumé, souvent construit par simple extraction de phrases d'un texte source, est très proche des produits fournis par les techniques d'extraction d'information comme dans les conférences Message Understanding Conference (MUC) [MUC 97].

Depuis quelques années, la numérisation des textes et leur disponibilité quasi immédiate sur des dispositifs techniques, (assistant personnel, téléphone mobile, ordinateurs portables connectés à Internet), sont à l'origine des demandes nouvelles dans le domaine du résumé. Il ne s'agit plus alors de résumer un texte mais de résumer plusieurs textes qui relatent un même évènement ou de produire un texte qui synthétise l'évolution temporelle d'une situation. Un système qui aurait fournit en décembre 2004 l'estimation du nombre de victimes causées par le tsunami, en résumant les différentes dépêches des agences de presse se serait situé dans cette problématique du résumé multidocuments. La prise en compte de la temporalité devient alors cruciale puisqu'en quelques heures le nombre de victimes était

multiplié par dix puis par cent. On constate d'ailleurs [CAR & al. 98, BAR & al. 02] l'émergence de nouveaux travaux cherchant à prendre en compte la temporalité (intra ou inter-documents) ainsi que des propositions pour des normes d'annotation temporelle [FER & al. 03].

Ces différents exemples montrent à quel point le terme de résumé recouvre en fait des notions et des attentes fort disparates. Ces disparités tiennent à la fois aux caractéristiques des textes à résumer (narratifs, journalistiques, etc.), aux attentes du lecteur (prendre connaissance du thème, décider de lire ou de ne pas lire le texte, connaître l'opinion ou identifier les arguments de l'auteur du texte, synthétiser une information du point de vue de son évolution temporelle, etc.) et aux contraintes technologiques (support imprimé ou numérique, résumé indépendant ou non du texte original). Le seul point commun à ces différentes conceptions, est qu'un résumé d'un texte est un autre texte qui reflète, de manière condensée, les idées, thèmes, méthodes, résultats, évènements, opinions, etc. exprimés dans le texte d'origine.

Les modèles et les méthodes à mettre en œuvre relèvent intégralement du Traitement Automatique des Langues (TAL) en mobilisant des traitements comme l'identification de la structure (identification des titres, paragraphes, phrases), l'analyse morpho-syntaxique, l'analyse syntaxique, sémantique et pragmatique. Les traitements qui relèvent du TAL mais qui sont spécifiques au domaine du résumé automatique concernent l'identification des éléments considérés comme saillants dans un texte, du point de vue de l'auteur ou du lecteur, et du processus de réduction ou de condensation à mettre en œuvre pour obtenir un texte plus court [MIN & DES 00, MAN 01, MIN 02].

En ce qui concerne l'identification des éléments saillants, deux approches ont été, et sont toujours, explorées. L'une postule qu'il est nécessaire de comprendre pour résumer, ce qui implique la construction préalable de modèles de compréhension d'un texte ; l'autre esquive explicitement ou implicitement ce problème de la compréhension, s'appuie soit sur des analyses linguistiques, soit sur des études statistiques de corpus, en vue d'attribuer à un segment textuel (une phrase, un paragraphe) un « score » numérique ou une annotation sémantique. Quant aux algorithmes de condensation, qui par construction rompent la linéarité du texte, ils vont entraîner des problèmes qui tiennent à la cohérence et plus spécifiquement à la rupture de la chronologie temporelle, notamment dans les textes qui relatent des évènements.

Depuis quelques années, les recherches dans le domaine du résumé automatique se sont élargies d'une part, vers la *fouille sémantique de textes*, en cherchant à appliquer des algorithmes de repérage d'informations saillantes pour annoter sémantiquement les textes, et d'autre part vers la *navigation textuelle* en

questionnant le principe même de la production d'un résumé pour s'orienter plutôt vers la notion de parcours de lecture. La fouille sémantique de textes associe les techniques de recherche d'information et de TAL. L'exemple des travaux menés par des chercheurs du Centre d'Etude de la Vie Politique Française (CEVIPOF) qui développe des analyses sur la politisation de la parole illustre ce type de besoin. Il s'agit de savoir comment est définie la politisation de la parole dite « populiste » en identifiant dans des retranscriptions d'interviews les passages qui relèvent du « récit anecdotique », de la « montée en généralité » ou de l' « expression d'un clivage ». L'étape suivante vise à repérer la prise de position de l'interviewé dans le clivage, c'est-à-dire le passage à une situation politisée, en identifiant des sous-catégories comme par exemple : « naturalisation du clivage », « désignation d'un responsable », « identification à un groupe impliqué », etc. On conçoit que le dépouillement de ces entretiens, qui représentent plusieurs dizaines de pages, est une tâche très lourde. L'hypothèse posée par les chercheurs du CEVIPOF, en collaboration avec des chercheurs du laboratoire LaLICC[3], est que le repérage de ces catégories peut s'appuyer en partie sur des marques linguistiques et qu'il convient ensuite d'exploiter des outils de navigation textuelle pour, par exemple, vérifier qu'un « clivage » est toujours encadré par une « montée en généralité ». Un outil de fouille sémantique fournit ainsi aux chercheurs du CEVIPOF[4], et plus généralement aux sociologues qui dépouillent des textes, un moyen de repérer rapidement des séquences textuelles qui caractérisent les catégories recherchées.

Qu'il s'agisse de résumer automatiquement ou de fouiller sémantiquement des textes, se pose la question d'identifier les divers rapports qu'entretiennent les différents besoins décrits ci-dessus et la nécessité de comprendre le texte pour répondre à ceux-ci de manière adéquate.

La première remarque tient à la notion de compréhension. Comme nous l'avons déjà souligné précédemment (cf. 4.1.), quand un système identifie automatiquement des unités simples, comme par exemple des entité nommées (le CNRS, la société PSA, etc.) ou des expressions temporelles simples (le 29 août 2004, mardi prochain, etc.), le fait même de pouvoir distinguer dans un ensemble textuel ces informations implique un niveau de compréhension.

La deuxième remarque concerne les interactions avec l'utilisateur que vont proposer, ou non, les systèmes de résumé ou de fouille de texte. Le fait de modéliser des connaissances qui vont assister ou guider le lecteur relève de notre point de vue d'une compréhension du texte.

---

[3] Laboratoire de recherche du CNRS et de l'Université Paris-Sorbonne.
[4] Laboratoire de recherche du CNRS et de la Fondation Nationale des Sciences Politiques.

4.2.2. Les solutions proposées pour le résumé automatique

4.2.2.1. L'approche fondée sur la construction de modèles

Les approches que l'on peut qualifier d' « approche par compréhension » postulent qu'un lecteur construit des représentations qui symbolisent le contenu du texte lu [DIJ & KIN 83, JOH 83]. Ces représentations sont mémorisées et associées à des connaissances du lecteur qui sont présentes dans sa mémoire. Il existe plusieurs modèles cognitifs qui s'appliquent à expliquer ces processus de compréhension et de mémorisation mais tous ont en commun les points suivants. Premièrement, le modèle doit permettre de construire plusieurs niveaux de représentation d'un texte. Ensuite, ces représentations sont utilisées pour établir des liens de cohérence entre, d'une part, les différents énoncés du texte et les connaissances du lecteur, d'autre part. Enfin, ces deux processus, sont contraints par le processus de mémorisation. Dans le domaine du résumé automatique, le modèle élaboré par Kintsch (1978, 1998), appelé construction/intégration (C/I), fait sans aucun doute référence.

Le modèle Construction/Intégration

Le cœur du modèle C/I de Kintsch repose sur le concept de proposition atomique. Dans une première version [KIN 74], une proposition atomique est définie comme une structure relationnelle simple composée d'un prédicat et d'un ou plusieurs arguments. Il faut souligner que Kintsch précise qu'il emprunte le terme de proposition à la logique mais qu'il l'emploie dans un sens étendu et que notamment la notion de valeur de vérité est sans objet [KIN 98, p. 60]. Par exemple, l'énoncé « *Marie donne un livre à Jean* »[5] sera encodé sous la forme de la proposition atomique suivante :

Donner [agent : Marie, Objet : Livre, But : Jean]

Dans une deuxième version du modèle [VAN & KIN 83], le concept de proposition complexe est introduit. Une proposition complexe est composée de plusieurs propositions atomiques qui sont subordonnées à une représentation propositionnelle principale (cf. figure 4.1.).

5. Tous les exemples sont extraits de [KIN 74 ; 98] et traduits par nos soins.

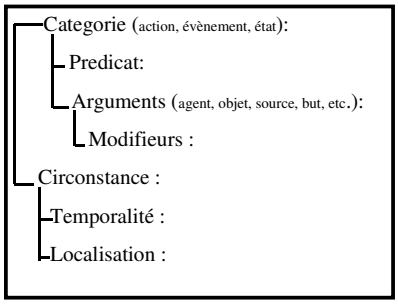

Figure 4.1. Proposition complexe

Ainsi, la phrase « *Hier, Marie a donné à Jean le vieux livre dans la bibliothèque* » sera représentée de la manière suivante :

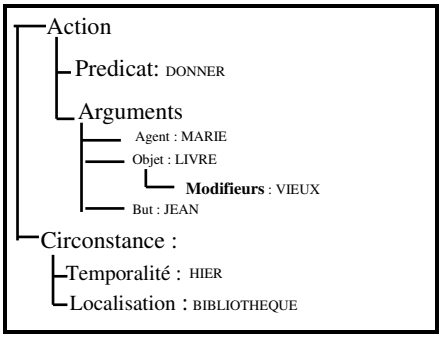

Figure 4.2. Représentation d'une phrase

Comme le montre cet exemple et comme Kintsch l'explique lui-même, toutes les marques, morphologiques, syntaxiques et sémantiques ne sont pas représentées. Kintsch insiste sur les choix pragmatiques qui guident la construction de telles représentations.

Le modèle C/I va s'appuyer sur cette représentation propositionnelle pour la représentation d'un texte. Le modèle distingue d'une part la microstructure et la macrostructure d'un texte et d'autre part la base du texte (*textbase*) et le modèle de situation. Les propositions qui sont construites, à l'aide du formalisme présenté ci-dessus, à partir de l'analyse des phrases qui composent le texte constituent la base du texte. Les connaissances nécessaires à la compréhension du texte et qui préexistent dans la mémoire du lecteur sont, elles aussi, représentées à l'aide de

propositions complexes. La composition de ces deux représentations constitue le modèle de situation.

La microstructure correspond à la structure locale du texte ; c'est un réseau de propositions construit à partir de la base du texte et des informations contenues dans la mémoire à long terme du lecteur. La macrostructure est un ensemble hiérarchiquement ordonné de propositions qui représente la structure globale du texte et qui est construit à partir de la microstructure. L'élaboration de ce réseau hiérarchisé s'obtient par la mise en œuvre, au cours même du traitement, de trois macro-règles :

- *élimination*. Une proposition ou une information qui n'est pas nécessaire à l'interprétation des autres propositions est éliminée ; on peut remarquer qu'une des conséquences de l'application d'une règle d'élimination est la suppression de la référence temporelle « *Hier*» dans la phrase « *Hier, Pierre a donné à Marie deux billets de cinéma* ».

- *généralisation*. Une proposition est remplacée par une proposition plus générale ;

- *condensation*. Une séquence de propositions est remplacée par une proposition plus englobante.

Pour Kintsch, un résumé idéal est le texte qui exprimerait la macrostructure [KIN 98, p. 50]. Il convient de remarquer que cette macrostructure n'est pas indépendante du lecteur puisque le processus de construction met en jeu des connaissances (le modèle de situation) propres au lecteur. Néanmoins, Kintsch souligne que les diverses et nombreuses expérimentations qui lui ont permis de tester ce modèle tendent à montrer que les variations entre les macrostructures élaborées par différents lecteurs sont minimes. Tous ces travaux ont permis de vérifier les qualités, du point de vue de sa capacité prédictive notamment, du modèle C/I. Les résultats sont satisfaisants à l'exception des deux cas suivants. Tout d'abord, lorsque les textes à comprendre sont très longs et ensuite lorsque le lecteur comprend mal le texte comme c'est le cas pour certains articles scientifiques, ce qui correspond à certaines situations rencontrées pour résumer un texte.

Du point de vue du TAL, ce modèle a été implanté pour traiter des textes courts dans des domaines très restreints [CUM & al. 88], mais aucun système apte à traiter n'importe quel type de texte n'a jamais été réalisé. La tâche d'analyse qui est au cœur du modèle, la construction de la base de texte est au-delà des capacités actuelles des analyseurs syntaxico-sémantiques. Afin de contourner ce problème, Kintsch propose dans son dernier ouvrage de substituer au concept de proposition[6],

6. « A concept or proposition can thus be thought of as a vector of numbers, each number indicating the strength with which the concept or proposition is linked to another concept or proposition » [KIN 98, p. 87].

le concept de vecteur en s'appuyant sur la LSA (Latent Semantic Analysis) [DEE & al. 90]. Cette technique, qui est issue des travaux menés en recherche d'informations construit par apprentissage sur des textes, des représentations vectorielles de la corrélation entre les mots du texte. Ce choix représente un tournant important puisqu'il revient à abandonner une représentation purement symbolique pour la remplacer par une représentation purement numérique.

Les modèles fondés sur la notion de réseau sémantique

Alors que les travaux de W. Kintsch avaient pour objet de construire un modèle cognitivement pertinent, d'autres recherches menées dans le cadre des travaux en intelligence artificielle se sont orientées vers la construction de représentations censées capter le contenu sémantique d'un texte mais sans postuler que celles-ci reflètent un processus de compréhension. Les modèles visent plutôt à fournir un cadre conceptuel aux représentations construites. Ainsi, la représentation du texte est le résultat d'une analyse syntaxique classique ou s'appuie sur l'analyse casuelle de Fillmore [FIL 68]. La forme de cette représentation varie selon les approches. Il peut s'agir d'une représentation causale des évènements [SCH 75], ou d'un graphe construit à partir de la séquence ordonnée des évènements conceptuels [ALT & al. 90]. On pourrait aussi citer d'autres formes de représentation dans la tradition de l'IA (pour plus de détails voir par exemple [ALT 91]).

La représentation du texte ainsi construite devient alors l'entrée d'un module qui procède à sa réduction au moyen d'une série d'opérations. Pour chacun de ces modèles, ces opérations de condensation se fondent sur des hypothèses concernant l'importance des parties de la représentation retenues pour le résumé final :

- pour Schank [SCH 75], les évènements « les plus intéressants » sont ceux qui correspondent à la succession narrative d'évènements reliés par des liens causaux : c'est le chemin critique de la représentation.

- l'importance d'une information exprimée dans la représentation est évaluée quantitativement dans d'autres approches. Par exemple, Alterman et Bookman [ALT & BOO 92] considèrent l'importance d'un évènement comme une fonction du nombre de nœuds associés à cet évènement dans la représentation. Ainsi, on peut fixer un seuil quantitatif d'importance à partir duquel « les évènements les plus importants » sont sélectionnés.

Le résultat obtenu à l'issue de cette étape est une représentation réduite aux parties les plus importantes de la représentation du texte initial. L'étape suivante consiste à engendrer un texte à partir de la représentation résultante. Ce texte est considéré comme le résumé du texte initial.

4.2.2.2. Synthèse sur la notion de compréhension dans les systèmes fondés sur la construction de modèles

Une hypothèse est commune à ces différentes approches : le processus de compréhension du texte en vue de le résumer, c'est-à-dire l'identification des parties importantes d'un texte, est modélisé comme la recherche d'un chemin dans un graphe. Les différences tiennent aux éléments constitutifs du graphe et aux critères utilisés pour calculer le chemin retenu. À ce titre il convient de remarquer que la représentation de la temporalité est plutôt ignorée ou, plus exactement, que le temps n'est pas considéré comme un paramètre crucial pour le calcul du chemin dans le graphe. La recherche d'un chemin présuppose la construction d'un modèle qui s'appuie sur la notion de représentation. Lassègue replace cette dernière dans une perspective historique :

« *(...)* on se rend compte que [la notion de représentation] n'est pas une notion qui serait donnée comme entièrement constituée sous la forme d'un objet abstrait, mais qu'elle est au contraire le résultat d'un processus ininterrompu d'interprétation théorique. (...). La notion de représentation n'a pas de statut scientifique définitif et n'est peut être pas destinée à en avoir un. (...) En postulant l'existence d'un niveau cognitif spécifique au mental, conçu comme pleinement autonome par rapport aux processus neuronaux, les sciences cognitives ont accordé une place capitale à la notion de représentation symbolique du réel. Elles considèrent en particulier que le domaine de la représentation peut être adéquatement décrit grâce à une méthodologie mise au point à partir de formalismes dérivés des mathématiques et de la logique : la représentation symbolique est en effet considérée comme susceptible d'un traitement de nature calculatoire. (...) Le domaine de la représentation apparaît comme assimilable à une machinerie logique et l'ordinateur apparaît naturellement comme le modèle le plus adéquat pour rendre compte du fonctionnement de ce domaine …» [LAS 93]

Les attitudes vis-à-vis de la question des représentations diffèrent, certains chercheurs allant même jusqu'à récuser le recours à cette notion [voir la discussion de [DEN 93]). Cette démarche implique que des « indicateurs objectifs » soient identifiés. Pour certains chercheurs, le langage apparaît justement comme pouvant servir de révélateur de propriétés cognitives générales. L'analyse détaillée des configurations linguistiques donnerait ainsi des indices précieux sur les processus mentaux qui les ont produites, faisant de la linguistique une discipline essentielle pour les sciences cognitives.

4.2.2.3. L'approche fondée sur des analyses de marques de surface

Parallèlement aux travaux fondés sur la construction de modèles, cognitifs ou conceptuels, d'autres recherches se sont orientées très tôt [EDM 69] vers

l'exploitation des marques de surface : typographiques, lexicales, grammaticales ou structurelles. Ces recherches sont généralement qualifiées d'approches par extraction, car elles construisent le résumé à partir des phrases qui composent le texte source. Néanmoins, comme nous le montrerons dans la suite de ce chapitre, des méthodes récentes [BAR & al. 02] fondées sur l'analyse de marqueurs peuvent produire des résumés composés de phrases qui n'existent pas telles quelles dans le ou les textes sources. L'hypothèse commune à toutes ces approches consiste à postuler l'existence de segments textuels (propositions, phrases, etc.) saillants, c'est-à-dire représentatifs du contenu du texte source. Cette hypothèse est controversée. D'une part certains travaux [END 98, SAG & LAP 02] montrent que les résumés, produits par des professionnels, sont souvent constitués de phrases, identiques ou tout du moins très proches lexicalement, des celles du texte source alors que d'autres travaux montrent au contraire une forte disparité entre les résumés produits par différents professionnels [RAT & al. 61, TEU & al. 97]. Les différences entre ces approches résident dans les algorithmes mis en œuvre pour calculer cette saillance : calcul d'un score numérique, calcul d'une annotation sémantique du segment textuel considéré ou bien encore calcul des relations entre les segments textuels.

Sélection de segments textuels par calcul d'un score numérique

Ce type de méthode consiste à calculer un score S pour chaque segment textuel puis à conserver les segments dont le score est supérieur à un certain seuil, ou à fixer un nombre absolu de segments à conserver en fonction d'un pourcentage de réduction.

Le score le plus couramment utilisé est une fonction de la fréquence du mot dans le texte ; il est issu des techniques utilisées dans les sciences de l'information. Le calcul de ce score appelé, *tf*idf,* nécessitant de connaître la fréquence moyenne d'un mot, il est nécessaire de construire un corpus de référence composé d'un ensemble de textes considérés comme représentatifs. L'hypothèse sous-jacente à ce type de méthode est que l'importance d'un segment textuel est une fonction des éléments lexicaux qu'il contient, moyennant un correctif lié au domaine traité et aux usages syntaxiques ; en d'autres termes, l'hypothèse est que les phrases qui contiennent les mots les plus fréquents sont représentatives de la thématique du texte.

Un autre mode de calcul de score est fondé sur le repérage d'expressions prototypiques (*cue-phrases*). Cette approche part d'observations effectuées sur des corpus de textes essentiellement scientifiques et techniques, et fait émerger des critères de sélection autres que ceux qui sont fondés uniquement sur la fréquence des mots :

- certains mots ou expressions du texte peuvent indiquer l'importance des phrases, indépendamment de considérations purement fréquentielles. Par exemple,

les phrases contenant des expressions du type « notre travail », « ce papier », « la présente recherche », etc., sont des marques placées par l'auteur pour présenter le thème de son article ;

- certaines expressions, en se référant à des passages précédents, font office de liens structurels entre les différentes parties d'un texte et peuvent donc être exploitées pour construire des résumés plus cohérents. C'est le cas d'expressions du type « présenté précédemment », « énoncé au-dessus », etc. ; la position des phrases dans un texte peut aussi être utilisée comme critère de sélection. Par exemple, les phrases de l'introduction, de la conclusion ou de certaines sections du texte ont une certaine importance par rapport à d'autres phrases du texte.

Les travaux de Paice [PAI 90] sont les plus représentatifs de cette approche. L'auteur propose des critères d'importance, en partie à partir de l'observation d'expressions linguistiques du type de l'exemple de la figure 1.3.

In this first **paper we** will **discuss** a simple **method** for...
+3 +2 +2 (+7)
In this investigation a new automatic process **is** briefly **discussed**...
+3 +2 (+5)

Figure 4.3. Exemples d'expressions importantes d'après [PAI 90]

À partir d'expressions similaires, l'auteur définit des règles qui se fondent sur deux types d'informations. Premièrement, la présence de mots ou de classes de mots. Ainsi, la classe de mots se référant à la discussion (« discuss words ») est constituée par les éléments suivants[7] : *discuss-, introduce, present, examine, describe, review, report, outline, consider, investigate, explore, assess-, analy-, synthesi-, styd-, survey, ask, simplif-, deal-, trace(-2), cover(-2)*. Deuxièmement, la présence des schémas d'expressions construits à partir d'une généralisation des expressions linguistiques déjà observées. Ces schémas d'expressions spécifient comment peuvent être associées les classes de mots et précisent, sous forme de nombre entiers, le nombre de mots qui peuvent s'intercaler dans l'expression.

De cette manière, à chaque phrase est attribué un score résultant du cumul des poids figurant dans les schémas d'expressions. Après la phase de calcul des scores intervient la phase d'extraction des phrases ayant cumulé les plus grands scores. Celles-ci sont extraites avec les phrases adjacentes lorsqu'elles contiennent des références à des éléments externes (voir chapitre sur la référence de B. Gaiffe).

7. Le symbole - indique que les différentes flexions du terme doivent être prises en compte et les nombres entre parenthèses sont des poids qui indiquent la fréquence d'utilisation des mots qui leur sont associés par rapport au reste de la liste.

Sélection de segments textuels par analyse de relations

Les travaux de Marcu [MAR 97] se fondent explicitement sur la Rhetorical Structure Theory (RST) développée par [THO & MAN 88] en vue de construire ce que l'auteur appelle un « analyseur rhétorique » (*rhetorical parser*). Cet analyseur considère le texte comme un ensemble d'unités élémentaires disjointes, reliées entre elles par deux types de relations binaires, les relations *paratactic* et les relations *hypotactic*. Les relations *paratactic* lient des unités d'égale importance, alors que les relations *hypotactic* lient un *nucleus*, c'est-à-dire une unité considérée comme essentielle par l'auteur, avec un satellite, unité considérée comme non essentielle. L'analyseur va d'abord construire un arbre binaire, l'arbre RS, en appliquant un algorithme fondé sur l'exploitation d'un ensemble de 450 marqueurs (essentiellement des connecteurs), résultat d'une étude d'un corpus de 7900 fragments de textes. Puisqu'il peut exister plusieurs relations entre les unités élémentaires, la détection de ces relations binaires étiquetées sur le texte conduit à la construction d'un ensemble d'arbres binaires. L'étape suivante consiste à choisir l'arbre optimum (voir MAR 97 pour une présentation détaillée de cet algorithme). La construction du résumé consiste alors à choisir, dans l'arbre sélectionné, les unités élémentaires qui sont considérées comme saillantes. Par construction, l'unité élémentaire qui se trouve à la racine est l'unité la plus saillante. Si la taille du résumé n'est pas atteinte, les nœuds fils sont sélectionnés, et ainsi de suite jusqu'à obtenir un résumé de taille adéquate.

Sélection de segments textuels par calcul d'une annotation sémantique

L'approche développée tout d'abord dans le projet SERAPHIN [LER & al. 94 ; BER & al. 96] puis étendue dans le projet Filtext [MIN & al. 01] est fondée sur la méthode d'exploration contextuelle [DES & al. 91, DES & MIN 05]. Cette méthode vise à identifier les connaissances linguistiques en les restituant dans leurs contextes et en les organisant en tâches spécialisées. Pour appliquer la méthode d'exploration contextuelle, le linguiste doit accumuler des marqueurs linguistiques qui peuvent être des unités lexicales simples ou complexes. Ces marqueurs linguistiques, lorsqu'on les identifie dans un texte constituent des indices de certaines notions discursives. Certains indices, ceux dont la valeur sémantique est à désambiguïser, sont appelés indices déclencheurs. Le linguiste doit ensuite spécifier, sous forme de règles d'exploration contextuelle, les dépendances contextuelles entre ces indices. La partie « Action » d'une règle spécifie le segment textuel et type d'annotation sémantique qui doit être attribué à celui-ci. La figure 4.4 présente un exemple d'une règle qui attribue à une phrase l'annotation « Annonce Thématique ». Chaque règle d'exploration contextuelle permet d'attribuer une étiquette aux phrases qui contiennent les indicateurs et les indices pertinents et contribue ainsi à produire une structure hiérarchisée « décorée » par des informations sémantiques. Une dizaine

d'annotations (« Annonce Thématique », « Conclusion », « Définition », etc.) ont ainsi été définies.

Certaines phrases étant étiquetées, il devient possible de construire des résumés qui répondent aux besoins spécifiques d'un utilisateur en appliquant différentes stratégies de sélection. Ces fonctionnalités ont ensuite été étendues par [CRI & COU 04], en développant des interfaces pour produire des « résumés dynamiques ».

Nom de la règle : Rthematique **;**
Tâche déclenchante : Thématique ;
Commentaire : capte un schéma du type : *Dans les lignes qui suivent … nous présentons …*
Classe de l'Indicateur : &verbe_presentatif ;
E1 := Créer_espace_de_ Recherche (voisinage Indicateur)
L1:= &partie_document3
L2:= &partie_document1
Condition : Il_existe_un_indice x appartenant_à E1 tel_que classe_de x appartient_a (L1) ;
Condition : Il_existe_un_indice y appartenant_à E1 tel_que classe_de y appartient_a (L2) ;
Precede (x,y, contrainte) ;
Actions :
1 : Attribuer(PhraseParent « Annonce_Thematique »)

Figure 4.4. Un exemple de règle d'exploration contextuelle [MIN & al. 01]

Ces dernières années d'autres systèmes de résumé automatique, qui attribuent des annotations à des segments textuels mais qui se fondent sur d'autres approches (transducteurs, analyse syntaxique partielle, etc.) que l'exploration contextuelle pour calculer ces annotations, ont été développés [SAG 02 ; TEU & MOE 98] sans toutefois mieux préciser le statut de ces annotations.

4.2.2.4. Synthèse sur la notion de compréhension dans l'approche fondée sur l'analyse de marques de surface

Comme l'illustre la règle d'exploration contextuelle présentée ci-dessus, les systèmes fondés sur l'analyse de marques de surface s'appuient sur des « raisonnements », ou plus exactement sur des heuristiques qui sont modélisées, suivant les systèmes, sous des formes plus ou moins explicites. Remarquons ainsi que la règle proposée pourrait s'écrire sous une forme procédurale ou sous la forme d'une expression régulière. Le résultat obtenu serait le même, mais par contre l'expressivité et la représentation des connaissances linguistiques mobilisées seraient fort différentes.

De fait, tous ces systèmes se sont positionnés hors de la problématique de la compréhension en arguant du fait que d'une part, ils ne proposent pas de modélisations élaborées des connaissances linguistiques utilisées, et que d'autre part ils ne construisent pas, pour effectuer leur calcul, une représentation du texte. Si cela était effectivement vérifié pour les tous premiers systèmes fondés notamment sur la récurrence lexicale (cf. 4.2.2.3), on peut remarquer que les systèmes plus récents construisent des représentations (arbres rhétoriques par exemple) ou que ces représentations sont enfouies dans les algorithmes qui construisent les extraits (cf. notion de résumé dynamique). Nous rejoignons ainsi les questions posées à la communauté par les organisateurs l'appel du colloque CIDE 7 qui écrivaient entre autre que :

« La mise en avant du « sens » a en effet longtemps été regardée avec beaucoup de scepticisme au profit de traitements dits « de surface », s'attachant à « la forme » par opposition au « contenu ». Cette perception est en train de changer. Des progrès significatifs ont été réalisés au cours des dernières années, d'abord sur le document textuel (extraction d'informations, réponse aux questions, résumé automatique...), puis relayés de plus en plus dans les autres médias (extraction d'information et indexation de documents sonores et vidéo par le contenu, résumé d'œuvres...). Un autre point de vue, plus radical, serait de considérer que même les traitements dits « de surface » ou « numériques » sont en fait, à y bien regarder, sémantiques. Si le « sens » ne se réduit pas à « l'information », produire de l'information, n'est-ce pas produire du sens ? » [CIDE 7, juin 2004].

4.3. Quelques perspectives proposées pour le résumé automatique

4.3.1. Comment assurer une continuité référentielle dans la lecture d'un résumé

Une des questions cruciales qui traversent l'ensemble des travaux sur la compréhension de textes concerne la nature des relations qui existent entre les propositions d'un texte. Ces relations renvoient au caractère intégratif du processus de compréhension, ce dernier soulevant alors le problème majeur de la structure à (re)construire dans le processus de compréhension [COI et al. 96, p. 254]. Dans un texte, typiquement, les liens entre propositions sont établis par des conjonctions (ou connecteurs) ou selon des procédures anaphoriques (comme les pronoms ou les ellipses) mais l'intégration peut aussi être envisagée en tenant compte d'autres liens comme ceux d'ordre métonymique ou synonymique, entre lexèmes verbaux et nominaux [HAL & HAS 76] ou encore en tenant compte des liens temporels entre les propositions. Par ailleurs, un aspect des processus intégratifs à l'œuvre dans la compréhension de textes concerne le fait que les processus intégratifs sont également « incrémentaux ». En d'autres termes, quand on lit un texte, l'intégration

de la proposition en cours de traitement dans la représentation provisoire d'un texte crée une représentation nouvelle, plus riche. L'ordre des propositions est essentiel, chaque proposition étant interprétée dans le contexte constitué par l'ensemble des propositions précédentes. Un des déterminants majeurs de la cohérence d'un texte est l'aisance avec laquelle les multiples références aux mêmes individus et objets peuvent être pistées. Le moyen le plus courant de rendre la référence facile à suivre consiste à assurer la continuité référentielle entre phrases consécutives. Garnham, Oakhill et Johnson-Laird (82) ont mis à l'épreuve le fait selon lequel il est plus difficile de comprendre et de se souvenir des histoires dont l'ordre des phrases a été bouleversé, que les histoires d'origine[8]. Garham et Oakhill (93) ont avancé l'idée que la difficulté de compréhension des histoires en désordre découlait, au moins en partie, de la présence de nombreuses expressions référentielles, dont les référents sont obscurs : le désordre détruit la continuité référentielle ainsi que la structure de l'histoire. La perception de la cohérence d'un texte (*i.e.* les critères d'acceptabilité) constitue un phénomène complexe ; la notion même de cohérence étant une « notion difficile à cerner mais apparemment bien ancrée dans l'expérience et l'intuition des sujets parlants » [CHA 99]. Pour sa part, L. Lundquist (80) la conçoit comme présupposée : « les phrases successives d'un texte révèlent sa cohérence ».

Les systèmes de résumé sont confrontés à ce problème car les textes produits (les résumés) se doivent d'être cohérents pour être intelligibles et donc utiles ; autrement dit, ils se doivent d'offrir une certaine continuité référentielle. Ils sont ainsi confrontés à deux difficultés majeures en rapport avec la notion de cohérence textuelle.

La première difficulté majeure à laquelle sont confrontés les systèmes de résumés concerne la temporalité, la structure temporelle d'un texte constituant en effet l'un des éléments essentiels de sa cohérence. Dans les approches à base de marqueurs, les systèmes de résumés mono-documents ne se sont pas confrontés à celui-ci puisque la stratégie adoptée consiste à extraire des phrases dans l'ordre dans lequel elles apparaissent dans le texte initial. Le problème apparaît cependant nécessairement quand il s'agit de systèmes de résumés multi-documents ; ceux-ci se doivent en effet de gérer l'imbrication temporelle des différentes phrases sujettes à l'extraction dans les différents documents, et, ce faisant, de gérer la diversité et l'imbrication des informations qui participent au calcul des références temporelles (cf. 4.3.2).

La deuxième difficulté majeure à laquelle sont confrontés les systèmes de résumés – mono-documents comme multi-documents – concerne la nécessité de résoudre le problème difficile des anaphores, pronominales et nominales (voir

8. Ce résultat avait été utilisé comme preuve de l'utilisation de grammaires de récits, lors de la compréhension.

chapitre sur la référence de B. Gaiffe). Le premier sous-problème concerne ainsi le repérage du référent des anaphores pronominales comme dans l'exemple: « *J. Chirac a inauguré le Salon de l'Agriculture. Il a rencontré… »*. Dans le cas ou la phrase qui commence par « Il a » serait sélectionnée pour figurer dans le résumé, il conviendrait de remplacer le pronom « il » par « J. Chirac » et non pas par « Salon ». Actuellement, ce problème est généralement résolu de manière satisfaisante, tout du moins pour des textes journalistiques ou scientifiques (mais beaucoup moins pour des textes narratifs)[9]. Les stratégies mises en œuvre s'appuient uniquement sur des connaissances syntaxiques et quelque fois sémantiques. Le deuxième sous-problème concerne l'anaphore nominale et est plus complexe à résoudre, exigeant d'utiliser des ontologies de domaines. Ainsi dans l'exemple « *J. Chirac a inauguré le Salon de l'Agriculture. Le chef de l'Etat a rencontré… »* l'association entre « Le chef de l'Etat » et « J. Chirac » exige des connaissances extra-linguistiques. La résolution de ce type d'anaphore, qualifiée d'anaphores infidèles par [LUN 05], est pourtant considérée par les enseignants des langues comme un indice de la bonne compréhension du texte. Une stratégie de contournement de ce problème de résolution des anaphores consiste à proposer des outils de navigation textuelle (cf. 4.3.3).

4.3.2. Intégrer l'analyse de la temporalité dans les systèmes de résumé

Actuellement, la temporalité dans les textes est appréhendée dans les traitements automatiques à deux principaux niveaux d'analyse et de représentation : l'un renvoie à la tâche d'ancrage des expressions temporelles dans un système calendaire (mise en relation avec des dates ou des périodes) ; l'autre renvoie à la tâche de calcul de l'ordonnancement temporel des évènements dans un texte. Historiquement, c'est en fait la deuxième – pourtant plus complexe – qui a fait l'objet des premiers travaux pour être ensuite peu à peu « délaissée » au profit de la première considérée comme plus réaliste et surtout comme indispensable dans le cadre du développement des systèmes de questions-réponses [TER 02] et de résumés multi-documents [MUL & TAN 04].

4.3.2.1. Rappels sur quelques éléments fondamentaux d'analyse de la temporalité dans les textes

Les traitements informatiques de la temporalité dans les textes sont depuis assez récemment le plus souvent désignés à l'aide du terme « d'annotation temporelle »,

[9] Dans les textes scientifiques, afin de faciliter la lecture, le référent lexical d'une anaphore pronominale se situe généralement dans la phrase précédente, ce qui restreint l'espace de calcul. En revanche, dans les textes narratifs, le référent lexical peut être situé beaucoup plus en amont, par exemple deux paragraphes avant l'occurrence de la marque anaphorique, en aval (cataphore) ou même de ne pas exister (anaphore résomptive).

renvoyant de manière générale à des tâches qui consistent à repérer dans les textes des informations d'ordre temporel et à les étiqueter sémantiquement. Cette manière de qualifier les traitements envisagés semble ramener là encore à l'opposition classique « approche en surface / approche en profondeur » en faisant des analyses temporelles des tâches relevant d'une approche en surface. En réalité, les faits linguistiques relevant de cette catégorie tout comme la manière dont les traitements les concernant sont envisagés sont d'une complexité telle qu'ils semblent échapper à cette « simple » opposition ; tout va finalement dépendre du type d'information d'ordre temporel que l'on cherche à repérer puis à annoter sémantiquement, depuis les expressions dites « calendaires » jusqu'aux relations temporelles entre propositions d'un texte, sachant que ces deux types de traitement sont étroitement liés. Ils renvoient en effet tous deux à un même problème fondamental, celui qui concerne l'interaction – au sein d'un texte comme d'un énoncé isolé – des différents modes d'expression de la temporalité dans la langue. Ces modes sont très divers et les marques qualifiées de temporelles peuvent être scindées en quatre types : les temps grammaticaux (auxquels correspondent différentes valeurs aspecto-temporelles), les connecteurs (qui participent explicitement au calcul des relations temporelles entre propositions), d'autres indices temporels tels que des dates ou des marqueurs de durées (qui peuvent inscrire les situations dans un système calendaire) et enfin des indices typographiques tels que le guillemet ou le point (qui ouvrent ou ferment des espaces de validation). Ces différents modes d'expression interagissent pour renvoyer à une certaine signification (aspecto-)temporelle, que l'on envisage cette dernière au niveau de l'énoncé isolé sous la forme d'une valeur aspecto-temporelle ou au niveau d'un texte sous la forme de relations temporelles entre propositions. Notons que ce type de démarche est actuellement principalement développée sur – et à partir – de corpus essentiellement journalistiques qui font beaucoup usage de ces expressions calendaires, contrairement à des corpus narratifs par exemple.

L'annotation des expressions calendaires dans les textes

La tâche de repérage des expressions temporelles dans un texte est complexe, non seulement du fait de la difficulté à délimiter parfois une expression temporelle (*cf.* exemples (1) et (2) de [VAZ 01]) mais aussi du fait de la difficulté liée à la décision de retenir ou non comme expression temporelle une expression contenant une unité calendaire (*cf.* exemple (3) de [VAZ 01]) [10].

[10] Dans tous nos exemples, nous soulignerons ce qui fonctionne comme indice déclencheur, c'est-à-dire l'expression calendaire, mettrons en gras ce qui est interprété comme un attribut, et enfin entre crochets ce qui correspond à l'expression temporelle complète. Remarquons ici que les systèmes annotent, non pas l'expression temporelle, mais la seule expression calendaire.

(1) « Le ministre est venu [3 minutes **après son porte-parole**] qui avait déjà annoncé la bonne nouvelle ».

(2) « La réunion a commencé et [3 minutes **après**] son porte-parole qui avait déjà annoncé la bonne nouvelle est parti pour la capitale ».

(3) « Trois mille séminaristes, jeunes prêtres et étudiants ont souhaité [vendredi] un bon anniversaire au pape qui fêtera [samedi] ses 21 ans de pontificat, en lui chantant en polonais "stolat" ("cent ans"), [**lors d'une** messe] dans la basilique Saint-Pierre ».

Aussi, la tâche d'annotation dans les textes des expressions qualifiées de temporelles renvoie en réalité principalement à une annotation des seules expressions calendaires. Elle vise à identifier dans les textes les expressions propres à être ancrées dans un système calendaire [SCH 02], c'est-à-dire dans un système qui permet de situer des évènements sur une échelle de temps, en fonction de la durée de ces évènements et selon une hiérarchie d'unités – encore appelées « grains ». On distingue alors classiquement deux sous-tâches : une première visant à repérer automatiquement les dites expressions calendaires ; la seconde visant à en réaliser l'ancrage sur une « ligne temporelle » sous forme de valeurs, le plus souvent notées en suivant le format standard ISO 8601 de manière à assurer une certaine « portabilité » des annotations. On peut estimer que ces deux sous-tâches obtiennent actuellement des résultats corrects, voir par exemple [MAN & WIL 00, SET & GAI 00, FIL & HOV 01].

En vue d'identifier les expressions calendaires, la plupart des systèmes actuels se fondent sur le repérage de certains lexèmes-clés (ou « indices déclencheurs ») que sont les « grains d'observation » classiques comme 'année', 'mois', 'jour'… Des marqueurs linguistiques tels que des articles déterminants ([**le** 12 juin]), des prépositions ([**en** 2005], [**à** 9 heures]) ou encore des connecteurs ([**pendant** 2005]) sont ensuite pris en compte dans les systèmes de représentation sous forme d'attributs qui permettent de distinguer les types d'ancrage réalisés.

En dehors du fait que la deuxième sous-tâche, à savoir celle d'ancrage des expressions calendaires, est confrontée au problème de la mise en œuvre d'un calcul dans le cas des expressions anaphoriques ([*L'*année *suivante*]) ou déictiques ([*avant-*hier]), elle reste surtout confrontée à deux problèmes principaux : le problème de l'ambiguïté de ces expressions ainsi que celui qui est lié aux changements de granularité.

Les exemples (4)-(4'), (5)-(5') et (6) illustrent différents types d'ambiguïté du point de vue de l'interprétation temporelle.[11]

(4) « [**Il y a** deux semaines], J. Chirac faisait une intervention télévisée remarquée ».

11. Les exemples (4) à (8) ainsi que les commentaire sur ceux-ci sont issus de [BAT & SCH 06].

(4') « Il a été déprimé [**pendant** deux semaines] ».

(5/5') « [Samedi], je **viendrai / suis venue** te voir ».

(6) « Je suis occupée [**les** jours] où tu ne l'es pas. »

Le couple (4)–(4') renvoie à une ambiguïté de l'unité calendaire «deux semaines» sur le plan de son interprétation temporelle (date *vs*. période). Par ailleurs, l'expression « il y a deux semaines » dans (4) renvoie à un autre type de problème qui a trait à la valeur calendaire précise à retenir (s'agit-il d'un renvoi à deux semaines environ ou à deux semaines jour pour jour ?). L'exemple (5) renvoie à une ambiguïté de l'unité calendaire «samedi» sur le plan de la valeur calendaire qui dépend du temps grammatical auquel est conjugué le verbe dans la phrase ; l'expression temporelle « les jours » dans (6) peut être interprétée comme un ensemble de jours contigus ou non et le contexte ne permet pas de désambiguïser. Si des cas d'ambiguïté comme (4) et (5) sont correctement traités par les systèmes, le cas de (6) reste « mal » traité selon nous. TIDES [FER & al. 03] invite en effet à choisir l'une des deux interprétations ; or, selon nous, cette expression n'est pas réellement ambiguë : quelle que soit la période d'occupation du mois - voire de l'année - qui est désignée, l'utilisation de cette expression vise seulement à oblitérer toute rencontre ce jour-là.

En ce qui concerne les changements de granularité, on peut en effet observer qu'il est très courant dans les textes, notre manipulation des unités calendaires nous permettant des effets de zoom ou d'éloignement, comme dans (7) et (8) – nous mettons entre parenthèses le type d'unité calendaire (D pour date ou H pour horaire) en jeu dans la proposition. Dans les deux cas, le « lendemain » correspond au 9 juin, bien que dans (8) le calcul de l'horaire dépasse largement les 24 heures que compte le 8 juin.

(7) « **Le** 8 juin(D), la machine à café a explosé ***à*** 8h(H). L'appartement a pris feu 10 minutes(H) **plus tard**. **Le** lendemain(D), j'habitais chez mon voisin. »

(8) « **Le** 8 juin(D), la fête battait son plein **à** 23h(H) ; 3h(H) **plus tard**, nous étions couchés. **Le** lendemain(D) fut difficile. »

Actuellement, ces types de problèmes sont correctement pris en charge par les systèmes, même si certaines critiques peuvent être formulées à l'égard de la manière dont sont modélisées certaines opérations relatives à l'ancrage calendaire. (Voir par exemple [BAT & SCH 06], utilisant [SCH 02], pour une proposition de modélisation alternative).

L'annotation des relations temporelles entre propositions dans un texte

Il est d'usage de distinguer clairement la succession linéaire des propositions d'un texte de l'ordonnancement temporel des situations qu'elles dénotent, cet

ordonnancement pouvant prendre des configurations très diverses. S'il arrive qu'il y ait isomorphisme exact entre les deux structures, c'est cependant loin d'être toujours le cas, qu'il s'agisse par exemple de retours en arrière (analepses) ou bien de relations de recouvrement partiel ou total entre situations. Le texte (9) montre bien cette absence d'isomorphisme. À partir des formes verbales conjuguées (*a donné, avoir passé* et *a attendu*) et des connecteurs (*après* et *puis*), le lecteur comprend que le ministre a d'abord passé des nuits à rédiger son projet, puis qu'il l'a donné à son porte-parole, et qu'il a ensuite attendu son avion. Dans ce type d'exemple, l'exploitation des seuls indices linguistiques permet le calcul de l'ordonnancement temporel entre les propositions. C'est aussi le cas pour un texte un peu plus long tel que le texte (10).

(9) *Le ministre a donné à son porte-parole le document rouge après avoir passé des nuits à le rédiger, puis il a attendu le départ de son avion pour Amsterdam.*

(10) *Pendant que je montais les marches du perron **(P1)** en cherchant mes clés **(P2)**, un souffle de vent bienvenu balaya la rue **(P3)**, chassa à travers la chaussée des papiers de bonbon **(P4)**, envoya des cannettes vides s'entrechoquer comme les cloches d'un carillon **(P5)**. Un vieux journal glissa sur le trottoir **(P6)** en chuchotant comme une maîtresse défunte **(P7)**.*[12]

Cependant, dans ce dernier, et contrairement au texte (9), les relations temporelles entre propositions ne sont pas toutes explicitement marquées à l'aide de connecteurs tels que *puis* ou *après* par exemple. Leur mise en évidence nécessite le recours à des règles d'ordonnancement temporel qui rendent compte du fait que les valeurs aspectuelles des propositions codent certaines instructions relatives à l'organisation temporelle des propositions entre elles, et ceci de manière indépendante des autres marqueurs explicites que sont les connecteurs temporels ou les locutions adverbiales (parmi elles, les expressions de type calendaire qui peuvent apparaître en position adverbiale). Prises deux à deux, l'ordre temporel entre les propositions P3, P4, P5 et P6 du texte (10) est une succession ; il est déduit de l'ordre linéaire du texte mais aussi de l'utilisation du passé simple qui renvoie ici à une valeur d'évènement [DES 95?] dans chacune des propositions. Changer le temps du verbe dans la proposition P4 et utiliser le plus que parfait par exemple conduirait à une valeur aspectuelle différente, ce qui amènerait à inverser l'ordre temporel de P4 et P3. En ce qui concerne l'interprétation du gérondif dans P2 et P7, une analyse possible est de considérer que la proposition le contenant « prend » la valeur aspectuelle de la proposition précédente, ce qui conduit à considérer que P2 et P7 sont concomitantes [CHA 05?].

Pour l'analyse de ce type de textes (c'est-à-dire essentiellement des narrations), plusieurs démarches d'ordre calculatoire ont été proposées visant toutes à tenter

12. Connolly J., *Tout ce qui meurt*, Presses de la Cité, p. 142. Cet exemple de texte est tiré du corpus d'analyse de [CHA 05].

d'expliciter des règles d'ordonnancement temporel des propositions (c'est-à-dire en la présence ou en l'absence de marqueurs explicites) – *cf.* par exemple [WEB 88, SON & COH 91, HWA & SCH 92, HIT & al. 95]. La tâche de constitution de ressources linguistiques dans ce domaine reste lourde et de nombreux travaux demeurent encore à réaliser pour prétendre à une certaine qualité des résultats à attendre des systèmes automatiques.

Un autre phénomène à prendre en compte dans les systèmes (ce n'est actuellement pas le cas) concerne les différentes prises en charge énonciatives des discours directs et indirects. Le texte (11) contient ainsi une première forme verbale *a indiqué* qui ne situe pas la situation correspondante dans le même paradigme temporel que celui qui correspond à la proposition contenant la forme verbale *attendait* ; cela conduit dans un premier temps à situer temporellement et de façon indépendante d'une part les propositions P1 et P5 et d'autre part P2, P3, et P4, pour, dans un deuxième temps, pouvoir ordonner l'ensemble des propositions du texte. Ce type de conceptualisation[13], qui explicite différents niveaux de structuration temporels articulés entre eux, permet de prendre en charge correctement des calculs d'ancrage de certaines expressions calendaires, comme dans (12a) et (12b) par exemple, où « *demain* » renvoie respectivement au « demain » d'un énonciateur second et au « demain » de l'énonciateur principal.

(11) *NTT DoCoMo, premier opérateur japonais de téléphonie mobile, a indiqué mardi **(P1)** qu'il attendait sur l'exercice en cours une croissance du nombre de ses abonnés inférieure de 30% à celle de l'exercice précédent **(P2)**. "Je pense **(P3)** que la croissance nette du nombre de nos abonnés atteindra cette année environ 70% du chiffre de l'an dernier **(P4)**", a déclaré mardi à Reuters Keiji Tachikawa, directeur général de DoCoMo **(P5)**.*[14]

(12) (a) « Hier, il a dit « je viendrai <u>demain</u> »

(b) « Il a dit qu'il viendrait <u>demain</u> »

Il reste à souligner que de manière générale les travaux sur la temporalité dans les textes se trouvent tous confrontés à des difficultés méthodologiques qui concernent deux questions inhérentes à l'activité d'annotation sémantique : quelles étiquettes retenir ? quelles unités textuelles (ou relations entre unités textuelles) annoter ? A ce titre, il convient de remarquer que toutes les approches achoppent en ce qui concerne la caractérisation des relations temporelles en elles-mêmes[15] – à

[13] Voir par exemple [BAT & al. 06] pour une proposition de prise en compte des phénomènes liés à l'analyse des discours directs et indirects dans une perspective d'automatisation, à la suite de [DES 95] et [CHA & al. 05].

[14] Extrait du corpus Mobile News propriété de l'entreprise Mondeca**.**

[15] Voir par exemple [BAT & al. 04] pour une présentation du problème de la caractérisation des types de relations temporelles.

savoir leur nature et leur nombre –, difficulté soulignée dans le cadre de la mise en place de protocoles d'évaluation (voir par exemple [MUL & TAN 04]).

4.3.2.2. Les systèmes de résumé multidocuments

La réalisation de systèmes de résumé multidocuments [MAN & BLO 97, CAR & GOL 98, BAR & al. 02, CHA & al. 04] a rendu indispensable la prise en compte de la temporalité. En effet, alors que les systèmes de résumé monodocument peuvent reproduire l'ordre des phrases tel qu'il apparaît dans le texte source, cette solution ne peut plus être mise en œuvre dans les systèmes de résumé multidocuments. Comme ceux-ci doivent construire un résumé à partir de sources différentes, il leur faut d'une part éliminer les informations redondantes et d'autre part ordonner temporellement les phrases. Par exemple, pour une biographie, le texte produit doit respecter l'ordre chronologique, de la naissance à l'époque présente. L'évaluation menée par [BAR & al. 02] démontre l'importance de l'ordonnancement des phrases pour une bonne compréhension du résumé par un lecteur humain. L'analyse de la temporalité est généralement restreinte au repérage de marqueurs explicites qui réfèrent à des évènements calendaires et à l'exploitation de la date d'émission du texte source. Néanmoins, il n'est pas toujours possible de considérer que la date d'émission du texte est valide pour toutes les phrases qui le composent. Dans leur article, [BAR & al. 02] donnent l'exemple de la phase suivante (traduit par nos soins) qui, comme le soulignent les auteurs, ne contient aucune marque temporelle explicite ; de plus, l'assertion qu'elle contient ne peut pas être considérée comme vraie uniquement à partir de la date d'émission du texte.

(13) « *La vaste région du Xingjiang, pratiquement inhabitée et déserte, possède de nombreuses installations nucléaires militaires ainsi que des installations minières civiles.* »

L'autre problème d'ordre temporel concerne l'ordonnancement des phrases qui vont constituer le résumé. Le calcul de l'ordonnancement s'appuie sur l'horodatage. Chaque texte est horodaté avec la date d'émission du texte et des blocs de phrases (constitutifs selon [BAR & al. 02] de thèmes) qui le composent sont considérés comme ordonnés suivant leur ordre d'apparition dans le texte. La réunion de ces ordonnancements issus des différents documents qui traitent d'un même évènement peut conduire à des conflits puisque l'ordre de ces blocs (ou thèmes) peut être différent selon les textes. En conséquence, les thèmes sont ordonnés par paires et un poids, fonction de leur fréquence d'apparition dans cet ordre, leur est affecté. L'union des ordonnancements constitue un graphe. Un algorithme est ensuite appliqué pour rechercher un ordre optimal dans ce graphe (voir [BAR & al. 02] pour une description détaillée). Le choix de l'algorithme est crucial pour obtenir des résumés de bonne qualité. Ainsi, la figue 4.5 présente le résultat où l'algorithme appliqué consiste à mettre en œuvre la stratégie du vote majoritaire qui privilégie

l'ordre des couples de blocs de phrases les plus fréquemment rencontrés dans les textes. Comme on peut le constater à la lecture du résumé obtenu, la cohérence de l'ensemble est mauvaise. Les recherches actuelles, dont relève [BAR & al. 02], se focalisent sur l'élaboration d'algorithmes qui prennent en compte la segmentation thématique du texte, laquelle consiste à considérer des blocs de phrases et non pas des phrases isolées ainsi que les liens qui existent entre ces différents segments thématiques. La figure 4.6 illustre le résultat obtenu avec un algorithme de ce type.

Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania.
Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large.
President Clinton said, "The intended victims of this vicious crime stood for everything that is right about our country and the world".
U.S. federal prosecutors have charged 17 people in the bombings.
Albright said that the mourning continues.
Kenyans are observing a national day of mourning in honor of the 215 people who died there.

Figure 4.5. Un exemple de résumé multi-document mal ordonné [BAR & al. 02][16]

Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania. Kenyans are observing a national day of mourning in honor of the 215 people who died there.
Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large.
U.S. federal prosecutors have charged 17 people in the bombings.
President Clinton said, "The intended victims of this vicious crime stood for everything that is right about our country and the world". Albright said that the mourning continues.

Figure 4.6. Un exemple de résumé multi-document correctement ordonné [BAR & al. 02][17]

[16] Des milliers de personnes ont assisté à la cérémonie à Nairobi qui commémorait le premier anniversaire des attentats à l'explosif contre l'Ambassade des Etats-Unis au Kenya et en Tanzanie. Le dissident Saoudien Osama bin Laden, accusé d'être le cerveau de l'attentat, ainsi que neuf autres personnes sont encore en fuite. Le Président Clinton a déclaré « Les victimes visées de ce crime brutal représentent l'ensemble des valeurs de notre pays et du monde ». Les procureurs de l'Etat ont inculpé 17 personnes pour cet attentat. Albright a déclaré que le deuil continuuait. Les Keynians observent un jour de deuil national en mémoire des 215 personnes qui sont mortes dans cet attentat.

Ces trois exemples démontrent l'importance de la prise en compte des marques aspecto-temporelles, de la sémantique verbale, des informations causales, etc. dans le calcul de l'ordonnancement temporel. Par exemple, dans la cinquième phrase du texte (13), la prise en compte du sens du verbe *posséder* et de la marque aspecto-temporelle associée au morphème du présent permettrait de caractériser l'intervalle sur lequel la phrase est validée temporellement. Pour Mani [MAN 04], les informations nécessaires à ce type de calcul ne sont actuellement pas disponibles et nécessiteraient la construction de ressources et de modèles proches de ceux qui ont mobilisé les chercheurs en IA dans les années 70. Nous pensons pour notre part que ce type de ressources n'est pas en dehors de notre portée puisque, contrairement aux premiers modèles de compréhension proposés en IA, les modèles actuellement développés privilégient le recours à des ressources essentiellement linguistiques et non pas liées à des domaines particuliers.

4.3.3. La navigation interactive et le résumé

Comme nous l'avons montré précédemment, dans le domaine du résumé automatique un grand nombre de travaux de recherche se sont focalisés sur les outils qui permettent de sélectionner des parties saillantes dans un texte. Une des caractéristiques communes à ces différentes approches est le peu d'importance apporté aux modes de représentation visuelle du résultat : un fragment textuel qui « résume » un texte. En fait, le résultat produit reste imprégné des contraintes imposées par une conception marquée par la prégnance technologique qui assimile texte et texte imprimé [VAN 99]. Certains travaux, notamment [HEA 99, JAC & JAR 02], ont proposé de développer des interfaces qui transforment les représentations visuelles, mais ces transformations ne s'appuient pas sur les marques linguistiques, c'est-à-dire plus précisément, sur les structures discursives construites par l'auteur.

Les logiciels de traitement textuel actuels offrent des possibilités extrêmement puissantes puisqu'ils disposent, en arrière plan, de la représentation structurelle du texte décrite avec les normes proposées par XML. Cette représentation structurelle peut notamment être annotée par des résultats issus de traitements linguistiques (avec repérage d'entités nommées, de structures discursives, de relations

[17] Des milliers de personnes ont assisté à la cérémonie à Nairobi qui commémorait le premier anniversaire des attentats à l'explosif contre l'Ambassade des Etats-Unis au Kenya et en Tanzanie. Les Keynians observent un jour de deuil national en mémoire des 215 personnes qui sont mortes dans cet attentat. Le dissident Saoudien Osama bin Laden, accusé d'être le cerveau de l'attentat, ainsi que neuf autres personnes sont encore en fuite. Les procureurs de l'Etat ont inculpé 17 personnes pour cet attentat. Le Président Clinton a déclaré « Les victimes visées de ce crime brutal représentent l'ensemble des valeurs de notre pays et du monde ». Albright a déclaré que le deuil continuait.

sémantiques, etc.). L'exploitation de cette structure annotée par des logiciels de présentation permet ainsi d'envisager de nouveaux modes de lecture sur les « *écrits d'écran* » [SOU 98]. Le fait qu'un texte soit maintenant numérisé et qu'il soit présenté au lecteur sur un écran peut être considéré comme une nouvelle mutation qui place le lecteur devant de nouvelles possibilités qui restent à explorer : « *Le texte [...] offre en effet une richesse sémiotique particulière, qui fournit de multiples objets d'interprétation et de multiples pistes d'actions [...] les lecteurs n'ont pas la même démarche envers l'objet ni la même définition de cet objet, ils ne « voient » pas la même chose* » [SOU & al. 03].

Des recherches en cours [MIN 02, COU & al. 04, COU & MIN 04] se proposent d'exploiter ces possibilités en considérant le résumé automatique et la fouille de textes comme un processus de recherche d'information guidé par les besoins spécifiques de l'utilisateur et par la prise en compte contextualisée des structures discursives dans lesquelles l'information recherchée est mise en discours. Des outils de visualisation et de navigation textuelle qui exploitent ces structures vont guider le lecteur dans sa fouille textuelle. Ainsi plutôt que de proposer un fragment textuel indépendant du texte, le résumé est conceptualisé comme un parcours de lecture [MIN 05] propre aux attentes du lecteur. Ainsi il y a potentiellement pour un même texte une multiplicité de parcours de navigation que l'on peut, en partie, comparer au cheminement déambulatoire dans les hyperdocuments proposé par [GER 02]. Ces principes de navigation se distinguent néanmoins de la navigation hypertextuelle puisque les opérations de navigation vont s'appuyer sur les marques sémiotiques et linguistiques du texte. La navigation proposée n'est donc pas guidée par l'auteur du texte comme dans le cas de la navigation hypertextuelle où les hyperliens sont placés par cet auteur.

4.3.3.1. Modélisation des connaissances de navigation

La navigation dans un texte [COU & MIN 04] est conceptualisée comme une opération qui relie une unité textuelle source, un fragment, une phrase, un syntagme, avec une unité textuelle cible. Ainsi dans la plate-forme *NaviTexte* développée pour mettre en œuvre ces hypothèses, une opération de navigation est définie par un *libellé* (ce qui sera visible par le lecteur), une *source*, une *cible*, une ou plusieurs *conditions*, et un *empan*. L'exécution d'une opération est ainsi soumise à des conditions qui contraignent les attributs (type d'unité textuelle, annotations morpho-syntaxiques, sémantiques ou pragmatiques, etc.) des unités textuelles sources et cibles considérées. L'*empan* de texte peut être spécifié, sous la forme d'un type d'unité textuelle, section, paragraphe, segment textuel, etc. de manière à restreindre l'espace de recherche des unités textuelles cibles. Chaque opération de navigation est typée avec une valeur qui appartient à l'ensemble {*Premier, Dernier, Suivant[i], Précédent[i]*}. Ces valeurs spécifient d'une part l'orientation, c'est-à-dire le sens (avant ou après l'unité textuelle source) dans lequel doit être effectué la recherche de

l'unité textuelle cible, et d'autre part le référentiel, absolu (*Premier, Dernier),* ou relatif (*Suivant[i], Précédent[i]*), par rapport à l'unité textuelle source. Dans le cas d'un référencement relatif, l'index *i* permet de spécifier le rang de la cible recherchée. Par exemple, l'expression « Type d'opération= Suivant[3] » s'interprète comme la recherche de la troisième unité textuelle située après l'unité textuelle source et qui satisfait aux conditions spécifiées. Les conditions expriment des contraintes sur les valeurs des attributs des unités textuelles source ou cible. Ces conditions simples peuvent être combinées entre elles avec les opérateurs logiques et relationnels. Un ensemble d'opérations de navigation, regroupées dans un module de navigation, modélise ainsi un parcours potentiel dans un texte.

Ces opérations de navigation sont proposées, interactivement, au lecteur sous forme de menu contextuel (cf. fig. 4.3). L'écran de visualisation et de navigation de la plate-forme offre différentes fonctionnalités qui sont le résultat de l'interprétation dynamique du contenu des modules décrits précédemment. Par conséquent, chaque module offre des possibilités différentes, ce qui rend possible des lectures adaptées à des lecteurs dont les compétences ou les besoins diffèrent. L'exemple de la figure 4.3 illustre l'exécution, déclenchée par un geste de lecture (une action sur un des boutons de la souris), de l'opération de navigation associée à l'unité textuelle, « Les publications récentes… », annotée dans le texte comme « Annonce Thématique ». Quatre pistes de lecture sont ainsi offertes au lecteur : « Lire Thématique », « Lire Récapitulation », « Lire Argumentation » et « Lire Conclusion » ; c'est le choix de l'opération de navigation par le lecteur qui conditionne le déplacement de la fenêtre de lecture sur le segment textuel pertinent.

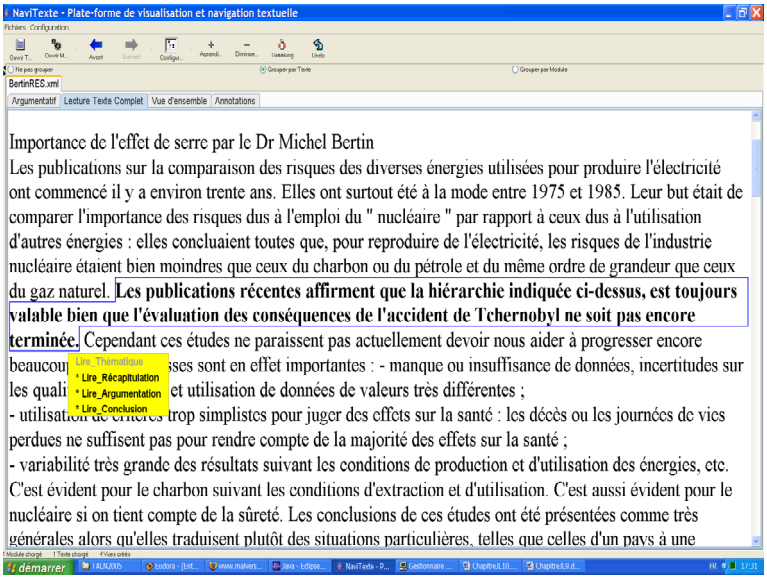

Figure 4.3. Exemple d'opération interactive de navigation dans la plate-forme *NaviTexte* [COU & MIN 04]

4.3.3.2. Compréhension de texte et navigation

Ce procédé, par lequel un lecteur navigue dans un texte en suivant ses différentes pistes de lecture, peut être utilisé pour la compréhension de textes. Ainsi, dans le projet de recherche *NaviLire* [COU & al. 05], qui s'appuie d'une part sur les travaux en linguistique textuelle menés par Lundquist [LUN 80, 99] et la plate-forme *NaviTexte*, c'est la navigation entre les pistes de cohérence – fondées par exemple, sur la référence, sur la prédication, les connecteurs, les marques temporelles, etc. – qui va être exploitée pour faire comprendre à un étudiant non francophone les organisations discursives complexes et leurs réalisations langagières dans les textes.

Par exemple, une piste de cohérence consiste à identifier les référents discursifs d'un texte et à établir les relations correctes entre les syntagmes nominaux qui y réfèrent ; en d'autres termes, de décider s'il s'agit d'une relation de coréférence ou d'une disjonction référentielle. Cette compétence cognitive est primordiale pour arriver à établir une représentation mentale cohérente et correcte du texte en question, condition de toute compréhension par les textes [KIN 98, p. 307]. Une autre piste de cohérence consiste à identifier « où veut en venir l'émetteur » du texte. Cette orientation – expressive, argumentative, etc. – qualifiée de « programme d'interprétation » par [LUN 90], qui fonctionne d'abord du *général au particulier*, et

ensuite *du spécifique au générique*, permet d'identifier des marques dans le texte qui « vont dans le même sens » [KIN 98, p. 50]. Cette identification de l'orientation, apportée en particulier par les prédications, est primordiale pour un déchiffrage correct de la cohérence sémantique et pragmatique du texte. Finalement, les connecteurs soulignent les relations rhétoriques à établir entre des propositions ou autres séquences du texte, ce qui contribue, évidemment, de manière essentielle à établir les relations nécessaires pour construire la représentation correcte du texte.

Comme l'illustre cet exemple, la compréhension d'un texte peut être en partie assimilée à l'apprentissage de la navigation dans celui-ci, en identifiant les marques, référentielles, argumentatives, etc., placées par l'auteur, et par conséquent la modélisation de cette navigation constitue, de notre point de vue, une autre manière de modéliser la compréhension.

4.4. Conclusion

Le problème du statut de la compréhension s'est posé dès les premières recherches sur le résumé automatique. Le développement des recherches a abouti rapidement à un clivage entre les équipes posant la notion de modèle et de représentations symboliques comme principe premier et celles proposant, allant même parfois jusqu'à le revendiquer, des approches strictement calculatoires. L'expérience a montré, comme c'est d'ailleurs cas dans le domaine du jeu d'échecs, que les secondes donnaient des résultats opérationnelles, robustes, et industrialisables.

On peut tout d'abord néanmoins noter que les objectifs des deux démarches diffèrent dans leurs ambitions. Les méthodes fondées sur les représentations ont des visées cognitives et cherchent à simuler des comportements humains. Ensuite, les méthodes fondées sur l'analyse des marques de surface, en s'attaquant à des problèmes de plus en plus complexes, comme les systèmes de résumés multi-documents ou les systèmes de résumés adaptés à différents lecteurs, sont amenées à intégrer des algorithmes complexes qui s'appuient essentiellement sur des ressources linguistiques. Cette intégration nécessite une réflexion sur l'architecture des systèmes et sur l'interopérabilité des différents modules. On voit ainsi que ces méthodes sont conduites à se poser des problèmes de représentation et qu'elles rejoignent ainsi, mais guidées par d'autres motivations, les premières approches. Enfin, la mutation technologique introduite par les « écrits d'écran » pourrait bien transformer la nature même du problème au coeur du résumé automatique, ou tout du moins à le déplacer vers la notion de parcours de lecture.

4.5. Bibliographie

[ALT 91] ALTERMAN R., « Understanding and summarisation», *Artificial Intelligence Review*, 5, p. 239-254, 1991.

[ALT & BOO 92] ALTERMAN R., BOOKMAN L. A.., « Reasoning About a Semantic Memory Encoding of the Connectivity of Events», *Cognitive Science*, 16, p. 205-232, 1992.

[BAR & al. 02] BARZILAY R., ELHADAD N., MCKEOWN K. R., « Inferring stategies for sentence ordering in multi-document news summarization», *Journal of Artificial Intelligence Research,* (17), p. 35-55*,* 2002.

[BAT & al. 04] BATTISTELLI D., MINEL J.-L., PICARD E., SCHWER S., «Temporalité linguistique et S-Langages», *TALN 04*, Fès, Maroc, 2004.

[BAT & al 06] BATTISTELLI D., CHAGNOUX M.,. DESCLES J.-P., «Référentiels et ordonnancements temporels dans les textes», *Cahiers Chronos*, à paraître en 2006.

[BAT & SCH 06] BATTISTELLI D., SCHWER S., «Modélisation des expressions calendaires dans les textes», soumis à *RFIA 05*, Tours, janvier 2006.

[BER & al. 96] BERRI J., CARTIER E., DESCLES J.-P., JACKIEWICZ A., MINEL J.-L., « SAFIR, système automatique de filtrage de textes », *Actes du colloque TALN'96*, p. 140-149, Marseille, 1996.

[CAR & al. 98] CARBONNEL J., GOLDSTEIN J., « The use of MMR, diversity-based reranking for reordering documents and producing summaries », *SIGIR* , p. 335-346, 1998.

[CHA & al. 04] CHAAR S.L., FERRET O., FLUHR J., « Filtrage pour la construction de résumés multidocuments guidée par un profil», *TAL.* Vol 45/1, Editions Hermès, p. 65-95 , 2004.

[CHA 2005] CHAGNOUX M., *Temporalité et aspectualité dans les textes français : modélisation sémantico-cognitive et traitement informatique*, Thèse de doctorat, Université Paris-Sorbonne, Paris, en cours.

[CHA & al. 05] CHAGNOUX M., DESCLES J.-P., MAIRE-REPPERT D., «Comment est structuré dynamiquement un texte ? », *à paraître*.

[CHA 99] CHAROLLES M., « Phrase, texte, discours », *Langue Française*, n° 121, p. 76-116, 1999.

[CRI & COU 04] CRISPINO G., COUTO, J., « Construction automatique de résumés. Une approche dynamique », *TAL.* Vol 45/1, Editions Hermès, p. 95-120 , 2004.

[COI & al. 96] COIRIER P., GAONACH D., PASSERAULT J.M., *Psycholinguistique textuelle,* Armand Colin, Paris, 1996.

[COU & al. 04] COUTO J., FERRET O., GRAU B., HERNANDEZ N., JACKIEWICZ A., MINEL J.-L., PORHIEL S., « RÉGAL, un système pour la visualisation sélective de documents. », Revue d'Intelligence Artificielle, Hermès, p. 481-514, 2004.

[COU & MIN 04] COUTO J., MINEL J.-L., « Outils dynamiques de fouilles textuelles », *Actes de RIAO,* Avignon, p. 420-430, 2004.

[COU & al. 05] COUTO J., LUNDQUIST L., MINEL J.-L., « Navigation interactive pour l'apprentissage en linguistique textuelle », *Actes de EIAH,* Montpellier, p. 45-56 , 2005.

[CUM & al. 88] CUMMINS D., KINTSCH W., REUSSER K., WEIMER R. « The role of understanding in solving word problems », *Cognitive Psycholoy*, 20, p. 505-438, 1988.

[DEE & al. 90] DEERWESTER S. , DUMAIS S. T., FURNAS G. W., LANDAUEUR T.K., HARSMAN R., « Indexing by Latent Semantic Analysis », *Journal of the American Society of Information Science*, *41*, p. 391-407, 1990.

[DEN 84] DENHIERE G., *Il était une fois … Compréhension et souvenir de récits*, Lille, Presses Universitaires de Lille.

[DEN 93] DENIS M., «Pour les représentations», *Modèles et concepts pour la science cognitive, Hommage à Jean-François Le Ny*, textes réunis par M. Denis et G. Sabah, Presses Universitaires de Grenoble, Coll. Sciences et Technologies de la Connaissance, p. 95-106.

[DES & al. 91] DESCLES J.-P., JOUIS C., OH H-G., MAIRE REPPERT D., « Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte », In *Knowledge modeling and expertise transfer*, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam, p. 371-400, 1991.

[DES 95] DESCLES J.-P., «Les référentiels temporels pour le temps linguistique», *Modèles linguistiques*, XVI (2), p. 9-36, 1995.

[DES & MIN 05] DESCLES J.-P., MINEL J.-L**.,** « Interpréter par exploration contextuelle », in *Interpréter en Contexte*, sous la direction de Francis Corblin, Paris, Editions Hermès, p. 305-328, 2005.

[DIJ & KIN 83] VAN DIJK T. A., KINTSCH W., *Strategies of discourse comprehension*, New York, Academic press, 1983.

[EDM 69] EDMUNDSON, H.P., « New Methods in Automatic Extracting », *Journal of the Association for Comptuting Machinery,* 16(2), p. 264-285, 1969.

[END 98] ENDRES-NIGGEMEYR B., *Summarizing Information*, Berlin, Springer, 1998.

[FER & al. 03] FERRO L., GERBER L., MANI I., SUNDHEIM B., WILSON G., TIDES 2003 Standard for the Annotation of Temporal Expressions, http ://www.mitre.org/work/techpapers/ techpapers04/ferrotides/.

[FIL & HOV 01] FILATOVA E., HOVY E., «Assigning Time-Stamps to Event-Clauses», *Workshop on Temporal and Spatial Information*, ACL 01, p. 88-95, 2001.

[FIL 68] FILLMORE C., « The case for case » in *Universals in linguistic theory,* B. Harms (éd), Holt, Rinehart and Winston, Chicago, p. 1-90, 1968.

[FRA 98] FRAENKEL J.-J., «Référence, référenciation et valeurs référentielles», *Sémiotiques*, 15, INALF, Didier-Érudition, p. 61-84, 1998.

[GAR & al. 82] GARNHAM A., OAKHILL J.V., JOHNSON-LAIRD, P. N., «Referential continuity and the coherence of discourse», *Cognition*, 11, p. 29-46.

[GAR & OAK 93] GARNHAM A., OAKHILL J.V., «Modèles mentaux et compréhension du langage», *Les modèles mentaux - approche cognitive des représentations*, coordonné par M.-F. Ehrlich, H. Tardieu, M. Cavazza, Introduction de P.N. Johnson-Laird, Masson, Paris, p. 23-44, 1993.

[GER 02] GERY M., « Un modèle d'hyperdocument en contexte pour la recherche d'information structurée sur le Web », *Revue des Sciences et Technologies de l'Information,* 7/2002*,* Hermès, Paris : 11-44, 2002.

[HAL & HAS 76] HALLIDAY M.A.K., HASAN R., *Cohesion in English*, London, Longman, 1976.

[HEA 99] HEARST M., «User Interfaces and Visualization. », In *Modern Information Retrieval* R. Baeta-Yates, B. Ribeiro-Neto (eds), Addison-Wesley, p. 257-322, 1999.

[HIT & al. 95] HITZEMAN J., MOENS M., GROVER C., «Algorithms for Analyzing the Temporal Structure of Discourse», *EACL 95*, p. 253-260, 1995

[HWA & al. 92] HWANG C.H., LENHART K.S., «Tense Trees as the 'Fine Structure' of Discourse», ACL 92, p. 232-240, 1992.

[JAC & JAR 02] JACQUEMIN C., JARDINO M., « *Multi-dimensional and Multi-scale Visualizer of Large XML Documents*», *Proceedings of EUROGRAPHICS,* Saarbrücken, Germany, 2002.

[JOH 83] JOHNSON-LAIRD P.N., *Mental models*, Cambridge, MA: Harvard University Press 1983.

[KIN 74] KINTSCH W., *The representation of meaning in memory*, Hillsdale, NJ, Erlbaum, 1974.

[KIN 78] KINTSCH W., VAN DIJK T. A., « Toward a model of text comprehension and production », *Psychological review*, 85, p. 363-394, 1978.

[KIN 98] KINTSCH W., *Comprehension. À Paradigm for Cognition*, Cambridge, Cambridge University Press, 1998/2003.

[LAS 93] LASSEGUE J., «Emploi et origine de la notion de représentation en sciences cognitives», *Intellectica,* 93/2, 17, p. 199-212.

[LER & al. 94] LEROUX D., MINEL J.-L., BERRI J., « SERAPHIN project », *First European Conference of Cognitive Science in Industry*, Luxembourg, p. 275-283, 1994.

[LUN 80] LUNDQUIST L., *La cohérence textuelle, syntaxe, sémantique, pragmatique*, Copenhagen, Nordisk Forlag, 1980.

[LUN 90] LUNDQUIST L., « Conditions de Production et Programmation Argumentative », *Verbum*, vol. 13, nr. 4, p. 237-264, 1990.

[LUN 99] LUNDQUIST L., « Le factum textus. Fait de grammaire, fait de linguistique ou fait de cognition? » *Langue française*, p. 56-75, 1999.

[LUN 05] LUNDQUIST L., « Noms, verbes et anaphores (in)fidèles. Pourquoi les Danois sont plus fidèles que les Français. », *Langue française*, 2005.

[MAN & BLO 97] MANI I., « Multi-document summarization by graph search and matching » *Fifteenth National Conference on Artificial Intelligence (AAAI 97)*, Providence, Rhode Island, p. 622-628, 1997.

[MAN & WIL 00] MANI I., WILSON G., «Robust Temporal Processing of News», *ACL 00*, p. 69-76, 2000.

[MAN 01] MANI I., *Automatic Summarization*, John Benjamins Publishing Company, Amsterdam, 2001.

[MAN 03] MANI I., « Recent Developments in Temporal Information Extraction», *RANLP 03*, 2003.

[MAN 04] MANI I., « Narrative Summarization », *TAL*, 45/1, Éditions Hermès, p. 15-38, 2004.

[MAR 97] MARCU D., « From discourse structures to text summaries», in *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, p. 82-88, 1997.

[MIN& DES 00] MINEL J.-L., DESCLES J.-P., *Résumé Automatique et Filtrage des textes*, in *Ingénierie des langues*, (sous la direction de J-M. Pierrel) Paris, Hermès, p. 253-270, 2000.

[MIN 01] MINEL J.-L., CARTIER E., CRISPINO G., DESCLES J.-P., BEN HAZEZ S., JACKIEWICZ. A., « Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText », *Technique et Science Informatiques,* 3, Paris, 2001.

[MIN 02] MINEL J.-L., *Filtrage sémantique de textes. Du résumé automatique à la fouille des textes,* Paris, Hermès, Paris, 2002.

[MIN 04] MINEL J.-L., « Résumé automatique, bilan et perspectives ». *TAL*, Vol 45/1, Editions Hermès, p. 7-14, 2004.

[MIN 05] MINEL J.-L., « Réflexions autour de l'identification, la modélisation et la visualisation de certaines organisations textuelles *», L'unité Texte* sous la direction de D. Klingler et S. Porhiel, Editions Pleyben, p. 231-250, 2005.

[MUC 97] Message Understanding Conferences, 1997 http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[MUL & TAN 04] MULLER P., TANNIER X., « Une méthode pour l'annotation de relations temporelles dans des textes et son évaluation», *TALN 04*, Fès, Maroc, 2004.

[PIT 85] PITRAT J., *Textes, ordinateurs et compréhension*, Eyrolles, Paris, 1985.

[PAI 90] PAICE C. D., «Constructing literature abstracts by computer : techniques and prospects», *Information processing management*, 26 (1), p. 171-186, 1990.

[RAT & al. 61] RATH G. J., RESNICK A., SAVAGE T., «The formation of abstracts by the selection of sentences», *American Documentation*, 12(2), p.139-143, 1961.

[SAB 88] SABAH G., *Intelligence artificielle et le langage*, Paris, Hermès, 1988.

[SAB 78] SABAH G., *Contribution à la compréhension effective d'un récit*, Thèse de doctorat d'état, Université Pierre et Marie Curie, Paris, 1978.

[SAG 02] SAGGION H., « SumUM*»,* in *Filtrage sémantique de textes. Du résumé automatique à la fouille des textes,* Paris, Hermès, 2002.

[SAG & LAP 02] SAGGION H., LAPALME G., « Generating Indicative-Informative Summaries with SumUM*», Computational Linguistics*. December 2002, 28(4), p. 497-526, 2002.

[SCH 75] SCHANK R., «The structure of episodes in memory», *Representation and understanding : Studies in cognitive science*, Bobrow D. & Collins A. (eds.), New York, Academic press, 1975.

[SCH & ABE 77] SCHANK R., ABELSON R., *Scripts, plans goals, and understanding*, Hillsdale N. J., Erlbaum, 1977.

[SCH 02] SCHWER S., «Formalizing Calendars with the Category of Ordinals», *Applied Intelligence*, 17 (3), p. 275–295, 2002.

[SET & GAI 00] SETZER A., GAIZAUSKAS G., «Annotating Events and Temporal Information in Newswire Texts», *LREC 00*, p. 64-66, 2000.

[SON & COH 91] SONH F., COHEN R., «Tense Interpretation in the Context of Narrative», *AAAI 91*, p. 131-136, 1991.

[SOU 98] SOUCHIER E., « L'image du texte. Pour une théorie de l'énonciation éditoriale. », *Les Cahiers de médiologie*, 6, Paris, 1998.

[SOU & al 03] SOUCHIER E., JEANNERET Y., LE MAREC J., *Lire, écrire, récrire : objets, signes et pratiques des médias informatisés,* Bibliothèque publique d'information, Paris, 2003.

[TER 02] TERQUAS, *Time and Event Recognition for Question Answering Systems*, an ARDA Workshop on Advanced Question Answering Technology, 2002, http ://www.timeml.org/terqas.

[TEU 97] TEUFEL S., MOENS M., «Sentence extraction as a classification task», *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, p. 58-65, 1997.

[TEU & MOE 98] TEUFEL S., MOENS M., «Sentence Extraction and rhetorical classification for flexible abstracts», *AAAI Spring Symposium on Intelligent Text summarization*, Stanford, 1998.

[THO & MAN 88] THOMPSON S. & W. MANN. «Rhetorical structure theory, a framework for the analysis of texts», *IPRA Papers in Pragmatics*, I, 1988, p. 79-105, 1988.

[VAN 99] VANDENDORPE C., *Du papyrus à l'hypertexte*, Paris, Editions la Découverte, 1999.

[VAZ 01] VAZOV N., «A System for Extraction of Temporal Expressions from French Texts», *TALN 01*, Tours, p. 315-324, 2001.

[WEB 88] WEBBER B.L., «Tense as Discourse Anaphor», *Computational Linguistics*, 14(2), p. 61-73, 1988.