Modélisation par contraintes pour la description et l'analyse automatique de la structure du discours

Antoine Widlöcher[1]
Université de Caen

Abstract

We focus on the problem of discourse structure analysis from the point of view of its formal description and its automatic processing. We intend to test the hypothesis that constraint-based approaches could enable various structures and clues of their presence to be taken into account. The CDML formalism which is introduced allows such an approach whose benefit is demontrated through two case studies: the analysis of temporal discourse frame and the analysis of contrast relation.

Keywords : discourse structure analysis, NLP, constraint-based approach.

Résumé

Nous abordons la question de l'analyse de la structure du discours, du point de vue de sa description formelle et de son traitement automatique. Nous envisageons l'hypothèse selon laquelle une approche par contraintes pourrait permettre la prise en charge de structures discursives variées d'une part, et de différents types d'indices de leur manifestation d'autre part. Le formalisme CDML que nous introduisons vise précisément une telle approche, dont nous présentons l'intérêt à travers deux études de cas : l'analyse de la portée des cadres de discours temporels et l'analyse des relations de contraste.

Mots-clés : analyse de la structure du discours, TAL, approche par contraintes.

1. Introduction

Des travaux récents au sein de la communauté TAL révèlent un intérêt croissant pour l'analyse de la structure du discours. Qu'il s'agisse de la décrire, de la formaliser et/ou de procéder à son analyse automatique, ou à son exploration expérimentale sur corpus, l'articulation entre la description linguistique et l'opérationnalisation en TAL des modèles élaborés pose problème. Et ce problème s'avère d'autant plus épineux que l'analyse du discours subsume une diversité d'approches, d'objets d'études et d'indices pourtant difficiles à appréhender dans une théorie unique. Certaines approches envisagent par exemple de formaliser la structure du discours, en privilégiant un niveau de granularité relativement réduit, inter-propositionnel ou inter-phrastique. Mais qu'en est-il alors du traitement automatique de telles organisations discursives ? Est-il possible d'appliquer ces modèles à d'autres échelles ? D'autres

[1] Laboratoire GREYC, CNRS UMR 6072, Université de Caen, awidloch@info.unicaen.fr.

travaux privilégient au contraire certaines tâches opérationnelles telles que la segmentation automatique et opèrent à un niveau de granularité plus élevé. Mais comment donner une description formelle des structures discursives ainsi étudiées ?

Nous envisageons ici les conditions de possibilité d'une approche générique de l'organisation du discours, en cherchant à concilier sa description linguistique *et* son analyse automatique. Après avoir abstrait les différents éléments et mécanismes textuels qu'une telle approche prend pour objets, nous mettrons en évidence la pertinence d'une modélisation du discours en termes de *contraintes* et introduisons CDML (*Constraint-based Discourse Modeling Language*) (Widlöcher 2006), un formalisme dédié à une telle modélisation des structures discursives et à leur exploration automatique sur corpus. Afin d'illustrer les bénéfices d'une telle démarche, nous procédons enfin à deux études de cas en évoquant l'analyse de la portée des cadres temporels et celle des relations rhétoriques de contraste.

2. Analyses du discours

Prenons tout d'abord la mesure de la diversité des approches possibles afin d'envisager la variété des formes que peut prendre la structure du discours. À travers les différentes approches, issues de la linguistique, du TAL et de la linguistique computationnelle, on constate en effet d'importantes variations en termes de visée (modélisation *vs.* traitement), d'objet d'étude, d'échelle ou de grain, d'indices utilisés, ou en termes de *plan* (point de vue macro-syntaxique, sémantique, pragmatique, rhétorique…). La figure 1 illustre les approches que nous évoquons ci-après.

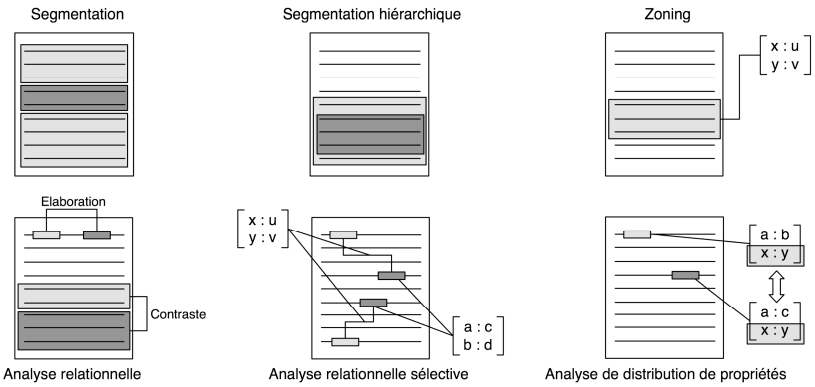

Figure 1. Différentes approches de la structure du discours

Différents travaux visent tout d'abord la *segmentation* du discours, c'est-à-dire la délimitation d'unités textuelles homogènes d'un certain point de vue. *La progression du discours* est alors considérée comme un enchaînement séquentiel de segments contigus présentant un certain degré de cohésion ou de cohérence. On pensera ici en particulier aux travaux apparentés au *Text-Tiling* (Hearst 1994).

D'autres travaux conçoivent l'organisation du discours de manière plus hiérarchique, et visent l'étude de cette *structuration* d'un point de vue par exemple thématique ou

rhétorique. L'analyse thématique proposée par Bilhaut (2006) en constitue un exemple. De telles approches, souvent adossées à une analyse sémantique du contenu, se distinguent des précédentes en ne visant plus un « pavage » complet du discours, mais un repérage de zones aux propriétés sémantiques et structurelles particulières.

Certaines approches visent l'identification de zones textuelles possédant certaines propriétés particulières, par exemple sémantiques ou rhétoriques, éventuellement retenues pour leur pertinence pour une tâche donnée comme l'extraction d'information ou le résumé. L'*argumentative zoning* (Teufel 1999), qui considère le discours d'un point de vue argumentatif et vise l'identification de zones textuelles correspondant à des intentions rhétoriques des auteurs, rentre dans cette catégorie.

Un autre point de vue sur l'organisation textuelle privilégie les *relations* entre les énoncés. Étant donné un ensemble de relations de discours fondamentales, on vise alors la connexion d'une série d'énoncés au sein d'une structure souvent arborescente. Les travaux entrant dans cette catégorie, dont la RST (*Rhetorical Structure Theory*) (Mann et Thompson 1987) constitue un exemplaire représentant, privilégient souvent un niveau inter-propositionnel ou inter-phrastique.

D'autres travaux, également consacrés à l'analyse relationnelle, portent leur attention sur certaines relations particulières. Il ne s'agit plus alors de saisir la trame relationnelle d'un ensemble donné d'énoncés, mais de privilégier l'étude des relations d'un type prédéfini entre des éléments susceptibles de porter ces relations. Des travaux tels que Lappin et Leass (1994) mettent ainsi l'accent sur des relations particulières (anaphoriques) entre des éléments eux aussi spécifiques (reprises anaphoriques).

Enfin, d'autres travaux portent sur l'étude de certains éléments textuels spécifiques ou certaines relations spécifiques, et sur l'analyse de la distribution, et en particulier de la récurrence, dans le texte, de certaines propriétés de ces éléments ou relations. Des travaux sur l'isotopie tels que Tanguy (1997) mettent ainsi l'accent sur la récurrence de propriétés sémantiques et sur les relations entre les éléments porteurs de ces propriétés et constituent, à ce titre, un bon exemple d'approche de ce type.

Ce parcours en largeur de l'univers de l'analyse discursive met en lumière un certain nombre d'éléments à prendre en compte pour la mise en place d'un cadre théorique aussi général que possible.

Certaines approches privilégient la description à un *niveau de granularité* relativement réduit, inter-propositionnel ou inter-phrastique. C'est par exemple le cas des travaux issus de la RST. Visant la description fine de phénomènes discursifs locaux, de telles approches s'articulent assez naturellement avec les analyses plus traditionnelles, par exemple en matière syntaxique, au niveau de la phrase ou de la proposition. D'autres approches, souvent motivées par des objectifs applicatifs telle qu'une segmentation du texte, considèrent celui-ci à un niveau plus élevé : phrases, paragraphes, etc. À ce niveau de grain, les approches *bottom-up* et *compositionnelles* acceptables à d'autres échelles sont souvent proscrites, en particulier pour des raisons combinatoires, au profit de perspectives plus *descendantes*, privilégiant des « faisceaux d'indices », à des

critères formels plus stricts. Pour notre part, nous visons la prise en compte de ces différents niveaux, et la définition d'un cadre théorique et d'un formalisme aussi indépendants que possible de cette granularité. Il nous paraît en effet nécessaire non seulement de pouvoir rendre compte de la diversité des modes d'organisation opérant à différentes échelles, mais aussi de permettre de décrire selon un modèle unifié des phénomènes discursifs de même nature bien qu'opérant à des échelles différentes. L'analyse des relations contrastives proposée ci-après illustrera ces points. La figure 2 évoque la possibilité de fonctions discursives remplies par des éléments de niveau différent.

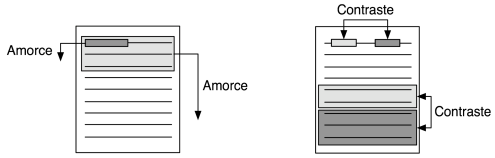

Figure 2. Variabilité du grain d'analyse discursive

Par ailleurs, nous pouvons distinguer les approches *orientées segments* et les approches *orientées relations*. Nous ne restreignons notre acception du discours ni à l'une ni à l'autre de ces perspectives et entendons formaliser sa structuration de ce double point de vue. L'adoption d'une perspective unique ne saurait en effet permettre de rendre compte de la diversité des phénomènes discursifs envisagés. De plus, si certains travaux privilégient effectivement l'une ou l'autre de ces dimensions, celle d'entre elles qui n'est pas retenue comme clef de voûte du modèle n'en demeure pas moins présente, quoique de manière sous-jacente. Quand la RST propose par exemple une modélisation de certaines relations rhétoriques en termes de noyau et de satellites, une notion d'unité textuelle est nécessaire à la compréhension du modèle, alors même que la prépondérance du concept de relation est ici aveuglante. Symétriquement, les approches de type *text-tiling* supposent l'existence d'une relation pouvant assurer la cohésion entre éléments, en mettant pourtant en exergue la notion de segment. Enfin, la prise en compte, au sein d'un même modèle, de cette double dimension nous semble nécessaire pour que ce modèle puisse rendre compte de l'irréductible variabilité du grain en matière d'analyse du discours. Faute de laisser une place essentielle à la notion d'unité ou de segment, au profit de celle de relation, on risque, subrepticement, d'adopter un grain par défaut, telle que la proposition par exemple, pour qualifier les unités reliées. Faute de mettre la notion de relation au cœur du système, au profit de celle d'unité, on risque de perdre de vue le fait qu'une unité tire sa *valeur* discursive du réseau des rapports au sein duquel elle intervient, et de négliger des unités de statut discursif similaire, au profit d'un type d'unité retenu implicitement comme archétype.

D'autre part, une formalisation purement *descriptive* de la structure du discours n'implique pas nécessairement que soient précisées les *conditions* sous lesquelles une structure donnée peut être identifiée comme telle. Au contraire, une approche plus *opérationnelle* mettra l'accent sur ces indices. Mais les besoins de l'implémentation conduisent souvent à privilégier des traitements *ad hoc*, en renonçant fréquemment à

la description formelle (par exemple en termes de grammaire). Nous proposons ici de (re)concilier les approches *descriptives* et *prescriptives* (*i.e. opératoires*) en cherchant à la fois à décrire de manière formelle certains phénomènes discursifs, et à exprimer les conditions d'identification (les *indices*) permettant de reconnaître ces phénomènes, et d'opérationnaliser, en TAL, cette identification.

À cet égard, les différentes approches envisagées laissent présager la remarquable *diversité des indices* utilisables pour la détection de telles structures. Ne considérons ici que certains de ces indices, dont la taille (caractère, mot, phrase...) et dont la nature (morphologique, syntaxique, sémantique...) peuvent varier. L'organisation discursive peut être révélée par des *connecteurs* ou par des *cue-phrases* caractéristiques. D'autres *formes de surface* peuvent être utilisées, par exemple pour mesurer la cohésion lexicale entre des énoncés. Des objectifs comme la recherche d'isotopie ou l'analyse de la coréférence exigent en revanche une approche plus *sémantique*. L'adoption d'un point de vue générique sur la structure discursive exige la prise en compte de cette variété d'indices, que la modélisation d'un unique motif discursif peut du reste imposer. Il sera nécessaire de pouvoir invoquer de manière homogène ces différentes informations et retenir un mode de formalisation traduisant cette homogénéité.

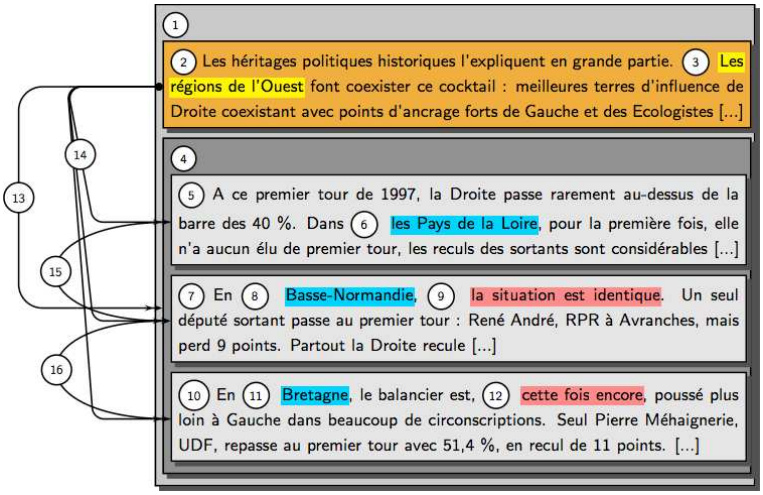

Figure 3. Exemple de structure énumérative

Afin d'illustrer ces différents aspects, nous pouvons nous appuyer sur l'exemple d'une structure discursive composée telle que la structure énumérative dont la figure 3[2] donne un exemple. Trois *segments* consécutifs (5, 7, 10) introduisent trois zones géographiques et constituent ainsi trois *items* d'une *énumération* (4). Celle-ci est incluse dans une structure de plus haut niveau (1), une *structure énumérative*, introduite par une *amorce hyperonymique* (2) indiquant la *classe* (3) dont les items sont des *instances*. Il existe donc un ensemble de *relations hyperonymiques* (14) entre l'amorce et les items. De plus, l'amorce entretient une *relation d'introduction* (13) avec l'énumération et présente une *thèse générale* ensuite déclinée pour chaque

[2] Extrait de P. Buléon. *Quarante années d'évolution politique de l'Ouest de la France : 1960-2000*.

région. Ainsi, une *relation de spécialisation* lie l'amorce et les items. Enfin différentes *cue-phrases* (9, 12) révèlent des *relations de similarité* (15, 16) entre (5), (7) et (10).

Cet exemple illustre la dimension *relationnelle* de l'organisation du discours en plus de sa composition en segments et met en évidence leurs rapports consubstantiels. Un segment de type amorce ne devient tel que par un jeu de relations hyperonymiques avec les items ; un item ne devient tel que par un ensemble de relations de coénumérabilité elles-mêmes guidées par leur rapport à l'hyperonyme commun. Cet exemple éclaire par ailleurs la possibilité de constructions imbriquées et met en lumière la variabilité du grain. Si la fonction qu'occupe par exemple l'amorce hyperonymique peut être décrite par un ensemble de critères objectifs, rien ne permet de savoir si cet élément sera de la taille d'une proposition ou d'un paragraphe. Cet exemple donne par ailleurs une idée de la variété des indices utilisables pour le repérage d'une telle structure. L'identification des expressions spatiales autour desquelles la structure est articulée peut nécessiter un travail sur les entités nommées et l'écriture d'une grammaire dédiée à ce problème particulier. La reconnaissance de relations hyperonymiques peut exiger l'accès à des connaissances. L'émergence des relations de similarité résulte pour sa part de l'emploi de *cue-phrases* (« la situation est identique »…) irréductibles à des ressources lexicales, et imposant la définition de *patrons* archétypaux, elle-même adossée à des informations lexicales sur les composants de ces expressions (« identique », « similaire », etc.). Enfin, cet exemple illustre un point décisif pour l'analyse discursive : la position des indices dans le flot textuel n'est pas toujours connue, leur ordre n'est pas toujours significatif et la distance entre différents indices peut varier d'une occurrence à l'autre. Si nous pouvons par exemple nous attendre à ce que les marques de similarité interviennent à proximité des expressions spatiales caractérisant les zones, il est difficile de « quantifier » cette proximité, et dangereux d'affirmer que leur ordre ne sera pas inversé à l'occasion : « La situation est identique en Basse-Normandie ».

3. Modèle du discours

3.1. Unités, relations et schémas

Adopter par rapport à l'organisation du discours un point de vue général exige que nous opérions une première abstraction visant à déterminer les éléments structurels fondamentaux sur lesquels s'élabore cette organisation. Nous proposons de distinguer trois « éléments » fondamentaux, irréductibles les uns aux autres et nécessaires à la description formelle des motifs discursifs envisagés : les ***unités discursives***, les ***relations discursives*** et les ***schémas discursifs*** que représente la figure 4.

Les *unités discursives* (DU) correspondent à des zones textuelles délimitées, dont le niveau de granularité peut varier. Elles sont caractérisées par un certain degré d'homogénéité de sens ou de fonction. Dans le cas de la structure énumérative envisagée ci-dessus, amorce, items et énumération constituent des telles unités.

Les *relations discursives* (DR) correspondent à l'existence d'un rapport entre DU, que celui-ci porte sur leurs propriétés internes respectives, ou sur la mise en relation par l'utilisation d'éléments externes tels que des connecteurs. Les relations de similarité de la structure énumérative illustrent cette ambivalence : la relation s'y établit d'une part par la présence d'expressions caractéristiques (« la situation est identique »…) faisant office de connecteurs indépendants des unités reliées (les items) et d'autre part par un certain degré de cohérence sémantique[3] entre les items, non représenté par la figure.

Les *schémas discursifs* (DS) sont des *patterns* discursifs de plus haut niveau, correspondant à une certaine unité fonctionnelle ou sémantique dépendante d'un ensemble de DU et d'un faisceau de DR entre ces unités. La structure énumérative constitue un tel motif. Irréductible fonctionnellement à l'une de ses parties, elle se caractérise par la présence de certaines unités (amorce, items, …) et par l'existence de relations entre ces unités (relation hyperonymique entre l'amorce et les items, …).

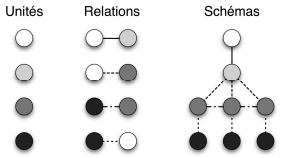

Figure 4. Unités, relations et schémas

Insistons sur l'*interdépendance* possible entre ces niveaux. Ainsi, par exemple, une énumération ne devient telle qu'à condition qu'un certain nombre d'éléments (les items) soient énumérés, mais ces derniers n'acquièrent ce statut qu'à condition de prendre place dans une énumération. Ajoutons que c'est précisément cette relative circularité qui rend problématique l'appréhension en TAL de ce type de structures.

Notons également que ce modèle du discours est par essence *récursif*. Tel schéma discursif donné pourra ainsi prendre part, en tant qu'unité, à une structure discursive de plus haut niveau. Ainsi, dans notre exemple, l'énumération, *stricto sensu*, prend place, comme unité, dans une structure énumérative liant amorce et énumération.

Enfin, insistons sur la distinction fondamentale entre *unité logique* et *inscription textuelle*. S'il est en effet tentant, en particulier aux niveaux de granularité inférieurs, de confondre l'unité sémantico-fonctionnelle avec son inscription de surface (pensons par exemple à la dualité du concept de *proposition*), il n'en demeure pas moins qu'une même unité de sens peut être *incarnée*, en discours, par des éléments distants.

Sur la base de ce triptyque, l'établissement d'une taxinomie des unités, relations et schémas fondamentaux pourrait être entrepris. Pour les relations, un lien assez naturel pourrait par exemple être tissé avec les typologies extensives proposées par la RST, ou avec des modèles visant la détermination d'un jeu restreint de relations structurelles fondamentales : subordination/coordination dans un certain nombre d'approches,

[3] Du point de vue de la situation politique dont le passage fait état.

dominance et *satisfaction-precedence* chez Grosz et Sidner (1986), etc. De façon similaire, la quête d'unités de discours fondamentales pourrait se nourrir des nombreuses typologies envisagées depuis les origines antiques de la rhétorique. Cependant, nous n'abordons pas ici cette question de l'établissement d'un tel jeu de types supposés primitifs aux différents niveaux du triptyque, et ce pour plusieurs raisons. Tout d'abord parce que la découverte d'un jeu exhaustif nous semble, comme à beaucoup d'autres, à la fois hors de propos et d'atteinte. Ensuite, parce que l'alternative consistant à ne rechercher que certaines catégories structurelles fondamentales nous semble risquer d'introduire une dépendance implicite à l'égard d'un certain niveau d'analyse. On peut par exemple se demander si le choix de la bipartition coordination/subordination n'est pas surdéterminé par une analyse conduite au niveau interpropositionnel, niveau où la valeur de cette bipartition est évidente. Si cette articulation se révèle particulièrement opérante à ce niveau, et pour des analyses ascendantes telles que celles proposées par exemple par la SDRT (Asher 1993), comment garantir leur pertinence à des niveaux de granularité plus élevés et/ou pour des approches plus descendantes ? Enfin, il nous semble préférable de réinterroger systématiquement cette taxinomie, en fonction du type de phénomène linguistique observé, pour éviter la tentation d'une adaptation forcée d'un problème à une solution.

3.2. Opérations de cohésion discursive

Se pose alors la question de la manière dont émergent, en discours, unités, relations et schémas et comment le « chaos discursif » est amené à s'articuler autour de ces éléments fondamentaux, articulation dont résulte sa structure, sa *texture*. Il s'agit ici d'étudier les mécanismes discursifs en œuvre dans l'*individuation* de ces objets pour savoir quels types de phénomènes le formalisme proposé devra permettre de décrire. Pour désigner ces mécanismes généraux par lesquels unités, relations et schémas sont révélés et rendus *prégnants*, nous parlerons d'*opérations de cohésion discursive*.

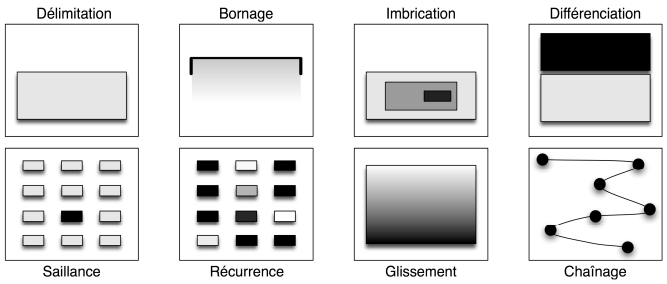

Figure 5. Opérations de cohésion discursive

Un premier type d'opération fondamentale, (*cf.* figure 5), concerne les mécanismes de *délimitation* par le biais desquels un ensemble d'éléments textuels acquiert une certaine unité : *un* paragraphe, *une* proposition, *un* item, *une* énumération, *un* mot…

Correspondant à une délimitation partielle, le *bornage* correspond à l'introduction d'une rupture au-delà de laquelle une unité peut être reconnue. L'amorce de structure énumérative et les introducteurs de cadres (*cf. infra*) en sont des exemples canoniques.

L'appartenance d'un élément à un autre élément et le transfert de propriétés qui en résulte découle du phénomène d'*imbrication*, *principe organisateur* fondamental. L'item par exemple, s'inscrit dans une énumération qui en détermine le sens.

L'opération de *différenciation* renvoie à l'identification d'une hétérogénéité entre des éléments considérés dès lors comme distincts et dans l'établissement d'une relation de contraste dont l'intensité pourra varier, en fonction de propriétés des parties et de leur connexion. On pensera par exemple à l'opposition classique entre thèse et antithèse.

Le phénomène de *saillance* correspond à l'établissement d'un focus particulier sur un objet rendu ainsi prédominant. Des degrés différents pourront résulter de la nature de l'objet et de son rapport au contexte dont il émerge. La présence d'une entité nommée établit par exemple un *focus* permettant l'ancrage d'autres éléments du texte.

Le phénomène de *récurrence* correspond à la mise en rapport d'objets distincts aux propriétés communes et à la constitution de *groupes* d'éléments potentiellement distants. Le phénomène de cohésion lexicale en fournit une bonne illustration.

L'opération de *glissement* correspond à l'établissement d'un cheminement, par une variation progressive d'une ou plusieurs propriétés de ces éléments, et/ou à l'aide de connecteurs manifestant cette progression. Le phénomène de portée d'un critère interprétatif fixé par un introducteur de cadre (*cf.* ci-après) est exemplaire à cet égard.

L'opération de *chaînage*, enfin, correspond au *maintien* ou au *rappel* d'un objet ou d'une propriété d'objet, à travers la mise en place de jalons indiquant cette persistance. Les mécanismes de chaînage anaphorique en constituent un excellent exemple.

3.3. Modélisation par contraintes de l'organisation du discours

Aux différentes *opérations* par lesquelles sont révélés les *éléments* (unités, relations et schémas) répondent différentes *fonctions* qui correspondent à la prise en charge de ces opérations, en contexte, par des éléments du texte. Une même fonction pouvant être remplie par des objets de natures différentes, nous distinguons également *fonction* et *réalisation*. Nous interrogeons ici la manière dont une *fonction* discursive peut être *réalisée* par des objets textuels devenant alors, à proprement parler, *indices*.

Considérons le matériau discursif comme un ensemble ordonné d'objets textuels. La capacité d'un objet ou d'un ensemble d'objets à occuper une fonction discursive particulière résulte nécessairement, soit de propriétés intrinsèques (par exemple morphologiques ou sémantiques) de ces objets, soit de leur rapport au contexte discursif (contraintes positionnelles ou syntaxiques par exemple).

Afin de modéliser les phénomènes discursifs, nous devons par ailleurs prendre en compte les spécificités suivantes, par lesquelles l'analyse au niveau discours se distingue du reste considérablement de l'analyse plus traditionnelle de la phrase.

Le *niveau de granularité* des phénomènes et des indices n'étant pas connu *a priori*, et un même phénomène pouvant être observé à différents niveaux, le formalisme proposé devra être aussi transparent que possible à l'égard du grain d'analyse.

D'autre part, du point de vue de la compréhension de la structure du discours, une analyse mot à mot, pas à pas, ascendante (*bottom-up*) n'est ni forcément nécessaire, ni nécessairement pertinente. Au contraire, il peut être utile de localiser et d'analyser des indices distants, sans les penser comme parties d'une suite d'éléments contigus. Le formalisme proposé devra tenir compte de cette *possible non-linéarité*.

De plus, l'ordre dans lequel les éléments apparaissent n'est pas nécessairement significatif. La coprésence, dans un ordre quelconque, d'éléments particuliers pourra ainsi valoir comme indice, et le mode de description que nous proposons devra tenir compte de cette *possible non-séquentialité*.

Dans la mesure où nous souhaitons concilier les aspects descriptifs et opératoires, nous sommes conduits à considérer la possibilité d'une approche en termes de *grammaire*. Les différentes contraintes envisagées ci-dessus exigent cependant que cette notion soit considérée avec précaution, et bien distinguée de l'analyse plus traditionnelle de la phrase. Si nous nous autorisons à parler de *grammaire de discours*, nous insistons donc sur les points suivants. Tout d'abord, aucune prétention à la générativité ne saurait y être associée : nos grammaires de discours sont *orientées description* et *détection*. De plus, la notion ne présuppose ici ni séquentialité ni linéarité.

Pour répondre à ces différentes exigences et faire face à la diversité des modes d'organisation et des indices de leur présence, nous proposons une *modélisation par contraintes des structures discursives*. On pensera ici en particulier aux *Grammaires de Propriétés* qui, dans l'univers certes différent de l'analyse syntaxique, ont montré la pertinence d'une approche par contraintes (Blache 2005) : les arguments avancés à ce niveau nous semblent à plus forte raison valables au niveau discours.

4. Le formalisme CDML

Le formalisme CDML (*Constraint-based Discourse Modeling Language*) a pour objectif de fournir un moyen formel et déclaratif de *décrire*, et d'*analyser automatiquement*, par contraintes, les structures du discours. En d'autres termes, la description formelle d'une structure discursive donnée par une grammaire CDML peut être directement utilisée par un analyseur afin de détecter celle-ci automatiquement. Son principe fondamental, fidèle aux analyses précédentes, consiste à permettre de préciser un ensemble de contraintes devant être satisfaites par des objets textuels pour devenir indices de la présence d'objets textuels de plus haut niveau.

Le discours est ici considéré comme une succession d'*objets de discours* (DO) sur lesquels les contraintes seront exprimées. Leur taille peut varier, de même que leur statut linguistique : unités morpho-syntaxiques, éléments syntagmatiques... Ils représentent toute l'information disponible à un certain stade de l'analyse, information

pouvant émaner de toute analyse préalable. Chaque DO est associé à une *structure de traits* (notée FS pour *feature-set*) qui en représente l'information pertinente (statut morpho-syntaxique, interprétation sémantique, etc.). Les DO entrent par ailleurs dans un ensemble de relations (syntaxiques, etc.), elles aussi représentées par de tels FS. L'ensemble des informations linguistiques utilisables sera rendu accessible par ce biais, et les contraintes porteront principalement sur ces représentations symboliques pour lesquelles nous utilisons une notation plate, dans laquelle les accolades permettent l'imbrication hiérarchique, comme l'illustre la figure 6.

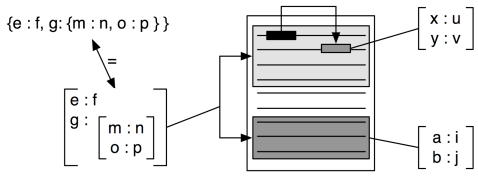

Figure 6. Représentation du discours

Une grammaire CDML est composée d'un ensemble de *règles*. Chacune a pour objet de décrire et de détecter un élément de discours, c'est-à-dire une unité (DU), une relation (DR) ou un schéma (DS). Nous nous limitons ici aux DU et DR. À chacun de ces éléments est associé un type de règle dédié. Chaque règle est composée d'un ensemble d'appels de contraintes. Sa structure fondamentale est :

```
RuleType:
        constraint-1
        constraint-2
        …
```

où *RuleType* peut être *Unit* ou *Relation* et où les contraintes utilisées sont choisies parmi celles disponibles pour ce type de règle. Une règle peut produire en sortie une représentation symbolique du phénomène décrit, à l'aide d'un FS :

```
Unit {a : {b : X,  c : {d : Y, e : Z } } }:
        unit-constraint-1
        uni-constraint-2
        …
```

La satisfaction de la règle, c'est-à-dire la détection d'un objet textuel satisfaisant les contraintes indiquées, génère un objet de discours (DO) qui pourra être utilisé par d'autres règles pour appliquer des contraintes d'ordre supérieur :

```
Relation rule1 {a:b} requires rule2, rule3:
        relation-constraint-1
Unit rule2:
        …
```

Les contraintes constituent un ensemble extensible de *primitives discursives* appelées pour filtrer certains candidats d'un espace de recherche. Un appel est de la forme :

```
constraint-name(arg-1:val-1, arg-2:val-2…)
```

où *arg-1* et *arg-2* sont des arguments nommés dépendant du type de contrainte.

Les contraintes portent fréquemment sur les FS associés aux DO, par le biais de paramètres précisant les motifs recherchés. L'exemple d'une contrainte assez intuitive illustrera cette idée. La grammaire suivante :

```
Unit:
        start(pattern:{type:"pronoun"})
```

décrit et accepte toute unité textuelle qui *commence* par un objet de discours (DO) dont le FS *unifie* (*cf.* ci-après) avec *{type:"pronoun"}*.

De plus, toutes les contraintes peuvent être préfixées par l'opérateur *not*, afin d'obtenir le complémentaire de l'ensemble des candidats filtrés. Ainsi, la grammaire suivante accepte toute unité ne commençant pas par un pronom :

```
Unit:
        not start(pattern:{type:"pronoun"})
```

Le *matching* entre FS opère par *unification*. Nous distinguons *unification standard* (notée ~) et *unification forte* (notée ≈). Contrairement à l'unification standard, l'unification forte (ou *filtrage*) considère l'un des éléments comme le *modèle* auquel l'autre doit se conformer. Ainsi, *{a:b} ~ {c:d}*, mais *{a:b}* !≈ *{c:d}*. La comparaison entre un *pattern* et le FS associé à un DO opère par unification forte. Ainsi :

```
Unit:
        start(pattern:{type:"sentence"})
```

accepte à la fois les DO représentés par *{type:"sentence"}* et *{type:"sentence", size:"short"}*. Mais la règle suivante n'accepte que les seconds :

```
Unit:
        start(pattern:{type:"sentence", size:"short"})
```

À l'inverse, le mécanisme d'unification de variables utilise l'unification standard et est déclenché implicitement pour chaque occurence de variable. La grammaire suivante identifie les segments commençant et se terminant par des phrases de « même taille ».

```
Unit segment {firstSentenceLength: $a}:
        start(pattern:{type:"sentence", size: $a})
        end(pattern:{type:"sentence", size: $a})
```

Notons que l'unification standard permet ici la *remontée d'information* : l'objet d'ordre supérieur créé par la règle *transfère* l'information issue de ses composants.

Dans le cas des règles visant la description d'une relation, précisons que les extrémités de la relation sont elles-mêmes définies à l'aide de contraintes. Ainsi, la grammaire :

```
Relation {type: "subject-verb"}:
        target(pattern:{type:"subject"})
        target(pattern:{type:"verb" })
```

décrit les relations entre un élément reconnu comme sujet et un élément reconnu comme verbe, quels que soient du reste l'ordre et la distance entre ce sujet et ce verbe.

Une règle décrivant une relation peut faire appel à une règle décrivant une unité, en indiquant cette dépendance et en posant des contraintes sur les traits de cette unité.

```
Relation {type: "subject-verb"} requires subject:
        target(pattern:{type:"subject"})
        target(pattern:{type:"verb" })
Unit subject {type: "subject"}:
        is(pattern:{type:"pronoun"})
```

De la même manière, une règle décrivant une unité peut faire appel à une règle décrivant une relation en utilisant une contrainte dédiée à cette tâche.

```
Unit {type: "subject-noun"} requires subjet-verb:
        is(pattern:{type:"noun"})
        inRelation(pattern:{type:"subject-verb"})
Relation subjet-verb {type: "subject-verb"}:
        …
```

Chaque règle peut par ailleurs préciser une *perspective d'analyse*, c'est-à-dire un certain point de vue sur le matériau discursif auquel elle s'applique. En particulier, elle peut préciser une MRU (*Maximal Relevant Unit*), soit un type d'unité au-delà duquel aucun candidat satisfaisant les contraintes ne devra être recherché. Pour décrire par exemple une relation sujet-verbe à l'intérieur de la phrase, on précisera :

```
Relation {type: "subject-verb"}:
@mru:['sentence']
        target(pattern:{type:"subject"})
        target(pattern:{type:"verb" })
```

Il est également possible de masquer certains éléments du flot textuel, soit en filtrant certains d'entre eux, soit au contraire en n'acceptant explicitement que certains autres,

pour se concentrer sur les objets effectivement pertinents pour une règle donnée. Par exemple, la règle suivante permet de décrire une phrase contenant trois propositions[4].

```
Unit {type: "3-clauses"}:
@accept:['clause']
        size(amount:3)
```

Enfin, si les constituants d'une unité textuelle sont par défaut accessibles, et s'il est également possible de distinguer ses bornes gauche et droite, il est cependant possible de la considérer au contraire comme un atome indivisible dont les constituants sont inaccessibles. Si nous souhaitons décrire une phrase dont le verbe de la principale est au présent, et dont le temps de la subordonnée est indifférent, nous pouvons interdire l'analyse du contenu de la subordonnée en en faisant un atome, un *token*.

```
Unit {type: "present-sentence"}:
@tokens:['subordinate-clause']
        contains(pattern:{type:"verb", tense:"present"})
```

Les exemples de grammaires fournis ci-après donneront une idée plus précise des possibilités de modélisation des structures du discours offertes par le formalisme CDML, et de certaines contraintes disponibles dans le langage. Pour une présentation plus formelle du système de contraintes, on pourra se reporter à Widlöcher (2006) ainsi qu'à la documentation du formalisme.

5. Exemple de l'analyse des cadres de discours temporels

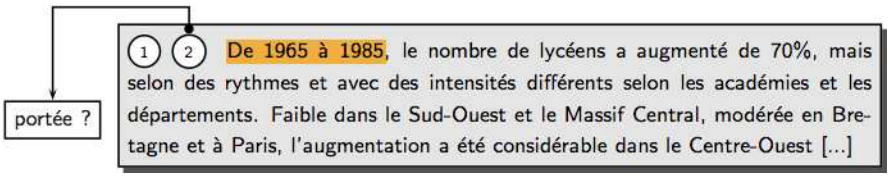

Figure 7. Exemple de cadre de discours temporel[5]

Afin de mettre en évidence la pertinence de cette approche pour l'analyse d'*unités discursives*, procédons à une première étude de cas. L'hypothèse psycho-linguistique de l'encadrement du discours (Charolles 1997) identifie des segments textuels, appelés *cadres de discours*, homogènes du point de vue d'un critère d'interprétation fixé dans une expression en position détachée, nommée *introducteur de cadre*. Nous nous limitons ici aux *cadres temporels* (*cf.* figure 7). L'opérationnalisation en TAL de ce modèle impose l'analyse des expressions temporelles, l'identification de la fonction d'introducteur de certaines d'entre elles et la détermination de la portée des

[4] Ce n'est bien entendu ni la seule manière d'imposer cette contrainte, ni d'ailleurs la plus élégante.
[5] Extrait de R. Hérin. et R. Rouault, *Atlas de la France scolaire de la maternelle au lycée*.

introducteurs. Nous nous intéressons ici à ce dernier et très épineux problème. Une analyse des expressions temporelles, des verbes et des introducteurs est supposée effectuée et différents FS rendent l'information associée disponible.

Trois types d'indices permettent de déterminer la portée de l'introducteur (Bilhaut *et al* 2003). Tout d'abord, des critères *énonciatifs* tels que les temps des verbes sont utilisables, un changement de temps indiquant souvent la fermeture du cadre courant. Par ailleurs, l'incompatibilité *sémantique* entre l'intervalle temporel fixé par l'introducteur et les autres expressions temporelles fournit un critère décisif de fermeture. Enfin, des indications *structurelles* doivent être considérées : un cadre n'est composé que de phrases complètes. Considérons à présent la grammaire suivante :

```
Unit frame {type:"frame", sub-type:"temporal"}:
        start(pattern:{type:"introducer"})
        end(pattern:{type:"sentence"})
        not absolutePresence(pattern:{type:"introducer"}, amount:2)
        homogeneity(comparator:scope)
        size(mode:#LONGEST)
Comparator scope ({type:"verb"} as $v1, {type:"verb"} as $v2):
        $v1/tense = $v2/tense
Comparator scope ({type:"introducer"} as $i, {type:"temporal"} as $t):
         (($i/start >= $t/start) and ($i/start <= $t/end))
        or
        (($i/end >= $t/start) and ($i/end <= $t/end))
```

Nous recherchons une unité textuelle commençant par un DO identifié, par une analyse préalable, comme introducteur. Cette unité devra se terminer par une phrase, c'est-à-dire n'être composée que de phrases complètes. La contrainte suivante interdit l'ouverture d'un cadre imbriqué. La contrainte d'homogénéité garantit que les relations de comparaison définies par le *Comparator* nommé *scope* sont vérifiées. L'analyse préliminaire des verbes est supposée avoir produit des FS de la forme *{tense:"present"}*. La première *signature* du comparateur de portée contraint les verbes à être au même temps. La seconde vérifie que les expressions temporelles désignent des intervalles compatibles avec l'introducteur. La dernière contrainte élimine les candidats cadres qui, pour un même introducteur, sont inclus dans un cadre plus grand. Les unités textuelles satisfaisant ces contraintes seront symboliquement représentées par le FS *{type:"frame", sub-type:"temporal"}*.

6. Exemple de l'analyse des relations de contraste

Afin d'illustrer les bénéfices d'une approche par contraintes pour la description et le traitement automatique de *relations discursives*, procédons à une seconde étude de cas, centrée sur la question des *relations de contraste*. À juste titre considéré comme un

principe fondamental d'organisation de l'information, et trouvant naturellement sa place dans la plupart des taxinomies de relations rhétoriques, le phénomène du contraste soulève un ensemble de questions caractéristiques en analyse du discours, en particulier en termes de granularité et de valeur sémantique des structures concernées.

6.1. Le contraste comme *différenciation*

Notre acception de la notion de contraste différant significativement du sens qu'on y accorde traditionnellement, précisons les principaux axes de démarcation. Tout d'abord, bon nombre d'approches considèrent essentiellement la notion de contraste à un niveau de granularité inter-propositionnel ou inter-phrastique finalement assez proche de la tradition logique mais sans retenir nécessairement la ligne vériconditionnelle associée à cette tradition. À ce niveau d'investigation, l'étude des connecteurs (en particulier « mais ») trouve naturellement une place de choix, même si les contrastes non ainsi marqués font également l'objet d'études. Fidèle à notre choix de privilégier la variabilité du grain d'analyse, nous ne restreignons pas notre analyse à ce niveau *mésostructurel* (*cf. infra*) et considérons que des phénomènes contrastifs opérant à d'autres échelles jouent un rôle décisif dans l'organisation du discours.

D'autre part, si la notion de contraste implique un certain degré d'identité entre les éléments mis en contraste, *axe* commun autour duquel le contraste est possible, reste que l'accent est traditionnellement mis sur la différence radicale, l'opposition entre les éléments liés. Au contraire, nous privilégions la dimension *distinctive* du contraste, c'est-à-dire sa propension à procéder à *l'individuation* d'objets textuels. Loin d'impliquer une interruption de la cohérence textuelle, le contraste constitue un mode particulier d'*agrégation* par lequel des objets textuels sont à la fois *rapprochés* et *distingués*. Nous conférons ainsi à la notion de contraste un spectre assez vaste acceptant divers degrés de différenciation, de la rupture textuelle entre éléments comparables différents mais non opposés, à l'opposition manifeste entre ces éléments.

Ainsi défini, le contraste peut être observé à différents *niveaux*. En toute rigueur, pour distinguer ces niveaux, il est nécessaire de considérer d'une part l'*échelle de la structure* contrastive et d'autre part l'*échelle des termes* de la relation. Une structure contrastive peut, par exemple, mettre en relation des éléments lexicaux (échelle de termes) distants dans le texte (échelle de la structure). Cependant, nous privilégions ici l'échelle de la structure. Nous faisons en effet l'hypothèse qu'une relation de grande échelle s'appuyant sur des éléments de petite échelle ne deviendra signifiante que si ces éléments reliés permettent de définir des structures intermédiaires constituant les termes effectifs de la relation. Par exemple, une relation de contraste entre deux termes antonymes distants de plusieurs paragraphes ne sera pas, selon cette hypothèse, signifiante en tant que telle. Si les termes antonymes peuvent en revanche être considérés comme *représentants* (thématiques, …) de phrases ou de paragraphes, le *rapprochement d'échelle* (nous parlerons aussi de *mise à l'échelle*) peut conduire à l'émergence d'une structure signifiante.

Nous distinguons différents niveaux de structuration contrastive que le formalisme proposé permet de représenter et dont certains seront illustrés ci-dessous. D'un point de vue *microstructurel*, nous pourrions nous intéresser aux rapports de différenciation existant, au niveau le plus local, entre mots, domaine bien balisé par la rhétorique classique à travers par exemple l'étude des *alliances de mots* et *oxymores*. Une analyse des relations antonymiques pourrait également être entreprise à ce niveau. Le niveau que nous qualifions de *mésostructurel* regroupe l'ensemble des structures contrastives opérant aux niveaux interpropositionnel et interphrastique, telles que décrites dans les approches traditionnelles déjà évoquées. À ce niveau, on s'intéressera en particulier aux articulations entre propositions ou phrases déterminées par l'usage des connecteurs (« mais », « cependant », …). Domaine de relations entre phrases et/ou groupe de phrases délimités structurellement (paragraphes, parties…) ou sémantiquement (cadre de discours, zone de cohésion lexicale…), le niveau *macrostructurel* révèle une diversité de relations contrastives d'autant plus importante que les unités de sens (à relier) existant à ce niveau sont variées. Une rupture de cohésion lexicale conduit par exemple à l'émergence de segments contigus en relation de contraste. Deux items d'une énumération, deux cadres de discours de même type, deux éléments à la fois comparables et différents, introduisent bien, dans leur succession même, un effet de contraste. D'autres motifs contrastifs pourront prendre appui sur des critères moins interprétatifs, tels que des parallélismes structurels. Nous qualifions de *superstructurelles* les relations intradocumentaires de plus haut niveau. Elles interviennent entre les éléments globaux telles que parties ou sous-parties, éléments correspondant au « plan » du document. À ce niveau, le contraste opère en particulier à travers des motifs tels que thèse / anthithèse, position critiquée / position défendue, etc. Le niveau *hyperstructurel* désigne des relations interdocumentaires que nous n'abordons pas ici. On pourrait cependant souhaiter les modéliser comme des structures intra-documentaires, ce que le formalisme proposé n'interdit pas *a priori*.

6.2. Similitude

Le contraste repose sur une certaine *similitude* et un certain degré de *différenciation*. En ce qui concerne la similitude, on distinguera similitudes *sémantique* et *structurelle*.

La *similitude sémantique* correspond au partage d'un certain nombre de propriétés entre les éléments en présence ou à leurs rapports communs avec un élément de sens extérieur. Un élément de type *région géographique* est comparable avec un autre élément de ce type. Un *bateau* pourra pour sa part être mis en relation de similarité avec une *voiture* par le biais de leur hyperonyme commun *véhicule*. Plus généralement, la similitude sémantique résulte de la coprésence des éléments considérés sur un même *axe sémantique* tel que *zone géographique* ou *véhicule*. Tel élément de sens appartenant à un axe pourra du reste qualifier une unité textuelle de grain plus important, comme c'est le cas pour chaque item de l'énumération envisagée ci-dessus (*cf.* figure 3). Dans le cas plus traditionnel des relations contrastives inter-propositionnelles, une nécessaire similitude lie là encore les propositions en contraste. Un énoncé du type « Jean est en retard, mais il ne semble pas pressé », n'est cohérent

que parce qu'un rapport de sens lie distinctement *être en retard* et *être pressé* sur l'axe du *rapport au temps*. La prise en compte de la diversité sans doute infinie des axes et relations sémantiques pose évidemment problème, en particulier dans la perspective d'un traitement automatique. Certains axes que nous qualifierons de conventionnels interviennent cependant de manière récurrente dans la structuration du texte. On sait par exemple le pouvoir fortement structurant du temps ou de l'évaluation axiologique pour l'organisation textuelle.

La *similitude structurelle* porte non plus sur les propriétés sémantiques des parties reliées, mais sur certains parallélismes structurels permettant de les assimiler. Deux énoncés répondant à un même motif syntaxique seront par exemple d'autant plus naturellement rapprochés que ce motif sera peu fréquent. Deux occurrences du schéma assez rare <Verbe infinitif–Verbe–SN> (« Voter est un devoir ») produiront ainsi un effet d'appel, éventuellement renforcé par la reprise d'éléments de surface occupant certaines fonctions (« Voter est un droit » / « Voter est un devoir »), ou par des éléments de même ordre occupant ces fonctions. La simple reprise de n-grammes, par exemple de mots, produit également cet effet de rappel, que la reprise soit d'ailleurs strictement séquentielle ou non. Ainsi, les énoncés « Résister, c'est dire oui » et « Mais résister, c'est aussi dire non » partagent un degré de similitude pouvant être reconnu à un niveau infra-syntaxique. À un niveau plus « superficiel[6] » encore, la taille des éléments comparés constitue un indice de similitude décisif. Unités textuelles et éléments de sens seront en effet d'autant plus comparables que leurs tailles seront *commensurables*, ce qui renvoie d'ailleurs à la contrainte de *mise à l'échelle* évoquée ci-dessus. La similitude structurelle peut également porter non plus sur la structure interne des objets considérés, mais sur leurs rapports au contexte. Deux objets étant dans des rapports similaires à leur contexte commun ou à leurs contextes respectifs partageront quelque similitude. Les rapports positionnels et fonctionnels au contexte jouent à ce niveau un rôle prépondérant. Deux introducteurs de cadres proches indiquent par exemple, par leur même position détachée en initiale de phrase, et éventuellement de paragraphe, un lien de similitude pouvant servir de base à une relation de contraste. Enfin, la manière dont chaque élément comparé s'inscrit dans le contexte de l'autre est un critère fondamental de rapprochement. La *proximité*, dans le flux textuel, des éléments concernés invite par exemple à les rapprocher en leur conférant d'emblée quelque similitude. De ce fait, les relations de contraste seront à chercher prioritairement dans le contexte immédiat de chaque terme de la relation.

6.3. Différenciation

Sur la base de cette *similitude*, différents degrés de *différenciation* sont possibles, couvrant un spectre allant de la similarité à l'opposition.

La similitude peut résulter, avons-nous dit, de la co-appartenance à un même *axe sémantique*. Une première différenciation résulte de la distance entre les éléments sur

[6] Privilégiant les éléments de surface, indépendamment de leur sens ou de leur fonction.

cet axe sémantique, même si cette différenciation n'implique en elle-même aucune opposition. À ce niveau, il est nécessaire de prendre en compte la nature des rapports entretenus par les différents éléments d'un même axe sémantique. Ainsi, nous pouvons considérer l'axe de l'évaluation axiologique, comme comportant une opposition franche entre évaluations positive et négative. L'axe de la localisation spatiale ne comporte en revanche pas de contradiction essentielle entre ses sous-parties. *Bien* et *mal* s'opposent plus intrinsèquement que *Bretagne* et *Basse-Normandie*, en vertu des propriétés fondamentales des axes sémantiques auxquels ils appartiennent.

Ce constat nous amène à aborder un aspect essentiel de la structuration argumentative du discours, en considérant l'existence de jeux d'opposition fondamentaux autour desquels le cheminement logique du texte est fréquemment articulé. Sur le plan de l'évaluation, l'opposition entre le bien et le mal, le valide et l'invalide est ainsi, par exemple, tout à fait canonique. Sur le plan temporel, l'opposition passé/présent/futur est elle aussi un levier fréquent de construction du parcours textuel. Pensons encore à l'opposition entre le possible et l'impossible, le descriptif et le prescriptif, etc.

Nous venons d'envisager les cas dans lesquels le degré de différenciation entre les éléments dépend de leur distance sur l'axe garantissant leur similitude. Similitude et différenciation peuvent également être prises en charge par des éléments textuels différents, le cheminement discursif opérant alors un déplacement sur plusieurs axes. Dans cette configuration, les jeux d'opposition fondamentaux évoqués ci-dessus jouent un rôle prépondérant, en tant que principes conventionnels de différenciation. Un déplacement sur l'axe temporel assurant la similitude pourra par exemple être doublé d'un déplacement sur l'axe de la possibilité assurant la différenciation : « En 1960, tel phénomène était possible […]. En 1980, il ne l'est plus. ». À cet égard, l'axe opposant l'identique et le différent occupe une position essentielle. Circuler sur cet axe reviendra à expliciter le rapport de différenciation existant entre les éléments en relation. La structure énumérative (*cf.* figure 3) en fournit un exemple caractéristique.

Si l'exploitation de ces stéréotypes discursifs est récurrente, la différenciation peut cependant résulter de principes différents. Reste que le processus de différenciation peut souvent être rapporté à ces stéréotypes. Par exemple, sur la base d'une similitude par reproduction de n-grammes, l'opposition peut être déterminée par les relations sémantiques entre des éléments entrant dans la composition de ces n-grammes.

6.4. Exemples de descriptions formelles de relations contrastives

Étant donné cet arrière-plan théorique, nous pouvons envisager différents exemples représentatifs de la diversité des organisations contrastives, et de l'homogénéité de description permise par une modélisation par contraintes. Il s'agit principalement ici de mettre en évidence l'expressivité de CDML et son adéquation avec les besoins de l'exploration d'un phénomène tel que le contraste. On ne s'étonnera donc pas de voir traités des exemples choisis pour leur caractère schématique.

6.4.1. Contraste entre homogénéités lexicales

Un premier exemple[7], illustré par la figure 8, porte sur l'émergence d'une relation de contraste entre deux unités textuelles présentant une homogénéité au niveau lexical.

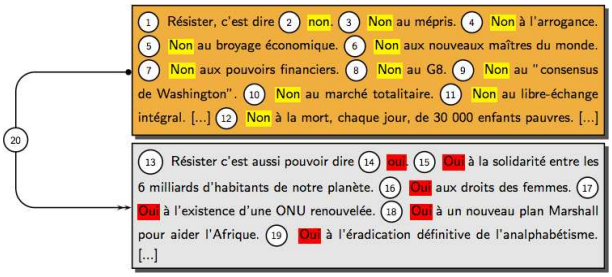

Figure 8. Contraste entre homogénéités au niveau lexical

Plus précisément, nous nous intéressons aux éléments lexicaux relatifs à l'expression du refus ou de l'acceptation (réduits pour l'exemple à « oui » et « non ») et décrivons des segments aussi longs que possible, composés de phrases complètes contenant plus de termes exprimant le refus (ou l'acceptation) que de termes exprimant l'acceptation (ou le refus). La grammaire CDML suivante décrit de telles structures.

```
Unit acceptanceSentence {type:"acceptance-sentence"}:
        is(pattern:{type:"sentence"})
        majority(pattern:{type:"lexical-acceptance"}, versus:[{type:"lexical-refusal"}])
Unit refusalSentence {type:"refusal-sentence"}:
        ...
Unit acceptanceSegment {type:"acceptance-segment"} requires refusalSentence, acceptanceSentence:
        setOf(pattern:{type:"sentence"})
        majority(pattern:{type:"acceptance-sentence"}, versus:[{type:"refusal-sentence"}])
        size(mode:#LONGEST)
Unit refusalSegment {type:"refusal-segment"} requires refusalSentence, acceptanceSentence:
        ...
Relation contrast {type:"contrast"} requires acceptanceSegment, refusalSegment:
        target(pattern:{type:"acceptance-segment"})
        target(pattern:{type:"refusal-segment"})
```

6.4.2. Contraste entre homogénéités au niveau des énoncés

On pourra souhaiter représenter des zones textuelles homogènes du point de vue de la présence d'*énoncés* prenant en charge refus ou acceptation (*cf.* figure 9). Cependant, les structures de haut niveau visées ici (segments et relation de contraste) restent inchangées, et la grammaire devra traduire cette transparence relative.

[7] Le passage est extrait du *Monde Diplomatique*.

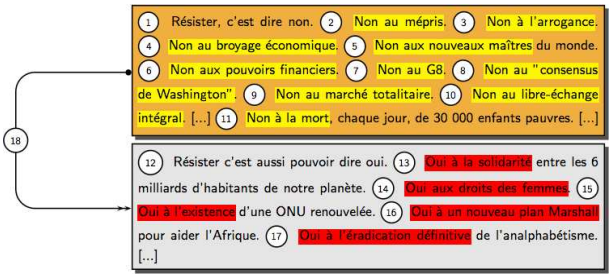

Figure 9. Contraste entre zones homogènes au niveau des énoncés

Par simplification, nous considérons qu'un énoncé exprimant refus ou acceptation est composé de l'adverbe « oui » ou « non » suivi d'une préposition et d'un syntagme nominal dont nous ignorons les éventuelles extensions autres qu'adjectivales. Nous souhaitons par ailleurs, pour un traitement ultérieur, représenter ces énoncés par une structure de traits de la forme *{type:"[acceptance/refusal]-statement", object:[tête du SN]}*. L'analyse des syntagmes nominaux est supposée ici avoir déjà été réalisée. La description des segments et de la relation de contraste ne différant que peu de ce qui a été vu, nous indiquons seulement ici la partie consacrée à la modélisation des énoncés.

```
Unit refusalStatement {type:"refusal-statement", object:$head}:
@mru:['sentence']
        start(pattern:{type:"refusal"})
        end(pattern:{type:"sn", head:$head})
        not absolutePresence(pattern:{type:"sn"}, amount:2)
Unit acceptanceStatement …
```

On notera que les motifs décrits sont intra-phrastiques (*@mru:['sentence']*) et que nous filtrons les compléments du nom en interdisant la présence de plus d'un unique SN. L'unification permet la remontée de l'information relative à la tête du SN.

Sur cette base, nous pourrions étudier les rapports entre les objets d'acceptation et de refus visés par ces énoncés. Par exemple, l'exploration d'énoncés symétriques refusant et acceptant des éléments de sens liés par une relation d'antonymie (« Oui à l'existence », « Non à la mort »), pourrait être entreprise par la grammaire suivante :

```
Relation antonym-statement {type:"antonym-statement"} requires refusal-statement, acceptance-statement:
        target(pattern:{type:"acceptance-statement", object:$ob1})
        target(pattern:{type:"refusal-statement", object:$ob2})
        antonymy(words:[$ob1,$ob2])
```

6.4.3. Contraste et parallélisme structurel

L'importance des rapports introduits par des parallélismes structurels a été évoquée. Envisageons ici une reprise de n-grammes telle que celle illustrée par la figure 10.

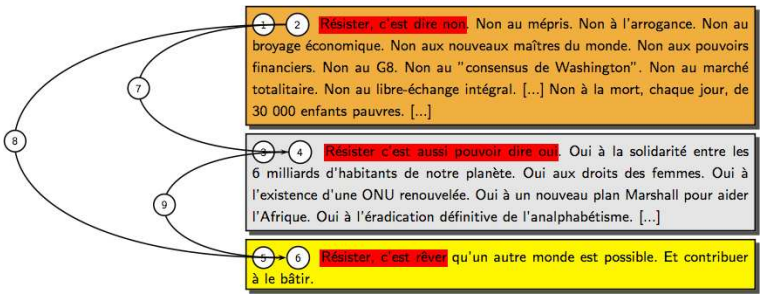

Figure 10. Contraste et parallélisme structurel

Considérons par exemple ici des bigrammes de mots pleins, soit des *séquences non-linéaires* de verbes, noms et adjectifs. Ces n-grammes seront représentés par une structure de traits retenant les formes lemmatisées de leur constituants. Ainsi, la séquence « Résister, c'est dire non » débute par un bigramme de la forme *{lemma-1:"résister", lemma-2:"être"}*. Nous interrogeons dès lors les relations entre ces n-grammes, en recherchant les motifs faisant l'objet de reprises au fil du texte, tels que 2, 4 et 6 dans l'exemple. Les n-grammes faisant l'objet d'une reprise acquièrent un rôle structurant particulier, le phénomène d'écho faisant de chacun d'eux une sorte de « jalon » dans le cheminement textuel. Nous les qualifions de *n-grammes structurants*. Enfin, nous décrivons les segments textuels tels que 1, 3 et 5 introduits par de telles articulations. La grammaire suivante permet 1) de repérer les bigrammes ; 2) de mettre en relation ceux qui se font écho ; 3) de définir ces derniers comme structurants ; 4) d'annoter les segments introduits par ces éléments, composés de phrases complètes et bornés par les ruptures introduites par un nouvel élément structurant.

```
Unit nGram {type:"nGram", length:2, elements:{lemma-1:$l1, lemma-2:$l2}}:
@mru:['sentence']
@accept:['relevantWord']
        consecution(patterns:[{type:"relevantWord",lemma:$l1}, {type:"relevantWord",lemma:$l2}])
Relation parallelism {type:"parallelism"} requires nGram:
@mru:['document']
        leftTarget(pattern:{type:"nGram",elements:{lemma-1:$l1, lemma-2:$l2}})
        rightTarget(pattern:{type:"nGram",elements:{lemma-1:$l1, lemma-2:$l2}})
Unit structuringNGram {type:"structuringNGram"} requires parallelism, nGram:
@mru:[document]
        is(pattern:{type:"nGram"})
        inRelation(pattern:{type:"parallelism"})
Unit segment {type: "segment"} requires structuringNGram:
        start(pattern: {type:"structuringNGram"})
        end(pattern: {type:"sentence"})
        not absolutePresence(pattern: {type:"structuringNGram"}, amount:2)
        size(mode:#LONGEST)
```

Les mots pleins (verbes, noms et adjectifs) sont supposés déjà identifiés en tant que *relevantWords*. Pour décrire les n-grammes, on ne considère que ces mots pleins

(*@accept:['relevantWord']*), et on ne prend en compte que les séquences de termes appartenant à une même phrase (*@mru:['sentence']*). Leur structure de traits indique les lemmes des constituants. La relation de parallélisme met en rapport des n-grammes situés à une distance quelconque (*@mru:['document']*) et dont les constituants sont contraints, par unification, à avoir les mêmes formes lemmatisées. Sont considérés comme *structuringNGrams*, tous les n-grammes en relation de parallélisme. Les segments sont définis comme unités composées de phrases complètes, introduites par un élément structurant et ne contenant pas d'autre élément de ce type.

6.4.4. Contraste temporel au niveau propositionnel

L'axe temporel fournit de nombreux exemples de relations contrastives. Entre propositions juxtaposées, par exemple, un changement de temps traduit souvent un changement d'état, porteur de contraste, comme l'illustre la figure 11[8].

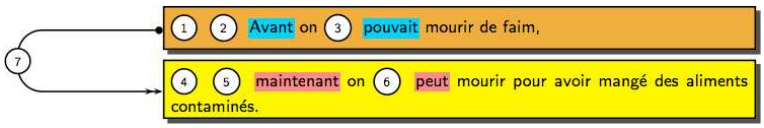

Figure 11. Contraste temporel au niveau propositionnel

Nous supposons que les propositions juxtaposées ont été analysées (par exemple en CDML), et qu'elles sont représentées par les traits *{type: "clause"*, *temporal-value: [past / present / future]}*. Nous nous concentrons ici sur la description du contraste.

```
Relation temporal-contrast {type:"temporal-contrast"}:
        within(pattern:{type:"sentence"})
        maxDistance(distance: 0)
        target(pattern:{type:"clause", temporal-value:"past"})
        target(pattern:{type:"clause", temporal-value:"present"})
```

Nous décrivons ici des relations intra-phrastiques entre propositions immédiatement juxtaposées. Nous fixons les valeurs "passé" et "présent", mais n'indiquons pas l'ordre dans lequel ces valeurs devront apparaître. Si nous souhaitions en revanche représenter des relations de contraste temporel *chronologique*, nous pourrions écrire par exemple :

```
Relation temporal-contrast {type:"temporal-contrast", order:"chonological"}:
        ...
        leftTarget(pattern:{type:"clause", temporal-value:"past"})
        rightTarget(pattern:{type:"clause", temporal-value:"present"})
```

[8] Ce passage est également extrait du *Monde Diplomatique*.

6.4.5. Contraste temporel au niveau macro-structurel

L'étude des structures temporelles s'avère également essentielle pour l'analyse argumentative à des niveaux de granularité plus macroscopiques. En particulier, l'organisation chronologique de parties d'un texte constitue un mode de structuration stéréotypique. On souhaitera repérer par exemple des zones textuelles homogènes du point de vue des temps verbaux ou du point de vue de catégories plus générales telles que {passé, présent, futur}. Pour déterminer cette homogénéité, on pourra privilégier les marques verbales présentes dans certaines positions particulières, en ignorant par exemple les temps verbaux des propositions subordonnées, ou ceux présents dans les appositions, les parenthèses, etc. Sur cette base, on recherchera alors des motifs discursifs de plus haut niveau consistant en une mise en relation de ces unités textuelles, à travers par exemple un cheminement chronologique. La grammaire suivante explore ces structures. On considère que le découpage en propositions et l'analyse des temps verbaux sont réalisés. Pour chaque phrase, on détermine son temps, *i.e.* le temps de sa principale, en ignorant les marques verbales présentes dans les autres propositions. Par simplification, nous ignorons les propositions coordonnées et juxtaposées. On définit ensuite des segments textuels homogènes temporellement, *i.e.* composés de phrases au même temps. Enfin on analyse les motifs chronologiques composés d'un enchaînement ordonné de séquences au passé, au présent et au futur.

```
Unit sentence {type:"sentence", tense:$tense}:
@tokens:['subordinate-clause', 'parenthesis', 'apposition']
        is(pattern:{type:"sentence"})
        absolutePresence(pattern:{type:"verb", tense:$tense})
Unit temporal-segment {type:"temporal-segment", tense:$tense} requires sentence:
        start(pattern:{type:"sentence", tense:$tense})
        end(start(pattern:{type:"sentence"}))
        size(mode:#LONGEST)
        homogeneity(comparator:tense)
Comparator tense ({type:"sentence"} as $sentence-1, {type:"sentence"} as $sentence-2):
        $sentence-1/tense == $sentence-2/tense
Unit chonology {type:"chronology"} requires temporal-segment:
        consecution(patterns:[{type:"temporal-segment", tense:past},
                              {type:"temporal-segment", tense:present},
                              {type:"temporal-segment", tense: future}])
```

On notera ici en particulier l'utilisation de la perspective d'analyse (*@tokens*) pour la détermination du temps de phrases. Considérer subordonnées, appositions et parenthèses comme des *tokens*, c'est aussi les rendre atomiques, indivisibles, et donc masquer les verbes présents dans ces unités, comme nous le souhaitons.

7. Conclusion

Notre objectif principal était de trouver un mode unifié de description et d'analyse des structures discursives, dont la diversité pose d'importants problèmes méthodologiques et computationnels, et exige des outils descriptifs et opératoires adaptés. Une approche par contraintes peut satisfaire ces exigences, par la prise en compte d'indices de nature variée, et en autorisant une vue non linéaire et non séquentielle du discours.

Dans cet esprit, le formalisme CDML permet la description et le traitement automatique des structures discursives. L'implémentation que nous en proposons, sur la base de la formalisation évoquée dans Widlöcher (2006), prend la forme d'un composant pour la plate-forme Linguastream[9] et tire avantage de ses principes (Widlöcher et Bilhaut 2005). Plate-forme générique de TAL, elle permet la mise en œuvre de chaînes de traitement articulant, sur corpus, des analyses de type et de niveau variés : morphologique, syntaxique, sémantique... Chaque étape produit de nouvelles informations sur lesquelles les étapes ultérieures peuvent s'appuyer. Une analyse CDML peut ainsi s'intégrer à une chaîne LinguaStream, les *discourse objects* sur lesquels portent les contraintes pouvant résulter de tout composant situé en amont.

Si la nature de « méta-modèle », c'est-à-dire de cadre formel pour l'expression de modèles linguistiques, du formalisme proposé rend délicate une évaluation autre que qualitative, en termes de pouvoir expressif dudit méta-modèle, les modèles exprimés à l'aide de ce dernier peuvent pour leur part être évalués. Ainsi, nous avons par exemple amorcé une démarche d'évaluation de l'analyseur de cadres de discours faisant intervenir l'analyse CDML présentée ci-dessus (Ferrari *et al.* 2005). Les autres exemples sont issus de travaux expérimentaux conduits sur les relations de contraste à l'aide du formalisme proposé. D'un point de vue opérationnel, cette méthode s'avère, en pratique, tout à fait utilisable. Insistons cependant sur le privilège accordé à l'expressivité du langage et à la manipulation expérimentale.

Références

ASHER N. (1993), *Reference to Abstract objects in Discourse*, Kluwer Academic Publishers, Dordrecht.

BILHAUT F. (2006), *Analyse automatique de structures thématiques discursives - Application à la recherche d'information*. Thèse de doctorat, Université de Caen.

BILHAUT F., HO-DAC L.-M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PERY-WOODLEY M.-P. et SARDA L. (2003), « Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique », in DAILLE B. (éd.), *Actes de TALN 2003*, ATALA IRIN, Batz-sur-Mer : 315-320.

BLACHE Ph. (2005), « Property grammars: A fully constraint-based theory », in CHRISTIANSEN H., SKADHAUGE P.R. et VILLADSEN J. (éds), *Constraint Solving and Language Processing*, volume LNAI 3438 : 1-16.

[9] http://www.linguastream.org

CHAROLLES M. (1997), « L'encadrement du discours : Univers, champs, domaines et espaces », in *Cahier de Recherche Linguistique* 6 : 1-73.

FERRARI S., BILHAUT F., WIDLÖCHER A. et LAIGNELET M. (2005), « Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels », in *Actes des 4èes Journées de Linguistique de Corpus*, Lorient (à paraître).

GROSZ B. J. et SIDNER C. L. (1986), « Attention, Intentions, and the Structure of Discourse », in *Computational Linguistics* 12 (3) : 175-204.

HEARST M. (1994), « Multi-paragraph segmentation of expository text », in *Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces : 9-16.

LAPPIN S. et LEASS H. (1994), « An algorithm for pronominal anaphora resolution », in *Computational Linguistics* 20 (4) : 535-561.

MANN W. C. et THOMPSON S. A. (1987), *Rhetorical Structure Theory : A theory of Text Organization*. Rapport interne ISI-RS-87-190, Information Sciences Institute, Marina del Rey, CA.

TANGUY L. (1997), *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle théorique informatique de la sémantique interprétative*, Thèse de doctorat, Université de Rennes 1.

TEUFEL S. (1999), *Argumentative Zoning : Information Extraction from Scientific Articles*, PhD thesis, University of Edinburgh.

WIDLÖCHER A. et BILHAUT F. (2005), « La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus », in M. JARDINO (éd.), *Actes de TALN 2005*, Dourdan : 517-522.

WIDLÖCHER A. (2006), « Analyse par contraintes de l'organisation du discours », in MERTENS P., FAIRON C., DISTER A. et WATRIN P. (éds), *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*. Presses universitaires de Louvain, Louvain-la-Neuve : 367-376 (*Cahiers du Cental*, 2.1).

Cahiers du Cental

La collection « Cahiers du Cental » est une publication du Centre de traitement automatique du langage de l'Université catholique de Louvain
http://www.uclouvain.be/cental

Hors-série

Purnelle G., Fairon C. et Dister A. (éds) (2004), *Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, 2 vols, Presses universitaires de Louvain, Louvain-la-Neuve, 1219 p.

Cahiers du Cental

1. Didier J.-J., Hambursin O., Moreau Ph. et Seron M. (éds) (2006), *« Le français m'a tuer », Actes du colloque L'orthographe française à l'épreuve du supérieur*, Bruxelles 27 mai 2005, Presses universitaires de Louvain, Louvain-la-Neuve, v-113 p.
2. Mertens P., Fairon C., Dister A. et Watrin P. (éds) (2006), *Verbum ex machina, Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles* (2006), Leuven, 10-13 avril 2006, Presses universitaires de Louvain, Louvain-la-Neuve, 2 vols, xviii-951 p.

3.1. Fairon C., Klein J. et Paumier S. (2006), *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête « Faites donc de vos SMS à la science »*, Presses universitaires de Louvain, Louvain-la-Neuve, viii-123 p.

3.2. Fairon C., Klein J. et Paumier S. (2006), *Le corpus SMS pour la science. Base de données de 30.000 SMS et logiciel de consultation*, CD-Rom, Presses universitaires de Louvain, Louvain-la-Neuve, v-44 p.

4. Fairon C., Naets, H., Kilgarriff A. et G.-M. de Schryver (éds) (2007), *Building and Exploring Web Corpora, Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, Presses universitaires de Louvain, Louvain-la-Neuve, viii-167 p.
5. Constant M., Dister A., Emirkanian L. et Piron S. (éds) (2008), *Description linguistique pour le traitement automatique du français*, Presses universitaires de Louvain, Louvain-la-Neuve, v-246 p.