Analyse par contraintes de l'organisation du discours

Antoine Widlöcher

Université de Caen - Basse-Normandie, GREYC - CNRS UMR 6072
awidloch@info.unicaen.fr

Résumé

Nous abordons ici 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.

Mots-clés : analyse du discours, formalisation du discours, approche par contraintes.

Abstract

We focus on the problem of discourse structure analysis from the point of view of its formal description and automatic processing. We intend to test the hypothesis that constraint-based approaches could enable various structures and cues of their presence to be taken into account. The CDML formalism which is here introduced allows such an approach.

Keywords: discourse analysis, discourse formalisation, constraint-based approach.

1. Introduction

Des travaux récents au sein de la communauté TAL révèlent un intérêt croissant pour l'analyse de la structure du discours. Certaines théories envisagent celle-ci dans le but de la décrire et de la formaliser, en privilégiant souvent un niveau de granularité relativement réduit, inter-propositionnel ou inter-phrastique. Cependant, différentes questions se posent à présent. Qu'en est-il du traitement automatique de telles organisations discursives ? Est-il envisageable d'appliquer ces modèles à d'autres échelles ? D'autres travaux privilégient au contraire certaines tâches opérationnelles telles que la segmentation automatique et considèrent le discours à un niveau de granularité plus élevé. Mais comment donner une description formelle des structures discursives ainsi envisagées ?

Nous interrogeons ici la description linguistique *et* l'analyse automatique de l'organisation du discours. Nous présentons tout d'abord certains problèmes fondamentaux liés à cette double perspective. Puis, nous introduisons CDML (*Constraint-based Discourse Modeling Language*), un formalisme déclaratif, descriptif et prescriptif reposant sur une modélisation par contraintes des structures discursives.

2. Analyse du discours

Prenons tout d'abord la mesure de la diversité des approches possibles. Différents travaux tels que le *text-tiling* (Hearst, 1994) visent 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. Des travaux tels que l'*argumentative zoning* (Teufel, 1999) considèrent le discours d'un point de vue argumentatif et visent l'identification de zones textuelles correspondant à différentes intentions rhétoriques des auteurs. Un autre point de vue sur l'organisation textuelle, représenté par la *Rhetorical Structure Theory* (Mann et Thompson, 1987), privilégie les *relations* entre les énoncés. D'autres travaux, tels que (Lappin et Leass, 1994), dédiés à l'anaphore, mettent l'accent sur certaines relations particulières (anaphoriques) entre des éléments distants spécifiques. Enfin, des travaux sur l'isotopie tels que (Tanguy, 1997) mettent l'accent sur la récurrence de propriétés sémantiques et sur les relations entre les éléments porteurs de ces propriétés.

2.1. Conditions d'une approche générique

Niveau de granularité : Parmi les différentes approches, certaines privilégient la description à un niveau relativement réduit, inter-propositionnel ou inter-phrastique. D'autres, souvent motivées par l'objectif d'une segmentation du texte, considèrent celui-ci à un niveau plus élevé : phrases, paragraphes... Pour notre part, nous visons la prise en compte de ces différents niveaux, et la définition d'un formalisme aussi indépendant que possible de cette granularité.

Segments et relations : D'autre part, nous pouvons distinguer les approches *orientées segments* et les approches *orientées relations*. Les premières considèrent les unités textuelles comme les clefs de voûte de l'organisation du discours, quand les secondes privilégient les relations entre les éléments textuels. 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.

Marques et indices : Les différents modèles envisagés ci-dessus mettent en lumière la remarquable diversité des indices utilisables pour la détection de telles structures. Ne considérons ici que certaines d'entre elles, dont la taille (caractère, mot, phrase...) et dont la nature (morphologique, syntaxique, sémantique...) peuvent varier. L'organisation discursive peut être révélée par des *connecteurs* explicites (rhétoriques...) ou par des *cue-phrases* caractéristiques. D'autres *formes de surface* pourront également être utilisées, par exemple pour estimer une cohésion lexicale. D'autres objectifs, comme la recherche d'isotopie ou l'analyse de la coréférence, exigent 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 linguistiques.

L'exemple des structures énumératives (figure 1[1]) éclairera cette discussion. Trois *segments* consécutifs (5) (7) (10) introduisent trois zones géographiques (6) (8) (11) 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 région. Ainsi, une *relation de spécialisation* lie l'amorce et les items. Enfin différentes *cue-phrases* (9) et (12) révèlent des *relations de similarité* (15) (16) entre (5) et (7) et entre (7) et (10). Cet exemple illustre la

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

dimension *relationnelle* de l'organisation du discours en plus de sa composition en segments, il éclaire la possibilité des constructions imbriquées et met en lumière la variabilité du grain (la taille d'un élément n'est pas donnée *a priori*) et la possible distance entre les indices pertinents.

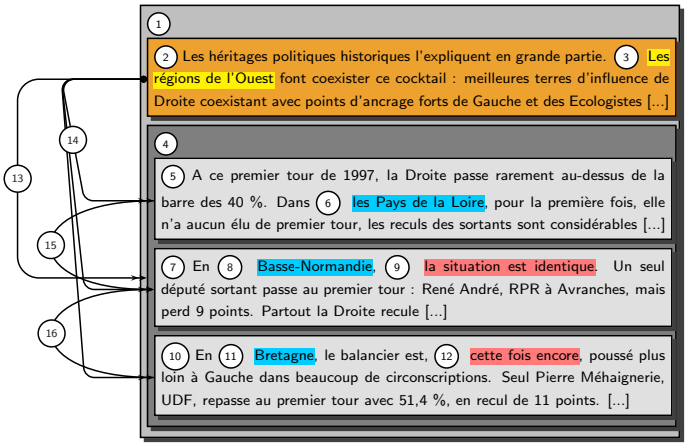

Figure 1. Un exemple de structure énumérative

2.2. Une approche descriptive et opératoire

Si nous considérons la nature de la description du discours proposée par les différentes approches, nous pouvons encore distinguer deux tendances opposées. Une formalisation purement descriptive de la structure du discours n'implique pas nécessairement que soient précisées les *conditions d'identification* 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* et conséquemment à perdre la pure description formelle (par exemple *grammar-like*). Le formalisme décrit ici propose de (re)concilier les approches *descriptives* et *prescriptives* (*i.e. opératoires*). Il offre un moyen, d'une part de décrire de manière formelle certains phénomènes discursifs, et d'autre part d'exprimer les conditions d'identification permettant de paramétrer un analyseur.

2.3. Spécificités de l'analyse du discours

Les différentes contraintes présentées ci-dessus exigent une méthodologie appropriée, tant en ce qui concerne les choix de formalisation, qu'en ce qui concerne les possibilités d'implémentation qui en résultent. Dans la mesure où nous souhaitons concilier les aspect formels-descriptifs et opératoires-prescriptifs, nous sommes conduits à considérer la possibilité d'une approche en termes de *grammaire*. Ce serait cependant une erreur de concevoir l'analyse du discours comme celle, plus traditionnelle, de la phrase, et ce pour trois raisons principales.

Niveau de granularité : Le niveau de granularité n'étant pas connu *a priori* et un même phénomène discursif pouvant être observé à différents niveaux, le formalisme proposé devra permettre la variation ou la non spécification du grain d'analyse.

Non-linéarité : 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.

Non-séquentialité : De plus, l'ordre dans lequel les éléments apparaissent n'est pas nécessairement significatif.

Ces propriétés exigent que l'approche *grammar-like* soit considérée avec précaution. Si nous nous autorisons à parler de *grammaire de discours*, nous insistons donc sur les points suivants. Tout d'abord, bien entendu, aucune prétention à la générativité ne saurait y être associée : nos grammaires de discours sont *orientées détection*. De plus, la notion ne présuppose ici ni séquentialité ni linéarité. Pour répondre à ces exigences, nous adoptons une *méthode à base de contraintes*. On pensera ici en particulier aux *Grammaires de Propriétés* qui, dans l'univers 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.

3. CDML

CDML 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.

3.1. Modèle sous-jacent du discours

Le modèle du discours sur lequel s'appuie CDML distingue trois « éléments » fondamentaux. Les **Unités Discursives (DU)** correspondent à des zones textuelles délimitées, dont le niveau de granularité peut varier, pour lesquelles certaines contraintes sont satisfaites. Items et énumération constituent de telles unités. Les **Relations Discursives (DR)** correspondent à des relations entre des DU et peuvent être définies par un ensemble de contraintes sur ces DU et sur leurs rapports. Les relations de similarité sont de telles DR. Les **Schémas Discursifs (DS)** correspondent à des *patterns* discursifs de plus haut niveau, définis par un ensemble de contraintes sur les DU et les DR existant entre ces DU. La structure énumérative constitue un tel motif.

3.2. Représentation du discours

Le discours est ici considéré comme une succession d'*objets de discours* ($\mathcal{DO}$) sur lesquels les contraintes seront exprimées. Leur taille peut varier, de même que leur statut linguistique : unités morpho-syntaxiques, éléments syntagmatiques... Imbrications et chevauchements sont possibles. Ils représentent toute l'information disponible à un certain stade de l'analyse, information pouvant émaner de toute analyse préalable. Chaque $\mathcal{DO}$ est associé à une *structure de traits* (notée FS pour *feature-set*) qui en représente l'information pertinente (statut morpho-syntaxique, interprétation sémantique...). Les $\mathcal{DO}$ entrent par ailleurs dans un ensemble de relations (syntaxiques...) 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 la notation suivante :

$\{a : \{b : X,\ c : \{d : Y,\ e : Z\}\}\}$ pour la structure $\left|\begin{array}{ll} a : & \left|\begin{array}{ll} b : & X \\ c : & \left|\begin{array}{ll} d : & Y \\ e : & Z \end{array}\right| \end{array}\right| \end{array}\right|$

3.3. Grammaires CDML

Une grammaire CDML est composée d'un ensemble de *règles*. Chaque règle 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). À chacun de ces éléments est associé un type de règle dédié. Nous nous limitons ici aux DU et aux DR, les DS étant toujours en cours de formalisation.

Chaque règle est composée d'un ensemble d'appels de contraintes. Les contraintes disponibles dépendent du type d'élément visé. La structure fondamentale d'une règle est :

```
RuleType:
    constraint1
    constraint2
    ...
```

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
```

La satisfaction de la règle génère un objet de discours ($\mathcal{DO}$) qui pourra être immédiatement 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 disponibles dépendent du type de règle. Elles constituent un ensemble extensible de *primitives discursives* appelées pour filtrer certains candidats d'un espace de recherche. Un appel de contrainte 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. Leur ordre est indifférent et leurs valeurs sont typées. Les contraintes portent fréquemment sur les FS associés aux $\mathcal{DO}$, par le biais de paramètres précisant les motifs recherchés. L'exemple d'une contrainte assez intuitive illustrera cette idée. La grammaire :

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

décrit et accepte n'importe quelle unité textuelle qui *commence* par un object de discours ($\mathcal{DO}$) dont le FS *unifie* avec `{type:"sentence"}`.

De plus, toutes les contraintes peuvent être préfixées par l'opérateur `not`, afin d'obtenir le complémentaire de l'ensemble des candidats filtrés.

Le *matching* entre les FS est établi par *unification*. Nous distinguons *unification standard* (notée ici $\sim$) et *unification forte* (notée ici $\approx$). Contrairement à l'unification standard, l'unification forte (ou *filtrage*) considère l'un des éléments unifiés comme le *modèle* auquel l'autre doit se conformer. Ainsi, $\{a : b\} \sim \{c : d\}$, mais $\{a : b\} \not\approx \{c : d\}$. La comparaison entre un *pattern* contraignant et le FS associé à un $\mathcal{DO}$ opère par unification forte. Ainsi :

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

accepte à la fois les $\mathcal{DO}$ représentés par `{type:"sentence"}` et `{type:"sentence", size:"short"}`. Mais :

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

n'accepte que les seconds. À l'inverse, le mécanisme d'unification de variables utilise l'unifi-

cation standard et est déclenché implicitement et automatiquement, pour chaque occurence de variable. La grammaire suivante :

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

identifie les segments textuels commençant et se terminant par des phrases de « même taille ». Notons que l'unification standard permet ici la *remontée d'information*.

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

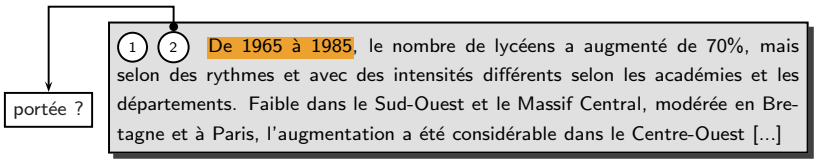

Figure 2. Un exemple de cadre de discours temporel[2]

L'hypothèse psycho-linguistique de l'encadrement du discours (Charolles, 1997) identifie des segments textuels, appelés *cadres de discours* (1), homogènes du point de vue d'un critère sémantique d'interprétation fixé dans une expression en position détachée, nommée introducteur de cadre (2). L'opérationnalisation en TAL de ce modèle peut être décomposée en trois tâches : analyse des expressions temporelles, identification de la fonction d'introducteur de certaines d'entre elles et détermination de la portée des introducteurs (Bilhaut *et al.*, 2003). 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 désigné par l'introducteur et les autres expressions temporelles contenues dans le cadre fournit un critère décisif de fermeture. Enfin, des indications *structurelles* doivent être prises en compte : un cadre n'est composé que de phrases complètes. Considérons à présent la grammaire CDML 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 $\mathcal{DO}$ 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 des relations de comparaisons définies par le `Comparator` nommé `scope` sont vérifiées. L'analyse préliminaire des verbes est supposée

[2] Extrait de : R. HÉRIN. et R. ROUAULT. *Atlas de la France scolaire de la maternelle au lycée*.

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 filtre 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"}`.

3.5. Formalisation

Nous abordons ici exclusivement la formalisation liée aux unités (DU), en nous limitant aux contraintes utilisées ci-dessus, afin d'indiquer notre méthodologie.

Définitions

Le discours est considéré comme un ensemble d'objets de discours ($\mathcal{DO}$). Nous identifions un $\mathcal{DO}$ à son index o. Soit k le nombre de $\mathcal{DO}$ présents dans le discours $\mathcal{D}$.

Soit $\mathbb{F}$ l'ensemble des structures de traits. Pour un objet de discours donné i, sa structure de traits est obtenue par $fs(i)$.

Une *séquence d'objets de discours* ($\mathcal{DOS}$) est un ensemble de $\mathcal{DO}$ consécutifs. Le discours complet est la $\mathcal{DOS}$ maximale et $|\mathcal{D}| = k$.

Une *unité candidate* u est un intervalle de $\mathcal{D}$. Cet intervalle est une $\mathcal{DOS}$ pouvant être identifiée par ses bornes, *i.e.* son premier et son dernier $\mathcal{DO}$.

$$u \in \{[a,b] \mid a \in [1,k], b \in [a,k]\}$$

Pour une u donnée, ses bornes de début et de fin peuvent être respectivement désignées par $s(u)$ et $e(u)$. Ainsi, par exemple, $fs(s(u))$ retourne le FS associé à la borne gauche de u.

Les unifications standard et forte entre x et y sont respectivement notées $x \sim y$ et $x \approx y$.

Une règle de description d'unité est composée d'un ensemble de contraintes $\mathcal{C} = \{c_1, c_2, ..., c_n\}$. Nous distinguons les *contraintes simples*, notées $Constraint_{uc}(u, \overline{p})$, où uc indique que la contrainte porte sur une unité, où u désigne une unité et $\overline{p}$ un ensemble de paramètres, dont la satisfaction peut être déterminée pour une unité donnée ; et les *méta-contraintes*, notées $MetaConstraint_{uc}(\mathcal{U}, u, \overline{p})$, dont la satisfaction est également relative à un ensemble d'unités $\mathcal{U}$. Nous considérons ici, par simplification, qu'une seule méta-contrainte sera appliquée à la solution de toutes les contraintes simples et non à un sous ensemble d'entre elles. Soit $\mathcal{C}^s$ l'ensemble des contraintes simples et μ une méta-contrainte tels que $\mathcal{C} = \mathcal{C}^s \cup \{\mu\}$. Soient $\mathcal{S}_*$ et $\mathcal{S}$ des ensembles d'unités :

$$\mathcal{S}_* = \{u_i \mid \forall c_j \in \mathcal{C}^s, c_j(u_i, \overline{p}) \text{ est satisfaite}\}$$
$$\mathcal{S} = \{u_i \mid \mu(\mathcal{S}_*, u_i, \overline{p}) \text{ est satisfaite}\}$$

Nous dirons que $\mathcal{S}$ est une solution de $\mathcal{C}$: $\mathcal{S} \models \mathcal{C}$.

Contraintes sur les unités

La contrainte $\mathbf{Start_{uc}}$ vérifie qu'un candidat débute par un $\mathcal{DO}$ *matchant* un *pattern* donné $p \in \mathbb{F}$ selon les conditions suivantes :

$$Start_{uc}(u, p) \equiv (fs(s(u)) \approx p)$$

La contrainte $\mathbf{AbsolutePresence_{uc}}$ ou $\mathbf{AP_{uc}}$ vérifie que le nombre d'instances matchant un pattern donné $p \in \mathbb{F}$ est supérieur ou égal à une limite fixée $q \in \mathbb{N}$:

$$S = \{i \mid s(u) \leq i \leq e(u), fs(i) \approx p\}$$
$$AP_{uc}(u, p, q) \equiv |S| \geq q$$

La contrainte $\mathbf{Longest_{uc}}$ est une méta-contrainte filtrant les unités imbriquées appartenant à un ensemble S, pour ne conserver que les plus longues.

$$Longest_{uc}(S, u) \equiv \left|\{v \mid v \in S, u \subsetneqq v\}\right| = 0$$

La contrainte $\mathbf{Homogeneity_{uc}}$ ($\mathbf{H_{uc}}$) est satisfaite si les conditions indiquées par un comparateur[3] le sont. Celui-ci est composé d'un ensemble de *signatures*. Chacune définit une condition de comparaison $cond$ pour deux $\mathcal{DO}$ matchant des patterns donnés. Pour un comparateur c, une signature est notée $(cond, fs_1, fs_2)_c$ avec $fs_1, fs_2 \in \mathbb{F}$. Si la condition $cond$ est satisfaite pour les $\mathcal{DO}$ o_a et o_b matchant les patterns fs_1 et fs_2, nous écrivons $(o_a, o_b) \models (cond, fs_1, fs_2)_c$. La contrainte d'homogénéité vérifie que tous les $\mathcal{DO}$ matchant les patterns indiqués par les différentes signatures satisfont les critères de comparaison correspondants.

$$H_{uc}(u, c) \equiv \forall (cond, fs_1, fs_2)_c, \forall (o_a, o_b) \text{ tels que } \begin{array}{r} o_a, o_b \in u, (cond, fs_1, fs_2)_c \in c \\ fs(o_a) \approx fs_1, fs(o_b) \approx fs_2 \\ (o_a, o_b) \models (cond, fs_1, fs_2)_c \end{array}$$

3.6. Mise en œuvre

La mise en œuvre que nous proposons, sur la base de cette formalisation, s'appuie sur un mécanisme de satisfaction de contraintes retenu pour sa généricité et pour la facilité d'introduction de nouvelles contraintes qui en découle, une optimisation pouvant par la suite être envisagée. Il consiste fondamentalement dans la construction puis le filtrage d'un espace de recherche. Dans la mesure où cet espace est composé de $\mathcal{DOS}$, une génération naïve (mais parfois inévitable) de son état initial produit $(n(n+1)/2)$ candidats avec n le nombre de $\mathcal{DO}$. Afin d'améliorer la procédure de résolution, nous tirons bénéfice de certaines contraintes (telles que $Start_{uc}$) permettant d'améliorer son initialisation. Avant d'appliquer l'ensemble des contraintes, le système recherche une telle contrainte favorable, et l'applique en premier lieu, pour produire l'espace de recherche minimal auquel les autres contraintes seront appliquées.

Notre implémentation prend la forme d'un composant pour la plate-forme Linguastream[4] et tire avantage de ses principes (Widlöcher et Bilhaut, 2005). Plate-forme générique dédiée au TAL, LinguaStream permet la mise en œuvre de chaînes de traitement procédant à l'enrichissement incrémental de documents électroniques, par une succession de composants de type et de niveau variés : morphologique, syntaxique, sémantique... Chaque étape découvre et produit de nouvelles informations sur lesquelles les étapes ultérieures peuvent s'appuyer. LinguaStream permet ainsi la réalisation d'expérimentations complexes sur corpus, en utilisant dès que possible des formalismes déclaratifs (favorables à la capitalisation du savoir linguistique) et en offrant un accès unifié aux annotations produites par les différents analyseurs. CDML et son implémentation souscrivent à ces principes fondamentaux.

[3] Ce mécanisme est utilisable par d'autres contraintes.
[4] http://www.linguastream.org.

Afin de procéder à l'analyse du discours, celui-ci doit être lu et représenté. Le corpus, les marquages et les annotations produits par les différents analyseurs utilisent les technologies XML. Le *parsing* de documents XML repose traditionnellement sur des méthodes *SAX-like* ou *DOM-like*. Les premières s'appuient sur une lecture événementielle, peu gourmande en mémoire, consistant à déclencher les traitements au fil de la lecture. Les secondes au contraire passent par la représentation en mémoire de l'intégralité du document. Comme les contraintes CDML peuvent être non séquentielles et non linéaires, une approche *SAX-like* est évidemment contre-indiquée ici. Une approche *DOM-like* satisferait à l'évidence ces besoins, mais la représentation en mémoire de la totalité du discours n'est pas souhaitable. En conséquence, nous proposons une méthode hybride nommée DOMBMS (*DOM-Based Markup System*). L'utilisateur définit une *Maximal Relevant Unit* (MRU) correspondant à la taille d'unité au-delà de laquelle aucune représentation *DOM-like* n'est nécessaire. Par exemple, nous pouvons décider que les structures à observer seront toutes comprises dans le paragraphe. DOMBMS génère alors une représentation DOM partielle pour chaque MRU, représentation rendue dès lors accessible de manière événementielle, *SAX-like*, la mémoire étant libérée au fur et à mesure.

Le système CDML repose sur un parseur généré par ANTLR[5] pour la lecture des grammaires et sur un solveur de contraintes écrit en Java. L'implémentation des contraintes consiste dans la description XML de leurs syntaxes et dans l'implémentation Java de leurs méthodes de filtrage conformément à certaines interfaces. Il est aisé d'ajouter de nouvelles contraintes.

4. Évaluation

Si la nature de "méta-modèle", c'est-à-dire de cadre formel théorique pour l'expression de modèles linguistiques, du formalisme proposé rend assez 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 effectivement 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 résultats obtenus par cette dernière sont tout à fait comparables, en temps de calcul et en termes d'objets annotés, à ceux obtenus à l'aide d'un composant logiciel dédié, avec un temps de mise en œuvre considérablement réduit. D'un point de vue opérationnel, cette méthode s'avère donc tout à fait utilisable, sur des corpus de taille moyenne et dans des temps raisonnables, les dangers de l'explosion combinatoire étant largement modérés à la fois par la possible non linéarité (permettant d'ignorer certains éléments) et par le mécanisme des MRU (limitant la taille des candidats). À titre indicatif, le temps de traitement de l'analyse CDML de la portée des cadres est de l'ordre de la minute pour un corpus d'environ 65.000 mots, sur une station de travail standard. Insistons cependant sur le privilège accordé à la manipulation expérimentale. Une implémentation *ad hoc* d'un modèle élaboré à l'aide de CDML permettra vraisemblablement d'optimiser les temps de résolution.

5. Conclusion

Notre objectif principal était de trouver une manière unifiée de décrire et détecter les structures du discours, dont la diversité met en lumière d'importants problèmes méthodologiques et computationnels, et exige des outils descriptifs et opératoires adaptés. Une approche par contraintes peut satisfaire ces exigences, en permettant la prise en compte d'indices de natures variées, et en

[5] http://www.antlr.org.

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 et permet d'ores et déjà, par exemple, l'analyse des cadres temporels présentée ci-dessus.

Références

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

BLACHE P. (2005). « Property Grammars : A Fully Constraint-Based Theory ». In H. Christiansen, P. R. Skadhauge et J. Villadsen (éds.), *Constraint Solving and Language Processing*, volume LNAI 3438 : Springer. p. 1-16.

CHAROLLES M. (1997). « L'Encadrement du discours : Univers, champs, domaines et espaces ». In *Cahier de Recherche Linguistique*, 6.

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 4èmes Journées de Linguistique de Corpus*. Lorient, France. À paraître.

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, New Mexico, p. 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, ISI : 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*. PhD thesis, 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 (Traitement automatique des langues naturelles)* : LIMSI. ATALA, Dourdan, France, p. 517-522.