TALN 2004, Fès, 19-21 avril 2004

Spécification et implantation informatique d'un langage de description des structures discursives

Gustavo Crispino, Agata Jackiewicz et Jean-Luc Minel

Laboratoire LaLICC (Paris IV, CNRS) – Université de Paris-Sorbonne (ISHA)
96, bd Raspail, 75006 Paris -France
Prénom.Nom@paris4.sorbonne.fr

Résumé – Abstract

Cet article présente le langage de représentation des connaissances linguistiques LangTex qui permet de spécifier d'une manière unifiée les descriptions linguistiques nécessaires au repérage d'objets textuels qui organisent les textes écrits.

This article presents LangTex, a language for linguistic knowledge representation. This language allows to specify in a unified way the linguistic descriptions necessary for the location of textual objects which organize written texts.

Keywords – Mots Clés

Représentation des structures discursives, langage de représentation des connaissances linguistiques.
Representation of discourse structures, linguistic knowledge language representation.

1 Introduction

Le développement de nouveaux modes d'accès aux documents numériques, que nous connaissons aujourd'hui, conduit à mettre l'accent sur l'exploitation de la structuration interne des documents et de sa signalisation. Dans cette perspective informatique, focalisée en particulier sur la fouille sémantique de textes, nous nous intéressons aux différentes familles de marques d'organisation discursive à l'écrit (Jackiewicz, 2002 ; Couto et al., 2004 ; Jackiewicz, Minel, 2003). En nous appuyant sur plusieurs études linguistiques portant sur l'énumération (Turco, Coltier, 1988), (Péry-Woodley, 2000), (Luc, Virbel, 2001), la reformulation (Rossari, 1997), l'auto-représentation de l'énonciation (Authier-Revuz, 1995), la prise en charge énonciative (Desclés, Guentchéva, 2000), nous cherchons à proposer un ensemble de critères (organisés idéalement dans une carte de valeurs sémantiques) permettant de rendre compte d'une manière unifiée des principales opérations énonciatives et méta-énonciatives rencontrées dans des textes écrits. Ces critères sont destinés à une procédure

automatique permettant d'annoter certains segments textuels ou certains lieux ou modes d'articulation de ces segments, l'annotation pouvant être exploitée notamment pour la navigation textuelle ou intra-documentaire ou la synthèse sélective de textes (Couto et al., 2004). Nous disposons actuellement de deux études linguistiques abouties portant sur les marques organisatrices de discours. La première s'intéresse aux marqueurs d'intégration linéaire et aux structures organisationnelles en série que ces marques permettent d'engendrer (Jackiewicz, 2002). La deuxième décrit le fonctionnement discursif des introducteurs de thématiques (Porhiel 2001).

Dans (Jackiewicz, Minel, 2003) nous avons présenté une première modélisation pour le repérage et la représentation des structures en série. Cette modélisation a été réalisée dans le cadre de la mise en œuvre de la plate-forme d'ingénierie linguistique ContextO (Desclés et al., 1997 ; Crispino et al., 1999, Minel et al., 2001). Cette plate-forme, dédiée au filtrage sémantique de textes et au développement d'outils interactifs de fouille de textes s'organise en plusieurs sous-systèmes qui coopèrent autour d'un modèle du texte (Crispino, 2003). C'est celui-ci qui va être décoré avec les résultats produits par le moteur d'exploration contextuelle en appliquant les règles d'exploration contextuelle.

Notre travail actuel se centre sur la conception d'un langage de représentation qui permette de décrire les modèles des différentes structures discursives, afin de proposer un outil de description générique en vue d'une implémentation informatique dans la plate-forme ContextO.

2 L'objet textuel « série », sa structure typique et ses variantes

Afin de préciser la terminologie employée, considérons un exemple prototypique d'une structure organisationnelle en série [1]. Cette structure se compose d'une **amorce** et d'une série de segments proprement dite, s'enchaînant linéairement dans le texte. L'amorce introduit la série, généralement (i) en explicitant son principe fédérateur, à l'aide notamment d'un **classifieur** (ici : *champ*), et (ii) en indiquant sa longueur (ici : *trois*). Chaque item de la série constitue un cadre organisationnel ; il est introduit par une expression désignée par le terme de **marqueur d'intégration linéaire** ou **MIL** (ici, *en premier lieu*, *en deuxième lieu* et *en troisième lieu*). Ces trois marques fonctionnent de concert. *En premier lieu* est un marqueur d'**ouverture**. *En deuxième lieu* et *en troisième lieu* sont des marqueurs « **relais** ». Un relais indique que le segment textuel introduit n'est pas le premier de la série et qu'une suite est possible. La série de l'exemple [1] ne comporte pas de marqueur de **clôture,** dont les exemples possibles sont *en dernier lieu, enfin*. Chaque MIL balise le début de l'item textuel qu'il introduit. Chaque MIL possède également une **portée**, correspondant à toutes les unités (segments, propositions, phrases) qui forment l'item. Précisons que ni le début de l'amorce ni la fin du dernier item, ne sont signalés par des marques discursives.

[1] *Quatre-vingts ans après, l'Union soviétique a fait naufrage, et le monde connaît une nouvelle grande mutation,[…].* En premier lieu*, dans le domaine technologique. L'informatisation de tous les secteurs d'activités[…].* En deuxième lieu*, dans le domaine économique. Les nouvelles technologies favorisent l'expansion de la sphère financière.[…].* En troisième lieu*, dans le domaine sociologique. Les deux bouleversements précédents mettent à mal les prérogatives traditionnelles de l'Etat-nation […].(LMD Oct 97 p.1)*

Les séries textuelles balisées par des marques discursives sont typiquement : (i) constituées de 2 ou de 3 items (au maximum 10 items attestés) ; (ii) introduites par des MIL hétérogènes, souvent manquants ou relayés par d'autres marques (notamment des connecteurs argumentatifs) ; (iii) structurées à un niveau, plus rarement à deux niveaux ; (iv) annoncées par une amorce. Ces séries ne comportent pas de clôture explicite, mais peuvent (ce qui est rare) être terminées par une rétro-évaluation. Ce résumé des caractéristiques saillantes des structures organisationnelles en séries ne peut être donné que sous une forme fortement modalisée : certaines propriétés sont fréquentes et typiques, d'autres sont données pour rares, d'autres enfin peuvent être perçues comme des variantes accidentelles (voire, comme une atteinte à une certaine norme d'usage). L'existence de plusieurs variantes s'écartant de la structure typique de l'objet « série discursive » a des incidences significatives sur les plans formel et informatique. En effet, si les scripteurs respectaient certaines normes d'écriture, en indiquant systématiquement par des marques lexicales ou visuelles explicites et homogènes les rangs des items, les différents niveaux d'emboîtement, ainsi que les clôtures de ces structures, la reconnaissance de celles-ci relèveraient d'une grammaire hors-contexte (type 2). Mais d'une part, certaines marques peuvent être absentes et de plus, certaines marques sont ambiguës au sens où elles peuvent appartenir potentiellement à plusieurs structures discursives.

3 Langage de représentation de connaissances linguistiques

Un grand nombre de logiciels construits en TALN peuvent être utilisés pour identifier des structures discursives. Mais de notre point de vue, ces logiciels ne fournissent pas les outils, formels ou techniques, qui permettent d'opérer une distinction claire entre les descriptions linguistiques et les opérations de repérage. A ce titre, des langages ou des standards appartenant ou non à la famille XML comme Xslt, Xpath… ne répondent pas à nos exigences, au sens où d'une part, ils s'appuient sur des concepts opératoires limitatifs, d'autre part, leur syntaxe les rend impropres à une utilisation par des non-spécialistes. Le langage LangTex (Crispino, 2003) constitue une réponse à nos besoins. LangTex est un langage de représentation de connaissances linguistiques destiné à exprimer, à l'origine, des règles de la méthode d'exploration contextuelle. Il s'agit d'un langage déclaratif, structuré en deux couches : CBase et CRegEC. Les règles d'exploration contextuelle capables d'attribuer une valeur sémantique à un segment textuel, sont écrites avec des expressions de la couche de niveau supérieur, CRegEC. Ainsi, par exemple, une construction syntaxique telle que :

```
Nomregle = AmorceStruc ;
IndicateursDeclencheurs = &classifieurs ;
Etiquette = AmorceClassifieurLongueur ;
{Existe :
        [EspaceRecherche : gauche(indicateur,2)]
        [&numeral ;]}
attribuerEtiquette :phraseParent(indicateur) ;
finRegle
```

exprime les conditions pour étiqueter une phrase avec l'étiquette sémantique « AmorceStruc ». La phrase à étiqueter est celle qui contient un indicateur de la classe *&classifieurs* (*facteur*, *champ*, *hypothèse*…). La condition imposée dans cette règle pour l'attribution d'une étiquette est celle de trouver dans un espace de recherche de deux positions

à gauche de l'indicateur une expression de la classe *&numeral* (*deux*, *trois*, *quatre*, …). Cette construction présente deux expressions de la couche CBase du langage : *gauche(indicateur, 2)*, qui délimite un espace de recherche à gauche de l'indicateur et *phraseParent(indicateur)* qui indique que le segment à étiqueter est celui de la phrase qui contient l'indicateur. De son côté, la couche de niveau inférieur, CBase, fournit les fonctionnalités élémentaires concernant la structure de base du texte. CBase contient les opérateurs et les opérations qui permettent de naviguer dans la structure textuelle, de se positionner dans un segment textuel déterminé, d'extraire certaines unités lexicales et de vérifier la position d'une unité lexicale dans le texte. LangTex s'appuie sur une représentation hiérarchique du texte, modélisable par un graphe orienté.

Nous avons montré (Jackiewicz, Minel, 2003) que le repérage des structures discursives nécessite deux niveaux d'analyse. Un premier niveau vise à identifier la fonction discursive (ou sémantique) d'un marqueur. Un deuxième niveau d'analyse nécessite la notion de modèle de référence. Il conjugue le repérage issu du premier niveau avec des contraintes exprimées dans ce modèle, propre à la structure discursive que l'on cherche à identifier. Par exemple, pour les séries de cadres organisationnels, ce modèle stipule les contraintes suivantes. Premièrement, les éléments de la structure obéissent à une organisation séquentielle (*en second lieu* apparaît nécessairement après en *premier lieu* et avant en *troisième lieu*) sachant que certaines marques lexicales peuvent être absentes. L'ensemble de ces éléments constitue une série paradigmatique. Deuxièmement, deux séries sont soient disjointes, soient enchâssées. Ces contraintes sont utilisées par le processus de reconnaissance pour calculer la portée et l'enchâssement des structures discursives notamment afin de pallier l'absence de certaines marques typographiques ou lexicales. A chaque structure discursive est donc associé un modèle qui décrit son organisation paradigmatique, les contraintes qui régissent les éléments appartenant à la structure et celles qui régissent les structures entre elles. Ces contraintes sont propres à chaque modèle. Ainsi, l'étude menée sur les séries de cadres organisationnels (Jackiewicz, 2002) montre que ce type de structure peut présenter au plus deux niveaux d'imbrication, alors que le niveau d'enchâssement des cadres spatiaux ou temporels peut être supérieur à deux (Charolles, 1997).

L'identification d'une structure discursive se traduit dans la représentation hiérarchique par la création d'un chemin entre les unités phrastiques qui composent la structure. Ce qui signifie que dans la représentation présentée précédemment nous ajoutons des arcs lesquels, au lieu de modéliser la linéarité matérielle du texte, modélisent sa « linéarité » discursive.

4 Langage de description des structures discursives et architecture informatique

Dans cette section nous présentons les éléments nécessaires pour la modélisation des structures discursives. Cette modélisation touche plusieurs aspects. Un premier élément de la modélisation est la **caractérisation des structures à repérer et à représenter**. Pour illustrer cet aspect, considérons l'exemple de la structure organisationnelle en série, évoqué dans la section 2. Dans notre représentation du texte comme graphe orienté, les nœuds sont étiquetés, sous la forme de paires <attribut, valeur>, par des informations concernant la structure textuelle et par des valeurs sémantiques attribuées par des règles d'exploration contextuelle. Par exemple, un nœud correspondant à la deuxième phrase du texte, qui contient le troisième

MIL (relais), comportera, à l'issue de la première phase de la procédure de reconnaissance de structures, entre autres, les étiquettes suivantes :

<type = "phrase">, <position = "2">, < etiquetteStruc = "relais" >, <rang = "3">.

Cette structure, désignée SCO_{AOR}[1], sera représentée par un chemin dans le graphe constitué par les phrases qui contiennent les différents composants de la série. Formellement, soit *p* le nœud que représente une phrase, *p.att* l'attribut *att* du nœud *p* et *p.att[n]* l'attribut *att* du énième nœud d'une liste de nœuds. La structure SCO_{AOR} est donc une liste de nœuds phrase (p_1, …,p_n) telle que :

- existe un arc de chaque nœud vers le suivant, c'est-à-dire, $p_i \rightarrow p_{i+1}$ pour i = 1,..,n-1
- p.etiquetteStruc[1] = "amorce"
- p.etiquetteStruc[2] = "ouverture"
- $\forall$ i, $3 \leq i \leq n$, p.etiquetteStruc[i] = "relais",
- $\forall$ (i,j), $3 \leq i \leq n$, tel que p.position[i] < p.position[j] alors p.rang[i] < p.rang[j].

Le deuxième élément de notre modélisation concerne **l'extension du langage LangTex avec une couche CRegSD** (couche de règles de structures discursives). Les règles de cette couche traitent les étiquettes ajoutées à la représentation hiérarchique du texte pendant le premier niveau d'analyse des structures discursives. Les composants d'une règle de cette couche sont : l'élément déclencheur (p.etiquetteStruc = "amorce" dans notre exemple), les conditions et l'action. Le troisième élément de la modélisation est constitué par les **contraintes générales des structures**.

Les spécifications décrites précédemment nécessitent, pour être implémentées dans la plate-forme ContextO, des modifications dans l'architecture de celle-ci. En effet, jusqu'à présent (cf. section 1), l'application des règles d'exploration contextuelle avait seulement pour résultat la décoration de la représentation hiérarchique du texte. Cet étiquetage correspond au premier niveau d'analyse décrit précédemment (cf. section 3). Le développement informatique nécessité par les traitements du deuxième niveau d'analyse implique l'ajout de deux modules
Un premier module a pour charge la création des chemins (cf. section 4), c'est-à-dire des arcs nommés entre les nœuds représentant les phrases dans la représentation du texte. Ce module s'appuie d'une part, sur l'étiquetage réalisé par l'étiqueteur sémantique du premier niveau, et d'autre part, sur une base de modèles des structures discursives, où chaque modèle décrit les composants de celles-ci. Le repérage d'une structure doit faire appel à des heuristiques qui cherchent une configuration qui réponde à la fois à des contraintes locales et globales. Par exemple, en cas d'indétermination, une marque de clôture telle que *Enfin* est toujours attachée au cadre organisationnel le plus proche ; ce qui est une contrainte locale. La règle voulant que les cadres organisationnels soient toujours enchâssés ou disjoints est un exemple d'une contrainte globale. Ces contraintes spécifiées sous forme déclarative sont traitées par un deuxième module, le résolveur de contraintes, qui se charge d'explorer l'espace des solutions possibles et de proposer, s'il en existe, une solution préférentielle.

5 Conclusion

Le travail que nous venons de présenter comporte plusieurs composantes : l'étude linguistique, la modélisation d'une solution à l'aide d'outils formels de représentation et l'architecture informatique. Sur le plan linguistique, nous avons entrepris plusieurs recherches

[1] SCO_{AOR} : Série de Cadres Organisationnels avec Amorce, Ouverture et Relais

sur des organisateurs discursifs, ayant abouti, entre autres, sur un ensemble de critères permettant de caractériser la sémantique et le fonctionnement discursif de ces marques. Sur le plan de la modélisation, nous travaillons à l'extension du langage de représentation LangTex, pour décrire les modèles des différentes structures discursives, afin de proposer un outil de description générique.

Références

AUTHIER-REVUZ J. (1995), Ces mots qui ne vont pas de soi. Boucles réflexives et non coïncidences du dire, *t.1 et t.2, Larousse*.

CHAROLLES M. (1997), L'encadrement du discours - Univers, champs, domaines et espace , Cahier de recherche linguistique, 6.

COUTO J., FERRET O., GRAU B., HERNANDEZ N., JACKIEWICZ A., MINEL J.-L., PORHIEL S. (2004), RÉGAL, un système pour la visualisation sélective de documents, *Revue d'Intelligence Artificielle*, *à paraître*.

CRISPINO G., BEN HAZEZ S., MINEL J.-L. (1999), Architecture logicielle de Context, plate-forme d'ingénierie linguistique, Actes de *TALN 99*, Cargèse, pp 327-332.

CRISPINO G., (2003), *Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes*, Thèse de doctorat, Université de Paris-Sorbonne.

DESCLÉS J.-P., CARTIER E., JACKIEWICZ A., MINEL J.-L. (1997), Textual Processing and Contextual Exploration Method, Actes de *CONTEXT 97*, Rio de Janeiro, pp. 189-197.

DESCLÉS J.-P., GUENTCHÉVA Z. (2000), *Enonciateur, locuteur, médiateur*, ed. A. Becquelin et Ph. Erikson, *Les rituels du dialogue*, Editions de l'Harmattan.

JACKIEWICZ A. (2002), Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes, Actes de *CIFT'02*, Hammamet, Tunisie, pp. 95-107.

JACKIEWICZ A., MINEL J.-L. (2003), L'identification des structures discursives engendrées par les cadres organisationnels, Actes de *TALN 2003*, Batz sur Mer, pp 155-164

LUC C., VIRBEL J. (2001), Le modèle d'architecture textuelle, fondements et expérimentation, *Verbum*, t. XXIII, n°1, Cohérence et relations de discours à l'écrit.

MINEL J.-L., CARTIER E., CRISPINO G., DESCLÉS J.-P., BEN HAZEZ S., JACKIEWICZ A. (2001), Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText, *Technique et Science Informatiques*, n°3,, pp. 369-395.

PÉRY-WOODLEY M.-P., (2000), *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle,* Mémoire d'habilitation, Université de Toulouse-le Mirail.

PORHIEL S. (2001), *Linguistic expressions as a tool to extract thematic information*, Actes Corpus Linguistic, Lancaster University.

ROSSARI C. (1997), *Les opérations de reformulation*, Peter Lang.

TURCO G., COLTIER D. (1988), Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire, *Pratiques*, n°57.