Segmentation et structuration de textes procéduraux pour l'aide à la modélisation de connaissances : le rôle de la structure visuelle

Amanda Bouffier, Thierry Poibeau

Segmentation et structuration de textes procéduraux pour l'aide à la modélisation de connaissances : le rôle de la structure visuelle

Amanda Bouffier et Thierry Poibeau

Laboratoire d'Informatique de Paris-Nord, Université Paris-Nord, 99 avenue Jean-Baptiste Clément 93430 Villetaneuse, http://www-lipn.univ-paris13.fr, amanda.bouffier@lipn.univ-paris13.fr, thierry.poibeau@lipn.univ-paris13.fr

Résumé

Dans cet article, nous étudions le rôle de la structure visuelle pour la segmentation de textes procéduraux. Nous nous focalisons sur un type de textes procéduraux particulier : les Guides de Bonnes Pratiques médicales. Une étude linguistique effectuée sur ce corpus montre la pertinence ainsi que les limites des indices visuels, pour délimiter des ensembles conditions-actions, qui forment des unités sémantiques de base pour la segmentation. Cette étude a permis de définir une architecture modulaire qui exploite ces indices pour segmenter et structurer les textes.

Mots clés : Aide à la modélisation, Traitement Automatique des Langues, linguistique textuelle, textes procéduraux, Guides de Bonnes Pratiques

1. Introduction

Les textes procéduraux sont des textes qui ont pour objectif de prescrire des actions au vu de certaines conditions. Ils reçoivent une attention croissante en entreprise car ils ont des conséquences importantes en termes de sécurité et en termes légaux (respect des contraintes légales). Ils sont pourtant souvent peu lus ou peu adaptés aux conditions de travail effectives (situations d'urgence, habitudes de travail difficiles à modifier).

Dès lors, le développement de systèmes facilitant l'accès aux instructions présentes dans les textes de manière adaptée aux situations de travail représenterait un bénéfice incontestable. Ces systèmes ont pour objectif d'aider l'acteur à prendre les bonnes décisions au vu du contexte dans lequel il se trouve, c'est à dire de manière conforme aux instructions présentes dans les textes.

Pour construire ces systèmes, il est nécessaire de modéliser les textes. C'est particulièrement le cas dans le domaine médical, notamment pour les Guides de Bonnes Pratiques (dorénavant GBP). Il s'agit d'ensembles de recommandations relatives aux pratiques médicales, écrites par des autorités en matière de santé et adressées aux médecins afin d'uniformiser leurs pratiques. Ces documents restent hélas trop peu consultés.

La modélisation de ces documents, c'est-à-dire le passage du texte brut à un modèle structuré (une DTD spécifique appelée GEM constitue un standard *de facto* pour les GBP) reste une étape largement manuelle, donc coûteuse. L'objectif de ce travail est de fournir une aide à la modélisation en proposant une première représentation structurée des textes. La structuration consiste à isoler les unités textuelles qui correspondent aux conditions et aux actions et à lier ces unités entre elles (cf. FIG 1).

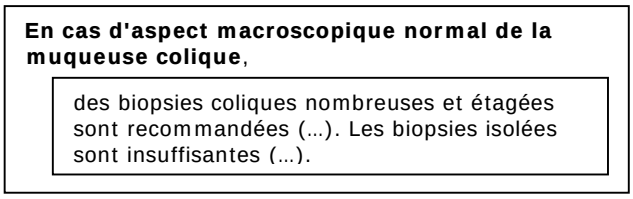

FIG.1-*Découpage d'un texte en cadres*

La tâche serait relativement simple si chaque condition était suivie dans la même phrase de l'action à laquelle elle est liée de manière linéaire. Cependant, différents phénomènes textuels rendent la tâche plus complexe. Il existe notamment des phénomènes de portée étendue concernant les unités exprimant une condition. La portée désigne le « champ d'action » d'une unité. Elle est au minimum égale à la phrase à laquelle appartient l'unité mais peut aussi s'étendre sur plusieurs phrases, comme le montre la figure (FIG.1).

Dans cet exemple, **en cas d'aspect macroscopique normal de la muqueuse colique**, est une condition qui doit être liée non seulement à l'action exprimée dans la même phrase (*des biopsies coliques…)* mais également à la remarque exprimée dans la phrase suivante (*Les biopsie isolées…*). La portée d'une unité-condition est représentée par un cadre. L'unité-condition est nommée introducteur de cadre. Nous disons qu'un cadre est minimal lorsqu'il ne dépasse pas la phrase de l'introducteur. Il est dit étendu lorsqu'il inclut plusieurs phrases.

La difficulté est de délimiter ces cadres, autrement dit de segmenter le texte en cadres, pour pouvoir ensuite

relier entre elles les différentes unités conditions et actions. Notre objectif est de proposer une méthode de segmentation fondée sur des indices textuels et pouvant être généralisable à l'ensemble des textes procéduraux. Nous nous concentrons dans cet article sur le rôle de la structure visuelle des textes (découpage en paragraphes, titres, énumérations etc.) pour la segmentation.

Dans la suite, nous présentons la tâche d'aide à la modélisation de textes procéduraux particuliers : les recommandations médicales. La troisième et quatrième partie sont alors dédiées à une étude basée sur ces textes, qui tente de cerner le rôle de la structure visuelle pour la segmentation en cadres. Enfin, des éléments d'implémentation et de validation sont présentés ainsi que des éléments de comparaison avec d'autres approches.

2 Contexte

Nous décrivons de manière plus détaillée dans cette partie le corpus d'étude et la segmentation envisagée.

2.1 Aider à la modélisation des GBP

Les GBP se présentent comme un catalogue de situations cliniques auxquelles sont associées des recommandations. Suite au constat que la simple diffusion de ces textes avaient peu d'impact sur les pratiques des médecins, de nombreux travaux dans le domaine de l'informatique médicale et de l'ingénierie des connaissances ont vu le jour. Ces travaux ont pour objectif de contribuer au développement d'outils d'aide à la décision fondés sur ces guides [15] afin d'améliorer l'observance des médecins aux pratiques thérapeutiques recommandées.

Pour construire les bases de connaissances venant alimenter ces systèmes, des travaux ont proposé des langages de représentation dédiés permettant d'élaborer le modèle formel [18]. Néanmoins ces langages supposent que les connaissances aient déjà été modélisées. Pour fournir une étape intermédiaire entre le texte brut et le modèle formel, un modèle de document a été proposé: GEM (Guideline Elements Model) [16]. Le modèle, basé sur une DTD XML, a pour but de représenter et structurer les éléments textuels pertinents pour la modélisation.

Même si cette étape intermédiaire représente déjà une aide à la modélisation, il n'en reste pas moins que le passage entre le texte et GEM est manuelle. Notre travail consiste à automatiser ce passage.

2.2 La tâche : du texte brut à une représentation basée sur GEM

L'objectif est de passer du texte à une représentation basée sur le modèle de GEM, comme l'illustre la figure (FIG.2). Nous nous focalisons sur certains éléments de la DTD qui permettent de structurer les recommandations en conditions et actions et laissons de coté (dans un premier temps en tout cas, car ces éléments sont moins centraux pour la tâche) des éléments comme ceux ayant traits à la méthodologie, au calcul du niveau de preuve etc.

Chez le sujet non immunodéprimé, **en cas d'aspect macroscopique normal de la muqueuse colique**, des biopsies coliques nombreuses et étagées sont recommandées (…). Les biopsies isolées sont insuffisantes (…).
L'exploration de l'iléon terminal est également recommandée (grade C). **En cas d'aspect normal de la muqueuse iléale (…)**, la réalisation de biospsies n'est pas systématique (accord professionnel).

Chez le sujet immunodéprimé, il est nécessaire de réaliser des biopsies systématiques (…)

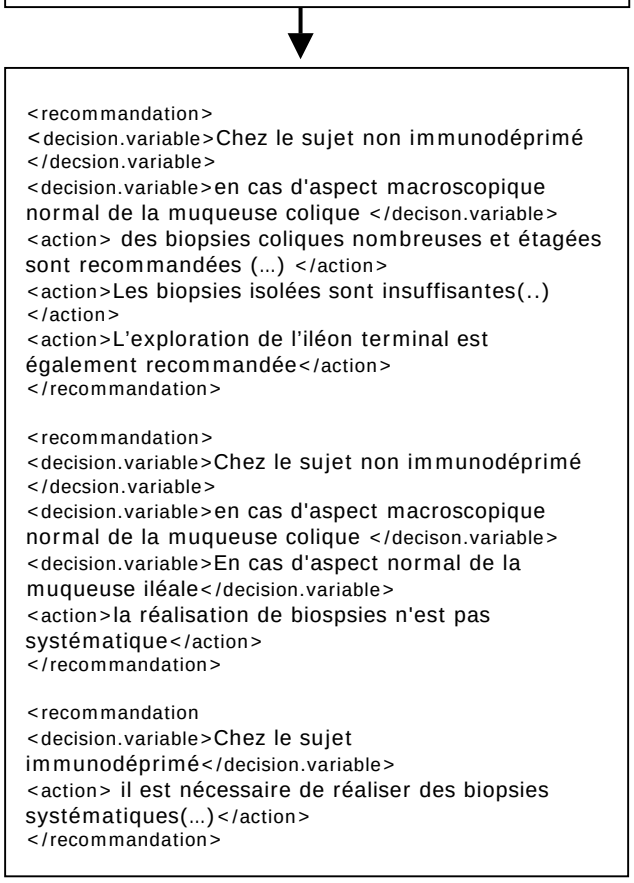

FIG. 2-*Du texte brut à une représentation basée sur GEM*

Dans cet exemple, sur un premier plan, le texte met en avant deux cas distincts (*Chez le sujet non immunodéprimé / chez le sujet immunodéprimé*). Une distinction de second plan réside dans l'imbrication d'un second cas (*En cas d'aspect normal de la muqueuse iléale*) dans le premier (*Chez le sujet non immunodéprimé*). On est donc en présence de trois cas, ce qui conduit à engendrer, au niveau de la représentation, trois unités « recommandation ». On peut remarquer que, dans le cas d'imbrication de cas, cette représentation peut se voir comme une mise à plat de ce qui est factorisé dans le texte.

2.3 Segmentation en cadres

Pour savoir quelles conditions sont reliées à quelles conditions et aboutir à la représentation présentée ci-dessus, la première étape consiste à segmenter le texte en cadres. Comme nous l'avons défini, un cadre représente la portée d'une unité-condition que l'on nomme introducteur de cadre (désormais introducteur). Nous nous concentrons dans un premier temps sur le problème du repérage de la fermeture des cadres, autrement dit, du calcul de la portée des unités-condition.

Le fait, pour une unité-condition, d'avoir une portée étendue (i.e supérieure à sa propre phrase), peut se réaliser de différentes manières sur le plan linguistique. Nous distinguons trois configurations possibles :

1. L'introducteur de cadre étendu est une expression non intégrée syntaxiquement à la phrase comme un titre ou certaines expressions détachées en début de phrase.

IV-En cas de diarrhée chronique

Chez le sujet non immunodéprimé, **en cas d'aspect macroscopique normal de la muqueuse colique**, des biopsies coliques nombreuses et étagées sont recommandées (…). Les biopsies isolées sont insuffisantes (…).
L'exploration de l'iléon terminal est également recommandée (grade C). **En cas d'aspect normal de la muqueuse iléale et compte tenu du (…)** , la réalisation de biospsies n'est pas systématique (accord professionnel).

Chez le sujet immunodéprimé, il est nécessaire de réaliser des biopsies systématiques (…)

FIG. 3-*Exemples de cadres introduits par un titre ou une expression détachée*

La figure (FIG.3) montre un exemple de cadre introduit par un titre (*IV-En cas de diarrhée chronique*) et plusieurs exemples de cadres introduits par des expressions détachées en début de phrase (toutes les autres expressions en gras).

La propriété d'être non intégré syntaxiquement à la phrase permet à l'introducteur d'avoir une portée étendue et par là même d'engendrer des unités homogènes, sans qu' aucune autre marque linguistique ne soit nécessaire pour rappeler cette homogénéité.

2. Dans d'autres cas l'introducteur est une expression intégrée syntaxiquement. Des marques linguistiques "relais" sont alors présentes pour signaler au lecteur qu'il se situe toujours dans le même cadre.

   Nous distinguons alors deux cas :

   a. L'introducteur est relayé par une expression anaphorique qui reprend l'introducteur et qui a pour but de signaler au lecteur que la proposition qui suit est toujours sous sa portée. La figure (FIG.4) montre un cas de ce type.

L'indication d'une insulinothérapie est recommandée **lorsque l'HbA1c est > 8%, sur deux contrôles successifs sous l'association de sulfamides/metformine à posologie optimale**, elle est laissée à l'appréciation par le clinicien du rapport bénéfices/inconvénients de l'insulinothérapie **lorsque l'HbA1c est comprise entre 6,6% et 8% sous la même association**. Dans les deux cas, la diététique aura au préalable été réévaluée et un facteur intercurrent de décompensation aura été recherchée (accord professionnel)

Stratégie de prise en charge du patient diabétique de type 2 à l'exclusion de la prise en charge des complications (2000)

FIG. 4-*Exemple de cadre introduit par une expression intégrée relayée par un anaphorique*

Dans cet exemple, *Dans les deux cas* est un anaphorique qui reprend les introducteurs (*lorsque l'HbA1c est >8% / lorsque l'HbA1c est comprise entre 6,6% et 8%)* et signale au lecteur que l'action *la diététique aura au préalable été réévaluée* tombe toujours sous la condition exprimée par les deux introducteurs.

   b. La portée étendue d'un introducteur s'établit de manière indirecte par une relation d'ordre temporelle entre deux actions, comme le montre la figure (FIG.5).

Un traitement médicamenteux de l'obésité ne doit être envisagé qu'**en cas d'échec des conseils diététiques**.
La poursuite de ce traitement au delà de 3 mois ne doit être envisagée que chez les patients répondeurs.

Stratégie de prise en charge du patient diabétique de type 2 à l'exclusion de la prise en charge des complications (2000)

FIG. 5-*Exemple de cadre établi de manière indirecte par une relation temporelle entre deux actions*

Dans cet exemple, l'introducteur *en cas d'échec des conseils diététiques*, bien qu'il soit intégré à la phrase, fait tomber sous sa portée non seulement l'action *un traitement médicamenteux de l'obésité doit être envisagé* mais également *La poursuite de ce traitement au delà de 3 mois ne doit être envisagée que chez les patients répondeurs.*

Le fait qu'il y ait une portée étendue n'est pas seulement dû au lien anaphorique entre les deux actions (*ce traitement*). En effet, on peut imaginer un lien anaphorique alors que l'on est dans deux cas différents (par exemple : *Un traitement médicamenteux de l'obésité doit être envisagé dans le cas x. Ce traitement peut être également proposé dans le cas y).* La portée est due au fait que la deuxième action (*La poursuite*

de ce traitement au delà de 3 mois) n'est rien d'autre que la première action (*un traitement médicamenteux de l'obésité doit être envisagé)*à un moment différent. Par conséquent, la population concernée par cette seconde action (i.e les patients répondeurs) représente nécessairement une sous-classe de la population de départ (i.e les personnes pour lesquelles le traitement non médicamenteux a échoué).

Pour délimiter la fin de ces cadres, de type différents, de nombreux indices interviennent comme le passage à un autre paragraphe, les constructions dites « parallèles » (*chez le sujet immunodéprimé / chez le sujet non immunodéprimé)* la présence de marqueurs discursifs (*En revanche, En effet*) etc. D'autres indices suggèrent la continuation d'un cadre, notamment les reprises et les marques anaphoriques.

Dans cet article, nous nous concentrons sur le rôle des indices relatifs à la structure visuelle du texte. Nous avons cherché à faire une estimation de la pertinence d'une segmentation exclusivement fondée sur ce type d'indices.

3 Segmenter en cadres : le rôle de la structure visuelle

La structure visuelle réfère à tous les procédés typo-dispositionnels utilisés pour structurer un texte : la structuration logique en paragraphes, les titres, les énumérations, la typographie etc. Nous faisons l'hypothèse qu'il s'agit d'un ensemble d'indices très pertinents pour les textes procéduraux. En effet, parce qu'ils doivent être lus rapidement et efficacement, ces textes sollicitent fortement la structuration visuelle.

Pour tester notre hypothèse, nous avons effectué une étude sur corpus dont l'objectif est de mettre en lumière les indices relatifs à la structure visuelle pertinents pour la segmentation, de connaître leur poids et la manière dont ils interagissent entre eux. Nous avons voulu ainsi estimer la pertinence mais aussi les limites d'une segmentation exclusivement fondée sur ce type d'indices.

3.1 Matériel

Le corpus qui a été utilisé dans cette étude comporte 18 GBP publiés par l'ANAES (*Agence Nationale d'Accréditation et d'Evaluation en Santé*) ou l'AFSSAPS (*Agence Francaise de Sécurité Sanitaire des Produits de Santé)* entre 2000 et 2005. Ils portent sur la prise en charge de diverses pathologies : diabète, hypertension, asthme etc. ou sur la pratique d'examens (endoscopie digestive basse). Ce corpus, homogène du point de vue du style d'écriture, représente environ 250 pages imprimables. On peut trouver ces guides sur http://www.anaes.fr ou http://affsaps.santé.fr.

3.2 Questions soulevées dans le cadre de l'étude

De manière préliminaire, on a voulu savoir s'il existe une corrélation entre le type de l'introducteur (expression intégrée ou détachée) et l'étendue de sa portée.

Concernant la structure logique (découpage en paragraphes), on a voulu savoir :

- S'il existe une corrélation entre les unités engendrées par les cadres et les unités engendrées par la structure logique.
- S'il existe une corrélation entre la position de l'introducteur dans le paragraphe et l'étendue de sa portée.

Concernant les titres et les énumérations, on a voulu savoir quel rôle pouvait jouer ces deux procédés de structuration pour la segmentation.

On a alors étudié les relations entre chacun de ces deux procédés de structuration et les introducteurs de type expression détachée. Plus précisément, on a voulu savoir :

- Quelles fonctions peut-on attribuer à ces trois procédés de structuration ?
- Sont-ils en interaction ou non ? Si oui, comment décrire cette interaction et quelles en sont les conséquences pour la segmentation ?

3.3 Méthode

500 introducteurs de cadre ont été isolés (titres, expressions détachées ou intégrées). Pour chaque introducteur, le cadre qu'il engendre a été délimité avec l'aide d'un expert (Catherine Duclos du laboratoire LIM&Bio de l'université Paris 13). Plusieurs paramètres, en relation avec la structure visuelle, jugés comme indices potentiellement pertinents pour la segmentation ont été retenus : la portée de l'introducteur par rapport à la structure logique, la position de celui-ci dans le paragraphe, la relation qu'il entretient avec le titre de la section dont il fait partie, la relation qu'il entretient avec l'énumération (dans le cas où il fait partie d'une énumération). Pour chaque introducteur, la valeur de chaque paramètre a été relevée. Pour tester la corrélation entre certains paramètres, un chi carré a été calculé.

Paramètres	Valeurs possibles
Type de l'introducteur de cadre	exp détachée, exp intégrée, titre
Portée de l'introducteur par rapport à la structure logique	= fin-phrase, <fin-paragraphe, = fin-paragraphe,> fin-paragraphe)
Position dans le paragraphe de l'introducteur	début, intérieur
Type du titre	condition, action, mixte
Relation titre/introducteur	redondance, divergence, sous-catégorisation

Relation énumération/introducteur	amorce, item

FIG. 6-*Liste des paramètres retenus pour l'étude*

Les paramètres retenus ainsi que l'ensemble de leurs valeurs possibles sont récapitulés dans le tableau de la figure (FIG.6).

3.4 Observations et commentaires

Les observations qui suivent sont structurées en 3 points :

Le rôle de la structure logique, le rôle des titres et le rôle des énumérations pour la segmentation.

Le rôle de la structure logique

La structure logique d'un texte correspond au découpage en chapitres, sections, paragraphes etc. Cette structure correspond à un arbre dont les feuilles sont les paragraphes élémentaires. On dira que deux unités sont au même niveau quand elles sont soeurs dans l'arbre.

Pour les expressions détachées, il existe une corrélation entre la segmentation en paragraphes et la segmentation en cadres. En effet, 60 % des introducteurs de ce type, engendrent un cadre qui se ferme à la fin du paragraphe dont l'introducteur fait partie. C'est donc l'unité du paragraphe qui domine dans ce cas. En revanche, 88% des expressions intégrées ont une portée qui se limite à leur propre phrase. C'est l'unité de la phrase qui domine dans ce cas.

Dans les 40 % des expressions détachées dont le cadre ne se ferme pas à la fin du paragraphe :

IV.1. Surveillance des maladies inflammatoires chroniques intestinales (maladie de Crohn et rectocolite ulcéro-hémmorragique*)*

En cas de colite chronique de type rectocolite ulcéro-hémmoragique ou maladie de Crohn,

Une surveillance endoscopique à la recherche de lésions néoplasiques est recommandée après 10 ans d'évolution en cas de pancolite (grade B) et après 15 ans d'évolution en cas de colite gauche (grade B), au rythme d'une coloscopie totale tous les 2-3ans (grade B).

Il est recommandé de réaliser des biopsies étagées tous les 10 cm de façon à obtenir un minimum de 30 biopsies (grade C).

En cas de dysplasie incertaine, un contrôle endoscopique avec biopsies est recommandé à 6 mois (accord professionnel).

Endoscopie digestive basse. Indications en dehors du dépistage en population (2004)

FIG. 7-*Un introducteur ayant une portée étendue à d'autres paragraphes*

- 17,5 % (6,8%du total) ont une portée qui dépasse la fin de leur paragraphes et inclut d'autres paragraphes de même niveau, comme dans l'exemple de la figure (FIG.7).

Dans cet exemple de la FIG 7, le cadre introduit par ***En cas de colite chronique de type rectocolite ulcéro-hémmoragique ou maladie de Crohn*** inclut les deux paragraphes de même niveau qui suivent. De manière surprenante il inclut le cadre introduit par ***en cas de dysplasie incertaine*** alors qu'ils sont au même niveau dans la hiérarchie visuelle.

Ce type de cas est intéressant car il montrerait une « contradiction » entre la structure logique et les cadres (déjà mise en lumière [19] sur le cas des énumérations non-parallèles). En effet, du fait qu'ils appartiennent à deux paragraphes de même niveau, ils s'excluent, tandis que du point de vue de la structure en cadres le deuxième est inclus dans le premier.

Cependant, il faut remarquer d'emblée que ce type de cas n'apparaît que dans une configuration précise : quand l'introducteur, de type expression détachée est redondant avec au moins une partie du titre. Dans notre exemple, on peut observer que ***En cas de colite chronique de type rectocolite ulcéro-hémmoragique ou maladie de Crohn***) reprend une partie du titre (*IV.1. Surveillance des maladies inflammatoires chroniques intestinales (maladie de Crohn et rectocolite ulcéro-hémmorragique*). L'expression détachée joue une sorte de « relais » par rapport au titre. La portée de l'expression détachée se confond dès lors avec celle du titre. Le rôle des titres pour calculer la portée d'une expression détachée est abordé dans le point suivant consacré aux titres.

Hormis ces cas particuliers qui ne représente que 6,8% des expressions détachées, aucun cas de portée étendue à d'autres paragraphes de même niveau n'a été trouvé. Ce résultat se comprend dans la mesure où une contradiction entre la structuration logique et la structuration en cadres auraient pour conséquence de perdre le lecteur.

- 46 % (18,1% du total, le reste fait référence à des cas où l'introducteur est lié à une énumération) ont une portée qui se situe avant la fin du paragraphe, c'est à dire entre la fin de la phrase et la fin du paragraphe. Nous avons alors cherché à savoir si des indices relatifs à des critères de positionnement dans le paragraphe n'étaient pas pertinents. Une corrélation entre la position de l'introducteur de type expression détachée dans le paragraphe (début ou intérieur du paragraphe) et l'étendue de sa portée a été cherchée mais aucun résultat significatif n'a été trouvé. En revanche, nous avons trouvé que beaucoup d' introducteurs qui à la fois se situaient à l'intérieur d'un paragraphe et n'étaient pas imbriqués dans d'autres cadres avaient souvent une portée limitée à la phrase. Peut-être est-ce parce que, dans ces cas, l'introducteur vient après une ou des

recommandations très générales et expriment un cas particulier. Il se trouve donc dans une relation de subordination, de dépendance vis-à-vis de ce qui le précède. La subordination tendrait alors à limiter la portée des introducteurs. Cette hypothèse demanderait à être testée sur un plus grand corpus.

Ces observations nous amènent à conclure à la pertinence de la structure logique comme indice de segmentation pour les expression détachées. En effet, il existe une correspondance assez forte, bien que non univoque, entre la segmentation en paragraphes et la segmentation en cadres pour ce type d'introducteurs.

Comme nous avons affaire à un indice très discriminant, nous proposons une segmentation par défaut correspondant au paragraphe dans le cas des expressions détachées et à la phrase dans le cas des expressions intégrées. Cette segmentation par défaut est remise en cause seulement si d'autres indices sont présents. Les titres et les énumérations font partie de ces indices pertinents qui sont examinés ci-après.

Le rôle des titres

Les titres et les expressions détachées sont deux procédés de structuration qui interagissent ensemble. En étudiant cette interaction, il a été observé que les titres peuvent jouer le rôle d'indice pour repérer la fermeture d'un cadre engendré par une expression détachée.

Un typage manuel des titres a montré que ceux-ci peuvent introduire trois aspects :

1. Dans un premier cas, les titres font référence à un type d'acte médical. Par exemple :

   Interrogatoire et examens initiaux
   (…)
   Place des examens morphologiques endoscopiques
   (…)

Le texte de la section va alors annoncer quels sont les actes médicaux, relatifs à la catégorie annoncée dans le titre, à appliquer dans tel et tel cas. Ils représentent 33% du total.

2. Les titres peuvent également exprimer une condition et être à ce titre des introducteurs de cadres qui nous intéressent. Par exemple :

   En cas de diarrhée chronique
   (…)
   II.2.2 Hypertension artérielle
   (…)

Ils représentent 12 % du total.

3. Ils peuvent combiner ces deux aspects. Par exemple :

   IV.1. Surveillance des maladies inflammatoires chroniques intestinales
   Ce cas représente 30 % du total (le reste concerne des titres autres : introduction, objectifs etc.*)*

Les titres de type 2 et 3 nous intéressent pleinement car, dans la mesure où ils expriment une condition, ce sont des introducteurs au même titre que les expressions intégrées ou détachées.

Du fait d'être tous les deux des introducteurs , les titres et les expressions détachées (nous laissons ici de coté les expressions intégrées) peuvent donc assumer la même fonction. Dès lors, comment interagissent-ils ?

Ce qui était attendu est un fonctionnement exclusif entre les titres et les expressions détachées, autrement dit il était attendu qu' ils ne puissent pas assumer la même fonction au même moment. Si on suit cette hypothèse, deux cas de figure sont possible : soit le titre introduit une dimension « action » (un type d'acte médicale) et alors la première expression détachée vient préciser les conditions sous lesquelles ce type d'acte doit être réalisée, soit le titre exprime une dimension « condition » et alors les expressions détachées viennent sous catégoriser la condition exprimée dans le titre (introduire des sous-cas du cas général annoncé par le titre). Mais en aucun cas, il n'est attendu qu'une expression détachée vienne reprendre la condition exprimée dans le titre. Les faits vont à l'encontre de cette hypothèse : les titres et expressions détachées ne fonctionnent pas de manière exclusive. Au contraire, dans 51 % des cas, le premier introducteur de type expression détachée présent après le titre est redondant totalement ou partiellement avec celui-ci, comme en témoigne l'exemple de la figure (FIG.8).

IV.1.Surveillance des maladies inflammatoires chroniques intestinales (maladie de Crohn et rectocolite ulcéro-hémmoragique)

En cas de colite chronique de type rectocolite ulcéro-hémorragique ou maladie de Crohn, une surveillance endoscopique à la recherche de lésions néoplasiques est recommandée après 10 ans (…).

Endoscopie digestive basse. Indications en dehors du dépistage en population (2004)

FIG. 8-*Un cas de redondance entre le titre et l'expression détachée*

Dans cet exemple, l'expression détachée ***En cas de colite*** reprend partiellement le titre. Nous faisons l'hypothèse que ce type de fonctionnement est lié aux textes procéduraux. D'une part, il est tout à fait compréhensible que les titres expriment souvent des conditions dans la mesure où c'est un point d'entrée pertinent pour le lecteur : l'information que possède le médecin est le cas qu'il a devant lui et l'information qu'il cherche est ce qu'il faut faire dans ce cas. D'autre part, l'expression détachée, quand elle reprend le titre est une sorte de relais et a pour fonction de rappeler au lecteur qu'il se situe dans le cas annoncé par le titre.

La conséquence de la redondance entre titres et expressions détachées est une confusion de la portée de l'expression détachée avec celle du titre. La similarité entre titre et expression détachée est alors un indice pertinent pour établir la portée de l'expression détachée. Cet indice permet de segmenter correctement le cas

évoqué à la figure (FIG.7) où la portée de l'expression détachée s'étendait sur plusieurs paragraphes de même niveau.

Le rôle des énumérations

Les énumérations sont très présentes dans les GBP. 37% des cas où la portée d'une expression détachée n'est pas égale à la fin du paragraphe sont liés à la présence d'une énumération. Il est donc important de pouvoir les prendre en compte. Si les énumérations peuvent être entièrement structurées par des moyens linguistiques, elles sont dans nos textes structurées par des moyens visuels (usages de puces notamment). Une énumération est composée d'une liste d'items qui est précédée, parfois, d'une amorce.

On peut distinguer deux cas où un introducteur interagit avec une énumération.

1. L'introducteur joue le rôle d'amorce de l'énumération, comme le montre l'exemple de la figure (FIG.9)

En cas de contrôle inacceptable :

- s'assurer qu'il s'agit bien d'un asthme, vérifier l'observance et la technique d'utilisation des dispositifs d'inhalation ;
- rechercher et traiter des facteurs aggravants, des pathologies associées ou des formes cliniques particulières ;

Recommandations pour le suivi médical des patients asthmatiques adultes et adolescents (2004)

FIG. 9-*Un introducteur amorce d'une énumération*

Dans cet exemple, **En cas de contrôle inacceptable** joue le rôle d'amorce de l'énumération. Dans ce cas, on sait que l'ensemble des items de l'énumération sont sous la portée de l'introducteur et que le cadre se ferme à la fin de l'énumération (i.e du dernier item). Repérer la fin de l'énumération revient donc à repérer la fin du cadre.

2. L'introducteur joue le rôle d'item de l'énumération, comme le montre l'exemple de la figure (FIG.10).

Malades ayant une corticothérapie inhalée et au moins un traitement additionnel :

- chez les malades sous CSI à dose faible et prenant un traitement additionnel, il est recommandé d'augmenter la dose de CSI ;

- chez les malades sous CSI à dose moyenne et prenant un traitement additionnel, il est recommandé d'augmenter la dose de CSI et d'ajouter un traitement additionnel ;

- (…)

Recommandations pour le suivi médical des patients asthmatiques adultes et adolescents (2004)

FIG. 10-*des introducteurs items d'une énumération*

Dans cet exemple, ***-chez les malades sous CSI à dose faible*** et ***-chez les malades sous CSI à dose moyenne*** sont items de l'énumération. Les items d'une énumération sont, par définition, au même niveau, c'est à dire qu'ils ne sont pas dépendants ni subordonnés les uns aux autres mais s'excluent les uns les autres. Dans ce cas, on sait que le cadre engendré par un introducteur se ferme au prochain item ou à la fin du dernier item. Par conséquent repérer les items d'une énumération revient à repérer la fin du cadre.

Dans la mesure où les introducteurs interagissent fréquemment avec les énumérations, ces dernières jouent donc le rôle d' indices pertinents pour la segmentation en cadres.

3.5 Bilan

Le tableau de la figure (FIG.11) récapitule la part que les différents indices relatifs à la structure visuelle qui viennent d'être abordés jouent pour la segmentation, selon le type de l'introducteur.

Type introducteur	§		**titres**		**enum**		**Struct visuelle**		**autres**		**Total**	
expression détachée	140	60,3	16	6,8	34	14,6	190	81,9	42	18,1	232	46,4
expression intégrée	19	11,5							145	88,4	164	32,8
titre											104	20,8

FIG. 11-*Tableau récapitulatif*

60 % des expressions détachées engendrent un cadre qui se ferme à la fin du paragraphe. Le découpage logique en paragraphes est un indice très discriminant et sert d'élément de découpage par défaut. Le repérage des autres types de structures (énumérations, titres…) doit malgré tout permettre d'améliorer la précision du découpage simple obtenu par simple détection des marques de fin de paragraphes. De manière globale, les indices relatifs à la structure visuelle sont très discriminants, même si on doit aussi prendre en compte ses limites.

4 Les limites d'une segmentation uniquement fondée sur la structure visuelle

La section précédente avait pour objectif de montrer que la structure visuelle donnait des indices très pertinents pour la segmentation en cadres dans les Guides de Bonnes Pratiques). Néanmoins, Pour 18% des expressions détachées et 11,5 % des expressions ces indices ne sont pas pertinents. On distingue deux cas :

1. Pour les expressions détachées, les cas qui ne peuvent être analysés correctement par des indices visuels correspondent aux cas où le cadre se ferme entre la phrase de l'introducteur et la fin du paragraphe.

2. Pour les expressions intégrées, il s'agit des cas où le cadre ne se ferme pas à la fin de la phrase de l'introducteur mais inclut plusieurs phrases.

Dans le cas 1, des marqueurs de relations sémantico-rhétoriques peuvent signaler la fermeture du cadre. Les marqueurs de relations de contraste tels que *En revanche, Cependant, néanmoins,* quand ils sont antéposés à un introducteur, sont pertinents pour marquer deux cadres incompatibles.

Les marqueurs de relations de type explicatif ou justificatif tels que *en effet, en fait, il en est pour preuve* sont également de bons indices de fermeture d'un cadre. En effet, la sémantique des introducteurs de condition interdit certains type d'énoncés comme les énoncés de type explicatif ou justificatif. Ce type d'introducteurs n'acceptent que des énoncés exprimant une action ou un état de fait. Par conséquent, les énoncés de type explicatif ou justificatif ne peuvent être sous la portée d'un introducteur (même si par ailleurs il peut être intéressant de les repérer). Par conséquent, ces marqueurs sont de bons indices de fermeture, comme le montre la figure (FIG.12)

Chez les patients ayant initialement une concentration très élevée de LDL-cholestérol, (…)

le prescripteur doit garder à l'esprit que la prescription de statine à fortes doses ou en association nécessite (…)

En effet, les fortes doses de statines et les bithérapies n'ont pas fait l'objet à ce jour d'une évaluation suffisante dans ces situations.

Prise en charge thérapeutique du patient dyslipidémique, (2005)

FIG.12-*Un marqueur de justification comme indice de rupture*

Dans le cas 2 (lorsqu' une expression intégrée étend sa portée au delà de sa propre phrase),les marqueurs de liens anaphoriques sont de très bons indices.

En effet, dans ce cas l'introducteur est relayé par une expression anaphorique. Celui-ci reprend l'introducteur et indique que la proposition qui suit est toujours sous sa portée, comme dans la figure (FIG.13)

Dans cet exemple, *Dans les deux cas* renvoie le lecteur aux deux conditions énoncées *auparavant (lorsque l'HbA1c est >8% / lorsque l'HbA1c est comprise entre 6,6% et 8%)* et indique que l'action qui suit tombe sous la portée de ces deux conditions.

L'indication d'une insulinothérapie est recommandée **lorsque l'HbA1c est > 8%, sur deux contrôles successifs sous l'association de sulfamides/metformine à posologie optimale**, elle est laissée à l'appréciation par le clinicien du rapport bénéfices/inconvénients de l'insulinothérapie **lorsque l'HbA1c est comprise entre 6,6% et 8% sous la même association**. Dans les deux cas, la diététique aura au préalable été réévaluée et un facteur intercurrent de décompensation aura été recherchée (accord professionnel)

Stratégie de prise en charge du patient diabétique de type 2 à l'exclusion de la prise en charge des complications (2000)

FIG.13-*Un anaphorique comme indice de continuation*

Pour pouvoir traiter ce type de cas, les indices marquant une relation anaphorique sont très pertinents, comme la reprise lexicale, la présence de pronoms démonstratifs et d'ellipses.

Deux types d'indices pertinents ont donc été mis en valeur : les marqueurs de relations sémantico-rhétoriques comme la justification ou le contraste, qui sont de bons indices de rupture dans le cas où la segmentation par défaut au paragraphe pour les expressions détachées se révèle trop tardive, et les marqueurs de relations anaphoriques, qui sont de bons indices dans le cas où la segmentation à la phrase pour les expressions intégrées se révèle à l'inverse trop précoce.

5 Vers une automatisation de la segmentation

L'étude linguistique présentée ci-avant a permis de cerner quels indices sont pertinents pour la segmentation en cadres. Sur la base de cette étude, nous avons défini une architecture modulaire qui exploite ces indices pour segmenter et structurer les GBP.

5.1 Implémentation

Une architecture modulaire composée des éléments suivants a été définie :

1. Repérage des introducteurs (unités-condition)
    a. Module générique de repérage des marques linguistiques de la condition
    b. Module spécifique dédié aux titres dont l'expression de la condition n'est pas marquée linguistiquement. Le typage se fait alors par le repérage de termes qui, dans le domaine, ont une propension à exprimer une condition.
2. Repérage des indices pertinents pour délimiter la fin des cadres
    a. Repérage de la structuration logique et des titres
    b. Repérage des énumérations
    c. Repérage des marqueurs de relations sémantico-rhétoriques
    d. Repérage des anaphoriques
3. Calcul de la délimitation des cadres sur la base d'heuristiques exploitant les indices.
4. Structuration des unités conditions-actions sur le modèle de GEM.
5. Génération d'une interface interactive permettant à l'utilisateur de consulter et valider les propositions du système

Les modules implémentés sont les suivants :

1. Le module générique de repérage des introducteurs de cadre. Ce module est fondé sur l'utilisation de transducteurs qui exploitent des classes de

marqueurs linguistiques ainsi que des étiquettes morpho-syntaxiques.
2. Le module de repérage des marqueurs de relations sémantico-rhétoriques également fondé sur l'utilisation de transducteurs.

Les transducteurs sont définis grâce au logiciel Unitex développé par l'Université de Marne-la-Vallée (`www-igm.univ-mlv.fr/~unitex`). On peut ainsi annoter les marqueurs repérés en ayant recours, de manière simple, à un ensemble varié d'informations linguistiques (lemmes, formes fléchies des mots, catégorie morpho-syntaxique, traits sémantiques…).

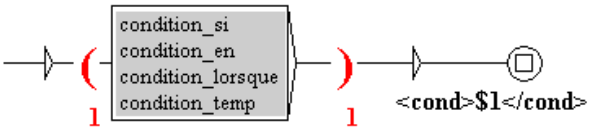

FIG.14-*Un exemple de transducteur Unitex*

Un programme Perl permet ensuite de désambiguïser certains marqueurs ambigus en fonction de leur contexte d'apparition, sur la foi d'indices locaux.

3. Le module de repérage des énumérations. Ce module est fondé sur le repérage d'éléments graphiques pertinents (puces etc.), de la structuration logique et des introducteurs. Ce module est écrit en Perl et XSLT. Tous les modules prennent en entrée un fichier XML et produisent en sortie les mêmes fichiers XML modifiés.

5.2 Validation

L'implémentation du système complet étant en cours, il n'est évidemment pas possible d'en donner une évaluation détaillée. La démarche générale et les modules implémentés ont toutefois été validés de la manière suivante. Des médecins et chercheurs du LIM&BIO-Paris 13 travaillant sur la modélisation des GBP ont validé notre approche. La stratégie de modélisation en deux étapes (segmentation en cadres et structuration d'après le modèle GEM) a notamment été considérée comme pertinente.

De plus, un schéma d'interface a été présenté et validé. Cette interface permet de valider de manière simple et interactive les propositions du système. Le principal critère d'évaluation est la réduction de temps que permet l'utilisation du système par rapport à une modélisation manuelle. Des mesures du temps nécessaire pour effectuer la tâche sont prévus avec les experts.

Les performances de deux modules implémentés ont été évaluées sur trois GBP, isolés au départ pour servir d'échantillon de test. Les résultats sont les suivants :

1. Le module de repérage des introducteurs (sans les titres) donne une précision de 98% et un rappel de 94% sur un total de 65 introducteurs .
2. Le module de repérage des énumérations donne une précision de 90 % et un rappel de 95% sur un total de 18 énumérations.

Si ces premiers résultats ne permettent évidemment pas de tirer de conclusions certaines, en raison du très petit nombre de données, le niveau atteint permet de poursuivre l'implémentation du processus de segmentation. Ces résultats seront complétés et affinés par la suite sur un nombre plus important de données.

6 Positionnement de l'approche

L'analyse de textes de spécialité est un des champs applicatifs du TAL qui a connu un grand succès ces dernières années : des travaux se sont concentrés sur les méthodes d'extraction de termes [3] ou de relations entre ces termes [7]. Leur objectif étant d'aider à modéliser, nous nous inscrivons pleinement dans la lignée de ces travaux. Néanmoins, nous nous en distinguons par la prise en compte d'un niveau d'analyse différent. En effet, les travaux sus-mentionnés s'appliquent à un niveau interne à la phrase tandis que notre niveau d'analyse est le texte en lui-même. On peut dire, pour reprendre la terminologie de Rastier [14], que notre étude délaisse partiellement la micro-sémantique, elle se situe à mi chemin entre la méso-sémantique (étude des cadres, de leur imbrication) et la macro-sémantique (comment ces cadres contribuent à l'organisation globale du texte).

Les travaux fondateurs de Halliday [8] sur la notion de cohérence textuelle constituent l'arrière plan de notre étude. Nous nous inspirons fortement de la théorie de l'encadrement du discours de Charolles [4], dont nous reprenons en partie la notion de cadre et de portée. Nous empruntons également aux travaux de Virbel et Luc, qui étudient le fonctionnement de la structure visuelle d'un texte [10] [19]. Nous portons enfin une attention toute particulière aux travaux de Pascual et Péry-Woodley [13a] [13b] qui a notamment travaillé sur l'interaction entre différentes structures textuelles.

Nous partageons assez largement l'arrière-plan commun à l'ensemble de ces études. Nous pensons que le texte est un objet social, qu'il vise à faire passer un certain nombre d'information au lecteur. L'organisation visuelle et les marqueurs rhétoriques contribuent à l'identification de cette structure [8]. Toutefois, la DTD GEM ne propose pas une structure globale figée (ce que l'on pourrait appeler une « grammaire de texte ») mais elle propose des éléments de modélisation qui, mis ensemble, donnent une vision globale du texte. Il est donc primordial d'identifier les séquences (méso-sémantique) à mettre en relation. Leur organisation et leur inter-dépendance est largement libre et s'organise autour des marqueurs linguistiques et autres indices visuels (cf. [19]).

Un des enjeux consiste à déterminer la généricité de ces marques. De fait, nous savons que certains des marqueurs identifiés sont spécifiques au domaine, que d'autres sont ambigus et doivent être maniés avec précaution (Péry-Woodley a montré ce phénomène pour les définitions au sein de textes procéduraux [13b], Adam et Revaz [21] pour certains marqueurs comme *enfin*). Ceci est particulièrement vrai pour les titres : une

identification et un typage propre est possible mais demande des ressources fines et adaptées. Des travaux, s'inspirant largement de l'étude de M.-P. Jacques [9] sont actuellement en cours sur ce point. Au-delà de l'analyse linguistique, nous nous démarquons des travaux précédents en visant le développement d'un outil permettant le repérage de ces structures et leur exploitation dans le cadre d'une tâche d'aide à la modélisation. La description est donc faite avec un souci d'opérationnalisation qui oblige à fonder l'analyse linguistique sur des indices repérables automatiquement en corpus.

Cet objectif d'opérationnalisation nous inscrit dans le cadre des systèmes d'accès à l'information qui exploitent la structure du texte – qu'elle soit de nature thématique [5] ou rhétorique [12] [17] – ou des structures spécifiques comme les cadres de discours [2]. Nous nous démarquons néanmoins de ce dernier par l'objectif applicatif final ainsi que par une étude manuelle approfondie des indices permettant de délimiter ces cadres.

7 Conclusion

Cet article a présenté une méthode de segmentation et de structuration des GBP pour une tâche d'aide à la modélisation. L'étape de segmentation consiste à délimiter les cadres engendrés par les unités qui dans le texte expriment une condition.

Nous nous sommes concentrés dans cet article sur le rôle de la structure visuelle pour la segmentation. Une étude linguistique, effectuée sur un corpus de textes de recommandations médicales, a montré la pertinence ainsi que les limites de ce type d'indices.

L'enjeu principal consiste désormais à étudier la portée de ce travail et sa généricité. Chaque texte procédural est spécifique et, si l'on espère qu'une partie de ce travail (marqueurs, indices visuels…) est réutilisable, une autre partie devra être adapté. Nous travaillons actuellement sur d'autres corpus, afin de déterminer des stratégies d'adaptation efficaces en fonction des textes et de la tâche envisagée.

Références

[1] J.-M. Adam et F. Revaz, F. Aspects de la Structuration du Texte Descriptif: Les Marqueurs d'Énumération et de Reformulation. *Langue Française*, 81. 1989. pp. 59-98.

[2] F. Bihaut & al. Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. *TALN 2003*.

[3] D. Bourigault. *LEXTER, un logiciel d'Extraction de TERminologie.Application à l'acquisition de connaissances à partir de textes.* Thèse de doctorat, EHESS, 1994.

[4] M. Charolles. L'encadrement du discours-univers, champs, domaines et espaces. *Cahier de recherche linguistique*,6, 1997.

[5] O. Ferret., B. Grau, J-L Minel, S. Porhiel. Repérage de structures thématiques dans des textes. *TALN*, Tours, 2001.

[6] G. Georg, B. Séroussi, J. Bouaud. Dérivation d'une base de connaissances à partir d'une instance GEM d'un guide de bonnes pratiques médicales textuel. In *IC' 2003*.

[7] N. Grabar et T. Hamon. Les relations dans les terminologies structurées :de la théorie à la pratique. In *Revue d'Intelligence Artificielle*. Vol 18/1. p57-85, 2004.

[8] M.A.K. Halliday et R. Hassan. *Cohesion in English. London*,1976.

[9] M.-P. Jacques. Structure matérielle et contenu sémantique du texte écrit. *Corela* , Vol. 3 n° 2. 2005.

[10] C. Luc. *Représentation et composition des structures rhétoriques et visuelles du texte. Approche pour la génération de textes formatés*. Thèse de doctorat. Université, Paul Sabatier, 2000.

[11] W.C. Mann et S.A. Thompson. *Rhetorical Structure Theory : Toward a functional theory of text organization*, 1998.

[12] J-L. Minel et G. Mourad. Filtrage automatique de textes, le cas de la citation. *CIDE'2000* p41-56, 2000.

[13a] E. Pascual et M-P. Péry-Woodley. Définition et actions dans les textes procéduraux. In Pascual E.,Nespoulous J-L et Virbel Editeurs *Le texte procédural: langage, action et cognition*, p 223-248, Mons, Gers. Prescot, 1997.

[13b] M.P. Péry-woodley. Modes d'organisation et de signalisation dans des textes procéduraux. *Langages* 141, p28-46.,1998.

[14] F. Rastier, M. Cavazza et A. Abeillé. *Sémantique pour l'analyse*. Masson. Paris. 1994.

[15] B. Séroussi,.J. Bouaud, H. Dréau., H. Falcoff., C. Riou., M. Joubert., G. Simon, A. Veno. ASTI : A Guideline-based drug-ordering system for primary care. I: Patel VL, Rogers R, Haux R (eds*). MedInfo* 84(1), p 528-532, 2001.

[16] RN. Shiffman et al. GEM : A proposal for a more comprehensive guideline documant using XML. *J Am Med Inform Assoc* 7(5) p488-98, 2000.

[17] S. Teufel S.*Argumentative zoning. Information Extraction from scientific articles*. PhD Thesis. University of Cambridge, 1999.

[18] SW. Tu et MA. Musen. Modeling Data and Knowledge in the EON Guideline Architecture.Proc.*MedInfo* 2001,London, UK, p280-284, 2001.

[19] J-L. Virbel et C. Luc. Le modèle d'Architecture Textuelle : fondements et expérimentation. *Verbum* , XXIII, 1, p103-123, 2001.