Une approche textuelle pour l'analyse de textes de recommandations médicales

Amanda Bouffier

Laboratoire d'Informatique de l'université Paris-Nord
UMR CNRS 7030
Institut Galilée - Université Paris-Nord
99, avenue Jean-Baptiste Clément
93430 Villetaneuse
amanda.bouffier@lipn.univ-paris13.fr

RÉSUMÉ. *Les systèmes d'analyse sémantique reposent le plus souvent sur des analyses locales du texte tandis qu'il devient de plus en plus évident que l'organisation du texte fait sens et doit être exploitée pour fournir un meilleur accès au contenu des documents. Cet article a pour objectif de montrer l'apport d'une approche textuelle au sein d'un cadre applicatif précis : la modélisation des Guides de Bonnes Pratiques Médicales. Nous proposons l'application GemFrame qui se fonde sur une approche textuelle pour fournir une première représentation structurée de ces textes. Le système a été validé sur trois aspects complémentaires : utilité, performances et pertinence de la méthode.*

ABSTRACT. *NLP is mainly focused on words and sentences even if most people agree that a better understanding of text structure could help extracting knowledge. In this article we show that, in certain domains, textual approaches are relevant for NLP, taking as an example a specific task related to the medical domain : the automatic modelling of health practice guidelines. Our system, GemFrame, is capable of automatically structuring Health Practice Guidelines. We propose a strategy based on the recognition of linguistic features. The system has been validated on three complementary aspects : usefulness, performances and relevance of the method.*

MOTS-CLÉS : *analyse discursive, linguistique textuelle, modélisation des connaissances, Guides de Bonnes Pratiques Médicales.*

KEYWORDS: *discursive analysis, text linguistics, knowledge modelling, Clinical Practice Guidelines.*

1. Introduction

Les systèmes d'analyse sémantique, que ce soit pour la recherche, l'extraction d'information ou la construction de ressources, reposent le plus souvent sur des analyses locales du texte (repérage d'entités nommées, extraction de termes, etc.) dépassant rarement le cadre de la phrase. L'organisation du texte, sa structure, reste peu exploitée en TAL alors qu'il devient de plus en plus évident qu'elle fait sens et doit être prise en compte pour fournir un meilleur accès au contenu des documents. Par exemple, dans le cas des systèmes de recherche d'information, analyser l'organisation d'un texte permettrait de fournir à l'utilisateur un fragment de texte cohérent et d'améliorer ainsi l'interprétation du résultat. Dans les systèmes de questions-réponses, certaines questions non factuelles comme les questions en « comment » supposent de pouvoir extraire des blocs de texte et non de simples phrases. Si certains systèmes prennent en compte une dimension textuelle dans leurs analyses, l'apport de ce type d'approche par rapport à une approche locale, en général moins coûteuse, reste peu évalué. L'objectif du travail que nous présentons dans cet article est de montrer et d'évaluer l'apport d'une approche textuelle au sein d'un cadre applicatif précis.

Montrer l'apport d'une approche textuelle ne peut se passer d'un cadre applicatif, garant d'une évaluation rigoureuse. En effet, un tel cadre permet de réduire les variations d'interprétation et ainsi d'établir un référentiel stable sur lequel faire reposer l'évaluation. Notre cadre applicatif est la modélisation de textes de recommandations dans le domaine médical : les Guides de Bonnes Pratiques Médicales. Ils font partie du type plus général des textes « incitatifs » dont la fonction est de « diriger nos actions, en nous disant quoi faire et comment le faire» (Adam, 2001). Ils sont rédigés par des autorités de santé[1] et adressés aux médecins afin d'assister leurs prises de décision. Ils sont néanmoins peu lus et exploités par ces derniers. Ce constat fit naître un champ de recherche en informatique médicale ayant pour objectif de développer des applications (Séroussi *et al.*, 2001 ; Shiffman *et al.*, 2000) permettant de faciliter l'accès des médecins aux connaissances contenues dans les GBPM. L'implication de chercheurs eux-mêmes médecins permet d'accéder à des référentiels stables sur lesquels fonder l'évaluation. Pour construire ces applications, les textes ont besoin d'être modélisés, ce qui reste une activité très coûteuse si elle est manuelle. Notre objectif est d'automatiser partiellement le processus de modélisation des GBPM en construisant une première représentation structurée de ces textes. Les GBPM étant essentiellement structurés autour de recommandations et de situations thérapeutiques sous lesquelles elles s'appliquent (que nous appelons des « conditions »), la tâche principale consiste à extraire automatiquement les segments exprimant une recommandation et ceux exprimant une condition puis de rattacher à chaque segment « condition » l'ensemble des segments « recommandation » qui dépendent de cette condition. Nous verrons que les relations de dépendance entre ces deux types de segments s'expriment pour une grande part au-delà de la phrase, au niveau textuel. Par conséquent, la modélisation

1. Notamment la Haute Autorité de Santé. http ://wwww.has.fr.

des GBPM offre un cadre applicatif pertinent pour montrer l'apport d'une approche textuelle en garantissant un cadre d'évaluation.

Cet article se focalise sur les spécificités du niveau d'analyse « textuel » concernant les indices linguistiques à prendre en compte ainsi que la méthode permettant de les obtenir. En effet, nous verrons que le niveau textuel met en exergue certaines propriétés linguistiques (notamment la sous-détermination et l'hétérogénéité des indices) qui ont des conséquences sur la méthode d'acquisition. Nous développerons donc ces aspects plutôt que la mise en œuvre concrète du système qui reste classique (nous décrirons néanmoins de manière succincte les formats d'entrée-sortie, modules et formalismes utilisés pour que le lecteur ait une vue d'ensemble du système).

L'article est structuré de la manière suivante : les deux premières parties ont pour objectif de présenter l'arrière-plan théorique et applicatif de ce travail et son positionnement par rapport à quelques travaux représentatifs. La partie 1 montre notre positionnement par rapport à la linguistique textuelle et les systèmes d'analyse automatique. La partie 2 relate du cadre applicatif, à savoir la modélisation des GBPM puis la tâche. Les deux dernières parties sont consacrées aux réalisations effectuées. La partie 3 présente les connaissances linguistiques pertinentes pour l'analyse automatique obtenues grâce à une méthode originale articulant analyse qualitative et quantitative. La partie 4 est dédié à la présentation succincte du système d'analyse automatique GemFrame que nous avons conçu pour l'analyse des GBPM et au travail d'évaluation dont il a fait l'objet.

2. Linguistique textuelle et systèmes d'analyse automatique

2.1. *Linguistique textuelle*

Pour la linguistique textuelle, un texte n'est pas qu'un ensemble de phrases mais un tout organisé et cohérent. Un texte est d'abord la trace matérielle d'un discours, que l'on définit avec (Cornish, 2006) comme une « séquence hiérarchisée et contextuellement située d'actes illocutoires, d'énonciation, de contenus propositionnels effectués dans la poursuite d'un but communicatif quelconque ». Le texte, lui, est la trace de ce discours, « une séquence connexe de signes verbaux et non verbaux en fonction de laquelle le discours est construit ». Le caractère partiellement décontextualisé du texte fait qu'il comporte un grand nombre d'indices linguistiques permettant de guider le lecteur dans son travail d'interprétation afin qu'il recrée du discours (Pery-Woodley, 2000).

2.2. *Systèmes d'analyse automatique*

Ces indices que l'on regroupe sous le terme de « marques de cohésion » vont justement permettre d'alimenter certains systèmes d'analyse automatique textuelle. Nous

définissons un système d'analyse textuelle comme un système dont le niveau d'analyse est supérieur à la phrase et qui s'intéresse explicitement aux relations qui existent entre des parties de texte. Pendant longtemps, les systèmes d'analyse du langage naturel se sont focalisés sur l'unité du syntagme ou de la phrase. Néanmoins, depuis une quinzaine d'années, l'intérêt pour le texte et ses articulations devient de plus en plus manifeste. Cet intérêt est né de besoins applicatifs nouveaux, liés notamment à l'accès de plus en plus facile aux grandes masses documentaires, et la nécessité, pour le lecteur, de pouvoir retrouver rapidement une information précise, synthétiser certaines informations et naviguer facilement dans les textes

2.2.1. *Applications*

Suite à une requête, les moteurs de recherche traditionnels renvoient à l'utilisateur un ensemble de documents sans dire beaucoup de leurs contenus. Les limites de ce type de système sont évidentes : pour vérifier la pertinence du document ou rechercher l'information dont il a besoin, l'utilisateur est obligé de lire le document en entier, ce qui peut se révéler fastidieux surtout dans le cas de documents longs. Les systèmes d'analyse textuelle sont nés précisément de ce besoin de recherche affiné dans le contenu des documents. On peut distinguer plusieurs besoins, notamment : 1) on cherche une information précise et localisée, et on aimerait d'une manière ou d'une autre connaître les zones de texte où l'information est le plus probablement susceptible de figurer. Dans ce type d'application, une approche textuelle est exploitée pour retrouver des unités textuelles pertinentes vers lesquelles diriger le lecteur (Bilhaut, 2006 ; Mondary *et al.*, 2007) ; 2) on veut prendre connaissance de manière synthétique de l'ensemble d'un document. L'application typique est alors le résumé automatique. Ce domaine applicatif bénéficie d'une meilleure maturité grâce à de nombreuses campagnes d'évaluation[2]. Les auteurs adoptent une approche textuelle soit sur la phase d'extraction (sélectionner les phrases importantes) (Marcu, 2000 ; Polanyi *et al.*, 2004), soit sur la phase de génération (recréer une cohérence entre les phrases extraites) (Saggion *et al.*, 2002) ; 3) on veut pouvoir passer facilement d'un grain à l'autre, c'est-à-dire avoir une vue d'ensemble mais aussi atteindre une partie plus précise du document et naviguer entre ces différents niveaux. Citons par exemple dans cette catégorie l'interface 3D–XV (Jacquemin et Jardino, 2002) ou (Choi, 2002).

Une des limites de ces travaux est que l'apport d'une approche textuelle (par rapport à d'autres méthodes qui ne le sont pas[3]), n'est souvent pas clairement évalué. (Marcu, 2000) donne l'évaluation certainement la plus aboutie quant à l'apport d'une approche textuelle dans le cadre du résumé automatique. Néanmoins, l'évaluation ne répond pas aux questions suivantes : cette approche est-elle plus pertinente

2. *Summarization Evaluation Conference* (SUMMAC) il y a quelques années et *Document Understanding Conference* aujourd'hui.

3. Toutes les applications précédemment citées ont fait l'objet d'approches non textuelles. Par exemple, le résumé automatique peut se fonder sur l'extraction de phrases sur des critères de récurrence lexicale ou d'identification de mots-clés (Teufel et Moens, 1999).

qu'une méthode non textuelle, fondée par exemple sur l'identification d'expressions-clés (Teufel et Moens, 1999) ? Dans ce travail, nous entendons amener des éléments de réponses tangibles à cette question en comparant, dans notre tâche, une approche textuelle avec une approche où l'analyse est restreinte à la phrase.

2.2.2. *Méthodes d'analyse*

Les auteurs s'accordent en général sur la multiplicité et l'hétérogénéité des indices à prendre en compte (Choi, 2002 ; Hernandez, 2004 ; Kurohashi et Nagao, 1994 ; Marcu, 1999). Peu, en revanche, font une analyse du rôle et de la contribution de ces différentes catégories d'indices. (Polanyi *et al.*, 2004), par exemple, ne justifie pas la prise en compte de tel ou tel indice. (Marcu, 1999) n'effectue pas non plus de bilan quant à la contribution des différents indices. Par rapport à ces travaux, notre approche se caractérise par un retour fin sur l'apport des différents indices exploités au moyen d'une analyse à la fois quantitative et qualitative. Si cette analyse n'est pas *de facto* généralisable car prenant pour objet un type de texte particulier, elle constitue néanmoins un premier pas vers une comparaison et une généralisation des connaissances à prendre en compte.

Un deuxième consensus règne autour du fait que les indices doivent être combinés pour être suffisamment fiables. La question la plus épineuse concerne selon nous les méthodes mises en œuvre pour acquérir ces combinaisons d'indices. (Polanyi *et al.*, 2004) ou (Kurohashi et Nagao, 1994) par exemple acquièrent manuellement leurs règles mais peu d'éléments méthodologiques sont décrits. On peut, par conséquent, s'interroger sur la validité des scores attribués aux différents indices ou sur l'ordonnancement des règles. Il est clair que plus les interactions sont nombreuses, plus elles sont difficilement appréhendables par la seule observation linguistique et plus elle nécessite le recours à des outils appropriés. (Choi, 2002 ; Hernandez, 2004 ; Marcu, 1999) en revanche recourent à une procédure d'apprentissage. Néanmoins, les différents indices pris en entrée ne font pas l'objet d'une justification *a priori*. Or, on connaît l'importance d'une bonne sélection et d'une bonne représentation des connaissances par rapport à la pertinence de l'apprentissage. Une analyse des résultats de l'apprentissage aurait également été intéressante pour déterminer la part de généricité. Face à ces questions, notre positionnement est le suivant : nous adoptons une approche qui allie l'apprentissage à l'observation linguistique de manière complémentaire. L'apprentissage nous donne un outil fiable pour acquérir les combinaisons d'indices pertinentes qui ne sont pas facilement appréhendables par la seule observation. L'observation de manière complémentaire permet de donner en entrée les indices et par la suite de faire un bilan des connaissances pertinentes.

3. Le cadre applicatif : la modélisation des Guides de Bonnes Pratiques médicales

3.1. *La tâche*

Le cadre applicatif est la modélisation des Guides de Bonnes Pratiques Médicales. Les GBPM sont des recommandations médicales écrites par des autorités en matière de santé et adressés aux médecins afin d'aider leurs prises de décision. Afin de faciliter l'accès des médecins aux connaissances contenues dans les guides, un champ de recherche en informatique médicale s'est développé ayant pour objectif de développer des applications d'aide à la consultation sous forme de systèmes de navigation ou de systèmes experts (Séroussi *et al.*, 2001). Notre travail s'inscrit dans le processus de modélisation des GBPM qui part des textes vers des modèles formels (par exemple (Peleg *et al.*, 2000)) venant alimenter les différentes applications finales. L'objectif applicatif de ce travail est de fournir une aide à la modélisation en construisant automatiquement une première représentation structurée de ces textes. Le cœur de la tâche consiste à 1) extraire automatiquement les segments exprimant une recommandation et ceux exprimant une condition puis 2) rattacher à chaque segment « condition » l'ensemble des segments « recommandation » qui dépendent de cette condition. Le modèle peut être représenté sous forme d'un arbre où les nœuds sont des conditions et les feuilles les actions recommandées. D'un point de vue référentiel, tous les fils d'un nœud « condition » sont vrais sous cette condition. La figure 1 montre l'exemple d'un fragment de texte structuré. Par exemple, dans l'arbre donné, la recommandation *(2)*

Chez les sujets non immunodéprimés *(a)*,(...) des biopsies nombreuses et étagées sont recommandées (...)*(1)*. Les biopsies isolées sont insuffisantes *(2)*.
L'exploration de l'iléon terminal est également recommandée *(3)*. En cas d'aspect normal de la muqueuse iléale (...) *(b)*, la réalisation de biopsies n'est pas systématique *(4)*.

Chez les sujets immunodéprimés *(c)*, il est recommandé de réaliser des biopsies systématiques (...) à la recherche d'une infection opportuniste *(5)*.

Endoscopie digestive basse. Indications en dehors du dépistage en population, 2004, p 13

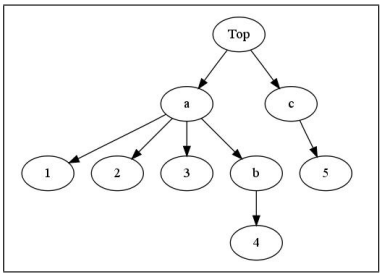


Figure 1. *Un extrait de GBPM modélisé*

est fille du segment conditionnel *(a)* car elle n'est valide que sous cette condition

On doit souligner que cette représentation est pertinente bien au-delà du cadre des GBPM. En effet, les relations entre segments « conditions » et « recommandations » (désormais nommées « relations conditionnelles ») sont propres au type des textes « incitatifs » (Adam, 2001) et non pas seulement aux GBPM (on les retrouve dans les manuels techniques, les textes juridiques, etc.). Par ailleurs, d'un point de vue formel,

une représentation arborescente reliant entre eux les différents segments est, quant à elle, encore plus générique et est équivalente à une modélisation de type RST (Mann et Thompson, 1988), formalisée et implémentée par la suite par D. Marcu (Marcu, 2000).

3.2. *Pertinence d'une approche textuelle*

Des travaux antérieurs (Georg, 2006 ; Hagerty *et al.*, 2004) ont eu pour objectif d'automatiser la modélisation des GBPM. Cependant, ils fondent leurs approches sur l'analyse de phrases isolées et ne calculent que les relations conditionnelles qui s'effectuent à l'intérieur de ce niveau. Or, nous avons observé que dans un grand nombre de cas, les relations conditionnelles ne s'établissent pas au travers du dispositif syntaxique mais au-delà de la phrase, au niveau textuel. Par exemple, dans la figure 1, on peut observer que les recommandations *(1)*, *(2)* et *(3)* sont liées à la condition *(a)* bien qu'elles fassent partie de phrases différentes. On doit également remarquer que certaines conditions peuvent être en relation de dépendance avec d'autres conditions, telles que *(b)* et *(a)* de sorte que l'on peut avoir plusieurs niveaux d'imbrication entre les conditions. Ce fait complexifie encore la tâche. On doit noter que (Kaiser *et al.*, 2005) étend l'analyse au niveau textuel mais ne décrit pas exhaustivement et de manière problématisée les connaissances linguistiques et la méthodologie d'acquisition.

Pour valider en amont la pertinence d'une approche textuelle, nous avons compté sur quatre GBPM le nombre de rattachements s'établissant au niveau phrastique et le nombre de rattachements s'établissant au niveau textuel. Pour ce faire, les quatre GBPM ont été annotés manuellement, selon la procédure décrite section 4.2. Les résultats obtenus sont récapitulés dans le tableau 1.

GBPM	**Nb rattachements**	**Nb rattachements textuels**	**Proportion**
Chimiothérapie	177	17	9,6 %
Dénutrition	182	71	39 %
AOMI	151	69	45,6%
Cancer du sein	169	48	28,4%
Total	679	205	30,7 %

Tableau 1. *Proportion des rattachements s'effectuant au niveau textuel dans quatre GBPM (résultat obtenu manuellement)*

D'après ce tableau, on constate qu'en moyenne 30,7 % des rattachements s'effectuent au-delà du niveau de la phrase. L'analyse textuelle permet par conséquent d'obtenir *a priori* plus de 30 % de rattachements corrects en plus par rapport à l'analyse centrée sur la phrase. Ce résultat valide la pertinence de l'approche.

Pour signaler ces relations, de nombreux indices linguistiques sont présents. Pour donner un exemple, le fait que *(a)* et *(3)* figurent dans le même paragraphe suggère qu'ils sont liés, de même que la non-intégration syntaxique de *(a)* ou l'occurrence de

l'adverbe de reprise *également* dans le segment *(3)*. De même la symétrie lexico-syntaxique entre *(c)* et *(a)* entre les deux expressions ainsi que la présence d'une marque d'antonymie entre (*immunodéprimés*/*non immunodéprimés*) suggère qu'ils ne le sont pas. Nous avons choisi d'exploiter ces indices linguistiques dans le traitement automatique pour des raisons de généricité.

4. L'acquisition des connaissances linguistiques

Pour acquérir les indices linguistiques pertinents pour le calcul des relations conditionnelles[4], nous avons mis en place une méthode d'acquisition qui articule analyse qualitative et quantitative. Cette méthode a été motivée par l'observation de deux propriétés récurrentes des indices linguistiques : leur hétérogénéité en terme de niveau d'analyse et leur sous-détermination. L'objectif de cette section est de présenter nos choix méthodologiques pour l'acquisition ainsi que les connaissances acquises.

Les indices peuvent être « profonds ». Nous avons observé que les indices (au sens très large d'éléments qui orientent l'interprétation du lecteur (Desclés, 1997)) sont hétérogènes du point de vue de leur niveau d'analyse. En effet, à côté de marques surfaciques, certains indices peuvent être profonds, c'est-à-dire être le résultat de calculs[5]. Par exemple, nous avons constaté le rôle d'un indice de similarité entre le titre d'une section et un segment conditionnel détaché, pour le calcul des relations conditionnelles. Il est illustré à la figure 2.

Dans cet exemple[6], on peut observer que le segment conditionnel détaché (en gras) est en partie redondant avec le titre qui le précède (sur les termes *rectocolite ulcéro-hémorragique ou maladie de Crohn*). Cette redondance entre le titre et le segment conditionnel suggère que tous les segments appartenant à la section sont en relation de dépendance avec le segment conditionnel, ce qui est loin d'être la situation la plus habituelle. La similarité entre le titre et le segment conditionnel est donc un indice pertinent pour calculer les relations de dépendance de ce dernier.

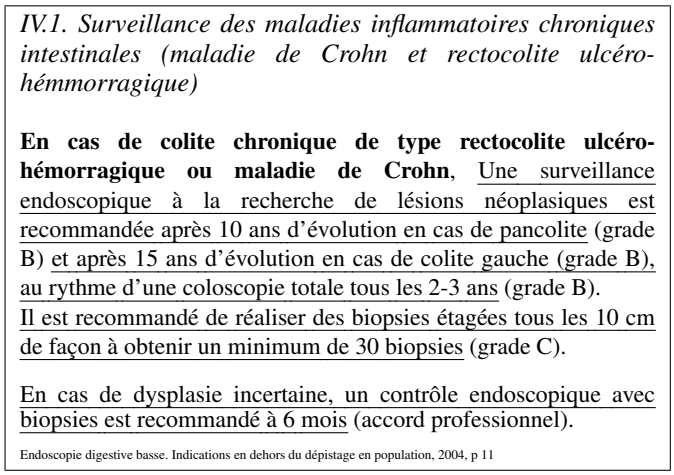
L'hétérogénéité des indices à prendre en compte se retrouve chez bon nombre d'auteurs en analyse discursive, ce qui souligne la généricité du phénomène. Par exemple, (Polanyi *et al.*, 2004) prend en compte dans son analyse, à côté d'indices surfaciques tels que des marqueurs lexicaux, un indice de parallélisme syntaxique ou d'information de centrage qui sont, au contraire, des indices profonds résultant de calculs.

---

4. Dans cette section, nous nous focalisons uniquement sur la deuxième étape de la tâche globale : le calcul des relations conditionnelles.

5. Dans la littérature, un indice est souvent synonyme de marque linguistique, ce qui, à notre sens, n'est pas exact.

6. Les segments en relation avec le segment conditionnel mis en gras sont soulignés. Ils constituent la « portée » du segment conditionnel.



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

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 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 2-3 ans (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, p 11

Figure 2. *Un indice profond : la simlarité entre un segment conditionnel détaché et le titre de la section*

Pour acquérir ce type d'indices, seule une phase d'observation linguistique experte est pertinente. En effet, il est illusoire de penser qu'une quelconque procédure d'apprentissage automatique puisse dégager automatiquement ce type d'indices à partir d'un texte annoté sur ses marques de surface. Aujourd'hui, le seul outil réellement pertinent reste l'observation experte du linguiste. De plus, certains phénomènes linguistiques peuvent être particuliers à un type de texte ou bien peu fréquents (cf. exemple précédent), de sorte qu'il est insuffisant de se reposer uniquement sur les indices présents dans l'état de l'art. Une observation spécifique du corpus reste nécessaire, même si elle possède également ses propres limites et nécessite d'être complétée par des outils quantitatifs comme nous l'argumentons un peu plus loin.

Les indices sont sous-déterminés. Une deuxième propriété récurrente des indices linguistiques est leur sous-détermination, à savoir le fait qu'aucun d'entre eux ne permet de manière isolée de conclure de façon certaine à une seule interprétation. La sous-détermination des indices fait qu'ils apparaissent souvent en cooccurrence et qu'ils interagissent entre eux : ils ont des poids plus ou moins grands suivant le contexte dans lesquels ils apparaissent. Par conséquent, l'analyse automatique doit reposer sur des combinaisons d'indices plutôt que sur des indices isolés pour atteindre un niveau de fiabilité suffisant. Or, l'observation pure ne permet pas de capturer les combinaisons tant les interactions sont complexes. Il devient nécessaire d'avoir recours à des outils quantitatifs appropriés. Ceux-ci permettront de déterminer le rôle de chacun des indices, la manière dont ils interagissent et s'influencent les uns les autres, ainsi que les combinaisons les plus discriminantes sur lesquelles faire reposer l'analyse automatique.

Sur la base de ces arguments, nous proposons une méthode d'acquisition qui articule analyse qualitative et quantitative de manière complémentaire, et que nous présentons ci-après.

4.1. *Principes généraux*

La méthode d'acquisition procède en deux temps :

– une première phase d'observation du corpus a pour objectif d'isoler les indices qui ont un rôle à jouer dans la détermination des relations de dépendance conditionnelles ;

– une phase d'analyse quantitative fait appel à une méthode d'apprentissage artificiel ainsi qu'à des outils statistiques et a le double objectif de déterminer le rôle de chaque indice de manière isolée et les meilleures combinaisons d'indices.

4.2. *Le corpus d'étude*

La phase d'acquisition exploite un corpus d'étude comportant 25 GBPM (environ 150 000 mots) publiés par l'ANAES (*Agence Nationale d'Accréditation et d'Évaluation en Santé*) ou l'AFSSAPS (A*gence Francaise de Sécurité Sanitaire des Produits de Santé*) et portant sur la prise en charge de diverses pathologies.

Le corpus a été annoté manuellement avec l'aide de médecins chercheurs travaillant également sur la modélisation des GBPM (Alain Venot et Catherine Duclos du laboratoire LIM & Bio de Paris 13). Un échantillon de trois GBPM a d'abord été annoté par deux annotateurs : un expert et un non-expert familier des activités d'annotation et de modélisation. Le bon accord entre annotateurs (cœfficient kappa[7] de 0,9) suggère que, dans un grand nombre de cas, des connaissances linguistiques générales suffisent pour interpréter correctement les GBPM, quand bien même des connaissances du domaine restent nécessaires dans certains cas. Compte tenu de ces observations et du fait que l'annotation des GBPM constitue une tâche très lourde pour les experts, nous avons eu recours à une double annotation, par deux linguistes experts en modélisation. La double annotation permet de prévenir au maximum les cas d'erreurs.

L'annotation a consisté à construire une structure arborescente (dans un tableur) où chaque nœud correspond à un segment exprimant une condition et chaque feuille correspond à un segment de type « recommandation » de façon similaire au format de sortie du système. Il a donc été aisé de comparer les deux lors de la phase d'évaluation.

7. Indice statistique variant entre 0 et 1 utilisé notamment pour évaluer le degré d'accord entre deux juges quant à la manière de classer un ensemble d'individus ou d'objets dans un certain nombre de catégories.

4.3. *La phase d'analyse qualitative*

La phase d'analyse qualitative nous a permis de relever l'ensemble des indices potentiellement pertinents pour le calcul des relations conditionnelles. Nous les avons caractérisés et organisés au sein d'une typologie que nous présentons ci-après.

Du point de vue de leur fonctionnement linguistique, nous reprenons à notre compte une distinction qui est apparu comme particulièrement pertinente : celle de (Charolles, 1995) entre deux systèmes de marques de cohésion : les marques « descendantes » et « ascendantes ». Cette distinction, tout à fait prégnante dans les GBPM, témoigne de deux manières de « mettre en texte » les relations de dépendances conditionnelles et correspond, selon nous, à des stratégies discursives et des modes d'organisation des connaissances différents.

Les **marques de cohésion « descendantes »** désignent les expressions qui ne sont pas intégrées syntaxiquement, soit qu'elles sont détachées à l'initial d'une phrase soit qu'elles font partie d'un titre, par exemple. Les segments conditionnels qui répondent à ces propriétés ont la capacité de lier plusieurs segments qui le suivent (d'où le terme de « descendant ») en les regroupant, comme le montre l'exemple de la figure 3.

> **En cas de traitement de l'hypertension artérielle**, il est recommandé d'utiliser préférentiellement la perfusion intraveineuse pour un ajustement tensionnel précis. Les voies intramusculaire et sublinguale sont à éviter. (...) ? Les agents hypersmolaires peuvent être utilisés pendant moins de 5 jours chez les patients dont l'état clinique se détériore du fait d'un œdème cérébral.
>
> Prise en charge initiale des patients atteints d'accident vasculaire cérébral, 2002, p 9

Figure 3. *Un segment conditionnel détaché syntaxiquement*

Dans cet exemple, tous les segments soulignés forment une unité homogène et se trouvent liés au segment conditionnel *en cas de traitement de l'hypertension artérielle*. Nous avons observé que si les segments détachés regroupent souvent plusieurs segments, les segments intégrés n'ont en général pas cette capacité sans recours à d'autres marques. D'autres travaux ont montré des propriétés similaires pour des expressions détachées, sémantiquement différentes des expressions conditionnelles : temporelles (Bilhaut, 2006), spatiales (Charolles, 1997), de prise en charge (Schrepfer-André, 2005).

Soulignons que sur le plan textuel, les relations conditionnelles sont exprimées, dans cette configuration, au travers d'un processus de segmentation plutôt que d'une mise en relation explicite. La position particulière (détachée, dans un titre) de ces segments conditionnels et leur capacité à segmenter en font des unités saillantes structurant le matériau textuel, contrairement aux expressions intégrées syntaxiquement.

Les **marques de cohésion « ascendantes »** forment un deuxième système de marques qui permet de signaler qu'un segment (recommandation ou condition) doit

être lié à une condition déjà introduite dans le texte (d'où le terme « ascendantes »). Elles sont multiples : relation de coréférence, marques lexicales, connecteurs, etc. À titre d'exemple, la figure 4 montre l'utilisation d'un anaphorique qui permet à l'auteur de lier des recommandations à un segment conditionnel en faisant référence à une condition précédemment évoquée.

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.

Stratégie de prise en charge du diabète de type 2 à l'exclusion des complications, 2000, p 10

Figure 4. *Une relation de coréférence entre deux segments conditionnels*

Dans cet exemple, *Dans les deux cas* renvoie aux deux conditions exprimées par *lorsque l'HbA1c est > 8 %, sur deux contrôles successifs sous l'association de sulfamides/metformine à posologie optimale*, ainsi que *lorsque l'HbA1c est comprise entre 6,6 % et 8 % sous la même association*. Par conséquent, l'auteur veut signifier que la recommandation liée à la condition *Dans les deux cas*, à savoir *la diététique aura au préalable été réévaluée et un facteur intercurrent de décompensation aura été recherchée* doit être du même coup liée aux deux conditions évoquées précédemment.

Contrairement aux marques descendantes, les marques ascendantes signalent les relations conditionnelles non pas au travers d'un processus de segmentation mais plus directement par une mise en relation.

On doit également souligner le rôle des indices relatifs à la structure visuelle, c'est-à-dire l'ensemble des moyens typo-dispositionnels d'un texte (structure en paragraphes ou sections, titres, mise en forme, etc.) pour signaler les relations conditionnelles dans les GBPM. La structure visuelle opère une hiérarchisation du contenu textuel à travers le découpage en sections, paragraphes ou listes. Il y a souvent une certaine correspondance entre la portée des segments conditionnels et la hiérarchisation visuelle qui peut aller d'une non-contradiction, à, dans certains cas, une isomorphie. Ainsi, dans l'exemple de la figure 5 le segment conditionnel qui correspond à un item d'une liste a une portée qui va jusqu'à l'item suivant.

Les indices observés ont été reportés et récapitulés dans le tableau 2 selon qu'ils signalent plutôt une relation ou une absence de relation entre un segment conditionnel (nommé $segConditionnel$) et un segment quelconque de type « condition » ou « recommandation » (nommé $segCandidat$). On peut observer qu'ils sont hétérogènes du point de vue de leur nature linguistique. En effet, on trouve :

– des indices relatifs à la structure visuelle ;

– des indices de coréférence ;

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, p 8

Figure 5. *Cas d'isomorphie entre portée et item d'une énumération*

– des connecteurs discursifs ;

– des indices lexicaux.

On doit également constater qu'ils sont hétérogènes en terme de profondeur d'analyse (un connecteur discursif est une marque de surface tandis qu'une relation de coréférence ou une similarité entre segments est issue d'un calcul) ainsi qu'en terme de grain d'analyse (la présence d'un terme est local tandis que la structure en paragraphes implique une vue globale, bien que simplifiée, du texte).

4.4. *La phase d'analyse quantitative*

Si l'analyse qualitative a permis d'isoler les indices pertinents, nous avons toutefois montré qu'aucun indice n'est fiable de manière isolée de sorte qu'il est nécessaire de s'appuyer sur des combinaisons de ceux-ci. Une observation purement qualitative ne permet pas de déterminer les interactions entre indices et la manière dont chacun contribue à la relation. C'est pourquoi, dans un deuxième temps, nous recourons à une analyse quantitative, utilisant le même corpus que l'analyse linguistique, en vue de calculer la contribution de chaque indice et de chaque catégorie d'indices et d'identifier les combinaisons d'indices les plus discriminantes. Le premier objectif est satisfait par l'utilisation d'une mesure statistique standard : le gain d'information que nous décomposons ensuite en deux mesures de fréquence et de précision (que nous définissons plus loin) afin de mieux capturer certains phénomènes. Le second objectif utilise une procédure d'apprentissage supervisé. Une des originalités de cette phase d'analyse réside dans l'utilisation des outils d'apprentissage non comme une procédure purement automatique d'induction de connaissances mais comme un outil d'aide à l'analyse linguistique.

4.4.1. *Système d'apprentissage du concept de relation conditionnelle*

Déterminer les meilleures combinaisons d'indices permettant de définir qu'un couple quelconque $(segCandidat, segConditionnel)$ satisfait ou non une relation conditionnelle peut se formuler de la manière suivante : il s'agit d'apprendre une fonction booléenne $estEnRelation$ permettant d'associer à tout couple de segments

Indices de relation
1. $segConditionnel$ est en position détachée
2. $segCandidat$ appartient au même paragraphe que $segConditionnel$
3. $segCandidat$ n'est pas dans la même position visuelle que $segConditionnel$
4. la phrase du $segCandidat$ contient un connecteur de subordination
5. $segCandidat$est en relation de coréférence avec une phrase précédente qui est liée au $segConditionnel$
6. $segCandidat$ possède un terme t' en expansion d'un terme t appartenant au $segConditionnel$
7. $segConditionnel$ est similaire au titre de la section de $segConditionnel$
Indices d'absence de relation
8. $segConditionnel$ est en position intégrée
9. $segCandidat$ est dans la même position visuelle que $segConditionnel$
10. $segCandidat$ n'appartient pas au même paragraphe que $segConditionnel$
11. $segCandidat$ a la même mise en forme que $segConditionnel$
12. $segCandidat$ a la même amorce lexicale que $segConditionnel$
13. $segCandidat$ possède un terme avec une expansion de longueur identique que $segConditionnel$
14. $segCandidat$ a un terme t qui a une marque d'antonymie avec un terme t' appartenant au $segConditionnel$
15. la phrase de $segCandidat$ possède un connecteur de coordination
16. $segConditionnel$ est lui même lié à un autre segment conditionnel

Tableau 2. *Récapitulatif des indices pour le calcul des relations conditionnelles*

$(segCandidat, segConditionnel)$ la valeur *vrai* « est en relation » ou la valeur *faux* « n'est pas en relation » à partir d'un échantillon d'exemples étiquetés (couples de segments liés et couples de segments non liés) représentés par l'ensemble d'indices mis en évidence lors de la phase d'analyse qualitative. Il s'agit donc d'un problème de classification et plus précisément d'apprentissage supervisé de concept où on cherche une fonction booléenne f qui classe correctement toute instance x soit en exemple positif du concept ($f(x) = vrai \Leftrightarrow x \in X^+$) soit en exemple négatif ($f(x) = faux \Leftrightarrow x \in X^-$) à partir d'un échantillon d'exemples positifs de X^+ et négatifs de X^-.

Pour mener à bien cet apprentissage, nous avons choisi une représentation des instances sous forme d'attributs booléens (reposant sur la logique des propositions) représentant les indices identifiés lors de la phase qualitative. En effet, chaque indice, qu'il implique l'un des segments (par exemple $segConditionnel$ contient un connecteur) ou les deux (par exemple $segConditionnel$ et $segCandidat$ sont dans le même paragraphe), est naturellement représenté par un attribut booléen comme le montre le tableau 3. Outre sa simplicité, la représentation propositionnelle a l'avantage d'être exploitable par des algorithmes efficaces.

Attributs	**Domaine**
est_intégré_syntaxiquement ($segConditionnel$) *est_dans_même_paragraphe* ($segCandidat, segConditionnel$) (...)	*{vrai,faux}*

Tableau 3. *Exemples d'indices représentés sous formes d'attributs booléens*

Selon cette représentation un indice correspond alors à un attribut instancié (recevant la valeur *vrai* ou *faux*). À partir des indices présentés tableau 2, nous aboutissons à un ensemble de 14 attributs (toutes les instanciations possibles ne correspondant pas à un indice, le nombre d'attributs n'est pas deux fois inférieur au nombre d'indices). D'un point de vue formel, la représentation des instances peut donc être décrite de la manière suivante. Soit $A = \{a_1, \ldots, a_{14}\}$ l'ensemble des attributs booléens représentant les indices, chaque couple de segments $(segCandidat, segConditionnel)$ est décrit par un vecteur $< v_1, \ldots, v_{14} >$ contenant les valeurs de vérité (*vrai* ou *faux*) pris par les attributs de A. Cette représentation permet de différencier $2^{14} = 16384$ couples de segments qui constituent l'ensemble X des instances du domaine.

Nous avons choisi une fonction de classification sous forme d'arbre de décision que l'on construit à partir des attributs de A. Un arbre de décision booléen est une organisation arborescente d'un ensemble de tests booléens où chaque nœud teste la valeur d'un attribut (par exemple $segConditionnel$ et $segCandidat$ sont-ils dans le même paragraphe ?). Pour construire l'arbre de décision, nous avons choisi l'algorithme C4.5 (Quinlan, 1993). On trouvera une description de l'algorithme de construction de l'arbre de décision dans (Cornuéjols et Miclet, 2002).

4.4.2. *Outils statistiques*

Nous utilisons une mesure de gain d'information pour calculer la contribution des indices au sein de la relation conditionnelle. Le gain d'information est aussi utilisé par $C4.5$ pour identifier l'attribut le plus discriminant à ajouter à l'arbre courant. Il peut être explicité de la manière suivante : soit E l'ensemble des exemples (couples de segments) décrits par l'ensemble d'attributs A, le gain d'information permet de savoir en quoi la connaissance de la valeur prise par un attribut a de A nous renseigne sur la classe d'un couple de segments, et donc quel est le pouvoir discriminant de a. Le gain d'information fait appel à la notion d'entropie qui, appliquée à notre problème peut se comprendre de cette manière : soit C la classe d'un couple de segments tiré au hasard dans E selon la loi uniforme. Comme nous l'avons montré auparavant, C prend ses valeurs dans $val(C) = \{+, -\}$. L'entropie notée $H_E(C)$ est la quantité d'information moyenne nécessaire pour identifier la classe d'un couple de segments tiré au hasard. De façon équivalente $H_E(C)$ représente l'incertitude moyenne sur la classe d'un couple de segments. $H_E(C)$ vaut :

$$-p(C = +).log_2(p(C = +)) - (1 - p(C = +)).log_2(1 - p(C = +)).$$

$H_{E_{a=vrai}}(C)$ (resp. $H_{E_{a=faux}}(C)$) est l'incertitude moyenne sur la classe d'un couple de segments sachant que la valeur d'un attribut est connue ($E_{a=vrai}$ (resp. $E_{a=vrai}$). En conséquence, l'entropie conditionnelle, ou incertitude moyenne sur la classe d'un couple de segments tiré au hasard dans E quand la valeur de a est connue est :

$$p(a = vrai).H_{E_{a=vrai}}(C) + (1 - p(a = vrai)).H_{E_{a=faux}}(C).$$

Le gain d'information associé à l'attribut a est $H_E(C) - H_E(C|a)$, représente la diminution d'entropie (ou d'incertitude) sur la classe d'un couple de segments due à la connaissance de l'attribut a. Autrement dit, le gain d'information représente l'information que a apporte sur la classe d'un couple de segments.

L'utilisation seule du gain d'information n'est pas suffisante pour notre objectif, c'est pourquoi nous recourons également à des mesures intermédiaires utilisées par le gain d'information. En effet, le gain d'information calcule la contribution moyenne d'un attribut. Or, ce qui nous intéresse plus précisément est la contribution d'un indice, ce qui correspond à un attribut instancié (à la valeur *vrai* ou faux). La contribution d'un indice se joue par rapport à deux facteurs :

– l'incertitude sur la classe (correspond à $H_{E_{a=vrai}}(C)$ du gain d'information). Nous apellons « précision » l'inverse de cette incertitude. $1/H_{E_{a=vrai}}(C)$;

– la fréquence ou probabilité d'occurrence de l'indice (correspond au facteur $p(a = vrai)$ du gain d'information). L'incertitude étant pondérée par la fréquence.

Il est important de distinguer ces deux éléments car deux indices peuvent être, en moyenne, discriminants de manière équivalente mais se distinguer par rapport à ces deux paramètres : l'un pouvant être fréquents mais peu précis et inversement. Comme nous le verrons par la suite, cette distinction est intéressante d'un point de vue linguistique et nous permettra notamment de poser la distinction entre des normes et des exceptions.

4.4.3. *Analyse des données*

En vue de l'apprentissage nous avons constitué 798 exemples (492 positifs et 306 négatifs) à partir du corpus annoté de 25 GBPM. Un couple de segments $(segCandidat, segConditionnel)$ du corpus annoté est un exemple positif (est lié) si $segCandidat$ est un enfant de $segConditionnel$, négatif dans le cas contraire. Cependant, nous n'avons pas considéré certains exemples négatifs : en effet seuls les exemples tels que $segCandidat$ et $segConditionnel$ sont frères, sont susceptibles de contenir des indices linguistiques pertinents de rupture. Prendre en compte l'ensemble des exemples négatifs aurait généré de nombreux exemples non informatifs.

Pour l'application des mesures et algorithmes, nous avons essentiellement utilisé la plate-forme WEKA (www.cs.wakaito.ac.nz/ml/weka) contenant un ensemble de programmes Java dédiés au Data Mining. Nous avons utilisé particulièrement le sélecteur *InfoGainEval* qui implémente la mesure de gain d'information et le programme *J48* qui implémente C4.5.

4.4.4. *Résultats et interprétation*

Nous présentons dans cette partie quelques résultats les plus significatifs suite à plusieurs expériences que nous avons menées : construction de l'arbre de décision après prise en compte de l'ensemble des indices ou seulement de certaines catégories (et calcul de la précision ainsi obtenue), calcul du gain d'information de chaque attribut et calcul de la précision et de la fréquence de chaque indice. Nous présentons une vue d'ensemble de ces résultats à la figure 6. Pour prendre connaissance précisément de tous les résultats, le lecteur peut se reporter à (Bouffier, 2008).

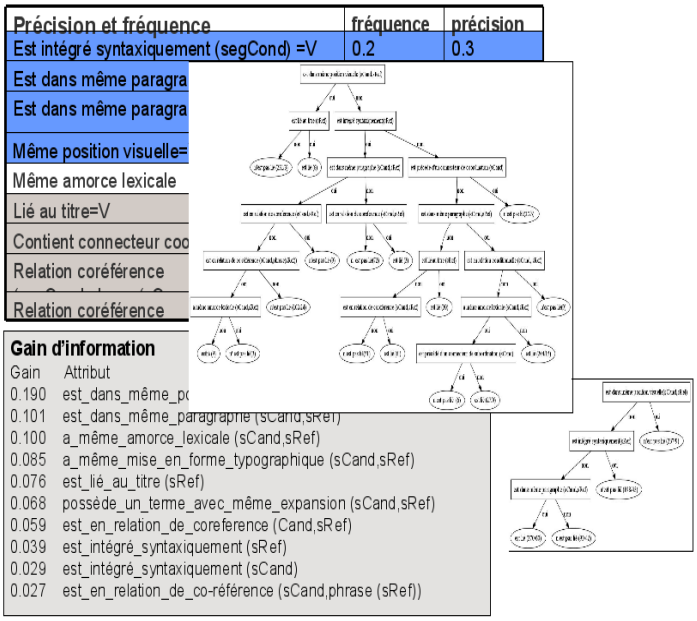


Figure 6. *Vue des résultats obtenus*

4.5. *Pouvoir discriminant des catégories d'indices*

Les indices relatifs à la structure visuelle sont très discriminants. Les indices relatifs à la structure visuelle sont les plus discriminants, comme le montre la mesure de gain d'information qui les place en tête. Pour confirmer ce résultat, nous avons appliqué *J48GainInfo* en prenant en compte uniquement les indices visuels. La précision de l'arbre obtenu est égale à 73 % (73 % des instances sont correctement classées). Cette précision est beaucoup plus élevée que pour n'importe quelle autre catégorie d'indices. Les indices les plus discriminants sont *est_dans_même_position_visuelle : vrai* ainsi que *est_dans_même_paragraphe : vrai*.

Les indices lexicaux sont corrélés aux indices visuels. Les indices lexicaux sont moins discriminants que les indices visuels, parce qu' ils sont largement corrélés à ces derniers et cela particulièrement pour les attributs lexicaux *a_la_même_expansion* et *possède_un_terme_en_relation_d'expansion*. En effet, si l'on ajoute ces attributs aux attributs liés à la structure visuelle dans l'algorithme d'apprentissage, une très faible amélioration de la précision de l'arbre produit est trouvée (moins de 1 %). Toutefois, ce constat est moins vrai pour l'attribut *a_même_amorce_lexicale* qui reste informatif et donc doit être pris en compte. Les autres peuvent être éliminés.

Les connecteurs discursifs et indices de coréférence sont intéressants. Les connecteurs discursifs ou les indices de coréférence sont, de manière isolée, peu discriminants. Cependant, combinés aux indices relatifs à la structure visuelle, ils permettent d'améliorer la précision de l'arbre obtenu de manière intéressante, et ce, davantage que les indices lexicaux. En effet, si l'on ajoute les connecteurs aux indices visuels dans C4.5, la précision de l'arbre produit augmente sensiblement (4 % pour les connecteurs, 2 % pour les anaphoriques contre moins de 1 % pour les indices lexicaux).

4.6. *Des normes et des exceptions*

Nous distinguons deux types d'indices : des normes et des exceptions.

Les indices relatifs à la structure visuelle sont ce que l'on appelle des « normes ». En effet, pris isolément, ils sont relativement précis mais surtout en moyenne plus fréquents que les autres catégories d'indices. On peut d'ailleurs observer qu'ils se conjuguent ensemble en premier pour former le haut de l'arbre de décision. Enfin, ils permettent de classer correctement 73 % des exemples. Le statut de norme des indices relatifs à la structure visuelle est dû au fait que les moyens de structuration visuels sont très exploités dans les GBPM, et plus généralement, dans les textes incitatifs.

Les autres indices sont ce que l'on peut appeller des « exceptions ». En effet, pris isolément, ils sont beaucoup moins fréquents. Cependant, ils ont la capacité de remettre en cause le classement suggéré par les indices « normes » et ainsi améliorent la précision de l'arbre. C'est pourquoi, ils se combinent aux indices visuels pour former le bas de l'arbre. Ces indices se distinguent entre eux par leur précision. Certains sont précis. C'est le cas par exemple de l'indice *est_lié_au_titre : vrai* ou de l'indice *est_précedé_d'un_connecteur_de_coordination : vrai*. Ces indices permettent par conséquent, à eux seuls, de remettre en cause *a_meme_amorce_lexicale : vrai*, sont des indices moins précis. C'est pourquoi ils se combinent entre eux pour atteindre une précision satisfaisante.

En résumé, sur la base de ces résultats, nous proposons les règles présentées à la figure 7, organisées autour de cette distinction entre normes et exceptions. Elles constituent une reformulation de l'arbre de décision global (les deux représentations

sont formellement équivalentes) mais la répartition des indices entre ces deux catégories permet d'en améliorer la lisibilité.

`De manière générale`

Par défaut, selon le principe de non-contradiction avec la structure visuelle, tout $segCandidat$ qui est dans la même position visuelle que $segConditionnel$ (à savoir qu'il appartiennent à des titres, des paragraphes, sous-paragraphes, des items d'énumération, etc. de même niveau) ne lui est pas lié, **sauf**

– si $segConditionnel$ est similaire au titre de la section

`À l'intérieur d'un même paragraphe`

Par défaut, si $segConditionnel$ est détaché alors $segCandidat$ est lié **sauf**

– si $segCandidat$ est précédé d'un connecteur de coordination ;

– si $segConditionnel$ est déjà lié à un segment conditionnel ;

– si $segCandidat$ est précédé d'un connecteur de subordination et qu'il possède la même amorce lexicale ;

Par défaut, si $segConditionnel$ est intégré alors $segCandidat$ n'est pas lié **sauf**

– si $segCandidat$ est un segment conditionnel qui coréfère à $egConditionnel$

– si $segCandidat$ reprend un référent de $segConditionnel$ sans posséder la même amorce lexicale.

Figure 7. *Règles pour le calcul des relations conditionnelles*

Ces règles ont été implémentées au sein du système GemFrame dédié à la modélisation des GBPM, dont nous décrivons certains principes de conception à la section suivante.

Le système GemFrame, dont cette section présente une brève description[8], permet de traiter un texte et d'en ressortir une représentation arborescente reliant les segments « condition » et « recommandation » tel qu'elle a été présentée à la figure 1 (l'arbre de sortie du système ne devant être confondu avec l'arbre de décision qui modélise les règles de rattachements). Les choix de conception découlent directement des conclusions de l'analyse linguistique.

La multiplicité des indices à prendre en compte nous a fait opté pour une architecture modulaire qui facilite la lisibilité du système et la réutilisabilité des compo-

8. Pour davantage de précisions, le lecteur peut consulter (Bouffier, 2008).

sants. Le système est découpé en deux modules principaux qui correspondent aux deux tâches majeures : 1) repérage des segments « condition » et « recommandation » 2) calcul des relations conditionnelles. Ces deux modules font appel à un ensemble de modules secondaires chargés de produire les connaissances nécessaires aux modules principaux, tels que le découpage en phrases, le repérage de la structure visuelle, le repérage des termes, etc. Lorsque cela était possible, des modules déjà existants ont été intégrés (tel que le Tree Tagger pour l'étiquetage morpho-syntaxique). Sinon des modules spécifiques ont été développés (tels que le repérage des segments conditionnels). L'ensemble des modules est présenté à la figure 8. Chaque module prend en entrée un document XML et rend en sortie le même document annoté.

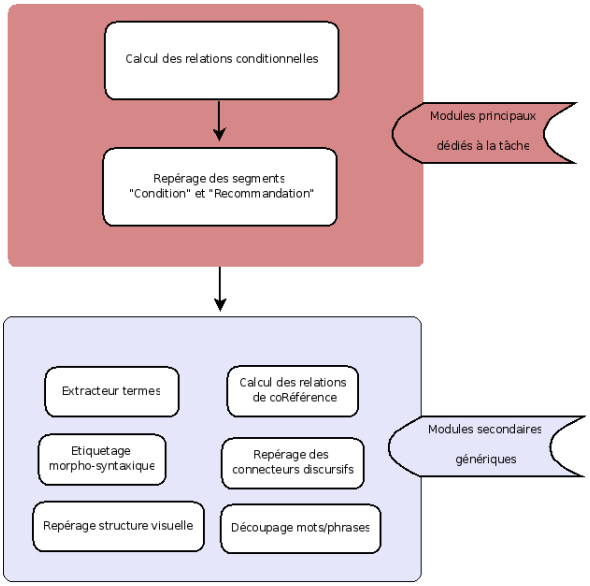


Figure 8. *Modules de GEMFrame. Les flèches signifient « fait appel à »*

L'hétérogénéité des indices à prendre en compte (du point de vue du grain, de la structure ou de la profondeur d'analyse) nous a amenés à devoir manipuler des représentations hétérogènes du texte : notamment accéder et modifier une représentation arborescente et accéder également à une représentation séquentielle grâce à des automates. C'est pourquoi, nous avons implémenté essentiellement les différents modules grâce au package Perl XML : : TWIG autorisant la définition de règles de type XPATH[9] permettant d'accéder à une représentation arborescente ainsi que d'expressions régulières permettent d'accéder à une représentation séquentielle. L'inconvénient majeur de ce langage est son manque de déclarativité. C'est pourquoi nous avons spécifié un langage de représentation des connaissances linguistiques qui nous

9. Syntaxe (non XML) pour désigner une portion d'un document XML.

a permis d'exprimer l'ensemble des règles linguistiques et qui est en cours d'implémentation.

Pour donner une brève description des deux modules principaux, le module de repérage des segments « condition » et « recommandation » s'appuie sur une trentaine de règles appliquées de manière non ordonnée et faisant appel à des marqueurs collectés manuellement par analyse de corpus. Par exemple, la règle suivante *estSuiviDe (@être, @préconisé, 3)*[10] et où les deux marqueurs doivent se suivre à plus ou moins trois mots.,exprimée à l'aide du langage précédemment mentionné, permet de repérer certaines marques de la recommandation.
Le cœur du module de calcul des relations conditionnelles est quant à lui un algorithme permettant de construire de manière incrémentale et descendante, à la manière de (Polanyi *et al.*, 2004). Il prend en entrée une liste de segments « condition » et « recommandation » préalablement identifiés. Chaque segment après avoir été dépilé est attaché à un nœud candidat de l'arbre. Le nœud d'attachement est décidé grâce à l'application des règles de rattachement obtenues par apprentissage et précédemment décrites figure 7. Voici l'exemple d'une règle, formalisée à l'aide du langage de représentation :
est_lié (segCandidat, segConditionnel) :- non_est_dans_même_phrase (segCandidat, segConditionnel), non est_intégré_syntaxiquement (segConditionnel), est_similaire (segConditionnel, titre), titre (titre), est_dans_meme_section (titre, segConditionnel)

Elle exprime que deux segments sont liés s'ils ne se situent pas dans la même phrase, que le segment conditionnel est détaché et si le segment conditionnel est redondant avec le titre de la section.

4.7. *Évaluation du travail*

La pertinence de la tâche a été validée par des médecins et chercheurs du LIM & Bio de l'université Paris 13 travaillant sur la modélisation des GBPM. C'est un domaine qui mobilise un grand nombre de travaux dans la communauté de l'informatique médicale. En outre, notre approche n'a pas fait l'objet de travaux antérieurs. Comme nous l'avons présenté en première section, la pertinence d'une approche textuelle a également été validée. Le troisième volet concerne donc l'évaluation des performances du système.

4.7.1. *Évaluation des performances du système*

Les performances du système ont été évaluées sur la sous-tâche consistant à repérer les segments « condition » et « recommandation » ainsi que celle consistant à calculer les relations conditionnelles. L'évaluation a utilisé un corpus de test composé de quatre GBPM absents du corpus d'étude et annotés manuellement avec l'aide de médecins suivant la même méthode que pour le corpus d'étude. Le tableau 4 récapi-

10. Où *@préconisé* est une classe de marqueurs tels que *préconisé*, *recommandé*, *indiqué*

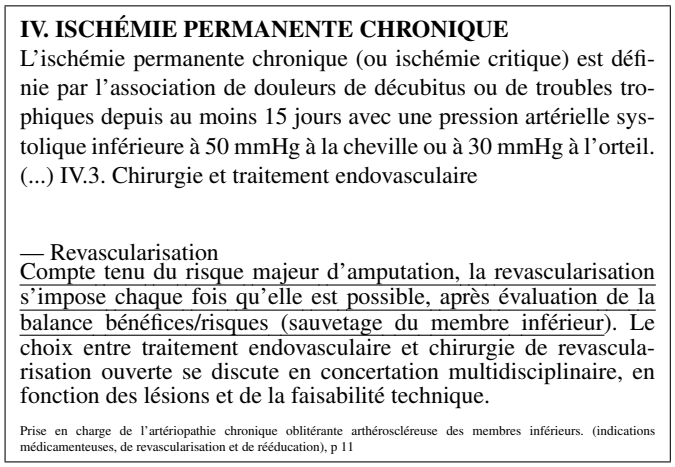
tule l'ensemble des résultats : les deux premières colonnes donnent la F-mesure [11] qui a été calculée pour le repérage des segments « conditions » et « recommandations ». La dernière colonne donne la précision obtenue pour le calcul des relations conditionnelles : un segment est considéré comme étant bien rattaché s'il est attaché dans l'arbre produit au même père que dans la référence. Il s'agit d'une décision binaire (rattachement correct ou non). Le rattachement de 638 segments a ainsi été évalué. Pour cette tâche, une mesure de rappel n'est pas pertinente car elle suppose des absences de rattachements, ce qui est impossible. Enfin, précisons que pour le calcul des relations conditionnelles, l'algorithme d'apprentissage appliqué en validation croisée nous donnait déjà des résultats d'évaluation (plus de 80 % de rattachements corrects). Cependant, ces résultats ne pouvaient pas prendre en compte les cas où une erreur proviendrait par exemple d'un mauvais repérage des indices utilisés. C'est pourquoi, il nous semblait essentiel de mener à bien une évaluation indépendante supplémentaire.

GBPM	Conditions	Recommandations	Relations conditionnelles
Chimiothérapie	0,92	0,93	0,76
Dénutrition	0,89	0,96	0,89
AOMI	0,85	0,82	0,7
Cancer du sein	0,88	0,94	0,75

Tableau 4. *Résultats pour les tâches principales du système sur quatre GBPM*

Concernant le repérage des conditions et recommandations, les résultats sont relativement élevés. C'est en fait la précision (aux environs de 0,96) qui est particulièrement bonne. Ce résultat n'est pas surprenant du fait que, les GBPM étant homogènes du point de vue du type de texte, les marques linguistiques sont faiblement ambiguës. Concernant le calcul des relations conditionnelles, les résultats sont moins hauts mais la tâche est aussi beaucoup plus difficile. On peut toutefois remarquer que ce résultat place le système dans la tranche haute des résultats obtenus avec les sytèmes d'analyse discursive cf. (Hernandez, 2004), bien que ce propos doive être pris avec précaution dans la mesure où il s'agit de tâches et de corpus différents. On constate également certaines variations entre les guides dues à des styles d'écriture et des complexités variables entre les guides. Les cas d'échec du système pour le rattachement sont notamment dus à la nécessité d'exploiter des connaissances du domaine. Par exemple, un mauvais rattachement peut être causé par un segment conditionnel manqué par le système faute de connaissances du domaine, comme dans l'exemple de la figure 9. Dans cet exemple, le titre *IV. Ischémie permanente chronique* joue le rôle de condition (en effet, toutes les recommandations présentes dans la section sont retreintes aux cas d'ischémie permanente chronique) mais elle n'est pas repérée comme telle car aucune marque linguistique explicite (comme *en cas de*, *si*, etc.) ne vient signaler son statut de conditionnel. Ce silence du système entraîne un mauvais rattachement de la recommandation soulignée.

11. Moyenne harmonique de la précision et du rappel.



IV. ISCHÉMIE PERMANENTE CHRONIQUE
L'ischémie permanente chronique (ou ischémie critique) est définie par l'association de douleurs de décubitus ou de troubles trophiques depuis au moins 15 jours avec une pression artérielle systolique inférieure à 50 mmHg à la cheville ou à 30 mmHg à l'orteil. (...) IV.3. Chirurgie et traitement endovasculaire

— Revascularisation
Compte tenu du risque majeur d'amputation, la revascularisation s'impose chaque fois qu'elle est possible, après évaluation de la balance bénéfices/risques (sauvetage du membre inférieur). Le choix entre traitement endovasculaire et chirurgie de revascularisation ouverte se discute en concertation multidisciplinaire, en fonction des lésions et de la faisabilité technique.

Prise en charge de l'artériopathie chronique oblitérante arthéroscléreuse des membres inférieurs. (indications médicamenteuses, de revascularisation et de rééducation), p 11

Figure 9. *Une erreur de rattachement due au non-repérage d'une condition*

5. Conclusion et perspectives

Dans cet article, nous avons montré l'apport d'une approche textuelle pour le TAL à travers un cadre applicatif précis : la modélisation des Guides de Bonnes Pratiques Médicales. Nous avons présenté et décrit l'application GemFrame qui permet d'automatiser partiellement le processus de modélisation en fournissant une première représentation structurée de ces textes. Nous avons opté pour une stratégie fondée sur l'exploitation de connaissances linguistiques obtenues par une méthode liant observation linguistique et apprentissage artificiel. Le système a été validé sur trois aspects : utilité, performances et pertinence de la méthode.

La question principale de ce travail reste l'étude de sa généricité, tant sur le versant applicatif (une analyse discursive est-elle utile dans d'autres cadres applicatifs ?) que sur le versant des connaissances et des traitements exploités (sont-ils « transférables » à d'autres types de textes ?). Concernant le versant applicatif, une expérience (Mondary *et al.*, 2007) a montré qu'une analyse des relations conditionnelles au niveau textuel est pertinente dans le cadre d'un système de navigation dédié aux GBPM. L'analyse discursive est utilisée dans ce cas pour focaliser l'utilisateur sur les segments textuels les plus pertinents, répondant aux contraintes de concision et de complétude. La perspective principale reste donc l'étude de la généricité du point de vue des connaissances exploitées. Nous disposons néanmoins d'ores et déjà d'éléments intéressants. Une étude linguistique détaillée et comparée aux observations présentes dans l'état de l'art (Bouffier, 2008) a montré que les indices linguistiques mis en avant sont moins dépendants du domaine (médical) que du type de texte (incitatif). En effet, beaucoup de propriétés des GBPM comme la prédominance des moyens de segmentation visuelle, la présence de marques conditionnelles ou l'hétérogénéité des types énonciatifs sont des propriétés qui ont été observées sur d'autres textes inci-

tatifs (manuels techniques, textes juridiques, etc.) (Adam, 2001). Par ailleurs, il faut souligner que les indices que nous avons retenus concernant le calcul des relations conditionnelles sont pour la plupart communs avec ceux retenus dans la littérature (Hernandez, 2004 ; Marcu, 1999) alors qu'ils portent sur d'autres types de textes (expositifs dans la plupart des cas) et concernent des relations discursives d'autres types. Notre hypothèse est que ce sont moins les indices qui varient selon le type de texte que leurs poids et leurs contributions respectives. Quelques éléments dans l'état de l'art permettent d'argumenter en faveur de cette hypothèse (par exemple, différemment de (Hernandez, 2004), nous avons trouvé que les indices lexicaux sont largement corrélés aux indices visuels, ces derniers se révélant prédominants). Cependant, les retours d'expérience dans la littérature sont trop peu nombreux pour permettre une réelle étude comparative. Cette hypothèse devra maintenant être testée par un travail expérimental consistant à appliquer notre analyse à d'autres types de texte.

6. Bibliographie

Adam J.-M., « Entre conseil et consigne : les genres de l'incitation à l'action », *Pratiques*, vol. Les textes de consigne, n 111-112, p. 7-38, 2001.

Bilhaut F., Analyse automatique de structures thématiques discursives. Application à la recherche d'information, PhD thesis, Université de Caen, 2006.

Bouffier A., Analyse discursive automatique de textes. Application à la modélisation de connaissances, PhD thesis, université Paris 13, Octobre, 2008.

Charolles M., « Cohésion, cohérence et pertinence du discours », *Travaux de Linguistique*, vol. 24, p. 125-151, 1995.

Charolles M., « L' encadrement du discours - Univers, champs, domaines et espaces », *Cahier de recherche linguistique*, 1997.

Choi F. Y. Y., Content-based Text Navigation, PhD thesis, Department of Computer Science, University of Manchester, 2002.

Cornish F., « Relations de cohérence en discours : critères de reconnaissance, caractérisation et articulation cohésion-cohérence », *CORELA*, 2006.

Cornuéjols A., Miclet L., *Apprentissage artificiel. Concepts et algorithmes*, Eyrolles, 2002.

Desclés J., « Systèmes d'exploration contextuelle », *in* C. Guimier (ed.), *Co-texte et calcul du sens*, Presses Universitaires de Caen, Caen, p. 215-232, 1997.

Georg G., Analyse informatique de Guides de Bonnes Pratiques Cliniques, PhD thesis, université Pierre et Marie Curie, Septembre, 2006.

Hagerty C., Chang J., Pickens D., Kulikowski C., Sonnenberg F., « Semi-automated Encoding of Guidelines », *Proceedings of Medinfo*, San Francisco, 2004.

Hernandez N., Détection et Description Automatique de Structures de Texte, PhD thesis, Université de Paris XI, 2004.

Jacquemin C., Jardino M., « Une interface 3D multi-échelle pour la visualisation et la navigation dans de grands documents XML », *Proceedings of IHM*, Poitiers, 2002.

Kaiser K., Akkaya C., Miksch S., « Gaining Process Information from Clinical Practice Guidelines Using Information Extraction », Aberdeen, UK, 2005.

Kurohashi S., Nagao M., « Automatic Detection of Discourse Structure by Checking Surface Information in Sentences », *COLING*, p. 1123-1127, 1994.

Mann W. C., Thompson S. A., « Rhetorical structure theory : a theory of text organization », *Text*, vol. 8, p. 243-281, 1988.

Marcu D., « A decision-based approach to rhetorical parsing », *The 37th Annual Meeting of the Association for Computational Linguistics (ACL' 99)*, The Association for Computer Linguistics, Maryland, US, p. 365-372, June, 1999.

Marcu D., « The rhetorical parsing of unrestricted texts : a surface-based approach », *Computational Linguistics*, vol. 26, n 3, p. 395-448, 2000.

Mondary T., Bouffier A., Nazarenko. A., « Between browsing and search, a new model for navigating through large documents », *EuroCogSci 2007*, Lawrence Erlbaum Associates, Delphes, Greece, 2007.

Peleg M., Boxwala A., Ogunyemi O., Zeng Q., Tu S., Lacson R., « GLIF3 : The Evolution of a Guideline Representation Format », *Proceedings of the American Medical Informatics Association*, Washington,US, p. 645–649, 2000.

Pery-Woodley M.-P., *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*, Université de Toulouse-LeMirail : ERSS, 2000.

Polanyi L., Culy C., van den Berg M., Thione G. L., Ahn D., « A Rule Based Approach to Discourse Parsing », *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*, Boston, Massachusetts, USA, May, 2004.

Quinlan R., *C4.5 : Programs for machine learning*, Morgan Kaufmann, Paris, 1993.

Saggion H., Farzindar A., Lapalme G., « Summaries with SumUM and its Expansion for Document Understanding Conference (DUC 2002) », *Workshop on Text Summarization in Document Understanding Conference (DUC)*, Philadelphia, Pennsylvania, USA, July 11-12, 2002.

Schrepfer-André G., « Les selon X énonciatifs. Portée phrastique et textuelle et indices de clôture », *Verbum*, 2005.

Séroussi B., Bouaud J., Dréau H., Falcoff H., CRiou, Joubert M., Simon G., Venot A., « ASTI : A Guideline-based drug-ordering system for primary care », *Proceedings of MedInfo*, p. 528–532, 2001.

Shiffman R., Karras B., Agrawal A., Chen R., Marenco L., Nath S., « GEM : A proposal for a more comprehensive guideline document model using XML », *Journal of the American Medical Informatics Assoc*, vol. , n 7, p. 488–498, 2000.

Teufel S., Moens M., « Argumentative classification of extracted sentences as a first step towards flexible abstracting », 1999.