Trois méthodes d'analyse pour conceptualiser le contenu de différentes sections des monographies des médicaments

Catherine Duclos, Jérôme Nobécourt, Alain Venot

Trois méthodes d'analyse pour conceptualiser le contenu de différentes sections des monographies des médicaments

Catherine Duclos, Jérôme Nobécourt et Alain Venot

Laboratoire d'Informatique Médicale et de BIOinformatique (Lim&Bio), Université Paris 13
catherine.duclos@avc.aphp.fr, {j.nobecourt,avenot}@smbh.univ-paris13.fr

Résumé : A partir de l'expérience issue de travaux de modélisation conceptuelle des connaissances contenues dans trois sections différentes des monographies des médicaments (indication, pharmacodynamie, pharmacocinétique), une analyse des méthodes de modélisation est proposée. Les différentes méthodes (pattern matching, modélisation ascendante et approche mixte) et les modalités de leur choix sont analysées en mettant en lumière des différences de nature entre les textes et l'existence de connaissances sur le domaine. Ceci nous conduit à proposer plusieurs indicateurs descriptifs de la nature du texte qui nous semblent susceptibles d'aider au choix d'une des trois méthodes proposées. Nous proposons aussi plusieurs méthodologies d'évaluation des modèles obtenus, elles aussi étant liées aux caractéristiques des textes initiaux.

Mots-clés : Médicament, Résumé des Caractéristiques Produit, TAL, Modélisation, Evaluation

1 Introduction

Pour tout médicament ayant une autorisation de mise sur le marché, un résumé des caractéristiques produit (RCP)[1] est élaboré. Ce document textuel, validé par l'autorité de régulation des produits de santé, décrit, en utilisant un vocabulaire contrôlé, les propriétés pharmaceutiques, cliniques, pharmacologiques et administratives du médicament. Le RCP est le document de référence pour le médecin qui cherche l'information légale sur le médicament.

Les RCP constituent ainsi le support essentiel pour la construction de bases de connaissances électroniques sur le médicament.

Outre un accès au RCP en simple consultation dans des sites Web ou des CDRoms, les différents éditeurs de bases de connaissances sur le médicament[2], proposent des fonctionnalités de sécurisation de la prescription comme la détection des interactions médicamenteuses ou celle des contre indications. Ceci est possible grâce à un effort de structuration de certains éléments d'information du RCP.

[1] Le RCP est aussi appelé monographie du médicament

[2] Vidal® http://www.vidalpro.net/, Banque Claude Bernard® http://www.resip.fr/

Pour les éditeurs de bases de connaissances sur le médicament, le choix des éléments à structurer est guidé par l'utilisation qui sera faite de l'information et par la facilité du travail de structuration. Par exemple, pour développer la fonctionnalité de détection des contre-indications qui implique une interopérabilité entre le dossier patient et la base de connaissances sur le médicament, les éditeurs français se sont limités à transcoder la pathologie décrite dans la contre indication à l'aide de la Classification Internationale des Maladies 10ème édition, alors que l'information y est beaucoup plus complexe comme l'ont montré (Liu *et al*.,1998).
Cette sélection des éléments d'information du RCP à structurer conduit à une perte d'information ou à des inexactitudes. Par exemple, le Cibadrex® comprimé (*hydrochlorothazide 10mg/bénazépril 12,5 mg*) a pour indication « traitement de l'hypertension artérielle, en cas d'échec d'une monothérapie par un inhibiteur de conversion » ; le fait de restreindre la structuration de cette indication à la seule pathologie « hypertension artérielle » dénature le sens de l'indication, entraîne une perte d'information et peut conduire à un mésusage de ce médicament.

Il existe peu de travaux qui, à partir d'une analyse des textes des monographies des médicaments, ont abouti à une modélisation des propriétés de ceux ci .

(Liu *et al*., 1998) ont modélisé les contre indications de 150 médicaments en réalisant une analyse sémantique manuelle de leur RCP. Si leur modèle n'a pas été évalué, il a permis de décrire avec précision les contre indications et de montrer que la plupart des terminologies médicales susceptibles d'être utilisées pour décrire le contexte clinique des contre indications étaient insuffisantes pour exprimer correctement cette information.

Dans le projet Drug ontology[3], la globalité des propriétés des médicaments ont été décrites afin d'obtenir une base de connaissances ayant une dénotation formelle basée sur le Common Reference Model d'OpenGalen[4] . Pour construire cette ontologie, les textes du British National Formulary[5] ont été étudiés manuellement, et les concepts utiles pour certaines fonctionnalités d'aide à la décision ont été retenus. Si certains concepts sont précisément décrits (comme le concept de forme pharmaceutique (Cimino *et al*, 1999)), d'autres propriétés comme la pharmacologie sont analysées très superficiellement (identification d'un unique concept « pharmacological feature »(Solomon *et al*, 1999)).

Il existe actuellement environ 5000 spécialités pharmaceutiques, l'ensemble de leurs monographies représentent un corpus de texte d'environ neuf millions de mots. Les expériences d'analyse manuelle des textes des monographies des médicaments montrent des limites liées au volume du corpus à analyser :

- Pour décrire avec précision les concepts contenus dans une section de la monographie, (Liu *et al*., 1998) se sont restreints à un corpus de texte « humainement » traitable. Pour cela, ils ont sélectionné les RCP à analyser selon des critères objectifs, mais en aucun cas ils n'ont fait mention d'une

[3] DRUG ONTOLOGY. (2005). http://www.cs.man.ac.uk/mig/projects/old/drugontology/.

[4] OPEN GALEN (2005) http://opengalen.org/.

[5] British National Formulary http://www.bnf.org/bnf/

représentativité du corpus échantillon traité par rapport au corpus global. L'universalité des concepts identifiés n'a donc pas été démontrée.

- En traitant la globalité du corpus, (Solomon *et al*., 1999) ont perdu en précision dans la description des concepts. Ainsi le concept de pharmacocinétique se limite à un « processus pharmacocinétique », alors que la lecture du texte semble véhiculer de plus riches informations qui permettraient de répondre, par exemple, aux questions suivantes : « quels antibiotiques diffusent dans l'os ? », « pour quels médicaments antiasthmatiques y a-t-il des données chez l'enfant ? ». L'exhaustivité des concepts identifiés dans cette expérience n'a donc pas été démontrée. Cet *a priori* sur le choix des concepts conduit de nouveau à une limitation des fonctionnalités pouvant être développées à partir de la structuration qui sera faite de la connaissance.

Il serait ainsi important de pouvoir disposer d'un modèle conceptuel susceptible de représenter précisément l'ensemble des propriétés des médicaments et qui aurait ainsi un caractère universel. Il faut, à la fois, pouvoir analyser la totalité des 5000 textes de RCP, trouver une méthodologie permettant un accès facilité aux éléments d'informations contenus dans ces textes et élaborer une méthode de validation des modèles conceptuels générés.

La communauté Ingénierie des Connaissances à partir de Textes (ICT) s'intéresse aux problèmes de traitement automatique du langage et de modélisation des connaissances textuelles. A partir des nombreux travaux réalisés, il est possible de proposer 3 points de vue méthodologiques.

- Les documents ont une structuration interne, ou des formats de reconnaissance (pattern) existent pour ces textes : il est possible, dans ce cas d'utiliser des outils de fouille de texte et d'utiliser la structuration du document. On se contente ici de retrouver le contexte d'une expression (groupe de mots) dans le texte (Alphonse *et al*, 2004) (Georg *et al*, 2004).
- Les documents ont été pré-traités et peuvent être passés à des outils de Traitement Automatique du Langage (TAL) pour extraire des candidats termes et des relations. Le modèle est constitué petit à petit via des regroupements conceptuels. On utilise ici des méthodes ascendantes basées sur la sémantique différentielle (Le Moigno *et al*, 2002).
- Des ressources existent pour le domaine (bases de connaissances sur le domaine, UMLS, GALEN (Trombert-Paviot *et al*, 2000*)*...). Des textes ont été analysés et des outils de TAL sont disponibles pour une analyse plus poussée. On peut alors adopter des méthodes permettant de raffiner successivement le modèle en réutilisant les outils de TAL et en spécifiant/généralisant les concepts/propriétés (Ceausu & Després, 2004).

Quel que soit le point de vue adopté, la méthode d'analyse repose sur des outils de TAL (étiqueteur, extracteur, analyseur ...) (Zweigenbaum, 1998) (Nazarenko, 2004) et sur des méthodes et/ou outils d'ingénierie des connaissances (utilisation de ressources textuelles, bases de connaissances, ontologie, création de modèles, validation, représentation formelle, opérationnalisation) (Charlet, 2002) (Biébow, 2004). Pour utiliser tous ces outils, il faut, néanmoins avoir la/les connaissance(s) d'un expert.

L'expertise fait partie intégrante de la méthode. Dans le domaine du médicament, des experts sont disponibles. Ils sont représentés par les producteurs (pharmacologue, toxicologues, cliniciens,…) et les utilisateurs (pharmaciens, médecins) de l'information.

En fonction de la nature du texte à traiter, il n'existe cependant pas de recommandations pour le choix d'une méthode particulière.

L'objectif de notre travail est de présenter différentes approches méthodologiques basées sur le traitement automatique du langage pour modéliser efficacement la totalité de l'information contenue dans les RCP des médicaments. En prenant comme illustration trois sections du RCP : la pharmacodynamie, les indications et la pharmacocinétique, nous montrons que chacune des méthodes connues en ingénierie des connaissances est applicable aux textes des RCP et que le choix de la méthode doit être guidé à la fois par la nature du texte et par la connaissances du domaine. Nous montrons également que pour évaluer la validité des modèles générés, il faut à la fois tenir compte de la nature du texte mais aussi de la complexité du modèle.

Dans la suite de cet article nous présenterons successivement les domaines de connaissance et les textes étudiés et les choix faits en terme de méthodologie de conceptualisation et d'évaluation.

2 Les domaines de connaissance

Les domaines relatifs à chaque section du RCP sont plus ou moins précisément définis : il peut exister une définition du domaine de connaissance spécifique de l'information textuelle contenue dans une section du RCP (cas de la pharmacodynamie des antibiotiques), une définition à l'échelle du thème de la section (cas de la pharmacocinétique), voire l'absence de définition précise du domaine (cas de l'indication qui recouvre le domaine de la médecine).

2.1 Connaissances spécifiques au texte d'une section du RCP

La rédaction du RCP suit obligatoirement la recommandation III/9163/89 qui définit les différentes parties du document et les informations attendues. Pour certaines classes pharmaco-thérapeutiques comme les benzodiazépines anxiolytiques et hypnotiques ou les antibactériens, il existe des instructions spécifiques de rédaction de certain points particuliers du RCP. Le but de ces recommandations est de présenter l'information qui nécessite une attention particulière selon un format commun. La section pharmacodynamie des antibiotiques est une section qui décrit l'action et l'activité des antibiotiques sur les bactéries. Elle présente la caractéristique d'être rédigée selon une de ces recommandations[6]. Le guide n°3BC5a donne une définition précise des différents chapitres que la section doit contenir et la façon dont l'information doit être présentée.

[6] European Commission, Pharmaceuticals http://pharmacos.eudra.org/F2/eudralex/Vol-3/home.htm

2.2 Connaissances spécifiques au thème d'une section du RCP

La pharmacocinétique décrit le devenir du médicament dans l'organisme. Elle correspond à une discipline scientifique à part entière (Wagner, 1993) et il existe des conceptualisations du domaine (modèle compartimental de la pharmacocinétique, modèle basé sur la physiologie). Les principaux concepts qui doivent être retrouvés dans la section pharmacocinétique du RCP sont définis dans la recommandation III/9163/89. Il n'existe pas, par contre, de description précise de la façon dont la connaissance doit être exprimée et présentée dans le texte du RCP.

2.3 Connaissances non spécifiques

La section « indication » est particulièrement importante car elle définit pour quelles pathologies le médicament peut être prescrit. Il n'existe pas de description précise de ce qui est attendu dans cette section du RCP. Son domaine couvre à la fois la pathologie et la stratégie thérapeutique.

3 Les textes issus des RCP

Les textes trouvés dans les différentes sections du RCP présentent des niveaux de langage différents : soit proches de l'énumération car il existe une terminologie contrôlée (cas des textes de pharmacodynamie) ou des expressions nominales ayant valeur de primitives conceptuelles (cas des textes d'indication) , soit utilisant la langue générale car les expressions nominales isolées n'ont de sens que dans un contexte exprimé par une phrase ou un paragraphe (cas des textes de pharmacocinétique).

3.1 Natures des textes issus de trois sections du RCP

Cents trois textes distincts de pharmacodynamie d'antibiotiques ont été identifiés dans la base des monographies de médicaments du Vidal®. Un exemple d'une section de pharmacocynamie d'un antibiotique est présentée en figure 1. L'ensemble des 103 textes constitue un corpus de 29789 mots.

SPECTRE D'ACTIVITE ANTIBACTERIENNE
Les concentrations critiques séparent les souches sensibles des souches de sensibilité intermédiaire et ces dernières, des résistantes : S < 4 mg/l et R > 16 mg/l
CMI pneumocoque : S < 0,5 mg/l et R > 2 mg/l (à titre provisoire)…
ESPÈCES SENSIBLES
Aérobies à Gram positif : *Corynebacterium diphtheriae, …, Streptococcus pneumoniae (15-35 %),*
Aérobies à Gram négatif : *Escherichia coli (10 -30 %), …*
Anaérobies : *Actinomyces, Bacteroides, Clostridium, Eubacterium, …*
Autres : *Bartonella, Borrelia, Leptospira, Treponema*
ESPÈCES MODEREMENT SENSIBLES (in vitro de sensibilité intermédiaire)
Aérobies à Gram positif : *Enterococcus faecium* (40-80%)
ESPÈCES RESISTANTES
Aérobies à Gram positif : *Staphylococcus Méti-R*

Fig. 1 - Extrait de la section pharmacodynamie de l'Augmentin® 500mg/62,5mg comprimé

Dans la base des monographies de médicament du Vidal®, 3046 textes distincts d'indication ont été identifiés (corpus de 143078 mots). . Un exemple d'une section des indications est présentée en figure 2.

De même, 1935 textes distincts de pharmacocinétique ont été identifiés dans la base des monographies de médicament du Vidal® (corpus de 318532 mots). Un exemple d'une section de pharmacocinétique est présentée en figure 3.

Elles sont limitées aux infections dues aux germes définis comme sensibles :
- chez l'adulte et l'enfant :
· en traitement initial des :
· pneumopathies aiguës
· surinfections de bronchites aiguës et exacerbation de bronchites chroniques
· infections ORL (otite, sinusite, angine) et stomatologiques
· maladie de Lyme : traitement de la phase primaire (érythème chronique migrant) et de la phase primo-secondaire (érythème chronique migrant associé à des signes généraux : asthénies, céphalées, fièvre, arthralgies...)
· en traitement de relais de la voie injectable des endocardites, septicémies

Fig. 2 - Extrait de la section indication du Clamoxyl® 500mg gélule

Absorption :Administré par voie orale, l'acébutolol est rapidement et presque complètement résorbé ; toutefois, l'effet de premier passage hépatique est important et la biodisponibilité est de 40%; .. ;
Métabolisme :La majorité de l'acébutolol est transformée au niveau hépatique en un dérivé N-acétylé, le diacétolol, qui est un métabolite actif ; le pic de concentration plasmatique de ce métabolite est atteint au bout de 4 heures environ, et les concentrations plasmatiques de diacétolol représentent le double de celles de l'acébutolol.
Distribution : Liaison aux protéines plasmatiques : la liaison aux protéines est faible : 9 à 11 % pour l'acébutolol, 6 à 9 % pour le diacétolol.
Elimination :L' acébutolol et le diacétolol circulants sont excrétés en majorité par le rein.
Insuffisance rénale :L'élimination urinaire est diminuée et les demi-vies de l'acébutolol, et plus encore du diacétolol, augmentent. ..

Fig. 3 - Extrait de la section pharmacocinétique du Sectral® 200mg comprimé

3.2 Spécificité des différents textes

En utilisant les outils de statistiques disponibles dans Cordial[7], il nous a été possible de décrire ces corpus de texte en terme de nombre de phrases, nombre de phrases verbales, nombre de formes, nombre d'occurrences, pourcentage de mots inconnus. Des traitements supplémentaires sur le corpus étiqueté ont permis de faire des statistiques à l'échelle du texte. Il a ainsi été possible de décrire la nature des textes d'indication, de pharmacodynamie et de pharmacocinétique. Les principaux résultats sont présentés dans le tableau 1.

Les textes des indications sont plus courts que les textes de pharmacodynamie ou de pharmacocinétique (moins de phrases, moins de mots). Les textes des indications et de pharmacodynamie sont plus proches de l'énumération que les textes de pharmacocinétique (part faible des phrases verbales). Le vocabulaire utilisé dans la pharmacocinétique est assez contrôlé puisqu'il y a une utilisation très fréquente d'un nombre limité de mots. Les textes de pharmacodynamie ont une part importante de

[7] Logiciel Cordial® http://www.synapse-fr.com/

vocabulaire inconnu (nom des bactéries écrit en latin). Le contenu des textes des indications est assez variable (peu de mots se retrouvent dans beaucoup de textes) alors qu'il existe une certaine régularité pour les textes de pharmacocinétique et de pharmacodynamie.

Table 1. Indicateurs statistiques calculés sur les 3 corpus de textes des indications, de la pharmacodynamie et de la pharmacocinétique

	Pharmacodynamie	**Indication**	**Pharmacocinétique**
Nombre de textes	103	3 046	1 935
Nombre de phrases	4 818	12 177	20 747
Nombre de phrases par texte	47 (19-456)	**4 (2-68)**	11 (2-108)
Part des phrases verbales	20%	24%	**78%**
Nombre de mots	**29 789**	**143 078**	**318 532**
Nombre moyen de mots par textes (minimum-maximum)	289 (185-2766)	**47 (2-578)**	194 (4 -1383)
Nombre de formes (noms,verbes,adverbes,adjectifs) (Part dans le corpus de leur total d'occurrences)	1135 (61%)	6215 (61%)	5838(**54%)**
Nombre de formes apparaissants plus de 100 fois (Part dans le corpus de leur total d'occurrence)	40 (27%)	142 (26%)	**269 (38%)**
Pourcentage de mots inconnus	**16%**	3%	5%
Nombre de formes apparaissant dans 25% des textes	69	**2**	32

4 Les méthodes de conceptualisation

Les méthodes utilisées sont d'une part fondées sur l'existence d'un modèle conceptuel initial complet ou partiel. Ensuite elles reposent sur l'analyse du format de présentation des données. Si le document est structuré, le pattern matching est applicable. Si le document n'est pas structuré, il faut pouvoir accéder aux éléments d'information du texte et les outils de TAL peuvent être utiles. L'étude des candidats termes produits par de tels outils peut être quasi exhaustive (cas des indications) ou être filtrée par la connaissance du domaine (cas de la pharmacocinétique). L'étude sémantique seule de ces candidats termes est efficace lorsque le texte n'est pas complexe (proche de l'énumération comme dans les indications), mais lorsque la part des phrases verbales devient importante, le sens porte sur les unités textuelles ; l'environnement contextuel des candidats termes doit alors être étudié.

4.1 Le pattern matching appliqué à la modélisation de la pharmacodynamie

Le guide n°3BC5a peut être vu comme un « gold standard » : les informations contenues dans ce standard sont toujours présentes dans le texte du RCP. Chaque chapitre, sous chapitre, définition contenue dans ce gold standard permet de proposer un modèle conceptuel de la connaissance supposée être trouvée dans la section pharmacodynamie des antibiotiques. Ce dernier est décrit dans (Duclos *et al*, 2004).

4.2 L'approche ascendante appliquée à la modélisation de l'indication

La découverte des concepts contenus dans le textes des indications s'est appuyée sur l'analyse des entités lexicales (noms, adjectifs, verbes, adverbes et unités complexes nominales (UCN)) produites par Nomino[8]. Les 10543 entités lexicales ont été étudiées par un expert pharmacien et rassemblées dans des groupes sémantiques distincts. L'analyse particulière des UCN par l'expert a permis de définir de nouveaux groupes sémantiques mais aussi les relations entre ces groupes (par exemple l'UCN « folliculite à *Trichophyton rubrum »* permet de définir la relation « étiologie » entre la pathologie « folliculite » et le champignon « *trichophyton rubrum »*).

Une fois découvert l'ensemble des groupes sémantiques et leurs relations, un travail de rapprochement des groupes, de déduction de concepts et d'abstraction a conduit à la production d'un modèle conceptuel (par exemple le concept de « puissance d'action du traitement » qui définit si le médicament est suffisant pour traiter seul la pathologie visée par l'indication ou s'il doit être associé à une autre thérapeutique, regroupe des groupes sémantiques proches comme « traitement d'appoint », « utilisé en association », « traitement complémentaire » ; le concept « puissance d'action du traitement » est une propriété du concept plus abstrait « degré d'efficacité du médicament »). Le modèle développé est décrit dans (Duclos et Venot, 2000).

4.3 L'approche mixte appliquée à la pharmacocinétique

Un noyau de concepts a été initialement établi. Ce noyau contient les concepts fondamentaux de la pharmacocinétique que l'on retrouve dans la recommandation III/9163/89 et dans des supports pédagogiques de pharmacocinétique (par exemple les concepts d'absorption, de distribution, de métabolisme et d'élimination).

La recherche de nouveaux concepts contenus dans le corpus de texte a servi à enrichir le modèle initial. La découverte de ces concepts s'est appuyée sur une analyse sélective portant sur 1127 des candidats termes (CT) parmi les 17520 produits par Lexter (Bourrigault, 1995). Ces CT ont été sélectionnés manuellement par un expert pharmacien car ils décrivaient spécifiquement un concept de pharmacocinétique (par exemple « métabolisme ») et ou semblaient importants (par exemple « insuffisance rénale »). Ces CT, étudiés hors contexte, ne permettaient pas de décrire correctement le domaine, par exemple les phrases « La [liaison aux protéines plasmatiques du principe actif Y] est de [10%]. », et « L'[administration du principe actif X] diminue la [liaison aux protéines plasmatiques du principe actif Y]. » contiennent des CT identiques pourtant elles n'ont pas le même sens, la première quantifie la liaison plasmatique, la deuxième explique un mécanisme d'interaction entre 2 principes actifs.

Pour découvrir les relations entre les CT, l'environnement lexical des CT ayant un sens similaire (par exemple « fixation », « liaison ») a été exploré. Le réseau terminologique de ces CT et les unités textuelles dans lesquelles ils apparaissaient a permis de découvrir les CT co-occurents. Ces CT co-occurrents ont été listés selon

[8] Logiciel Nomino® http://www/ling.uqam.ca/nomino

leur fréquence de co-occurrence. Pour des CT de sens conceptuellement voisin, les listes des CT co-occurrents ont été comparées et des concepts sous-jacents ont été déduits (par exemple les CT « absorption » et « métabolisme » qui représentent tous les deux un processus peuvent être comparés. « Absorption » apparaît avec « estomac », métabolisme apparaît avec « hépatique », et on peut déduire que le concept « processus » a une propriété qui est « localisation »). Les concepts et propriétés ainsi découverts ont été organisés pour enrichir le modèle initial par spécialisation ou par généralisation. Le modèle développé est décrit dans (Duclos-Cartolano et Venot, 2003).

5 Les méthodes d'évaluation des modèles conceptuels

L'évaluation tend à qualifier la validité du modèle conceptuel (totalité de la connaissance représentée, non déformation de la connaissance, précision des concepts, utilité des concepts, non redondance des concepts). Elle a pour principe commun de présenter à un ou plusieurs experts la connaissance structurée en utilisant les concepts du modèle et le texte initial pour qu'il(s) effectue(nt) une comparaison entre les deux représentations.

L'étape de production des données est plus ou moins longue, ceci étant conditionné à la possibilité de réaliser une extraction automatique, et si ce n'est pas le cas, à la longueur et à la complexité des textes à étudier.

L'évaluation nécessite de recourir à un nombre plus ou moins grand d'experts. Le nombre d'experts sollicités dépend de la lourdeur de la tâche d'évaluation. Plus le modèle est complexe et plus les textes sont longs, plus le travail de l'expert est fastidieux.

5.1 Cas du modèle de la pharmacodynamie

La méthode d'évaluation du modèle a tiré profit de la présence de mots, de phrases ou d'éléments de ponctuation caractéristiques délimitant des sections d'information. Un algorithme de pattern matching fondé sur le repérage de 40 structures caractéristiques a permis d'extraire les éléments d'information décrits dans le modèle conceptuel et de générer automatiquement une base de connaissances structurée de la pharmacodynamie des antibiotiques (la production des données a durée moins d'une heure). Les informations automatiquement extraites ont été confrontées aux informations manuellement saisies par un expert du domaine lisant les textes des monographies. Les taux de rappel et de précision ont été respectivement de 97,9% et 96,2%. L'ensemble de ce processus d'évaluation a pris moins d'une journée.

5.2 Cas du modèle des indications

Pour réaliser cette évaluation, un échantillon de 100 textes d'indication a été tiré aléatoirement parmi les 3046 disponibles. Deux experts travaillant indépendamment ont typé les informations contenues dans ces textes en utilisant les concepts contenus dans le modèle des indications. Ce travail a duré 2 jours. L'évaluation de la variabilité

entre ces experts dans l'utilisation du modèle pour transcrire l'indication a permis de qualifier la précision des définitions de concepts, l'intervalle de confiance (IC) à 95% du pourcentage de concepts précisément défini est de 89,3% à 94 ,7%. Ce travail de production des données a aussi permis d'identifier les concepts inutiles (non utilisés) et les concepts redondants (répétition d'éléments structurés dans des concepts différents). Un troisième expert a été sollicité pour définir si le modèle était capable de couvrir l'information dans l'indication. Pour cela il a donné, à chaque texte d'indication, un score compris entre 0 et 2 (0= pas de correspondance raisonnable, 2= correspondance totale). Ce critère d'évaluation développé par Chute (Chute *et al,* 1996) a qualifié à la fois la complétude du modèle et la déformation du sens de l'information que le modèle est capable de produire. Le score final obtenu a été de 1,95. Il a fallu moins d'une journée pour réaliser cette évaluation

5.2.1 La pharmacocinétique

Pour réaliser cette évaluation, un échantillon de 100 textes de pharmacocinétique a été tiré aléatoirement parmi les 1935 disponibles. Un expert a typé les informations contenues dans ces textes en utilisant les concepts contenus dans le modèle de la pharmacocinétique. Ce travail a duré 3 mois. Devant la lourdeur de la production des données, il n'était pas envisageable de solliciter un 2ème expert pour renouveler cette tâche comme dans le cas de l'évaluation du modèle des indications.
Pour pallier à ce défaut de double production des données, il a été décidé de réaliser une double évaluation en aveugle de chacun des textes. Le volume du corpus à évaluer étant trop important, il a été décidé qu'un expert n'évaluerait que 25 textes attribués aléatoirement. Au total 8 experts ont été sollicités pour cette évaluation.
L'utilisation du modèle pour décrire entièrement le texte produisant en moyenne plus de 100 informations à enregistrer par un expert, il a été décidé de faire l'évaluation au niveau de la phrase. Ainsi, pour conduire cette évaluation, chaque texte a été découpé en phrases qui ont été évaluées selon 2 critères : un critère de complétude (complètement, presque complètement, peu, pas structuré) et un critère de distorsion (entièrement déformé, déformé de façon importante, peu, pas déformé). A l'issue de cette analyse par phrase, l'expert était alors capable d'affecter pour un texte, une valeur globale pour chacun de ces critères. Comme chaque texte était évalué 2 fois, si l'appréciation des experts divergeait, une recherche de consensus était tentée en utilisant la méthode Delphi (elle consiste à refaire une évaluation à la lumière de l'ensemble des résultats obtenus). Sur les 100 textes, 93 évaluations ont été consensuelles dès le 1er tour d'évaluation, et les 7 autres dès le 2ème. Pour 95 à 100% des textes (IC à 95%) l'information n'était pas déformée et pour 83 à 95% des textes (IC à 95%), l'ensemble des concepts était représenté par le modèle. Ce travail d'évaluation a pris une semaine.

6 Discussion et conclusion

L'analyse comparative des méthodes utilisées pour modéliser l'information contenue dans différentes sections du RCP laisse entrevoir un possible scénario

méthodologique pour choisir un mode de traitement des textes en fonction de leur nature et de la connaissances du domaine. Quand le domaine de connaissance est défini à l'échelle du texte du RCP, la terminologie est contrôlée ainsi que l'organisation du texte. L'application du pattern matching permet la production automatique d'une base de connaissance mais est sensible à la qualité du texte. Toute faute d'orthographe, abréviation non prévue, variance d'expression peut altérer les performances de cette méthode. Lorsque le domaine est défini à l'échelle du thème d'une section du RCP, il est possible d'utiliser efficacement les CT issus des techniques de TAL en s'affranchissant du bruit lié à l'extraction terminologique par un filtrage des CT spécifiques du domaine. Ces CT sont le point d'ancrage de l'analyse des autres CT qui lui sont rattachés dans des contextes d'occurrence. Ce sont ici les relations entre CT qui sont importantes pour la construction du modèle. Enfin, lorsque le domaine n'est pas précisément défini, les techniques de TAL peuvent encore être utilisées efficacement si le texte est proche de l'énumération car les CT suffisent pour la construction du modèle ; par contre il n'y a pas moyen de s'affranchir du bruit associé à l'extraction terminologique

Il existe une difficulté à trouver une méthode parfaitement adaptée à l'évaluation de modèles conceptuels : les critères développés par (Guarino & Christopher 2004, Gomez-Perez, 2004) sont plus destinés à valider des ontologies, alors que les critères définis en informatique médicale par le Canon Group (Evans *et al*, 1994) sont plus destinés à valider des systèmes terminologiques. Les méthodes d'évaluation doivent être adaptées en fonction de la complexité du texte mais plus les textes se complexifient plus le coût de l'expertise augmente en terme d'organisation, de durée, de financement et d'analyse.

Essayer d'identifier la totalité des concepts contenus dans une section de RCP permet de disposer d'une ressource pour développer de nouvelles fonctionnalités. La structuration de la pharmacodynamie des antibiotiques nous conduit actuellement à développer un outil inédit de comparaison des spectres bactériens avec visualisation des prévalences de résistance, destiné au médecin généraliste, avec le soutien financier de la Haute Autorité de Santé.

Références

ALPHONSE E., AUBIN, S. BESSIERES P., BISSON G., HAMON T., LAGARRIGUE S., NAZARENKO A., MANINE A.P, NEDELLEC C., VETAH M., POIBEAU T. & WEISSENBACHER D. (2004). Event-based Information Extraction for the biomedical domain: the Caderige project. In *Proceedings of the International Workshop on Natural language Processing in Biomedicine and its Applications (JNLPBA)*. p. 43-49. Geneva. Suisse.

BIEBOW B. (2004). De Daseart à Terminae : Voyage au pays de la modélisation des connaissances à partir de textes. Habilitation à diriger des recherches. Université Paris-Nord.

BOURRIGAULT D (1995). LEXTER, a terminology extraction software for knowledge acquisition from texts. *In 9th Knowledge Acquisition for Knowledge Based System Workshop* , Banff, Canada.

CEAUSU V. & DESPRES S. (2004). Une approche mixte pour la construction d'une ressource terminologique (Tome 1). In *Actes des 15èmes journées francophones d'Ingénierie des Connaissances*. p. 211-223. Lyon.

CHARLET J. (2002). L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Habilitation à diriger des recherches - CHU Pitié-Salpêtrière. Université Pierre et Marie Curie.

CHUTE, C., COHN, S., CAMPBELL, K., OLIVER, D., & CAMPBELL, J. (1996). The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. Journal of the American Medical Association*, 3*(3), p.224 - 233.

CIMINO, J., MCNAMARA, T., MEREDITH, T., BROVERMAN, C., ECKERT, K., MOORE, M., & TYREE, D. (1999). Evaluation of a proposed method for representing drug terminology. *Proc AMIA annu Fall Symp*, p.47-51.

DUCLOS, C., & VENOT, A. (2000). Structured representation of drug indications: lexical and semantic analysis of drug indications. *Methods of Information in Medicine, 39*(1), p.83 - 87.

DUCLOS-CARTOLANO, C., & VENOT, A. (2003). Building and evaluation of a structured representation of pharmacokinetics information presented in SPCs: from existing conceptual views of pharmacokinetics associated with natural language processing to object-oriented design. *JAMIA, 10*(3), p. 271-280.

DUCLOS C, CARTOLANO GL, GHEZ M, VENOT A. (2004). Structured representation of the pharmacodynamics section of the Summary of Product Characteristics for antibiotics : Application for automated extraction and visualization of their antimicrobial activity spectra. *JAMIA*, 11(4) , p.285 – 293.

EVANS D, CIMINO J, HERSCH W, HUFF S, BELL D. (1996) : Toward a medical concept representation language. The Canon Group. *JAMIA*, 1(3), p.207-217.

GEORG G., SEROUSSI B., BOUAUD J. (2004). Synthesis of Elementary Single-Disease Recommandations to Support Guideline-Based Therapeutic Decision for Complex Polythological Patients. In *Proceedings of MEDINFO 2004*, p. 38-42. Amsterdam.

GOMEZ-PEREZ A (2004). Ontology Evaluation.An overview of OntoClean*. In Handbooks on ontologies*. S. Staab, R. Studer eds. Springer, Berlin , p 251-273.

GUARINO N, CHRISTOPHER AW (2004). An overview of OntoClean*. In Handbooks on ontologies*. S. Staab, R. Studer eds. Springer, Berlin , p. 151-171.

LE MOIGNO S., CHARLET J., BOURIGAULT D. & JAULENT M.C. (2002). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. In *Actes des 13ème journées francophones d'Ingénierie des Connaissances.* p. 229-238. Rouen.

LIU, J. H., MILSTEIN, C., SENE, B., & VENOT, A. (1998). Object-oriented modeling and terminologies for drug contraindications. *Methods of Information in Medicine, 37*(1), p. 45-52.

NAZARENKO A. (2004). Donner accès au contenu des documents textuels : Acquisition de connaissances et analyse de corpus spécialisés. Habilitation à diriger des recherches. Université Paris-Nord.

SOLOMON, W. D., WROE, C. J., RECTOR, A. L., ROGERS, J. E., FISTEIN, J. L., & JOHNSON, P. (1999). A refence terminology for drugs. *Proc AMIA Symp*, p. 152-156.

TROMBERT-PAVIOT B., RODRIGUES J.M., ROGERS J.E., BAUD R., VAN DER HARING E., RASSINOUX A.M., ABRIAL, V., CLAVEL L. & IDIR H. (2000). GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *International Journal of Medical Informatics*. 58-59, p. 71-85.

WAGNER J. (1993) Pharmacokinetics for the pharmaceutical Scientist. Basel, Technomic Publishing Company.

ZWEIGENBAUM P. (1998). Traitement automatique de la langue médicale. Habilitation à diriger des recherches - CHU Pitié-Salpêtrière. Université Paris-Nord.