# Contraindre le fond et la forme en domaine contraint: la normalisation de documents

**Aurélien Max**

*Groupe Langues, Information et Représentations (LIR)*
*LIMSI-CNRS et Université Paris-Sud 11*
*Orsay, France*
*aurelien.max@limsi.fr*

RÉSUMÉ. *La normalisation de documents en domaine contraint tels que des notices de médicaments peut être poussée plus loin que les technologies actuelles ne le permettent, et sans alourdir significativement la tâche du rédacteur technique. Nous proposons une nouvelle approche de normalisation de documents en domaine contraint qui utilise des représentations du contenu bien formées des documents en termes de buts communicatifs et des textes normalisés associés. Cette approche combine une analyse automatique et une phase de négociation interactive. La création de documents normalisés par les méthodes traditionnelles de saisie du texte est présentée et critiquée, et le paradigme émergent de la création de documents par spécification du contenu, dans lequel s'ancre notre approche, est introduit.*

ABSTRACT. *The present article shows that the normalization of documents in limited domains, such as drug leaflets, can be improved over what can be done with current technologies and without putting more constraints on the technical writer's work. We propose a novel approach to document normalization that uses well-formed content representations expressed at the level of communicative goals and associated normalized texts. This approach combines automatic analysis and interactive negociation with an expert of the domain. Traditional document authoring techniques based on text input are reviewed and critiziced, and the emerging approach to symbolic document authoring, to which our approach belongs, is introduced.*

MOTS-CLÉS : *documents techniques ; documents normalisés ; création de documents contrôlée ; analyse du contenu ; normalisation de documents.*

KEYWORDS: *technical documents ; normalized documents ; controlled document authoring ; content analysis ; document normalization.*

TAL. Volume 47 – n°2/2006, pages 199 à 224

## 1. Introduction : situation et motivations

La rédaction d'un document impose des efforts particuliers à son auteur. Celui-ci doit s'assurer qu'il transmet toute l'information qu'il souhaite partager, dans une langue et une structure d'exposition qui feront qu'il sera compris de ses lecteurs, et que ces derniers trouveront toute l'information à laquelle ils s'attendent, contraintes qui sont particulièrement importantes dans le cadre d'écrits techniques (Alamargot *et al.*, 2005).

Une démarche importante de normalisation a été entreprise pour palier le problème de la variabilité et de l'incompatibilité des formats de représentation des documents (Bonhomme, 2000), pour essentiellement dissocier les informations d'ordre typographique de celles concernant l'organisation logique des documents et éliminer les contraintes liées à l'utilisation d'un format propriétaire d'encodage particulier. En *domaine contraint*, c'est-à-dire dans des domaines de discours spécialisés et délimités, la limitation du champ communicationnel permet de plus l'énumération des types d'éléments pouvant apparaître dans les documents, ainsi qu'une certaine description de ce qui constitue un contenu bien formé. Si les efforts de normalisation peuvent bénéficier à la fois au producteur et au consommateur des documents, les normes sont en pratique souvent difficiles à définir et à appliquer. En effet, les formalismes et les outils d'assistance existants pouvant s'intégrer efficacement dans des pratiques documentaires sont limités. Or, dans le contexte de la rédaction de documents en domaine contraint, il semble que le rôle essentiel de l'intervenant humain soit la spécification du contenu, et que le respect des normes devrait être obtenu par une assistance plus ou moins poussée de la machine. Lorsque l'information est contenue dans un document rédigé, il peut être utile de l'extraire automatiquement, de l'organiser dans une structure normalisée, éventuellement de faire intervenir un expert pour lever les ambiguïtés d'analyse ou compléter un document, et finalement produire la version normalisée du document. C'est cette problématique de la *normalisation de documents en domaine contraint* qui nous intéresse dans le cadre de cet article.

Nous commençons par définir le champ d'application de la normalisation à des documents appartenant à des classes de documents bien définies. L'étude d'un corpus de documents appartenant à la classe des notices de médicaments destinées au patient et un exemple de normalisation illustreront les besoins de normalisation à différents niveaux, et serviront de guide pour établir les propriétés des documents normalisés en domaine contraint, organisées autour des *buts communicatifs*. Nous préciserons alors un processus de normalisation utilisant le formalisme d'un système de création de documents par spécification du contenu. Dans un premier temps, une analyse automatique du document à normaliser produit des représentations du contenu candidates, ordonnées par similarité décroissante avec le document. La similarité entre documents est mesurée en exploitant les prédictions faites par un générateur à partir de représentations du contenu bien formées. Dans un second temps, une phase de négociation interactive avec un expert de la classe de documents permet d'isoler progressivement la représentation du contenu véhiculant au mieux le contenu du document à normaliser, qui peut finalement être utilisée pour produire le texte normalisé. Nous présenterons

une implémentation d'un système de normalisation de documents ainsi qu'une évaluation préliminaire, et nous exposerons les limitations et perspectives de l'approche proposée.

Nous décrirons ensuite brièvement les modes de création de documents normalisés possibles. La création par saisie du texte utilise des technologies très répandues dans le domaine de la rédaction industrielle, mais laisse au rédacteur une responsabilité importante pour le respect des normes telles que nous les définissons. L'approche émergente par spécification du contenu sémantique ouvre des perspectives intéressantes : l'auteur spécifie le contenu sémantique d'un document appartenant à une classe, et la machine se charge de produire le document. Cependant, cette approche se limite aujourd'hui à des prototypes de recherche qui connaissent plusieurs limitations, telles que la difficulté de la construction des ressources utilisées et la rigidité du mode de création. Nous conclurons en décrivant les points forts et les limitations de l'approche de normalisation proposée et nous présenterons des perspectives de ce travail.

## 2. Documents normalisés en domaine contraint

### 2.1. *Documents en domaine contraint et classes de documents*

Nous disons qu'un domaine documentaire est *contraint* lorsqu'il est possible d'énumérer les concepts qui peuvent y apparaître *a priori*. Par exemple, les descriptions explications et instructions dans une notice de médicaments peuvent être énumérées par un pharmacologue pour un ensemble de médicaments. À l'inverse, le rédacteur d'un rapport boursier aura parfois à introduire des concepts qui sont *a priori* en dehors du domaine boursier. L'énumération des concepts d'une classe fermée de documents fait uniquement appel aux connaissances d'un rédacteur technique, alors que l'énumération (partielle) des concepts d'une classe ouverte de documents fait de plus appel à son imagination. Une *classe de documents en domaine contraint* est donc un ensemble de documents dont l'ensemble des concepts peut être circonscrit et pour lesquels les lecteurs peuvent avoir des attentes spécifiques. Le lecteur de la notice pour un comprimé contre le mal de tête s'attendra notamment à lire une section relative aux effets indésirables, ainsi qu'un message lui indiquant la posologie adaptée à l'évolution de sa douleur. La *bonne formation* de documents d'une classe en domaine contraint peut être considérée selon les propriétés suivantes :

– *La pertinence thématique du contenu*. La condition principale pour qu'une information soit exprimée dans un document est qu'elle contribue à changer le modèle de discours du lecteur en fonction des *buts communicatifs* à atteindre (ex. (Alamargot *et al.*, 2005)). Dans des classes de documents techniques, il s'agit typiquement d'*assertions* pour informer le lecteur et de *commandes* pour l'inciter à agir.

– *La cohérence du contenu*. Les messages exprimés dans un même document doivent être compatibles entre eux, ce qui requiert une expertise sur le domaine et

est responsable en grande partie du *contenu informatif* d'un document.[1]

– *La complétude du contenu*. Il existe un accord tacite sur l'ensemble des messages présents dans un document en domaine contraint entre son producteur et ses lecteurs, sur lequel peut peser des réglementations légales.[2]

– *La bonne formation de la structure d'exposition*. Le découpage du contenu en unités thématiques cohérentes et une progression motivée d'actes de parole visant à remplir un même but communicatif, facilitent la compréhension ainsi que la recherche d'informations.

– *La compréhensibilité de la langue*. Puisque le lecteur doit être capable d'identifier les buts communicatifs à l'origine des énoncés, les formulations doivent être univoques et la terminologie claire, en ayant recours à des formulations « étalon or », dont la compréhensibilité a été testée, et à des glossaires pour décrire les termes utilisés.

– *La cohérence d'emploi de la langue*. Utiliser les mêmes phraséologie et terminologie dans les documents d'une même classe facilite la comparaison de contenus entre documents.

Les pratiques documentaires sont aujourd'hui souvent influencées par le besoin de produire les documents en domaine contraint en plusieurs langues, bien que le contexte de rédaction demeure souvent monolingue. Typiquement, un brouillon initial est produit dans une première langue et il est ensuite traduit dans plusieurs langues cibles. Des modifications ultérieures sur la spécification des produits peuvent engendrer des mises à jour du document maître, entraînant alors une perte de correspondance avec les traductions de ses versions précédentes. Les versions finales ne peuvent donc être obtenues qu'à l'issue de la validation complète du document maître, rendant le processus essentiellement sériel, ce qui a de lourdes implications sur les délais de production de la documentation. Ceci souligne l'importance de la *traductibilité* et de l'*évolutivité* de ce type de documents.

En outre, ces documents partageant par nature une grande partie de leur contenu[3], il est important d'assurer leur *réutilisabilité* et donc de gérer les documents et ce dont ils sont constitués comme des objets réutilisables afin de pouvoir dériver de nouveaux documents à partir de documents existants. L'*adaptabilité* des documents est également importante pour pouvoir produire des versions personnalisées en fonction du type de lecteur et de ses besoins. Enfin, il est souhaitable de faciliter d'autres exploitations correspondant à des tâches fréquemment appliquées sur des documents ou des collections de documents, telles que le résumé ou la recherche d'informations.

1. Il existe, par exemple, une dépendance dans une notice pharmaceutique entre la forme d'un médicament et ses modes d'administration possibles.
2. En France les documents accompagnant les médicaments doivent par exemple contenir la mention « *Ne pas laisser à la portée des enfants* » (Agence du Médicament, 1996).
3. Cette caractéristique est particulièrement marquée pour les notices d'un même médicament vendu à des concentrations de principe actif différentes.

### 2.2. *Étude d'un corpus de documents en domaine contraint*

Afin de pouvoir confronter les caractéristiques souhaitables des documents en domaine contraint que nous venons de décrire à des exemples réels, nous avons fait l'étude d'une classe de documents particulière, celle des notices pharmaceutiques destinées au patient (Max, 2003a).[4] Notre corpus d'étude est constitué de 50 notices en anglais pour des médicaments analgésiques contenant de l'aspirine.

Cette étude a révélé plusieurs types de variations. Les structures thématiques diffèrent dans les parties qui sont rendues sous forme de sections ainsi que dans l'ordre d'apparition de ces sections. Certaines sections présentant les mêmes contenus portent des noms différents (ex. : *Directions* et *Dosage and Administration*), et les messages n'appartiennent pas toujours à la même section (ex. : certains avertissements peuvent apparaître dans les sections *Warnings* ou *Drug interactions*), alors que certains messages appropriés n'apparaissent pas dans tous les documents comparables (comme par exemple des mises en garde relatives à la consommation d'alcool conjointe à la prise d'aspirine). Les notices utilisent parfois des formulations différentes pour ce qui semble être le même but communicatif sous-jacent, révélant des choix différents des rédacteurs. Notre corpus s'est révélé assez homogène pour les choix linguistiques concernant l'expression, ce qui est vraisemblablement lié au biais introduit par une source d'information commune. Une autre étude sur la variation stylistique dans un corpus de 342 notices pharmaceutiques en anglais (Paiva, 2000) a révélé deux facteurs de variation principaux, opposant d'une part l'abstraction (par ex. utilisation de passifs sans agents) à l'engagement (par ex. utilisation de la $2^e$ personne et de l'impératif), et d'autre part la référence complète à la référence pronominale. Enfin, certaines phrases de notre corpus sont assez longues et donc parfois difficiles à comprendre.

Des réglementations légales existent et sont décrites et illustrées dans des documents tels que (Agence du Médicament, 1996) visant à assister l'industrie pharmaceutique dans l'interprétation et le respect de ces règles. Des directives précisent la structuration thématique du contenu en spécifiant les titres des sections attendues et la nature du contenu obligatoire et optionnel. Toutefois, la plupart des directives imposent la présence d'un contenu sans imposer de formulation (à l'exception des termes standards de la Pharmacopée européenne pour les formes pharmaceutiques et les voies d'administration), se limitant pour l'essentiel à des recommandations générales (ex. : utiliser un style direct, faire des phrases courtes sans subordonnées, donner les ins-

4. L'identification de classes de documents pour lesquelles il est possible d'obtenir un nombre significatif de documents, présentant en outre des variations, s'avère difficile. Le choix de la classe des notices de médicaments se justifie notamment par la disponibilité sur Internet de différentes notices pour des mêmes produits sur des sites anglo-saxons autorisés à faire de la vente en ligne, ainsi que par l'existence de collections normalisées et de directives officielles reflétant un réel besoin de normalisation. Des requêtes automatisées portant sur des noms de médicament permettent de récupérer des notices en anglais sur plusieurs sites marchands. Néanmoins, les documents comparables récupérés partagent souvent une même source, vraisemblablement produite par le laboratoire producteur, et présentent de ce fait moins de variations que s'ils avaient été rédigés indépendamment.

tructions avant des explications éventuelles, commencer les énumérations avec les cas les moins fréquents ou les plus spécifiques, etc.). Les rédacteurs de notices disposent donc d'une certaine latitude, et il est admis que les notices puissent dévier sur certains points des recommandations tant que celles-ci obtiennent de bons scores en compréhensibilité.

### 2.3. *Exemple de normalisation*

La figure 1 présente les avertissements d'une notice de notre corpus et une version normalisée correspondante.[5] La normalisation manuelle impose tout d'abord de reconnaître les *buts communicatifs* présents dans le document, qui sont parfois identifiés grâce à des fragments de textes discontinus. Certaines réalisations ne correspondent qu'approximativement à des buts communicatifs appartenant à la norme utilisée, voire ne correspondent à aucun de ceux-ci, ce qui nécessitera l'arbitrage d'un expert. Une structure de document doit alors être construite à partir des buts communicatifs retenus, et il peut être nécessaire de préciser ou d'imposer certains buts communicatifs rendus nécessaires. Enfin, une version normalisée du document peut être obtenue en associant des formulations aux buts communicatifs de la structure.

La partie très générique de la notice intitulée *Warnings* a été scindée en plusieurs sous-sections. Lorsque des buts communicatifs distincts étaient exprimés dans une même phrase, ils ont été réexprimés dans des phrases indépendantes. Par exemple, la contre-indication en cas d'allergie à l'aspirine, qui se trouvait dans la phrase complexe *Do not take this product if you have asthma, an allergy to aspirin, stomach problems...*, a été reformulée dans la section d'avertissements sur le produit sous forme normalisée (*DO NOT TAKE THIS DRUG IF YOU ARE ALLERGIC TO ASPIRIN*).

L'avertissement concernant le risque de syndrome de Reye est exprimé dans une phrase longue (*Children and teenagers should not use ...*). Si l'on considère qu'aucun autre but communicatif pour cette classe de documents ne peut être mis en compétition avec celui-ci dès lors que le terme *Reye's syndrome* est impliqué, son identification peut se fonder sur quelques indicateurs informatifs et ne pas requérir d'analyse plus fine que nécessaire. Cependant, la mise en correspondance des buts communicatifs exprimés dans un document avec ceux d'une norme pour la classe demandera en général des connaissances expertes qu'il sera difficile d'encoder et d'appliquer automatiquement. Nous présentons dans la section suivante une approche pour la normalisation de documents mettant en œuvre des prédictions fortes sur les documents d'une classe, couplée à une phase de négociation interactive avec un expert du domaine.

5. La normalisation effectuée est inspirée de différentes recommandations (par ex. (Agence du Médicament, 1996)) mais n'a pas pu être évaluée de façon détaillée auprès de professionnels de la santé.

**Drug Interaction Precautions :** Do not take this product if you are taking a prescription drug for anticoagulation (thinning the blood) diabetes or gout unless directed by a doctor.

**Warnings :** Children and teenagers should not use this medicine for chicken pox or flu symptoms before a doctor is consulted about Reye syndrome a rare but serious illness reported to be associated with aspirin. Do not take this product if you have asthma an allergy to aspirin stomach problems (such as heartburn upset stomach or stomach pain) that persist or recur ulcers or bleeding problems or if ringing in the ears or a loss of hearing occurs unless directed by a doctor. Do not take this product for pain for more than 10 days unless directed by a doctor. If pain persists or gets worse if new symptoms occur or if redness or swelling is present consult a doctor because these could be signs of a serious condition. As with any drug. If you are pregnant or nursing a baby seek the advice of a health professional before using this product. It is especially important not to use aspirin during the last 3 months of pregnancy unless specifically directed to do so by a doctor because it may cause problems in the unborn child or complications during delivery. Keep this and all drugs out of the reach of children. In case of accidental overdose seek professional assistance or contact a poison control center immediately.

**Alcohol Warning :** If you consume 3 or more alcoholic drinks every day ask you doctor whether you should take aspirin or other pain relievers or fever reducers. Aspirin may cause stomach bleeding.

**WARNINGS**
**Product warnings.** DO NOT TAKE THIS DRUG IF YOU ARE ALLERGIC TO ASPIRIN. Do not take this product for more than 10 days unless directed by a health professional. Consult your doctor if pain persists or gets worse.

**Alcohol.** Do not take alcohol when you take this drug or ask your doctor for an alternative pain reducer.

**Particular conditions.** A doctor should be consulted before taking this drug if you have any of the following conditions :
- a respiratory disorder
- stomach problems
- ulcers
- bleeding problems

**Children and teenagers.** CONSULT A DOCTOR BEFORE ADMINISTERING THIS PRODUCT TO A CHILD OR A TEENAGER AS IT CAN INCREASE THE RISKS OF A SERIOUS ILLNESS CALLED REYE'S SYNDROME.

**Pregnancy.** Consult a doctor before taking this drug if you are pregnant. Using aspirin during the last 3 months of pregnancy may cause problems to the unborn child or complications during delivery.

**Overdose.** Stop taking this drug immediately and call a poison control control center or a health professional if you have taken too much of this drug.

**Figure 1.** *Section d'avertissements brute (en haut) et normalisée (en bas) d'une notice pour un médicament analgésique*

## 3. Normalisation de documents par génération inversée floue et négociation interactive

Dans cette section, nous présentons notre travail sur le développement d'un système permettant de normaliser des documents appartenant à une classe définie. Le niveau d'analyse visé est celui des buts communicatifs. Utiliser des approches d'analyse sémantique prédictive sur le texte d'un document impose qu'il soit possible de construire des représentations du contenu bien formées de documents normalisés. Utiliser des approches d'analyse par mesure d'une similarité de contenu impose qu'il soit possible de mesurer une similarité entre un document et l'ensemble des documents possibles pour sa classe de documents, auxquels aurait été associée une description normalisée. Une formalisation de la représentation du contenu des documents normalisés apparaît donc comme centrale pour pouvoir normaliser des documents en domaine contraint. Pour chaque représentation du contenu, il existe un ensemble de textes, dont l'analyse correspond à cette représentation, parmi lesquels se trouve un texte normalisé, comme illustré sur la figure 2. Analyser un document quelconque d'une classe modélisée correspond donc à trouver sa représentation du contenu parmi celles possibles, et sa version normalisée peut être produite par génération du texte normalisé à partir de la représentation du contenu trouvée.

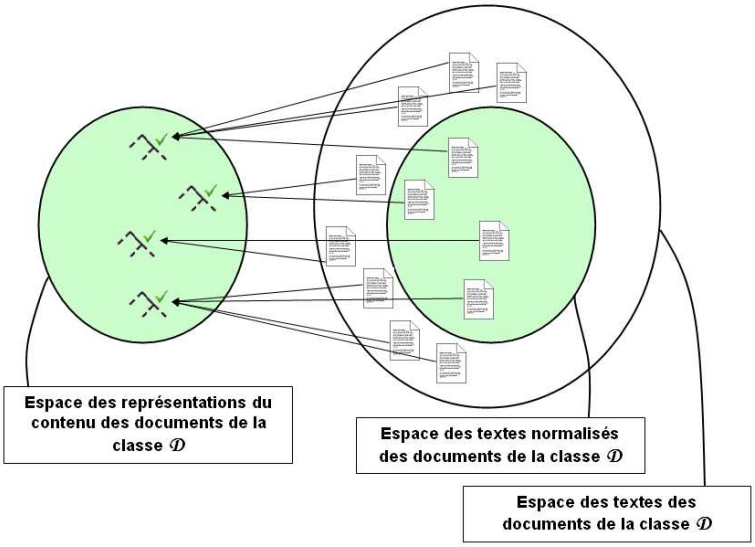


**Figure 2.** *Analyse de documents appartenant à une classe définie*

### 3.1. *Cadre de normalisation*

Normaliser un document permet de réexploiter son contenu dans un nouveau format requis, pour des raisons propres au producteur des documents, pour des raisons d'exploitation, ou encore des raisons légales.[6] Le document normalisé doit être valide par rapport à la spécification des normes pour sa classe de documents, ce qui peut impliquer l'ajout de nouveaux buts communicatifs (par ex., ajout d'une nouvelle disposition légale) ainsi qu'une restructuration. Un but communicatif présent dans un document (par ex., un message publicitaire) peut ne pas appartenir au répertoire de la norme utilisée, et sa non-sélection par le processus de normalisation s'apparente à un résumé par sélection du contenu. Des remplacements de terminologie et de phraséologies peuvent également être apportés pour améliorer la compréhensibilité et la cohérence des documents.

À la lumière des propriétés défendues en 2.1 et des observations résultant de notre étude de corpus, il est important de définir un cadre permettant de normaliser les documents d'une classe bien définie.[7] Nous proposons de considérer la normalisation d'un document en domaine contraint selon trois axes : les buts communicatifs, leur agencement pour former des documents normalisés, et les formulations en langue qui leur sont associées.

L'identification des buts communicatifs exprimables dans une classe de documents permet de n'autoriser que ceux-ci dans les documents, et formalise ainsi le contenu informationnel qui est propre à cette classe. Les buts communicatifs devant être retenus sont ceux qui jouent un rôle informatif dans leur classe de documents, et ils peuvent être décrits à un niveau d'abstraction assez élevé, indépendamment de toute formulation. Un but communicatif a un certain type, il peut être atomique ou complexe lorsqu'il met en jeu d'autres buts communicatifs, et il peut être paramétré lorsque des valeurs spécifient sa portée.

La structure des représentations du contenu sémantique des documents met en jeu les buts communicatifs de la classe dans des structures complètes et cohérentes, et correspond à la structure thématique arborescente normalisée du document. Par exemple, les normes du Vidal de la famille (Vidal, 2006) définissent la structure thématique d'une notice de médicament comme se composant des sections *Présentation*, *Composition*, *Indications*, etc. Les sections sont ici assimilables à des types de buts communicatifs : en effet, la section *Présentation* vise à remplir le but communicatif complexe *Informer sur le nom du produit, sa forme pharmaceutique, son conditionnement, etc.*

Des formulations doivent enfin être associées aux buts communicatifs avec une progression rhétorique appropriée, et être utilisées de façon cohérente dans tous les documents de la classe (pour un même lectorat). Par exemple, le but communicatif *In-*

6. Cela permet également de créer des documents normalisés par normalisation de documents écrits très librement, à contre-courant des approches rigides de création par saisie du texte décrites en 4.1.
7. La présentation des documents est par exemple abordée dans (Bouayad-Agha *et al.*, 2000).

*diquer que le médicament ne peut pas être pris en cas d'allaitement* peut être formulé comme une instruction suivie d'une explication, ex. : *Consult a doctor before taking this drug if you are pregnant. Using aspirin during the last 3 months of pregnancy may cause problems to the unborn child or complications during delivery.*

Nous avons utilisé le formalisme des Grammaires d'Interaction de l'approche MDA (Dymetman *et al.*, 2000; Brun et Dymetman, 2002) dérivé des Grammaires à Clauses Définies de Prolog, et qui offre un couplage fort entre représentation sémantique et réalisation textuelle, pour décrire les documents normalisés d'une classe. Un système MDA permet à un auteur de spécifier le contenu de documents bien formés en domaine contraint via des *textes de contrôle*, comme illustré sur la figure 3. Un texte de contrôle contient des *ancres* qui permettent à l'auteur de spécifier des valeurs sémantiques valides non encore spécifiées dans la représentation sous-jacente du document. Les représentations du contenu des documents sont utilisées par un générateur pour produire des textes de contrôle en plusieurs langues correspondant à la spécification courante du contenu du document.

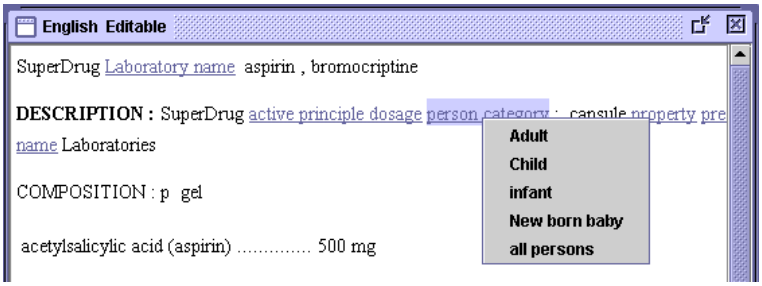


**Figure 3.** *Interface MDA de création de notice de médicament*

La figure 4 présente un extrait d'une grammaire décrivant une notice de médicament en anglais ayant servi pour produire la version normalisée du document de la figure 1. Le formalisme permet de définir la bonne formation sémantique en termes de types autorisés et de valeurs possibles pour ces types, et des dépendances sémantiques entre ces valeurs peuvent être établies. Par exemple, `warnings` est un objet sémantique de type `leafletWarnings` paramétré par une variable `MAIngr` (*main active ingredient*), composé d'objets de types `productWarnings`, `particularConditionsWarnings`, etc. Les éléments commençant par une majuscule dénotent des variables sémantiques ; ainsi, par exemple, `PartCondW` est de type `particularConditionsWarnings`. Le paramètrage des types offre la possibilité d'exprimer des dépendances sémantiques par unification de variables.

Le caractère récursif du formalisme est utilisé pour représenter des structures de listes où un type d'objet peut apparaître plusieurs fois. Le type `listOfParticularConditions_rec1` peut, par exemple, être composé d'un seul objet de type `condition` (dénotant une seule condition du patient), ou d'un objet de type `condition` suivi d'un objet de son propre type (dénotant une liste d'au moins

```
% main warning section
warnings(ProdW PartCondW DrugIntW PossSEW)::leafletWarnings(MAIngr)-e-[] -->
    ['<section title="WARNING">']
    ProdW::productWarnings(MAIngr)-e-[]
    PartCondW::particularConditionsWarnings(MAIng)-e-[]
    DrugIntW::drugInteractionsWarnings(MAIngr)-e-[]
    PossSEW::possibleSideEffects(MAIngr)-e-[]
    ['</section>'].

% subsection on particular conditions
noParticularCondition::particularConditionsWarnings(MAIng)-e-[] --> [].
particularConditions(LOPC)::particularConditionsWarnings(MAIng)-e-[] -->
    ['<subsection title="Particular conditions">A doctor should be consulted before taking
    this drug if you have any of the following conditions:']
    LOPC::listOfParticularConditions_rec1(MAIng)-e-[]
    ['</subsection>'].

moreThanOneCondition(PC PCs)::listOfParticularConditions_rec1(MAIng)-e-[]  -->
    PC::condition(MAIng)-e-[]
    PCs::listOfParticularConditions_rec1(MAIng)-e-[].
oneCondition(PC)::listOfParticularConditions_rec1(MAIng)-e-[]  -->
    PC::condition(MAIng)-e-[].

% conditions
asthma::condition(_)-e-[] --> ['asthma'].
asthma_1::condition(_)-e-[] --> ['a respiratory disease'].
asthma_2::condition(_)-e-[] --> ['a respiratory disorder'].
stomachProblems::condition(_)-e-[] --> ['stomach problems'].
```

**Figure 4.** *Extrait de grammaire MDA pour une section d'avertissements*

deux conditions). Le marqueur de type `_recN` a été ajouté pour nos besoins comme marqueur de type récursif afin de limiter la profondeur de l'analyse, en acceptant au plus `N` occurrences d'un même objet sémantique dans une structure récursive.

Le formalisme couple à la syntaxe sémantique abstraite une syntaxe concrète pour une langue donnée, l'anglais dans cet exemple (indiqué par le suffixe de type `-e`). Les objets sémantiques sont ordonnés comme l'impose la réalisation dans une syntaxe concrète particulière, et des fragments de texte (entre crochets) peuvent être intercalés. La génération du texte d'un document est un processus compositionnel qui parcourt récursivement les nœuds de son arbre abstrait, concatène les chaînes de caractères et remonte le résultat au nœud parent. La chaîne associée à la racine d'un arbre abstrait représente donc le texte de l'ensemble du document.[8] Nous avons étendu le formalisme pour qu'il autorise une notation indexée pour les objets sémantiques. Les objets `asthma`, `asthma_1` et `asthma_2` sont par exemple tous sémantiquement équivalents, mais sont associés à des formulations différentes (dans ce cas, extraites des

8. Dans le mode de création des systèmes MDA, les objets sémantiques non instanciés sont représentés par des chaînes de caractères prédéfinies décrivant les types (l'étiquette d'une ancre) et les valeurs possibles (les choix présentés dans un menu quand on sélectionne l'objet) (voir figure 3).

ensembles de synonymes de WordNet). Nous expliquerons plus loin la façon dont ce non-déterminisme est utilisé en analyse.

L'utilisation de Grammaires d'Interaction pour la normalisation automatique de documents donne à la fois accès à un formalisme ayant de bonnes caractéristiques et à des ressources existantes. Elle permet en outre de considérer la normalisation d'un document sous un angle particulier : un expert lit le document à normaliser puis utilise une grammaire en mode de création MDA. Pour chaque ancre dans le texte de contrôle, il cherche à répondre à la question sous-jacente (par exemple, *existe-t-il des contre-indications pour les enfants ?*) d'après sa compréhension du document et les choix sémantiques qui sont encore possibles. La simulation de cette méthodologie consiste alors à recréer automatiquement les choix qu'aurait faits un auteur pour décrire le document dont le contenu est le plus proche de celui du document analysé.

### 3.2. *Une approche par recréation du contenu automatique et interaction*

Notre approche consiste à utiliser une grammaire MDA décrivant les documents normalisés d'une classe pour faire des prédictions sur le texte des documents, qui seront utilisées pour mesurer une similarité avec le document devant être normalisé. La motivation principale pour cette approche est que l'espace des représentations du contenu valides de documents normalisés est beaucoup plus restreint que l'ensemble des textes pouvant exprimer ces contenus, et donc qu'une analyse du texte d'un document plus fine que nécessaire pour discriminer serait contre-productive.[9] Il faut tout d'abord simuler la construction de représentations du contenu à l'aide de la grammaire utilisée, puis mesurer une similarité avec le document. Pour cela, nous utilisons des prédictions faites par un générateur à partir de représentations du contenu produisant des *descripteurs* du contenu. L'analyse du document à normaliser au même niveau de description permet alors de mesurer une similarité floue entre le contenu de ce document et le contenu d'un document normalisé. Une recherche heuristique dans l'espace des représentations du contenu possibles peut guider la construction des représentations les plus prometteuses d'abord. La figure 5 illustre ce processus de recherche : différentes représentations du contenu complètes sont trouvées par score de similarité décroissante entre les textes qui leur sont associés et le document à normaliser. La génération du texte associé à la représentation obtenant le meilleur score produit une version normalisée du document obtenue de façon automatique.

Nous appelons ce type de recréation du contenu sémantique des documents *génération inversée*, puisqu'elle utilise les prédictions d'un générateur pour mesurer des similarités de contenu en vue de faire de l'analyse. Un générateur ne pouvant que sous-générer relativement à l'espace des descripteurs possibles pour un même contenu communicatif, nous rendons cette génération inversée *floue* afin qu'elle soit résistante

9. Cela n'exclut pas le recours à des techniques d'analyse fines dans une deuxième passe comme par exemple celles décrites dans (Blanchon, 2002) et (Brun et Hagège, 2003), si des ressources adéquates sont disponibles.

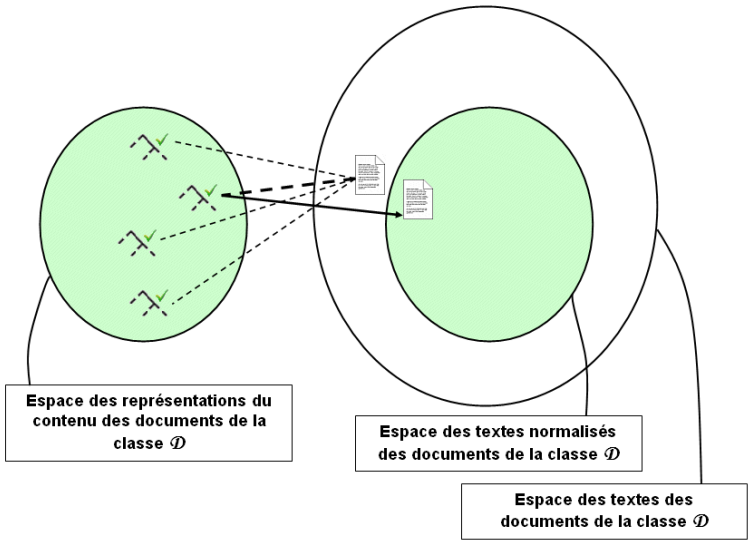


**Figure 5.** *Recherche heuristique de la représentation du contenu obtenant la meilleure similarité avec un document à normaliser*

à la variabilité linguistique rencontrée dans les documents. Il s'agit donc d'une forme de *réversibilité floue* d'un processus de génération à des fins d'analyse, et c'est la structure de représentations du contenu bien formées qui guide l'analyse. De plus, cette approche réutilise des ressources existantes (couplage fort des représentations sémantiques et des réalisations textuelles des grammaires MDA) et ne requiert pas la construction de nouvelles connaissances.

Compte tenu des limites des mesures pouvant être utilisées pour mesurer une similarité sémantique basée sur les textes, le processus produit un ensemble de représentations du contenu candidates à l'analyse d'un document. Ces représentations sont ensuite factorisées en une structure unique révélant les sous-spécifications résultant de l'analyse. Nous proposons une approche pour la sélection de la représentation du contenu représentant au mieux la version normalisée du document par négociation interactive avec un expert de la classe de documents.

### 3.3. *Analyse automatique d'un document par génération inversée floue*

La génération inversée floue repose sur l'hypothèse qu'il existe un document dans l'ensemble des documents virtuels d'une Grammaire d'Interaction qui soit suffisamment proche du document analysé. L'espace sémantique des représentations bien formées du contenu d'une classe de documents étant potentiellement immense, une énu-

mération exhaustive des documents virtuels n'est pas envisageable. Nous considérons donc l'identification de cette représentation du contenu comme un problème de recherche d'un maximum global dans l'espace sémantique décrit par la grammaire utilisée. Nous utilisons une recherche heuristique *admissible* (par ex. (Nilsson, 1998)) qui garantit que le premier document virtuel trouvé est le plus similaire avec le document à analyser pour une fonction d'évaluation donnée. Cette fonction d'évaluation correspond à une mesure de similarité de contenu entre documents. La procédure est proche d'une analyse syntaxique descendante où des représentations partielles du contenu des documents sont itérativement construites en tenant compte de la mesure de similarité avec le contenu du document.

Un nœud dans l'espace de recherche est une représentation du contenu d'un document, soit un arbre sémantique abstrait typé. Les successeurs d'un nœud sont obtenus en appliquant un pas de dérivation à l'arbre partiel : une variable non instanciée de l'arbre est tout d'abord choisie, puis l'ensemble des objets compatibles avec son type et menant à une structure valide est extrait et permet de créer une liste de successeurs qui sont alors insérés dans la liste ordonnée de recherche. La modélisation de la classe de documents opère ainsi un filtrage qui limite l'espace de recherche en ne produisant que des représentations du contenu valides.

Une procédure de recherche heuristique admissible utilise une approche en *meilleur d'abord* qui fait l'expansion du nœud le plus prometteur à chaque étape de la recherche. La fonction d'évaluation utilisée doit donner une indication de la plus forte similarité pouvant être obtenue entre une représentation du contenu accessible depuis le nœud et le document à normaliser. Pour que la recherche soit admissible la fonction d'évaluation doit être *optimiste*, elle doit donc surestimer la valeur réelle de la similarité entre le document analysé et tout document virtuel pouvant être produit à partir d'une représentation partielle du contenu. De plus, la fonction d'évaluation doit décroître au fur et à mesure de la progression de la recherche. Dans notre implémentation, nous utilisons une fonction basée sur une similarité entre les documents virtuels de l'espace de recherche et le document à normaliser. La similarité d'une représentation partielle avec le document doit être au moins égale à chacune des mesures de similarité entre l'ensemble des représentations complètes accessibles depuis cette structure partielle et le document. Cette valeur doit donc décroître au fur et à mesure qu'une représentation partielle est davantage spécifiée, en tendant vers la valeur réelle de similarité pour chaque représentation complète produite à partir de cette représentation partielle.

En recherche d'information, les documents sont typiquement représentés par des descripteurs de leur contenu. En fonction du type visé (formes de surface, lemmes, termes, syntagmes, etc.), l'extraction automatique des descripteurs peut être réalisée avec plus ou moins de précision. Si nous ne voulons avoir recours à d'autres connaissances sur le domaine des documents que celles contenues dans la grammaire utilisée, nous pouvons extraire les lemmes avec une bonne précision au moyen d'un analyseur

| Objets sémantiques | Profils partiels (nombres d'occurrences) |
|---|---|
| warnings | {"warning" : 1 "reye" : 1 "doctor" : 3 "consult" : 3 "condition" : 1 "asthma" : 1 "respiratory" : 1 "disease" : 1 . . . } |
| particularConditions | {"doctor" : 1 "consult" : 1 "condition" : 1 "asthma" : 1 "respiratory" : 1 "disease" : 1 . . . } |
| asthma | {"asthma" : 1} |
| asthma_1 | {"respiratory" : 1 "disease" : 1} |

**Figure 6.** *Profils partiels pour des objets de la grammaire de la figure 4*

morpho-syntaxique tel que le TreeTagger.[10] Un document est représenté par un *profil* qui associe aux descripteurs qu'il contient un poids qui est le produit de leur nombre d'occurrences dans le document par leur informativité. Cette informativité correspond à la fréquence inverse d'apparition dans un corpus de la classe de documents, afin que les descripteurs les moins fréquents soient considérés comme plus significatifs.[11]

Afin de pouvoir calculer le score d'un nœud de notre espace de recherche, nous devons pouvoir calculer la similarité entre la représentation partielle du contenu associée à ce nœud et le document en cours d'analyse, ce qui impose de généraliser la notion de profil de descripteurs au cas d'une représentation partielle. Une représentation partielle représente une *potentialité* de textes virtuels obtenus en la complétant de toutes les façons possibles. Nous pouvons donc propager une certaine connaissance des profils de descripteurs de ces textes virtuels vers les représentations du contenu partielles et l'utiliser pour mesurer une borne supérieure de la similarité entre ces textes virtuels et le texte d'un document à normaliser.

Le profil d'un type peut donner une mesure du poids maximal d'un descripteur pouvant être obtenu en dérivant ce type de toutes les façons possibles. Cela représente le fait que, quelle que soit la dérivation faite pour un type, un descripteur donné peut apparaître au plus un certain nombre de fois. Le profil d'un type sémantique s'obtient en prenant pour chaque descripteur son poids maximal dans les profils de chacun des objets de ce type. On peut alors construire le profil d'un sous-arbre sémantique dont la racine est un objet sémantique. Un profil initial est construit, contenant l'ensemble des descripteurs présents dans les chaînes de caractères en partie droite de la définition de l'objet dans la grammaire (son *profil constant*). Ensuite, les profils des éléments fils de l'objet sémantique dans l'arbre sémantique sont construits récursivement, et l'union de leur profil avec le profil constant donne le profil de l'objet sémantique. La figure 6 donne quelques exemples de profils partiels d'objets sémantiques construits à partir de la grammaire de la figure 4.

Le profil associé à un arbre sémantique abstrait partiel ou complet s'obtient donc en prenant l'union des profils associés à chacune de ses feuilles (objets sémantiques

10. Nous envisageons également une extraction de termes et un repérage de leurs variantes à l'aide de l'outil Fastr (Jacquemin, 2001).
11. Ce type de pondération est issu de la mesure *tf.idf* utilisée en recherche d'information.

ou variables typées) et des profils constants associés à chacun de ses nœuds intermédiaires. Le profil d'un arbre partiel correspond ainsi à une surestimation du contenu en termes de descripteurs de chacun des documents virtuels atteignables. Le profil d'un arbre complet correspond, quant à lui, exactement au profil d'un document virtuel particulier. Les profils de descripteurs des objets et des types sémantiques d'une Grammaire d'Interaction sont calculés de façon statique par un processus de précompilation utilisant une approche à point fixe (Max, 2003a), dont le but est d'associer des profils aux objets et aux types, et de diffuser progressivement ces profils dans la grammaire jusqu'à ce que le profil du type sémantique racine de la grammaire soit construit.

La fonction suivante, qui répond aux critères ci-dessus, est utilisée comme fonction d'évaluation de la recherche heuristique :[12]

$$sim(P_{doc}P_{rep}) \quad = \quad \sum_{d \in \mathcal{D}} min(poids(dP_{doc})poids(dP_{rep}))$$

$P_{doc}$ et $P_{rep}$ représentent respectivement le profil du document et le profil d'une représentation du contenu. $d$ est un descripteur et $\mathcal{D}$ est l'ensemble des descripteurs présents dans les deux profils. La fonction $poids$ retourne le poids d'un descripteur dans un profil.

La mesure de similarité automatique utilisée ne peut prétendre correspondre précisément à la similarité de contenu telle qu'elle serait évaluée par un expert humain. Cette approximation de la similarité du contenu impose de continuer la recherche pour extraire les $N$ documents virtuels par score de similarité décroissant, pour les soumettre à la validation d'un expert.

Les Grammaires d'Interaction ont en outre été rendues non-déterministes, ce qui permet d'associer plusieurs textes à un même objet sémantique afin de mieux couvrir l'espace des textes. Ceci est réalisé en associant aux objets sémantiques des objets secondaires (`asthma_1` relativement à `asthma` dans la figure 4). Ces derniers respectent la syntaxe abstraite de leur objet sémantique de référence, mais peuvent produire de nouveaux textes, et donc avoir des profils de descripteurs différents. Ces objets sémantiques entrent en compétition les uns avec les autres lors de la génération inversée floue, mais une fois les arbres sémantiques abstraits candidats obtenus, ils sont normalisés en leur objet de référence. Ce mécanisme permet notamment de faire de l'apprentissage supervisé de nouvelles formulations dans un système de normalisation.

### 3.4. *Sélection d'une représentation du contenu par négociation interactive*

Compte tenu du caractère sensible de la tâche de normalisation de documents, le compromis concernant le document normalisé qui véhicule le mieux le contenu du document d'origine ne peut être effectué que par un expert de la classe de documents.

12. Voir (Max, 2003a) pour une démonstration de l'admissibilité de la recherche.

Il existe au moins deux façons pour l'expert de consulter les résultats de l'analyse automatique : il peut soit lire les documents normalisés générés à partir de leur représentation du contenu, ce qui peut être long et laborieux, soit consulter ces représentations dans un format qu'il peut interpréter. Les arbres sémantiques candidats peuvent être factorisés dans une structure unique, révélant ainsi des incertitudes d'analyse pour les parties distinctes de différents candidats.

Lorsque l'expert rencontre une sous-spécification pour un type sémantique dans l'arbre sémantique abstrait factorisé, il peut identifier l'objet en compétition qui lui semble correct et le valider. L'opération de *validation* a pour effet d'élaguer la représentation factorisée en éliminant non seulement les sous-arbres dominés par les objets en compétition directe avec l'objet validé, mais également l'ensemble des sous-arbres n'appartenant pas à des représentations globales bien formées compatibles avec la présence de l'objet validé. L'opération d'*invalidation* permet au contraire d'indiquer qu'un objet n'appartient pas à la représentation globale du document normalisé, et a pour effet d'éliminer de la représentation factorisée les sous-arbres qui ne sont compatibles qu'avec cet objet.

L'expert doit donc pouvoir naviguer dans la structure factorisée. En descendant dans cette structure et en passant éventuellement par plusieurs sous-spécifications, l'expert peut remettre à plus tard des décisions qu'il souhaiterait prendre sur la base d'informations supplémentaires. Des opérations de validation ou d'invalidation peuvent entraîner des éliminations de sous-arbres non compatibles de façon remontante. Il peut par ailleurs s'avérer intéressant pour l'expert d'avoir une idée globale de l'importance d'une sous-spécification particulière. Les sous-spécifications présentes peuvent être énumérées dans une liste accessible à l'expert ordonnée par importance décroissante afin de minimiser le nombre d'opérations à effectuer pour identifier l'arbre résultat. La fonction calculant l'importance d'une sous-spécification attribue un score d'autant plus fort qu'en moyenne les scores des objets sémantiques en compétition sont forts et qu'ils appartiennent à peu d'arbres sémantiques, favorisant ainsi les sous-spécifications impliquant les objets sémantiques les plus probables ainsi que celles qui permettent de faire des résolutions qui conservent le moins d'arbres sémantiques candidats (la validation d'un objet appartenant à peu d'arbres candidats aura pour effet d'éliminer davantage d'arbres candidats, donc éventuellement davantage de sous-spécifications).

### 3.5. *Implémentation d'un système de normalisation interactive et évaluation*

Nous avons développé un système de normalisation interactive implémentant la génération inversée floue et la négociation interactive présentées. L'utilisateur charge tout d'abord un document à normaliser, puis il peut sélectionner une grammaire pour le normaliser et spécifier un degré de confiance du système qui définit un nombre de représentations candidates maximal à trouver. À l'issue de l'analyse par génération inversée floue, l'interface présente plusieurs vues permettant à l'utilisateur de consulter les résultats et de résoudre les sous-spécifications présentes et qui sont mises à jour

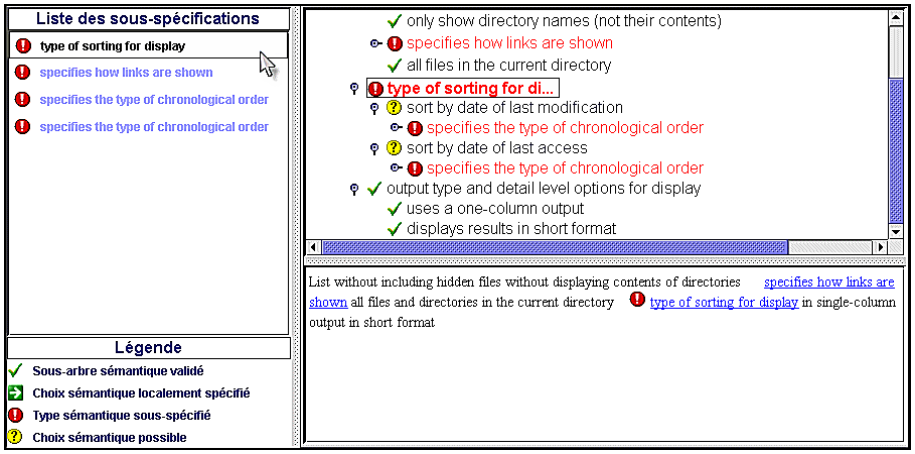


**Figure 7.** *Interface du système de normalisation après analyse automatique*

après chaque résolution. La figure 7 montre les principaux éléments de l'interface du système après analyse d'un document court correspondant à la description en langue d'une commande UNIX.[13]

La partie gauche présente une liste ordonnée des sous-spécifications d'analyse restantes par ordre d'importance décroissante. Cliquer dans cette vue déclenche des dialogues de négociation, qui pourront être traités individuellement ou en cascade. Cette dernière possibilité permet à l'utilisateur de normaliser un document uniquement en répondant à des questions ordonnées par le système.

Une vue sémantique factorisée (en haut à droite) présente sous forme arborescente les éléments de la structure factorisée contenant des sous-spécifications et les différentes valeurs possibles ordonnées par score décroissant. Cette vue permet d'utiliser l'opération de validation pour confirmer qu'une hypothèse appartient bien à la solution, et l'opération d'invalidation pour exclure une hypothèse et toutes celles qui lui sont liées. Des pictogrammes, décrits dans la légende en bas à gauche, sont utilisés pour aider l'utilisateur à repérer les sous-spécifications nécessitant son intervention.

Enfin, une vue MDA (en bas à droite) reprend en partie le mode de création des système MDA, en présentant sous forme textuelle les éléments du document qui ont été reconnus et normalisés et sous forme d'ancres les types sous-spécifiés les plus proches de la racine de l'arbre sémantique. Le texte de contrôle permet ainsi à l'utilisateur de confronter la version textuelle normalisée à la représentation du contenu qui est affichée dans la vue sémantique factorisée. Un clic sur une ancre présente un menu

13. Ce type de documents permet de trouver facilement des utilisateurs experts, ce qui nous a été beaucoup plus difficile pour le domaine pharmaceutique.

contenant les objets sémantiques en compétition pour la variable typée correspondante par score décroissant, ce qui permet de faire une opération de validation avec l'objet choisi.

Les différentes vues permettent ainsi à un expert de reconstruire progressivement la représentation du contenu d'un document appartenant à une classe pour laquelle on dispose d'une Grammaire d'Interaction. Cette représentation permet de générer le texte d'une version normalisée du document qui apparaît progressivement dans la vue MDA. Il est également possible d'utiliser une grammaire *parallèle* à la grammaire utilisée pour la normalisation (même syntaxe abstraite mais syntaxe concrète différente) pour produire d'autres versions du document normalisé, notamment dans d'autres langues, ce qui revient à faire une *traduction normalisatrice* du document de départ (Max, 2003b).

L'évaluation de ce type de traitement se confronte à plusieurs problèmes. Tout d'abord, il est difficile d'obtenir des corpus de test en quantité significative et présentant suffisamment de variations avec le corpus de développement utilisé pour la conception des grammaires. En outre, il faudrait idéalement que des experts du domaine participent à la conception des grammaires (Brun *et al.*, 2003), sinon les grammaires développées risquent d'être peu naturelles pour des experts et d'être biaisées pour les documents du corpus de développement. Alors que certains buts communicatifs peuvent être identifiés de façon assez fiable (comme par exemple celui concernant le risque de syndrome de Reye évoqué plus haut), certains phénomènes tels que l'absence d'élément ou la négation ne pourront être capturés sans recours à des techniques d'analyse plus fines.

On constate empiriquement que le document finalement retenu, via l'interface du système, a presque toujours été produit par le processus de génération inversée floue pour des nombres de candidats trouvés demandant un temps d'analyse raisonnable. Dès lors, il semble approprié de mesurer la performance de l'approche proposée par une évaluation par la tâche. Nous avons mené des expériences informelles sur la tâche de normalisation de documents courts correspondant à la description en anglais de commandes UNIX. Ces commandes devaient être saisies (parmi un sous-ensemble des commandes principales), puis normalisées avec le système par des utilisateurs ayant une expertise allant de novice à expert.

Les novices ont apprécié le type d'aide proposé, puisqu'il leur permettait de préciser de façon incomplète ce que devait réaliser une commande et le système les guidait ensuite pour compléter les éléments manquants. De plus, la reformulation de la vue MDA offre un moyen efficace pour contrôler ce qui a déjà été reconnu. En général, ils utilisent peu la vue sémantique factorisée, car elle présente vraisemblablement trop d'informations, alors que le mode de dialogue de négociation en cascade représente la méthode la plus souple pour ce profil d'utilisateur. Enfin, on note un recours parfois important à la fonction d'annulation (restauration d'un état antérieur à choisir parmi une liste). Si en moyenne le temps nécessaire pour normaliser un document est important, les utilisateurs se sont en général déclarés satisfaits des résultats obtenus.

Les utilisateurs experts se sont montrés plus aptes à utiliser la vue sémantique factorisée et à y identifier les sous-spécifications qui pouvaient accélérer l'analyse, ils ont donc délaissé le mode de résolution par dialogue de négociation. Ces utilisateurs se sont montrés critiques sur les erreurs faites par le système, mais ont néanmoins apprécié la propagation de résolutions par le jeu des dépendances sémantiques, notamment via les opérations d'invalidations.

Il ressort que la normalisation d'un document, même court, avec le système présenté requiert un temps parfois important, mais qu'en général les utilisateurs affirment avoir obtenu une version normalisée acceptable. Si ce système ne semble pas adapté pour des normalisations occasionnelles par des utilisateurs non initiés, il promet néanmoins de se montrer utile pour des volumes de documents importants ou lorsque les dérivations possibles du résultat (documents normalisés annotés, traductions en plusieurs langues, commandes pour un système, etc.) justifient le temps investi. Finalement, plusieurs voies d'amélioration du prototype ont été suggérées, notamment :

– surligner des éléments dans le texte pouvant permettre de faire un choix pour un type sémantique sous-spécifié, surtout lorsque l'utilisateur n'est pas l'auteur du document ;

– mettre en évidence dans les vues factorisée et MDA les éléments modifiés lors du dernier rafraîchissement, et éventuellement produire des messages en langue pour expliquer pourquoi certains éléments ont disparu ;

– apprendre dynamiquement et de façon supervisée de nouvelles formulations à ajouter à la grammaire pouvant améliorer les analyses futures.

## 4. Approches pour la création de documents normalisés

Dans cette dernière section, nous situons nos apports par rapport aux approches existantes pouvant être adaptées pour la création de documents normalisés. En particulier, nous discutons de l'utilisation des approches prépondérantes de création de documents par saisie du texte, puis nous décrivons le champ émergent de la création de documents par spécification du contenu dans lequel s'incrit notre approche.

### 4.1. *Création par saisie du texte*

La méthode traditionnelle de création de documents se fait par saisie du texte par un rédacteur. Des DTDs (Document Type Definitions) ou schémas XML permettent de décrire des éléments de contenu assimilables à des buts communicatifs, ainsi que, dans une certaine mesure, la structure du contenu de documents bien formés sous forme d'arbres XML. Il existe cependant des limitations concernant la description avec une DTD de dépendances sémantiques entre deux sous-arbres (Dymetman *et al.*, 2000) : permettre l'établissement d'une dépendance entre, par exemple, des formes pharmaceutiques et des modes d'administration nécessiterait la spécialisation des types d'éléments utilisés. La propagation de tels paramètres dans des sous-structures plus impor-

tantes et l'augmentation du nombre de paramètres rendraient cette approche difficile en pratique. Il revient donc implicitement aux rédacteurs de garantir une partie de la cohérence des documents.

Par ailleurs, il n'existe pas de mécanisme générique qui puisse certifier qu'un contenu linguistique est bien du type de la balise qui l'encadre (qui pourrait être sa description sémantique normalisée). Les langues contrôlées ont été proposées pour diminuer les ambiguïtés d'analyse, mais leur utilisation est souvent perçue par les rédacteurs techniques comme un exercice fastidieux et contraignant (Régnier et Dauphin, 2002) et les outils de vérification ne permettent pas toujours de propagation pour des corrections souvent difficiles à apporter. La consultation de mémoires de rédaction pourrait apporter une aide significative pour la rédaction de documents normalisés, et pourrait être combinée avec l'utilisation de mémoires de traduction, largement adoptées pour la traduction de documents techniques. Néanmoins, l'organisation de la chaîne rédactionnelle conserverait la séquentialité des tâches de rédaction puis de traduction pour la diffusion des modifications.

La création de documents normalisés par des méthodes de saisie du texte est donc possible, mais elle repose sur le respect de contraintes par le rédacteur pour lesquelles les outils actuels n'offrent pas toujours d'assistance. En outre, aucune aide n'est apportée pour la normalisation de documents existants.

### 4.2. *Création par spécification du contenu*

Des techniques de rendu automatique, qui laissent le rédacteur se concentrer sur la spécification du contenu (incluant ici la forme linguistique) en déléguant les détails de rendu visuel à la machine, sont aujourd'hui largement utilisées. Cela peut être transposé à la rédaction en séparant la spécification du contenu et la façon dont il est organisé et exprimé, tâches qui sont typiquement dissociées dans les systèmes de génération automatique (Reiter et Dale, 2000). Un paradigme récent de rédaction technique à visée multilingue (Hartley et Paris, 1995; Hartley et Paris, 1997) permet à un auteur technique monolingue d'interagir avec un système de génération automatique qui produit un brouillon de document en plusieurs langues. L'auteur apporte ses connaissances sur ce qui est l'objet d'un document, et le système apporte ses connaissances linguistiques ainsi que celles sur une classe de documents particulière. Cette déviation de la traduction vers la production de textes multilingues dérive du scénario de la *traduction automatique pour auteurs monolingues* initialement décrit par Kay en 1973 (Kay, 1997) qui a notamment été implémenté par les approches de *traduction automatique fondée sur le dialogue* (Boitet, 1989) et de *traduction sans texte source* (Somers *et al.*, 1990) dans lesquelles un auteur entretient un dialogue avec un système visant à expliciter les informations nécessaires pour produire le texte cible à partir d'un texte source rédigé. La plupart des systèmes de création par spécification du contenu développés par la suite sont des prototypes de recherche dont l'objectif principal est de permettre une production multilingue de documents par interaction avec un auteur. Ils pourraient être adaptés pour permettre, dans une certaine mesure, la création de

documents normalisés. La création de documents normalisés à partir d'un document existant, telle que proposée par notre approche de normalisation, requiert toutefois une interprétation du contenu du document.

Les systèmes de création symbolique de document (*symbolic authoring systems*) (Caldwell et Korelsky, 1994; Paris *et al.*, 1995; Power et Cavallotto, 1996; Sheremetyeva *et al.*, 1996; Isard *et al.*, 2003) permettent à un auteur de créer un document par édition d'une base de connaissances sans nécessiter de compétences particulières dans le formalisme de représentation. La gamme des concepts exprimables est restreinte, leur structuration sous forme de plan est généralement réalisée par le système, et l'expression linguistique de surface est produite en plusieurs langues par un générateur ou à partir de patrons. Le mode de spécification du contenu à l'aide de menus en cascade dans des formulaires apparaît assez éloigné d'une tâche de rédaction, et les liens entre d'une part les parties spécifiées et les noms symboliques utilisés et, d'autre part, le texte final ne sont pas toujours évidents pour l'utilisateur.

La génération suivante de systèmes utilise des *textes de contrôle* (*feedback texts*) qui autorisent la manipulation du contenu par le texte en évolution d'un document. Chaque opération de l'utilisateur sur le texte de contrôle met à jour la représentation du contenu sous-jacente, et le texte de contrôle est regénéré pour refléter les changements dans les langues supportées par le système. Ce que l'utilisateur lit dans le texte de contrôle correspond à la représentation du contenu qu'il a construite par le biais de choix sémantiques. Ce paradigme est ainsi appelé WYSIWYM (What You See Is What You Meant), et trouve son origine dans de nombreux projets de l'ITRI (Power et Scott, 1998; Power *et al.*, 1998).

Le paradigme WYSIWYM a été implémenté dans des prototypes couvrant de nombreux domaines, notamment : création de manuels d'utilisation de logiciels (DRAFTER-2 (Power *et al.*, 1998), DRAFTER-3 (van Deemter et Power, 1998)), spécifications de requêtes pour un système de régulation maritime (MILE (Piwek *et al.*, 2000)), création de notices de médicaments (ICONOCLAST (Bouayad-Agha *et al.*, 2000), PILLS (Bouayad-Agha *et al.*, 2002)), bulletins météorologiques (MultiMétéo (Coch, 1996)). ICONOCLAST s'est plus particulièrement concentré sur la séparation du contenu et du style dans les documents. Le projet PILLS définit quant à lui la notion de modèle de référence (*master model*) contenant l'ensemble des informations requises pour dériver plusieurs types de documents qui partagent des informations. MultiMétéo vise à laisser plus de contrôle à l'utilisateur sur la forme linguistique des textes produits en autorisant des nuances assez fines par le biais d'une édition WYSIWYM sans que les règles à l'origine de leur choix n'aient eu à être formalisées.[14] Les améliorations successives du paradigme ont notamment concerné l'établissement de coréférences (van Deemter et Power, 1998), la possibilité de créer des entités multiples (Piwek *et al.*, 2000), ou encore le support de structures complexes telles que l'implication, l'obligation ou la négation (Kibble *et al.*, 1999).

14. Les données utilisées pour produire les bulletins étant déjà spécifiées, il s'agit davantage de ce que ses auteurs appellent un système de génération multilingue interactive.

Une approche plus récente met l'accent sur la notion de bonne formation du contenu des documents. Si les formalismes grammaticaux sous-jacents sont plus complexes et permettent d'exprimer plus de contraintes sur le contenu des documents, les interfaces proposées partagent beaucoup de caractéristiques avec l'approche WYSIWYM. Les systèmes de type GF (Grammatical Framework) (Ranta, 2003) et MDA (Mutilingual Document Authoring) (Dymetman *et al.*, 2000) se basent sur un typage fort des objets sémantiques et sur la notion de type dépendant. Des arbres sémantiques abstraits typés, représentant le contenu de documents, sont utilisés par un processus de génération compositionnelle pour produire le texte des documents. Des grammaires parallèles décrivent des syntaxes concrètes différentes pour une même syntaxe abstraite, et permettent donc, par exemple, de générer des versions en plusieurs langues pour un même document. Typiquement, la granularité des représentations sémantiques se limite aux sous-constituants qui jouent un rôle au niveau de la variabilité communicationnelle dans une classe de documents particulière. Ainsi, la spécification d'un but communicatif peut être réalisée en un seul choix par l'utilisateur lorsque la spécification en sous-constituants serait contre-productive. Divers types d'applications ont été implémentés, notamment : édition de spécifications logicielles (Hähnle *et al.*, 2002), notices de médicaments, rapports de protocoles d'expériences biologiques (Brun *et al.*, 2003), et descriptifs de produits dangereux (Brun et Hagège, 2003).

Les aides au développement et à la validation des ressources utilisées par ces systèmes constituent une voie de recherche importante. Une intégration plus radicale entre la création de documents par le contenu et des bases de connaissances externes comme proposée par (Dymetman, 2002) permet d'enrichir les connaissances utilisées par le système avec l'information spécifiée par l'auteur lors de la création d'un document. Il est en effet difficile de garantir qu'une modélisation a priori d'un domaine ainsi que la couverture d'un générateur associé puissent être complètes. Des travaux récents (Paris *et al.*, 2005) visent à extraire automatiquement ce type de connaissances et à en permettre la manipulation via des éditeurs de modèles de domaines.

Les paradigmes de création de documents présentés sont radicalement différents des approches traditionnelles, et l'intégration effective dans des contextes de production de documents soulève plusieurs difficultés. Notamment, la création peut être laborieuse du fait de la manière itérative dont les documents sont construits : la comparaison informelle entre le temps nécessaire pour rédiger un texte par saisie directe et par un système WYSIWYM donne un rapport d'environ 4 à 6 (Power *et al.*, 2003). Un niveau de granularité adéquat pour les choix faits par l'auteur peut néanmoins accélérer considérablement le travail de création en plusieurs langues pour des classes de documents particulières. Le fait que l'auteur ait à identifier dans des menus possiblement grands la formulation correspondant à son choix, suggère un agencement motivé des choix dans les menus, par exemple par probabilité d'apparition dans un contexte local appris sur corpus (Nickerson, 2003). La relative complexification des formalismes pour permettre la spécification d'expressions telles que l'implication (Kibble *et al.*, 1999) semble indiquer que la granularité choisie peut limiter le répertoire d'expressions disponibles pour l'auteur et donc la complexité des textes produits. Finale-

ment, la formulation en langue n'est pas sous le contrôle direct de l'auteur des documents, et l'expressivité est bornée par ce que le système de génération utilisé peut produire, mais cela est souhaitable lorsqu'il s'agit de produire des documents normalisés. Toutefois, ces systèmes gagneraient parfois à autoriser la saisie de texte par l'utilisateur,[15] ce qui est essentiellement ce qu'essaie de faire notre approche, qui pourrait être utilisée pour créer des documents normalisés par normalisation du texte saisi par un auteur.

## 5. Conclusions et perspectives

Dans cet article, nous avons proposé une approche permettant de normaliser des documents appartenant à une classe définie qui réutilise le formalisme et les ressources d'un système de création contrôlée de documents, et n'impose pas le recours à une technique d'analyse fine. Cette approche se base sur une négociation interactive avec un expert du domaine pour résoudre les sous-spécifications de l'analyse au niveau des buts communicatifs qui seraient difficiles à résoudre automatiquement de façon fiable. En outre, l'analyse par génération inversée floue implémentée est robuste à la variation dans les textes, ce qui permet d'envisager un mode de création de documents normalisés par normalisation de documents pouvant aller de brouillons sténographiques à des documents plus ou moins rédigés.

L'implémentation proposée permet de garantir les propriétés souhaitables pour les documents normalisés qui ont été dégagées dans la section 2.1 : pertinence thématique, cohérence et complétude du contenu, et compréhensibilité et cohérence d'emploi de la langue. La construction des ressources nécessaires et leur évolution demeurent néanmoins des sujets de recherche, notamment pour permettre l'application des techniques présentées à des domaines moins contraints. La complexité de description des normes et le temps nécessaire pour normaliser des documents ne doivent cependant pas masquer les applications rendues possibles sur des documents normalisés dont l'analyse a été validée par un expert du domaine : la possibilité d'annoter sémantiquement les documents produits ainsi que de les dériver automatiquement en plusieurs langues et pour des lecteurs ayant des niveaux d'expertise différents offrent des perspectives en recherche d'information et en accès à l'information et participent de façon importante à la pérennité des documents.

## Remerciements

La majeure partie de ce travail a été effectuée à XRCE ainsi qu'au GETA-CLIPS à Grenoble. L'auteur remercie Marc Dymetman et Christian Boitet pour leur aide et leurs conseils.

15. Cela est permis dans une certaine mesure avec des systèmes utilisant GF, mais il faut alors que l'entrée puisse être analysée par les grammaires réversibles du système, ce qui n'offre pas une solution très générale.

## 6. Bibliographie

Agence du Médicament, *Notice et étiquetage des médicaments à usage humain : réglementation et recommandations*, Ministère du Travail et des Affaires Sociales, Paris, 1996.

Alamargot D., Terrier P., Cellier J.-M., *Production, compréhension et usages des écrits techniques au travail*, Octarès Éditions, 2005.

Blanchon H., « A Pattern-based Analyzer for French in the Context of Spoken Language Translation : First Prototype and Evaluation », *Actes de COLING-02, Taïwan*, 2002.

Boitet C., « Speech Synthesis and Dialogue Based Machine Translation », *Actes de ATR Symposium on Basic Research for Telephone Interpretation, Kyoto, Japon*, 1989.

Bonhomme P., « Codage et normalisation de ressources textuelles », *Ingénierie des Langues*, Hermès Science Publications, p. 173-191, 2000.

Bouayad-Agha N., Power R., Scott D., Belz A., « PILLS : Multilingual Generation of Medical Information Documents with Overlapping Content », *Actes de LREC-02, Las Palmas, Espagne*, 2002.

Bouayad-Agha N., Scott D., Power R., « Integrating Content and Style in Documents : a Case Study of Patient Information Leaflets », *Information Design Journal*, vol. 9, p. 161-176, 2000.

Brun C., Dymetman M., « Rédaction multilingue assistée dans le modèle MDA », *Multilinguisme et Traitement de l'Information*, Hermès Science Publications, p. 129-152, 2002.

Brun C., Dymetman M., Fanchon E., Lhomme S., « Controlled Authoring of Biological Experiment Reports », *Actes de EACL-03, Demonstrations and Research Notes, Budapest, Hongrie*, 2003.

Brun C., Hagège C., « Normalization and Paraphrasing using Symbolic Methods », *Actes de International Workshop on Paraphrasing (IWP2003), ACL'03, Sapporo, Japon*, 2003.

Caldwell D. E., Korelsky T., « Bilingual Generation of Job Descriptions from Quasi-Conceptual Forms », *Actes de ANLP-94, Stuttgart, Allemagne*, 1994.

Coch J., « Evaluating and Comparing Three Text-Production Techniques », *Actes de COLING-96, Copenhague, Danemark*, 1996.

Dymetman M., « Text Authoring, Knowledge Acquisition and Description Logics », *Actes de COLING-02, Taïwan*, 2002.

Dymetman M., Lux V., Ranta A., « XML and Multilingual Document Authoring : Convergent Trends », *Actes de COLING 2000, Saarbruck, Allemagne*, 2000.

Hartley A. F., Paris C. L., « Supporting Multilingual Document Production : Machine Translation or Multilingual Generation ? », *Actes de IJCAI-95 Workshop on Multiligual Text Generation, Montréal, Canada*, 1995.

Hartley A. F., Paris C. L., « Multilingual Document Production - From Support for Translating to Support for Authoring », *Machine Translation*, vol. 12, p. 109-128, 1997.

Hähnle R., Johannisson K., Ranta A., « An Authoring Tool for Informal and Formal Requirement Specifications », *Actes de ETAPSFASE-02*, 2002.

Isard A., Oberlander J., Androutsopoulos I., Matheson C., « Speaking the Users' Languages », *IEEE Intelligent Systems*, vol. 18, p. 40-45, 2003.

Jacquemin C., *Spotting and Discovering Terms through Natural Language Processing*, MIT Press, Boston, 2001.

Kay M., « The Proper Place of Men and Machines in Language Translation », *Machine Translation*, vol. 12, p. 3-23, 1997.

Kibble R., Power R., van Deemter K., « Editing Logically Complex Discourse Meanings », *Actes de International Workshop on Computational Semantics, Tilburg, Pays-Bas*, 1999.

Max A., De la création de documents normalisés à la normalisation de documents en domaine contraint, Thèse de doctorat, Université Joseph Fourier, Grenoble, 2003a.

Max A., « Multi-language Machine Translation through Interactive Document Normalization », *Actes de European Association for Machine Translation Workshop, EACL-03, Budapest, Hongrie*, 2003b.

Nickerson J., « Statistical Models for Organizing Semantic Options in Knowledge Editing Interfaces », *Actes de AAAI Spring Symposium Series Workshop on Natural Language Generation in Spoken and Written Dialogue*, 2003.

Nilsson N. J., *Artificial Intelligence : a New Synthesis*, Morgan Kaufmann, 1998.

Paiva D. S., « Investing Style in a Corpus of Pharmaceutical Leaflets : Result of a Factor Analysis », *Actes de Student Research Workshop de l'ACL-2000, Hong Kong*, 2000.

Paris C., Colineau N., Lu S., vander Linden K., « Automatically Generating Effective Online Help », *International Journal on E-Learning*, vol. 4, p. 83-103, 2005.

Paris C., Linden K. V., Fischer M., Hartley A., Pemberton L., Power R., Scott D., « A Support Tool for Writing Multilingual Instructions », *Actes de IJCAI-95, Montréal, Canada*, 1995.

Piwek P., Evans R., Cahill L., Tipper N., « Natural Language Generation in the MILE System », *Actes de Workshop IMPACTS in NLG, Schloss Dagstuhl, Allemagne*, 2000.

Power R., Cavallotto N., « Multilingual Generation of Administrative Forms », *Actes de IWNLG-96, Herstmonceux Castle, Royaume-Uni*, 1996.

Power R., Scott D., « Multilingual Authoring using Feedback Texts », *Actes de COLING/ACL-98, Montréal, Canada*, 1998.

Power R., Scott D., Evans R., « What You See Is What You Meant : Direct Knowledge Editing with Natural Language Feedback », *Actes de ECAI-98, Brighton, Royaume-Uni*, 1998.

Power R., Scott D., Hartley A., « Multilingual Generation of Controlled Languages », *Actes de EAMT/CLAW-03, Dublin, Irlande*, 2003.

Ranta A., « Grammatical Framework : a Type-Theoretical Grammar Formalism », *The Journal of Functional Programming*, 2003.

Régnier S., Dauphin E., « Aide à la production de documentation technique multilingue », *Multilinguisme et traitement de l'information*, Hermès Science, p. 153-180, 2002.

Reiter E., Dale R., *Building Natural Language Generation Systems*, Cambridge University Press, 2000.

Sheremetyeva S., Nirenburg S., Nirenburg I., « Generating Patent Claims from Interactive Input », *Actes de IWNLG, Herstmonceux, Royaume-Uni*, 1996.

Somers H., Tsujii J.-I., Jones D., « Machine Translation without a Source Text », *Actes de COLING-90, Helsinki, Finlande*, vol. 3, p. 217-276, 1990.

van Deemter K., Power R., « Coreference in Knowledge Editing », *Actes de Workshop on the Computational Treatment of Nominals de COLING-ACL-98, Montréal, Canada*, 1998.

Vidal, *Le dictionnaire*, Vidal, Paris, 2006.