Thèse Catherine DOMINGUES 2001

Etude d'outils informatiques et linguistiques pour l'aide à la recherche automatique d'information dans un corpus documentaire.

THESE

Etude d'outils informatiques et linguistiques pour l'aide à la recherche automatique d'information dans un corpus documentaire

vol. I

Catherine DOMINGUES

IGM 2001 – 04

UNIVERSITE DE MARNE-LA-VALLEE

2001

Thèse pour obtenir le grade de
DOCTEUR DE L'UNIVERSITE DE MARNE-LA-VALLEE
discipline : informatique
présentée et soutenue publiquement par
Catherine DOMINGUES
le 16 mars 2001

Etude d'outils informatiques et linguistiques pour l'aide à la recherche automatique d'information dans un corpus documentaire

directeur de thèse : Maurice GROSS (LADL)

jury : Bruno BACHIMONT (INA)
Marc BOURDEAU (CSTB)
Pierre LAFON (ENS)
Eric LAPORTE (UMLV)
Max SILBERZTEIN (GRELIS et IBM)

Contexte de l'étude, objectifs, plan du travail

Cette thèse a été réalisée sous la direction de Maurice GROSS, dans le cadre d'un accord entre le ministère de l'équipement, des transports et du logement (METL) et en particulier de sa direction de la recherche et celle du personnel et des services, le centre scientifique et technique du bâtiment (CSTB) et l'université de Marne la Vallée. La structure d'accueil a été double : l'Institut Gaspard Monge à Marne la Vallée, et le Laboratoire d'automatique documentaire et linguistique (LADL) de Paris 7.

Le CSTB est un établissement public à caractère industriel et commercial, sous tutelle du METL. Il emploie environ 600 personnes réparties sur plusieurs sites : Paris et Champs sur Marne, Grenoble, Nantes et Sophia-Antipolis. Son organigramme témoigne de la diversité des thèmes de travail qui sont abordés en relation avec le bâtiment et la construction :

- physique du bâtiment et confort : acoustique ; sécurité feu ; aérodynamique et environnement ; énergie, environnement intérieur et automatismes ;
- économie et sciences sociales ;
- matériaux et techniques de construction : structures ; matériaux ; enveloppes légères et transferts ; technologie des revêtements et toitures ; enduits, mortiers et colles ;
- équipements : hydraulique et équipements sanitaires ; eaux, air et environnement ;
- industries de l'information : informatique et bâtiment.

Un des métiers du CSTB est de diffuser des connaissances, et en particulier mettre à la disposition des professionnels tous les documents réglementaires et techniques de la construction (lois, arrêtés, notes techniques, ...) nécessaires à la conception, au suivi et à la réalisation des bâtiments et des ouvrages. Pour cela, le CSTB publie depuis 1946 le *Reef*, une encyclopédie qui rassemble sous forme papier :

- les textes officiels : textes législatifs et réglementaires ;
- les règles d'exécution : les documents techniques unifiés (DTU) et les normes-DTU, les documents généraux d'avis techniques ;
- les principales normes applicables au Bâtiment ;
- des règles de calcul ;
- des exemples de solutions et des solutions techniques.

L'encyclopédie complète totalise 25 volumes, et près de 25 000 pages contenant des textes, des graphiques et des tableaux. Elle fait l'objet de quatre mises à jour annuelles, qui concernent l'apparition de nouveaux textes ainsi que la modification de ceux existants. Ces mises à jour occupent une documentaliste environ cinquante jours par an.

Le public concerné par le *Reef* est composé de tous les professionnels du bâtiment : les maîtres d'œuvre, les maîtres d'ouvrage, les bureaux d'études et de vérifications, les organismes de contrôle, les entrepreneurs et artisans des différents corps de métiers du bâtiment, ...

Le *Reef* qui intéresse potentiellement tous ces acteurs du bâtiment est pourtant distribué d'une manière très inégale. Afin d'en faciliter la diffusion et l'utilisation, depuis 1991 le CSTB édite ce document sous forme de cédérom, le *CD-Reef*. La recherche d'information peut s'y développer de cinq façons différentes :

- par contexte : il y a quatre critères de sélection : le type d'ouvrage, les exigences, les contraintes administratives, la destination (i.e. le type de bâtiment : habitation, bureaux, ...) qu'il est possible de croiser. Pour ce type de recherche, la totalité du corpus a été indexée manuellement ;
- par produits et matériaux : c'est une recherche équivalente à la précédente mais qui passe par un choix technologique ;
- par catalogue : les documents sont accessibles par familles (textes législatifs, réglementaires, documents techniques, ...), puis par leur référence ;
- par mots-clés contenus dans les titres des documents. environ 1438 termes (mots simples et composés) sont indexés. Il est possible d'utiliser l'opérateur booléen *et* pour écrire la requête de recherche ;
- par construction (puis consultation) d'un index associé à un ensemble de documents sélectionnés par l'utilisateur et qui regroupe les termes significatifs identifiés automatiquement dans les documents. Ces termes sont reconnus par consultation d'un dictionnaire général de la langue. Cette opération est réalisée à l'aide la boîte à outils SYLEX[1].

Ainsi, hormis pour une recherche de type pattern-matching proposée à l'intérieur du document en cours de visualisation, l'utilisateur ne peut saisir le texte qu'il veut rechercher. Les seuls mots utilisables pour la recherche d'information sont des mots-clés. Ceux-ci ont été ou bien sélectionnés manuellement par des documentalistes (dans le texte intégral ou seulement les titres selon le type d'utilisation envisagée) à partir d'une liste spécifique aux métiers du bâtiment, ou bien identifiés à partir d'un dictionnaire de base du français.

L'objectif de cette thèse est donc de conforter des compétences en ingénierie linguistique, déjà opérationnelles au CSTB, en y apportant une réflexion plus théorique grâce à l'expérience et aux compétences du directeur de thèse et du LADL. En particulier, une des retombées attendues est d'améliorer l'accès à l'information contenue dans le *Cd-Reef* afin que la réglementation soit mieux connue et mieux appliquée. Dans le cadre de ce travail, il a été décidé de circonscrire le corpus aux textes concernant la sécurité incendie des établissements recevant du public (ERP). En effet, cette réglementation a la réputation d'être peu lisible et difficile ; elle occupe un volume important, contient des textes qui admettent des prolongements tentaculaires, couvre différents domaines d'application, fait intervenir plusieurs ministères (METL et ministère de l'Intérieur). En outre, elle concerne un domaine sensible : la sécurité des personnes. Elle constitue donc un champ d'application intéressant pour y mettre au point et tester de nouveaux outils.
L'étude de cette réglementation nécessite aussi deux types de compétences. D'un point de vue linguistique, ce travail est fondé sur les travaux, méthodes et outils développés au LADL. Le système de dictionnaires électroniques et les programmes d'analyse lexicale permettant le traitement linguistique du corpus sont ceux fournis par le logiciel INTEX (M. Silberztein, 1993).
Et, dans le prolongement du travail linguistique, l'intervention d'un expert du domaine (qui recouvre aussi des compétences variées : sécurité incendie et phénomènes physiques liés au feu) permet de confronter les constructions linguistiques aux coutumes langagières et syntaxiques observées chez les différents intervenants : auteurs des textes réglementaires, auteurs des textes techniques, maîtres d'œuvre, maîtres d'ouvrages, artisans, ... En particulier, pour construire les dictionnaires de métier à partir des expressions candidates fournies par les outils linguistiques, nous avons bénéficié de l'aide de Michel CURTAT, responsable de la partie "consultance" dans le service "Sécurité feu" du CSTB.

---

1 SYLEX a été développé par INGENIA Langage Naturel, et est maintenant distribué par SYSTAL.

Dans le **premier chapitre** de notre travail, nous présenterons brièvement **l'évolution des systèmes de recherche d'information et d'extraction de texte**, avant de préciser **les bases de travail** que le **LADL** s'est choisies. Ensuite, nous examinerons en détail une étude réalisée en 1999 par le CSTB et dont l'objectif était d'évaluer l'intérêt d'utiliser des outils linguistiques pour des recherches automatiques d'informations dans un corpus.

Améliorer l'accès à la réglementation peut se concrétiser sur différents aspects :

a) Le nombre d'interlocuteurs concernés par la sécurité incendie est important, et leurs origines diverses. En particulier, les métiers du bâtiment ont un vocabulaire très riche dans lequel on peut trouver, pour désigner le même objet ou le même concept, des expressions qui varient selon le corps de métier. Dans le **deuxième chapitre** nous présenterons une **étude du corpus** destinée à identifier le vocabulaire du domaine, et à l'organiser sous forme de listes de mots simples et de mots composés.

b) Le **troisième chapitre** regroupe différents thèmes autour de **l'utilisation d'automates pour la réalisation de grammaires locales.**
Nous présentons d'abord une étude comparative de quatre noms : *appareil*, *dispositif*, *installation* et *système*, considérés comme des classifieurs pour la construction de mots composés destinés à désigner des objets concrets. Ceux-ci réalisent des fonctions spécifiques aux métiers :

appareil (à gaz + de commande)
dispositif (de commande + de fixation)
installation au gaz
système de fixation

Les classifieurs paraissent plus ou moins interchangeables dans la formation des noms composés. Cette intuition doit être vérifiée avant d'étudier ses implications dans la construction des dictionnaires de mots composés.

Ensuite, afin de préciser le vocabulaire de la sécurité incendie, on s'est intéressé aux deux noms *type* et *catégorie* qui, dans le corpus, cumulent des définitions du vocabulaire courant avec celles liées à la spécialité (ce sont les deux caractéristiques d'un établissement qui conditionnent l'application de la réglementation : le type d'un établissement dépend de l'activité qu'il abrite, et la catégorie de l'effectif du public autorisé). Leurs emplois sont décrits afin de désambiguïser les groupes nominaux (*GN*) dans lesquels ils sont impliqués.

La dernière partie du chapitre montre les différentes formulations du type et de la catégorie d'un établissement :

un établissement de type R *de 1ère catégorie*
un collège *acceptant plus de 1500 élèves*

L'objectif est de recenser ces formulations équivalentes et de proposer, sous forme de transducteurs, les transformations qui permettent d'associer à un établissement son type et sa catégorie, chaque fois que l'information est disponible.

L'étude destinée à évaluer l'intérêt d'utiliser des outils linguistiques pour améliorer la recherche d'information dans un corpus a montré que la sélection des documents répondant à une question se fait par rapprochement entre les mots de la requête et ceux du texte. Les causes de silence sont multiples ; nous en avons choisi deux sur lesquelles nous nous sommes proposée de travailler :

c) La première concerne l'emploi d'une coordination dans les *GN* :

dans les établissements et dans les locaux présentant des risques particuliers d'incendie

En effet, si l'on interroge la documentation avec l'expression *établissements présentant des risques particuliers d'incendie*, le paragraphe contenant le *GN* complet ne sera pas jugé pertinent car le nom tête *établissement* est séparé de son modifieur *présentant des risques particuliers d'incendie*. Le but du travail présenté dans le **quatrième chapitre**, est de mettre au point des règles permettant de reconstruire les *GN* complets (en l'occurrence *dans les établissements présentant des risques particuliers d'incendie* et *dans les locaux présentant des risques particuliers d'incendie*) afin de réduire le silence dû à la **coordination des modifieurs droits dans les *GN***.

d) La deuxième est illustrée par la paire d'exemples suivante :

la stabilité au feu de la porte est de degré 1h
la porte est stable au feu 1h

L'expression d'un concept élémentaire peut se réaliser grâce à différentes formulations. En l'occurrence, il s'agit de la stabilité au feu, une caractéristique qui est définie pour bon nombre d'éléments de bâtiment et qui se traduit par une quantité exprimée en unités de temps. Elle s'exprime par une phrase simple à laquelle on peut appliquer des **transformations syntaxiques** classiques dans le domaine des grammaires transformationnelles. A ces transformations s'ajoutent les constructions propres au vocabulaire et à la syntaxe du domaine technique étudié. L'objectif est d'identifier toutes ces variations et de les ramener à une expression canonique, plus facile à traiter et à comparer. Ces variations sont étudiées dans le **cinquième chapitre**.

Enfin, nous présentons des conclusions en confrontant les résultats obtenus avec les différents outils mis au point, aux intentions initiales ; puis nous resituons notre travail par rapport aux tendances observées dans les systèmes que nous avons étudiés dans le chapitre *Etat de l'art*. Nous concluons enfin sur les prolongements de ce travail.

La problématique de ce travail pourrait être définie de la manière suivante : constituer des ressources linguistiques propres au domaine de la sécurité incendie (dictionnaires de mots simples et mots composés, grammaires locales, ...) afin de faciliter la recherche d'information à l'intérieur d'un corpus spécifique à ce domaine.
Dans ce contexte, nous allons présenter des travaux réalisés dans deux domaines spécifiques liés à notre problématique[1] : la recherche documentaire et la mise en évidence de groupes nominaux complexes, et leur évolution à travers les expériences déjà menées.

1. Recherche automatique d'information à l'intérieur d'un corpus

Pour un observateur humain, la recherche automatique d'information dans un corpus est subordonnée au problème plus général que constitue la compréhension d'un texte. Dans ce domaine, différentes techniques se sont succédées, apportant chacune ses caractéristiques propres. Nous allons donner un aperçu de ces méthodes.

Pour construire des systèmes informatiques réalisant automatiquement cette recherche d'information, la première difficulté, à laquelle nous ne donnons pas de réponse, consiste à préciser ce que l'on appelle compréhension d'un texte pour un système automatique : on se contentera de définir l'analyse automatique d'un texte comme la transformation d'une suite de caractères organisés linéairement en une représentation logique et conceptuelle (qui est elle aussi à définir) qui permette ensuite d'effectuer certaines opérations comme rechercher des informations, compléter une base de données, écrire un résumé ...

Cette transformation peut s'appuyer sur une analyse fouillée et globale du texte (comme l'envisageaient les systèmes du début des années 80) ou sur des analyses locales qui sont mises en place pour répondre à des besoins particuliers (ce qui caractérise les outils plus récents des années 90). On présentera ensuite un troisième point de vue qui procède, en même temps, des deux premiers.

1.1. Analyse fouillée et globale du texte

Les systèmes de compréhension de texte élaborés dans la deuxième moitié des années 80 ont pour objectif d'établir une analyse de la totalité du texte, à l'aide de ressources lexicales et linguistiques qui soient les plus larges possible. Les techniques employées doivent emprunter aux méthodes de la

[1] Cet état de l'art emprunte sa présentation chronologique et des références à l'article écrit par Thierry POIBEAU et Adeline NAZARENKO (1999) et à celui de Benoît HABERT et Christian JACQUEMIN (1993).

linguistique mais aussi de l'intelligence artificielle pour reconstituer des informations plus ou moins explicites, réaliser des inférences, ...

Les analyses morphologique et syntaxique ne manquent pas de laisser des ambiguïtés qui sont conservées dans les étapes ultérieures, ce qui accroît la complexité du traitement. Et sur cette analyse linguistique, se greffent des outils sémantiques qui ont pour objet, à travers un formalisme qui a varié selon les expériences, une représentation sémantique profonde du texte dans sa totalité.

On peut citer différents projets qui mêlent donc linguistique et intelligence artificielle : KALIPSOS (1984) au Centre scientifique d'IBM France, ACORD (1991) dans le cadre d'un projet ESPRIT qui regroupait des équipes universitaires et industrielles dans différents pays européens, TACITUS, ...

TACITUS appartient à ce type d'outils. Il est développé à la fin des années 80 par une équipe du Stanford Research Institute (SRI). Ces concepteurs lui reconnaissent des qualités en termes de compréhension de texte, mais déplorent sa lenteur d'exécution qu'ils attribuent au temps passé pour analyser des portions entières de texte qui n'ont, finalement, que peu de rapport avec la tâche définie initialement. Ces outils génériques ont finalement montré peu de succès dans des applications en vraie grandeur pour différentes raisons :

- bien qu'ils aient été conçus comme génériques, les spécificités d'un corpus de spécialité sont assez importantes pour nécessiter des adaptations conséquentes : les dictionnaires spécifiques et surtout le lexique sémantique, mais aussi les règles syntaxiques qui doivent être modifiées pour tenir compte des particularités d'un langage propre à un domaine ;

- "comprendre" un texte est une tâche qui, outre des connaissances linguistiques : phonétiques, morphologiques, lexicales, syntaxiques, ... exigent des informations souvent implicites, ou du moins difficiles à définir et encore plus à formaliser. Définir et constituer l'ensemble des connaissances nécessaires pour réaliser un système automatique de compréhension de texte s'avère une tâche difficile. Et, de plus, les concepteurs de ces systèmes génériques ont constaté que tenter de rendre compte, d'un point de vue sémantique, de la totalité d'un texte ne contribuait pas nécessairement à améliorer le filtrage de documents ou le repérage d'informations ciblées.

Finalement, les auteurs de TACITUS vont abandonner son développement, en particulier parce que ses temps d'exécution sont trop mauvais pour pouvoir être présentés à une conférence portant sur l'évaluation des systèmes destinés à la compréhension de texte, mais vont utiliser leur expérience pour mettre au point un autre outil, plus modeste FASTUS (cf. paragraphe suivant). Et en effet, ces projets de grande ampleur, malgré leurs performances médiocres, ont permis d'explorer une approche de la compréhension des textes dans laquelle des connaissances linguistiques et sémantiques doivent collaborer pour rendre compte de la totalité des informations contenues dans un texte. Mais leurs difficultés ont aussi montré que la démarche employée était peut-être à revoir : construire des systèmes génériques qui réalisent une analyse profonde et globale du texte.

1.2. Analyse locale

A la suite des travaux précédents, un objectif plus modeste a contribué à définir la compréhension de texte comme un processus qui consiste à extraire des informations du texte pour réaliser une tâche définie à l'avance : par exemple, compléter une base de données ou remplir un formulaire.

Cette approche a fait l'objet d'une série de conférences d'évaluation (conférences MUC pour Message Understanding Conferences) organisées de 1987 à 1998. Des applications en vraie grandeur y ont été présentées dans lesquelles le texte se présente comme un message court et factuel, et l'objectif est de remplir un formulaire déjà pré-établi. L'extraction se fait à partir de patrons syntaxiques (formalisés à l'aide d'automates). Une partie des automates sont construits à partir de connaissances générales sur la langue et sont utilisés pour toutes les applications ; mais pour tenir compte du vocabulaire, des expressions et des constructions spécifiques d'un domaine d'activité, on a recours à un corpus d'entraînement qui permet de mettre au point des automates spécifiques (et donc non réutilisables).

FASTUS fait partie des outils qui ont été présentés dans ces conférences MUC. C'est un outil d'extraction d'information, développé aussi par le SRI à partir de 1991-1992. Il a ensuite évolué et participé aux campagnes d'évaluation pendant toute la série des conférences. Ses concepteurs insistent sur la distinction entre systèmes d'extraction d'information pour lesquels :

- seulement une fraction du texte est pertinente ;
- l'information est reformulée dans une représentation prédéfinie, ciblée par rapport à la tâche à effectuer, et relativement simple ;
- les nuances incluses dans les propos de l'auteur ne sont pas d'un intérêt essentiel pour ma mise à jour de la base de données ;

et systèmes de compréhension de textes qui présentent les caractéristiques suivantes :

- l'objectif consiste à rendre compte du texte tout entier ;
- la représentation doit pouvoir traduire toutes les complexités de la pensée et les objectifs du rédacteur.

FASTUS, comme la majorité des outils évalués lors des MUC, utilisent des techniques robustes et déjà largement éprouvées : automates et pattern-matching, ce qui lui confère des avantages non négligeables pour ses auteurs il est conceptuellement simple, (d'après les comparaisons avec les autres outils présentés aux MUC) efficace et rapide, et sa rapidité d'exécution le rend aussi plus rapide à développer. Les automates sont plus ou moins nombreux et complexes, regroupés ou enchaînés en cascades. Mais là aussi il est nécessaire de les adapter lorsqu'on change de corpus.

Les performances présentées pour l'ensemble des systèmes testés lors de la 5ème conférence MUC (1993) sont moyennes : aux alentours de 57% de rappel et 64% de précision (qui peuvent s'élever à 74% quand il s'agit des informations principales contenues dans le texte).

Historiquement, ces deux approches que l'on pourraient qualifier d'opposées ont donc été expérimentées sans donner des résultats totalement satisfaisants et se heurtant finalement au même problème : le coût de mise au point des ressources lexicales, syntaxiques, ... nécessaires au fonctionnement des systèmes. Pour rentabiliser ces investissements, il devient nécessaire de concevoir des outils réutilisables ou du moins adaptables à de nouvelles applications, dans des conditions de temps et d'expertise que l'on pourra juger raisonnables. Une nouvelle conception de ce type de techniques essentiellement fondée sur la réutilisabilité est donc étudiée. Ces nouveaux outils doivent permettre de formaliser et de réutiliser à la fois des connaissances linguistiques indépendantes des applications envisagées, et des informations spécifiques propres à un domaine d'étude.

1.3. Construction d'outils réutilisables

Les systèmes d'analyse de textes ne peuvent faire l'économie de connaissances morphologiques, lexicales et syntaxiques, générales même si l'ambition de ce type d'outils n'est pas de rendre compte de la totalité d'un texte. D'autre part, les informations spécifiques à un domaine sont indispensables chaque fois que l'on veut construire une application en vraie grandeur, mais aussi longues et coûteuses à développer. Une solution pour néanmoins continuer à construire des systèmes destinés à l'extraction automatique d'information consiste donc d'une part à définir et construire l'ensemble des "informations génériques" nécessaires pour toute application dans ce domaine, et d'autre part de privilégier des techniques d'acquisition de connaissances qui permettent des adaptations semi-automatiques des systèmes existants.

Les réseaux neuronaux sont des outils déjà expérimentés dans d'autres domaines pour leur capacité d'apprentissage et d'adaptation. Ils ont aussi permis d'obtenir de bons résultats en filtrage d'informations. Par exemple, dans les travaux de M. Stricker et al. (2000) des dépêches sont filtrées, en temps réel, selon leur pertinence par rapport à des sujets donnés, en utilisant à la fois des réseaux neuronaux et des méthodes probabilistes. La méthode repose sur la construction d'une base d'apprentissage et la sélection de descripteurs pertinents pour représenter le thème que l'on doit illustrer. Un moteur de recherche est utilisé pour choisir des documents pertinents, et un premier réseau neuronal sommaire permet d'estimer la probabilité pour qu'il y ait eu une erreur sur ce choix, on procède itérativement jusqu'à construire complètement la base d'apprentissage. L'autre étape consiste à construire la liste la plus courte possible de descripteurs du thème : la liste initiale, composée de tous les mots du texte, est d'abord réduite sur des considérations statistiques, puis elle est classée, et de nouveau réduite, en utilisant itérativement la méthode d'orthogonalistion de Gram-Schimdt, chaque texte étant représenté par le vecteur de ses descripteurs ; enfin, le critère d'arrêt, est élaboré à partir d'un vecteur aléatoire des descripteurs qui est classé par la même méthode. La fonction d'activation des neurones et l'architecture sont choisies de manière à optimiser conjointement le rappel et la précision sur une base de validation.
Les résultats obtenus (95% de rappel, 88% de précision dans une des expériences rapportées) sont discutés selon différents critères (structure du réseau, taille de la base d'apprentissage, lemmatisation des mots).

Les travaux de J. Sénellart (1998) illustrent l'intérêt d'utiliser des automates pour extraire des informations. L'objectif est de construire et mettre à jour semi-automatiquement (avec vérification manuelle) une base de donnés contenant des noms propres associés à des professions. Les séquences pertinentes par rapport à ce type d'information sont décrites à l'aide de transducteurs et par consultation de dictionnaires. L'utilisation de transducteurs, qui met en œuvre des reformulations, permet de répondre à des requêtes, par des séquences dans lesquelles ne figure aucun mot de la question.

2. Mise en évidence de groupes nominaux complexes

Les approches mises en œuvre pour mettre en évidence, délimiter, désambiguïser, identifier des groupes nominaux complexes sont multiples. On peut dissocier :

- les techniques structurelles fondées sur l'utilisation de grammaire formelles. Elles nécessitent la mise en œuvre de connaissances préalables importantes (qui s'expriment le plus souvent par la constitution de grammaires et de lexiques), mais demandent peu de

travail de vérification a posteriori. On détaillera cette approche à travers le travail effectué au LADL (§ 3) ;

- les méthodes numériques ou connexionnistes (ou mixtes : numériques et connexionnistes). Elles imposent peu de connaissances linguistiques avant leur mise en œuvre mais exigent des vérifications post traitement importantes. Les résultats obtenus, puisqu'ils ne s'appuient sur aucune analyse structurée préalable, mélangent différents niveaux d'analyse. Les méthodes permettent aussi, quand elles s'appliquent à des données textuelles volumineuses, de faire ressortir des caractéristiques difficiles à saisir pour un observateur humain.

2.1. L'information mutuelle et d'autres méthodes statistiques

Une méthode statistique qui met en évidence des dépendances entre termes est fondée sur le calcul de l'information mutuelle. Celle-ci se calcule de manière suivante : dans une fenêtre de texte de longueur donnée, on associe à tous couples de mots une quantité d'autant plus élevée qu'ils sont présents, en même temps, plus souvent que ne le prévoit une distribution aléatoire.

La formule n'est pas symétrique car l'ordre des mots est aussi significatif.
Des fenêtre étroites (de taille un à cinq mots) permettent d'identifier des associations lexicales (qui sont candidates pour devenir des mots composés) ou des morceaux de mots composés. Des fenêtres plus larges (de cinq à mille mots) mettent en lumière des affinités sémantiques qui ne se traduisent pas nécessairement pas la formation de noms composés.
Les calculs effectués sont rapides, mais les résultats ne peuvent se passer d'un filtrage effectué par un observateur humain avant d'être utilisables.

Des variantes sont possibles selon les applications :

- variation sur la largeur de la fenêtre qui, au lieu d'être fixée numériquement, peut se raccrocher à un découpage logique du texte : la phrase ou le paragraphe ;

- prise en compte, ou non, de l'ordre d'apparition des mots dans la fenêtre ;

- variation sur la méthode de calcul des probabilités des paires étudiées.

Dans tous les cas, ces méthodes ne protègent pas contre des associations hétérogènes de mots puisqu'elles ne s'appuient pas sur des connaissances morphologiques ou syntaxiques de la langue. Les résultats présentés permettent, par exemple, la sélection concomitante des paires : *avant garde* et *est un*.

Une autre variante est fondée sur le fait de tenir compte ou pas des mots grammaticaux dans la segmentation des mots composés. Par exemple, si on ne tient pas compte des prépositions, on identifiera les expressions *verre de vin* et *verre à vin* ce qui, en termes de mise en évidence de mots composés, n'est pas un très bon choix mais permet de rendre compte d'un lien d'hyperonymie entre *verre* et *verre à lait*, et aussi de la composition sémantique entre *verre à vin* et *verre*.

Dans sa thèse, B. Daille (1994) montre l'utilisation conjointe de connaissances lexicales et syntaxiques avec des méthodes statistiques. Son objectif est de définir une méthode qui lui permette de mettre en évidence, dans un corpus technique, des mots composés. Pour cela, elle relève des expressions candidates à l'aide de patrons syntaxiques (qui opèrent donc sur un texte préalablement

étiqueté) puis utilise différentes mesures statistiques pour calculer la quantité caractéristique de chaque expression. Ce calcul permet alors de classer les expressions, et à chaque mesure est associée une liste ordonnée d'expressions candidates. La comparaison de ces listes avec un dictionnaire déjà construit de mots composés spécifiques du domaine permet de choisir les méthodes statistiques qui paraissent les plus performantes dans la sélection de mots composés attestés par le dictionnaire (B. Daille conclut que ce sont la fréquence et le coefficient de vraisemblance).

2.2. Les méthodes connexionnistes

Les réseaux neuronaux sont aussi utilisés pour le traitement de mots composés.

Par exemple, pour désambiguïser des syntagmes nominaux complexes contenant plusieurs modifieurs *: N Mod1 Mod2 Mod3*[2] où le modifieur peut s'écrire *Prép N*, l'objectif réside alors dans le rattachement de chaque modifieur avec le nom dont il dépend.

Une autre utilisation citée dans la littérature concerne l'identification de mots composés qui puissent admettre des variantes lexicales, des insertions, etc. Dans ce cas, le but est de reconnaître des expressions différentes comme, par exemple, *dispositif de fermeture automatique* et *système de fermeture automatique* et de pouvoir les associer à la même forme, par exemple *dispositif de fermeture automatique.*

3. L'approche du LADL

3.1. Le lexique-grammaire

Les travaux du LADL s'inscrivent dans une culture linguistique, et reposent sur le refus de la distinction entre lexique et grammaire parce que comme le présente A. Guillet (1991), cette séparation n'est pas opératoire : toutes les règles syntaxiques sont lexicalement contraintes. Il prend l'exemple de la règle de passivation qui s'applique à tout verbe transitif :

Max mange une pomme
La pomme est mangée par Max

Apparaît immédiatement une liste d'exceptions (une centaine, par rapport aux 6 000 verbes recensés) :

Cette pièce pue le tabac
**Le tabac est pué par cette pièce*

Puis il faut rajouter une contrainte supplémentaire pour pouvoir rendre compte des exceptions suivantes :

(1) *Des milliers de gens regardent la TV*

---

[2] On utilise les notations suivantes : *N* désigne un nom, *Mod* un modifieur, *Prép* une préposition.

La TV est regardée par des milliers de gens

(2) *Cette histoire regarde Max*
**Max est regardé par cette histoire*

Et si l'on veut étendre ce type de comportement à tous les verbes qui acceptent un sens propre et un sens figuré, comme *regarder*, on peut trouver d'autres exemples qui contredisent l'exception de la règle :

Max a dérangé les papiers
Les papiers sont dérangés par Max

Cette histoire dérange Max
Max est dérangé par cette histoire

Déranger admet deux sens, propre et figuré, et la passivation dans les deux cas, contrairement à *regarder.*

Cette série d'exceptions fonde finalement une règle : une règle syntaxique valide indépendamment du lexique n'existe pas, et les exceptions à ces règles de grammaire peuvent concerner non pas des mots, mais certains emplois de ces mots (comme on l'a constaté dans les exemples précédents). Il faut donc tenir compte de cette notion de sens ou d'emploi qui n'appartient pas du tout à la syntaxe mais au lexique. Le LADL se propose alors de créer un nouvel outil, le lexique-grammaire. Dans celui-ci sont consignées les propriétés qui concernent toutes les parties du discours : verbes, adjectifs, noms, adverbes, ... Le format choisi est celui d'une matrice : chaque ligne est constituée par une phrase élémentaire à laquelle sont associés des emplois possibles (qui pourraient traduire "les sens" de la lexicographie).

Pour décrire les emplois de *regarder*, nous avons donc les informations suivantes :

(1) *regarder* *sujet humain*
objet non humain
passif accepté

(2) *regarder* *sujet non humain*
objet humain
passif interdit

Ce qui se traduit de la manière suivante dans le lexique grammaire :

Sujet humain	Sujet non humain		Objet humain	Objet non humain	Passif
+	-	*regarder* (1)	-	+	+
-	+	*regarder* (2)	+	-	-

3.2. La notion de transformation

Les travaux du LADL se fondent aussi sur la notion de transformation telle qu'elle est présentée par Z.S. Harris (1952). M. Gross (1990) considère cette utilisation comme *"un progrès logique : des définitions rigoureuses et minimales, autrement dit, une limitation à la seule description combinatoire*

de la langue, permettent de construire des grammaires cohérentes qui rendent compte de l'aspect mécanique du comportement syntaxique des locuteurs et de le simuler sur ordinateur".

Dans ce cadre théorique, une partie très importante des travaux du LADL est consacrée à l'étude des noms, et en particulier aux noms composés. En effet, ceux-ci sont très nombreux (ils couvrent, selon les auteurs, un cinquième de la surface d'un texte), présentent des caractéristiques particulières et, dans le cadre d'une recherche d'information, sont moins ambigus et plus pertinents que des noms simples. La définition d'un nom composé varie sensiblement selon les écoles, mais on observe un consensus sur le fait de considérer la composition comme une échelle de figement. Un des objectifs du LADL est de construire des dictionnaires de noms composés à large couverture (plusieurs centaines de milliers d'entrées déjà recensées) et de créer des outils de reconnaissance de ces expressions (en particulier à l'aide d'automates à états finis).
De nombreuses études leur ont été consacrées : G. Gross (1990), A. Poncet-Montange (1991), A. Monceaux (1993), ...

Les principes seulement de l'approche du LADL ont été présentés ici, sans indiquer les réalisations plus opérationnelles qu'a permises cette description de la langue. M. Gross (1990) conclut son article en insistant sur le fait que *"les mécanismes et représentations utilisés sont entièrement dégagés du sens, mais que toute l'approche consiste à réduire les variations de forme qui cachent la distribution du sens, but (encore éloigné) de la description linguistique. Les méthodes lexicales et syntaxiques disponibles sont loin d'avoir épuisé leur domaine d'application. En fait, aucune langue à ce jour n'a été décrite de façon raisonnablement complète au moyen de ces méthodes. Il apparaît néanmoins que des progrès ont été réalisés quand la description du sens a pu être séparée de la description des formes, alors que rien de cohérent ou de général ne peut être affirmé aujourd'hui pour la moindre unité de sens. On peut seulement espérer que des méthodes de traitement du sens pourront être développées à partir des systèmes lexico-syntaxiques en cours de construction."*

4. Tests effectués par le CSTB : étude MediaConstruct

4.1. Contexte de l'étude

MediaConstruct est une association loi 1901 qui regroupe des professionnels de la construction, des institutions et des entreprises offrant des services informatiques ou télématiques. Elle a pour mission de promouvoir l'usage des nouvelles technologies pour l'information et la communication dans le domaine du bâtiment, de préparer les normes et les standards d'échange d'informations, de référencer et labelliser les produits et les services électroniques pour la construction ...
Cette association a chargé le CSTB de conduire une étude dans le but d'évaluer l'intérêt d'outils linguistiques pour la recherche d'information dans des corpus documentaires.

Pour cela, le CSTB a mis au point un test dans lequel il s'agit de simuler le fonctionnement d'un service d'assistance téléphonique déjà existant, en faisant une recherche automatique d'information à l'intérieur d'un corpus, la recherche d'information se faisant à l'aide d'outils informatiques utilisant des

connaissances linguistiques.
Ce service d'assistance téléphonique existe réellement ; il est proposé par le CATED[3] à des abonnés. Ceux-ci appartiennent à toutes les catégories d'intervenants dans le domaine du bâtiment : les maîtres d'ouvrage, les maîtres d'œuvre, les installateurs, les bureaux de contrôle, ... L'abonné pose sa question à son premier interlocuteur qui identifie le domaine du bâtiment concerné et adresse le demandeur à l'expert attaché à ce domaine. Cet expert dialogue avec son correspondant, reformule éventuellement la question posée, grâce à sa connaissance du domaine, puis la traite. La réponse est donnée au téléphone, immédiatement ou après un délai, et les documents qui justifient cette réponse sont ensuite envoyés par courrier. L'envoi des documents constitue d'ailleurs une étape importante dans la mesure où le CATED est souvent sollicité pour régler des différends entre intervenants, entre bureau d'étude et maître d'œuvre par exemple, et ces justificatifs permettent de fonder les positions des uns et des autres.

Pour ce test, le CSTB a utilisé des questions réellement exprimées par les abonnés du CATED, en restreignant le domaine à celui de la sécurité incendie. Ces questions sont ensuite "posées à un système informatique", par l'intermédiaire d'une interface spécifique à chaque système, celui-ci intégrant ou non des outils linguistiques à un moteur de recherche qui interroge le corpus documentaire choisi, en l'occurrence le CD-Reef.

L'objectif est de vérifier si les réponses apportées par ces systèmes informatiques sont comparables, en termes d'exhaustivité et de pertinence, à celles fournies par l'expert. Ces réponses sont constituées par une collection d'articles renvoyés par le moteur parce qu'ils sont considérés comme pertinents par rapport à la question posée. On peut ensuite les confronter à celles fournies par l'expert, examiner et justifier les éventuelles différences, proposer, quand cela est possible, des aménagements, et finalement conclure sur l'intérêt d'utiliser des outils linguistiques dans la recherche automatique d'information à l'intérieur d'un corpus.
Nous insistons sur le fait que le but de cette étude n'est pas de comparer entre eux les outils informatiques et linguistiques mis en œuvre pour le test, mais d'évaluer l'intérêt d'utiliser des outils utilisant ce type de techniques pour retrouver de l'information dans des textes rédigés en langage naturel.

4.2. Détail de l'étude

4.2.1. Le corpus

Le corpus de référence est constitué par la version électronique du Reef, soit environ 1 300 documents pour près de 25 000 pages et contenant des textes, des graphiques et des tableaux. La version utilisée pour l'étude est au format HTML (compatible avec tous les moteurs de recherche utilisés pour ce test) et regroupe 75 000 unités documentaires pour environ 80 Mo. Ces unités ont été découpées dans le texte initial de manière que la réponse soit significative pour un interlocuteur humain. En effet, si l'unité documentaire renvoyée par le moteur est trop longue, le lecteur humain a du mal à apprécier la pertinence d'une réponse.
D'autre part, le format HTML permet d'attacher un entête à chaque unité documentaire. Le titre du document et les titres des niveaux supérieurs ont été recopiés dans cet entête (voir en annexe un exemple de découpage).

---

[3] Le CATED est un des participants de MediaConstruct.

4.2.2. Les questions

Elles sont exprimées en langage naturel. Elles proviennent de questions réelles posées à l'expert sécurité incendie par les abonnés au service d'assistance téléphonique du CATED. Une cinquantaine a été retenue pour former le test, parmi lesquelles les questions suivantes (la liste complète figure en annexe) :

Quelle doit être la réaction au feu d'un podium d'une salle d'exposition de type T ?
Faut-il prévoir une sonnerie d'alarme cabine pour l'ascenseur d'un ERP ?

4.2.3. Les outils testés et les scénarios mis en place

Deux sociétés proposant une gamme d'outils linguistiques et informatiques complète[4], outils couplés à un moteur de recherche, ont été sélectionnées pour effectuer ce test. Il s'agit de LexiQuest et T-GID.

Les outils de chaque gamme se combinent entre eux, tout en conservant leurs caractéristiques, pour former les différents scénarios de test (les trois composantes qui varient dans un scénario sont le moteur de recherche, l'utilisation des outils linguistiques et/ou de la terminologie spécialisée). Ceux-ci ont été construits afin de mettre en évidence l'importance des différents composants qui interviennent dans l'obtention des réponses : moteurs de recherche, intégration ou non d'outils linguistiques, utilisation ou non d'une terminologie spécialisée propre au domaine (elle concerne la sécurité incendie et le bâtiment).
Il paraît incohérent de tester l'intérêt d'une terminologie spécialisée si on n'a pas d'outils linguistiques pour la mettre en œuvre, il n'y a donc pas de scénario de ce genre.
D'autre part, le moteur de recherche utilisé par T-GID, Spirit, intégrant ses propres outils linguistiques, il n'est pas possible de dissocier son utilisation de celle d'outils linguistiques.
Enfin, une partie de l'expérimentation consiste à mesurer l'intérêt d'ajouter aux dictionnaires une terminologie spécifique au domaine des questions (en l'occurrence la sécurité incendie) pour la recherche d'informations dans le corpus. Cette terminologie a été construite à l'aide d'INTEX en extrayant des expressions correspondant à des patrons syntaxiques, puis ces expressions candidates ont été validées ou rejetées par un expert du domaine pour donner finalement environ 500 termes et 700 liens (cf. chapitre *Etude du corpus* § 2. Un extrait de cette liste est donné en annexe).
Chacune des entreprises concernées a consacré moins de trois jours à l'examen et l'intégration, à ses propres outils, de ces informations supplémentaires. T-GID a considéré nécessaire d'ajouter à son dictionnaire de base une cinquantaine de termes (choisis dans la terminologie) plus autant de liens non typés. LexiQuest a finalement constitué un lexique personnalisé d'environ 630 mots et 1 600 liens ; ce travail qui paraît plus important a été effectué à l'aide des outils d'extraction de sa gamme LexiGuide.

Finalement, les scénarios mis en œuvre pour interroger le corpus sont les suivants :

scénario	moteur de recherche	outils linguistiques	terminologie spécialisée

4 On considère que ces gammes d'outils sont complètes dans la mesure où elles permettent de formaliser et de capitaliser des connaissances linguistiques sous forme de dictionnaires, réseaux sémantiques, etc ; elles proposent des interfaces qui rendent possible le fait de formuler une question à l'adresse d'un corpus ; elles intègrent des règles morphologiques, lexicales, syntaxiques, qui autorisent la reformulation des questions avant d'utiliser un moteur de recherche ; et enfin elles disposent d'outils linguistiques et informatiques qui permettent d'associer une pertinence à une réponse, par rapport à la question posée, de classer et d'afficher ces réponses.

1	Fulcrum	non	non
2	Fulcrum	oui (ceux de Lexiquest)	non
3	Fulcrum	oui (ceux de Lexiquest)	oui
4	Verity	non	non
5	Verity	oui (ceux de Lexiquest)	non
6	Verity	oui (ceux de Lexiquest)	oui
7	Spirit	oui (ceux de Spirit)	non
8	Spirit	oui (ceux de Spirit)	oui

Dans la suite du chapitre, on examinera d'abord, en détail, les caractéristiques des différents outils proposés par chacune des deux sociétés qui sont intervenues dans ce test. En fait, tous les outils développés par une même société partagent les mêmes principes ; on parlera donc de deux familles d'outils : celle de LexiQuest et celle de T-GID.
Puis, en se fondant sur les résultats obtenus, on expliquera et commentera, pour chacune des deux familles, les réponses obtenues en les rapprochant des principes de construction de ces produits.
Enfin, on essaiera de présenter certaines difficultés inhérentes à la recherche automatique d'information dans un corpus, autour de trois thèmes : l'utilisation d'un vocabulaire spécialisé, l'analyse des questions et la représentation sémantique d'une phrase.

4.3. Caractéristiques de la gamme LexiQuest

L'environnement de travail LexiQuest propose des composants modulaires qui permettent de construire une application complète fondée sur un moteur de recherche du marché, en l'occurrence Fulcrum ou Verity, et des dictionnaires généraux (fournis par LexiQuest). Dans cette architecture, on peut assembler différents composants :

- un module qui, couplé au moteur de recherche, permet l'interrogation du fonds documentaire en langage naturel ;
- un autre qui analyse un corpus afin d'en extraire des nouveaux mots, simples ou composés spécifiques d'un domaine ;
- un environnement de travail destiné à gérer les connaissances linguistiques de l'utilisateur ;
- des dictionnaires spécialisés qui existent déjà et peuvent être ajoutés aux dictionnaires standards.

Les différents modules de la gamme adoptent des principes communs que l'on va préciser.

4.3.1. Alphabet

L'alphabet pris en compte par le logiciel est composé de lettres (il y a équivalence entre majuscules et minuscules), chiffres et séparateurs (blanc, signes de ponctuation, apostrophe). Mais les caractères diacritiques ne sont pas pris en compte. Par exemple, les deux mots des paires suivantes:

haler et hâler
sur et *sûr*
mais et maïs

sont analysés de la même manière.

Les deux notions de mot simple et mot composé existent : un mot simple est une chaîne de caractères, un mot composé contient un séparateur. Par exemple :

sécurité est un mot simple
sécurité incendie, s'emparer sont des mots composés

4.3.2. Dictionnaires

LexiQuest fournit des dictionnaires standard auxquels l'utilisateur peut rajouter des dictionnaires personnalisés.
Les dictionnaires concentrent, pour chaque entrée, des informations lexicales, syntaxiques et sémantiques que l'on va détailler.

4.3.2.1. Jeu de catégories grammaticales

Les étiquettes disponibles sont les suivantes : *nom, nom propre, acronyme, adjectif, verbe, adverbe* pour les mots simples comme pour les mots composés.
La structure des mots composés n'est pas précisée. Pour les noms composés par exemple, la seule étiquette disponible est celle de *nom* : *(n)*. On code donc de la même manière :

machine à aléser (n)
poteau d'incendie (n)
potentiel calorifique (n)

Le choix de ces étiquettes pour les composants des composés, a aussi une influence sur les mécanismes de recherche. Les composants des composés peuvent recevoir les étiquettes précédentes ou les suivantes :

- *chaîne* et *typo* qui sont affectées aux composants d'un mot composé qui doivent être recherchés tels quels ;
- *lexique* qui est associée aux composants d'un mot composé qui sont pris en compte lors de l'analyse de la question mais pas recherchés dans les documents.

Par exemple, dans les expressions suivantes : *permis de construire, type L, mètre cube/heure,*

- *construire* et *L* doivent être codés *chaîne*, et non *verbe* pour *construire* sinon on recherchera aussi ses formes conjuguées ;
- */* porte l'étiquette *typo* : on recherchera *cube* près de *heure* ;
- *de* est dans la catégorie *lexique* ce qui permettra de rechercher aussi *permis pour construire.*

4.3.2.2. Indicateur de pertinence d'un mot

A chaque terme du lexique de référence ou du lexique personnalisé, est associé un indicateur de pertinence qui peut prendre cinq valeurs : *élevée, moyenne, normale, basse* et *nulle*. Son niveau de pertinence influe sensiblement sur la manière dont le mot est exploité lors du traitement d'une question.

Un mot du lexique personnalisé aura par défaut une pertinence élevée. Ceux du lexique standard ont des valeurs a priori (et qui sont modifiables selon les applications de l'utilisateur), définies selon leur fréquence dans la langue et leur niveau d'ambiguïté : un terme appartenant au domaine général, est jugé peu pertinent et sa pertinence est donc *moyenne, normale* ou *basse* ; un mot très polysémique a une pertinence *basse* alors qu'un mot monosémique a la valeur *moyenne*. Un mot de pertinence *nulle* dans une question n'est pas recherché dans le texte de référence.

4.3.2.3. Relations sémantiques entre termes

A un terme peut être associé dans le dictionnaire un réseau sémantique formé d'autres mots liés au premier par des relations sémantiques, quelle que soit leur catégorie grammaticale : à un mot simple peut correspondre un mot composé ; à un nom peuvent être associés un verbe, un adverbe, ...
Les relations susceptibles d'être codées sont les suivantes :

- la synonymie ; en particulier la forme développée et l'acronyme ou l'abréviation d'un nom doivent être liés par des relations de synonymie ;
- l'hyperonymie et l'hyponymie ;
- l'association.

Le logiciel définit automatiquement les relations réciproques : si un lien de synonymie ou d'association est établi explicitement entre les noms *N1* et *N2*, le lien réciproque entre *N2* et *N1* est créé automatiquement. De même si *N1* est noté comme hyperonyme de *N2*, alors automatiquement un lien d'hyponymie est établi entre *N2* et *N1*.

Selon les différents sens d'un mot, le réseau sémantique concerné doit être différent. Dans les dictionnaires fournis par LexiQuest, la polysémie de certains termes peut ainsi être représentée grâce à la prise en compte de réseaux sémantiques différents pour un même mot.

Ces extensions permettent, quand elles sont activées, d'élargir le champ de recherche des documents.

4.3.2.4. Domaines

Pour chaque mot du lexique standard existe l'information du domaine d'appartenance. Il existe 21 grands domaines, eux-mêmes divisés d'une manière arborescente en une cinquantaine de domaines prédéfinis et plus restreints ; il n'est pas possible d'en créer de nouveaux.
L'analyse sémantique, qui repose en grande partie sur cette information, a la possibilité de calculer des intersections de domaines mais la vérification manuelle est difficile car cette indication du domaine ne figure pas dans les résultats d'analyse de la question.
On peut associer différents domaines à un mot et à chaque domaine faire correspondre un réseau sémantique aussi bien pour les entrées du lexique standard que celles du lexique personnalisé. Ajouter une nouvelle acception à une entrée du dictionnaire revient ainsi à lui associer un nouveau domaine d'appartenance possible, éventuellement précisé par un réseau sémantique.

4.3.2.5. Mots vides et mots pleins

Les mots grammaticaux sont notés dans le dictionnaire comme des mots vides ; leur effectif est d'une centaine. Les autres sont des mots pleins.

4.3.2.6. Flexion des nouvelles entrées

Quand on crée un mot simple appartenant aux catégories *nom commun*, *verbe* ou *adjectif*, il faut aussi indiquer un modèle, de la même catégorie grammaticale, qui permet de calculer les formes fléchies de la nouvelle entrée. LexiGuide détermine ainsi les formes fléchies du mot.

Pour un mot composé, ses différents composants doivent exister au préalable dans l'un des dictionnaires. La flexion du mot composé est formulée à partir de celle de ses composants. Par conséquent, le lemme (et donc la forme figurant dans le lexique) du nom composé : *structure porteuse* est *structure (n) porteur (adjectif)*.

4.3.2.7. Définition d'une nouvelle entrée du dictionnaire

Un mot est défini par son lemme (simple ou composé) et sa catégorie grammaticale.
Le niveau de pertinence est une information obligatoire : il est déjà indiqué pour les entrées du dictionnaire standard, et par défaut calé à *élevée* pour les entrées du lexique personnalisé. Le niveau de pertinence d'un mot composé est indépendant de celui de ses composants.
Le terme peut être précisé par la construction d'un réseau sémantique.
L'indication d'un domaine auquel le terme appartient n'est pas une information obligatoire pour les mots du dictionnaire standard ni pour le lexique personnalisé.

4.3.2.8. Contrôle sur les dictionnaires

Il n'est pas possible de supprimer un terme du lexique standard, mais on peut lui attribuer une pertinence *basse* ou bien le recréer dans un lexique personnalisé en lui donnant des caractéristiques adéquates (ce qui aura pour effet de supprimer les valeurs initiales des paramètres).
On peut ainsi visualiser le réseau sémantique d'un mot du dictionnaire standard pour le recréer en le modifiant dans le lexique personnalisé.
Par exemple, dans le lexique de base *bâtiment* et *immeuble* sont considérés comme synonymes. Il faudra désactiver ce lien pour notre application sur la sécurité incendie où *bâtiment* a une acception plus technique qu'*immeuble*.

Le dictionnaire standard contient 60 000 lemmes dont 1 500 mots composés correspondant à 66 000 sens (i.e. réseaux sémantiques) différents. Les effectifs des différentes catégories syntaxiques sont les suivants :

- 38 000 noms
- 11 000 adjectifs
- 4 700 verbes
- 1 500 adverbes.

D'un point de vue sémantique, ces entrées regroupent :

- 8 000 liens générique/spécifique
- 20 000 liens de synonymie
- 20 000 liens d'association.

Tous les champs associés aux entrées du dictionnaire : étiquette grammaticale, pertinence, domaine d'appartenance, extension sémantique sont constants, ils ne varient pas avec la prise en compte des textes qui constituent les ressources documentaires de la nouvelle application. En revanche, l'utilisateur peut modifier et ajuster les informations contenues dans le dictionnaire standard en les reprenant et en les modifiant dans son dictionnaire personnalisé.

4.3.3. Traitement d'une question par les outils Lexiquest

La question formulée en langage naturel par l'utilisateur est analysée en trois étapes par les outils linguistiques et informatiques de la gamme Lexiquest, puis transformée en une requête booléenne.

4.3.3.1. Analyse de la question

La première analyse est morphologique ; elle a pour but de :

- identifier les mots simples et les mots composés à l'aide des dictionnaires;
- faire des prédictions sur la catégorie grammaticale des mots inconnus ;
- sélectionner une étiquette syntaxique quand il y a ambiguïté.

Les prédictions concernant le choix de la catégorie grammaticale d'un mot inconnu ou ambigu sont mises en œuvre grâce à l'utilisation de règles qui prennent en compte la grammaire du français.

Les ambiguïtés concernant en particulier la segmentation en mots simples et mots composés sont traitées grâce à des règles de grammaire.

L'analyse syntaxique s'appuie sur l'analyse morphologique afin de :

- segmenter la phrase en groupes fonctionnels ;
- hiérarchiser ces groupes de mots et leur affecter une fonction dans la phrase, quand les règles du système permettent de conclure.

Enfin, la désambiguïsation sémantique tente d'éliminer certaines des significations associées aux mots de la question. En effet, dans les dictionnaires fournis par LexiQuest, la polysémie est prise en compte sous la forme de réseaux sémantiques différents associés à un même terme. Dans cette étape, il faut diminuer l'ambiguïté sémantique liée aux différentes interprétations des mots, celles-ci étant rendues possibles par les réseaux, en éliminant ceux non pertinents.

Les règles utilisées aussi bien lors de l'analyse morphologique que syntaxique ou sémantique, résultent de l'expertise de LexiQuest. Elles ne sont pas accessibles à l'utilisateur. Le principe adopté consiste à conserver les ambiguïtés détectées le plus longtemps possible, afin de ne pas être à l'origine de silences. Concrètement, cela conduit à traiter toutes les analyses possibles là où il peut y avoir ambiguïté morphologique, lexicale ou syntaxique, et à choisir, sur des critères qui ne sont pas visibles pour l'utilisateur, une analyse avant de formuler les clauses de recherche.

L'analyse mise en œuvre pour interpréter la question se limite à du pattern-matching, en excluant toute aide syntaxique ou terminologique, quand le texte de la question dépasse une centaine de caractères.

4.3.3.2. Construction des clauses de recherche

Après l'analyse sémantique de la question, celle-ci est soumise à deux traitements : la décomposition graduelle et l'extension sémantique des termes afin de créer des clauses de recherche utilisables par le moteur.

4.3.3.2.1. Décomposition graduelle

La décomposition graduelle de la question consiste à l'écrire sous forme de combinaisons de mots, appelées clauses de recherche. La première clause contient tous les mots de la question, les clauses suivantes certains des mots de la question, et les dernières, chacune un mot de la question. Pour les clauses correspondant aux deux premières catégories la recherche se fait *en contexte* ; dans le dernier cas, elle est *hors contexte*.
Dans une recherche en contexte, les mots identifiés comme des mots grammaticaux sont considérés comme des mots vides et ne sont pas pris en compte. Les mots coordonnés sont comptés comme un seul mot, et la longueur de la clause est limitée à quatre mots.

Dans une recherche hors contexte, les mots de pertinence basse ne sont pas recherchés. Et les documents atteints de cette manière ont un score moins élevé que ceux retrouvés grâce à une recherche en contexte.

Par exemple, une question posée sous la forme du groupe nominal suivant :

évolution de l'indice des prix en France

en tenant compte :

- du découpage en mots simples et mots composées tel qu'il est proposé par les entrées des lexiques ;
- de l'information mot vide/mot plein ;
- de l'indicateur de pertinence associé à chaque mot lemmatisé de la question et associé à une entrée du dictionnaire ;

sera décomposée de la manière suivante :

Recherche en contexte	Recherche hors contexte
évolution, indice des prix, France	*indice des prix*
évolution, indice des prix	*France*
indice des prix, France	*évolution*

La clause *(évolution, France)* ne figure pas dans la liste car son score de pertinence, calculé lors de l'analyse de la question, est trop bas. Et, en effet, *(évolution, France)* aurait concerné les textes qui évoquent l'évolution de la France dans tous les domaines et, bien que construite, elle est écartée avant d'être appliquée au corpus.

4.3.3.2.2. Extension sémantique

L'extension sémantique a pour but de réduire le silence : au lieu de se limiter aux mots de la question, LexiQuest va interroger le corpus documentaire en recherchant aussi les termes liés aux mots d'origine grâce aux réseaux sémantiques associés. Pour contrôler cette extension, une distance sémantique est mise en place qui va mesurer l'éloignement entre deux termes, cette distance dépendant à la fois :

- du niveau de pertinence du mot initial
- du type de lien parcouru entre le terme initial et celui atteint
- du niveau de pertinence du mot obtenu par l'intermédiaire du réseau sémantique
- du nombre de mots de la question.

Il est alors possible de limiter l'extension sémantique en fixant :

- une distance sémantique maximale au-delà de laquelle on ne retiendra pas le nouveau mot,
- un nombre maximal de mots, obtenus par ce procédé, que l'on prendra en compte dans la nouvelle recherche.

Dans la dernière version du produit, l'utilisateur a la possibilité de visualiser les domaines et réseaux sémantiques pris en compte pour chaque terme d'une question, et il peut valider ou rejeter certains domaines. Cette vérification permet de contrôler l'analyse automatique produite.

4.3.3.3. Pondération des clauses de recherche

Cette valeur de pondération est associée à chaque clause de recherche produite lors de l'analyse de la question, pour tenir compte de la représentativité de cette clause par rapport aux mots identifiés dans la question. La valeur maximale est donnée à la clause composée à partir de la totalité des mots de la question, les autres se voient attribuer des valeurs qui varient en fonction de :

- la pertinence des mots figurant dans la clause ;
- le nombre de mots qui composent la clause ;
- la fonction grammaticale des mots ou des groupes de mots, identifiés lors de l'analyse syntaxique, qui figurent dans la clause.

La valeur de pondération associée à une clause est dégradée en fonction de la distance entre la requête d'origine et celle obtenue en utilisant les différentes relations qui existent dans les dictionnaires pour chacun des mots de la question.

4.3.8. Recherche des documents-réponses

Une fois les clauses de recherche construites, elles sont transmises au moteur de recherche qui n'est pas un outil de la gamme LexiQuest, mais un moteur du marché, en l'occurrence Fulcrum ou Verity. Les scénarios proposés lors de l'élaboration du test mettent ces deux moteurs dans les mêmes conditions d'utilisation.

4.3.8.1. Algorithmes de recherche proposés par les moteurs Fulcrum et Verity

Fulcrum peut appliquer différents algorithmes de recherche :

- le premier algorithme ne tient pas compte du nombre d'occurrences du mot figurant dans la question pour une unité documentaire donnée ;
- pour calculer la pertinence d'un document, Fulcrum utilise un algorithme statistique de fréquence (plus nombreuses sont les occurrences du terme recherché dans le texte, plus le texte est jugé pertinent) et de fréquence inverse (plus nombreuses sont les occurrences du terme recherché sont présentes dans le texte, moins celui-ci est jugé pertinent);
- un autre algorithme consiste à construire un vecteur censé représenter la question et les documents du corpus : rechercher une information consiste alors à trouver les textes les plus proches de la question, au sens de la distance entre deux vecteurs.

LexiQuest, après avoir expérimenté ces différents algorithmes, a choisi d'utiliser par défaut le premier présenté : la procédure choisie rend seulement compte de la présence ou de l'absence d'un mot de la question dans le document examiné, sans tenir compte du nombre de ses occurrences dans le document.

4.3.8.2. Calcul de la pertinence d'un document

La pertinence d'un document est la valeur de pertinence attribuée à la combinaison de mots (la clause dans le vocabulaire LexiQuest) qui a sélectionné le document examiné. Cette quantité est ensuite utilisée pour ordonner ces documents selon leur pertinence présumée.

4.4. Caractéristiques de la gamme T-GID

Les outils T-GID utilisent deux types d'informations:

- les informations constantes, indépendantes du fonds documentaire utilisé, organisées autour des dictionnaires standards, "dictionnaire public" selon la terminologie T-GID ;
- les informations qui dépendent de l'application ; elles sont fondées sur "la base", c'est-à-dire l'ensemble des textes qui forment le corpus documentaire.

4.4.1. Dictionnaires

Les dictionnaires concentrent les informations constantes.

4.4.1.1. Mots simples et mots composés

Les deux notions de mot simple et mot composé existent ; le qualificatif "mot composé" recouvre aussi bien des expressions complètement figées : *un homme-grenouille, un cordon bleu, une oie blanche, prendre le taureau par les cornes, ...* que des expressions composées qu'on pourra choisir de faire rentrer dans les dictionnaires : *à l'occasion de, chaudière à vapeur, ...*

4.4.1.2. Alphabet

L'alphabet pris en compte par les outils T-GID est composé de lettres (il n'y a pas équivalence entre majuscules et minuscules), chiffres et séparateurs (blanc, signes de ponctuation).
Les lettres isolées ne sont pas significatives.[5]

4.4.1.3. Jeu de catégories grammaticales

Il existe cinq étiquettes principales : *nom, adjectif, verbe, adverbe, adverbe* mais la représentation de la langue étant fondée sur une grammaire positionnelle, ces cinq étiquettes ont été affinées jusqu'à donner finalement 115 étiquettes syntaxiques (qui n'ont pas été communiquées). Par exemple, dans les expressions suivantes :

une robe bleue
une belle robe

les deux adjectifs *bleue* et *belle* ne reçoivent pas la même étiquette syntaxique.

Un mot composé dispose d'une étiquette qui lui est propre, et tous ses composants doivent être reconnus comme des mots simples du dictionnaire.

[5] Pour ce test qui concerne l'application de la réglementation sur la sécurité incendie, T-GID a dû modifier cette caractéristique de ses outils. En effet, la totalité de la réglementation sécurité incendie repose sur la notion de classement des établissements selon deux critères : le type de l'activité hébergée par l'établissement et l'effectif de public admis. Or, le type de l'établissement est donné par une lettre : *un établissement de type L.*

4.4.1.4. Mots vides – mots pleins

Le vocabulaire du dictionnaire est divisé en mots vides et mots pleins.
Chaque entrée du dictionnaire public se voit attribuer un de ces deux états. La valeur de ce champ est fondée sur une étude de la langue pratiquée chez T-GID ; elle est modifiable par les fabricants du produit, mais pas par le client pour son application propre. Cette quantité est utilisée lors de l'analyse de la question, et de la recherche d'information à l'intérieur de la base.

4.4.1.5. Relations sémantiques

A une entrée du dictionnaire, on peut associer différents autres termes sans contrainte sur la nature grammaticale du mot initial ni des termes associés. Les relations prises en compte dans le dictionnaire sont classiquement la synonymie, l'hypéronymie, l'hyponymie, l'association, ...
Ces liaisons permettent aussi, pour des mots qui n'existent que dans des expressions composées, de faire le lien entre les mots simples de l'expression et l'expression complète. Par exemple, la recherche de l'expression *au fur et à mesure* sera déclenchée dès que le système aura identifié les mots *fur* ou *mesure* avec les étiquettes syntaxiques adéquates parce qu'existent, dans le dictionnaire, un lien entre chacun des deux mots simples et l'adverbe composé *au fur et à mesure*.

4.4.1.6. Les entrées des dictionnaires

Le dictionnaire standard contient 500 000 formes fléchies. Chaque entrée est définie par les informations suivantes :

- un lemme ;
- une étiquette syntaxique ;
- un poids ;
- une indication quant à l'état : mot plein ou mot vide ;
- une indication quant à l'état : mot de liaison ou mot de rupture (voir plus loin 3.4.2.2.).

Chaque forme fléchie constitue une entrée autonome du dictionnaire.
Les entrées du dictionnaire privé sont construites sur le même modèle, mais elles ne renseignent pas le champ : mot de liaison ou mot de rupture.

4.4.1.7. Contrôle sur les dictionnaires

Les entrées sont regroupées dans différents dictionnaires, dont il est possible de gérer les priorités d'utilisation. En particulier, le dictionnaire privé est plus prioritaire que le dictionnaire public.

4.4.2. Informations fondées sur la base

Puisqu'elle est formée de l'ensemble des textes utilisés pour une application donnée, la base prend en compte les spécificités lexicales et syntaxiques du domaine. Ainsi les informations fondées sur ces textes doivent être recalculées pour chaque base qui constitue de fait une nouvelle application.

4.4.2.1. Indicateur de pertinence

A chaque entrée du dictionnaire est associée une pertinence qui est quantifiée par la valeur d'un poids qui est d'autant plus élevé que le mot est moins courant.

Les mots inconnus ont par définition un poids nul, il en est de même pour les mots vides.
Ce poids varie donc selon l'application ; il est recalculé puis associé à chaque entrée du dictionnaire public, comme du dictionnaire privé.

Puisque chaque lemme a son poids, les différentes formes, si elles sont rattachées à des lemmes différents, peuvent alors être créditées de poids différents. Par exemple, selon que l'on relie dans les dictionnaires *établissements* au lemme *établissement* ou au lemme *établissements*, il pourra être crédité d'un poids différent.

Il est donc nécessaire de désambiguïser les mots des questions comme ceux de la base. Cette désambiguïsation se fait à l'aide de règles qui constituent l'expertise de T-GID ; ces règles ne sont pas accessibles au client, ni modifiables.
Si un mot n'est pas désambiguïsé après application des règles, on lui associe le poids correspondant au lemme le plus pertinent parmi les lemmes auxquels il peut être raccordé.

4.4.2.2. Notion de dépendance

Elle est différente de celle de mot composé. Elle permet de mesurer le lien sémantique entre des mots topologiquement voisins dans la phrase, et qui ne constituent pas une entrée du dictionnaire. Elle est calculée avec les mêmes règles pour les mots de la question et ceux de la base, ce qui, pour T-GID, assure une cohérence des traitements : même si l'analyse est fautive, le fait qu'elle soit répétée de la même manière dans la question et dans le corpus permet de diminuer le bruit et même de trouver des réponses pertinentes.
Par rapport à cette notion, existent deux types de mots : les mots de liaison et les mots de rupture. Cette caractéristique s'applique à tous les mots du dictionnaire public, quelle que soit leur étiquette syntaxique. Deux mots séparés par un mot vide seront considérés comme reliés par une relation de dépendance si ce mot est un mot de liaison ; si le mot vide est un mot de rupture, cette possibilité est définitivement exclue.

La dépendance n'est possible que pour des mots proches : la distance maximale prise en compte est de un. Par exemple, dans *chaudière à vapeur*, *chaudière* et *vapeur* sont considérés comme liés par une relation de dépendance parce qu'ils sont séparés par un mot vide *à* qui est ignoré lors de la détection de la dépendance. En revanche, dans l'expression *chaudière autonome à vapeur*, *chaudière* et *vapeur* trop éloignés (distance de 2) ne pourront être détectés comme dépendants, mais ce sera possible pour *chaudière autonome*. De plus, si *chaudière autonome* est entré dans le dictionnaire sous la forme d'un mot composé, il est ensuite considéré comme un seul mot, et donc *vapeur* ne se trouvant plus qu'à une distance de 1 de *chaudière autonome*, les deux entités *chaudière autonome* et *vapeur* seront liées par une relation de dépendance.

4.4.3. Recherche d'information

Spirit associe son propre moteur de recherche à ses outils linguistiques. Très classiquement, Spirit traite la question avant de la reformuler pour la confronter aux textes de la base.

4.4.3.1. Analyse de la question

L'analyse est fondée sur l'exploitation des dictionnaires et la mise en œuvre d'un ensemble de règles syntaxiques puis statistiques, qui permettent de désambiguïser les mots qui composent la question.

Quand des dépendances sont détectées dans une question, à l'aide de l'algorithme détaillé en 3.4.2.2., un poids important est associé au n-uplet identifié : le système considérera comme très pertinent l'expression ainsi formée et recherchera cette association de termes dans le corpus.
Les utilisateurs ne peuvent pas accéder aux règles qui permettent d'établir des dépendances, a fortiori les modifier.
Quand une dépendance est détectée entre deux formes, elle est en fait associée aux lemmes représentées par ces deux formes : *moyens de secours* étant reliés par une dépendance, le poids associé à la paire (*moyens, secours*) est important, il est égal au poids de (*moyen, secours*). L'ordre d'apparition des termes n'étant pas pris en compte, les couples (*secours, moyens*) et (*secours, moyen*) sont aussi pertinents que (*moyen, secours*).

4.4.3.2. Construction des clauses de recherche

Le moteur de recherche de Spirit intègre des outils linguistiques qui assurent, entre autres, le traitement de la question afin d'identifier les formes interrogatives, éliminer les mots de la phrase qui sont considérés comme non significatifs, puis rechercher les dépendances, etc. La question posée initialement est ainsi retraduite en un ensemble de n-uplets dont les occurrences vont être recherchées dans la base. Par exemple, pour la question suivante :

(60) Quelle est la périodicité des visites de la commission de sécurité dans un lycée de 2 000 élèves ?

le traitement initial reconnaît la forme *quelle ...est ... ?* et la supprime, identifie l'expression numérique *2 000 élèves* comme non significative et conserve finalement le *GN* suivant :

périodicité des visites de la commission de sécurité dans un lycée

Spirit applique ensuite des règles de calcul des dépendances pour former toutes les dépendances possibles. Celles-ci sont conservées, et donc présentées pour le classement des textes pertinents, seulement si Spirit trouve au moins une occurrence de cette paire dans la base ; sinon, cette paire est considérée comme non pertinente et n'est pas prise en compte dans la présentation des résultats.
Pour la question (5),

(5) Quel doit être le niveau de sécurité incendie d'un local d'archives situé dans un ERP contigu à un parc de stationnement ?

les paires considérées comme pertinentes, i.e. les couples de deux termes identifiés comme dépendants dans la question et retrouvés dans la base, sont les suivants :

doit-niveau, parc-stationnement,

auxquels s'ajoutent les mots simples de la question :

sécurité, incendie, situé, contigu.

Cet ensemble de mots caractérise la classe 1 des documents trouvés, i.e. celle qui contient les documents les plus pertinents. On peut remarquer que la paire *doit-niveau* trouvée dans la question a été validée, à tort, parce qu'elle figure aussi dans la base ; d'ailleurs l'unité documentaire trouvée en réponse à cette question ne contient pas la réponse telle qu'elle est formulée par l'expert.

4.4.3.3. Utilisation de l'extension sémantique

Chaque entrée du dictionnaire a la possibilité d'être reliée à d'autres termes par l'intermédiaire d'un lien. Tous ces liens sont ensuite exploités au moment de l'analyse de la question afin d'étendre les mots employés dans le libellé de la question à ceux qui leur sont associés par des relations variées. Toutes les relations sont utilisées de la même manière lors de cette extension sémantique.

D'autre part, afin de limiter le bruit, un seul niveau d'extension sémantique est mis en œuvre. Par exemple : si *A* est lié à *B* et *B* est lié à *C*, on ne rajoutera, dans les clauses de recherche construites sur la question contenant *A*, que des clauses obtenues avec le terme *B*, mais pas celles formées sur *C*.

4.4.3.4. Calcul de la pertinence d'un document

La pertinence d'un document est une fonction calculée à partir des poids associés aux mots de la question, mots simples et mots composés, qui sont trouvés dans le document. Le score ainsi obtenu se traduit, pour le document, par l'appartenance à une classe. Tous les documents appartenant à une classe sont alors considérés comme ayant le même intérêt par rapport à la question posée et sont présentés comme tels en réponse à la question posée.

Les caractéristiques des deux gammes d'outils, LexiQuest et Spirit, ont été détaillées dans les paragraphes précédents 3.3. et 3.4. Après avoir rapidement présenté les résultats du test mis en place par le CSTB pour évaluer l'intérêt d'utiliser des méthodes linguistiques et informatiques lors d'une recherche automatique d'information, nous allons essayer d'analyser, pour chaque gamme, les conséquences de ces choix constructifs, d'une part, dans la construction et la maintenance des ressources lexicales spécifiques à un domaine technique comme la sécurité incendie, d'autre part dans la recherche automatique d'informations dans un corpus documentaire, en illustrant nos propos par des exemples tirés du test.
Dans une deuxième partie, nous développerons des commentaires sur certaines des difficultés rencontrées dans ce test et qui, nous semble-t-il, se seraient manifestées quels que soient les outils linguistiques et informatiques employés.

4.5. Commentaires sur les conséquences des choix constructifs

Le premier commentaire concerne les résultats obtenus dans le test mis en place par le CSTB. Ensuite, indépendamment de ces résultats, les commentaires aborderont des thèmes qui diffèreront selon la gamme concernée puisque les principes sur lesquels reposent les outils sont distincts. Les outils LexiQuest sont fondés sur des dictionnaires qui concentrent une grande partie de l'information qui sera mise en œuvre pour analyser le corpus et les questions. Les produits T-GID en revanche, même s'ils utilisent les lexiques standard ou spécialisé, font aussi un large usage d'algorithmes ; nos commentaires porteront plutôt sur ces algorithmes.

4.5.1. Présentation des résultats

Le test mis en place par le CSTB contient 47 questions. Sept scénarios ont été définis qui diffèrent par le moteur de recherche utilisé, les outils linguistiques mis en œuvre, la terminologie spécialisée ajoutée aux dictionnaires. Dans chaque scénario, pour chaque question, on compare les réponses

fournies par l'expert et celles proposées par le système informatique[6]. Les résultats sont synthétisés, d'une manière quantitative, par deux valeurs *C1* et *C2* définies de la manière suivante :

C1 = nombre de bonnes réponses obtenues parmi les trente[7] premières / nombre de bonnes réponses attendues

C2 = total des scores des bonnes réponses obtenues / total des scores associés à la totalité des bonnes réponses attendues

La définition de *C1* est celle du rappel, telle qu'on peut la trouver dans la littérature concernant la recherche automatique d'informations dans des textes ; celle de *C2* ne correspond pas vraiment à la précision des réponses. Avec cette formule, C2 dépend à la fois du rappel et de la précision dont la formule suit :

précision = total des scores des bonnes réponses obtenues / total des scores associés à la totalité des réponses (bonnes ou mauvaises) obtenues

En effet, C2 varie si le nombre de bonnes réponses augmente, alors qu'il ne varie pas si c'est le nombre de mauvaises réponses obtenues qui augmente et que ces mauvaises réponses sont classées en premier (et dans ce cas, la précision elle diminue fortement).
Le calcul de *C2* n'est pas immédiat : à chaque bonne réponse trouvée est associé un score qui représente le rang de la réponse par rapport à toutes celles obtenues. Meilleur est le rang, plus élevé est le score. Pour les réponses de T-GID, le calcul est un peu différent parce que les réponses sont regroupées en classes, et toutes les réponses de la même classe ont le même rang, et donc le même score.

Pour chaque question du CATED, ces deux coefficients sont calculés, puis les valeurs moyennes de *C1* et *C2* pour chaque scénario. Toutes ces variables sont regroupées dans un tableau présenté page suivante. Sur cette présentation des résultats, on peut remarquer que les rappel s'améliore sensiblement lorsqu'on passe du scénario qu'on peut qualifier de minimum où le moteur de recherche est utilisé seul, à un scénario plus élaboré où on ajoute d'abord des outils logiciels linguistiques, puis une terminologie spécifique à la sécurité incendie. La valeur de C2 en revanche ne suit pas la même évolution : elle augmente, en même temps que le rappel, quand on apporte des connaissances linguistiques à Fulcrum, mais décroît très légèrement avec l'addition de la terminologie. Pour Verity, la précision varie dans le même sens que le rappel, mais avec des améliorations moindres. Et pour Spirit, ~~la précision~~ décroît avec l'utilisation des dictionnaires spécialisés. En revanche, il est plus difficile de qualifier la précision des réponses puisque la formule de C2 est inadéquate et ne donne pas d'information précise.
De manière plus globale, on peut remarquer que les réponses ne sont pas très bonnes : un seul scénario sur sept dépasse 50% de taux de rappel ; la précision, elle, est meilleure (entre 55 et 77%) mais elle s'applique à un petit nombre de bonnes réponses (puisque le rappel est faible) et concerne les trente premiers documents, ce qui oblige à lire un grand nombre de textes avant d'espérer reconnaître une information pertinente.

[6] On désigne par ce terme l'ensemble des ressources mis en œuvre par un scénario : moteur de recherche, plus éventuellement des outils linguistiques, plus éventuellement une terminologie spécialisée.

[7] Dans tous les cas le nombre de réponses attendues est inférieur à cette limite arbitraire.

N° question	Nbre de réponses attendues	Sc.1 Fulcrum		Sc.4 Verity		Sc.2 Fulcrum+LQ		Sc.5 Verity+LQ		Sc.7 Spirit		Sc.3 Fulcrum+LQ +termino		Sc.6 Verity+LQ +termino		Sc.8 Spirit +termino	
		C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
01	4	0	0	0	0	0,25	1	0,25	1	1	0,94	1	0,89	1	0,96	1	0,94
02	2	0	0	0	0	0,5	0,17	0	0	0,5	0,68	0,5	1	0	0	0,5	0,41
03	5	0	0	0,2	0,97	0	0	0	0	0,2	0,43	0	0	0,2	0,57	0,2	0,45
04	1	0	0	1	1	0	0	0	0	1	0,5	0	0	1	0,7	1	0,38
05	4	0	0	0	0	0	0	0	0	0,5	0,21	0	0	0	0	0	0
06	2	0	0	0	0	0	0	0	0	0	0	0,5	1	0,5	0,97	0	0
08	3	0	0	0,33	0,63	0	0	0	0	0	0	0	0	0	0	0	0
09	2	0	0	1	0,56	0	0	0,5	0,87	0,5	0,8	0	0	0,5	1	1	0,62
10	19	0	0	0,16	0,24	0,11	0,86	0,16	0,89	0,11	0,48	0,21	0,65	0,74	0,61	0,37	0,35
11	2	0	0	0,5	1	0,5	0,97	0,5	1	0,5	1	0,5	0,93	0,5	0,97	0,5	1
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0,38
15	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0,9	1	0,22
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0,38
18	1	1	0,97	0	0	1	0,97	0	0	0	0	1	1	1	0,97	1	0,44
20	1	0	0	0	0	0	0	0	0	0	0	1	0,85	0	0	1	0,5
23	1	0	0	0	0	1	0,47	1	0,8	0	0	0	0	1	0,97	0	0
25	1	0	0	1	0,3	0	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	12	0,08	0,17	0,42	0,6	0,25	0,63	0,25	0,87	0,25	0,73	0,42	0,57	0,25	0,72	0,33	0,68
28	10	0,1	0,3	0,1	0,9	0,1	0,73	0,3	0,43	0,4	0,96	0,2	0,73	0,2	0,78	0,4	0,86
29	2	0,5	1	0,5	1	0	0	0	0	0,5	1	0,5	0,67	0	0	0,5	1
30	4	0	0	0	0	0	0	0	0	0,25	0,27	0	0	0,5	0,8	0,25	0,92
31	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	2	0	0	0,5	0,87	0,5	0,4	0,5	0,07	0,5	0,55	0,5	0,5	0	0	0,5	1
34	4	0,5	0,49	0,5	0,78	0,75	0,8	0,75	0,97	0,75	0,68	0,75	0,97	0,75	0,99	0,75	0,79
35	3	0	0	0	0	0	0	0	0	1	0,75	0,33	0,37	0	0	1	0,75
37	3	0,33	0,03	0,33	0,97	0,33	0,9	0,33	1	0,33	1	0,33	1	0,33	0,97	0,33	1
38	24	0,12	0,49	0	0	0,42	0,71	0,25	0,59	0,08	0,73	0,29	0,39	0,62	0,78	0,58	0,61
39	3	0	0	1	0,62	0	0	1	0,52	1	0,76	0	0	0	0	1	0,75
43	3	1	0,71	1	1	1	0,93	1	0,97	1	1	1	0,21	0,67	1	0,67	1
44	1	1	0,67	1	1	1	1	1	1	1	1	1	1	1	0,77	0	0
45	4	0	0	0,5	0,69	0,5	0,42	0,5	0,9	0,75	0,72	0,75	0,86	0,25	0,93	0,5	0,93
46	5	0,4	0,68	0,4	0,75	0,8	0,19	0,8	0,46	0,8	0,5	0	0	0,8	0,93	0,8	0,7
47	1	1	1	1	0,97	1	1	1	1	1	1	1	0,97	1	1	1	1
48	9	0	0	0	0	0,11	0,93	0	0	0,11	0,43	0	0	0	0	0,11	0,56
49	10	0	0	0	0	0	0	0	0	0	0	0,6	0,58	0,3	0,55	0	0
50	1	0	0	0	0	0	0	1	0,87	1	0,95	1	0,9	1	0,23	1	1
51	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0,17	1	0,54
52	9	0,11	0,17	0,44	0,63	0,89	0,69	1	0,94	0,67	0,92	0,33	0,66	1	0,85	0,78	0,95
53	1	0	0	0	0	0	0	1	0,6	1	1	1	0,97	1	0,97	1	1
54	1	1	0,63	1	0,63	1	0,73	1	0,7	1	0,74	1	0,63	1	0,63	1	0,71
55	1	1	0,73	1	0,97	1	0,9	1	0,47	1	0,43	1	0,93	1	0,53	1	0,43
56	1	1	0,53	0	0	0	0	0	0	0	0	1	0,47	1	0,87	1	0,75
57	1	1	0,1	0	0	1	0,57	1	0,13	1	1	1	1	0	0	1	1
58	1	1	0,37	1	1	0	0	0	0	1	0,86	0	0	1	1	1	0,86
59	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	1	0	0	0	0	1	0,77	1	0,33	1	0,46	1	0,63	1	0,77	1	0,46
C1 moyen		0,24		0,32		0,32		0,36		0,46		0,42		0,49		0,60	
C2 moyen			0,55		0.73		0.69		0,74		0,75		0.68		0.77		0,71

Légende :

0,00 <valeur < 1,00

valeur = 1,00

4.5.2. La gamme LexiQuest

4.5.2.1. Les informations contenues dans les dictionnaires

4.5.2.1.1. Non-prise en compte de la syntaxe

A une forme donnée ne peut correspondre qu'une seule entrée du dictionnaire ; on ne peut donc différencier les mêmes formes d'un mot en utilisant les constructions qui peuvent lui être associées. Par exemple, une seule entrée est consacrée au verbe *voler* sans distinguer ses constructions transitive et intransitive :

l'avion vole la nuit
le brigand vole les voyageurs

En revanche, pour les dictionnaires fournis par LexiQuest, des réseaux sémantiques différents sont associés aux différents sens d'un mot. Cette gestion de la polysémie n'est pas possible pour les entrées du dictionnaire personnalisé.

4.5.2.1.2. Mise à jour des dictionnaires

La granularité des dictionnaires ne peut être homogène parce qu'elle repose sur des précisions et des améliorations ponctuelles, et non sur l'observation et le traitement de phénomènes réguliers. Par exemple, dans le domaine de la sécurité incendie, la notion d'*accès* est essentielle, elle se traduit par plusieurs types de constructions :

les locaux doivent être accessibles aux moyens de secours
l'accès des locaux aux moyens de secours
les moyens de secours doivent pouvoir accéder aux locaux

et si l'on veut limiter le silence sur ce thème il faudra faire des liens dans le lexique entre les lemmes *accès*, *accessible*, *accéder*, *accessibilité*, etc. Mais si la question est posée avec des synonymes, et dans tous les cas synonymes seulement dans ce contexte,

les locaux doivent être ouverts aux moyens de secours
l'ouverture des locaux aux moyens de secours
les moyens de secours doivent pouvoir (entrer + pénétrer) dans les locaux

il faut mettre en place de nouveau "manuellement" des liens sémantiques et qui dépendent toujours du contexte, pour tous les autres termes concernés.

D'autre part, il est difficile de capitaliser les modifications ou les ajouts apportés aux lexiques pour une étude particulière car la conception du système ne facilite pas la gestion simultanée de l'ensemble des emplois d'un même terme.

4.5.2.2. Exploitation des liens sémantiques

4.5.2.2.1. Synonymie

En termes de vocabulaire, les liens de synonymie établis entre mots dans les dictionnaires généraux sont assez peu efficaces puisque dans le vocabulaire technique, les mots n'ont généralement pas de synonymes.
La synonymie est un lien difficile à mettre en œuvre en cas de polysémie d'un nom. Il est hasardeux de considérer comme synonymes par exemple des mots comme *sauvegarde* et *enregistrement.*
Une autre difficulté réside dans les différences de vocabulaire : les mots employés par les hommes de

terrain dans les questions sont souvent assez éloignés de ceux du législateur que l'on trouve dans les réponses.

4.5.2.2.2. Exploitation de l'extension sémantique

L'extension sémantique effectuée à partir des réseaux sémantiques des mots reconnus dans la question peut produire beaucoup de bruit.
Les liens de synonymie, hormis pour les mots techniques quand ces liens existent, ne peuvent être exacts : une *voiture* n'est pas toujours une *automobile* - l'acception *wagon* liée à *voiture* est encore vivante - et un *bâtiment* n'est pas nécessairement un *immeuble*, le sens prédicatif de *bâtiment* est aussi présent dans la langue. Si les modifieurs ne sont pas pris en compte pour former des noms composés et associer un synonyme à l'expression complète et non à chacun de ses constituants, ces liens peuvent être à l'origine de nombreux contresens qui encombrent la liste des documents trouvés.

4.5.2.3. Interprétation des questions

Les questions posées sont le plus souvent longues parce qu'elles sont précises et donc, dépassant la limite à partir de laquelle LexiQuest renonce à toute analyse syntaxique ou linguistique, obtiennent des réponses non pertinentes.
D'autre part, quand la question est reformulée en autorisant des extensions sémantiques, le nombre de réponses proposées augmente et le point clé devient la mesure de la pertinence d'une réponse par rapport à la question posée. Et dans ce cas, il n'est pas rare que le rang des bonnes réponses devienne plus mauvais parce que les réponses pertinentes sont précédées, dans le classement, par celles obtenues grâce aux extensions sémantiques des mots de la question.

4.5.2.4. Calcul de la pertinence d'un document

Les mots très polysémiques ont une pertinence basse, ceux monosémiques une pertinence moyenne ou élevée. Cette dichotomie paraît excessive. Les mots monosémiques appartiennent le plus souvent à des terminologies scientifiques ou techniques : la définition d'un *groupe abélien* est précise en mathématique, celle d'une *appendicectomie* ne fait pas de doute pour un chirurgien ; et ces termes ne sont pas employés hors de leur contexte spécialisé. Il n'en est pas de même pour des mots courants employés avec des acceptions particulières : par exemple en sécurité incendie, les mots *type* et *établissement* ont une définition technique et donc une pertinence qui doit être élevée comme dans : *un établissement de type M*. Mais cela n'empêche pas des constructions dans lesquelles ces mots retrouvent une acception concourante, et donc imprécise, et qui varie selon le contexte. On trouve donc, toujours dans le corpus de la sécurité incendie, les expressions suivantes :

les responsables d'établissement publics et privés
l'établissement (du diagnostic + d'une ligne de conduite[8]*)*
(un mur + un ventilateur + ...) du type ...
un foyer type

dans lesquelles les mots *établissement* et *type* n'ont pas "l'acception sécurité incendie". De même en terme de construction syntaxique, on rencontre en même temps les deux expressions :

en fonction du type (de l'activité ferroviaire + d'élément de construction)
et *en fonction du type d'activité et de l'effectif du public*

---

[8] dans le corpus *Le Monde* (1994).

qui sont identiques et donc lesquelles la pertinence de *type* par rapport à la sécurité incendie est complètement différente : dans la première ligne, on peut remplacer *en fonction du type* par *en fonction (du genre + de la nature)* ; dans le deuxième exemple, *type* n'accepte pas de synonyme parce qu'il a l'acception liée au domaine de la sécurité incendie.

Un défaut du modèle est sous-jacent à ces dysfonctionnements : l'association d'une pertinence à des formes (monosémiques ou polysémiques) et non à des entrées lexicales (monosémiques par définition).

4.5.2.5. Application des règles fondées sur la syntaxe

Des règles fondées sur la syntaxe constituent le noyau de la gamme LexiQuest et permettent, entre autres, de désambiguïser les mots de la phrase. Ces règles sont des heuristiques, elles concentrent le savoir-faire du constructeur et ne sont pas modifiables par l'utilisateur.
En particulier, la coordination est traitée par ces règles de telle manière, par exemple, que le *GN* : *handicapés et transports en commun* n'est pas analysé de la même manière que : *transports en commun et handicapés.* En effet, dans le premier cas, la coordination *et* est reconnue comme reliant les deux noms têtes *handicapés et transports ;* dans le deuxième *GN*, comme unissant les deux modifieurs *handicapés* et *en commun. Handicapés* est au masculin pluriel et donc compatible, en terme d'accord en genre et nombre, avec *transports*. Le deuxième *GN* permet alors d'écrire deux clauses de recherche : *transports en commun* et *transports handicapés.*
Selon les réseaux sémantiques activés par les règles mises en œuvre dans chaque cas, on peut aboutir à des réponses d'une part très différentes selon l'ordre des mots, d'autre part, tout à fait hors sujet pour une des clauses.
On peut conclure dans ce cas qu'il serait très utile, voire indispensable, que l'utilisateur soit informé et conscient du fait que des heuristiques (et non des règles "zéro erreur" qui de toutes façons n'existeraient pas) sont mises en œuvre pour désambiguïser les requêtes. Et la possibilité de visualiser les clauses de recherche produites par une question procède de cette idée puisqu'ainsi l'utilisateur peut comprendre les clauses obtenues et modifier sa requête en conséquence, même s'il ne peut accéder aux règles qui ont permis ces désambiguïsations.
D'autre part, vu les heuristiques de désambiguïsation appliquées, il paraît nécessaire que la notion de mot composé soit mieux prise en compte dans les règles de désambiguïsation et évidemment que les dictionnaires de mots composés soient complétés en fonction du domaine que l'on vise.

4.5.3. La gamme T-GID

4.5.3.1. Calcul de dépendance

La dépendance entre des mots se calcule à partir de la position relative de ces mots et du statut des mots qui les séparent (mot de liaison ou mot de rupture). Dans l'exemple suivant :

(5) *Quel doit être le niveau de sécurité incendie d'un local d'archives situé dans un ERP contigu à un parc de stationnement ?*

être n'est pas un mot de rupture, mais un mot vide, de même pour le déterminant *le* : le système va donc déclarer une dépendance entre *doit* et *niveau*, associer à cette paire un poids important et rechercher des paragraphes qui la contiennent, ce qui n'a aucun intérêt.

(26) *Combien de marches doit comporter une volée d'escalier dans un ERP ?*

En appliquant la même règle à (26), on identifie le triplet *marches-doit-comporter* comme pertinent et on lui associe un poids important par rapport à *volée*, *escalier* et *ERP* les autres mots identifiés dans la phrase, ce qui n'a, là non plus, aucun intérêt. L'identification de la paire (*volée*, *escalier*) ou du triplet (*volée*, *escalier*, *ERP*) serait beaucoup plus pertinente.

4.5.3.2. Dépendance et construction syntaxique

Le calcul de la dépendance entre deux mots tient compte uniquement de la proximité des mots dans la phrase indépendamment des contraintes imposées par la syntaxe. Par exemple, si l'on prend la préposition *de* : en tant que mot de liaison, elle va indiquer une dépendance entre les deux noms qui l'entourent et cette dépendance supplée effectivement à l'absence de mots composés comme *appareil de chauffage*, *point d'éclair* ou *trappe de visite*.
En revanche, cette représentation des relations entre les divers composants de la phase ne permet pas de prendre en compte les multiples emplois de *de* qui ne rapproche pas nécessairement les deux mots qui l'entourent. Par exemple :

- les verbes qui admettent un complément d'objet indirect introduit par *de*, par exemple *approcher* :
  dès qu'on approche du maximum de température

  les prépositions composées, qui contiennent *de* :

  à l'approche d'une flamme

  vont indiquer des dépendances inutiles *approche-maximum* et *approche-flamme* alors qu'elle est importante dans

  réserve d'approche

- de même, les dépendances *base-bois* et *base-unité* ou *base-unité-passage* n'ont pas intérêt dans :
  à base de bois
  sur la base d'une unité de passage

  alors que *essai-base* et *volume-base-atrium* ou *volume-base* sont essentielles dans :

  l'essai de base
  le volume de base de l'atrium

- les verbes construits avec deux compléments ainsi que les noms prédicatifs qui en dérivent et qui admettent comme modifieurs ces compléments du verbe :
  pour éviter que les flammes ou gaz chauds passent d'un étage à l'autre
  pour éviter le passage rapide des flammes ou des gaz chauds d'un étage à l'autre

  rendent possible la construction de la dépendance *gaz-chauds-étage* qui n'est pertinente.

Les dépendances inutiles ne font pas qu'introduire du bruit, elles peuvent aussi être cause de silence dans la mesure où elles sélectionnent comme appropriés des textes qui ne le sont pas, et donc peuvent repousser plus loin dans le classement des textes qui sont, eux, effectivement pertinents par rapport à la question posée.

4.5.3.3. Justification des réponses

Spirit propose des réponses classés par ordre de pertinence. Pour la question :

(34) Quelle doit être la surface utile des exutoires dans une galerie marchande de 2 000 m^2 ?

après traitements, la question se trouve résumée de la manière suivante :

surface utile des exutoires dans une galerie marchande

et les résultats s'affichent ainsi :

classe nbe documents expressions retrouvées

1	4	*<surface-utile-exutoires>*
2	1	*<surface-utile >, <galerie-marchande>, exutoires*
3	1	*<galerie-marchande>,surface*
4	6	*<surface-utile >, exutoires*
5	2	*surface, utile, galerie*

Cette présentation est très utile parce qu'elle permet à l'utilisateur de comprendre le classement des textes en fonction des mots dont il s'est servi dans sa question et donc de modifier les termes ou les tournures utilisés en conséquence. Néanmoins, cette progression par essai-erreur jusqu'à construire une liste d'expressions significatives satisfaisante, n'est possible que pour la recherche d'information mais ne peut s'appliquer au corpus, où une analyse fautive produite par les outils linguistiques, même quand elle est repérée, ne peut dans ce cas être rectifiée par une reformulation.

4.5.4. Commentaires non liés aux outils testés

Ces commentaires ne prétendent pas à l'exhaustivité sur les sujets abordés, mais ils se sont imposés après l'observation des résultats du test mis en place par le CSTB. Ils sont organisés en trois points : d'abord des confusions liées au vocabulaire, ensuite des problèmes associés à l'analyse des questions et enfin des questions posées par la représentation sémantique d'une phrase.

4.5.4.1. Difficultés liées aux variantes de vocabulaire

4.5.4.1.1. Le vocabulaire varie avec le locuteur

Les utilisateurs du service d'aide téléphonique du CATED sont des professionnels du bâtiment, mais on a vu que cette dénomination recouvre une multitude de métiers, des spécialistes qui appartiennent à un corps de métier, ou des généralistes comme des architectes. Dans ces contextes variés, les vocabulaires employés sont différents : un spécialiste pourra plus facilement utiliser un jargon de son métier, avec éventuellement des termes ambigus pour un généraliste, mais pas dans son domaine, alors qu'un architecte aura besoin, avec un seul document, et donc les mêmes termes pour tous, de se faire comprendre de tous ses interlocuteurs qui appartiennent pourtant à des spécialités différentes. Par exemple, dans la question suivante :

(32) *Quelles doivent être les caractéristiques de résistance au feu des gaines de désenfumage mécanique dans un ERP ?*

le mot *gaine* est impropre, mais ne pose pas de problème d'interprétation grâce au contexte. Le terme exact serait *conduit.*
Une solution consisterait à lier ces deux termes dans les dictionnaires de spécialité, par une relation d'association.

4.5.4.1.2. Le vocabulaire de spécialité est hiérarchisé

Comme on l'a vu précédemment, le vocabulaire technique admet peu de synonymes. En revanche, il est hiérarchisé et ce classement peut engendrer des différences notables de vocabulaire entre les questions et les textes du corpus. Par exemple, examinons l'organisation arborescente des noms de locaux dans la réglementation : on définit d'abord des établissements par le type d'activité qu'ils hébergent ; puis pour chaque type, il existe des noms d'établissement, puis dans chaque établissement, des noms de locaux qui sont spécifiés dans les textes par les risques spécifiques encourus à cause de ce qui est contenu dans ces locaux : les *établissements de type R* sont identifiés dans les textes à *établissements d'enseignement, colonies de vacances.*
Dans ce groupe, on peut ajouter aussi des établissements désignés par des noms comme *crèche, halte-garderie, jardin d'enfants, école maternelle, école primaire, collège, lycée, internat, classe de*

découverte, classe de nature, centre de loisirs, ...
Puis chacun de ces établissements peut abriter un *local à risques particuliers* : *cuisine, chaufferie, local de machinerie d'ascenseur, ...* eux-mêmes classés en *locaux à risques moyens* et *locaux à risques importants.*

Dans les questions qui suivent :

(15) *Faut-il que le plancher haut d'une cuisine collective d'un ERP de type R soit coupe-feu ?*
(17) *Quel est le degré coupe-feu d'un bloc-porte de machinerie d'ascenseur dans un hôtel ?*
(31) *Quel type de cloison doit-on installer entre une ensemble de grande cuisine et les autres locaux d'un établissement d'enseignement ?*

il faut, pour trouver des réponses pertinentes, donner des indications lexicales supplémentaires au moteur de recherche, en particulier, substituer *local à risque moyen* à *cuisine collective, grande cuisine* et *machinerie d'ascenseur.* De plus, d'après les réponses fournies par l'expert, les précisions concernant les types d'établissements sont, dans ce cas, non pertinentes parce que la réglementation concernant ces locaux s'applique quel que soit l'établissement dans lequel ceux-ci sont inclus.

Cette notion de hiérarchie du vocabulaire ne concerne pas que les noms de locaux. Dans les questions (12) et (38), on se heurte au même type de difficulté sur des thèmes différents :

(12) *Les clapets de ventilation d'un ERP de 1ère catégorie doivent-ils être pourvus de sondes ?*
(38) *Quelle est la réglementation concernant les sources centrales électriques de sécurité pour un hôtel de 1ère catégorie ?*

Cette fois-ci, il faut substituer *dispositif de détection* à *sonde* et *dispositif actionné de sécurité* (en abrégé *DAS*) à *clapet de ventilation.* Pour (38), une *source centrale électrique de sécurité* peut être soit un *groupe thermique générateur*, soit une *batterie d'accumulateurs* ; ensuite, la réglementation détaille, pour chacun de ces cas, les conditions d'installation, d'utilisation, etc.
Même si cette organisation du vocabulaire est indiquée dans les textes, une recherche vraiment automatique d'information dépasse le cadre linguistique et nécessite, pour retrouver les unités documentaires qui contiennent les réponses appropriées, que les systèmes informatiques utilisés aient des capacités d'inférence, ce qui n'est pas pour le moment. Cette situation a pu être observée plusieurs fois dans le cadre de cette étude.

4.5.4.1.3. Liens entre les mots utilisés et le concept exprimé

Les mots utilisés pour formuler la question sont différents de ceux que l'on trouve dans la réponse non pas seulement en termes de vocabulaire mais aussi parce qu'ils font implicitement référence à des objets ou des notions différents de ceux du corpus, et qui donc s'expriment avec des termes différents. Les exemples suivants illustrent cette difficulté :

(5) *Quel doit être le niveau de sécurité incendie d'un local d'archives situé dans un ERP contigu à un parc de stationnement ?*

Pour des locaux contigus, l'objectif est qu'un feu déclaré dans un des locaux ne puisse se propager rapidement dans l'autre. Le point de réglementation évoqué dans ce cas va donc s'exprimer en termes d'isolement latéral constitué par la paroi située entre les locaux, et dont on va exiger certaines caractéristiques de stabilité au feu. De plus, le libellé *local d'archives* renvoie à la dénomination *local à risques particuliers* et le type de l'ERP va influer sur le classement de ce local d'archives en *local à risques moyens ou local à risques importants.*

(31) *Quel type de cloison doit-on installer entre un ensemble de grande cuisine et les autres locaux d'un établissement d'enseignement ?*

L'expression *type de cloison* renvoie à la notion de stabilité au feu d'une paroi, et donc aux caractéristiques *pare-flammes* ou *coupe-feu*. Dans cette question s'ajoutent des problèmes que l'on a déjà vus précédemment : il faut identifier *établissement d'enseignement* et *établissement de type R*, puis *ensemble de grande cuisine* comme un *local à risques particuliers*. Comme pour l'exemple précédent, la gravité des risques (moyens ou importants) est ensuite déterminée en fonction du type de l'établissement.

(39) a. Quelles sont les mesures d'isolement entre un ERP de la 5ème catégorie et un tiers ?
(39) Quelles sont les mesures d'isolement entre un ERP de la 5ème catégorie et un commerce ?

Le mot composé *mesures d'isolement* désigne des caractéristiques liées cette fois encore à la stabilité au feu, essentiellement le coupe-feu.
La réglementation pour la sécurité incendie a pour objectif de protéger les personnes ; une des manières de réaliser cet objectif est de confiner le feu, le plus longtemps possible, dans le local où il s'est déclaré. Dans cet esprit, la réglementation désigne par *tiers* tout local non soumis aux contraintes existant pour les ERP mais qui peut être atteint par un incendie ayant débuté dans un ERP. On peut ainsi, dans les textes, désigner par *tiers*, selon les cas, un bâtiment d'habitation, un commerce, une aire de stationnement, ... Si on pose la même question sous la forme (39a), il faudra, pour trouver des réponses pertinentes, faire un lien lexical entre *commerce* et *tiers*, lien qui n'est pertinent que dans ce contexte. En revanche, les expressions *magasins de vente, centres commerciaux* qui figurent dans l'article donnant le classement des établissements ne doivent pas être remplacées par *tiers*, mais par *établissement de type M*.

(49) a. Comment doit être assurée la surveillance d'un ERP de type L (salle de conférence) de 1ère catégorie (1 600 personnes) ?

La réponse donnée par le CATED fait référence aux articles MS 45 à MS 48 de la réglementation. Le contenu de l'article MS 45 est le suivant : *La surveillance des établissements doit être assurée pendant la présence du public*. Puis immédiatement l'article MS 46 débute : *Le service de sécurité incendie doit être assuré suivant* ... Dans ce contexte, la synonymie entre *surveillance et service de sécurité incendie* est réelle mais implicite, et sans le recours à une réécriture manuelle de la question, les articles pertinents sont difficiles à identifier.

4.5.4.1.4. Des outils spécifiques sont nécessaires pour exprimer ces variations de mots

L'identification des paragraphes pertinents par rapport à une question, repose essentiellement sur le rapprochement entre les mots employés dans la question et ceux figurant dans le corpus. Dans ces conditions, il est nécessaire d'adopter des méthodes spécifiques afin d'améliorer ce rapprochement. On peut ranger dans cette catégorie, la construction de dictionnaires contenant les mots composés spécifiques au domaine ainsi que l'utilisation de grammaires locales permettant de reconnaître des séquences et de les reformuler en d'autres termes plus canoniques.

4.5.4.1.4.1. Dictionnaires de mots composés

La nécessité de la prise en compte des noms composés se présente comme une évidence. Elle peut se faire à travers des dictionnaires ou des grammaires locales en particulier pour tenir compte de la combinatoire des constituants des expressions techniques, souvent complexes. Les exemples sont nombreux :

- *canton* est employé dans le corpus avec deux contextes distincts :
  . d'un point de vue administratif, il désigne une division territoriale de l'arrondissement ;

. dans le domaine de la sécurité incendie, l'expression complète est un nom composé *canton de désenfumage*, souvent utilisé avec le seul nom tête *canton* .

- *catégorie* a des acceptions multiples dans le corpus et les modifieurs qui l'accompagnent varient alors selon les contextes :

  . pour indiquer la catégorie d'un établissement, les modifieurs sont des numéraux, cardinaux ou ordinaux :

  les établissements de (5ème catégorie + catégorie 5)

  . la catégorie d'un matériau est comprise entre M0 et M4 :

  un matériau M0

  . utilisés pour désigner la catégorie d'un système de sécurité incendie, les modifieurs sont classés de A à E :

  un système de sécurité incendie classé A est exigé

4.5.4.1.4.2. Grammaires locales pour les notions propres à la sécurité incendie

La différence entre le vocabulaire de la question et celui des réponses s'illustre d'une manière prévisible pour des thèmes propres à la sécurité incendie, de telle sorte qu'il paraît nécessaire de prévoir des grammaires locales détaillant l'expression de ces concepts, en particulier la réaction au feu, question (30), et la résistance au feu, exemples (27) et (32) :

(30) *Quelle doit être la réaction au feu des isolations intérieures dans un hôpital de 1ère catégorie ?*
(27) *Quelle doit être la résistance au feu des conduits de diamètre inférieur ou égal à 125 ?*
(32) *Quelles doivent être les caractéristiques de résistance au feu des gaines de désenfumage mécanique dans un ERP ?*

En effet, la réponse à une question portant sur la réaction au feu d'un élément de décoration ou de mobilier, d'un revêtement, etc, concerne les matériaux utilisés pour fabriquer cet élément et la réaction au feu s'exprime sous forme d'une catégorie associée au matériau :

matériau de catégorie (M0 + M1 + M2 +M3 +M4)

Les questions qui ont trait à la résistance au feu s'appliquent à des éléments de construction du bâtiment et les réponses s'expriment sous différentes formes selon le niveau de sécurité exigé. Les notions mises en jeu sont celles de *stabilité au feu, pare-flammes, coupe-feu* :

les conduits de diamètre nominal supérieur à 75 mm et inférieur ou égal à 315 mm doivent être pare-flammes de traversée 30mn ...
Les conduits doivent être réalisés en matériaux incombustibles et être SF de degré 1/4h.

4.5.4.2. Analyse des questions

Les techniques utilisées pour analyser les questions sont les mêmes que celles mises en œuvre pour traiter le corpus, ce qui peut paraître comme un gage de cohérence des réponses. De fait, aux difficultés déjà rencontrées dans l'analyse du corpus s'ajoutent des problèmes résultant de la forme des questions, qui doivent adopter une tournure syntaxique imposée par la langue.

4.5.4.2.1. La forme interrogative des questions influe sur l'utilisation de mots grammaticaux

Les mots qui reviennent souvent dans les questions n'ont qu'une valeur formelle liée à la construction au mode interrogatif des phrases en français et sont non significatifs :

(3) *Quelle est la réglementation incendie applicable pour* ***le parc de stationnement à l'air libre d'un gymnase de 4ème catégorie*** *?*

(35) *Quelle est la réglementation applicable au* ***parc de stationnement annexe d'un hôtel*** *?*

(38) *Quelle est la réglementation concernant* ***les sources centrales électriques de sécurité pour un hôtel de 1ère catégorie*** *?*

(2) *Quelles doivent être les caractéristiques des* ***cloisons entre les salles de classe et les circulations d'un lycée*** *?*

(5) *Quel doit être le niveau de sécurité incendie d'un* ***local d'archives situé dans un ERP contigu à un parc de stationnement*** *?*

Les *GN* pertinents pour interroger la documentation par l'intermédiaire du moteur de recherche sont imprimés en gras. En plus des problèmes lexicaux ou sémantiques liés au vocabulaire, il est nécessaire de pré-traiter les questions pour reconnaître ces formes interrogatives qui n'ont aucun intérêt dans la recherche d'information.

4.5.4.2.2. Les mots inutiles de la question introduisent du bruit dans les réponses

Les spécialistes connaissent les principes généraux, en particulier type et catégorie, qui sous-tendent la réglementation et orientent systématiquement la question en conséquence. Mais parfois ces précisions sont inutiles parce que les textes s'appliquent pour tous les établissement ou quelle que soit la catégorie. Dans ce cas, les réponses sont bruitées parce qu'elles sont encombrées avec des informations généralement importantes dans le contexte de la sécurité incendie, mais dans le cas précis de la question, inutiles. Par exemple, pour les questions suivantes :

(3) *Quelle est la réglementation incendie applicable pour le parc de stationnement à l'air libre d'un gymnase de 4ème catégorie ?*

(17) *Quel est le degré coupe-feu d'un bloc-porte de machinerie d'ascenseur dans un hôtel ?*

Le fait que l'ERP soit un gymnase dans (3) ou un hôtel dans (17) n'a aucune influence sur la réponse. Au contraire, à cause de cette précision, des paragraphes verront leur pertinence augmenter parce qu'ils contiennent le mot *hôtel*, alors qu'ils n'ont aucun intérêt par rapport à la question posée, et vont parvenir ainsi à devancer des paragraphes plus pertinents mais qui ne font pas référence à *hôtel*.

4.5.4.2.3. Les mots effacés dans la question ou dans le corpus modifient la pertinence des réponses

Les questions posées sont formulées par des experts, elles sont souvent elliptiques du contexte, ce qui peut se traduire d'un point de vue lexical par l'utilisation de *GN* tronqués dans lesquels il ne subsiste plus que le nom tête et où les modifieurs ont été effacés. Et dans ce cas, la recherche automatique de documents pertinents sera rendue plus difficile puisqu'on trouvera toujours "plus de mots" dans le corpus que dans la question.

A l'inverse, l'effacement des termes peut aussi avoir une influence sur le calcul de la pertinence d'un document quand l'auteur du texte a voulu alléger son style et n'utilise plus que la tête du nom composé dans le corps du texte alors qu'il a spécifié le nom complet dans le titre. Puisque la pertinence associée à un nom composé est plus importante que celle accordée au nom tête seul ou à un modifieur isolé, un texte qui contient seulement le nom tête sera jugé peu pertinent par rapport au texte d'une question où aucun composant n'a été effacé, et un texte qui spécifie un nom complet sera considéré coome trop spécifique par rapport à une question dans laquelle un ou plusieurs composants ont été effacés.

4.5.4.3. La représentation sémantique d'une phrase n'est pas réductible à l'étude de son vocabulaire ajoutée à celle de sa syntaxe

Il est nécessaire de développer les outils linguistiques qui permettent d'améliorer le traitement des questions et l'analyse du corpus, grâce à une meilleure étude du vocabulaire et de la syntaxe des phrases. Néanmoins, même si les phrases produites à l'issue de ces opérations étaient découpées de manière parfaitement correcte, ne comportaient aucune erreur d'étiquetage syntaxique, ni ambiguïté lexicale, l'interprétation sémantique poserait encore des problèmes tels que la recherche dans un corpus de référence des réponses à une question ne pourrait donner des résultats pleinement satisfaisants. Nous allons développer cette remarque autour de thèmes spécifiques au test mis en place ou à la sécurité incendie, comme la reformulation initiale des questions et le calcul de la catégorie d'un établissement, et de thèmes plus généraux comme la prise en compte des nombres ou la sélection de documents quand la réponse à la question ne figure pas dans le corpus.

4.5.4.3.1. Reformulation initiale des questions

Les questions telles qu'elles ont été posées aux différents systèmes testés (et telles qu'elles sont rapportées en annexe) n'ont rien à voir avec les demandes qui ont été formulées par les usagers du service d'assistance téléphonique. Elles sont le fruit d'un échange entre ce professionnel du bâtiment qui a besoin d'une information et un expert du domaine concerné par la question. Cette maïeutique préliminaire constitue une étape essentielle parce qu'elle filtre la demande initiale de l'usager à travers la connaissance de l'expert afin d'en supprimer les aspects non pertinents et y ajouter, grâce à un dialogue, les éléments qui manquent et qui sont nécessaires pour pouvoir appliquer la réglementation. La demande initiale de même que les différentes formulations proposées à l'abonné avant d'aboutir à la formulation de la question finalement jugée convenable par l'expert du CATED, ne sont pas disponibles dans les archives du CATED. Elles n'ont pu donc être étudiées.
Mais dans tous les cas, ce travail essentiel ne se limite pas à l'application de techniques linguistiques. Il ne fait pas non plus l'objet de cette étude.

4.5.4.3.2. Calcul de la catégorie d'un établissement

La réglementation repose sur les notions de type d'activité et de catégorie de l'établissement ; cette dernière est calculée exclusivement à partir de l'effectif de public admis dans l'établissement et de l'activité abritée par le bâtiment. Dans la documentation, le mode de calcul de la catégorie est donné dans un des tout premiers textes et ensuite il n'est plus fait référence à cet effectif qu'à travers des expressions concernant la catégorie de l'établissement. Or, dans les questions posées par les utilisateurs, l'activité et la catégorie sont le plus souvent données "en clair" : c'est l'expert interrogé qui transforme ensuite les indications textuelles en caractéristiques "codées" par rapport au domaine s'il doit effectivement faire une recherche (manuelle actuellement) dans la documentation.
Si l'on veut automatiser le système de collecte d'informations, il faut ou bien avertir l'utilisateur et le former afin qu'il fasse lui-même et explicitement la conversation, ou bien utiliser des filtres qui reconnaissent et transforment les informations concernant l'activité abritée par l'établissement et l'effectif maximal de public admissible, en un type et une catégorie d'établissement. Ainsi, les deux exemples suivants, malgré des libellés différents, sont équivalents et doivent produire les mêmes réponses :

(49) *Comment doit être assurée la surveillance d'une salle de conférence d'une capacité de 1 600 personnes ?*

(49) b. *Comment doit être assurée la surveillance d'un ERP de type L 1ère catégorie ?*

Une autre difficulté subsiste dans le libellé de la catégorie. Comme on l'a vu précédemment, la réglementation se présente sous la forme de livres :

livre 1er : Dispositions applicables à tous les ERP

livre II : Dispositions applicables aux ERP des quatre premières catégories
livre III : Dispositions applicables aux établissements de cinquième catégorie
livre IV : Dispositions applicables aux établissements spéciaux

Il est donc nécessaire de disposer d'outils pour transformer l'expression *ERP de type L 1ère catégorie* de telle manière que les dispositions des livres I et II soient applicables, mais pas celles des livres III ni IV.

4.5.4.3.3. Interprétation des nombres

La réglementation utilise souvent des nombres : pour déterminer les catégories, donner les dimensions limites des éléments de construction, indiquer des quantités mesurables, ...
Pourtant, les suites de caractères numériques sont considérées, par les analyseurs, comme des chaînes de caractères et non comme des nombres. La relation d'ordre qui existe entre les nombres pour un locuteur humain ne peut donc pas s'appliquer à ces chaînes de caractères et la notion de comparaison numérique est absente.

(34) *Quelle doit être la surface utile des exutoires dans une galerie marchande de 2 000 m^2 ?*
(27) *Quelle doit être la résistance au feu des conduits de diamètre inférieur ou égal à 125 mm ?*

Dans le corpus figure l'expression *locaux de surface supérieure à 1 000 mètres carrés* mais aucun des moteurs testés ne peut rapprocher, sur des critères de comparaison numérique, ce libellé de celui de la question (34) *galerie marchande de 2 000 m^2*.
De même, pour la question (27), des réponses correctes (i.e. des articles pertinents par rapport à la question posée) sont trouvées par les moteurs de recherche parce que figurent dans le corpus les expressions *conduits de diamètre nominal inférieur ou égal à 125 mm* mais quand la réponse se subdivise en *diamètre nominal supérieur à 75 mm et inférieur ou égal à 315 mm* et les autres cas, les alinéas pertinents n'apparaissent pas dans la liste des articles proposés comme réponses.

4.5.4.3.4. La réponse est négative ou ne figure pas dans le corpus

La recherche des réponses à une question posée se fait par le rapprochement des mots contenus dans la question avec le contenu des paragraphes du corpus en tenant compte des associations de mots, des expressions composées identifiées, du poids de certains termes, etc ; elle repose donc sur une similitude de mots. Quand la réponse à la question posée est négative, comme dans les exemples suivants :

(46) *Est-ce que tous les ERP doivent être vérifiés par des personnes ou des organismes agréés ?*
(47) a. *Est-ce qu'un écran de cantonnement peut être en verre ?*

plusieurs cas de figures peuvent se présenter :

- dans le cas le plus favorable, les mots de la question figurent dans le corpus, mais avec une négation. Le moteur de recherche peut donc appliquer ses méthodes d'investigation habituelles et l'unité documentaire concernée pourra être repérée et proposée comme réponse ;

mais il peut aussi être difficile de trouver la réponse pour plusieurs types de raisons :

- on ne fait pas allusion dans le corpus au problème posé dans la question, c'est à dire qu'on ne peut pas trouver une unité documentaire contenant la totalité des termes contenus dans la demande initiale. Les réponses présentées seront alors non pertinentes parce qu'elles ne contiendront qu'une partie des mots de la question ;
- on trouve, dans le corpus, une partie des mots de la question en association avec d'autres termes que ceux de la question mais qui pourraient, pour le domaine de spécialité considéré, appartenir au "même groupe sémantique" (même si ces groupes n'ont d'existence

qu'intuitive). Ce serait le cas pour (47a). En effet, la réponse apportée par l'expert du CATED s'appuie sur une instruction technique formulée ainsi :

Un écran de cantonnement est constitué :

. *soit par des parois en matériaux incombustibles et SF de degré 1/4h ;*
. *soit par des éléments de structure ;*
. *soit par tout autre dispositif ayant fait l'objet d'un avis favorable de la commission centrale de sécurité.*

et la réponse est donc *oui, à condition que l'écran respecte une SF 1/4 h.* Mais il n'y a aucune chance pour qu'un moteur de recherche identifie comme pertinent ce paragraphe. En revanche quand la question est écrite sous la forme suivante:

(47) *Qu'est-ce qu'un écran de cantonnement ?*

l'article approprié de l'instruction technique est repéré dans presque tous les scénarios.

- le paragraphe qui contient les mots de la question est inclus dans un document plus long qui comporte une phrase d'en-tête destinée à indiquer la portée des articles qui suivent. Dans ce cas, quand un article est pertinent pour une réponse, il faut vérifier qu'il est effectivement valide pour le cas dans lequel on se trouve. Par exemple, dans le corpus documentaire applicable aux ERP, la deuxième partie s'intitule *Règlement de sécurité* et comporte deux livres : le premier *Dispositions applicables à tous les établissements recevant du public*, le deuxième *Dispositions applicables aux établissements des quatre premières catégories*. Les questions de sécurité incendie se posent dans les mêmes termes quelle que soit la catégorie, il est donc tout à fait envisageable qu'un paragraphe paraisse pertinent par rapport à une question concernant un établissement de la cinquième catégorie alors que ce paragraphe appartient au livre II de la documentation. Dans ce cas, ou bien chaque unité documentaire contient une référence aux différents découpages logiques du texte et la pertinence d'un paragraphe doit être jugée aussi, voire en priorité, par rapport à ce titre, ou bien ce problème doit avoir été expliqué en préalable à tout utilisateur de l'outil automatique afin qu'il en tienne compte lors de l'examen des réponses jugées pertinentes par le moteur de recherche.

Enfin, tout comme le processus déductif des systèmes experts, la méthode de recherche de l'information essentiellement fondée, après l'application de procédés lexicaux ou syntaxiques, sur la coïncidence de chaînes de caractères n'a souvent rien à voir avec le mode de raisonnement et donc le procédé de recherche de l'information dans le corpus mis en œuvre par l'expert.

4.6. Conclusions

L'objectif de cette étude consistait à évaluer l'intérêt d'ajouter des ressources linguistiques (en l'occurrence expertise syntaxique, règles de désambiguïsation, terminologie spécifique) dans un système destiné à rechercher automatiquement de l'information dans un corpus, afin d'améliorer les résultats de ce système. Si l'on étudie le taux de rappel, il est flagrant que ces nouveaux outils, quand ils sont intégrés dans la chaîne de traitement avant le travail exécuté par le moteur de recherche, augmentent le nombre de bonnes réponses. mais celles-ci sont le plus souvent égarées au milieu des réponses non pertinentes comme en témoigne le taux de précision des réponses. Et surtout, le rappel étant dans la majorité des cas inférieur à 50%, les documents proposés comme réponses représentent moins de la moitié des documents attendus.

Une autre conclusion que l'on peut tirer de cette étude et de son travail préparatoire est que les outils linguistiques et informatiques proposés sur le marché mettent à profit une des deux approches existantes ou les combinent selon les cas pour retrouver ou extraire de l'information d'un corpus :

- la première consiste à opérer sur le corpus sans connaissance préalable du type des données textuelles disponibles, mais sur une base statistique ;

- la deuxième utilise des connaissances préalables, connaissances terminologiques sur le domaine technique du corpus, et linguistiques sur le fonctionnement général de la langue.

Par exemple, la gamme T-GID met en œuvre des connaissances linguistiques générales sur la langue pour établir des règles de désambiguïsation, des connaissances lexicales rassemblées des dictionnaires, mais n'utilise pas la notion de mot composé qui est simulée par des considérations statistiques pour l'établissement de dépendances entre mots du texte. Les produits lexiQuest, au contraire, concentrent leurs efforts sur les aspects linguistiques, et plus particulièrement lexicaux, en utilisant des dictionnaires de mots simples et de mots composés, des règles de désambiguïsation, de segmentation, etc.
On peut donc remarquer qu'il n'existe pas de méthode consensuelle et efficace pour, à partir des mots qui le composent, analyser un texte, c'est-à-dire, pour les aspects qui nous préoccupent, en extraire de l'information. On observe plutôt une multitude de techniques qui sont évaluées, testées et ajustées par ceux qui les mettent en œuvre en fonction des résultats obtenus dans un domaine ou un autre.

Le problème essentiel consiste donc à produire une analyse sémantique d'un texte mais, comme le montre le test réalisé, l'ajout de connaissances lexico-sémantiques nouvelles (sous forme de dictionnaires, dictionnaires sémantiques, réseaux sémantiques, etc) ne constitue pas une solution réellement efficace pour pallier les difficultés rencontrées lorsqu'elles s'appuient sur des analyses morphologique, lexicale et syntaxique insuffisantes voire défaillantes, car ces connaissances produisent du bruit qui compense le bénéfice qu'elles procurent.

L'intérêt d'outils informatiques et linguistiques pour rechercher automatiquement de l'information à l'intérieur d'un corpus consiste donc à examiner en fonction des contraintes de mise en service qui les accompagnent : profil des utilisateurs, niveau de satisfaction attendu, portée et utilisation des réponses, ...
Dans certains contextes, en particulier pour ceux qui sont concernés par l'application de textes réglementaires, le bruit mais surtout le silence observés dans les réponses fournies par des systèmes automatiques peut avoir des conséquences désastreuses. En revanche, considérer ces applications comme des outils supplémentaires mis à la disposition d'experts du domaine peut constituer une alternative intéressante à la création d'applications tout public. En effet, ces experts, utilisant ces nouveaux produits pour retrouver une information qu'ils connaissent déjà, maîtrisent les principes techniques, réglementaires, ... du domaine concerné, ses concepts et son vocabulaire ; ils peuvent évaluer la pertinence des réponses obtenues et vérifier que l'information retrouvée lors d'une recherche automatique n'est pas gâtée par des silences dommageables.

Il est irréaliste d'espérer traiter la sémantique d'un texte en faisant l'économie d'analyses lexicale et syntaxique rigoureuses. Les risques que l'on court se matérialisent sous la forme de contresens, de bruit dans les réponses mais aussi de silence, ce qui pour l'application de textes réglementaires n'est pas acceptable.

Enfin, nous avons montré que la partie la plus importante du travail est consacrée au traitement des questions et aux moyens de rapprocher les informations contenues dans une question avec celles provenant du corpus documentaire, afin de trouver les réponses les plus pertinentes à la question posée. Ce traitement est essentiellement sémantique mais il s'appuie sur des analyses morphologiques et syntaxiques qui sont ambiguës voire défaillantes. Cette méthode de travail peut s'avérer dangereuse.

Etude du corpus

L'objectif d'améliorer la recherche automatique d'informations dans un corpus relevant de la sécurité incendie se réalise aussi par la mise en place de dictionnaires de spécialité. L'étude du corpus grâce à un outil comme INTEX détecte peu de mots inconnus. On en déduit donc que les mots caractéristiques de ce domaine, ne sont pas des créations mais plutôt des emprunts au vocabulaire courant, employés avec des acceptions particulières ou précisés par des modifieurs ou des contextes spécifiques. La recherche des mots caractéristiques du domaine se traduit par la mise en évidence de séquences composées plus ou moins figées. On dispose pour cela d'outils linguistiques qui utilisent des critères morphologiques et syntaxiques. On s'intéresse aussi aux mots simples. Les outils cités précédemment sont alors inefficaces pour mettre en évidence ces mots simples ; on essaiera de les repérer en utilisant plutôt des critères numériques.

1. Les mots inconnus

La taille du corpus concernant la sécurité incendie est d'à peu près 2 Mo. Ces textes couvrent des aspects législatifs, réglementaires et techniques. Quand on applique les dictionnaires du système DELA à ces textes, le nombre de mots inconnus est faible : 308 mots sur les 413 863 (dont 8703 différents) que comporte le texte.

	taille du corpus	nombre de mots différents	nombre de mots inconnus
texte "sécurité incendie"	2 Mo	8 703	308
Le Monde	3 Mo	37 216	6950

Si l'on compare le taux de mots inconnus selon les corpus, la couverture des dictionnaires paraît tout à fait satisfaisante pour le corpus spécialisé ; en revanche, si l'on veut analyser les textes d'un journal comme *Le Monde*, il faut sûrement l'améliorer, en particulier pour les noms propres qui constituent l'essentiel des mots inconnus : 3.5% de mots inconnus dans les corpus technique, 19% dans le quotidien.

1.1. Fautes de frappe et fautes d'accord

échantilons
depression
détâchement
dévêtissement
dizièmes
Danmark

Les mots mal orthographiés et qui ne peuvent pas être rattachés à un lemme sont rejetés sous la forme de mots inconnus.

les équipements nécessaire à la diffusion de ce message*
les salles polyvalentes à dominante sportive dont l'air d'activité est supérieure ou égale à 1 200 m2*

Pour ces deux derniers exemples, les fautes ne conduisent pas à refuser des mots puisque ceux-ci, même fautifs, sont trouvés dans les dictionnaires *DELAF*. Ainsi, ces fautes ne peuvent être détectées par simple consultation des dictionnaires. Néanmoins, après examen d'une partie des textes, elles devraient être assez peu nombreuses pour ne pas poser d'autre problème que celui de la correction de la langue, et en particulier ne pas fausser les comptages des occurrences.

1.2. Acronymes

On trouve dans la réglementation des acronymes français ou étrangers passés dans le langage courant :

ASCII,American Standard Code for Information Interchange.A:ms:mp:fs:fp
HT/BT,haute tension-basse tension.N:ms:mp:fs:fp

Les acronymes permettent aussi de désigner des organismes dont l'activité est liée au bâtiment ou à la sécurité incendie. La forme développée est toujours employée la première fois, accompagnée de l'acronyme qui est ensuite utilisé seul :

attribué par l'Association Française de Normalisation (AFNOR)
application des DTU ou des normes expérimentales AFNOR

Voici un extrait du dictionnaire qui permet de relier l'acronyme à sa forme développée :

ASTM,American Society for Testing and Materials.A:ms:mp:fs:fp
ATG,Association Technique de l'industrie du Gaz.N:fs
ACERMI,Association pour la Certification des Matériaux Isolants.N:fs
AIA,Association des Ingénieurs-Architectes.N:fs
CECMI,Comité d'Etude et de Classification des Matériaux et éléments de Construction par rapport au danger d'incendie.N:ms
CERFA,Centre d'Enregistrement et de Révision des Formulaires Administratifs.N:ms

La sécurité incendie est une activité issue du domaine du bâtiment. Les auteurs des textes considèrent comme connus des termes spécifiques à ce domaine :

BAEL,béton armé aux états limites.N:ms:mp
BPEL,béton précontraint aux états limites.N:ms:mp
HEA,poutrelle en h à ailes larges.N:ms:mp
IPE,profilé d'acier en forme de I.N:ms:mp
IPN,profilé d'acier en forme de N.N:ms:mp

La réglementation définit des acronymes pour désigner des objets spécifiques à la sécurité incendie. Dans les textes figurent aussi bien l'acronyme que la forme développée du nom composé.

détecteur autonome déclencheur,DAD.N:ms
dispositif actionné de sécurité,DAS.N:ms
dispositif adaptateur de commande,DAC.N:ms
système de détection incendie,SDI.N:ms
système de mise en sécurité incendie,SMSI.N:ms
système de sécurité incendie,SSI.N:ms
alimentation électrique de sécurité,AES.N:fs

1.3. Noms propres

bombe de Parr
cône d'Abrams
méthode de Dumas
tube de Pitot (double+E)

feu de Bengale

La réglementation utilise des termes techniques ou non, de patron syntaxique *N de N-propre*. Ces termes sont des noms composés dont le deuxième nom est un nom propre et le nom tête, un nom courant appartenant déjà aux dictionnaires de mots simples. Les séquences complètes constituent des entrées du dictionnaire de noms composés spécifique au domaine de la sécurité incendie au même titre que les noms communs composés.
Le composé *feu de Bengale* fait lui partie du langage courant.

1.4. Mots étrangers

Instituto Eduardo Torroja de Ciencias de la Construccion (IETCC), c Serrano Galvache s n - Costillares-Chamartin, E-28033 Madrid (Espagne)
SITAC, Swedish Institute for Technical, Approval in Construction, Svenskt ByggodkÌnnande AB, Box 553, S-371 23 Karlskrona (Suède).

Les textes de spécialité contiennent des mots étrangers (noms communs, noms propres, prépositions, ...) essentiellement parce qu'ils font référence à des organismes étrangers dont ils indiquent les noms et adresses dans la langue d'origine ou en anglais. Il n'y a donc pas d'information utilisable dans des dictionnaires bilingues sous la forme de traduction de termes techniques français en leurs homologues en d'autres langues.
En revanche, il peut y avoir l'emprunt direct d'un mot étranger, qui devient alors un terme de la langue et figure donc dans le dictionnaire de spécialité : c'est le cas de *sprinkler* (voir 1.7.b).

1.5. Noms d'unités

kN
kPa
daN m
daN mètre linéaire

Les parties techniques et même celles réglementaires font référence à des calculs dans lesquels les quantités sont accompagnées de leur unité. Celles-ci sont, le plus souvent, notées dans le système international de mesures mais des auteurs utilisent aussi des unités ou des notations d'unités consacrées par l'expérience mais dont le nom n'obéit pas aux règles de construction du système international. Nous avons détaillé ces unités dans un transducteur qui contient les noms du système international comme les noms usuels. Ce transducteur constitue un dictionnaire, présenté en annexe.

1.6. Symboles

Dans les textes réglementaires, les références aux articles s'expriment par un nom plus un code. Par exemple, les généralités sont traitées dans les articles *GE 1* à *GE 9* ; ce qui concerne l'éclairage se

trouve dans les articles *EC 1* à *EC 21*. Ces symboles peuvent être considérés comme des entrées du dictionnaire car les experts s'en servent pour désigner des séries d'articles.

Les types d'établissements sont aussi notés par un code : lettres *L* à *P* et *R* à *Y* pour les codes qui concernent les établissements installés dans un bâtiment, et code à deux ou trois lettres pour les établissements spéciaux : *PA*, *CTS*, *SG* ... Ces codes contiennent une information très souvent utilisée dans les textes et dans les questions des utilisateurs sur le corpus. On choisit donc de traiter ces symboles comme des mots ; l'équivalence entre code et libellés d'établissement se fait grâce à un transducteur.

1.7. Mots nouveaux

Après avoir éliminé les fautes de frappe, traité les mots étrangers, les acronymes, les noms d'unités et les symboles, on peut considérer que les mots inconnus sont bien des mots nouveaux. Parmi ceux-là, on peut encore distinguer plusieurs cas :

a) les mots nouveaux qui ne sont que des variantes orthographiques d'un mot déjà existant. Le concept, l'objet, ... sont désignés par un mot qui subit dans les textes que l'on examine une variante orthographique :

bouchepore,A
bouche-pores,A

b) les mots nouveaux ou étrangers qui apparaissent directement dans les textes, avec une ou plusieurs orthographes (selon les auteurs), et accompagnés ou non de leur traduction française :

sprinkler,N
sprinkleur,N
unité de toiture monobloc (roof-top)

c) les mots nouveaux formés par assemblage de mots et de préfixes existants et sans trait d'union :

cylindroconique,A
thermovélocimétrique,A

Cylindro-conique et *cylindro-ogival* existent, la forme soudée est une création. De même, *thermo-déformable*, *thermo-phonique* sont des formes répertoriées dans les dictionnaires mais pas les formes soudées. Avec le premier mot *vélo*, les mots soudés existent *vélocifère*, *vélocipède*, ... *Thermovélocimétrique* est une création par soudure de trois mots existants : *thermo*, *véloci* et *métrique*. Il adopte l'étiquette syntaxique du dernier composant *métrique* : adjectif.

On peut aussi classer dans cette catégorie des mots nouveaux qui sont créés à partir de formes existantes et directement avec plusieurs orthographes. C'est le cas de *coupe-feu* qui admet aussi les graphies *coupe feu* (sans trait d'union) et *CF*.

d) les mots nouveaux formés selon des schémas morphologiques productifs, à partir de racines identifiées :

désenfumé,A
désenfumage,N

réallumage,N
extrudage,N
exigentiel,A
filetable,A

Les créations correspondent à des opérations bien connues : ajout du préfixe privatif *dé-* (qui se transforme en *dés-* devant une voyelle), nominalisation des verbes par le suffixe *-age* pour exprimer l'objet de l'action, adjectivation des verbes par les suffixe *-el*, adjectivation des verbes transitifs pour indiquer la possibilité de faire l'action exprimée par le verbe grâce au suffixe *-able*.
Le sens de ces nouveaux termes est déductible de celui des mots-racines.

e) certains mots sont des néologismes qui permettent de préciser le vocabulaire : le vocabulaire technique a parfois besoin de distinguer une action de son résultat, alors que le mot existant désigne les deux aspects dans la langue courante : *mesurage* (création pour le procès *mesurer*) figure dans le corpus technique alors que *mesure* (à la fois pour le résultat du procès et le procès *mesurer*) existe et n'est plus utilisé que pour le résultat numérique du procès.

f) en revanche, certaines créations doublent simplement un mot déjà existant dans la langue :

majorateur alors que *majorant* existe
minorateur alors que *minorant* existe

g) les mots que nous ne savons pas interpréter ni rattacher à une racine connue par un processus de dérivation :

gradinage,N
pareclose,N
tavaillon,N

Ce sont le plus souvent des mots désignant des objets du bâtiment, ou plus rarement de la sécurité incendie ; ils portent en principe l'étiquette nom, mais on constate aussi des emplois d'adjectifs. Ils sont recensés dans des dictionnaires spécialisés du bâtiment comme le DICOBAT.
Néanmoins, l'élaboration d'un dictionnaire du bâtiment utilisable par la totalité des corps de métiers paraît être une tâche considérable et difficile à cerner à cause du nombre important d'objets manipulés et donc de mots dans chaque métier, et aussi parce qu'un même objet ou une même tâche peut être désigné, selon le corps de métier, par des mots différents.

L'ensemble des mots inconnus, et donc non étiquetés par INTEX à sa première analyse, a été examiné. Les noms propres non reconnus n'ont pas tous été entrés dans les dictionnaires adéquats, essentiellement parce qu'ils ne sont pas recherchés dans les patrons syntaxiques que l'on a construits. Les nouvelles entrées des dictionnaires figurent en annexe.

Les dictionnaires généralistes du système DELA détectent peu de mots inconnus, pourtant il est peu vraisemblable que le domaine de la sécurité incendie n'ait pas un vocabulaire spécifique. Dans ce cas, très classiquement, il serait constitué de mots simples et de mots composés. La fin du chapitre est consacrée à la construction de ces deux dictionnaires, d'abord les mots composés, puis les mots simples.

2. Recherche des mots composés

Les mots composés, formés essentiellement à partir de mots connus du vocabulaire généraliste (puisqu'il y a peu de mots simples inconnus[1]) permettent en particulier de désigner des concepts nouveaux ou spécifiques du domaine. Ils sont le plus souvent formés par composition de mots simples courants mais leur sens n'est pas toujours "calculable" à partir du sens de chaque composant, ou bien repose sur une définition technique. On citera comme exemples :

colonne humide, N+NA:fs
colonne sèche, N+NA:fs
écran de cantonnement, N+NDN:ms
unité de passage, N+NDN:fs

A côté de ces expressions qui s'apparentent plutôt à des expressions figées, on s'intéresse aussi, dans le contexte de la sécurité incendie, aux expressions libres et aux constructions productives qu'il peut être intéressant de repérer dans les textes parce qu'elles sont employées dans ce domaine spécifique, et par les auteurs des documentations, et par les utilisateurs.

Dans ce paragraphe, nous détaillons les opérations nécessaires à la construction de dictionnaires de mots composés spécifiques au domaine de la sécurité incendie. Ces tâches sont effectuées essentiellement à partir du corpus grâce aux fonctionnalités offertes par INTEX qui permet de rechercher des séquences de mots repérés grâce à leur étiquette, le corpus ayant été préalablement étiqueté. Les méthodes mises en œuvre quant au repérage des séquences candidates à devenir des noms composés, et aux critères de sélection parmi ces séquences candidates, ont déjà été détaillées dans de nombreux travaux (G. GROSS 1990, A. PONCET-MONTANGE 1991, A. MONCEAUX 1993).

2.1. Utilisation de patrons syntaxiques

On a d'abord repérer automatiquement, à l'aide de patrons syntaxiques, les noms composés. Ensuite les séquences sélectionnées ont été examinées afin de :

- éliminer les ambiguïtés et rétablir des classements syntaxiques corrects des expressions repérées :
- prolonger éventuellement les expressions de structure particulière qui n'avaient pas été sélectionnées complètement avec le patron ;
- choisir les expressions pertinentes en fondant ce choix, dans tous les cas subjectif, sur les connaissances générales de langue et celles particulières au domaine.

Les patrons classiques de mots composés ont été construits, sous forme d'automates, à partir des étiquettes proposées par INTEX et donc déjà utilisées dans les dictionnaires ; les patrons ont ensuite été appliqués au corpus. Ces patrons se justifient par l'expérience et par le fait que des traitements

[1] Dans le reste de l'exposé, les dictionnaires utilisés par INTEX sont complétés par les mots qui étaient considérés comme inconnus lors de la première utilisation du logiciel. L'ensemble de ces mots constituent une partie du dictionnaire de spécialité.

informatiques existent qui permettent les flexions quasi-automatiques des listes de noms composés établies sur ces critères de structure :

- patron de recherche des *NA* pour repérer les noms suivis d'un adjectif comme *adaptateur multiple* ou *air neuf* ;
- patron de recherche des *NDN* pour repérer les séquences *nom de nom* comme *aérotherme à gaz* ou *boîte à graisse* ;
- patron de recherche des *NPN* pour repérer les séquences *nom préposition nom* comme *mise en pression* ou *agent de sécurité* ;
- patron de recherche des *NN* pour repérer les séquences *nom nom* comme *bec papillon* ou *bloc porte*.

2.1.1. Expressions de type *NA*

L'automate utilise évidemment les étiquettes *nom* et *adjectif*, mais on a aussi admis en position d'adjectif des participes passés à valeur adjectivale comme *saturé* ou *qualifié* utilisés dans *vapeur saturée* ou *agent qualifié*. Dans ces expressions, l'adjectif est un modifieur du nom et prend donc le genre et le nombre de ce nom.

Certaines expressions, dans les textes, figurent exclusivement au pluriel comme :

les conditions générales
les conditions particulières
les directives générales
des combles inaccessibles

on les a conservées telles quelles dans les dictionnaires ; les expressions au singulier, *un comble inaccessible* par exemple, ne sont donc pas reconnues.
Une expression comme *arrêtés conjoints* a aussi été conservée exclusivement au pluriel. Au singulier, l'expression devrait être précisée : *un arrêté conjoint (à celui du ... + à la décision ...)* ou *un arrêté pris conjointement avec ...*

On a aussi rangé dans cette catégorie des expressions qui ne sont pas véritablement de structure *NA*, mais qui peuvent s'y rattacher comme *chauffe-eau instantané* ou *garde-corps plein* dans lesquelles le nom tête est lui même un nom composé formé d'un préfixe suivi d'un nom : *XI N*.

En observant les résultats de la sélection par automate, on a ensuite élargi la recherche aux séquences dont le nom ou l'adjectif sont précédés de l'adverbe *non* comme dans *cirque non forain* ou *non conformité grave*.

Certaines expressions de type *NA* appellent nécessairement un complément explicite ou précisé dans une autre partie du contexte. C'est par exemple le cas pour :

système conforme	*(aux dispositions ... + à la norme ... + à l'arrêté ... + aux prescriptions ...)*
dégagement non réservé	*aux seuls électriciens*
local accessible	*au public*
façade exposée	*au feu*

Ces constructions peuvent être très productives. On a néanmoins ajouté dans les listes les expressions *NA* sans complément, mais aussi certaines séquences, essentiellement parce qu'elles recouvrent des notions importantes pour la sécurité incendie, ont été dupliquées dans des listes de structures plus

complexes, en particulier *Nom Adjectif Préposition Nom*, comme : *local réservé au sommeil* ou *local accessible au public.*

Ces différents composants : préfixe accompagnant le nom tête, négation devant le nom, négation devant l'adjectif, complément de l'adjectif, peuvent se combiner. Selon leur intérêt par rapport au domaine, on les a ou non répertoriés dans les dictionnaires.

En utilisant ces filtres *NA*, on a isolé des *GN* qui font en fait partie d'une préposition composée. Elles ont été rajoutées dans les dictionnaires appropriées. C'est le cas par exemple de :

de façon (non + E) rigide
de manière (non + E) exhaustive
de type (non + E) permanent
de type (non + E) défini
dans un cas (non + E) prévu

2.1.2. Expressions de type *NDN*

Les mots composés de ce type ont été obtenus à l'aide d'un automate ; le deuxième nom peut ou non être précédé d'un déterminant qui éventuellement se combine avec la préposition *de*. On obtient ainsi des noms comme : *armature de sécurité* ou *foyer des artistes.*

Le deuxième nom peut être un nom propre ; dans ce cas, le deuxième nom n'est pas précédé d'un déterminant, mais dans le contexte le nom composé est accompagné d'un déterminant défini. Dans le corpus étudié, ces structures permettent essentiellement de nommer des inventions : *la bombe de Parr* ou *la méthode de Dumas*, mais on trouve aussi *arbre de Noël* et *feu de Bengale* (ou *bengale*) qui sont plusieurs fois évoqués dans la réglementation.

L'un des noms peut être précédé de la négation *non* : *dispositif de non arrêt [automatique], exigences de non transmission [du feu].* La possibilité de placer une négation devant le nom tête n'est pas attestée dans le corpus.

Le corpus peut proposer pour les mêmes séquences *N de N* des formes différentes, quant au nombre et à la présence d'un déterminant, pour le deuxième nom alors que le nom tête reste inchangé comme pour : *bureau de vérification, bureau des vérifications.* Et, même si la forme au singulier est unique, cette question se pose lors du passage au pluriel du nom tête : le nom du modifieur reste ou non au singulier. Le plus souvent, les entrées sont dédoublées dans le dictionnaire : *bureau de vérification* et *bureau des vérifications*, avec les flexions suivantes au pluriel : *bureaux de vérification* et *bureaux des vérifications.*
Certaines expressions ont leur deuxième nom au pluriel, même quand le nom tête est au singulier, mais cette décision se prend sur des critères sémantiques et non formels ; c'est le cas de *batterie de portes* ou *aggloméré de fibres.*
Un même nom situé en position de modifieur peut se comporter de manières différentes selon le composé auquel il est intégré : *coup de feu* fait son pluriel en *coups de feu*, alors que *classe de feu* devient *classes de feux.* Là aussi, ce sont des critères sémantiques et non formels qui permettent d'établir les flexions convenables.

Des mots composés peuvent être attestés dans le langage courant, mais se révéler non pertinents dans un langage spécialisé. C'est le cas de *bris de glace*, qui n'est pas retenu dans le vocabulaire de la sécurité incendie parce que dans le corpus, il est uniquement employé dans le nom *boîtier à bris de glace* pour désigner un objet particulier au domaine. Dans ce contexte, le nom complet de structure

NPNPN (où *P* désigne une préposition) doit être recherché d'une manière prioritaire par rapport à la séquence tronquée *NPN*.

On n'a pas retenu pour le dictionnaire de noms composés les expressions dont le deuxième nom est *classe*, *catégorie*, *type* car celles-ci doivent nécessairement être précisées par la valeur de la classe, la catégorie ou le type. On obtient ainsi des schémas productifs :

N de classe valeurClasse
N de catégorie valeurCategorie
N de type valeurType

mais où *valeurClasse*, *valeurCategorie* et *valeurType* ne peuvent être listés a priori parce qu'ils dépendent du nom tête *N*.

2.1.3. Expressions de type *N à N*

L'automate construit pour repérer les mots composés de structure *N à N* permet de sélectionner des expressions comme *chaudière à vapeur* ou *mise à la terre*.

Dans ces expressions, le deuxième nom est généralement au singulier, et ne varie pas quand le nom composé est mis au pluriel ; le nom tête prend seul la marque du pluriel : *des chaudières à vapeur* ou *des essais à froid*.
Mais cette remarque ne peut être érigée en règle : des composés *N à N* où le nom tête est au singulier demande un nom modifieur au pluriel comme *local à skis* ou *palier à aiguilles*.
Comme on l'a vu précédemment, ce sont des critères sémantiques et non formels qui permettent de déterminer, quand seules les formes avec les deux noms au pluriel figurent dans le corpus, le lemme du nom composé et les flexions admissibles. Dans les cas douteux, les entrées du dictionnaire sont dupliquées pour rendre compte des différentes possibilités d'orthographe.
L'utilisation ou non d'un déterminant devant le nom du modifieur conduit aussi à multiplier les entrées. C'est le cas, par exemple, pour les composés formés sur le nom tête *accès*. Dans le contexte de la sécurité incendie, la notion d'accessibilité à certains locaux ou objets est réglementée, d'où l'intérêt de recenser toutes ces expressions et d'ajouter celles qui paraissent pertinentes dans le dictionnaire de noms composés :

accès à un local	accès à un objet
accès à la réserve	*accès aux machines-outils*
accès aux emplacements de stockage	*accès aux moyens de secours*
accès à la chaufferie	*accès à l' organe de coupure*
accès au sas	*accès à la trappe de (service+visite)*
accès au volume-recueil	
accès à l'ascenseur	
accès à (la cage d' +l') escaliers	
accès au logement du gardien	
accès aux chambres	
accès à la zone de locaux à sommeil	

Pour toutes les expressions précédentes, la présence d'un déterminant devant le nom du modifieur est obligatoire, mais ce déterminant peut être défini ou non. Cette possibilité croisée avec le fait que, pour le lemme du nom composé, l'ensemble (modifieur plus déterminant) peut se trouver au singulier ou au pluriel conduit à multiplier le nombre d'entrées du dictionnaire.

Le corpus ne donne pas d'exemples d'expressions de type *N à N* qui acceptent l'insertion d'une négation devant le nom tête, devant la préposition *à* ou devant le nom du modifieur. Pour qu'une

négation soit sémantiquement acceptable, il faut qu'elle porte sur une propriété associée au nom tête ou au nom du modifieur (cf. § 2.1.6).

Des expressions construites sur ce schéma peuvent être doublées avec des expressions synonymes construites sur un autre schéma :

accessoire (de + nécessaire à la) distribution
immeuble (d' + à usage d') habitation

2.1.4. Expressions de type *N Préposition N*

On regroupe dans cette catégorie les noms composés construits sur le schéma *N Préposition N* où la préposition n'est ni *de* ni *à*. Sur la manière de repérer ces expressions, les difficultés à lemmatiser l'expression à cause des variations sur l'emploi des déterminants, le nombre du modifieur, l'insertion éventuelle d'un adverbe *non*, les remarques que l'on peut faire sont identiques à celles concernant les expressions *N de N* et *N à N*.

2.1.5. Expressions de type *NN*

Les mots composés de type *NN* constituent la classe d'effectif le plus réduit, mais ces mots sont le plus souvent très spécifiques au domaine, voire à l'auteur du texte, et difficiles à interpréter sans contexte, même pour un spécialiste. Classiquement les deux noms peuvent être juxtaposés ou liés par trait d'union ; mais dans le corpus étudié, on trouve aussi des expressions où les noms sont séparés par une barre oblique. Plusieurs formes peuvent coexister. C'est le cas de :

construction standard	et	*construction-standard*
marque NF		*marque NF*

Ces séquences sont le plus souvent obtenues par effacement de la préposition ou de la relative qui devrait joindre les deux noms. La préposition effacée n'est pas identique dans toutes les expressions *NN* comme on peut le voir dans les exemples suivants où elle est insérée [entre crochets] :

sécurité incendie	*pour*	*sécurité [en cas d']incendie*
bec papillon		*bec [en forme de] papillon*
information matériaux		*information [sur les] matériaux*
filetage gaz		*filetage [pour le] gaz*
gaine acier		*gaine [en] acier*
gaz réseau		*gaz [du] réseau*
marquage NF		*marquage [par] NF*
plancher plafond		*construction [qui est aussi un] plafond*
arrêté type		*arrêté [qui est du] type [attendu]*

Pour les derniers exemples, la reconstitution d'une expression complète à partir d'une séquence *NN* ne se limite pas à rajouter une préposition, ni même une relative qui pourrait toujours être *qui est* pour obtenir un *GN : N qui est N*. On a aussi :

mur rideau	pour	*mur [qui constitue un] rideau*

Dans d'autres constructions, la séquence *NN* pourrait être réécrite sous la forme *N et N*. C'est le cas par exemple pour :

nickel-chrome	pour	*nickel et chrome*
amiante-ciment		*amiante et ciment*
ventilation-réfrigération		*ventilation et réfrigération*

production émission *production et émission*

Dans ces cas où la juxtaposition des deux noms correspond à l'effacement d'une conjonction de coordination, il est possible de trouver dans le corpus à la fois les deux expressions *N1N2* et *N2N1*, même si une forme est consacrée par l'usage. *Chrome-nickel* n'est pas employé, pas plus que *ciment-amiante*, mais on trouve aussi dans le corpus étudié les expressions *réfrigération-ventilation*, *émission-production*, *logement-foyer*, *foyer-logement*, ...

Certaines tournures sont très productives.
Pour *type*, les emplois des constructions *type valeurType* se superposent à celles de patron *de type valeurType* vues précédemment (les deux constructions sont identiques après l'effacement de *de*). Dans les deux cas, ces expressions ne sont que des modifieurs d'un nom tête situé plus à gauche dans la phrase et avec lequel il forme un *GN* de structure *N Mod* où *Mod* se développe en *de type valeurType* ou bien *type valeurType*.

N (de+E) type (Alsace +OA + conduit + exutoire + sprinkleur)

On trouve aussi bon nombre de mots composés *NN* formés sur *bloc*, où le deuxième *N* est un élément de la construction :

bloc (salle + porte + cuisine + scène)

Dans le corpus est utilisée plusieurs fois l'expression *3 plis okoumé*, de structure apparente *Déterminant N N*. Elle provient en fait d'une transformation de l'expression initiale :

contre-plaqué [en] okoumé [formé de] 3 plis

dans laquelle la préposition *en* et l'adjectif qui introduit le complément *3 plis* ont été effacés pour donner :

contre-plaqué okoumé 3 plis

puis le nom classifieur *contre-plaqué* a été effacé après l'inversion des deux noms modifieurs. Sur ce modèle, on peut construire d'autres noms composés formés avec d'autres noms de bois ou différentes structures de contre-plaqué : *3 plis iroko*, *5 plis okoumé*, ...

On observe en utilisant des patrons pour retrouver les formes binaires : *NA* et *N Préposition N* que bon nombre de noms composés présentent une structure plus complexe. On a déjà vu que chacun des mots sémantiquement pleins peut admettre une négation. Le vocabulaire technique ajoute aussi des modifieurs à chacun de ces mots pour donner des formes ternaires, voire plus complexes. Dans la suite de ce paragraphe, on va faire quelques remarques sur chacune de ces structures.

2.1.6. Expressions de type *NAA*

Les expressions de type *NA* peuvent se prolonger avec un autre adjectif, ou un participe passé employé comme adjectif. On a construit des listes de ce type *NAA* qui contiennent par exemple : *eau chaude surchauffée* et *agrément technique européen*.
Tout comme les mots composés *NA*, l'un des composants peut être précédé par une négation, et l'adjectif peut être précisé par un complément : *allumage électrique fixé sur le dispositif* ou *bâtiments voisins non isolés entre eux*.

On a intégré dans les dictionnaires des expressions non attestées directement mais qui s'imposaient parce qu'elles provenaient de la transformation d'une expression qui figurait explicitement dans les textes. C'est le cas pour :

tâches techniques liées à la sécurité
couche combustible apparente
diffuseur sonore autonome
fluide frigorigène inflammable

qui sont ajoutées dans les dictionnaires bien que non attestées parce que :

tâches techniques non liées à la sécurité
couche combustible non apparente
diffuseur sonore non autonome
fluide frigorigène non inflammable

existent. Néanmoins, la construction systématique des expressions obtenues en transformant, selon la situation initiale par ajout ou effacement de l'adverbe *non* devant un composant, ne produit pas que des expressions acceptables. Certaines sont incongrues :

faute professionnelle non grave
instances officielles non compétentes

d'autres inexactes parce que le contraire d'un modifieur n'est pas, dans ce contexte de la sécurité incendie, obtenu en ajoutant *non* devant ce modifieur : construite à partir de *escaliers tournants normaux*, l'expression *escaliers tournants non normaux* n'est pas pertinente car le contraire de *normal* dans ce contexte est *supplémentaire*, et on trouve d'ailleurs dans le corpus l'expression *escalier tournant supplémentaire*.

Des mots techniques comme *organe* ou *alarme* admettent un nombre important de modifieurs qui peuvent se juxtaposer dans un ordre libre. Dans ces cas-là, on a choisi de décrire ces combinaisons de noms et modifieurs possibles, sous la forme d'automates.

2.1.7. Expressions de type *NPNPN*

Cette étiquette permet de regrouper des expressions qui admettent deux modifieurs, introduit chacun par une préposition, et sans tenir compte de la portée de chacun d'eux : *débit de renouvellement d'air*, *arrêté de réaction au feu* ou *électrovanne à ouverture sous tension*.

Certaines expressions peuvent admettrent plusieurs prépositions :

classement (de + en) réaction au feu

ou l'effacement d'un déterminant :

débit de renouvellement (de l' + d') air

Afin de ne pas multiplier inutilement le nombre de catégories[2], on donne cette étiquette à des expressions dont la structure ne paraît pas immédiatement être celle indiquée, mais que l'on retrouve en rétablissement la préposition auparavant effacée :

façade (E + sous forme de) rideau (à grille + à remplissage)

[2] L'intérêt essentiel de ces catégories est de permettre de fléchir automatiquement, à partir de dictionnaires et d'automates déjà écrits pour INTEX, des listes de mots composés ainsi étiquetés.

2.1.8. Expressions de type *NAPN* et *NPNA*

On classe dans cette catégorie toutes les expressions *N Préposition N* dans lesquelles l'un des noms est précisé par un modifieur exprimé sous forme d'un adjectif ou d'un participe passé employé comme un adjectif. Les difficultés concernant le nombre du nom du modifieur sont les mêmes que celles évoquées pour les expressions binaires *N Prép N*.

Pour la structure *NPNA*, le modifieur *A* se rapporte le plus souvent au nom qui le précède immédiatement. Dans ce cas, la plus grande partie de ces expressions ne figure pas déjà, sous une forme tronquée, dans une catégorie *N Prep N* parce que c'est le modifieur qui rend l'expression pertinente :

*aérotherme à combustible (*E + liquide + gazeux + solide)*
*exposition à caractère (*E + commercial)*
*bâtiment à occupations (*E + multiples)*
*sas à portes (*E + pleines)*

Mais on trouve aussi des exemples qui contredisent cette remarque :

porte à tambour (E + automatique)
appareil à combustion (E + étanche)

Pour ces exemples, l'adjectif se rapporte soit au nom tête comme dans *porte automatique*, soit au nom du modifieur comme *combustion étanche*, et l'expression tronquée *N à N* est aussi pertinente.

Il est aussi possible d'ajouter une négation devant l'adjectif : *appareil à combustion non étanche*, *porte à tambour non automatique*.

Pour la structure *NAPN*, l'adjectif se rapporte sans équivoque au nom tête du composé. Il peut comme pour les structures *NPNA* être nécessaire à la pertinence du mot composé, ou intervenir comme un modifieur supplémentaire dans une expression déjà attestée :

*volume (*E + libre) de fumée*
exercice (E + obligatoire) d'évacuation

On range aussi dans cette catégorie des expressions effectivement construites sur ce schéma, mais aussi les variantes éventuelles qui modifient la structure apparente de l'expression :

conduit spécial (pour le gaz + E) gaz

Dans l'exemple précédent, le codage *NAPN* permet de rendre compte de la flexion pour le pluriel : *conduits spéciaux gaz*, où le modifieur *Prép N* ne varie pas quand le nom passe au pluriel.

En revanche, des expressions comme *local à sommeil* et *local réservé au sommeil*, bien que synonymes, sont rangées dans des catégories distinctes (*NàN* pour le premier et *NAPN* pour le deuxième) parce que les consignes à appliquer pour obtenir les flexions de ces deux expressions sont différentes.

2.1.9. Les expressions non classées

Malgré la multiplicité des catégories déjà construites, il n'est pas toujours possible de trouver l'étiquette adéquate :

lampe à décharge à cathode froide
réseau fixe d'extinction autonome à eau
eau surchauffée à basse température

distribution intérieure en eau chaude

D'une manière générale, il ne paraît pas nécessaire de créer de nouvelles catégories puisqu'en tenant compte des seuls composants qui se fléchissent, on peut étiqueter ces expressions afin d'obtenir l'ensemble de leurs formes fléchies. En particulier pour les exemples précédents, on a donné l'étiquette : *NPNPN* pour la première expression puisqu'au pluriel seul le nom se fléchit en *lampes*, et *NAPN* pour les trois autres afin que le groupe *nom adjectif* passe au pluriel pour donner la forme pluriel du nom composé.

Lorsque les modifieurs sont indépendants et tous reliés sémantiquement au nom tête, leur ordre d'apparition dans l'expression peut varier selon les auteurs ; ils peuvent apparaître soit juxtaposés, soit coordonnés :

réseau fixe d'extinction autonome à eau
réseau d'extinction fixe autonome à eau
réseau d'extinction à eau fixe (et + E) autonome

Selon le nombre des modifieurs que l'on veut prendre en compte, la présentation sous forme d'automates peut devenir la plus facile à mettre en œuvre.

2.2. Repérage d'une coordination dans un *GN*

Le corpus que l'on a étudié est à vocation technique et décrit des méthodes et des outils spécifiques, à l'aide de noms concrets ou abstraits auxquels sont rattachés plusieurs modifieurs coordonnés qui précisent chacun une fonction, une caractéristique technique, une contrainte d'utilisation, ... :

appareils à combustible solide et liquide (ou à alcool solidifié)
appareils de chauffage indépendants électriques ou à combustible gazeux
une ventilation naturelle haute et basse permanente

Ce même procédé de coordination permet aussi d'associer à deux noms têtes, une liste de modifieurs qui décrivent des caractéristiques qu'ils partagent :

aménagements et installations techniques
bouches et poteaux de incendie privés
bouches ou poteaux de incendie normalisés
logements, bureaux ou zones accessibles au public, contigus

Il apparaît donc nécessaire, afin de compléter les dictionnaires de mots composés construits comme on vient de le décrire, de repérer des *GN* qui contiennent une coordination entre des modifieurs, ou entre des nom têtes, et de développer, selon des règles à établir, ces *GN* afin d'obtenir des expressions construites sur un nom tête suivi d'un ou plusieurs modifieurs, et sans coordination. Le travail sur la coordination a été développé dans le chapitre 4, et les automates construits à cette occasion ont été utilisés pour compléter les dictionnaires de noms composés. Les expressions obtenues ne présentent pas de caractéristiques différentes de celles sélectionnées par les automates de ce chapitre : les critères de classement, les incertitudes concernant la flexion des modifieurs, l'effacement éventuel du déterminant devant le nom du modifieur, les possibilités d'insertion d'une négation, sont identiques quel que soit le mode de sélection des composés. Les listes obtenues ont été fusionnées, sans mentionner l'origine de la sélection, et sont présentées en annexe.

2.3. Utilisation d'automates pour l'écriture de noms composés

On a vu précédemment qu'il peut être utile de présenter certaines expressions qui admettent plusieurs modifieurs dont l'ordre n'est pas fixé, sous forme d'automates. On va recenser les cas dans lesquels on a présenté les dictionnaires sous forme d'automates.

2.3.1. Le nom tête est un nom générique par rapport au domaine

Dans le domaine de la réglementation incendie, les contraintes imposées aux établissements diffèrent à la fois selon la nature et la durée de la manifestation qu'ils abritent, et l'effectif admissible dans les locaux. Dans ce contexte, *manifestation* peut être considéré comme un nom générique d'une classe dont chaque élément admet des modifieurs qui peuvent caractériser cette manifestation ou exprimer une notion de durée. Ainsi les composés formés à partir des noms de cette classe et avec un ou plusieurs modifieurs peuvent être construits à l'aide d'automate :

noms têtes : *activité, exposition, manifestation*

nature de l'activité : *sportive, florale, commerciale, à caractère commercial, de type typeActivité ...*

annexe, accessoire, normale, particulière, principale

durée : *temporaire + provisoire + de courte durée + à caractère temporaire*

effectif : *à faible densité*

Le nom *local* est aussi un nom générique pour le domaine et tous les modifieurs qu'il accepte peuvent aussi être employés avec des instances de *local* : *chambre, dégagement, niveau, salle, sous-sol, zone.* On a donc traité les expressions qui admettent *local* comme nom tête à l'aide d'une grammaire locale.

Dans le même contexte, la notion d'accessibilité au public est une notion clé, très largement évoquée dans la réglementation. Elle peut être exprimée positivement ou à l'aide d'une négation, et s'applique à tous les noms qui désignent des lieux ou des objets utilisés dans la lutte contre l'incendie et les phénomènes qu'il entraîne:

bâtiment + dégagement + local + niveau + salle + sous-sol + toiture + volume-recueil + zone + ... — *(non + E) accessible au public*

bouche d'incendie + poteau d'incendie + prise d'air + — *(non + E) accessible au public*

La mise en application de la sécurité incendie nécessite l'utilisation, et donc la description, d'objets spécifiques à ce domaine :

dispositif extérieur d'arrêt de la admission du combustible gazeux ou liquide
dispositif sonore à commande manuelle ou automatique
installation d'extinction automatique ou à commande manuelle
installation fixe d'extinction automatique à eau
système d'extraction forcée de l'air et des fumées

Ces composés sont construits sur un nom tête et une liste de modifieurs dont l'ordre n'est pas fixé. Le chapitre *Utilisation de grammaires locales pour la construction du dictionnaire de mots composés* détaille l'utilisation de quatre noms têtes *appareil, dispositif, installation* et *système* dans la formation de composés.

2.3.2. Les mots composés comme composants d'autres noms composés

une unité de passage coupe-feu de degré 1 heure et à fermeture automatique
des conditions d'étanchéité et d'isolation thermique
un réseau fixe d'extinction autonome à eau

Dans les exemples précédents, les modifieurs *à fermeture automatique*, *d'isolation thermique* sont eux-mêmes formés sur des noms composés, le nom tête *réseau fixe* est un nom composé. Le codage de ces noms composés comme des noms simplifierait l'écriture des nouveaux composés dont ils sont des composants. Pourtant, on n'en tient pas compte lors du codage des nouvelles expressions : *unité de passage à fermeture automatique*, *conditions d'isolation thermique* et *réseau fixe d'extinction autonome* pour plusieurs raisons. En effet, il est nécessaire que ces groupes de mots perdent leur statut de noms composés en s'intégrant à un autre composé plus complexe pour plusieurs raisons.
D'abord, le marquage d'un mot composé à l'intérieur d'un nouveau composé empêcherait de construire les flexions correctes avec les programmes existants habituellement utilisés pour cette fonction. Par exemple l'assimilation du nom tête *réseau fixe* à un nom dans *réseau fixe d'extinction autonome à eau* ne permettrait pas d'obtenir *réseaux fixes* dans la forme pluriel du nouveau composé.
D'autre part, comme on le verra lors de l'étude de la coordination à l'intérieur des *GN*, la reconnaissance et le marquage d'un nom composé peut empêcher d'identifier ou de reconstruire un autre composé avec lequel le premier partage un modifieur ou un nom tête. Par exemple si *conditions d'étanchéité* est marqué comme un nom, comment analyser *et d'isolation thermique* et quel est alors le statut de la conjonction de coordination *et* ?
Enfin, lorsque plusieurs noms composés accompagnés de modifieurs se regroupent pour former un nouveau composé, le marquage des noms composés rendrait plus difficile l'analyse et les flexions du nom complet :

formes singulier	formes pluriel
air froid	*airs froids*
courant d'air	*courants d'air*
courant d'air direct	*courants d'air directs*
courant d'air froid direct	*courants d'air froid directs*

3. Recherche des mots simples

Comme on l'a vu dans le premier paragraphe, il y a peu de mots simples non reconnus par les dictionnaires généralistes. Comme il n'y a pas de raison de postuler a priori qu'il n'y a pas de mots simples techniques spécifiques au domaine, on en conclut que ces mots simples techniques sont de deux sortes :

a) des **mots simples qui sont des abréviations de mots composés**. L'abréviation peut se faire en reprenant le nom tête comme en (31) ou un modifieur comme dans (30) :

(30) *un (établissement + E) tiers*
(31) *un établissement (recevant du public + E)*

On exclut de cet emploi les noms composés qui sont formulés complètement au début de l'unité textuelle considérée (par exemple le paragraphe) et qui sont repris dans la suite de l'énoncé par un seul composant du *GN*, en général le nom tête :

Tous les locaux doivent être équipés de détecteurs automatiques d'incendie sensibles aux fumées et aux gaz de combustion, à l'exception de la cuisine qui doit être équipée de détecteurs thermovélocimétriques.

Dans l'exemple précédant, le *GN* initial est *détecteurs automatique d'incendie sensibles aux fumées et aux gaz de combustion* qui est repris dans la suite de la phrase par le seul nom tête *détecteurs* auquel on rajoute l'épithète *thermovélocimétriques*.

b) ou bien des **mots du vocabulaire courant utilisés dans des acceptions particulières**, par exemple :

un établissement de type T

On peut alors dégager deux méthodes de recherche des noms simples, en utilisant les remarques précédentes :

- rechercher des mots simples caractéristiques du domaine en décomposant les expressions composées mises en évidence au paragraphe 2 ;
- comparer la fréquence des emplois des mots dans le corpus de la sécurité incendie par rapport à la fréquence dans un corpus qu'on pourrait considérer comme généraliste.

Dans le cadre de ce travail, la première méthode n'a pas été mise en œuvre. On n'a recherché les mots simples caractéristiques du corpus qu'en comparant les fréquences des emplois entre deux ensembles de textes.

3.1. La méthode

La méthode que l'on se propose de mettre en application est fondée sur la comparaison de deux corpus : l'un est considéré comme le corpus de référence, et l'autre constitue le corpus caractéristique de la spécialité à étudier. Pour cela, on va comparer les nombres d'occurrences des mots dans ces deux corpus. La légitimité de cette comparaison repose sur une idée simple : si un mot est très utilisé, par rapport au corpus de référence, dans les textes spécialisés, on peut en déduire que le concept ou l'objet désigné par ce mot est largement évoqué ou décrit, et qu'il est donc caractéristique de la spécialité étudiée.

L'objectif à réaliser pour pouvoir comparer[3] deux corpus, est de produire la liste des lemmes contenus dans chacun d'eux, chaque lemme étant associé à son nombre d'occurrences dans ce texte. Cette information ne peut pas s'obtenir directement en utilisant INTEX, il faut y ajouter d'autres opérations que l'on détaillera tout au long du § 3.
La première étape consiste en la formation des corpus (§ 3.2). On utilise ensuite INTEX associé à une série de programmes annexes (écrits en PERL et en C-shell sous UNIX) qui permettent d'aligner en face de chaque lemme reconnu sa fréquence. A la fin de cette étape (§ 3.3, 3.4 et 3.5), on produit pour chaque corpus la liste des lemmes, chacun accompagné de sa fréquence. On peut alors comparer les deux corpus grâce à un programme (§ 3.6) qui, à partir des deux listes de lemmes produites à l'étape précédente, donne la liste des mots simples caractéristiques du domaine, par rapport à un corpus de référence. On commente ensuite les résultats obtenus par la comparaison au § 3.7, puis on fait varier les différents éléments impliqués dans la comparaison au § 3.8.

3.2. Choix des corpus

Pour le corpus spécifique, on a rassemblé tous les textes dont on disposait sur le domaine, soit environ 2 Mo et 414 000 mots. Pour le corpus de référence on a utilisé le quotidien français *Le Monde,* pour les cinq premiers mois de l'année 1994. Il n'a pas paru pertinent d'utiliser comme corpus de référence

[3] La comparaison des corpus se faisant à l'aide d'INTEX, l'analyse adopte et prolonge les principes adoptés par ce logiciel pour la segmentation des phrases en mots, le traitement des lettres majuscules et minuscules, le comptage des occurrences, l'application des dictionnaires et la lemmatisation des termes.

un ouvrage littéraire tant la différence des vocabulaires employés, des thèmes abordés, du niveau de langue, des exigences stylistiques est flagrante. En revanche, des mots comme *chaudière*, *vapeur*, *hydrocarbure*, font partie du vocabulaire courant et quotidien d'un locuteur du 20ème siècle et se retrouvent dans le vocabulaire d'un journal.

Il est évident que le choix du corpus de référence n'est pas neutre dans la mise en évidence des mots spécifiques du domaine et une partie de l'expérimentation consiste à faire varier ce corpus.

3.3. Que comparer ?

En premier lieu, il faut segmenter le texte en mots, un mot étant une séquence de caractères compris entre deux séparateurs. Le découpage automatique nécessite de lever des ambiguïtés ; les algorithmes utilisés sont ceux mis en œuvre par les automates d'INTEX et conduisent aussi à des choix parmi les différentes segmentations possibles. Faire varier les algorithmes de segmentation n'appartient pas à la comparaison de corpus qu'on se propose de réaliser.

La première information qu'on peut extraire d'un corpus est le nombre d'occurrences de chaque mot, après lemmatisation. Par exemple la fréquence du mot *établissement* dans les deux corpus est obtenue en comptabilisant ensemble toutes les formes que peut prendre ce lemme : les mots *établissement*, *établissements*, *Etablissement*, *Etablissements*. On a donc de nouveau recours à INTEX pour étiqueter et lemmatiser le texte à l'aide des dictionnaires.

3.4. Lemmatisation

La lemmatisation ne peut se borner à utiliser INTEX en appliquant les dictionnaires aux textes étudiés. Elle comporte différentes étapes dont certaines ne se réalisent pas automatiquement. Ces opérations sont détaillées dans les paragraphes suivants.

3.4.1. Emploi des lettres majuscules et minuscules

Les lettres majuscules n'admettent, en principe, pas de signes diacritiques. Donc, quand le mot *établissement* figure dans un titre ou débute une phrase, il est transformé totalement ou partiellement en majuscules et donne *ETABLISSEMENT* ou *Etablissement* puisque la lettre *é* perd son accent en devenant majuscule. INTEX sait traiter automatiquement ces transformations : *Etablissement* sera implicitement identifié à *établissement* si cette dernière forme existe dans les dictionnaires utilisés ; par ailleurs, *Algérie* restera un mot inconnu s'il n'y a pas d'entrée *Algérie* (ni *algérie, âlgérie, àlgérie ou älgérie*) dans les dictionnaires.
Cette transformation implicite oblige à rajouter une étape de mise en correspondance entre une forme observée dans le texte et sa réécriture sous "forme canonique" quant à l'emploi de majuscules et de minuscules. Pour les mots simples, cette règle consiste à dire que tous les mots sont écrits entièrement en minuscules, avec une variation pour les noms propres qui admettent une initiale majuscule. Et c'est sous cette forme qu'on les retrouve en entrées des dictionnaires d'INTEX.

Donc, INTEX va faire implicitement la transformation des majuscules en minuscules (en considérant par exemple que la transformée de *E* peut être *é ê è* ou *ë*) et si celle-ci produit un mot qui peut être associé à un lemme des dictionnaires, INTEX va considérer que cette transformation est légitime ce qui permet de se passer d'une étape de confirmation des transformations typographiques et permet

d'automatiser le traitement. Mais ceci engendre systématiquement du bruit. Par exemple, à la forme *DU*, INTEX propose d'associer quatre lemmes à cause de l'équivalence entre *U* et *u* ou *û* :

du,du.DET:ms
du,du.PREPDET:ms
dû,devoir.V:Kms
dû,dû.N:ms

Ce principe de réécriture s'applique sans grand risque si le texte fait peu usage de majuscules et contient peu de noms propres, ce qui est par exemple le cas dans le texte technique spécialisé ; il en va tout autrement pour le texte de notre corpus de référence qui fait allusion à de nombreuses personnalités, lieux, associations, etc, qui sont tous désignés par des noms à initiale majuscule. L'association au nom commun par transformation de l'initiale majuscule en minuscule est alors abusive. D'autres associations sont possibles et, dans ce contexte, l'emploi de majuscules génère de nouvelles ambiguïtés :

a) la séquence peut représenter un mot étranger ou français :

   THE peut signifier *thé* (nom commun français) ou *the* (déterminant défini en anglais). Dans ce cas, INTEX propose systématiquement le lemme *thé*, puisque le dictionnaire anglais n'est pas utilisé ;

b) un mot comme *Français*, *Espagnol*, *Anglais*, ... peut être le nom des habitants d'un pays, le nom de la langue du pays ou l'adjectif dérivé du nom de pays. On peut aussi trouver ces mots en écriture majuscule *FRANCAIS* ou avec la seule initiale majuscule *Espagnol*, avec les mêmes ambiguïtés que la forme toute en minuscules.

c) un même nom propre peut désigner, selon le contexte, deux personnages différents :

   Barbie peut être un nom propre qui désigne un jouet : *une (poupée+E) Barbie* ou le patronyme d'un nom de personne : *Klaus Barbie*

d) le mot peut être inclus dans un nom propre composé, il en devient alors une partie et ne doit pas être comptabilisé comme un nom commun simple :

   ETATS, *Etats* peuvent être dérivés du lemme *état* (nom simple) ou constituer un mot du nom propre *Etats-Unis* ou *Etats-Unis d'Amérique* (avec ou sans trait d'union) ;

   dans *Association des Architectes Ingénieurs*, les trois noms sont des parties d'un nom propre ;

   REUNION, *Réunion* peuvent représenter le lemme *réunion* ou un mot du nom propre composé *La Réunion* ;

   AIR et *Air* peuvent être des dérivés du lemme *air* ou un mot du nom propre *Air France* ;

   BARRE et Barre peuvent représenter un nom commun ou le patronyme d'un ancien ministre ;

e) le mot peut être homographe d'un nom propre :

   TOTAL et *Total* peuvent être un nom propre qui désigne la compagnie pétrolière ou un adjectif qualificatif ; de même pour *Candide* et *candide*.

   Avec *BANDERAS* et *Banderas* l'ambiguïté porte entre le patronyme d'un nom propre et une forme conjuguée de verbe.

   Dans tous les exemples de d) et e), l'analyse conduite par INTEX est évidemment celle qui donne les noms *état*, *réunion*, *air* et *barre*, les adjectifs *candide* et *total*, et le verbe *bander*. Pour qu'il en soit autrement il faudrait que ces noms propres soient régulièrement ajoutés aux dictionnaires.

3.4.2. Correspondance entre forme et lemme

Le résultat produit par INTEX après la consultation des dictionnaires consiste en une liste dont chaque ligne contient l'entrée lexicale correspondant à la forme "observée"[4] dans le texte à analyser, suivie du lemme associé et de différentes informations syntaxiques. Par exemple :

établissements,établissement.N:mp
établissements,établissements.N:mp

Il faut maintenant apparier un mot du texte à un lemme d'un dictionnaire. A la forme *établissements* on peut associer, d'après les dictionnaires consultés, deux lemmes : *établissement* et *établissements*. D'autre part, on ne trouve pas *Etablissement* en tête d'une ligne de la liste forme+lemme produite par INTEX. On ne peut donc pas retrouver par simple consultation de cette liste tous les lemmes qu'il est possible de relier à un mot si celui-ci contient une ou plusieurs majuscules. Il faut établir ce lien "manuellement"[5]. Si l'on continue avec l'exemple de *Etablissements*, il faut décider si cette forme est un nom propre ou une partie de nom propre et dans ce cas le lemme reste *Etablissement*. Ou bien, on considère que la majuscule est due à la position du mot en début de phrase, et il faut transcrire le mot *Etablissement* sous la forme qui figure dans les dictionnaires, dans ce cas *établissement*. L'opération est la même pour *Etablissements* que l'on doit relier à *Etablissements* ou *établissements*.
A moins d'un examen cas par cas des contextes, incompatible avec le traitement de grands corpus, cette décision est prise en même temps pour toutes les occurrences de *Etablissement* et *Etablissements*. Pour notre exemple, relier *Etablissements* à *établissements* revient à dire que toutes les occurrences de *Etablissements* sont en tête de phrase (ce qui est sûrement abusif). On décide d'associer *Etablissement* à *établissement*, ainsi que *Etablissements* à *établissements*.

En revanche, les choses sont différentes pour les mots répertoriés dans les dictionnaires de noms propres d'INTEX. En tant que telle, l'entrée correspondante du dictionnaire comporte une initiale majuscule dans les dictionnaires et INTEX peut afficher le lemme proposé de la manière suivante :

France,France.N+Top:fs

Le problème se pose de la même manière que précédemment pour relier la forme *FRANCE* au lemme *France* : INTEX propose une entrée *France* après l'étape de consultation des dictionnaires, mais il faut faire "manuellement"[6] le lien entre la forme *FRANCE* du texte et cette entrée *France*.

En conclusion, pour tous les mots écrits totalement en minuscules dans le texte à analyser, l'association entre forme du texte et entrée du dictionnaire est immédiate et automatisable[7]. Pour tous

---

[4] Il ne s'agit pas exactement de la forme observée dans le texte, mais d'une forme déduite de cette forme initiale. La nouvelle graphie est obtenue par transformation des majuscules en minuscules, selon les règles et avec les ambiguïtés évoquées en 3.4.1.

[5] En fait, ce traitement n'est pas manuel, il est partiellement traité par programme (PERL) en reliant *é*, *ê*, ... à *E*. Mais, comme cela est développé plus loin, si les propositions de lemmes peuvent se faire automatiquement grâce à ces formules de transcription, le choix entre les différentes graphies ne peut être établi que par un intervenant humain (ou par consultation des dictionnaires, mais cette hypothèse ne peut être retenue car on ne dispose pas de dictionnaires de noms propres suffisants).

[6] comme cela est expliqué dans la note précédente.

[7] Elle pourra néanmoins produire un résultat ambigu si les dictionnaires proposent plusieurs analyses.

les autres mots, c'est-à-dire ceux contenant au moins une majuscule, en principe l'initiale mais aussi la totalité des lettres si ce mot figure dans un titre, l'association entre forme et lemme implique de lever, manuellement ou non, des ambiguïtés car :

- ou bien, la forme peut représenter un nom propre ou un nom commun ;
- ou bien, ce mot ne figure (ou ne devrait figurer après mise à jour du dictionnaire concerné) que dans un dictionnaire de noms composés.

Dans tous les cas, les erreurs lors de cette mise en correspondance semi-automatique entre une forme observée dans le texte et une entrée du dictionnaire sont nombreuses puisque l'association se fait pour toutes les occurrences en même temps sans aucune référence au corpus.

Lorsque l'association entre forme du texte et entrée du dictionnaire est faite, on dispose pour chaque mot d'une étiquette syntaxique. Dans un premier temps, on décide de comparer les corpus en ne tenant compte que des noms (étiquette *N*).

3.5. Calcul du nombre d'occurrences pour chaque lemme

On a vu dans 3.4 que la correspondance entre forme observée et lemme peut être entachée d'erreur (erreur due aux ambiguïtés), celle-ci se répercute sur la correspondance entre lemme et nombre d'occurrences de ce lemme. Si on reprend l'exemple de *établissement*, on a les informations suivantes :

Etablissement	*1*
établissement	*687*
Etablissements	*44*
établissements	*1245*

On a décidé d'associer le lemme *établissement* au mot *Etablissement, et établissements* à *Etablissements.* Il faut donc ajouter le nombre d'occurrences de *Etablissement* à celui de *établissement,* et celui de *Etablissements* à *établissements.* Les occurrences des lemmes deviennent alors les suivantes :

établissement	*688*
établissements	*1289*

Mais ce décompte n'est pas exact puisqu'une autre erreur systématique de comptage se produit, due à l'ambiguïté de l'étiquetage syntaxique. Par exemple, si on s'intéresse à la forme *porte* dans le journal *Le Monde*, celle-ci peut représenter : le nom *porte*, l'adjectif *porte*,[8] ou le verbe *porter* à une forme conjuguée ou comme premier mot de noms composés comme *porte-parole*, *porte-drapeau*, ... Dans ce corpus, la forme *porte* est créditée de 187 occurrences et puisque *porte* est ambigu, on va créditer les trois lemmes *porte* (*N*), *porte* (*A*) et le verbe *porter* du nombre d'occurrences de la forme *porte*. Cette décision crée du bruit puisqu'elle va faire apparaître au même niveau de fréquences le nom courant *porte* et l'adjectif très spécifique du domaine médical *porte*.

[8] essentiellement utilisé en anatomie comme dans *veine porte*, *système porte*, ...

Avec cette méthode, les noms composés sont comptabilisés non pas comme des unités autonomes mais pour autant de mots que l'expression en comporte. Par exemple, le nom *sapeur-pompier*, qui compte 127 occurrences dans le corpus spécialisé, est en fait décomposé en deux unités : *sapeurs* et *pompiers*, chacune créditée de 127 occurrences. Les noms simples *sapeur* et *pompier* n'y sont pas employés séparément dans les textes (au singulier ni au pluriel), sauf une fois pour une abréviation dans la mention *gaine pompiers chaufferie*. Au contraire, dans le corpus généraliste, l'expression *sapeur-pompier* n'est jamais mentionnée, pas plus que le mot simple *sapeur,* mais le nom *pompier* est utilisé deux fois à la place de *sapeur-pompier* et sans que le nom complet ne soit mentionné auparavant. Le nom *pompier* est donc, dans cet emploi, une abréviation de *sapeur-pompier*. On peut donc déduire de ces observations que, pour le corpus spécialisé dans le domaine de la sécurité incendie, les textes utilisent un vocabulaire précis quant à ce domaine (il existe aussi des *marins-pompiers*), et le mot *pompier* au lieu de l'expression complète *sapeur-pompier* n'a pas sa place[9], alors que dans un corpus non spécialisé, l'appellation imprécise *pompier* suffit aux auteurs et aux lecteurs.

La séparation des composants des noms composés a aussi des conséquences pour les noms dont le sens n'est pas compositionnel. En effet, pour des séquences comme *queue de cochon* ou *langue de chat*[10], le nombre d'occurrences de chacun des mots *queue*, *cochon*, *langue* et *chat* va être incrémenté alors que le sens de ces expressions n'a rien à voir avec celui de chaque constituant. Les métiers du bâtiment utilisent beaucoup de séquences de ce genre pour désigner des outils ou des procédés de confection, dont le sens n'est pas compositionnel. Et de plus, les mêmes séquences dans des métiers du bâtiment différents peuvent être employées pour désigner des objets différents : une *queue de cochon* est un outil pour un plombier, et une tige hélicoïdale dans une pompe à mortier pour le maçon ; de même, une *langue de chat* est une sorte de truelle pour un maçon ou un carreleur mais nomme aussi l'ornement d'un mur ou d'une corniche.
En revanche, la sécurité incendie est une discipline récente dont le vocabulaire spécialisé sert à désigner des objets technologiques dont les noms sont purement techniques sans aucune dimension métaphorique ou poétique. Les textes spécialisés sur lesquels nous avons travaillé contiennent peu d'expressions techniques dont le sens n'est pas compositionnel. On a relevé les mots composés suivants (dont seulement les deux derniers sont spécifiques de la sécurité incendie) :

main courante	*8 occurrences*
bec de cane	*4*
nez de dalle	*2*
voie-échelle	*6*
voie-engins	*6*

La comparaison entre les deux corpus se fait par le rapprochement entre les nombres d'occurrences de mots que l'on considère identiques parce qu'ils ont la même orthographe. Pour des termes qui supportent une acception technique à côté d'un sens courant, la comparaison du nombre d'occurrences dans un corpus avec celui concernant l'autre corpus n'a pas grand sens. Des mots comme *chandelle*, *chien*, *crinoline*, *dépouille*, *martyr*, *oiseau*, *polka*, *servante* ont un sens dans le bâtiment qui n'est en rien comparable avec celui qu'ils peuvent avoir dans les articles d'un quotidien, et même à l'intérieur

---

[9] si on considère l'expression *gaine pompiers chaufferie* comme lexicalisée.

[10] Les métiers du bâtiment utilisent quantité d'outils dont les noms font référence à leur forme par l'intermédiaire de l'anatomie des animaux. On obtient ainsi des schémas productifs : *langue de (bœuf+carpe+vache+ ...), queue de (carpe+morue+mouton+paon+ ...), pied de (biche+chèvre), pied d'aile.* D'autres outils portent directement le nom d'animaux : *une chèvre, un chien, un furet, un hérisson, un ouistiti, un singe*, etc

du corpus spécialisé ils peuvent être employés parfois dans leur sens courant et parfois avec l'acception technique. Cet inconvénient, non négligeable pour des textes sur les métiers du bâtiment, n'a pas vraiment d'importance dans le domaine de la sécurité incendie dont le vocabulaire spécifique ne contient pas d'expressions métaphoriques.

3.6. Comparaison de deux corpus

La comparaison de deux corpus est fondée sur celle des listes (lemme+occurrence) correspondant à chaque ensemble de textes et construites lors de l'étape précédente.

Pour chaque corpus, on connaît :

- le nombre d'occurrences de noms dans ce corpus ;
- la fréquence de chaque nom-lemme.

On peut alors calculer :

$v_{ref\ i}$ = nombre d'occurrences du nom i dans le corpus de référence / nombre d'occurrences de noms dans le corpus de référence

$v_{cs\ i}$ = nombre d'occurrences du nom i dans le corpus spécialisé / nombre d'occurrences de noms dans le corpus de spécialité

et mettre en place le test suivant :

si $v_{cs\ i} / v_{ref\ i} > écartComparaison$
alors le mot i appartient au domaine spécialisé

Les mots qui n'apparaissent que dans le corpus de spécialité et non dans le corpus de référence, ne peuvent être traités de la même manière. On a choisi de les considérer comme spécifiques au domaine si leur nombre d'occurrences dépasse un certain seuil (paramètre *seuilDeSpecialité* que l'on peut faire varier).

Le détail des programmes, des descriptions des fichiers et les fichiers de données eux-mêmes figurent en annexe.

3.7. Résultats

On va dans cette partie exposer les résultats expérimentaux et justifier des décisions prises en cours de travail et qui n'étaient pas examinées lors de l'exposé de la méthode.

Dans un premier temps, on a voulu conduire le test comme il a été décrit au § 3.6. Mais le nombre des noms-candidats qui apparaissent dans les deux corpus, c'est-à-dire ceux pour lesquels l'écart d'emploi correspond au critère établi, est trop important pour être traité manuellement. D'autre part, si on observe la liste ainsi obtenue, on remarque des noms très peu employés dans l'un et l'autre des corpus :

nom	occurrences dans le corpus spécialisé	occurrences dans le corpus de référence	écart
absorption	1	1	0.84
barème	1	1	0.84
déviation	1	1	0.84

freinage	1	1	0.84
maternité	1	1	0.84
natation	1	1	0.84
urée	1	1	0.84
voûte	1	1	0.84
zinc	1	1	0.84

dont on pourrait considérer l'emploi comme anecdotique, au moins dans le texte de spécialité. On décide donc d'introduire un paramètre supplémentaire, *seuilDesOccurrences*, qui permettra d'éliminer a priori les mots dont le nombre d'occurrences est inférieur à ce seuil et qui ne pourront, de ce fait, être considérés comme significatifs. On en fait varier la valeur de la même manière que *seuilDeSpécialité*, le nombre minimum d'occurrences pour les noms apparaissant seulement dans le corpus de spécialité.

Le tableau suivant récapitule numériquement les résultats obtenus lors des premiers tests.

	test 1.1	test 1.2	test 1.3	test 1.4	test 1.5	test 1.6
nombre minimum d'occurrences dans chaque corpus[11]	20	20	20	20	20	20
écart d'emplois entre les 2 corpus[12]	0.5	1	10	50	80	100
nombre minimum d'occurrences dans le corpus à comparer[13]	20	20	20	20	20	20
comparaison 1 (par rapport à *Le Monde*) :						
noms significatifs dans les 2 corpus	587	488	151	28	14	9
noms significatifs dans le corpus spécialisé seul	132	132	132	132	132	132
noms non significatifs dans les 2 corpus	1798	1897	2234	2357	2371	2376
noms non significatifs dans le corpus spécialisé seul	755	755	755	755	755	755

3.7.1. Les effectifs des corpus à comparer

	taille du corpus	mots	occurrences de noms	mots différents	noms lemmatisés	motMAJ
texte "sécurité incendie"	2 Mo	413 863	142 368	8 703	3 377	1 007
Le Monde	3 Mo	651 352	136 686	37 216	9 923	9 895

[11] C'est le nombre minimum d'occurrences dans chaque corpus - quand le nom considéré appartient effectivement aux deux ensembles de textes - à partir duquel le nom peut être sélectionné comme candidat à devenir une entrée du glossaire ou de l'index du domaine. Cette quantité est représentée par la variable *seuilDesOccurrences*.

[12] C'est la variable *écartComparaison*

[13] C'est la variable *seuilDeSpécialité*

On peut remarquer, à la lecture du tableau précédant, que si les nombres de mots sont bien dans le même rapport que celui des tailles des corpus, il en va tout autrement pour le nombre de mots différents : le vocabulaire des textes spécialisés est beaucoup plus rétréci que celui des textes à portée générale. En revanche, les textes techniques comportent beaucoup plus de noms que les textes non spécialisés.

Le traitement des mots contenant des lettres majuscules a été détaillé dans le § 3.4, mais il n'a été appliqué qu'au corpus spécifique de la sécurité incendie. Dans le corpus de référence, 9895 mots contenant des majuscules ont été détectés, mais on a décidé de les ignorer : ils ne sont pas lemmatisés, ni pris en compte dans le cumul du nombre de noms figurant dans chaque corpus (cette quantité est utilisée dans le calcul du coefficient de comparaison des emplois dans les deux corpus).
Ce nombre de mots contenant une majuscule, c'est-à-dire écrits tout en majuscules ou commençant par une majuscule, (notés motMAJ) varie notablement selon le corpus :

- 11.6 % dans le texte de spécialité (si on compare avec les mots différents) ;
- 26.6 % dans le corpus de référence.

Cette différence étant évidemment due au fait que l'utilisation de noms propres est beaucoup plus courante dans un journal que dans un texte technique. Néanmoins, cette différence a une conséquence sur le calcul de l'écart des emplois puisqu'en négligeant le traitement des mots contenant des majuscules dans le texte de référence, on fait varier la valeur de l'écart de plus de 25 % mais puisque cette erreur d'évaluation se réalise de la même manière pour tous les calculs d'écarts, elle n'intervient pas dans les valeurs relatives de ces écarts. On décide donc que la variation de ce paramètre (le nombre de mots lemmatisés dans les deux ensembles de textes) ne fait pas partie de l'expérimentation.

3.7.2. Les mots significatifs

Le but poursuivi en recherchant les mots simples significatifs du corpus est de fournir une aide à un expert du domaine pour déterminer les entrées d'un glossaire sur la sécurité incendie. Dans cette démarche, l'expert peut s'aider de l'index d'un document existant (*Règlement de sécurité contre l'incendie relatif aux ERP*, 1993). Cet index comporte 66 entrées représentant 69 mots simples et termes composés, il admet des entrées hiérarchisées mais ne peut, pour l'expert consulté, être considéré comme exhaustif du domaine. Il constitue néanmoins un point de comparaison pour notre méthode. On en donne ci-après un extrait (la liste complète des termes figure en annexe) :

tribunes
velum
ventilateurs
vestiaires
voilages, tentures, portières, rideaux
volet

A la consultation de cette liste, une difficulté apparaît immédiatement, liée à la synonymie : il est peu probable que les textes spécialisés emploient d'une manière significativement élevée à la fois les quatre mots *voilages*, *tentures*, *portières*, *rideaux*. C'est la compétence de l'expert qui permet de regrouper dans une même entrée des mots qui, bien que non synonymes, évoquent pour le domaine étudié le même concept. La difficulté corollaire réside dans le choix de l'entrée : *voilages*, *tentures*, *portières* ou *rideaux*. Dans ce cas, ce choix paraît assez arbitraire : *rideau* ou *tenture* paraîtrait plus générique que *voilages*.

Cet index donne aussi une indication quant au nombre de mots significatifs du domaine. On peut envisager que la liste de mots que l'on construit soit "du même ordre de grandeur". Si on considère que cet ordre de grandeur correspond à un rapport compris entre un et dix, cela donne pour le nombre de mots significatifs une fourchette de valeurs comprises entre 7 et 660 mots simples.

Dans une première expérimentation, on a utilisé pour les deux paramètres variables les valeurs suivantes :

- *écartComparaison* = 0.5
- *seuilDeSpecialité* = 20

Avec ces valeurs, on obtient 587 noms considérés donc comme significatifs parce que leur fréquence d'emploi est plus élevée, d'une manière qu'on a jugée notable, que celle relevée dans le corpus généraliste. Un extrait, choisi arbitrairement, est présenté ici :

sonore	*tente*	*usages*
sortie	*tenu*	*utile*
source	*terrasse*	*utilisation*
spéciale	*textile*	*valeur*
sportif	*thermique*	*validité*
stabilisation	*tiers*	*vapeur*
stabilité	*tirage*	*vapeurs*
stand	*titre*	*variation*
station	*toile*	*vase*
stationnement	*toiture*	*véhicule*
stockage	*total*	*vente*
structure	*totale*	*vérification*
suffisant	*totalité*	*verre*
suite	*toxique*	*verticale*
suivant	*tracé*	*vide*
superficie	*traitement*	*vieillissement*
supérieur	*transfert*	*vigueur*
support	*transformation*	*vis*
surface	*trappe*	*visé*
surveillance	*travaux*	*visée*
synthèse	*traversée*	*visible*
système	*tribune*	*visite*
tableau	*trottoir*	*vitesse*
tâche	*tube*	*voie*
technicien	*tuyau*	*voies*
technique	*tuyauterie*	*voisin*
téléphone	*type*	*voisins*
température	*unitaire*	*volet*
teneur	*unité*	*volume*
tension	*usage*	*zone*

Ce nombre de mots est trop important pour qu'on puisse examiner un à un les candidats-noms d'une manière cohérente ; en effet, il est difficile de reproduire, à l'examen du 600ème nom, le même comportement présidant à la sélection d'un nom représentatif d'un domaine et le risque est d'obtenir une liste non homogène quant aux critères de choix pris en compte pour l'établir. L'objectif va être de réduire ce nombre de mots-candidats en jouant, simultanément ou non, sur deux paramètres : la valeur de l'écart d'emplois entre les deux corpus à partir duquel les noms sont considérés comme significatifs du domaine et le nombre minimum d'occurrences à partir duquel on considère que l'utilisation d'un nom n'est pas anecdotique. Les résultats sont récapitulés dans le paragraphe 3.8.2.

La comparaison entre deux corpus permet de mettre aussi en évidence des mots qu'on juge spécifiques parce qu'ils n'apparaissent que dans le corpus de spécialité. Pour que ces emplois soient jugés significatifs, on ajoute une contrainte concernant le nombre minimum d'occurrences (indépendant de la taille du corpus). En mettant ce seuil à partir duquel le nombre d'occurrences est significatif à 20, on obtient les 92 mots suivants :

aération 21
aéraulique 21
aérotherme 28
accessoire 92
accumulateur 48
agrément 61
alinéa 63
an 66
appareillage 38
art 502
assujetti 62
atrium 108
atténuation 43
aval 28
azote 30
b 394
balayage 21
balisage 36
bois 174
brûleur 116
but 25
butane 52
cabine 60
caisson 48
calfeutrement 24
canalisation 247
caniveau 21
cantonnement 22
certification 27
chlore 88
clapet 82
combustion 161
compartiment 59
compteur 20
conditionnement 22
contenance 24
couvertures 26
cuisines 60
défaillance 38
dégagement 540
désenfumage 413
déversoir 24
dalle 38
desservant 103
diamètre 100
durabilité 21
e 127
encloisonnement 22
exutoire 36
flasque 22
fluide 92
franchissement 26
fusible 25
généralités 197
gaine 139
gradin 64
humidité 36
hydrocarbure 103
ignifugation 32
inflammation 68
installateur 44
intercommunication 28
isolation 46
jonction 21
marquage 28
monobloc 25
nota 21
obturation 28
orifice 66
ossature 26
ouvertures 21
ouvrants 25
parement 20
paroi 380
pascal 27
patio 20
plénum 23
plan 129
point 212
portatif 36
propane 51
réalimentation 22
récipient 101
radiant 71
recoupement 41
retrait 28
sapeur 127
signalisation 78
soudure 24
soufflerie 24
spécification 48
sprinkleur 28

On peut, sans analyse supplémentaire, supprimer :

- les lettres isolées comme *b* ou *e* ;
- des mots qui font référence au découpage matériel des textes comme : *alinéa*, *art* (abréviation typographique d'*article*) ou logique : atténuation (dans l'expression *en atténuation des dispositions de l'article ...*), *but*, *généralités*, *nota* ;
- les mots liés au contexte d'obligation des textes réglementaires : *an*, *contenance*, *diamètre*. En effet, les textes fixent des périodicités de visite ou des durées de validité, exprimées à l'aide de *an* ; de même les dimensions réglementaires des objets sont fixées par les textes à l'aide des mots *contenance* et *diamètre* ;
- les noms d'unités comme *pascal*.

D'autres ne sont qu'une partie (nom tête ou modifieur) d'un nom composé :

- *cantonnement* pour *écran de cantonnement*
- *sapeur* pour *sapeur-pompier*

Accessoire peut aussi constituer une partie d'un nom composé. Ses 92 occurrences obligent à un traitement particulier en étudiant ses concordances. On détaille les emplois suivants :

- adjectif dans *(activité + dégagement + escalier) accessoire*
- nom accompagné d'un modifieur dans *accessoire (de tuyauterie + de distribution de vapeur + des organes terminaux)* ou *petits accessoires* (expression figée au pluriel).

Dans aucun de ces emplois, *accessoire* ne peut être considéré comme un nom simple.

De même, l'étude des concordances construites sur *point* (212 occurrences) ne permet pas de le considérer comme un mot simple du domaine :

- il peut être une partie d'un adverbe composé :
  (du + d'un + de ce dernier) point de vue ou (de + en) tout point ;
- ou nom tête de mots composés
  point (d'accès + d'éclair + d'eau + de fixation + de vente + ...)
  point (bas + singulier + accessible + ...)

Le mot *caisson* est employé d'une manière notablement élevée dans le corpus de spécialité. Pourtant il ne désigne pas un objet spécifique de ce contexte mais plutôt un élément largement utilisé pour la fonction qu'il réalise.

On peut s'étonner de la présence du mot *sapeur* alors que *pompier* n'apparaît pas dans cette liste alors que les emplois sont concomitants dans *sapeur-pompier*. En fait les deux mots sont traités comme des mots simples mais alors *sapeur* n'apparaît que dans le texte spécialisé et sa fréquence d'emploi élevé justifie qu'il apparaisse dans cette liste alors que *pompier* est aussi employé, mais seul, dans les textes de portée générale et il est traité d'une autre façon.

Il n'est pas surprenant de trouver dans un tel contexte des mots comme *combustion* ou *inflammation* mais qui tels quels n'apportent pas d'information parce qu'ils sont trop vagues. Ils ne sont qu'un des constituants de noms composés :

appareil à combustion directe
(gaz + produit) de combustion
combustion (directe + vive)
inflammation des gaz
(délai + dispositif + temps) d'inflammation

Si on s'intéresse aux mots les plus fréquents (plus de cent occurrences), on trouve les noms suivants :

atrium 108
bois 174
brûleur 116
canalisation 247
combustion 161
dégagement 540
désenfumage 413
desservant 103
hydrocarbure 103
paroi 380
plan 129
récipient 101
sapeur 127

Dans ce contexte, les emplois de *desservant* sont ceux d'un participe présent[14] : *les escaliers normaux desservant les locaux*, et n'ont pas d'intérêt pour construire un glossaire.

Pour le mot *bois*, c'est le mode de dénombrement des occurrences qui conduit à ce nombre élevé de 174 emplois. En fait, *bois* se rencontre 87 fois dans le corpus spécialisé. Mais pour ce mot, INTEX propose plusieurs analyses :

bois,boire.V+t:P1s:P2s:Y2s
bois,bois.N:mp
bois,bois.N:ms:mp

Lors de la sélection des lemmes, l'acception en tant que verbe est éliminée puisqu'on n'étudie que les noms et les deux acceptions de noms sont conservées et créditées chacune, puisque la forme *bois*[15] rencontrée dans le texte est ambiguë, de 87 occurrences. Ensuite le cumul du nombre d'occurrences associé à chaque lemme se fait seulement en tenant compte du lemme, et nom de l'ensemble lemme plus informations syntaxiques ; à ce niveau, on ne fait donc plus la différence entre ces deux lemmes dont les nombres d'occurrences sont cumulés pour donner (2*87=) 174 occurrences. Néanmoins le nom *bois* est employé d'une manière significative par rapport à des textes généralistes, ce qui n'est ni surprenant ni intéressant dans un corpus concernant la sécurité incendie.
D'autre part, il n'est pas pertinent de modifier la formule de calcul concernant le cumul des occurrences : en effet, hors contexte, il est impossible de différencier les deux acceptions du mot et donc de choisir le lemme adéquat, même "manuellement". Donc, conserver les informations syntaxiques et en tenir compte lors du calcul du nombre d'occurrences d'un lemme ne garantit pas une information de meilleure qualité. Dans le corpus, cette source d'erreur se retrouve pour des mots comme :

secours,secours.N:mp
secours,secours.N:ms:mp
sol,sol.N:ms
sol,sol.N:ms:mp

où les formes *secours*[16] et *sol*[17] sont ambiguës et peuvent être rattachées à deux noms différents. De même pour la forme *frais*[18] :

frais,frai.N:mp
frais,frais.A+d:ms:mp

---

[14] L'étiquette *nom* correspond au sens suivant : *Ecclésiastique qui dessert une cure, une chapelle, une paroisse*, et est donc hors sujet.

[15] Les deux entrées du DELAS correspondent à plusieurs définitions pour chaque forme : au pluriel, ce peut être *instruments à vent en bois (ou en métal), les poteaux de but au football* ou *les cornes caduques d'un cervidé.* Pour les formes homographes au singulier et au pluriel, les différents sens se regroupent autour de : *synonyme de forêt*, ou de la *matière ligneuse et compacte des arbres.*

[16] Pour l'acception toujours au pluriel, *secours* peut être synonyme de *renfort* ou de *soins donnés à un malade ou blessé dans un état dangereux.* Quand il signifie : *aide, réconfort* ou *aumône*, il peut être singulier ou pluriel.

[17] La note de musique est invariable ; pour les autres sens, les formes singulier et pluriel se différencient par le *s* final.

[18] *frais* peut être le pluriel de *frai* (*ponte des œufs chez les poissons* ou *usure des monnaies en circulation*) ou de *frais* (*fraîcheur*). Quand il est toujours au pluriel, il est synonyme de *dépenses.*

frais,frais.ADV
frais,frais.N:mp
frais,frais.N:ms:mp

La sélection exclusive des noms permet d'éliminer les adjectifs et les adverbes. Les occurrences de *frais* rattachées au lemme *frai* ne vont pas se cumuler avec celles rattachées au lemme *frais* (puisque les lemmes sont différents), mais entre les deux derniers lemmes, il y a addition des occurrences, dans tous les cas abusive, sur le lemme *frais*.

Brûleur n'est pas un mot simple dans ce contexte, car il est le plus souvent précisé par un modifieur : *à gaz + électrique.*

Les 247 occurrences de *canalisations* évoquent toutes le même objet de la famille des conduits et tuyaux. Les différents emplois sont aussi précisés par un ou plusieurs modifieurs, sans altérer le nom tête :

canalisation (électrique + de gaz + de combustible gazeux)
(générale + E) (d'alimentation + d'évacuation + de branchement)
(de l'éclairage +des canalisation + E) de sécurite

Le cas de *dégagement* est similaire. Le mot est essentiellement employé dans des expressions liées à l'idée d'évacuation des personnes en cas d'incendie, et donc dans une acception spécifique au domaine, avec éventuellement les modifieurs suivants :

dégagement (accessoire + accessible au public + en cloisonné + en surnombre + normal + protégé + supplémentaire)

Les cas où *dégagement* est inclus dans une expression *dégagement de fumée*, non significative du domaine, représente moins de dix occurrences.

Dans les emplois d'*hydrocarbure*, le nom est éventuellement accompagné d'un modifieur , mais on peut considérer que les utilisations répétées du nom simple insistent sur les nombreux points dans lesquels les hydrocarbures sont évoqués et, par là, confortent l'idée qu'une entrée dans l'index du domaine est nécessaire.

Pour *paroi* et *récipient*, les occurrences correspondent à des lemmes uniques et qui, même s'ils constituent une partie de nom composé, renvoient toujours à la même acception de sens. Il faut éventuellement préciser les emplois à l'aide de grammaires locales :

paroi (d'essai + d'isolement + horizontale + verticale)
récipient (de butane commercial + de propane commercial + de chlore + de gaz + de peinture + fixe + mobile + transportable + vide)

Pour *plan* (129 occurrences) les emplois dans le corpus correspondent à des acceptions différentes du nom et qui sont donc cumulées à tort :

plan (d'eau + de l'éprouvette + de la façade)
plan (d'ensemble+E) de (l'établissement+la chaufferie ...) (détaillé+simplifié+E)
plan (de formation + d'intervention + d'orientation + d'essais)
plan mobile

Aucun de ces mots composés n'est spécifique de la sécurité incendie, néanmoins l'entrée *plan de l'établissement* doit dans tous les cas apparaître dans l'index parce que l'expression désigne un objet qui existe et est utilisé dans le langage courant mais qui est aussi très largement évoqué dans la réglementation parce que celle-ci crée des obligations à son égard.

Enfin, des mots comme *atrium*, *balayage*, *balisage*, *brûleur*, *certification*, *compartiment*, *désenfumage*, *encloisonnement*, *exutoire*, *hydrocarbure*, *intercommunication*, *marquage*, *recoupement*, *signalisation* sont effectivement spécifiques au domaine parce qu'ils sont employés dans

une acception particulière à la réglementation incendie ou bien parce qu'ils sont largement évoqués par la réglementation à cause de sujétions qui les concernent et qui sont propres au contexte.

L'examen des mots n'apparaissant que dans le texte spécialisé mais rejetés parce que leur nombre d'occurrences est inférieur au seuil minimal fixé dans le programme d'exploitation est aussi très utile pour fixer empiriquement la ou les valeurs jugées admissibles pour ce seuil. Si on place ce seuil à trente occurrences, on obtient 84 noms non significatifs (et donc 803 noms significatifs) ; si le seuil diminue à vingt occurrences, le nombre de mots passe à 132 (et donc 755 noms significatifs) dont un extrait figure ci-dessous.

aérosol 6
abies 1
abord 6
aboutage 1
abréviation 2
acétone 1
acétylène 3
accessibilité 18
accident 6
accord 16
accotement 2
adaptateur 2
adhésif 3
adiabatique 1
adjuvant 1
agencement 14
agenouilloir 1
aggloméré 3
agrandissement 3
ailette 2
alpinisme 1
alumine 1
améliorations 1
ammoniaque 1
amorçage 4
ampère 2
ampèremètre 2
ampliation 1
analogie 12
ancienneté 1
anesthésique 3
angulaire 2
antidérapant 1
applicateur 9
apposition 4
apprêt 3
apprêts 2
arènes 1
arête 12
archivage 1
armure 1
attaches 5
autocollant 2
automatisme 4
auvent 3
avertisseur 10
avitaillement 1
bâchage 1
bûcher 1
bagagerie 1
balancement 2
balcon 18
bardage 7
bergerie 5
blanchisserie 5
bombement 1
bornes 10
bornier 4
bourrage 1
brancard 2
branchée 4
brasure 3
bris 7
brochure 6
brome 12
brossage 4
buanderie 3
building 4
buse 4
butée 4
but 25
calage 2
caloporteur 9
calorifuge 7
calorifugeage 1
calorimètre 6
calorimétrie 4
candélabre 1
cane 2
caniveau 21
cannelure 1
capot 2
captation 3
caractérisation 1
carbonate 1
carneau 2
cathode 1
cedex 3
ceinture 5
celluloïd 1
certificateur 4
châssis 9
chaînage 1
chalumeau 3
chandelle 1
charpente 4
chaussé 1
chaux 1
chemisage 2
chevron 2
chlorofibre 2
chloromètre 1
chlorure 2
chrome 1
chronomètre 4
cierge 1
cintrage 1
cintre 2
cisaillement 6
clavecin 8
client 4
climatisation 1

On voit que la majorité des mots sont employés un petit nombre de fois, une ou deux. Néanmoins, si on laisse ce seuil très bas, par exemple trois ou quatre occurrences, le nombre de mots à examiner dépasse des limites raisonnables où il est facile d'examiner cas par cas les mots jugés significatifs. On a donc décidé de fixer ce seuil à 20 occurrences. Cette valeur est tout à fait arbitraire. Si on examine la liste des mots écartés, des noms comme *accessibilité*, *balcon* ou *fluorescence* (respectivement 18, 18

et 10 occurrences) sont un peu gênants parce que l'*accessibilité* (des locaux, des points d'eau, des moyens de secours, ...) est une idée largement développée dans la réglementation incendie ; les *balcons* constituent un dégagement accessoire qui permet de mettre à l'abri ou d'évacuer du public, un autre point clé de la réglementation et la *fluorescence* est la propriété des équipements de sécurité d'être visibles même en cas de limitation de l'utilisation de l'électricité.

Dans tous les cas, l'utilisation d'une méthode automatique ne peut fournir que des candidats à devenir des entrées d'un dictionnaire de spécialité. La validation de ces entrées possibles doit se faire manuellement en étudiant leur contexte. Des grammaires locales s'avèrent nécessaires pour préciser les contextes d'utilisation de ces mots et réaliser un glossaire sur le domaine. Néanmoins, cette méthode permet d'obtenir une base de travail sous la forme de listes de mots qui permettent de guider le travail de l'expert du domaine. Une partie importante de ce travail, et qui ne paraît guère automatisable, consiste à regrouper sur un même thème, et donc peut-être sous une même entrée de glossaire, des mots qui évoquent le même concept, le même point de la réglementation ou qui y sont reliés. C'est ainsi par exemple que si on regarde les mots *tenture*, *rideau*, *voilage* et *portière*, on peut considérer que *tenture* et *rideau* sont significatifs dans le domaine, pourtant il est essentiel de les relier, au moins, aux deux autres termes pour évoquer complètement le problème de l'inflammabilité des objets mobiliers utilisés à l'intérieur des locaux recevant du public. C'est la compétence de l'expert qui permet de faire ces associations d'idées et de les synthétiser en choisissant une entrée. Dans le document existant, c'est l'entrée *voilages* qui a été choisie, à laquelle sont reliés les mots *tentures*, *portières*, *rideaux*. Si on utilise les résultats de notre étude, on peut observer que c'est le mot *rideau*[19] qui est utilisé de la manière la plus significative dans le corpus et choisir plutôt cette entrée pour l'index ou le glossaire. Par ailleurs c'est l'expérience de l'expert qui lui permet de construire à partir de l'entrée *tentures, rideaux, portières et voilages* (qui figure telle quelle dans le texte ou provient d'un assemblage assez immédiat des mots du texte), l'entrée *inflammabilité des objets mobiliers* qui ne figure pas dans le texte (pas plus qu'*objets mobiliers* ; quant à *inflammabilité*, il est employé trois fois dans le corpus et comme un mot simple) mais qui synthétise les problèmes attachés au mobilier, en sécurité incendie. Conserver les deux entrées dans l'index du corpus permet d'accéder au même texte de deux façons : par l'expression qui exprime le problème de l'inflammabilité ou bien par les noms des objets qui sont concernés par ce problème.

3.7.3. Les résultats annexes

Le "traitement manuel" des mots du corpus de spécialité contenant au moins une majuscules, soit 1007 occurrences, a permis de former des dictionnaires de noms propres, de sigles et de symboles utilisés dans le domaine. Les dictionnaires correspondant sont donnés en annexe.
Ce traitement n'a pas été fait pour le corpus de référence parce qu'il concernait plus de 9 500 mots commençant par une majuscule et qui n'étaient pas associables automatiquement à une entrée du dictionnaire.

[19] *Tenture* est employé 29 dans le corpus spécialisé et n'apparaît pas dans celui de référence. *Rideau* apparaît dans les deux ensembles de textes, avec un écart significatif 10.7 alors que l'écart limite varie antre 0.5 et 0.8

3.8. Variantes de la méthode

3.8.1. Choix du corpus de référence

Le premier test est effectué avec la parution quotidienne durant cinq mois du journal *Le Monde*. Dans un deuxième test, on a choisi comme corpus de référence le mensuel *Le Monde Diplomatique*. Ses centres d'intérêt sont plus restreints que dans le quotidien, et dans ce sens il réalise moins bien cette notion de référence. En revanche, il utilise une langue plus soutenue, et par là un vocabulaire plus précis.

Cette archive contient des listes de fréquences de mots français, calculées à partir du *Monde Diplomatique* (1987-1997) par Jean Véronis (Université de Provence - *www.up.univ-mrs.fr/~veronis*).
La totalité du texte comportait 11 139 376 occurrences (y compris les ponctuations), se réduisant à 150 340 formes distinctes si l'on prend en compte la différence minuscule-majuscule, ou 127 452 si on l'ignore.
Le fichier se présente sous la forme suivante :

fréquence	**forme**
689124	,
500213	*de*
344747	.
282885	*la*
230694	*l'*
201304	*et*
194431	*les*
184129	*des*
180163	*à*
176758	*le*
173456	"
154527	*d'*
146863	-
128392	*en*
113848	*du*

Malheureusement, on ne peut soumettre directement le fichier des formes à l'analyse d'INTEX parce qu'il dépasse les capacités d'indexation du logiciel. Il faut donc réduire sa taille : puisque, dans notre application, on considère qu'un mot qui apparaît moins d'un certain nombre de fois (variable *seuilDesOccurrences*) dans le corpus ne peut être considéré comme significatif, si on supprime de la liste initiale des formes toutes celles dont la fréquence et inférieure à cette variable, on n'éliminera pas ainsi des mots qui pourraient être ensuite considérés comme représentatifs du domaine.
Néanmoins, cette opération n'est pas non plus sans conséquence sur le choix des mots représentatifs puisqu'en supprimant des mots, on modifie le nombre total d'occurrences qui intervient dans le calcul de l'écart entre l'emploi d'un terme dans le corpus de spécialité et celui dans le corpus de référence. Si on élimine d'office les mots qui apparaissent moins de dix fois dans ce corpus, on passe de 150 340 formes à 33 371 formes distinctes, ce qui est compatible avec les limites d'indexation d'INTEX.

Le tableau suivant récapitule les premiers résultats concernant les corpus étudiés :

	taille du corpus	nbe de mots	nbe de mots différents	nbe de noms lemmatisés	nbe occurences de noms	nbe de motMAJ
texte "sécurité incendie"	2 Mo	413 863	8 703	3 377	148 774	1 007
Le Monde	3 Mo	651 352	37 216	9 923	133 559	9 895
Le Monde diplomatique	?	35 376	33 371	9 200	2 277 632	7 270

On n'a pas calculé la taille du corpus du *Monde Diplomatique* qui correspond aux mots retenus après le traitement effectué sur le fichier des fréquences.
Les valeurs des différentes grandeurs observées sont cohérentes les unes par rapport aux autres. Il apparaît cependant une exception concernant les occurrences des noms : bien que *Le Monde diplomatique* contienne moins de noms que le quotidien *Le Monde* (9 200 par rapport à 9 923), ces noms totalisent à peu près cinq fois plus d'occurrences : 2 277 632 par rapport à 133 559. Il n'y a pas d'explication immédiate de cette discordance dont l'explication nécessiterait une analyse à part.

Quand on applique la méthode décrite, on obtient les résultats suivants (les résultats obtenus avec *Le Monde* comme corpus de référence sont rappelés ensuite) :

	test 1.1	test 12	test 1.3	test 1.4	test 1.5	test 1.6
nombre minimum d'occurrences dans chaque corpus[20]	20	20	20	20	20	20
écart d'emplois entre les 2 corpus[21]	0.5	1	10	50	80	100
nombre minimum d'occurrences dans le corpus à comparer[22]	20	20	20	20	20	20
comparaison 2 (par rapport à *Le Monde Diplomatique*) :						
lemmes significatifs dans les 2 corpus	613	497	169	42	19	16
lemmes significatifs dans le corpus spécialisé seul	115	115	115	115	115	115
lemmes non significatifs dans les 2 corpus	1765	1881	2209	2336	2359	2362
	779	779	779	779	779	779
comparaison 1 (par rapport à *Le Monde*) :						
lemmes significatifs dans les 2 corpus	587	488	151	28	14	9
lemmes significatifs dans le corpus spécialisé seul	132	132	132	132	132	132
lemmes non significatifs dans les 2 corpus	1798	1897	2234	2357	2371	2376
lemmes non significatifs dans le corpus spécialisé seul	755	755	755	755	755	755

Numériquement, les résultats sont comparables. Malgré la différence concernant le nombre d'occurrences des noms employés, le nombre de mots significatifs dans la comparaison entre corpus de référence et corpus de spécialité varie peu : 702 en comparant avec *Le Monde Diplomatique* et 689 par

---

[20] C'est le nombre minimum d'occurrences dans chaque corpus - quand le nom considéré appartient effectivement aux deux ensembles de textes - à partir duquel le nom peut être sélectionné comme candidat à devenir une entrée du glossaire ou de l'index du domaine. Cette quantité est représentée par la variable *seuilDesOccurrences*.

[21] C'est la variable *écartComparaison*.

[22] C'est la variable *seuilDeSpécialité*.

rapport à *Le Monde*. Pour évaluer les contenus de ces listes, on a procédé à un autre traitement (*scriptComparaisonCorpusEvaluation*) dont les résultats[23] sont les suivants :

- 528 noms sont significatifs du domaine de spécialité quel que soit le corpus de référence ;
- 2632 noms ne sont pas significatifs quel que soit le corpus de référence ;
- 218 noms sont significatifs par rapport à un seul corpus de référence (et donc non significatifs par rapport à l'autre corpus).

Le début de cette liste figure ici (la totalité est donnée en annexe) :

a
abri
accessoire
accumulateur
accès
accéléré
acier
actes
admis
agent
aggravation
agrément
agréé
air
aire
alarme
alerte
alimentation
alinéa
allège
ambiance
amenée
aménagement
annexe
août
appareil
appareillage
application
approbation
armé
arrêt
arrêté
art
article
ascenseur
assemblage
assujetti
assuré
atelier
atrium
attente
attestation
atténuation
automatique
autonome
autorisation
avis
axe
b
baie
balisage
bande
bar
bas
basse
batterie
bloc
bois
bouche
bouches
bouteille
branchement
brûleur
bureau
butane
bâtiment
béton
c
cabine
cage
caisson
calcul
canalisation
canton
caractéristique
carré
cas
catégorie
centimètre
certificat
chaleur
chambre
chapitre
charge

[23] toutes choses étant égales par ailleurs : c'est-à-dire, *seuilDesOccurrence*, *écartComparaison* et *seuilDeSpécialité* conservent les mêmes valeurs dans les deux comparaisons.

On peut faire les mêmes remarques que celles du § 3.7.2. concernant la liste des mots significatifs figurant seulement dans le corpus de spécialité (d'autant plus que certains mots de la liste précédente ne figurent que dans le corpus de spécialité et proviennent donc de la liste du § 3.7.2.). Ces listes constituent un point de départ pour un travail plus précis, et non automatisable en totalité, qui consiste à supprimer :

- les lettres isolées comme *a*, *b*, *c* ... ;
- les mots qui font référence au découpage des textes ou à leur argumentation logique : *alinéa*, *art* (abréviation d'*article*), *cas*, *chapitre* ;
- les noms de mois qui font partie d'une date et qui apparaissent souvent dans le texte technique parce qu'ils référencent une loi : arrêtés du *25 juin 1980*, du *22 décembre 1981*, du *2 février 1993*, ou désignent des documents officiels : *J.O. NC du 4 mai 1982*, ... La totalité des emplois de noms de mois représentent plus de 2 100 occurrences ;
- les noms d'unités comme *centimètre* ;
- les mots qui peuvent porter l'étiquette *nom* comme : *autonome*, *automatique*, *bas*, *basse*, ... et qui, dans le texte à étudier, appartiennent à l'évidence à d'autres catégories syntaxiques : adjectif ou adverbe ;
- les mots qui seuls ne peuvent apporter aucune précision : *calcul*, *caractéristique*, *cas*, *carré*, ... et dont on doit, éventuellement, étudier les concordances.

Enfin, certains mots font partie d'un nom composé : *cabine* dans *cabine d'ascenseur*, *certificat* dans *certificat de conformité*, *canton* dans *canton de désenfumage*. Pour ces mots qui peuvent être des parties de mots effectivement significatifs pour le domaine, ou bien l'expérience dans le domaine de spécialité de l'opérateur lui permet intuitivement de rejeter le mot simple, ou bien de le compléter pour former le mot composé, ou bien il a recourt aux concordances construites sur le mot simple pour décider du statut du mot. Avec par exemple le nom *certificat*, il faut ajouter à *certificat de conformité* les expressions *certificat de qualification, certificat d'installation et d'épreuve* et *certificat d'aptitude professionnelle* qui apparaissent dans les concordances. Le mot *certificat* peut apparaître seul dans le corpus, mais il est toujours précédé d'un démonstratif ou d'un déterminant défini qui permettent de faire référence à un certificat désigné plus haut par le nom composé complet et repris ensuite seulement par le nom tête. La sélection du mot simple *certificat* n'est en fait que la conséquence du nombre significatif de mots composés formés avec ce nom tête (et ce procédé est tout à fait général).

3.8.2. Variations des valeurs des paramètres

On a fait varier les différents paramètres de la méthode :

- *écartComparaison* qui mesure quantitativement les emplois dans un corpus par rapport à un autre (c'est le rapport des fréquences : au numérateur dans le corpus de spécialité, au dénominateur dans le corpus de référence) ;
  écartComparaison =
  nbe d'occurrences dans le corpus spécialisé / nbe total de noms dans le corpus spécialisé
  / *nbe d'occurrences dans le corpus de référence / nbe total de noms dans le corpus de référence*

- *seuilDeSpécialité :* c'est le nombre minimum d'occurrences dans le corpus de spécialité pour que le nom candidat, qui figure dans le seul corpus de spécialité, soit pris en compte. Au-dessous de ce nombre d'occurrences, on considère que l'emploi de ce nom n'est pas significatif ;

- *seuilDesOccurrences* : c'est le nombre minimum d'occurrences dans chacun des corpus pour que le nom candidat, qui figure dans chaque corpus, soit pris en compte. Au-dessous de ce nombre d'occurrences, on considère que l'emploi de ce nom n'est pas significatif.

Nous avons calculé le rapport des fréquences pour tous les mots identifiés et classé les mots en fonction de ce rapport. La liste ordonnée des trente premières occurrences est présentée à la suite (une liste plus complète est donnée en annexe). Nous pouvons formuler certaines remarques :

- le nombre de noms à examiner manuellement est relativement important : (689+98=) 787 candidats dans le cas le moins sélectif (*écartComparaison* = 0.5 et nombre minimum d'occurrences quels que soient les corpus concernés égal à 20). Si l'on veut réduire ce nombre, on peut jouer sur *écartComparaison* ou sur le nombre minimal d'occurrences. Dans les deux cas, la diminution a lieu : -15% si *écartComparaison* passe à 0.9 (585 candidats, le nombre de candidats n'appartenant qu'au corpus de spécialité reste évidemment constant) et -21% si on augmente le nombre minimum d'occurrences à 30. Les diminutions restent dans le même ordre de grandeur, on peut donc conclure que ces deux paramètres ont un poids comparable dans la sélection des candidats ;

- le choix du corpus de référence (*Le Monde* ou *Le Monde Diplomatique*) ne modifie pas fondamentalement la liste de noms-candidats. Par exemple dans le test 1.1, le nombre de mots significatifs est de 787 par rapport à *Le Monde* et 788 par rapport à *Le Monde Diplomatique* ; et parmi ceux-ci 761 sont communs aux deux listes.

mot	**rapport des fréquences**
protection	171,08
largeur	154,82
matériaux	140,92
éclairage	127,25
revêtement	109,58
emplacement	108,87
détection	107,46
éprouvette	94,73
chaussée	92,61
matériau	88,25
annexe	87,31
chaufferie	85,54
évacuation	79,88
rez	78,47
extraction	74,23
généralité	69,63
escalier	66,31
agréé	62,21
extincteur	62,21
bouches	59,38
visé	57,62
raccordement	55,14
propagation	53,02
verticale	53,02

litre	52,31
superficie	52,31
classement	48,43
étanche	47,37
tuyauterie	46,66
extinction	46,42

D'un point de vue quantitatif, l'ensemble des résultats est récapitulé dans le tableau suivant :

	test 1.1	test 1.2	test 1.3	test 1.4	test 1.5	test 1.6	test 2.1	test 2.2	test 2.3	test 2.4	test 2.5	test 2.6
nombre minimum d'occurrences dans chaque corpus[24]	20	20	20	20	20	20	30	30	30	30	30	30
écart d'emplois entre les 2 corpus[25]	0.5	1	10	50	80	100	0.5	1	10	50	80	100
nombre minimum d'occurrences dans le corpus à comparer[26]	20	20	20	20	20	20	30	30	30	30	30	30
comparaison 1 (par rapport à *Le Monde*) :												
noms significatifs dans les 2 corpus	587	488	151	28	14	9	477	397	134	28	14	9
noms significatifs dans le corpus spécialisé seul	132	132	132	132	132	132	84	84	84	84	84	84
noms non significatifs dans les 2 corpus	1798	1897	2234	2357	2371	2376	1908	1988	2251	2357	2371	2376
noms non significatifs dans le corpus spécialisé seul	755	755	755	755	755	755	803	803	803	803	803	803
comparaison 2 (par rapport à *Le Monde Diplomatique*) :												
noms significatifs dans les 2 corpus	613	497	169	42	19	16	494	404	146	42	19	16
noms significatifs dans le corpus spécialisé seul	115	115	115	115	115	115	75	75	75	75	75	75
noms non significatifs dans les 2 corpus	1765	1881	2209	2336	2359	2362	1884	1974	2232	2336	2359	2362
noms non significatifs dans le corpus spécialisé seul	779	779	779	779	779	779	819	819	819	819	819	819
noms significatifs dans les 2 comparaisons	694	571	232	107	93	89	543	447	182	75	61	57
noms significatifs dans une seule comparaison	55	527	101	101	92	92	40	62	73	77	68	68
noms non significatifs dans les 2 comparaisons	2517	2168	2934	3059	3082	3086	2683	2757	3012	3115	3138	3142
noms non significatifs dans une seule comparaison	73	77	68	101	92	92	40	62	73	77	68	68

[24] C'est le nombre minimum d'occurrences dans chaque corpus - quand le nom considéré appartient effectivement aux deux ensembles de textes - à partir duquel le nom peut être sélectionné comme candidat à devenir une entrée du glossaire ou de l'index du domaine. Cette quantité est représentée par la variable *seuilDesOccurrences*.

[25] C'est la variable *écartComparaison*.

[26] C'est la variable *seuilDeSpécialité*.

3.8.3 Choix des étiquettes syntaxiques concernées par la comparaison

Dans la comparaison initiale, on a choisi de ne sélectionner que les mots pouvant supporter l'étiquette syntaxique *nom*. Les étiquettes sont appliquées d'après les dictionnaires d'INTEX et reproduisent donc les éventuelles ambiguïtés d'analyse ; par exemple pour le mot *porte*, les étiquettes proposées sont les suivantes :

porte,porte.A+d:ms:fs
porte,porte.N+Abst:fs
porte,porte.N+Conc:fs
porte,porter.V+i+31R:P1s:P3s:S1s:S3s:Y2s[27]
porte,porter.V+i+35ST:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+pi+2:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+pi+31H:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+t+10:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+t+11:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+t+32R3:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+t+38L0:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+t+38L:P1s:P3s:S1s:S3s:Y2s

Dans la première partie de l'expérimentation, on n'a recherché que les noms dont une acception peut être propre au domaine étudié. En particulier, dans l'exemple précédent, seule la deuxième analyse a été conservée. Dans ce paragraphe au contraire, en appliquant une variante de la méthode, on essaie d'isoler les verbes employés spécifiquement en sécurité incendie. Pour l'exemple précédent, on conserve les trois dernières analyses proposées, et on va déterminer avec la même méthode que pour les noms, si la fréquence d'emploi du verbe *porter* est significative dans le corpus technique par rapport au texte de référence.
La question du choix des dictionnaires se pose d'une manière plus précise : il paraît inutile d'utiliser un dictionnaire qui différencie les entrées selon leurs codes syntactico-sémantiques. L'analyse proposée par le dictionnaire morpho-syntaxique *delafm* est plus judicieuse. On aurait donc pour l'exemple précédent les possibilités suivantes :

porte,porte.A+d:ms:fs
porte,porte.N:fs
porte,porter.V+i:P1s:P3s:S1s:S3s:Y2s
porte,porter.V+t:P1s:P3s:S1s:S3s:Y2s
porte,porter.V:P1s:P3s:S1s:S3s:Y2s

et le choix de l'étiquette *verbe*, sans indication sur la construction du verbe puisque cette information n'est pas utilisée par notre programme, permettrait de ne conserver que la dernière analyse.

L'étude complète se déroule de la même manière que celle déjà détaillée pour les noms, en utilisant le logiciel INTEX :

- transcription des lettres majuscules en minuscules ;
- mise en correspondance entres les formes trouvées dans le texte et les lemmes proposés par les dictionnaires ;
- calcul du nombre d'occurrences pour chaque lemme ;
- comparaison des corpus ;

et avec les mêmes imprécisions.

Les comparaisons ont été faites avec les deux mêmes corpus de référence : le quotidien *Le Monde* et le mensuel *Le Monde Diplomatique.*

On présente les résultats de la même manière que pour les noms mais ces résultats n'ont pas été exploités. Nous avons seulement vérifié que les tendances entre les deux sélections étaient comparables : influence de la valeur minimale des occurrences, du coefficient qui mesure les écarts d'emploi entre les deux corpus, ...

Une particularité : le nombre de verbes significatifs est très largement inférieur à celui des noms : dans les mêmes conditions de comparaison (corpus de référence : *Le Monde diplomatique* et écart d'emploi égal à 10), 18 verbes significatifs et 146 noms.

	test 1.1	test 1.2	test 1.3	test 1.4	test 1.5	test 1.6	test 2.1	test 2.2	test 2.3	test 2.4	test 2.5	test 2.6
nombre minimum d'occurrences dans chaque corpus	20	20	20	20	20	20	20	30	30	30	30	30
écart d'emplois entre les 2 corpus	0.5	1	10	50	80	100	0.5	1	10	50	80	100
nombre minimum d'occurrences dans le corpus à comparer	20	20	20	20	20	20	30	30	30	30	30	30
comparaison 1 (par rapport à *Le Monde*) :												
verbes significatifs dans les 2 corpus	114	95	13	4	3	3	100	84	13	4	3	
verbes significatifs dans le corpus spécialisé seul	3	3	3	3	3	3	2	2	2	2	2	2
verbes non significatifs dans les 2 corpus	290	309	391	400	401	401	304	320	391	400	401	401
verbes non significatifs dans le corpus spécialisé seul	54	54	54	54	54	54	55	55	55	55	55	55
comparaison 2 (par rapport à *Le Monde Diplomatique*) :												
verbes significatifs dans les 2 corpus	114	89	18	6	4	3	101	79	18	6	4	3
verbes significatifs dans le corpus spécialisé seul	3	3	3	3	3	3	2	2	2	2	2	2
verbes non significatifs dans les 2 corpus	272	297	368	380	382	383	285	307	368	380	382	383
verbes non significatifs dans le corpus spécialisé seul	72	72	72	72	72	72	73	73	73	73	73	73
verbes significatifs dans les 2 comparaisons	100	77	14	3	3	3	114	88	15	4	4	4
verbes significatifs dans une seule comparaison	3	11	5	6	3	2	4	12	5	6	3	2
verbes non significatifs dans les 2 comparaisons	357	372	441	451	454	455	342	360	440	450	453	454
verbes non significatifs dans une seule comparaison	3	11	5	6	3	2	4	12	5	6	3	2

3.8.4. Autres variantes possibles

D'autres variantes sont imaginables afin de pouvoir évaluer l'effet d'un paramètre sur la liste de mots simples sélectionnés. Elles n'ont pas été mises en œuvre dans le cadre de ce travail, mais pourraient facilement être testées puisque les programmes de comparaison sont déjà écrits. Ces variantes pourraient concerner :

- le choix des textes à l'intérieur du corpus de spécialité
  Dans le corpus actuel, on a pris la totalité des textes qui concernent la sécurité incendie des ERP. On pourrait envisager de ne conserver que les textes techniques, ou seulement ceux à caractère réglementaire, ..., ou bien ajouter aussi ceux qui traitent de la sécurité incendie des immeubles de grande hauteur, ceux qui s'appliquent sur les lieux de travail, ...

- les éléments comparés
  Dans le test mis en place, on a comparé les nombres d'occurrences d'un même lemme dans les deux corpus en présence. On pourrait choisir plutôt de regarder les rangs des lemmes identifiés (avec les mêmes approximations et incertitudes que pour le calcul du nombre d'occurrences associé à chaque lemme) dans chaque corpus et de déduire de la comparaison l'appartenance ou non d'un mot au vocabulaire spécifique du domaine.

  On peut mettre en place ainsi le même type de calcul avec les rangs, que celui utilisé pour les nombres d'occurrences :

  $\rho_{ref\ i}$ = rang du mot i dans le corpus de référence
  / nombre de lemmes considérés dans le corpus de référence

  $\rho_{cs\ i}$ = rang du mot i dans le corpus spécialisé
  / nombre de lemmes considérés dans le corpus spécialisé

  si $|\rho_{cs\ i} / \rho_{ref\ i}| > seuil$
  alors le mot i appartient au domaine spécialisé

L'objectif de cette partie de l'étude était d'obtenir des listes de mots qui pourraient être considérés comme significatifs du domaine étudié : la réglementation incendie concernant les établissements recevant du public. Ces mots devaient ensuite être utilisés pour former le glossaire du métier ainsi que les entrées de l'index attaché à la réglementation.
Le vocabulaire du métier comprend, a priori, des mots simples et des mots composés. Les méthodes utilisées pour mettre en évidence les uns et les autres sont différentes, elles ont été développées et justifiées tout au long de ce chapitre. La majeure partie du travail, jusqu'à l'établissement des listes de mots candidats pour l'index et le glossaire révisées avec des traitements linguistiques et informatiques, peut être menée par un non expert du domaine technique concerné. En revanche, la fin de l'étude (choix des mots du glossaire et de l'index) doit être conduite par un expert du domaine, éventuellement aidé par un lexicographe. Le travail linguistique et lexicographique préalable fournit à l'expert des arguments, parfois indispensables, pour fonder ses choix. Dans tous les cas, les listes obtenues dépendent de l'expert : dans le nombre de termes jugés acceptables, le choix et la hiérarchisation des entrées (un mot composé plutôt que le mot simple qui constitue la tête de l'expression auxquels sont rattachés différents modifieurs qui permettent de former les expressions composées), la répartition ou la répétition des termes entre l'index et le glossaire, ...

Utilisation de grammaires locales pour la construction du dictionnaire de mots composés

1. Objectifs

L'écriture de grammaires locales concernant des termes ou des expressions employés dans le contexte de la sécurité incendie contribue à une meilleure connaissance du domaine par différents aspects que l'on détaillera dans les paragraphes qui vont suivre.

Les textes du corpus décrivent de nombreux objets concrets de construction *N-classifieur A*, *N-classifieur Prép N* et *N-classifieur N*. Les propriétés concernant ces modifieurs sont classiques : variantes lexicales et orthographiques, possibilités d'insertions, inversion et effacement des modifieurs, ... et contribuent à augmenter considérablement le nombre d'expressions à prendre en compte pour la mise au point d'un dictionnaire de noms composés spécifique à la sécurité incendie. Les automates permettent de rendre compte plus facilement des différentes combinaisons possibles et donc simplifient l'écriture des dictionnaires de noms composés. L'étude de quatre classifieurs est détaillée dans le § 2.

Certains mots, comme *classe* ou *catégorie*, ont une définition tout à fait générale et un sens plus précis dans le contexte de la sécurité incendie. Néanmoins, même à l'intérieur du domaine, *classe* et *catégorie* sont ambigus dans la mesure où ils peuvent s'appliquer à différents groupes d'objets concrets et prendre des valeurs différentes selon ces groupes. Des grammaires locales spécifiques contribuent à clarifier les emplois de ces deux classifieurs et organiser les points de vue sur ce thème. On étudiera cet aspect dans le § 3.

Les paragraphes 4 et 5 ne sont que des ébauches qui illustrent l'idée suivante : les grammaires locales écrites sous forme de transducteurs permettent de reconnaître des formulations liées à l'expression d'une idée afin de les retranscrire sous une autre forme équivalente sémantiquement.
Dans le paragraphe 4, nous esquissons les étapes nécessaires pour traduire les informations précisant l'effectif d'un établissement en un libellé de catégorie pour cet établissement[1]. Les deux informations sont sémantiquement équivalentes mais s'expriment lexicalement et syntaxiquement de manières fort différentes. Une grammaire qui permet de passer d'une formulation à l'autre est utile pour unifier l'expression de ce concept.
Dans le paragraphe 5, nous ne faisons que présenter, sans y apporter aucune solution, un problème lié à l'utilisation des nombres dans des textes : la relation d'ordre qui permet de classer tous les nombres

[1] La catégorie d'un établissement est une caractéristique qui se calcule, à partir de certaines informations, selon des règles exposées dans la réglementation. Cette information est capitale car les obligations de construction, équipement, surveillance, etc, qui s'imposent aux établissements varient en fonction de cette catégorie.

pour un lecteur humain, et en particulier d'établir que *500* est plus petit que *1000*, *1 000* ou *mille* doit être décrite pour un système automatique de recherche d'information.

2. Comparaison de noms classifieurs

Dans leurs textes, les auteurs doivent pouvoir désigner des objets concrets du domaine du bâtiment ou de la sécurité incendie : des assemblages de pièces mécaniques, d'éléments du bâtiment, etc, qui réalisent des fonctions. A cette fin, ils disposent de la structure syntaxique de groupe nominal : *N1 Prép N2* où *N1* est le nom tête que l'on considère comme un classifieur, synonyme d'*assemblage*, et *N2* qualifie cet assemblage en donnant des indications sur sa fonction ou sur ses composants. Le rôle du modifieur peut aussi être tenu par un adjectif ou un participe passé (notés respectivement *A* et *V:K*) ou un autre nom (*N*) ; on obtient alors les schémas : *N-classifieur A et N-classifieur N.*
En observant les textes, on a isolé quatre noms qui jouent le rôle de classifieurs : *appareil, dispositif, installation, système*, et étudié leurs modifieurs. Les propriétés concernant ces modifieurs sont classiques : variantes lexicales et orthographiques, possibilités d'insertions, permutation et effacement des modifieurs, etc, et contribuent à augmenter considérablement le nombre d'expressions à prendre en compte pour la mise au point d'un dictionnaire de noms composés spécifique à la sécurité incendie. Ecrire les dictionnaires de noms composés sous forme d'automates permet de rendre compte plus facilement des différentes combinaisons possibles et donc en simplifie l'écriture.

L'objectif de ce paragraphe est donc de faire une étude comparative des emplois de ces classifieurs afin de préciser leurs utilisations dans le domaine de la sécurité incendie et de renseigner le dictionnaire spécialisé de mots composés, d'une manière adéquate. Dans le langage courant, le classifieur paraît interchangeable dans la formation des *GN*, ou du moins il y a plusieurs classifieurs possibles pour chaque séquence *N Prép N*. On souhaite préciser cette intuition, et vérifier si, dans la langue de spécialité concernant la sécurité incendie, cette hypothèse a une quelconque validité. Pour cela, on comparera les rôles des classifieurs selon différents critères : effectifs, formation de noms composés, productivité de la composition, caractère compositionnel des expressions formées, degré de figement, effacement du nom classifieur ou du modifieur, différences d'emplois des expressions ...

2.1. Justification du choix des classifieurs

On a examiné les fréquences des mots qui pouvaient rendre compte de cette idée d'assemblage. Les résultats sont les suivants :

	sing.	pluriel	total
aménagement	59	164	221
appareil	**157**	**737**	**884**
appareillage	37	0	37
assemblage	29	14	43
dispositif	**381**	**285**	**666**
engin	0	46	46
ensemble	158	19	177
installation	**327**	**659**	**986**
machine	0	38	38
machinerie	0	9	9

matériel	55	60	115
mécanisme	0	12	12
système	**368**	**96**	**464**

Les effectifs les plus importants concernent : *appareil, dispositif, installation, système.*

La consultation des concordances de ces différents noms permet de vérifier si ces classifieurs figurent comme tête dans des *GN* qui désignent un assemblage auquel est associée une fonction. On a recherché essentiellement des séquences de la forme *N Prép N*, *N A* ou *N N* mais en conservant les séquences qui contiennent des modifieurs ou des éléments adverbiaux (adverbe, adjectif, négation) et dans lesquelles les modifieurs sont eux-mêmes simples ou complexes :

appareil de levage
dispositif sonore à commande manuelle ou automatique
système porte-éprouvette

La dernière vérification est fondée sur la consultation de dictionnaires (ROBERT électronique, LAROUSSE Lexis, DICOBAT, dictionnaire du bâtiment) : on a cherché les définitions et les synonymes des noms de la liste précédente afin de vérifier que d'autres termes n'étaient pas cités comme synonymes de plusieurs noms appartenant à la liste, ce qui aurait pu fournir d'autres candidats classifieurs.

Le Robert électronique, dans l'ensemble des synonymes qu'il propose, ne fournit pas de terme candidat supplémentaire. Les noms que l'on considère comme classifieurs et qu'on choisit de comparer sont donc les suivants : *appareil, dispositif, installation, système.*
Les définitions du Robert sont confirmées par celles du Larousse. Une mention spéciale est faite des emplois sans modifieur de *appareil* qui selon le contexte peut désigner trois objets concrets différents : un téléphone (*qui est à l'appareil ?*), un avion (*l'appareil s'est écrasé au sol*), un appareil dentaire (*porter un appareil*). Dans le contexte de la sécurité incendie, les téléphones sont des objets évoqués plusieurs fois car ils sont utilisés pour l'appel des secours mais, vraisemblablement à cause du nombre important d'appareils cités dans la réglementation, ils ne sont jamais désignés par le nom simple *appareil*, mais plutôt par *appareil téléphonique* et plus rarement par *téléphone*.

2.2. Comparaison des effectifs

En partant des concordances de ces noms classifieurs, on a supprimé les phrases correspondant à des emplois de noms isolés (i.e. non accompagnés de modifieur) pour retenir les expressions de constructions *N Prép N*, *N (Adj + V:K)* et *N N*. En particulier, les occurrences suivantes n'ont pas été comptabilisées :

Cette interdiction doit être rappelée à proximité de l'appareil
Cette conformité doit être attestée par l'apposition sur chaque appareil de la marque NF correspondante
Les classements sont valables pour un dispositif déterminé
Ce dispositif est complété par un élément pare-flamme 1/4 h
après vérification de l'installation par une personne ou un organisme agréé

Pour dénombrer les occurrences, lorsqu'une expression était présente plusieurs fois, on n'a conservé qu'une seule phrase. En revanche, on a retenu les *GN* qui contenaient des insertions et ceux dans lesquels la préposition est remplacée par une variante formelle. On obtient alors les effectifs suivants :

	concordances	expressions retenues
appareil	884	140
dispositif	666	177
installation	986	142
système	464	81

2.3. Formation des noms composés

Dans le corpus, les mots composés à l'aide de ces classifieurs sont exclusivement des noms dont le classifieur est la tête. Ils sont formés, essentiellement, sur les patrons suivants : *N-classifieur Prép N*, *N-classifieur (non + E) (Adj + V:K)*, *N-classifieur N* mais ils admettent aussi des variantes que l'on va préciser.

2.3.1. Les noms : *N-classifieur Prép N*

(10) *appareil (à gaz + de levage + en service normal)*
(11) *dispositif (à fonctionnement automatique + d'accès + en fosse ouverte + par fusible + sans clé)*
(12) *installation (au gaz + de cuisson + en cabine)*
(13) *système (à double flux + de fixation)*

Les exemples précédents attestent les prépositions indiquées. En revanche, on n'a pas trouvé dans le corpus d'exemples où *appareil* ou *installation* sont les têtes de *GN* avec les prépositions *par* et *sans* ; de même pour *système* et les prépositions *en, par* et *sans*. Néanmoins, ces constructions ne paraissent pas impossibles et les expressions : *appareil sans clé*, *installation sans clé*, *système sans clé*, *système en service normal*, *système par fusible* semblent acceptables.

La préposition n'est pas toujours fixée :

appareil (à + de) grande capacité

ou bien elle peut être effacée quand le modifieur est intégré à un modifieur composé :

appareil (de production + d'émission)

devient *appareil de production-émission*

2.3.2. Les noms : *N-classifieur (non + E) (Adj + V:K)*

(20) *appareil (distributeur + raccordé + non raccordé)*
(21) *dispositif (sonore + thermique)*
(22) *installation (frigorifique + semi-permanente)*
(23) *système (rigide + spécifique)*

Les classifieurs peuvent être suivis d'un adjectif ou d'un participe passé à construction adjectivale, précédé éventuellement de l'adverbe *non*.

2.3.3. Les noms : *N-classifieur N*

(30) *appareil détecteur enregistreur*
(31) *système (obturateur + porte-éprouvette)*

Le statut de nom de ces mots : *détecteur*, *enregistreur*, *obturateur*, *porte-éprouvette* n'est pas tout à fait clair, et on pourrait hésiter entre nom et adjectif mais le fait que, précédés d'un déterminant, ils puissent former un groupe nominal : *un détecteur*, *un enregistreur*, *un obturateur*, *un porte-éprouvette* postule pour qu'ils soient considérés comme des noms.
On n'a pas recensé dans le corpus d'expressions dans lesquelles *dispositif* ou *installation* sont suivis d'un nom mais les *GN* non attestés *dispositif obturateur* ou *installation porte-éprouvette* paraissent acceptables.

2.4. Les variantes formelles des prépositions

(40) *appareil de cuisson (utilisant un + à) combustible liquide*
appareil (à usage + E) domestique
(dispositif + système) (à fonctionnement + E) automatique
système de ventilation (du type+ à) double flux
(41) *appareil (nécessitant l'emploi de + à) combustible solide*
appareil (d'utilisation du + à) gaz
(42) *dispositif de commande accompagnée (fonctionnant à l'aide d'une + à) clé*
système (permettant de communiquer + de communication) avec le concierge

On trouve dans le corpus à la fois des formes développées utilisant une paraphrase ou une locution qui joue le rôle d'une préposition pour introduire le modifieur, et aussi des formes courtes où la paraphrase est remplacée par une préposition sans modification du sens de l'expression entière.

Dans les exemples suivants, seules les expressions soulignées ne sont pas attestées par le corpus :

(43) *dispositif d'ouverture (qui doit être + E) aisément manœuvrable*
(44) *dispositifs de désenfumage (qui ne sont pas obligatoirement automatiques + non obligatoirement automatiques + non automatiques obligatoirement)*

Dans les exemples (43) et (44) on a noté en première position dans la parenthèse la forme avec la paraphrase, relevée dans les textes, et ensuite la ou les formes synonymes. Ces nouveaux *GN* sont obtenus en appliquant une transformation classique où la relative en *être* est effacée.
L'exemple (43), *un dispositif d'ouverture aisément manœuvrable* est forgé à partir de l'exemple attesté dans le corpus : *un dispositif d'arrêt d'urgence de l'alimentation facilement accessible*. Les deux expressions sont construites sur le même patron *dispositif de N Adv A* où *N* est le nom simple *ouverture* dans (43) et le nom composé *arrêt d'urgence de l'alimentation* dans l'autre.
Pour (44), la forme négative du verbe *être* de la relative a pour variante l'adverbe *non* s'appliquant à l'adverbe *obligatoirement* ou à l'adjectif *automatiques* qui appartiennent à la relative et sont conservés dans l'expression transformée.

(45) a. *appareil qui produit et émet de la chaleur*
(45) b. *appareil de production émission de chaleur*
(45) c. *Un appareil de production-émission est un appareil indépendant qui produit et émet de la chaleur dans le local où il est installé*

Avec la série (45), on peut comparer différentes formulations pour désigner le même appareil. Dans (45a), l'expression est de la forme *N Mod* où le modifieur est formé par la coordination de deux relatives : *qui produit et émet de la chaleur*. La première est elliptique du complément d'objet direct, la deuxième du pronom relatif. (45b) est obtenu par nominalisation des verbes de la relative : les deux

propositions relatives en partie elliptiques sont devenues un seul nom composé *production émission* qui admet comme complément le complément d'objet direct de la phrase initiale. Enfin en (45c), on a une définition qui met en jeu et identifie deux formulations différentes :

appareil de production-émission
appareil indépendant qui produit et émet de la chaleur

et autorise donc à écrire les égalités suivantes :

appareil de production-émission de chaleur

= *appareil (indépendant + E) qui produit et émet de la chaleur*

dans lesquelles la première formulation est figée, la seconde libre.

Enfin, la phrase (45c) permet implicitement d'ajouter les deux formulations suivantes à la liste des expressions désignant le même appareil :

appareil indépendant de production-émission

= *appareil indépendant de production-émission de chaleur*

(46) *un dispositif coupant automatiquement l'arrivée du gaz*
(46) *a.* *un dispositif automatique de coupure de l'arrivée du gaz*
(46) *b.* *un dispositif de coupure automatique de l'arrivée du gaz*

L'exemple (46) montre un modifieur *coupant automatiquement l'arrivée du gaz* construit sur un participe présent qui admet un complément d'objet comme dans les exemples (40), (41) et (42), avec cette fois en plus un adverbe *automatiquement*. La transformation en un complément prépositionnel doit donc tenir compte de cet adverbe : le participe présent *coupant* devient un complément *de coupure*, et l'adverbe qui modifiait le participe présent devient un adjectif automatique qui peut se rapporter aussi bien au nom classifieur *dispositif* qu'au nom prédicatif *de coupure*. La construction (46b) qui évite la succession immédiate des trois compléments de nom est plus heureuse à entendre que (46a).
On trouve le même genre de transformations avec les exemples ci-dessous où les participes présents peuvent être remplacés par des relatives :

un dispositif provoquant automatiquement l'arrêt des moteurs
un dispositif provoquant l'arrêt automatique des moteurs
un dispositif d'arrêt automatique des moteurs
un dispositif automatique d'arrêt des moteurs

Une autre contrainte peut favoriser l'emploi d'une variante formelle de la préposition devant le nom prédicatif. Dans l'exemple (47b), le *dispositif d'arrêt automatique* concerne les *moteurs*, sujet de la phrase. Pour rendre explicite cette coréférence il faut ajouter un déterminant possessif devant *arrêt automatique*, ce qui oblige à utiliser une formulation longue où la préposition *de* est remplacée par un participe présent. La formulation (47b) aussi explicite est plus concise.

(47) a. *Les moteurs ... doivent être équipés d'un dispositif (permettant + provoquant) leur arrêt automatique en cas d'échauffement*
(47) b. *Les moteurs ... doivent être équipés d'un dispositif d'arrêt automatique en cas d'échauffement*

2.5. Les modifieurs

(48) a. *installation de gaz*
(48) b. *installation fixe de gaz*

(49) a. *dispositif de protection*
(49) b. *dispositif spécial éventuel de protection*
(49) c. *dispositif de protection différentiel*

(44) a. *système de désenfumage*
(44) b. *système de désenfumage mécanique*
(44) c. *système unique de désenfumage mécanique*

Les séquences *N-classifieur de N*, *N-classifieur A* et *N-classifieur N* acceptent l'insertion d'un modifieur supplémentaire (*souligné* dans les exemples), en plus du complément initial. Ce modifieur peut s'appliquer au nom tête comme en (48b), (49b), (49c) et (44c), ou au complément du patron initial comme en (44b). Quand il s'applique au nom tête, il peut être placé à côté de celui-ci - (48b) *installation fixe* - ou rejeté en queue de séquence, et donc plus près du complément initial, comme en (49c) : *dispositif différentiel*. Le modifieur supplémentaire placé en fin de séquence peut aussi se rattacher au complément initial, comme en (44b) *désenfumage mécanique*. Il n'y a donc pas de critère positionnel qui permette de savoir à quel nom rattacher un modifieur placé en fin de séquence.
Avec (49b), on observe l'insertion de deux modifieurs épithètes du nom tête et positionnés à côté de lui. Les séquences *éventuel dispositif spécial de protection*, et *éventuel dispositif de protection spécial* sont également acceptables, tandis que *spécial éventuel dispositif de protection*, *éventuel spécial dispositif de protection*, et *spécial dispositif de protection éventuel*, dans lesquels *spécial* est antéposé, nécessitent un contexte. L'acceptabilité de ces séquences dépend du modifieur à insérer et n'est pas spécifique des noms classifieurs étudiés. On obtient la même liste d'acceptabilité avec les classifieurs *système*, *installation* ou *appareil* :

(appareil + dispositif + intallation + système) spécial éventuel de protection
éventuel (appareil + dispositif + intallation + système) spécial de protection
**spécial éventuel (appareil + dispositif + intallation + système) de protection*
**éventuel spécial (appareil + dispositif + intallation + système) de protection*
**spécial (appareil + dispositif + intallation + système) éventuel de protection*

2.5.1. Insertions et prépositions

L'introduction d'une insertion peut avoir des conséquences sur la construction du *GN* initial :

(51) a. *installation de chauffage*
(51) b. *installation fixe destinée au chauffage*
(52) a. *système de coupure*
(52) b. *système assurant la coupure de l'arrivée du combustible*

Pour les deux paires précédentes, l'insertion d'un modifieur provoque le remplacement de la préposition *de* par une paraphrase. Pour (51a-b), le modifieur *fixe* s'applique à la tête et la préposition *de* est remplacée par la locution *destinée à*, suivie d'un déterminant. Pour (52a-b), le modifieur ajouté *de l'arrivée du combustible* s'applique au modifieur initial et l'adverbe *assurant* se substitue à la préposition *de*, sans doute pour pallier une difficulté stylistique : la construction avec trois compléments de noms introduits par *de* n'est pas agréable à lire ni à entendre. La forme *système de coupure* existe dans les textes, mais pas *système de coupure de l'arrivée du combustible* ; elle est néanmoins acceptable de même que *installation fixe de chauffage*.

2.5.2. Insertion et négation

Il est rarement acceptable sémantiquement d'insérer, dans un composé de structure *N Prép N*, une négation devant le modifieur :

installation de gaz
**installation de non gaz*
système d'extinction
**système de non extinction*

En revanche, la négation peut se combiner avec l'insertion d'un modifieur supplémentaire, mais cet ajout de l'adverbe *non* devant le complément du classifieur ne se fait pas avec une portée identique dans les deux positions du modifieur et donc intervient dans l'acceptabilité des expressions produites :

dispositif de coupure
**dispositif de non coupure*

dispositif automatique de coupure
**dispositif automatique de non coupure*

dispositif non automatique de coupure
dispositif de coupure automatique

En fait, le corpus atteste seulement deux expressions :

dispositif de non arrêt automatique de la cabine [d'ascenseur]
exigences de non transmission du feu par les façades

Et si l'on examine le premier exemple, la négation a bien une portée différente selon la position de l'insertion :

dispositif de non arrêt automatique de la cabine [d'ascenseur]
≠ *dispositif automatique de non arrêt de la cabine [d'ascenseur]*

Le premier GN est utilisé dans le corpus et décrit un appareillage existant[2], alors que le deuxième ne caractérise aucun dispositif connu en sécurité incendie et ne paraît pas très sensé.

2.5.3. Contraintes orthographiques

L'orthographe des expressions telle qu'elle s'exprime dans les textes donne des informations sur les dépendances entre noms et modifieurs à l'intérieur du *GN,* mais ces accords paraissent parfois fautifs, ou du moins imprévisibles :

système de commande (accompagné + accompagnée) fonctionnant à clé

On trouve les deux orthographes dont l'une est certainement fautive sans que les effectifs dans les textes de l'une ou l'autre orthographe permettent de trancher.

---

[2] En effet, la phrase complète est la suivante : *un dispositif de non arrêt automatique de la cabine à l'étage sinistré doit être asservi à la détection automatique d'incendie*. Ce dispositif permet ainsi de supprimer les échanges entre une partie de l'établissement dans lequel un incendie s'est déclaré, et le reste de l'établissement.

(52) *dispositif de coupure manuel incorporé*
(57) a *dispositif de commande manuelle*
(57) b *dispositifs sonores à commande manuelle ou automatique*

Si on compare les exemples (52), (57 a) et (57 b), les accords sont différents : dans (52), la forme de *manuel* le rattache nécessairement à *dispositif* ; dans (57 a), *manuelle* se rapporte sans équivoque à *commande*, mais l'expression *dispositif de commande manuel*, correcte aussi, serait interchangeable avec (57 a). Dans l'exemple (57 b), *manuelle* et *automatique* sont également accordés à *commande*. Pourtant les structures apparentes de ces expressions sont identiques : *dispositif de N Mod*. On peut alors avancer deux hypothèses. Dans (52), le nom composé initial est *dispositif de coupure*, avec un premier modifieur *incorporé* qui se rapporte à *dispositif* - par l'accord et par la sémantique : un *dispositif* peut être *incorporé* mais pas une *coupure* - et qui est rejeté en fin de *GN* à cause de la présence du deuxième modifieur *manuel*. En effet, les expressions *dispositif manuel incorporé de coupure* et *dispositif incorporé manuel de coupure* sont difficiles à entendre. De plus, le lien de *manuel* avec le nom auquel il se rapporte est plus flou : un *dispositif* peut être *manuel*, tout comme une *coupure*. Finalement, le modifieur *manuel* se rapporte, par "aspiration" de l'autre modifieur *incorporé*, au même nom *dispositif*.
Avec (57 a) au contraire, *commande manuelle* est un nom composé - tout comme *commande manuelle (et + ou) automatique* - et le schéma de construction de (57 a) est *dispositif de N* où *N* est le nom composé *commande manuelle*. Le même type d'explication peut être avancé pour (57 b) : *dispositif sonore* et *commande manuelle et automatique* sont des noms composés ; donc (57 b) est construit sur le patron *N1 à N2* où *N1* et *N2* sont des noms composés.

(56) a. *le dispositif de ventilation mécanique ou naturelle*
(56) b. *Le dispositif de ventilation est mécanique*
(56) c. **Le dispositif de ventilation est naturel*

Dans (56 a) l'expression coordonnée *mécanique ou naturelle* se rapporte à *ventilation*, et donc *mécanique* en particulier est épithète de *ventilation*. Dans (56 b), ce rattachement est impossible puisque *mécanique* est par construction attribut de *dispositif*. Quant à la phrase (56 c), elle est quasiment inacceptable. Ces différences d'acceptabilité paraissent contradictoires mais sont pourtant consacrées par l'usage.

(54) *dispositifs de régulation automatique*
(55) *dispositifs automatiques de régulation de tirage*

Les accords du modifieur *automatique* sont incohérents : alors que les deux *GN* désignent le même objet, dans (54) *automatique* se rapporte à *régulation*, alors que dans (55), il se rapporte à *dispositifs*. Les deux expressions se retrouvent pourtant dans le corpus et sont acceptables pour des spécialistes de la sécurité incendie.

Si on compare les deux groupes d'expressions (52), (57 a) et (57 b) d'une part, et (54) et (55) d'autre part, on peut ébaucher la règle suivante : les adjectifs qui peuvent qualifier aussi bien le classifieur (en l'occurrence *dispositif* et *système*) que la fonction accordée à l'assemblage comme *coupure*, *commande*, *régulation*, ... s'accordent avec le nom dont ils sont le plus proches.

Néanmoins dans (52), l'accord de *manuel* contredit cette règle. Mais on peut alors considérer qu'il obéit à une autre règle de priorité supérieure : quand, dans un *GN* de patron *N-classifieur Prép N*, deux modifieurs qui prennent la marque du genre et du nombre sont juxtaposés, que l'un d'entre eux se rattache nécessairement à un des noms et que l'autre peut se rapporter aux deux, l'accord du modifieur le moins contraint se fait de la même manière que celui dont l'accord est obligatoire.

On a essayé de vérifier la règle précédente sur le corpus. Pour cela, on a recherché des *GN* de la forme *(dispositif + système) de N (A + V:K)* pour lesquels l'accord du modifieur *(A + V:K)* paraît sous différentes formes. Dans ce type d'emploi, on a trouvé essentiellement, avec *dispositif* comme avec *système*, un seul adjectif : *automatique*. Ponctuellement, avec les deux compléments de noms *commande* et *déverrouillage*, on a d'autres adjectifs : *manuelle*, *manuelle ou automatique* pour *commande* et *électromagnétique* pour *déverrouillage*. On peut résumer ces emplois à l'aide du tableau ci-dessous :

dispositif	*(d'arrêt+ de condamnation+ de coupure+ de déclenchement+ de détection+ de détection d'incendie+ de réglage+ de régulation+ de renouvellement d'air+ d'introduction d'air frais+ d'obturation+ d'ouverture)*	*automatique*
dispositif	*automatique*	*(d'arrêt+ de condamnation+ de coupure+ de déclenchement+ de détection+ de détection d'incendie+ de réglage+ de régulation+ de renouvellement d'air+ d'introduction d'air frais+ d'obturation+ d'ouverture)*
dispositif	*de commande*	*(automatique + manuelle + manuelle ou automatique)*
dispositif	*(automatique + manuelle + manuelle ou automatique)*	*de commande*
dispositif	*de déverrouillage*	*électromagnétique*
dispositif	*électromagnétique*	*de déverrouillage*
dispositif	*électromagnétique*	*de déverrouillage*
dispositif	*de déverrouillage*	*électromagnétique*
système	*(de commande + de détection + de détection incendie + de fermeture + de d'extinction + de d'irrigation + de d'obturation + de d'ouverture)*	*automatique*
système	*automatique*	*(de commande + de détection + de détection incendie + de fermeture + de d'extinction + de d'irrigation + de d'obturation + de d'ouverture)*

Pour chacun des *GN* précédents, les deux formulations *(dispositif + système) de N Mod* et *(dispositif + système) Mod de N* sont acceptables, désignent le même objet et acceptent des variantes d'orthographe (qui se manifestent quand le nom tête est au pluriel, et le modifieur parfois au pluriel et parfois au singulier).

2.5.4. Synonymie des modifieurs

dispositif déclencheur obturateur

dispositif d'obturation
dispositif de déclenchement
système obturateur

Dispositif d'obturation et *dispositif obturateur* sont des synonymes où le modifieur est l'adjectif *obturateur* ou bien le complément de nom *d'obturation* obtenu par nominalisation de *obturateur*, mais seule la première expression est attestée par le corpus. Il en est de même avec *dispositif de déclenchement* et *dispositif déclencheur*. Pourtant, on trouve la séquence *dispositif déclencheur obturateur* et non *dispositif de déclenchement et d'obturation*. Enfin le corpus atteste *dispositif d'obturation* et *système obturateur*, mais pas *dispositif obturateur* ni *système d'obturation* ; toutes ces séquences paraissent néanmoins tout à fait acceptables.

dispositif de coupure des fluides
dispositif de coupure des gaz

Ces expressions sont synonymes dans le domaine de la sécurité incendie alors que pour le vocabulaire général, les gaz ne constituent qu'un sous-ensemble de l'ensemble des fluides.

Au contraire, des expressions qui peuvent être synonymes dans le langage courant, prennent de fait le sens précis qu'elles ont dans des domaines spécialisés dès lors qu'elles sont employées dans un contexte technique :

dispositif de charge
dispositif de chargement

Dans le vocabulaire courant, *charge* et *chargement* peuvent être synonymes pour certaines acceptions liées au poids, au fardeau. Dans ce contexte, ils ne sont pourtant pas synonymes : *charge* fait référence à un système électrique tandis que *chargement* s'emploie dans le domaine de la mécanique où le *dispositif de chargement* doit rendre compte de la *charge*, dans le sens de *force*, qui lui est appliquée.

Dans la suite de ce paragraphe, on s'intéresse aux expressions formées sur les noms classifieurs *appareil* et *système*. Dans ce cas, des modifieurs employés avec le même classifieur et qui paraissent synonymes désignent des objets différents. Ces *GN* sont donc des expressions figées.

*2.5.4.1. **Classifieur** appareil*

La sécurité incendie instaure une réglementation en particulier pour les matériels qui utilisent le gaz, à cause des risques qu'ils présentent, et en distinguent deux sortes : ceux qui concernent la cuisson et ceux utilisés pour le chauffage. Les modifieurs qualifiant les appareils ne sont pas identiques dans chacun de ces domaines et l'usage a consacré l'utilisation de certaines expressions pour un domaine, et pas pour l'autre.

chauffage	cuisson
appareil à combustible gazeux	*(appareil + installation de cuisson) fonctionnant au gaz*
appareil à gaz	*appareil (d'usage + E) domestique*
appareil de combustion à gaz	*appareil à flamme d'alcool sans pression*
appareil fonctionnant au gaz	*appareil alimenté par des récipients de gaz*
appareil mobile électrique ou à gaz	*appareil de cuisson (électrique + à gaz +*

	utilisant un combustible gazeux)
appareil générateur d'air chaud	*appareil ménager à (butane-propane + gaz)*

Le *GN, appareil (mobile + E) (alimenté + fonctionnant) au gaz* est utilisé à la fois pour le chauffage et la cuisson.

Ces expressions ne sont pas vraiment compositionnelles dans la mesure où l'usage leur accorde un sens beaucoup plus précis que celui que leur confère la stricte composition des différents éléments qui la composent : par exemple, il n'est pas explicite en examinant les modifieurs que *appareil à gaz* est utilisé pour le chauffage alors qu'un *appareil ménager à gaz* sert à la cuisson.

On trouve aussi l'expression étrange *appareil électrique ou gazeux*, à la place de *appareil à combustible électrique ou gazeux* : le nom du modifieur *combustible* a été effacé, ainsi que la préposition qui l'accompagne, pour ne laisser que les adjectifs qui s'y rapportent.

Enfin, les modifieurs décrivent des propriétés sémantiquement différentes ; il est donc possible d'ajouter plusieurs séquences pour un objet désigné. De plus, l'ordre d'apparition des modifieurs dans le *GN* n'est pas figé :

appareil indépendant de cuisson électrique
appareil de cuisson électrique indépendant
appareil électrique de cuisson indépendant
appareil électrique indépendant de cuisson

Les différentes classes de propriétés sont les suivantes :

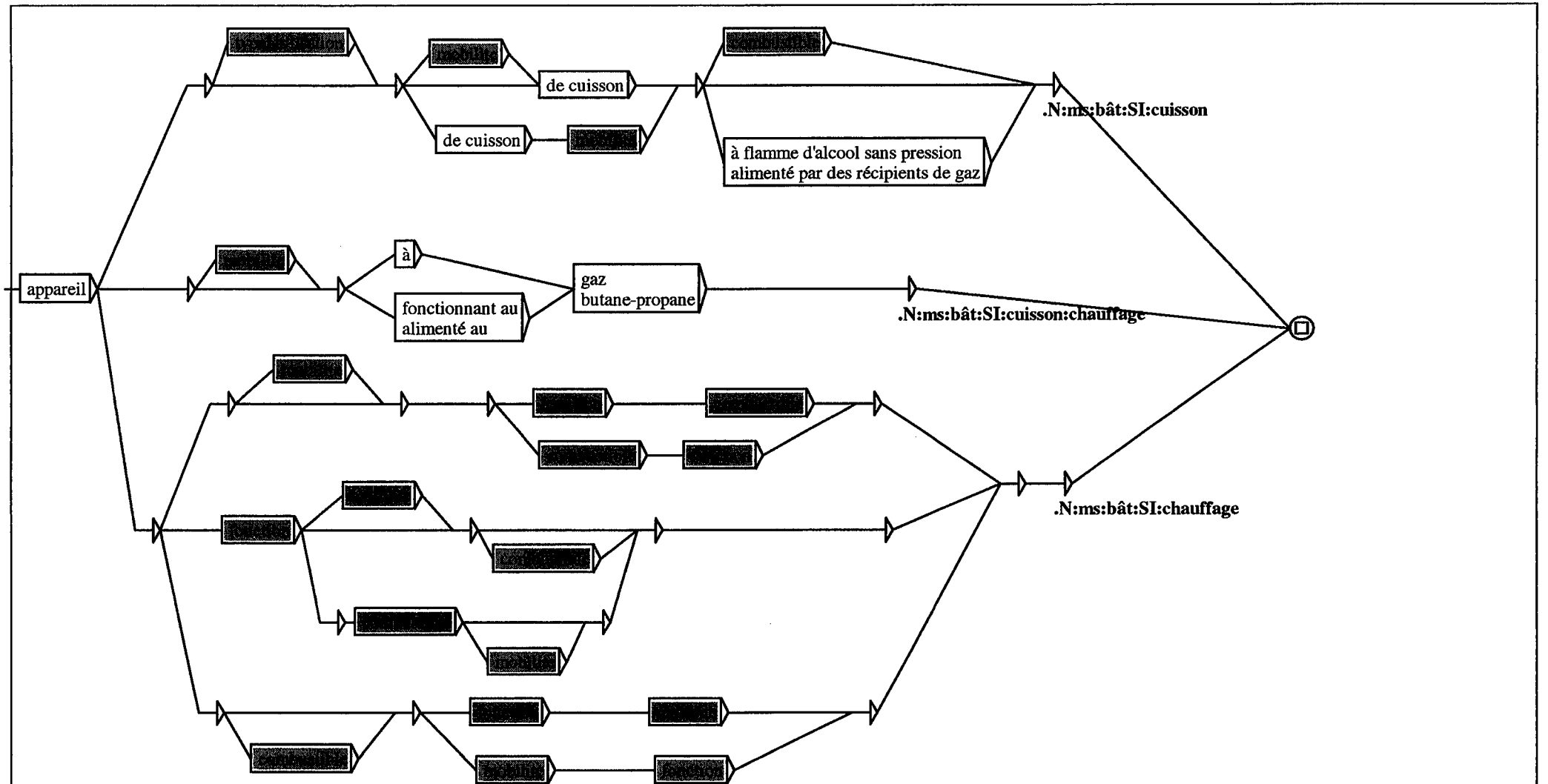
- le type d'utilisation. Les modifieurs possibles sont :
  domestique
  ménager

- la mobilité. Les modifieurs possibles sont :
  mobile
  indépendant

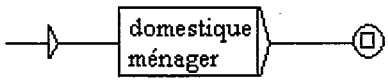
- la fonction. Les modifieurs possibles sont :
  de combustion
  de cuisson
  générateur d'air chaud

- le type de combustible. Les modifieurs possibles sont :
  électrique
  à gaz

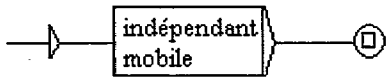
Toutes ces informations sont regroupées dans les grammaires locales qui suivent (*GlappareilCuissonChauffage.grf* qui appelle *typeUtilisation.grf, mobilite.grf, combustible.grf, fonction.grf*) et dont le langage généré est donné en annexe : 1902 ont été produites, avec le seul classifieur *appareil* (au singulier).



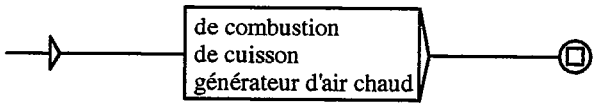
GLappareilCuissonChauffage.grf
Thu Oct 12 20:07:56 2000



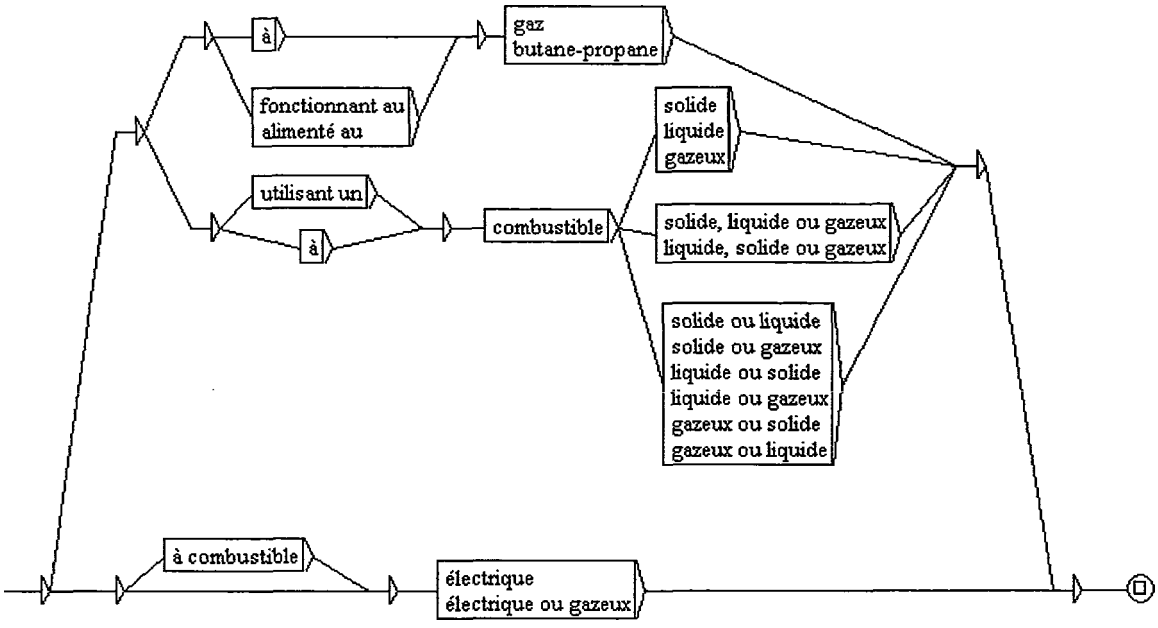
typeUtilisation.grf : grammaire locale donnant les différentes utilisations d'un appareil de cuisson ou de chauffage



mobilite. grf : grammaire locale donnant les différentes caractéristiques de mobilité d'un appareil de cuisson ou de chauffage



fonction. grf : grammaire locale donnant les différentes fonctions d'un appareil de cuisson ou de chauffage



combustible.grf grf : grammaire locale donnant les différents combustibles d'un appareil de cuisson ou de chauffage

2.5.4.2. Classifieur système

Un autre aspect développé dans la réglementation sur la sécurité incendie concerne le contrôle de l'air respiré par le public et les personnels des ERP. Ces vérifications s'exercent dans deux types de situations différentes : en situation normale, la vérification de l'air ambiant qui est soumis à des conditions d'hygiène, et le contrôle des fumées en cas d'incendie. Pour ces deux domaines, les expressions utilisées ne sont pas tout à fait identiques et leur formation provient de la description des phénomènes physiques qui se manifestent lors de ces deux types d'opérations.
Pour renouveler l'air d'un local, il faut combiner deux opérations : éliminer l'air intérieur afin de pouvoir amener de l'air extérieur[3]. On parlera donc d'un *système de ventilation* ou *système de désenfumage* qui est composé de deux *dispositifs* ou *appareils* distincts. L'arrivée d'air frais peut se faire ou bien à l'aide d'un moteur, ou bien grâce à des orifices percés dans les parois. L'élimination de l'air intérieur au local peut être naturelle : dans le cas des fumées, ce sont des gaz chauds qui montent et sortent (à condition que de l'air frais pénètre) ou mécanique : l'évacuation est forcée par un moteur. On peut donc former quatre modifieurs composés pour décrire ce système de ventilation, le premier adjectif qualifie l'élimination de l'air intérieur, et le deuxième l'arrivée de l'air extérieur :

[3] Les deux opérations sont évidemment liées : pour que de l'air extérieur puisse entrer, il faut que le gaz situé à l'intérieur du local sorte.

système (naturel/naturel + naturel/mécanique[4] *+ mécanique/naturel + mécanique/mécanique).*

Cette description des phénomènes est didactique mais les textes du corpus sont rédigés par plusieurs auteurs dont les objectifs varient selon les textes. On peut donc trouver des expressions moins précises et qui ne correspondent pas vraiment à la description plus rigoureuse qui peut être faite dans une autre partie du corpus. En particulier, la relation d'inclusion entre les différents classifieurs n'est pas toujours prise en compte et on pourra trouver les *GN* suivants :

(système + appareil + dispositif) de (désenfumage + ventilation)

a) Pour assurer le contrôle des fumées :

Le système complet est appelé *système de désenfumage* ou *de contrôle des fumées*. Le gaz éliminé est essentiellement composé de fumées. On trouve aussi bien *dispositif*, *appareil* que *système* comme classifieurs. Les *GN* sont donc les suivants :

système	*(de désenfumage + de contrôle des fumées)*
(système + appareil + dispositif)	*pour l'évacuation des fumées*
	d'évacuation des fumées

Le dispositif qui séparerait les fumées de l'air respirable serait parfaitement efficace, mais il n'existe pas ; le modifieur *fumées* est donc souvent remplacé par *air et fumées* (le modifieur *fumées et air* sémantiquement équivalent n'est pas attesté). Les *GN* deviennent :

(système + appareil + dispositif)	*pour l'évacuation de l'air et des fumées*
	d'évacuation de l'air et des fumées

Si l'élimination des fumées se fait à l'aide d'un moteur, le terme consacré est *extraction* ; le classifieur *appareil* n'est plus du tout utilisé à ce niveau de description, on obtient donc :

(système + dispositif)	*pour l'extraction (des fumées + de l'air et des fumées)*
	d'extraction (des fumées + de l'air et des fumées)

Bien que la combinaison *N A* soit redondante dans ce cas, *extraction* peut être accompagné de *forcée* ou *mécanique* :

(système + dispositif)	*pour l'extraction (forcée+mécanique) (des fumées+de l'air et des fumées)*
	d'extraction (forcée + mécanique) (des fumées + de l'air et des fumées)

L'air extérieur qui vient prendre la place de l'air éliminé s'appelle *air frais*. Cette opération se fait grâce à un *dispositif d'amenée d'air*. Le classifieur *système* n'est pratiquement pas utilisé avec ce modifieur ; et *dispositif* est le plus souvent effacé devant *amenée d'air*. Sur les 92 occurrences de l'expression *(dispositif d' + E) amenée d'air*, aucune n'est complétée par le modifieur *frais*. L'expression *amenée d'air frais* semble donc considérée comme redondante. Le *GN* peut néanmoins être accompagné du modifieur *mécanique* ou *naturelle* selon la technologie utilisée :

*dispositif d'amenée d'air (E + *frais)*
(dispositif d' + E) amenée d'air (E + naturelle + mécanique)

L'objet concret qui assure l'amenée d'air est un *ouvrant* ; le *GN amenée d'air* est aussi employé pour désigner cet objet :

---

[4] Ce système n'existe pas pour les ERP et n'est donc pas mentionné dans les textes.

un ouvrant (E + d'amenée d'air)
= *une amenée d'air*

b) Pour le contrôle des conditions d'hygiène concernant l'air respiré en situation normale :

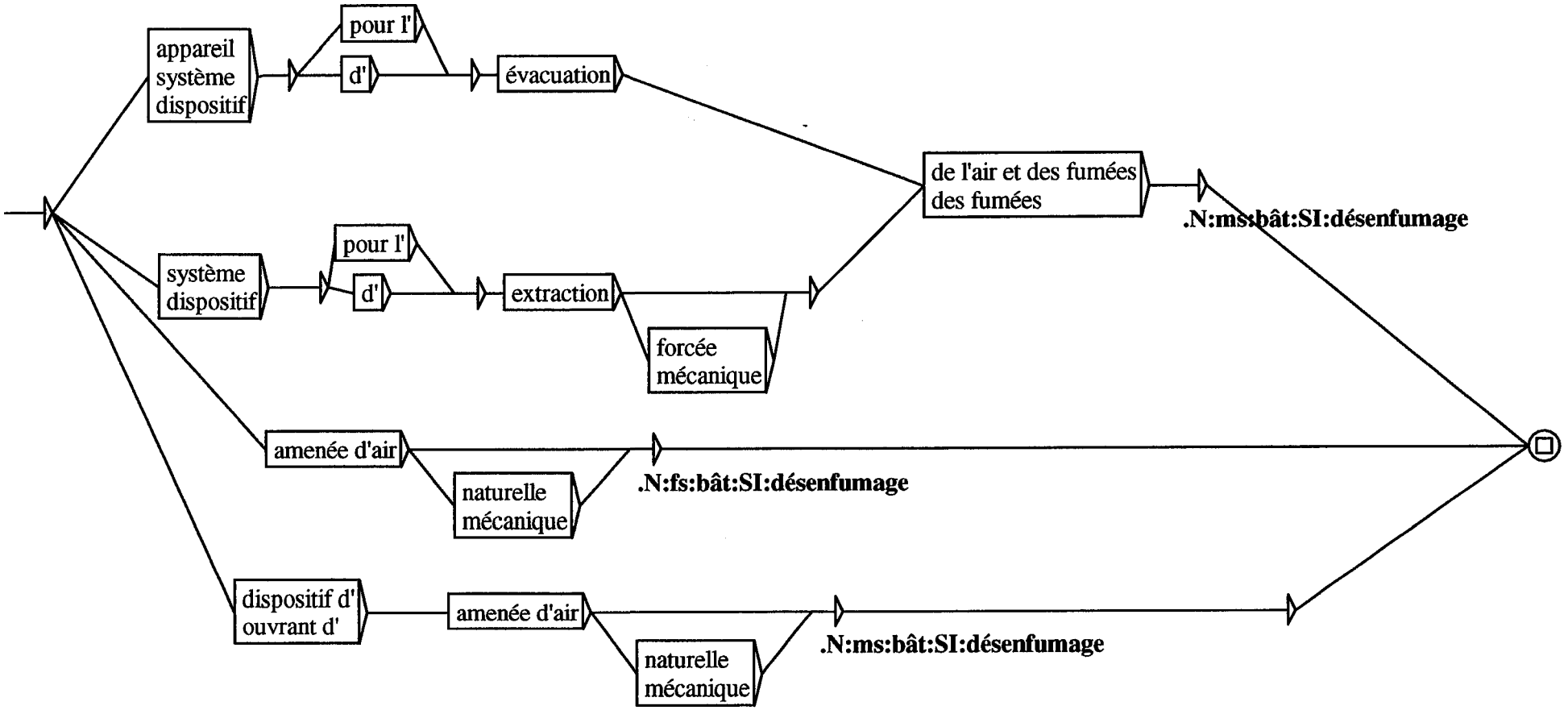
Le système ou dispositif mis en place a seulement pour but de remplacer l'air d'un local par de l'air venant de l'extérieur. Dans ce cas, l'air à remplacer est qualifié d'*air vicié*. On a donc les *GN* suivants :

(système + dispositif) *pour l'extraction (d' + de l') air vicié*
d'extraction (d' + de l') air vicié

Le *système naturel/mécanique* destiné à renouveler l'air vicié s'appelle une *ventilation mécanique contrôlée* abrégé en *VMC*.

Les règles de formation des *GN* concernant le désenfumage et la ventilation sont regroupées dans une grammaire locale, le dictionnaire généré est présenté en annexe.

Enfin, on n'a pas trouvé dans le corpus de séquences *N Mod* où *N* se dérive en *appareil, dispositif, installation, système* et pour lesquelles *Mod* prend une acception différente selon le classifieur.


appareil
système
dispositif
pour l'
d'
évacuation
de l'air et des fumées
des fumées
.N:ms:bât:SI:désenfumage
système
dispositif
pour l'
d'
extraction
forcée
mécanique
amenée d'air
naturelle
mécanique
.N:fs:bât:SI:désenfumage
dispositif d'
ouvrant d'
amenée d'air
naturelle
mécanique
.N:ms:bât:SI:désenfumage

Dans ce paragraphe, on a étudié des modifieurs synonymes et on a vérifié que cette synonymie ne dépendait pas du classifieur utilisé. D'autre part, on s'est intéressé à des cas particuliers : on a relevé les expressions liées au chauffage et à la cuisson construits sur le classifieur *appareil*. Ces expressions sont consacrées par l'usage et l'échange de *appareil* avec un autre classifieur paraît difficile dans le contexte, même s'il produit des *GN* tout à fait intelligibles. Pour le classifieur *système*, deux domaines d'utilisation ont été étudiés : le contrôle de l'air ambiant et celui des fumées en cas d'incendie, qui conduisent à former des expressions spécifiques. On peut conclure à partir de ces éléments que, hormis dans les contextes que l'on a évoqués et pour lesquels les expressions n'ont pas vraiment un sens compositionnel, les modifieurs sont utilisables avec plusieurs, voire la totalité des classifieurs, et ne changent pas d'acception selon ce classifieur.

2.5.5. Permutation des modifieurs

(49) a. *dispositif de protection*
(49) b. *dispositif spécial éventuel de protection*
(49) c. *éventuel dispositif spécial de protection*
(49) d. *éventuel dispositif de protection spécial*

(53) a. *dispositif dynamique de recoupement*
(53) b. *dispositif de recoupement dynamique*
(53) c. *système (dynamique + statique + de recoupement)*

(58) a. *dispositif automatique de freinage*
(58) b. *dispositif de freinage automatique*

(59) a. *dispositif déclencheur ou obturateur de sécurité*
(59) b. *dispositif obturateur ou déclencheur de sécurité*
(59) c. *dispositif de sécurité déclencheur ou obturateur*
(59) d. *dispositif de sécurité obturateur ou déclencheur*

Avec la série (49), on voit des *GN* qui contiennent trois modifieurs dont l'ordre d'apparition dans l'expression complète peut varier, même si seule la forme *b* est observée dans les textes.
Dans les séries d'exemples (53) et (58), on énonce des séquences *N-classifieur de N A* et *N-classifieur A de N* qui désignent le même objet et qui diffèrent seulement par l'ordre des modifieurs. Avec (53a-b), aucune des expressions *dispositif dynamique* ni *dispositif de recoupement* n'est attestée par les textes, alors que l'expression complète *dispositif dynamique de recoupement* existe. Avec (53 c) au contraire, on montre l'existence pour le classifieur *système* des deux séquences *NA* et de l'expression *N de N* sans que les séquences complètes *NA de N* ni *N de NA,* qui désignent le même objet, n'existent dans le corpus même si elles paraissent tout à fait acceptables.
Dans (59a), le premier modifieur résulte d'une coordination entre les deux noms (ou adjectifs) *déclencheur* et *obturateur* et cet exemple est le seul attesté par le corpus sans qu'aucun des trois autres désignant le même objet ne paraissent inacceptables.
Dans tous les cas, la possibilité de coordonner, juxtaposer et permuter les modifieurs ne semble pas liée au choix du nom classifieur.

2.5.6. Coordination des modifieurs

(90) *appareil de commande et de signalisation*
(91) *appareil à combustible solide ou liquide*
(92) *dispositif de commande ou d'alimentation*

(93) *dispositif de commande ou de protection*
(94) *installation de gaz combustible et d'hydrocarbures liquéfiés*
(95) *système préfabriqué ou industrialisé*
(96) *système d'extraction forcée de l'air et des fumées*

Ce phénomène de coordination des modifieurs est très courant dans le domaine technique où l'objet désigné par le nom tête assure plusieurs fonctions ou caractéristiques indiquées chacune par un modifieur, ce qui se traduit par les patrons suivants[5]: *N-classifieur de N1 (et + ou) (de + E) N2* et *N-classifieur A1 (et + ou) A2*. Le corpus en fournit de nombreux exemples. Dans (90) et (95), ce sont les modifieurs, adjectifs *A1* et *A2* ou compléments de nom *N1* et *N2*, qui sont coordonnés ; dans (91), les deux modifieurs coordonnés se rapportent au premier modifieur *combustible*. La coordination peut aussi se faire par *ou* comme en (92) et (93). En (94), chacun des modifieurs coordonnés *gaz* et *hydrocarbures* est accompagné de son propre modifieur, ce qui donne à l'expression complète une construction symétrique. En (96), le deuxième modifieur *forcée* se rapporte au premier *extraction*, ainsi que le troisième obtenu par coordination de deux noms *air* et *fumées*.

installations de chauffage et de cuisson
installations du chauffage ou de cuisson

Les expressions précédentes contiennent les mêmes modifieurs coordonnés mais les articles et conjonctions de coordination diffèrent sans que le sens de l'expression globale en soit significativement modifié.

appareil à combustible solide, liquide ou gazeux
les appareils d'éclairage, de projection, de sonorisation
les installations de gaz, d'éclairage, de chauffage et de secours contre l'incendie

Les modifieurs peuvent aussi être regroupés dans une énumération. Les règles de reconstruction des *GN* non elliptiques construits sur les nom têtes *appareil*, *dispositif*, *installation*, *système* (qui permettraient d'obtenir par exemple les expressions *appareil à combustible solide*, *appareil à combustible liquide* ou *appareil à combustible gazeux*) sont les mêmes que celles concernant tous les *GN*, et sont développées dans le chapitre concernant la coordination des modifieurs droits à l'intérieur des *GN*.

2.5.7. Abréviation des modifieurs

Comme on l'a vu dans le § précédent, il est très courant qu'un *GN* compte plusieurs modifieurs associés à un nom tête. Dans ce cas, à la première évocation de l'objet, l'auteur citera chacune des propriétés de l'objet qui se traduit par l'ajout d'un modifieur, juxtaposé ou coordonné aux précédents. Mais dans la suite de l'exposé, il se contentera de désigner l'objet avec un minimum de modifieurs tout en évitant l'ambiguïté. On peut donc observer dans les textes différents *GN* qui, au fil du discours, désignent tous le même objet alors que pris isolément hors contexte ils nomment des objets avec des propriétés différentes.

(67) *Lorsque la protection contre les contacts indirects est assurée par des dispositifs de protection à courant différentiel résiduel, il est admis de regrouper les circuits d'éclairage des locaux accessibles au public de façon à n'utiliser pour ces locaux que deux dispositifs de protection*

[5] Un nom tête peut aussi être suivi de plusieurs modifieurs de constructions différentes : complément de nom, adjectif, ..., et qui ne sont pas coordonnés mais juxtaposés.

différentiels tout en respectant, dans les locaux pouvant recevoir plus de cinquante personnes, la règle générale de l'alinéa ci-dessus.

(68) *Les installations destinées à assurer l'extraction mécanique de l'air vicié des locaux (système de ventilation courante ou inversée) doivent être conçues de manière à éviter la propagation du feu et des fumées dans tout local autre que celui où le feu a pris naissance. Les systèmes dans lesquels les débits de soufflage sont limités à 200 mètres cubes par heure par local sont considérés comme des systèmes à double flux.*

Dans les deux exemples précédents, l'unité de texte est le paragraphe. En début de §, pour l'exemple (67), l'objet est nommé avec la totalité de ses modifieurs : *dispositifs de protection à courant différentiel résiduel*. Plus loin dans le §, l'auteur doit en préciser le nombre : le déterminant indéfini est remplacé par un numéral *deux*, le nom tête et le premier modifieur sont conservés : *dispositifs de protection* et le modifieur *différentiel* qui se rapportait au premier modifieur *de protection* est maintenant raccroché au nom tête *dispositifs* (comme en témoigne son accord) grâce à un mécanisme qu'on a déjà détaillé au § 2.5.3.
Pour (68), le processus est le même. Le nom est complet au début du § : *système de ventilation courante ou inversée* puis tronqué mais en utilisant un déterminant défini pour faire référence aux objets nommés précédemment : *les systèmes*, au lieu de *les systèmes de ventilation courante ou inversée*. Dans la dernière phrase qui constitue une définition l'expression *systèmes à double flux* vaut en fait pour *systèmes de ventilation courante ou inversée à double flux* : le modifieur initial n'est pas repris, il est implicite.

appareil de secours (contre l'incendie+E)
dispositif de secours (contre l'incendie+E)

On a relevé dans le corpus une abréviation qui pourrait être ambiguë. En effet, le modifieur *de secours*, surtout quand il est employé dans un *GN* dont la tête désigne un accessoire automobile, signifie *de rechange* ; on trouve les expressions : *roue de secours, pare-brise de secours.* Dans le corpus, il est parfois l'abréviation de *de secours contre l'incendie*. On notera aussi que l'expression tronquée n'est jamais, dans les textes relatifs à la sécurité incendie, synonyme de *de rechange*. Enfin des expressions comme *porte de secours, issue de secours, sortie de secours* s'emploient toujours telles quelles et sont en fait équivalentes aux expressions complètes et non employées : *porte de secours contre l'incendie, issue de secours contre l'incendie, sortie de secours contre l'incendie.*

Dans tous les cas, cette possibilité d'abréger un *GN* contenant plusieurs modifieurs en une séquence en contenant un seul, dépend de la liste des modifieurs et non pas du classifieur utilisé.

2.6. Coordination des noms têtes

(80) *dispositifs et commandes de sécurité*
(81) *installation ou matériel électriques*
(82) *appareils et installations fixes*
appareils ou dispositifs d'extinction et d'alerte
(83) *dispositifs ou systèmes banalisés*

Tout comme les modifieurs, les noms têtes peuvent aussi être coordonnés avec d'autres noms appartenant à notre liste de classifieurs comme dans les exemples (82) à (83). Les classifieurs étudiés ne sont donc pas interchangeables, au moins pour ces expressions. Mais ils acceptent aussi la proximité d'autres types de noms comme en (80) et (81). Les possibilités de coordination comme les règles de reconstruction des expressions non elliptiques semblent être les mêmes que pour le cas général.

2.7. Séquences contenant plusieurs unités lexicales autonomes

(70) *appareil ménager à butane-propane*
(71) *dispositif extérieur d'arrêt de l'admission du combustible gazeux ou liquide*
(72) *dispositif sonore à commande manuelle ou automatique*
(49) b. *dispositif spécial éventuel de protection*
(73) *système d'extinction automatique à eau*
(74) *système d'extraction forcée de l'air et des fumées*

On l'a vu : les expressions formées sur les classifieurs étudiés comportent souvent plus d'un modifieur. Mais les noms têtes peuvent aussi admettre des modifieurs. Ces séquences *N Mod* peuvent être autonomes dans la mesure où elles existent en dehors des expressions étudiées ici : elles ont déjà été recensées dans le corpus et désignent une entité bien réelle. Enfin, employées dans des expressions de construction plus complexe, ces séquences autonomes peuvent ou bien se juxtaposer ou bien se mélanger au reste du *GN*.
Pour toutes ces raisons, on obtient ainsi des expressions très longues qu'on ne sait pas toujours comment segmenter ; en particulier, la délimitation du nom tête n'est pas toujours certaine parce que plusieurs possibilités se présentent.

Dans (70), *appareil ménager* est un nom composé, tout comme *appareil à butane-propane.* Deux segmentations sont possibles : le nom tête *appareil ménager* admet un modifieur *à butane-propane*, ou bien le nom tête *appareil à butane-propane* est interrompu par le modifieur *ménager.* Puisque la construction :

N à (Dét+E) Combustible
avec *Combustible ::- butane + propane +gaz + combustible (solide+liquide+gazeux) + ...*

est productive, on considèrera que *appareil ménager* est le nom tête.
La construction de l'expression (71) est complexe : *combustible gazeux ou liquide* peut être considéré comme un nom composé, il en est de même pour *dispositif d'arrêt* ; l'adjectif *extérieur* est une insertion dans le nom composé tête ; *de l'admission* est un modifieur qui s'applique à *arrêt* et admet lui-même un modifieur *du combustible gazeux ou liquide.* Et dans ce cas, le nom tête est *dispositif d'arrêt.* Une autre analyse possible reviendrait à segmenter le *GN* en une tête *dispositif extérieur* suivie de trois compléments de nom : *d'arrêt, de l'admission* et *du combustible gazeux ou liquide* qui modifient chacun le modifieur précédent. Cette analyse est sûrement moins pertinente parce que *dispositif extérieur* est moins précis que *dispositif d'arrêt* et l'analyse de *extérieur*, qui peut s'appliquer à différents types de dispositifs, comme un modifieur inséré à l'intérieur d'un nom composé paraît plus légitime.
Dans le *GN* (72), on peut reconnaître les deux noms composées *dispositif sonore* et *dispositif à commande manuelle ou automatique*, d'où un arbitrage difficile dans le choix du nom tête : *dispositif, dispositif sonore* ou *dispositif à commande manuelle ou automatique.*
Les difficultés sont les mêmes avec l'exemple (73) formé à partir de plusieurs noms composés : *système d'extinction, système automatique* et *système à eau.*
Pour (74), dans le contexte de la sécurité incendie, l'air et les fumées constituent un ensemble de gaz qu'il faut traiter (extraire, aspirer, filtrer, régénérer, ...) de la même façon ; il paraît donc justifié de considérer que l'expression coordonnée *air et fumées* constitue un nom composé. *Système d'extraction forcée* est aussi un nom composé formé sur le schéma productif : *système de N* où *N* est le nom composé *extraction forcée* de la forme *N A*. L'expression (74) peut donc être considérée comme un nom composé de la forme *N1 de N2 Mod* où *N2* et *Mod* sont eux-mêmes des noms composés.

Les expressions contenant plusieurs unités lexicales autonomes admettent aussi des insertions (les modifieurs rajoutés sont soulignés) :

(75) *installation d'extinction*
(75) a. *installation d'extinction à eau*
(75) b. *installation fixe d'extinction à eau*
(75) c. *installation d'extinction automatique*
(75) d. *installation fixe d'extinction automatique*
(75) e. *installation d'extinction automatique à eau*
(75) f. *installation fixe d'extinction automatique à eau*
(75) g. *installation d'extinction automatique ou à commande manuelle*

En partant de (75) de schéma *installation de N*, on peut rajouter un modifieur en queue de *GN*, sans que l'on sache d'ailleurs très bien s'il se rapporte au nom tête *installation* ou au modifieur *N*, comme en (75a-c), puis en réitérant l'opération on obtient les exemples (75e-g). Dans ces expressions contenant déjà plusieurs unités autonomes, on insère l'adjectif *fixe*, en tant que modifieur du nom tête pour obtenir (75b-d-f). A chaque niveau de construction, l'expression désigne un objet unique auquel sont associées autant de propriétés que de modifieurs.

Lorsqu'on rajoute un ou plusieurs modifieurs, les prépositions de la construction initiale peuvent être modifiées :

(76) *a.* *systèmes à double flux*
(76) *b.* *systèmes simple ou double flux*
(76) *c* *systèmes de ventilation courante ou inversée, simple ou double flux*

Quand le modifieur de (76a) est intégré à un modifieur plus complexe, la préposition *à* est effacée et le modifieur coordonné *simple ou double flux* est juxtaposée au classifieur. Lorsqu'on rajoute le modifieur composée *de ventilation courante ou inversée* au *GN* (76b) la préposition *à* n'est pas réintroduite et le nouveau *GN* contient deux modifieurs de structures complexes, séparés par une virgule. On pourrait appliquer les mêmes règles de construction aux autres classifieurs :

(système + dispositif + appareil) à double flux
(système + dispositif + appareil) simple ou double flux
(système + dispositif + appareil) de ventilation courante ou inversée, simple ou double flux

2.8. Effacement du classifieur

(30) *un (appareil + E) détecteur enregistreur*
(31) *un (système + E) (obturateur + porte-éprouvette)*
(49) *a.* *un dispositif de protection*
(49) *d.* *une protection*
(77) a. *(un appareil+un dispositif+une installation+un système) de ventilation*
(77) b. *une ventilation*
(78) a. *un dispositif d'amenée d'air*
(78) b. *une amenée d'air*
(44) *d.* *(un dispositif + une installation + un système) de désenfumage*
(44) *e.* *un désenfumage*

Dans chacune des paires précédentes, il est possible d'effacer le classifieur sans changer le sens de l'expression initiale. Ces expressions sont de constructions *N-classifieur A* et *N-classifieur de N*. On observe que dans chaque cas, on peut énoncer la propriété suivante : *N-classifieur est un N* et aussi *N est un N-classifieur*. Par exemple, pour (30), un *appareil détecteur enregistreur est un détecteur enregistreur* et un *appareil détecteur enregistreur est un appareil*, et pour (77 a) un *appareil de ventilation* est une *ventilation*, de même qu'un *dispositif*, une *installation* ou un *système*. Quand le classifieur est effacé, le déterminant s'accorde alors avec *N* et peut donc changer de genre comme dans les paires (49) et (78).

*un (dispositif + *E) luminescent*
*une (installation + *E) stable*
*un (système + *E) statique*
*un (appareil + *E) respiratoire*

Les *GN* précédents sont de construction *N-classifieur A*, le modifieur est un adjectif et donc la propriété *N-classifieur est un N* ne peut, par définition, être valide puisque *A* n'est pas un *N-classifieur* mais désigne une propriété associée à *N-classifieur*. Il est donc impossible d'effacer le modifieur, quelle que soit sa valeur.

(41) *a.* *appareil à combustible solide*
dispositif à courant différentiel résiduel
installation au gaz
système à panneaux

Dans ces expressions de construction *N-classifieur à N*, *N-classifieur* n'est pas un *N* : par exemple, l'*installation* n'est pas du *gaz*. L'effacement du classifieur est impossible.

(70) *appareil ménager à butane-propane*
(70) a. *appareil à butane-propane*
(70) b. *un butane-propane*

L'exemple (70) semble contredire la justification précédente puisque, bien que le modifieur soit introduit par la préposition *à*, l'effacement du classifieur *appareil* paraît acceptable. En fait existe déjà, en manière d'abréviation, dans le langage courant la possibilité d'effacer le nom classifieur dans (70 a) pour donner (70b).

Pour les deux séries d'exemples qui suivent, on a souligné le modifieur inséré :

(71) *un dispositif extérieur d'arrêt de l'admission du combustible gazeux ou liquide*
(71) a. *un arrêt extérieur de l'admission du combustible gazeux ou liquide*
(49) b. *un dispositif spécial éventuel de protection*
(49) c. *une protection spéciale éventuelle*
(73) *un système d'extinction automatique à eau*
(73) a. *une extinction automatique à eau*
(152) *(un dispositif+une installation+un système) automatique de désenfumage*
(152) a. *un désenfumage automatique*

On observe que l'insertion d'un adjectif se rapportant au nom classifieur n'est pas un obstacle à l'effacement de ce classifieur, s'il était déjà possible de l'effacer dans la construction de base. L'adjectif inséré se rapporte, après l'effacement, à la nouvelle tête du *GN,* et en prend donc les caractéristiques en genre et en nombre.

(74) *un système d'extraction forcée de l'air et des fumées*
(74) a. *une extraction forcée de l'air et des fumées*
(42) *un dispositif de commande accompagnée (fonctionnant à l'aide d'une + à) clé*
(42) a. *une commande accompagnée (fonctionnant à l'aide d'une + à) clé*
(56) a. *un dispositif de ventilation mécanique ou naturelle*
(56) b. *une ventilation mécanique ou naturelle*

Que le modifieur admette lui-même la présence d'un modifieur, quelle qu'en soit sa construction, ne constitue pas non plus un obstacle à l'effacement du nom classifieur.

A la fin de ce paragraphe, on peut donc conclure que si l'effacement est possible, il peut se faire quel que soit le nom classifieur. Cette propriété est donc liée au modifieur et non au choix du classifieur.

2.9. Cas particulier du classifieur *installation*

(80) *installation électrique*
(81) *installation d'alarme et d'alerte*
(82) *installation au gaz*
(83) *installation de détection*

(84) *installation de locaux accessibles aux élèves*
(85) *installation de paratonnerres*
(86) *installation de systèmes de détection*
(87) *installation de gaz*

Installation, comme la plupart des noms en *-tion* formés sur des verbes, peut désigner à la fois l'action exprimée par le verbe et le résultat de cette action. Cette ambiguïté se retrouve aussi dans les textes techniques : pour les exemples (80) à (83), *installation* désigne le résultat de l'action d'installer qui prendra la forme concrète d'un objet, d'un assemblage, ... alors que dans les expressions (84) à (87) *installation* représente l'action d'installer. Cette caractéristique est propre à *installation* et ne se retrouve pas chez les autres classifieurs étudiés : *appareil*, *dispositif*, *système*. Ceci explique aussi la différence entre le nombre d'occurrences d'*installation* dans le texte (986 occurrences : l'effectif le plus important parmi les quatre classifieurs étudiés) et le nombre proportionnellement restreint d'expressions composées retenues (142).

2.10. Intersections des emplois pour les modifieurs

Le corpus étudié atteste la possibilité d'ajouter certains modifieurs à n'importe lequel des quatre classifieurs en conservant l'unicité de l'objet désigné.

(appareil + dispositif + installation + système)	*de contrôle*
	de sécurité
	de sûreté
	de ventilation mécanique contrôlée
	de ventilation mécanique inversée
	d'éclairage
	d'éclairage de sécurité
	d'éclairage électrique

On a aussi recensé les modifieurs partagés par trois ou deux classifieurs, ainsi que ceux qui accompagnent exclusivement un seul modifieur.

Modifieurs partagés par *appareil*, *dispositif* et *installation*

(appareil + dispositif + installation)	*de contrôle*

Modifieurs partagés par *appareil*, *dispositif* et *système*

(appareil + dispositif + système)	*(d' + pour l') évacuation des fumées*
	de conditionnement d'air
	de coupure
	de désenfumage
	de mesure de la distance

de réglage
de ventilation

Modifieurs partagés par *appareil*, *installation* et *système*

(appareil + installation + système)

d'alimentation électrique
de chauffage

Modifieurs partagés par *dispositif*, *installation* et *système*

(dispositif + installation + système)

d'alarme
d'alerte
d'appel des secours
de commande
de désenfumage
de détection (E+automatique+automatique d'incendie)
d'extinction automatique (E + à eau)

Modifieurs partagés par deux classifieurs

(appareil + dispositif) /

(appareil + installation) *de cuisson*[6]

(appareil + système) /

(dispositif + installation)

technique
thermique

(dispositif + système)

d'accrochage
de fermeture (E + automatique)
de fixation
(d'+pour l') extraction (E+forcé+mécanique)
(des fumées +de l'air et des fumées)
d'intercommunication
d'obturation
d'ouverture
mécanique

(installation + système) /

Modifieurs exclusifs

pour appareil

appareil à (circuit étanche + combustible liquide + combustion directe)
appareil à (gaz + vapeur)
appareil d'utilisation du gaz
appareil (portatif + respiratoire)

---

[6] On trouve aussi l'expression *installation d'appareils de cuisson* qui ne diffère pas vraiment, sémantiquement, des deux précédentes.

pour installation

installation automatique
installation alimentant l'éclairage de sécurité
installation frigorifique
installation (neuve + normale + nouvelle + sensible)

pour dispositif

dispositif automatique
dispositif (d'accès + d'accès et de sortie)
dispositif de vidange et de purge d'air
dispositif d'inflammation

pour système

système constructif
système de montage
système de pinces
système de recoupement
système de réglage
système d'irrigation
système d'isolation

Ajout de combinaisons aux dictionnaires

dispositif d'attache
dispositif d'assemblage
(système+ dispositif) d'accrochage

On a des modifieurs qui ne sont attestés qu'avec un seul classifieur, mais qui peuvent être regroupés dans une même "famille sémantique". Dans ce cas, il paraît légitime de combiner tous ces modifieurs avec chacun des modifieurs attestés dans au moins une occurrence. Par exemple, pour les trois *GN* précédents, *attache, assemblage* et *accrochage* procèdent de l'idée de couplage entre deux éléments. Les deux premiers exemples sont construits avec *dispositif*, le troisième à partir de *système* et *dispositif* mais il n'y a pas de raison syntaxique ni sémantique de refuser les expressions *système (d'attache + d'assemblage).* Ces deux expressions sont donc rajoutées aux dictionnaires au même titre que les quatre autres attestées par le corpus.

2.11. Les expressions "figées"

Comme on l'a vu dans le paragraphe précédent, le corpus atteste des modifieurs qui peuvent être employés avec plusieurs des noms classifieurs étudiés sans que l'expression composée désigne un objet différent. On ne trouvera donc pas d'expressions figées construites à partir de ce groupe de modifieurs. Les expressions candidates sont celles pour lesquelles le modifieur n'est pas employé avec un autre classifieur. On considère aussi que les séquences *N A* et *N Prép N* qui figurent comme telles dans les dictionnaires utilisés et bénéficient d'une définition propre constituent des expressions figées. D'autre part, les expressions dont les initiales forment un acronyme que l'on trouve comme tel dans les textes sont de fait considérées comme des expressions figées.

Expressions figées construites sur ***installation***

installation classée
installation dangereuse
installation frigorifique
installation sportive
installation temporaire

Expressions figées construites sur ***dispositif***

dispositif actionné de sécurité (DAS)
dispositif adaptateur de commande (DAC)
dispositif d'essai
dispositif mécanique

Expressions figées construites sur ***système***

système constructif
système de construction
système de détection incendie (SDI)
système de forces
système de montage
système de sécurité incendie (SSI)
système rigide
système statique

Expressions figées construites sur ***appareil***

Le terme *appareil* est extrêmement vague et par conséquent se prête mal à la construction d'expressions figées. Néanmoins, malgré ce manque de précision, certaines expressions rencontrées dans le corpus, si elles ne sont pas figées, semblent du moins consacrées par l'usage pour nommer des assemblages souvent imprécis, dont les détails de construction ne sont pas connus du locuteur, mais qui sont explicites par leur fonction ou leur destination, indiquée par le modifieur. On peut retenir, avec cette définition du figement, les expressions suivantes :

appareil de chauffage (électrique + E)
appareil de cuisson (électrique + à gaz + E)
appareil de projection (cinématographique + E)
appareil d'eau chaude (sanitaire + E)
appareil électrique
appareil frigorifique
appareil ménager
appareil sanitaire
appareil téléphonique

2.12. Conclusions

Le corpus atteste l'utilisation d'expressions composées à partir des quatre noms classifieurs *appareil, dispositif, installation, système*. Les modifieurs relevés dans les textes se répartissent en plusieurs classes : ceux qui sont communs aux quatre classifieurs, ceux qui s'emploient avec trois, deux ou bien un seul classifieur. On remarque aussi que certaines classes sont vides : par exemple il n'y a pas de modifieurs communs seulement aux classifieurs *appareil* et *système* ou *dispositif* et *installation.*

A priori et si on excepte les expressions figées, l'utilisation d'un classifieur plutôt qu'un autre paraît avoir peu de signification par rapport aux objets désignés : par exemple, comme on l'a vu dans le § 2.10., on trouve dans les textes *(dispositif + système) d'accrochage* et *dispositif (d'assemblage + d'attache)* mais les expressions *système (d'assemblage + d'attache)* ne sont pas attestées alors qu'elles paraissent strictement équivalentes. Pourtant l'étude détaillée de ces classifieurs et des modifieurs qui leur sont accolés révèlent une réalité plus complexe ; en particulier, il paraît difficile, alors qu'on pouvait l'imaginer en première analyse, de combiner la totalité des modifieurs avec chacun des classifieurs pour composer les entrées des dictionnaires de noms composés spécifiques à la sécurité incendie.

L'étude précédente permet de dégager deux idées qui ont des conséquences dans la construction des dictionnaires :

- les propriétés étudiées (utilisation d'une paraphrase à la place de la préposition ; possibilité d'insérer un modifieur ; synonymie, inversion et coordination des modifieurs ; abréviation des *GN* ; contraintes orthographiques) pour les expressions de constructions *N-classifieur Prép N*, *N-classifieur N* ou *N-classifieur (A + V:K)* ne varient pas selon le classifieur, mais plutôt selon le modifieur.

- si un modifieur intervient dans une expression *N-classifieur Mod*, les autres expressions construites avec le même modifieur et en changeant de classifieur ne désigneront pas un objet concret différent.

Il paraît donc légitime de ne pas se cantonner, dans les dictionnaires, aux seules entrées relevées dans les textes, mais il faut définir les règles qui permettent de fabriquer les entrées du dictionnaire à partir des expressions attestées par le corpus.
Selon les définitions des dictionnaires, ces classifieurs respectent une hiérarchie en termes d'inclusion : *système* et *installation* sont "plus larges" que *appareil* et *dispositif*. On pourrait donc considérer que, sauf pour les expressions figées, *système* et *installation* d'une part et *appareil* et *dispositif* d'autre part, représentent le même type de classifieur. Ainsi chaque modifieur attesté avec un classifieur des éléments d'une paire pourrait peut se combiner avec l'autre. En fait, à cause de la double signification de *installation*, cette démarche conduit à des expressions bizarres voire inacceptables :

*(installation+*appareil+?*dispositif+?*système) d'ascenseurs*
*(installation+*appareil+?*dispositif+?*système) de paratonnerres*

On pourrait reconduire le même raisonnement aux trois autres classifieurs, on obtient là-aussi des expressions difficilement acceptables :

*(?*appareil+dispositif+système) d'ancrage*
*(?*appareil+dispositif+système) d'irrigation*

Si on se limite au rapprochement de seulement deux modifieurs, il faut évaluer "les affinités", mesurées deux à deux, entre *appareil*, *dispositif* et *système*. En observant les listes de modifieurs, on peut faire les observations suivantes :

nombre de modifieurs partagés par appareil et dispositif : 16
nombre de modifieurs partagés par appareil et système : 17
nombre de modifieurs partagés par dispositif et système : 31

On conclut donc que, dans les textes, ce sont *dispositif* et *système* qui partagent le plus grand nombre de modifieurs. Et on extrapole en disant que *dispositif* et *système* représentent les deux occurrences d'un même classifieur. On considère alors que tous les modifieurs de l'un peuvent se combiner avec l'autre.

On adopte donc, pour construire les dictionnaires de mots composés propres au domaine de la sécurité incendie, les règles suivantes :

- on conserve telles quelles les expressions qu'on a considérées comme figées dans les § 2.6, 2.3.7.1 et 2.3.7.2.

- on tient compte dans les dictionnaires de la possibilité pour la préposition qui introduit le modifieur d'être remplacée par une paraphrase (cf. § 2.3.4)

- on tient compte dans les dictionnaires de la possibilité d'insertion à l'intérieur d'une séquence composée (cf. § 2.7). Ces insertions peuvent prendre différentes formes :
  . un autre modifieur qui se rapporte au modifieur initial, et placé à sa gauche ;
  . un autre modifieur qui se rapporte au nom tête et qui pourra se trouver immédiatement à sa gauche ou à sa droite, ou bien rejeté en fin de *GN* ;
  . un adverbe quand le modifieur est un adjectif ou un participe passé.

- on tient compte des variantes orthographiques dans le cas où un modifieur peut s'accorder avec le nom tête ou un autre modifieur (cf. § 2.5.3)

- on tient compte dans la composition des expressions de l'utilisation de synonymes pour les modifieurs (cf. § 2.5.4). On ajoute, en particulier, les modifieurs obtenus par nominalisation de l'adjectif et on note les alternances *adjectif* ou *de N*

- on tient compte des possibilités d'inversion des modifieurs (cf. § 2.5.5)

- si une expression contenant plusieurs modifieurs constituant des unités autonomes existe, on décide que les expressions construites à partir d'une combinaison quelconque de ces modifieurs doivent aussi figurer dans la liste des expressions composées. Par exemple, le corpus atteste la formation des expressions :
  (appareil + dispositif + installation + système) *de ventilation mécanique contrôlée*

  on rajoute donc aussi aux dictionnaires les expressions non attestées suivantes :

  (appareil + dispositif + installation + système) *de ventilation*
  (appareil + dispositif + installation + système) *de ventilation mécanique*

- on combine tous les modifieurs communs à *système* et *dispositif* avec chacun de ces deux classifieurs.

Les principes d'élaboration des dictionnaires ont été présentés mais tous les dictionnaires n'ont pas été construits.

3. Des grammaires locales pour préciser le contexte d'utilisation

Dans cette partie, on s'intéresse aux deux noms : *classe* et *catégorie*. Dans le vocabulaire général, d'après le Robert électronique, le mot *classe* est utilisé pour regrouper des *unités présentant une caractéristique dont la valeur se situe entre certaines limites*, et dans ce sens admet un synonyme *catégorie*. Dans le domaine de la sécurité incendie, ces deux mots sont aussi utilisés pour rendre compte d'un classement qui a été attribué à l'objet qu'ils qualifient, mais les emplois de *classe* et de *catégorie* ne sont pas interchangeables. Ceux-ci renvoient en particulier à des objets, des domaines de définition, différents : par exemple, on parlera de la *classe d'un feu*, et non de sa catégorie, alors que l'expression *catégorie d'un établissement* n'est en rien synonyme de la *classe de l'établissement*. D'autre part, les valeurs que peuvent prendre la classe et la catégorie dépendent de l'objet auquel elles

sont associées : par exemple, *un équipement d'alarme est de classe 2a*, alors qu'*une installation électrique est de classe IIc* ; de même pour la *catégorie*, *un matériau est de catégorie M1* alors qu'*un système de sécurité incendie est de catégorie A*. Le paragraphe suivant a pour objectif d'étudier les contextes d'applications de ces deux types de classement propres à la sécurité incendie.

3.1. La grammaire locale de *classe*

Le terme *classe* fait référence à une mise en ordre selon des critères déterminés ; les valeurs que prennent le ou les critères mesurés indiquent le rang, le classement, dans lequel l'élément à classer est rangé. Dans cette étude, on ne s'intéressera pas aux critères qui permettent de déterminer le classement mais aux résultats de ce classement. On identifiera les éléments qui donnent matière à un classement qui se décrit avec le terme *classe* et on recherchera les différentes valeurs que peuvent prendre les classes en fonction du type de l'élément classé :

a) pour les feux et les extincteurs

la classe de feu correspondant à l'établissement
des feux de classe 34 B 1 ou B 2 au moins

extincteur à poudre polyvalente de classe minimum 5 A - 34 B
extincteur portatif de classe 21 B au moins
extincteur portatif homologué pour feux de classe 34 B
extincteurs portatifs pour feux de classe 34 B 1 ou B 2 au moins

b) pour les contraventions

les contraventions de la 5ème classe
les contraventions de la 5ème classe en récidive

c) pour les locaux

ces salles sont rangées en "classe 2"

d) pour les équipements d'alarme

Cette norme classe les équipements d'alarme en quatre types par ordre de sévérité décroissante 1, 2a ou 2b, 3 et 4

e) pour les appareils et les installations

Les appareils de la classe O doivent être protégés
Le dispositif de déviation optique éventuel doit être de la classe I ou II
Le système installé doit être de la classe III A 1
les appareils de la classe I
une installation de la classe I A
une installation de la classe II C

f) pour les emplois non spécifiques de *classe*

une classe de température
la classe d'isolation du moteur

les caractéristiques minimales de classe et d'indice de propagation

3.2. La grammaire locale de *catégorie*

Le nom *catégorie* appartient au vocabulaire courant du français. Dans le corpus technique que l'on a utilisé, il cumule les emplois ordinaires avec des acceptations techniques qui varient selon le nom concret auquel il se rapporte :

a) pour les établissements et les groupements d'établissements

La catégorie d'un établissement est une caractéristique d'un établissement recevant du public. Sa définition est précisée dans le texte : c'est l'effectif maximum du public autorisé à accéder aux locaux. Les emplois dans le texte sont les suivants :

(400) *les établissements de la 4e catégorie*
(401) *dispositions applicables aux ERP de la 5ème catégorie*
(402) *les établissements recevant du public de 1ère, 2ème, 3ème catégorie*
(403) *dispositions applicables aux établissements des quatre premières catégories*

Les catégories sont donc au nombre de cinq et désignées, le plus souvent, par un ordinal libellé en chiffres. Il y a des variantes formelles : *4e*, *4ème* ou *quatrième*. La catégorie peut être citée exactement comme en (400) et (401), à l'intérieur d'une énumération en (402) : *1ère, 2ème, 3ème catégorie*, ou comme faisant partie d'un intervalle *les quatre premières catégories* dans l'exemple (403).
Le bâtiment auquel cette catégorie se rapporte peut être désigné par le nom complet *établissement recevant du public* comme dans l'exemple (402), par le nom abrégé *établissement* : (400) et (403), ou par l'acronyme *ERP* en (401).

On choisit une forme canonique :

ERP de OrdinalCatégorie catégorie

où

OrdinalCatégorie =: première | deuxième | troisième | quatrième

à laquelle on va ramener toutes les expressions de la catégorie qui concernent un établissement ou un groupement d'établissements. Pour un *GN* comme (402), on devra obtenir l'expression développée suivante :

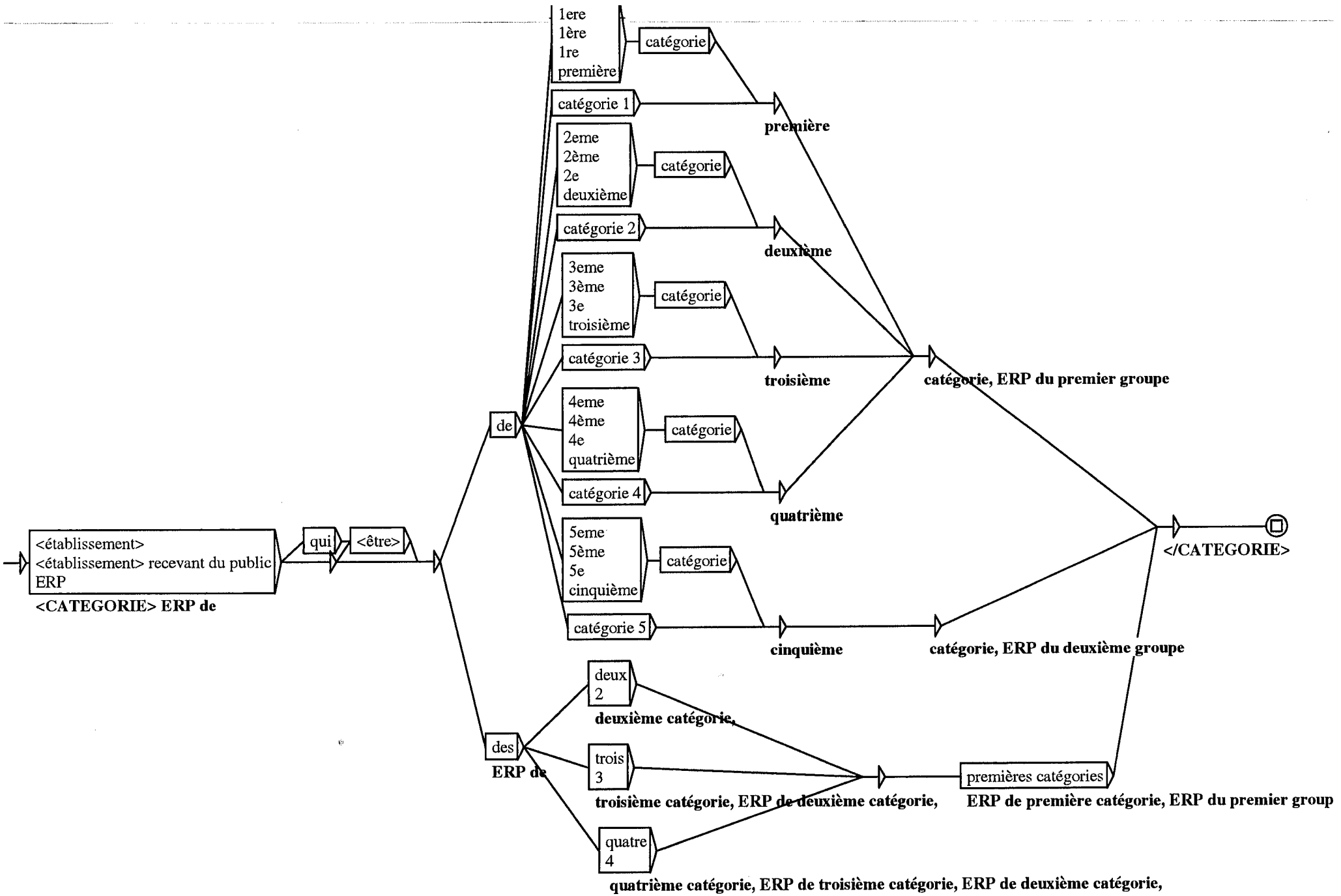
les ERP de première catégorie, ERP de deuxième catégorie, ERP de troisième catégorie, ERP de quatrième catégorie

Les valeurs de la catégorie ne rendent pas compte d'une gradation dans les exigences ou les propriétés de l'établissement. Il n'y a donc pas lieu de prévoir l'insertion d'un prédéterminant ou d'un postdéterminant numéral :

une porte stable au feu de degré 1h (E + au moins)
*un ERP de première catégorie (E + *au moins)*

A la division en catégories, se superpose celle en groupes : les quatre premières catégories constituent le premier groupe, et la cinquième, le deuxième groupe.

Toutes ces informations sont regroupées dans le transducteur suivant qui reconnaît l'expression de la catégorie d'un établissement et la transforme en une expression de forme canonique.


<établissement>
<établissement> recevant du public
ERP
<CATEGORIE> ERP de
qui
<être>
de
des
1ere
1ère
1re
première
catégorie
catégorie 1
première
2eme
2ème
2e
deuxième
catégorie
catégorie 2
deuxième
3eme
3ème
3e
troisième
catégorie
catégorie 3
troisième
4eme
4ème
4e
quatrième
catégorie
catégorie 4
quatrième
5eme
5ème
5e
cinquième
catégorie
catégorie 5
cinquième
catégorie, ERP du premier groupe
catégorie, ERP du deuxième groupe
</CATEGORIE>
deux
2
deuxième catégorie,
ERP de
trois
3
troisième catégorie, ERP de deuxième catégorie,
quatre
4
quatrième catégorie, ERP de troisième catégorie, ERP de deuxième catégorie,
premières catégories
ERP de première catégorie, ERP du premier group

b) pour les matériaux et les éléments de construction

(410) *Les matériaux sont répartis en cinq catégories : M 0, M 1, M 2, M3, M4*
(411) *Les revêtements muraux doivent être de la catégorie M1 au moins*
(412) *Les matériaux sont de catégorie M4 ou non classés*

(413) *(des aménagements scéniques + un conduit de raccordement + un écran de projection + un écran isolant + un objet de décoration + le rembourrage des sièges + un revêtement extérieur de façade + un panneau radiant) de catégorie M2*
(414) *(l'âme du matériau + la structure des sièges + l'enveloppe + les unités de traitement d'air de toiture) de catégorie M3*

D'après les exemples précédents, la notion de catégorie peut aussi être rattachée à un matériau ou un élément de construction. Dans ce cas, elle peut prendre six valeurs *M0, M1, M2 ,M3, M4* ou *non classé*. En effet, d'après (412), un matériau qui n'a pas été classé, au sens des tests de résistance au feu du CSTB, doit être rangé dans la catégorie *non classé*. On remarque que, dans ce cas, l'adjectif *non classé* s'accorde avec le nom de l'élément à classer et non avec *catégorie*.
Dans (413) et (414), on a regroupé les éléments de bâtiment et de construction auxquels est rattachée, dans les textes, une mention de catégorie. Néanmoins tous les éléments de bâtiment et de construction peuvent être classés au feu. Les entrées correspondantes dans les dictionnaires doivent porter la mention de ce sous-domaine.

Il y a des variantes lexicales : *M 1* en (410) ou *M1* en (411). Pour les codes qui comportent des zéros, il faut aussi prévoir dans les grammaires l'erreur assez fréquente qui consiste à écrire la lettre majuscule O à la place d'un zéro 0. Les libellés se comportent comme un nombre cardinal et sont placés après le nom *catégorie*.
La locution *Mi* (où *i* représente un nombre) se comporte comme un numéral et peut donc être modifiée par un prédéterminant ou postdéterminant numéral comme en (411). On peut trouver dans cette position des adverbes codés *ADV* dans les tables du lexique-grammaire :

au plus + au moins + exactement + seulement

Des adverbes comme *presque, à peu près* ou *environ* qui sont aussi codés dans cette classe ne peuvent être employés dans ce contexte où le choix de la catégorie fait l'objet de tests précis.
Les catégories sont implicitement classées en tenant compte de la relation d'ordre entre les nombres : *M4* est la catégorie la moins contraignante, et *M0* celle qui donne le plus de garanties. Les expressions comme *catégorie M1 au moins* employée en (411) permettent donc de désigner plusieurs catégories, en l'occurrence les *catégories M0 et M1*.

c) pour les systèmes de sécurité incendie

système de sécurité incendie de catégorie A, B, C, D ou E

d) pour les câbles

câbles souples de la catégorie C 2
câbles de catégorie C 1,C 3
câbles des catégories CR 1-C 1 et CR 1-C
des câbles de catégorie CR 1 classés CR 1-C 1 de catégorie CR 2
Ces câbles sont classés en deux catégories: CR 1 et CR 2

e) pour les générateurs électriques, les ventilateurs, les liquides inflammables

(440) *un générateur ou groupe de générateurs de 1re catégorie ou de deuxième catégorie*
(441) *les ventilateurs de catégorie 1, 2, 3 et 4*

(442) *les liquides inflammables de (1^{re} + 2^{e}) catégorie et alcools*
(443) *combustible liquide de (première + deuxième) catégorie*

Pour ces éléments, la catégorie est un nombre compris entre 1 et 4. Le cardinal peut se transformer en ordinal libellé en chiffres ou en lettres comme en (441). Les formulations (442) et (443) sont attestées dans le corpus : l'énumération des catégories se fait sous une forme pseudo-algébrique à l'aide du signe opératoire + et avec des libellés d'ordinaux : *(1^{re} + 2^{e}) catégorie.*

f) pour les emplois non spécifiques de *catégorie*

catégorie (d'affaires + d'ouvrages + de constructions + de fonctionnaires + de personnel)

Pour tous ces emplois, les règles d'accord sont appliquées de manière variable. On trouvera, selon les auteurs, l'un ou l'autre accord sans que l'on puisse en déduire quoi que ce soit sur la portée des déterminants numéraux.

les gares de la 1^{re}, $2^{ème}$ et $3^{ème}$(catégorie + catégories)
les ventilateurs de (catégorie + catégories) 1, 2, 3 et 4

Ces emplois doivent être précisés par des grammaires locales qui permettent d'identifier et marquer les groupes nominaux concernés, et contribuent ainsi à la segmentation de la phrase.

4. Une grammaire locale pour reformuler une information : grammaire de passage entre catégories, effectifs et types

La réglementation incendie des ERP repose sur la classification de l'établissement ou du groupement d'établissements (au sens de la définition des ERP), concerné. Ce classement est établi à partir du type d'activité de l'établissement et de l'effectif qu'il doit recevoir.
Dans une recherche documentaire automatique, une cause de silence tient à ce que la question posée, même si elle contient l'information suffisante, ne la formule pas de la même manière que dans les textes. En particulier, la question peut contenir une expression précisant l'effectif d'un établissement sans en mentionner la catégorie :

Une salle polyvalente recevant 2000 personnes doit-elle posséder une ligne directe avec le centre de secours des sapeurs-pompiers ?
Quelle doit être la réaction au feu des isolations intérieures dans un établissement de type U pouvant contenir plus de 3000 personnes ?
Comment doit être assurée la surveillance d'une salle de conférence pouvant accueillir 1600 personnes ?

ou bien la question donne le libellé de l'activité sans en préciser le type, au sens de la réglementation :

> *Quelle doit être la réaction au feu des isolations intérieures dans un hôpital ?*
> *Où doit-on placer l'organe de coupure gaz d'une salle de travaux pratiques pour un établissement d'enseignement de 1ère catégorie ?*

Le type d'un établissement est donné par la réglementation (arrêté du 25 juin 1980) en fonction de l'activité de cet établissement de la manière suivante :

> *Les établissements sont classés en types, selon la nature de leur exploitation :*
> *a) Etablissements installés dans un bâtiment :*
> *L : salles d'audition, de conférences, de réunions, de spectacles ou à usage multiple ;*
> *M : magasins de vente, centres commerciaux ;*
> ...

On peut donc déterminer le type de l'établissement à partir du libellé de son activité, à condition que l'expression employée dans la question soit inscrite explicitement dans la liste des activités inscrites dans le texte de l'arrêté.

La manière de calculer la catégorie d'un établissement est aussi précisée dans les textes :

> *1re catégorie : au-dessus de 1 500 personnes ; - 2e catégorie : de 701 à 1 500 personnes ; - 3e catégorie : de 301 à 700 personnes ; - 4e catégorie : 300 personnes et au-dessous, à l'exception des établissements compris dans la 5e catégorie ; - 5e catégorie : établissements faisant l'objet de l'article R. 123-14 dans lesquels l'effectif du public n'atteint pas le chiffre minimum fixé par le règlement de sécurité pour chaque type d'exploitation.*

On peut ainsi calculer la catégorie d'un établissement à partir de l'effectif du public autorisé, sauf si celui-ci est inférieur ou égal à 300 personnes car, dans ce cas, intervient aussi le type de l'établissement.

L'objectif de ce paragraphe est d'écrire une grammaire permettant de passer d'une formulation "en clair" à une autre qui précise le type et la catégorie de l'établissement selon les codes donnés dans les textes.
On écrira un transducteur qui permettra de passer d'un libellé d'activité à un type d'établissement (au sens de l'arrêté du 25 juin 1980). Puis, à la suite du §3.2.1, on étudiera les différentes formulations qui associent un effectif à un local ou un établissement afin de le transformer en une forme canonique. On établira ensuite les transformations de la forme canonique en une catégorie en utilisant les textes réglementaires. On pourra ainsi connaître le type et la catégorie d'un établissement des quatre premières catégories.
Enfin, pour les établissements dont l'effectif est inférieur à 300 personnes, il faudra croiser les informations concernant le type et l'effectif pour déterminer la catégorie.

Dans un premier temps, on ne s'intéresse qu'aux cas simples dont l'activité est explicitement prévue dans les textes, et on suppose qu'il est toujours pertinent, pour une recherche documentaire automatique, de formuler la catégorie et le type de l'établissement. Pourtant, cette situation admet de nombreuses exceptions : des textes peuvent s'appliquer à un établissement d'une catégorie quel que soit son type, ou bien, des établissements, quel que soit l'effectif du public reçu, se voient appliquer une réglementation à cause du type de l'activité qu'ils abritent.

En fait, ce travail n'a été qu'ébauché. Pour le mener à bien, il faudrait réaliser les étapes suivantes :

- étudier les différentes formulations de l'affectif associé à un établissement. Le transducteur *ERPeffectif.grf* analyse les différentes formulations de cet effectif (ces expressions sont attestées dans le corpus mais devraient être complétées par d'autres formes relevées dans d'autres ensembles de textes). On a en particulier relevé l'importance des adverbes qui peuvent précéder le numéral ou suivre le *GN* qui contient un numéral : ils doivent être traités en détail parce qu'ils peuvent provoquer un changement de catégorie pour un établissement :

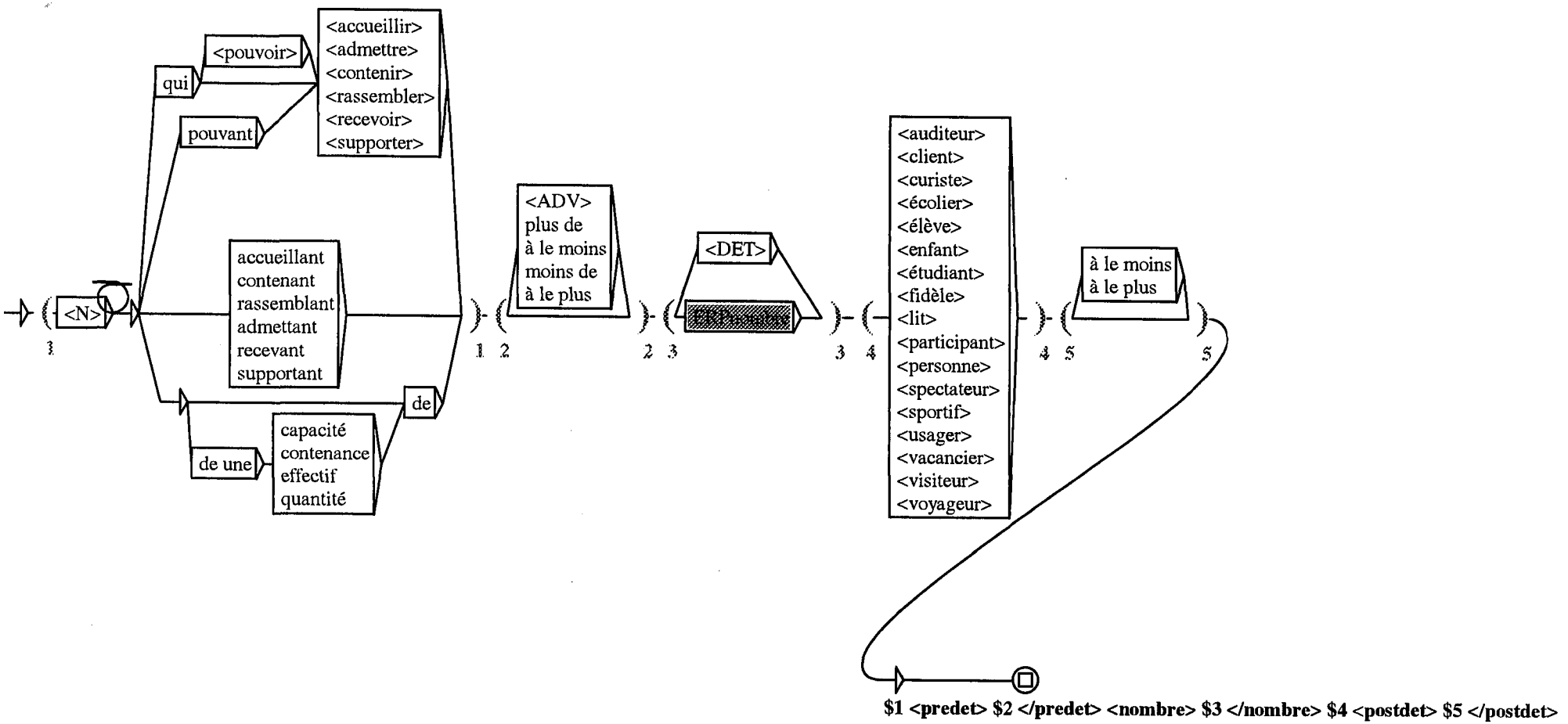
  un collège pouvant recevoir 1 500 élèves
  un collège pouvant recevoir moins de 1 500 élèves
  un collège pouvant recevoir au moins 1 500 élèves
  un collège pouvant recevoir 1 500 élèves au moins

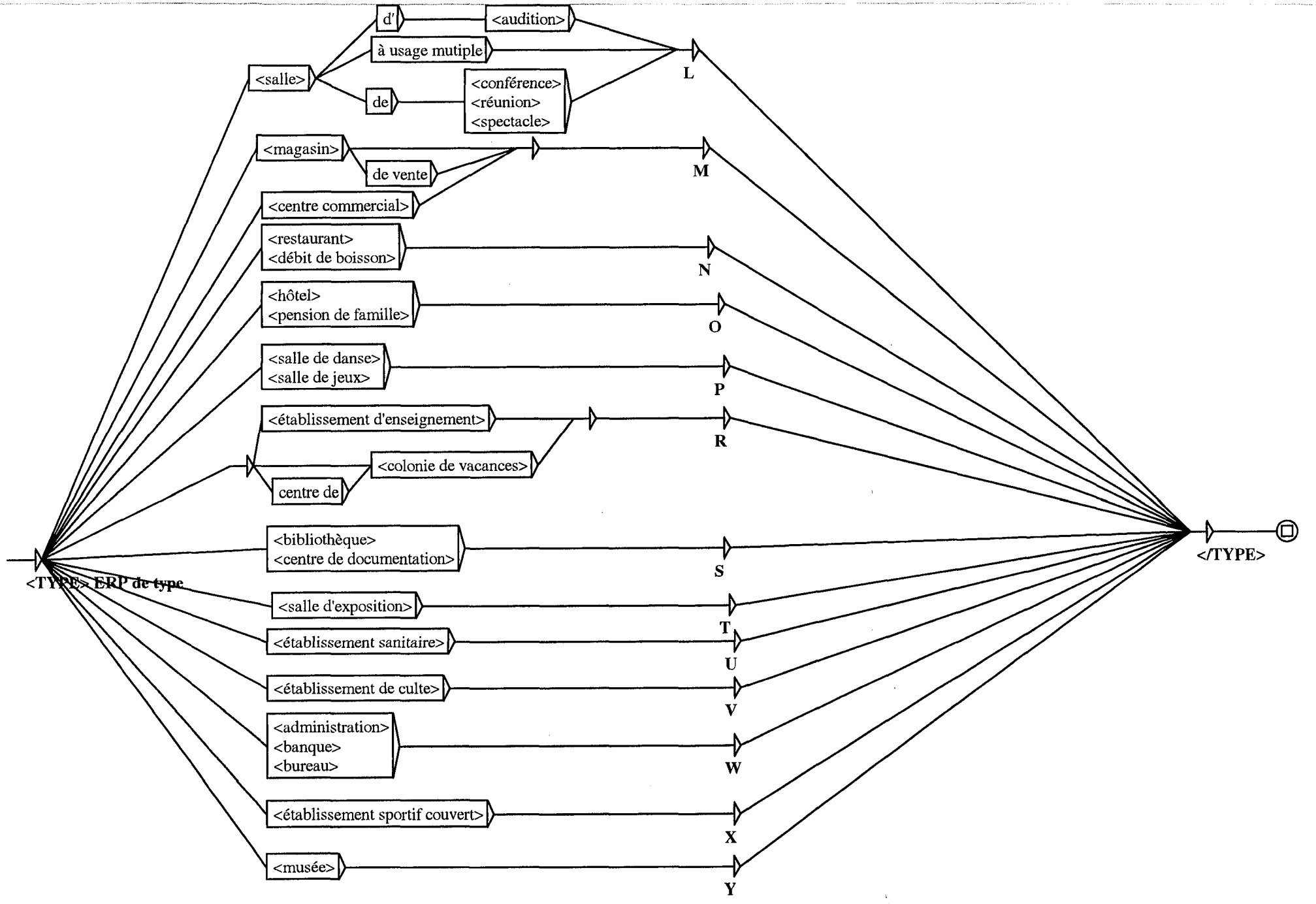
  Les deux premiers *GN* permettent de classer le collège dans la deuxième catégorie, les deux suivants dans la première ;

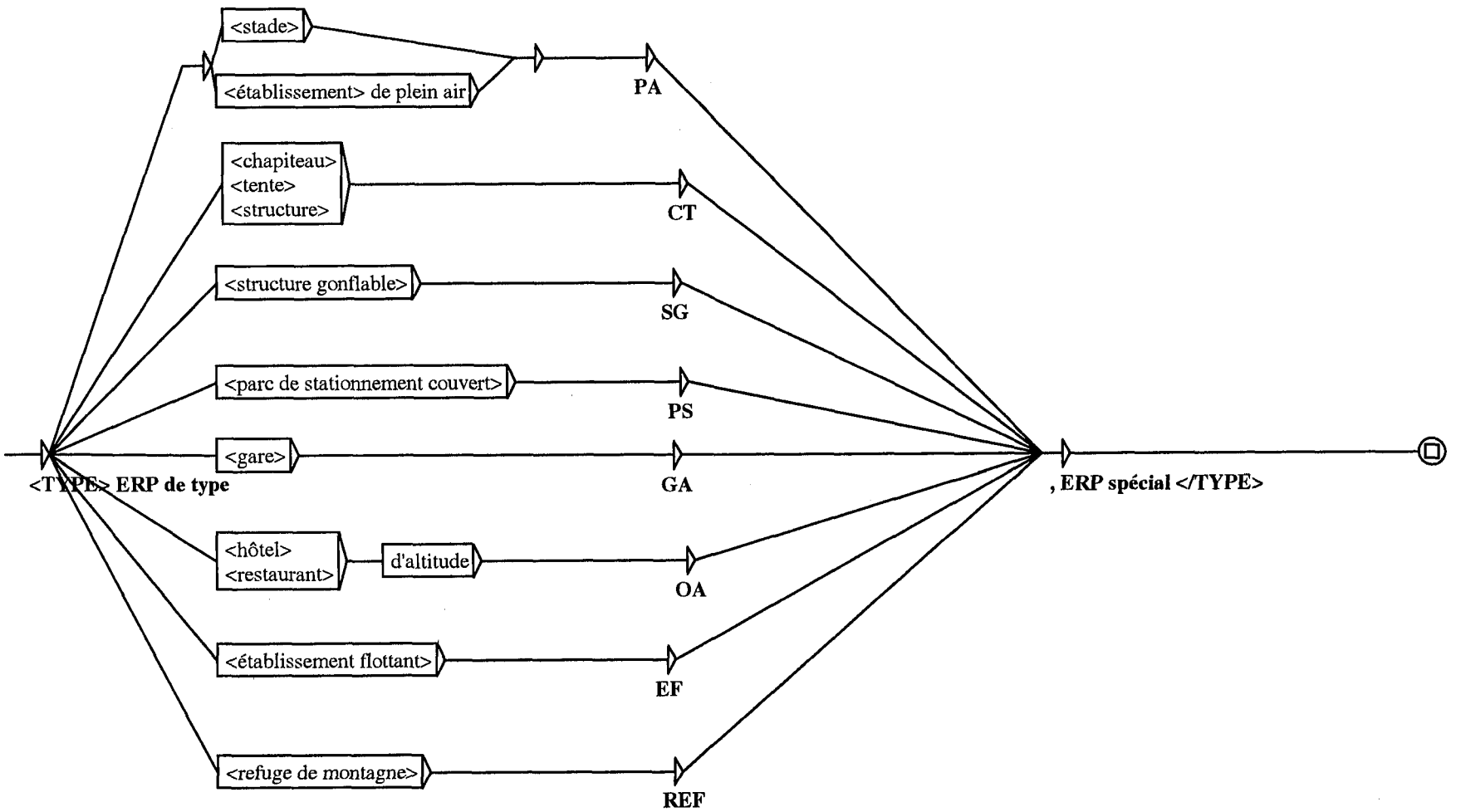
- traduire par un algorithme la règle de calcul d'un catégorie à partir d'un effectif en s'appuyant sur les éléments identifiés dans l'étape précédente. Si les numéraux sont écrits en lettres, il faut les transcrire (à l'aide d'un transducteur) en une formulation chiffrée ;

- repérer l'expression du type de l'établissement ou son activité afin de la traduire en un type. C'est l'objectif des deux transducteurs *ERPtype.grf* et *ERPtypeSpecial.grf*. Il faudrait les compléter en ajoutant des synonymes (par exemple *salle de sport*s pour *établissement sportif couvert*) et des hyponymes (*école maternelle*, *école primaire*, *collège*, *lycée* pour *établissement d'enseignement*) des *GN* déjà identifiés ;

- enfin, traiter le problème des établissements de 5ème catégorie, c'est-à-dire traduire par un algorithme la réglementation qui fait intervenir, pour classer un établissement dans cette catégorie, à la fois l'activité et l'effectif. Pour cette partie, nous ne présentons aucune réalisation.


<accueillir>
<admettre>
<contenir>
<rassembler>
<recevoir>
<supporter>
<pouvoir>
qui
pouvant
<N>
accueillant
contenant
rassemblant
admettant
recevant
supportant
de
de une
capacité
contenance
effectif
quantité
<ADV>
plus de
à le moins
moins de
à le plus
<DET>
<auditeur>
<client>
<curiste>
<écolier>
<élève>
<enfant>
<étudiant>
<fidèle>
<lit>
<participant>
<personne>
<spectateur>
<sportif>
<usager>
<vacancier>
<visiteur>
<voyageur>
à le moins
à le plus
1 2
2 3
3 4
4 5
5
$1 <predet> $2 </predet> <nombre> $3 </nombre> $4 <postdet> $5 </postdet>


<TYPE> ERP de type
d'
<audition>
à usage mutiple
<salle>
de
<conférence>
<réunion>
<spectacle>
L
<magasin>
de vente
M
<centre commercial>
<restaurant>
<débit de boisson>
N
<hôtel>
<pension de famille>
O
<salle de danse>
<salle de jeux>
P
<établissement d'enseignement>
R
<colonie de vacances>
centre de
<bibliothèque>
<centre de documentation>
S
</TYPE>
<salle d'exposition>
T
<établissement sanitaire>
U
<établissement de culte>
V
<administration>
<banque>
<bureau>
W
<établissement sportif couvert>
X
<musée>
Y


<stade>
<établissement> de plein air
PA
<chapiteau>
<tente>
<structure>
CT
<structure gonflable>
SG
<parc de stationnement couvert>
PS
<gare>
GA
<TYPE> ERP de type
, ERP spécial </TYPE>
<hôtel>
<restaurant>
d'altitude
OA
<établissement flottant>
EF
<refuge de montagne>
REF

5. Une grammaire locale pour traiter les comparaisons de nombres

Lors d'une recherche automatique d'informations qui pourraient répondre à une question posée par l'utilisateur, une cause de silence est que les nombres ne sont pas identifiés, ou s'ils sont identifiés ne sont pas considérés comme ordonnés. Et dans un texte technique contenant beaucoup de nombres, la perte d'informations est alors considérable. Par exemple, si une question contient le *GN* suivant :

> *un local de 1 500 mètres carrés de surface*

le nombre 1500 n'est pas comparé à 1000 pour conclure qu'il est plus grand et que la réglementation dont le titre suit est donc applicable :

> *application du paragraphe 6.2.3 [2°] relatif aux locaux de surface supérieure à 1 000 mètres carrés*

Aucun des outils utilisés dans l'étude MédiaConstruct présentée dans le premier chapitre, pas plus que ceux existant pour le moment dans le commerce, ne repèrent ces informations numériques. Pour certains domaines ces informations peuvent être secondaires, dans notre corpus elles sont essentielles. Il faudrait écrire une grammaire locale pour reconnaître ces expressions et un programme de traitement qui lui serait couplé pour simuler les comparaisons intuitivement mises en œuvre quand on compare deux énoncés contenant des nombres, et les restituer selon une forme utilisable pour les outils qui procèdent à la recherche d'information dans le corpus.

6. Conclusions

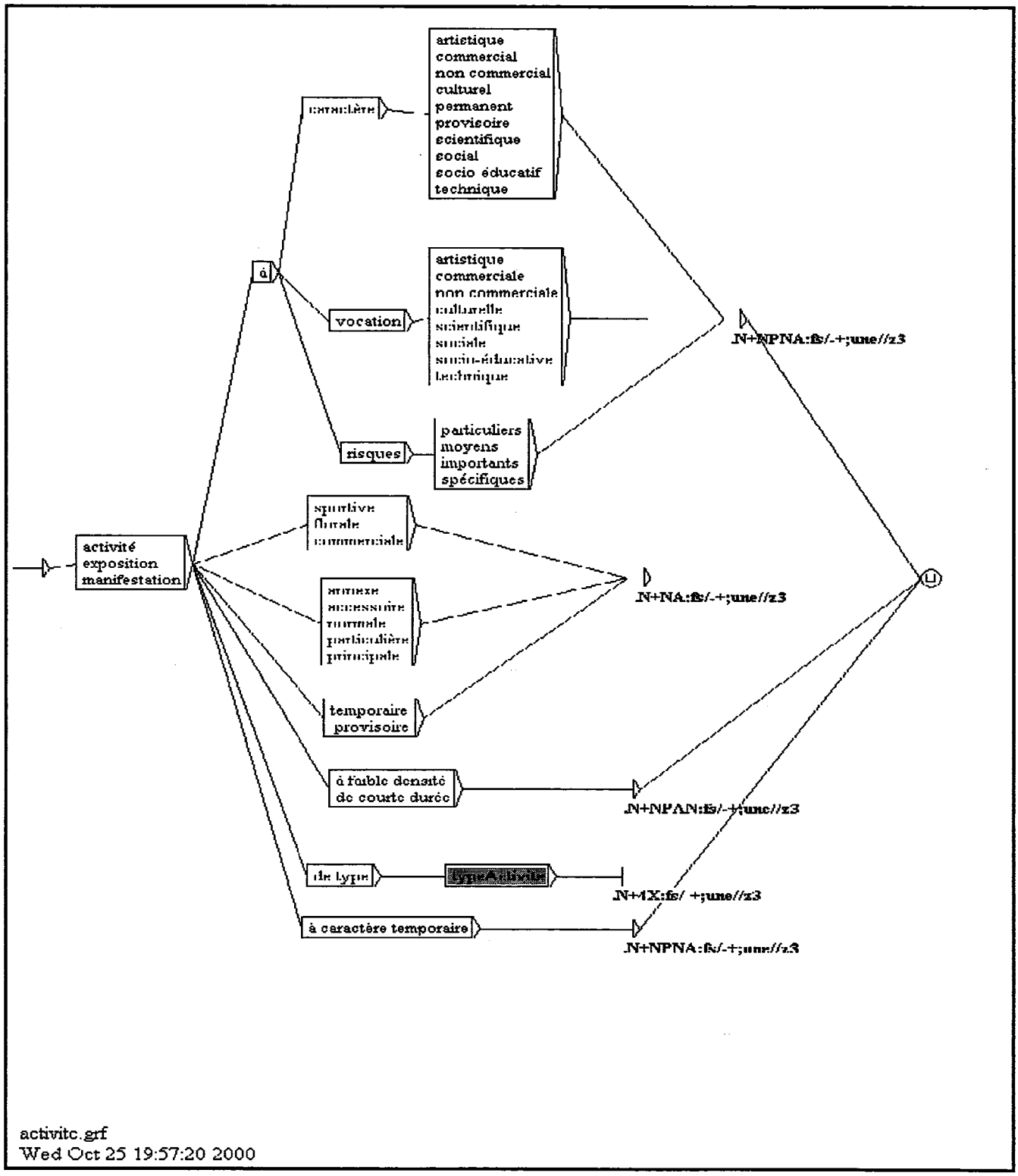
L'objectif de ce chapitre était de montrer l'intérêt d'écrire des grammaires locales pour formaliser des connaissances lexicales et syntaxiques concernant le domaine de la sécurité incendie.

En effet, comme on l'a déjà vu dans le chapitre concernant l'emploi de différents classifieurs dans la construction de noms composés, la représentation sous forme d'automates ou de transducteurs permet de rendre compte de la combinaison des modifieurs dans la formation des noms composés.

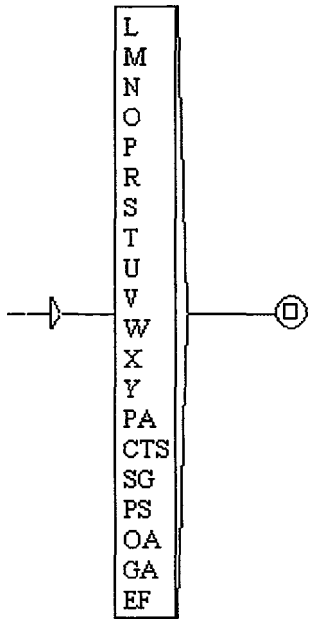
En outre, les grammaires locales qui traitent d'un aspect de la réglementation à travers des connaissances lexicales et syntaxiques permettent de synthétiser et d'organiser les connaissances sur l'aspect traité. Le développement arborescent des automates contribue à la hiérarchisation des informations. De ce point de vue, les grammaires locales, couplées à un glossaire et à une description thématique des aspects abordés par le domaine de spécialité, permettent à un public non spécialiste de se familiariser avec les concepts manipulés par ce nouveau domaine grâce à une présentation hiérarchisée des connaissances. Nous présentons une grammaire pour illustrer ce propos. Elle concerne la description d'une activité, au sens de la SI[7] : elle permet à un néophyte de visualiser les différents éléments pris en compte pour qualifier une activité, et ainsi fournit une aide dans l'apprentissage de la réglementation.

---

[7] Nous rappelons que l'activité qu'abrite un établissement donne son type qui est une caractéristique essentielle pour la réglementation sécurité incendie.



activite.grf : transducteur récapitulant les caractéristiques pertinentes, par rapport à la sécurité incendie, d'une activité



typeActivite.grf : automate d'identification du code désignant l'activité d'un établissement

De même, quand elle concerne un objet concret, une grammaire locale peut être assimilée à une note technique sur cet objet parce qu'elle propose une organisation des éléments lexicaux qui, si on le désire, peut coïncider avec le découpage mécanique de l'objet étudié. Nous présentons en annexe une grammaire locale bâtie sur le nom *porte*.

Enfin, des grammaires, comme celle ébauchée en 4, permettent de repérer différentes informations pour les reformuler sous une forme canonique et recensent ainsi des formulations différentes que l'on peut rattacher à un même concept. Pour une recherche automatique d'information, elles permettent ainsi de rapprocher le libellé de la question, de textes qui contiennent des expressions formellement très différentes, mais qui sont pourtant pertinents par rapport à la question posée.

Traitement de la coordination à l'intérieur des groupes nominaux

1. Objectif

*dans les **établissements** et dans les **locaux** présentant des risques particuliers d'incendie*
*les locaux "**groupe électrogène**" et **transformateurs***
*les établissements **existants** et **à modifier***
*des écrans ou carters **fixés** et bien **adaptés***

Les groupes nominaux précédant comportent au moins une coordination reliant des noms et des modifieurs. Dans le premier, c'est le modifieur qui est commun aux deux noms. Dans les deux suivants, ce sont les noms têtes de groupe qui sont partagés par les deux modifieurs. Dans le dernier, chacun des noms se combine avec chacun des modifieurs. Chacune des phrases est donc elliptique d'un nom ou d'un modifieur.
En termes quantitatifs, la coordination par *et* et *ou* est sensiblement plus utilisée dans le corpus technique, que dans un corpus généraliste comme *Le Monde* : 2,19 % des mots du texte sont des occurrences de *et* ou *ou* dans le premier, pourcentage à comparer avec 1.62 % dans le deuxième.

De plus, la coordination dans le corpus technique est surtout utilisée à l'intérieur des groupes nominaux pour lier des modifieurs alors que dans *Le Monde*, on la trouvera qui relie des groupes nominaux qui partagent la même fonction dans la phrase, ou des propositions de même nature :

La porte d'entrée et plusieurs vitrines ont été soufflées, et de nombreux véhicules stationnés alentour ont été détériorés
Par ailleurs, le Club de Paris pourra "traiter" ou rééchelonner le stock résiduel de cette dette

Ce travail a pour objectif de définir des règles permettant de reconstruire automatiquement une phrase complétée, à partir d'une phrase qui comporte au moins une coordination. Dans une première étape, la reconstruction de groupes nominaux complétés s'appuie sur une segmentation de la phrase faite au préalable, mais ensuite elle contribue aussi, par l'application de règles que l'on établira, à cette segmentation en donnant des indices sur la structure des groupes nominaux.
Les coordinations étudiées se font essentiellement par *et* ou *ou* et relient au moins deux modifieurs ou deux noms à l'intérieur du groupe nominal.
Ces règles sont fondées sur l'étude des exemples trouvés dans le corpus. De nombreuses formulations sont ambiguës et il n'est évidemment pas question de résoudre les ambiguïtés mieux que ce qu'un interlocuteur humain non spécialiste du domaine concerné ne pourrait le faire.

Le corpus utilisé dans ce chapitre est hétérogène. Il contient la réglementation concernant l'incendie dans les établissements recevant du public (ERP), un ensemble de lois, décrets, arrêtés, notes techniques représentant 820 ko de textes, ainsi qu'une norme, la norme ICS. Celle-ci contient environ un millier d'expressions organisées de manière à constituer un classement hiérarchique des noms d'activités. Ce classement sert ensuite de structure à des catalogues et des documents normatifs. Ces

activités concernent des domaines aussi variés que les mathématiques, les soins de santé, l'agriculture ou l'économie domestique.

Notations et grammaires utilisées pour la recherche de motifs syntaxiques:

GN =: *(Det + <E>) (Adjectif + <E>) (non + <E>) N (Mod + <E>) (Et GN + <E>))*

Et =: *et + ou*

Adjectif =: *(non + <E>) (Adv+ <E>) (A+V:K)*

Compl =: *Prep GN*

Mod =: *(Adjectif (Compl+ <E>)) +*
(Prep (Det + <E>) (Adjectif + <E>) N (Adjectif + <E>)) +
V:G (Compl +<E>) |+
Prep (Adv + <E>) V:W

où :

Prep désigne	une préposition
N	un nom
Det	un déterminant
Adv	un adverbe
A	un adjectif
V:K	un participe passé
V:G	un participe présent
V:W	un verbe à l'infinitif

Dans la suite de l'exposé, on considérera qu'on dispose de dictionnaires de mots simples et de mots composés complets pour les termes du bâtiment et de la sécurité incendie.

2. Typologie des coordinations

Proposer une typologie des groupes nominaux revient à considérer que l'on sait segmenter le texte et identifier les *GN,* quelle que soit leur structure. Le problème est loin d'être résolu en particulier pour les structures récursives :

> *(les plans et caractéristiques) ((du système d'extraction) et (des conduits d'évacuation (des buées et fumées)))*

Néanmoins, les outils fournis par INTEX (étiquetage syntaxique, construction de transducteurs applicables au corpus) permettent de travailler à cette identification des *GN.*
Dans cette étude, le point sur lequel est articulée la segmentation est la conjonction de coordination, élément à partir duquel on recherche les parties gauche et droite qui supportent la coordination et, le cas échéant, les groupes de mots mis en facteur grâce à cette coordination.
La même structure de groupe nominal contenant une ou plusieurs coordinations conduit à des règles de réécriture très différentes. On proposera une typologie fondée sur la distributivité du ou des éléments mis en facteur par rapport aux groupes de mots coordonnés. A une forme coordonnée unique

peuvent correspondre différentes formes développées : la forme initiale des types 1 et 2 est identique mais conduit à des formes développées différentes.

La typologie proposée est la suivante :

type 1 : ***N1* et *N2 Mod*** équivalent à ***N1* et (*N2* Mod)**

dont le modèle est :

acoustique et mesurage acoustique

pour :

acoustique et (mesurage acoustique)

type 2 : ***N1* et *N2 Mod*** équivalent à **(*N1 Mod*) et (*N2* Mod)**

dont le modèle est :

barres et ronds en acier

pour

(barres en acier) et (ronds en acier)

type 3 : ***N1 Mod1* et *Mod2*** équivalent à **(*N1 Mod1*) et (*N1 Mod2*)**

dont le modèle est :

communications téléphoniques et télégraphiques

pour

(communications téléphoniques) et (communications télégraphiques)

type 4 : ***N1 Mod1* et MOT*** équivalent à **(*N1 Mod1*) et (MOT*)**

dans lequel on reconnaît au début de *MOT** une séquence que l'on peut analyser sous la forme *N Mod*. Les expressions classées dans cette catégorie sont donc analysées de la manière suivante :

N1 Mod1* et MOT équivalent à **(*N1 Mod1*) et (*N2 Mod2*)**

dont le modèle est :

appareils diviseurs et dispositifs de préhension

pour

(appareils diviseurs) et (dispositifs de préhension)

type 5 : ***N1* et *N2 Mod1* et *Mod2*** équivalent à **(*N1 Mod1*) et (*N1 Mod2*) et (*N2 Mod1*) et (*N2 Mod2*)**

dont le modèle est :

écrans ou carters fixés et bien adaptés

pour

((écrans fixés) et (écrans bien adaptés)) ou ((carters fixés) et (carters bien adaptés))

On va dans la suite de ce paragraphe détailler les exemples et les propriétés qui illustrent chacun des types identifiés.

2.1. type 1 : *N1 et N2 Mod* équivalent à *N1 et (N2 Mod)*

	N1	*et*	*N2*	*Mod*
(1)	*acoustique*	*et*	*mesurage*	*acoustique*
(2)	*béton*	*et*	*produits*	*en béton*
(3)	*aluminium*	*et*	*alliages*	*d'aluminium*
(4)	*générateur*	*ou*	*groupe*	*de générateurs*
(5)	*non allumage*	*ou*	*extinction*	*fortuite*

On n'observe pas de restriction sur le modifieur. Ce peut être un adjectif, un participe passé, ... S'il est rattaché à *N2* par une préposition, celle-ci n'est pas contrainte : *de, en, à base de, ...*

Dans les exemples (1) à (3), *Mod* contient *N1* ou un dérivé de *N1*. Pour (1), le nom acoustique est devenu un adjectif ; dans (2) et (3), *N1* est répété, dans le modifieur, sous la même forme : un nom relié à *N2* par une préposition (cf. règle 3.1).

Le remplacement du modifieur par un adjectif possessif (cf. règle 2.1) n'est pas toujours possible :

(1) a	*l'acoustique*	*et*	*son mesurage*
(2) *a*	**le béton*	*et*	*ses produits*
(3) *a*	*l'aluminium*	*et*	*ses alliages*
(4) *a*	**le générateur*	*ou*	*son groupe*

2.2. type 2 : *N1 et N2 Mod* équivalent à *N1 Mod et N2 Mod*

	N1	*et*	*N2*	*Mod*
(10)	*façade*	*et*	*baie*	*accessibles*
(11)	*dans les établissements*	*et*	*dans les locaux*	*présentant des risques particuliers d'incendie*
(12)	*barres*	*et*	*ronds*	*en acier*
(13)	*les canalisations*	*et*	*autres matériels électriques*	*des locaux présentant des risques*
(14)	*composants*	*et*	*accessoires*	*pour matériel de télécommunication*

Comme pour le type 1, on n'observe pas de restriction sur la construction du modifieur. Il ne pourra donc y avoir de critère contrastif portant sur la forme de ce modifieur qui permette, pour un groupe nominal de forme *N1* et *N2* Mod, de choisir entre un développement de type 1 ou de type 2.

2.3. type 3 : *N1 Mod1 et Mod2* équivalent à *(N1 Mod1) et (N1 Mod2)*

	N	*Mod1*	*et*	*Mod2*
(20)	*dégagements*	*accessoires*	*et*	*supplémentaires*
(21)	*dispositifs*	*d'éclairage*	*et*	*de signalisation*
(22)	*machines*	*à aléser*	*et*	*à fraiser*
(23)	*une ossature*	*en matériaux de catégorie M3*	*et*	*en bon état*
(24)	*un endroit*	*accessible en permanence*	*et*	*bien signalé*
(25)	*les établissements*	*existants*	*et*	*à modifier*
(26)	*une cage*	*aux parois incombustibles*	*et*	*de degré coupe-feu égal à 2 h*
(27)	*une paroi*	*en matériaux incombustibles*	*et*	*CF de degré deux heures*
(28)	*un raccord*	*spécifique au gaz distribué*	*et*	*portant l'identité de ce gaz*
(29)	*les locaux*	*"groupe électrogène"*	*et*	*transformateurs*

Statistiquement, c'est ce type qui compte l'effectif le plus important.

2.3.1. Construction des modifieurs

Les modifieurs peuvent être des adjectifs comme dans (20), (24) et (27) ou des noms (29), commencer par une préposition *de*, *à*, *en*, ... (exemples (21), (22), (23), (26)) ou un participe présent (28).

En termes de construction, les modifieurs peuvent être identiques (par exemple (20) à (23)) : deux adjectifs ou deux compléments introduits par la même préposition.
Dans (23) cette symétrie est plus artificielle puisque les deux modifieurs sont introduits par la même préposition mais *en matériaux de catégorie M3* est une séquence libre alors que *en bon état* est un adverbe composé.
Dans (24), les deux modifieurs sont des adjectifs, mais eux-mêmes modifiés par des adverbes, l'un antéposé, l'autre postposé.
Pour les expressions (25) à (28), la symétrie de la construction - une coordination reliant deux modifieurs qui s'appliquent à la tête du groupe nominal - est masquée par le fait que les deux modifieurs n'ont pas la même structure.
Dans l'exemple (29), la construction de la phrase, bien qu'elle soit attestée dans le corpus, n'est pas très claire. Une expression complétée et plus facilement compréhensible utiliserait des prépositions et s'écrirait :

(29) a. les locaux du "groupe électrogène" et des transformateurs

2.3.2. Mots composés et coordination des modifieurs

	N	***Mod1***	***et***	***Mod2***
(21)	*dispositifs*	*d'éclairage*	*et*	*de signalisation*

Les dictionnaires de noms composés d'INTEX donnent la liste suivante :

dispositif d'appel,.N+NDN+Conc+z0:ms/un
dispositif d'assemblage,.N+NDN+Conc+z0:ms/un;Méca
dispositif d'évacuation,.N+NDN+Conc+z0:ms/un;Mil
dispositif de commande,.N+NDN+Conc+z0:ms/un;Méca
dispositif de coupure,.N+NDN+Conc+z0:ms/un
dispositif de protection,.N+NDN+Conc+z0:ms/un;Mil
dispositif de réglage,.N+NDN+Conc+z0:ms/un;Méca
dispositif de sécurité,.N+NDN+Conc+z0:ms/un;Méca

Il apparaît que *dispositif d'éclairage* et *dispositif d'appel* sont, au même titre, des noms composés à rajouter dans les dictionnaires.

On peut alors décrire le procédé qui permet de coordonner deux noms composés dont la tête est identique et qui partagent la même structure – ici, *NDN* – :

(21) a.	*dispositifs d'éclairage*	*et*	*dispositifs*	*de signalisation*
(21) b.	*dispositifs d'éclairage*	*et*	*ε*	*de signalisation*
(21)	*dispositifs d'éclairage*	*et*		*de signalisation*

La tête du nom composé de la partie droite est effacée sans que le sens de l'expression complète ne soit modifié.

Si l'on étend la dénomination *nom composé* à toutes les expressions de type *NDN* formées sur *dispositif* que l'on a trouvées dans le corpus, on rajoute au dictionnaire[1] les entrées suivantes :

dispositif de (<Det> + ε) N

dispositif d'arrêt d'urgence
dispositif de désenfumage
dispositif de régulation

et on peut former :

dispositif de désenfumage et d'éclairage
dispositif d'éclairage et de désenfumage
dispositif de régulation et de signalisation
dispositif de signalisation et de régulation
...

Ce procédé de coordination et effacement s'applique à la totalité des expressions *NDN* que l'on peut former sur *dispositif* et il est largement attesté dans le corpus que l'on a utilisé. Les possibilités de coordination entre différents types de noms composés (*NA* et *NDN, NA* et *NN, NN* et *NDN,* ...) demanderaient une étude complète.

Néanmoins, cette opération – coordination et effacement – n'est pas généralisable à tous les noms composés.

En effet, si on prend par exemple les mots composés du vocabulaire généraliste formés sur *homme*, ce procédé de coordination-effacement n'est pas toujours possible :

mots composés de type *NA* :

homme-grenouille
homme-oiseau

le mot composé obtenu en coordonnant les modifieurs est impossible :

**homme grenouille et oiseau*
**homme oiseau et grenouille*

En revanche, avec :

homme public
homme privé

la coordination paraît plus acceptable :

homme public et privé
**? homme privé et public*

mots composés de type *NDN* :

homme de théâtre
homme de science
homme de lettres
homme de plume
homme d'église

**? homme de théâtre et de lettres*
homme de lettres et de théâtre
homme de science et de lettres
homme de lettres et de science
homme d'église et de théâtre

1 L'ensemble des noms composés que l'on peut former sur le nom classifieur *dispositif* est étudié dans le chapitre 3, paragraphe 2.

? homme de théâtre et d'église
homme d'église et de lettres
homme de lettres et d'église
homme d'église et de science
homme de science et d'église

homme d'épée
homme de guerre
homme de main

**homme d'épée et de guerre*
homme de guerre et d'épée
**homme d'épée et de main*
**homme de main et d'épée*
**homme de guerre et de main*
**homme de main et de guerre*

La possibilité de prévoir, à l'intérieur du vocabulaire général, l'acceptabilité des expressions coordonnées où le nom tête de groupe est effacé dans la partie droite, paraît assez limitée. Elle dépend visiblement du degré de figement des expressions initiales, du fait que leur sens soit ou non compositionnel, de la proximité sémantique des éléments désignés et du locuteur.
En revanche, les noms composés des vocabulaires techniques formés à partir d'un nom tête identique et d'un même schéma sont très faciles à coordonner de cette manière. Parce que le sens de ces expressions est tout à fait compositionnel, il est facile d'effacer le nom tête de la partie droite de la coordination.

2.3.3. Coordination des modifieurs et unicité de l'objet désigné par la tête de groupe

Dans toutes les expressions de types 1 ou 2 (i.e. *N1 et N2 Mod*), il y a nécessairement deux objets désignés par *N1* et *N2*. Les expressions de type 3 (i.e. *N Mod1 et Mod2*) peuvent, elles, qualifier un ou deux objets.

Les expressions (20) à (22) et (29) peuvent ou bien nommer une entité unique partageant *Mod1* et *Mod2*, ou bien représenter les deux objets (*N Mod1*) et (*N Mod2*) que l'on désigne plus commodément par *N Mod1 et Mod2*.
Les exemples (23) à (28) se rapportent à l'évidence à une seule entité possédant, dans ce cas, les deux propriétés *Mod1* et *Mod2*, même si existent les expressions isolées (et les objets correspondants) *N Mod1* et *N Mod2*.

Pour les expressions suivantes, il n'est pas possible de dissocier partie droite et partie gauche de la coordination :

le ministre délégué au commerce et à l'artisanat
le ministre délégué au logement et au cadre de vie
Code de la Construction et de l'Habitation
Comité français du butane et du propane

Les deux premières expressions sont, en fait, équivalentes à des noms propres et par là non modifiables ou dissociables. Mais la formule peut changer selon la composition des gouvernements, ou l'expression être dissociée si on ne s'intéresse qu'à une partie des compétences d'un ministre.

Pour les deux dernières, il n'est pas possible de distribuer le nom tête sur les modifieurs parce que les expressions désignent des entités aux propriétés multiples par définition. Le *Code de la Construction*

ni *Code de l'Habitation* n'existent en tant que tels ; de même, il n'y a pas de *Comité français du butane* ni de *Comité français du propane*. En revanche, on pourra trouver, en manière d'abréviation, *Code de la Construction* ou *Comité français du butane* à la place des expressions complètes.

le clos et le couvert

L'expression est issue du domaine de la construction, et est utilisée telle quelle dans le langage courant. *Clos et couvert* sont dans ce cas des noms ; l'expression complète n'existe pas au pluriel. On retrouve *couvert* employé comme nom dans deux autres expressions : *le vivre et le couvert, le gîte et le couvert.*

(75) *les flammes et gaz chauds ou inflammables*
(75) a. *étanchéité aux flammes et aux gaz chauds ou inflammables*
(75) b. *le passage rapide des flammes ou des gaz chauds*

(76) *les fumées et gaz de combustion*
(76) a. *détection sensible aux fumées et aux gaz de combustion.*

(77) *le risque d'incendie et de panique*
(77) a. *les risques d'incendie et de panique*

Toutes les expressions précédentes sont liées au domaine de la sécurité incendie.
Les flammes ou les gaz chauds ou inflammables sont à considérer de la même façon quand on s'intéresse aux propriétés pare-feu d'un élément de construction. L'expression est toujours au pluriel. On observe dans (75 b) une forme abrégée en *des flammes ou des gaz chauds.*
Dans (76), les objets visés sont ceux se rapportant au désenfumage et donc *fumées et gaz de combustion*, est toujours sous une forme plurielle.
Pour (77), on trouvera la forme singulier ou pluriel pour le mot *risque*. Le singulier insiste plus sur sa prise en compte globale par la sécurité civile, même si dans la réalité il s'agit de deux dangers différents.

le règlement et les normes
examen de conformité au règlement et aux normes
conditions fixées par les règlements et normes en vigueur
les dispositions du présent règlement

Dans des textes à visée législative et réglementaire, la référence est constituée par *les règlement et normes*, que l'on précisera éventuellement en ajoutant l'adverbe *en vigueur. Règlements et normes* fixent des conditions ; si l'on s'en tient à des dispositions, on se référera au seul *règlement.*

aptitude physique et connaissances techniques
normes d'hygiène et de sécurité
règles techniques et de sécurité
contrôles et vérifications techniques

Les quatre expressions précédentes désignent des entités souvent évoquées dans des textes réglementaires, et qu'on considérera comme des mots composés.

2.4. type 4 : *N1 Mod1 et <MOT>** équivalent à *(N1 Mod1) et (N2 Mod2)*

On classe dans le type 3 des séquences où on sait identifier dans la partie droite de la coordination un modifieur ; l'expression se présente alors sous la forme suivante : *N1 Mod1 et N2 Mod2*, qui est équivalent à *(N1 Mod1) et (N2 Mod2)*.
Dans le type 4, on regroupe les expressions pour lesquelles parties gauche et droite de la coordination sont indépendantes : aucun terme de l'une ne se distribue sur un terme de l'autre.

	N1	*Mod1*	*et*	*N2*	*Mod2*
(30)	*appareils*	*mobiles*	*et*	*moyens*	*divers*
(31)	*développement*	*de logiciels*	*et*	*documentation*	*des systèmes*
(32)	*appareils*	*diviseurs*	*et*	*dispositifs*	*de préhension*
(33)	*appareils*	*de coupure*	*et*	*dispositifs*	*de commande*
(34)	*bloc*	*moteur*	*et*	*composants*	*internes*

Ce type de coordination paraît symétrique : chaque nom possède son modifieur ; la coordination lie deux *GN* qui partagent la même fonction grammaticale dans la phrase. Mais, selon la couverture des dictionnaires de noms composés, ces expressions peuvent perdre leur symétrie apparente.

Les termes suivants :

appareil diviseur
appareil de coupure
dispositif de préhension
dispositif de commande

même s'ils n'appartiennent pas aux dictionnaires généralistes d'INTEX sont à rapprocher d'autres expressions construites sur le même type : *appareil de N* et *dispositif de N*, et donc à intégrer à des dictionnaires spécialisés de noms composés.

On fera de même avec les expressions suivantes :

appareil mobile
bloc moteur
développement de logiciel

Avec ces nouveaux mots composés, (30), (31) et (34) perdent leur symétrie ; leur structure devient : *N1* et *N2 Mod* et se pose alors le problème de la distributivité de *Mod* par rapport à *N1* et *N2*.

2.5. type 5 : *N1 et N2 Mod1 et Mod2* équivalent à *(N1 Mod1) et (N1 Mod2) et (N2 Mod1) et (N2 Mod2)*

	N1	*et*	*N2*	*Mod1*	*et*	*Mod2*
(40)	*comportement au feu*	*et*	*facilité d'allumage*	*des matériaux*	*et*	*des produits*
(41)	*écrans*	*ou*	*carters*	*fixés*	*et*	*bien adaptés*

2.5.1. Mots composés désolidarisés et reconstruits

(42)	*les appareils*	*et*	*les installations*	*de chauffage*	*et*	*de ventilation*
(43)	*appareils*	*ou*	*dispositifs*	*d'extinction*	*et*	*d'alerte*

Appareils de chauffage, *installations de chauffage*, *dispositifs d'extinction* et *dispositifs d'alerte* sont assez classiquement des noms composés. Les expressions *appareils de ventilation*, *installations de ventilation*, *appareils d'extinction*, *appareils d'alarme* sont moins utilisées dans le langage courant

mais sont rajoutées aux dictionnaires de noms composés parce qu'elles proviennent de compositions productives : *appareil de N* et *installation de N*

Dans les exemples (40) à (44), les expressions composées ont été désolidarisées, et les différents composants recombinés pour créer des expressions coordonnées complexes mais qui manifestent une certaine compacité et évitent les répétitions.

appareils d'extinction ou dispositifs d'extinction et appareils d'alerte ou dispositifs d'alerte
(appareils ε ou dispositifs) d'extinction et (appareils ε ou dispositifs) d'alerte
(appareils ou dispositifs) d'extinction et (appareils ou dispositifs) d'alerte
(appareils ou dispositifs) d'extinction et <E> d'alerte
(appareils ou dispositifs) (d'extinction et d'alerte)

L'opération inverse – distribution du nom tête et recombinaison - permet d'éviter une mauvaise interprétation de l'exemple suivant.

(44) *compresseurs et circuits d'admission et d'échappement*

Les deux expressions *circuits d'admission* et *circuits d'échappement* sont des mots composés. Elles ont été coordonnées, désolidarisées de leur nom tête, et regroupées dans la séquence *circuits d'admission et d'échappement*. En reconnaissant prioritairement ces deux mots composés, on évite une interprétation fausse qui consisterait à classer (44) dans le type 5, et donc à le développer sous la forme :

**compresseurs d'admission et compresseurs d'échappement et circuits d'admission et circuits d'échappement*

Le développement correct est :

(compresseurs) et (circuits d'admission) et (circuits d'échappement)

2.5.2. Ajout d'un troisième modifieur

La structure des expressions coordonnées de type 5 se prête bien à l'ajout d'un troisième modifieur qui est mis en facteur et qui, dans l'expression reconstruite, se distribue sur chacune des expressions reconstituées.

N1 et N2 Mod1 et Mod2 Mod3 équivalent à
(N1 Mod1 Mod3) et (N1 Mod2 Mod3) et (N2 Mod1 Mod3) et (N2 Mod2 Mod3)

N1	***et N2***	***Mod1***	***et Mod2***	***Mod3***
bâtiments	*et installations*	*pour le traitement*	*et l'entreposage*	
des productions agricoles				
les plans	*et caractéristiques*	*du système d'extraction*	*et des conduits d'évacuation*	
des buées et fumées				

3. Description des règles de réécriture

Dans cette partie, on va déduire à la fois de la morphologie des éléments et de leur position dans le *GN*, et des règles syntaxiques générales de la langue, des indices qui permettent de proposer un développement du *GN* coordonné. La conjonction de ces indices s'exprime sous la forme de règles concernant :

- l'accord de l'adjectif ou du participe passé à valeur adjectivale. Elle permet de choisir entre les types 1 et 2 ;

- l'utilisation de l'adjectif possessif à droite de la coordination qui permet de répéter la partie gauche de la coordination afin de constituer un *GN* complet ;

- la répétition de la tête du *GN* dans le modifieur à droite de la coordination qui indique plutôt un *GN* coordonné de type 1 ;

- l'effacement du déterminant ou de la préposition dans la partie droite de la coordination qui permet de répartir les *GN* identifiés entre les types 2 et 3 ;

- la symétrie de la construction entre les composants droite et gauche de la coordination.

A la fin de ce paragraphe, on s'intéressera à des mots particuliers de la langue dont couramment les emplois permettent de classer a priori le *GN* coordonné selon la typologie (et donc proposer un développement), sans devoir l'analyser complètement.

notation : les éléments mis en facteur sont écrits en *italique*, ceux sur lesquels portent la coordination sont en **gras**. *(e)* désigne l'expression initiale et *(er)* l'expression reconstruite.

3.1. Accord de l'adjectif

(51) *construction navale et structure* maritimes
(52) *façade et baie* accessibles
(53) *corps gras d'*origines *animale et végétale*
(54) *élevage et reproduction animale*
(55) *latex et caoutchouc brut*

(56) *couleurs et mesurage de lumière*
(57) *grandeurs et unités* spécifiques *à certains domaines*

(58) salles *de sciences et de travaux pratiques*

Pour les modifieurs qui en prennent la marque, l'accord en genre et en nombre donne des indications catégoriques sur les noms auxquels ces modifieurs se rapportent.

Dans les exemples (51) et (52), l'adjectif mis en facteur est au pluriel, alors que le nom qui le précède immédiatement est au singulier, on peut donc en déduire que l'adjectif se rapporte aussi au nom qui se trouve du côté gauche de la coordination. Les expressions sont de type 2.

Dans (54) et (55), l'adjectif est au singulier, donc il ne peut se rapporter qu'au nom immédiatement à sa gauche et pas à celui qui se trouve de l'autre côté de la coordination. Ces deux expressions appartiennent donc au type 1.
La même règle joue dans (53) : l'adjectif est au singulier, le nom qui le précède au pluriel donc le modifieur qui est à droite de la coordination se rapporte aussi au même nom.

Pour les exemples (56) et (57), la syntaxe n'apporte pas d'information sur la distributivité éventuelle du modifieur : dans (56) parce que le modifieur ne prend pas la marque du genre ni du nombre, dans (57) parce que le modifieur est au pluriel, mais le nom qui le précède l'est aussi.

Règle 1.1 :

si *(e)* =: *N1 et N2 Mod*
Mod prend la marque du genre et du nombre
Mod est au pluriel
N2 est au singulier
alors *(e)* est de type 2
(er) =: *N1 Mod et N2 Mod*

Règle 1.2 :

si *(e)* =: *N1 et N2 Mod*
Mod prend la marque du genre et du nombre
Mod est au singulier
N2 est au singulier
alors *(e)* est de type 1
(er) =: *N1 et (N2 Mod)*

Règle 1.3 :

si *(e)* =: *N1 et N2 Mod*
Mod prend la marque du genre et du nombre
Mod est au masculin
N2 est au féminin
alors *(e)* est de type 2
(er) =: *N1 Mod et N2 Mod*

Règle 1.4 :

si *(e)* =: *N Mod1 et Mod2*
Mod1 prend la marque du genre et du nombre
Mod1 est au singulier
N est au pluriel
alors *(e)* est de type 3
(er) =: *N Mod1 et N Mod2*

3.2. Utilisation d'un adjectif possessif

(61) *les revêtements et leurs procédés*

(62) *les coques et leurs éléments de structure*
(63) *le bruit et ses effets sur l'homme*
(64) *les vibrations et leurs effets sur l'homme*

(65) *les entraînements par câble et leurs composants*
(66) *les entraînements par chaîne et leurs composants*
(67) *les entraînements par courroie et leurs composants*

Quand un possessif se trouve en tête du 2ème groupe nominal coordonné, la coordination porte sur les noms têtes de groupes nominaux et non, le cas échéant, sur les modifieurs. L'expression composée peut être classée dans le type 1.

Règle 2.1 :

si *(e)* =: *N1 Mod1 et DetPoss N2 Mod2*
alors *(er)*=: *N1 Mod1 et DetDef N2 Mod2 de DetDef N1 Mod1*

où

DetDef =: *le* | *la* | *les*

DetPoss =: *mon* | *ma* | *mes* | *ton* | *ta* | *tes* | *son* | *sa* | *ses* | *notre* | *nos* | *votre* | *vos* | *leur* | *leurs*

L'expression coordonnée n'est pas ambiguë et son développement n'apporte rien quant à la compréhension du texte. En revanche, si l'on utilise un procédé automatique, développer l'adjectif possessif et le coordination permet d'"attraper" un groupe nominal coordonné qui n'aurait pas été reconnu sinon, parce qu'il est elliptique dans la formulation initiale.

Le développement du *GN* coordonné contenant un possessif en partie droite peut néanmoins être délicat. En effet, si l'on observe l'exemple suivant :

avertir de l'existence d'un sinistre et de sa localisation

Le possessif *sa* ne se réfère pas au nom tête du *GN existence* mais à son modifieur *sinistre*. Et le développement produit en utilisant la règle 2.1 serait inepte :

**avertir de l'existence d'un sinistre et de l'existence de sa localisation*

On pourrait faire la même remarque pour l'exemple qui suit où *son* fait référence au modifieur *fabricant* et non au nom tête *adresse* :

adresse du fabricant ou de son mandataire

3.3. Répétition de la tête du groupe nominal (ou d'un dérivé) dans le modifieur

acoustique et mesurage acoustique
aluminium et alliages d'aluminium
béton et produits en béton
viande et produits à base de viande

plafonds et plafonds suspendus
filtres ou ensemble de filtration
plafonds et faux plafonds pleins
escaliers et escaliers mécaniques
cordon continu ou cordons discontinus
matériaux rigides de toute épaisseur et matériaux souples d'épaisseur supérieure à 5 mm
bois massifs et panneaux dérivés du bois

On peut construire[2] d'autres exemples illustrant le même procédé où le nom tête du premier groupe nominal est repris dans le 2ème groupe nominal coordonné au premier, pour former le modifieur :

chasse et armes de chasse
cuisines et ustensiles de cuisine

Cette remarque permet de classer, dans le type 1, les expressions construites sur ce modèle : le modifieur ne s'applique qu'au nom tête du 2ème groupe nominal et non aux deux groupes nominaux coordonnés.

On peut utiliser dans le 2ème groupe nominal non pas exactement le nom tête du 1er groupe nominal mais un terme dérivé :

chasse et abris de chasseurs
chasseur et armes de chasse
présidence et prérogatives présidentielles

---

[2] Ce type d'expressions se retrouvera dans des titres de rubriques ou de chapitres, plus que dans le corps d'un texte.

Les expressions suivantes sont plus difficiles à repérer, même si elles procèdent de la même idée :

chaudières et échanges de chaleur
viande et produits carnés
musique et instruments [de musique] à vent

Dans les deux premiers exemples, le modifieur appartient à la "même famille" mais n'a pas la même racine étymologique. Dans le dernier, la tête du 1er groupe nominal *musique* qui appartient aussi au modifieur est effacé.

Règle 3.1 :

si *(e)* =: *N1 et N2 Mod*
dans *Mod*, *A* ou *N* appartiennent à la même famille que *N1*
alors *(e)* est de type 1
(er) =: *N1 et (N2 Mod)*

NB : la notion de "même famille" n'est pas décrite dans ce travail, ni donc exploitée.

3.4. Effacement du déterminant ou de la préposition dans la partie droite de la coordination

(81) *tous les **plans et documents*** nécessaires
(82) *les **plans et renseignements de détail*** concernant les installations techniques
(83) *dans les **locaux et dégagements*** accessibles au public
(84) comportement au feu *des **matériaux et éléments de construction***

Dans (81) et (82), le déterminant qui précède le *GN* situé à droite de la coordination est effacé ; pour (83), la préposition qui introduit les deux noms coordonnés n'existe plus dans la partie droite ; dans (84), c'est *des*, contraction de *de-les*, devant le deuxième nom coordonné qui est effacé. Aucun des noms coordonnés ne possède de modifieur qui lui soit propre. La structure des *GN* pour lesquels cet effacement est le plus fréquent, est plus simple que la structure générale décrite en 1 :

GN* =: *Det N Et N Mod

Ce procédé stylistique a pour effet de relier plus étroitement la partie droite de la coordination à la partie gauche. En allégeant la construction de la phrase, cette mesure est d'autant plus utile que les modifieurs qui s'appliquent aux têtes du *GN* coordonné sont nombreux.

*les **(plans et descriptifs)** (de la distribution et du stockage) du combustible*
*les **(plans et caractéristiques)** (du système d'extraction et des conduits d'évacuation) (des **buées et fumées**)*

Cette construction confirme l'intuition que ce sont bien les mots immédiatement situés à droite et à gauche de la coordination qui sont coordonnés.

Règle 4.1 :

si *(e)* =: *Det N1 Et N2 Mod*
alors *(e)* est de type 2
(er) =: *(Det N1 Mod) Et (Det N2 Mod)*

Règle 4.2 :

si *(e)* =: *N Prep Det N1 Et N2*

alors *(e)* est de type 3
(er) =: *(N Prep Det N1 Mod) Et (N Prep Det N2 Mod)*

3.5. Symétrie de la construction à l'intérieur des groupes nominaux

(91)	*les* dégagements	*accessoires*	*et*	*supplémentaires*
(92)	dispositifs	*d'éclairage*	*et*	*de signalisation*
(93)	machines	*à aléser*	*et*	*à fraiser*
(94)	*une* ossature	*en matériaux de catégorie M3*	*et*	*en bon état*
(95)	salles	*de sciences*	*et*	*de travaux pratiques*
(96)	*la* ventilation	*de la cuisine*	*et*	*de la salle polyvalente*
(97)	*les* locaux	*"groupe électrogène"*	*et*	*transformateurs*

De chaque côté de la coordination, se trouvent des éléments qui partagent la même nature grammaticale et le même schéma de construction. Dans (91), ce sont deux adjectifs ; dans (92), *et* relie deux compléments de noms ; dans (93) et (94), ce sont des compléments introduits par la même préposition : *à*, *en* .
Pour des expressions du type : *N1 Mod1 et (MOT)**, la symétrie de la construction entre *Mod1* et *(MOT)** postule pour la reconnaissance de *(MOT)** comme un deuxième modifieur, coordonné au précédent si le début de la séquence peut être identifié comme un modifieur. On obtient alors une expression de la forme *N1 Mod1 et Mod2* que l'on peut classer dans le type 3.

Cependant, la symétrie de la construction peut être masquée, et la mise en évidence à droite de la coordination d'un modifieur de même statut que celui de gauche, présente alors quelques difficultés :

a) les deux modifieurs présentent des structures différentes :

les établissements ***existants** et **à modifier***
une unité de passage ***coupe-feu** de degré une heure et **à fermeture automatique***
les locaux ***appelés**"grandes cuisines" et **répondant**, selon le cas, aux dispositions ...*

b) les deux modifieurs sont eux-mêmes prolongés par des modifieurs :

des conduits ***cheminant** à l'extérieur du bâtiment et **pénétrant** directement dans les locaux*
les aménagements ***accessibles** au public et **situés** en élévation*

c) l'un seulement des modifieurs se prolonge par son modifieur :

utilisation *de **lampes** mobiles et **de bougies***
le temps nécessaire ***à l'alarme** et **à l'évacuation** des occupants de l'établissement et des locaux*

d) un nom (tête du groupe nominal ou du modifieur) est un nom composé :

les épreuves *de résistance mécanique et d'étanchéité*
accès *au local ou à l'emplacement* de stockage

e) un modifieur commence par une négation :

un emplacement ***non accessible** au public et **surveillé** pendant les heures d'exploitation de l'établissement*

f) les préposition ou déterminant précédant le 2ème modifieur ont été effacés :
conduits d'évacuation ***des buées et fumées***

g) l'utilisation de dictionnaires de mots composés masque la symétrie des constructions :

(30) ***appareils mobiles*** *et moyens divers*
(50) ventilation ***de la cuisine*** *et* ***de la salle polyvalente***

Dans (30), *appareils mobiles* est clairement un nom composé *N* (de type *NA*), alors que *moyens divers* forme une séquence *NA* beaucoup moins figée, sémantiquement vague et qui s'apparente plutôt à un *etc.* Si on conserve cette analyse, la structure de l'expression est : *N et N A* et masque complètement des symétries visuelle, auditive et grammaticale pourtant bien réelles.

Pour (50), le raisonnement est inversé : en première approche, l'expression est de la forme *N de N et de N A*. Mais *salle polyvalente* peut être reconnu comme un nom composé *N* (de type *NA*) et l'expression apparaît alors comme une structure symétrique : *N de N et de N* où la coordination relie deux modifieurs du nom tête.

On voit donc que, selon les cas, il est préférable de considérer la segmentation de la phrase faite en recherchant prioritairement les noms composés, alors que, dans d'autres situations, l'analyse en mots simples permet de reconnaître des symétries dans la structure des groupes nominaux, symétries qui disparaissent quand on utilise des dictionnaires de noms composés. Comme il n'est pas possible de trancher a priori sur les priorités d'utilisation des dictionnaires, on conduira en parallèle toutes les analyses possibles et on essaiera de mettre en évidence les éventuelles symétries. Reste à définir un critère de choix entre les différentes interprétations possibles : la reconnaissance d'une symétrie dans la construction n'est pas le gage de l'exactitude de l'analyse.

3.6. La partie droite de la coordination débute par un mot composé

le passage rapide *des flammes ou des* ***gaz chauds***
escaliers et ***escaliers mécaniques***
escaliers mécaniques et ***trottoirs roulants***
faux plafonds et ***plafonds suspendus***
poids et ***pouvoirs calorifiques***

les trappes et ***portes de visite*** *pratiquées*
comportement au feu des conducteurs et ***câbles électriques***
contrôle et ***vérifications techniques***

Quand la partie droite de la coordination commence par un nom composé de structure *N Mod*, il y a ambiguïté sur la manière de rattacher le modifieur *Mod* à la coordination : il peut se rapporter seulement au nom qui le précède immédiatement, ou bien aux deux noms constituant les parties gauche et droite de la coordination. Pour les cinq premiers exemples présentés, le membre droit est bien un nom composé et le modifieur se rapporte exclusivement au dernier nom. Dans ce cas, marquer les noms composés empêche de distribuer le modifieur sur la partie gauche de la coordination et prévient des réécritures fautives[3] (comme *flammes chaudes*, *faux plafonds suspendus* ou *poids*

[3] Si on développe *mécaniques* sur les parties gauche et droite de la coordination pour écrire *escaliers mécaniques et escaliers mécaniques*, la réécriture est fautive non pas parce que l'expression formée ne désigne aucun objet réel mais parce qu'elle répète une partie du *GN* initial.

calorifiques). Dans les trois derniers cas, les membres droits constituent bien des mots composés : *portes de visite*, *câbles électriques* et *vérifications techniques* mais il serait peut être aussi judicieux de distribuer le modifieur sur la partie gauche des coordinations pour former : *trappes de visite*, *conducteurs électriques* et *contrôle technique*.

3.7. Mise en facteur de deux modifieurs

*la fermeture des **éventuels** clapets et volets **propres** à ces locaux*
*ventilation **naturelle** haute et basse **permanente***

Dans les *GN* précédents, deux des modifieurs sont coordonnés *clapets et volets* pour le premier exemple et *haute et basse* pour le deuxième, deux autres modifieurs sont disposés de part et d'autre de la coordination (sans conjonction ni ponctuation) et s'appliquent aux deux éléments coordonnés.

Pour le deuxième exemple, des réorganisations des modifieurs semblent acceptables mais non attestées par le corpus : *ventilation haute et basse naturelle et permanente, ventilation naturelle permanente haute et basse* et *ventilation naturelle et permanente haute et basse.* Une variante comme *ventilation naturelle haute et basse permanente* paraît improbable, mais sur des critères sémantiques. Dans ce cas, il paraît plus approprié de consigner ces expressions complexes dans un dictionnaire plutôt que d'essayer de reconstituer des *GN* avec des règles générales de réécriture.

3.8. Indices sur la structure du *GN*, fondés sur le vocabulaire

On a observé que certains modifieurs induisaient, par leur valeur sémantique, une construction plutôt qu'une autre lorsqu'ils étaient employés dans un groupe nominal qui contient une coordination. Ce paragraphe a pour objet de montrer les indications sur la structure des groupes nominaux coordonnés – et donc sur la manière de construire l'expression complète – que l'on peut tirer des mots qui composent le groupe nominal.
Selon les cas, on peut soit proposer des règles de développement différentes de celles admises dans le cas général, soit considérer qu'il est possible d'ignorer le mot en question dans l'expression.

3.8.1. L'adjectif *autre*

3.8.1.1. L'adjectif *autre* dans une expression coordonnée à deux éléments

Les expressions que l'on se propose de traiter particulièrement sont construites sur le schéma suivant :

(Det+ε) N1 Et (Det+ε) <autre> N2 (Mod+ε)

les hottes ou autres dispositifs de captation
les divers locaux techniques et autres locaux à risques
les fusibles et autres dispositifs de protection contre les surintensités

Dans un groupe nominal qui contient deux éléments cordonnés par *et* ou *ou*, la reconnaissance de l'adjectif *autre* (précédé ou non par un déterminant) en tête de partie droite donne une indication sur la distributivité du modifieur éventuellement présent à droite de *N2* : le groupe nominal *N2 Mod* désigne, en général, toute une classe d'objets dont *N1* n'est qu'un représentant. Ce modifieur ne se rapporte alors qu'à *N2* et non à *N1 ;* et l'expression est de type 1.
Sémantiquement, *autre* fonctionne un peu comme un *etc* mais en indiquant en plus un lien

d'hyperonymie entre ce qui se trouve de part et d'autre de la coordination : *N2* est un hyperonyme de *N1* et ce qui va être rapporté s'applique à tous les *N2*, et donc aussi à *N1*.
Ainsi, *les hottes* ne sont qu'un *des dispositifs de captation*, *les locaux techniques* un exemple de *locaux à risques* et *les fusibles* un des *dispositifs de protection contre les surintensités*.

Si on utilise des dictionnaires de noms composés, *dispositifs de captation*, *locaux à risques* et *dispositifs de protection* sont reconnus comme des noms composés, et la question de la distributivité du modifieur par rapport à *N1* et *N2* disparaît pour les deux premiers exemples, et est repoussée au modifieur suivant *contre les surintensités* pour le dernier.

Prep (Det+ε) N1 Et (Prep+ε) (Det+ε) <autre> N2 (Mod+ε)

dans les locaux et les autres dégagements
en étages et autres ouvrages en élévation
dans les salles d'exposition et autres locaux accessibles au public

Dans les trois exemples précédents, *autre* appartient à un groupe nominal introduit par une préposition. Celle-ci figure devant *N1* mais est effacée devant *N2*, la partie droite du *GN* coordonné. Pour les deux derniers, et la préposition, et le déterminant, sont effacés devant *N2*.

Règle 6.1 :

si *(e) =: (Prep+ε)(Det+ε) N1 Et (Prep+ε)(Det+ε) <autre> N2 Mod2*
alors *(e)* est de type 1
(er) =: ((Prep+ε)(Det+ε)N1) Et ((Prep+ε)(Det+ε) <autre> N2 Mod2)

(Det+ε) N1 (Prep+ε) (Det+ε) N Et (Prep+ε) (Det+ε) <autre> N2 Mod2

(70) *chaque porte de chambre, ou de tout autre local accessible au public*
(71) *les plafonds suspendus des salles omnisports, et autres grands volumes assimilables*
(72) *les tuyaux de raccordement en métal ou autres matériaux incombustibles à paroi mince*

Dans (70) à (72), la coordination relie les modifieurs du nom tête du *GN*. Les expressions formées sur ce schéma sont alors de type 3 : le nom tête (nom composé *plafonds suspendus* dans (71) et *tuyaux de raccordement* dans (72)) est mis en facteur et doit être distribué sur chacun des modifieurs.
Pour (70), la préposition *de* qui introduit chacun des deux modifieurs est présente ; dans (71) et (72), et les prépositions (respectivement *de* et *en*), et les déterminants définis ont été effacés.

Règle 6.2 :

si *(e) =: (Det1+ε) N1 Prep1 (Det11+ε) N11 Et (Prep2+ε) (Det2+ε) <autre> N2 Mod2*
alors *(e)* est de type 3
(er) =: (Det1+ε) N1 Prep1 Det11 N11 Et (Det1+ε) N1 (Prep2+ε) (Det2+ε) <autre> N2 Mod2

dans les salles d'exposition et autres locaux accessibles au public
les plafonds suspendus des salles omnisports, et autres grands volumes assimilables

Une difficulté apparaît. Si, dans la règle 6.1., *N1* admet un modifieur *Mod1*, l'expression initiale s'écrit :

(e_1) =: *(Prep+ε)(Det+ε) N1 Mod1 Et <autre> N2 Mod2*

Si, dans la règle 6.2, *Prep2* et *Det2* dans la partie droite de la coordination sont effacés, l'expression initiale s'écrit :

(e_2) =: *(Det1+ε) N1 Prep1 (Det1.1+ε) N1.1 Et <autre> N2 Mod2*
(Det1+ε) N1 Mod1 Et <autre> N2 Mod2
où *Prep1 (Det1.1+ε) N1.1* constitue le modifieur *Mod1*.

Les deux expressions initiales (*e1*) et (*e2*) sont alors identiques et il devient impossible de trancher entre les deux développements possibles du groupe nominal.
On peut néanmoins proposer deux pistes pour limiter le nombre de cas ambigus :

- faire l'analyse des expressions en utilisant prioritairement des dictionnaires de noms composés. Ainsi le groupe nominal *N Prep Det N* peut, le cas échéant, être reconnu comme un nom composé et le développement de l'expression peut se faire selon la règle 6.1 ;

- utiliser la ponctuation. Dans (70) et (71) où la coordination relie les modifieurs et non les têtes de groupe nominal, une virgule sépare le premier modifieur de la conjonction de coordination. On peut formaliser cette indication en modifiant la règle 6.2 :

Règle 6.2' :

si *(e)* =: *(Det1+ε) N1 Prep1 (Det1.1+ε) N1.1 (,+ε) Et (Prep2+ε) (Det2+ε) <autre> N2 Mod2*
alors *(e)* est de type 3
(er) =: *(Det1+ε) N1 Prep1 Det1.1 N1.1 Et (Det1+ε) N1 (Prep2+ε) (Det2+ε) <autre> N2 Mod2*

3.8.1.2. *Autre* dans une énumération :

(73) *à l'intérieur des chambres, des dortoirs et autres locaux recevant du public*
(74) *les socles des prises de courant, les interrupteurs et autres appareillages installés dans les locaux accessibles au public*

Dans une énumération (i.e. une suite de noms accompagnés ou non de modifieurs, ayant la même fonction grammaticale et séparés par des virgules, sauf pour le dernier qui est relié au groupe précédant par une coordination), l'adjectif *autre* conserve la même fonction sémantique : indiquer que ce qui est à gauche de *autre* est un élément de l'ensemble des objets dont la définition suit immédiatement *autre*. Dans (73), *chambres et dortoirs* sont des exemples de *locaux recevant du public*. Le groupe de mots à droite de *et* est donc un modifieur au même titre que ce qui précède *et*, et la préposition composée *à l'intérieur de* est en facteur par rapport à chacun des modifieurs.

On pourrait aussi trouver les formulation suivantes avec la même signification et la même acceptabilité :

(73) a. *à l'intérieur des chambres, dortoirs et autres locaux recevant du public*
(73) b. *à l'intérieur des chambres, des dortoirs et des autres locaux recevant du public*

Règle 6.3 :

si *(e)* =: *Det1 N1 (Mod1+ε), (Det2+ε) N2 (Mod2+ε), ...(Detn+ε) Nn (Modn+ε) Et (Det+ε) <autre> N (Mod+ε)*

alors *(e)* est de type 3'
(er) =: *Det1 N1 Mod1*
Et Det2 N2 Mod2
...

Et Detn Nn Modn
Et Det N (Prep+ε)(Det+ε) <autre> Mod

Règle 6.4 :

si *(e) =: (Prep+ε)(Det+ε) N Mod1, Mod2, ... Modn*
Et (Prep+ε)(Det+ε) <autre> Mod

alors *(e)* est de type 3'
(er) =: (Prep+ε)(Det+ε) N Mod1
Et (Prep+ε)(Det+ε) N Mod2
...
Et (Prep+ε)(Det+ε) N Modn
Et (Prep+ε)(Det+ε) N (Prep+ε)(Det+ε) <autre> Mod

En revanche dans (73c) et en cohérence avec l'utilisation ordinaire de l'adjectif *autre*, les *chambres* sont toujours des *locaux recevant du public* mais pas les *dortoirs* qui échappent à la portée de *autre*. Les différentes formulations (73c), (73d), (73e) sont sémantiquement équivalentes :

(73) c. à l'intérieur des chambres et autres locaux recevant du public, des dortoirs
(73) d. *à l'intérieur des chambres et des autres locaux recevant du public, des dortoirs*
(73) e. *à l'intérieur des chambres et autres locaux recevant du public, dortoirs*

Enfin, les différentes règles précédentes ne permettent pas de développer correctement les exemples suivants :

matériel de pont et autres équipements et installations
les jeux (billards et autres jeux électriques ou électroniques)

Ces règles de reformulation n'ont pas été testées sur le corpus. Dans un premier temps, on a choisi d'ignorer l'apparition de *autre* dans la partie droite de la coordination et de considérer qu'une expression construite ainsi, est du type 1.

Dans la suite de ce paragraphe, on a simplement relevé des exemples, issus du corpus, dans lesquels les mots *correspondant*, *associé*, *similaire*, *assimilable*, et la préposition *entre* donnent des indications sur la manière de réécrire le *GN* complété. Les règles ou heuristiques de réécriture que l'on pourrait en déduire n'ont pas été formalisées.

3.8.2. L'adjectif *correspondant*

essais non destructifs et matériel correspondant
les incidents possibles et les risques correspondants

Dans ces exemples, l'adjectif *correspondant* placé à droite d'un nom - et cet ensemble nom plus adjectif constituant lui-même la partie droite d'une coordination - se comporte un peu comme un déterminant possessif. Une réécriture possible serait la suivante :

essais non destructifs et leur matériel correspondant
les incidents possibles et leurs risques correspondants

pour donner finalement :

essais non destructifs et le matériel (de + correspondant à) les essais non destructifs
les incidents possibles et les risques (de + correspondant à) les incidents possibles

Cette construction possible de l'adjectif *correspondant* - c'est à dire sans complément explicite - est à différencier d'un autre emploi où *correspondant* est un participe présent[4] et n'est pas nécessairement lié à l'utilisation d'une coordination :

l'installation doit être calculée pour le niveau correspondant au plus grand débit
le total de unités de passage correspondant aux effectifs cumulés
les plans correspondant aux installations

même si cela est possible :

le nombre et la largeur des dégagements correspondant à l'effectif du public admis

Enfin, l'emploi de *correspondant* peut se conjuguer avec la répétition de la partie gauche de la coordination dans la partie droite :

ce produit ou la famille de produits correspondante

et rend la réécriture automatique hasardeuse.

3.8.3. Les adjectifs *assimilable, associé, similaire, ...*

Ces adjectifs peuvent être considérés comme sémantiquement voisins de l'adjectif *correspondant* et sont aussi employés en partie droite d'une coordination :

machines et volumineux biens d'équipement assimilables
industries du pétrole et technologies associées
l'atmosphère des salles d'opération et des salles d'anesthésie associées
surfaces commerciales ou d'exposition ou locaux similaires

Ils pourraient aussi être réécrits en répétant la partie gauche de la coordination sous la forme d'un complément de nom ajouté à la partie droite pour donner (les parties rajoutées sont soulignées) :

machines et volumineux biens d'équipement assimilables à des machines
industries du pétrole et technologies associées aux industries du pétrole
l'atmosphère des salles d'opération et des salles d'anesthésie associées aux salles d'opération
surfaces commerciales ou d'exposition ou locaux similaires à des surfaces commerciales ou d'exposition

Mais les emplois de ces adjectifs sont aussi plus variés et rendent plus difficile la reformulation automatique des *GN* concernés :

est assimilable à une sous-station un local abritant un générateur ...
les équipement associées à la batterie
les deux faces ne sont pas similaires

Enfin ces adjectifs peuvent aussi se combiner avec *autre* :

salles omnisports et autres grands volumes assimilables
courroies de transmission, amortisseurs et autres éléments similaires

---

[4] La réécriture des expressions contenant l'adjectif *correspondant* proposée précédemment revient à construire une phrase où *correspondant* est transformé en participe présent.

3.8.4. Les cas où il est inutile de réécrire le *GN* contenant une coordination

Pour ne pas produire des réécritures non pertinentes ou fautives, il est utile de repérer les cas où les *GN* contiennent une coordination qu'il est, systématiquement ou presque, inutile de reformuler. On a relevé, et la liste est très loin d'être exhaustive, des éléments lexicaux qui induisent ce comportement.

La préposition *entre* laisse présager la présence d'au moins deux éléments à comparer, opposer, lier, ... et ces éléments vont le plus souvent être désignés en utilisant une coordination[5] qu'il n'est pas pertinent de réécrire. Le *GN* ainsi construit *entre GN1 et GN2* peut être classé a priori dans le type 2 :

entre ces lambris et les parois
calfeutrement entre conduit et paroi traversée
entre une bouche d'extraction des fumées et une amenée d'air

De même s'il s'agit d'une liste de *GN* :

entre le diamètre intérieur, la largeur entre flasques et le diamètre extérieur

L'expression *l'ensemble de* produit le même type de *GN* :

l'ensemble de la trappe et de ce mécanisme constitue un dispositif ...
l'ensemble des réserves et des locaux d'emballage installés

4. Coordination et juxtaposition

Le corpus présente des *GN* dans lesquels le nom tête est accompagné d'une liste de modifieurs séparés par des virgules ou des conjonctions de coordination (essentiellement *et* et *ou*), ou bien plusieurs noms têtes, coordonnés ou juxtaposés, admettent le même modifieur :

chapiteaux, tentes et structures itinérants
roue, arbre et volute en acier
injecter tout produit germicide, désinfectant ou désodorisant dans le flux d'air,
sous fourreau continu, résistant, étanche et ouvert à une extrémité

Il est important de repérer cette liste de modifieurs avant de procéder à une quelconque réécriture. En effet, développer la coordination qui lie, le plus souvent, les deux derniers noms têtes dans le premier cas augmente la distance entre les autres noms têtes et le modifieur partagé d'une part, et d'autre part, la réécriture fait perdre toute trace de l'éventuelle symétrie de la construction et donc tout moyen de repérer la mise en commun du modifieur. Dans les constructions où le nom tête est suivi d'une liste de modifieurs il est difficile, si on ne repère pas la coordination, d'identifier le nom tête auquel chacun des modifieurs se rapporte.
Ainsi, dans les exemples précédents, si on développe *tentes et structures itinérants* en *tentes itinérantes*[6] *et structures itinérantes*, on perd toute possibilité de repérer que *chapiteaux* peut et doit accepter aussi *itinérants* comme modifieur.
Enfin, dans la séquence *fourreau continu, résistant, étanche et ouvert à une extrémité,* il faut repérer la structure de liste pour pouvoir associer le nom tête *fourreau* à chacun de ses modifieurs.

---

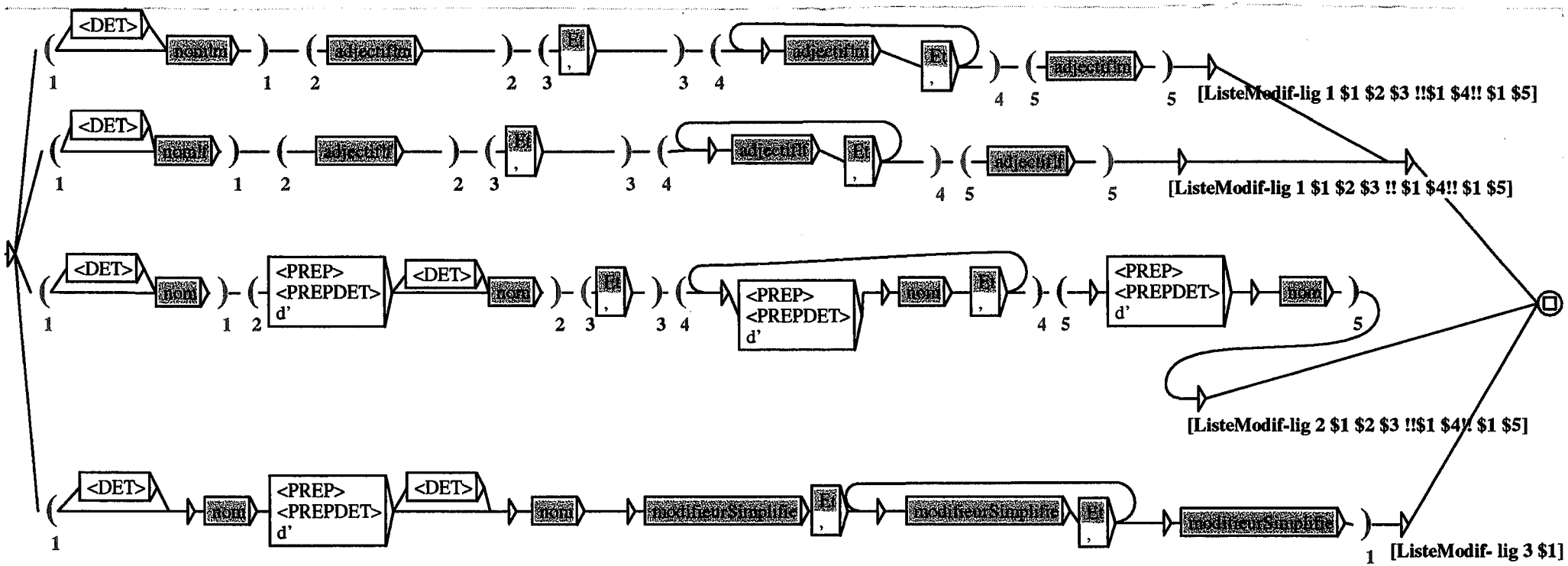
[5] Les éléments liés par la préposition *entre* peuvent être désignés en utilisant des déterminants au pluriel : *entre ces deux limites de format, entre ces formats,* ... et dans ce cas l'emploi de *entre* n'appelle plus une coordination.

[6] On considère ici comme résolu le problème de l'accord entre le modifieur et chacun des noms têtes auxquels il peut, éventuellement, se rapporter.

On utilise cette fois encore un transducteur pour procéder à ces identification et réécriture. Si on repère simplement les listes de modifieurs et le nom qui les précède immédiatement afin de distribuer chacun des modifieurs sur ce nom, on peut aussi fabriquer beaucoup de bruit :

les accessoires de distribution de vapeur, fluide liquide ou air chaud
le circuit d'extraction d'air vicié, de buées et de graisses
des courants d'air froids, directs, gênants pour le personnel
des fibres de coton neuves, non teintes et douces

En effet, dans les deux premiers exemples, chacune des listes de modifieurs se rapporte au nom qui précède immédiatement le premier modifieur : *distribution* et *extraction*. Dans les deux derniers, comme en témoignent les accords des modifieurs, ceux-ci concernent les noms têtes : respectivement *courants* et *fibres*. Ces quatre expressions ayant la même structure, on en déduit qu'il n'est pas possible de dégager des critères formels qui permettent de développer, d'une manière qui produit des résultats sémantiquement corrects, des *GN* dont le nom tête ne se limite pas à un nom seul (éventuellement précisé par un déterminant).


<DET>
nom/m
adjectif/m
Et
,
adjectif/m
1 1 2 2 3 3 4 4 5 5
[ListeModif-lig 1 $1 $2 $3 !!$1 $4!! $1 $5]
<DET>
nom/f
adjectif/f
Et
,
adjectif/f
1 1 2 2 3 3 4 4 5 5
[ListeModif-lig 1 $1 $2 $3 !! $1 $4!! $1 $5]
<DET>
nom
<PREP>
<PREPDET>
d'
<DET>
nom
Et
,
<PREP>
<PREPDET>
d'
nom
Et
,
<PREP>
<PREPDET>
d'
nom
1 1 2 2 3 3 4 4 5 5
[ListeModif-lig 2 $1 $2 $3 !!$1 $4!! $1 $5]
<DET>
nom
<PREP>
<PREPDET>
d'
<DET>
nom
nom
modifieurSimple/fe
Et
,
modifieurSimple/fe
Et
,
modifieurSimple/fe
1 1
[ListeModif- lig 3 $1]

Les automates présentés (le précédant et le suivant) ont deux objectifs :

- d'abord, identifier les expressions dont la délimitation du nom tête ne peut se faire d'une manière formelle, afin de ne pas leur appliquer de règles de réécriture.
  Les expressions trop complexes telles que celles décrites précédemment sont reconnues et non transformées par les transducteurs. Les *GN* de structure plus simple :

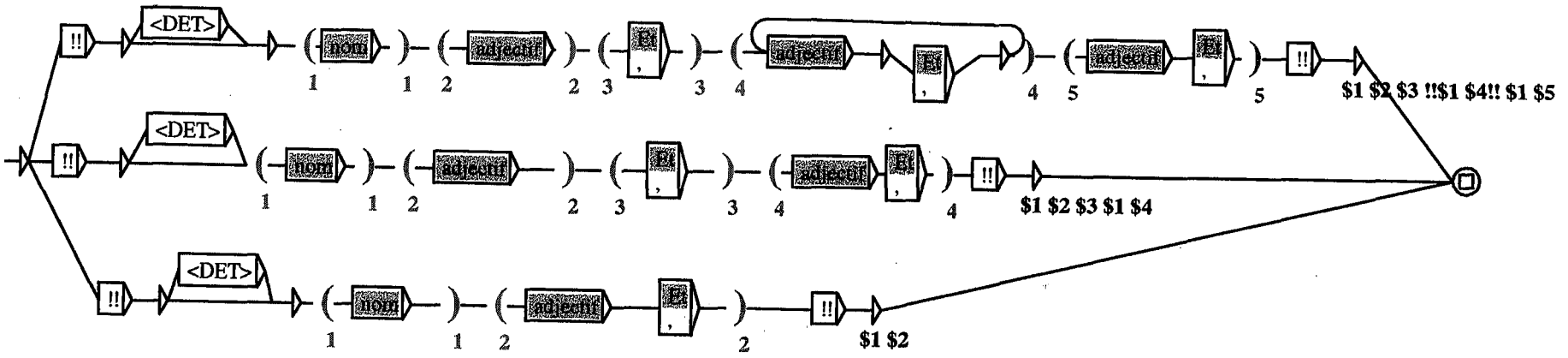
  une activité de conception, d'exécution ou d'expertise
  ses installations de chauffage, de refroidissement et d'aération

  sont réécrits en :

  une activité de conception, une activité d'exécution ou une activité d'expertise
  ses installations de chauffage, ses installations de refroidissement et ses installations d'aération

- ensuite, identifier et marquer les *GN* qui sont candidats à une reformulation. Il n'est pas possible, puisqu'on ne connaît pas a priori le nombre de modifieurs que comporte la liste, de construire d'automates qui permettent d'identifier et de réécrire le *GN* en une seule passe. Il faut réitérer, sur le texte reformulé, le procédé de transformation exprimé par un deuxième automate, tant que le transducteur reconnaît une expression qu'il peut transformer.
  Pour les listes de modifieurs qui comporte plus de trois éléments, dans la première étape on identifie la totalité de l'expression, on en développe une partie et on marque une sous-liste de modifieurs comme restant à développer. Dans les passes suivantes, on réitère le procédé dans la partie marquée : on distribue une partie des modifieurs et on marque la sous-liste qui doit être traitée. Les résultats intermédiaire et final obtenus sont les suivants (l'expression complète est entourée de *[crochets]* et la séquence marquée, et qui doit être traitée dans une passe ultérieure, est encadrée par *!!*) :

  [un système mécanique , !!un système électrique, pneumatique, hydraulique ou!! un système faisant] appel ...
  [un système mécanique , système électrique , !!système pneumatique,!! système hydraulique ou un système faisant] appel ...


<DET>
!!
nom
adjectif
Et
,
$1 $2 $3 !!$1 $4!! $1 $5
$1 $2 $3 $1 $4
$1 $2

L'ambiguïté concernant l'étiquette syntaxique, nom ou adjectif, qui est attachée à certains mots engendre des erreurs dans la segmentation et l'identification des séquences dans lesquelles le mot ambigu joue le rôle d'un adjectif, mais est étiqueté nom. Cela se produit dans la phrase suivante où la séquence [nom tête suivi d'une liste de modifieurs] est marquée entre crochets dans la phrase initiale :

l'organe [technique d'étude, de contrôle, d'information] du Préfet et du Maire

puis développée en :

l'organe [technique d'étude , !!technique de contrôle,!! technique d'information] du Préfet et du Maire

La séquence n'est guère plus explicite que le *GN* initial par rapport à une recherche automatique d'information, et éloigne le nom tête *organe* de ses compléments *du Préfet et du Maire.*

Même quand il n'y a pas d'erreur d'étiquetage ni de distribution entre tête et modifieur, la rusticité du transducteur qui, afin en particulier de ne pas distribuer à tort sur la totalité de la liste un modifieur qui ne concerne que le dernier d'entre eux, n'identifie pas les compléments des modifieurs fait que la réécriture du *GN* n'est pas complète et pourrait être plus explicite. C'est le cas des deux exemples suivants :

[travaux d'entretien , travaux de réparations ou travaux de modifications] effectués ...
[tout produit germicide, tout produit désinfectant ou tout produit désodorisant] dans le flux d'air

Une autre conséquence de la rusticité du constructeur est que l'identification de la liste s'interrompt dès qu'un modifieur a une structure plus complexe qu'un adjectif (au sens large comme le montre l'automate utilisé) ou un complément de structure *de N.* Dans la séquence suivante, *bandes de plastique* n'est pas reconnu comme un modifieur donc le nom tête *ligne* n'est pas distribué devant lui, ni devant le modifieur suivant *de grillage* :

[une ligne de brique, une ligne de tuiles, une ligne de bandes] de plastique ou de grillage

On pourrait effectuer et présenter le même travail pour la coordination des noms têtes, ce qui n'a pas été fait dans ce mémoire.

5. Stratégie de traitement de la coordination : "heuristiques simples" et "heuristiques confirmées"

L'ambition de ce travail est de diminuer le silence lors d'une recherche automatique d'information à l'intérieur d'un corpus en traitant la coordination à l'intérieur des *GN.* Pour cela, on s'est donné comme premier objectif de construire des *GN* complets, ou plus ou moins complets, mais pas de reformuler en totalité les phrases comportant des coordinations. En effet, pour des exemples tels que le suivant :

*toutefois pour les **établissements ou services spécialisés** pour recevoir des enfants en bas âges, des personnes handicapées ou des personnes âgées non hébergées dans des logements-foyers, le calcul se fera sur la base d'une personne pour deux lits, au titre des visiteurs.*

la construction du *GN* complet conduirait à dédoubler la phrase, pour obtenir les deux phrases indépendantes :

*toutefois pour les **établissements spécialisés** pour recevoir des ... le calcul ... visiteurs.*
*toutefois pour les **services spécialisés** pour recevoir des . le calcul ... visiteurs.*

Dans ce cadre, on propose deux stratégies pour reconstruire des *GN* complétés :

- ou bien des "heuristiques simples" fondées seulement sur la mise en œuvre des règles énoncées en 3. Dans ce cas, il n'est pas vérifié que les *GN* reconstruits appartiennent aux dictionnaires de mots composés existants. On peut ainsi repérer des séquences candidates à devenir des noms composés ;

- ou bien des "heuristiques confirmées", i.e. où la reconstruction des *GN* s'appuie à la fois sur les règles énoncées au paragraphe 3 et aussi sur les dictionnaires de mots composés existants. Dans ce cas, la distribution des modifieurs sur le nom tête, ou des noms têtes sur le modifieur n'est validée que si les *GN* ainsi complétés figurent dans les dictionnaires de mots composés. Cette reconstruction des *GN* ne consiste en fait qu'en une nouvelle présentation d'une information existant déjà dans les dictionnaires, et ne permet pas de repérer de nouveaux noms composés.

Dans un cas comme dans l'autre, la limite que l'on se propose de traiter pour la profondeur de la récursivité des *GN* considérés s'impose de la manière exposée au début de ce paragraphe.

Dans la suite du chapitre, on va examiner ces différentes règles et heuristiques, les conditions dans lesquelles on peut les appliquer et les résultats qu'elles permettent d'obtenir.

6. Mise en œuvre des règles de réécriture dans des heuristiques simples

Les règles présentées dans le paragraphe 3 sont traduites par des automates qui s'appliquent à un texte étiqueté au préalable à l'aide des **dictionnaires de mots simples** uniquement. Toutes ces opérations se font à l'aide d'INTEX.
Nous discutons d'abord du passage de règles de syntaxe (qui s'imposent au locuteur d'une langue), à leur traduction sous forme d'automates. Ceux-ci à cause des approximations nécessaires provoquent des erreurs d'application de ces règles pouvant aboutir à des réécritures des *GN* qui sont non pertinentes, voire fautives. Nous justifierons ensuite les différents détails de mise en œuvre pour chacun des transducteurs.

6.1. Heuristiques 1 : accord du modifieur

Les règles d'accord du modifieur avec le ou les noms auquel il se rapporte ne souffre pas d'exception. Mais leur traduction sous forme d'automates peut entraîner des erreurs. En effet, les règles d'accord s'appuie sur la délimitation des groupes syntaxiques, en particulier la segmentation en modifieur et nom tête, or ces syntagmes ne sont pas aisés à délimiter automatiquement : le nom qui précède la coordination peut être le nom tête ou bien du *GN* ou bien du modifieur de la tête du *GN*. Par exemple, pour les trois exemples suivants de structure *N Mod1 (et+ou) Mod2 Mod3* :

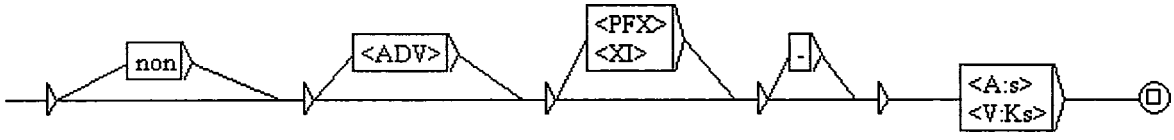
la liste de ces normes et de ces spécifications techniques
les façades en bordure de voies ou d'espaces libres
les mesures de prévention et de sauvegarde propres à assurer la sécurité des personnes

dans le premier, *Mod3 (techniques)* se rapportent à chacun des modifieurs qui le précèdent *Mod2* et *Mod1 ;* dans le deuxième, *Mod3 (libres)* concerne seulement *Mod2* et dans le troisième *Mod3* devrait

être rattaché au nom tête. Et c'est la segmentation proposée de fait par le transducteur qui, bien que celui-ci s'appuie sur les indices d'accord donnés par ces règles exactes, peut introduire des erreurs. On passe ainsi de la certitude d'une règle à l'approximation d'une heuristique.

D'autre part, seuls apportent des indications sur leurs genre et nombre les modifieurs qui supportent effectivement les marques de ces accords ; la grammaire proposée sous le nom *modifieur* doit donc être réécrite pour décrire les modifieurs au singulier et ceux au pluriel. On a indiqué que seuls les adjectifs qualificatifs et les participes passés employés comme adjectifs prenaient ces marques d'accord. Pour tenir compte des modifieurs qui peuvent être attachés à un adjectif, on a ajouté la possibilité qu'un adjectif soit précédé d'un adverbe comme dans *extinction non fortuite* ou *deux points bien visibles*, ou d'un préfixe comme dans *pré-enregistré*.

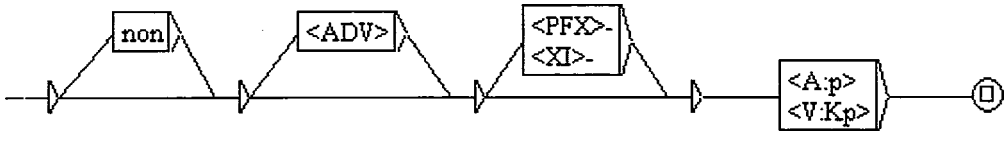
Les noms utilisés dans le texte peuvent, comme les adjectifs, être précédés de l'adverbe *non* comme dans *non arrêt*. Ils sont parfois formés d'un préfixe suivi d'un nom comme : *pré-nettoyage*, *sous-face* ou *demi-heure*. Enfin, selon les cas, on a besoin de préciser le genre ou le nombre, voire les deux. Les automates que l'on a construits en tiennent compte :



adjectif!s.grf : automate d'identification d'un adjectif au singulier

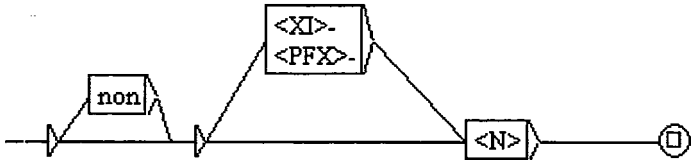
rappel :

<A:s>	désigne	un adjectif au singulier
<V:ks>		un participe passé au singulier
<PFX>		un préfixe
<XI>		un composant



adjectif!p.grf *:* automate d'identification d'un «adjectif au pluriel

<A:p>	désigne	un adjectif au pluriel
<V:kp>		un participe passé au pluriel



nom.grf *:* automate d'identification d'un nom

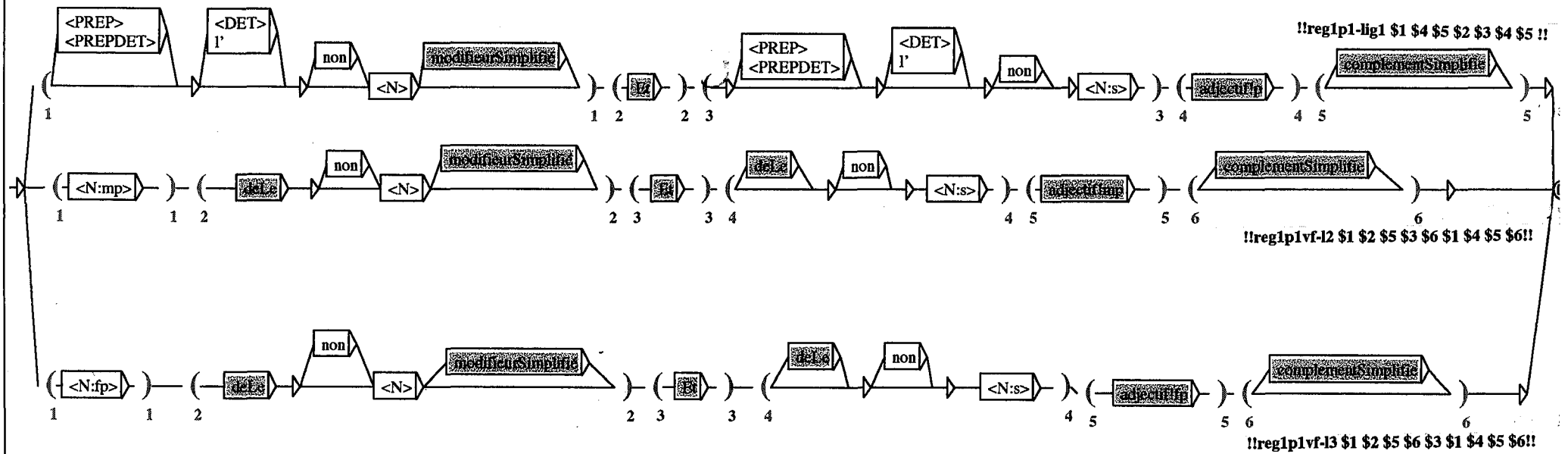
Pour tenir compte de ce qu'un *GN* peut constituer un complément prépositionnel, on a ajouté la possibilité que le nom tête soit précédé d'une préposition et d'un déterminant. Du point de vue de la recherche d'information, cette option ne se justifie pas mais elle apporte des informations utiles quant à la segmentation de la phrase :

avec les locaux et dégagement accessibles au public
sur certains aspects et coloris choisis par le laboratoire
par une personne ou un organisme agréés

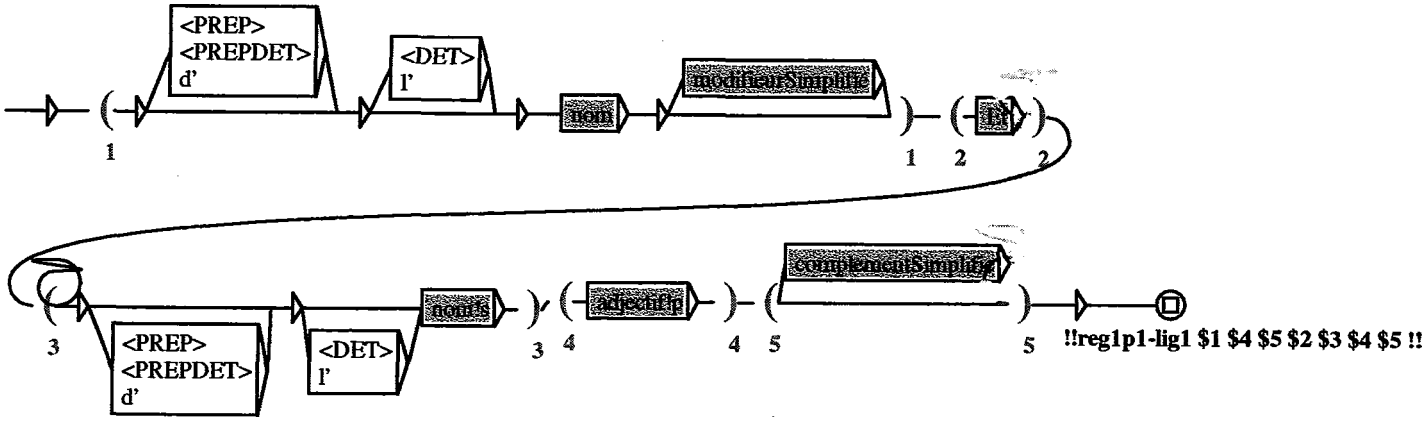
Enfin, le premier nom tête peut être accompagné d'un modifieur qui lui est propre *;* on a donc aussi rajouté cette possibilité dans les transducteurs.

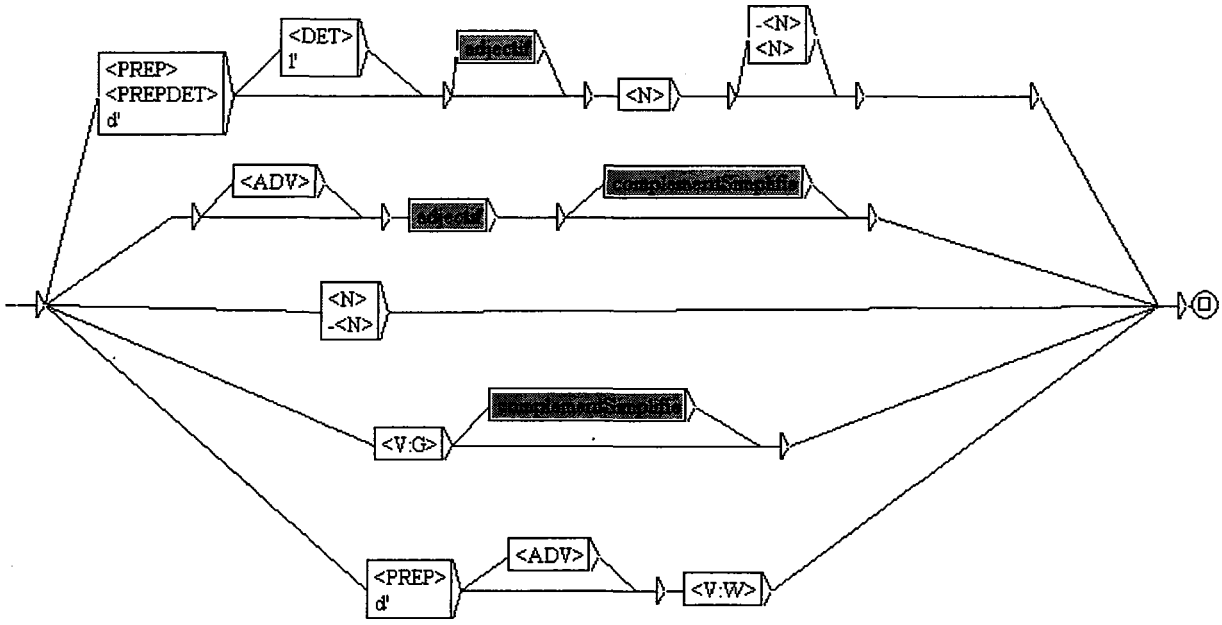
6.1.1. Heuristique 1.1

Le transducteur qui traduit, à l'intérieur du *GN,* les indications concernant la coordination et fondées sur les règles d'accord entre le modifieur et le nom auquel il se rapporte, est donc le suivant :

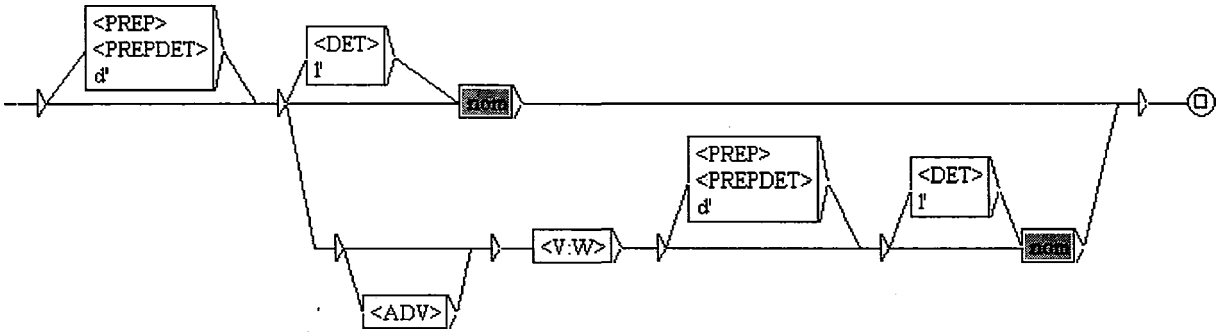

<PREP>
<PREPDET>
<DET>
l'
non
<N>
modifieurSimplifié
!!reg1p1-lig1 $1 $4 $5 $2 $3 $4 $5 !!
<N:mp>
delà
<N:fp>
<N:s>
adjectifmp
adjectiffp
complementSimplifie
!!reg1p1vf-l2 $1 $2 $5 $3 $6 $1 $4 $5 $6!!
!!reg1p1vf-l3 $1 $2 $5 $6 $3 $1 $4 $5 $6!!

Cependant, ce transducteur comporte une récursivité croisée : *GN* appelle *Compl* (par l'intermédiaire de *Mod*) qui appelle lui-même l'automate *GN*. INTEX ne permet pas d'appliquer ce transducteur au corpus en des temps raisonnables. On a donc choisi d'opérer des simplifications et de construire des automates moins généraux dans les descriptions des modifieurs et des compléments. Finalement les transducteurs utilisés pour appliquer la règle 1.1 sont les suivants :





modifieurSimplifie.grf : automate d'identification d'un modifieur simplifié



complementSimplifie.grf : automate d'identification d'un complément simplifié

Ce transducteur apporte des réécritures satisfaisantes, mais produit aussi du bruit. En particulier, la préposition admissible devant le nom tête de la partie droite de la coordination, et nécessaire pour repérer des *GN* comme :

avec les locaux et dégagement accessibles au public
sur certains aspects et coloris choisis par le laboratoire
par une personne ou un organisme agréés

fait que l'on reconnaît aussi les séquences suivantes tronquées (la partie non reconnue est indiquée entre crochets *[partie tronquée]*) :

[les quantités] d'azote ou de chlore contenues dans les matériaux ...
[les conditions] d'implantation et d'isolement prescrites au règlement ...
[des accroissements] de la largeur et de la hauteur compris ...

Dans les trois exemples précédents la préposition *de* introduit le premier modifieur du nom placé à gauche de ce modifieur (immédiatement ou séparé par un adjectif comme *dans le passage rapide des flammes ou des gaz chauds*), et non un complément circonstanciel du verbe introduit par une préposition comme *avec*, *sur* et *par* reconnus dans les trois premiers exemples. Les genre et nombre de l'adjectif situé en fin de *GN* sont alors ceux du nom tête puisque c'est à lui que l'adjectif se rapporte, et non ceux du modifieur.
Même s'il est possible que des prépositions autres que *de* introduisent aussi un modifieur du nom[7], on modifie le transducteur de la règle 1.1 de telle manière que si la première préposition est *de*, alors on recherche le nom situé à gauche de cette proposition et on vérifie son accord avec l'adjectif pour conclure que le *GN* est effectivement du type *(N Mod1 et Mod2 adjectif)* et se développe sous la forme *(N Mod1 adjectif) et (N Mod2 adjectif).* Cette heuristique permet de mieux segmenter les *GN* mais peut aussi produire du bruit, par exemple pour tous les verbes à construction transitive locative :

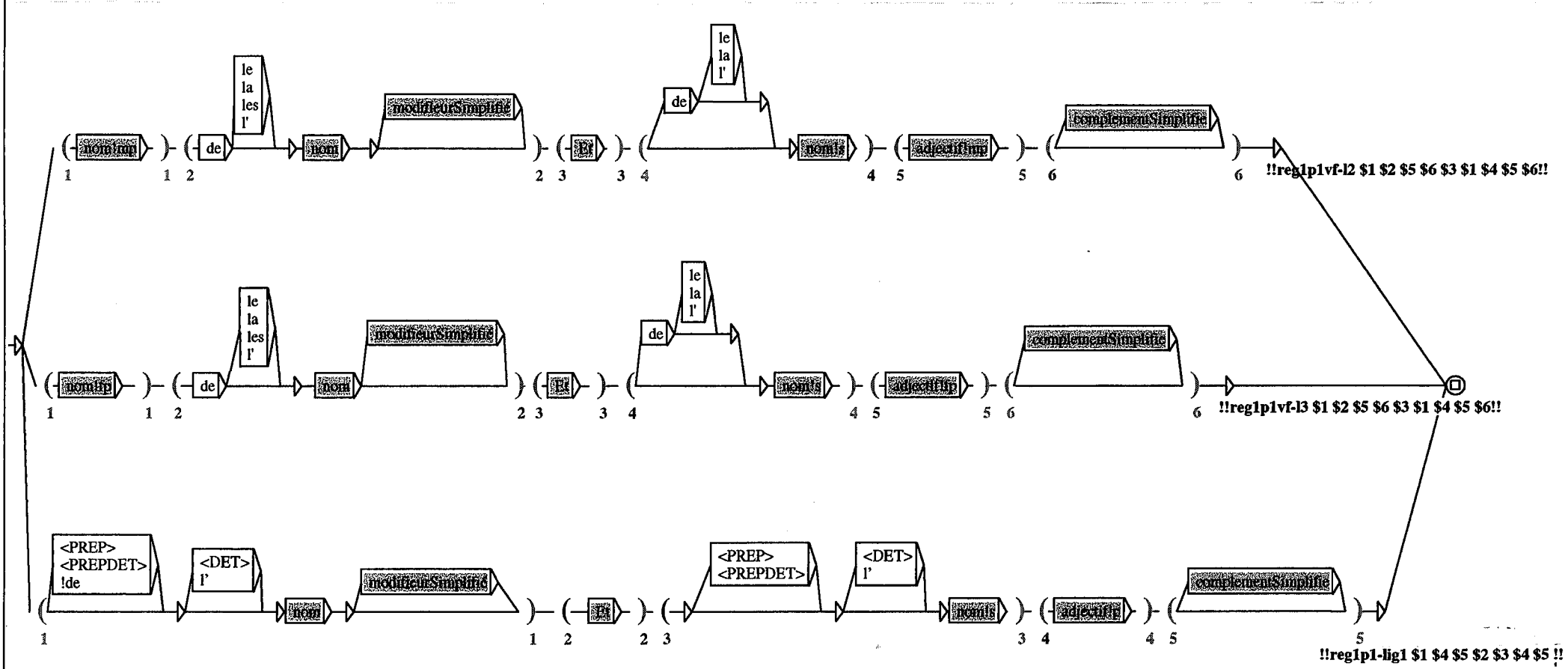
les secours éloignent les enfants des pièce et escalier enfumés

Dans cet exemple, le transducteur qui traduit la règle 1.1 rattachera *enfumés* à *enfants* pour construire deux *GN* coordonnés : *les enfants de la pièce enfumée* et *les enfants de l'escalier enfumés*

La dernière version proposée pour traduire la règle 1.1 est donc la suivante :

---

[7] comme par exemple *déjeuner sur l'herbe* ou *preuve par neuf.*


nom1mp
de
le
la
les
l'
nom
modifieurSimplifié
Et
nom1s
adjectif1mp
complementSimplifié
!!reg1p1vf-l2 $1 $2 $5 $6 $3 $1 $4 $5 $6!!
nom1fp
adjectif1fp
!!reg1p1vf-l3 $1 $2 $5 $6 $3 $1 $4 $5 $6!!
<PREP>
<PREPDET>
!de
<DET>
l'
adjectif1p
!!reg1p1-lig1 $1 $4 $5 $2 $3 $4 $5 !!

L'heuristique proposée pour la préposition *de* produit de bons résultats :

> *conditions d'implantation prescrites au règlement et conditions d'isolement prescrites au règlement*
> *quantités d'azote contenues dans les matériaux ou quantités de chlore contenues dans les matériaux*
> *qualités de réaction appropriées aux risques et qualités de résistance appropriées aux risques*

mais aussi des réécritures erronées ou incomplètes (les séquences reconnues par l'automate et réécrites sont soulignées) :

> *le tracé des dégagements et les mesures complémentaires de prévention proposées et complémentaires de protection proposées*
> *de la nature et de la hauteur des stockages de marchandises exposés ou stockages d'objets exposés ou de matériels entreposés*

Dans les réécritures précédentes, les *GN* qui contiennent plusieurs coordinations ou plusieurs modifieurs en cascade ne sont pas segmentés correctement.

> *Les éléments de décoration flottants ou éléments de habillage flottants tels que*
> *Les éléments flottants de décoration intérieurs ou flottants de habillage intérieurs tels que*

Ici, à cause de l'ambiguïté de *flottant* qui peut être étiqueté *nom* ou *adjectif*, le développement de *éléments de décoration ou d'habillage flottants* et *éléments flottants de décoration ou d'habillage* ne produit pas le même résultat.

Dans le cas où ni la préposition, ni le déterminant ne figurent dans la partie droite de la coordination, le *GN* reconstruit paraît sibyllin. Si, au lieu de :

> *quantités d'azote ou de chlore contenues dans les matériaux*

le texte initial est rédigé ainsi (la préposition est effacée devant le deuxième modifieur *chlore*) :

> *quantités d'azote ou chlore contenues dans les matériaux*

le *GN* est reconstruit de la manière suivante :

> *quantités d'azote contenues dans les matériaux ou quantités chlore contenues dans les matériaux*

6.1.2. Heuristique 1.2

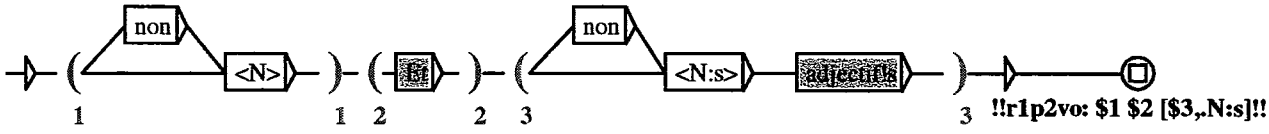
Elle est complémentaire de la précédente, ne donne lieu à aucune réécriture puisque ni le nom tête du *GN,* ni l'un des modifieurs ne se distribuent, mais peut donner des indications quant à la délimitation des composants à l'intérieur des *GN.*

Si l'on s'en tient à des contraintes minimales sur la structure du *GN,* par exemple avec le graphe suivant, on ne peut plus reconnaître les listes de modifieurs juxtaposés ou coordonnés et dépendant d'un même nom tête :

> *fourreau continu, résistant, étanche et ouvert à une extrémité*
> *groupe moteur thermique-générateur*

En effet, puisque l'ensemble (tête suivie du premier modifieur) *fourreau continu* est isolé, les autres modifieurs de la liste ne sont plus rattachables à aucun nom tête. Il faut donc repérer et traiter les

listes de modifieurs rattachés au même nom tête, et les listes de noms têtes admettant le même modifieur, avant d'appliquer cette heuristique.



Par ailleurs, les règles d'accord le plus souvent observées dans le corpus montrent que si la tête du *GN* est constituée par deux noms coordonnés par *ou*, c'est avec le premier nom seul que sont effectués les accords des modifieurs :

> *fourreau ou gaine continu, résistant, étanche et ouvert à une extrémité*
> *local ou dégagement accessible au public*

alors qu'en fait, les accords devraient être :

> *fourreau ou gaine continus, résistants, étanches et ouverts à une extrémité*
> *local ou dégagement accessibles au public*

puisque les modifieurs se distribuent sur les deux noms têtes pour donner :

fourreau continu, résistant, étanche et ouvert à une extrémité ou gaine continu, résistant, étanche et ouvert à une extrémité
local accessible au public ou dégagement accessible au public

On en conclut que, dans ce corpus au moins, il ne faut pas appliquer l'heuristique 1.2 aux *GN* coordonnés par *ou*.

Enfin, on se heurte au même problème que pour la règle 1.1 si l'on n'admet pas de modifieur à droite du membre gauche de la coordination : dans ce cas, le nom modifieur est considéré comme le nom tête de la partie gauche de la coordination.

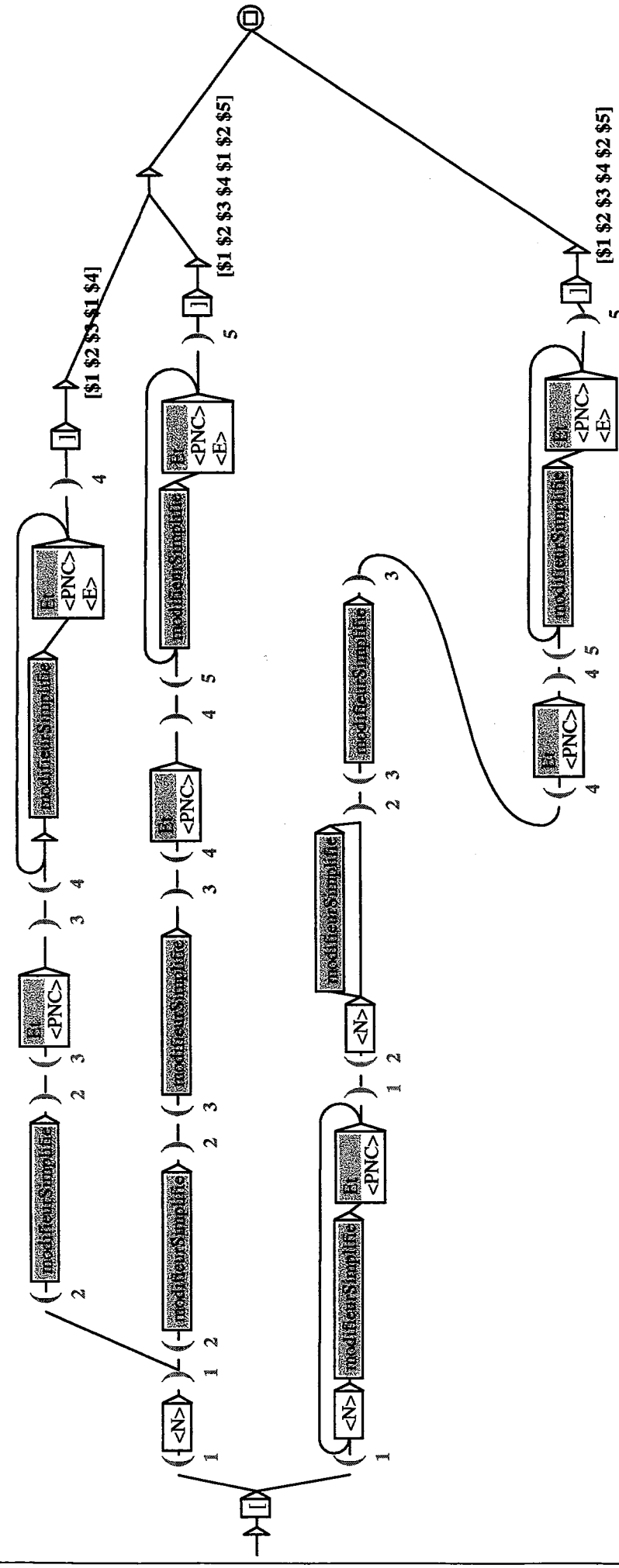
[économie d']énergie et isolation thermique
[batterie d']accumulateurs ou groupe moteur thermique

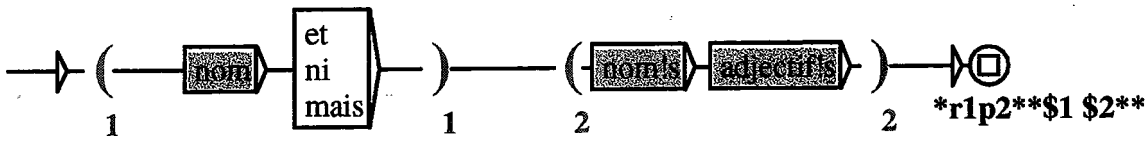
Dans les exemples précédents, le transducteur reconnaît *énergie* et *isolation* comme deux noms coordonnés par *et*, et essaie de rattacher *thermique*. Deux solutions sont possibles : au seul nom qui le précède immédiatement ou aux deux noms coordonnés. Le fait que *thermique* soit au singulier (ce qui est cohérent avec le fait que *isolation thermique* soit un nom composé) indique qu'il ne se rapporte pas à *énergie*. Ces essais de segmentation même fautive n'entraînent pas, cette fois-ci, une mauvaise distribution du modifieur de la partie gauche de la coordination (du moins tant qu'il reste au singulier) puisque si ce dernier modifieur se rapportait aussi à un nom situé en partie gauche, il porterait la marque du pluriel.

On construit donc deux autres transducteurs (présentés dans les deux pages suivantes) pour tenir compte de ces remarques :

- un pour reconnaître les listes de modifieurs (pour être complet, il faudrait aussi repérer les listes de noms têtes, cela n'a pas été fait ici) ;

- un autre pour rendre compte de l'heuristique 1.2


[$1 $2 $3 $1 $4]
[$1 $2 $3 $4 $1 $2 $5]
[$1 $2 $3 $4 $2 $5]
modifieurSimplifié
Et
<PNC>
<E>
<N>



On obtient ainsi des expressions correctement parenthésées (entre *[crochets]*) :

conduit et [paroi traversée,.N:s]
produit ni [forme harmonisée,.N:s]
surlargeur et [rayon intérieur,.N:s]

mais aussi des segmentations fautives (les parties entre *<chevrons>* ne sont pas reconnues par les transducteurs)

<ventilation naturelle> haute et [basse permanente,.N:s]
<un endroit accessible en> permanence et [bien signalé,.N:s]
<une rangée de sièges> relevés et [une rangée,.N:s] <de sièges inclinés>

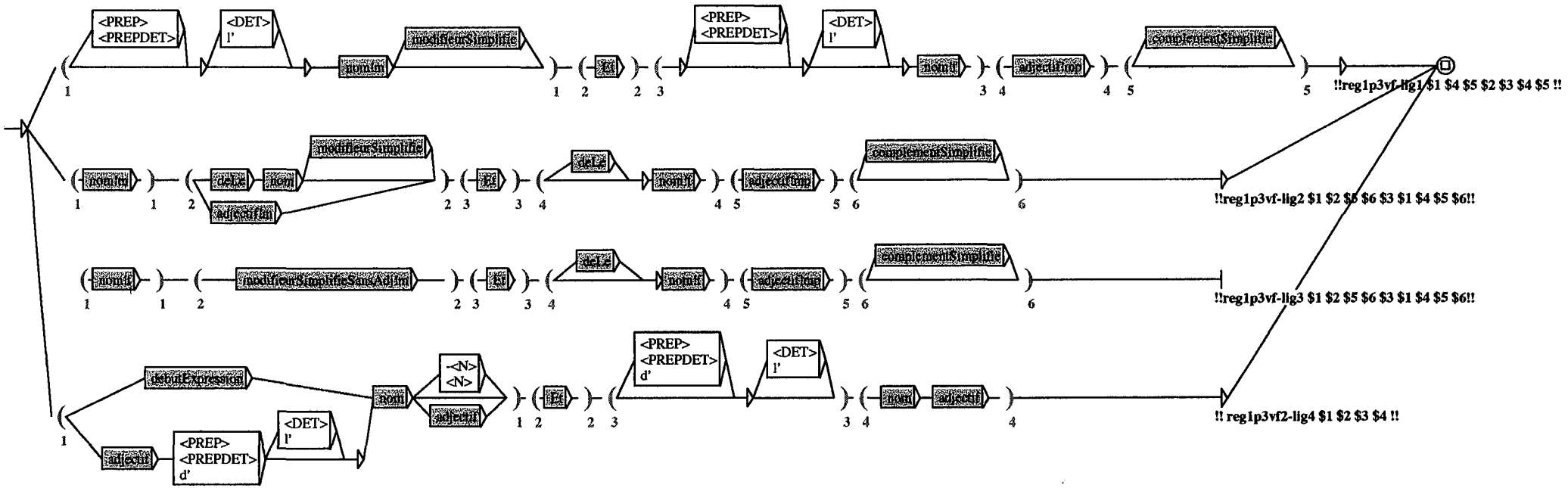
tout en ayant perdu (à cause de la restriction sur la coordination) :

avis ou [justification expérimentale,.N:s]
magasin ou [centre commercial,.N:s]
examen ou [intervention quelconque,.N:s]

Les segmentations fautives sont dues à des ambiguïtés d'étiquetage : l'étiquette *N* est possible et retenue, alors que dans les exemples précédents *haute et basse* sont des adjectifs, *bien* un adverbe et *une* un déterminant.

6.1.3. Heuristique 1.3

Pour construire ce transducteur on va se fonder sur la règle 1.3. et l'expérience de la règle 1.1. On obtient ainsi le transducteur suivant :


<PREP>
<PREPDET>
<DET>
l'
nom/m
modifieurSimplifié
Et
<PREP>
<PREPDET>
<DET>
l'
nom/f
adjectif/mp
complémentSimplifié
!!reg1p3vf-lig1 $1 $4 $5 $2 $3 $4 $5 !!
nom/m
deLe
nom
modifieurSimplifié
adjectif/m
Et
deLe
nom/f
adjectif/mp
complémentSimplifié
!!reg1p3vf-lig2 $1 $2 $5 $6 $3 $1 $4 $5 $6!!
nom/f
modifieurSimplifiéSansAdj/m
Et
deLe
nom/f
adjectif/mp
complémentSimplifié
!!reg1p3vf-lig3 $1 $2 $5 $6 $3 $1 $4 $5 $6!!
début Expression
adjectif
<PREP>
<PREPDET>
d'
<DET>
l'
nom
<N>
<N>
adjectif
Et
<PREP>
<PREPDET>
d'
<DET>
l'
nom
adjectif
!! reg1p3vf2-lig4 $1 $2 $3 $4 !!

Celui-ci sélectionne bien les expressions que l'on veut atteindre avec cette règle :

de fragments ou de gouttes enflammés
des coffrets ou de armoires fixés à des éléments
des gaz et fumées produits par la combustion
en coffret ou en niche réalisés dans le mur
isolement et distribution intérieurs
les branchements et canalisations situés à l'intérieur

et réalise un nombre important de réécritures correctes :

de fragments enflammés ou de gouttes enflammés
des coffrets fixés à des éléments ou de armoires fixés à des éléments
des gaz produits par la combustion et fumées produits par la combustion
en coffret réalisés dans le mur ou en niche réalisés dans le mur
isolement intérieurs et distribution intérieurs
les branchements situés à l'intérieur et canalisations situés à l'intérieur

Cependant du fait de l'ambiguïté en genre des adjectifs comme *complémentaire*, *électrique*, *métallique*, *automatique*, *sensible*, *inflammable*, *fixe*, *multiple* ... des sélections abusives sont faites qui conduisent à des réécritures fautives.

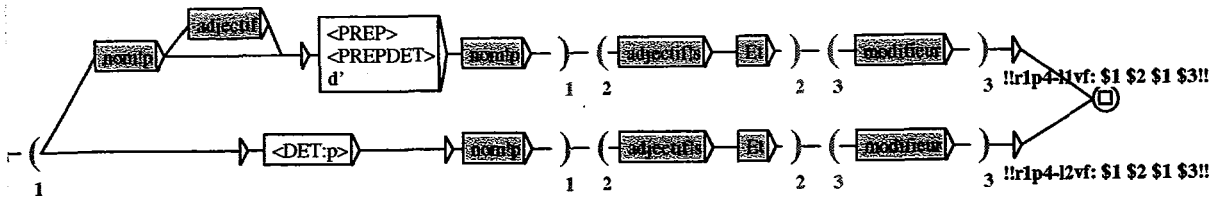
du présent article et des instructions complémentaires
les installations de désenfumage et aux installations électriques
parois verticales d'isolement ou aires libres d'isolement
armé de fibres de verre ou à armatures métalliques

En effet, si les formes au masculin et aux féminin étaient différentes pour les adjectifs précédents, ceux-ci, puisqu'ils ne se rapportent qu'au nom immédiatement situé à leur gauche, auraient une forme féminine et les conditions d'application de cette heuristiques ne seraient pas satisfaites. A cause de cette ambiguïté, l'heuristique est applicable et conduit aux réécritures suivantes :

du présent article complémentaires et des instructions complémentaires
les installations de désenfumage électriques et aux installations électriques
parois verticales d'isolement libres d'isolement ou aires libres d'isolement
armé de fibres de verre métalliques ou à armatures métalliques

6.1.4. Heuristique 1.4

Le transducteur correspondant à la règle 1.4 est celui-ci :



Des ajustements ont été faits en étudiant les résultats obtenus, dans le but de restreindre le nombre d'expressions sélectionnées. En particulier, des noms comme *accès*, *bois*, *gaz* étant ambigus quant à leur nombre, on a rajouté des contraintes sur le déterminant qui les précède.

Cet automate permet ainsi d'identifier des expressions telles que :

constructions des secteurs sanitaire et social
les couleurs rouge et orange étant interdites
les planchers haut et bas du local
quantités d'acides chlorhydrique et cyanhydrique
zones d'accès particulièrement difficile ou défavorable

de les développer correctement en :

constructions des secteurs sanitaire et constructions des secteurs social
les couleurs rouge et les couleurs orange étant interdites
les planchers haut et les planchers bas du local
quantités de acides chlorhydrique et quantités de acides cyanhydrique
zones de accès particulièrement difficile ou zones de accès défavorable

mais sélectionnent aussi (les séquences sélectionnées sont <u>soulignées</u>) :

le coefficient de massivité <u>des éléments soit inférieur ou égal</u> à celui de l'essai
plusieurs <u>exploitations de types divers ou de types similaires</u>
<u>installations de gaz combustible ou d'hydrocarbures liquéfiés</u>

pour les reconstruire de la manière suivante :

le... massivité <u>des éléments soit inférieur ou des éléments égal</u> à celui de l'essai
plusieurs <u>exploitations de types divers ou exploitations de types de types similaires</u>
<u>installations de gaz combustible ou installations de gaz d'hydrocarbures liquéfiés</u>

On peut déduire des remarques précédentes que le repérage de constructions symétriques dans les modifieurs comme *de types ... ou de types ...* doit être effectué avant de mettre en œuvre les heuristiques.

6.2. Heuristiques 2

La règle 2.1 concerne l'emploi d'un possessif à droite de la coordination. La difficulté réside dans la détermination du référent du possessif. Les exemples trouvés dans le corpus montre de nombreux cas de figures :

l'adresse du fabricant ou de son mandataire
une bouche d'extraction et son raccordement
ces appareils et leurs canalisations
pour leur accès et leur évacuation, ils sont ...

Pour le premier exemple, le possessif *son* fait référence au modifieur *fabricant* du nom tête *adresse*. Dans le deuxième, il fait référence au nom tête *bouche*. Pour le troisième exemple, *leurs* est coréférent au nom situé à gauche de la coordination. Enfin dans le dernier exemple, le déterminant possessif placé à droite de la coordination *leur* fait référence au sujet de la phrase, et non au nom situé en partie gauche de la coordination. Dans ce dernier cas, le fait que ce dernier nom soit lui-même précédé d'un possessif est un indice pour que le possessif de la partie droite n'y fasse pas référence.

Pour construire le transducteur, on va donc essayer de repérer les *GN* qui contiennent un nom tête accompagné de plusieurs modifieurs à la suite, puis une coordination suivie d'un déterminant possessif et on décide de ne pas développer ces *GN* puisqu'il est impossible, sur des critères formels, de déterminer à quel nom rattacher le possessif. On ne transforme pas non plus le possessif en un complément de nom quand le déterminant qui précède ce nom est lui-même un possessif ; dans ce cas en effet, il est plus probable que les deux possessifs soient coréférents au même mot, et non le deuxième possessif coréférent au nom qui le précède.
Enfin, lorsqu'on peut vérifier que le nom qui précède le possessif n'est pas un composant d'un modifieur, on choisit de reconstruire le *GN* selon la règle 2.1.

Dans les résultats obtenus, une partie importante des expressions sélectionnées n'est pas destinée à être reformulée pour les raisons détaillées au-dessus. On laisse donc telles quelles :

un seul niveau de sous-sol accessible à le public et son point le plus bas doit être ...
une visite périodique doit être effectuée par l'utilisateur ou son représentant
les conduits de évacuation doivent être construits en matériaux incombustibles et leur face intérieure ...
température des ailettes du moteur et de ses paliers
situés à l'intérieur des bâtiments de habitation ou de leur dépendances
une paroi présentant une trémie obturée par un conduit et son calfeutrement périphérique

mais on développe correctement :

celles-ci, du nombre de robinets ou d'orifices à desservir et de la hauteur de orifices à desservir
Le support fait un angle de 45° avec le plan horizontal et la base de plan horizontal est à 250 mm de
le présent arrêté et les annexes de présent arrêté ont pour objet

Il demeure des cas où, bien que le rattachement du possessif se fasse correctement, l'information contenue dans l'expression produite est tronquée ou devient inexacte. C'est le cas lorsque le nom qui suit le possessif est en fait un nom composé comme dans :

le laser et son dispositif de déviation optique

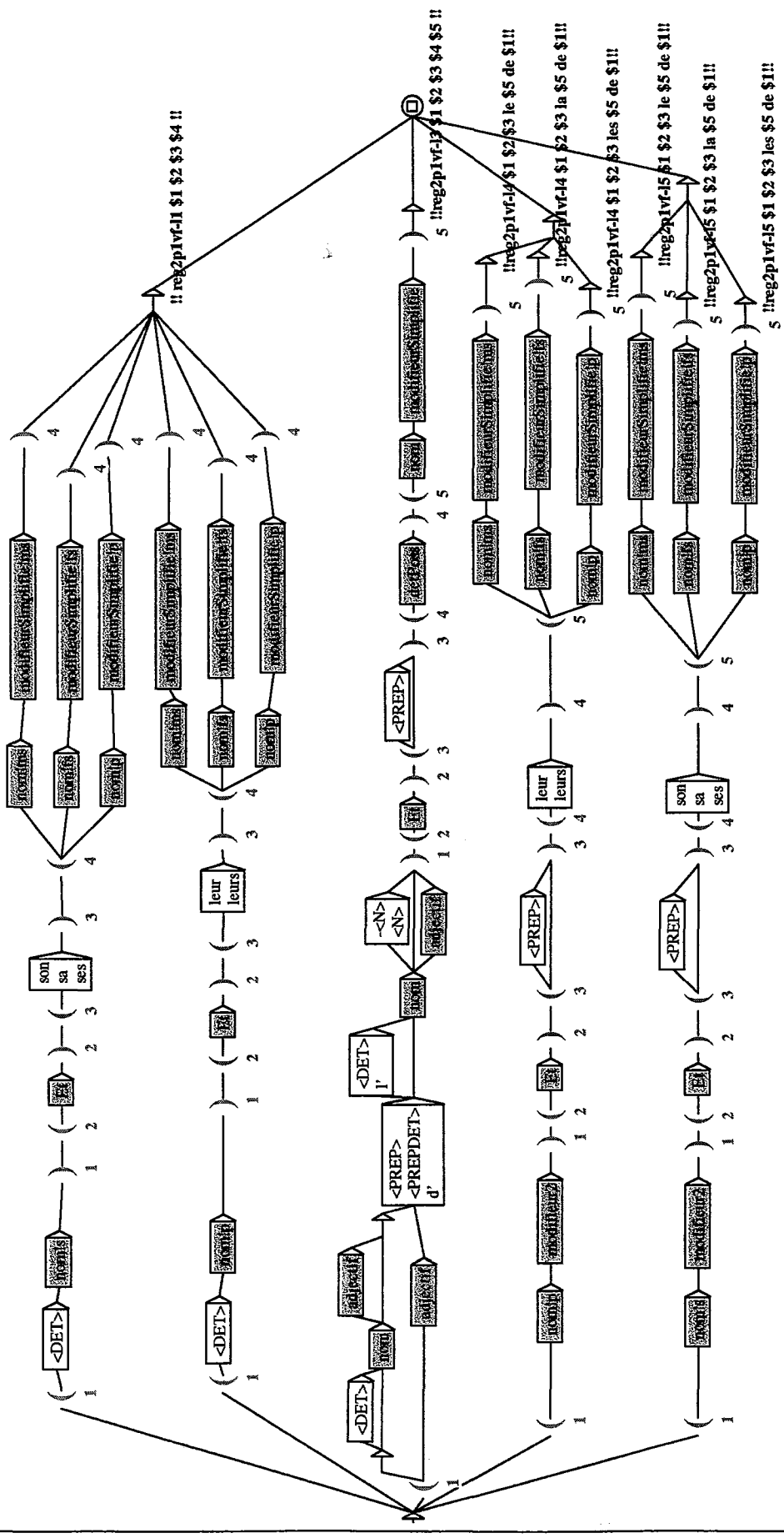
La réécriture proposée est la suivante :

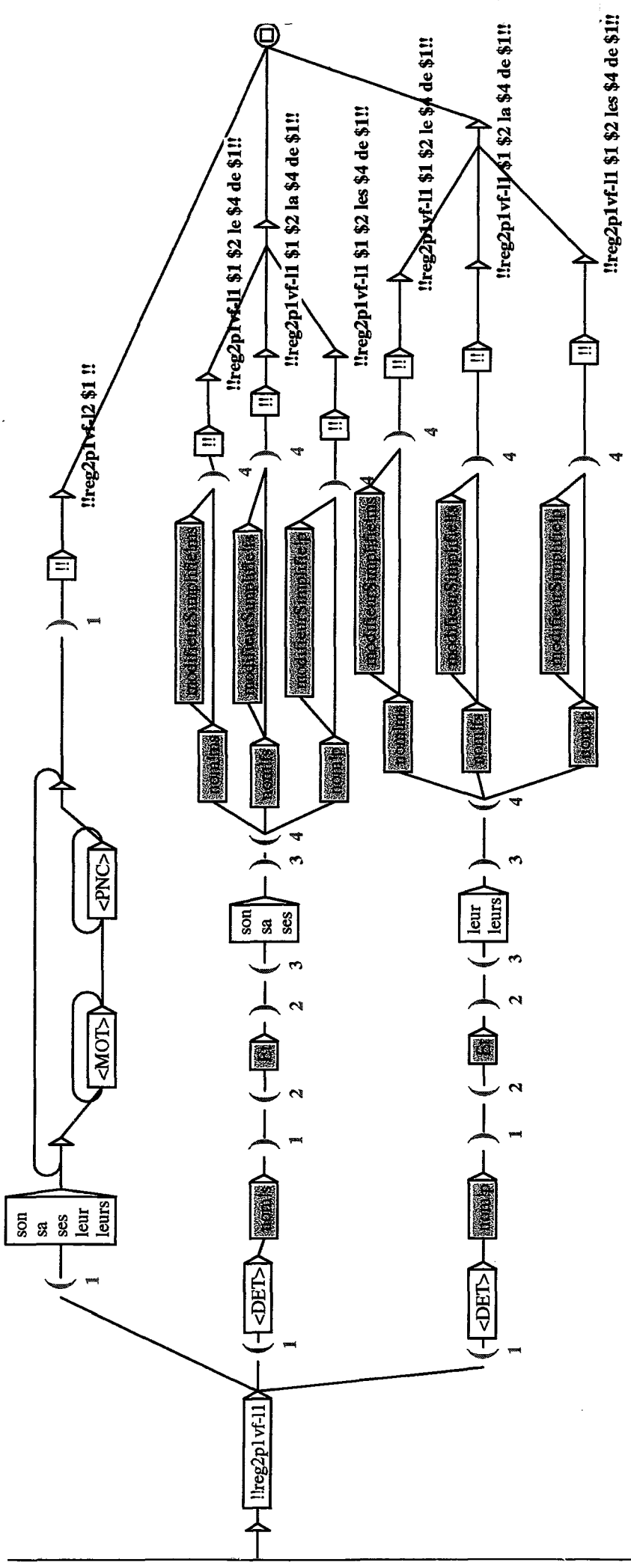
le laser et le dispositif du laser de déviation optique

alors que la reformulation correcte serait :

le laser et le dispositif de déviation optique du laser

Les réécritures sont réalisées en deux passes à l'aide des deux automates qui sont présentés à la suite.


!! reg2p1vf-l1 $1 $2 $3 $4 !!
!!reg2p1vf-l3 $1 $2 $3 $4 $5 !!
!!reg2p1vf-l4 $1 $2 $3 le $5 de $1!!
!!reg2p1vf-l4 $1 $2 $3 la $5 de $1!!
!!reg2p1vf-l4 $1 $2 $3 les $5 de $1!!
!!reg2p1vf-l5 $1 $2 $3 le $5 de $1!!
!!reg2p1vf-l5 $1 $2 $3 la $5 de $1!!
!!reg2p1vf-l5 $1 $2 $3 les $5 de $1!!
<DET>
<PREP>
<PREPDET>
d'
l'
<N>
son
sa
ses
leur
leurs


son sa ses leur leurs
<PNC>
<MOT>
!!reg2p1vf-l2 $1 !!
<DET>
!!reg2p1vf-l1
!!reg2p1vf-l1 $1 $2 le $4 de $1!!
!!reg2p1vf-l1 $1 $2 la $4 de $1!!
!!reg2p1vf-l1 $1 $2 les $4 de $1!!
!!reg2p1vf-l1 $1 $2 le $4 de $1!!
!!reg2p1vf-l1 $1 $2 la $4 de $1!!
!!reg2p1vf-l1 $1 $2 les $4 de $1!!

6.3. Heuristiques 3

Ces heuristiques ont pour objectif de formaliser la règle 3.1 qui concerne l'utilisation dans la partie droite de la coordination d'une partie du membre gauche comme dans :

béton et produits en béton

mais aussi :

chasse et abris de chasseurs
filtres ou ensembles de filtration
plafonds et faux plafonds pleins

Dans ce travail, nous n'avons pas essayé de formaliser la notion de "même mot" ou "même famille" étymologique ou sémantique. Nous n'avons donc pas construit les automates correspondant.

6.4. Heuristiques 4

Cette heuristique met à profit l'effacement de la préposition et du déterminant en tête de la partie droite de la coordination, pour distribuer le modifieur de cette partie aussi sur le nom tête de la partie gauche de la coordination. Dans les heuristiques exploitant les règles d'accord du modifieur (cf. règles 1 au § 3.1), on a déjà utilisé certains traits morphologiques des composants pour prendre des décisions de ce type, mais on se limitait aux modifieurs qui présentaient des formes différentes selon leurs genre et nombre. Ici, on se fonde simplement sur l'effacement des préposition et déterminant pour proposer une réécriture qui peut donc s'appliquer aussi à des modifieurs qui ne varient pas selon les genre et nombre du nom auquel ils se rattachent. Les expressions présentant conjointement les deux caractéristiques (i.e. formes différentes selon les genres et nombres, et effacement de la préposition et du déterminant devant le nom tête de la partie droite de la cordination) sont donc aussi sélectionnées par les heuristiques 1.

Du fait, pour certains mots, de l'ambiguïté morphologique entre l'adjectif et le nom, cette heuristique produit aussi du bruit :

deux demi-conduits coupés suivant une génératrice et plaqués contre le conduit à protéger

plaqués est analysé comme un nom coordonné à celui de la partie gauche de la coordination, alors qu'en fait ce sont les deux participes passés à valeur adjectivale *coupés* et *plaqués* qui sont coordonnés.
Ce type d'ambiguïté se reproduit souvent : *placé*, *relatif*, *sensible*, *tenu* peuvent être analysés comme des noms ou des adjectifs. On décide donc de modifier en conséquence le dictionnaire de mots prioritaires en forçant l'analyse *adjectif* dans le dictionnaire de mots prioritaires[8].

Pour l'exemple suivant, le problème rencontré a déjà été évoqué pour les heuristiques 1 (en particulier, 1.1 et 1.3) : si le *GN* coordonné est marqué à partir de la première préposition rencontrée à gauche de la coordination, on prend le risque que cette préposition ne serve qu'à introduire un modifieur du nom tête ; dans ce cas, la coordination peut aussi bien relier deux modifieurs de ce nom tête, que deux noms têtes situés de part et d'autre de la coordination :

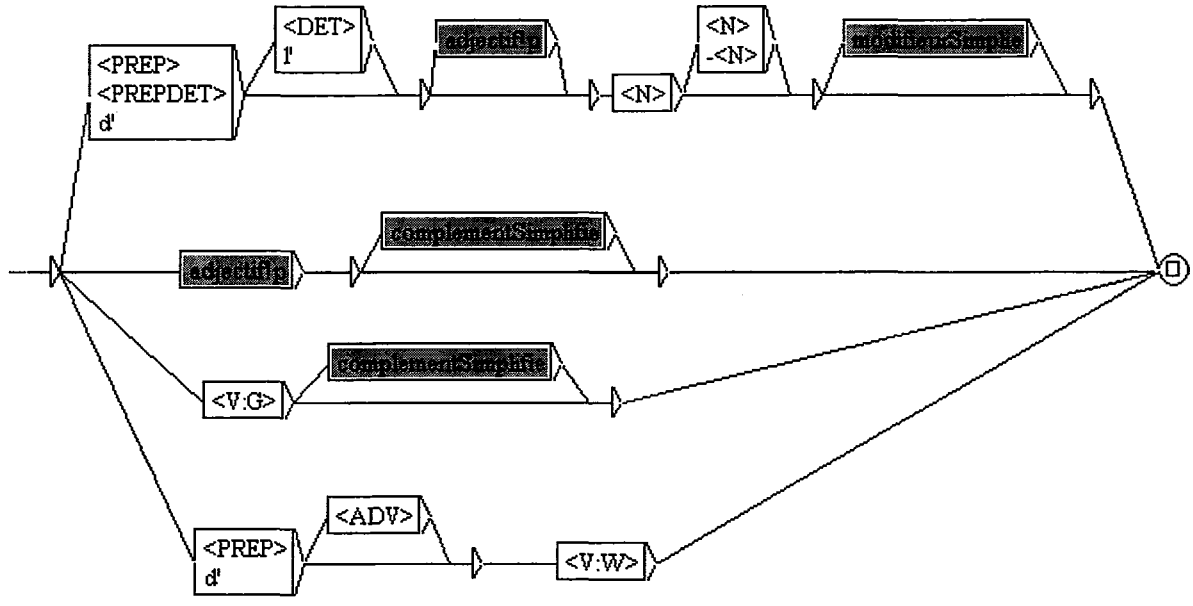
matériaux rigides de toute épaisseur et matériaux souples de épaisseur supérieure à 5 mm

---

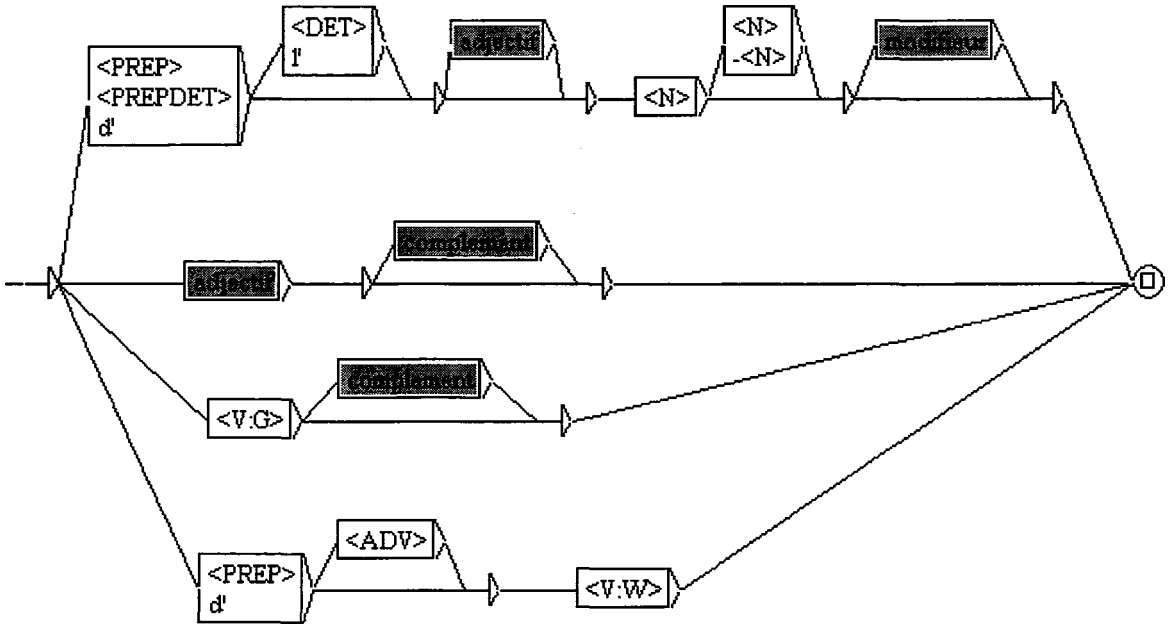
[8] Il en est de même pour les formes au pluriel : *placés*, *relatifs*, *sensibles*, *tenus* pour lesquelles on proposera aussi l'étiquette adjectif, ainsi que pour les formes au féminin et féminin pluriel *relative* et *sensible*, *relatives* et *sensibles*.

On décide donc d'appliquer la même décision que pour les heuristiques 1 : si on identifie un nom ou un ensemble (nom adjectif) immédiatement à gauche de cette préposition, on ne prend aucune décision quant à la réécriture du *GN*.
Mais cette contrainte écarte à tort des *GN* introduits par des prépositions composées comme *en communication avec*, ou *en l'absence de* que l'on a donc rajoutées explicitement dans le transducteur.

D'autre part, on peut utiliser, dans l'application de cette heuristique, les remarques faites quant à l'accord obligatoire entre adjectif et nom. L'objectif est, dans certaines conditions, de distribuer le modifieur situé dans la partie droite de la coordination aux noms têtes de gauche et droite, car si ce modifieur est un adjectif, il doit porter la marque du pluriel. On peut donc ajouter cette contrainte dans le transducteur (cf. lignes 1 et 2 de *modifieur3!p.grf*).

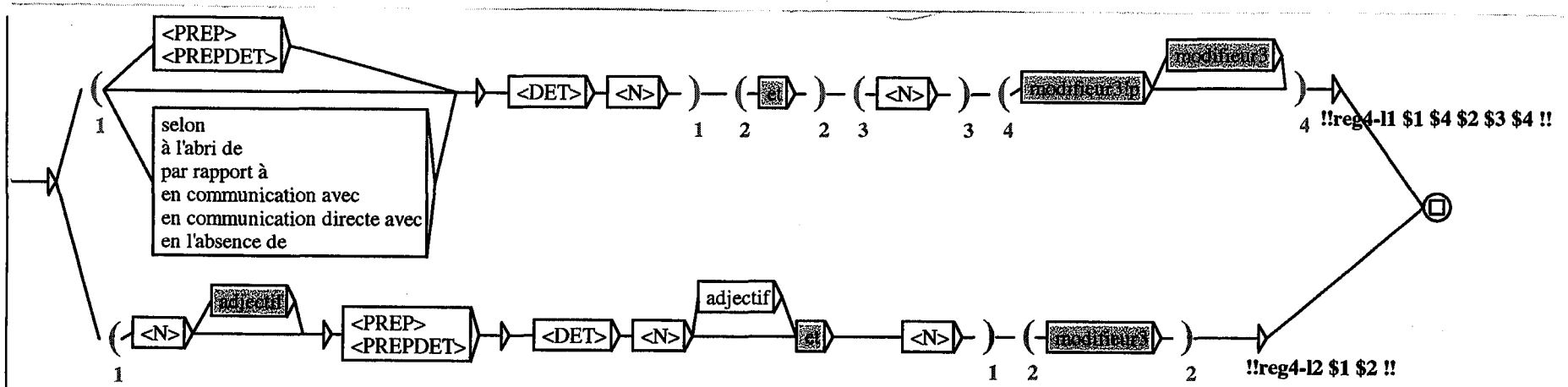


modifieur3!p.grf : automate d'identification d'un modifieur simplifié au pluriel



modifieur3.grf : automate d'identification d'un modifieur

Avec les constructions précédentes, le transducteur finalement construit est le suivant :


<PREP>
<PREPDET>
selon
à l'abri de
par rapport à
en communication avec
en communication directe avec
en l'absence de
<DET>
<N>
et
<N>
modifieur3n
modifieur3
!!reg4-l1 $1 $4 $2 $3 $4 !!
<N>
adjectif
<PREP>
<PREPDET>
<DET>
<N>
adjectif
et
<N>
modifieur3
!!reg4-l2 $1 $2 !!

Il a fallu mettre en place là aussi des modifieurs simplifiés car les transducteurs récursifs étaient difficilement utilisables par INTEX.

Avec ces contraintes, les résultats corrects que l'on peut obtenir sont les suivants :

par rapport aux locaux fréquentés par le public et dégagements fréquentés par le public
bureaux ou aux locaux accessibles à le public et dégagements accessibles au public
conformes aux textes en vigueur et normes en vigueur
doit s'effectuer dans des gaines suffisamment ventilés ou caniveaux suffisamment ventilés
depuis les volumes adjacents mis en surpression ou cantons adjacents mis en surpression

La contrainte ajoutée pour ne pas sélectionner les *GN* dans lesquels on ne peut déterminer formellement si la coordination lie des modifieurs ou des noms têtes, permet d'identifier et de ne pas développer les expressions suivantes :

entretien des détecteurs sensibles aux fumées et gaz de combustion

où *et* pourrait lier *détecteurs* et *gaz*, ou bien *fumées* et *gaz*. C'est ce dernier cas qui en l'occurrence est juste pour cet exemple, et il faudrait développer en *fumées de combustion et gaz de combustion*.
On peut faire la même remarque pour :

façades accessibles et dessertes par des voies ou espaces libres
vérifications nécessaires par des organismes ou personnes agréées
la qualité de ces matériaux et éléments fait l'objet d'essais.
à condition que la puissance utile de chaque appareil ou groupe d'appareils isolé

qui sont identifiés comme trop complexes et non développés.

Néanmoins, demeurent des cas où la réécriture n'est pas correcte :

procès-verbaux et comptes-rendus des vérifications sont transmis au Maire à la disposition de la commission et tenus à la disposition de la commission de sécurité (R 123-44).

est disposé perpendiculairement aux solives à l'aide de clous et fixé à l'aide de clous

6.5. Heuristiques 5

Dans le paragraphe 4.5, on a présenté et étudié des *GN* qui présentent des aspects symétriques. Cette symétrie peut se manifester d'un point de vue morphologique, syntaxique, visuel, auditif ... mais les transducteurs construits pour identifier ces symétries ne peuvent prendre en compte que des aspects morphologiques et syntaxiques.

Comme on l'a vu précédemment, l'un des écueils de l'application des transducteurs est de produire du bruit en développant des *GN* de manière incongrue. On va donc, en priorité, essayer de reconnaître des *GN* dont la structure est complexe - essentiellement ceux qui contiennent juxtaposés, plusieurs modifieurs et noms candidats à devenir des noms têtes - afin de ne pas les développer tant qu'on ne possède pas d'indice supplémentaire sur leur structure. Et on ne cherchera à reformuler que des *GN* dont la construction est, à peu près, identifiée. On verra que, malgré ces précautions, l'application des règles de réécriture au corpus produit encore quelques expressions fautives.

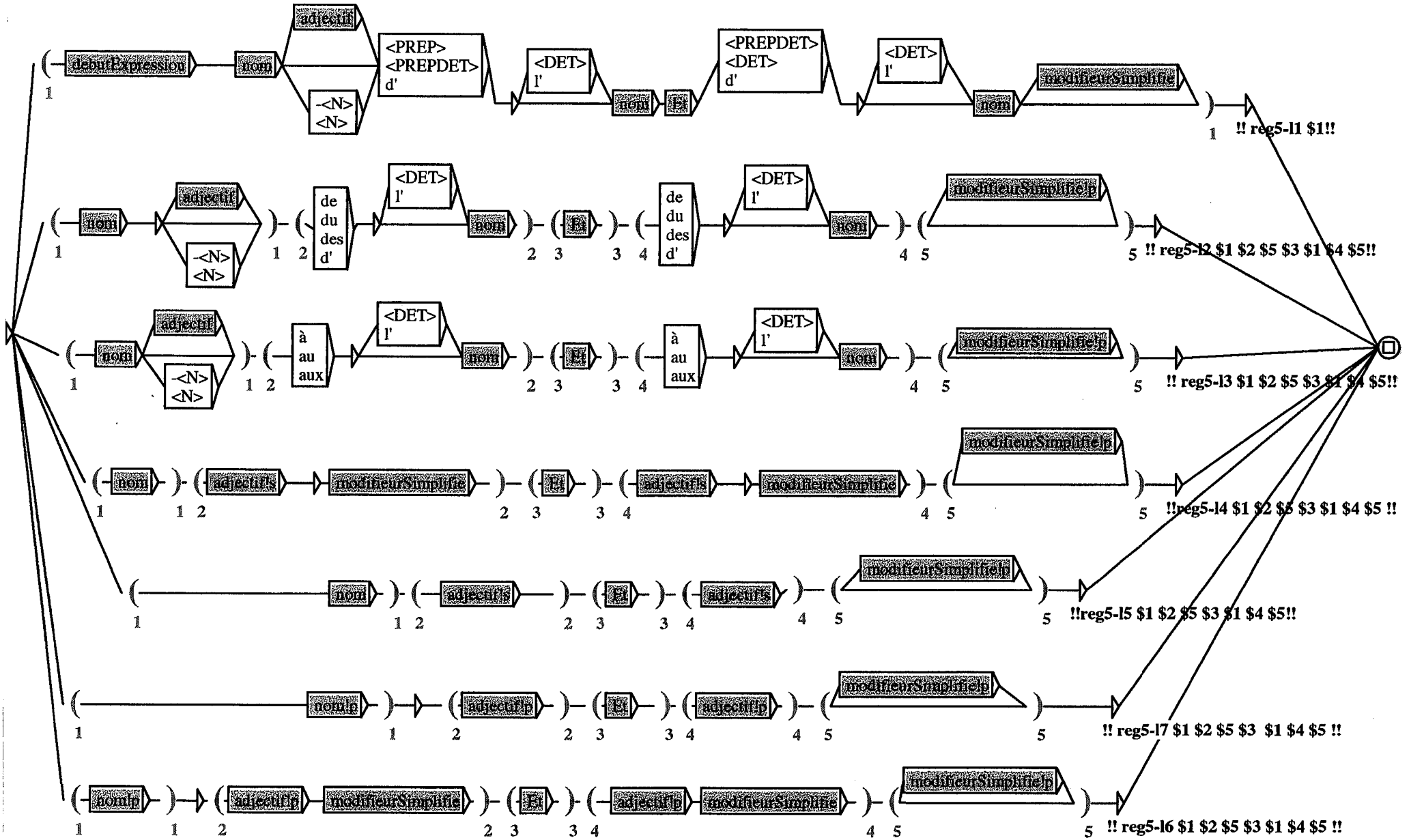
D'autres règles étudiées dans les paragraphes précédents exploitaient aussi implicitement la symétrie des expressions étudiées. C'est en particulier le cas des règles concernant l'accord du modifieur. Cela

explique que l'on sélectionne, dans l'application de ces règles, des *GN* qui ont déjà pu être réécrits auparavant.

Dans le transducteur, on n'a choisi de réécrire que des structures complètement symétriques : *N A et A*, *N de N et de N*, *N à N et à N* bien que l'on ait vu au paragraphe 4.5 que la coordination peut lier, par exemple, deux modifieurs qui se rapportent au même nom mais de construction différente :

les épreuves de résistance mécanique et d'étanchéité
les établissements existants et à modifier

Dans le transducteur proposé, la première ligne marque les *GN* trop complexes pour tenter de les développer et les réécrit à l'identique, tandis que les lignes suivantes identifient des symétries et proposent des réécritures :


debutExpression
nom
adjectif
-<N>
<N>
<PREP>
<PREPDET>
d'
<DET>
l'
Et
<PREPDET>
<DET>
d'
<DET>
l'
modifieurSimplifie
!! reg5-l1 $1!!
de
du
des
d'
modifieurSimplifiep
!! reg5-l2 $1 $2 $5 $3 $1 $4 $5!!
à
au
aux
!! reg5-l3 $1 $2 $5 $3 $1 $4 $5!!
adjectifs
!!reg5-l4 $1 $2 $5 $3 $1 $4 $5 !!
!!reg5-l5 $1 $2 $5 $3 $1 $4 $5!!
nomp
adjectifp
!! reg5-l7 $1 $2 $5 $3 $1 $4 $5 !!
!! reg5-l6 $1 $2 $5 $3 $1 $4 $5 !!

Ce transducteur sélectionne 1 871 expressions qui se répartissent de la manière suivante :

- 560 expressions reconnues comme complexes et pour lesquelles aucune réécriture n'est proposée :

  approbation des règles de sécurité et des modalités de contrôle
  bouches d'amenées d'air et d'extraction des fumées
  [en] fonction du type d'activité et de l'effectif de public reçu
  la notice descriptive des conditions d'entretien et de fonctionnement
  approuvés après avis du comité d'études et de classification des matériaux

- des expressions de structure *N de N et de N* ou *N à N et à N*, éventuellement suivies par un dernier modifieur qui se rapporte à chacune des parties gauche et droite de la coordination. Le nom tête peut aussi régir un adjectif juxtaposé avant le modifieur qui constitue la partie gauche de la coordination. On dénombre et réécrit 675 expressions de type *N de N et de N* parmi lesquelles :

  appareil de cuisson ou appareil de chauffage
  encloisonnement d'un escalier ou encloisonnement d'un ascenseur
  éléments de décoration flottants ou éléments d'habillage flottants
  [à l']abri des intempéries dans un espace clos ou abri du climat dans un espace clos
  locaux d'emballage de déchets et locaux de manipulation de déchets
  mesure complémentaires de prévention proposées et mesure complémentaires de protection proposées

  Pour la structure *N à N et à N* , on identifie et réécrit 72 expressions parmi lesquelles :

  palier à billes ou palier à aiguilles
  lampes électriques à piles ou lampes électriques à accumulateurs
  temps nécessaire à l'alarme des occupants et temps nécessaire à l'évacuation des occupants

- des expressions de structure *N A et A*. Pour pouvoir reconnaître plus d'expressions, on admet des modifieurs des adjectifs à partir du moment où leurs emplois dans le *GN* sont symétriques : *N A modifieurAdjectif et A modifieurAdjectif ;* et on reconnaît aussi et on distribue sur les membres gauche et droit de la coordination, un éventuel modifieur situé en queue du *GN*. Pour essayer de limiter les sélections fautives, on utilise les contraintes imposées par les règles d'accord entre le nom et ses modifieurs et on les rajoute dans les transducteurs.
  On dénombre ainsi 474 expressions *N A et A modifieurAdjectif* :

  vapeurs inflammables ou vapeurs toxiques
  vapeurs toxiques ou vapeurs corrosives
  endroits bien visibles et endroits facilement accessibles
  parois horizontales résistant au feu ou parois verticales résistant au feu
  parties translucides incorporées dans les plafonds et parties transparentes incorporées dans les plafonds

  et 54 expressions N A *modifieurAdjectif* et A *modifieurAdjectif modifieur* :

  circuit électrique distinct et circuit protégé contre les surintensités
  âme combustible simple isolante ou âme multicouche en mousse isolante
  aménagements accessibles à le public et aménagements situés en élévation
  luminaires installés à poste fixe ou luminaires suspendus d'une façade
  dimensions extérieures des huisseries ou dimensions bâtis en bois

Les heuristiques appliquées dans ce paragraphe sont celles qui donnent lieu au plus grand nombre de réécritures, et aussi à un grand nombre d'erreurs.

7. Analyse des causes d'erreurs dans la réécriture des GN

La réécriture des *GN* coordonnés concernent 1960 expressions[9] et crée des libellés inexacts ou incomplets par rapport à la compréhension qu'un lecteur, même non spécialiste du domaine, pourrait avoir du corpus. Le plus souvent, ces erreurs se retrouvent tout au long de ce travail et quelle que soit l'heuristique utilisée. Nous allons donc détailler ces causes d'erreur et essayer de proposer des solutions quand cela paraît possible.

7.1. Erreur sur l'étiquetage syntaxique

INTEX propose pour chaque mot du corpus la totalité des étiquettes qui sont autorisées à partir des dictionnaires utilisés pour l'analyse. Certains mots sont donc largement ambigus. Les transducteurs utilisent ces étiquettes pour sélectionner des expressions et si, pour un mot, une étiquette est possible cette contrainte devient applicable. On a déjà relevé cette ambiguïté dans le paragraphe 6.1 et étudié ces conséquences pour l'heuristique 1.2 en particulier, mais on peut généraliser ces remarques à l'ensemble des heuristiques présentées. Parmi les exemples suivants, le premier est réécrit à partir de l'heuristique 1.1, les deux suivants à partir de 1.2 et le dernier en utilisant la symétrie de la construction (les parties soulignées sont celles reconnues par les transducteurs) :

agrément du ministre de la Santé avant et de la Famille avant le 1er décembre 1978
un endroit accessible en permanence et bien signalé
les parois verticales et une demi heure pour les plafonds
les garanties de sécurité et de bon fonctionnement

Avant admet l'étiquette adjectif pluriel, il est précédé d'un nom au singulier, il se réécrit donc après les parties gauche et droite de la coordination.
Bien est analysé comme un nom singulier suivi de l'adjectif au singulier *signalé,* la séquence *bien signalé* est donc un *GN* dont *bien* constitue la tête. L'interprétation est la même pour *une* et *demi* étiquetés respectivement nom et adjectif singulier.
Pour le dernier exemple, *bon* peut être un nom et dans ce cas la symétrie de la construction ferait que le *GN* serait constitué d'un nom tête *garanties*, suivi de deux compléments de nom *de sécurité* et *de bon, fonctionnement* en tant que nom peut constituer un modifieur simplifié d'après la définition donnée dans le transducteur correspondant, *modifieurSimplifie.grf.*

Plusieurs solutions existent pour limiter ces cas d'ambiguïtés : d'abord écrire des grammaires locales qui permettent de désambiguïser le texte (A. Dister 2000), ensuite forcer l'analyse de certains mots en indiquant explicitement dans un dictionnaire prioritaire les étiquettes admissibles. C'est cette solution que a été choisie dans le paragraphe 6.4. pour éliminer l'étiquette *nom* associée aux mots *relatif, relative, placé, sensible, ...* qui, dans le contexte de la sécurité incendie, ont a priori une étiquette

[9] Ce calcul qui n'est que la somme, pour toutes les heuristiques, du nombre d'expressions identifiées n'est pas tout à fait exact puisque certaines expressions sont sélectionnées par plusieurs transducteurs. L'ensemble des réécritures obtenues, classé par règle, est présenté en annexe.

adjectif. Mais, pour que cette opération soit réellement efficace, il faudrait pratiquer une étude exhaustive du vocabulaire du texte pour lister tous les mots dont le statut est ambigu, ce qui n'a pas été fait ici, et prendre une décision quant aux emplois possibles dans ce contexte. Dans tous les cas, ni l'une ni l'autre de ces méthodes, ni leur utilisation conjointe ne peuvent supprimer automatiquement toute ambiguïté syntaxique dans le texte.

7.2. Complexité des *GN*

La construction des transducteurs essaie d'obéir à un double objectif : ne pas produire de réécritures incorrectes et construire des *GN* les plus complets possible.

On considère qu'une récriture est incorrecte quand l'information contenue dans la formulation finale est moins explicite que celle figurant dans le texte initial. Cette notion est sûrement subjective et surtout difficilement automatisable, mais néanmoins peut se représenter en considérant la pertinence de l'expression produite par rapport à un *GN* qui serait utilisé lors d'une recherche automatique d'information. Par exemple, on trouve dans le texte l'expression :

les conduits et canalisations non incorporés dans une gaine

qui est développée en :

les conduits non incorporés dans une gaine et canalisations non incorporés dans une gaine

et cette reformulation pourrait être utile si l'on recherche les paragraphes concernant *les conduits non incorporés à une gaine*. Malheureusement, le *GN* complet est le suivant :

les conduits et canalisations non incorporés dans une gaine ou non encastrés

et le fait de réécrire la partie gauche du *GN* conduit à éloigner encore plus le nom *conduits* du modifieur *non encastrés* de telle manière qu'il devient encore plus difficile d'identifier dans le *GN* reconstruit l'expression *conduits non encastrés* qui est pourtant sémantiquement aussi pertinente que les expressions obtenues grâce à la réécriture du *GN* initial.
Dans certains cas, ce type de fonctionnement des transducteurs peut tellement dissocier des composants pourtant liés du *GN*, en particulier les modifieurs, que la réécriture de *GN* complétés revient paradoxalement à détruire de l'information. Par exemple avec l'heuristique 1.4, on réécrit les deux *GN* initiaux :

installations de gaz combustible ou d'hydrocarbures liquéfiés
exploitations de types divers ou de types similaires

sous la forme :

installations de gaz combustible ou installations de gaz d'hydrocarbures liquéfiés
exploitations de types divers ou exploitations de types de types similaires

Les deux expression que l'on souhaitait former : *installations d'hydrocarbures liquéfiés* et *exploitations de types similaires* sont absentes après la reformulation[10] et les expressions produites ne contribuent qu'à polluer le texte initial.

Pour minimiser le risque de reformulations fautives, il faut aussi limiter la complexité des expressions qu'on se propose de traiter : si l'on recherche des *GN* de structure *N de N*, on peut ou bien se limiter à ce patron et se contenter de *appareil de chauffage*, ou bien décider de rechercher des formes *N de N modifieur* pour pouvoir trouver *appareil de chauffage mobile* et *appareil de chauffage à gaz*

[10] On peut néanmoins remarquer que ces expressions sont réécrites correctement en utilisant d'autres heuristiques.

où les modifieurs *mobile* et *à gaz* se rapportent à *appareil*, mais aussi *appareil de chauffage central* où *central* renvoie à *chauffage* et non à *appareil*.

D'autre part, pour construire des *GN* les plus complets possible, il faut identifier le plus grand nombre possible de leurs composants afin de les réécrire au mieux, quant à leur structure et leur position, dans les *GN* reformulés. Ce principe induit les conséquences suivantes dans l'écriture des transducteurs :

- on a multiplié le nombre de modifieurs qu'il est possible de relier aux noms têtes et aux noms inclus dans les modifieurs ;
- les structures recherchées pour chaque modifieur sont les plus générales et variées possible. Mais cela conduit à écrire des automates récursifs et dans ce cas, on a été limité par les possibilités de traitement du logiciel ;
- on a essayé d'autoriser l'insertion d'un modifieur du nom partout où cela était possible.

On a peut donc ainsi reconnaître des *GN* de structure complexe et, grâce à cela, les réécrire en tenant compte de l'imbrications des modifieurs et de la portée de la coordination à l'intérieur de ce *GN*.

Néanmoins, si l'on se donne pour objectif d'identifier des structures complexes, on prend aussi le risque de confondre les fonctions des différents composants du *GN* et, par exemple, de ne pas rattacher les modifieurs identifiés au nom tête concerné. Cette difficulté a déjà été évoquée lors de l'étude des règles de développement de la coordination, liées à l'accord en genre et nombre entre le nom et ses modifieurs. En effet, si l'on veut coordonner deux *GN* de construction *N de N* : *N de N1* et *N de N2*, sous la forme *N de N1 et N2* et quand on considère que le nom tête *N* peut accepter d'autres modifieurs par exemple un adjectif *A* et un complément de nom *de N3*, on en vient à identifier des *GN* de structure *N de N3* A *de N1 et N2* dans lesquels il n'est pas possible de déterminer formellement la portée de la coordination et les relations entre les différents composants. C'est par exemple le cas pour les deux exemples suivants de même structure apparente mais pour lesquels les dépendances entre composants diffèrent :

appareil de production d'eau chaude et de chauffage
appareil de chauffage d'appoint mobile et à gaz

Puisqu'il n'y a pas de critère formel de réécriture, il faudrait les reconstruire de la même manière, et cela conduirait, au moins dans un cas, à produire des *GN* ineptes. On a donc pris la décision, pour toutes les heuristiques, de recopier le *GN* à l'identique quand on identifiait plusieurs modifieurs juxtaposés, et cela pour la majorité des heuristiques.

7.3. Erreurs de segmentation

L'identification, entre autres, des *GN* repose sur une segmentation en phrases correcte. Ce découpage est effectué par un transducteur appliqué au texte initial par INTEX. Ce transducteur est modifiable pour prendre en compte d'éventuelles caractéristiques formelles du texte mais ne peut pas remédier aux fautes d'écriture de ce texte. Par exemple, le corpus concernant la sécurité incendie contient de nombreux titres formulés ainsi :

Chapitre IV cas des conduits à risques importants ou spéciaux annexe VI (annexe particulière aux clapets)

aménagements intérieurs, décorations et mobilier arrêté du 25 juin 1980 portant approbation des dispositions ...
code de la construction et de l'habitation protection contre les risques d'incendie

Il n'y a, presque systématiquement, pas de ponctuation forte à la fin des titres de chapitres ou d'articles. Les phrases ne sont donc pas arrêtées convenablement et les transducteurs poursuivent leur exploration du texte en passant à la phrase ou au titre suivants. Ceci donne des développements fautifs de la coordination :

... cas des conduits à risques importants annexe ou risques spéciaux annexe VI ...
... décorations arrêté et mobilier arrêté du 25 juin 1980 portant approbation ...
code de la construction protection et de l'habitation protection contre les risques d'incendie

Le vocabulaire employé dans la phrase peut donner des indices quant à sa segmentation, et notamment autour des conjonctions de coordination. En particulier, des déterminants composés comme *un ou plusieurs*, ou des adjectifs composés comme *inférieur ou égal à*, *supérieur ou égal à*, et leurs flexions au féminin ou au pluriel, ne doivent pas provoquer la même distribution des parties gauche et droite de la coordination que des *GN* effectivement libres qui contiennent une coordination. Par exemple :

par une ou plusieurs gaines prélevant l'air directement
épaisseurs inférieures ou égales à 0,5 mm

ne doivent pas être reconstruits selon les heuristiques précédemment présentées qui conduiraient à écrire :

par une ou plusieurs gaines prélevant l'air directement
épaisseurs inférieures ou épaisseurs égales à 0,5 mm

ce qui, sans être préjudiciable, ne présente aucun intérêt.

L'adverbe *non*, lui, doit être traité avec plus de précautions : la réécriture, selon les heuristiques précédentes, des *GN* qui contiennent une ellipse liée à l'emploi de *non* comme quantifieur d'un adjectif contribuent à désorganiser l'expression et à perdre de l'information. Le *GN* :

des portes coulissantes ou non destinées à obturer ces baies

est réécrit, en utilisant les heuristiques concernant la symétrie de la construction des modifieurs dans une expression contenant une coordination, de la manière suivante :

des portes coulissantes à obturer ou portes non destinées à obturer ces baies

dans laquelle il n'y a pas un *GN* plus complet grâce à la réécriture.

7.4. Erreurs induites par la réécriture des GN

La réécriture sous forme de *GN* complétés des expressions initiales, provoque des erreurs systématiques dans la réécriture des *GN* sélectionnés par les transducteurs.

7.4.1. Fautes d'accord entre nom tête et modifieurs

On a vu que les heuristiques développées au § 1 concernent l'accord en genre et en nombre entre le nom tête du *GN* et ses modifieurs identifiés, et que l'on a fait coïncider ce type de modifieur avec les adjectifs qualificatifs et les participes passés à valeur adjectivale. Mais, même quand ces heuristiques conduisent à une reformulation correcte, il y a une faute d'orthographe systématique sur les modifieurs. En effet, la forme répétée derrière chacune des têtes de la coordination est la forme du

modifieur telle qu'elle est repérée dans le *GN* initial et elle est alors nécessairement au pluriel pour satisfaire les contraintes des différentes heuristiques. Or chacune des têtes derrière lesquelles le modifieur est répété peut figurer au singulier. Ou bien l'adjectif est au masculin parce que l'une des têtes est au masculin, mais lorsqu'il est réécrit il ne présente pas une forme satisfaisante parce qu'il associé à l'autre nom tête coordonné qui est au féminin. C'est ce type d'erreurs que l'on retrouve dans les exemples suivants :

Certificats ou factures relatifs à la teneur en azote
chute de fragments ou de gouttes enflammés
quantités d'acides chlorhydrique et cyanhydrique
les couleurs rouge et orange étant interdites

qui sont réécrits sous les formes (les ensembles (nom plus modifieur) dont l'accord est fautif sont **en gras**) :

Certificats relatifs à la teneur ou ***factures relatifs*** *à la teneur en azote*
chute de fragments enflammés ou de ***gouttes enflammés***
quantités ***d'acides chlorhydrique*** *et quantités* ***d'acides cyanhydrique***
les couleurs rouge *et* ***les couleurs orange*** *étant interdites*

Pour les deux derniers exemples, c'est le nom qui devrait passer au singulier pour donner *acide chlorhydrique*, *acide chlorhydrique*, *la couleur rouge* et *la couleur orange*. Cette erreur se produit pour les modifieurs dont la forme varie effectivement selon les genre et nombre. Mais, lorsque les formes du texte sont ambiguës : par exemple pour les adjectifs qui ont la même forme au masculin et au féminin, ou au singulier et au pluriel, les contraintes de l'heuristique sont considérées comme étant satisfaites, mais le plus souvent à tort, ce qui donne lieu à des développements non pertinents voire fautifs :

combustibles métalliques dans le local ou des poudres métalliques dans le local
les concessionnaires permanents des locaux et les locataires permanents des locaux

alors que les formes initiales étaient :

combustibles ou des poudres métalliques dans le local
les concessionnaires et les locataires permanents des locaux

Dans le premier exemple, *métalliques* est analysé comme un adjectif mais ses genre et nombre sont ambigus puisque toutes les flexions sont identiques. La contrainte : adjectif au masculin pluriel précédé d'un nom féminin, est donc satisfaite et l'heuristique est appliquée. Pour l'expression suivante, c'est le nom *locataires* qui est ambigu mais puisqu'il admet le genre féminin, on considère que la condition est là aussi remplie.

Tout au long de ce chapitre, on a rétabli manuellement l'orthographe correcte dans les *GN* reformulés, sauf évidemment lorsque ces fautes d'accord sont, comme ici, le sujet de l'exposé.

7.4.2. Effacement de la préposition devant le modifieur du nom

Lorsque la préposition et le déterminant, ou seulement le déterminant, sont effacés dans la partie droite du *GN* initial contenant une coordination, les transducteurs mettent à profit cet effacement pour en tirer certaines conclusions mais ne peuvent pas toujours rétablir la préposition et le déterminant. Les *GN* réécrits paraissent alors étranges et la réécriture non pertinente (la préposition est ajoutée en **[gras]**, entre le nom et le modifieur) :

au moyen de suspentes de nature identiques aux cas réels et suspentes ***[d']****espacement identiques aux cas réels*

Néanmoins, si on fait une recherche automatique d'information, les "mots grammaticaux" tels que la préposition *de* ne sont pas pris en compte, et donc, même si elle n'est pas complète, la réécriture peut néanmoins être utile.

7.4.3. Apparition de majuscules à l'intérieur des *GN* complétés

L'heuristique prenant en compte la présence d'un déterminant possessif en tête de la partie droite du *GN* coordonné (§ 4.2) procède à une reformulation en répétant le nom tête de la partie gauche précédé de son déterminant. Si ceux-ci figurent en début de phrase, le déterminant porte une initiale majuscule qui est ensuite répétée telle qu'elle à l'intérieur du *GN* reconstruit. Comme dans le cas de l'effacement de la préposition, le *GN* obtenu paraît étrange mais l'efficacité de la reformulation n'en est pas affectée si les logiciels ne différencient pas majuscules et minuscules.

Ces colonnes et les dispositifs de alimentation de Ces colonnes ...
Ces constructions et les escaliers de accès de Ces constructions ...
Ces appareils et la canalisation de Ces appareils ...

8. Heuristiques confirmées

Dans cette partie, on exposera l'autre volet du travail sur l'étude de la coordination à l'intérieur des *GN*. Après avoir décrit des règles qui se fondent sur la morphologie, le lexique et la syntaxe des *GN*, on va les utiliser, comme dans le paragraphe précédant, pour construire des *GN* complétés mais, cette fois, on se propose de valider ces constructions en exploitant les dictionnaires : les *GN* transformés ne seront acceptés que si les expressions qui figurent après réécriture de part et d'autre de la coordination sont aussi présentes dans les dictionnaires de noms composés. Ces transformations ne constituent en fait qu'une nouvelle présentation des expressions figurant dans les dictionnaires et ne permettent pas de repérer de nouveaux termes, mais elles ont pour but de limiter le bruit dû au développement des coordinations à l'intérieur des *GN*.

La validation de ces réécritures repose donc sur les dictionnaires de mots composés qui décrivent le vocabulaire de la langue générale, mais surtout les concepts, objets, outils, méthodes de travail propres au domaine. Avec ce travail, on ne peut espérer valider toutes les réécritures produites, d'une part parce que, comme on l'a développé dans le paragraphe 7, certaines sont fautives, d'autre part parce qu'un grand nombre des reformulations, même si elles sont correctes, concernent des expressions libres et qui donc n'apparaîtront pas dans des dictionnaires de mots composés. Par exemple, les *GN* suivants ont été obtenus après le repérage de la coordination et la réécriture de l'expression initiale :

les conduits non incorporés dans une gaine et canalisations non incorporées dans une gaine
les aménagements réalisés dans les établissements ou modifications réalisées dans les établissements

or ces expressions n'appartiennent pas au dictionnaire de noms composés et rien ne justifie leur ajout. Donc, bien que la développement du *GN* initial en *GN* complété soit correct, il n'est pas retenu et on conserve, dans le corpus, la phrase initiale.

D'un point de vue opérationnel, ce sont les mêmes transducteurs que ceux décrivant les heuristiques étudiées dans le paragraphe précédant qui sont appliqués au corpus, à la seule différence que les expressions qui satisfont les conditions de déclenchement de ces transducteurs sont marquées en

même temps qu'elles sont réécrites. Le corpus initial dans lequel certaines séquences sont remplacées par d'autres constitue un nouveau corpus qui est soumis de nouveau à INTEX. Il faut ensuite appliquer à ce nouveau corpus les automates qui effectuent deux tâches : valider les expressions candidates par application de dictionnaires (eux aussi écrits sous forme d'automates) et "nettoyer" le corpus en ôtant les marques posées pour les traitements précédents.
Ces opérations n'ont été testées que sur un ensemble d'expressions réduit, issu du corpus initial. En effet, plusieurs problèmes soulevés dans le paragraphe précédent n'ont pas été résolus et empêchent une vérification "en vraie grandeur" de la nouvelle méthode proposée.

On a déjà évoqué au sujet des heuristiques 1.1, 1.3 et 1.4 la faute d'accord systématique entre noms têtes au singulier et modifieur au pluriel, ou bien entre nom tête au pluriel et modifieurs au singulier. Pour que l'accord correct entre nom et modifieurs soit rétabli[11], il faudrait disposer d'un dictionnaire inverse qui à partir de la forme fléchie du modifieur permettrait d'obtenir, en donnant les caractéristiques en genre et nombre du nouveau nom concerné par le modifieur, la forme convenable. Mais cet outil n'est pas disponible[12] actuellement dans le logiciel INTEX. Il est donc impossible, dans le corpus composé automatiquement avec les expressions coordonnées réécrites, de faire figurer des expressions orthographiquement correctes : les séquences sont formulées à partir des mots fléchis figurant déjà dans le texte initial :

par une personne agréés ou un organisme agréés
de fragments enflammés ou de gouttes enflammés
les planchers haut et les planchers bas du local

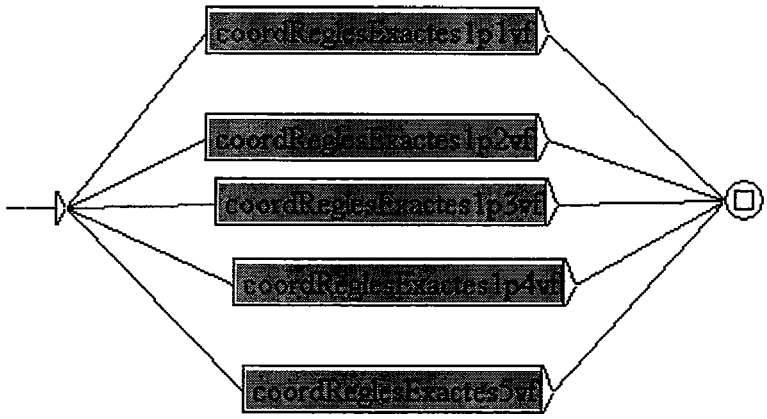
Et donc, même si les expressions *personne agréée*, *plancher haut* ou *plancher bas* constituent des entrées du dictionnaire, les séquences fautives en termes d'accord : *personne agréés*, *planchers haut* et *planchers bas* ne pourront être reconnues à l'aide de ces dictionnaires.
On peut contourner cette difficulté dans la phase de test en fabriquant, à l'aide d'automates, des dictionnaires dans lesquels chacune des flexions du nom tête peut se combiner avec chacune des flexions du modifieur. Mais cette méthode n'est guère satisfaisante dans une phase réellement opérationnelle. Cependant, on peut tempérer l'impact de cette difficulté en remarquant que le problème ne se pose que pour les composés de structure *NA*, *NAA* et *NAPN*.

Les transducteurs utilisés sont les suivants :

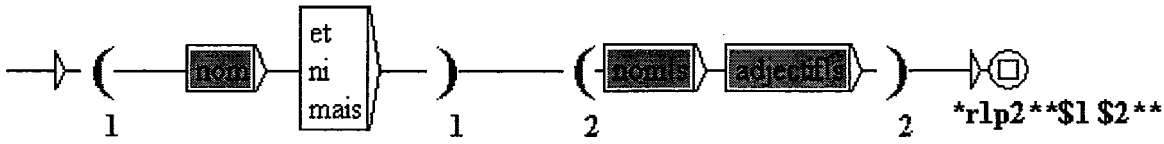
- ceux concernant la mise en œuvre des heuristiques sont identiques à ceux étudiés précédemment, on a juste rajouté des marques pour délimiter l'expression marquée et/ou réécrite ; on a aussi conservé les références de l'heuristique utilisée pour développer la coordination. On ne présente à la suite que l'automate qui récapitule l'application des différentes heuristiques confirmées, et non les transducteurs de ces heuristiques eux-mêmes, puisque ceux-ci sont identiques aux transducteurs initiaux, et, à titre d'exemple, la règle exacte correspondant à la règle 1.2 :

[11] Dans tout l'exposé précédent, on a artificiellement gommé cette difficulté en rétablissant manuellement les accords corrects entre noms têtes et modifieurs, afin de ne pas rendre plus difficile la lecture.

[12] Ceci pour des questions de protection des dictionnaires, et non pour des difficultés à concevoir l'outil adéquat.



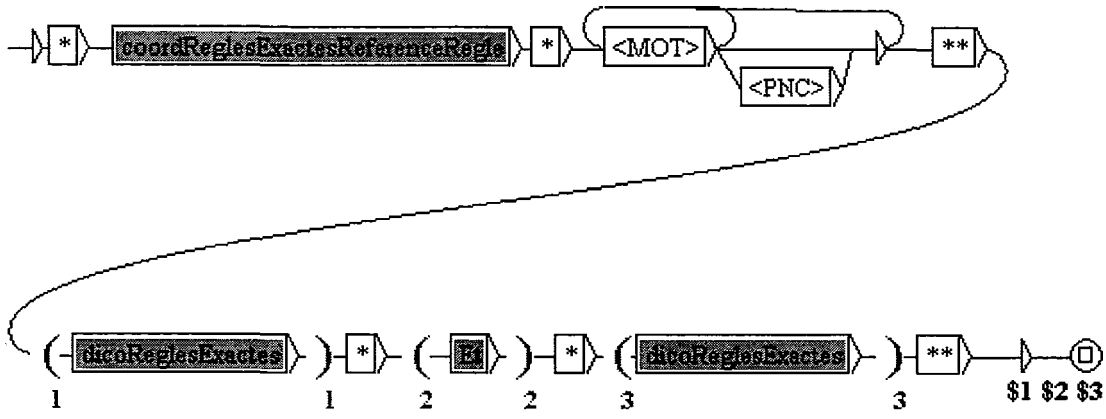
coordReglesExactesRecap.grf : transducteur récapitulant l'appel des règles exactes utilisées



coordReglesExactes1p2vf.grf : transducteur traduisant la règle 1.2 avec le marquage de la réécriture

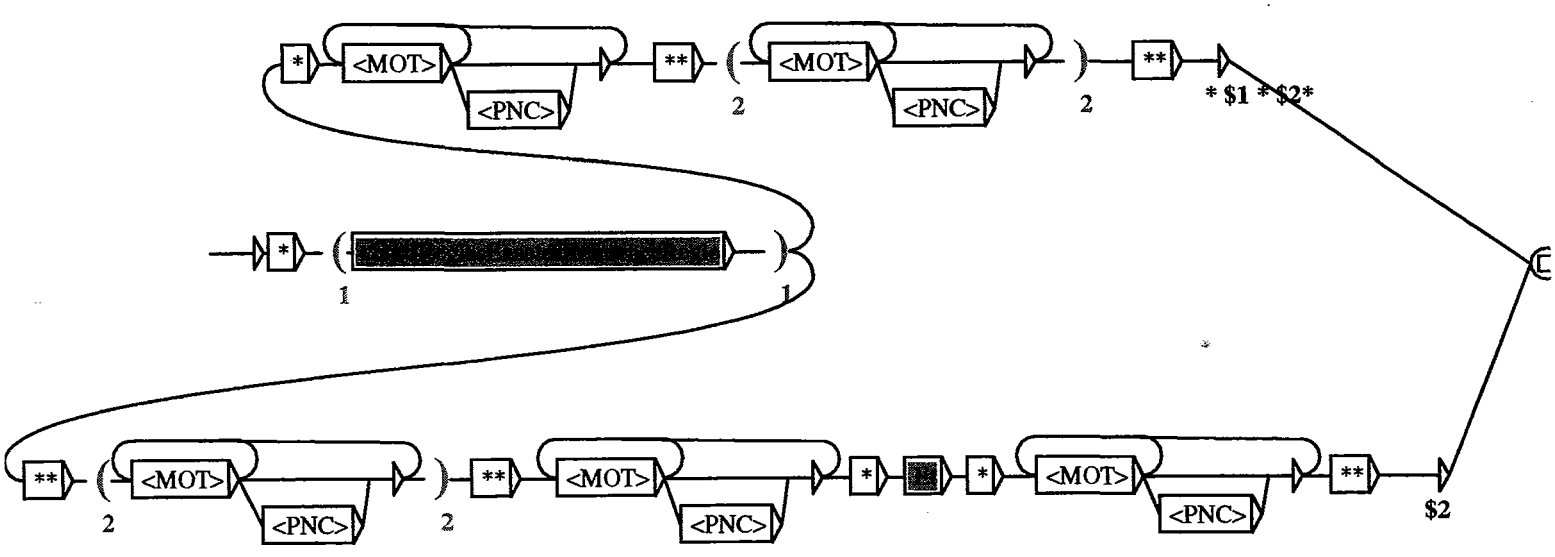
- pour vérifier si les expressions produites figurent dans les dictionnaires (pour pouvoir retourner à la formulation initiale, il faut conserver à côté de l'expression réécrite, l'expression d'origine) :

coordReglesExactesNettoyage1

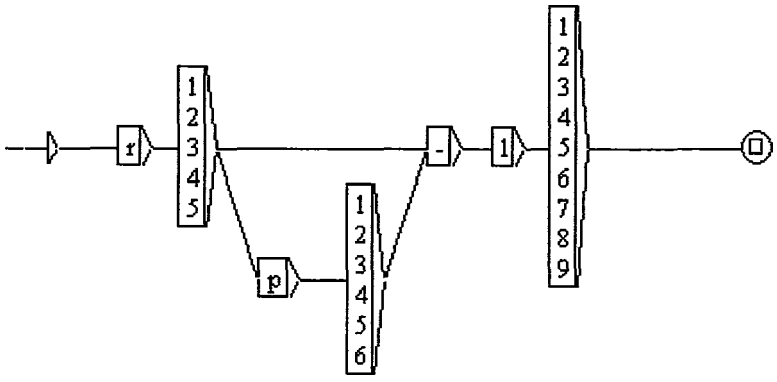


coordReglesExactesNettoyage1.grf

- pour conserver l'expression satisfaisante (pour que la formulation réécrite soit considérée comme satisfaisante, il faut que chacune des expressions produites figure dans les dictionnaires. On pourrait modifier cette règle, et accepter la reformulation dès que la réécriture de l'une des parties de la coordination appartient à un dictionnaire) :



coordReglesExactesNettoyage2.grf : transducteur montrant le choix de l'expression réécrite



coordReglesExactesReferenceRegle.grf : automate d'identification des références d'une règle

9. Ordre d'application des règles de réécriture

9.1. Règles de réécriture prioritaires

Comme on l'a vu dans le paragraphe 4, l'identification et la réécriture des *GN* composés d'un nom tête et d'une liste de modifieurs doivent être traités en priorité et toutes les heuristiques des paragraphes 6 et 7 devraient s'appliquer à un texte dans lequel ces *GN* auront déjà été réécrits.

Les éléments lexicaux qui induisent une construction particulière doivent aussi être repérés avant d'essayer d'appliquer les heuristiques sinon celles-ci peuvent produire des reformulations fautives, ou pur le moins non pertinentes. Par exemple dans le cas du nom *l'ensemble de*, si cette tournure n'est pas repérée, les heuristiques fondées sur la symétrie de la construction proposeront la réécriture suivante :

l'ensemble de la trappe et ensemble de ce mécanisme constitue un dispositif ...
l'ensemble des réserves et ensemble des locaux d'emballage installés

De même, si la préposition composée *en fonction de* est accompagnée de plus d'un *GN,* ceux-ci, juxtaposés ou coordonnées, sont réécrits grâce aux heuristiques fondées sur la symétrie de la construction pour produire des expressions confuses :

en fonction de la nature des aléas ou fonction de l'importance des aléas
en fonction de sa hauteur du public reçu et fonction de l'effectif du public reçu

alors que la formulation initiale était la suivante :

en fonction de la nature ou de l'importance des aléas
en fonction de sa hauteur et de l'effectif du public reçu

9.2. Mise en parallèle des règles de réécriture

Les heuristiques 1 tiennent compte des règles d'accord entre nom et modifieurs, elles n'ont donc aucun effet sur des expressions où les modifieurs sont introduits par des prépositions, là où les heuristiques

fondées sur l'identification d'une symétrie dans la construction du *GN* (heuristiques 5) peuvent avoir de l'intérêt. On peut donc considérer que ces règles sont applicables en parallèle sur le corpus.

Dans d'autres cas, les conditions d'application d'heuristiques distinctes peuvent être satisfaites par la même expression. Par exemple le *GN* :

revêtements de mur tendus ou collés

remplit les conditions des heuristiques 1.1 (accord entre modifieur et nom) et 1.5 (symétrie de la construction). Les expressions produites sont dans ce cas identiques :

revêtements de mur tendus ou revêtements de mur collés

En revanche, avec l'expression suivante :

combles et dessous compris

qui est concernée par les deux heuristiques 1.1 et 1.2 à cause de l'ambiguïté en nombre de *dessous* et *compris* (dont la même forme peut représenter respectivement un nom masculin singulier ou pluriel, et un adjectif masculin singulier ou pluriel), les *GN* réécrits en appliquant les deux transducteurs correspondants aux deux règles, sont différents :

combles et (dessous compris)
(combles compris) et (dessous compris)

La première réécriture est fautive même si elle ne crée pas de silence supplémentaire par rapport à la formulation initiale dans laquelle la portée de *compris* est ambiguë. Seule la deuxième expression est convenable.
Les ambiguïtés sur les genre et nombre présentés par une forme étant nombreuses, il est préférable d'appliquer les règles 1 (concernant l'accord du modifieur) à la suite et non en parallèle. Dans ces cas ambigus, plusieurs réécritures pourront être produites et l'utilisation de dictionnaires de mots composés (on se retrouve ainsi dans le cas des "heuristiques confirmées" développé au paragraphe 8) permettra de trancher pour l'une ou l'autre d'entre elles.

Un travail complémentaire (non présenté ici) est à faire pour proposer des optimisations dans l'application des règles de réécriture.

10. Conclusions

Ce travail n'est pas réellement terminé. Deux points particuliers pourraient donner matière à des développements utiles : d'une part, le repérage et le traitement des éléments lexicaux comme *autre*, *entre*, *en fonction de*, ... en préalable à tout traitement de réécriture, d'autre part, l'étude des conflits possibles entre plusieurs heuristiques qui pourraient être applicables à un *GN* donné. Cette étude complémentaire permettrait peut-être de proposer un ordre plus adéquat pour l'application des heuristiques, ou bien d'établir des critères formels de choix entre les différentes reformulations produites pour un même *GN*. Nous pouvons néanmoins dégager plusieurs idées de notre travail :

Il faut se fixer une limite a priori pour la complexité des groupes nominaux que l'on souhaite traiter. Il est illusoire d'espérer réécrire automatiquement tous les groupes nominaux contenant une coordination, quelle que soit la récursivité de leur construction. De plus, il est difficile de définir des schémas de réécriture généraux parce que de nombreuses constructions sont complètement dépendantes du vocabulaire employé.

Une autre limitation a été apportée au problème que l'on se proposait de traiter : la coordination des noms têtes n'est pas prise en compte dans les heuristiques qui utilisent la symétrie de la construction pour proposer une reformulation du *GN*, tout comme le traitement des listes de plusieurs noms têtes auxquels se rapporte un modifieur parce que ces heuristiques introduisent de nombreuses réécritures non pertinentes voire fautives. M. Salkoff (1979) a déjà insisté sur l'*ambiguïté structurelle introduite par une conjonction. En effet, il s'avère le plus souvent impossible de préciser formellement, i.e. au moyen d'un algorithme et univoquement, la structure à laquelle la conjonction doit être rattachée. ... elle peut souvent être rattachée à plusieurs structures dans l'arbre, bien que le témoin humain qui est familier avec le domaine technique trouve qu'il n'y a qu'une seule liaison conjonctionnelle qui soit sémantiquement convenable.* ... En particulier, la séquence *Dét N et Dét N* peut être analysée de multiples façons :

Pierre mange la poire et la pomme
Pierre mange la poire et la pomme reste dans son assiette
Pierre mange un morceau de la poire et la pomme toute entière
Pierre mange des morceaux de la poire et la pomme

sans que l'on puisse montrer de critère formel discriminant. On a donc renoncé à cette partie de l'étude.

L'identification du type de l'expression coordonnée (par rapport à la typologie proposée) est très dépendante de la couverture en mots composés des dictionnaires utilisés. Par exemple, si l'on donne priorité à la reconnaissance des noms composés, un modifieur pourra ne plus être reconnu en tant que tel et isolé, et donc devenir non applicable à une autre tête que celle qui le précède immédiatement à l'intérieur du *GN* coordonné. En effet, si l'on considère l'exemple suivant :

règles techniques et de sécurité applicables au stockage

identifier le nom composé *règles techniques* rend difficile le rattachement des deux modifieurs suivants qui ont, de plus, des portées différentes : *de sécurité* qualifie le même nom tête *règles* pour composer un autre nom composé *règles de sécurité* coordonné au précédent, et *applicables au stockage* se rapporte à chacun des deux composés précédents.
En revanche, marquer le nom composé *gaz chauds* dans l'expression *fumées et gaz chauds* permettrait d'éviter l'application des heuristiques fondées sur la symétrie de la construction, qui produisent le *GN* absurde suivant :

fumées chaudes et gaz chauds

Mais il n'est pas possible d'appliquer en même temps, et d'une manière sélective et automatique, ces deux principes. On a donc choisi de ne pratiquer une analyse qu'en mots simples et de projeter ensuite sur le corpus ainsi étiqueté les transducteurs de réécriture. On prend ainsi le risque de distribuer à tous les noms d'un *GN* des modifieurs qui sont en fait spécifiques d'un seul d'entre eux. La manière de pallier cette faiblesse est d'appliquer la méthode développée au paragraphe 8 concernant les heuristiques confirmées.

Par ailleurs, le fait d'ajouter des contraintes fait que les transducteurs ont un champ d'application moins large. Parfois, les règles permettent de marquer des expressions comme ne devant pas être développées, mais pas toujours pour de bonnes raisons. On peut néanmoins en conclure que les contraintes ajoutées permettent de sélectionner seulement des *GN* relativement simples. Ceux dont la

construction est complexe et récursive[13] se trouvent écartés et cela diminue d'autant le risque de réécritures difficiles à établir précisément du fait de cette complexité, et donc le plus souvent fautives.

Enfin, à condition de résoudre les problèmes d'orthographe rencontrés lors de la réécriture automatique des *GN*, cette étude contribue à :

- diminuer le silence lors d'une recherche automatique d'information à l'intérieur d'un corpus en reformulant automatiquement des *GN* là où la coordination permettait d'écrire des *GN* elliptiques ;

- proposer de nouveaux termes candidats à devenir des mots composés, termes issus de *GN* contenant des coordinations et sur lesquels les patrons utilisés dans le chapitre *Etude du corpus* ne pouvaient opérer ;

- améliorer la segmentation de la phrase en décrivant partiellement la syntaxe des modifieurs droits à l'intérieur d'un *GN*.

[13] Comme on l'a détaillé tout au long de ce chapitre, les raisons de cette complexité peuvent être multiples : plusieurs modifieurs sont imbriqués ou juxtaposés, le *GN* contient des coordinations concernant à la fois les noms têtes et les modifieurs.

Grammaires de reformulation

Une des possibilités offertes à des utilisateurs par la recherche documentaire, est d'interroger un corpus par l'intermédiaire d'une question exprimée en langue naturelle. La "réponse" se présente sous la forme d'un ou plusieurs textes évalués, par le système, comme pertinents par rapport à la question posée. Dans ce contexte, une des causes du silence – c'est à dire la non-présentation d'un texte pertinent – est la différence de formulation (due aux constructions syntaxiques ou de vocabulaire), pour une même idée ou un concept identique, entre la question et le corpus.

Cet article présente, pour le corpus particulier de la réglementation incendie des établissements recevant du public, des grammaires qui, appliquées à des questions posées par l'utilisateur, identifient un énoncé initial et en construisent d'autres formulations en employant des transformations linguistiques : variations du verbe support, restructuration du groupe nominal, variantes lexicales ...

Les grammaires de reformulation proposées concernent une notion technique caractéristique de la sécurité incendie ; l'exemple choisi est celui de l'expression de la stabilité au feu, mais les procédés linguistiques sont généraux et applicables à d'autres notions techniques du même domaine ou non. Les phrases suivantes illustrent la variété des emplois de cette notion :

les structures principales doivent avoir un degré de SF d'au moins deux heures
les conduits doivent être réalisés en matériaux incombustibles et être SF de degré 1/4 h
complexe classé stable au feu une heure
une structure stable au feu de degré 1/2 heure
une stabilité au feu de degré une heure est exigée
aucune exigence de stabilité au feu n'est imposée aux structures des bâtiments
le temps requis pour la stabilité au feu du bâtiment
leur durée de stabilité au feu est estimée conformément au DTU précité

On remarque que l'expression de la stabilité au feu s'exprime à l'aide du nom *stabilité au feu*, ou de l'adjectif *stable au feu* ; les constructions utilisent les auxiliaires *être* ou *avoir* ; *stabilité au feu* peut avoir le rôle du nom tête de groupe nominal, ou bien complément du nom tête.

Afin d'étudier les différentes formulations des expressions permettant l'expression de la stabilité au feu, on définira une classe d'équivalence qui permettra de reconnaître et de regrouper les phrases ou les *GN* concernant cette notion, puis on recherchera de manière quasi systématique les différentes formulations de cet invariant de sens, on construira ensuite les grammaires correspondant à ces expressions. Dans le paragraphe suivant, on essaiera d'utiliser ces grammaires de reformulation pour reconnaître des expressions concernant d'autres notions employées en sécurité incendie, comme le coupe-feu et le pare-flammes, on étudiera ensuite les formulations qui combinent dans la même phrase des expressions liées à deux notions (comparaison des stabilités au feu, coupe-feu ou pare-flammes associés à deux éléments différents du bâtiment ou bien comparaison de deux propriétés, parmi les trois énoncées, concernant ou non le même élément du bâtiment).

1. Définition de la classe d'équivalence

1.1. Domaine d'application

Le concept de stabilité au feu s'applique à des éléments de construction : conduit, paroi, porte, élément de structure, élément de plancher, dalle, plancher, ou à des assemblages d'éléments de construction : structure, bâtiment, établissement.
On pourrait ajouter un trait sémantique "élément de construction" dans les dictionnaires pour distinguer les entrées auxquelles on peut appliquer plusieurs notions caractéristiques de la construction : stabilité au feu, pare-flammes, coupe-feu, ...

1.2. Nom de la classe

La même idée de caractérisation de la stabilité au feu d'un élément de bâtiment peut s'exprimer de différentes manières avec des étiquettes syntaxiques différentes : nom, adjectif, constructions en *être* ou en *avoir*, ... On considère que toutes les formulations liées à l'expression de la stabilité au feu appartiennent à la même classe d'équivalence qui représente un invariant de sens et à laquelle on donne le nom de *SF* (*stabilité au feu*). On passe d'une formulation à une autre grâce à l'application de transducteurs.

1.3. Elément représentatif - choix de la forme canonique

(1) *La porte a une stabilité au feu de degré une heure*

La forme la plus simple et la plus complète de l'expression de la stabilité au feu est la suivante :

(2) N_0 *<avoir> une stabilité au feu de degré <quantification de la stabilité au feu>*

On décidera qu'elle constitue la forme canonique des éléments de la classe.

Le cas échéant, la forme canonique peut être tronquée si l'expression initiale ne comporte pas d'indication de durée. Mais dans ce cas, il faut qualifier d'une autre manière la stabilité au feu énoncée :

(3) N_0 *<avoir>(*une + *la + une certaine) stabilité au feu*

(4) a. *La porte a une stabilité au feu comparable à celle de ...*

(4) b. *Les planchers en bois sont considérés comme stables au feu tant que la température du plénum ne dépasse pas 300°C*

Il existe aussi une forme adjectivale parallèle :

La porte est stable au feu de degré 1h

Le nom composé *stabilité au feu* et l'adjectif *stable au feu* admettent une variante dans la forme *SF*, qui peut se substituer à eux, dans tous les cas.
Cette forme *SF* est donc ambiguë : elle peut aussi bien être employée pour le nom *stabilité au feu* que pour l'adjectif *stable au feu*.

(1) *La porte a une stabilité au feu de degré une heure*
= (30) *La porte est stable au feu de degré une heure*

Ces transformations qui permettent de passer d'une formulation à l'autre correspondent à celles étudiées dans A. Meunier 1981.
D'un point de vue lexical, on observe encore l'abréviation *SF* pour *stable au feu*, et ceci quels que soient le genre et le nombre de l'adjectif.

Toutes les variations de la phrase canonique déjà étudiées peuvent se combiner avec la transformation en construction adjectivale :

- choix d'un autre classifieur que *degré*, ou bien effacement de ce classifieur :

(30) *La porte est stable au feu (E + de degré + de classement + de durée + d'exigence) une heure*

- variations sur la quantification de la durée :

La porte est stable au feu (une heure + 1 h + 1 heure)

2.2.1. Variantes du verbe support *être*

(32) a.		*se présente*	*(<E>+comme)*	
(32)	N_0	*est*	*<E>*	*stable au feu de degré une heure*
(33) a.			*appelée*	
(33) b.			*dite*	
(33) c.			*réputée*	
(34)			*estimée*	
(35)			*exigée*	
(36)			*classée (<E>+comme)*	
(37)			*définie (<E>+comme*	
(38)			*considérée comme*	
(39)			*présentée comme*	

se présenter est une variante classique de *être*.

Les autres exemples, de (33) à (39), sont considérés comme sémantiquement équivalents et présentent une construction grammaticale superficielle proche, même si les structures profondes des phrases diffèrent.
En effet, si (33a-c) sont effectivement des variantes du verbe support *être* avec seulement des nuances modales : la phrase est une plutôt une définition pour (33a-b), et l'affirmation n'est pas certaine pour (33c), les autres exemples sont en fait issus de transformations de phrases complexes :

(34) a.	*on estime que*	N_0 *est*	*stable au feu de degré une heure*
(35) a.	*on exige que*	N_0 *soit*	
(36) a.	*on classe*	N_0 *comme*	
(37) a.	*on définit*	N_0 *comme*	
(38) a.	*on considère*	(N_0 *comme + que* N_0 *est*)	
(39) a.	*on présente*	N_0 *comme*	

avec une variante de mode pour (35a) : le verbe de la complétive doit être au subjonctif.

2.2.2. Variantes dans l'expression de la durée de stabilité au feu

2.2.2.1. Autres classifieurs

Les variantes concernant le nom classifieur s'appliquent aussi à la construction adjectivale :

La porte est stable au feu (de degré+de classement+de durée+d'exigence) 1h
La porte est stable au feu de degré 1(heure+h)

On peut aussi envisager une autre construction, même si elle n'est pas attestée dans le corpus :

La porte est stable au feu avec (un degré+un classement+une durée+une exigence) 1h

alors que la construction initiale en *avoir* dont elle serait issue paraît stylistiquement peu adroite :

La porte a une stabilité au feu ?(avec (un degré+un classement+une durée+une exigence)) (de+d'+<E>) 1h*

2.2.2.2. Effacement du classifieur

Avec la construction nominale :

(5)	*une stabilité au feu de degré*	*une heure*
(5) a.	*une stabilité au feu de*	*une heure*
(5) b.	*une stabilité au feu d'*	*une heure*

il est possible d'effacer le classifieur tout en conservant la préposition comme dans (5a), éventuellement élidée dans (5b). Dans la construction adjectivale, cette possibilité n'existe pas :

*la porte est stable au feu <E> (*de+*d'+<E>) 1(heure+h)*

2.2.2.3. Quantification de la durée

(4) *La porte a une certaine stabilité au feu*
(31) *La porte est stable au feu*

On remarque que la phrase (31), produite par la transformation adjectivale de (4), est acceptable sans modifieur de *stable au feu* alors que la phrase initiale (4) nécessitait une qualification, même vague, de la stabilité au feu, par exemple par *certaine*. La variante de sens obtenue lors du passage de la construction nominale - où la présence du modifieur pour le classifieur est obligatoire - à la construction adjectivale peut aussi s'observer pour les exemples déjà évoqués en 2.1.2 :

la porte a une certaine largeur
≠ *La porte est large*
La porte est large signifie *La porte a une grande largeur*

de même avec :

la porte a une certaine hauteur
≠ *La porte est haute*
La porte est haute signifie *La porte a une grande hauteur*

et

la porte a une certaine couleur
≠ *La porte est colorée*
La porte est colorée signifie *La porte a beaucoup de couleurs*

En revanche, pour l'expression de la température, le passage de la construction nominale en *avoir* à la construction adjectivale en *être* ne peut se faire à l'aide d'un seul adjectif dérivé de *température* et nécessite un choix entre *chaud, froid, tempéré*, ... selon la valeur de la température.

la pièce a une certaine température

= *la pièce est (chaude + froide + tempérée)*

(4) *La porte a une certaine stabilité au feu*
?= (31) *La porte est stable au feu*

Dans la première paire, les deux phrases sont synonymes. Dans la deuxième, le passage de la forme nominale à la construction adjectivale confère à l'adjectif composé *stable au feu* une valeur absolue, nuance qui n'est pas présente dans la phrase *(4)* initiale.

2.3. Variations de la construction nominale en *avoir*

La phrase canonique est construite sur le verbe *avoir*. Celui-ci admet des variantes obtenues en le remplaçant par d'autres verbes supports (§ 2.3.1) ou par des constructions équivalentes (§ 2.3.2).

2.3.1. Variantes du verbe support *avoir* et verbes supports spécifiques

La forme canonique retenue est formée à l'aide du verbe support *avoir* :

(1) N_0 *<avoir> une stabilité au feu de degré <quantification de la durée>*

Dans cette phrase, *avoir* admet des variantes classiques pour une construction en *avoir* et d'autres plus spécifiques qui introduisent des variations stylistiques et modales, plus ou moins sensibles. On ne peut utiliser directement les transformations présentées par M. Gross (1999) concernaant la fonction sémantique des verbes supports. En effet, la stabilité au feu est une grandeur mesurable à valeurs entières discrètes[1], et l'association entre stabilité au feu et verbe support ne produit pas toujours le même sens que dans un texte généraliste, ou bien nécessite un contexte lié aux conditions de mesure de la stabilité au feu.

Verbes supports neutres

La porte	*correspond à*	*une stabilité au feu de degré une heure*
	montre	
	observe	
	offre	
	présente	
	répond à	
	respecte	
	satisfait	

Verbes supports intensifs

La porte (assume+assure+garantit+procure+pourvoit à) une stabilité au feu de degré une heure

La stabilité au feu traduisant une résistance par rapport à un agent agresseur extérieur, des verbes comme *opposer* ou *préserver* peuvent, dans ce cas, traduire une stabilité au feu renforcée :

---

[1] La valeur de la stabilité au feu est mesurée par l'intermédiaire d'essais normalisés. Même si cela est théoriquement possible, il n'y a pas un continuum de valeurs exprimées : on trouvera dans le corpus *la porte est stable au feu 10 mn*, mais pas *la porte est stable au feu 9 mn*, ou *9 mn 30 s*. Les valeurs prises par la stabilité au feu sont donc entières en minutes, transformées parfois en fractions d'heure : *une demi-heure* au lieu de *15 minutes*.

La porte (oppose+préserve) une stabilité au feu de degré une heure

Aspect duratif

La phrase :

*La porte (?*garde+?*conserve) une stabilité au feu de degré une heure*

nécessite un contexte qui évoque le processus de classement concernant la stabilité au feu pour devenir acceptable, par exemple :

malgré le changement des conditions de mesure

Aspect inchoatif

*La porte (*prend+?*acquiert) une stabilité au feu de degré une heure*

Aspect terminatif

*La porte (?*perd) sa stabilité au feu de degré une heure*

On note ici la présence obligatoire d'un élément qui précise *stabilité au feu* : le possessif *sa* ou un modifieur *qui lui avait été accordée lors des précédents essais.*

Répétition

*La porte (*répète+*reproduit) une stabilité au feu de degré une heure*

L'emploi d'un verbe support qui indique la répétition paraît difficile même avec un contexte.

Verbes liés à un contexte technique ou traduisant une mesure

La porte (admet+supporte+tolère) une stabilité au feu de degré une heure

On retrouve ces variantes de *avoir* dans le vocabulaire mathématique :

La fonction admet un maximum en 0
Cette valeur tolère une erreur de 10%

Verbes traduisant une définition

La porte (caractérise+définit+détermine) une stabilité au feu de degré une heure

Ces phrases sont en fait des phrases complexes dérivées de :

La *caractérisation* *de la stabilité au feu de la porte est une heure*
définition
détermination

Verbes traduisant l'obligation

La porte (requiert+nécessite+réclame) une stabilité au feu de degré une heure

Ces phrases sont en fait des phrases complexes :

Il est (requis+nécessaire+réclamé) que la porte ait une stabilité au feu de degré une heure

Verbes supports causatifs

La porte (accorde+donne+préserve+respecte) une stabilité au feu de degré une heure

Dans les phrases précédentes, *la porte* est considérée comme un élément d'un ensemble (un local par exemple) ; elle ne peut être la cause de la stabilité au feu mais participe au classement de l'ensemble tout entier quant à la stabilité au feu. Ces phrases peuvent être restructurées sous la forme :

La porte fait (une stabilité au feu de degré une heure est (accordée+donnée+préservée+respectée))
La stabilité au feu de la porte fait (une stabilité au feu de degré une heure est (accordée au +donnée au +préservée pour +respectée par)) le local

2.3.2. Autres transformations du verbe dans la phrase à verbe support *avoir*

Les travaux antérieurs sur les lexiques grammaires ont décrit de nombreuses transformations des phrases à verbes supports (J. Labelle 1983 ; M. Gross 1996) ; mais toutes ne sont pas applicables à la forme canonique de l'expression de la stabilité au feu :

(2) *N_0 <avoir> une stabilité au feu de degré <quantification de la durée>*

Les transformations détaillées à la suite constituent des variantes neutres du verbe support.

2.3.2.1. être de ***à la place de*** *avoir*

ce portrait a un relief saisissant — *cette porte a une certaine stabilité au feu*

ce portrait est d'un relief saisissant — *cette porte est d'une certaine stabilité au feu*

cette porte a une stabilité au feu de degré 1h

cette porte est d'une stabilité au feu de degré 1h

2.3.2.2. ***Effacement de*** *qui être*

Luc a dessiné un portrait (qui est+<E>) d'un relief saisissant
le maître d'œuvre demande une porte (qui soit+<E>) de stabilité au feu de degré 1h
le maître d'œuvre demande une porte (qui soit+<E>) à stabilité au feu de degré 1h

2.3.2.3. être à ***à la place de*** *avoir*

ce couteau a une lame rentrante — *cette porte a une stabilité au feu de degré 1h*

ce couteau est à lame rentrante — *cette porte est à stabilité au feu de degré 1h*

2.3.2.4. ***Transformations en formes adverbiales :*** *être avec* ***et*** *être sans*

*?*ce portrait est avec relief* — *?*cette porte est avec stabilité au feu*

cette porte est avec stabilité au feu de degré 1h

ce portrait est sans relief — ** cette porte est sans stabilité au feu*

ce portrait est sans aucun relief — *cette porte est sans aucune stabilité au feu*

cette porte est sans stabilité au feu de degré une heure

2.3.2.5. ***Forme en :*** *il y a*

Luc a un amour-propre qui le désavantage
Il y a chez Luc un amour-propre qui le désavantage

Dans le cas de l'expression de la stabilité au feu, cette transformation n'est jamais possible quelle que soit la préposition choisie pour introduire le sujet porte :

Cette porte a une stabilité au feu de degré 1h
*Il y a (?*avec+*en+*chez +?*pour) cette porte une stabilité au feu de degré 1h*

On pourrait construire une tournure impersonnelle à sujet *on* sous la forme :

Cette porte a une stabilité au feu de degré 1h
On (mesure + constate + vérifie) pour cette porte une stabilité au feu de degré 1h

2.3.3. Variantes sur le verbe support combinées avec celles sur l'expression de la durée de stabilité au feu

Les variations répertoriées pour l'expression de la durée de stabilité au feu sont admissibles avec les variantes lexicales ou syntaxiques sur le verbe support *avoir*. On peut fabriquer les exemples suivants :

Cette porte est d'une stabilité au feu (de degré + de classement + de durée + d'exigence + <E>) 1h
Le maître d'œuvre demande une porte de stabilité au feu (de degré + de classement + de durée + d'exigence + <E>) 1h
Cette porte est à stabilité au feu (de degré + de classement + de durée + d'exigence + <E>) de 1h
Cette porte est avec stabilité au feu (de degré + de classement + de durée + d'exigence + <E>) 1h
Cette porte est sans stabilité au feu (de degré + de classement + de durée + d'exigence + <E>) 1h

2.3.4. Restructuration des groupes nominaux

On peut appliquer à la phrase en *avoir* différentes transformations qui restructurent les *GN* objet ou sujet, ou qui modifient les positions et les fonctions grammaticales des deux *GN*, tout en conservant l'invariant de sens.

2.3.4.1. Restructuration du GN *objet*

(1) *La porte a une stabilité au feu de degré une heure*
(40) *La porte a un degré de stabilité au feu d'une heure*

On passe de (1) à (40) en restructurant le *GN* objet :

stabilité au feu de <classifieur> <quantification de la durée>

devient :

<classifieur> de stabilité au feu de < quantification de la durée >

2.3.4.1.1. Variations sur l'expression de la durée

(40) *La porte a un degré de stabilité au feu d'une heure*

On peut faire varier le nom classifieur devenu objet de la transformée (son effacement, toujours possible, n'a plus d'objet sinon on retomberait sur la forme (1b) inventoriée en 2.1.3) :

(40) a. *La porte a (un degré+un classement+une durée+une exigence) de stabilité au feu de 1h*

On peut exprimer la durée avec les variantes listées en 2.1.2 :

(40) b. *La porte a un degré de stabilité au feu* *(de+d'+<E>)* *un quart d'heure*
30 mn
2 h

et en 2.1.3 :

(40) b. *La porte a un degré de stabilité au feu* *de un quart d'heure*
d'un quart d'heure

2.3.4.1.2. Variations du verbe support

Toutes les transformations listées en 2.3.1, 2.3.2 et 2.3.3 peuvent se combiner avec la restructuration du *GN* objet. Elles sont regroupées selon le plan utilisé pour les étudier dans les paragraphes précédants et listées à la suite :

Autres verbes supports neutres ou porteurs de modalité :

La porte (assume+présente+répond à+respecte+satisfait+supporte+assure) un degré de stabilité au feu de 1h

Variantes du verbe dans la phrase à verbe support *avoir* :

(40) *La porte a un degré de stabilité au feu d'une heure*

A partir de (40) qui contient *avoir* comme verbe support, on peut appliquer toutes les transformations listées en 2.3.2. :

La porte est d'un degré de stabilité au feu de 1h

Le maître d'œuvre demande une porte (qui soit+<E>) d'un degré de stabilité au feu de 1h

La porte est à degré de stabilité au feu de 1h

**La porte est avec degré de stabilité au feu*
**La porte est avec degré de stabilité au feu de 1h*
La porte est sans degré de stabilité au feu
La porte est sans aucun degré de stabilité au feu
La porte est sans degré de stabilité au feu de 1h

2.3.4.1.3. Transformation en construction adjectivale

(40) a. *La porte a (un degré + un classement + une durée + une exigence) de stabilité au feu de 1h*

Selon le nom classifieur utilisé, la transformation en construction adjectivale est ou non possible : ni *degré* ni *durée* n'ont d'adjectif dérivé qui supportent cette transformation. En revanche, pour *classement* et *exigence*, on peut trouver un adjectif ou participe passé à emploi adjectival, *classé* et *exigé*, qui permettent de transformer la construction nominale avec le classifieur de (40) en construction adjectivale avec *être*.

(40) a. *La porte a (un classement + une exigence) de stabilité au feu de 1h*
(43) *La porte est classée 1h de stabilité au feu*
(44) *La porte est exigée de stabilité au feu 1h*

L'inversion entre le *GN* assurant l'expression de la durée et le nom *stabilité au feu* paraît obligatoire pour accorder l'acceptabilité à la phrase transformée (44).

2.3.4.2. Restructurations où stabilité au feu *devient sujet*

(1) *La porte a une stabilité au feu de degré une heure*
(41) *la stabilité au feu de la porte est (de degré+d'+de+<E>) une heure*

On passe de (1) à (41) par une équivalence de phrases à verbe support ; N_0 devient modifieur du nouveau sujet *stabilité au feu* et le modifieur de *stabilité au feu* dans la phrase initiale prend la position d'attribut.

N_0 <avoir> une stabilité au feu de <modifieurSF>
La stabilité au feu de N_0 est de <modifieurSF>

On observe l'alternance obligatoire devant *stabilité au feu* entre le déterminant indéfini *une* de la forme canonique et le déterminant défini *la* de la transformée.

2.3.4.2.1. Variations sur l'expression de la durée

(60) *la stabilité au feu de la porte est (de degré+d'+de+<E>) une heure*

On peut là aussi faire varier le nom classifieur. De plus, dans ce cas, on peut aussi l'effacer :

(60) a. *la stabilité au feu de la porte est (de degré+de classement+de durée+d'exigence+<E>) 1h*

Les variantes sur l'expression de la durée étudiées en 2.1.2. sont valides :

(60) b. *la stabilité au feu de la porte est (de degré+<E>)* *un quart d'heure*

ainsi que celles de 2.1.3.

(60) b. *la stabilité au feu de la porte est* *de une heure*
d'une heure

2.3.4.2.2. Variations du verbe support

(60)	*la stabilité au feu de la porte est*	*<E>*	*(de degré+d'+de+<E>)*	*une heure*
(62)		*réputée*	*(de degré+<E>)*	*une heure*
(63)		*estimée*	*(de degré+à+<E>)*	*une heure*
(64)		*exigée*	*de degré*	*une heure*
(65)		*classée*	*de degré*	*une heure*
(66)		*définie*	*de degré*	*une heure*
(67)		*considérée comme étant*	*de degré*	*une heure*
(68)		*présentée comme étant*	*de degré*	*une heure*

On a déjà étudié les variantes du verbe support *être* dans 2.2.1. Mais, quand elles accompagnent la transformation où le *GN* objet *stabilité au feu* devient sujet, elles paraissent, souvent, un peu artificielles même si elles ne sont pas vraiment inacceptables. Dans aucune des transformations de (62) à (68), on ne peut utiliser la préposition *de* (ou *d'*) seule pour introduire l'attribut. De (63) à (68), la variante du verbe support *être* donne la même construction que la mise au passif de phrases complexes qui seraient :

> *On (estime+exige+considère) que la stabilité au feu de la porte (est+soit) (de degré+d'+de+<E>) une heure*
> *On (classe+définit+présente) la stabilité au feu de la porte comme étant (de degré+d'+de+<E>) une heure*

avec alternance du mode indicatif ou subjonctif pour le verbe de la complétive.

Dans (63), la préposition *à* est plus naturelle que le classifieur pour introduire l'expression de la durée.

Dans (65) et (66), l'adverbe de comparaison *comme* disparaît lors de la transformation.

Dans (67) et (68), la phrase paraît plus acceptable quand on conserve le verbe *être* pour introduire l'attribut du sujet.

Dans tous les exemples, sauf (65), une construction plus naturelle consisterait à faire passer le participe passé à gauche du verbe *être* où il aurait une fonction d'adjectif modifieur de *stabilité au feu* :

(62) a.	*La stabilité au feu*	*réputée*	*de la porte*	*est*	*(de degré+<E>)*	*une heure*
(63) a.		*estimée*				
(64) a.		*exigée*				
(66) a.		*définie*				
(67) a.		*considérée*				
(68) a.		*présentée*				

2.3.4.3. Restructuration du GN *sujet*

(1) *La porte a une stabilité au feu de degré une heure*
(60) *La stabilité au feu de la porte est (de degré+d'+de+<E>) une heure*

(40) *La porte a un degré de stabilité au feu d'une heure*
(50) *Le degré de stabilité au feu de la porte est (d'+de+<E>) une heure*

On a obtenu (40) à partir de la phrase canonique en restructurant le *GN* objet (cf. 2.3.4.1.), on applique maintenant à (40) la transformation étudiée en 2.3.4.2. où le *GN* objet devient sujet de la transformée. Il faut donc une double série de transformations pour passer de (1) à (50).
On peut aussi obtenir (50) à partir de (60) par une transformation identique à celle qui permettrait de passer de (1) à (40).

On observe là aussi l'alternance obligatoire devant le classifieur *degré* entre le déterminant indéfini *un* (quand il est complément d'objet) dans (40) et le déterminant défini *le* (quand il est sujet) dans (50).

2.3.4.3.1. Variations sur l'expression de la durée

Les autres classifieurs étudiés peuvent, à l'occasion de cette transformation, devenir le sujet de la phrase :

(50) a. *(le degré+le classement+l'exigence+la durée) de stabilité au feu de la porte est (d'+de+<E>) une heure*

Toutes les variations sur la quantification de la durée de *stabilité au feu* énoncées en 2.1.2. sont acceptables, ainsi que les variantes sur la préposition *de* devant la quantifieur de la durée.

(50) b.	*Le degré de stabilité au feu de la porte est*	*<E>*	*une heure*
		d'	*une heure*
		de	*une heure*

2.3.4.3.2. Variations du verbe support

(50)	*Le degré de stabilité au feu de la porte*	*est*			*(d'+de+<E>)*	*1h*
(52)		*est*	*réputé*	*comme étant*	*(d'+de+<E>)*	*1h*
(53)		*est*	*estimé*		*(à+<E>)*	*1h*
(55)		*est*	*classé*			*1h*
(56)		*est*	*défini*			*1h*
(57)		*est*	*considéré*			*1h*
(58)		*est*	*présenté*			*1h*
(54)		*est*	*exigé*			*1h*

Comme en 2.3.4.2., l'introduction des variantes du verbe support *être* dans une phrase obtenue par double transformation de la forme canonique conduit à des formulations assez peu naturelles. Elles ne sont pas observées dans le corpus, on peut néanmoins les considérer comme acceptables en rajoutant, sauf pour (54), la forme *comme étant* devant la quantification de la durée.

On peut, de la même manière qu'en 2.3.4.2., modifier l'ordre des éléments de la phrase afin d'obtenir une phrase plus naturelle : le participe passé du verbe variante de *être* peut passer à gauche du verbe en position d'adjectif modifieur du sujet classifieur *degré*.

(52)	*Le degré*	*réputé*	*de stabilité au feu de la porte*	*est*	*(d'+de+<E>) 1h*
(53)		*estimé*			
(54)		*exigé*			
(55)		*classé*			

(56) *défini*
(57) *considéré*
(58) *présenté*

Pour récapituler les transformations concernant les groupes nominaux de la forme canonique (2), on peut considérer que l'invariant de sens peut prendre quatre formes différentes :

(2) N_0 *<avoir> une stabilité au feu de degré <quantification de la durée>*
(10) N_0 *<avoir> un degré de stabilité au feu (de+d'+<E>) <quantification de la durée>*
(51) *le degré de stabilité au feu de* N_0 *est (de+d'+<E>) <quantification de la durée>*
(60) *la stabilité au feu de* N_0 *<être> (de degré+de+<E>) <quantification de la durée>*

Ces transformations qui conduisent à des formulations différentes pour des phrases sémantiquement équivalentes sont classiques dans l'expression d'une **grandeur physique** que l'on veut quantifier. Si on transpose mot à mot ces schémas, par exemple de manière à exprimer une **longueur**, on a les équivalences suivantes :

stabilité au feu	correspond à	*longueur*
degré		*mesure*
unitéTemps		*unitéLongueur*

et les schémas de construction prennent les formes suivantes :

N_0 *<avoir> une longueur de mesure Dnum unitéLongueur*
N_0 *<avoir> une mesure de longueur (de+d'+<E>) Dnum unitéLongueur*
La mesure de longueur de N_0 *est (de+d'+<E>) Dnum unitéLongueur*
La longueur de N_0 *<être> (de mesure+de+<E>) Dnum unitéLongueur*

Les exemples sont les suivants :

La corde a une longueur de (mesure+<E>) 4 mètres
(la mesure de + <E>) la longueur de la corde est 4 mètres
La longueur de la corde est (de mesure + <E>) 4 mètres

Les deux phrases suivantes, qui traduisent l'existence d'un verbe distributionnel, existent pour l'expression de la longueur, mais pas pour la stabilité au feu :

La corde mesure 4 mètres de longueur
La corde mesure 4 mètres

2.4. Remarques sur les déterminants

2.4.1. Déterminants acceptables devant *stabilité au feu*

(1) *La porte a (une+*la) stabilité au feu de degré 1h*
(60) *(la+*une) stabilité au feu de la porte est (de degré+de+<E>) <quantification de la durée>*

On a vu en 2.3.4.2 l'alternance obligatoire des déterminants indéfini et défini devant *stabilité au feu* selon que le *GN* est complément d'objet direct en (1) ou sujet en (60).

(40) *La porte a un degré de (<E>+*une+*la) stabilité au feu de 1h*

(50) *Le degré de (?*la+*une) stabilité au feu de la porte est de 1 h*

En revanche, quand *stabilité au feu* devient complément du nom classifieur le déterminant, qu'il soit défini ou indéfini, est interdit devant *stabilité au feu.*

2.4.2. Déterminants et classifieurs

(40) *La porte a un degré de (<E>+*une+*la) stabilité au feu de 1h*

(50) *Le degré de (?*la+*une) stabilité au feu de la porte est de 1 h*

On a vu en 2.3.4.2. l'alternance obligatoire entre déterminant défini et déterminant indéfini devant le classifieur selon qu'il est sujet comme en (50) ou complément d'objet direct (40).

(1) *La porte a (une+*la) stabilité au feu de degré 1h*

(60) *(la+*une) stabilité au feu de la porte est (de degré+de+<E>) <quantification de la durée>*

On observe, comme dans le paragraphe précédent, que le déterminant est interdit quand le classifieur est modifieur de *stabilité au feu*, qu'il soit épithète (1) ou attribut (60).

3. Les grammaires : construction, dénombrement, utilisation

Le choix de la forme canonique et l'étude de ses transformations donnent l'ossature des grammaires de reformulation. Néanmoins ces grammaires doivent être complétées en tenant compte des possibilités d'insertion qu'offre la syntaxe du français afin de permettre aux automates de reconnaître toutes les phrases du corpus. On s'intéressera ainsi à deux types de fonctions syntaxiques : les adverbes et les modifieurs, et on détaillera, pour chaque schéma de phrase simple représentant une transformation possible, les emplacements auxquels on peut les trouver afin de compléter les automates en conséquence.

3.1. Ajout d'un adverbe

3.1.1. L'adverbe est un prédéterminant ou un postdéterminant numéral

Les structures principales doivent avoir un degré de stabilité au feu d'au moins 2 heures
Les éléments principaux de la structure peuvent être seulement SF de degré une demi-heure

On peut trouver dans cette position des adverbes codés *ADV+* dans les tables du lexique-grammaire :

au plus + exactement + seulement + presque + à peu près + environ ...

3.1.2. Formes jouant un rôle d'adverbe

On peut aussi penser à d'autres formes qu'on peut trouver en position d'adverbe :

N_0 est	*soi-disant*	*stable au feu*
	dit (<E>+ comme)	
	donné (<E>+ comme)	

3.1.3. Possibilité d'insertion d'un adverbe

On représentera par une flèche verticale (↑) l'emplacement où il est possible d'insérer, en l'occurrence, un adverbe :

(2) N_0 <avoir> *une stabilité au feu* *de degré* <*quantification de la durée*>

(32) N_0 <être> *stable au feu* *de degré* <*quantification de la durée*>

(40) c. N_0 <avoir> *un degré de stabilité au feu* *de* <*quantification de la durée*>

(60) c *la stabilité au feu de* N_0 *est* *de degré* <*quantification de la durée*>

(50) c *le degré de stabilité au feu de* N_0 *est* *de* <*quantification de la durée*>

3.2. Ajout d'un modifieur

(1) *La porte a une stabilité au feu de degré 1h*

(62) a. *la stabilité au feu réputée de la porte est (de degré+<E>) une heure*

(52) *le degré réputé de stabilité au feu de la porte est (d'+de+<E>) 1h*
les éléments principaux de la structure ont un degré minimal de stabilité au feu égal au degré coupe-feu de ce plancher.

On a déjà étudié différents modifieurs qui faisaient partie de la phrase canonique ou de ses transformations :

- la quantification de la durée qui assure la fonction de modifieur du nom classifieur, dans la forme canonique (2), comme dans (1) ;
- le participe passé des verbes variantes de <*être*> qui figurent dans les constructions adjectivales où les *GN* sont restructurés. Dans ces phrases, le participe passé joue le rôle de modifieur du sujet *stabilité au feu* comme dans (62a), ou du sujet *degré* comme dans (52).

On s'intéresse maintenant d'une façon systématique aux positions qu'un modifieur peut occuper dans la phrase : en tant que modifieur du sujet, de l'expression "stabilité au feu", du nom ou du *GN* permettant la quantification de la durée.

(2) N_0 <*avoir*> *une stabilité au feu* *de degré* <*quantification de la durée*>

(32) N_0 <*être*> *stable au feu* *de degré* <*quantification de la durée*>

(40) c. N_0 <*avoir*> *un degré de stabilité au feu* *de* <*quantification de la durée*>

(60) c *la stabilité au feu* *de N_0* *est* *de degré* *<quantification de la durée>*

(50) c *le degré* *de stabilité au feu* *de N_0* *est* *<quantification de la durée>*

Pour (61c) et (51c) la possibilité d'insérer un modifieur après le verbe *être* correspond en fait à l'insertion du participe passé du verbe variante de *être,* étudiée respectivement en 2.3.4.2.2 et 2.3.4.3.2.

3.3. Noms classifieurs pour *stabilité au feu*

(50)	*le degré de stabilité au feu de la porte est*	*(d'+de+<E>)*	*une heure*
(1)	*La porte a une stabilité au feu*	*de degré*	*une heure*
	La porte a une stabilité au feu	*<E>*	*une heure*

On a vu en 2.1.1 les différents noms classifieurs spécifiques à l'expression de la stabilité au feu : *classement, degré, durée, exigence.* Lorsque la phrase canonique est restructurée, le classifieur peut se retrouver en position sujet, le déterminant est alors effacé entre le classifieur et l'expression *stabilité au feu* :

*(le classement+le degré+la durée+l'exigence) de (?*la+<E>) stabilité au feu de la porte est (d'+de+<E>) une heure*

ou bien en position de modifieur de *stabilité au feu,* là aussi le déterminant disparaît entre *stabilité au feu* et le classifieur :

*La porte a une stabilité au feu (de+*du) degré une heure*

On peut trouver d'autres noms employés comme classifieurs dans l'expression de la stabilité au feu :

- des noms que l'on peut utiliser aussi pour introduire d'autres grandeurs physiques mesurables ou pour qualifier des objets mathématiques ou physiques :

  caractéristique+définition+détermination+qualité+propriété

  la caractéristique de (la+<E>) stabilité au feu
  la définition de la chaleur massique
  la détermination de la masse volumique
  la qualité de la réfringence du milieu optique
  la propriété d'un espace vectoriel

- ou bien des classifieurs produits par nominalisation des verbes variantes du verbe *<être>,* verbes porteurs de modalité :

  <DET> *estimation* *de* *(la+<E>)stabilité au feu est* *(d'+de+<E>)* *1h*
  évaluation
  garantie
  mesure
  nécessité
  observation
  respect

Quand on utilise pour l'expression de la stabilité au feu, des classifieurs non spécifiques, l'emploi du déterminant défini devant *stabilité au feu* est plus naturel que son effacement.

Dans tous les cas, que le classifieur soit spécifique ou général, son utilisation n'ajoute rien au sens de la phrase : parce que *stabilité au feu* ne peut s'exprimer que par *un classement, un degré, une durée ou une exigence* (qui sont ici synonymes) quand il est spécifique ; parce qu'il n'apporte aucune information supplémentaire quand il est général. Seuls les classifieurs produits à partir des verbes porteurs de modalité modifient le sens de la phrase considérée.

3.4. Dénombrement des différentes formulations

L'étude des transformations applicables à la phrase canonique (2) et qui lui conserve un invariant de sens laisse supposer un nombre important de formulations différentes pour une même idée. Dans ce paragraphe, on essaiera de dénombrer ces formulations. On comptabilisera d'abord séparément les variantes pour chaque groupe de mots de la phrase canonique, puis on combinera ces transformations en séparant le traitement des constructions nominales de celui des constructions adjectivales.

3.4.1. Variantes sur les groupes de la phrase

3.4.1.1. Variantes lexicales

- concernant l'expression *stabilité au feu : GN* complet ou abréviation 2
- concernant l'expression *stable au feu :* groupe modifieur complet ou abréviation 2

On a donc chaque fois deux variantes lexicales, dans les constructions nominales pour *stabilité au feu*, et dans celles adjectivales pour *stable au feu*.

3.4.1.2. Variantes sur l'expression de la durée : un degré de 1h

- concernant le classifieur : 4 classifieurs possibles ou effacement 5
- concernant l'unité de temps : mot complet ou abréviation 2
- concernant la préposition quand le classifieur est effacé : de ou d' ou <*E*> 3

On obtient donc pour l'expression de la durée : 5*2 + 3*2=16 formules possibles.

3.4.1.3. Variantes du verbe :

- le verbe *avoir* ou ses variantes (verbes porteurs de modalité) 8
- transformations de *avoir* 4
- le verbe <*être*> ou ses variantes (verbes porteurs de modalité) 8

On obtient donc pour le groupe verbal :

- pour les constructions nominales en *avoir*, en comptabilisant *avoir*, ses variantes et les transformations auxquelles il se prête : 8+4=12 expressions possibles.
- pour les constructions adjectivales en *être*, en comptabilisant *être* et ses variantes 8 expressions possibles.

3.4.2. Constructions nominales

La phrase initiale est la phrase canonique :

(1) *La porte a une stabilité au feu de degré une heure*

3.4.2.1. Variantes de la phrase canonique

On peut appliquer à cette phrase les variantes suivantes :

- concernant la stabilité au feu : 2
- concernant l'expression de la durée : 16
- sur le verbe *avoir :* 12

La phrase canonique a donc 2*16*12=384 formulations sémantiquement équivalentes.

3.4.2.2. Restructurations du GN objet

(40) *La porte a un degré de stabilité au feu d'une heure*

On peut appliquer à cette phrase les transformations :

- concernant l'expression *stabilité au feu :* mot complet ou abréviation 2
- concernant l'expression de la durée : 16
- concernant le verbe *avoir* : 12

La phrase en *avoir* où le *GN* objet est restructuré accepte donc : 2*16*12=384 formulations sémantiquement équivalentes.

3.4.2.3. Restructurations où stabilité au feu est sujet

(60) *la stabilité au feu de la porte est (de degré+d'+de+<E>) une heure*

On peut appliquer à cette phrase les transformations :

- concernant l'expression *stabilité au feu :* mot complet ou abréviation 2
- concernant l'expression de la durée : 16
- concernant le verbe *être* : 8

La phrase en *être* où *stabilité au feu* est sujet accepte donc : 2*16*8=256 formulations sémantiquement équivalentes.

3.4.2.4. Restructurations du GN sujet

(50) *le degré de stabilité au feu de la porte est (de+d' +<E>) une heure*

On peut appliquer à cette phrase les transformations :

- concernant l'expression *stabilité au feu :* mot complet ou abréviation 2
- concernant l'expression de la durée : 16
- concernant le verbe *être* : 8

La phrase (50) correspondant à une restructuration du *GN* sujet accepte donc : 2*16*8=256 formulations sémantiquement équivalentes.

3.4.3. Constructions adjectivales

La phrase initiale est la suivante :

(30) *La porte est stable au feu de degré une heure*

On peut appliquer à cette phrase les transformations :

- concernant l'expression *stabilité au feu :* mot complet ou abréviation 2
- concernant l'expression de la durée : 16
- concernant le verbe *être* : 8

La construction adjectivale correspondant à la phrase (30) admet donc 2*16*8=256 formulations sémantiquement équivalentes.

Le nombre total de formes, nominales ou adjectivales, présentant le même invariant de sens que la phrase canonique s'élève donc à : 384+384+256+256+256=1536.

3.5. Génération des phrases

Pour illustrer le dénombrement des différentes formes, on a décrit dans un automate les différentes variantes de la phrase canonique :

(1) *La porte a une stabilité au feu de degré une heure*

La description des variantes s'appuie sur l'étude précédente. Pour que l'automate soit lisible, on a utilisé un niveau intermédiaire dans l'écriture :

- *variantesAvoir3èPS =: assume+assure+présente+répond à+respecte+satisfait+supporte*

  variantesEtre3èPSf =: réputée+estimée+exigée+imposée+requise+
  (classée + définie) (<E>+comme (étant+<E>)) +
  (considérée+présentée) (comme (étant+<E>))

  variantesAvoir3èPS et *variantesEtre3èPSf* contiennent les variantes lexicales de *être* et *avoir* qui introduisent des différences modales dans la phrase canonique. Ce sont les variantes étudiées en 2.2.1 et 2.3.1. Tous les accords (verbe, participe passé) se font avec le sujet *porte*.

- *variantesEtre3èPSm =: réputé+estimé+exigé+imposé+requis+*
  (classé+ défini) (<E>+comme (étant+<E>)) +
  (considéré+présenté) (comme (étant+<E>))

  Pour générer les phrases dans lesquelles le classifieur est sujet, il faut disposer des participes passés des variantes de être au masculin singulier : *variantesEtre3èPSm.*

- *transformationsAvoir3èPS =: est d' +d'+est à +est avec +est sans*
  transformationsAvoir3èPS répertorie les transformations syntaxiques de *avoir* étudiées en :

- *LEclassifieur, LAclassifieur, DEclassifieur, Unclassifieur* permettent de tenir compte du genre du classifieur, de l'alternance entre déterminant défini et déterminant indéfini et de son effacement devant la préposition *de :*
  LEclassifieur =: le classement + le degré
  Laclassifieur =: la durée + l'exigence
  Declassifieur =: de classement + de degré + de duré + d'exigence
  Unclassifieur =: un classement + un degré + une duré + une exigence

L'automate complet permet de générer les formes dénombrées dans 3.4.1. Si l'on élimine les variantes lexicales de *être* et *avoir* pour ne conserver que les transformations syntaxiques de *avoir*, en tenant compte de l'alternance entre construction nominale et construction adjectivale, des différentes formes que peuvent prendre le classifieur et l'expression de la durée, on obtient 636 phrases (dont un extrait figure en annexe).

3.6. Utilisation des grammaires de reformulation

L'objectif, chaque fois qu'on reconnaît dans un énoncé une expression se rapportant à la formulation d'une stabilité au feu, est de générer automatiquement toutes les formes sémantiquement équivalentes à l'expression initiale et de les ajouter à l'énoncé recherché.
Dans le cadre de la recherche documentaire, on peut utiliser ces grammaires de deux manières différentes :

- ou bien les appliquer à la question posée par l'utilisateur. Dans ce cas, à une question correspond différentes formulations que l'on considère comme sémantiquement équivalentes et que l'on recherche en parallèle dans le corpus pour répondre à la question initiale ;

- ou bien les appliquer directement au corpus. Ces transformations peuvent alors se faire lors d'une phase de pré-traitement. On ajoute à l'énoncé initial toutes les formulations obtenues grâce à l'application de ces grammaires, et en signalant les ajouts. Lorsqu'un utilisateur pose une question, le texte de sa requête n'est pas modifié mais il est comparé aux différentes formulations que l'on peut trouver dans le corpus.

4. Utilisation de ces grammaires pour d'autres notions de la sécurité incendie

On a défini, pour l'expression de la stabilité au feu, une phrase canonique permettant d'exprimer cette notion et un ensemble de transformations qui peuvent s'appliquer à cette phrase canonique pour produire des formes différentes tout en conservant le même sens à la phrase transformée. On essaie d'utiliser le même contexte et la même méthode pour étudier d'autres notions utilisées en sécurité incendie : le pare-flammes et le coupe-feu.

4.1. L'expression du pare-flammes

La porte est pare-flammes de degré 1/4 heure
ce dispositif est complété par un élément pare-flammes 1/4 d'heure.
le recoupement des couloirs doit être effectué par une porte pare-flammes de degré 1/2 h
un degré pare-flammes de 1/4 heure
le pare-flammes est assuré

La notion de pare-flammes s'exprime à l'aide du même mot *pare-flammes* qui peut être nom ou adjectif, c'est donc une forme ambiguë.

Forme canonique

(101) *La porte a un pare-flammes de degré* *1/4 heure*
(102) *N_0 <avoir> un pare-flammes de degré* *<quantification de la durée>*

Forme canonique tronquée (sans indication de durée)

(103) *N_0 <avoir> (*un + *le + un certain) pare-flammes*
(104) *N_0 <avoir> un pare-flammes comparable à ...*

Variantes orthographiques

(101) a. *La porte a un pare-flammes* *de degré* *1/4 heure*
(101) b. *La porte a un PF* *de degré* *1/4 heure*

Le nom *pare-flammes* admet deux variantes orthographiques : la forme *pare-flammes* et la forme abrégée *PF*, qui peut se substituer à la forme *pare-flammes,* qu'elle soit nominale ou adjectivale.

4.1.1. Variations sur l'expression de la durée du pare-flammes

La démarche ainsi que les résultats sont identiques à ceux développés en 2.1.

4.1.1.1. Classifieurs et effacement du classifieur

(102) a. *N_0 <avoir> un pare-flammes* *de degré* *<quantification de la durée>*
de durée
de classement
d'exigence
(103) *<E>*

4.1.1.2. Elision et effacement de la préposition

Quand le classifieur est effacé, la préposition devant le quantifieur de durée peut être : *de*, ou *d'* ou être effacée :

un pare-flammes <E> *de 1/2 h*
d'
<E>

4.1.1.3. Quantification de la durée

un pare-flammes (de degré+<E>) *quinze minutes*
1/2 h
30 mn

4.1.2. Construction adjectivale en *être*

(101) a. *La porte a un pare-flammes* *de degré* *1/4 heure*
= (130) *La porte est pare-flammes* *de degré* *1/4 heure*

Comme dans la paire *stabilité au feu/stable au feu*, les emplois nominal et adjectival de *pare-flammes* ne sont pas complètement symétriques :

(105) *La porte a un certain pare-flammes*
?= (131) *La porte est pare-flammes*

4.1.2.1. Variantes du verbes support être

(130) *La porte* *est* *pare-flammes* *de degré* *1/4 heure*
est *réputée*
est *estimée*
est *exigée*
est *classée(<E>+ comme)*
est *définie (<E>+ comme)*
est *considérée comme*
est *présentée comme*

4.1.2.2. Variantes dans l'expression de la durée du pare-flammes

sur le classifieur :

(130) *La porte est pare-flammes (de degré+de durée+de classement+d'exigence)* *1/4 h*
<E> *(*de+*d'+<E>)* *1/4 h*

sur la quantification de la durée :

Comme dans 2.2.2.3 on peut observer que la construction adjectivale sans quantification de la durée peut être acceptable alors que la phrase nominale dont elle est issue ne l'est pas. Mais le sens n'est pas conservé non plus dans le passage d'une construction à l'autre :

(131) *La porte est pare-flammes*
**(la porte a un pare-flammes)*

4.1.3. Variations sur la construction nominale en *avoir*

4.1.3.1. Verbes supports spécifiques

Ces verbes spécifiques, neutres ou porteurs de modalité, sont les mêmes que ceux utilisés pour la stabilité au feu. Ils sont en effet spécifiques d'un contexte où une caractéristique mesurée par des essais est accordée ou non à un objet concret.

(101) b. *La porte* *a* *un PF de degré1/4 h*
neutres *(assume+présente+répond à+respecte+satisfait+supporte)*
modalité *assure*

4.1.3.2. Variantes du verbe dans la phrase à verbe support avoir

(101) a. *La porte* *a* *un PF* *de degré1/4 h*
est *d' un PF*
<E> *d' un PF*
est *à PF*
est *(avec+sans) PF*

4.1.3.3. Variantes du verbe support avoir combinées avec celles de l'expression de la durée du pare-flammes

On peut combiner toutes les variantes précédentes avec celles de la durée : (classifieur, préposition, unité).

La porte est d' un PF ((de degré+de durée+de classement+d'exigence+<E>)+(de+d'+<E>)) 1/4 h
<E> d' un PF

est à PF
est avec PF

4.1.3.4. Restructuration des groupes nominaux

4.1.3.4.1. Restructuration du GN *objet*

On passe de la phrase canonique (101) à la transformée (140) :

(101) a. *La porte a un pare-flammes de degré 1/4 h*
(140) *La porte a un degré de pare-flammes d'1/4 h*

Variations de l'expression de la durée

(140) b. *La porte a (un degré+une durée+un classement+une exigence) de PF (de+d'+<E> 1/4 h*
La porte a <E> un PF (de+d'+<E>) 1/4 h
La porte a un degré de PF(de+d'+<E>) 1/4(h+heure)

Variations du verbe support

La porte (assume+présente+répond à+respecte+satisfait+supporte+assure) un degré de PF d'1/4 h

A partir de (140) qui est construite sur le verbe support *avoir*, on peut appliquer toutes les transformations listées en 2.3.2 :

La porte est d'un degré de PF d'1/4 h

une porte (qui soit+<E>) d'un degré de PF d'1/4 h

La porte est à degré de PF d'1/4 h

**La porte est avec degré de PF*
**La porte est avec degré de PF d'1/4 h*
La porte est sans degré de PF
La porte est sans aucun degré de PF
La porte est sans degré de PF d'1/4 h

Transformation en construction adjectivale

A partir de (140b) :

(140) b. *La porte a (un degré+une durée+un classement+une exigence) de PF (de+d'+<E>)1/4 h*

on peut ou non, selon le classifieur, obtenir les phrases transformées :

(150) *La porte est (classée+exigée) PF (de+d'+<E>)1/4 h*

4.1.3.4.2. Restructuration où pare-flammes devient sujet

(101) *La porte a un pare-flammes de degré 1/4 heure*
(105) *le pare-flammes de la porte est de degré 1/4 heure*
(106)

Variations sur l'expression de la durée

classifieur :

(105) *le pare-flammes de la porte est (de degré+de durée+de classement+d'exigence) 1/4 heure*

unité :

(105) *le pare-flammes de la porte est (de degré+de durée+de classement+d'exigence) 1/4 (heure+h)*

préposition :

(105) *le pare-flammes de la porte est (de+d'+<E>) 1/4 (heure+h)*

Variations du verbe support

(105) *le pare-flammes de la porte est* *de degré 1/4 heure*
réputé
exigé
classé (<E>+comme
défini (<E>+comme
considéré comme
présenté comme

4.1.3.4.3. Restructuration du GN sujet

(101) a. *La porte a un pare-flammes de degré 1/4 h*
(140) a. *La porte a un degré de pare-flammes de 1/4 h*
(150) *le degré de pare-flammes de la porte est 1/4 h*
(151)

Variations de l'expression de la durée

classifieur :

(150) a. *(le degré+la durée+le classement+l'exigence) pare-flammes de la porte est 1/4 heure*

unité :

(150) b. *le degré de pare-flammes de la porte est 1/4 (heure+h)*

préposition :

(150) c. *le degré de pare-flammes de la porte est (de+d'+<E>) 1/4 (heure+h)*

Variations du verbe support

(150) *le degré de pare-flammes de la porte est* *1/4 h*
réputé
exigé
classé (<E>+comme)
défini (<E>+comme)
considéré comme
présenté comme

4.2. L'expression du coupe-feu

La paroi est coupe-feu de degré deux heures
Le critère coupe-feu exigé est de 1/4 heure

La notion de coupe-feu s'exprime là aussi à l'aide d'un mot unique *coupe-feu* qui peut être nom ou adjectif, et constitue donc une forme ambiguë.

Forme canonique

(201) *La porte a un coupe-feu de degré 1/4 heure*
(202) N_0 *<avoir> un coupe-feu de degré* *<quantification de la durée>*

Forme canonique tronquée (sans indication de durée)

(203) *N_0 <avoir> (*un + *le + un certain) coupe-feu*

(204) *N_0 <avoir> un coupe-feu comparable à ...*

Variantes orthographiques

(201) a. *La porte a un coupe-feu de degré 1/4 heure*

(201) b. *La porte a un CF de degré 1/4 heure*

Le nom *coupe-feu* admet deux variantes orthographiques : la forme *coupe-feu* et la forme abrégée *CF*, qui peuvent se substituer à la forme *coupe-feu,* qu'elle soit nominale ou adjectivale.

4.2.1. Variations sur l'expression de la durée du coupe-feu

La démarche ainsi que les résultats sont identiques à ceux développés en 2.1.

4.2.1.1. Classifieurs et effacement du classifieur

(202) *N_0 <avoir> un coupe-feu* *de degré* *<quantification de la durée>*
de durée
de classement
d'exigence

(203) *<E>*

4.2.1.2. Elision et effacement de la préposition

(202) *N_0 <avoir> un coupe-feu* *<E>* *(de+d'+<E>) 1/4 h*

4.2.1.3. Quantification de la durée

un coupe-feu (de degré+<E>) *quinze minutes*
1/2 h
30 mn

4.2.2. Construction adjectivale en <être>

(201) a. *La porte a un coupe-feu de degré 1/4 heure*

(230) *La porte est coupe-feu de degré 1/4 heure*

4.2.2.1. Variantes du verbes support <être>

(230) *La porte* *est* *coupe-feu* *de degré* *1/4 heure*
est réputée
est estimée
est exigée
est classée(<E>+ comme)
est définie (<E>+ comme)
est considérée comme
est présentée comme

4.2.2.2. Variantes dans l'expression de la durée de coupe-feu

du classifieur :

(230) *La porte est coupe-feu (de degré+de durée+de classement+d'exigence)* *1/4 h*
<E> *(*de+*d'+<E>) 1/4 h*

de la quantification de la durée :

Comme dans 2.2.2.3 on peut observer que la construction adjectivale sans qualification de la durée peut être acceptable alors que la phrase nominale dont elle est issue ne l'est pas. Mais le sens n'est pas conservé non plus dans le passage d'une construction à l'autre :

(231) *La porte est coupe-feu*
**(la porte a un coupe-feu)*

4.2.3. Variations de la construction nominale en *avoir*

4.2.3.1. Verbes supports spécifiques

Ces verbes spécifiques, neutres ou porteurs de modalité, sont les mêmes que ceux utilisés pour la stabilité au feu. Ils sont en effet spécifiques d'un contexte où une caractéristique mesurée par des essais est accordée ou non à un objet concret.

(201) b. *La porte* *a* *un CF de degré1/4 h*
neutres *(assume+présente+répond à+respecte+satisfait+supporte)*
modalité *assure*

4.2.3.2. Variantes du verbe dans la phrase à verbe support avoir

(201) a. *La porte* *a* *un CF* *de degré1/4 h*
est *d' un CF*
<E> *d' un CF*
est *à CF*
est *(avec+sans) CF*

4.2.3.3. Variantes du verbe support avoir combinées avec celles de l'expression de la durée de coupe-feu

On peut combiner toutes les variantes précédentes avec celles de la durée : classifieur, préposition, unité.

La porte *est* *d' un CF* *((de degré+de durée+de classement+d'exigence+<E>)+(de+d'+<E>)) 1/4 h*
<E> d' un CF
est *à CF*
est *(avec+sans) CF*

4.2.3.4. Restructuration des groupes nominaux

4.2.3.4.1. Restructuration du GN *objet*

On passe de la phrase canonique (201) à la transformée 240) :

(201) a. *La porte* *a* *un coupe-feu de degré 1/4 h*
(240) *La porte* *a* *un degré de coupe-feu d'1/4 h*

Variations de l'expression de la durée

(240) b. *La porte* *a (un degré+une durée+un classement+une exigence) de CF (de+d'+<E> 1/4 h*
un CF (de+d'+<E>) *1/4 h*
un degré *de CF(de+d'+<E>) 1/4(h+heure)*

Variations du verbe support

La porte (assume+présente+répond à+respecte+satisfait+supporte+assure) un degré de CF d'1/4 h

A partir de (240) qui est construite sur le verbe support *avoir*, on peut appliquer toutes les transformations listées en 2.3.2 :

La porte est d'un degré de CF d'1/4 h

une porte (qui soit+<E>) d'un degré de CF d'1/4 h

La porte est à degré de CF d'1/4 h

**La porte est avec degré de CF*
**La porte est avec degré de CF d'1/4 h*
La porte est sans degré de CF
La porte est sans aucun degré de CF
La porte est sans degré de CF d'1/4 h

Transformation en construction adjectivale

A partir de (240b) :

(240) b. *La porte a (un degré+une durée+un classement+une exigence) de CF (de+d'+<E>)1/4 h*

on peut ou non, selon le classifieur, obtenir les phrases transformées :

(250) *La porte est (classée+exigée) CF (de+d'+<E>)1/4 h*

4.2.3.4.2. Restructuration où coupe-feu devient sujet

(201) *La porte a un coupe-feu de degré 1/4 heure*
(205) *le coupe-feu de la porte est de degré 1/4 heure*

Variations de l'expression de la durée

classifieur :

(205) *le coupe-feu de la porte est (de degré+de durée+de classement+d'exigence) 1/4 heure*

unité :

(205) *le coupe-feu de la porte est (de degré+de durée+de classement+d'exigence) 1/4 (heure+h)*

préposition :

(205) *le coupe-feu de la porte est (de+d'+<E>) 1/4 (heure+h)*

Variations du verbe support

(150) *le coupe-feu de la porte est de degré 1/4 h*
est réputé
est exigé
est classé (<E>+comme)
est défini (<E>+comme)
est considéré comme
est présenté comme

4.2.3.4.3. Restructuration du GN *sujet*

(201) a. *La porte a un pare-flammes de degré 1/4 h*
(240) a. *La porte a un degré de pare-flammes de 1/4 h*
(250) *le degré de pare-flammes de la porte est 1/4 h*

(251)

Variations de l'expression de la durée

classifieur :

(250) a. *(le degré+la durée+le classement+l'exigence) coupe-feu de la porte est 1/4 heure*

unité :

(250) b. *le degré de coupe-feu de la porte est 1/4 (heure+h)*

préposition :

(250) c. *le degré de coupe-feu de la porte est (de+d'+<E>) 1/4 (heure+h)*

Variations du verbe support

(205)	*le degré de coupe-feu de la porte*	*est*	*1/4 heure*
		est réputé	
		est exigé	
		est classé (<E>+comme	
		est défini (<E>+comme	
		est considéré comme	
		est présenté comme	

5. Notions combinées

Dans ce paragraphe, on s'intéresse aux phrases dans lesquelles la notion de stabilité au feu est combinée avec d'autres caractéristiques du domaine de la sécurité incendie : pour le quantifier comme en 5.1 ou pour produire un autre critère en 5.2.

5.1. Comparaison de deux critères

On a vu précédemment que l'expression de la stabilité au feu comportait nécessairement une quantification de la durée, ou au moins un modifieur qui s'applique, selon le cas, au classifieur, au nom *stabilité au feu* ou à l'adjectif *stable au feu*. On présente ici une autre manière de quantifier la stabilité au feu en la confrontant avec une grandeur physique comparable. Les exemples issus du corpus constituent des phrases complexes qu'il n'est pas question d'analyser automatiquement. On s'intéresse seulement à la comparaison entre deux caractéristiques distinctes (stabilité au feu et coupe-feu, stabilité au feu et pare-flammes, ou coupe-feu et pare-flammes) ou bien, pour deux éléments de bâtiment différents, à la comparaison du même critère (stabilité au feu, coupe-feu ou pare-flammes).

5.1.1. *stabilité au feu* et *coupe-feu*

(510) *Ces parois doivent* ***être*** *coupe-feu d'un degré égal au degré de stabilité au feu exigé pour la structure avec un minimum d'une demi-heure.*

(511) *Une cage aux parois incombustibles et* ***de*** *degré coupe-feu égal à celui de la stabilité au feu du bâtiment.*

(512) *Des exploitations ou locaux présentant des risques particuliers d'incendie doivent* ***avoir****, dans la hauteur de ces locaux, un degré de stabilité au feu égal au degré coupe- feu du plancher d'isolement supporté.*

(513) *Les parois d'encloisonnement doivent* ***avoir*** *un degré coupe-feu égal au degré de stabilité au feu de la structure du bâtiment.*

(514) *Les éléments principaux de la structure* ***ont*** *un degré minimal de stabilité au feu égal au degré coupe-feu de ce plancher.*

On observe dans le corpus l'alternance entre les constructions en *être* (510) et *avoir* (512) à (514). Dans (511), on a affaire à des transformations successives étudiées en 2.3.2. : le verbe *avoir* est transformé en *être de*, puis *être* est effacé.
L'expression complète du coupe-feu est *le degré de coupe-feu*, mais l'expression *coupe-feu* est aussi un adjectif, et on trouve systématiquement - de (511) à (514) - l'expression *le degré coupe-feu* où la préposition est effacée devant *coupe-feu*. Ce n'est pas le cas pour la stabilité au feu, où la préposition ne peut être effacée.
Dans la comparaison, on peut mettre en première position le coupe-feu (510) ou la stabilité au feu (514).
Les expressions sont précises : les éléments mis en balance sont les degrés de stabilité au feu et de coupe-feu, et non directement la stabilité au feu et le coupe-feu dont on ne sait comparer que les mesures.
La comparaison fait le plus souvent intervenir un deuxième élément du bâtiment et les *GN* de la phrase ont schématiquement les structures suivantes (on qualifie de *comparaison* l'expression linguistique permettant de comparer les classifieurs mis en présence) :

(515) *<classifieur> (de+d'+<E>) CF de N1 comparaison <classifieur>de SF de N2*
(516) *CF (de+d'+<E>) <classifieur> N1 comparaison classifieur de SF de N2*

Pour la phrase complète on a les schémas de construction suivants :

(517) *N1 <avoir> un CF de <classifieur> comparaison <classifieur>de SF de N2*
(518) *N1 <avoir> un <classifieur> (de+<E>) CF comparaison <classifieur>de SF de N2*
(519) *N1 <être> CF de <classifieur> comparaison <classifieur>de SF de N2*

Dans les trois constructions précédentes, on peut exactement intervertir *CF* et *SF*, sauf en (518) où la préposition *de* ne peut être effacée entre le nom classifieur et *SF*. Les transformations de *être* et *avoir* étudiées précédemment sont encore applicables à ces constructions concernant la comparaison des expressions de la résistance au feu.

5.1.2. stabilité au feu et pare-flammes

(520) *Les locaux doivent* ***avoir*** *un degré de résistance au feu défini en fonction du degré de stabilité au feu exigé pour la structure du bâtiment.*

5.1.3. *coupe-feu* et *pare-flammes*

(530) *Les portes* ***ont*** *un degré pare-flammes égal à la moitié du degré coupe-feu des parois*
(531) *Les blocs-portes de la cage d'ascenseur doivent* ***être*** *CF de degré un quart d'heure ou PF de degré une demi-heure.*
(532) *des parois CF de degré une demi-heure et des blocs portes PF de même degré*
(533) *une cage coupe-feu de degré 1h 1/2 et pare-flammes de degré 2h*

Toutes les remarques faites en 5.1.1. sont encore valables dans ce cas. De plus, la forme *pare-flammes* pouvant, comme *coupe-feu,* représenter aussi un adjectif, on observe l'effacement de la préposition *de* devant *pare-flammes* dans l'expression *degré pare-flammes* en (530).

5.1.4. la comparaison se fait sur le même critère

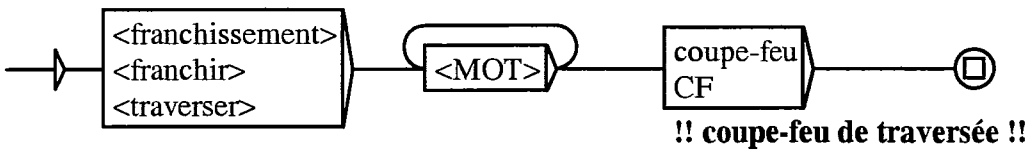
(540) *un bloc-porte à va-et-vient et pare-flammes du même degré que la paroi où il est installé*
(541) *Elle doit présenter le même degré pare-flammes et d'étanchéité que l'ensemble du système*

5.2.1. coupe-feu de traversée

(550) *le clapet **assure** un coupe-feu de traversée de degré 1/4 heure*
(551) *le classement " pare-flammes de traversée " ou " coupe-feu de traversée "*
(552) *le coupe-feu de traversée doit être égal au degré coupe-feu de la paroi franchie*
(553) *une gaine en matériaux incombustibles de coupe-feu de traversée égal au degré coupe-feu de la paroi franchie avec un maximum de 60 minutes*
(554) *le coupe-feu de traversée de la gaine ou du conduit doit **être** égal au degré coupe-feu de la paroi franchie*
(555) *une gaine ou un conduit traversant la paroi coupe-feu séparant deux locaux **satisfait** au critère coupe-feu exigé entre ces deux locaux,*
(556) *s' ils traversent des locaux tiers, ils doivent **assurer** dans la traversée de ces locaux un coupe-feu de degré 1 heure*

La notion de coupe-feu de traversée concerne des éléments du bâtiment qui traversent des parties ou d'autres éléments du bâtiment ayant eux-mêmes des propriétés de coupe-feu. L'expression peut être explicite comme dans les phrases (550) à (554), ou implicite comme dans (555) et (556). Dans ces deux derniers exemples, le coupe-feu à attribuer aux sujets est un coupe-feu de traversée.

Dans le contexte d'une recherche documentaire automatisée à l'intérieur d'un corpus, il serait intéressant de pouvoir identifier cette situation et de reconstruire l'expression complète, automatiquement. En effet, cela permettrait de réduire notablement le bruit dans la mesure où la notion de coupe-feu ou pare-flammes est très courante dans la réglementation, alors que les paragraphes concernant ces mêmes critères lors de la traversée (ou le franchissement) d'un élément par un autre sont beaucoup plus rares. Mais les phrases du corpus mettant en évidence l'intérêt de cette transformation automatique sont complexes et il paraît illusoire d'espérer analyser automatiquement toute phrase illustrant ce critère. Dans un premier temps, on peut alors envisager d'ajouter au texte d'origine (celui du corpus ou de la question posée pour rechercher une information) l'expression *coupe-feu de traversée* à côté de celle présente initialement. L'identification des *GN* initiaux et l'ajout d'information (il est nécessaire d'ajouter l'information et non de la substituer à la suite de mots initiale) se fait à l'aide du transducteur suivant :



5.2.2. pare-flammes de traversée

(550) *le conduit **est** pare-flammes de traversée 30 mn*
(551) *L'exigence pare-flammes de traversée 30 minutes est **réputée satisfaite***
(552) *le calfeutrement entre conduit et paroi traversée doit **être** pare-flammes selon les critères ...*

La notion de pare-flammes de traversée est symétrique à celle de coupe-feu de traversée. Elle est explicitement évoquée en (550) et (551), et implicite en (552).
On choisit d'ajouter l'information *pare-flammes de traversée* de la même manière que pour le coupe-feu de traversée[2].

6. Conclusions

La stabilité au feu est une notion caractéristique de la sécurité incendie ; elle se définit avec des mots du vocabulaire courant mais des constructions et des sens différents de ceux d'un corpus non technique. Dans le cas de la recherche documentaire, les règles syntaxiques et sémantiques de transformation ou d'équivalence établies pour le contexte généraliste ne peuvent s'appliquer.

L'étude précédente montre que cette notion technique simple et précise, de formulation non ambiguë, peut se traduire en un nombre important d'expressions formellement toutes différentes et qui conservent toutes le même invariant de sens.

Ces deux phénomènes, notion caractéristique d'un domaine conduisant à un emploi particulier de mots courants ajoutée à la multiplicité de formulations présentant toutes un invariant de sens, sont souvent observables dans des contextes techniques. Les notions correspondantes doivent être repérées dans le cas d'une recherche documentaire automatisée et les expressions correspondantes réécrites sous forme canonique, ou du moins identifiées, dans le corpus comme dans les questions, afin de diminuer le silence lors de l'interrogation, silence dû à la multiplicité des formulations possibles.

Nous présentons à la suite un prolongement envisagé pour ce travail mais qui n'a pas été réalisé. Il consisterait en la combinaison des connaissances linguistiques sur l'expression de la résistance au feu avec des connaissances techniques sur la sécurité incendie.
En effet, en sécurité incendie, la résistance au feu d'un élément qualifie à la fois la capacité de cet élément à limiter l'extension du feu, et celle, toujours pour le même élément, de conserver ses caractéristiques en cas d'incendie. Les caractéristiques de stabilité au feu, pare-flammes et coupe-feu sont, dans cet ordre, de plus en plus exigeantes. Elles correspondent à différents degrés d'une rubrique plus générique : la résistance au feu.

Un élément peut donc être :

- non stable au feu
- ou stable au feu. Dans ce cas, il peut être :
  + non pare-flammes
  + ou pare-flammes, et alors :
    . non coupe-feu
    . ou coupe-feu

On pourrait donc mettre au point des grammaires qui permettent, dans le cas d'une question portant sur la résistance au feu d'un élément et qui n'apporte pas de réponses suffisantes, de transformer

automatiquement cette question en une autre portant sur le pare-flammes, puis, toujours si on n'obtient pas de réponses satisfaisantes, sur le coupe-feu et enfin sur la stabilité au feu, la caractéristique la moins exigeante.

Conclusions

L'objectif de cette thèse consistait en l'amélioration des accès à l'information contenue dans un corpus documentaire, par la mise au point d'outils linguistiques et informatiques. Partant d'une étude réelle menée à l'aide d'outils existants, nous avons mis en évidence des difficultés en essayant d'en analyser les causes. Notre travail nous a permis de montrer que les possibilités d'amélioration étaient multiples et s'appuyaient à la fois sur des connaissances lexicales et syntaxiques, certaines dépendant du domaine de spécialité, d'autres décrivant des mécanismes généraux de la langue.

En premier lieu, il est essentiel de construire des dictionnaires de spécialité (en mots simples et en mots composés). Comme on l'a vu dans le premier chapitre, la prise en considération d'une terminologie et des règles syntaxiques propres à une spécialité, permettent d'augmenter le rappel lors d'une recherche automatique d'information. Mais ces progrès sont insuffisants, voire trompeurs, s'ils ne s'accompagnent pas d'une amélioration conjointe de la précision : il ne suffit pas de trouver des textes pertinents, il faut aussi éliminer ceux sans rapport avec la question, ou moins pertinents pour que les bonnes réponses ne soient pas rejetées en fin de sélection. L'amélioration doit donc porter, en même temps, sur le rappel et la précision, et des ressources lexicales spécifiques sont nécessaires à cette tâche, mais sûrement pas suffisantes pour produire des améliorations réelles. Les chapitres 2 et 3 montrent la mise au point de ces outils d'aide à la construction d'une terminologie : extraction des mots simples par une méthode numérique, sélection des mots composés par des patrons lexicaux, rôle des noms têtes classifieurs dans la formation de noms concrets du domaine.

Par ailleurs, cette terminologie du domaine est aussi indispensable pour procéder à des vérifications. En effet, l'amélioration des réponses passe par une reformulation de séquences complètes (identifiées dans les questions ou le corpus). Le chapitre 4 a développé cette idée en l'appliquant à la coordination des modifieurs droits dans les groupes nominaux où ceux-ci sont identifiés par rapport à une typologie, puis réécrits en appliquant des transducteurs. Mais ces transformations produisent aussi du bruit car l'identification des expressions s'appuie sur la segmentation de la phrase qui n'est pas encore une tâche complètement automatisée. Un moyen de réduire le bruit produit par la réécriture réside dans la confrontation de chaque expression produite avec des dictionnaires, généraux ou de spécialité : que cette expression soit un mot composé recensé constitue un indice que la reformulation est correcte, même si cela n'en fournit pas la preuve. Nous avons également appliqué les mêmes règles de réécriture pour construire des groupes nominaux candidats à être sélectionnés comme des noms composés. Dans cette expérience, le mécanisme de vérification n'est évidemment pas mis en œuvre.

Le chapitre 5, où on étudie les variations lexicales et syntaxiques autour de l'expression de la stabilité au feu, propose une autre application de cette idée de transformation, mais les mécanismes de vérification y sont plus difficiles à mettre en œuvre parce qu'ils ne concernent pas seulement des expressions lexicalisées du domaine mais plutôt des structures syntaxiques de la langue. En revanche, même si les exemples étudiés sont très spécifiques du domaine de la sécurité incendie, la démarche (i.e. relevé d'expressions associées à des constructions syntaxiques propres au domaine de spécialité pour former des phrases très différentes entre-elles mais présentant toutes la même information) est généralisable à d'autres domaines techniques qui ont en commun d'avoir créé chacune sa langue de spécialité.

Nous avons observé, en utilisant INTEX, que nous repérions plus d'expressions lorsque les règles de réécriture, concernant les groupes nominaux coordonnés par exemple, étaient appliquées que sans ces

réécritures, et que ces expressions étaient pertinentes. Nous avons démontré que les développements présentés au cours de ce travail améliorent les résultats des outils de recherche automatique d'information. Nous avons quantifié cette amélioration pour l'ajout d'une terminologie spécifique, bien que celle-ci soit de taille restreinte. Une perspective intéressante ouverte par nos expériences serait de refaire le test présenté dans l'état de l'art en intégrant tous les outils étudiés (une terminologie complète, des dictionnaires de mots composés, la réécriture des groupes nominaux contenant des modifieurs coordonnés, la reformulation des expressions concernant la stabilité au feu, le coupe-feu et le pare-flamme, la reformulation des expressions concernant le type et la catégorie des établissements, l'identification et le traitement des nombres) et en l'étendant éventuellement à d'autres logiciels du commerce.

D'un point de vue opérationnel, cette thèse témoigne de l'intérêt de la collaboration entre laboratoire universitaire et service de l'administration (ou entreprise) : le premier peut fournir l'expérience de la recherche, un cadre de réflexion et de travail, un environnement pluridisciplinaire, des références théoriques et bibliographiques sur le domaine ; le second un contexte opérationnel qui permet des réalisations concrètes à partir des résultats obtenus et des moyens financiers. Les prolongements les plus immédiats de ce travail consisteraient à mettre en œuvre à grande échelle les outils décrits dans ce document, et à utiliser les méthodes présentées ici pour constituer des ressources linguistiques liées à un autre domaine technique. Citions quelques autres pistes ouvertes par nos expériences : compléter la construction des dictionnaires sous forme d'automates afin de tenir compte de la combinatoire des modifieurs, mettre en place des outils qui permettent de reconnaître l'expression de certaines caractéristiques réglementaires (en particulier le type et la catégorie) des établissements afin de les mettre sous forme canonique, repérer et décrire d'autres notions propres au domaine et qui donnent lieu à des variations lexicales et syntaxiques, identifier et traiter les expressions numériques afin de pouvoir établir des comparaisons entre nombres, quantifier le gain obtenu, en termes de rappel et précision, grâce à l'utilisation des outils décrits dans cette thèse, ...

Ainsi les outils linguistiques et informatiques que nous avons pu tester ou étudier, doivent et peuvent améliorer les performances obtenues lors d'une recherche automatique d'information dans un texte, que leur utilisation soit réservée à des experts ou à des néophytes. Mais pour cela, ces outils ne peuvent faire l'économie de la formalisation de connaissances syntaxiques générales de la langue, même s'ils doivent en même temps porter leurs efforts sur la construction de connaissances spécifiques au domaine. Et ces deux types d'informations sont conjointement indispensables pour se rapprocher de l'objectif poursuivi : associer à un texte sa représentation sémantique, même partielle, et par là rendre possible des tâches comme la recherche automatique d'information dans un texte.

Bibliographie

APPEL, Douglas ; Jerry HOBBS ; John BEAR ; Davis ISRAEL ; Megumi KAMEYANA ; Mabri TYSON. 1993. FASTUS : a finite-state processor for information extraction from real-world text. In Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'93

BOONS Jean-Paul ; Christian LECLERE ; Alain GUILLET. 1976. La structure des phrases simples en français - Constructions intransitives ". Libraire Droz, Genève-Paris

BOTTEQUIN, A.. *Subtilités et Délicatesses de langage*, p. 236.

BOURDEAU, Marc, SIB, Division Systèmes d'information (Sophia-Antipolis), PUCA (Patr.), Medi@Construct (Paris). 1999. Apport des technologies linguistiques à la recherche documentaire. Sophia-Antipolis, CSTB, août 1999, rapport DSI 1748

BOURIGAULT, Didier. 1993, "Analyse syntaxique locale pour le repérage de termes complexes dans un texte", *T.A.L. Traitements automatiques de la composition nominale,* volume 34, 1993, numéro 2, ATALA, Paris

DAILLE, Béatrice. 1994, *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Thèse de doctorat en informatique fondamentale. Université de Paris VII.

DISTER, Anne. 2000. Construire des grammaires de levée d'ambiguïtés pour Intex. *Lingvisticæ Investigationes, Analyse syntaxique et lexicale. Le système INTEX*, Amsterdam/Philadelphia, John Benjamins, , p. 231-245. à paraître

DRILLON, Jacques. 1991. *Traité de la ponctuation française.* collection TEL Gallimard

FOTOPOULOU, Aggeliki. 1995, Construction d'un lexique des termes de télécommunications. Institut national des télécommunications (EVRY)

GARRIGUES, Mylène. 1992. Dictionnaires hiérarchiques du français, *Langue française* 96 "la productivité lexicale. coordonné par André Dugas et Christian Molinier ,Larousse, Paris

GILLET, Alain. 1991, "Dictionnaires électroniques et lexique-grammaire" p.117-128
UAM vol.16 POZNAN 1991

GIRY-SCHNEIDER, Jacqueline. 1978. *Les nominalisations en français – l'opérateur "faire" dans le lexique.* Librairie DROZ

GROSS, Gaston. 1988. Degré de figement des noms composés. *Les expressions figées. Langages* 90, p. 57-72

GROSS, Maurice 1996. Les formes *être PREP X* du français. *Linguisticae Investigationes* XX:2, Amsterdam: John Benjamins, pp. 217-270

GROSS, Maurice. 1986. *Grammaire transformationnelle du français – syntaxe du nom.* Editions Cantilène

GROSS, Maurice. La fonction sémantique des verbes supports. Travaux de linguistique – n°37. éd. B-LAMIROY

GUILLET, Alain. Dictionnaires électroniques et lexique-grammaire. Studia Romanica Posnaniensia Uam vol. 16

HABERT, Benoît ; Christian JACQUEMIN. 1993, "Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques", *T.A.L.* volume 34, 1993, numéro 2 "Traitements automatiques de la composition nominale", ATALA, Paris

LABELLE, Jacques 1983. Verbes supports et opérateurs dans les constructions en *avoir* à un ou deux compléments, *Linguisticae Investigationes* VII:2, Amsterdam: John Benjamins, pp. 237-260

LE PESANT, Denis ; Michel MATHIEU-COLAS. 1998. Introduction aux classes d'objets, *Langages* 131, Larousse, Paris.

LEBART L., A. SALEM. *Statistique textuelle.* DUNOD

LECLERE, Christian ; Alain GUILLET. 1992. *La structure des phrases simples en français - Constructions transitives locatives.* Libraire Droz, Genève-Paris

MATHEZ, Joseph ; Claude MOYE. septembre 1988. Présentation de la réglementation française en sécurité incendie dans les bâtiments in *Cahiers du centre technique et scientifique du bâtiment*, Paris, livraison 292, cahier 2268

MEUNIER, Annie

MONCEAUX Anne. 1993. *La formation des noms composés de structure nom adjectif.* Thèse de doctorat. Université de Marne la Vallée

MORIN, Emmanuel. 1999. Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Multilinguisme, T.A.L.* 1999, volume 40, numéro 1, pp.143-166

POIBEAU, Thierry ; Adeline NAZARENKO. 1999. L'extraction d'information, une nouvelle conception de la compréhension de texte ? *VARIA, T.A.L.* volume 40, 1999, numéro 2, ATALA, Paris

PONCET-MONTANGE, Anne. 1991. *Classification des noms composés N A N et N A V.* Thèse de Doctorat, Villetaneuse: Université ParisXIII.

REY Alain. La terminologie. PUF

Salkoff, Morris. 1979. Analyse syntaxique du français : Grammaire en chaîne; vol.2. *Lingvisticae Investigationes : Supplementa*, La conjonction pp. 15-67, Amsterdam/Philadelphia: John Benjamins.

SENELLART, Jean. 1998. Locating noun phrases with finite state transducers. In Proceedings of ACL-COLING'98

SILBERZTEIN, Max. 1993. Dictionnaires électroniques et analyse automatique de textes - Le système INTEX. MASSON

STRICKER, Mathieu ; Frantz VICHOT ; Gérard DREYFUS ; Francis WOLINSKI. 2000, RFIA 2000, p. 129-137

VAN PETEGHEM, Marleen. 1997. communication au 16ème colloque international sur les lexiques-grammaires comparés des langues romanes de Louvain

VIVES, Robert. 1986. Les noms composés de forme NN. In *Rapport Technique du LADL* n° 25, Paris: Université Paris 7.

VIVES, Robert. 1988b. Comment étudier les noms composés NN du français. *In Rapport Scientifique pour la coopération : Un lexique structuré des noms composés du français pour un traitement en intelligence artificielle*, Paris/Montréal: Université Paris 13 / UQAM.

Dictionnaires

VIGAN (de), Jean. 1996. DICOBAT. Editions ARCATURE

Dictionnaire de linguistique et des sciences du langage 1994 Larousse

Dictionnaire de la langue française Lexis 1993 Larousse

Dictionnaire ROBERT électronique

Règlement de sécurité contre l'incendie relatif aux établissements recevant du public. 1993. Ministère de l'Intérieur, Direction de la Sécurité civile France-Sélection,

Rapports internes - Laboratoire Institut Gaspard Monge - Année 1998

IGM 98-01 Renaud Vérin
Algorithmes de recherches de motifs dans les séquences d'ADN

IGM 98-02 Nicolas Bedon
Langages reconnaissables de mots indexés par des ordinaux

IGM 98 - 03 Marie-Pierre Béal, Dominique Perrin
Une équivalence faible entre les systèmes de type fini

IGM 98 - 04 Elsa Sklavounou
Etude Comparée de la Nominalisation des Adjectifs en Grec Moderne et en Français

IGM 98 - 05 M. Crochemore - F. Mignosi - A. Restivo
Automata and Forbidden Words

IGM 98 - 06 Frédéric Toumazet
Products and symmetrised powers oj irreducible representations of SO (2n)*

IGM 98 - 07 Sun-Woo Choi
Some statistical properties of Korean text corpus and Zipf's law

IGM 98 - 08 Bun Chan Vorac Ung
Fonctions symétriques non commutatives

IGM 98 - 09 Xavier Droubay, Jacques Justin et Giuseppe Pirillo
Episturmian Words and some constructions of de Luca and Rauzy

IGM 98 - 10 M. Crochemore, F. Mignosi, A. Restivo
Text Compression Using Antidictionaries

IGM 98 - 11 S. Veigneau
Ace

IGM 98 - 12 D. Arquès, A. Giorgetti
Counting rooted maps on a surface

IGM 98 - 13 M.P. Béal
On rotationally invariant codes

IGM 98 - 14 B.C.V. Ung
Quasi-differential operators

IGM 98 - 15 B.C.V. Ung
Combinatorial indentities for series of quasi-symmetric functions

IGM 98 - 16 B. Leclerc, J.Y. Thibon, E. Vasserot
Zelevinsky's involution at roots of unity

IGM 98 - 17 O. Foda, B. Leclerc, M. Okado, J.Y. Thibon
Ribbon tableaux and q-analogues of fusion rules in WZW conformal field theories

IGM 98 - 18 B Leclerc, J.Y. Thibon
Littlewood-Richardson coefficients and Kazhdan-Lusztig polynomials

IGM 98 - 19 J.F. Béraud
Etude topologique des cartes, équations fonctionnelles et énumérations

Rapports internes – Institut Gaspard Monge – Année 1999

IGM 99 – 01 J. F. Beraud
MAP, un package Maple pour compter les cartes pointées

IGM 99 – 02 V. Prosper
Polynômes multivariés

IGM 99 – 03 A. Giorgetti
Combinatoire bijective et énumérative des cartes pointées sur une surface

IGM 99 – 04 B. Piranda
Rendu réaliste et surfaces complexes : application à la simulation du milieu maritime

IGM 99 – 05 R. Incitti
The local growth of centralizers

IGM 99 – 06 M.P. Béal, O. Carton
*Asynchronous slinding block maps**

IGM 99 – 07 J. Justin, G. Pirillo
Fractional powers in sturmian words

IGM 99 – 08 C. Allauzen, M. Crochemore, M. Raffinot
Oracle des facteurs, oracle des suffixes

IGM 99 – 09 O. Carton, R. Maceiras
Computing the Rabin Index of a Parity Automaton

IGM 99 – 10 F. Hivert
Combinatoire des fonctions quasi-symétriques

IGM 99 – 11 C. Allauzen, M. Raffinot
Oracle des facteurs d'un ensemble de mots

IGM 99 – 12 M.P. Béal, O. Carton
Determinization of transducers over finite and infinite words

IGM 99 – 13 M. Crochemore, Z. Tronicek
Directed Acyclic Subsequence Graph for multiple texts

IGM 99 – 14 C. Allauzen, M. Raffinot
Algorithme simple et optimal de recherche d'un mot dans un texte

IGM 99 – 15 M.P. Béal, F. Mignosi, A. Restivo, M. Sciortino
Forbidden Words in Symbolic Dynamics

IGM 99 – 16 S.W. Choi
Implantation de dictionnaires électroniques du coréen par automates finis

IGM 99 – 17 B . Gauthier
Calcul symbolique sur les séries hypergéométriques

IGM 99 18 – M. Raffinot
Structures pour la localisation de motifs

IGM 99 19 – J. Berstel, L. Boasson
Shuffle factorization is unique

Rapports internes – Institut Gaspard Monge – Année 2000

IGM 2000 – 01 Frédérique Bassino, Marie-Pierre Béal, Dominique Perrin
Length distributions and regular sequences

IGM 2000 – 02 Cyril Allauzen,
Calcul efficace du shuffle de k mots

IGM 2000 – 03 M. Crochemore, F. Mignosi, A. Restivo, S. Salemi
Data Compression using Antidictionaries

IGM 2000 – 04 Nicolas Bedon
Logic over words on denumerable ordinals

IGM 2000 – 05 Isabelle Icart
Modèles d'illumination locaux pour les couches et multicouches prenant en compte les phénomènes interférentiels

IGM 2000 – 06 Jean Berstel, Luc Boasson
XML Grammars

IGM 2000 – 07 Jean-Paul Fallot
Propriétés statistiques des séquences biologiques et leurs simulations

IGM 2000 – 08 Olivier Carton, Marie-Pierre Béal
Computing the prefix of an automaton

IGM 2000 – 09 Thèse Sabria Benhamida
Mots interdits dans les séquences biologiques

IGM 2000 – 10 Rémi Forax, Etienne Duris, Gilles Roussel
Java Multi-Method Framework

IGM 2000 – 11 Olivier Carton, Wolfgang Thomas
The Monadic Theory of Morphic Infinite Words and Generalizations

IGM 2000 – 12 Véronique Bruyere, Olivier Carton
Automata on linear orderings

IGM 2000 – 13 Maxime Crochemore, Marie-France Sagot
Motifs in sequences : localization and extraction

IGM 2000 – 14 Nadia Pisanti, Marie-France Sagot
Further thoughts on the syntenic distance between genomes

Rapports Internes – Institut Gaspard-Monge – année 2001

IGM 2001 – 01 G. Duchamp, E. Laugerotte, J-G. Luque
Extending the scalars of minimization

IGM 2001 – 02 Rémi Forax, Etienne Duris, Gilles Roussel
A Simple Dispatch Technique for Pure Java Multi-Methods

IGM 2001 – 03 Jean Berstel, Stefano Crespi Reghizzi, Gilles Roussel, Pierluigi San Pietro
A Scalable Formal Method for Design and Automatic Checking of User Interfaces #

IGM 2001 – 04 Catherine Domingues
Etude d'outils informatiques et linguistiques pour l'aide à la recherche automatique d'information dans un corpus documentaire