Université Paris VII

COMPUTER COMMUNICATION and VISION (C2V)

TALANA

Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques

Béatrice DAILLE

Thèse de doctorat en informatique fondamentale
Directeur de thèse : Laurence Danlos
Février 1994

Jury :
Laurence DANLOS
Maurice GROSS
Pierre LAFON
Jean-Marc LANGE
Tony McENERY

Université Paris VII

COMPUTER COMMUNICATION and VISION (C2V)

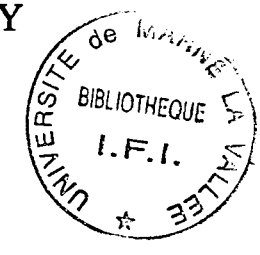
TALANA

Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques

Béatrice DAILLE

Thèse de doctorat en informatique fondamentale
Directeur de thèse : Laurence Danlos
Février 1994

Jury :
Laurence DANLOS
Maurice GROSS
Pierre LAFON
Jean-Marc LANGE
Tony McENERY



Remerciements

Je tiens avant tout à remercier Laurence Danlos pour sa patience, sa direction scientifique et ses encouragements à me guider vers le jour de la soutenance.

Je remercie Maurice Gross, qui après avoir été mon directeur de DEA, me fait l'honneur de présider le jury, ainsi que Pierre Lafon, Jean-Marc Langé et Tony McEnery, pour avoir accepté de participer à la soutenance.

Je remercie les partenaires du projet ET-10/63 : l'université d'Essex, Louisa Sadler, Tim Nicolas et Wim Peters, l'université de Lancaster, Roger Garside, Tony McEnery et Paul Rayson et, en particulier, l'équipe d'IBM-France, Élisabeth Bonnet, Jean-Marc Langé, Éric Gaussier et Frédéric Meunier car cette thèse n'aurait pu être écrite sans le travail de cette équipe, et sans les fonds de la Communauté Européenne.

Je remercie Gaëlle Recourcé, Max Silberztein, Frédéric Meunier, et Éric Gaussier pour leur relecture attentive et critique de ce mémoire.

Enfin, je remercie C_2V qui m'a offert d'excellentes conditions matérielles et un accueil chaleureux.

Table des matières

Table des figures

Introduction

La terminologie est un problème crucial dans le monde scientifique et technique. Chaque domaine particulier use de l'aide d'un nombre important de termes qui en reflètent les concepts, autrement dit la connaissance. Ces termes sont principalement des noms composés qu'un expert doit recenser dans une banque terminologique. Cet expert n'a en général aucune connaissance linguistique et son recensement s'appuie principalement sur l'intuition. Il en résulte un manque de cohérence dans les banques terminologiques et une absence totale de structures pour les termes relevés. Cette difficulté dans la création de banques terminologiques se ressent cruellement en traitement automatique du langage naturel (TALN), car aucun programme d'analyse ou de traduction ne peut aboutir sans une liste (structurée) de termes. À l'inverse, le linguiste élaborant un programme d'analyse ou de traduction n'a en général aucune intuition sur les concepts (termes) d'un domaine technique. Il lui est donc difficile de compléter ou d'affiner les banques de données fournies par les experts.
Au vu de ces difficultés tant dans le domaine industriel que dans celui du TALN, il est devenu urgent d'élaborer des méthodes permettant de créer automatiquement des banques terminologiques. Le stockage de plus en plus systématique des textes sur support informatique fournit un nouvel outil de travail : les corpus qui décrivent implicitement en contexte réel la connaissance du domaine. Il s'agit donc de tenter d'extraire automatiquement les termes d'un domaine à partir de corpus. Pour cela, la recherche s'est orientée massivement depuis quelques années vers les méthodes statistiques qui ont fait leur preuve en reconnaissance de la parole. Les méthodes purement statistiques d'extraction automatique de terminologie ont donné des résultats qui ont été jugés en première approximation comme satisfaisants, les listes obtenues par des calculs statistiques contenant effectivement un nombre non négligeable de termes du domaine. Mais ces listes incluent aussi un bruit important, i.e. des séquences qui ne correspondent pas à des termes. Pour réduire l'excès bruit dans les résultats obtenus à partir de calculs purement statistiques, nous nous sommes proposé d'élaborer une méthode combinant données linguistiques et calculs statistiques. Plus précisément, à partir d'une étude linguistique rigoureuse des noms composés terminologiques, nous avons mis au point des filtres linguistiques qui permettent une première sélection des séquences susceptibles, sur le plan morphosyntaxique, d'être des noms

composés. C'est aux séquences ainsi sélectionnées que nous appliquerons divers modèles statistiques avant d'en évaluer les résultats. Cette évaluation nous amène à la conclusion suivante: le meilleur critère statistique, i.e. le critère qui fournit une liste de noms composés en minimisant au mieux le bruit et le silence, est un test du rapport de vraisemblance où les événements fréquents sont pris en compte. Cette conclusion va à l'encontre de nombreux travaux sur l'extraction de ressources lexicales qui proclament que leurs critères d'association (information mutuelle, nombreuses mesures définies à partir de tableaux de contingence) sont de meilleurs indicateurs que la fréquence. La liste des noms composés fournie par le critère de vraisemblance n'est évidemment pas parfaite. Elle contient en particulier une part importante de silence puisque les séquences n'apparaissant qu'une fois sont éliminées *a priori*. Elle contient aussi une part de bruit, bien que celui-ci soit réduit. Nous verrons que ce bruit est la plupart du temps dû aux problèmes de surcomposition et de modification des noms composés. La surcomposition et la modification sont des phénomènes relativement connus sur le plan de la linguistique théorique mais qui posent des problèmes empiriques souvent insolubles. En effet, il est aussi difficile pour un linguiste que pour un expert du domaine étudié (ici les télécommunications), de statuer sur la séquence *antenne parabolique de réception*: est-ce une surcomposition de *antenne parabolique* et de *antenne de réception* ou est-ce une modification de *antenne de réception* par l'insertion de l'adjectif *parabolique*? Les calculs statistiques ne fournissent guère de réponse à ce type de question, d'où l'introduction de bruit. Signalons à ce propos que la surcomposition et la modification des noms composés rendent les textes "techniques" aussi délicats à traiter que les textes "littéraires" : des problèmes classiques, chers aux linguistes, tels que les dépendances non bornées, le temps et l'aspect n'y sont effectivement présents que sous forme simplifiée, mais les problèmes liés à la terminologie n'ont reçu à l'heure actuelle aucune solution théorique ou empirique vraiment satisfaisante.

Notre travail s'est effectué dans le cadre du projet de recherche et développement piloté par la Commission Européenne, ET-10/63. Ce projet auquel ont collaboré le Centre Scientifique d'IBM-France, l'université de Lancaster, l'université d'Essex et C_2V-TALANA, avait pour objectif d'étudier l'apport des statistiques en traduction automatique. Nous nous sommes concentrée sur l'extraction automatique de la terminologie monolingue, mais nous évoquerons brièvement l'extraction de terminologie bilingue effectuée par IBM-France et l'extraction de structures argumentales effectuée par l'Université de Lancaster.

La première partie examine les principales applications des statistiques dans le traitement automatique du langage naturel, leur domaine précis d'utilisation, ainsi que les résultats qu'elles produisent.
La deuxième partie est consacrée à une étude linguistique de la terminologie. Dans une première étape, nous examinons quelques travaux effectués sur la composition, puis sur la composition nominale. Puis nous appliquons les propriétés

dégagées par les linguistes à notre domaine technique: nous présentons une typologie des noms composés terminologiques de type élémentaire, nous étudions comment de nouveaux noms composés se créent à partir des noms composés de type élémentaire, puis nous en recensons les principales variantes.
La troisième partie décrit la modélisation de notre corpus de manière à pouvoir utiliser les spécifications linguistiques des noms composés dans un programme d'extraction, puis le stockage du corpus dans une base de données relationnelle.
La quatrième partie concerne notre méthodologie: ses principes, la technique choisie pour filtrer et compter les co-occurrences des noms composés, les modèles statistiques, l'évaluation de ces modèles, et la justification du choix final de n'en retenir qu'un seul. Nous terminerons ce chapitre en évoquant l'extension de la méthodologie à d'autres ressources lexicales: l'extraction de terminologie bilingue et l'extraction de structures argumentales.

Chapitre 1

Les modèles statistiques en Traitement du Langage Naturel : état de l'art

Avec l'arrivée de techniques permettant la gestion et le traitement de grands corpus, les méthodes statistiques sont devenues très présentes dans le traitement du langage naturel. Ignorées ou mésestimées par les linguistes, celles-ci occupent néanmoins une place de plus en plus importante. Il suffit d'examiner les compte-rendus de colloques internationaux sur le traitement informatique des langues pour se rendre à l'évidence. Outre cet intérêt scientifique, des produits de pointe utilisent des modèles statistiques. L'exemple du logiciel de reconnaissance de la parole commercialisé par IBM montre leur efficacité. Il s'agit toutefois de préciser leur domaine d'utilisation. Pour cela, nous examinerons, dans ce chapitre, leurs principales applications. Nous aborderons successivement l'étiquetage grammatical, l'analyse syntaxique, l'extraction de ressources lexicales monolingues, l'alignement de phrases, l'extraction de ressources lexicales bilingues, et pour conclure, la traduction automatique.

1.1 Assignation d'étiquettes grammaticales

L'assignation d'étiquettes grammaticales est une première étape dans l'analyse automatique d'un texte. Elle consiste à affecter au mot sa catégorie grammaticale dans le contexte où il apparaît. Les étiquettes grammaticales correspondent aux traditionnelles parties du discours auxquelles sont intégrées des informations supplémentaires telles que le nombre et le genre pour les noms, les adjectifs et les participes passés, la personne pour le verbe, etc. Le nombre d'étiquettes varie suivant les programmes et les applications envisagées ; par exemple, la détermination de classes sémantiques pour les noms sera plus facile avec des étiquettes dissociant par exemple, les noms de pays, les noms de ville, les noms propres et les noms

temporels qu'avec une étiquette générale regroupant n'importe quel type de nom. Il existe de nombreux programmes stochastiques d'étiquetage pour l'anglais: [Jelinek, 1985], [Church, 1988], [Cutting *et al.*, 1992], [Garside *et al.*, 1987], etc. Pour le français, seul semble véritablement performant celui qui est développé par l'équipe de recherche d'IBM dans le domaine de la reconnaissance de la parole ([Dérouault et Merialdo, 1986], ..., [El-Bèze, 1993]). Ces programmes stochastiques se déroulent généralement en deux phases:

- **Reconnaissance** des mots du texte et **assignation** des étiquettes grammaticales par l'intermédiaire soit d'un dictionnaire, soit d'un analyseur morphologique. L'un comme l'autre regroupe les **règles lexicales.**

- **Choix** d'une étiquette lorsque plusieurs sont possibles en examinant le contexte. Cette tache est effectué par les **règles contextuelles.**

Examinons maintenant ces deux types de règle:

1.1.1 Règles lexicales

L'identification d'un mot du texte s'effectue avec un dictionnaire ([Church, 1988], [Dérouault et Merialdo, 1986], etc.), ou avec un analyseur morphologique ([Garside *et al.*, 1987], [Cutting *et al.*, 1992], etc.).

1.1.1.1 Dictionnaire

Le dictionnaire est une liste de formes fléchies. Chaque forme fléchie est accompagnée d'une ou plusieurs étiquettes grammaticales. Ainsi, par exemple, le mot *passé* reçoit les étiquettes NOM et PARTICIPE-PASSÉ. Certains dictionnaires attachent à ces étiquettes une "probabilité lexicale" qui rend compte de l'usage plus ou moins fréquent de l'unité syntaxique (mot, étiquette) considérée. Lorsque le mot n'est associé qu'à une seule étiquette, la probabilité est égale à 1 ; lorsque plusieurs étiquettes sont possibles, c'est la somme des probabilités assignées à chacune des étiquettes qui est égale à 1. Par exemple, la forme fléchie française *sommes* reçoit quatre étiquettes différentes dans le dictionnaire du programme d'IBM:

- AUXE5 pour l'auxiliaire *être* à la première personne du pluriel
  P(AUXE5 | *sommes*) = 0,2

- VERB2 pour le verbe *sommer* à la deuxième personne du singulier
  P(VERB2 | *sommes*) = 0,01

- SUBSFP pour le nom féminin *somme* au pluriel dont l'un des sens est: *résultat d'une addition*
  P(SUBSFP | *sommes*) = 0,78

- SUBSMP pour le nom masculin *somme* au pluriel utilisé dans l'expression *faire un somme*
  P(SUBSMP | *sommes*) = 0,01

L'étiquette la plus probable pour le mot *sommes* est donc SUBSFP.
Ces probabilités sont généralement assignées automatiquement à l'aide d'un corpus d'apprentissage (voir section 1.1.3). Ces dictionnaires sont obligatoirement très importants: 200 000 formes, par exemple, pour celui qu'[El-Bèze, 1993] a utilisé pour le français.

1.1.1.2 Analyseur morphologique

Dans les programmes utilisant un analyseur morphologique, le dictionnaire est de taille réduite: dans [Garside *et al*, 1987], le dictionnaire contient 7 200 unités lexicales dont les mots outils de la langue anglaise, les mots les plus fréquents et les exceptions aux règles morphologiques utilisées. Ces programmes consultent d'abord le dictionnaire, et dans le cas où le mot ne s'y trouve pas, ils examinent sa terminaison. Toujours dans [Garside *et al.*, 1987], la liste des suffixes les plus courants consiste en 720 terminaisons de une à cinq lettres. Par exemple, le suffixe *-ness* assigne l'étiquette NOM, *-mp* les étiquettes NOM ou VERBE sauf pour les exceptions *damp* et *lump* enregistrées dans le dictionnaire. Dans le cas où la terminaison d'un mot s'unifie avec plusieurs suffixes, c'est le suffixe le plus grand qui est retenu. Ainsi, pour la forme *available*, seul *-able* ADJECTIF est retenu parmi les suffixes avec les étiquettes correspondantes: *-able* ADJECTIF, *-ble* NOM ou VERBE, *-le* NOM existant dans la liste. Un traitement particulier est appliqué pour certains suffixes tel que le *-s* marque du pluriel et *-er* qui demande l'examen de la racine du mot. Lorsqu'un suffixe introduit plusieurs étiquettes, [Garside *et al.*, 1987] n'associent pas à ces étiquettes de "probabilités lexicales" comme dans le cas des programmes d'étiquetage avec dictionnaire. Ces étiquettes sont classées suivant un ordre séquentiel respectant les choix les plus probables avec des marqueurs spéciaux pour celles qui correspondent aux cas rares.[1] L'avantage de l'analyseur morphologique sur le dictionnaire concerne le traitement des mots inconnus. Alors que les programmes utilisant des dictionnaires attribuent une catégorie "fourre-tout" aux mots inconnus, par exemple NPRO (signifiant nom propre) dans [El-Bèze, 1993], l'analyseur morphologique effectue des prédictions en utilisant la terminaison du mot. Les programmes avec analyseur morphologique sont donc plus portables d'un type de corpus à un autre. En revanche, les programmes utilisant un dictionnaire sont beaucoup plus rapides, la consultation s'effectuant soit en un temps logarithmique, soit en un temps proportionnel à la longueur du mot et indépendamment de la taille du dictionnaire.

[1] La stratégie de [Cutting *et al.*, 1992] semble être identique à celle de [Garside *et al*, 1987] mais aucune précision n'est donnée quant à la taille du dictionnaire, la liste des suffixes utilisés, etc.

1.1.2 Règles contextuelles

Lorsqu'un mot accepte plusieurs étiquettes grammaticales, l'utilisation de règles contextuelles s'impose. Il s'agit en examinant le contexte local dans lequel le mot apparaît d'en déduire sa catégorie syntaxique. Ces règles se présentent sous la forme suivante :

$$X\,?\,Y \rightarrow A$$

où ? représente le mot avec plusieurs étiquettes possibles et X et Y les étiquettes de ses contextes gauche et droite. Cette règle signifie que l'étiquette du mot ambigu est A si elle est précédée de X et suivie par Y.
Les règles de désambiguïsation lexicale dans les programmes stochastiques sont généralement exprimées en terme de biclasses ou de triclasses. Un "biclasse" est une règle comportant deux étiquettes grammaticales apparaissant séquentiellement et à laquelle est associée un poids ; un "triclasse" comporte trois étiquettes séquentielles et un poids. Par exemple, le poids associé à la séquence : VERBE + ARTICLE + NOM est supérieur au poids associé à la séquence : NOM + ARTICLE + NOM. Le calcul de ce poids est expliqué dans la section suivante (1.1.3). Beaucoup de biclasses et triclasses sont reconnaissables comme des structures syntaxiques simples. Ces règles syntaxiques qui ne portent que sur trois mots au plus ne couvrent évidemment ni les dépendances non bornées, ni les contraintes syntaxiques sur de longues distances comme [Chomsky, 1957] l'a fait remarquer ; [Church, 1988] le reconnaît mais constate aussi que ces phénomènes syntaxiques sont plutôt rares, surtout dans les textes techniques et que les essais qui ont été faits avec des règles contextuelles de plus grande taille n'améliorent pas de façon significative les résultats, tout en étant beaucoup plus difficiles à obtenir et à appliquer. Si [Church, 1988] et [Cutting *et al.*, 1992] se contentent de triclasses, [Garside *et al.*, 1987] et [El-Bèze, 1993] utilisent biclasses et triclasses. Cette dernière méthode est plus souple puisque si un triclasse n'existe pas, il y a quand même une possibilité que le biclasse existe. Ainsi, si le triclasse X Y Z n'apparaît pas dans les règles, les biclasses X Y ou Y Z peuvent exister.

1.1.3 Calcul des probabilités lexicales et génération des règles contextuelles

Les probabilités lexicales comme les règles contextuelles sont calculées ou générées automatiquement à partir d'un "corpus d'apprentissage". Pour l'anglais, le premier et le plus utilisé des corpus d'apprentissage est le *Brown Corpus* (Brown University Standard Corpus of Present-Day American English) [Francis et Kucera, 1982] ; il contient 1 000 000 mots étiquetés à la main. Pour le français, il n'existe pas de corpus similaire au *Brown Corpus* et il faut appliquer une procédure semi-automatique comme nous le verrons ci-dessous.

1.1.3.1 Probabilités lexicales

Les probabilités lexicales sont estimées automatiquement à partir d'un corpus d'apprentissage et rendent compte de la fréquence d'apparition des différentes étiquettes grammaticales d'un mot donné. Par exemple, [Church, 1988] définit la probabilité lexicale de la manière suivante:

$$P(Y \mid mot_i) = \frac{freq(Y \mid mot_i)}{freq(mot_i)}$$

où la probabilité que mot_i reçoive l'étiquette Y est exprimée comme le rapport des fréquences du mot_i associé à l'étiquette Y et des fréquences du mot_i associé à n'importe quelle étiquette.

1.1.3.2 Modèles biclasse ou triclasse

Les modèles biclasse ou triclasse sont générés automatiquement à l'aide d'un corpus d'apprentissage, et les scores qui sont alloués aux règles dépendent de leur fréquence d'apparition. Le corpus d'apprentissage est soit préalablement étiqueté ([Church, 1988], [Dérouault et Merialdo, 1986], [Garside *et al*, 1987], etc), soit "brut" ([Jelinek, 1985]). Ainsi, [Church, 1988] pour l'anglais utilise le *Brown Corpus*, corpus d'apprentissage étiqueté, et génère les probabilités contextuelles de la manière suivante: étant données trois étiquettes E_1, E_2, E_3 et un mot M, la probabilité que le mot M ait l'étiquette E_1 est égale à:

$$P(E_1 \mid M) = \frac{freq(E_1E_2E_3)}{freq(E_2E_3)}$$

où $freq(E_1E_2E_3)$ représente la fréquence de la séquence E_1 E_2 E_3 et $freq(E_2E_3)$ la fréquence de la séquence E_2 E_3. La procédure d'apprentissage est entièrement automatique.
Le français ne possédant pas de grand corpus étiqueté, [Dérouault et Merialdo, 1986] étiquettent manuellement une petite partie d'un texte (2 000 mots) et en relèvent les biclasses et les triclasses résultants. Puis, avec le modèle de Markov utilisant ce petit nombre de biclasses et de triclasses, et à l'aide de l'algorithme de Viterbi, ils étiquettent automatiquement une autre portion du texte mais cette fois beaucoup plus importante (16 000 mots). À la suite de cet étiquetage, les biclasses et les triclasses sont relevés automatiquement et ajoutés aux biclasses déjà obtenus dans la première étape. Ce nouveau modèle est appliqué sur une nouvelle partie du texte, de taille encore plus importante (47 000 mots). L'étiquetage résultant sera corrigé à la main, et c'est à partir du résultat final que seront relevées les probabilités contextuelles du programme. La procédure d'apprentissage s'est déroulée semi-automatiquement du fait de la non-disponibilité

d'un grand corpus d'apprentissage du français.
La modélisation et les algorithmes utilisés par les autres programmes stochastiques sont principalement : Chaîne de Markov ou Chaîne de Markov cachées pour les modèles, algorithmes de Vitterbi ou de Baum-Welch (Forward-Backward) [Baum, 1972].

1.1.4 Évaluation de ces programmes

Le pourcentage d'assignations correctes d'étiquettes grammaticales d'un programme stochastique est de l'ordre de 95 à 98 %. Le score de 98% est donné pour celui de [Church, 1988] ou de [Foster, 1991], 97 % pour celui de [Garside *et al.*, 1987], plus de 95 % pour celui de [Dérouault et Merialdo, 1986]. L'évaluation est effectuée en prenant un morceau de texte étiqueté au hasard dans le corpus et en le vérifiant à la main. Ces programmes sont particulièrement performants : pour donner une élément de comparaison, un programme d'étiquetage, qui n'utilise pas de statistiques mais uniquement des règles grammaticales comme celui de [Green et Rubin, 1971] obtient 77 % d'assignations correctes.
Les erreurs résiduelles des programmes pré-cités ont été étudiées par [Macklovitch, 1992]. Elles correspondent à des situations où le modèle assigne des probabilités égales à des séquences concurrentes. [Macklovitch, 1992] montre qu'avec quelques règles de grammaire, les erreurs d'un programme stochastique sont aisément corrigées. Le programme d'étiquetage idéal serait donc un programme stochastique pour effectuer l'essentiel du travail et des techniques linguistiques pour les cas litigieux.
[Brill, 1992] a conçu un programme d'assignation d'étiquettes grammaticales qui est à la croisée des programmes stochastiques et des programmes à base de règles : les probabilités sont présentes dans le dictionnaire mais non dans les règles contextuelles. Par ailleurs, les règles contextuelles ne sont pas utilisées pour étiqueter le texte mais pour corriger l'étiquetage obtenu avec le dictionnaire. Les règles contextuelles examinent l'étiquette qui précède et celle qui suit le mot courant. Elles sont générées automatiquement à partir d'un corpus d'apprentissage. Cette phase d'apprentissage est assez longue et requiert des connaissances linguistiques puisque les règles doivent être classées suivant un ordre d'efficacité décroissante ; une règle étant considérée comme efficace si elle corrige beaucoup d'erreurs sans en induire. La qualité de l'affectation d'étiquettes avoisine celle des programmes stochastiques (95 %).
L'étiquetage morphologique qui assigne à chaque mot du texte son lemme peut être vu comme comme un sous-produit immédiat de l'étiquetage grammatical puisque, dans la grande majorité des cas, un lemme correspond à une forme fléchie unique dès que sa classe grammaticale est connue ([El-Bèze, 1993]). Les programmes d'étiquetage morphologique sont souvent intégrés aux programmes d'étiquetage grammatical.

1.2 Analyse syntaxique

Une grammaire probabiliste contient des informations statistiques concernant la fréquence d'utilisation de ses règles. Ces statistiques, exprimées sous forme de probabilités, "guident" l'analyseur syntaxique: lorsque celui-ci a le choix entre plusieurs règles, il choisit celle qui est la plus probable. Les probabilités sont calculées grâce à un corpus d'apprentissage, c'est-à-dire un corpus préalablement analysé. Si déjà l'élaboration à la main d'un corpus d'apprentissage étiqueté est onéreux (voir section 1.1.3), celle d'un corpus analysé est pire encore. [Garside *et al.*, 1987] affirment que l'élaboration de leur banque d'arbres le *LOB tree bank*, comprenant 50 000 mots et 2 284 phrases, a pris deux ans. Certains, comme [Briscoe et Carroll, 1993], utilisent l'analyseur syntaxique qu'ils veulent transformer en analyseur probabiliste, pour créer le corpus d'apprentissage. Lorsque plusieurs analyses sont possibles, l'analyseur laisse la main au linguiste pour choisir la bonne analyse. Les analyses obtenues sont plus cohérentes puisqu'elles sont souvent attribuées automatiquement. Les analyseurs probabilistes utilisent soit uniquement une grammaire probabiliste ([Baker, 1979], [Briscoe et Carroll, 1993], [Fujisaki *et al.*, 1989], [Pereira et Schabes, 1992], [Sharman *et al.*, 1990]), soit conjointement une grammaire probabiliste et un programme stochastique d'étiquetage grammatical ([Dérouault, 1985], [Garside *et al.*, 1987], [Magerman et Marcus, 1991]).

1.2.1 Approche grammaticale

L'aventure de la grammaire probabiliste a commencé avec [Baker, 1979] qui proposait d'utiliser en reconnaissance de la parole une grammaire probabiliste "context-free" (CF), sous forme normale de Chomsky (CNF).
Plus récemment, [Fujisaki *et al.*, 1989] ont repris la technique de [Baker, 1979]. Ils ont assignées aux règles CF des probabilités acquises automatiquement: dans un premier temps, toutes les probabilités sont nulles, puis elles sont itérativement ré-estimées en utilisant l'algorithme de Baum-Welh sur les phrases analysées. Ces phrases analysées ne sont pas vérifiées ; les statistiques obtenues prennent donc en compte à la fois des analyses correctes et des analyses incorrectes. Sur 84 phrases d'au plus 11 mots, l'analyse avec la probabilité la plus forte est une analyse correcte pour 72 d'entre-elles, soit 85 %. Néanmoins, comme aucune indication n'est donnée sur le corpus, les phrases pourraient ne comprendre que des constructions simples répétées.
[Sharman *et al.*, 1990] ont conduit une expérience similaire mais avec un formalisme grammatical plus élaboré, capable de différencier les règles de dominance et les règles de précédence (ID/LP). Les probabilités initiales, toujours établies en fonction des fréquences d'utilisation des règles, ont été extraites automatiquement d'un corpus d'un million de mots analysé à la main. Sur 42 phrases de 30 mots au plus, 38 (88 %) cumulent une probabilité maximale et une analyse correcte.

[Pereira et Schabes, 1992] ont montré sur une grammaire lexicalisée que la manière dont étaient acquises les probabilités associées aux règles, c'est-à-dire soit automatiquement, soit sous contrôle humain, déterminent la fiabilité de l'analyseur. En mode contrôlé, une analyse est acceptée si elle est cohérente: elle peut ne pas correspondre à l'analyse donnée par le corpus d'apprentissage. Les résultats sont frappants: lorsque l'apprentissage s'effectue automatiquement, 35 % des phrases ayant obtenu la probabilité la plus haute correspondent à une analyse correcte contre 78 % avec un apprentissage contrôlé.
[Briscoe et Carroll, 1993] prennent en compte les résultats obtenus par [Pereira et Schabes, 1992] et assignent les probabilités semi-automatiquement. De plus, de manière à prendre en compte le contexte et la profondeur de l'arbre, ils utilisent une grammaire d'unification, c'est-à-dire un formalisme plus puissant que les règles CF simples. Les pourcentages de phrases correctement analysées sont de 76 % pour les phrases extraites du corpus d'apprentissage et de 75 % pour les nouvelles phrases.
Toutes ces expériences ont été réalisées sur l'anglais.

1.2.2 Combinaison des approches markovienne et grammaticale

L'équipe de recherche d'IBM avec [Dérouault, 1985] ont été les premiers et d'ailleurs les seuls, à notre connaissance, à expérimenter l'analyse probabiliste sur le français. Ses travaux s'effectuant dans le cadre de la reconnaissance de la parole, [Dérouault, 1985] a comparé sur un même texte oral les performances d'un modèle markovien et d'une grammaire probabiliste. Le modèle markovien donne de meilleurs résultats avec un temps d'exécution beaucoup plus rapide et de plus, ne requiert pas l'écriture de règles de grammaire; 93,2 % de reconnaissance correcte de phrases pour le modèle markovien, 89,4 % pour l'analyseur probabiliste. Mais, comme ces deux approches, markovienne et grammaticale, sont complémentaires, l'une prédisant le contexte local et l'autre la structure globale, [Dérouault, 1985] a tenté d'associer ces deux approches et de mélanger les probabilités des triclasses et biclasses aux probabilités des règles de grammaire. Les résultats obtenus sont supérieurs à ceux du modèle Markovien (94,5 % pour le modèle mixte).
[Garside *et al.*, 1987] ont aussi expérimenté une solution mixte entre le modèle Markovien et la grammaire probabiliste. Sur un corpus étiqueté par leur programme d'assignation d'étiquettes grammaticales (voir section 1.1), ils ont inséré des crochets indexés entre des étiquettes indiquant des frontières de constituants. Par exemple pour la séquence VERBE ARTICLE, ils insèrent les crochets indexés suivants:

$$\text{VERBE } {}_{V}] \; [_{N} \text{ ARTICLE}$$

où $_V$] signifie fin d'un verbe et [$_N$ début de groupe nominal. Malheureusement, ces règles qui s'appuient sur un contexte local marquent les constituants à leur entrée mais, pas à leur fin ou alors d'une manière ambiguë. Le programme utilisant une grammaire CF probabiliste se charge de déterminer les positions finales des constituants ouverts grâce au modèle markovien. Il examine toutes les positions possibles de fermeture d'un constituant et ne retient que celle associée à la probabilité la plus forte. Le calcul de la probabilité d'un constituant est normalisé en fonction de la profondeur de l'arbre et de la longueur de la branche, et est égal au produit des probabilités de chacun des ses constituants, ceux-ci pouvant être des éléments lexicaux. Le programme commence par fermer les constituants les plus profonds de l'arbre et passe ensuite aux constituants supérieurs. Les probabilités associées aux règles CF ont été calculées automatiquement à partir d'un corpus de 50 000 mots analysé à la main. Les résultats sur 250 phrases tirées au hasard du corpus d'apprentissage donnent 50 % d'analyses correctes. Les analyseurs probabilistes sont donc beaucoup moins performants que les programmes stochastiques d'étiquetage grammatical et malgré les nouveaux efforts développés dans ce domaine, aucun véritable analyseur syntaxique de bonne qualité n'est disponible actuellement.
Les statistiques ont aussi été utilisées pour aider à résoudre certains problèmes rencontrés au cours de l'analyse syntaxique, tels que l'attachement prépositionnel [Hindle et Rooth, 91] ou la référence anaphorique [Dagan et Itai, 1990]. Ces travaux relevant des ressources lexicales monolingues, ils sont présentés dans la section suivante.
Un système de règles de préférence a été mis au point par [Dologlou *et al.*, 1991] de façon à choisir parmi les analyses produites par un analyseur syntaxique de type stratificationnel, au niveau de l'analyse en constituants immédiats, laquelle est la plus probable. Ces règles de préférences sont établies à la suite d'un examen statistique d'un ensemble d'analyses de phrases, produites par l'analyseur.

1.3 Ressources lexicales monolingues

L'application de modèles statistiques à des corpus monolingues apporte des informations quantitatives et qualitatives sur les affinités lexicales que peuvent présenter certains mots entre eux. Ces co-occurrences révèlent des propriétés syntaxiques ou sémantiques importantes pour le traitement automatique d'une langue. Ces informations lexicales diffèrent selon qu'elles ont été obtenues à partir d'un corpus non traité ou d'un corpus préalablement étiqueté ou analysé. En France, des recherches sur l'application et le développement de méthodes statistiques lexicales ont été menées dès les années soixante, par exemple au Trésor de la Langue Française à Nancy, sous la direction de P. Imb et à Besançon sous la direction de B. Quemada. Les travaux français que nous connaissons sur les co-occurrences sont ceux de [Lafon, 1984] et sont antérieurs à ceux

de [Church et Hanks, 1989] sur l'anglais ; si les méthodologies et les mesures statistiques employées diffèrent selon les auteurs, les résultats obtenus sont tout à fait similaires. Ce sont pourtant les travaux de [Church et Hanks, 1989] qui sont à l'origine du regain d'intérêt de la communauté scientifique pour la statistique lexicale. Nous allons successivement présenter les travaux sur les co-occurrences lexicales de [Lafon, 1984] sur le français, de [Church et Hanks, 1989] sur l'anglais et de [Calzolari et Bindi, 1990] sur l'italien, qui s'appliquent à des corpus non traités, puis nous examinerons quelques travaux ultérieurs tous effectués sur l'anglais et qui s'appuient sur des corpus étiquetés ou analysés. Nous terminerons ce tour d'horizon par la présentation des travaux de [Dagan *et al.*, 1991] qui utilisent plusieurs corpus en différentes langues.
L'équipe de lexicométrie de [Lafon, 1984] a expérimenté trois méthodes pour sélectionner les co-occurrences d'un texte, seule la dernière a été retenue et appliquée à des corpus. Cette méthode mesure le nombre de co-occurrences d'une paire composée de deux unités lexicales dans la phrase. Nous allons expliquer comment le nombre d'occurrences de deux unités lexicales dans le corpus est calculé ; ce nombre d'occurrences est utilisé ensuite pour mesurer la co-occurrence d'une paire. Soit F et G, les unités lexicales d'une paire, S, un délimiteur de phrase, et la configuration suivante rencontrée dans une portion de corpus (seules les apparitions de F, G et S sont indiquées, et donc un certain nombre de mots peuvent apparaître avant, entre et après F et G) :

S S F G S F F G S S F G F S S S F S G F S S

Nous remarquons que les occurrences de F et de G alternent de telle sorte que parfois F est suivie de G et d'autres fois G est suivie de F. Ces deux orientations possibles de la paire (F, G) sont distinguées par les deux couples suivants auxquels est associé un chiffre représentant le nombre d'occurrences rencontrées dans la portion de corpus ci-dessus :

- si F précède G, $(F \rightarrow G) = 3$
- si G précède F, $(G \rightarrow F) = 2$

Le nombre d'occurrences de la paire (F, G) correspond à la somme des nombres d'occurrences des deux couples : $(F, G) = (F \rightarrow G) + (G \rightarrow F) = 5$. Pour chaque occurrence d'un couple, la distance séparant les deux unités lexicales est relevée. Un couple ou une paire sera caractérisé par son nombre d'occurrences et par la distance moyenne séparant ses unités lexicales, sans tenir compte de l'orientation. Un modèle probabiliste dont la variable de décision se rapproche d'une loi binomiale est ensuite appliquée aux paires et aux couples. Nous ne présenterons pas ce modèle qui comporte un développement assez long ; nous retiendrons uniquement que la probabilité affectée à chaque paire (resp. chaque

couple) permet de juger la co-occurrence de ses unités lexicales. Cette méthode est appliquée sur un corpus d'environ 8 000 unités lexicales et on relève les paires (resp. les couples) avec un nombre d'occurrences supérieur ou égal à 2 et dont la probabilité de co-occurrence dépasse un seuil arbitraire. Les paires sélectionnées (environ une pour six) correspondent soit à des co-occurrences non orientées, soit à des co-occurrences orientées, ces dernières étant, à peu près, deux fois plus nombreuses. Ces co-occurrences sont d'une extrême diversité: parmi les co-occurrences non-orientées se trouvent des associations sémantiques référant à des appels thématiques à assez longue distance comme par exemple la paire (*divergence*, *revendications*) et parmi les occurrences orientées, des mots composés comme *état d'esprit* mais aussi des associations fonctionnelles comme le couple (*par*, *d'*) qu'il est difficile de caractériser.
[Church et Hanks, 1989] sur l'anglais, [Calzolari et Bindi, 1990] sur l'italien ont extrait des co-occurrences lexicales d'un corpus en utilisant une mesure proche du concept d' "information mutuelle" : le "score d'association" (association ratio). La formule exacte du score d'association est donnée dans notre chapitre IV, section 4.3.2. Retenons seulement pour l'instant que le score d'association évalue le lien que peuvent avoir deux mots entre eux ; un score positif indique que les deux mots apparaissent plus fréquemment ensemble que séparément ; un score négatif que ces deux mots sont en distribution complémentaire. La technique employée est la suivante : sur un corpus monolingue, on déplace une fenêtre de taille paramétrable i. Le premier mot m_1 de la fenêtre est associé à tous les mots m_2, m_3, ..., m_i, qui le suivent à l'intérieur de cette fenêtre. Des couples, (m_1, m_2), ..., (m_1, m_i), sont ainsi formés. Chaque couple est accompagné de son nombre d'occurrences dans le corpus, ainsi que de la distance séparant les deux mots pour chaque occurrence. Les couples de mots regroupent toutes sortes d'associations en terme de parties du discours: (nom, nom), (nom, verbe), (verbe, préposition), etc. Ils sont ensuite classés en fonction des valeurs décroissantes de score d'association. Les plus grandes valeurs désignent :

- des couples regroupant soit des mots composés, soit la totalité ou une sous-partie d'une expression figée, des noms propres, etc., lorsque la variance des distances entre les deux éléments du couple est faible,

- des couples regroupant des associations sémantiques, lorsque la variance est plus importante.

Ces résultats corroborent ceux de [Lafon, 1984] bien que le modèle statistique employé soit différent, que les co-occurrences soient orientées et que la méthode soit moins souple, la taille de la fenêtre étant fixe. Le score d'association apparaît être une mesure appréciable pour l'identification des co-occurrences lexicales et sera utilisé dans de nombreux travaux ultérieurs. Néanmoins et sans vouloir dès à présent dévoiler notre appréciation de cette mesure, il est intéressant de garder en tête la remarque de [Calzolari et Bindi, 1990] précisant que les couples avec

un score d'information élevé partagent tous un faible nombre d'occurrences.
Les autres travaux se concentrent sur des questions plus spécifiques et réutilisent pour la plupart, la mesure introduite par [Church et Hanks, 1989] :
[Dagan et Itai, 1990] utilisent les données statistiques sur les co-occurrences lexicales extraites d'un corpus analysé pour lever les ambiguïtés des références anaphoriques apparaissant à l'intérieur d'une phrase. Le pronom étudié est le pronom *it*. L'expérience se déroule en deux phases :

- un corpus de 28 millions de mots est analysé par un analyseur traditionnel et les occurrences des relations syntaxiques (sujet, verbe), (verbe, objet direct), (adjectif, nom) sont relevées et intégrées dans une base de donnée,

- un ensemble de phrases contenant au moins une occurrence du pronom *it* est extrait au hasard du corpus. Le pronom doit être une anaphore ambiguë appartenant à l'une des relations syntaxiques ci-dessus. Pour chaque occurrence du pronom, ses référents possibles sont identifiés à la main, puis substitués au pronom dans sa relation syntaxique. Les fréquences des relations syntaxiques impliquant chacun des référents possibles sont examinées et la plus élevée est retenue.

Sur 38 occurrences ambiguës du pronom *it*, relevées manuellement, ce simple modèle statistique fondé sur la fréquence réussit à identifier le bon référent pour 33 d'entre-elles, soit 87 %.
[Hindle, 1990] réalise un classement sémantique des noms par l'application d'une mesure statistique définie à partir du score d'association ([Church et Hanks, 1989]). L'idée reprise est celle de l'hypothèse distributionnelle de [Harris, 1968], à savoir que chaque nom peut être caractérisé par l'ensemble des verbes avec lesquels il est employé. L'auteur examine donc les noms dans les positions syntaxiques sujet ou objet direct du verbe. Un corpus de 6 millions de mots est analysé, toujours par un analyseur traditionnel, et les relations syntaxiques (sujet, verbe) et (verbe, objet direct) sont relevées. Pour chaque couple (*nom*, *verbe*) dans une des deux relations syntaxiques, le score d'association est calculé. La mesure de similarité qui caractérise un couple formé de deux noms (nom_1, nom_2) est la somme des scores d'association obtenus par les couples (*nom*, *verbe*) où nom_1 et nom_2 sont employés avec le même verbe dans une même relation syntaxique. Cette mesure est calculée pour tous les couples de noms du corpus. Pour obtenir une classe d'équivalence, il suffit de choisir un nom et de lister tous les couples où il apparaît. Les résultats sont encourageants : pour beaucoup de noms, les classes obtenues sont sémantiquement cohérentes même si parfois des éléments perturbateurs, dont le nombre n'est pas précisé, réussissent à s'infiltrer.
Le travail de [Brent, 1991a et 1991b] diffère des travaux précédents, au sens où il cherche à obtenir un classement sémantique des verbes, puis à déterminer leur complémentation à partir d'un corpus non-traité.
Les deux expériences partagent une même première étape : l'identification des

verbes du corpus. Le programme considère qu'un mot est un verbe si dans son entourage immédiat se trouve un pronom ou un nom propre. Les verbes ainsi reconnus ont été vérifiés : sur un corpus de 2,6 millions de mots (*Wall Street Journal*), où 5 000 mots ont été étiquetés comme verbes, 28 seulement n'en n'étaient pas. [Brent, 1991b] utilise ensuite une grammaire locale exprimée à l'aide d'automates finis déterministes pour extraire les actants des verbes. Il n'est évidemment pas question de chercher à reconnaître n'importe quel type de groupe nominal. Seuls sont identifiés les groupes nominaux exprimés par un pronom ! La grammaire utilise donc les verbes étiquetés et les classes lexicales à éléments finis tels que les pronoms, les déterminants, les prépositions et les verbes auxiliaires et isole les relations syntaxiques suivantes : (verbe, objet direct), (verbe, objet direct, complétive), (verbe, objet direct, infinitive), (verbe, complétive), (verbe, infinitive). Les complémentations obtenues sont vérifiées à la main et un taux d'erreur de 2 % est constaté. L'étape suivante consisterait à calculer le score d'association et d'utiliser cette mesure statistique pour éliminer les mauvaises complémentations.

Pour conclure ce tour d'horizon des recherches effectuées sur l'extraction automatique de ressources lexicales monolingues à partir de corpus monoligues, nous retiendrons que les corpus sont effectivement riches d'informations lexicales et que l'extraction de celles-ci grâce à des modèles statistiques tels que le score d'association et d'informations linguistiques est une expérience à mener sur le français ; ce que nous ferons dans le chapitres IV.
Nous allons maintenant présenter les travaux de [Dagan *et al.*, 1991] qui utilisent plusieurs corpus monolingues pour tenter de découvrir la meilleure traduction d'un mot dans le contexte d'une phrase. En traduction automatique, et plus exactement au cours du transfert lexical, un mot dans une langue peut avoir plusieurs traductions possibles dans une autre langue. Comment déterminer quelle est la meilleure traduction d'un mot dans le contexte d'une phrase? [Dagan *et al.*, 1991] cherchent à résoudre ce problème, et plus généralement le problème de l'ambiguïté lexicale au cours de l'analyse, grâce à l'examen statistique des traductions possibles d'un mot dans une autre langue. Ils affirment que :

> "les différents sens d'un mot sont déterminés par les différentes traductions de ce mot dans une autre langue".

Les ambiguïtés qu'ils veulent traiter sont celles qu'une analyse syntaxique n'a pu résoudre. Plutôt que de recourir à des calculs sémantiques et pragmatiques, ils proposent d'utiliser des informations statistiques sur la traduction : selon eux, les ambiguïtés lexicales qui apparaissent dans un langue, n'en sont plus dans une autre langue. L'expérience menée tente de résoudre des ambiguïtés lexicales en Hébreu et en Allemand grâce à un modèle statistique appliqué sur un corpus anglais. Cette expérience se déroule en deux étapes : 1) Préparation du corpus

anglais, 2) Application du modèle statistique.

1. Les corpus de l'anglais qu'ils ont à leur disposition sont : *Washington Post* (40 millions de mots), *Associated Press* (24 millions de mots), *Hansard* (80 millions de mots). L'expérience nécessite un corpus analysé : l'analyseur traditionnel utilisé n'acceptant que les phrases comportant au plus 25 mots, la taille du corpus est réduite à 55 millions de mots anglais. De plus, cet analyseur échoue pour 35 Cette analyse isole les relations syntaxiques de la phrase, i.e. verbe et arguments (sujet, objet direct, objet indirect), verbe et modifieurs (adverbe, groupe prépositionnels), nom et arguments, nom et modifieurs (adjectif, groupe prépositionnel).

2. Le traitement préalable du corpus de la langue cible étant terminé, une phrase comprenant au moins un mot ambigu est extraite soit du corpus Hébreu (dix paragraphes tirés de la presse israélienne), soit du corpus Allemand (douze paragraphes tirés de la presse allemande). Cette phrase est traduite à la main en anglais et les diverses options de traduction qui résultent des mots ambigus sont conservées. Par exemple, dans la proposition en hébreu : *l-hassagat hitqaddmut*, le mot *hitqaddmut* (objet direct du verbe *l-hassagat* = *to achieve*) a trois traductions possibles en anglais : *progress, advance, advancement.* Pour déterminer la bonne traduction, des statistiques sont effectuées sur le corpus anglais analysé, pour les trois relations syntaxiques suivantes :

   - (verbe, objet-direct *achieve progress*),
   - (verbe, objet-direct *achieve advance*),
   - (verbe, objet-direct *achieve advancement*).

   Les fréquences d'apparition de ces trois relations syntaxiques sont calculées et données en entrée au modèle statistique qui calcule une mesure appelée : "rapport de chance" (*odds ratio*). La relation syntaxique avec le rapport de chance le plus élevé est retenue comme la bonne traduction. Pour l'exemple ci-dessus, c'est la relation : (verbe-objet direct *achieve progress*) qui sera choisie.

La levée des ambiguïtés des mots ambigus de l'allemand ou de l'hébreu dans une phrase est donc effectuée grâce à l'examen statistique de leurs traductions en anglais sur un corpus préalablement analysé. Les résultats de l'expérience sont donnés en terme d'applicabilité et de sélection correcte. L'applicabilité correspond aux couples de mots dans une certaine relation syntaxique de la langue source dont les traductions conservent la même relation syntaxique dans la langue cible. La sélection correcte correspond à la bonne traduction.

- Pour Hébreu-Anglais :
  - Applicabilité du modèle : 70 %

- Sélection correcte à 92 %

- Pour Allemand-Anglais :

  - Applicabilité du modèle : 59 %
  - Sélection correcte à 75 %

C'est l'inapplicabilité de cette méthode qui est source majeure d'échecs. Les résultats obtenus sont de 18 % supérieurs à ceux d'un modèle statistique où serait choisi le couple le plus fréquent. Néanmoins, le corpus de départ étant de 144 millions de mots et le corpus de base analysé sur lequel le modèle statistique s'applique de 35 millions de mots, il est étonnant que les auteurs considèrent que c'est le manque de données qui est la cause principale d'échec ; ces données sont déjà très importantes ! L'utilisation d'un corpus bilingue où le même domaine sémantique est représenté dans les deux langues améliorerait obligatoirement les résultats.

1.4 Alignement de phrases

Les chercheurs en traduction automatique (voir section 1.6) et en ressources lexicales bilingues (voir section 1.5) se sont récemment intéressés à l'étude de textes parallèles, comme par exemple les débats du Parlement canadien écrits en français et en anglais (*Hansards*) ou les publications de l'Union des Banques Suisses (*UBC*) disponibles en français, anglais et allemand. Les textes parallèles ont été utilisés en premier par [Brown *et al.*, 1988] pour leur système de traduction automatique (voir section 1.6). Les corpus parallèles sont des corpus alignés, c'est-à-dire où un segment d'un texte est la traduction d'un segment du même texte dans une langue différente. L'unité de segmentation atteinte est généralement la phrase. L'alignement de textes bilingues effectué à la main est très coûteux et demande de bonnes connaissances linguistiques dans les deux langues. Les techniques utilisées pour obtenir un alignement phrase à phrase sont, soit lexicales et s'appliquent aux mots comme [Catizone *et al.*, 1989] et [Warwick et Russel, 1990], soit statistiques et s'appliquent alors à la phrase comme [Brown *et al.*, 1991] et [Gale et Church, 1991a]. Examinons ces techniques :

1. L'approche lexicale de [Catizone *et al.*, 1989] s'inspire de l'intuition humaine qui, pour aligner des phrases, se penche sur les mots qui les composent. Le programme de [Catizone *et al.*, 1989] construit donc les correspondances de phrases sur la base de la fréquence statistique d'un mot et de sa traduction. Le programme utilise un dictionnaire bilingue et ce sont les paires de mots dont l'un est la traduction de l'autre qui déterminent les alignements. La potentialité de réussite de cette approche est forte, mais le programme est long et inutilisable sur de gros corpus. [Warwick *et al.*, 1992]

reconnaissent l'échec de cette approche: un même mot n'est pas toujours traduit de la même façon, le rédacteur par souci stylistique n'utilise pas toujours le même signifié pour un même signifiant. Elles préféreront pour des raisons d'efficacité l'approche statistique de [Gale et Church, 1991a].

2. Les programmes de [Brown *et al.*, 1991] et [Gale et Church, 1991a] utilisent un modèle probabiliste simple basé sur la longueur de la phrase. Ils ont constaté que les phrases longues (resp. les phrases courtes) d'une langue se traduisent généralement par des phrases longues (resp. des phrases courtes) dans une autre langue, mais ils diffèrent sur la manière de mesurer la longueur de la phrase: [Brown *et al.*, 1991] comptent le nombre de mots, [Gale et Church, 1991a] le nombre de caractères. L'alignement de phrases passe par l'alignement de parties plus importantes telles que le paragraphe. Cette première étape est indispensable à la bonne application du modèle statistique. Le plus souvent, une phrase d'une langue est traduite par une phrase dans une autre langue : c'est l'alignement 1-1. Néanmoins, quelques cas sont plus difficiles:

   - alignement 1-0, où une phrase dans une langue est simplement omise dans l'autre langue,
   - alignement 2-1, où deux phrases dans une langue sont traduites par une seule phrase dans une autre langue,
   - alignement 2-2, où deux phrases dans une langue correspondent à deux phrases dans une autre langue sans qu'il existe de correspondances 1-1 entre ces quatre phrases,
   - etc.

   Leurs programmes sont simples, rapides, pratiques, n'utilisent aucune information lexicale et ont un pourcentage élevé de réussite.

Nous allons décrire plus précisément les programmes d'abord de [Brown *et al*, 1991], puis de [Gale et Church, 1991a]:
[Brown *et al.*, 1991] ont utilisé cette technique pour aligner plusieurs millions de phrases extraites du compte-rendu des débats du Parlement Canadien. Ils espèrent obtenir de 96 % à 97 % de bons alignements. La méthode est la suivante: ils réalisent d'abord des alignements de régions de texte autour de "points d'ancrage"; les points d'ancrage apparaissent dans les deux corpus et sont, par exemple, le nom des intervenants, les dates, les commentaires, etc. Ces régions alignées rassemblent en moyenne une dizaine de phrases. Puis, considérant que le nombre de mots d'une phrase dans une langue est en corrélation avec le nombre de mots de sa traduction dans une autre langue, ils estiment les probabilités suivantes grâce à un modèle de Markov à deux états: probabilité qu'une phrase anglaise de longueur l_e donne une phrase française de longueur l_f, probabilité que

deux phrases anglaises de longueur ll_e donnent une phrase française de longueur l_f, etc. Ces probabilités sont obtenues à partir du corpus pré-aligné en comptant le nombre de fois qu'une phrase de longueur l_e a comme correspondant possible une phrase de longueur l_f dans une même région. Malheureusement, ce programme n'a été réellement évalué que sur mille alignements 1-1 français-anglais ; sur ces mille alignements, six sont mauvais. Le score donné de 99 % d'alignements corrects est donc à prendre avec circonspection puisqu'il ne concerne que 0,03 % des alignements obtenus et que de plus, aucun résultat n'est donné pour les cas plus complexes, pourtant prévus par leur programme, tels que les alignements : 1-0, 2-1, 2-2, 3-1, 3-2.
[Gale et Church, 1991a] ont, eux aussi, utilisé leur programme pour aligner 90 millions de mots extraits du compte-rendu des débats du Parlement Canadien. Leur technique est très proche de celle de [Brown *et al.*, 1991]. Le programme se déroule en deux étapes : en premier, les paragraphes sont alignés, puis à l'intérieur de ces paragraphes, les phrases. L'alignement des paragraphes se fait automatiquement, ceux-ci étant généralement bien délimités. Une vérification manuelle garantit le bon déroulement de la deuxième phase du programme. Comme le remarque [Warwick *et al.*, 1992] , il suffit qu'une section manque dans un texte, ou que les deux textes ne suivent pas exactement la même structure hiérarchique pour que l'alignement des phrases soit un échec total. Une fois les paragraphes correctement alignés, un modèle statistique simple prenant en compte la longueur de la phrase en terme de caractères est appliqué. Une probabilité est assignée à chacun des alignements possibles à l'intérieur d'un même paragraphe. C'est l'alignement qui aura la probabilité la plus grande qui sera retenu. Ces probabilités prennent en compte deux paramètres : c, le nombre moyen de caractères dans L_2 pour un caractère dans L_1 et s^2, la variance du nombre de caractères dans L_2 pour un nombre de caractères dans L_1. Ces paramètres sont estimés à partir du corpus pré-aligné : c en comptant le nombre de caractères d'un paragraphe de L_1 et en le divisant par le nombre de caractères correspondants dans le paragraphe de L_2, s^2 en prenant en compte tous les paragraphes alignés. Le programme a été évalué sur le corpus trilingue composé des rapports économiques produits par l'Union des Banques Suisse ; 1 316 alignements anglais-français et anglais-allemand ont été examinés ; 96 % de bons alignements sont obtenus et 6 types d'alignements différents sont pris en compte : 1-0, 1-1, 2-1, 2-2, 3-1, 3-2. Si le programme gère presque parfaitement les alignements 1-1, les alignements complexes, par contre, sont plus difficiles à obtenir ; en particulier, les alignements 1-0, 3-1 et 3-2 ne sont jamais reconnus. Nous présentons dans notre chapitre III, section 3.1.5, la technique de [Simard *et al.*, 1992] qui permet d'améliorer la qualité des alignements obtenus grâce à l'utilisation de *cognates*.

1.5 Ressources lexicales bilingues

L'extraction de ressources lexicales bilingues s'effectue à partir de corpus bilingues alignés phrases à phrases (voir section 1.4). Outils pour les lexicographes ou pour les linguistes étudiant la traduction, ces ressources peuvent aussi servir à la construction d'une banque de connaissances bilingues.

[Gale et Church, 1991b] proposent deux outils pour extraire des mots dans des textes alignés. Ces outils utilisent le corpus des débats du Parlement Canadien. Le premier outil donne toutes les phrases alignées dans lequel un mot et ses traductions envisagées apparaissent. Par exemple, l'utilisateur entre le mot anglais *drug* et les mots français *drogue* et *médicament*, et toutes les phrases alignées comprenant *drug* dans la phrase anglaise et, soit *drogue*, soit *médicament* dans la phrase française correspondante sont extraites du corpus. Ces alignements sont donnés en entrée d'un modèle probabiliste qui se charge de relever des clés contextuelles associées à chaque traduction : pour notre exemple, les mots qui apparaissent dans le contexte de *drug*, dans le sens *médicament*, sont *prices*, *prescription*, *patent*, ..., et dans le sens *drogue*, *abuse*, *paraphernalia*, *illicit*, ... Si l'utilisateur propose ensuite au programme une phrase anglaise contenant le mot *drug*, celui-ci utilise ces clés contextuelles et en déduit la traduction la plus probable. Cette traduction est effectivement la bonne dans 94 % à 95 % des cas.

Le second outil, un peu plus élaboré, propose pour un mot donné par l'utilisateur, une liste des traductions probables rencontrées dans le corpus, accompagnées de leurs fréquences d'apparition. La liste des traductions possibles d'un mot est obtenue en relevant les alignements probables des mots à l'intérieur de phrases appariées de la manière suivante : à l'intérieur de chaque phrase du corpus aligné, toutes les correspondances mot anglais/mot français sont relevées exhaustivement et comptabilisées. Puis à l'aide d'un tableau de contingence (voir chapitre IV - section 4.3.2), une modèle statistique Φ^2, est calculé pour chaque alignement (mot_e, mot_f) rencontré dans le corpus. La valeur du Φ^2 est élevée lorsque les deux mots sont fortement associés, et donc qu'ils peuvent être la traduction l'un de l'autre (Φ^2 est très proche du score d'association présenté précédemment). Chaque mot du corpus possède donc une liste de ses traductions les plus probables établie en fonction de la valeur décroissante du Φ^2. Ce programme est plus performant que celui fondé sur l'appariement des phrases puisqu'il n'a pas besoin de connaître la liste des traductions possibles du mot, parfois difficiles à deviner si elles n'ont pas été recensées dans un dictionnaire, et qu'il permet d'isoler les instances du mot qui n'ont pas de correspondance littérale comme par exemple les cas de pronominalisation. Pour l'alignement de mots, proprement dit, à l'intérieur de phrases, le programme choisit, dans la liste de candidats, le mot de la phrase cible ayant la valeur du Φ^2 la plus forte. L'appariement des mots n'a pas été évalué et les quelques exemples de listes de traductions d'un mot qui sont donnés ne permettent pas véritablement de juger ce programme.

[Gaussier *et al.*, 1992] proposent une méthode générale d'appariement des mots à partir de corpus alignés et en déduisent pour chaque mot du corpus une liste associée de traductions probables. Dans un premier temps, la méthode est identique à celle que [Gale et Church, 1991b] utilisent pour construire leurs listes, à ceci près que la mesure statistique calculée n'est pas le Φ^2 mais le score d'association. Puis ces listes sont modifiées selon le critère suivant : un mot anglais est mis en correspondance avec un mot français si le mot français proposé comme la traduction la plus probable ne possède pas un score d'association plus élevé dans une autre correspondance. Par exemple, les traductions proposées pour le mot anglais *prime* sont entre autres *bureau* (avec un score d'association de 5,6) et *premier* (avec un score d'association de 5,4). Le candidat retenu devrait donc être *bureau*; cependant, avant d'entériner cette décision, on vérifie que *bureau* n'est pas associé à un autre nom anglais avec un score plus important, et effectivement, dans la liste des traductions de *desk*, on trouve *bureau* avec un score d'association de 5,8. Donc, *bureau* a plus de chance d'être la traduction de *desk* que de *prime*, et *bureau* est supprimé de la liste des traductions possibles de *prime*. Cette méthode d'alignement des mots a été évaluée sur 3 000 phrases alignées extraites des débats du Parlement Canadien. Seuls les mots qui ne sont pas des mots grammaticaux sont pris en compte. Pour 65 % des mots, la liste des candidats à la traduction est cohérente et ne contient donc que des traductions possibles; pour 25 % des mots, il n'existe pas de liste de candidats et pour 10 %, les listes contiennent de mauvais candidats.

Les programmes d'appariement fondé sur les mots de [Gale et Church, 1991b] et [Gaussier *et al.*, 1992] n'aboutissent qu'à des correspondances 1-1 ; c'est-à-dire qu'à un mot anglais correspond un mot français. Les appariements de mots plus difficiles tels que les appariements 1-2, 2-2, 1-3, etc., ne sont pas traités, contrairement aux alignements complexes de phrases. Ces alignements complexes sont pourtant essentiels et permettraient d'éviter l'obtention d'alignements incomplets tels que l'appariement du verbe anglais *to ignore* avec le substantif français *fi*, présenté dans [Gale et Church, 1991b]. Si un locuteur français est effectivement capable de retrouver l'expression complète à partir de l'un de ses éléments, comme *faire fi* à partir de *fi*, l'opération se révèle moins triviale pour un locuteur étranger. Les traductions que nous fournissent ces programmes restent approximatives, et nécessitent une vérification manuelle avant d'être intégrées à une base lexicale. Nous verrons plus loin que notre travail sur les noms composés binaires permet d'obtenir des appariements complexes de type 2-2, et représente donc un réel progrès dans la recherche sur l'appariements des mots.

[Klavans et Tzoukermann, 1990] se sont intéressées aux verbes de mouvement anglais et à leur traduction en français. La langue anglaise incorpore mouvement et cause dans une unique entité lexicale alors que le français en utilise plusieurs. Par exemple, le verbe anglais *to march* lorsqu'il est construit avec un objet direct possède un caractère causatif. Cet aspect causatif s'exprime en français par

l'opérateur factitif *faire* introduisant une infinitive:

The soldiers marched past the general
Les soldats ont défilé devant le général

The sergeant marched the soldiers past the general
Le sergent a fait défiler les soldats devant le général

Leur objectif est la construction d'un dictionnaire bilingue électronique contenant les traductions du *Robert & Collins*, et d'autres traductions extraites à l'aide d'un modèle statistique des débats du Parlement Canadien alignés phrases à phrase. L'expérience est conduite de la façon suivante: la liste exhaustive des verbes de mouvement de l'anglais étant fournie au programme, celui-ci extrait du texte anglais les phrases en contenant au moins un. Pour chaque verbe de mouvement, les traductions possibles en français données par le *Robert & Collins* sont comparées avec le texte français correspondant. Si l'une des traductions proposées est effectivement présente dans la phrase française, le programme l'enregistre. À la fin de l'examen de toutes les phrases, le programme annote les traductions du *Robert & Collins* de leur nombre d'occurrences dans le corpus. Les nouvelles traductions, i.e. non proposées par le *Robert & Collins*, sont isolées par un modèle probabiliste non décrit et ajoutées aux précédentes. Ces nouvelles définitions incluent des expressions figées construites à partir d'un verbe de mouvement anglais comme par exemple l'expression anglaise *to dance to (the) tune (of)* traduite en français dans leur programme par *se mettre au diapason* ou *compléter le quatuor*. Il n'est pas précisé comment la traduction de l'expression anglaise est identifiée dans la phrase française et, sachant que l'appariement des mots à l'intérieur d'une phrase n'est pas une tache triviale, il aurait été intéressant que [Klavans et Tzoukermann, 1990] expliquent plus clairement comment elles procèdent. D'autre part, je ne suis vraiment pas convaincue que *se mettre au diapason* et *compléter le quatuor* aient la même signification!

1.6 Traduction automatique

L'idée d'un système de traduction automatique par ordinateur (TAO) fondé sur un modèle statistique a été émise il y a de nombreuses années par [Weaver, 1949]. Celui-ci pensait que l'approche mathématique qui avait si bien réussi pour la cryptologie s'adapterait parfaitement à la TAO. Malheureusement, cette tentative a été un échec et l'approche statistique est tombée aux oubliettes. Il faut dire que les grands noms de la linguistique, pour ne citer que Chomsky, avaient démontré par des critères formels l'inefficacité des statistiques dans ce domaine. L'un de ces critères était l'impossibilité des statistiques à rendre compte des contraintes syntaxiques tel que l'accord en nombre sur des longues

distances ([Chomsky, 1957]). Quarante ans après, devant la "non-réussite" des systèmes de traduction automatique fondés sur la connaissance des langues, et fort de la réussite des méthodes statistiques dans la reconnaissance de la parole ([Dérouault et Merialdo, 1986], etc.), [Brown *et al.*, 1988] relancent l'idée d'un système TAO purement statistique, c'est-à-dire sans connaissances linguistiques. L'environnement a changé : ils ont maintenant à leur disposition d'énormes corpus bilingues alignés, principalement le Hansard, et des calculateurs nettement plus puissants que dans les années 50. Ce sont ces facilités techniques qui d'après [Brown *et al.*, 1988, 1990] faisaient défaut à [Weaver, 1949]. Néanmoins, après six ans d'expérience, leur tentative de réaliser un système TAO purement statistique n'ayant pas abouti, ils reconnaissent que les facilités techniques ne résolvent pas tous les problèmes de la traduction. À ce jour, cette approche a donc encore à faire ses preuves.

Leur expérience sur ce système statistique est décrite par trois articles : [Brown *et al.*, 1988, 1990 et 1992]. Ayant remarqué qu'il existe souvent plusieurs manières de traduire une phrase, les auteurs adoptent le postulat suivant : toutes les phrases d'un texte sont les traductions possibles de n'importe quelle phrase du texte traduit dans une autre langue ; ainsi, chaque phrase S_s d'une langue source a la probabilité $Pr(S_c \mid S_s)$ de se traduire par la phrase S_c, dans un langue cible. Lorsque les deux phrases n'ont aucune chance d'être la traduction l'une de l'autre comme par exemple lorsque S_s = *le matin, je me brosse les dents* et S_c = *President Lincoln was a good layer*, $Pr(S_c \mid S_s)$ est faible. Lorsque S_s = *Le Président Lincoln était un bon avocat* et S_c = *President Lincoln was a good layer*, $Pr(S_c \mid S_s)$ est forte. Le problème de la TAO se résume donc à choisir pour une phrase de la langue source la phrase de la langue cible qui la probabilité maximale. Cette probabilité est donnée par le théorème de Bayes :

$$Pr(S_c \mid S_s) = \frac{Pr(S_c)Pr(S_s \mid S_c)}{Pr(S_s)}$$

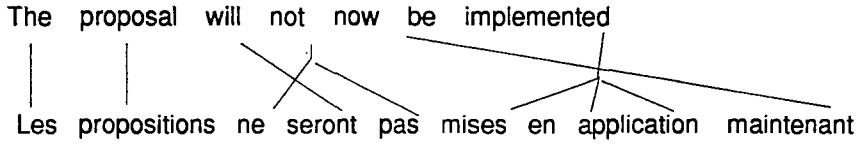
Comme le dénominateur de droite de l'équation ne dépend pas de la phrase cible S_c, il suffit de choisir S_c telle que le produit $Pr(S_c)Pr(S_s \mid S_c)$ soit maximal. Le premier facteur de ce produit représente la probabilité d'avoir dans la langue cible une phrase S_c et le second facteur la probabilité qu'une phrase de la langue cible S_c soit la traduction de S_s. Plus clairement, la **probabilité de traduction** revient à déterminer les mots de la phrase source qui auraient produit les mots de la phrase cible, et la **probabilité du modèle de langue** comment articuler ces mots dans S_c. Donc leur système de TAO a besoin d'une méthode pour calculer les probabilités du modèle de langue, d'une méthode pour calculer les probabilités de traduction, et finalement exige de déterminer parmi les possibles phrases cible S_c celle qui a la probabilité $Pr(S_c)Pr(S_s \mid S_c)$ maximale. Le modèle statistique utilisé pour déterminer les probabilités associées à une langue est le modèle Markovien à trois états ou triclasse. Ce modèle permet d'ordonner

les mots obtenus après le processus de traduction de manière à former une phrase. Les paramètres du modèle sont établis à partir d'un corpus monolingue de la langue source. La probabilité de traduction associée à chaque couple de phrases est calculée en fonction des probabilités de chaque correspondance de mots à l'intérieur des phrases. Pour obtenir les probabilités liées aux couples (mot_s, mot_c), il faut effectuer l'appariement des mots à l'intérieur des phrases alignées; le problème de l'appariement des mots a été présenté dans la partie précédente à propos de l'extraction de ressources lexicales bilingues (section 1.5). Dans [Brown *et al.*, 1988], ils envisagent de découper la phrase source en une série de "locutions fixes"; les locutions fixes comprenant les mots composés, les expressions figées, l'expression de la négation, etc. L'appariement s'effectuerait donc au niveau de ces locutions fixes et non plus au niveau des mots. Néanmoins, devant la difficulté à obtenir une division de la phrase source en locutions d'une manière unique, cette approche est abandonnée et le problème de la reconnaissance de ces locutions fixes est intégré dans [Brown *et al.*, 1990] à celui d'apparier les mots, en prenant en compte les deux facteurs suivants:

- les correspondances de mots ne sont pas tous de type 1-1 (c'est-à-dire un mot de la langue source correspond à un mot de la langue cible); un mot de la langue source peut donner 0, 1, 2, 3 ou plus mots dans la langue cible: c'est la "fertilité". Ainsi, de l'anglais vers le français, *John* étant traduit par *Jean* , la fertilité est égale à 1, alors que lorsque *beat* est traduit par *est battu*, la fertilité est égale à 2.

- l'ordre des mots n'est pas toujours le même dans la langue source et dans la langue cible; un mot se trouvant à une certaine place dans la phrase source ne correspond pas toujours au mot se trouvant à la même place dans la phrase cible: c'est la "distorsion".

Leur programme d'appariement des mots à l'intérieur de phrases alignées prend en compte la probabilité de fertilité, la probabilité de distorsion et la probabilité lexicale qu'un mot_s se traduise par mot_c. Un exemple d'appariement des mots obtenus grâce à ces probabilités est le suivant:



L'appariement des mots étant effectué à l'intérieur des phrases, les correspondances (mot_s, mot_c) sont relevées et produisent un dictionnaire bilingue. Dans [Brown *et al.*, 1988], l'expérience n'étant vieille que d'un an, ils n'obtiennent pas de résultats de traduction mais des résultats intermédiaires pour la création du dictionnaire bilingue et l'arrangement des séquences de mots cibles. Le chiffre

de 79 % de réussite est avancé pour l'arrangement de phrases anglaises de 10 mots au plus ; néanmoins il n'est pas précisé sur combien de phrases tests porte ce pourcentage. Ils sont très confiants en leur approche, même si ils envisagent déjà d'intégrer des informations syntaxiques à leur système. Après trois années d'expérience, [Brown *et al.*, 1990] obtiennent des traductions mais les résultats ne sont pas ceux escomptés : sur 73 phrases françaises de moins de dix mots traduites automatiquement en anglais, 35 sont correctes ce qui donne un score de 48 % de réussite. Ils envisagent donc d'améliorer leur système en apportant les modifications suivantes :

- utilisation de données supplémentaires ; l'expérience menée n'utilisait que 1 % des données bilingues et que 10 % des données monolingues,
- leur modèle statistique de création de dictionnaire bilingue accepte qu'un mot source donne plusieurs mots cibles, mais pas le contraire ; la possibilité que plusieurs mots source donnent un mot cible sera donc intégrée,
- utilisation d'un modèle probabiliste plus performant pour déterminer la manière d'ordonner les mots cible en une phrase,
- utilisation d'un modèle morphologique,
- utilisation éventuelle d'une grammaire obtenue automatiquement à partir de corpus,
- prise en compte de l'agencement dans la phrase de certains mots par rapport à d'autres.

Dans leur dernier article [Brown *et al.*, 1992], la pure approche statistique est déjà abandonnée. De manière à améliorer leur système, ils ont adopté l'architecture des systèmes de traduction automatique de la deuxième génération : analyse de la phrase source produisant une représentation intermédiaire, transfert de cette représentation intermédiaire source en une représentation intermédiaire cible, enfin génération de la phrase cible. La nécessité de cette représentation intermédiaire, couramment appelée structure d'interface dans les systèmes classiques, est motivée par le fait qu'avant de procéder à la traduction des mots, il faut aplanir les différences de structures existant entre les langues. Les opérations linguistiques que proposent [Brown *et al.*, 1992] pour obtenir une structure de phrase commune aux deux langues n'ont aucun caractère général et ne peuvent remplacer une analyse syntaxique. Pour obtenir cette représentation intermédiaire, ils utilisent un analyseur morphologique et un programme d'assignation d'étiquettes grammaticales. L'évaluation du système est faite sur la traduction en anglais de 100 phrases françaises de 10 mots au plus : avec le système précédent, ils obtenaient 39 % de réussite[2]et sur ce nouveau système 60 %. Le fait d'introduire des

[2]Score encore inférieur à celui précédemment donné de 48 % de réussite

éléments linguistiques améliore donc considérablement le système. Celui-ci pourrait sans doute être encore plus performant si une réelle analyse (resp. génération) syntaxico-sémantique était menée. Dans ce cas, le système n'aurait plus rien de "statistique" si ce n'est le module de transfert construit automatiquement à partir d'un corpus bilingue aligné phrases à phrases. Les résultats de l'expérience de l'équipe de Brown semblent donc confirmer le fait que les statistiques sont utiles pour les ressources lexicales mais un peu limitées pour la traduction.

Chapitre 2

Aspects linguistiques de la terminologie

Nous avons fait le choix d'utiliser les apports de la linguistique pour développer notre système d'extraction de terminologie. Une banque terminologique est constituée pour l'essentiel de noms composés. C'est sur ce sujet que se sont penchés de nombreux linguistes. Un tour d'horizon de leurs travaux montre que ceux-ci convergent tous vers une tentative de définition de la composition. Toutefois, on constate qu'il y a pratiquement autant de définitions que d'écoles linguistiques. Ainsi, si ces travaux nous sont très profitables, en particulier pour l'établissement d'un classement des noms composés en fonction de leurs structures morphosyntaxiques, ils ne nous fournissent pas de recettes magiques pour la reconnaissance. Il n'existe donc pas une définition unique du mot composé mais un certain nombre de propriétés communes. Notre propos n'est pas ici d'ajouter une ligne à la liste des définitions existantes, mais plutôt de faire le point sur les données linguistiques utilisables. Ayant de plus fait le choix d'utiliser des corpus relevant d'un domaine technique particulier, nos conclusions ne seront valides que pour ce type de textes. En cela, notre perspective est légèrement différente des travaux qui nous servent de référence puisque ceux-ci sont rarement orientés vers l'extraction automatique à partir de corpus. Après avoir présenté quelques travaux de linguistes sur la composition, nous appliquerons les propriétés ainsi rassemblées aux corpus du domaine des télécommunications.

2.1 Les diverses définitions proposées

Notre objectif étant l'extraction automatique à partir d'un corpus de noms composés incluant les termes d'un domaine technique, nous donnerons dans cette section un rappel des principales définitions de la composition et plus particulièrement de la composition nominale et nous présenterons quelques travaux linguistiques effectués dans ce domaine. Nous verrons qu'il y a autant

de définitions différentes de la composition que d'écoles linguistiques. Nous ne prétendons pas passer en revue de façon exhaustive les différents travaux sur la composition. Après avoir rappelé quelques définitions sur la composition en général, nous nous concentrerons sur la composition nominale, en examinant les travaux sur la synapsie de [Benveniste, 66], puis ceux sur la co-occurrence lexicale restreinte de [Mel'čuk *et al.*, 1984] et enfin sur les travaux du LADL.

2.1.1 Composition

Qu'est-ce qu'un "mot composé" ? La réponse à cette question est toujours en suspens et les nombreux linguistes qui se sont plongés dans le monde de la composition n'ont pas véritablement réussi ni à dégager des critères d'identification opérationnels, ni à s'accorder sur une même définition. Chaque linguiste utilise sa propre terminologie pour définir les mots composés et propose ses propres critères. Ainsi, [Grévisse et Goosse, 1986 :261] différencient "composé" et "locution" :

> "Un composé est une unité lexicale formée à partir de deux mots existants qui sont soit coagulés soit joints par un trait d'union." (...) "Une locution est une suite de mots qui sont séparés par des blancs dans l'écriture et qui forme pourtant une unité lexicale. Selon la nature des mots simples avec lesquels la locution peut commuter, on parlera de locution nominale (*chemin de fer*), adjectivale (*comme il faut*), pronominale (*quelque chose*), verbale (*avoir lieu*), adverbiale (*tout à fait*), prépositionnelle (*quant à*), conjonctive (*bien que*), interjective (ou locution-phrase : *Par exemple !*)."

La locution est identifiée principalement par le critère sémantique de l'unité d'image :

> " Le sentiment d'unité est fondé :
>
> 1. sur des critères paradigmatiques : une locution peut commuter avec un mot simple : *Il voyage en chemin de fer / ... en voiture*,
> 2. sur des critères syntaxiques, c'est-à-dire le fait que la locution ne respecte pas les règles ordinaires de la syntaxe,
> 3. sur des critères sémantiques, c'est-à-dire sur le fait que le sens de la locution n'équivaut pas à l'addition des sens des constituants : *bande dessinée*, ou le fait que la locution représente une réalité unique et qu'elle équivaut à un mot simple : *avoir lieu = arriver*,
> 4. sur le fait qu'un locuteur moyen est incapable d'analyser les composantes : *Il y a belle lurette.*"

Comme le font remarquer les auteurs, il est rare que ces critères soient réunis ; ceci simplement par le fait que les critères 2. et 3. concernent une minorité de locutions.

Reste le critère 1., aussi dénommé le critère de commutation ([Martinet, 1960], [Picoche, 1977]), et le critère 4. qui regroupe en fait deux critères: le critère de la non-compositionalité des sens et le critère de l'unité d'image. Le critère de non-compositionalité a souvent été confondu avec celui de l'unité d'image. S'il est exact que le premier s'applique à des unités dont le figement est souvent incontestable (*chemin de fer*), le deuxième, interprété dans le sens à signifié unique, s'applique à un nombre beaucoup plus important d'unités. Des locutions comme *antenne de réception* ou *moulin à vent* qui ne répondent pas au critère de non-compositionalité possèdent un signifié unique et constant. Reste à savoir si notre interprétation de ce critère de l'unité d'image correspond à celle de ces grammairiens.
Les travaux effectués au Laboratoire d'Automatique Documentaire et Linguistique de Paris 7 (LADL) et plus particulièrement ceux effectués dans le cadre du projet de recherche sur les formes composées du français, mené conjointement avec le Laboratoire de Linguistique Informatique de Paris 13 depuis 1982, diffèrent des autres études effectuées sur ce sujet, le cadre théorique dans lequel ils se placent étant celui du lexique-grammaire. Dans ce cadre, qui est fondé sur la théorie transformationnelle de [Harris, 1976], l'unité minimale de sens est la phrase et non pas le mot et les formes composés sont étudiées à travers leur insertion dans les phrases simples observables en fonction de leurs propriétés syntaxiques. Les formes composées étudiées sont des chaînes de caractères qui incluent au moins deux mots simples et un séparateur. [M. Gross, 1989:40] propose une définition générale:

> " Lorsque deux éléments d'une construction sont fixes l'un par rapport à l'autre, alors la construction est figée."

Le LADL a donc entrepris une étude systématique de toutes les formes composées du français; citons par exemple les travaux de [M. Gross, 1986] sur les adjectifs et les adverbes composés, [Piot, 1988] sur les conjonctions composées, [Danlos, 1980] sur certains couples verbe-complément, [M. Gross, 1988] sur les verbes composés et les phrases idiomatiques, etc. Les études sur les noms composés seront examinées plus en détail dans la section suivante consacrée à la composition nominale.
Pour l'école traditionaliste comme pour le LADL, les composés sont des mots séparés par des blancs dans l'écriture. L'objectif du LADL étant l'analyse automatique, les composés soudés ne sont pas considérés comme des mots composés et il est exact que ceux-ci ne présentent pas, en général, de difficulté pour un analyseur lexical; mais, les mots composés comportant un trait d'union ou une apostrophe sont aussi facilement identifiables. Même si le trait d'union est facultatif pour certains types de composés, la mise en correspondance des graphies avec et sans trait d'union peut s'opérer assez aisément. D'autre part, la distinction de [Grévisse et Goosse, 1986] entre composé et locution, calculée uniquement sur la présence ou l'absence d'un "blanc" entre les unités, sensée refléter un comportement sémantique est loin d'être convaincante; ce n'est pas la présence ou

l'absence de trait d'union qui atteste d'une lexicalisation avancée ou en cours : *pomme de terre* a un caractère aussi figé que *rond-point.*
L'école fonctionnaliste avec [Martinet, 1985] se détache de la graphie. Les "synthèmes" sont soit des mots simples comme *gaieté*, soit plusieurs mots simples comme *jeune fille.*

> "On parle de composition lorsque des monèmes libérables sont accolés ou reliés par quelques éléments de liaison.
> On appelle synthème un signe linguistique que la commutation révèle comme résultant de la combinaison de plusieurs signes minima, mais qui se comporte vis-à-vis des autres monèmes de la chaîne comme un morphème unique."

Les synthèmes peuvent résulter de figements et avoir la même forme qu'une succession de monèmes libre ; il est donc important de différencier le synthème du syntagme, et les critères syntaxiques suivants sont avancés :

- "le synthème fonctionne comme un monème unique" et "il a toutes les compatibilités des monèmes d'une certaine classe" : *chemin de fer* peut commuter dans une phrase avec *avion* ou *voiture*,

- "aucune de ses parties constitutives n'entre dans des rapports particuliers avec un monème qui ne fait pas partie du synthème" ; un synthème n'accepte pas de modification sur l'un des ses éléments constitutifs ; s'il y a modification, c'est sur le seul synthème :

  un vieux chemin de fer
  **un chemin creux de fer*
  **un chemin de fer forgé*

L'auteur reconnaît que ce dernier critère est à manier avec précaution le caractère figé d'un synthème pouvant varier suivant les locuteurs : *la corne de l'Afrique/la corne orientale de l'Afrique.* Une attitude de "doute" est donc de rigueur et les cas litigieux sont à enregistrer. Cette attitude qui ne tient pas compte de la graphie du composé nous paraît beaucoup plus juste ; néanmoins, regrouper dans la même classe des synthèmes, des unités telles que les dérivés comme *mélange*, *amoral* et des unités résultantes de mots préexistants comme *chemin de fer*, nous paraît quelque peu excessif. Ce tour d'horizon pourrait continuer avec les "unités lexicalisées" de [Galisson et Coste, 1976], la "lexie" de [Pottier, 1985], etc. qui donnent eux aussi leurs propres définitions et leurs propres critères d'identification des mots composés. Le but de cette présentation n'est pas d'exposer toutes les définitions existantes mais de montrer combien la notion de mot composé varie d'un linguiste à un autre, tant du point de vue des unités en jeu, que des critères qui les caractérisent. Le terme de "mot composé" ne doit donc être considéré que comme une commodité de langage ; il regroupe des unités lexicales simples ou

complexes appartenant à toutes les catégories grammaticales et, pour les unités lexicales complexes, présentant des degrés de figement variables.

2.1.2 Composition nominale

Sans vouloir entrer dans la polémique sur la composition, nous nous concentrerons dans cette section sur les noms composés de plusieurs unités lexicales; ceux-ci étant au moins quatre fois plus nombreux que les noms simples d'après une estimation de [G. Gross *et al.*, 1986]. Même si, dans le pire des cas, tous les noms simples se révélaient être des noms composés, les noms composés complexes, i.e. regroupant plusieurs unités lexicales séparées par des blancs dans l'écriture, seraient donc toujours aussi nombreux et, il paraît donc légitime de s'intéresser à eux plus particulièrement. D'autre part, les noms composés de plusieurs unités lexicales séparées par des blancs dans l'écriture, posent un réel problème lors de l'analyse automatique à la différence des noms simples. Le comportement syntaxique du nom composé complexe est identique à celui d'un nom simple et il peut apparaître aussi bien en position sujet, qu'en position objet direct, objet indirect, etc. Le problème principal reste donc de les différencier des syntagmes nominaux libres et d'essayer de les caractériser par des critères syntaxiques ou sémantiques.

2.1.2.1 Synapsie

Certains noms composés complexes ont été baptisés "synapsies" par [Benveniste, 1966:172]:

> "Une synapsie consiste en un groupe entier de lexèmes, reliés par divers procédés, et formant une désignation constante et spécifique. On trouve le noyau initial dans des exemples déjà anciens comme: *pomme de terre*, *robe de chambre*, *clair de lune*, *plat à barbe*. Le fait nouveau et important est que ce type de composition prend aujourd'hui une extension considérable et qu'il est appelé à une productivité indéfinie: il sera la formation de base dans les nomenclatures techniques."

La synapsie est caractérisée par les traits suivants:

1. la nature syntaxique (non morphologique) de la liaison entre les membres;
2. l'emploi de joncteurs à cet effet, notamment *de* et *à*;
3. l'ordre déterminé + déterminant des membres;
4. leur forme lexicale pleine, et le choix libre de tout substantif ou adjectif;
5. l'absence d'article devant le déterminant;

6. la possibilité d'expansion pour l'un ou l'autre membre;
7. le caractère unique et constant du signifié.

Cette définition est intéressante car c'est la première qui prend en compte l'extrême productivité des formes N_1 PREP N_2 et N ADJ dans les nomenclatures techniques. Cette productivité sera confirmée par les travaux du LADL sur le dictionnaire DELAC des composés du français (section 2.1.2.3). Les critères donnés sont à la fois d'ordre sémantique et d'ordre syntaxique.

2.1.2.2 Co-occurrence lexicale restreinte (CRL)

Le terme anglais *collocation* semble avoir été introduit par R. Firth (1951) qui fut membre de l'École Contextualiste anglaise. Ce terme a été repris par [Hausman, 1985] et fait référence aux contraintes de co-occurrence de deux unités lexicales: les deux unités ne sont pas associées totalement librement, mais l'une d'elle détermine sémantiquement l'apparition de l'autre. Par exemple, dans *beurre rance*, le substantif *beurre* ne peut pas commuter avec d'autres éléments lexicaux de la famille des produits laitiers **fromage rance*; de même, dans *désir ardent*, l'adjectif ne peut pas être remplacé par un synonyme **désir embrasé*, **désir enflammé*, **désir incandescent*.
Le terme français équivalent, utilisé dans [Mel'čuk *et al.*, 1984:4], est "co-occurrence lexicale restreinte (CRL)". La définition proposée n'équivaut pas à la définition donnée précédemment par [Hausman, 1985] et est une re-formulation du critère de non-compositionnalité:

> "Par co-occurrence lexicale, on entend la capacité des lexèmes de se combiner en syntagmes pour exprimer un sens donné. On parle de co-occurrence lexicale restreinte si un lexème *A* signifiant 'A' et un lexème *B* signifiant 'B' ne peuvent pas se combiner pour exprimer le sens composé 'A + B', cela n'étant pas interdit par la syntaxe."

Néanmoins, la définition de [Hausman, 1985] caractérise beaucoup mieux les diverses catégories de CRL donnée par [Mel'čuk *et al.*, 1984] que sa propre définition: *essaim d'abeilles* a un sens totalement compositionnel et est pourtant donné comme exemple de CRL. Mais comme nous allons le voir, nous n'en sommes pas à une contradiction près. Dans le Dictionnaire Explicatif et Combinatoire du Français Contemporain, les CRL sont ensuite décrites en terme de fonctions lexicales: une fonction lexicale prend comme argument un lexème (mot associé à un sens) et lui associe un ou plusieurs lexèmes exprimant le sens indiqué par la fonction lexicale. Un exemple est la fonction **Culm** signifiant "culmination de ..." et qui appliquée au substantif *joie* donne: **Culm**(*joie*) = *comble* [*de la joie*].
Les 37 fonctions lexicales décrites sont loin de former un groupe homogène et décrivent aussi bien des relations paradigmatiques que des relations syntagma-

tiques. Les fonctions lexicales caractérisant une relation paradigmatique sont par exemple:

Syn (synonyme): **Syn**(*espoir*) = *espérance*
Anti (antonyme): **Anti**(*respect*) = *irrespect*

Ces relations paradigmatiques ne caractérisent aucunement des co-occurrences lexicales. Nous avons tenté de classer les fonctions lexicales caractérisant des relations syntagmatiques à l'intérieur d'un groupe nominal en deux types: celles qui mettent en jeu deux noms et celles qui mettent en jeu un nom et un adjectif:

1. F(N) = N caractérisant N PREP N

   Sing un quantum < une portion > régulière(e) de ...
   Sing(*riz*) = *grain* [*de riz*]

   Mult ensemble régulier de ...
   Mult(*abeille*) = *essaim* [*d' abeilles*]

   Magn très, intensément, à un degré élevé
   Magn(*mémoire*) = [*mémoire d'*] *éléphant*

2. F(N) = Adj caractérisant N ADJ

   Magn très, intensément, à un degré élevé
   Magn(*mémoire*) = *prodigieuse*

   Ver tel qu'il doit être, correct
   Ver(*souhait*) = *légitime*

   Bon bon - expression qu'on emploie comme une louange standard codifiée par la langue
   Bon(*conseil*) = *précieux*

   Pos évaluation positive
   Pos(*opinion*) = *favorable*

Nous avons écarté certaines fonctions lexicales comme:

Culm culmination de ...
Culm(*gloire*) = *apogée* [*de la gloire*]
Culm(*colère*) = *paroxysme* [*de la colère*]
Culm(*joie*) = *comble* [*de la joie*]

qui caractérisent des groupes nominaux qui ont peu d'emplois en dehors de leurs constructions verbales figées de type *être* PREP X (*Marie est à l'apogée de sa gloire*) et de ses variantes aspectuelles (*Marie arrive à l'apogée de sa gloire*). D'autre part, les éléments obtenus en appliquant les fonctions **Sing** et **Mult** ressemblent fort aux classificateurs *partie d'un tout* et *groupement*

d'éléments. Le fait que les auteurs s'en défendent est d'autant plus bizarre que le classificateur ne se combine qu'avec une classe réduite d'unités lexicales, et caractérise parfaitement une contrainte de sélection. Il règne une telle confusion dans les travaux de [Mel'čuk *et al.*, 1984] que malgré la richesse des informations enregistrées sous chaque entrée lexicale, elles sont difficilement utilisables dans un traitement automatique. Les fonctions lexicales ne répondent à aucun critère syntaxique et relèvent seulement d'intuitions sémantiques.

2.1.2.3 Les travaux du LADL

Rappelons que les travaux du LADL ont pour objectif de décrire systématiquement la langue française en recensant non seulement les formes lexicales mais aussi les structures syntaxiques élémentaires dans lesquelles elles peuvent apparaître. La description des noms composés passe par leur recensement, et pour ceux qui sont prédicatifs, leur insertion dans une phrase à verbe support. Face aux divergences de définitions et de critères qui règnent au sein des linguistes, et afin d'éviter de donner des définitions trop riches, [G. Gross, 1988] décide d'obtenir une classification plus fine des noms composés en les classant suivant les catégories syntaxiques de leur constituants. Le travail d'analyse s'effectue alors dans le cadre de chaque classe ainsi obtenue. Sont ainsi présentés 26 types de composés nominaux, mais l'entière typologie réalisée par [Mathieu-Colas, 1988] contient plus de 500 types élémentaires regroupés dans 17 classes, auxquelles s'ajoutent 8 classes complémentaires qui décrivent des composés complexes comprenant plus de deux noms ou adjectifs. Quelques exemples de ces types binaires sont :

1. **N de N** : *pomme de terre, coup de force*
2. **N Adj** : *cordon bleu, cercle vicieux*
3. **Adj N** : *blanc-bec, grand ensemble*
4. **NN** : *café-filtre, cheval-vapeur*
5. **N Participe présent** : *poisson volant, chat huant*
6. **N par N** : *preuve par neuf*
7. **N en N** : *arc-en-ciel, entrée en fonction*
8. **N à N** : *pelle à gâteau*
9. **N Prep N** : *sculpture sur bois*
10. **V N** : *gratte-papier*
11. **V Prep Inf** : *pince-sans-rire*

D'après [G. Gross, 1988], parmi tous ces types, seuls ceux qui partagent la même structure qu'un groupe nominal simple posent problème et sont principalement les classes 1, 2, 3, 5. Cette position est peut-être un peu rapide : l'étude des composés N N menée par [Noailly, 1990] montre que l'emploi épithète du substantif est désormais un véritable phénomène syntaxique ; phénomène qui a été découvert justement par l'étude des composés correspondants. De plus, les classes 6 et 7 ne sont pas seulement caractéristiques des noms composés : l'absence de déterminant est aussi très répandue dans les groupes nominaux libres de même structure. La méthodologie est ensuite la même pour chaque classe : les propriétés syntaxiques qui caractérisent la structure libre associée sont énumérées et serviront de critères pour évaluer le degré de figement du groupe nominal. En effet, pour [G. Gross *et al.*, 1987],

> " ... la majorité des noms composés sont des groupes nominaux en voie de lexicalisation qui attestent souvent d'un paradigme sur l'un ou l'autre des constituants ..."

Les principaux travaux linguistiques concernent les N ADJ et les N N dans [G. Gross *et al.*, 1986 et 1987], et les N *à* N et N *à* V par [Poncet-Montange, 1991]. Les noms composés ainsi recensés sont intégrés au dictionnaire électronique morphologique des mots composés, le dictionnaire DELAC, construit par [Silberztein, 1989]. Toute entrée du DELAC est associée à une catégorie grammaticale et à un type élémentaire. L'entrée du nom composé précise de plus son genre et sa flexion. Ce code flexionnel associé permettra de le fléchir et d'obtenir ainsi la liste DELACF des formes pouvant apparaître dans les textes. Le dictionnaire DELAC contient à ce jour 90 000 noms composés. La reconnaissance morphologique des noms composés dans des textes a été traitée par [Silberztein, 1989 et 1993]. Cette analyse lexicale signale dans un texte la présence de séquence de mots susceptibles d'être des noms composés en les encadrant à l'aide de crochets indicés. Cette identification est faite grâce au dictionnaire DELACF. Lorsque le nom composé a une structure identique à celle d'un groupe nominal libre, aucune ambiguïté n'est levée ; ce sera le rôle de l'analyse syntaxique ultérieure. [Jacquemin, 1991] s'est concentré sur l'analyse des noms composés ayant subi des "transformations", telles que les modifications adjectivale ou adverbiale, la coordination, la nominalisation et l'adjectivation. Ces transformations s'appliquent principalement aux noms composés représentatifs d'un domaine technique. L'identification s'effectue par l'intermédiaire d'un dictionnaire des mots simples. Une série de règles lexicales présentes dans l'entrée du nom simple permettent de lui rattacher les noms composés associés. Par exemple le nom composé *homme d'état* est rattaché à l'entrée lexicale du nom *état*. Ce rattachement est réalisé par un algorithme approché.
Revenons aux propriétés syntaxiques énoncées par [G. Gross, 1988] qui caractérisent la relation entre un adjectif et un nom dans le cadre du groupe nominal

et dont la non-observation permet de calculer le degré de figement du nom composé de type N ADJ. Ces propriétés sont d'ordre soit morphosyntaxique, soit sémantique :

Critères morphosyntaxiques

1. "variation en nombre"
   Un groupe nominal libre accepte généralement les flexions singulier et pluriel. L'absence de l'une ou l'autre de ces flexions est un indice de figement.
2. "adjonction d'un adverbe" et "adjonction d'un autre adjectif"
   Ces propriétés précisent la possible modification du groupe nominal soit par insertion d'un adverbe, soit par coordination avec un autre adjectif.
3. "la nominalisation" et "adjectifs et compléments de noms"
   Les nominalisations d'adjectifs se font par l'intermédiaire de verbes supports ([Meunier, 1981]) et ne peuvent caractériser que les adjectifs prédicatifs :

   un gentil garçon
   ce garçon est gentil
   ce garçon a une certaine gentillesse
   la gentillesse de ce garçon

   L'adjectif relationnel peut être reconnu par l'équivalence ADJ = *de* N :
   une faute (grammaticale + de grammaire)

Critères sémantiques

1. "la règle d'identité" associée au test :

   un N ADJ est un N

   permet d'émettre un jugement sur le caractère figé ou non du nom composé. Les noms composés qui n'acceptent pas ce test répondent au critère de non-compositionalité et sont totalement figés. À l'inverse, les noms composés qui acceptent ce test sont liés par une relation d'hyperonymie au nom simple tête de la structure composée.
2. les règles de "rupture paragdigmatique" et "figement du premier terme" caractérisent les co-occurrences lexicales restreintes (CRL) (section 2.1.2.2) d'une manière beaucoup plus rigoureuse que celle de [Mel'čuk *et al.*, 1984]. À ces deux propriétés s'ajoute la propriété ci-dessous permettant d'isoler une certaine catégorie de CRL :

3. "la prédicativité"
   Même si certains adjectifs qualificatifs ne peuvent pas être employés comme attribut, la non-acceptabilité du test :

   Ce N est ADJ

   permet d'isoler des noms composés où l'un des éléments (ou les deux) possède un caractère métaphorique. Ces noms composés au sens semi-compositionnel appartiennent à la classe des CRL.

[Poncet-Montange, 1991] a étudié les noms composés de type N *à* N et N *à* V en vue de leur intégration dans le DELAC ; le genre et le nombre du composé, son comportement flexionnel, le type de substantif, les variantes elliptiques ou morphosyntaxiques attestées et le domaine d'emploi ont été indiqués pour chacune des 5 000 formes étudiées. Puis, un classement des noms composés concrets a été adopté en fonction des classes sémantiques d'objets telles que les classes d'instruments, de vêtements, d'aliments, de contenants, etc. Le constat est le suivant : la structure N *à* N est productive.

2.1.3 Conclusion

Les travaux du LADL ont permis de faire un grand bond en avant dans la connaissance des composés du français. L'étude systématique menée sur les différents types de composés a révélé le rôle important de ces unités qui ne peuvent plus maintenant être ignorées dans un traitement automatique sérieux. Le seul problème reste la solution retenue pour le traitement des composés : le recensement. Étant donnée la productivité des structures nominales de type N PREP N et N ADJ, le recensement rencontre de lui-même ses limites. Comme le fait remarquer [Poncet-Montange, 1991:158] pour les N *à* N, qui sont pourtant beaucoup moins productifs que les N ADJ ou les N *de* N :

> "Si l'on considére uniquement les groupes nominaux présentant une relation de **tout** à **partie**, leur enregistrement nécessiterait pour chaque nom d'objet d'un inventaire des noms des parties constituantes. D'autre part, nous avons observé que les compléments prépositionnels sont théoriquement coordonnables ou simplement apposables. À supposer que l'on dispose des données nécessaires, un listage se solderait rapidement par une explosion combinatoire."

La solution réside peut-être dans un traitement différent pour les noms composés les plus figés qui seraient reconnus grâce à un dictionnaire et les plus productifs par des règles de grammaire identiques à celles qui traitent les groupes nominaux libres, c'est-à-dire en analysant la nature du lien entre la "tête" du nom composé et le reste de la structure, soit par des règles syntaxiques, soit par des règles sémantiques : par exemple, dans une structure N_1 PREP N_2, si le nom de tête

N_1 est un nom prédicatif, la séquence PREP N_2 peut être analysée en terme d'argument ou de modifieur de N_1. Même en admettant que cette distinction est possible, et elle est loin de l'être, il reste que, dans la majorité des cas, les règles qui analysent la structure interne du nom composé pourront tout aussi bien s'appliquer à n'importe quel groupe nominal libre. Si de plus, l'analyse de la structure interne du nom composé s'inscrit dans le cadre plus général de l'analyse de la phrase, ces règles de grammaire introduisent beaucoup trop d'ambiguïtés, comme nous allons le montrer.
L'attachement prépositionnel en français, mais aussi dans les autres langues romanes, soulève beaucoup de problèmes en analyse automatique; dans une phrase comme:

Max a commandé une glace à la fraise à la serveuse

sans traitement sémantique indiquant que *serveuse* est un être humain au contraire de *fraise*, les attachements prépositionnels possibles sont:

1. (*une glace à la fraise à la serveuse*)$_1$: objet direct du verbe *commander* avec *à la fraise* et *à la serveuse* tous les deux modifieurs de *glace* dans le sens "*une glace à la fraise à l'italienne*" ;

2. (*une glace à la fraise*)$_1$ objet direct et (*à la serveuse*)$_2$ objet indirect du verbe *commander*;

3. (*une glace à la fraise*)$_1$ objet direct et (*à la serveuse*) complément circonstanciel du verbe *commander* comme dans : "*commander une glace à la fraise au repas de midi*" ;

4. (*une glace*)$_1$ objet direct et (*à la fraise à la serveuse*)$_2$ objet indirect du verbe *commander* comme dans : "*commander une glace à la serveuse aux lunettes*" ;

5. (*une glace*)$_1$ objet direct et (*à la fraise à la serveuse*) complément circonstanciel du verbe *commander* comme dans : "*commander une glace à la terrasse à la vigne vierge*".

Si la reconnaissance des noms composés est laissée au soin de la grammaire, les ambiguïtés sur le statut composé ou non d'un groupe nominal se rajouteront aux ambiguïtés des attachements prépositionnels. Dans ce cas, aux analyses obtenues pour la phrase précédente, il faudra rajouter celles où *glace à la vanille* est éventuellement un nom composé, celles où *fraise à la serveuse* est éventuellement un nom composé et celles où les deux sont des noms composés. Maintenant, si les règles de construction des noms composés utilisent les mêmes règles syntaxiques que celles des groupes nominaux libres - c'est-à-dire que l'on veuille statuer sur le rôle syntaxique du groupe prépositionnel à l'intérieur du nom composé comme à l'intérieur d'un groupe nominal libre -, le problème de l'attachement

prépositionnel se rencontrera non seulement au niveau des structures syntaxiques mais aussi au niveau de la structure interne du nom composé. Prenons quelques exemples où les groupes nominaux sont analysés comme des noms composés potentiels :

- pour *glace à la fraise*, *glace* étant un nom concret, *à la fraise* est analysé comme modifieur et il n'y a qu'une structure interne possible,

- pour *amendement des députés*, *amendement* étant un nom prédicatif, *des députés* sera analysé soit comme argument soit comme modifieur de *amendement* puisque seul un traitement sémantique est capable de lever de telles ambiguïtés. Deux structures internes sont donc possibles, trois si on considère que *députés* peut être soit sujet, soit objet d'*amendement.*

Ce traitement paraît donc particulièrement coûteux puisqu'il introduit des ambiguïtés non présentes dans la phrase : si, en effet, *glace à la fraise* peut effectivement être considéré comme un nom composé, il n'y a aucune ambiguïté pour *fraise à la serveuse* qui n'en est pas un. La présence des noms composés dans un dictionnaire simplifie le travail de l'analyseur, réduit considérablement le nombre des ambiguïtés et permet d'obtenir des analyses plus fiables et plus rigoureuses. Reste le problème d'avoir *glace à la fraise* dans le dictionnaire et pas simplement *glace à la vanille.* C'est en fait ce problème que nous nous proposons de résoudre : obtenir une liste exhaustive des noms composés du domaine d'étude de manière à pouvoir les intégrer dans un dictionnaire électronique. Cette liste des noms composés incluant les termes techniques du domaine sera créée automatiquement à partir d'un corpus étiqueté et lemmatisé.

2.2 Typologie, composition et modification des noms composés terminologiques du domaine des télécommunications

Dans cette partie, nous allons mener une étude linguistique sur les formes composées nominales du domaine des télécommunications. Bien qu'il existe déjà une typologie générale des noms composés du français ([Mathieu-Colas, 1988]), il nous semble important de préciser quels types élémentaires de noms composés sont effectivement présents dans ce domaine technique. Ensuite, nous nous reverrons les différents critères d'identification donnés par les linguistes et présentés dans la partie précédente, et nous nous poserons les questions suivantes : existe-t-il réellement des restrictions morphosyntaxiques qui permettent d'isoler les mots composés, telles que l'absence d'une flexion ou la non-présence d'un déterminant, et ce pour quels types élémentaires? Les modifications qui affectent les noms composés sont-elles aussi nombreuses que le prétend [Jacquemin, 1991] ou au contraire la non-modification est-elle une propriété comme le proclame [Martinet, 1985] ? Les noms composés ont-ils la capacité de se surcomposer ? Le travail se présente en quatre parties : la première consacrée à la typologie des types élémentaires, la deuxième à la surcomposition, à la modification, et à la coordination, la troisième aux variantes et la dernière aux difficultés que nous appréhendons avec un traitement statistique. De manière à confirmer à quel point la composition est un phénomène aléatoire d'une langue à une autre, nous donnerons la traduction anglaise correspondante pour chaque exemple de nom composé. Cette étude des noms composés est donc accompagnée en quelque sorte d'une étude contrastive avec l'anglais. Le domaine étudié est celui des télécommunications et cette étude linguistique a été effectuée principalement sur notre premier corpus : le "manuel des télécommunications par satellite" noté MTS et présenté ci-dessous.

2.2.1 Classification des types élémentaires

Le "manuel des télécommunications par satellite" est disponible en français et en anglais. La taille du corpus français est de 200 000 mots. La méthodologie est la suivante : nous avons examiné le corpus et extrait des suites de mots susceptibles d'apparaître dans des positions syntaxiques variées : sujet, objet direct, objet indirect, etc., et qui appartiennent à l'un des types élémentaires décrits par [Mathieu-Colas, 1988], puis nous avons recherché la traduction de ces suites de mots dans le corpus anglais. Les indications graphiques (trait d'union, lettres capitales, etc) ou morphosyntaxiques (nombre de flexions réduites, absence de déterminant, etc.), nous ont été particulièrement utiles pour déceler les termes du domaine des télécommunications. Nous avons tiré profit de la traduction anglaise correspondante pour certains types de noms composés ; en effet, la plupart des

termes de structure N_1 PREP N_2 en français se traduisent par des termes de structure N_2 N_1 structure morphosyntaxique caractéristique du *compound*; en revanche, pour les termes ADJ N et N ADJ, la traduction n'apporte aucune indication sur le fait qu'un groupe nominal est un terme ou non puisqu'en anglais, les modifieurs du nom apparaissent toujours devant lui. Les critères graphiques et morphosyntaxiques ont été utilisés conjointement avec les critères sémantiques de non-compositionalité et de référé unique. Ces critères sémantiques sont néanmoins difficiles à appliquer à un domaine dans lequel nous ne sommes pas experte. Nous avons donc interprété le critère de référé unique par celui de traduction unique. Un nom composé français répondant à ce critère sera toujours traduit de la même manière, le plus souvent par un nom composé ou par un nom simple anglais. Le domaine des télécommunications étant un domaine technique, les noms composés répondent en partie à la définition de la synapsie de [Benveniste, 66] et sont les termes caractéristiques du domaine. Dans cette typologie, les contraintes morphologiques relevées, comme les absences de flexions, sont valables uniquement pour notre corpus. Le genre des noms composés est généralement le genre du premier nom qui apparaît dans la structure, à quelques exceptions près. Le genre du nom composé est indiqué pour chaque exemple sous la forme suivante: *(m)* si le nom composé est masculin, *(f)* si féminin. Les noms composés sont classés suivant leur longueur. Pour compter la longueur du composé, seuls sont pris en compte les mots non grammaticaux tels que les noms, les adjectifs, les verbes et les adverbes séparés par des blancs. Ni les déterminants, ni les prépositions ne sont pris en compte dans le calcul de la longueur (à l'exception des prépositions apparaissant dans la structure PREP N). Nous appellerons ces mots non grammaticaux, les unités lexicales pleines.

2.2.1.1 Noms composés de longueur 1

Nous avons adopté la possibilité d'avoir des noms composés de longueur 1. Ceux-ci sont formés à partir de plusieurs unités lexicales (généralement deux) reliées entre elles soit par un trait d'union, soit par une apostrophe. Le premier élément en jeu est soit une unité lexicale non autonome comme un préfixe, soit une unité lexicale autonome comme un verbe, un adjectif, un adverbe, un nom ou une préposition. Le deuxième élément est principalement un nom. Pour certaines structures, le trait d'union est optionnel. Pour les structures construites avec un préfixe, le trait d'union s'efface parfois et les deux éléments se retrouvent agglomérés. Pour certains préfixes, tel que *auto*, seule la forme agglomérée existe. Ces formes agglomérées ont donc été intégrées à cette classe. Pour d'autres structures où la première unité lexicale n'est pas un préfixe, le trait d'union laisse place à un blanc; le problème de ces variantes orthographiques de longueur 2 est abordé dans la section 2.2.3.2.

PRÉFIXE(-)N

Le pluriel est obtenu par la mise au pluriel du substantif ; le préfixe est invariable.

- autocommutateur(s) (automatic exchange) (*m*)
- demi-circuit(s) (half circuit) (*m*)
- semi-conducteur(s) (semiconductor) (*m*)

ADJ-N

Les structures ADJ-N rencontrés dans le corpus caractérisent uniquement des noms. Le pluriel est obtenu par la mise au pluriel du substantif. L'adjectif reste invariable.

- court-circuit(s) (short circuiting) (*m*)
- plate-forme(s) (platform) (*f*)

N-ADJ

Les noms composés de structure N-ADJ rencontrés dans le corpus sont soit des noms propres soit des noms communs. Le pluriel des noms communs est obtenu par la mise au pluriel à la fois du substantif et de l'adjectif.

1. N-ADJ qui sont des noms propres
   - États-Unis ou Etat-Unis
   - Royaume-Uni (*m*)
2. N-ADJ qui sont des noms communs
   - station-terrestre, stations-terrestres (Earth station) (*f*)

   Cet exemple de composé est le seul du corpus et il n'apparaît qu'une fois. Étant donné le grand nombre d'occurrences du composé sans trait d'union, cette graphie correspond certainement plus à une erreur de frappe ou de traduction qu'à une réelle variante orthographique de *station terrestre*. Les N-ADJ qui sont des noms communs n'existent donc pas dans notre corpus.

N_1-N_2

Les noms composés de structure N_1-N_2 rencontrés dans le corpus sont soit des noms propres soit des noms communs. Les noms propres sont invariables. Le pluriel des noms communs est obtenu par la mise au pluriel des deux substantifs.

1. N_1-N_2 qui sont des noms propres
   - Reed-Solomon
   - Pleumeur-Bodou
2. N_1-N_2 qui sont des noms communs
   - aller-retour, allers-retours (round trip) (*m*)
   - entrée-sortie, entrées-sorties (input-output) (*f*)
   - modulateur-démodulateur, modulateurs-démodulateurs (modulator/demodulator) (*m*)

V-N

Lorsque le nom composé n'est pas invariable, le pluriel est obtenu par la mise au pluriel du substantif.

- brise-glace (invariable)(ice-breaker(s)) (*m*)
  Le genre masculin de ce nom composé ne correspond pas au genre du substantif *glace*. C'est donc une exception à la règle affirmant que le genre du composé est celui du premier nom rencontré.
- porte-conteneur(s) (container vessel) (*m*)

PREP-N

Le pluriel est obtenue par la mise au pluriel du substantif.

- après-midi (invariable) (afternoon) (*f*)
- contre-réaction(s) (negative feedback) (*f*)
- sous-ensemble(s) (sub-unit) (*m*)
- sous-système(s) (sub-system) (*m*)

N_1-*à*-N_2

Il n'existe aucun terme de structure N_1-*à*-N_2 sauf peut-être *Vis-à-vis* où *vis* n'est plus un substantif en français. *Vis-à-vis* peut-être une préposition ou un nom. Dans notre corpus, c'est toujours une préposition. Il n'existe donc pas de noms composés de structure N_1-*à*-N_2.

En conclusion, la liste des noms composés de longueur 1 présentée ci-dessus n'est pas exhaustive. D'autres structures sont sans doute possibles mais nous nous contenterons de celles-ci, la reconnaissance de ces unités ne posant pas de problème. Nous retiendrons la diversité des structures internes de ces noms composés, bien que certaines structures ne soient pas représentées (N-ADJ, N_1-*à*-N_2), et l'hétérogénéité de l'ensemble de leurs traductions en anglais en terme de nombre d'unités et d'orthographe ; un nom composé de longueur 1 en français a comme équivalent anglais un nom simple, un nom composé de longueur 1 (où les deux unités sont soit agglomérées, soit reliées par un trait d'union ou même une barre oblique (/)) ou un nom composé de longueur 2.

2.2.1.2 Noms composés de longueur 2

Les noms composés de longueur 2 sont formés de plusieurs unités lexicales séparées par des blancs et comprennent deux unités lexicales pleines, ou éventuellemet, un ou deux noms composés de longueur 1. Rappelons que nous avons défini les unités lexicales pleines comme les unités lexicales qui ne sont pas les mots grammaticaux.

N ADJ

Ce type élémentaire de nom composé est très fréquent. Nous avons inclu dans cette catégorie, les structures N PARTICIPE-PASSÉ et N PARTICIPE-PRÉSENT. Ces structures nous paraissent peu représentés et ne justifient donc pas une catégorie particulière. Le participe-passé s'accorde en nombre et genre avec le nom qu'il modifie, et il est souvent difficile de le différencier de l'adjectif lorsqu'il occupe une position épithête. Le participe-présent ne s'accordant pas en genre et en nombre, il n'y a ambiguïté avec l'adjectif qu'au masculin singulier.

- N ADJ (uniquement au singulier)
  - Service national (Domestic Service) (*m*)
  - signal modulant (modulating signal) (*m*)
  - téléphonie rurale (rural telephony) (*f*)
- N ADJ (uniquement au pluriel)
  - lobes latéraux (side lobes) (*m*)

- N ADJ (pluriel régulier)
  Le pluriel est obtenu par la mise au pluriel des deux unités lexicales, c'est-à-dire le nom et l'adjectif.
  - contour(s) formé(s) (shaped contour) (*m*)
  - élément(s) rayonnant(s) (radiator) (*m*)
  - faisceau(x) modelé(s) (shaped beam) (*m*)
  - fréquence(s) radioélectrique(s) (radio-frequency) (*f*)
  - orbite(s) géostationnaire(s) (geostationary orbit) (*f*)
  - satellite(s) géostationnaire(s) (geostationary satellite) (*m*)
  - station(s) brouilleuse(s) (interfering station) (*f*)
  - station(s) terrienne(s) (Earth station/earth station/earth-station) (*f*)

ADJ N

Les séquences ADJ N qui n'apparaissent qu'au singulier sont la plupart du temps des sous-séquences de groupes prépositionnels adverbiaux de structure PREP ADJ N comme par exemple : *à large bande (wideband)*. Les autres séquences ADJ N relevées dans le corpus ne sont ni des noms composés techniques, ni des co-occurrences lexicales restreintes. Il n'existe donc pas de noms composés élémentaire de structure ADJ N où l'adjectif est séparé du substantif par un blanc.

N_1 N_2

Les noms composés de structure N_1 N_2 peuvent être des variantes morphosyntaxiques d'une structure N_1 PREP (DET) N_2 où la préposition et éventuellement le déterminant ont été effacés (voir section 2.2.3.3). De plus contrairement aux autres structures qui se caractérisent par la présence ou l'absence de trait d'union, cette structure accepte dans certains cas les deux graphies. Lorsque la graphie avec trait d'union est attestée, la graphie sans trait d'union est considérée comme une variante orthographique (voir section 2.2.3.2). Pour les exemples présentés ci-dessous, la graphie avec trait d'union n'a pas été rencontrée dans le corpus. Le pluriel est obtenu soit en mettant au pluriel le premier nom, soit les deux noms.

- N_1 N_2 (pluriel sur le premier nom)
  - diode tunnel, diodes tunnel (tunnel diode) (*f*)
  - génératrice diesel, génératrices diesel (diesel generator) (*f*)
- N_1 N_2 (pluriel sur les deux noms)
  - canal sémaphore, canaux sémaphores (common channel) (*m*)

- mémoire tampon, mémoires tampons (buffer) (*f*)
- voie support, voies supports (bearer channel) (*f*)

N_1 *à* DET N_2

Le pluriel est obtenu par la mise au pluriel du premier nom ; le deuxième nom, dans le corpus, est généralement invariable (soit singulier, soit pluriel).

1. N_1 *à* N_2 (DET=0)
   - N_1 *à* N_2 (uniquement au singulier)
     - mise à jour (updating) (*f*)
     - pilote à quartz (quartz-cristal oscillator) (*m*)
   - N_1 *à* N_2 (pluriel régulier)
     - antenne(s) à réflecteur (reflector antenna) (*f*)
     - réflecteur(s) à grille (grid reflector) (*m*)
     - réseau(x) à satellite, réseau(x) à satellites (satellite network) (*m*)
       N_2 dans ce nom composé est rencontré au singulier et au pluriel. C'est donc une exception à la règle affirmant le caractère fixe du N_2.
     - système(s) à modulation (modulation system) (*m*)
2. N_1 *à* DET N_2 (DET=LE **avec** LE={*le,la,l',les*})
   Aucun nom composé avec le déterminant pluriel *les* n'a été rencontré.
   - assignation à la demande (demand-assignment) (*f*)
   - mise(s) au point (finalization) (*f*)
   - station au sol (Ground station) (*f*)
   - voie au repos (idle channel) (*f*)

N_1 *de* DET N_2

1. N_1 *de* N_2 (DET=0)
   Le pluriel est obtenu par la mise au pluriel du premier nom ; le deuxième nom, dans le corpus, est généralement invariable (soit singulier, soit pluriel).
   - N_1 *de* N_2 (uniquement singulier)
     - modulation de fréquence (frequency modulation) (*f*)
   - N_1 *de* N_2 (pluriel régulier)
     - bande(s) de fréquences (frequency band) (*f*)
       fréquence est toujours au pluriel.

- capacité(s) de transmission (transmission capacity) (*f*)
- service(s) de radiodiffusion (broadcasting service) (*m*)
- structure(s) de trame (frame structure) (*f*)
- système(s) de signalisation (signalling system) (*m*)
- zone(s) de couverture (coverage zone) (*f*)
- zone(s) de service (service zone or service area) (*f*)

2. N_1 *de* DET N_2 (DET=LE **avec** LE={*le,la,l',les*})

   - N *de* LE N (uniquement singulier)
     Ces noms composés sont rares et ressemblent à des variantes elliptiques de noms surcomposés obtenus par juxtaposition où le dernier élément a été omis (voir section 2.2.2.1).
     - spectre des fréquences (frequency spectrum) (*m*)
     - synchronisation des paquets (burst synchronisation) (*f*)
   - N *de* LE N (pluriel régulier)
     Ces noms composés à pluriel régulier sont rares.
     - traitement(s) des signaux (signal processing) (*m*)

N_1 *en* N_2

Les exemples relevés sont tous au singulier :

- multiplexage en fréquence (frequency division multiplexing) (*m*)
- réponse en fréquence (frequency response) (*f*)

N_1 *par* N_2

Le pluriel est obtenue par la mise au pluriel du premier nom ; le deuxième nom est généralement invariable (soit singulier, soit pluriel). Les séquences N_1 par N_2 n'apparaissant qu'au singulier sont souvent une sous-séquence d'un nom composé de type élémentaire de longueur 3 ou d'un nom surcomposé de longueur ≥ 3 (voir section 2.2.2.1). Les véritables noms composés de structure N_1 par N_2 acceptent le pluriel :

- circuit(s) par satellite (satellite circuit) (*m*)
- liaison(s) par satellite (satellite link) (*f*)
- service(s) par satellite (satellite service) (*m*)

N_1 PREP DET N_2 (PREP $\neq$ *à, de, en, par*)

Le pluriel est obtenu par la mise au pluriel du premier nom ; le deuxième nom est généralement invariable dans le corpus.

1. N_1 *sur* N_2 (PREP=sur et DET=0)
   - masse sur orbite (uniquement au singulier) (orbit mass) (*f*)
   - satellite(s) sur orbite (satellite in orbit) (*m*)
2. N_1 *dans* DET N_2 (PREP=*dans* et DET=LE avec LE={*le,la,l',les*})
   - répartition dans le temps (time division) (uniquement au singulier) (*f*)

2.2.1.3 Noms composés de longueur ≥ 3

Les noms composés de type élémentaire de longueur ≥ 3 sont rares et particulièrement difficiles à identifier ; en effet, ils ne doivent pas être amalgamés avec des noms composés surcomposés, modifiés ou coordonnés. Si aucune des transformations possibles que peut subir un nom composé de type élémentaire de longueur 1 ou 2 ne s'applique à une sous-séquence de longueur 1 ou 2 de la séquence de longueur 3 considérée, alors cette dernière est décrétée de type élémentaire. La difficulté réside dans le fait que la formation des noms composés élémentaires de longueur ≥ 3 s'est effectuée par l'application des règles de formation de nouveaux noms composés (ces règles sont présentées dans la section 2.2.2). La différence entre les surcomposés, les modifiés ou les coordonnés et les types élémentaires se situent au niveau du concept terminologique stabilisé de ces derniers. Un bon indice de cette stabilisation du concept est exprimé par la présence d'une abréviation. (les abréviations sont décrites dans notre section 2.2.3.1 consacrée aux variantes des noms composés). Mais, l'existence d'une abréviation pour un nom composé n'implique absolument pas que la structure morphosyntaxique de ce dernier soit figée et, les noms composés de type élémentaire de longueur ≥ 3 acceptent des variantes de la même façon que les surcomposés ou les modifiés. De la manière dont nous avons définie les noms composés terminologiques de type élémentaire, les noms composés de longueur ≥ 3 les plus représentés sont les composés de type élémentaire de longueur 1 ou 2 modifiés ou surcomposés, une abréviation n'étant pas introduite pour chaque terme. Les noms composés de longueur ¿ 3 ont été estimés par [Nkwenti-Azeh, 1992] pour l'anglais à moins de 5 % de l'ensemble des termes du domaine ; son domaine étant le même que le notre, c'est-à-dire les communications par satellite. Cette estimation nous donne une idée de la marginalité de ces constructions de type élémentaire de longueur ≥ 3.

NOMS COMPOSÉS DE LONGUEUR 3

Les noms composés de longueur 3 sont construits avec trois unités lexicales pleines ou éventuellement deux unités lexicales pleines et une préposition composée. Les exemples donnés ci-dessous décrivent quelques structures rencontrées accompagnées pour la plupart de leurs abréviations :

1. N_1 $PREP_{comp}$ (DET) N_2
   $PREP_{comp}$ signifie que la préposition apparaissant à l'intérieur du nom composé est elle-même composée.

   - assignation en fonction de la demande (demand-assignement - DA) (uniquement singulier) (*f*)

2. N *non* ADJ

   - rayonnement(s) non essentiel(s) (spurious emission) (*m*)

   Notons que la traduction anglaise ne porte pas de signe de négation.

3. N ADJ_1 ADJ_2

   - bande latérale unique - BLU (single side-band - SSB) (*f*)

4. N_1 ADJ PREP N_2

   - service(s) fixe(s) par satellite - SFS (satellite fixed service - FSS) (*m*)

5. N_1 $PREP_1$ N_2 $PREP_2$ N_3

   - modulation par déplacement de phase - MDP (phase-shift keying - PSK) (uniquement singulier dans le corpus) (*f*)

6. N_1 CONJ ($PREP_1$) (DET) N_2 $PREP_2$ N_3
   Cette structure de nom composé suit les règles générales de détermination. En conséquence, la préposition et l'article qui apparaissent juste avant N_2 sont optionnels. La présence de la préposition dépend de la position syntaxique du nom composé dans la phrase; la présence et le type de déterminant de N_2 dépendent de ceux de N_1. Ainsi, si ce nom composé est le premier argument d'un nom prédicatif et si N_1 n'a pas de déterminant, alors DET=0 comme dans *le service d'assemblage et de désassemblage de paquets*; s'il est le premier argument d'un nom prédicatif et si le déterminant de N_1 est un article défini, alors DET=*le* comme dans *la gestion de l'assemblage et du désassemblage de paquets.*

   - N_1 et N_2 de N_3 : assemblage et (de + du + 0) désassemblage de paquets - ADP (packet assembly/desassembly - PAD) (*m*)

NOMS COMPOSÉS DE LONGUEUR >3

Voici deux exemples :

- N ADJ_1 ADJ_2 ADJ_3 : puissance isotrope rayonnée équivalente - PIRE (equivalent isotropically radiated power - EIRP) (*f*)

- N_1 ADJ PREP_1 N_2 PREP_2 N_3 : équipement terminal de traitement de donnée - ETTD (data terminal equipment - DTE)

2.2.2 Surcomposition, modification et coordination

De nouveaux noms composés se créent à partir de noms composés. Nous avons distingué trois opérations permettant de passer d'un nom composé de longueur 2 à un nom composé de longueur ≥ 3 : la surcomposition, la modification et la coordination. Néanmoins, statuer sur le caractère terminologique des noms composés surcomposés ou modifiés reste difficile, comme nous le verrons. Avant de décrire ces trois opérations, nous allons rappeler les travaux de [Jacquemin, 1991] sur la formation de nouveaux noms composés et nous nous positionnerons par rapport à son approche.

[Jacquemin, 1991] a étudié la surcomposition, c'est-à-dire la manière dont de nouveaux noms composés se créent à partir de noms composés déjà existants. La surcomposition utilise deux opérations :

- la **juxtaposition**, où la structure du ou des nom(s) composé(s) déjà existant(s) utilisé(s) dans cette opération reste identifiable. Par exemple, si *champ électrique* est un nom composé attesté, *champ électrique statique* est un nouveau nom composé obtenu par juxtaposition de l'adjectif *statique* au nom composé *champ électrique*. La structure du nom composé *champ électrique* n'est pas altérée.

- le **recouvrement**, où l'une des structures des noms composés déjà existants utilisés dans cette opération peut être altérée. Par exemple, si les noms composés *panneau de comptage* et *comptage d'électricité* sont des noms composés attestés, ils peuvent s'imbriquer sur leur élément en commun *comptage* pour donner le nouveau nom composé *panneau de comptage d'électricité* ; les structures des deux composés ne sont pas ici altérées car le recouvrement s'effectue par le "milieu". Au contraire, si le recouvrement s'effectue, non pas sur l'élément du milieu mais sur l'élément de tête, comme pour *antenne parabolique* et *antenne de réception* sur le nom *antenne* pour produire *antenne parabolique de réception*, la structure de *antenne de réception* est altérée.

La juxtaposition, telle qu'elle est définie, inclut la modification de noms composés par un adjectif ou un groupe prépositionnel antéposé. Il ne nous semble pas aussi évident de différencier un nom composé modifié d'un surcomposé obtenu par juxtaposition : pourquoi *champ électrique statique* est-il considéré comme un surcomposé et pas comme un nom composé modifié? Cette décision a peut-être été prise en vue du caractère technique de l'adjectif *statique*. Mais, même un adjectif très commun comme *unique* peut donner lieu à de nouveaux noms composés, comme le montre le nom composé *bande latérale unique* dont le statut terminologique est attesté par la présence d'une abréviation. D'autre part, certains recouvrements, comme le nom composé *courant d'air frais* obtenu par recouvrement de la tête *courant* en utilisant les deux noms composés *courant d'air* et *courant frais*, nous paraissent bizarres : il ne suffit pas que deux formes soient attestées et qu'elles aient un élément en commun pour qu'il y ait recouvrement. Enfin, ces deux opérations reposent sur la présence attestée ou non des noms composés dans le lexique : s'il est difficile de passer outre ces hypothèses, que nous allons d'ailleurs reprendre, cela implique d'abord contruire un lexique des noms composés binaires du domaine, Notre définition de la surcomposition s'inspire des travaux de [Jacquemin, 1991], mais impose des critères plus sélectifs ; par exemple, nous n'acceptons pas les structures modifiées comme des structures surcomposées. De plus, nous considérons que la modification est autant productrice de nouveaux noms composés que la surcomposition. La coordination, quant à elle l'est beaucoup moins et est rapidement abordée. Les exemples sont extraits de nos textes : le "Manuel des Communications par Satellite" (MTS) et le "Livre Bleu du CCITT" (LBC)

2.2.2.1 Surcomposition sur les types élémentaires

Nous distinguons deux types de surcomposition : la surcomposition par juxtaposition et la surcomposition par substitution. La première respecte la structure interne du composé élémentaire utilisé, la seconde est susceptible de la modifier. Nous définissons la surcomposition à partir, principalement, de noms composés binaires attestés. Cette hypothèse de travail nous pose néanmoins problème puisque nous ne possédons une telle liste.

Juxtaposition

Un nom surcomposé obtenu par juxtaposition, que nous appellerons un juxtaposé, est construit avec au moins un nom composé de type élémentaire et se caractérise par les propriétés suivantes :

- les éléments de la structure du ou des composé(s) de type élémentaire restent solidaires,

- lorsque un nom simple se juxtapose à un nom composé, c'est le plus souvent le nom simple qui précède le composé,
- la juxtaposition s'effectue par l'intermédiaire d'une préposition,
- les enchevêtrements à l'intérieur de la structure juxtaposée ne réfèrent pas à des noms composés de type élémentaire. Cette propriété est illustrée dans les exemples qui suivent.

Nous allons maintenant donner quelques exemples de noms juxtaposés ; le ou les noms composés de type élémentaire apparaissent entre crochets :

1. Juxtaposition d'un nom composé de type élémentaire et d'un nom simple $\mathbf{N_1}$ PREP$_1$ [$\mathbf{N_2}$ PREP$_2$ $\mathbf{N_3}$] (longueur 3)
   - modulation par [déplacement de phase]
     ([phase shift keying]
     Ni *modulation par déplacement*, ni *modulation de phase* ne sont des noms composés de type élémentaire.
   - connection de [réseaux de microstations]
     ([very small aperture terminal network]
     Ni *connection de réseaux*, ni *connection de microstations* ne sont des noms composés de type élémentaire.
2. Juxtaposition de deux noms composés de type élémentaire
   (a) [$\mathbf{N_1}$ $\mathbf{Adj_1}$] PREP$_1$ [$\mathbf{N_2}$ PREP$_2$ **(Det)** $\mathbf{N_3}$] (longueur 4)
   - [accès multiple] avec [assignation à la demande]
     ([demand assigned] [multiple access])
     Ni *accès avec assignation*, ni *accès à la demande* ne sont des noms composés de type élémentaire.
   - [accès multiple] par [répartition dans le temps]
     ([time division] [multiple access])
     Ni *accès par répartition*, ni *accès dans le temps* ne sont des noms composés de type élémentaire.

   (b) [$\mathbf{N_1}$ PREP$_1$ $\mathbf{N_2}$] PREP$_2$ **(Det)** [$\mathbf{N_3}$ PREP$_3$ $\mathbf{N_4}$] (longueur 4)
   - [code d'identification] de [réseau de données]
     ([data network] [identification code])
     Ni *code de réseau*, ni *identification de réseau*, ni *code de données*, et ni *identification de données* ne sont des noms composés de type élémentaire.

Substitution

La substitution est définie de la manière suivante : étant donné un nom composé de longueur 2, l'une des unités lexicales pleines est substituée par un nom composé dont la tête est cette unité lexicale. Par exemple, dans la structure de type élémentaire N_1 PREP$_1$ N_2, N_1 peut être substitué par un nom composé de structure N_1 PREP$_2$ N_3, pour donner un surcomposé de structure N_1 PREP$_2$ N_3 PREP$_1$ N_2 comme par exemple le nom *réseau* dans le nom composé *réseau à satellite(s)* est substitué par le nom composé *réseau de transit* pour former le surcomposé *réseau de transit à satellite(s)* La substitution se différencie de la juxtaposition de deux façons :

- elle demande obligatoirement l'emploi de deux noms composés élémentaires, dans notre exemple ci-dessus *réseau à satellite* et *réseau de transit*,
- elle ne respecte pas toujours leurs structures internes : dans notre exemple ci-dessus, la structure de *réseau à satellite* est altérée.

Nous avons appelé cette opération substitution, [Jacquemin, 1991] la nomme recouvrement partiel. Le terme de substitution nous semble plus adéquat que le terme de recouvrement partiel ; il existe en effet un lien sémantique entre le substitué (hyponyme) et le substituant (hyperonyme) identifié comme une relation d'hyperonymie. Voici quelques exemples de noms surcomposés obtenus par substitution :

1. N_1 PREP N_2 + N_2 ADJ → N_1 PREP N_2 ADJ
   - réseau à satellite(s) + satellite(s) géostationnaire(s) → réseau(x) à satellites géostationnaires
     (satellite network + geostationnary satellite → geostationnary satellite network(s))
   - spectre des fréquences + fréquences radioélectriques → spectre des fréquences radioélectriques
     (frequency spectrum + radio-frequency → radio-frequency spectrum)
   - signalisation par canal + canal sémaphore → signalisation par canal sémaphore
     (channel signalling + common channel → common channel signalling)
2. N_1 PREP$_1$ N_2 + N_2 PREP$_2$ N_3 → N_1 PREP$_1$ N_2 PREP$_2$ N_3
   - spectre de bruit + bruit d'intermodulation → spectre de bruit d'intermodulation
     (noise spectrum + intermodulation noise → intermodulation noise spectrum)

3. N_1 PREP N_2 + N_1 ADJ → N_1 ADJ PREP N_2

   - densité de bruit + densité spectrale → densité spectrale de bruit
     (noise density + spectral density → noise spectral density)

4. N_1 $PREP_1$ N_2 + N_1 $PREP_2$ N_3 → N_1 $PREP_1$ N_3 $PREP_2$ N_2

   - antenne du satellite + antenne de réception → antenne de réception du satellite
     (satellite antenna + receiving antenna → satellite receiving antenna)

Pour certaines structures, il est difficile de trancher entre substitution et modification : par exemple, si N_1 prep N_2 est un nom composé, N_1 peut être substitué par un hyperonyme de type N_1 ADJ, comme par exemple dans *réseau à satellites*, *réseau* est substitué par *réseau national* pour former le surcomposé suivant : *réseau national à satellites* ; l'autre interprétation consiste à dire que le terme *réseau à satellites* est modifié par l'adjectif *national*. Choisir entre ces deux interprétations revient à statuer sur le caractère d'hyperonyme ou non de *réseau national* par rapport au nom *réseau* employé à l'intérieur du composé *réseau à satellite* ; ce qui n'est pas toujours trivial.

2.2.2.2 Modification sur les types élémentaires

Les modifieurs apparaissent avant, à l'intérieur ou après le nom composé. Les modifieurs qui apparaissent avant ou après ne cassent pas la structure, au contraire de ceux qui apparaissent à l'intérieur. Ces deux cas de modifications peuvent donner lieu à de nouveaux noms composés. Il est presque impossible de statuer sur le caractère composé ou non du nom composé modifié. L'attitude qui est adoptée consiste à enregistrer les modifieurs rencontrés ainsi que leur place. Nous décrivons en premier la modification qui prend place à l'intérieur du nom composé, puis celle qui prend place avant ou après le nom composé.

Insertion de modifieurs

Les modifieurs qui peuvent être insérés à l'intérieur d'une structure de nom composé de type élémentaire sont principalement les adjectifs et les adverbes.

1. **Insertion de modifieurs à l'intérieur d'un composé de longueur 2**

   (a) **Insertion d'un adjectif**
   L'adjectif s'insère à l'intérieur de la structure N_1 PREP (DET) N_2 juste après le N_1. En ce qui concerne la structure de nom composé binaire N ADJ, nous n'avons pas accepté la possibilité d'avoir une insertion d'adjectif après le N. L'adjectif lorsqu'il modifie une structure N ADJ apparaît systématiquement en post-position. Certaines apparences sont pourtant trompeuses : nous avons relevés les termes *bande*

étroite et *bande passante étroite* sans rencontrer *bande passante*. Une conclusion hâtive serait de considérer que *passante* est une adjectif modifieur inséré ; conclusion qui est contredite par l'examen de la traduction anglaise, respectivement *narrow band* et *narrow passband*, où le concept de *bande passante* (*passband*) est clairement identifié. La non-possibilité d'avoir un adjectif inséré dans une structure N ADJ a été confirmée par le fait que nous n'avons pas pu trouver d'exemple de structure N ADJ_1 ADJ_2 française à laquelle correspondrait une structure $\textbf{adj}_1$ ADJ_2 N anglaise. La seule structure retenue pour l'insertion de l'adjectif est donc :

- N_1 PREP (DET) $N_2 \rightarrow N_1$ **Adj** PREP (DET) N_2
  - réseaux **mondiaux** de télécommunication
    (global telecommunications networks)
  - liaisons **multiples** par satellite
    (multiple satellite links)
  - précision **globale** de pointage
    (overall pointing accuracy)

(b) **Insertion d'un adverbe**
Cette insertion a lieu à l'intérieur de la structure de base N ADJ juste après le N. Rappelons que dans la structure N ADJ, l'adjectif représente aussi bien un adjectif, qu'un participe passé ou un participe-présent employé comme adjectif.

- N ADJ $\rightarrow$ N **Adv** ADJ
  - orbite **presque** géostationnaire
    (near [geostationnary orbit])
  - réseaux **entièrement** numériques
    (all [digital networks])
  - bruit **non** thermique
    (non-thermal noise)

2. **Insertion de modifieurs à l'intérieur de noms composés de longueur 3**
Les noms composés de type élémentaire étant peu nombreux, les modifications qui les affectent sont assez rares. Voici celles que nous avons rencontrées :

(a) **Insertion dans la structure N_1 ADJ PREP N_2**

N_1 ADJ PREP $N_2 \rightarrow N_1$ ADJ **Adj** PREP N_2

- services fixes **nationaux** par satellite
  (domestic [fixed satellite service])

(b) **Insertion dans la structure** N_1 PREP$_1$ N_2 PREP$_2$ N_3

N_1 PREP$_1$ N_2 PREP$_2$ N_3 → N_1 **Adj** PREP$_1$ N_2 PREP$_2$ N_3

- services **communautaires** de radiodiffusion par satellite
  (community [broadcasting-satellite services])

Antéposition et postposition des modifieurs

Les modifieurs qui apparaissent avant le nom composé ne prennent généralement pas part à la formation de nouveaux noms composés. Les adjectifs antéposés forment une classe fermée, dont les plus utilisés sont: *divers, nombreux, tel, autre, principal, etc.* Les adjectifs qui apparaissent après le nom composé et qui modifient le terme entier s'accordent avec le composé. En voici quelques exemples:

1. [N_1 ADJ$_1$] **Adj$_2$** (longueur 3)
   - [débit binaire] élevé
     (high [bit rate])
   - [station terrienne] brouilleuse
     (interfering [earth(-)station])
2. [N_1 PREP$_1$ N_2] **Adj** (longueur 3)
   - [largeur de bande] occupée
     ([bandwidth] occupancy)
   - [satellite de télécommunication] géostationnaire
     (geostationay [telecommunication satellite])
3. [N_1 N_2] **Adj** (longueur 3)
   - [modulation delta] adaptable
     (adaptative [delta modulation])
   - [interfaces usager-réseau] polyvalentes
     (multi-purpose [user network interface])

Un autre type de modifieur apparaît en postposition: ce sont les groupes prépositionnels adverbiaux de structure PREP ADJ N ou PREP N ADJ. Ce groupe prépositionnel occupe une position épithète mais reste néanmoins invariable, soit singulier, soit pluriel. La préposition et le nom sont généralement fixes l'un par rapport à l'autre; seuls un petit nombre d'adjectifs sont employés avec une préposition et un nom donnés. Ces groupes prépositionnels sont issus de phrases de forme N_1 être PREP X et acceptent donc les restructurations suivantes:

N_1 [PREP ADJ N_2]
N_1 être PREP adj N_2
DET$_{def}$ ADJ N_2 de LE N_1

Le nom modifié par ce groupe prépositionnel peut engendrer un nouveau nom composé. Des exemples sont les suivants :

1. **Modification d'un nom simple**

   - amplificateur(s) [à faible bruit]
     ([low noise] amplifier(s))
   - antenne [à contours formés]
     ([shaped beam] antenna)
   - station(s) [de petite taille]
     (small station(s))
   - système(s) [à faible capacité]
     ([low capacity] system(s))

2. **Modification d'un nom composé**

   (a) [$\mathbf{N_1}$ $\mathbf{ADJ_1}$] [PREP ADJ N]

   - [câble(s) [sous-marin(s)]] [à large bande]
     ([wideband] [[submarine] cable(s)])

   (b) [$\mathbf{N_1}$ $\mathbf{N_2}$] [PREP ADJ N]

   - [interface(s) usager-réseau] [à usage multiple]
     ([multipurpose] [user-network interface(s)])

   (c) [$\mathbf{N_1}$ $\mathbf{PREP_1}$ $\mathbf{N_2}$] [PREP ADJ N]

   - [liaison(s) de télécommunication] [à grande vitesse]
     ([high speed] [communication link(s)])

Si de nombreux groupes prépositionnels acceptent la transformation ci-dessus, certains dont les groupes prépositionnels introduits par la préposition *par* la refusent. Voici un exemple :

- réutilisation des fréquences [par double polarisation]
  ([frequency re-use] [by dual polarization])

Pour d'autres, cette transformation est à la limite de l'acceptabilité :

- réseaux [à double sens]
  $\leftrightarrow$ ces réseaux sont à double sens
  $\leftrightarrow$?le double sens des réseaux

2.2.2.3 Modification et surcomposition

La modification et la surcomposition s'appliquent récursivement. Nous allons présenter quelques exemples de surcomposition de composé, puis de modification de modifié, et enfin un exemple où ces opérations s'enchevêtrent.

Surcomposition de composé(s)

La surcomposition de niveau 2 s'opère nettement plus difficilement que la surcomposition de niveau 1. Nous n'examinerons pas les surcompositions de niveau ¿ 2 ; en effet, les abréviations sont souvent utilisées dans la surcomposition ¿ 2 (les abréviations de noms composés sont présentées dans la section 2.2.3.1).

1. **Juxtaposition d'un surcomposé et d'un nom simple**

   - procédure d' [accès à la liaison symétrique] (longueur 4)
     ([bilateral link access] procedure)

2. **Juxtaposition d'un surcomposé et d'un nom composé de type élémentaire**

   - [réseau public pour données] avec [commutation de paquets] (longueur 5)
     ([packet switch] [public data network])

3. **Juxtaposition de deux surcomposés de niveau 1**

   - [service de circuit virtuel] du [RNIS] (longueur 4)
     ([ISDN] [virtual circuit service])
     où *RNIS* est l'abréviation du terme : *réseau numérique à intégration de service* obtenu par juxtaposition (*ISDN = integrated service digital network*).

Modification de modifié(s)

La modification comme la surcomposition peut s'appliquer de manière récursive. Néanmoins les modifiés de niveau ¿2 sont beaucoup plus rares. En voici quelques exemples :

1. N ADJ ADJ → N ADJ ADJ **Adj** (longueur 4)

   - stations terriennes numériques **petites**
     (small sized digital earth station)

2. N ADJ ADJ → N ADJ **Adv** ADJ (longueur 4)

   - bandes interdites **très** étroites
     (very narrow forbidden bands)
   - terminaux terriens **facilement** transportables
     (readily transportable earth terminals)

3. N_1 N_2 ADJ $\rightarrow$ N_1 N_2 ADJ **Adj** (longueur 4)

   - signal vidéo monochrome **composite**
     (monochrome composite video signal)

4. N_1 ADJ PREP N_2 $\rightarrow$ N_1 ADJ **Adj** PREP N_2 (longueur 4)

   - densité spectrale **maximale** de puissance
     (maximum spectral power density)
   - services publics **nationaux** de télécommunications
     (Domestic public telecommunication services)

5. N_1 ADJ PREP N_2 $\rightarrow$ N_1 **Adj** ADJ PREP N_2 (longueur 4)

   - services **occasionnels** internationaux de télévision
     (occasional use international televison services)

Modification de Surcomposé(s) et Surcomposition de Modifié(s)

Modification et surcomposition s'enchevêtrent pour former de nouveaux noms composés. Nous allons donner quelques exemples des enchevêtrements possibles :

1. **Surcomposition de modifié(s)**

   Modification : N ADJ $\rightarrow$ N **Adv** ADJ
   réseaux numériques $\rightarrow$ réseaux **entièrement** numériques
   Substitution : N_1 PREP N_2 + N_1 **Adv** ADJ $\rightarrow$ [N_1 **Adv** ADJ] PREP N_2

   - réseaux par satellite + réseaux **entièrement** numériques $\rightarrow$ [réseaux entièrement numériques] par satellite
     (all digital satellite network)

2. **Modification de surcomposé(s)**

   Surcomposition : N_1 PREP N_2 + N_1 ADJ $\rightarrow$ N_1 ADJ PREP N_2
   réseau de Terre + réseau public $\rightarrow$ réseau public de Terre (terrestrial public network)
   Modification : N_1 ADJ PREP N_2 $\rightarrow$ [N_1 ADJ PREP N_2] **Adj**

   - [réseaux publics de Terre] **locaux**
     (local terrestrial public network)

3. **Un exemple un peu plus complexe**

Modification : N ADJ → [N ADJ] **Adj**
interface numérique → interface numérique **directe**
Substitution : N_1 PREP N_2 + N_2 ADJ ADJ → N_1 PREP [N_2 ADJ ADJ]
module d'interface + interface numérique directe → module d'interface numérique directe
Modification : N_1 PREP N_2 ADJ_1 ADJ_2 → N_1 **Adj** PREP N_2 ADJ_1 ADJ_2
module d'interface numérique directe → module **transparent** d'interface numérique directe
Modification : N_1 ADJ_3 PREP N_2 ADJ_1 ADJ_2 → N_1 ADJ_3 **Adj** PREP N_2 ADJ_1 ADJ_2
module transparent **spécial** d'interface numérique directe
(special transparent direct digital interface module)

2.2.2.4 Coordination

La coordination de noms composés a été traitée dans [Jacquemin, 1991]. Ce phénomème étant relativement complexe et ne donnant généralement pas naissance à un nouveau nom composé, nous ne donnerons ci-dessous que quelques exemples de coordination rencontrés :

Coordination morphologique

Un seul exemple a été relevé :

- mono-, bi-, ou tridimensionnel
  (one-, -, or three-dimensional)

Coordination de deux noms composés de type élémentaire partageant une même structure

1. Schémas de coordination pour deux structures de type N_1 PREP N_2 où la préposition est identique

   (a) Coordination sur le N_1
   [N_1 $PREP_1$ N_2], [N_1 $PREP_2$ N_3] → N_1 $PREP_1$ N_2 CONJ ($PREP_2$) N_3
   - équipement d'émission et de transmission
     (transmit and receive equipment)
   - systèmes de surveillance, d'alarme et de commande
     (monitoring, alarm and control systems)

   (b) Coordination sur le N_2
   [N_1 PREP N_2], [N_3 PREP N_2] → N_1 CONJ N_3 PREP N_2
   - élévateurs et abaisseurs de fréquence
     (up and down converters)

Coordination de deux noms composés de type élémentaire de structures différentes

1. [N_1 PREP$_1$ N_2], [N_1 PREP$_2$ N_3] → N_1 PREP$_1$ N_2 CONJ PREP$_2$ N_2

   - couplage avant amplification ou après amplification (post-HPA versus pre-HPA)
   - systèmes en câble et à satellites (cable and satellite systems)

2. [N_1 PREP N_2], [N_1 ADJ] → N_1 ADJ CONJ PREP N_2

   - services spatiaux et de Terre (space and terrestrial services)

Coordinations diverses

- signalisation et interconnexion des liaisons par satellite (signalling and interconnecting satellite links)
- service fixe par satellite et de radiodiffusion par satellite (fixed-satellite and broadcasting-satellite service)

2.2.2.5 Conclusion

Le plongeon que nous venons de faire, a mis en évidence un important brouillage des données entre surcomposition et modification. Ces deux phénomènes ont des définitions linguistiques différentes mais les données étudiées révèlent que la frontière réelle est floue. La linguistique est ici un peu décevante: la morphosyntaxe ne permet pas véritablement de trancher. L'étude de notre corpus met bien en valeur le caractère lexical des phénomènes de surcomposition et de modification, au contraire de la coordination qui relève clairement du domaine de la syntaxe. Malheureusement, ceci revient à reconnaître la faiblesse de l'apport linguistique dans ce domaine précis. La lexicalité des phénomènes les rende très difficiles à prendre en compte dans un modèle sans dictionnaire comme le notre. Nous aurons l'occasion par la suite d'exposer les difficultés rencontrées dans l'extraction à cause de la présence de surcomposés et de modifiés. Face aux manques de critères, nous nous trouvons placée devant la nécessité de trancher relativement arbitrairement entre surcomposition et modification.

2.2.3 Variantes

Nous avons considéré quatre catégories principales de variantes: les abréviations, les variantes orthographiques, les variantes morphosyntaxiques et les va-

riantes elliptiques. Seules les abréviations peuvent être plus fréquemment utilisées que le nom composé auxquelles elles réfèrent.

2.2.3.1 Abréviation

Cette section est divisée en deux sous-sections : l'une pour les abréviations de termes et l'autre pour les abréviations d'organisations, de compagnies, etc., aussi appelées sigles. Seule la première catégorie d'abréviation soulève quelques problèmes.

Abréviation d'un terme

Les abréviations attestent en quelque sorte du caractère terminologique du composé. Nous avons vu en 2.2.1.3 comment elles étaient utilisées pour isoler les noms composés de type élémentaire de longueur ≥ 3. Les abréviations ne sont pas forcément utilisées parce que le terme est fréquemment employé ou s'il est long : des abréviations de noms composés binaires existent, de même que des abréviations qui n'apparaissent qu'une fois dans le corpus. L'abréviation d'un terme est obtenue en prenant la première lettre de chaque unité lexicale pleine. Dans les tableaux ci-dessous, nous donnons quelques exemples d'abréviations accompagnées du terme qu'elles désignent, ainsi que du nombre d'occurrences des deux graphies dans nos deux textes MTS et LBC ; la traduction anglaise peut là encore se révéler particulièrement utile lorsqu'une même abréviation est partagée par deux termes différents à l'intérieur d'un même texte.

	réseaux numériques avec intégration des services (*integrated services digital networks*)	*RNIS* (*ISDN*)
MTS	1	96
LBC	169	235

	équipement(s) termin(al/aux) de traitements de données (*data terminal equipment(s)*)	*ETTD* (*DTE*)
MTS	9	23
LBC	13	2888

	service fixe par satellite (*fixed-satellite service*)	*SFS* (*FSS*)
MTS	130	130
LBC	0	0

Examinons maintenant un cas plutôt amusant d'abréviation possédant deux référents possibles :

	Assemblée(s) plénière(s) (*plenary assembly*)	*amplificateur(s) de puissance* (*high power amplifier*)	*AP* (*PA/HPA*)
MTS	2	135	10
LBC	44	0	72

Dans le corpus LBC, le terme *amplificateur de puissance* n'apparaît pas ; *AP* désigne donc *Assemblée(s) plénière(s)*. Par contre dans MTS, il n'est pas possible de trancher sur ce que désigne *AP* ; cela peut-être soit *amplificateur de puissance* soit *Assemblée(s) plénière(s)*. Heureusement dans ce cas, la traduction anglaise lève l'ambiguïté.
Les abréviations dépendent non seulement du domaine technique, mais plus encore du texte technique où elles ont été introduites. De manière à éviter des erreurs dans la mise en correspondance avec leur référent, il faut conserver l'information indiquant le texte où l'abréviation a été introduite.
Les abréviations sont utilisées comme des unités lexicales dans le corpus : elles gardent les propriétés syntaxiques des noms simples, sauf bien entendu l'accord. En conséquence, l'abréviation d'un terme est considérée comme une variante lexicale de longueur 1 avec les propriétés suivantes :

- elle peut apparaître plus fréquemment que le terme auquel elle réfère,
- elle peut être utilisée à la place du terme pour donner lieu à un surcomposé par juxtaposition (voir section 2.2.2.3).

Sigle

Quelques exemples de sigles rencontrés :

	CCITT[a] (*CCITT*)	*CCEP*[b] (*CCPS*)	*UPU*[c] (*UPU*)	*CEI*[d] (*IEC*)
MTS	108	0	0	0
LBC	362	9	3	38

[a]Conseil Consultatif International Télégraphique et Téléphonique
[b]Conseil Consultatif des Études Postales
[c]Union Postale Universelle
[d]Commission Électrotechnique Internationale

2.2.3.2 Variantes orthographiques

Les variantes orthographiques d'un nom composé sont de trois types:

- variation en nombre de N_2 normalement interdite dans les structures de type élémentaire N_1 Prep N_2,
- l'un des composants du nom composé a plusieurs orthographes possibles,
- caractère optionnel du trait d'union.

1. **La flexion du N_2 n'est pas fixe**

   Dans la section 2.2.1.2, pour les noms composés de type N_1 Prep N_2, nous avons affirmé que la flexion en nombre du N_2 était fixe, c'est-à-dire soit singulier, soit pluriel. Il existe néanmoins quelques exceptions, comme:

   - réseau(x) à satellite, réseau(x) à satellites
     (satellite network(s))

2. **Variation graphique**
   L'utilisation des lettres capitales pour certains noms composés est optionnelle:

   - **S**ervice de transmission des données/service de transmission des données
     (data transmission service)
   - **S**ervice national/service national
     (Domestic Service/domestic service)

   En fait, l'emploi d'une majuscule initiale semble être dépendante du corpus. En effet, dans MTS, **S***ervice national* apparaît toujours avec cette graphie et peut être ainsi différencié de *service national de communications* qui apparaît toujours sans lettre capitale. Dans LBC, la séquence **s***ervice national*, qui représente soit le nom composé de type élémentaire soit la sous-séquence d'un plus grand terme, n'apparaît jamais avec une lettre capitale. Les lettres capitales peuvent donc être utilisées pour identifier des noms composés uniquement à l'intérieur d'un même corpus et sous la condition que le même nom composé sans lettres capitales n'existe pas.

3. **Trait d'union facultatif**
   Le trait d'union est généralement optionnel pour les structures N_1 N_2 présentées dans les sections 2.2.1.1 et 2.2.1.2 comme pour:

   - mode-paquet/mode paquet (packet-mode/packet mode)

Lorsque les deux graphies existent, le nom composé avec blanc est considéré comme une variante orthographique de celui avec trait d'union. Le trait d'union est interdit pour les noms composés de type élémentaire dont les structures sont les suivantes: N ADJ et N_1 PREP N_2. Le blanc est interdit pour les noms composés de type élémentaire: ADJ-N.

2.2.3.3 Variantes morphosyntaxiques

Les variantes morphosyntaxiques du nom composé sont de trois types:

- simplification/complication de la structure du nom composé par l'effacement/l'ajout de la préposition ou/et du déterminant qui apparaît à l'intérieur du nom composé,
- relation de synonymie entre deux structures de nom composé qui diffèrent seulement par l'une de leurs unités lexicales pleines,
- variation de la préposition.

1. **Simplification/complication de la structure du nom composé**
   Certaines structures de nom composé se simplifient par l'effacement de la préposition ou/et du déterminant du nom composé. De la même manière, une préposition ou/et un déterminant peuvent être introduits à l'intérieur de la structure. Comme cet effacement (resp. ajout) ne prend pas en compte les unités lexicales pleines, la structure simplifiée ou compliquée est considérée comme une variante morphosyntaxique et non pas comme une variante elliptique (les variantes elliptiques sont présentées dans la section suivante 2.2.3.4).

   - N_1 N_2 = N_1 de N_2
     tension hélice = tension d'hélice
   - N_1 de N_2 ADJ_2 PREP DET ABR = N_1 de N_2 ADJ_2 ABR
     service de circuit virtuel du RNIS = service de circuit virtuel RNIS (ISDN virtual circuit service)

2. Relation de synonymie
   Nous avons observé des relations de synonymie entre des structures différentes comme (N ADJ) et (N_1 de N_2) ou encore (N ADJ PARTICIPE-PASSÉ) et (N_1 ADJ à N_2):

(N ADJ) PARTICIPE-PASSÉ	LBC	(N_1 ADJ) à N_2	LBC
réseau public commuté	11	*réseau public à commutation*	1
réseaux publics commutés	2	*réseaux publics à commutation*	0
(public switched network(s))		*(public switched network(s))*	

Les relations de synonymie sont néanmoins à examiner très soigneusement : des séquences qui semblent référer à une même entité sont en fait deux termes différents comme l'indique la traduction anglaise dans l'exemple ci-dessous :

N adj	MTS	N_1 de N_2	MTS
station terrienne	402	*station de Terre*	9
stations terriennes	982	*stations de Terre*	14
(earth station(s))		*(terrestrial station(s))*	

3. **Changement de préposition**
   Une exemple de variation de préposition :

   - pour → de
     - N_1 pour N_2 → N_1 de N_2
       réseau pour données → réseau de données
       (data network)
     - N_1 ADJ pour N_2 → N_1 ADJ de N_2
       réseau public pour données → réseau public de données
       (public data network)

   Cet exemple montre que la variation de préposition est peut-être maintenue pendant la surcomposition.

2.2.3.4 Variantes elliptiques

Il arrive qu'un long nom composé soit évoqué par une sous-séquence de sa structure originale. Même si les textes techniques préfèrent les abréviations aux variantes elliptiques, nous en avons identifié quelques-unes, principalement lors de phénomènes de discours.
Un nom composé peut être évoqué par un nom composé elliptique où une ou plusieurs de ses unités lexicales pleines ont disparu. Cette variante elliptique est généralement soit la première unité lexicale soit une structure simplifiée. Si c'est le plus souvent l'élément de queue qui disparaît, les autres cas sont néanmoins

possibles. Quand la préposition ou l'article appartenant à la structure du nom composé disparaît, nous avons considéré ces variantes comme morphosyntaxiques (voir 2.2.3.3).

Variante elliptique de longueur 1

Voici quelques exemples :

1. variante elliptique se rapportant à un nom composé de longueur 2.
   Cette modification dépend du degré de lexicalisation du nom composé. Par exemple, il est impossible d'utiliser *mise* pour *mise à jour* et *modulation* pour *modulation de fréquence.* Par contre, pour de nombreux cas de noms composés de structure N ADJ, N peut être utilisé comme variante elliptique comme : *débit* pour *débit binaire*, etc.

2. schéma anaphorique où la première unité lexicale réfère à un terme plus long évoqué dans le même paragraphe :

   - *ces réseaux (these networks)* réfère à *réseaux de télécommunications par satellite (satellite telecomunication networks)*,
   - *cette zone* réfère à *zone des équipements de télécommunication (telecomunication equipment area)*

Ces variantes de longueur 1 ne pourront pas être mise en corrélation avec le terme qu'elles désignent mais, elles ne poseront pas de problème d'identification puisqu'elles sont de longueur 1.

Variante elliptique de longueur 2

Les variantes elliptiques de longueur 2 ne devraient être employées que lorsque qu'elles ne génèrent pas d'ambiguïtés sur leur hyponyme. L'utilisation des abréviations est donc beaucoup moins risquée et les textes techniques les emploient plus volontiers. Néanmoins, en voici quelques exemples :

- N_1 Prep N_2 ADJ $\rightarrow$ N_1 Prep N_2
  spectre des fréquences radioélectriques $\rightarrow$ spectre des fréquences

- N_1 ADJ Prep N_2 $\rightarrow$ N_1 Prep N_2
  service fixe par satellite $\rightarrow$ service par satellite

- N_1 ADJ Prep N_2 $\rightarrow$ N_1 ADJ
  service fixe par satellite $\rightarrow$ service fixe

- N_1 ADJ $PREP_1$ N_2 $PREP_2$ DET N_3 $\rightarrow$ N_1 ADJ PREP N_2
  accès multiple avec assignation à la demande $\rightarrow$ accès multiple avec assignation

- N_1 ADJ_1 ADJ_2 PREP N_2 → N_1 ADJ_1 PREP N_2
  conduit numérique fictif de référence → conduit numérique de référence

2.2.4 Conclusion

Cette étude linguistique des noms composés de notre corpus montre que les noms composés techniques ne sont pas des structures morphosyntaxiques figées et qu'ils subissent de nombreuses transformations. L'extraction automatique des noms composés se heurte au problème suivant, qui est celui de la reconnaissance de ces unités lexicales complexes :

- il sera impossible de décider si une séquence morphosyntaxique constitue un nom composé de longueur 2 avant d'avoir identifié les noms composés de longueur 3. Par exemple, la séquence *service fixe* n'est pas un nom composé de longueur 2 mais une sous-séquence du nom composé de longueur 3 *service fixe par satellite*,

- à l'inverse, il sera impossible de décider du statut d'une séquence de longueur 3 avant d'avoir déterminé les noms composés de longueur 2 ; en effet, comment reconnaître dans la séquence *régénération des lobes latéraux* un nom composé si *lobes latéraux* n'a pas été préalablement identifié comme un nom composé de longueur 2.

Il y a là en apparence un cercle vicieux qu'il va falloir briser. La section ci-dessous présente les principales séquences morphosyntaxiques de longueur 2 et 3, caractéristiques des noms composés et résume leurs ambiguïtés structurelles.

Ambiguïtés

Nous avons présenté quelles étaient les structures des noms composés de type élémentaire et comment de nouveaux noms composés se créaient à partir de noms composés attestés. Notre tâche consistant à extraire d'un corpus les noms composés du domaine, nous ne possédons évidement pas la liste des noms composés élémentaires. Sans véritablement dévoiler la manière dont nous allons procéder, nous nous appuierons sur leurs structures morphosyntaxiques. Il reste que ces structures sont ambigues comme l'examen des séquences qui suit le montre :

1. **Séquences de longueur 2**
   Les séquences extraites du corpus et appartenant à l'une des structures morphosyntaxiques de type élémentaire de longueur 2 décrites dans la première partie de cette étude linguistique, représentent l'une des entités ci-dessous :

   - un nom composé de type élémentaire,

- un groupe nominal qui n'a pas le statut de composé,
- une variante elliptique d'un nom composé de longueur ≥ 3,
- une sous-séquence d'un nom composé de longueur ≥ 3

2. **Séquences de longueur 3**
Les séquences morphosyntaxiques de longueur 3 les plus représentées dans nos textes sont énumérées ci-dessous et annotées des différentes analyses quelles peuvent recevoir. À ces analyses, il faut bien entendu rajouter celle d'un groupe nominal libre qui n'a pas de statut de composé.

(a) N *non* ADJ
- Nom composé de type élémentaire de longueur 3 ;
- Nom composé de type élémentaire de longueur 2 modifié par un adverbe négatif *non*; dans ce cas, N ADJ devrait exister sans modifieur ;
- Variante elliptique d'un nom composé de longueur ≥ 4 ;
- Sous-séquence d'un nom composé de longueur ≥ 4.

(b) N ADJ_1 ADJ_2
- Nom composé de type élémentaire de longueur 3 ;
- Nom composé de type élémentaire de longueur 2 de structure N ADJ_1 modifié par un adjectif; N_1 ADJ_1 devrait exister sans modifieur ;
- Variante elliptique d'un nom composé de longueur ≥ 4 ;
- Sous-séquence d'un nom composé de longueur ≥ 4.

(c) N_1 ADJ PREP N_2
- Nom composé de type élémentaire de longueur 3 ;
- Nom composé de type élémentaire de longueur 2 de structure N_1 PREP N_2 modifiée par un adjectif; N_1 PREP N_2 devrait exister sans modifieur ;
- Surcomposé par substitution du N_1; N_1 PREP N_2 et N_1 ADJ doivent être des noms composés de type élémentaire,
- Surcomposé par juxtaposition ; seul N_1 ADJ existe comme composé et pas N_1 PREP N_2 ;
- Variante elliptique d'un nom composé de longueur ≥ 4 ;
- Sous-séquence d'un nom composé de longueur ≥ 4.

(d) N_1 PREP N_2 ADJ
- Nom composé de type élémentaire de longueur 3 ;

- Nom composé de type élémentaire de longueur 2 de structure N_1 PREP N_2 modifié par un adjectif; N_1 PREP N_2 devrait être un nom composé de type élémentaire; l'adjectif doit s'accorder en nombre et en genre avec N_1 ;
- Surcomposé par juxtaposition ; seul N_2 ADJ devrait être un nom composé de type élémentaire et pas N_1 PREP N_2 ;
- Variante elliptique d'un nom composé de longueur ≥ 4 ;
- Sous-séquence d'un nom composé de longueur ≥ 4.

(e) N_1 $PREP_1$ N_2 $PREP_2$ N_3

- Nom composé de type élémentaire de longueur 3 ;
- Surcomposé par substitution du N_2 ; N_1 PREP N_2 et N_2 $PREP_2$ N_3 doivent être des noms composés de type élémentaire;
- Surcomposé par juxtaposition ; soit N_1 $PREP_1$ N_2, soit N_2 $PREP_2$ N_3 devrait être un nom composé de type élémentaire mais pas les deux ;
- Variante elliptique d'un nom composé de longueur ≥ 4 ;
- Sous-séquence d'un nom composé de longueur ≥ 4.

(f) N_1 CONJ ($PREP_1$) (DET) N_2 $PREP_2$ N_3

- Nom composé de type élémentaire de longueur 3 ;
- Coordination de deux noms composés de type élémentaire; N_1 $PREP_1$ N_3 et N_2 $PREP_2$ N_3 doivent être des noms composés de type élémentaire ;
- Variante elliptique d'un nom composé de longueur ≥ 4 ;
- Sous-séquence d'un nom composé de longueur ≥ 4.

Les ambiguïtés attachées aux séquences de longueur 3 sont donc nombreuses. À ce problème s'ajoute celui des variantes des noms composés.

Noms composés et leurs variantes

Pour être parfait, ce travail d'extraction de terminologie devrait indiquer pour chaque nom composé ses variantes. En ce qui concerne les variantes orthographiques, le lien est assez aisé : des procédures simples devraient nous permettre de relier par exemple les structures N_1 N_2 et N_1-N_2. Pour les variantes morphosyntaxiques, le traitement est déjà moins trivial. Certaines variantes seront identifées : celles présentant une variation de préposition ou une simplification/complication de structure. Par contre, comment relier *réseau public commuté* et *réseau public à commutation*? Le critère de la traduction unique est-il suffisant ? Pour conclure, il nous semble très difficile de relier un nom composé à ces variantes elliptiques dans la cadre des ressources monolingues. Il faudra sans

doute faire appel à des informations bilingues pour traiter ce problème.
L'extraction automatique des noms composés d'un texte est donc loin d'être une tache triviale. Cette étude linguistique a permis de faire un tour d'horizon des problèmes à prendre en considération. Nous briserons le cercle vicieux qui demande pour déterminer les noms composés de longueur 3 d'avoir identifié ceux de longueur 2 et pour déterminer ceux de longueur 2 d'avoir identifié ceux de longueur 3, en acceptant l'hypothèse suivante:

> La majorité des noms composés de longueur 3 sont construits à partir de noms composés de longueur 2 par surcomposition, modification ou coordination.

La stratégie adoptée sera donc l'extraction des séquences morphosyntaxiques caractéristiques des types élémentaires de longueur 2, ces structures pouvant être modifiées ou coordonnées. Ces séquences morphosyntaxiques constitueront une liste de noms composés potentiels de longueur 2. Cette liste de candidats sera soumise à divers calculs statistiques. Ces calculs permettront-ils de déterminer le statut de la séquence rencontrée? D'autre part, cette étude a été effectuée conjointement sur le français et sur l'anglais. Un modèle statistique bilingue sera-t-il plus ou moins efficace qu'un modèle uniquement monolingue?

Chapitre 3

Traitement et stockage de corpus

La méthode que nous avons choisie pour extraire d'un corpus des noms composés terminologiques combinant modèles statistiques et données linguistiques, nous avons besoin de reconnaître les séquences morphosyntaxiques caractéristiques des noms composés. Reconnaître ces dernières, implique un traitement préliminaire du corpus, au cours duquel toutes les unités lexicales reçoivent une étiquette grammaticale. De plus, pour obtenir un échantillonnage optimum, il est plus efficace de travailler sur les lemmes que sur les formes fléchies : nous attribuons à toutes les unités lexicales du corpus, leurs lemmes. Un prolongement de ce travail concerne l'extraction de noms composés terminologiques bilingues ; pour cela, le corpus bilingue sera aligné phrases à phrases.
Ce chapitre regroupe donc, dans une première partie, tous les traitements que nous avons fait subir au corpus : le nettoyage, la reconnaissance des unités lexicales, l'assignation d'étiquettes grammaticales et morphologiques, et en vue d'une extraction de noms composés bilingues, l'alignement phrases à phrases.
La deuxième partie présente le modèle de base de données choisie pour stocker les corpus ainsi traités.

3.1 Traitement des corpus

Dans cette partie, nous présentons le traitement préliminaire que nous avons fait subir à nos corpus en vue de l'extraction de ressources lexicales monolingues et bilingues. Nous décrirons successivement, le nettoyage et la synchronisation, l'identification des items, puis les programmes que nous avons utilisés pour effectuer : l'assignation d'étiquettes grammaticales pour le français et l'anglais, l'assignation d'étiquettes morphologiques pour le français et l'anglais, enfin l'alignement de phrases.

Les textes sources ont été fournis par l'ITU (International Telecommunications Union) et appartiennent au domaine des télécommunications. Ces textes sont

disponibles en plusieurs langues et nous avons travaillé sur des textes bilingues anglais-français :

- le "Manuel des Télécommunications par Satellite" (MTS), 200 000 mots en français,
- un extrait du " Livre Bleu du CCITT" (LBC), 800 000 mots en français.

Notre corpus de départ comprend donc 1 million de mots. Le travail de préparation des corpus est un travail d'équipe, réalisé dans le cadre du projet ET-10/63.

3.1.1 Nettoyage et synchronisation des corpus

Le nettoyage d'un corpus consiste à éliminer les parties de textes qui ne sont pas intéressantes pour un traitement linguistique, c'est-à-dire certains caractères de contrôle, les tableaux, les figures, les équations, etc. Les caractères de contrôle sont utilisés par les traitements de texte pour coder par exemple les sauts de pages, les en-têtes et fins de fichier, les espaces insécables. Ces caractères de contrôle sont à identifier et éventuellement à supprimer.

Les corpus fournis comprennent les mêmes textes disponibles en français et en anglais. Néanmoins, le découpage en fichiers des corpus n'est pas synchronisé : par exemple, le corpus MTS français est composé de 37 fichiers alors que le corpus MTS anglais en comprend 41. Un fichier français peut donc correspondre à plusieurs fichiers anglais. Un parallélisme entre les deux corpus au niveau des fichiers est une condition nécessaire pour l'obtention d'un alignement correct des phrases. Pour obtenir des fichiers synchronisés, nous prenons en compte le sectionnement du texte : par exemple, deux fichiers, l'un français, l'autre anglais, formés avec le titre 1.1, contiennnent uniquement le titre et le texte affecté à cette section.

3.1.2 Identification des items

L'identification des items intervient à la suite du nettoyage et du découpage du corpus en fichiers. L'unité de base de notre modèle de base de données étant le mot, le signe de ponctuation, l'apostrophe, etc., il faut découper le corpus en fonction de ces unités. Nous appelons ces unités des items. La reconnaissance des items ainsi que celle des phrases s'effectue grâce à un analyseur qui utilise une grammaire écrite sous forme BNF. Le mot est identifié comme toute suite de caractères comprise entre deux séparateurs typographiques. Les séparateurs typographiques sont le caractère espace, les ponctuations (, .;:!?), les signes (/ ( ) [ ] « » '), le retour chariot, la fin de paragraphe, la fin de fichier, le caractère de tabulation. Les séparateurs phrastiques sont soit un séparateur typographique comme le point, soit une suite de séparateurs typographiques comme deux points suivi d'un retour chariot, etc. Le concept de phrase regroupe n'importe quelle séquence d'item comprise entre deux séparateurs phrastiques ; le

concept de phrase classique, défini comme une séquence d'items commençant par une majuscule et se terminant par un point, est abandonné car il est beaucoup trop restrictif.
Les items sont normalisés et se présentent sous un format unique sans majuscules, ni typographie spéciale. Certains caractères de contrôle présents dans les textes originaux ont été conservés et codés de la manière suivante :

< **EOL** > retour à la ligne
< **TAB** > tabulation
< **ESP** > trois espaces ou plus
< **NL** > saut de ligne
< **FF** > saut de page

Chaque item est accompagné d'une fiche signalétique comportant : sa position dans le corpus (nom du fichier, index de phrase dans le fichier, index d'unité dans la phrase), la manière dont il est séparé de l'item précédant (i.e. son contexte gauche), ses caractéristiques typographiques, et ses caractéristiques graphiques. Détaillons ces trois dernières informations.
Le contexte gauche de l'item est codé ainsi :

s un ou deux espaces
ns aucun séparateur
esp au moins trois espaces
eol retour à la ligne
nl saut de ligne (retour à la ligne + ligne blanche)
tab tabulation
ff saut de page
DEB première unité d'un fichier

L'item n'est précédée d'aucun séparateur (**nf**) lorsqu'il est aggloméré à l'item précédent : c'est le cas de certaines ponctuations, comme par exemple la virgule. Certaines caractéristiques typographiques ont été conservées. Il s'agit des caractères italiques et gras codés par :

g gras
i italique
x ni gras, ni italique (normal)

L'item est décrit en termes de majuscules et minuscules. Le codage utilise une expression régulière de l'alphabet {u, l, +, x} où **u** signifie une lettre en majuscule, **l** une lettre en minuscule, + signifie " la fin de la graphie du mot est identique à la graphie de la lettre précédente", et **x** signifie que le signe rencontré n'est pas une lettre. Quelques exemples :

Item dans le corpus	Item normalisé	Codage de la graphie
CEE	cee	u+
Paris	paris	ul+
MacDonald	macdonald	ullul+

3.1.3 Assignation d'étiquettes grammaticales

À ce stade du traitement, les fichiers des corpus dans les deux langues sont nettoyés, synchronisés, découpés en phrases et contiennent des items normalisés et accompagnés de fiches signalétiques. Chaque item du corpus est maintenant étiqueté. Le programme d'assignation d'étiquettes grammaticales pour le français est celui qu'IBM a utilisé en reconnaissance de la parole et pour l'anglais, CLAWS développé par l'Université de Lancaster. Présentons ces programmes :

Programme d'assignation d'étiquettes grammaticales pour le français

L'étiquette grammaticale d'un mot est assignée par le programme stochastique du Centre de Recherche Scientifique d'IBM développé par l'équipe de recherche sur la reconnaissance de la parole et plus particulièrement par [Dérouault, 1985] et [El-Bèze, 1993]. Ce programme a été décrit dans notre chapitre I (section 1.1). Rappelons brièvement les principaux modules de ce programme :

1. un dictionnaire comprenant 200 000 formes fléchies qui pour chacune d'entre elles indique les étiquettes grammaticales possibles. Chaque étiquette grammaticale est accompagnée d'un score. Par exemple, nous trouvons dans le dictionnaire l'entrée suivante :
   `courant ADJEMS 1 PPRE 1 SUBSMS 41`
   qui signifie que le mot *courant* est soit un adjectif au masculin singulier (`ADJEMS`), soit un participe-présent (`PPRE`) ou encore un nom au masculin singulier (`SUBSMS`) ; l'étiquette la plus probable étant `SUBSMS`.

2. des règles de désambiguïsation lexicale exprimées en termes de biclasses et triclasses.

Le programme d'assignation d'étiquettes grammaticales examine pour un mot les étiquettes données par le dictionnaire et, en fonction du contexte du mot, décide de l'étiquette la plus probable. Ce programme utilise 103 étiquettes syntaxiques différentes dont celles qui nous permettent d'isoler les séquences morphosyntaxiques des termes du domaine. Cette liste d'étiquettes est donnée en annexe A. Ce programme a été amélioré de manière à reconnaître certains mots composés construits à l'aide d'un trait d'union tel que *entrée-sortie*. Pour ceux-ci, le premier élément du mot composé reçoit l'étiquette correspondant à l'étiquette du mot composé, et ses autres éléments une étiquette vide notée `c`. Par exemple, l'adverbe *ci-dessus* est codé ainsi (`ADVE` signifiant adverbe) :

```
ci  ADVE
-   c
dessus  c
```

Cette représentation a été calquée sur la représentation des mots composés anglais produite par CLAWS (voir ci-dessous). Néanmoins, un réel traitement des composés reste à envisager puisque seuls les mots composés construits avec un trait d'union sont pris en compte. Le programme ignore donc les mots composés de plusieurs unités lexicales séparées par des blancs, telles que les prépositions composées (*au lieu de, en fonction de*) pourtant très nombreuses dans le corpus. La non-reconnaissance des mots composés les plus figés génère des erreurs dans notre programme d'extraction de terminologie.
Le programme d'assignation d'étiquettes grammaticales d'IBM était précédemment utilisé sur des textes du domaine médical. Son dictionnaire, quoique conséquent puisqu'il contenait 200 000 mots, comprenait majoritairement des termes médicaux et ne couvrait pas le domaine des télécommunications. Il a donc fallu injecter dans ce dictionnaire certains mots du français courant et un nombre important de mots spécifiques de notre domaine. Les mots que nous avons rajoutés ont, soit été empruntés à d'autres dictionnaires appartenant à IBM, soit été décrits manuellement ; dans ce cas, et lorsque le mot est ambigu, nous avons estimé le poids de chaque étiquette.
Le dictionnaire ayant été complété et corrigé pour certaines entrées, le programme d'assignation d'étiquettes grammaticales a de nouveau été appliqué. Aucune évaluation sur les performances de ce programme n'a été faite et nous rappelons simplement le taux de 95 % d'assignations correctes donné par [Dérouault et Merialdo, 1986].

Programme d'assignation d'étiquettes grammaticales pour l'anglais

Le programme d'assignation d'étiquettes grammaticales utilisé pour l'anglais a été développé à l'université de Lancaster ([Garside *et al.*, 1987]). Ce programme a lui aussi été décrit dans la section 1.1 du chapitre I . Il utilise un analyseur morphologique qui assigne les étiquettes grammaticales possibles en fonction de la terminaison du mot. Des règles de désambiguïsation lexicale exprimées en terme de biclasses et triclasses choisissent ensuite, parmi ces étiquettes, celle qui est la plus probable. Le nombre d'étiquettes grammaticales s'élève à 169 parmi lesquelles certaines ne seront pas utiles pour notre application. Cette liste d'étiquettes est donnée en annexe B. Les mots composés reçoivent un traitement particulier: CLAWS possède un dictionnaire de formes figées et à chaque fois qu'une forme figée est rencontrée dans une phrase, des probabilités additionnelles sont calculées de manière à déterminer si la forme figée est plus ou moins probable que la forme libre associée. Les formes figées prises en compte peuvent être construites à l'aide de traits d'union ou de blancs. Le codage du nom composé

est le suivant :

1. lorsque le mot composé est construit avec un trait d'union, la première unité lexicale du composé reçoit l'étiquette grammaticale du composé et ses autres éléments l'étiquette vide notée c. Le nom *re-use* est codé ainsi (NN1 signifiant nom commun au singulier) :

   ```
   re NN1
   - c
   use c
   ```

2. lorsque le mot composé comprend plusieurs unités lexicales séparées par des blancs, chaque unité lexicale reçoit l'étiquette grammaticale du composé, accompagnée d'un index indiquant la relation de l'étiquette avec le reste de la séquence. Par exemple, la préposition composée *according to* est codée de la manière suivante (IN signifiant préposition) :

   ```
   according IN21
   to IN22
   ```

   IN21 signifie que le mot *according* est la première unité lexicale d'une préposition composée de deux unités lexicales ; IN22 signifie que *to* en est la deuxième unité lexicale. D'une même façon, l'adverbe *none the less* sera codé (RB correspondant à l'étiquette adverbe) :

   ```
   none RB31
   the RB32
   less RB33
   ```

CLAWS, à la différence du programme d'IBM, ne possède pas d'étiquette réservée aux mots inconnus puisque l'analyseur effectue des prédictions sur la catégorie grammaticale du mot par rapport à sa terminaison. Néanmoins, des étiquettes pour les noms propres existent : elles sont assignées aux noms dont la première lettre est une majuscule et qui n'apparaissent pas en début de phrase. Cette définition du nom propre engendre des erreurs qui ont été relevées sur le corpus étiqueté. Le seul moyen pour éviter que ces erreurs ne se reproduisent, consiste à enregistrer ces noms propres ou ces faux noms propres dans un dictionnaire. CLAWS ne possédant pas de dictionnaire prévu à cet effet, la solution *ad hoc* retenue a été l'intégration des noms propres de nos textes dans le dictionnaire des idiomes. L'étiquetage des noms propres ayant été vérifié et corrigé, une estimation a été faite sur 100 phrases et a donné un taux de 98 % de bonnes assignations pour tous les mots.

Programme d'assignation d'étiquettes de partie du discours pour le français et l'anglais

Afin de faciliter la lecture et l'interprétation des étiquettes grammaticales, nous avons introduit des étiquettes supplémentaires caractérisant les parties du discours. Une étiquette de partie du discours comme par exemple `NOM` regroupera les étiquettes grammaticales françaises : `SUBSMS` (nom masculin singulier), `SUBSMP` (nom masculin pluriel), etc., et les étiquettes grammaticales anglaises : `NN1` (nom singulier), `NN2` (nom pluriel), etc. Les étiquettes de partie du discours sont donc communes au français et à l'anglais et facilitent les opérations de recherche dans la base de données. La liste des étiquettes de partie du discours est la suivante :

ADJ adjectif

ADV adverbe

CHIF chiffre

DET article

NPRO nom propre

NOM nom

PONC ponctuation

PREP préposition

PRO pronom

VER verbe

Ces étiquettes sont néanmoins trop générales pour l'extraction de terminologie.

3.1.4 Assignation d'étiquettes morphologiques

De manière à obtenir de meilleures statistiques, il est préférable de travailler sur les lemmes plutôt que sur les formes fléchies. Ces lemmes sont déduits de la forme fléchie et de l'étiquette grammaticale assignée automatiquement, à l'aide des programmes décrits ci-dessous.

Programme d'assignation d'étiquettes morphologiques pour le français

L'étiquetage morphologique du français est réalisé par un programme appartenant lui aussi à l'équipe de recherche sur la reconnaissance de la parole d'IBM ([El-Bèze, 1993]). Ce programme utilise un dictionnaire de formes fléchies associées à une étiquette grammaticale, de 360 000 entrées. Ce programme n'a été utilisé que pour les noms, les verbes et les adjectifs. Pour les autres catégories grammaticales, le lemme assigné est le mot lui-même. Il n'y a généralement qu'un

lemme correspondant à une forme fléchie dès lors que sa classe grammaticale est connue : par exemple, la forme fléchie *couvent* associée à l'étiquette grammaticale SUBSMS (nom masculin singulier) n'a comme lemme possible que *couvent*. Le programme ne résout pas les ambiguïtés lorsque plusieurs lemmes sont possibles et se contente d'indiquer les divers lemmes candidats indiqués dans le dictionnaire comme le montrent les exemples suivants extraits de notre corpus :

Forme fléchie	Classe grammaticale	Lemmes associés
cours	substantif masculin pluriel	*cours/cour*
exitatrice	substantif féminin singulier	*exitatrice/exitateur*
faut	verbe à la 3^{me} personne du singulier	*faillir/falloir*
fois	substantif féminin pluriel	*foi/fois*
fondent	verbe à la 3^{me} personne du pluriel	*fonder/fondre*
frais	substantif masculin pluriel	*frai/frais*
matériaux	substantif masculin pluriel	*matériaux/matériau*
mise	participe passé féminin singulier	*miser/mettre*
seconde	substantif féminin singulier	*second/seconde*

Plusieurs de ces ambiguïtés sont suspectes, voire fausses : le verbe *faillir* peut être considérer comme obsolète en français contemporain, le poisson *frai* ne doit pas être particulièrement priser des textes techniques, et il nous semble que le participe passsé du verbe *miser* est plutôt *misée* que *mise*.
Pour les noms ou adjectifs composés construits avec un trait d'union, le lemme associé est très approximatif, voire erroné. Le calcul du lemme du composé s'effectue en calculant le lemme du premier élément et en associant aux éléments suivants leurs formes fléchies comme lemmes. Par exemple, le nom composé *court-circuit* reçoit le lemme *court-circuit*, sous le format suivant où le premier champ réfère à la forme fléchie, le deuxième au lemme et le troisième à l'étiquette grammaticale (SUBSMS signifiant substantif masculin singulier) :

```
court  court  SUBSMS
-  -  c
circuit circuit c
```

Dans l'exemple ci-dessus, le lemme associé est correct, bien qu'il faille le reconstituer, mais ce cas est plutôt une exception. Il suffit que l'un des éléments du mot composé associé à l'étiquette vide soit fléchi (soit en nombre pour un nom, soit en genre ou/et nombre pour un adjectif) pour que le lemme soit incorrect. Par exemple, le nom composé *convertisseurs-élévateurs* au pluriel se voit associé le lemme *convertisseur-élévateurs* (SUBSMP signifiant substantif masculin pluriel) :

```
convertisseurs  convertisseur  SUBSMP
-  -  c
```

```
élévateurs  élévateurs  c
```

Un autre problème rencontré avec ce calcul du lemme concerne les noms ou adjectifs composés à l'aide d'une préposition. Seul le lemme du premier élément du composé est calculé mais comme il l'est à l'aide de l'étiquette grammaticale qui caractérise le composé en entier, la préposition est considérée comme un nom et par conséquent, se voit régularisée. Ainsi, dans l'exemple ci-dessous, le nom composé *sous-système* se voit attribuer le lemme *sou-système*:

```
sous  sou  SUBSMS
-  -  c
système système c
```

Ce traitement peu orthodoxe s'explique par l'absence de dictionnaire de mots composés: la forme fléchie du premier élément du composé est recherchée dans le dictionnaire des mots simples associé à la classe grammaticale du composé. Il aurait peut-être mieux valu ne pas tenter de lemmatiser le mot composé, plutôt que d'obtenir, dans le meilleur des cas, ces moitiés de lemme. Le problème du calcul des lemmes des mots composés s'inscrit dans le problème plus général de l'absence de traitement des mots composés dans les programmes d'IBM, que cela soit au niveau de l'étiquetage grammatical ou morphologique. Les mots composés avec trait d'union ne sont donc pas correctement lemmatisés et les statistiques les accompagnant sont faussées. Heureusement, ceux-ci n'étaient pas trop nombreux dans nos textes.

Programme d'assignation d'étiquettes morphologiques pour l'anglais

L'étiquetage morphologique de l'anglais est réalisé par le programme de l'Université de Lancaster ([Garside *et al.*, 1987]). Ce programme examine l'étiquette grammaticale attribuée précédemment par CLAWS. À chaque étiquette grammaticale est associé un ensemble de règles. Chaque règle comporte deux terminaisons:

- la terminaison qui doit s'unifier avec celle de la forme fléchie, et dont celle-ci sera amputée,
- la terminaison qu'il va falloir rajouter à la racine pour obtenir le lemme correspondant.

Lorsque la terminaison de la forme fléchie s'unifie avec plusieurs suffixes, c'est le suffixe le plus grand qui est retenu. Par exemple, les règles chargées de construire le lemme des verbes anglais rencontrés à la troisième personne du singulier sont les suivantes:

Terminaison de la forme fléchie	Terminaison du lemme	Exemples
ches	*ch*	*reaches* → *reach*
shes	*sh*	*flashes* → *flash*
ies	*y*	*studies* → *study*
sses	*ss*	*passes* → *pass*
xes	*x*	*relaxes* → *relax*
s		*reads* → *read*

Les exceptions à ces ensembles de règles sont enregistrées dans un dictionnaire. Le programme le consulte en premier et si la forme fléchie accompagnée de son étiquette grammaticale n'y est pas présente, les terminaisons sont examinées. Aucune indication n'est fournie sur le traitement des formes fléchies ambiguës et, d'ailleurs, notre corpus anglais lemmatisé n'en possède pas, soit le nombre d'étiquettes grammaticales plus important que pour le français permet de lever ce type d'ambiguïté, soit l'anglais a la chance d'être une langue morphologiquement non ambiguë !
Le lemme des mots composés est calculé de la manière suivante :

1. lorsque le mot composé est construit avec un trait d'union, le premier élément du composé reçoit le lemme du composé et ses autres éléments l'étiquette vide `c`. Le participe passé *re-used* est codé ainsi (`VVN` signifie participe-passé) :

   ```
   re  re-use  VVN
   -  c  c
   used c  c
   ```

2. lorsque le mot composé comprend plusieurs unités lexicales séparées par des blancs, aucune lemmatisation ne s'effectue et le lemme de chaque élément du composé est égal à sa forme fléchie. Ainsi la préposition composée *depending on* est lemmatisée de la manière suivante (`II` signifie préposition) :

   ```
   depending depending II21
   on  on  II22
   ```

Nous ne pouvons que remarquer l'attention que CLAWS a porté au traitement des mots composés et à espérer que les programmes stochastiques pour le français feront un jour aussi bien.

3.1.5 Alignement de phrases

Nous considérerons que deux éléments sont alignés si l'un des éléments est une traduction de l'autre. L'alignement s'applique aux textes, aux paragraphes, aux phrases, aux sous-phrases et aux mots. Les textes de départ étant bilingues, ils

sont considérés comme alignés. L'alignement des paragraphes, appelé synchronisation, a été décrit dans la section 3.1.1. Le programme utilisé pour l'alignement des phrases à partir de nos corpus synchronisés est celui de [Brown *et al.*, 1991] que nous avons présenté dans notre chapitre I (voir section 1.4). Ce programme qui émet, pour chaque couple de phrases candidates à l'alignement, une probabilité calculée sur la longueur de ces phrases en termes de nombre de mots, a été modifié de manière à prendre en compte des informations linguistiques de type *cognate*. Les *cognates* sont des couples de mots partageant certains traits typographiques. [Simard *et al.*, 1992] les ont définis ainsi :

> "Un couple de *cognates* est un couple de mots bilingues vérifiant l'un des critères suivants :
>
> - les deux mots ne contiennent que des lettres (pas de chiffres) et ont les mêmes quatre premiers caractères,
> - les deux mots ont au moins un caractère numérique et sont identiques,
> - les deux mots sont un caractère de ponctuation et sont identiques."

Nous avons élargi cette définition en rajoutant le point suivant qui n'est valable que pour le couple de langues français-anglais :

> les deux mots ne contiennent que des lettres et partagent une terminaison en *ion*

Les couples de *cognates* ne sont recherchés dans les deux phrases candidates à l'alignement que lorsque le programme probabiliste de [Brown *et al.*, 1991] ne parvient pas à trancher. Cette approche combinée permet d'une part, de conserver un temps rapide d'exécution puisque la plus grande partie de l'alignement s'effectue grâce au modèle probabiliste, d'autre part d'améliorer sensiblement la qualité et le nombre de bons alignements obtenus. Le corpus aligné phrases à phrases se présente de la manière suivante : les paires de phrases alignées sont séparées les unes des autres par le séparateur : `###############` suivi de l'index du fichier, de l'index de la phrase anglaise dans le fichier, de l'index de la phrase française dans le fichier et d'un entier indiquant le type d'alignement rencontré. Cet entier prend une valeur dans l'intervalle [1,6] correspondant à :

alignement codé : (1) alignement 1-1

Exemple : `############### 5790 : 121 - 124 (1)`

Une phrase dans une langue correspond à une phrase dans une autre langue. Dans l'exemple ci-dessus, la phrase anglaise d'index `121` correspond à la phrase française d'index `124`. Les deux phrases apparaissent sous le délimiteur.

alignement codé :(2) alignement 1-0

Exemple : `################ 5796 : 26 - 27 (2)`

Une phrase en anglais n'a pas d'équivalent en français. Dans ce cas, l'index de la phrase française (27) est égal à l'index de l'avant-dernier alignement et seule la phrase anglaise apparaît sous le délimiteur.

alignement codé : (3) alignement 0-1

Exemple : `################ 5914 : 30 - 33 (3)`

Une phrase en français n'a pas d'équivalent en anglais. Dans ce cas, l'index de la phrase anglaise (30) est égal à l'index de l'avant-dernier alignement et seule la phrase française apparaît sous le délimiteur.

alignement codé : (4) alignement 2-1

Exemple : `################ 5790 : 4 - 3 (4)`

Deux phrases anglaises contiguës correspondent à une phrase française. L'index de la phrase anglaise prend la valeur de l'index de la deuxième phrase anglaise en jeu dans cet alignement. L'exemple ci-dessus est à interpréter de la manière suivante : dans le fichier d'index 5790, les phrases 3 et 4 anglaises s'alignent avec la phrase 3 française. Les deux phrases anglaises contiguës sont concaténées pour ne former qu'une seule phrase et apparaissent accompagnées de la phrase française sous le délimiteur.

alignement codé : (5) alignement 1-2

Exemple : `################ 5789 : 1 - 2 (5)`

Une phrase anglaise correspond à deux phrases françaises contiguës. L'index de la phrase française prend la valeur de l'index de la deuxième phrase française en jeu dans cet alignement. L'exemple ci-dessus est à interpréter de la manière suivante : dans le fichier d'index 5789, la phrase 1 anglaise s'aligne avec les phrases 1 et 2 françaises. Les deux phrases françaises sont concaténées pour ne former qu'une seule phrase et apparaissent accompagnées de la phrase anglaise sous le délimiteur.

alignement codé : (6) alignement 2-2

Exemple : `################ 5803 : 34 - 34 (6)`

Deux phrases anglaises correspondent à deux phrases françaises sans qu'il y ait pour autant d'alignements 1-1 entre les quatre phrases. Les index indiqués pour les phrases françaises et anglaises sont les index des deux deuxièmes phrases en jeu dans cet alignement. L'exemple ci-dessus est à interpréter de la manière suivante : dans le fichier d'index 5803, les phrases 33 et 34 anglaises s'alignent avec les phrases 33 et 34 françaises. Les deux phrases françaises comme les deux phrases anglaises sont concaténées pour ne former qu'une seule phrase et apparaissent sous le délimiteur.

Chaque item apparaissant dans ces phrases est accompagné de son lemme et de son étiquette grammaticale. Voici un extrait du corpus LBC aligné phrases à

phrases :

```
################ 5790 : 121 - 124 (1)

1977_1977_MC (_(_( june_june_NP1 )_)_) :_:_: establishment_establishment_NN
of_of_IO the_the_AT EUTELSAT_eutelsat_NNJ organization_organization_NNJ
with_with_IW 17_17_MC administrations_administration_NN2 as_as_RG
initial_initial_JJ signatories_signatory_NN2 ._._.

1977_1977_NPRO (_(_YAAA juin_juin_SUBSMS )_)_YAAA :_:_YAAA
création_création_SUBSFS ,_,_YAAA avec_avec_PREP 17_17_NPRO
administrations_administration_SUBSFP fondatrices_fondateur_ADJEFP
signataires_signataire_SUBSFP ,_,_YAAA de_de_PDEA l@'_l@'_DETRFS
organisation_organisation_SUBSFS EUTELSAT_eutelsat_XSOC ._._AAAA

################ 5790 : 123 - 125 (4)

1977_1977_MC (_(_( aug_aug_NPM1 ._._. )_)_) :_:_: launching_launching_NN1
of_of_IO the_the_AT sirio_sirio_NP1 satellite_satellite_NN1
(_(_( italy_italy_NP1 )_)_) ._._. first_first_MD experimental_experimental_JJ
communication_communication_NN1 satellite_satellite_NN1 using_use_VVG
frequencies_frequency_NN2 above_above_II 15_15_MC ghz_ghz_NP1
(_(_( 17_17_MC /_/_CC 11_11_MC ghz_ghz_NP1 )_)_) ._._. 1977_1977_NPRO
(_(_YAAA aot_aot_SUBSMS )_)_YAAA :_:_YAAA

lancement_lancement_SUBSMS du_du_PREPMS satellite_satellite_SUBSMS
sirio_sirio_NPRO (_(_YAAA italie_italie_XPAYFS )_)_YAAA ,_,_YAAA
premier_premier_ADJEMS satellite_satellite_SUBSMS
expérimental_expérimental_ADJEMS de_de_PDEA
télécommunication_télécommunication_SUBSFS
utilisant_utiliser_PPRE des_des_PDES fréquences_fréquence_SUBSFP
supérieures_supérieur_ADJEFP à_à_PDEA 15_15_NPRO ghz_ghz_SUBSMS
(_(_YAAA 17_17_NPRO /_/_YAAA 11_11_NPRO ghz_ghz_SUBSMS )_)_YAAA ._._AAAA
```

Nous avons évalué ces alignements en extrayant au hasard un échantillon de cent paires de phrases alignées et nous avons recensé un pourcentage d'erreur de 2 %. Nous rappelons que chaque unité du corpus, accompagnée de sa fiche signalétique et des ses étiquettes, est intégrée dans la base de données, de même que les alignements de phrases. Nous n'avons pas utilisé la base de donnée pour notre application, mais il nous a paru intéressant de présenter un modéle de stockage bilingue de gros corpus. La section qui suit présente donc le modéle que nous avons choisi, et peut être considéré comme une parenthése technique à notre travail.

3.2 Un modèle de base de données pour les corpus bilingues

La question que nous abordons ici concerne le modèle de base de données où est stocké le corpus traité. Nous présentons nos conclusions sur l'adéquation du stockage en base de données relationnelle. Ceci est le fruit d'un travail présenté in extenso dans [Daille et McEnery, 1993]. Il faut bien noter que le choix de tel ou tel mode de stockage est indépendant des techniques d'extraction et de reconnaissance présentées dans le chapitre IV.

3.2.1 Présentation générale

Rappelons les caractéristiques de notre corpus :

- il est bilingue (anglais-français),
- aligné phrases à phrases (chaque phrase anglaise est mise en correspondance avec une phrase française qui est sa traduction)
- annoté (chaque item du corpus est accompagnée de sa classe grammaticale, son lemme, etc.)
- d'une taille importante (environ 1 million de mots dans chaque langue).

Les corpus sont généralement stockés en machine sous le format fichier standard. Sous ce format, dans lequel le corpus est divisé en fichiers textes, seul le nom du fichier peut être utilisé pour organiser le stockage. Par exemple, le corpus LOB ([Johansson *et al.*, 1978]) est réparti dans plusieurs fichiers, chaque fichier reflétant un type spécifique de textes. La seule organisation du corpus LOB est donc une division de base. Ce format impose des limites aux fonctionalités du système. Un corpus stocké se doit de présenter une réelle adéquation fonctionnelle. Les avantages d'une base de données sur un système de fichiers sont les suivants :

- toutes les données implicites ou explicites du corpus, c'est-à-dire les mots, mais aussi les étiquettes grammaticales, les lemmes ou autres éléments d'informations, sont présentes dans la base de données d'une manière non redondante,
- les données sont les mêmes pour les programmeurs et les utilisateurs,
- les données sont structurées, ce qui facilitera le calcul de mesures statistiques simples,
- le format est unique pour toutes les données du corpus (élargi éventuellement à d'autres corpus si ce format est reconnu comme standard),

- la base de données possède des modes d'interrogation qui permettront de consulter le corpus, suivant des critères décidés par l'utilisateur,

- le langage SQL, standard pour un grand nombre de bases de données, permet d'écrire des librairies d'exploitation de la base qui sont portables.

Ces éléments soulignent l'adéquation fonctionnelle de la base de données, que son utilisateur soit linguiste ou informaticien. Il est important de préciser que la motivation première pour le stockage des corpus en base de données repose sur une utilisation et une consultation importantes de ceux-ci. Le modèle que nous allons présenter peut être modifié ou étendu dans le cas où d'autres applications seraient envisagées. L'adéquation formelle est un concept important à prendre en compte pour le stockage du corpus. Elle s'obtient en imposant au corpus d'être stocké :

- de la manière la plus efficace qui soit du point de vue de l'usage envisagé,

- d'une façon qui assure que toutes les données sont accessibles, i.e. aucune n'est perdue.

La première étape consiste à concevoir notre base de données. Les données issues du corpus et des programmes périphériques (programme d'assignation d'étiquettes grammaticales, etc.) sont modélisées suivant le Modèle Entité/Association (MEA) normalisé, puis ce modèle est transformé de manière à produire un Modèle Relationnel. Tous les concepts qui peuvent être modélisés dans un MEA normalisé peuvent être représentés dans un modèle relationnel, sans redondance.

3.2.2 Un Modèle Entité/Association (MEA)

Un MEA rassemble entités et associations. Ces entités et associations sont conçues ensembles et le résultat est une série de spécifications pour le stockage des données. Ces spécifications sont ensuite rassemblées dans un schéma entité/association.

3.2.2.1 Les entités

Les objets qui sont clairement délimités sont appelés les entités. Nous avons développé cinq types d'entités pour le stockage d'un corpus multilingue :

- le corpus,

- les items (mots, signes de ponctuations, etc.),

- les lemmes,

- les étiquettes grammaticales,

- les fichiers.

Nous n'avons pas introduit d'entité de type caractère ; les caractères sont considérés comme subordonnés aux entités de type item.
Chaque entité possède des propriétés appelées attributs. Les valeurs prises par cet attribut constituent son domaine. Décrire les domaines des attributs d'une entité n'est pas toujours nécessaire : le domaine d'un attribut est parfois suffisant pour caractériser complètement l'entité, par exemple une entité de type étiquette grammaticale est entièrement caractérisée par son étiquette grammaticale. Cet attribut est appelé attribut-clé. L'ensemble des valeurs d'un attribut-clé, par exemple la liste des étiquettes grammaticales, est appelé la clé d'un type d'entité. Une clé ne doit pas comporter de redondance et ne doit pas être nulle. De manière à clarifier ces concepts, il est utile d'examiner les différents attributs des types d'entités développées et leurs clés. La description de chaque entité est accompagnée d'un schéma qui résume ses attributs : l'entité y est représentée par un rectangle, et ses attributs par des cercles ; le cercle en gras correspond à l'attribut-clé.

Corpus

Une entité de type **Corpus** se caractérise par deux ensembles de données :

1. les données déduites directement des textes,
2. les données obtenues avec les programmes d'assignation d'étiquettes grammaticales et d'assignation d'étiquettes morphologiques qui caractérisent chaque item du corpus.

Les attributs qui correspondent aux premières données sont :

- l'item (les mots, les signes de ponctuations, les chiffres, ...),
- la position de l'item dans le corpus (fichier, phrase, item),
- la typographie,
- le contexte gauche de l'item.

Les attributs qui correspondent aux secondes données sont :

- l'étiquette grammaticale,
- la partie du discours associée à chaque étiquette grammaticale (POS),
- le lemme.

Quoique chaque entité de type **Corpus** soit définie de manière unique, aucune clé primaire n'apparaît, i.e. aucune des valeurs des attributs ci-dessus n'est unique. Une clé primaire pour chaque entité de type **Corpus** est à introduire : cet attribut correspondra à l'index de l'item dans le corpus. Les attributs de l'entité de type **Corpus** sont résumés dans la Figure 3.1.

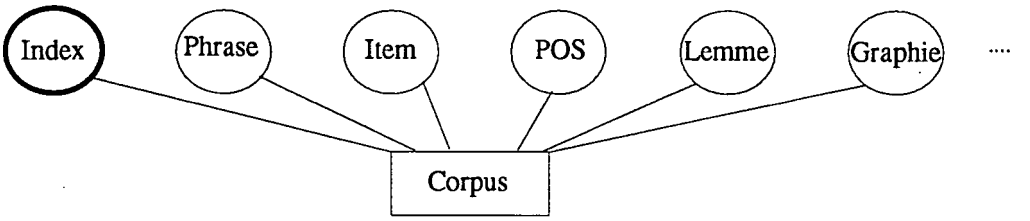


FIG. 3.1 - Corpus

Item et Lemme

L'entité de type **Item** et l'entité de type **Lemme** ne possèdent chacune qu'un attribut qui pourra être leur clé (voir Figure 3.2).

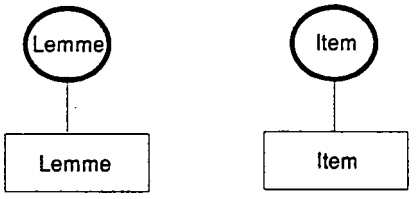


FIG. 3.2 - Lemme et Item

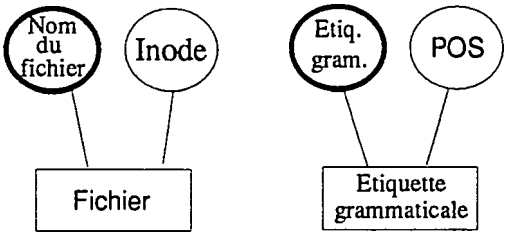


FIG. 3.3 - Fichier et Étiquette grammaticale

Fichier et Étiquette grammaticale

Chaque entité de type **Fichier** est associé à un identificateur UNIX ; ses deux attributs sont donc le nom du fichier qui pourra éventuellement servir de clé et son identificateur Unix (inode).
Chaque entité de type **Étiquette grammaticale** est associée à une étiquette de partie du discours ; ses deux attributs sont donc l'étiquette grammaticale qui

pourra éventuellement servir de clé et l'étiquette de partie du discours (POS). Ces deux entités sont représentées dans la Figure 3.3.

3.2.2.2 Les associations

Les entités ne sont indépendantes : par exemple, l'entité de type **Corpus** est liée à l'entité de type **Étiquette grammaticale** puisque chaque mot du corpus a reçu une étiquette grammaticale. La représentation d'un lien entre plusieurs entités est appelée "association". Si une clé a été définie pour chacune des entités participant à l'association, l'association sera définie grâce aux valeurs de ces clés. Les associations identifiées pour le corpus monolingue entre les différentes entités développées sont :

Association (Corpus, Item)

Le corpus est composé d'items qui peuvent apparaître plusieurs fois dans le corpus, mais un item doit apparaître au moins une fois pour être identifiée comme tel. L'association (Corpus, Item) est une association fonctionnelle ou encore N:1 : chaque entité de type **Corpus** est reliée à une et une seule entité de type **Item** ; inversement, chaque entité de type **Item** est associée à au moins une entité de type **Corpus**.

Association (Corpus, Fichier)

Le corpus est divisé en un certain nombre de fichiers dont aucun n'est vide et l'ensemble des fichiers représente le corpus. Chaque entité de type **Corpus** est liée à une et une seule entité de type **Fichier** ; inversement, chaque entité de type **Fichier** est associée à au moins une entité de type **Corpus**. Cette association est de type N:1.

Association (Corpus, Étiquette grammaticale)

Rappelons que chaque item du corpus a reçu une étiquette grammaticale, mais hors contexte, un même item a souvent plusieurs étiquettes grammaticales possibles. Par exemple, le mot *compte* dans la phrase :

L'installation compte habituellement deux canaux.

est un verbe et reçoit l'étiquette `VERB3` (signifiant verbe à la troisième personne du singulier). Par contre, *compte* dans la phrase :

IMMARSAT les exploitera pour son propre compte.

est un nom et reçoit l'étiquette `SUBSMS` (signifiant nom masculin singulier). Le programme d'assignation d'étiquettes grammaticales a choisi parmi les différentes

étiquettes possibles d'un mot en examinant son contexte et chaque item dans le corpus ne reçoit donc qu'une et une seule étiquette.
Chaque entité de type **Corpus** est liée à une et une seule entité de type **Étiquette grammaticale** ; inversement, chaque entité de type **Étiquette grammaticale** est associée à zéro, une ou plusieurs entités de type **Corpus**. Cette association est encore de type N:1.

Association (Corpus, Lemme)

Rappelons que chaque item rencontrée dans le corpus a été lemmatisée. Ce lemme correspond soit à la forme canonique de l'item (pour les noms, les verbes et les adjectifs), soit à l'item lui-même (pour les adverbes, les prépositions, ...). Un item peut avoir plusieurs lemmes possibles comme l'exemple ci-dessus le montre : dans la première phrase, le lemme de *compte* est *compter* et dans la deuxième *compte*. L'association d'un mot et d'une étiquette grammaticale permet néanmoins de n'obtenir qu'un seul lemme pour chaque couple (mot, étiquette grammaticale) à quelques exceptions près. Ces exceptions énumérées dans la section 3.1.4 et qui concernent principalement les verbes ne seront pas prises en compte : les lemmes associés seront considérés comme une seule unité. Nous considérons donc que chaque entité de type **Corpus** est liée à une et une seule entité de type **Lemme**, inversement chaque entité de type **Lemme** est associée à au moins une entité de type **Corpus**. Cette association est de type N:1.

3.2.2.3 Schéma

Le schéma d'entité/association est construit avec les éléments suivants :

- une entité est représentée par un rectangle et les attributs par des cercles (le cercle en gras représente la clé). Chaque cercle est relié au rectangle par un trait plein. Les domaines ne sont pas marqués.
- une association est représentée par un losange et est reliée aux entités par des traits pleins.

Ce schéma est présenté dans la Figure 3.4.

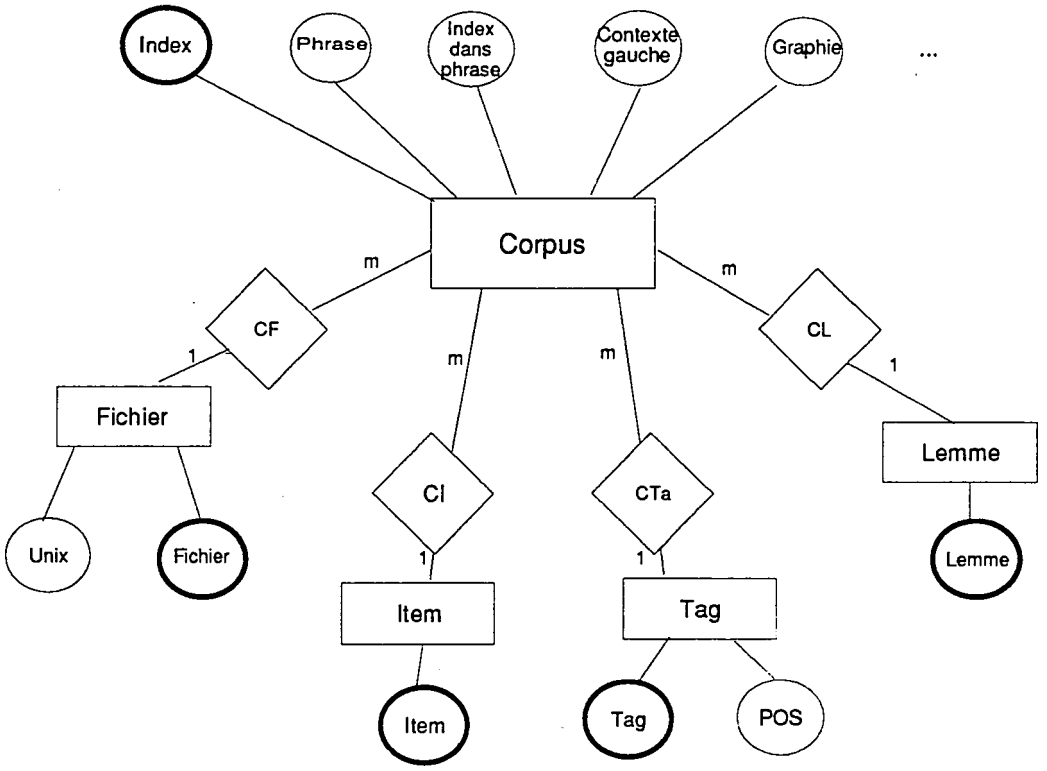


FIG. 3.4 - Schéma entité/association

3.2.3 Modèle relationnel

Le modèle entité/association élaboré précédemment est normalisé : toutes les associations sont de type N:1. Un modèle relationnel peut être déduit automatiquement : les types d'entités deviennent des tables, une entité correspondant à un enregistrement d'une table et les associations se traduisant par des liens entre les différentes tables. Nous verrons ci-dessous comment sont exprimés ces liens. Si ce modèle suffit pour stocker les corpus monolingues, il reste à prévoir le stockage des alignements phrase par phrase. Cette table est annoncée dans la section 3.2.3.2.

3.2.3.1 Les tables monolingues

- le corpus est divisé en fichiers :
  - la table **Fichier**
- la table initiale contient toutes les items qui apparaissent dans le corpus :
  - la table **Corpus**

- les tables construites avec les informations apportées par le programme d'assignation d'étiquettes grammaticales et le programme d'assignation d'étiquettes morphologiques :
  - la table **Tag** pour les étiquettes grammaticales,
  - la table **Lemme**

La table **Corpus** ne comporte pas de redondances grâce aux tables suivantes :

- la table **Fichier**
  Chaque fichier sera associé à un identificateur numérique. C'est cet identificateur qui apparaît dans la table Corpus.

- la table **Item** pour les items
  Un item peut apparaître plusieurs fois dans le corpus et ainsi être enregistrée plusieurs fois dans la table **Corpus**. La table **Item** ne contient qu'une entrée par item, que celle-ci ait été rencontrée une ou plusieurs fois dans le corpus. À cet item unique est associé un identificateur numérique de type entier. C'est cet entier qui apparaît dans la table **Corpus**.

- la table **Tag** pour les étiquettes grammaticales
  Chaque étiquette grammaticale est associée à un entier qui la remplace dans la table **Corpus**.

- la table **Lemme**
  Un lemme est associé à un entier qui le représente dans la table **Corpus**.

Ces différentes tables sont les implémentations des entités du modèle entité/association. Les liens avec la table **Corpus** présentés dans le schéma d'entité/association sont traduits grâce à l'utilisation de ces identificateurs uniques, souvent appelés dans la littérature des clés étrangères.

3.2.3.2 Les tables bilingues

La table contenant les résultats du programme d'alignement de phrase :

- la table **Synchronisation des phrases**
  Cette table permet de retrouver les alignements de phrases bilingues éffectués précedemment. Pour chaque fichier, les alignements phrase par phrase sont enregistrés en utilisant leur numéro et le type d'alignement (voir section 3.1.5) Dans cette table, les différentes phrases d'un fichier sont associées aux phrases anglaises correspondantes.

Chapitre 4

Méthodologie utilisée pour l'extraction automatique de noms composés terminologiques

Nous présentons dans ce chapitre la méthode que nous avons choisie pour extraire d'un corpus des noms composés terminologiques, qui se propose de combiner modèles statistiques et données linguistiques.
Dans la première partie, nous développons notre méthodologie et définissons notre objectif. Nous prenons en compte les résultats de l'étude linguistique du chapitre II et exposons quels sont les noms composés qui paraissent s'adapter à une démarche quantitative.
Dans la deuxième partie, nous décrivons le programme chargé de d'extraire les co-occurrences et de relever leur fréquence: ce programme utilise des patrons sous automate. Après une brève description de ce programme, nous examinons les co-occurrences relevées et nous commentons les résultats obtenus.
La troisième partie expose les modèles statistiques testés, puis l'évaluation graphique qui permet de n'en retenir que quelques uns.
Dans la quatrième partie, les classements proposés par ces modèles sont examinés, ainsi que les informations apportées par la diversité et les mesures de distances. Un seul critère sera alors retenu.
Enfin, la dernière partie présente l'extension de la méthodologie à d'autres ressources lexicales: l'extraction de terminologie bilingue et l'extraction de structures argumentales.

4.1 Principes de la méthodologie

L'objectif de notre travail consiste à définir une méthode pour l'automatisation de l'extraction de noms composés terminologiques à partir de corpus en utilisant des modèles statistiques. Cette méthode est définie à partir de corpus

monolingues. Nous verrons, à la fin de ce chapitre, section 4.6, comment on peut l'étendre à la terminologie bilingue.
Nous avons vu dans notre chapitre I que les modèles statistiques étaient maintenant très présents dans le traitement du langage naturel et qu'ils donnent de bons résultats dans certains domaines comme l'assignation d'étiquettes grammaticales. Dans le cadre plus spécifique de l'extraction des ressources lexicales monolingues, nous avons présenté les travaux de [Lafon, 1984] sur le français, de [Calzolari et Bindi, 1990] sur l'italien et de nombreux autres sur l'anglais, comme par exemple [Church et Hanks, 1990] (section 1.3). Nous avons vu que certains modèles statistiques appliqués à des corpus apportent des informations qualitatives et quantitatives sur les affinités lexicales que peuvent présenter certains mots entre eux. Ces affinités sont appelées co-occurrences. Le problème intrinsèque de ces modèles, quels qu'ils soient, se situe dans l'extrême diversité des associations extraites. Rappelons que parmi les co-occurrences extraites par le modèle statistique de [Lafon, 1984] se trouvent des associations sémantiques, des associations fonctionnelles parmi lesquelles on rencontre les noms composés, et des associations diverses. Le but que nous nous fixons est d'extraire les noms composés et d'ignorer les autres catégories de co-occurrences. Le problème à résoudre est double :

- d'abord, "guider" linguistiquement les modèles statistiques sur les co-occurrences que nous voulons extraire. Nous avons décidé d'extraire les noms composés binaires. En effet, les noms composés de longueur 2 sont de loin les plus nombreux. De plus, la majorité des noms composés de longueur supérieure ou égale à 3, par ailleurs peu représentés en comparaison des premiers, sont construits à l'aide de noms composés binaires. L'utilisation d'un modèle statistique demande une bonne représentation des échantillonnages relevés ; les noms composés de longueur 2 semblent suffisamment représentés dans nos textes pour s'accommoder d'un modèle statistique. Pour cela, nous utilisons les spécifications linguistiques (établies dans le chapitre II), en termes de structures morphosyntaxiques (patrons). Ces structures morphosyntaxiques s'expriment à l'aide d'expressions rationnelles et donc peuvent être extraites du corpus étiqueté à l'aide d'automates finis. Les co-occurrences extraites sont donc toutes susceptibles d'être des noms composés sur le plan morphosyntaxique.

- puis, utiliser des modèles statistiques pour distinguer parmi ces co-occurrences lesquelles sont effectivement des noms composés. Nous examinerons un certain nombre de modèles statistiques : les modèles déjà utilisés dans le traitement du langage naturel, mais aussi d'autres, utilisés dans des domaines scientifiques comme la biologie, Nous les évaluerons et nous déterminerons quel est le meilleur pour notre application : c'est-à-dire un modèle qui concentre ses valeurs fortes sur les noms composés de notre

domaine technique et ses valeurs faibles sur les co-occurrences qui n'en sont pas.

Nous allons maintenant préciser comment nous avons décidé de guider les modèles statistiques. La section suivante (section 4.2) présente le programme utilisé pour extraire et compter les co-occurrences. La suite de ce chapitre est consacrée aux modèles statistiques que nous avons retenus ou introduits (section 4.3.5), à leur évaluation (section 4.4) et aux résultats qu'ils produisent (section 4.5).

4.2 Extraction linguistique et relevé des fréquences

Les automates que nous utilisons ne reconnaissent que les noms composés de longueur 2, ce qui, rappelons-le, est motivé tant linguistiquement que statistiquement. Nous allons rappeler brièvement le cadre théorique dans lequel se situent les automates à l'aide de deux définitions, d'abord celle d'une expression rationnelle, puis celle d'un automate. Avant de décrire les automates des noms composés pour chaque type élémentaire, nous préciserons comment nous avons traité les problèmes de surcomposition et de modification. Le programme sera ensuite rapidement présenté et nous donnerons le format sous lequel apparaissent les couples extraits.

4.2.1 Cadre théorique

4.2.1.1 Expressions rationnelles

Une expression rationnelle est définie sur un ensemble fini appelé "alphabet" dont les éléments sont appelés des "symboles". L'alphabet que nous utilisons correspond à un sous-ensemble de l'ensemble des étiquettes grammaticales affectées aux items de notre corpus. Les expressions rationnelles sont définies par récurrence :

- le mot vide noté ϵ est une expression rationnelle. Le langage correspondant à l'expression rationnelle ϵ est exactement le singleton $\{\epsilon\}$;
- tout symbole S de l'alphabet est une expression rationnelle. Le langage correspondant est le singleton $\{S\}$;
- si X_1 et X_2 sont deux expressions rationnelles, alors X_1X_2 (concaténation) est une expression rationnelle. Par exemple, l'expression N ADJ représente le singleton {N ADJ}. La séquence N ADJ est obtenue par concaténation des deux symboles N et ADJ ;

- si X_1 et X_2 sont deux expressions rationnelles, alors $X_1 + X_2$ (union) est une expression rationnelle. L'expression N PREP N + N PREP DET N représente le langage constitué de deux séquences {N PREP N, N PREP DET N} ;
- si X est une expression rationnelle alors X^* (opération de Kleene) est une expression rationnelle. L'opération de Kleene peut être interprétée comme une série formelle de séquences, c'est-à-dire une union de concaténations répétée de façon non bornée: ADJ* = ϵ + ADJ + ADJ ADJ + ADJ ADJ ADJ + ...

Des parenthèses peuvent être utilisées dans les expressions. Par exemple, l'expression suivante: N (ADJ + PPAS) représente le langage {N ADJ, N PPAS}.

4.2.1.2 Automates finis

Les automates finis, comme les expressions rationnelles, permettent de représenter des langages de Kleene. Un automate est défini par un quintuplet: { A, E, I, F, T } avec A, l'alphabet, E un ensemble fini d'états, I une partie I non vide de l'ensemble E d'états initiaux, F une partie F non vide de l'ensemble E d'états finals, et T l'ensemble des transitions ; une transition étant définie par un triplet (e_i, e_j, l) où e_i et e_j sont des états et où l appartient à l'alphabet. Un patron syntaxique est reconnu par un automate s'il appartient au langage représenté par l'automate. Le principe général de la reconnaissance par automate est le suivant :

> On part du nœud initial, et on lit la séquence d'étiquettes grammaticales tout en se déplaçant dans l'automate en suivant les transitions étiquetées par l'étiquette courante. Si un état terminal de l'automate est atteint, la séquence reconnue par l'automate est enregistrée, sinon la séquence n'est pas enregistrée.

Le lien entre expressions rationnelles et automates est démontré par le théorème de Kleene :

> Un ensemble de mots est reconnu par un automate fini si et seulement s'il peut être décrit par une expression rationnelle.

Automates finis et expressions rationnelles sont donc équivalents, mais nous préférons décrire nos patrons syntaxiques à l'aide d'automates pour augmenter la lisibilité.

4.2.2 Définition des noms composés en termes de co-occurrences

Une co-occurrence est une association de deux unités lexicales constatée dans un corpus. Les noms composés sont des co-occurrences particulières qui possèdent

les propriétés présentées dans notre chapitre II : ils se définissent par rapport à leur structure morphosyntaxique ; ils acceptent des modifications qui peuvent être à l'origine de nouveaux noms composés ; ils admettent des variantes. Une co-occurrence, si elle caractérise un nom composé, répond aux conditions suivantes :

1. elle est orientée et suit l'ordre linéaire du texte,
2. elle met en jeu deux unités lexicales pleines,
3. elle doit apparaître dans l'une des structures morphosyntaxiques des noms composés binaires de type élémentaire.

Réunir ces conditions implique un traitement préliminaire du corpus où toutes les unités lexicales sont accompagnées de leurs étiquettes grammaticales. Chaque co-occurrence relevée caractérise un couple : ce couple est formé des deux lemmes des deux unités lexicales principales d'une structure morphosyntaxique. Il est préférable d'utiliser les lemmes plutôt que les formes fléchies de manière à ne pas obtenir de sous-représentation des couples. Ce choix implique d'associer chaque unité lexicale du corpus à son lemme. Chaque co-occurrence est quantifiée de la même façon : nous considérons qu'il y a équiprobabilité des apparitions.
Nous avons décidé d'extraire les noms composés binaires. Faut-il prendre en compte les opérations de modification, de surcomposition et de coordination qui affectent ceux-ci dans nos relevés de fréquences ? Ces relevés sont essentiels puisque c'est à partir de ceux-ci que vont s'appliquer les modèles statistiques. Un mauvais calcul de fréquence peut engendrer des scores statistiques faux ou non pertinents dans notre démarche. Ce choix est difficile et important. La décision dépend de la nature de la transformation qui affecte le nom composé binaire. Examinons successivement comment nous comptabilisons les co-occurrences suivant les différents cas de figures :

1. Noms composés de type élémentaire de longueur 2 :
   si la séquence *antenne de réception* est rencontrée, le nombre d'occurrences du couple (*antenne*, *réception*) est incrémenté de 1 ; si c'est la séquence *réception de l'antenne* qui est rencontrée, c'est le nombre d'occurrences du couple (*réception*, *antenne*) qui est incrémenté de 1.

2. Noms composés binaires surcomposés, modifiés ou coordonés :
   examinons dans quelle mesure la surcomposition, la modification et la coordination sont prises en compte dans le relevé des fréquences :

   (a) Surcomposition
   Nous ne prendrons pas en compte la surcomposition. Ainsi, si on rencontre la séquence : *antenne de réception du satellite*, correspondant à la séquence morphosyntaxique N_1 PREP N_2 PREP N_3, nous relevons les occurrences des couples (*antenne*, *réception*) et (*réception*, *satellite*), et

nous ne relevons pas l'occurrence du couple (*antenne*, *satellite*). Cette décision est motivée, d'une part, car nous voulons d'abord extraire les noms composés binaires, puisqu'il est impossible de statuer sur les noms composés ternaires avant de connaître les binaires, et d'autre part, si nous relevons les occurrences des couples (*antenne*, *réception*) et (*antenne*, *satellite*), nous faussons le relevé des fréquences ; en effet, dans ce cas, une seule occurrence d'*antenne* générerait deux occurrences : une pour le couple (*antenne*, *réception*) et l'autre pour (*antenne*, *satellite*). De manière à ne pas fausser le compte des fréquences et à éviter d'augmenter artificiellement le nombre d'occurrences de certains couples, nous ne relevons pas les fréquences des surcomposés.

(b) Post-modification
Nous ne prendrons pas en compte la post-modification. Considérons la séquence : *bande passante étroite* qui correspond soit à un surcomposé par substitution mettant en jeu les noms composés *bande passante* et *bande étroite*, soit au nom composé *bande passante* post-modifié par l'adjectif *étroite* ; *passante* n'est pas considéré comme un adjectif modifieur de *bande étroite* puisque l'insertion d'un adjectif dans un nom composé binaire de type N ADJ n'est pas permise. Le statut des séquences *bande passante* et *bande étroite* n'étant pas connu, nous considérons que l'adjectif *étroite* post-modifie la séquence *bande passante*. Dans ce cas, nous relevons uniquement l'occurrence du nom composé : le nombre d'occurrences du couple (*bande*, *passante*) est incrémenté. Pareillement, si l'on rencontre la séquence *antenne de réception parabolique*, seule l'occurrence du couple (antenne, réception) est relevée.

(c) Insertion de modifieur
L'insertion d'un adjectif dans une séquence N_1 PREP N_2 nous pose problème car elle est très fréquente. Nous ne pouvons pas nous permettre de ne pas la prendre en compte, mais si nous la prenons en compte, le calcul des fréquences sera erroné : avec la séquence *antenne parabolique de réception*, le nombre d'occurrences des couples (*antenne*, *parabolique*) et (*antenne*, *réception*) est incrémenté de 1. La solution que nous avons retenue consiste à procéder en plusieurs étapes pour l'extraction des noms composés binaires : nous extrayons les noms composés de structure N_1 (PREP (DET)) N_2 séparément des noms composés de structure N ADJ. Ainsi, pour la séquence ci-dessus, le nombre d'occurrence du couple (*antenne*, *parabolique*) est incrémenté une fois lors du relevé d'occurrences des noms composés de type N ADJ et le nombre d'occurrences du couple (*antenne*, *réception*) est incrémenté une fois lors du relevé d'occurrences des noms composés de type N_1 (PREP (DET)) N_2. Accepter qu'une structure N_1 PREP N_2 soit modifiée par

un adjectif inséré, implique l'extraction de véritables noms composés ternaires de structure N_1 ADJ PREP N_2 qui sont considérés comme deux noms composés binaires de structures N ADJ et N_1 PREP N_2. Cependant, nous verrons que les modèles statistiques utilisés permettrons de différencier les ternaires des binaires de structure N_1 PREP N_2 modifié par un adjectif. Par exemple, le couple (service, satellite) apparaît dans nos listes de noms composés binaires de structure N_1 (PREP (DET)) N_2, et le couple (service, fixe) apparaît dans nos listes de structure N ADJ, mais le programme indique que le couple (service, satellite) apparaît le plus fréquemment dans la séquence *service fixe par satellite.* Nous pourrons donc en déduire que *service fixe par satellite* est un ternaire et que *service fixe* n'est pas un binaire. Signalons néanmoins que la situation décrite pour (service, satellite) et (service, fixe) est assez rare. Le cas le plus fréquent est celui illustré par *antenne parabolique de réception* où le couple (*antenne*, *réception*) apparaît le plus souvent dans la séquence *antenne de réception* et non dans la séquence *antenne parabolique de réception.* Statuer sur la séquence *antenne parabolique de réception*, surcomposé avec (antenne, parabolique) et (antenne, réception) tous les deux binaires, ou modifié avec uniquement (antenne, réception) nom composé et *parabolique* modifieur, est difficile. Devant ce problème, nous comptons *antenne parabolique* et *antenne de réception* comme deux noms composés binaires possibles sans statuer ni sur la séquence *antenne parabolique*, ni sur la séquence *antenne parabolique de réception.*

Nous sommes consciente que les décisions que nous avons dû prendre ici ne sont pas les seules possibles. En particulier, un autre traitement possible aurait pu être le suivant : enregistrer la séquence morphosyntaxique N ADJ dans les couples correspondants si cette séquence n'est pas une sous-séquence du patron N_1 (PREP (DET)) N_2. Or, procéder ainsi entraînerait, par contre coup, la reconnaissance de beaucoup de surcomposés au détriment de la reconnaissance des noms composés binaires. Par exemple, pour la séquence *antenne parabolique du satellite*, l'occurrence du couple (antenne, satellite) serait relevée dans le schéma N_1 (PREP (DET)) N_2 et l'occurence du couple (antenne, parabolique) ne serait pas relevé dans le schéma N ADJ. C'est toujours parce que vous désirons avant tout reconnaître les noms composés binaires que nous avons distingué ces deux structures : N ADJ et N_1 (PREP (DET)) N_2, malgré les inconvénients de cette solution, la non-présence des N ADJ dans un classement conceptuel global.

D'une manière générale, les autres types de modifieurs insérés sont acceptés s'il ne faussent pas le relevé des fréquences, comme l'adverbe dans la structure N ADJ. Les modifieurs permis pour chaque structure sont précisés dans la section 4.2.3.

(d) Coordination
Considérons le fragment de texte: *équipements de modulation et de démodulation.* Les nombres d'occurrences des couples (*équipement, modulation*) et (*équipement, démodulation*) sont tout deux incrémentés de 1. Dans ce cas, même si *équipement* n'apparaît effectivement qu'une fois dans la séquence, il nous semble naturel de considérer qu'il apparaît deux fois puisque nous aurions pu rencontrer aussi bien la séquence *équipement de modulation et équipement de démodulation.* Ce relevé de fréquence n'est faux que si la structure N_1 PREP N_2 CONJ (PREP) N_3, où CONJ est une conjonction de coordination, est un nom composé élémentaire de longueur 3. Or, ces noms composés étant rarissimes, nous pouvons admettre cette marge d'erreur.

Il nous reste à exposer la technique qui permet de comptabiliser les co-occurrences que nous désirons. Nous avons choisi d'utiliser des automates finis. Les automates nous permettent d'extraire les séquences morphosyntaxiques caractéristiques des noms composés binaires de type élémentaire. Les occurrences de ces séquences sont classées sous l'entrée de couples: un couple est orienté, et composé de deux lemmes et regroupent toutes les séquences où les formes fléchies des deux lemmes apparaissent dans l'un de nos patrons syntaxiques: soit N_1 (PREP (DET)) N_2, soit N ADJ. Ce sont sur ces couples que s'appliquent les mesures statistiques présentées dans la section 4.3.5.

4.2.3 Automates des noms composés de type élémentaire

Le choix qui est fait ici d'utiliser des automates n'est pas innocent: en effet, la plupart des systèmes réalisés jusqu'ici et présentés dans notre chapitre I (section 1.3) utilisent un fenêtre déplacée sur le corpus. Nous voyons un inconvénient à cette technique et son importance nous semble largement justifier son abandon au profit des automates. Tout d'abord, la fenêtre utilisée a nécessairement une taille arbitraire. Aucun tri n'est donc réalisé dans les séquences relevées, ni du point de vue de l'organisation syntaxique, ni dans les positions relatives des paires de mots constituées. Nous pensons que la linguistique ne peut que rendre plus efficace le relevé des occurrences.
Les automates que nous avons écrits pour reconnaître les séquences morphosyntaxiques caractéristiques des noms composés de longueur 2 ne sont pas déterministes. Plusieurs états initiaux et plusieurs états finals sont possibles pour certains patrons. Ces automates ont été mis au point à partir de notre corpus étiqueté et lemmatisé et ne peuvent donc être appliqués que sur un corpus utilisant le même ensemble d'étiquettes grammaticales (la liste d'étiquettes pour le français est donnée en Annexe A). Ces automates tiennent compte de certains défauts du programme d'assignation d'étiquettes grammaticales d'IBM: si ces automates

sont utilisés avec la même liste d'étiquettes mais avec un programme différent, ils devront être modifiés sur les points que nous allons préciser. L'alphabet de base de notre automate est un sous-ensemble de l'ensemble des étiquettes grammaticales. À cet alphabet, nous avons ajouté quelques symboles particuliers comme certains lemmes, certaines formes fléchies, et certains codages de graphie. Nous avons intégré à ces automates une procédure chargée de vérifier les règles d'accord en genre et en nombre.
Les automates sont considérés comme des filtres linguistiques pour le relevé des fréquences des co-occurrences. Nous allons représenter les automates sous forme de graphe où :

- l'entrée de l'automate est indiquée par une flèche (états initiaux),
- la sortie de l'automate (états finals) est indiquée par un double carré,
- les états sont représentées par des rectangles, les transitions sont représentés par des traits partant du bord extérieur droit d'un état et arrivant au bord extérieur gauche d'un autre état ; mais, pour plus de lisibilité, l'étiquette grammaticale qu'il a fallu rencontrer pour atteindre un état est indiqué sur l'état atteint.
- les paires de lemmes enregistrées correspondent aux états encadrés par un trait épais ; les états encadrés par un trait épais en pointillé correspondent à des états finals intermédiaires : l'occurrence du couple dont l'un des lemmes est associé à un état final intermédiaire est enregistrée si un état final est atteint par la suite.

Ces automates ont été modifiés de nombreuses fois de manière à optimiser l'extraction des couples : nous avons, en effet, essayé de minimiser la prise en compte de mauvais candidats en interdisant certaines séquences morphosyntaxiques qui introduisaient plus de mauvais candidats que de bons. Des automates ont été aussi écrits pour les noms composés binaires de l'anglais (N_2 N_1, ADJ N, N_1 PREP N_2) ; ceux-ci ne sont pas détaillés ci-dessous. Les résultats de ces automates anglais sont néanmoins utilisés dans notre programme de statistique bilingue (section 4.6).

4.2.3.1 Automate des N_1 N_2

La figure 4.1 représente l'automate pour la reconnaissance des noms composés de structure N_1 N_2 ; il est précisé qu'aucun des noms ne doit être écrit en majuscule : cette condition est imposée pour ne pas prendre en compte les abréviations. L'utilisation d'une abréviation dans une construction de type élémentaire implique une surcomposition et nous voulons pour l'instant uniquement isoler les noms composés de type élémentaire. Dans les figures suivantes présentant d'autres types d'automates, cette condition n'est plus représentée sur le schéma, mais elle

a été effectivement imposée. Malheureusement, cette condition graphique imposée aux noms ne permet pas d'éliminer toutes les abréviations puisque certaines apparaissent en minuscules : certaines surcompositions sont donc prises en compte.

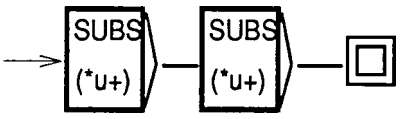


FIG. 4.1 - N_1 N_2

Cet automate est très simple : les noms composés de type N_1 N_2 sont les plus figés des noms composés. Nous n'avons recensé aucune altération de leur structure par aucune des opérations que nous avons décrites. Cet aspect figé sera éventuellement à remettre en question pour d'autres domaines techniques.

4.2.3.2 Automates des N_1 PREP N_2

Par souci de lisibilité, nous présentons séparément les automates pour les noms composés de type élémentaire N_1 *de* (DET) N_2, N_1 *à* (DET) N_2 et N_1 PREP N_2. Nous n'avons représenté dans les figures 4.2 et 4.3 certaines coordinations qui sont pourtant extraites. Ce sont :

- les coordinations à droite avec des virgules traitant une séquence comme : *circuits de commande, d'affichage et d'alarme* ; celles reconnues dans les véritables automates doivent se plier aux conditions suivantes :
  - chacun des noms coordonnés à l'aide de la virgule doit être accompagné de la préposition *de* (resp. *à*) et éventuellement d'un déterminant : une séquence comme *circuits de commande, affichage et alarme* n'est pas reconnue par l'automate,
  - le dernier élément coordonné doit l'être à l'aide d'une conjonction de coordination autre que la virgule ; une séquence comme *circuits de commande, d'affichage, d'alarme* n'est pas acceptée,
- les coordinations à gauche à l'aide d'une conjonction de coordination autre que la virgule comme dans la séquence suivante : *convertisseurs-élévateurs et abaisseurs de fréquence* : seules les séquences qui sont précédées par un article ou une préposition contractée sont acceptées.

Ces restrictions ont été introduites de manière à réduire la prise en compte de séquences ne correspondant pas à de réelles coordinations. Les coordinations impliquant des noms composés de type élémentaire de structures différentes ne sont pas traitées. Un déterminant optionnel est permis pour ces deux patrons entre la préposition *de* (resp. *à*) et le N_2.

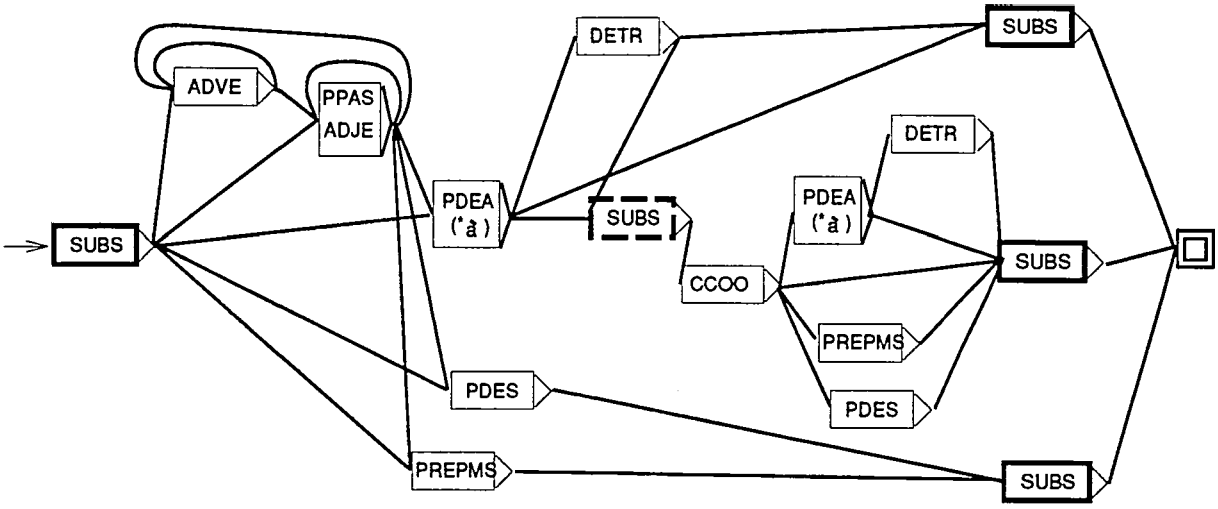


FIG. 4.2 - N_1 de (DET) N_2

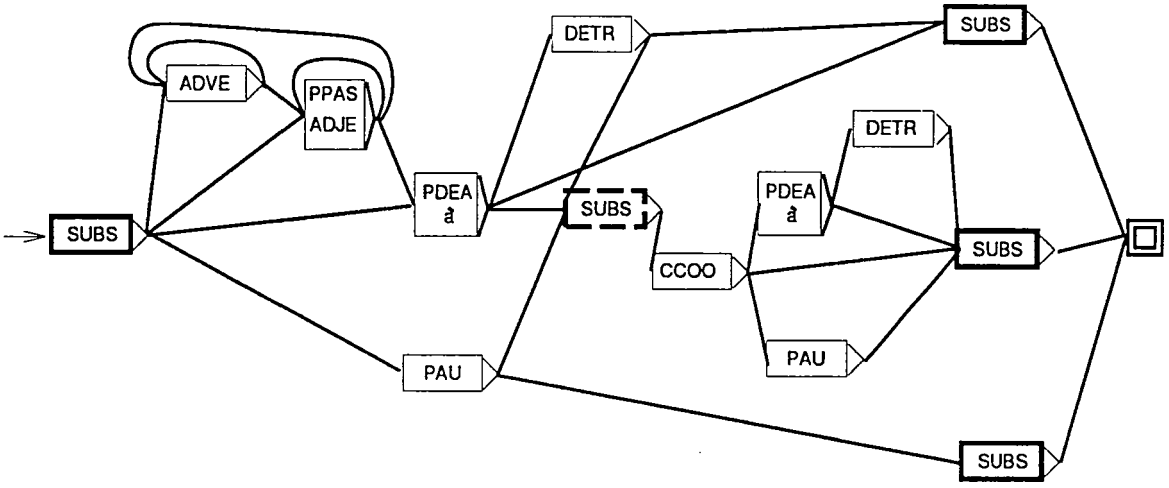


FIG. 4.3 - N_1 à (DET) N_2

Dans le patron morphosyntaxique des N_1 PREP N_2 où PREP n'est ni la préposition *à*, ni la préposition *de*, un déterminant optionnel n'est pas permis. Nous avons en effet remarqué que nous récupérions plus de mauvais candidats que de bons si nous permettions la présence d'un déterminant avant le N_2 dans ce patron.

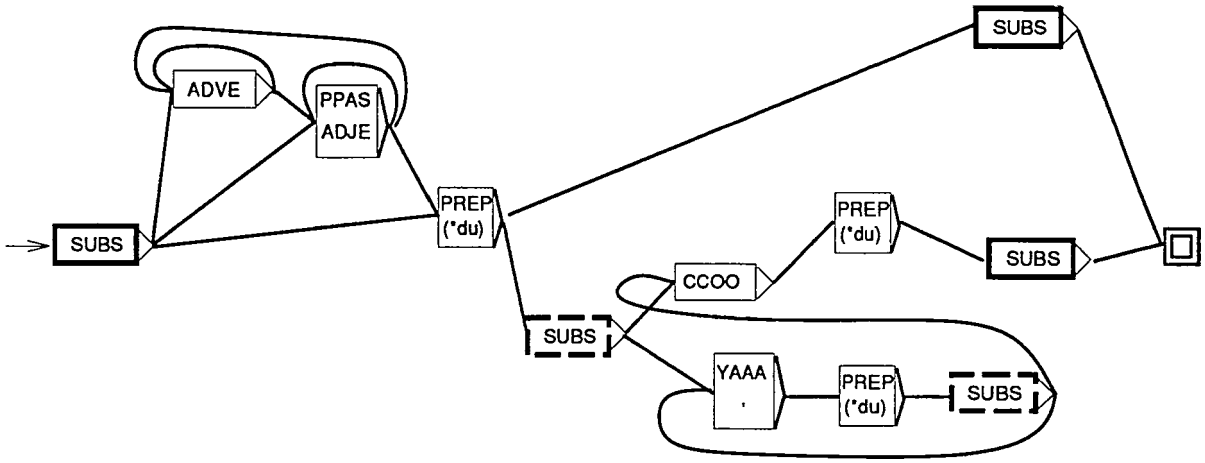


FIG. 4.4 - N_1 PREP N_2

4.2.3.3 Automate des N ADJ

Dans l'automate des noms composés de structure N ADJ, seul le lemme de l'adjectif (étiquette ADJE) est retenu comme deuxième lemme possible d'un couple. Le lemme d'un participe-passé (étiquette PPAS) n'est pas accepté car les couples de type (N, PPAS) seraient, pour la plupart, invalides. Cette décision découle du comportement du programme d'assignation d'étiquettes grammaticales vis-à-vis des participes passés : les participes passés employés comme adjectifs et non suivis par une préposition reçoivent l'étiquette ADJE à quelques exceptions près ; ceux suivis par une préposition reçoivent l'étiquette PPAS. Par exemple, *rayonné* reçoit l'étiquette ADJE dans la séquence *puissance rayonnée et ...* et reçoit l'étiquette PPAS dans la séquence *champ rayonné en espace libre.* D'autre part, nous n'avons pas pris en compte les coordinations à gauche comme *services et systèmes spatiaux*; celles-ci sont en effet ambiguës dans la plupart des cas. Nous avons accepté la possibilité de rencontrer l'adjectif en position attributive : cette transformation ne caractérise pas les noms composés les plus figés, ni les co-occurrences lexicales restreintes, mais elle est acceptée par les noms composés techniques : par exemple, le nom composé *antenne parabolique* accepte la restructuration : *cette antenne est parabolique.* Le relevé de ces co-occurrences n'est pris en compte que pour l'extraction de terminologie monolingue ; pour l'extraction de terminologie bilingue, ces co-occurrences sont ignorées.

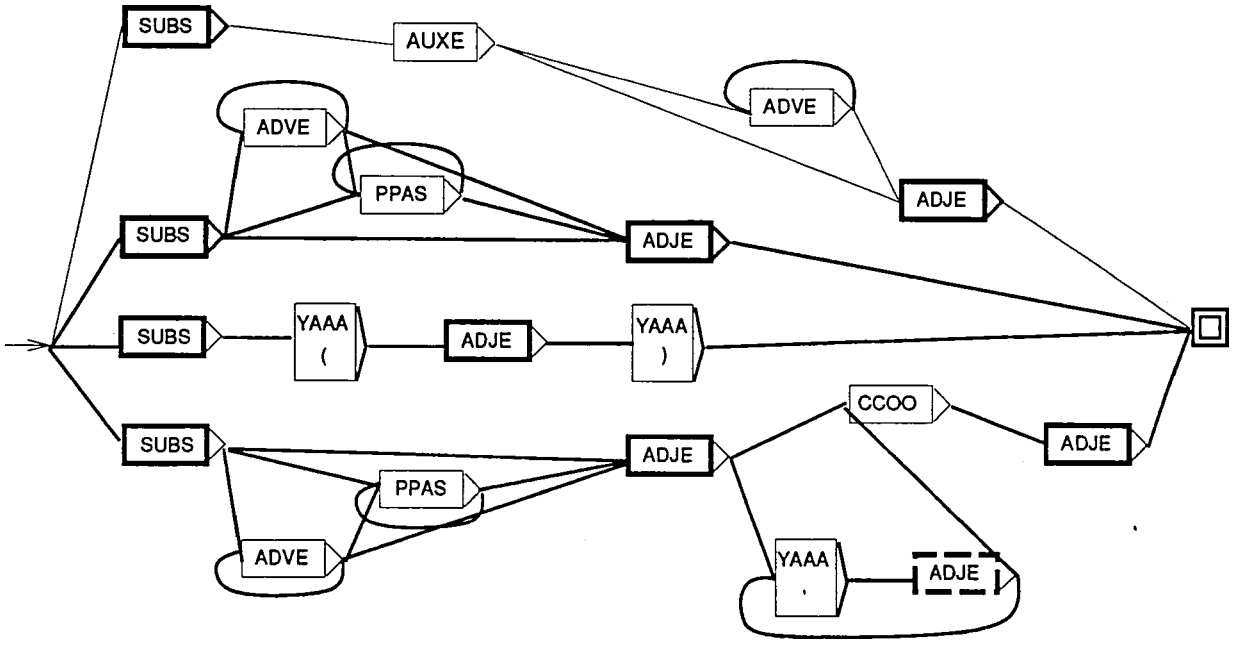


FIG. 4.5 - N ADJ

4.2.4 Présentation du programme

Le programme que nous avons écrit, utilise les automates à la fois pour relever les fréquences des co-occurrences et pour les extraire. Toutes les séquences morphosyntaxiques caractéristiques d'un couple et rencontrées dans le corpus pour un patron donné sont listées sous l'entrée de ce couple, accompagnées de leur propre nombre d'occurrences. De plus, chaque occurrence d'un couple est accompagnée de sa position dans le corpus : cette information, permet d'une part, un retour aisé au texte, et d'autre part, peut être utilisée dans le programme de statistique bilingue. Voici le format sous lequel apparaissent les couples extraits :

```
no 65 charge trafic nbc= 6  dist= 3.166667  mdist= 2.166667  var= 0.138889
occ1= charge de trafic patt1= SUBSFS PDEA SUBSMS nbo1= 5
Pos1=  5865/4/7/9 - 5865/29/16/18 - 5965/21/34/36 - 6243/14/19/21 - 6243/16/18/20
occ2= charge normale de trafic patt2= SUBSFS ADJEFS PREPMS SUBSMS nbo2= 1
Pos2=  5965/22/67/70

no 70 bande garde nbc= 6  dist= 3.000000  mdist= 2.000000  var=0.000000
occ1= bandes de garde patt1= SUBSFP PDEA SUBSFS nbo1= 3
Pos1=5872/10/7/9 - 5918/15/31/33 - 6033/9/52/54 -
occ2= bande de garde patt2= SUBSFS PDEA SUBSFS nbo2= 3
Pos2=  5889/6/10/12 - 5889/7/8/10 - 6239/6/42/44
```

Quelques commentaires sur ce format :

- `no 65` : identificateur numérique du couple,
- *charge* et *trafic* correspondent aux deux lemmes du couple,
- `nbc= 6` : nombre d'occurrences du couple,
- `dist= 3.166667 mdist= 2.166667 var= 0.138889` : ces nombres correspondent aux mesures de distance calculées pour chaque couple (voir 4.3.4 et 4.5.4),
- le couple est apparu dans deux séquences morphosyntaxiques différentes (`patt1`, `patt2`) correspondant aux deux portions de textes (`occ1`, `occ2`). Le nombre d'occurrences de chaque séquence est précisé (`nbo1`, `nbo2`). Les positions dans le corpus apparaissent sous le format : numéro de fichier, numéro de phrase, position du premier item et du dernier item de la séquence extraite.

Le programme utilise des arbres binaires associés à une fonction de hachage pour construire la liste des couples et les diverses informations qui les caractérisent : le temps d'exécution est donc très rapide : par exemple, l'extraction des 8 000 couples du corpus MTS, pour la structure N_1 *de* (DET) N_2, prend deux minutes sur une station Sparc ELC (SS1) sous la version Sun-Os 4.1.3.. Le programme est écrit en C norme ANSI POSIX.

4.2.5 Résultats du programme

Nous avons relevé les co-occurrences pour nos deux corpus MTS (200 000 mots) et LBC (800 000 mots) et pour les deux patrons, N_1 (PREP (DET)) N_2 (les parenthèses indiquent l'optionalité d'une ou de plusieurs étiquettes syntaxiques) et N ADJ. Une occurrence d'un couple correspond à une co-occurrence où les deux éléments du couple entrent dans un de ces deux patrons syntaxiques. Les tableaux ci-dessous résument les fréquences de co-occurrences exprimées en nombre de couples ; ainsi pour le corpus MTS et le patron N ADJ, nous avons relevé 4 483 couples dont 3 144 n'ont qu'une occurrence, 655 deux occurrences et 684 plus de deux occurrences.

MTS	1 occurrence	2 occurrences	Plus de 2 occurrences	Total
N ADJ	3 144	655	684	4 483
N_1 (PREP (Det)) N_2	6 834	1 503	1 616	9 953

LBC	1 occurrence	2 occurrences	Plus de 2 occurrences	Total
N ADJ	5 201	1 507	2 113	8 821
N_1 (PREP (Det)) N_2	12 167	3 481	6 288	21 936

Les co-occurrences relevées pour le patron syntaxique N_1 (PREP (Det)) N_2 ne nous donnent aucune information sur la représentativité des différents types élémentaires: N_1 *de* (DET) N_2, N_1 *à* (DET) N_2, N_1 PREP N_2 (avec PREP $\neq$ *de* et *à*) et N_1 N_2. Nous avons donc relevé les co-occurrences de chacun de ces types élémentaires. Les résultats en terme de nombre de couples sont résumés dans les tableaux suivants:

MTS	1 occurrence	2 occurrences	Plus de 2 occurrences	Total
N_1 DE[a] N_2	5 393	1 195	1 374	7 962
N_1 À[b] N_2	1 156	161	115	1 432
N_1 Prep[c] N_2	891	132	116	1 139
N_1 N_2	309	66	60	435
Total	7 749	1 554	1 665	10 968

LBC	1 occurrence	2 occurrences	Plus de 2 occurrences	Total
N_1 DE[a] N_2	9 558	2 804	5 308	17 670
N_1 À[b] N_2	2 358	492	471	3 321
N_1 Prep[c] N_2	1 474	341	439	2 254
N_1 N_2	682	180	352	1 214
Total	14 072	3 817	6 570	24 459

[a] DE = {*de, d', du, des, de la*}
[b] À = {*à, au, aux, à la*}
[c] Prep $\neq$ {DE, À}

Quelques remarques immédiates concernant le détail des sous-patrons des N_1 (PREP (DET)) N_2:

- ce sont de loin les couples relevés à l'aide du patron N_1 *de* (DET) N_2 qui sont les plus nombreux,
- les couples relevés à l'aide du patron N_1 *à* (DET) N_2 sont à peine plus nombreux que ceux relevés à l'aide du patron N_1 PREP N_2 (avec PREP $\neq$ *de* et *à*),
- les couples sous patron N_1 N_2 sont les moins représentés.

D'autre part, si nous additionnons les nombres des couples correspondant aux différents types élémentaires, nous obtenons un nombre supérieur à celui des couples relevés à l'aide du patron général N_1 (PREP (DET)) N_2. Ce résultat s'explique par le fait que de nombreux couples partagent les mêmes lemmes et varient uniquement avec la présence ou non d'une préposition ou avec la forme de la préposition. Par exemple, sous le couple (**circuit, hyperfréquence**) extrait avec le patron général N_1 (PREP (DET)) N_2, nous trouvons des co-occurrences

relevant de types élémentaires différents: *circuit hyperfréquence* appartient au type élémentaire N_1 N_2, *circuit à hyperfréquences* au type élémentaire N_1 *à* (DET) N_2. Ce patron général a donc l'avantage de recenser sous l'entrée d'un couple certaines variantes morphosyntaxiques; dans d'autres cas, il groupe parfois sous l'entrée d'un même couple:

- des co-occurrences qui ne sous-tendent pas le même concept: sous le couple (**centre, origine**), nous trouvons les co-occurrences *centre d'origine* et *centre à l'origine*. Dans cette dernière, *à l'origine* est une sous-séquence de la préposition composée *à l'origine de* et la co-occurrence extraite ne représente aucun concept spécifique et donc, ne correspond pas à l'entité évoquée par *centre d'origine*.

- des co-occurrences relevant de divers types élémentaires et pour lesquelles il est difficile de décider si elles réfèrent ou non au même concept. En voici deux exemples:

  - sous le couple (**liaison, satellite**), les co-occurrences *liaison par satellite* appartenant au type élémentaire N_1 *par* N_2, et *liaison à satellites* au type N_1 *à* (DET) N_2 réfèrent-elles à la même entité?
  - sous le couple (**terminal, interface**), nous nous posons la même question pour les co-occurrences *terminal d'interface* et *terminal à interface*.

  Décider du statut conceptuel de différentes co-occurrences peut s'effectuer par un retour au texte.

Le problème est général et se rencontre même pour des co-occurrences appartenant à un même type élémentaire: le couple (**couleur, fond**) pour le patron N_1 *de* (DET) N_2, regroupe *couleur de fond* et *couleur du fond* qui ne nous semblent pas référer à la même entité. Nous nous trouvons devant le dilemme suivant: faut-il relever les co-occurrences régies par un patron général comme N_1 (PREP (DET)) N_2 quitte à réunir sous un même couple plusieurs entités ou faut-il relever les co-occurrences suivant un type élémentaire précis, perdre ainsi les liaisons entre ces types, surtout limiter la hiérachisation des concepts, et de toute façon rencontrer le même problème mais à une échelle un peu moindre? Ce problème de signification du dénombrement est inhérent à notre approche quantitative et nous préférons avoir plus de représentativité, quitte à enregistrer sous un couple un nombre, qui espérons-le est limité, d'occurrences référant à des entités différentes, que de limiter trop les relevés et perdre nombre d'informations essentielles.
S'il est important d'être conscient du problème lié au dénombrement exposé ci-dessus, le problème des co-occurrences pertinentes oubliées n'est pas moins crucial. Nous avons vu précédemment que, pour ne pas fausser les relevés d'occurrences, les noms composés binaires surcomposés et pré- ou post- modifiés ne sont pas pris en compte. Cette position implique la non-reconnaissance de certaines

co-occurrences : pour la séquence *ondes à polarisation rectiligne perpendiculaires*, les occurrences des couples (**onde**, **polarisation**) pour le patron N_1 (PREP (DET)) N_2 et (**polarisation**, **rectiligne**) pour le patron N ADJ sont relevées, mais l'occurrence du couple (**onde**, **perpendiculaire**) est ignorée. Cependant, l'adjectif *perpendiculaires* modifie *ondes à polarisation rectiligne* et pas véritablement le nom *onde* et le non-relevé de cette co-occurrence est plutôt une bonne chose. D'autres co-occurrences relevant pourtant des patrons syntaxiques de noms composés binaires nous échappent ; en voici quelques exemples :

- certaines co-occurrences sont séparées par des éléments perturbateurs comme par exemple la séquence *une composante (x) brouilleuse* ; la présence entre *composante* et *brouilleuse* d'un élément entre parenthèses ne permet pas la reconnaissance de cette occurrence du couple (**composante**, **brouilleuse**),

- le programme d'étiquetage grammatical commet certaines erreurs qui entraînent la non-reconnaissance de certaines séquences morphosyntaxiques ; par exemple, dans la séquence *les polarisations quasi circulaires*, l'occurrence du couple (**polarisation**, **circulaire**) n'est pas reconnue. L'automate demande l'accord en genre et en nombre entre le nom et l'adjectif qui le modifie : dans cette séquence *polarisations* reçoit une étiquette correcte (nom féminin pluriel), mais *circulaires* reçoit une étiquette incorrecte, à savoir : adjectif masculin pluriel. Comme l'accord en genre demandé par l'automate n'est pas respecté entre le nom et l'adjectif, l'occurrence n'est pas enregistrée.

Ces cas où les co-occurrences n'ont pas été comptabilisées montrent que :

1. il est impossible de prendre en compte toutes les séquences morphosyntaxiques où apparaissent des co-occurrences correctes sans comptabiliser beaucoup de co-occurrences incorrectes,

2. la sélection des co-occurrences repose sur l'étiquetage morphologique ; les erreurs d'étiquetage peuvent entraîner la non-sélection de bonnes co-occurrences comme la prise en compte de mauvaises co-occurrences.

Nous n'évaluerons les modèles statistiques présentés dans la section suivante que sur les couples présentant au moins deux occurrences. Ce seuil de fréquence est le même que celui de [Lafon, 1984] mais est bas comparé au seuil de cinq occurrences préconisé par [Smadja et McKeown, 1990] ou [Church et Hanks, 1990]. Nous avons pris en compte le filtrage linguistique préliminaire : le "bruit" est considérablement diminué. Ce seuil est fixé arbitrairement ; cependant, les valeurs des modèles statistiques sont effectivement faussées pour les couples n'ayant qu'une occurrence. En ne tenant pas en compte des co-occurrences uniques, nous retenons en moyenne un couple sur deux. Il reste que ces co-occurrences peuvent très bien contenir des termes et nous sommes consciente que cette élimination engendre du "silence".

Nous abordons maintenant la partie statistique de notre travail. Nous avons présenté comment nous "guidions" les modèles statistiques: nous avons filtré les co-occurrences et celles-ci sont toutes, morphosyntaxiquement, susceptibles d'être des noms composés. Nous allons tout d'abord présenter un certain nombre de modèles statistiques (section 4.3.5); dans la section 4.4, ces modèles seront ensuite évalués par une méthode graphique sur un sous-ensemble de nos couples: ceux de structure N_1 *de* (DET) N_2 extraits du corpus MTS. Puis, nous prendrons en compte les résultats de cette évaluation et nous examinerons sur nos deux patrons plus généraux, N_1 (PREP (DET)) N_2 et N ADJ, les classements conceptuels proposés par ces modèles (section 4.5). Nous analyserons aussi, dans la section 4.5, les résultats de quelques modèles périphériques qui ne mesurent pas la force du lien entre les éléments du couple, mais qui apportent d'autres types d'informations.

4.3 Modèles statistiques

L'intégration de modèles statistiques passe par un recours à des mesures bien connues des statisticiens et dont nous rappelons ici les définitions.
Quatre types de caractéristiques numériques sont calculées:

- les fréquences,
- les critères d'association,
- la diversité,
- les mesures de distance.

À ces caractéristiques numériques, on peut ajouter une mesure un peu particulière qui s'appuie sur des données bilingues: l'affinité, sur laquelle nous ne nous étendrons pas. Elle est étudiée extensivement dans [Gaussier, 1994]. Ces caractéristiques ne sont pas susceptibles de jouer le même rôle: les fréquences correspondent aux paramètres des critères d'association, et normalement ne devraient pas jouer de rôles discriminatoires; les critères d'association, l'affinité y compris, mesurent la force du lien entre les deux lemmes d'un couple et propose un classement conceptuel des couples; la diversité et les mesures de distances apportent d'autres types d'information.

4.3.1 Fréquences

Les occurrences relevées pour une structure morphosyntaxique donnée sont:

- le nombre d'occurrences d'un couple; celui-ci a été calculé par le programme de relevé et d'extraction de co-occurrences. Les fréquences suivantes sont calculées sur l'ensemble des couples extraits:

- le nombre d'occurrences des couples où un lemme donné apparaît comme premier élément du couple,
- le nombre d'occurrences des couples où un lemme donné apparaît comme deuxième élément,
- le nombre total d'occurrences des couples (pour chaque patron syntaxique).

Ces quatre comptes sont utilisés par tous les modèles statistiques que nous avons retenu, ce qui justifie le soin apporté au relevé du nombre d'occurrences d'un couple.

4.3.2 Critères d'association

D'un point de vue statistique, les deux lemmes qui forment un couple sont considérés comme deux variables qualitatives dont il s'agit de tester la liaison. Les données se représentent sous la forme d'un tableau croisé, appelé tableau de contingence et défini à partir des comptes précédents. Un tableau de contingence est associé à chaque couple de lemmes (L_i, L_j) :

	L_j	$L_{j'}$ avec $j' \neq j$
L_i	a	b
$L_{i'}$ avec $i' \neq i$	c	d

Les valeurs a, b, c et d résument les occurrences d'un couple :

a = le nombre d'occurrences du couple (L_i, L_j),

b = le nombre d'occurrences des couples où L_i est le premier élément d'un couple et L_j n'est pas le second,

c = le nombre d'occurrences des couples où L_j est le second élément du couple et L_i n'est pas le premier,

d = le nombre d'occurrences de couples où ni L_i ni L_j n'apparaissent.

La somme $a + b + c + d$, notée N, est le nombre total d'occurrences de tous les couples trouvés pour un patron morphosyntaxique.

La littérature statistique regorge de mesures destinées à "tester l'indépendance" ou à "mesurer la liaison" ou encore à "mesurer le degré de similitude ou d'affinité" entre deux variables régies par un tableau de contingence. Ces mesures sont soit des mesures de liaison (SMC), soit des tests statistiques (Φ^2) que nous utilisons comme mesures de liaison. Nous ne faisons pas la différence entre les mesures de liaison et les tests statistiques. D'ailleurs, nous utilisons ces derniers, non pour

statuer sur la dépendance ou l'indépendance de deux variables aléatoires, mais uniquement comme mesure de liaison : les deux lemmes qui composent le couple appartiennent à une structure morphosyntaxique particulière et ne sont donc pas indépendants. Dans nos hypothèses, ces tests statistiques s'expriment aussi en fonction des valeurs d'un tableau de contingence. Nous n'examinons pas si les valeurs prises par les mesures sont statistiquement significatives : aucun couple n'est rejeté. Nous allons maintenant énumérer les mesures statistiques que nous avons testées sur nos couples :

- certaines ont déjà été utilisées dans le domaine de la statistique lexicale : coefficient de proximité simple (équation 4.1), coefficient du Φ^2 (équation 4.7), coefficient de vraisemblance (équation 4.11), score d'association (équation de 4.8),
- certaines ont été utilisées dans d'autres domaines comme par exemple la biologie : coefficient de Kulczinsky (équation 4.2), coefficient d'Ochiai (équation 4.3), coefficient de Fager et McGowan (équation 4.4), coefficient de Yule (équation 4.5), coefficient de McConnoughy (équation 4.6),
- nous en avons introduite une nouvelle pour des raisons que nous préciserons : score d'association au cube (équation 4.9).

Coefficient de Proximité Simple (SMC)

Ce score est symétrique entre L_1 et L_2, et varie de 0 à 1.

$$SMC = \frac{a+d}{a+b+c+d} \tag{4.1}$$

Coefficient de Kulczinsky (KUC)

Ce score varie de 0 à 1. Quand L_1 (resp. L_2) est observé uniquement avec L_2 (resp. L_1), la valeur du coefficient de Kulczinsky est supérieure à 0,5.

$$KUC = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right) \tag{4.2}$$

Coefficient d'Ochiai (OCH)

Ce score varie de 0 à 1.

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{4.3}$$

Coefficient de Fager et McGowan (FAG)

Ce score varie d'une valeur non définie *a priori* qui peut être négative à 1 (cette borne supérieure n'etant pas atteinte).

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+b}} \tag{4.4}$$

Coefficient de Yule (YUL)

Ce score varie de -1 à 1 et est égal à 1 lorsque l'un des lemmes apparaît toujours avec l'autre.

$$YUL = \frac{ad-bc}{ad+bc} \tag{4.5}$$

Coefficient de McConnoughy (MCC)

Ce score varie de -1 à +1.

$$MCC = \frac{a^2-bc}{(a+b)(a+c)} \tag{4.6}$$

On vérifie que:

$$MCC = 2KUL - 1$$

et donc, que le coefficient de McConnoughy et celui de Kulczinsky induisent la même distribution. Pour cette raison, nous ne présentons pas l'histogramme de McConnoughy puisqu'il est identique à celui de Kulczinsky.

Coefficient du Φ^2 (PHI)

Ce score a été utilisé par [Gale et Church, 1991b] pour l'alignement de mots à l'intérieur de phrases appariées.

$$PHI = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)} \tag{4.7}$$

Score d'association

Cette méthode pour mesurer le poids d'association de deux mots a été décrite par [Brown *et al.*, 1988] dans le cadre de l'extraction de ressources lexicales bilingues et par [Church et Hanks, 1990] pour l'extraction monolingue. Le score d'association d'un couple de lemmes (L_1, L_2) est fondé sur la notion théorique d'information mutuelle. Ce score, noté IM, compare la probabilité d'observer deux lemmes ensembles avec la probabilité d'observer ces deux lemmes séparément. Sa définition est la suivante :

$$IM(L_1, L_2) = \log_2 \frac{P(L_1, L_2)}{P(L_1)P(L_2)}$$

où P est la probabilité.
Soit encore en termes de nos indices de référence :

$$IM = \log_2 \frac{a}{(a+b)(a+c)} \tag{4.8}$$

Le score d'association tel qu'il est défini, comme nous le verrons dans la section 4.5.2.1, donne trop de poids aux événements rares. Nous avons donc introduit la mesure suivante :

Score d'Association au cube (IM^3)

Cette formule résulte d'une étude expérimentale : nous avons voulu donner plus de poids aux événements fréquents et nous avons essayé toutes les puissances de a de 2 à 10. Le cube a été retenu car il apparaît comme un bon compromis entre ne retenir que les événements rares et trop les négliger.

$$IM^3 = \log_2 \frac{a^3}{(a+b)(a+c)} \tag{4.9}$$

Coefficient de vraisemblance (Loglike)

Ce coefficient, introduit par [Dunning, 1993], est le test du rapport de vraisemblance appliqué à une loi binomiale. Il correspond dans nos hypothèses à une information mutuelle généralisée et peut être exprimé avec nos indices de tableau de contingence.

$$\begin{aligned} Loglike = {} & a\log a + b\log b + c\log c + d\log d - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) \\ & -(c+d)\log(c+d) + N\log N \end{aligned} \tag{4.10}$$

La valeur prise par ces mesures croie avec le degré de liaison des lemmes.

4.3.3 Diversité de Shannon

Cette mesure a été introduite par Shannon (1948) et est utilisée en biologie pour classer des individus parmi des espèces. Elle nous a paru intéressante car elle permet de mesurer le rôle d'un lemme à une place fixe à l'intérieur du couple par rapport à l'ensemble des couples. L'idée étant qu'un lemme qui apparaît dans un nombre important de couples à proportion égale, en première position du couple, correspond soit à un élément systématiquement constitutif d'un terme, comme par exemple le nom *système*, soit le contraire, c'est-à-dire un nom systématiquement non-constitutif d'un terme comme par exemple le nom *caractéristique*. La diversité est une mesure qui caractérise la distribution marginale d'un des lemmes d'un couple parmi l'ensemble des couples. Son calcul s'appuie sur un tableau de contingence de dimension $n \times m$ dont la représentation théorique est la suivante :

$x_i\ y_j$	y_1	y_2	..	y_j	..	y_m	Total
x_1	n_{11}	n_{12}	..	n_{1j}	..	n_{1m}	$n_{1.}$
x_2	n_{21}	n_{22}	..	n_{2j}	..	n_{2m}	$n_{2.}$
.	.	.	.	.	.	.	.
x_i	n_{i1}	n_{i2}	..	n_{ij}	..	n_{im}	$n_{i.}$
.	.	.	.	.	.	.	.
x_k	n_{k1}	n_{k2}	..	n_{kj}	..	n_{km}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	..	$n_{.j}$	..	$n_{.m}$	$n_{..}$

avec X et Y les deux caractères considérés, x_1, x_2, ..., x_i, ..., x_k les k modalités de X ; y_1, y_2, ..., y_i, ..., y_m les m modalités de Y. Les n observations sont réparties suivant les modalités de X et Y. Dans notre application, X représente l'étiquette grammaticale du premier lemme du couple (N), Y l'étiquette grammaticale du deuxième lemme du couple (N ou ADJ). Le nombre n_{ij} figurant à l'intersection de la ligne i et de la colonne j du tableau est le nombre d'occurrences des couples contenant le lemme x_i avec l'étiquette X et le lemme y_j avec l'étiquette Y. Voici un extrait du tableau de contingence associé au schéma syntaxique N ADJ :

$N_i\ Adj_j$	*progressif*	*circulaire*	*porteur*	...	Total
onde	19	4	6	...	$nb_{(onde,.)}$
limiteur	9	0	0	...	$nb_{(limiteur,.)}$
cornet	0	2	0	...	$nb_{(cornet,.)}$
...	...	...	...	...	...
Total	$nb_{(.,progressif)}$	$nb_{(.,circulaire)}$	$nb_{(.,porteur)}$	...	$nb_{(.,.)}$

Les totaux $n_{i.}$ des lignes, placés dans la marge de droite, représentent la distribution des adjectifs par rapport à un substantif donné. Les totaux $n_{.j}$ des colonnes, placés en dernière ligne, représentent la distribution des noms par rapport à un adjectif donné. Ce sont ces distributions qui sont appelées "distributions marginales" des noms et des adjectifs pour le patron syntaxique N ADJ.
La diversité est calculée pour chaque modalité d'un caractère considéré, i.e. chaque lemme apparaissant dans un couple, grâce à la formule suivante :

$$\begin{aligned} H_i &= n_{i.} \log n_{i.} - \sum_{j=1}^{s} n_{ij} \log n_{ij} \\ H_j &= n_{.j} \log n_{.j} - \sum_{i=1}^{s} n_{ij} \log n_{ij} \end{aligned} \tag{4.11}$$

Par exemple, sur le tableau de contingence du schéma syntaxique N ADJ donné ci-dessus, la diversité du nom *onde* est donné par :

$$\begin{aligned} H_{(onde,.)} = nb_{(onde,.)} \log nb_{(onde,.)} - (&nb_{(onde,progressif)} \log nb_{(onde,progressif)} + \\ &nb_{(onde,circulaire)} \log nb_{(onde,circulaire)} + \ldots) \end{aligned}$$

Les diversités calculées pour les premiers lemmes des couples seront notées H_1 ; celles qui concernent les deuxièmes lemmes des couples H_2. La valeur de la diversité associée à l'un des lemmes du couple devient encore plus parlante lorsqu'elle est normalisée par le nombre d'occurrences du couple :

$$\begin{aligned} h_i &= \frac{Hi}{n_{ij}} \\ h_j &= \frac{Hj}{n_{ij}} \end{aligned} \tag{4.12}$$

Les diversités normalisées h_1 et h_2 sont donc définies à partir de H_1 et H_2. Nous verrons par la suite tout l'intérêt de la diversité normalisée qui permet de corriger certaines erreurs commises par le programme d'étiquetage grammatical.

4.3.4 Mesures de distance

Les noms composés terminologiques acceptent fréquemment des modifications : la distance séparant les deux lemmes du couple est variable. Pour chaque occurrence du couple, deux distances sont calculées :

- le nombre total d'éléments apparaissant entre les deux lemmes,
- le nombre d' "éléments pleins" ; les éléments pleins sont présentés dans le chapitre II (section 2.2.1) comme les noms, adjectifs, adverbes, etc et d'une manière générale, n'incluent pas les mots grammaticaux.

Pour chaque couple, la moyenne arithmétique de ces deux distances est calculée. Pour mesurer la dispersion de ces distributions statistiques, il paraît naturel de calculer les écarts entre les valeurs prises par la variable et sa moyenne arithmétique. Plus ces écarts sont importants en moyenne, plus la dispersion de la distribution est forte. Rappelons que le calcul de l'écart-type se fait en deux étapes : on calcule les écarts à la moyenne, puis les carrés de ces écarts et on en fait la somme. La moyenne arithmétique des carrés des écarts à la moyenne est la variance, notée $V(X)$ pour la variable x ; l'écart-type, noté σ, est la racine carrée de la variance, avec $\bar{x}$ la moyenne arithmétique :

$$V(X) = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\sigma(X) = \sqrt{V(X)}$$

L'écart-type de la distance séparant les deux lemmes du couple est calculé pour chaque couple et permettra ainsi de qualifier les modifications dont sont l'objet les termes candidats.

4.3.5 Affinité

L'appariement des mots à l'intérieur de phrases bilingues alignées donne pour chaque mot d'une langue une liste associée de traductions probables. Cette méthode pour apparier les mots a été présentée dans notre chapitre I section 1.5 ([Gaussier *et al.*, 1992]). Ces listes de traductions probables permettent d'introduire une nouvelle mesure d'association : l'affinité présentée dans [Gaussier et Langé, 1993]. L'idée est la suivante : si deux mots apparaissent fréquemment ensembles, alors leurs listes de traductions probables possèdent inévitablement des éléments en commun. Par exemple, dans les listes associées à *pomme* et à *terre* se trouverait le nom anglais *potato*. L'affinité est en fait une mesure de liaison entre mots définie à partir d'une mesure de liaison entre listes.

4.4 Évaluation graphique des modèles statistiques

Chaque modèle statistique propose un classement conceptuel des couples. Ce classement, néanmoins, peut très bien mettre plus en avant des noms composés figés ou des contraintes de sélection qui appartiennent au langage courant que véritablement des termes du domaine. Notre but étant d'obtenir une liste des termes du domaine des télécommunications, il est essentiel d'évaluer la corrélation entre les valeurs des modèles et les couples, et ainsi de déterminer quel modèle est

le plus adapté à l'extraction de terminologie. Nous avons donc décidé de comparer les valeurs obtenues pour chaque modèle à une liste de référence des termes du domaine. Cette liste de référence est présentée ci-dessous. Cette évaluation s'effectue sur les couples de structure syntaxique N_1 *de* (DET) N_2 extraits du corpus MTS. Si le couple apparaît dans la liste de référence, il est considéré comme un bon candidat ; sinon comme un mauvais candidat. Nous pourrons ainsi examiner les valeurs associées des différentes mesures calculées et décider laquelle ou lesquelles sont les plus adéquates à la détection des concepts terminologiques.

4.4.1 Liste de référence

Une liste de référence des termes du domaine des télécommunications pourrait être construite manuellement à partir de notre corpus MTS comme [Van der Eijk, 1993] l'a fait pour son programme d'extraction de terminologie bilingue. Une alternative à ce travail consiste à acquérir une banque terminologique déjà existante de notre domaine. Nous avons utilisé la banque de donnée terminologique de la Commission Européenne, section des télécommunications d'Eurodicautom, élaborée par des experts des télécommunications. Cette banque terminologique est disponible pour les neuf langues de la Communauté européenne. Nous avons obtenu les données françaises qui contiennent environ 6 000 termes. Voici un extrait de cette banque :

```
caractéristiques fondamentales d'une attribution de fréquence
température apparente du ciel
température cinétique
température de bruit d'un récepteur
température de bruit d'un système de réception
température de bruit de fonctionnement
température de bruit de fond
température de bruit équivalente
température de bruit quantique
température de bruit thermique
température de régime
température équivalente d'antenne
température équivalente pour le trajet descendant
température moyenne de rayonnement
```

Cet extrait montre que tous les termes de cette banque terminologique sont "à plat" : la banque terminologique n'est aucunement structurée et chaque terme occupe simplement une ligne d'un fichier texte. Nous pouvons remarquer que six termes sont construits avec le nom composé *température de bruit* sans que celui-ci soit listé ; la même remarque est valable pour *température équivalente*, *bruit de fonctionnement*, *trajet descendant*, *température moyenne*. De plus, certaines séquences comme *température apparente du ciel* ou *caractéristiques fondamentales*

d'une attribution de fréquence ressemblent plus à des groupes nominaux libres qu'à des termes du domaine. Nous avons donc comparé nos couples à cette liste de référence en acceptant que nos candidat N_1 *de* (DET) N_2 puissent ne former qu'une sous-séquence d'un terme de la liste. Cette première évaluation a donné de mauvais résultats : seuls 300 couples sur les 2 200 étaient considérés comme des bons candidats. Nous avons rejeté en partie la faute sur la liste de référence et nous avons décidé de lui adjoindre une liste de termes extraits directement de notre corpus MTS. Nous avons donné à trois experts du domaine la liste des couples qui n'apparaissaient pas dans Eurodicautom, soit 1 900 candidats, et nous leur avons demandé de porter un jugement sur leur valeur compositionnelle : ces experts devaient indiquer pour chaque couple associé au patron N_1 *de* (DET) N_2 s'ils le considéraient comme un nom composé, s'ils ne le considéraient pas comme un nom composé ou s'ils ne pouvaient trancher sur sa nature compositionnelle. Nous n'avons retenu que les couples sur lesquels deux experts au moins avaient voté favorablement, soit environ 900 couples (seuls 300 couples ont fait l'unanimité). Ainsi, sur nos 2 200 couples de patron N_1 *de* N_2 possédant au moins deux occurrences dans le texte, nous admettrons que 1 200 environ sont des noms composés dont la plupart des termes du domaine des télécommunications.

4.4.2 Comparaison graphique

Chaque mesure fournit comme résultat une liste de candidats qui sont rangés par ordre décroissant de la valeur de la mesure. Dans cette liste, nous avons défini des classes d'équivalence qui regroupent en général 50 éléments successifs de la liste (les classes qui comportent plus de 50 éléments seront précisées ci-dessous). Les résultats d'une mesure sont représentés sous forme d'histogramme : pour chaque classe d'équivalence, l'histogramme indique le rapport du nombre de candidats appartenant à la liste de référence sur le nombre de candidats de la classe, i.e. généralement 50. Si tous les couples d'une classe apparaissent dans la liste de référence, nous obtenons le score maximum de 1 ; si aucun couple d'une classe n'apparaît dans la liste de référence, le score minimum de 0 est atteint.
La mesure idéale devrait être telle que ses valeurs les plus fortes (resp. faibles) soient associées à des "bons" (resp. "mauvais") candidats, i.e. des candidats appartenant (resp. n'appartenant pas) à la liste de référence. Autrement dit, l'histogramme d'une mesure idéale devrait associer aux classes d'équivalence regroupant les valeurs les plus fortes (resp. faibles) de la mesure un rapport proche de 1 (resp. 0). De plus, une mesure idéale devrait permettre de définir un seuil, ou éventuellement plusieurs, distinguant les bons des mauvais candidats. L'histogramme d'une mesure idéale devrait donc comporter zéro ou peu de classes prenant des valeurs moyennes. Un histogramme d'une mesure idéale ne montrant qu'un seuil serait le suivant :

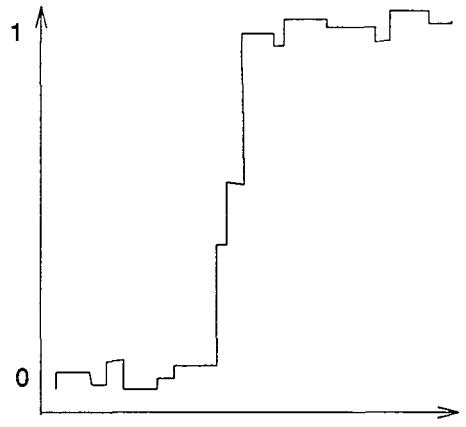


FIG. 4.6 - Histogramme idéal

Pour chaque histogramme, nous essayerons de dégager la courbe des tendances et d'examiner si celle-ci se rapproche de la courbe d'une mesure idéale comme la suivante :

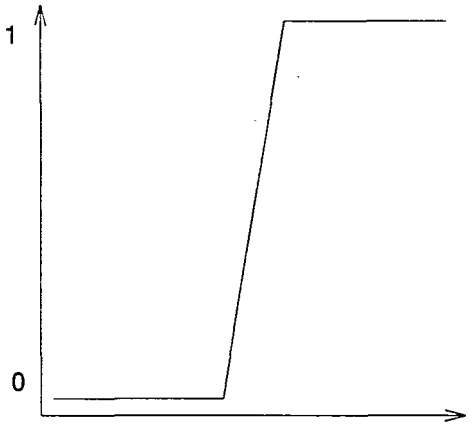


FIG. 4.7 - Courbe idéale

Lorsqu'un histogramme est irrégulier comme le suivant :

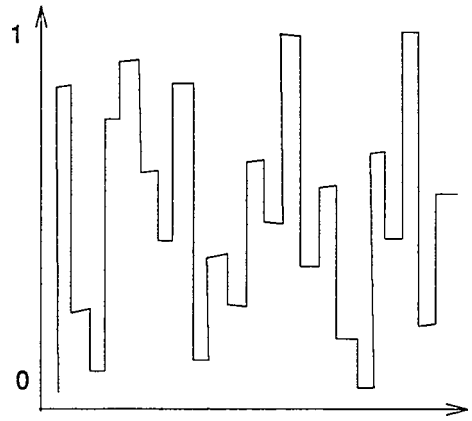


FIG. 4.8 - Histogramme irrégulier

il est impossible de dégager une courbe et nous savons alors que la mesure ne permet pas de sélectionner les "bons" candidats.
Avant d'examiner les histogrammes des mesures que nous avons présentées dans la section 4.3.5, précisons les deux points suivants :

- nous avons défini les classes d'équivalence comme regroupant en général 50 éléments successifs de la liste des candidats fournie par une mesure. Cependant, deux classes d'équivalence contiguës ne doivent pas contenir des candidats qui partagent une même valeur de la mesure. Par exemple, pour le score d'association, la classe d'effectif comprenant les valeurs les plus élevées, entre 13,65 et 11,19, contient 51 candidats puisque le cinquantième et le cinquante-et-unième partagent une même valeur. Ces classes sont généralement assez homogènes sauf pour la fréquence d'un couple : la classe des couples de deux occurrences regroupe en effet 1 500 couples.

- la comparaison avec la liste de référence n'a été effectuée que sur un sous-ensemble des couples. L'évaluation porte sur 2 600 couples et la liste de référence a été établie sur 2 200 couples. La représentativité d'une classe n'excède donc jamais 0,8 et des mesures dont la valeur maximale tend vers 0,6 sont retenues.

Nous donnons maintenant les histogrammes des mesures suivantes :

- Figure 4.9

  N1 Nombre d'occurrences des couples où un lemme donné apparaît comme premier élément du couple

N2 Nombre d'occurrences des couples où un lemme donné apparaît comme deuxième élément du couple

NC Nombre d'occurrences d'un couple

PHI2 Coefficient du Φ^2

OCH Coefficient d'Ochiai

YUL Coefficient de Yule

- Figure 4.10

FAG Coefficient de Fager et MacGowan

AFF Affinité

h1 Diversité appliquée à N_1 normalisée

h2 Diversité appliquée à N_2 normalisée

H1 Diversité appliquée à N_1

H2 Diversité appliquée à N_2

- Figure 4.11

LOG Coefficient de vraisemblance

MI Score d'association

MI2 Score d'association (numérateur au carré)

MI3 Score d'association (numérateur au cube)

KUC Coefficient de Kulczinsky

SMC Coefficient de Proximité Simple

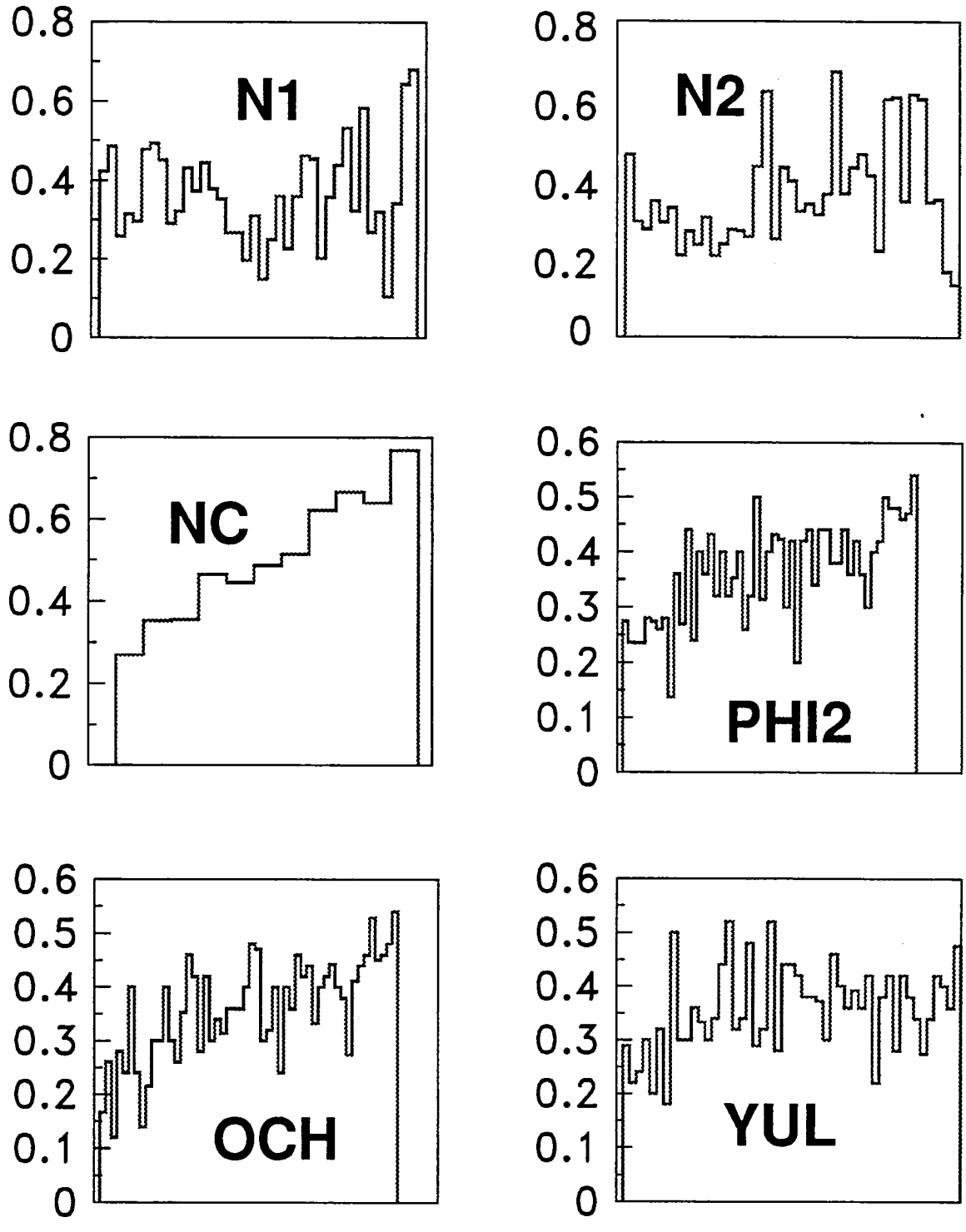


FIG. 4.9 - Histogrammes des modèles (partie 1/3)

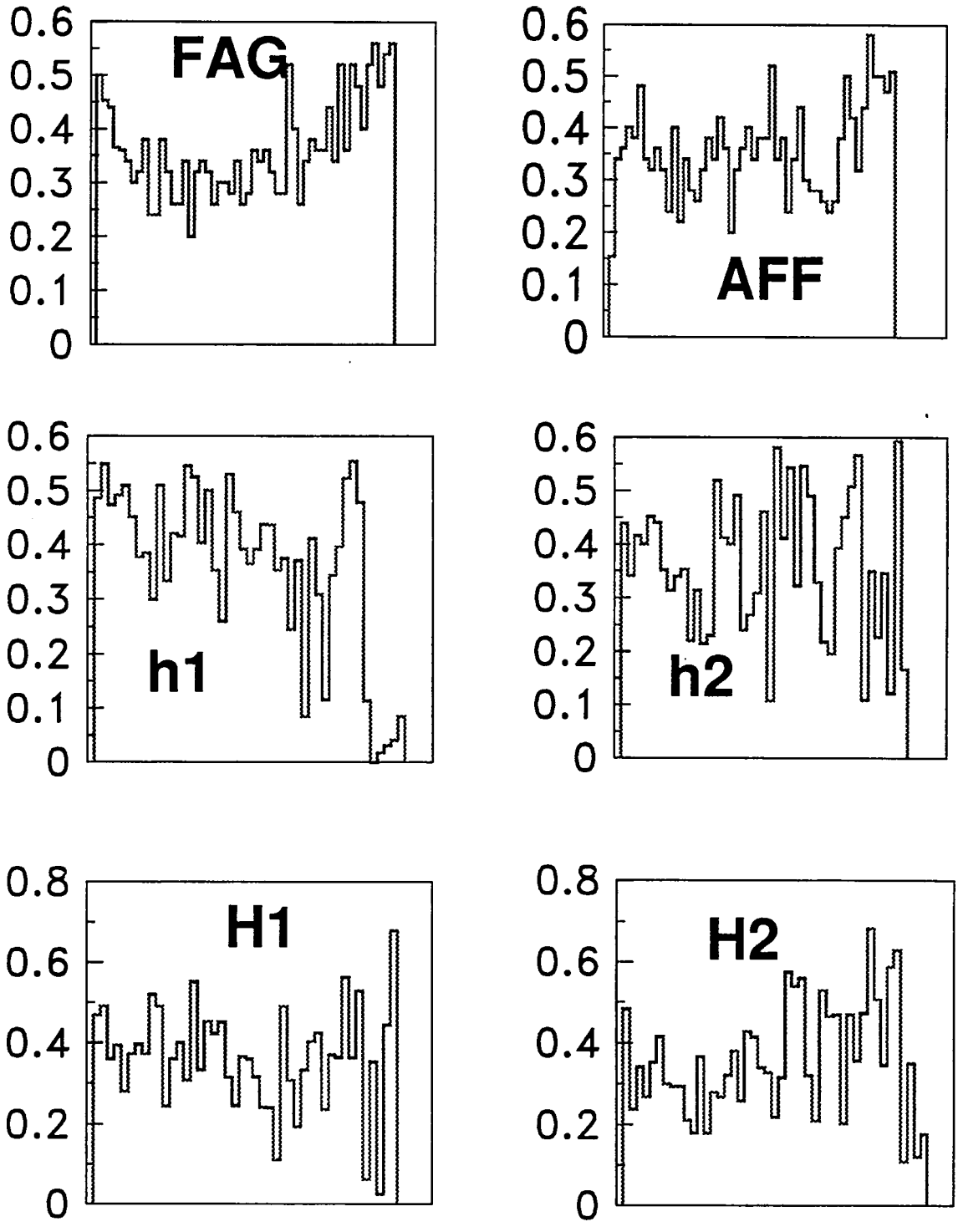


FIG. 4.10 - Histogrammes des modèles (partie 2/3)

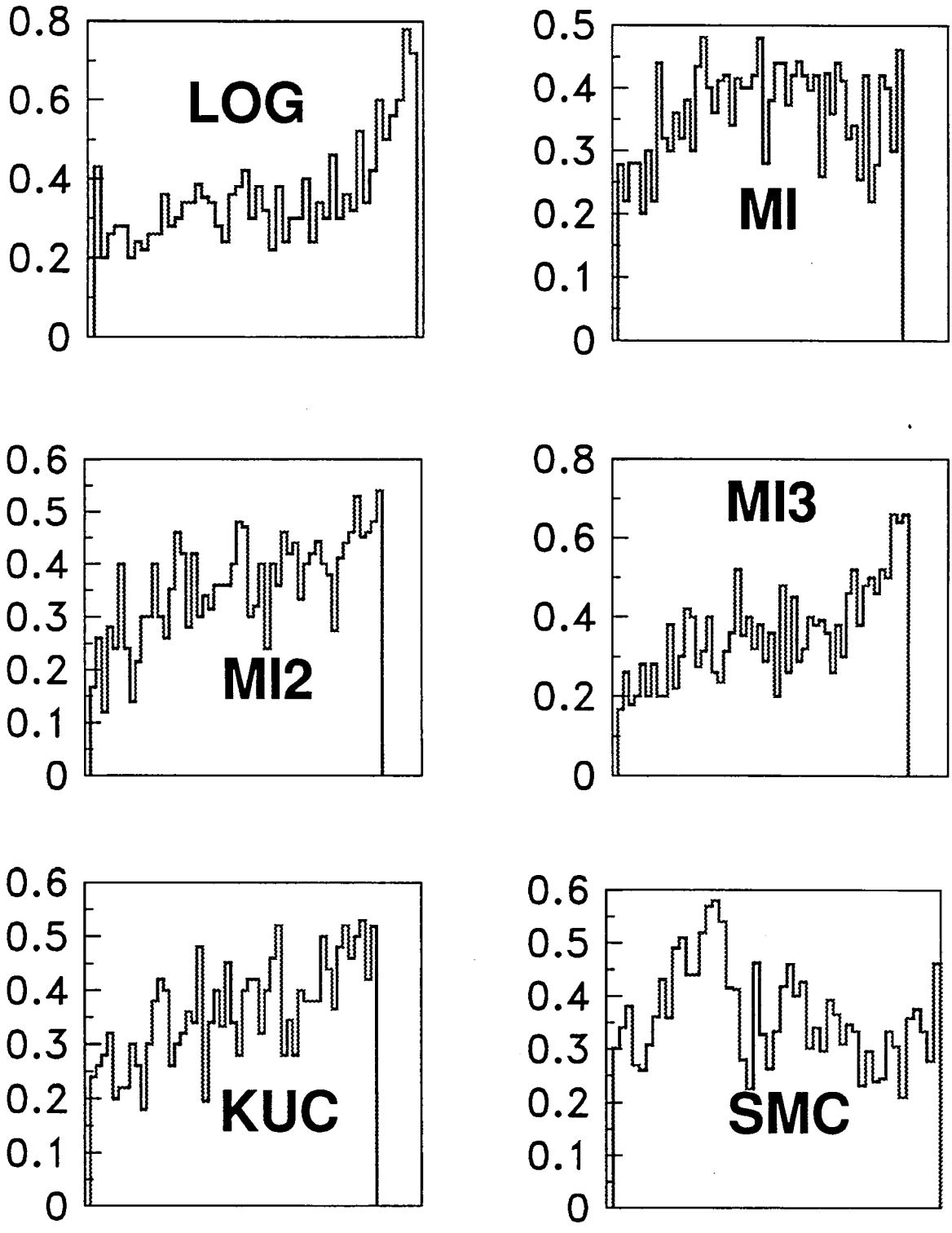


FIG. 4.11 - Histogrammes des modèles (partie 3/3)

L'examen des histogrammes montre qu'il y a bien croissance dans la plupart des cas. Cette croissance est plus ou moins marquée et présente des écarts plus ou moins importants. Examinons plus en détail les allures des histogrammes :

- Pour SMC, nous observons d'abord une croissance, puis une décroissance avec une stabilisation autour d'une valeur moyenne égale à 0,3. Ce palier signifie que dans les valeurs hautes se trouvent plus de mauvais candidats que de bons. La présence d'un pic dans les valeurs faibles montre que cette mesure rejette de bons candidats. Cette mesure n'est pas retenue.

- Pour KUC, la croissance est nette mais présente de gros écarts tantôt au-dessus, tantôt au-dessous de la pente. Il est très difficile avec une telle courbe de déterminer un seuil à partir duquel la proportion de bons candidats serait importante et ne varierait pas. Cette mesure n'est donc pas discriminatoire et n'est pas retenue.

- Pour OCH, la croissance est peu marquée, surtout dans les valeurs hautes, et de nombreux écarts apparaissent. Cette mesure n'est pas retenue.

- Pour FAG, la courbe décroît, se stabilise autour d'une valeur moyenne égale à 0,3 et croît à nouveau fortement. Cette courbe isole deux régions de valeurs, les faibles contenant une faible proportion de bons candidats et les fortes avec une proportion plus importante. Cette mesure peut avoir un rôle discriminatoire et nous l'avons retenue.

- Pour YUL, cette courbe est presque plate, de valeur moyenne 0,3, avec des écarts. Il très difficile d'exploiter une telle mesure et nous ne l'avons donc pas retenue.

- Pour PHI2, la courbe se décompose en trois parties : une croissance marquée avec de nombreux écarts, une croissance plus douce et enfin une forte croissance sur la fin. Comme pour FAG, il serait possible d'utiliser cette mesure pour les valeurs les plus hautes, mais un tel seuil ne retiendrait que peu de candidats. Nous n'avons pas retenu cette mesure.

- Pour LOG, excepté un écart dans les plus faibles valeurs, la croissance est faible, puis nettement marquée pour les valeurs hautes. Nous observons une réelle opposition entre valeurs faibles et fortes. De plus, la représentativité des candidats tend vers son maximum, soit 0,8. Cette mesure est retenue.

- Pour MI, la courbe est plate, de valeur moyenne 0,35 avec des écarts. Cette mesure n'est pas retenue.

- Pour MI2, la croissance est peu marquée avec de nombreux écarts. Cette mesure n'est pas retenue.

- Pour MI3, la croissance est bien marquée, surtout vers la fin, avec une opposition très contrastée entre les valeurs faibles et les valeurs fortes. Cette mesure est retenue.

- L'histogramme de NC (nombre d'occurrences du couple) est celui qui se rapproche le plus de la courbe idéale: croissance prononcée et uniforme vers une valeur maximale. Mais contrairement aux autres mesures, l'effectif des classes n'est pas constant et un examen approfondi de celles-ci montre qu'un nombre important de bons candidats se cachent dans la classe la plus faible (deux occurrences du couple). La fréquence n'est donc pas la mesure idéale, mais elle a tout de même l'avantage de préciser quels candidats elle rejette. Cette mesure est retenue.

- Les histogrammes de N1 et de N2 ne montrent aucune croissance: le fait qu'un lemme apparaisse souvent soit en première, soit en deuxième position dans un couple ne peut, en aucun cas, être discriminatoire.

- Pour AFF, nous remarquons deux paliers: l'un pour les valeurs faibles autour de 0,3 et l'autre pour les valeurs fortes autour de 0,45. Cette mesure n'est pas retenue.

- Pour h1, h2, H1, H2
  Ces histogrammes ne présentent aucune croissance. L'histogramme de h1 montre que ces valeurs fortes caractérisent des mauvais candidats. Cette mesure n'est pas retenue, mais pourra éventuellement être utilisée comme filtre.

L'étude précédente montre que peu de mesures sont susceptibles de servir notre objectif. Les meilleures sont LOG, coefficient de vraisemblance, FAG, coefficient de Fager et MacGowan, MI3, score d'association avec le numérateur au cube, et NC, le nombre d'occurrences du couple. En ne nous fiant qu'aux histogrammes, nous serions tentée de ne retenir que la fréquence; néanmoins, comme nous l'avons fait remarquer précédemment, la fréquence, par définition, ne peut pas retenir les termes peu fréquents. Il nous reste à déterminer si les quatre mesures que nous avons choisies sont complémentaires ou au contraire fortement liées, i.e. isolent-elles les mêmes candidats dans l'ensemble des couples? Nous avons donc étudié ces mesures d'un point de vue statistique, par le biais d'une matrice de co-variance. Les résultats de cette étude montrent que les mesures sont peu corrélées entre elles, le maximum (0,35) est atteint avec MI3 et FAG. Il semble donc difficile de pousser plus avant dans la sélection d'une mesure et d'éliminer certaines mesures parmi les restantes. Cette évaluation graphique montre qu'aucune mesure ne réalise véritablement l'objectif fixé: aucun histogramme ne laisse apparaître un seuil clair au-dessus duquel nous sommes sûre de ne rencontrer que des noms composés et en-dessous duquel il n'y aurait que

des couples ne correspondant pas à des noms composés. Nous allons maintenant examiner plus en détail le classement des couples proposés par ces quatre modèles statistiques. Nous examinerons aussi les résultats du score d'association, quoique celui-ci n'ait pas été retenu, ainsi que les résultats de la diversité normalisée.

4.5 Examen des résultats

L'évaluation graphique précédente ne retient que quatre modèles sur les dix-huit testés. L'élimination de ces modèles s'explique simplement : nous avons démontré que l'un des meilleurs critères de sélection d'un terme est son nombre important d'occurrences dans le corpus, et la plupart des critères d'association, dont le score d'association, privilégient les couples peu fréquents. Les critères d'association qui sont retenus, possèdent l'avantage de ne pas éliminer systématiquement ces couples fréquents, comme nous allons le voir. Les modèles statistiques ont été appliqués sur nos deux ensembles de couples : ceux de structures N_1 (PREP (DET)) N_2 et N ADJ. Nous conservons les deux corpus MTS (200 000 mots) et LBC (800 000 mots) pour évaluer si le paramètre de la taille des données est, ou non, important. Nous examinerons donc le classement proposé par la fréquence, puis ceux proposés par les critères d'association : ceux fournis par le score d'association et par le score d'association avec le numérateur au cube, par le critère de vraisemblance et par le critère de Fager et MacGowan. Nous présenterons, pour chacun de ces critères, les couples qui reçoivent les valeurs les plus fortes. Ces observations nous amèneront à ne retenir que le critère de vraisemblance ; un échantillon des listes de couples classés par cette mesure est donné en annexe. Nous terminerons par l'examen des informations apportées par la diversité normalisée et les mesures de distance. Ces deux dernières mesures ne sont pas discriminatoires mais sont à utiliser conjointement avec le critère d'association retenu.

4.5.1 Fréquence

La fréquence se révèle donc comme l'une des mesures les plus performantes pour détecter les termes d'un domaine. Ce résultat contredit un certain de nombres de travaux effectués dans le cadre des ressources lexicales monolingues dont ceux de [Church et Hanks, 1989] qui affirmaient que le score d'association était beaucoup plus efficace que la fréquence pour extraire les co-occurrences d'un corpus. Tous les couples fréquents que nous obtenons partagent tous un faible score d'association, et pourtant, il n'y a aucun doute sur leur caractère terminologique. Ce mauvais comportement du score d'association est un résultat important et est commenté un peu plus loin dans cette section (voir 4.5.2.1). Si les résultats de la fréquence remettent en cause un certain nombre de travaux effectués en statistique lexicale, ils confirment, par ailleurs, le bien fondé des travaux de

[Enguehard, 1992] sur l'apprentissage automatique de concepts qui prend uniquement en compte la fréquence de mots dans les textes. Par ailleurs, si le nombre important d'occurrences d'un couple est un élément déterminant, le nombre important de couples où un lemme donné apparaît en première position (resp. en deuxième position) ne l'est absolument pas. Ces résultats pourraient contredire ceux de [Bourigault, 1994] qui considère que la productivité d'un nom de tête est un élément déterminant pour statuer sur le caractère terminologique d'un groupe nominal ; cependant, l'auteur n'extrayant pas uniquement les noms composés binaires, il est exact que certains noms, comme dans notre domaine *système*, produisent systématiquement des noms composés binaires et des surcomposés : *système d'alimentation*, *système à débit binaire*, *système de télécommunications par satellite*, etc.. D'ailleurs, nous avons presque envie de qualifier ces composés binaires de surcomposés, tant la présence du nom *système* détermine leur caractère composé. Ce comportement n'est néanmoins pas partagé par tous les noms apparaissant fréquemment en tête d'un groupe nominal : *caractéristique*, *fonctionnement*, *utilisation*, *information*, etc, ne produisent pas ou rarement des composés.
Le classement proposé par la fréquence intègre très rapidement des couples qui ne correspondent pas à des noms composés : par exemple, le premier mauvais candidat est le couple (**cas, transmission**) et apparaît en 56ème position pour le corpus MTS. Pour fournir un élément de comparaison, ce couple correspond aussi au premier mauvais candidat du classement proposé par le coefficient de vraisemblance mais à la 176ème place. Nous verrons comment une partie de ces mauvais candidats peut être éliminée en utilisant les informations fournies par la diversité. Nous allons présenter les couples qui partagent les scores les plus importants de nombre d'occurrences.

N_1 (PREP (DET)) N_2

Le nombre d'occurrences des couples s'échelonne entre 2 et 223 pour MTS et entre 2 et 1 188 pour LBC. Pour chaque couple, nous préciserons pour plus de lisibilité la séquence morphosyntaxique la plus représentée accompagnée de sa fréquence, indiquée entre parenthèses ; nous ne préciserons ici ni les déterminants, ni les variations de préposition, ni les modifieurs susceptibles d'apparaître sous un couple lorsque ceux-ci sont peu représentés (les séquences morphosyntaxiques apparaissant sous le couple sont commentées en partie avec les mesures de distances (section 4.5.4)). La fréquence décroît très vite : dans le corpus MTS, un seul couple possède plus de 200 occurrences, quatre couples possèdent entre 100 et 200 occurrences ; dans le corpus LBC, un seul couple possède plus de 1 000 occurrences, 3 entre 500 et 1 000, 14 entre 200 et 500, ... Les couples les plus fréquents sont listés ci-dessous. Nous utilisons les notations suivantes :

Nbc Nombre d'occurrences du couple

MI Score d'association associé au couple

Corpus	Couple de structure N_1 (PREP (DET)) N_2	Séquence la plus fréquente	Nbc	MI
MTS	(largeur, bande)	*largeur de bande* (197)	**223**	5,73
	(bande, base)	*bande de base* (142)	**145**	5,52
	(amplificateur, puissance)	*amplificateur(s) de puissance* (137)	**137**	5,66
	(température, bruit)	*température de bruit* (110)	**126**	6,18
	(système, satellite)	*système(s) à satellites* (89)	**108**	1,81
	(télécommunication, satellite)	*télécommunication(s) par satellite* (88)	**99**	4,09
	(réseau, satellite)	*réseau(x) à satellites* (62)	**97**	2,58
	(temps, propagation)	*temps de propagation* (93)	**94**	6,89
	(bande, fréquence)	*bande(s) de fréquences* (85)	**89**	3,54
	(système, signalisation)	*système(s) de signalisation* (82)	**85**	2,81
	(liaison, satellite)	*liaison(s) par satellites* (73)	**82**	2,72

Corpus	Couple de structure N_1 (PREP (DET)) N_2	Séquence la plus fréquente	Nbc	MI
LBC	(canal, sémaphore)	*canal/canaux sémaphores* (1 188)	**1 188**	4,77
	(système, signalisation)	*système(s) de signalisation* (839)	**847**	3,60
	(point, sémaphore)	*point(s) sémaphore(s)* (677)	**679**	3,57
	(accusé, réception)	*accusé(s) de réception* (592)	**592**	6,37
	(signal, fin)	*signal/signaux de fin* (385)	**391**	4,20
	(trame, sémaphore)	*trame(s) sémaphore(s)* (354)	**354**	4,55
	(message, adresse)	*message(s) d'adresse* (187) *message initial d'adresse* (120)	**337**	4,14
	(réception, signal)	*réception du signal* (143) *réception d'un signal* (126)	**332**	3,45
	(unité, signalisation)	*unité(s) de signalisation* (320)	**320**	3,63
	(réseau, sémaphore)	*réseau(x) sémaphore(s)* (283)	**283**	3,27
	(connexion, sémaphore)	*connexion(s) sémaphore(s)* (274)	**274**	3,16

N ADJ

Le nombre d'occurrences des couples s'échelonne entre 2 et 750 pour MTS et entre 2 et 340 pour LBC. Les couples N ADJ les plus fréquents et présentés ci-dessous ne sont pas tous des noms composés de type élémentaire: dans le corpus MTS, *service fixe* est une sous-séquence d'un nom composé de type élémentaire de longueur 3: *service fixe par satellite.* Notons, là encore, que le score d'association prend des valeurs moyennes et ne considère donc pas ces couples fréquents comme des termes du domaine. Les valeurs du score d'association sont cependant plus régulières que pour les N_1 (PREP (DET)) N_2.

Corpus	Couple de structure N ADJ	Nbc	MI	Corpus	Couple de structure N ADJ	Nbc	MI
MTS	(station, terrien)	**750**	3,37	LBC	(service, supplémentaire)	**340**	4,33
	(débit, binaire)	**134**	5,32		(centre, international)	**325**	3,63
	(voie, téléphonique)	**118**	4,75		(équipement, terminal)	**275**	5,43
	(accès, multiple)	**105**	5,66		(considération, général)	**256**	5,38
	(liaison, montant)	**88**	5,17		(circuit, international)	**213**	2,23
	(secteur, spatial)	**79**	4,80		(réseau, national)	**208**	3,26
	(liaison, descendant)	**77**	5,22		(niveau, relatif)	**202**	4,16
	(service, fixe)	**66**	5,33		(entité, fonctionnel)	**199**	5,42
	(engin, spatial)	**57**	5,28		(caractère, graphique)	**196**	5,86
	(station, spatial)	**56**	1,17		(adresse, complète)	**183**	5,30
	(station, distant)	**44**	2,88		(effet, local)	**169**	5,43

4.5.2 Critères d'Association

Des dix critères présentés dans la section 4.3.2, l'évaluation graphique n'a permis d'en retenir que trois : le score d'association avec le numérateur au cube (formule 4.9), le coefficient de vraisemblance (formule 4.11) et le critère de Fager et MacGowan (formule 4.4). Ces critères ont été retenus car, contrairement aux autres critères, ils n'éliminent pas systématiquement les couples fréquents. Chacun de ces critères propose un traitement conceptuel des couples. Nous allons successivement examiner ces trois critères et essayer de déterminer quelles sont les particularités de leurs classements.

4.5.2.1 Score d'association et score d'association avec le numérateur au cube

Les mauvais résultats du score d'association nous ont surpris : les couples partageant les valeurs les plus fortes possédaient tous un faible nombre d'occurrences (2 à 3) et isolaient plus des noms composés figés comme *aiguille d'une montre*, *béton armé*, des adverbes figés comme *dos à dos* que véritablement des termes du domaine des télécommunications. La valeur forte découlait du fait que les deux lemmes apparaissaient uniquement ensemble et jamais dans d'autres couples. L'introduction de filtres linguistiques et l'acceptation des occurrences basses a sans doute influé sur les résultats des travaux obtenus sur des corpus non traités, sans filtrage préliminaire et à partir d'un seuil d'occurrences de 5. Ces mauvais résultats nous ont amené à modifier la formule du score d'association de manière à ce qu'il ne rejette pas systématiquement les couples fréquents. Nous sommes consciente que la modification de la formule du score d'association relève de l'empirisme. Les valeurs faibles du score d'association et du score d'association au cube sont réservées aux couples dont les deux lemmes apparaissent rarement ensemble et fréquemment séparément comme les couples (**système**, **terre**), (**code**, **signalisation**), (**bande**, **bruit**).

Nous ne présenterons pas les couples qui possèdent les valeurs les plus fortes de score d'association au cube car ce sont les mêmes couples qui se voient attribuer

les valeurs les plus fortes du coefficient de vraisemblance. Nous notons simplement que pour le patron N_1 (PREP (DET)) N_2, les valeurs du score d'association au cube s'échelonnent entre -1,01 et 21,34 pour le corpus MTS et entre -4,51 et 25,20 pour le corpus LBC, et pour le patron N ADJ, entre -0,7 et 22,4 pour le corpus MTS et entre -2,28 et 21,89 pour le corpus LBC.

4.5.2.2 Coefficient de vraisemblance

Le coefficient de vraisemblance sélectionne les mêmes couples que le score d'association au cube dans ses valeurs fortes. Au contraire, il n'est pas défini si l'un des deux lemmes d'un couple apparaît seulement dans ce couple. Nous pouvons remarquer que lorsque le coefficient de vraisemblance n'est pas défini, le coefficient de Yule (équation 4.5) prend sa valeur maximale, c'est-à-dire 1, que les valeurs du score d'association et du coefficient de Fager et MacGowan sont parmi les plus fortes et que l'une des diversités associée à l'un des lemmes du couple prend la valeur 0. La diversité est donc plus précise puisqu'elle indique si les deux lemmes apparaissent uniquement ensembles comme pour (**océan**, **indien**) (H_1=h_1=H_2=h_2=0), ou sinon, lequel des deux lemmes n'apparaît qu'avec l'autre comme pour (**réseau**, **maillé**) (H_2=h_2=0 donc *maillé* n'apparaît qu'avec *réseau*) ou pour (**codeur**, **idéal**) (H_1=h_1=0 où le nom *codeur* n'apparaît qu'avec l'adjectif *idéal*). D'autres exemples sont : (**île**, **salomon**), (**hélium**, **gazeux**), (**suppresseur**, **écho**), (**retour**, **chariot**). Ces couples pour lesquels le coefficient de vraisemblance n'est pas défini regroupent beaucoup de noms composés figés et de co-occurrences lexicales restreintes de la langue courante.

N_1 (PREP (DET)) N_2

Les valeurs du coefficient de vraisemblance s'échelonnent entre 0+ et 1328 pour le corpus MTS et 0+ et 5738 pour le corpus LBC. L'amplitude des valeurs dépend du nombre d'occurrences du couple: plus le couple est fréquent plus la valeur du coefficient de vraisemblance tend à être élevée et ce, indépendamment du nombre total de couples extraits. Les couples pour lesquels ce coefficient n'est pas défini sont au nombre de 167 pour MTS et 169 pour LBC. Les notations suivantes sont utilisées dans les tableaux ci-dessous:

Logl Coefficient de vraisemblance (Loglike)

Mi3 Score d'association au cube

Nbc Nombre d'occurrences du couple

Corpus	Couple de structure N_1 (PREP (DET)) N_2	Séquence la plus fréquente	**Logl**	Mi3	Nbc
MTS	(largeur, bande)	*largeur de bande* (197)	**1328**	21,34	223
	(température, bruit)	*température de bruit* (110)	**777**	20,13	126
	(bande, base)	*bande de base* (142)	**745**	19,88	145
	(amplificateur, puissance)	*amplificateur(s) de puissance* (137)	**728**	19,86	137
	(temps, propagation)	*temps de propagation* (93)	**612**	19,80	94
	(règlement, radiocommunication)	*règlement des radiocommunications* (60)	**521**	19,96	60
	(produit, intermodulation)	*produit(s) d'intermodulation* (61)	**458**	19,32	61
	(taux, erreur)	*taux d'erreur* (70)	**420**	18,61	70
	(mise, œuvre)	*mise en œuvre* (47)	**355**	18,60	47
	(télécommunication, satellite)	*télécommunication(s) par satellite* (88)	**353**	17,35	99
	(bilan, liaison)	*bilan(s) de liaison* (37)	**344**	17,99	55

Corpus	Couple de structure N_1 (PREP (DET)) N_2	Séquence la plus fréquente	**LogL**	Mi3	Nbc
LBC	(canal, sémaphore)	*canal/canaux sémaphores* (1 188)	**5 738**	25,20	1 188
	(accusé, réception)	*accusé de réception* (558)	**3 983**	24,78	592
	(système, signalisation)	*système(s) de signalisation* (82)	**2 417**	23,05	85
	(complément, étude)	*complément d'étude* (242)	**1985**	23,74	245
	(point, sémaphore)	*point(s) sémaphore(s)* (677)	**1 822**	22,38	679
	(intervalle, temps)	*intervalle(s) de temps* (249)	**1 782**	23,19	251
	(trame, sémaphore)	*trame(s) sémaphore(s)* (354)	**1 444**	21,48	354
	(signal, fin)	*signal/signaux de fin* (385)	**1 407**	21,43	391
	(sou-système, utilisateur)	*sous-système utilisateur* (195)	**1 226**	22,10	195
	(bout, bout)	*bout en bout* (136)	**1 155**	22,58	137
	(contrôle, continuïté)	*contrôle(s) de continuïté* (171)	**1 116**	21,89	171

N ADJ

Les valeurs du coefficient de vraisemblance s'échelonnent entre 0+ et 2934 pour le corpus MTS et entre 0+ et 1435 pour le corpus LBC. Les couples pour lesquels ce coefficient n'est pas défini sont au nombre de 147 pour MTS et 176 pour LBC.

Corpus	Couple de structure N ADJ	**Logl**	Mi3	Nbc
MTS	(station, terrien)	**2934**	22,47	750
	(débit, binaire)	**716**	19,45	134
	(accès, multiple)	**605**	19,10	105
	(voie, téléphonique)	**512**	18,52	118
	(liaison, montant)	**457**	17,50	88
	(liaison, descendant)	**408**	17,50	77
	(secteur, spatial)	**341**	17,41	79
	(service, fixe)	**326**	17,42	66
	(lobe, latéral)	**299**	17,93	40
	(faisceau, hertzien)	**244**	17,22	35
	(puissance, surfacique)	**205**	16,76	35

Corpus	Couple de structure N ADJ	**Logl**	Mi3	Nbc
LBC	(équipement, terminal)	**1425**	21,63	275
	(considération, général)	**1385**	21,38	256
	(service, supplémentaire)	**1275**	21,15	340
	(télégraphie, harmonique)	**1250**	21,89	152
	(étude, ultérieur)	**1171**	21,52	169
	(caractère, graphique)	**1112**	21,09	196
	(entité, fonctionnel)	**999**	20,70	199
	(centre, international)	**964**	20,32	325
	(adresse, complet)	**874**	20,33	183
	(effet, local)	**865**	20,23	169
	(station, mobile)	**855**	20,41	164

4.5.2.3 Critère de Fager et MacGowan

Les valeurs fortes du critère de Fager et MacGowan sont attribuées principalement aux couples dont les deux lemmes apparaissent souvent ensemble et rarement séparément. Cette mesure est très proche du score d'association, mais ne refuse pas systématiquement les termes fréquents. Les valeurs négatives du critère de Fager et MacGowan sont attribuées aux couples dont le premier lemme n'apparaît qu'avec un second lemme employé dans un grand nombre de couples, i.e. la diversité sur le premier lemme est nulle, comme pour (prestataire, service), (cheminement, information), (ingénierie, trafic).
Les valeurs du critère de Fager et MacGowan s'échelonnent dans un intervalle maximal de]-1,+1[.

N_1 (PREP (DET)) N_2

Les valeurs du critère de Fager et MacGowan s'échelonnent entre -0,310 et 0,849 pour MTS et -0,326 et 0,827 pour LBC. Les couples extraits caractérisent des termes du domaine des télécommunications comme *moteur d'apogée*, des noms composés de la langue courante comme *accusé de réception*, des adverbes figés comme *bout à bout*. Nous retrouvons des noms composés déjà isolés par le score d'association comme *aiguille d'une montre* et *glossaire du fascicule*. Les notations suivantes sont utilisées :

Fag Critère de Fager et MacGowan

Nbc Nombre d'occurrences du couple

Corpus	Couple de structure N_1 (PREP (DET)) N_2	Séquence la plus fréquente	Fag	Nbc
MTS	(arséniure, gallium)	*arséniure de gallium* (11)	**0,849**	11
	(mémoire, tampon)	*mémoire(s) tampon(s)* (35)	**0,823**	35
	(égalité, droit)	*égalité de droits* (4)	**0,776**	5
	(règlement, radiocommunication)	*règlement des radiocommunications* (60)	**0,748**	60
	(batterie, accumulateur)	*batterie(s) d'accumulateurs* (11)	**0,711**	11
	(reportage, actualité)	*reportage(s) d'actualités* (2)	**0,711**	3
	(registre, décalage)	*registre à décalage* (7)	**0,698**	7
	(largeur, bande)	*largeur de bande* (197)	**0,691**	223
	(moteur, apogée)	*moteur d'apogée* (26)	**0,691**	28
	(aiguille, montre)	*aiguilles d'une montre* (2)	**0,646**	2
	(glossaire, fascicule)	*glossaire du fascicule* (2)	**0,646**	2

Corpus	Couple de structure N_1 (PREP (DET)) N_2	Séquence la plus fréquente	Fag	Nbc
LBC	(isolement, processeur)	*isolement de processeur* (24) *isolement du processeur* (11) *isolement des processeurs* (6)	**0,827**	41
	(complément, étude)	*complément d'étude* (242)	**0,731**	245
	(dos, dos)	*dos à dos* (3)	**0,711**	3
	(force, son)	*force des sons* (35)	**0,690**	38
	(accusé, réception)	*accusé de réception* (558)	**0,683**	592
	(compensation, dérive)	*compensation de dérive* (11)	**0,663**	13
	(microphone, charbon)	*microphone(s) à charbon* (35)	**0,658**	35
	(bout, bout)	*bout en bout* (136)	**0,640**	137
	(recommandation, série)	*recommandation de la série* (100)	**0,629**	123
	(tiers, octave)	*tiers d'octave* (14)	**0,625**	15
	(rapidité, modulation)	*rapidité(s) de modulation* (112)	**0,596**	117

N ADJ

Les valeurs du critère de Fager et MacGowan s'échelonnent entre -0,281 et 0,896 pour MTS et -0,306 et 0,885 pour LBC. Les couples retenus sont des termes du domaine des télécommunications comme *lobes latéraux* ou *station terrienne*, des noms propres comme *île Salomon* ou *océan indien*, des conjonctions de subordination composées comme *compte tenu*, des co-occurrences lexicales restreintes comme *période probatoire*. Seul le candidat *étude ultérieure* n'est pas franchement convaincant, mais peut éventuellement être considéré comme une co-occurrence lexicale restreinte.

Corpus	Couple de structure N ADJ	Fag	Nbc
MTS	(compte, tenu)	**0,896**	23
	(station, terrien)	**0,842**	750
	(publication, anticipée)	**0,823**	8
	(stabilisation, triaxiale)	**0,811**	7
	(rafraîchissement, conditionnel)	**0,811**	7
	(île, salomom)	**0,750**	4
	(distorsion, intersymbole)	**0,737**	6
	(couple, perturbateur)	**0,715**	7
	(hélium, gazeux)	**0,711**	3
	(lobe, latéral)	**0,699**	40
	(atmosphère, clair)	**0,693**	12

Corpus	Couple de structure N ADJ	Fag	Nbc
LBC	(télégraphie, harmonique)	**0,885**	152
	(compte, tenu)	**0,852**	66
	(polynôme, générateur)	**0,849**	11
	(faute, matériel)	**0,828**	38
	(période, probatoire)	**0,800**	75
	(assemblée, plénière)	**0,796**	6
	(accord, bilatéral)	**0,787**	80
	(signe, diacritique)	**0,755**	31
	(océan, indien)	**0,750**	4
	(étude, ultérieur)	**0,735**	169
	(prise, simultané)	**0,683**	78

Pour conclure, après l'examen des classements proposés par ces trois critères, nous n'avons envie de ne retenir que celui du critère de vraisemblance. En effet, le critère de Fager et MacGowan donne trop d'importance aux couples dont les deux lemmes apparaissent souvent ensemble et peu séparément ; le score d'association

au cube donne de bons résultats mais, sa formule relevant de l'empirisme, nous répugnons à la choisir. Le critère de vraisemblance, proposé par [Dunning, 1993], possède les qualités suivantes :

- il est un véritable test statistique,
- il propose un classement qui prend en compte la fréquence du couple,
- il se comporte aussi bien sur un corpus moyen (MTS) que sur un corpus plus important (LBC),
- il n'est pas défini pour un certain nombre de couples qui admettent une diversité nulle sur l'un des éléments : ces couples qui sont retenus systématiquement par les autres mesures peuvent posséder un caractère figé et appartiennent le plus souvent à la langue générale. Il est donc important qu'ils soient isolés.

Nous donnons en annexe C un échantillon du classement proposé par le critère de vraisemblance pour nos deux corpus et nos deux structures syntaxiques. Nous incluons la liste des couples pour lesquels ce coefficient n'est pas défini et dont la moitié correspond effectivement à des termes du domaine. Ces listes comprennent les couples, ainsi que l'expression rationnelle correspondante lorsque le couple n'a été rencontré que sous une ou au maximum deux séquences différentes. Pour les autres couples, i.e. qui admettent plus de deux formes différentes, aucune séquence n'est indiquée. D'autre part, pour fournir un élément de comparaison, nous avons indiqué en plus de la valeur du critère de vraisemblance, la fréquence, la valeur du score d'association au cube, et la valeur du coefficient de Fager et MacGowan. Nous avons aussi indiqué les valeurs prises par la diversité normalisée pour les raisons que nous allons maintenant précisées.

4.5.3 Diversité

La diversité normalisée par le nombre d'occurrences du couple donne des informations intéressantes sur la distribution des lemmes du couple dans l'ensemble des couples. Un lemme ayant une diversité importante signifie qu'il est employé dans un grand nombre de couples à proportion relativement égale ; à l'inverse, un lemme employé uniquement dans un couple possède une diversité nulle (valeur minimale) et ceci quelle que soit la fréquence du couple. Nous avons déjà vu quelques exemples de couples caractérisés par une diversité nulle sur l'un de leurs lemmes : ce sont les couples pour lesquels le coefficient de vraisemblance n'est pas défini, le coefficient de Yule prend sa valeur maximale, etc. En voici de nouveau deux exemples : étant donné l'adjectif *anormal*, dans le schéma syntaxique N ADJ, seul le nom *fonctionnement* a été rencontré et on a $H_2 = H_{anormal} = 0$; le nom *fonctionnement* peut néanmoins très bien apparaître avec d'autres adjectifs. Inversement, étant donné le nom *éclaircissement*, toujours dans le schéma syntaxique

N ADJ, seul l'adjectif *nécessaire* a été rencontré et on a $H_1 = H_{claicissement} = 0$; ce qui n'implique évidement pas que l'adjectif *nécessaire* n'apparaît pas avec d'autres noms. La diversité a été calculée pour chaque élément apparaissant dans un couple à une place fixe.

N_1 (PREP (DET)) N_2

1. Diversité appliquée à N_1

   - Pour le corpus MTS, les dix scores les plus importants sont assignés aux noms : *fonctionnement*, *fonction*, *moyen*, *type*, *partie*, *compte*, *cas*, *raison* , *exemple*, *signal*, *caractéristique*.
   - Pour le corpus LBC, les dix scores les plus élevés sont assignés aux noms : *fonctionnement*, *cas*, *partie*, *signalisation*, *utilisation*, *moyenne*, *compte*, *moyen*, *exemple*, *procédure*.

   Alors que les valeurs élevées des mesures précédentes ne caractérisaient pas les même couples au travers des deux corpus, il est intéressant de noter ici que les valeurs de grande diversité pour le N_1 sont partagées par les mêmes noms dans les deux corpus et ce, malgré leur différence de taille. De plus, une diversité importante sur le N_1 caractérise soit des noms apparaissant à l'intérieur d'une préposition composée comme *en fonction de*, *au moyen de*, *en raison de*, *en cas de*, *en tenant compte de* ou *compte tenu de*, soit des quantificateurs comme *nombre de*, des classificateurs comme *partie de*, *type de*.

2. Diversité appliquée à N_2

   - Pour le corpus MTS, les dix scores les plus importants sont assignés aux noms : *système*, *station*, *réseau*, *signal*, *service*, *équipement*, *fonctionnement*, *antenne*, *fréquence*, *niveau*.
   - Pour le corpus LBC, les dix scores les plus élevés sont assignés aux noms : *système*, *service*, *fonctionnement*, *équipement*, *niveau*, *fonction*, *réseau*, *commutateur*, *circuit*, *signaleur*.

   Là encore, malgré la différence de taille des corpus, les listes sont presque identiques. Les noms avec une forte diversité sur N_2 semblent caractéristiques du domaine (à part certains comme *fonction* ou *niveau*), nous pourrions les appeler mots-clés du domaine des télécommunications.

N ADJ

1. Diversité appliquée à ADJ
   La diversité appliquée à l'adjectif du patron N ADJ permet d'identifier des

adjectifs n'entrant pas dans la composition du terme comme les adjectifs: *nécessaire, suivant, important, différent, particulier, tel, possible, ...* Là encore il est intéressant de remarquer que ce sont les mêmes adjectifs qui apparaissent dans les deux listes associées aux deux corpus.

- Pour le corpus MTS, les dix scores de diversité les plus importants sont assignés aux adjectifs: *nécessaire, suivant, différentiel, numérique, général, important, supplémentaire, particulier, relatif, différent.*
- Pour le corpus LBC, les dix scores de diversité les plus importants sont assignés aux adjectifs: *suivant, différentiel, différent, correspondant, particulier, possible, spécifique, tel, nécessaire, normal.*

2. Diversité appliquée à N
   Les noms qui possèdent les plus grandes valeurs de diversité sous le patron N ADJ là aussi semblent caractéristiques du domaine à l'exception de quelques-uns.

   - Pour le corpus MTS, les dix scores de diversité les plus importants sont assignés aux noms: *valeur, fonctionnement, système, réseau, caractéristique, équipement, niveau, signal, antenne, satellite.*
   - Pour le corpus LBC, les dix scores de diversité les plus importants sont assignés aux noms: *signal, fonctionnement, façon, signalisation, fonction, système, valeur, mesure, caractéristique, méthode.*

Dans [Daille, 1993], nous avions montré comment en utilisant la diversité, i.e. en éliminant un certain nombre de couples qui recevaient une valeur forte de diversité, nous arrivions à repousser le seuil de fréquence à partir duquel nous n'avions que des bons candidats. Avec le classement proposé par le coefficient de vraisemblance, les couples avec une grande diversité ne se voient pas attribuer les valeurs les plus fortes, et le bruit est d'une autre nature. Néanmoins, celle-ci pourrait se révéler utile plus avant dans le classement et nous avons décider d'indiquer les valeurs prises par la diversité normalisée appliquée au N_1 pour les couples de structure N_1 (PREP (DET)) N_2, et la diversité appliquée à l'adjectif pour les couples N ADJ.

4.5.4 Moyenne et variance des distances

À ces mesures statistiques, il faut rajouter les mesures de distances qui peuvent nous aider à répondre aux questions suivantes: un candidat nom composé de patron principal N_1 PREP (DET) N_2 est-il toujours employé sans déterminant devant le N_2, avec un déterminant, ou bien ce déterminant est-il optionnel? Nous allons voir que les couples relevés avec le patron N_1 (PREP (DET)) N_2 acceptent, en moyenne, beaucoup plus de variations en distance que les couples relevés avec

le patron N ADJ. Ces mesures de distances donnent des indications intéressantes sur certaines variations morphosyntaxiques des occurrences des couples, mais ces indications ne permettent en aucun cas de juger du statut de nom composé des couples retenus. Un couple qui ne connaît pas de variation en distance, et ceci quelle que soit la distance, est ou non, un nom composé: voici des exemples de couples qui ne varient pas en distance et qui ne sont pas des noms composés: *côté abonné*, *paire de signal*, *type d'antenne*, *organigramme de la figure*. La même remarque est valable pour les couples qui admettent des variations en distance. Néanmoins, pour illustrer plus clairement les propriétés des mesures de distances, nous avons principalement sélectionné dans les exemples ci-dessous, des couples référant à des termes du domaine.

N_1 (PREP (DET)) N_2

Les mesures de distance permettent de classer le terme dans différentes structures de type élémentaire: soit N_1 N_2, soit N_1 PREP N_2, soit N_1 PREP DET N_2, soit N_1 ADJ PREP (DET) N_2, ou encore N_1 PREP N_2 CONJ PREP N_3. Ce classement s'effectue uniquement sur la distance et des variations sont possibles: avec la structure N_1 PREP N_2 sur la préposition, avec la structure N_1 PREP DET N_2 sur le déterminant et/ou la préposition. Plus la distance entre les deux noms est importante, plus les variations de structures augmentent.
Dans les exemples ci-dessous, nous allons utiliser les notations suivantes: *Dist*, moyenne des longueurs des occurrences relevées pour chaque couple, *Var*, la variance attachée à ces distances, *MDist*, moyenne des longueurs des occurrences relevées pour chaque couple sans prise en compte des mots grammaticaux. Nous indiquons, pour chaque couple, les flexions et les variations observées.

1. **Couples n'acceptant aucune variation** ($Var = 0$)
   Ces couples représentent 70 % de l'ensemble des couples pour le corpus MTS et 65 % pour le corpus LBC.

   (a) N_1 N_2: $Dist = 2$ $MDist = 2$
   - *côté usager*
   - *signal/signaux multifréquence*
   - *liaison sémaphore*, *liaisons sémaphores*
   - *canal support*, *canaux support*, *canaux supports*

   (b) N_1 PREP N_2: $Dist = 3$ $MDist = 2$
   - *mise en œuvre*
   - *accusé(s) de réception*
   - *refroidissement à air*, *refroidissement par air*
   - *catégorie(s) de réponse*, *catégories de réponses*
   - *commande à codage*, *commande de codage*, *commande par codage*

(c) N_1 PREP DET N_2 : $Dist = 4$ $MDist = 2$

- *encombrement à la réception*
- *mise au point*
- *sensibilité au bruit*
- *blocage de l'organe*
- *reconnaissance des signaux*
- *principe(s) du codage, principes d'un codage*

(d) N_1 ADJ PREP N_2 : $Dist = 4$ $MDist = 3$

- *adresse complète hors bande*
- *période probatoire d'urgence*
- *réseau local de lignes, réseaux locaux de lignes*
- *effet local avec masquage, effet local par masquage*
- *information émise en réponse, information envoyée en réponse*

(e) N_1 ADJ PREP DET N_2 : $Dist = 5$ $MDist = 3$

- *retour obligatoire à l'option, retour facultatif à l'option*
- *longueur fixe d'un octet*
- *voies(s) dépendante(s) du code, voie(s) indépendante(s) du code*

(f) N_1 PREP N_2 CONJ PREP N_3 : $Dist = 6$ $MDist = 3$

- *interface de commande et de saisie*
- *procédure(s) de blocage et de déblocage*

2. **Couples acceptant des variations** ($Var \neq 0$)
Ces couples représentent 30 % de l'ensemble des couples pour le corpus MTS et 35 % pour LBC. Ce score confirme notre étude linguistique et celle de [Jacquemin, 1991] : les termes d'un domaine technique ne sont pas des structures morphosyntaxiques figées. Quelques exemples :

- (demande, trafic)
  demande de trafic
  demandes en trafic
  demande réelle en trafic
- (liaison, satellite)
  liaison par satellite, liaisons par satellite
  liaisons (très rapides + numériques + téléphoniques nationales) par satellite
  liaisons numériques par satellites
  liaisons satellite
  liaisons entre satellites

- (signal, fin)
  signal de fin , signaux de fin
  signal (local + national + valide + périodique) de fin
  signal émis à des fins
  signal numérique utilisé à des fins
  Les deux dernières occurrences illustrent le problème présenté au début de ce chapitre: nous n'avons aucune assurance qu'un couple regroupe des co-occurrences désignant un seul et unique concept. Les polysémies de cet ordre restent rares.
- (ligne, abonné)
  ligne d'abonné, lignes d'abonné
  ligne de l'abonné, lignes de l'abonné
  ligne d'abonnés, lignes des abonnés
  ligne(s) (téléphonique(s) + numériques(s) + analogique(s)) d'abonné
  ligne(s) (numérique(s) + analogique(s)) de l'abonné
  lignes et services d'abonné

Contrairement à certaines idées reçues, les modifieurs pouvant s'insérer à l'intérieur d'un nom composé donné sont en nombre réduit. L'enregistrement de ces modifieurs, comme des autres altérations que subit la structure de base, y compris les différentes flexions rencontrées, n'est donc pas une tâche incommensurable surtout si celle-ci est effectuée automatiquement. Ces informations lexicales sont présentes sous l'entrée de chaque couple et pourront donc être directement intégrées dans un dictionnaire.

N ADJ

Les candidats noms composés de type N ADJ qui acceptent des modifications sont beaucoup moins nombreux que les N_1 (PREP (DET)) N_2. Ceci découle du fait que nous n'avons pas permis aux adjectifs de s'insérer à l'intérieur de cette structure. Les couples qui acceptent des variations représentent 13 % de l'ensemble de couples pour le corpus MTS et 11 % pour le corpus LBC, toujours sur les couples ayant au moins deux occurrences. Voici quelques exemples de couples acceptant et n'acceptant pas de variations :

1. **Couples n'acceptant aucune variation** ($Var = 0$)
   Ces couples sont à 99 % de structure N ADJ comme *satellite artificiel, rayonnement solaire, cadre supérieur.* Les autres exemples de structures de longueur fixe sont :

   (a) N *non* ADJ
   - *amplificateur(s) non linéaire(s)*
   - *numéro non valable*

- *rayonnement(s) non essentiel(s)*

(b) N *adv* ADJ

- *filtre passe-bas idéal*
- *onde avant absente, onde arrière absente*

(c) N ADJ CONJ ADJ

- *idéogrammes chinois et japonais*
- *ondes métriques et décimètriques*
- *organisme(s) scientifique(s) ou industriel(s)*

Certains couples comme (**codage**, **décimal**) sont uniquement déduits à partir de constructions attributives : *codage est décimal.*

2. **Couples acceptant des variations** ($Var \neq 0$)
Les couples dont l'adjectif apparaît aussi bien en position épithète qu'en position attributive représentent 20 % de l'ensemble des couples acceptant des variations en distance pour le corpus MTS et 30 % pour le corpus LBC ; la variation la plus fréquente reste l'insertion d'un adverbe.

- (**circuit**, **numérique**)
circuit numérique, circuits numériques
circuits entièrement numériques
circuits analogiques et numériques
circuits sont numériques
- (**effet**, **local**)
effet local
effet purement local
effet uniquement local
- (**ligne**, **téléphonique**)
ligne téléphonique, lignes téléphoniques
ligne (téléphonique), lignes (téléphoniques)

Pour la liste de couples donnée en annexe, nous n'avons pas indiqué les valeurs prises par les mesures de distances. Par contre, pour plus de lisibilité, nous avons précisé, après certains couples, l'expression rationnelle correspondant aux différentes séquences de texte rencontrées.

4.5.5 Conclusion

En conclusion de l'examen des modèles statistiques qui ont été sélectionnés par l'évaluation graphique, nous retenons que la fréquence d'un couple est un très bon indicateur de son caractère terminologique. Le problème de la fréquence est qu'elle ne permet pas d'isoler les noms composés rares et que le classement qu'elle

propose intègre très rapidement du bruit, même si celui-ci peut être diminué en utilisant les résultats de la diversité de Shannon normalisée. Entre la fréquences et les trois critères d'association, nous avons choisi de ne retenir que le coefficient de vraisemblance, ou plus exactement le test du rapport de vraisemblance appliqué à une loi binomiale, pour les qualités déjà invoquées précédemment à savoir : sa nature de test statistique, sa tendance générale à prendre en compte la fréquence du couple, son bon comportement quelle que soit la taille du corpus, et la valeur indéfinie qu'il assigne aux couples possédant une diversité nulle sur l'un des lemmes du couple. Néanmoins, malgré toutes les qualités de ce coefficient, le classement qu'il propose comporte un certain bruit, dont il n'est pas responsable, que nous analysons ainsi :

1. quelques couples sans statut doivent leur sélection à des erreurs d'étiquetage grammatical :
   - (**pas, traduction**) (*pas de traduction*) où l'adverbe de négation *pas* est étiqueté comme nom.
2. quelques couples dont l'un des lemmes est un nom composé de longueur 1 construit avec un trait d'union apparaissent en double dans les listes ; ceci à cause du programme d'assignation d'étiquettes morphologiques qui ne lemmatise pas correctement les noms composés avec trait d'union (voir section 3.1.4). Voici un exemple de doublon :
   - (**sou- système, utilisateur**), (**sou-systèmes, utilisateur**)
3. certains couples ne sont pas des noms composés mais :
   - des adverbes ou des sous-séquences d'un groupe prépositionnel de type adverbial :
     - (**bout, bout**) (*bout en bout*)
     - (**titre, exemple**) (*à titre d'exemple*),
     - (**plupart, cas**) (*dans la plupart des cas*)
     - (**heure, actuelle**) (*à l'heure actuelle*)
   - des unités de mesures :
     - (**ko, bit**) (*ko bits*)
4. un nombre déjà plus important de couples correspondent à une sous-séquence d'un nom composé de longueur ≥ 3 . Ce bruit découle directement du problème de la surcomposition et de la modification. Quelques exemples :
   - (**rapport**, *porteur*) par exemple sous-séquence du nom composé *rapport porteuse/bruit*,
   - (**type, limiteur**) sous-séquence de *linéariseur de type à limiteur*,

- (augmentation, espacement) sous-séquence de *augmentation de l'espacement angulaire*
- (accès, étalement) sous-séquence de *accès multiple par étalement du spectre*
- (service, fixe) sous-séquence de *service fixe par satellite*,
- (circuit, fictif) sous-séquence de *circuit fictif de référence*,
- (bande latérale) sous-séquence de *bande latérale unique.*

À l'inverse, le fait d'avoir permis l'insertion de modifieurs dans la structure des noms composés binaires, implique la reconnaissance de certains noms composés ternaires, i.e de longueur 3. Par exemple, le couple (service, satellite) correspond au nom composé *service fixe par satellite* et le couple (accès, répartition) au nom composé *accès multiple par répartition.* Nous devons aussi préciser que certains noms composés n'appartiennent pas au domaine spécifique des télécommunications mais plutôt au langage courant comme (feuille, papier) (*feuille(s) de papier*), ou encore (mise, page) (*mise en page*). Ces résultats ne sont donc pas parfaits, mais le bruit y est quand même minime par rapport aux listes obtenues uniquement par l'application de modèles statistiques sans filtrage linguistique. Le choix de ne retenir que le coefficient de vraisemblance entraîne la non sélection d'un certain nombre de couples qui sont pourtant des noms composés. Cependant, quelle que soit la mesure choisie, nous aurions été confrontée au même problème, i.e. l'élimination de certains couples qui correspondent à des noms composés du domaine, sans obtenir pour autant un classement conceptuel aussi performant.
D'autre part, l'objectif de ce travail d'extraction était de relier le nom composé à ses principales variantes. Nous avons obtenu les variantes suivantes :

- **variantes orthographique :**
  sous l'entrée d'un couple se trouve : les flexions et les graphies du nom composé rencontrées dans le corpus. Pour le trait d'union facultatif dans la structure N_1 N_2, nous avons considére que les noms composés qui admettaient les deux graphies étaient des variantes du nom composé de longueur 1, i.e. construit avec un trait d'union. Nous avons donc vérifié quels noms composés admettaient les deux graphies : seuls deux noms composés dans le corpus MTS *écart-type/ écart type* et *schéma-type/schéma type* et trois noms composés dans le corpus LBC *écart-type/ écart type*, *mémoire-tampon/mémoire tampon* et *mode-paquet/mode paquet* acceptent un trait d'union facultatif. Ce résultat est très intéressant : le trait d'union dans le domaine des télécommunications possède un caractère figé qu'il n'a sans doute pas dans d'autres domaines techniques.

- **variantes morphosyntaxiques :**
  nous avons sous l'entrée d'un couples N_1 (PREP (DET)) N_2, les variantes qui

introduisent une simplification ou une complication de la structure, et les changements éventuels de préposition. Par contre, nous n'avons pas identifié les relations de synonymie ; celles-ci devront pouvoir être reconnues lors de l'extraction de terminologie bilingue.

- **variantes elliptiques :**
  l'identification des variantes elliptiques n'était absolument pas prévue. Néanmoins, le fait d'avoir extrait certains noms composés ternaires de structure N_1 ADJ PREP N_2 a permis de lister certaines variantes comme, par exemple, la variante elliptique *service par satellite* du nom composé ternaire *service fixe par satellite.*

- **abréviations :**
  les principales abréviations de nos corpus ont été relevées manuellement. Elles devront être ajoutées aux autres variantes.

4.6 Extension de la méthode à d'autres ressources lexicales

Cette méthode ayant donné des résultats concluants dans le cadre de l'extraction de terminologie monolingue, il nous a paru intéressant de l'étendre à l'extraction d'autres ressources lexicales, en particulier, l'extraction de terminologie bilingue et l'extraction de structures argumentales.

4.6.1 Extraction de terminologie bilingue

L'étude linguistique monolingue nous a permis de définir les spécifications linguistiques des noms composés pour le français et pour l'anglais en termes de structures morphosyntaxiques. Nous avons ensuite utilisé ces spécifications linguistiques pour extraire des corpus les séquences morphosyntaxiques qui caractérisent les noms composés binaires : ce sont nos candidats. Nous possédons donc, des listes de candidats français et anglais. Il ne s'agit plus ici d'appliquer un modèle statistique sur les candidats monolingues pour déterminer lesquels sont les bons (resp. les mauvais), mais de trier les candidats monolingues en s'appuyant sur un critère bilingue : dans un même domaine technique, un nom composé n'a qu'une et une seule traduction possible dans une autre langue. Ce critère de traduction reprend en quelque sorte le critère de référé unique : un nom composé terminologique correspond à un concept unique et universel.

Chaque candidat se présente sous forme de couple composé des deux lemmes des deux unités lexicales principales d'une structure morphosyntaxique. Les patrons utilisés dans cette section ne sont plus les patrons généraux utiles pour obtenir un classement conceptuel des couples monolingues ; nous utilisons des

patrons plus précis correspondant à chaque type élémentaire, c'est-à-dire pour le français: N_1 *de* (DET) N_2, (N_1 *à* (DET) N_2 + N_1 PREP N_2), N_1 N_2 et N ADJ, et pour l'anglais: N_2 N_1 et ADJ N. Le problème est maintenant de mettre en correspondance un candidat français et un candidat anglais en utilisant le corpus aligné phrases à phrases. Il devrait être aisé d'aligner les noms composés français et leur traduction simplement en utilisant le critère de traduction unique. Avant de présenter la méthode utilisée pour apparier les candidats à l'intérieur de phrases alignées, nous présentons quelques données linguistiques bilingues. Cette courte étude examine les correspondances de structures syntaxiques entre le français et l'anglais.

4.6.1.1 Données linguistiques bilingues

Nous avons examiné les noms composés français et leur traduction en anglais et nous avons établi des correspondances de structures syntaxiques. Certaines correspondances sont régulièrement rencontrées ; d'autres sont déjà plus rares. Nous avons aussi relevé certaines correspondances problématiques qui correspondent à l'emploi d'une variante elliptique ou d'une abréviation.

Régularités dans l'alignement des patrons

Du point de vue de la traduction, les noms composés terminologiques français de structure N_1 *de* (DET) N_2 se traduisent régulièrement par des noms composés anglais de structure N_2 N_1. Quelques exemples :

1. Fr. N_1 *de* N_2 ⇔ Eng. N_2 N_1

distance de séparation	*separation distance*
modulation de fréquence	*frequency modulation*
fonction d'exploitation	*operation function*

2. Fr. N_1 *de* DET N_2 ⇔ Eng. N_2 N_1

concentration des conversations	*speech interpolation*
affectation des voies	*channel assignment*
effet de la propagation	*propagation effect*
étalement du spectre	*spectrum spreading*
intégrité du réseau	*network integrity*

Deux remarques :

- la flexion en nombre portée par le N_1 français est répercutée sur le N_1 anglais,
- la flexion en nombre et la détermination du N_2 français est perdue au cours du processus de traduction ; dans la majorité des cas, le N_2 anglais est au singulier.

Nous avons néanmoins rencontré quelques exceptions à ces régles avec les exemples suivants :

règlement des radiocommunications	*radio regulations*
équipements de télécommunications	*communication(s) equipment*

Dans le premier exemple, le N_1 français est singulier alors que le N_1 anglais est pluriel ; dans le second exemple, le N_2 français est toujours rencontré au pluriel au contraire du N_2 anglais qui varie librement en nombre.

De la même manière, les noms composés terminologiques français de structure N ADJ se traduisent régulièrement par des noms composés anglais de structure ADJ N. En voici, quelques exemples :

- Fr. N ADJ ⇔ Eng. ADJ N

Service national	*Domestic Service*
cornet pyramidal	*pyramidal horn*
canal adjacent	*adjacent channel*
dilatation thermique	*thermal expansion*
fréquence intermédiaire	*intermediate frequency*
cellules solaires	*solar cells*

Cette correspondance structurelle n'implique pas toujours une correspondance mot à mot entre les unités lexicales pleines du nom composé. Si nous examinons le couple : *concentration des conversations/speech interpolation*, nous remarquons qu'il n'y a pas de réelle correspondance de sens entre les différents éléments de ces noms composés : le nom anglais *interpolation* se traduit normalement par le nom français *interpolation* et non pas par *concentration.*
Les autres régularités, mais déjà moins fréquentes, dans l'alignement des structures morphosyntaxiques sont :

1. Fr. N_1 PREP N_2 ⇔ Eng. N_2 N_1

codage par transition	transition coding
couverture à l'émission	emission coverage
perte par étalement	spread loss

2. Fr. N_1 N_2 ⇔ Eng. N_2 N_1

équipement radio	radio equipment
antenne multifaisceau	multibeam antenna
circuit hyperfréquence	microwave circuit

Irrégularités dans l'alignement des patrons

Les irrégularités dans l'alignement des patrons sont principalement de deux sortes : celles qui proviennent d'une altération de la structure du nom composé

dans une langue et pas dans l'autre, et celles qui mettent en jeu des alignements non réguliers de patrons.

Altération de la structure de l'un des noms composés

Les altérations de structures qui posent problème sont celles qui mettent en jeu les unités lexicales pleines du nom composé, c'est-à-dire les abréviations et les variantes elliptiques.

1. **Abréviation**

   Un nom composé terminologique admet parfois une abréviation comme variante (les abréviations des noms composés ont été présentées dans la section 2.2.3.1). Quand un nom composé et sa traduction admettent dans les deux langues une abréviation, son utilisation dans une langue n'implique absolument pas son utilisation dans l'autre langue. Étant donné une occurrence de l'abréviation *IF* (*intermediate frequency*) en anglais, rien ne nous permet d'être sûr que la forme française correspondante sera *FI* (*fréquence intermédiaire*). Au contraire, nous avons remarqué que l'anglais utilisait presque systématiquement les abréviations introduites, au contraire du français qui préfère la forme entière. Donc, si dans le texte français, on rencontre *fréquence intermédiaire* et dans le texte anglais *FI*, l'alignement des patrons n'est plus de type: Fr. N ADJ $\Leftrightarrow$ Eng. ADJ N, mais de type: Fr. N ADJ $\Rightarrow$ Eng. N.

2. **Ellipse**

   Les ellipses introduisent le même type d'ambiguïté que les abréviations. Il est possible que, dans une langue, une partie du nom composé n'apparaisse pas dans l'autre langue; ainsi l'anglais semble éliminer le nom *equipment* dans le terme *thermal control* qui se traduit par *régulation thermique des équipements*, et qui sous une forme entière aurait dû être *thermal equipment control*.

Changement de patrons

Nous avons vu précédemment dans la section 4.6.1.1 qu'il existe des régularités dans l'alignement des patrons. Il existe, bien entendu, aussi des irrégularités et d'autres alignements sont possibles:

1. Fr. N_1 *de* N_2 $\Leftrightarrow$ Eng. ADJ N

émetteur de Terre	*terrestrial transmitter*
élément de recharge	*spare part*

2. Fr. N_1 PREP (DET) N_2 $\Leftrightarrow$ Eng. ADJ N

voie au repos	*idle channel*
impression à distance	*remote printing*

3. Fr. N ADJ $\Leftrightarrow$ Eng. N_2 N_1

alimentation électrique	*power supply*
créneau temporel	*time slot*

Dans ces exemples d'alignement de patrons, la remarque que nous avons faite précédemment est encore valable: la correspondance de sens entre les unités lexicales pleines n'est absolument pas obligatoire. Il faut même parfois se méfier des erreurs que pourrait engendrer une correspondance un-un, i.e. une traduction mot à mot. Par exemple, le terme français *émetteur de Terre* ne se traduit pas par *Earth transmitter* mais par *terrestrial transmitter*; de plus, *Earth transmitter* qui correspond à une traduction mot à mot de *émetteur de Terre* réfère à un terme différent. Un autre exemple de traduction non-compositionnelle est l'alignement *créneau temporel/time slot.*
D'autres exemples d'alignement de patrons mais plutôt rares:

1. Fr. N_1 PREP N_2 $\Leftrightarrow$ Eng. N
   mise à jour *updating*

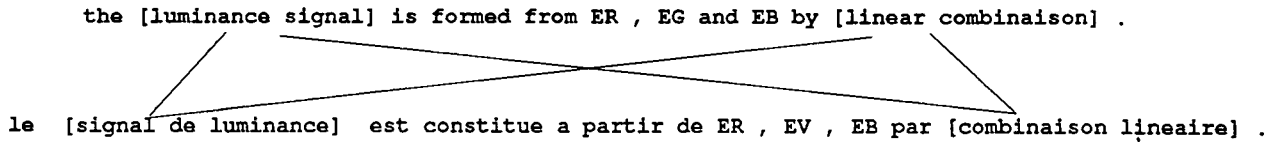
2. Fr. N_1 PARTICIPE-PRÉSENT $\Leftrightarrow$ Eng. N
   élément rayonnant *radiator*

3. Fr. N $\Leftrightarrow$ Eng. N_2 N_1
   télécopie *facsimile transmission*

Les irrégularités dans l'alignement des patrons posent problème lorsque l'alignement ne met plus en jeu deux composés binaires (alignement 2-2). Lorsqu'un nom composé anglais de structure N_2 N_1 n'est pas mis en correspondance avec un nom composé français de structure N_1 *de* (DET) N_2 mais, avec un autre nom composé binaire, l'alignement reste de type 2-2, et comme, nous avons normalement relevé tous les candidats binaires de chaque langue, il est relativement aisé de retrouver l'alignement correct. Par contre, lorsqu'un nom composé binaire anglais (resp. français) est traduit par un nom simple ou une variante de longueur 1, l'alignement n'est plus de type 2-2 mais de type 2-1 (resp. 1-2). Cet alignement ne sera pas identifié car nous n'avons pas inclu dans la liste des candidats les noms simples ou les abréviations. Si ces dernières peuvent néanmoins être rajoutées assez facilement, l'intégration de tous les noms simples pouvant être les variantes elliptiques potentielles des noms composés risque d'introduire beaucoup de bruit dans les alignements bilingues.

4.6.1.2 Méthode d'alignement de termes

Nous rappelons qu'à la suite du programme d'extraction de structures morphosyntaxiques caractérisant les noms composés binaires en français et en anglais, on associe à chaque candidat la liste de ses positions dans le corpus ; une position correspond à une occurrence du couple. Retrouver pour chaque alignement de phrases, les candidats monolingues y apparaissant est immédiat. Il s'agit maintenant d'apparier les candidats à l'intérieur de phrases alignées. Chaque candidat français apparaissant à l'intérieur d'une phrase est aligné avec tous les candidats anglais de la phrase appariée. Ainsi, pour les phrases appariées ci-dessous :



nous relevons pour les candidats français :

- (signal, luminance), une occurrence du couple (luminance, signal) et une occurrence du couple (linear, combinaison),
- (combinaison, linéaire), les mêmes occurrences, c'est-à-dire une occurrence de (luminance, signal) et une occurrence de (linear, combinaison)

Dans un premier temps, aucun filtrage n'est effectué, ni en termes d'alignement de structure morphosyntaxique, ni en termes de position dans la phrase. On relève les occurrences des alignements candidat à candidat (2-2) pour chaque alignement de phrases du corpus MTS. À la fin de cette opération, nous obtenons pour chaque candidat français, une liste de candidats anglais ; à chaque candidat anglais est associé un score qui correspond au nombre d'occurrences d'alignements dans le corpus. Les candidats anglais sont ensuite classés suivant la valeur décroissante de cette "fréquence d'alignement". Puis, nous appliquons le critère décrit dans [Gaussier *et al.*, 1992] (présenté dans la section 1.5) qui consiste à exclure d'une liste d'un candidat français, les candidats anglais qui possèdent une fréquence d'alignement plus élevée dans une autre liste. Ce critère permet de réduire le nombre d'alignements terme à terme de 37 000 à 1 200. Cette méthode, qui ne prend en compte que la fréquence d'alignement, donne déjà des résultats satisfaisants : pour la moitié des candidats français, le candidat anglais associé à la fréquence d'alignement la plus élevée est sa traduction. Ces résultats peuvent, bien entendu, être améliorés de plusieurs manières : soit en intégrant des données linguistiques bilingues, soit en utilisant un modèle statistique un peu plus complexe que la fréquence, soit en combinant modèle statistique et données linguistiques bilingues. Ces trois expériences seront largement décrites dans [Gaussier, 1994]. Nous nous contenterons de présenter rapidement la première méthode prenant en compte les informations linguistiques bilingues présentées au début de cette section.

Nous avons vu qu'un candidat anglais de structure N_2 N_1 se traduit le plus souvent en français par un candidat de structure N_1 *de* (DET) N_2 et beaucoup plus rarement par des candidats de structures N_1 PREP N_2, N ADJ, ou encore N_1 N_2. Nous n'examinons pas pour l'instant les correspondances plus complexes, de type 2-1 ou 1-2, comme par exemple celle qui associe un candidat anglais de structure N_2 N_1 à un nom simple ou à son abréviation française. Ces correspondances structurales sont exprimées en termes de probabilités d'alignement de structures et sont estimées à partir d'une portion du corpus aligné. Les alignements de candidats ne sont plus maintenant équiprobables mais prennent en compte les probabilités d'alignement de structures. Par exemple, pour la phrase alignée précédente, le candidat français suivant associé à sa structure syntaxique, ((signal, luminance), N_1 *de* (DET) N_2), ayant plus de chance d'être traduit par le candidat anglais de structure N_2 N_1 que par celui de structure ADJ N , on assigne au couple ((signal, luminance), N_1 *de* (DET) N_2) un score de 0,7 et au couple ((linear, combinaison), ADJ N) un score de 0,3. En pondérant les fréquences d'alignement par la probabilité d'alignement de patrons, nous améliorons les résultats précédents : nous passons de 50 % à 60 % d'alignements corrects. Ces alignements sont évalués par une méthode identique à celle présentée en section 4.4, c'est-à-dire en utilisant une liste de référence de termes bilingues. Voici quelques exemples d'alignements bilingues obtenus :

(earth, station)	(station, terrien)
(multiple, acces)	(accès, multiple)
(noise, temperature)	(température, bruit)
(error, ratio)	(taux, erreur)
(frequency, band)	(bande, fréquence)
(bit, rate)	(débit, binaire)
(satellite, transponder)	(répéteur, satellite)
(intermodulation, product)	(produit, intermodulation)
(space, segment)	(secteur, spatial)
(power, amplifier)	(amplificateur, puissance)
(radio, regulation)	(règlement, radiocommunication)
(satellite, system)	(système, satellite)
(data, transmission)	(transmission, donnée)
(link, budget)	(bilan, liaison)
(channel, unit)	(unité, voie)
(traffic, capacity)	(capacité, trafic)
(reference, station)	(station, référence)
(energy, dispersal)	(dispersion, énergie)
(signalling, system)	(système, signalisation)
(antenna, gain)	(gain, antenne)

4.6.2 Extraction de structures argumentales

Nous avons expérimenté deux méthodes pour l'extraction des structures argumentales des verbes du français. La première s'appuie sur l'analyseur probabiliste développé pour l'anglais par l'Université de Lancaster, la deuxième sur le programme d'extraction de séquences morphosyntaxiques présenté dans la section 4.2.

4.6.2.1 Analyseur markovien

L'objet de cette étude étant d'automatiser l'extraction des structures argumentales des verbes, nous n'avons pris en compte que les verbes les plus fréquents de notre corpus. Nous utilisons un analyseur markovien qui est chargé d'identifier dans la phrase les groupes nominaux. Le programme plus complet qui analyse la phrase en constituants immédiats a été présenté dans la section 1.2. Rappelons brièvement le principe : sur le corpus étiqueté, le programme insère des crochets autour des groupes nominaux. Il utilise pour cela les étiquettes grammaticales assignées aux items du corpus, et des régles de grammaire probabilistes. Ces règles décrivent toutes les structures possibles d'un groupe nominal rencontrées dans un corpus d'apprentissage analysé. À chaque règle est associée une probabilité qui rend compte de la fréquence de la structure dans le corpus d'apprentissage. L'expérience se déroule en deux étapes :

1. Identification des groupes nominaux en utilisant l'analyseur markovien,
2. Extraction de certaines structures argumentales des verbes les plus fréquents : les compléments d'objet direct et indirect.

Nous avons introduit plusieurs classes de complément d'objet indirect : ceux introduit par la préposition *de*, ceux introduit par la préposition *à* et ceux introduit par une autre préposition. Les compléments d'objet direct et indirect sont exprimés en termes de patrons, et leurs occurrences sont relevées dans le corpus. Seuls sont examinés les patrons les plus fréquents pour un verbe donné. Nous avons ainsi extrait les structures argumentales d'une vingtaine de verbes. Les résultats sont loin d'être satisfaisants. En effet, les structures argumentales indiquant un objet direct ou un objet prépositionnel introduit par la préposition *de* restent imprécises quand elles ne sont pas tout simplement incorrectes. Les raisons découlent des mauvais comportements des deux programmes stochastiques :

- le programme d'assignation d'étiquettes grammaticales
  La particule partitive *de* constitutive de l'article partitif : *du*, *de la* et l'article indéfini pluriel *des* lorsqu'il introduit un objet direct sont systématiquement étiquetés comme préposition. Les étiquettes grammaticales ne permettent donc pas de distinguer les compléments d'objet direct introduit par les articles partitifs *du, de la* ou par l'article indéfini pluriel *des* des compléments d'objet indirect introduit par la préposition *de.*

- l'analyseur markovien
  L'analyseur markovien ne reconnaît pas les groupes nominaux maximaux et en particulier, les noms composés. Nous avons vu précédemment combien était nombreux les noms composés de structure N_1 *de* (DET) N_2 dans le corpus. Chaque fois que l'analyseur rencontre une séquence N_1 *de* (DET) N_2, il l'analyse comme "groupe nominal + *de* + groupe nominal". De nombreux verbes se voient ainsi attribués un complément prépositionnel introduit par la préposition *de* qui n'existe pas.

Nous ne chercherons donc pas à identifier les compléments prépositionnels introduit par *de*. L'analyseur ne précisant pas si le verbe de la phrase est employé à l'actif ou au passif, nous ne retiendrons pas non plus les compléments prépositionnels introduit par *par* qui correspondent, pour leur majorité, à des sujets profonds présents à la voix passive. Concernant les autres patrons, rien ne nous indique qu'ils réfèrent plus à un complément prépositionnel qu'à un complément circonstanciel (excepté certains compléments prépositionnels introduits par *à* qui peuvent être identifiés en utilisant les pronoms datifs (voir section suivante)). Pour prendre une décision sur le statut du groupe prépositionnel extrait, il faut nécessairement examiner le texte. Cette méthode ne permet donc pas d'automatiser l'extraction de structures argumentales : nous allons illustrer ce résultat par l'examen des patrons extraits de quelques verbes :

- **assurer** (295 occurrences)
  - *assurer* + (NP) + *à* NP (22) (to provide)
    Le complément prépositionnel est confirmé par la rencontre d'un pronom datif. Quelques contextes :
    les répartiteurs assurent les connexions ...
    le couplage ... assure aux opérateurs une bonne souplesse d'utilisation
- **donner** (401 occurrences)
  - *donner* + (NP) + *à* NP (39) (to assign)
    Le complément prépositionnel est confirmé par la rencontre d'un pronom datif.
    ... donner une valeur au TEB ....
  - *donner* + *dans* NP (21)
    Cette entrée n'existe pas : le groupe prépositionnel *dans* NP correspond toujours à un complément circonstanciel comme dans l'exemple : *cette indication est donnée dans le paragraphe ....*
  - *donner* + (NP) + *pour* NP (16)
    Cette entrée n'existe pas : *pour* NP correspond toujours à un modifieur du verbe *donner* ou à une expression verbale figée comme : *donner pour exemple.*

- **obtenir** (195 occurrences)

  - *obtenir* + (NP) + *à* NP (36)
    Le complément d'objet direct est confirmé par la rencontre dans le corpus d'un pronom objet direct. Par contre, aucun pronom datif n'a été rencontré. La complémentation entière d'ailleurs n'existe pas: la séquence *à* NP correspond toujours à un complément circonstanciel comme dans: ... *obtenir à l'aide du système de codage MIC* ...
  - *obtenir* + *en* NP (20)
    Cette entrée n'existe pas et la séquence *en* NP correspond toujours à un modifieur comme dans: *On l'obtient en ajoutant* ...
  - *obtenir* + (NP) + *avec* NP (15)
    Cette entrée n'existe pas et la séquence *avec* NP correspond toujours à un modifieur comme dans: *les capacités que l'on peut obtenir avec différents systèmes*

  La seule structure argumentale du verbe *obtenir* rencontrée dans le corpus est la suivante: *obtenir* + NP (to obtain)
  ... *pour obtenir une grande fiabilité de système* ...

- **utiliser** (839 occurrences)

  - *utiliser* + (NP) + *pour* NP (143) (to use/utilize for)
    Ces moyens sont utilisés pour assurer, depuis le sol, les opérations de soutien logistique des satellites
  - *utiliser* + (NP) + *dans* NP (69) (to use in)
    les techniques les plus couramment utilisées dans les télécommunications par satellite
  - *utiliser* + (NP) + *à* NP (41)
    Seul le complément d'objet direct fait effectivement partie de la structure argumentale du verbe *utiliser*; cet objet direct est d'ailleurs confirmé par la rencontre d'un pronom objet direct. Aucun pronom datif n'a été rencontré.
    L'entrée complète n'existe pas: la séquence *à* NP correspond toujours à un modifieur comme dans: *L'algorithme utilisé à cet effet* ...
  - *utiliser* + (NP) + *en* NP (37) (to use in)
    les codages correcteurs sont utilisés en transmission numérique

L'examen des ces quelques structures argumentales des verbes les plus fréquents du corpus MTS montre clairement que les structures argumentales extraites ne peuvent pas être directement intégrées dans un dictionnaire. La fréquence ici n'est pas vraiment pertinente et chaque occurrence relevée demande d'être vérifiée par un examen du texte. Plus de la moitié des structures extraites ne

sont pas caractéristiques de la structure argumentale du verbe mais réfèrent à des modifieurs. Les résultats obtenus pourraient être améliorés si les groupes nominaux étaient correctement identifiés et si certains déterminants n'étaient pas analysés comme des prépositions. Cependant, même avec ces améliorations, le problème de l'attachement prépositionnel ne serait pas pour autant réglé et chaque structure extraite devra toujours être manuellement vérifiée.

4.6.2.2 Programme d'extraction de patrons

Nous avons utilisé le programme réalisé pour extraire et compter les séquences morphosyntaxiques qui caractérisent les noms composés pour extraire, d'abord les verbes employés avec des pronoms pré-verbaux, puis pour tenter d'isoler des constructions à verbes supports ou des expressions verbales figées les plus simples.

Pronoms pré-verbaux

Seuls les pronoms pré-verbaux objet direct (*le*, *la*, *les*, *l'*) et les pronoms pré-verbaux datifs (*lui*, *leur*) sont non-ambigus et peuvent être utiliser pour déduire la complémentation ou une partie de la complémentation des verbes. Les pronoms pré-verbaux *y* et *en* sont inutilisables puisqu'ils réfèrent aussi bien à un argument du verbe qu'à d'autres types de groupes prépositionnels. Les résultats sont les suivants :

1. **Verbe + Objet direct** (93 verbes)
   acheminer, *adapter*, *affecter*, *aider*, *aiguiller*, *ajouter*, *ajuster*, *allonger*, *amener*, *amplifier*, *analyser*, *appeler*, *appliquer*, *apprécier*, *assister*, *associer*, *attacher*, *circulariser*, *classer*, *comparer*, *composer*, *concentrer*, *concevoir*, *connecter*, *conserver*, *constater*, *constituer*, *consulter*, *convertir*, *décomprimer*, *définir*, *démoduler*, *démultiplexer*, *désirer*, *désservir*, *dire*, *diriger*, *distinguer*, *distribuer*, *diviser*, *écrire*, *éloigner*, *employer*, *envoyer*, *éviter*, *examiner*, *expliquer*, *exploiter*, *exprimer*, *faire*, *falloir*, *grouper*, *indiquer*, *informer*, *installer*, *interconnecter*, *interpréter*, *juger*, *limiter*, *lire*, *mettre*, *montrer*, *multiplexer*, *noter*, *obtenir*, *ouvrir*, *permettre*, *précéder*, *présenter*, *prévoir*, *protéger*, *ramener*, *réaliser*, *recombiner*, *reconnaître*, *reconstituer*, *recueillir*, *réduire*, *réémettre*, *relier*, *remplacer*, *rendre*, *répartir*, *restituer*, *retransmettre*, *supprimer*, *synchroniser*, *transmettre*, *transporter*, *transposer*, *utiliser*, *voir*

2. **Verbe + *à* NP** et **Verbe + NP + *à* NP** (24 verbes)
   Les verbes employés à la voix passive avec un un pronom datif correspondent aux verbes qui acceptent un complément d'objet direct et un complément prépositionnel introduit par *à* : ces verbes sont marqués dans le liste ci-dessous par un signe plus (+) :
   affecter (+), *ajouter* (+), *appliquer*, *assigner* (+), *associer* (+), *assurer*,

attacher, *attribuer* (+), *conserver*, *correspondre*, *destiner*, *donner*, *faire*, *imposer*, *indiquer*, *intégrer* (+), *lier* (+), *limiter*, *parvenir*, *permettre*, *relier* (+), *représenter*, *rester*, *signaler* (+)

Nous avons vérifié ces résultats obtenus automatiquement. Sur les 94 verbes rencontrés avec un pronom objet direct, un seul, le verbe *aller* n'accepte pas d'objet direct. Cette mauvaise analyse est introduite par une erreur du programme d'assignation étiquettes grammaticales : la séquence *l'aller* a été analysée comme pronom pré-verbal + verbe au lieu de recevoir l'analyse correcte : article +nom. Ce type d'erreurs ne devrait pas être trop important en français, le nombre de substantifs et de verbes partageant une même graphie n'étant pas trop nombreux. Par contre, il serait intéressant de vérifier les résultats obtenus en anglais. L'utilisation des pronoms pour déduire une partie de la structure argumentale du verbe donne de bons résultats quoique ceux-ci soient limités. Néanmoins, ils ont l'avantage d'être fiables.

Constructions à verbes supports et constructions verbales figées

Les constructions à verbes supports et les constructions verbales figées peuvent être considérées comme des co-occurrences de type (Verbe, Nom). Il n'est évidement pas question de chercher à extraire toutes les co-occurrences d'un tel couple : il peut avoir des discontinuités entre le verbe et le nom et leur ordre respectif peut être variable lorsque, par exemple, ces constructions se rencontrent au passif, dans une relative, etc. Pour identifier toutes les co-occurrences (Verbe, Nom), il est nécessaire d'analyser syntaxiquement la phrase. Nous avons relevé ces co-occurrences qu'à l'actif et nous n'avons permis qu'a un petit nombre d'éléments de s'insérer entre le verbe et le nom (déterminant, préposition, adverbe, adjectif). Nous espérons, néanmoins, recueillir quelques couples dont ceux présentant un réel figement entre le verbe et le nom.
Nous avons écrits deux automates : l'un pour extraire les séquences V N, l'autre pour les séquences Vsup N où Vsup est l'un des verbes supports suivants : *avoir, donner, faire, mettre, tenir, prendre, maintenir.* Les modèles statistiques que nous avons appliqué sur les couples (V, N) ou (Vsup, N) sont la fréquence et le score d'association. Nous avons ensuite examiné les valeurs prises par ces deux mesures et essayé de déterminer si elles permettaient de détecter les expressions verbales figées ou les constructions à verbes supports. La fréquence donne de bons résultats pour les couples (Vsup, N) mais pas pour les couples (V, N). Les couples (V, N) les plus fréquents ne correspondent ni à des expressions verbales figées, ni à des constructions à verbes supports comme l'atteste le couple (**voir**, **figure**) (71 occurrences) (*voir (la + sur la + 0 + à la) figure*). À l'inverse, le score d'association n'apporte aucune information sur les couples (Vsup, N) mais permet d'isoler parmi l'ensemble des couples (V, N) des constructions verbales figées qui sont peu fréquentes. Les résultats obtenus pour les couples (V, N) et (Vsup, N)

sont résumés dans le tableau ci-dessous :

MTS	1 occurrence	2 occurrences	Plus de 2 occurrences	Total
V (PREP) N	4 492	621	412	5 525
Vsup (PREP) N	195	46	60	301
donner (PREP) N	47	19	15	81
avoir (PREP) N	68	14	25	107
faire (PREP) N	22	1	7	30
mettre (PREP) N	22	2	10	34
tenir (PREP) N	3	1	1	5
prendre (PREP) N	11	6	2	19
maintenir (PREP) N	22	3	0	25

Les couples (Vsup, N) les plus fréquents sont les suivants (le nombre d'occurrences du couple est indiqué entre parenthèses) :

- *mettre en œuvre* (81)
- *tenir compte* (78)
- *faire l'objet* (34)
- *faire appel* (29)
- *prendre en considération* (24)
- *avoir titre* (19)
- *donner (des + un + quelques + pour + plusieurs + l') exemple(s)* (18)
- *avoir besoin* (16)
- *mettre en œuvre* (15)
- *avoir lieu* (15)
- *mettre en place* (14)
- *avoir accès* (13)
- *mettre au point* (12)
- *prendre en compte* (11)
- *avoir recours* (11)

Les couples (V, N) qui reçoivent une valeur élevée du score d'association sont (le nombre d'occurrences du couple est indiqué entre parenthèses) :

- *attendre (d' + un) accusé* (2)
- *dégager (la + quelques grandes) tendance(s)* (2)
- *revenir en arrière* (2)
- *lever l'ambiguïté* (3)
- *annexer à la circulaire* (6)
- *distinguer la parole* (2)
- *afficher (sur + sur un) écran* (2)
- *dissiper (la + une + de la) chaleur* (2)
- *jouer un rôle* (7)

Nous extrayons ainsi quelques constructions verbales figées et quelques constructions à verbes supports présentes dans le corpus. Ces constructions ne sont évidement qu'un sous-ensemble des constructions verbales du corpus. Néanmoins, la fréquence lorsque la structure syntaxique est précise donne de meilleurs résultats que le score d'information.

Conclusion

Face à la pléthore de travaux fondés sur l'exploitation des modèles statistiques pour le traitement automatique du langage naturel, il nous a semblé important d'évaluer le degré réel d'efficacité de ces méthodes. Une première conclusion s'impose : les statistiques donnent de bons résultats dans certaines applications comme l'étiquetage grammatical ou l'alignement phrases à phrases, mais des limites apparaissent vite dès qu'on aborde l'analyse syntaxique et *a fortiori* la traduction automatique. Cet état de l'art des systèmes existants nous a aussi permis de penser que les statistiques trouveraient un bon domaine d'application dans les ressources lexicales et plus particulièrement dans l'extraction de terminologie. Une deuxième conclusion a suivi : l'introduction de données linguistiques améliore sensiblement la qualité des systèmes stochastiques. C'est la raison pour laquelle nous avons décidé d'exploiter ce que la linguistique pouvait nous apporter comme connaissance sur les noms composés. L'examen de quelques-uns des travaux linguistiques effectués dans ce domaine a montré l'utilité certaine des structures morphosyntaxiques mises à jour par les linguistes, mais laisse aussi un certain nombre de problèmes en suspens dont le moindre n'est pas la formation de nouveaux noms composés à partir de noms composés déjà existants, principalement par surcomposition ou modification. Cette faiblesse est un peu décevante, car on pouvait s'attendre à ce que ce problème soit partiellement résolu par les linguistes. C'est sur la base des données ainsi recueillies que nous avons mis au point notre méthodologie d'extraction de terminologie. De manière à n'extraire et ne compter que les co-occurrences lexicales susceptibles d'être, morphosyntaxiquement, des noms composés, nous avons préalablement étiqueté et lemmatisé le corpus. Ces deux opérations se sont effectuées à l'aide de programmes stochastiques. Ces programmes sont performants et n'engendrent que peu d'erreurs pour notre tâche. Néanmoins, les quelques erreurs commises par ces programmes introduisent du bruit dans nos résultats. La qualité des résultats de notre technique d'extraction repose sur la qualité de l'étiquetage grammatical et morphologique. Nous avons utilisé des automates finis pour relever les co-occurrences. Leur performance est supérieure à la technique de la fenêtre qui est généralement utilisée dans les travaux antérieurs au nôtre. Les automates permettent de préciser dans quel contexte morphosyntaxique les co-occurrences doivent apparaître, sans restreindre *a priori* la distance entre les deux mots. Nous

avons regroupé les co-occurrences extraites par couples : un couple est composé de deux lemmes qui correspondent aux deux unités lexicales pleines principales de la structure d'un nom composé. Les co-occurrences présentes sous l'entrée d'un couple relèvent le plus souvent d'un même concept sauf, pour certains cas, où le couple regroupe des concepts différents. Ce problème de signification du dénombrement est inhérent à l'approche quantitative, et l'utilisation de filtres linguistiques ne l'élimine pas totalement. Nous avons ensuite évalué un certain nombre de modèles statistiques dont ceux qui ont déjà été utilisés pour l'extraction de ressources lexicales comme par exemple le score d'association, proche du concept d'information mutuelle. Les résultats de cette évaluation se sont révélés surprenants : le score d'association n'a pas donné de bons résultats à l'inverse de la fréquence. Ce résultat est à l'encontre, en quelque sorte, de nombreux travaux précédents qui montraient le meilleur comportement du score d'association par rapport à la fréquence. Ce résultat contradictoire s'explique sans doute par l'introduction de filtres linguistiques. Il reste que les couples fréquents caractérisent sans aucune équivoque les termes d'un domaine, au contraire des valeurs fortes du score d'association qui isolent plutôt des noms composés figés appartenant à la langue courante. Nous n'avons pas pourtant élu la fréquence comme meilleur modèle statistique, mais le coefficient de vraisemblance, ou plutôt le test du rapport de vraisemblance appliqué à une loi binomiale. Celui-ci intègre beaucoup moins de bruit que la fréquence et présente un classement conceptuel performant où l'on retrouve les termes les plus fréquents mais aussi d'autres, moins fréquents mais tout aussi corrects, au moins dans les valeurs les plus fortes.

Le système que nous avons construit à base d'automates finis a donc permis une amélioration sensible des performances dans l'extraction automatique de ressources lexicales et a permis d'établir l'efficacité de l'enrichissement des systèmes statistiques par la linguistique. Le problème principal, qui est non résolu au terme de cette étude et qui est le principal responsable du bruit présent restant dans nos résultats, est celui de la surcomposition et de la modification. Ni la linguistique, ni la statistique n'ont pour l'instant apporté de solution véritable. Il nous semble toutefois que c'est en approfondissant les spécifications linguistiques qu'on pourra minimiser ce problème. Nous avons présenté rapidement deux possibilités d'extension de notre méthode : d'une part les ressources lexicales bilingues et d'autre part, l'extraction de structures argumentales. Autant le premier domaine semble prometteur et fera l'objet de futurs développements ([Gaussier, 1994]), autant les résultats dans l'extraction de structures argumentales sont peu encourageants. On retrouve ici les limites des méthodes statistiques dans l'analyse automatique.

En conclusion, ce travail a établi l'intérêt de l'exploitation des corpus pour la création de bases lexicales. Pour réaliser des systèmes avec applications réelles, l'usage des corpus est à notre avis indispensable de même que l'exploitation des modèles statistiques. Leur efficacité certaine ne doit pas masquer leurs limites : les domaines d'application sont restreints et ils ne sont réellement performants qu'en proportion de l'apport linguistique qui est introduit.

Bibliographie

[Baker, 1979] Baker (J.K.). – Trainable grammars for speech recognition. *In : Speech communication papers presented at the 97th Meeting of the Acoustical Society of America*, éd. par Klatt (D.H.) et J.J.Wolf. – Cambridge, MA, MIT, Juin 1979.

[Baum, 1972] Baum (L. E.). – An inequality and associated maximization technique in statistical estimation for probalistics function of a markov process. *Inequalities*, vol. 3, 1972, pp. 1–8.

[Benveniste, 66] Benveniste (Émile). – Formes nouvelles de la composition nominale. *In : Problèmes de linguistique générale*, pp. 163–173. – Paris, Gallimard, 1966.

[Bourigault, 1994] Bourigault (Didier). – *Acquisition de terminologie.* – Thèse, EHESS, 1994.

[Brent, 1991b] Brent (Michael R.). – Automatic aquisition of subcategorization frames from untagged text. *In : 29th Annual Meeting of the ACL.* ACL. – Berkeley, California, Juin 1991.

[Brent, 1991a] Brent (Michael R.). – Automatic semantic classification of verbs from their syntactic contexts : An implemented classifier for stativity. *In : Proceedings of the 5th Conference of the European Chapter of the ACL*, pp. 222–226. – Berlin, Germany, Avril 1991.

[Brill, 1992] Brill (Eric). – A simple rule-based part of speech tagger. *In : Proceedings of the Third Conference on Applied Natural Language Processing (ANLP-92)*, pp. 152–155. – Trento, Italy, Avril 1992.

[Briscoe et Carroll, 1993] Briscoe (Ted) et Carroll (John). – Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, vol. 19, n° 1, Mars 1993, pp. 25–61.

[Brown *et al.*, 1988] Brown (Peter F.), Cocke (John), Pietra (Stephen A. Della), Pietra (Vincent J. Della), Jelinek (Frederik), Mercer (Robert L.) et Roossin (Paul S.). – A statistical approach to language translation. *In : Proceeding of*

the 12th International Conference on Computational Linguistics (Coling-88). – Budapest, Hungary, Août 1988.

[Brown *et al.*, 1990] Brown (P.), Cocke (J.), Pietra (S. Della), Pietra (V. Della), Jelinek (F.), Lafferty (J.), Mercer (R.) et Roossin (P.). – A statistical approach to machine translation. *Computational Linguistics*, vol. 16, n° 2, Juin 1990.

[Brown *et al.*, 1991] Brown (Peter F.), Lai (Jennifer C.) et Mercer (Robert L.). – Aligning sentences in parallel corpora. *In : Proceedings of the 29th Annual Meeting of the ACL.* – Berkeley, California, Juin 1991.

[Brown *et al.*, 1992] Brown (Peter F.), Cocke (John), Pietra (Stephen A. Della), Pietra (Vincent J. Della), Jelinek (Frederick), Lafferty (John D.), Mercer (Robert L.) et Roossin (Paul S.). – Analysis, statistical transfer, and synthesis in machine translation. *In : Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, pp. 83–100. – Montréal, Canada, Juin 1992.

[Calzolari et Bindi, 1990] Calzolari (Nicoletta) et Bindi (Remo). – Acquisition of lexical information from a large textual italian corpus. *In : Proceedings of the Thirteenth International Conference on Computational Linguistics.* – Helsinki, Finland, 1990.

[Catizone *et al.*, 1989] Catizone (Roberta), Russell (Graham) et Warwick (Susan). – Deriving translation data from bilingual texts. *In : Proceeding of the First International Workshop on Lexical Acqusition (IJCAI'89).* – Detroit, Mi, Août 1989.

[Chomsky, 1957] Chomsky (N.). – *Syntactic Structures.* – Mouton, 1957.

[Church, 1988] Church (K.). – A stochastic parts program and noun phrase paser for unrestricted texts. *In : Proceedings of Second Conference on Applied Natural Language Processing.* – Austin, Texas, 1988.

[Church et Hanks, 1989] Church (Kenneth Ward) et Hanks (Patrick). – Word association norms, mutual information, and lexicography. *In : Proceeding of the 27th Annual Meeting of the ACL.* – Vancouver, Canada, Juin 1989.

[Church et Hanks, 1990] Church (Kenneth Ward) et Hanks (Patrick). – Word Association Norms, Mutual Information, and Lexicography. *In : Computational Linguistics*, volume 16-1, pages 22–29.– Mars 1990.

[Cutting *et al.*, 1992] Cutting (Doug), Kupiec (Julian), Pedersen (Jan) et Sibun (Penelope). – A practical part-of-speech tagger. *In : Proceedings of the Third Conference on Applied Language Processing (ANLP-92).* – Trento, Italy, Avril 1992.

[Dagan et Itai, 1990] Dagan (Ido) et Itai (Alon). – Automatic processing of large corpora for the resolution of anaphora references. *In : Proceedings of the Thirteenth International Conference on Computational Lingusitics (Coling-90).* – Helsinki, Finland, Août 1990.

[Dagan *et al.*, 1991] Dagan (Ido), Itai (Alon) et Schwall (Ulrike). – Two languages are more informative than one. *In : Proceedings of the 29th Annual Meeting of the ACL.* – Berkeley, California, Juin 1991.

[Daille, 1993] Daille (Béatrice). – Extraction automatique de terminologie monolingue. *In : Colloque Informatique et Langue Naturelle.* Nantes, Décembre 93.

[Daille et McEnery, 1993] Daille (Béatrice) et McEnery (Tony). – *Database Design for Corpus Storage : The ET10-63 Data Model.* – Rapport technique n° 1, Unit for Computer Research on the English Language - Lancaster University, 1993.

[Danlos, 1980] Danlos (Laurence). – *Représentation d'informations linguistiques : les constructions N être Prep X.* – Thèse de doctorat en Linguistique, Université de Paris 7, 1980.

[Darmesteter, 1894] Darmesteter (Arsène). – *Traité de la Formation des noms composés.* – Paris, Bouillon, 1894.

[Dérouault, 1985] Dérouault (Anne-Marie). – *Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques.* – Thèse, Université Paris VII, 1985.

[Dérouault et Merialdo, 1986] Dérouault (Anne Marie) et Merialdo (B.). – Natural language modeling for phoneme-to-text transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, n° 6, Novembre 1986, pp. 742–749.

[Dologlou *et al.*, 1991] Dologlou (Yannis), Malnati (Giovanni) et Paggio (Patrizia). – A preference mecanism based on multiple criteria resolution. *In : Proceeding of Fith Conference of the European Chapter of the Association for Computational Linguistics*, pp. 281–286.

[Dunning, 1993] Dunning (Ted). – Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. 19, n° 1, Mars 1993.

[El-Bèze, 1993] El-Bèze (Marc). – *Les Modèles de Langage Probabilistes : Quelques Domaines d'Applications.* – Habilitation à diriger des recherches, LIPN : Université Paris-Nord, Janvier 1993.

[Enguehard, 1992] Enguehard (Chantal). – *ANA, Apprentissage Naturel Automatique d'un réseau sémantique.* – Thèse, Université de technologie de Compiègne, 1992.

[Francis et Kucera, 1982] Francis (W. N.) et Kucera (F.). – *Frequency Analysis of English Usage.* – Houghton Mifflin, 1982.

[Foster, 1991] Foster (George F.). – *Statistical Lexical Disambiguation.* – Montreal, Canada, Thèse, McGill University, School of Computer Science, 1991.

[Fujisaki *et al.*, 1989] Fujisaki (T.), Jelinek (F.), Cocke (J.), Black (E.) et Nishino (T.). – A probabilistic method for sentence disambiguation. *In : Proceedings of the 1st International Workshop on Parsing Technologies*, pp. 105–114. – Carnegie-Mellon University, Pittsburg, PA, 1989.

[Gale et Church, 1991a] Gale (William A.) et Church (Kenneth W.). – A program for aligning sentences in bilingual corpora. *In : Proceedings of the 29th Annual Meeting of the ACL.* – Berkeley, California, Juin 1991.

[Gale et Church, 1991b] Gale (William A.) et Church (Kenneth W.). – Concordances for parallel texts. *In : Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, Using Corpora*, pp. 40–62. – Oxford, U.K., 1991.

[Galisson et Coste, 1976] Galisson (R.) et Coste (D.). – *Dictionnaire Didactique des Langues.* – Paris, Hachette, 1976.

[Garside *et al.*, 1987] Garside (Roger.), Leech (Geoffrey) et Sampson (Geoffrey). – *The Computational Analysis of English.* – Longman, 1987.

[Gaussier *et al.*, 1992] Gaussier (Eric), Langé (Jean Marc) et Meunier (Frederic). – Towards bilingual terminology. *In : Proceedings of ALLC/ACH Conference.* – Oxford, England, 1992.

[Gaussier et Langé, 1993] Gaussier (Eric) et Langé (Jean Marc). – Construction de termes monolingues à l'aide d'information bilingue. *In : Actes des Troisièmes Journées Scientifiques LTT - AUPELF-UREF.* – Montréal, Québec, 1993.

[Gaussier, 1994] Dérouault (Anne-Marie). – *Introduction des probabilités dans le module de transfert en traduction automatique.* – Thèse, Université Paris VII, 1994.

[Green et Rubin, 1971] Greene (B.B.) et Rubin (G. M.). – *Automatic grammatical tagging of English.* – Rapport technique, Providence, Rhode Island, Department of Linguistics, Brown University, 1971.

[Grévisse et Goosse, 1986] Grévisse (Maurice) et Goosse (André). – *Le Bon Usage.* – Paris-Gembloux, Duculot, Mars 1986.

[G. Gross *et al.*, 1986] Gross (Gaston), Chaurand (Jacques), Vivès (Robert), Mathieu-Colas (Michel) et Billy (Pierre). – *Typologie des noms composés.* – Rapport technique A.T.P.- Nouvelles Recherches sur le Langage - Université Paris 13, 1986.

[G. Gross *et al.*, 1987] Gross (Gaston), Jung (René) et Mathieu-Colas (Michel). – *Noms composés.* – Rapport technique n° 5, Paris, Programme de recherches Coordonnées "Informatique Linguistique", CNRS, 1987.

[G. Gross, 1988] Gross (Gaston). – Degré de figement des noms composés. *Langages*, vol. 90, 1988. – Larousse, Paris.

[M. Gross, 1986] Gross (Maurice). – *Les adjectifs composés du Français.* – Rapport technique n° 3, Paris, Programme de recherches Coordonnées "Informatique Linguistique", CNRS, 1986.

[M. Gross, 1988] Gross (Maurice). – Sur les phrases figées complexes du français. *Langue Française*, vol. 77, 1988. – Larousse, Paris.

[M. Gross, 1989] Gross (Maurice). – *Grammaire Transformationnelle du Français : 3) Syntaxe de l'adverbe.* – Paris, Cantilène, 1989.

[Harris, 1968] Harris (Zelig S.). – *Mathematical Structures of Language.* – New York, Wiley, 1968.

[Harris, 1976] Harris (Zellig S.). – *Notes du cours de syntaxe.* – Paris, Le Seuil, 1976.

[Hausman, 1985] Hausmann (Franz Josef). – Kollokationen im deutschen wörterbuch. ein beitrag zur theorie des lexikographischen beispiels. *In : Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Worterbuch*, pp. 118–129. – Henning Bergenholtz/Joachin Mudgan, 1985. Lexicographica. Series Maior 3.

[Heid et Raab, 1989] Heid (Ulrich) et Raab (Sybille). – Collocations in multilingual generation. *In : Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics.* – Manchester, England, April 1989.

[Hindle, 1990] Hindle (D.). – Noun classification from predicate-argument structures. *In : Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics.* pp. 268–275. – Pittsburgh, 1990.

[Hindle et Rooth, 91] Hindle (D.) et Rooth (M.). – Structural ambiguity and lexical relations. *In : Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 229–236. – Berkeley, California, Juin 1991.

[Jacquemin, 1991] Jacquemin (Christian). – *Transformations des noms composés.* – Thèse de doctorat en Informatique Fondamentale, Université Paris 7, 1991.

[Jelinek, 1985] Jelinek (F.). – The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, vol. 73, n° 11, Novembre 1985.

[Johansson *et al.*, 1978] Johansson (S.), Leech (G.N.) et Goodluck (H.). – *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* – Departement d'Anglais, Université d'Oslo, 1978.

[Klavans et Tzoukermann, 1990] Klavans (Judith) et Tzoukermann (Evelyne). – The BICORD system : Combining lexical information from bilingual corpora and machine readable dictionaries. *In : Proceeding of the 13th International Conference on Computational Linguistics (Coling-90).* – Helsinki, Finland, Août 1990.

[Lafon, 1984] Lafon (Pierre). – *Dépouillements et Statistiques en Lexicométrie.* – Genève, Slatkine - Champion, 1984.

[Macklovitch, 1992] Macklovitch (Elliott). – Where the taggers falters. *In : Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, pp. 113–126. – Montréal, Canada, Juin 1992.

[Magerman et Marcus, 1991] Magerman (D. M.) et Marcus (M. P.). – Pearl: A probabilistic chart parsing. *In : Proceedings of European ACL.*

[Martinet, 1960] Martinet (André). – *Éléments de linguistique générale.* – Paris, Colin, 1960.

[Martinet, 1979] Martinet (André). – *Grammaire Fonctionnelle du Français.* – Paris, Crédif, 1979.

[Martinet, 1985] Martinet (André). – *Syntaxe Générale.* – Paris, Armand Colin, 1985.

[Mathieu-Colas, 1988] Mathieu-Colas (Michel). – *Typologie des noms composés.* – Rapport technique n° 7, Paris, Programme de Recherches Coordonées "Informatique et Linguistique", Université Paris 13, 1988.

[Mel'čuk *et al.*, 1984] Mel'čuk (Igor), Arbatchewsky-Jumarie (Nadia), Elnitsky (Léo), Iordanskaja (Lidija) et Lessard (Adèle). – *Dictionnaire explicatif et combinatoire du français contemporain.* – Montréal, Les Presses de L'Université de Montréal, 1984.

[Meunier, 1981] Meunier (Annie). – *Nominalisations d'adjectifs par verbes supports.* – Thèse de doctorat en Linguistique, Université Paris 7, 1981.

[Nkwenti-Azeh, 1992] Nkwenti-Azeh (Basile). – *Positional and Combinational Characteristics of Satellite Communications Terms.* – Rapport technique, Manchester, UK, CCL-UMIST, 1992.

[Noailly, 1990] Noally (Michèle). – *Le substantif épithète.* – Paris, PUF, 1990.

[Pereira et Schabes, 1992] Pereira (Fernando) et Schabes (Yves). – Inside-outside reestimation from partially braketed corpora. *In : 20th Meeting of the Association for Computional Linguistics (ACL'92).* – Newark, Delaware, 1992.

[Picoche, 1977] Picoche (Jacqueline). – *Précis de Lexicologie Française.* – Paris, Nathan, 1977.

[Piot, 1988] Piot (Mireille). – Conjonctions de subordination et figement. *Langages*, vol. 90, 1988. – Larousse, Paris.

[Poncet-Montange, 1991] Poncet-Montange (Anne). – *Les groupes nominaux de structure NAN et NAV.* – Thèse de Doctorat en Linguistique, Université Paris 13, 1991.

[Pottier, 1985] Pottier (B.). – *Linguistique Générale. Théorie et Description.* – Paris, Klincksieck, 1985.

[Sharman *et al.*, 1990] Sharman (R.), Jelinek (F.) et Mercer (R.). – Generating a grammar for statistical training. *In : DARPA Speech and Natural Language Workshop*, pp. 267–274. – Hidden Valley, PA, 1990.

[Silberztein, 1989] Silberztein (Max Dan). – *Dictionnaires électroniques et Reconnaissance lexicale automatique.* – Thèse de doctorat en Informatique Fondamentale, Université Paris 7, 1989.

[Silberztein, 1993] Silberztein (Max Dan). – *Dictionnaires électroniques et Analyse automatique de texte - Le système INTEX.* – Paris, Masson, 1993.

[Simard *et al.*, 1992] Simard (Michel), Forster (Georges F.) et Isabelle (Pierre). – Using Cognates to Align Sentences in Bilingual Corpora. *In : Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-92). – Montréal, Canada, June 1992.

[Smadja et McKeown, 1990] Smadja (Frank A.) et McKeown (Kathleen R.). – Automatically extracting and representing collocations for language generation. *In : Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252–259.

[Warwick et Russel, 1990] Warwick (S.) et Russel (G.). – Bilingual concordancing and bilingual lexicography. *In : EURALEX 4^{th} International Congress.* – Malaga, Spain, 1990.

[Van der Eijk, 1993] Van der Eijk (Pim). – Automating the acquisition of bilingual terminology. *In : Proceedings of the Sixth Conference of the European Chapter of the ACL - EACL-93.* – Utrecht, Hollande, 1993.

[Warwick *et al.*, 1992] Warwick-Armstrong (S.), Hajic (J.) et Winarske (A.). – Tagging and alignment of parralel texts : Current status of bcp. *In : Proceedings of the Third Conference on Applied Language Processing*, pp. 227–229. – Trento, Italy, 1992.

[Weaver, 1949] Weaver (Warren). – Translation. *In : Machine Translation of Languages*, éd. par Locke et Booth. – Cambridge, MA, MIT Press, 1955.

Annexe A

Listes des étiquettes grammaticales du français

AAAA Ponctuation

ADJEFP Adjectif Genre Féminin Nombre Pluriel

ADJEFS Adjectif Genre Féminin Nombre Singulier

ADJEMP Adjectif Genre Masculin Nombre Pluriel

ADJEMS Adjectif Genre Masculin Nombre Singulier

ADJIFP Adjectif Indéfini Genre Féminin Nombre Pluriel

ADJIFS Adjectif Indéfini Genre Féminin Nombre Singulier

ADJIMP Adjectif Indéfini Genre Masculin Nombre Pluriel

ADJIMS Adjectif Indéfini Genre Masculin Nombre Singulier

ADVE Adverbe

AUXA Auxiliaire *avoir*

AUXA1 Auxiliaire *avoir* Personne 1. Nombre Singulier

AUXA2 Auxiliaire *avoir* Personne 2. Nombre Singulier

AUXA3 Auxiliaire *avoir* Personne 3. Nombre Singulier

AUXA4 Auxiliaire *avoir* Personne 1. Nombre Pluriel

AUXA5 Auxiliaire *avoir* Personne 2. Nombre Pluriel

AUXA6 Auxiliaire *avoir* Personne 3. Nombre Pluriel

AUXE Auxiliaire *être*

AUXE1 Auxiliaire *être* Personne 1. Nombre Singulier

AUXE2 Auxiliaire *être* Personne 2. Nombre Singulier

AUXE3 Auxiliaire *être* Personne 3. Nombre Singulier

AUXE4 Auxiliaire *être* Personne 1. Nombre Pluriel

AUXE5 Auxiliaire *être* Personne 2. Nombre Pluriel

AUXE6 Auxiliaire *être* Personne 3. Nombre Pluriel

CCOO Conjonction de Coordination

CHIF Chiffre

CSUB Conjonction de Subordination

DETRFP Déterminant Genre Féminin Nombre Pluriel

DETRFS Déterminant Genre Féminin Nombre Singulier

DETRMP Déterminant Genre Masculin Nombre Pluriel

DETRMS Déterminant Genre Masculin Nombre Singulier

DINTFP Déterminant Interrogatif Genre Féminin Nombre Pluriel

DINTFS Déterminant Interrogatif Genre Féminin Nombre Singulier

DINTMP Déterminant Interrogatif Genre Masculin Nombre Pluriel

DINTMS Déterminant Interrogatif Genre Masculin Nombre Singulier

NE Négation

NPRO Nom Propre

PAS *pas, plus*

PAU Préposition contractée it au Nombre Singulier

PAUX Préposition contracté aux Nombre Pluriel

PDEA Préposition *de, d', à* et *es*

PDES Préposition contracté *des* Nombre Pluriel

PDETFP Pronom Démonstratif Genre Féminin Nombre Pluriel

PDETFS Pronom Démonstratif Genre Féminin Nombre Singulier

PDETMP Pronom Démonstratif Genre Masculin Nombre Pluriel

PDETMS Pronom Démonstratif Genre Masculin Nombre Singulier

PINDFS Pronom Indéfini Genre Féminin Nombre Singulier

PINDFP Pronom Indéfini Genre Féminin Nombre Pluriel

PINDMS Pronom Indéfini Genre Masculin Nombre Singulier

PINDMP Pronom Indéfini Genre Masculin Nombre Pluriel

PINTFS Pronom Interrogatif Genre Féminin Nombre Singulier

PINTFP Pronom Interrogatif Genre Féminin Nombre Pluriel

PINTMS Pronom Interrogatif Genre Masculin Nombre Singulier

PINTMP Pronom Interrogatif Genre Masculin Nombre Pluriel

PPASFP Participe Passé Genre Féminin Nombre Pluriel

PPASFS Participe Passé Genre Féminin Nombre Singulier

PPASMP Participe Passé Genre Masculin Nombre Pluriel

PPASMS Participe Passé Genre Masculin Nombre Singulier

PPER1 Pronom Personnel 1. Nombre Singulier

PPER2 Pronom Personnel 2. Nombre Singulier

PPER3F Pronom Personnel 3. Genre Féminin Nombre Singulier

PPER3M Pronom Personnel 3. Genre Masculin Nombre Singulier

PPER4 Pronom Personnel 1. Nombre Pluriel

PPER5 Pronom Personnel 2. Nombre Pluriel

PPER6F Pronom Personnel 3. Genre Féminin Nombre Pluriel

PPER6M Pronom Personnel 3. Genre Masculin Nombre Pluriel

PPOBFP Pronom Personnel Objet Genre Féminin Nombre Pluriel

PPOBFS Pronom Personnel Objet Genre Féminin Nombre Singulier

PPOBMP Pronom Personnel Objet Genre Masculin Nombre Pluriel

PPOBMS Pronom Personnel Objet Genre Masculin Nombre Singulier

PPRE Participe Présent

PREFFP Pronom Réfléchi Genre Féminin Nombre Pluriel

PREFFS Pronom Réfléchi Genre Féminin Nombre Singulier

PREFMP Pronom Réfléchi Genre Masculin Nombre Pluriel

PREFMS Pronom Réfléchi Genre Masculin Nombre Singulier

PRELFP Pronom Relatif Genre Féminin Nombre Pluriel

PRELFS Pronom Relatif Genre Féminin Nombre Singulier

PRELMP Pronom Relatif Genre Masculin Nombre Pluriel

PRELMS Pronom Relatif Genre Masculin Nombre Singulier

PREP Préposition

PREPMS Préposition contracté *du* Genre Masculin Nombre Singulier

SUBSFP Substantif Genre Féminin Nombre Pluriel

SUBSFS Substantif Genre Féminin Nombre Singulier

SUBSMP Substantif Genre Masculin Nombre Pluriel

SUBSMS Substantif Genre Masculin Nombre Singulier

VERB1 Verbe Personne 1 Nombre Singulier

VERB2 Verbe Personne 2 Nombre Singulier

VERB3 Verbe Personne 3 Nombre Singulier

VERB4 Verbe Personne 1 Nombre Pluriel

VERB5 Verbe Personne 2 Nombre Pluriel

VERB6 Verbe Personne 3 Nombre Pluriel

VINF Verbe Infinitif

XFAMIL Nom de Famille

XPAYFP Nom de Pays Féminin Pluriel

XPAYFS Nom de Pays Féminin Singulier

XPAYMP Nom de Pays Masculin Pluriel

XPAYMS Nom de Pays Masculin Singulier

XPREF Prénom Féminin

XPREM Prénom Masculin

XSOC Nom de Société

XVILLE Nom de Ville

YAAA Ponctuation Faible

ZTRM Retour chariot

Annexe B

Listes des étiquettes grammaticales de l'anglais

! punctuation tag - exclamation mark (*!*)

" punctuation tag - quotes (*"*)

(punctuation tag - left bracket (*(*)

) punctuation tag - right bracket (*)*)

, punctuation tag - comma (,)

- punctuation tag - dash (-)

. punctuation tag - full-stop (.)

... punctuation tag - ellipsis (...)

: punctuation tag - colon (*:*)

; punctuation tag - semicolon (*;*)

? punctuation tag - question mark (*?*)

APPGE possessive pronoun, pre-nominal (*e.g. my, your, our*)

AT article (*e.g. the, no*)

AT1 singular article (*e.g. a, an. every*)

BCL before-clause marker (*e.g. in order (that), in order (to)*)

CC coordinating conjunction (*e.g. and, or*)

CCB adversative coordinating conjunction (*but*)

CS subordinating conjunction (*e.g. if, because, unless, so, for*)

CSA *as* (as conjunction)

CSN *than* (as conjunction)

CST *that* (as conjunction)

CSW *whether* (as conjunction)

DA after-determiner or post-determiner capable of pronominal function (*e.g. such, former, same*)

DA1 singular after-determiner (*e.g. little, much*)

DA2 plural after-determiner (*e.g. few, several, many*)

DAR comparative after-determiner (*e.g. more, less, fewer*)

DAT superlative after-determiner (*e.g. most, least, fewest*)

DB before determiner or pre-determiner capable of pronominal function (*all, half*)

DB2 plural before-determiner (*both*)

DD determiner (capable of pronominal function) (*e.g any, some*)

DD1 singular determiner (*e.g. this, that, another*)

DD2 plural determiner (*these,those*)

DDQ wh-determiner (*which, what*)

DDQGE wh-determiner, genitive (*whose*)

DDQV wh-ever determiner (*whichever, whatever*)

EX existential there

FO formula

FW foreign word

GE germanic genitive marker (*'* or *'s*)

IF *for* (as preposition)

II general preposition

IO *of* (as preposition)

IW *with, without* (as prepositions)

JJ general adjective

JJR general comparative adjective (*e.g. older, better, stronger*)

JJT general superlative adjective (*e.g. oldest, best, strongest*)

JK catenative adjective (*able in be able to, willing in be willing to*)

MC cardinal number,neutral for number (*two, three..*)

MCGE genitive cardinal number, neutral for number (*two's, 100's*)

MC-MC hyphenated number (*40-50, 1770-1827*)

MC1 singular cardinal number (*one*)

MC2 plural cardinal number (*tens, hundreds*)

MD ordinal number (*e.g. first, second, next, last*)

MF fraction,neutral for number (*e.g. quarters, two-thirds*)

ND1 singular noun of direction (*e.g. north, southeast*)

NN common noun,neutral for number (*e.g. sheep, cod, headquarters*)

NN1 singular common noun (*e.g. book, girl*)

NN2 plural common noun (*e.g. books, girls*)

NNA following noun of title (*e.g. M.A.*)

NNB preceding noun of title (*e.g. Mr., Prof.*)

NNJ organization noun, neutral for number (*e.g. council, department*)

NNJ2 organization noun, plural (*e.g. governments, committees*)

NNL1 singular locative noun (*e.g. island, street*)

NNL2 plural locative noun (*e.g.islands, streets*)

NNO numeral noun, neutral for number (*e.g. dozen, hundred*)

NNO2 numeral noun, plural (*e.g. hundreds, thousands*)

NNT1 temporal noun,singular (*e.g. day, week, year*)

NNT2 temporal noun,plural (*e.g. days, weeks, years*)

NNU unit of measurement,neutral for number (*e.g. in, cc*)

NNU1 singular unit of measurement (*e.g. inch, centimetre*)

NNU2 plural unit of measurement (*e.g. ins., feet*)

NP proper noun, neutral for number (*e.g. IBM, Andes*)

NP1 singular proper noun (*e.g. London, Jane, Frederick*)

NP2 plural proper noun (*e.g. Browns, Reagans, Koreas*)

NPD1 singular weekday noun (*e.g. Sunday*)

NPD2 plural weekday noun (*e.g. Sundays*)

NPM1 singular month noun (*e.g. October*)

NPM2 plural month noun (*e.g. Octobers*)

PN indefinite pronoun, neutral for number (*none*)

PN1 indefinite pronoun, singular (*e.g. anyone, everything, nobody, one*)

PNQO objective wh-pronoun (*whom*)

PNQS subjective wh-pronoun (*who*)

PNQV wh-ever pronoun (*whoever*)

PNX1 reflexive indefinite pronoun (*oneself*)

PPGE nominal possessive personal pronoun (*e.g. mine, yours*)

PPH1 3rd person sing. neuter personal pronoun (*it*)

PPHO1 3rd person sing. objective personal pronoun (*him, her*)

PPHO2 3rd person plural objective personal pronoun (*them*)

PPHS1 3rd person sing. subjective personal pronoun (*he, she*)

PPHS2 3rd person plural subjective personal pronoun (*they*)

PPIO1 1st person sing. objective personal pronoun (*me*)

PPIO2 1st person plural objective personal pronoun (*us*)

PPIS1 1st person sing. subjective personal pronoun (*I*)

PPIS2 1st person plural subjective personal pronoun (*we*)

PPX1 singular reflexive personal pronoun (*e.g. yourself, itself*)

PPX2 plural reflexive personal pronoun (*e.g. yourselves, themselves*)

PPY 2nd person personal pronoun (*you*)

RA adverb, after nominal head (*e.g. else, galore*)

REX adverb introducing appositional constructions (*namely, e.g.*)

RG degree adverb (*very, so, too*)

RGQ wh- degree adverb (*how*)

RGQV wh-ever degree adverb (*however*)

RGR comparative degree adverb (*more, less*)

RGT superlative degree adverb (*most, least*)

RL locative adverb (*e.g. alongside, forward*)

RP prep. adverb, particle (*e.g about, in*)

RPK prep. adv., catenative (*about in be about to*)

RR general adverb

RRQ wh- general adverb (*where, when, why, how*)

RRQV wh-ever general adverb (*wherever, whenever*)

RRR comparative general adverb (*e.g. better, longer*)

RRT superlative general adverb (*e.g. best, longest*)

RT quasi-nominal adverb of time (*e.g. now, tomorrow*)

TO infinitive marker (*to*)

UH interjection (*e.g. oh, yes, um*)

VB0 *be* base form (finite i.e. imperative, subjunctive)

VBDR *were*

VBDZ *was*

VBG *being*

VBI *be* infinitive (*To be or not... It will be ..*)

VBM *am*

VBN *been*

VBR *are*

VBZ *is*

VD0 *do* base form (finite)

VDD *did*

VDG *doing*

VDI *do* infinitive (*I may do... To do...*)

VDN *done*

VDZ *does*

VH0 *have* base form (finite)

VHD *had* (past tense)

VHG *having*

VHI *have* infinitive

VHN *had* (past participle)

VHZ *has*

VM modal auxiliary (*can, will, would, etc.*)

VMK modal catenative (*ought, used*)

VV0 base form of lexical verb (*e.g. give, work*)

VVD past tense of lexical verb (*e.g. gave, worked*)

VVG -ing participle of lexical verb (*e.g. giving, working*)

VVGK -ing participle catenative (*going in be going to*)

VVI infinitive (*e.g. to give... It will work...*)

VVN past participle of lexical verb (*e.g. given, worked*)

VVNK past participle catenative (*e.g. bound in be bound to*)

VVZ -s form of lexical verb (*e.g. gives, works*)

XX *not, n't*

ZZ1 singular letter of the alphabet (*e.g. A,b*)

ZZ2 plural letter of the alphabet (*e.g. A's, b's*)

Annexe C

Classement des couples proposé par le coefficient de vraisemblance

C.1 Corpus MTS

N_1 (PREP (DET)) N_2

Valeurs décroissantes du critère de vraisemblance

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
514	largeur	bande		223	1327	0.691	21.34	0.38
5110	température	bruit		126	777	0.592	20.13	0.65
6	bande	base	*(bande + bandes) de base*	145	745	0.515	19.88	2.19
7194	amplificateur	puissance	*(amplificateur + amplificateurs) de puissance*	137	727	0.517	19.86	1.37
2937	temps	propagation		94	611	0.611	19.80	1.80
3904	règlement	radiocommunication	*règlement des radiocommunications*	60	521	0.803	19.96	0.49
2604	produit	intermodulation	*(produit + produits) d' intermodulation*	61	457	0.626	19.31	0.63
3256	taux	erreur	*taux d' erreur*	70	420	0.448	18.61	1.02
1	mise	oeuvre	*mise en oeuvre*	47	355	0.563	18.60	2.16
2	télécommunication	satellite		99	353	0.228	17.35	1.05
678	bilan	liaison		55	349	0.390	17.99	0.26
1297	mémoire	tampon	*mémoires tampons + mémoire tampon*	35	343	0.823	19.30	0.70
3810	concentration	conversation		39	315	0.630	18.71	1.08
761	diagramme	rayonnement	*(diagramme + diagrammes) de rayonnement*	40	306	0.578	18.51	1.47
7383	angle	site		37	300	0.622	18.59	1.69
7776	propagation	groupe	*propagation de groupe*	38	292	0.574	18.43	1.52
3423	correction	erreur		46	280	0.352	17.49	0.75
8617	moteur	apogée		28	253	0.683	18.49	1.22
1200	titre	exemple	*titre d' exemple*	29	252	0.645	18.40	1.08
5672	dispersion	énergie	*dispersion d' énergie*	33	249	0.453	17.72	0.38
6116	répéteur	satellite		76	246	0.177	16.41	1.44
3324	réduction	puissance		52	245	0.261	16.84	1.41
5718	assignation	demande	*assignation à la demande*	30	233	0.534	17.91	1.31
9774	capacité	trafic		53	230	0.262	16.75	2.90
3	bande	fréquence		89	223	0.187	16.50	2.19
3649	puissance	sortie		57	222	0.241	16.54	2.99
8616	transmission	donnée		65	221	0.216	16.46	3.06
9349	objectif	qualité		34	216	0.352	17.08	1.06

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
8411	diamètre	antenne		39	194	0.200	16.00	0.71
5746	alimentation	énergie	*alimentation (en + d') énergie*	33	186	0.304	16.57	2.06
2546	code	convolution	*(code + codes) à convolution*	24	183	0.416	16.84	2.71
8854	schéma	principe	*(schéma + schémas) de principe*	23	180	0.471	17.18	2.59
8972	unité	voie		42	180	0.219	16.05	2.52
9472	accusé	réception	*(accusé + accusés) de réception*	28	179	0.240	16.10	0.15
8135	suppresseurs	écho	*suppresseurs d' écho*	25	175	0.375	16.85	0.64
4268	multiplication	circuit		25	171	0.307	16.47	0.54
2852	antenne	station		62	171	0.159	15.70	2.74
6119	multiplexage	répartition	*multiplexage par répartition*	23	170	0.412	16.96	1.54
9381	accès	répartition	*accès multiple (par + à) répartition*	28	170	0.327	16.48	2.66
824	canal	sémaphore	*canal sémaphore + canaux sémaphores*	26	166	0.344	16.57	2.27
7011	couplage	amplification		21	165	0.453	17.09	1.32
6584	programme	télévision	*(programme + programmes) de télévision*	28	163	0.267	16.17	1.77
7111	guide	onde		20	153	0.345	16.50	0.30
5881	réseau	satellite		97	153	0.133	15.78	2.94
9163	système	signalisation		85	150	0.143	15.63	3.66
466	gain	antenne		46	149	0.153	15.32	2.54
5218	zone	couverture	*(zone + zones) de couverture*	22	147	0.348	16.37	2.04
5708	maintien	position	*maintien en position*	18	145	0.449	16.86	1.35
6284	service	radiodiffusion		28	144	0.236	15.47	3.23
1269	traitement	signal		39	143	0.159	15.25	2.25
1365	liaison	satellite		82	141	0.124	15.43	2.92
2925	modulation	delta	*modulation delta*	19	137	0.325	15.84	2.52
3935	méthode	modulation		29	134	0.204	15.37	3.04
2241	train	bit	*(train + trains) de bits*	21	132	0.282	15.90	1.37
8913	facteur	qualité	*(facteur + facteurs) de qualité*	27	131	0.206	15.37	2.50
2777	rapport	porteur		42	129	0.150	14.98	**3.38**
7915	durée	vie	*durée de vie*	16	127	0.411	16.34	2.85
5292	station	référence		36	126	0.165	14.91	3.51
6657	amplification	puissance	*amplification de puissance*	23	126	0.150	14.95	0.57
9097	système	alimentation	*(système + systèmes) d' alimentation*	47	123	0.142	14.76	3.66
8485	puissance	entrée		38	123	0.153	14.84	2.99
4246	pourcentage	temps		20	122	0.232	15.51	1.49
7480	batterie	accumulateur	*(batterie + batteries) d' accumulateurs*	11	121	0.715	17.46	0.75
148	code	bloc		21	120	0.254	15.41	2.71
1028	mise	place	*mise en place*	18	119	0.291	15.50	2.16
7377	tube	onde	*(tube + tubes) à ondes*	19	118	0.249	15.53	1.31
4771	trame	sémaphore	*trame sémaphore + trames sémaphores*	17	116	0.292	15.78	1.73
954	module	interface		16	115	0.303	15.85	1.33
554	étalement	spectre	*étalement (de + du) spectre*	13	114	0.379	16.24	0.26
7861	centre	commutation	*(centre + centres) de commutation*	27	113	0.171	14.79	2.79
7813	liaison	connexion	*(liaison + liaisons) de connexion*	24	112	0.192	14.72	2.92
248	faisceau	antenne		32	112	0.120	14.49	2.40
6425	modulation	déplacement	*modulation par déplacement*	15	107	0.282	15.14	2.52
7329	radiodiffusion	satellite		26	105	0.057	13.73	0.31
1431	déplacement	phase		16	104	0.217	15.19	0.97
337	orbite	transfert		15	104	0.303	15.59	1.22
3601	voie	retour	*(voie + voies) de retour*	18	102	0.219	14.69	3.68
4154	distance	coordination	*distance de coordination*	15	102	0.271	15.45	1.77
1699	commande	orientation	*commande d' orientation*	16	102	0.262	15.10	3.35
4326	bruit	intermodulation	*bruit d' intermodulation*	28	101	0.147	14.36	3.39
100	bout	bout	*bout en bout*	11	101	0.475	16.46	1.12
239	intervalle	temps	*(intervalle + intervalles) de temps*	17	99	0.187	14.85	1.51
453	rayonnement	antenne		22	97	0.098	14.08	1.30
7592	démodulateur	seuil	*(démodulateur + démodulateurs) à seuil*	13	97	0.307	15.60	1.65
7985	bit	contrôle	*(bit + bits) de contrôle*	18	97	0.207	14.76	3.00
2128	point	vue	*point de vue*	18	96	0.207	14.64	3.36
8438	système	satellite		108	95	0.110	15.32	3.66

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
5971	réseau	terre		44	95	0.111	14.21	2.94
7330	service	satellite		66	95	0.095	14.54	3.23
6696	équipement	multiplexage		24	92	0.148	14.04	3.61
9671	entrée	récepteur	*entrée du récepteur*	16	91	0.212	14.67	2.93
4781	poursuite	échelon	*poursuite par échelons*	11	90	0.376	15.73	2.00
1271	interface	terre	*(interface + interfaces) de terre*	23	90	0.112	14.00	2.40
8936	flux	donnée		19	89	0.113	14.07	1.43
4697	répartition	fréquence	*répartition en (fréquence + fréquences)*	23	88	0.084	13.77	1.48
3602	erreur	pointage		18	88	0.173	14.31	3.24
2203	signalisation	enregistreur	*signalisation (entre + d') enregistreurs*	15	87	0.192	14.04	3.80
6198	train	impulsion	*train d' impulsions*	13	86	0.254	14.98	1.37
6402	équipement	télécommunication	*(équipement + équipements) de télécommunication*	38	86	0.106	13.92	3.61
8344	centre	contrôle	*centre de contrôle*	18	86	0.169	14.18	2.79
5705	satellite	télécommunication		30	85	0.107	13.82	3.77
4979	excursion	fréquence		20	83	0.073	13.58	1.04
4284	registre	décalage	*registre à décalage*	7	83	0.698	16.89	0.38
3040	affaiblissement	espace	*affaiblissement en espace*	13	82	0.232	14.67	2.45
2879	boucle	verrouillage	*boucle à verrouillage*	9	80	0.404	15.64	2.04
7112	onde	polarisation		14	79	0.166	14.30	2.28
3467	récupération	porteur		16	79	0.101	13.79	1.29
4166	modulation	fréquence	*modulation de fréquence*	33	79	0.085	13.60	2.52
4197	rapport	signal		35	78	0.093	13.65	**3.38**
8474	système	couplage	*(système + systèmes) de couplage*	20	76	0.124	13.19	3.66
9636	chaîne	amplification		14	76	0.179	14.13	2.19
7074	longueur	onde	*(longueur + longueurs) d' onde*	14	76	0.156	14.10	2.29
6223	exploitation	réseau	*exploitation (du réseau + des réseaux)*	21	74	0.086	13.42	2.29
6681	algorithme	décodage	*(algorithme + algorithmes) de décodage*	10	74	0.260	14.92	1.82
8487	signal	entrée		30	73	0.100	13.42	3.95
9109	modèle	référence	*modèle de référence*	12	72	0.136	14.03	1.37
8307	densité	puissance		17	72	0.078	13.35	1.30
6112	système	télécommunication		46	71	0.093	13.68	3.66
975	conduit	référence		11	71	0.133	14.07	0.86
9076	débit	information		15	71	0.132	13.68	2.61
4987	matrice	commutation	*(matrice + matrices) de commutation*	12	70	0.129	13.91	0.98
5415	tube	hyperfréquence		11	70	0.204	14.32	1.31
7340	égaliseur	temps		11	69	0.124	13.92	1.04
1004	table	erlang	*tables d' erlang*	6	69	0.613	16.42	0.41
3072	terre	espace	*terre vers espace*	11	69	0.204	14.21	2.40
6176	stabilisation	rotation	*stabilisation par rotation*	8	68	0.270	15.02	0.64
8768	réduction	débit	*réduction (de + du) débit*	15	68	0.131	13.51	1.41
223	axe	faisceau		13	68	0.132	13.71	2.40
1645	point	multipoint	*point à multipoint*	11	67	0.195	13.79	3.36
1413	effet	champ	*effet de champ*	10	67	0.228	14.20	3.58
6243	service	exploration	*service d' exploration*	12	66	0.161	13.29	3.23
4789	poursuite	mono-impulsion	*poursuite mono-impulsion*	8	66	0.325	14.93	2.00
995	verrouillage	phase	*verrouillage de phase*	10	65	0.124	13.87	0.90
152	mot	code	*(mot + mots) de code*	10	65	0.151	14.00	1.43
7790	générateur	secours	*(génératrice + génératrices) de secours*	10	64	0.182	14.10	2.12
5779	réseau	communication		29	64	0.088	13.09	2.94
3761	niveau	brouillage		20	64	0.100	13.04	3.60
4664	modulation	amplitude	*modulation d' amplitude*	14	62	0.130	13.22	2.52
4914	synchronisation	trame		12	62	0.132	13.51	2.59
4251	contour	coordination	*(contour + contours) de coordination*	8	62	0.177	14.32	0.64
1361	liaison	terre		31	62	0.080	13.08	2.92
4623	ligne	alimentation		14	62	0.103	13.20	2.69
9683	chaîne	émission		18	61	0.080	12.93	2.19
1345	affaiblissement	pluie		9	61	0.217	13.99	2.45
5354	gestion	réseau		17	60	0.064	12.83	2.22

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
733	calcul	bilan		10	60	0.182	13.71	3.13
9435	sou-système	utilisateur	*sous-système utilisateur*	10	60	0.170	13.76	2.68
2616	appareil	mesure	*appareils de (mesure + mesures)*	7	60	0.268	14.82	0.94
4661	commutation	paquet		13	60	0.111	13.20	2.78
5488	architecture	réseau		14	60	0.054	12.84	1.31
336	signal	bande		38	59	0.076	13.17	3.95
6423	précision	pointage	*précision (de + du) pointage*	10	59	0.139	13.58	2.36
8736	communication	donnée	*(communication + communications) de données*	21	58	0.073	12.74	2.55
5220	procédure	coordination		11	58	0.134	13.34	2.94
7613	accroissement	température		8	58	0.182	14.10	1.51
5671	communication	satellite		34	57	0.052	12.89	2.55
1143	écoulement	trafic	*écoulement (du + de) trafic*	9	57	0.045	13.04	0.33
7707	sortie	amplificateur		15	57	0.095	12.81	3.30
1179	lancement	satellite		16	57	-0.001	12.12	0.78
8822	possibilité	correction	*(possibilité + possibilités) de correction*	10	56	0.143	13.43	2.91
4837	bit	information	*(bit + bits) d' information*	14	56	0.100	12.87	3.00
1665	échange	programme		8	56	0.166	13.95	1.19
6516	centre	exploitation		16	56	0.094	12.75	2.79
6897	distribution	télévision		13	56	0.085	12.85	2.50
3677	bruit	récepteur		14	55	0.111	12.72	3.39
6902	établissement	communication		13	55	0.074	12.76	2.35
486	signal	luminance		12	55	0.122	12.56	3.95
9352	traitement	donnée		19	54	0.066	12.52	2.25
2144	plupart	cas	*plupart des cas*	10	53	0.133	13.13	3.02
8601	suppression	écho	*(suppression + suppressions) d' écho*	9	53	0.112	13.28	1.75
8206	annuleurs	écho	*annuleurs d' écho*	8	53	0.111	13.47	0.67
8770	continuité	service	*continuité (de + du) service*	9	53	0.015	12.58	0.33
1566	orbite	satellite	*(orbite + orbites) des satellites*	21	52	0.023	12.22	1.22
6117	zone	service	*(zone + zones) de service*	16	52	0.060	12.44	2.04
339	refroidissement	air	*refroidissement (à + par) air*	6	52	0.312	14.58	2.09
7220	équipement	commutation	*(équipement + équipements) de commutation*	21	51	0.079	12.40	3.61
2732	cas	transmission		28	51	0.067	12.60	**4.12**
8866	qualité	transmission		21	51	0.061	12.40	3.16
8093	débit	symbole		8	51	0.177	13.37	2.61
2329	compensation	mouvement	*compensation (de + du) mouvement*	7	51	0.184	13.85	1.64
539	perte	gain		10	51	0.109	12.90	2.98
520	surface	terre	*surface de la terre*	12	51	0.041	12.40	1.65
2530	transport	message	*transport de (messages + message)*	6	50	0.157	13.99	0.41
5475	architecture	étoile	*architecture en étoile*	7	50	0.202	13.80	1.31
4444	signal	sortie		24	49	0.073	12.40	3.95
2851	station	type		21	49	0.075	12.33	3.51
355	période	éclipse		6	49	0.261	14.01	2.71
9291	technique	accès	*(technique + techniques) d' accès*	12	48	0.090	12.44	3.44
4206	communication	entreprise		10	48	0.120	12.63	2.55
9912	qualité	fonctionnement	*qualité de fonctionnement*	18	48	0.064	12.24	3.16
3509	ordinateur	serveur	*ordinateur serveur*	4	48	0.576	15.83	0.50
8921	caractéristique	qualité	*(caractéristique + caractéristiques) de qualité*	18	48	0.073	12.21	3.92
3225	intégration	service	*intégration (des + de) services*	10	47	0.026	12.27	1.40
6260	réseau	distribution	*(réseau + réseaux) de distribution*	13	47	0.096	12.11	2.94
7382	nombre	voie	*nombre (de + des) voies*	21	47	0.065	12.24	**3.91**
5421	calcul	rapport		12	47	0.077	12.33	3.13
2179	voisinage	saturation		7	47	0.146	13.40	2.08
4557	ambiguïté	phase	*ambiguïté de phase*	7	47	0.058	12.97	0.68
3096	espace	terre	*espace vers terre*	9	47	0.008	12.22	0.72
5507	type	limiteur	*type à limiteur*	9	46	0.121	12.17	**4.23**
3155	km	altitude	*km d' altitude*	4	46	0.507	15.56	0.50
1457	centre	commande	*centre de commande*	14	46	0.075	12.14	2.79
2919	emplacement	station		12	46	0.006	11.88	1.50
3837	saut	répéteur	*saut de répéteur*	9	46	0.045	12.41	1.27
9442	chaîne	réception		15	45	0.054	12.05	2.19

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
9433	objet	accusé		7	45	0.161	13.19	2.86
3648	puissance	bruit		25	45	0.062	12.27	2.99
1572	point	point	*point à point*	12	45	0.087	12.17	3.36
64	bit	échantillon	*bits par échantillon*	8	45	0.138	12.71	3.00
6478	convertisseur	fréquence		10	45	-0.014	11.79	0.69
3802	équipement	multiplication	*(équipement + équipements) de multiplication*	12	45	0.094	11.99	3.61
5757	fiche	notification	*(fiche + fiches) de notification*	4	45	0.453	15.34	0.50
18	ko	bit	*ko bits*	6	45	0.065	13.10	0.41
4044	antenne	réception		21	45	0.060	12.11	2.74
3781	transmission	programme		13	44	0.085	11.98	3.06
1779	bord	satellite		16	44	-0.001	11.63	1.50
6758	réseau	microstations	*(réseau + réseaux) de microstations*	13	44	0.087	11.90	2.94
4681	paquet	référence		11	44	0.068	12.15	2.65
5093	sortie	démodulateur	*sortie du démodulateur*	9	44	0.110	12.37	3.30
5052	commutation	bord		8	43	0.119	12.62	2.78
1106	lobe	antenne		10	43	-0.005	11.76	1.08
1268	bruit	brouillage		16	43	0.068	11.91	3.39
9161	numéro	séquence	*(numéro + numéros) de séquence*	5	43	0.216	14.02	1.15
6357	discrimination	polarisation		7	43	0.065	12.68	1.16
2702	bruit	quantification	*bruit de quantification*	10	43	0.101	12.02	3.39
5347	valeur	rapport		11	43	0.067	12.04	3.09
2330	alimentation	courant	*(alimentation + alimentations) en courant*	8	42	0.120	12.45	2.06
3572	station	réception		24	42	0.060	12.10	3.51
6237	service	télécommunication		23	42	0.059	12.04	3.23
4806	convertisseur-élévateur	fréquence	*convertisseur-élévateur de fréquence*	12	41	0.008	11.63	1.05
5299	réseau	télécommunication		26	41	0.061	12.14	2.94
4706	rapport	horizon	*rapport à l' horizon*	8	41	0.118	12.04	**3.38**
9095	objectif	disponibilité		7	41	0.130	12.66	1.06
898	différence	couleur	*différence de couleur*	5	41	0.231	13.73	2.37
7560	préambule	paquet		6	41	0.041	12.66	0.56
2416	prise	ressource	*prise de ressource*	5	41	0.208	13.74	1.73
2855	type	modulation		18	40	0.064	11.79	**4.23**
9174	dégradation	qualité		9	40	0.051	12.01	2.23
8144	câble	fibre	*(câble + câbles) à fibres*	4	40	0.358	14.66	0.97
6864	site	lancement	*site de lancement*	6	40	0.133	13.00	1.82
5561	augmentation	espacement	*augmentation de l' espacement*	5	40	0.225	13.38	3.16
7171	antenne	dimension		8	40	0.112	11.96	2.74
3235	linéariseur	type	*linéariseur du type*	7	40	0.027	12.19	1.12
8717	travail	génie	*travaux de génie*	4	39	0.342	14.43	1.89
5508	pays	membre	*pays membres + pays membre*	5	39	0.203	13.49	2.20
3619	détecteur	parole	*détecteur de parole*	6	39	0.112	12.81	1.85
8926	efficacité	utilisation	*efficacité d' utilisation*	6	39	0.115	12.81	2.31
9223	région	océan	*région de l' océan*	4	39	0.333	14.24	2.06
3822	sens	transmission	*sens de transmission*	10	39	0.013	11.61	1.87
854	circuit	référence		13	39	0.061	11.65	3.88
1474	centre	transit	*centre de transit*	8	39	0.106	11.99	2.79
1136	mode	fonctionnement	*(mode + modes) de fonctionnement*	15	38	0.048	11.60	3.01
4685	processus	application	*processus d' application*	7	38	0.101	12.26	3.12
9100	conception	système		14	38	0.022	11.47	2.42
5795	mode	multidestination	*mode multidestination*	6	38	0.142	12.26	3.01
1701	équipement	traitement		14	37	0.068	11.50	3.61
6936	point	saturation	*point de saturation*	9	37	0.083	11.68	3.36
4573	signalisation	ligne	*signalisation de ligne*	12	37	0.071	11.46	3.80
3432	récepteur	station		13	37	0.011	11.34	2.18
5644	boucle	réaction	*boucle de réaction*	5	36	0.171	12.98	2.04
4141	disposition	règlement	*(disposition + dispositions) du règlement*	5	36	0.175	12.87	3.19
5882	réflecteur	antenne		8	36	-0.035	11.24	0.89
6069	écran	visualisation	*écran de visualisation*	3	36	0.500	15.23	0.56
9096	système	poursuite	*(système + systèmes) de poursuite*	18	36	0.064	11.46	3.66
1763	cas	défaillance	*cas de défaillance*	10	36	0.078	11.40	**4.12**
9020	ligne	abonné		6	36	0.122	12.28	2.69

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
3152	trafic	répéteur		10	36	0.039	11.45	2.85
21	foi/fois	an	*fois par an*	4	36	0.254	13.77	2.03
9603	excursion	crête	*excursion de crête*	5	35	0.162	12.83	1.04
3629	niveau	bruit		19	35	0.047	11.55	3.60
9270	facteur	activité	*facteur d' activité*	6	35	0.124	12.11	2.50
3746	gain	multiplication		8	35	0.086	11.61	2.54
4962	sémaphore	message	*sémaphore de message*	6	35	0.085	12.21	2.17
2916	station	émission		21	35	0.052	11.61	3.51
1105	fonctionnement	mode	*fonctionnement en mode*	12	35	0.067	11.26	**4.82**
463	itinéraire	trafic	*(itinéraire + itinéraires) de trafic*	6	34	-0.025	11.61	0.74
4520	plan	fréquence		15	34	0.022	11.26	2.91
2901	émission	station		17	34	0.027	11.35	3.09
8205	système	commande		21	34	0.059	11.53	3.66
9670	entrée	amplificateur		10	34	0.051	11.33	2.93
968	combineurs	filtre	*combineurs à filtres*	6	34	0.061	12.00	1.85
7685	maintien	poste	*maintien à poste*	5	34	0.137	12.53	1.35
1768	défaillance	équipement		8	34	0.028	11.42	2.64
3639	appel	offre	*(appel + appels) d' offres*	4	34	0.218	13.02	3.28
5377	réseau	étoile		9	33	0.078	11.13	2.94
2142	loi	illumination	*loi d' illumination*	3	33	0.408	14.64	1.01
4596	publication	renseignement		3	33	0.362	14.64	0.56
820	code	rendement		6	33	0.113	11.78	2.71
8930	tonalité	essai	*tonalité d' essai*	4	33	0.133	13.20	0.96
916	fréquence	échantillonnage	*fréquence d' échantillonnage*	9	33	0.072	11.23	3.90
684	choix	emplacement		5	33	0.142	12.26	3.10
5568	énergie	bit		7	33	0.050	11.57	2.62
2694	message	résultat	*message de résultat*	4	33	0.207	13.08	2.13
6318	récepteur	poursuite	*récepteur de poursuite*	7	33	0.055	11.55	2.18
9350	accès	assignation	*accès multiple avec assignation*	7	33	0.084	11.53	2.66
1262	volume	trafic	*volume de trafic*	6	33	-0.024	11.44	1.00
8236	cadre	système		11	33	0.002	11.01	2.13
1798	méthode	calcul	*(méthode + méthodes) de calcul*	7	32	0.085	11.48	3.04
3518	détecteur	mouvement	*détecteur de mouvement*	5	32	0.094	12.31	1.85
5511	emploi	technique		6	32	0.086	11.80	3.27
3368	détection	erreur		8	32	0.008	11.17	1.97
6285	modulation	impulsion	*modulation par impulsions*	8	32	0.072	11.25	2.52
1833	pompe	chaleur	*(pompe + pompes) à chaleur*	3	32	0.280	14.23	0.56
4941	corrélation	erreur	*corrélation des erreurs*	6	32	-0.032	11.28	0.85
2995	alimentation	antenne		15	31	0.024	11.10	2.06
2986	affaiblissement	transmission	*affaiblissement de transmission*	12	31	0.023	11.02	2.45
7169	dimension	antenne		9	31	-0.015	10.89	1.94
1587	mise	point	*mise au point*	9	31	0.057	11.09	2.16
5834	transmission	satellite		39	31	0.044	12.32	3.06
1745	signal	image		12	31	0.058	10.95	3.95
4910	répartition	code	*répartition en code*	7	31	0.047	11.30	1.48

Valeur indéfinie du critère de vraisemblance

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
4551	abaisseur	fréquence	*(abaisseur + abaisseurs) de fréquence*	12	∞	0.011	12.69	0.00
5556	adaptateur	interface	*adaptateur d' interfaces*	2	∞	-0.175	10.49	0.00
3098	adjudication	contrat	*adjudication du contrat*	2	∞	0.354	14.47	0.00
9240	agilité	fréquence	*agilité (en + de) fréquence*	7	∞	-0.070	11.13	0.00
1466	aiguille	montre	*aiguilles d' une montre*	2	∞	0.646	15.47	0.00
5358	allumage	moteur	*allumage (du moteur + des moteurs)*	5	∞	0.207	14.36	0.00
4942	amortissement	nutation	*amortissement de la nutation*	2	∞	0.463	14.89	0.00
8374	angle	obliquité	*angle d' obliquité*	2	∞	0.116	10.52	1.69
9037	annulation	écho	*annulation d' écho*	3	∞	-0.094	11.34	0.00
8388	annuleur	écho	*annuleur d' écho*	5	∞	0.028	12.81	0.00
3052	antenne	tore	*(antenne + antennes) tore*	2	∞	0.065	8.86	2.74
996	arséniure	gallium	*arséniure de gallium*	11	∞	0.849	17.93	0.00
502	asservissement	antenne	*asservissement de l' antenne*	2	∞	-0.286	7.70	0.00
2272	atop	linéariseur	*atop avec linéariseur*	5	∞	0.567	16.11	0.00
4228	attente	numérotation	*attente après numérotation*	2	∞	-0.020	12.30	0.00
771	axe	déclinaison	*axe de déclinaison*	2	∞	0.148	11.22	2.40
9723	banque	donnée	*banques de données*	3	∞	-0.184	9.54	0.00
171	base	tarif	*base des tarifs*	2	∞	0.107	10.28	3.84
2140	bas	page	*bas de page*	2	∞	0.457	14.47	1.04
7986	bit	parité	*(bit + bits) de parité*	3	∞	0.132	11.20	3.00
4288	bouton	poussoir	*bouton poussoir*	2	∞	0.373	13.89	1.56
133	câblage	baie	*câblage entre baies*	2	∞	0.457	14.47	1.04
5029	caractérisation	information	*caractérisation de (ces informations + l' information)*	2	∞	-0.209	9.89	0.00
8115	carte	crédit	*cartes de crédit*	2	∞	0.373	13.89	1.56
2658	centrage	spectre	*centrage du spectre*	2	∞	-0.145	10.95	0.00
1476	centre	télémaintenance	*centre de télémaintenance*	2	∞	0.080	9.45	2.79
431	chemin	roulement	*chemin de roulement*	2	∞	0.646	15.47	0.00
550	circuit	silencieux	*circuit de silencieux*	2	∞	0.079	9.42	3.88
4202	circulateur	ferrite	*circulateur à ferrite*	2	∞	0.528	14.89	0.64
3562	codage	voix	*codage de la voix*	2	∞	0.120	10.61	3.34
792	codeur-décodeur	récurrence	*codeur-décodeur à récurrence*	2	∞	0.108	10.30	2.31
9107	collège	formation	*(collège + collèges) de formation*	2	∞	0.181	13.66	0.00
5352	collecteur	dépression	*collecteur (en + à) dépression*	3	∞	0.436	14.64	1.67
7378	commission	étude	*commission d' études*	6	∞	0.358	15.39	0.00
2589	compresseur	émission	*compresseur à l' émission*	2	∞	-0.260	8.63	0.00
7997	conductivité	sol	*conductivité du sol*	2	∞	-0.059	11.95	0.00
2853	conférence	radiocommunication	*conférence administrative (extraordinaire + mondiale) des radiocommunications*	4	∞	-0.014	12.30	0.00
4894	conseil	gouverneur	*conseil des gouverneurs*	2	∞	0.409	14.15	1.05
4787	convention	télécommunication	*convention internationale des télécommunications*	2	∞	-0.264	8.50	0.00
8051	côte	côte	*côte à côte*	2	∞	0.646	15.47	0.00
7557	cristal	glace	*cristaux de glace*	2	∞	0.646	15.47	0.00
1493	croissance	trafic	*croissance du trafic*	3	∞	-0.165	10.03	0.00
1187	débordement	mémoire	*débordement de la mémoire*	2	∞	-0.029	12.22	0.00
3540	découplage	polarisation	*découplage (des polarisations + de polarisation)*	2	∞	-0.205	9.96	0.00
2984	déformation	antenne	*(déformations + déformation) de l' antenne*	2	∞	-0.286	7.70	0.00
9191	démultiplexeur	entrée	*démultiplexeur d' entrée*	5	∞	-0.049	11.76	0.00
1658	déploiement	satellite	*déploiement du satellite*	2	∞	-0.310	6.44	0.00
2723	db	octave	*db par octave*	2	∞	0.152	11.30	2.94

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
3825	dent	scie	*dent de scie*	2	∞	0.646	15.47	0.00
724	diode	laser		2	∞	0.289	13.15	2.03
1732	diode	tunnel	*diodes tunnel*	2	∞	0.289	13.15	2.03
6655	dissipation	chaleur		2	∞	0.146	13.47	0.00
1157	dizaine	mbit	*dizaines de mbit*	2	∞	0.373	13.89	1.33
197	dizaine	millisecondes	*dizaines de millisecondes*	2	∞	0.373	13.89	1.33
8266	durée	reconnaissance	*(durée + durées) de reconnaissance*	2	∞	0.122	10.66	2.85
4369	effet	masque	*effet de masque*	3	∞	0.146	11.49	3.58
4196	égalisation	temps		11	∞	0.150	14.46	0.00
9545	égalité	droit	*égalité (de + des) droits*	5	∞	0.776	16.79	0.00
5626	élément	contact	*élément de contact*	2	∞	0.141	11.08	2.83
5111	élévation	température		3	∞	-0.028	12.18	0.00
8932	embrouillage	donnée	*embrouillage de données*	2	∞	-0.268	8.37	0.00
8532	entité	propriétaire	*entité propriétaire*	2	∞	0.354	14.47	0.00
5581	énergie	désembrouillage	*énergie par désembrouillage*	2	∞	0.162	11.47	2.62
1663	équinoxe	automne	*équinoxe d' automne*	3	∞	0.551	15.32	0.67
1542	équinoxe	printemps	*équinoxe de printemps*	2	∞	0.409	14.15	0.67
6381	évanouissement	propagation	*(évanouissement + évanouissements) de propagation*	5	∞	-0.039	11.92	0.00
6621	extenseur	réception	*extenseur à la réception*	2	∞	-0.263	8.54	0.00
813	fibre	carbone	*fibres de carbone*	2	∞	0.528	14.89	0.64
6543	filtre	miroir	*filtres miroirs*	2	∞	0.141	11.08	3.09
6429	focalisation	bobine	*focalisation (à bobine + par bobines)*	2	∞	0.457	14.47	0.69
2401	fonction	tri	*fonction de tri*	2	∞	0.065	8.83	4.52
9193	formatage	donnée		2	∞	-0.268	8.37	0.00
1897	glossaire	fascicule	*glossaire du fascicule*	2	∞	0.646	15.47	0.00
2779	indice	modulation	*indice de modulation*	4	∞	-0.071	11.51	0.00
8740	interconnectivité	station	*interconnectivité entre stations*	3	∞	-0.213	8.62	0.00
708	interface	saisie	*interface de commande et de saisie*	2	∞	0.120	10.61	2.40
6521	interliaison	réseau	*interliaison de réseaux*	2	∞	-0.275	8.12	0.00
3298	intertrame	mouvement	*intertrame à mouvement*	2	∞	-0.100	11.52	0.00
5858	inversion	phase	*inversion de phase*	5	∞	-0.008	12.36	0.00
7232	jonction	orthomode	*(jonction + jonctions) orthomode*	3	∞	0.418	15.06	0.00
7564	juridiction	partie	*juridiction d' une partie*	3	∞	0.174	13.83	0.00
1077	laps	temps	*laps de temps*	2	∞	-0.226	9.54	0.00
8406	largeur	fenêtre	*largeur de la fenêtre*	2	∞	0.059	8.56	0.38
9062	levée	ambiguïté		4	∞	0.644	16.15	0.00
897	lien	codage	*liens entre codage*	2	∞	-0.197	10.11	0.00
8145	ligne	flèche	*lignes à une flèche*	2	∞	0.127	10.77	2.69
5956	luminance	amplitude	*luminance à l' amplitude*	2	∞	-0.152	10.86	0.00
7822	mécanisme	captage	*(mécanisme + mécanismes) de captage*	2	∞	0.236	12.56	2.21
7264	méthode	évaluation	*méthodes d' évaluation*	2	∞	0.089	9.76	3.04
7287	méthode	détermination	*(méthode + méthodes) de détermination*	3	∞	0.120	10.93	3.04
3045	matériel	démonstration		2	∞	0.457	14.47	1.04
1510	maximum	vraisemblance	*maximum de vraisemblance*	5	∞	0.481	15.41	1.73
9119	modèle	fiche	*modèles de fiches*	2	∞	0.204	12.15	1.37
5809	modulation	inversion	*(modulation + modulations) par inversion*	5	∞	0.151	12.06	2.52
2787	modulation	multiphase	*modulation multiphase*	2	∞	0.079	9.42	2.52
4512	mono-impulsion	multimode	*mono-impulsion multimode*	2	∞	0.457	14.47	0.69
3259	monture	antenne	*(monture + montures) d' antenne*	2	∞	-0.286	7.70	0.00
9341	moteur	périgée	*moteur de périgée*	2	∞	0.143	11.11	1.22
1827	moyen	navette	*moyen de la navette*	2	∞	0.075	9.27	4.24
6488	multiplexeur	sortie	*multiplexeur de sortie*	5	∞	-0.052	11.71	0.00
789	orbite	parking	*orbite de parking*	2	∞	0.141	11.08	1.22
2447	ordre	grandeur	*ordre de grandeur*	3	∞	0.251	13.06	2.61
4709	ouverture	mi-puissance	*ouverture à mi-puissance*	2	∞	0.254	12.77	1.38
7299	personne	juridiction	*personne sous juridiction*	2	∞	0.646	15.47	0.00
8856	pièce	rechange	*pièces de rechange*	3	∞	0.366	14.83	0.00
3613	plan	équateur	*plan de l' équateur*	3	∞	0.155	11.66	2.91
4542	plan	allotissement	*plan d' allotissement*	2	∞	0.115	10.49	2.91
2334	plan	découpage	*plan de découpage*	2	∞	0.115	10.49	2.91

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
546	plaque	diélectrique	*plaque diélectrique*	6	∞	0.616	16.32	0.90
3086	plaque	quart	*plaque quart*	3	∞	0.390	14.32	0.90
6938	point	interception	*point d' interception*	7	∞	0.186	13.03	3.36
7067	polariseur	onde	*polariseur quart d' onde*	2	∞	-0.204	10.00	0.00
4530	pompage	fréquence		2	∞	-0.290	7.52	0.00
5261	porteur	plage	*porteuse par plage*	2	∞	0.088	9.70	3.35
2475	pression	rayonnement	*pression du rayonnement*	2	∞	-0.171	10.56	0.00
4684	processus	interrogation	*processus d' interrogation*	2	∞	0.138	11.01	3.12
829	profondeur	décharge	*profondeur de décharge*	2	∞	0.528	14.89	0.64
9897	pureté	polarisation	*pureté de polarisation*	7	∞	0.088	13.58	0.00
5914	quadrature	phase	*quadrature de phase*	3	∞	-0.122	10.89	0.00
7338	quart	onde	*quart d' onde*	3	∞	-0.105	11.17	0.00
4961	rail	azimut	*rail d' azimut*	2	∞	0.073	13.01	0.00
2157	rapport	découplage	*rapport de découplage*	2	∞	0.063	8.74	3.38
6448	rapport	isolement	*rapport d' isolement*	2	∞	0.063	8.74	3.38
1009	récurrence	redondance	*récurrence avec redondance*	2	∞	-0.045	12.08	0.00
518	refroidissement	hélium	*refroidissement (à + par) hélium*	2	∞	0.204	12.15	2.09
815	rendement	combustible	*rendement du combustible*	2	∞	0.143	11.11	2.75
2102	repliement	spectre	*repliement du spectre*	2	∞	-0.145	10.95	0.00
8545	reportage	actualité		3	∞	0.711	16.06	0.00
5026	réorganisation	acheminement	*réorganisation de l' acheminement*	2	∞	0.012	12.56	0.00
7064	répéteur	écrêteur	*répéteurs à écrêteur*	2	∞	0.088	9.70	1.44
6355	réseau	collecte	*(réseau + réseaux) de collecte*	3	∞	0.066	9.18	2.94
6277	réussite	lancement		2	∞	-0.095	11.56	0.00
4696	réutilisation	fréquence		27	∞	0.137	15.03	0.00
4384	risque	accident	*risques d' accident*	2	∞	0.289	13.15	1.47
3778	rotation	faraday	*rotation de faraday*	2	∞	0.204	12.15	2.51
9004	saisie	donnée	*saisie de données*	2	∞	-0.268	8.37	0.00
4855	saut	cycle	*(saut + sauts) de cycle*	2	∞	0.215	12.30	1.27
3516	serveur	station	*serveur de la station*	2	∞	-0.292	7.45	0.00
3990	service	radiocommunication	*(service(s) de radiocommunication) + (services de radiocommunications)*	10	∞	0.164	13.08	3.23
5832	signalisation	décade	*signalisation à décade*	3	∞	0.075	9.57	3.80
3928	silence	conversation	*silences de la conversation*	2	∞	-0.165	10.66	0.00
1760	socle	antenne	*socle de l' antenne*	3	∞	-0.206	8.87	0.00
3680	sortie	multiplicateur	*sortie du multiplicateur*	2	∞	0.092	9.86	3.30
9395	sou-réseau	accès	*sous-réseau d' accès*	2	∞	-0.195	10.15	0.00
3878	station	correspondance	*stations en correspondance*	3	∞	0.072	9.44	3.51
8600	suppresseur	écho	*suppresseur d' écho*	13	∞	0.267	15.57	0.00
7083	synthétiseur	fréquence		12	∞	0.011	12.69	0.00
8246	système	avertissement	*(système + systèmes) d' avertissement*	2	∞	0.035	7.08	3.66
194	télésurveillance	alarme	*télésurveillance des alarmes*	2	∞	-0.065	11.89	0.00
2053	technologie	invar	*technologie de l' invar*	2	∞	0.254	12.77	2.35
212	tec	arséniure	*tec à arséniure*	3	∞	0.486	15.32	0.00
5118	température	brillance	*température de brillance*	2	∞	0.077	9.32	0.65
8700	temps	montée	*temps de montée*	2	∞	0.077	9.33	1.80
8056	théorème	réciprocité	*théorème de réciprocité*	2	∞	0.457	14.47	0.69
401	trajet	descendant	*trajets descendants*	2	∞	0.155	11.34	2.74
3700	transistor	effet	*(transistor + transistors) à effet*	10	∞	0.516	16.66	0.00
2119	transit	satellite	*transit à satellites*	6	∞	-0.128	9.61	0.00
192	tri	cliques/clique	*tri (des + de) cliques*	2	∞	0.646	15.47	0.00
6019	type	famille	*types de famille*	2	∞	0.053	8.25	4.23
1484	union	république	*union des républiques*	2	∞	0.409	14.15	1.05
1499	véhicule	lancement	*(véhicule + véhicules) de lancement*	3	∞	0.028	12.73	0.00
5081	vérification	carte	*vérification (de + des) cartes*	2	∞	0.305	13.30	1.89
8884	visée	antenne	*visée de l' antenne*	2	∞	-0.286	7.70	0.00
7187	voie	conséquence	*voie de conséquence*	2	∞	0.065	8.83	3.68
6528	volant	inertie	*volant d' inertie*	3	∞	0.418	15.06	0.00
25	vol	bit	*vol de bits*	2	∞	-0.195	10.15	0.00
7451	zone	ouest	*zone ouest*	2	∞	0.116	10.52	2.04

N ADJ

Valeurs décroissantes du critère de vraisemblance

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
1708	station	terrien	*stations terriennes + station terrienne*	750	2933	0.842	22.47	0.24
3680	débit	binaire	*débit binaire + débits binaires*	134	715	0.675	19.45	1.24
4116	accès	multiple	*accès multiple*	105	605	0.664	19.09	1.74
3588	voie	téléphonique		118	511	0.512	18.52	2.24
2847	liaison	montant	*liaison montante + liaisons montantes*	88	456	0.519	18.09	0.26
1282	liaison	descendant		77	407	0.491	17.75	0.14
1580	secteur	spatial	*secteur spatial*	79	341	0.409	17.41	2.21
1795	service	fixe	*service fixe + services fixes*	66	326	0.467	17.42	1.19
3836	lobe	latéral	*lobes latéraux*	40	299	0.699	17.93	0.72
1952	faisceau	hertzien	*faisceau hertzien + faisceaux hertziens*	35	244	0.586	17.22	0.40
1157	puissance	surfacique	*puissance surfacique + puissances surfaciques*	35	231	0.502	16.76	0.13
2013	polarisation	circulaire		36	204	0.443	16.49	1.15
3106	oscillateur	local	*oscillateurs locaux + oscillateur local*	32	200	0.434	16.45	2.94
1515	bruit	thermique		35	183	0.393	16.15	1.92
2936	polarisation	rectiligne	*polarisation rectiligne + polarisations rectilignes*	28	181	0.450	16.16	0.29
1499	erreur	binaire	*erreur binaire + erreurs binaires*	40	158	0.263	15.40	1.24
2207	espace	libre	*espace libre*	18	156	0.678	16.83	0.63
1364	groupe	primaire		27	152	0.385	15.82	2.00
3546	amplificateur	paramétrique	*amplificateur paramétrique + amplificateurs paramétriques*	19	142	0.534	16.22	1.05
1022	signal	vocal	*signal vocal + signaux vocaux*	32	140	0.286	15.09	1.41
188	commande	automatique	*commande automatique*	20	137	0.421	15.80	2.48
3082	réflecteur	secondaire	*réflecteur secondaire + réflecteurs secondaires*	24	135	0.360	15.51	1.50
2810	décision	souple		15	131	0.642	16.44	0.73
5	orbite	géostationnaire		18	123	0.442	15.70	0.69
3261	démodulation	cohérent		15	118	0.521	15.94	1.44
939	équipement	terminal	*équipements terminaux + équipement terminal*	19	117	0.361	15.06	0.70
4255	atmosphère	clair	*atmosphère claire*	12	116	0.693	16.42	0.56
1193	groupe	secondaire		22	114	0.304	14.95	1.50
1794	courant	continu	*courant continu*	18	114	0.350	15.22	2.17
889	filtre	passe-bande	*filtre passe-bande + filtres passe-bande*	15	113	0.455	15.41	0.23
3736	système	national		39	111	0.197	14.43	2.55
3160	polarisation	orthogonal	*polarisations orthogonales + polarisation orthogonale*	19	109	0.322	14.75	0.87
84	circuit	fictif	*circuit fictif + circuits fictifs*	20	108	0.303	14.64	0.78
3466	onde	progressif	*ondes progressives*	19	104	0.314	14.81	1.68
2606	section	spécial	*sections spéciales + section spéciale*	14	101	0.415	15.31	2.32
3615	réflecteur	principal	*réflecteur principal*	22	100	0.235	14.40	3.20
2087	transmission	numérique		38	97	0.151	14.01	3.66
1586	station	distant	*stations distantes + station distante*	44	96	0.160	13.80	1.35
700	fréquence	intermédiaire	*fréquence intermédiaire + fréquences intermédiaires*	21	96	0.243	14.10	1.48
678	exemple	typique	*exemples typiques + exemple typique*	17	95	0.286	14.57	2.54
1581	destination	multiple	*destinations multiples + destination multiple*	20	91	0.178	13.90	1.74
2246	utilisateur	final	*utilisateurs finals + utilisateur final*	12	88	0.415	15.08	1.84
1780	codage	correcteur	*codage correcteur + codages correcteurs*	16	86	0.272	14.26	1.63
877	refroidissement	thermoélectrique	*refroidissement thermoélectrique*	9	85	0.606	15.64	0.33
3135	organisation	international	*organisations internationales + organisation internationale*	18	84	0.180	13.79	3.28

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
1662	heure	actuel	*heure actuelle*	12	80	0.304	14.46	2.77
2040	code	correcteur	*code correcteur + codes correcteurs*	16	78	0.237	13.87	1.63
3610	règle	général	*règle générale*	14	78	0.162	13.57	3.58
3518	énergie	électrique	*énergie électrique*	14	78	0.257	14.06	2.44
679	satellite	géostationnaire		18	78	0.211	13.56	0.69
4372	densité	spectral	*densité spectrale + densités spectrales*	11	76	0.317	14.44	2.00
315	affaiblissement	atmosphérique	*affaiblissements atmosphériques + affaiblissement atmosphérique*	10	76	0.385	14.74	1.71
2583	fréquence	porteur	*fréquences porteuses + fréquence porteuse*	17	73	0.201	13.33	1.44
4172	compatibilité	technique	*compatibilité technique*	13	72	0.196	13.67	3.00
634	distance	minimal	*distance minimale*	12	71	0.264	13.96	2.50
2724	mode	continu		14	70	0.207	13.58	2.17
2706	bande	étroit		15	67	0.201	13.23	1.12
3842	état	solide	*état solide*	8	66	0.444	14.76	1.07
269	conduit	numérique	*conduit numérique + conduits numériques*	15	66	0.066	12.49	3.66
1409	émission	brouilleuse	*émission brouilleuse*	8	65	0.416	14.66	0.88
4128	câble	coaxial	*câble coaxial + câbles coaxiaux*	8	65	0.426	14.61	0.82
1130	circuit	téléphonique	*circuits téléphoniques + circuit téléphonique*	29	65	0.120	13.03	2.24
2639	transmission	analogique		20	65	0.144	12.99	3.16
2860	position	orbital	*position orbitale + positions orbitales*	9	64	0.325	14.17	1.55
896	gain	nominal	*gain nominal*	11	63	0.244	13.61	2.16
2628	limiteur	progressif	*limiteur progressif*	9	63	0.257	13.90	1.68
931	élément	rayonnant	*élément rayonnant + éléments rayonnants*	9	63	0.316	13.80	0.60
2156	polarisation	elliptique	*polarisation elliptique*	11	59	0.220	13.03	0.95
2053	courant	alternatif	*courant alternatif*	7	57	0.387	14.08	0.38
2660	faisceau	étroit		12	57	0.190	12.92	1.12
4063	qualité	subjectif		9	56	0.256	13.42	1.87
161	surface	équivalent	*surface équivalente + surfaces équivalentes*	8	55	0.284	13.73	1.91
2921	réseau	public		12	55	0.177	12.51	0.79
1595	station	central		40	55	0.118	12.89	2.05
1083	panneau	solaire	*panneaux solaires + panneau solaire*	8	54	0.244	13.56	2.36
1132	moment	cinétique	*moment cinétique*	5	54	0.629	15.16	0.45
3531	transmission	télévisuel	*transmissions télévisuelles + transmission télévisuelle*	11	54	0.193	12.77	1.38
2139	enveloppe	constant	*enveloppe constante*	6	53	0.378	14.31	2.01
1847	niveau	admissible	*niveaux admissibles + niveau admissible*	10	51	0.197	12.68	1.27
4083	signal	vidéo	*signaux vidéo + signal vidéo*	14	51	0.148	12.28	2.12
1579	porteur	multiple	*porteuses multiples + porteuse multiple*	20	51	0.104	12.35	1.74
1711	question	relatif	*questions relatives*	9	51	0.129	12.73	**3.43**
3007	poursuite	automatique	*poursuite automatique*	8	51	0.163	13.02	2.48
1009	onde	triangulaire	*onde triangulaire*	8	51	0.247	13.00	0.76
2120	charge	essentiel		9	50	0.195	12.92	2.49
2471	algorithme	probabiliste	*algorithmes probabilistes + algorithme probabiliste*	6	49	0.373	13.99	0.90
4465	efficacité	spectral	*efficacité spectrale*	7	49	0.206	13.25	2.00
4361	schéma	fonctionnel	*schéma fonctionnel + schémas fonctionnels*	8	49	0.224	13.05	1.96
3334	antenne	isotrope	*antenne isotrope + antennes isotropes*	8	48	0.232	12.91	0.88
3583	ouverture	angulaire	*ouverture angulaire*	8	48	0.186	12.95	2.37
3863	système	mondial	*système mondial + systèmes mondiaux*	14	48	0.136	12.08	2.09
1980	signal	transmis		12	48	0.148	12.06	1.36
2191	plan	équatorial	*plan équatorial*	6	47	0.344	13.70	1.00
1641	radiocommunication	spatial	*radiocommunication spatiale*	11	47	0.045	11.85	2.21
3922	signal	reçu	*signal reçu*	10	47	0.163	12.03	0.79
3062	décodage	séquentiel	*décodage séquentiel*	6	47	0.335	13.63	1.00
3507	dimension	moyen	*dimensions moyennes + dimension moyenne*	8	46	0.134	12.62	3.05

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
2687	assignation	préalable	*assignation préalable*	5	45	0.412	14.16	1.15
2136	occupation	spectral		6	45	0.184	13.10	2.00
1796	service	mobile	*service mobile + services mobiles*	12	44	0.136	11.86	1.84
2627	fréquence	central	*fréquence centrale + fréquences centrales*	18	44	0.106	11.96	2.05
3193	longueur	variable		6	44	0.254	13.34	2.32
1011	espacement	angulaire	*espacements angulaires + espacement angulaire*	9	44	0.152	12.34	2.37
4157	polynôme	générateur	*polynôme générateur*	4	44	0.576	14.72	0.50
2722	réseau	national		21	43	0.097	11.99	2.55
745	caractéristique	technique	*caractéristiques techniques*	14	43	0.106	11.88	3.00
2712	segment	spatial	*segment spatial*	10	42	0.029	11.54	2.21
3360	formule	suivant		10	42	0.087	11.91	**3.80**
1049	onde	lent	*ondes lentes*	7	42	0.208	12.42	1.03
3694	bande	latéral	*bande latérale + bandes latérales*	13	42	0.115	11.79	0.72
3289	gestion	opérationnel	*gestion opérationnelle*	6	41	0.202	12.89	2.68
458	fréquence	supérieur	*fréquences supérieures + fréquence supérieure*	14	41	0.110	11.66	2.99
4125	système	existant	*systèmes existants + système existant*	11	40	0.126	11.55	1.69
2720	régulation	thermique	*régulation thermique*	7	40	0.108	12.20	1.92
2943	réseau	terrestre	*réseaux terrestres + réseau terrestre*	12	40	0.118	11.60	2.40
1697	charge	utile	*charge utile + charges utiles*	7	40	0.175	12.35	1.95
1284	onde	radioélectrique	*onde radioélectrique + ondes radioélectriques*	11	40	0.105	11.70	2.64
2891	détection	cohérent	*détection cohérente*	5	39	0.189	12.99	1.44
2192	rapport	axial	*rapport axial*	5	39	0.303	13.13	0.60
724	valeur	typique	*valeur typique + valeurs typiques*	12	39	0.107	11.62	2.54
2434	température	ambiant	*température ambiante + températures ambiantes*	5	39	0.291	12.94	0.45
397	mot	unique	*mot unique + mots uniques*	7	39	0.122	12.18	3.09
3864	république	fédéral	*république fédérale*	5	39	0.286	12.89	0.45
882	république	socialiste	*république socialiste + républiques socialistes*	5	39	0.286	12.89	0.45
1700	alimentation	périscopique	*alimentation périscopique*	6	39	0.219	12.39	1.00
2488	température	physique	*température physique*	7	39	0.160	12.22	2.34
1280	traitement	numérique	*traitement numérique*	14	38	0.040	11.34	3.66
328	cas	particulier	*cas particulier + cas particuliers*	9	38	0.106	11.76	**3.47**
2058	téléphonie	rural		6	38	0.189	12.57	1.84
3375	représentation	schématique	*représentation schématique*	4	38	0.430	13.87	0.50
1549	alimentation	électrique	*alimentation électrique*	9	38	0.119	11.76	2.44
680	satellite	national		18	38	0.086	11.61	2.55
1296	formule	approximatif	*formule approximative*	5	37	0.264	12.80	0.90

Valeur indéfinie du critère de vraisemblance

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
2178	absorption	atmosphérique	*absorption atmosphérique*	5	∞	0.233	13.42	1.71
1927	accroissement	apparent	*accroissement apparent*	2	∞	0.279	13.04	1.33
2923	administration	notificatrice	*administration notificatrice*	3	∞	0.390	13.21	0.00
925	agent	local	*agents locaux + agent local*	7	∞	0.070	12.28	2.94
323	ailette	métallique	*ailettes métalliques + ailette métallique*	2	∞	0.354	13.36	1.04
922	aimant	permanent	*aimant permanent + aimants permanents*	7	∞	0.351	14.39	2.36
1560	alignement	plésiochrone	*alignement plésiochrone*	2	∞	0.146	12.36	1.73
3003	alimentation	frontal	*alimentation frontale*	6	∞	0.284	12.98	0.00
1625	alimentation	ininterruptible	*alimentation ininterruptible*	9	∞	0.365	14.15	0.00
3199	amplificateur	excitateur	*amplificateur excitateur*	3	∞	0.208	11.40	0.00
1029	antenne	imparfait		2	∞	0.129	9.72	0.00
2208	article	manquant	*articles manquants*	2	∞	0.457	13.36	0.00
4353	assemblée	plénier	*assemblée plénière*	2	∞	0.646	14.36	0.00
1516	atop	seul	*atop seul*	4	∞	0.566	14.78	0.87
2197	attention	particulier	*attention particulière*	3	∞	-0.074	10.51	3.47
3827	autorité	responsable	*autorités responsables*	2	∞	0.224	12.78	1.56
3668	béton	armé	*béton armé*	2	∞	0.646	14.36	0.00
4059	bande	appariées	*bandes appariées*	3	∞	0.113	9.64	0.00
2410	bande	interdit	*bandes interdites*	2	∞	0.084	8.47	0.00
1490	bande	passant	*bande passante + bandes passantes*	10	∞	0.244	13.11	0.00
41	bobine	électromagnétique	*bobines électromagnétiques + bobine électromagnétique*	4	∞	0.417	14.19	0.69
2401	brouillage	potentiel	*brouillages potentiels + brouillage potentiel*	2	∞	0.152	10.19	0.00
2353	bruit	cosmique	*bruit cosmique*	2	∞	0.093	8.76	0.00
2154	bruit	impulsif	*bruit impulsif*	9	∞	0.254	13.10	0.00
3912	câble	sous-marins	*câbles sous-marins*	4	∞	0.354	13.19	0.00
3698	câble	sous-marin	*câble sous-marin*	5	∞	0.409	13.84	0.00
3657	cavité	couplées	*cavités couplées*	2	∞	0.409	13.04	0.00
754	cellule	solaire	*cellules solaires*	5	∞	0.172	13.01	2.36
3580	circuit	interrupteur	*circuit interrupteur*	2	∞	0.083	8.43	0.00
220	circulaire	hebdomadaire	*circulaire hebdomadaire*	5	∞	0.614	15.01	0.00
4234	coût	modique	*coût modique*	2	∞	0.179	10.66	0.00
140	codage	adaptatif		2	∞	0.128	9.69	0.00
472	code	auto-orthogonaux	*codes auto-orthogonaux*	5	∞	0.212	11.94	0.00
4084	code	poinçonnés		3	∞	0.151	10.47	0.00
2334	combineurs	hybride	*combineurs hybrides*	2	∞	0.181	12.56	1.75
1713	compresseur-extenseur	syllabique	*compresseur-extenseur syllabique*	2	∞	0.094	12.04	1.47
2776	compte	tenu	*compte tenu*	23	∞	0.896	17.89	0.00
2046	concentration	numérique	*concentration numérique*	34	∞	0.211	14.95	3.66
560	conférence	administratif	*conférence administrative*	4	∞	0.353	13.90	1.77
2939	consultation	officiel	*consultations officielles*	2	∞	0.346	12.56	0.00
3192	consultation	officieux	*consultations officieuses*	2	∞	0.346	12.56	0.00
3991	cornet	cannelé	*cornet cannelé*	2	∞	0.254	11.66	0.00
1672	couple	perturbateur	*couple perturbateur + couples perturbateurs*	7	∞	0.715	15.81	0.00
3376	courrier	électronique	*courrier électronique*	3	∞	0.144	12.53	1.89
1688	couverture	hémisphérique		6	∞	0.382	13.83	0.00
3369	couverture	zonale		5	∞	0.340	13.31	0.00
2821	décision	ferme	*décisions fermes*	2	∞	0.199	10.97	0.00
3595	déflection	mécanique	*déflection mécanique*	2	∞	-0.091	10.50	2.98
4325	description	général	*description générale*	4	∞	-0.086	10.14	3.58
846	diaphonie	intelligible	*diaphonie intelligible*	2	∞	0.646	14.36	0.00
4201	directeur	général	*directeur général*	3	∞	-0.147	9.31	3.58
4071	disponibilité	accrue	*disponibilité accrue*	2	∞	0.276	11.90	0.00

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
2065	dissipation	thermique	*dissipation thermique*	2	∞	-0.177	9.36	1.92
2234	distorsion	intersymbole	*distorsion intersymbole*	6	∞	0.737	15.73	0.00
2400	domaine	fréquentiel	*domaine fréquentiel*	2	∞	0.289	12.04	0.00
1061	éclaircissement	nécessaire	*éclaircissements nécessaires*	2	∞	-0.185	9.23	**3.82**
2430	égaliseur	accordable		2	∞	0.528	13.78	0.00
849	embrouilleur	pseudo-aléatoire	*embrouilleur pseudo-aléatoire*	3	∞	0.234	13.07	1.50
2887	engin	spatial	*engins spatiaux + engin spatial*	57	∞	0.392	16.94	2.21
4453	étude	poussées	*études poussées*	2	∞	0.236	11.46	0.00
2511	événement	sportif	*événements sportifs*	2	∞	0.346	12.56	0.00
327	fibre	optique	*fibres optiques + fibre optique*	9	∞	0.608	15.80	1.20
3612	focalisation	magnétique	*focalisation magnétique*	2	∞	0.181	12.56	0.96
2909	fonctionnement	anormal	*fonctionnement anormal*	2	∞	0.115	9.39	0.00
3697	génie	civil	*génie civil*	6	∞	0.612	15.36	1.00
1198	gradient	thermique	*gradient thermique + gradients thermiques*	2	∞	-0.177	9.36	1.92
4435	groupe	électrogène	*groupe électrogène + groupes électrogènes*	2	∞	0.120	9.50	0.00
2378	hélium	gazeux	*hélium gazeux*	3	∞	0.711	14.95	0.00
4365	île	salomon	*îles salomon*	4	∞	0.750	15.36	0.00
3926	interconnectivité	total	*interconnectivité totale*	2	∞	-0.203	8.90	3.06
2554	interrogation	préalable	*interrogation préalable*	2	∞	0.118	12.19	1.15
1274	invar	mince	*invar mince*	2	∞	0.463	13.78	0.64
3597	isolation	phonique	*isolation phonique*	2	∞	0.528	13.78	0.00
3320	langue	maternel	*langue maternelle*	2	∞	0.409	13.04	0.00
403	liaison	hétérodyne	*liaison hétérodyne*	2	∞	0.055	7.25	0.00
2280	liste	complet	*liste complète*	2	∞	-0.059	10.84	2.60
770	logique	combinatoire	*logique combinatoire*	2	∞	0.457	13.36	0.00
1620	mélangeur	double		3	∞	0.289	13.36	1.68
1978	microstations	distant	*microstations distantes*	2	∞	-0.178	9.34	1.35
2045	mission	multiple	*missions multiples*	2	∞	-0.249	7.84	1.74
3277	mode	semi-continu		2	∞	0.148	10.11	0.00
2651	module	défectueux	*modules défectueux*	2	∞	0.323	12.36	0.00
1372	mois	quelconque	*mois quelconque*	12	∞	0.548	15.89	1.92
1696	monture	polaire	*monture polaire + montures polaires*	2	∞	0.354	13.36	1.04
1661	navette	spatial	*navette spatiale*	5	∞	-0.088	9.92	2.21
3392	niveau	hiérarchique	*niveaux hiérarchiques + niveau hiérarchique*	4	∞	0.159	10.89	0.00
1351	note	relatif	*notes relatives*	2	∞	-0.202	8.92	**3.43**
4193	observation	général	*observations générales*	15	∞	0.188	13.96	3.58
4204	océan	atlantique	*océan atlantique*	2	∞	0.457	13.36	0.00
4123	océan	indien	*océan indien*	2	∞	0.457	13.36	0.00
1537	octet	indicateur	*octet indicateur*	2	∞	0.354	13.36	0.69
2384	opération	réversible	*opération réversible*	2	∞	0.195	10.90	0.00
2678	oscillation	journalier	*oscillation journalière*	2	∞	0.224	12.78	1.56
3943	pôle	mécanique	*pôle mécanique*	2	∞	-0.091	10.50	2.98
2629	palier	central	*palier central*	2	∞	-0.206	8.84	2.05
2063	paragraphe	suivant	*paragraphes suivants*	4	∞	-0.067	10.47	**3.80**
2571	partage	interrégional	*partage interrégional*	3	∞	0.466	13.73	0.00
2094	partie	gauche	*partie gauche*	2	∞	0.148	10.11	0.00
1375	partie	intégrant	*partie intégrante*	8	∞	0.378	14.11	0.00
3258	personnel	compétent	*personnel compétent*	2	∞	0.210	11.11	0.00
3591	perte	ohmiques	*pertes ohmiques*	2	∞	0.210	11.11	0.00
970	point	nodal	*point nodal*	2	∞	0.222	11.28	0.00
3083	polariseur	quart	*polariseur quart*	2	∞	0.346	12.56	0.00
3480	pondération	psophométrique	*pondération psophométrique*	4	∞	0.566	14.78	0.64
4376	porteur	modulé	*porteuses modulées*	2	∞	0.097	8.90	0.00
346	préjudice	économique	*préjudice économique*	7	∞	0.302	14.12	2.64
2381	proportion	important		2	∞	-0.192	9.10	**3.57**
4219	publication	anticipé	*publication anticipée*	8	∞	0.823	16.36	0.00

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
3990	puissance	rayonné	*puissance rayonnée*	2	∞	0.086	8.54	0.00
1233	référence	bibliographique	*références bibliographiques*	3	∞	0.390	13.21	0.00
738	région	océanique	*régions océaniques*	2	∞	0.236	11.46	0.00
2487	répartiteur	numérique	*répartiteur numérique*	2	∞	-0.281	6.77	3.66
587	réponse	impulsionnelle	*réponse impulsionnelle*	8	∞	0.622	15.56	0.00
284	république	démocratique	*république démocratique*	8	∞	0.432	14.50	0.00
881	république	populaire	*république populaire*	9	∞	0.464	14.84	0.00
4446	réseau	maillés	*réseaux maillés*	5	∞	0.132	10.57	0.00
3820	rôle	prédominant	*rôle prédominant*	2	∞	0.289	12.04	0.00
598	rafraîchissement	conditionnel	*rafraîchissement conditionnel*	7	∞	0.811	16.17	0.00
2067	rdcp	classique	*rdcp classiques*	4	∞	0.074	12.11	2.75
461	séquence	clef	*séquence clef*	3	∞	0.318	12.63	0.00
3	satellite	artificiel	*satellite artificiel + satellites artificiels*	3	∞	0.104	9.39	0.00
1371	satellite	géostationnaires		26	∞	0.387	15.62	0.00
3010	secousse	sismique	*secousses sismiques*	2	∞	0.354	13.36	1.04
294	signal	discret	*signal discret*	3	∞	0.088	8.93	0.00
196	signal	original	*signal original + signaux originaux*	2	∞	0.066	7.76	0.00
3734	sous-systèmes	utilisateurs	*sous-systèmes utilisateurs*	3	∞	0.466	13.73	0.00
2992	soutien	logistique	*soutien logistique*	2	∞	0.646	14.36	0.00
16	srs	communautaire	*srs (communautaire*	2	∞	-0.166	9.53	2.93
3126	stabilisation	triaxiale	*stabilisation triaxiale*	7	∞	0.811	16.17	0.00
4065	stabilité	dimensionnel	*stabilité dimensionnelle*	2	∞	0.528	13.78	0.00
3029	station	brouilleuses	*stations brouilleuses*	4	∞	0.048	7.45	0.00
2201	symétrie	axial	*symétrie axiale*	2	∞	0.181	12.56	0.60
3751	système	sous-régionaux	*systèmes (nationaux et + régionaux ou) sous-régionaux*	2	∞	0.064	7.70	0.00
296	terme	dépendant	*terme dépendant*	5	∞	0.656	15.20	0.00
1508	tirant	parti	*tirant parti*	3	∞	0.577	14.53	0.56
1129	titre	indicatif	*titre indicatif*	3	∞	0.356	12.95	0.00
1145	total	disponible	*total disponible*	6	∞	0.118	12.67	3.11
268	trafic	sporadique		4	∞	0.212	11.72	0.00
3328	transaction	préliminaire	*transaction préliminaire*	2	∞	0.024	11.56	2.04
1684	transition	brutal	*transition brutale + transitions brutales*	2	∞	0.528	13.78	0.00
2051	transposition	régénérateur		2	∞	0.279	13.04	1.05
1707	travail	intérimaire	*travail intérimaire*	3	∞	0.466	13.73	0.00
1526	usage	exclusif	*usage exclusif*	2	∞	0.323	12.36	0.00
32	voisinage	immédiat	*voisinage immédiat*	2	∞	0.224	12.78	1.56
3643	vue	éclatée	*vue éclatée*	2	∞	0.409	13.04	0.00
3575	zone	hydrométéorologiques	*zones hydrométéorologiques*	4	∞	0.234	12.01	0.00
2524	zone	radioclimatiques	*zones radioclimatiques*	2	∞	0.143	10.01	0.00

C.2 Corpus LBC

N_1 (PREP (DET)) N_2

Valeurs décroissantes du critère de vraisemblance

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
1364	canal	sémaphore	*canaux sémaphores + canal sémaphore*	1188	5738	0.559	25.20	0.31
20898	accusé	réception	*(accusé + accusés) de réception*	592	3983	0.683	24.78	0.07
20092	système	signalisation		847	2417	0.309	23.05	2.41
16310	complément	étude	*complément d' (étude + études)*	245	1984	0.731	23.74	0.48
3280	point	sémaphore		679	1822	0.273	22.38	2.69
510	intervalle	temps		251	1781	0.593	23.19	0.98
10249	trame	sémaphore	*trame sémaphore + trames sémaphores*	354	1443	0.266	21.48	1.08
870	signal	fin		391	1407	0.262	21.43	4.26
20851	sou-système	utilisateur	*sous-système utilisateur*	195	1226	0.458	22.10	1.52
6	bout	bout	*bout (en + à) bout*	137	1155	0.640	22,58	0.66
16967	contrôle	continuité	*(contrôle + contrôles) de continuité*	171	1116	0.457	21.89	2.89
6985	taux	erreur		156	1092	0.468	21.90	1.08
9478	message	adresse		337	1068	0.234	20.94	3.83
15899	indicatif	pays		142	1038	0.524	22.04	2.07
19104	recommandation	série		123	1024	0.628	22.36	1.69
6403	temps	propagation	*temps de propagation*	146	1001	0.410	21.34	3.43
3173	objet	complément	*objet d' un complément*	120	987	0.566	22.00	2.75
19306	rapidité	modulation		117	976	0.596	22.17	0.80
12935	mise	oeuvre	*(mise + mises) en oeuvre*	130	973	0.479	21.62	2.50
19926	degré	distorsion		118	963	0.579	22.09	1.08
3016	centre	transit		203	916	0.281	20.76	2.63
1851	liaison	réserve	*(liaison + liaisons) de réserve*	167	912	0.327	20.89	2.43
18576	unité	signalisation	*(unité + unités) de signalisation*	320	903	0.179	20.28	1.94
14607	lancement	opération		110	875	0.522	21.73	0.77
14764	établissement	communication		184	859	0.279	20.65	2.04
19709	ligne	abonné		189	849	0.270	20.58	3.44
19808	liaison	donnée		256	832	0.209	20.29	2.43
4506	sémaphore	secours	*(sémaphore + sémaphores) de secours*	150	827	0.321	20.69	2.81
8732	section	connexion		172	826	0.264	20.47	1.99
8614	réception	signal		332	804	0.177	20.21	3.16
15106	flux	information	*flux d' (information + informations)*	150	769	0.257	20.25	1.25
10283	sou-couche	transaction		104	759	0.446	21.21	0.64
4155	route	sémaphore	*routes sémaphores + route sémaphore*	199	749	0.176	19.67	0.94
3278	groupe	circuit		193	737	0.205	19.93	2.52
16349	jeu	caractère		113	719	0.333	20.57	1.43
20181	numéro	séquence	*(numéro + numéros) de séquence*	140	711	0.280	20.24	3.09
7375	plan	numérotage		119	709	0.330	20.55	2.26
18593	abonné	demandeur	*(abonné + abonnés) demandeur*	142	708	0.267	20.21	1.87
5257	transport	message		147	708	0.208	19.73	0.94
13051	couche	liaison		136	704	0.276	20.25	2.04
19057	sou-système	transport	*sous-système transport*	125	703	0.314	20.44	1.52
10086	libération	garde	*libération de garde*	122	695	0.317	20.43	2.62
5760	enregistreur	localisation	*(enregistreur + enregistreurs) de localisation*	99	681	0.405	20.83	1.86
5821	bruit	salle	*bruit de salle*	75	669	0.537	21.20	2.58
14332	élément	information		193	667	0.192	19.70	3.01
7084	niveau	puissance		111	662	0.307	20.14	3.68
11781	gestion	réseau		185	661	0.194	19.68	2.90
13685	réseau	sémaphore	*réseau sémaphore + réseaux sémaphores*	283	638	0.149	19.56	2.97

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
7506	correction	erreur		84	593	0.332	20.21	0.77
15818	connexion	sémaphore		274	584	0.141	19.36	2.99
15219	onde	signalisation	*(onde + ondes) de signalisation*	165	575	0.133	18.80	0.88
17792	code	en-tête	*(code + codes) d' en-tête*	87	558	0.289	19.62	3.40
7234	station	navire		71	553	0.447	20.67	1.71
8715	émission	signal		177	535	0.145	18.94	2.79
7735	verrouillage	trame	*verrouillage de trame*	74	533	0.332	20.04	0.79
3279	faisceau	route		90	532	0.292	19.77	1.56
17446	suppresseurs	écho	*suppresseurs d' écho*	81	531	0.325	20.02	0.75
1032	faisceau	circuit		155	529	0.160	19.00	1.56
17055	abonné	télex		116	522	0.206	19.25	1.87
3582	code	point		150	520	0.173	19.04	3.40
16600	centre	départ		158	518	0.167	19.00	2.63
1329	bande	fréquence		91	511	0.245	19.47	2.29
15505	signal	libération		219	511	0.141	18.87	4.26
4923	fonction	commande		177	497	0.151	18.87	4.27
9052	récepteur	signal		123	491	0.143	18.59	1.51
7100	disposition	recommandation		81	483	0.264	19.49	2.19
13563	isolement	processeur		41	481	0.827	21.66	0.11
6540	distorsion	affaiblissement		73	471	0.312	19.71	2.83
514	faisceau	canal		119	467	0.172	18.80	1.56
17804	identité	ligne		104	463	0.179	18.81	2.25
5462	passage	canal		106	461	0.181	18.83	2.68
8198	réception	message		215	457	0.128	18.73	3.16
18166	valeur	paramètre		101	456	0.200	18.91	3.86
19863	numéro	référence		118	453	0.173	18.72	3.09
17805	tête	ligne		74	447	0.195	18.80	0.51
17445	état	repos		83	444	0.230	18.95	3.94
13964	transfert	information		151	443	0.141	18.54	2.67
21126	service	télétex		110	439	0.175	18.56	3.72
1633	document	commande	*(document + documents) de commande*	87	439	0.189	18.78	1.72
19106	réduction	écho		62	437	0.311	19.62	1.52
20275	sémaphore	donnée		160	437	0.136	18.51	2.81
19464	qualité	transmission		113	429	0.156	18.51	2.10
4456	étiquette	acheminement	*(étiquette + étiquettes) d' acheminement*	62	424	0.287	19.42	1.48
3653	mise	jour	*(mise + mises) à jour*	59	418	0.303	19.24	2.50
3749	force	son		38	416	0.690	21.08	0.23
977	signal	réponse		190	415	0.125	18.34	4.26
18438	centre	tête		70	411	0.228	18.65	2.63
20759	réseau	télex		127	407	0.145	18.33	2.97
16226	établissement	appel		135	406	0.133	18.28	2.04
20990	centre	arrivée		123	405	0.148	18.32	2.63
3543	signal	raccrochage	*(signal + signaux) de raccrochage*	101	397	0.139	17.72	4.26
881	reconnaissance	signal		103	396	0.120	17.99	1.69
8359	signal	invitation		92	393	0.140	17.59	4.26
19152	service	télex	*service télex + services télex*	142	392	0.133	18.20	3.72
17603	répertoire	caractère		66	388	0.214	18.77	1.59
15385	équipement	commutation		115	388	0.147	18.22	3.37
17602	caractère	commande	*(caractère + caractères) de commande*	102	383	0.147	18.21	3.46
18066	contrôle	flux	*contrôle de flux*	67	380	0.234	18.73	2.89
7797	surveillance	taux	*surveillance du taux*	46	378	0.417	19.90	2.33
10844	microphone	charbon	*(microphone + microphones) à charbon*	35	375	0.658	20.78	1.92
3235	personne	personne	*personne à personne*	37	363	0.512	20.28	0.45
5036	ordre	succession	*ordre de succession*	43	361	0.373	19.41	3.24
11376	chute	temporisation		40	359	0.383	19.67	0.22
8411	fin	numérotation		96	353	0.146	18.01	3.81
6638	partage	charge		44	352	0.345	19.47	1.28
15071	mode	connexion		114	351	0.126	17.90	3.08
16537	propagation	groupe	*propagation (de + du) groupe*	53	351	0.225	18.68	1.06
16908	polarité	arrêt		45	347	0.356	19.47	1.42
4302	point	vue	*(point + points) de vue*	75	347	0.164	17.81	2.69

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
9707	appel	instance	*(appel + appels) en instance*	49	346	0.264	18.59	4.01
4807	tentative	appel		86	346	0.130	17.85	2.03
11600	fin	inhibition		62	344	0.214	18.36	3.81
9785	ligne	exploration		59	341	0.218	18.33	3.44
2773	liaison	satellite		79	336	0.162	17.91	2.43
7501	service	télématique		78	335	0.159	17.81	3.72
15045	signal	confirmation		96	335	0.124	17.34	4.26
19478	essai	continuité		71	333	0.177	18.10	3.82
7989	refus	connexion		65	331	0.145	17.91	1.31
11716	sou-couche	composant	*sous-couche composant*	52	330	0.251	18.73	0.64
14308	adaptateur	jonction	*(adaptateur + adaptateurs) de jonctions*	36	329	0.439	19.84	0.63
21932	extrémité	départ		91	329	0.134	17.78	2.65
19007	syntaxe	donnée		54	327	0.129	17.65	0.09
2137	élément	image	*(élément + éléments) d' image*	63	327	0.192	18.09	3.01
15155	établissement	connexion		108	325	0.119	17.68	2.04
5673	appareil	mesure		67	324	0.174	18.08	2.81
2281	mode	fonctionnement		98	323	0.126	17.70	3.08
12146	mise	page	*mise en page*	61	321	0.197	18.15	2.50
1698	trafic	sémaphore	*trafic sémaphore*	134	319	0.094	17.54	2.59
2859	chiffre	langue	*(chiffre + chiffres) de langue*	41	317	0.337	19.10	2.31
9583	utilisateur	sou-couche		42	315	0.324	19.05	2.72
17132	extrémité	arrivée	*extrémité d' arrivée*	81	310	0.137	17.68	2.65
7364	service	support		85	310	0.137	17.55	3.72
11022	extension	code	*extension (du + de) code*	48	308	0.206	18.33	1.65
5404	message	demande		103	306	0.120	17.42	3.83
1124	signal	adresse	*(signal + signaux) d' adresse*	207	299	0.100	17.87	4.26
19308	transition	état		56	298	0.164	17.90	1.85
4205	télécopie	groupe		45	297	0.198	18.22	1.23
504	encombrement	faisceau		53	296	0.197	18.12	2.49
13809	dispositif	réduction	*(dispositif + dispositifs) de réduction*	42	294	0.270	18.49	3.68
16732	centre	commutation		99	294	0.117	17.41	2.63
8560	code	identification	*(code + codes) d' identification*	70	292	0.148	17.54	3.40
3387	point	transfert		130	292	0.105	17.44	2.69
7555	localisation	visiteur	*localisation pour visiteurs*	27	291	0.589	20.12	2.08
825	largeur	bande		41	291	0.256	18.62	1.31
13209	multiplexage	répartition	*multiplexage (par + à) répartition*	30	286	0.486	19.78	1.88
6523	signalisation	multifréquence	*(signalisation + signalisations) multifréquence*	88	285	0.115	17.03	4.51
17992	changement	état		55	285	0.156	17.74	1.97
15955	information	adresse		120	284	0.104	17.35	4.04
18122	télétex	base		44	283	0.182	18.01	1.31
16262	période	étude	*(période + périodes) d' études*	56	280	0.164	17.73	3.09
7504	détecteur	parole		41	278	0.234	18.37	1.65
21916	capacité	mémoire		42	274	0.242	18.24	3.40
7958	sens	transmission		65	273	0.116	17.27	2.73
21345	message	accusé		105	273	0.107	17.14	3.83
205	temps	transfert		94	270	0.108	17.17	3.43
21105	écho	départ		52	268	0.137	17.46	1.34
15569	équipement	protection	*(équipement + équipements) de protection*	57	267	0.155	17.36	3.37
4279	envoi	signal		99	266	0.086	17.01	2.40
12095	interface	usager-réseau	*(interface + interfaces) usager-réseau*	45	264	0.204	17.92	3.26
318	mot	code		40	262	0.182	17.89	1.00
9654	modulation	fréquence		55	260	0.137	17.38	2.18
21853	action	entité		42	260	0.217	18.02	3.06
8522	signal	signalisation	*signaux de signalisation*	2	259	-0.008	-4.51	4.26
20255	temporisateur	inactivité	*temporisateur d' inactivité*	26	256	0.469	19.59	1.00
3502	tableau	code		45	253	0.162	17.60	2.32
6593	retour	canal		68	251	0.110	17.02	2.60
143	classe	trafic	*classe de trafic*	52	249	0.138	17.30	2.10
1012	perte	synchronisme		33	249	0.293	18.46	2.73
8392	transmission	signal		139	246	0.083	17.11	3.63
48	usager	usager	*usager à usager*	61	246	0.121	17.07	3.23

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
10102	modulation	amplitude		32	245	0.297	18.46	2.18
5716	appareil	essai		67	243	0.107	16.93	2.81
14653	retard	formation		26	242	0.432	19.27	2.26
5721	résultat	essai		58	240	0.112	16.97	2.73
142	catégorie	demandeur		53	239	0.122	17.07	2.73
3153	nombre	chiffre		63	239	0.123	16.91	**4.21**
18225	point	repère		43	238	0.144	16.66	2.69
17559	taille	fenêtre		26	238	0.412	19.19	1.87
19214	chiffre	numéro		46	238	0.157	17.37	2.31
19021	transmission	donnée		106	237	0.088	16.86	3.63
5677	circuit	conversation		73	237	0.110	16.79	4.15
21339	signal	accusé		122	236	0.091	16.84	4.26
9965	fréquence	signalisation		125	235	0.071	16.83	3.13
19386	schéma	codage	*(schéma + schémas) de codage*	38	234	0.176	17.62	1.92
10132	commutation	paquet		39	232	0.194	17.60	2.72
5845	circuit	support	*circuit support + circuits supports*	66	232	0.115	16.76	4.15
17762	impédance	entrée		39	231	0.173	17.54	2.36
10631	opérateur	assistance	*(opératrice + opératrices) d' assistance*	26	230	0.337	18.46	2.64
5414	trajet	conversation	*(trajet + trajets) de conversation*	47	227	0.141	17.10	2.73
8929	signal	prise	*(signal + signaux) de prise*	89	226	0.093	16.46	4.26
9585	cas	utilisation		54	223	0.126	16.71	**4.63**
10767	utilisateur	téléphonie	*utilisateur téléphonie*	28	222	0.293	18.20	2.72
21857	qualité	fonctionnement		66	222	0.098	16.65	2.10
7645	service	vidéotex	*(services + service) vidéotex*	53	222	0.126	16.64	3.72
10860	sémaphore	message		122	222	0.080	16.81	2.81
2389	valeur	défaut		40	221	0.170	17.14	3.86
589	perte	verrouillage	*perte (de + du) verrouillage*	34	220	0.212	17.70	2.73
3424	affaiblissement	sonie	*affaiblissement (en + de) sonie*	36	218	0.193	17.44	2.81
2671	interfonctionnement	télématique	*interfonctionnement télématique*	36	218	0.187	17.49	2.60
1731	liaison	sémaphore	*liaison sémaphore + liaisons sémaphores*	167	218	0.077	17.12	2.43
15610	voie	télégraphie		48	218	0.136	16.79	3.55
7226	station	base	*(station + stations) de base*	46	218	0.123	16.90	1.71
16643	connexion	liaison		82	217	0.092	16.56	2.99
17801	élément	arrêt		45	215	0.142	16.82	3.01
13393	zone	localisation	*(zone + zones) de localisation*	37	214	0.170	17.32	2.57
11439	commutation	circuit		71	214	0.080	16.43	2.72
5498	message	blocage	*(message + messages) de blocage*	100	213	0.088	16.57	3.83
8165	transfert	message		115	213	0.078	16.68	2.67
12077	cycle	répétition	*(cycle + cycles) de répétition*	26	212	0.311	18.42	2.19
2597	caractéristique	interface		53	211	0.117	16.65	4.07
19136	expiration	délai		26	211	0.301	18.37	1.31
19735	tonalité	retour		35	210	0.181	17.35	2.73
99	charge	trafic		41	209	0.124	16.88	2.16
17769	durée	reconnaissance		37	209	0.171	17.08	3.67
6269	répétition	tentative		27	208	0.268	18.12	2.29
3146	noeud	origine		40	206	0.136	16.94	2.12
7951	équipement	signalisation		177	206	0.073	17.10	3.37
125	classe	protocole		39	206	0.145	16.99	2.10
11907	échec	appel		52	203	0.080	16.32	1.99
6199	progression	appel		38	202	0.074	16.32	0.79
12442	diagramme	transition		36	201	0.154	17.05	2.44
17338	télex	demandeur		49	200	0.100	16.47	2.76
16354	forme	impulsion		37	199	0.152	16.94	3.70
13266	chiffre	discrimination		26	198	0.259	17.76	2.31
11374	opération	classe	*(opération + opérations) de classe*	32	198	0.184	17.30	3.48
2341	écoulement	trafic	*écoulement (du + en) trafic*	30	197	0.128	16.97	0.93
21856	polarité	départ		43	196	0.099	16.49	1.42
12378	inhibition	canal		37	196	0.097	16.55	1.42
17678	domaine	paramètre		46	196	0.113	16.52	2.92
19430	signal	numéro	*(signal + signaux) de numéro*	97	196	0.083	16.26	4.26

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
19461	qualité	service		74	195	0.075	16.22	2.10
11890	régulation	surcharge		25	195	0.265	18.00	1.41
13094	inversion	lettre	*(inversion + inversions) lettres*	23	194	0.315	18.33	1.60
18689	numéro	ordre	*(numéro + numéros) d' ordre*	45	193	0.122	16.43	3.09
1691	séquence	échappement	*(séquence + séquences) d' échappement*	30	192	0.183	16.94	3.72
2143	mise	phase	*mise en phase*	38	191	0.142	16.63	2.50
3046	analyse	chiffre	*analyse (des + de) chiffres*	33	191	0.152	16.99	2.84
5583	salve	parole		23	191	0.196	17.60	0.33
8358	signal	occupation		71	190	0.086	15.92	4.26
3515	personnel	maintenance		29	189	0.159	17.16	1.50
21543	entité	couche		41	188	0.116	16.50	2.60
2893	diagramme	flux		31	188	0.172	17.09	2.44
403	blocage	faute		29	188	0.194	17.13	3.26
15143	identificateur	transaction		43	187	0.115	16.42	3.05
18670	indication	état		58	187	0.089	16.17	**4.02**
14908	mise	relation	*mise en relation*	34	187	0.155	16.69	2.50
11047	message	libération	*(message + messages) de libération*	96	186	0.080	16.25	3.83
8221	expiration	temporisation		27	185	0.188	17.34	1.31
14705	terminal	navire	*terminal (de + du) navire*	32	185	0.161	16.91	3.44
12044	tiers	octave		15	182	0.625	19.67	0.23
7321	fonction	gestion		78	181	0.081	16.09	4.27
4774	registre	emplacement	*(registre + registres) d' emplacement*	16	181	0.563	19.44	0.80
14713	identificateur	lancement	*(identificateur + identificateurs) de lancement*	31	181	0.165	16.81	3.05
7057	détection	erreur		40	181	0.105	16.34	3.00
4377	circuit	satellite		53	181	0.098	16.04	4.15
14336	intervention	opérateur		24	181	0.211	17.56	1.75
4536	pays	destination		42	181	0.100	16.27	2.45
21753	entité	gestion		44	180	0.097	16.22	2.60
20404	facteur	activité		22	180	0.284	17.98	2.23
9716	référence	appel	*(référence + références) d' appel*	61	178	0.071	15.95	3.09
14122	noeud	destination		39	177	0.101	16.28	2.12
11942	traitement	appel		62	177	0.070	15.94	3.11
7165	intégration	service		32	177	0.053	15.87	0.39
856	code	verrouillage	*(code + codes) de verrouillage*	41	177	0.116	16.14	3.40
19248	enregistreur	arrivée		47	177	0.090	16.08	1.86
344	commande	quartz	*commande à quartz*	28	176	0.160	16.43	4.26
4556	db	rapport	*db par rapport*	25	176	0.198	17.34	1.62
2654	mise	garde	*mise en garde*	47	176	0.099	16.08	2.50
8025	demande	service		85	175	0.069	16.04	3.70
13821	renvoi	appel		47	174	0.068	15.88	2.08
13759	jour	localisation		24	174	0.180	17.26	1.31
15673	niveau	onde		45	174	0.104	16.02	3.68
4778	indication	alarme		35	174	0.132	16.36	**4.02**
6460	affaiblissement	transmission		57	173	0.073	15.90	2.81
15115	enregistreur	départ		50	173	0.082	15.95	1.86
16468	module	coopération	*(module + modules) de coopération*	17	172	0.440	18.82	1.44
10301	fonction	fréquence		79	172	0.076	15.97	4.27
2839	calcul	équivalent		23	171	0.225	17.49	2.82
15598	libération	connexion		65	170	0.072	15.87	2.62
15582	mode	exploitation	*(mode + modes) d' exploitation*	46	169	0.097	15.96	3.08
19708	donnée	utilisateur		50	169	0.089	15.91	3.93
15599	identification	ligne		60	169	0.072	15.83	3.11
7503	niveau	bruit		50	168	0.092	15.87	3.68
2626	long	terme	*long terme*	18	168	0.339	18.34	1.48
16033	circuit	jonction		41	168	0.107	15.90	4.15
9608	champ	application	*(champ + champs) d'application*	38	168	0.106	16.13	2.77
3193	phase	établissement		34	168	0.112	16.29	2.86
4887	train	bit		25	168	0.159	16.96	1.58
7032	position	conversation	*position de conversation*	40	167	0.101	16.04	3.47
3886	anneau	garde	*anneau de garde*	23	166	0.144	16.92	1.00

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
3193	phase	établissement		34	168	0.112	16.29	2.86
4887	train	bit		25	168	0.159	16.96	1.58
7032	position	conversation	*position de conversation*	40	167	0.101	16.04	3.47
3886	anneau	garde	*anneau de garde*	23	166	0.144	16.92	1.00
2478	titre	exemple	*titre d' (exemple + exemples)*	19	163	0.304	17.90	2.52
2788	avis	remise		22	163	0.154	17.00	1.09
3178	point	relais	*point relais*	43	163	0.098	15.65	2.69
2523	procédure	transfert		82	163	0.072	15.88	4.34
6517	gestion	couche		50	163	0.086	15.78	2.90
15492	signal	intervention	*signal d' intervention*	46	162	0.085	15.25	4.26
2649	mémoire	tampon		16	161	0.420	18.55	2.54
15109	compensation	dérive		13	161	0.663	19.61	0.69
21323	entrée	récepteur		28	161	0.144	16.52	3.17
8468	signal	enregistreur		84	160	0.073	15.75	4.26
13764	centre	destination		65	160	0.076	15.73	2.63
4	procédure	commande		89	158	0.069	15.90	4.34
5708	bruit	circuit		58	158	0.060	15.59	2.58
14329	position	opérateur		32	158	0.121	16.13	3.47
8131	courant	fuite	*(courant + courants) de fuite*	17	158	0.335	17.95	2.78
12666	notification	remise		27	158	0.128	16.45	2.62
15506	confirmation	libération	*confirmation de libération*	38	156	0.086	15.82	2.15
14350	commutateur	origine		38	156	0.094	15.85	3.36
2154	relais	mise		25	156	0.137	16.59	1.84
2894	équivalent	référence	*(équivalent + équivalents) de référence*	28	156	0.093	16.14	1.48
4045	combinaison	multifréquence	*combinaison multifréquence*	29	155	0.131	16.27	3.31
16552	envoi	message		65	155	0.058	15.59	2.40
12352	composant	rejet		20	155	0.239	17.40	2.68
13774	confirmation	appel	*confirmation d' appel*	49	155	0.061	15.54	2.15
17949	sécurité	fonctionnement	*sécurité de fonctionnement*	28	154	0.073	15.92	1.46
849	trajet	effet		28	153	0.135	16.26	2.73
1304	gamme	fréquence		32	153	0.084	15.88	2.25
18968	distorsion	non-linéarité	*(distorsion + distorsions) de non-linéarité*	19	153	0.230	16.98	2.83
966	signal	encombrement	*(signal + signaux) d' encombrement*	82	153	0.071	15.64	4.26
14355	redémarrage	point		27	152	0.075	15.94	0.82
12602	libération	communication		52	152	0.072	15.56	2.62
11149	essai	validation	*(essai + essais) de validation*	29	152	0.131	16.04	3.82
21338	gestion	système		66	152	0.067	15.60	2.90
3122	point	référence	*(point + points) de référence*	79	152	0.070	15.70	2.69
19016	transfert	donnée		82	152	0.063	15.73	2.67
6536	affaiblissement	adaptation	*affaiblissement d' adaptation*	22	152	0.189	16.67	2.81
21802	connexion	accès	*(connexion + connexions) d' (accès + accès)*	56	151	0.076	15.54	2.99
12386	réinitialisation	bande	*réinitialisation de bande*	24	150	0.134	16.49	1.35
17599	sémaphore	état		61	149	0.071	15.52	2.81
21865	qualité	écoulement		20	149	0.195	16.57	2.10
76	réponse	demande		40	149	0.088	15.61	3.61
9838	feuille	papier	*(feuille + feuilles) de papier*	14	148	0.380	18.43	0.46
15680	gestionnaire	transaction	*gestionnaire de (transactions + transaction)*	45	148	0.083	15.51	3.01
17124	écho	arrivée		32	148	0.082	15.76	1.34
17042	numéro	abonné		59	147	0.070	15.49	3.09
4214	terminaison	dialogue		22	147	0.167	16.71	2.14
12100	régulation	trafic	*régulation de trafic*	28	147	0.088	15.92	1.41
5080	structure	trame		34	146	0.089	15.70	3.25
6023	section	section	*section par section*	29	146	0.120	15.94	1.99
8203	réception	déblocage	*réception de déblocage*	37	144	0.094	15.41	3.16
7581	niveau	sortie		43	144	0.084	15.43	3.68
21735	chaîne	circuit		41	144	0.049	15.28	1.84
13690	collision	front	*(collision + collisions) de front*	12	144	0.591	19.15	1.18
8393	signal	sélection	*(signal + signaux) de sélection*	70	143	0.070	15.35	4.26
21030	accès	service		66	143	0.058	15.44	3.63
14900	remise	message		46	143	0.048	15.24	2.27
11390	dispositif	protection	*(dispositif + dispositifs) de protection*	29	143	0.116	15.85	3.68
9207	source	bruit		24	143	0.118	16.21	2.69

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
6509	affaiblissement	équilibrage	*(affaiblissement + affaiblissements) d' équilibrage*	22	142	0.166	16.37	2.81
20886	position	caractère		42	142	0.073	15.41	3.47
6139	protocole	couche		36	142	0.084	15.52	3.25
4658	signalisation	enregistreur		73	141	0.069	15.41	4.51
21414	connexion	entrée	*(connexion + connexions) d' entrée*	48	141	0.077	15.34	2.99
14952	action	noeud	*(action + actions) au noeud*	22	140	0.161	16.38	3.06
876	affaiblissement	effet		27	139	0.119	15.86	2.81
4787	acheminement	appel		47	139	0.054	15.23	2.51
18469	indice	netteté	*(indice + indices) de netteté*	12	139	0.516	18.74	1.59
7505	voie	support		42	138	0.081	15.31	3.55
13323	service	réseau		96	138	0.062	15.74	3.72
150	déblocage	faute		17	138	0.238	17.27	1.30
8032	adaptation	impédance		15	138	0.281	17.71	1.46
10660	information	taxation		44	137	0.079	15.24	4.04
19944	efficacité	fonction	*efficacité en fonction*	31	137	0.083	15.53	2.52
7923	concentration	conversation		21	136	0.108	16.18	0.98
17419	intensité	trafic		21	136	0.078	15.91	0.85
9210	connexion	sortie		43	136	0.078	15.23	2.99
6387	traduction	appellation		32	135	0.047	15.20	2.02
601	transfert	charge		38	135	0.084	15.26	2.67
13503	pavillon	écouteur		14	134	0.303	17.84	0.90
737	effet	produit	*effet produit*	19	134	0.187	16.53	3.53
14895	message	remise		59	134	0.069	15.19	3.83
9254	nature	circuit		37	134	0.043	15.07	2.16
7206	service	base		65	134	0.065	15.30	3.72
11172	sémaphore	remplissage		25	133	0.119	15.50	2.81
20375	efficacité	émission	*(efficacités + efficacité) à l' émission*	32	133	0.074	15.38	2.52
10307	information	signalisation		169	133	0.060	16.52	4.04
16686	modification	appel	*modification d' appel*	38	132	0.048	15.09	2.84
20729	unité	conversion	*unité de conversion*	29	132	0.103	15.44	1.94
17687	état	ligne		69	132	0.059	15.32	3.94
21462	ordre	priorité	*(ordre + ordres) de priorité*	21	132	0.154	16.13	3.24
6896	db	niveau		25	132	0.084	15.67	1.62
3946	route	secours	*(route + routes) de secours*	31	132	0.089	15.42	0.94
7457	voie	transmission		70	132	0.059	15.32	3.55
18185	largeur	fenêtre		17	130	0.209	16.88	1.31
15274	sonie	émission	*sonie à l' émission*	27	130	0.075	15.47	2.05
6748	adresse	station		28	130	0.096	15.51	3.33
16983	bit	contrôle	*bits de contrôle*	33	130	0.079	15.29	3.68
3522	cas	défaillance	*cas de (défaillance + défaillances)*	38	130	0.081	15.10	**4.63**
16601	circuit	départ		69	129	0.061	15.28	4.15
11980	écart	type	*écart type*	24	129	0.072	15.55	1.13
3538	signal	échec		53	128	0.068	14.87	4.26
942	signal	blocage		100	128	0.063	15.61	4.26
20579	unité	enregistrement	*(unité + unités) d' enregistrement*	29	128	0.097	15.30	1.94
2386	signe	voyelle	*signe de voyelle*	12	127	0.408	18.08	2.23
9037	couche	transport	*couche transport*	34	127	0.077	15.19	2.04
904	transfert	signal		99	127	0.053	15.63	2.67
5902	prise	circuit		37	127	0.041	14.93	2.59
18770	délai	temporisation	*(délai + délais) de temporisation*	25	126	0.103	15.58	3.02
11223	erreur	procédure	*(erreur + erreurs) de procédure*	29	126	0.084	15.33	3.38
6938	taxe	base	*(taxe + taxes) de base*	19	126	0.067	15.72	0.79
9763	signalisation	ligne		134	126	0.063	16.07	4.51
2959	couleur	fond		11	125	0.466	18.58	0.66
13395	relation	sémaphore	*relations sémaphores + relation sémaphore*	46	125	0.028	14.76	2.23
11002	étage	commutation	*(étage + étages) de commutation*	21	124	0.064	15.53	0.97
20556	compatibilité	couche		18	124	0.088	15.94	1.08
13040	définition	terme		20	124	0.145	15.96	3.97
17377	délai	garde	*(délai + délais) de garde*	28	124	0.082	15.28	3.02
12119	couche	réseau		50	123	0.052	14.96	2.04
7372	service	transport	*(service + services) de transport*	52	123	0.065	14.98	3.72

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
16924	nom	paramètre		24	123	0.082	15.46	2.82
15315	non-réception	signal		30	123	0.014	14.61	0.89
11014	domaine	information		47	123	0.052	14.94	2.92
19217	pays	arrivée	*pays d' arrivée*	33	123	0.064	15.05	2.45
12541	télégraphie	modulation		18	123	0.114	16.11	1.79
6439	trajet	transmission		43	123	0.052	14.92	2.73
21527	objectif	qualité		20	122	0.119	15.92	2.96
11601	test	inhibition	*test d' inhibition*	19	122	0.137	16.09	2.44
12408	transmission	document		40	122	0.071	14.94	3.63
12354	temps	maintien		26	121	0.101	15.15	3.43
20068	modèle	référence	*(modèle + modèles) de référence*	28	121	0.066	15.14	2.72
8726	signalisation	voie		110	121	0.061	15.70	4.51
17206	durée	indisponibilité		22	121	0.121	15.49	3.67
18643	télex	destination		32	120	0.065	15.00	2.76
19689	seuil	audibilité	*(seuil + seuils) d' audibilité*	12	119	0.331	17.47	2.83
15778	inversion	chiffre	*inversion chiffres*	20	119	0.099	15.70	1.60
5485	acheminement	message		47	118	0.041	14.78	2.51
9421	identification	circuit		57	118	0.047	14.92	3.11
13692	niveau	décision	*(niveau + niveaux) de décision*	22	118	0.112	15.18	3.68
21234	spécification	système		41	118	0.053	14.82	3.39
15072	identificateur	composant		28	118	0.083	15.09	3.05
945	réponse	signal		73	118	0.046	15.11	3.61
8010	pays	origine	*pays d' origine*	28	117	0.073	15.06	2.45
4928	activation	fonction		24	117	0.072	15.21	1.64
14883	échange	information		36	117	0.045	14.76	2.37
835	signal	signal		7	116	-0.006	1.73	4.26
5525	distorsion	quantification	*distorsion de quantification*	16	116	0.175	16.05	2.83
19349	communication	télex	*communications télex + communication télex*	41	116	0.056	14.79	3.67
10395	dispositif	coupure	*(dispositif + dispositifs) de coupure*	20	116	0.126	15.53	3.68
13004	mode	paquet		28	116	0.086	14.97	3.08
21917	caractéristique	efficacité		24	115	0.101	15.15	4.07
4724	activation	canal		28	115	0.048	14.85	1.64
3903	trafic	transit		37	114	0.061	14.78	2.59
10100	processus	application	*processus d' application*	22	114	0.081	15.30	3.03
1819	combinaison	code		30	114	0.065	14.88	3.31
13300	trait	union	*(trait + traits) d' union*	10	113	0.467	18.38	1.32
21428	opérateur	arrivée	*(opératrice + opératrices) d' arrivée*	31	113	0.059	14.81	2.64
13337	disposition	exploitation		25	113	0.079	15.08	2.19
21407	altération	durée		13	113	0.157	16.64	0.60
8125	passage	contrainte	*passage sous contrainte*	16	113	0.164	15.85	2.68
10408	appel	cour/cours		29	113	0.080	14.84	4.01
5135	prolongement	appel		20	112	0.006	14.61	0.30
3945	groupe	usager		40	112	0.058	14.70	2.52
10078	coupure	ligne		29	112	0.042	14.71	2.25
9354	type	message		83	112	0.050	15.21	4.26
14854	point	destination		65	112	0.057	14.93	2.69
8705	message	progression		26	112	0.083	14.46	3.83
18799	relève	dérangement	*relève (des dérangements + du dérangement)*	12	112	0.166	16.75	0.27
21440	indicatif	abonné		36	111	0.055	14.68	2.07
8774	fournisseur	service		22	111	0.011	14.57	0.71
14560	détermination	équivalent		16	111	0.146	16.06	2.99
18407	période	polarité		19	110	0.122	15.49	3.09
4183	déblocage	groupe		21	110	0.062	15.12	1.30
7987	connexion	travers		28	110	0.081	14.69	2.99
6235	retour	appel		46	110	0.044	14.63	2.60
15948	interrogation	groupe	*interrogation de groupe*	15	110	0.045	15.39	0.23
14123	valeur	temporisateur	*valeur du temporisateur*	18	109	0.125	15.22	3.86
15022	détection	défaillance		21	109	0.095	15.21	3.00
7382	effet	distorsion		24	109	0.078	14.96	3.53
2372	changement	ligne	*(changement + changements) de ligne*	32	108	0.043	14.58	1.97

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
8540	signalisation	signal	*signalisation (du signal + par signaux)*	3	108	-0.008	-1.67	4.51
2426	valeur	affaiblissement		30	108	0.073	14.68	3.86
15145	identificateur	dialogue	*identificateur de dialogue*	23	108	0.091	15.00	3.05
12492	alimentation	énergie	*(alimentation + alimentations) en énergie*	11	108	0.326	17.51	2.04
1253	cas	interfonctionnement		39	108	0.064	14.57	**4.63**
9477	fourniture	service		24	108	0.018	14.50	1.46
4405	contrainte	route	*contrainte sur route*	15	107	0.108	15.89	1.63
3700	compte	fait		17	107	0.136	15.50	4.45
20229	durée	émission		38	106	0.056	14.56	3.67
21257	niveau	entrée		38	106	0.061	14.53	3.68
10045	application	procédure		25	105	0.070	14.77	3.75
18927	probabilité	dépassement	*probabilité de dépassement*	10	105	0.349	17.40	2.95
4983	commande	appel		68	105	0.048	14.87	4.26
337	objet	accord		19	105	0.112	15.16	2.75
2554	volume	trafic		19	104	0.052	15.00	1.84
9019	connexion	transport	*(connexion + connexions) de transport*	41	104	0.059	14.50	2.99
21065	facteur	efficacité	*facteur d' efficacité*	15	104	0.143	15.90	2.23
638	adresse	demandeur		33	104	0.051	14.49	3.33
20192	essai	compatibilité		21	104	0.097	14.87	3.82
11363	combinaison	marche	*(combinaison + combinaisons) de marche*	14	104	0.172	15.86	3.31
8607	usager	service		47	103	0.041	14.50	3.23
6863	signalisation	type		87	103	0.056	15.12	4.51
857	flux	trafic		30	102	0.054	14.49	1.25
12967	titre	option	*titre d' option*	13	102	0.193	16.30	2.52
10490	champ	information		40	102	0.042	14.41	2.77
7943	passage	position	*passage en position*	22	102	0.086	14.82	2.68
15615	zone	image		19	102	0.096	15.10	2.57
9835	cas	non-réception	*cas de non-réception*	18	102	0.103	14.61	**4.63**
8672	voie	radio	*(voies + voie) radio*	20	102	0.099	14.72	3.55
11687	réseau	télécommunication		26	101	0.077	14.47	2.97
21337	système	transmission		84	101	0.052	15.07	2.41
8802	attribution	code		22	101	0.058	14.70	2.91
15333	impulsion	comptage		14	100	0.158	15.89	3.02
15563	point	signalisation	*points de signalisation*	2	100	-0.012	-3.26	2.69
14485	identificateur	corrélation	*identificateur de corrélation*	13	100	0.164	15.54	3.05
13892	appel	provenance	*(appel + appels) en provenance*	24	100	0.079	14.56	4.01
4227	procédure	établissement		45	100	0.056	14.42	4.34
15886	détecteur	interruption	*détecteur d' interruption*	15	100	0.132	15.65	1.65
5627	programme	essai		18	100	0.022	14.64	1.15
9881	domaine	application	*(domaine + domaines) d' application*	27	100	0.062	14.48	2.92
20087	interconnexion	système	*interconnexion (des + de) systèmes*	23	100	0.033	14.45	2.08
1811	document	télécopie	*documents (par + de) télécopie*	19	100	0.091	14.99	1.72
16416	fanion	ouverture	*(fanion + fanions) d' ouverture*	9	99	0.411	17.82	1.72
9898	demande	déconnexion	*demande de déconnexion*	20	99	0.096	14.71	3.70
16672	transit	message		30	99	0.021	14.18	2.46
11210	procédure	passage		31	99	0.066	14.28	4.34
20845	objet	accusé		30	99	0.057	14.39	2.75
9796	interface	commutateur		28	99	0.057	14.42	3.26
3988	blocage	groupe		31	98	0.053	14.35	3.26
7955	réception	blocage	*réception de blocage*	51	98	0.052	14.47	3.16
6683	produit	intermodulation	*produits d' intermodulation*	9	98	0.397	17.77	2.09
20973	contrôle	validation	*(contrôle + contrôles) de validation*	21	98	0.088	14.64	2.89
9538	circuit	type		51	98	0.051	14.46	4.15
1971	écart	fréquence		20	97	0.042	14.60	1.13
836	temps	reconnaissance	*temps (de + après) reconnaissance*	24	97	0.077	14.39	3.43
3550	bobine	sonde	*bobine sonde*	10	96	0.203	16.81	0.69
7685	sortie	récepteur	*sortie du récepteur*	19	96	0.084	14.81	3.26
10022	procédure	redémarrage	*procédure de redémarrage*	21	96	0.084	14.28	4.34
17494	état	encombrement		34	95	0.057	14.22	3.94

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
15639	interruption	voie		26	95	0.041	14.27	2.89
3957	télécopieur	groupe		21	94	0.050	14.49	1.48
4110	poste	demandeur	*poste demandeur*	24	94	0.042	14.31	2.45
8437	succession	signal		26	94	0.001	13.89	1.45
2808	abandon	échantillon	*abandon d' (échantillon + échantillons)*	11	94	0.211	16.46	1.84
8692	voie	signalisation	*(voie + voies) de signalisation*	115	93	0.044	15.45	3.55
270	côté	usager	*côté usager*	24	93	0.046	14.28	2.62
5012	format	base	*(format + formats) de base*	29	93	0.049	14.19	3.27
8880	niveau	pression		20	93	0.085	14.36	3.68
8364	signal	message		4	93	-0.007	-0.29	4.26
20482	sou-système	application	*sous-système application*	29	92	0.056	14.19	1.52
19605	tonalité	numérotation	*(tonalité + tonalités) de numérotation*	26	92	0.047	14.21	2.73
9105	remarque	figure		11	92	0.139	16.13	1.24
17325	carte	crédit	*(carte + cartes) de crédit*	8	92	0.403	17.88	0.76
3812	nombre	circuit		72	92	0.044	14.72	**4.21**
3868	centre	origine		40	92	0.052	14.17	2.63
3867	nombre	octet		25	92	0.069	14.17	**4.21**
8362	signal	incitation	*(signal + signaux) d' incitation*	21	91	0.063	13.37	4.26
2618	procédure	alignement	*(procédure + procédures) d' alignement*	22	91	0.076	14.11	4.34
18938	indisponibilité	faisceau		18	91	0.065	14.65	2.61
11906	architecture	contenu	*architecture de contenu*	11	91	0.184	16.26	1.90
9800	commutateur	destination		29	91	0.048	14.13	3.36
757	blocage	organe	*blocage de l' organe*	15	91	0.117	14.96	3.26
11449	étude	titre	*étude au titre*	9	91	0.320	17.19	2.41
12835	tampon	retransmission		11	91	0.170	16.21	1.46
16879	état	sou-système		31	91	0.057	14.10	3.94
6258	activation	faisceau		18	91	0.064	14.63	1.64
5323	message	déblocage		32	91	0.060	13.96	3.83
4527	plupart	cas	*plupart des cas*	14	91	0.093	15.25	2.63
10865	message	sémaphore	*messages sémaphores*	2	91	-0.010	-3.14	3.83
3233	bit	poids	*(bit + bits) de poids*	13	91	0.143	15.26	3.68
12336	communication	référence		32	90	0.047	14.10	3.67
8186	message	supervision	*(message + messages) de supervision*	29	90	0.062	13.92	3.83
17724	séquence	caractère		37	90	0.047	14.11	3.72
18761	code	problème		16	90	0.099	14.39	3.40
10944	vérification	acheminement		20	90	0.055	14.39	2.99
8528	impédance	équilibrage	*(impédance + impédances) d' équilibrage*	13	90	0.138	15.45	2.36
14431	spécification	recommandation		27	90	0.050	14.11	3.39
8106	variation	équivalent		14	89	0.110	15.22	3.02
5854	domaine	diagnostic		13	89	0.139	15.17	2.92
17450	délai	attente		18	89	0.080	14.58	3.02
13935	action	centre		22	89	0.059	14.27	3.06
15723	tentative	établissement		23	89	0.055	14.22	2.03
16242	commutateur	transit		28	89	0.048	14.09	3.36
5179	ordre	passage		19	89	0.078	14.47	3.24
20278	entité	application		23	89	0.056	14.21	2.60
14320	opérateur	langue		15	89	0.107	14.93	2.64
12934	type	document		33	89	0.053	14.03	**4.26**
13416	épaisseur	trait		8	88	0.348	17.55	0.98
7585	puissance	bruit		20	88	0.060	14.32	3.06
3289	moment	établissement	*moment de l' établissement*	19	88	0.055	14.37	3.42
4085	messagerie	personne	*messagerie de personne*	31	88	0.059	13.86	3.85
12937	primitif	gestion		23	88	0.043	14.12	2.63
21688	point	accès		51	87	0.049	14.23	2.69
361	circuit	garde	*circuit de garde*	38	87	0.051	14.03	4.15
13653	zone	affichage	*zone d' affichage*	14	87	0.116	15.05	2.57
3525	échec	établissement		20	87	0.054	14.27	1.99
1911	phase	transfert		26	87	0.038	13.99	2.86
7649	sortie	voie		27	87	0.038	13.98	3.26

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
6374	cadre	question	*cadre de la question*	13	87	0.107	15.27	2.97
19625	rangée	chiffre	*rangée (des + de) chiffres*	11	87	0.054	15.30	0.43
19145	soumission	télex		13	87	-0.005	14.38	0.26
16081	réinitialisation	circuit		25	86	0.013	13.81	1.35
8808	concentration	parole		14	86	0.065	14.85	0.98
2716	cellule	affaiblissement		10	85	0.069	15.54	0.30
5318	orientation	message	*orientation des messages*	20	85	-0.009	13.71	1.36
11078	condition	libération		34	85	0.045	13.95	4.29
21806	capacité	terminal		22	85	0.059	14.09	3.40
18418	donnée	protocole		28	85	0.053	13.94	3.93
5454	code	destination	*(code + codes) de destination*	42	85	0.046	14.04	3.40
16796	onde	mesure	*onde de mesure*	24	85	0.049	14.01	0.88
19403	message	sou-système		42	85	0.051	13.98	3.83
2042	attribut	écran		10	85	0.203	16.10	3.00
4087	indice	force		9	85	0.251	16.67	1.59
7913	service	couche		43	84	0.049	14.02	3.72
16253	émetteur	indicatif		14	84	0.077	14.84	2.37
561	temps	réponse		40	84	0.046	13.98	3.43
7837	autorisation	transfert	*autorisation de transfert*	15	84	0.010	14.25	0.96
14902	transaction	origine		18	84	0.049	14.26	2.74
17882	entité	liaison		27	84	0.040	13.89	2.60
3592	point	origine		47	84	0.049	14.07	2.69
16554	primitif	service		30	84	0.024	13.79	2.63
3644	onde	pilote	*ondes pilotes + onde pilote*	12	83	0.138	15.10	0.88
5570	tentative	prise		19	83	0.063	14.18	2.03
19588	tonalité	occupation		17	83	0.072	14.34	2.73
5079	erreur	protocole	*(erreur + erreurs) de protocole*	21	83	0.054	14.03	3.38
15280	demande	modification	*demande de modification*	17	82	0.083	14.15	3.70
6655	quart	vitesse		8	82	0.241	16.88	0.89
16053	appareil	groupe		27	82	0.044	13.83	2.81
7340	fin	sélection	*fin de sélection*	26	82	0.054	13.82	3.81
13242	fonction	relais		25	82	0.060	13.74	4.27
21245	système	télégraphie		30	81	0.055	13.72	2.41
19189	série	recommandation		21	81	0.043	13.93	3.26
14109	verrouillage	multitrame	*verrouillage de multitrame*	9	81	0.220	16.01	0.79
12664	support	télégraphie		14	81	0.081	14.67	2.91
14976	signal	réinitialisation		37	81	0.051	13.65	4.26
6229	lettre	alphabet		9	81	0.197	16.32	1.64
14349	interdiction	transfert	*interdiction de transfert*	14	81	0.003	14.15	0.91
11330	rappel	opérateur		10	81	0.080	15.46	0.99
8887	niveau	signal		82	81	0.040	14.73	3.68
3646	méthode	calcul	*(méthode + méthodes) de calcul*	15	81	0.093	14.34	4.00
16156	opérateur	centre		20	81	0.052	13.98	2.64
16840	communication	cour/cours		20	80	0.062	13.95	3.67
4015	rétablissement	synchronisme	*rétablissement du synchronisme*	12	80	0.114	15.05	2.66
12286	longueur	plage		10	80	0.163	15.25	3.23
14396	point	initiateur	*point initiateur*	17	79	0.073	13.59	2.69
21454	poste	abonné		21	79	0.032	13.80	2.45
5435	parasite	bande	*parasites hors bande*	10	79	0.044	15.10	0.45
15864	description	étape		14	79	0.094	14.49	3.51
3437	défaillance	canal		25	79	0.029	13.68	2.91
8796	affectation	voie		17	79	0.018	13.90	2.04
2211	filet	espacement	*filet sans espacement*	8	79	0.168	16.36	0.50
16982	contrôle	inactivité		17	78	0.075	13.99	2.89
16785	gestion	route	*gestion des routes*	25	78	0.053	13.69	2.90
20346	mesure	durée		19	78	0.063	13.89	4.31
17537	continuité	circuit		25	78	0.012	13.54	2.53
15658	maintenance	réseau	*maintenance (du réseau + des réseaux)*	23	78	0.017	13.60	2.65
10135	commutation	liaison		27	78	0.037	13.66	2.72
5619	appel	essai		41	77	0.038	13.82	4.01
18241	caractère	mise	*(caractère + caractères) de mise*	23	77	0.051	13.71	3.46
10300	répartition	fréquence	*répartition (en fréquence + des fréquences)*	17	77	0.025	13.87	2.27

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
10291	échelle	notation	*(échelle + échelles) de notation*	7	77	0.358	17.26	2.06
700	fin	période		18	77	0.070	13.82	3.81
19940	efficacité	cabine		11	77	0.132	15.00	2.52
15005	imitation	signal		17	77	-0.037	13.19	0.59
11345	objet	étude		25	77	0.044	13.65	2.75
12415	dérive	fréquence		12	77	-0.005	14.16	0.72
21298	système	multiplexage		24	76	0.057	13.51	2.41
9423	méthode	mesure	*(méthode + méthodes) de mesure*	27	76	0.044	13.63	4.00
7606	effet	bruit		19	76	0.050	13.82	3.53
3731	interfonctionnement	vidéotex	*interfonctionnement vidéotex*	15	76	0.073	14.19	2.60
5539	introduction	message		23	76	0.002	13.41	2.46
5053	échantillon	parole		12	76	0.050	14.57	2.00
218	synchronisme	bit		12	76	0.054	14.59	1.78
3192	commande	mise	*(commande + commandes) de mise*	28	76	0.048	13.58	4.26
118	usager	demandeur		29	75	0.036	13.58	3.23
11784	structure	réseau		30	75	0.026	13.55	3.25
17853	caractère	classe		19	75	0.059	13.73	3.46
12754	classe	architecture		10	75	0.146	15.05	2.10
18601	donnée	taxation	*données de taxation*	23	74	0.051	13.57	3.93
748	effort	écoute	*effort d' écoute*	7	74	0.201	16.59	0.38
13120	procédure	retour	*(procédure + procédures) de retour*	28	74	0.050	13.48	4.34
5334	message	ordre	*(message + messages) d' ordre*	33	74	0.048	13.49	3.83
13662	codage	zone	*codage par zones*	15	74	0.071	14.02	3.68
10369	présence	bruit		17	74	0.048	13.81	3.62
18128	temps	traversée	*temps de traversée*	14	74	0.084	13.82	3.43
8986	temps	voie		43	73	0.039	13.76	3.43
3877	adresse	renvoi	*adresse de renvoi*	14	73	0.080	14.12	3.33
17392	stratégie	surcharge		9	73	0.107	15.46	1.33
4229	moyen	procédure		22	73	0.045	13.54	**4.37**
15162	message	assignation	*message d' assignation*	19	73	0.061	13.26	3.83
12242	encombrement	réseau		30	73	0.025	13.47	2.49
8188	message	incohérence	*message d' incohérence*	14	73	0.069	13.12	3.83
15043	information	gestion		44	73	0.041	13.76	4.04
6639	courbe	réponse		14	73	0.019	13.90	2.06
8495	message	refus		24	73	0.054	13.30	3.83
8520	restriction	identification		13	72	0.053	14.20	2.74
4666	demande	activation	*(demande + demandes) d' activation*	17	72	0.067	13.65	3.70
8504	compteur	bloc		17	72	0.052	13.74	3.99
12103	échec	tentative	*échec (d' une + de la) tentative*	13	72	0.077	14.24	1.99
21878	négociation	capacité		10	72	0.105	15.04	2.28
8516	demande	connexion	*(demande + demandes) de connexion*	45	72	0.035	13.77	3.70
1834	code	raison		17	72	0.067	13.56	3.40
19376	état	occupation		23	72	0.052	13.44	3.94
6267	test	faisceau		15	72	0.045	13.87	2.44
10685	référence	session		19	72	0.049	13.58	3.09
17467	télex	origine	*télex d' origine*	20	71	0.040	13.51	2.76
7711	niveau	écoute		16	71	0.070	13.58	3.68
5779	résultat	mesure		20	71	0.039	13.50	2.73
6715	caractéristique	transmission	*caractéristiques de transmission*	41	71	0.033	13.64	4.07
15722	conclusion	accord		7	71	0.148	16.16	0.38
11156	fiche	essai		12	71	-0.024	13.68	0.50
3564	signal	ligne		118	71	0.050	15.25	4.26
10589	information	usager		44	71	0.040	13.71	4.04
6865	rapport	terre	*rapport à la terre*	12	71	0.095	14.23	4.08
20693	préfixe	accès	*(préfixe + préfixes) d' (accès + accès)*	11	71	0.010	14.19	1.28
1457	charge	référence	*(charge + charges) de référence*	19	71	0.030	13.48	2.16

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
1530	champ	commande	*(champ + champs) de commande*	29	70	0.029	13.40	2.77
1294	référence	document		20	70	0.044	13.48	3.09
1503	ensemble	liaison		27	70	0.034	13.39	**4.04**
8095	couche	session		20	70	0.047	13.47	2.04
14790	méthode	masquage	*méthode de masquage*	11	70	0.106	14.18	4.00
4022	combinaison	élément		19	70	0.043	13.49	3.31
15081	gestion	dialogue		20	70	0.054	13.42	2.90
2946	note	opinion		6	70	0.306	17.15	0.41
5537	expression	code		10	70	-0.008	14.17	0.57
17685	paramètre	primitif		17	69	0.059	13.55	4.16
18388	activité	maintenance	*(activité + activités) de maintenance*	12	69	0.035	14.10	2.23
20983	accès	entrant	*accès entrante*	11	69	0.103	14.11	3.63
8061	protocole	clavier	*(protocole + protocoles) clavier*	13	69	0.081	13.96	3.25
17697	annuleurs	écho	*annuleurs d' écho*	10	69	0.018	14.38	0.64
12111	nombre	tentative	*nombre (des + de) tentatives*	20	69	0.055	13.35	**4.21**
16540	supervision	groupe		17	69	0.029	13.51	2.34
15015	affaiblissement	insertion	*affaiblissement d' insertion*	11	69	0.104	14.28	2.81
6178	coupure	porteur	*coupure de la porteuse*	11	69	0.089	14.45	2.25
17578	diagramme	état		21	69	0.029	13.35	2.44
13336	provenance	réseau		26	69	0.020	13.27	2.99
10812	commande	extension	*commande d' extension*	17	69	0.062	13.43	4.26
5356	fonction	orientation		13	69	0.076	13.36	4.27
2612	procédure	refus	*procédure de refus*	20	69	0.056	13.25	4.34
13322	réseau	intégration		14	68	0.074	13.49	2.97
15452	sonie	réception		21	68	0.012	13.22	2.05
643	urgence	charge	*urgence de la charge*	11	68	0.050	14.27	1.64
1730	nombre	maximum	*nombre maximum*	19	68	0.056	13.30	**4.21**
21701	point	extrémité		35	68	0.043	13.38	2.69
8628	langage	description	*langage de description*	7	68	0.185	16.14	1.26
17895	remise	zéro	*remise à zéro*	9	68	0.139	14.88	2.27
14456	élément	présentation	*(élément + éléments) de présentation*	19	68	0.054	13.31	3.01
10255	attention	administration	*attention des administrations*	8	67	0.116	15.44	1.30
21346	chaîne	connexion	*(chaîne + chaînes) de (connexion + connexions)*	22	67	0.016	13.20	1.84
17650	champ	paramètre		21	67	0.037	13.30	2.77
13650	zone	numérotage		19	67	0.033	13.34	2.57
14807	protection	situation	*protection en situation*	9	67	0.113	15.01	2.23
9684	distorsion	phase		15	67	0.059	13.60	2.83
250	détournement	trafic		10	67	-0.019	13.92	0.57
9485	oeuvre	service		25	67	0.013	13.17	3.28
9389	sélection	circuit		29	67	0.017	13.23	3.05
12124	maintien	rythme	*maintien du rythme*	10	67	0.091	14.61	2.73
16549	point	mesure	*(point + points) de mesure*	43	67	0.041	13.58	2.69
8687	voie	acheminement	*(voie + voies) d' acheminement*	30	67	0.040	13.31	3.55
7644	sélection	cadran		10	67	0.112	14.50	3.05
1084	chiffre	adresse		26	67	0.025	13.22	2.31
19665	extrémité	circuit		45	67	0.029	13.61	2.65
4216	diagramme	figure		13	67	0.065	13.83	2.44
16140	seuil	réservation	*(seuil + seuils) de réservation*	8	67	0.169	15.36	2.83
15151	côté	réseau		24	66	0.016	13.16	2.62
4231	ensemble	route		18	66	0.047	13.34	**4.04**
6462	gestion	trafic	*gestion du trafic*	32	66	0.036	13.32	2.90
6408	distorsion	temps		23	66	0.032	13.22	2.83
13231	différence	niveau		13	66	0.019	13.66	2.49
7933	touche	fonction	*touches de fonction*	11	66	0.006	13.89	1.35
6901	temps	coupure	*temps de coupure*	16	66	0.062	13.32	3.43
8363	message	signal	*message (du + de) signal*	2	66	-0.010	-2.74	3.83
7389	voie	voie		40	66	0.035	13.48	3.55
9893	pays	taille	*pays de taille*	10	66	0.108	14.35	2.45
203	équipement	terminaison	*(équipement + équipements) de terminaison*	18	66	0.056	13.21	3.37
8274	gamme	niveau		16	66	0.028	13.38	2.25

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
14935	valeur	temporisation		21	66	0.046	13.20	3.86
6972	commande	soumission	*commande de soumission*	12	65	0.081	13.53	4.26
9150	lettre	accent	*(lettre + lettres) avec accent*	6	65	0.320	16.72	1.64
4388	défaillance	liaison		20	65	0.025	13.18	2.91
19056	identité	station		17	65	0.045	13.32	2.25
14706	gauche	droite	*gauche à droite*	6	65	0.291	16.74	1.48
15957	spécification	essai		27	65	0.027	13.16	3.39
19858	plan	lèvre	*plan des lèvres*	10	65	0.104	14.16	2.26
5572	réponse	message		47	64	0.027	13.61	3.61
18272	système	satellite		31	64	0.042	13.19	2.41
13773	convertisseur	loi		6	64	0.300	16.64	1.84
12725	défaut	adaptation	*(défaut + défauts) d' adaptation*	8	64	0.133	15.17	2.75
12957	titre	question		11	64	0.067	14.03	2.52
498	intervalle	silence	*(intervalle + intervalles) de silence*	12	64	0.078	13.71	0.98
18399	base	donnée	*(base + bases) de données*	32	64	0.023	13.21	4.27
808	période	urgence	*période probatoire d' urgence*	13	64	0.064	13.61	3.09
9021	réglage	affaiblisseur	*(réglages + réglage) de l' affaiblisseur*	6	64	0.280	16.22	2.62
20933	accès	base		27	64	0.033	13.14	3.63
9467	comptage	octet	*comptage d' octets*	10	63	0.068	14.23	2.54
3476	abandon	dialogue		10	63	0.062	14.21	1.84
21218	modèle	intégration	*modèle d' intégration*	9	63	0.116	14.54	2.72
1868	cas	échec		22	63	0.046	13.04	**4.63**
17321	marge	sécurité	*(marge + marges) de sécurité*	8	63	0.135	15.07	2.60
8600	service	signalisation		5	63	-0.012	1.06	3.72
5437	protocole	transport	*protocole de transport*	20	63	0.036	13.11	3.25
4369	circuit	amont		16	63	0.058	13.08	4.15
2145	mise	place	*mise en place*	10	63	0.095	13.76	2.50
17092	paramètre	qualité		18	63	0.045	13.14	4.16
10296	conception	expérience	*conception de l' expérience*	6	63	0.278	16.26	2.55
8599	signalisation	service	*signalisation (des services + du service)*	2	63	-0.008	-2.68	4.51
3458	séquence	signal		57	63	0.028	13.83	3.72
12221	référence	bouche		12	63	0.072	13.64	3.09
3104	noeud	relais	*noeuds relais + noeud relais*	12	63	0.063	13.69	2.12
13398	utilisateur	service		29	62	0.018	13.08	2.72
2392	seconde/second	maximum	*(seconde + secondes) au maximum*	9	62	0.079	14.50	2.54
18937	disponibilité	faisceau		13	62	0.034	13.49	2.67
6801	pente	courbe		5	62	0.385	17.39	0.45
12413	rejet	composant		11	62	0.040	13.83	2.64
1220	fréquence	référence		26	62	0.032	13.07	3.13
15952	initiateur	essai	*initiateur (de l' + d' un) essai*	11	62	-0.033	13.31	0.86
20570	nombre	unité		23	62	0.043	13.02	**4.21**
5474	fonction	acheminement		35	62	0.038	13.24	4.27
11853	enregistrement	localisation		12	62	0.041	13.63	3.09
10018	simulateur	charge	*simulateur de charge*	10	62	0.040	14.03	1.75
9541	type	connexion		47	62	0.033	13.57	**4.26**
5357	désactivation	canal		12	62	-0.022	13.23	1.37
18343	durée	signal		53	62	0.026	13.70	3.67
7897	réception	signalisation		3	62	-0.013	-0.98	3.16
813	encombrement	équipement		19	62	0.031	13.06	2.49
7293	décibel	rapport	*décibels par rapport*	7	62	0.033	14.90	0.38
21069	efficacité	réception		24	62	0.017	12.98	2.52
18476	demande	contrôle	*demande de contrôle*	26	62	0.037	13.05	3.70
1970	axe	pavillon	*axe du pavillon*	6	62	0.260	16.32	1.39
14628	association	section		9	61	0.059	14.35	1.75
542	encombrement	canal		24	61	0.022	12.99	2.49
15082	élément	départ	*(élément + éléments) de départ*	38	61	0.033	13.31	3.01
3394	rétablissement	canal		18	61	0.012	12.97	2.66
1086	texte	commande	*texte de commande*	19	61	0.012	12.95	2.66
13704	déplacement	position		9	61	0.056	14.29	1.86
5338	bloc	message	*bloc message + blocs messages*	31	61	0.015	13.05	3.05

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
7264	rythme	réception		17	61	-0.001	12.92	2.48
2519	valeur	seuil		15	61	0.056	13.12	3.86
4732	canal	signalisation		5	61	-0.012	1.10	0.31
15116	trafic	départ		29	61	0.028	13.05	2.59
15051	signal	réseau	*signaux reçus (du + au) réseau*	2	60	-0.007	-2.63	4.26
8352	identification	plan	*identification du plan*	14	60	0.056	13.23	3.11
129	protocole	classe		15	60	0.048	13.16	3.25
8422	affaiblissement	fonction	*affaiblissement en fonction*	21	60	0.031	12.96	2.81
728	commande	suppresseurs	*commande des suppresseurs*	15	60	0.057	13.04	4.26
3018	cour/cours	période	*cours (d' une + de la) période*	11	60	0.069	13.67	3.45
6605	temps	émission		33	60	0.034	13.13	3.43
6025	multifréquence	enregistreur		10	60	0.021	13.78	1.88
19334	applicabilité	recommandation	*applicabilité des recommandations*	11	60	0.004	13.47	1.30
4466	temps	transit		32	60	0.034	13.10	3.43
19110	seuil	intelligibilité		7	60	0.167	15.14	2.83
8409	identification	station		18	60	0.040	12.97	3.11
16370	opération	borne		9	60	0.101	14.14	3.48
5361	message	suspension		17	60	0.052	12.70	3.83
2980	minute	an	*minutes par an*	5	60	0.385	17.05	1.73
13099	liste	définition	*liste des définitions*	10	60	0.081	13.86	3.43
21432	impulsion	polarité	*(impulsion + impulsions) de polarité*	10	59	0.072	13.89	3.02
3402	élément	signal		67	59	0.031	14.03	3.01
13818	dispositif	détection	*(dispositif + dispositifs) de détection*	12	59	0.066	13.37	3.68
17568	ensemble	règle		10	59	0.087	13.71	**4.04**
7877	fil/fils	connexion	*fils (de la + d' une) connexion*	17	59	0.001	12.85	2.56
5563	message	réponse		51	59	0.038	13.60	3.83
16318	présentation	texte		11	59	0.045	13.59	3.00
6929	réception	indication		22	59	0.043	12.84	3.16
2891	transfert	sémaphore	*transfert (transfert + sémaphore)*	81	59	0.033	14.36	2.67
10768	séparateur	information	*(séparateur + séparateurs) d' information*	11	59	-0.050	12.91	0.66
10743	synchronisation	trame	*synchronisation de trame*	12	59	0.015	13.29	2.47
18778	début	mesure		19	59	0.031	12.90	3.69
12441	réseau	mode		26	59	0.038	12.91	2.97
10002	longueur	ligne		25	58	0.019	12.87	3.23
6768	pas	traduction	*pas de traduction*	8	58	0.083	14.50	2.02
8079	passage	urgence	*passage d' urgence*	13	58	0.055	13.18	2.68
17900	lieu	difficulté		7	58	0.154	14.94	3.77
11109	accord	administration		7	58	0.089	14.95	1.83
13491	relation	phase	*(relation + relations) de phase*	11	58	0.054	13.50	2.23
11994	fonctionnement	signalisation		12	58	-0.008	4.58	4.69
20959	analyse	numéro	*analyse (du numéro + des numéros)*	14	58	0.031	13.06	2.84
14149	profil	souscription	*profil de souscription*	6	58	0.204	15.80	1.43
7058	erreur	syntaxe	*(erreur + erreurs) de syntaxe*	8	58	0.116	14.16	3.38
5317	format	domaine	*format du domaine*	12	57	0.059	13.25	3.27
3553	absence	défaillance	*absence de (défaillance + défaillances)*	13	57	0.044	13.14	3.66
21040	retour	contrôle	*retour sous contrôle*	19	57	0.031	12.82	2.60
3972	signe	arithmétique	*signe arithmétique + signes arithmétiques*	6	57	0.211	15.57	2.23
8065	service	session		28	57	0.038	12.88	3.72
815	méthode	codage	*(méthode + méthodes) de codage*	21	57	0.034	12.81	4.00
329	code	nom		13	57	0.060	12.92	3.40
374	blocage	circuit		35	57	0.020	13.06	3.26
9855	libération	enregistreur		20	57	0.035	12.80	2.62
2095	ligne	codage	*ligne (en + de) codage*	25	57	0.035	12.84	3.44
17394	état	canal		40	57	0.031	13.22	3.94
8283	message	test		18	56	0.047	12.56	3.83
3137	opinion	usager		10	56	-0.019	13.23	1.35
21672	temporisation	contrôle	*(temporisation + temporisations) de contrôle*	14	56	0.025	12.93	2.94
3805	chiffre	indicatif		14	56	0.040	12.96	2.31
19838	règle	application		14	56	0.027	12.93	3.23
7751	détection	dérangement		12	56	0.048	13.17	3.00
12383	demande	réinitialisation	*demande de réinitialisation*	16	56	0.046	12.78	3.70

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
16822	fanion	fermeture		5	56	0.302	16.27	1.72
745	période	silence	*(période + périodes) de silence*	10	56	0.073	13.48	3.09
16826	technique	extension	*(technique + techniques) d' extension*	9	56	0.072	13.83	3.40
21423	entrée	voie		20	56	0.018	12.71	3.17
15519	méthode	évaluation		9	56	0.088	13.48	4.00
15375	fonction	inversion	*(fonction + fonctions) d' inversion*	14	56	0.055	12.68	4.27
11754	rapport	valeur		15	56	0.039	12.85	**4.08**
12998	réseau	aval		15	55	0.051	12.74	2.97
5285	format	message		34	55	0.017	12.97	3.27
20435	installation	abonné		12	55	-0.005	12.95	2.38
17208	accès	abonné		27	55	0.027	12.81	3.63
5753	spectre	bruit		10	55	0.019	13.40	2.09
2988	période	minute	*période probatoire d' une minute*	10	55	0.072	13.44	3.09
18873	amélioration	qualité	*amélioration de la qualité*	8	55	0.026	14.00	1.89
12585	section	amont		11	55	0.065	13.20	1.99
16100	borne	sortie		11	55	0.029	13.23	2.60
13344	restriction	présentation		9	55	0.056	13.76	2.74
13453	confirmation	récupération	*confirmation de récupération*	9	55	0.084	13.69	2.15
6347	propagation	boucle	*propagation en boucle*	9	55	0.065	13.74	1.06
2712	élément	procédure		24	55	0.034	12.72	3.01
3504	absence	onde		13	55	0.039	12.96	3.66
3936	poste	haut-parleur		7	55	0.132	14.49	2.45
4609	redémarrage	trafic	*redémarrage du trafic*	12	55	0.003	12.94	0.82
411	domaine	octet	*(domaine d' un + domaines de l') octet*	13	54	0.047	12.92	2.92
10308	partie	transaction	*partie transaction*	20	54	0.033	12.65	**4.35**
14992	présentation	identification	*présentation d' identification*	11	54	0.030	13.14	3.00
18447	procédure	contrôle		32	54	0.035	12.88	4.34
16671	communication	transit		22	54	0.027	12.65	3.67
3479	indépendance	égard	*indépendance à l' égard*	5	54	0.226	16.22	1.30
6718	nature	adresse		17	54	0.006	12.58	2.16
3747	poste	poste	*poste à poste*	11	54	0.040	13.12	2.45
12530	horloge	maintien	*horloge de maintien*	7	53	0.077	14.43	1.60
9118	niveau	seuil		15	53	0.047	12.62	3.68
14276	équipement	multiplexage	*(équipement + équipements) de multiplexage*	16	53	0.045	12.58	3.37
4981	ordre	bloc		15	53	0.035	12.70	3.24
10634	assistance	opérateur		7	53	0.017	14.13	1.23
10065	procédure	interrogation	*(procédure + procédures) d' interrogation*	12	53	0.058	12.58	4.34
4590	appel	destination		27	53	0.029	12.72	4.01
5488	usager	session		16	53	0.034	12.62	3.23
7055	équivalent	fonction		12	53	0.010	12.83	1.48
8983	attente	numérotation	*(attente + attentes) après numérotation*	14	53	0.010	12.65	2.74
2292	gamme	valeur		11	53	0.034	13.03	2.25
8412	régénération	signal		11	52	-0.082	12.04	0.29
2489	champ	adresse		24	52	0.019	12.60	2.77
20355	évaluation	qualité		9	52	0.030	13.39	2.97
18287	durée	intervalle		13	52	0.048	12.70	3.67
15738	opérateur	départ		20	52	0.016	12.52	2.64
3212	demande	établissement		23	52	0.031	12.57	3.70
1887	plage	blanc		5	52	0.237	15.93	2.08
17594	bord	fenêtre	*bord (inférieur + supérieur) de la fenêtre*	6	52	0.084	14.84	1.63
7915	équipement	multiplication	*(équipement + équipements) de multiplication*	10	52	0.066	12.75	3.37
9442	octet	service	*(octet + octets) de service*	19	52	-0.002	12.42	2.77
11848	composant	lancement		9	52	0.054	13.41	2.68
4748	classe	ressource		10	52	0.056	13.13	2.10
9224	partie	équipement		22	52	0.028	12.55	**4.35**
10264	format	champ		13	52	0.040	12.71	3.27
307	code	échappement	*(code + codes) d' échappement*	13	51	0.051	12.55	3.40
6559	absence	perturbation	*absence de perturbations*	9	51	0.068	13.33	3.66
13836	filtre	élimination	*filtre à élimination*	5	51	0.235	15.62	2.36

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
15075	générateur	signal		19	51	-0.018	12.29	2.01
10330	choix	administration		9	51	0.059	13.35	3.73
14233	recherche	ligne		13	51	-0.017	12.47	2.28
17409	renseignement	état	*renseignements d' état*	11	51	-0.009	12.74	2.41
2148	remise	fonctionnement	*(remise + remises) en fonctionnement*	18	51	0.012	12.45	2.27
15050	signal	information		4	51	-0.006	0.40	4.26
21255	traversée	commutateur		8	51	-0.027	13.27	1.18
14679	identificateur	terminal		16	51	0.032	12.50	3.05
9736	protocole	application		17	51	0.028	12.48	3.25
19242	numéro	station		20	51	0.034	12.47	3.09
18262	espacement	caractère		12	51	-0.001	12.63	2.36
18754	suppression	écho		10	51	0.011	12.97	2.61
1901	valeur	tableau		14	51	0.044	12.53	3.86
16048	interruption	procédure		13	51	0.024	12.63	2.89
9951	séquence	émission		25	50	0.027	12.56	3.72
18551	taxation	abonné		13	50	0.001	12.53	2.35
5744	essai	laboratoire	*(essai + essais) en laboratoire*	8	50	0.084	13.13	3.82
17720	système	référence		42	50	0.033	13.09	2.41
13545	abandon	transaction		10	50	0.013	12.93	1.84
7472	niveau	parole		22	50	0.033	12.47	3.68
2190	demande	mise		20	50	0.032	12.44	3.70
8605	description	service		27	50	0.013	12.57	3.51
14683	facteur	coopération	*facteur de coopération*	7	50	0.094	14.02	2.23
1635	événement	interfonctionnement		9	50	0.014	13.14	2.40
7155	objet	recommandation		20	50	0.025	12.44	2.75
1262	interface	usager		20	50	0.021	12.43	3.26
4421	titre	variante		6	50	0.140	14.50	2.52
10044	procédure	activation	*(procédure + procédures) d' activation*	16	50	0.043	12.33	4.34
4234	intervalle	minute	*(intervalles + intervalle) d' une minute*	10	50	0.059	12.89	0.98
18901	bloc	donnée		24	50	0.011	12.46	3.05
15187	organe	commutation	*(organe + organes) de commutation*	12	49	-0.000	12.54	2.53

Valeur indéfinie du critère de vraisemblance

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
20444	accès	rpdcp	*accès au rpdcp*	4	∞	0.074	11.91	3.63
9577	affaiblisseur	bruit	*affaiblisseurs de bruit*	2	∞	-0.267	10.53	0.00
17655	aide	psophomètre	*aide d' un psophomètre*	2	∞	0.090	11.89	3.80
2271	allumage	lampe		2	∞	0.279	16.27	0.00
4454	analyse	variance		4	∞	0.144	13.82	2.84
3916	angle	inclinaison	*angle d' inclinaison*	2	∞	0.373	16.01	1.33
10437	annulation	marque		2	∞	0.152	13.42	2.55
309	appartenance	usager	*appartenance (de l' + d' un) usager*	2	∞	-0.294	9.43	0.00
18969	boîte	lettre	*(boîte + boîtes) aux lettres*	2	∞	-0.174	12.64	0.00
7610	bruit	friture	*bruits de friture*	3	∞	0.082	11.93	2.58
16273	bulletin	exploitation	*bulletin d' exploitation*	3	∞	-0.185	11.63	0.00
10694	cadrage	document	*cadrage du document*	2	∞	-0.269	10.46	0.00
1456	cahier	charge	*(cahier des + cahiers de) charges*	3	∞	-0.163	12.19	0.00
16550	calendrier	mesure		4	∞	-0.149	11.98	0.00
11499	capsule	microphone	*(capsule + capsules) du microphone*	2	∞	-0.171	12.69	0.00
20156	caractère	substitution	*(caractère + caractères) de substitution*	3	∞	0.064	11.24	3.46
7333	carrier	transmission	*carrier transmission*	2	∞	-0.312	8.45	0.00
16878	cas	être	*cas être*	2	∞	0.029	8.63	4.63
707	cas	conflit	*cas de conflit*	2	∞	0.029	8.63	4.63
17105	cas	insuccès	*cas d' insuccès*	2	∞	0.029	8.63	4.63
11010	cheminement	information		2	∞	-0.313	8.34	0.00
21664	clé	chiffrement	*clés de chiffrement*	3	∞	-0.178	11.81	0.00
5800	clef	appel	*clef d' appel*	2	∞	-0.316	8.10	0.00
15694	client	service	*client du service*	2	∞	-0.318	8.00	0.00
19803	collecteur	donnée	*collecteur de données*	3	∞	-0.242	9.34	0.00
16577	commission	étude	*(commission + commissions) d' études*	19	∞	0.110	16.53	0.00
20233	communauté	conception	*communauté de conception*	3	∞	0.089	15.37	0.00
20527	commutateur	frontière		2	∞	0.057	10.61	3.36
20682	compte	progrès	*compte tenu des progrès*	2	∞	0.055	10.49	4.45
12720	concentrateur	ligne	*concentrateur de (lignes + ligne)*	4	∞	-0.189	10.54	0.00
20288	condition	applicabilité	*conditions d' applicabilité*	5	∞	0.082	12.44	4.29
14170	condition	entraînement	*conditions d' entraînement*	2	∞	0.043	9.79	4.29
13500	conduction	travers	*conduction à travers*	2	∞	-0.207	12.05	0.00
18079	contrôle	vraisemblance		9	∞	0.115	14.05	2.89
13921	contraste	bruit	*contraste de bruit*	2	∞	-0.267	10.53	0.00
11310	cour/cours	préparation		2	∞	0.075	11.38	3.45
8804	créneau	voie	*créneau temporel d' une voie*	2	∞	-0.302	9.03	0.00
10775	crochet	commutateur	*crochet commutateur*	11	∞	0.016	14.89	0.00
12675	définition	vocabulaire	*définitions du vocabulaire*	2	∞	0.064	10.91	3.97
20240	dérapage	entrée		6	∞	-0.062	13.54	0.00
3269	déséquilibre	impédance	*déséquilibre d' impédance*	2	∞	-0.174	12.64	0.00
17119	degré	complexité	*(degré + degrés) de complexité*	3	∞	0.101	12.55	1.08
8577	demande	reconnexion		2	∞	0.038	9.41	3.70
8875	destruction	signal	*destruction des signaux*	2	∞	-0.326	7.25	0.00
15193	différence	potentiel	*différence de potentiel*	2	∞	0.148	13.34	2.49
17483	distance	kilomètre	*distance en kilomètres*	2	∞	0.244	14.78	2.54
3619	donne	lieu	*donne lieu*	2	∞	0.055	15.01	0.00
10	dos	dos	*dos à dos*	3	∞	0.711	18.18	0.00
16892	durée	vie	*durée de vie*	2	∞	0.044	9.86	3.67
15466	échelle	graduation	*échelles à graduation*	2	∞	0.215	14.42	2.06
9274	effet	masque	*(effet + effets) de masque*	9	∞	0.189	15.48	3.53
7033	effet	obstruction	*effet d' obstruction*	2	∞	0.069	11.14	3.53
10278	emploi	conducteur	*emploi de conducteurs*	2	∞	0.075	11.37	4.19
2524	enregistrement	faute	*enregistrement (des + de) fautes*	2	∞	0.113	12.57	3.09
16387	épaisseur	micron	*épaisseurs en microns*	2	∞	0.264	15.01	0.98
4072	établissement	décompte	*établissement des décomptes*	5	∞	0.075	12.18	2.04
13265	étalon	condensateur	*étalons à condensateur*	3	∞	0.418	17.18	0.00

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
20001	état	non-occupation	*état de non-occupation*	2	∞	0.033	9.05	3.94
11823	exploitant	réseau		3	∞	-0.240	9.48	0.00
1603	fiche	jack	*fiche du jack*	3	∞	0.318	15.86	0.50
981	file	attente	*(file + files) d' attente*	29	∞	0.372	19.24	0.00
2349	foi/fois	mois	*fois par mois*	3	∞	0.200	14.51	3.39
11658	fonctionnement	multidestination	*fonctionnement multidestination*	2	∞	0.023	7.99	4.69
8426	fonction	affaiblissement	*fonction de l' affaiblissement*	3	∞	0.035	9.51	4.27
5027	fonction	appartenance	*fonction de l' appartenance*	2	∞	0.026	8.34	4.27
5429	fonction	convergence	*fonction de convergence*	3	∞	0.035	9.51	4.27
15740	formation	queue	*formation de queues*	36	∞	0.539	20.23	2.55
15259	frai/frais	maintenance	*frais de maintenance*	3	∞	-0.188	11.56	0.00
5935	front	attaque	*front d' attaque*	3	∞	0.466	16.95	1.48
17319	général	qualité	*généraux de qualité*	3	∞	-0.157	12.34	0.00
1290	gigue	phase	*gigue de phase*	2	∞	-0.220	11.78	0.00
6417	guise	support	*guise de support*	2	∞	-0.263	10.65	0.00
11828	home	position	*home position*	2	∞	-0.229	11.59	0.00
20803	hystérésis	compensation	*hystérésis de la compensation*	2	∞	0.181	15.78	0.00
20741	impossibilité	libération	*impossibilité de libération*	4	∞	-0.165	11.48	0.00
2526	ingénierie	trafic	*ingénierie (de + du) trafic*	2	∞	-0.292	9.54	0.00
11863	ingénieur	transmission		3	∞	-0.237	9.62	0.00
18608	intelligibilité	logatomes	*intelligibilité des logatomes*	3	∞	0.390	16.44	1.17
8063	intercommunication	service		2	∞	-0.318	8.00	0.00
8795	interférence	voie	*interférences entre voies*	4	∞	-0.177	11.03	0.00
14424	itinéraire	télécommunication	*itinéraire de (télécommunications + télécommunication)*	3	∞	-0.107	13.25	0.00
5029	jet	parole	*jets de (parole + paroles)*	3	∞	-0.177	11.84	0.00
2244	laps	temps	*laps de temps*	18	∞	0.067	15.89	0.00
20323	levée	garde	*levée de garde*	18	∞	0.122	16.65	0.00
1285	ligne	ajustement	*lignes d' ajustement*	3	∞	0.048	10.42	3.44
3133	ligne	diagonale	*ligne en diagonale*	2	∞	0.036	9.25	3.44
9962	ligne	traction		5	∞	0.068	11.89	3.44
15993	liste	abréviation	*liste (d' + des) abréviations*	6	∞	0.164	14.62	3.43
742	majuscule	cédille	*majuscule avec cédille*	5	∞	0.501	17.65	1.42
13795	manuel	pays	*(manuel + manuels) du pays*	2	∞	-0.263	10.66	0.00
9289	mesure	isolation		4	∞	0.082	12.20	4.31
11420	mesure	possible	*mesure du possible*	3	∞	0.067	11.37	4.31
12642	mise	dérivation	*mise en dérivation*	2	∞	0.042	9.70	2.50
5008	modulation	rate	*modulation rate*	2	∞	0.073	11.32	2.18
18210	moyen	distorsiomètre	*moyen d' un distorsiomètre*	3	∞	0.071	11.54	4.37
9452	multiplicateur	circuit	*multiplicateur de circuit*	5	∞	-0.170	10.50	0.00
9097	multiplication	circuit	*multiplication de (circuit + circuits)*	17	∞	-0.022	14.03	0.00
8459	mutilation	signal		9	∞	0.442	17.93	2.41
7484	niveau	automatisation	*niveau d' automatisation*	3	∞	0.044	10.18	3.68
7822	niveau	hiérarchie		4	∞	0.054	11.01	3.68
19884	numéro	feuillet	*(numéro + numéros) (de + des) (feuillet + feuillets)*	2	∞	0.035	9.20	3.09
18790	numéro	quadrant	*(numéro + numéros) de quadrant*	2	∞	0.035	9.20	3.09
817	objet	entente	*objet d' une entente*	2	∞	0.051	10.27	2.75
20578	octroi	priorité	*octroi d' une priorité*	2	∞	-0.157	12.89	0.00
1074	organigramme	figure	*(organigramme + organigrammes) de la figure*	5	∞	-0.002	14.56	0.00
739	orifice	conduit	*orifice du conduit*	2	∞	0.073	15.13	0.00
13589	période	accumulation	*période d' accumulation*	2	∞	0.063	10.87	3.09
5425	perforation	bande	*perforation (de la + d' une) bande*	5	∞	-0.078	13.36	0.00
9731	plan	symétrie	*plan de symétrie*	7	∞	0.137	14.27	2.26
4582	pont	alimentation	*(pont + ponts) d' alimentation*	7	∞	0.139	16.18	0.00
10809	potentiel	stockage	*potentiel de stockage*	3	∞	0.065	15.18	0.00
16630	préambule	recommandation	*préambule de la recommandation*	2	∞	-0.287	9.79	0.00
21413	présélection	arrivée	*présélection en arrivée*	2	∞	-0.296	9.37	0.00
14539	présentation	fiche	*présentation des fiches*	2	∞	0.112	12.53	3.00
8030	prescription	recommandation	*prescriptions de la recommandation*	2	∞	-0.287	9.79	0.00
7485	prestataire	service	*prestataire (du + de) service*	2	∞	-0.318	8.00	0.00
12817	procédure	audit	*procédure d' audit*	2	∞	0.026	8.35	4.34
11565	procédure	expédition	*(procédure + procédures) d' expédition*	6	∞	0.056	11.52	4.34
13222	radiodiffusion	réception	*radiodiffusion pour réception*	3	∞	-0.238	9.55	0.00

Index	N_1	N_2	Expression R.	NC	LOG	FAG	MI3	h1
20637	raison	commodité	*raisons de commodité*	3	∞	0.092	12.29	4.18
10058	raison	réglementation		2	∞	0.069	11.12	4.18
19710	raison	simplicité	*raisons de simplicité*	3	∞	0.092	12.29	4.18
16009	réception	positif	*réception en positif*	2	∞	0.027	8.43	3.16
19933	réciprocité	champ	*réciprocité en champ*	2	∞	-0.230	11.55	0.00
9909	réémetteur	bande	*réémetteur à bande*	3	∞	-0.176	11.89	0.00
14970	réémission	signal	*réémission de (ces signaux + un signal)*	2	∞	-0.326	7.25	0.00
8492	refus	message	*refus de message*	12	∞	0.298	17.13	1.31
3413	région	monde	*régions du monde*	3	∞	0.503	17.18	1.01
20234	rentrée	opérateur		5	∞	-0.045	13.94	0.00
11445	réoccupation	circuit	*réoccupation du circuit*	2	∞	-0.319	7.85	0.00
16753	répartiteur	groupe		21	∞	0.099	16.45	0.00
16735	réserve	conclusion	*réserve de la conclusion*	2	∞	0.187	14.01	2.64
4681	restriction	divulgation	*restriction de divulgation*	3	∞	0.169	14.03	2.74
7490	retour	chariot		40	∞	0.358	19.19	2.60
16230	risque	lésion	*risques de lésions*	2	∞	0.183	13.95	2.81
21624	rôle	séparateur	*rôle de séparateur*	3	∞	0.205	14.59	2.61
2913	sémaphore	esclave	*sémaphores esclaves*	2	∞	0.032	8.96	2.81
14497	service	inter-réseaux	*(services + service) inter-réseaux*	2	∞	0.025	8.26	3.72
9662	shift	modulation	*shift modulation*	2	∞	-0.262	10.69	0.00
17793	sigle	caractère		3	∞	-0.217	10.57	0.00
8385	signal	égalisation	*signaux d' égalisation*	2	∞	0.015	6.65	4.26
15585	signe	ponctuation	*(signe + signes) de ponctuation*	11	∞	0.425	18.05	2.23
8360	sou-champ	identification	*sous-champ d' identification*	2	∞	-0.258	10.83	0.00
7377	souscription	abonnement	*souscription de l' abonnement*	2	∞	0.118	15.42	0.00
9099	ssut	testeur	*ssut testeur*	3	∞	0.072	15.24	0.00
7147	stimulus	fait	*stimulus du fait*	2	∞	-0.124	13.34	0.00
6105	stock	papier	*stock de papier*	4	∞	0.039	15.01	0.00
2207	substance	élément	*substance de l' élément*	2	∞	-0.271	10.38	0.00
10247	sujet	intérêt	*sujets d' intérêt*	2	∞	0.132	13.01	3.08
20178	superposition	caractère	*superposition (de + des) caractères*	2	∞	-0.295	9.40	0.00
18720	suppresseur	écho	*suppresseur d' écho*	77	∞	0.443	20.86	0.00
10361	symbole	connecteur		3	∞	0.269	15.37	2.69
20099	système	multivoie	*système multivoie*	2	∞	0.024	8.04	2.41
9750	tache	exploration	*tache d' exploration*	2	∞	-0.208	12.02	0.00
13874	technique	transformation	*techniques de transformation*	2	∞	0.102	12.25	3.40
12714	téléimprimeur	simplex	*téléimprimeur simplex*	2	∞	0.191	14.07	2.34
573	temps	descente		5	∞	0.063	11.69	3.43
18995	temps	montée	*temps de montée*	5	∞	0.063	11.69	3.43
12667	temps	présélection	*temps de présélection*	3	∞	0.045	10.22	3.43
10035	titre	explication	*titre d' explication*	2	∞	0.104	12.33	2.52
8800	tolérance	fabrication	*(tolérance + tolérances) de fabrication*	7	∞	0.363	17.08	2.18
1388	tolérance	gigue	*tolérance de gigue*	6	∞	0.330	16.63	2.18
15156	transposition	élément	*transpositions d' éléments*	2	∞	-0.271	10.38	0.00
9415	type	mesure	*(type + types) de (mesure + mesures)*	3	0	-0.013	3.17	4.26
7616	type	stimulus	*type stimulus*	7	∞	0.078	12.64	4.26
7412	voie	témoin	*voie témoin + voies témoins*	7	∞	0.079	12.70	3.55
8786	voix	femme	*voix de femme*	3	∞	0.342	16.06	1.48
15478	voix	homme		6	∞	0.541	18.06	1.48

N ADJ

Valeurs décroissantes du critère de vraisemblance

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
1893	équipement	terminal	*équipements terminaux + équipement terminal*	275	1424	0.603	21.63	2.01
7964	considération	général	*considérations générales*	256	1384	0.566	21.38	2.86
3792	service	supplémentaire	*services supplémentaires + service supplémentaire*	340	1275	0.459	21.15	2.39
6880	télégraphie	harmonique	*télégraphie harmonique*	152	1250	0.885	21.89	0.37
4835	étude	ultérieur	*étude ultérieure + études ultérieures*	169	1170	0.735	21.52	1.57
7326	caractère	graphique	*caractères graphiques + caractère graphique*	196	1112	0.591	21.09	1.06
7684	entité	fonctionnel	*entités fonctionnelles + entité fonctionnelle*	199	999	0.503	20.70	2.48
2564	centre	international	*centre international + centres internationaux*	325	963	0.345	20.32	3.25
3328	adresse	complet		183	873	0.460	20.33	2.77
1703	effet	local		169	865	0.458	20.23	3.34
3158	station	mobile	*station mobile + stations mobiles*	164	854	0.504	20.41	1.54
7286	accord	bilatéral	*accords bilatéraux + accord bilatéral*	80	682	0.787	20.68	0.30
4075	niveau	relatif	*niveau relatif + niveaux relatifs*	202	671	0.326	19.48	3.12
3799	message	initial	*message initial + messages initiaux*	168	626	0.340	19.33	2.43
1224	ligne	appelant	*ligne appelante*	105	622	0.481	19.64	0.68
7767	prise	simultané	*prise simultanée + prises simultanées*	78	608	0.683	20.27	1.67
153	intervalle	unitaire	*intervalle unitaire + intervalles unitaires*	80	563	0.595	19.90	0.91
5966	appellation	global	*appellation globale + appellations globales*	87	546	0.505	19.58	2.43
6986	débit	binaire	*débit binaire + débits binaires*	105	515	0.358	18.93	2.19
1259	élément	binaire		118	488	0.318	18.72	2.19
5327	réseau	national	*réseaux nationaux + réseau national*	208	479	0.239	18.66	3.32
4223	pression	acoustique	*pression acoustique + pressions acoustiques*	65	471	0.501	19.23	2.74
3700	bouche	artificiel		61	422	0.475	19.00	1.45
4045	position	actif	*position active*	78	413	0.376	18.61	2.07
2295	livre	rouge	*livre rouge*	53	408	0.557	19.13	0.35
6180	enregistreur	international		105	406	0.206	17.62	3.25
2644	groupe	primaire	*groupe primaire + groupes primaires*	72	401	0.393	18.62	1.94
4927	page	blanc	*page blanche*	53	398	0.563	19.23	1.68
1010	combinaison	binaire	*combinaison binaire + combinaisons binaires*	81	398	0.307	18.21	2.19
1901	poste	téléphonique		111	391	0.234	17.90	3.07
7362	paramètre	obligatoire		69	380	0.377	18.41	1.86
2019	fréquence	vocal	*fréquence vocale + fréquences vocales*	105	378	0.260	17.99	2.34
8018	libération	forcé	*libération forcée + libérations forcées*	50	374	0.543	19.00	0.84
6195	longueur	variable		59	373	0.434	18.58	1.50
2674	bit	indicateur	*bit indicateur + bits indicateurs*	44	370	0.637	19.28	0.60
6499	voie	commun		82	366	0.301	18.00	2.48
3906	voix	artificiel	*voix artificielle + voix artificielles*	54	348	0.408	18.43	1.45
7248	sou-systèmes	utilisateurs	*sous-systèmes utilisateurs*	41	344	0.629	19.16	0.66
7085	donnée	exprès	*données exprès*	44	340	0.496	18.54	0.11
3041	alignement	initial	*alignement initial*	61	337	0.296	17.81	2.43
6272	signal	composite	*signal composite + signaux composites*	81	335	0.253	17.45	1.14
768	procédure	logique	*procédure logique + procédures logiques*	60	334	0.352	18.01	1.53
329	alphabet	télégraphique	*alphabet télégraphique*	72	333	0.276	17.76	2.71
7167	entité	homologue	*entité homologue + entités homologues*	53	332	0.376	18.00	0.59

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
3403	courant	porteur	*courants porteurs + courant porteur*	57	331	0.378	18.18	1.48
5689	réseau	public		77	320	0.251	17.36	1.03
1213	valeur	nominal	*valeurs nominales + valeur nominale*	97	304	0.217	17.35	2.67
3416	courant	continu	*courant continu*	55	301	0.338	17.84	2.53
1030	livre	bleu	*livre bleu*	40	301	0.472	18.28	0.43
1775	oreille	artificiel	*oreille artificielle + oreilles artificielles*	47	301	0.371	18.02	1.45
8744	groupe	fermé	*groupe fermé*	41	283	0.413	17.94	0.62
8293	unité	fonctionnel	*unités fonctionnelles + unité fonctionnelle*	53	274	0.235	17.10	2.48
4225	bruit	impulsif		37	273	0.415	17.80	0.12
6580	appareil	téléphonique		97	272	0.178	16.98	3.07
2514	circuit	international		213	272	0.166	17.70	3.25
4112	communication	fictif	*communications fictives + communication fictive*	47	269	0.324	17.44	1.04
8605	autocommutateur	privé	*autocommutateur privé*	36	268	0.386	17.85	1.89
6193	situation	anormal	*situations anormales + situation anormale*	41	263	0.368	17.78	1.85
6407	recommandation	pertinent	*recommandation pertinente + recommandations pertinentes*	43	262	0.360	17.68	2.19
4807	commutateur	numérique	*commutateur numérique + commutateurs numériques*	114	260	0.170	16.98	3.73
8559	modèle	fonctionnel	*modèles fonctionnels + modèle fonctionnel*	53	259	0.223	16.95	2.48
4743	zone	imprimable	*zone imprimable + zones imprimables*	32	258	0.532	18.37	0.98
3607	répétition	automatique	*répétition automatique*	48	258	0.212	16.80	3.77
6292	transfert	interdit	*transfert interdit*	40	252	0.362	17.56	1.04
5793	exploitation	semi-automatique		49	251	0.287	17.19	1.29
3962	niveau	absolu	*niveau absolu + niveaux absolus*	58	239	0.223	16.67	1.54
636	interface	analogique	*interfaces analogiques + interface analogique*	52	233	0.231	16.84	3.43
2126	noeud	intermédiaire	*noeud intermédiaire + noeuds intermédiaires*	37	232	0.328	17.38	2.17
5649	service	complémentaire	*services complémentaires + service complémentaire*	62	227	0.196	16.37	1.57
4753	condition	anormal		53	217	0.217	16.53	1.85
7763	numéro	national	*numéro national + numéros nationaux*	97	214	0.149	16.46	3.32
3121	appel	sortant	*appels sortants + appel sortant*	60	214	0.196	16.42	1.84
7566	dialogue	structuré		21	207	0.669	18.47	0.51
8740	tonalité	spécial	*tonalité spéciale*	37	204	0.263	16.79	3.55
1430	objectif	nominal	*objectifs nominaux + objectif nominal*	38	203	0.217	16.50	2.67
7724	extrémité	distant	*extrémité distante*	42	203	0.240	16.60	2.73
716	champ	magnétique	*champ magnétique + champs magnétiques*	29	202	0.377	17.27	1.24
2523	puissance	moyen	*puissance moyenne*	43	202	0.220	16.51	2.82
4155	message	multiple		57	191	0.176	16.04	2.60
3449	action	suivant	*actions suivantes*	68	181	0.128	15.81	**4.43**
2970	appareil	arythmique	*appareils arythmiques + appareil arythmique*	44	180	0.196	16.08	2.42
2845	signalisation	absent		33	178	0.220	15.82	0.45
6696	instant	significatif	*instants significatifs + instant significatif*	27	177	0.301	16.79	1.96
7146	état	inactif		27	176	0.291	16.38	0.56
7750	numéro	complet		69	175	0.143	15.86	2.77
6657	mesure	objectif		26	175	0.326	16.69	1.00
5475	couche	physique	*couche physique*	34	170	0.227	16.19	2.88
5471	type	téléphonique		58	170	0.126	15.62	3.07
1889	référence	local	*références locales + référence locale*	46	167	0.143	15.65	3.34
599	trafic	différé	*trafic différé*	29	166	0.265	16.29	1.36
7259	état	significatif	*état significatif + états significatifs*	37	165	0.202	15.91	1.96
3791	couche	supérieur		38	164	0.186	15.88	3.17
6864	commutation	manuel	*commutation manuelle*	31	163	0.231	16.17	2.93
4947	zone	géographique	*zone géographique*	21	162	0.403	17.05	1.28

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
7084	numéro	reçu	*numéro reçu*	42	162	0.178	15.69	2.04
6168	exploitation	bidirectionnel	*exploitation bidirectionnelle*	27	160	0.257	16.04	0.96
6189	longueur	fixe	*longueur fixe*	35	158	0.198	15.82	2.60
2542	onde	résiduel	*onde résiduelle*	20	158	0.393	16.88	0.62
211	transfert	restreint	*transfert restreint*	27	156	0.258	16.12	1.65
1180	nombre	maximal	*nombre maximal + nombres maximaux*	47	153	0.149	15.49	3.24
3639	partie	récepteur	*partie réceptrice*	34	152	0.191	15.73	2.39
151	fonctionnement	attendu	*fonctionnement attendu*	28	152	0.222	15.67	0.82
3682	voie	télégraphique	*voies télégraphiques + voie télégraphique*	61	151	0.132	15.47	2.71
4038	circuit	sortant	*circuit sortant + circuits sortants*	62	151	0.135	15.39	1.84
5859	retransmission	cyclique	*retransmission cyclique*	18	149	0.437	17.17	1.68
8720	degré	conventionnel	*degré conventionnel*	17	147	0.470	17.28	1.28
6990	règle	général	*règles générales + règle générale*	45	147	0.121	15.25	2.86
8098	administration	intéressé	*administrations intéressées*	17	147	0.464	17.19	0.86
4783	information	relatif		68	145	0.120	15.42	3.12
3491	sélection	direct	*sélection directe*	23	145	0.240	16.12	3.30
5861	exploitation	automatique		57	145	0.122	15.32	3.77
2697	distorsion	isochrone	*distorsion isochrone*	22	144	0.270	15.92	0.64
900	titre	facultatif	*titre facultatif*	28	144	0.182	15.66	3.16
7706	réserve	prêt	*réserve prête*	17	143	0.443	17.11	1.02
8212	observation	général	*observations générales*	31	142	0.112	15.06	2.86
6591	centre	directeur		42	142	0.151	15.20	2.20
755	livre	jaune	*livre jaune*	20	142	0.306	16.16	0.68
1905	signal	vocal		76	139	0.120	15.40	2.34
5957	information	nécessaire	*informations nécessaires + information nécessaire*	53	138	0.127	15.20	**3.93**
8749	heure	chargé	*heures chargées + heure chargée*	16	138	0.452	17.09	1.32
2803	appel	entrant	*appels entrants*	31	137	0.179	15.23	1.62
5263	nombre	entier	*nombre entier + nombres entiers*	23	136	0.233	15.56	0.95
2175	son	vocal	*sons vocaux*	28	134	0.138	15.22	2.34
6517	synchronisation	mineur	*synchronisation mineure*	13	132	0.583	17.47	0.26
8027	abonné	libre	*abonné libre*	30	131	0.172	15.30	2.28
230	codage	binaire	*codage binaire*	36	129	0.123	14.99	2.19
935	microphone	linéaire	*microphones linéaires + microphone linéaire*	15	129	0.423	16.92	1.97
8103	système	maritime	*système maritime + systèmes maritimes*	29	128	0.173	15.08	1.67
5881	notification	facultatif		22	126	0.167	15.38	3.16
104	accent	aigu	*accent aigu*	14	125	0.442	16.77	0.38
8786	capacité	essentiel	*capacités non essentielles*	20	125	0.254	15.84	2.21
8704	tonalité	audible	*tonalité audible + tonalités audibles*	18	125	0.300	16.07	1.50
6770	distorsion	propre	*distorsion propre*	24	125	0.203	15.32	2.26
1508	remise	positif	*remise positive*	19	125	0.255	15.94	2.34
7105	acheminement	détourné	*acheminement détourné + acheminements détournés*	18	124	0.293	15.96	0.87
2509	personnel	technique	*personnel technique*	17	123	0.242	15.92	3.38
7294	durée	minimal	*durée minimale + durées minimales*	34	123	0.143	14.92	3.03
2426	distorsion	arythmique	*distorsion arythmique*	32	122	0.147	14.96	2.42
8599	densité	spectral	*densité spectrale*	14	121	0.393	16.73	1.93
7490	caractère	alphabétique		22	121	0.190	14.90	0.52
7166	principe	général	*principes généraux + principe général*	36	120	0.099	14.67	2.86
6400	message	simple	*message simple + messages simples*	34	119	0.140	14.71	2.64
7711	donnée	anisochrones	*données anisochrones*	17	119	0.270	15.60	0.52
479	abri	acoustique	*abris acoustiques + abri acoustique*	17	118	0.168	15.35	2.74
291	protocole	générique	*protocoles génériques + protocole générique*	17	117	0.282	15.94	1.76
1571	fréquence	caractéristique	*fréquence caractéristique + fréquences caractéristiques*	26	116	0.165	14.81	1.73
6972	durée	égal		30	115	0.145	14.79	3.40
2252	résultat	négatif		20	115	0.218	15.41	1.98
8395	description	général	*description générale + descriptions générales*	34	115	0.094	14.54	2.86
57	alphabet	latin	*alphabet latin*	20	115	0.218	15.34	1.50

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
27	jeu	primaire	*jeu primaire + jeux primaires*	23	114	0.159	15.05	1.94
4885	commutateur	principal	*commutateurs principaux + commutateur principal*	30	112	0.142	14.60	2.54
5053	transfert	subséquent	*transfert subséquent + transferts subséquents*	24	112	0.171	14.94	1.81
4754	condition	normal		55	112	0.102	14.73	3.91
5470	communication	téléphonique		60	111	0.093	14.74	3.07
7321	durée	taxable	*durée taxable*	18	110	0.213	14.99	0.66
4764	document	normal	*documents normaux + document normal*	28	107	0.097	14.43	3.91
6888	commutation	automatique		34	107	0.095	14.41	3.77
2415	fréquence	moyen	*fréquences moyennes + fréquence moyenne*	40	107	0.111	14.48	2.82
7139	état	actif	*état actif*	32	107	0.122	14.50	2.07
4055	normalisation	futur	*normalisation future*	14	105	0.273	15.87	2.71
2885	usage	régional		14	105	0.294	15.62	0.84
4315	modulation	nominal	*modulation nominale*	24	104	0.112	14.51	2.67
4957	version	analogique	*version analogique*	27	104	0.118	14.49	3.43
7022	son	voisé		13	103	0.339	16.02	1.20
3212	option	national	*options nationales + option nationale*	30	103	0.071	14.12	3.32
8292	réseau	privé		43	102	0.108	14.33	1.89
7516	onde	sinusoïdal	*onde sinusoïdale + ondes sinusoïdales*	18	101	0.194	15.03	1.28
45	marge	net	*marge nette*	11	100	0.427	16.44	0.76
4905	procédure	normal	*procédures normales + procédure normale*	46	100	0.093	14.36	3.91
3648	bruit	pondéré		14	99	0.235	14.93	0.24
8193	accès	sortant	*accès sortant*	24	99	0.127	14.44	1.84
5935	courant	vocal	*courants vocaux*	35	99	0.098	14.25	2.34
649	champ	libre	*champ libre*	22	98	0.140	14.52	2.28
3066	alimentation	électrique	*alimentation électrique*	14	97	0.176	15.15	2.77
6989	extrémité	éloigné	*extrémité éloignée*	19	97	0.174	14.68	2.25
8104	système	synchrone		26	97	0.130	14.23	2.22
8657	usage	privé	*usage privé*	22	97	0.132	14.46	1.89
494	conduit	numérique	*conduit numérique + conduits numériques*	23	95	0.057	13.86	3.73
803	procédure	applicable		30	95	0.112	14.16	3.16
1922	bloc	fonctionnel	*blocs fonctionnels + bloc fonctionnel*	25	94	0.084	14.06	2.48
5380	couche	inférieur		25	93	0.114	14.20	3.39
5056	fréquence	porteur	*fréquence porteuse + fréquences porteuses*	28	93	0.117	14.09	1.48
1348	élément	unitaire		26	93	0.120	14.13	0.91
6948	adresse	incomplet	*adresse incomplète*	22	91	0.137	14.15	1.66
4945	liaison	normal	*liaison normale + liaisons normales*	37	91	0.086	14.04	3.91
747	réponse	positif		19	91	0.150	14.41	2.34
1357	temps	mort	*temps mort*	15	90	0.198	14.61	0.65
1378	référence	intermédiaire	*référence intermédiaire*	21	90	0.125	14.24	2.17
1938	idéogramme	chinois	*idéogrammes chinois*	8	90	0.612	17.09	0.50
2370	intervalle	significatif	*intervalle significatif + intervalles significatifs*	21	88	0.128	14.16	1.96
1635	diaphonie	intelligible	*diaphonie intelligible*	8	88	0.579	16.90	0.35
6069	nombre	suffisant	*nombre suffisant*	20	88	0.140	14.11	2.74
6206	commutateur	précédent	*commutateur précédent*	32	87	0.102	13.89	3.20
4022	appel	infructueux	*appels infructueux + appel infructueux*	18	87	0.147	13.96	0.92
1686	réponse	négatif		17	86	0.157	14.36	1.98
3161	résultat	partiel		17	86	0.153	14.38	2.69
8062	administration	européen	*administrations européennes*	10	86	0.347	15.71	0.54
2906	usage	national		33	85	0.064	13.72	3.32
7753	abonné	mobile		28	84	0.091	13.80	1.54
8627	nombre	limité		15	84	0.170	14.16	1.24
7972	extrémité	récepteur	*extrémité réceptrice*	22	84	0.113	13.93	2.39
1781	équivalent	global	*équivalent global + équivalents globaux*	15	84	0.115	14.29	2.43
6804	route	normal	*route normale*	21	83	0.067	13.71	3.91
6746	distorsion	total	*distorsion totale*	27	83	0.097	13.79	3.13
3565	trame	multiple	*trames multiples*	19	83	0.115	14.03	2.60
4043	essai	subjectif	*essai subjectif + essais subjectifs*	19	83	0.133	14.00	2.31

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
7309	durée	supérieur		29	83	0.094	13.76	3.17
1350	temps	nécessaire		30	82	0.086	13.75	3.93
7169	paramètre	facultatif	*paramètre facultatif + paramètres facultatifs*	29	82	0.093	13.74	3.16
300	chiffre	décimal	*chiffres décimaux + chiffre décimal*	14	82	0.183	14.48	1.89
7083	abonné	appelé	*abonné appelé + abonnés appelés*	20	81	0.120	13.90	2.73
4811	bruit	ambiant	*bruit ambiant*	12	81	0.204	14.36	0.51
5812	service	semi-automatique		32	79	0.095	13.56	1.29
2539	transit	international	*transit international*	32	79	0.042	13.39	3.25
202	codage	unidimensionnel	*codage unidimensionnel*	11	79	0.238	14.75	0.73
792	titre	provisoire	*titre provisoire*	12	79	0.208	14.73	2.01
597	caractéristique	électrique		23	78	0.103	13.64	2.77
5365	secteur	mort	*secteur mort*	8	77	0.369	16.03	0.65
1158	valeur	maximal	*valeur maximale + valeurs maximales*	39	77	0.084	13.69	3.24
6830	longueur	total	*longueur totale + longueurs totales*	25	76	0.090	13.56	3.13
4697	message	subséquent	*messages subséquents + message subséquent*	25	76	0.100	13.45	1.81
4271	accord	mutuel	*accord mutuel*	12	76	0.198	14.42	1.59
4518	destination	difficile	*destination difficile + destinations difficiles*	10	76	0.254	14.82	0.79
8388	écouteur	couplé		7	76	0.528	16.49	0.38
7677	système	téléphonique		56	76	0.073	14.04	3.07
7408	temps	réel	*temps réel*	20	76	0.107	13.63	3.29
4197	recommandation	relatif	*recommandations relatives + recommandation relative*	31	75	0.067	13.48	3.12
3593	partie	émetteur	*partie émettrice*	18	75	0.118	13.70	2.31
6019	niveau	vocal	*niveau vocal + niveaux vocaux*	47	75	0.080	13.83	2.34
8524	instant	idéal	*instant idéal*	11	75	0.215	14.74	2.04
95	jeu	supplémentaire		24	75	0.059	13.36	2.39
868	impédance	nominal	*impédance nominale*	19	75	0.075	13.53	2.67
4403	gfu	préférentiel	*gfu préférentiel*	8	74	0.376	15.82	1.22
154	nom	générique	*nom générique + noms génériques*	10	73	0.211	14.89	1.76
914	échelon	national	*échelon national*	20	73	0.031	13.11	3.32
7604	activité	vocal	*activité vocale + activités vocales*	15	73	0.055	13.48	2.34
7931	essai	sinusoïdal	*essai sinusoïdal*	17	73	0.121	13.63	1.28
2239	circuit	téléphonique		91	72	0.084	14.75	3.07
2454	valeur	moyen	*valeur moyenne + valeurs moyennes*	36	72	0.081	13.51	2.82
5394	commutateur	intermédiaire	*commutateurs intermédiaires + commutateur intermédiaire*	28	72	0.088	13.37	2.17
7857	extrémité	virtuel	*extrémités virtuelles + extrémité virtuelle*	15	72	0.134	13.74	1.75
7188	signal	sinusoïdal	*signal sinusoïdal + signaux sinusoïdaux*	25	72	0.094	13.20	1.28
8607	sens	opposé	*sens opposé*	10	72	0.232	14.71	1.43
5615	indication	contraire	*indication contraire + indications contraires*	12	71	0.171	14.21	1.62
1890	point	équivalent	*point équivalent + points équivalents*	16	71	0.119	13.65	2.57
5430	réseau	téléphonique	*réseau téléphonique + réseaux téléphoniques*	80	71	0.080	14.51	3.07
527	circuit	télégraphique	*circuits télégraphiques + circuit télégraphique*	58	71	0.080	13.98	2.71
992	trajet	radioélectrique	*trajet radioélectrique*	15	70	0.115	13.70	2.17
1701	interfonctionnement	international	*interfonctionnement international*	24	70	0.020	12.89	3.25
4710	condition	applicable	*conditions applicables*	27	70	0.085	13.30	3.16
4275	affectation	automatique		15	70	0.030	13.14	3.77
1957	préfixe	interurbain	*préfixe interurbain + préfixes interurbains*	11	70	0.162	14.33	1.78
7455	contrôle	dynamique	*contrôle dynamique*	9	70	0.261	14.97	1.90
4024	concentration	numérique	*concentration numérique*	15	70	0.008	12.88	3.73
8733	pas	sûr	*pas sûr*	6	69	0.567	16.65	0.85
8374	numéro	incomplet	*numéro incomplet*	19	69	0.105	13.27	1.66
3164	type	différent	*type différent + types différents*	20	69	0.087	13.32	4.24
575	signal	électrique		32	69	0.084	13.27	2.77

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
5879	voie	radioélectrique	*voies radioélectriques + voie radioélectrique*	23	69	0.092	13.22	2.17
3880	signal	transmis		22	69	0.095	13.04	1.98
8263	commutateur	privé	*commutateur privé + commutateurs privés*	28	69	0.084	13.25	1.89
965	interface	numérique		33	68	0.057	13.28	3.73
6360	définition	formel	*définition formelle*	10	68	0.206	14.31	1.07
3722	transmission	satisfaisant	*transmission satisfaisante*	17	68	0.109	13.36	2.62
5318	signalisation	présent		20	68	0.099	13.11	2.37
8393	système	ouvert	*systèmes ouverts + système ouvert*	15	67	0.122	13.27	1.50
1906	limite	supérieur	*limite supérieure*	19	67	0.080	13.23	3.17
5181	utilisateur	terminal	*utilisateur terminal + utilisateurs terminaux*	20	67	0.056	13.10	2.01
5650	réseau	fixe	*réseaux fixes + réseau fixe*	30	66	0.082	13.15	2.60
4250	essai	automatique		32	66	0.063	13.21	3.77
836	bord	inférieur	*bord inférieur*	11	66	0.071	13.64	3.39
3909	connexion	complet		28	65	0.063	13.12	2.77
3047	version	numérique	*version numérique*	28	65	0.050	13.05	3.73
1911	impédance	complexe	*impédances complexes + impédance complexe*	10	65	0.182	14.21	1.91
6298	translation	régénérateur	*translation régénératrice + translations régénératrices*	6	64	0.489	16.17	0.74
2844	accent	grave	*accent grave*	8	64	0.267	14.77	1.10
5146	écran	complet	*écran complet*	13	64	0.023	12.99	2.77
5102	transmission	phototélégraphiques	*transmissions phototélégraphiques*	12	64	0.142	13.46	1.37
8338	façon	transparent	*façon transparente*	13	64	0.130	13.44	2.23
4682	fréquence	parasite	*fréquence parasite*	12	64	0.134	13.19	0.70
2623	centre	intermédiaire	*centres intermédiaires + centre intermédiaire*	28	64	0.079	13.07	2.17
6887	source	sonore	*source sonore + sources sonores*	9	64	0.203	14.45	1.97
2416	taille	moyen	*taille moyenne*	13	64	0.065	13.31	2.82
4587	information	supplémentaire	*informations supplémentaires + information supplémentaire*	43	63	0.066	13.44	2.39
4520	trame	fondamental	*trame fondamentale + trames fondamentales*	13	63	0.112	13.43	2.80
2532	mode	facultatif		21	63	0.073	13.00	3.16
5013	processeur	distant	*processeur distant*	11	62	0.086	13.58	2.73
5240	interligne	partiel	*interligne partiel*	8	62	0.138	14.36	2.69
485	cas	particulier	*cas particulier + cas particuliers*	26	62	0.072	13.01	4.15
5721	trait	noir	*traits noirs*	9	62	0.178	14.29	2.25
1720	république	socialiste	*république socialiste + républiques socialistes*	7	62	0.298	14.84	0.38
526	république	démocratique	*république démocratique*	7	62	0.298	14.84	0.38
4456	condition	requis	*conditions requises + condition requise*	16	62	0.102	12.95	2.37
7872	manière	suivant	*manière suivante + manières suivantes*	33	61	0.047	13.04	**4.43**
7679	polarité	permanent	*polarité permanente*	9	61	0.138	14.11	3.23
6597	répartition	spectral	*répartition spectrale*	8	61	0.191	14.55	1.93
1902	nombre	total	*nombre total*	24	61	0.073	12.94	3.13
522	signal	télégraphique	*signaux télégraphiques + signal télégraphique*	50	61	0.072	13.57	2.71
2346	indicatif	interurbain	*indicatif interurbain + indicatifs interurbains*	11	61	0.127	13.62	1.78
2681	distorsion	individuel	*distorsion individuelle*	14	61	0.109	13.11	2.70
1009	symbole	graphique	*symbole graphique + symboles graphiques*	14	60	0.061	13.04	1.06
2324	groupe	secondaire	*groupes secondaires + groupe secondaire*	12	60	0.128	13.28	2.34
5809	fonctionnement	intempestif	*fonctionnement intempestif + fonctionnements intempestifs*	15	60	0.103	12.89	2.16
5820	impulsion	bref	*impulsions brèves*	7	60	0.280	14.97	1.57
8604	compatibilité	descendant	*compatibilité descendante + compatibilités descendantes*	6	59	0.393	15.54	0.56
3999	courant	alternatif	*courant alternatif + courants alternatifs*	9	59	0.169	13.50	0.60

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
753	emplacement	physique	*emplacements physiques + emplacement physique*	10	58	0.087	13.49	2.88
2562	note	explicatif	*notes explicatives*	6	58	0.373	15.50	0.67
6636	livre	vert	*livre vert*	9	58	0.171	13.59	1.04
1260	terminal	élémentaire	*terminal élémentaire*	11	58	0.127	13.35	2.43
3181	usager	distant		17	58	0.077	12.83	2.73
3599	retard	excessif	*retard excessif + retards excessifs*	9	58	0.163	13.88	2.53
4393	utilisation	national		27	58	0.042	12.76	3.32
3267	disposition	relatif	*dispositions relatives*	25	58	0.049	12.75	3.12
7250	règle	applicable	*règles applicables + règle applicable*	16	57	0.074	12.82	3.16
92	trafic	frontalier	*trafic frontalier*	9	57	0.160	13.36	0.84
8282	extrémité	initiateur	*extrémité initiatrice*	8	56	0.176	13.47	0.35
2873	destination	inaccessible		8	56	0.192	13.85	1.18
5943	information	irrationnel	*information irrationnelle + informations irrationnelles*	14	56	0.099	12.70	1.77
528	circuit	interurbain	*circuit interurbain + circuits interurbains*	22	56	0.078	12.52	1.78
8513	façon	satisfaisant	*façon satisfaisante*	14	56	0.093	12.82	2.62
5032	motion	temporaire	*motion temporaire*	7	56	0.164	14.35	2.52
5336	conversion	longitudinal	*conversion longitudinale*	9	56	0.103	13.57	2.80
1656	fond	noir	*fond noir*	8	55	0.159	14.00	2.25
8319	règle	suivant	*règles suivantes + règle suivante*	29	55	0.039	12.72	**4.43**
7268	état	auxiliaire	*état auxiliaire + états auxiliaires*	14	55	0.095	12.77	2.08
3317	seconde/second	consécutif	*secondes consécutives*	7	55	0.143	14.22	3.12
4426	validation	positif	*validation positive*	8	55	0.108	13.78	2.34
7170	délai	maximal	*délai maximal + délais maximaux*	16	55	0.056	12.63	3.24
1788	espacement	vertical	*espacement vertical*	10	55	0.102	13.29	2.80
1542	mode	latin	*mode latin*	12	55	0.105	12.91	1.50
3086	test	favorable	*test est favorable*	6	54	0.309	15.00	1.12
3227	heure	actuel	*heure actuelle*	9	54	0.122	13.52	3.04
7866	crédit	initial	*crédit initial*	10	54	-0.000	12.60	2.43
102	symbole	monétaire	*symboles monétaires + symbole monétaire*	6	54	0.283	14.54	0.41
6811	extension	futur	*extensions futures + extension future*	8	53	0.122	13.72	2.71
6350	voix	humain	*voix humaine + voix humaines*	10	53	0.117	13.17	2.11
2230	envoi	unidirectionnel	*envoi unidirectionnel*	8	53	0.116	13.68	2.34
6532	formule	suivant	*formule suivante + formules suivantes*	15	53	-0.007	12.12	**4.43**
4562	utilisation	réussi	*utilisation réussie*	8	52	0.166	13.47	1.35
2319	caractéristique	fondamental	*caractéristiques fondamentales*	15	52	0.082	12.54	2.80
1313	affaiblissement	réglable	*affaiblissement réglable*	7	52	0.199	13.72	0.68
7143	bande	latéral	*bande latérale*	7	52	0.202	13.80	0.80
7206	zéro	préliminaire	*zéros préliminaires*	5	52	0.323	15.35	2.08
6816	récepteur	téléphonique	*récepteur téléphonique + récepteurs téléphoniques*	18	52	0.016	12.23	3.07
504	paragraphe	précédent	*paragraphes précédents + paragraphe précédent*	11	52	0.065	12.79	3.20
7005	sou-système	marqué	*sous-système marqué*	6	51	0.260	14.34	0.56
3019	essai	rapide	*essais rapides + essai rapide*	13	51	0.087	12.54	3.03
2684	résultat	complet	*résultat complet*	20	51	0.043	12.37	2.77
1941	point	nodal	*point nodal + points nodaux*	7	51	0.176	13.32	0.38
6551	intervention	manuel	*intervention manuelle*	8	51	0.060	13.21	2.93
2293	charge	moyen		14	51	0.051	12.45	2.82
6153	couche	sous-jacente		7	51	0.175	13.31	0.38
1835	chiffre	nécessaire	*chiffres nécessaires*	18	50	0.048	12.36	**3.93**
3397	numérotage	mondial	*numérotage mondial*	11	50	0.102	12.39	1.19
5123	fréquence	central	*fréquence centrale + fréquences centrales*	13	50	0.089	12.36	2.09
1006	principe	fondamental	*principes fondamentaux + principe fondamental*	11	50	0.086	12.70	2.80
5965	condition	suivant	*conditions suivantes + condition suivante*	50	50	0.055	13.29	**4.43**
7063	abonné	demandé	*abonné demandé*	10	49	0.110	12.71	1.47
8081	réserve	synchronisé		6	49	0.233	14.14	0.90
1211	valeur	minimal	*valeur minimale + valeurs minimales*	24	49	0.064	12.39	3.03

Valeur indéfinie du critère de vraisemblance

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
1961	accent	circonflexe	*accent circonflexe*	24	∞	0.649	18.52	0.00
7187	accroissement	égal	*accroissements égaux*	4	∞	-0.065	12.00	3.40
7331	acuité	auditif	*acuité auditive*	3	∞	0.109	13.80	1.71
5670	administration	exploitant	*administration exploitante + administrations exploitantes*	3	∞	0.188	12.62	0.00
59	annuaire	public	*annuaire public*	2	∞	-0.219	10.09	1.03
4092	appareil	imprimeurs	*appareils imprimeurs*	2	∞	0.059	8.97	0.00
4534	appel	malveillant	*appels malveillants + appel malveillant*	22	∞	0.207	15.12	0.00
1673	approche	probabiliste	*approche probabiliste*	4	∞	0.474	15.55	0.00
5595	argument	suivant	*arguments suivants*	4	∞	-0.188	8.86	**4.43**
7352	artère	principal	*artères principales*	2	∞	-0.190	10.65	2.54
8622	assemblée	plénier	*assemblée plénière + assemblées plénières*	6	∞	0.796	17.46	0.00
5302	bande	étroit		6	∞	0.230	13.87	0.00
2935	bande	passant	*bandes passantes + bande passante*	12	∞	0.349	15.87	0.00
2833	bruit	rose	*bruit rose*	8	∞	0.176	13.42	0.00
2957	bus	passif	*bus passif*	4	∞	0.209	14.63	1.94
7765	câble	sous-marins	*câbles sous-marins*	6	∞	0.521	16.24	0.00
2468	cabine	téléphonique	*cabine téléphonique + cabines téléphoniques*	16	∞	0.009	13.08	3.07
3833	cachet	dateur	*cachets dateurs + cachet dateur*	2	∞	0.646	15.87	0.00
6801	cas	échéant	*cas échéant*	123	∞	0.656	20.73	0.00
6554	cercle	concentrique	*cercles concentriques*	2	∞	0.373	14.29	0.00
1608	champ	diffus	*champ diffus*	4	∞	0.144	12.12	0.00
4311	champ	lointain	*champ lointain*	2	∞	0.088	10.12	0.00
523	choc	acoustique	*chocs acoustiques + choc acoustique*	5	∞	-0.064	11.90	2.74
7990	circuit	exploité	*circuits exploités*	3	∞	0.041	8.19	0.00
7796	circuit	loué	*circuit loué*	5	∞	0.057	9.67	0.00
201	codage	bidimensionnel		11	∞	0.286	15.19	0.00
8527	codeur	idéal	*codeur idéal*	3	∞	0.022	13.09	2.04
3238	communication	payable	*communications payables*	2	∞	0.056	8.82	0.00
416	compétence	national	*compétence nationale*	13	∞	-0.017	12.50	3.32
8333	composant	passé	*composants passés*	2	∞	0.183	12.23	0.00
1591	compresseur-extenseurs	syllabique	*compresseurs-extenseurs syllabiques*	2	∞	0.463	15.29	0.64
6923	compression-extension	logarithmique		2	∞	0.354	14.87	0.69
8699	comptabilité	international	*comptabilité internationale*	4	∞	-0.198	8.34	3.25
1522	compte	rendu	*compte rendu*	8	∞	0.260	14.55	0.00
5179	compte	tenu	*compte tenu*	66	∞	0.852	20.64	0.00
4981	condition	climatique	*conditions climatiques*	3	∞	0.065	9.55	0.00
5358	confusion	possible	*confusions possibles + confusion possible*	3	∞	-0.150	10.75	**4.02**
7353	connexité	numérique	*connexité numérique*	12	∞	-0.021	12.42	3.73
4219	contenu	informatif	*contenu informatif + contenus informatifs*	7	∞	0.429	15.85	0.00
7510	contrôle	audiométrique	*contrôle audiométrique*	3	∞	0.205	12.87	0.00
6401	couche	sous-jacentes	*couches sous-jacentes*	5	∞	0.154	12.53	0.00
2498	coupleur	acoustique	*coupleur acoustique*	2	∞	-0.253	9.25	2.74
1054	créneau	temporel	*créneaux temporels + créneau temporel*	9	∞	0.422	16.51	2.19
8529	décodeur	idéal	*décodeur idéal*	4	∞	0.109	13.92	2.04
1513	découpe	fonctionnel	*découpe fonctionnelle*	3	∞	-0.211	9.10	2.48
8057	délai	logistiques	*délais logistiques*	2	∞	0.108	10.70	0.00
8808	dernier	inchangé	*dernières sont transmises inchangées*	2	∞	0.409	14.55	0.00
3985	description	sommaire	*description sommaire*	2	∞	0.106	10.67	0.00
2506	diagramme	synoptique	*diagramme synoptique*	3	∞	0.371	14.59	0.00
7503	difficulté	sérieux	*difficultés sérieuses*	2	∞	0.305	13.70	0.00
6347	diffusion	local	*diffusion locale*	14	∞	0.023	13.34	3.34
5409	diminution	progressif	*diminutions progressives + diminution progressive*	3	∞	0.551	15.72	0.00
4320	dissymétrie	inférieur	*dissymétrie inférieure*	4	∞	-0.107	11.26	3.39

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
2673	distinction	net		2	∞	0.012	12.97	0.76
4376	distorsion	biais	*distorsion biaise*	17	∞	0.257	15.41	0.00
7845	diversité	double	*diversité double*	2	∞	-0.020	12.70	2.17
7082	donnée	attaché	*données attachées*	2	∞	0.075	9.66	0.00
2740	duplex	intégral	*duplex intégral*	3	∞	0.486	15.72	0.95
7502	durée	annuel	*durée annuelle*	2	∞	0.058	8.94	0.00
1515	écran	cathodique	*écran cathodique*	2	∞	0.222	12.79	0.00
1843	élément	intelligent	*élément intelligent*	2	∞	0.057	8.87	0.00
7795	enregistreur	oubliés	*enregistreurs oubliés*	2	∞	0.085	10.00	0.00
8025	ensemble	constitué	*ensemble constitué*	3	∞	0.200	12.80	0.00
8020	ensemble	ordonné	*ensemble ordonné*	4	∞	0.243	13.63	0.00
7549	entente	bilatéral	*entente bilatérale*	2	∞	-0.200	10.47	0.30
2669	espace	insécable	*espace insécable*	2	∞	0.215	12.70	0.00
6626	essai	comparatif	*essais comparatifs*	2	∞	0.067	9.34	0.00
6821	étude	approfondi	*étude approfondie*	2	∞	0.065	9.24	0.00
5690	exploitation	conjoint	*exploitation conjointe*	2	∞	0.059	8.95	0.00
3804	faisceau	hertzien	*faisceaux hertziens + faisceau hertzien*	4	∞	0.226	13.42	0.00
2735	faute	matériel	*faute matérielle*	38	∞	0.828	19.85	0.79
385	femme	artificiel	*femme artificielle*	2	∞	-0.248	9.37	1.45
769	fibre	optique	*fibres optiques*	2	∞	0.279	14.55	1.33
5730	fonction	utilisatrices	*fonctions utilisatrices*	2	∞	0.069	9.42	0.00
6525	force	électromoteur	*force électromotrice*	3	∞	0.616	16.04	0.00
7847	frontière	national	*frontières nationales*	3	∞	-0.230	8.26	3.32
4281	graduation	millimétrique	*graduation millimétrique*	2	∞	0.646	15.87	0.00
384	idéogramme	japonais	*idéogrammes chinois et japonais*	3	∞	0.371	14.59	0.00
6827	incertitude	total	*incertitude totale*	2	∞	-0.256	9.16	3.13
7196	inclinaison	latéral	*inclinaison latérale*	2	∞	0.094	13.55	0.80
1811	index	local	*index local*	3	∞	-0.216	8.89	3.34
2340	indicatif	exploitable	*indicatif exploitable*	2	∞	0.124	11.12	0.00
6059	initiateur	inconnu		5	∞	0.233	14.93	1.76
4054	isolation	acoustique	*isolation acoustique*	5	∞	-0.064	11.90	2.74
2236	jeton	disponible	*jetons disponibles*	2	∞	-0.232	9.79	3.64
8326	lèvre	virtuel	*lèvres virtuelles*	3	∞	-0.041	12.43	1.75
3129	laboratoire	différent	*laboratoires différents*	2	∞	-0.248	9.39	**4.24**
1665	langue	allemand	*langue allemande*	3	∞	0.290	13.87	0.00
1666	langue	espagnol	*langue espagnole*	3	∞	0.290	13.87	0.00
5448	lettre	capital	*lettres capitales*	2	∞	0.131	11.26	0.00
3695	lettre	majuscule		22	∞	0.599	18.18	0.00
5916	lettre	minuscule		15	∞	0.482	17.07	0.00
713	ligne	imaginaire	*ligne imaginaire*	2	∞	0.050	8.48	0.00
783	livre	orange	*livre orange*	6	∞	0.167	12.95	0.00
5490	logique	central	*logique centrale*	5	∞	0.149	14.35	2.09
1123	machine	parlant	*machine parlante*	5	∞	0.622	16.71	0.60
1201	main	libre	*mains libres*	2	∞	-0.228	9.90	2.28
8373	manière	incrémentielle	*manière incrémentielle*	2	∞	0.082	9.92	0.00
4497	matrice	pseudo-aléatoire	*matrice pseudo-aléatoire*	2	∞	0.346	14.07	0.00
8572	message	accusé	*message accusé*	8	∞	0.100	11.80	0.00
3035	mesure	lue	*mesure lue*	2	∞	0.076	9.71	0.00
1324	mode	conversationnel	*mode conversationnel*	2	∞	0.075	9.65	0.00
1335	mode	menu	*mode menu*	2	∞	0.075	9.65	0.00
1356	mode	opératoire	*mode opératoire + modes opératoires*	17	∞	0.296	15.82	0.00
1927	monde	extérieur	*monde extérieur*	2	∞	-0.071	12.23	2.51
8739	mot	clé	*mot clé*	3	∞	0.616	16.04	0.00
1	nation	uni	*nations unies*	3	∞	0.577	16.04	0.56
5191	nombre	pair	*nombre pair*	3	∞	0.078	10.07	0.00
8180	océan	indien	*océan indien*	4	∞	0.750	16.87	0.00
8732	occupation	injustifié		2	∞	0.264	13.29	0.00
1184	ohm	réactif	*ohms non réactifs*	2	∞	0.646	15.87	0.00
6807	onde	décamétriques	*ondes décamétriques*	4	∞	0.165	12.50	0.00
6826	onde	décimétriques	*ondes métriques et décimétriques*	9	∞	0.274	14.84	0.00
2630	onde	métrique	*ondes métriques*	17	∞	0.398	16.67	0.00
8556	opération	lancé	*opération lancée*	4	∞	0.161	12.43	0.00
2251	organisme	scientifique	*organisme scientifique + organismes scientifiques*	2	∞	0.323	13.87	0.00
298	période	probatoire	*période probatoire + périodes probatoires*	75	∞	0.800	20.63	0.00

Index	N	Adj	Expression R.	NC	LOG	FAG	MI3	h2
8130	périodicité	minimal	*périodicité minimale*	2	∞	-0.240	9.60	3.03
7539	paramètre	dernier	*paramètre dernier*	2	∞	0.060	8.99	0.00
3023	partie	intégrant	*partie intégrante*	9	∞	0.198	13.89	0.00
3829	plan	tangent	*plan tangent*	2	∞	0.114	10.87	0.00
8140	polynôme	générateur	*polynôme générateur*	11	∞	0.849	18.33	0.00
6638	poste	payeur	*poste payeur*	2	∞	0.066	9.31	0.00
5558	principale	caractéristique	*principales caractéristiques*	2	∞	-0.159	11.15	1.73
7982	privés	numérique	*privés numériques*	2	∞	-0.303	7.25	3.73
7827	probabilité	normal	*probabilité normale + probabilités normales*	2	∞	-0.291	7.89	3.91
1461	procédure	rattrapable	*procédure non rattrapable*	2	∞	0.055	8.78	0.00
114	prolongement	national	*prolongements nationaux*	2	∞	-0.306	7.09	3.32
8736	prothèse	auditif	*prothèses auditives + prothèse auditive*	6	∞	0.358	15.80	1.71
4064	puissance	psophométriques	*puissances psophométriques*	2	∞	0.089	10.15	0.00
984	qualificatif	supplémentaire	*qualificatif supplémentaire*	2	∞	-0.296	7.63	2.39
6747	réémetteur	télégraphique		2	∞	-0.282	8.26	2.71
3456	récepteur	perforateur	*récepteur perforateur + récepteurs perforateurs*	2	∞	0.143	11.52	0.00
5920	régulation	automatique	*régulation automatique*	17	∞	0.053	13.92	3.77
7573	république	fédéral	*république fédérale*	5	∞	0.262	14.06	0.00
1769	république	islamique	*république islamique*	3	∞	0.186	12.59	0.00
1717	république	populaire	*république populaire*	15	∞	0.508	17.23	0.00
8425	réseau	maillé	*réseau maillé*	7	∞	0.079	10.95	0.00
6587	réservation	sélectif	*réservation sélective*	8	∞	0.401	16.29	2.25
1813	refus	local	*refus local*	2	∞	-0.294	7.72	3.34
209	revêtement	dur	*revêtement dur*	2	∞	0.279	14.55	1.33
6317	séparation	fonctionnel	*séparation fonctionnelle*	3	∞	-0.211	9.10	2.48
3430	section	suivant	*sections suivantes + section suivante*	3	0	-0.044	3.17	**4.43**
5702	serveur	antérieur	*serveur antérieur*	2	∞	-0.065	12.29	2.67
4122	service	fourni	*service fourni*	2	∞	0.032	7.16	0.00
7832	seuil	dépassé	*seuil dépassé*	2	∞	0.195	12.42	0.00
1966	signal	assimilable		2	∞	0.033	7.25	0.00
7933	signal	déphasés	*signaux déphasés*	3	∞	0.044	8.42	0.00
981	signe	diacritique	*signes diacritiques + signe diacritique*	31	∞	0.755	19.29	0.00
7399	sonomètre	conforme	*sonomètre conforme*	2	∞	-0.222	10.02	3.75
3289	source	ponctuel	*source ponctuelle*	2	∞	0.155	11.75	0.00
6267	station	côtier	*station côtière + stations côtières*	18	∞	0.206	14.84	0.00
3500	station	terrien	*station terrienne + stations terriennes*	113	∞	0.557	20.14	0.00
7222	statique	égal	*statique égale*	4	∞	-0.065	12.00	3.40
5477	statistique	utile		2	∞	-0.059	12.35	2.77
7975	stylo	levé	*stylo levé*	3	∞	0.466	15.24	0.00
8685	stylo	posé	*stylo posé*	2	∞	0.346	14.07	0.00
6421	subdivision	fonctionnel	*subdivision fonctionnelle*	5	∞	-0.123	10.58	2.48
4622	substitution	correspondant	*substitution correspondante + substitutions correspondantes*	2	∞	-0.242	9.54	**4.21**
8216	syntaxe	général	*syntaxe générale*	4	∞	-0.173	9.48	2.86
1370	téléimprimeur	bilingue	*téléimprimeur bilingue + téléimprimeurs bilingues*	14	∞	0.648	17.85	0.00
1249	téléimprimeur	monolingue	*téléimprimeur monolingue*	3	∞	0.246	13.40	0.00
1726	tableau	récapitulatif	*tableau récapitulatif*	3	∞	0.329	14.24	0.00
8714	texte	communiqué	*texte communiqué*	7	∞	0.313	14.93	0.00
2967	tirant	parti	*tirant parti*	2	∞	0.463	15.29	0.64
794	tiret	mou	*tiret mou*	2	∞	0.457	14.87	0.00
7994	transfert	couronné	*transfert couronné*	2	∞	0.070	9.47	0.00
1134	transfert	intercellulaire	*transfert intercellulaire + transferts intercellulaires*	29	∞	0.376	17.19	0.00
4253	transmetteur	automatique	*transmetteur automatique*	2	∞	-0.294	7.74	3.77
6934	vibration	mécanique	*vibrations mécaniques*	2	∞	0.094	13.55	1.83
1399	vierge	britannique	*vierges britanniques*	2	∞	0.528	15.29	0.00
8679	vitesse	cadencé	*vitesse cadencée*	4	∞	0.320	14.42	0.00
864	vocabulaire	électrotechnique	*vocabulaire électrotechnique*	5	∞	0.567	16.52	0.66
97	voisinage	immédiat	*voisinage immédiat*	2	∞	-0.086	12.07	2.65
6471	voix	caverneux	*voix caverneuse*	3	∞	0.143	11.84	0.00
7536	voltmètre	vocal	*voltmètre vocal*	3	∞	-0.200	9.46	2.34